การตรวจเทียบภายนอกหาการลักลอกในงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความ

นายศุภวัจน์ แต่รุ่งเรือง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรดุษฎีบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2560

EXTRINSIC PLAGIARISM DETECTION IN ACADEMIC TEXTS USING A SUPPORT VECTOR MACHINE MODEL AND TEXT SIMILARITY MEASUREMENT

Mr. Supawat Taerungruang

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Linguistics
Department of Linguistics
Faculty of Arts
Chulalongkorn University
Academic Year 2017

หัวข้อวิทยานิพนธ์	การตรวจเทียบภายนอกหาการลักลอกในงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความ
โดย	นายศุภวัจน์ แต่รุ่งเรือง
สาขาวิชา	ภาษาศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาดุษฎีบัณฑิต

.....คณบดีคณะอักษรศาสตร์

(รองศาสตราจารย์ ดร.กิ่งกาญจน์ เทพกาญจนา)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.กิ่งกาญจน์ เทพกาญจนา)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล)

.....กรรมการ

(รองศาสตราจารย์ ดร.วิวัฒน์ วัฒนาวุฒิ)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.วิภาส โพธิแพทย์)

.....กรรมการภายนอกมหาวิทยาลัย

(ดร.เทพชัย ทรัพย์นิธิ)

ศุภวัจน์ แต่รุ่งเรือง : การตรวจเทียบภายนอกหาการลักลอกในงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความ (EXTRINSIC PLAGIARISM DETECTION IN ACADEMIC TEXTS USING A SUPPORT VECTOR MACHINE MODEL AND TEXT SIMILARITY MEASUREMENT) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร.วิโรจน์ อรุณมานะกุล, หน้า.

งานวิจัยชิ้นนี้มีวัตถุประสงค์ 4 ประการ ประการแรกคือ เพื่อวิเคราะห์หาลักษณ์ทางภาษาที่จะใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก ประการต่อมาคือ เพื่อพัฒนาระบบต้นแบบสำหรับตรวจเทียบภายนอกหาการลักลอกงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความ ประการที่ 3 คือ เพื่อประเมินประสิทธิภาพของระบบต้นแบบที่พัฒนาขึ้นใน 2 แง่มุม ได้แก่ ความเหมาะสมของลักษณะของข้อมูลรับเข้าที่จะใช้ในระบบ และความเหมาะสมของลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก และประการสุดท้ายคือ เพื่อเปรียบเทียบวิธีวัดค่าความละม้ายของข้อความที่มีประสิทธิภาพ เหมาะสมจะนำมาใช้ระบบตรวจหาการลักลอกมากที่สุด

ในด้านการดำเนินการวิจัย การศึกษาครั้งนี้ได้เพิ่มขั้นตอนเพื่อศึกษาเกี่ยวกับกลวิธีลักลอกงานวิชาการภาษาไทย โดยเก็บข้อมูลจากการจำลองสถานการณ์การลักลอกแล้วนำมาวิเคราะห์ด้วยแนวคิดทางภาษาศาสตร์ ผลจากการศึกษาในขั้นนี้ได้ถูกนำมาใช้ประโยชน์ในการออกแบบและสร้างคลังข้อมูล ตลอดจนนำมาใช้อ้างอิงในการอภิปรายข้อค้นพบในขั้นต่อไป นอกจากนี้ ยังมีการออกแบบ สร้าง และตรวจสอบคุณภาพของคลังข้อมูลด้วยความรอบคอบและรัดกุม เพื่อให้ผลการศึกษาที่ได้มาในตอนท้ายมีความหนักแน่นน่าเชื่อถือ

ผลการศึกษาในด้านการวิเคราะห์หาลักษณ์ทางภาษาสำหรับใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกปรากฏว่า สามารถวิเคราะห์หาลักษณ์ทางภาษาโดยอาศัยความรู้ทางภาษาศาสตร์มาประยุกต์เข้ากับวิธีการทางการประมวลภาษาธรรมชาติได้ทั้งหมด 51 ลักษณ์ ซึ่งแบ่งเป็นลักษณ์ทางศัพท์ 25 ลักษณ์ ลักษณ์ทางวากยสัมพันธ์ 23 ลักษณ์ ลักษณ์ทางความหมาย 2 ลักษณ์ และลักษณ์ทางวากยสัมพันธ์และความหมาย 1 ลักษณ์

ส่วนผลการศึกษาในด้านการประเมินประสิทธิภาพของระบบต้นแบบที่พัฒนาขึ้นนั้น ในแง่การประเมินประสิทธิภาพของระบบเมื่อใช้ข้อมูลรับเข้าที่ต่างกันปรากฏว่า เมื่อทดสอบการจำแนกประเภทข้อมูลการลักลอกทุกประเภทแล้ว ข้อมูลรับเข้าประเภทย่อหน้ามีความเหมาะสมที่ใช้ในระบบมากกว่าข้อมูลรับเข้าประเภทหน่วยปริจเฉทพื้นฐาน ส่วนในแง่การประเมินประสิทธิภาพของลักษณ์ ปรากฏว่าลักษณ์ที่ให้ประสิทธิภาพสูงที่สุดเป็นลักษณ์ทางศัพท์ คือลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (F = 0.9870) และเมื่อพิจารณาผลในภาพรวมแล้ว พบว่าลักษณ์ทางศัพท์และลักษณ์ทางอักขระให้ประสิทธิภาพสูงกว่าลักษณ์ทางวายสัมพันธ์และลักษณ์ทางความหมาย ทั้งนี้ สาเหตุหลักเป็นเพราะลักษณ์ทางศัพท์และลักษณ์ทางอักขระเป็นการแทนรูปคำและอักขระที่ชัดเจน ในขณะที่ลักษณ์ทางวากยสัมพันธ์และลักษณ์ทางความหมายเป็นการแทนรูปความสัมพันธ์ของหน่วยทางภาษาซึ่งมีความเป็นนามธรรมกว่า

ส่วนผลการเปรียบเทียบประสิทธิภาพของวิธีวัดค่าความละม้ายของข้อความ พบว่าค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำสามารถให้ค่าความละม้ายได้สอดคล้องกับค่าความละม้ายที่ให้โดยผู้เชี่ยวชาญทางภาษาไทยมากที่สุด (r = 0.9124) จึงถือว่าเป็นวิธีวัดค่าความละม้ายของข้อความที่มีประสิทธิภาพ สามารถนำมาใช้แทนการระบุค่าความละม้ายโดยมนุษย์ในระบบตรวจหาการลักลอกได้ สาเหตุที่ผลปรากฏเป็นเช่นนี้อาจเป็นเพราะผู้เชี่ยวชาญพิจารณาความละม้ายของข้อความจากลำดับของรูปคำเช่นเดียวกับวิธีการวัดค่าความละม้ายข้างต้น

ภาควิชา ภาษาศาสตร์ ลายมือชื่อนิสิต .....

สาขาวิชา ภาษาศาสตร์ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2560

5380193222 : MAJOR LINGUISTICS
KEYWORDS: PLAGIARISM / THAI PLAGIARISM DETECCTION / SUPPORT VECTOR MACHINE / TEXT SIMILARITY / NATURAL LANGUAGE PROCESSING

SUPAWAT TAERUNGRUANG: EXTRINSIC PLAGIARISM DETECTION IN ACADEMIC TEXTS USING A SUPPORT VECTOR MACHINE MODEL AND TEXT SIMILARITY MEASUREMENT. ADVISOR: ASSOC. PROF. WIROTE AROONMANAKUN, Ph.D., pp.

This research is based on 4 objectives: first, to analyze the linguistic features used to classify plagiarized text and non-plagiarized text. The next is to develop a prototype system for extrinsic academic plagiarism detecting using a support vector machine model and text similarity measurement. The third is to evaluate the effectiveness of the prototype system developed in 2 aspects: the suitability of the input characteristics to be used in the system and the suitability of the features used to classify plagiarized text and non-plagiarized text. And lastly, to compare the effectiveness of the text similarity measurement methods for use in the system.

In conducting this research, the analysis of plagiarism strategies in Thai academic texts, which collect data from the simulated plagiarism situation and analyzed them with linguistic concepts, is added in the research phase. The results of this analysis were used to design and construct a corpus. In addition, to make the final findings more credible, a corpus used for this research is also designed, created, and validated with care and circumspection.

The result of the analysis of linguistic features used to classify plagiarized text and non-plagiarized text shows that all 51 linguistic features are analyzed, based on linguistic knowledge applied to the methods of natural language processing, including 25 lexical features, 23 syntactic features, 2 semantic features, and 1 syntactic and semantic features.

For the results of the study on the effectiveness evaluation of the developed prototype system, in terms of the effectiveness of the input data, it is found that, when testing the classification of all types of plagiarized data, paragraph type input was more appropriate for the system than EDU type input. In terms of effectiveness of the features, it appears that the most effective feature is lexical feature i.e. Sørensen–Dice similarity coefficient of word bigram (F = 0.9870). Considering the overall results, lexical features and character features are more effective than syntactic features and semantic features. The main reason is that the lexical features and character features are derived from the representation of word and character form that is more tangible than syntactic features and semantic features, which derived from the representation of the linguistic relations.

And for the results of effectiveness evaluation of the text similarity measurement methods, it is found that the normalized longest common subsequence of word can calculated similarity correlated with Thai language experts the most (r = 0.9124). The reason for this may be because the experts consider the similarity of texts from the sequence of words, as well as the method of the normalized longest common subsequence of word.

Department: Linguistics
Field of Study: Linguistics
Academic Year: 2017

Student's Signature .....
Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยความกรุณาของรองศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล อาจารย์ที่ปรึกษา ที่ได้สละเวลาให้ความรู้และข้อคิดเห็นทั้งทางด้านวิชาการและการดำเนินงานวิจัย ตลอดระยะเวลาหลายปีของการศึกษา หากผู้วิจัยไม่ได้รับความเมตตาและความทุ่มเทเอาใจใส่จากอาจารย์ก็คงยากยิ่งที่จะทำวิทยานิพนธ์ฉบับนี้ให้สำเร็จได้ ผู้วิจัยขอกราบขอบพระคุณและรำลึกถึงพระคุณตลอดไป

ผู้วิจัยขอกราบขอบพระคุณรองศาสตราจารย์ ดร.กิ่งกาญจน์ เทพกาญจนา รองศาสตราจารย์ ดร.วิวัฒน์ วัฒนาวุฒิ ผู้ช่วยศาสตราจารย์ ดร.วิภาส โพธิแพทย์ และ ดร.เทพชัย ทรัพย์นิธิ ที่กรุณารับเป็นกรรมการสอบวิทยานิพนธ์ รวมถึงให้ข้อเสนอแนะและข้อแก้ไขอันมีประโยชน์ยิ่ง ยังผลให้วิทยานิพนธ์ฉบับนี้มีความถูกต้องและสมบูรณ์ยิ่งขึ้น

ขอกราบขอบพระคุณคณาจารย์ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย และคณาจารย์ภาควิชาภาษาไทย คณะมนุษยศาสตร์ มหาวิทยาลัยเชียงใหม่ ที่ประสิทธิ์ประสาทความรู้ทางภาษาศาสตร์และภาษาไทยอันเป็นรากฐานที่สำคัญยิ่งในการทำวิทยานิพนธ์ฉบับนี้

ขอขอบคุณบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย ที่ได้มอบทุนอุดหนุนการศึกษาเฉพาะค่าเล่าเรียนประเภท 60/40 และทุน 90 ปีจุฬาลงกรณ์มหาวิทยาลัย กองทุนรัชดาภิเษกสมโภช เพื่อสนับสนุนการทำวิจัยในวิทยานิพนธ์ฉบับนี้

ขอขอบคุณผู้จำลองข้อมูลการลักลอกทุกคนที่ได้ให้ความร่วมมือและสละเวลา ช่วยสร้างข้อมูลอันเป็นประโยชน์ยิ่งต่อการทดลองในวิทยานิพนธ์ฉบับนี้ ขอขอบคุณคุณธีรุตม์ สุขสกุลวัฒน์ ที่คอยให้คำปรึกษาและคำแนะนำเกี่ยวกับวิธีการทางสถิติ ซึ่งได้นำมาใช้ประโยชน์ในหลายขั้นตอนของการทำวิจัยในวิทยานิพนธ์ฉบับนี้

ขอขอบคุณคุณนลินี อินต๊ะชาว คุณศิวพร ทวนไธสง ดร.ธารทอง แจ่มไพบูลย์ ดร.นัชชา ถิระสาโรช และนิสิตร่วมที่ปรึกษาอีกหลายคนที่ไม่ได้เอ่ยนาม ที่ให้ความช่วยเหลือและช่วยแลกเปลี่ยนความเห็นในประเด็นต่างๆ มาโดยตลอด ขอขอบคุณเพื่อนและพี่น้องชาวภาษาศาสตร์ จุฬาฯ ทุกคนที่คอยให้กำลังใจตลอดระยะเวลาของการศึกษาและการทำวิทยานิพนธ์ โดยเฉพาะอย่างยิ่ง ดร.ชาฎินี มณีนาวาชัย ผู้คอยช่วยเหลือและให้ข้อเสนอแนะเรื่อยมาตั้งแต่ครั้งที่ผู้วิจัยพัฒนาหัวข้อวิทยานิพนธ์

และสุดท้าย ขอขอบคุณแม่และพี่สาวทั้งสองคนของผู้วิจัยที่สนับสนุนเส้นทางที่ผู้วิจัยเลือก และเฝ้ารอความสำเร็จของผู้วิจัยด้วยความอดทน

สารบัญ

หน้า

หน้า

หน้า

หน้า

หน้า

หน้า

หน้า

หน้า

สารบัญตาราง

หน้า

หน้า

สารบัญภาพ

หน้า

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบัน การลักลอกงานวิชาการ (plagiarism) ถือว่าเป็นปัญหาที่พึงตระหนักเป็นอย่างยิ่ง เพราะการกระทำดังกล่าวนอกจากจะทำให้ความน่าเชื่อถือของสถาบันการศึกษาลดลงแล้ว ยังส่งผลให้กระบวนการแลกเปลี่ยนเรียนรู้ในประชาคมวิชาการลดลงหรือหยุดลงได้อีกด้วย (บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2554, บทนำ) (บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2554, บทนำ) ยิ่งในยุคปัจจุบันที่เทคโนโลยีสารสนเทศได้รับการพัฒนาให้ก้าวหน้าด้วยแล้ว ปัญหาการลักลอกงานวิชาการก็ยิ่งทวีความรุนแรงมากขึ้นตามไปด้วย

ในการป้องปรามปัญหาข้างต้นได้มีผู้พยายามพัฒนาระบบตรวจหาการลักลอก (plagiarism detection system) ขึ้นเป็นจำนวนมากโดยอาศัยความรู้ทางวิทยาศาสตร์คอมพิวเตอร์และการประมวลผลภาษาธรรมชาติทำให้ประหยัดเวลาและทรัพยากรในการตรวจหาการลักลอกได้มาก อย่างไรก็ตาม ระบบตรวจหาการลักลอกที่ใช้อย่างแพร่หลายในปัจจุบันก็ยังไม่สามารถให้ผลได้ในระดับที่น่าพอใจ โดยเฉพาะในกรณีที่ข้อความถูกลักลอกโดยการเปลี่ยนลำดับของคำหรือวลี การถอดความ หรือการสรุปความ ทั้งนี้เนื่องมาจากระบบตรวจหาการลักลอกที่พัฒนาขึ้นโดยมากนั้นจะใช้วิธีตรวจหาแบบอิงสายอักขระ ซึ่งแม้จะเป็นวิธีที่มีขั้นตอนก่อนการประมวลผลไม่ซับซ้อนและไม่จำเป็นต้องประยุกต์ความรู้ทางภาษาศาสตร์ ทำให้ง่ายต่อการพัฒนาระบบ แต่ก็มีประสิทธิภาพตรวจหาในระดับสายอักขระพื้นผิวเท่านั้น (Alzahrani, Salim, & Abraham, 2012, p. 143) ยกตัวอย่างเช่นข้อความต่อไปนี้

> <u>ข้อความ (1)</u> สงคราม คือ ความขัดแย้งเป็นวงกว้าง และก่อให้เกิดผลกระทบอย่างร้ายแรง สงครามนั้นเกิดขึ้นเมื่อเกิดความขัดแย้งและไม่สามารถแก้ไขด้วยวิธีสันติ สุดท้ายจึงลงเอยด้วยการทำสงครามหรือการใช้กำลัง เพื่อลิดรอนหรือกำจัดบทบาททางการเมืองของรัฐอื่น สงครามนั้นเกิดขึ้นตลอดช่วงเวลาในประวัติศาสตร์ของมนุษยชาติ สงครามนั้นมีตั้งแต่ระดับ รัฐ ชาติและจักรวรรดิ[1]

[1] คัดลอกจาก http://th.wikipedia.org/wiki/สงคราม

> <u>ข้อความ (2)</u> สงคราม หมายถึง ความขัดแย้งในวงกว้าง ซึ่งก่อให้เกิดผลกระทบที่ร้ายแรง ทั้งนี้ สงครามจะเกิดได้เมื่อมีความขัดแย้งซึ่งแก้ไขไม่ได้โดยใช้สันติวิธี นำไปสู่การใช้กำลังหรือการก่อสงครามเพื่อลดทอนหรือขจัดบทบาททางการเมืองของประเทศอื่น ในประวัติศาสตร์ของมนุษยชาตินั้นสงครามเป็นสิ่งที่เกิดขึ้นเสมอมาโดยมีทั้งในระดับรัฐ ชาติ รวมถึงจักรวรรดิ

จากตัวอย่างข้างต้น ข้อความ (1) เป็นข้อความที่ผู้วิจัยคัดลอกมาจากอินเทอร์เน็ต ส่วนข้อความ (2) ผู้วิจัยได้ดัดแปลงจากข้อความ (1) โดยการแทรกคำ ลบคำ เปลี่ยนคำ และสลับตำแหน่งของคำและข้อความ เมื่อทดลองนำข้อความ (1) และข้อความ (2) ไปตรวจสอบความละม้ายกันของเอกสารในระบบ Anti-Kobpae[2] ซึ่งพัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ผลปรากฏว่าระบบตรวจพบว่าข้อความ (1) ละม้ายกับข้อความในเว็บไซต์อื่นๆ ในขณะที่ข้อความ (2) นั้น ระบบตรวจไม่พบว่ามีความละม้ายกับข้อความในเว็บไซต์ใด กรณีตัวอย่างนี้แสดงให้เห็นอย่างชัดเจนว่าเทคนิควิธีตรวจหาที่ใช้อยู่ยังไม่มีประสิทธิภาพเพียงพอที่จะใช้ตรวจหาการลักลอกบางประเภท

หากพิจารณาถึงเทคนิควิธีตรวจหาอื่นแล้ว ผู้วิจัยเห็นว่าการเรียนรู้ของเครื่องแบบมีการสอน (supervised machine learning) เป็นอีกเทคนิควิธีหนึ่งที่มีประสิทธิภาพเหมาะจะนำมาประยุกต์ใช้ในระบบตรวจหาการลักลอก เนื่องจากเทคนิควิธีนี้จะช่วยให้ระบบที่พัฒนาขึ้นสามารถเรียนรู้จากกรณีตัวอย่างที่จัดเตรียมให้และสร้างสมมติฐานทั่วไปเพื่อทำนายหรือตัดสินใจเกี่ยวกับกรณีตัวอย่างในอนาคตได้ (Kotsiantis, 2007, p. 249) ในงานวิจัยชิ้นนี้ ผู้วิจัยจึงได้เลือกใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines: SVMs) ซึ่งเป็นแบบจำลองการเรียนรู้แบบมีการสอน (Supervised learning model) ในการทำหน้าที่เป็นเสมือนตัวกรอง (filter) จำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกโดยตรวจเทียบกับข้อความต้นฉบับ ทั้งนี้ ในการแยกประเภทดังกล่าวจำเป็นต้องใช้ลักษณ์ต่างๆ ในการตัดสินใจ ดังนั้น ในงานวิจัยชิ้นนี้จึงต้องมีการวิเคราะห์หาลักษณ์ที่เหมาะสมเพื่อให้การทำงานของซัพพอร์ตเวกเตอร์แมชชีนเป็นไปอย่างมีประสิทธิภาพด้วย

ภารกิจขั้นต่อมาที่จำเป็นต้องมีในระบบเพื่อเพิ่มประสิทธิภาพนั่นก็คือ การระบุว่าข้อความต้องสงสัยมีปริมาณการลักลอกมากน้อยเพียงใด ในขั้นนี้ การวัดค่าความละม้ายของข้อความก็เป็นอีกแนวคิดหนึ่งที่เหมาะจะนำมาประยุกต์ใช้ เนื่องจากหลักในการทำงานของแนวคิดดังกล่าวนี้คือการระบุค่าเป็นตัวเลขว่าข้อความ 2 ข้อความในฐานะข้อมูลรับเข้ามีความสมมูลกันทางความหมาย (semantic equivalence) มากน้อยเพียงใด ด้วยหลักในทำงานดังกล่าวนี้จึงทำการวัดความละม้ายของข้อ

---

[2] ตรวจสอบเมื่อวันที่ 3 สิงหาคม 2555 ในเว็บไซต์ http://www.anti-kobpae.in.th/

ความถูกนำไปใช้ในโปรแกรมประยุกต์ต่างๆ อย่างแพร่หลาย ไม่ว่าจะเป็นการทำเหมืองข้อความ (Atkinson-Abutridy, Mellish, & Aitken, 2004) การค้นคืนสารสนเทศ (Feng, Zhou, & Martin, 2008; Ganguly, Leveling, & Jones, 2011; E.-K. Park, Ra, & Jang, 2005) ระบบคำถามคำตอบ (De Boni & Manandhar, 2003; Y. Li, Bandar, McLean, & O'Shea, 2004) การจำแนกประเภทข้อความ (Yang & Wen, 2007) การสรุปย่อข้อความ (Aliguliyev, 2009; Liang, Wang, & Huang, 2010; C.-Y. Lin & Hovy, 2003; P.-y. Zhang & Li, 2009) การแปลด้วยเครื่อง (Kauchak & Barzilay, 2006; Liu & Zong, 2004) ตลอดจนการประเมินผลอัตโนมัติของการแปลด้วยเครื่อง (Papineni, Roukos, Ward, & Zhu, 2002) และเช่นเดียวกับขั้นตอนการจำแนกประเภทส่วนของข้อความที่มีการลักลอกและไม่มีการลักลอก ในขั้นตอนการวัดค่าความละม้ายระหว่างส่วนของข้อความนี้ ผู้วิจัยก็จำเป็นต้องวิเคราะห์และเปรียบเทียบหาวิธีการวัดความละม้ายที่เหมาะสมกับข้อมูลที่มีการลักลอกให้มากที่สุด

ในแง่การประมวลผล ความยาวของส่วนที่เป็นข้อมูลรับเข้าเป็นปัจจัยสำคัญประการหนึ่งที่ต้องพิจารณา โดยทั่วไปแล้วนั้น การจำแนกประเภทและการวัดความละม้ายสามารถทำได้ตั้งแต่ระดับคำไปกระทั่งถึงระดับเอกสาร อย่างไรก็ตาม งานที่เกี่ยวข้องในระยะหลัง ไม่ว่าจะเป็นการรู้จำการถอดความหรือการสรุปย่อข้อความจะนิยมใช้ข้อมูลรับเข้าที่เป็นประโยคเป็นส่วนใหญ่ ด้วยเหตุผลที่ว่าข้อมูลรับเข้าที่เป็นประโยคซึ่งมีใจความสมบูรณ์นั้นจะมีขนาดสั้นกว่าข้อมูลรับเข้าที่เป็นบรรทัดหรือย่อหน้า ทำให้ง่ายต่อการประมวลผลมากกว่า เนื่องจากประโยคเป็นตัวแทนของหน่วยทางความหมายที่เป็นธรรมชาติมากกว่าข้อความขนาดยาว (Ganguly et al., 2011, p. 63) นอกจากนั้นแล้ว ประโยคหนึ่งๆนั้นก็มีแนวโน้มจะเกี่ยวข้องกับหัวเรื่องใดหัวเรื่องหนึ่งโดยเฉพาะ ในขณะที่ข้อมูลระดับบรรทัด ย่อหน้า หรือทั้งเอกสาร มักเกี่ยวข้องกับหัวเรื่องที่หลากหลาย (Feng et al., 2008, p. 833) อย่างไรก็ตาม ในภาษาไทยนั้นไม่สามารถระบุขอบเขตของประโยคได้ชัดเจน ในงานวิจัยชิ้นนี้ ผู้วิจัยจึงยึดแนวคิดตามทฤษฎีโครงสร้างวาทะ (rhetorical structure theory: RST) ของแมนน์และทอมป์สัน (Mann & Thompson, 1988) ที่มองว่าปริจเฉทประกอบไปด้วยหน่วยย่อยๆ เรียกว่าหน่วยปริจเฉทพื้นฐาน (elementary discourse units: EDUs) ซึ่งเป็นหน่วยสร้างที่เล็กที่สุดในปริจเฉทที่มีข้อมูลเชิงเนื้อหาและความหมาย และจะได้เลือกใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้าในการประมวลแทนประโยค

ทั้งนี้ แนวคิดในทฤษฎีโครงสร้างวาทะได้ถูกประยุกต์ใช้งานกับภาษาไทยในหลายแขนง ไม่ว่าจะเป็นงานของธนา สุขวารี และคณะ (Sukvaree, Charoensuk, Wattanamethanont, & Kultrakul, 2004) ที่ได้เสนอการสรุปย่อข้อความ (text summarization) ในวงการเกษตรโดยอิงทฤษฎีโครงสร้างวาทะ ในขณะที่จิรวรรณ เจริญสุข (2549) ได้เสนอวิธีการตัดอนุพากย์ด้วยวิธีการผสม

ระหว่างการเรียนรู้ด้วยเครื่องและการใช้กฎ โดยกำหนดขอบเขตหน่วยปริจเฉทพื้นฐานในภาษาไทย โดยอิงทฤษฎีโครงสร้างวาทะ หรืองานของเมทินี วัฒนะเมธานนท์ และคณะ (Wattanamethanont, Suvakree, & Kawtrakul, 2005) ที่เสนอวิธีการระบุความสัมพันธ์ในปริจเฉทโดยใช้แบบจำลองนาอีฟเบย์เป็นตัวแยกประเภทจากลักษณ์ที่ประกอบไปด้วยตัวบ่งชี้ความสัมพันธ์ในปริจเฉท (discourse relation marker) วลีสำคัญ (key phrase) และการเกิดร่วมของคำ นอกจากนี้ยังมีงานของสมนึก สินธุปวน และโอม ศรนิล (Sinthupoun & Sornil, 2010) ที่ได้เสนอแนวทางการวิเคราะห์โครงสร้างวาทะ (rhetorical structure) ในภาษาไทยโดยได้ใช้แบบจำลองฮิดเดนมาร์คอฟในการตัดแบ่งหน่วยปริจเฉทพื้นฐานด้วย

นอกจากหน่วยปริจเฉทพื้นฐานแล้ว ข้อความอีกประเภทหนึ่งที่ผู้วิจัยสนใจเลือกใช้ในงานชิ้นนี้คือย่อหน้า (paragraph) ทั้งนี้เนื่องมาจากโดยธรรมชาติของการลักลอกแล้วอาจมีการตัดหรือยุบรวมข้อความในระดับประโยคให้กลายเป็นหน่วยอื่นๆในระดับย่อหน้าได้ จึงน่าสนใจว่าข้อมูลรับเข้าที่เป็นประโยคจะให้ประสิทธิภาพในการจำแนกประเภทและการวัดความละม้ายได้ดีหรือไม่ในกรณีของการลักลอก ด้วยเหตุนี้ ในงานชิ้นนี้ผู้วิจัยจึงเลือกใช้ข้อความ (text) 2 ประเภทเป็นข้อมูลรับเข้า ได้แก่ ย่อหน้าและหน่วยปริจเฉทพื้นฐาน เพื่อประเมินว่าข้อความชนิดใดเป็นข้อมูลรับเข้าที่ให้ประสิทธิภาพดีที่สุดในระบบ โดยในการประเมินนั้น ผู้วิจัยได้แบ่งประเภทของการลักลอกงานวิชาการออกเป็น 4 ประเภทตามความยากของการตรวจหา (difficulty of detection) และระดับความคลุมเครือ (degree of obfuscation) ได้แก่ การคัดลอกโดยตรง (exact copy) การคัดลอกโดยใกล้เคียง (near copy) การคัดลอกโดยดัดแปลง (modified copy) และการถอดความ (paraphrase) ทั้งนี้ เพื่อจะได้ประเมินผลการตรวจหาในการลักลอกประเภทต่างๆ ด้วย

ด้วยเหตุผลดังได้กล่าวมาข้างต้น ในงานชิ้นนี้ ผู้วิจัยจึงสนใจพัฒนาระบบตรวจเทียบภายนอกหาการลักลอกงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนร่วมกับการวัดความละม้ายของข้อความ โดยมีการวิเคราะห์หาลักษณ์ที่จะใช้ในแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและวิเคราะห์หาวิธีวัดความละม้ายที่เหมาะสม รวมทั้งวิเคราะห์หาประเภทข้อมูลรับเข้าที่เหมาะสมกับระบบที่พัฒนาขึ้นด้วย ซึ่งเป็นไปได้ว่าระบบที่พัฒนาขึ้นในงานชิ้นนี้จะมีประสิทธิภาพในการตรวจหาและให้ผลที่น่าพึงพอใจกว่าวิธีตรวจหาแบบอิงสายอักขระ ทั้งนี้ ผลจากงานวิจัยชิ้นนี้นอกจากจะแสดงถึงการประยุกต์ใช้ความรู้ทางภาษาศาสตร์ให้เกิดประโยชน์อย่างเป็นรูปธรรมแล้ว ยังเป็นแนวทางในการพัฒนาระบบตรวจหาการลักลอกให้มีประสิทธิภาพต่อไปด้วย

1.2 วัตถุประสงค์ของการวิจัย

1) วิเคราะห์หาลักษณ์ทางภาษาที่จะใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก
2) พัฒนาระบบต้นแบบตรวจเทียบภายนอกหาการลักลอกงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความ
3) ประเมินประสิทธิภาพของระบบตรวจเทียบภายนอกหาการลักลอกงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความที่พัฒนาขึ้นในประเด็นต่อไปนี้
   3.1) ประเภทข้อมูลรับเข้าที่เหมาะสมจะใช้ในการตรวจหาการลักลอกแต่ละประเภท
   3.2) ลักษณ์ที่เหมาะสมจะใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก
4) เปรียบเทียบวิธีวัดค่าความละม้ายของข้อความเพื่อหาวิธีวัดที่มีประสิทธิภาพมากที่สุด

1.3 สมมติฐาน

1) ลักษณ์ทางภาษาที่สามารถนำมาใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก มีดังต่อไปนี้
   1.1) ลักษณ์ทางความหมาย ได้แก่ ค่าความละม้ายจากเวกเตอร์ทางความหมาย
   1.2) ลักษณ์ทางวากยสัมพันธ์ ได้แก่ POS common subsequence, POS trigram, และความสัมพันธ์แบบพึ่งพาในข้อความ
   1.3) ลักษณ์ทางศัพท์ ได้แก่ ลำดับเหมือนที่ยาวที่สุดในชุดคำ (longest common subsequence: *lcs*), ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ด (Jaccard similarity coefficient), และไตรแกรมของคำ
2) ข้อมูลรับเข้าที่เป็นหน่วยปริจเฉทพื้นฐานสามารถตรวจหาการลักลอกแบบคัดลอกโดยตรง การลักลอกแบบคัดลอกโดยใกล้เคียง และการลักลอกแบบคัดลอกโดยดัดแปลง ได้ดีกว่าข้อมูลรับเข้าที่เป็นย่อหน้า ในขณะที่ข้อมูลรับเข้าที่เป็นย่อหน้าสามารถตรวจหาการลักลอกแบบถอดความได้ดีกว่าข้อมูลรับเข้าที่เป็นหน่วยปริจเฉทพื้นฐาน
3) ลักษณ์ที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับลึกจะมีประสิทธิภาพในการตรวจหาการลักลอกได้ดีกว่าลักษณ์ที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับผิว และลักษณ์ที่ไม่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ ตามลำดับ

4) วิธีการวัดค่าความละม้ายที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับลึกจะให้ผลดีกว่าวิธีการวัดที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับผิว

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) เป็นแนวทางในการพัฒนาระบบตรวจหาการลักลอกเพื่อนำไปใช้จริง
2) เป็นแนวทางการประยุกต์ใช้ความรู้ทางภาษาศาสตร์ให้เกิดประโยชน์อย่างเป็นรูปธรรม

1.5 เครื่องมือที่ใช้ในการวิจัย

1) โปรแกรมภาษาไพทอน (Python) เวอร์ชัน 3.6.3 พัฒนาโดยมูลนิธิซอฟต์แวร์ไพทอน (Python Software Foundation: PSF)
2) Anaconda Distribution เวอร์ชัน 5.1.0 พัฒนาโดย Anaconda, Inc.
3) ซอฟต์แวร์ไลบรารีภาษาไพทอน scikit-learn เวอร์ชัน 0.19.1 (Pedregosa et al., 2011)
4) ชุดเครื่องมือ Gensim เวอร์ชัน 3.3.0 พัฒนาโดยบริษัท RaRe Technologies
5) แพ็กเกจภาษาไพทอน tltk เวอร์ชัน 0.3.5 พัฒนาโดยวิโรจน์ อรุณมานะกุล

1.6 นิยามคำศัพท์

เพื่อความเข้าใจที่ตรงกัน ผู้วิจัยจึงได้นิยามความหมายของคำศัพท์ซึ่งเป็นคำสำคัญที่ใช้ในงานวิจัยชิ้นนี้ รายละเอียดมีดังต่อไปนี้

1) การลักลอก (plagiarism)

การลักลอก หมายถึง การนำผลงาน ข้อเขียน คำพูด ความคิด หรือการแสดงเนื้อหาออกมาในรูปแบบอื่นๆ ไม่ว่าจะเป็นรูปภาพ แผนภูมิ ตาราง สมการ หรือสิ่งอื่นใด ทั้งที่เป็นของผู้อื่นหรือเป็นผลผลิตเก่าของตนเองมานำเสนอในลักษณะเสมือนว่าเป็นผลผลิตใหม่ของตน โดยมิได้แสดงการรับรู้ความเป็นเจ้าของผลงานและระบุถึงแหล่งที่มาของผลงานหรือความคิดนั้นๆ ให้ชัดเจนด้วยวิธีการอ้างอิงที่เป็นที่ยอมรับโดยทั่วไป

2) การตรวจเทียบภายนอกหาการลักลอก (extrinsic plagiarism detection)

การตรวจเทียบภายนอกหาการลักลอกเป็นแนวทางหนึ่งของการตรวจหาการลักลอกงานวิชาการด้วยเครื่อง การตรวจหาการลักลอกด้วยแนวทางนี้ เครื่องจะเปรียบเทียบเอกสารที่ต้องสงสัยว่าเป็นการลักลอกกับข้อมูลเอกสารภายนอกอื่นๆ ที่ต้องสงสัย ข้อมูลดังกล่าวอาจมาจากแหล่งเดียวหรือเกิดจากการรวบรวมจากที่มาหลายแหล่งก็ได้ (Alzahrani et al., 2012, p. 137) ด้วยแนวคิดนี้

ระบบจะประเมินว่าเอกสารที่ต้องสงสัยมีความละม้ายกับข้อมูลที่นำมาเปรียบเทียบหรือไม่ อย่างไร จากนั้นจึงรายงานผลดังกล่าวออกมา

3) การตรวจเทียบภายในหาการลักลอก (intrinsic plagiarism detection)

การตรวจเทียบภายนอกหาการลักลอกเป็นแนวทางหนึ่งของการตรวจหาการลักลอกงานวิชาการด้วยเครื่อง การตรวจหาการลักลอกด้วยแนวทางนี้ เครื่องจะวิเคราะห์ว่าลีลา (style) การเขียนที่ปรากฏภายในเอกสารที่ต้องสงสัยมีการเปลี่ยนแปลงหรือไม่ โดยอาศัยแนวคิดพื้นฐานว่าเอกสารที่เขียนโดยผู้เขียนคนเดียวกันย่อมมีลีลาการเขียนที่คงที่ในระดับหนึ่ง (Stein, Lipka, & Prettenhofer, 2011, p. 64) ด้วยแนวคิดนี้ ระบบจะประเมินว่าเอกสารที่ต้องสงสัยนั้นเขียนโดยผู้เขียนคนเดียวกันตลอดทั้งเอกสารหรือไม่ จากนั้นจึงรายงานผลดังกล่าวออกมา

4) แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (support vector machines: SVMs)

แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนเป็นวิธีเรียนรู้เชิงสถิติของเครื่องที่ใช้ในการจำแนกประเภทข้อมูล 2 หรือหลายประเภทออกจากกันโดยอาศัยแนวคิดเรื่องปริภูมิเวกเตอร์ (vector space) และใช้ประโยชน์จากระนาบหลายมิติเพื่อสร้างเส้นแบ่งประเภทของข้อมูลที่ดีที่สุด (optimal separating hyperplane)

5) การเรียนรู้ของเครื่อง (machine learning)

การเรียนรู้ของเครื่องเป็นการศึกษาและการสร้างอัลกอริทึมที่สามารถเรียนรู้ชุดข้อมูลและทำนายข้อมูลได้ อัลกอริทึมของการเรียนรู้ของเครื่องจะต้องอาศัยแบบจำลองที่สร้างมาจากชุดข้อมูลรับเข้า เพื่อทำนายหรือตัดสินใจในภายหลัง แทนที่จะทำงานตามลำดับของคำสั่งโปรแกรมคอมพิวเตอร์

6) ลักษณ์ (feature)

ในการเรียนรู้ของเครื่องและการรู้จำรูปแบบ (pattern recognition) ลักษณ์คือคุณสมบัติหรือลักษณะ (characteristic) ของปรากฏการณ์ใดๆ ที่สามารถสังเกตและวัดค่าได้ ในงานวิจัยชิ้นนี้ลักษณ์มีบทบาทเป็นข้อมูลรับเข้าในการฝึกฝนและทดสอบประสิทธิภาพของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน โดยแบบจำลองจะใช้ลักษณ์ในการสร้างสมมติฐานทั่วไปเพื่อตัดสินแบ่งแยกประเภทของข้อมูล

7) การวัดค่าความละม้ายของข้อความ (text similarity measurement)

การวัดความละม้ายของข้อความ คือ การคำนวณเพื่อเปรียบเทียบและระบุว่าข้อความ 2 ข้อความใดๆ ในฐานะข้อมูลรับเข้ามีการสมมูลทางความหมาย (semantic equivalence) กันใน

ระดับใด โดยจะให้ค่าความละม้ายเป็นค่าตัวเลขตั้งแต่ 0 ถึง 1 ในกรณีที่ค่าความละม้ายเท่ากับ 0 หมายความข้อความทั้ง 2 ข้อความที่นำมาเปรียบเทียบกันไม่มีความละม้ายกันโดยสิ้นเชิง ในขณะที่หากค่าความละม้ายเท่ากับ 1 หมายความว่าข้อความทั้ง 2 ข้อความที่นำมาเปรียบเทียบกันเหมือนกันทุกประการ

8) คู่หน่วยเทียบ

คู่หน่วยเทียบ หมายถึง คู่ของหน่วยทางข้อความในระดับต่างๆ ที่อยู่ในระดับที่เท่ากัน ซึ่งต้องสงสัยว่าเป็นคู่ลักลอกของกันและกัน โดยข้อความหนึ่งเป็นข้อความที่ต้องสงสัยว่าเป็นข้อความต้นฉบับที่ใช้สำหรับการลักลอก ส่วนอีกข้อความหนึ่งเป็นข้อความที่ต้องสงสัยว่าเป็นข้อความที่เกิดจากการลักลอกข้อความต้นฉบับ ในงานวิจัยชิ้นนี้มีคู่หน่วยเทียบ 2 ประเภท ได้แก่ คู่หน่วยเทียบที่เป็นคู่ของย่อหน้า และคู่หน่วยเทียบที่เป็นคู่ของหน่วยปริจเฉทพื้นฐาน

9) คู่ลักลอก

คู่ลักลอก หมายถึง คู่ของหน่วยทางข้อความในระดับต่างๆ ที่อยู่ในระดับที่เท่ากัน ซึ่งพิสูจน์ได้ว่ามีการลักลอกเกิดขึ้น โดยข้อความหนึ่งเป็นข้อความต้นฉบับที่ใช้สำหรับการลักลอก ส่วนอีกข้อความหนึ่งเป็นข้อความที่เกิดจากการลักลอกข้อความต้นฉบับ

10) โครงสร้างวาทะ (rhetorical structure)

ตามทฤษฎีโครงสร้างวาทะฉบับของคาร์ลสันและคณะ (Carlson, Marcu, & Okurowski, 2001) โครงสร้างวาทะ คือ โครงสร้างทางปริจเฉทที่ประกอบขึ้นจากหน่วยปริจเฉทพื้นฐานหลายๆ หน่วย หน่วยปริจเฉทพื้นฐานแต่ละหน่วยในโครงสร้างวาทะจะมีสถานะความสำคัญแตกต่างกัน และมีความสัมพันธ์กับหน่วยปริจเฉทพื้นฐานอื่นๆ ภายในโครงสร้างวาทะเดียวกัน

11) หน่วยปริจเฉทพื้นฐาน (elementary discourse units: EDUs)

ตามทฤษฎีโครงสร้างวาทะฉบับของคาร์ลสันและคณะ (Carlson et al., 2001) หน่วยปริจเฉทพื้นฐาน คือ หน่วยที่เล็กที่สุดภายในโครงสร้างวาทะ ซึ่งสามารถสื่อข้อมูลเชิงเนื้อหาและความหมายได้สมบูรณ์ ทั้งนี้ หน่วยปริจเฉทพื้นฐานแต่ละหน่วยจะมีสถานะความสำคัญ (nuclearity status) และมีความสัมพันธ์กับหน่วยปริจเฉทพื้นฐานอื่นๆ ภายในโครงสร้างวาทะเดียวกัน

บทที่ 2
ทบทวนวรรณกรรม

งานวิจัยชิ้นนี้มีส่วนเกี่ยวข้องกับศาสตร์ต่างๆ หลายแขนง ไม่ว่าจะเป็นการลักลอกงานวิชาการ ประเภทและลักษณะของการลักลอกงานวิชาการ แนวทางตรวจหาการลักลอก วิธีตรวจเทียบภายนอกหาการลักลอก การตรวจวัดความละม้ายของข้อความ ทฤษฎีโครงสร้างวาทะ รวมถึงแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ในบทนี้ ผู้วิจัยได้ค้นคว้าและรวบรวมเอกสารและงานวิจัยที่เกี่ยวข้องกับเรื่องดังกล่าวไว้ เพื่อเป็นการเชื่อมโยงความรู้ทางวิชาการในอดีตกับงานวิจัยชิ้นนี้และเพื่อใช้กำหนดแนวทางในการดำเนินการวิจัย โดยได้จัดประเภทของทฤษฎีและงานวิจัยที่เกี่ยวข้องไว้ดังนี้

2.1 นิยามของการลักลอก

นิยามของการลักลอก (plagiarism) นั้นสามารถตีความได้หลากหลายตามปริบทที่เกี่ยวข้อง ด้วยเหตุนี้จึงไม่ปรากฏนิยามทั่วไปของการลักลอกที่สอดคล้องเหมาะสมกับทุกสถานการณ์ (Sutherland-Smith, 2008, p. 57) อย่างไรก็ดี จากการทบทวนวรรณกรรมสามารถแบ่งนิยามของการลักลอกได้เป็น 3 กลุ่ม ดังนี้

กลุ่มแรกเป็นกลุ่มที่นิยามการลักลอกงานวิชาการว่าเป็นการขโมยหรือโจรกรรมความคิด ผลงาน ข้อเขียน หรือคำพูดของผู้อื่นมาใช้ในงานของตัวเอง ผู้ที่นิยามการลักลอกในกลุ่มนี้จะเห็นว่างานวิชาการเป็นทรัพย์สินประเภทหนึ่งที่สามารถขโมยได้ ด้วยเหตุนี้จึงอาจพิจารณาได้ว่านิยามการลักลอกในกลุ่มแรกนี้ปรากฏเจตนาของผู้ลักลอกอย่างชัดเจน ดังที่ปรากฏในนิยามที่ให้โดยพาร์ก (C. Park, 2003, p. 472), รอสส์และโทมัส (Ross & Thomas, 2003, p. 211), ศรีคเณศห์และไอเยอร์ (Sriganesh & Iyer, 2007, p.146) และยง ภู่วรวรรณ (2553 อ้างถึงใน กัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์, 2554, น. 7) รวมถึงในพจนานุกรมจำนวนหนึ่งเช่น *The Compact Edition of the Oxford English Dictionary* (Henry, 1971, p. 2192) หรือศัพท์บัญญัติสาขาวรรณกรรม (ราชบัณฑิตยสถาน, 2545)

ส่วนกลุ่มที่สองนั้นได้ให้นิยามการลักลอกในเชิงการหลอกลวงและเสแสร้ง กล่าวคือพิจารณาว่าการลักลอกเป็นการนำความคิด ข้อเขียน หรือคำพูดของผู้อื่นมาใช้ในงานของตนโดยผู้ลักลอกมีเจตนาหลอกลวง เสแสร้ง หรือชักจูงให้ผู้อ่านเชื่อว่าเป็นประหนึ่งงานของผู้ลักลอกเอง นิยามของการลักลอกในแง่นี้ถูกใช้โดยแพร่หลายในวงการการศึกษา ดังได้ปรากฏในการให้นิยามของทิกเนอร์และฟิลด์ส (Ticknor & Fields, 1989 อ้างถึงใน บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2554, น. 5),

วอร์น (Warn, 2007, p. 196), เพโครารี (Pecorari, 2008, p. 4), บรีแทกและมาห์มุด (Bretag & Mahmud, 2009, p. 50) และวิไลวรรณ ศรีสงคราม (2554, น. 11)

นิยามกลุ่มสุดท้ายพิจารณาว่าการลักลอกเป็นการใช้ซ้ำซึ่งความคิด คำพูด หรือข้อเขียนของผู้อื่นอย่างไม่เหมาะสม การใช้ซ้ำอย่างไม่เหมาะสมในที่นี้หมายถึง การใช้ข้อมูลของผู้อื่นโดยไม่แสดงแหล่งที่มา หรือใช้ผลงานของผู้อื่นโดยไม่แสดงการรับรู้ (unacknowledged use) นิยามในแง่หลังนี้นี้เป็นที่แพร่หลายในงานวิชาการในยุคหลังๆ โดยเฉพาะอย่างยิ่งในสาขาเทคโนโลยีสารสนเทศ เนื่องจากการทำความเข้าใจนิยามในแง่นี้ไม่ต้องอาศัยการตีความเจตนาที่อยู่เบื้องหลังการกระทำของผู้ลักลอก ยกตัวอย่างเช่นนิยามที่ปรากฏในงานของสินธุ.แอลและคณะ (Sindhu.L, Thomas, & Idicula, 2011, p. 65), บาร์รอน-เซเดโญและคณะ (Barrón-Cedeño, Vila, Martí, & Rosso, 2013, p. 918), โรนัลด์และซูฮาร์จีโต (Ronald & Suharjito, 2014, p. 168) และโมฮ์ตัจญ์และคณะ (Mohtaj, Asghari, & Zarrabi, 2015, p. 1)

เมื่อพิจารณานิยามของการลักลอกทั้ง 3 กลุ่มข้างต้นแล้ว จะเห็นได้ว่าการลักลอกเป็นพฤติกรรมที่เกี่ยวข้องกับเจตนาและการให้เกียรติเจ้าของผลงาน อย่างไรก็ตาม ผู้วิจัยเห็นว่าในงานวิจัยชิ้นนี้ควรหลีกเลี่ยงการยึดเจตนาของการลักลอกเป็นส่วนหนึ่งของคำจำกัดความการลักลอก เนื่องจากเจตนาการลักลอกนั้นเป็นสิ่งที่ไม่อาจตรวจจับได้โดยง่าย ในทางตรงกันข้าม ผู้วิจัยเห็นว่าควรให้นิยามของการลักลอกในเชิงลักษณะของข้อความ (textual characteristics) มากกว่า ด้วยลักษณะทางข้อความนั้นสามารถพิจารณาด้วยวิธีการทางภาษาศาสตร์และการประมวลผลภาษาธรรมชาติได้

ด้วยเหตุผลที่กล่าวมา ในงานวิจัยชิ้นนี้ ผู้วิจัยจึงสรุปนิยามของการลักลอกขึ้นใหม่โดยอิงตามนิยามในกลุ่มที่ 3 ซึ่งมุ่งพิจารณาลักษณะทางข้อความและการแสดงการรับรู้ความเป็นเจ้าของผลงานมากกว่าการพิจารณาเจตนาของผู้ลักลอกว่า การลักลอก คือ การนำผลงาน ข้อเขียน คำพูด ความคิด หรือการแสดงเนื้อหาออกมาในรูปแบบอื่นๆ ไม่ว่าจะเป็นรูปภาพ แผนภูมิ ตาราง สมการ หรือสิ่งอื่นใด ทั้งที่เป็นของผู้อื่นหรือเป็นผลผลิตเก่าของตนเองมานำเสนอในลักษณะเสมือนว่าเป็นผลผลิตใหม่ของตน โดยมิได้แสดงการรับรู้ความเป็นเจ้าของผลงานและระบุถึงแหล่งที่มาของผลงานหรือความคิดนั้นๆ ให้ชัดเจนด้วยวิธีการอ้างอิงที่เป็นที่ยอมรับโดยทั่วไป

2.2 ประเภทและลักษณะของการลักลอกงานวิชาการ

ประเภทและลักษณะของการลักลอกงานวิชาการนั้นถือเป็นสิ่งสำคัญเบื้องต้นที่ต้องศึกษาเพื่อใช้ในการพัฒนาระบบการตรวจหาการลักลอกงานวิชาการ เนื่องจากความเข้าใจในลักษณะทางภาษาและรูปแบบของข้อมูลในข้อความที่ต้องการตรวจหาจะส่งผลให้สามารถเลือกใช้วิธีการตรวจหาที่

สอดคล้องกันได้ นอกจากนี้ ความรู้ในส่วนนี้ยังจำเป็นต่อการออกแบบข้อมูลเพื่อใช้ในการฝึกฝนและทดสอบระบบซึ่งจะส่งผลให้การประเมินประสิทธิภาพน่าเชื่อถือยิ่งขึ้นด้วย ในหัวข้อนี้ ผู้วิจัยได้รวบรวมการจัดแบ่งประเภทและลักษณะของการลักลอกไว้ดังนี้

การจัดประเภทและลักษณะของการลักลอกงานวิชาการนั้นมีจุดเริ่มต้นจากวงการการศึกษา โดยเป็นไปเพื่อศึกษากลวิธีลักลอกและช่วยให้ผู้ตรวจ (examiner) สามารถตรวจจับการลักลอกที่ปรากฏอยู่ในงานเขียนได้ ประเภทและลักษณะของการลักลอกตามมุมมองนี้จึงมักอิงกับกลวิธีการเขียนที่ใช้ในการลักลอกและเจตนาในการลักลอกของผู้เขียน

ทั้งนี้ สามารถยกตัวอย่างการจัดประเภทของการลักลอกที่อิงตามเจตนาของผู้ลักลอกได้ เช่น งานของเพโครารี (Pecorari, 2008, pp. 1-7) ซึ่งจัดประเภทลักลอกทางข้อความ (textual plagiarism) ไว้เป็น 2 ประเภท ได้แก่ การลักลอกต้นแบบ (prototypical plagiarism) และการเขียนแบบปะติดปะต่อ (patch writing) การลักลอกทั้ง 2 ประเภทนี้มีความแตกต่างกันในด้านเจตนาในการลักลอก กล่าวคือ ในกรณีของการลักลอกต้นแบบนั้น ผู้ลักลอกจะมีเจตนาปิดบังแหล่งที่มาของข้อมูล ลักษณะของข้อความจะปรากฏการลักลอกจากแหล่งข้อมูล โดยมีการลบคำทิ้ง มีการปรับเปลี่ยนโครงสร้างทางไวยากรณ์ การแทนที่คำหนึ่งด้วยคำไวพจน์อีกคำหนึ่ง ส่วนการเขียนแบบปะติดปะต่อนั้น ผู้เขียนไม่มีเจตนาจะปิดบังหรือหลอกลวง ลักษณะของข้อความที่สามารถสังเกตได้จะมีความละม้ายกับข้อมูลที่ปรากฏในต้นฉบับ แต่อาจขาดลักษณะบางประการในการเขียนไป เช่น ขาดเครื่องหมายปรัศนี เป็นต้น หรือกล่าวอีกนัยหนึ่งคือผู้ที่ลักลอกโดยเจตนานั้นย่อมมีความพยายามที่จะดัดแปลงแก้ไขข้อความอันเกิดจากการลักลอกให้แตกต่างไปจากต้นฉบับมากกว่าผู้ลักลอกโดยไม่เจตนา

จากการจัดประเภทข้างต้น แม้ว่าการลักลอกที่เกิดขึ้นโดยเจตนาจะปรากฏลักษณะที่สามารถตรวจจับได้ แต่รายละเอียดทางลักษณะของข้อความที่เพโครารีกล่าวไว้ก็ยังมีไม่มากพอที่จะใช้ตัดสินได้ว่าการลักลอกนั้นๆ เกิดขึ้นจากความตั้งใจหรือไม่ ทั้งนี้เป็นเพราะการเขียนแบบปะติดปะต่อนั้นก็อาจเกิดจากความตั้งใจจะลักลอกได้เช่นกัน

งานอีกหนึ่งชิ้นที่กล่าวถึงประเภทของการลักลอกในเชิงเจตนาคืองานของกัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์ (2554) งานชิ้นนี้ได้เสนอให้จัดแบ่งประเภทของการลักลอกงานวิชาการไว้โดยใช้เกณฑ์ 3 เกณฑ์ ได้แก่ จัดแบบตามระดับเจตนา จัดแบ่งตามแหล่งที่มาของข้อมูล และจัดแบ่งตามวิธีการ

ในส่วนของการจัดแบบตามระดับเจตนานั้น กัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์ (2554, น. 12) ได้จัดแบ่งการลักลอกงานวิชาการเป็น 2 ประเภทคือ การลักลอกงานวิชาการแบบ

ตั้งใจ อันเป็นการที่บุคคลตั้งใจหรือมีเจตนาจะลอกเลียนงานของผู้อื่นทั้งที่ทราบดีว่าเป็นความผิดทางวิชาการ และเสนองานนั้นเสมือนว่าเป็นงานของตนโดยปราศจากการอ้างอิงแหล่งที่มาที่ถูกต้อง ส่วนอีกประเภทหนึ่งคือ การลักลอกงานวิชาการแบบไม่ตั้งใจ ซึ่งหมายถึงการที่บุคคลไม่มีเจตนาจะลอกเลียนงานของผู้อื่น แต่ไม่เข้าใจขอบข่ายของการลักลอกว่าการกระทำใดที่ถือว่าเข้าข่ายการลักลอกตลอดจนอาจขาดความรู้เรื่องหลักการอ้างอิงที่ถูกต้อง สำหรับการจัดแบ่งประเภทด้วยวิธีนี้ ผู้วิจัยเห็นด้วยว่ามีความเป็นไปได้ที่ผู้ลักลอกจะกระทำการที่เข้าข่ายการลักลอกได้โดยไม่มีเจตนาตามที่กัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์ ได้เสนอไว้

ในประเด็นที่เกี่ยวข้องกับเจตนานี้ ผู้วิจัยเห็นว่าเจตนาการลักลอกโดยตั้งใจหรือไม่ตั้งใจนั้นสามารถสะท้อนผ่านลักษณะของข้อความ (textual feature) ได้ เช่น พฤติกรรมการลักลอกที่ใช้ความพยายามมากอย่างการถอดความหรือการสรุปความก็ย่อมปรากฏร่องรอยที่ชี้ให้เห็นว่าเกิดจากความตั้งใจลักลอก เช่น การเลือกหาคำไวพจน์มาแทนที่คำในต้นฉบับ การสลับตำแหน่งของประโยคที่มีประเด็นความคิดในต้นฉบับ การใช้คำผิดในตำแหน่งเดียวกันบ่อยครั้ง ลักษณะเหล่านี้สามารถพิจารณาได้ถึงลีลาการเขียนซึ่งเป็นลักษณะเฉพาะตัวของผู้ลักลอกที่ยังปรากฏอยู่โดยเกิดจากความตั้งใจ

อย่างไรก็ดี หากพิจารณาในแง่การศึกษาวิจัยเกี่ยวกับการลักลอกงานวิชาการ โดยเฉพาะการตรวจหาการลักลอกแล้ว จะเห็นได้ว่าวิธีจัดแบ่งประเภทของการลักลอกงานวิชาการตามเจตนานี้อาจทำให้เกิดปัญหาและข้อโต้แย้งได้ง่ายเมื่อเข้าสู่กระบวนการศึกษาวิจัยหรือกระบวนการตัดสิน ทั้งนี้เป็นเพราะประเภทของการลักลอกที่จัดโดยใช้มุมมองนี้มักมีลักษณะที่เหลื่อมซ้อน ไม่สามารถแยกขาดออกจากกันได้ (Bretag & Mahmud, 2009, p. 51) ด้วยเหตุนี้ การตรวจจับจากเจตนาหรือความตั้งใจของผู้ลักลอกนั้นจึงไม่สามารถทำได้โดยง่าย และอาจต้องอาศัยความรู้จากหลายสาขามาใช้ในการพิจารณา เช่น จิตวิทยา และภาษาศาสตร์ เป็นต้น จึงอาจกล่าวได้ว่าการตรวจหาการลักลอกจากเจตนานั้นน่าจะเป็นวิธีการที่เหมาะจะใช้โดยมนุษย์มากกว่าเครื่อง

ทั้งนี้ เมื่อกลับมาพิจารณาเกณฑ์การจัดแบ่งประเภทการลักลอกของกัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์ (2554) ที่เหลืออีก 2 เกณฑ์ ได้แก่ จัดแบ่งตามแหล่งที่มาของข้อมูล และจัดแบ่งตามวิธีการ แล้วก็จะพบว่าเกณฑ์ทั้งสองมีความน่าสนใจให้อภิปรายเช่นกัน

การจัดแบ่งประเภทของการลักลอกงานวิชาการตามแหล่งที่มาของข้อมูลนั้น กัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์ (2554, น. 10) ได้แบ่งย่อยเป็น 2 ประเภท ได้แก่ การลักลอกงานวิชาการของผู้อื่น และการลักลอกงานวิชาการของตนเอง (self-plagiarism) ในส่วนของการลักลอกงานวิชาการของผู้อื่นนั้นเป็นที่เข้าใจได้ว่าตรงกับนิยามของการลักลอกโดยทั่วไป ส่วนการลักลอกงาน

วิชาการของตนเองนั้นได้อธิบายว่าเป็นการนำงานบางส่วนของตนหรือเป็นงานของตนที่ผลิตร่วมกับผู้อื่นมาใช้ซ้ำให้ดูเหมือนเป็นงานใหม่โดยไม่ระบุว่าเคยปรากฏในแหล่งอื่นมาแล้ว ทั้งนี้ พฤติกรรมที่เข้าข่ายการลักลอกงานวิชาการของตนเองนั้นแบ่งย่อยได้ 3 ประเภท ตั้งแต่การตีพิมพ์เกินความจำเป็น กล่าวคือ นำผลงานของตนที่ตีพิมพ์แล้วมาตีพิมพ์ซ้ำโดยแก้ไขเล็กน้อย การตีพิมพ์ซ้ำ ได้แก่ การแก้ไขรายละเอียดในงานเล็กน้อย เช่น เปลี่ยนชื่อเรื่องหรือลำดับเนื้อหาแล้วส่งตีพิมพ์ในแหล่งต่างๆ พร้อมกันให้ดูเหมือนมีงานหลายชิ้น จนกระทั่งถึงการแบ่งซอยงานออกเป็นส่วนๆ ให้มีหลายชิ้น อย่างไรก็ตาม ในส่วนพฤติกรรมการลักลอกงานของตนเองที่กล่าวไว้นี้ ผู้วิจัยเห็นว่ายังมีความก้ำกึ่งกันอยู่ระหว่างประเภทย่อยต่างๆ เช่น การตีพิมพ์ซ้ำและการแบ่งซอย เป็นต้น

ส่วนการจัดแบ่งประเภทของการลักลอกงานวิชาการตามวิธีการนั้นก็น่าสนใจอยู่ไม่น้อย เนื่องจากได้จัดแบ่งประเภทย่อยของการลักลอกไว้ถึง 8 ประเภท (กัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์, 2554, น. 13) ดังนี้

1) การลักลอกแบบคำต่อคำ (word-by-word plagiarism หรือ verbatim หรือ outright copying)
2) การลักลอกแบบนำวลีหลายๆ วลีมาปะติดปะต่อกัน (patchwork plagiarism) โดยวลีนั้นอาจมีที่มาจากแหล่งเดียวกันหรือหลายแหล่งก็ได้
3) การแปลงานจากภาษาที่ปรากฏในต้นฉบับมาเป็นงานของตนเอง
4) การอ้างอิงและใส่ข้อมูลเท็จ คือการลอกข้อมูลมาแล้วอ้างอิงแหล่งที่มา แต่การอ้างอิงนั้นจงใจใส่ให้เป็นเท็จ
5) การถอดความ (paraphrase) คือการที่ผู้ลักลอกได้ถอดความหรือเปลี่ยนโครงสร้างของประโยคเดิมในแหล่งที่มาโดยยังคงความคิดของเจ้าของงานไว้
6) การนำข้อความจากแหล่งที่มามาใช้โดยไม่ค่อมอัญประกาศแต่ไม่ได้อ้างอิงถึงแหล่งที่มา
7) การนำโครงสร้างของประโยคในแหล่งที่มามาใช้โดยเปลี่ยนคำในประโยคและไม่ได้อ้างอิงแหล่งที่มา
8) การคัดลอกคำบางคำที่มีความหมายเหมาะสม (apt term) หรือนำแนวความคิดจากแหล่งอื่นมาใช้เป็นเนื้องานส่วนใหญ่ของตนเอง ไม่ว่าจะอ้างอิงแหล่งที่มาหรือไม่ก็ตาม

จากการจัดแบ่งประเภทการลักลอกงานวิชาการตามวิธีการข้างต้น จะเห็นได้ว่ามีความแตกต่างจากวิธีการจัดแบ่งประเภทจากที่กล่าวมาก่อนหน้านี้ กล่าวคือ การจัดแบ่งประเภทการลักลอกงานวิชาการตามวิธีการนี้ทำให้เห็นความเกี่ยวข้องกับลักษณะของภาษาและข้อความที่เกิดจากลักลอกมากกว่าวิธีการจัดแบ่งประเภทที่ผ่านมา อย่างไรก็ตาม เมื่อพิจารณาแล้วจะพบว่าการจัดแบ่งประเภทการลักลอกงานวิชาการตามวิธีการนี้ยังอาจนำไปใช้ประโยชน์ต่อได้ในแง่การตรวจหาการลักลอกได้ไม่

มากนัก เนื่องจากประเภทของการลักลอกยังมีลักษณะคลุมเครือและมีลักษณะเป็นอัตวิสัยสูง จำเป็นต้องใช้วิจารณญาณของผู้เชี่ยวชาญในการตัดสินว่าลักษณะที่ปรากฏเข้าข่ายการลักลอกหรือไม่ ยกตัวอย่างเช่น วิธีการที่ 8) ซึ่งต้องตีความว่าการนำแนวคิดจากผู้อื่นมาใช้เป็นเนื้อหานั้นต้องมีปริมาณมากน้อยเพียงใดจึงจะเข้าข่ายการลักลอกงานวิชาการ นอกจากนี้ ยังปรากฏประเภทของการลักลอกที่ทับซ้อน ได้แก่ วิธีการที่ 5) และวิธีการที่ 7) ซึ่งสามารถพิจารณาเป็นการถอดความได้ทั้งสองวิธี

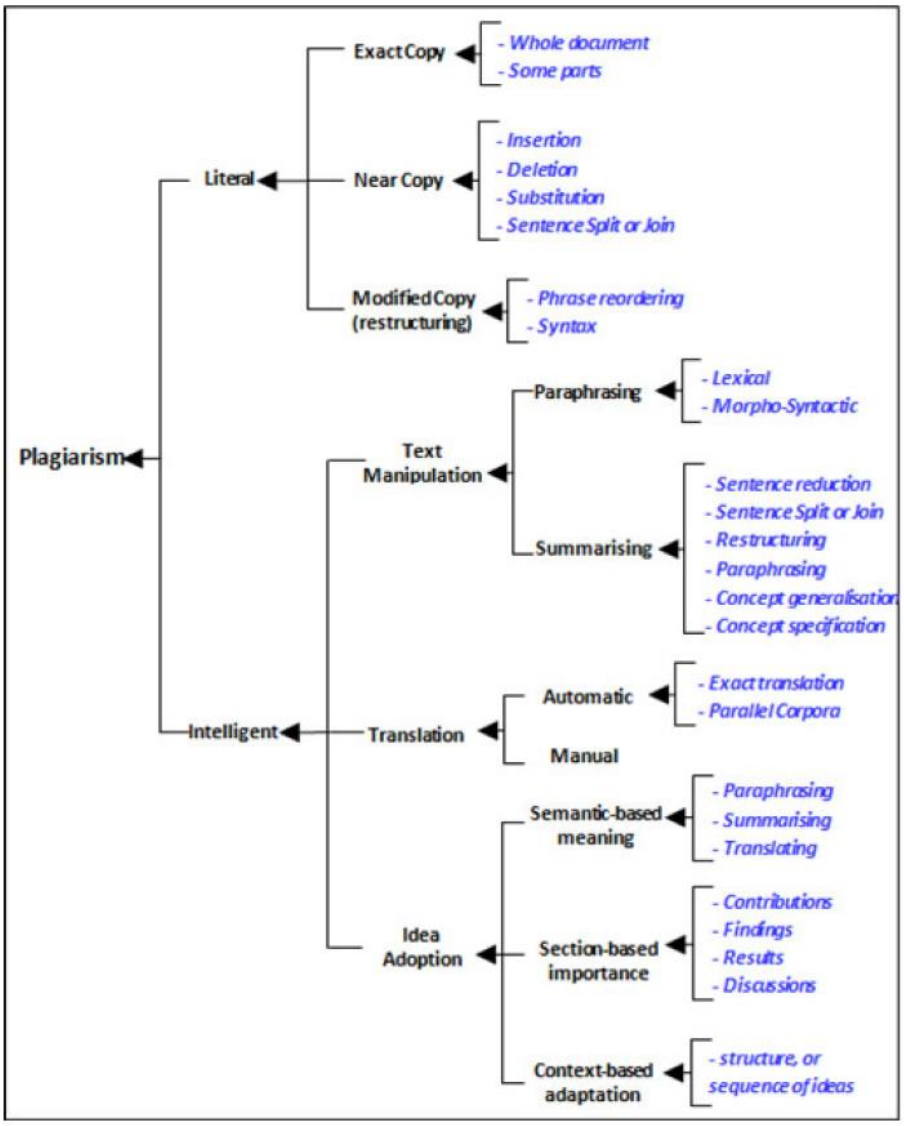
วิธีการจัดแบ่งประเภทการลักลอกงานวิชาการอีกวิธีหนึ่งที่มีรายละเอียดสอดคล้องเป็นไปในทางเดียวกันกับงานของกัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์ (2554) นั้นปรากฏอยู่ในงานของอัลซะฮ์รานิและคณะ (Alzahrani et al., 2012, p. 134) ที่ได้จัดแบ่งประเภทการลักลอกงานวิชาการโดยอิงกับพฤติกรรมการลักลอกของผู้ลักลอก เริ่มต้นจากเสนอให้แบ่งการลักลอกงานวิชาการออกเป็น 2 ประเภทใหญ่ ได้แก่ การลักลอกตามตัวอักษร (literal plagiarism) และการลักลอกความคิด (intelligent plagiarism)

ในส่วนของการลักลอกตามตัวอักษรนั้นเป็นวิธีการลักลอกที่ทำได้ง่ายและใช้เวลาไม่มาก ยกตัวอย่างเช่น การคัดลอกและวาง (copy & paste) จากอินเทอร์เน็ตที่ปรากฏในการลักลอกแบบคัดลอกโดยตรง (exact copy) การแทรกหรือลบคำบางคำปรากฏในการลักลอกแบบคัดลอกโดยใกล้เคียง (near copy) รวมถึงการสลับลำดับคำในวลีใหม่ที่ปรากฏในการลักลอกแบบคัดลอกโดยดัดแปลง (modified copy)

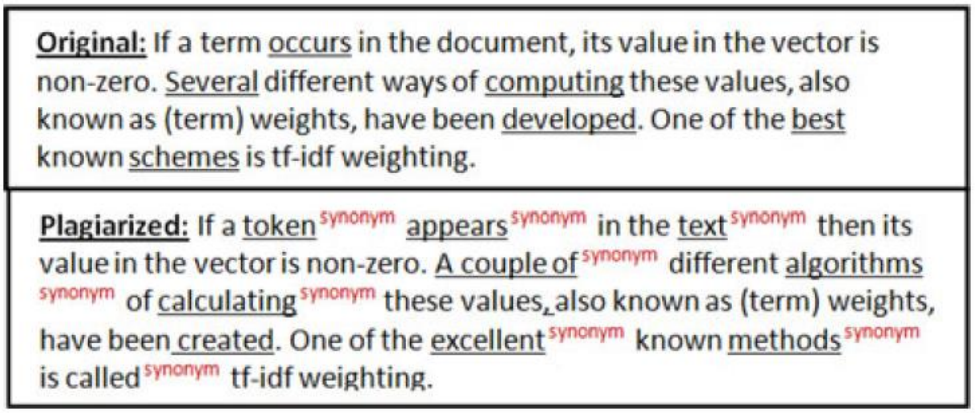
ส่วนการลักลอกความคิดนั้นเป็นการกระทำที่ตั้งใจจะให้ผู้อ่านเชื่อว่าข้อความเป็นผลงานของตนโดยเปลี่ยนแปลงข้อความไปจากต้นฉบับ วิธีการลักลอกประเภทนี้สามารถแบ่งย่อยได้อีก 3 วิธี วิธีแรกคือ การจัดการกับข้อความ (text manipulation) เช่น การถอดความและสรุปความข้อความซึ่งเป็นต้นฉบับ วิธีถัดมาคือ การแปล (translation) ข้อความจากภาษาหนึ่งเป็นภาษาอื่นโดยอาจใช้การแปลด้วยเครื่องหรือแปลด้วยตัวผู้ลักลอกเอง ส่วนอีกวิธีหนึ่งคือ การรับความคิด (idea adoption) ซึ่งถือเป็นการลักลอกที่ร้ายแรงและตรวจหายากมากที่สุด ในวิธีนี้ผู้ลักลอกจะเลือกเอาส่วนใดส่วนหนึ่งของผลงานมาใช้ไม่ว่าจะเป็นผลการวิจัย ข้อค้นพบ หรือข้อสรุป โดยวิธีการลักลอกก็เป็นไปเช่นเดียวกับการจัดการกับข้อความคือใช้วิธีถอดความและสรุปความเป็นหลัก

ภาพที่ 2.1 แสดงให้เห็นถึงภาพรวมของการจัดแบ่งประเภทการลักลอกงานวิชาการในงานของอัลซะฮ์รานิและคณะ (Alzahrani et al., 2012, pp. 133-137) เมื่อพิจารณาแล้วจะเห็นได้ว่างานชิ้นนี้ไม่เพียงจัดประเภทการลักลอกตามพฤติกรรมการลักลอกเท่านั้น แต่ยังให้รายละเอียดของวิธีการลักลอกที่เกี่ยวกับการแก้ไขลักษณะของข้อความและรูปแบบของภาษาในพฤติกรรมการลักลอกนั้นๆ ด้วย ทั้งนี้ จะสังเกตได้ว่างานชิ้นก็ได้จัดการถอดความเป็นพฤติกรรมหรือวิธีการลักลอกอย่างหนึ่ง

เช่นเดียวกับงานของ กัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์ (2554) อย่างไรก็ดี ในงานชิ้นนี้นั้นได้แสดงลักษณะการถอดความที่เข้าข่ายการลักลอกไว้อย่างชัดเจน ดังจะเห็นได้จากภาพที่ 2.2 ซึ่งแสดงให้เห็นถึงลักษณะของการถอดความที่แม้จะเป็นการเปลี่ยนแปลงด้านคำศัพท์และวากยสัมพันธ์ แต่ก็ยังคงไว้ซึ่งลำดับคำและลำดับโครงสร้างของงานต้นฉบับไว้มาก ผู้ลักลอกเพียงเปลี่ยนแค่คำไวพจน์และเพิ่มเติมคำเชื่อมในบางตำแหน่งเท่านั้น การถอดความในลักษณะนี้ก็ถือเป็นการถอดความที่ไม่เหมาะสมตามหลักการเขียนด้วย เมื่อเปรียบเทียบกับลักษณะของการถอดความที่เสนอโดยกัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์ (2554, น. 13) แล้ว จะเห็นได้ว่าอัลซะฮ์รานิและคณะได้แสดงตัวอย่างของการถอดความที่เข้าข่ายเป็นการลักลอกได้ชัดเจนกว่า



ภาพที่ 2.1 การจัดแบ่งประเภทการลักลอกที่เสนอโดยอัลซะฮ์รานิและคณะ (Alzahrani et al., 2012, p. 134)



ภาพที่ 2.2 รูปแบบของการลักลอกความคิด (การถอดความ) (Alzahrani et al., 2012, p. 135)

อีกประเด็นหนึ่งที่น่าสนใจในงานของอัลซะฮ์รานิและคณะนั้น ผู้วิจัยขอให้กลับมาพิจารณารูป 2.1 อีกครั้ง จะสังเกตเห็นได้ว่าการถอดความ (paraphrasing) นั้นปรากฏอยู่ในกิ่งต่างๆ ในแผนภาพถึง 3 กิ่งในลำดับชั้นที่แตกต่างกันตั้งแต่เป็นประเภทหนึ่งของการจัดการกับข้อความ (text manipulation) เป็นรายละเอียดวิธีการภายใต้การสรุปความ (summarizing) และเป็นรายละเอียดวิธีการของการรับความคิดในประเภทการเลือกโดยอิงความหมายอีกชั้นหนึ่ง ลักษณะดังกล่าวแสดงเห็นถึงลักษณะสำคัญประการหนึ่งของวิธีการลักลอก คือ การมีความสัมพันธ์แบบส่วนย่อย-ส่วนใหญ่ (part-whole relation) ระหว่างวิธีการลักลอกต่างๆ โดยวิธีการลักลอกในลำดับชั้นที่ลึกกว่าจะเป็นส่วนประกอบของวิธีการลักลอกในลำดับชั้นที่เหนือขึ้นไป ทั้งนี้ ความสัมพันธ์ดังกล่าวไม่ได้เกิดขึ้นภายในประเภทเดียวกันเท่านั้น แต่ยังเกิดขึ้นระหว่างประเภทย่อยของการลักลอกด้วย ดังจะเห็นได้ว่าการลักลอกที่ใช้ความสามารถสูงกว่าอย่างการรับความคิดนั้นก็มีการถอดความเป็นส่วนหนึ่งของวิธีการ รวมถึงการสรุปความที่เชื่อได้ว่าใช้ความสามารถในการเขียนมากกว่าการถอดความก็ยังปรากฏว่ามีการถอดความเป็นส่วนหนึ่งของวิธีการสรุปความ และหากพิจารณาจากมุมมองที่ว่าการถอดความเป็นทักษะที่ใช้ความสามารถสูงกว่าแล้ว การถอดความก็มีทักษะของการคัดลอกโดยใกล้เคียงและการคัดลักลอกโดยดัดแปลงเป็นส่วนหนึ่งของวิธีการถอดความด้วย ไม่ว่าจะเป็นการเพิ่มและลบคำในข้อความหรือการเรียงลำดับวลีในข้อความใหม่ และหากเป็นกรณีที่ข้อความต้นฉบับใช้ภาษาได้สื่อความหมายชัดเจนดีแล้ว ในการถอดความก็อาจเลือกคัดลอกข้อความบางส่วนจากต้นฉบับดังกล่าวโดยตรงได้

ความสัมพันธ์แบบส่วนประกอบที่มีในการลักลอกงานวิชาการประเภทต่างๆ ที่ปรากฏผ่านการจัดประเภทของอัลซะฮ์รานิและคณะนั้นได้แสดงให้เห็นว่าการลักลอกงานวิชาการแต่ละประเภทใช้ระดับความสามารถและทักษะในการลักลอกแตกต่างกัน ตั้งแต่สามารถทำได้ง่ายอย่างการคัดลอกโดยตรงจนกระทั่งถึงการลักลอกที่ต้องอาศัยทักษะการเขียนในขั้นสูงอย่างการรับความคิด หากพิจารณาในแง่ของการตรวจหาการลักลอกในงานวิชาการแล้ว ก็เป็นที่แน่นอนว่าการลักลอกประเภท

ที่ใช้ความสามารถในการลักลอกระดับต่ำย่อมมีความยากของการตรวจหา (difficulty of detection) ต่ำกว่าการลักลอกประเภทที่ใช้ความสามารถในการลักลอกในระดับสูง หรืออาจกล่าวเปรียบเทียบให้เข้าใจโดยง่ายได้ว่าการลักลอกด้วยวิธีคัดลอกโดยตรงย่อมถูกตรวจหาได้ง่ายกว่าการลักลอกด้วยวิธีถอดความนั่นเอง เหตุผลที่เป็นเช่นนี้เพราะการลักลอกประเภทที่ใช้ความสามารถในการลักลอกต่ำนั้นจะทำให้ข้อความที่ได้จากการลักลอกมีระดับความคลุมเครือ (degree of obfuscation) ต่ำไปด้วยเช่นเดียวกัน

ทั้งนี้ ระดับความคลุมเครือถือเป็นตัวชี้วัดที่สำคัญตัวหนึ่งที่ใช้ในการประเมินประสิทธิภาพของของระบบตรวจหาการลักลอกงานวิชาการ กล่าวคือ ยิ่งระบบตรวจหาการลักลอกสามารถตรวจหาการลักลอกในข้อความที่มีระดับความคลุมเครือได้ผลมากเท่าไรก็ยิ่งแสดงว่าระบบดังกล่าวมีประสิทธิภาพเท่านั้น ด้วยเหตุนี้ ระดับความคลุมเครือจึงเป็นปัจจัยหนึ่งที่สัมพันธ์สอดคล้องไปกับประเภทและรูปแบบทางภาษาของการลักลอก และยังถือเป็นปัจจัยที่ผู้พัฒนาระบบตรวจหาการลักลอกต้องคำนึงถึงเมื่อสร้างคลังข้อมูลฝึกฝนหรือทดสอบด้วย

งานกลุ่มหนึ่งที่แสดงความสัมพันธ์ระหว่างประเภทและรูปแบบทางภาษาของการลักลอกงานวิชาการกับระดับความคลุมเครือของข้อความได้แก่งานของพอตแทสต์และคณะ (Potthast, Barrón-Cedeño, Stein, & Rosso, 2011; Potthast, Eiselt, Barrón-Cedeño, Stein, & Rosso, 2011; Potthast, Hagen, Völske, & Stein, 2013; Potthast, Stein, Barrón-Cedeño, & Rosso, 2010; Potthast, Stein, Eiselt, Barrón-Cedeño, & Rosso, 2009) งานกลุ่มนี้มีวัตถุประสงค์เพื่อประเมินและเปรียบเทียบประสิทธิภาพการตรวจหาระหว่างแนวการตรวจหาการลักลอกต่างๆ ที่ถูกพัฒนาขึ้นเพื่อให้บรรลุวัตถุประสงค์ดังกล่าว พอตแทสต์และคณะต้องสร้างคลังข้อมูลที่ใช้ในการประเมินประสิทธิภาพของระบบซึ่งประกอบด้วยข้อความที่มีระดับความคลุมเครือต่างกัน งานกลุ่มนี้ได้แสดงให้เห็นว่าข้อความที่สร้างขึ้นด้วยเครื่องโดยใช้แบบจำลองการค้นคืนซึ่งโดยมากจะมีความละม้ายสูงอย่างการคัดลอกโดยตรงนั้นจะมีค่าความคลุมเครือต่ำกว่าข้อความที่สร้างโดยมนุษย์ที่ได้จากแปลข้อความข้ามภาษา ลักษณะดังกล่าวจึงยืนยันได้ว่าประเภทและรูปแบบทางภาษาของการลักลอกงานวิชาการที่ใช้ความสามารถในการลักลอกสูงก็ย่อมมีระดับความคลุมเครือสูงตามไปด้วย

นอกจากนี้ ในแง่ของความยากของการตรวจหา (difficulty of detection) นั้นยังมีงานอีกชิ้นหนึ่งซึ่งช่วยยืนยันแนวคิดเรื่องความสัมพันธ์ระหว่างประเภทและรูปแบบทางภาษาของการลักลอกงานวิชาการกับความยากของการตรวจหาได้คืองานของกิปป์และคณะ (Gipp, Meuschke, & Beel, 2011) ที่ต้องการจะเปรียบเทียบประสิทธิภาพของการตรวจหาการลักลอกระหว่างแนวทางการตรวจหาแบบอิงข้อความ (text-based plagiarism detection approaches) กับแนวทางการตรวจหาแบบอิงการอ้างอิง (citation-based plagiarism detection approaches) ในงานชิ้นนี้

กิปป์และคณะได้สร้างคลังข้อมูลที่ประกอบด้วยข้อความจากการลักลอก 4 ประเภท ได้แก่ การคัดลอกและวาง (copy & paste) การลักลอกแบบปลอมแปลง (disguised plagiarism) การลักลอกความคิด/โครงสร้าง (idea/structure plagiarism) และการแปลภาษา ทั้งนี้ หากพิจารณาการจัดประเภทการลักลอกงานวิชาการในงานชิ้นนี้แล้วจะเห็นได้ว่ามีลักษณะคล้ายกับการจัดประเภทงานของอัลซะฮ์รานิและคณะ ต่างกันที่งานชิ้นนี้จะแบ่งประเภทย่อยไม่ละเอียดเท่ากับงานดังกล่าว กล่าวคือ ประเภทของการลักลอกระหว่างการคัดลอกและวางและการลักลอกความคิดและโครงสร้างนั้น กิปป์และคณะแทนด้วยการลักลอกแบบปลอมแปลงทั้งหมด ซึ่งในประเภทนี้จะรวมตั้งแต่การลบหรือเพิ่มคำไปจนถึงการถอดความ จนอาจกล่าวได้ว่าการจัดประเภทเช่นนี้อาจหยาบเกินกว่าที่จะให้ผลการวิเคราะห์ที่ถูกต้องและแม่นยำ อย่างไรก็ตาม ผลการวิเคราะห์ที่ออกมาก็ช่วยทำให้เห็นความสัมพันธ์ระหว่างประเภทและรูปแบบทางภาษาของการลักลอกงานวิชาการกับความยากของการตรวจหาได้

Plagiarism type	**Text-based**	**Citation-based**
Copy&paste	~ 70 % Good results even for short fragments	Unsuitable as short fragments cannot be detected
Disguised plagiarism	< 10 %	Depending on the fragments length ~ 30 %
Idea / structure plagiarism	0 %	Some cases could be identified
Translated plagiarism	< 5 %	~ 80 %. 13 out of 16 fragments could be identified.

ภาพที่ 2.3 ผลการเปรียบเทียบประสิทธิภาพของระบบตรวจหาการลักลอกโดยกิปป์และคณะ (Gipp et al., 2011, p. 257)

ภาพที่ 2.3 แสดงผลการเปรียบเทียบประสิทธิภาพของระบบตรวจหาการลักลอกระหว่างแนวทางการตรวจหาแบบอิงข้อความกับแนวทางการตรวจหาแบบอิงการอ้างอิง จะเห็นได้ว่าระบบการตรวจหาการลักลอกแบบอิงข้อความซึ่งเป็นแนวทางที่ใช้กันโดยทั่วไปนั้นสามารถตรวจหาการลักลอกประเภทการคัดลอกและวางได้ดีที่สุด แม้จะมีปัญหาในแนวทางการตรวจหาแบบอิงการอ้างอิงก็ตาม รองลงมาคือการลักลอกแบบปลอมแปลง การลักลอกโดยการแปล และการลักลอกความคิด/โครงสร้าง หากพิจารณาเฉพาะผลจากการลักลอกแบบอิงข้อความแล้วจะเห็นว่าความยากของการตรวจหานั้นมีทิศทางสอดคล้องไปกับประเภทและรูปแบบทางภาษาของการลักลอกงานวิชาการ อีกทั้งยังสอดคล้องกับระดับความคลุมเครือดังได้กล่าวมาแล้วอีกด้วย และลักษณะดังกล่าวยังสอดคล้องกับลักษณะที่เป็นลำดับชั้นของประเภทการลักลอกตามที่อัลซะฮ์รานิและคณะได้เสนอไว้อีกด้วย

หากนำผลการศึกษาข้างต้นของกิปป์และคณะ (Gipp et al., 2011) มาเปรียบเทียบกับการจัดแบ่งประเภทการลักลอกงานวิชาการในงานของอัลซะฮ์รานิและคณะ (Alzahrani et al., 2012) แล้ว จะเห็นได้ว่าการจัดประเภทของอัลซะฮ์รานิและคณะนั้นไม่เพียงแต่จะให้รายละเอียดที่มากพอจะใช้ในการศึกษาวิจัยด้านการลักลอกงานวิชาการต่อไปเท่านั้น แต่ลำดับชั้นในการจัดประเภทดังกล่าวยังสอดคล้องไปกับระดับความคลุมเครือ (degree of obfuscation) และความยากของการตรวจหา (difficulty of detection) จากน้อยไปหามาก ซึ่งลักษณะดังกล่าวนี้ยังตรงกับเกณฑ์การประเมินประสิทธิภาพของระบบตรวจหาการลักลอกที่ยอมกันทั่วไปด้วย ด้วยเหตุดังกล่าวนี้ผู้วิจัยจึงเห็นด้วยกับงานของอัลซะฮ์รานิและคณะเป็นพิเศษ เนื่องจากสามารถนำไปปรับประยุกต์ใช้ในการศึกษาวิจัยด้านการลักลอกงานวิชาการได้เป็นอย่างดี โดยเฉพาะอย่างยิ่งในแขนงงานตรวจหาการลักลอกทางวิชาการ

จากเหตุผลที่ได้กล่าวข้างต้น ในงานวิจัยชิ้นนี้ ผู้วิจัยจึงออกแบบข้อมูลการลักลอกเพื่อใช้ทดสอบระบบโดยยึดแนวคิดของอัลซะฮ์รานิและคณะเป็นหลัก โดยจะลดทอนรายละเอียดของประเภทการลักลอกในงานชิ้นดังกล่าวลงเพื่อให้เหมาะกับแนวทางการตรวจหาการลักลอก ทั้งนี้ประเภทของการลักลอกงานวิชาการที่สรุปได้ใหม่นั้นสามารถแบ่งได้เป็น 4 ประเภท ดังนี้

1) การคัดลอกโดยตรง (Exact copy)

การคัดลอกโดยตรงถือได้ว่าเป็นการลักลอกที่ทำได้ง่ายและใช้เวลาน้อยที่สุด ทั้งนี้ ผู้ลักลอกอาจทำได้โดยคัดลอกข้อความจากอินเทอร์เน็ตและวางในเอกสาร (copy & paste: C&P) ลักษณะของข้อความที่ลักลอกจะมีลักษณะเหมือนกับต้นฉบับแบบคำต่อคำโดยไม่มีการอ้างอิงแหล่งที่มาของข้อมูล ทั้งนี้การคัดลอกโดยตรงอาจเกิดขึ้นทั้งหมดของเอกสารหรือปรากฏเฉพาะบางส่วนในเอกสารก็ได้ ในแง่การตรวจหาการลักลอกนั้น การคัดลอกโดยตรงจะสามารถตรวจหาได้ง่ายที่สุดโดยใช้วิธีพื้นฐานอย่างเช่นการจับคู่สายอักขระพื้นฐาน

2) การคัดลอกโดยใกล้เคียง (Near copy)

การคัดลอกโดยใกล้เคียงนั้นแตกต่างจากการคัดลอกโดยตรงในแง่ของการแก้ไข กล่าวคือแทนที่จะคัดลอกและวางเพียงขั้นตอนเดียว การคัดลอกโดยใกล้เคียงจะใช้วิธีแทรกคำ (insertion) ลบคำ (deletion) ในแง่การตรวจหาการลักลอก การลักลอกโดยใกล้เคียงยังสามารถถูกตรวจหาได้จากการจับคู่สายอักขระพื้นฐานหากประยุกต์ใช้ค่าระยะการแก้ไข (edit distance) ของชุดอักขระทั้งสองชุด

3) การคัดลอกโดยดัดแปลง (Modified copy)

การคัดลอกโดยดัดแปลงนั้นจะมุ่งแก้ไขข้อความต้นฉบับในระดับวลีและอนุพากย์เป็นหลัก กล่าวคือ ผู้ลักลอกจะแทรก ลบ และเรียงลำดับวลีและวลีในตัวฉบับใหม่ให้แตกต่างจากข้อความต้นฉบับ ทั้งนี้ในบางกรณีอาจรวมเอาการแทรกคำ ลบคำ หรือสลับตำแหน่งของคำเอาไว้ด้วย ทำให้ในการตรวจหาการลักลอกจำเป็นต้องใช้วิธีที่ซับซ้อนมากยิ่งขึ้น ซึ่งผู้วิจัยจะกล่าวถึงในหัวข้อวิธีตรวจเทียบภายนอกหาการลักลอก

4) การถอดความ (Paraphrase)

การถอดความถือได้ว่าเป็นการลักลอกที่ใช้ทักษะและความสามารถในการเขียนสูงกว่าการลักลอกทั้ง 3 ประเภทที่ได้กล่าวมาข้างต้น กล่าวคือ ผู้ลักลอกจะต้องเปลี่ยนแปลงคำศัพท์และลักษณะทางวากยสัมพันธ์ของข้อความต้นฉบับไป ในขณะเดียวกันก็ต้องคงไว้ซึ่งลำดับคำและลำดับโครงสร้างด้วย โดยการเปลี่ยนแปลงที่เกิดขึ้นดังกล่าวจะเป็นเพียงการแทนที่คำเดิมด้วยคำไวพจน์และเพิ่มเติมคำเชื่อมในบางจุดเท่านั้น จึงกล่าวได้ว่า การถอดความมีความเกี่ยวข้องสัมพันธ์กับระดับต่างๆ ทางภาษาหลายระดับ ทั้งระดับคำศัพท์ ระดับวากยสัมพันธ์ และระดับความหมาย ซึ่งเป็นเหตุให้ไม่สามารถตรวจหาการถอดความในฐานะการลักลอกงานวิชาการได้โดยง่าย ด้วยเหตุนี้ ในหัวข้อถัดไปผู้วิจัยจึงได้รวบรวมแง่มุมต่างๆ ที่เกี่ยวข้องกับการถอดความไว้โดยละเอียดเพื่อเป็นประโยชน์ในการพัฒนาระบบตรวจหาการลักลอกต่อไป

2.3 การถอดความ

การถอดความ (paraphrase) ถือได้ว่าเป็นปรากฏการณ์ทางภาษาที่น่าสนใจทั้งในเชิงภาษาศาสตร์และในการประมวลภาษาธรรมชาติ ในด้านภาษาศาสตร์นั้นสามารถพิจารณาการถอดความเป็นการปริวรรต (transform) ข้อความต้นฉบับไปสู่รูปอื่นๆ การปริวรรตนี้อาจเกิดได้ทั้งในระดับคำศัพท์และระดับประโยค อย่างไรก็ตาม กระบวนการถอดความที่เกิดขึ้นนี้ต้องพยายามรักษาข้อมูลในระดับความหมายเอาไว้ให้คงที่หรือใกล้เคียงกับข้อความต้นฉบับมากที่สุด ส่วนในด้านการประมวลผลภาษาธรรมชาตินั้นได้เป็นที่ยอมรับกันโดยทั่วไปว่า การถอดความคือทางเลือกที่รูปภาษาใดๆ ไม่ว่าจะเป็นวลี ประโยค หรือข้อความขนาดยาว ใช้ถ่ายทอดข้อมูล (information) ที่เหมือนกัน (Androutsopoulos & Malakasiotis, 2010, p. 135) ด้วยลักษณะของการถอดความที่เกี่ยวกับโครงสร้างและความหมายในรูปภาษาดังกล่าวนี้ จึงทำให้เป็นการยากที่จะตรวจหาการลักลอกในข้อความที่ผ่านการถอดความมา ดังนั้น เพื่อเป็นการประมวลความรู้มาประยุกต์ใช้ในการสร้างชุดข้อมูลฝึกฝนและออกแบบระบบตรวจหาการลักลอกต่อไป ในหัวข้อนี้ ผู้วิจัยจึงจะกล่าวถึงข้อแตกต่าง

ระหว่างการลักลอกงานวิชาการกับการถอดความ หลักทั่วไปที่ใช้ในการถอดความ รวมถึงงานที่ศึกษาเกี่ยวกับหมวดหมู่และรูปแบบของการถอดความในภาษาต่างๆ รายละเอียดมีดังนี้

2.3.1 การถอดความกับการลักลอกงานวิชาการ

หากกล่าวถึงการเรียบเรียงและนำเสนอความคิดในงานวิชาการแล้ว การถอดความถือเป็นเครื่องมือที่มีบทบาทสำคัญอย่างหนึ่งในกระบวนการดังกล่าว กล่าวคือ การถอดความจะช่วยให้ผู้ผลิตงานวิชาการเข้าใจเอกสารอันเป็นแหล่งที่มาได้อย่างแจ่มแจ้ง และยังช่วยให้ผู้ผลิตงานวิชาการนำเอาสาระสำคัญที่ปรากฏในเอกสารแหล่งที่มานั้นๆ ไปสู่ผู้อ่านได้ (Behrens & Rosen, 2008, p. 36) จึงอาจกล่าวได้ว่าการถอดความเป็นทักษะสำคัญประการหนึ่งที่ผู้ผลิตงานวิชาการพึงมี อย่างไรก็ตาม การถอดความและการลักลอกงานวิชาการนั้นก็มีความคาบเกี่ยวกันอยู่ไม่น้อย ในที่นี้ผู้วิจัยจึงจะชี้ให้เห็นถึงข้อแตกต่างระหว่างการลักลอกงานวิชาการและการถอดความที่ถูกต้องตามหลักการ ซึ่งแบ่งได้เป็น 3 ประการ ดังนี้

ข้อแตกต่างระหว่างการลักลอกงานวิชาการและการถอดความประการแรกนั้นสามารถพิจารณาได้จาก การอ้างอิงแหล่งที่มาของเอกสารต้นฉบับ ทั้งนี้ ผู้เชี่ยวชาญด้านการเขียนเชิงวิชาการหลายคน ไม่ว่าจะเป็นเฮฟเฟอร์นันและลินคอล์น (Heffernan & Lincoln, 1982, p. 457), สแปตต์ (Spatt, 1987, p. 92), เกลาและจาคอบเซน (Glau & Jacobsen, 2001, p. 57), มัลเวนีย์และโจล์ลิฟฟ์ (Mulvaney & Jolliffe, 2005, p. 335) ต่างก็กล่าวตรงกันว่าในการถอดความจำเป็นต้องมีการอ้างอิงให้ถูกต้อง เพราะแม้จะมีหรือไม่มีเจตนาในการลักลอกก็ตาม หากข้อเขียนที่ถอดความมาจากผู้อื่นไม่มีการอ้างอิงที่ถูกต้องแล้ว ข้อเขียนดังกล่าวก็จะเข้าข่ายการลักลอกทันที

ข้อแตกต่างประการต่อมาที่สังเกตได้จากการลักลอกและการถอดความคือ พฤติกรรมการเขียนของผู้ผลิตงานวิชาการซึ่งจะแสดงผ่านลักษณะของข้อความ กล่าวคือ ในการเขียนเพื่อลักลอกนั้น ผู้ลักลอกสามารถเขียนได้หลายวิธี ตั้งแต่คัดลอกข้อความจากเอกสารต้นฉบับ การเลือกนำวลีหรือข้อความจากเอกสารต้นฉบับมาเขียนปะติดปะต่อ (patch writing) ในงานของตน การถอดความแบบนำวลีหลายๆ วลีมาปะติดปะต่อกัน (patchwork paraphrasing) (กัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์, 2554, น. 13) จนกระทั่งถึงสรุปความจากเอกสารต้นฉบับ ดังนั้นข้อความที่ปรากฏในเอกสารที่มาจากการลักลอกก็อาจมีลักษณะการใช้ภาษาที่ไม่สม่ำเสมอ เช่น มีการสะกดคำคำเดียวกันผิดหรือมีการใช้คำเชื่อมบางคำถี่สูงเฉพาะในบางตำแหน่งของเอกสาร เป็นต้น ในขณะที่การเขียนเพื่อถอดความนั้น ผู้ถอดความจะต้องพิจารณาหาประเด็นความคิดหลักในระดับประโยคหรือย่อหน้าเท่านั้นเสียก่อน (Behrens & Rosen, 2008, p. 37) จากนั้นจึงสร้างรูปภาษาที่เป็นการถอดความจากประเด็นความคิดในประโยคหรือย่อหน้านั้น ด้วยวิธีดังกล่าวนี้จะทำให้การถอดความที่ได้จึงมีลักษณะ

การใช้ภาษาของตัวผู้ถอดความเอง ลักษณะการใช้ภาษาดังกล่าวจะสอดคล้องสม่ำเสมอกันทั้งเอกสาร อย่างไรก็ดี ในประเด็นนี้อาจเกิดคำถามเกิดขึ้นได้ว่า หากผู้ลักลอกใช้การถอดความเพื่อลักลอกทั้งหมดแล้วข้อแตกต่างระหว่างการลักลอกและการถอดความจะพิจารณาได้จากลักษณะใด ในลักษณะเช่นนี้ยังสามารถพิจารณาความแตกต่างจากลักษณะภาษาของรูปถอดความได้

ลักษณะภาษาของรูปถอดความเป็นลักษณะประการสุดท้ายที่สามารถบ่งชี้ให้เห็นข้อแตกต่างระหว่างการลักลอกและการถอดความได้ โดยเฉพาะในกรณีที่การลักลอกนั้นใช้วิธีการถอดความเป็นเครื่องมือในการลักลอก ทั้งนี้ ฟริกก์ (Frick, 2005) และกัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์ (2554, น. 24-25) ได้ยกกรณีตัวอย่างที่แสดงให้เห็นว่า แม้ข้อเขียนจะเป็นการถอดความจากต้นฉบับแต่หากไม่เรียบเรียงด้วยลักษณะการใช้ภาษาของตนเองก็ถือว่าข้อเขียนนั้นเป็นการลักลอกเช่นเดียวกัน ดังนี้

> ข้อความต้นฉบับ : The rise of industry, the growth of cities, and the expansion of the population were the three great developments of late nineteenth century American history. As new, larger, steam-powered factories became a feature of the American landscape in the East, they transformed farm hands into industrial laborers, and provided jobs for a rising tide of immigrants. With industry came urbanization the growth of large cities (like Fall River, Massachusetts, where the Bordens lived) which became the centers of production as well as of commerce and trade.
>
> ข้อความ (1) : The increase of industry, the growth of cities, and the explosion of the population were three large factors of nineteenth century America. As steam- driven companies became more visible in the eastern part of the country, they changed farm hands into factory workers and provided jobs for the large wave of immigrants. With industry came the growth of large cities like Fall River where the Bordens lived which turned into centers of commerce and trade as well as production.

> ข้อความ (2) : Fall River, where the Borden family lived, was typical of northeastern industrial cities of the nineteenth century. Steam-powered production had shifted labor from agriculture to manufacturing, and as immigrants arrived in the US, they found work in these new factories. As a result, populations grew, and large urban areas arose. Fall River was one of these manufacturing and commercial centers (Joyce Williams et al., 1981, p.1).

จากข้อความ (1) เป็นข้อความที่ถือเป็นการลักลอกแม้ว่าจะมีการถอดความ ทั้งนี้เนื่องมาจากเหตุผล 2 ประการ ได้แก่ ประการแรก ผู้เขียนถอดความโดยไม่ได้ใช้ลักษณะภาษาของตัวเอง เป็นเพียงเปลี่ยนถ้อยคำและวลีบางวลีหรือเปลี่ยนลำดับประโยคเท่านั้น นอกจากนี้ การใช้คำว่า *companies* แทนคำว่า *factories* ก็ยังสื่อถึงความหมายที่แตกต่างกัน เมื่อพิจารณาโครงสร้างโดยรวมจะเห็นว่ามีความละม้ายกับข้อความต้นฉบับมาก และที่สำคัญคือข้อความนี้ไม่มีการอ้างอิงแหล่งที่มา แตกต่างจากข้อความ (2) ซึ่งนอกจากจะมีการอ้างอิงถูกต้องแล้ว ในส่วนของการถอดความยังเรียบเรียงด้วยภาษาของตนเอง ซึ่งจะทำให้ข้อเขียนที่ได้มีลักษณะภาษาที่สม่ำเสมอดังที่ผู้เขียนได้กล่าวไปแล้ว

อย่างไรก็ดี ข้อแตกต่างระหว่างการลักลอกงานวิชาการกับการถอดความนั้นอาจจะไม่สามารถพิจารณาได้จากลักษณะใดเพียงลักษณะเดียวจากที่กล่าวมาทั้ง 3 ประการ ทั้งนี้ ในกรณีของงานวิชาการนั้นจำเป็นต้องพิจารณาลักษณะทั้ง 3 ประการร่วมกันทุกครั้งจึงจะสามารถทำให้เห็นข้อแตกต่างระหว่างการลักลอกงานวิชาการกับการถอดความจากต้นฉบับที่ยอมรับในงานวิชาการได้อย่างชัดเจน

จากข้อแตกต่างระหว่างการลักลอกงานวิชาการกับการถอดความที่กล่าวมาสามารถทำให้เห็นได้ว่าการถอดความที่ไม่ถูกต้องตามหลักการเขียนนั้นอาจส่งผลให้ข้อเขียนเข้าข่ายการลักลอกงานวิชาการได้ นอกจากนี้ การถอดความยังถือเป็นวิธีอ้างอิงที่ไม่อาจจะหลีกเลี่ยงได้ในการเรียบเรียงงานวิชาการ เพราะแม้ว่าเราจะสามารถยกเอาคำพูด (quote) ของต้นฉบับมาไว้ในงานได้ แต่หากข้อเขียนต้นฉบับนั้นมีการใช้ภาษาที่ไม่ชัดเจน เป็นนามธรรม หรือเป็นภาษาที่ล้าสมัย ก็อาจก่อให้เกิดความสับสนแก่ผู้อ่านได้ การถอดความด้วยภาษาที่เรียบเรียงใหม่จึงแก้ปัญหาดังกล่าวได้ดีกว่า (Behrens & Rosen, 2008, p. 34) ด้วยเหตุนี้ การเรียนรู้ถึงหลักการถอดความที่ถูกต้องจึงถือเป็นสิ่งที่ผู้ผลิตงานวิชาการพึงปฏิบัติ

2.3.2 หลักปฏิบัติในการถอดความ

หลักปฏิบัติที่ใช้กันในการถอดความที่ปรากฏอยู่ในตำราการเขียนเชิงวิชาการไม่ว่าจะเป็นงานของสแปตต์ (Spatt, 1987), โกลด์และคณะ (Gould, DiYanni, & Smith, 1989), แมเรียสและวีเนอร์ (Marius & Wiener, 1994), เกลาและจาคอบเซน (Glau & Jacobsen, 2001), ลันส์ฟอร์ดและบริดจ์ส (Lunsford & Bridges, 2005), ครูเซียสและชาเนลล์ (Crusius & Channell, 2006), เบห์เรนส์และโรเซน (Behrens & Rosen, 2008), และเคลาส์ (Clouse, 2008) นั้นต่างก็มีเนื้อหาสอดคล้องไปในทางเดียวกัน กล่าวคือ ขั้นแรก ผู้ถอดความจำเป็นต้องทำความเข้าใจข้อเขียนต้นฉบับให้แจ่มแจ้งเสียก่อน ทั้งนี้เพราะการถอดความนั้นจำเป็นต้องถ่ายทอดความคิดหรือประเด็นสำคัญที่ปรากฏในข้อเขียนต้นฉบับแบบประเด็นต่อประเด็น ทั้งนี้ เบห์เรนส์และโรเซน (Behrens & Rosen, 2008, p. 37) ได้กล่าวว่า โดยทั่วไปแล้วการถอดความจะอิงตามย่อหน้าหรือประโยคที่เข้าใจยากหรือมีความสำคัญในข้อเขียนต้นฉบับ หลักปฏิบัติในขั้นต่อมาจึงเป็นการเรียบเรียงย่อหน้าหรือประโยคในข้อเขียนต้นฉบับออกมาด้วยลักษณะการใช้ภาษาของผู้ถอดความ ในขั้นตอนนี้ ครูเซียสและชาเนลล์ (Crusius & Channell, 2006, p. 112) ได้ให้ข้อควรคำนึงไว้ว่า การถอดความไม่ใช่การรักษาลำดับคำในต้นฉบับไว้แล้วแทนที่ด้วยคำไวพจน์ ผู้ถอดความอาจต้องแยกประโยคที่ซับซ้อนออกมาและเรียบเรียงใหม่ให้เป็นประโยคพื้นฐาน 2-3 ประโยคโดยยังคงความหมายของประเด็นความคิดสำคัญไว้ นอกจากนี้ยังไม่ควรตกเป็นทาสของอรรถาภิธาน (thesaurus) หากคำที่ปรากฏในข้อเขียนเป็นคำที่มีความหมายพื้นฐานที่เข้าใจง่ายแล้วก็ไม่จำเป็นต้องหาคำไวพจน์อื่นมาแทนที่ ด้วยวิธีดังกล่าวนี้นำไปสู่หลักปฏิบัติประการต่อมาคือ ให้นำเสนอประเด็นความคิดตามที่ปรากฏในข้อเขียนต้นฉบับโดยไม่มีการเลือก ย่อ หรือเพิ่มเติมความคิดลงไป ด้วยวิธีนี้การถอดความที่ได้อาจมีความยาวเท่ากับข้อเขียนต้นฉบับหรืออาจยาวกว่าต้นฉบับได้ในกรณีที่ต้องอธิบายความคิดที่ซับซ้อน และในกรณีที่ถอดความจากภาษาต่างประเทศ ผู้ถอดความต้องพยายามเลือกใช้ประโยคและคำศัพท์ที่สื่อความหมายได้ตรงกับข้อเขียนต้นฉบับ

ส่วนหลักการเรียบเรียงภาษาในการถอดความนั้น สถาบันภาษา จุฬาลงกรณ์มหาวิทยาลัย (Chulalongkorn University Language Institute, n.d.) ได้เสนอขั้นตอนการปฏิบัติไว้ 3 ขั้นตอนดังนี้ ขั้นแรกให้ใช้คำไวพจน์หรือคำที่มีความหมายใกล้เคียง ขั้นที่สองให้เปลี่ยนโครงสร้างของประโยคด้วยวิธีต่างๆ เช่น การใช้โครงสร้าง There + V_{be} + NP การลดฐานะของอนุพากย์ให้เป็นวลี การย้ายที่คุณานุประโยค หรือการใช้รูปกรรตุหรือกรรมวาจกของกริยา เป็นต้น ส่วนในขั้นตอนสุดท้ายนั้นคือทำการเปลี่ยนแปลงหน่วยอื่นๆ ให้อยู่ในตำแหน่งที่เหมาะสม เช่น การย้ายส่วนขยายในประโยค การสับตำแหน่งของอนุพากย์ในประโยค เป็นต้น

จากที่กล่าวมาสามารถสรุปข้อควรคำนึงในการถอดความเพิ่มเติมได้อีก 3 ประเด็น ประเด็นแรกคือ ขนาดของข้อความที่ได้จากถอดความต้องมีความยาวเท่ากับหรือใกล้เคียงกับต้นฉบับเพื่อจะเก็บรายละเอียดของต้นฉบับให้ครบทุกประเด็นความคิด ประเด็นที่สองคือ การถอดความควรเป็นไปตามโครงสร้างและลำดับของประเด็นความคิดในข้อเขียนต้นฉบับ ซึ่งจะเป็นตัวกำหนดลักษณะโครงสร้างของการถอดความ ดังจะเห็นได้จากหลักปฏิบัติในการถอดความที่ได้กล่าวไปข้างต้นที่กล่าวถึงการเปลี่ยนแปลงโครงสร้างในย่อหน้าของประโยคเป็นสำคัญ และประเด็นสุดท้ายคือ การใช้ภาษาในการถอดความที่ผู้ถอดความต้องใช้สำนวนภาษาของตนเองไม่ว่าจะเป็นระดับคำหรือประโยค และภาษาที่เลือกใช้นั้นต้องสื่อความหมายได้ใกล้เคียงกับต้นฉบับ

2.3.3 การจัดประเภทการถอดความ

การถอดความถือเป็นปรากฏการณ์ที่ได้รับความสนใจจากนักภาษาศาสตร์มานานแล้ว โดยพยายามอธิบายธรรมชาติของกระบวนการถอดความดังที่ปรากฏในงานของโนแลน (Nolan, 1970) เป็นต้น กระทั่งถึงปัจจุบัน ประเด็นการวิจัยที่เกี่ยวเนื่องกับการถอดความไม่ได้จำกัดอยู่ในแวดวงภาษาศาสตร์เท่านั้น แต่ยังขยายไปถึงสาขาวิทยาศาสตร์คอมพิวเตอร์และการประมวลผลภาษาธรรมชาติด้วย โดยปัจจุบัน ความสนใจเกี่ยวกับการถอดความจะมุ่งไปที่การสรุปลักษณะร่วมและการจัดประเภทรูปแบบของการถอดความเพื่อประยุกต์ใช้ในงานคอมพิวเตอร์เป็นหลัก

วิทยานิพนธ์ปริญญาดุษฎีบัณฑิตของฟุจิตะ (Fujita, 2005) เรื่อง *Automatic generation of syntactically well-formed and semantically appropriate paraphrases* เป็นตัวอย่างหนึ่งของความพยายามจัดแบ่งหมวดหมู่ของการถอดความทางคำศัพท์และโครงสร้างในภาษาญี่ปุ่นเพื่อนำไปใช้ในการสร้างการถอดความอัตโนมัติ

ในการสร้างการถอดความนั้น ขั้นตอนที่สำคัญขั้นตอนหนึ่งคือการจัดแบ่งหมวดหมู่ของการถอดความทางคำศัพท์และโครงสร้างในภาษาญี่ปุ่น ฟุจิตะได้รวบรวมปรากฏการณ์ทางภาษาญี่ปุ่นไว้อย่างหลากหลายและบางส่วนก็ได้รวบรวมจากผู้ที่เคยศึกษาเรื่องการถอดความไว้แล้ว จากนั้นจึงวิเคราะห์การถอดความในแง่ของคำศัพท์และโครงสร้างเท่านั้น มิได้คำนึงถึงสถานการณ์ในการสื่อสารจริง ด้วยการวิเคราะห์โดยใช้คำศัพท์เป็นเกณฑ์ทำให้พบว่า ในภาษาญี่ปุ่นนั้นมีการรักษาความหมายในกระบวนการปริวรรตเป็นรูปถอดความไว้ด้วยการใช้รูปไวพจน์ ส่วนการวิเคราะห์โดยอาศัยโครงสร้างเป็นเกณฑ์นั้นก็ทำให้ฟุจิตะพิจารณาโครงสร้างในแง่วากยสัมพันธ์ได้ชัดเจนขึ้น เนื่องจากภาษาญี่ปุ่นเป็นภาษาที่มีระบบการก ดังจะเห็นได้จากตัวอย่างคู่ถอดความต่อไปนี้

Src.	*kare-wa*	*kikai-**sousa-ga***		***jouzu-da***.
	he-TOP	machine-operation-NOM		be good-COP
	He is good at operating the machine.			
Pa.	*kare-wa*	*kikai-**o***	***jouzu-ni***	***sousa-suru***.
	he-TOP	machine-DAT	well-ADV	to operate-PRES
	He operates the machine well.			

จากตัวอย่างคู่ถอดความ คำว่า *kikai-sousa* ‘machine-operation’ นั้นเป็นคำนามที่เป็นกรรมของประโยคสังเกตได้จะตัวบ่งการก *–ga* แต่เมื่อปริวรรตสู่รูปถอดความแล้วคำว่า *kikai* ‘machine’ นั้นกลายเป็นคำที่ถูกบ่งการกกรรมรอง *-o* ส่วนคำ sousa ‘to operate’ นั้นกลับกลายเป็นกริยาของประโยคที่ถูกระบุการกปัจจุบันกาล *–suru* ตัวอย่างดังกล่าวนี้แสดงเห็นว่าการวิเคราะห์จัดประเภทการถอดความในภาษาญี่ปุ่นโดยอิงโครงสร้างในกรอบของระบบการกนั้นทำให้การถอดความของรูปประสมได้อย่างชัดเจน

ด้วยการวิเคราะห์โดยอาศัยเกณฑ์ดังกล่าวข้างต้นทำให้ฟุจิตะสามารถจัดประเภทการถอดความในภาษาญี่ปุ่นได้เป็น 6 ประเภทใหญ่ ได้แก่ การถอดความโดยใช้คำไวพจน์ของคำนาม การถอดความโดยใช้รูปเชิงหน้าที่ เช่น การถอดความคำหน้าที่หรือรูปทัศนภาวะ (modality expression) การถอดความรูปประสมของนาม กริยา หรือตัวบ่งชี้การก การถอดความในโครงสร้างของอนุพากย์ เช่น การเปลี่ยนรูปแสดงการเปรียบเทียบหรือรูปแสดงการก การถอดความระดับพหุอนุพากย์ (muti-clausal paraphrase) และการถอดความรูปที่มีลักษณะเฉพาะ (idiosyncratic expression) เช่น สำนวนเฉพาะ

โดยส่วนตัวแล้ว ผู้วิจัยเห็นว่าการจัดประเภทการถอดความจากต้นฉบับที่ปรากฏในงานของฟุจิตะนั้นมีข้อได้เปรียบกว่าการจัดประเภทการถอดความในงานชิ้นอื่นๆ ในด้านเลือกวิเคราะห์และจัดประเภทภาษาญี่ปุ่นซึ่งเป็นภาษาที่ระบบการกชัดเจน ในการวิเคราะห์เพื่อจัดประเภทจึงไม่จำเป็นต้องพึ่งพาทฤษฎีทางภาษาศาสตร์ในเชิงลึกมากเท่าใดนัก ข้อน่าสนใจอีกประการในงานชิ้นนี้คือการศึกษาการถอดความในระดับพหุอนุพากย์ คือการยุบรวมหรือแตกอนุพากย์หรือประโยค 2 ประโยคในการถอดความ ซึ่งเป็นลักษณะที่เกิดขึ้นจริงในการปรากฏการณ์ทางภาษาแต่มักถูกละเลยในการศึกษาการถอดความ อย่างไรก็ตาม หากจะนำกรอบการวิเคราะห์ดังกล่าวมาประยุกต์ใช้กับภาษาไทยก็คงเป็นไปโดยไม่สะดวกนัก เนื่องด้วยภาษาไทยไม่มีระบบบ่งชี้การกดังเช่นที่ภาษาญี่ปุ่นมี

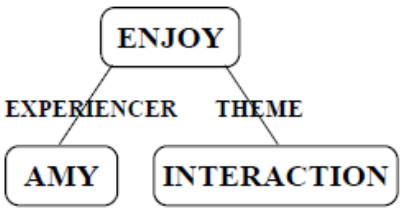
งานที่จัดประเภทการถอดความชิ้นต่อไปเป็นงานของคอสโลว์สกีและคณะ (Kozlowski, McCoy, & Vijay-Shanker, 2003) ที่ต้องการสร้างการถอดความเช่นเดียวกับงานของฟุจิตะ (Fujita,

2005) ในงานชิ้นนี้คอสโลว์สกีและคณะได้จัดประเภทการถอดความโดยอาศัยการพิจารณาจากลำดับชั้นในโครงสร้างแบบเพรดิเคต/อาร์กิวเมนต์เป็นหลัก กล่าวอีกนัยหนึ่งคือพิจารณาการปรากฏของเพรดิเคตและอาร์กิวเมนต์ระหว่างคู่ของประโยคโดยอิงเกณฑ์บทบาททางความหมายของหน่วยต่างๆ ที่เกิดร่วมกับคำกริยาในประโยค หากพบว่าคู่ของประโยคมีเพรดิเคตและอาร์กิวเมนต์ปรากฏในบทบาทตรงกันก็จะถือว่าเป็นการถอดความ ยกตัวอย่างเช่นคู่ถอดความ

Src. Amy enjoyed the interaction.

Pa. The interaction pleased Amy.

ตัวอย่างคู่ถอดความนี้ คอสโลว์สกีและคณะ (Kozlowski et al., 2003, p. 2) ได้จัดให้เป็นการถอดความประเภทหนึ่งเรียกว่าตำแหน่งที่ต่างกันของการปรากฏของอาร์กิวเมนต์ (different placement of argument realizations) จะสังเกตได้ว่าทั้ง 2 ประโยคมีอาร์กิวเมนต์ที่เป็นคำนามคำเดียวกันแต่มีตำแหน่งที่ต่างกัน ทั้งนี้หากพิจารณาในโครงสร้างแบบพึ่งพาจะเห็นได้ว่ากริยา *enjoy* ในประโยคต้นฉบับเป็นส่วนหลัก และคำนาม *Amy* และ *the interaction* เป็นส่วนพึ่งพา ตำแหน่งของอาร์กิวเมนต์ในประโยคดังกล่าวเป็นผลมาจากเงื่อนไขว่าส่วนพึ่งพา (อาร์กิวเมนต์) ต้องมีบทบาททางความหมายสอดคล้องกับส่วนหลักดังได้แสดงในภาพที่ 2.4 ด้านล่างที่จะเห็นได้ว่า *Amy* มีบทบาททางความหมายเป็นผู้มีประสบการณ์ และ *the interaction* ก็มีบทบาทเป็นธีม ในทางกลับกันเมื่อเปลี่ยนเป็นรูปถอดความแล้ว ผู้มีประสบการณ์ *Amy* ก็สามารถถูกวางในตำแหน่งกรรมของประโยคได้ และ *the interaction* ประธานในประโยคได้ในกรณีที่มีบทบาทสอดคล้องกับกริยา *please* ซึ่งเป็นส่วนหลัก (เพรดิเคต)



ภาพที่ 2.4 บทบาททางความหมายที่แสดงในโครงสร้างแบบพึ่งพา (Kozlowski et al., 2003, p. 2)

ด้วยเกณฑ์การจัดประเภทดังได้กล่าวมาทำให้คอสโลว์สกีและคณะสามารถจัดประเภทการถอดความในภาษาอังกฤษได้ 9 ประเภทด้วยกัน ซึ่งแบบได้เป็น 2 กลุ่ม คือ กลุ่มที่พิจารณาความหมายในระดับคำ เช่น การใช้คำไวพจน์ การใช้คำที่มีความหมายทับซ้อนกัน ส่วนอีกกลุ่มหนึ่งจะเน้นพิจารณาลักษณะทางวากยสัมพันธ์ เช่น การถอดความที่ปรากฏผ่านประเภททางวากยสัมพันธ์ (syntactic categories) ที่แตกต่างกัน หรือการสลับตำแหน่งของหน่วยในประโยค เป็นต้น ทั้งนี้ลักษณะที่น่าสนใจในการจัดประเภทดังกล่าวมี 2 ประเด็นคือ ประเด็นแรก คอสโลว์สกีและคณะได้

แสดงให้เห็นว่าการถอดความอาจไม่จำเป็นต้องอยู่ในของรูปภาษาต่อรูปภาษาเป็นคู่เสมอไป การถอดความอาจทำให้เกิดรูปภาษามากกว่า 1 รูปได้ เช่น การถอดความประเภทที่มีคำที่มีความหมายทับซ้อนอย่างประโยค 3 ประโยคต่อไปนี้

(a) Charles flew across the ocean.

(b) Charles crossed the ocean by plane.

(c) Charles went across the ocean by plane.

จากตัวอย่างประโยค (a)-(c) จะเห็นได้ว่ากริยา *fly* นั้นไม่ได้มีความหมายว่า 'ไป' โดยตรง แต่มีความว่า 'ไปโดยเครื่องบิน' อย่างที่ประโยค (c) ได้แสดงให้เห็นรูปถอดความ เช่นเดียวกันกับกริยา *cross* ก็เป็นหมายถึงการไปและแสดงความหมายส่วนหนึ่งของคำว่า *across* เช่นกัน ลักษณะของประโยคที่ประกอบด้วยคำที่มีความหมายทับซ้อนกันเช่นนี้จึงแสดงให้เห็นว่าการถอดความ 1 รูปไม่จำเป็นต้องสร้างรูปถอดความได้เพียงรูปเดียวเสมอไป

นอกจากนี้ยังมีงานวิจัยอื่น ๆ อีกที่มุ่งจัดประเภทการถอดความอีก ได้แก่ งานของวิลาและคณะ (Vila, Martí, & Rodríguez, 2011) ที่เสนอแบบลักษณ์การถอดความ (paraphrase typology) โดยใช้การวิเคราะห์ทางภาษาร่วมกับแนวทางคอมพิวเตอร์เพื่อให้สามารถประยุกต์ใช้ในงานคอมพิวเตอร์ได้อย่างมีประสิทธิภาพ ในงานชิ้นนี้ วิลาและคณะได้เสนอว่า ในการถอดความนั้นมีความสัมพันธ์ที่เกี่ยวเนื่องกันระหว่างเนื้อหาประพจน์ (propositional content) กับการเลือกใช้ถ้อยคำ โดยอาศัยแนวคิดดังกล่าวนี้ วิลาและคณะได้จัดแบบลักษณ์ของการถอดความออกเป็น 5 ประเภทตามลำดับชั้นของหน่วยทางภาษา โดยแต่ละประเภทก็มีกลไกการถอดความที่แตกต่างกันไป

นอกจากงานของวิลาและคณะแล้ว ยังมีงานอีกชิ้นที่นำเสนอการจัดประเภทการถอดความได้แก่งานของภาคัตและโฮวี (Bhagat & Hovy, 2013) จุดเด่นของงานชิ้นนี้คือ นอกจากคณะผู้วิจัยจากเสนอประเภทของการถอดความ 25 ประเภทแล้ว ยังตรวจสอบขอบเขตและความถูกต้องของประเภทการถอดความที่เสนอมาโดยการวิเคราะห์คลังข้อมูลการถอดความด้วย

เมื่อพิจารณางานของวิลาและคณะ (Vila et al., 2011) และภาคัตและโฮวี (Bhagat & Hovy, 2013) เปรียบเทียบกันแล้ว ผู้วิจัยเห็นว่าประเภทย่อยของการถอดความที่งานทั้งสองชิ้นนำเสนอนั้นคล้ายคลึงกัน เพียงแต่งานของวิลาและคณะมีการจัดกลุ่มประเภทของการถอดความโดยอิงกับระดับของหน่วยทางภาษา ทำให้เอื้อต่อการทำความเข้าใจมากกว่า อย่างไรก็ตาม งานทั้ง 2 ชิ้นนี้ก็ไม่ได้ระบุกรอบทฤษฎีไวยากรณ์ที่ใช้ในการวิเคราะห์อย่างชัดเจน อีกทั้งยังมุ่งวิเคราะห์ภาษาอังกฤษเพียงภาษาเดียว จึงเป็นข้อน่าสังเกตว่าหากนำกรอบการวิเคราะห์นี้มาประยุกต์ใช้กับภาษาไทยอาจจะพบข้อจำกัดบางประการอันเนื่องมาจากลักษณะภาษาที่แตกต่างกัน

อย่างไรก็ดี ยังมีงานที่มุ่งวิเคราะห์ประเภทของการถอดความในภาษาไทยปรากฏอยู่ชิ้นหนึ่ง คืองานของกฤตาพร พัชระสุภา และพรฤดี เนติโสภากุล (Phucharasupa & Netisopakul, 2012) ในงานชิ้นนี้ คณะผู้วิจัยได้ยึดบทบาททางความหมายและชนิดของคำกริยาเป็นแนวทางในการวิเคราะห์ กล่าวคือคณะผู้วิจัยจะพิจารณาบทบาททางความหมายของอาร์กิวเมนต์ที่เกิดร่วมกับกริยา ซึ่งเป็นวิธีที่คล้ายกับงานของคอสโลว์สกีและคณะ (Kozlowski et al., 2003) แต่แตกต่างกันในส่วนที่งานชิ้นนี้ใช้ชนิดของกริยาและบทบาททางความหมายที่ถูกกำหนดไว้แล้วในงานของ ภาณุ สังขะวร (2527) ที่แบ่งบทบาททางความหมายไว้ 16 บทบาทและแบ่งชนิดของกริยาออกเป็น 18 ชนิด ยกตัวอย่างการวิเคราะห์คู่ถอดความเช่น

Src. ตำรวจ/a policeman-Agent ต่อสู้/fight-ReciprocityAction
กับ/with-ParallelMarker อาชญากร/a criminal-Participant
A policeman fight with a criminal.

Pa. อาชญกร/a criminal-Agent ต่อสู้/fight-ReciprocityAction
กับ/with-ParallelMarker ตำรวจ/a policeman-Participant
A criminal fight with a policeman.

จากตัวอย่างจะเห็นได้ว่าประโยคต้นฉบับและประโยคที่ถูกถอดความนั้นมีการสลับตำแหน่งระหว่างอาร์กิวเมนต์ *ตำรวจ* และ *อาชญากร* ที่เป็นเช่นนี้ได้เพราะกริยาในประโยคดังกล่าวเป็นกริยาประเภทการกระทำซึ่งกันและกัน (reciprocity action) ซึ่งแสดงการกระทำระหว่างอาร์กิวเมนต์อย่างเท่าเทียม ทำให้สามารถจัดประเภทการถอดความประเภทนี้ออกมาได้เรียกว่าการสลับอาร์กิวเมนต์ในการกระทำซึ่งกันและกัน (argument switching in reciprocity action) จึงแสดงให้เห็นว่าการวิเคราะห์ตามกรอบบทบาททางความหมายและชนิดของกริยาช่วยให้สามารถวิเคราะห์และจัดประเภทการถอดความได้

ด้วยวิธีนี้ กฤตาพร พัชระสุภา และพรฤดี เนติโสภากุลสามารถจัดประเภทของการถอดความในภาษาไทยได้อย่างละเอียดโดยแบ่งออกถึง 14 ประเภท ตั้งแต่ระดับคำศัพท์พื้นฐานที่สามารถแทนที่ได้ด้วยคำไวพจน์ การแสดงเจตนาในระดับประโยค ไปจนกระทั่งถึงระดับพหุอนุพากย์ (multi-clause level)

กรอบการวิเคราะห์ที่ปรากฏในงานของกฤตาพร พัชระสุภา และพรฤดี เนติโสภากุล นี้ถือเป็นวิธีการวิเคราะห์ที่น่าสนใจ โดยเฉพาะอย่างยิ่งกับการวิเคราะห์การถอดความในภาษาไทยที่การแสดงความสัมพันธ์ระหว่างนามและกริยาไม่ปรากฏให้เห็นเด่นชัดเมื่อเปรียบเทียบกับภาษาญี่ปุ่นหรือภาษาอังกฤษในงานที่ได้กล่าวตอนต้น

เมื่อกล่าวถึงงานที่วิจัยให้ความสนใจกับการถอดความแล้ว งานวิจัยอีกแง่มุมหนึ่งที่ต้องกล่าวถึงในที่นี้คืองานที่มุ่งวิเคราะห์ความสัมพันธ์ระหว่างการถอดความกับการลักลอกโดยตรง งานชิ้นดังกล่าวได้แก่งานของบาร์รอน-เซเดโญและคณะ (Barrón-Cedeño et al., 2013) ซึ่งเสนอว่าการถอดความเป็นกลไกทางภาษาที่อยู่เบื้องหลังการลักลอก ทั้งยังเป็นกระบวนการที่การลักลอกใช้เป็นหลัก การทำความเข้าใจปรากฏการณ์ทางภาษาในการถอดความจึงช่วยแก้ไขปัญหาการตรวจจับการลักลอกได้ ทั้งนี้ บาร์รอน-เซเดโญและคณะได้แบ่งขั้นการวิเคราะห์เป็น 2 ขั้นตอน ขั้นแรกคือเสนอแบบลักษณ์การลักลอก โดยได้ประยุกต์ใช้แบบลักษณ์การลักลอกที่วิลาและคณะ (Vila et al., 2011) เสนอไว้แต่จัดแบ่งการลักลอกใหม่เป็น 3 ประเภท ขั้นต่อมาจึงสร้างคลังข้อมูลการลักลอกขึ้น และใช้กลไกทางภาษาที่ปรากฏในแบบลักษณ์การลักลอกในการกำกับข้อมูล จากการวิเคราะห์คลังข้อมูลการลักลอกทำให้คณะผู้วิจัยได้ข้อค้นพบ 3 ประการ ประการแรก ความยากง่ายในการตรวจจับการลักลอกขึ้นอยู่ความซับซ้อนของกลไกที่ใช้ในการถอดความ ประการต่อมา การแทนที่ด้วยคำศัพท์เป็นกลไกการถอดความที่พบมากที่สุดในการลักลอก และประการสุดท้าย กลไกการถอดความมีแนวโน้มที่จะทำให้ข้อความที่ถูกลักลอกมีขนาดสั้นลงจากข้อความต้นฉบับ

จากการทบทวนการจัดประเภทการถอดความจากต้นฉบับที่ผ่านมา พบว่าการถอดความนั้นมุ่งวิเคราะห์รูปภาษาทั้งระดับคำศัพท์ ระดับประโยค และระดับความหมายเหมือนกัน และประเภทของการถอดความที่งานทุกชิ้นกล่าวถึงตรงกัน ได้แก่ การถอดความโดยใช้คำไวพจน์แทนที่ การถอดความประโยคกรรตุเป็นประโยคกรรม การถอดความโดยการทำให้เป็นคำนาม และการถอดความในการกกรรมรอง ส่วนข้อแตกต่างนั้นเป็นรายละเอียดปลีกย่อยถึงแสดงถึงลักษณะเฉพาะของแต่ละภาษา นอกจากนี้ ข้อแตกต่างอีกประการหนึ่งคืองานแต่ละชิ้นนั้นอาศัยความรู้ทางภาษาศาสตร์ในการวิเคราะห์และจัดประเภทมากน้อยแตกต่างกันไป ทั้งนี้ จะสังเกตได้ว่างานที่พึ่งพาทฤษฎีภาษาศาสตร์ในการจัดประเภทมากก็จะได้จำนวนประเภทที่ละเอียดและมีความน่าเชื่อถือมากกว่า

ส่วนในแง่การลักลอกนั้น การถอดความยังถือเป็นกลวิธีสำคัญที่ถูกใช้ในการลักลอก อย่างไรก็ตาม ผู้วิจัยใคร่ขอชี้ให้เห็นถึงข้อน่าสังเกตประการหนึ่งคือ งานทุกชิ้นที่กล่าวมาข้างต้นมุ่งวิเคราะห์การถอดความที่เกิดขึ้นในระดับประโยคเท่านั้น ทั้งนี้ ในการสถานการณ์การลักลอกจริงแล้ว ผู้ลักลอกมิได้ลักลอกโดยถอดความประโยคต่อประโยค หากแต่เป็นการจัดการกับข้อความในระดับย่อหน้าหรือปริจเฉท ประเด็นนี้เองที่จูงใจให้ผู้วิจัยสนใจทำความเข้าใจกลวิธีลักลอกในระดับปริจเฉทโดยใช้ทฤษฎี Rhetorical Structure Theory ซึ่งจะได้กล่าวในรายละเอียดต่อไป

นอกจากนี้ ประเด็นสำคัญอีกประเด็นหนึ่งที่เชื่อมโยงการลักลอกและการถอดความไว้ด้วยกันคือปรากฏการณ์ที่เกิดขึ้นในเชิงวากยสัมพันธ์และคำศัพท์ จากงานของวิลาและคณะ (Vila et al., 2011) ก็ดี งานของภาคัตและโฮวี (Bhagat & Hovy, 2013) ก็ดี หรืองานของบาร์รอน-เซเดโญและ

คณะ (Barrón-Cedeño et al., 2013) ก็ดี งานทุกชิ้นล้วนกล่าวถึงกลวิธีถอดความได้อย่างน่าสนใจ ผู้วิจัยสามารถยึดเป็นแนวทางในการวิเคราะห์และจำลองการลักลอกได้ โดยเฉพาะอย่างยิ่ง งานของกฤตาพร พัชระสุภา และพรฤดี เนติโสภากุล (Phucharasupa & Netisopakul, 2012) ที่ใช้บทบาททางความหมายเป็นกรอบการวิเคราะห์นั้นถือเป็นแนวทางที่เอื้อต่อการวิเคราะห์การลักลอกในภาษาไทย ด้วยเหตุนี้ ในงานวิจัยชิ้นนี้ ผู้วิจัยจึงเลือกใช้แนวคิดเรื่องบทบาททางความหมายที่อาร์กิวเมนต์มีต่อภาคแสดงเป็นกรอบในการวิเคราะห์และทำความเข้าใจลักษณะทางวากยสัมพันธ์และคำศัพท์ของการถอดความเป็นสำคัญ

2.4 แนวทางตรวจหาการลักลอกงานวิชาการ

แนวทางในการตรวจหาการลักลอกงานวิชาการนั้นสามารถแบ่งออกได้เป็น 2 แนวทางหลัก ได้แก่ แนวทางตรวจหาการลักลอกงานวิชาการโดยมนุษย์ และแนวทางตรวจหาการลักลอกงานวิชาการด้วยเครื่อง ในหัวข้อนี้ ผู้วิจัยได้รวบรวมรายละเอียดของการตรวจหาทั้ง 2 แนวทางเอาไว้ ดังนี้

2.4.1 แนวทางตรวจหาการลักลอกงานวิชาการโดยมนุษย์

สุวิมล ว่องวาณิช และวิไลวรรณ ศรีสงคราม (2553 อ้างถึงใน กัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์, 2554, น. 76) ได้กล่าวถึงวิธีการตรวจหาการลักลอกในแง่ที่ผู้ตรวจมิได้พึ่งพาซอฟต์แวร์ที่ใช้ในการตรวจหาการลักลอกงานวิชาการไว้ว่า ผู้ตรวจควรใช้หลายวิธีควบคู่กันไปทั้งจากประสบการณ์ของผู้ตรวจเอง รวมถึงการวิเคราะห์ทางด้านภาษา เพื่อให้ได้ประสิทธิภาพในการตรวจหาที่ถูกต้องและแม่นยำ

ในแง่การตรวจหาโดยใช้ประสบการณ์ส่วนตัวหรือใช้สามัญสำนึก (common sense) นั้น สุวิมลและวิไลวรรณได้กล่าวว่าเป็นวิธีที่เหมาะกับอาจารย์ผู้สอนที่มีความรู้และคุ้นเคยกับสาขาวิชาที่ตนสอนอยู่ เพราะบุคคลกลุ่มนี้ได้อ่านและประเมินงานของนักศึกษามาพอสมควร จึงสามารถใช้ความรู้สึก ความรู้ และประสบการณ์ที่มีอยู่สืบจับงานที่ลอกมาได้ อย่างไรก็ตาม สุวิมลและวิไลวรรณก็ได้เสนอว่าการตรวจหาด้วยวิธีนี้อาจจำเป็นต้องใช้ตัวชี้วัดในด้านภาษาเข้ามาร่วมในการพิจารณาด้วย เพราะลักษณะทางภาษาก็เป็นอีกส่วนหนึ่งที่สะท้อนการลักลอกงานของนักศึกษาได้

การวิเคราะห์ทางด้านภาษา (linguistic analysis) เป็นวิธีการตรวจหาการลักลอกงานอีกวิธีหนึ่งที่สุวิมลและวิไลวรรณได้เสนอไว้ โดยตรวจสอบคุณภาพของลักษณะภาษาที่ใช้ในเชิงปริมาณภายใต้แนวคิดที่ว่าผู้เขียนแต่ละคนย่อมมีลีลา (style) การเขียนเป็นของตนเอง หากรู้สึกว่าลีลาบางตอนผิดแผกไปจากภาพรวมของเอกสาร ก็ให้ตั้งข้อสงสัยได้ว่างานชิ้นดังกล่าวมีการลักลอกปนอยู่ การตรวจหาในแง่นี้สอดคล้องกับข้อสังเกตของเมเยอร์ ซู ไอส์เซนและคณะ (Meyer zu Eissen, Stein, &

Kulig, 2007, p. 360) ที่กล่าวว่า ผู้อ่านงานที่เป็นมนุษย์นั้นอาจระบุได้ว่าข้อเขียนใดเป็นข้อเขียนที่ต้องสงสัยว่ามีการลักลอกได้แม้จะไม่มีแหล่งข้อมูลของเอกสารอ้างอิงอยู่ในใจ ทั้งนี้ การเปลี่ยนแปลงของภาษาจากที่สละสลวยกลายเป็นการใช้ภาษาที่คลุมเครือหรือก่อให้เกิดความสับสนในข้อเขียน หรือการเปลี่ยนการเรียกแทนตัวของผู้เขียน เหล่านี้ล้วนเป็นการบอกใบ้ถึงการลักลอกทั้งสิ้น

แนวคิดด้านการวิเคราะห์ภาษาโดยไม่ใช้เอกสารอ้างอิงเพื่อตรวจเทียบดังกล่าวข้างต้นได้พัฒนาต่อมาเป็นแนวคิดหลักของแนวทางการตรวจหาการลักลอกงานวิชาการด้วยเครื่องอีกแนวทางหนึ่งที่เรียกว่า การตรวจหาการลักลอกภายใน (intrinsic plagiarism detection) ในที่นี้ ผู้วิจัยจะขอยกตัวอย่างลักษณ์ที่ใช้ในการตรวจหาการลักลอกภายนอกมาแสดงให้เห็นเพื่อให้ประยุกต์ใช้เป็นแนวทางการวิเคราะห์ภาษาในกรณีที่ผู้ตรวจเป็นมนุษย์ได้ ทั้งนี้ ลักษณ์ดังกล่าวเป็นลักษณ์ที่ใช้ในการพิจารณาลีลาการเขียนเป็นสำคัญเรียกว่าลักษณ์ลีลามาตร (stylometric feature) (Meyer zu Eissen et al., 2007, p. 361) ลักษณ์ดังกล่าวสามารถพิจารณาได้จากรายละเอียดต่อไปนี้

1) ลักษณ์ทางข้อความ ลักษณ์ประเภทนี้จะพิจารณาในระดับอักขระเป็นหลัก ในกรณีที่ผู้ตรวจเป็นมนุษย์อาจพิจารณาได้จากการสังเกตความถี่บ่อยของการใช้เครื่องหมายวรรคตอนของผู้เขียน เช่น เครื่องหมายจุลภาค หรือเครื่องหมายปรัศนี เป็นต้น
2) ลักษณ์ทางวากยสัมพันธ์ เป็นการพิจารณาจากลีลาการเขียนในระดับประโยค ทั้งนี้ ผู้ตรวจอาจสังเกตได้จากความยาวของประโยค การผูกประโยคให้ซับซ้อน หรือการใช้คำหน้าที่ ได้ เช่น หากภาพรวมของเอกสารนั้น ผู้เขียนใช้ประโยคที่มีขนาดสั้นและผู้ประโยคง่ายมาโดยตลอด ส่วนของเอกสารที่มีประโยคขนาดยาวหรือมีการผูกประโยคที่ขยายอย่างซับซ้อนก็ให้ตั้งข้อสงสัยได้ว่าส่วนดังกล่าวมีการลักลอกปนอยู่
3) ลักษณ์หมวดคำ ลักษณ์ประเภทนี้เป็นการให้พิจารณาจำนวนของการใช้คำในแต่ละหมวดเปรียบเทียบกันว่าแต่ละส่วนของเอกสารมีปริมาณเฉลี่ยของการใช้คำในแต่ละหมวดผิดแผกไปจากภาพรวมของทั้งเอกสารหรือไม่ ลักษณ์ประเภทนี้อาจใช้พิจารณาได้ไม่สะดวกเท่าใดนักในกรณีที่ผู้ตรวจเป็นมนุษย์ อย่างไรก็ตาม การพิจารณาในส่วนที่คำคุณศัพท์หรือคำสรรพนามก็ยังอาจจะทำได้ เช่น การใช้สรรพนามเรียกแทนตัวผู้เขียนที่มีการเปลี่ยนแปลงหรือใช้ไม่สม่ำเสมอ เป็นต้น
4) ลักษณ์ชุดหมวดคำปิด ลักษณ์ประเภทนี้ให้ผู้ตรวจสังเกตลีลาการใช้คำพิเศษในการเขียนเป็นหลัก ไม่ว่าจะเป็นคำหยุด (stop word) คำภาษาต่างประเทศ หรือคำยาก ลีลาการเขียนที่ผิดแผกไปนั้นอาจเกิดจากการจงใจใช้คำประเภทนี้โดยไม่มีความจำเป็น ซึ่งจะสังเกตพบได้โดยเฉพาะกรณีที่ลักลอกโดยการถอดความที่ผู้เขียนต้องการจะใช้คำไวพจน์แทนที่คำในต้นฉบับ

5) ลักษณ์เชิงโครงสร้าง ลักษณ์ประเภทสุดท้ายนี้พิจารณาภาพสะท้อนของการเรียบเรียงข้อความ ผู้ตรวจสามารถสังเกตได้จากขนาดและความยาวของย่อหน้าและบทว่ามีความผิดแผกไปจากภาพรวมหรือไม่ เช่น หากย่อหน้าหรือบทใดมีขนาดยาวหรือสั้นกว่าปกติให้สงสัยได้ว่าย่อหน้าหรือข้อความดังกล่าวมีการลักลอกปะปนอยู่

ในส่วนของการตรวจหาการลักลอกที่ผ่านการถอดความมาโดยเฉพาะนั้น เซาซา-ซิลวาและคณะ (Sousa-Silva, Grant, & Maia, 2010) ได้วิเคราะห์หาลักษณ์ที่ใช้การระบุการถอดความในการลักลอกงานวิชาการไว้โดยเฉพาะโดยเก็บข้อมูลจากงานของนักศึกษาโดยตรง ผลปรากฏว่าลักษณ์ที่บ่งชี้ถึงการถอดความในการลักลอกได้ดีนั้นได้แก่ การแทนที่คำในสาระสำคัญด้วยคำที่มีความหมายสัมพันธ์กัน และการจงใจเปลี่ยนแปลงลำดับคำในประโยคอย่างเห็นได้ชัด ตามลำดับ จะเห็นได้ว่าลักษณ์ทั้ง 2 ประเภทดังกล่าวนั้นตรงกับหลักที่ใช้ในการถอดความโดยทั่วไป ทั้งนี้ ผู้ตรวจสามารถใช้ลักษณ์ดังกล่าวในการตรวจหาการลักลอกได้ กล่าวคือ ผู้ตรวจอาจพิจารณาประโยคใจความสำคัญของข้อเขียนว่ามีการใช้คำที่ไม่คุ้นตาหรือไม่ ทั้งนี้ คำที่ไม่คุ้นตาดังกล่าวอาจเป็นคำไวพจน์ที่ถูกแทนที่มาด้วยการถอดความ หรืออาจพิจารณาการเรียงลำดับคำในประโยคว่าเป็นการเรียงลำดับคำให้แปลกตาโดยไม่จำเป็นหรือไม่ หากพบลักษณะดังกล่าวก็ให้ตั้งข้อสงสัยได้ว่าเป็นการลักลอกที่ผ่านการถอดความมา

วิธีการตรวจหาการลักลอกวิธีสุดท้ายที่ผู้วิจัยจะกล่าวถึงในที่นี้เป็นวิธีที่ปรากฏในงานของเมาเรอร์และคณะ (Maurer, Kappe, & Zaka, 2006, p. 1058) วิธีดังกล่าวได้แก่การค้นหาวลีที่มีลักษณะเฉพาะด้วยตนเอง (manual search of characteristic phrases) วิธีนี้ทำได้โดยให้ผู้ตรวจเลือกวลีหรือประโยคที่นำเสนอความคิดรวบยอดของงานวิชาการออก จากนั้นจึงใช้เครื่องมือค้นหา (search engine) ค้นหาวลีหรือประโยคดังกล่าวในอินเทอร์เน็ต หากค้นหาพบก็ให้สงสัยได้ว่างานชิ้นดังกล่าวมีการลักลอกมา ด้วยวิธีนี้ผู้ตรวจอาจพบทั้งการลักลอกแบบคำต่อคำไปจนถึงการลักลอกโดยการถอดความ ผู้ตรวจจึงอาจต้องใช้ลักษณ์ในการวิเคราะห์ลีลาการเขียนที่กล่าวมาข้างต้นในการพิจารณาการถอดความอีกชั้นหนึ่ง

ทั้งนี้ ดังได้กล่าวไปแล้วในตอนต้นว่าในการตรวจหาการลักลอกงานวิชาการหรือการถอดความนั้น เพื่อให้เกิดความถูกต้องและความแม่นยำมากที่สุด ตัวผู้ตรวจเองควรตรวจหาด้วยวิธีที่หลากหลายทั้งการใช้สามัญสำนึกส่วนตัวในฐานะผู้คุ้นเคยกับสาขาวิชานั้นๆ และวิธีการวิเคราะห์ทางภาษาด้วยลักษณ์ที่แสดงถึงลีลาการเขียนส่วนตัวของผู้เขียน หรืออาจพึ่งพาการค้นข้อมูลจากอินเทอร์เน็ตด้วยก็ย่อมได้ อย่างไรก็ตาม แนวทางการตรวจหาการลักลอกงานวิชาการโดยมนุษย์นี้ก็มีข้อจำกัดหลายด้าน ไม่ว่าจะเป็นด้านทรัพยากรบุคคลที่อาจมีไม่เพียงพอหรือไม่มีมาตรฐานในการตรวจหาที่คงที่ในกรณีที่ต้องตรวจหาการลักลอกในปริมาณมาก หรือด้านเวลาที่แน่นอนว่าการ

ตรวจหาการลักลอกโดยมนุษย์ย่อมกินเวลาพอสมควร ซึ่งข้อจำกัดที่ได้กล่าวมานี้ก็เป็นปัจจัยสำคัญประการหนึ่งที่ผลักดันให้เกิดงานวิจัยชิ้นนี้ขึ้น

2.4.2 แนวทางตรวจหาการลักลอกงานวิชาการด้วยเครื่อง

แนวทางตรวจหาการลักลอกงานวิชาการด้วยเครื่องเป็นแนวทางที่ใช้ในการพัฒนาระบบตรวจหาการลักลอกเพื่อหลีกเลี่ยงข้อจำกัดที่อาจเกิดขึ้นได้ในการตรวจหาการลักลอกโดยมนุษย์ ทั้งนี้เป็นที่ทราบกันดีว่า ในการพัฒนาระบบตรวจหาการลักลอกนั้น ก่อนอื่นต้องทำความเข้าใจถึงแนวทางการตรวจหาการลักลอกด้วยเครื่องเสียก่อน เพื่อจะได้เลือกใช้ประเภทที่เหมาะสมกับลักษณะของข้อมูลที่ต้องการตรวจหาได้ แนวทางการตรวจหาการลักลอกด้วยเครื่องนี้สามารถแบ่งย่อยออกได้เป็นอีก 2 แนวทาง ได้แก่ การตรวจเทียบภายนอกหาการลักลอก (extrinsic plagiarism detection) และการตรวจเทียบภายในหาการลักลอก (intrinsic plagiarism detection)

สำหรับการตรวจเทียบภายนอกหาการลักลอกนั้น แนวคิดที่สำคัญในด้านระเบียบวิธีที่เกี่ยวข้องกับการตรวจหาคือ เพื่อเปรียบเทียบเอกสารที่ต้องสงสัยว่าเป็นการลักลอกกับแหล่งข้อมูลอื่นๆ โดยแหล่งข้อมูลดังกล่าวอาจมาจากแหล่งเดียวหรือเกิดจากการรวบรวมจากที่มาหลายแหล่งก็ได้ (Alzahrani et al., 2012, p. 137) ด้วยแนวคิดนี้ระบบจะประเมินว่าเอกสารที่ต้องสงสัยมีความละม้ายกับแหล่งข้อมูลที่นำมาเปรียบเทียบหรือไม่อย่างไร จากนั้นจึงรายงานผลดังกล่าวออกมา

ส่วนแนวคิดที่ใช้ในการตรวจเทียบภายในหาการลักลอกคือ เพื่อวิเคราะห์ว่าเอกสารที่ต้องสงสัยนั้นมีความเปลี่ยนแปลงในด้านลีลาการเขียนหรือไม่ โดยมีความเชื่อพื้นฐานว่าผู้เขียนคนเดียวกันย่อมมีลีลาการเขียนที่คงที่ในระดับหนึ่ง (Stein, Lipka, & Prettenhofer, 2011, p. 64) ด้วยวิธีนี้การตรวจหาจึงสามารถทำได้โดยมีเอกสารที่ต้องสงสัยเพียงชิ้นเดียวเท่านั้น เพราะการเปรียบเทียบที่เกิดขึ้นจะเป็นไปภายในตัวเอกสารเอง ไม่จำเป็นต้องเปรียบเทียบจากแหล่งอื่น

ทั้งนี้ หากพิจารณาถึงวัตถุประสงค์ของงานวิจัยชิ้นนี้แล้ว ผู้วิจัยเห็นว่าแนวทางตรวจเทียบภายนอกนั้นเหมาะสมจะใช้ตรวจหาการลักลอกในงานวิชาการมากกว่าแนวทางการตรวจเทียบภายใน ทั้งนี้เนื่องจากงานวิชาการนั้นจำเป็นต้องอ้างอิงข้อมูลจากแหล่งอื่นเป็นส่วนใหญ่ การตรวจเทียบภายในเฉพาะลีลาการเขียนอาจไม่ครอบคลุมถึงการลักลอกข้อมูลจากแหล่งอื่น ดังนั้นการตรวจเทียบภายนอกซึ่งเป็นการตรวจเทียบในเชิงเนื้อหาของข้อความจึงตรงกับวัตถุประสงค์ของงานวิจัยมากกว่า

2.5 วิธีตรวจเทียบภายนอกหาการลักลอกงานวิชาการ

ดังได้กล่าวไปแล้วในหัวข้อที่ผ่านมาว่างานวิจัยชิ้นนี้จะใช้แนวทางการตรวจเทียบภายนอกหาการลักลอก ดังนั้นในหัวข้อนี้ ผู้วิจัยจะกล่าวถึงวิธีตรวจหาการลักลอกที่นิยมใช้โดยทั่วไปในแนวทางการตรวจเทียบภายนอก โดยมีรายละเอียดดังต่อไปนี้

2.5.1 วิธีอิงสายอักขระ (String-based method)

วิธีตรวจหาโดยอิงสายอักขระถือได้ว่าเป็นวิธีตรวจหาที่ได้รับความนิยมและใช้อย่างกว้างขวางในการตรวจหาการลักลอก อัลกอริทึมการตรวจหาที่ใช้วิธีนี้จะเปรียบเทียบเอกสารสอบถาม (query document) d_q กับเอกสารที่เป็นไปได้ (candidate document) $d_x \epsilon D_x$ ในระดับอักขระของคำหรือประโยค ซึ่งการจับคู่สายอักขระในปริบทอาจเป็นไปในแบบตรงตัว (exact) หรือแบบใกล้เคียง (approximate) ก็ได้

ในกรณีของการจับคู่สายอักขระแบบตรงตัวระหว่างสายอักขระนั้น การจับคู่จะดำเนินการโดยมีแนวคิดว่าสายอักขระทั้งสองมีอักขระเหมือนกันในลำดับที่เหมือนกัน ยกตัวอย่างเช่นในกรณีของสายอักขระที่มี 8 แกรม สายอักขระ x = "*aaabbbcc*" จะสามารถจับคู่แบบแม่นยำได้กับ "*aaabbbcc*" แต่ไม่สามารถจับคู่ได้กับสายอักขระ y = "*aaabbbcd*" ส่วนในอีกกรณีหนึ่งที่เป็นการจับคู่สายอักขระแบบใกล้เคียงนั้นจะแสดงค่าความเหมือนหรือไม่เหมือนกันระหว่างสายอักขระ 2 สาย เช่นในกรณีของอักขระ 9 แกรม อย่าง x = "*aaabbbccc*" และ y = "*aaabbbccd*" สายอักขระทั้งสองจะถือได้ว่ามีค่าความละม้ายกันสูง เพราะอักขระทุกตัวสามารถจับคู่กันได้ยกเว้นตัวสุดท้าย ค่าความละม้ายนี้ได้นำไปสู่แนวคิดของการวัดค่าระยะระหว่างสายอักขระ "$d(x,y)$" เพื่อใช้พิจารณาว่าเอกสารที่ต้องสงสัยถูกคัดลอกมาจากเอกสารต้นฉบับที่เป็นไปได้หรือไม่ โดยให้นิยามของค่าดังกล่าวไว้ว่าเป็นค่าที่น้อยที่สุดในการแปลง x ไปสู่ y (Navarro, 2001, p. 37) ด้วยแนวคิดนี้จึงทำให้การจับคู่สายอักขระแบบใกล้เคียงสามารถรองรับการตรวจวัดความละม้ายระหว่างสายอักขระ 2 สายในกรณีที่มีการแทรก (insertion) การลบ (deletion) การสลับที่ (substitution) หรือการย้ายที่ (transportation) อักขระในสายอักขระได้ การจับคู่ลักษณะนี้ใช้อย่างกว้างขวางในการตรวจหาการลักลอกและการใช้ตัวบทซ้ำ ทั้งนี้ สามารถยกตัวอย่างค่าการแก้ไขระยะที่ใช้กันอย่างแพร่หลายได้ เช่น ระยะการแก้ไขเลเวนชเตย์น (Levenshtein edit distance) และระยะของลำดับย่อยร่วมที่ยาวที่สุด (longest common subsequence distance) เป็นต้น ส่วนในด้านการนำไปประยุกต์ใช้นั้น งานตรวจหาการลักลอกที่ใช้วิธีนี้สามารถยกตัวอย่างได้เช่นงานของบาซีเลและคณะ (Basile et al., 2009) หรืองานของกรอเซียและคณะ (Grozea, Gehl, & Popescu, 2009) เป็นต้น

แม้วิธีอิงสายอักขระจะเป็นที่นิยมใช้กันอย่างแพร่หลายในระบบตรวจหาการลักลอก เนื่องจากมีขั้นตอนที่ไม่ซับซ้อนทั้งในขั้นก่อนการประมวลผล (pre-processing) หรือการจัดการคลังข้อมูล อีกทั้งไม่จำเป็นต้องประยุกต์ใช้ความรู้ทางภาษาศาสตร์ แต่ก็พบว่าระบบที่ใช้วิธีนี้ไม่สามารถให้ผลได้ในระดับที่น่าพอใจ โดยเฉพาะในกรณีที่ข้อความถูกลักลอกโดยการเปลี่ยนลำดับของคำหรือวลี การถอดความ หรือการสรุปความ

2.5.2 วิธีอิงเวกเตอร์ (Vector-based method)

เช่นเดียวกับวิธีที่ได้กล่าวไปก่อนหน้านี้ วิธีอิงเวกเตอร์ก็มีแนวคิดสำคัญอยู่ที่การตรวจหาการลักลอกจากค่าความละม้ายที่วัดได้จากข้อมูลรับเข้า ค่าดังกล่าวสามารถคำนวณได้เป็นค่าสัมประสิทธิ์ความละม้ายของเวกเตอร์ การตรวจหาตามวิธีนี้มักทำในระดับคำ กล่าวคือ คำในข้อมูลรับเข้าจะถูกพิจารณาเป็นเอ็นแกรมของคำ แล้วจึงแปลงเอ็นแกรมของคำดังกล่าวให้เป็นเวกเตอร์ของ *n* token หรือเวกเตอร์ของเอ็นแกรมของคำ จากนั้นจะสามารถประเมินความเหมือนกันได้โดยใช้การจับคู่ ทั้งนี้ วิธีการวัดความละม้ายของเวกเตอร์ที่นิยมใช้กันในการตรวจหาการลักลอกนั้นมี 2 วิธี ได้แก่ สัมประสิทธิ์ความละม้ายโคไซน์ (cosine similarity coefficient) และสัมประสิทธิ์ความละม้ายแจ็กการ์ด (Jaccard similarity coefficient) ยกตัวอย่างเช่นที่ปรากฏในงานของบาร์รอน-เซเดโญและคณะ (Barrón-Cedeño, Basile, Degli Esposti, & Rosso, 2010) เป็นต้น

วิธีการตรวจหาการลักลอกแบบอิงเวกเตอร์นี้เป็นวิธีการตรวจหาอีกวิธีหนึ่งที่ผู้วิจัยเห็นว่าควรนำมาพัฒนาต่อยอดให้มีประสิทธิภาพมากขึ้น เพราะเดิมนั้นผู้ที่เลือกใช้วิธีนี้มักเลือกแปลงเวกเตอร์จากสายอักขระหรือสายคำ ทำให้ประสิทธิภาพของการตรวจหาการลักลอกถูกจำกัดอยู่ในระดับการคัดลอกโดยดัดแปลงเท่านั้น ทั้งนี้ หากได้ทดลองแทนรูปข้อมูลรับเข้าด้วยวิธีอื่นๆ ที่อาศัยความรู้ทางภาษาศาสตร์มากกว่าเดิมและแปลงให้เป็นเวกเตอร์ ก็เชื่อได้ว่าวิธีนี้จะให้ประสิทธิภาพในการตรวจหาการลักลอกที่มากขึ้น

2.5.3 วิธีอิงวากยสัมพันธ์ (Syntax-based method)

วิธีตรวจหาการลักลอกโดยอิงวากยสัมพันธ์นั้นกล่าวได้ว่ามีความแตกต่างจาก 2 วิธีที่กล่าวถึงไปข้างต้น เนื่องจากมีการนำความรู้ทางภาษาศาสตร์มาใช้วัดค่าความละม้ายในกระบวนการตรวจหา นั่นคือข้อมูลทางวากยสัมพันธ์ ทั้งนี้ ข้อมูลทางวากยสัมพันธ์ที่จะกล่าวถึงในที่นี้ได้แก่หมวดคำ (Part of speech: POS) และลำดับคำ (word order)

ในการตรวจหาการลักลอกด้วยหมวดคำนั้นจะต้องนำข้อมูลทั้งหมดมากำกับหมวดคำ (POS tagging) ก่อนตามสายอักขระที่ปรากฏในเอกสาร จากนั้นจึงวิเคราะห์และคำนวณค่าความละม้ายระหว่างเอกสารออกมา โดยมีแนวคิดพื้นฐานว่าเอกสารที่เหมือนกันจากการคัดลอกโดยตรงย่อมมี

ลำดับของป้ายกำกับทางหมวดคำเหมือนกัน ในการนี้อาจมีการผลิตตัวกำกับหมวดคำที่มีลักษณะเฉพาะเพื่อจุดประสงค์ในการวัดค่าความเหมือนกันโดยเฉพาะด้วยเช่นกัน ยกตัวอย่างเช่นงานของเอลฮาดีและอัลโตบี (Elhadi & Al-Tobi, 2008) ที่ได้ใช้โปรแกรม TreeTagger กำกับหมวดคำในข้อมูลและใช้อัลกอริทึมลำดับร่วมที่ยาวที่สุด (longest common subsequences: *lcs*) ในการจับคู่สายหมวดคำ เพื่อความเข้าใจที่ชัดเจนยิ่งขึ้น ผู้วิจัยขอยกตัวอย่างการตรวจหาการลักลอกด้วยหมวดคำในข้อความภาษาไทยต่อไปนี้

S_1 = “แมว|ขโมย|ปลา|ของ|น้อง|ไป|กิน”

S_2 = “วิฬาร์|ขโมย|มัจฉา|ของ|น้อง|ไป|กิน”

จากตัวอย่างนี้ข้างต้น จะเห็นได้ว่าข้อความ S_1 และ S_2 มีความหมายเหมือนกัน แต่ข้อความ S_2 เลือกใช้คำไวพจน์ “วิฬาร์” และ “มัจฉา” แทนที่คำว่า “แมว” และ “ปลา” ในข้อความ S_1 ตามลำดับ การประยุกต์ใช้แนวคิดเรื่องหมวดคำสามารถช่วยแก้ไขปัญหาข้างต้นได้ส่วนหนึ่ง ในที่นี้จะแสดงให้เห็นโดยการแทนรูปข้อความ S_1 และ S_2 ข้างต้นให้อยู่ในรูปลำดับของหมวดคำ ดังนี้

$S_{1\text{-POS}}$ = “NOUN|VERB|NOUN|ADP|NOUN|VERB|VERB”

$S_{2\text{-POS}}$ = “NOUN|VERB|NOUN|ADP|NOUN|VERB|VERB”

จากตัวอย่าง $S_{1\text{-POS}}$ และ $S_{2\text{-POS}}$ เป็นลำดับของหมวดคำที่แทนรูปได้จากข้อความ S_1 และ S_2 ตามลำดับ จะเห็นได้ว่าเมื่อแทนรูปเป็นหมวดคำแล้ว ลำดับของหมวดคำของข้อความ S_1 และ S_2 เหมือนกันทุกประการ ด้วยวิธีการดังกล่าวนี้แทนที่ระบบจะตรวจหาข้อมูลในระดับสายอักขระหรือสายคำ ข้อมูลที่ระบบตรวจหาจะมีความเป็นนามธรรมในระดับมโนทัศน์ที่สะท้อนผ่านคำได้มากขึ้น เพราะระบบอาจตรวจจับการแทนที่คำอื่นในหมวดคำเดียวกันแต่สะท้อนมโนทัศน์เดียวกันหรือใกล้เคียงกันได้ ทำให้สามารจับลีลาของผู้เขียนหรือข้อมูลทางความหมายบางประการได้

ส่วนลำดับคำ (word order) นั้นก็สามารถนำมาประยุกต์ใช้ในการตรวจหาการลักลอกในฐานะข้อมูลทางวากยสัมพันธ์ได้ โดยใช้ในการเปรียบเทียบความละม้ายในระดับประโยค ยกตัวอย่างเช่นในงานของลีและคณะ (Y. Li et al., 2004; Y. Li, McLean, Bandar, O'Shea, & Crockett, 2006) ที่เปรียบเทียบความละม้ายของประโยคโดยกำหนดลำดับให้คำแต่ละคำในประโยค จากนั้นจึงแปลงเป็นเวกเตอร์ของลำดับคำเพื่อใช้หาความละม้ายของประโยคต่อไป ยกตัวอย่างเช่น

T_1: RAM keeps things being worked with.

T_2: The CPU uses RAM as a short-term memory store.

T= [RAM, keeps, things, being, worked, with, The, CPU, uses, as, a, short-term, memory, store]

จากข้างต้น จะเห็นได้ว่าข้อความ T_1 และ T_2 มีคำที่ปรากฏอยู่ร่วมกันจำนวนหนึ่งแต่ปรากฏในลำดับที่แตกต่างกัน ในการนี้ ลีและคณะได้เสนอให้สร้างชุดคำรวมของคู่หน่วยเทียบ T แล้ว จากนั้นสร้างเวกเตอร์ของลำดับคำ r_1 และ r_2 ขึ้นโดยเทียบกับคำที่ปรากฏในประโยค T_1 และ T_2 และลำดับของคำในชุดคำรวมของคู่หน่วยเทียบ T ดังนี้

r_1 = {1 2 3 4 5 6 0 3 3 0 0 0 1 1}

r_2 = {4 0 3 0 0 0 1 2 3 5 6 7 8 9}

เมื่อได้สร้างเวกเตอร์ของลำดับคำ r_1 และ r_2 แล้วจึงเวกเตอร์ทั้งสองไปวัดค่าความละม้ายระหว่างเวกเตอร์ต่อไป ทั้งนี้ โดยส่วนตัวของผู้วิจัยแล้ว ผู้วิจัยมีความเห็นว่าการใช้ข้อมูลทางวากยสัมพันธ์อันได้แก่ลำดับคำนี้มีความเหมาะสมจะใช้ในงานวิจัยชิ้นนี้ โดยอาจรวมเข้ากับข้อมูลชนิดอื่นเพื่อให้ผลที่ได้มีความถูกต้องมากขึ้น

ในด้านประสิทธิภาพการตรวจหาการลักลอกผลงานนั้น วิธีอิงวากยสัมพันธ์จะสามารถตรวจหาการลักลอกที่เกิดจากการคัดลอกโดยตรง การคัดลอกโดยใกล้เคียง รวมทั้งการคัดลอกโดยดัดแปลงที่มีการเรียงลำดับวลีที่มีภายในเอกสารใหม่ได้ (Alzahrani et al., 2012, p. 144) อย่างไรก็ตาม ข้อจำกัดของวิธีนี้ก็มีไม่น้อย ไม่ว่าจะเป็นข้อจำกัดที่เกิดขึ้นในขั้นตอนการประมวลผลก่อน (pre-processing) อันได้แก่การตัดคำและการกำกับหมวดคำ ซึ่งในกระบวนการของการพัฒนาตัวกำกับหมวดก็ยังต้องวิเคราะห์เพิ่มเติมว่าลักษณะใดที่ควรกำกับลงไปจึงจะเหมาะกับการนำมาใช้ในการหาค่าความละม้ายกัน และที่สำคัญตัวกำกับหมวดคำที่จะพัฒนาขึ้นมานั้นย่อมต้องใช้ความรู้ทางภาษาศาสตร์ที่ตั้งอยู่บนฐานของภาษาใดภาษาหนึ่ง ซึ่งนั่นหมายความว่าระบบที่พัฒนาด้วยวิธีนี้จะตรวจหาการลักลอกได้ภาษาเดียวเท่านั้น

2.5.4 วิธีอิงความหมาย (Semantic-based method)

วิธีอิงความหมายนี้มีแนวคิดพื้นฐานว่าสามารถพิจารณาประโยคเป็นกลุ่มของคำที่เรียงกันตามลำดับ ดังนั้นประโยค 2 ประโยคจึงสามารถมีความละม้ายกันทางความหมายได้แม้จะมีโครงสร้างต่างกัน ยกตัวอย่างเช่นในกรณีของการใช้รูปกรรตุวาจกและกรรมวาจก หรือความแตกต่างของประโยคจากการหลากคำ เป็นต้น ในส่วนของการตรวจหานั้น ข้อมูลทางความหมายที่นำมาใช้ถือเป็นหัวใจสำคัญของวิธีนี้ ข้อมูลทางความหมายในที่นี้ได้แก่ การเป็นคำไวพจน์ การเป็นคำความหมายตรงกันข้าม การเป็นคำจ่ากลุ่ม และการเป็นคำลูกกลุ่ม ซึ่งได้จากฐานข้อมูลคำศัพท์

ในด้านการทำงานนั้น วิธีนี้จะใช้การเปรียบเทียบความละม้ายกันทางความหมายของเอกสาร ยกตัวอย่างงานของลีและคณะ (Y. Li et al., 2006) ที่คำนวณค่าความละม้ายกันทางความหมายระหว่างประโยคโดยใช้ความรู้จากฐานข้อมูลคำศัพท์ที่ถูกสร้างขึ้น ทั้งนี้ ค่าความละม้ายกันของ 2

ประโยคจะได้จากการเปรียบเทียบเวกเตอร์ของคำเฉพาะ (unique term) และคำไวพจน์จากเครือข่ายคำ (WordNet) เทียบกับค่าน้ำหนักจากคลังข้อมูล นอกจากนี้ยังปรากฏในงานของซัตซาโรนิสและคณะ (Tsatsaronis, Varlamis, Giannakoulopoulos, & Kanellopoulos, 2010) ที่ได้วัดค่าความเหมือนกันทางความหมายในเอกสารโดยใช้การคำนวณระยะทางระหว่างคำไวพจน์ที่มีลักษณะเป็นลำดับชั้นจากเครือข่ายคำแล้วเปรียบเทียบกันระหว่างเอกสารด้วย

หากพิจารณาในแง่ของประสิทธิภาพการตรวจหาการลักลอกจะพบว่าวิธีอิงความหมายสามารถตรวจหาได้ตั้งแต่การลักลอกที่เกิดจากการคัดลอกโดยตรง การคัดลอกโดยใกล้เคียง การคัดลอกโดยดัดแปลง รวมไปถึงการลักลอกจากการถอดความที่รวมการใช้คำไวพจน์และการเปลี่ยนตำแหน่งของคำในประโยคไว้ ทั้งนี้เนื่องมาจากวิธีนี้ไม่ได้คำนวณค่าความละม้ายกันของเอกสารจากการเปรียบเทียบอักขระในระดับผิว แต่เป็นการเปรียบเทียบความหมายแก่นที่มีประจำอยู่ในคำศัพท์ซึ่งสามารถสะท้อนความหมายออกมาได้ในระดับประโยคนั่นเอง อย่างไรก็ตาม วิธีนี้ก็มีข้อจำกัดอยู่ไม่น้อย ทั้งในด้านการคิดค้นรูปแทนทางความหมายที่จะต้องใช้ในการคำนวณค่าความละม้าย หรือการต้องประยุกต์กระบวนการตรวจหาให้ทำงานร่วมกับระบบเครือข่ายคำ ซึ่งแน่นอนว่าจะทำให้เกิดข้อจำกัดด้านภาษาตามมา เนื่องจากเครือข่ายคำที่ใช้นั้นย่อมเป็นภาษาเดียว การตรวจจับที่เกิดขึ้นจึงจำเพาะเพียงภาษาใดภาษาหนึ่งไปด้วย ด้วยเหตุดังกล่าวจึงทำให้วิธีนี้ได้รับความสนใจน้อย

กล่าวโดยสรุป จะเห็นว่าการข้อมูลทางภาษาศาสตร์ที่ปรากฏใช้ในวิธีตรวจหาแบบอิงวากยสัมพันธ์และอิงความหมายนั้นเหมาะสมจะใช้ในงานวิจัยชิ้นนี้ เนื่องจากการตรวจหาการลักลอกจากข้อมูลดังกล่าวย่อมจะเข้าถึงข้อมูลทางความหมายได้มากกว่าการตรวจหาจากอักขระระดับผิวหรือเวกเตอร์ของคำ ซึ่งทำให้สามารถตรวจหาการลักลอกได้ถึงระดับที่เป็นการลักลอกด้วยการถอดความ (Alzahrani et al., 2012, p. 145) ดังตารางสรุปวิธีตรวจหาการลักลอกและประสิทธิภาพในการตรวจหาการลักลอกประเภทต่างๆ ดังตารางที่ 2.1

จากตารางที่ 2.1 จะเห็นได้ว่าวิธีตรวจหาโดยอิงความหมายนั้นมีประสิทธิภาพในการตรวจหามากที่สุด รองลงมาคือวิธีตรวจหาโดยอิงความหมายอิงวากยสัมพันธ์และอิงเวกเตอร์ ผู้วิจัยจึงเห็นว่าควรเลือกพัฒนาระบบตรวจหาการลักลอกแบบผสมระหว่างวิธีอิงวากยสัมพันธ์และอิงความหมายเพื่อให้ระบบให้ผลการตรวจหาที่ถูกต้องและแม่นยำมากที่สุด

ตารางที่ 2.1 วิธีตรวจหาการลักลอกผลงานและประสิทธิภาพในการตรวจหาการลักลอกประเภทต่างๆ

วิธีตรวจหา	ภาษา	ประเภทของการลักลอก				
		คัดลอกโดยตรง	คัดลอกโดยใกล้เคียง	คัดลอกโดยดัดแปลง	ถอดความ	สรุปความ
อิงสายอักขระ	ใดๆ	✓	✓			
อิงเวกเตอร์	ใดๆ	✓	✓	✓		
อิงวากยสัมพันธ์	จำเพาะ	✓	✓	✓		
อิงความหมาย	จำเพาะ	✓	✓	✓	✓	

จากวิธีตรวจเทียบภายนอกหาการลักลอกงานวิชาการจะเห็นได้ว่าหลักสำคัญที่ทุกวิธีล้วนใช้พิจารณาร่วมกันว่าข้อความเป็นการลักลอกหรือไม่นั้นก็คือความละม้ายของข้อความ ดังนั้น ในหัวข้อต่อไปผู้วิจัยจะได้กล่าวถึงเทคนิคและวิธีที่ใช้ในการตรวจวัดความละม้ายระหว่างข้อความโดยละเอียด ซึ่งมีทั้งแนวทางที่อาศัยการเรียนรู้ของเครื่องและไม่อาศัยการเรียนรู้ของเครื่อง ทั้งนี้ ในส่วนของแนวทางการตรวจวัดความละม้ายด้วยการเรียนรู้ของเครื่องนั้น เป็นที่น่าสังเกตว่ายังไม่ปรากฏว่ามีการใช้แนวทางดังกล่าวในงานพัฒนาระบบตรวจหาการลักลอกเลย ในหัวข้อรองสุดท้ายของการทบทวนวรรณกรรมนี้ ผู้วิจัยจึงจะกล่าวถึงแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนซึ่งเป็นแบบจำลองหนึ่งของการเรียนรู้ของเครื่องแบบมีการสอนที่ผู้วิจัยเลือกใช้ในงานชิ้นนี้ด้วย

2.6 คลังข้อมูลที่เกี่ยวข้องกับการลักลอก

เมื่อกล่าวถึงการเรียนรู้ของเครื่องแบบมีผู้สอนแล้ว ปัจจัยหนึ่งที่มีความสำคัญยิ่งต่อประสิทธิภาพของเครื่องคือข้อมูลที่ใช้ในการฝึกฝน และเนื่องจากงานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อทดสอบประสิทธิภาพของการตรวจหาการลักลอกงานวิจัยโดยใช้แบบจำลองการเรียนรู้ของเครื่องแบบมีการสอน จึงจำเป็นต้องสร้างคลังข้อมูลสำหรับฝึกฝนและทดสอบแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนซึ่งเป็นแบบจำลองการเรียนรู้ของเครื่องแบบหนึ่งที่ผู้วิจัยเลือกใช้ในงานชิ้นนี้ ในหัวข้อนี้ ผู้วิจัยได้รวบรวมงานว่าด้วยการสร้างคลังข้อมูลที่ใช้ในการฝึกฝนและทดสอบประสิทธิภาพของระบบตรวจหาการลักลอกเอาไว้ ทั้งนี้ เพื่อทำความเข้าใจกรอบคิดและประยุกต์เป็นแนวทางในการสร้างคลังข้อมูลสำหรับใช้ประโยชน์ในวิจัยชิ้นนี้

ก่อนหน้าที่วงวิชาการจะให้ความสนใจศึกษาเกี่ยวกับการลักลอกอย่างแพร่หลายเช่นในปัจจุบัน ในวงการวิทยาศาสตร์คอมพิวเตอร์และสารสนเทศได้มีความสนใจเกี่ยวกับการประมวลข้อความเกิดขึ้น ในช่วงดังกล่าว หัวข้อการศึกษาหนึ่งที่ใกล้เคียงกับการลักลอกคือการศึกษาทำความ

เข้าใจเกี่ยวกับการใช้ข้อความซ้ำ (text reuse) ด้วยความสนใจดังกล่าวนี้ ไกเซาสกัสและคณะ (Gaizauskas et al., 2001) ได้สร้างคลังข้อมูล METER ขึ้นสำหรับใช้ศึกษาและวิเคราะห์การใช้ข้อความซ้ำในทางวารสารศาสตร์

คลังข้อมูล METER ประกอบขึ้นจากชุดของข่าวที่เขียนโดยสมาคมสื่อมวลชนอังกฤษ (Press Association: PA) ข่าวดังกล่าวเป็นข่าวที่มีเนื้อหาเกี่ยวกับเรื่องหรือเหตุการณ์เดียวกันแต่ถูกเขียนโดยหนังสือพิมพ์ต่างฉบับกันจำนวน 9 ฉบับ ในการรวบรวบข่าวเข้าเป็นคลังข้อมูลนั้น แต่ละข่าวจากหนังสือพิมพ์จะได้รับการกำหนดระดับของการใช้ข้อความซ้ำโดยพิจารณาอิงจากจำนวนข้อความของเหตุการณ์เดียวกันที่เขียนขึ้นโดยสมาคมสื่อมวลชน แบ่งเป็น 3 ระดับ ได้แก่ นำมาทั้งหมด (wholly derived), นำมาบางส่วน (partially derived), และไม่นำมาเลย (non-derived) เนื่องด้วยลักษณะและวิธีการแบ่งข้อมูลออกเป็นระดับต่างๆ เช่นที่กล่าวมา จึงทำให้คลังข้อมูล METER ถูกใช้ในการประเมินประสิทธิภาพของระบบตรวจหาการลักลอกอย่างแพร่หลายในช่วงเวลาต่อมา (Cheema et al., 2015, p. 2) อย่างไรก็ตาม การสร้างคลังข้อมูลนี้ต้องอาศัยแรงงานมนุษย์ในการคัดเลือกข้อความและกำกับข้อมูลในเอกสารแต่ละชิ้น จึงส่งผลให้คลังข้อมูลมีขนาดเล็ก กล่าวคือ ประกอบไปด้วยข้อความจำนวน 1,716 ข้อความ หรือคิดเป็นจำนวนคำประมาณ 50,000 คำเท่านั้น

ในระยะเวลาต่อมา การลักลอกได้กลายเป็นหัวข้อที่มีผู้สนใจศึกษาในหลากหลายสาขาวิชา หนึ่งในหัวข้อที่ได้รับการพัฒนาไปมากที่สุดหัวข้อหนึ่งก็คือการตรวจหาการลักลอก ในการดังกล่าวคลังข้อมูลการลักลอกได้ถูกสร้างขึ้นเป็นจำนวนมากเพื่อใช้เป็นเกณฑ์กลางในการเปรียบเทียบประสิทธิภาพของการระบบตรวจหาการลักลอกที่ได้รับการพัฒนาขึ้น และคลังข้อมูลการลักลอกงานวิชาการภาษาอังกฤษกลุ่มหนึ่งที่เป็นที่ยอมรับกันอย่างกว้างขวางในวงการได้แก่คลังข้อมูลในชุด PAN-PC (Potthast, Eiselt, et al., 2011; Potthast et al., 2010; Potthast et al., 2009) ซึ่งถูกใช้เป็นเกณฑ์เปรียบเทียบสมรรถนะของระบบตรวจหาการลักลอกในการแข่งขันซึ่งเป็นส่วนหนึ่งของการประชุมวิชาการการวิเคราะห์การลักลอก การระบุตัวตนของผู้เขียน และการตรวจจับการทำซ้ำโดยใกล้เคียง (Plagiarism analysis, Authorship identification, and Near-duplicate detection: PAN) ซึ่งจัดขึ้นเป็นประจำทุกปี

ในการสร้างคลังข้อมูล PAN-PC-10 พอตแทสต์และคณะ (Potthast et al., 2010, pp. 4-6) ได้เสนอแนวคิดเรื่องลักษณะของแท้ของการลักลอก (plagiarism authenticity) ขึ้นเพื่อประโยชน์ในการรวบรวบข้อมูลเข้าในคลังข้อมูล พอตแทสต์และคณะได้กล่าวว่าลักษณะของแท้ของการลักลอกสามารถแบ่งได้เป็น 3 ระดับ ได้แก่

1) *การลักลอกจริง (real plagiarism)* อันได้แก่กรณีการลักลอกที่เกิดขึ้นในสถานการณ์จริง

2) *การลักลอกเลียนแบบ (simulated plagiarism)* อันได้แก่กรณีการลักลอกที่จำลองขึ้นด้วยน้ำมือมนุษย์
3) *การลักลอกประดิษฐ์ (artificial plagiarism)* อันได้แก่กรณีการลักลอกที่สร้างขึ้นโดยอัลกอริทึมของเครื่อง

ในการรวบรวมกรณีการลักลอกเพื่อสร้างเป็นคลังข้อมูลนั้น พอตแทสต์และคณะ (Potthast et al., 2010, p. 4) ได้ชี้ให้เห็นข้อจำกัดการหาข้อมูลการลักลอกจริงหลายประการ ไม่ว่าจะเป็นข้อจำกัดด้านปริมาณของข้อมูลที่มีน้อย ข้อจำกัดด้านราคาอันเนื่องมาจากค่าธรรมเนียมในการเข้าถึงเอกสารที่เป็นความลับ ข้อจำกัดด้านการยินยอมจากผู้ลักลอก ข้อจำกัดด้านการเผยแพร่คลังข้อมูลซึ่งอาจมานำมาซึ่งการตั้งคำถามด้านกฎหมายและจริยธรรม นอกจากนี้ การปกปิดตนตัวของผู้ลักลอกยังทำได้ยาก เนื่องจากเทคโนโลยีด้านการสืบค้นข้อมูลที่พัฒนาไปมากในปัจจุบัน

จากข้อจำกัดดังกล่าว พอตแทสต์และคณะจึงได้เลือกใช้กรณีการลักลอกจากการลักลอกเลียนแบบและการลักลอกประดิษฐ์เป็นข้อมูลในคลังข้อมูล อย่างไรก็ตาม พอตแทสต์และคณะก็ได้ให้ข้อสังเกตเกี่ยวกับข้อมูลทั้ง 2 ประเภทดังกล่าวไว้ในใน 2 แง่มุม กล่าวคือ ในแง่จิตวิทยา ผู้ที่จำลองการลักลอกขึ้นมีทัศนคติทางจิต (mental attitude) ที่แตกต่างจากผู้ลักลอกในสถานการณ์จริง ซึ่งอาจส่งผลต่อความเป็นธรรมชาติของข้อมูลการลักลอกได้ ส่วนในด้านภาษานั้นก็ยังไม่อาจพิสูจน์ได้แน่ชัดว่าลักษณะทางภาษาที่ปรากฏในการลักลอกจริงจะแตกต่างจากการลักลอกจำลองหรือไม่อย่างไร

ส่วนการสร้างข้อมูลนั้น ในส่วนการลักลอกเลียนแบบ พอตแทสต์และคณะ (Potthast et al., 2010, pp. 4-5) ได้ว่าจ้างผู้จำลองการลักลอกผ่านเว็บไซต์ Amazon's Mechanical Turk (AMT) ซึ่งเป็นเว็บไซต์สำหรับรวบรวมกลุ่มคนเพื่อสร้างสรรค์ผลงาน (crowdsource) โดยในเบื้องต้น ผู้จำลองการลักลอกต้องกรอกข้อมูลส่วนบุคคลของต้นอันประกอบไปด้วยอายุ การศึกษา เพศ และภาษาแม่ จากนั้นจึงให้ผู้จำลองการลักลอกเขียนข้อความที่กำหนดให้ขึ้นใหม่โดยให้จินตนาการว่าตนเองกำลังลอกงานของเพื่อนในชั้นเรียน หรือเป็นผู้ลักลอกที่พยายามจะลักลอกงานของผู้อื่นโดยปราศจากการอ้างอิงแหล่งที่มาอย่างเหมาะสม

ส่วนการลักลอกประดิษฐ์ พอตแทสต์และคณะ (Potthast et al., 2010, pp. 5-6) ได้เสนอกลวิธีประดิษฐ์การลักลอกด้วยเครื่อง 3 ประเภท ได้แก่

1) *การดำเนินการสุ่มข้อความ (random text operations)* ได้แก่ การสับเปลี่ยน ลบ แทรก แทนที่ คำหรือวลีสั้นๆ โดยสุ่ม

2) *การหลากคำ (semantic word variation)* ได้แก่ การสุ่มแทนที่คำด้วยคำไวพจน์ คำตรงกันข้าม คำจ่ากลุ่ม คำลูกกลุ่ม
3) *การสับเปลี่ยนคำโดยคงหมวดคำ (POS-preserving word shuffling)* ได้แก่ การสุ่มสับเปลี่ยนคำโดยรักษาลำดับของหมวดคำไว้ตามเดิม

ด้วยวิธีการสร้างข้อมูลการลักลอกดังกล่าวข้างต้น คลังข้อมูล PAN-PC-10 จึงประกอบไปด้วยกรณีการลักลอกที่หลากหลายซึ่งประกอบด้วยข้อมูลที่ได้มาจากทั้งมนุษย์และเครื่อง และด้วยเหตุนี้ PAN-PC-10 จึงเป็นคลังข้อมูลการลักลอกที่มีขนาดใหญ่กว่าคลังข้อมูลการลักลอกใดๆ ที่ได้เคยสร้างขึ้น กล่าวคือ ประกอบด้วยข้อความจำนวน 27,073 ข้อความซึ่งรวมกรณีลักลอกไว้ถึง 68,558 กรณี

คลังข้อมูลในชุด PAN ที่ได้รับการสร้างขึ้นในช่วงเวลาต่อมาต่างก็ได้รับการปรับปรุงสัดส่วนของข้อมูลให้สอดรับกับภารกิจที่ใช้ในการแข่งขันเปรียบประสิทธิภาพของระบบตรวจหาการลักลอกที่แตกต่างกันไปในแต่ละปี อย่างไรก็ตาม แนวคิดพื้นฐานที่ใช้ในการสร้างคลังข้อมูลยังอิงจากแนวคิดในการสร้างคลังข้อมูล PAN-PC-10 เป็นสำคัญ ยกตัวอย่างเช่น คลังข้อมูลการลักลอกที่สร้างโดยโมฮ์ตัจญ์และคณะ (Mohtaj et al., 2015) ซึ่งได้จำลองการลักลอกอัตโนมัติโดยอาศัยการค้นคืนข้อความที่มีค่าความละม้ายกันตามที่กำหนดไว้ จากนั้นจึงนำข้อความดังกล่าวมาจับคู่กันเป็นกรณีลักลอก

ในทางตรงกันข้ามกับคลังข้อมูลในชุด PAN ซึ่งข้อมูลส่วนใหญ่สร้างขึ้นด้วยเครื่อง ก็ยังปรากฏคลังข้อมูลการลักลอกที่ข้อมูลทั้งหมดในคลังข้อมูลสร้างขึ้นด้วยน้ำมือของมนุษย์ คลังข้อมูลการลักลอกคำตอบขนาดสั้น (Clough & Stevenson, 2009, 2011) เป็นคลังข้อมูลที่สร้างขึ้นเพื่อใช้เป็นเกณฑ์เปรียบเทียบประสิทธิภาพของระบบตรวจหาการลักลอกงานวิชาการภาษาอังกฤษโดยเฉพาะ

ในการสร้างคลังข้อมูลการลักลอกคำตอบขนาดสั้นขึ้นนั้น คลอฟและสตีเวนสัน (Clough & Stevenson, 2011, p. 6) ได้ตั้งข้อสังเกตเกี่ยวกับคลังข้อมูลการลักลอกที่มีใช้อยู่ในขณะนั้นซึ่งรวมถึงคลังข้อมูล PAN-PC-09 (Potthast et al., 2009, p. 3) ด้วยว่าขาดรายละเอียดที่ชัดเจนเกี่ยวกับวิธีการสร้างและให้ความสำคัญเฉพาะการลักลอกลอกประเภทใดประเภทหนึ่ง ด้วยเหตุดังกล่าวนี้ คลอฟและสตีเวนสันจึงตัดสินใจสร้างคลังข้อมูลการลักลอกขึ้นโดยอาศัยมนุษย์เป็นผู้จำลองข้อมูลการลักลอกทั้งหมด โดยได้ว่าจ้างนักศึกษาให้ตอบคำถามในสาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ความยาว 200-300 คำ โดยคลอฟและสตีเวนสันจะเตรียมตัวบทสำหรับใช้เป็นคำตอบไว้ให้นักศึกษาผู้จำลองการลักลอก ทั้งนี้ ผู้จำลองการลักลอกจะได้รับมอบหมายให้ตอบคำถามโดยใช้ความรู้จากตัวบทในระดับที่แตกต่างเป็น 4 ระดับตั้งแต่ใช้ความรู้จากตัวบททั้งหมดจนถึงไม่ใช้เลย (Clough & Stevenson, 2011, p. 11) ได้แก่

1) *การคัดลอกโดยใกล้เคียง (near copy)* ได้แก่ การใช้ข้อมูลจากตัวบทเตรียมไว้ให้ทั้งหมดในการตอบคำถาม โดยไม่แก้ไขภาษาจากตัวบทที่เตรียมไว้ให้เลย
2) *การแก้ไขอย่างเบา (light revision)* ได้แก่ การใช้ข้อมูลจากตัวบทเตรียมไว้ในการตอบคำถาม โดยปรับเปลี่ยนข้อความจากตัวบทด้วยวิธีพื้นฐาน เช่น แทนที่คำหรือวลีด้วยรูปไวพจน์ หรือปรับเปลี่ยนโครงสร้างทางไวยากรณ์ด้วยการถอดความ
3) *การแก้ไขอย่างหนัก (heavy revision)* ได้แก่ การใช้ข้อมูลจากตัวบทเตรียมไว้ในการตอบคำถาม แต่ให้แก้ไขภาษาที่ใช้ให้แตกต่างจากตัวบทโดยคงความหมายเดิมไว้ด้วยการใช้คำและโครงสร้างที่แตกต่างจากเดิม รวมถึงการแยกหรือรวมประโยคใดๆ ในตัวบทที่เตรียมไว้ให้
4) *ไม่ลักลอก (non-plagiarism)* ได้แก่ การตอบคำถามโดยใช้ความรู้ที่นักศึกษามีอยู่แล้ว หรือใช้ข้อมูลที่นักศึกษาได้เตรียมมาล่วงหน้า

ด้วยวิธีการดังกล่าวข้างต้น คลอฟและสตีเวนสันสามารถสร้างคลังข้อมูลการลักลอกคำตอบขนาดสั้นในภาษาอังกฤษได้ โดยประกอบด้วยข้อความที่เป็นคำตอบขนาดสั้นจำนวนทั้งหมด 100 ข้อความ แบ่งเป็นคำตอบที่ได้จากผู้จำลองการลักลอก 95 ข้อความ และข้อความที่เป็นตัวบทต้นฉบับให้ผู้จำลองการลักลอกอีก 5 ข้อความ อย่างไรก็ตาม จะสังเกตได้ว่าคลังข้อมูลของคลอฟและสตีเวนสันนี้มีขนาดค่อนข้างเล็ก อันเป็นผลเนื่องมาจากกระบวนการสร้างข้อมูลการลักลอกที่ใช้มนุษย์

นอกจากคลังข้อมูลการลักลอกภาษาอังกฤษที่ได้กล่าวไปทั้งหมดข้างต้นแล้ว ยังมีคลังข้อมูลการลักลอกในภาษาอื่นๆ อีกที่สร้างขึ้นเพื่อตอบวัตถุประสงค์ในการวิเคราะห์และตรวจหาการลักลอก ยกตัวอย่างเช่น คลังข้อมูลการลักลอกแบบถอดความภาษาอูรดู (Urdu Paraphrase Plagiarism Corpus: UPPC) (Sharjeel, Rayson, & Nawab, 2016) อย่างไรก็ตาม จากการทบทวนวรรณกรรมที่เกี่ยวข้อง ก็ยังไม่พบว่ามีการพัฒนาคลังข้อมูลการลักลอกภาษาไทยขึ้นสำหรับใช้ประโยชน์ในการประเมินประสิทธิภาพของระบบตรวจหาการลักลอกงานวิชาการ

จากการทบทวนวรรณกรรมที่เกี่ยวข้องในหัวข้อนี้ จะเห็นได้ว่าในการสร้างคลังข้อมูลสำหรับใช้ประโยชน์ในการตรวจหาการลักลอกงานวิชาการนั้น การแบ่งประเภทของข้อมูลในคลังข้อมูลเป็นเรื่องหนึ่งที่ต้องให้ความสำคัญยิ่ง อย่างไรก็ตาม จะเห็นได้ว่าคลังข้อมูลการลักลอกแต่ละคลังที่กล่าวไปข้างต้นต่างก็มีวิธีการจัดประเภทและสร้างข้อมูลการลักลอกของตนเอง เป็นต้นว่าพอตแทสต์และคณะ (Potthast et al., 2010) แบ่งข้อมูลตามวิธีการสร้างด้วยเครื่องและสร้างโดยมนุษย์แต่ก็ไม่ได้มีการกำหนดปริมาณการลักลอกในแต่ละกรณีการลักลอกให้มากน้อยแตกต่างกันอย่างชัดเจน ในขณะที่คลอฟและสตีเวนสัน (Clough & Stevenson, 2009, 2011) แม้จะได้จัดประเภทข้อมูลการลักลอกภายในคลังข้อมูลตามระดับความเข้มข้นของการลักลอกแล้ว แต่กลวิธีที่ใช้ในการลักลอกก็ยังไม่ชัดเจน เช่น การปรับเปลี่ยนโครงสร้างทางไวยากรณ์ในข้อมูลประเภทการแก้ไขอย่างเบากับการแก้ไขทาง

โครงสร้างในข้อมูลประเภทการแก้ไขอย่างหนักก็ไม่อาจระบุได้โดยชัดเจนว่าแตกต่างกันอย่างไร ทั้งนี้ หากเปรียบเทียบกับประเภทของการลักลอกที่เสนอโดยอัลซะฮ์รานิและคณะ (Alzahrani et al., 2012, pp. 133-137) ที่ได้กล่าวไปในหัวข้อที่ 2.3.3 แล้ว ก็ดูเหมือนจะมีความชัดเจนกว่าในทั้งแง่ทฤษฎีทางภาษาศาสตร์และในแง่การปฏิบัติการสร้างคลังข้อมูล ด้วยการจัดประเภทของอัลซะฮ์รานิและคณะนั้นยึดโยงอยู่กับลำดับชั้นทางภาษาที่เรียงไปตั้งแต่การดัดแปลงแก้ไขหน่วยทางภาษาดังแต่หน่วยขนาดเล็กไปกระทั่งถึงหน่วยขนาดใหญ่ ลักษณะดังกล่าวยังสะท้อนให้เห็นถึงความยากง่ายที่แตกต่างกันระหว่างประเภทของการลักลอกทั้งในแง่การสร้างข้อมูลและการตรวจหาการลักลอก ผู้วิจัยจึงสามารถยืนยันได้อย่างมั่นใจว่าจะเลือกใช้กรอบคิดของของอัลซะฮ์รานิและคณะในการทำความใจและสร้างข้อมูลการลักลอกในงานวิจัยชิ้นนี้

ด้วยเหตุผลด้านความขาดแคลนคลังข้อมูลการลักลอกงานวิชาการภาษาไทยสำหรับใช้ในการประเมินประสิทธิภาพของระบบดังได้กล่าวไปข้างต้น งานวิจัยชิ้นนี้จึงจำเป็นต้องสร้างคลังข้อมูลการลักลอกภาษาไทยที่มีขนาดใหญ่พอที่ฝึกฝนและทดสอบประสิทธิภาพของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน อย่างไรก็ดี งานสร้างคลังข้อมูลการลักลอกที่ได้กล่าวถึงในหัวข้อนี้ก็ได้ทำให้แนวคิดและกระบวนการที่เลือกใช้ในการสร้างข้อมูลการลักลอกทั้งแบบที่ใช้เครื่องและทำโดยมนุษย์ ผู้วิจัยสามารถยึดเอาแนวทางดังกล่าวมาปรับใช้ในการสร้างคลังข้อมูลให้มีขนาดที่เหมาะสมแก่การใช้ประโยชน์ในการวิจัยได้ ทั้งนี้ ผู้วิจัยจะได้กล่าวถึงรายละเอียดของการออกแบบและสร้างคลังข้อมูลการลักลอกงานวิชาการภาษาไทยในบทต่อไป

2.7 ความละม้ายของข้อความ

ความละม้ายของข้อความ (text similarity) เป็นแนวคิดหนึ่งที่ใช้ในการประมวลผลข้อความในระดับความหมาย งานวิจัยชิ้นนี้ได้เลือกใช้แนวคิดดังกล่าวเป็นส่วนหนึ่งของระบบตรวจหาการลักลอก ดังนั้นในหัวข้อนี้ ผู้วิจัยจึงได้รวบรวมงานที่เกี่ยวกับความละม้ายของข้อความไว้ โดยในหัวข้อย่อยแรก ผู้วิจัยจะกล่าวถึงมโนทัศน์พื้นฐานเกี่ยวกับความละม้ายของข้อความ และจากการทบทวนงานแขนงต่างๆ ทำให้พบว่า ในการตรวจวัดความละม้ายของข้อความด้วยเครื่องนั้นสามารถแบ่งเทคนิควิธีที่ใช้ออกได้เป็น 2 แนวทางใหญ่ๆ ได้แก่ การจำแนกประเภทข้อความที่มีความละม้ายโดยใช้การเรียนรู้ของเครื่อง และการวัดค่าความละม้ายระหว่างข้อความ ในหัวข้อย่อยอีก 2 หัวข้อ ผู้วิจัยได้รวบรวมและสรุปรายละเอียดของเทคนิคและวิธีการตรวจวัดทั้งสองแนวทางไว้อย่างชัดเจน พร้อมทั้งยังได้อภิปรายถึงการนำความรู้ทางภาษาศาสตร์มาประยุกต์ใช้ในแต่ละแนวทางอีกด้วย รายละเอียดดังกล่าวผู้วิจัยจะนำเสนอเป็นลำดับดังต่อไปนี้

2.7.1 มโนทัศน์พื้นฐานเกี่ยวกับความละม้ายของข้อความ

ความละม้ายของข้อความเป็นแนวคิดหนึ่งในงานประมวลผลภาษาธรรมชาติที่ถูกใช้อย่างแพร่หลายในแขนงงานต่างๆ ไม่ว่าจะเป็นการค้นคืนสารสนเทศ (information retrieval) การจำแนกประเภทข้อความ (text classification) การรวมกลุ่มเอกสาร (document clustering) การตรวจหาหัวข้อ (topic detection) การติดตามหัวข้อ (topic tracking) การสร้างคำถาม (questions generation) การตอบคำถาม (question answering) การให้คะแนนความเรียง (essay scoring) การให้คะแนนคำตอบขนาดสั้น (short answer scoring) การแปลภาษาด้วยเครื่อง (machine translation) การสรุปความจากข้อความ (text summarization) การระบุการถอดความ (paraphrase identification) การรู้จำการโยงไปสู่ข้อสรุปทางข้อความ (recognizing textual entailment: RTE) รวมถึงการตรวจหาการลักลอกงานวิชาการ (Gomaa & Fahmy, 2013, p. 13)

จุดประสงค์ของการวัดความละม้ายนั้นเป็นไปเพื่อเปรียบเทียบและระบุว่าข้อความ 2 ข้อความใดๆ ในฐานะข้อมูลรับเข้ามีการสมมูลทางความหมายกันหรือไม่ ทั้งนี้ การวัดความละม้ายของข้อความสามารถอิงกับหน่วยทางภาษาได้หลายระดับ

ปัจจุบัน แนวโน้มของการวัดความละม้ายมักนิยมวัดในระดับวลีหรือประโยคระหว่างข้อความที่มีขนาดใกล้เคียงกัน (Agirre, Cer, Diab, & Gonzalez-Agirre, 2012; Agirre, Cer, Diab, Gonzalez-Agirre, & Guo, 2013; Agirrea et al., 2014) อย่างไรก็ตาม ระดับทางภาษาที่ถูกใช้อย่างแพร่หลายตลอดมาสำหรับการวัดความละม้ายคือระดับคำ (Finkelstein et al., 2002; Rubenstein & Goodenough, 1965) นอกจากระดับทางภาษาที่กล่าวมายังปรากฏการวัดความละม้ายในระดับสายอักขระ (Barrón-Cedeño, Rosso, Agirre, & Labaka, 2010) บ้างประปราย

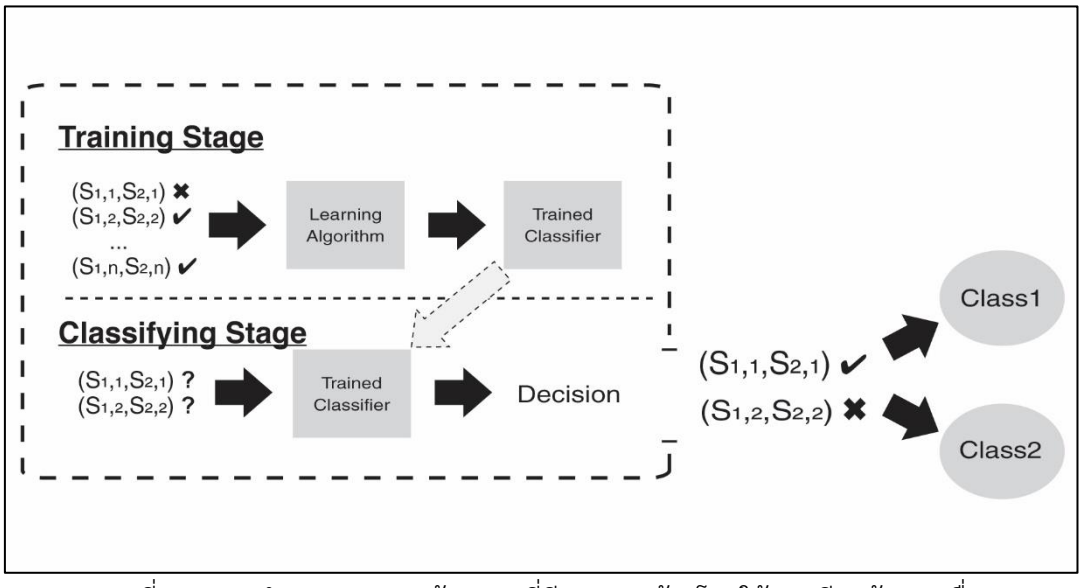
ในหัวข้อย่อยถัดไป ผู้วิจัยจะได้กล่าวถึงวิธีการวัดความละม้ายด้วยใน 2 แนวทาง โดยอิงจากรูปทางภาษาในระดับต่างๆ ดังได้กล่าวมาเป็นตัวอย่างข้างต้นนำเสนอเป็นลำดับดังต่อไปนี้

2.7.2 การจำแนกข้อความที่มีความละม้ายกันโดยใช้การเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่อง (machine learning) ถือว่าเป็นเทคนิคหนึ่งที่ได้ถูกนำมาใช้ในงานต่างๆ ในสาขาการประมวลผลภาษาธรรมชาติอย่างหลากหลาย ไม่ว่าจะเป็นการกำกับหมวดคำ (Brill, 1995; Toutanova, Klein, Manning, & Singer, 2003; Tseng, Jurafsky, & Manning, 2005) การแก้ปัญหาความกำกวมของคำ (Bruce & Wiebe, 1994; Leacock, Towell, & Voorhees, 1993; Palmer, Fellbaum, Cotton, Delfs, & Dang, 2001) หรือการรู้จำชื่อเฉพาะ (Lee et al., 2006; Turian, Ratinov, & Bengio, 2010; นัชชา ถิระสาโรช, 2553) ในส่วนของการตรวจวัดความละม้ายของข้อความนั้น การใช้เทคนิคการเรียนรู้ของเครื่องจะมีบทบาทมากในงานการระบุการถอดความและ

การรู้จำการโยงไปสู่ข้อสรุปทางข้อความ กล่าวคือ การเรียนรู้ของเครื่องจะมีหน้าที่จำแนกประเภทข้อความที่มีความละม้ายและไม่มีความละม้ายออกจากกันนั่นเอง

การจำแนกประเภทข้อความที่มีความละม้ายและข้อความไม่มีความละม้ายออกจากกันดังกล่าวข้างต้นจะเกิดขึ้นภายใต้แนวทางการเรียนรู้ของเครื่องแบบมีการสอน (supervised machine learning) ซึ่งเป็นการให้แบบจำลองการเรียนรู้แบบมีการสอน (supervised learning model) เรียนรู้จากกรณีตัวอย่างที่จัดเตรียมให้ จากนั้นแบบจำลองจึงจะสร้างสมมติฐานทั่วไปเพื่อทำนายหรือตัดสินใจเกี่ยวกับกรณีตัวอย่างในอนาคต (Kotsiantis, 2007, p. 249) กรณีตัวอย่างที่จำเป็นต้องจัดเตรียมไว้ให้เครื่องเรียนรู้ในที่นี้ ได้แก่ ลักษณ์ (feature) ต่างๆ ที่ใช้ในการตัดสินใจและค่าการตัดสินใจว่าคู่ของข้อความมีความละม้ายกันหรือไม่



ภาพที่ 2.5 การจำแนกประเภทข้อความที่มีความละม้ายโดยใช้การเรียนรู้ของเครื่อง

จากภาพที่ 2.5 จะเห็นได้ว่าการทำงานของการเรียนรู้ของเครื่องแบบมีการสอนนั้นแบ่งออกเป็น 2 ขั้น ได้แก่ ขั้นการฝึกฝน (training stage) และขั้นการจำแนกประเภท (classifying stage) ในขั้นการฝึกฝนนั้น แบบจำลองการเรียนรู้จะรับเข้าส่วนของข้อความ (segment of text) ซึ่งปกติแล้วจะเป็นประโยคเข้าไปเป็นคู่ คู่ของส่วนของความดังกล่าวนี้จะถูกกำกับหรือแทนรูปด้วยลักษณ์ชนิดต่างๆ ที่ใช้ในการตัดสินใจ และยังถูกกำกับค่าการตัดสินใจในการจำแนกประเภทไว้โดยผู้สอน (มนุษย์) ด้วยว่าเป็นคู่ของส่วนของข้อความที่มีความละม้ายกันหรือเป็นคู่ของส่วนของข้อความที่ไม่มีความละม้ายกัน จากนั้นคู่ของส่วนของความเหล่านี้จะผ่านเข้าสู่อัลกอริทึมการเรียนรู้ (learning algorithm) เพื่อเก็บและคำนวณข้อมูลเชิงสถิติจากลักษณ์และค่าการตัดสินใจที่ถูกกำกับมาเพื่อสร้างเป็นสมมติฐานทั่วไปในการตัดสินใจ เมื่อเสร็จสิ้นจากขั้นตอนแล้วก็จะได้ตัวจำแนกประเภทที่ถูกฝึกฝน (trained classifier) ซึ่งรวบรวมข้อมูลเชิงสถิติไว้เพียงพอต่อการตัดสินใจแล้ว ตัวจำแนกประเภทที่

ถูกฝึกฝนแล้วนี่เองจะถูกนำมาใช้ในขั้นตอนการจำแนกประเภทข้อความที่มีความละม้ายหรือไม่มีความละม้ายออกจากกัน ในขั้นนี้แบบจำลองจะรับเอาคู่ของส่วนของข้อความที่ถูกกำกับหรือแทนรูปด้วยลักษณ์เรียบร้อยแล้วเข้ามา แต่ค่าการตัดสินใจในการจำแนกประเภทนั้นจะไม่ถูกกำกับไปด้วย เพราะเป็นหน้าที่ที่แบบจำลองจะต้องตัดสินใจจากข้อมูลที่มีอยู่แล้วในตัวจำแนกประเภท เมื่อส่วนของข้อความถูกส่งเข้าไปยังตัวจำแนกประเภทที่ถูกฝึกฝนแล้ว หากแบบจำลองตัดสินว่าคู่ของส่วนของข้อความมีความละม้ายกัน คู่ดังกล่าวก็จะถูกจัดไว้เป็นประเภท (class) หนึ่ง ในทางตรงกันข้าม คู่ของส่วนของข้อความที่ระบบตัดสินว่าไม่มีความละม้ายกันก็จะถือจัดไว้เป็นอีกประเภทหนึ่ง ด้วยวิธีนี้จึงทำให้สามารถตรวจวัดความละม้ายของข้อความได้ ทั้งนี้ ประสิทธิภาพของการใช้เทคนิควิธีนี้ขึ้นอยู่กับลักษณ์ชนิดต่างๆ ให้เครื่องได้เรียนรู้และฝึกฝนการตัดสินใจ จึงอาจกล่าวได้ว่านอกจากการเลือกใช้แบบจำลองการเรียนรู้แล้ว ลักษณ์ก็ถือเป็นหัวใจสำคัญของการเรียนรู้ของเครื่องแบบมีการสอนด้วยเช่นกัน

ดังได้กล่าวไปแล้วว่าลักษณ์ต่าง ๆ ที่ผู้สอนกำหนดให้นั้นมีผลโดยต่อประสิทธิภาพการตัดสินใจของแบบจำลองการเรียนรู้ ในที่นี้ ผู้วิจัยจึงได้รวบรวมและสรุปประเภทของลักษณ์ที่ใช้ในการตัดสินใจจำแนกประเภทข้อความที่มีความละม้ายที่ปรากฏในงานต่างๆ เอาไว้เป็นกลุ่มและได้แสดงไว้เป็นลำดับด้านล่าง ทั้งนี้ ลักษณ์ประเภทต่างๆ นั้นก็อาศัยความรู้ทางภาษาศาสตร์ในการสร้างหรือสกัดมากน้อยแตกต่างกันไป ผู้วิจัยจะได้อภิปรายถึงประเด็นนี้ด้วย

1) ลักษณ์ทางข้อความ (textual feature)

ลักษณ์ชนิดนี้เป็นลักษณ์ที่ไม่ได้อาศัยความรู้ทางภาษาศาสตร์ในการสร้าง เพราะเป็นเพียงการแทนลักษณะของข้อความออกมาในเชิงปริมาณของสายอักขระหรือสายคำเท่านั้น ได้แก่ ความยาวของข้อความ ไม่ว่าจะเป็นความยาวของประโยค ความยาวของย่อหน้า ความยาวของบท หรือความยาวของข้อเขียน โดยปกติแล้วลักษณ์ชนิดนี้จะปรากฏใช้ในการตรวจจับการลักลอกงานวิชาการภายใน (intrinsic plagiarism detection) เป็นส่วนใหญ่ดังที่ปรากฏในงานของเมาเรอร์และคณะ (Maurer et al., 2006) และเมเยอร์ ซู ไอส์เซนและคณะ (Meyer zu Eissen et al., 2007)

ทั้งนี้ ผู้วิจัยเห็นว่าลักษณ์ประเภทนี้สามารถนำมาประยุกต์ใช้ในงานวิจัยชิ้นนี้ได้ โดยจะใช้เป็นลักษณ์การตรวจหาการลักลอกซึ่งอาจช่วยเสริมประสิทธิภาพการทำงานเมื่อฝึกฝนร่วมกับลักษณ์ชนิดอื่นๆ ได้

2) ลักษณ์อักขระ (character feature)

ลักษณ์อักขระเป็นลักษณ์อีกประเภทหนึ่งที่นิยมใช้ในการตรวจจับการลักลอกงานวิชาการภายใน ลักษณ์เหล่านี้ ได้แก่ ความถี่ของอักขระ จำนวนอักขระแต่ละชนิด เช่น ตัวอักษร ตัวเลข

เครื่องหมายวรรคตอน ความถี่ของอักขระพิเศษ รวมถึงความถี่ของอักขระเอ็นแกรม ทั้งนี้ เช่นเดียวกับลักษณ์ทางข้อความ ลักษณ์ชนิดนี้ก็เป็นลักษณ์ที่ไม่ได้อาศัยความรู้ทางภาษาศาสตร์ในการสร้างเช่นกัน เนื่องจากเป็นเพียงการแทนรูปข้อความตามอักขระเท่านั้น

3) ลักษณ์ทางศัพท์ (lexical feature)

ลักษณ์ทางศัพท์นี้เป็นลักษณ์ที่อาศัยความรู้ทางภาษาศาสตร์เล็กน้อยในระดับคำ กล่าวคือ ในการจะสร้างลักษณ์ประเภทนี้ได้ต้องใช้ความรู้ในการตัดแบ่งคำซึ่งอาจต้องพึ่งพาพจนานุกรม หรืออาจต้องใช้ความรู้ในด้านวิทยาหน่วยคำในกรณีที่ภาษาต้นทางเป็นภาษาที่มีระบบวิภักติปัจจัยเพื่อจะได้ทำให้คำอยู่ในรูปต้นเค้า (stemming) อย่างไรก็ตาม ความรู้ทางภาษาศาสตร์ที่ใช้นั้นจะไม่เกินเลยไปถึงระดับวากยสัมพันธ์ ทั้งนี้เพราะลักษณ์ทางศัพท์ในที่นี้จะอิงแนวคิดเรื่องถุงใส่คำ (bag-of-words) ซึ่งพิจารณาคำในข้อความเป็นชุดของคำที่ไม่มีลำดับและไม่สนใจตำแหน่งที่แท้จริงของคำในประโยค (Jurafsky & Martin, 2009, p. 641)

ทั้งนี้ เมื่อข้อความได้รับการตัดแบ่งคำแล้ว ลักษณ์ทางศัพท์สามารถหาได้จากค่าคำทับซ้อนที่ปรากฏร่วมกันระหว่างส่วนของข้อความ ดังที่ปรากฏในงานของโคซาเรวาและมอนโตโย (Kozareva & Montoyo, 2006) ที่ได้ระบุการถอดความด้วยลักษณ์ชนิดนี้ โดยอาศัยแนวคิดที่ว่ายิ่งจำนวนคำที่มีร่วมกันในทั้งสองส่วนของข้อความมีสูงเท่าใด ยิ่งชี้ให้เห็นว่าส่วนของข้อความทั้งคู่นั้นมีความละม้ายกัน ทั้งนี้ สมมติให้มีส่วนของข้อความ S_1 และ S_2 ค่าคำทับซ้อนของส่วนของข้อความทั้งคู่อาจหาอย่างง่ายได้จากการเปรียบเทียบสัดส่วนของจำนวนคำรวมของทั้งสองส่วนของข้อความ S_1/S_2 หรือมิเช่นนั้น การหาค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ด (Jaccard similarity coefficient) ก็เป็นอีกแนวทางที่ใช้เป็นลักษณ์ในกลุ่มนี้ โดยหาได้จากการนำจำนวนคำที่ S_1 และ S_2 มีร่วมกันหารด้วยจำนวนคำทั้งหมดของ S_1 และ S_2 ก็จะได้ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดดังสมการ

$$sim_{Jaccard}(S_1,S_2)=\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

อย่างไรก็ตาม ในกรณีที่ประโยคมีคำที่เหมือนกัน เช่นในกรณีของประโยค “Mary calls the police” และ “The police calls Mary” ซึ่งแม้ทั้งสองประโยคจะมีคำที่เหมือนกันแต่ก็มิได้แสดงถึงความหมายอย่างเดียวกัน ในกรณีนี้ โคซาเรวาและมอนโตโย (Kozareva & Montoyo, 2006, p. 526) จึงได้เสนอให้หลีกเลี่ยงลักษณะที่เป็นข้อจำกัดของลำดับคำโดยการใช้ค่า Skip-grams และการหาลำดับเหมือนที่ยาวที่สุดในชุดอักขระ (longest common subsequence*: lcs*)

Skip-grams เป็นการพิจารณาปริบทโดยกระโดดข้ามหน่วย[3]ที่อยู่ระหว่างกลางไป ยกตัวอย่าง หากพิจารณาเป็น skip bigram ของคำก็จะเป็นการพิจารณาคู่ของคำในลำดับประโยคซึ่งย่อมให้มีคำอื่นแทรกอยู่ระหว่างคู่คำดังกล่าว (Jurafsky & Martin, 2009, p. 806) ในงานของโคซาเรวาและมอนโตโยเลือกใช้ค่า skip-grams เป็นลักษณ์ในการจำแนกการถอดความด้วยเครื่อง โดยจะพิจารณาลำดับที่ไม่ต่อเนื่องกันของคำซึ่งย่อมให้เกิดช่องว่างขึ้นระหว่างลำดับเปรียบเทียบกับผลรวมทั้งหมดของคำที่สามารถปรากฏในประโยคได้ ซึ่งสามารถหาได้จากการวัดค่า 2 อย่าง ได้แก่ $\text{skip-gram}\,S_1 = \frac{\text{skip-gram}(S_1,S_2)}{\text{C(n, skip-gram}(S_1,S_2))}$ และ $\text{skip-gram}\,S_2 = \frac{\text{skip-gram}(S_1,S_2)}{\text{C(m, skip-gram}(S_1,S_2))}$ ทั้งนี้ $\text{skip-gram}(S_1,S_2)$ คือจำนวนของ skip grams *(*คู่ของคำตามลำดับในประโยคซึ่งยอมให้เกิดช่องว่างขึ้นใดๆได้*)* ซึ่งพบใน S_1 และ S_2 และ $\text{C(n, skip-gram}(S_1,S_2))$ คือ combinatorial function ซึ่ง *n* เป็นจำนวนคำในส่วนของข้อความ S_1 ในขณะเดียวกัน *m* ก็แทนจำนวนคำใน S_2 ความยาวสูงสุดของการคำนวณ Skip-gram ถูกจำกัดถึงสี่เท่านั้น เพราะลำดับที่สูงกว่านี้ไม่ปรากฏบ่อยครั้ง

การหาลำดับเหมือนที่ยาวที่สุดในชุดอักขระ (longest common subsequence: *lcs*) ก็เป็นอีกหนึ่งวิธีที่สามารถใช้เป็นลักษณ์ทางศัพท์ได้ ทั้งนี้ *lcs* จะหาค่าลำดับเหมือนของคำขนาดยาวใน 2 ประโยคออกมา ค่าดังกล่าวจะแสดงให้เห็นได้ว่ามีความละม้ายกันระหว่างประโยคมากน้อยเท่าใด

นอกจากนี้แล้ว จางและแพตทริก (Y. Zhang & Patrick, 2005) ยังได้ใช้ลักษณ์อีกตัวหนึ่งในงาน ได้แก่ การหาระยะการแก้ไข (edit distance*)* ระหว่างส่วนของข้อความ วิธีการนี้จะให้รายละเอียดถึงจำนวนครั้งของการแก้ไข ไม่ว่าจะเป็นการเพิ่ม การลบ การย้ายที่ ของคำ ซึ่งเกิดจากการแก้ไขข้อความต้นฉบับให้กลายเป็นข้อความที่ถูกลักลอกหรือถอดความได้

ลักษณ์ทางศัพท์อีกชนิดถูกเสนอในงานของชงและคณะ (Chong, Specia, & Mitkov, 2010) ได้แก่การหาค่าความละม้ายของไตรแกรม วิธีการนี้ทำได้โดยหาไตรแกรมของคำในประโยค ยกตัวอย่างเช่นประโยค ["This is an example."] จะสามารถทำเป็นไตรแกรมได้เป็น ["This", "is", "an"] ["is", "an", "example"] ["an", "example", "."] เมื่อได้ชุดของไตรแกรมออกมาแล้วก็สามารถนำมาหาค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดได้ด้วยสมการ $sim_{Jaccard}(A,B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|}$ โดยให้ *S*(A) และ *S*(B) ชุดของไตรแกรมในส่วนของข้อความที่ต้องการวัดความละม้าย

นอกจากการพิจารณาค่าความละม้ายของทั้งข้อความดังได้กล่าวมาทั้งหมดแล้ว โคซาเรวาและมอนโตโย (Kozareva & Montoyo, 2006, p. 527) ยังได้เสนอให้ใช้ลักษณ์อีกประเภทได้แก่ ชื่อ

---

[3] ในกรณีนี้คือคำ

เฉพาะ ด้วย ลักษณ์ชนิดนี้สร้างได้จากการจับคู่ชื่อเฉพาะที่ตรงกันในส่วนของข้อความ หากพบว่ามีชื่อเฉพาะที่ตรงกันจะกำหนดค่าให้เป็น 1 ในขณะเดียวกันก็จะกำหนดค่าให้เป็น 0 ในกรณีที่คู่ของส่วนของข้อความไม่ปรากฏชื่อเฉพาะที่ตรงกัน อย่างไรก็ตาม การจะสร้างลักษณ์ชนิดนี้ได้ระบบก็จำเป็นต้องพึ่งพาความรู้เกี่ยวกับการรู้จำชื่อเฉพาะมาใช้ในระดับด้วย ซึ่งแน่นอนว่าต้องอาศัยความรู้ทางภาษาศาสตร์เกี่ยวกับโครงสร้างของคำที่มีลักษณะเป็นชื่อเฉพาะและโครงสร้างของคำที่ไม่ใช่ชื่อเฉพาะด้วย

4) ลักษณ์ทางวากยสัมพันธ์ (syntactic feature)

ลักษณ์ทางวากยสัมพันธ์เป็นลักษณ์พิจารณาเอาความสัมพันธ์ระหว่างส่วนของประโยคมาใช้เป็นลักษณ์ในการฝึกฝนแบบจำลองการเรียนรู้ ทั้งนี้ งานที่ใช้ลักษณ์ประเภทนี้สามารถยกตัวอย่างได้เช่นงานของวานและคณะ (Wan, Dras, Dale, & Paris, 2006) ที่เลือกใช้ค่าความแม่นยำจากความสัมพันธ์แบบพึ่งพาและค่าความครบถ้วนจากความสัมพันธ์แบบพึ่งพาเป็นลักษณ์ในการฝึกฝน ในกรณีนี้แต่ละประโยคจะถูกแจงส่วน (parse) เพื่อให้ได้ชุดของชุดของความสัมพันธ์แบบพึ่งพา (1 ชุดต่อประโยค) ความสัมพันธ์แบบพึ่งพานี้จะอยู่ในรูปคู่ของคำในแผนผังต้นไม้แบบพึ่งพาในฐานะส่วนหลักและส่วนพึ่งพา ทั้งนี้ค่าความแม่นยำและค่าความครบถ้วนสามารถคำนวณได้จากสมการที่ 2.1 และ 2.2 ตามลำดับ

$$\text{precision}_d = \frac{|\text{relations}(\text{sentence}_1) \cap \text{relations}(\text{sentence}_2)|}{|\text{relations}(\text{sentence}_1)|} \quad (2.1)$$

$$\text{recall}_d = \frac{|\text{relations}(\text{sentence}_1) \cap \text{relations}(\text{sentence}_2)|}{|\text{relations}(\text{sentence}_2)|} \quad (2.2)$$

นอกจากจะใช้ค่าความแม่นยำและค่าความครบถ้วนของความสัมพันธ์แบบพึ่งพาแล้ว วานและคณะยังเสนอให้ค่าระยะแก้ไขต้นไม้ (tree edit distance) เป็นลักษณ์อีกชนิดหนึ่งด้วย

งานของมาลากาซิโอติส (Malakasiotis, 2009) เป็นงานอีกชิ้นหนึ่งที่ใช้ประโยชน์จากความสัมพันธ์แบบพึ่งพาในการจำแนกประเภทการถอดความ โดยนอกจากจะอาศัยค่าความแม่นยำและค่าความครบถ้วนดังเช่นที่ปรากฏในงานของวานและคณะ งานชิ้นนี้ยังได้เพิ่มค่า F เป็นลักษณ์อีกตัวหนึ่งด้วยดังปรากฏในสมการที่ 2.3

$$F = \frac{2 \cdot \text{precision}_d \cdot \text{recall}_d}{\text{precision}_d + \text{recall}_d} \quad (2.3)$$

ทั้งนี้ เพื่อความเข้าใจที่ชัดเจนขึ้น ขอให้สังเกตคู่ของประโยคในตัวอย่างต่อไปนี้

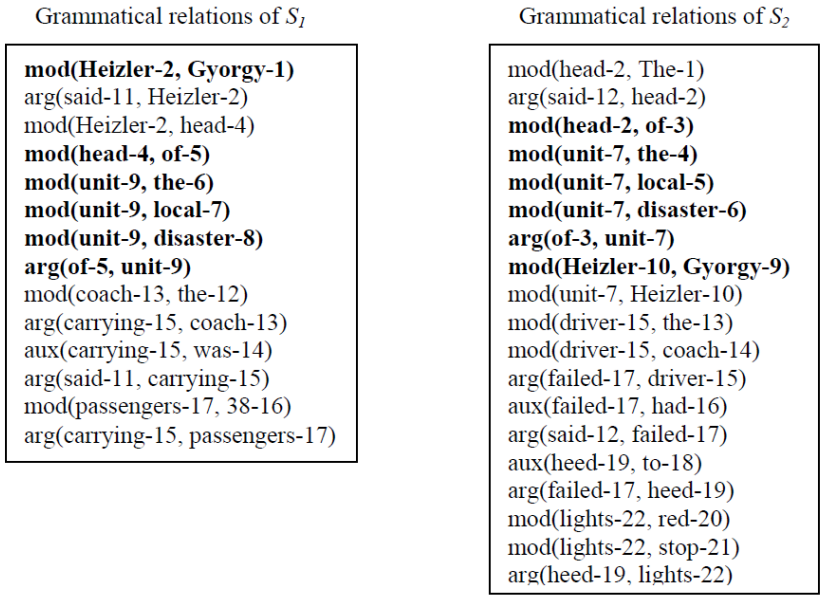
ตัวอย่าง 1:

S_1: Gyorgy Heizler, head of the local disaster unit, said thecoach was carrying 38 passengers.

S_2: The head of the local disaster unit, Gyorgy Heizler, said the coach driver had failed to heed red stop lights.

$precision_d$ = 0.43, $recall_d$ = 0.32, F = 0.36

ประโยค S_1 และ S_2 ในตัวอย่างนี้ไม่ได้เป็นการถอดความซึ่งกันและกันทำให้ค่าคะแนนที่ได้ออกมาต่ำ ดังจะเห็นได้จากความสัมพันธ์แบบพึ่งพาที่แสดงให้เห็นในภาพที่ 2.6



ภาพที่ 2.6 ความสัมพันธ์ทางไวยากรณ์ในตัวอย่าง 1 (Malakasiotis, 2009, p. 30)

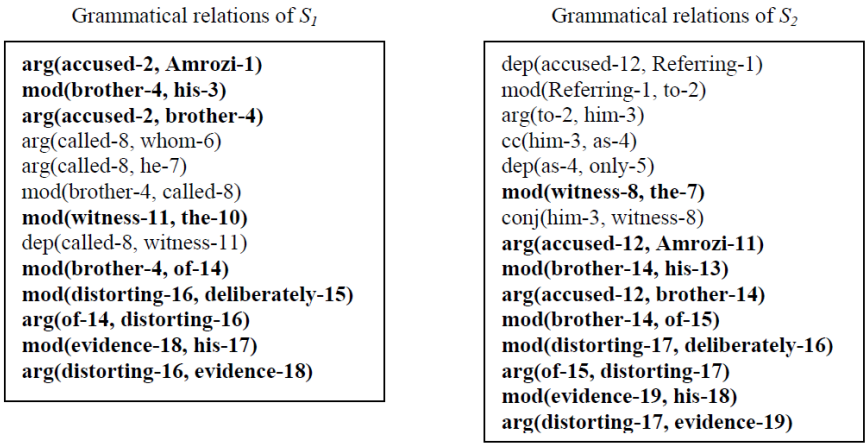
ตัวอย่าง 2:

S_1: Amrozi accused his brother, whom he called “the witness”, of deliberately distorting his evidence.

S_2: Referring to him as only “the witness”, Amrozi accused his brother of deliberately distorting his evidence.

$precision_d$ = 0.69, $recall_d$ = 0.6, F = 0.64

ในตัวอย่างที่ 2 นี้จะเห็นได้ว่าประโยค S_1 และ S_2 นี้มีความหมายเกือบเหมือนกัน ค่าคะแนนที่วัดได้จึงออกมาสูงมาก ซึ่งลักษณะดังกล่าวก็สอดคล้องไปกับความสัมพันธ์ทางไวยากรณ์ที่แจงส่วนได้ดังปรากฏในภาพที่ 2.7



ภาพที่ 2.7 ความสัมพันธ์ทางไวยากรณ์ในตัวอย่าง 2 (Malakasiotis, 2009, p. 30)

จากที่กล่าวมาจะเห็นได้ว่าการสร้างลักษณ์ทางวากยสัมพันธ์นี้จำเป็นต้องอาศัยความรู้ทางภาษาศาสตร์พอสมควร ทั้งนี้เนื่องจากในการทำความเข้าใจความสัมพันธ์ระหว่างหน่วยต่างๆ ในประโยคนั้นต้องใช้ความรู้เกี่ยวกับทฤษฎีไวยากรณ์ ยกตัวอย่างเช่นที่ปรากฏในงานของวานและคณะ (Wan et al., 2006); มาลากาซิโอติส (Malakasiotis, 2009); และชงและคณะ (Chong et al., 2010) คณะผู้วิจัยก็จำเป็นต้องอาศัยความรู้เกี่ยวกับทฤษฎีไวยากรณ์พึ่งพาในการแจงส่วนประโยคเช่นกัน

5) ลักษณ์ทางความหมาย (semantic feature)

ลักษณ์ทางความหมายเป็นลักษณ์อีกประเภทหนึ่งต้องอาศัยความรู้ทางภาษาศาสตร์ในการสร้างหรือสกัดมาก เนื่องมาจากลักษณ์ชนิดนี้จะมุ่งพิจารณาความสัมพันธ์ทางความหมายเป็นสำคัญ เพื่อที่จะตรวจจับความละม้ายให้ได้ถึงระดับความหมายของส่วนของข้อความได้ ความสัมพันธ์ทางความหมายดังกล่าวจะถูกพิจารณาในระดับคำเป็นหลัก ไม่ว่าจะเป็นคำไวพจน์ คำที่มีความหมายตรงกันข้าม คำจ่ากลุ่ม คำลูกกลุ่ม ทั้งนี้ การสร้างลักษณ์ประเภทนี้ได้ ระบบจำเป็นต้องพึ่งพาเครือข่ายคำ (WordNet) ซึ่งมีการจัดลำดับชั้นและความสัมพันธ์ตามความหมายของคำเอาไว้

ในส่วนของงานที่ใช้ลักษณ์ทางความหมายในการฝึกฝนแบบจำลองการเรียนรู้นั้นสามารถยกตัวอย่างได้เช่นงานของบรอกเคตต์และโดแลน (Brockett & Dolan, 2005) ซึ่งได้ใช้ลักษณ์หลายประเภทในการระบุการถอดความ ลักษณ์ประเภทหนึ่งที่งานนี้เลือกใช้ได้แก่การจับคู่ทางศัพท์ใน

เครือข่ายคำ (WordNet lexical mapping) ที่จะพิจารณาจับคู่คำไวพจน์ในคลังข้อมูลขนาดใหญ่ จากนั้นจึงใช้คู่ของคำที่มีความหมายเหมือนกันเหล่านั้นมาใช้เป็นลักษณ์ในการจำแนกประเภท

นอกจากนี้ แล้วยังมีงานของโคซาเรวาและมอนโตโย (Kozareva & Montoyo, 2006) ซึ่งนอกจากจะใช้ลักษณ์ทางศัพท์ดังได้กล่าวไปแล้ว งานชิ้นนี้ยังใช้ลักษณ์ทางความหมายในการจำแนกประเภทข้อความที่มีการถอดความอีกด้วย โดยประยุกต์ใช้ค่าความละม้ายของคำปรากฏในงานของลิน (D. Lin, 1998) ซึ่งหาได้จากเครือข่ายคำเพื่อคำนวณหาอัตราส่วนความละม้ายของคำนาม/คำกริยาระหว่าง 2 ประโยค จากสมการ $sim_{Lin}=\frac{\sum_{i=1}^{n} sim(T_1,T_2)_{Lin}}{n}$ ทั้งนี้ ค่าความละม้ายของคำของลินนั้นสามารถเรียกใช้ได้จาก WordNet::Similarity package (Pedersen, Patwardhan, & Michelizzi, 2004) ซึ่งเป็นแพ็กเกจที่ใช้หาค่าความละม้ายของคำในเครือข่ายคำ (WordNet)

ทั้งนี้ หากพิจารณาจากความรู้ทางภาษาศาสตร์ที่ลักษณ์ทางความหมายแล้วอาจทำให้เข้าใจได้ว่าลักษณ์ประเภทนี้จะสามารถจับ (capture) ข้อมูลในระดับความหมายซึ่งเป็นตัวบ่งชี้ความละม้ายได้ดีกว่าลักษณ์ประเภทอื่นๆ อย่างไรก็ตาม ในการใช้งานจริงนั้นลักษณ์ทางความหมายอาจไม่ให้ประสิทธิภาพที่ดีที่สุดเสมอไป ดังจะเห็นได้จากงานของโคซาเรวาและมอนโตโย (Kozareva & Montoyo, 2006, p. 530) ที่เปรียบเทียบการระบุการถอดความโดยใช้ลักษณ์ 2 ประเภท ได้แก่ ลักษณ์ทางศัพท์ และลักษณ์ทางความหมาย ผลปรากฏว่าลักษณ์ทางศัพท์สามารถระบุการถอดความได้มีประสิทธิภาพดีกว่าลักษณ์ทางความหมาย

อย่างไรก็ตาม จะเห็นได้ว่าลักษณ์ทางความหมายที่ปรากฏใช้ในงานที่กล่าวมาข้างต้นนั้นเป็นการประยุกต์ใช้เพียงค่าความละม้ายของคำ ในแง่นี้ ผู้วิจัยเห็นว่าหากนำค่าความละม้ายของคำดังกล่าวมาคำนวณรวมในระดับประโยคแล้วใช้เป็นลักษณ์ในการจำแนกประเภทข้อความที่มีความละม้ายได้ก็อาจทำให้ได้ผลการจำแนกในระดับที่น่าพอใจมากขึ้น ดังนั้น ในงานวิจัยชิ้นนี้ ผู้วิจัยจะใช้ค่าความละม้ายทางความหมายที่วัดได้จากการคำนวณรวมค่าความละม้ายของคำในประโยคเป็นลักษณ์ในการฝึกฝนด้วย

6) ลักษณ์ค่าการประเมินผลการแปลภาษาด้วยเครื่อง (MT evaluation metric feature)

ลักษณ์ค่าการประเมินผลการแปลภาษาด้วยเครื่องนี้เป็นลักษณ์ที่ได้จากการประยุกต์ใช้เทคนิคการประเมินผลของการแปลภาษาด้วยเครื่อง โดยมีแนวคิดว่าในการแปลภาษาต้นทางไปสู่ภาษาเป้าหมายด้วยเครื่องที่มีประสิทธิภาพนั้น ผล (output) ที่ได้ออกมาต้องมีความความใกล้ชิด (closeness) กับประโยคที่ระบบรับเข้าไป (input) ความใกล้ชิดดังกล่าวเป็นเทคนิคที่ลักษณ์ประเภทนี้นำมาใช้จับ (capture) ความหมายในการสมมูลกันของประโยค ในที่นี้จะยกตัวอย่างงานที่ใช้ลักษณ์

ประเภทนี้ 2 ชิ้นได้แก่งานของฟินช์และคณะ (Finch, Hwang, & Sumita, 2005) และมัทนานีและคณะ (Madnani, Tetreault, & Chodorow, 2012)

งานของฟินช์และคณะ (Finch et al., 2005) ได้ประยุกต์ใช้เทคนิคการประเมินผลของการแปลภาษาด้วยเครื่อง โดยอาศัยหลักการที่ว่าการแปลที่ดีนั้น ข้อมูลส่งออกจากการแปลภาษาด้วยเครื่องจะต้องมีความใกล้เคียงกับการแปลอ้างอิง (reference translation) ด้วยแนวคิดดังกล่าว งานชิ้นนี้จึงได้เลือกใช้ค่าการประเมินภาษาด้วยเครื่อง 4 วิธี ได้แก่ การใช้ค่าคะแนน WER (Word error rate), PER (Position- independent word error rate) BLEU (Bilingual evaluation understudy), และ NIST (National institute of standards and technology) ในการฝึกฝนแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนพิจารณาว่าคู่เทียบของประโยคเป็นคู่ถอดความกันหรือไม่ โดยให้ประโยคถอดความแทนการแปลอ้างอิง (reference translation) และประโยคเทียบแทนข้อความที่ได้จากการแปลภาษาด้วยเครื่อง และใช้ชุดข้อมูลจากคลังข้อมูล 2 แหล่ง ได้แก่ Microsoft Research Paraphrase Corpus (MSRP) และ PASCAL Challenge's entailment recognition corpus (PASCAL) ทั้งนี้ ผลจากการศึกษาได้ชี้ให้เห็นว่าเทคนิคการประเมินผลของการแปลด้วยเครื่องสามารถนำมาใช้เป็นลักษณ์ในการจำแนกการถอดความได้เป็นอย่างดี

มัทนานีและคณะ (Madnani et al., 2012) ได้นำแนวคิดที่ปรากฏอยู่ในงานของฟินช์และคณะ (Finch et al., 2005) มาพัฒนาต่อโดยเพิ่มค่าการประเมินผลการแปลภาษาด้วยเครื่องที่ได้จากการทบทวนงานของผู้อื่นเข้าไปอีก 6 ตัว ได้แก่ TER, TERp, METEOR, SEPIA, BADGER, และ MAXSIM นอกจากนี้ยังได้ขยายขอบเขตของการวิจัยเพิ่มเติมขึ้นโดยแทนที่จะจำแนกความละม้ายในการถอดความเพียงอย่างเดียว งานชิ้นนี้ยังได้เพิ่มการจำแนกความละม้ายในการลักลอกงานวิชาการเข้าไปด้วย ซึ่งผลที่ได้ออกมานั้นปรากฏว่าให้ประสิทธิภาพในระดับที่น่าพอใจมากกว่างานของฟินช์และคณะโดยประเมินจากค่าความถูกต้องและค่า F ที่ได้ถึง 77.4 และ 84.1 ตามลำดับ เปรียบเทียบกับงานของฟินช์และคณะที่ได้เพียง 75.0 และ 82.7 ตามลำดับเท่านั้น

อย่างไรก็ดี การที่งานของมัทนานีและคณะให้ผลที่มีประสิทธิภาพดีกว่างานของฟินช์และคณะนั้นอาจเป็นผลมาจากการที่ค่าการประเมินผลที่มัทนานีและคณะใช้เป็นลักษณ์บางตัวมีการประยุกต์ใช้ความรู้ทางภาษาศาสตร์มากกว่า เช่น TERp ที่ใช้พิจารณาการแก้ไขโดยอิงการทำให้เป็นต้นเค้า การมีความหมายเหมือนกันของคำ และการถอดความ หรือ METEOR ที่อาศัยข้อมูลคำไวพจน์จากเครือข่ายคำ เป็นต้น ทั้งนี้ จะสังเกตได้ว่าจากเดิมที่ลักษณ์ในงานของฟินช์และคณะใช้ความรู้ทางภาษาศาสตร์ในระดับคำก็ได้พัฒนาสู่การใช้ความรู้ทางภาษาศาสตร์ในระดับความหมายในงานของมัทนานีและคณะ ลักษณะดังกล่าวจึงเป็นข้อสนับสนุนหนึ่งที่ชี้ให้เห็นได้ว่าการประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับลึกกว่าย่อมให้ผลในการจำแนกประเภทที่ดีกว่า

ตารางที่ 2.2 การเปรียบเทียบประสิทธิภาพในการตรวจวัดความละม้ายด้วยการจำแนกประเภทโดยใช้การเรียนรู้ของเครื่อง

Reference	Description	Features	Acc.	F
Finch et al. (2005)	combination of MT evaluation measures as features	MT matric	75.0	82.7
Kozareva and Montoyo (2006)	combination of lexical and semantic features	Lexical + semactic	76.6	76.9
Qiu, Kan, and Chua (2006)	sentence dissimilarity classification	syntactic	72.0	81.6
Wan et al. (2006)	dependency-based features	syntactic	75.6	83.0
Madnani et al. (2012)	combination of eight MT metrics	MT matric	**77.4**	**84.1**

จากลักษณ์ประเภทต่างๆ ที่ได้กล่าวถึงมาทั้งหมดนี้จะทำให้ได้ว่าการตรวจวัดความละม้ายข้อความด้วยแนวทางการจำแนกประเภทข้อความที่มีความละม้ายโดยใช้การเรียนรู้ของเครื่องอาศัยการประยุกต์ใช้ความรู้ทางภาษาศาสตร์มากน้อยแตกต่างกันไป ตั้งแต่ระดับที่ไม่ได้อาศัยความรู้ทางภาษาศาสตร์ใดๆ เป็นเพียงวิธีอิงสถิติ เรื่อยมาจนกระทั่งใช้ความรู้ในระดับคำ ระดับวากยสัมพันธ์ และระดับความหมาย แน่นอนว่าการประยุกต์ใช้ความรู้ทางภาษาศาสตร์ที่แตกต่างกันย่อมให้ประสิทธิภาพในจำแนกประเภทเพื่อตรวจวัดความละม้ายของข้อความแตกต่างกันด้วย ดังแสดงให้เห็นในตารางที่ 2.2 ซึ่งยกตัวอย่างงานที่จำแนกประเภทความละม้ายในการถอดความซึ่งใช้ชุดข้อมูลจากแหล่งเดียวกันได้แก่ Microsoft Research Paraphrase Corpus (MSRP) (Dolan, Quirk, & Brockett, 2004) ซึ่งประกอบด้วยคู่ประโยคจำนวน 5,801 คู่ แบ่งเป็นชุดข้อมูลฝึกฝนจำนวน 4,076 คู่ของประโยค (คู่ของประโยค 2,753 คู่ หรือคิดเป็นร้อยละ 67.5 ได้รับการตัดสินโดยมนุษย์ว่าเป็นการถอดความจริง) และเป็นชุดข้อมูลทดสอบจำนวน 1,725 คู่ของประโยค (คู่ของประโยค 1,147 คู่ หรือคิดเป็นร้อยละ 66.5 ได้รับการตัดสินโดยมนุษย์ว่าเป็นการถอดความจริง)

ทั้งนี้ จากตารางที่ 2.2 จะเห็นได้ลักษณ์ที่มีประสิทธิภาพดีที่สุดในการจำแนกประเภทข้อความที่มีความละม้ายในกรณีของการถอดความนั้น ได้แก่ ลักษณ์ค่าการประเมินผลการแปลภาษาด้วยเครื่อง ลักษณ์ทางวากยสัมพันธ์ และลักษณ์ทางศัพท์ ตามลำดับ จึงสรุปได้ว่าการอาศัยความรู้ทางภาษาศาสตร์ในระดับที่แตกต่างกันย่อมมีผลต่อประสิทธิการจำแนกประเภท กล่าวคือ ลักษณ์ที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับสูงกว่าย่อมให้ผลการจำแนกที่ดีกว่า ดังจะเห็นได้จากงาน

ของโคซาเรวาและมอนโตโย (Kozareva & Montoyo, 2006) และมัทนานีและคณะ (Madnani et al., 2012) ที่ได้นำความรู้ในระดับความหมายมาใช้ดังได้กล่าวไปแล้วข้างต้น

อย่างไรก็ดี ในการเรียนรู้ของเครื่องนั้น ผู้สอนจำเป็นต้องทดลองใช้ลักษณ์หลายประเภทเพื่อทดสอบว่าลักษณ์ประเภทใดให้ประสิทธิภาพในการจำแนกที่ดีที่สุด ซึ่งโดยส่วนใหญ่แล้วจะมักไม่ปรากฏว่างานชิ้นใดใช้ลักษณ์ประเภทใดประเภทหนึ่งเพียงประเภทเดียว กล่าวคือ ลักษณ์ที่ดีสุดในการจำแนกประเภทนั้นอาจผสมกันระหว่างลักษณ์หลายตัวหรือหลายประเภทได้

2.7.3 การวัดค่าความละม้ายของข้อความ

ในหัวข้อ 2.7.2 ที่ผ่านมา ผู้วิจัยได้กล่าวถึงแนวทางการจำแนกข้อความที่มีความละม้ายกันโดยอาศัยเทคนิคการเรียนรู้ของเครื่องไปแล้ว ซึ่งเป็นการตัดสินว่าข้อความคู่ที่เทียบกันนั้นมีความละม้ายหรือไม่ ในหัวข้อย่อยนี้ ผู้วิจัยจะได้กล่าวถึงแนวทางการตรวจวัดความละม้ายของข้อความซึ่งไม่จำเป็นต้องอาศัยเทคนิคการจำแนกประเภทด้วยการเรียนรู้ของเครื่อง แต่ใช้วิธีการวัดค่าความละม้ายระหว่างข้อออกมาเป็นตัวเลข และตัดสินว่าข้อความละม้ายกันหรือไม่จากการกำหนดค่าขีดแบ่ง (threshold) แนวทางดังกล่าวนี้เรียกว่า การวัดค่าความละม้ายของข้อความ (text similarity measure) อย่างไรก็ตาม ค่าความละม้ายที่จะกล่าวถึงในหัวข้อนี้บางชนิด เช่น ค่าความละม้ายจากคำศัพท์ อาจปรากฏใช้เป็นลักษณ์ในการฝึกฝนการเรียนรู้ของเครื่องด้วย ดังได้กล่าวไปในหัวข้อ 2.7.2 แล้ว

การวัดค่าความละม้ายของข้อความ (text similarity measure) เป็นแนวคิดที่มีใช้มานานในด้านประมวลผลภาษาธรรมชาติและสาขาที่เกี่ยวข้อง โดยเฉพาะอย่างยิ่งในด้านการค้นคืนสารสนเทศ (information retrieval) จุดประสงค์ของการวัดความละม้ายนั้นเป็นไปเพื่อระบุข้อความ 2 ข้อความที่รับเข้ามีความสมมูลกันทางความหมายหรือไม่ (Achananuparp, Hu, & Shen, 2008, p. 1) โดยคะแนนค่าความละม้ายที่ได้จากการวัดนั้นจะเป็นค่าจำนวนจริงที่ทำให้เป็นบรรทัดฐาน (normalized real number value) จาก 0 ถึง 1 ด้วยเหตุนี้ การวัดความละม้ายจึงอาจเป็นที่รู้จักในชื่ออื่นด้วย เช่น การวัดค่าความละม้ายทางความหมาย (semantic similarity measure) หรือความสัมพันธ์ทางความหมาย (semantic relatedness) เป็นต้น

ในการวัดค่าความละม้ายระหว่างข้อความใดๆ นั้นมีปัจจัยที่ต้องพิจารณาอยู่ 3 ประการด้วยกัน ประการแรก ได้แก่ ความยาวของข้อความ ทั้งนี้อาจกล่าวได้ความยาวที่เหมาะสมของข้อความนั้นขึ้นอยู่กับประเภทและวัตถุประสงค์ของงานที่นำไปใช้ เช่น ในงานค้นคืนสารสนเทศอาจจะต้องหาค่าความละม้ายของเอกสารทั้งฉบับกับข้อคำถาม (query) ซึ่งอาจเป็นเพียงคำหรือวลีสั้น ในขณะที่งานบางประเภทอย่างการรู้จำการถอดความจะต้องการข้อความที่มีความยาวในระดับ

ประโยคเท่านั้น ปัจจัยที่ต้องพิจารณาประการต่อมาได้แก่ การแทนรูปข้อความ (text representation) กล่าวได้ว่าการแทนรูปข้อความเป็นส่วนสำคัญที่ต้องพิจารณาในการวัดค่าความละม้าย เนื่องจากเกี่ยวข้องโดยตรงกับวิธีที่จะใช้วัด กล่าวคือ รูปแทนบางชนิดอาจทำงานได้ดีกับวิธีการวัดบางวิธีเท่านั้น ทั้งนี้ ข้อความนั้นสามารถแทนรูปได้หลายแบบ (Metzler, Dumais, & Meek, 2007, p. 17) ไม่ว่าจะเป็นรูปอักขระพื้นผิวอันเป็นรูปแทนที่เรียบง่ายที่สุด หรืออาจแทนเป็นรูปต้นเค้า (stemmed representation) ตลอดจนการแทนรูปด้วยข้อมูลทางวากยสัมพันธ์หรือความหมาย และปัจจัยประการสุดท้ายที่ต้องพิจารณาคือ วิธีการวัดค่าความละม้าย ซึ่งมีความแตกต่างกันขึ้นอยู่กับความต้องการของงานแต่ละประเภท

ดังกล่าวไปแล้วข้างต้นว่าแนวคิดเรื่องการวัดค่าความละม้ายระหว่างข้อความนั้นเป็นที่แพร่หลายไปในแขนงงานต่างๆ ดังนั้นในส่วนของวิธีการวัดค่าความละม้ายนี้ ผู้วิจัยจึงได้รวบรวมวิธีการวัดค่าความละม้ายจากงานแขนงต่างๆ ไว้ ไม่ว่าจะเป็นการค้นคืนสารสนเทศ การระบุการถอดความ การรู้จำการโยงไปสู่ข้อสรุปทางข้อความ และการลักลอกงานวิชาการ โดยได้จัดประเภทวิธีการวัดค่าใหม่ พร้อมกันนั้นจะได้อภิปรายถึงการนำความรู้ทางภาษาศาสตร์มาใช้ในแต่ละวิธีอีกด้วย ทั้งนี้วิธีการวัดค่าความละม้ายต่างๆ มีรายละเอียดดังต่อไปนี้

1) การวัดค่าความละม้ายจากคำศัพท์

การวัดค่าความละม้ายจากคำศัพท์นี้เป็นวิธีพื้นฐานที่ใช้ทั่วไปในการวัดความละม้ายของข้อความ ทั้งนี้สามารถแบ่งเป็นประเภทย่อยได้ 2 ประเภท ได้แก่ การวัดค่าคำทับซ้อน และการวัดค่าความละม้ายระหว่างสายคำ

ในส่วนของการวัดค่าคำทับซ้อน (word overlap similarity measure) นั้นเป็นการคำนวณจำนวนคำที่สองข้อความมีร่วมกัน ในการคำนวณแบบง่ายนั้นจะคำนวณจากจำนวนคำที่สองข้อความมีร่วมกันแล้วหารด้วยจำนวนคำในข้อความแรกดังสมการต่อไปนี้

$$sim_{overlap}(S_1,S_2) = \frac{|S_1 \cap S_2|}{|S_1|} \quad (2.4)$$

สมการที่ 2.4 กำหนดให้ $sim_{overlap}(S_1,S_2)$ เป็นค่าความละม้ายของข้อความที่คำนวณจากคำทับซ้อนใน 2 ข้อความ ในขณะที่ S_1 และ S_2 เป็นคำที่มีอยู่ข้อความ 2 ข้อความตามลำดับ

นอกจากการคำนวณค่าคำทับซ้อนแบบง่ายแล้ว ค่าคำทับซ้อนอีกตัวที่เป็นที่นิยมใช้คือค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ด (Jaccard similarity coefficient) ซึ่งเป็นการเปรียบเทียบจำนวน

คำที่ 2 ข้อความมีร่วมกันกับจำนวนคำทั้งหมดของทั้งสองข้อความ ดังได้แสดงสมการให้เห็นไปบ้างแล้วในหัวข้อ 2.7.2

จากวิธีการวัดค่าต้นจะเห็นได้ว่าด้วยเหตุที่วัดค่าความละม้ายของข้อความจากคำทับซ้อนทำได้โดยไม่ซับซ้อน ค่าความละม้ายที่ได้ออกมาจึงไม่ตรงกับความเป็นจริงเท่าใดนัก ทั้งนี้เพราะเป็นการเทียบอัตราส่วนระหว่างจำนวนคำในสองข้อความเท่านั้น

วิธีการวัดค่าความละม้ายทางศัพท์อีกประเภทหนึ่ง ได้แก่ การวัดค่าความละม้ายระหว่างสายคำ (string similarity measure between word) ในวิธีนี้นอกจากจะวัดได้จากการหาลำดับเหมือนที่ยาวที่สุดในชุดอักขระ (longest common subsequence: *lcs*) และการหาสายคำเหมือนที่ยาวที่สุดในชุดอักขระ (longest common substring) ดังได้กล่าวมาในข้างต้นบ้างแล้ว ยังอาจประยุกต์การวัดดังกล่าวเพิ่มเติมได้ด้วย ดังเช่นที่ปรากฏใช้ในงานของอิสลามและอิงเพน (Islam & Inkpen, 2009) ที่ได้หาค่าที่ทำให้เป็นบรรทัดฐานของ *lcs* จากนั้นจึงใช้อัลกอริทึม maximal consecutive longest common subsequence (MCLSC) เพื่อหาช่วงที่ต่อเนื่องกันมากที่สุดของสายคำ

จากที่กล่าวมาจะเห็นได้ว่าการวัดความละม้ายด้วยวิธีนี้นั้นได้อาศัยความรู้ทางภาษาศาสตร์ในระดับคำเท่านั้น กล่าวคือ ต้องสามารถแบ่งคำได้ หรือสามารถหารูปต้นเค้าของคำได้ด้วยความรู้ด้านวิทยาหน่วยคำ

2) การวัดค่าความละม้ายจากค่าน้ำหนักคำ

การให้ค่าน้ำหนักคำในข้อความเป็นอีกแนวคิดหนึ่งที่ใช้ในการวัดค่าความละม้ายของข้อความ วิธีการนี้มีแนวคิดพื้นฐานที่ว่าคำแต่ละคำในข้อความนั้นมีความสำคัญไม่เท่ากัน คำที่ปรากฏถี่สูงในเอกสารที่เกี่ยวข้องหรือในคลังข้อมูลย่อมมีความสำคัญมากกว่า

ค่า tf*idf (term frequency–inverse document frequency) เป็นวิธีการหาค่าน้ำหนักของคำที่สนใจในเอกสารวิธีหนึ่งที่นิยมใช้อย่างแพร่หลาย โดยค่า tf คือจำนวนครั้งที่คำที่สนใจ (term) ที่ปรากฏในเอกสาร (document) หากค่า tf มากจะแสดงว่าคำที่สนใจปรากฏในเอกสารมาก ในขณะที่ df คือจำนวนของเอกสารที่ปรากฏคำที่สนใจอยู่ หากค่า df น้อยจะแสดงว่าคำที่สนใจปรากฏอยู่ในเอกสารไม่มาก ไม่ได้เป็นคำที่ปรากฏโดยทั่วไปในเอกสาร ลักษณะนี้อำนาจจำแนก (discrimination power) จะสูง การคิดค่า idf จะเป็นค่าแทนอำนาจจำแนกดังกล่าว โดยคำนวณได้จากสมการที่ 2.5 ซึ่งกำหนดให้ D เป็นจำนวนเอกสารทั้งหมดที่มีอยู่ในคลังข้อมูล ส่วน $d: t \in d$ นั้นหมายถึงจำนวนของเอกสารซึ่งมีคำ t ปรากฏอยู่ ทั้งนี้ เมื่อหาค่า idf ได้แล้วก็จะสามารถหาค่า tf*idf ได้จากสมการที่ 2.6

$$\mathrm{idf(t)} = \log \frac{|D|}{|\{d: t \in d\}|} \tag{2.5}$$

$$\mathrm{tf} \times \mathrm{idf}(t,d) = \mathrm{tf}(t,d) \times \mathrm{idf}(t) \quad (2.6)$$

ในการวัดค่าความละม้ายของข้อความด้วยค่า tf*idf นี้จำเป็นต้องการแทนรูปข้อความในแบบจำลองปริภูมิเวกเตอร์ (vector space model) กล่าวคือเป็นการแทนขนาดแต่ละมิติของเวกเตอร์ด้วยค่าน้ำหนัก tf*idf จึงคำนวณค่าความละม้ายจากค่าความละม้ายโคไซน์ *(*cosine similarity*)* ระหว่างเวกเตอร์ของ 2 ข้อความ

มิฮัลเซียและคณะ (Mihalcea, Corley, & Strapparava, 2006) ได้ประยุกต์แนวคิดเรื่องการให้น้ำหนักคำ *idf* มาใช้ในการวัดค่าความละม้ายแบบคำต่อคำ (word-to-word similarity) ด้วยสมการที่ 2.7 ซึ่งกำหนดให้ *T*1 และ *T*2 เป็นประโยคที่ต้องการหาค่าความละม้าย idf(w) เป็นค่าน้ำหนักที่ได้จากการคำนวณในคลังข้อมูลขนาดใหญ่ (ในงานชิ้นนี้ได้ใช้คลังข้อมูล BNC) ส่วน maxSim นั้นเป็นค่าความละม้ายแบบคำต่อคำ (word-to-word similarity) ที่สูงที่สุด ซึ่งงานชิ้นนี้ มิฮัลเซียและคณะได้ทดลองหาค่าดังกล่าวจากวิธีการวัด 8 วิธีเพื่อให้ได้ค่าที่สุดซึ่งรวมถึงวิธีการวัดที่อาศัยเครือข่ายคำด้วย อย่างไรก็ตาม ผลปรากฏว่าค่าที่ใช้ในการหาค่า maxSim ที่มีประสิทธิภาพมากที่สุดนั้น ได้แก่ PMI-IR ที่ได้จากสมการที่ 2.8

$$sim(T1,T2) = \tfrac{1}{2}\left(\frac{\sum_{w\in\{T1\}}(\max Sim(w,T2)*idf(w))}{\sum_{w\in\{T1\}} idf(w)} + \frac{\sum_{w\in\{T2\}}(\max Sim(w,T1)*idf(w))}{\sum_{w\in\{T2\}} idf(w)}\right) \quad (2.7)$$

$$PMI-IR(w1,w2) = \log_2 \frac{hits(w1 AND w2)*WebSize}{hits(w1)*hits(w2)} \quad (2.8)$$

หากพิจารณาจากภาพรวมแล้ววิธีนี้ถือว่าเป็นวิธีการหาค่าความละม้ายที่อิงสถิติของคำเป็นหลัก ดังนั้นจึงมีการประยุกต์ใช้ความรู้ทางภาษาศาสตร์น้อย อย่างไรก็ดี พบว่าวิธีการวัดค่าความละม้ายแบบคำต่อคำ (word-to-word similarity) ที่ปรากฏในงานของมิฮัลเซียและคณะ (Mihalcea et al., 2006) นี้ให้ประสิทธิภาพดีกว่าวิธีการวัดบางวิธีที่ประยุกต์ใช้ความรู้ทางวากยสัมพันธ์และอรรถศาสตร์ ในประเด็นนี้ผู้วิจัยจะได้กล่าวถึงในตอนท้ายของการสรุปวิธีการวัดค่าความละม้ายของข้อความ

3) การวัดค่าความละม้ายทางวากยสัมพันธ์

การวัดค่าความละม้ายทางวากยสัมพันธ์นี้เป็นการใช้ประโยชน์จากข้อมูลทางวากยสัมพันธ์ได้แก่ตำแหน่งของคำในข้อความเป็นหลัก ในด้านการใช้ลำดับของคำเพื่อวัดค่าความละม้ายของประโยคนั้น ลีและคณะ (Y. Li et al., 2006, p. 1143) ได้เสนอว่า การใช้แนวคิดถุงใส่คำ (bag-of-word) เพียงอย่างเดียวจะทำให้ไม่สามารถวัดความละม้ายของประโยคได้อย่างมีประสิทธิภาพ เพราะ

แม้จะสามารถตัดคำออกมาได้แต่ลำดับของคำในประโยคส่งมีผลทำให้ความแตกต่างกัน ยกตัวอย่างเช่นประโยค T_1 และ T_2 ต่อไปนี้

T_1: A quick brown dog jumps over the lazy fox.

T_2: A quick brown fox jumps over the lazy dog.

จากประโยค T_1 และ T_2 จะเห็นได้ว่าแม้รูปคำในทั้ง 2 ประโยคจะเหมือนกันแต่การสลับตำแหน่งของคำว่า dog และ fox กลับทำให้ความหมายของประโยคต่างกัน หากใช้แนวคิดถุงคำในการวัดค่าความละม้ายของประโยคทั้งสอง ผลที่ออกมาก็จะแสดงว่าประโยค T_1 และ T_2 สมมูลกัน ดังนั้นลีและคณะจึงได้เลือกแปลงประโยค T_1 และ T_2 ไปเป็นเวกเตอร์ของลำดับคำ r_1 และ r_2 โดยเทียบคำชุดคำรวมของทั้ง 2 ประโยค T ทั้งนี้ แต่ละตำแหน่งของคำในเวกเตอร์ของลำดับคำ r_i จะได้มาโดยการคำนวณค่าความละม้ายของคำระหว่างคำที่สนใจ w กับทุกคำในประโยค s_i ด้วยการคิดค่าน้ำหนักคำเช่นนี้ ตำแหน่งของคำใน s_i ที่ให้ค่าความละม้ายของคำได้ใกล้เคียงกับ w มากที่สุดจะถูกเลือกมาใช้ในเวกเตอร์ ดังตัวอย่างด้านล่าง

T: [A quick brown dog jumps over the lazy fox]

r_1 = [1 2 3 4 5 6 7 8 9]

r_2 = [1 2 3 9 5 6 7 8 4]

ด้วยเหตุที่เวกเตอร์ของลำดับเป็นข้อมูลทางโครงสร้างพื้นฐานที่ปรากฏอยู่ในประโยค การวัดความละม้ายของประโยคจึงเป็นไปเพื่อวัดว่าลำดับของคำใน 2 ประโยคนั้นแตกต่างกันอย่างไร โดยใช้สมการที่ 2.9 โดย S_r แทนค่าความละม้ายของลำดับคำระหว่างเวกเตอร์ของลำดับคำ r_1 ที่แปลงได้จากประโยค T_1 กับเวกเตอร์ของลำดับคำ r_2 ที่แปลงได้จากประโยค T_2

$$S_r = 1 - \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \qquad (2.9)$$

จากที่กล่าวมาจะเห็นได้ว่าการวัดค่าความละม้ายทางวากยสัมพันธ์นี้สามารถแก้ไขปัญหาที่เกิดจากการใช้วิธีการวัดความละม้ายที่ประยุกต์ใช้แนวคิดถุงใส่คำได้ ผู้วิจัยจึงเห็นว่าควรนำวิธีการวัดค่าความละม้ายจากลำดับคำมาทดสอบในการวิจัยชิ้นนี้ด้วย ทั้งนี้ ผู้วิจัยจะได้กล่าวถึงการประยุกต์ใช้วิธีการวัดค่าความละม้ายนี้โดยละเอียดในบทที่ 6

4) การวัดค่าความละม้ายทางความหมาย

ในการวัดค่าความละม้ายระหว่างข้อความด้วยวิธีการวัดประเภทนี้ ข้อมูลทางความหมายถือเป็นสิ่งสำคัญที่ใช้ในการวัด ทั้งนี้ ข้อมูลทางความหมายที่ใช้พิจารณาในที่นี้ได้แก่ ความสัมพันธ์ทาง

ความหมายระหว่างคำซึ่งได้จากการอาศัยเครือข่ายคำ (WordNet) ในส่วนของงานที่ใช้วิธีการวัดค่าประเภท ผู้วิจัยจะยกตัวอย่างงาน 2 ชิ้น ได้แก่งานของลีและคณะ (Y. Li et al., 2004; Y. Li et al., 2006) และงานของเฟอร์นันโดและสตีเวนสัน (Fernando & Stevenson, 2008)

งานของลีและคณะ (Y. Li et al., 2004; Y. Li et al., 2006) นั้น นอกจากจะเสนอวิธีการวัดค่าความละม้ายของประโยคจากลำดับคำของประโยคดังได้กล่าวไปแล้ว งานชิ้นนี้ยังได้เสนอวิธีการวัดค่าความละม้ายทางความหมายของประโยคอีกด้วย โดยได้เสนอให้ใช้เวกเตอร์ทางความหมาย (semantic vector) ของประโยคที่สร้างขึ้นจากค่าความละม้ายของแต่ละคำในประโยค ในวิธีนี้ค่าน้ำหนักคำจะหาได้จากคะแนนความละม้ายทางความหมายสูงสุดระหว่างคำในเวกเตอร์ลักษณะ (feature vector) กับคำในชุดคำรวมของทั้ง 2 ประโยคที่ตรงกัน จากนั้นจึงหาค่าความละม้ายทางความหมายระหว่างประโยคได้จากค่าสัมประสิทธิ์ความละม้ายโคไซน์ (cosine similarity) ระหว่างเวกเตอร์ทางความหมายของสองประโยค ดังปรากฏในสมการที่ 2.10 โดย S_s แทนค่าสัมประสิทธิ์ความละม้ายโคไซน์ระหว่างเวกเตอร์ทางความหมายของประโยคที่ 1 (s_1) กับเวกเตอร์ทางความหมายของประโยคที่ 2 (s_2)

$$S_s = \frac{s_1 \cdot s_2}{||s_1|| \cdot ||s_2||} \tag{2.10}$$

ทั้งนี้ เพื่อให้ได้ค่าความละม้ายของประโยคที่มีความถูกต้องและแม่นยำมากขึ้น ลีและคณะ (Y. Li et al., 2004, p. 4; Y. Li et al., 2006, p. 1144) ได้รวมการวัดค่าทางความหมายกับการวัดค่าทางวากยสัมพันธ์เข้าด้วยกันโดยอาศัยเวกเตอร์ทางความหมายและเวกเตอร์ลำดับคำ ดังปรากฏในสมการที่ 2.11 ด้านล่าง โดย $S(T_1,T_2)$ แทนค่าความละม้ายทางวากยสัมพันธ์และความหมายของประโยค T1 และ T2 ซึ่งคำนวณได้จากการนำสมการที่ 2.9 และ 2.10 มารวมกันโดยมีสัมประสิทธ์ δ เป็นตัวควบคุมความสัมพันธ์เกื้อหนุน (relative contribution) ของค่าความละม้ายจากเวกเตอร์ของลำดับคำและค่าความละม้ายทางความหมาย ทั้งนี้ ลีและคณะแนะนำให้ค่าสัมประสิทธิ์ควบคุมความสัมพันธ์เกื้อหนุนดังกล่าวมีค่าอยู่ระหว่าง 0.5 ถึง 1 โดยลีและคณะได้ใช้ค่า $\delta = 0.85$

$$S(T_1,T_2) = \delta S_s + (1-\delta)S_r = \delta \frac{S_1 \times S_2}{||S_1|| \times ||S_2||} + (1-\delta)\frac{||r_1 - r_2||}{||r_2 + r_1||} \tag{2.11}$$

ส่วนงานของเฟอร์นันโดและสตีเวนสัน (Fernando & Stevenson, 2008) นั้นเป็นการสร้างเมทริกซ์เพื่อวัดความละม้ายของคำระหว่างทุกคู่ของคำในคู่ของประโยคที่ต้องการวัดค่าความละม้ายโดยอาศัยค่าจากเครือข่ายคำ จากนั้นจึงตั้งค่าขีดแบ่ง (threshold) เพื่อระบุการถอดความ อย่างไรก็ดีงานชิ้นนี้มีข้อน่าสังเกตที่พึงชี้ให้เห็นอยู่ 2 ประการ ได้แก่ ประการแรก งานชิ้นนี้ไม่ให้ความสนใจกับ

โครงสร้างทางวายสัมพันธ์ของข้อความ เนื่องจากการวัดในเมทริกซ์นั้นเป็นการวัดเปรียบเทียบคำเดี่ยวแบบคู่ต่อคู่ ลักษณะดังกล่าวนี้แตกต่างจากงานของลีและคณะ (Y. Li et al., 2004; Y. Li et al., 2006) ที่พยายามจะใช้ความรู้ทั้งทางวากยสัมพันธ์และอรรถศาสตร์ในการวัดค่าความละม้าย นอกจากนี้แล้ว งานชิ้นนี้ยังไม่ให้ความสนใจกับลักษณะอื่นๆ ของคำในเมทริกซ์ ยกตัวอย่างเช่น การเปรียบเทียบสายอักขระ หรือค่าน้ำหนักของคำ หรือกล่าวอีกนัยหนึ่งคือ งานชิ้นนี้สนใจคำนวณพียงแค่ความหมายของคำเท่านั้น

จากงานที่ยกตัวอย่างมาทำให้เห็นได้ว่าวิธีการวัดค่าความละม้ายทางความหมายนี้มีทั้งส่วนที่ใช้ความรู้ทางภาษาศาสตร์ไล่เรียงเรื่อยมาตั้งแต่ลำดับคำ วากยสัมพันธ์ และอรรถศาสตร์ ดังที่ปรากฏให้เห็นในงานของลีและคณะ (Y. Li et al., 2004; Y. Li et al., 2006) และส่วนที่ใช้ความรู้เฉพาะเรื่องความสัมพันธ์ทางความหมายของคำเพียงอย่างเดียวโดยไม่สนใจความรู้ทางวากยสัมพันธ์ใดๆ

จากวิธีการวัดค่าความละม้ายระหว่างข้อความวิธีต่างๆ ที่ได้กล่าวถึงมาทั้งหมดจะเห็นได้ว่าการตรวจวัดความละม้ายด้วยแนวทางนี้ก็อาศัยการประยุกต์ใช้ความรู้ทางภาษาศาสตร์มากน้อยแตกต่างกัน ตั้งแต่ระดับที่ไม่ได้อาศัยความรู้ทางภาษาศาสตร์ใดๆ จนกระทั่งถึงระดับคำ ระดับวากยสัมพันธ์ และระดับความหมาย ทั้งนี้ หากเปรียบเทียบผลจากตัวอย่างงานวัดค่าความละม้ายในการถอดความซึ่งใช้ชุดข้อมูลในการทดสอบจากแหล่งเดียวกันได้แก่ Microsoft Research Paraphrase Corpus (MSRP) (Dolan et al., 2004) แล้วจะเห็นได้ว่าวิธีการวัดค่าความละม้ายที่มีการประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในขั้นสูงกว่านั้นย่อมให้มีประสิทธิภาพในการวัดที่ดีกว่าดังได้แสดงในตารางที่ 2.3

ตารางที่ 2.3 การเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายระหว่างข้อความ

Reference	Description	method	Acc.	F
Mihalcea et al. (2006)	combination of several word similarity measures	word weight	70.3	81.3
Islam and Inkpen (2007)	combination of semantic and string similarity	Lexical + semantic	72.6	81.3
Fernando and Stevenson (2008)	WordNet similarity with matrix	semantic	**74.1**	**82.4**
Rus et al. (2008)	dependency graph subsumption	syntactic	70.6	80.5

จากตารางที่ 2.3 ด้านบน หากพิจารณาจากค่าความถูกต้อง (accuracy) แล้วจะเห็นได้ว่า วิธีวัดค่าความละม้ายระหว่างข้อความที่อาศัยความรู้ทางภาษาศาสตร์ในขั้นสูงกว่าอันได้แก่งานเฟอร์นัน

โดและสตีเวนสัน (Fernando & Stevenson, 2008) ที่ใช้ความสัมพันธ์ทางความหมายในการวัดค่านั้นให้ค่าความถูกต้องมากที่สุด รองลงมาเป็นงานของอิสลามและอิงเพน (Islam & Inkpen, 2009) ที่ประยุกต์ใช้การวัดค่าทางความหมายเข้ากับการวัดค่าทางศัพท์ อันดับต่อมาจึงเป็นงานของรัสและคณะ (Rus, McCarthy, Lintean, McNamara, & Graesser, 2008) ที่ใช้โครงสร้างทางวากยสัมพันธ์ในการวัดค่าความละม้าย และงานของมิฮัลเซียและคณะ (Mihalcea et al., 2006) ที่ใช้การวัดค่าน้ำหนักคำอิงจากคลังข้อมูล ตามลำดับ ผลที่ได้นี้จึงเป็นเครื่องยืนยันได้ในทางหนึ่งว่าในการวัดค่าความละม้ายระหว่างข้อความนั้น หากเลือกใช้วิธีการวัดที่อาศัยความรู้ทางภาษาศาสตร์สูงแล้วก็จะทำให้การวัดมีความถูกต้องมากกว่าวิธีที่อาศัยความรู้ทางภาษาศาสตร์ในระดับต้นเช่นระดับคำศัพท์ อย่างไรก็ดี หากพิจารณาจากค่า F ซึ่งเป็นค่าเฉลี่ยฮาร์มอนิกระหว่างค่าความแม่นยำกับค่าความครบถ้วนแล้วจะเห็นได้ว่าวิธีการวัดแต่ละวิธีให้ประสิทธิภาพที่ใกล้เคียงกัน ยกเว้นงานของของรัสและคณะ (Rus et al., 2008) ที่ให้ค่า F ค่อนข้างต่ำกว่างานอื่นๆ ทั้งนี้ ผู้วิจัยเห็นว่าน่าจะเป็นผลมาจากระเบียบวิธีในการวัดที่ไม่ได้ให้ความสำคัญกับโครงสร้างทางวากยสัมพันธ์ซึ่งทำให้วิธีนี้มีประสิทธิภาพต่ำลงดังที่ผู้วิจัยได้กล่าวถึงไปแล้วข้างต้น

อย่างไรก็ตาม ตารางที่ 2.3 ข้างต้นไม่รวมงานของลีและคณะ (Y. Li et al., 2004; Y. Li et al., 2006) ไว้ในการเปรียบเทียบด้วย เนื่องจากใช้ชุดข้อมูลในการทดสอบคนละชุดกัน แต่จากผลการทดสอบประสิทธิภาพที่ปรากฏในงานของพลากร อาชานานุภาพ และคณะ (Achananuparp et al., 2008, pp. 312-315) ก็ช่วยสนับสนุนได้ว่าวิธีการวัดค่าความละม้ายจากเวกเตอร์ของลำดับคำและเวกเตอร์ทางความหมายที่ลีและคณะเสนอนั้นก็ให้ประสิทธิภาพในระดับที่น่าพอใจเช่นกัน ด้วยเหตุนี้ผู้วิจัยจึงสนใจจะประยุกต์ใช้แนวคิดของลีและคณะในวิจัยชิ้นนี้ด้วยโดยจะกล่าวถึงรายละเอียดอีกครั้งในบทที่ 6

เมื่อได้แสดงเทคนิควิธีการตรวจวัดความละม้ายของข้อความไปทั้ง 2 แนวทางแล้ว ในส่วนต่อไป ผู้วิจัยจะกล่าวถึงการเปรียบเทียบประสิทธิภาพในการตรวจวัดความละม้ายจากทั้งสองแนวทางพร้อมกันนั้นจะได้ชี้ให้เห็นถึงข้อดีและข้อจำกัดของแต่ละวิธีด้วย ทั้งนี้ ขอให้พิจารณาตารางที่ 2.4 ต่อไปนี้

ตารางที่ 2.4 การเปรียบเทียบประสิทธิภาพของวิธีการตรวจวัดความละม้ายของข้อความ

Reference	Description	ML	Acc.	F
Finch et al. (2005)	combination of MT evaluation measures as features	+	75.0	82.7
Kozareva & Montoyo (2006)	combination of lexical and semantic features	+	76.6	76.9
Mihalcea et al. (2006)	combination of word similarity measures	-	70.3	81.3
Qiu et al. (2006)	sentence dissimilarity classification	+	72.0	81.6
Wan et al. (2006)	dependency-based features	+	75.6	83.0
Islam and Inkpen (2007)	combination of semantic and string similarity	-	72.6	81.3
Fernando & Stevenson (2008)	WordNet similarity with matrix	-	**74.1**	**82.4**
Rus et al. (2008)	dependency graph subsumption	-	70.6	80.5
Socher et al. (2011)	recursive auto encoder with dynamic pooling	+	76.8	83.6
Madnani et al. (2012)	combination of eight MT metrics	+	**77.4**	**84.1**

ตารางที่ 2.4 ข้างต้นเป็นตารางแสดงการเปรียบเทียบประสิทธิภาพในการระบุการถอดความโดยใช้แหล่งข้อมูลทดสอบเดียวกันคือ Microsoft Research Paraphrase Corpus (MSRP) (Dolan et al., 2004) ซึ่งประกอบด้วยคู่ประโยคจำนวน 5,801 คู่ แบ่งเป็นชุดข้อมูลฝึกฝนจำนวน 4,076 คู่ของประโยค (คู่ของประโยค 2,753 คู่ หรือคิดเป็นร้อยละ 67.5 ได้รับการตัดสินโดยมนุษย์ว่าเป็นการถอดความจริง) และเป็นชุดข้อมูลทดสอบจำนวน 1,725 คู่ของประโยค (คู่ของประโยค 1,147 คู่ หรือคิดเป็นร้อยละ 66.5 ได้รับการตัดสินโดยมนุษย์ว่าเป็นการถอดความจริง) ทั้งนี้ จะเห็นได้ว่าวิธีการตรวจวัดความละม้ายของข้อความที่ใช้ในการระบุการถอดความที่ปรากฏในตารางนี้มีทั้งที่เป็นแนวทางการจำแนกประเภทข้อความที่มีความละม้ายโดยใช้การเรียนรู้ของเครื่อง (+ML) และแนวทางการวัดค่าความละม้ายระหว่างข้อความ (-ML) จากตารางดังกล่าวจะเห็นได้อย่างชัดเจนว่า แนวทางการจำแนกประเภทข้อความที่มีความละม้ายโดยใช้การเรียนรู้ของเครื่องนั้นให้ประสิทธิภาพในการตรวจวัดความละม้ายที่ดีกว่าแนวทางการวัดค่าความละม้ายระหว่างข้อความ ดังจะเห็นได้จากงานของ

มัทนานีและคณะ (Madnani et al., 2012) ที่ได้ค่าความถูกต้องและค่า F ถึง 77.4 และ 84.1 ตามลำดับ ในขณะที่งานของเฟอร์นันโดและสตีเวนสัน (Fernando & Stevenson, 2008) นั้นได้ค่าความถูกต้องและค่า F เพียง 74.1 และ 82.4 ตามลำดับ ค่าตัวเลขดังกล่าวนี้สามารถยืนยันได้ว่าแนวทางการจำแนกประเภทข้อความที่มีความละม้ายโดยใช้การเรียนรู้ของเครื่องให้ประสิทธิภาพที่ดีกว่า แม้กระนั้นเองทั้งสองแนวทางดังกล่าวก็มีข้อดีและข้อจำกัดให้ผู้พัฒนาระบบต้องพิจารณาในการเลือกใช้เช่นเดียวกัน

หากกล่าวถึงข้อดีของแนวทางการจำแนกประเภทข้อความที่มีความละม้ายโดยใช้การเรียนรู้ของเครื่องแล้วจะพบว่าประสิทธิภาพของแนวทางนี้ขึ้นอยู่กับการเลือกใช้และเรียงลำดับลักษณ์ในการฝึกฝนเป็นสำคัญ การที่สามารถเลือกใช้ลักษณ์ได้อย่างหลากหลายจึงถือเป็นข้อดีของแนวทาง ซึ่งลักษณ์ดังกล่าวก็สามารถรวมเอาการวัดค่าความละม้ายระหว่างข้อความด้วยวิธีต่างๆ มาใช้ได้ด้วยเช่นกัน กล่าวได้ว่าลักษณะเช่นนี้เองที่เป็นปัจจัยทำให้แนวทางนี้มีประสิทธิภาพเหนือกว่าแนวทางการวัดค่าความละม้ายระหว่างข้อความ อย่างไรก็ตาม แนวทางนี้ก็มีข้อจำกัดที่เห็นเด่นชัดอยู่ 2 ประการ กล่าวคือ ประการแรก แนวทางนี้ไม่สามารถคืนค่าความละม้ายระหว่างข้อความที่ต้องการออกมาเป็นตัวเลขได้จริง ผู้ที่เลือกใช้แนวทางนี้จึงไม่อาจทราบได้ว่าคู่ของข้อความที่ส่งให้ระบบตรวจจับความละม้ายนั้นมีความละม้ายคล้ายหรือไม่คล้ายกันมากน้อยเพียงใด เพราะระบบได้ทำเพียงจำแนกประเภทจากสมมติฐานที่ได้จากการเรียนรู้ฝึกฝนเท่านั้น ข้อจำกัดอีกประการหนึ่งปรากฏในงานของอิสลามและอิงเพน (Islam & Inkpen, 2009, p. 292) ที่ได้กล่าวไว้ว่า การหาลักษณ์ที่มีประสิทธิภาพและการจะได้ค่าสำหรับลักษณ์จากประโยคนั้นเป็นเรื่องที่ปฏิบัติได้ยากและอาจต้องใช้เวลานาน ในประเด็นนี้ ผู้วิจัยก็เห็นคล้อยตามกับความคิดของอิสลามและอิงเพน อย่างไรก็ดี ยังปรากฏว่ามีงานหลายชิ้นที่ใช้แนวทางนี้ออกมาอย่างสม่ำเสมอและงานเหล่านั้นก็ได้ให้ประสิทธิภาพในขั้นที่น่าพอใจ ผู้วิจัยจึงเห็นว่าการหาลักษณ์ที่มีประสิทธิภาพนั้นคงไม่เป็นการยากเกินไปนัก

อย่างไรก็ตาม ในขั้นตอนการวิเคราะห์เพื่อหาลักษณ์ที่จะนำมาใช้ในระบบจำแนกข้อความที่มีการลักลอกและไม่ลักลอกออกจากกันนั้น เนื่องด้วยในภาษาไทยยังไม่มีผู้ศึกษาการลักลอกไว้ในเชิงภาษาศาสตร์ ผู้วิจัยจึงไม่อาจปรับประยุกต์ข้อค้นพบงานวิจัยใดๆ มาสร้างเป็นลักษณ์ ฉะนั้นแล้ว ผู้วิจัยจึงเห็นจำเป็นต้องวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทยจากมุมมองเชิงภาษาศาสตร์ เพื่อสามารถนำข้อค้นพบที่ได้จากการวิจัยมาประยุกต์ใช้ในการสร้างลักษณ์ ตลอดจนใช้อภิปรายทำความเข้าใจผลอันจะเกิดขึ้นจากการใช้ลักษณ์ทางภาษานั้นๆ

ในอีกด้านหนึ่ง หากพิจารณาข้อดีที่ได้จากการใช้แนวทางการวัดค่าความละม้ายระหว่างข้อความนั้น จะเห็นได้ว่าการวัดค่าความละม้ายด้วยแนวทางนี้ ระบบสามารถคืนค่าความละม้ายออกมาเป็นตัวเลขได้ตั้งแต่ 0 ถึง 1 ทำให้ทราบได้ว่าคู่ของข้อความที่ส่งเข้าไปในระบบเพื่อวัดความ

ละม้ายนั้นมีความละม้ายคล้ายกันมากหรือน้อยเพียงใดในเชิงปริมาณ อย่างไรก็ตาม แนวทางนี้ก็มีข้อจำกัดที่ผู้วิจัยใคร่ชี้ให้เห็น ได้แก่ ประการแรก จะสังเกตได้ว่างานที่ใช้แนวทางนี้แล้วได้ผลที่มีประสิทธิภาพสูงนั้นจะอาศัยการพิจารณาความสัมพันธ์ทางความหมายของคำจากเครือข่ายคำทั้งสิ้น การจะพัฒนาการตรวจวัดความละม้ายในแนวทางนี้ให้มีประสิทธิภาพทัดเทียมกับงานอื่นๆ จึงถือเป็นข้อจำกัดประการหนึ่งในภาษาที่ยังไม่มีเครือข่ายคำที่สมบูรณ์ไว้ใช้งาน โดยเฉพาะอย่างยิ่ง ในกรณีของภาษาไทย ซึ่งเป็นภาษาที่งานวิจัยชิ้นนี้มุ่งพัฒนาระบบตรวจหาการลักลอก

แม้ว่าการตรวจวัดความละม้ายด้วยแนวทางการจำแนกประเภทข้อความที่มีความละม้ายโดยใช้การเรียนรู้ของเครื่องและแนวทางการวัดค่าความละม้ายระหว่างข้อความจะมีข้อดีและข้อจำกัดให้เลือกพัฒนาแตกต่างกันไป แต่จะเห็นได้ว่าลักษณ์ส่วนหนึ่งที่ใช้ในแนวทางการเรียนรู้ของเครื่องนั้นก็ได้มาจากวิธีการวัดค่าความละม้ายระหว่างข้อความ ในขณะเดียวกันลักษณ์ที่ใช้ในแนวทางการเรียนรู้ของเครื่องบางตัวก็เสนอวิธีการวัดค่าความละม้ายระหว่างข้อความที่ไม่เคยปรากฏใช้ในแนวทางทางการวัดค่าความละม้ายระหว่างข้อความมาก่อน ลักษณะดังกล่าวถือเป็นลักษณะที่ทั้งสองแนวทางมีร่วมกัน ซึ่งสามารถนำวิธีการวัดค่าความละม้ายนี้มาปรับประยุกต์ใช้ซึ่งกันและกันได้หากผู้พัฒนาระบบได้ทบทวนงานทั้งแนวทางไว้อย่างเพียงพอ นอกจากนี้แล้วยังอาจกล่าวได้ว่าข้อดีและข้อจำกัดที่ทั้งสองแนวทางมีต่างกันนั้นก็เป็นประเด็นวิจัยที่น่าสนใจในการพัฒนาระบบ ทั้งนี้ ผู้พัฒนาระบบอาจศึกษาการใช้ทั้ง 2 แนวทางร่วมกันในระบบเดียวเพื่อให้ข้อดีของแต่ละแนวทางช่วยส่งเสริมซึ่งกันและกันก็ได้ หรือมิเช่นนั้น ผู้พัฒนาระบบอาจศึกษาเปรียบเทียบประสิทธิภาพในการตรวจวัดความละม้ายของข้อความของทั้งสองแนวทางโดยใช้ข้อมูลชุดเดียวก็ย่อมได้เช่นกัน

อย่างไรก็ดี จะเห็นได้ว่างานที่ผู้วิจัยยกมาอ้างอิงถึงประสิทธิภาพในการตรวจวัดความละม้ายในตารางที่ *2.4* ข้างต้นนั้นเป็นงานที่ตรวจจับการถอดความทั้งหมด ทั้งนี้ หากได้ประยุกต์ใช้แนวทางการตรวจวัดดังกล่าวในข้อความที่มีระดับความคลุมเครือ (degree of obfuscation) ต่ำกว่าการถอดความ เช่น ข้อความที่มีการคัดลอกจากกันและกันโดยตรง หรือข้อความที่มีการคัดลอกโดยเปลี่ยนแปลงบางส่วน ผลที่ได้ก็ย่อมมีประสิทธิภาพสูงขึ้นไปอีก ดังที่ปรากฏในงานของมัทนานีและคณะ (Madnani et al., 2012) ที่นอกจากตรวจวัดความละม้ายในข้อความที่มีการถอดความจาก Microsoft Research Paraphrase Corpus แล้ว งานชิ้นนี้ยังตรวจวัดความละม้ายจากข้อความที่มีการลักลอกประเภทต่างๆ จากคลังข้อมูลการลักลอก PAN อีกด้วย ซึ่งได้ผลที่มีประสิทธิภาพสูงกว่าการตรวจวัดความละม้ายในข้อความที่มีการถอดความ ข้อค้นพบดังกล่าวนี้จึงช่วยสนับสนุนได้ว่าสมควรนำระเบียบวิธีในงานระบุการถอดความมาประยุกต์ใช้ในงานตรวจจับการลักลอกงานวิชาการด้วยเช่นกัน

ด้วยเหตุผลดังกล่าวมาทั้งหมดนี้เอง ในงานวิจัยชิ้นนี้ ผู้วิจัยจึงได้ตัดสินใจพัฒนาระบบตรวจหาการลักลอกงานวิชาการโดยใช้ทั้งแนวทางการจำแนกประเภทข้อความที่มีความละม้ายโดยใช้การเรียนรู้ของเครื่องและแนวทางการวัดค่าความละม้ายระหว่างข้อความร่วมกัน ซึ่งเชื่อได้ว่าระบบที่พัฒนาขึ้นจะให้ประสิทธิภาพในการตรวจหาการลักลอกที่ดีกว่างานชิ้นอื่นๆ ที่ผ่านมา

2.8 ทฤษฎีโครงสร้างวาทะ

ดังได้กล่าวไปในตอนท้ายของหัวข้อที่แล้ว การจะทำเข้าใจการลักลอกงานวิชาการภาษาไทยเพื่อสร้างคลังข้อมูลการลักลอก วิเคราะห์ลักษณ์ที่ใช้สำหรับจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอก กำหนดลักษณะของข้อมูลรับเข้า ตลอดจนอภิปรายผลที่ได้จากใช้งานระบบตรวจหาการลักลอกที่พัฒนาขึ้นนั้น จำเป็นอย่างยิ่งที่จะต้องอาศัยข้อค้นพบที่ได้จากการวิเคราะห์กลวิธีลักลอกงานวิชาภาษาไทยในเชิงภาษาศาสตร์ ดังนั้นในหัวข้อนี้ ผู้วิจัยจะได้กล่าวถึงทฤษฎีโครงสร้างวาทะ (rhetorical structure theory: RST) ซึ่งเป็นทฤษฎีหลักที่งานวิจัยชิ้นนี้เลือกใช้ในการศึกษาทำความเข้าใจกลวิธีลักลอกงานวิชาการภาษาไทย

ทฤษฎีโครงสร้างวาทะเป็นทฤษฎีว่าด้วยการจัดองค์ประกอบของข้อความที่พัฒนาขึ้นโดยแมนน์และทอมป์สัน (Mann & Thompson, 1988) เพื่อเป็นส่วนหนึ่งของโครงการสร้างข้อความอิงคอมพิวเตอร์ (computer-based text generation) แนวคิดหลักของทฤษฎีนี้คือการพิจารณาความสัมพันธ์ระหว่างหน่วยย่อยในปริจเฉทโดยยึดเจตนาของผู้ส่งสารและสาระของตัวสาร และแสดงออกมาโดยใช้แผนผังต้นไม้

ในทางทฤษฎีนั้น แมนน์และทอมป์สันได้กำหนดให้ช่วงของข้อความ[4] (text span) ใดๆ มีความสัมพันธ์กันในฐานะที่ช่วงของข้อความหนึ่งมีสถานะเป็นแกนกลาง (nucleus) และอีกช่วงของข้อความหนึ่งมีสถานะเป็นบริวาร (satellite) ช่วงของข้อความที่มีสถานะแกนกลางจะเป็นใจความหลัก เป็นส่วนที่จำเป็นและไม่สามารถละทิ้งได้ ในขณะที่ช่วงของข้อความที่มีสถานะบริวารคือข้อความส่วนที่ทำหน้าที่ขยายความให้ช่วงของข้อความที่มีสถานะแกนกลาง เป็นส่วนที่สามารถละทิ้งได้โดยไม่กระทบใจความหลัก ช่วงของข้อความแต่ละช่วงไม่ว่าจะมีสถานะใดก็ตามเมื่อประกอบเข้าด้วยกันในโครงสร้างปริจเฉทจะแสดงความสัมพันธ์หรือมีวาทสัมพันธ์ (rhetorical relation) อย่างใดอย่างหนึ่งต่อกันเสมอ โดยแมนน์และทอมป์สันเสนอให้มีวาทสัมพันธ์ทั้งหมด 23 แบบ

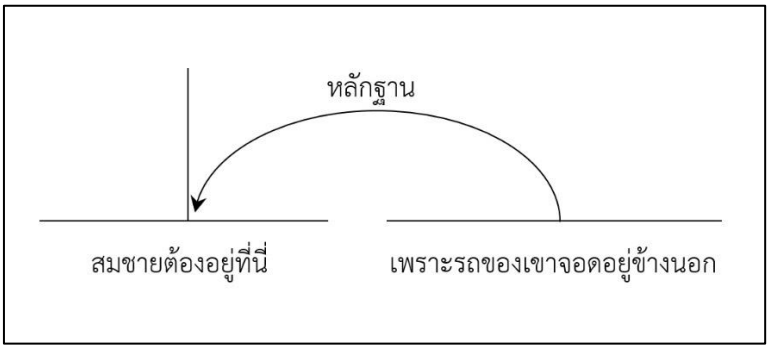
[4] ช่วงของข้อความในทฤษฎีฉบับดั้งเดิมมักอยู่ในรูปของอนุพากย์หรือประโยค

ในทฤษฎีฉบับดั้งเดิมของแมนน์และทอมป์สัน วาทสัมพันธ์จะได้รับการนิยามด้วยชุดของข้อบังคับ (constraint) ของแกนกลาง (nucleus: N) และบริวาร (satellite: S) ซึ่งเกี่ยวข้องกับเป้าหมายและความเชื่อของผู้เขียน (writer: W) และผู้อ่าน (reader: R) อันจะก่อให้เกิดผลกระทบ (effect) ต่อผู้อ่าน (reader: R) เพื่อความเข้าใจที่ชัดเจนยิ่งขึ้นขอให้พิจารณาข้อความตัวอย่างข้อบังคับของวาทสัมพันธ์แบบหลักฐาน (evidence) และแผนผังต้นไม้ต่อไปนี้

ข้อความตัวอย่าง : สมชายต้องอยู่ที่นี่ เพราะรถของเขาจอดอยู่ข้างนอก

ชื่อความสัมพันธ์ :	หลักฐาน
ข้อบังคับของ N :	R อาจจะไม่เชื่อ N ในระดับที่ W พอใจ
ข้อบังคับของ S :	R เชื่อมั่นใน S หรือพบว่า S เชื่อถือได้
ข้อบังคับของ N+S :	ความเข้าใจของ R ใน S ช่วยเพิ่มความเชื่อมั่นที่มีต่อ N
ผลกระทบ :	ความเชื่อมั่นของ R ใน N เพิ่มมากขึ้น

ข้อความตัวอย่างข้างต้นสามารถแบ่งช่วงของข้อความได้ 2 ช่วง ได้แก่ ช่วงของข้อความ [สมชายต้องอยู่ที่นี่] ซึ่งมีสถานะแกนกลาง (N) และช่วงของข้อความ [เพราะรถของเขาจอดอยู่ข้างนอก] หากพิจารณาตามข้อบังคับจะเห็นได้ว่าผู้อ่าน (R) อาจไม่เชื่อว่าสมชายอยู่ในตำแหน่งตามที่ข้อความระบุในระดับที่ผู้เขียนพอใจ แต่ในขณะเดียวข้อความ [เพราะรถของเขาจอดอยู่ข้างนอก] (S) ก็สามารถช่วยเป็นหลักฐานยืนยันความน่าเชื่อถือของข้อความ [สมชายต้องอยู่ที่นี่] (N) ได้ตามข้อบังคับของ N+S เป็นผลให้ความเชื่อมั่นของผู้อ่าน (R) ที่มีต่อข้อความ [สมชายต้องอยู่ที่นี่] (N) เพิ่มมากขึ้น ความสัมพันธ์ของช่วงของข้อความทั้งสองนี้สามารถแสดงให้ได้ดังแผนผังต้นในภาพที่ 2.8



ภาพที่ 2.8 แผนผังต้นไม้โครงสร้างวาทะแสดงวาทสัมพันธ์แบบหลักฐาน

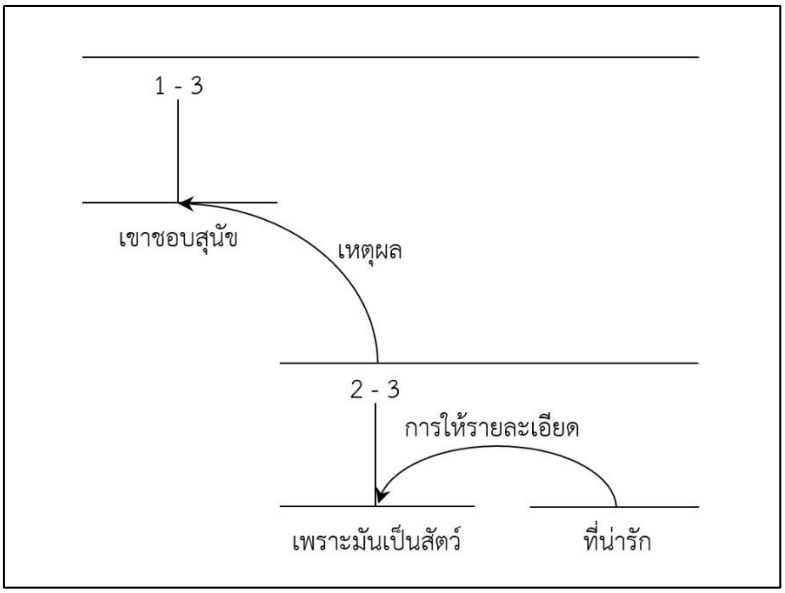
ภาพที่ 2.8 เป็นการแสดงความสัมพันธ์ระหว่างหน่วยย่อยในปริจเฉทผ่านแผนผังต้นไม้โครงสร้างวาทะ จะเห็นได้ว่าข้อความ “สมชายต้องอยู่ที่นี่ เพราะรถของเขาจอดอยู่ข้างนอก” สามารถแบ่งเป็นช่วงของข้อความได้ 2 ช่วง สังเกตเห็นได้จากเส้นตรงแนวนอนที่มีข้อความกำกับอยู่ ช่วงของข้อความที่มีสถานะแกนกลางจะมีเส้นตรงตั้งฉากกับเส้นตรงแนวนอนที่มีข้อความของช่วงนั้น ส่วนช่วง

ของข้อความที่มีสถานะบริวารจะมีเส้นโค้งที่ปลายด้านหนึ่งเป็นรูปลูกศรกำกับอยู่ โดยหัวลูกศรนี้จะชี้ไปยังเส้นแนวตั้งของช่วงของข้อความที่มีสถานะแกนกลางเพื่อแสดงสถานะความเป็นบริวารของอีกช่วงของข้อความหนึ่ง และมีชนิดของวาทสัมพันธ์ “หลักฐาน” กำกับอยู่เหนือเส้นโค้งดังกล่าว จากแผนผังในภาพที่ 2.8 จึงแปลความได้ว่าช่วงของข้อความ [เพราะรถของเขาจอดอยู่ข้างนอก] มีสถานะบริวาร ทำหน้าที่เป็นหลักฐานสนับสนุนช่วงของข้อความ [สมชายต้องอยู่ที่นี่] ซึ่งมีสถานะแกนกลาง

แนวคิดเชิงทฤษฎีของทฤษฎีโครงสร้างวาทะได้รับการพัฒนาให้สามารถนำไปใช้จริงได้มากยิ่งขึ้นเมื่อคาร์ลสันและคณะ (Carlson et al., 2001) ได้เสนอว่า ในโครงสร้างปริจเฉทมีหน่วยที่เล็กที่สุดซึ่งสามารถสื่อข้อมูลเชิงเนื้อหาและความหมายได้สมบูรณ์ เรียกว่าหน่วยปริจเฉทพื้นฐาน (elementary discourse units: EDUs) และได้ใช้แนวคิดเรื่องหน่วยปริจเฉทพื้นฐานนี้แทนแนวคิดเรื่องช่วงของข้อความ (text span) ที่เสนอไว้ในทฤษฎีฉบับดั้งเดิม อย่างไรก็ดี คาร์ลสันและคณะยังคงแนวคิดเรื่องสถานะความสำคัญ (nuclearity status) ระหว่างหน่วยปริจเฉทพื้นฐานแต่ละหน่วยไว้เช่นเดิม กล่าวคือ หน่วยปริจเฉทพื้นฐานแต่ละหน่วยในโครงสร้างปริจเฉทยังคงมีความสัมพันธ์ในสถานะแกนกลางหรือบริวารอยู่เช่นเดิม แต่สิ่งที่ถือว่าเป็นการพัฒนาในทฤษฎีฉบับของคาร์ลสันและคณะนี้คือการกำหนดหลักการในการแบ่งขอบเขตของหน่วยปริจเฉทพื้นฐานที่ชัดเจน (Carlson & Marcu, 2001) หลักการดังกล่าวสัมพันธ์กับชุดของวาทสัมพันธ์ที่คาร์ลสันและคณะได้เสนอขึ้นใหม่ทั้งหมด 16 กลุ่ม แบ่งเป็นชนิดของวาทสัมพันธ์ได้ 78 แบบ ยกตัวอย่างเช่น วาทสัมพันธ์แบบเปรียบต่าง (contrast) วาทสัมพันธ์แบบลักษณะ (attribution) วาทสัมพันธ์แบบรายการ (list) วาทสัมพันธ์แบบภูมิหลัง (background) เป็นต้น และด้วยเหตุที่ทฤษฎีฉบับนี้ใช้การพิจารณาสถานะความสำคัญควบคู่ไปกับประเภทของวาทสัมพันธ์จึงทำให้ไม่ต้องอาศัยข้อบังคับ (constraint) ในการกำหนดประเภทของวาทสัมพันธ์อีกต่อไป

ภาพที่ 2.9 แสดงความสัมพันธ์ระหว่างหน่วยย่อยในปริจเฉทของข้อความ “เขาชอบสุนัขเพราะมันเป็นสัตว์ที่น่ารัก” ผ่านแผนผังต้นไม้โครงสร้างวาทะ จะเห็นได้ว่าข้อความดังกล่าวสามารถตัดแบ่งหน่วยปริจเฉทพื้นฐานได้เป็น 3 หน่วย ได้แก่ [เขาชอบสุนัข] [เพราะมันเป็นสัตว์] และ [ที่น่ารัก] โดยระหว่างปริจเฉทพื้นฐาน [เพราะมันเป็นสัตว์] กับ [ที่น่ารัก] มีวาทสัมพันธ์แบบการให้รายละเอียด (elaboration) ซึ่งคาร์สันและมาร์คู (Carlson & Marcu, 2001, p. 33) ได้ระบุไว้ว่าในวาทสัมพันธ์แบบนี้ หน่วยปริจเฉทที่มีสถานะเป็นบริวารจะให้ข้อมูลเพิ่มเติมเกี่ยวกับเนื้อหาของหน่วยปริจเฉทพื้นฐานที่มีสถานะเป็นแกนกลาง ในที่นี้ หน่วยปริจเฉทพื้นฐาน [ที่น่ารัก] จึงมีสถานะเป็นบริวาร เนื่องจากทำหน้าให้ข้อมูลเพิ่มเติมเกี่ยวกับ “สัตว์” ในหน่วยปริจเฉทพื้นฐาน [เพราะมันเป็นสัตว์] และในลำดับชั้นที่สูงขึ้นไป หน่วยปริจเฉทพื้นฐาน [เพราะมันเป็นสัตว์] และ [ที่น่ารัก] ก็มีสถานะเป็นบริวารของหน่วยปริจเฉทพื้นฐาน [เขาชอบสุนัข] ตามที่คาร์สันและมาร์คู (Carlson & Marcu,

2001, p. 65) ระบุไว้ว่าในวาทสัมพันธ์แบบเหตุผล (reason) นั้น หน่วยปริจเฉทพื้นฐานที่มีสถานะเป็นแกนกลางจะต้องเป็นการกระทำของสิ่งมีชีวิต และหน่วยปริจเฉทพื้นฐานที่มีสถานะเป็นบริวารจะแสดงเหตุผลของการกระทำดังกล่าว จากตัวอย่างนี้จึงเห็นได้ว่าทฤษฎีฉบับของคาร์ลสันและคณะนั้นได้พรรณนาหลักการในการตัดสินสถานะความสำคัญของหน่วยปริจเฉทพื้นฐานไปพร้อมกับการระบุวาทสัมพันธ์ของหน่วยปริจเฉทพื้นฐานคู่นั้นๆ จึงไม่ต้องอาศัยข้อบังคับในการทำความเข้าใจประเภทของวาทสัมพันธ์ดังเช่นทฤษฎีฉบับดั้งเดิม



ภาพที่ 2.9 แผนผังต้นไม้โครงสร้างวาทะแสดงวาทสัมพันธ์ตามแบบที่คาร์ลสันและคณะเสนอ

ด้วยทฤษฎีฉบับที่เสนอโดยคาร์ลสันและคณะนี้มีหลักการในการกำหนดขอบเขตของหน่วยปริจเฉทพื้นฐานที่ชัดเจน อีกทั้งยังช่วยลดความยุ่งยากจากการตีความข้อบังคับในขั้นตอนการระบุประเภทของวาทสัมพันธ์ ทฤษฎีฉบับจึงเป็นที่ยอมรับและถูกนำไปประยุกต์ใช้อย่างแพร่หลายในหลายภาษา ในส่วนของภาษาไทยนั้น นลินี อินต๊ะซาว และวิโรจน์ อรุณมานะกุล (Intasaw & Aroonmanakun, 2013) ก็ได้กำหนดหลักการในการตัดแยกขอบเขตของหน่วยปริจเฉทพื้นฐานในภาษาไทยเอาไว้

ตลอดระยะเวลานับตั้งแต่มีการเสนอทฤษฎีโครงสร้างวาทะขึ้นมาจนกระทั่งปัจจุบัน ทฤษฎีดังกล่าวได้ถูกนำไปประยุกต์ใช้เป็นกรอบการวิเคราะห์ข้อมูลทางภาษาเรื่อยมา ไม่เฉพาะสาขาภาษาศาสตร์คอมพิวเตอร์เท่านั้น แต่ยังรวมถึงสาขาปริจเฉทวิเคราะห์ ภาษาศาสตร์เชิงทฤษฎี และภาษาศาสตร์จิตวิทยาด้วย (Taboada & Mann, 2006) ในส่วนที่เกี่ยวข้องกับภาษาไทยนั้นก็มีงานวิจัยหลายชิ้นที่นำแนวคิดจากทฤษฎีโครงสร้างวาทะไปใช้ ไม่ว่าจะเป็นงานที่เสนอแนวทางการตัดแยกอนุพากย์ในภาษาไทยโดยอิงขอบเขตของหน่วยปริจเฉทพื้นฐาน (Ketui, Theeramunkong, &

Onsuwan, 2012; จิรวรรณ เจริญสุข, 2549; นลินี อินต๊ะซาว, 2556) งานของธนา สุขวารี และคณะ (Sukvaree et al., 2004) ที่ได้เสนอการสรุปย่อข้อความ (text summarization) ในวงการเกษตร โดยอิงทฤษฎีโครงสร้างวาทะ หรืองานของเมทินี วัฒนะเมธานนท์ และคณะ (Wattanamethanont et al., 2005) ที่เสนอเทคนิคการระบุวาทสัมพันธ์โดยใช้แบบจำลองนาอีฟเบย์เป็นตัวแยกประเภทจากลักษณ์ที่ประกอบไปด้วยตัวบ่งชี้ปริจเฉทสัมพันธ์ วลีสำคัญ และการเกิดร่วมของคำ งานของสมนึก สินธุปวน และโอม ศรนิล (Sinthupoun & Sornil, 2010) ที่ได้เสนอแนวทางการวิเคราะห์โครงสร้างปริจเฉทในภาษาไทยโดยได้ใช้แบบจำลองฮิดเดนมาร์คอฟในการตัดแยกหน่วยปริจเฉทพื้นฐานด้วย นอกจากนี้ยังมีงานของนงนุช เกตุ้ย และคณะ (Ketui, Theeramunkong, & Onsuwan, 2015) ที่นำเสนอแนวทางการสรุปย่อพหุเอกสาร (multi-document) โดยอิงจากหน่วยปริจเฉทพื้นฐานภาษาไทย (Thai Elementary Discourse Units: TEDUs) ตัวอย่างงานเหล่านี้เป็นเครื่องยืนยันได้เป็นอย่างดีถึงความนิยมและความน่าเชื่อถือของทฤษฎีโครงสร้างวาทะในการประยุกต์ใช้กับภาษาไทย

ในส่วนที่เกี่ยวข้องกับงานวิจัยชิ้นนี้ ดังได้กล่าวไปในตอนต้นหัวข้อนี้ว่าการคลังข้อมูลการลักลอก การวิเคราะห์ลักษณ์ที่ใช้สำหรับจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอก การกำหนดลักษณะของข้อมูลรับเข้า ตลอดจนการอภิปรายผลที่ได้จากใช้งานระบบตรวจหาการลักลอกที่พัฒนาขึ้นนั้น จำเป็นอย่างยิ่งที่จะต้องอาศัยข้อค้นพบที่ได้จากการวิเคราะห์กลวิธีลักลอกงานวิชาภาษาไทยในเชิงภาษาศาสตร์ อีกทั้งงานที่ศึกษาเกี่ยวกับการลักลอกหรือการถอดความที่ผ่านมานั้นไม่ได้ให้ความสำคัญกับการวิเคราะห์ในระดับปริจเฉทเท่าที่ควร เป็นผลให้มุมมองเกี่ยวกับการลักลอกถูกจำกัดไว้เพียงระดับประโยคเท่านั้น ด้วยเหตุนี้ ผู้วิจัยจึงได้เลือกใช้ทฤษฎีโครงสร้างวาทะซึ่งเป็นทฤษฎีที่ใช้พิจารณาความสัมพันธ์ระหว่างหน่วยต่างๆ ในปริจเฉทในการศึกษากลวิธีลักลอกและประยุกต์ใช้ในระบบตรวจหาการลักลอก ทั้งนี้ เหตุผลสนับสนุนที่ทำให้ผู้วิจัยเลือกใช้ทฤษฎีดังกล่าวนั้นยังมีอยู่อีก 2 ประการ ดังนี้

ประการแรก หน่วยที่ใช้ในการวิเคราะห์การลักลอกหรือการถอดความนั้นควรเป็นหน่วยที่มีเนื้อหาประพจน์ (propositional content) สมบูรณ์เพื่อใช้แทนข้อมูลที่เป็นความคิดสำคัญ (idea) เพียงความคิดเดียว ทั้งนี้ เป็นที่เข้าใจกันว่าหน่วยดังกล่าวได้แก่ประโยค อย่างไรก็ตาม ประโยคในภาษาไทยนั้นไม่อาจระบุขอบเขตได้โดยชัดเจน หากนำประโยคมาใช้เป็นหน่วยในการวิเคราะห์ก็ย่อมก่อให้เกิดความยุ่งยากในการให้นิยามและระบุขอบเขตของประโยคซึ่งอาจมีได้หลากหลายตามแนวทฤษฎีไวยากรณ์สำนักต่างๆ ผู้วิจัยจึงเลือกใช้หน่วยปริจเฉทพื้นฐานเป็นหน่วยในการวิเคราะห์เนื่องจากเป็นหน่วยสร้างที่เล็กที่สุดที่มีข้อมูลเชิงเนื้อหาและความหมายสมบูรณ์ตามแนวคิดของทฤษฎีโครงสร้างวาทะ

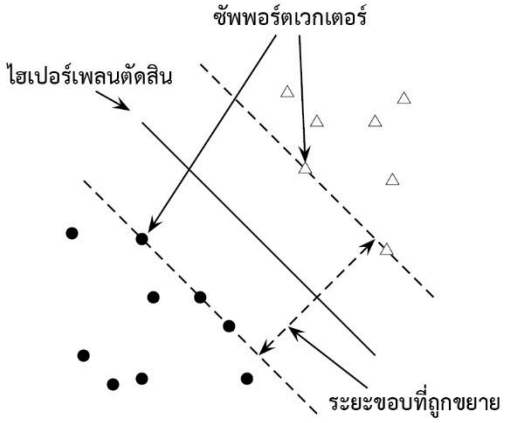
เหตุผลอีกประการหนึ่งนั้นเป็นผลสืบเนื่องมาจากเหตุผลประการแรก ทั้งนี้ หากพิจารณาในแง่วากยสัมพันธ์แล้ว หน่วยปริจเฉทพื้นฐานจะได้แก่อนุพากย์และวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น อันเป็นหน่วยที่สามารถระบุขอบเขตได้แน่นอน ลักษณะที่ชัดเจนของหน่วยปริจเฉทพื้นฐานนี้เอื้อให้สามารถนำหน่วยดังกล่าวไปใช้ได้จริงในขั้นประมวลผลข้อความ หรือหากจะมีผู้นำข้อค้นพบจากงานวิจัยชิ้นนี้ไปพัฒนาต่อในเชิงคอมพิวเตอร์ก็ย่อมทำได้โดยสะดวก ด้วยมีความเข้าใจเกี่ยวกับหน่วยที่ใช้ในการประมวลผลตรงกัน

จากที่กล่าวมาจะเห็นได้ว่าทฤษฎีโครงสร้างวาทะมีลักษณะที่เอื้อต่องานวิจัยชิ้นนี้ทั้งในเชิงแนวคิดทฤษฎีและการนำไปประยุกต์ใช้จริง ทั้งนี้ ผู้เขียนจะกล่าวถึงการนำทฤษฎีดังกล่าวมาใช้เป็นกรอบวิเคราะห์การลักลอกในบทถัดไป

2.9 แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (support vector machines: SVMs) เป็นวิธีเรียนรู้เชิงสถิติของเครื่องที่ใช้ในการจำแนกประเภทข้อมูล เช่นเดียวกับแบบจำลองอื่นๆ ไม่ว่าจะเป็นแบบจำลองต้นไม้ตัดสินใจ (decision tree model) แบบจำลองถดถอยแบบลอจิสติก (logistic regression model) หรือแบบจำลองโครงข่ายประสาทเทียม (neural network model) แต่แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนนั้นจะอาศัยแนวคิดเรื่องปริภูมิเวกเตอร์ (vector space) เพื่อแบ่งข้อมูล 2 ประเภทใดๆ ออกจากกันโดยสร้างระยะห่างที่กว้างที่สุด กล่าวคือ แบบจำลองนี้จะใช้ลักษณ์ (feature) ในข้อมูลฝึกฝนสร้างเวกเตอร์ในปริภูมิลักษณะ (feature space) ออกเป็นหลายมิติ (*n* มิติ) จากนั้นจึงหาไฮเปอร์เพลน (hyperplane) หรือระนาบเกินที่ใช้ในการแบ่งข้อมูลออกจากกัน (Manning, Raghavan, & Schütze, 2008, p. 293) โดยระนาบนี้จะถูกสร้างให้มีระยะห่างจากข้อมูลที่ต้องการแบ่งให้มากที่สุด (optimal margin hyperplane)

ภาพที่ 2.10 แสดงการทำงานของซัพพอร์ตเวกเตอร์แมชชีนที่สร้างไฮเปอร์เพลนสำหรับตัดสินแบ่งข้อมูลขึ้นก่อน จากนั้นจึงขยายขอบ (margin) ออกไปเรื่อยๆ จนกระทั่งสัมผัสซัพพอร์ตเวกเตอร์ทั้ง 5 จุด



ภาพที่ 2.10 การหาไฮเปอร์เพลนของซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีนสามารถใช้แยกข้อมูลได้ทั้งปัญหาที่แยกแบบเชิงเส้นได้และปัญหาที่ไม่สามารถแยกแบบเชิงเส้นได้ (ณรงค์ บุญสิริสัมพันธ์, 2546, น. 4) ในการแยกข้อมูลแบบเชิงเส้นนั้น สมมติให้มีเซตของข้อมูล *D* ที่ประกอบด้วยตัวอย่างจำนวน *l* ตัวในปริภูมิอันดับที่มี 2 ประเภทคือ +1 และ -1

$$D = \{(x_k, y_k) | k \in \{1, \dots, l\}, x_k \in \mathrm{R}^n, y_k \in \{+1, -1\} \quad (2.12)$$

ระนาบหลายมิติในปริภูมิอันดับ *n* ถูกกำหนดโดย (w, b) เมื่อ w คือเวกเตอร์ในปริภูมิอันดับที่ *n* ที่ตั้งฉากกับระนาบหลายมิติ และ *b* เป็นค่าคงที่ ระนาบหลายมิติ $(\mathrm{w} \cdot \mathrm{x}) + b$ จะแบ่งข้อมูลได้ก็ต่อเมื่อ

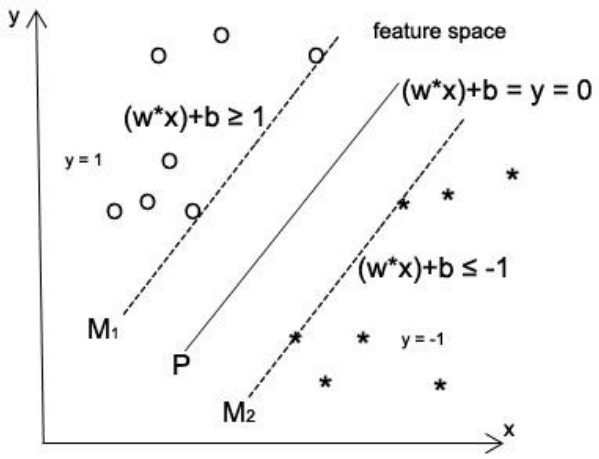
$$\begin{aligned} (\mathrm{w} \cdot \mathrm{x}_i) + b > 0 \text{ ถ้า } y_i = +1 \\ (\mathrm{w} \cdot \mathrm{x}_i) + b < 0 \text{ ถ้า } y_i = -1 \end{aligned} \quad (2.13)$$

หากต้องการค่า w และ b ที่ทำให้จุดที่อยู่ใกล้ระนาบหลายมิติมากที่สุดที่มีระยะห่าง $\frac{1}{|w|}$ แล้ว จะได้

$$\begin{aligned} (w \cdot x_i) + b \geq 1 \text{ ถ้า } y_i = +1 \\ (w \cdot x_i) + b \leq -1 \text{ ถ้า } y_i = -1 \end{aligned} \quad (2.14)$$

ซึ่งจะเท่ากับ

$$y_i[(w \cdot x_i) + b] \geq 1 \qquad \forall i \quad (2.15)$$



ภาพที่ 2.11 ไฮเปอร์เพลนที่สอดคล้องกับสมการที่ 2.14

ในการค้นหาไฮเปอร์เพลนหลายมิติที่ใช้แบ่งข้อมูลที่ดีสุดนั้นจะต้องค้นหาไฮเปอร์เพลนหลายมิติที่มีระยะห่างระหว่างข้อมูลที่ใช้สอนกับไฮเปอร์เพลนหลายมิติที่น้อยที่สุดมีค่ามากที่สุด ระยะห่างระหว่างข้อมูลตัวอย่างสองตัวจากประเภทที่แตกต่างกันมีค่าเท่ากับ

$$d(\mathrm{w},b)=\min_{\{\mathrm{x}_i|y_i=1\}}\frac{(\mathrm{w}\cdot\mathrm{x}_i)+b}{|\mathrm{w}|}-\max_{\mathrm{x}_i|y_i=-1}\frac{(\mathrm{w}\cdot\mathrm{x}_i)+b}{|\mathrm{w}|} \tag{2.16}$$

ส่วนกรณีของการแยกข้อมูลที่ไม่อาจแบ่งแบบเชิงเส้นได้ ซัพพอร์ตเวกเตอร์แมชชีนจะแก้ปัญหาโดยอาศัยหลักการแปลงข้อมูลจากปริภูมิขาเข้า (input space) ให้เป็นปริภูมิลักษณะ (feature space) ที่มีมิติสูงขึ้น (นันทณัฐ พันธุ์สีดา, 2556, น. 18-19) โดยมีสมมติฐานว่าข้อมูลที่ไม่สามารถแบ่งแบบเชิงเส้นได้ในปริภูมิอันดับต่ำ เมื่อส่งแปลงไปอยู่ในปริภูมิอันดับสูงจะมีความสัมพันธ์แบบเชิงเส้นได้ ในการนี้ ซัพพอร์ตเวกเตอร์แมชชีนจะใช้ฟังก์ชันการแปลง (map) ของฟังก์ชันเคอร์เนล (kernel mapping function) โดยมีฟังก์ชันเคอร์เนลที่นิยมใช้มีอยู่ 3 ประเภท ดังนี้

ฟังก์ชันโพลิโนเมียล (Polynomial function) คำนวณได้จากสมการที่ 2.17

$$K(\vec{x}_i,\vec{x}_j)=\left[\overrightarrow{x_i}\cdot\overrightarrow{x_j}+1\right]^d \tag{2.17}$$

ฟังก์ชันเรเดียลเบสิส (Radial Basis Function: RBF) คำนวณได้จากสมการที่ 2.18

$$K(\vec{x}_i,\vec{x}_j)=\exp(-\left|\left|\overrightarrow{x_i}-\overrightarrow{x_j}\right|\right|^2/(2\sigma^2)) \tag{2.18}$$

ฟังก์ชันซิกมอยด์ (Sigmoid) คำนวณได้จากสมการที่ 2.19

$$K(\vec{x}_i,\vec{x}_j)=\tanh\left[\gamma\left(\overrightarrow{x_i}-\overrightarrow{x_j}\right)+c\right] \tag{2.19}$$

ในส่วนของตัวอย่างงานประมวลผลภาษาธรรมชาติที่เลือกใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนนั้นก็มีอยู่ไม่น้อย ยกตัวอย่างเช่นงานของฟินช์และคณะ (Finch et al., 2005) ที่ได้ประยุกต์ใช้เทคนิคการประเมินผลของการแปลด้วยเครื่อง 4 วิธี ได้แก่ การใช้ค่าคะแนน BLEU NIST WER และ PER ในการฝึกฝนซัพพอร์ตเวกเตอร์แมชชีนเพื่อให้ทำนายความสมมูลทางความหมายระหว่างประโยค ทั้งนี้ ผลจากการศึกษาได้ชี้ให้เห็นว่าเทคนิคการประเมินผลของการแปลด้วยเครื่องสามารถนำมาใช้เป็นลักษณ์ในการจำแนกการถอดความได้เป็นอย่างดี

นอกจากงานของฟินช์และคณะแล้วยังมีงานของฉิวและคณะ (Qiu et al., 2006) ที่ใช้การประมวลผล 2 ขั้นเพื่อรู้จำการถอดความจากการตรวจหาความไม่ละม้ายกัน (dissimilarity) ระหว่างประโยค โดยในขั้นแรกจะเป็นการหาก้อนข้อมูล (information nuggets) หรือหน่วยที่เป็นเนื้อหาสำคัญทางความหมายในแต่ละประโยคจากแผนผังต้นไม้ทางวากยสัมพันธ์แล้วจับคู่ที่เหมือนกัน หากมีก้อนข้อมูลใดๆ ที่ไม่เกี่ยวข้องหลงเหลืออยู่จากการจับคู่ ก้อนข้อมูลดังกล่าวจะถูกนำมาใช้เป็นลักษณ์ในการตัดสินการถอดความระหว่างประโยคในขั้นที่ 2 กล่าวคือ หากคู่ของประโยคไม่มีก้อนข้อมูลที่ไม่ได้รับการจับคู่บรรจุอยู่หรือก้อนข้อมูลทั้งหมดไม่มีนัยสำคัญจะแสดงว่าประโยคคู่นั้น ๆ เป็นการถอดความ ในการนี้ ฉิวและคณะได้เลือกใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน SVM^{Light} ในการแยกประเภทคู่ของประโยคที่มีการถอดความและไม่มีการถอดความ ทั้งนี้ ก้อนข้อมูลในที่นี้คือทูเพิลที่แสดงความสัมพันธ์แบบเพรดิเคต-อาร์กิวเมนต์ (predicate-argument tuples) ที่เปรียบเทียบโดยใช้เทคนิคการจับคู่ศัพท์แบบพื้นฐาน (simple lexical matching technique)

ในส่วนการประมวลผลภาษาไทย นลินี อินต๊ะชาว (2556) ได้ใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนในการแยกอนุพากย์ภาษาไทย โดยใช้ฟังก์ชั่น SMO ของโปรแกรมวีก้า (Weka) และฟังก์ชันเคอร์เนลที่ใช้คือโพลีโนเมียล เพื่อให้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนตัดสินใจว่าข้อมูลรับเข้าในระดับคำเป็นคำขอบเขตเริ่มต้นอนุพากย์หรือไม่ การตัดสินใจของแบบจำลองอาศัยลักษณ์ทางภาษา ได้แก่ ลักษณ์หมวดคำปัจจุบัน หมวดคำก่อนหน้า หมวดคำตามหลัง รายการคำเชื่อมอนุพากย์ ความน่าจะเป็นของช่องว่างที่จะเป็นตัวแบ่งอนุพากย์ และเครื่องหมายวรรคตอน การเปรียบเทียบประสิทธิภาพของแต่ละลักษณ์ทำโดยกำหนดชุดของลักษณ์รูปแบบต่างๆ แล้วนำไปทดสอบ ผลปรากฏว่ารูปแบบของลักษณ์ที่ส่งผลต่อประสิทธิภาพของระบบมากที่สุด คือการใช้ทุกลักษณ์ร่วมกันทั้งหมด โดยให้ค่า F เท่ากับ 81.17 จากนั้น นลินีได้ทดลองปรับค่าพารามิเตอร์ของฟังก์ชันโพลีโนเมียลให้สูงขึ้น พบว่าระบบมีประสิทธิภาพดีขึ้น คือให้ค่า F เท่ากับ 84.74 เมื่อปรับค่าพารามิเตอร์ไว้ที่ D = 4

นอกจากนี้ยังปรากฏงานที่ใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนในการตรวจการลักลอกงานวิชาการภาษาไทยด้วย ได้แก่ งานของศิวพร ทวนไธสง (2556) ที่พัฒนาระบบการตรวจเทียบ

ภายในหาการลักลอกงานวิชาการในภาษาไทยโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนของของโปรแกรมวีก้า (Weka) ในงานชิ้นนี้ ศิวพรได้เปรียบเทียบประสิทธิภาพระบบที่สร้างขึ้นในแง่ของลักษณะข้อมูลรับเข้า ระหว่างข้อมูลรับเข้าที่เป็นคำกับข้อมูลรับเข้าที่เป็นตัวอักษร นอกจากนี้ยังทดสอบผลของขนาดความยาวของสายคำที่ลักลอกมาจากต้นฉบับที่มีต่อค่าความแม่นยำในการตรวจจับ ผลการวิจัยพบว่าชุดลักษณ์ที่ให้ผลดีที่สุดในการตรวจหาย่อหน้าที่มีการลักลอกคือชุดลักษณ์ทางสถิติ จำนวน 7 ลักษณ์ จากข้อมูลรับเข้าแบบคำ กล่าวคือ สามารถตรวจจับย่อหน้าที่ลักลอกได้ถูกต้อง 318 ย่อหน้า จาก 735 ย่อหน้า มีค่าความครบถ้วนเท่ากับ 0.43 ส่วนการทดลองกับลักษณ์ทางภาษาที่เปรียบเทียบค่าเฉลี่ยคำที่มีความถี่สูงสุด การเลือกใช้คำ และชุดคำเขียนผิดพบว่า ลักษณ์ประเภทนี้ไม่สามารถแยกประเภทของย่อหน้าที่มีการลักลอกและไม่มีการลักลอก ในประเด็นนี้ ศิวพรให้ความเห็นว่าลักษณ์กลุ่มดังกล่าวมีการกระจายที่ไม่สม่ำเสมอในคลังข้อมูล ส่วนปัจจัยเรื่องความยาวของย่อหน้าลักลอกต่อการตรวจเทียบหาการลักลอกภายในนั้น ผลจากการทดลองยังไม่สามารถระบุถึงความสัมพันธ์ของความยาวย่อหน้าที่มีต่อความแม่นยาในการตรวจจับได้ เพราะย่อหน้าลักลอกที่ตรวจจับได้ถูกต้องมากที่สุดในการทดลองคือย่อหน้าลักลอกขนาดกลางและขนาดยาว ซึ่งมีผลตรวจจับผิดพลาดร้อยละ 16.55 และ 36.67 ตามลำดับ ส่วนย่อหน้าขนาดสั้นนั้นไม่สามารถตรวจจับได้เลย

จากที่กล่าวมาจะเห็นได้ว่าแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนนั้นเหมาะสมที่จะนำมาประยุกต์ใช้ในงานวิจัยชิ้นนี้ เนื่องจากสามารถใช้แยกข้อมูล 2 ประเภทออกจากกันได้ซึ่งในงานจะใช้ข้อความที่มีการลักลอกและไม่มีลักลอกออกจากกัน นอกจากนี้ งานด้านการประมวลภาษาธรรมชาติที่ยกมาในหัวข้อนี้ยังชี้ให้เห็นอีกว่าแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนสามารถใช้แยกประเภทข้อมูลทางภาษาที่มีความเกี่ยวกันในระดับวากยสัมพันธ์และระดับความหมายได้เป็นอย่างดี ประเด็นดังกล่าวนี้ทำให้ผู้วิจัยคาดหวังว่าจะสามารถใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกข้อความที่ถูกลักลอกในลักษณะต่างๆ ได้อย่างมีประสิทธิภาพ

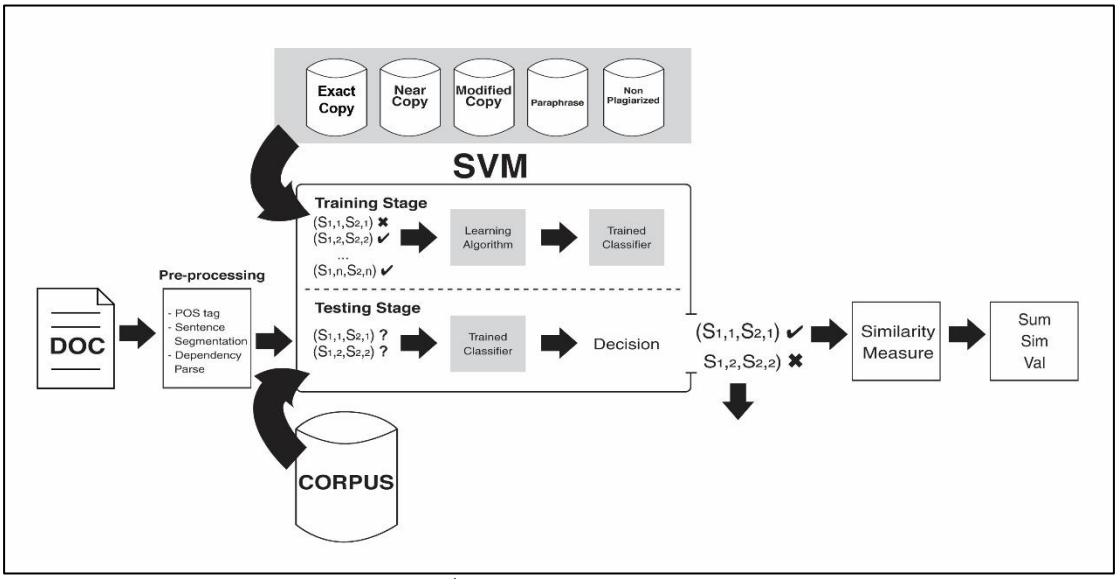
อย่างไรก็ตาม ผลจากงานของศิวพร ทวนไธสง (2556) ซึ่งใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนในการตรวจหาการลักลอกงานวิชาการภาษาไทยนั้นได้ชี้ให้เห็นว่าการออกแบบและสร้างคลังข้อมูลเพื่อใช้สำหรับฝึกฝนและทดสอบประสิทธิภาพของระบบเป็นขั้นตอนหนึ่งที่ส่งผลต่อประสิทธิภาพของระบบเป็นอย่างยิ่ง ดังนั้นในงานวิจัยชิ้นนี้จึงได้เพิ่มขั้นตอนการวิเคราะห์และทดสอบคุณภาพของคลังข้อมูล (corpus analysis & validation) ด้วย เพื่อให้มั่นใจได้ว่าคลังข้อมูลที่สร้างขึ้นจะเอื้อให้ระบบสามารถทำงานได้อย่างมีประสิทธิภาพมากที่สุด ทั้งนี้ รายละเอียดจะได้กล่าวถึงในบทต่อไป

บทที่ 3
วิธีการวิจัย

การดำเนินการพัฒนาระบบตรวจเทียบภายนอกหาการลักลอกงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความนั้นมีขั้นตอนที่เกี่ยวข้องหลายขั้นตอน ไม่ว่าจะเป็นภาพรวมของระบบ การวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทย การสร้างและออกแบบคลังข้อมูล การฝึกฝนและทดสอบระบบ ตลอดจนวิธีประเมินประสิทธิภาพของระบบที่พัฒนาขึ้น ดังนั้นในบทนี้ ผู้วิจัยจะได้กล่าวถึงขั้นตอนการดำเนินการวิจัยดังกล่าวซึ่งมีรายละเอียดตามลำดับต่อไปนี้

3.1 ภาพรวมของระบบ

ระบบตรวจเทียบภายนอกหาการลักลอกงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความที่พัฒนาขึ้นในงานวิจัยชิ้นนี้มีขั้นตอนการประมวลผลหลักอยู่ 2 ขั้นตอน (phase) ได้แก่ การจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก และการวัดค่าความละม้ายระหว่างส่วนของข้อความ ดังจะเห็นได้จากภาพรวมของระบบที่แสดงอยู่ในภาพต่อไปนี้



ภาพที่ 3.1 ภาพรวมของระบบ

จากภาพที่ 3.1 จะเห็นได้ว่ากระบวนการทั้งหมดของระบบเริ่มต้นจากการส่งเอกสารที่ต้องสงสัยว่ามีลักลอกเข้าสู่ระบบ จากนั้นเอกสารจะเข้าสู่ขั้นก่อนการประมวลผล (pre-processing) ได้แก่

การตัดคำ การกำกับหมวดคำ การตัดเอกสารเพื่อให้ได้ส่วนของข้อความ (segment of text) ซึ่งเป็นหน่วยเทียบที่จะเข้าสู่การประมวลผลในขั้นแรกคือการจำแนกประเภทส่วนของข้อความที่มีการลักลอกและไม่มีการลักลอก ในขั้นนี้ หน่วยเทียบที่เป็นข้อมูลรับเข้าจะถูกจับคู่กับหน่วยเทียบที่ค้นคืนได้จากคลังข้อมูลเอกสารต้นฉบับขนาดใหญ่ คู่หน่วยเทียบนี้จะเข้าสู่แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่ได้รับการฝึกฝนแล้วซึ่งทำหน้าที่เป็นเสมือนตัวกรอง (filter) ที่จะจำแนกประเภทคู่ของส่วนของข้อความที่มีการลักลอกและไม่มีการลักลอกออกจากกัน คู่หน่วยเทียบที่ถูกจำแนกประเภทว่าไม่มีการลักลอกจะหลุดออกจากระบบในขั้นนี้ ส่วนคู่หน่วยเทียบที่ถูกจำแนกประเภทว่ามีการลักลอกจะคงอยู่ในระบบต่อไปเพื่อเข้าสู่การประมวลในขั้นที่ 2 ได้แก่การวัดค่าความละม้ายของข้อความ ในขั้นนี้ คู่หน่วยเทียบจะเข้าสู่อัลกอริทึมที่ใช้ในการวัดค่าความละม้ายระหว่างกันและกัน ค่าที่ได้คำนวณได้ออกมาจะเป็นค่าจำนวนจริงที่ทำให้เป็นบรรทัดฐาน (normalized real number value) ตั้งแต่ 0 ถึง 1 ซึ่งระบบจะเก็บค่าความละม้ายของแต่ละคู่หน่วยเทียบเอาไว้จนกระทั่งไม่มีข้อมูลรับเข้าจากเอกสารต้องสงสัยแล้วจึงจะนำค่าความละม้ายดังกล่าวมารวมกันเพื่อระบุว่าเอกสารต้องสงสัยที่รับเข้ามาในระบบนั้นมีการลักลอกในปริมาณมากน้อยเพียงใด จากนั้นระบบจึงจะรายงานผลดังกล่าวออกมาถือเป็นจุดสิ้นสุดของกระบวนการทั้งหมด

อย่างไรก็ดี ผู้วิจัยเห็นว่าในขั้นก่อนการประมวลผลซึ่งเป็นขั้นตอนของการกำกับหมวดคำและตัดเอกสารออกเป็นส่วนนั้นได้มีผู้อื่นศึกษาวิจัยไว้เพียงพอแล้ว ดังนั้นในงานชิ้นนี้จะศึกษาในการประมวลผลเท่านั้น กล่าวคือ จะเริ่มพัฒนาระบบเฉพาะขั้นตอนการจำแนกประเภทการลักลอกด้วยซัพพอร์ตเวกเตอร์แมชชีนไปจนกระทั่งถึงขั้นตอนการวัดค่าความละม้ายของข้อความเท่านั้น ทั้งนี้ข้อมูลที่เข้าสู่ระบบจะเป็นส่วนของข้อความที่ผู้วิจัยได้ตัดไว้เรียบร้อยแล้วจึงไม่ต้องอาศัยกระบวนการการค้นคืนคลังข้อมูลเอกสารต้นฉบับขนาดใหญ่

3.2 การวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทย

ดังได้กล่าวไปในบทที่แล้วว่าในขั้นตอนการสร้างคลังข้อมูลการลักลอก วิเคราะห์หาลักษณ์ที่ใช้สำหรับจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอก กำหนดลักษณะของข้อมูลรับเข้าตลอดจนอภิปรายผลที่ได้จากใช้งานระบบตรวจหาการลักลอกที่พัฒนาขึ้นนั้น จำเป็นอย่างยิ่งที่จะต้องอาศัยองค์ความรู้เกี่ยวกับกลวิธีลักลอกงานวิชาภาษาไทยในเชิงภาษาศาสตร์ ซึ่งยังไม่ปรากฏว่ามีผู้ใดได้ศึกษาไว้ ดังนั้นในงานวิจัยชิ้นนี้จึงจำเป็นต้องวิเคราะห์กลวิธีการลักลอกงานวิชาการในภาษาไทยเพื่อจะได้นำไปประยุกต์ใช้ในขั้นตอนการวิจัยที่ได้กล่าวไปข้างต้น ซึ่งจะส่งผลให้ข้อค้นพบในภาพรวมของงานวิจัยชิ้นนี้มีความหนักแน่นน่าเชื่อถือ

ในหัวข้อนี้จะกล่าวถึงขั้นตอนการวิเคราะห์กลวิธีการลักลอกงานวิชาการภาษาไทย อันประกอบด้วยการสร้างแบบสอบถาม การเก็บรวมรวบข้อมูลจากกลุ่มตัวอย่าง ตลอดจนวิธีวิเคราะห์ข้อมูลขั้นต่างๆ โดยมีรายละเอียดดังต่อไปนี้

3.2.1 การเก็บรวบรวมข้อมูล

ในการเก็บรวบรวมข้อมูลการลักลอกเพื่อนำมาวิเคราะห์นั้น ย่อมเป็นไปได้ยากยิ่งที่ผู้วิจัยจะหาข้อความที่ผ่านการลักลอกที่เกิดขึ้นในสถานการณ์จริงได้ ด้วยเหตุดังกล่าวนี้เอง ผู้วิจัยจึงจำเป็นต้องจำลองการลักลอกขึ้น โดยเก็บข้อมูลการลักลอกจากแบบสอบถาม 2 ชุด แต่ละชุดจะประกอบด้วยย่อหน้าจำนวนที่มีขนาดแตกต่างกัน 5 ย่อหน้า และมีคำสั่งให้ผู้ตอบแบบสอบถามสวมบทบาทสมมติเป็นผู้ลักลอกดังนี้

> *“ในส่วนต่อไปนี้ ผู้วิจัยได้เตรียมข้อความต้นฉบับไว้จำนวน 5 ข้อความ สมมติให้ท่านเป็นนิสิตที่กำลังทำรายงานส่งอาจารย์ ให้ท่านนำข้อความต้นฉบับมาเขียนขึ้นใหม่ให้สื่อความหมายได้ใกล้เคียงกับข้อความต้นฉบับมากที่สุด ทั้งนี้ ท่านต้องระมัดระวังมิให้อาจารย์สงสัยหรือตรวจสอบได้ว่าท่านลอกข้อความต้นฉบับมา”*

ในส่วนที่มาของย่อหน้าต้นฉบับที่ปรากฏในแบบสอบถามนั้น ผู้วิจัยได้คัดมาจากย่อหน้าในวิทยานิพนธ์ของจุฬาลงกรณ์มหาวิทยาลัย 10 เล่ม เพื่อให้เป็นตัวแทนของงานเขียนทางวิชาการในภาษาไทย โดยแบ่งเป็นข้อความจากวิทยานิพนธ์สาขาวิทยาศาสตร์ 5 ย่อหน้า และข้อความจากวิทยานิพนธ์สาขามนุษยศาสตร์และสังคมศาสตร์ 5 ย่อหน้า ทั้งนี้ ในส่วนเนื้อหาของย่อหน้าต้นฉบับผู้วิจัยได้เลือกคัดเฉพาะย่อหน้าที่มีเนื้อหาเป็นกลางซึ่งไม่จำเป็นต้องอาศัยความรู้เฉพาะทางในศาสตร์นั้น ๆ ในการอ่านทำความเข้าใจ ทั้งนี้ก็เพื่อหลีกเลี่ยงปัญหาความลำเอียงอันอาจเกิดขึ้นจากความรู้เฉพาะศาสตร์ที่ผู้ตอบแบบสอบถามแต่ละคนมีอยู่ก่อน จากนั้น ผู้วิจัยจึงนำย่อหน้าทั้ง 2 สาขาวิชามาคละกันเพื่อสร้างเป็นแบบสอบถาม 2 ชุด โดยเรียงลำดับย่อหน้าตามขนาดยาวและสั้นสลับกัน เพื่อเป็นการช่วยลดความเหนื่อยล้าซึ่งอาจเกิดจากการลักลอกย่อหน้าที่มีขนาดยาวติดต่อกันให้แก่ผู้ตอบแบบสอบถาม

ส่วนกลุ่มตัวอย่างนั้น ผู้วิจัยได้เก็บข้อมูลจากผู้ได้รับการศึกษาตั้งแต่ระดับปริญญาตรีขึ้นไป จำนวน 50 คน โดยผู้ตอบแบบสอบถามทั้งหมดจะได้รับแบบสอบถามคนละ 1 ชุด หรือกล่าวคือ ผู้ตอบแบบสอบถาม 1 คนจะต้องลักลอกย่อหน้าทางวิชาการ 5 ย่อหน้าในสถานที่ที่ผู้วิจัยจัดเตรียมไว้ให้ ภายในเวลาที่กำหนดให้ 1 ชั่วโมง ด้วยวิธีนี้ทำให้ผู้วิจัยได้ย่อหน้าที่ผู้ตอบแบบสอบถามเขียนขึ้นทั้งสิ้น 250 ย่อหน้า ซึ่งจะถูกนำไปวิเคราะห์ขั้นตอนต่อไป

3.2.2 วิธีวิเคราะห์ข้อมูล

เมื่อได้ย่อหน้าที่ผู้ตอบแบบสอบถามเขียนเรียบร้อยแล้ว ขั้นตอนแรกที่ผู้วิจัยต้องทำคือการคัดแยกย่อหน้าที่ไม่ถือว่าเป็นการลักลอกออก ทั้งนี้ การคัดแยกจะพิจารณาจากเนื้อหาที่ผู้ตอบแบบสอบถามเขียนมา กล่าวคือ หากย่อหน้าที่เขียนมาไม่สื่อความได้ใกล้เคียงกับเนื้อหาที่ย่อหน้าต้นฉบับมีอยู่เดิมเลยก็จะไม่ถือว่าเป็นการลักลอกและถูกคัดออกไป ดังตัวอย่างต่อไปนี้

(ก)$_{src}$ การเรียนกวดวิชาแม้จะก่อให้เกิดประโยชน์ แต่การเรียนกวดวิชาก็ได้ก่อให้เกิดผลกระทบในด้านลบพอสมควร โดยเฉพาะการสร้างภาระค่าใช้จ่ายแก่ผู้ปกครองอย่างมาก นอกจากการสร้างภาระค่าใช้จ่ายกับผู้ปกครองแล้ว การกวดวิชายังก่อให้เกิดผลกระทบต่อนักเรียน เนื่องจากนักเรียนต้องใช้เวลาว่างไปกับการเรียนกวดวิชา ก่อให้เกิดความห่างเหินระหว่างเด็กกับผู้ปกครองมีผลต่อความอบอุ่นในครอบครัว

(ข)$_{plg}$ ถึงทำให้ไม่มีความอบอุ่นแก่ครอบครัว แต่ทำให้เด็กมีความรู้มีวิชาติดตัว ถึงจะสร้างภาระค่าใช้จ่ายแก่ผู้ปกครองมาก็ตาม แต่ผู้ปกครองก็ยอมให้ลูกได้มีการเรียนการสอนที่ดี

จากตัวอย่างข้างต้น จะเห็นได้ว่าเนื้อหาในย่อหน้าที่ผู้ตอบแบบสอบถามเขียน (ข) ไม่สามารถสื่อความให้ใกล้เคียงกับเนื้อหาที่ย่อหน้าต้นฉบับ (ก) ได้ จึงไม่ถือว่าเป็นย่อหน้าที่ผ่านการลักลอกหรือมีการลักลอกเกิดขึ้น และจะถูกคัดออกไม่นำไปวิเคราะห์ในขั้นต่อไป ในทางตรงกันข้าม ย่อหน้าที่ผู้ตอบแบบสอบถามเขียนขึ้นและสามารถสื่อความได้ใกล้เคียงกับเนื้อหาที่ย่อหน้าต้นฉบับ จะถือว่าเป็นย่อหน้าที่ผ่านการลักลอกหรือมีการลักลอกเกิดขึ้น ย่อหน้าดังกล่าวจะถูกนำไปวิเคราะห์ในขั้นตอนต่อไป

ดังได้กล่าวไปแล้วในหัวข้อที่ 2.8 ว่า หน่วยปริจเฉทพื้นฐานเป็นหน่วยที่เล็กที่สุดในโครงสร้างปริจเฉทที่สามารถสื่อข้อมูลเชิงเนื้อหาและความหมายได้สมบูรณ์ จึงเหมาะสำหรับใช้วิเคราะห์กลวิธีลักลอกที่เกิดขึ้นระหว่างข้อความต้นฉบับกับข้อความที่ผ่านการลักลอก ในขั้นตอนนี้ ผู้วิจัยจึงจะนำย่อหน้าต้นฉบับและย่อหน้าที่ผ่านการลักลอกมาตัดแยกขอบเขตของหน่วยปริจเฉทพื้นฐานตามหลักการที่นลินี อินต๊ะซาว และวิโรจน์ อรุณมานะกุล (Intasaw & Aroonmanakun, 2013) ได้เสนอไว้ โดยเฉพาะในส่วนย่อหน้าต้นฉบับ ผู้วิจัยจะได้ระบุสถานะความสำคัญและชนิดของวาทสัมพันธ์ที่หน่วยปริจเฉทพื้นฐานแต่ละหน่วยในย่อหน้าต้นฉบับมีต่อกันไว้ด้วย โดยยึดหลักที่คาร์ลสันและมาร์คู (Carlson & Marcu, 2001) เสนอไว้ ทั้งนี้ เพื่อให้ง่ายต่อการวิเคราะห์กลวิธีลักลอกที่เกี่ยวข้องกับความสัมพันธ์ในระดับปริจเฉท

หลังจากที่ได้ตัดแยกขอบเขตของหน่วยปริจเฉทพื้นฐานแล้ว ในขั้นต่อมา ผู้วิจัยจะดำเนินการหาคู่ลักลอก (plagiarism pair) อันเป็นหน่วยหลักที่จะใช้ในการวิเคราะห์กลวิธีลักลอกตามวัตถุประสงค์ของงานวิจัยชิ้นนี้ คู่ลักลอกดังกล่าวนี้เกิดจากการจับคู่กันระหว่างหน่วยปริจเฉทพื้นฐานเดิมก่อนเกิดการลักลอกในย่อหน้าต้นฉบับกับหน่วยปริจเฉทพื้นฐานที่มีเนื้อหาเดียวกันแต่ผ่านการลักลอกจากผู้ลักลอก

อย่างไรก็ตาม การดำเนินการหาคู่ลักลอกนั้นจำเป็นต้องมีหลักที่ใช้พิสูจน์เพื่อยืนยันว่าหน่วยปริจเฉทพื้นฐานที่นำมาจับคู่กันเป็นคู่ลักลอกนั้นเป็นต้นฉบับและรูปที่เกิดจากลักลอกของกันและกันจริง ในขั้นตอนนี้เองที่ผู้วิจัยได้นำแนวคิดเรื่องบทบาททางความหมายที่อาร์กิวเมนต์มีต่อภาคแสดงมาประยุกต์ใช้

ทั้งนี้ ผู้วิจัยได้กล่าวในหัวข้อที่ 2.8 แล้วว่า หากพิจารณาในแง่วากยสัมพันธ์แล้ว หน่วยปริจเฉทพื้นฐานก็คืออนุพากย์หรือวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น คุณสมบัติการเป็นอนุพากย์ของหน่วยปริจเฉทพื้นฐานนี้เองที่ช่วยให้ผู้วิจัยสามารถนำบทบาททางความหมายที่คำกริยามีต่ออาร์กิวเมนต์ต่างๆ ภายในอนุพากย์มาเปรียบเทียบกันได้ เพื่อพิสูจน์และช่วยยืนยันว่าคู่ลักลอกของหน่วยปริจเฉทพื้นฐานนั้นเป็นต้นฉบับและรูปที่เกิดจากลักลอกของกันและกันหรือไม่ โดยในที่นี้ ผู้วิจัยได้เลือกเอาชุดบทบาททางความหมายที่พรพิลาส เรืองโชติวิทย์ (2545) ได้เสนอไว้มาประยุกต์ใช้ ดังจะแสดงให้เห็นในตัวอย่างต่อไปนี้

(ก)$_{src}$	เด็ก	มักกลัว	เข็ม
	E		A
(ข)$_{plg}$	เด็ก	อาจจะกลัว	เข็ม
	E		A

จากตัวอย่างคู่ของหน่วยปริจเฉทพื้นฐาน (ก) และ (ข) ข้างต้น จะเห็นได้ว่าหน่วยปริจเฉทพื้นฐาน (ข) ซึ่งถูกลักลอกมาจากหน่วยปริจเฉทพื้นฐาน (ก) มีบทบาททางความหมายตรงกันทุกประการ กล่าวคือ “เด็ก” มีบทบาททางความหมายเป็นผู้ประสบ (E) และ “เข็ม” มีบทบาททางความหมายเป็นผู้ทำ (A) ซึ่งบทบาททางความหมายทั้งสองสัมพันธ์กับคำกริยา “กลัว” นอกจากนี้แล้ว เมื่อพิจารณาความหมายโดยรวมของหน่วยปริจเฉทพื้นฐานทั้งสองก็ยังสื่อความหมายสอดคล้องไปในทำนองเดียวกัน จึงสรุปได้ว่าคู่เทียบหน่วยปริจเฉทพื้นฐาน (ก) และ (ข) เป็นคู่ลักลอกของกันและกัน ซึ่งจะถูกนำไปวิเคราะห์กลวิธีที่ใช้ในการลักลอกในขั้นต่อไป

การประยุกต์ใช้แนวคิดเรื่องบทบาททางความหมายที่กล่าวไปข้างต้น ไม่เพียงแต่จะมีประโยชน์ในการใช้พิสูจน์คู่ลักลอกเท่านั้น แต่ยังเอื้อประโยชน์ในขั้นตอนการวิเคราะห์กลวิธีลักลอกอีก

ด้วย กล่าวคือ แม้โครงสร้างของหน่วยปริจเฉทพื้นฐานจะเปลี่ยนแปลงอันเนื่องมาจากการลักลอก แต่บทบาททางความหมายของอาร์กิวเมนต์ที่ปรากฏในหน่วยปริจเฉทพื้นฐานจะยังคงอยู่เช่นเดิม ลักษณะดังกล่าวนี้ช่วยให้ผู้วิจัยวิเคราะห์กลวิธีลักลอกที่มุ่งเปลี่ยนแปลงโครงสร้างทางวากยสัมพันธ์ได้ชัดเจนขึ้น ดังตัวอย่างต่อไปนี้

(ก)$_{src}$	เด็ก	มักกลัว	เข็ม	
	E		A	
(ข)$_{plg}$	เข็ม	มักทำให้	เด็ก	กลัว
	A		E	

จากตัวอย่างนี้ จะเห็นได้ว่าแม้จะมีการสลับตำแหน่งกันระหว่างประธานกับกรรมในหน่วยปริจเฉทพื้นฐาน (ก) จนอยู่ในรูปหน่วยปริจเฉทพื้นฐาน (ข) แต่บทบาททางความหมายที่ภาคแสดงและอาร์กิวเมนต์มีต่อกันก็ยังคงไม่เปลี่ยนแปลง จึงกล่าวได้ว่าการกำกับบทบาททางความหมายให้กับอาร์กิวเมนต์ในคู่ลักลอกทำให้เห็นได้อย่างชัดเจนว่าคู่ลักลอกข้างต้นใช้การสลับวาจกเป็นกลวิธีในการลักลอก

นอกจากด้านวากยสัมพันธ์แล้ว การกำกับบทบาททางความหมายในคู่ลักลอกยังช่วยให้ผู้วิจัยวิเคราะห์กลวิธีลักลอกทางที่มุ่งเปลี่ยนแปลงคำศัพท์ในข้อความต้นฉบับด้วย ดังตัวอย่างต่อไปนี้

(ก)$_{src}$	มนุษย์	อาศัยอยู่	ในบ้าน
	O		L
(ข)$_{plg}$	มนุษย์	อาศัยอยู่	ในที่พัก
	O		L

จากตัวอย่างคู่ลักลอกข้างต้น จะเห็นว่าผู้ลักลอกตั้งใจแทนที่คำว่า “บ้าน” ในหน่วยปริจเฉทพื้นฐานต้นฉบับ (ก) ด้วยคำว่า “ที่พัก” ตามที่ปรากฏในหน่วยปริจเฉทพื้นฐาน (ข) ทั้งนี้ หากพิจารณาบทบาททางความหมายร่วมด้วยแล้ว จะพบว่าทั้ง “บ้าน” และ “ที่พัก” ต่างก็มีบทบาททางความหมายเป็นบริเวณ (L) ด้วยกันทั้งคู่ ซึ่งบทบาททางความหมายดังกล่าวก็สอดคล้องกับกริยา “อาศัยอยู่” และคำนาม “มนุษย์” ที่มีบทบาททางความหมายเป็นผู้มีสภาพ (O) จึงสรุปได้ว่าในคู่ลักลอกนี้ ผู้ลักลอกใช้กลวิธีแทนที่คำศัพท์ในขั้วเดียวกัน

หลักจากได้วิเคราะห์กลวิธีลักลอกตามกระบวนการที่ได้กล่าวมาข้างต้นแล้ว ในขั้นต่อมาผู้วิจัยจะนับปริมาณการใช้กลวิธีลักลอกแต่ละวิธีที่ถูกใช้โดยผู้ลักลอก เพื่อวิเคราะห์ว่าพฤติกรรมการลักลอกงานวิชาการในภาษาไทยมีลักษณะอย่างไร

นอกจากนี้แล้ว ผู้วิจัยยังจะวิเคราะห์รูปแบบการใช้กลวิธีลักลอก โดยพิจารณาว่าคู่ลักลอกแต่ละคู่มีการใช้กลวิธีลักลอกร่วมกันหรือไม่ หากมีการใช้กลวิธีลักลอกร่วมกัน ผู้ลักลอกใช้กลวิธีลักล อกร่วมกันกี่กลวิธี เป็นกลวิธีใดบ้าง และถูกใช้ร่วมกันในปริมาณเท่าใด

ด้วยวิธีการวิจัยทั้งหมดที่ได้กล่าวมาในหัวข้อนี้ ทำให้ได้ข้อค้นพบที่น่าสนใจและสามารถนำมาประยุกต์ใช้ในขั้นตอนการวิจัยส่วนอื่นๆ ของงานวิจัยชิ้นนี้ได้ ทั้งนี้ จะได้กล่าวถึงรายละเอียดของผลการวิเคราะห์ในบทต่อไป

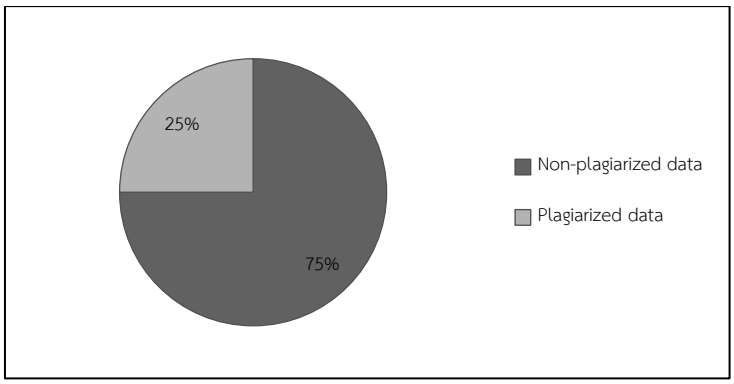
3.3 การออกแบบและสร้างคลังข้อมูล

กล่าวได้ว่าคลังข้อมูลถือเป็นหัวใจสำคัญของงานวิจัยชิ้นนี้ ในแง่การฝึกฝนการจำแนกประเภทด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนนั้นจำเป็นต้องใช้ข้อความที่มีการลักลอกและไม่มีการลักลอกเป็นจำนวนมากในการฝึกฝนให้เครื่องสามารถจำแนกประเภทได้อย่างมีประสิทธิภาพ คลังข้อมูลที่ออกแบบและสร้างขึ้นโดยประกอบด้วยข้อมูลการลักลอกที่หลากหลายก็จะเอื้อให้ระบบทำงานได้อย่างมีประสิทธิภาพมากขึ้น และในแง่ของการทดสอบประสิทธิภาพของระบบที่สร้างขึ้น คลังข้อมูลที่ออกแบบและสร้างตามขั้นตอนและกระบวนที่รัดกุมก็จะช่วยให้ผลการทดสอบประสิทธิภาพของระบบมีความหนักแน่นและน่าเชื่อถือ ในหัวข้อย่อยนี้ ผู้วิจัยจะกล่าวถึงกระบวนออกแบบและการสร้างข้อมูลดังกล่าวเพื่อใช้ในการฝึกฝนและทดสอบประสิทธิภาพของการจำแนกประเภท โดยมีรายละเอียดดังนี้

3.3.1 การออกแบบคลังข้อมูล

ดังได้กล่าวไปในหัวข้อที่ 2.6 แล้วว่าการจะได้มาซึ่งข้อมูลการลักลอกที่เกิดขึ้นในสถานการณ์จริงนั้นมีข้อจำกัดหลายประการ (Potthast et al., 2010, p. 4) ในงานวิจัยชิ้นนี้ ผู้วิจัยจึงได้ออกแบบให้สร้างคลังข้อมูลโดยจำลองการลักลอกงานวิชาการขึ้น และเพื่อให้การฝึกฝนและทดสอบระบบตรวจหาการลักลอกเป็นไปอย่างมีประสิทธิภาพ จึงกำหนดให้คลังข้อมูลมีขนาด 50,000 คู่หน่วยเทียบ ทั้งนี้ คู่หน่วยเทียบในที่นี้คือคู่ของย่อหน้า ย่อหน้าหนึ่งแทนข้อความต้องสงสัยว่าเป็นข้อความต้นฉบับในการลักลอก ส่วนอีกย่อหน้าหนึ่งแทนข้อความต้องสงสัยว่าเป็นการลักลอกจากข้อความต้นฉบับ

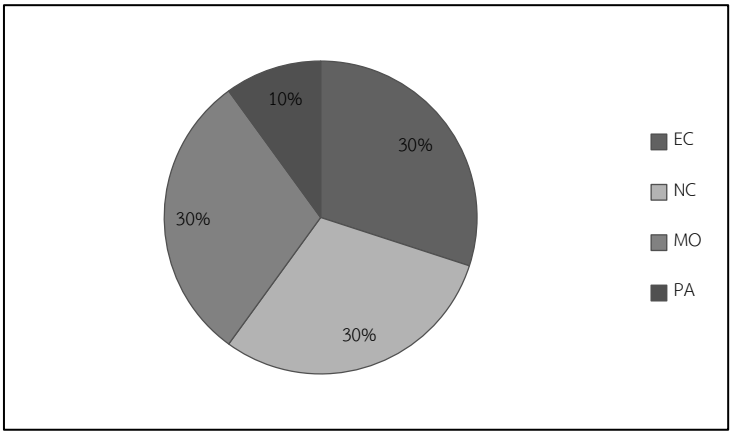
ข้อมูลทั้ง 50,000 คู่หน่วยเทียบข้างต้นแบ่งออกเป็น 2 กลุ่ม ได้แก่ ข้อมูลที่ไม่มีการลักลอก จำนวน 37,500 คู่หน่วยเทียบ คิดเป็นร้อยละ 75 ของคลังข้อมูล และข้อมูลที่มีการลักลอก จำนวน 12,500 คู่หน่วยเทียบ คิดเป็นร้อยละ 25 ของคลังข้อมูล ดังจะเห็นได้จากแผนภูมิวงกลมในภาพที่ 3.2



ภาพที่ 3.2 แผนภูมิแสดงสัดส่วนของข้อมูลในคลังข้อมูล

ในส่วนของข้อมูลที่มีการลักลอก จำนวน 12,500 คู่หน่วยเทียบนั้น ยังแบ่งย่อยออกเป็น 4 ประเภท ได้แก่

1) ข้อมูลที่มีการลักลอกแบบคัดลอกโดยตรง (Exact Copy: EC) จำนวน 3,750 คู่หน่วยเทียบ หรือคิดเป็นร้อยละ 30 ของข้อมูลที่มีการลักลอกทั้งหมด
2) ข้อมูลที่มีการลักลอกแบบคัดลอกโดยใกล้เคียง (Near Copy: NC) จำนวน 3,750 คู่หน่วยเทียบ หรือคิดเป็นร้อยละ 30 ของข้อมูลที่มีการลักลอกทั้งหมด
3) ข้อมูลที่มีการลักลอกแบบคัดลอกโดยดัดแปลง (Modified Copy: MO) จำนวน 3,750 คู่หน่วยเทียบ หรือคิดเป็นร้อยละ 30 ของข้อมูลที่มีการลักลอกทั้งหมด
4) ข้อมูลที่มีการลักลอกแบบถอดความ (paraphrase: PA) จำนวน 1,250 คู่หน่วยเทียบ หรือคิดเป็นร้อยละ 10 ของข้อมูลที่มีการลักลอกทั้งหมด



ภาพที่ 3.3 แผนภูมิแสดงสัดส่วนของข้อมูลที่มีการลักลอกแต่ละประเภท

3.3.2 การเก็บรวบรวมข้อมูล

เนื่องด้วยงานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อพัฒนาระบบตรวจหาการลักลอกในงานวิชาการเป็นหลัก ผู้วิจัยจึงได้เลือกเก็บรวบรวมข้อมูลจากวิทยานิพนธ์ระดับปริญญามหาบัณฑิตและดุษฎีบัณฑิต

ของจุฬาลงกรณ์มหาวิทยาลัยที่ตีพิมพ์เป็นภาษาไทยใน 4 ภาคการศึกษา ได้แก่ ภาคการศึกษาปลาย ปีการศึกษา 2556 จำนวน 1,069 เล่ม, ภาคการศึกษาต้น ปีการศึกษา 2557 จำนวน 258 เล่ม, ภาคการศึกษาปลาย ปีการศึกษา 2557 จำนวน 1,080 เล่ม, และภาคการศึกษาต้น ปีการศึกษา 2558 จำนวน 217 เล่ม รวมจำนวนทั้งสิ้น 2,624 เล่ม โดยได้รับความอนุเคราะห์ข้อมูลส่วนดังกล่าวในรูปแบบไฟล์สกุล .docx จากบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย จากนั้น ผู้วิจัยจะแบ่งข้อมูลวิทยานิพนธ์ทั้งหมดออกเป็น 2 กลุ่ม ตามสาขาวิชาที่เกี่ยวข้องได้ สาขาวิชาวิทยาศาสตร์ (SC) และสาขาวิชามนุษยศาสตร์และสังคมศาสตร์ (HS)

ในขั้นต่อมา ผู้วิจัยได้แปลงวิทยานิพนธ์ในรูปแบบไฟล์สกุล .docx ทั้งหมดให้อยู่ในรูปแบบไฟล์ข้อความเรียบ (plain text) สกุล .txt เข้ารหัสตัวอักษรแบบ UTF-8 โดยใช้โปรแกรม Doxillion Document Converter เวอร์ชัน 2.43 จากนั้นจึงกำกับขอบเขตของคำภาษาไทยในไฟล์ดังกล่าวด้วยโปรแกรม Thai word segmentation เวอร์ชัน 2.2 (Aroonmanakun, 2002)

หลังจากนั้น ผู้วิจัยจะนำไฟล์ข้อความมาคัดเลือกย่อหน้าที่มีขนาดตามที่ได้กำหนดไว้ว่าเป็นขนาดที่เหมาะสมของการเขียนย่อหน้า (จุไรรัตน์ ลักษณะศิริ และบาหยัน อิ่มสำราญ, 2548, น. 194) โดยแบ่งเป็น 3 ขนาด ได้แก่ ย่อหน้าขนาดสั้น (S) มีความยาว 50-100 คำ, ย่อหน้าขนาดกลาง (M) มีความยาว 101-150 คำ และย่อหน้าขนาดยาว (L) มีความยาว 151-200 คำ จากขั้นตอนนี้ทำให้ได้ย่อหน้าข้อมูลดิบสำหรับใช้สร้างคลังข้อมูลจำนวนทั้งสิ้น 489,713 ย่อหน้า ดังมีรายละเอียดแจกแจงไว้ในตารางที่ 3.1 ทั้งนี้ แต่ละย่อหน้าจะถูกแยกบันทึกไว้ในรูปแบบข้อความล้วนสกุล .txt ย่อหน้าละไฟล์

ตารางที่ 3.1 จำนวนของย่อหน้าข้อมูลดิบจำแนกตามสาขาวิชาที่เกี่ยวข้องและขนาด

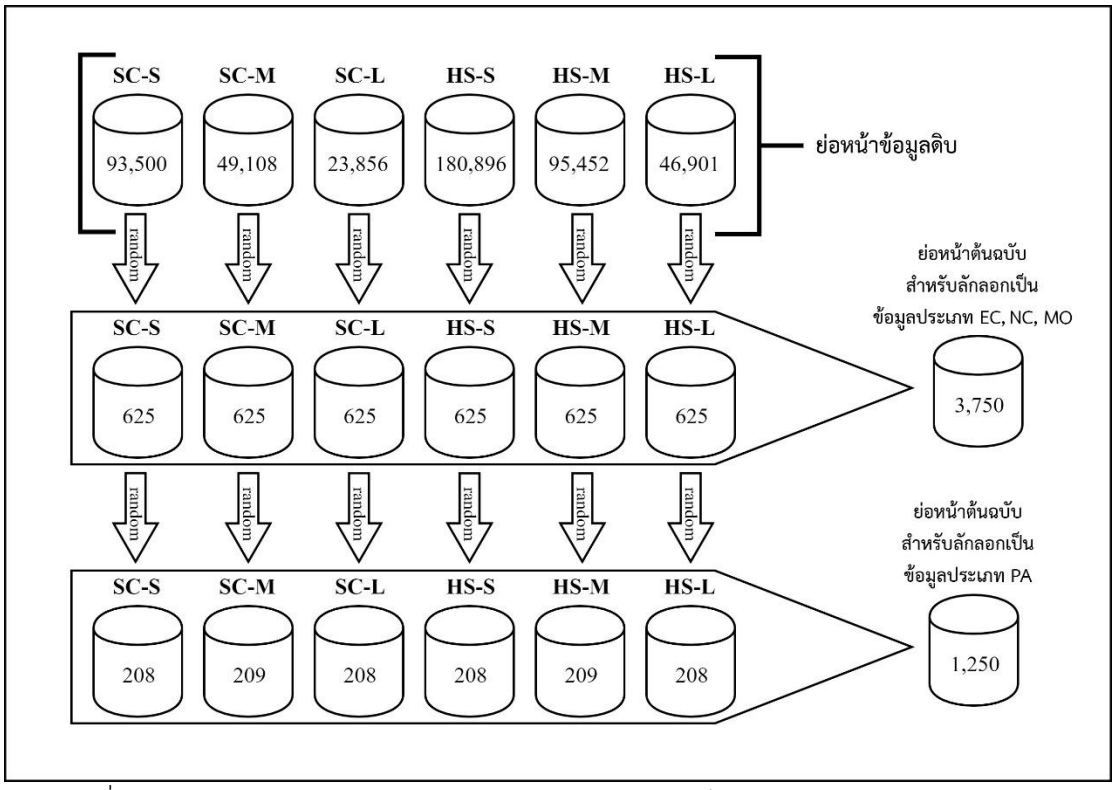
สาขาวิชาและขนาดของย่อหน้าข้อมูลดิบ					
วิทยาศาสตร์			มนุษยศาสตร์และสังคมศาสตร์		
ขนาดสั้น	ขนาดกลาง	ขนาดยาว	ขนาดสั้น	ขนาดกลาง	ขนาดยาว
93,500	49,108	23,856	180,896	95,452	46,901

ย่อหน้าข้อมูลดิบทั้งหมดที่ได้จากการเก็บรวบรวมข้อมูลในขั้นตอนนี้จะถูกนำไปใช้เป็นข้อมูลตั้งต้นสำหรับจำลองข้อมูลการลักลอกและข้อมูลที่ไม่มีการลักลอก ดังจะกล่าวในรายละเอียดในหัวข้อย่อยถัดไป

3.3.3 การจำลองข้อมูลการลักลอก

หลังจากเก็บรวบรวมย่อหน้าข้อมูลดิบได้แล้ว ในขั้นตอนต่อมา ผู้วิจัยจะได้จำลองข้อมูลการลักลอกสำหรับบรรจุในคลังข้อมูลจากย่อหน้าข้อมูลดิบดังกล่าว โดยในขั้นตอนแรก ผู้วิจัยจะสุ่มเลือก

ย่อหน้าต้นฉบับสำหรับใช้ลักลอก ดังปรากฏในภาพที่ 3.4 ย่อหน้าข้อมูลดิบแต่ละสาขาวิชาและขนาดได้รับการเลือกโดยการสุ่มกลุ่มละ 625 ย่อหน้า สำหรับใช้เป็นต้นฉบับในการจำลองข้อมูลการลักลอกประเภทคัดลอกโดยตรง (EC) คัดลอกโดยใกล้เคียง (NC) และคัดลอกโดยดัดแปลง ในขั้นตอนนี้จะทำให้ได้ย่อหน้าต้นฉบับทั้งหมด 3,250 ย่อหน้า จากนั้นจะนำย่อหน้ากลุ่มดังกล่าวมาสุ่มเลือกอีกครั้งตามสาขาวิชาและขนาด กลุ่มละ 208 ย่อหน้าสำหรับย่อหน้าขนาดสั้นและขนาดยาว และกลุ่มละ 209 ย่อหน้าสำหรับย่อหน้าขนาดกลาง ในขั้นนี้จะได้ย่อหน้าต้นฉบับทั้งหมด 1,250 ย่อหน้าสำหรับใช้เป็นต้นฉบับในการจำลองข้อมูลการลักลอกประเภทถอดความ (PA)



ภาพที่ 3.4 กระบวนการสุ่มเลือกย่อหน้าข้อมูลดิบสำหรับใช้เป็นข้อความต้นฉบับในการลักลอกประเภทต่างๆ

หลังจากได้ย่อหน้าต้นฉบับเรียบร้อยแล้ว ผู้วิจัยจะนำย่อหน้าดังกล่าวไปสร้างเป็นข้อมูลการลักลอกประเภทต่างๆ ดังมีรายละเอียดต่อไปนี้

3.3.3.1 การสร้างข้อมูลลักลอกประเภทคัดลอกโดยตรง (Exact Copy: EC)

ข้อมูลการลักลอกประเภทคัดลอกโดยตรงจะสร้างด้วยเครื่องโดยอัตโนมัติ เนื่องจากข้อมูลประเภทนี้ไม่ต้องอาศัยความรู้ทางภาษาศาสตร์ในการสร้าง ทั้งนี้ ในกระบวนการสร้าง ผู้วิจัยได้เขียนโปรแกรมให้คัดลอกข้อความในย่อหน้าต้นฉบับแต่ละข้อความไปบันทึกไว้ในไฟล์ใหม่ ด้วยแนวคิดดังกล่าวนี้ การสร้างข้อมูลการลักลอกประเภทคัดลอกโดยตรงจะเป็นการจำลองการลักลอกแบบ

คัดลอกและวาง (copy & paste) ที่เกิดขึ้นในสถานกาณ์การลักลอกจริง ดังจะเห็นได้จากภาพที่ 3.5 ซึ่งแสดงตัวอย่างของย่อหน้าต้นฉบับสาขาวิชามนุษยศาสตร์และสังคมศาสตร์ ขนาดสั้น และย่อหน้าลักลอกที่จำลองขึ้นตามแนวคิดนี้

ข้อความต้นฉบับ (SR) นอกจากผลการวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้จากการศึกษาภาคสนามดังกล่าวข้างต้น ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้
ข้อความคัดลอกโดยตรง (EC) นอกจากผลการวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้จากการศึกษาภาคสนามดังกล่าวข้างต้น ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

ภาพที่ 3.5 ตัวอย่างข้อความต้นฉบับและข้อความที่ลักลอกด้วยการคัดลอกโดยตรง

จากนั้น ผู้วิจัยจะนำย่อหน้าที่ได้จากการจำลองการลักลอกทั้งหมด 3,750 ย่อหน้าไปจัดเรียงเทียบกับย่อหน้าต้นฉบับให้อยู่ในรูปคู่หน่วยเทียบของไฟล์ และบรรจุเข้าในคลังข้อมูลต่อไป

3.3.3.2 การสร้างข้อมูลลักลอกประเภทคัดลอกโดยใกล้เคียง (Near Copy: NC)

ข้อมูลการลักลอกประเภทคัดลอกโดยใกล้เคียงเป็นตัวแทนของการลักลอกที่เกิดขึ้นในระดับคำ ได้แก่ การแทรกและการลบคำจากข้อความต้นฉบับ ทั้งนี้ ข้อค้นพบจากการวิเคราะห์กลวิธีลักลอกได้ชี้ให้เห็นว่าผู้ลักลอกจะเลือกใช้กลวิธีดังกล่าวเป็นอันดับต้นๆ ในการลักลอก[5] อีกทั้ง การลักลอกประเภทนี้ยังมีลักษณะสอดคล้องกับลำดับชั้นข้อเสนอที่ผู้วิจัยได้กล่าวในการทบวรรณกรรมหัวข้อที่ 2.6 ให้สร้างข้อมูลการลักลอกอิงกับลำดับชั้นของหน่วยทางภาษาเพื่อประโยชน์ในประเมินประสิทธิภาพของลักษณ์ทางภาษาที่ใช้ในการจำแนกข้อความลักลอกและไม่ลักลอก ด้วยเหตุนี้ ผู้วิจัยจึงกำหนดประเภทการลักลอกประเภทนี้ขึ้นเป็นกลุ่มหนึ่งต่างหาก

ในการสร้างข้อมูลการลักลอกประเภทนี้ ผู้วิจัยกำหนดให้สร้างกระบวนการกึ่งมนุษย์และกึ่งเครื่อง กล่าวคือ ในขั้นต้น ผู้วิจัยได้จัดเตรียมกฎสำหรับแก้ไขคำในข้อความต้นฉบับซึ่งอยู่ในรูปแบบ

[5] รายละเอียดจะกล่าวถึงในบทที่ 4

ของรายการคำพร้อมปริบทในการปรากฏของคำดังกล่าวทั้งด้านซ้ายและด้านขวา[6] โดยการแทรกหรือลบคำดังกล่าวในข้อความต้นฉบับจะไม่ส่งผลกระทบให้ความหมายโดยรวมของข้อความเปลี่ยนแปลงไป จากนั้น ผู้วิจัยจึงเขียนโปรแกรมให้เครื่องประยุกต์กฎดังกล่าว โดยกำหนดให้เครื่องค้นหาคำและปริบทด้านซ้ายและขวาในข้อความต้นฉบับที่เตรียมไว้ หากเครื่องพบคำในปริบทด้านซ้ายและขวาตามที่กำหนดไว้รายการก็จะดำเนินการแทรกหรือลบคำดังกล่าวโดยอัตโนมัติ โดยมีเงื่อนไขบังคับว่าทุกย่อหน้าต้องได้การแก้ไขด้วยกฎข้างต้นอย่างน้อย 1 ครั้ง ดังภาพที่ 3.6 ที่แสดงตัวอย่างของย่อหน้าต้นฉบับสาขาวิชามนุษยศาสตร์และสังคมศาสตร์ ขนาดสั้น และย่อหน้าลักลอกมีการแทรกและลบคำตามกฎที่ได้กำหนดไว้

ข้อความต้นฉบับ (SR)

นอกจากผลการวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้จากการศึกษาภาคสนามดังกล่าวข้างต้น ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

ข้อความคัดลอกโดยใกล้เคียง (NC)

นอกจากผล***จาก***(แทรก)การวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่(ลบ)ได้***มา***(แทรก)จากการศึกษาภาคสนามดังกล่าวข้างต้นข้างต้น(ลบ)***แล้ว***(แทรก) ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

ภาพที่ 3.6 ตัวอย่างข้อความต้นฉบับและข้อความที่ลักลอกด้วยการคัดลอกโดยใกล้เคียง

ด้วยกระบวนการดังกล่าวข้างต้นทำให้ได้ข้อมูลการลักลอกประเภทคัดลอกโดยใกล้เคียงทั้งหมด 3,750 ย่อหน้าตามข้อความต้นฉบับที่เตรียมไว้ จากนั้น ผู้วิจัยจะนำย่อหน้าที่ได้จากการจำลองการลักลอกทั้งหมดไปจัดเรียงเทียบกับย่อหน้าต้นฉบับให้อยู่ในรูปคู่หน่วยเทียบของไฟล์ และบรรจุเข้าในคลังข้อมูลต่อไป

3.3.3.3 การสร้างข้อมูลลักลอกประเภทคัดลอกโดยดัดแปลง (Modified Copy: MO)

ข้อมูลการลักลอกประเภทคัดลอกโดยใกล้เคียงเป็นตัวแทนของการลักลอกที่เกิดขึ้นในระดับคำและวากยสัมพันธ์ ทั้งนี้ ผลการวิเคราะห์กลวิธีการลักลอกงานวิชาการภาษาไทยได้ชี้ให้ว่ากลวิธีลักลอกส่วนหนึ่งที่ผู้ลักลอกเลือกใช้นั้นเกิดขึ้นในระดับอนุพากย์ เช่น แทรก ลบ ย้าย แยก หรือรวม

[6] ดูภาคผนวก ข

อนุพากย์ที่มีอยู่ในข้อความต้นฉบับ โดยเฉพาะกลวิธีตัดทอนเนื้อหาในระดับอนุพากย์นั้นมีผู้ลักลอกใช้มากเป็นอันดับที่ 2[7] ด้วยเหตุนี้ ผู้วิจัยจึงได้เสนอจึงกำหนดประเภทการลักลอกประเภทนี้ขึ้นอีกกลุ่มหนึ่งเพื่อเป็นตัวแทนของลำดับชั้นทางภาษาที่สูงขึ้นและซับซ้อนระดับคำในข้อมูลการลักลอกประเภทคัดลอกโดยใกล้เคียง

ข้อมูลประเภทนี้ ผู้วิจัยได้กำหนดให้สร้างโดยผู้จำลองการลักลอกจำนวน 3 คน โดยผู้จำลองการลักลอกต้องมีคุณสมบัติคือ ต้องสำเร็จการศึกษาระดับปริญญาโททางภาษาไทยหรือภาษาศาสตร์และต้องสอนวิชาเกี่ยวกับการเขียนภาษาไทยในระดับมหาวิทยาลัย

ข้อความต้นฉบับ (SR)

นอกจากผลการวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้จากการศึกษาภาคสนามดังกล่าวข้างต้น ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

ข้อความคัดลอกโดยดัดแปลง (MO)

นอกจากผล***จาก***การวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่***ได้มา***จากการศึกษาภาคสนามดังกล่าวข้างต้น***แล้ว***ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน***ที่เกิดขึ้นจากการมีส่วนร่วมของคนในสังคม***ดังตารางที่ 4.2 ต่อไปนี้

ภาพที่ 3.7 ตัวอย่างข้อความต้นฉบับและข้อความที่ลักลอกด้วยการคัดลอกโดยดัดแปลง (การแทรก)

ในส่วนของกระบวนการสร้างข้อมูล ผู้จำลองการลักลอกจะต้องสร้างข้อมูลการลักลอกขึ้นโดยแทรก ลบ หรือย้าย วลีหรืออนุพากย์ในข้อความลักลอกประเภทคัดลอกโดยใกล้เคียง (NC) ข้อความละ 1 ครั้ง โดยผู้วิจัยได้เตรียมคำชี้แจงลักษณะของวลีและอนุพากย์ที่สามารถแก้ไขในการลักลอก[8]และวิธีการกำกับข้อมูลการลักลอกไว้เพื่อให้ผู้จำลองการลักลอกอ่านและทำเข้าใจ ลักษณะและขอบเขตของวลีและอนุพากย์ดังกล่าวนี้ ผู้วิจัยได้กำหนดขึ้นโดยอิงจากลักษณะของหน่วยปริจเฉทพื้นฐานที่ระบุไว้ในงานของนลินี อินต๊ะซาว และวิโรจน์ อรุณมานะกุล (Intasaw & Aroonmanakun, 2013) ขอให้สังเกตภาพที่ 3.7 ที่แสดงการแทรกอนุพากย์ใหม่เข้าไปในข้อความต้นฉบับ จะเห็นได้ว่าอนุพากย์ดังกล่าวเป็นคุณานุประโยคที่ทำหน้าที่ขยาย อันเป็นลักษณะหนึ่งของหน่วยปริจเฉทพื้นฐานตามที่นลินี

[7] ดูรายละเอียดในหัวข้อที่ 4.2

[8] ดูภาคผนวก ค

และวิโรจน์ได้ระบุไว้ ในขณะภาพที่ 3.8 แสดงการลบคุณานุประโยคขยายนามในตอนต้นย่อหน้าทิ้งไป และภาพที่ 3.9 แสดงการย้ายวลี “ดังตารางที่ 4.2 ต่อไปนี้” ที่เดิมปรากฏอยู่ท้ายย่อหน้าไปอยู่ในตำแหน่งต้นย่อหน้าแทน ทั้งนี้ จะสังเกตเห็นได้ว่าวลีดังกล่าวเป็นวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น (strong marker) จึงถือได้ว่าวลีดังกล่าวเป็นหน่วยปริจเฉทพื้นฐาน (Intasaw & Aroonmanakun, 2013, p. 495) ที่สามารถแก้ไขได้ตามข้อกำหนดของการสร้างข้อมูลประเภทนี้

ข้อความต้นฉบับ (SR)

นอกจากผลการวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้จากการศึกษาภาคสนามดังกล่าวข้างต้น ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

ข้อความคัดลอกโดยดัดแปลง (MO)

นอกจากผล***จาก***การวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้~~มา~~จากการศึกษาภาคสนามดังกล่าวข้างต้น~~ข้างต้น~~***แล้ว*** ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

ภาพที่ 3.8 ตัวอย่างข้อความต้นฉบับและข้อความที่ลักลอกด้วยการคัดลอกโดยดัดแปลง (การลบ)

ข้อความต้นฉบับ (SR)

นอกจากผลการวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้จากการศึกษาภาคสนามดังกล่าวข้างต้น ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

ข้อความคัดลอกโดยดัดแปลง (MO)

ดังตารางที่ 4.2 ต่อไปนี้ นอกจากผล***จาก***การวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้~~มา~~จากการศึกษาภาคสนามดังกล่าวข้างต้น ~~ข้างต้น~~***แล้ว*** ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ~~ดังตารางที่ 4.2 ต่อไปนี้~~

ภาพที่ 3.9 ตัวอย่างข้อความต้นฉบับและข้อความที่ลักลอกด้วยการคัดลอกโดยดัดแปลง (การย้าย)

ทั้งนี้ ผู้จำลองการลักลอกจะได้ค่าตอบแทนย่อหน้าละ 5 บาท โดยผู้วิจัยจะตรวจทานย่อหน้าที่สร้างขึ้นก่อนจ่ายค่าตอบแทน และมีสิทธิ์ร้องขอให้ผู้จำลองการลักลอกกลับไปแก้ไขย่อหน้าลักลอกใหม่ในกรณีที่ผู้วิจัยเห็นว่าย่อหน้านั้นๆ มีลักษณะไม่ถูกต้องตามที่ระบุในคำชี้แจง

ด้วยกระบวนการดังกล่าวข้างต้นทำให้ได้ข้อมูลการลักลอกประเภทคัดลอกโดยดัดแปลงทั้งหมด 3,750 ย่อหน้าตามข้อความต้นฉบับที่เตรียมไว้ จากนั้น ผู้วิจัยจะนำย่อหน้าที่ได้จากการจำลองการลักลอกทั้งหมดไปจัดเรียงเทียบกับย่อหน้าต้นฉบับให้อยู่ในรูปคู่หน่วยเทียบของไฟล์ และบรรจุเข้าในคลังข้อมูลต่อไป

3.3.3.4 การสร้างข้อมูลลักลอกประเภทถอดความ (Paraphrase: PA)

ข้อมูลการลักลอกประเภทถอดความเป็นตัวแทนของการลักลอกที่เกิดขึ้นในระดับคำ วากยสัมพันธ์ และความหมาย ผลการวิเคราะห์กลวิธีการลักลอกงานวิชาการภาษาไทยได้ชี้ให้ว่าผู้ลักลอกใช้การแทนที่ด้วยหน่วยทางศัพท์ (lexical unit) หรือหน่วยทางหน้าที่ (functional unit) เพื่อคงความหมายที่ใกล้เคียงกันของต้นฉบับไว้มากเป็นอันดับที่ 4 ของกลวิธีลักลอกทั้งหมดที่วิเคราะห์พบ[9] ลักษณะดังกล่าวนี้ชี้ให้เห็นว่าการแก้ไขเปลี่ยนแปลงข้อความต้นฉบับในเชิงความหมายเป็นกลวิธีลักลอกลำดับแรกๆ ที่ผู้ลักลอกเลือกใช้ นอกจากนี้ เมื่อพิจารณาการปรากฏร่วมกันของกลวิธีลักลอกยังปรากฏว่าการวิธีดังกล่าวยังปรากฏร่วมกับการแก้ไขข้อความต้นฉบับในระดับคำซึ่งได้แก่การ แทรกและการลบคำเป็นอันดับต้นๆ[10] ลักษณะดังกล่าวถือได้ว่าตรงกับนิยามของการถอดความตามที่ได้ทบทวนวรรณกรรมไว้ ในการสร้างข้อมูลสำหรับการฝึกฝนและทดสอบระบบตรวจหาการลักลอกนี้ ผู้วิจัยจึงได้กำหนดให้มีข้อมูลส่วนหนึ่งเป็นข้อมูลการลักลอกประเภทการถอดความเพื่อยังผลให้ระบบที่ได้รับการฝึกฝนด้วยข้อมูลประเภทดังกล่าวสามารถตรวจจับการลักลอกด้วยการถอดความได้อย่างมีประสิทธิภาพ

การสร้างข้อมูลการลักลอกประเภทถอดความ (PA) เป็นไปในลักษณะที่คล้ายคลึงกับการสร้างข้อมูลการลักลอกประเภทคัดลอกโดยดัดแปลง (MO) กล่าวคือ สร้างโดยผู้จำลองการลักลอกจำนวน 3 คน โดยผู้จำลองการลักลอกต้องมีคุณสมบัติเช่นเดียวกันกับผู้จำลองการลักลอกประเภทคัดลอกโดยดัดแปลง ทั้งนี้ ผู้จำลองการลักลอกจะได้รับคำชี้แจงในการลักลอกโดยถอดความและการกำกับข้อมูลการลักลอกที่ผู้วิจัยเตรียมไว้ โดยอาศัยคุณสมบัติของผู้จำลองการลักลอกซึ่งถือเป็นผู้เชี่ยวชาญด้านการเขียนภาษาไทย ในการถอดความข้อความต้นฉบับเพื่อให้ได้ข้อมูลการลักลอก

[9] ดูรายละเอียดในหัวข้อที่ 4.2

[10] ดูรายละเอียดในหัวข้อที่ 4.3

ผู้วิจัยได้เปิดโอกาสให้ผู้จำลองการลักลอกสามารถใช้กลวิธีการเขียนใดๆ ก็ได้ที่ไม่ส่งผลกระทบให้ใจความสำคัญในข้อความต้นฉบับเปลี่ยนแปลงไป อย่างไรก็ดี ข้อมูลการลักลอกทุกชิ้นที่ผู้จำลองการลักลอกถอดความนั้นจะได้รับการตรวจทานโดยผู้วิจัย หากผู้วิจัยเห็นว่าข้อมูลการลักลอกส่วนใดหรือชิ้นใดมีความไม่ถูกต้องเหมาะสมทั้งได้ด้านปริมาณและกลวิธีถอดความ ผู้วิจัยจะร้องขอให้ผู้จำลองการลักลอกกลับไปแก้ไขข้อความส่วนหรือชิ้นดังกล่าวอีกครั้ง ทั้งนี้ ผู้จำลองการลักลอกจะได้รับค่าตอบแทน 20 บาทต่อการลักลอกด้วยการถอดความย่อหน้าต้นฉบับ 1 ย่อหน้า

ภาพที่ 3.10 แสดงตัวอย่างข้อมูลการลักลอกประเภทถอดความ จากภาพดังกล่าวจะเห็นได้ว่าข้อความลักลอกมีกลวิธีการแก้ไขต้นฉบับที่หลากหลาย ไม่ว่าเป็นการแทนที่ด้วยหน่วยทางศัพท์และหน่วยทางหน้าที่เดียวกัน การแทรกคำ การลบคำ และการย้ายตำแหน่งวลี ลักษณะการแก้ไขดังกล่าวนี้เกิดขึ้นโดยผู้จำลองการลักลอกทั้งสิ้น แต่ด้วยเหตุที่การใช้ผู้จำลองการลักลอกในการถอดความในระดับย่อหน้านั้นต้องอาศัยระยะเวลาและกำลังสติปัญญาค่อนข้างมาก ผู้วิจัยจึงได้กำหนดให้ข้อมูลประเภทนี้มีปริมาณเพียงร้อยละ 10 จากจำนวนข้อมูลการลักลอกทั้งหมด

ข้อความต้นฉบับ (SR)

นอกจากผลการวิเคราะห์องค์ประกอบของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ที่ได้จากการศึกษาภาคสนามดังกล่าวข้างต้น ยังมีผลสรุปรวมจากการวิเคราะห์การเสริมสร้างความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ดังตารางที่ 4.2 ต่อไปนี้

ข้อความถอดความ (PA)

ย้ายที่ ไวพจน์ ไวพจน์ แทรก ลบ แทรก ลบ แทรก ไวพจน์ แทรก แทรก

ดังตารางที่ 4.2 ต่อไปนี้ นอกจากผลการวิเคราะห์***ปัจจัยพื้นฐาน***ของทุนทางสังคมในกรณีศึกษาสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน ***ซึ่ง***ได้***มา***จากการศึกษาภาคสนามดังกล่าวข้างต้น ***นอกจากนี้*** ยังมีผลสรุปรวมจากการวิเคราะห์***เกี่ยวกับ***การ***สร้างเสริม***ความเข้มแข็งของทุนทางสังคมในสังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน***อีกด้วย***

ภาพที่ 3.10 ตัวอย่างข้อความต้นฉบับและข้อความที่ลักลอกด้วยการถอดความ

ด้วยกระบวนการดังได้กล่าวไปข้างต้น ย่อหน้าต้นฉบับที่ได้จากขั้นตอนการเก็บรวบรวมข้อมูลทั้งหมด 1,250 ย่อหน้าจะถูกลักลอกโดยการถอดความเป็นข้อมูลการลักลอก จากนั้น ผู้วิจัยจะนำย่อหน้าที่ได้จากการจำลองการลักลอกทั้งหมดไปจัดเรียงเทียบกับย่อหน้าต้นฉบับให้อยู่ในรูปคู่หน่วยเทียบของไฟล์ และบรรจุเข้าในคลังข้อมูลต่อไป

3.3.4 การจำลองข้อมูลที่ไม่มีการลักลอก

ข้อมูลที่ไม่มีการลักลอกเป็นข้อมูลส่วนใหญ่ของคลังข้อมูลที่ออกแบบขึ้นเพื่อใช้ในงานวิจัยชิ้นนี้ ข้อมูลส่วนดังกล่าวนี้สร้างขึ้นเพื่อใช้ฝึกฝนระบบตรวจหาการลักลอกให้เรียนรู้ลักษณะของข้อความที่ไม่มีการลักลอก ดังนั้น ผู้วิจัยจึงได้ออกแบบให้คู่หน่วยเทียบของข้อมูลกลุ่มนี้มีลักษณะเป็นคู่ของย่อหน้า โดยสมมติให้ข้อความหนึ่งเป็นข้อความต้องสงสัยว่าเป็นต้นฉบับที่ใช้ในการลักลอก ส่วนอีกข้อความหนึ่งเป็นข้อความที่ต้องสงสัยว่าลักลอกจากข้อความแรก

เพื่อให้เป็นไปตามแนวคิดดังกล่าวข้างต้น ในการสร้างข้อมูลประเภทนี้ ผู้วิจัยได้ใช้เครื่องสร้างขึ้นโดยอัตโนมัติ โดยเขียนโปรแกรมให้เครื่องจับคู่ย่อหน้าข้อมูลดิบที่ได้จากขั้นตอนการเก็บรวบรวมข้อมูล โดยกำหนดเงื่อนไขว่าย่อหน้าที่จะจับคู่กันได้นั้นต้องมีชุดของคำที่เหมือนกันในระดับหนึ่ง ในการนี้ ผู้วิจัยจึงได้ประยุกต์ใช้ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ (Sørensen-Dice coefficient similarity)[11] เพื่อหาความละม้ายของไตรแกรมของคำระหว่างย่อหน้าซึ่งเป็นข้อมูลดิบทั้งหมด 489,713 ย่อหน้า โดยแบ่งตามสาขาวิชาและขนาดของย่อหน้า หากย่อหน้าคู่ใดมีค่าความละม้ายอยู่ระหว่าง 0.2 กับ 0.5 เครื่องจะคัดคู่ของย่อหน้านั้นมาใช้เป็นข้อมูลส่วนที่ไม่มีการลักลอก ทั้งนี้ เครื่องจะจับคู่ของย่อหน้าลักษณะดังกล่าวนี้จนกระทั่งได้คู่ของย่อหน้าในแต่ละสาขาวิชาและขนาดในปริมาณที่ใกล้เคียงกัน

ข้อความ (ก)

จากตารางข้างต้น สามารถสรุปเป็นผลการเปลี่ยนแปลงความสามารถในการสร้างสัมพันธภาพระหว่างครูกับเด็กปฐมวัยที่เป็นผลจากกระบวนการจัดการเรียนรู้ตามแนวคิดจิตตปัญญาศึกษาและแนวคิดประสบการณ์ที่นำไปสู่ศักยภาพการเรียนรู้สูงสุดเพื่อพัฒนาความสามารถในการสร้างสัมพันธภาพระหว่างครูกับเด็กปฐมวัยดังในตาราง

ข้อความ (ข)

คำถามวิจัยข้อที่ 2 กระบวนการจัดการเรียนรู้ตามแนวคิดจิตตปัญญาศึกษาและแนวคิดประสบการณ์ที่นำไปสู่ศักยภาพการเรียนรู้สูงสุดเพื่อพัฒนาความสามารถในการสร้างสัมพันธภาพระหว่างครูกับเด็กปฐมวัย ส่งผลต่อครูอย่างไร ในด้าน 1) การเคารพ ยอมรับเด็ก 2) ความรัก ความผูกพัน และ 3) การสร้างบรรยากาศที่เอื้อต่อการเรียนรู้

ภาพที่ 3.11 ตัวอย่างข้อมูลที่ไม่มีการลักลอก

[11] รายละเอียดเกี่ยวกับการคำนวณแสดงในหัวข้อที่ 3.3.5

ภาพที่ 3.11 แสดงตัวอย่างข้อมูลที่ไม่มีการลักลอกเป็นคู่ของย่อหน้าที่มีค่าความละม้ายเท่ากับ 0.5 จะสังเกตได้ว่าข้อความ (ก) ซึ่งต้องสงสัยว่าเป็นข้อความต้นฉบับ และข้อความ (ข) ที่ต้องสงสัยว่าเป็นข้อความที่ถูกลักลอกจากข้อความ (ก) มีเนื้อหาใกล้เคียงกัน กล่าวคือ กล่าวถึงการสร้างสัมพันธภาพระหว่างครูกับเด็กปฐมวัยเช่นเดียวกัน ทั้งนี้เพราะอัลกอริทึมของเครื่องได้จับคู่ย่อหน้าดังกล่าวโดยพิจารณาจากค่าความละม้ายของไตรแกรมของคำระหว่างย่อหน้า อย่างไรก็ตาม หากพิจารณาในเนื้อหาแล้ว จะพบว่าย่อหน้าทั้งสองมีใจความสำคัญที่แตกต่างและไม่จัดว่าเป็นการลักลอก

ด้วยกระบวนการที่กล่าวไปข้างต้นทำให้ได้คู่ของย่อหน้าที่ไม่มีการลักลอกครบ 37,500 ย่อหน้าตามที่ได้กำหนดไว้ในขั้นตอนการออกแบบคลังข้อมูล ผู้วิจัยได้จัดคู่ของย่อหน้าดังกล่าวให้อยู่ในลักษณะคู่หน่วยเทียบ จากนั้นจึงบรรจุเข้าในคลังข้อมูล

3.3.5 การวิเคราะห์และตรวจสอบคลังข้อมูล

ดังได้กล่าวไปในตอนท้ายของการทบทวนวรรณกรรมหัวข้อ 2.9 ว่าคลังข้อมูลถือเป็นหัวใจสำคัญของระบบตรวจหาการลักลอกงานวิชาการที่พัฒนาขึ้นในงานวิจัยชิ้นนี้ ด้วยคลังข้อมูลนั้นเป็นปัจจัยสำคัญที่ส่งผลต่อประสิทธิภาพการตรวจหาการลักลอก กล่าวคือ ระบบจะเรียนรู้ลักษณ์ที่สกัดได้จากข้อมูลภายในคลังข้อมูล ในแง่นี้ คลังข้อมูลที่ออกแบบและสร้างขึ้นโดยประกอบด้วยข้อมูลการลักลอกในรูปแบบและลักษณะที่หลากหลายก็จะเอื้อให้ระบบทำงานได้อย่างมีประสิทธิภาพมากขึ้น และอีกแง่มุมหนึ่งคือการทดสอบประสิทธิภาพของระบบนั้น คลังข้อมูลที่ออกแบบและสร้างตามขั้นตอนและกระบวนที่รัดกุมก็จะช่วยให้ผลการทดสอบประสิทธิภาพของระบบมีความหนักแน่นและน่าเชื่อถือ

ในหัวข้อนี้ ผู้วิจัยจะกล่าวถึงวิธีการวิเคราะห์และตรวจสอบคลังข้อมูล รวมถึงผลที่ได้จากการสร้างคลังข้อมูลตามกระบวนออกแบบและสร้างที่ได้กล่าวไว้ในหัวข้อที่ 3.3.1–3.3.4 โดยจะแบ่งเนื้อหาออกเป็น 2 หัวข้อใหญ่ ได้แก่ คุณสมบัติของคลังข้อมูล และการวิเคราะห์และตรวจสอบคลังข้อมูล รายละเอียดมีดังต่อไปนี้

3.3.5.1 คุณสมบัติของคลังข้อมูล (corpus properties)

หลังจากได้สร้างคลังข้อมูลจำลองการลักลอกตามกระบวนออกแบบและสร้างที่ได้กล่าวไว้ในหัวข้อที่ 3.3.1-3.3.4 แล้ว ในหัวข้อนี้ ผู้วิจัยจะได้กล่าวถึงคุณสมบัติที่แสดงถึงลักษณะโดยรวมของคลังข้อมูลที่สร้างขึ้น รายละเอียดมีดังต่อไปนี้

คลังข้อมูลที่สร้างขึ้นประกอบด้วยข้อความทั้งหมด 91,250 ย่อหน้า ตารางที่ 3.2 แสดงจำนวนข้อความทั้งหมดในคลังข้อมูลจำแนกตามประเภทและขนาดของข้อความ จากตารางดังกล่าวจะเห็นได้ว่าคลังข้อมูลประกอบด้วยข้อความ 3 ประเภทหลัก ได้แก่ ประเภทแรก ข้อความที่ใช้เป็นต้นฉบับในการลักลอก มีจำนวน 3,750 ย่อหน้า ข้อความประเภทต่อมาคือข้อความลักลอก ซึ่งเป็น

ข้อความที่เกิดจากการลักลอกข้อความต้นฉบับ (SR) แบ่งออกเป็น 4 ประเภทย่อย ได้แก่ ข้อความลักลอกประเภทคัดลอกโดยตรง (EC) ข้อความลักลอกประเภทคัดลอกโดยใกล้เคียง (NC) ข้อความลักลอกประเภทคัดลอกโดยดัดแปลง (MO) และข้อความลักลอกประเภทถอดความ (PA) ข้อความประเภทนี้มีทั้งหมด 12,500 ย่อหน้า และข้อความประเภทสุดท้ายคือข้อความที่ไม่มีการลักลอก (NA และ NB) ข้อความประเภทนี้ได้มาจากการจับคู่ข้อความที่มีความละม้ายกันในกลุ่มย่อหน้าข้อมูลดิบตามที่กำหนดไว้ในหัวข้อที่ 3.3.4 เป็นข้อความที่มีสัดส่วนมากที่สุดในคลังข้อมูลคือ 75,000 ย่อหน้า

ตารางที่ 3.2 จำนวนข้อความในคลังข้อมูล

ประเภทข้อความ	**ประเภทย่อยของข้อความ**	**จำนวนข้อความ (ย่อหน้า)**		
		ย่อหน้าขนาดสั้น (S)	**ย่อหน้าขนาดกลาง (M)**	**ย่อหน้าขนาดยาว (L)**
ต้นฉบับ	SR	1,250	1,250	1,250
ลักลอก	EC	1,250	1,250	1,250
	NC	1,250	1,250	1,250
	MO	1,250	1,250	1,250
	PA	416	418	416
ไม่ลักลอก	NA	13,444	13,444	10,612
	NB	13,444	13,444	10,612

ตารางที่ 3.3 จำนวนเฉลี่ยของคำในย่อหน้าจำแนกตามขนาดและประเภทของข้อมูล

ประเภทข้อความ	**ประเภทย่อยของข้อความ**	**จำนวนคำเฉลี่ย**		
		ย่อหน้าขนาดสั้น (S)	**ย่อหน้าขนาดกลาง (M)**	**ย่อหน้าขนาดยาว (L)**
ต้นฉบับ	$SR_{EC, NC, MO}$	75.01	122.10	173.18
	SR_{PA}	74.92	122.67	173.21
ลักลอก	EC	75.01	122.10	173.18
	NC	75.78	123.27	174.41
	MO	72.60	118.16	167.68
	PA	73.64	116.98	164.52
ไม่ลักลอก	NA	73.62	118.84	171.45
	NB	74.17	121.95	171.63

เมื่อพิจารณาจากจำนวนคำ คลังข้อมูลนี้จะมีขนาด 11,933,188 คำ (word token) ซึ่งประกอบขึ้นจากแบบชนิดของคำ (word type) จำนวน 51,874 คำ ตารางที่ 3.3 แสดงจำนวนเฉลี่ยของคำในย่อหน้าจำแนกตามขนาดและประเภทของข้อมูล

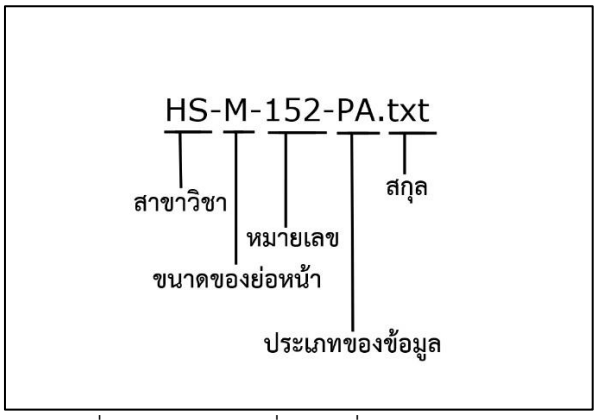
ทั้งนี้ เป็นที่น่าสังเกตว่าย่อหน้าที่มีการลักลอกจะมีขนาดสั้นกว่าย่อหน้าต้นฉบับ ลักษณะดังกล่าวนี้สอดคล้องกับข้อค้นพบของบาร์รอน-เซเดโญและคณะ (Barrón-Cedeño et al., 2013, p. 943) รวมถึงข้อค้นพบของผู้วิจัยเองที่ได้กล่าวไปแล้วในหัวข้อที่ 4.2 ว่าข้อความที่ผ่านการลักลอกย่อมมีแนวโน้มที่จะมีขนาดสั้นกว่าข้อความต้นฉบับ ด้วยเหตุนี้จึงอาจกล่าวได้ว่าข้อมูลที่สร้างขึ้นเพื่อบรรจุในคลังข้อมูลมีลักษณะใกล้เคียงกับการลักลอกที่เกิดขึ้นในสถานการณ์จริง อย่างไรก็ตาม หากพิจารณาจำนวนคำเฉลี่ยของย่อหน้าประเภทลักลอกแบบคัดลอกโดยใกล้เคียง (NC) จะพบว่าย่อหน้าประเภทดังกล่าวมีขนาดยาวกว่าย่อหน้าประเภทต้นฉบับ เกี่ยวกับลักษณะดังกล่าว ผู้วิจัยเห็นว่าเป็นผลมาจากกระบวนการที่ใช้ในการสร้างข้อมูลชนิดนี้ที่กำหนดให้แทรกและลบคำจากข้อความต้นฉบับโดยอิงจากรายการและปริบทด้านซ้ายและขวาของคำที่ได้เตรียมไว้ กล่าวคือ อาจเป็นไปได้ว่ารายการคำดังกล่าวเอื้อให้เกิดการแทรกคำในข้อความต้นฉบับเดิมในปริมาณที่มากกว่ากว่าการลบคำจากข้อความต้นฉบับ

ข้อความในคลังข้อมูลทั้งหมดจะถูกจับคู่ให้อยู่ในรูปของคู่หน่วยเทียบ ในกรณีของข้อมูลลักลอก คู่หน่วยเทียบจะได้จากการจับคู่ของย่อหน้าต้นฉบับที่ใช้ในการลักลอกกับย่อหน้าที่ได้จากการลักลอกย่อหน้าต้นฉบับดังกล่าว ส่วนกรณีของข้อมูลที่ไม่มีการลักลอก คู่หน่วยเทียบจะเป็นได้จากการจับคู่กันของย่อหน้าที่ค่าความละม้ายกันอยู่ระหว่าง 0.2 กับ 0.5 อันเป็นผลมาตั้งแต่การสร้างข้อมูลชนิดนี้ตามที่ระบุไว้ในหัวข้อที่ 3.3.4 คู่หน่วยเทียบประเภทนี้จะเป็นการสมมติให้ย่อหน้าหนึ่งเป็นข้อความต้องสงสัยว่าเป็นข้อความต้นฉบับที่ใช้ในการลักลอก ส่วนอีกย่อหน้าหนึ่งเป็นข้อความต้องสงสัยว่าเกิดจากลักลอกข้อความต้นฉบับดังกล่าว

ทั้งนี้ เมื่อจัดข้อความในคลังข้อมูลทั้งหมดให้อยู่ในรูปคู่หน่วยเทียบแล้ว จะได้คู่หน่วยเทียบทั้งหมด 50,000 คู่ ตามที่ได้กำหนดและอธิบายไว้ในหัวข้อที่ 3.3.1 ดังตารางที่ 3.4 ที่แสดงจำนวนของคู่หน่วยเทียบที่บรรจุในคลังข้อมูลจำแนกตามประเภท สาขาวิชาที่เกี่ยวข้อง และขนาดของย่อหน้า

ตารางที่ 3.4 จำนวนคู่หน่วยเทียบที่บรรจุเข้าในคลังข้อมูล

ประเภทของข้อมูล	ประเภทของคู่หน่วยเทียบ	สาขาวิชาและขนาดของย่อหน้าในคู่หน่วยเทียบ						รวม
		วิทยาศาสตร์			มนุษยศาสตร์และสังคมศาสตร์			
		สั้น	กลาง	ยาว	สั้น	กลาง	ยาว	
ลักลอก	SR-EC	625	625	625	625	625	625	3,750
	SR-NC	625	625	625	625	625	625	3,750
	SR-MO	625	625	625	625	625	625	3,750
	SR-PA	208	209	208	208	209	208	1,250
ไม่ลักลอก	NA-NB	7,194	7,194	4,362	6,250	6,250	6,250	37,500

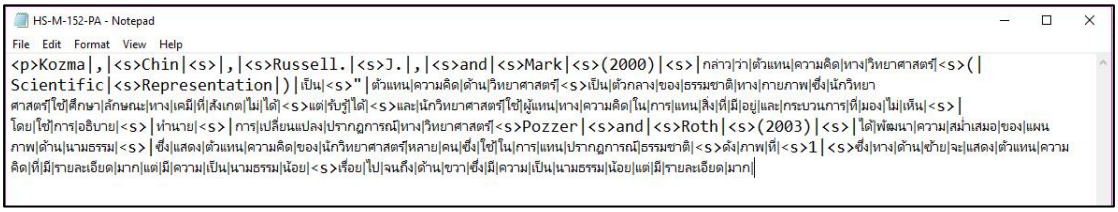


ภาพที่ 3.12 ตัวอย่างชื่อไฟล์ที่บรรจุในคลังข้อมูล

คลังข้อมูลที่สร้างขึ้นจะได้รับบันทึกในรูปข้อความล้วน (plain text) สกุล .txt เข้ารหัสแบบ UTF-8 โดยแต่ละย่อหน้าจะได้รับการแยกบันทึกเป็นหนึ่งไฟล์ ชื่อของไฟล์แต่ละไฟล์จะถูกตั้งให้แสดงเมทาดาทา (metadata) ของย่อหน้านั้นๆ ซึ่งเชื่อมโยงกับไฟล์ซึ่งบันทึกข้อมูลย่อหน้าที่เป็นคู่เทียบ ดังตัวอย่างในภาพที่ 3.12 ซึ่งแสดงการตั้งชื่อไฟล์ข้อความบันทึกย่อหน้าลักลอกประเภทถอดความ จากภาพดังกล่าวจะเห็นได้ว่าอักษร 2 ตัวแรกของชื่อไฟล์จะแสดงสาขาวิชาที่เกี่ยวข้องของเนื้อหาในย่อหน้า ในที่นี้คืออักษร “HS” บ่งชี้ว่ามีเนื้อหาในสาขาวิชามนุษยศาสตร์และสังคมศาสตร์ อักขระตัวที่ 3 แสดงขนาดของย่อหน้า ในที่นี้คืออักษร “M” บ่งชี้ว่าเป็นย่อหน้าขนาดกลาง อักษรตัวที่ 4-6 แสดงลำดับหมายเลขของย่อหน้า ทั้งนี้ เพื่อประโยชน์ในการจัดเป็นคู่หน่วยเทียบ ในที่นี้จึงกำหนดให้ลำดับหมายเลขนี้เป็นหมายเลขเดิมที่ตรงกันกับย่อหน้าต้นฉบับ อักษร 2 ตัวถัดมาแสดงประเภทของข้อมูล

ในที่นี้คืออักษร “PA” บ่งชี้ว่าเป็นข้อมูลลักลอกประเภทถอดความ และอักษร 4 ตัวสุดท้ายของชื่อไฟล์แสดงรูปแบบของไฟล์ข้อความเรียบที่บันทึกในสกุล txt

ส่วนการกำกับข้อมูลในคลังข้อมูลนั้น ในเบื้องต้น ผู้วิจัยได้กำกับขอบเขตของคำในข้อความทุกย่อหน้า ดังตัวอย่างในภาพที่ 3.13



ภาพที่ 3.13 ตัวอย่างการกำกับข้อมูลในคลังข้อมูล

3.3.5.2 ผลการวิเคราะห์และตรวจสอบคลังข้อมูล

ในการตรวจหาการลักลอก แนวคิดหนึ่งที่ถูกใช้อย่างแพร่หลายในการพัฒนาระบบตรวจหาการลักลอกคือการวัดค่าความละม้ายระหว่างข้อความที่ต้องสงสัยว่าเป็นต้นฉบับกับข้อความที่ต้องสงสัยว่าได้มาจากการลักลอก งานวิจัยชิ้นนี้ได้ประยุกต์แนวคิดดังกล่าวมาใช้เป็นเกณฑ์ในการวิเคราะห์คลังข้อมูลโดยอิงตามแนวทางที่เสนอโดยคลอฟและสตีเวนสัน (Clough & Stevenson, 2011, p. 17) ที่ใช้ค่าความละม้ายในการกำหนดและแยกแยะประเภทต่างๆ ของข้อมูลที่รวมเข้าเป็นคลังข้อมูลการลักลอก หากผลการวิเคราะห์ระบุว่าข้อมูลแต่ละประเภทสามารถแยกออกจากกันได้อย่างเด็ดขาด ก็จะสามารถยืนยันได้ว่ากระบวนการที่ใช้ในการสร้างคลังข้อมูลและคุณภาพของข้อมูลในคลังข้อมูลนั้นๆ มีคุณภาพ กล่าวคือ กระบวนการสร้างคลังข้อมูลที่ได้รับการออกแบบขึ้นนั้นสามารถจำลองการลักลอกแต่ละประเภทได้โดยไม่ทับซ้อนกัน

โดยอาศัยแนวคิดข้างต้น งานวิจัยชิ้นนี้ได้เลือกใช้วิธีวัดค่าความละม้าย 2 วิธีในการวิเคราะห์ข้อมูลในระดับที่แตกต่างได้แก่

1) ลำดับย่อยร่วมที่ยาวที่สุด (longest common subsequence: *lcs*)

ลำดับย่อยร่วมที่ยาวที่สุด (longest common subsequence: *lcs*) เป็นแนวคิดพื้นฐานที่ถูกใช้อย่างกว้างขวางในงานด้านการประมวลผลภาษาธรรมชาติ แนวคิดของลำดับย่อยร่วมที่ยาวที่สุดคือการเปรียบเทียบลำดับของสายอักขระ 2 สายจากซ้ายไปขวาแล้วคืนค่าเป็นลำดับของอักขระที่ยาวที่สุดที่สายอักขระทั้งสองสายนั้นมีร่วมกัน

ในส่วนการประยุกต์ใช้เป็นค่าความละม้าย คลอฟและสตีเวนสัน (Clough & Stevenson, 2011, p. 17) ได้เสนอให้ทำลำดับย่อยร่วมที่ยาวที่สุดเป็นค่าบรรทัดฐาน (normalize) โดยนำค่าความยาวของลำดับย่อยร่วมที่ยาวที่สุดของข้อความ 2 ข้อความมาหารด้วยค่าความยาวของข้อความใด

ข้อความหนึ่งจากทั้งสองข้อความ วิธีการคำนวณเช่นนี้ได้ให้ค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุด (normalized *lcs*: lcs_{norm}) ซึ่งมีค่าเป็นตัวเลขอยู่ระหว่าง 0 ถึง 1 หากค่า lcs_{norm} เท่ากับ 0 แปลว่าคู่ของข้อความนั้นๆ ไม่มีความละม้ายกันเลย แต่หากค่า lcs_{norm} เท่ากับ 1 แปลว่าคู่ของข้อความนั้นๆ เหมือนกันทุกประการ

ในแง่ของข้อมูลที่บรรจุในคลังข้อมูลการลักลอกนั้น ค่า lcs_{norm} ที่แตกต่างกันย่อมแสดงถึงระดับของการแก้ไขข้อความต้นฉบับที่แตกต่างกันของคู่หน่วยเทียบแต่ละคู่ โดยอาศัยแนวคิดนี้ ผู้วิจัยจะวัดค่า lcs_{norm} ในระดับอักขระของคู่หน่วยเทียบแต่ละคู่ ทั้งหมด 50,000 คู่ เพื่อเป็นตัวแทนของการแก้ไขข้อความต้นฉบับให้กลายเป็นข้อความที่ต้องสงสัยว่าลักลอกในระดับอักขระ จากนั้น ผู้วิจัยจะนำค่าความละม้ายที่ได้ไปวิเคราะห์เชิงสถิติต่อไป

ตารางที่ 3.5 ค่าสถิติของค่าความละม้ายของคู่หน่วยเทียบในระดับอักขระ

ประเภทของข้อมูล	**ประเภทของคู่หน่วยเทียบ**	**ค่าความละม้าย (lcs_{norm})**	
		ค่าเฉลี่ย	**ส่วนเบี่ยงเบนมาตรฐาน**
ลักลอก	SR-EC	1.0000	0.0000
	SR-NC	0.9788	0.1616
	SR-MO	0.9493	0.4909
	SR-PA	0.8932	0.0741
ไม่ลักลอก	NA-NB	0.5298	0.1068

ตารางที่ 3.5 แสดงค่าสถิติซึ่งเป็นผลจากการวิเคราะห์ค่าความละม้ายระหว่างคู่หน่วยเทียบในคลังข้อมูล เมื่อพิจารณาค่าเฉลี่ยของค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดแล้ว จะเห็นได้ว่าค่าความละม้ายในกลุ่มข้อมูลลักลอกนั้นลดลงเป็นลำดับจากคู่หน่วยเทียบลักลอกประเภทคัดลอกโดยตรง (SR-EC) ไปหาคู่หน่วยเทียบลักลอกประเภทถอดความ (SR-PA) ลักษณะดังกล่าวนี้สอดคล้องกับข้อสรุปจากการทบวรรณกรรมในหัวข้อที่ 2.2 ที่ระบุว่าการลักลอกที่อาศัยการประยุกต์ใช้ความรู้ทางภาษาในระดับต่ำย่อมตรวจหาได้ง่ายกว่า ดังจะเห็นได้จากคู่หน่วยเทียบลักลอกประเภทคัดลอกโดยตรง (SR-EC) ซึ่งเป็นการคัดลอกข้อความจากต้นฉบับโดยไม่มีการเปลี่ยนแปลงมีค่าเฉลี่ยของค่าความละม้ายเท่ากับ 1 ในขณะที่คู่หน่วยเทียบลักลอกประเภทถอดความ (SR-PA) ที่ต้องอาศัยกลไกทางภาษาที่หลากหลายในการลักลอกมีค่าเฉลี่ยเท่ากับ 0.89 จากลักษณะดังกล่าวนี้ หากนำคู่หน่วยเทียบทั้งสองประเภทไปตรวจหาการลักลอกด้วยเครื่อง ย่อมเป็นที่แน่นอนว่าคู่หน่วยเทียบลักลอกประเภทคัดลอกโดยตรง (SR-EC) จะถูกตรวจหาพบได้ง่ายกว่า ในทางตรงกันข้าม คู่หน่วยเทียบ

ประเภทไม่มีการลักลอก (NA-NB) มีค่าเฉลี่ยของค่าความละม้ายเพียง 0.53 เท่านั้น ค่าดังกล่าวแสดงให้เห็นว่า หากพิจารณาในระดับอักขระแล้ว หน่วยเทียบประเภทไม่มีการลักลอก (NA-NB) มีความละม้ายกันประมาณครึ่งหนึ่งเท่านั้น

อย่างไรก็ตาม เป็นที่น่าสังเกตว่าค่าเฉลี่ยของค่าความละม้ายของคู่หน่วยเทียบประเภทลักลอกแต่ละประเภทย่อยนั้นมีความใกล้เคียงกันมาก ลักษณะดังกล่าวนำไปสู่ข้อสังเกตประการหนึ่งว่า ในกรณีการตรวจหาการลักลอกในสถานการณ์จริง การประยุกต์ใช้วิธีการตรวจในระดับอักขระอาจให้ผลไม่เป็นที่น่าพอใจ โดยเฉพาะอย่างยิ่งในกรณีที่ข้อมูลลักลอกมีความคล้ายคลึงกับข้อมูลที่ไม่มีการลักลอก และจากข้อสังเกตดังกล่าวนี้ จึงคาดได้ว่าหากใช้คลังข้อมูลที่สร้างขึ้นเป็นเกณฑ์มาตรฐานในการเปรียบเทียบประสิทธิภาพของระบบตรวจหาการลักลอก ระบบที่ประยุกต์ใช้เพียงวิธีการตรวจหาระดับอักขระในขั้นพื้นฐานอาจไม่สามารถผ่านการประเมินประสิทธิภาพได้ในระดับที่น่าพอใจ อย่างไรก็ดี ลักษณะดังกล่าวก็ถือเป็นข้อยืนยันได้ถึงคุณภาพของคลังข้อมูลที่สร้างขึ้นเพื่อใช้ในงานวิจัยชิ้นนี้ได้เช่นกัน

2) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์

ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ (Sørensen-Dice coefficient similarity) (Dice, 1945; Sørensen, 1948) สถิติที่ใช้ในการเปรียบเทียบความละม้ายของสองตัวอย่าง แนวคิดพื้นฐานของค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์นั้นคล้ายกับค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ด (Jaccard coefficient similarity) ซึ่งถูกใช้อย่างแพร่หลายในงานค้นคืนสารสนเทศ กล่าวคือ สามารถคำนวณได้จากจำนวนสมาชิกที่ทั้งสองตัวอย่างมีร่วมกันต่อจำนวนสมาชิกทั้งหมด แต่หลักการคำนวณค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์จะลดผลกระทบจากการมีสมาชิกร่วมกันของกลุ่มตัวอย่าง ดังนั้นจึงสามารถพิจารณาค่าที่คำนวณได้ในฐานะค่าความละม้ายของข้อมูลทั้งชุด

ในงานชิ้นนี้ได้ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ในการวัดความละม้ายระหว่างข้อความในระดับคำ ทั้งนี้ กำหนดให้เอ็นแกรมของคำมีความยาว n, $S(A, n)$ แทนชุดของเอ็นแกรมของคำในข้อความ A, $S(B, n)$ แทนชุดของเอ็นแกรมของคำในข้อความ B, QS_n ซึ่งแทนผลหารของความละม้าย (quotient of similarity) ระหว่างข้อความ A กับข้อความ B จะหาได้จากสมการที่ 3.1 ค่า QS_n เป็นตัวเลขระหว่าง 0 ถึง 1 เช่นเดียวกันกับค่า lcs_{norm} หากค่า QS_n เท่ากับ 0 แปลว่าข้อความ A และข้อความ B ไม่ปรากฏชุดของเอ็นแกรมร่วมกันเลย แต่หากค่า QS_n เท่ากับ 0 แปลว่าข้อความ A และข้อความ B เหมือนกันทุกประการ

$$QS_n = \frac{2|S(A,n) \cap S(B,n)|}{|S(A,n)| + |S(B,n)|} \tag{3.1}$$

ทั้งนี้ เพื่อให้ครอบคลุมการแก้ไขข้อความต้นฉบับในปริมาณที่มากน้อยแตกต่างกันและเพื่อระบุระดับความละม้ายของการลักลอกประเภทต่างๆ ที่มีบรรจุเข้าในคลังข้อมูลนี้ การวิเคราะห์ในขั้นนี้จะเปรียบเทียบค่าความละม้ายระหว่างคู่ของหน่วยเทียบในคลังข้อมูลทั้งหมด 50,000 คู่ โดยกำหนดค่าเอ็นแกรมของคำไว้ตั้งแต่ 1 ถึง 5 จากนั้น ผู้วิจัยจะนำค่าความละม้ายที่ได้ไปวิเคราะห์เชิงสถิติต่อไป

ตารางที่ 3.6 ค่าสถิติของค่าความละม้ายของคู่หน่วยเทียบในระดับคำ

ประเภทของข้อมูล	**ประเภทของคู่หน่วยเทียบ**	**สถิติ**	**ค่าความละม้าย**				
			QS_1	QS_2	QS_3	QS_4	QS_5
ลักลอก	SR-EC	Mean	1.0000	1.0000	1.0000	1.0000	1.0000
		SD	0.0000	0.0000	0.0000	0.0000	0.0000
	SR-NC	Mean	0.9660	0.9259	0.8871	0.8500	0.8145
		SD	0.2108	0.0416	0.0620	0.0890	0.0162
	SR-MO	Mean	0.9220	0.8731	0.8272	0.7838	0.7425
		SD	0.0564	0.0658	0.0787	0.0926	0.1065
	SR-PA	Mean	0.8808	0.7691	0.6839	0.6124	0.5510
		SD	0.0727	0.1225	0.1554	0.1784	0.1948
ไม่ลักลอก	NA-NB	Mean	0.5598	0.3830	0.3025	0.2458	0.2029
		SD	0.0884	0.0870	0.0809	0.0811	0.0834

ตารางที่ 3.6 แสดงค่าสถิติของความละม้ายที่วัดได้เปรียบเทียบกันระหว่างข้อมูลประเภทต่างๆ ภายในคลังข้อมูล จากตารางดังกล่าวจะได้เห็นว่าผลการวิเคราะห์ค่าความละม้ายระหว่างข้อมูลประเภทต่างๆ ในคลังข้อมูลมีแนวโน้มเป็นเช่นเดียวกันกับผลการวิเคราะห์ค่าความละม้ายในระดับอักขระ กล่าวคือ ระดับความละม้ายมีค่าลดลงเมื่อระดับของการลักลอกสูงขึ้น ลักษณะดังกล่าวนี้ชี้ให้เห็นว่าวิธีการลักลอกในระดับที่สูงกว่ามีแนวโน้มที่จะแบ่งแยกลำดับของคำที่ปรากฏในต้นฉบับเดิมมากกว่าวิธีการลักลอกในระดับที่ต่ำกว่า ในขณะเดียวกัน ค่าความละม้ายของข้อมูลลักลอกก็มีความแตกต่างอย่างเห็นได้ชัดจากค่าความละม้ายของข้อมูลที่ไม่มีการลักลอก

นอกจากนี้ยังพบว่าเมื่อขนาดของเอ็นแกรมของคำเพิ่มขึ้น ค่าความละม้ายจะลดต่ำลง ดังจะเห็นได้จากค่าความละม้ายในกลุ่มข้อมูลลักลอกว่าค่าเฉลี่ยของค่าความละม้ายของ 5 แกรมของคำ (QS_5) มีค่าต่ำกว่าค่าเฉลี่ยของค่าความละม้ายของ 4 แกรมของคำ (QS_4) ค่าเฉลี่ยของค่าความละม้าย

ของไตรแกรมของคำ (QS_3) ค่าเฉลี่ยของค่าความละม้ายของไบแกรมของคำ (QS_2) และค่าเฉลี่ยของค่าความละม้ายของยูนิแกรมของคำ (QS_1) เป็นลำดับลดหลั่นกันไป ลักษณะเช่นนี้นำไปสู่ข้อสังเกตว่าในกรณีที่คลังข้อมูลที่สร้างขึ้นเป็นเกณฑ์มาตรฐานในการเปรียบเทียบประสิทธิภาพของระบบตรวจหาการลักลอก ระบบตรวจหาการลักลอกที่พัฒนาขึ้นโดยประยุกต์ใช้องค์ความรู้ทางภาษาศาสตร์ที่หลากหลายและซับซ้อนกว่าย่อมให้ผลการตรวจหาการลักลอกที่น่าพอใจกว่าระบบตรวจหาการลักลอกที่พัฒนาขึ้นโดยประยุกต์ใช้วิธีการตรวจหาขั้นพื้นฐานวิธีใดเพียงวิธีเดียว

เมื่อคำนวณได้ค่าความละม้ายทั้งในระดับอักขระและระดับคำมาแล้ว ขั้นตอนต่อมา ผู้วิจัยจะนำค่าที่ได้มาวิเคราะห์เชิงสถิติ โดยผู้วิจัยจะนำค่าความละม้ายของข้อมูลในคลังข้อมูลแต่ละประเภทข้างต้นมาเปรียบเทียบกัน เพื่อทดสอบว่าข้อมูลแต่ละกลุ่มแตกต่างกันอย่างมีนัยสำคัญทางสถิติหรือไม่ ทั้งนี้ แม้ผลการวิเคราะห์ค่าความละม้ายในระดับอักขระและระดับคำจะแสดงให้เห็นความแตกต่างของข้อมูลแต่ละประเภทในคลังข้อมูล แต่ผลดังกล่าวก็ยังไม่สามารถยืนยันได้ว่าข้อมูลแต่ละประเภทที่บรรจุเข้าในคลังข้อมูลจะไม่ได้มาจากกลุ่มประชากรเดียวกัน ทั้งนี้ หากข้อมูลภายในคลังข้อมูลมาจากกลุ่มประชากรเดียวกันก็อาจส่งผลให้การทดสอบประสิทธิภาพของระบบตรวจหาการลักลอกผิดพลาดได้ เนื่องจากระบบจะไม่สามารถจำแนกประเภทข้อความที่ลักลอกและไม่มีการลักลอกออกจากกันได้อย่างเด็ดขาด อันเป็นผลมาจากการที่ระบบเรียนรู้ข้อมูลที่ไม่มีความแตกต่างกันนั่นเอง ด้วยเหตุนี้ การวิเคราะห์คลังข้อมูลในขั้นตอนสุดท้ายจึงเป็นไปเพื่อพิสูจน์ว่าข้อมูลแต่ละประเภทที่บรรจุเข้าในคลังข้อมูลมีความแตกต่างกันทางสถิติอย่างมีนัยสำคัญหรือไม่

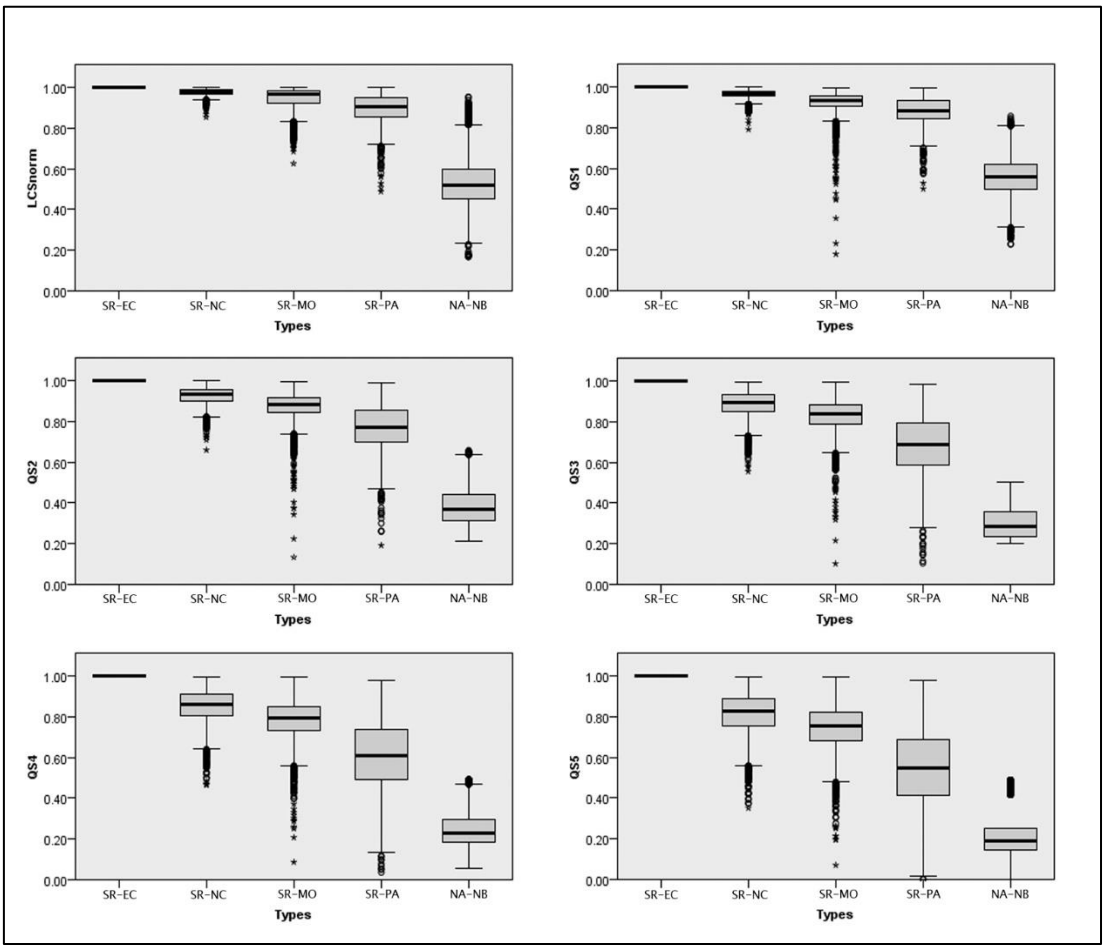
โดยทั่วไปแล้ว การทดสอบว่ากลุ่มตัวอย่าง 2 กลุ่มหรือมากกว่ามีความแตกต่างกันหรือไม่นั้นสามารถทำได้โดยเปรียบเทียบค่าเฉลี่ยของกลุ่มตัวอย่างแต่ละกลุ่ม หรือที่รู้จักกันในฐานะการวิเคราะห์ความแปรปรวนทางเดียว (one-way analysis of variance: one-way ANOVA) ซึ่งอาจใช้ค่าเฉลี่ยของค่าความละม้ายในการทดสอบได้ อย่างไรก็ตาม การใช้ค่าเฉลี่ยของค่าความละม้ายที่ได้ข้อมูลในคลังข้อมูลนี้ถือว่าเป็นการฝ่าฝืนข้อตกลงเบื้องต้นของการวิเคราะห์ความแปรปรวนทางเดียว เนื่องจากกลุ่มตัวอย่างที่ใช้ในการทดสอบมีความแปรปรวนไม่เท่ากัน ลักษณะดังกล่าวค่าความละม้ายของสมาชิกทุกตัวในกลุ่มข้อมูลลักลอกประเภทคัดลอกโดยตรง (SR-EC) มีค่าเท่ากับ 1 เสมอ ดังนั้นในกรณีนี้จึงไม่สามารถใช้การวิเคราะห์ความแปรปรวนทางเดียวในการทดสอบความแตกต่างของข้อมูลแต่ละประเภทในคลังข้อมูลได้

ด้วยเหตุดังกล่าวข้างต้น ในขั้นนี้ ผู้วิจัยจึงได้เลือกใช้การทดสอบครัสคาล-วอลลิส (Kruskal-Wallis test) (Kruskal & Wallis, 1952) แทน การทดสอบครัสคาล-วอลลิส หรือที่รู้จักกันในอีกชื่อว่าการวิเคราะห์ความแปรปรวนทางเดียวตามลำดับ (one-way ANOVA on ranks) เป็นการทดสอบทางสถิติแบบนอนพาราเมตริกเพื่อเปรียบเทียบกลุ่มตัวอย่างอิสระ 2 กลุ่มหรือมากกว่าที่มีขนาดของ

กลุ่มตัวอย่างเท่ากันหรือต่างกัน แนวคิดพื้นฐานของสถิติดังกล่าวคือทดสอบว่าการทดสอบว่ากลุ่มตัวอย่าง k กลุ่มมีค่ามัธยฐานเท่ากันหรือไม่ โดยอาศัยแนวคิดดังกล่าวนี้ การทดสอบครัสคาล-วอลลิสจึงมักถูกใช้เมื่อไม่สามารถทดสอบด้วยสถิติแบบพาราเมตริกได้

ด้วยวิธีการทดสอบทางสถิติดังได้กล่าวไปข้างต้น ผู้วิจัยได้นำค่าความละม้ายทั้ง 6 ค่าที่คำนวณได้จากคู่หน่วยเทียบแต่ละประเภท (กลุ่มตัวอย่าง) ได้แก่ ค่าความละม้ายจากค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุด (lcs_{norm}) ค่าความละม้ายของยูนิแกรมของคำ (QS_1) ค่าความละม้ายของไบแกรมของคำ (QS_2) ค่าความละม้ายของไตรแกรมของคำ (QS_3) ค่าความละม้ายของ 4 แกรมของคำ (QS_4) และความละม้ายของ 5 แกรมของคำ (QS_5) มาทดสอบแยกต่างหากทีละค่า เพื่อเปรียบเทียบว่าค่าความละม้ายดังกล่าวมีความแตกต่างกันระหว่างประเภทของข้อมูลหรือไม่

ผลการทดสอบแสดงให้เห็นว่าค่าความละม้ายทั้ง 6 ค่ามีความแตกต่างกันระหว่างประเภทข้อมูลอย่างมีนัยสำคัญ (ทดสอบหลังการวิเคราะห์ด้วยวิธี Bonferroni, $p < 0.05$) ภาพที่ 3.14 แสดงการแจกแจงของข้อมูลในคลังข้อมูลที่ระดับค่าความละม้ายต่างๆ



ภาพที่ 3.14 การแจงแจกของข้อมูลในคลังข้อมูลที่ค่าความละม้ายระดับต่างๆ

ผลการวิเคราะห์คลังข้อมูลดังได้กล่าวมาทั้งหมดสามารถเป็นเครื่องยืนยันได้ว่าวิธีการที่ใช้ในออกแบบและสร้างคลังข้อมูลที่ผู้วิจัยได้เสนอเพื่อใช้ในงานวิจัยชิ้นนี้เป็นไปอย่างรัดกุมและมีประสิทธิภาพ กล่าวคือ วิธีการดังกล่าวสามารถจำลองข้อมูลที่มีลักลอกและข้อมูลที่ไม่มีการลักลอกออกมาได้อย่างแตกต่างกันอย่างมีนัยสำคัญทางสถิติ ด้วยเหตุนี้ คลังข้อมูลที่สร้างขึ้นเพื่อใช้ในงานวิจัยชิ้นนี้จึงสามารถใช้ฝึกฝนและทดสอบระบบตรวจหาการลักลอกได้อย่างมีประสิทธิภาพ

3.4 การวิเคราะห์หาลักษณ์ทางภาษาสำหรับจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอก

หลังจากสร้างคลังข้อมูลการลักลอกเสร็จเรียบร้อยแล้ว ในขั้นตอนการวิเคราะห์หาลักษณ์ทางภาษาสำหรับจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกนี้ ผู้วิจัยได้ประยุกต์ความรู้ที่ได้จากการทบทวนวรรณกรรมและการวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทยมาใช้ในการวิเคราะห์หาลักษณ์ โดยจะแบ่งกลุ่มลักษณ์ทางภาษาออกเป็น 3 กลุ่ม ดังต่อไปนี้

3.4.1 การวิเคราะห์หาลักษณ์ทางศัพท์

ลักษณ์ทางศัพท์จัดเป็นลักษณ์ทางภาษาที่วิเคราะห์จากข้อความในระดับคำ ทั้งนี้ จากการทบทวนวรรณกรรมหัวข้อที่ 2.7.2 ว่าด้วยการจำแนกประเภทข้อความที่มีความละม้ายกันโดยใช้การเรียนรู้ของเครื่อง พบว่ามีการจำนวนไม่น้อยใช้ลักษณ์ทางศัพท์ในการจำแนกประเภทข้อความที่มีความละม้ายกันแล้วได้ผลเป็นที่น่าพอใจ อีกทั้งข้อค้นพบที่ได้จากการวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทยในหัวข้อที่ 4.2 ก็ชี้ให้ว่าผู้ลักลอกมีแนวโน้มจะแก้ไขข้อความต้นฉบับในระดับคำเป็นอันดับต้นๆ ดังนั้นในงานวิจัยชิ้นนี้จึงจะวิเคราะห์หาลักษณ์ทางศัพท์เพื่อนำมาใช้ในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอก

การวิเคราะห์หาลักษณ์ทางศัพท์จำเป็นต้องพิจารณาขอบเขตของคำในภาษาไทย ในงานวิจัยชิ้นนี้ จึงได้กำกับข้อมูลขอบเขตของคำในคลังข้อมูลไว้ตั้งแต่ต้นเพื่อให้สะดวกต่อการวิเคราะห์หาและสกัดลักษณ์ ทั้งนี้ หากพิจารณาจากการทบทวนวรรณกรรมและผลการวิเคราะห์กลวิธีลักลอกแล้ว ลักษณ์ทางศัพท์สามารถวิเคราะห์หาได้โดยประยุกต์แนวคิดที่หลากหลายซึ่งสามารถแสดงวิธีวิเคราะห์หาลักษณ์จำแนกตามกลุ่มแนวคิดได้ดังต่อไปนี้

1) ลำดับย่อยร่วมที่ยาวที่สุด (longest common subsequence: *lcs*)

ลักษณ์ชนิดนี้วิเคราะห์หาโดยพิจารณาลำดับคำที่เหมือนกันที่ปรากฏในข้อความในคู่หน่วยเทียบ ทั้งนี้ งานวิจัยชิ้นนี้ ผู้วิจัยจะเขียนโปรแกรมเพื่อหาลำดับย่อยร่วมที่ยาวที่สุดในชุดคำของคู่หน่วย

เทียบแล้วทำให้เป็นค่าบรรทัดฐานซึ่งจะคืนค่าออกมาเป็นตัวเลข ค่าดังกล่าวจะแสดงให้เห็นได้ว่ามีความละม้ายกันระหว่างประโยคมากน้อยเท่าใด

2) จำนวนคำ (word count)

ลักษณ์ชนิดนี้วิเคราะห์หาการนับจำนวนคำในข้อความ ดังนั้นในขั้นต้นของการวิเคราะห์หาจึงจำเป็นกำกับขอบเขตของคำ จากนั้น ผู้วิจัยจะเขียนโปรแกรมสำหรับนับจำนวนคำในข้อความต้นฉบับและข้อความต้องสงสัยว่าลักลอก จากนั้นจึงจะนำค่าที่ได้ออกมาใช้เป็นลักษณ์

3) เอ็นแกรมของคำ (word *n*-gram)

เอ็นแกรมเป็นแนวคิดที่ถูกใช้อย่างแพร่หลายในการประมวลภาษาธรรมชาติ และถูกประยุกต์ใช้ในการแทนรูปภาษาในหลายระดับ ไม่ว่าจะเป็นระดับอักขระ คำ หรือวายสัมพันธ์ ในส่วนที่เกี่ยวข้องกับหัวข้อนี้คือการแทนรูปเอ็นแกรมในระดับคำ เอ็นแกรมหนึ่งๆ นั้นคือเอ็นโทเค็น (*n*-token) ชุดของคำ (Jurafsky & Martin, 2009, p. 83) ยกตัวอย่างเช่น ในข้อความ “This is an example.” ไบแกรมของข้อความดังกล่าวจะเป็นชุดของคำชุดละ 2 คำ ได้แก่ [“This”, “is”], [“is”, “an”], [“an”, “example”], และ [“example”, “.”] แต่หากพิจารณาเป็นไตรแกรม ไตรแกรมของข้อความข้างต้นจะได้เป็นชุดของคำชุดละ 3 คำ ได้แก่ [“This”, “is”, “an”], [“is”, “an”, “example”], [“an”, “example”, “.”]

แนวคิดเรื่องเอ็นแกรมดังได้กล่าวไปข้างต้นนี้ยังสามารถประยุกต์ในการวิเคราะห์หาลักษณ์ได้อีกหลายรูปแบบ เช่น ประยุกต์ใช้กับการให้น้ำหนักคำแบบ *tf-idf* หรือประยุกต์ใช้กับการวัดค่าความละม้ายในระดับเอ็นแกรมแล้วนำค่าความละม้ายดังกล่าวมาใช้เป็นลักษณ์

อย่างไรก็ตาม เนื่องด้วยข้อจำกัดของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่ไม่สามารถรับเข้าลักษณ์ที่ไม่เป็นตัวเลขได้ ดังนั้นในหาวิเคราะห์ลักษณะด้วยแนวคิดเอ็นแกรมนี้จึงจำเป็นต้องแปลงเอ็นแกรมที่วิเคราะห์ได้ให้เป็นเวกเตอร์ก่อน จากนั้นจึงนำไปดำเนินการต่อ

4) ค่าความละม้ายที่วัดได้ในระดับคำ

จากการทบทวนวรรณกรรมหัวข้อที่ 2.7.2 จะเห็นได้ว่ามีงานจำนวนหนึ่งที่นำค่าความละม้ายที่วัดได้ในระดับคำมาใช้เป็นลักษณ์ในการจำแนกประเภทของข้อความที่มีความละม้ายกัน ในงานวิจัยชิ้นนี้ ผู้วิจัยจึงได้ประยุกต์ใช้แนวคิดดังกล่าว โดยนำค่าความละม้ายที่คำนวณได้ในระดับคำมาใช้เป็นลักษณ์ในการจำแนกข้อความลักลอกและข้อความที่ไม่มีการลักลอก ในการนี้ ผู้วิจัยจะเขียนโปรแกรมสำหรับวัดค่าความละม้ายในระดับคำด้วยวิธีการวัดต่างๆ ขึ้น จากนั้นจึงนำค่าตัวเลขที่แสดงความละม้ายระหว่างคู่หน่วยเทียบมาใช้เป็นลักษณ์

เมื่อวิเคราะห์หาลักษณ์ทางศัพท์ได้จากวิธีการดังได้กล่าวมาข้างต้นแล้ว ผู้วิจัยจะเขียนโปรแกรมสกัดลักษณ์แต่ละตัวจากข้อมูลทั้งหมดในคลังข้อมูล และนำข้อมูลดังกล่าวไปฝึกฝนและทดสอบระบบตรวจการลักลอกต่อไป

3.4.2 การวิเคราะห์หาลักษณ์ทางวากยสัมพันธ์

ลักษณ์ทางวากยสัมพันธ์เป็นลักษณ์ทางภาษาที่วิเคราะห์จากข้อความในระดับวลีและประโยค จากการทบทวนวรรณกรรมหัวข้อที่ 2.7.2 พบว่ามีงานวิจัยจำนวนหนึ่งที่นำความสัมพันธ์ภายในประโยคมาประยุกต์ใช้เป็นลักษณ์ในการจำแนกประเภทข้อความที่มีความละม้ายกัน อีกทั้งผลจากการวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทยในหัวข้อ 4.2 ได้ชี้ให้เห็นว่ากลวิธีลักลอกกลุ่มหนึ่งที่ผู้ลักลอกเลือกใช้เป็นการแก้ไขข้อความในระดับวากยสัมพันธ์ เช่น การแยกอนุพากย์ การซ้อนความ การรวมความ ด้วยเหตุนี้ ในงานวิจัยชิ้นนี้จึงจะวิเคราะห์หาลักษณ์ทางวากยสัมพันธ์เพื่อนำทดสอบประสิทธิภาพการจำแนกประเภทข้อความลักลอกที่มีการแก้ไขในระดับวากยสัมพันธ์

จากการทบทวนวรรณกรรมทำให้เห็นว่าลักษณ์ทางวากยสัมพันธ์สามารถวิเคราะห์หาได้โดยประยุกต์แนวคิดที่หลากหลายซึ่งสามารถแสดงวิธีวิเคราะห์หาลักษณ์จำแนกตามกลุ่มแนวคิดได้ดังต่อไปนี้

1) หมวดคำ (Part of Speech: POS)

กล่าวได้ว่าหมวดคำมีความสำคัญยิ่งในงานด้านการประมวลผลภาษา เนื่องด้วยหมวดคำนั้นให้ข้อมูลเกี่ยวกับคำและปริบทแวดล้อมคำได้เป็นอย่างมาก (Jurafsky & Martin, 2009, p. 123) ทั้งนี้เพราะข้อมูลที่อยู่เบื้องหลังหมวดคำสามารถแสดงถึงหน้าที่เชิงวากยสัมพันธ์ของคำ การปรากฏของคำในประโยค และคุณสมบัติภายในของคำนั้นๆ ส่วนในแง่การลักลอก การแก้ไขโดยเปลี่ยนคำในต้นฉบับไปใช้คำในหมวดคำเดียวกันยังเป็นกลวิธีหนึ่งที่ผู้ลักลอกเลือกใช้ด้วย ในงานวิจัยชิ้นนี้ ผู้วิจัยจึงจะกำกับหมวดคำให้แก่ข้อมูลทั้งหมดภายในคลังข้อมูล จากนั้นจึงนำมาชุดของหมวดคำที่ได้จากคู่หน่วยเทียบไปดำเนินการแปลงเป็นลักษณ์ต่างๆ โดยอาศัยแนวคิดต่างๆ ที่ได้ใช้ในการวิเคราะห์หาลักษณ์ในระดับคำมาประยุกต์ใช้ร่วมด้วย เช่น ลำดับร่วมที่ยาวที่สุด (*lcs*) เอ็นแกรม หรือการวัดค่าความละม้าย

2) ความสัมพันธ์แบบพึ่งพา (dependency relation)

ลักษณ์ชนิดนี้เป็นการนำแนวคิดของวานและคณะ (Wan et al., 2006) และมาลากาซิโอติส (Malakasiotis, 2009) มาใช้ กล่าวคือเป็นการพิจารณาความสัมพันธ์แบบพึ่งพาระหว่างคู่หน่วยเทียบโดยความสัมพันธ์ดังกล่าวได้จากการแจงส่วนประโยคแบบพึ่งพา ในกรณีนี้แต่ละข้อความในหน่วยคู่เทียบจะถูกแจงส่วน (parse) เพื่อให้ได้ชุดของชุดของความสัมพันธ์แบบพึ่งพา (1 ชุดต่อประโยค)

ความสัมพันธ์แบบพึ่งพานี้จะอยู่ในรูปคู่ของคำในแผนผังต้นไม้แบบพึ่งพาในฐานะส่วนหลักและส่วนพึ่งพา ทั้งนี้ ลักษณ์ชนิดนี้จะนำค่าความแม่นยำ ค่าความครบถ้วน และค่า f ซึ่งคำนวณได้จากสมการที่ 2.1, 2.2, และ 2.3 ในหัวข้อที่ 2.7.2 ตามลำดับ และนำค่าดังกล่าวมาใช้เป็นลักษณ์

อย่างไรก็ตาม เมื่อได้สำรวจเครื่องมือที่ใช้สำหรับแจกส่วนประโยคและกำกับความสัมพันธ์แบบพึ่งพาในภาษาไทยที่มีอยู่ในปัจจุบันแล้ว พบว่ายังไม่มีเครื่องมือชิ้นใดที่ให้ผลการแจกส่วนและกำกับความสัมพันธ์ในระดับที่น่าพอใจ ทั้งนี้ หากนำเครื่องมือเหล่านั้นมาใช้ก็อาจส่งผลต่อผลในภาพรวมของงานวิจัยชิ้นนี้ ด้วยเหตุนี้ ผู้วิจัยจึงตัดสินใจไม่ใช้ความสัมพันธ์แบบพึ่งพาเป็นลักษณ์สำหรับจำแนกข้อความลักลอกและไม่ลักลอกในงานวิจัยชิ้นนี้ และจะวิเคราะห์หาลักษณ์ชนิดอื่นที่คาดว่าจะให้ประสิทธิภาพในระดับที่ทัดเทียมกันมาใช้แทน

3) ค่าความละม้ายที่วัดได้ในระดับวากยสัมพันธ์

เช่นเดียวกับกับลักษณ์ทางศัพท์ ข้อมูลเชิงวากยสัมพันธ์ที่กล่าวมาข้างต้นก็สามารถนำมาคำนวณเป็นค่าความละม้ายของคู่หน่วยเทียบได้ด้วยวิธีต่างๆ ในส่วนนี้ ผู้วิจัยจะเขียนโปรแกรมสำหรับวัดค่าความละม้ายในระดับวากยสัมพันธ์ด้วยวิธีการวัดต่างๆ ขึ้น จากนั้นจึงนำค่าตัวเลขที่แสดงความละม้ายระหว่างคู่หน่วยเทียบมาใช้เป็นลักษณ์

3.4.3 การวิเคราะห์หาลักษณ์ทางความหมาย

ลักษณ์ทางความหมายถือเป็นลักษณ์ที่มีความท้าทายมากที่สุดในแง่การวิเคราะห์หา จากการทบทวนวรรณกรรมจะเห็นได้ว่างานวิจัยที่ใช้ลักษณ์ทางความหมายในการจำแนกประเภทข้อความที่มีความละม้ายก็ดี หรือใช้ในการตรวจหาการลักลอกก็ดี ล้วนแล้วแต่พึ่งพาการคืนค่าทางความหมายจากเครือข่ายคำ (WordNet) ทั้งสิ้น ในกรณีของภาษาไทยนั้น ไม่ปรากฏว่ามีเครือข่ายคำที่เสร็จสมบูรณ์พร้อมใช้งาน ในการวิเคราะห์หาลักษณ์ทางความหมายนี้ ผู้วิจัยจึงจำเป็นต้องอาศัยแนวคิดอื่นที่สามารถเป็นตัวแทนของความหมายได้โดยอ้อม ได้แก่ การวิเคราะห์ความหมายแอบแฝง (Latent Semantic Analysis: LSA) และการฝังคำ (word embedding) โดยผู้วิจัยจะนำค่าทางความหมายของแต่ละคำที่ได้จากแนวคิดดังกล่าวมาคำนวณให้ค่าโดยรวมของข้อความจากนั้นจึงนำไปใช้เป็นลักษณ์ต่อไป

ทั้งนี้ ผู้วิจัยจะได้แสดงผลที่ได้จากการวิเคราะห์หาลักษณ์ ซึ่งได้แก่แนวคิดที่ใช้ในการแทนรูปลักษณ์ วิธีการสกัดลักษณ์จากข้อมูล รวมถึงการอภิปรายถึงลักษณะเด่นของลักษณ์แต่ละตัว โดยละเอียดในบทที่ 5

3.5 การประเมินประสิทธิภาพของระบบ

การประเมินประสิทธิภาพของระบบตรวจหาการลักลอกที่ถูกสร้างขึ้นถือเป็นวัตถุประสงค์หลักอีกข้อหนึ่งของงานวิจัยชิ้นนี้ ในหัวข้อนี้ ผู้วิจัยจะได้กล่าวถึงวิธีการประเมินประสิทธิภาพของระบบโดยละเอียด อันประกอบไปด้วยเกณฑ์การประเมินประสิทธิภาพการจำแนกประเภทของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน การฝึกฝนและทดสอบประสิทธิภาพการจำแนกประเภทด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน การประเมินประสิทธิภาพของข้อมูลรับเข้าแต่ละประเภท และการประเมินประสิทธิภาพของลักษณ์ที่ใช้ในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอก รายละเอียดมีดังต่อไปนี้ดังต่อไปนี้

3.5.1 เกณฑ์การประเมินประสิทธิภาพการจำแนกประเภทของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

ในหัวข้อย่อยนี้จะกล่าวถึงเกณฑ์ที่ใช้ในการจำแนกประเภทของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ซึ่งจะถูกใช้เป็นค่าอ้างอิงประสิทธิภาพทั้งในขั้นตอนการทดสอบประสิทธิภาพของข้อมูลรับเข้าและการทดสอบประสิทธิภาพของลักษณ์ที่ใช้ในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอก ทั้งนี้ เพื่อให้เข้าใจเกณฑ์การประเมินดังกล่าว ผู้วิจัยจะขอกล่าวถึงค่าพื้นฐานอันเป็นที่มาของการคำนวณค่าที่ใช้ในเป็นเกณฑ์ประเมินประสิทธิภาพก่อน ค่าดังกล่าวเป็นผลของการจำแนกประเภทข้อความลักลอกและไม่ลักลอกโดยเครื่องเปรียบเทียบกับผลของการจำแนกการลักลอกและไม่ลักลอกโดยมนุษย์ ได้แก่

1) *ผลบวกจริง (true positive: tp)* คือจำนวนตัวอย่างของข้อความที่เครื่องและมนุษย์จำแนกว่าเป็นการลักลอกเหมือนกัน
2) *ผลบวกปลอม (false positive: fp)* คือจำนวนตัวอย่างของข้อความที่เครื่องจำแนกว่าเป็นการลักลอก แต่มนุษย์จำแนกว่าไม่เป็นการลักลอก
3) *ผลลบจริง (true negative: tn)* คือจำนวนตัวอย่างของข้อความที่เครื่องและมนุษย์จำแนกว่าไม่เป็นการลักลอกเหมือนกัน
4) *ผลลบปลอม (false negative: fn)* คือจำนวนตัวอย่างของข้อความที่เครื่องจำแนกว่าไม่เป็นการลักลอก แต่มนุษย์จำแนกว่าเป็นการลักลอก

ทั้งนี้ ผลการจำแนกประเภททั้ง 4 ค่าข้างต้นสามารถแสดงในรูปตารางได้ดังตารางที่ 3.7

ตารางที่ 3.7 ค่าตั้งต้นสำหรับคำนวณประสิทธิภาพของระบบ

ผลโดยเครื่อง / ผลโดยมนุษย์	ลักลอก	ไม่ลักลอก
ลักลอก	ผลบวกจริง (true positive: tp)	ผลลบปลอม (false negative: fn)
ไม่ลักลอก	ผลบวกปลอม (false positive: fp)	ผลลบจริง (true negative: tn)

จากค่าตั้งต้นทั้ง 4 ค่าข้างต้น สามารถนำมาคำนวณค่าที่ใช้เป็นเกณฑ์ในการประเมินประสิทธิภาพของระบบได้ดังต่อไปนี้

1) ค่าความแม่นยำ (precision)

ค่าความแม่นยำคืออัตราส่วนระหว่างผลการจำแนกข้อความลักลอกที่ถูกต้องต่อผลที่เครื่องจำแนกว่าลักลอกทั้งหมด ในแง่นี้ ค่าความแม่นยำจึงสะท้อนความถูกต้อง (accuracy) ของการจำแนกประเภท (L. Li, Wu, & Ye, 2015, p. 77) ทั้งนี้ สามารถคำนวณได้จากสมการที่ 3.2

$$precision = \frac{tp}{tp + fp} \tag{3.2}$$

2) ค่าความครบถ้วน (recall)

ค่าความครบถ้วนคืออัตราส่วนระหว่างผลการจำแนกข้อความลักลอกที่ถูกต้องต่อผลที่มนุษย์จำแนกว่าลักลอกทั้งหมด ในแง่นี้ ค่าความครบถ้วนจึงสะท้อนความครอบคลุม (comprehensiveness) ของการจำแนกประเภท (L. Li et al., 2015, p. 77) ทั้งนี้ สามารถคำนวณได้จากสมการที่ 3.3

$$recall = \frac{tp}{tp + fn} \tag{3.3}$$

3) ค่า F (F-measure or F_1)

ค่า F ถือเป็นวิธีการวัดแบบบูรณาการของค่าความแม่นยำและค่าความครบถ้วน แนวคิดในการคำนวณค่าดังกล่าวนี้คือการหาค่าเฉลี่ยฮาร์มอนิกแบบถ่วงน้ำหนักของค่าความแม่นยำและค่าความครบถ้วน (Manning et al., 2008, p. 144) ทั้งนี้ สามารถคำนวณได้จากสมการที่ 3.4

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.4}$$

ทั้งนี้ ในการประเมินประสิทธิภาพของระบบในงานวิจัยชิ้นนี้ ผู้วิจัยจะพิจารณาค่า F เป็นค่าหลักในการเปรียบเทียบประสิทธิภาพของข้อมูลรับเข้าหรือลักษณ์ หากพบว่าค่า F มีค่าเท่ากัน ผู้วิจัยจึงจะพิจารณาตัดสินประสิทธิภาพจากค่าความแม่นยำและค่าความครบถ้วน

3.5.2 การฝึกฝนและทดสอบประสิทธิภาพการจำแนกประเภทด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

ในหัวข้อย่อยนี้จะกล่าวถึงขั้นตอนการฝึกฝนและทดสอบประสิทธิภาพการจำแนกประเภทด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน รวมถึงการตั้งค่าแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนซึ่งเป็นถือเป็นตัวแปรควบคุมในการทดลองเปรียบเทียบประสิทธิภาพของข้อมูลรับเข้าและการทดลองเปรียบเทียบประสิทธิภาพของลักษณ์ที่ใช้ในการจำแนกข้อความลักลอกและข้อความที่ไม่มีการลักลอก

แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่ผู้วิจัยเลือกใช้ในงานวิจัยชิ้นนี้คือ SVC ซึ่งเป็นคลาสหนึ่งในไลบรารี Scikit-learn เวอร์ชัน 0.19.1 ซึ่งเป็นไลบรารีการเรียนรู้ของเครื่องสำหรับโปรแกรมภาษาไพทอน (python)

คลาส SVC ใน Scikit-learn นั้นทำให้เกิดผล (implement) โดยอิงจากไลบรารี libsvm ซึ่งพัฒนาโดยชางและลิน (Chang & Lin, 2011) มีประสิทธิภาพในการเป็นตัวจำแนกประเภทชุดข้อมูลทั้งในแบบ 2 ประเภท (2 classes classification) และแบบหลายประเภท (multi-class classification)

ในงานวิจัยชิ้นนี้ ผู้วิจัยใช้คลาส SVC ในการจำแนกแบบ 2 ประเภท คือจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอก โดยใช้ค่าพารามิเตอร์ปริยายของคลาส ซึ่งมีรายละเอียดดังแสดงในตารางที่ 3.8

ตารางที่ 3.8 การกำหนดค่าพารามิเตอร์ของคลาส SVC ในไลบรารี Scikit-learn

พารามิเตอร์	การตั้งค่า	คำอธิบาย
C	1.0	ค่าควบคุมการหาระนาบ (hyperplane) ที่ดีที่สุดของการจำแนกข้อมูลในปริภูมิเวกเตอร์
kernel	‘rbf’	ฟังก์ชันเคอร์เนลของแบบจำลอง ในที่นี้ตั้งค่าปริยายเป็นฟังก์ชันเรเดียลเบสิส
degree	3	องศาของฟังก์ชันเคอร์เนลโพลีโนเมียล ในที่นี้ไม่สนใจ (ignore) เนื่องจากใช้ฟังก์ชันเรเดียลเบสิส
gamma	auto	สัมประสิทธิ์ของฟังก์ชันเคอร์เนล ในกรณีที่ตั้งค่าเป็น ‘auto’ จะใช้ค่า 1/*n* feature แทน

พารามิเตอร์	การตั้งค่า	คำอธิบาย
coef0	0.0	ข้อตกลงอิสระในฟังก์ชันเคอร์เนล มีนัยสำคัญเฉพาะในฟังก์ชันโพลีโนเมียลและซิกมอยด์
probability	False	เปิดใช้การประมาณความน่าจะเป็น ในที่นี้คือไม่เปิดใช้งาน
shrinking	True	เปิดใช้งานการแก้ปัญหาแบบ shrinking (shrinking heuristic) ในที่นี้คือเปิดใช้งาน
tol	1e-3	ค่าความคลาดเคลื่อนยินยอม
cache_size	-	จำกัดขนาดแคชของเคอร์เนล (เป็น MB) ในที่นี้ไม่ระบุ
class_weight	-	การกำหนดค่าน้ำหนักให้ประเภทที่จำแนก ในที่นี้ไม่ระบุ ทุกประเภทจึงมีค่าน้ำหนักเดียวกัน
verbose	False	เปิดใช้งานการแสดงผลข้อมูลออกแบบ verbose ในที่นี้ไม่เปิดใช้งาน
max_iter	-1	ค่าขีดจำกัดในการแก้ปัญหา ในที่นี้ระบุเป็น -1 คือไม่จำกัด
decision_function_shape	‘ovr’	ระบุประเภทของฟังก์ชันการตัดสินใจ ในที่นี้ระบุเป็น ‘ovr’ หรือ “one-vs-rest” เครื่องจะตัดสินใจแบบ n_samples, n_classes
random_state	None	ระบุค่าของการสุ่มเมื่อต้องสับเปลี่ยนข้อมูล ในที่นี้ระบุว่าไม่มี

ในการฝึกฝนและทดสอบประสิทธิภาพด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน แบบจำลองต้องอาศัยชุดข้อมูล (data set) ซึ่งประกอบด้วยลักษณ์ (feature) และคำตอบที่ถูกต้องของการจำแนกประเภทโดยมนุษย์เป็นข้อมูลรับเข้า (input) ในการตัดสินจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอก ในงานวิจัยชิ้นนี้ ลักษณ์แต่ละตัวจะสกัดได้จากโปรแกรมที่ผู้วิจัยเขียนขึ้นโดยอิงจากผลการวิเคราะห์หาลักษณ์ตามหัวข้อที่ 3.5.1 ทั้งนี้ ลักษณ์และคำตอบที่ถูกต้องที่ใช้ในแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนจำเป็นต้องมีลักษณะเป็นตัวเลข (numeric) เท่านั้น

ในการป้อนชุดข้อมูลให้แก่แบบจำลอง ลักษณ์และคำตอบที่ถูกต้องในชุดข้อมูลต้องได้รับการแปลงให้อยู่ในรูปของ numpy array ก่อน ในขั้นตอนนี้ ผู้วิจัยจะบันทึกลักษณ์ที่สกัดได้และคำตอบที่ถูกต้องให้อยู่รูปไฟล์สกุล .csv เมื่อเข้าสู่ก่อนตอนก่อนการฝึกฝน จึงใช้คำสั่งในไลบรารี pandas เพื่ออ่านไฟล์ .csv ออกมาในรูปกรอบข้อมูล (data frame) จากนั้นจึงใช้คำสั่งแปลงกรอบข้อมูลดังกล่าวให้อยู่ในรูป numpy array 2 ชุด ชุดหนึ่งเป็น numpy array ที่บรรจุข้อมูลของลักษณ์ ส่วนอีกชุดบรรจุข้อมูลที่เป็นคำตอบที่ถูกต้องของการจำแนกประเภท โดยจำนวนกรณีของลักษณ์และจำนวน

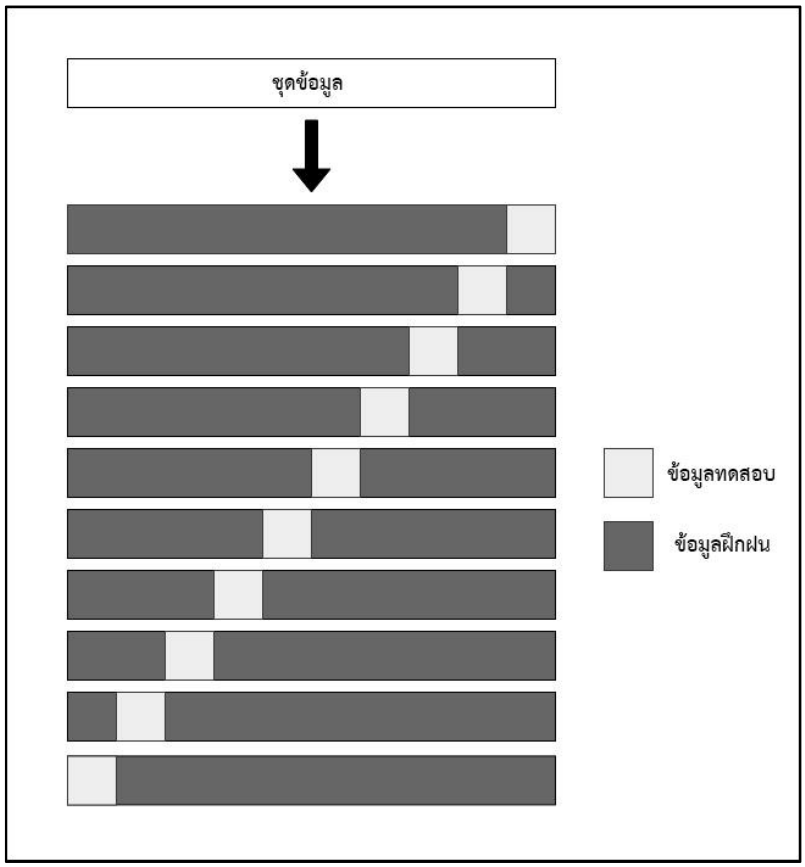
กรณีของคำตอบต้องเท่ากันเสมอ ทั้งนี้ numpy array ที่บรรจุข้อมูลของลักษณ์สามารถบรรจุข้อมูลของลักษณ์ได้มากกว่า 1 ชนิด ซึ่งลักษณะดังกล่าวยังเอื้อต่อการทดลองรวมชุดของลักษณ์ในการจำแนกประเภท

Case No.	Text A	Text B	lcs_w	Ans
1	HS-L-001-SR.txt	HS-L-001-EC.txt	163	1
2	HS-L-002-SR.txt	HS-L-002-EC.txt	196	1
3	HS-L-003-SR.txt	HS-L-003-EC.txt	183	1
4	HS-L-004-SR.txt	HS-L-004-EC.txt	184	1
5	HS-L-005-SR.txt	HS-L-005-EC.txt	197	1
6	HS-L-006-SR.txt	HS-L-006-EC.txt	193	1
7	HS-L-007-SR.txt	HS-L-007-EC.txt	156	1
8	HS-L-008-SR.txt	HS-L-008-EC.txt	180	1
9	HS-L-009-SR.txt	HS-L-009-EC.txt	152	1
10	HS-L-010-SR.txt	HS-L-010-EC.txt	194	1
11	HS-L-011-SR.txt	HS-L-011-EC.txt	180	1
12	HS-L-012-SR.txt	HS-L-012-EC.txt	190	1
13	HS-L-013-SR.txt	HS-L-013-EC.txt	152	1
14	HS-L-014-SR.txt	HS-L-014-EC.txt	189	1
15	HS-L-015-SR.txt	HS-L-015-EC.txt	184	1
16	HS-L-016-SR.txt	HS-L-016-EC.txt	189	1
17	HS-L-017-SR.txt	HS-L-017-EC.txt	163	1
18	HS-L-018-SR.txt	HS-L-018-EC.txt	182	1
19	HS-L-019-SR.txt	HS-L-019-EC.txt	166	1
20	HS-L-020-SR.txt	HS-L-020-EC.txt	188	1
⋮	⋮	⋮	⋮	⋮

ภาพที่ 3.15 ตัวอย่างชุดข้อมูล 20 กรณีแรกของคลังข้อมูล

ภาพที่ 3.15 แสดงชุดข้อมูลซึ่งบันทึกค่าลักษณ์และคำตอบที่ถูกต้องในการจำแนกประเภทโดยมนุษย์ในรูปไฟล์ .csv และแสดงผลในโปรแกรมไมโครซอฟต์เอกซ์เซล (Microsoft Excel) จากภาพจะเห็นกรณีต้องสงสัย 20 กรณีแรกของคลังข้อมูลแสดงในคอลัมน์ “Case No.” จากนั้นจะเป็นชื่อไฟล์ของคู่หน่วยเทียบ คอลัมน์ “Text A” แสดงย่อหน้าต้นฉบับ ส่วนคอลัมน์ “Text B” แสดงย่อหน้าลักลอกประเภทคัดลอกโดยตรง ต่อมาในคอลัมน์ “lcs_w” เป็นลักษณ์ โดยลักษณ์ในที่นี้คือค่า

ความยาวของลำดับร่วมที่ยาวที่สุดในชุดคำของคู่หน่วยเทียบ และคอลัมน์ “Ans” เป็นคำตอบที่ถูกต้องของการจำแนกประเภทโดยมนุษย์ หากกำหนดค่าเป็น 1 หมายถึงลักลอก หากกำหนดค่าเป็น 0 หมายถึงไม่ลักลอก



ภาพที่ 3.16 การแบ่งข้อมูลฝึกฝนและทดสอบในการตรวจสอบไขว้ 10 ทบ

เมื่อแปลงลักษณ์และคำตอบที่ถูกต้องให้อยู่ในรูป numpy array เรียบร้อยแล้ว ขั้นตอนต่อมาคือการฝึกฝนและทดสอบแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนด้วยลักษณ์และคำตอบดังกล่าว ทั้งนี้ เพื่อความน่าเชื่อถือของการประเมินประสิทธิภาพ ผู้วิจัยได้ประยุกต์ใช้การตรวจสอบไขว้ k ทบ (k-fold cross validation) (Fushiki, 2011) ในการฝึกฝนและทดสอบแบบจำลอง กล่าวคือ ข้อมูลในชุดข้อมูลทั้งหมดจะถูกแบ่งออกโดยการสุ่มเป็น k ส่วนเท่าๆ กัน โดยข้อมูลจำนวน k-1 ส่วนจะถูกกำหนดให้เป็นข้อมูลสำหรับฝึกฝน และข้อมูลที่เหลืออีก $\frac{1}{k}$ ส่วนจะถูกกำหนดให้เป็นข้อมูลสำหรับทดสอบ กระบวนการฝึกฝนและทดสอบทั้งหมดจะทำซ้ำทั้งหมด k ครั้ง โดยจะเวียนข้อมูลทั้ง k ส่วนเป็นข้อมูลทดสอบ 1 ครั้ง การฝึกฝนและทดสอบแต่ละครั้งจะคืนค่าการประเมินประสิทธิภาพออกมา 4 ค่า ได้แก่ ค่าความถูกต้อง ค่าความแม่นยำ ค่าความครบถ้วน และค่า F ดังได้กล่าวไปในหัวข้อที่

3.5.1 ผู้วิจัยจะนำค่าการประเมินแต่ละค่าที่ได้จากการทดสอบทั้ง *k* ครั้งมาหาค่าเฉลี่ยและใช้เป็นผลลัพธ์ของการประเมินประสิทธิภาพ

ในงานวิจัยชิ้นนี้ได้กำหนดให้ฝึกฝนและทดสอบแบบจำลองด้วยการตรวจสอบไขว้ 10 ทบ (*k* = 10) ดังแสดงในภาพที่ 3.16 จากภาพจะเห็นได้ว่าชุดข้อมูลที่ป้อนให้แก่แบบจำลองถูกแบบออกเป็น 10 ส่วน และดำเนินการฝึกฝนและทดสอบทั้งหมด 10 ครั้ง แต่ละครั้งข้อมูล 9 จาก 10 ส่วนจะถูกกำหนดให้เป็นข้อมูลสำหรับฝึกฝนและข้อมูล 1 จาก 10 ส่วนจะถูกกำหนดให้เป็นข้อมูลทดสอบ ข้อมูลทั้ง 10 ส่วนจะถูกเวียนเพื่อใช้เป็นข้อมูลทดสอบ 1 ครั้ง ด้วยวิธีการนี้ การตรวจสอบไขว้แต่ละรอบจะคืนค่าการประเมินประสิทธิภาพทั้ง 4 ค่าออกมาทั้งหมด 10 ครั้ง ผู้วิจัยจะหาค่าเฉลี่ยทั้ง 10 ครั้งของค่าการประเมินประสิทธิภาพแต่ละค่าและใช้เป็นผลลัพธ์ของการประเมิน

3.5.3 การประเมินประสิทธิภาพของระบบเมื่อใช้ข้อมูลรับเข้าที่ต่างกัน

กล่าวได้ว่าลักษณะของข้อมูลรับเข้า (input) เป็นปัจจัยหนึ่งที่ส่งผลต่อประสิทธิภาพการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก ทั้งนี้ จากการทบทวนวรรณกรรมในหัวข้อที่ 2.8 ที่ว่าด้วยทฤษฎีโครงสร้างวาทะนั้น ผู้วิจัยได้แสดงให้เห็นถึงลักษณะเด่นของหน่วยปริจเฉทพื้นฐานอันเป็นหน่วยย่อยในการวิเคราะห์โครงสร้างและความสัมพันธ์ภายในปริจเฉทตามทฤษฎีดังกล่าว กระทั่งนำมาสู่การวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทยโดยพิจารณาการแก้ไขข้อความต้นฉบับในระดับหน่วยปริจเฉทพื้นฐานดังได้กล่าวไปแล้วในหัวข้อที่ 3.2.2 อีกทั้งผลจากการวิเคราะห์กลวิธีการลักลอกงานวิชาการภาษาไทยยังชี้ให้เห็นว่าผู้ลักลอกใช้กลวิธีลักลอกทั้งภายในหน่วยปริจเฉทพื้นฐานเดียวกันและระหว่างหน่วยปริจเฉทพื้นฐาน[12] ด้วยเหตุนี้ ผู้วิจัยจึงเกิดความสนใจและต้องการพิสูจน์ว่าระหว่างย่อหน้าซึ่งเป็นหน่วยของข้อมูลรับเข้าที่ใช้ในการออกแบบและสร้างคลังข้อมูลในงานวิจัยชิ้นนี้กับหน่วยปริจเฉทพื้นฐาน ข้อมูลรับเข้าลักษณะใดจะเอื้อให้เกิดประสิทธิภาพที่ดีกว่าในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก จนนำมาสู่การออกแบบการทดลองเพื่อเปรียบเทียบประสิทธิภาพของข้อมูลรับเข้าทั้ง 2 ลักษณะดังกล่าว

ในหัวข้อย่อยนี้ ผู้วิจัยจะกล่าวถึงวิธีการประเมินประสิทธิภาพของระบบเมื่อใช้ข้อมูลรับเข้าที่ต่างกันในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก ซึ่งประกอบด้วยปัจจัยและขั้นตอนย่อยที่เกี่ยวข้องดังต่อไปนี้

[12] ดูรายละเอียดในหัวข้อที่ 4.1

3.5.3.1 ชุดข้อมูลทดลอง

ขั้นตอนแรกของการประเมินประสิทธิภาพของระบบเมื่อใช้ข้อมูลรับเข้าที่ต่างกันคือการเตรียมชุดข้อมูลทดลอง ชุดข้อมูลทดลองในที่นี้จะถูกใช้ในการสกัดลักษณ์และให้คำตอบที่ถูกต้องจากการจำแนกประเภทโดยมนุษย์ อย่างไรก็ตาม เนื่องมาจากในกรณีของภาษาไทยนั้นยังไม่มีเครื่องมือสำหรับตัดแยกขอบเขตของหน่วยปริจเฉทพื้นฐานที่มีประสิทธิภาพเป็นที่น่าพอใจ ฉะนั้นการทดลองในขั้นนี้จึงไม่สามารถใช้ข้อมูลจากคลังข้อมูลการลักลอกที่ผู้วิจัยสร้างขึ้นทั้งหมดได้ เพราะจำนวนข้อมูลมีมากเกินกว่าจะตัดแยกขอบเขตของหน่วยปริจเฉทพื้นฐานได้โดยตัวผู้วิจัยเองและนำมาใช้เป็นข้อมูลรับเข้าในการทดลอง ด้วยเหตุผลนี้ ผู้วิจัยจึงได้สร้างชุดข้อมูลขนาดย่อมขึ้นเพื่อใช้ในการทดลอง

ข้อมูลที่นำมาสร้างเป็นชุดข้อมูลทดลองนั้นได้มาจากการสุ่มตัวอย่างข้อมูลจากคลังข้อมูลการลักลอกที่สร้างขึ้นตามที่ได้กล่าวไปในหัวข้อที่ 3.3 โดยผู้วิจัยจะสุ่มตัวอย่างคู่หน่วยเทียบตามสาขาวิชา ขนาดของย่อหน้า และประเภทของการลักลอก ให้ได้จำนวนทั้งสิ้น 150 คู่หน่วยเทียบ ดังได้แสดงรายละเอียดไว้ในตารางที่ 3.9

ตารางที่ 3.9 รายละเอียดของข้อมูลในชุดข้อมูลทดลอง

สาขาวิชา	ขนาดของย่อหน้า	ประเภทของข้อมูลการลักลอก	จำนวน (คู่)
วิทยาศาสตร์ (SC)	ยาว (L)	คัดลอกโดยตรง (EC)	5
		คัดลอกโดยใกล้เคียง (NC)	5
		คัดลอกโดยดัดแปลง (MO)	5
		ถอดความ (PA)	5
		ไม่ลักลอก (NO)	5
	กลาง (M)	คัดลอกโดยตรง (EC)	5
		คัดลอกโดยใกล้เคียง (NC)	5
		คัดลอกโดยดัดแปลง (MO)	5
		ถอดความ (PA)	5
		ไม่ลักลอก (NO)	5
	สั้น (S)	คัดลอกโดยตรง (EC)	5
		คัดลอกโดยใกล้เคียง (NC)	5
		คัดลอกโดยดัดแปลง (MO)	5
		ถอดความ (PA)	5

สาขาวิชา	ขนาดของย่อหน้า	ประเภทของข้อมูลการลักลอก	จำนวน (คู่)
		ไม่ลักลอก (NO)	5
มนุษยศาสตร์และสังคมศาสตร์ (HS)	ยาว (L)	คัดลอกโดยตรง (EC)	5
		คัดลอกโดยใกล้เคียง (NC)	5
		คัดลอกโดยดัดแปลง (MO)	5
		ถอดความ (PA)	5
		ไม่ลักลอก (NO)	5
	กลาง (M)	คัดลอกโดยตรง (EC)	5
		คัดลอกโดยใกล้เคียง (NC)	5
		คัดลอกโดยดัดแปลง (MO)	5
		ถอดความ (PA)	5
		ไม่ลักลอก (NO)	5
	สั้น (S)	คัดลอกโดยตรง (EC)	5
		คัดลอกโดยใกล้เคียง (NC)	5
		คัดลอกโดยดัดแปลง (MO)	5
		ถอดความ (PA)	5
		ไม่ลักลอก (NO)	5
		รวม	**150**

หลังจากการสุ่มตัวอย่างข้อมูลจากคลังข้อมูลการลักลอกได้ตามรายละเอียดข้างต้นแล้ว ในขั้นตอนต่อมา ผู้วิจัยจะใช้ข้อมูลดังกล่าวสร้างเป็นชุดข้อมูลทดลอง 2 ชุดตามลักษณะของข้อมูลรับเข้าที่ต้องการศึกษา ได้แก่ ชุดข้อมูลทดลอง PRG และชุดข้อมูลทดลอง EDU ชุดข้อมูลทดลอง PRG เป็นตัวแทนของข้อมูลรับเข้าเป็นย่อหน้า ผู้วิจัยจะคงรูปแบบของคู่หน่วยเทียบที่เป็นย่อหน้าไว้เช่นเดิมตามที่สุ่มได้จากคลังข้อมูลการลักลอก ส่วนชุดข้อมูลทดลอง EDU นั้น ผู้วิจัยจะนำย่อหน้าที่อยู่ในรูปคู่หน่วยเทียบมาตัดแยกขอบเขตของหน่วยปริจเฉทพื้นฐานตามหลักการของนลินี อินต๊ะซาว และวิโรจน์ อรุณมานะกุล (Intasaw & Aroonmanakun, 2013) ทั้งย่อหน้าต้นฉบับและย่อหน้าที่ผ่านการลักลอก เพื่อใช้เป็นตัวแทนของข้อมูลรับเข้าที่เป็นหน่วยปริจเฉทพื้นฐาน

เมื่อได้ชุดข้อมูลทดลองทั้งสองชุดข้างต้น ผู้วิจัยจะสกัดลักษณ์และให้คำตอบที่ถูกต้องของการจำแนกจากชุดข้อมูลทดลองทั้งสองชุดเป็นขั้นตอนต่อไป

3.5.3.2 ลักษณ์และการให้คำตอบ

ดังได้กล่าวไปในหัวข้อที่ 3.5.2 แล้วว่าข้อมูลรับเข้าในกระบวนการฝึกฝนและทดสอบประสิทธิภาพด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนนั้นได้แก่ลักษณ์และคำตอบที่ถูกต้องจากการจำแนกประเภทโดยมนุษย์ ในการทดลองเปรียบเทียบประสิทธิภาพของข้อมูลรับเข้าในตอนต้นของระบบตรวจหาการลักลอกอันได้แก่ย่อหน้าและหน่วยปริจเฉทพื้นฐานจึงต้องกระทำผ่านลักษณ์และคำตอบที่ได้จากข้อมูลรับเข้าทั้งสองประเภท

ในการทดลองขั้นนี้ ประเภทของลักษณ์ถือเป็นตัวแปรควบคุม ผู้วิจัยได้เลือกใช้ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ (Sørensen-Dice coefficient similarity) ที่เอ็นแกรมของคำตั้งแต่ 1-5 แกรมเป็นลักษณ์ที่ใช้ฝึกฝนและทดสอบประสิทธิภาพของข้อมูลรับเข้า

ในขั้นตอนการสกัดลักษณ์นั้น ในชุดข้อมูลทดลอง PRG ผู้วิจัยจะเขียนโปรแกรมเพื่อวัดค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ระหว่างย่อหน้าของคู่หน่วยเทียบขึ้น และใช้ค่าความละม้ายดังกล่าวเป็นลักษณ์ที่ได้จากข้อมูลรับเข้าที่เป็นย่อหน้า ดังนั้นลักษณ์ที่ได้ในขั้นนี้จะมีจำนวนทั้งหมด 600 ค่าเท่ากับจำนวนของคู่หน่วยเทียบที่เป็นคู่ของย่อหน้าคูณด้วยจำนวนประเภทของลักษณ์ทั้ง 5 ตัว ส่วนชุดข้อมูลทดลอง EDU ซึ่งเป็นตัวแทนของข้อมูลรับเข้าที่เป็นหน่วยปริจเฉทพื้นฐานนั้น ผู้วิจัยจะใช้โปรแกรมวัดค่าความละม้ายระหว่างหน่วยปริจเฉทพื้นฐานที่มาจากย่อหน้าต้นฉบับกับหน่วยปริจเฉทพื้นฐานที่มาจากย่อหน้าลักลอกแบบพบกันทั้งหมด กล่าวคือ หน่วยปริจเฉทพื้นฐานที่มาจากย่อหน้าต้นฉบับแต่ละหน่วยจะถูกเทียบกับหน่วยปริจเฉทพื้นฐานที่มาจากย่อหน้าลักลอกแต่ละหน่วยและวัดค่าความละม้ายออกมาใช้เป็นลักษณ์ คู่หน่วยเทียบในขั้นนี้จึงเป็นคู่ของหน่วยปริจเฉทพื้นฐานซึ่งเป็นตัวสะท้อนประสิทธิภาพของระบบที่ใช้ข้อมูลรับเข้าที่เป็นคู่หน่วยเทียบ

SRC EDU CODE	SOURCE TEXT	PLG EDU CODE	PLAGIARIZED TEXT
HS-S-497-SR_01.txt	ในการทดลองผลงานครั้งที่ 2	HS-S-497-PA_01.txt	ในการทดลองผลงานครั้งถัดมา
HS-S-497-SR_02.txt	ผู้วิจัยได้แสดงผลงานให้กับนักศึกษา นิสิต สาขาวิชานาฏยศิลป์ ภาควิชานาฏยศิลป์ คณะศิลปกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย	HS-S-497-PA_02.txt	ผู้วิจัยได้แสดงผลงานให้กับนิสิตนักศึกษาสาขาวิชานาฏยศิลป์ ภาควิชานาฏยศิลป์ คณะศิลปกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
		HS-S-497-PA_03.txt	เพื่อสอบถามความคิดเห็นและข้อเสนอแนะ
HS-S-497-SR_03.txt	และได้นำข้อเสนอแนะต่าง ๆ มารวมกับข้อคิดเห็นจากผู้เชี่ยวชาญทางด้านนาฏยศิลป์	HS-S-497-PA_04.txt	และได้นำความคิดเห็นและข้อเสนอแนะ
		HS-S-497-PA_05.txt	ที่ได้
HS-S-497-SR_03.txt	และได้นำข้อเสนอแนะต่าง ๆ มารวมกับข้อคิดเห็นจากผู้เชี่ยวชาญทางด้านนาฏยศิลป์	HS-S-497-PA_06.txt	มารวมกับข้อคิดเห็นจากผู้เชี่ยวชาญทางด้านนาฏยศิลป์
HS-S-497-SR_04.txt	ในที่นี้		
HS-S-497-SR_05.txt	จึงสามารถสรุปปัญหาและวิธีการแก้ไขในประเด็นต่าง ๆ ตามองค์ประกอบทางด้านนาฏยศิลป์	HS-S-497-PA_07.txt	จึงสามารถสรุปปัญหาและวิธีการแก้ไขในประเด็นต่าง ๆ ตามองค์ประกอบทางด้านนาฏยศิลป์
HS-S-497-SR_06.txt	ดังนี้	HS-S-497-PA_08.txt	ดังนี้

ภาพที่ 3.17 การจัดแนวเทียบหาคู่ลักลอกในข้อมูลรับเข้าประเภทหน่วยปริจเฉทพื้นฐาน

ส่วนการให้คำตอบของการจำแนกประเภทนั้น ในชุดข้อมูลทดลอง PRG สามารถใช้คำตอบเดิมที่ได้จากการกำกับข้อมูลในขั้นการสร้างคลังข้อมูลการลักลอกได้ทันที ส่วนชุดข้อมูลทดลอง EDU นั้น เนื่องจากมีการเทียบคู่หน่วยเทียบแบบพบกันทั้งหมด ผู้วิจัยจึงต้องให้คำตอบของการจำแนก

ประเภทใหม่ โดยในขั้นแรก ผู้วิจัยจะจัดแนว (align) เทียบระหว่างหน่วยปริจเฉทพื้นฐานในย่อหน้าต้นฉบับกับหน่วยปริจเฉทพื้นฐานในย่อหน้าลักลอกเพื่อหาคู่ลักลอก (plagiarism pair) ด้วยวิธีที่ได้กล่าวไปแล้วในหัวข้อที่ 3.2.2 ดังตัวอย่างในภาพที่ 3.17 เมื่อได้คู่ลักลอกเรียบร้อยแล้ว จึงนำคู่ลักลอกที่ได้ไปค้นหาในกลุ่มของคู่หน่วยเทียบ หากปรากฏว่าตรงกันแสดงว่าคู่หน่วยเทียบนั้นเป็นคู่ลักลอก และสามารถให้จัดประเภทเป็นการลักลอกได้ดังตัวอย่างในภาพที่ 3.18

PLG EDU CODE	SRC EDU CODE	PLAGIARIZED TEXT	SOURCE TEXT	ANS
HS-S-497-PA_01.txt	HS-S-497-SR_01.txt	ในการทดลองผลงานครั้งถัดมา	ในการทดลองผลงานครั้งที่ 2	1
HS-S-497-PA_01.txt	HS-S-497-SR_02.txt	ในการทดลองผลงานครั้งถัดมา	ผู้วิจัยได้แสดงผลงานให้กับนักศึกษา นิสิต สาขาวิชานาฏยศิลป์ ภ	0
HS-S-497-PA_01.txt	HS-S-497-SR_03.txt	ในการทดลองผลงานครั้งถัดมา	และได้นำข้อเสนอแนะต่าง ๆ มารวมกับข้อคิดเห็นจากผู้เชี่ยวชาญ	0
HS-S-497-PA_01.txt	HS-S-497-SR_04.txt	ในการทดลองผลงานครั้งถัดมา	ในที่นี้	0
HS-S-497-PA_01.txt	HS-S-497-SR_05.txt	ในการทดลองผลงานครั้งถัดมา	จึงสามารถสรุปปัญหาและวิธีการแก้ไขในประเด็นต่าง ๆ ตามองค์ป	0
HS-S-497-PA_01.txt	HS-S-497-SR_06.txt	ในการทดลองผลงานครั้งถัดมา	ดังนี้	0
HS-S-497-PA_02.txt	HS-S-497-SR_01.txt	ผู้วิจัยได้แสดงผลงานให้กับนิสิตนักศึกษาสาขาวิชานา	ในการทดลองผลงานครั้งที่ 2	0
HS-S-497-PA_02.txt	HS-S-497-SR_02.txt	ผู้วิจัยได้แสดงผลงานให้กับนิสิตนักศึกษาสาขาวิชานา	ผู้วิจัยได้แสดงผลงานให้กับนักศึกษา นิสิต สาขาวิชานาฏยศิลป์ ภ	1
HS-S-497-PA_02.txt	HS-S-497-SR_03.txt	ผู้วิจัยได้แสดงผลงานให้กับนิสิตนักศึกษาสาขาวิชานา	และได้นำข้อเสนอแนะต่าง ๆ มารวมกับข้อคิดเห็นจากผู้เชี่ยวชาญ	0
HS-S-497-PA_02.txt	HS-S-497-SR_04.txt	ผู้วิจัยได้แสดงผลงานให้กับนิสิตนักศึกษาสาขาวิชานา	ในที่นี้	0
HS-S-497-PA_02.txt	HS-S-497-SR_05.txt	ผู้วิจัยได้แสดงผลงานให้กับนิสิตนักศึกษาสาขาวิชานา	จึงสามารถสรุปปัญหาและวิธีการแก้ไขในประเด็นต่าง ๆ ตามองค์ป	0
HS-S-497-PA_02.txt	HS-S-497-SR_06.txt	ผู้วิจัยได้แสดงผลงานให้กับนิสิตนักศึกษาสาขาวิชานา	ดังนี้	0
HS-S-497-PA_03.txt	HS-S-497-SR_01.txt	เพื่อสอบถามความคิดเห็นและข้อเสนอแนะ	ในการทดลองผลงานครั้งที่ 2	0
HS-S-497-PA_03.txt	HS-S-497-SR_02.txt	เพื่อสอบถามความคิดเห็นและข้อเสนอแนะ	ผู้วิจัยได้แสดงผลงานให้กับนักศึกษา นิสิต สาขาวิชานาฏยศิลป์ ภ	0
HS-S-497-PA_03.txt	HS-S-497-SR_03.txt	เพื่อสอบถามความคิดเห็นและข้อเสนอแนะ	และได้นำข้อเสนอแนะต่าง ๆ มารวมกับข้อคิดเห็นจากผู้เชี่ยวชาญ	0
HS-S-497-PA_03.txt	HS-S-497-SR_04.txt	เพื่อสอบถามความคิดเห็นและข้อเสนอแนะ	ในที่นี้	0
HS-S-497-PA_03.txt	HS-S-497-SR_05.txt	เพื่อสอบถามความคิดเห็นและข้อเสนอแนะ	จึงสามารถสรุปปัญหาและวิธีการแก้ไขในประเด็นต่าง ๆ ตามองค์ป	0
HS-S-497-PA_03.txt	HS-S-497-SR_06.txt	เพื่อสอบถามความคิดเห็นและข้อเสนอแนะ	ดังนี้	0
HS-S-497-PA_04.txt	HS-S-497-SR_01.txt	และได้นำความคิดเห็นและข้อเสนอแนะ	ในการทดลองผลงานครั้งที่ 2	0
HS-S-497-PA_04.txt	HS-S-497-SR_02.txt	และได้นำความคิดเห็นและข้อเสนอแนะ	ผู้วิจัยได้แสดงผลงานให้กับนักศึกษา นิสิต สาขาวิชานาฏยศิลป์ ภ	0
HS-S-497-PA_04.txt	HS-S-497-SR_03.txt	และได้นำความคิดเห็นและข้อเสนอแนะ	และได้นำข้อเสนอแนะต่าง ๆ มารวมกับข้อคิดเห็นจากผู้เชี่ยวชาญ	1
HS-S-497-PA_04.txt	HS-S-497-SR_04.txt	และได้นำความคิดเห็นและข้อเสนอแนะ	ในที่นี้	0
HS-S-497-PA_04.txt	HS-S-497-SR_05.txt	และได้นำความคิดเห็นและข้อเสนอแนะ	จึงสามารถสรุปปัญหาและวิธีการแก้ไขในประเด็นต่าง ๆ ตามองค์ป	0
HS-S-497-PA_04.txt	HS-S-497-SR_06.txt	และได้นำความคิดเห็นและข้อเสนอแนะ	ดังนี้	0

ภาพที่ 3.18 การให้คำตอบของการจำแนกประเภทการลักลอกในคู่หน่วยเทียบที่เป็นหน่วยปริจเฉทพื้นฐาน

เมื่อได้ลักษณ์และคำตอบ จากชุดข้อมูลทดลองทั้งสองชุดแล้ว ผู้วิจัยจะนำลักษณ์และคำตอบดังกล่าวไปฝึกฝนและทดสอบประสิทธิด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนต่อไป

3.5.3.3 การฝึกฝนและทดสอบประสิทธิภาพแบบจำลอง

เมื่อได้ลักษณ์และคำตอบที่ถูกต้องจากการจำแนกประเภทโดยมนุษย์จากชุดข้อมูลทดลองทั้งสองชุดซึ่งเป็นตัวแทนของข้อมูลรับเข้าประเภทย่อหน้าและข้อมูลรับเข้าประเภทหน่วยปริจเฉทพื้นฐานแล้ว ในขั้นตอนนี้ ผู้วิจัยจะนำลักษณ์และคำตอบทั้งสองชุดมาฝึกฝนและทดสอบด้วยแบบจำลองซัพพอร์เวกเตอร์แมชชีนทีละชุดตามวิธีการที่ได้กล่าวไปในหัวข้อที่ 3.5.2

การทดสอบจากให้ค่าการประเมินประสิทธิภาพของแบบจำลองที่ใช้ลักษณ์จากข้อมูลรับเข้าที่ต่างกันในการฝึกฝนทั้งหมด 3 ค่า ได้แก่ ค่าความแม่นยำ ค่าความครบถ้วน และค่า F และเนื่องมาจากลักษณ์ที่ใช้ฝึกฝนและทดสอบในขั้นตอนนี้มีด้วยกัน 5 ตัวตามจำนวนเอ็นแกรมของคำของค่าความละม้ายที่ได้กำหนดไว้ในหัวข้อที่แล้ว ดังนั้นนอกจากจะเปรียบเทียบประสิทธิภาพเป็นรายลักษณ์แล้ว ผู้วิจัยจะได้ฝึกฝนและทดสอบแบบจำลองโดยรวมชุดของลักษณ์ทั้ง 5 ตัวเข้าด้วยกันด้วยเพื่อพิจารณา

ประสิทธิภาพของข้อมูลรับเข้าในภาพรวม จากนั้นจึงสรุปและอภิปรายผลการทดลองเปรียบเทียบประสิทธิภาพของข้อมูลรับเข้าต่อไป

3.5.4 การประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน

การประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกันในการจำแนกประเภทข้อความที่มีการลักลอกและข้อความที่ไม่มีการลักลอกเป็นวัตถุประสงค์ที่สำคัญอีกประการของงานวิจัยชิ้นนี้ ในขั้นตอนนี้ ผู้วิจัยจะนำลักษณ์ที่วิเคราะห์ได้จากวิธีการที่แสดงในหัวข้อที่ 3.4 มาเปรียบเทียบและประเมินประสิทธิภาพ โดยสกัดลักษณ์แต่ละตัวจากข้อมูลทั้งหมดภายในคลังข้อมูลที่ได้สร้างไว้ตามวิธีการในหัวข้อที่ 3.3 ซึ่งประกอบด้วยคู่หน่วยเทียบซึ่งมีลักษณะข้อมูลรับเข้าเป็นคู่ของย่อหน้าจำนวน 50,000 คู่ ทั้งนี้ ผู้วิจัยได้แบ่งขั้นตอนการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกันออกเป็น 3 ขั้น รายละเอียดมีดังต่อไปนี้

3.5.4.1 การประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์เป็นรายลักษณ์

การทดลองในขั้นนี้มีวัตถุประสงค์เพื่อประเมินประสิทธิภาพของลักษณ์ที่มีผลในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกแต่ละลักษณ์เป็นรายลักษณ์ทั้งลักษณ์ทางภาษาและลักษณ์อิงอักขระ

ในขั้นตอนนี้ ผู้วิจัยจะนำลักษณที่ไม่ใช่ทางภาษา อันได้แก่ลักษณ์ที่สกัดได้จากลักษณะด้านความยาวของข้อความ และลักษณ์ที่สกัดได้จากลักษณะทางอักขระของข้อความ รวมถึงลักษณ์ที่วิเคราะห์ได้จากวิธีการที่แสดงในหัวข้อที่ 3.4 แต่ละตัวมาฝึกฝนและทดสอบประสิทธิภาพด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนเป็นรายตัวตามวิธีการที่ได้กล่าวไปแล้วในหัวข้อที่ 3.5.2

ลักษณ์แต่ละตัวที่ผ่านการทดสอบในขั้นนี้จะคืนค่าการประเมินทั้ง 3 ค่า ได้แก่ ค่าความแม่นยำ ค่าความครบถ้วน และค่า F ผู้วิจัยจะพิจารณาค่าการประเมินเหล่านี้ตามวิธีที่ระบุไว้ในหัวข้อที่ 3.5.1 และเปรียบเทียบประสิทธิภาพของลักษณ์แต่ละตัวจากค่าประเมินดังกล่าว จากนั้นจึงอภิปรายผลการประเมิน

3.5.4.2 การประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์แบบรวมชุด

การทดลองในขั้นนี้มีวัตถุประสงค์เพื่อวิเคราะห์ว่าควรจัดชุดรวมของลักษณ์โดยประกอบด้วยลักษณ์ตัวใดบ้างจึงจะส่งผลให้แบบจำลองสามารถจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกได้ผลดีที่สุด

ทั้งนี้ หลังจากทดสอบและประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์เป็นรายลักษณ์ตามขั้นตอนในหัวข้อที่ 3.5.4.1 แล้วจะทำให้ทราบอันดับของลักษณ์ที่มีประสิทธิภาพในการจำแนก

ประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกมากไปหาน้อย ในขั้นตอนนี้ ผู้วิจัยจะนำลักษณ์ที่มีประสิทธิภาพดีที่สุด 10 อันดับแรกมารวมเป็นชุดของลักษณ์และนำฝึกฝนและทดสอบประสิทธิภาพด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนตามขั้นตอนในหัวข้อที่ 3.5.2 จากนั้นจะทดลองรวมชุดของลักษณ์ใหม่โดยตัดลักษณ์ที่มีประสิทธิภาพอยู่ในอันดับที่ต่ำที่สุดออกไป 1 ลักษณ์ และนำฝึกฝนและทดสอบประสิทธิภาพด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนอีกรอบ ผู้วิจัยจะจัดชุดของลักษณ์ใหม่และนำไปฝึกฝนและทดสอบเช่นนี้จนกระทั่งเหลือลักษณ์ที่มีประสิทธิภาพอยู่ในอันดับที่สูงที่สุดเพียงตัว จากวิธีการดังกล่าวนี้จะทำให้ได้ชุดของลักษณ์ทั้งหมด 10 ชุดที่ผ่านการฝึกและทดสอบด้วยแบบจำลอง

การฝึกฝนและทดสอบประสิทธิภาพด้วยแบบจำลองแต่ละรอบตามขั้นตอนข้างต้นจะให้ค่าการประเมิน 3 ค่าที่เป็นตัวบ่งชี้ประสิทธิภาพของระบบเมื่อใช้ชุดรวมของลักษณ์แต่ละชุดออกมา ด้วยวิธีการดังกล่าวนี้ จะทำให้สามารถประเมินได้ว่าการรวมชุดของลักษณ์ตัวใดบ้างที่ให้ประสิทธิภาพในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกดีที่สุด จากนั้นจึงอภิปรายผลการประเมินในขั้นนี้

3.5.4.3 การประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ทางภาษา

การประเมินประสิทธิภาพของลักษณ์ในขั้นสุดท้ายนี้มีวัตถุประสงค์เพื่อประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ทางภาษาในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอก ด้วยงานวิจัยชิ้นนี้ตั้งสมมติฐานในตอนต้นไว้ว่าลักษณ์ที่ประยุกต์ใช้ความรู้ทางภาษาในการสร้างและสกัดจะมีประสิทธิภาพในการตรวจหาการลักลอกได้ดีกว่าลักษณ์ที่ไม่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ ด้วยเหตุนี้ ผู้วิจัยจึงต้องการทดลองรวมชุดของลักษณ์ทางภาษาและนำไปฝึกฝนและทดสอบด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนเพื่อประเมินประสิทธิภาพของลักษณ์ทางภาษา

ในการรวมชุดของลักษณ์ทางภาษานั้น ผู้วิจัยจะใช้ลักษณ์ 3 ตัวตามระดับของหน่วยภาษา 3 กลุ่ม ได้แก่ ลักษณ์ทางศัพท์ ลักษณ์ทางวายสัมพันธ์ และลักษณ์ทางความหมาย โดยจะนำลักษณ์ที่มีประสิทธิภาพสูงที่สุดของแต่ละกลุ่มจากการทดลองในหัวข้อที่ 3.5.4.1 กลุ่มละ 1 ตัว มาจัดเป็นชุดทั้งหมด 4 ชุด ได้แก่

1) *ลักษณ์ทางภาษาชุดที่ 1 (LF_1)* ประกอบด้วยลักษณ์ทางศัพท์และลักษณ์ทางวายสัมพันธ์
2) *ลักษณ์ทางภาษาชุดที่ 2 (LF_2)* ประกอบด้วยลักษณ์ทางศัพท์และลักษณ์ทางความหมาย
3) *ลักษณ์ทางภาษาชุดที่ 3 (LF_3)* ประกอบด้วยลักษณ์ทางวายสัมพันธ์และลักษณ์ทางความหมาย

4) *ลักษณ์ทางภาษาชุดที่ 4 (LF_4)* ประกอบด้วยลักษณ์ทางศัพท์ ลักษณ์ทางวายสัมพันธ์ และลักษณ์ทางความหมาย

เมื่อนำลักษณ์ทางภาษาทั้ง 4 ชุดไปฝึกฝนและทดสอบประสิทธิภาพด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนตามขั้นตอนในหัวข้อที่ 3.5.2 แล้ว ผู้วิจัยประเมินประสิทธิภาพของลักษณ์ทางภาษาแต่ละชุดโดยพิจารณาจากค่าการประเมินประสิทธิภาพตามที่กล่าวไว้ในหัวข้อที่ 3.5.1 จากนั้นจึงอภิปรายผลการประเมิน

3.6 การเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความ

การเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความเป็นวัตถุประสงค์ประการสุดท้ายของงานวิจัยชิ้นนี้ ทั้งนี้ ดังได้กล่าวไปในหัวข้อที่ 3.1 แล้วว่าระบบตรวจหาการลักลอกงานวิชาการที่ออกแบบไว้นั้นประกอบด้วยขั้นตอนการตรวจหา 2 ขั้น ขั้นแรกเป็นการตรวจหาโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก ซึ่งสามารถประเมินประสิทธิภาพได้ดังได้กล่าวไปแล้วในหัวข้อที่ 3.5.3 และ 3.5.4 ส่วนอีกขั้นนั้นเป็นการวัดค่าความละม้ายของข้อความที่ระบบขั้นแรกจำแนกประเภทได้ว่าเป็นการลักลอก ด้วยระบบในขั้นแรกนั้นไม่สามารถคืนค่าเป็นตัวเลขบ่งชี้ปริมาณการลักลอกในเอกสารได้ จึงจำเป็นตัววัดค่าความละม้ายระหว่างข้อความต้นฉบับและข้อความลักลอกในขั้นที่ 2 อย่างไรก็ตาม ยังไม่ปรากฏผลการศึกษาที่ชี้ชัดว่าวิธีการวัดค่าความละม้ายวิธีใดเป็นวิธีที่มีประสิทธิภาพมากที่สุด ด้วยเหตุนี้จึงจำเป็นต้องเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความเพื่อหาวิธีการวัดค่าความละม้ายของข้อความที่มีประสิทธิภาพ สามารถใช้แทนมนุษย์ได้ มาใช้ในระบบ

แนวคิดสำคัญที่ใช้ในการเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความในขั้นนี้คือการนำค่าที่ได้วิธีการวัดแต่ละวิธีมาเปรียบเทียบกับผลการวัดค่าความละม้ายโดยมนุษย์ หรือกล่าวอีกนัยหนึ่งคือใช้ค่าความละม้ายที่ให้โดยมนุษย์เป็นบรรทัดฐาน หากวิธีการวัดค่าความละม้ายวิธีใดให้ค่าได้ใกล้เคียงกับค่าความละม้ายที่ให้โดยมนุษย์มากที่สุด จะถือว่าวิธีการวัดค่าความละม้ายวิธีดังกล่าวมีประสิทธิภาพเหมาะสมจะใช้ในระบบตรวจหาการลักลอกมากที่สุด

เพื่อให้บรรลุตามแนวคิดข้างต้น ในขั้นแรกของการทดลอง ผู้วิจัยจะให้ผู้เชี่ยวชาญด้านภาษาไทย 3 คนเป็นผู้ระบุค่าความละม้ายของคู่หน่วยเทียบในชุดข้อมูลทดลอง โดยผู้เชี่ยวชาญต้องมีคุณสมบัติคือต้องสำเร็จการศึกษาระดับปริญญาโทขึ้นไปทางภาษาไทยหรือภาษาศาสตร์ และสอนวิชาเกี่ยวกับการใช้ภาษาไทยในระดับอุดมศึกษา ส่วนชุดข้อมูลทดลองนั้น ผู้วิจัยจะใช้ชุดข้อมูลทดลอง PRG ที่สร้างขึ้นตามขั้นตอนที่ระบุไว้ในหัวข้อที่ 3.5.3.1 ทั้งนี้ ผู้วิจัยได้แก้ไขชุดข้อมูลทดลองดังกล่าว

เล็กน้อยโดยสุ่มลำดับของคู่หน่วยเทียบใหม่เพื่อไม่ให้ประเภทการลักลอกและขนาดของคู่หน่วยเทียบที่เหมือนกันอยู่ติดกัน ซึ่งอาจส่งผลให้ให้ผู้เชี่ยวชาญคาดเดาค่าความละม้ายได้ตามรูปแบบประเภทของการลักลอก

ผู้เชี่ยวชาญทางภาษาไทยทั้ง 3 คนจะได้รับมอบหมายให้อ่านคู่หน่วยเทียบในชุดข้อมูลทดลองและระบุค่าความละม้ายเป็นร้อยละตั้งแต่ 0 ถึง 100 ซึ่งบ่งชี้ถึงความไม่ละม้ายกันและความเหมือนกันทุกประการ โดยมีคำชี้แจงกำกับไว้ดังนี้

> ในส่วนต่อไปนี้ ผู้วิจัยได้เตรียมคู่ของย่อหน้าไว้จำนวน 150 คู่ ให้ท่านอ่านย่อหน้าเปรียบเทียบกันภายในคู่และพิจารณาว่าแต่ละคู่มีความเหมือนกันมากน้อยเพียงใด จากนั้น **ให้ท่านระบุตัวเลขตั้งแต่ 0% ถึง 100%** โดย 0% หมายถึงคู่ของย่อหน้าไม่เหมือนกันเลย ส่วน 100% หมายถึงคู่ของย่อหน้าเหมือนกันทุกประการ

เมื่อได้ค่าความละม้ายจากผู้เชี่ยวชาญครบทุกคนแล้ว ในขั้นต่อมา ผู้วิจัยจะนำค่าความละม้ายของที่ได้จากผู้เชี่ยวชาญทุกคนมาทดสอบความเป็นเอกพันธ์ (test of homogeneity) ทั้งนี้ หากพบว่าค่าความละม้ายของข้อความที่ได้จากผู้เชี่ยวชาญทุกคนมีความเป็นเอกพันธ์ ผู้วิจัยจะหาค่าเฉลี่ยของค่าความละม้ายที่ให้โดยผู้เชี่ยวชาญทั้ง 3 คนในแต่ละคู่หน่วยเทียบแล้วนำไปเปรียบเทียบกับค่าความละม้ายที่ได้จากวิธีวัดค่าความละม้ายวิธีต่างๆ แต่ถ้าหากพบว่าค่าความละม้ายของที่ได้จากผู้เชี่ยวชาญทุกคนไม่มีความเป็นเอกพันธ์ ผู้วิจัยจะนำค่าความละม้ายที่ให้โดยผู้เชี่ยวชาญแต่ละคนมาเปรียบเทียบกับค่าความละม้ายที่ได้จากวิธีวัดค่าความละม้ายวิธีต่างๆ เป็นรายบุคคลแล้วหาค่าเฉลี่ย

ส่วนค่าความละม้ายที่ได้จากวิธีวัดวิธีต่างๆ นั้น ผู้วิจัยจะใช้ค่าความละม้ายทุกประเภทที่วิเคราะห์ได้และถูกใช้เป็นลักษณ์ในการจำแนกประเภทข้อความลักลอกและไม่ลักลอกตามขั้นตอนในหัวข้อที่ 3.4 และ 3.5.4 ตามลำดับ ค่าความละม้ายเหล่านี้จะให้ระหว่าง 0 กับ 1 ซึ่ง หากค่าเท่ากับ 0 จะบ่งชี้ถึงความไม่ละม้ายกันเลยของคู่หน่วยเทียบ และหากค่าเท่ากับ 1 จะบ่งชี้ว่าคู่หน่วยเทียบของย่อหน้าเหมือนกันทุกประการ โดยในขั้นนี้ ผู้วิจัยได้จะแบ่งกลุ่มวิธีการวัดค่าความละม้ายเป็น 5 กลุ่มตามลำดับขั้นของหน่วยทางภาษา เพื่อให้สอดคล้องกับการแนวทางการวิเคราะห์หาลักษณ์ และเพื่อให้สะดวกแก่การวิเคราะห์และอภิปรายผล ดังนี้

1) วิธีวัดค่าความละม้ายอิงอักขระ
2) วิธีวัดค่าความละม้ายอิงคำและศัพท์
3) วิธีวัดค่าความละม้ายอิงวายสัมพันธ์
4) วิธีวัดค่าความละม้ายอิงความหมาย

5) วิธีวัดค่าความละม้ายอิงวายสัมพันธ์และความหมาย

โดยอาศัยแนวคิดข้างต้น ผู้วิจัยจะใช้โปรแกรมที่ได้เขียนไว้ก่อนแล้วในขั้นตอนการวิเคราะห์และสกัดหาลักษณ์มาวัดค่าความละม้ายของชุดข้อมูลทดลอง PRG และนำค่าที่ได้มาเปรียบเทียบค่าความละม้ายของข้อความที่ได้จากผู้เชี่ยวชาญ

ส่วนขั้นตอนการเปรียบเทียบค่าความละม้ายที่ได้จากผู้เชี่ยวชาญกับค่าความละม้ายที่ได้จากวิธีวัดวิธีต่างๆ นั้น ผู้วิจัยจะนำค่าความละม้ายที่วัดได้จากแต่ละวิธีวัดค่าความละม้ายวิธีต่างๆ กับค่าความละม้ายที่ประเมินโดยผู้เชี่ยวชาญมาหาวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์ (correlation analysis) เพื่อหาความสัมพันธ์ระหว่างค่าความละม้ายที่วัดได้จากแต่ละวิธีการวัดกับค่าความละม้ายที่ประเมินโดยผู้เชี่ยวชาญ หากผลปรากฏว่าค่าความละม้ายที่วัดได้จากวิธีวัดค่าความละม้ายวิธีใดสอดคล้องสัมพันธ์กับค่าความละม้ายที่ประเมินโดยผู้เชี่ยวชาญมากที่สุด จะถือว่าวิธีวัดค่าความละม้ายดังกล่าวมีประสิทธิภาพ เหมาะสมจะใช้ระบุค่าความละม้ายแทนมนุษย์ในระบบตรวจหาการลักลอกงานวิชาการได้

บทที่ 4
กลวิธีลักลอกงานวิชาการภาษาไทย[13]

ดังได้กล่าวไปในแล้วว่าในขั้นตอนการสร้างคลังข้อมูลการลักลอก การวิเคราะห์หาลักษณ์ที่ใช้สำหรับจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอก การกำหนดลักษณะของข้อมูลรับเข้าตลอดจนการอภิปรายผลที่ได้จากใช้งานระบบตรวจหาการลักลอกที่พัฒนาขึ้นนั้น จำเป็นอย่างยิ่งที่จะต้องอาศัยองค์ความรู้เกี่ยวกับกลวิธีลักลอกงานวิชาการภาษาไทยในเชิงภาษาศาสตร์ ดังนั้นในบทนี้ผู้วิจัยจะนำเสนอผลการวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทย ซึ่งประกอบไปด้วยเนื้อหาสำคัญ 4 ส่วน ได้แก่ ประเภทของกลวิธีลักลอก ปริมาณการใช้กลวิธีลักลอก รูปแบบการใช้กลวิธีลักลอก และการประยุกต์ข้อค้นพบที่ได้ในงานวิจัยชิ้น

ทั้งนี้ จากการคัดแยกย่อหน้าที่ไม่ถือว่าเป็นการลักลอกในขั้นต้นของการวิเคราะห์ข้อมูลตามวิธีการที่ได้ระบุไว้ในหัวข้อที่ 3.2.2 ทำให้เหลือย่อหน้าที่จะนำมาวิเคราะห์กลวิธีลักลอกจำนวน 188 ย่อหน้า หรือคิดเป็นร้อยละ 75.2 ของย่อหน้าที่มีผู้ตอบแบบสอบถามเขียนขึ้นใหม่จากย่อหน้าต้นฉบับที่กำหนดให้ จากนั้นผู้วิจัยได้ดำเนินการวิเคราะห์กลวิธีลักลอกตามวิธีการวิเคราะห์ที่ได้กำหนดไว้ กระทั่งได้ข้อค้นพบดังจะกล่าวต่อไปนี้

4.1 ประเภทของกลวิธีลักลอก

จากการวิเคราะห์คู่ลักลอก พบว่าผู้ลักลอกใช้กลวิธีลักลอกทั้งหมด 15 ประเภท ซึ่งแบ่งได้เป็นกลวิธีลักลอกภายในหน่วยปริจเฉทพื้นฐาน 5 ประเภท และกลวิธีลักลอกระหว่างหน่วยปริจเฉทพื้นฐาน 10 ประเภท ทั้งนี้ กลวิธีลักลอกดังกล่าวมีรายละเอียดดังนี้

4.1.1 กลวิธีลักลอกภายในหน่วยปริจเฉทพื้นฐาน

กลวิธีลักลอกภายในหน่วยปริจเฉทพื้นฐานนี้เป็นกลวิธีที่ผู้ลักลอกใช้เพื่อเปลี่ยนแปลงองค์ประกอบต่างๆ ภายในหน่วยปริจเฉทพื้นฐาน กลวิธีประเภทนี้ประกอบด้วยกลวิธีย่อย 5 ประเภท ได้แก่

---

[13] ข้อค้นพบส่วนหนึ่งที่แสดงในบทนี้ได้รับการตีพิมพ์เผยแพร่แล้วในบทความเรื่อง “กลวิธีลักลอกงานวิชาการภาษาไทย: การวิเคราะห์ทางภาษาศาสตร์” (ศุภวัจน์ แต่รุ่งเรือง และวิโรจน์ อรุณมานะกุล, 2558)

1) การแทนที่ในขั้วเดียวกัน

การแทนที่ในขั้วเดียวกัน หมายถึง การแทนที่หน่วยใดๆ ในข้อความต้นฉบับด้วยหน่วยอื่น แต่หน่วยดังกล่าวยังคงคุณสมบัติเช่นเดียวกันกับหน่วยที่ถูกแทนที่ไป ในที่นี้ได้แก่ การแทนที่ด้วยหน่วยทางศัพท์ (lexical unit) หรือหน่วยทางหน้าที่ (functional unit) เพื่อคงความหมายที่ใกล้เคียงกันของคู่ลักลอกไว้ กลวิธีนี้อาจปรากฏในรูปการแทนที่ด้วยคำไวพจน์ การแทนที่ด้วยคำที่มีความหมายทั่วไปกับคำที่มีความหมายเฉพาะ การแทนที่ด้วยสรรพนามหรือรูปแสดงการอ้างถึง ดังจะเห็นได้จากตัวอย่างที่ 4.1 ที่แสดงการแทนที่ด้วยหน่วยทางศัพท์ในขั้วเดียว และตัวอย่างที่ 4.2 ที่แสดงการแทนที่ด้วยหน่วยทางหน้าที่ชนิดเดียวกัน ซึ่งในที่นี้ได้แก่คำเชื่อม

(4.1) (ก)$_{src}$ [โดยเฉพาะ*คน*ในเมืองใหญ่][14]

(ข)$_{plg}$ [โดยเฉพาะ*มนุษย์*ในเมืองใหญ่]

(4.2) (ก)$_{src}$ [*นอกจากนี้*ป่าชายเลนยังเป็นแหล่งกักเก็บคาร์บอน]

(ข)$_{plg}$ [*และ*ป่าชายเลนยังเป็นแหล่งกักเก็บคาร์บอน]

2) การแทนที่ในขั้วตรงกันข้าม

การแทนที่ในขั้วตรงกันข้าม หมายถึง การแทนที่หน่วยใดๆ ในข้อความต้นฉบับด้วยหน่วยอื่น โดยหน่วยดังกล่าวจะมีคุณสมบัติตรงกันข้ามกับหน่วยเดิมที่ถูกแทนที่ ในที่นี้ได้แก่ การแทนที่ด้วยหน่วยทางศัพท์หรือหน่วยทางหน้าที่ เพื่อคงความหมายที่ใกล้เคียงกันของคู่ลักลอกไว้ โดยจะปรากฏร่วมกับคำแสดงการปฏิเสธ ดังตัวอย่างที่ 4.3

(4.3) (ก)$_{src}$ [จะมีรูปร่างหน้าตา*น่าเกลียดน่ากลัว*]

(ข)$_{plg}$ [หน้าตาจะ*ไม่งาม*]

3) การแทรก

การแทรก คือ การเพิ่มเติมหน่วยทางศัพท์หรือหน่วยทางหน้าที่ เช่นในตัวอย่างที่ 4.4 มีการแทรกคำนาม “ก๊าซธรรมชาติ” ที่เดิมไม่มีอยู่ในต้นฉบับ พร้อมกันนั้นยังเป็นการแทรกหน่วยที่ทำหน้าที่เป็นประธานด้วย อย่างไรก็ตาม การใช้กลวิธีนี้อาจมีผลให้ความหมายเปลี่ยนแปลงไปจากเดิมบ้างเล็กน้อย ยกตัวอย่างเช่นในกรณีที่มีการแทรกคำขยายหรือแทรกคำช่วยหน้ากริยา เช่น ควร จะ อาจจะ เป็นต้น

[14] เครื่องหมาย [...] ในที่นี้ใช้เพื่อระบุขอบเขตของหน่วยปริจเฉทพื้นฐานแต่ละหน่วย

(4.4) (ก)$_{src}$ [และเป็นพลังงานอีกทางเลือกหนึ่ง]

(ข)$_{plg}$ [และ*ก๊าซธรรมชาติ*เป็นพลังงานอีกทางเลือกหนึ่ง]

4) การลบและการละ

กลวิธีลักลอกนี้ได้แก่การลบหรือละหน่วยทางศัพท์หรือหน่วยทางหน้าที่ออกไป เช่นในตัวอย่างที่ 4.5 ที่แสดงการลบคำนาม “บ้านเรือน” ออกไป หรือในตัวอย่างที่ 4.6 ที่แสดงให้เห็นถึงการละหน่วยที่ทำหน้าที่กรรมไปในหน่วยปริจเฉทพื้นฐานที่ 2

(4.5) (ก)$_{src}$ [มลพิษทางอากาศภายในอาคาร *บ้านเรือน* เป็นปัญหาหนึ่ง]

(ข)$_{plg}$ [มลพิษทางอากาศภายในอาคารเป็นปัญหาหนึ่ง]

(4.6) (ก)$_{src}$ [จึงทำให้มีพฤติกรรมต่อต้าน]$_1$ [ขัดขวาง*การฉีดยาชา*]$_2$

(ข)$_{plg}$ [จึงทำให้มีพฤติกรรมต่อต้าน]$_1$ [ขัดขวาง]$_2$

5) การเปลี่ยนลำดับ

การเปลี่ยนลำดับในที่นี้ได้แก่ การเปลี่ยนลำดับของคำ วลี หรืออนุพากย์ ที่ทำหน้าที่ต่างๆ ภายในหน่วยปริจเฉทพื้นฐาน เช่นในตัวอย่างที่ 4.7 ซึ่งแสดงให้เห็นถึงการเปลี่ยนแปลงลำดับคำในหน่วยปริจเฉทพื้นฐาน หรือในตัวอย่างที่ 4.8 ที่แสดงให้เห็นการสลับตำแหน่งกันระหว่างประธานกับส่วนเติมเต็มของประธานโดยอาศัยคุณสมบัติของสัมพันธกริยา (copula) “เป็น” และ “คือ”

(4.7) (ก)$_{src}$ [ชอบกิน*เนื้อมนุษย์*และ*ซากศพ*]

(ข)$_{plg}$ [ชอบกิน*ซากศพ*และ*เนื้อมนุษย์*]

(4.8) (ก)$_{src}$ [*การฉีดยาชาเฉพาะที่*เป็น*สิ่ง*]$_1$ [ที่ทันตแพทย์เกือบทั้งหมดทำ]$_2$

(ข)$_{plg}$ [*สิ่ง*]$_{1.1}$ [ที่ทันตแพทย์เกือบทั้งหมดทำ]$_2$ [คือ*การฉีดยาชาเฉพาะที่*]$_{1.2}$

4.1.2 กลวิธีลักลอกระหว่างหน่วยปริจเฉทพื้นฐาน

กลวิธีลักลอกระหว่างหน่วยปริจเฉทพื้นฐาน ได้แก่ กลวิธีลักลอกที่ผู้ลักลอกใช้เพื่อเปลี่ยนแปลงความสัมพันธ์ระหว่างปริจเฉทพื้นฐานที่มีอยู่เดิมโดยอาจใช้กลไกทางภาษาในด้านคำศัพท์หรือวากยสัมพันธ์เป็นเครื่องมือ กลวิธีประเภทนี้ประกอบด้วยกลวิธีย่อย 10 ประเภท ดังนี้

1) การซ้อนความ

การซ้อนความ (subordination) ในที่นี้ได้แก่ การทำให้หน่วยปริจเฉทพื้นฐานที่เดิมเป็นอนุพากย์อิสระ (independent clause) กลายเป็นอนุพากย์ไม่อิสระ (dependent clause) กล่าวคือ

ทำให้กลายเป็นอนุพากย์ซ้อน (embedded clause) กลวิธีนี้อาจเกิดขึ้นกับกริยาตัวใดตัวหนึ่งในหน่วยปริจเฉทพื้นฐานดังได้แสดงให้เห็นในตัวอย่างที่ 4.9 หรือเกิดขึ้นกับหน่วยปริจเฉทพื้นฐานทั้งหน่วยเช่นในตัวอย่างที่ 4.10 ก็ได้

(4.9) (ก)$_{src}$ [จะมีรูปกาย*งดงาม*]$_1$

(ข)$_{plg}$ [จะมีรูปกาย]$_1$ [*ที่งดงาม*]$_2$

(4.10) (ก)$_{src}$ [มีบริษัทผลิตละครโทรทัศน์รายใหญ่]$_1$ [*เป็นที่รู้จักของประชาชนอยู่หลายบริษัท*]$_2$

(ข)$_{plg}$ [มีบริษัทรายใหญ่]$_{1.1}$ [*ที่เป็นที่รู้จักของประชาชน*]$_2$ [หลายบริษัท]$_{1.2}$

2) การยุบเลิกความซ้อน

การยุบเลิกความซ้อน คือ การทำให้หน่วยปริจเฉทพื้นฐานซึ่งเดิมเป็นอนุพากย์ซ้อนกลายเป็นอนุพากย์อิสระ ยกตัวอย่างเช่นในตัวอย่างที่ 4.11 เป็นต้น

(4.11) (ก)$_{src}$ [ถ้าการได้รับมลพิษในอากาศของประชากรในเขตเมืองจะเกิดขึ้นภายในอาคารมากกว่า]$_1$ [*ที่เกิดขึ้นขณะดำเนินกิจกรรมอยู่ภายนอกอาคาร*]$_2$

(ข)$_{plg}$ [หากการได้รับมลพิษในอากาศของประชากรในเขตเมืองเกิดขึ้นภายในอาคารมากกว่า*การทำกิจกรรมอยู่ภายนอกอาคาร*]$_1$

3) การเชื่อมความ

การเชื่อมความ (coordination) เป็นการเชื่อมหน่วยปริจเฉทพื้นฐานมากกว่า 1 หน่วยที่เดิมเป็นอนุพากย์ใดๆ ที่เป็นอิสระต่อกันให้กลายเป็นอนุพากย์ความรวม (coordinate clause) ที่มีความเท่าเทียมกันในเชิงหน้าที่โดยแทรกคำเชื่อมเข้าไปให้ชัดเจน ดังจะเห็นได้จากตัวอย่างที่ 4.12 ที่กล่าวถึงประโยชน์ของการฝึกสุนัข เป็นต้น

(4.12) (ก)$_{src}$ [ดมกลิ่นระเบิด]$_1$ [ตรวจค้นหาสารเสพติด เป็นต้น]$_2$

(ข)$_{plg}$ [ค้นหาวัตถุระเบิด]$_1$ [หรือตรวจหาสารเสพติด เป็นต้น]$_2$

4) การยุบเลิกความรวม

การยุบเลิกความรวม ได้แก่ การทำให้หน่วยปริจเฉทพื้นฐานซึ่งเดิมเป็นอนุพากย์ความรวมกลายเป็นอนุพากย์ใดๆ ซึ่งไม่ได้แสดงความเท่าเทียมกันในเชิงหน้าที่ ดังตัวอย่างที่ 4.13 ผู้ลักลอกได้ยุบหน่วยปริจเฉทพื้นฐานที่ 2 ซึ่งเป็นความรวมที่มีความเท่าเทียมกันในเชิงหน้าที่กับหน่วยปริจเฉทพื้นฐานที่ 1 ในข้อความต้นฉบับ (ก) ลงโดยนำมาซ้อนความไว้ด้วยกัน ดังเช่นที่ปรากฏให้เป็นหน่วยปริจเฉทพื้นฐานที่ 2 ในข้อความลักลอก (ข)

(4.13) (ก)$_{src}$ [เพราะนอกจากจะ*เป็นสาเหตุสำคัญของปัญหาครอบครัวแตกแยก*]$_1$ [*และนำมาซึ่งปัญหาสังคมอีกมากมายแล้ว*]$_2$

(ข)$_{plg}$ [เพราะเป็นสาเหตุ]$_1$ [*ที่ทำให้เกิดปัญหาครอบครัวแตกแยกและปัญหาสังคมอื่นๆ*]$_2$

5) การแปลงเป็นนามวลี

การแปลงให้เป็นนามวลี คือ กลวิธีที่ผู้ลักลอกใช้เพื่อแปลงหน่วยหน่วยปริจเฉทพื้นฐานให้กลายเป็นนามวลี ดังจะเห็นได้จากตัวอย่างที่ 4.14 ที่แสดงให้เห็นการแปลงหน่วยปริจเฉทพื้นฐานที่ 3 ในข้อความ (ก) ให้กลายเป็นนามวลีแล้วนำไปรวมกับหน่วยปริจเฉทพื้นฐานที่ 2 จนได้เป็นหน่วยปริจเฉทพื้นฐานที่ 2 ในข้อ (ข)

(4.14) (ก)$_{src}$ [รวมไปถึงอาจมีประสบการณ์]$_1$ [ที่ไม่ดีจากการฉีดวัคซีน]$_2$ [*เจาะเลือด*]$_3$

(ข)$_{plg}$ [รวมไปถึงอาจมีประสบการณ์]$_1$ [ที่ไม่ดีจากการฉีดวัคซีนหรือ*การเจาะเลือด*]$_2$

6) การแปลงเป็นส่วนเติมเต็ม

การแปลงเป็นส่วนเติม คือ การแปลงหน่วยปริจเฉทพื้นฐานหนึ่งให้กลายเป็นส่วนเติมเต็มของอีกหน่วยปริจเฉทพื้นฐานหนึ่ง ดังได้แสดงให้เห็นในตัวอย่างที่ 4.15 ซึ่งจะเห็นได้ว่ามีการแปลงหน่วยปริจเฉทพื้นฐานที่ 2 ในข้อ (ก) ให้เป็นบุพบทวลีเพื่อนำมาใช้เป็นส่วนเต็มเติมในข้อ (ข)

(4.15) (ก)$_{src}$ [พลังงานจากก๊าซธรรมชาติมีบทบาทมากขึ้น]$_1$ [โดยเป็นแนวทางหนึ่งในการลดการใช้พลังงานจากน้ำมัน]$_2$

(ข)$_{plg}$ [พลังงานจากก๊าซธรรมชาติมีบทบาทมากขึ้น*ในการลดการใช้พลังงานจากน้ำมันมาเป็นตัวเลือก*]$_1$

7) การเพิ่มเติมเนื้อหา

การเพิ่มเติมเนื้อหา คือ การเพิ่มเติมเนื้อหาที่ไม่มีอยู่ในข้อความต้นฉบับลงไปในข้อความลักลอก โดยข้อมูลดังกล่าวอาจได้จากประสบการณ์หรือความรู้อันเป็นภูมิหลังของผู้ลักลอกเอง ทั้งนี้ ดังได้กล่าวไปในตอนต้นแล้วว่า หน่วยปริจเฉทพื้นฐาน 1 หน่วยสามารถใช้แทนข้อมูลเชิงเนื้อหาที่สมบูรณ์ได้ 1 ข้อมูล ฉะนั้นแล้ว หากในขั้นการวิเคราะห์พบว่า ข้อความลักลอกมีจำนวนหน่วยปริจเฉทพื้นฐานที่ไม่สามารถจับคู่เทียบกับหน่วยปริจเฉทพื้นฐานในต้นฉบับได้มากเท่าใดก็แสดงว่าผู้ลักลอกเพิ่มเติมเนื้อหาขึ้นเองมากเท่านั้น ดังตัวอย่างที่ 4.16 จะเห็นได้ว่าผู้ลักลอกไม่ได้เปลี่ยนแปลงเนื้อหาในต้นฉบับ (ก) เลย แต่เลือกใช้กลวิธีเพิ่มเติมเนื้อหาภายนอกเข้ามาเป็นจำนวนมากถึง 9 ข้อมูล

(4.16) (ก)$_{src}$ [ในวรรณคดีสันสกฤต]$_1$ [ยักษ์เป็นอมนุษย์]$_2$ [ที่มีรูปร่างหน้าตาอัปลักษณ์]$_3$ [ดุร้าย]$_4$ [ชอบกินเนื้อมนุษย์และซากศพ]$_5$ [แต่ในทางพุทธศาสนา]$_6$ [ยักษ์มีทั้งพวก]$_{7.1}$ [ที่มีรูปกายงดงาม]$_8$ [และพวก]$_{7.2}$ [ที่มีรูปร่างหน้าตาอัปลักษณ์]$_9$ [โดยยักษ์]$_{10.1}$ [ที่เคยสร้างกุศล]$_{11}$ [จะมีรูปกายงดงาม]$_{10.2}$ [ทรงสง่าราศี]$_{12}$ [ส่วนยักษ์]$_{13.1}$ [ที่เคยสร้างบาปกรรม]$_{14}$ [จะมีรูปร่างหน้าตาน่าเกลียดน่ากลัว]$_{13.2}$

(ข)$_{plg}$ [ในวรรณคดีสันสกฤต]$_1$ [ยักษ์เป็นอมนุษย์]$_2$ [ที่มีรูปร่างหน้าตาอัปลักษณ์]$_3$ [ดุร้าย]$_4$ [ชอบกินเนื้อมนุษย์และซากศพ]$_5$ [แต่ในทางพุทธศาสนา]$_6$ [ยักษ์มีทั้งพวก]$_{7.1}$ [ที่มีรูปกายงดงาม]$_8$ [และพวก]$_{7.2}$ [ที่มีรูปร่างหน้าตาอัปลักษณ์]$_9$ [โดยยักษ์]$_{10.1}$ [ที่เคยสร้างกุศล]$_{11}$ [จะมีรูปกายงดงาม]$_{10.2}$ [ทรงสง่าราศี]$_{12}$ [ส่วนยักษ์]$_{13.1}$ [ที่เคยสร้างบาปกรรม]$_{14}$ [จะมีรูปร่างหน้าตาน่าเกลียดน่ากลัว]$_{13.2}$ *[นั้นคือ ยักษ์ในวรรณคดี]$_{15}$ [แต่เทียบในปัจจุบัน]$_{16}$ [พุทธศาสนาสั่งสอนให้ทุกศาสนาเป็นคนดี]$_{17}$ [ทำความดี]$_{18}$ [ละเว้นความชั่ว]$_{19}$ [สร้างสมบุญเก่า]$_{20}$ [เกิดมาอีกชาติภพอีกครั้ง]$_{21}$ [ก็จะส่งผลทำให้มนุษย์เกิดมาตามบุญและกรรม]$_{22}$ [ที่ทำมาแต่ชาติปางก่อน]$_{23}$*

8) การตัดทอนเนื้อหา

การตัดทอนเนื้อหา คือ กลวิธีที่ผู้ลักลอกใช้เพื่อตัดข้อมูลที่เป็นเนื้อหาบางส่วนที่มีอยู่เดิมในข้อความต้นฉบับออกไป โดยเป็นการตัดหน่วยปริจเฉทพื้นฐานออกไปทั้งหน่วย ดังตัวอย่างที่ 4.17 ซึ่งจะเห็นได้ว่าผู้ลักลอกตัดหน่วยปริจเฉทพื้นฐานที่ 3 ในข้อความต้นฉบับ (ก) ออกไป

(4.17)

(ก)$_{src}$	(ข)$_{plg}$
[ความรุนแรงในสังคมและความรุนแรงในครอบครัวมีความคล้ายคลึงกัน]$_1$	[ความรุนแรงในสังคมและความรุนแรงในครอบครัวมีความคล้ายคลึงกัน]$_1$
[ที่เหยื่อ คือผู้อ่อนแอและมีสถานภาพต่ำกว่า]$_2$	[คือ เหยื่อเป็นผู้อ่อนแอและสถานภาพต่ำกว่า]$_2$
[เช่น สตรี เด็ก ฯลฯ]$_3$	*Ø*
[แต่ผลกระทบของความรุนแรงในครอบครัวรุนแรงและยาวไกลกว่าความรุนแรงในสังคม]$_4$	[แต่ความรุนแรงในครอบครัวร้ายแรงกว่าความรุนแรงในสังคม]$_3$

9) การย้ายลำดับเนื้อหา

การย้ายลำดับเนื้อหา คือ การเรียบเรียงเนื้อหาของข้อความลักลอกให้ต่างไปจากเดิมในข้อความต้นฉบับ โดยย้ายลำดับของหน่วยปริจเฉทพื้นฐานไปไว้ด้านหน้าหรือหลังตำแหน่งเดิม ขอให้สังเกตจากตัวอย่างที่ 4.18 ที่หน่วยปริจเฉทพื้นฐานที่ 6 และ 7 ในข้อความต้นฉบับ (ก) ถูกย้ายลำดับไปเป็นหน่วยปริจเฉทพื้นฐานที่ 5 และ 4 ในข้อความลักลอก (ข) ตามลำดับ

(4.18) (ก)$_{src}$ [การเรียนกวดวิชาแม้จะก่อให้เกิดประโยชน์]$_1$ [แต่การเรียนกวดวิชาก็ได้ก่อให้เกิดผลกระทบในด้านลบพอสมควร]$_2$ [โดยเฉพาะการสร้างภาระแก่ผู้ปกครองอย่างมาก]$_3$ [นอกจากการสร้างภาระค่าใช้จ่ายกับผู้ปกครองแล้ว]$_4$ [การกวดวิชายังก่อให้เกิดผลกระทบต่อนักเรียน]$_5$ *[เนื่องจากนักเรียนต้องใช้เวลาว่างไปกับการเรียนกวดวิชา]$_6$ [ก่อให้เกิดความห่างเหินระหว่างเด็กกับผู้ปกครอง]$_7$* [มีผลต่อความอบอุ่นในครอบครัว]$_8$

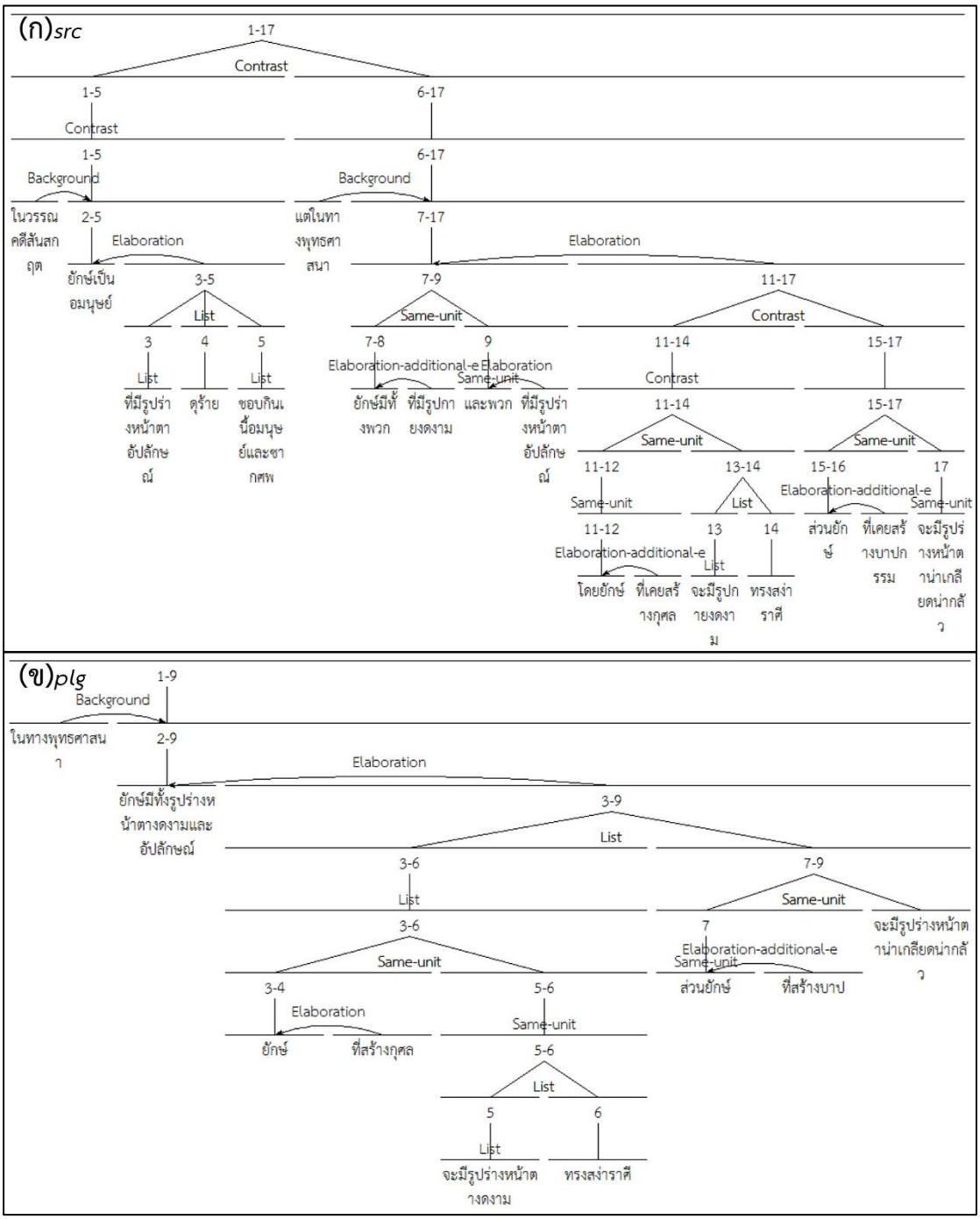
(ข)$_{plg}$ [การเรียนกวดวิชาแม้จะมีประโยชน์]$_1$ [แต่ขณะเดียวกันก็ก่อให้เกิดผลในด้านลบด้วย]$_2$ [เพราะสร้างภาระค่าใช้จ่ายแก่ผู้ปกครอง]$_3$ *[อีกทั้งยังก่อให้เกิดความเหินห่างระหว่างเด็กกับผู้ปกครอง]$_4$ [เพราะเด็กต้องใช้เวลาว่างไปกับการเรียนกวดวิชาแทนการใช้เวลาร่วมกับคนในครอบครัว]$_5$*

10) การเปลี่ยนแปลงความสัมพันธ์

การเปลี่ยนแปลงความสัมพันธ์ ได้แก่ การเปลี่ยนสถานะความสำคัญ (nuclearity) และการเปลี่ยนวาทสัมพันธ์ (rhetorical relation) ที่มีอยู่ระหว่างหน่วยปริจเฉทพื้นฐาน ซึ่งเป็นผลเนื่องมากจากการแทรกหรือแทนที่คำเชื่อมในหน่วยปริจเฉทพื้นฐาน การเพิ่มเติมเนื้อหา การตัดทอนเนื้อหา หรือการย้ายลำดับเนื้อหา ทั้งนี้ เมื่อความสัมพันธ์ระหว่างหน่วยปริจเฉทพื้นฐานในข้อความต้นฉบับถูกเปลี่ยนแปลงไปก็จะมีผลทำให้เจตนาเดิมที่ผู้เขียนต้องการถ่ายทอดสู่ผู้อ่านเปลี่ยนแปลงไปด้วย ดังตัวอย่างที่ 4.19 ที่แสดงให้เห็นการเปลี่ยนแปลงความสัมพันธ์อันเป็นผลสืบเนื่องมาจากการใช้กลวิธีตัดทอนเนื้อหาเป็นกลวิธีหลักในการลักลอก

(4.19) (ก)$_{src}$ [ในวรรณคดีสันสกฤต]$_1$ [ยักษ์เป็นอมนุษย์]$_2$ [ที่มีรูปร่างหน้าตาอัปลักษณ์]$_3$ [ดุร้าย]$_4$ [ชอบกินเนื้อมนุษย์และซากศพ]$_5$ [แต่ในทางพุทธศาสนา]$_6$ [ยักษ์มีทั้งพวก]$_{7.1}$ [ที่มีรูปกายงดงาม]$_8$ [และพวก]$_{7.2}$ [ที่มีรูปร่างหน้าตาอัปลักษณ์]$_9$ [โดยยักษ์]$_{10.1}$ [ที่เคยสร้างกุศล]$_{11}$ [จะมีรูปกายงดงาม]$_{10.2}$ [ทรงสง่าราศี]$_{12}$ [ส่วนยักษ์]$_{13.1}$ [ที่เคยสร้างบาปกรรม]$_{14}$ [จะมีรูปร่างหน้าตาน่าเกลียดน่ากลัว]$_{13.2}$

(ข)$_{plg}$ [ในทางพุทธศาสนา]$_1$ [ยักษ์มีทั้งรูปร่างหน้าตางดงามและอัปลักษณ์]$_2$ [ยักษ์]$_{3.1}$ [ที่สร้างกุศล]$_4$ [จะมีรูปร่างหน้าตางดงาม]$_{3.2}$ [ทรงสง่าราศี]$_5$ [ส่วนยักษ์]$_{6.1}$ [ที่สร้างบาป]$_7$ [จะมีรูปร่างหน้าตาน่าเกลียดน่ากลัว]$_{6.2}$



ภาพที่ 4.1 แผนผังต้นไม้โครงสร้างวาทะจากข้อความในตัวอย่างที่ 4.19

จากตัวอย่างที่ 4.19 ข้างต้นจะเห็นได้ว่าเจตนาของผู้เขียนข้อความต้นฉบับ (ก) คือต้องการแสดงการเปรียบต่างระหว่างยักษ์ในวรรณคดีสันสกฤตกับยักษ์ในความเชื่อทางพุทธศาสนา ดังจากเห็น

ได้จากภาพที่ 4.1 ที่แสดงถึงวาทสัมพันธ์เปรียบต่าง (contrast) ที่ถูกกำกับไว้ในจุดแตกกิ่งบนสุดในแผนผังต้นไม้โครงสร้างวาทะ แต่ในข้อความลักลอก (ข) ผู้ลักลอกได้เลือกตัดทอนเนื้อหาส่วนที่กล่าวถึงยักษ์ในวรรณคดีสันสกฤตในหน่วยปริจเฉทพื้นฐานที่ 1-5 ของข้อความต้นฉบับ (ก) ออกไปทั้งหมด คงเหลือเฉพาะเนื้อหาเกี่ยวกับยักษ์ในความเชื่อทางพุทธศาสนา เจตนาหลักของข้อความลักลอก (ข) จึงเปลี่ยนเป็นการแสดงรายละเอียดของยักษ์ 2 ประเภทตามความเชื่อของพุทธศาสนาแทน โดยใช้วาทสัมพันธ์การให้รายละเอียด (elaboration)

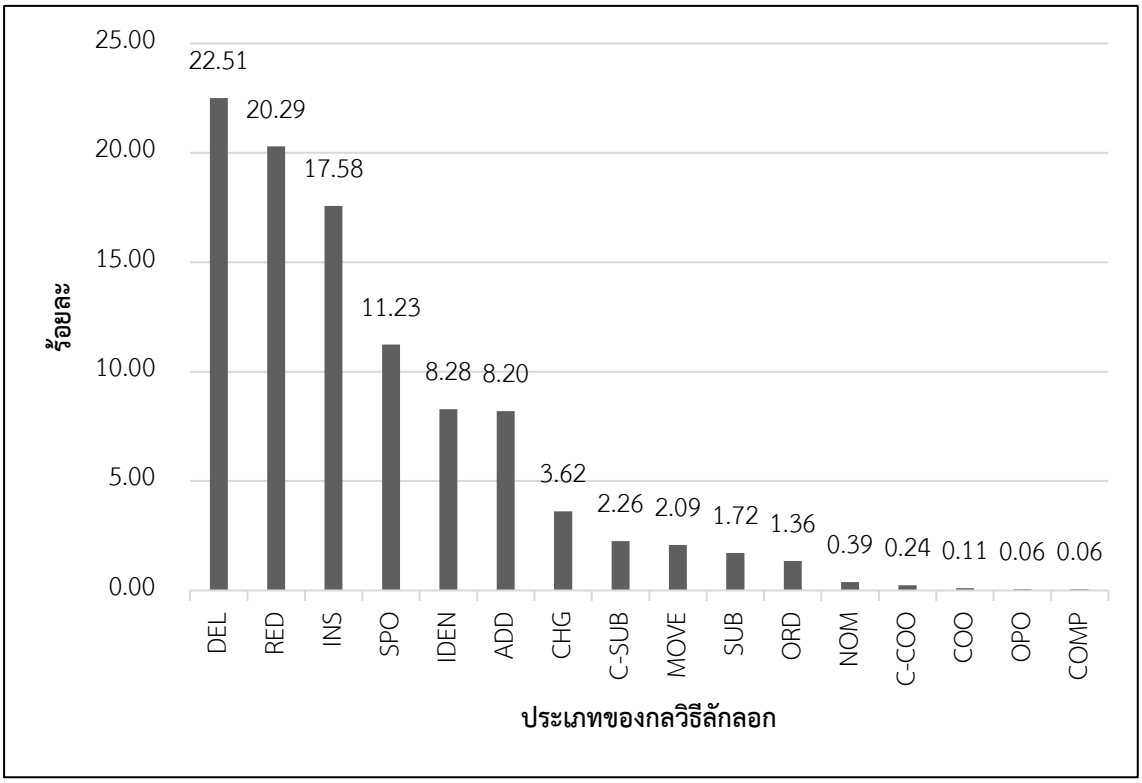
จากประเภทของกลวิธีลักลอกทั้งหมดที่ได้กล่าวไปแล้ว ในส่วนท้ายของในหัวข้อนี้ผู้วิจัยจะได้สรุปรายการประเภทของกลวิธีทั้งหมด พร้อมทั้งกำกับสัญลักษณ์ไว้เพื่อให้ง่ายต่อการแสดงผลการวิเคราะห์ในหัวข้อต่อไป ดังแสดงในตารางที่ 4.1 ต่อไปนี้

ตารางที่ 4.1 รายการประเภทของกลวิธีลักลอก

ประเภท	ประเภทย่อย	สัญลักษณ์
กลวิธีลักลอกภายในหน่วยปริจเฉทพื้นฐาน	การแทนที่ในขั้วเดียวกัน	SPO
	การแทนที่ในขั้วตรงกันข้าม	OPO
	การแทรก	INS
	การลบและการละ	DEL
	การเปลี่ยนลำดับ	ORD
กลวิธีลักลอกระหว่างหน่วยปริจเฉทพื้นฐาน	การซ้อนความ	SUB
	การยุบเลิกความซ้อน	C-SUB
	การเชื่อมความ	COO
	การยุบเลิกความรวม	C-COO
	การแปลงเป็นนามวลี	NOM
	การแปลงเป็นส่วนเติมเต็ม	COMP
	การเพิ่มเติมเนื้อหา	ADD
	การตัดทอนเนื้อหา	RED
	การย้ายลำดับเนื้อหา	MOVE
	การเปลี่ยนแปลงความสัมพันธ์	CHG

4.2 ปริมาณการใช้กลวิธีลักลอก

หลังจากวิเคราะห์ได้กลวิธีลักลอกมาเรียบร้อยแล้ว ผู้วิจัยได้นับจำนวนครั้งของการใช้กลวิธีลักลอกแต่ละวิธีที่ถูกใช้ในคู่ลักลอก ผลปรากฏว่ามีการใช้กลวิธีลักลอกทั้งหมด 4,647 ครั้ง ทั้งนี้สามารถแจกแจงปริมาณการใช้กลวิธีลักลอกแต่ละวิธีเป็นร้อยละได้ดังภาพที่ 4.2



ภาพที่ 4.2 แผนภูมิแสดงร้อยละของปริมาณการใช้กลวิธีลักลอกเป็นรายกลวิธี

จากแผนภูมิในภาพที่ 4.2 ข้างต้น จะเห็นได้ว่าผู้ลักลอกใช้กลวิธีการลบ/การละ (DEL) ในการลักลอกมากที่สุด คิดเป็นร้อยละ 22.51 ของปริมาณการใช้กลวิธีลักลอกทั้งหมด กลวิธีที่ผู้ลักลอกใช้มากรองลงมา ได้แก่ การตัดทอนเนื้อหา (RED) ซึ่งคิดเป็นร้อยละ 20.29 และกลวิธีที่ผู้ลักลอกใช้มากเป็นอันดับที่ 3 ได้แก่ การแทรก (INS) คิดเป็นร้อยละ 17.58 จากนั้นจึงเป็นกลวิธีอื่นๆ ไล่เรียงกันตามลำดับ

ข้อค้นพบข้างต้นนี้ได้เผยให้เห็นถึงธรรมชาติที่แท้จริงของการลักลอกว่า ในสถานการณ์การลักลอกที่ผู้ลักลอกไม่ได้ถูกกำหนดให้ลักลอกหรือถอดความประโยคต่อประโยค แต่เป็นการลักลอกในระดับข้อความนั้น ผู้ลักลอกย่อมเลือกใช้กลวิธีที่ง่ายและซับซ้อนน้อยกว่าเป็นลำดับแรกๆ แต่ในขณะเดียวกันผู้ลักลอกก็ต้องการให้ข้อความต้นฉบับกับข้อความลักลอกแตกต่างกันอย่างชัดเจน ดังจะเห็นได้ว่าการลบ/การละ (DEL) และการแทรก (INS) ที่มีปริมาณการใช้มากเป็นอันดับต้นๆ นั้นเป็น

กลวิธีที่อาศัยความรู้ทางภาษาศาสตร์ในระดับคำศัพท์เท่านั้น ในขณะที่กลวิธีอื่นๆ ที่มุ่งเปลี่ยนแปลงโครงสร้างทางวากยสัมพันธ์กลับถูกเลือกใช้ในปริมาณที่ต่ำมาก ส่วนการตัดทอนเนื้อหา (RED) อันเป็นกลวิธีที่ถูกใช้มากเป็นอันดับที่ 2 นั้นก็เป็นกลวิธีที่ส่งผลให้ความยาวของข้อความลักลอกสั้นลงจากข้อความต้นฉบับอย่างชัดเจน ซึ่งข้อค้นพบเรื่องความยาวของข้อความนี้ยังสอดคล้องกับข้อสรุปของบาร์รอน-เซเดโญและคณะ (Barrón-Cedeño et al., 2013, p. 943) ที่ระบุว่าข้อความที่ถูกลักลอกมีแนวโน้มที่จะสั้นลงจากข้อความต้นฉบับอีกด้วย

อย่างไรก็ตาม ข้อค้นพบข้างต้นก็มีส่วนที่ไม่สอดคล้องกับงานวิจัยที่เกี่ยวข้องในบางประเด็น ดังจะกล่าวต่อไปนี้

ประเด็นแรก งานวิจัยที่เกี่ยวข้องส่วนใหญ่ ไม่ว่าจะเป็นงานของฟุจิตะ (Fujita, 2005), วิลาและคณะ (Vila et al., 2011), ภาคัตและโฮวี (Bhagat & Hovy, 2013), หรือ บาร์รอน-เซเดโญและคณะ (Barrón-Cedeño et al., 2013) ล้วนแล้วแต่ให้ความสำคัญกับการแทนที่ในขั้วเดียวกัน (SPO) หรือการแทนที่ด้วยคำไวพจน์ในฐานะกลวิธีพื้นฐานในการลักลอกและถอดความ และในกรณีที่มีการนับปริมาณการใช้ก็จะปรากฏเป็นกลวิธีที่ใช้มากที่สุด ลักษณะดังกล่าวแตกต่างกับผลวิเคราะห์ในงานวิจัยชิ้นนี้ที่พบปริมาณการใช้กลวิธีแทนที่ในขั้วเดียวกัน (SPO) มากเป็นอันดับที่ 4 ในประเด็นนี้ผู้วิจัยเห็นว่าอาจเป็นผลมาจากการที่ในงานเหล่านั้น ผู้ลักลอกถูกกำหนดให้ลักลอกหรือถอดความประโยคต่อประโยค ทั้งนี้ แน่นอนว่าประโยคย่อมมีปริมาณเนื้อหาให้ดัดแปลงแก้ไขน้อยกว่าข้อความในระดับย่อหน้า ผู้ลักลอกจึงต้องมุ่งแทนที่คำศัพท์ที่มีอยู่ในประโยคด้วยคำในขั้วเดียวกันเพื่อให้เกิดความแตกต่างระหว่างประโยคต้นฉบับกับประโยคปลายทาง

อีกประเด็นที่ผู้วิจัยจะยกมากล่าวในที่นี้คือ จากการวิเคราะห์ไม่พบว่าผู้ลักลอกใช้กลวิธีสลับวาจก (voice alternation) เลย ต่างกับงานวิจัยที่เกี่ยวข้องชิ้นอื่นๆ ที่วิเคราะห์พบกลวิธีดังกล่าว ทั้งนี้ผู้วิจัยเห็นว่าสาเหตุที่ไม่พบกลวิธีสลับวาจกนั้น เนื่องมาจากในภาษาไทย การใช้รูปประโยคกรรมวาจกมักจะถูกต่อต้าน โดยให้เหตุผลว่าเป็นรูปภาษาต่างประเทศ (อมรา ประสิทธิ์รัฐสินธุ์, 2542, น. 132-133) ยิ่งในปริบทเชิงวิชาการที่ต้องการความน่าเชื่อถือเช่นในสถานการณ์การลักลอกที่กำหนดให้ในการเก็บข้อมูลด้วยแล้ว ผู้ลักลอกยิ่งต้องหลีกเลี่ยงการใช้รูปประโยคดังกล่าว นอกจากนั้นแล้ว การสลับวาจกยังทำให้เกิดโครงสร้างประโยคที่ผู้ใช้ภาษาไทยทั่วไปไม่ใช้กันโดยปรกติ ด้วยเหตุนี้ กลวิธีดังกล่าวจึงไม่ถูกนำมาใช้ อย่างไรก็ดี งานวิจัยของ ภาคัตและโฮวี (Bhagat & Hovy, 2013, p. 470) และบาร์รอน-เซเดโญและคณะ (Barrón-Cedeño et al., 2013, p. 932) ต่างก็พบการใช้กลวิธีสลับวาจกในปริมาณที่ต่ำมากเช่นกัน

นอกจากนี้แล้ว ยังมีอีกข้อน่าสังเกตอีกประการหนึ่งที่ผู้วิจัยใคร่ชี้ให้เห็น ได้แก่ การคงรูปคู่ลักลอกให้เหมือนเดิมทุกประการ (IDEN) ซึ่งจากภาพที่ 4.2 จะได้เห็นว่าปรากฏในปริมาณเป็นอันดับที่ 5 เมื่อนับรวมกับกลวิธีลักลอกอื่นๆ ทั้งนี้ ผู้วิจัยเห็นว่าลักษณะดังกล่าวเป็นผลมาจากการกำหนดให้ผู้ลักลอกลักลอกในระดับย่อหน้าเช่นกัน เนื่องจากเมื่อกำหนดให้ลักลอกในระดับย่อหน้า ผู้ลักลอกย่อมมีโอกาสเลือกที่จะดัดแปลงแก้ไขเนื้อหาได้หลากหลายกว่าการกำหนดให้ลักลอกในระดับประโยค ในขณะเดียวกัน ผู้ลักลอกก็อาจเลือกคงรูปภาษาและเนื้อหาที่เห็นว่าไม่จำเป็นต้องดัดแปลงแก้ไขไว้ได้เช่นกัน

4.3 รูปแบบการใช้กลวิธีลักลอก

จากการวิเคราะห์รูปแบบการใช้กลวิธีลักลอก ผู้วิจัยพบว่า ในคู่ลักลอกแต่ละคู่นั้น ผู้ลักลอกอาจเลือกใช้กลวิธีใดกลวิธีหนึ่งในการลักลอก หรือใช้หลายกลวิธีร่วมกันก็ได้ ทั้งนี้ สามารถแสดงจำนวนกลวิธีที่ปรากฏในคู่ลักลอกและความถี่ในการปรากฏได้ดังตารางที่ 4.2

ตารางที่ 4.2 จำนวนกลวิธีลักลอกที่ปรากฏในคู่ลักลอกและความถี่ในการปรากฏ

จำนวนกลวิธีลักลอกที่ปรากฏในคู่ลักลอก (ประเภท)	**ความถี่ในการปรากฏ (ครั้ง)**	**ร้อยละ**
5	10	0.37
4	78	2.86
3	194	7.10
2	348	12.74
1	2,102	76.93
รวม	2,732	100.00

จากตารางที่ 4.2 จะเห็นได้ว่าในคู่ลักลอก 1 คู่สามารถปรากฏกลวิธีลักลอกร่วมกันได้ถึง 5 ประเภท อย่างไรก็ดี หากพิจารณาความถี่ในการปรากฏร่วมด้วยแล้ว จะเห็นว่าการลักลอกที่ใช้กลวิธีร่วมกันน้อยกว่าปรากฏในความถี่สูงกว่าการลักลอกที่ใช้กลวิธีร่วมกันมาก ลักษณะดังกล่าวสะท้อนให้เห็นถึงพฤติกรรมของผู้ลักลอกได้ว่า ในการลักลอกข้อมูลที่เป็นความคิดสำคัญครั้งหนึ่งๆ ผู้ลักลอกมีแนวโน้มจะใช้กลวิธีลักลอกร่วมกันในจำนวนน้อยหรือเลือกใช้กลวิธีใดกลวิธีหนึ่งเท่านั้น

อย่างไรก็ดี หากพิจารณาในด้านรูปแบบการปรากฏร่วมกันของกลวิธีลักลอกนั้นแล้ว จะพบว่าผู้ลักลอกมีแนวโน้มที่จะเลือกใช้การลบ/การละ (DEL) และกลวิธีการแทรก (INS) ร่วมกันเป็นรูปแบบพื้นฐาน และหากผู้ลักลอกต้องการเลือกใช้กลวิธีอื่น ๆ เพิ่มขึ้น ก็จะเลือกใช้การแทนที่ในขั้วเดียวกัน

(SPO) การเปลี่ยนแปลงความสัมพันธ์ (CHG) และการยุบเลิกความซ้อน (C-SUB) เพิ่มเข้ามา อันจะสังเกตได้จากรูปแบบที่ปรากฏเป็นอันดับที่ 1 ของจำนวนกลวิธีลักลอกที่ปรากฏร่วมกัน 2, 3, 4 และ 5 ประเภท ตามลำดับ ดังตารางที่ 4.3

ตารางที่ 4.3 กลวิธีลักลอกที่ปรากฏร่วมกันสูงสุด 3 รูปแบบแรก

จำนวนกลวิธีลักลอกที่ปรากฏร่วมกัน	อันดับ	รูปแบบ	ร้อยละ
5	1	DEL, INS, SPO, CHG, C-SUB	30.00
	2	DEL, INS, SPO, ORD, MOVE	20.00
	3	DEL, INS, SPO, ORD, SUB	10.00
		DEL, INS, SPO, SUB, CHG	10.00
		DEL, INS, SPO, C-SUB, MOVE	10.00
		DEL, INS, SPO, ORD, CHG	10.00
		DEL, SPO, SUB, C-COO, MOVE	10.00
4	1	DEL, INS, SPO, CHG	17.95
	2	DEL, INS, SPO, C-SUB	11.54
	3	DEL, INS, SPO, MOVE	10.26
	...	อื่นๆ	60.25
3	1	DEL, INS, SPO	23.71
	2	DEL, INS, CHG	11.34
	3	DEL, SPO, C-SUB	8.25
	...	อื่นๆ	56.70
2	1	DEL, INS	21.84
	2	DEL, SPO	13.79
	3	INS, SPO	12.07
	...	อื่นๆ	52.30

เมื่อพิจารณารูปแบบการปรากฏร่วมกันของกลวิธีลักลอกทั้งหมดแล้ว พบว่ากลวิธีลักลอกทั้งหมดสามารถปรากฏร่วมกับกลวิธีลักลอกประเภทอื่นได้ ยกเว้นการตัดทอนเนื้อหา (RED) และการเพิ่มเติมเนื้อหา (ADD) ทั้งนี้ เนื่องมาจากกลวิธีดังกล่าวเป็นการดัดแปลงแก้ไขหน่วยปริจเฉทพื้นฐานทั้งหน่วย

อย่างไรก็ดี เมื่อไม่นับรวมกลวิธีการตัดทอนเนื้อหาและการเพิ่มเติมเนื้อหาแล้ว จะพบว่าพฤติกรรมการเลือกกลวิธีลักลอกเพื่อให้ปรากฏร่วมกันนั้นสอดคล้องกับปริมาณการใช้กลวิธีลักลอกในภาพที่ 3 กล่าวคือ กลวิธีที่มีปริมาณการใช้มากกว่าจะถูกผู้ลักลอกเลือกเข้ามาให้ปรากฏร่วมกันก่อนกลวิธีที่มีปริมาณการใช้น้อยกว่า ลักษณะดังกล่าวช่วยยืนยันให้เห็นชัดเจนยิ่งขึ้นว่าผู้ลักลอกย่อมเลือกใช้กลวิธีที่ซับซ้อนน้อยกว่าก่อนจะเลือกกลวิธีที่ซับซ้อนมากขึ้น

4.4 การประยุกต์ใช้กลวิธีลักลอกในการพัฒนาระบบการลักลอกงานวิชาการ

ในหัวข้อนี้ ผู้วิจัยจะกล่าวถึงการนำข้อค้นพบที่ได้จากการวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทยมาประยุกต์ใช้ในงานวิจัยชิ้นนี้ ทั้งนี้ ผู้วิจัยได้แบ่งการประยุกต์ใช้ออกเป็น 3 ด้าน ได้แก่ ด้านการสร้างคลังข้อมูล ด้านการวิเคราะห์หาลักษณ์ที่ใช้ในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอก และด้านการอภิปรายผลที่ได้จากงานวิจัยชิ้นนี้ รายละเอียดมีดังต่อไปนี้

4.4.1 ด้านการสร้างคลังข้อมูล

จากข้อค้นพบด้านปริมาณการใช้กลวิธีลักลอกที่ได้แสดงข้างต้น จะเห็นได้ว่ากลวิธีลักลอกที่ผู้ลักลอกมีแนวโน้มจะใช้ในอันดับต้นๆ นั้นสอดคล้องกับการออกแบบคลังข้อมูลในเบื้องต้นที่ผู้วิจัยได้สรุปไว้จากการทบทวนวรรณกรรม กล่าวคือ ออกแบบให้ข้อมูลส่วนที่มีการลักลอกแบ่งเป็นหลายประเภทตามลำดับขั้นของหน่วยทางภาษาในการแก้ไขข้อความต้นฉบับ

หากพิจารณาลักษณะของข้อมูลลักลอกประเภทคัดลอกโดยตรง (Exact Copy: EC) ที่ได้กล่าวถึงไปในหัวข้อที่ 3.3.3.1 แล้ว จะเห็นได้ว่าลักษณะของข้อมูลประเภทดังกล่าวคือการคัดลอกข้อความต้นฉบับโดยไม่มีการเปลี่ยนแปลงใดๆ ลักษณะดังกล่าวนี้สอดคล้องกับกลวิธีการคงรูปคู่ลักลอกให้เหมือนเดิมทุกประการ (IDEN) ที่ปรากฏใช้มากเป็นอันดับที่ 5 ของปริมาณการใช้กลวิธีลักลอกทั้งหมด ข้อค้นพบประการนี้จึงช่วยยืนยันได้ว่าการออกแบบให้มีข้อมูลลักลอกประเภทคัดลอกโดยตรงในคลังข้อมูลนั้นไม่ได้มีความคลาดเคลื่อนไปจากธรรมชาติของการลักลอกโดยที่กระทำโดยมนุษย์แต่อย่างใด อีกทั้งกลวิธีดังกล่าวยังถูกใช้ในปริมาณมากจนสามารถจูงใจให้ผู้วิจัยใช้กลวิธีดังกล่าวนี้สร้างเป็นข้อมูลลักลอกประเภทหนึ่งในคลังข้อมูลได้

เมื่อพิจารณาข้อค้นพบด้านปริมาณการใช้กลวิธีลักลอกเพิ่มเติมอีกจะเห็นได้ว่ากลวิธีที่ถูกใช้มากที่สุด 3 อันดับแรกนั้นเป็นการแก้ไขข้อความต้นฉบับในระดับคำและวากยสัมพันธ์ ด้วยเหตุนี้ผู้วิจัยจึงกำหนดให้ใช้กลวิธีแทรกคำ (INS) และกลวิธีลบคำ (DEL) เป็นกลวิธีหลักที่ใช้ในการจำลองข้อมูลลักลอกประเภทคัดลอกโดยใกล้เคียง (Near Copy: NC) ด้วยกลวิธีทั้งสองดังกล่าวถูกใช้มากกว่า

ร้อยละ 40 ของกลวิธีทั้งหมด ซึ่งปริมาณดังกล่าวมากพอให้แยกข้อมูลส่วนดังกล่าวเป็นประเภทหนึ่งของข้อมูลลักลอกในคลังข้อมูลได้

ส่วนกลวิธีการตัดทอนเนื้อหา (RED) ที่ปรากฏใช้มากเป็นอันดับที่ 2 ของปริมาณการใช้กลวิธีลักลอกทั้งหมดนั้น แม้ในแง่มุมหนึ่งจะสามาถพิจารณาให้เป็นการแก้ไขข้อความต้นฉบับในระดับปริจเฉทได้ แต่ด้วยวิธีการวิเคราะห์กลวิธีในงานชิ้นนี้อิงกับทฤษฎีโครงสร้างวาทะที่มีหน่วยพื้นฐานในที่ใช้ในการวิเคราะห์คือหน่วยปริจเฉทพื้นฐาน จึงกล่าวได้กลวิธีการตัดทอนเนื้อหาเป็นการแก้ไขข้อความต้นฉบับในระดับวากยสัมพันธ์ได้เช่นกัน ด้วยหน่วยปริจเฉทพื้นฐานอยู่ในรูปอนุพากย์หรือวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น การตัดทอนเนื้อหาใดๆ ในข้อความต้นฉบับจึงเป็นการตัดหน่วยปริจเฉทพื้นฐานในข้อความต้นฉบับออกไป และเพื่อให้มีการแก้ไขข้อความต้นฉบับในระนาบเดียวกันทางลำดับชั้นของหน่วยทางภาษา ผู้วิจัยจึงได้นำกลวิธีอีก 2 กลวิธีที่สัมพันธ์กันคือการเพิ่มเติมเนื้อหา (ADD) และการย้ายลำดับเนื้อหา (MOVE) มาใช้ร่วมกับกลวิธีการตัดทอนเนื้อหา (RED) ในการจำลองข้อมูลลักลอกประเภทคัดลอกโดยดัดแปลง (Modified Copy: MO) โดยดัดแปลงจากข้อมูลลักลอกประเภทคัดลอกโดยใกล้เคียง (Near Copy: NC) อีกชั้นหนึ่ง ในแง่นี้ข้อมูลลักลอกประเภทคัดลอกโดยดัดแปลงจะมีความซับซ้อนมากขึ้นจากข้อมูลลักลอก 2 ประเภทแรกที่กล่าวไป และรวมกลวิธีที่ใช้ในการลักลอกมากถึง 5 กลวิธี

ส่วนข้อมูลลักลอกประเภทสุดท้ายที่ออกแบบให้บรรจุเข้าในคลังข้อมูลคือข้อมูลลักลอกประเภทถอดความ (Paraphrase: PA) นั้น ดังได้กล่าวไว้ในแล้วในหัวข้อข้อที่ 3.3.3.4 แม้ผู้วิจัยจะไม่ได้กำหนดให้ผู้จำลองการลักลอกใช้กลวิธีใดโดยตรง เป็นเพียงการกำหนดกรอบโดยกว้างว่าให้ถอดความข้อความต้นฉบับ เพื่อเปิดโอกาสให้จำลองการลักลอกได้ใช้กลวิธีต่างๆ อย่างหลากหลาย อย่างไรก็ดี ผู้วิจัยก็พบว่าผู้จำลอกการลักลอกถอดความข้อความต้นฉบับโดยใช้กลวิธีที่หลากหลายสอดคล้องกับข้อค้นพบจากการวิเคราะห์กลวิธีลักลอกที่ได้กล่าวไปแล้ว ทั้งนี้ ผู้วิจัยยังสังเกตพบว่าผู้จำลองการลักลอกเหลือใช้กลวิธีแทนที่ในขั้วเดียวกัน (SPO) ในปริมาณมาก รวมถึงมีการแก้ไข้ข้อความต้นฉบับทั้งในระดับคำ วากยสัมพันธ์ และความหมาย ลักษณะดังกล่าวถือได้ว่าตรงกับนิยามของการถอดความตามที่ได้ทบทวนวรรณกรรมไว้และสอดคล้องกับเจตนาเบื้องหลังการออกแบบคลังข้อมูลที่ตั้งไว้แต่ต้นที่ตั้งใจจะให้ข้อมูลลักลอกประเภทนี้มีระดับความคลุมเครือสูงที่สุดและตรวจหาการลักลอกได้ยากที่สุด

4.4.2 ด้านการประยุกต์ใช้ลักษณ์ในการตรวจหาการลักลอก

จากข้อค้นพบด้านประเภทของกลวิธีลักลอกทำให้ผู้วิจัยสามารถเข้าใจถึงลักษณะการแก้ไขข้อความต้นฉบับและนำมาสู่ความเข้าใจด้านการประยุกต์ใช้ลักษณ์ต่างๆ ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกได้

จากการทบทวนวรรณกรรมในหัวข้อที่ 2.5 ว่าด้วยวิธีตรวจเทียบภายนอกหาการลักลอกงานวิชาการ และหัวข้อที่ 2.7.3 ว่าด้วยแนวทางการวัดค่าความละม้ายของข้อความ จะเห็นได้ว่าวิธีการตรวจหาการลักลอกภายนอกและวิธีการวัดค่าความละม้ายของข้อความนั้นอิงกับลำดับชั้นของหน่วยทางภาษาเช่นเดียวกันกับกลวิธีการลักลอก ลักษณะดังกล่าวจูงใจให้ผู้วิจัยเลือกนำแนวทางการตรวจหาและค่าความละม้ายที่วัดได้จากหน่วยทางภาษาในลำดับชั้นต่างๆ มาใช้เป็นลักษณ์ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก

ทั้งนี้ หากพิจารณากลวิธีลักลอกที่วิเคราะห์พบแล้ว จะเห็นได้ว่ากลวิธีดังกล่าวสามารถตรวจหาได้ด้วยลักษณ์ประเภทต่างๆ กล่าวคือ กลวิธีการแทรก และการลบ/การละ นั้นสามารถตรวจหาการลักลอกได้โดยใช้ลักษณ์ทางศัพท์ ไม่ว่าจะเป็นลักษณ์ที่ได้จากลำดับร่วมที่ยาวที่สุดของชุดคำ ลักษณ์ที่ได้จากการประยุกต์ใช้แนวคิดเรื่องเอ็นแกรมของคำ ตลอดจนลักษณ์ที่นำค่าความละม้ายที่วัดได้ในระดับคำมาประยุกต์

ส่วนกลวิธีการซ้อนความ การยุบเลิกความซ้อน การเชื่อมความ การยุบเลิกความรวม การแปลงเป็นนามวลี การแปลงเป็นส่วนเติมเต็ม นั้น นอกจากจะสามารถตรวจหาได้จากลักษณ์ทางศัพท์แล้ว ยังอาจตรวจหาได้ด้วยลักษณ์ทางวายสัมพันธ์ได้ด้วย เช่น ลักษณ์ที่ได้จากลำดับคำในข้อความหมวดคำ ความสัมพันธ์แบบพึ่งพา และค่าความละม้ายที่ได้วัดจากลักษณะทางวายสัมพันธ์ดังกล่าว

ส่วนกลวิธีลักลอกที่แก้ไขข้อความต้นฉบับในระดับความหมายนั้น ผู้วิจัยคิดว่าลักษณ์ทางวากยสัมพันธ์บางชนิดที่ประยุกต์ใช้ลักษณะการเรียงลำดับของหมวดคำก็อาจจะตรวจหาการลักลอกที่ใช้กลวิธีการแทนที่ในขั้วเดียวกันได้ หากผู้ลักลอกแทนที่คำในข้อความต้นฉบับด้วยคำในหมวดคำเดียวกัน นอกจากนี้ ลักษณ์ที่ได้จากลักษณะทางความหมาย เช่น การวิเคราะห์ความหมายแฝง ก็น่าจะมีประสิทธิภาพในการตรวจหาการลักลอกในระดับความหมายได้

อย่างไรก็ตาม หากพิจารณากลวิธีการเพิ่มเติมเนื้อหา การตัดทอนเนื้อหา และการเปลี่ยนแปลงความสัมพันธ์ แล้ว จะเห็นได้ว่ากลวิธีดังกล่าวมุ่งเปลี่ยนแปลงข้อความต้นฉบับในระดับปริจเฉท ซึ่งในงานวิจัยชิ้นนี้ยังมีข้อจำกัดด้านการวิเคราะห์หาลักษณ์จากหน่วยทางภาษาในระดับดังกล่าว การตรวจหาการลักลอกในระดับปริจเฉทในงานวิจัยชิ้นนี้จึงอาจต้องอาศัยลักษณ์ที่ได้จากหน่วยทางภาษาระดับอื่นๆ แทน อย่างไรก็ดี ดังได้กล่าวไปในหัวข้อย่อยที่แล้วว่าหน่วยพื้นฐานในที่ใช้

ในการวิเคราะห์กลวิธีลักลอกในงานวิจัยชิ้นนี้คือหน่วยปริจเฉทพื้นฐานซึ่งเป็นอยู่ในรูปของอนุพากย์หรือวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น ฉะนั้นแล้ว ผู้วิจัยจึงคาดว่ากลวิธีลักลอกประเภทที่มุ่งแก้ไขข้อความต้นฉบับในระดับปริจเฉทน่าจะสามารถตรวจหาได้โดยใช้ลักษณ์ทางวากยสัมพันธ์ได้

จากที่กล่าวมาข้างต้นจะเห็นได้ว่างานวิจัยชิ้นนี้ได้วางแนวทางในการตรวจหาการลักลอกที่เกิดจากกลวิธีลักลอกต่างๆ ไว้อย่างรัดกุม ครอบคลุมลักษณะการลักลอกทั้งหมดที่เกิดขึ้นในสถานการณ์จริง โดยการประยุกต์ใช้ลักษณ์ที่ได้จากหน่วยทางภาษาระดับต่างๆ ในการตรวจหา ด้วยเหตุนี้ ผู้วิจัยจึงมั่นใจว่าระบบตรวจหาการลักลอกงานวิชาการที่พัฒนาขึ้นตามแนวทางและองค์ความรู้ที่เสนอในงานวิจัยชิ้นนี้จะมีประสิทธิภาพสามารถตรวจหาการลักลอกได้ผลเป็นที่น่าพอใจ

4.4.3 ด้านการอภิปรายผลการวิจัย

เมื่อพิจารณาข้อค้นพบด้านประเภทของกลวิธีลักลอกและปริมาณการใช้กลวิธีลักลอกแล้ว จะเห็นได้ว่าข้อค้นพบดังกล่าวเป็นประโยชน์อย่างยิ่งในฐานะฐานความรู้ที่ใช้ในการอภิปรายผลการวิจัย โดยเฉพาะอย่างยิ่ง ในกรณีของผลการประเมินประสิทธิภาพของระบบตรวจหาการลักลอกที่พัฒนาขึ้นตามแนวทางที่เสนอในงานวิจัยชิ้นนี้

ในด้านการประเมินประสิทธิภาพของข้อมูลรับเข้า ผลการวิเคราะห์กลวิธีลักลอกช่วยให้ผู้วิจัยเข้าใจธรรมชาติของการลักลอกที่เกิดขึ้นทั้งภายในหน่วยปริจเฉทพื้นฐานและระหว่างหน่วยปริจเฉทพื้นฐาน อันเป็นลักษณ์ที่เกี่ยวข้องกับข้อมูลรับเข้าที่งานวิจัยชิ้นนี้สนใจศึกษาเปรียบเทียบประสิทธิภาพ ดังนั้นองค์ความรู้ในส่วนนี้จึงประโยชน์อย่างต่อการทำความเข้าใจและอภิปรายผลการเปรียบเทียบประสิทธิภาพของข้อมูลรับเข้า

ส่วนในด้านการประเมินประสิทธิภาพของลักษณ์ที่ใช้ในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกนั้น นอกจากข้อค้นพบด้านประเภทของกลวิธีลักลอกจะช่วยในการอภิปรายประสิทธิภาพของลักษณ์แต่ละตัวได้แล้ว ข้อค้นพบในด้านปริมาณการใช้กลวิธีลักลอกยังช่วยในการอภิปรายประสิทธิภาพในการตรวจหาการลักลอกในภาพรวมด้วย กล่าวคือ ในกรณีที่ผลการประเมินประสิทธิภาพของลักษณ์ตัวที่ดีที่สุดไม่เป็นที่น่าพอใจ เช่น ให้ค่า F ไม่ถึง 0.5 ผู้วิจัยจะประเมินผลการตรวจหาข้อมูลลักลอกแยกเป็นรายประเภทจากทั้งหมด 4 ประเภท ทั้งนี้ ผลจากการวิเคราะห์ปริมาณการใช้กลวิธีลักลอกชี้ให้เห็นว่าประมาณร้อยละ 50 ของปริมาณกลวิธีที่ผู้ลักลอกเลือกใช้เป็นกลวิธีที่แก้ไขข้อความต้นฉบับในระดับคำและวากยสัมพันธ์ ได้แก่ กลวิธีการลบ/ละ (DEL) การตัดทอนเนื้อหา (RED) และการแทรก (INS) กลวิธีทั้งสามนี้ปรากฏใช้อยู่ในข้อมูลลักลอกประเภทคัดลอกโดยใกล้เคียง (NC) และข้อมูลลักลอกประเภทคัดลอกโดยดัดแปลง (MO) ดังนั้น การทดสอบ

ประสิทธิภาพของลักษณ์ในข้อมูลลักลอกทั้งสองประเภทนี้จะช่วยให้เข้าใจประสิทธิภาพการจำแนกประเภทในเชิงลึกมากยิ่งขึ้น ซึ่งจะช่วยให้อภิปรายผลได้การวิจัยในเบื้องต้นได้ชัดเจนมากขึ้น

จากที่กล่าวมาทั้งหมดจะเห็นได้ว่าการวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทยไม่ได้เป็นเพียงการทำความเข้าใจการลักลอกในเชิงภาษาศาสตร์เท่านั้น แต่ข้อค้นพบที่ได้จากการวิเคราะห์ในบทนี้ยังเป็นฐานความรู้ที่สำคัญในการพัฒนาระบบตรวจหาการลักลอกในขั้นต่อไป ทั้งในด้านการสร้างคลังข้อมูลการลักลอกที่ใช้ในการฝึกฝนและทดสอบระบบ การใช้ลักษณ์ที่เหมาะสมในการตรวจหาการลักลอก รวมถึงการอภิปรายผลการวิจัยในด้านการประเมินประสิทธิภาพของระบบอีกด้วย

บทที่ 5
ลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก

ลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกถือเป็นปัจจัยสำคัญอีกประการหนึ่งที่ส่งผลต่อประสิทธิภาพของระบบตรวจหาการลักลอกงานวิชาการที่พัฒนาขึ้นในงานวิจัยชิ้นนี้ กล่าวคือ หากลักษณ์ที่นำมาใช้สร้างขึ้นจากลักษณะ (characteristics) ที่ชัดเจนและปรากฏอย่าสม่ำเสมอในข้อความที่มีการลักลอกและข้อความไม่มีการลักลอกแล้ว เครื่องก็จะสามารถเรียนรู้รูปแบบและลักษณะของข้อความที่มีการลักลอกและข้อความไม่มีการลักลอกนั้นผ่านลักษณ์ที่ป้อนเข้าสู่ระบบได้ จนกระทั่งส่งผลให้เครื่องสามารถจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกออกจากได้อย่างถูกต้องแม่นยำ ด้วยเหตุดังกล่าวนี้ การวิเคราะห์หาลักษณ์ทางภาษาเพื่อใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกจึงเป็นวัตถุประสงค์อีกประการหนึ่งที่ผลักดันให้เกิดงานวิจัยชิ้นนี้ขึ้น ด้วยผู้วิจัยเชื่อว่าลักษณ์ที่วิเคราะห์และสร้างขึ้นจากลักษณะทางภาษานั้นจะให้ผลในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกเป็นที่น่าพอใจ

ในบทนี้ ผู้วิจัยจะกล่าวถึงลักษณ์ทั้งหมดที่ถูกนำมาวิเคราะห์เพื่อนำไปใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกด้วยแบบจำลองซัพพอร์เวกเตอร์แมชชีน จำนวนทั้งสิ้น 71 ลักษณ์ โดยจะแบ่งการนำเสนอออกเป็น 2 กลุ่มใหญ่ ได้แก่ ลักษณ์อิงอักขระ และลักษณ์ทางภาษา ลักษณ์อิงอักขระเป็นลักษณ์ที่ผู้วิจัยนำมาใช้ในฐานะเส้นฐาน (baseline) ในการทดลองเปรียบเทียบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกันในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกตามวัตถุประสงค์ของการวิจัย ส่วนลักษณ์ทางภาษานั้นเป็นผลที่ได้จากการวิเคราะห์หาตามวิธีการวิจัยที่ได้กล่าวไปแล้วในหัวข้อที่ 3.4 ในส่วนนี้จึงเป็นการนำเสนอผลการวิจัยตามวัตถุประสงค์ประการหนึ่งของงานวิจัยชิ้นนี้ด้วย ทั้งนี้ รายละเอียดของลักษณ์ทั้งหมดมีดังมีต่อไปนี้

5.1 ลักษณ์อิงอักขระ

ลักษณ์อิงอักขระคือลักษณ์ที่ไม่ต้องประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในการวิเคราะห์และสร้าง กล่าวคือ ในการวิเคราะห์และสร้างลักษณ์ประเภทนี้ไม่ต้องอาศัยความเข้าใจเรื่องขอบเขตของคำและลักษณะของคำ ลักษณะและความสัมพันธ์ทางวากยสัมพันธ์ รวมถึงลักษณะและความสัมพันธ์ทางความหมาย เมื่อกล่าวเช่นนี้ ลักษณะเพียงประการที่จะนำมาประยุกต์ใช้เพื่อวิเคราะห์และสร้างลักษณ์ประเภทนี้ได้คือลักษณะของอักขระ (character) ในข้อความ

ในหัวข้อนี้ ผู้วิจัยจะกล่าวถึงลักษณ์อิงอักขระที่นำมาใช้ในการทดลองเปรียบเทียบประสิทธิภาพของลักษณ์ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน รวมจำนวนทั้งหมด 20 ลักษณ์ โดยจะแบ่งการนำเสนอลักษณ์ดังกล่าวออกเป็น 8 กลุ่ม ตามวิธีวิเคราะห์หาและสร้าง ดังต่อไปนี้

5.1.1 ขนาดของคู่หน่วยเทียบ (pair size: $size_{char}$[15])

การใช้ขนาดของย่อหน้าในคู่หน่วยเทียบเป็นลักษณ์เป็นแนวคิดขั้นพื้นฐานในการเปรียบเทียบความแตกต่างระหว่างย่อหน้า 2 ย่อหน้า ด้วยแนวคิดนี้ ย่อหน้าที่ผ่านการลักลอกย่อมมีความยาวแตกต่างจากย่อหน้าต้นฉบับ ดังจะเห็นได้จากข้อค้นพบของบาร์รอน-เซเดโญและคณะ (Barrón-Cedeño et al., 2013, p. 943) และข้อค้นพบที่ได้จากการวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทยที่ได้กล่าวถึงไปแล้วหัวข้อที่ 4.2 ว่าข้อความที่ถูกลักลอกย่อมมีแนวโน้มที่จะมีขนาดสั้นลงจากข้อความที่ปรากฏเป็นต้นฉบับ

ลักษณ์ขนาดของคู่หน่วยเทียบวิเคราะห์หาได้จากการนับจำนวนอักขระของย่อหน้าที่เป็นคู่หน่วยเทียบของกันและกันทั้ง 2 ย่อหน้าแยกต่างหากออกจากกัน ลักษณ์ที่ได้จึงมีลักษณะเป็นจำนวนนับ เมื่อนำมาใช้ทดลองจำแนกประเภทของข้อความ ความยาวของทั้งสองย่อหน้าจะถูกป้อนเข้าเครื่องพร้อมกันเป็นคู่

Case No.	src	plg	len(src)	len(plg)
11301	HS-L-158-SR.txt	HS-L-158-PA.txt	766	759
11302	HS-L-160-SR.txt	HS-L-160-PA.txt	692	681
11303	HS-L-165-SR.txt	HS-L-165-PA.txt	779	571
11304	HS-L-167-SR.txt	HS-L-167-PA.txt	848	776
11305	HS-L-168-SR.txt	HS-L-168-PA.txt	992	968
11306	HS-L-169-SR.txt	HS-L-169-PA.txt	770	757
11307	HS-L-170-SR.txt	HS-L-170-PA.txt	683	643
11308	HS-L-174-SR.txt	HS-L-174-PA.txt	964	895
11309	HS-L-176-SR.txt	HS-L-176-PA.txt	809	805
11310	HS-L-183-SR.txt	HS-L-183-PA.txt	753	745

ภาพที่ 5.1 ตัวอย่างของลักษณ์ขนาดของคู่หน่วยที่เทียบ (อักขระ)

ภาพที่ *5.1* แสดงตัวอย่างของลักษณ์ขนาดของคู่หน่วยเทียบที่วิเคราะห์หาได้จากการนับจำนวนอักขระในย่อหน้าขนาดยาวของคู่หน่วยเทียบประเภทลักลอกแบบถอดความ 10 กรณีในคลังข้อมูล จากภาพ คอลัมน์ “len(src)” และ “len(plg)” แสดงจำนวนอักขระที่นับได้จากข้อความต้นฉบับในคอลัมน์ “src” และข้อความลักลอกในคอลัมน์“plg” ตามลำดับ จากภาพจะเห็นได้ว่า

[15] สัญลักษณ์ภายในวงเล็บนี้ ผู้วิจัยได้กำหนดขึ้นสำหรับใช้ในงานวิจัยชิ้นนี้เท่านั้น ทั้งนี้ เพื่อให้สะดวกแก่การอ้างถึงและนำเสนอผลการวิจัยในส่วนต่อไป

จำนวนอักขระในข้อความลักลอกมีจำนวนลดลงจากจำนวนอักขระในข้อความต้นฉบับ อันเป็นผลเนื่องมาจากการแก้ไขของผู้จำลองการลักลอก ผู้วิจัยจะนำจำนวนอักขระดังกล่าวไปใช้เป็นลักษณ์ในการจำแนกประเภทข้อความที่มีการลักลอกและข้อความที่ไม่มีการลักลอกได้ทันที

ทั้งนี้ ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.1.2 ผลต่างของขนาดของคู่หน่วยเทียบ (difference of pair size: $diff_{Char}$)

ลักษณ์ผลต่างของขนาดของคู่หน่วยเทียบเป็นลักษณ์มุ่งพิจารณาความแตกต่างระหว่างคู่หน่วยเทียบเป็นสำคัญ โดยมีแนวคิดเบื้องหลังคือให้เครื่องเรียนรู้จากลักษณ์ที่เป็นตัวแทนของความแตกต่างกันด้านความยาวของย่อหน้าในคู่เทียบโดยตรง ต่างจากลักษณ์ขนาดของคู่หน่วยเทียบ ($size_{char}$) ที่เครื่องต้องเรียนรู้ขนาดของย่อหน้าทั้งสองย่อหน้าภายในคู่หน่วยเทียบ

ส่วนการวิเคราะห์หาและสร้างลักษณ์ชนิดนี้ ผู้วิจัยจะนำจำนวนอักขระที่นับได้จากย่อหน้าข้อความต้นฉบับลบด้วยจำนวนอักขระที่นับได้จากย่อหน้าข้อความลักลอก ค่าที่ได้จากการคำนวณจะมีทั้งจำนวนเต็มบวกในกรณีที่ข้อความในย่อหน้าลักลอกมีขนาดสั้นกว่าข้อความในย่อหน้าต้นฉบับ จำนวนเต็มลบในกรณีที่ข้อความในย่อหน้าลักลอกมีขนาดยาวกว่าข้อความในย่อหน้าต้นฉบับ และมีค่าเป็น 0 ในกรณีที่ข้อความในย่อหน้าต้นฉบับและข้อความในย่อหน้าลักลอกมีขนาดเท่ากัน

Case No.	src	plg	$diff_{Char}$
11301	HS-L-158-SR.txt	HS-L-158-PA.txt	7
11302	HS-L-160-SR.txt	HS-L-160-PA.txt	11
11303	HS-L-165-SR.txt	HS-L-165-PA.txt	208
11304	HS-L-167-SR.txt	HS-L-167-PA.txt	72
11305	HS-L-168-SR.txt	HS-L-168-PA.txt	24
11306	HS-L-169-SR.txt	HS-L-169-PA.txt	13
11307	HS-L-170-SR.txt	HS-L-170-PA.txt	40
11308	HS-L-174-SR.txt	HS-L-174-PA.txt	69
11309	HS-L-176-SR.txt	HS-L-176-PA.txt	4
11310	HS-L-183-SR.txt	HS-L-183-PA.txt	8

ภาพที่ 5.2 ตัวอย่างลักษณ์ผลต่างของขนาดของคู่หน่วยเทียบ

ภาพที่ 5.2 แสดงตัวอย่างของลักษณ์ผลต่างของขนาดของคู่หน่วยเทียบประเภทลักลอกแบบถอดความ 10 กรณีเดียวกันกับตัวอย่างในภาพที่ 5.1 แต่ในกรณีนี้ ลักษณ์ที่นำมาใช้จะมีลักษณะเป็นค่าเดียว ต่างจากลักษณ์ขนาดของคู่หน่วยเทียบ ($size_{char}$) ที่ความยาวของย่อหน้าทั้งสองของคู่หน่วยเทียบจะถูกป้อนเข้าเครื่องพร้อมกันเป็นคู่ ในแง่นี้จึงเป็นที่น่าสนใจว่าระหว่างลักษณ์ที่ประยุกต์ใช้แนวคิดเรื่องขนาดของข้อความในลักษณะที่เป็นคู่เช่นในกรณีของลักษณ์ขนาดของคู่หน่วยเทียบกับ

ลักษณะที่เป็นค่าเดียวเช่นในกรณีนี้ ลักษณ์ประเภทใดจะให้ประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกที่ดีกว่ากัน

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.1.3 ค่าระยะการแก้ไขเลเวนชเตย์นของอักขระ (Levenshtein edit distance of character: LD_{Char})

ค่าระยะการแก้ไขเลเวนชเตย์น (Levenshtein edit distance) (Levenshtein, 1965) เป็นแนวคิดที่ถูกใช้อย่างแพร่หลายในวงการวิทยาศาสตร์คอมพิวเตอร์และสารสนเทศศาสตร์ เพื่อใช้วัดความแตกต่างระหว่างลำดับของสายอักขระ 2 สาย โดยมีแนวคิดพื้นฐานในการเปรียบเทียบลำดับของสายอักขระว่าการแก้ไขโดยการแทรก ลบ หรือแทนที่อักขระ 1 ครั้งจะมีค่าน้ำหนักเท่ากับ 1

หากจะยกตัวอย่างที่สอดคล้องแนวคิดดังกล่าวข้างต้น ผู้วิจัยขอให้พิจารณากรณีการหาค่าระยะการแก้ไขของสายอักขระ “kitten” และ “sitting” ในกรณีนี้ สายอักขระทั้งสองจะมีค่าระยะการแก้ไขเลเวนชเตย์นเท่ากับ 3 เนื่องจากมีกระบวนการแก้ไขเกิดขึ้นทั้งหมด 3 ครั้ง ได้แก่

(1) *k*itten ⟶ *s*itten มีการแก้ไขโดยแทนที่ “k” ด้วย “s”

(2) sitt*e*n ⟶ sitt*i*n มีการแก้ไขโดยแทนที่ “e” ด้วย “i”

(3) sittin ⟶ sitting มีการแก้ไขโดยแทรก “g” ข้างท้ายสายอักขระ

จากตัวอย่างที่ยกมาข้างต้น จะเห็นได้ว่าการหาค่าระยะการแก้ไขเลเวนชเตย์นเป็นวิธีที่สามารถสะท้อนความแตกต่างระหว่างสายอักขระได้ดีวิธีหนึ่ง ด้วยเหตุนี้ ผู้วิจัยจึงนำแนวคิดดังกล่าวมาประยุกต์ใช้กับการเปรียบเทียบความแตกต่างของข้อความที่เกิดจากการลักลอกโดยนำค่าระยะการแก้ไขเลเวนชเตย์นของอักขระมาใช้เป็นลักษณ์สำหรับจำแนกประเภทของข้อความที่มีการลักลอกและไม่มีการลักลอก

ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้เขียนโปรแกรมคำนวณค่าระยะการแก้ไขเลเวนชเตย์นขึ้นด้วยภาษาไพทอน จากนั้นจึงใช้โปรแกรมดังกล่าวคำนวณหาค่าระยะการแก้ไขของคู่หน่วยเทียบทุกคู่ในคลังข้อมูลแล้วนำมาค่าที่ได้มาใช้เป็นลักษณ์

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.1.4 ความยาวของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ (length of longest common subsequence of character: len(lcs_{Char}))

ลำดับย่อยร่วมที่ยาวที่สุด (longest common subsequence: *lcs*) เป็นอีกหนึ่งแนวคิดพื้นฐานที่ถูกใช้อย่างกว้างขวางในงานด้านการประมวลผลภาษาธรรมชาติ แนวคิดของลำดับย่อยร่วมที่ยาวที่สุดคือการเปรียบเทียบลำดับของสายอักขระ 2 สายจากซ้ายไปขวาโดยไม่จำเป็นต้องมาจากลำดับที่ติดกัน แล้วคืนค่าเป็นลำดับของอักขระที่ยาวที่สุดที่สายอักขระทั้งสองสายนั้นมีร่วมกัน

เพื่อความเข้าใจที่ชัดเจนยิ่งขึ้น ในที่นี้จะยกตัวอย่างเกี่ยวกับลำดับย่อยและลำดับย่อยร่วมยาวสุดที่ยาว เช่น กำหนดให้มีสายอักขระ “ACBBDDCC” ลำดับย่อย (subsequence) ของสายอักขระดังกล่าวสามารถปรากฏเป็นสายอักขระได้หลากหลาย เช่น “ACB”, “ADDC”, “D”, “AC”, “ACBBDC” จะเห็นได้ว่าลำดับย่อยมีการเรียงลำดับไล่จากด้านซ้ายไปด้านขวาโดยไม่จำเป็นต้องมีมีลำดับที่ติดกันได้ ในขณะที่สายอักขระ “DDA” ไม่ถือว่าเป็นลำดับย่อยของสายอักขระดังกล่าว เนื่องจากหลังอักขระ “D” ไม่ปรากฏอักขระ “A” จึงไม่ใช่ลำดับย่อยที่เรียงจากด้านซ้ายไปด้านขวา

ส่วนลำดับร่วมที่ยาวที่สุด (longest common subsequence) นั้นหาได้จากลำดับร่วมที่สายอักขระ 2 สายมีร่วมกัน เช่น

S_1 = AAACCBDBABDDADDCBDDCDABAA

S_2 = CACCCCDAABBDACCDDDBBDDC

$lcs(S_1,S_2)$ = ACCDABDACDDDB

จากตัวอย่างข้างต้น จะเห็นได้ว่าลำดับย่อยร่วมที่ยาวที่สุด $lcs(S_1,S_2)$ เกิดจากการนำลำดับย่อยที่สายอักขระ S_1 และสายขระ S_2 มีร่วมกันมาเรียงต่อกันจากด้านซ้ายไปด้านขวาตามลำดับการปรากฏ

ในแง่การตรวจหาการลักลอกที่เป็นการเปรียบเทียบข้อความต้นฉบับกับข้อความที่ผ่านการลักลอก ลำดับย่อยร่วมที่ยาวที่สุดจะสะท้อนให้เห็นส่วนของข้อความที่คงเหลืออยู่จากการแก้ไข ทั้งนี้เนื่องจากลักษณ์ที่เป็นข้อมูลรับเข้าสำหรับแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนนั้นต้องเป็นค่าตัวเลขเท่านั้น ในที่นี้ ผู้วิจัยจึงประยุกต์แนวเรื่องลำดับย่อยร่วมที่ยาวที่สุดโดยนำจำนวนอักขระที่ปรากฏในลำดับย่อยร่วมที่ยาวที่สุดมาใช้เป็นลักษณ์

ในการสร้างลักษณ์ข้างต้น ขั้นแรก ผู้วิจัยต้องเขียนโปรแกรมเพื่อหาลำดับย่อยร่วมที่ยาวที่สุดของอักขระภายในคู่หน่วยเทียบก่อน จากนั้นจึงหาความยาวของลำดับย่อยร่วมดังกล่าว โดยเขียนโปรแกรมเพื่อนับจำนวนอักขระปรากฏในลำดับย่อยร่วมที่ยาวที่สุดที่หาได้ กระบวนการหาลำดับย่อย

ร่วมที่ยาวที่สุดของอักขระและหาความยาวของลำดับย่อยร่วมที่ยาวที่สุดจะดำเนินไปจนกระทั่งได้ข้อมูลครบจากทุกคู่หน่วยเทียบในคลังข้อมูล

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.1.5 ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ (normalized longest common subsequence of character: $lcs_{\text{norm-Char}}$)

ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ (normalized longest common subsequence of character) เป็นวิธีการวัดค่าความละม้ายที่เสนอโดยคลอฟและสตีเวนสัน (Clough & Stevenson, 2011, p. 17) เพื่อใช้ในการตรวจสอบคลังข้อมูลการลักลอกที่สร้างขึ้น

ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระคำนวณได้จากหาลำดับย่อยร่วมที่ยาวที่สุดของสายอักขระ 2 สายก่อนในขั้นแรก จากนั้นจึงแปลงลำดับย่อยร่วมที่ยาวที่สุดให้เป็นค่าบรรทัดฐาน (normalize) โดยนำค่าความยาวของลำดับย่อยร่วมที่ยาวที่สุดของสายอักขระ 2 สายมาหารด้วยค่าความยาวของสายอักขระใดสายอักขระหนึ่งจากทั้งสองสายอักขระ (งานชิ้นนี้กำหนดให้หารด้วยค่าความยาวของข้อความลักลอกเสมอ) ดังสมการที่ 5.1 ค่าที่ได้จากการคำนวณนี้จะมีลักษณะเป็นค่าความละม้ายระหว่าง 0 ถึง 1 หากมีค่าเท่ากับ 0 แปลว่าคู่ของสายอักขระไม่มีความละม้ายกันเลย แต่หากค่าเท่ากับ 1 แปลว่าคู่ของสายอักขระเหมือนกันทุกประการ

$$lcs_{norm-Char} = \frac{len(lcs_{Char}(S_{src}, S_{plg}))}{len(S_{plg})} \quad (5.1)$$

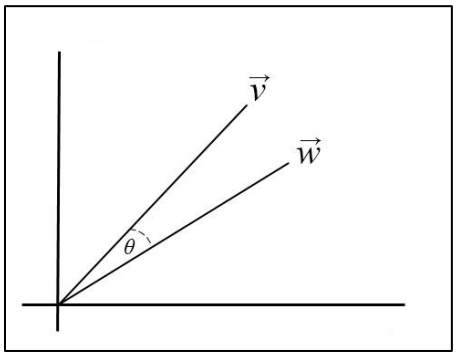
เนื่องจากค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระมีฐานะเป็นค่าความละม้าย จึงสามารถสะท้อนความแตกต่างของข้อความที่ผ่านการลักลอกกับต้นฉบับออกมาในเชิงปริมาณได้ ด้วยเหตุนี้ ผู้วิจัยจึงสนใจนำค่าดังกล่าวมาใช้เป็นลักษณ์ในการทดสอบประสิทธิภาพการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก

ส่วนการสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้แก้ไขโปรแกรมสำหรับหาลำดับย่อยร่วมที่ยาวที่สุดของอักขระภายในคู่หน่วยเทียบที่มีอยู่แล้ว โดยเพิ่มเติมส่วนของการคำนวณค่าความละม้ายตามแนวคิดที่ได้กล่าวไปแล้วในตอนต้น จากนั้นจึงนำไปใช้หาค่า $lcs_{\text{norm-Char}}$ ของคู่หน่วยเทียบทุกคู่ในคลังข้อมูล แล้วนำค่าที่ได้มาใช้เป็นลักษณ์

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.1.6 ค่าความละม้ายโคไซน์ของเอ็นแกรมของอักขระ (cosine similarity of character *n*-gram: $\cos_{Char}$)

ค่าความละม้ายโคไซน์ของเอ็นแกรมของอักขระ (cosine similarity of character *n*-gram) เป็นการประยุกต์แนวคิด 2 แนวในการประมวลผลภาษาธรรมชาติเข้าไว้ด้วยกัน ได้แก่ แนวคิดเรื่องความละม้ายโคไซน์ (cosine similarity) และแนวคิดเรื่องเอ็นแกรม (*n*-gram)



ภาพที่ 5.3 ความละม้ายโคไซน์ระหว่างเวกเตอร์ v กับ w

ความละม้ายโคไซน์หรือความละม้ายเชิงมุมเป็นแนวคิดที่แพร่หลายในงานด้านการค้นคืนสารสนเทศ (information retrieval) แนวคิดหลักในการคำนวณค่าดังกล่าวคือการวัดค่าความละม้ายระหว่างเวกเตอร์ไม่ศูนย์ (non-zero vector) 2 เวกเตอร์ในพื้นที่ผลคูณภายใน เวกเตอร์ที่ละม้ายกันจะทำมุมระหว่างกันน้อย ภาพที่ 5.3 แสดงตัวอย่างการวัดค่าความละม้ายโคไซน์ระหว่างเวกเตอร์ v กับ w จากภาพจะเห็นได้ว่าเวกเตอร์ v และ w ทำมุมระหว่างกัน (θ) เป็นมุมแคบ แสดงให้เห็นว่าเวกเตอร์ทั้งสองมีความละม้ายกันสูงในระดับหนึ่ง

ในแง่การคำนวณ ความละม้ายโคไซน์สามารถคำนวณได้จากการนำผลคูณ (dot product) ของเวกเตอร์ 2 เวกเตอร์แล้วนำมาหารด้วยขนาดของทั้งสองเวกเตอร์ วิธีการนี้จะทำให้ได้ค่าบรรทัดฐานของผลคูณ (normalized dot product) ซึ่งให้ค่าเดียวกันกับค่าโคไซน์ของมุมระหว่าง 2 เวกเตอร์ (Jurafsky & Martin, 2009, p. 665) ดังแสดงในสมการที่ 5.2 ความละม้ายโคไซน์จะมีค่าเป็นตัวเลขตั้งแต่ 0 ถึง 1 หากมีค่าเท่ากับ 0 แปลว่าคู่ของสายอักขระไม่มีความละม้ายกันเลย แต่หากค่าเท่ากับ 1 แปลว่าคู่ของสายอักขระเหมือนกันทุกประการ

$$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \tag{5.2}$$

ส่วนเอ็นแกรม (*n*-gram) นั้นเป็นแนวคิดที่ประยุกต์ความรู้เรื่องความน่าจะเป็นมาสร้างเป็นแบบจำลองซึ่งใช้ทำนายหน่วยถัดไปจากหน่วยที่ $n - 1$ (Jurafsky & Martin, 2009, p. 83) หากกำหนดให้หน่วยดังกล่าวเป็นอักขระ เอ็นแกรมจะเป็นลำดับ *n* โทเค็นของอักขระ ดังตัวอย่างต่อไปนี้

กำหนดให้สายอักขระ *S* = “AABBBCCDDD”

2 แกรม (โดยทั่วไปเรียกว่า “ไบแกรม”) ของสายอักขระ *S* จะได้แก่ “AA”, “AB”, “BB”, “BB”, “BC”, “CC”, “CD”, “DD”, และ “DD”

3 แกรม (โดยทั่วไปเรียกว่า “ไตรแกรม”) ของสายอักขระ *S* จะได้แก่ “AAB”, “ABB”, “BBB”, “BBC”, “BCC”, “CCD”, และ “DDD”

ในกรณีของค่าความละม้ายโคไซน์ของเอ็นแกรมของอักขระนี้ ผู้วิจัยได้ประยุกต์แนวคิดทั้ง 2 แนวที่ได้กล่าวมาข้างต้นมาใช้เป็นลักษณ์ โดยแปลงเอ็นแกรมของสายอักขระของข้อความในคู่หน่วยเทียบแต่ละข้อความให้อยู่ในรูปของเวกเตอร์จำนวนนับ (count vector) ของเอ็นแกรมของอักขระ ยกตัวอย่างเช่นในกรณีของสายอักขระ *S* ข้างต้นจะสามารถแปลงเป็นเวกเตอร์จำนวนนับของไบแกรมของอักขระได้ดังนี้

อักขระไบแกรม	AA	AB	BB	BC	CC	CD	DD
เวกเตอร์	1	1	2	1	1	1	2

S_{bigram} = [1 1 2 1 1 1 2]

จากตัวอย่างข้างต้น จะเห็นได้ว่าจำนวนนับของอักขระไบแกรม S_{bigram} เกิดจากการนับจำนวนครั้งของรูปไบแกรมของอักขระที่เกิดขึ้นในสายอักขระ

ผู้วิจัยได้อาศัยแนวคิดข้างต้นนี้สร้างเวกเตอร์จำนวนนับของเอ็นแกรมของอักขระของแต่ละย่อหน้าในคู่หน่วยเทียบขั้น โดยใช้ชุดของรูปอักขระเอ็นแกรมจากย่อหน้าทั้ง 2 ย่อหน้าร่วมกันเพื่อให้เวกเตอร์ที่ได้มีขนาดเท่ากัน จากนั้นจึงนำเวกเตอร์ที่ได้จากแต่ละย่อหน้ามาคำนวณค่าความละม้ายโคไซน์ตามสมการที่ 5.2 แล้วนำค่าความละม้ายที่วัดได้จากคู่หน่วยเทียบแต่ละคู่มาใช้เป็นลักษณ์

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 5 ลักษณ์ ตามช่วงของเอ็นแกรมตั้งแต่ 1 ถึง 5 แกรม ได้แก่

1) ค่าความละม้ายโคไซน์ของยูนิแกรมของอักขระ (cosine similarity of character unigram: $\cos_{\text{Char1}}$)
2) ค่าความละม้ายโคไซน์ของไบแกรมของอักขระ (cosine similarity of character bigram: $\cos_{\text{Char2}}$)

3) ค่าความละม้ายโคไซน์ของไตรแกรมของอักขระ (cosine similarity of character trigram: cos_{Char3})
4) ค่าความละม้ายโคไซน์ของ 4 แกรมของอักขระ (cosine similarity of character 4-gram: cos_{Char4})
5) ค่าความละม้ายโคไซน์ของ 5 แกรมของอักขระ (cosine similarity of character 5-gram: cos_{Char5})

จะเห็นได้ว่าลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของอักขระทั้ง 5 ลักษณ์ข้างต้นได้ประยุกต์ใช้ขนาดของเอ็นแกรมที่ต่างกัน ผู้วิจัยเชื่อว่าลักษณ์แต่ละตัวย่อมให้ตรวจจับการแก้ไขข้อความที่ผ่านการลักลอกได้ในระดับที่ต่างกัน อันจะส่งผลต่อประสิทธิภาพในจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกที่แตกต่างกันด้วย

5.1.7 ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของเอ็นแกรมของอักขระ (Jaccard similarity coefficient of character *n*-gram: J_{Char})

ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ด (Jaccard similarity coefficient) (Jaccard, 1908, 1912) เป็นสถิติที่ประยุกต์แนวคิดในทฤษฎีเซตเพื่อนำมาใช้เปรียบเทียบความละม้ายและความหลากหลายของกลุ่มตัวอย่าง เมื่อแรกเริ่มค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดถูกเสนอขึ้นเพื่อเปรียบเทียบความหลากหลายในเชิงพฤกษศาสตร์ ต่อมาจึงแพร่หลายไปสู่วงการอื่นๆ โดยเฉพาะอย่างยิ่ง ในงานค้นคืนสารสนเทศ

แนวคิดของค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดคือการวัดค่าความละม้ายระหว่างกลุ่มประชากร 2 กลุ่ม โดยคำนวณจากขนาดของประชากรที่ทั้งสองกลุ่มตัวอย่างมีร่วมกัน (อินเตอร์เซกชันในทฤษฎีเซต) หารด้วยขนาดของประชากรทั้งหมดจากทั้งสองกลุ่มตัวอย่าง (ยูเนียนในทฤษฎีเซต)

ในขั้นนี้ ผู้วิจัยได้ทำโปรแกรมเพื่อนำค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดมาใช้วัดความละม้ายระหว่างข้อความภายในคู่หน่วยเทียบในระดับเอ็นแกรมของอักขระ จากนั้นจึงนำค่าที่คำนวณได้มาใช้เป็นลักษณ์

กำหนดให้เอ็นแกรมของอักขระมีความยาว n, $S(A, n)$ แทนชุดของเอ็นแกรมของอักขระในข้อความ A, $S(B, n)$ แทนชุดของเอ็นแกรมของอักขระในข้อความ B, J_n ซึ่งแทนค่าความละม้ายระหว่างข้อความ A กับข้อความ B จะหาได้จากสมการที่ 5.3 ค่า J_n ที่คำนวณได้จะเป็นตัวเลขระหว่าง 0 ถึง 1 หากค่า J_n เท่ากับ 0 แปลว่าข้อความ A และข้อความ B ไม่ปรากฏชุดของเอ็นแกรมของอักขระร่วมกันเลย แต่หากค่า J_n เท่ากับ 0 แปลว่าข้อความ A และข้อความ B เหมือนกันทุกประการ

$$j_{n} = \frac{|S(A,n) \cap S(B,n)|}{|S(A,n) \cup S(B,n)|} \quad (5.3)$$

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 5 ลักษณ์ ตามช่วงของเอ็นแกรมตั้งแต่ 1 ถึง 5 แกรม ได้แก่

1) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของอักขระ (Jaccard similarity coefficient of character unigram: J_{Char1})
2) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของอักขระ (Jaccard similarity coefficient of character bigram: J_{Char2})
3) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของอักขระ (Jaccard similarity coefficient of character trigram: J_{Char3})
4) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของอักขระ (Jaccard similarity coefficient of character 4-gram: J_{Char4})
5) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของอักขระ (Jaccard similarity coefficient of character 5-gram: J_{Char5})

ทั้งนี้ เช่นเดียวกับค่าความละม้ายโคไซน์ของเอ็นแกรมของอักขระ ลักษณ์ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดได้ประยุกต์ใช้ขนาดของเอ็นแกรมของอักขระที่ต่างกัน ลักษณ์แต่ละตัวย่อมให้ตรวจจับการแก้ไขข้อความที่ผ่านการลักลอกได้ในระดับที่ต่างกัน และมีประสิทธิภาพในจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกที่แตกต่างกัน

5.1.8 ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของเอ็นแกรมของอักขระ (Sørensen–Dice coefficient of character *n*-gram: QS_{Char})

ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ (Sørensen-Dice coefficient similarity) (Dice, 1945; Sørensen, 1948) เป็นสถิติที่ใช้ในการเปรียบเทียบความละม้ายของกลุ่มตัวอย่าง 2 กลุ่ม แนวคิดพื้นฐานของค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์นั้นคล้ายกับค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ด กล่าวคือ ประยุกต์ใช้ทฤษฎีเซตในการคำนวณค่าความละม้ายจากจำนวนสมาชิกที่ทั้งสองตัวอย่างมีร่วมกันต่อจำนวนสมาชิกทั้งหมด แต่หลักการคำนวณค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์จะลดผลกระทบจากการมีสมาชิกร่วมกันของกลุ่มตัวอย่าง ดังนั้นจึงสามารถพิจารณาค่าที่คำนวณได้ในฐานะค่าความละม้ายของข้อมูลทั้งชุด

ในขั้นนี้ ผู้วิจัยได้ทำโปรแกรมเพื่อนำค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์มาใช้วัดความละม้ายระหว่างข้อความภายในคู่หน่วยเทียบในระดับเอ็นแกรมของอักขระ จากนั้นจึงนำค่าที่คำนวณได้มาใช้เป็นลักษณ์

ทั้งนี้ กำหนดให้เอ็นแกรมของอักขระมีความยาว *n*, *S*(*A*, *n*) แทนชุดของเอ็นแกรมของอักขระในข้อความ A, *S*(*B*, *n*) แทนชุดของเอ็นแกรมของอักขระในข้อความ B, QS_n ซึ่งแทนผลหารของความละม้าย (quotient of similarity) ระหว่างข้อความ A กับข้อความ B จะหาได้จากสมการที่ 5.4 ค่า QS_n เป็นตัวเลขระหว่าง 0 ถึง 1 หากค่า QS_n เท่ากับ 0 แปลว่าข้อความ A และข้อความ B ไม่ปรากฏชุดของเอ็นแกรมร่วมกันเลย แต่หากค่า QS_n เท่ากับ 0 แปลว่าข้อความ A และข้อความ B เหมือนกันทุกประการ

$$QS_n = \frac{2|S(A,n) \cap S(B,n)|}{|S(A,n)| + |S(B,n)|} \tag{5.4}$$

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 5 ลักษณ์ ตามช่วงของเอ็นแกรมตั้งแต่ 1 ถึง 5 แกรม ได้แก่

1) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของอักขระ (Sørensen–Dice similarity coefficient of character unigram: QS_{Char1})
2) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของอักขระ (Sørensen–Dice similarity coefficient of character bigram: QS_{Char2})
3) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของอักขระ (Sørensen–Dice similarity coefficient of character trigram: QS_{Char3})
4) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของอักขระ (Sørensen–Dice similarity coefficient of character 4-gram: QS_{Char4})
5) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของอักขระ (Sørensen–Dice similarity coefficient of character 5-gram: QS_{Char5})

จะเห็นได้ว่าลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ได้ประยุกต์ใช้ขนาดของเอ็นแกรมของอักขระที่ต่างกัน จึงอาจส่งผลให้มีประสิทธิภาพตรวจจับการแก้ไขข้อความที่ผ่านการลักลอกได้ในระดับที่ต่างกัน และมีประสิทธิภาพในจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกที่แตกต่างกัน

5.2 ลักษณ์ทางภาษา

ลักษณ์ทางภาษาคือลักษณ์ที่วิเคราะห์หาและสร้างขึ้นโดยอิงจากลักษณะทางภาษาของข้อความที่มีการลักลอกและไม่มีการลักลอก ในหัวข้อนี้ ผู้วิจัยจะนำเสนอผลที่ได้จากการวิเคราะห์หาลักษณ์ทางภาษาตามวิธีการวิจัยที่ได้กล่าวไปแล้วในหัวข้อที่ 3.4 เพื่อนำมาใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก

ลักษณ์ทางภาษาที่จะนำเสนอต่อไปนี้ถือเป็นข้อค้นพบที่ได้จากการวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทย การสังเกตลักษณะทางภาษาของข้อความที่มีการลักลอกและไม่มีการลักลอก แล้วนำมาประยุกต์เข้ากับองค์ความรู้และเทคนิควิธีทางการประมวลผลภาษาธรรมชาติที่มีใช้อยู่ในปัจจุบัน ในการนำเสนอ ผู้วิจัยได้แบ่งลักษณ์ทั้งหมด 51 ลักษณ์ออกเป็น 4 ประเภทใหญ่ตามระดับของหน่วยทางภาษา ได้แก่ ลักษณ์ทางศัพท์ ลักษณ์ทางวายสัมพันธ์ ลักษณ์ทางความหมาย และลักษณ์ทางวากยสัมพันธ์และความหมาย รายละเอียดมีดังต่อไปนี้

5.2.1 ลักษณ์ทางศัพท์

ลักษณ์ทางศัพท์คือลักษณ์ที่วิเคราะห์หาและสร้างขึ้นโดยอาศัยความรู้ทางภาษาศาสตร์ในระดับคำ ในแง่การตรวจหาการลักลอก ลักษณ์ทางศัพท์ต้องเอื้อให้เครื่องได้เรียนรู้ถึงขอบเขต ลักษณะและคุณสมบัติของคำในภาษา และด้วยเหตุนี้ กระบวนการวิเคราะห์หาและสร้างลักษณ์ประเภทจึงต้องอาศัยเครื่องมือช่วยในการกำกับขอบเขตของคำในขั้นก่อนการประมวลผล

ในหัวข้อย่อยนี้ ผู้วิจัยจะนำเสนอรายละเอียดของลักษณ์ทางศัพท์ทั้งหมด 25 ลักษณ์ซึ่งเป็นผลจากการวิเคราะห์หาตามวิธีการวิจัยที่ได้กำหนดไว้ โดยแบ่งการนำเสนอออกเป็น 9 กลุ่มย่อย ดังรายละเอียดต่อไปนี้

5.2.1.1 ขนาดของคู่หน่วยเทียบ (pair size: $Size_W$)

แนวคิดหลักของการใช้ขนาดของคู่หน่วยเทียบในระดับเป็นลักษณ์ในที่นี้เป็นเช่นเดียวกันกับการนำขนาดของคู่หน่วยเทียบในระดับอักขระมาใช้เป็นลักษณ์ที่กล่าวไปแล้วในหัวข้อ 5.1.1 กล่าวคืออิงจากข้อค้นพบที่ว่าย่อหน้าที่ผ่านการลักลอกย่อมมีความยาวแตกต่างจากย่อหน้าต้นฉบับ การนำขนาดของแต่ละย่อหน้าในคู่หน่วยเทียบมาใช้เป็นลักษณ์จึงอาจเอื้อให้เครื่องสามารถจำแนกประเภทข้อความที่มีการลักลอกและไม่ลักลอกออกจากกันได้

ในระดับคำ ขนาดของคู่หน่วยเทียบได้มากจากการนับจำนวนคำที่ปรากฏในย่อหน้าแต่ละย่อหน้าของคู่หน่วยเทียบ เนื่องจากคลังข้อมูลที่สร้างขึ้นสำหรับใช้ในงานวิจัยชิ้นนี้มีการกำกับข้อมูลขอบเขตของคำไว้แล้ว ในการวิเคราะห์และสร้างลักษณ์ชนิดนี้ ผู้วิจัยจึงสามารถเขียนโปรแกรมเพื่อนับ

จำนวนคำที่ปรากฏในแต่ละย่อหน้าของคู่หน่วยเทียบแล้วนำมาใช้เป็นลักษณ์ได้ทันที ลักษณ์ที่ได้จึงมีลักษณะเป็นจำนวนนับ เมื่อนำมาใช้ทดลองจำแนกประเภทของข้อความ ความยาวของทั้งสองย่อหน้าจะถูกป้อนเข้าเครื่องพร้อมกันเป็นคู่

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.1.2 ผลต่างของขนาดของคู่หน่วยเทียบ (difference of pair size: $diff_W$)

ลักษณ์ผลต่างของขนาดของคู่หน่วยเทียบในระดับคำก็เป็นลักษณ์ที่ประยุกต์ขึ้นตามแนวคิดเดียวกันกับลักษณ์ผลต่างของขนาดของคู่หน่วยเทียบในระดับอักขระ กล่าวคือ เป็นลักษณ์มุ่งพิจารณาความแตกต่างระหว่างคู่หน่วยเทียบเป็นสำคัญ โดยอาศัยความแตกต่างกันด้านความยาวของย่อหน้าในคู่เทียบมาใช้เป็นลักษณ์ให้เครื่องเรียนรู้โดยตรง

ในการวิเคราะห์หาและสร้างลักษณ์ชนิดนี้ ผู้วิจัยจะนำจำนวนคำที่นับได้จากย่อหน้าข้อความต้นฉบับลบด้วยจำนวนคำที่นับได้จากย่อหน้าข้อความลักลอก ค่าที่ได้จากการคำนวณจะมีทั้งจำนวนเต็มบวกในกรณีที่ข้อความในย่อหน้าลักลอกมีขนาดสั้นกว่าข้อความในย่อหน้าต้นฉบับ จำนวนเต็มลบในกรณีที่ข้อความในย่อหน้าลักลอกมีขนาดยาวกว่าข้อความในย่อหน้าต้นฉบับ และมีค่าเป็น 0 ในกรณีที่ข้อความในย่อหน้าต้นฉบับและข้อความในย่อหน้าลักลอกมีขนาดเท่ากัน

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.1.3 ค่าระยะการแก้ไขเลเวนชเตย์นของคำ (Levenshtein edit distance of word: LD_W)

ลักษณ์ค่าระยะการแก้ไขเลเวนชเตย์นของคำเป็นลักษณ์ที่วิเคราะห์หาและสร้างโดยประยุกต์ใช้แนวคิดเรื่องระยะการแก้ไขเลเวนชเตย์น (Levenshtein edit distance) เช่นเดียวกันกับลักษณ์ค่าระยะการแก้ไขเลเวนชเตย์นของอักขระที่ได้กล่าวไปแล้วในหัวข้อ 5.1.3 แต่ปรับเปลี่ยนให้พิจารณาการแก้ไขรูปคำแทนการแก้ไขรูปอักขระ

การประยุกต์แนวคิดเรื่องระยะการแก้ไขเลเวนชเตย์นในระดับคำนี้จะช่วยสะท้อนให้เห็นถึงความแตกต่างระหว่างข้อความต้นฉบับกับข้อความที่ถูกลักลอกซึ่งเป็นผลจากการแก้ไขของผู้ลักลอกได้โดยพิจารณาจากรูปคำที่ปรากฏอยู่ในข้อความต้นฉบับและข้อความลักลอก ยกตัวอย่างเช่น ข้อความ “เขามีสุนัขตัวใหญ่” และ “เขาเลี้ยงสุนัขใหญ่ไว้ในบ้าน”ข้อความทั้งสองนี้จะมีค่าระยะการแก้ไขเลเวนชเตย์นเท่ากับ 5 เนื่องจากมีกระบวนการแก้ไขเกิดขึ้นทั้งหมด 5 ครั้ง ได้แก่

1) “เขา|*มี*|สุนัข|ตัว|ใหญ่” ⟶ “เขา|*เลี้ยง*|สุนัข|ตัว|ใหญ่” มีการแก้ไขโดยแทนที่คำว่า “มี” ด้วย “เลี้ยง”
2) “เขา|เลี้ยง|สุนัข|*ตัว*|ใหญ่” ⟶ “เขา|เลี้ยง|สุนัข|ใหญ่” มีการแก้ไขโดยลบคำว่า “ตัว” ออกไป
3) “เขา|เลี้ยง|สุนัข|ใหญ่” ⟶ “เขา|*เลี้ยง*|สุนัข|ใหญ่|*ไว้*” มีการแก้ไขโดยแทรกคำว่า “ไว้” ท้ายข้อความ
4) “เขา|เลี้ยง|สุนัข|ใหญ่|ไว้” ⟶ “เขา|เลี้ยง|สุนัข|ใหญ่|ไว้|*ใน*” มีการแก้ไขโดยแทรกคำว่า “ใน” ท้ายข้อความ
5) “เขา|เลี้ยง|สุนัข|ใหญ่|ไว้|ใน” ⟶ “เขา|เลี้ยง|สุนัข|ใหญ่|ไว้|ใน|*บ้าน*” มีการแก้ไขโดยแทรกคำว่า “บ้าน” ท้ายข้อความ

ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้เขียนโปรแกรมคำนวณค่าระยะการแก้ไขเลเวนชเตย์นในระดับคำขึ้นด้วยภาษาไพทอน จากนั้นจึงใช้โปรแกรมดังกล่าวคำนวณหาค่าระยะการแก้ไขของคู่หน่วยเทียบทุกคู่ในคลังข้อมูลแล้วนำมาค่าที่ได้มาใช้เป็นลักษณ์

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.1.4 ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ (length of longest common subsequence of word: len(lcs_w))

ลักษณ์ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำเป็นลักษณ์ที่วิเคราะห์หาและสร้างโดยประยุกต์ใช้แนวคิดเรื่องลำดับย่อยร่วมที่ยาวที่สุด (longest common subsequence: *lcs*) เช่นเดียวกันกับลักษณ์ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของสายอักขระที่ได้กล่าวไปแล้วในหัวข้อที่ 5.1.4 แต่ลักษณ์ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำนี้ได้ปรับเปลี่ยนระดับของภาษาที่ใช้พิจารณาให้สูงคือจากการพิจารณารูปอักขระเป็นการพิจารณารูปคำแทน โดยมีแนวคิดว่าลำดับย่อยร่วมที่ยาวที่สุดของคำจะแสดงคำที่คงเหลืออยู่จากการแก้ไข ยกตัวอย่างเช่นข้อความ “เขามีสุนัขตัวใหญ่” และ “เขาเลี้ยงสุนัขใหญ่ไว้ในบ้าน” ดังนี้

S_1 = “<u>เขา</u>|มี|<u>สุนัข</u>|ตัว|<u>ใหญ่</u>”
S_2 = “<u>เขา</u>|เลี้ยง|<u>สุนัข</u>|<u>ใหญ่</u>|ไว้|ใน|บ้าน”
$lcs_w(S_1,S_2)$ = “เขา|สุนัข|ใหญ่”
$\text{len}(lcs_W(S_1,S_2))$ = 3

จากตัวอย่างข้อความข้างต้น จะเห็นได้ว่าลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำของข้อความ S_1 และ S_2 ($lcs_w(S_1,S_2)$) คือ “เขา|สุนัข|ใหญ่” ลำดับย่อยร่วมนี้ประกอบด้วยคำจำนวน 3 คำ ดังนั้น ความยาวของลำดับย่อยร่วมดังกล่าว ($\text{len}(lcs_W(S_1,S_2))$) จึงเท่ากับ 3

ส่วนการสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้เขียนโปรแกรมเพื่อหาลำดับย่อยร่วมที่ยาวที่สุดของคำภายในคู่หน่วยเทียบก่อน จากนั้นจึงหาความยาวของลำดับย่อยร่วมดังกล่าว โดยเขียนโปรแกรมเพื่อนับจำนวนคำปรากฏในลำดับย่อยร่วมที่ยาวที่สุดที่หาได้ กระบวนการหาลำดับย่อยร่วมที่ยาวที่สุดของคำและหาความยาวของลำดับย่อยร่วมที่ยาวที่สุดจะดำเนินไปจนกระทั่งได้ข้อมูลครบจากทุกคู่หน่วยเทียบในคลังข้อมูล จากนั้นจึงนำค่าความยาวที่หาได้มาใช้เป็นลักษณ์

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.1.5 ค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ (normalized longest common subsequence of word: $lcs_{\text{norm-W}}$)

ลักษณ์ค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำเป็นลักษณ์ที่อาศัยแนวคิดเรื่องการนำลำดับย่อยร่วมยาวสุดที่ยาวที่สุดมาคำนวณเป็นค่าความละม้ายเช่นเดียวกันกับลักษณ์ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระที่ได้กล่าวไปแล้วในหัวข้อที่ 5.1.5 แต่ปรับเปลี่ยนระดับของหน่วยทางภาษาที่ใช้พิจารณาให้สูงคือจากการพิจารณารูปอักขระเป็นการพิจารณารูปคำ

ดังได้กล่าวไปแล้วในหัวข้อที่ 5.1.5 ว่าค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดคำนวณได้จากการนำความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดมาหารด้วยค่าความยาวของข้อความใดข้อความหนึ่งจากทั้งสองจากทั้งสองข้อความ ซึ่งในงานชิ้นนี้จะกำหนดให้หารด้วยค่าความยาวของข้อความลักลอกเสมอ ทั้งนี้ เพื่อให้เข้าใจชัดเจนยิ่งขึ้น จะขอตัวอย่างการคำนวณค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำระหว่างข้อความ “เขามีสุนัขตัวใหญ่” กับ “เขาเลี้ยงสุนัขใหญ่ไว้ในบ้าน” ดังนี้

S_1 = “<u>เขา</u>|มี|<u>สุนัข</u>|ตัว|<u>ใหญ่</u>”

S_2 = “<u>เขา</u>|เลี้ยง|<u>สุนัข</u>|<u>ใหญ่</u>|ไว้|ใน|บ้าน”

$lcs_w(S_1,S_2)$ = “เขา|สุนัข|ใหญ่”

$\text{len}(lcs_W(S_1,S_2))$ = 3

$\text{len}(S_2)$ = 7

$lcs_{\text{norm-W}} = \text{len}(lcs_W(S_1,S_2))/\text{len}(S_2) = 3/7 = 0.4286$

ตัวอย่างข้างต้นเป็นการคำนวณค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของคำโดยการนำค่าความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของทั้งสองข้อความ ($\text{len}(lcs_W(S_1,S_2))$) มาหารด้วยค่าความยาวของข้อความลักลอก ($\text{len}(S_2)$) ผลที่ได้ออกมาจึงเท่ากับ 0.4286

เนื่องด้วยค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของคำเป็นค่าความละม้ายจึงสามารถแสดงความแตกต่างระหว่างข้อความลักลอกกับข้อความต้นฉบับที่เกิดจากการแก้ไขในระดับคำออกมาในเชิงปริมาณได้ ด้วยเหตุนี้ ผู้วิจัยจึงนำค่าดังกล่าวมาใช้เป็นลักษณ์ในการทดสอบประสิทธิภาพการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก

ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้แก้ไขโปรแกรมสำหรับหาลำดับย่อยร่วมที่ยาวที่สุดของคำภายในคู่หน่วยเทียบที่มีอยู่เดิมให้สามารถคำนวณค่าความละม้ายได้ดังตัวอย่างที่ยกมาข้างต้นได้ จากนั้นจึงนำไปใช้หาค่า $lcs_{\text{norm-W}}$ ของคู่หน่วยเทียบทุกคู่ในคลังข้อมูลแล้วนำค่าที่ได้มาใช้เป็นลักษณ์

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.1.6 ค่าความละม้ายโคไซน์ของเอ็นแกรมของคำ (cosine similarity of word n-gram: cos_W)

ลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของคำอาศัยแนวคิดในการวิเคราะห์หาและสร้างเช่นเดียวกันกับลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของอักขระ กล่าวคือ เป็นการประยุกต์แนวคิดเรื่องความละม้ายโคไซน์ (cosine similarity) และแนวคิดเรื่องเอ็นแกรม (n-gram) เข้าไว้ด้วยแต่เอ็นแกรมที่ใช้ในลักษณ์ประเภทนี้เป็นเอ็นแกรมของคำที่ปรากฏในข้อความ

เช่นเดียวกันกับค่าความละม้ายโคไซน์ของเอ็นแกรมของอักขระ ในการสร้างลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของคำ ผู้วิจัยต้องแปลงเอ็นแกรมของคำของข้อความในคู่หน่วยเทียบแต่ละข้อความให้อยู่ในรูปของเวกเตอร์จำนวนนับ (count vector) ของเอ็นแกรมของคำก่อนในขั้นแรก ยกตัวอย่างเช่น ในกรณีของข้อความ “เขามีสุนัขตัวใหญ่” และ “เขาเลี้ยงสุนัขตัวใหญ่” จะสามารถแปลงเป็นเวกเตอร์จำนวนนับของไบแกรมของคำได้ดังนี้

S_1 = “เขา|มี|สุนัข|ตัว|ใหญ่”
S_2 = “เขา|เลี้ยง|สุนัข|ตัว|ใหญ่”

ไบแกรมของคำ	“ตัว|ใหญ่”	“มี|สุนัข”	“สุนัข|ตัว”	“เขา|มี”	“เขา|เลี้ยง”	“เลี้ยง|สุนัข”
เวกเตอร์ของ S_1	1	1	1	1	0	0
เวกเตอร์ของ S_2	1	0	1	0	1	1

$\overrightarrow{S_1}$ = [1 1 1 1 0 0]

$\overrightarrow{S_2}$ = [1 0 1 0 1 1]

จากตัวอย่างข้างต้น จะเห็นได้ว่าข้อความ S_1 และ S_2 มีชุดของไบแกรมของคำร่วมกันอยู่ 6 ชุด ได้แก่ “ตัว|ใหญ่”, “มี|สุนัข”, “สุนัข|ตัว”, “เขา|มี”, “เขา|เลี้ยง”, และ “เลี้ยง|สุนัข” เวกเตอร์จำนวนนับ $\overrightarrow{S_1}$ และ $\overrightarrow{S_2}$ ได้มาจากการนับจำนวนชุดของไบแกรมดังกล่าวที่ปรากฏในข้อความ S_1 และ S_2 ตามลำดับ

จากนั้น ผู้วิจัยจะนำเวกเตอร์ที่แปลงได้ทั้งสองเวกเตอร์ข้างต้นมาคำนวณค่าความละม้ายโคไซน์ตามสมการที่ 5.2 ที่ได้แสดงในหัวข้อที่ 5.1.6 แล้ว จะได้ค่าความละม้ายระหว่างข้อความทั้งสองเท่ากับ 0.5 และนำค่าที่ได้นี้มาใช้เป็นลักษณ์ กระบวนการสร้างลักษณ์ดังกล่าวมานี้จะดำเนินไปจนกระทั่งได้ค่าความละม้ายจากทุกคู่หน่วยเทียบในคลังข้อมูลครบ

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 5 ลักษณ์ ตามช่วงของเอ็นแกรมตั้งแต่ 1 ถึง 5 แกรม ได้แก่

1) ค่าความละม้ายโคไซน์ของยูนิแกรมของคำ (cosine similarity of word unigram: $\cos_{W1}$)
2) ค่าความละม้ายโคไซน์ของไบแกรมของคำ (cosine similarity of word bigram: $\cos_{W2}$)
3) ค่าความละม้ายโคไซน์ของไตรแกรมของคำ (cosine similarity of word trigram: $\cos_{W3}$)
4) ค่าความละม้ายโคไซน์ของ 4 แกรมของคำ (cosine similarity of word 4-gram: $\cos_{W4}$)
5) ค่าความละม้ายโคไซน์ของ 5 แกรมของคำ (cosine similarity of word 5-gram: $\cos_{W5}$)

ลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของคำทั้ง 5 ลักษณ์ข้างต้นได้ประยุกต์ใช้ขนาดของเอ็นแกรมที่ต่างกัน ผู้วิจัยเชื่อว่าลักษณ์แต่ละตัวย่อมให้ตรวจจับการแก้ไขข้อความที่ผ่านการลักลอกได้ในระดับที่ต่างกัน นอกจากนี้แล้ว การประยุกต์ใช้ขนาดของเอ็นแกรมที่ต่างกันนี้ยังช่วยแก้ไขปัญหาที่เกิดขึ้นจากแนวคิดเรื่องถุงใส่คำ (bag-of-words) ด้วย

ถุงใส่คำ (bag-of-words) หมายถึงชุดของคำที่ไม่เรียงตามลำดับการปรากฏจริงในข้อความ (Jurafsky & Martin, 2009, p. 641) แนวคิดดังกล่าวนี้เป็นผลมาจากการจัดการกับข้อความโดยมุ่ง

ความสำคัญไปที่รูปคำโดยไม่สนใจลำดับในการปรากฏของคำในข้อความซึ่งเป็นการพิจารณาข้อความในระดับวากยสัมพันธ์ ด้วยเหตุนี้จึงอาจก่อให้เกิดปัญหาบางประการตามมา ยกตัวอย่างเช่นกรณีของการคำนวณค่าความละม้ายของยูนิแกรมของคำระหว่างข้อความต่อไปนี้

S_1 = “แมว|ขโมย|ปลา|ของ|น้อง|ไป|กิน”

S_2 = “น้อง|ขโมย|ปลา|ของ|แมว|ไป|กิน”

ยูนิแกรมของคำ	“กิน”	“ของ”	“ขโมย”	“น้อง”	“ปลา”	“แมว”	“ไป”
เวกเตอร์ของ S_1	1	1	1	1	1	1	1
เวกเตอร์ของ S_2	1	1	1	1	1	1	1

$\overrightarrow{S_1}$ = [1 1 1 1 1 1 1]

$\overrightarrow{S_2}$ = [1 1 1 1 1 1 1]

$\cos_{W1}(\overrightarrow{S_1}, \overrightarrow{S_2})$ = 1.0

ตัวอย่างข้างต้นแสดงให้เห็นถึงปัญหาของแนวคิดเรื่องถุงใส่คำในกรณีของการวัดค่าความละม้ายโคไซน์ของยูนิแกรมของคำ จากตัวอย่างจะเห็นได้ว่าข้อความ S_1 และ S_2 แม้จะประกอบขึ้นจากรูปคำชุดเดียวกันแต่ก็มีความหมายแตกต่างกันเนื่องจากลำดับของคำไม่เหมือนกัน แต่เนื่องด้วยการวัดค่าความละม้ายจากยูนิแกรมของคำนั้นไม่สนใจลำดับของคำที่ปรากฏในแต่ละข้อความจึงทำให้แปลงเวกเตอร์จำนวนนับออกมาได้เหมือนกัน และเมื่อคำนวณค่าความละม้ายแล้วก็ให้ผลเท่ากับ 1 ซึ่งแปลว่าข้อความ S_1 และ S_2 เหมือนกันทุกประการ

อย่างไรก็ดี การประยุกต์ใช้ช่วงของเอ็นแกรมที่แตกต่างกันในการคำนวณค่าความละม้ายสามารถช่วยแก้ปัญหาข้างต้นได้ ยกตัวอย่างเช่นกรณีของการคำนวณค่าความละม้ายของไบแกรมของคำระหว่างข้อความ S_1 กับ S_2 ดังต่อไปนี้

ไบแกรมของคำ	“ของ|น้อง”	“ของ|แมว”	“ขโมย|ปลา”	“น้อง|-ขโมย”	“น้อง|ไป”	“ปลา|ของ”	“แมว|ขโมย”	“แมว|ไป”	“ไป|กิน”
เวกเตอร์ของ S_1	1	0	1	0	1	1	1	0	1
เวกเตอร์ของ S_2	0	1	1	1	0	1	0	1	1

$\overrightarrow{S_1}$ = [1 0 1 0 1 1 1 0 1]

$\overrightarrow{S_2}$ = [0 1 1 1 0 1 0 1 1]

$\cos_{W2}(\overrightarrow{S_1}, \overrightarrow{S_2})$ = 0.5

จากตัวอย่างข้างต้น จะเห็นได้ว่าเมื่อปรับช่วงของเอ็นแกรมของคำจากยูนิแกรมเป็นไบแกรมแล้ว ค่าความละม้ายโคไซน์ของข้อความ S_1 และ S_2 มีค่าเท่ากับ 0.5 ซึ่งสะท้อนความเป็นจริงว่าข้อความทั้งสองมีความหมายแตกต่างกัน จึงกล่าวได้ว่าการประยุกต์ใช้ช่วงของเอ็นแกรมที่แตกต่างกันสามารถแก้ปัญหาถุงใส่คำได้

อย่างไรก็ตาม การประยุกต์ใช้ช่วงของเอ็นแกรมที่แตกต่างกันในการวัดค่าความละม้ายย่อมให้ค่าความละม้ายที่แตกต่าง ดังตัวอย่างข้อความ S_1 และ S_2 ข้างต้นสามารถคำนวณค่าความละม้ายโคไซน์ของไตรแกรมของคำ, 4 แกรมของคำ, และ 5 แกรมของคำ ได้เท่ากับ 0.8, 1.0, และ 1.0 ตามลำดับ ทั้งนี้ สาเหตุที่ทำให้ค่าความละม้ายกลับเพิ่มมากขึ้นนั้นเนื่องมาจากจำนวนคำในแต่ละข้อความมีจำนวนเพียง 7 คำ เมื่อจำนวนของเอ็นแกรมเพิ่มขึ้น ชุดของเอ็นแกรมที่ได้จากข้อความจึงมีไม่มากพอที่จะสะท้อนความแตกต่างของข้อความได้ ด้วยเหตุนี้เอง ในงานวิจัยชิ้นนี้จึงได้กำหนดช่วงของเอ็นแกรมไว้ตั้งแต่ 1-5 แกรม เพื่อทดลองว่าเมื่อนำค่าความละม้ายมาใช้เป็นลักษณ์แล้ว ขนาดของเอ็นแกรมใดที่ให้ประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกดีที่สุด โดยเฉพาะอย่างยิ่ง ในกรณีของคลังข้อมูลที่ใช้ในงานวิจัยชิ้นนี้ที่ประกอบด้วยข้อความขนาดแตกต่างกัน

5.2.1.7 ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของเอ็นแกรมของคำ (Jaccard similarity coefficient of word n-gram: J_W)

ลักษณ์ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของเอ็นแกรมของคำเป็นลักษณ์ที่อาศัยแนวคิดเรื่องค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดและแนวคิดเรื่องเอ็นแกรมมาวิเคราะห์หาและสร้างเป็นลักษณ์เช่นเดียวกันกับค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของเอ็นแกรมของอักขระที่ได้กล่าวไปแล้ว

ในหัวข้อที่ 5.1.7 แต่ปรับเปลี่ยนระดับของหน่วยทางภาษาที่ใช้พิจารณาให้สูงคือจากการพิจารณารูปอักขระเป็นการพิจารณารูปคำ

ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้เขียนโปรแกรมเพื่อนำค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดมาใช้ในฐานะค่าความละม้ายในระดับเอ็นแกรมของคำระหว่างข้อความภายในคู่หน่วยเทียบ จากนั้นจึงนำค่าที่คำนวณได้มาใช้เป็นลักษณ์ โดยอาศัยการคำนวณตามสมการที่ 5.3 ที่ได้แสดงไปแล้วในหัวข้อที่ 5.1.7 แต่ปรับเปลี่ยนให้ใช้ชุดของเอ็นแกรมของคำในข้อความในการคำนวณแทนชุดของเอ็นแกรมของอักขระ กระบวนการสร้างลักษณ์จะดำเนินไปเช่นนี้จนกระทั่งได้ค่าความละม้ายจากคู่หน่วยเทียบครบทุกคู่

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 5 ลักษณ์ ตามช่วงของเอ็นแกรมตั้งแต่ 1 ถึง 5 แกรม ได้แก่

1) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของคำ (Jaccard similarity coefficient of word unigram: J_{W1})
2) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของคำ (Jaccard similarity coefficient of word bigram: J_{W2})
3) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของคำ (Jaccard similarity coefficient of word trigram: J_{W3})
4) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของคำ (Jaccard similarity coefficient of word 4-gram: J_{W4})
5) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของคำ (Jaccard similarity coefficient of word 5-gram: J_{W5})

ลักษณ์ดังกล่าวข้างต้นนี้ได้ประยุกต์ใช้ขนาดของเอ็นแกรมของคำที่ต่างกัน เช่นเดียวกับลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของคำ ด้วยเหตุนี้ ผู้วิจัยจึงเชื่อว่าลักษณ์แต่ละตัวย่อมในกลุ่มนี้จะให้ประสิทธิภาพในจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกที่แตกต่างกัน และสามารถแก้ไขปัญหาถุงใส่คำได้เช่นเดียวกับลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของคำ

5.2.1.8 ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของเอ็นแกรมของคำ (Sørensen–Dice coefficient of word n-gram: QS_W)

ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของเอ็นแกรมของคำเป็นลักษณ์อีกชนิดที่นำแนวคิดเรื่องค่าความละม้ายมาประยุกต์เข้ากับแนวคิดเรื่องเอ็นแกรม แนวคิดในการวิเคราะห์หาและสร้างลักษณ์ชนิดนี้คล้ายกับค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของเอ็นแกรมของอักขระที่

ได้กล่าวไปแล้วในหัวข้อที่ 5.1.8 แต่ปรับเปลี่ยนระดับของหน่วยทางภาษาที่พิจารณาจากระดับอักขระเป็นระดับคำ ด้วยเหตุนี้ ผู้วิจัยจึงเชื่อลักษณชนิดนี้จะสามารถจะจำแนกข้อความลักลอกที่ผ่านแก้ไขในระดับคำได้อย่างมีประสิทธิภาพ

ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้เขียนโปรแกรมเพื่อวัดค่าความละม้ายระหว่างคู่หน่วยเทียบในคลังข้อมูลในระดับเอ็นแกรมของคำจากนั้นจึงนำค่าที่คำนวณได้มาใช้เป็นลักษณ์ ในการนี้ วิธีการคำนวณค่าความละม้ายจะเป็นไปตามสมการที่ *5.4* ที่ได้แสดงไปแล้วในหัวข้อที่ 5.1.8 แต่ปรับเปลี่ยนให้ใช้ชุดของเอ็นแกรมของคำในข้อความในการคำนวณแทนชุดของเอ็นแกรมของอักขระ กระบวนการสร้างลักษณ์จะดำเนินไปเช่นนี้จนกระทั่งได้ค่าความละม้ายจากคู่หน่วยเทียบในคลังข้อมูลครบทุกคู่

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 5 ลักษณ์ ตามช่วงของเอ็นแกรมตั้งแต่ 1 ถึง 5 แกรม ได้แก่

1) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ (Sørensen–Dice similarity coefficient of word unigram: QS_{W1})
2) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (Sørensen–Dice similarity coefficient of word bigram: QS_{W2})
3) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของคำ (Sørensen–Dice similarity coefficient of word trigram: QS_{W3})
4) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของคำ (Sørensen–Dice similarity coefficient of word 4-gram: QS_{W4})
5) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของคำ (Sørensen–Dice similarity coefficient of word 5-gram: QS_{W5})

จะเห็นได้ว่าลักษณ์ในกลุ่มข้างต้นนี้ได้ประยุกต์ใช้ขนาดของเอ็นแกรมของคำที่ต่างกันเช่นเดียวกับลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของคำและลักษณ์ค่าความละม้ายแจ็กการ์ดของเอ็นแกรมของคำ ลักษณ์แต่ละตัวย่อมจึงให้ประสิทธิภาพในจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกที่แตกต่างกัน นอกจากนี้ยังสามารถแก้ไขปัญหาถุงใส่คำได้เช่นเดียวกับลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของคำและลักษณ์ค่าความละม้ายแจ็กการ์ดของเอ็นแกรมของคำ

5.2.1.9 ค่าความละม้ายโคไซน์ของน้ำหนักเอ็นแกรมของคำแบบ tf–idf (cosine similarity of tf-idf term weight n-gram: $cos_{tf\text{-}idf}$)

ในงานด้านการค้นคืนสารสนเทศ การให้น้ำหนักคำถือเป็นแนวคิดหนึ่งที่ได้รับการนำไปประยุกต์ใช้ในหลายแขนง ทั้งนี้ การให้น้ำหนักแบบ tf-idf (term frequency–inverse document frequency) (Spärck Jones, 1972) ก็เป็นวิธีการให้น้ำหนักคำวิธีหนึ่งที่นิยมใช้อย่างแพร่หลาย

แนวคิดเบื้องหลังของการให้น้ำหนักคำแบบ tf-idf คือการพิจารณาว่าคำแต่ละคำในเอกสารนั้นมีความสำคัญไม่เท่ากัน โดยคำที่ปรากฏน้อยหรือปรากฏอยู่อย่างจำกัดในเอกสารที่สนใจจะมีความสำคัญ สามารถใช้เป็นตัวแทนสำหรับจำแนกเอกสารเหล่านั้นจากเอกสารทั้งหมดที่เหลือ ส่วนคำที่ปรากฏมากในทุกๆ เอกสารจะไม่มีความสำคัญในแง่นี้ (Jurafsky & Martin, 2009, p. 771) โดยการอาศัยแนวคิดดังกล่าวนี้ การให้น้ำหนักคำคำหนึ่งจึงพิจารณาจากความถี่ในการปรากฏของคำ (term frequency: tf) ในเอกสารและจำนวนเอกสารทั้งหมด (document frequency: df) ที่มีคำนั้นๆ ปรากฏอยู่

ในแง่การคำนวณ tf คือความถี่ในการปรากฏของคำที่สนใจ (term) ในเอกสาร (document) หากค่า tf มากจะแสดงว่าคำที่สนใจปรากฏในเอกสารมาก ในขณะที่ df คือจำนวนของเอกสารที่ปรากฏคำที่สนใจอยู่ หากค่า df น้อยจะแสดงว่าคำที่สนใจปรากฏอยู่ในเอกสารไม่มาก ไม่ได้เป็นคำที่ปรากฏโดยทั่วไปในเอกสาร คำในลักษณะนี้จะมีอำนาจจำแนก (discrimination power) สูง การคิดค่า idf (inverse document frequency) จะเป็นค่าแทนอำนาจจำแนกดังกล่าว โดยคิดจากสมการที่ 5.5 จากสมการกำหนดให้ D เป็นจำนวนเอกสารทั้งหมดที่มีอยู่ในคลังข้อมูล ส่วน $d: t \in d$ นั้นหมายถึงจำนวนของเอกสารซึ่งมีคำ t ปรากฏอยู่ ทั้งนี้เมื่อหาค่า idf ได้แล้วก็จะสามารถหาค่า tf-idf ได้จากสมการที่ 5.6

$$idf(t) = \log \frac{|D|}{|\{d: t \in d\}|} \tag{5.5}$$

$$tf - idf(t,d) = tf(t,d) \times idf(t) \tag{5.6}$$

ในการวิเคราะห์หาลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักเอ็นแกรมของคำแบบ tf–idf นี้ ผู้วิจัยได้ประยุกต์ใช้แนวคิด 3 แนวคิดเข้าด้วยกัน ได้แก่ แนวคิดเรื่องการให้น้ำหนักคำแบบ tf-idf ที่ได้กล่าวไปข้างต้น แนวคิดเรื่องการวัดค่าความละม้ายโคไซน์ และแนวคิดเรื่องเอ็นแกรม โดยแทนรูปคำในคลังข้อมูลด้วยค่าน้ำแบบ tf-idf แล้วนำเอ็นแกรมของค่าน้ำหนักดังกล่าวมาวัดค่าความละม้ายโคไซน์ ด้วยวิธีดังกล่าวนี้ ลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักเอ็นแกรมของคำแบบ tf–idf จึงสามารถพิจารณาข้อความที่จะได้รับการจำแนกประเภทผ่านความสำคัญของแต่ละคำในข้อความที่

ผ่านการให้ค่าน้ำหนัก ในแง่นี้ ในกรณีที่ผู้ลักลอกแก้ไขข้อความต้นฉบับด้วยการแทนที่คำเดิมด้วยคำที่มีค่าน้ำหนักเท่ากัน หรือเปลี่ยนแปลงคำในข้อความด้วยคำที่มีค่าน้ำหนักแตกต่างออกไป ลักษณ์ชนิดนี้ก็จะตรวจหาการแก้ไขดังกล่าวได้

ส่วนการสร้างลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักเอ็นแกรมของคำแบบ tf–idf นั้น ในขั้นแรก ผู้วิจัยได้ใช้โปรแกรมเพื่อหาค่าน้ำหนักของแต่ละคำที่ปรากฏในคลังข้อมูลก่อน ในขั้นตอนนี้ต้องนำเข้าข้อมูลทั้งหมดคลังข้อมูลเพื่อสร้างเป็นคลังข้อมูลฝึกฝนของการให้น้ำหนักเอ็นแกรมของคำ เมื่อได้คลังข้อมูลฝึกฝนของการให้น้ำหนักคำแล้ว ในขั้นต่อมา ผู้วิจัยจะนำคู่หน่วยเทียบแต่ละคู่ไปแปลงเป็นเวกเตอร์ของน้ำหนักเอ็นแกรมของคำ เวกเตอร์ของน้ำหนักเอ็นแกรมของคำนี้เกิดจากการเทียบเอ็นแกรมของคำที่ปรากฏในข้อความกับเอ็นแกรมของคำที่มีอยู่คลังข้อมูลฝึกฝนของการให้น้ำหนักคำ เมื่อเครื่องพบว่ารูปเอ็นแกรมของคำที่ปรากฏในข้อความกับรูปเอ็นแกรมของคำที่มีอยู่คลังข้อมูลฝึกฝนของการให้น้ำหนักคำปรากฏตรงก็จะดึงเอาน้ำหนักเอ็นแกรมของคำของรูปเอ็นแกรมของคำที่เก็บไว้ในคลังข้อมูลฝึกฝนมาใช้แทนรูปเอ็นแกรมของคำในเวกเตอร์ ด้วยวิธีการนี้จึงทำให้แต่ละข้อความในคู่หน่วยเทียบถูกแปลงเป็นเวกเตอร์ของน้ำหนักเอ็นแกรมของคำ จากนั้นจึงนำเวกเตอร์ที่ได้มาวัดค่าความละม้ายโคไซน์ตามสมการที่ 5.2 แล้วนำค่าความละม้ายที่วัดได้มาใช้เป็นลักษณ์ กระบวนการสร้างลักษณ์ดังกล่าวนี้จะดำเนินไปจนกระทั่งได้ค่าความละม้ายจากคู่หน่วยเทียบในคลังข้อมูลครบทุกคู่

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 5 ลักษณ์ ตามช่วงของเอ็นแกรมตั้งแต่ 1 ถึง 5 แกรม ได้แก่

1) ค่าความละม้ายโคไซน์ของน้ำหนักยูนิแกรมของคำแบบ tf-idf (cosine similarity of tf-idf term weight unigram: $cos_{tf\text{-}idf\text{-}1}$)
2) ค่าความละม้ายโคไซน์ของน้ำหนักไบแกรมของคำแบบ tf-idf (cosine similarity of tf-idf term weight bigram: $cos_{tf\text{-}idf\text{-}2}$)
3) ค่าความละม้ายโคไซน์ของน้ำหนักไตรแกรมของคำแบบ tf-idf (cosine similarity of tf-idf term weight trigram: $cos_{tf\text{-}idf\text{-}3}$)
4) ค่าความละม้ายโคไซน์ของน้ำหนัก 4 แกรมของคำแบบ tf-idf (cosine similarity of tf-idf term weight 4-gram: $cos_{tf\text{-}idf\text{-}4}$)
5) ค่าความละม้ายโคไซน์ของน้ำหนัก 5 แกรมของคำแบบ tf-idf (cosine similarity of tf-idf term weight 5-gram: $cos_{tf\text{-}idf\text{-}5}$)

ลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักเอ็นแกรมของคำแบบ tf-idf ทั้ง 5 ลักษณ์ข้างต้นได้ประยุกต์ใช้ขนาดของเอ็นแกรมที่ต่างกันเพื่อประโยชน์ในการทดสอบประสิทธิภาพการจำแนกประเภท

ในแต่ละขนาดของเอ็นแกรม และแก้ไขปัญหาถุงใส่คำได้เช่นเดียวกับลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของคำ ลักษณ์ค่าความละม้ายแจ็กการ์ดของเอ็นแกรมของคำ และลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของเอ็นแกรมของคำ

5.2.2 ลักษณ์ทางวากยสัมพันธ์

ลักษณ์ทางวากยสัมพันธ์คือลักษณ์ที่วิเคราะห์หาและสร้างขึ้นโดยอาศัยความรู้ทางภาษาศาสตร์ในระดับวลี อนุพากย์ และประโยค ในแง่การตรวจหาการลักลอก ลักษณ์ทางวากยสัมพันธ์จะเอื้อให้เครื่องได้เรียนรู้ถึงหน้าที่ ตำแหน่งในการปรากฏ และความสัมพันธ์ระหว่างหน่วยย่อยในวลี อนุพากย์ และประโยค และด้วยเหตุนี้ กระบวนการวิเคราะห์หาและสร้างลักษณ์ประเภทจึงต้องอาศัยเครื่องมือช่วยในการกำกับหมวดคำและแทนรูปความสัมพันธ์ทางวากยสัมพันธ์ในขั้นก่อนการประมวลผล

ในหัวข้อย่อยนี้ ผู้วิจัยจะนำเสนอรายละเอียดของลักษณ์ทางวากยสัมพันธ์ทั้งหมด 23 ลักษณ์ ซึ่งเป็นผลจากการวิเคราะห์หาตามวิธีการวิจัยที่ได้กำหนดไว้ โดยแบ่งการนำเสนอออกเป็น 11 กลุ่มย่อย ดังรายละเอียดต่อไปนี้

5.2.2.1 ค่าระยะการแก้ไขเลเวนชเตย์นของหมวดคำ (Levenshtein edit distance of POS: LD_{POS})

ลักษณ์ค่าระยะการแก้ไขเลเวนชเตย์นของหมวดคำเป็นลักษณ์ทางวากยสัมพันธ์ที่ประยุกต์ใช้แนวคิด 2 แนวคิดมาวิเคราะห์หาและสร้างเป็นลักษณ์ ได้แก่ แนวคิดเรื่องระยะการแก้ไขเลเวนชเตย์นที่ได้กล่าวในรายละเอียดไปแล้วในหัวข้อที่ 5.1.3 และแนวคิดเรื่องหมวดคำ (part of speech: POS)

หมวดคำหรือชนิดของคำถือเป็นประเภททางไวยากรณ์ที่มีบทบาทสำคัญในงานประมวลผลภาษาธรรมชาติแขนงต่างๆ เนื่องจากหมวดคำได้ให้ข้อมูลอันหลากหลายเกี่ยวกับคำรวมทั้งคำที่ปรากฏใกล้เคียงกัน (Jurafsky & Martin, 2009, p. 123) ยกตัวอย่างเช่น หน้าที่ของคำ คุณสมบัติของคำ ความสัมพันธ์ระหว่างคำภายในประโยค เป็นต้น ด้วยเหตุนี้ ข้อมูลที่ได้จากหมวดคำนั้นสามารถนำไปใช้ประโยชน์ได้หลากหลาย ไม่ว่าจะเป็นการแจงส่วนประโยค การแก้ไขความกำกวมทางความหมายของคำ การรู้จำชื่อเฉพาะ หรือการศึกษาวิจัยภาษาอิงคลังข้อมูล

ด้วยคุณสมบัติของหมวดคำดังได้กล่าวไปข้างต้น ในงานวิจัยชิ้นนี้จึงได้เลือกนำหมวดคำมาแทนรูปเป็นลักษณ์สำหรับจำแนกประเภทของข้อความที่มีการลักลอกและไม่มีการลักลอก ทั้งนี้ ในแง่การตรวจหาการลักลอกนั้น ข้อความที่ถูกแทนรูปให้อยู่ในรูปหมวดคำย่อมมีสภาพที่เป็นนามธรรมมากขึ้น จึงเอื้อต่อการตรวจหาการลักลอกที่มีการแก้ไขโดยแทนที่หรือถอดความคำเดิมในข้อความต้นฉบับ

ด้วยคำในหมวดคำเดียวกันหรือคำที่ทำหน้าที่เดียวกัน เพื่อความเข้าใจที่ชัดเจนยิ่งขึ้น ผู้วิจัยขอให้พิจารณาตัวอย่างข้อความต่อไปนี้

S_1 = “แมว|ขโมย|ปลา|ของ|น้อง|ไป|กิน”
S_2 = “วิฬาร์|ขโมย|มัจฉา|ของ|น้อง|ไป|กิน”

จากตัวอย่างนี้ข้างต้น จะเห็นได้ว่าข้อความ S_1 และ S_2 มีความหมายเหมือนกัน แต่ข้อความ S_2 เลือกใช้คำไวพจน์ “วิฬาร์” และ “มัจฉา” แทนที่คำว่า “แมว” และ “ปลา” ในข้อความ S_1 ตามลำดับ ในกรณีนี้ หากเครื่องตรวจหาการลักลอกมีการประยุกต์ใช้เทคนิคการตรวจเพียงในระดับอักขระหรือระดับคำก็อาจให้ผลการตรวจหาว่าข้อความทั้งสองไม่เป็นคู่ลักลอกของกันและกัน เนื่องด้วยเครื่องตีความได้เพียงในรูปอักขระหรือรูปคำ ไม่สามารถตีความได้ว่า วิฬาร์” และ “มัจฉา” และ “แมว” และ “ปลา” เป็นคำไวพจน์ของกันและกัน ตามลำดับ

อย่างไรก็ดี การประยุกต์ใช้แนวคิดเรื่องหมวดคำสามารถช่วยแก้ไขปัญหาข้างต้นได้ส่วนหนึ่ง ในที่นี้จะแสดงให้เห็นโดยการแทนรูปข้อความ S_1 และ S_2 ข้างต้นให้อยู่ในรูปลำดับของหมวดคำ ดังนี้

$S_{1\text{-POS}}$ = “NOUN|VERB|NOUN|ADP|NOUN|VERB|VERB”
$S_{2\text{-POS}}$ = “NOUN|VERB|NOUN|ADP|NOUN|VERB|VERB”

จากตัวอย่าง $S_{1\text{-POS}}$ และ $S_{2\text{-POS}}$ เป็นลำดับของหมวดคำที่แทนรูปได้จากข้อความ S_1 และ S_2 ตามลำดับ จะเห็นได้ว่าเมื่อแทนรูปเป็นหมวดคำแล้ว ลำดับของหมวดคำของข้อความ S_1 และ S_2 เหมือนกันทุกประการ ในแง่นี้ จึงอาจกล่าวได้ว่าเครื่องที่ประยุกต์ใช้แนวคิดเรื่องหมวดคำในการตรวจหาการลักลอกจะสามารถตรวจพบได้ว่าข้อความ S_1 และ S_2 เป็นคู่ลักลอกของกันและกัน

สำหรับค่าระยะการแก้ไขเลเวนชเตย์นแล้ว การประยุกต์ใช้แนวคิดเรื่องหมวดคำจะยังผลให้ตรวจหาความแตกต่างระหว่างข้อความต้นฉบับกับข้อความที่ถูกลักลอกซึ่งเป็นผลจากการแก้ไขของผู้ลักลอกได้โดยพิจารณาจากลำดับของหมวดคำที่ถูกแทนรูปอยู่ในข้อความต้นฉบับและข้อความลักลอก

ในส่วนลักษณ์ค่าระยะการแก้ไขเลเวนชเตย์นของหมวดคำนี้ ผู้วิจัยได้ประยุกต์ใช้โมดูล tltk เวอร์ชั่น 0.3.5 ซึ่งเป็นแพ็กเกจภาษาไพทอนสำหรับประมวลผลภาษาไทยในการกำกับหมวดคำ ชุดกำกับหมวดคำ (tag set) ของแพ็กเกจ tltk นี้อิงตามชุดกำกับหมวดคำสากล (Petrov, Das, & McDonald, 2012) ของโครงการ Universal Dependencies (UD) ซึ่งพัฒนาการกำกับคลังต้นไม้ข้ามภาษาสำหรับหลายภาษา (De Marneffe et al., 2014) ชุดกำกับหมวดคำดังกล่าวประกอบด้วยหมวดคำและป้ายกำกับหมวดคำทั้งหมด 17 หมวด ดังตารางที่ 5.1

ตารางที่ 5.1 ชุดกำกับหมวดคำสากล

ลำดับที่	ป้ายกำกับ	ชื่อหมวดคำภาษาอังกฤษ	ชื่อหมวดคำภาษาไทย
1	ADJ	adjective	คำคุณศัพท์
2	ADP	adposition	คำสังโยคบท
3	ADV	adverb	คำวิเศษณ์
4	AUX	auxiliary	คำช่วยกริยา
5	CCONJ	coordinating conjunction	คำเชื่อมอนุพากย์ความรวม
6	DET	determiner	คำบอกกำหนด
7	INTJ	interjection	คำอุทาน
8	NOUN	noun	คำนาม
9	NUM	numeral	คำบอกจำนวน
10	PART	particle	คำอนุภาค
11	PRON	pronoun	คำสรรพนาม
12	PROPN	proper noun	คำวิสามานยนาม
13	PUNCT	punctuation	เครื่องหมายวรรคตอน
14	SCONJ	subordinating conjunction	คำเชื่อมอนุพากย์ความซ้อน
15	SYM	symbol	สัญลักษณ์
16	VERB	verb	คำกริยา
17	X	other	อื่นๆ

เมื่อกำกับหมวดให้แก่ข้อความในคู่หน่วยเทียบเรียบร้อยแล้ว ผู้วิจัยจะดึงเอาลำดับข้อหมวดคำที่ได้จากการกำกับมาคำนวณค่าระยะการแก้ไขเลเวนชเตย์นด้วยวิธีการเช่นเดียวกับลักษณ์ค่าระยะการแก้ไขเลเวนชเตย์นของอักขระ และลักษณ์ค่าระยะการแก้ไขเลเวนชเตย์นของคำ ที่ได้กล่าวไปแล้วในหัวข้อที่ 5.1.3 และ 5.2.1.3 ตามลำดับ กระบวนการคำนวณค่าระยะการแก้ไขเลเวนชเตย์นจากลำดับของหมวดคำจะดำเนินไปเช่นนี้จนกระทั่งได้ค่าความละม้ายจากคู่หน่วยเทียบในคลังข้อมูลครบทุกคู่

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.2.2 ความยาวของลำดับย่อยร่วมที่ยาวที่สุดของหมวดคำ (length of longest common subsequence of POS: len(lcs_{POS}))

ลักษณ์ความยาวของลำดับย่อยร่วมที่ยาวที่สุดของหมวดคำเป็นลักษณ์ที่วิเคราะห์หาและสร้างโดยประยุกต์ใช้แนวคิดเรื่องลำดับย่อยร่วมที่ยาวที่สุด (longest common subsequence: *lcs*) เช่นเดียวกันกับลักษณ์ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของสายอักขระ และความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ ที่ได้กล่าวไปแล้วในหัวข้อที่ 5.1.4 และ 5.2.1.4 ตามลำดับ แต่ลักษณ์ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของหมวดคำนี้มุ่งพิจารณาการลักลอกที่ปรากฏให้เห็นในระดับวากยสัมพันธ์โดยอาศัยความแตกต่างของลำดับของหมวดคำระหว่างข้อความต้นฉบับและข้อความลักลอกแทน

ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้แทนรูปคำให้เป็นหมวดคำโดยใช้แพ็กเกจ tltk ก่อนในลำดับแรก จากนั้นจึงหาลำดับย่อยร่วมที่ยาวที่สุดของหมวดคำภายในดังกล่าว แล้ววัดความยาวของลำดับย่อยร่วมนั้นโดยนับจำนวนป้ายกำกับหมวดคำที่ปรากฏในลำดับย่อยร่วม ดังตัวอย่างต่อไปนี้

S_1 = “เขา|มี|สุนัข|ตัว|ใหญ่”
S_2 = “เขา|เลี้ยง|สุนัข|ใหญ่|ไว้|ใน|บ้าน”
$S_{1\text{-POS}}$ = “PRON|VERB|NOUN|NOUN|ADJ”
$S_{2\text{-POS}}$ = “PRON|VERB|NOUN|ADJ|ADV|ADP|NOUN”
$lcs_{POS}(S_{1\text{-POS}},S_{2\text{-POS}})$ = “PRON|VERB|NOUN|ADJ”
$\text{len}(lcs_{POS}(S_{1\text{-POS}},S_{2\text{-POS}}))$ = 4

จากตัวอย่างข้างต้นกำหนดให้มีข้อความ S_1 และ S_2 คือ “เขามีสุนัขตัวใหญ่” และ “เขาเลี้ยง|สุนัขใหญ่ไว้ในบ้าน” ตามลำดับ จากนั้นจึงแทนรูปคำในข้อความ S_1 และ S_2 เป็นหมวดคำ ได้เป็นชุดของหมวดคำ $S_{1\text{-POS}}$ และ $S_{2\text{-POS}}$ ตามลำดับ ขั้นต่อมาจึงหาลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของหมวดคำในชุดของหมวดคำทั้งสอง ($lcs_{POS}(S_{1\text{-POS}},S_{2\text{-POS}})$) ซึ่งประกอบไปด้วยป้ายกำกับหมวดคำจำนวน 4 ป้าย ดังนั้น ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของหมวดคำของข้อความ S_1 และ S_2 ($\text{len}(lcs_{POS}(S_{1\text{-POS}},S_{2\text{-POS}}))$) จึงเท่ากับ 4 ทั้งนี้ หากเปรียบเทียบกับความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำที่ได้กล่าวไปแล้วในหัวข้อที่ 5.2.1.4 จะเห็นได้ว่า เมื่อใช้ข้อความ S_1 และ S_2 เหมือนกัน ค่าความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ ($\text{len}(lcs_W)$) กับค่าความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของหมวดคำ ($\text{len}(lcs_{POS})$) จะมีค่าแตกต่างกัน ทั้งนี้เป็นผลจากการแทนรูปคำกริยา “มี” ในข้อความ S_1 และ “เลี้ยง” ในข้อความ S_2 ออกมาเป็นป้ายกำกับคำกริยา

“VERB” เหมือนกัน จึงกล่าวได้ว่าลักษณ์ชนิดนี้สามารถสะท้อนความสัมพันธ์ทางวากยสัมพันธ์ออกมา ทำให้สามารถตรวจจับการแทนที่คำด้วยคำในหมวดคำเดียวกันได้

กระบวนการหาลำดับย่อยร่วมที่ยาวที่สุดของหมวดคำและหาความยาวของลำดับย่อยร่วมจะดำเนินไปจนกระทั่งได้ข้อมูลครบจากทุกคู่หน่วยเทียบในคลังข้อมูล จากนั้นจึงนำค่าความยาวที่วัดได้มาใช้เป็นลักษณ์

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.2.3 ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของหมวดคำ (normalized longest common subsequence of POS: $lcs_{\text{norm-POS}}$)

ลักษณ์ค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของหมวดคำอาศัยแนวคิดเรื่องการนำลำดับย่อยร่วมยาวสุดที่ยาวที่สุดมาคำนวณเป็นค่าความละม้ายเช่นเดียวกันกับลักษณ์ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ และลักษณ์ค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ ที่ได้กล่าวไปแล้วในหัวข้อที่ 5.1.5 และ 5.2.1.5 ตามลำดับ แต่ปรับเปลี่ยนระดับของหน่วยทางภาษาที่ใช้พิจารณาเป็นหมวดคำ

ทั้งนี้ เนื่องด้วยค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของหมวดคำมีฐานะเป็นค่าความละม้ายจึงสามารถแสดงความแตกต่างระหว่างข้อความลักลอกกับข้อความต้นฉบับที่เกิดจากการแก้ไขออกมาในเชิงปริมาณได้ ในแง่ของการประยุกต์หมวดคำมาใช้เป็นลักษณ์ด้วยแล้ว ลักษณ์ชนิดจึงสามารถสะท้อนการแก้ไขข้อความต้นฉบับด้วยการแทรก ลบ หรือแทนที่ ด้วยคำอื่นที่ทำหน้าที่เดียวกันได้ ด้วยเหตุนี้ ผู้วิจัยจึงนำค่าดังกล่าวมาใช้เป็นลักษณ์ในการทดสอบประสิทธิภาพการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก

ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้ใช้โปรแกรมสำหรับหาลำดับย่อยร่วมที่ยาวที่สุดของคำภายในคู่หน่วยเทียบที่มีอยู่เดิมในการคำนวณค่าความละม้ายของลำดับของหมวดคำดังได้กล่าวในรายละเอียดไปแล้วในหัวข้อที่ 5.1.5 และ 5.2.1.5 จากนั้นจึงนำค่าที่ได้ไปใช้จากคู่หน่วยเทียบทุกคู่ในคลังข้อมูลมาใช้เป็นลักษณ์

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.2.4 ค่าความละม้ายโคไซน์ของเอ็นแกรมของหมวดคำ (Cosine similarity of POS n-gram: cos_{POS})

ลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของหมวดคำอาศัยแนวคิดในการวิเคราะห์หาและสร้างเช่นเดียวกันกับลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของอักขระ และลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของคำ ที่กล่าวไปแล้วในหัวข้อที่ 5.1.6 และ 5.2.1.6 ตามลำดับ ทั้งนี้นอกจากแนวคิดเรื่องความละม้ายโคไซน์และแนวคิดเรื่องเอ็นแกรมแล้ว ลักษณ์ชนิดนี้ยังประยุกต์แนวคิดเรื่องหมวดคำเข้าไว้ด้วย โดยเป็นการวัดค่าความละม้ายจากเอ็นแกรมของหมวดคำ

ในการสร้างลักษณ์ชนิดนี้ ในขั้นแรก ผู้วิจัยจะแทนรูปคำให้เป็นหมวดคำโดยใช้แพ็กเกจ tltk ก่อน ขั้นตอนหลังจากนั้นจะเป็นเช่นเดียวกันกับการสร้างลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของอักขระและลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของคำ กล่าวคือ ผู้วิจัยจะแปลงเอ็นแกรมของหมวดคำของที่แทนรูปจากข้อความในคู่หน่วยเทียบแต่ละข้อความให้อยู่ในรูปของเวกเตอร์จำนวนนับ (count vector) ของเอ็นแกรมของหมวดคำ จากนั้นจึงนำเวกเตอร์ที่แปลงได้ทั้งสองเวกเตอร์มาคำนวณค่าความละม้ายโคไซน์ตามสมการที่ 5.2 ที่ได้แสดงในหัวข้อที่ 5.1.6 แล้วนำค่าที่ได้นี้มาใช้เป็นลักษณ์ กระบวนการสร้างลักษณ์ดังกล่าวมานี้จะดำเนินไปจนกระทั่งได้ค่าความละม้ายจากทุกคู่หน่วยเทียบในคลังข้อมูลครบ

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 5 ลักษณ์ ตามช่วงของเอ็นแกรมตั้งแต่ 1 ถึง 5 แกรม ได้แก่

1) ค่าความละม้ายโคไซน์ของยูนิแกรมของหมวดคำ (cosine similarity of POS unigram: cos_{POS1})
2) ค่าความละม้ายโคไซน์ของไบแกรมของหมวดคำ (cosine similarity of POS bigram: cos_{POS2})
3) ค่าความละม้ายโคไซน์ของไตรแกรมของหมวดคำ (cosine similarity of POS trigram: cos_{POS3})
4) ค่าความละม้ายโคไซน์ของ 4 แกรมของหมวดคำ (cosine similarity of POS 4-gram: cos_{POS4}
5) ค่าความละม้ายโคไซน์ของ 5 แกรมของหมวดคำ (cosine similarity of POS 5-gram: cos_{POS5})

จะเห็นได้ว่าลักษณ์ในกลุ่มดังกล่าวข้างต้นได้ประยุกต์ใช้ขนาดของเอ็นแกรมของหมวดคำที่ต่างกัน ในแง่นี้ ลักษณ์แต่ละจะให้ประสิทธิภาพในการตรวจหาการลักลอกได้แตกต่างกันตามขนาด

ของข้อความ นอกจากนี้ การประยุกต์ใช้ขนาดของเอ็นแกรมที่ต่างกันยังสามารถแก้ไขปัญหาถุงใส่คำได้ด้วย

5.2.2.5 ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของเอ็นแกรมของหมวดคำ (Jaccard similarity coefficient of POS n-gram: J_{POS})

ลักษณ์ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของเอ็นแกรมของหมวดคำเป็นลักษณ์ที่อาศัยแนวคิดเรื่องค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดและแนวคิดเรื่องเอ็นแกรมมาวิเคราะห์หาและสร้างเป็นลักษณ์เช่นเดียวกันกับค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของเอ็นแกรมของอักขระและลักษณ์ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของเอ็นแกรมของคำ ที่ได้กล่าวไปแล้วในหัวข้อที่ 5.1.7 และ 5.2.1.7 ตามลำดับ แต่ลักษณ์ชนิดได้ประยุกต์ใช้แนวคิดเรื่องหมวดคำเป็นหน่วยที่ใช้ในการพิจารณาแทน ซึ่งจะยังผลให้สามารถตรวจหาการแก้ไขข้อความต้นฉบับด้วยการแทรก ลบ หรือแทนที่ ด้วยคำอื่นที่ทำหน้าที่เดียวกันได้

ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้แทนรูปคำให้เป็นหมวดคำโดยใช้แพ็กเกจ tltk ก่อน จากนั้นจึงวัดค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดจากเอ็นแกรมของหมวดคำระหว่างข้อความภายในคู่หน่วยเทียบ แล้วจึงนำค่าที่คำนวณได้มาใช้เป็นลักษณ์ โดยอาศัยการคำนวณตามสมการที่ 5.3 ที่ได้แสดงไปแล้วในหัวข้อที่ 5.1.7 แต่ปรับเปลี่ยนให้ใช้ชุดของเอ็นแกรมของหมวดคำในข้อความแทนชุดของเอ็นแกรมของอักขระ กระบวนการสร้างลักษณ์จะดำเนินไปเช่นนี้จนกระทั่งได้ค่าความละม้ายจากคู่หน่วยเทียบครบทุกคู่

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 5 ลักษณ์ ตามช่วงของเอ็นแกรมตั้งแต่ 1 ถึง 5 แกรม ได้แก่

1) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของหมวดคำ (Jaccard similarity coefficient of POS unigram: J_{POS1})
2) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของหมวดคำ (Jaccard similarity coefficient of POS bigram: J_{POS2})
3) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของหมวดคำ (Jaccard similarity coefficient of POS trigram: J_{POS3})
4) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของหมวดคำ (Jaccard similarity coefficient of POS 4-gram: J_{POS4})
5) ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของหมวดคำ (Jaccard similarity coefficient of POS 5-gram: J_{POS5})

ลักษณ์ดังกล่าวข้างต้นนี้ได้ประยุกต์ใช้ขนาดของเอ็นแกรมของหมวดคำที่ต่างกัน ด้วยเหตุนี้ ลักษณ์แต่ละตัวจึงให้ประสิทธิภาพในจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกที่แตกต่างกัน และสามารถแก้ไขปัญหาถุงใส่คำได้

5.2.2.6 ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของเอ็นแกรมของหมวดคำ (Sørensen–Dice coefficient of POS n-gram: QS_{POS})

ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของเอ็นแกรมของหมวดคำเป็นลักษณ์ที่นำแนวคิดเรื่องค่าความละม้ายมาประยุกต์เข้าและแนวคิดเรื่องเอ็นแกรมมาใช้ในการวิเคราะห์หาและสร้าง เช่นเดียวกันกับลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของเอ็นแกรมของอักขระและค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของเอ็นแกรมของคำ ที่ได้กล่าวไปแล้วในหัวข้อที่ 5.1.8 และ 5.2.1.8 แต่ได้ปรับประยุกต์ให้ใช้หมวดคำเป็นหน่วยที่ใช้ในการพิจารณาการลักลอกแทน ด้วยเชื่อว่าลักษณ์ชนิดนี้จะสามารถตรวจหาการแก้ไขข้อความต้นฉบับด้วยการแทรก ลบ หรือแทนที่ ด้วยคำอื่นที่ทำหน้าที่เดียวกันได้

การสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้แทนรูปคำให้เป็นหมวดคำโดยใช้แพ็กเกจ tltk ก่อน จากนั้นจึงวัดค่าความละม้ายระหว่างคู่หน่วยเทียบในคลังข้อมูลในระดับเอ็นแกรมของหมวดคำจาก แล้วนำค่าที่คำนวณได้มาใช้เป็นลักษณ์ โดยมีวิธีการคำนวณค่าความละม้ายตามสมการที่ 5.4 ที่ได้แสดงไปแล้วในหัวข้อที่ 6.1.8 แต่ปรับเปลี่ยนให้ใช้ชุดของเอ็นแกรมของหมวดคำในข้อความแทนชุดของเอ็นแกรมของอักขระ กระบวนการสร้างลักษณ์จะดำเนินไปเช่นนี้จนกระทั่งได้ค่าความละม้ายจากคู่หน่วยเทียบในคลังข้อมูลครบทุกคู่

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 5 ลักษณ์ ตามช่วงของเอ็นแกรมตั้งแต่ 1 ถึง 5 แกรม ได้แก่

1) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของหมวดคำ (Sørensen–Dice similarity coefficient of POS unigram: QS_{POS1})
2) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของหมวดคำ (Sørensen–Dice similarity coefficient of POS bigram: QS_{POS2})
3) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของหมวดคำ (Sørensen–Dice similarity coefficient of POS trigram: QS_{POS3})
4) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของหมวดคำ (Sørensen–Dice similarity coefficient of POS 4-gram: QS_{POS4})

5) ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของหมวดคำ (Sørensen–Dice similarity coefficient of POS 5-gram: QS_{POS5})

ลักษณ์ข้างต้นนี้ได้ประยุกต์ใช้ขนาดของเอ็นแกรมของหมวดคำที่ต่างกันเช่นเดียวกับลักษณ์ค่าความละม้ายโคไซน์ของเอ็นแกรมของหมวดคำและค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของเอ็นแกรมของหมวดคำ ด้วยเหตุนี้ ลักษณ์แต่ละตัวจึงให้ประสิทธิภาพในจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกที่แตกต่างกันตามขนาดของข้อความ อีกทั้งยังสามารถแก้ไขปัญหาถุงใส่คำได้เช่นกัน

5.2.2.7 ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของคำ (cosine similarity of word n-gram range: $\cos_{W123}$)

ลักษณ์ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของคำเป็นลักษณ์ที่วิเคราะห์และสร้างขึ้นโดยมีแนวคิดเบื้องหลังที่ต้องการตรวจหาการลักลอกในกรณีที่ผู้ลักลอกใช้วิธีเปลี่ยนลำดับของคำเพื่อแก้ไขข้อความต้นฉบับ โดยไม่ต้องอาศัยการตรวจหาด้วยลักษณ์หลายชนิดที่ประยุกต์ใช้ขนาดของเอ็นแกรมที่ต่างกัน

โดยอาศัยแนวคิดดังกล่าวข้างต้น ผู้วิจัยจึงได้สร้างลักษณ์ที่รวมชุดของเอ็นแกรมหลายขนาดเข้าด้วยกันภายในลักษณ์ตัวเดียวโดยเรียกว่า “ช่วงของเอ็นแกรม” ทั้งนี้ เพื่อความเข้าใจที่ชัดเจนยิ่งขึ้น ขอให้พิจารณาตัวอย่างข้อความ S_1 และ S_2 ต่อไปนี้

S_1 = “ผู้ใหญ่|หลาย|ต่อ|หลาย|คน|เชื่อ|ว่า|เด็ก|ควร|ถูก|ตี|เมื่อ|ทำ|ผิด”

S_2 = “ผู้ใหญ่|หลาย|ต่อ|หลาย|คน|เชื่อ|ว่า|เมื่อ|ทำ|ผิด|เด็ก|ควร|ถูก|ตี”

จากตัวอย่างข้อความ S_1 และ S_2 สมมติให้ข้อความ S_2 เป็นข้อความที่ได้จากการลักลอกข้อความ S_1 โดยเปลี่ยนลำดับของคำ ในกรณีนี้จะเห็นได้ว่าใจความที่ข้อความทั้งสองสื่อออกมานั้นเหมือนกัน อย่างไรก็ตาม หากเครื่องประยุกต์ใช้การตรวจหาในระดับคำโดยอิงจากยูนิแกรมของคำในข้อความก็จะพบปัญหาจากแนวคิดถุงใส่คำ ดังได้แสดงให้เห็นเป็นตัวอย่างไปแล้วในหัวข้อที่ 5.2.1.6 กล่าวคือผลการตรวจหาการลักลอกที่ได้จะระบุว่าข้อความเหมือนกันทุกประการ ซึ่งไม่ได้สะท้อนลักษณะที่เป็นจริงของข้อความทั้งสองที่มีความละม้ายกันบางส่วนเท่านั้น ทั้งนี้ หากต้องการให้เครื่องรายงานผลว่าข้อความทั้งสองมีความละม้ายกันบางส่วนตามสภาพที่ปรากฏจริง ก็จำเป็นต้องอาศัยแนวคิดอื่นๆ ในการตรวจหา เช่นการใช้ลักษณ์ที่ประยุกต์ใช้ขนาดของเอ็นแกรมของคำที่แตกต่างกันหลายๆ ลักษณ์

ด้วยสาเหตุดังได้กล่าวไปข้างต้น ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยจึงได้ใช้แนวคิดช่วงของเอ็นแกรมของคำเพื่อให้สามารถสะท้อนลำดับคำที่เปลี่ยนแปลงไปในข้อความต้นฉบับและข้อความลักลอก

โดยผู้วิจัยได้เลือกรวมชุดของยูนิแกรม ไบแกรม และไตรแกรม ของคำไว้ในลักษณ์เดียวแล้วแปลงช่วงของเอ็นแกรมที่ได้เป็นเวกเตอร์จำนวนนับ และนำไปวัดค่าความละม้ายโคไซน์ระหว่างเวกเตอร์ดังกล่าว

ทั้งนี้ จากข้อความ S_1 และ S_2 ข้างต้นจะได้ช่วงของเอ็นแกรมของคำที่ประกอบด้วยยูนิแกรม ไบแกรม และไตรแกรมของคำ ทั้งหมด 44 ชุด ได้แก่ “คน”, “คน|เชื่อ”, “คน|เชื่อ|ว่า”, “ควร”, “ควร|ถูก”, “ควร|ถูก|ตี”, “ตี”, “ตี|เมื่อ”, “ตี|เมื่อ|ทำ”, “ต่อ”, “ต่อ|หลาย”, “ต่อ|หลาย|คน”, “ถูก”, “ถูก|ตี”, “ถูก|ตี|เมื่อ”, “ทำ”, “ทำ|ผิด”, “ทำ|ผิด|เด็ก”, “ผิด”, “ผิด|เด็ก”, “ผิด|เด็ก|ควร”, “ผู้ใหญ่”, “ผู้ใหญ่|หลาย”, “ผู้ใหญ่|หลาย|ต่อ”, “ว่า”, “ว่า|เด็ก”, “ว่า|เด็ก|ควร”, “ว่า|เมื่อ”, “ว่า|เมื่อ|ทำ”, “หลาย”, “หลาย|คน”, “หลาย|คน|เชื่อ”, “หลาย|ต่อ”, “หลาย|ต่อ|หลาย”, “เชื่อ”, “เชื่อ|ว่า”, “เชื่อ|ว่า|เด็ก”, “เชื่อ|ว่า|เมื่อ”, “เด็ก”, “เด็ก|ควร”, “เด็ก|ควร|ถูก”, “เมื่อ”, “เมื่อ|ทำ”, “เมื่อ|ทำ|ผิด” และเมื่อนำข้อความ S_1 และ S_2 ไปแปลงเป็นเวกเตอร์จำนวนนับจากช่วงของเอ็นแกรมดังกล่าวจะได้ดังนี้

$$S_{1\text{-}W123} = [1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 1\ 1\ 1\ 1\ 1\ 1]$$

$$S_{2\text{-}W123} = [1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 1\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 0\ 1\ 1\ 1\ 1\ 1\ 1]$$

เมื่อนำเวกเตอร์ $S_{1\text{-}W123}$ และ $S_{2\text{-}W123}$ ไปวัดค่าความละม้ายโคไซน์ตามสมการที่ 5.2 ผลปรากฏว่าได้ค่าความละม้ายเท่ากับ 0.1463 ผลดังกล่าวนี้ได้สะท้อนให้เห็นว่าแนวคิดนี้สามารถตรวจจับการเปลี่ยนแปลงลำดับของคำในข้อความต้นฉบับที่ถูกแก้ไขให้เปลี่ยนแปลงไปเป็นลำดับคำที่ปรากฏในข้อความลักลอกได้และรายงานผลที่สะท้อนสภาพที่ปรากฏจริงของข้อความทั้งว่ามีความละม้ายกันบางส่วน ไม่ได้เหมือนกันทุกประการ อย่างไรก็ดี ผู้วิจัยสังเกตได้ว่าค่าความละม้ายที่วัดได้จากวิธีการดังกล่าวนี้จะใกล้เคียงกับสภาพจริงของคู่ลักลอกมากขึ้นเมื่อข้อความมีความยาวที่เหมาะสม เช่น มีความยาวในระดับย่อหน้าตามปรากฏในคลังข้อมูลที่สร้างขึ้นเพื่อใช้ในงานวิจัยชิ้นนี้

กระบวนการสร้างลักษณ์จะดำเนินเช่นนี้จนกระทั่งได้ ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของคำจากคู่หน่วยเทียบในคลังข้อมูลครบทุกคู่

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.2.8 ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf (cosine similarity of tf-idf term weight n-gram range: $cos_{tf\text{-}idf\text{-}123}$)

ลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf เป็นลักษณ์ที่ประยุกต์แนวคิดด้านการประมวลผลภาษา 4 แนวคิดเข้าไว้ด้วย ได้แก่ แนวคิดเรื่องค่าความละม้ายโคไซน์ แนวคิดเรื่องการให้น้ำหนักคำ แนวคิดเรื่องเอ็นแกรม และแนวคิดเรื่องความสัมพันธ์ระหว่างคำในประโยคที่สะท้อนผ่านลำดับคำ โดยนำช่วงของเอ็นแกรมตั้งแต่ 1-3 แกรมไปหาค่าน้ำหนักแบบ tf-idf จากนั้นจึงแปลงเป็นเวกเตอร์จำนวนนับแล้วนำไปคำนวณค่าความละม้ายโคไซน์

โดยอาศัยแนวคิดที่ใช้วิเคราะห์และสร้างลักษณ์ข้างต้น ลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf จึงไม่เพียงแต่จะสะท้อนการเปลี่ยนแปลงด้านลำดับในคู่หน่วยเทียบเท่านั้น แต่ยังสะท้อนความสำคัญของคำในคู่หน่วยเทียบผ่านค่าน้ำหนักที่ถูกกำหนดให้ด้วย ในแง่นี้จึงกล่าวได้ว่าลักษณ์ชนิดนี้สามารถตรวจจับได้ทั้งการแก้ไขลำดับคำและการแก้ไขด้วยการแทรก ลบ หรือแทนที่ คำที่มีน้ำหนักเดียวกัน

ในการสร้างลักษณ์ดังกล่าวนี้ ในขั้นแรก ผู้วิจัยได้ใช้โปรแกรมเพื่อหาค่าน้ำหนักของช่วงเอ็นแกรมทั้งหมดในคลังข้อมูลตั้งแต่ 1-3 แกรมเพื่อสร้างเป็นคลังข้อมูลฝึกฝนของการให้น้ำหนัก เมื่อได้คลังข้อมูลฝึกฝนแล้ว ในขั้นต่อมา ผู้วิจัยจะนำคู่หน่วยเทียบแต่ละคู่ไปแปลงเป็นเวกเตอร์ของน้ำหนักช่วงเอ็นแกรมของคำ โดยเทียบชุดของเอ็นแกรมของคำที่ปรากฏในข้อความแต่ละข้อความกับชุดของเอ็นแกรมในคลังข้อมูลฝึกฝน เมื่อเครื่องพบว่าชุดของเอ็นแกรมตรงกันจึงคืนค่าน้ำหนักของชุดเอ็นแกรม จากนั้น ผู้วิจัยจะนำเวกเตอร์ที่แปลงได้จากคู่หน่วยเทียบไปวัดค่าความละม้ายโคไซน์ตามสมการที่ 5.2 แล้วนำค่าความละม้ายที่วัดได้มาใช้เป็นลักษณ์ กระบวนการสร้างลักษณ์ดังกล่าวนี้จะดำเนินไปจนกระทั่งได้ค่าความละม้ายจากคู่หน่วยเทียบในคลังข้อมูลครบทุกคู่

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.2.9 ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของหมวดคำ (cosine similarity of POS n-gram range: cos_{POS123})

ลักษณ์ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของหมวดคำอาศัยแนวคิดในการวิเคราะห์หาและสร้างคล้ายคลึงกับลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf กล่าวคือ เป็นการประยุกต์แนวคิดเรื่องค่าความละม้ายโคไซน์ การให้น้ำหนักคำ และเอ็นแกรม เข้าด้วยกัน แต่สิ่งที่แตกต่างกันคือลักษณ์ชนิดเลือกพิจารณาช่วงเอ็นแกรมของหมวดคำแทนช่วงเอ็นแกรมของน้ำหนักคำ

ในแง่การตรวจหาการลักลอก ลักษณ์ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของหมวดคำจะสามารถตรวจจับการแก้ไขลำดับคำและการแก้ไขด้วยการแทรก ลบ หรือแทนที่คำ ได้เช่นเดียวกับลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf แต่คำที่ถูกแก้ไขนั้นจะถูกพิจารณาในเชิงหน้าที่และความสัมพันธ์กับคำอื่นภายในข้อความ เนื่องจากคำในข้อความได้รับการแทนรูปให้อยู่รูปหมวดคำ ในแง่นี้ ลักษณ์ชนิดจึงเหมาะแก่การตรวจหาการลักลอกที่ผู้ลักลอกแก้ไยข้อความโดยการแทรก ลบ หรือแทนที่คำตามการทำหน้าที่ในประโยค พร้อมกับเปลี่ยนตำแหน่งของคำนั้นๆ ด้วย

ในการสร้างลักษณ์ชนิดนี้ ในขั้นแรก ผู้วิจัยจะแทนรูปคำให้เป็นหมวดคำโดยใช้แพ็กเกจ tltk ก่อน จากนั้นจึงจัดชุดของหมวดคำในข้อความตามช่วงของเอ็นแกรมตั้งแต่ 1-3 แกรม แล้วแปลงช่วงของเอ็นแกรมของหมวดคำดังกล่าวให้อยู่ในรูปของเวกเตอร์จำนวนนับ จากนั้นจึงนำเวกเตอร์ที่แปลงได้ทั้งสองเวกเตอร์มาคำนวณค่าความละม้ายโคไซน์ตามสมการที่ 5.2 ที่ได้แสดงในหัวข้อที่ 5.1.6 แล้วนำค่าที่ได้นี้มาใช้เป็นลักษณ์ กระบวนการสร้างลักษณ์ดังกล่าวมานี้จะดำเนินไปจนกระทั่งได้ค่าความละม้ายจากทุกคู่หน่วยเทียบในคลังข้อมูลครบ

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.2.10 ค่าความละม้ายของลำดับคำ (word order similarity: sim_{wo})

การวัดค่าความละม้ายของลำดับคำเป็นวิธีการคำนวณค่าความละม้ายทางวากยสัมพันธ์ที่ถูกเสนอโดยลีและคณะ (Y. Li et al., 2004; Y. Li et al., 2006) เพื่อแก้ปัญหาถุงใส่คำที่ส่งผลให้ค่าความละม้ายที่วัดได้คลาดเคลื่อนไปจากลักษณะของข้อความที่ปรากฏจริงดังที่งานวิจัยชิ้นนี้ได้แสดงให้เห็นแล้วหัวข้อที่ 5.2.1.6 โดยลีและคณะได้รวมคำพร้อมลำดับการปรากฏของข้อความ 2 ข้อความที่ต้องการวัดค่าความละม้ายเข้าด้วยกัน จากนั้นจึงกำหนดลำดับของคำให้ใหม่จากชุดรวมของคำดังกล่าว แล้วแปลงข้อความทั้งสองข้อความเป็นเวกเตอร์ของลำดับคำ และวัดค่าความละม้ายระหว่างเวกเตอร์ดังกล่าว

ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้ดัดแปลงแนวคิดที่ลีและคณะเสนอไว้สำหรับการสร้างเวกเตอร์ของลำดับคำ เพื่อให้รองรับกรณี 2 กรณีที่งานของลีและคณะไม่ได้ระบุ ได้แก่

1) มีการปรากฏซ้ำของคำในข้อความเดียวกัน
2) จำนวนคำของทั้งสองข้อความไม่เท่ากัน

ทั้งนี้ เพื่อความเข้าใจที่ชัดเจนยิ่งขึ้น ขอให้พิจารณาตัวอย่างการสร้างเวกเตอร์ของลำดับคำต่อไปนี้

T_1 = “ช่วง|นี้|เป็น|ช่วง|ของ|การ|ปรับ|เปลี่ยน|โครงสร้าง|และ|รูปแบบ|การ|ทำงาน|ของ|หน่วยงาน”

T_2 = “ช่วง|นี้|เป็น|ช่วง|ที่|มี|การ|ปรับ|โครงสร้าง|และ|เปลี่ยน|รูปแบบ|การ|ทำงาน|ของ|หน่วยงาน|ฝ่าย|ต่าง|ๆ”

T_j = [1: “ช่วง”, 2: “นี้”, 3: “เป็น”, 4: “ของ”, 5: “การ”, 6: “ปรับ”, 7: “เปลี่ยน”, 8: “โครงสร้าง”, 9: “และ”, 10: “รูปแบบ”, 11: “ทำงาน”, 12: “หน่วยงาน”, 13: “ที่”, 14: “มี”, 15: “ฝ่าย”, 16: “ต่าง”, 17: “ๆ”]

จากตัวอย่างข้างต้น จะเห็นได้ว่าข้อความ T_1 และ T_2 ประกอบด้วยคำและลำดับของคำที่เหมือนกันบางส่วน แต่จำนวนของคำในข้อความทั้งสองมีไม่เท่ากัน คือ ข้อความ T_1 มี 14 คำ ในขณะที่ข้อความ T_2 มี 19 คำ การเวกเตอร์ของลำดับคำในขั้นแรกทำได้โดยรวมคำที่ปรากฏในข้อความ T_1 และ T_2 เข้าในชุดรวมคำ T_j ตามลำดับจากด้านซ้ายไปขวา โดยดึงเอานำคำและลำดับคำจากข้อความ T_1 ก่อนแล้วตามด้วยคำและลำดับคำจากข้อความ T_2 หากปรากฏว่ามีคำซ้ำกันจะไม่เก็บคำนั้นเข้าในชุดรวมคำ T_j ซ้ำอีก จากนั้นจึงกำหนดหมายเลขลำดับให้คำแต่ละในชุดรวมคำ T_j จากนั้นจึงสร้างเวกเตอร์ของคำลำดับขึ้น โดยเทียบรูปคำในข้อความ T_1 และ T_2 กับรูปคำในชุดรวมคำ T_j หากปรากฏตรงกันจะดึงเอาหมายเลขลำดับของรูปคำมาแทนรูปคำในเวกเตอร์ของลำดับคำ

r_1 = [1 2 3 1 4 5 6 7 8 9 10 5 11 4 12]

r_2 = [1 2 3 1 13 14 5 6 8 9 7 10 5 11 4 12 15 16 17]

อย่างไรก็ตาม จะเห็นได้ว่าเวกเตอร์ r_1 และ r_2 ที่แปลงได้จากข้อความ T_1 และ T_2 ตามลำดับนั้นมีขนาดไม่เท่ากัน จึงไม่สามารถดำเนินการบวกหรือลบกันได้ ในขั้นตอนต่อมา ผู้วิจัยจึงปรับขนาดของเวกเตอร์เพื่อแก้ปัญหาเวกเตอร์มีขนาดไม่เท่ากันเนื่องจากมีจำนวนการปรากฏของคำไม่เท่ากัน โดยการแทรกหมายเลขลำดับ 0 เข้าไปในเวกเตอร์ทั้งสองตามจำนวนที่ได้จากการเทียบกับรูปคำที่ปรากฏในชุดรวมคำ T_j แต่ไม่ปรากฏในข้อความ (เป็นคำที่ปรากฏอยู่ในอีกข้อความ)

r_1 = [1 2 3 1 4 5 6 7 8 9 10 5 11 4 12 0 0 0 0 0]

r_2 = [1 2 3 1 13 14 5 6 8 9 7 10 5 11 4 12 15 16 17]

ในขั้นตอนนี้จะเห็นได้ว่าเวกเตอร์ r_1 ได้รับแทรกหมายเลขลำดับ 0 เข้าไป 5 ครั้ง เนื่องจากมีคำที่ปรากฏในชุดรวมคำ T_j แต่ไม่ปรากฏในข้อความ T_1 5 คำ ได้แก่ “ที่”, “มี”, “ฝ่าย”, “ต่าง”, และ “ๆ” จากนั้นจะวัดขนาดเวกเตอร์อีกครั้ง หากเวกเตอร์ทั้งสองมีขนาดเท่ากันจะเสร็จสิ้นกระบวนการสร้างเวกเตอร์ของลำดับ แต่ถ้าหากเวกเตอร์ทั้งสองยังคงมีขนาดไม่เท่ากัน เครื่องจะดำเนินการแทรก

หมายเลขลำดับ 0 เข้าไปในเวกเตอร์ที่มีจำนวนสมาชิกน้อยกว่าจนกระทั่งเวกเตอร์ทั้งสองขนาดเท่ากันในที่สุด

r_1 = [1 2 3 1 4 5 6 7 8 9 10 5 11 4 12 0 0 0 0 0]

r_2 = [1 2 3 1 13 14 5 6 8 9 7 10 5 11 4 12 15 16 17 0]

เมื่อได้เวกเตอร์ที่มีขนาดเท่ากันแล้ว ในขั้นตอนต่อมาจะได้นำเวกเตอร์ลำดับคำทั้งสองมาคำนวณค่าความละม้ายตามวิธีที่ลีและคณะ (Y. Li et al., 2006, p. 1143) เสนอไว้ดังสมการที่ 5.7

$$sim_{WO}(r_1, r_2) = 1 - \frac{||r_1 - r_2||}{||r_1 + r_2||} \quad (5.7)$$

เมื่อใช้วิธีการคำนวณข้างต้นแล้ว ตัวอย่างข้อความ T_1 และ T_2 จะมีค่าความละม้ายของลำดับคำเท่ากับ 0.4197 จะเห็นได้ว่าวิธีการวัดค่าความละม้ายดังกล่าวสามารถสะท้อนค่าความละม้ายของข้อความ 2 ข้อความที่มีรูปคำและลำดับของคำแตกต่างกันได้อย่างมีประสิทธิภาพ

ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยจะดำเนินการวัดค่าความละม้ายของลำดับตามวิธีการที่ได้กล่าวมาข้างต้นและนำค่าที่ได้มาใช้เป็นลักษณ์ โดยกระบวนการดังกล่าวมานี้จะดำเนินไปจนกระทั่งได้ค่าความละม้ายจากทุกคู่หน่วยเทียบในคลังข้อมูลครบ

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.2.11 ค่าความละม้ายโคไซน์ของลำดับคำ (cosine similarity of word order: cos_{WO})

ลักษณ์ค่าความละม้ายโคไซน์ของลำดับคำเป็นลักษณ์ที่ประยุกต์ใช้แนวคิดเรื่องการเรียงลำดับของรูปคำในข้อความมาหาค่าความละม้ายเช่นเดียวกันกับลักษณ์ค่าความละม้ายของลำดับคำ แต่ในกรณีของลักษณ์ชนิดนี้ ผู้วิจัยได้ปรับเปลี่ยนวิธีการวัดค่าความละม้ายระหว่างเวกเตอร์ของลำดับคำไปใช้ค่าความละม้ายโคไซน์แทน ทั้งนี้ เป็นที่น่าสนใจว่าวิธีการวัดค่าความละม้ายแบบใดจะให้ประสิทธิภาพในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกดีกว่ากัน

ในการสร้างลักษณ์ชนิดนี้ ขั้นแรก ผู้วิจัยได้สร้างเวกเตอร์ของลำดับคำของข้อความในคู่หน่วยเทียบขึ้นก่อนโดยใช้วิธีตามที่ได้กล่าวไปแล้วในหัวข้อที่ 5.2.2.10 เมื่อได้เวกเตอร์ของลำดับคำแล้ว ขั้นตอนต่อมาคือการวัดค่าความละม้ายโคไซน์ของเวกเตอร์ทั้งสองตามสมการที่ 5.2 ที่ได้แสดงในหัวข้อที่ 5.1.6 แล้วนำค่าความละม้ายที่ได้มาใช้เป็นลักษณ์ กระบวนการสร้างลักษณ์ดังกล่าวมานี้จะดำเนินไปจนกระทั่งได้ค่าความละม้ายจากทุกคู่หน่วยเทียบในคลังข้อมูลครบ

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.3 ลักษณ์ทางความหมาย

ลักษณ์ทางความหมายคือลักษณ์ที่วิเคราะห์หาและสร้างขึ้นโดยอาศัยความรู้ทางภาษาศาสตร์ในระดับความหมาย ในแง่การตรวจหาการลักลอก ลักษณ์ทางความหมายจะเอื้อให้เครื่องได้เรียนรู้ถึงความสัมพันธ์ทางความหมายระหว่างคำในข้อความ ด้วยเหตุนี้ กระบวนการวิเคราะห์หาและสร้างลักษณ์ประเภทจึงต้องอาศัยเครื่องมือช่วยในการวัดหรือคำนวณความใกล้เคียงกันของความหมายระหว่างคำต่างๆ ในข้อความ

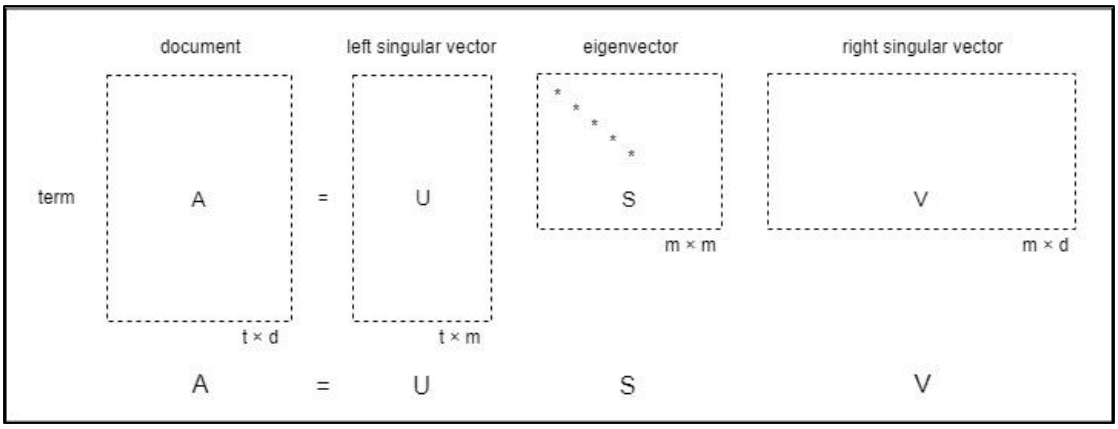
อย่างไรก็ตาม ดังได้กล่าวไปในหัวข้อที่ 3.4.3 ซึ่งว่าด้วยวิธีการวิเคราะห์ลักษณ์ทางความหมายแล้วว่า จากการทบทวนวรรณกรรมพบว่างานวิจัยที่ประยุกต์ใช้ลักษณ์ทางความหมายในการจำแนกประเภทข้อความที่มีความละม้ายก็ดี หรือใช้ในการตรวจหาการลักลอกก็ดี ล้วนแล้วแต่พึ่งพาการคืนค่าทางความหมายจากเครือข่ายคำ (WordNet) ทั้งสิ้น แต่ในกรณีของภาษาไทยนั้น ไม่ปรากฏว่ามีเครือข่ายคำที่เสร็จสมบูรณ์พร้อมใช้งาน ในการวิเคราะห์หาลักษณ์ทางความหมายนี้ ผู้วิจัยจึงจำเป็นต้องอาศัยแนวคิดหรือเทคนิควิธีอื่นที่สามารถใช้เป็นตัวแทนของความหมายได้โดยอ้อม

ในหัวข้อย่อยนี้ ผู้วิจัยจะนำเสนอรายละเอียดของลักษณ์ทางความหมายทั้งหมด 2 ลักษณ์ซึ่งเป็นผลจากการวิเคราะห์หาตามวิธีการวิจัยที่ได้กำหนดไว้ ดังรายละเอียดต่อไปนี้

5.2.3.1 ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง (Cosine similarity of latent semantic vector: cos_{LSA})

ลักษณ์ค่าความละม้ายโคไซน์จากเวกเตอร์ทางความหมายแอบแฝงเป็นลักษณ์ที่ประยุกต์ใช้แนวคิด 2 แนวคิดในการวิเคราะห์หาและสร้างเป็นลักษณ์ ได้แก่ แนวคิดเรื่องความละม้ายโคไซน์ และแนวคิดเรื่องการวิเคราะห์ความหมายแอบแฝง

การวิเคราะห์ความหมายแอบแฝง (Latent Semantic Analysis: LSA) หรือการทำดัชนีความหมายแอบแฝง (Latent Semantic Indexing: LSI) เป็นแนวคิดที่ประยุกต์ความรู้ด้านพีชคณิตเชิงเส้นและวิธีการคำนวณทางสถิติที่มีประสิทธิภาพสูง เพื่อแยกความหมายของคำจากข้อความหรือกลุ่มข้อความที่มีขนาดใหญ่ (Deerwester, Dumais, & Harshman, 1990 อ้างถึงใน ปิยธิดา อินทร์รักษ์, 2552, น. 5) โดยอาศัยแนวคิดที่ว่าคำที่มีความหมายคล้ายกันจะปรากฏอยู่ในเอกสารที่มีลักษณะคล้ายกัน (นัชชา ถิระสาโรช, 2559, น. 24) ในแง่นี้ การวิเคราะห์ความหมายแอบแฝงจึงนำมาใช้หาความสัมพันธ์ทางความหมายระหว่างคำต่างๆ ได้



ภาพที่ 5.4 การแยกค่าเชิงเดี่ยว

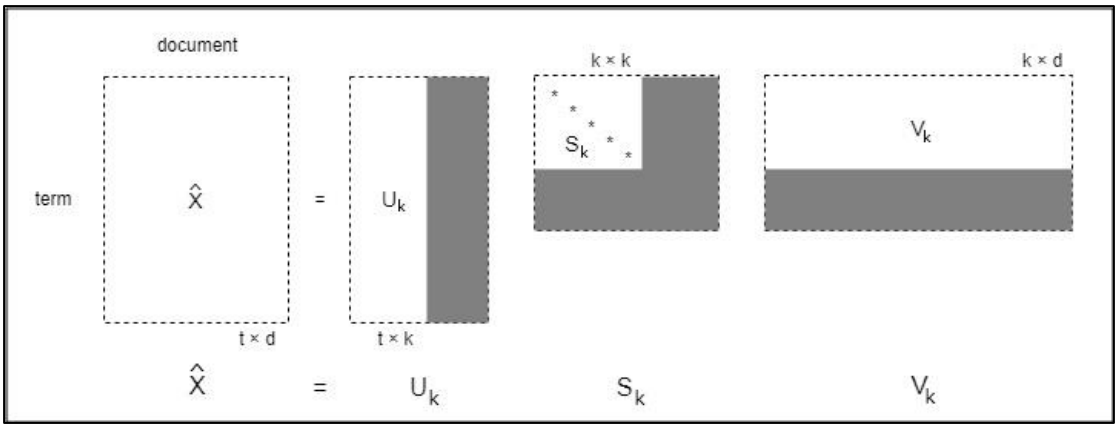
ในการแยกความหมายของคำนั้น ทำได้โดยสร้างเมทริกซ์ *A* ของคำและเอกสาร (term-document matrix) ขนาด $t \times d$ ขึ้น โดยกำหนดให้คอลัมน์ของเมทริกซ์แสดงเอกสารที่นำมาวิเคราะห์จำนวน *d* เอกสาร และแต่ละแถวของเมทริกซ์แสดงคำ แต่ละคำในข้อความไม่ซ้ำ (word type) จำนวน *t* คำ และแต่ละช่อง (cell) แสดงความถี่ของคำ $a_{i,j}$ ที่ปรากฏในแต่ละเอกสาร จากนั้นจะประยุกต์ใช้การแยกค่าแบบเดี่ยว (Singular Value Decomposition: SVD) แยกค่าเมทริกซ์ *A* ตามสมการที่ 5.8 ออกเป็น 3 เวกเตอร์ ได้แก่ เวกเตอร์เดี่ยวซ้าย (left singular vector) เวกเตอร์เดี่ยวขวา (right singular vector) และเวกเตอร์ไอเกน (eigenvectors) ซึ่งเป็นเวกเตอร์แนวทแยง ดังแสดงในภาพที่ 5.4

$$A = USV \tag{5.8}$$

จากนั้นจึงลดขนาดมิติ (dimension reduction) ของเวกเตอร์ทั้งสามเพื่อกำจัดสิ่งรบกวน (noise) ที่ไม่เป็นประโยชน์กับข้อมูล โดยการเลือกขนาดของ *k* ที่เหมาะสมเพื่อให้เหลือเพียงส่วนที่ความหมายกับเอกสาร ขั้นตอนนี้จะได้เมทริกซ์ของความสัมพันธ์ใหม่ที่มีขนาดเหมาะสม (the least square best fit) $\hat{X}$ ตามสมการที่ 5.9 เมทริกซ์นี้จะแสดงระนาบที่พยากรณ์ความถีฺ่ที่เหมาะสมของแต่ละคำที่มีแนวโน้มการเกิดขึ้นในแต่ละเอกสารได้โดยมีขนาดเท่ากับ *k*

$$\hat{X} = \sum_{i=1}^{k} U_i \cdot S_i \cdot V_i \tag{5.9}$$

เมื่อลดขนาดมิติของเมทริกซ์แล้ว ขั้นตอนต่อมา จะนำเมทริกซ์ทั้งสามมาคูณกัน วิธีการนี้จะทำให้เมทริกซ์ใหม่เกิดที่แสดงค่าความสัมพันธ์ระหว่างคำกับชุดข้อความอย่างมีนัย (ปิยธิดา อินทร์รักษ์, 2552, น. 6-9) ดังภาพที่ 5.5



ภาพที่ 5.5 เมทริกซ์ใหม่ที่ได้จากการคูณเมทริกซ์ที่ผ่านการลดขนาดมิติทั้งสาม

จากภาพที่ 5.5 จะเห็นได้ว่าเมทริกซ์ *U*, *S*, และ *V* ปรากฏพื้นที่สีขาวและพื้นที่สีทึบ พื้นที่สีขาวคือพื้นที่ที่คงเหลือจากการลดขนาดมิติ ส่วนพื้นที่สีทึบคือขนาดมิติที่ถูกลดไป เมทริกซ์ $\hat{X}$ เกิดขึ้นจากการคูณพื้นที่สีขาวเข้าด้วยกัน เมทริกซ์ $\hat{X}$ นี้จะถูกนำไปใช้ไปคำนวณค่าความสัมพันธ์กับระหว่างคำคำหนึ่งกับคำอื่นๆ ลักษณะเช่นนี้คือการแสดงความสัมพันธ์ที่แอบแฝง (latent) อยู่ออกมา

ในการสร้างลักษณ์ค่าความละม้ายโคไซน์จากเวกเตอร์ทางความหมายแอบแฝงนี้ ผู้วิจัยได้นำเข้าข้อมูลภายในคลังข้อมูลทั้งหมดเพื่อคำนวณค่าความสัมพันธ์ทางความหมายของคำที่ปรากฏในคลังข้อมูล โดยแปลงคำให้อยู่ในรูปของค่าน้ำหนักคำแบบ tf-idf แล้วสร้างเป็นเมทริกซ์ จากนั้นจึงนำเมทริกซ์ดังกล่าวเข้าสู่กระบวนการแยกค่าเชิงเดี่ยวและลดขนาดมิติตามวิธีการที่ได้กล่าวไปข้างต้น โดยใช้ฟังก์ชัน Truncated SVD ในไลบรารี scikit-learn เวอร์ชัน 0.19.1 ด้วยวิธีการนี้ เมทริกซ์ที่ได้ออกมาจะมีฐานะเป็นเสมือนคลังข้อมูลฝึกฝนสำหรับเทียบค่าความสัมพันธ์ทางความหมายของคำ จากนั้นจึงเป็นการวัดค่าความละม้ายของคู่หน่วยเทียบ ในขั้นนี้ ผู้วิจัยจะแปลงข้อความแต่ละข้อความในคู่หน่วยเทียบให้อยู่ในรูปเวกเตอร์ของน้ำหนักแบบ tf-idf ก่อนตามวิธีการที่ได้กล่าวไปแล้วในหัวข้อที่ 5.2.1.9 เมื่อได้เวกเตอร์ของน้ำหนักแบบ tf-idf แล้วจะนำเวกเตอร์ดังกล่าวไปลดขนาดมิติโดยเทียบกับเมทริกซ์ต้นแบบที่ได้สร้างไว้ใช้เป็นคลังข้อมูลฝึกฝนก่อนหน้านี้ วิธีการนี้จะทำให้ได้เวกเตอร์ที่บรรจุค่าความสัมพันธ์ระหว่างคำออกมา จากนั้นจึงนำเวกเตอร์ดังกล่าวไปวัดค่าความละม้ายโคไซน์ตามสมการที่ 5.2 ที่ได้แสดงในหัวข้อที่ 5.1.6 แล้วนำค่าความละม้ายที่ได้มาใช้เป็นลักษณ์ กระบวนการสร้างลักษณ์ดังกล่าวมานี้จะดำเนินไปจนกระทั่งได้ค่าความละม้ายจากทุกคู่หน่วยเทียบในคลังข้อมูลครบ

ด้วยวิธีการดังกล่าวนี้ ลักษณ์ชนิดนี้จึงสามารถสะท้อนความสัมพันธ์ทางความหมายระหว่างคำได้ และในแง่การตรวจหาการลักลอก คาดว่าลักษณ์ชนิดจะสามารถตรวจจับการแทนที่คำด้วยคำที่มีความหมายใกล้เคียงกันได้

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

5.2.3.2 ค่าความละม้ายทางความหมายของเวกเตอร์ของคำ (semantic similarity of word vector: sim_{WV})

ลักษณ์ค่าความละม้ายทางความหมายของเวกเตอร์ของคำเป็นลักษณ์ที่ผู้วิจัยวิเคราะห์หาได้จากการประยุกต์แนวคิดเรื่องการฝังคำ (word embedding) และการวัดค่าความละม้ายทางความหมายที่เสนอโดยลีและคณะ (Y. Li et al., 2004; Y. Li et al., 2006) เข้าด้วยกัน โดยหลักการแล้ว ลักษณ์ชนิดนี้จะสามารถเปรียบเทียบและคำนวณความสัมพันธ์ทางความหมายแบบคำต่อคำระหว่างข้อความในคู่หน่วยเทียบแล้วรายงานผลค่าความละม้ายของคู่หน่วยเทียบออกมาได้

การฝังคำ (word embedding) เป็นแนวคิดว่าด้วยการสร้างแบบจำลองทางภาษาและวิธีการเรียนรู้ลักษณ์ (feature learning technique) ในการประมวลผลภาษาธรรมชาติ โดยอาศัยการจับคู่คำกับเวกเตอร์ของจำนวนจริงในปริภูมิเวกเตอร์ การจับคู่ดังกล่าวนี้อาศัยการพิจารณาการปรากฏของคำนั้นๆ กับคำอื่นที่ปรากฏแวดล้อม ด้วยแนวคิดดังกล่าวนี้ คำที่ถูกแวดล้อมด้วยปริบทของคำที่คล้ายกันจึงควรมีความหมายใกล้เคียงกัน

ในงานวิจัยชิ้นนี้ ผู้วิจัยได้เลือกใช้แบบจำลอง Word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) ซึ่งเป็นแบบจำลองกลุ่มหนึ่งที่ใช้สร้างการฝังคำ เพื่อคำนวณค่าความสัมพันธ์ทางความหมายระหว่างคำ แบบจำลอง Word2vec ที่ผู้วิจัยใช้นี้ได้รับการพัฒนาให้เป็นส่วนหนึ่งของไลบรารี Gensim ซึ่งเป็นชุดเครื่องมือจำลองปริภูมิเวกเตอร์ในภาษาไพทอน (Python) อย่างไรก็ตาม การใช้งานแบบจำลอง Word2vec ต้องอาศัยข้อมูลเฉพาะภาษาเพื่อสร้างเป็นแบบจำลองการเรียนรู้ซึ่งบรรจุเวกเตอร์ของคำภายในปริภูมิเวกเตอร์ ในงานชิ้นนี้ ผู้วิจัยได้ใช้แบบจำลองซึ่งสร้างขึ้นจากข้อมูลภายในคลังข้อมูลภาษาไทยแห่งชาติ (Thai National Corpus: TNC) เป็นแบบจำลองการเรียนรู้ของ Word2vec

ในการคำนวณค่าความสัมพันธ์ทางความหมายระหว่างคำ แบบจำลอง Word2vec จะใช้หลักการเปรียบเทียบเวกเตอร์ทางความหมายของคำทั้ง 2 คำแล้วคืนค่าออกมาเป็นตัวเลขตั้งแต่ -1 ถึง 1 ซึ่งบ่งชี้ความใกล้เคียงทางความหมายเป็นระดับตั้งแต่น้อยไปหามาก ยกตัวอย่างเช่น คำว่า “ชอบ” มีความใกล้เคียงกับคำว่า “ถูกใจ” มากที่สุด โดยมีค่าความสัมพันธ์ทางความหมายระหว่างคำเท่ากับ 0.6246

ส่วนวิธีการวัดค่าความละม้ายที่เสนอโดยลีและคณะ (Y. Li et al., 2006, pp. 1142-1143) นั้นเป็นการสร้างเวกเตอร์ทางความหมายของแต่ละข้อความขึ้นจากการเปรียบเทียบความสัมพันธ์ทาง

ความหมายแบบคำต่อคำระหว่างข้อความ 2 ข้อความภายในคู่หน่วยเทียบโดยใช้ค่าระยะทางในเครือข่ายคำ (WordNet) ในการคำนวณค่าความสัมพันธ์ทางความหมาย จากนั้นจึงนำเวกเตอร์ที่ได้มาวัดค่าความละม้ายโคไซน์ระหว่างกัน อย่างไรก็ตาม เนื่องจากในกรณีของภาษาไทยยังไม่มีเครือข่ายที่เสร็จสมบูรณ์พร้อมใช้ในงาน ในการสร้างลักษณ์ประเภทนี้ ผู้วิจัยจึงได้ใช้ค่าความสัมพันธ์ทางความหมายระหว่างคำที่คำนวณได้จากเวกเตอร์ของคำตามแนวคิดการฝังคำแทน ดังตัวอย่างที่จะแสดงต่อไปนี้

T_1 = “น้อง|ขโมย|ไส้กรอก|ของ|สุนัข|ไป|รับประทาน”

T_2 = “น้อง|แย่ง|ไส้กรอก|ของ|หมา|ไป|กิน”

T_j = [“น้อง”, “ขโมย”, “ไส้กรอก”, “ของ”, “สุนัข”, “ไป”, “รับประทาน”, “น้อง, “แย่ง, “ไส้กรอก”, “ของ”, “หมา”, “ไป”, “กิน”]

จากตัวอย่างข้างต้น กำหนดให้มีข้อความ T_1 และ T_2 ที่ต้องการวัดค่าความละม้ายระหว่างข้อความทั้งสอง ในการสร้างเวกเตอร์ทางความหมายขั้นแรกนั้น ลีและคณะได้กำหนดให้สร้างชุดรวมคำ T_j ขึ้นโดยมีสมาชิกประกอบไปด้วยคำทุกคำจากข้อความ T_1 และ T_2 เรียงตามลำดับ จากนั้นในขั้นต่อมาจะสร้างตารางเพื่อคำนวณค่าความละม้ายแบบคำต่อขึ้น 2 ตาราง โดยกำหนดให้ส่วนคอลัมน์เป็นคำแต่ละคำในชุดรวมคำ T_j ส่วนแถวเป็นคำแต่ละคำในข้อความ T_1 หรือ T_2 จากนั้นจึงคำนวณค่าใช้ค่าความสัมพันธ์ทางความหมายระหว่างคำไขว้กันแบบคำต่อคำระหว่างคำในคอลัมน์และคำในแถว ด้วยวิธีการนี้จะได้ค่าความสัมพันธ์ทางความหมายระหว่างคำแบบคำต่อคำ จากนั้นจึงนำค่าความสัมพันธ์ทางความหมายที่มากที่สุดของแต่ละคอลัมน์มาสร้างเป็นเวกเตอร์ทางความหมายของข้อความแต่ละข้อความ

ภาพที่ 5.6 และภาพที่ 5.7 แสดงตารางคำนวณค่าความสัมพันธ์ทางความหมายแบบคำต่อคำของข้อความ T_1 และ T_2 ตามลำดับ จากภาพจะเห็นได้ว่าค่าความสัมพันธ์ทางความหมายระหว่างคำแบบคำต่อคำที่มากที่สุดของแต่ละคอลัมน์ได้ถูกดึงออกมาสร้างเป็นเวกเตอร์ทางความหมาย $T_{1\text{-}WV}$ และ $T_{2\text{-}WV}$ ของข้อความ T_1 และ T_2 ตามลำดับ

$T_{1\text{-}WV}$ = [1 1 1 1 1 1 1 1 0.6235 1 1 0.7758 1 0.8337]

$T_{2\text{-}WV}$ = [1 0.6235 1 1 0.7758 1 0.8337 1 1 1 1 1 1]

	น้อง	ขโมย	ไส้กรอก	ของ	สุนัข	ไป	รับประทาน	น้อง	แย่ง	ไส้กรอก	ของ	หมา	ไป	กิน
น้อง	1.0000	0.2040	0.0359	-0.2264	0.3238	0.0853	0.1387	1.0000	0.2084	0.0359	-0.2264	0.4441	0.0853	0.1939
ขโมย	0.2040	1.0000	0.0859	-0.1279	0.3109	-0.0304	0.2039	0.2040	0.6235	0.0859	-0.1279	0.3903	-0.0304	0.3847
ไส้กรอก	0.0359	0.0859	1.0000	-0.2806	0.3467	-0.0077	0.2599	0.0359	0.0931	1.0000	-0.2806	0.2957	-0.0077	0.3896
ของ	-0.2264	-0.1279	-0.2806	1.0000	-0.1148	0.0743	-0.1657	-0.2264	0.0046	-0.2806	1.0000	-0.1786	0.0743	-0.2590
สุนัข	0.3238	0.3109	0.3467	-0.1148	1.0000	0.1498	0.1051	0.3238	0.1277	0.3467	-0.1148	0.7758	0.1498	0.2328
ไป	0.0853	-0.0304	-0.0077	0.0743	0.1498	1.0000	0.0422	0.0853	0.0692	-0.0077	0.0743	0.1462	1.0000	0.0798
รับประทาน	0.1387	0.2039	0.2599	-0.1657	0.1051	0.0422	1.0000	0.1387	0.2707	0.2599	-0.1657	0.0618	0.0422	0.8337
	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩
$T_{1\text{-WV}}$ =	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.6235	1.0000	1.0000	0.7758	1.0000	0.8337

ภาพที่ 5.6 ตารางคำนวณค่าความสัมพันธ์ทางความหมายแบบคำต่อคำของตัวอย่างข้อความ T_1

	น้อง	ขโมย	ไส้กรอก	ของ	สุนัข	ไป	รับประทาน	น้อง	แย่ง	ไส้กรอก	ของ	หมา	ไป	กิน
น้อง	1.0000	0.2040	0.0359	-0.2264	0.3238	0.0853	0.1387	1.0000	0.2084	0.0359	-0.2264	0.4441	0.0853	0.1939
แย่ง	0.2084	0.6235	0.0931	0.0046	0.1277	0.0692	0.2707	0.2084	1.0000	0.0931	0.0046	0.2328	0.0692	0.3747
ไส้กรอก	0.0359	0.0859	1.0000	-0.2806	0.3467	-0.0077	0.2599	0.0359	0.0931	1.0000	-0.2806	0.2957	-0.0077	0.3896
ของ	-0.2264	-0.1279	-0.2806	1.0000	-0.1148	0.0743	-0.1657	-0.2264	0.0046	-0.2806	1.0000	-0.1786	0.0743	-0.2590
หมา	0.4441	0.3903	0.2957	-0.1786	0.7758	0.1462	0.0618	0.4441	0.2328	0.2957	-0.1786	1.0000	0.1462	0.2948
ไป	0.0853	-0.0304	-0.0077	0.0743	0.1498	1.0000	0.0422	0.0853	0.0692	-0.0077	0.0743	0.1462	1.0000	0.0798
กิน	0.1939	0.3847	0.3896	-0.2590	0.2328	0.0798	0.8337	0.1939	0.3747	0.3896	-0.2590	0.2948	0.0798	1.0000
	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩
$T_{2\text{-WV}}$ =	1.0000	0.6235	1.0000	1.0000	0.7758	1.0000	0.8337	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

ภาพที่ 5.7 ตารางคำนวณค่าความสัมพันธ์ทางความหมายแบบคำต่อคำของตัวอย่างข้อความ T_2

เมื่อได้เวกเตอร์ทางความหมาย $T_{1\text{-WV}}$ และ $T_{2\text{-WV}}$ ขั้นตอนสุดท้ายคือการวัดค่าความละม้ายระหว่างเวกเตอร์ทั้งสอง ในการนี้ ลีและคณะ (Y. Li et al., 2006, p. 1142) ได้เสนอให้ใช้วิธีวัดค่าความละม้ายโคไซน์ในการวัดเช่นเดียวกับสมการที่ 5.2 ที่ได้แสดงในหัวข้อที่ 5.1.6 ทั้งนี้ ผลจากการวัดค่าความละม้ายโคไซนระหว่างเวกเตอร์ทั้งสองเท่ากับ 0.9827

อย่างไรก็ตาม เมื่อผู้วิจัยนำวิธีการที่กล่าวมาทั้งหมดข้างต้นไปทดลองวัดค่าความละม้ายระหว่างข้อความภายในคู่หน่วยเทียบจากคลังข้อมูลที่ใช้ในการทดสอบ ผลปรากฏว่าเครื่องใช้เวลาในการคำนวณค่าความสัมพันธ์ทางความหมายแบบคำต่อคำนานเกินไป

ในกรณีทดลองสร้างลักษณ์ ผู้วิจัยได้ทดลองคำนวณค่าความสัมพันธ์ทางความหมายระหว่างข้อความที่ไม่มีการลักลอก (NA-NB) ของคู่หน่วยเทียบที่เป็นย่อหน้าสาขามนุษยศาสตร์และสังคมศาสตร์ ขนาดยาว (HS-L) ดังนี้

NA = “ความ|มุ่งหมาย|ของ|บท|นี้|คือ|การ|ชี้|ให้|เห็น|ภาพรวม|งาน|ด้าน|สังคม|ของ|พระ|ศาสนจักร|คาทอลิก|ไทย|ก่อน|สภา|สังคายนา|วาติกัน|ครั้ง|ที่|2|กล่าวคือ|การ|ดำเนินงาน|หรือ|ภารกิจ|ด้าน|สังคม|ของ|พระ|ศาสนจักร|คาทอลิก|ไทย|ใน|ช่วง|พ.ศ.|2500-2515|ซึ่ง|เป็น|ช่วง|เวลา|ก่อน|อยู่|ระหว่าง|และ|หลัง|เสร็จสิ้น|การ|ประชุม|สภา|สังคายนา|วาติกัน|ครั้ง|ที่|2|ซึ่ง|เป็น|ช่วง|เวลา|สำคัญ|ที่|สะท้อน|ให้|เห็น|ความคิด|และ|การ|ทำงาน|ด้าน|สังคม|ของ|พระ|ศาสนจักร|คาทอลิก|ไทย|ก่อน|ที่|จะ|มี|การ|จัดตั้ง|หน่วยงาน|ด้าน|สังคม|อย่าง|เป็น|ทางการ|ตาม|แนวทาง|ที่|สภา|

สังคายนา|วาติกัน|ครั้ง|ที่|2|ให้|ไว้|ดังนั้น|การ|จะ|บรรลุ|จุดมุ่งหมาย|นี้|จำเป็น|ต้อง|เริ่มต้น|จาก|การศึกษา|และ|ทำ|ความ|เข้าใจ|ประวัติศาสตร์|การ|เผยแพร่|ศาสนา|และ|โครงสร้าง|ของ|พระ|ศาสนจักร|คาทอลิก|สากล|อันเป็น|ที่มา|ของ|การ|กำเนิด|สถาปนา|ฐานานุกรม|(|Hierarchy|)|และ|การ|ดำเนิน|ภารกิจ|ใน|ด้าน|ต่าง|ๆ|ของ|พระ|ศาสนจักร|คาทอลิก|ใน|ประเทศ|ไทย|โดยเฉพาะ|อย่างยิ่ง|ภารกิจ|หรือ|งาน|ด้าน|สังคม|ที่|เป็น|ส่วน|สำคัญ|ที่|ขาด|ไม่|ได้|ใน|การ|เผยแพร่|ศาสนา|หรือ|การ|ประกาศ|พระ|วร|สาร|(|Evangelization|)|ของ|บรรดา|ธรรม|ทูต|(|Missionaries|)|"

NB = "ความ|มุ่งหมาย|ของ|บท|นี้|คือ|การ|ชี้|ให้|เห็น|ภาพรวม|ของ|การ|ทำงาน|ด้าน|สังคม|ของ|พระ|ศาสนจักร|คาทอลิก|ไทย|หลัง|สภา|สังคายนา|สากล|วาติกัน|ครั้ง|ที่|2|กล่าวคือ|เป็น|การ|นำ|เสนอ|ความ|เป็น|มา|ของ|การ|ทำงาน|ด้าน|สังคม|ของ|พระ|ศาสนจักร|คาทอลิก|ไทย|ตั้งแต่|พ.ศ.|2516-2543|เนื่องจาก|เป็น|ช่วง|เวลา|ของ|การ|เริ่มต้น|จัดตั้ง|หน่วยงาน|ที่|รับผิดชอบ|งาน|ด้าน|สังคม|อย่าง|เป็น|ระบบ|ใน|ระดับ|ชาติ|ครั้ง|แรก|ของ|พระ|ศาสนจักร|คาทอลิก|ไทย|นั่น|ก็คือ|คณะกรรมการ|สังคมสงเคราะห์|และ|ใน|เวลา|ต่อมา|ได้|เปลี่ยน|มา|เป็น|สภา|คาทอลิก|แห่ง|ประเทศ|ไทย|เพื่อ|การ|พัฒนา|ซึ่ง|เป็น|หน่วยงาน|ที่|กำเนิด|ขึ้น|ตาม|ข้อแนะนำ|ของ|สำนัก|วาติกัน|ภายหลัง|เสร็จสิ้น|การ|ประชุม|สภา|สังคายนา|ฯ|(|ค.ศ.|1962-1965/|พ.ศ.|2505-2508)|ที่|ได้|เสนอ|ให้|พระ|ศาสนจักร|ท้องถิ่น|ตั้ง|คณะกรรมการ|ว่า|ด้วย|การ|พัฒนา|สังคม|จนกระทั่ง|ถึง|ช่วง|เวลา|ที่|มี|การ|ปรับ|เปลี่ยน|โครงสร้าง|การ|ทำงาน|ของ|หน่วยงาน|ฝ่าย|ต่าง|ๆ|ของ|สภา|พระ|สังฆราช|คาทอลิก|แห่ง|ประเทศ|ไทย|ใน|ปี|ค.ศ.|2000/พ.ศ|2543|ซึ่ง|เป็น|หน่วยงาน|บริหาร|สูงสุด|ของ|พระ|ศาสนจักร|คาทอลิก|ไทย|"

จากการทดลองคำนวณค่าความสัมพันธ์ทางความหมายระหว่างข้อความพบว่า เครื่องใช้เวลาในคำนวณค่าความสัมพันธ์ทางความหมายแบบคำต่อคำ ข้อความละประมาณ 50 นาที ทั้งนี้ เป็นผลเนื่องมาจากข้อความแต่ละข้อความในคู่หน่วยเทียบประกอบด้วยคำจำนวนมาก และมีคำปรากฏซ้ำกันเพียงบางส่วนเท่านั้น เครื่องจึงต้องเวียนคำนวณค่าความสัมพันธ์ทางความหมายแบบคำต่อคำหลายรอบ เป็นเหตุให้ใช้เวลานานในการสร้างตารางคำนวณค่าความสัมพันธ์ทางความหมายแบบคำต่อคำให้เสร็จสมบูรณ์

ลักษณะที่เกิดในกรณีทดลองข้างต้นแสดงให้เห็นว่าเป็นไปได้ยากหากต้องสร้างเวกเตอร์ทางความหมายจากคู่หน่วยเทียบทั้งหมดในคลังข้อมูลจำนวน 50,000 คู่หน่วยเทียบ ด้วยเหตุนี้จึงสรุปได้ว่าการสร้างลักษณ์ค่าความละม้ายทางความหมายของเวกเตอร์ของคำเป็นอันล้มเหลว

อย่างไรก็ดี ผู้วิจัยเห็นว่าในการตรวจหาการลักลอกนั้น การประยุกต์ใช้ค่าความสัมพันธ์ทางความหมายจากเวกเตอร์ของคำตามแนวคิดการฝังคำยังมีความน่าสนใจและควรหาแนวทางพัฒนาต่อไป ทั้งนี้ เนื่องจากเห็นได้ว่าค่าความละม้ายที่วัดได้จากการทดลองวิเคราะห์หาลักษณ์ชนิดนี้ค่อนข้างสะท้อนสภาพจริงของข้อความที่มีความละม้ายกันสูง

5.2.4 ลักษณ์ทางวากยสัมพันธ์และความหมาย

ลักษณ์ทางวากยสัมพันธ์และความหมายคือลักษณ์ที่วิเคราะห์หาและสร้างขึ้นโดยอาศัยความรู้ทางภาษาศาสตร์ในระดับวลี อนุพากย์ และประโยค และระดับความหมาย ในแง่การตรวจหาการลักลอก ลักษณ์ทางวากยสัมพันธ์และความหมายจะเอื้อให้เครื่องได้เรียนรู้ความสัมพันธ์เชิงหน้าที่ระหว่างหน่วยต่างๆ ในข้อความพร้อมกับความสัมพันธ์ทางความหมายของคำในข้อความ

ลักษณ์ทางวากยสัมพันธ์และความหมายที่วิเคราะห์หาและสร้างในงานวิจัยชนิดมีจำนวน 1 ลักษณ์ ได้แก่

5.2.4.1 ค่าความละม้ายของเวกเตอร์ทางความหมายและลำดับคำ (similarity of semantic and word order: sim_{sem+wo})

ค่าความละม้ายของเวกเตอร์ทางความหมายและลำดับคำเป็นวิธีการที่วัดค่าความละม้ายที่เสนอโดยลีและคณะ (Y. Li et al., 2004; Y. Li et al., 2006) โดยเป็นการรวมเอาค่าความละม้ายของลำดับคำและค่าความละม้ายของเวกเตอร์ทางความหมายเข้าด้วยกัน ด้วยเหตุนี้ ในเชิงหลักการลักษณ์ชนิดนี้จึงมีคุณสมบัติในการตรวจจับข้อความที่ลักลอกที่ผ่านการแก้ไขโดยแทนที่คำเดิมด้วยคำที่มีความหมายใกล้เคียงกันและถูกเปลี่ยนลำดับของคำได้

อย่างไรก็ตาม ดังได้กล่าวไปแล้วในหัวข้อที่ 5.2.3.2 แล้วว่าในกรณีของภาษาไทยยังไม่มีเครือข่ายคำที่สมบูรณ์พร้อมใช้ในงาน การวิเคราะห์วิเคราะห์และสร้างลักษณ์ชนิดนี้ ผู้วิจัยจึงดัดแปลงโดยการใช้ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝงที่ได้แสดงการวิเคราะห์หาและสร้างไปแล้วในหัวข้อที่ 5.2.3.1 แทนค่าความละม้ายของเวกเตอร์ทางความหมาย และนำมารวมกับค่าความละม้ายของลำดับคำที่ได้แสดงการวิเคราะห์หาและสร้างไปแล้วในหัวข้อที่ 5.2.2.10

ส่วนการรวมค่าความละม้ายของลำดับคำและค่าความละม้ายของเวกเตอร์ทางความหมายเข้าด้วยกัน ลีและคณะ (Y. Li et al., 2006, p. 1144) ได้เสนอให้นำค่าความละม้ายทั้ง 2 ประเภทมาบวกกันโดยมีสัมประสิทธิ์ถ่วงน้ำหนักค่าความละม้ายทั้งสองไว้ ดังสมการที่ 5.10

$$\begin{aligned} sim_{sem+WO} &= \delta sim_{sem} + (1-\delta) sim_{WO} \\ &= \delta \frac{s_1 \cdot s_2}{||s_1|| \cdot ||s_2||} + (1-\delta) \frac{||r_1 - r_2||}{||r_1 + r_2||} \end{aligned} \tag{5.10}$$

โดย $\delta \leq 1$ และเป็นค่าสัมประสิทธิ์ควบคุมความสัมพันธ์ที่เกื้อหนุน (relative contribution) ระหว่างการวัดค่าความละม้ายทางความหมายและการวัดค่าความละม้ายของลำดับคำ ทั้งนี้ ในที่นี้ได้กำหนดให้ค่าดังกล่าวเท่ากับ 0.85 ตามที่ลีและคณะ (Y. Li et al., 2006, p. 1145) แนะนำไว้

ในการสร้างลักษณ์ชนิดนี้ ผู้วิจัยจะนำค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝงและค่าความละม้ายของลำดับคำของทุกคู่หน่วยเทียบในคลังข้อมูลที่ได้คำนวณไว้ก่อนนี้มารวมกันตามสมการที่ 5.10 ข้างต้น จากนั้นจึงนำมาใช้เป็นลักษณ์

ในการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน ลักษณ์ที่ประยุกต์แนวคิดดังกล่าวข้างต้นมีจำนวน 1 ลักษณ์

เมื่อได้กล่าวถึงรายละเอียดเกี่ยวกับแนวคิด การวิเคราะห์หา และการสร้าง ลักษณ์อิงอักขระและลักษณ์ทางภาษาเรียบร้อยแล้ว ในส่วนนี้ ผู้วิจัยจะสรุปรายการลักษณ์ที่วิเคราะห์หาได้ทั้งหมด 71 ลักษณ์อีกครั้ง ดังแสดงในตารางที่ 5.2 ต่อไปนี้

ตารางที่ 5.2 รายการลักษณ์ที่วิเคราะห์หาได้สำหรับใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก

#	ประเภท	ชื่อภาษาไทย	ชื่อภาษาอังกฤษ	สัญลักษณ์
1	อักขระ	ขนาดของคู่หน่วยเทียบ (อักขระ)	Pair size (characters)	$\text{Size}_{\text{Char}}$
2	อักขระ	ผลต่างของขนาดของคู่หน่วยเทียบ (อักขระ)	Difference of pair size (character)	$\text{diff}_{\text{Char}}$
3	อักขระ	ค่าระยะการแก้ไขเลเวนชเตย์นของอักขระ	Levenshtein edit distance of character	LD_{Char}
4	อักขระ	ความยาวของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ	Length of longest common subsequence of character	$\text{len}(lcs_{\text{Char}})$
5	อักขระ	ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ	Normalized longest common subsequence of character	$lcs_{\text{norm-Char}}$
6	อักขระ	ค่าความละม้ายโคไซน์ของยูนิแกรมของอักขระ	Cosine similarity of character unigram	$\cos_{\text{Char1}}$
7	อักขระ	ค่าความละม้ายโคไซน์ของไบแกรมของอักขระ	Cosine similarity of character bigram	$\cos_{\text{Char2}}$
8	อักขระ	ค่าความละม้ายโคไซน์ของไตรแกรมของอักขระ	Cosine similarity of character trigram	$\cos_{\text{Char3}}$

#	ประเภท	ชื่อภาษาไทย	ชื่อภาษาอังกฤษ	สัญลักษณ์
9	อักขระ	ค่าความละม้ายโคไซน์ของ 4 แกรมของอักขระ	Cosine similarity of character 4-gram	$\cos_{Char4}$
10	อักขระ	ค่าความละม้ายโคไซน์ของ 5 แกรมของอักขระ	Cosine similarity of character 5-gram	$\cos_{Char5}$
11	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของอักขระ	Jaccard similarity coefficient of character unigram	J_{Char1}
12	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของอักขระ	Jaccard similarity coefficient of character bigram	J_{Char2}
13	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของอักขระ	Jaccard similarity coefficient of character trigram	J_{Char3}
14	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของอักขระ	Jaccard similarity coefficient of character 4-gram	J_{Char4}
15	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของอักขระ	Jaccard similarity coefficient of character 5-gram	J_{Char5}
16	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของอักขระ	Sørensen–Dice coefficient of character unigram	QS_{Char1}
17	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของอักขระ	Sørensen–Dice coefficient of character bigram	QS_{Char2}
18	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของอักขระ	Sørensen–Dice coefficient of character trigram	QS_{Char3}
19	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของอักขระ	Sørensen–Dice coefficient of character 4-gram	QS_{Char4}
20	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของอักขระ	Sørensen–Dice coefficient of character 5 gram	QS_{Char5}
21	ศัพท์	ขนาดของคู่หน่วยเทียบ (คำ)	Pair size (words)	$Size_W$
22	ศัพท์	ผลต่างของขนาดของคู่หน่วยเทียบ (คำ)	Difference of pair size (words)	$diff_W$
23	ศัพท์	ค่าระยะการแก้ไขเลเวนชเตย์นของคำ	Levenshtein edit distance of word	LD_W
24	ศัพท์	ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ	Length of longest common subsequence of word	$\text{len}(lcs_W)$
25	ศัพท์	ค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ	Normalized longest common subsequence of word	$lcs_{norm\text{-}W}$

#	ประเภท	ชื่อภาษาไทย	ชื่อภาษาอังกฤษ	สัญลักษณ์
26	ศัพท์	ค่าความละม้ายโคไซน์ของยูนิแกรมของคำ	Cosine similarity of word unigram	cos_{W1}
27	ศัพท์	ค่าความละม้ายโคไซน์ของไบแกรมของคำ	Cosine similarity of word bigram	cos_{W2}
28	ศัพท์	ค่าความละม้ายโคไซน์ของไตรแกรมของคำ	Cosine similarity of word trigram	cos_{W3}
29	ศัพท์	ค่าความละม้ายโคไซน์ของ 4 แกรมของคำ	Cosine similarity of word 4-gram	cos_{W4}
30	ศัพท์	ค่าความละม้ายโคไซน์ของ 5 แกรมของคำ	Cosine similarity of word 5-gram	cos_{W5}
31	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของคำ	Jaccard similarity coefficient of word unigram	J_{W1}
32	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของคำ	Jaccard similarity coefficient of word bigram	J_{W2}
33	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของคำ	Jaccard similarity coefficient of word trigram	J_{W3}
34	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของคำ	Jaccard similarity coefficient of word 4-gram	J_{W4}
35	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของคำ	Jaccard similarity coefficient of word 5-gram	J_{W5}
36	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ	Sørensen–Dice coefficient of word unigram	QS_{W1}
37	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ	Sørensen–Dice coefficient of word bigram	QS_{W2}
38	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของคำ	Sørensen–Dice coefficient of word trigram	QS_{W3}
39	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของคำ	Sørensen–Dice coefficient of word 4-gram	QS_{W4}
40	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของคำ	Sørensen–Dice coefficient of word 5 gram	QS_{W5}
41	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักยูนิแกรมของคำแบบ tf-idf	Cosine similarity of tf-idf trem weight unigram	$cos_{tf\text{-}idf\text{-}1}$
42	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักไบแกรมของคำแบบ tf-idf	Cosine similarity of tf-idf trem weight bigram	$cos_{tf\text{-}idf\text{-}2}$

#	ประเภท	ชื่อภาษาไทย	ชื่อภาษาอังกฤษ	สัญลักษณ์
43	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักไตรแกรมของคำแบบ tf-idf	Cosine similarity of tf-idf trem weight trigram	$\cos_{\text{tf-idf-3}}$
44	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนัก 4 แกรมของคำแบบ tf-idf	Cosine similarity of tf-idf trem weight 4-gram	$\cos_{\text{tf-idf-4}}$
45	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนัก 5 แกรมของคำแบบ tf-idf	Cosine similarity of tf-idf trem weight 5-gram	$\cos_{\text{tf-idf-5}}$
46	วากยสัมพันธ์	ค่าระยะการแก้ไขเลเวนชเตย์นของหมวดคำ	Levenshtein edit distance of POS	LD_{POS}
47	วากยสัมพันธ์	ความยาวของลำดับย่อยร่วมที่ยาวที่สุดของหมวดคำ	Length of longest common subsequence of POS	$\text{len}(lcs_{\text{POS}})$
48	วากยสัมพันธ์	ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของหมวดคำ	Normalized longest common subsequence of POS	$lcs_{\text{norm-POS}}$
49	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของยูนิแกรมของหมวดคำ	Cosine similarity of POS unigram	$\cos_{\text{POS1}}$
50	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของไบแกรมของหมวดคำ	Cosine similarity of POS bigram	$\cos_{\text{POS2}}$
51	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของไตรแกรมของหมวดคำ	Cosine similarity of POS trigram	$\cos_{\text{POS3}}$
52	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของ 4 แกรมของหมวดคำ	Cosine similarity of POS 4-gram	$\cos_{\text{POS4}}$
53	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของ 5 แกรมของหมวดคำ	Cosine similarity of POS 5-gram	$\cos_{\text{POS5}}$
54	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของหมวดคำ	Jaccard similarity coefficient of POS unigram	J_{POS1}
55	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของหมวดคำ	Jaccard similarity coefficient of POS bigram	J_{POS2}
56	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของหมวดคำ	Jaccard similarity coefficient of POS trigram	J_{POS3}
57	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของหมวดคำ	Jaccard similarity coefficient of POS 4-gram	J_{POS4}
58	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของหมวดคำ	Jaccard similarity coefficient of POS 5-gram	J_{POS5}
59	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของหมวดคำ	Sørensen–Dice coefficient of POS unigram	QS_{POS1}

#	ประเภท	ชื่อภาษาไทย	ชื่อภาษาอังกฤษ	สัญลักษณ์
60	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของหมวดคำ	Sørensen–Dice coefficient of POS bigram	QS_{POS2}
61	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของหมวดคำ	Sørensen–Dice coefficient of POS trigram	QS_{POS3}
62	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของหมวดคำ	Sørensen–Dice coefficient of POS 4-gram	QS_{POS4}
63	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของหมวดคำ	Sørensen–Dice coefficient of POS 5 gram	QS_{POS5}
64	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของคำ	Cosine similarity of word n-gram range	$\cos_{W123}$
65	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf	Cosine similarity of tf-idf term weight n-gram range	$\cos_{tf\text{-}idf\text{-}123}$
66	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของหมวดคำ	Cosine similarity of POS n-gram range	$\cos_{POS123}$
67	วากยสัมพันธ์	ค่าความละม้ายของลำดับคำ	Word order similarity	sim_{WO}
68	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของลำดับคำ	Cosine similarity of word order	$\cos_{WO}$
69	ความหมาย	ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง	Cosine similarity of latent semantic vector	$\cos_{LSA}$
70	ความหมาย	ค่าความละม้ายทางความหมายของเวกเตอร์ของคำ[16]	semantic similarity of word vector	sim_{wv}
71	วากยสัมพันธ์+ความหมาย	ค่าความละม้ายของเวกเตอร์ทางความหมายและลำดับคำ	Similarity of semantic and word order	sim_{sem+wo}

จากรายละเอียดว่าด้วยการวิเคราะห์หาและการสร้างลักษณ์ที่กล่าวมาทั้งหมดในบทนี้จะเห็นได้ว่าผู้วิจัยได้วิเคราะห์ลักษณะของข้อความที่มีการลักลอกและไม่มีการลักลอกทั้งในแง่ที่เป็นลักษณะทางภาษาและไม่ใช่ลักษณะทางภาษาแล้วนำมาประยุกต์ให้เข้ากับวิธีการทางการประมวลผลภาษาธรรมชาติที่มีใช้แพร่หลายในปัจจุบันจนสามารถสร้างเป็นลักษณ์ชนิดต่างๆ ดังได้กล่าวมา

ในแง่ของลักษณะที่ไม่ใช่ลักษณะทางภาษานั้น ผู้วิจัยได้วิเคราะห์หาลักษณ์จากรูปอักขระ ความยาวของอักขระ ลำดับของอักขระ ในฐานะเส้นฐาน (baseline) ในการทดลองเปรียบเทียบประสิทธิภาพของลักษณ์ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก ลักษณ์ใน

[16] ล้มเหลวในขั้นตอนการสร้างลักษณ์

กลุ่มนี้จะถูกใช้เป็นเกณฑ์เปรียบเทียบให้เห็นถึงประสิทธิภาพในการจำแนกของลักษณ์ทางภาษาซึ่งเป็นลักษณ์ที่งานวิจัยชิ้นนี้สนใจ

ส่วนในแง่ของลักษณะทางภาษา เพื่อตอบคำถามวิจัยตามที่ปรากฏในวัตถุประสงค์ข้อแรกของงานวิจัยชิ้นนี้ ในขั้นตอนก่อนหน้านี้ ผู้วิจัยได้วิเคราะห์กลวิธีการลักลอกงานวิชาการในภาษาไทยก่อน ส่งผลให้ถึงลักษณะและกลไกทางภาษาที่ใช้ในการลักลอก จนนำมาสู่การวิเคราะห์หาลักษณ์ทางภาษาโดยแบ่งประเภทของลักษณ์ตามระดับของหน่วยทางภาษา

ในระดับคำ ผู้วิจัยได้วิเคราะห์หาและสร้างลักษณ์ทางศัพท์ขึ้น ลักษณ์ในกลุ่มนี้ได้วิเคราะห์หาได้จากลักษณะของรูปคำ ขอบเขตของคำ และลำดับของคำ (โดยผ่านการประยุกต์ใช้แนวคิดเรื่องระยะการแก้ไข ลำดับร่วมที่ยาวที่สุด และเอ็นแกรม)

ในระดับวลีและประโยค ผู้วิจัยได้วิเคราะห์หาและสร้างลักษณ์ทางวากยสัมพันธ์ขึ้น โดยวิเคราะห์หาจากหมวดคำ และการเรียงลำดับของคำและหมวดคำในข้อความ อย่างไรก็ตาม ในส่วนการแทนรูปความสัมพันธ์แบบพึ่งพาระหว่างหน่วยต่างๆ ในข้อความนั้น จากการสำรวจเครื่องมือที่ใช้สำหรับแจกส่วนประโยคและกำกับความสัมพันธ์แบบพึ่งพาในภาษาไทยที่มีอยู่ในปัจจุบันแล้ว พบว่ายังไม่มีเครื่องมือชิ้นใดที่ให้ผลการแจกส่วนและกำกับความสัมพันธ์ในระดับที่น่าพอใจ หากนำเครื่องมือเหล่านั้นมาใช้ก็อาจส่งผลกระทบต่อผลการวิจัยในภาพรวม ด้วยเหตุนี้ ผู้วิจัยจึงตัดสินใจไม่ใช้ความสัมพันธ์แบบพึ่งพาเป็นลักษณ์สำหรับจำแนกข้อความลักลอกและไม่ลักลอกในงานวิจัยชิ้นนี้ อย่างไรก็ดี ลักษณ์ที่วิเคราะห์หาและสร้างได้ในกลุ่มนี้ก็สะท้อนสะท้อนความสัมพันธ์ในระดับวากยสัมพันธ์ได้ในระดับหนึ่ง ซึ่งอาจเพียงต่อการใช้จำแนกการลักลอก

ในระดับความหมาย ผู้วิจัยได้ประยุกต์แนวคิดต่างๆ ในการแทนรูปความสัมพันธ์ทางความหมายระหว่างคำในข้อความออกมาได้ และนำมาสร้างเป็นเวกเตอร์ทางความหมายเพื่อใช้วัดค่าความละม้ายของข้อความ อย่างไรก็ตาม เป็นที่น่าเสียดายว่าในขั้นตอนการสร้างลักษณ์จากเวกเตอร์ทางความหมายของคำตามแนวคิดการฝังคำนั้นต้องใช้เวลาค่อนมากในการคำนวณค่าความสัมพันธ์แบบคำต่อคำ จึงทำให้ไม่สามารถนำลักษณ์ค่าความละม้ายทางความหมายของเวกเตอร์ของคำมาใช้ในการทดสอบได้ อย่างไรก็ดี ผู้วิจัยเห็นควรให้พัฒนาวิธีการตรวจหาการลักลอกโดยอาศัยแนวคิดการฝังคำต่อไป เนื่องจากแนวคิดดังกล่าวสามารถคืนค่าที่เป็นรูปแทนในระดับความหมายได้โดยไม่ต้องพึ่งพาการตัดสินความหมายโดยมนุษย์ จึงช่วยลดปัญหาอันเกิดจากอัตวิสัยของมนุษย์ลงได้ ซึ่งลักษณะดังกล่าวนี้เป็นแนวทางที่ควรดำเนินไปในการสร้างนวัตกรรมในอนาคต

ลักษณ์ที่วิเคราะห์หาและสร้างได้ตามที่กล่าวไปในบทนี้จะเข้าสู่กระบวนการทดสอบประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกในลำดับต่อไป ทั้งนี้ผู้วิจัยจะได้นำเสนอผลการทดสอบประสิทธิภาพของลักษณ์ในบทถัดไป

บทที่ 6
ผลการประเมินประสิทธิภาพของระบบ

การประเมินประสิทธิภาพของระบบตรวจหาการลักลอกที่ถูกสร้างขึ้นถือเป็นวัตถุประสงค์หลักอีกประการหนึ่งของงานวิจัยชิ้นนี้ ในบทนี้ ผู้วิจัยจะได้นำเสนอผลประเมินประสิทธิภาพของระบบตามวิธีการประเมินที่ได้กล่าวถึงไปแล้วในหัวข้อที่ 3.5 โดยจะแบ่งการนำเสนอออกเป็น 2 หัวข้อใหญ่ ได้แก่ ผลการประเมินประสิทธิภาพของข้อมูลรับเข้า และผลการประเมินประสิทธิภาพของลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก รายละเอียดมีดังต่อไปนี้

6.1 ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ข้อมูลรับเข้าที่ต่างกัน

ลักษณะของข้อมูลรับเข้าถือเป็นปัจจัยหนึ่งที่ส่งผลโดยตรงต่อประสิทธิภาพการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกของระบบ การทดลองในขั้นนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกของระบบเมื่อใช้ข้อมูลรับเข้า 2 ประเภทที่แตกต่างกัน ได้แก่ ย่อหน้า และหน่วยปริจเฉทพื้นฐาน โดยได้ดำเนินการทดลองตามวิธีการที่ได้กล่าวไปในหัวข้อที่ 3.5.3

ในการทดลอง ผู้วิจัยได้ควบคุมให้จำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกในชุดข้อมูลทดลองที่สร้างขึ้นเป็นการเฉพาะ และใช้ลักษณ์ในการจำแนกประเภทเป็นรายลักษณ์ทั้งหมด 5 ลักษณ์ ได้แก่ ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ (QS_{W1}), ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}), ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของคำ (QS_{W3}), ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของคำ (QS_{W4}), และค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของคำ (QS_{W5})

ในส่วนของข้อมูลที่ใช้ในการทดลองนั้น เนื่องมาจากในกรณีของภาษาไทยนั้นยังไม่มีเครื่องมือสำหรับตัดแยกขอบเขตของหน่วยปริจเฉทพื้นฐานที่มีประสิทธิภาพเป็นที่น่าพอใจ ฉะนั้นการทดลองในขั้นนี้จึงไม่สามารถใช้ข้อมูลจากคลังข้อมูลการลักลอกที่ผู้วิจัยสร้างขึ้นทั้งหมดได้ ด้วยเหตุนี้ ผู้วิจัยจึงได้สร้างชุดข้อมูลขนาดย่อมขึ้นเพื่อใช้ในการทดลอง ชุดข้อมูลทดลองนี้ได้มาจากสุ่มตัวอย่างคู่หน่วยเทียบตามสาขาวิชา ขนาดของย่อหน้า และประเภทของการลักลอก ให้ได้จำนวนทั้งสิ้น 150 คู่หน่วยเทียบ จากนั้นจึงนำมาสร้างชุดข้อมูลทดลอง 2 ชุดตามลักษณะของข้อมูลรับเข้าที่ต้องการศึกษา ได้แก่ ชุดข้อมูลทดลอง PRG และชุดข้อมูลทดลอง EDU ชุดข้อมูลทดลอง PRG เป็นตัวแทนของข้อมูลรับเข้าเป็นย่อหน้า ผู้วิจัยจะคงรูปแบบของคู่หน่วยเทียบที่เป็นย่อหน้าไว้เช่นเดิมตามที่สุ่มได้จาก

คลังข้อมูลการลักลอก ส่วนชุดข้อมูลทดลอง EDU นั้น ผู้วิจัยจะนำย่อหน้าที่อยู่ในรูปคู่หน่วยเทียบมาตัดแยกขอบเขตของหน่วยปริจเฉทพื้นฐานตามหลักการของนลินี อินต๊ะซาว และวิโรจน์ อรุณมานะกุล (Intasaw & Aroonmanakun, 2013) ทั้งย่อหน้าต้นฉบับและย่อหน้าที่ผ่านการลักลอก เพื่อใช้เป็นตัวแทนของข้อมูลรับเข้าที่เป็นหน่วยปริจเฉทพื้นฐาน

ในส่วนชุดข้อมูลทดลอง PRG ซึ่งเป็นข้อมูลตัวแทนของหน่วยรับเข้าประเภทย่อหน้านั้น ผู้วิจัยจะเขียนโปรแกรมเพื่อวัดค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ระหว่างย่อหน้าของคู่หน่วยเทียบขึ้น และใช้ค่าความละม้ายดังกล่าวเป็นลักษณ์ที่ได้จากข้อมูลรับเข้าที่เป็นย่อหน้า ดังนั้นจำนวนของค่าที่ใช้เป็นลักษณ์ที่ได้ในขั้นนี้จะมีจำนวนทั้งหมด 600 ค่าตามจำนวนของคู่หน่วยเทียบที่เป็นคู่ของย่อหน้า 150 คู่คูณด้วยจำนวนประเภทของลักษณ์ทั้ง 5 ลักษณ์ ส่วนการให้คำตอบสำหรับการเรียนรู้ของเครื่องจากคู่หน่วยเทียบทั้งหมด 150 คู่นั้น คู่หน่วยเทียบของย่อหน้าที่ไม่เป็นคู่ลักลอกจะได้รับการกำหนดค่าคำตอบเป็น 0 ส่วนหน่วยเทียบของย่อหน้าที่เป็นคู่ลักลอกจะได้รับการกำหนดค่าคำตอบเป็น 1

ส่วนชุดข้อมูลทดลอง EDU ซึ่งเป็นตัวแทนของหน่วยรับเข้าประเภทหน่วยปริจเฉทพื้นฐานนั้นเป็นการเทียบคู่ของหน่วยปริจเฉทพื้นฐานในคู่ของย่อหน้าแบบพบกันทั้งหมด ผู้วิจัยจึงต้องให้คำตอบของการจำแนกประเภทใหม่ โดยในขั้นแรก ผู้วิจัยจะจัดแนว (align) เทียบระหว่างหน่วยปริจเฉทพื้นฐานในย่อหน้าต้นฉบับกับหน่วยปริจเฉทพื้นฐานในย่อหน้าลักลอกเพื่อให้คำตอบสำหรับการเรียนรู้ของเครื่อง ในการกรณีของคู่หน่วยเทียบของหน่วยปริจเฉทพื้นฐานที่ไม่เป็นคู่ลักลอกจะได้รับการกำหนดค่าคำตอบเป็น 0 ส่วนคู่หน่วยเทียบของหน่วยปริจเฉทพื้นฐานที่เป็นคู่ลักลอกจะได้รับการกำหนดค่าคำตอบเป็น 1 ด้วยวิธีการนี้จะทำให้ได้ค่าคำตอบทั้งหมด 56,340 ค่าเท่ากับจำนวนคู่หน่วยเทียบของหน่วยปริจเฉทพื้นฐานที่ได้จากคู่ของย่อหน้าตั้งต้น 150 คู่ และได้จำนวนค่าที่ใช้เป็นลักษณ์ทั้งหมด 281,700 ค่าตามจำนวนของคู่หน่วยเทียบที่เป็นคู่ของหน่วยปริจเฉทพื้นฐาน 56,340 คู่คูณด้วยจำนวนประเภทของลักษณ์ทั้ง 5 ลักษณ์

ชุดข้อมูลทดลอง PRG และชุดข้อมูลทดลอง EDU จะถูกนำมาใช้ฝึกฝนและทดสอบประสิทธิภาพของระบบแยกต่างหากออกจากกันเพื่อเปรียบเทียบประสิทธิภาพของลักษณะของข้อมูลรับเข้าตามวิธีการที่กล่าวไปแล้วในหัวข้อที่ 3.5.2

ตารางที่ 6.1 แสดงผลการประเมินประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกของระบบที่ลักษณ์ต่างๆ ทั้ง 5 ลักษณ์ เมื่อใช้ข้อมูลรับเข้า 2 ประเภทที่แตกต่างกัน ได้แก่ ย่อหน้า (PRG) และหน่วยปริจเฉทพื้นฐาน (EDU) จากตารางดังกล่าวจะเห็นได้ว่าในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกด้วยลักษณ์ทั้ง 5 ลักษณ์ ค่า F ที่ได้

จากการใช้ย่อหน้าเป็นข้อมูลรับเข้าสูงกว่าค่า F ที่ได้จากการใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้าทั้งหมดทุกลักษณ์

ตารางที่ 6.1 ผลการประเมินประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกของระบบเมื่อใช้ข้อมูลรับเข้าประเภทย่อหน้าและหน่วยปริจเฉทพื้นฐาน

ลักษณ์	ข้อมูลรับเข้า	ค่าการประเมิน		
		Prec.	Rec.	F
QS_{W1}	PRG	1.0000	0.9833	0.9913
	EDU	0.8443	0.8151	0.8263
QS_{W2}	PRG	1.0000	0.9750	0.9870
	EDU	0.8449	0.7835	0.8089
QS_{W3}	PRG	0.9917	0.9667	0.9786
	EDU	0.8671	0.6920	0.7678
QS_{W4}	PRG	0.9763	0.9417	0.9559
	EDU	0.8823	0.5882	0.7034
QS_{W5}	PRG	0.9780	0.9417	0.9562
	EDU	0.8842	0.4882	0.6268

ทั้งนี้ ผู้วิจัยเห็นว่าการที่ประสิทธิภาพในการจำแนกประเภทของระบบที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้าต่ำกว่าระบบที่ใช้ย่อหน้าเป็นข้อมูลรับเข้านั้น เนื่องมาจากสาเหตุหลัก 3 ประการ ดังนี้

สาเหตุประการแรกคือ ขนาดของหน่วยปริจเฉทพื้นฐาน ทั้งนี้ เป็นที่ทราบแล้วว่าหน่วยปริจเฉทพื้นฐานนั้นประกอบด้วยจำนวนคำที่น้อยกว่าย่อหน้า บางหน่วยปริจเฉทอาจประกอบด้วยคำเพียง 1-3 คำเท่านั้น ลักษณะดังกล่าวนี้ทำให้เมื่อประยุกต์ใช้ลักษณ์ค่าความละม้ายที่คำนวณจากเอ็นแกรมของคำที่มากขึ้นเรื่อยๆ ค่าความละม้ายก็อาจเท่ากับ 0 ได้ ยกตัวอย่างเช่น

$$T_1 = [\text{การ}|\text{แสดง}|\text{นี้}|\text{มิใช่}|\text{การ}|\text{ประกาศ}|\text{ชักชวน}|\text{ให้}|\text{ศิลปิน}]_{T1\text{-}1}\ [\text{ที่}|\text{สนใจ}]_{T1\text{-}2}$$

$$T_2 = [\text{การ}|\text{แสดง}|\text{นี้}|\text{มิใช่}|\text{การ}|\text{ประกาศ}|\text{ชักชวน}|\text{ให้}|\text{นักร้อง}]_{T2\text{-}1}\ [\text{ที่}|\text{สนใจ}]_{T2\text{-}2}$$

จากตัวอย่าง สมมติให้ข้อความ T_1 และ T_2 เป็นข้อความลักลอก จะเห็นได้ว่าข้อความทั้งสองประกอบด้วยหน่วยปริจเฉทพื้นฐานข้อความละ 2 หน่วย ในกรณีที่กำหนดให้ข้อมูลรับเข้าเป็นหน่วยปริจเฉทพื้นฐาน หน่วยปริจเฉทพื้นฐานทุกหน่วยจะถูกนำมาวัดค่าความละม้ายแบบพบกันทั้งหมดเพื่อ

หานำค่าความละม้ายดังกล่าวมาใช้เป็นลักษณ์ ในกรณีนี้ คู่หน่วยเทียบของหน่วยปริจเฉทพื้นฐาน T1-2 และ T2-2 จะให้เป็นคู่ลักลอกที่เหมือนกันทุกประการ แต่ผลการวัดค่าความละม้ายกลับปรากฏว่าคู่ลักลอกดังกล่าวมีค่า QS_{W1}, QS_{W2}, QS_{W3}, QS_{W4}, และ QS_{W5} เท่ากับ 1, 1, 0, 0, และ 0 ตามลำดับ สาเหตุที่ค่าความละม้ายตั้งแต่ไตรแกรมขึ้นไปเท่ากับ 0 นั้น เนื่องมาจากหน่วยปริจเฉทพื้นฐานแต่ละหน่วยในคู่หน่วยเทียบประกอบด้วยคำเพียง 2 คำ คือ “ที่” และ “สนใจ” ด้วยเหตุนี้ เมื่อวัดค่าความละม้ายในระดับไตรแกรมของคำจึงคืนค่าเป็น 0 ค่าความละม้ายที่เท่ากับ 0 ตามตัวอย่างนี้ส่งผลให้การตัดสินใจจำแนกประเภทของเครื่องเบี่ยงเบนไป แม้ว่าคู่หน่วยเทียบในตัวอย่างนี้จะได้รับการกำหนดคำตอบในการเรียนรู้ของเครื่องว่าเป็นการลักลอก แต่ลักษณ์ที่ใช้เป็นค่าในการตัดสินใจกลับเท่ากับ 0 เหมือนกันกับคู่หน่วยเทียบอื่นๆ ที่ได้รับการกำหนดคำตอบว่าไม่ลักลอก

ลักษณะดังกล่าวข้างต้นสอดคล้องกับผลการประเมินประสิทธิภาพในตารางที่ 6.1 ที่ค่า F ของการจำแนกที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้า ดังจะเห็นได้ว่าเมื่อขนาดของเอ็นแกรมของคำเพิ่มมากขึ้น ค่า F กลับต่ำลงอย่างเห็นได้ชัด ด้วยเหตุผลด้านจำนวนคำในหน่วยปริจเฉทพื้นฐานดังได้กล่าวมานี้ ในขณะที่ ค่า F ของการจำแนกที่ใช้ย่อหน้าเป็นข้อมูลรับเข้ากลับต่ำลงเพียงเล็กน้อยเมื่อขนาดของเอ็นแกรมของคำเพิ่มมากขึ้น

ส่วนสาเหตุประการต่อมานั้นเป็นสาเหตุที่เกี่ยวกับแนวคิดเชิงทฤษฎี ดังได้กล่าวไปในหัวข้อที่ 2.8 ว่าด้วยทฤษฎีโครงสร้างวาทะแล้วว่า ตามทฤษฎีดังกล่าว หน่วยปริจเฉทพื้นฐานแต่ละหน่วยที่ประกอบกันขึ้นเป็นปริจเฉทย่อมมีวาทสัมพันธ์อย่างใดอย่างหนึ่งต่อกัน ในกรณีของการนำหน่วยปริจเฉทพื้นฐานมาใช้เป็นข้อมูลรับเข้าในการตรวจหาการลักลอก ผู้วิจัยเห็นว่าหากตัดแบ่งเฉพาะหน่วยปริจเฉทพื้นฐานมาใช้โดยไม่กำกับและพิจารณาวาทสัมพันธ์ระหว่างหน่วยต่างๆ หน่วยปริจเฉทพื้นฐานก็มีสถานะเป็นเพียงข้อความที่มีขนาดสั้นกว่าย่อหน้า ไม่สามารถสะท้อนความสัมพันธ์ภายในปริจเฉทอันเป็นแง่มุมหนึ่งที่ควรพิจารณาในการลักลอกและการตรวจหาการลักลอกได้ และจะส่งผลต่อประสิทธิภาพของการตรวจหาการลักลอกดังได้กล่าวไปแล้วข้างต้น

สาเหตุประการสุดท้ายเป็นสาเหตุอันเนื่องมาการกลวิธีที่ผู้ลักลอกใช้ในลักลอก ทั้งนี้ กลวิธีลักลอกวิธีหนึ่งที่ผู้ลักลอกสามารถทำได้คือการแทรกหรือลบเนื้อหาในข้อความในระดับหน่วยปริจเฉทพื้นฐานดังได้กล่าวไปแล้วในหัวข้อที่ 4.1.2 ในบทที่ 4 ทั้งนี้ ในการวัดค่าความละม้ายเพื่อนำมาใช้เป็นลักษณ์ คู่หน่วยเทียบของหน่วยปริจเฉทพื้นฐานที่จับคู่จากหน่วยปริจเฉทพื้นฐานที่ถูกแทรกเข้าไปใหม่ย่อมคืนค่าความละม้ายเป็น 0 เสมอ ในขณะที่การลบหน่วยปริจเฉทพื้นฐานก็จะส่งผลให้ไม่สามารถจับคู่ลักลอกที่แท้จริงได้ เนื่องจากหน่วยปริจเฉทพื้นเดิมที่ปรากฏอยู่ในข้อความต้นฉบับได้ถูกลบออกไปในข้อความลักลอก อย่างไรก็ตาม การกำหนดคำตอบสำหรับการเรียนรู้ของเครื่องจะกำหนดว่าลักษณะเช่นนี้ถือเป็นการลักลอก กรณีเช่นนี้ก็จะส่งผลให้การตัดสินใจจำแนกประเภทของเครื่อง

เบี่ยงเบนไป เนื่องจากค่าความละม้ายที่เครื่องเรียนรู้มีลักษณะเหมือนกันทั้งจากการกำหนดคำตอบว่าเป็นการลักลอกและไม่เป็นการลักลอก เหตุผลประการสุดท้ายนี้ยิ่งเป็นการแสดงให้ข้อจำกัดของหน่วยรับเข้าประเภทหน่วยปริจเฉทพื้นฐานที่ไม่ได้รับการกำกับวาทสัมพันธ์มาพร้อมกันก่อนนำมาใช้เป็นข้อมูลรับเข้า

ด้วยเหตุดังได้กล่าวไปข้างต้นนี้ เมื่อพิจารณาในภาพรวม โดยเฉพาะในกรณีที่มีข้อมูลการลักลอกหลายประเภทปะปนกันเช่นในกรณีของชุดข้อมูลทดลอง ย่อหน้าจึงเหมาะสมจะใช้เป็นข้อมูลรับเข้าในระบบตรวจหาการลักลอกมากกว่าหน่วยปริจเฉทพื้นฐาน

ทั้งนี้ เพื่อตอบวัตถุประสงค์ของการวิจัยข้อที่ 3.1 ว่าด้วยความเหมาะสมของการใช้ข้อมูลรับเข้าแต่ละประเภทในการตรวจหาข้อมูลลักลอกแต่ละประเภทให้ชัดเจนยิ่งขึ้น ในการทดลองขั้นตอนต่อมา ผู้วิจัยได้จัดแบ่งชุดข้อมูลทดสอบเดิมออกเป็น 4 ชุดย่อย แต่ละชุดย่อยจะประกอบด้วยข้อมูลลักลอกเฉพาะประเภทกับข้อมูลที่ไม่มีการลักลอก ได้แก่ ข้อมูลลักลอกประเภทคัดลอกโดยตรงและข้อมูลที่ไม่มีการลักลอก (EC-NO), ข้อมูลลักลอกประเภทคัดลอกโดยใกล้เคียงและข้อมูลที่ไม่มีการลักลอก (NC-NO), ข้อมูลลักลอกประเภทคัดลอกโดยดัดแปลงและข้อมูลที่ไม่มีการลักลอก (MO-NO), และข้อมูลลักลอกประเภทถอดความและข้อมูลที่ไม่มีการลักลอก (PA-NO) จากนั้นจึงนำชุดข้อมูลทดลองย่อยแต่ละชุดมาทดลองเปรียบเทียบประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกของระบบ โดยใช้ข้อมูลรับเข้า 2 ประเภทที่แตกต่างกัน ได้แก่ ย่อหน้าและหน่วยปริจเฉทพื้นฐาน และใช้ลักษณ์ค่าความละม้ายโซเรนเซน-ไดซ์จากเอ็นแกรมของคำทั้ง 5 ลักษณ์ เช่นเดียวกันกับการทดลองในขั้นแรก

ตารางที่ 6.2 แสดงผลการประเมินประสิทธิภาพของระบบในการจำแนกประเภทข้อมูลลักลอกแต่ละประเภทในชุดทดสอบย่อย 4 ชุด เปรียบเทียบระหว่างระบบที่ใช้ย่อหน้าเป็นข้อมูลรับเข้ากับระบบที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้า จากตารางจะเห็นได้ว่าเมื่อใช้ลักษณ์ค่าความละม้ายโซเรนเซน-ไดซ์ของเอ็นแกรมของคำตั้งแต่ 1-5 แกรมในการทดสอบ ผลปรากฏว่าระบบที่ใช้ย่อหน้าเป็นข้อมูลรับเข้าให้ค่า F สูงกว่าระบบที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้าทั้งหมดในข้อมูลลักลอกทุกประเภท นอกจากนี้ หากพิจารณาค่า F เปรียบเทียบกันระหว่างระบบที่ใช้ย่อหน้าเป็นข้อมูลรับเข้ากับระบบที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้า จะพบว่าเมื่อขนาดของเอ็นแกรมของคำเพิ่มมากขึ้น ค่า F ที่ได้จากระบบที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้าจะต่ำกว่าค่า F ที่ได้ระบบที่ใช้ย่อหน้าเป็นข้อมูลรับเข้าอย่างเห็นได้ชัด ลักษณะดังกล่าวนี้สอดคล้องกับข้ออภิปรายของผู้วิจัยที่ได้กล่าวไปในตอนต้นเกี่ยวกับข้อจำกัดอันเกิดจากขนาดของหน่วยปริจเฉทพื้นฐาน

ตารางที่ 6.2 ผลการประเมินประสิทธิภาพในการจำแนกประเภทข้อมูลลักลอกเฉพาะประเภทและข้อมูลที่ไม่มีการลักลอก เปรียบเทียบระหว่างข้อมูลรับเข้าประเภทย่อหน้ากับหน่วยปริจเฉทพื้นฐาน

ลักษณ์	ประเภทข้อมูล	ข้อมูลรับเข้า	ค่าการประเมิน		
			Prec.	Rec.	F
QS_{W1}	EC-NO	PRG	1.0000	1.0000	1.0000
		EDU	0.7812	1.0000	0.8700
	NC-NO	PRG	1.0000	1.0000	1.0000
		EDU	0.7014	0.8825	0.7725
	MO-NO	PRG	1.0000	0.9667	0.9800
		EDU	0.7295	0.8615	0.7808
	PA-NO	PRG	1.0000	0.9667	0.9800
		EDU	0.6407	0.5126	0.5449
QS_{W2}	EC-NO	PRG	1.0000	1.0000	1.0000
		EDU	0.8096	0.9566	0.8728
	NC-NO	PRG	1.0000	1.0000	1.0000
		EDU	0.7257	0.7631	0.7337
	MO-NO	PRG	1.0000	0.9667	0.9800
		EDU	0.7627	0.7546	0.7497
	PA-NO	PRG	1.0000	0.9333	0.9600
		EDU	0.6595	0.4486	0.5123
QS_{W3}	EC-NO	PRG	1.0000	1.0000	1.0000
		EDU	0.8211	0.8756	0.8444
	NC-NO	PRG	1.0000	1.0000	1.0000
		EDU	0.7781	0.6311	0.6910
	MO-NO	PRG	1.0000	0.9667	0.9800
		EDU	0.7657	0.6598	0.6997
	PA-NO	PRG	0.9750	0.9000	0.9257
		EDU	0.6616	0.3681	0.4525
QS_{W4}	EC-NO	PRG	1.0000	1.0000	1.0000
		EDU	0.8562	0.7608	0.8035

ลักษณ์	ประเภทข้อมูล	ข้อมูลรับเข้า	ค่าการประเมิน		
			Prec.	Rec.	F
	NC-NO	PRG	1.0000	1.0000	1.0000
		EDU	0.8099	0.5416	0.6467
	MO-NO	PRG	1.0000	0.9667	0.9800
		EDU	0.8006	0.5548	0.6484
	PA-NO	PRG	0.9250	0.8333	0.8557
		EDU	0.7155	0.3003	0.4092
QS_{W5}	EC-NO	PRG	1.0000	1.0000	1.0000
		EDU	0.8769	0.6367	0.7332
	NC-NO	PRG	1.0000	1.0000	1.0000
		EDU	0.8137	0.4412	0.5697
	MO-NO	PRG	1.0000	0.9667	0.9800
		EDU	0.7963	0.4578	0.5726
	PA-NO	PRG	0.9250	0.8000	0.8357
		EDU	0.7331	0.2473	0.3514

ผลการทดลองข้างต้นนี้ไม่เป็นไปตามสมมติฐานของการวิจัยข้อที่ 2 ที่ได้ตั้งไว้ในตอนต้นว่า ระบบที่ใช้ข้อมูลรับเข้าที่เป็นหน่วยปริจเฉทพื้นฐานจะสามารถตรวจหาการลักลอกแบบคัดลอกโดยตรง การลักลอกแบบคัดลอกโดยใกล้เคียง และการลักลอกแบบคัดลอกโดยดัดแปลง ได้ดีกว่าระบบที่ใช้ข้อมูลรับเข้าที่เป็นย่อหน้า ในขณะที่ระบบที่ใช้ข้อมูลรับเข้าที่เป็นย่อหน้าสามารถตรวจหาการลักลอกแบบถอดความได้ดีกว่าระบบที่ใช้ข้อมูลรับเข้าที่เป็นหน่วยปริจเฉทพื้นฐาน เพราะผลการทดลองได้ชี้ให้เห็นว่าข้อมูลรับเข้าที่เป็นย่อหน้าเอื้อต่อการตรวจหาการลักลอกทุกประเภท ไม่ว่าจะเป็นการลักลอกแบบคัดลอกโดยตรง การลักลอกแบบคัดลอกโดยใกล้เคียง และการลักลอกแบบคัดลอกโดยดัดแปลง หรือการลักลอกแบบถอดความ ทั้งนี้ สาเหตุที่ทำให้ข้อมูลรับเข้าประเภทย่อหน้าให้ประสิทธิภาพในการตรวจหาที่ดีกว่านั้นก็เนื่องมาจากข้อจำกัดทั้ง 3 ประการของข้อมูลรับเข้าที่เป็นหน่วยปริจเฉทพื้นฐานที่ได้กล่าวไปแล้วในตอนต้น ในทางตรงกันข้าม ข้อมูลรับเข้าประเภทย่อหน้าก็มีข้อได้เปรียบในด้านขนาดความยาวที่มากกว่า ซึ่งเอื้อให้เครื่องเรียนรู้ความแตกต่างของข้อความในคู่หน่วยเทียบได้ชัดเจนกว่าคู่หน่วยเทียบที่เป็นคู่ของหน่วยปริจเฉทพื้นฐาน

อย่างไรก็ดี ในประเด็นที่เกี่ยวข้องกับการออกแบบการทดลองข้างต้น อาจมีความเห็นแย้งว่าในทดลองนี้ใช้ลักษณ์เพียง 5 ลักษณ์ และเป็นลักษณ์ที่ประยุกต์ใช้แนวคิดเรื่องค่าความละม้ายของข้อความในระดับคำ หากเปลี่ยนไปใช้ลักษณ์ชนิดอื่น ประสิทธิภาพของจำแนกของระบบที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้าอาจเพิ่มขึ้นหรือดีกว่าระบบที่ใช้ย่อหน้าเป็นข้อมูลรับเข้า ในประเด็นนี้ ผู้วิจัยเห็นว่าอาจเป็นไปได้ที่ลักษณ์ชนิดอื่นจะเอื้อให้ระบบที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้ามีประสิทธิภาพในการจำแนกสูงขึ้น แต่ลักษณ์ดังกล่าวอาจต้องเป็นลักษณ์ที่แทนรูปข้อความในระดับความหมายหรือระดับความสัมพันธ์ทางปริจเฉท ทั้งนี้ ลักษณ์ที่วิเคราะห์หาและสร้างขึ้นในงานวิจัยชิ้นนี้ต่างก็มีหน่วยพื้นฐานในการวิเคราะห์อยู่ในระดับคำ แม้จะเป็นลักษณ์ทางวากยสัมพันธ์หรือลักษณ์ทางความหมาย ลักษณ์ทั้งสองกลุ่มนี้ก็เกิดจากการวิเคราะห์หาและแทนรูปลักษณ์โดยพิจารณาความสัมพันธ์ทางวายสัมพันธ์และความหมายในหน่วยทางภาษาระดับคำก่อน นอกจากนี้ ผลจากการวิเคราะห์กลวิธีการลักลอกงานวิชาการภาษาไทยยังชี้ให้เห็นว่าผู้ลักลอกมีแนวโน้มแก้ไขข้อความในระดับคำก่อนเป็นลำดับแรก ด้วยเหตุนี้ ผู้วิจัยจึงเห็นว่าการใช้ค่าความละม้ายโซเรนเซน-ไดซ์จากเอ็นแกรมของคำทั้ง 5 ชนิดเป็นลักษณ์ในการทดลองจำแนกประเภท มีความเหมาะสม และสามารถเป็นตัวแทนของการใช้ลักษณ์ชนิดอื่นๆ ได้อย่างครอบคลุมแล้ว

ด้วยผลการทดลองที่ได้กล่าวทั้งหมดในหัวข้อนี้ จึงสามารถยืนยันได้ว่าในการพัฒนาระบบตรวจหาการลักลอกงานวิชาการ ประเภทของข้อมูลรับเข้าที่เหมาะสมจะใช้ในระบบคือย่อหน้า

6.2 ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน

ลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกถือเป็นปัจจัยสำคัญอีกประการหนึ่งที่ส่งผลต่อประสิทธิภาพของระบบตรวจหาการลักลอกงานวิชาการที่พัฒนาขึ้นในงานวิจัยชิ้นนี้ ในหัวข้อนี้ ผู้วิจัยจะได้นำเสนอผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่วิเคราะห์หาและสร้างได้ดังมีรายละเอียดปรากฏในบทที่ 6 แล้ว

เพื่อประเมินว่าลักษณ์ชนิดใดเอื้อต่อประสิทธิภาพในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอก เหมาะสมจะนำมาใช้ในการพัฒนาระบบตรวจหาการลักลอกงานวิชาการ ผู้วิจัยจึงได้แบ่งการทดลองออกเป็น 3 ขั้น ได้แก่ การประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์เป็นรายลักษณ์ การประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์แบบรวมชุด และการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ทางภาษา โดยมีข้อมูลที่ใช้ทดสอบเป็นคู่หน่วยเทียบของย่อหน้าทั้งหมดในคลังข้อมูลจำนวน 50,000 คู่หน่วยเทียบ ตามวิธีการทดลองที่ได้กล่าวไว้แล้วในหัวข้อที่ 3.5.4 ทั้งนี้ ผลการประเมินประสิทธิภาพในการทดลองแต่ละขั้นมีดังต่อไปนี้

6.2.1 ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์เป็นรายลักษณ์

การทดลองในขั้นนี้มีวัตถุประสงค์เพื่อประเมินประสิทธิภาพของลักษณ์ที่มีผลในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกแต่ละลักษณ์เป็นรายลักษณ์ทั้งลักษณ์ทางภาษาและลักษณ์อิงอักขระ

ในการทดลองขั้นนี้ ผู้วิจัยจะนำลักษณ์ที่วิเคราะห์และสร้างได้สำเร็จรวมจำนวนทั้งหมด 70 ลักษณ์ ตามรายละเอียดที่ได้กล่าวไว้บทที่ 5 มาฝึกฝนและทดสอบประสิทธิภาพการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกทั้งหมดในคลังข้อมูลด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน โดยแยกทดสอบเป็นรายลักษณ์ ตามวิธีการที่ได้กล่าวไปแล้วในหัวข้อที่ 3.5.2

ตารางที่ 6.3 แสดงผลการประเมินประสิทธิภาพของระบบในการจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอกเมื่อใช้ลักษณ์เป็นรายลักษณ์ เรียงตามลำดับจากประสิทธิภาพสูงไปหาประสิทธิภาพสูงต่ำ เมื่อพิจารณาค่า F จากตารางดังกล่าวแล้ว จะเห็นว่าลักษณ์ที่เอื้อให้ระบบมีประสิทธิภาพสูงสุด 10 อันดับแรก ประกอบด้วยลักษณ์ทางศัพท์ ซึ่งจัดเป็นลักษณ์ทางภาษาใน 6 อันดับแรก โดยลักษณ์ที่ให้ประสิทธิภาพสูงที่สุดคือลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) รองลงมาคือลักษณ์ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของคำ (J_{W2}) และลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของคำ (QS_{W3}) ตามลำดับ ส่วน 4 อันดับที่เหลือเป็นลักษณ์อิงอักขระ

อย่างไรก็ดี เมื่อพิจารณาลักษณ์ที่ให้ประสิทธิภาพสูงในอันดับที่ 7 และ 8 ซึ่งได้แก่ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของอักขระ (QS_{Char4}) และลักษณ์ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของอักขระ (J_{Char4}) ตามลำดับแล้ว จะสังเกตได้ว่าลักษณ์ทั้งสองอยู่ในอันดับรองจากลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ (QS_{W1}) และค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของคำ (J_{W1}) ในประเด็นนี้ ผู้วิจัยเห็นว่าเป็นเพราะอักขระ 4 ตัวนั้นมีขนาดใกล้เคียงกับขนาดของคำ 1 คำ แต่ด้วยขนาด 4 แกรมของอักขระดังกล่าวอาจมากกว่าหรือน้อยกว่าขนาดยูนิแกรมของคำที่ได้รับการกำกับขอบเขตของคำตามจริง จึงทำให้ลักษณ์ที่สร้างจาก 4 แกรมของอักขระให้ประสิทธิภาพเป็นรองลักษณ์ที่สร้างจากยูนิแกรมของอักขระ

ดังนั้น หากอาศัยแนวคิดที่กล่าวไปข้างต้น ก็อาจหาข้อสรุปในเบื้องต้นได้ว่าลักษณ์ค่าความละม้ายที่วิเคราะห์หาและสร้างโดยอิงแนวคิดเรื่องเอ็นแกรมของคำ โดยเฉพาะ ไบแกรมและไตรแกรมจะให้ประสิทธิภาพในจำแนกประเภทของที่มีการลักลอกและไม่มีการลักลอกได้ดีที่สุด ในขณะเดียวกันก็อาจสรุปได้ด้วยว่าลักษณ์ที่วิเคราะห์หาและสร้างโดยอิงรูปคำโดยตรงย่อมให้ประสิทธิภาพดีเช่นกัน

ตารางที่ 6.3 ผลการประเมินประสิทธิภาพของระบบในการจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอกเมื่อใช้ลักษณ์เป็นรายลักษณ์ เรียงตามลำดับจากประสิทธิภาพสูงไปหาประสิทธิภาพสูงต่ำ

ที่	ประเภท	ลักษณ์	สัญลักษณ์	ค่าการประเมิน		
				Prec.	Rec.	F
1	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ	QS_{W2}	0.9998	0.9766	0.9870
2	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของคำ	J_{W2}	0.9998	0.9760	0.9867
3	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของคำ	QS_{W3}	1.0000	0.9758	0.9865
4	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของคำ	J_{W3}	1.0000	0.9752	0.9861
5	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ	QS_{W1}	0.9959	0.9770	0.9859
6	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของคำ	J_{W1}	0.9961	0.9766	0.9858
7	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของอักขระ	QS_{Char4}	0.9965	0.9755	0.9849
8	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของอักขระ	J_{Char4}	0.9967	0.9750	0.9847
9	อักขระ	ค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ	$lcs_{norm\text{-}W}$	0.9898	0.9806	0.9846
10	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของอักขระ	QS_{Char5}	0.9974	0.9744	0.9846
11	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของอักขระ	QS_{Char3}	0.9954	0.9754	0.9844
12	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของอักขระ	J_{Char3}	0.9958	0.9750	0.9844

ที่	ประเภท	ลักษณ์	สัญลักษณ์	ค่าการประเมิน		
				Prec.	Rec.	F
13	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของอักขระ	J_{Char5}	0.9976	0.9739	0.9844
14	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของอักขระ	J_{Char2}	0.9923	0.9724	0.9814
15	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของอักขระ	QS_{Char2}	0.9913	0.9732	0.9813
16	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของคำ	QS_{W4}	1.0000	0.9618	0.9775
17	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของคำ	J_{W4}	1.0000	0.9614	0.9773
18	อักขระ	ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ	$lcs_{norm\text{-}Char}$	0.9834	0.9725	0.9772
19	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของคำ	QS_{W5}	0.9929	0.9500	0.9660
20	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของคำ	J_{W5}	0.9929	0.9492	0.9654
21	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักไตรแกรมของคำแบบ tf–idf	$\cos_{tf\text{-}idf\text{-}3}$	0.9781	0.9545	0.9619
22	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักไบแกรมคำของแบบ tf–idf	$\cos_{tf\text{-}idf\text{-}2}$	0.9798	0.9503	0.9609
23	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf	$\cos_{tf\text{-}idf\text{-}123}$	0.9811	0.9486	0.9603
24	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนัก 4 แกรมของคำแบบ tf–idf	$\cos_{tf\text{-}idf\text{-}4}$	0.9769	0.9507	0.9585
25	อักขระ	ค่าความละม้ายโคไซน์ของไบแกรมของอักขระ	$\cos_{Char2}$	0.9617	0.9542	0.9552
26	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักยูนิแกรมของคำแบบ tf–idf	$\cos_{tf\text{-}idf\text{-}1}$	0.9689	0.9462	0.9549

ที่	ประเภท	ลักษณ์	สัญลักษณ์	ค่าการประเมิน		
				Prec.	Rec.	F
27	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนัก 5 แกรมของคำแบบ tf–idf	$\cos_{\text{tf-idf-5}}$	0.9770	0.9444	0.9540
28	อักขระ	ค่าความละม้ายโคไซน์ของ 4 แกรมของอักขระ	$\cos_{\text{Char4}}$	0.9662	0.9474	0.9528
29	อักขระ	ค่าความละม้ายโคไซน์ของไตรแกรมของอักขระ	$\cos_{\text{Char3}}$	0.9630	0.9487	0.9522
30	อักขระ	ค่าความละม้ายโคไซน์ของยูนิแกรมของอักขระ	$\cos_{\text{Char1}}$	0.9589	0.9500	0.9518
31	อักขระ	ค่าความละม้ายโคไซน์ของ 5 แกรมของอักขระ	$\cos_{\text{Char5}}$	0.9677	0.9442	0.9513
32	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของคำ	$\cos_{\text{W123}}$	0.9656	0.9370	0.9463
33	ศัพท์	ค่าความละม้ายโคไซน์ของไตรแกรมของคำ	$\cos_{\text{W3}}$	0.9760	0.9298	0.9459
34	ศัพท์	ค่าความละม้ายโคไซน์ของ 4 แกรมของคำ	$\cos_{\text{W4}}$	0.9765	0.9294	0.9454
35	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของอักขระ	QS_{Char1}	0.9806	0.9198	0.9452
36	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของอักขระ	J_{Char1}	0.9815	0.9191	0.9451
37	ศัพท์	ค่าความละม้ายโคไซน์ของไบแกรมของคำ	$\cos_{\text{W2}}$	0.9660	0.9329	0.9445
38	อักขระ	ค่าระยะการแก้ไขเลเวนชเตย์นของอักขระ	LD_{char}	0.9643	0.9269	0.9423
39	ศัพท์	ค่าระยะการแก้ไขเลเวนชเตย์นของคำ	LD_{W}	0.9669	0.9250	0.9422
40	ศัพท์	ค่าความละม้ายโคไซน์ของ 5 แกรมของคำ	$\cos_{\text{W5}}$	0.9747	0.9242	0.9406

ที่	ประเภท	ลักษณ์	สัญลักษณ์	ค่าการประเมิน		
				Prec.	Rec.	F
41	ศัพท์	ค่าความละม้ายโคไซน์ของยูนิแกรมของคำ	COS_{W1}	0.9490	0.9389	0.9400
42	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของหมวดคำ	QS_{POS5}	0.9706	0.9225	0.9391
43	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของหมวดคำ	J_{POS5}	0.9743	0.9187	0.9386
44	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของหมวดคำ	J_{POS4}	0.9714	0.9166	0.9353
45	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของหมวดคำ	QS_{POS4}	0.9672	0.9197	0.9350
46	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของหมวดคำ	QS_{POS3}	0.9637	0.9116	0.9287
47	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของหมวดคำ	J_{POS3}	0.9664	0.9093	0.9284
48	วากยสัมพันธ์	ค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของหมวดคำ	$lcs_{norm\text{-}POS}$	0.9278	0.9143	0.9186
49	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของหมวดคำ	QS_{POS2}	0.9555	0.8942	0.9154
50	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของหมวดคำ	J_{POS2}	0.9586	0.8902	0.9144
51	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของ 5 แกรมของหมวดคำ	COS_{POS5}	0.9425	0.8729	0.8939
52	วากยสัมพันธ์	ค่าระยะการแก้ไขเลเวนชเตย์นของหมวดคำ	LD_{POS}	0.9333	0.8594	0.8862
53	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของหมวดคำ	COS_{POS123}	0.9320	0.8550	0.8807
54	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของ 4 แกรมของหมวดคำ	COS_{POS4}	0.9324	0.8574	0.8807

ที่	ประเภท	ลักษณ์	สัญลักษณ์	ค่าการประเมิน		
				Prec.	Rec.	F
55	วากยสัมพันธ์ +ความหมาย	ค่าความละม้ายของเวกเตอร์ทางความหมายและลำดับคำ	sim_{sem+wo}	0.9003	0.8770	0.8800
56	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของไตรแกรมของหมวดคำ	cos_{POS3}	0.9244	0.8446	0.8694
57	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของไบแกรมของหมวดคำ	cos_{POS2}	0.9083	0.8406	0.8606
58	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของยูนิแกรมของหมวดคำ	QS_{POS1}	0.9319	0.8151	0.8521
59	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของหมวดคำ	J_{POS1}	0.9353	0.8110	0.8507
60	ความหมาย	ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง	cos_{LSA}	0.8253	0.8782	0.8400
61	วากยสัมพันธ์	ค่าความละม้ายของลำดับคำ	sim_{WO}	0.8921	0.6932	0.7593
62	ศัพท์	ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ	$len(lcs_{W})$	0.8927	0.6202	0.6986
63	ศัพท์	ขนาดของคู่หน่วยเทียบ (คำ)	$Size_{W}$	0.7153	0.6946	0.6893
64	ศัพท์	ผลต่างของขนาดของคู่หน่วยเทียบ (คำ)	$diff_{W}$	0.7315	0.6640	0.6360
65	อักขระ	ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของอักขระ	$len(lcs_{Char})$	0.8578	0.5104	0.5909
66	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของลำดับคำ	cos_{WO}	0.8793	0.4744	0.5562
67	อักขระ	ขนาดของคู่หน่วยเทียบ (อักขระ)	$Size_{Char}$	0.5885	0.4315	0.4700

ที่	ประเภท	ลักษณ์	สัญลักษณ์	ค่าการประเมิน		
				Prec.	Rec.	F
68	วากยสัมพันธ์	ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของหมวดคำ	$\text{len}(lcs_{POS})$	0.8692	0.3474	0.4527
69	อักขระ	ผลต่างของขนาดของคู่หน่วยเทียบ (อักขระ)	diff_{Char}	0.6497	0.3902	0.4265
70	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของยูนิแกรมของหมวดคำ	$\cos_{POS1}$	0.0000	0.0000	0.0000

เมื่อพิจารณาประสิทธิภาพของระบบเมื่อใช้ลักษณ์ในการทดสอบตามประเภทของลักษณ์ จะพบว่าในภาพรวมแล้วลักษณ์ทางศัพท์และลักษณ์อิงอักขระ ซึ่งเป็นลักษณ์ที่อาศัยแนวคิดด้านรูปคำและรูปอักขระในการสร้าง จะให้ประสิทธิภาพในการจำแนกประเภทข้อความลักลอกและไม่ลักลอกสูงกว่าลักษณ์ที่อาศัยแนวคิดทางภาษาศาสตร์ในระดับสูงกว่าในการสร้างอย่างลักษณ์ทางวากยสัมพันธ์และลักษณ์ทางความหมาย ลักษณะดังกล่าวนี้อาจเป็นผลจากคลังข้อมูลที่สร้างขึ้นในงานชิ้นนี้สร้างโดยอาศัยการแก้ไขในระดับคำและวลีเป็นส่วนใหญ่ ในขณะที่การแก้ไขในระดับความหมายอย่างการถอดความปรากฏเป็นข้อมูลเพียงร้อยละ 10 ของข้อมูลการลักลอกทั้งหมด จึงส่งผลให้ลักษณ์ทางวากยสัมพันธ์และลักษณ์ทางความหมายให้ค่า F ในระดับที่ต่ำกว่าลักษณ์ทางศัพท์และลักษณ์อิงอักขระ

เพื่อพิสูจน์ว่าข้อสังเกตดังกล่าวข้างต้นเป็นจริงหรือไม่ ผู้วิจัยจึงเลือกลักษณ์ที่อาศัยแนวคิดทางภาษาศาสตร์ในระดับสูงในการสร้าง จำนวน 2 ลักษณ์ ได้แก่ ลักษณ์ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง (cos_{LSA}) และลักษณ์ค่าความละม้ายของเวกเตอร์ทางความหมายและลำดับคำ (sim_{sem+wo}) และลักษณ์ทางศัพท์ จำนวน 1 ลักษณ์ คือลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) มาทดลองเปรียบเทียบประสิทธิภาพของระบบในการจำแนกข้อมูลการลักลอกเป็นรายประเภท ได้แก่ ข้อมูลลักลอกประเภทคัดลอกโดยตรงและข้อมูลที่ไม่มีการลักลอก (EC-NO), ข้อมูลลักลอกประเภทคัดลอกโดยใกล้เคียงและข้อมูลที่ไม่มีการลักลอก (NC-NO), ข้อมูลลักลอกประเภทคัดลอกโดยดัดแปลงและข้อมูลที่ไม่มีการลักลอก (MO-NO), และข้อมูลลักลอกประเภทถอดความและข้อมูลที่ไม่มีการลักลอก (PA-NO) ทั้งนี้ หากข้อสังเกตดังกล่าวข้างต้นเป็นจริง ค่า F ที่ได้จากการจำแนกข้อมูลลักลอกประเภทถอดความและข้อมูลที่ไม่มีการลักลอก (PA-NO) จะต้องสูงกว่าค่า F ที่ได้จากการจำแนกข้อมูลทั้งหมดในคลังข้อมูลดังปรากฏในตารางที่ 6.3 เนื่องจากลักษณ์ที่อาศัยแนวคิดทางภาษาศาสตร์ในระดับสูงในการสร้างอาจมีประสิทธิภาพสูงเมื่อตรวจหาการลักลอกในระดับความหมาย

ตารางที่ 6.4 แสดงผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ทั้ง 3 ประเภทข้างต้นในการจำแนกประเภทข้อมูลลักลอกเฉพาะประเภทและข้อมูลที่ไม่มีการลักลอก เมื่อพิจารณาค่า F จากตารางดังกล่าว จะเห็นได้ว่าในส่วนลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) นั้น ค่า F มีแนวโน้มค่อยๆ ลดลงเมื่อจำแนกประเภทข้อความลักลอกที่มีการแก้ไขในระดับที่สูงขึ้น ลักษณะดังกล่าวนี้สะท้อนธรรมชาติของข้อมูลในคลังข้อมูลตามที่ได้ออกแบบไว้ในขั้นตอนการสร้างคลังข้อมูล ซึ่งตั้งใจให้ข้อมูลในคลังข้อมูลมีระดับความยากง่ายในการตรวจหาที่แตกต่างกัน ทั้งนี้ ผู้วิจัยเห็นว่าลักษณะการลดลงของค่า F เช่นนี้เป็นลักษณะร่วมที่ปรากฏร่วมกันในลักษณ์ทุกลักษณ์ที่ผ่านการทดสอบจำแนกประเภทข้อมูลลักลอกด้วยคลังข้อมูลนี้ ลักษณะดังกล่าวยัง

สอดคล้องกับค่า F ของลักษณ์ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแฝง (cos_{LSA}) และลักษณ์ค่าความละม้ายของเวกเตอร์ทางความหมายและลำดับคำ (sim_{sem+wo}) ที่ลดลงตามลำดับเช่นกัน ด้วยลักษณะดังกล่าวนี้จึงพิสูจน์ได้ว่าลักษณ์ที่อาศัยแนวคิดทางภาษาศาสตร์ในระดับสูงกว่าในการสร้างไม่ได้ให้ประสิทธิภาพในการจำแนกประเภทที่ดีกว่าลักษณ์ที่อาศัยแนวคิดด้านรูปคำและรูปอักขระในการสร้าง แม้ในข้อมูลที่มีการลักลอกในระดับวากยสัมพันธ์หรือความหมายก็ตาม

ตารางที่ 6.4 ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ 3 ประเภทในการจำแนกประเภทข้อมูลลักลอกเฉพาะประเภทและข้อมูลที่ไม่มีการลักลอก

ลักษณ์	ประเภทข้อมูล	ค่าการประเมิน		
		Prec.	Rec.	F
QS_{W2}	EC-NO	1.0000	1.0000	1.0000
	NC-NO	0.9989	1.0000	0.9995
	MO-NO	0.9895	0.9568	0.9726
	PA-NO	0.9538	0.8400	0.8908
cos_{LSA}	EC-NO	0.8299	1.0000	0.8978
	NC-NO	0.7620	0.8763	0.7997
	MO-NO	0.0000	0.0000	0.0000
	PA-NO	0.0000	0.0000	0.0000
sim_{sem+wo}	EC-NO	1.0000	1.0000	1.0000
	NC-NO	0.9072	0.8963	0.8995
	MO-NO	0.8325	0.6651	0.7259
	PA-NO	0.0000	0.0000	0.0000

อย่างไรก็ตาม จากตารางที่ 6.4 ข้างต้น จะเห็นได้ว่ามีค่าความแม่นยำ ค่าความครบถ้วน และค่า F ของบางกรณีที่เท่ากับ 0 ลักษณะดังกล่าวนี้เป็นผลเนื่องมาจากวิธีการคำนวณค่าดังกล่าว กล่าวคือ ในการคำนวณค่าความแม่นยำและค่าความครบถ้วน จะใช้ผลบวกจริง (true positive: tp) เป็นตัวตั้งแล้วหารด้วยผลรวมของผลบวกจริง (true positive: tp) กับผลบวกปลอม (false positive: fp) และผลรวมของผลบวกจริง (true positive: tp) กับผลลบปลอม (false negative: fn) ตามลำดับ ในกรณีที่เครื่องจำแนกประเภทกรณีลักลอกไม่ตรงกับคำตอบที่ให้โดยมนุษย์เลยเช่นในกรณีนี้ ผลบวกจริง (true positive: tp) จะเท่ากับ 0 ส่งผลให้เมื่อนำมาหารแล้ว ค่าความแม่นยำและค่าความครบถ้วนจึงมีค่าเท่ากับ 0 ตามไปด้วย อีกทั้งยังส่งต่อไปให้ค่า F ซึ่งเป็นค่าเฉลี่ยฮาร์มอนิกของค่าความ

แม่นยำและค่าความครบถ้วนมีค่าเป็น 0 ตามไปด้วยเช่นกัน คำอธิบายนี้ยังใช้อธิบายผลการประเมินประสิทธิภาพของค่าความละม้ายโคไซน์ของยูนิแกรมของหมวดคำ (COS_{POS1}) ที่ปรากฏอยู่ในอันดับที่ 70 ในตารางที่ 6.1 ได้ด้วย

6.2.2 ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์แบบรวมชุด

การทดลองในขั้นนี้มีวัตถุประสงค์เพื่อวิเคราะห์ว่าควรจัดชุดรวมของลักษณ์โดยประกอบด้วยลักษณ์ใดบ้างจึงจะส่งผลให้แบบจำลองสามารถจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกได้ผลดีที่สุด

ในการทดลองขั้นตอนนี้ ผู้วิจัยจะนำลักษณ์ที่ให้ประสิทธิภาพดีที่สุด 10 อันดับแรกจากการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์เป็นรายลักษณ์ในหัวข้อที่ 6.2.1 มารวมเป็นชุดของลักษณ์และนำฝึกฝนและทดสอบประสิทธิภาพด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนตามขั้นตอนในหัวข้อที่ 3.5.2 จากนั้นจะทดลองรวมชุดของลักษณ์ใหม่โดยตัดลักษณ์ที่มีประสิทธิภาพอยู่ในอันดับที่ต่ำที่สุดออกไป 1 ลักษณ์และนำฝึกฝนและทดสอบประสิทธิภาพด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนอีกรอบ ผู้วิจัยจะจัดชุดของลักษณ์ใหม่และนำไปฝึกฝนและทดสอบเช่นนี้จนกระทั่งเหลือลักษณ์ที่ให้ประสิทธิภาพอยู่ในอันดับที่สูงที่สุดเพียงลักษณ์เดียว จากวิธีการดังกล่าวนี้จะทำให้ได้ชุดของลักษณ์ทั้งหมด 10 ชุดที่ผ่านการฝึกและทดสอบด้วยแบบจำลอง

ตารางที่ 6.5 แสดงผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์แบบรวมชุดในการจำแนกประเภทข้อมูลลักลอกเฉพาะประเภทและข้อมูลที่ไม่มีการลักลอก โดยในเบื้องต้น ผู้วิจัยได้ทดลองสลับลำดับของลักษณ์ภายในชุดของลักษณ์แต่ละชุดแล้ว ปรากฏว่าไม่มีผลทำให้ค่าการประเมินทั้ง 3 ค่าเปลี่ยนแปลงไป

เมื่อพิจารณาค่า F ในตารางที่ 6.5 แล้ว จะเห็นได้ว่าชุดของลักษณ์ที่ให้ประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและข้อความที่ไม่มีการลักลอกดีที่สุดคือชุดที่ 10 ซึ่งประกอบด้วยลักษณ์เพียงลักษณ์เดียวคือลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) สาเหตุที่เป็นเช่นนี้ ผู้วิจัยเห็นว่าเป็นเพราะลักษณ์อื่นๆ ที่นำมารวมชุดกับลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำล้วนแล้วแต่มีประสิทธิภาพในการจำแนกประเภทที่ด้อยกว่าลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำเป็นเหตุให้ประสิทธิภาพโดยรวมของชุดของลักษณ์ต่ำลง

ตารางที่ 6.5 ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์แบบรวมชุดในการจำแนกประเภทข้อความที่มีการลักลอกและข้อความที่ไม่มีการลักลอก

ชุดที่	ชุดของลักษณ์	ค่าการประเมิน		
		Prec.	Rec.	F
1	QS_{W2}, J_{W2}, QS_{W3}, J_{W3}, QS_{W1}, J_{W1}, QS_{Char4}, J_{Char4}, $lcs_{norm\text{-}W}$, QS_{Char5}	0.9981	0.9775	0.9866
2	QS_{W2}, J_{W2}, QS_{W3}, J_{W3}, QS_{W1}, J_{W1}, QS_{Char4}, J_{Char4}, $lcs_{norm\text{-}W}$	0.9981	0.9774	0.9866
3	QS_{W2}, J_{W2}, QS_{W3}, J_{W3}, QS_{W1}, J_{W1}, QS_{Char4}, J_{Char4}	0.9998	0.9738	0.9853
4	QS_{W2}, J_{W2}, QS_{W3}, J_{W3}, QS_{W1}, J_{W1}, QS_{Char4}	0.9998	0.9738	0.9853
5	QS_{W2}, J_{W2}, QS_{W3}, J_{W3}, QS_{W1}, J_{W1}	0.9999	0.9739	0.9854
6	QS_{W2}, J_{W2}, QS_{W3}, J_{W3}, QS_{W1}	1.0000	0.9745	0.9858
7	QS_{W2}, J_{W2}, QS_{W3}, J_{W3}	1.0000	0.9755	0.9864
8	QS_{W2}, J_{W2}, QS_{W3}	1.0000	0.9758	0.9865
9	QS_{W2}, J_{W2}	0.9998	0.9763	0.9868
10	QS_{W2}	0.9998	0.9766	0.9870

ด้วยเหตุที่กล่าวมาข้างต้นนี้ จึงสรุปได้ว่าในการสร้างระบบตรวจหาการลักลอกงานวิชาการอาจไม่มีความจำเป็นต้องรวมลักษณ์เป็นชุดเพื่อใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและข้อความที่ไม่มีการลักลอก เนื่องจากจะมีผลให้ประสิทธิภาพของลักษณ์ที่ดีที่สุดต่ำลง

6.2.3 ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ทางภาษา

การประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกันในขั้นสุดท้ายนี้มีวัตถุประสงค์เพื่อประเมินประสิทธิภาพในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกเมื่อใช้ลักษณ์ทางภาษา ด้วยงานวิจัยชิ้นนี้ตั้งสมมติฐานในตอนต้นไว้ว่าลักษณ์ที่ประยุกต์ใช้ความรู้ทางภาษาในการสร้างจะให้ประสิทธิภาพในการตรวจหาการลักลอกได้ดีกว่าลักษณ์ที่ไม่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ ด้วยเหตุนี้ ผู้วิจัยจึงต้องการทดลองรวมชุดของลักษณ์ทางภาษาและนำไปฝึกฝนและทดสอบด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

ในการรวมชุดของลักษณ์ทางภาษานั้น ผู้วิจัยจะใช้ลักษณ์ 3 ลักษณ์ตามระดับของหน่วยภาษา 3 กลุ่ม ได้แก่ ลักษณ์ทางศัพท์ ลักษณ์ทางวายสัมพันธ์ และลักษณ์ทางความหมาย โดยจะนำลักษณ์ที่

ให้ประสิทธิภาพสูงที่สุดของแต่ละกลุ่มจากการทดลองประเมินประสิทธิภพรายลักษณ์ในหัวข้อที่ 6.2.1 กลุ่มละ 1 ลักษณ์ มาจัดเป็นชุดทั้งหมด 4 ชุด ได้แก่

1) ลักษณ์ทางภาษาชุดที่ 1 (LF_1) ประกอบด้วยลักษณ์ทางศัพท์และลักษณ์ทางวายสัมพันธ์ ได้แก่ ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) และลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf ($\cos_{tf\text{-}idf\text{-}123}$) ตามลำดับ
2) ลักษณ์ทางภาษาชุดที่ 2 (LF_2) ประกอบด้วยลักษณ์ทางศัพท์และลักษณ์ทางความหมาย ได้แก่ ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) และลักษณ์ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง ($\cos_{LSA}$) ตามลำดับ
3) ลักษณ์ทางภาษาชุดที่ 3 (LF_3) ประกอบด้วยลักษณ์ทางวายสัมพันธ์และลักษณ์ทางความหมาย ได้แก่ ลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf ($\cos_{tf\text{-}idf\text{-}123}$) และลักษณ์ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง ($\cos_{LSA}$) ตามลำดับ
4) ลักษณ์ทางภาษาชุดที่ 4 (LF_4) ประกอบด้วยลักษณ์ทางศัพท์ ลักษณ์ทางวายสัมพันธ์ และลักษณ์ทางความหมาย ได้แก่ ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) ลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf ($\cos_{tf\text{-}idf\text{-}123}$) และลักษณ์ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง ($\cos_{LSA}$) ตามลำดับ

ตารางที่ 6.6 ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ทางภาษาแบบรวมชุดในการจำแนกประเภทข้อความที่มีการลักลอกและข้อความที่ไม่มีการลักลอก

ชุด	ประเภทของลักษณ์	ชุดของลักษณ์	ค่าการประเมิน		
			Prec.	Rec.	F
LF_1	ศัพท์, วากยสัมพันธ์	QS_{W2}, $\cos_{tf\text{-}idf\text{-}123}$	0.9992	0.9750	0.9857
LF_2	ศัพท์, ความหมาย	QS_{W2}, $\cos_{LSA}$	0.9999	0.9748	0.9859
LF_3	วากยสัมพันธ์, ความหมาย	$\cos_{tf\text{-}idf\text{-}123}$, $\cos_{LSA}$	0.9798	0.9491	0.9598
LF_4	ศัพท์, วากยสัมพันธ์, ความหมาย	QS_{W2}, $\cos_{tf\text{-}idf\text{-}123}$, $\cos_{LSA}$	0.9995	0.9745	0.9855

ตารางที่ 6.6 แสดงผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ทางภาษาแบบรวมชุดในการจำแนกประเภทข้อความที่มีการลักลอกและข้อความที่ไม่มีการลักลอก เมื่อพิจารณาค่า F ในตารางดังกล่าวจะเห็นได้ว่าชุดรวมลักษณ์ทางภาษาชุดที่ 2 (LF_2) ซึ่งประกอบด้วยลักษณ์ทางศัพท์และลักษณ์ทางความหมาย ได้แก่ ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) และลักษณ์ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง (cos_{LSA}) ตามลำดับ เป็นชุดรวมลักษณ์ที่ให้ประสิทธิภาพดีที่สุดในการจำแนกประเภท กล่าวคือมีค่า F เท่ากับ 0.9859 อย่างไรก็ตาม ค่าดังกล่าวก็ยังต่ำกว่าค่า F ที่ได้จากการจำแนกประเภทโดยใช้ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) เพียงลักษณ์เดียว ซึ่งมีค่า F เท่ากับ 0.9870 ดังได้แสดงให้เห็นในการประเมินประสิทธิภาพรายลักษณ์ในตารางที่ 6.3 ด้วยเหตุนี้จึงอาจกล่าวได้ว่าลักษณ์ทางศัพท์ซึ่งเป็นลักษณ์ทางภาษาประเภทหนึ่งให้ประสิทธิภาพในการจำแนกข้อความที่มีการลักลอกและข้อความที่ไม่มีการลักลอกสูงกว่าชุดรวมของลักษณ์ทางภาษาประเภทอื่นๆ

จากข้อค้นพบที่กล่าวมาทั้งหมดในหัวข้อนี้ สามารถสรุปได้ว่าลักษณ์ทางศัพท์ซึ่งอาจถือได้ว่าลักษณ์ที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับผิวให้ประสิทธิภาพในการตรวจหาการลักลอกได้ดีกว่าลักษณ์ที่ไม่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ คือลักษณ์อิงอักขระ และลักษณ์ที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับลึก อันได้แก่ลักษณ์ทางวากยสัมพันธ์และลักษณ์ทางความ ตามลำดับ ข้อสรุปดังกล่าวนี้ไม่เป็นไปตามสมมติฐานของการวิจัยข้อที่ 3 ที่ตั้งไว้ในตอนต้น

ในแง่ที่ผลปรากฏดังได้กล่าวไปข้างต้นนั้น ผู้วิจัยเห็นว่าเป็นผลมาจากสาเหตุ 3 ประการ ได้แก่

ประการแรก คือ การประยุกต์ใช้แนวคิดเรื่องค่าความละม้ายในการวิเคราะห์และสร้างลักษณ์ ดังจะเห็นได้ว่าลักษณ์ทางศัพท์ที่ใช้ในการทดลองครั้งนี้วิเคราะห์หาและสร้างจากรูปคำที่ปรากฏโดยตรง และแทนรูปออกมาเป็นค่าความละม้ายซึ่งมีช่วงของปริมาณตั้งแต่ 0 ถึง 1 ลักษณะเช่นนี้จึงเอื้อให้เครื่องเรียนรู้และตัดสินใจได้อย่างแม่นยำ ทั้งนี้ หากเปรียบเทียบกับลักษณ์ชนิดอื่นที่ไม่ได้ประยุกต์ใช้แนวคิดเรื่องค่าความละม้าย เช่น ลักษณ์ขนาดของคู่หน่วยเทียบ ลักษณ์ผลต่างของขนาดของคู่หน่วยเทียบ หรือลักษณ์ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุด ลักษณ์เหล่านี้เป็นค่าตัวเลขดิบที่ได้จากการคำนวณ ไม่ได้อยู่ในรูปช่วงของปริมาณที่จำกัด จึงเป็นการยากที่เครื่องจะใช้ในสร้างสมมติฐานทั่วไปในการตัดสินใจจำแนกประเภท

ประการต่อมา คือ การแทนรูปลักษณ์จากรูปในระดับผิว ดังข้อค้นพบที่ปรากฏให้เห็นว่าลักษณ์ทางศัพท์และลักษณ์ทางอักขระให้ประสิทธิภาพในการจำแนกข้อความลักลอกและข้อความที่ไม่มีการลักลอกดีกว่าลักษณ์ทางวากยสัมพันธ์และลักษณ์ทางความหมายซึ่งมุ่งแทนรูปความสัมพันธ์

ทางภาษาศาสตร์ในระดับลึก ทั้งนี้ เป็นไปได้ว่าลักษณ์ทางศัพท์เป็นการแทนรูปจากคำซึ่งเป็นหน่วยทางภาษาที่มีขอบเขตชัดเจน เมื่อนำมาแปลงเป็นค่าตัวเลขแล้ว ค่าที่ได้ก็แสดงมีความแตกต่างกันอย่างชัดเจน เมื่อเข้าสู่อัลกอริทึมของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนแล้ว แบบจำลองจึงสามารถจำแนกประเภทได้อย่างถูกต้องและแม่นยำ

สาเหตุประการสุดท้ายเป็นเหตุที่มาจากลักษณะของข้อมูลที่ใช้ในการทดลอง ดังได้กล่าวไปในหัวข้อที่ 6.2.1 ในส่วนการอภิปรายผลการทดสอบประสิทธิภาพของระบบเมื่อใช้ลักษณ์เป็นรายลักษณ์แล้วว่าคลังข้อมูลที่สร้างขึ้นในงานชิ้นนี้สร้างโดยอาศัยการแก้ไขในระดับคำและวลีเป็นส่วนใหญ่ ในขณะที่การแก้ไขในระดับความหมายอย่างการถอดความปรากฏเป็นข้อมูลเพียงร้อยละ 10 ของข้อมูลที่เป็นการลักลอก ด้วยเหตุนี้ ลักษณ์ทางศัพท์จึงให้ประสิทธิภาพในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกได้ดีกว่าลักษณ์ประเภทอื่นๆ อย่างเห็นได้ชัด อย่างไรก็ดี คลังข้อมูลที่สร้างขึ้นเพื่อใช้ในงานวิจัยชิ้นนี้ก็ออกแบบโดยอิงจากผลการวิเคราะห์กลวิธีการลักลอกงานวิชาการภาษาไทย เพื่อให้สอดคล้องกับสถานการณ์การลักลอกที่เกิดขึ้นจริง ด้วยเหตุนี้เอง ผู้วิจัยจึงมั่นใจว่าลักษณ์ทางศัพท์เป็นลักษณ์ประเภทที่มีประสิทธิภาพในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอก เหมาะสมจะใช้ในการพัฒนาระบบตรวจหาการลักลอกงานวิชาการต่อไป

บทที่ 7
ผลการเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความ

การเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความเป็นวัตถุประสงค์ประการสุดท้ายของงานวิจัยชิ้นนี้ ดังได้กล่าวไปในตอนต้นแล้วว่า ระบบตรวจหาการลักลอกงานวิชาการที่ออกแบบไว้ในงานวิจัยชิ้นนี้ประกอบด้วยขั้นตอนการตรวจหา 2 ขั้น ขั้นแรกเป็นการตรวจหาโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก ซึ่งมีผลการประเมินประสิทธิภาพดังได้กล่าวไปในบทที่แล้ว ส่วนอีกขั้นนั้นเป็นการวัดค่าความละม้ายของข้อความที่ระบบขั้นแรกจำแนกประเภทได้ว่าเป็นการลักลอก ด้วยระบบในขั้นแรกนั้นไม่สามารถคืนค่าเป็นตัวเลขที่บ่งชี้ปริมาณการลักลอกในเอกสารได้ จึงจำเป็นต้องวัดค่าความละม้ายระหว่างข้อความต้นฉบับและข้อความลักลอกในขั้นนี้ อย่างไรก็ตาม ยังไม่ปรากฏผลการศึกษาที่ชี้ชัดว่าวิธีการวัดค่าความละม้ายวิธีใดเป็นวิธีที่มีประสิทธิภาพมากที่สุด ด้วยเหตุนี้จึงจำเป็นต้องเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความเพื่อหาวิธีการวัดค่าความละม้ายของข้อความที่มีประสิทธิภาพใกล้เคียงกับมนุษย์มากที่สุดเพื่อนำมาใช้ในระบบ

แนวคิดสำคัญที่ผู้วิจัยใช้ในการเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความในขั้นนี้คือ การนำค่าที่ได้วิธีการวัดค่าความละม้ายแต่ละวิธีมาเปรียบเทียบกับผลการวัดค่าความละม้ายโดยมนุษย์ หรือกล่าวอีกนัยหนึ่งคือใช้ค่าความละม้ายที่ให้โดยมนุษย์เป็นบรรทัดฐาน หากวิธีการวัดค่าความละม้ายวิธีใดให้ค่าได้ใกล้เคียงกับค่าความละม้ายที่ให้โดยมนุษย์มากที่สุด จะถือว่าวิธีการวัดค่าความละม้ายวิธีดังกล่าวมีประสิทธิภาพเหมาะสมจะใช้ในระบบตรวจหาการลักลอกมากที่สุด

เพื่อให้บรรลุผลตามแนวคิดข้างต้น ในการทดลองขั้นแรก ผู้วิจัยจะให้ผู้เชี่ยวชาญด้านภาษาไทย 3 คน ซึ่งมีคุณสมบัติตรงตามที่ได้ระบุไว้ในวิธีการวิจัยหัวข้อที่ 3.6 กล่าวคือ ต้องสำเร็จการศึกษาระดับปริญญาโทขึ้นไปทางภาษาไทยหรือภาษาศาสตร์ และสอนวิชาเกี่ยวกับการใช้ภาษาไทยในระดับอุดมศึกษา เป็นผู้ระบุค่าความละม้ายของคู่หน่วยเทียบในชุดข้อมูลทดลองซึ่งประกอบไปด้วยคู่หน่วยเทียบของย่อหน้าสาขาวิชาและขนาดต่างๆ จำนวนทั้งหมด 150 คู่ โดยกำหนดให้ระบุเป็นร้อยละของความละม้ายตั้งแต่ 0 ถึง 100 เมื่อได้ร้อยละของความละม้ายดังกล่าวมาผู้วิจัยจะแปลงร้อยละดังกล่าวให้เป็นค่าความละม้ายด้วยการนำมาหารด้วย 100 เพื่อให้ค่ามีอยู่ในช่วงตั้งแต่ 0 ถึง 1 และนำค่าความละม้ายดังกล่าวไปเปรียบเทียบกับค่าความละม้ายที่เครื่องวัดได้จากวิธีต่างๆ

ในบทนี้ ผู้วิจัยจะได้นำเสนอผลเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความดังกล่าว เพื่อระบุว่าวิธีการวัดค่าความละม้ายวิธีใดมีประสิทธิภาพเหมาะสมจะนำมาใช้ในระบบ ทั้งนี้ ผู้วิจัยได้แบ่งการนำเสนอผลออกเป็น 2 หัวข้อ หัวข้อแรกว่าด้วยผลการทดสอบความเป็นเอกพันธ์ของค่าความละม้ายที่ให้โดยผู้เชี่ยวชาญ ส่วนอีกหัวข้อหนึ่งจะนำเสนอผลการวิเคราะห์ความสัมพันธ์ระหว่างค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญกับค่าความละม้ายที่เครื่องวัดได้จากวิธีการวัดค่าความละม้ายวิธีต่างๆ รายละเอียดมีดังต่อไปนี้

7.1 ผลการทดสอบความเป็นเอกพันธ์ของค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญ

การทดลองในขั้นนี้มีวัตถุประสงค์เพื่อทดสอบความเป็นเอกพันธ์ (test of homogeneity) ของค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญทั้ง 3 คน เพื่อพิสูจน์ว่าค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญทั้ง 3 คน คนละ 150 ตัวอย่างจากชุดข้อมูลทดลอง มีความแตกต่างกันทางสถิติหรือไม่

ทั้งนี้ หากพบว่าค่าความละม้ายของข้อความที่ระบุโดยผู้เชี่ยวชาญทุกคนมีความเป็นเอกพันธ์ ผู้วิจัยจะหาค่าเฉลี่ยของค่าความละม้ายที่ให้โดยผู้เชี่ยวชาญทั้ง 3 คนในแต่ละคู่หน่วยเทียบแล้วนำไปวิเคราะห์ความสัมพันธ์กับค่าความละม้ายที่ได้จากวิธีวัดค่าความละม้ายวิธีต่างๆ แต่ถ้าหากพบว่าค่าความละม้ายของที่ระบุโดยผู้เชี่ยวชาญทุกคนไม่มีความเป็นเอกพันธ์ ผู้วิจัยจะนำค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญแต่ละคนมาวิเคราะห์ความสัมพันธ์กับค่าความละม้ายที่ได้จากวิธีวัดค่าความละม้ายวิธีต่างๆ เป็นรายบุคคล

การทดสอบความเป็นเอกพันธ์ในที่นี้ ผู้วิจัยพิจารณาด้วยการวิเคราะห์ความแปรปรวนแบบทางเดียว (one-way ANOVA) เพื่อทดสอบว่าตัวแปรตาม ซึ่งในกรณีนี้คือค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญ แตกต่างกันตามตัวแปรอิสระ ซึ่งในกรณีนี้คือผู้เชี่ยวชาญแต่ละคนหรือไม่

จากการวิเคราะห์ความแปรปรวนแบบทางเดียวพบว่า ณ ระดับนัยสำคัญ $\alpha = .05$ ค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญทั้ง 3 คนไม่แตกต่างกัน ($F = .856$, $p = .426$) หรือกล่าวอีกนัยหนึ่งได้ว่าค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญทั้ง 3 คนมีความเป็นเอกพันธ์ ด้วยเหตุนี้ จึงสามารถนำค่าเฉลี่ยของค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญทั้ง 3 คนในแต่ละคู่หน่วยเทียบไปวิเคราะห์ความสัมพันธ์กับค่าความละม้ายที่เครื่องวัดได้จากวิธีต่างๆ ได้

ผลจากการทดสอบความเป็นเอกพันธ์ข้างต้นได้สะท้อนให้เห็นว่าภายใต้กรอบที่กำหนดขึ้นอย่างกว้างๆ ในการให้ระบุค่าความละม้าย ผู้เชี่ยวชาญด้านภาษาไทยทั้ง 3 คนก็สามารถระบุค่าความละม้ายได้สอดคล้องเป็นไปในทิศทางเดียวกัน ลักษณะดังกล่าวนี้ช่วยยืนยันถึงความเชื่อถือได้ (reliability) ของผู้เชี่ยวชาญทั้ง 3 คน อันจะส่งผลให้ผลการวิเคราะห์ความสัมพันธ์ระหว่างค่าความ

ละม้ายที่ระบุโดยผู้เชี่ยวชาญกับค่าความละม้ายที่เครื่องวัดได้จากวิธีวัดค่าความละม้ายวิธีต่างๆ ที่จะกล่าวถึงในหัวข้อต่อไปนั้นมีความหนักแน่นน่าเชื่อถือมากยิ่งขึ้น

7.2 ผลการวิเคราะห์ความสัมพันธ์ของค่าความละม้าย

ในหัวข้อนี้ ผู้วิจัยจะนำเสนอผลการวิเคราะห์ความสัมพันธ์ระหว่างค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญกับค่าความละม้ายที่เครื่องวัดได้จากวิธีวัดค่าความละม้ายวิธีต่างๆ

การทดลองในขั้นนี้มีวัตถุประสงค์เพื่อระบุว่าวิธีวัดค่าความละม้ายวิธีใดจากจำนวนทั้งหมด 60 วิธี สามารถวัดค่าความละม้ายของคู่หน่วยเทียบทั้งหมด จำนวน 150 คู่ ในชุดข้อมูลทดลองได้สอดคล้องเป็นไปในทิศทางเดียวกับค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญมากที่สุด ทั้งนี้ เพื่อจะได้นำวิธีการวัดค่าความละม้ายดังกล่าวมาใช้ในระบบตรวจหาการลักลอกงานวิชาการที่พัฒนาขึ้น

ในขั้นต้น ผู้วิจัยจะนำค่าความละม้ายที่วัดได้จากวิธีการวัดทั้ง 60 วิธี ซึ่งเป็นผลที่ได้มาจากขั้นตอนการวิเคราะห์หาและลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก ที่ได้กล่าวไปแล้วในบทที่ 5 มาเป็น 5 กลุ่มตามการประยุกต์ใช้ระดับของหน่วยในการคำนวณค่าความละม้าย เพื่อให้สอดคล้องกับการแนวทางการวิเคราะห์หาลักษณ์ และเพื่อให้สะดวกแก่การวิเคราะห์และอภิปรายผล ดังนี้

1) ค่าความละม้ายอิงอักขระ
2) ค่าความละม้ายอิงศัพท์
3) ค่าความละม้ายอิงวายสัมพันธ์
4) ค่าความละม้ายอิงความหมาย
5) ค่าความละม้ายอิงวายสัมพันธ์และความหมาย

ตารางที่ 7.1 แสดงรายการค่าความละม้ายที่วัดได้จากวิธีการวัดทั้ง 60 วิธีที่จะนำมาวิเคราะห์ความสัมพันธ์กับค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญ จำแนกตามระดับของหน่วยที่ประยุกต์ใช้ในการคำนวณ

ตารางที่ 7.1 รายการค่าความละม้ายที่วัดได้จากวิธีการวัดต่างๆ จำแนกตามระดับของหน่วยที่ประยุกต์ใช้ในการคำนวณ

ที่	ระดับ	ค่าความละม้าย	สัญลักษณ์
1	อักขระ	ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ	$lcs_{\text{norm-Char}}$
2	อักขระ	ค่าความละม้ายโคไซน์ของยูนิแกรมของอักขระ	$\cos_{\text{Char1}}$
3	อักขระ	ค่าความละม้ายโคไซน์ของไบแกรมของอักขระ	$\cos_{\text{Char2}}$
4	อักขระ	ค่าความละม้ายโคไซน์ของไตรแกรมของอักขระ	$\cos_{\text{Char3}}$

ที่	ระดับ	ค่าความละม้าย	สัญลักษณ์
5	อักขระ	ค่าความละม้ายโคไซน์ของ 4 แกรมของอักขระ	COS_{Char4}
6	อักขระ	ค่าความละม้ายโคไซน์ของ 5 แกรมของอักขระ	COS_{Char5}
7	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของอักขระ	J_{Char1}
8	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของอักขระ	J_{Char2}
9	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของอักขระ	J_{Char3}
10	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของอักขระ	J_{Char4}
11	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของอักขระ	J_{Char5}
12	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของอักขระ	QS_{Char1}
13	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของอักขระ	QS_{Char2}
14	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของอักขระ	QS_{Char3}
15	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของอักขระ	QS_{Char4}
16	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของอักขระ	QS_{Char5}
17	ศัพท์	ค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ	$lcs_{norm\text{-}W}$
18	ศัพท์	ค่าความละม้ายโคไซน์ของยูนิแกรมของคำ	COS_{W1}
19	ศัพท์	ค่าความละม้ายโคไซน์ของไบแกรมของคำ	COS_{W2}
20	ศัพท์	ค่าความละม้ายโคไซน์ของไตรแกรมของคำ	COS_{W3}
21	ศัพท์	ค่าความละม้ายโคไซน์ของ 4 แกรมของคำ	COS_{W4}
22	ศัพท์	ค่าความละม้ายโคไซน์ของ 5 แกรมของคำ	COS_{W5}
23	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของคำ	J_{W1}
24	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของคำ	J_{W2}

ที่	ระดับ	ค่าความละม้าย	สัญลักษณ์
25	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของคำ	J_{W3}
26	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของคำ	J_{W4}
27	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของคำ	J_{W5}
28	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ	QS_{W1}
29	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ	QS_{W2}
30	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของคำ	QS_{W3}
31	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของคำ	QS_{W4}
32	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของคำ	QS_{W5}
33	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักยูนิแกรมของคำแบบ tf-idf	$\cos_{\text{tf-idf-1}}$
34	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักไบแกรมของคำแบบ tf-idf	$\cos_{\text{tf-idf-2}}$
35	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักไตรแกรมของคำแบบ tf-idf	$\cos_{\text{tf-idf-3}}$
36	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนัก 4 แกรมของคำแบบ tf-idf	$\cos_{\text{tf-idf-4}}$
37	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนัก 5 แกรมของคำแบบ tf-idf	$\cos_{\text{tf-idf-5}}$
38	วากยสัมพันธ์	ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของหมวดคำ	$lcs_{\text{norm-POS}}$
39	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของยูนิแกรมของหมวดคำ	$\cos_{POS1}$
40	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของไบแกรมของหมวดคำ	$\cos_{POS2}$
41	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของไตรแกรมของหมวดคำ	$\cos_{POS3}$
42	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของ 4 แกรมของหมวดคำ	$\cos_{POS4}$
43	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของ 5 แกรมของหมวดคำ	$\cos_{POS5}$
44	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของหมวดคำ	J_{POS1}

ที่	ระดับ	ค่าความละม้าย	สัญลักษณ์
45	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของหมวดคำ	J_{POS2}
46	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของหมวดคำ	J_{POS3}
47	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของหมวดคำ	J_{POS4}
48	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของหมวดคำ	J_{POS5}
49	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของหมวดคำ	QS_{POS1}
50	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของหมวดคำ	QS_{POS2}
51	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของหมวดคำ	QS_{POS3}
52	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของหมวดคำ	QS_{POS4}
53	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของหมวดคำ	QS_{POS5}
54	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของคำ	cos_{W123}
55	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf	$cos_{tf\text{-}idf\text{-}123}$
56	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของหมวดคำ	cos_{POS123}
57	วากยสัมพันธ์	ค่าความละม้ายของลำดับคำ	sim_{WO}
58	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของลำดับคำ	cos_{WO}
59	ความหมาย	ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง	cos_{LSA}
60	วากยสัมพันธ์+ความหมาย	ค่าความละม้ายของเวกเตอร์ทางความหมายและลำดับคำ	sim_{sem+wo}

ในการวิเคราะห์ความสัมพันธ์ระหว่างค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญกับค่าความละม้ายที่ได้จากวิธีการวัดค่าความละม้ายวิธีต่างๆ ผู้วิจัยได้ประยุกต์ใช้การวิเคราะห์สัมประสิทธิ์

สหสัมพันธ์ (correlation coefficient) ซึ่งเป็นวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร 2 ตัวว่ามีความสอดคล้องกันในระดับใด วิธีการวิเคราะห์ดังกล่าวนี้จะให้ผลเป็นค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญกับค่าความละม้ายที่ได้จากวิธีการวัดค่าความละม้ายวิธีต่างๆ แต่ละวิธี ค่าสัมประสิทธิ์สหสัมพันธ์ที่ได้นี้จะเป็นตัวเลขตั้งแต่ -1 ถึง 1 โดยที่ค่าที่อยู่ใกล้ -1 หรือ 1 ถือว่ามีความสัมพันธ์กันมากที่สุด หากค่าเป็นจำนวนบวกหมายความว่าตัวแปรทั้งสองมีความสัมพันธ์เป็นไปในทิศทางเดียวกัน แต่หากค่าเป็นจำนวนลบหมายความว่าตัวแปรทั้งสองมีความสัมพันธ์เป็นไปในทิศทางตรงกันข้าม ส่วน 0 หมายความว่าตัวแปรทั้งสองไม่มีความสัมพันธ์กันโดยสิ้นเชิง

ตารางที่ 7.2 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญกับค่าความละม้ายที่ได้จากวิธีการวัดค่าความละม้ายวิธีต่างๆ เรียงลำดับตามความสัมพันธ์ตั้งแต่มากไปหาน้อย

ที่	ระดับ	ค่าความละม้ายจากวิธีการวัดค่าความละม้ายโดยเครื่อง	สัญลักษณ์	ค่าสัมประสิทธิ์สหสัมพันธ์ (*r*)
1	ศัพท์	ค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ	$lcs_{\text{norm-W}}$	0.9124
2	อักขระ	ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ	$lcs_{\text{norm-Char}}$	0.9112
3	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ	QS_{W1}	0.8932
4	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของอักขระ	QS_{Char3}	0.8910
5	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของอักขระ	QS_{Char4}	0.8886
6	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของอักขระ	QS_{Char5}	0.8819
7	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของอักขระ	QS_{Char2}	0.8801
8	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของคำ	J_{W1}	0.8777
9	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ	QS_{W2}	0.8732

ที่	ระดับ	ค่าความละม้ายจากวิธีการวัดค่าความละม้ายโดยเครื่อง	สัญลักษณ์	ค่าสัมประสิทธิ์สหสัมพันธ์ (*r*)
10	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของอักขระ	J_{Char2}	0.8683
11	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของอักขระ	J_{Char3}	0.8668
12	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักยูนิแกรมของคำแบบ tf-idf	$cos_{tf\text{-}idf\text{-}1}$	0.8607
13	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของอักขระ	J_{Char4}	0.8566
14	อักขระ	ค่าความละม้ายโคไซน์ของ 4 แกรมของอักขระ	cos_{Char4}	0.8558
15	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักไบแกรมของคำแบบ tf-idf	$cos_{tf\text{-}idf\text{-}2}$	0.8504
16	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของอักขระ	cos_{Char5}	0.8488
17	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ	cos_{W1}	0.8457
18	อักขระ	ค่าความละม้ายโคไซน์ของไตรแกรมของอักขระ	cos_{Char3}	0.8450
19	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของอักขระ	J_{Char5}	0.8438
20	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของคำ	cos_{W123}	0.8413
21	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf	$cos_{tf\text{-}idf\text{-}123}$	0.8384
22	ศัพท์	ค่าความละม้ายโคไซน์ของไบแกรมของคำ	cos_{W2}	0.8323
23	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของคำ	QS_{W3}	0.8315
24	วากยสัมพันธ์	ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของหมวดคำ	$lcs_{norm\text{-}POS}$	0.8283

ที่	ระดับ	ค่าความละม้ายจากวิธีการวัดค่าความละม้ายโดยเครื่อง	สัญลักษณ์	ค่าสัมประสิทธิ์สหสัมพันธ์ (*r*)
25	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของคำ	J_{W2}	0.8251
26	อักขระ	ค่าความละม้ายโคไซน์ของไบแกรมของอักขระ	cos_{Char2}	0.8198
27	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนักไตรแกรมของคำแบบ tf-idf	$cos_{tf\text{-}idf\text{-}3}$	0.8173
28	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของคำ	QS_{W4}	0.7944
29	ศัพท์	ค่าความละม้ายโคไซน์ของไตรแกรมของคำ	cos_{W3}	0.7923
30	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนัก 4 แกรมของคำแบบ tf-idf	$cos_{tf\text{-}idf\text{-}4}$	0.7815
31	วากยสัมพันธ์+ความหมาย	ค่าความละม้ายของเวกเตอร์ทางความหมายและลำดับคำ	sim_{sem+wo}	0.7801
32	อักขระ	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของอักขระ	J_{Char1}	0.7717
33	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของคำ	J_{W3}	0.7693
34	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของคำ	QS_{W5}	0.7585
35	ศัพท์	ค่าความละม้ายโคไซน์ของ 4 แกรมของคำ	cos_{W4}	0.7557
36	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของหมวดคำ	QS_{POS3}	0.7548
37	ศัพท์	ค่าความละม้ายโคไซน์ของน้ำหนัก 5 แกรมของคำแบบ tf-idf	$cos_{tf\text{-}idf\text{-}5}$	0.7496
38	ความหมาย	ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง	cos_{LSA}	0.7462
39	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของหมวดคำ	QS_{POS4}	0.7442

ที่	ระดับ	ค่าความละม้ายจากวิธีการวัดค่าความละม้ายโดยเครื่อง	สัญลักษณ์	ค่าสัมประสิทธิ์สหสัมพันธ์ (*r*)
40	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของหมวดคำ	QS_{POS5}	0.7358
41	อักขระ	ค่าความละม้ายโคไซน์ของยูนิแกรมของอักขระ	COS_{Char1}	0.7318
42	ศัพท์	ค่าความละม้ายโคไซน์ของ 5 แกรมของคำ	COS_{W5}	0.7293
43	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไตรแกรมของหมวดคำ	J_{POS3}	0.7284
44	อักขระ	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของอักขระ	QS_{Char1}	0.7264
45	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของคำ	J_{W4}	0.7238
46	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของหมวดคำ	J_{POS2}	0.7118
47	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 4 แกรมของหมวดคำ	J_{POS4}	0.7042
48	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของ 4 แกรมของหมวดคำ	COS_{POS4}	0.7025
49	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของหมวดคำ	QS_{POS2}	0.7022
50	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของ 5 แกรมของหมวดคำ	COS_{POS5}	0.7017
51	ศัพท์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของคำ	J_{W5}	0.6843
52	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของ 5 แกรมของหมวดคำ	J_{POS5}	0.6842
53	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของไตรแกรมของหมวดคำ	COS_{POS3}	0.6779
54	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของช่วงเอ็นแกรมของหมวดคำ	COS_{POS123}	0.6286

ที่	ระดับ	ค่าความละม้ายจากวิธีการวัดค่าความละม้ายโดยเครื่อง	สัญลักษณ์	ค่าสัมประสิทธิ์สหสัมพันธ์ (*r*)
55	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของยูนิแกรมของหมวดคำ	J_{POS1}	0.6149
56	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของยูนิแกรมของหมวดคำ	cos_{POS1}	0.5972
57	วากยสัมพันธ์	ค่าความละม้ายของลำดับคำ	sim_{WO}	0.5509
58	วากยสัมพันธ์	ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของหมวดคำ	QS_{POS1}	0.5499
59	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของไบแกรมของหมวดคำ	cos_{POS2}	0.5446
60	วากยสัมพันธ์	ค่าความละม้ายโคไซน์ของลำดับคำ	cos_{WO}	0.5026

ตารางที่ 7.2 แสดงค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญกับค่าความละม้ายที่ได้จากวิธีการวัดค่าความละม้ายวิธีต่างๆ เรียงลำดับตามความสัมพันธ์ตั้งแต่มากไปหาน้อย จากตารางดังกล่าวจะเห็นได้ว่าค่าความละม้ายที่มีความสัมพันธ์เป็นไปในทิศทางเดียวกับค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญมากที่สุด 3 วิธีแรก ได้แก่ ค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ ($lcs_{norm\text{-}W}$), ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ ($lcs_{norm\text{-}Char}$), และค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ (QS_{W1}) ซึ่งมีค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.9124, 0.9112, และ 0.8932 ตามลำดับ

ด้วยเหตุนี้ จึงสามารถสรุปได้ว่าวิธีการวัดค่าความละม้ายด้วยการหาค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ ($lcs_{norm\text{-}W}$) เป็นวิธีวัดค่าความละม้ายที่มีประสิทธิภาพใกล้เคียงกับการระบุค่าความละม้ายโดยมนุษย์ เหมาะสมที่จะใช้ระบุแทนการระบุค่าคะแนนความละม้ายโดยมนุษย์ในข้อความที่พบว่ามีการลักลอกในระบบตรวจหาการลักลอกงานวิชาการมากที่สุด เนื่องจากให้ผลค่าของความละม้ายสัมพันธ์สอดคล้องเป็นไปในทิศทางเดียวกับผู้เชี่ยวชาญมากที่สุด

ทั้งนี้ หากพิจารณาตารางที่ 7.2 ในภาพรวม จะเห็นได้ว่าค่าความละม้ายที่วัดได้จากวิธีการวัดในระดับศัพท์และอักขระ ซึ่งถือได้ว่าเป็นวิธีการวัดค่าความละม้ายที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับผิวและไม่ได้ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ ตามลำดับนั้น มีความสัมพันธ์กับค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญมากกว่าวิธีการวัดในระดับวากยสัมพันธ์และความหมาย ซึ่งถือได้ว่าเป็นวิธีการวัดค่าความละม้ายที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับผิวลึก ดังจะ

สังเกตเห็นได้ว่าวิธีการวัดในระดับวากยสัมพันธ์และความหมายส่วนใหญ่มีค่าสัมประสิทธิ์สหสัมพันธ์ปรากฏอยู่ในอันดับครึ่งหลังของตารางดังกล่าว

ข้อค้นพบดังกล่าวข้างต้นนี้ไม่เป็นไปตามสมมติฐานของการวิจัยข้อที่ 4 ที่ระบุไว้ในตอนต้นว่าวิธีการวัดค่าความละม้ายที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับลึกจะให้ผลดีกว่าวิธีการวัดที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับผิว ทั้งนี้ สาเหตุที่ทำให้ผลการวิจัยดังกล่าวไม่เป็นไปตามสมมติฐานที่ตั้งเอาไว้นั้น ผู้วิจัยเห็นว่าอาจเป็นผลเนื่องมาจากวิธีการพิจารณาตัดสินค่าความละม้ายของผู้เชี่ยวชาญซึ่งเป็นมนุษย์ กล่าวคือ ในการพิจารณาค่าละม้ายของคู่หน่วยเทียบ ผู้เชี่ยวชาญทั้ง 3 คนมุ่งพิจารณาความแตกต่างด้านรูปคำและรูปอักขระมากกว่าพิจารณาการความแตกต่างด้านความสัมพันธ์ในระดับวากยสัมพันธ์และความหมายของข้อความในคู่หน่วยเทียบ ทั้งนี้ อาจเป็นรูปคำและรูปอักขระนั้นปรากฏให้เห็นชัดเจนและสังเกตเห็นความแตกต่างได้ง่ายกว่าการวิเคราะห์ความสัมพันธ์ทางวากยสัมพันธ์และความหมาย

อย่างไรก็ดี จะเห็นได้ว่าวิธีการวัดค่าความละม้ายที่มีประสิทธิภาพตามข้อค้นพบของการวิเคราะห์ในขั้นนี้ประยุกต์ใช้แนวคิดการคำนวณในระดับคำศัพท์และอักขระเช่นเดียวกับระดับของหน่วยที่ใช้ในการสร้างลักษณ์สำหรับจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกซึ่งให้ประสิทธิภาพดีที่สุดตามข้อค้นพบในบทที่แล้ว ดังจะเห็นได้จากการประเมินประสิทธิภาพของลักษณ์เป็นรายลักษณ์ในหัวข้อที่ 6.2.1 ของบทที่ 6 ซึ่งพบว่าค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ ($lcs_{\text{norm-W}}$), ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ ($lcs_{\text{norm-Char}}$), และค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ (QS_{W1}) ซึ่งเป็นวิธีการวัดค่าความละม้ายที่มีประสิทธิภาพสูงสุดในการทดลองในขั้นนี้ก็มีประสิทธิภาพในการจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอกอยู่ในอันดับ 9 (F = 0.9846), 18 (F = 0.9773), และ 5 (F = 0.9859) ตามลำดับ จากอันดับทั้งหมด 70 อันดับ เมื่อพิจารณาในแง่นี้ จะเห็นได้ว่าค่าความละม้ายที่วัดได้จากวิธีการวัดค่าความละม้ายทั้ง 3 วิธีดังกล่าวต่างก็มีประสิทธิภาพอยู่ในอันดับต้นๆ ของการประเมินประสิทธิภาพของลักษณ์

ลักษณะที่ได้กล่าวมาข้างต้นชี้ให้เห็นว่าลักษณ์ทางศัพท์และลักษณ์ทางอักขระเป็นประเภทของลักษณ์ที่ให้ผลดีทั้งในขั้นการตรวจหาการลักลอกและขั้นการบ่งชี้ปริมาณการลักลอกในระบบตรวจหาการลักลอก ด้วยเหตุนี้ จึงอาจกล่าวได้ว่า ในการพัฒนาระบบต้นแบบเพื่อตรวจหาการลักลอกงานวิชาการนั้น แนวคิดในการตรวจหาที่อิงรูปอักขระและรูปศัพท์ยังมีความสำคัญ ไม่ควรละเลย แต่ควรพัฒนาไปพร้อมกับวิธีการตรวจหาที่อิงจากความสัมพันธ์ทางภาษาในระดับที่ลึกกว่า

บทที่ 8
สรุป อภิปรายผล และข้อเสนอแนะ

งานวิจัยชิ้นนี้มีที่มาจากปัญหาการลักลอกงานวิชาการที่มีความรุนแรงมากขึ้นในปัจจุบัน ปัญหาดังกล่าวไม่เพียงก่อให้เกิดผลกระทบต่อสถาบันการศึกษาเท่านั้น แต่ยังทำให้กระบวนการสร้างสรรค์องค์ความรู้ของวงวิชาการต้องสะดุดหยุดลงด้วย ในขณะเดียวกัน เมื่อพิจารณาถึงแนวทางป้องปรามปัญหาดังกล่าว กลับพบว่าระบบตรวจหาการลักลอกที่มีให้ใช้งานอยู่ก็ไม่มีประสิทธิภาพในระดับที่น่าพึงพอใจ

สาเหตุที่กล่าวมาข้างต้นได้กลายมาเป็นแรงจูงใจให้ผู้วิจัยศึกษาเกี่ยวกับการพัฒนาระบบตรวจหาการลักลอกงานวิชาภาษาไทย โดยมีวัตถุประสงค์ของการศึกษา 4 ประการ ประการแรกคือ เพื่อวิเคราะห์หาลักษณ์ทางภาษาที่จะใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก ประการต่อมาคือ เพื่อพัฒนาระบบต้นแบบตรวจเทียบภายนอกหาการลักลอกงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความ ประการที่ 3 คือ เพื่อประเมินประสิทธิภาพของระบบต้นแบบที่พัฒนาขึ้นใน 2 แง่มุม ได้แก่ ความเหมาะสมของลักษณะของข้อมูลรับเข้าที่จะใช้ในระบบ และความเหมาะสมของลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก และประการสุดท้ายคือ เพื่อเปรียบเทียบวิธีวัดค่าความละม้ายของข้อความที่มีประสิทธิภาพ เหมาะสมจะนำมาใช้ในระบบตรวจหาการลักลอกมากที่สุด

เพื่อให้บรรลุวัตถุประสงค์ของการวิจัยที่ได้กล่าวมาข้างต้น ผู้วิจัยได้รวบรวมเอกสารและงานวิจัยที่เกี่ยวข้องเพื่อประมวลเอาองค์ความรู้และข้อค้นพบต่างๆ มากำหนดแนวทางในการดำเนินการวิจัย และเพื่อให้ผลของการวิจัยมีความหนักแน่นน่าเชื่อถือ ผู้วิจัยได้ศึกษาเกี่ยวกับกลวิธีลักลอกงานวิชาการภาษาไทย โดยเก็บข้อมูลจากการจำลองสถานการณ์การลักลอกแล้วนำมาวิเคราะห์ด้วยแนวคิดทางภาษาศาสตร์ ผลจากการศึกษาในขั้นนี้ได้ถูกนำมาใช้ประโยชน์ในการออกแบบและสร้างคลังข้อมูล ตลอดจนนำมาใช้อ้างอิงในการอภิปรายข้อค้นพบในขั้นต่อไป นอกจากนี้ผู้วิจัยยังได้ออกแบบ สร้าง และตรวจสอบคุณภาพของคลังข้อมูล ซึ่งถือได้ว่าเป็นหัวใจหลักของระบบด้วยความรอบคอบและรัดกุม ยังผลให้เชื่อได้ว่าผลการศึกษาที่ได้มาในตอนท้ายจะมีความหนักแน่นและเป็นที่เชื่อถือ สามารถอ้างอิงและนำไปพัฒนาต่อยอดได้ในอนาคต

ในบทสุดท้ายนี้ ผู้วิจัยได้กล่าวสรุปถึงผลของการวิจัยทั้งหมดที่ได้นำเสนอมาในงานวิจัยชิ้นนี้ จากนั้นจึงเป็นการอภิปรายผลการวิจัยอิงตามประเด็นที่ตั้งไว้ในสมมติฐานของการวิจัย แล้วจึงเป็นการ

เสนอข้อเสนอแนะอันจะเป็นประโยชน์ต่องานวิจัยอื่นในลำดับสุดท้าย ทั้งนี้ ผู้วิจัยได้เสนอรายละเอียดของประเด็นต่างๆ ดังกล่าวเป็นหัวข้อตามลำดับดังต่อไปนี้

8.1 สรุปผลการวิจัย

จากวัตถุประสงค์ของการวิจัยทั้ง 4 ข้อ ผู้วิจัยได้ดำเนินการวิจัยจนกระทั่งได้ข้อค้นพบดังมีรายละเอียดต่อไปนี้

8.1.1 ผลการวิเคราะห์หาลักษณ์

ลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกถือเป็นปัจจัยสำคัญอีกประการหนึ่งที่ส่งผลต่อประสิทธิภาพของระบบตรวจหาการลักลอกงานวิชาการที่พัฒนาขึ้นในงานวิจัยชิ้นนี้ ในขั้นนี้ ผู้วิจัยได้วิเคราะห์หาลักษณ์โดยพิจารณาจากลักษณะ (characteristics) ที่ชัดเจนและปรากฏอย่างสม่ำเสมอในข้อความที่มีการลักลอกและข้อความไม่มีการลักลอก แล้วนำลักษณะดังกล่าวมาประยุกต์เข้ากับองค์ความรู้และเทคนิควิธีทางการประมวลผลภาษาธรรมชาติที่มีใช้อยู่ในปัจจุบันแล้วสร้างเป็นลักษณ์ประเภทต่างๆ ขึ้น

เพื่อให้เอื้อต่อการนำไปทดสอบประสิทธิภาพของลักษณ์ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก ซึ่งเป็นวัตถุประสงค์อีกข้อของงานวิจัยชิ้นนี้ ผู้วิจัยได้แบ่งกลุ่มของลักษณ์จำนวนทั้งหมด 71 ลักษณ์ออกเป็น 2 กลุ่มใหญ่ ได้แก่ ลักษณ์อิงอักขระ และลักษณ์ทางภาษา

ลักษณ์อิงอักขระถือเป็นลักษณ์ที่ไม่ต้องประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในการวิเคราะห์หาและสร้าง เมื่อกล่าวเช่นนี้ ลักษณะเพียงประการเดียวที่จะนำมาประยุกต์ใช้เพื่อวิเคราะห์และสร้างลักษณ์ประเภทนี้ได้คือลักษณะของอักขระ (character) ในข้อความ ในงานวิจัยชิ้นนี้ ผู้วิจัยได้นำลักษณ์อิงอักขระมาใช้ในฐานะเส้นฐาน (baseline) ในการทดลองเปรียบเทียบประสิทธิภาพของลักษณ์ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกตามวัตถุประสงค์ของการวิจัยข้อที่ 3.2 ลักษณ์ประเภทนี้มีจำนวนทั้งหมด 20 ลักษณ์ ซึ่งวิเคราะห์ได้จากลักษณะทางรูปของอักขระและขนาดความยาวของข้อความอิงจากรูปของอักขระ

ส่วนลักษณ์ทางภาษานั้นคือลักษณ์ที่วิเคราะห์หาและสร้างขึ้นโดยอิงจากลักษณะทางภาษาของข้อความที่มีการลักลอกและไม่มีการลักลอก ในการวิเคราะห์หาลักษณะประเภทนี้ ผู้วิจัยจะพิจารณาข้อค้นพบที่ได้จากการวิเคราะห์กลวิธีลักลอกงานวิชาการภาษาไทยประกอบกับการสังเกตลักษณะทางภาษาของข้อความที่มีการลักลอกและไม่มีการลักลอก แล้วนำมาประยุกต์เข้ากับองค์ความรู้และเทคนิควิธีทางการประมวลผลภาษาธรรมชาติที่มีใช้อยู่ในปัจจุบัน

ผลของการวิจัยปรากฏว่า สามารถวิเคราะห์หาลักษณ์ทางภาษาเพื่อใช้ในการจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอกได้ รวมทั้งหมด 51 ลักษณ์ ลักษณ์เหล่านี้สามารถแบ่งออกตามระดับของหน่วยทางภาษาได้เป็น 4 ประเภท ได้แก่ ลักษณ์ทางศัพท์ ลักษณ์ทางวากยสัมพันธ์ ลักษณ์ทางความหมาย และลักษณ์ทางวากยสัมพันธ์และความหมาย

ลักษณ์ทางศัพท์เป็นลักษณ์ที่วิเคราะห์หาและสร้างขึ้นโดยอาศัยความรู้ทางภาษาศาสตร์ในระดับคำ ในแง่การตรวจหาการลักลอก ลักษณ์ทางศัพท์จะเอื้อให้เครื่องได้เรียนรู้ถึงขอบเขต ลักษณะและคุณสมบัติของคำในภาษา ลักษณ์ในกลุ่มนี้สามารถวิเคราะห์หาและสร้างได้จากลักษณะของรูปคำ ขอบเขตของคำ และลำดับของคำ โดยผ่านการประยุกต์ใช้แนวคิดทางการประมวลผลภาษาธรรมชาติเรื่องระยะการแก้ไข ลำดับร่วมที่ยาวที่สุด เอ็นแกรม และการวัดความละม้ายของข้อความ ทั้งนี้ จากการวิเคราะห์หาทำให้สามารถสร้างลักษณ์ทางศัพท์ได้ รวมจำนวนทั้งหมด 25 ลักษณ์

ลักษณ์ทางวากยสัมพันธ์คือลักษณ์ที่วิเคราะห์หาและสร้างขึ้นโดยอาศัยความรู้ทางภาษาศาสตร์ในระดับวลี อนุพากย์ และประโยค ในแง่การตรวจหาการลักลอก ลักษณ์ทางวากยสัมพันธ์จะเอื้อให้เครื่องได้เรียนรู้ถึงหน้าที่ ตำแหน่งในการปรากฏ และความสัมพันธ์ระหว่างหน่วยย่อยในวลี อนุพากย์ และประโยค ผู้วิจัยได้วิเคราะห์หาและสร้างลักษณ์ทางวากยสัมพันธ์ขึ้นโดยวิเคราะห์หาจากหมวดคำ และการเรียงลำดับของคำและหมวดคำในข้อความ ผู้วิจัยได้วิเคราะห์หาและสร้างลักษณ์ทางวากยสัมพันธ์ขึ้นโดยวิเคราะห์หาจากหมวดคำ และการเรียงลำดับของคำและหมวดคำในข้อความ ร่วมกับการประยุกต์ใช้แนวคิดทางการประมวลผลภาษาธรรมชาติเรื่องระยะการแก้ไขของหมวดคำ ลำดับร่วมที่ยาวที่สุดของหมวดคำ เอ็นแกรมของหมวดคำ และการวัดความละม้ายของข้อความ ทั้งนี้ จากการวิเคราะห์หาทำให้สามารถสร้างลักษณ์ทางศัพท์ได้ รวมจำนวนทั้งหมด 23 ลักษณ์

ลักษณ์ทางความหมายเป็นลักษณ์ที่วิเคราะห์หาและสร้างขึ้นโดยอาศัยความรู้ทางภาษาศาสตร์ในระดับความหมาย ในแง่การตรวจหาการลักลอก ลักษณ์ทางความหมายจะเอื้อให้เครื่องได้เรียนรู้ถึงความสัมพันธ์ทางความหมายระหว่างคำในข้อความ ทั้งนี้ ผู้วิจัยได้ประยุกต์แนวคิดต่างๆ ในการแทนรูปความสัมพันธ์ทางความหมายระหว่างคำในข้อความออกมาได้ ได้แก่ แนวคิดเรื่องการวิเคราะห์ความหมายแอบแฝง (Latent Semantic Analysis: LSA) และแนวคิดเรื่องการฝังคำ (word embedding) จากนั้นจึงนำความสัมพันธ์ทางความหมายระหว่างคำที่คำนวนได้มาสร้างเป็นเวกเตอร์ทางความหมายเพื่อใช้วัดค่าความละม้ายของข้อความ ทั้งนี้ จากการวิเคราะห์หาลักษณ์โดยอิงความหมายนี้ทำให้สามารถสร้างลักษณ์ทางศัพท์ได้ รวมจำนวนทั้งหมด 2 ลักษณ์ อย่างไรก็ตามเป็นที่น่าเสียดายว่าในขั้นตอนการสร้างลักษณ์จากเวกเตอร์ทางความหมายของคำโดยอาศัยแนวคิด

เรื่องการฝังคำนั้นต้องใช้เวลาค่อนข้างมากในการคำนวณค่าความสัมพันธ์แบบคำต่อคำ จึงทำให้ไม่สามารถนำลักษณ์ค่าความละม้ายทางความหมายของเวกเตอร์ของคำมาใช้ในการทดสอบได้

ส่วนลักษณ์ทางวากยสัมพันธ์และความหมายนั้นเป็นลักษณ์ที่วิเคราะห์หาและสร้างขึ้นโดยอาศัยความรู้ทางภาษาศาสตร์ในระดับวลี อนุพากย์ และประโยค และระดับความหมาย ในแง่การตรวจหาการลักลอก ลักษณ์ทางวากยสัมพันธ์และความหมายจะเอื้อให้เครื่องได้เรียนรู้ความสัมพันธ์เชิงหน้าที่ระหว่างหน่วยต่างๆ ในข้อความพร้อมกับความสัมพันธ์ทางความหมายของคำในข้อความ การวิเคราะห์หาและสร้างลักษณ์ชนิดนี้ ผู้วิจัยได้ประยุกต์ใช้แนวคิดของลีและคณะ (Y. Li et al., 2004; Y. Li et al., 2006) โดยเป็นการรวมเอาค่าความละม้ายของลำดับคำและค่าความละม้ายของเวกเตอร์ทางความหมายเข้าด้วยกัน ด้วยเหตุนี้ ในเชิงหลักการ ลักษณ์ชนิดนี้จึงมีคุณสมบัติในการตรวจจับข้อความที่ลักลอกที่ผ่านการแก้ไขโดยแทนที่คำเดิมด้วยคำที่มีความหมายใกล้เคียงกันและถูกเปลี่ยนลำดับของคำได้ อย่างไรก็ตาม ลักษณ์ในกลุ่มนี้สามารถวิเคราะห์หาและสร้างได้จำนวนเพียง 1 ลักษณ์

8.1.2 ผลการพัฒนาระบบต้นแบบตรวจหาการลักลอกงานวิชาการ

ระบบต้นแบบตรวจหาการลักลอกงานวิชาการที่พัฒนาขึ้นในงานวิจัยชิ้นนี้ทำให้เกิดผล (implement) ขึ้นโดยใช้ภาษาไพทอน เวอร์ชัน 3.6.3 ประกอบด้วยกระบวนการทำงาน 2 ส่วนหลักที่ออกแบบให้ทำงานต่อเนื่องกัน ได้แก่ ส่วนของการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก และส่วนของการวัดค่าความละม้ายของข้อความที่ได้รับการจำแนกประเภทว่ามีการลักลอก

ในส่วนของการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกนั้น ผู้วิจัยได้ประยุกต์วิธีการเรียนรู้ของเครื่อง โดยเลือกใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่ผู้วิจัยเลือกใช้ในงานวิจัยชิ้นนี้คือ SVC ซึ่งเป็นคลาสหนึ่งในไลบรารี Scikit-learn เวอร์ชัน 0.19.1 ซึ่งเป็นไลบรารีการเรียนรู้ของเครื่องสำหรับโปรแกรมภาษาไพทอน (python) มีประสิทธิภาพในการเป็นตัวจำแนกประเภทชุดข้อมูลทั้งในแบบ 2 ประเภทและแบบหลายประเภท ส่วนการตั้งค่าพารามิเตอร์ของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนนั้น ผู้วิจัยกำหนดค่าตามค่าปริยายของคลาส SVC ดังรายละเอียดที่ได้แสดงในตารางที่ 3.3 บทที่ 3

อีกส่วนหนึ่งของระบบเป็นส่วนของการวัดค่าความละม้ายของข้อความ ในส่วนนี้จะมีหน้าที่รับคู่ของข้อความที่ได้รับการจำแนกว่ามีการลักลอกจากส่วนแรกมาวัดค่าความละม้าย เนื่องมาจากการจำแนกประเภทในส่วนแรกนั้นไม่สามารถบ่งชี้ปริมาณการลักลอกได้ ส่วนนี้จึงทำหน้าที่วัดค่าความ

ละม้ายและรายงานค่าความละม้ายระหว่างข้อความออกมาเป็นค่าแทนปริมาณการลักลอก อย่างไรก็ตาม ในการเลือกวิธีการวัดค่าความละม้ายที่จะใช้ในขั้นนี้จำเป็นต้องเปรียบเทียบประสิทธิภาพของวิธีวัดค่าความละม้าย เพื่อจะได้ค่าความละม้ายที่มีประสิทธิใกล้เคียงกับค่าความละม้ายที่ระบุโดยมนุษย์มากที่สุด ทั้งนี้ ผลเปรียบเทียบประสิทธิภาพของวิธีวัดค่าความละม้ายจะได้กล่าวถึงในหัวข้อที่ 8.1.4 ต่อไป

8.1.3 ผลการประเมินประสิทธิภาพของระบบ

การประเมินประสิทธิภาพของระบบตรวจหาการลักลอกที่สร้างขึ้นเป็นวัตถุประสงค์อีกประการหนึ่งของงานวิจัยชิ้นนี้ ทั้งนี้ การประเมินประสิทธิภาพของระบบได้ถูกแบ่งออกเป็น 2 ส่วน ได้แก่การประเมินประสิทธิภาพของระบบเมื่อใช้ข้อมูลรับเข้าที่ต่างกัน และการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกันในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก ผลการประเมินประสิทธิภาพของระบบทั้ง 2 ส่วนสามารถสรุปได้ดังนี้

8.1.3.1 ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ข้อมูลรับเข้าที่ต่างกัน

ลักษณะของข้อมูลรับเข้าถือเป็นปัจจัยหนึ่งที่ส่งผลโดยตรงต่อประสิทธิภาพการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกของระบบ ในขั้นนี้ การวิเคราะห์ในขั้นนี้จึงเป็นการเปรียบเทียบประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกของระบบเมื่อใช้ข้อมูลรับเข้า 2 ประเภทที่แตกต่างกัน ได้แก่ ย่อหน้า และหน่วยปริจเฉทพื้นฐาน โดยควบคุมให้ใช้ลักษณ์ในการจำแนกชุดเดียวกัน ได้แก่ ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ (QS_{W1}), ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}), ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของคำ (QS_{W3}), ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 4 แกรมของคำ (QS_{W4}), และค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของ 5 แกรมของคำ (QS_{W5})

ผลการเปรียบเทียบประสิทธิภาพของระบบเมื่อใช้ประเภทของข้อมูลรับเข้าที่แตกต่างกันปรากฏว่า เมื่อทดลองให้จำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกในคลังข้อมูลทั้งหมด ค่า F ที่ได้จากการใช้ย่อหน้าเป็นข้อมูลรับเข้าสูงกว่าค่า F ที่ได้จากการใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้าทั้งหมดทุกลักษณ์ แสดงให้เห็นว่าในภาพรวมแล้ว ระบบที่ใช้ย่อหน้าเป็นข้อมูลรับเข้าจะมีประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกดีกว่าระบบที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้า และเมื่อทดลองให้ระบบจำแนกประเภทข้อมูลลักลอกแต่ละประเภท โดยแบ่งชุดข้อมูลทดสอบออกเป็น 4 ชุดย่อย แต่ละชุดย่อยจะประกอบด้วยข้อมูลลักลอกเฉพาะประเภทกับข้อมูลที่ไม่มีการลักลอก ได้แก่ ข้อมูลลักลอกประเภทคัดลอกโดยตรง

และข้อมูลที่ไม่มีการลักลอก (EC-NO), ข้อมูลลักลอกประเภทคัดลอกโดยใกล้เคียงและข้อมูลที่ไม่มีการลักลอก (NC-NO), ข้อมูลลักลอกประเภทคัดลอกโดยดัดแปลงและข้อมูลที่ไม่มีการลักลอก (MO-NO), และข้อมูลลักลอกประเภทถอดความและข้อมูลที่ไม่มีการลักลอก (PA-NO) ผลปรากฏว่าระบบที่ใช้ย่อหน้าเป็นข้อมูลรับเข้าให้ค่า F สูงกว่าระบบที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้าทั้งหมดในข้อมูลลักลอกทุกประเภท แสดงให้เห็นว่าย่อหน้าเป็นข้อมูลรับเข้าที่เหมาะสมจะใช้ในการตรวจหาการลักลอกทุกประเภท

8.1.3.2 ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ที่ต่างกัน

ลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกถือเป็นปัจจัยสำคัญอีกประการหนึ่งที่ส่งผลต่อประสิทธิภาพของระบบตรวจหาการลักลอกงานวิชาการที่พัฒนาขึ้น การทดลองประเมินประสิทธิภาพของระบบในขั้นนี้ ผู้วิจัยได้นำลักษณ์ที่วิเคราะห์และสร้างได้สำเร็จรวมจำนวนทั้งหมด 70 ลักษณ์ ซึ่งแบ่งเป็นลักษณ์อิงอักขระและลักษณ์ทางภาษา มาทดสอบประสิทธิภาพในการจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอก โดยแบ่งการทดลองออกเป็น 3 ขั้น ได้แก่ การประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์เป็นรายลักษณ์ การประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์แบบรวมชุด และการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ทางภาษา

ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์เป็นรายลักษณ์ปรากฏว่า ลักษณ์ที่ให้ประสิทธิภาพสูงสุด 10 อันดับแรก ประกอบด้วยลักษณ์ทางศัพท์ ซึ่งจัดเป็นลักษณ์ทางภาษาใน 6 อันดับแรก โดยลักษณ์ที่มีประสิทธิภาพสูงที่สุดคือลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) รองลงมาคือลักษณ์ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของคำ (J_{W2}) และลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของคำ (QS_{W3}) ซึ่งให้ค่า F เท่ากับ 0.9870, 0.9867, และ 0.9865 ตามลำดับ ส่วน 4 อันดับที่เหลือเป็นลักษณ์ทางอักขระ ซึ่งจัดเป็นลักษณ์ที่ไม่ได้ประยุกต์ใช้ความรู้ทางภาษาในการสร้าง

ส่วนการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์แบบรวมชุดนั้น ผู้วิจัยจะนำลักษณ์ที่ให้ประสิทธิภาพดีที่สุด 10 อันดับแรกจากการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์เป็นรายลักษณ์มารวมเป็นชุดของลักษณ์และนำฝึกฝนและทดสอบประสิทธิภาพในการจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอก จากนั้นจะทดลองรวมชุดของลักษณ์ใหม่โดยตัดลักษณ์ที่ให้ประสิทธิภาพอยู่ในอันดับที่ต่ำที่สุดออกไป 1 ลักษณ์และนำมาฝึกฝนและทดสอบประสิทธิภาพด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนอีกรอบ ผู้วิจัยจะจัดชุดของลักษณ์ใหม่และนำไปฝึกฝนและทดสอบเช่นนี้

จนกระทั่งเหลือลักษณ์ที่ให้ประสิทธิภาพอยู่ในอันดับที่สูงที่สุดเพียงลักษณ์เดียว จากวิธีการดังกล่าวนี้จะทำให้ได้ชุดของลักษณ์ทั้งหมด 10 ชุดที่ผ่านการฝึกและทดสอบด้วยแบบจำลอง

ผลการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์แบบรวมชุดปรากฏว่า ชุดของลักษณ์ที่ให้ประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและข้อความที่ไม่มีการลักลอกดีที่สุดคือชุดที่ประกอบด้วยลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) เพียงลักษณ์เดียว ซึ่งให้ค่า F เท่ากับ 0.9870

และในส่วนการประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ทางภาษา ผู้วิจัยได้ทดลองรวมชุดของลักษณ์ทางภาษาและนำไปฝึกฝนและทดสอบด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ลักษณ์ 3 ลักษณ์ที่ให้ประสิทธิภาพรายลักษณ์สูงที่สุดตามระดับของหน่วยภาษา 3 กลุ่ม ได้แก่ ลักษณ์ทางศัพท์ ลักษณ์ทางวายสัมพันธ์ และลักษณ์ทางความหมาย มารวมเป็นชุดทั้งหมด 4 ชุด ได้แก่

1) ลักษณ์ทางภาษาชุดที่ 1 (LF_1) ประกอบด้วยลักษณ์ทางศัพท์และลักษณ์ทางวายสัมพันธ์ ได้แก่ ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) และลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf ($cos_{tf\text{-}idf\text{-}123}$) ตามลำดับ
2) ลักษณ์ทางภาษาชุดที่ 2 (LF_2) ประกอบด้วยลักษณ์ทางศัพท์และลักษณ์ทางความหมาย ได้แก่ ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) และลักษณ์ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง (cos_{LSA}) ตามลำดับ
3) ลักษณ์ทางภาษาชุดที่ 3 (LF_3) ประกอบด้วยลักษณ์ทางวายสัมพันธ์และลักษณ์ทางความหมาย ได้แก่ ลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf ($cos_{tf\text{-}idf\text{-}123}$) และลักษณ์ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง (cos_{LSA}) ตามลำดับ
4) ลักษณ์ทางภาษาชุดที่ 4 (LF_4) ประกอบด้วยลักษณ์ทางศัพท์ ลักษณ์ทางวายสัมพันธ์ และลักษณ์ทางความหมาย ได้แก่ ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) ลักษณ์ค่าความละม้ายโคไซน์ของน้ำหนักช่วงเอ็นแกรมของคำแบบ tf–idf ($cos_{tf\text{-}idf\text{-}123}$) และลักษณ์ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง (cos_{LSA}) ตามลำดับ

ผลประเมินประสิทธิภาพของระบบเมื่อใช้ลักษณ์ทางภาษาปรากฏว่า ชุดรวมลักษณ์ทางภาษาชุดที่ 2 (LF_2) ซึ่งประกอบด้วยลักษณ์ทางศัพท์และลักษณ์ทางความหมาย ได้แก่ ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) และลักษณ์ค่าความละม้ายโคไซน์ของเวกเตอร์ทางความหมายแอบแฝง (cos_{LSA}) ตามลำดับ เป็นชุดรวมลักษณ์ที่ให้ประสิทธิภาพดีที่สุดใน

การจำแนกประเภท กล่าวคือมีค่า F เท่ากับ 0.9859 อย่างไรก็ตาม ค่าดังกล่าวก็ยังต่ำกว่าค่า F ที่ได้จากการจำแนกประเภทโดยใช้ลักษณ์ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) เพียงลักษณ์เดียว ซึ่งมีค่า F เท่ากับ 0.9870

8.1.4 ผลการเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความ

การทดลองในขั้นนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความเพื่อหาวิธีการวัดค่าความละม้ายของข้อความที่มีประสิทธิภาพใกล้เคียงกับการระบุค่าความละม้ายโดยมนุษย์เพื่อนำมาใช้ในระบบ ในการทดลองนั้น ผู้วิจัยได้สร้างชุดข้อมูลทดลองขึ้นประกอบไปด้วยคู่หน่วยเทียบของย่อหน้าขนาดต่างๆ 150 ย่อหน้า จากนั้นจึงให้ผู้เชี่ยวชาญทางภาษาไทย 3 คนเป็นผู้ระบุค่าความละม้ายของคู่หน่วยเทียบดังกล่าว เพื่อนำมาเปรียบเทียบกับค่าความละม้ายที่วัดได้จากวิธีการวัดค่าความละม้ายของข้อความทั้งหมด 60 วิธี ซึ่งเป็นผลที่ได้มาจากขั้นตอนการวิเคราะห์หาและลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก ทั้งนี้ วิธีการวัดค่าความละม้ายของข้อความที่ให้ค่าละม้ายที่สอดคล้องสัมพันธ์กับค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญมากที่สุดจะถือว่ามีประสิทธิภาพใกล้เคียงกับมนุษย์มากที่สุด

ทั้งนี้ เมื่อได้ค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญมาแล้ว ในขั้นแรก ผู้วิจัยจะนำค่าความละม้ายดังกล่าวมาทดสอบความเป็นเอกพันธ์ด้วยการวิเคราะห์ความแปรปรวนแบบทางเดียว (one-way ANOVA) ผลการทดสอบในขั้นนี้ปรากฏว่า ณ ระดับนัยสำคัญ $\alpha = .05$ ค่าความละม้ายที่ให้โดยผู้เชี่ยวชาญทั้ง 3 คนไม่แตกต่างกัน ($F = .856$, $p = .426$) หรือกล่าวอีกนัยคือค่าความละม้ายที่ให้โดยผู้เชี่ยวชาญทั้ง 3 คนมีความเป็นเอกพันธ์ ด้วยเหตุนี้ จึงสามารถนำค่าเฉลี่ยของค่าความละม้ายที่ให้โดยผู้เชี่ยวชาญทั้ง 3 คนในแต่ละคู่หน่วยเทียบไปวิเคราะห์ความสัมพันธ์กับค่าความละม้ายที่เครื่องวัดได้จากวิธีต่างๆ ได้

ในขั้นต่อมา ผู้วิจัยนำค่าความละม้ายที่วัดได้จากชุดข้อมูลทดลองโดยใช้วิธีการวัดทั้ง 60 วิธีมาแบ่งเป็น 5 กลุ่มตามการประยุกต์ใช้ระดับของหน่วยในการคำนวณค่าความละม้าย ดังนี้

1) ค่าความละม้ายอิงอักขระ
2) ค่าความละม้ายอิงศัพท์
3) ค่าความละม้ายอิงวายสัมพันธ์
4) ค่าความละม้ายอิงความหมาย
5) ค่าความละม้ายอิงวายสัมพันธ์และความหมาย

จากนั้นจึงวิเคราะห์ความสัมพันธ์ระหว่างค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญกับค่าความละม้ายที่ได้จากวิธีการวัดค่าความละม้ายวิธีต่างๆ ผู้วิจัยได้ประยุกต์ใช้การวิเคราะห์สัมประสิทธิ์

สหสัมพันธ์ (correlation coefficient) ทั้งนี้ ค่าความละม้ายที่มีความสัมพันธ์เป็นไปในทิศทางเดียวกับค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญมากที่สุด 3 วิธีแรก ได้แก่ ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของคำ ($lcs_{norm\text{-}W}$), ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ ($lcs_{norm\text{-}Char}$), และค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ (QS_{W1}) ซึ่งมีค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.9124, 0.9112, และ 0.8932 ตามลำดับ

ด้วยเหตุนี้ จึงสามารถสรุปได้ว่าวิธีการวัดค่าความละม้ายด้วยการหาค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ ($lcs_{norm\text{-}W}$) เป็นวิธีวัดค่าความละม้ายที่มีประสิทธิภาพใกล้เคียงกับการระบุค่าความละม้ายโดยมนุษย์ เหมาะสมที่จะใช้ระบุค่าคะแนนความละม้ายในข้อความที่พบว่ามีการลักลอกในระบบตรวจหาการลักลอกงานวิชาการ เนื่องจากให้ผลค่าของความละม้ายสัมพันธ์สอดคล้องเป็นไปในทิศทางเดียวกับผู้เชี่ยวชาญมากที่สุด

8.2 อภิปรายผลการวิจัย

จากผลการวิจัยที่ได้กล่าวไปในหัวข้อที่ 8.1 ผู้วิจัยมีข้ออภิปรายเกี่ยวกับผลการวิจัยดังกล่าวอิงตามสมมติฐานของวิจัยในประเด็นต่างๆ ดังต่อไปนี้

8.2.1 ลักษณ์ทางภาษาที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก

จากผลการวิเคราะห์หาลักษณ์ที่ได้กล่าวไปในหัวข้อที่ 8.1.1 จะเห็นได้ว่า ในการวิเคราะห์หาลักษณ์ทางภาษาสำหรับใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกนั้น ผู้วิจัยได้นำลักษณะทางภาษาของข้อความที่มีการลักลอกและไม่มีการลักลอกประยุกต์เข้ากับเทคนิควิธีทางการประมวลผลภาษาธรรมชาติที่มีใช้อยู่ในปัจจุบันอย่างหลากหลาย ไม่ว่าจะเป็นแนวคิดระยะการแก้ไข ลำดับร่วมที่ยาวที่สุด เอ็นแกรมของหน่วยทางภาษาระดับคำและหมวดคำ การให้น้ำหนักคำ การแทนรูปหน่วยทางภาษาในระดับต่างๆ ให้เป็นเวกเตอร์ หรือการวัดความละม้ายของข้อความ

ในแง่การพิสูจน์สมมติฐานนั้น กล่าวได้ว่าผลการวิจัยในขั้นนี้เป็นไปตามสมมติฐานที่ได้ตั้งไว้ กล่าวคือ ไม่ว่าจะเป็นการวิเคราะห์ลักษณ์ในระดับคำ ระดับวากยสัมพันธ์ หรือในระดับความหมาย ก็สามารถวิเคราะห์หาลักษณ์ตามที่ระบุไว้ในสมมติฐานข้อที่ 1 ได้หมดทั้งสิ้น ยิ่งไปกว่านั้น งานวิจัยชิ้นนี้ยังสามารถสร้างลักษณ์ทางภาษาโดยประยุกต์ใช้แนวคิดอื่นๆ นอกเหนือไปจากที่ได้ระบุไว้ในสมมติฐานของการวิจัยอีกด้วย ด้วยการพิจารณาลักษณะทางภาษาและการคาดการณ์ถึงประสิทธิภาพและผลที่ได้จะได้รับจากการใช้ลักษณ์แต่ละตัว ทำให้สามารถวิเคราะห์หาลักษณ์ทางภาษารวมเป็นจำนวนทั้งหมดได้ถึง 51 ลักษณ์

อย่างไรก็ตาม เป็นที่น่าเสียดายว่าในขั้นตอนการสร้างลักษณ์จากเวกเตอร์ทางความหมายของคำตามแนวคิดการฝังคำนั้นต้องใช้เวลาค่อนมากในการคำนวณค่าความสัมพันธ์แบบคำต่อคำ จึงทำให้ไม่สามารถนำลักษณ์ค่าความละม้ายทางความหมายของเวกเตอร์ของคำมาใช้ในการทดสอบประสิทธิภาพของลักษณ์ได้ ทั้งนี้ ในประเด็นนี้ ผู้วิจัยเห็นว่าเป็นผลเนื่องมาจากข้อความแต่ละข้อความในคู่หน่วยเทียบที่ใช้จริงในคลังข้อมูลนั้นประกอบด้วยคำจำนวนมาก และมีคำปรากฏซ้ำกันเพียงบางส่วนเท่านั้น เครื่องจึงต้องเวียนคำนวณค่าความสัมพันธ์ทางความหมายแบบคำต่อคำหลายรอบ เป็นเหตุให้ใช้เวลานานในการสร้างตารางคำนวณค่าความสัมพันธ์ทางความหมายแบบคำต่อคำให้เสร็จสมบูรณ์

8.2.2 ประเภทของข้อมูลรับเข้า

จากผลการประเมินประสิทธิภาพของระบบเมื่อใช้ข้อมูลรับเข้าที่ต่างกันที่ได้กล่าวไปในหัวข้อที่ 8.1.3.1 จะเห็นได้ว่าย่อหน้าเป็นข้อมูลรับเข้าที่เหมาะสมจะใช้ในการตรวจหาการลักลอกทุกประเภท ข้อค้นพบดังกล่าวนี้ไม่เป็นไปตามสมมติฐานของการวิจัยข้อที่ 2 ที่ได้ระบุไว้ว่าข้อมูลรับเข้าที่เป็นหน่วยปริจเฉทพื้นฐานสามารถตรวจหาการลักลอกแบบคัดลอกโดยตรง การลักลอกแบบคัดลอกโดยใกล้เคียง และการลักลอกแบบคัดลอกโดยดัดแปลง ได้ดีกว่าข้อมูลรับเข้าที่เป็นย่อหน้า ในขณะที่ข้อมูลรับเข้าที่เป็นย่อหน้าสามารถตรวจหาการลักลอกแบบถอดความได้ดีกว่าข้อมูลรับเข้าที่เป็นหน่วยปริจเฉทพื้นฐาน

ทั้งนี้ ผู้วิจัยเห็นว่าการที่ประสิทธิภาพในการจำแนกประเภทของระบบที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้าต่ำกว่าระบบที่ใช้ย่อหน้าเป็นข้อมูลรับเข้านั้น เนื่องมาจากสาเหตุหลัก 4 ประการ ได้แก่

ประการแรก สาเหตุจากขนาดของหน่วยปริจเฉทพื้นฐาน ทั้งนี้ เป็นที่ทราบแล้วว่าหน่วยปริจเฉทพื้นฐานนั้นประกอบด้วยจำนวนคำที่น้อยกว่าย่อหน้า บางหน่วยปริจเฉทอาจประกอบด้วยคำเพียง 1-3 คำเท่านั้น ลักษณะดังกล่าวนี้ทำให้เมื่อประยุกต์ใช้ลักษณ์ค่าความละม้ายที่คำนวณจากเอ็นแกรมของคำที่มากขึ้นเรื่อยๆ ค่าความละม้ายก็อาจเท่ากับ 0 ได้ การได้ลักษณ์ค่าความละม้ายที่มีค่าเป็น 0 นี้ย่อมส่งผลให้การตัดสินใจจำแนกประเภทของเครื่องเบี่ยงเบนไป เนื่องจากคู่หน่วยเทียบของหน่วยปริจเฉทพื้นฐานที่ลักลอกและไม่ลักลอกต่างก็มีค่าความละม้ายเท่ากับ 0 เหมือนกัน

สาเหตุประการต่อมาเป็นสาเหตุที่เกี่ยวกับแนวคิดเชิงโครงสร้างวาทะ ตามทฤษฎีดังกล่าวหน่วยปริจเฉทพื้นฐานแต่ละหน่วยที่ประกอบกันขึ้นเป็นปริจเฉทย่อมมีวาทสัมพันธ์อย่างใดอย่างหนึ่งต่อกัน ในกรณีของการนำหน่วยปริจเฉทพื้นมาใช้เป็นข้อมูลรับเข้าในการตรวจหาการลักลอก ผู้วิจัยเห็นว่าหากตัดแบ่งเฉพาะหน่วยปริจเฉทพื้นฐานมาใช้โดยไม่กำกับและพิจารณาวาทสัมพันธ์ระหว่าง

หน่วยต่างๆ หน่วยปริจเฉทพื้นฐานก็มีสถานะเป็นเพียงข้อความที่มีขนาดสั้นกว่าย่อหน้า ไม่สามารถสะท้อนความสัมพันธ์ภายในปริจเฉทอันเป็นแง่มุมหนึ่งที่ควรพิจารณาในการลักลอกและการตรวจหาการลักลอกได้ และจะส่งผลต่อประสิทธิภาพของการตรวจหาการลักลอกดังได้กล่าวไปแล้วข้างต้น

สาเหตุอีกประการหนึ่งเป็นสาเหตุอันเนื่องมาการกลวิธีที่ผู้ลักลอกใช้ในลักลอก ทั้งนี้ กลวิธีลักลอกวิธีหนึ่งที่ผู้ลักลอกสามารถทำได้คือการแทรกหรือลบเนื้อหาในข้อความในระดับหน่วยปริจเฉทพื้นฐาน ในการวัดค่าความละม้ายเพื่อนำมาใช้เป็นลักษณ์ คู่หน่วยเทียบของหน่วยปริจเฉทพื้นฐานที่จับคู่จากหน่วยปริจเฉทพื้นฐานที่มีอยู่เดิมกับหน่วยปริจเฉทพื้นฐานที่ถูกแทรกเข้าไปใหม่ อาจให้ค่าความละม้ายที่สูงในระดับที่ใกล้เคียงกับคู่หน่วยเทียบที่เป็นคู่ลักลอกจริง ในขณะที่การลบหน่วยปริจเฉทพื้นฐานก็จะส่งผลให้ไม่สามารถจับคู่ลักลอกที่แท้จริงได้ เนื่องจากหน่วยปริจเฉทพื้นฐานเดิมที่ปรากฏอยู่ในข้อความต้นฉบับได้ถูกลบออกไปในข้อความลักลอก ในแง่นี้ค่าความละม้ายที่ได้จะต่ำมาก อย่างไรก็ตาม การกำหนดคำตอบสำหรับการเรียนรู้ของเครื่องจะกำหนดว่าลักษณะเช่นนี้ถือเป็นการลักลอก กรณีเช่นนี้จะส่งผลให้การตัดสินใจจำแนกประเภทของเครื่องเบี่ยงเบนไป เนื่องจากค่าความละม้ายที่เครื่องเรียนรู้มีลักษณะเหมือนกันทั้งจากการกำหนดคำตอบว่าเป็นการลักลอกและไม่เป็นการลักลอก เหตุผลประการสุดท้ายนี้ยิ่งเป็นการแสดงให้เห็นข้อจำกัดของหน่วยรับเข้าประเภทหน่วยปริจเฉทพื้นฐานที่ไม่ได้รับการกำกับวาทสัมพันธ์มาพร้อมกันก่อนนำมาใช้เป็นข้อมูลรับเข้า

และสาเหตุประการสุดท้ายเป็นสาเหตุอันเนื่องมาจากการสร้างข้อมูลที่ใช้ในการทดสอบ ด้วยลักษณะของข้อมูลที่ใช้ในการทดสอบนั้นสร้างขึ้นจำลองการลักลอกทั้งย่อหน้า มิใช่การแทรกข้อความลักลอกเข้าไปในย่อหน้าเพียงบางส่วน ในกรณีเช่นนี้ การตรวจโดยอิงจากค่าความละม้ายของข้อความอาจมีปัญหาได้ เพราะข้อความส่วนใหญ่ในคู่หน่วยเทียบจะไม่ละม้ายกัน และส่งผลโดยตรงต่อประสิทธิภาพในการจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอก กล่าวคือทำให้ค่า F ต่ำ ดังนั้นแล้ว ผู้วิจัยจึงเห็นควรให้มีการทดสอบเพิ่มเติมในปริบทที่การลักลอกเป็นเพียงบางส่วนของย่อหน้า การตรวจหาในลักษณะนี้ต้องทำข้อความแต่ละส่วน ซึ่งอาจเป็นไปได้ว่าหน่วยรับเข้าประเภทหน่วยปริจเฉทพื้นฐานจะมีความเหมาะสมกับการตรวจหาในลักษณะเช่นนี้มากกว่า

8.2.3 ประสิทธิภาพของระบบเมื่อลักษณ์ใช้ที่ต่างกันในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก

จากผลการประเมินประสิทธิภาพของลักษณ์ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกที่ได้กล่าวไปในหัวข้อที่ 8.1.3.2 จะเห็นได้ว่าจากการทดลองประเมินประสิทธิภาพของลักษณ์ทั้ง 3 ขั้นตอน ผลปรากฏตรงกันว่าลักษณ์ทางศัพท์ซึ่งอาจถือได้ว่าลักษณ์ที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับผิวมีประสิทธิภาพในการตรวจหาการลักลอกได้ดีกว่าลักษณ์ที่ไม่

ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ คือลักษณ์อิงอักขระ และลักษณ์ที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับลึก อันได้แก่ลักษณ์ทางวากยสัมพันธ์และลักษณ์ทางความ ตามลำดับ ข้อสรุปดังกล่าวนี้ไม่เป็นไปตามสมมติฐานของการวิจัยข้อที่ 3 ที่ระบุไว้ว่าลักษณ์ที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับลึกจะมีประสิทธิภาพในการตรวจหาการลักลอกได้ดีกว่าลักษณ์ที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับผิว และลักษณ์ที่ไม่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ ตามลำดับ ในแง่นี้ ผู้วิจัยเห็นว่าเป็นผลมาจากสาเหตุ 3 ประการ ได้แก่

ประการแรก คือ การประยุกต์ใช้แนวคิดเรื่องค่าความละม้ายในการวิเคราะห์และสร้างลักษณ์ ดังจะเห็นได้ว่าลักษณ์ทางศัพท์ที่ใช้ในการทดลองครั้งนี้วิเคราะห์หาและสร้างจากรูปคำที่ปรากฏโดยตรง และแทนรูปออกมาเป็นค่าความละม้ายซึ่งมีช่วงของปริมาณตั้งแต่ 0 ถึง 1 ลักษณะเช่นนี้จึงเอื้อให้เครื่องเรียนรู้และตัดสินใจได้อย่างแม่นยำ ทั้งนี้หากเปรียบเทียบกับลักษณ์ชนิดอื่นที่ไม่ได้ประยุกต์ใช้แนวคิดเรื่องค่าความละม้าย เช่น ลักษณ์ขนาดของคู่หน่วยเทียบ ลักษณ์ผลต่างของขนาดของคู่หน่วยเทียบ หรือลักษณ์ความยาวของลำดับย่อยร่วมยาวสุดที่ยาวที่สุด ลักษณ์เหล่านี้เป็นค่าตัวเลขดิบที่ได้จากการคำนวณ ไม่ได้อยู่ในรูปช่วงของปริมาณที่จำกัด จึงเป็นการยากที่เครื่องจะใช้ในสร้างสมมติฐานทั่วไปในการตัดสินใจจำแนกประเภท

ประการต่อมา คือ การแทนรูปลักษณ์จากรูปในระดับผิว ดังข้อค้นพบที่ปรากฏให้เห็นว่าลักษณ์ทางศัพท์และลักษณ์ทางอักขระให้ประสิทธิภาพในการจำแนกข้อความลักลอกและข้อความที่ไม่มีการลักลอกดีกว่าลักษณ์ทางวากยสัมพันธ์และลักษณ์ทางความหมายซึ่งมุ่งแทนรูปความสัมพันธ์ทางภาษาศาสตร์ในระดับลึก ทั้งนี้ เป็นไปได้ว่าลักษณ์ทางศัพท์เป็นการแทนรูปจากคำซึ่งเป็นหน่วยทางภาษาที่มีขอบเขตชัดเจน เมื่อนำมาแปลงเป็นค่าตัวเลขแล้ว ค่าที่ได้ก็แสดงมีความแตกต่างกันอย่างชัดเจน เมื่อเข้าสู่อัลกอริทึมของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนแล้ว แบบจำลองจึงสามารถจำแนกประเภทได้อย่างถูกต้องและแม่นยำ

สาเหตุประการสุดท้ายเป็นเหตุที่มาจากลักษณะของข้อมูลที่ใช้ในการทดลอง กล่าวคือ คลังข้อมูลที่สร้างขึ้นในงานชิ้นนี้สร้างโดยอาศัยการแก้ไขในระดับคำและวลีเป็นส่วนใหญ่ ในขณะที่การแก้ไขในระดับความหมายอย่างการถอดความปรากฏเป็นข้อมูลเพียงร้อยละ 10 ของข้อมูลที่เป็นการลักลอก ด้วยเหตุนี้ ลักษณ์ทางศัพท์จึงให้ประสิทธิภาพในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอกได้ดีกว่าลักษณ์ประเภทอื่นๆ อย่างเห็นได้ชัด อย่างไรก็ดี คลังข้อมูลที่สร้างขึ้นเพื่อใช้ในงานวิจัยชิ้นนี้ก็ออกแบบโดยอิงจากผลการวิเคราะห์กลวิธีการลักลอกงานวิชาการภาษาไทย เพื่อให้สอดคล้องกับสถานการณ์การลักลอกที่เกิดขึ้นจริง ด้วยเหตุนี้เอง ผู้วิจัยจึงมั่นใจว่าลักษณ์ทางศัพท์เป็นลักษณ์ประเภทที่มีประสิทธิภาพในการจำแนกประเภทข้อความลักลอกและข้อความที่ไม่มีการลักลอก เหมาะสมจะใช้ในการพัฒนาระบบตรวจหาการลักลอกงานวิชาการต่อไป

8.2.4 ประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความ

จากผลการเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความ จะเห็นได้ว่าค่าความละม้ายที่วัดได้จากวิธีการวัดในระดับศัพท์และอักขระ ซึ่งถือได้ว่าเป็นวิธีการวัดค่าความละม้ายที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับผิวและไม่ได้ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ตามลำดับนั้น มีความสัมพันธ์กับค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญมากกว่าวิธีการวัดในระดับวากยสัมพันธ์และความหมาย ซึ่งถือได้ว่าเป็นวิธีการวัดค่าความละม้ายที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับผิวลึก ข้อค้นพบดังกล่าวข้างต้นนี้ไม่เป็นไปตามสมมติฐานของการวิจัยข้อที่ 4 ที่ระบุไว้ว่าวิธีการวัดค่าความละม้ายที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับลึกจะให้ผลดีกว่าวิธีการวัดที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับผิว

ทั้งนี้ สาเหตุที่ทำให้ผลการวิจัยดังกล่าวไม่เป็นไปตามสมมติฐานที่ตั้งเอาไว้นั้น ผู้วิจัยเห็นว่าอาจเป็นผลเนื่องมาจากวิธีการพิจารณาตัดสินค่าความละม้ายของผู้เชี่ยวชาญซึ่งเป็นมนุษย์ กล่าวคือในการพิจารณาค่าละม้ายของคู่หน่วยเทียบ ผู้เชี่ยวชาญทั้ง 3 คนมุ่งพิจารณาความแตกต่างด้านรูปคำและรูปอักขระมากกว่าพิจารณาการความแตกต่างด้านความสัมพันธ์ในระดับวากยสัมพันธ์และความหมายของข้อความในคู่หน่วยเทียบ ทั้งนี้ อาจเป็นได้ว่ารูปคำและรูปอักขระนั้นปรากฏให้เห็นชัดเจนและสังเกตเห็นความแตกต่างได้ง่ายกว่าการวิเคราะห์ความสัมพันธ์ทางวากยสัมพันธ์และความหมาย

8.3 ข้อเสนอแนะ

จากผลการวิจัยที่กล่าวได้ไปข้างต้น ผู้วิจัยเห็นว่ามีประเด็นสำคัญหลายประเด็นที่สามารถนำไปศึกษาและพัฒนาต่อในอนาคตได้ ในที่นี้ ผู้วิจัยจะขอกล่าวถึงประเด็นดังกล่าวโดยสังเขปดังนี้

8.3.1 การวิเคราะห์กลวิธีลักลอกในงานวิชาการภาษาไทย

จากผลการวิเคราะห์กลวิธีลักลอกงานวิชาที่ได้แสดงไปในบทที่ 4 ผู้วิจัยเห็นว่ามีประเด็นที่ควรนำเสนอเพื่อนำไปพัฒนาต่อ กล่าวคือ ในเชิงทฤษฎีนั้น ผู้วิจัยเห็นว่าสามารถนำทฤษฎีโครงสร้างวาทะไปใช้ศึกษาเพิ่มเติมได้ในหลายแง่มุม ในการวิเคราะห์ข้อความลักลอกดังเช่นที่ปรากฏในงานวิจัยชิ้นนี้ หากสามารถกำกับวาทสัมพันธ์ให้กับข้อมูลทั้งหมดได้ ก็จะช่วยขยายขอบเขตผลการศึกษาให้ลึกซึ้งยิ่งขึ้น กล่าวคือจะทำให้ทราบว่ารูปแบบความสัมพันธ์ในเนื้อหาระหว่างข้อความต้นฉบับกับข้อความลักลอกมีลักษณะเป็นเช่นไร หรือในด้านการวิเคราะห์องค์ประกอบของข้อความในการถอดความ การเลือกกำหนดสถานะความสำคัญให้กับหน่วยปริจเฉทพื้นฐานก็เป็นอีกหัวข้อที่น่าศึกษาวิจัยเพิ่มเติม

เนื่องจากจะเอื้อประโยชน์ในการทำความเข้าใจเจตนาของผู้เขียนที่ถ่ายทอดผ่านข้อความได้ อีกทั้งยังทำให้ทราบถึงกลไกทางภาษาที่ใช้ในการถ่ายทอดเจตนาที่แตกต่างกันด้วย

8.3.2 การออกแบบและสร้างคลังข้อมูลจำลองการลักลอกงานวิชาการ

เมื่อกล่าวถึงแง่มุมต่างๆ ที่เกี่ยวข้องกับการออกแบบและสร้างคลังข้อมูลเพื่อใช้ในการฝึกฝนและทดสอบประสิทธิภาพของระบบตรวจหาการลักลอกแล้ว ผู้วิจัยมีข้อเสนอแนะในแง่มุมต่างๆ ดังต่อไปนี้

ในแง่ขนาดของคลังข้อมูล เป็นที่ทราบกันโดยทั่วไปว่าคลังข้อมูลที่มีขนาดใหญ่กว่าย่อมให้ประสิทธิภาพในงานที่เกี่ยวข้องได้สูงกว่า เช่นเดียวกันกับในกรณีของคลังข้อมูลจำลองการลักลอกงานวิชาการในงานวิจัยชิ้นนี้ คลังข้อมูลที่ใช้งานในแง่นี้ก็สมควรได้รับการขยายขนาดอยู่เสมอ ทั้งนี้เนื่องจากข้อมูลในโลกวิชาการนั้นไม่เคยหยุดนิ่ง ข้อมูลที่บรรจุในคลังข้อมูลแบบสถิตย์เช่นที่ใช้ในงานวิจัยชิ้นนี้ก็อาจพ้นสมัยได้ในวันหนึ่ง ดังนั้นแล้ว การปรับปรุงและขยายขนาดคลังข้อมูลอยู่เสมอจะช่วยให้สามารถพัฒนาระบบตรวจหาการลักลอกที่มีประสิทธิภาพได้ การเพิ่มข้อมูลเข้าในคลังข้อมูลอาจไม่จำเป็นต้องเป็นข้อมูลจากวิทยานิพนธ์เพียงเท่านั้น แต่อาจเลือกจากงานวิชาการประเภทอื่นๆ เช่น ตำราวิชาการ บทความวิชาการ บทความวิจัย หรือแม้กระทั่งข้อมูลจากสารานุกรม ก็สมควรนำมาบรรจุเข้าในคลังข้อมูล ยิ่งไปกว่านั้น งานวิจัยที่นำมาใช้ควรเลือกจากแหล่งที่หลากหลาย ทั้งนี้ ในอนาคตอาจสร้างความร่วมมือระหว่างสถาบันการศึกษาเพื่อสร้างคลังข้อมูลจำลองการลักลอกงานวิชาการรวมกัน เพื่อจะได้ใช้ประโยชน์ในการตรวจหาการลักลอกร่วมกันต่อไป

ในแง่ของสาขาวิชาของข้อมูลที่ประกอบเข้าเป็นคลังข้อมูล ผู้วิจัยเห็นว่าควรจัดแบ่งสาขาวิชาให้ละเอียดมากกว่าที่ปรากฏในงานวิจัยชิ้นนี้ การทำเช่นนี้จะช่วยข้อมูลในคลังข้อมูลมีความสมดุลมากขึ้น และจะนำไปสู่การปรับปรุงคลังข้อมูลให้ครอบคลุมถึงสาขาวิชาการที่หลากหลายมากขึ้น ทั้งนี้ เมื่อข้อมูลแต่ละสาขาวิชาเฉพาะมีมากพอแล้ว ก็อาจพัฒนาให้ระบบตรวจหาการลักลอกในสาขาวิชาเฉพาะได้

ในแง่ของการกำกับข้อมูลในคลังข้อมูล เพื่ออำนวยความสะดวกให้เกิดแก่ผู้จะนำคลังข้อมูลในงานวิจัยชิ้นนี้ไปพัฒนาต่อ ผู้วิจัยเห็นควรให้กำกับข้อมูลอื่นๆ เพิ่มเติมขึ้น ไม่ว่าจะเป็นการกำกับหมวดคำ ความสัมพันธ์แบบพึ่งพา บทบาททางความหมาย หรือความสัมพันธ์ทางปริจเฉท

และในแง่การตรวจสอบคลังข้อมูลก่อนนำไปใช้งานจริง ผู้วิจัยเห็นว่าควรมีทดลองจำแนกประเภทข้อมูลภายในคลังข้อมูลโดยแบบจำลองการเรียนรู้ของเครื่องแบบพื้นฐาน เช่น แบบจำลองนาอีฟเบย์ (Naive Bayes classifier) หรือการแบ่งกลุ่มข้อมูลแบบเคมีน (*k*-means clustering) ก่อน ทั้งนี้ เพื่อพิสูจน์ให้แน่ใจว่าในเบื้องต้น ข้อมูลแต่ละประเภทในคลังข้อมูลมีความแตกต่างกันเป็นกลุ่ม

หรือไม่ หากปรากฏว่าข้อมูลมีความแตกต่างกันเป็นกลุ่มแล้ว ก็จะช่วยให้แน่ใจได้ว่าคลังข้อมูลมีคุณภาพสามารถใช้ในการฝึกฝนจำแนกประเภทได้ต่อไป

8.3.3 หน่วยปริจเฉทพื้นฐานในฐานะข้อมูลรับเข้าของระบบตรวจหาการลักลอกงานวิชาการ

แม้ผลการทดสอบประสิทธิภาพของระบบเมื่อใช้ข้อมูลรับเข้าที่ต่างกันจะชี้ให้เห็นว่าระบบตรวจหาการลักลอกที่ใช้ย่อหน้าเป็นข้อมูลรับเข้ามีประสิทธิภาพที่เหนือกว่าระบบตรวจหาการลักลอกที่ใช้หน่วยปริจเฉทพื้นฐานเป็นข้อมูลรับเข้าก็ตาม แต่ผู้วิจัยเห็นว่าไม่ควรหยุดพัฒนาการใช้หน่วยปริจเฉทพื้นฐานในฐานะข้อมูลรับเข้าของระบบ สาเหตุที่กล่าวเช่นนี้เพราะหน่วยปริจเฉทพื้นฐานยังปรากฏข้อเด่นให้เห็นอยู่ 2 ประการ

ข้อเด่นประการแรกของหน่วยปริจเฉทพื้นฐานคือขนาดและขอบเขตที่มีความจำเพาะกว่าย่อหน้า ทั้งนี้ หากต้องการสร้างระบบตรวจหาการลักลอกที่มีความสมบูรณ์พร้อมใช้งานอย่างเต็มรูปแบบ ระบบดังกล่าวควรรายงานผลการตรวจหาได้ทั้งในเชิงปริมาณการลักลอกและข้อความที่ต้องสงสัยว่าเป็นการลักลอก ในกรณีของระบบที่ใช้ย่อหน้าเป็นข้อมูลรับเข้า การรายงานผลข้อความที่ต้องสงสัยว่าเป็นการลักลอกจะเป็นไปโดยลำบาก เพราะหน่วยพื้นฐานที่ระบบจะแสดงผลคือย่อหน้า แม้ว่าย่อหน้าดังกล่าวอาจปรากฏการลักลอกเพียงในระดับคำหรือวลี ระบบก็จะรายงานผลเป็นย่อหน้าที่มีการลักลอก แต่หากใช้หน่วยรับเข้าที่มีขนาดและขอบเขตที่จำเพาะเจาะจงกว่าเช่นหน่วยปริจเฉทพื้นฐาน การรายงานผลการตรวจหาการลักลอกของระบบ ก็จะสามารถเน้นให้ผู้ใช้งานเห็นถึงหน่วยปริจเฉทพื้นฐานที่ต้องสงสัยว่ามีการลักลอกได้ ในแง่นี้จึงเอื้อความสะดวกให้แก่ผู้ใช้งานมากกว่า

ข้อเด่นอีกประการหนึ่งของหน่วยปริจเฉทพื้นฐานคือ ในทางทฤษฎีแล้ว หน่วยปริจเฉทพื้นฐานย่อมปรากฏพร้อมกับวาทสัมพันธ์ที่มีระหว่างหน่วยปริจเฉทพื้นฐานอื่นๆ ในข้อความ หากในอนาคตสามารถกำกับวาทสัมพันธ์ดังกล่าวได้ ผู้วิจัยเชื่อว่าจะช่วยให้การตรวจหาการลักลอกเป็นไปอย่างมีประสิทธิภาพยิ่ง ด้วยแนวคิดนี้จะแทนรูปข้อความให้อยู่รูปโครงสร้างของปริจเฉทที่มีลักษณะเป็นลำดับชั้นสูงต่ำลดหลั่นกันไป ในแง่การตรวจหาการลักลอกอาจทำได้ด้วยการคำนวณระยะการแก้ไข (edit distance) ระหว่างลำดับชั้นที่แตกต่างกันในคู่ของข้อความที่ต้องสงสัยว่าลักลอก หรือการตรวจการลักลอกจะการแทนที่คำที่บ่งชี้วาทสัมพันธ์เดียวกันก็อาจเป็นไปได้

ด้วยข้อเด่นของหน่วยปริจเฉทพื้นฐานทั้งสองประการที่กล่าวมาข้างต้นนี้ ผู้วิจัยจึงเห็นควรให้พัฒนาเครื่องมือที่ใช้ในการกำกับขอบเขตของหน่วยปริจเฉทพื้นฐานและกำกับวาทสัมพันธ์ที่มีอยู่ระหว่างหน่วยปริจเฉทพื้นฐานให้มีความสมบูรณ์ เพื่อจะได้นำมาประยุกต์ใช้ในการตรวจหาการลักลอกในอนาคต

8.3.4 ลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก

จากการทดสอบประสิทธิภาพของลักษณ์ที่ใช้ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอก แม้ผลการทดสอบจะชี้ให้เห็นว่าลักษณ์ทางศัพท์และลักษณ์ทางอักขระจะเป็นลักษณ์ที่เอื้อให้ระบบจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกได้ดีกว่าลักษณ์ทางวากยสัมพันธ์และความหมายก็ตาม แต่การพัฒนาระบบตรวจหาการลักลอกต่อไปในอนาคต ผู้วิจัยเห็นว่าไม่ควรละเลยลักษณ์ทางภาษาทั้ง 2 ประเภทดังกล่าว

ในส่วนลักษณ์ทางวากยสัมพันธ์ ลักษณ์ประเภทหนึ่งที่ผู้วิจัยสนใจแต่ยังไม่สามารถสร้างได้สำเร็จคือลักษณ์ที่ได้จากการแทนรูปความสัมพันธ์แบบพึ่งพาในข้อความ ทั้งนี้ หากสร้างลักษณ์ชนิดนี้ได้สำเร็จ ผู้วิจัยเห็นว่าจะเอื้อให้เครื่องสามารถพิจารณารูปแบบความสัมพันธ์แบบพึ่งพาที่คงอยู่หรือเปลี่ยนแปลงไปอันเป็นผลจากการลักลอกได้ ในแง่นี้ ผู้วิจัยเชื่อว่าลักษณ์ดังกล่าวจะช่วยให้ระบบตรวจหาการลักลอกจะมีประสิทธิภาพในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกสูงขึ้น

ส่วนลักษณ์ทางความหมายนั้น ดังได้กล่าวไปแล้วว่า เป็นที่น่าเสียดายที่การสร้างลักษณ์จากเวกเตอร์ทางความหมายของคำโดยอาศัยแนวคิดเรื่องการฝังคำนั้นเป็นอันล้มเหลวในขั้นตอนการสร้าง จึงทำให้ไม่สามารถนำลักษณ์ค่าความละม้ายทางความหมายของเวกเตอร์ของคำมาใช้ในการทดสอบได้ อย่างไรก็ดี ผู้วิจัยเห็นควรให้พัฒนาวิธีการตรวจหาการลักลอกโดยอาศัยแนวคิดการฝังคำต่อไป เนื่องจากแนวคิดดังกล่าวสามารถคืนค่าที่เป็นรูปแทนในระดับความหมายได้โดยไม่ต้องพึ่งพาการตัดสินความหมายโดยมนุษย์ จึงช่วยลดปัญหาอันเกิดจากอัตวิสัยของมนุษย์ลงได้ ซึ่งลักษณะดังกล่าวนี้เป็นแนวทางที่ควรดำเนินไปในการสร้างนวัตกรรมในอนาคต

นอกจากลักษณ์ทางวากยสัมพันธ์และลักษณ์ทางความหมายที่ได้กล่าวไปแล้ว ยังมีลักษณ์ทางภาษาอีกประเภทหนึ่งที่งานวิจัยชิ้นนี้ไม่ได้กล่าวถึงเนื่องมาจากความไม่พร้อมของเครื่องมือที่ใช้ในการวิเคราะห์หาและสร้าง ลักษณ์ประเภทดังกล่าวคือลักษณ์ทางปริจเฉท ในอนาคต เมื่อเครื่องมือมีความพร้อมแล้ว ผู้วิจัยหวังว่าจะสามารถนำลักษณะของข้อความในระดับปริจเฉทมาใช้เป็นลักษณ์ได้ ไม่ว่าจะเป็นลักษณ์ที่แทนรูปความสัมพันธ์ภายในปริจเฉท หรือลักษณ์ที่อาศัยการพิจารณาคำเชื่อมในการบ่งชี้เจตนาของผู้เขียนหรือหน้าที่ทางปริจเฉทของข้อความได้ หากสร้างลักษณ์ประเภทนี้ได้สำเร็จ ระบบก็อาจมีประสิทธิภาพถึงในระดับที่สามารถตรวจจับเจตนาที่เปลี่ยนแปลงไปของผู้ลักลอกได้

8.3.5 การวัดค่าความละม้ายของข้อความ

แม้ผลจากการทดสอบประสิทธิภาพของวิธีการวัดค่าความละม้ายจะชี้ให้เห็นว่า วิธีการวัดค่าความละม้ายด้วยการหาค่าบรรทัดฐานของลำดับย่อยร่วมยาวสุดที่ยาวที่สุดของคำ ($lcs_{\text{norm-W}}$) เป็นวิธี

วัดค่าความละม้ายที่มีประสิทธิภาพใกล้เคียงกับการระบุค่าความละม้ายโดยมนุษย์มากที่สุด และเหมาะสมจะใช้แทนการระบุค่าความละม้ายโดยมนุษย์ในระบบตรวจหาการลักลอกงานวิชาการ เนื่องจากให้ผลค่าของความละม้ายสัมพันธ์สอดคล้องเป็นไปในทิศทางเดียวกับผู้เชี่ยวชาญมากที่สุด แต่หากพิจารณาค่าสหสัมพันธ์ระหว่างค่าความละม้ายที่ระบุโดยผู้เชี่ยวชาญกับค่าความละม้ายที่ได้จากวิธีการวัดค่าความละม้ายวิธีต่างในตารางที่ 7.2 แล้ว จะเห็นได้ว่าค่าสหสัมพันธ์ที่ปรากฏอยู่ในอันดับต้นๆ นั้นไม่ได้แตกต่างกันมากนัก กรณีเช่นนี้ ในการพัฒนาระบบตรวจหาการลักลอก หากต้องการลดความซับซ้อนของระบบ ผู้วิจัยเห็นว่าอาจใช้วิธีการวัดค่าความละม้ายที่เลือกใช้เป็นลักษณ์ที่ใช้ในขั้นตอนการจำแนกข้อความที่มีการลักลอกและไม่มีการลักลอกมาใช้วัดค่าความละม้ายของข้อความในส่วนหลังของระบบได้เช่นกัน ซึ่งวิธีการวัดค่าความละม้ายดังกล่าวตามที่ปรากฏเป็นผลในงานวิจัยชิ้นนี้คือค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2})

นอกจากนี้แล้ว จากผลการทดลอง จะเห็นได้ว่าวิธีการวัดค่าความละม้ายที่มีประสิทธิภาพตามข้อค้นพบของการวิเคราะห์ในขั้นนี้ประยุกต์ใช้แนวคิดการคำนวณในระดับคำศัพท์และอักขระด้วยเหตุนี้ จึงอาจกล่าวได้ว่า ในการพัฒนาระบบต้นแบบเพื่อตรวจหาการลักลอกงานวิชาการนั้น แนวคิดในการตรวจหาที่อิงรูปอักขระและรูปศัพท์ยังมีความสำคัญ ไม่ควรละเลย แต่ควรพัฒนาไปพร้อมกับวิธีการตรวจหาที่อิงจากความสัมพันธ์ทางภาษาในระดับที่ลึกกว่า

8.3.6 แนวทางการพัฒนาระบบตรวจหาการลักลอก

จากผลการวิจัยที่ได้กล่าวมาทั้งหมดข้างต้น สามารถยืนยันได้ว่าข้อมูลรับเข้าที่เหมาะสมจะใช้ในระบบที่สร้างขึ้นคือย่อหน้า ในขณะที่ค่าความละม้ายของข้อความที่คำนวณได้จากลักษณะทางคำและอักขระนั้นมีประสิทธิภาพโดดเด่น ทั้งในแง่การใช้เป็นลักษณ์ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน และในแง่การใช้วัดค่าความละม้ายของข้อความเพื่อบ่งชี้ปริมาณการลักลอกหลังจากการที่ได้รับการจำแนกประเภทแล้ว

อย่างไรก็ตาม เนื่องจากแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่นำมาใช้ทดลองในงานวิจัยชิ้นนี้มีการตั้งค่าพารามิเตอร์ตามค่าปริยาย ดังนั้นจึงมีความเป็นไปได้ว่าหากปรับเปลี่ยนค่าพารามิเตอร์แล้ว ก็อาจส่งผลให้เครื่องมีประสิทธิภาพสูงขึ้นหรือต่ำลงได้ โดยเฉพาะอย่างยิ่ง ในกรณีที่ใช้ข้อมูลรับเข้าหรือใช้ลักษณ์ชนิดอื่นในการฝึกฝนและทดสอบ ในแง่นี้ ผู้วิจัยเห็นควรให้ศึกษาวิจัยต่อไป เพื่อให้ได้ระบบตรวจหาการลักลอกโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนที่มีประสิทธิภาพมากที่สุด

ในอีกแง่หนึ่ง ด้วยประสิทธิภาพที่โดดเด่นของค่าความละม้ายของข้อความที่คำนวณได้จากลักษณะทางคำและอักขระดังได้กล่าวมาข้างต้น ผู้วิจัยจึงสนใจว่าหากพัฒนาระบบตรวจหาการลักลอกขึ้นโดยอิงเฉพาะค่าความละม้ายที่วัดได้จากข้อความในระดับคำและอักขระ และตัดขั้นตอนการ

จำแนกประเภทการลักลอกด้วยวิธีการเรียนรู้ของเครื่องออกไป ระบบจะสามารถให้คำตอบได้ถูกต้องในระดับใด

เพื่อพิสูจน์ข้อสังเกตข้างต้น ผู้วิจัยได้นำวิธีวัดค่าความละม้ายที่มีประสิทธิภาพดีที่สุด 3 อันดับแรกจากการใช้เป็นลักษณ์ในการจำแนกประเภทข้อความที่มีการลักลอกและไม่มีการลักลอกโดยเครื่อง ได้แก่ ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}), ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของคำ (J_{W2}), และค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของคำ (QS_{W3}) และวิธีวัดค่าความละม้ายที่มีประสิทธิภาพดีที่สุด 3 อันดับแรกจากการทดลองเปรียบเทียบประสิทธิภาพของวิธีการวัดค่าความละม้ายของข้อความ ได้แก่ ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของคำ ($lcs_{norm\text{-}W}$), ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ ($lcs_{norm\text{-}Char}$), และค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ (QS_{W1}) มาหาช่วงของค่าความละม้ายของข้อความที่มีการลักลอกและไม่มีการลักลอกจากข้อมูลคู่หน่วยเทียบทั้งหมดในคลังข้อมูล เพื่อที่ว่าอาจจะสามารถนำช่วงของค่าความละม้ายดังกล่าวมาใช้ประโยชน์ต่อในการพัฒนาระบบตรวจหาการลักลอกในอนาคต

ตารางที่ 8.1 แสดงค่าสถิติของค่าความละม้าย 6 ค่าที่วัดได้จากข้อมูลลักลอกและข้อมูลไม่ลักลอกในคลังข้อมูล ได้แก่ ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}), ค่าสัมประสิทธิ์ความละม้ายแจ็กการ์ดของไบแกรมของคำ (J_{W2}), ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไตรแกรมของคำ (QS_{W3}), ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของคำ ($lcs_{norm\text{-}W}$), ค่าบรรทัดฐานของลำดับย่อยร่วมที่ยาวที่สุดของอักขระ ($lcs_{norm\text{-}Char}$), และค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ (QS_{W1}) ทั้งนี้ หากพิจารณาเฉพาะช่วงของค่าความละม้ายของข้อมูลลักลอกและข้อมูลไม่ลักลอกแล้ว จะเห็นได้ว่าช่วงของค่าความละม้ายของข้อมูลทั้งสองประเภทค่อนข้างกว้าง (ค่าพิสัยสูง) และมีช่วงของค่าที่ทับซ้อนกันอยู่ระหว่างข้อมูลลักลอกกับข้อมูลไม่ลักลอก ด้วยลักษณะเช่นนี้อาจเป็นการยากหากจะนำมาใช้ทำนายการลักลอกในระบบตรวจหาการลักลอก อย่างไรก็ดี หากพิจารณาจากค่าเฉลี่ยแล้ว จะเห็นได้ว่าค่าเฉลี่ยของของข้อมูลทั้งสองประเภทแตกต่างกันมาก อีกทั้งเมื่อพิจารณาส่วนเบี่ยงเบนมาตรฐานประกอบด้วยแล้ว จะเห็นได้ว่าค่าดังกล่าวน้อยมาก แสดงให้เห็นว่าค่าความละม้ายทั้งจากข้อมูลที่มีการลักลอกและข้อมูลที่ไม่มีการลักลอกนั้นมีการกระจายต่ำ ในขณะที่ค่าพิสัยของค่าความละม้ายแต่ละกลุ่มค่อนข้างมาก ในแง่นี้หมายความว่ามีค่าความละม้ายที่มีความถี่ในการปรากฏต่ำอยู่ในทุกกลุ่ม ซึ่งในแง่การใช้เป็นลักษณ์ในการตรวจหาการลักลอกแล้ว ค่าความละม้ายที่มีความถี่ในการปรากฏต่ำนี้ก็อาจส่งผลให้เครื่องตรวจจับการลักลอกผิดพลาดด้วย

ตารางที่ 8.1 สถิติของค่าความละม้ายของคู่ลักลอกและคู่ไม่ลักลอก

ประเภทคู่หน่วยเทียบ	สถิติ	ค่าความละม้ายของข้อความ					
		QS_{W2}	J_{W2}	QS_{W3}	$lcs_{norm\text{-}W}$	$lcs_{norm\text{-}Char}$	QS_{W1}
ลักลอก	range[17]	0.1322 ถึง 1.0000 (0.8678)	0.0708 ถึง 1.0000 (0.9292)	0.1008 ถึง 1.0000 (0.8992)	0.3889 ถึง 1.0000 (0.6111)	0.4863 ถึง 1.0000 (0.5137)	0.1789 ถึง 1.0000 (0.8211)
	mean	0.9166	0.8575	0.8827	0.9544	0.9676	0.9545
	S.D.	0.0904	0.1386	0.1201	0.0595	0.0488	0.0560
ไม่ลักลอก	range	0.2105 ถึง 0.6580 (0.4475)	0.1176 ถึง 0.4903 (0.3727)	0.2000 ถึง 0.5000 (0.3000)	0.1546 ถึง 0.9434 (0.7888)	0.2095 ถึง 0.9544 (0.7449)	0.2273 ถึง 0.8571 (0.6298)
	mean	0.3830	0.2405	0.3025	0.4367	0.5279	0.5598
	S.D.	0.0870	0.0691	0.0809	0.1128	0.1071	0.0884

ด้วยเหตุผลดังได้กล่าวมาข้างต้น การพัฒนาระบบตรวจหาการลักลอกขึ้นโดยอิงเฉพาะค่าความละม้ายของข้อความและตัดขั้นตอนการจำแนกประเภทการลักลอกด้วยวิธีการเรียนรู้ของเครื่องออกไปก็อาจเป็นไปได้ แต่ต้องปรับช่วงของค่าความละม้ายดังกล่าวโดยพิจารณาตัดค่าความละม้ายที่มีความถี่ในการปรากฏต่ำออกไปก่อน ในกรณีนี้ ผู้วิจัยแนะนำให้ใช้ค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของยูนิแกรมของคำ (QS_{W1}) หรือค่าสัมประสิทธิ์ความละม้ายโซเรนเซน-ไดซ์ของไบแกรมของคำ (QS_{W2}) เนื่องจากเมื่อพิจารณาจากสถิติแล้ว วิธีวัดค่าความละม้ายดังกล่าวให้พิสัยของค่าความละม้ายทั้งในข้อมูลลักลอกและไม่ลักลอกอยู่ในเกณฑ์ที่รับได้ ในขณะเดียวกัน ค่าเบี่ยงเบนมาตรฐานของค่าความละม้ายที่ได้ทั้งในข้อมูลลักลอกและไม่ลักลอกยังต่ำที่สุดในภาพรวมเมื่อเปรียบเทียบกับวิธีการวัดค่าความละม้ายวิธีอื่นๆ ที่เหลือ

นอกจากนี้ อีกประเด็นหนึ่งที่ผู้วิจัยใคร่ขอเสนอเพื่อเป็นแนวทางในการพัฒนาระบบตรวจหาการลักลอกต่อไปคือประเด็นเรื่องการตรวจหาการลักลอกโดยอิงจากหน่วยปริจเฉทพื้นฐานหรือ

[17] ตัวเลขในวงเล็บคือค่าพิสัย

ประโยค ทั้งนี้ ดังได้กล่าวไปบ้างในหัวข้อที่ 8.3.3 ว่า ในการพัฒนาระบบตรวจหาการลักลอกเพื่อใช้งานจริงนั้นจำเป็นต้องรายงานทั้งปริมาณการลักลอกและข้อความในส่วยที่ระบุว่าเป็นการลักลอก หากการตรวจหาการลักลอกดำเนินการบนหน่วยรับเข้าประเภทย่อหน้า ค่าความละม้ายที่ได้และการรายงานผลข้อความส่วนที่ลักลอกก็จะบ่งชี้ในระดับย่อหน้า มิใช่ในระดับวลีหรือประโยคที่มีการลักลอกเกินขึ้น ด้วยเหตุนี้ ระบบการตรวจหาการลักลอกในอนาคตจึงควรพัฒนาให้มีการตรวจหาการลักลอกโดยอิงจากหน่วยปริจเฉทพื้นฐานหรือประโยค เพื่อในขั้นรายงานผลจะสามารถรายงานค่าความละม้ายในฐานะปริมาณบ่งชี้การลักลอกและรายงานผลส่วนของข้อความที่มีการลักลอกที่ได้อย่างจำเพาะจงเจาะ

รายการอ้างอิง

ภาษาไทย

กัญจนา บุณยเกียรติ และประไพพิศ มงคลรัตน์. (2554). *การลักลอกงานวิชาการและวรรณกรรม (Plagiarism)*. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

จิรวรรณ เจริญสุข. (2549). *การแบ่งขอบเขตอนุพากย์ปริจเฉทในภาษาไทยโดยใช้คำระบุนัยและข้อสนเทศเชิงวากยสัมพันธ์*. (วิทยานิพนธ์ปริญญามหาบัณฑิต), มหาวิทยาลัยเกษตรศาสตร์.

จุไรรัตน์ ลักษณะศิริ และบาหยัน อิ่มสำราญ. (2548). *ภาษากับการสื่อสาร*. นครปฐม: โครงการตำราและหนังสือคณะอักษรศาสตร์ มหาวิทยาลัยศิลปากร.

ณรงค์ บุญสิริสัมพันธ์. (2546). *ซัพพอร์ตเวกเตอร์แมชชีนแบบหลายประเภทโดยการแตกครึ่งแบบสมดุล*. (วิทยานิพนธ์ปริญญามหาบัณฑิต), จุฬาลงกรณ์มหาวิทยาลัย.

นลินี อินต๊ะซาว. (2556). *การแยกอนุพากย์ภาษาไทยด้วยการใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน*. (วิทยานิพนธ์ปริญญามหาบัณฑิต), จุฬาลงกรณ์มหาวิทยาลัย.

นัชชา ถิระสาโรช. (2553). การรู้จำชื่อเฉพาะภาษาไทย : *การใช้แบบจำลองคอนดิชันนอลแรนดอมฟิลด์ส*. (วิทยานิพนธ์ปริญญามหาบัณฑิต), จุฬาลงกรณ์มหาวิทยาลัย.

นัชชา ถิระสาโรช. (2559). *การศึกษาการแยกความหมายของคำหลายความหมายในภาษาไทยโดยใช้วิธีการวิเคราะห์ความหมายแอบแฝง*. (วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต), จุฬาลงกรณ์มหาวิทยาลัย.

นันทณัฐ พันธุ์สีดา. (2556). *การจำลองข้อมูลเพื่อเปรียบเทียบความแม่นยำในการพยากรณ์ ระหว่างวิธีโครงข่ายประสาทเทียมกับวิธีซัพพอร์ตเวกเตอร์แมชชีน*. (วิทยานิพนธ์ปริญญามหาบัณฑิต), จุฬาลงกรณ์มหาวิทยาลัย.

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย. (2554). *การคัดลอกผลงานทางวิชาการ ผลงานวิจัย สิ่งพิมพ์ วิทยานิพนธ์ (Academic Plagiarism) "ประเด็นที่เราควรตระหนัก"*. กทม.: จุฬาลงกรณ์มหาวิทยาลัย.

ปิยธิดา อินทร์รักษ์. (2552). *การประยุกต์ใช้การวิเคราะห์ความหมายแฝงกับการจำแนกประเภทอารมณ์ในข้อความภาษาไทย*. (วิทยานิพนธ์ปริญญามหาบัณฑิต), จุฬาลงกรณ์มหาวิทยาลัย.

พรพิลาส เรืองโชติวิทย์. (2545). *ความสัมพันธ์ระหว่างความหมายของคำกริยากับโครงสร้างประโยคในภาษาไทย* (รายงานการวิจัย). เชียงใหม่: ภาควิชาภาษาไทย คณะมนุษยศาสตร์ มหาวิทยาลัยเชียงใหม่.

ภาณุ สังขะวร. (2527). *ความสัมพันธ์ทางอรรถศาสตร์ระหว่างคำนามกับคำกริยาในประโยคภาษาไทย*. (วิทยานิพนธ์ปริญญามหาบัณฑิต), จุฬาลงกรณ์มหาวิทยาลัย.

ราชบัณฑิตยสถาน. (2545). *ศัพท์บัญญัติราชบัณฑิตยสถาน*. ค้นคืนเมื่อ 11 กันยายน 2555 http://rirs3.royin.go.th/coinages/webcoinage.php

วิไลวรรณ ศรีสงคราม. (2554). *การพัฒนาความรู้ความเข้าใจเรื่องการลอกเลียนวรรณกรรมของนักศึกษาระดับปริญญาตรีบนพื้นฐานผลการวิจัยสำรวจและการวิเคราะห์เอกสาร*. (วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต), จุฬาลงกรณ์มหาวิทยาลัย.

ศิวพร ทวนไธสง. (2556). *การตรวจเทียบภายในหาการลักลอกงานวิชาการภาษาไทยโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน*. (วิทยานิพนธ์ปริญญามหาบัณฑิต), จุฬาลงกรณ์มหาวิทยาลัย.

ศุภวัจน์ แต่รุ่งเรือง และวิโรจน์ อรุณมานะกุล. (2558). กลวิธีลักลอกงานวิชาการภาษาไทย: การวิเคราะห์ทางภาษาศาสตร์. *ภาษาและภาษาศาสตร์, 34*(1), 38-65.

อมรา ประสิทธิ์รัฐสินธุ์. (2542). *ภาษาในสังคมไทย : ความหลากหลาย การเปลี่ยนแปลง และการพัฒนา* (พิมพ์ครั้งที่ 2). กรุงเทพฯ: โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

<u>ภาษาอังกฤษ</u>

Achananuparp, P., Hu, X., & Shen, X. (2008). *The Evaluation of Sentence Similarity Measures*. Paper presented at the 10th international conference on Data Warehousing and Knowledge Discovery, Turin, Italy.

Agirre, E., Cer, D., Diab, M., & Gonzalez-Agirre, A. (2012). *SemEval*-2012 *task* 6: *A pilot on semantic textual similarity*. Paper presented at the 6th international workshop on semantic evaluation (SemEval-2012), Montréal, Canada.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., & Guo, W. (2013). **SEM* 2013 *shared task: Semantic textual similarity, including a pilot on typed-similarity*. Paper presented at the second joint conference on lexical and computational semantics (*SEM), Atlanta, GA.

Agirrea, E., Baneab, C., Cardiec, C., Cerd, D., Diabe, M., Gonzalez-Agirrea, A., . . . Wiebe, J. (2014). *SemEval*-2014 *task* 10: *Multilingual semantic textual similarity*. Paper presented at the 8th international workshop on semantic evaluation (SemEval 2014), Dublin, Ireland.

Aliguliyev, R. M. (2009). A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Syst. Appl.,* 36(4), 7764-7772. doi: 10.1016/j.eswa.2008.11.022

Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 42(2), 133-149. doi: 10.1109/TSMCC.2011.2134847

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of artificial intelligence research,* 38, 135-187.

Aroonmanakun, W. (2002). Collocation and Thai Word Segmentation. In T. Theeramunkong & V. Sornlertlamvanich (Eds.), *Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop* (pp. 68-75). Pathumthani: Sirindhorn International Institute of Technology.

Atkinson-Abutridy, J., Mellish, C., & Aitken, S. (2004). Combining information extraction with genetic algorithms for text mining. *IEEE Intelligent Systems,* 19(3), 22-30.

Barrón-Cedeño, A., Basile, C., Degli Esposti, M., & Rosso, P. (2010). *Word Length n-Grams for Text Re-use Detection.* Paper presented at the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2010), Berlin, Heidelberg.

Barrón-Cedeño, A., Rosso, P., Agirre, E., & Labaka, G. (2010). *Plagiarism Detection across Distant Language Pairs*. Paper presented at the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing.

Barrón-Cedeño, A., Vila, M., Martí, M. A., & Rosso, P. (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational linguistics,* 39(4), 917-947. doi: 10.1162/COLI_a_00153

Basile, C., Matematica, D., Benedetto, D., Caglioti, E., Cristadoro, G., & Esposti, M. D. (2009). *A plagiarism detection procedure in three steps: Selection, Matches and "Squares"*. Paper presented at the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09), Donostia, Spain.

Behrens, L., & Rosen, L. J. (2008). *Writing and reading across the curriculum* (10th ed.). New York: Pearson Longman.

Bhagat, R., & Hovy, E. (2013). What is a paraphrase? *Computational linguistics,* 39(3), 463-472.

Bretag, T., & Mahmud, S. (2009). A model for determining student plagiarism: Electronic detection and academic judgement. *Journal of University Teaching & Learning Practice,* 6(1), 49-60.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational linguistics,* 21(4), 543-565.

Brockett, C., & Dolan, B. (2005). *Support vector machines for paraphrase identification and corpus construction*. Paper presented at the third International Workshop on Paraphrasing (IWP2005), Jeju island, Korea.

Bruce, R., & Wiebe, J. (1994). *Word-sense disambiguation using decomposable models*. Paper presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Las Cruces, New Mexico.

Carlson, L., & Marcu, D. (2001). *Discourse tagging reference manual*. CA: University of Southern California Information Sciences Institute.

Carlson, L., Marcu, D., & Okurowski, M. E. (2001). *Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory*. Paper presented at the Second SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology,* 2(3), 1-27. doi: 10.1145/1961189.1961199

Cheema, W. A., Najib, F., Ahmed, S., Bukhari, S. H., Sittar, A., & Nawab, R. M. A. (2015). *A Corpus for Analyzing Text Reuse by People of Different Groups—Notebook for PAN at CLEF* 2015. Paper presented at the CLEF 2015 Evaluation Labs and Workshop, Toulouse, France.

Chong, M., Specia, L., & Mitkov, R. (2010). *Using natural language processing for automatic detection of plagiarism*. Paper presented at the 4th international plagiarism conference, Newcastle-upon-Tyne, UK.

Chulalongkorn University Language Institute. (n.d.). How to paraphrase. Retrieved October 7, 2011, from http://www.culi.chula.ac.th/expeng/howtoparaphrase/index.htm

Clough, P., & Stevenson, M. (2009). *Creating a corpus of plagiarised academic texts*. Paper presented at the Corpus Linguistics Conference (CL2009), University of Liverpool, UK

Clough, P., & Stevenson, M. (2011). Developing a corpus of plagiarised short answers. *Language Resources and Evaluation,* 45(1), 5-24. doi: 10.1007/s10579-009-9112-1

Clouse, B. F. (2008). *The student writer : editor and critic* (7th ed.). Boston: McGraw-Hill.

Crusius, T. W., & Channell, C. E. (2006). *The aims of argument: a brief guide* (5th ed.). Boston: McGraw-Hill.

De Boni, M., & Manandhar, S. (2003). *The use of sentence similarity as a semantic relevance metric for question answering*. Paper presented at the AAAI symposium on new directions in question answering, Stanford University.

De Marneffe, M.-C., Silveira, T. N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). *Universal Stanford Dependencies: a Cross-Linguistic Typology*. Paper presented at the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology,* 26(3), 297-302. doi: 10.2307/1932409

Dolan, B., Quirk, C., & Brockett, C. (2004). *Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources*. Paper presented at the 20th international conference on Computational Linguistics, Geneva, Switzerland.

Elhadi, M., & Al-Tobi, A. (2008). *Use of text syntactical structures in detection of document duplicates*. Paper presented at the third International Conference on Digital Information Management, London.

Feng, J., Zhou, Y., & Martin, T. (2008). *Sentence Similarity based on Relevance.* Paper presented at the IPMU 2008, 2008.

Fernando, S., & Stevenson, M. (2008). *A semantic similarity approach to paraphrase detection*. Paper presented at the Computational Linguistics UK (CLUK. 2008) 11th Annual Research Colloquium, Oxford, UK.

Finch, A., Hwang, Y.-S., & Sumita, E. (2005). *Using machine translation evaluation techniques to determine sentence-level semantic equivalence.* Paper presented at the 3rd International Workshop on Paraphrasing (IWP2005), Jeju Island, Korea.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.,* 20(1), 116-131. doi: 10.1145/503104.503110

Frick, T. (2005). How to recognize plagiarism. Retrieved October 16, 2011, from https://www.indiana.edu/~istd/example5paraphrasing.html

Fujita, A. (2005). *Automatic generation of syntactically well-formed and semantically appropriate paraphrases.* (Doctoral dissertation), Nara Institute of Science and Technology, Nara, Japan.

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing,* 21(2), 137-146. doi: 10.1007/s11222-009-9153-8

Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P., & Piao, S. (2001). *The METER Corpus: A corpus for analysing journalistic text reuse.* Paper presented at the Corpus Linguistics 2001 conference, Lancaster University, UK.

Ganguly, D., Leveling, j., & Jones, G. (2011). *Query expansion for language modeling using sentence similarities.* Paper presented at the 2nd information retrieval facility (IRF) conference, Vienna, Austria.

Gipp, B., Meuschke, N., & Beel, J. (2011). *Comparative evaluation of text- and citation-based plagiarism detection approaches using Guttenplag*. Paper presented at the 11th annual international ACM/IEEE joint conference on Digital libraries, Ottawa, Ontario, Canada.

Glau, G. R., & Jacobsen, C. B. (2001). *Scenarios for writing: issues, analysis, and response*. Boston: McGraw-Hill.

Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications,* 68(13), 13-18.

Gould, E., DiYanni, R., & Smith, W. (1989). *The act of writing*. New York: Random House.

Grozea, C., Gehl, C., & Popescu, M. (2009). *ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection*. Paper presented at the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09), Donostia, Spain.

Heffernan, J. A. W., & Lincoln, J. E. (1982). *Writing: a college handbook*. Toronto: W.W.Norton & company.

Henry, J. A. (Ed.) (1971) The compact edition of the Oxford English dictionary. Oxford: Oxford University Press.

Intasaw, N., & Aroonmanakun, W. (2013). *Basic principles for segmenting Thai EDUs*. Paper presented at the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27), Taipei, Taiwan.

Islam, A., & Inkpen, D. Z. (2009). Semantic similarity of short texts. In N. Nicolov, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing V: Selected papers from RANLP* 2007 (pp. 227-236). Amsterdam, Netherlands: John Benjamins Publishing Company.

Jaccard, P. (1908). Nouvelles Recherches Sur La Distribution Florale. *Bulletin de la Société vaudoise des Sciences Naturelles,* 44, 223-270.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist,* 11, 37-50. doi: 10.1111/j.1469-8137.1912.tb05611.x

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, N.J.: Pearson Prentice Hall.

Kauchak, D., & Barzilay, R. (2006). *Paraphrasing for automatic evaluation*. Paper presented at the Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York, New York.

Ketui, N., Theeramunkong, T., & Onsuwan, C. (2012). *A rule-based method for Thai Elementary Discourse Unit Segmentation (TED-Seg)*. Paper presented at the 7th International Conference on Knowledge Information and Creativity Support Systems (KICSS 2012), Melbourne, Australia.

Ketui, N., Theeramunkong, T., & Onsuwan, C. (2015). An EDU-based approach for Thai multi-document summarization and its application. *ACM Transactions on Asian and Low-Resource Language Information Processing,* 14(1), 4:1-4:26. doi: http://dx.doi.org/10.1145/2641567

Kotsiantis, S. B. (2007). Supervised machine learning: a review of classification techniques. *Informatica,* 31(3), 249-268.

Kozareva, Z., & Montoyo, A. (2006). *Paraphrase Identification on the Basis of Supervised Machine Learning Techniques.* Paper presented at the 5th International Conference on NLP, FinTAL 2006, Turku, Finland.

Kozlowski, R., McCoy, K. F., & Vijay-Shanker, K. (2003). *Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources.* Paper presented at the the second international workshop on paraphrasing, Sapporo, Japan.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association,* 47(260), 583-621. doi: 10.1080/01621459.1952.10483441

Leacock, C., Towell, G., & Voorhees, E. (1993). *Corpus-based statistical sense resolution*. Paper presented at the Proceedings of the workshop on Human Language Technology, Princeton, New Jersey.

Lee, C., Hwang, Y.-G., Oh, H.-J., Lim, S., Heo, J., Lee, C.-H., . . . Jang, M.-G. (2006). *Fine-Grained Named Entity Recognition Using Conditional Random Fields for Question Answering*, Berlin, Heidelberg.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR,* 163(4), 845-848.

Li, L., Wu, Y., & Ye, M. (2015). Experimental Comparisons of Multi-class Classifiers. *Informatica,* 39(1), 71-85.

Li, Y., Bandar, Z., McLean, D., & O'Shea, J. (2004). *A method for measuring sentence similarity and its application to conversational agents*. Paper presented at the 17th International Florida Artificial Intelligence Research Society (FLAIRS) Conference, Florida, USA.

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on knowledge and data engineering,* 18(8), 1138-1150. doi: 10.1109/TKDE.2006.130

Liang, X., Wang, D., & Huang, M. (2010). *Improved Sentence Similarity Algorithm based on VSM and its application in Question Answering System*. Paper presented at the 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems, Xiamen.

Lin, C.-Y., & Hovy, E. (2003). *Automatic evaluation of summaries using N-gram co-occurrence statistics*. Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, Edmonton, Canada.

Lin, D. (1998). *An Information-Theoretic Definition of Similarity*. Paper presented at the Fifteenth International Conference on Machine Learning, Madison, Wisconsin, USA.

Liu, Y., & Zong, C. (2004). *Example-based Chinese-English MT*. Paper presented at the 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), The Hague, Netherlands.

Lunsford, R. F., & Bridges, B. (2005). *The Longwood guide to writing* (3rd ed.). New York: Longman.

Madnani, N., Tetreault, J., & Chodorow, M. (2012). *Re-examining machine translation metrics for paraphrase identification*. Paper presented at the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, Canada.

Malakasiotis, P. (2009). *Paraphrase recognition using machine learning to combine similarity measures*. Paper presented at the Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, Suntec, Singapore.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text,* 8(3), 243-281.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

Marius, R., & Wiener, H. S. (1994). *The McGraw-Hill college handbook* (4th ed.). New York: McGraw-Hill.

Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism - A Survey. *Journal of Universal Computer Science,* 12(8), 1050-1084.

Metzler, D., Dumais, S., & Meek, C. (2007). *Similarity measures for short segments of text*. Paper presented at the 29th European conference on IR research, Rome, Italy.

Meyer zu Eissen, S., Stein, B., & Kulig, M. (2007). *Plagiarism Detection Without Reference Collections*, Berlin, Heidelberg.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). *Corpus-based and knowledge-based measures of text semantic similarity*. Paper presented at the Proceedings of the 21st national conference on Artificial intelligence - Volume 1, Boston, Massachusetts.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. Paper presented at the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada.

Mohtaj, S., Asghari, H., & Zarrabi, V. (2015). *Developing monolingual English corpus for plagiarism detection using human annotated paraphrase corpus*. Paper presented at the Conference and Labs of the Evaluation Forum (CLEF 2015), Toulouse, France.

Mulvaney, M. K., & Jolliffe, D. A. (2005). *Academic writing: genres, samples, and resources*. New York: Pearson Longman.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Comput. Surv.,* 33(1), 31-88. doi: 10.1145/375360.375365

Nolan, R. (1970). *Foundations for an adequate criterion of paraphrase*. Paris: Mouton.

Palmer, M., Fellbaum, C., Cotton, S., Delfs, L., & Dang, H. T. (2001). *English tasks: all-words and verb lexical sample*. Paper presented at the The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, France.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation.* Paper presented at the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania.

Park, C. (2003). In Other (People's) Words: Plagiarism by university students--literature and lessons. *Assessment & Evaluation in Higher Education,* 28(5), 471-488. doi: 10.1080/02602930301677

Park, E.-K., Ra, D.-Y., & Jang, M.-G. (2005). Techniques for improving web retrieval effectiveness. *Inf. Process. Manage.,* 41(5), 1207-1223. doi: 10.1016/j.ipm.2004.08.002

Pecorari, D. (2008). *Academic writing and plagiarism: a linguistic analysis*. London: Continuum.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). *WordNet::Similarity: measuring the relatedness of concepts*. Paper presented at the Demonstration Papers at HLT-NAACL 2004, Boston, Massachusetts.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* 12, 2825-2830.

Petrov, S., Das, D., & McDonald, R. (2012). *A Universal Part-of-Speech Tagset*. Paper presented at the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.

Phucharasupa, K., & Netisopakul, P. (2012). *Classification of Thai sentence paraphrase*. Paper presented at the joint international symposium on natural language processing and agricultural ontology service 2011 (SNLP-AOS 2011), Bangkok, Thailand.

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation,* 45(1), 45-62. doi: 10.1007/s10579-009-9114-z

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). *Overview of the 3rd international competition on plagiarism detection.* Paper presented at the CLEF 2011 Labs and Workshops, Amsterdam.

Potthast, M., Hagen, M., Völske, M., & Stein, B. (2013). *Crowdsourcing interaction logs to understand text reuse from the web*. Paper presented at the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria.

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). *An evaluation framework for plagiarism detection*. Paper presented at the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.

Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., & Rosso, P. (2009). Overview of the 1st International Competition on Plagiarism Detection. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, & E. Agirre (Eds.), *SEPLN* 09 *Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN* 09) (pp. 1-9). Valencia, Spain: CEUR-WS.org.

Qiu, L., Kan, M.-Y., & Chua, T.-S. (2006). *Paraphrase recognition via dissimilarity significance classification*. Paper presented at the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia.

Ronald, A., & Suharjito. (2014). Plagiarism detection algorithm using natural language processing based on grammar analyzing. *Journal of Theoretical and Applied Information Technology,* 63(1), 168-180.

Ross, C., & Thomas, A. (2003). *Writing for real : A handbook for writers in community service*. New York: Longman.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Commun. ACM,* 8(10), 627-633. doi: 10.1145/365628.365657

Rus, V., McCarthy, P. M., Lintean, M. C., McNamara, D. S., & Graesser, A. C. (2008). *Paraphrase identification with lexico-syntactic graph subsumption*. Paper presented at the 21th International Florida Artificial Intelligence Research Society Conference (FLAIRS-21), Coconut Grove, FL.

Sharjeel, M., Rayson, P., & Nawab, R. M. A. (2016). *UPPC - Urdu paraphrase plagiarism corpus*. Paper presented at the Language Resource and Evaluation Conference (LREC) 2016, Portorož, Slovenia.

Sindhu.L, Thomas, B. B., & Idicula, S. M. (2011). A Study of Plagiarism Detection Tools and Technologies. *IJART,* 1(1), 64-70.

Sinthupoun, S., & Sornil, O. (2010). Thai rhetorical structure analysis. *International Journal of Computer Science and Information Security (IJCSIS),* 7(1), 95-105.

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., & Manning, C. D. (2011). *Dynamic pooling and unfolding recursive autoencoders for paraphrase detection*. Paper presented at the 24th International Conference on Neural Information Processing Systems, Granada, Spain.

Sørensen, T. J. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab,* 5(4), 1-34.

Sousa-Silva, R., Grant, T., & Maia, B. (2010). *'I didn't mean to steal someone else's words!': A Forensic Linguistic Approach to Detecting Intentional Plagiarism*. Paper presented at the 4th international plagiarism conference, Newcastle.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation,* 28(1), 11-21. doi: 10.1108/eb026526

Spatt, B. (1987). *Writing from sources* (2nd ed.). New York: St. Martin's.

Sriganesh, V., & Iyer, P. (2007). Plagiarism and medical writing. *Indian Journal of Radiology and Imaging,* 17(3), 146-147. doi: 10.4103/0971-3026.34716

Stein, B., Lipka, N., & Prettenhofer, P. (2011). Intrinsic plagiarism analysis. *Lang. Resour. Eval.,* 45(1), 63-82. doi: 10.1007/s10579-010-9115-y

Sukvaree, T., Charoensuk, J., Wattanamethanont, M., & Kultrakul, A. (2004). *RST based text summarization with ontology driven in agriculture domain*. Bangkok: Department of Computer Engineering, Kasetsart University.

Sutherland-Smith, W. (2008). *Plagiarism, the internet, and student learning: improving academic integrity*. New York: Routledge.

Taboada, M., & Mann, W. C. (2006). Applications of Rhetorical Structure Theory. *Discourse Studies,* 8(4), 567-588.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). *Feature-rich part-of-speech tagging with a cyclic dependency network*. Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for

Computational Linguistics on Human Language Technology - Volume 1, Edmonton, Canada.

Tsatsaronis, G., Varlamis, I., Giannakoulopoulos, A., & Kanellopoulos, N. (2010). *Identifying free text plagiarism based on semantic similarity*. Paper presented at the 4th International Plagiarism Conference (IPC 2010), Newcastle.

Tseng, H., Jurafsky, D., & Manning, C. D. (2005). *Morphological features help POS tagging of unknown words across language varieties.* Paper presented at the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea

Turian, J., Ratinov, L., & Bengio, Y. (2010). *Word representations: a simple and general method for semi-supervised learning*. Paper presented at the Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden.

Vila, M., Martí, M. A., & Rodríguez, H. (2011). Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural,* 46, 83-90.

Wan, S., Dras, M., Dale, R., & Paris, C. (2006). *Using dependency-based features to take the "para-farce" out of paraphrase.* Paper presented at the 2006 Australasian Language Technology Workshop (ALTW2006), Sancta Sophia College, Sydney.

Warn, J. (2007). Plagiarism software: no magic bullet! *Higher Education Research & Development,* 25(2), 195-208. doi: 10.1080/07294360600610438

Wattanamethanont, M., Suvakree, T., & Kawtrakul, A. (2005). *Discourse relation recognition by using Naïve Bayesian classifier*. Paper presented at the 9th National Computer Science and Engineering Conference (NCSEC 2005), University of the Thai Chamber of Commerce (UTCC), Bangkok, Thailand.

Yang, C., & Wen, J. (2007). *Text categorization based on a similarity approach*. Paper presented at the International conference on intelligence systems and knowledge engineering (ISKE), Chengdu, China.

Zhang, P.-y., & Li, C.-h. (2009). *Automatic text summarization based on sentences clustering and extraction*. Paper presented at the 2nd IEEE international conference on computer science and information, Beijing, China.

Zhang, Y., & Patrick, J. (2005). *Paraphrase Identification by Text Canonicalization*. Paper presented at the the Australasian Language Technology Workshop 2005, Sydney, Australia.

ภาคผนวก

ภาคผนวก ก

ตัวอย่างแบบสอบถามสำหรับเก็บข้อมูลการลักลอกงานวิชาการภาษาไทย

แบบสอบถามชุดนี้จัดทำขึ้นเพื่อเก็บข้อมูลเพื่อใช้ในการทำวิทยานิพนธ์เรื่อง “การตรวจเทียบภายนอกหาการลักลอกในงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความ” ของนายศุภวัจน์ แต่รุ่งเรือง นิสิตระดับดุษฎีบัณฑิต สาขาวิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ผู้วิจัยใคร่ขอความร่วมมือจากท่าน กรุณาตอบแบบสอบถามให้สมบูรณ์ ข้อมูลทั้งหมดที่ท่านตอบจะเป็นประโยชน์อย่างยิ่งสำหรับงานวิจัยครั้งนี้ ทั้งนี้ ผู้วิจัยขอขอบคุณท่านมา ณ ที่นี้ด้วย

<u>ตอนที่ 1</u> : สถานภาพทั่วไปของผู้ตอบแบบสอบถาม

<u>คำชี้แจง</u> : โปรดกรอกข้อมูลส่วนตัวของท่าน โดยทำเครื่องหมาย ✓ ลงในช่องที่กำหนดให้ต่อไปนี้

1. ระดับการศึกษาสูงสุด ☐ ต่ำกว่าปริญญาตรี ☐ ปริญญาตรี ☐ ปริญญาโท ☐ ปริญญาเอก
2. สถานภาพการศึกษา ☐ สำเร็จการศึกษาแล้ว ☐ กำลังศึกษา

<u>ตอนที่ 2</u> : การลักลอกงานวิชาการ

<u>คำชี้แจง</u> : ในส่วนต่อไปนี้ ผู้วิจัยได้เตรียมข้อความต้นฉบับไว้จำนวน 5 ข้อความ สมมติให้ท่านเป็นนิสิตที่กำลังทำรายงานส่งอาจารย์ ให้ท่านนำข้อความต้นฉบับมาเขียนขึ้นใหม่ให้สื่อความหมายได้ใกล้เคียงกับข้อความต้นฉบับมากที่สุด ทั้งนี้ ท่านต้องระมัดระวังมิให้อาจารย์สงสัยหรือตรวจสอบได้ว่าท่านลอกข้อความต้นฉบับมา

ข้อความ 1/5

มลพิษทางอากาศภายในอาคาร บ้านเรือน เป็นปัญหาหนึ่งที่สำคัญต่อมนุษย์ เนื่องจากมนุษย์ใช้เวลาส่วนใหญ่อยู่ในอาคาร บ้านเรือน หรือ สถานที่ทำงาน พบว่ามนุษย์โดยเฉพาะคนในเมืองใหญ่ใช้เวลาประมาณ 89% ในอาคาร บ้านเรือน จึงไม่ใช่เรื่องที่น่าประหลาดใจ ถ้าการได้รับมลพิษในอากาศของประชากรในเขตเมืองจะเกิดขึ้นภายในอาคารมากกว่าที่เกิดขึ้นขณะดำเนินกิจกรรมอยู่ภายนอกอาคาร จากข้อเท็จจริงดังกล่าวทำให้องค์กรพิทักษ์สิ่งแวดล้อมของประเทศสหรัฐอเมริกากำหนดให้ความเสี่ยงทางสุขภาพของมนุษย์อันเนื่องมาจากคุณภาพอากาศภายในอาคารอยู่ใน 5 อันดับแรกของความเสี่ยงทางสุขภาพเนื่องจากสภาวะแวดล้อมด้านต่าง ๆ

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

ข้อความ 2/5

ป่าชายเลนมีความสำคัญทางด้านอนุรักษ์พื้นที่ชายฝั่งทะเล โดยทำหน้าที่เป็นปราการตามธรรมชาติป้องกันลมพายุ ป้องกันชายฝั่งไม่ให้ถูกกัดเซาะจากกระแสคลื่น ช่วยในการรักษาคุณภาพสิ่งแวดล้อม โดยช่วยดักกรองของเสีย และขยะบริเวณชายฝั่งก่อนลงสู่ทะเล นอกจากนี้ป่าชายเลนยังเป็นแหล่งกักเก็บคาร์บอนที่มีความสำคัญอีกด้วย

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

ข้อความ 3/5

เงินเป็นสื่อกลางที่ใช้ในการแลกเปลี่ยนสินค้าและบริการ เงินเปรียบเสมือนปัจจัยหล่อลื่นเศรษฐกิจของประเทศ การพัฒนาประเทศจะดำเนินไปด้วยดีและความมั่นคง ย่อมขึ้นอยู่กับการมีปริมาณเงินและเครดิตในปริมาณที่เหมาะสม ถ้าในขณะใดขณะหนึ่งมีปริมาณเงินน้อยเกินไปก็จะก่อให้เกิดปัญหาเงินฝืด แต่ถ้ามีมากเกินไปก็จะก่อให้เกิดปัญหาเงินเฟ้อ ซึ่งมีผลเสียต่อเศรษฐกิจของประเทศทั้งสิ้น และผลเสียดังกล่าวก็จะเกิดผลกระทบต่อธุรกิจในประเทศเช่นกัน ด้วยเหตุผลดังกล่าว ธนาคารกลางซึ่งเป็นสถาบันการเงินที่สำคัญของประเทศที่มีอำนาจหน้าที่เกี่ยวข้องกับระบบการเงินและเครดิตของประเทศจึงต้องเข้ามาเกี่ยวข้อง โดยใช้นโยบายการเงินเข้าควบคุมปริมาณเงินและเครดิตของประเทศให้มีปริมาณที่เหมาะสมตามสถานการณ์

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

ข้อความ 4/5

ความรุนแรงในสังคมและความรุนแรงในครอบครัวมีความคล้ายคลึงกันที่เหยื่อ คือผู้อ่อนแอและมีสถานภาพต่ำกว่า เช่น สตรี เด็ก ฯลฯ แต่ผลกระทบของความรุนแรงในครอบครัวรุนแรงและยาวไกลกว่าความรุนแรงในสังคม เพราะนอกจากจะเป็นสาเหตุสำคัญของปัญหาครอบครัวแตกแยกและนำมาซึ่งปัญหาสังคมอีกมากมายแล้ว ความรุนแรงในครอบครัวยังเป็นมรดกถ่ายทอดไปลูกหลานต่อ ๆ ในอนาคตอย่างไม่รู้จบสิ้น

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

ข้อความ 5/5

ในปัจจุบันความสำคัญในการใช้สุนัขล่าสัตว์ และช่วยหาอาหารนั้นได้ลดความสำคัญลงไป แต่มนุษย์จะเลี้ยงสุนัขเพื่อใช้เป็นเพื่อน และใช้ประโยชน์อย่างอื่น ที่นอกเหนือไปจากการเฝ้าบ้าน การป้องกันขโมย ใช้ต้อนฝูงสัตว์ หรือเลี้ยงเพื่อการค้า ในบางครั้งสุนัขที่ได้รับการฝึกหัด สามารถจะใช้ในงานสะกดรอยติดตามจับตัวผู้ร้าย ดมกลิ่นระเบิด ตรวจค้นหาสารเสพติด เป็นต้น

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

ภาคผนวก ข
รายการคำและปริบทสำหรับใช้แทรกและลบคำ
ในการสร้างข้อมูลลักลอกประเภทคัดลอกโดยใกล้เคียง

รายการคำและปริบทสำหรับใช้แทรก

คำที่	คำ	ปริบท
1	ก็	และความจริง_____มีอย่างหลากหลาย
2	กับ	เผชิญ_____ปัญหา
3	การ	หลัง_____ปลูกถ่ายไต
4	การ	วิธี_____ดำน้ำตื้น
5	การ	พฤติกรรม_____บริโภคอาหาร
6	เกี่ยวกับ	ข้อมูล_____การแสดงละครหลวงนฤมิตร
7	ขอ	ในบทนี้จะ_____กล่าวถึง
8	ของ	ในด้าน_____
9	ของ	ในส่วน_____
10	ของ	ในแง่_____
11	ของ	ในมุม_____
12	ของ	อาคารสำนักงานราชการ_____สีเขียว
13	ของ	สถานภาพ_____ผู้หญิง
14	ของ	วัฏจักรชีวิต_____สารสนเทศ
15	ข้อต่อ	การเคลื่อนไหวผ่าน_____หลายข้อต่อ
16	ข้างต้น	ดังกล่าว_____
17	ข้างต้น	จากข้อสรุป_____
18	ขึ้น	เกิด_____จาก
19	เขา	ดาบที่_____ใช้เป็นอาวุธ
20	เข้า	_____ร่วมกิจกรรม
21	คอย	หาผู้เล่น_____เฝ้ากระป๋อง
22	ค่อย	ถ้าพบความแตกต่างจึง_____เปรียบเทียบความแตกต่าง
23	คิด	ห่อ _____เป็นเงินมูลค่า
24	แค่	_____เพียงทั้งหมด
25	จะ	_____เห็นได้ว่า
26	จะ	โดยรวม_____พบว่า
27	จะ	มัก_____มี
28	จะ	ควร_____เลือกใช้เม็ดมีดแบบเคลือบผิว
29	จะ	_____ได้ผลิตภัณฑ์เป็น
30	จะ	ผู้วิจัย_____ใช้
31	จาก	_____ผลการวิจัยพบว่า
32	จาก	_____ผลการวิเคราะห์พบว่า
33	จึง	ต่อไป_____เป็นขั้นลงมือปฏิบัติงาน
34	จึง	ดังนั้น_____ทำให้การทดสอบสมมติฐาน
35	ชั้น	เด็ก_____อนุบาล
36	ชิ้น	งานวิจัย_____นี้
37	เช่น	โดยสภาพ_____นี้
38	ซึ่ง	_____คิดเป็นร้อยละ
39	ซึ่ง	ผู้_____เป็น
40	ซึ่ง	ผู้_____มี
41	ดัง	ในเรื่อง_____ต่อไปนี้
42	ดังกล่าว	วิธี_____นี้
43	ด้าน	เก็บไว้_____ใน
44	ด้าน	ทาง_____ดนเปรียบเทียบ
45	ด้าน	ข้อกำหนดทาง_____กฎหมาย
46	เด็ก	_____นักเรียน
47	ได้	ผู้วิจัย_____สร้าง
48	ได้	ผู้วิจัย_____ศึกษา

คำที่	คำ	ปริบท
49	ได้	ผู้วิจัย_____พัฒนา
50	ได้	ผู้วิจัย_____วิเคราะห์
51	ได้	ผู้วิจัย_____พบ
52	ได้	ผู้วิจัย_____ทบทวน
53	ได้	ผู้วิจัย_____สังเคราะห์
54	ได้	ผู้วิจัย_____สกัด
55	ได้	ผู้วิจัย_____แยก
56	ได้	ผู้วิจัย_____จำแนก
57	ได้	ผู้วิจัย_____แบ่ง
58	ได้	ผู้วิจัย_____สร้าง
59	ได้	จะเห็น_____ว่า
60	ได้	เนื่องจาก_____มี
61	ได้	ผู้วิจัย_____สรุปผล
62	ได้	ถือ_____ว่า
63	ได้	_____ให้ความหมาย
64	ได้	เข้าใจ_____ง่าย
65	ได้	ผู้วิจัย_____วางแผนรายละเอียด
66	ต่อไป	ดัง_____นี้
67	ตั้ง	การจัด_____สภาชุมชน
68	ถึง	สะท้อนให้เห็น_____กระบวนการดูแล
69	ถือ	อีกด้านหนึ่งยัง_____เป็น
70	ทาง	ทักษะ_____ดนตรี
71	ทาง	ชื่อ_____การค้า
72	ทำ	ก่อน_____การทดลอง
73	ทำการ	ได้_____ศึกษา
74	ทำการ	ได้_____ทดลอง
75	ทำการ	ได้_____วิเคราะห์
76	ทำการ	ได้_____วิจัย
77	ทำให้	จากการศึกษา_____พบว่า
78	ทำให้	จากการทดลอง_____พบว่า
79	ทำให้	จากการวิเคราะห์_____พบว่า
80	ทำให้	จากการวิจัย_____พบว่า
81	ที่	เพื่อ_____จะ
82	ที่	สูง_____สุด

คำที่	คำ	ปริบท
83	ที่	ต่ำ_____สุด
84	ที่	ผู้_____เกี่ยวข้อง
85	ที่	จำเลยเป็นผู้_____ไร้ศีลธรรม
86	ที่	หน้าที่_____ต้องทำ
87	ที่	อันดับ_____หนึ่ง
88	ที่	กลุ่ม_____ออกกำลังกายแบบสลับ
89	ที่จะ	สามารถ_____
90	ที่จะ	ต้องการ_____
91	ที่จะ	อยาก_____
92	ที่จะ	ผลที่ร้ายแรง_____ตามมา
93	น้อย	ลด_____ลง
94	นี้	สารที่เกิดการดูดซับที่พื้นผิว_____
95	เนื่องจาก	_____ทำให้ผู้สอนทราบว่าสอนเพื่ออะไร
96	ใน	ตั้งวง_____ระดับ
97	ใน	ทักษะ_____การ
98	ใน	ป่าเต็งรัง_____บริเวณ
99	ใน	หรือเขียนได้ดัง_____สมการที่
100	ใน	ส่วนประสมทางการตลาด_____ด้านราคา
101	ใน	ภาวะ_____การดูดซับคาร์บอนไดออกไซด์
102	ในการ	มาใช้_____แก้ไขความผิดพลาด
103	เป็น	ซึ่ง_____ระบบที่ฝ่ายบริหารหรือคณะรัฐมนตรี
104	เป็น	เป็นหลักเกณฑ์และ_____กติกา
105	เป็น	ผู้ป่วย_____โรค
106	ไป	ประกอบ_____ด้วย
107	ไป	การเติมลง_____ในสารเคลือบผิว
108	ไป	ก่อนนำ_____เข้าเครื่องวิเคราะห์
109	เพราะ	_____ฉะนั้น
110	เพื่อ	_____แสดงสัญฐานของชั้นแพร่แก๊ส
111	เพื่อ	วิเคราะห์_____หา
112	มา	ตาม_____ด้วย
113	มา	แมสซีเตอร์มีขนาดใหญ่ขึ้น_____บ้าง
114	มาก	เพิ่ม_____ขึ้น
115	มี	ดัง_____รายละเอียด
116	มี	โดย_____คาร์บอนโมโนลิท

คำที่	คำ	ปริบท
117	มีความ	ผิวจราจร_____กว้าง
118	มีความ	ยืนเป้าหมายที่_____สนใจ
119	มีความ	กระบวนการที่_____ซับซ้อนและมีหลายมิติ
120	รวบรวม	เก็บ_____ข้อมูล
121	รอบ	ล้อม_____
122	ระยะ	เป็น_____เวลา
123	เรื่อง	แนวคิด_____ตัวชี้วัดผลลัพธ์การพยาบาล
124	โรค	_____ภูมิแพ้
125	ลด	ลัดวงจร_____น้อยลง
126	แล้ว	นอกจากนี้_____
127	แล้ว	กล่าวมา_____
128	แล้ว	จากนั้น_____
129	แล้ว	โดยมาก_____พ่อแม่
130	และ	ทางตรง_____ทางอ้อม

คำที่	คำ	ปริบท
131	และ	ทบทวนเอกสาร _____งานวิจัย
132	ว่ามี	ไม่พบ_____ความผิดปกติ
133	ไว้	ระบุ_____ว่า
134	ศึกษา	การ_____วิจัย
135	โศก	ความเศร้า_____
136	หมายเลข	สารฟู_____
137	หาก	_____แต่
138	ให้	_____ใช้ผ้า
139	ให้	มอบหมายงาน_____ทีมการพยาบาล
140	ใหม่	นวัตกรรม_____
141	ให้เห็น	ดังแสดง_____ในสมการ
142	ออก	แบ่ง_____เป็น
143	อัน	_____ได้แก่
144	อาทิ	_____เช่น

รายการคำและปริบทสำหรับใช้ลบ

คำที่	คำ	ปริบท
1	กระทั่ง	การพูดกระทบ_____
2	กับ	เพิ่มให้_____
3	การ	โดย_____ทำปฏิกิริยาระหว่างไกลโคไลซ์
4	การทดลอง	ในการทดลองจะแบ่ง_____
5	เกิน	_____ไป
6	แก่	ให้สินบน_____
7	แก่	ให้_____
8	ขอ	จะ_____
9	ขอ	ในบทนี้จะ_____กล่าวถึง
10	ข้อ	_____คำถาม
11	ของ	ในส่วน_____
12	ของ	ในด้าน_____
13	ของ	ในแง่_____
14	ของ	ในมุม_____
15	ข้างต้น	ดังกล่าว_____
16	ขึ้น	เกิด_____จาก
17	เข้า	_____ร่วมกิจกรรม

คำที่	คำ	ปริบท
18	เข้า	_____ร่วม
19	คง	ยัง_____
20	ความมี	จาก_____สาระประโยชน์
21	แค่นั้นแหละ	เขาก็จะเป็นอันตราย_____
22	งาน	ผู้ใช้_____
23	จะ	โดย_____
24	จะ	_____เห็นได้ว่า
25	จะ	_____เป็นการสัมภาษณ์กึ่งทางการ
26	จัด	_____ทำ
27	จาก	ผล_____การทำ
28	จำนวน	ในการเพิ่ม_____รถของบริษัท
29	ชิ้น	งานวิจัย_____นี้
30	ซึ่ง	_____ใน
31	ซึ่ง	_____เมื่อ
32	ซึ่ง	_____การ
33	ซึ่ง	_____จาก
34	ซึ่ง	_____พบว่า

คำที่	คำ	ปริบท
35	ซึ่ง	ผู้_____เป็น
36	ซึ่ง	ผู้_____มี
37	ซึ่ง	____น้ำที่ไหลผ่านจะกลายเป็นน้ำชะ
38	ซึ่ง	_____หากปรับตั้ง
39	ดัง	ในเรื่อง_____ต่อไปนี้
40	ด้าน	ทาง_____การ
41	เด็ก	_____นักเรียน
42	โดย	_____แบ่ง
43	โดย	_____มีรายละเอียด
44	โดย	_____จาก
45	โดย	_____ใน
46	ได้	สรุป_____ว่า
47	ได้	ผู้วิจัย_____
48	ได้	จะเห็น_____ว่า
49	ได้	เนื่องจาก_____มี
50	ได้	ผู้วิจัย_____
51	ได้	_____ให้
52	ได้	_____สรุป
53	ได้	_____ศึกษา
54	ได้	_____วิเคราะห์
55	ได้	_____ทบทวน
56	ได้	_____พบ
57	ได้	ที่_____
58	ได้	_____ให้ความหมาย
59	ได้	พบ_____
60	ต่อ	ละเมิด_____
61	ต่อ	ตอบสนอง_____
62	ต่อไป	ดัง_____นี้
63	ตามลำดับ	ที่ใดๆ ก็ตาม
64	ถึง	แสดงให้เห็น_____
65	ถึง	การตรวจสอบ_____การเตรียมความพร้อม
66	ทั้งนี้	_____จะ
67	ทั้งนี้	_____จะ
68	ทาง	_____ด้าน
69	ทำการ	ที่ใดๆ ก็ตาม

คำที่	คำ	ปริบท
70	ทำหน้าที่	ผู้รับผิดชอบต้อง_____จัดเตรียมเส้นทาง
71	ที่	โดย_____
72	ที่	เพื่อ_____จะ
73	ที่	หลังจาก_____
74	ที่	เป็นไปตาม_____
75	ที่	เน้น_____
76	ที่	เส้นทาง_____ให้
77	ที่	ผู้_____มีรายได้
78	ที่	เป็นเทคโนโลยี_____ใหม่
79	ที่	อันดับ_____หนึ่ง
80	ที่	อยู่ในระดับ_____เหมาะสม
81	ที่	สารอาหาร_____สาหร่ายจะใช้ในการเจริญ
82	ที่	การ_____ทำสุขภาพร่างกายให้แข็งแรง
83	ที่จะ	ก่อน_____
84	ที่จะ	สามารถ_____
85	ที่จะ	ต้องการ_____
86	ที่จะ	อยาก_____
87	ที่จะ	พยายาม_____
88	ที่จะ	ย่อม_____
89	ที่เป็น	ในฐานะ_____
90	ที่มี	สูงกว่ากลุ่มตัวอย่าง_____ที่มีอายุ
91	ที่อยู่	ความยาวคลื่น_____ในวงเล็บ
92	น้อย	ลด_____ลง
93	นั้น	นี้_____
94	นำ	_____เสนอ
95	เน้น	มุ่ง_____
96	เนื่อง	นับ_____
97	ใน	_____เชิง
98	ใน	_____ทาง
99	ใน	ทักษะ_____การ
100	ในการ	เวลา_____ชะละลายฟอสฟอรัส
101	เบื้อง	ตัวแปร_____ต้น
102	เป็น	_____โสด
103	เป็นการ	ปรากฏว่า_____จบที่
104	เป็นต้น	ที่ใดๆ ก็ตาม

คำที่	คำ	ปริบท
105	เป็นที่	คือ_____เหมาะ
106	เป็นอัน	_____มาก
107	ไป	ประกอบ_____ด้วย
108	ผล	จาก_____การทดสอบสมมติฐานข้อที่
109	ฝัง	แฝง_____
110	เพราะ	_____ฉะนั้น
111	เพื่อ	ใช้_____หาความสัมพันธ์ระหว่าง
112	ภาค	_____ส่วน
113	มัน	พม่า_____จะ
114	มาก	เพิ่ม_____ขึ้น
115	มี	ดัง_____รายละเอียดต่อไปนี
116	มีการ	ที่ใดๆ ก็ตาม
117	มีความ	ที่ใดๆ ก็ตาม
118	มีค่า	กรดด่างของน้ำทะเล_____ค่อนข้างคงที่
119	รวบรวม	เก็บ_____ข้อมูล
120	รอบ	ล้อม_____
121	ระยะ	เป็น_____เวลา
122	รายการ	ดัง_____ต่อไปนี้
123	เรื่อง	ศึกษา_____ผล
124	เรือน	บ้าน_____
125	ลง	สำเร็จ_____ได้
126	ลักษณะ	_____นิสัย
127	แล้ว	กล่าวมา_____
128	แล้ว	จากนั้น_____
129	และ	_____นอกจากนี้
130	และ	กลยุทธ์_____ความเป็นผู้นำ
131	และ	_____หลังจาก
132	ไว้	ระบุ_____ว่า
133	ศึกษา	การ_____วิจัย
134	โศก	ความเศร้า_____
135	สาย	รับ_____โทรศัพท์
136	เสีย	_____ก่อน

คำที่	คำ	ปริบท
137	หลัง	บ้าน_____นี้
138	หาก	_____แต่
139	เหนือ	นอกเหนือ_____จาก
140	แห่งนี้	สนามบินเลิงนกทา_____จะเป็น
141	ให้การ	ในการ_____บำบัดทางการพยาบาล
142	ให้ความ	_____เมตตา
143	ใหม่	นวัตกรรม_____
144	ให้มี	จัด_____
145	ให้เห็น	แสดง_____ว่า
146	อย่าง	_____ยิ่ง
147	อยู่	แยกส่วนกัน_____ระหว่าง
148	ออก	ถัด_____มา
149	ออก	ฟองน้ำยื่น_____ไปสัมผัสขวด
150	ออก	แบ่ง_____เป็น
151	อาทิ	_____เช่น
152	อีก	_____ด้วย
153	เอง	นี้_____
154	เอา	นำ_____
155	ๆ	ทั่ว_____ไป

ภาคผนวก ค

คำชี้แจงสำหรับผู้จำลองการลักลอกโดยดัดแปลง

เนื่องด้วยผู้วิจัยกำลังทำวิทยานิพนธ์หัวข้อ “การตรวจเทียบภายนอกหาการลักลอกในงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความ (Extrinsic Plagiarism Detection in Academic Texts Using a Support Vector Machine Model and Text Similarity Measurement)” ในขั้นตอนการวิจัยหัวข้อดังกล่าว ผู้วิจัยจำเป็นต้องทดลองสร้างคลังข้อมูลจำลองการลักลอกขึ้น เพื่อนำมาทดลองวัดประสิทธิภาพของวิธีการตรวจจับการลักลอกที่นำเสนอในงานวิทยานิพนธ์

ในการนี้ ผู้วิจัยจึงขอรบกวนให้ท่านจำลองการลักลอกขึ้น โดย**ดัดแปลง**ข้อความที่ผู้วิจัยได้เตรียมไว้ให้เป็นย่อหน้า โดยมีรายละเอียดดังต่อไปนี้

- **ข้อมูลต้นฉบับ**

ข้อมูลต้นฉบับสำหรับจำลองการลักลอกที่ท่านได้รับจะมีลักษณะเป็นย่อหน้าของข้อความเชิงวิชาการ จำนวนทั้งสิ้น 1,250 ย่อหน้า จำแนกตามสาขาวิชาและขนาดของย่อหน้า ตามรายละเอียดดังนี้

สาขาวิชา	วิทยาศาสตร์			มนุษยศาสตร์และสังคมศาสตร์		
ขนาด	สั้น	กลาง	ยาว	สั้น	กลาง	ยาว
จำนวนคำ/ย่อหน้า	50-100	101-150	151-200	50-100	101-150	151-200
จำนวนย่อหน้าที่ได้รับ	208	209	208	208	209	208

- **ไฟล์ข้อมูลต้นฉบับ**

ย่อหน้าที่ท่านจะได้รับจะถูกบันทึกในรูปแบบไฟล์สกุล .txt ไฟล์ละ 1 ย่อหน้า โดยชื่อไฟล์ถูกตั้งเป็นอักขระ 8 หลัก ตามด้วยข้อความ “-marked” ทั้งนี้ อักขระแต่ละตัวบ่งชี้ข้อมูลต่อไปนี้

1) อักขระหลักที่ 1-2 ใช้จำแนกสาขาวิชาของย่อหน้าต้นฉบับ ดังนี้

อักขระ SC หมายถึง ย่อหน้าที่มีเนื้อหาทางวิทยาศาสตร์

อักขระ HS หมายถึง ย่อหน้าที่มีเนื้อหาทางมนุษยศาสตร์และสังคมศาสตร์

2) อักขระหลักที่ 3 ใช้จำแนกขนาดของย่อหน้าต้นฉบับ ดังนี้

อักขระ S หมายถึง ย่อหน้าขนาดสั้น มีความยาว 50-100 คำ

อักขระ M หมายถึง ย่อหน้าขนาดกลาง มีความยาว 101-150 คำ

อักขระ L หมายถึง ย่อหน้าขนาดยาว มีความยาว 151-200 คำ

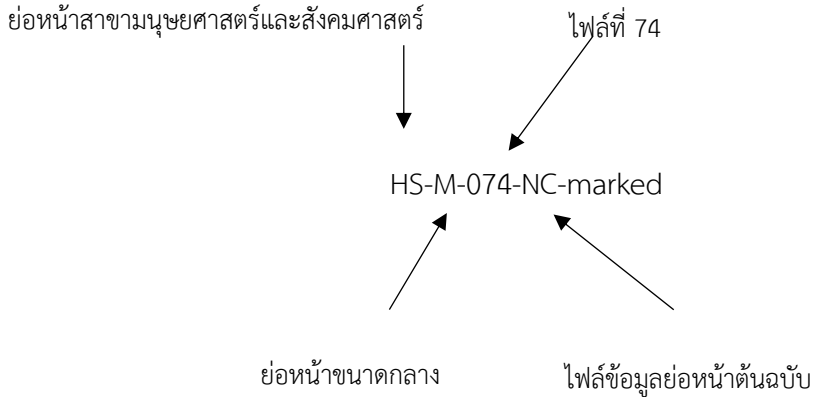
3) อักขระหลักที่ 4-6 ใช้จำแนกลำดับที่ของย่อหน้าต้นฉบับในสาขาวิชาและขนาดเดียวกัน

4) อักขระหลังที่ 7-8 ใช้จำแนกความเป็นข้อมูลต้นฉบับหรือข้อมูลที่ผ่านการลักลอก ดังนี้

อักขระ NC หมายถึง ย่อหน้าต้นฉบับ

อักขระ MO หมายถึง ย่อหน้าที่ผ่านการดัดแปลงแล้ว

<u>ตัวอย่าง</u>



- **ภาระงานโดยสังเขป**

ให้ท่าน**<u>ดัดแปลง</u>**ย่อหน้าต้นฉบับแต่ละไฟล์ตามขั้นตอนดังนี้

1) เปิดไฟล์ย่อหน้าต้นฉบับด้วยโปรแกรม Notepad, Notepad++, หรือ Microsoft Word ตามแต่ท่านสะดวก
2) ดัดแปลงข้อความที่ปรากฏในย่อหน้า
3) บันทึกไฟล์เป็นไฟล์ใหม่สกุล .txt โดยตั้งชื่อไฟล์อักขระที่ 1-6 ตามไฟล์ย่อหน้าต้นฉบับที่ลักลอก ส่วนอักขระที่ 7-8 ให้เปลี่ยนจากอักขระ NC เป็น MO
4) จัดเก็บไฟล์ที่ดัดแปลงแล้วแยกตามสาขาวิชาและขนาด

- **ลักษณะของข้อความที่ดัดแปลงได้**

การดัดแปลงข้อความในที่นี้จะทำโดยการ**<u>แทรก ลบ หรือย้ายที่</u>** ข้อความต้นฉบับในระดับ**<u>วลีหรืออนุพากย์</u>** ดังนั้นในหัวข้อนี้ ผู้วิจัยจะได้ชี้แจงถึงลักษณะของวลีและอนุพากย์ที่สามารถดัดแปลงได้โดยอิงตามนิยามที่ระบุไว้โดยนลินี อินต๊ะซาว (2556, น. 30-43) และให้ดัดแปลงวลีและอนุพากย์เหล่านั้นด้วยการแทรก ลบ และย้ายที่ จากนั้นจึงจะแสดงให้เห็นถึงกลวิธีดัดแปลงข้อความในลำดับถัดไป

1) ลักษณะของวลีที่ดัดแปลงได้

วลีที่สามารถดัดแปลงด้วยการแทรก ลบ หรือย้ายที่ นั้นจำต้องเป็นวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น (strong maker) ทั้งนี้ จากการศึกษาหน่วยเชื่อมโยงในภาษาไทย พบว่ามีคำเชื่อมเด่นอยู่เพียง 2 ประเภท คือ คำเชื่อมแจกแจงสมาชิกหรือแสดงตัวอย่าง เช่น “ได้แก่”, “เช่น”, “ตัวอย่างเช่น”, “อาทิ”, “เป็นต้นว่า” และคำเชื่อมแสดงวัตถุประสงค์ เช่นคำว่า “เพื่อ”

<u>ตัวอย่าง (1)</u>

> [ตำนานปรัมปราเป็นการอธิบายถึงกาเนิดของจักรวาล โครงสร้างและระบบของจักรวาล มนุษย์ สัตว์ ปรากฏการณ์ทางธรรมชาติ]$_1$ *[เช่น ลม ฝน กลางวัน กลางคืน ฟ้าร้อง ฟ้าผ่า]*$_2$

จากตัวอย่าง (1) จะเห็นได้ว่าข้อความหมายเลข 2 เป็นวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น “เช่น” อันเป็นคำเชื่อมแสดงตัวอย่าง วลีดังกล่าวจึงสามารถถูกดัดแปลงได้ตามหลักการในงานวิจัยชิ้นนี้

<u>ตัวอย่าง (2)</u>

> [เจ้าฟ้ากรมพระยานริศรานุวัดติวงศ์ทรงนิพนธ์บทละครดึกดำบรรพ์ไว้หลายเรื่อง]$_1$ *[ได้แก่ สังข์ทอง คาวี อิเหนา สังข์ศิลป์ชัย อุณรุท รามเกียรติ์ กรุงพาณชมทวีป]*$_2$ [นอกจากนี้ยังมีบทละคร]$_{3.1}$ [ที่ทรงแต่ง]$_4$ [อีกหลายเรื่องด้วย]$_{3.2}$

ในตัวอย่าง (2) ข้างต้นจะเห็นได้ว่าข้อความหมายเลข 2 เป็นวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น “ได้แก่” อันเป็นคำเชื่อมแจกแจงสมาชิก จึงสามารถถูกดัดแปลงวลีดังกล่าวได้ตามหลักการในงานวิจัยชิ้นนี้

ตัวอย่าง (3)

[จอมพลสฤษฎ์ ธนรัชต์ ได้ริเริ่มแนวทางการพัฒนาประเทศไปสู่ความทันสมัย]$_1$ [โดยเพิ่มประสิทธิภาพการผลิต]$_2$ [พัฒนาโครงสร้างพื้นฐาน]$_3$ *[เพื่อพัฒนาอุตสาหกรรม]*$_4$ [และพยายามทำลายขบวนการเคลื่อนไหวของพรรคคอมมิวนิสต์แห่งประเทศไทย]$_5$

จากตัวอย่าง (3) จะเห็นได้ว่าข้อความหมายเลข 4 เป็นวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น “เพื่อ” อันเป็นคำเชื่อมแสดงวัตถุประสงค์ วลีดังกล่าวจึงสามารถถูกดัดแปลงได้ตามหลักการในงานวิจัยชิ้นนี้

2) ลักษณะของอนุพากย์ที่ดัดแปลงได้

อนุพากย์ที่สามารถดัดแปลงด้วยการแทรก ลบ หรือย้ายที่ ตามขอบเขตของงานชิ้นนี้มีอยู่ 2 ประเภทใหญ่ ได้แก่ อนุพากย์ที่มีกริยาแท้ (finite clause) อนุพากย์เติมเต็มของกริยาที่มีกริยาแท้ (finite clausal complement of verb) ดังมีรายละเอียดต่อไปนี้

2.1) อนุพากย์ที่มีกริยาแท้ (finite clause)

กริยาแท้ หมายถึง กริยาที่สามารถทำหน้าที่เป็นภาคแสดง (predicate) ของอนุพากย์ สามารถแสดงข้อมูลทางไวยากรณ์ (grammatical information) ต่าง ๆ ได้ เช่น กาล การณ์ลักษณะวาจก ทัศนภาวะ ฯลฯ ในภาษาไทยซึ่งเป็นภาษารูปคำโดด (isolating language) กริยาแท้สามารถให้ข้อมูลทางไวยากรณ์ต่าง ๆ ที่กล่าวมาได้โดยการเติมกริยาช่วยประเภทต่าง ๆ ไว้ข้างหน้ากริยาแท้ เช่น “จะ” แสดงการณ์ลักษณะไม่สมบูรณ์, “เคย” แสดงสถานการณ์สมบูรณ์, “กำลัง” แสดงการณ์ลักษณะที่กำลังดำเนินอยู่, “ถูก” แสดงกรรมวาจก, “ควร” แสดงทัศนภาวะ ฯลฯ

อนุพากย์ที่มีกริยาแท้ (finite clause) สามารถแบ่งออกเป็นอนุพากย์อิสระ (independent clause) และอนุพากย์ไม่อิสระ (dependent clause) ในขณะที่อนุพากย์อิสระสามารถอยู่ได้ด้วยตัวเองอย่างมีใจความสมบูรณ์ อนุพากย์ไม่อิสระต้องพึ่งพาอนุพากย์อิสระเสมอ เพราะไม่สามารถอยู่ได้ด้วยตัวเอง อนุพากย์ทั้งสองประเภทสามารถเชื่อมเข้าด้วยกันด้วยหน่วยเชื่อมโยงปริจเฉทหรือคำเชื่อม ทั้งนี้ยังสามารถจัดแบ่งอนุพากย์ไม่อิสระที่สามารถดัดแปลงได้อีก 3 ประเภทย่อย ดังนี้

2.1.1) คุณานุประโยคที่มีกริยาแท้ (finite relative clause)

คุณานุประโยคที่มีกริยาแท้เป็นอนุพากย์ไม่อิสระประเภทหนึ่ง ในทางหน้าที่คุณานุประโยคเป็นหน่วยสร้างที่มีหน้าที่ขยายคำนามหลัก (head noun) และในทางความหมาย ถือว่าเป็นอนุพากย์ที่ทำให้คำนามหลักที่ถูกอ้างอิงถึงมีความเฉพาะจงเจาะยิ่งขึ้น หน่วยสร้างประเภทนี้จะปรากฏหลังคำนามหลักเท่านั้น และอาจจะมีหรือไม่มีตัวบ่งชี้ “ที่”, “ซึ่ง”, “อัน” นำหน้าก็ได้

ตัวอย่าง (4)

[เจ้าฟ้ากรมพระยานริศรานุวัดติวงศ์ทรงนิพนธ์บทละครดึกดำบรรพ์ไว้หลายเรื่อง]$_1$ [ได้แก่ สังข์ทอง คาวี อิเหนา สังข์ศิลป์ชัย อุณรุท รามเกียรติ์ กรุงพาณชมทวีป]$_2$ [นอกจากนี้ยังมีบทละคร]$_{3.1}$ *[ที่ทรงแต่ง]*$_4$ [อีกหลายเรื่องด้วย]$_{3.2}$

จากตัวอย่าง (4) ข้างต้นจะเห็นได้ว่าข้อความหมายเลข 4 เป็นคุณานุประโยคที่มีกริยาแท้ทำหน้าที่ขยายคำนามหลัก “บทละคร” ในข้อความหมายเลข 3.1 ให้มีความหมายจำเพาะเจาะจงมากยิ่งขึ้น ในงานวิจัยชิ้นนี้จึงถือว่าสามารถดัดแปลงอนุพากย์ดังกล่าวได้

2.1.2) วิเศษณานุประโยค (adverbial clause)

วิเศษณานุประโยคเป็นอนุพากย์ไม่อิสระประเภทหนึ่งที่ช่วยขยายกริยาหลักในอนุพากย์อิสระเกี่ยวกับเงื่อนไข เหตุผล เวลา เป็นต้น สามารถปรากฏตำแหน่งหน้าหรือหลังอนุพากย์อิสระได้

วิเศษณานุประโยคสามารถแบ่งออกเป็นประเภทต่าง ๆ ตามความหมายและหน้าที่ซึ่งแต่ละประเภทจะใช้คำเชื่อมแสดงปริจเฉทสัมพันธ์ที่แตกต่างกันออกไป เช่น วิเศษณานุประโยคบอกเวลา (time adverbial clause) ขึ้นต้นด้วยคาเชื่อม “จนกระทั่ง”, “ในขณะที่”, “ขณะที่”, “ขณะ”, “ในระหว่างที่”, “ในระหว่าง” เป็นต้น วิเศษณานุประโยคบอกเหตุ (causal adverbial clause) ขึ้นต้นด้วยคำเชื่อม”เพราะ”, “เพราะว่า”, “เนื่องจาก”, “เนื่องจากว่า” เป็นต้น วิเศษณานุประโยคแสดงเงื่อนไข (conditional adverbial clause) ขึ้นต้นด้วยคำเชื่อม “ถ้า”, “หาก”, “ถ้าหาก”, “หากว่า”, “ถ้าหากว่า” เป็นต้น วิเศษณานุประโยคเงื่อนไขเพื่อยืนยัน (concessive adverbial clause) ขึ้นต้นด้วยคำเชื่อม “ถึงแม้ว่า”, “แม้ว่า” เป็นต้น คำเชื่อมเหล่านี้สามามารถใช้เป็นตัวบ่งชี้จุดเริ่มต้นของวิเศษณานุประโยคได้

ตัวอย่าง (5)

[เขาต้องถูกส่งตัวกลับมาก่อน]$_1$ *[เพราะป่วยเป็นวัณโรค]*$_2$ [ด้วยเหตุนี้เขาจึงไม่ได้อยู่ประจำการที่ต่างประเทศตามเวลา]$_3$ [ที่กำหนดไว้เดิม]$_4$

ตัวอย่าง (6)

[หากพิเคราะห์ให้ถี่ถ้วนแล้ว]$_1$ [บทความชิ้นนี้ดูจะมีกลิ่นไอของวิธีคิดเรื่องตัวตนและอัตลักษณ์แบบสำนักคิดหลังสมัยใหม่]$_2$

จากตัวอย่างข้างต้น ข้อความหมายเลข 2 ในตัวอย่าง (5) เป็นวิเศษณานุประโยคแสดงเหตุผลขึ้นต้นอนุพากย์ด้วยคำเชื่อม “เพราะ” และข้อความหมายเลข 1 ในตัวอย่าง (6) เป็นวิเศษณานุ

ประโยคแสดงเงื่อนไข ขึ้นต้นด้วยคำเชื่อม “หาก” จึงถือว่าวิเศษณานุประโยคทั้งสองสามารถถูกดัดแปลงได้ตามหลักของงานวิจัยชิ้นนี้

2.1.3) อนุพากย์ความรวม (coordinate clause)

อนุพากย์ความรวม คือ อนุพากย์มากกว่าหนึ่งอนุพากย์ที่มีความเท่าเทียมกันในเชิงหน้าที่เชื่อมโยงเข้าด้วยกันในโครงสร้างระดับเดียวกันด้วยสันธานประสานหรือคำเชื่อม เช่น “และ” “หรือ” “แต่” ฯลฯ ผู้วิจัยกำหนดให้อนุพากย์ที่เชื่อมด้วยคำเชื่อมเหล่านี้เป็นอนุพากย์ที่สามารถถูกดัดแปลงได้

อย่างไรก็ตาม อนุพากย์ความรวมอาจมีรูปแบบของโครงสร้างที่คล้ายกับกริยาวลีที่เชื่อมเข้าด้วยกันด้วยสันธานประสาน เรียกว่า กริยาวลีความรวม (coordinate verb phrase) ซึ่งผู้วิจัยไม่ถือว่าเป็นอนุพากย์ที่สามารถดัดแปลงได้ เนื่องจากกริยาในกริยาวลีไม่ได้มีความสัมพันธ์ทางปริจเฉทต่อกัน

วิธีสังเกตว่าข้อความนั้นๆ เป็น อนุพากย์ความรวม (coordinate clause) หรือ กริยาวลีความรวม (coordinate verb phrase) สามารถพิจารณาได้จากกริยาหลัก กล่าวคือ ในอนุพากย์ความรวม กริยาของแต่ละอนุพากย์จะไม่ใช้กรรม ส่วนขยาย ส่วนเติมเต็ม หรือส่วนเสริม ร่วมกัน ส่วนกริยาในกริยาวลีความรวมจะใช้กรรม ส่วนขยาย ส่วนเติมเต็ม หรือส่วนเสริม ร่วมกันได้

ตัวอย่าง (7)

[เกิดในโตเกียว]$_1$ *[และมาเติบโตที่โอซากา]*$_2$

ตัวอย่าง (8)

[แต่หลายส่วนลอกและเพิ่มเติมมาจากกฎหมายตราสามดวง]$_1$

ตัวอย่าง (7) เป็นอนุพากย์ความรวมที่เชื่อมกันด้วยคำเชื่อม “และ” กริยาของแต่ละอนุพากย์ไม่ได้ใช้ส่วนเสริมร่วมกัน นั่นคือ “ในโตเกียว” เป็นส่วนเสริมของกริยา “เกิด” และ “ที่โอซากา” เป็นส่วนเสริมของกริยา “มาเติมโต” ในที่นี้ ผู้จำลองการลักลอกจึงอาจเลือกดัดแปลงข้อความหมายเลข 1 หรือ 2 ก็ได้ ส่วนตัวอย่าง (8) เป็นกริยาวลีความรวมที่กริยา “ลอก” และ “เพิ่มเติม” เชื่อมเข้าด้วยกันด้วยคำเชื่อม “และ” ใช้ส่วนเติมเต็ม “จากกฎหมายตราสามดวง” ร่วมกัน ดังนั้นจึงไม่ถืออนุพากย์ความรวมที่สามารถดัดแปลงได้ในที่นี้

2.2) อนุพากย์เติมเต็มของกริยาที่มีกริยาแท้ (finite clausal complement of verb)

อนุพากย์เติมเต็มเป็นอนุพากย์ที่ทำหน้าที่เป็นส่วนเติมเต็มให้กับกริยาหรือคำนามที่ต้องการส่วนเติมเต็ม อนุพากย์เติมเต็มอาจอยู่ในรูปที่มีกริยาแท้หรือไม่มีกริยาแท้ก็ได้ หากมีกริยาแท้ อนุพากย์นั้นก็จะถูกดัดแปลงได้ตามแนวคิดของงานวิจัยชิ้นนี้

อนุพากย์เติมเต็มของกริยาที่มีกริยาแท้มักเป็นอนุพากย์ที่เป็นส่วนเติมเต็มของกริยาแสดงการรับรู้ (cognitive verb) เช่น “คิด”, “เชื่อ”, “รู้”, “จินตนาการ”, “สมมติ”, “หวัง”, “คาด”, “ฝัน” ฯลฯ และกริยาที่ใช้ในการรายงานคำพูด (verb in reported speech) เช่น “พูด”, “ประกาศ”, “ชี้แจง”, “แนะนำ”, “รายงาน”, “อธิบาย”, “ถาม”, “บอก”, “กล่าว” ฯลฯ นักภาษาศาสตร์บางคนเรียกกริยาทั้ง 2 ประเภทนี้ว่า “กริยาลักษณะประจำ (attributive verb)” ทั้งนี้ อนุพากย์ชนิดนี้มักจะขึ้นต้นด้วยตัวบ่งชี้ส่วนเติมเต็ม “ว่า”

ตัวอย่าง (11)

[สรุปได้]$_1$ *[**ว่า**สามารถจำแนกออกได้เป็น 2 แบบด้วยกัน]*$_2$

ตัวอย่าง (11) ข้างต้นนี้ แสดงอนุพากย์เติมเติมของกริยา และอนุพากย์ดังกล่าวมีขึ้นต้นด้วยตัวบ่งชี้ส่วนเติมเต็ม “ว่า” ในงานวิจัยชิ้นนี้จึงถือว่าข้อความหมายเลข 2 เป็นอนุพากย์ที่สามารถดัดแปลงได้

อนึ่ง มีหน่วยสร้าง 2 ประเภทที่ผู้จำลองการลักลอกพึงสังเกต ได้แก่ อนุพากย์เติมเต็มของกริยาที่ไม่มีกริยาแท้ (non-finite clausal complement of verb) และอนุพากย์เติมเต็มของนาม (clausal complement of noun) เนื่องจากหน่วยสร้างทั้ง 2 ประเภท**ไม่สามารถตัดแบ่งเป็นวลีหรืออนุพากย์เพื่อดัดแปลงได้**

อนุพากย์เติมเต็มของกริยาที่ไม่มีกริยาแท้ (non-finite clausal complement of verb) คือส่วนเติมเต็มของกริยาที่มีโครงสร้างเป็นอนุพากย์ที่มักปรากฏหลังตัวบ่งชี้ “ที่จะ” และ “จะ” กริยาของอนุพากย์ที่ตามหลังตัวบ่งชี้เหล่านี้จะไม่สามารถแสดงข้อมูลทางไวยากรณ์ได้ เพราะมีสถานะเป็นกริยาไม่แท้ คำกริยาหลักที่ต้องการส่วนเต็มเต็มประเภทนี้ ได้แก่ กริยาแสดงความปรารถนา เช่น “อยาก”, ““ชอบ”, “ต้องการ” ฯลฯ กริยาที่สื่อความหมายโดยนัย (implicative verb) เช่น “พยายาม”, “ลอง” ฯลฯ กริยาช่วย (modal verb) เช่น “สามารถ”, “ควร” ฯลฯ

ตัวอย่าง (9)

[แต่ก็เลือก**ที่จะ**ละเลยเนื้อหามาตรฐานของพุทธศาสนา]$_1$

ตัวอย่าง (9) ข้างต้น เป็นตัวอย่างของหน่วยสร้างที่ประกอบไปด้วยอนุพากย์เติมเต็มของกริยาที่ไม่มีกริยาแท้ จะเห็นว่า “ที่จะละเลยเนื้อหา...” เป็นส่วนเติมเต็มของกริยา “เลือก” ดังนั้นหากจะดัดแปลงข้อความดังกล่าวจึงจำเป็นต้องดัดแปลงทั้งข้อความ เพราะถือเป็นอนุพากย์ที่มีกริยาแท้ทั้งข้อความ ไม่สามารถเลือกดัดแปลงเฉพาะส่วนของข้อความหลังคำว่า “ที่จะ” ได้

ส่วนหน่วยสร้างที่ผู้จำลองการลักลอกพึงระวังอีกประเภท ได้แก่ อนุพากย์เติมเต็มของนาม (clausal complement of noun) อนุพากย์ชนิดนี้มักจะขึ้นต้นด้วยตัวบ่งชี้ “ที่ว่า” “ที่จะ” ทำหน้าที่เป็นส่วนเติมเต็มให้กับคำนามหลัก หน่วยสร้างชนิดไม่สามารถตัดแบ่งเป็นส่วนของข้อความเพื่อดัดแปลงได้

ตัวอย่าง (10)

[โดยมีวัตถุประสงค์**ที่จะ**ปรับประยุกต์เข้ากับตลาดทุนนิยม]$_1$

ข้อความ “ที่จะปรับประยุกต์...” ในตัวอย่าง (10) ทำหน้าที่เป็นอนุพากย์เติมเต็มของคำนาม “ประสงค์” ดังนั้นจึงไม่สามารถเลือกดัดแปลงเฉพาะส่วนของข้อความหลังคำว่า “ที่จะ” ได้ เพราะถือเป็นอนุพากย์ที่มีกริยาแท้ หากจะดัดแปลงข้อความดังกล่าวจึงจำเป็นต้องดัดแปลงทั้งข้อความ

- **กลวิธีดัดแปลงข้อความและการกำกับข้อมูล**

ดังได้กล่าวไปแล้วในตอนต้นว่า การลักลอกโดยดัดแปลงข้อความที่กำหนดไว้ในงานวิจัยชิ้นนี้จะใช้กลวิธีหลัก 3 วิธี เพื่อดัดแปลงข้อความในระดับวลีหรืออนุพากย์ ดังนั้นในหัวข้อนี้ ผู้วิจัยจะได้แสดงถึงรายละเอียดของกลวิธีที่ใช้ดัดแปลงข้อความ ดังนี้

1) การแทรก

การแทรก คือ การเพิ่มเติมข้อความในระดับวลีหรืออนุพากย์เข้าไปในข้อความต้นฉบับที่มีอยู่เดิม ทั้งนี้ ผู้จำลองการลักลอกควรอ่านข้อความต้นฉบับทั้งย่อหน้าให้เข้าใจก่อน จากนั้นจึงเลือกแทรกวลีหรืออนุพากย์ที่มีลักษณะตามที่ได้กล่าวถึงไปในหัวข้อที่แล้วเข้าไปในข้อความต้นฉบับเพียง 1 ครั้ง โดยกำกับข้อมูลที่แทรกเพิ่มเติมเข้าไปด้วยป้ายกำกับ **<xins>...</xins>** ดังตัวอย่าง

ตัวอย่าง (11)

(11)$_{src}$ การแบ่งประเภทของการท่องเที่ยวมีพัฒนาการมาจากการเดินทางในยุคอดีต ปัจจุบันการท่องเที่ยวมีหลากหลายประเภท อาทิเช่น การเดินทางท่องเที่ยวเชิงการค้า ธุรกิจ การเดินทางเชิงศาสนา เพื่อแสวงบุญ การท่องเที่ยวพักผ่อน เพื่อความเพลิดเพลิน (Leiper, 2004: 4-10) อย่างไรก็ตาม การจัดแบ่งประเภทการท่องเที่ยวไม่ได้เป็นการแบ่งแยกอย่างเด็ดขาด เพราะในการเดินทางหนึ่งครั้งอาจจะเกี่ยวข้องหรือผสมผสานการเดินทางหลากหลายประเภท สำหรับผู้สูงอายุที่ท่องเที่ยวอย่างจริงจัง อาจจะเคยเดินทางท่องเที่ยวหลากหลายประเภทหรือสามารถที่จะเดินทางท่องเที่ยวได้

ทุกรูปแบบ ทุกประเภท หากแต่ความสนใจหลักของผู้สูงอายุหากแต่ละคนอาจจะแตกต่างกัน

$(11)_{plg}$ การแบ่งประเภทของการท่องเที่ยวมีพัฒนาการมาจากการเดินทางในยุคอดีต ปัจจุบันการท่องเที่ยวมีหลากหลายประเภท อาทิเช่น การเดินทางท่องเที่ยวเชิงการค้า ธุรกิจ การเดินทางเชิงศาสนา เพื่อแสวงบุญ **<xins>เช่น การเข้าวัด การไปฟังเทศน์ การไปเยือนศาสนสถานสำคัญประจำถิ่น เป็นต้น</xins>** การท่องเที่ยวพักผ่อน เพื่อความเพลิดเพลิน (Leiper, 2004: 4-10) อย่างไรก็ตาม การจัดแบ่งประเภทการท่องเที่ยวไม่ได้เป็นการแบ่งแยกอย่างเด็ดขาด เพราะในการเดินทางหนึ่งครั้งอาจจะเกี่ยวข้องหรือผสมผสานการเดินทางหลากหลายประเภท สำหรับผู้สูงอายุที่ท่องเที่ยวอย่างจริงจัง อาจจะเคยเดินทางท่องเที่ยวหลากหลายประเภทหรือสามารถที่จะเดินทางท่องเที่ยวได้ทุกรูปแบบ ทุกประเภท หากแต่ความสนใจหลักของผู้สูงอายุหากแต่ละคนอาจจะแตกต่างกัน

ตัวอย่าง (11) ข้างต้นแสดงการแทรกวลีที่ขึ้นต้นด้วยคำเชื่อมเด่น “เช่น การเข้าวัด การไปฟังเทศน์ การไปเยือนศาสนสถานสำคัญประจำถิ่น เป็นต้น” เข้าไปข้อความต้นฉบับ $(11)_{src}$ เพื่อทำหน้าที่แสดงตัวอย่างของ “การเดินทางเชิงศาสนา” เมื่อแทรกวลีดังกล่าวเข้าไปแล้วจึงกำกับป้ายกำกับ <xins> และ </xins> คร่อมข้อความที่แทรกเข้าไปเพื่อระบุขอบเขตของข้อความที่ถูกแทรก ดังแสดงให้เห็นในข้อความที่ถูกดัดแปลง $(11)_{plg}$

ตัวอย่าง (12)

$(12)_{src}$ สุมนทิพย์ จิตสว่าง (2554) ได้กล่าวถึงการกระทำผิดของเด็กและเยาวชนไทยไว้ด้วยว่าเด็กและเยาวชนที่กระทำผิดในประเทศไทยในปัจจุบันมีอายุระหว่าง 7-18 ปี โดยเด็กไทยยุคใหม่มีพฤติกรรมเบี่ยงเบนตลอดจนการกระทำความผิดที่สำคัญ อาทิ อัพยา การกระทำผิดเกี่ยวกับยาเสพติด โดยเฉพาะการเสพยาเสพติดของเด็กและเยาวชนไทยที่มีจำนวนมากขึ้น ท้าตี การกระทำผิดของเด็กและเยาวชนที่เกี่ยวข้องกับการยกพวกตีกัน การจัดตั้งกลุ่มอันธพาลในการยกพวกตีกัน หรือทำร้ายผู้อื่น ฟรีเซ็กส์ การมีพฤติกรรมในเรื่องการมีเพศสัมพันธ์เป็นเรื่องธรรมดา รวมทั้งอาจนำไปสู่การค้าประเวณี ตลอดจนปัญหาการทำแท้ง โดย อาชญากรเด็ก ส่วนใหญ่มีอายุ 15 – 18 ปี โดยเป็นเพศชายมากกว่าหญิงประมาณ 10 เท่า โดย

สามารถที่จะแบ่งประเภทความผิดที่เด็กและเยาวชนกระทำผิดออกเป็น 7 กลุ่มความผิดซึ่งเป็นไปตามการกำหนดของกรมพินิจและคุ้มครองเด็กและเยาวชน

$(12)_{plg}$ สุมนทิพย์ จิตสว่าง (2554) ได้กล่าวถึงการกระทำผิดของเด็กและเยาวชนไทยไว้ด้วยว่าเด็กและเยาวชนที่กระทำผิดในประเทศไทยในปัจจุบันมีอายุระหว่าง 7-18 ปี โดยเด็กไทยยุคใหม่มีพฤติกรรมเบี่ยงเบนตลอดจนการกระทำความผิดที่สำคัญ อาทิ อัพยา การกระทำผิดเกี่ยวกับยาเสพติด โดยเฉพาะการเสพยาเสพติดของเด็กและเยาวชนไทยที่มีจำนวนมากขึ้น ท้าตี การกระทำผิดของเด็กและเยาวชนที่เกี่ยวข้องกับการยกพวกตีกัน การจัดตั้งกลุ่มอันธพาลในการยกพวกตีกัน หรือทำร้ายผู้อื่น ฟรีเซ็กส์ การมีพฤติกรรมในเรื่องการมีเพศสัมพันธ์เป็นเรื่องธรรมดา**<xins>ที่สามารถพบได้โดยทั่วไปในเด็กและเยาวชนกลุ่มนี้</xins>** รวมทั้งอาจนำไปสู่การค้าประเวณี ตลอดจนปัญหาการทำแท้ง โดย อาชญากรเด็ก ส่วนใหญ่มีอายุ 15 – 18 ปี โดยเป็นเพศชายมากกว่าหญิงประมาณ 10 เท่า โดยสามารถที่จะแบ่งประเภทความผิดที่เด็กและเยาวชนกระทำผิดออกเป็น 7 กลุ่มความผิดซึ่งเป็นไปตามการกำหนดของกรมพินิจและคุ้มครองเด็กและเยาวชน

ตัวอย่าง (12) นี้แสดงการแทรกอนุพากย์ที่มีกริยาแท้ประเภทคุณานุประโยค “ที่สามารถพบได้โดยทั่วไปในเด็กและเยาวชนกลุ่มนี้” เข้าไปข้อความต้นฉบับ $(12)_{src}$ เพื่อทำหน้าที่ขยายคำนาม “เรื่องธรรมดา” เมื่อแทรกอนุพากย์ดังกล่าวเข้าไปแล้วจึงกำกับป้ายกำกับ <xins> และ </xins> คร่อมข้อความที่แทรกเข้าไปเพื่อระบุขอบเขตของข้อความที่ถูกแทรก ดังแสดงให้เห็นในข้อความที่ถูกดัดแปลง $(12)_{plg}$

2) การลบ

การลบถือเป็นกลวิธีที่ถูกใช้บ่อยที่สุดในสถานการณ์การลักลอกที่เกิดขึ้นจริง ทั้งนี้ การลบสามารถทำได้โดยตัดทอนวลีหรืออนุพากย์ที่มีลักษณะต้องตามที่ระบุไว้ในหัวข้อ “ลักษณะของข้อความที่ดัดแปลงได้” ที่ได้กล่าวไปแล้วข้างต้น การดัดแปลงข้อมูลด้วยกลวิธีการลบนี้กำหนดให้ผู้จำลองการลักลอกเลือกลบวลีหรืออนุพากย์ในข้อความต้นฉบับเพียง 1 ครั้งเท่านั้น โดยผู้จำลองการลักลอกไม่ต้องลบข้อความที่เลือกนั้นออกจากต้นฉบับจริง ๆ เพียงแค่กำกับป้ายกำกับ **<xdel>** และ **</xdel>** คร่อมข้อความที่ต้องการลบเท่านั้น ดังตัวอย่างต่อไปนี้

ตัวอย่าง (13)

$(13)_{src}$ ขั้นที่1 (การตระหนักรู้) กลุ่มการปรึกษาเชิงจิตวิทยาแนวความหมายในชีวิต เริ่มด้วยการให้สมาชิกที่ร่วมกลุ่มตระหนักในสถานการณ์หรือประสบการณ์ชีวิตของตนเอง ตระหนักในลักษณะเฉพาะตน ความต้องการที่จะต่างๆ และคุณค่า ความหมายในชีวิต ที่มีอยู่ในประสบการณ์หนึ่งๆ รวมถึง ความรู้สึกด้านลบ อาทิเช่น ความท้อแท้หมดหวังในชีวิต รู้สึกว่าตนเองไม่มีคุณค่า ความกังวลและความกลัว ในการดำเนินกลุ่มครั้งนี้ ผู้นำกลุ่มได้เริ่มต้นโดยให้เยาวชนที่เป็นสมาชิกกลุ่มตระหนักในสุขภาวะของตนเอง ผ่านการบอกเล่าเรื่องราว ตามแบบฝึกหัด"การเดินทางของรถ" โดยเปรียบตัวเยาวชนหากแต่ละคนเป็น "รถ" คันหนึ่ง จากนั้นแล้วให้เขียนเป็นความเรียง หรือวาดเป็นภาพ ก็ได้ และสุดท้าย เยาวชนที่เป็นสมาชิกกลุ่มหากแต่ละคนได้แลกเปลี่ยนประสบการณ์การสำรวจและตระหนักในสุขภาวะของตนในกลุ่ม

$(13)_{plg}$ ขั้นที่1 (การตระหนักรู้) กลุ่มการปรึกษาเชิงจิตวิทยาแนวความหมายในชีวิต เริ่มด้วยการให้สมาชิกที่ร่วมกลุ่มตระหนักในสถานการณ์หรือประสบการณ์ชีวิตของตนเอง ตระหนักในลักษณะเฉพาะตน ความต้องการที่จะต่างๆ และคุณค่า ความหมายในชีวิต ที่มีอยู่ในประสบการณ์หนึ่งๆ รวมถึง ความรู้สึกด้านลบ **<xdel>อาทิเช่น ความท้อแท้หมดหวังในชีวิต รู้สึกว่าตนเองไม่มีคุณค่า ความกังวลและความกลัว</xdel>** ในการดำเนินกลุ่มครั้งนี้ ผู้นำกลุ่มได้เริ่มต้นโดยให้เยาวชนที่เป็นสมาชิกกลุ่มตระหนักในสุขภาวะของตนเอง ผ่านการบอกเล่าเรื่องราว ตามแบบฝึกหัด" การเดินทางของรถ" โดยเปรียบตัวเยาวชนหากแต่ละคนเป็น "รถ" คันหนึ่ง จากนั้นแล้วให้เขียนเป็นความเรียง หรือวาดเป็นภาพ ก็ได้ และสุดท้าย เยาวชนที่เป็นสมาชิกกลุ่มหากแต่ละคนได้แลกเปลี่ยนประสบการณ์การสำรวจและตระหนักในสุขภาวะของตนในกลุ่ม

ตัวอย่างที่ 13 ข้างต้นแสดงให้เห็นการกำกับข้อความที่ต้องการลบในข้อความต้นฉบับ $(13)_{src}$ โดยกำกับวลีที่ขึ้นด้วยคำเชื่อมเด่น “อาทิเช่น ความท้อแท้หมดหวังในชีวิต รู้สึกว่าตนเองไม่มีคุณค่า ความกังวลและความกลัว” ที่ทำหน้าที่แสดงตัวอย่างของคำนาม “ความรู้สึกด้านลบ” ด้วยป้ายกำกับ <xdel> และ </xdel> คร่อมวลีที่ขึ้นต้นด้วยคำเชื่อมเด่นดังกล่าวจนได้เป็นข้อความที่ถูกดัดแปลง $(13)_{plg}$

ตัวอย่าง (14)

$(14)_{src}$ จิตรกรรมไทยมีมาตั้งหากแต่ในสมัยอดีต ปรากฏเป็นงานภาพเขียนหรือรูปเขียน สามารถที่จะวาดด้วยสีเดียวหรือสีมากกว่าสองสีก็ได้ โดยภาพเกิดขึ้นจากความคิดจินตนาการของศิลปินหรือจิตรกร โดยอาจมีลักษณะเป็นภาพเสมือนจริงตามธรรมชาติหรือเป็นภาพในจินตนาการก็ได้ อีกประเภทหนึ่งคืองานจิตรกรรมไทยที่มีลักษณะเป็นงานช่างฝีมือ เรียกว่าประณีตศิลป์ หรือศิลปหัตถกรรมบางประเภทจะใช้วิธีการที่แตกต่างจากงานวาดรูประบายสี อาทิ งานมุกประดับ งานลงรักปิดทอง ที่เรียกว่า ลายรดน้ำ

$(14)_{plg}$ จิตรกรรมไทยมีมาตั้งหากแต่ในสมัยอดีต ปรากฏเป็นงานภาพเขียนหรือรูปเขียน สามารถที่จะวาดด้วยสีเดียวหรือสีมากกว่าสองสีก็ได้ โดยภาพเกิดขึ้นจากความคิดจินตนาการของศิลปินหรือจิตรกร โดยอาจมีลักษณะเป็นภาพเสมือนจริงตามธรรมชาติหรือเป็นภาพในจินตนาการก็ได้ อีกประเภทหนึ่งคืองานจิตรกรรมไทยที่มีลักษณะเป็นงานช่างฝีมือ เรียกว่าประณีตศิลป์ หรือศิลปหัตถกรรมบางประเภทจะใช้วิธีการที่แตกต่างจากงานวาดรูประบายสี อาทิ งานมุกประดับ งานลงรักปิดทอง **<xdel>ที่เรียกว่าลายรดน้ำ</xdel>**

ตัวอย่าง (14) นี้แสดงการกำกับอนุพากย์ที่มีกริยาแท้ประเภทคุณานุประโยค “ที่เรียกว่า ลายรดน้ำ” ที่ต้องการจะลบออกจากข้อความต้นฉบับ $(14)_{src}$ อนุพากย์ดังกล่าวทำหน้าที่ขยายคำนาม “งานลงรักปิดทอง” สอดคล้องกับลักษณะของอนุพากย์ที่สามารถดัดแปลงได้ตามหลักของงานวิจัยชิ้นนี้ จะสังเกตได้ว่าอนุพากย์ดังกล่าวถูกกำกับป้ายกำกับ <xdel> และ </xdel> คร่อมด้านหน้าและหลังของข้อความเพื่อระบุขอบเขตของข้อความที่ต้องการลบ ดังแสดงให้เห็นในข้อความที่ถูกดัดแปลง $(14)_{plg}$

3) การย้ายที่

การย้ายที่ถือเป็นอีกกลวิธีหนึ่งที่สามารถใช้ในการดัดแปลงข้อความได้ โดยผู้จำลองการลักลอกสามารถเลือกย้ายวลีหรืออนุพากย์ที่มีลักษะณะต้องตามที่ระบุไว้ในหัวข้อ “ลักษณะของข้อความที่ดัดแปลงได้” ได้ตามความเหมาะสม อย่างไรก็ดี การย้ายที่ดังกล่าวไม่ควรทำให้ใจความสำคัญของข้อความผิดเพี้ยนไปจากเดิมมาก

ในการย้ายที่ข้อความ ให้ผู้จำลองการลักลอกกำกับป้ายกำกับข้อความ **<mvor>** และ **</mvor>** คร่อมข้อความที่เลือกย้ายที่ จากนั้นจึงคัดลอกข้อความดังกล่าวไปวางไว้ยังตำแหน่งใหม่ที่

ต้องการย้ายไป พร้อมทั้งกำกับป้ายกำกับข้อความ **<mved>** และ **</mved>** คร่อมข้อความที่คัดลอกไปวางไว้ในตำแหน่งใหม่ ดังตัวอย่างต่อไปนี้

ตัวอย่าง (15)

$(15)_{src}$ เนื่องจากบทละครเปลี่ยนบริบทมาเป็นสังคมไทยร่วมสมัย เครื่องหากแต่งกายที่ปรากฏจึงมีลักษณะของการเป็นเครื่องหากแต่งกายที่สามารถที่จะสะท้อนให้เห็นถึงรูปทรงของเครื่องหากแต่งกายเฉพาะของชายรักเพศเดียวกันที่เป็นอยู่ในปัจจุบัน ที่ใส่ใจในรายละเอียดที่เกี่ยวข้องกับแฟชั่นและการหากแต่งกายค่อนข้างสูง เครื่องหากแต่งกายที่ปรากฏจึงนอกจากที่จะต้องสามารถที่จะสะท้อนลักษณะของความนิยมของกระแสแฟชั่นเครื่องหากแต่งกายชายในปัจจุบันแล้วนั้น ยังต้องสามารถที่จะแสดงรสนิยมทางเพศของผู้สวมใส่ กล่าวคือ ต้องสามารถที่จะเสนอลักษณะของความโดเด่นในส่วนของที่เกี่ยวข้องกับรสนิยมด้านแฟชั่นของตัวละครได้ โดยเฉพาะยิ่งในตัวละครของ บาส ที่มีอาชีพเป็นบรรณาธิการนิตยสารแฟชั่น การปรากฏตัวของบาสนั้นจะอยู่ในเครื่องหากแต่งกายที่แตกต่างกันอย่างสิ้นเชิง คล้ายกับเขาพยายามหากแต่งตัวให้มีลักษณะเฉพาะไปในหากแต่ละวันไม่เหมือนกัน

$(15)_{plg}$ **<mvor>เนื่องจากบทละครเปลี่ยนบริบทมาเป็นสังคมไทยร่วมสมัย</mvor>** เครื่องหากแต่งกายที่ปรากฏจึงมีลักษณะของการเป็นเครื่องหากแต่งกายที่สามารถที่จะสะท้อนให้เห็นถึงรูปทรงของเครื่องหากแต่งกายเฉพาะของชายรักเพศเดียวกันที่เป็นอยู่ในปัจจุบัน ที่ใส่ใจในรายละเอียดที่เกี่ยวข้องกับแฟชั่นและการหากแต่งกายค่อนข้างสูง **<mved>เนื่องจากบทละครเปลี่ยนบริบทมาเป็นสังคมไทยร่วมสมัย</mved>** เครื่องหากแต่งกายที่ปรากฏจึงนอกจากที่จะต้องสามารถที่จะสะท้อนลักษณะของความนิยมของกระแสแฟชั่นเครื่องหากแต่งกายชายในปัจจุบันแล้วนั้น ยังต้องสามารถที่จะแสดงรสนิยมทางเพศของผู้สวมใส่ กล่าวคือ ต้องสามารถที่จะเสนอลักษณะของความโดเด่นในส่วนของที่เกี่ยวข้องกับรสนิยมด้านแฟชั่นของตัวละครได้ โดยเฉพาะยิ่ง ในตัวละครของ บาส ที่มีอาชีพเป็นบรรณาธิการนิตยสารแฟชั่น การปรากฏตัวของบาสนั้นจะอยู่ในเครื่องหากแต่งกายที่แตกต่างกันอย่างสิ้นเชิง คล้ายกับเขาพยายามหากแต่งตัวให้มีลักษณะเฉพาะไปในหากแต่ละวันไม่เหมือนกัน

ตัวอย่างที่ 15 ข้างต้นแสดงให้เห็นถึงการย้ายอนุพากย์ที่มีกริยาแท้ประเภทวิเศษณานุประโยค “เนื่องจากบทละครเปลี่ยนบริบทมาเป็นสังคมไทยร่วมสมัย” จากเดิมที่อยู่ในตำแหน่งต้นย่อหน้าเป็นตำแหน่งท้ายประโยคที่มีกริยาวลี “จึงมีลักษณะ” อันเป็นกริยาวลีที่วิเศษณานุประโยคดังกล่าวทำหน้าที่ขยายอยู่ และจะเห็นได้ว่าการย้ายตำแหน่งอนุพากย์ดังกล่าวมิได้ทำให้ใจความสำคัญของย่อหน้าต้นฉบับเปลี่ยนแปลงไป ทั้งนี้ ขอให้ผู้จำลองการลักลอกสังเกตการกำกับข้อมูลในย่อหน้าที่ถูกดัดแปลง $(15)_{plg}$ จะเห็นว่าป้ายกำกับ <mvor> และ </mvor> ถูกกำกับคร่อมวิเศษณานุประโยคที่ต้องการย้ายในตำแหน่งเดิมคือตอนต้นของย่อหน้า และการกำกับป้ายกำกับ <mved> และ </mved> ถูกกำกับคร่อมวิเศษณานุประโยคที่ถูกย้ายไปยังตำแหน่งใหม่ในย่อหน้า

ตัวอย่าง (16)

$(16)_{src}$ จากการศึกษาจะเห็นได้ว่า เมื่อเก็บรักษาไมโครแคปซูลเป็นระยะเวลา 3 เดือน ไมโครแคปซูลยังสามารถที่จะในการต้านออกซิเดชัน (ภาพที่ 4.16 และภาพที่ 4.18) ทั้งนี้เนื่องจากไมโครแคปซูลเมื่อเริ่มต้นเก็บรักษามีฤทธิ์การต้านออกซิเดชันสูง ดังนั้นกระบวนการเอนแคปซูเลชันจะช่วยรักษาเสถียรภาพของสารต้านออกซิเดชัน โดยกระบวนการเอนแคปซูเลชันจะช่วยชะลอการลดน้อยลงของฤทธิ์ต้านออกซิเดชันอันเนื่องมาจากสิ่งแวดล้อมรอบด้านความร้อน แสง และอากาศ ซึ่งสอดคล้องกับการศึกษาของ Kirby et al. (1991) ที่พบว่ากระบวนการเอนแคปซูเลชันจะช่วยยืดอายุการเก็บรักษาของสารต้านออกซิเดชันเมื่อเปรียบเทียบกับการเก็บรักษาในสภาวะสารละลาย

$(16)_{plg}$ จากการศึกษาจะเห็นได้ว่า เมื่อเก็บรักษาไมโครแคปซูลเป็นระยะเวลา 3 เดือน ไมโครแคปซูลยังสามารถที่จะในการต้านออกซิเดชัน (ภาพที่ 4.16 และภาพที่ 4.18) ทั้งนี้เนื่องจากไมโครแคปซูลเมื่อเริ่มต้นเก็บรักษามีฤทธิ์การต้านออกซิเดชันสูง **<mved>โดยกระบวนการเอนแคปซูเลชันจะช่วยชะลอการลดน้อยลงของฤทธิ์ต้านออกซิเดชันอันเนื่องมาจากสิ่งแวดล้อมรอบด้านความร้อน แสง และอากาศ<mved>** ดังนั้นกระบวนการเอนแคปซูเลชันจะช่วยรักษาเสถียรภาพของสารต้านออกซิเดชัน **<mvor>โดยกระบวนการเอนแคปซูเลชันจะช่วยชะลอการลดน้อยลงของฤทธิ์ต้านออกซิเดชันอันเนื่องมาจากสิ่งแวดล้อมรอบด้านความร้อน แสง และอากาศ<mvor>** ซึ่งสอดคล้องกับการศึกษาของ Kirby et al. (1991) ที่พบว่ากระบวนการเอนแคปซูเลชันจะช่วยยืดอายุการเก็บ

รักษาของสารต้านออกซิเดชันเมื่อเปรียบเทียบกับการเก็บรักษาในสภาวะสารละลาย

จากตัวอย่าง (16) ข้างต้นจะเห็นได้มีการย้ายอนุพากย์ที่มีกริยาแท้ “โดยกระบวนการเอนแคปซูเลชันจะช่วยชะลอการลดน้อยลงของฤทธิ์ต้านออกซิเดชันอันเนื่องมาจากสิ่งแวดล้อมรอบด้าน ความร้อน แสง และอากาศ” ไปไว้ในตำแหน่งก่อนหน้าตำแหน่งเดิมโดยอาศัยหลักความเป็นเหตุผลที่สามารถกล่าวถึงเรื่องใดก่อนก็ได้ และแม้ว่าอนุพากย์ที่ถูกย้ายที่จะเป็นอนุพากย์ที่ประกอบขึ้นจากอนุพากย์ย่อยมากกว่า 1 อนุพากย์ แต่ในงานวิจัยชิ้นก็ถือว่าสามารถดัดแปลงโดยการย้ายที่ได้หากได้ข้อความที่สมเหตุสมผลและไม่เปลี่ยนใจความสำคัญของย่อหน้าต้นฉบับ ทั้งนี้ ขอให้ผู้จำลองการลักลอกสังเกตการกำกับข้อมูลในย่อหน้าที่ถูกดัดแปลง $(16)_{plg}$ จะเห็นว่าป้ายกำกับ <mvor> และ </mvor> ถูกกำกับคร่อมอนุพากย์ที่มีกริยาแท้ที่ต้องการย้ายในตำแหน่งเดิม และการกำกับป้ายกำกับ <mved> และ </mved> ถูกกำกับคร่อมอนุพากย์ที่ถูกย้ายไปยังตำแหน่งใหม่ในย่อหน้า

ในส่วนท้ายของหัวข้อนี้ ผู้วิจัยขอสรุปป้ายกำกับข้อความที่ใช้ในการดัดแปลงข้อความอีกครั้งหนึ่งเพื่อให้ผู้จำลองการลักลอกเห็นภาพรวมของงานที่ชัดเจน ดังนี้

กลวิธีดัดแปลงข้อความ		**ป้ายกำกับต้นข้อความ**	**ป้ายกำกับท้ายข้อความ**
1) การแทรก		<xins>	</xins>
2) การลบ		<xdel>	</xdel>
3) การย้ายที่	กำกับข้อความในตำแหน่งเดิม	<mvor>	</mvor>
	กำกับข้อความในตำแหน่งใหม่ที่ถูกย้ายไป	<mved>	</mved>

ไฟล์ข้อมูลต้นฉบับที่ท่านได้รับจะมีการกำกับข้อมูลบางประเภทอยู่บ้างแล้ว ท่านสามารถดัดแปลงข้อความได้ตามปกติโดยไม่ต้องสนใจการกำกับข้อมูลดังกล่าว แต่ให้คงการกำกับข้อมูลที่มีอยู่ในไฟล์ต้นฉบับไว้

- **ค่าตอบแทน**

ในการดัดแปลงข้อความในย่อหน้าต้นฉบับแต่ละย่อหน้า ท่านจะได้รับค่าตอบแทนย่อหน้าละ 5 บาททั้งนี้ ย่อหน้าที่ท่านดัดแปลงข้อความจะได้รับการตรวจสอบโดยผู้วิจัย และผู้วิจัยอาจร้องขอให้ท่านแก้ไขย่อหน้าที่ผ่านการดัดแปลงแล้วใหม่ได้ย่อหน้าละ 1 ครั้ง

- **รายการอ้างอิง**

นลินี อินต๊ะซาว. (2556). *การแยกอนุพากย์ภาษาไทยด้วยการใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน*. (วิทยานิพนธ์ปริญญามหาบัณฑิต), จุฬาลงกรณ์มหาวิทยาลัย.

ภาคผนวก ง

คำชี้แจงสำหรับผู้จำลองการลักลอกโดยถอดความ

เนื่องด้วยผู้วิจัยกำลังทำวิทยานิพนธ์หัวข้อ “การตรวจเทียบภายนอกหาการลักลอกในงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความ (Extrinsic Plagiarism Detection in Academic Texts Using a Support Vector Machine Model and Text Similarity Measurement)” ในขั้นตอนการวิจัยหัวข้อดังกล่าว ผู้วิจัยจำเป็นต้องทดลองสร้างคลังข้อมูลจำลองการลักลอกขึ้น เพื่อนำมาทดลองวัดประสิทธิภาพของวิธีการตรวจจับการลักลอกที่นำเสนอในงานวิทยานิพนธ์

ในการนี้ ผู้วิจัยจึงขอรบกวนให้ท่านจำลองการลักลอกขึ้น โดย**ถอดความ (paraphrase)** ข้อความที่ผู้วิจัยได้เตรียมไว้ให้เป็นย่อหน้า โดยมีรายละเอียดดังต่อไปนี้

- **ข้อมูลต้นฉบับ**

ข้อมูลต้นฉบับสำหรับจำลองการลักลอกที่ท่านได้รับจะมีลักษณะเป็นย่อหน้าของข้อความเชิงวิชาการ จำนวนทั้งสิ้น 625 ย่อหน้า จำแนกตามสาขาวิชาและขนาดของย่อหน้า ตามรายละเอียดดังนี้

สาขาวิชา	วิทยาศาสตร์			มนุษยศาสตร์และสังคมศาสตร์		
ขนาด	สั้น	กลาง	ยาว	สั้น	กลาง	ยาว
จำนวนคำ/ย่อหน้า	50-100	101-150	151-200	50-100	101-150	151-200
จำนวนย่อหน้าที่ได้รับ	104	104	104	104	105	104

- **ไฟล์ข้อมูลต้นฉบับ**

ย่อหน้าที่ท่านจะได้รับจะถูกบันทึกในรูปแบบไฟล์สกุล .txt ไฟล์ละ 1 ย่อหน้า โดยชื่อไฟล์ถูกตั้งเป็นอักขระ 8 หลัก ซึ่งบ่งชี้ข้อมูลต่อไปนี้

1) อักขระหลักที่ 1-2 ใช้จำแนกสาขาวิชาของย่อหน้าต้นฉบับ ดังนี้

อักขระ SC หมายถึง ย่อหน้าที่มีเนื้อหาทางวิทยาศาสตร์

อักขระ HS หมายถึง ย่อหน้าที่มีเนื้อหาทางมนุษยศาสตร์และสังคมศาสตร์

2) อักขระหลักที่ 3 ใช้จำแนกขนาดของย่อหน้าต้นฉบับ ดังนี้

อักขระ S หมายถึง ย่อหน้าขนาดสั้น มีความยาว 50-100 คำ

อักขระ M หมายถึง ย่อหน้าขนาดกลาง มีความยาว 101-150 คำ

อักขระ L หมายถึง ย่อหน้าขนาดยาว มีความยาว 151-200 คำ

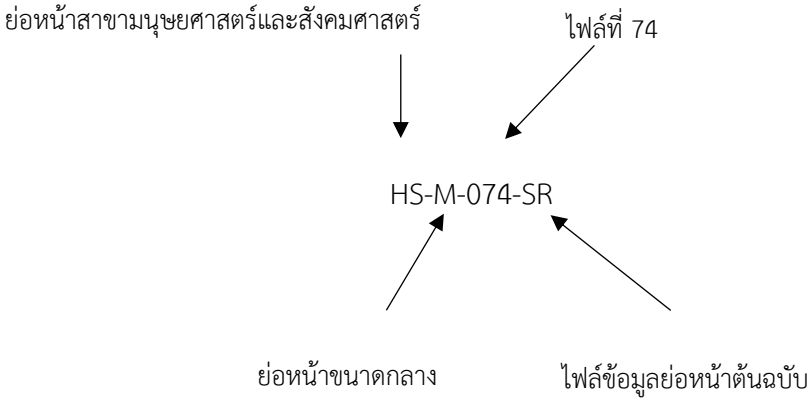
3) อักขระหลักที่ 4-6 ใช้จำแนกลำดับที่ของย่อหน้าต้นฉบับในสาขาวิชาและขนาดเดียวกัน

4) อักขระหลังที่ 7-8 ใช้จำแนกความเป็นข้อมูลต้นฉบับหรือข้อมูลที่ผ่านการลักลอก ดังนี้

อักขระ SR หมายถึง ย่อหน้าต้นฉบับ

อักขระ PA หมายถึง ย่อหน้าที่ผ่านการถอดความแล้ว

<u>ตัวอย่าง</u>



- **ภาระงาน**

ให้ท่าน**<u>ถอดความ (paraphrase)</u>** ต้นฉบับแต่ละไฟล์ตามขั้นตอนดังนี้

1) เปิดไฟล์ย่อหน้าต้นฉบับด้วยโปรแกรม Notepad, Notepad++, หรือ Microsoft Word ตามแต่ท่านสะดวก
2) ถอดความข้อความที่ปรากฏในย่อหน้า
3) กรณีที่ข้อความมีการอ้างอิงทางวิชาการหรือมีภาษาต่างประเทศปน ท่านสามารถพิจารณาตัดข้อความส่วนดังกล่าวออกไปได้
4) บันทึกไฟล์เป็นไฟล์ใหม่สกุล .txt โดยตั้งชื่อไฟล์อักขระที่ 1-6 ตามไฟล์ย่อหน้าต้นฉบับที่ลักลอก ส่วนอักขระที่ 7-8 ให้เปลี่ยนจากอักขระ SR เป็น PA
5) จัดเก็บไฟล์ที่ถอดความแล้วแยกตามสาขาวิชาและขนาด

- **ตัวอย่างการถอดความ**

ตัวอย่าง (1)

ข้อความต้นฉบับ

บทบาทแบบคนภายนอก หมายถึง การเข้าไปสู่สนามของผู้วิจัยเพื่อการสังเกตโดยตรงว่าสมาชิกในสังคมนั้นพูดคุยสนทนาประพฤติปฏิบัติอะไร อย่างไร ตลอดจนทำการสัมภาษณ์พูดคุยกับสมาชิกในกลุ่มที่ผู้วิจัยเข้าไปสังเกต บทบาทนี้ผู้วิจัยเปิดเผยตนเอง โดยการสังเกตเฝ้ามองอยู่ห่างๆ ไม่เข้าไปร่วมกิจกรรมด้วย จุดอ่อนของบทบาทแบบคนนอก คือ คุณภาพและความลึกซึ้งของข้อมูลซึ่งขึ้นอยู่กับความไว้วางใจของผู้ให้ข้อมูลและระยะเวลาในการอยู่ในพื้นที่ของผู้วิจัย

ข้อความที่ผ่านการถอดความ

บทบาทแบบคนภายนอก คือ การที่ผู้วิจัยเข้าไปสังเกตโดยตรงในพื้นที่ว่าสมาชิกในสังคมนั้นๆ พูดคุยกันเรื่องอะไร ปฏิบัติตัวอย่างไร นอกจากนี้ยังต้องสัมภาษณ์สมาชิกที่เข้าไปสังเกตด้วย แต่ผู้วิจัยจะสังเกตโดยเฝ้าดูอยู่ห่าง ๆ ไม่เข้าร่วมกิจกรรมโดยตรง อย่างไรก็ตาม บทบาทแบบคนนอกก็มีข้อด้อย คือ คุณภาพและความลึกซึ้งของข้อมูลจะขึ้นอยู่กับระยะเวลาของการอยู่ในพื้นที่ รวมถึงความไว้วางใจที่ผู้ให้ข้อมูลมีต่อผู้วิจัย

ตัวอย่าง (2)

ข้อความต้นฉบับ

พื้นที่สีเขียวที่ทำการศึกษาภายในโครงการเพอร์เฟค เพลส รัตนาธิเบศร์ มีทั้งหมด 9 แห่ง เป็นพื้นที่สวนสาธารณะ จำนวน 1 แห่ง พื้นที่สวนขนาดกลาง จำนวน 1 แห่ง, พื้นที่สวนหย่อมที่มีลักษณะตำแหน่งที่ตั้งอยู่ติดถนนหลัก หน้าโครงการ จำนวน 1 แห่ง, พื้นที่สวนหย่อมที่มีลักษณะที่ตั้งอยู่ในบริเวณใกล้บ้านพักอาศัย จำนวน 3 แห่ง, พื้นที่สวนหย่อมที่มีลักษณะที่ตั้งอยู่บริเวณถนนปลายตัน จำนวน 2 แห่ง และสวนชั่วคราวซึ่งเป็นพื้นที่ขายในอนาคต จำนวน 1 แห่ง ส่วนพื้นที่สีเขียวที่ทำการศึกษาภายในโครงการเพอร์เฟคเพลส ราชพฤกษ์ มีทั้งหมด 5 แห่ง เป็นพื้นที่สวนสาธารณะ จำนวน 1 แห่ง, พื้นที่สวนขนาดกลาง จำนวน 1 แห่ง, พื้นที่

สวนหย่อมที่มีลักษณะตำแหน่งที่ตั้งอยู่ติดถนนหลัก หน้าโครงการ จำนวน 1 แห่ง, พื้นที่สวนหย่อมที่มีลักษณะที่ตั้งอยู่ในบริเวณใกล้บ้านพักอาศัย จำนวน 1 แห่ง และพื้นที่สวนหย่อมที่มีลักษณะที่ตั้งอยู่บริเวณถนนปลายตัน จำนวน 1 แห่ง

ข้อความที่ผ่านการถอดความ

พื้นที่สีเขียวที่ศึกษาในโครงการเพอร์เฟค เพลส รัตนาธิเบศร์ มีทั้งสิ้น 9 ที่ แบ่งเป็นพื้นที่สวนสาธารณะ 1 ที่ พื้นที่สวนขนาดกลาง 1 ที่ พื้นที่สวนหย่อมที่อยู่ติดถนนหลักหน้าโครงการ 1 ที่พื้นที่สวนหย่อมที่มีอยู่ใกล้บ้านพักอาศัย 3 ที่ พื้นที่สวนหย่อมที่มีอยู่บริเวณถนนปลายตัน 2 ที่ และสวนชั่วคราวซึ่งจะเป็นพื้นที่ขายในอนาคตอีก1 ที่ ส่วนพื้นที่สีเขียวที่จะศึกษาภายในโครงการเพอร์เฟคเพลส ราชพฤกษ์ มี 5 ที่ ได้แก่ พื้นที่สวนสาธารณะ 1 ที่ พื้นที่สวนขนาดกลาง จำนวน 1 ที่ พื้นที่สวนหย่อมที่อยู่ติดถนนหลักหน้าโครงการ 1 ที่ พื้นที่สวนหย่อมที่อยู่ใกล้บ้านพักอาศัย 1 ที่ และพื้นที่สวนหย่อมที่อยู่บริเวณถนนปลายตัน 1 ที่

ตัวอย่าง (3)

ข้อความต้นฉบับ

ในวรรณคดีสันสกฤต ยักษ์เป็นอมนุษย์ที่มีรูปร่างหน้าตาอัปลักษณ์ดุร้าย ชอบกินเนื้อมนุษย์และซากศพ แต่ในทางพุทธศาสนา ยักษ์มีทั้งพวกที่มีรูปกายงดงามและพวกที่มีรูปร่างหน้าตาอัปลักษณ์ โดยยักษ์ที่เคยสร้างกุศลจะมีรูปกายงดงาม ทรงสง่าราศี ส่วนยักษ์ที่เคยสร้างบาปกรรมจะมีรูปร่างหน้าตาน่าเกลียดน่ากลัว

ข้อความที่ผ่านการถอดความ

ยักษ์แบ่งออกเป็น 2 ประเภท คือ ยักษ์ในทางวรรณคดีสันสกฤตและยักษ์ในทางพุทธศาสนา ยักษ์ในวรรณคดีสันสกฤตเป็นอมนุษย์ที่มีรูปร่างหน้าตาอัปลักษณ์ ดุร้าย ชอบกินเนื้อมนุษย์และซากศพ ส่วนยักษ์ในทางพุทธศาสนาแบ่งเป็นพวกที่มีรูปกายงดงาม สง่า มีราศี ซึ่งเกิดจากกรรมดีที่เคยสร้างไว้ และพวกที่มีหน้าตาอัปลักษณ์ น่าเกลียดน่ากลัว ซึ่งเกิดจากกรรมชั่วที่เคยสร้างไว้

- **ค่าตอบแทน**

ในการถอดความย่อหน้าต้นฉบับแต่ละย่อหน้า ท่านจะได้รับค่าตอบแทนย่อหน้าละ 20 บาท ทั้งนี้ ย่อหน้าที่ท่านถอดความจะได้รับการตรวจสอบโดยผู้วิจัย และผู้วิจัยอาจร้องขอให้ท่านแก้ไขย่อหน้าที่ผ่านการถอดความแล้วใหม่ได้ย่อหน้าละ 1 ครั้ง

ภาคผนวก จ

ตัวอย่างชุดข้อมูลทดลองสำหรับให้ผู้เชี่ยวชาญระบุค่าความละม้ายของข้อความ

ชุดข้อมูลนี้จัดทำขึ้นเพื่อเก็บข้อมูลเพื่อใช้ในการทำวิทยานิพนธ์เรื่อง “การตรวจเทียบภายนอกหาการลักลอกในงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความ” ของนายศุภวัจน์ แต่รุ่งเรือง นิสิตระดับดุษฎีบัณฑิต สาขาวิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ผู้วิจัยใคร่ขอความร่วมมือจากท่าน กรุณาตอบแบบสอบถามให้สมบูรณ์ ข้อมูลทั้งหมดที่ท่านตอบจะเป็นประโยชน์อย่างยิ่งสำหรับงานวิจัยครั้งนี้ ทั้งนี้ ผู้วิจัยขอขอบคุณท่านมา ณ ที่นี้ด้วย

<u>คำชี้แจง</u>

ในส่วนต่อไปนี้ ผู้วิจัยได้เตรียมคู่ของย่อหน้าไว้จำนวน 150 คู่ ให้ท่านอ่านย่อหน้าเปรียบเทียบกันภายในคู่และพิจารณาว่าแต่ละคู่มีความเหมือนกันมากน้อยเพียงใด จากนั้น **<u>ให้ท่านระบุตัวเลขตั้งแต่ 0% ถึง 100%</u>** โดย 0% หมายถึงคู่ของย่อหน้าไม่เหมือนกันเลย ส่วน 100% หมายถึงคู่ของย่อหน้าเหมือนกันทุกประการ

Case: 001

Text A:

4) ภาวะที่ทำให้เกิดการสูญเสียไบคาร์บอเนตออกจากร่างกาย ได้แก่การสูญเสียทางระบบทางดินอาหาร เช่น อุจจาระร่วง อาเจียนรุนแรง หรือการสูญเสียทางน้ำดี การสูญเสียทางระบบไต เช่น ไตวาย การได้รับสารยับยั้งเอนไซม์คาร์บอนิคแอนไฮเดรส เช่น ยากลุ่ม อะเซทาโซลาไมด์ (acetazolamide)

Text B:

4) ภาวะที่ทำให้เกิดการสูญเสียไบคาร์บอเนตออกจากร่างกาย อันได้แก่การสูญเสียทางระบบทางดินอาหาร อาทิเช่น อุจจาระร่วง อาเจียนรุนแรง หรือการสูญเสียทางน้ำดี การสูญเสียทางระบบไต อาทิเช่น ไตวาย การได้รับสารยับยั้งเอนไซม์คาร์บอนิคแอนไฮเดรส อาทิเช่น ยากลุ่ม อะเซทาโซลาไมด์ (acetazolamide)

เปอร์เซนต์ความเหมือน

Case: 002

Text A:

เฟดเดอริค เฮิร์ซเบิร์ก(23) ศึกษาเกี่ยวกับทัศนคติที่มีต่องานของวิศวกรและนักบัญชี จำนวน 200 คน ผลการศึกษาพบว่า ความพึงพอใจในการทำงาน และความไม่พึงพอใจในที่ทำงาน มิได้มีสาเหตุจากปัจจัยกลุ่มเดียว ทว่ามีสาเหตุมาจากปัจจัยสองกลุ่มโดย เฮิร์ซเบิร์กเสนอว่า ประกอบไปด้วย สิ่งที่ทำให้เกิดความพึงพอใจ (ปัจจัยจูงใจ) และสิ่งที่หากขาดไปจะทำให้เกิดความไม่พึงพอใจ (ปัจจัยค้ำจุน) โดยรายละเอียดของทฤษฎี 2 ปัจจัย มีดังนี้

Text B:

เฟดเดอริค เฮิร์ซเบิร์ก(23) ศึกษาเกี่ยวกับทัศนคติที่มีต่องานของวิศวกรและนักบัญชี จำนวน 200 คน ผลการศึกษาพบว่า ความพึงพอใจในการทำงาน และความไม่พึงพอใจในที่ทำงาน มิได้มีสาเหตุจากปัจจัยกลุ่มเดียว ทว่ามีสาเหตุมาจากปัจจัยสองกลุ่ม ประกอบไปด้วย สิ่งที่ทำให้เกิดความพึงพอใจ (ปัจจัยจูงใจ) และสิ่งที่หากขาดไปจะทำให้เกิดความไม่พึงพอใจ (ปัจจัยค้ำจุน) โดยรายละเอียดของทฤษฎี 2 ปัจจัย มีดังต่อไปนี้

เปอร์เซนต์ความเหมือน

Case: 003

Text A:

มีผู้สูงอายุที่ไม่มีปัญหาด้านการทำงานของร่างกาย 9 คน คิดเป็นร้อยละ 36 มีปัญหาในการได้ยิน 3 คน คิดเป็นร้อยละ 12 โดยมีลักษณะที่เกิดขึ้นคือหูตึง 2 คน (ร้อยละ 8) และหูได้ยินเป็นครั้งคราว 1 คน (ร้อยละ 4) มีปัญหาในการมองเห็น 4 คน คิดเป็นร้อยละ 16 โดยมีลักษณะที่เกิดขึ้นคือ ตามัว 1 คน (ร้อยละ 4) มองเห็นไม่ชัด 2 คน (ร้อยละ 8) และตาเป็นต้อ 1 คน (ร้อยละ 4) และมีปัญหาในการเคลื่อนไหว 9 คน คิดเป็นร้อยละ 36 โดยมีลักษณะที่เกิดขึ้นคือ การเดิน 3 คน (ร้อยละ 12) การลุก 5 คน (ร้อยละ 20) และท่าเดินผิดปกติ 1 คน (ร้อยละ 4)

Text B:

มีผู้สูงอายุที่ไม่มีปัญหาด้านการทำงานของร่างกาย 9 คน คิดเป็น 36% ผู้สูงอายุมีปัญหาด้านการได้ยิน 3 คน คิดเป็น 12% (แบ่งเป็นผู้สูงอายุที่หูตึง 2 คน คิดเป็น 8% และผู้สูงอายุที่หูได้ยินเป็นครั้งคราว 1 คน คิดเป็น 4%) ผู้สูงอายุที่มีปัญหาด้านการมองเห็น 4 คน คิดเป็น 16% (แบ่งเป็นผู้สูงอายุที่ตาพร่ามัว 1 คน คิดเป็น 4% ผู้สูงอายุที่มองเห็นไม่ชัด 2 คน คิดเป็น 8% และผู้สูงอายุที่ตาเป็นต้อ 1 คน คิดเป็น 4%) และผู้สูงอายุที่มีปัญหาด้านการเคลื่อนไหว 9 คน คิดเป็น 36% (แบ่งเป็นผู้สูงอายุที่มีปัญหาด้านการเดิน 3 คน คิดเป็น 12% ผู้สูงอายุที่มีปัญหาด้านการลุก 5 คน คิดเป็น 20% และผู้สูงอายุที่มีท่าเดินผิดปกติ 1 คน คิดเป็น 4%)

เปอร์เซนต์ความเหมือน

Case: 004

Text A:

ปัจจัยภายนอกที่สำคัญอีกประการหนึ่ง นั่นคือบทบาทของประเทศมหาอำนาจ และข้อพิจารณาทางการเมืองระหว่างประเทศต่อการทำความตกลงการค้าเสรี ซึ่งนักวิชาการหลายท่านพบว่าการทำความตกลงการค้าเสรีเป็นเครื่องมือหรือยุทธศาสตร์ในการดำเนินความสัมพันธ์ระหว่างประเทศ ทั้งของมหาอำนาจและประเทศอื่นๆ เช่น Sisira Jayasuriya และ Gary Magee จากหนังสือชื่อ Negotiating a preferential trading agreement: Issues, constraints and practical options ที่เชื่อว่ารัฐต่างๆ ใช้ความตกลงการค้าเสรีเพื่อทดสอบความสัมพันธ์อันเหนียวแน่นระหว่างกัน ส่งผลให้ผู้นำของประเทศที่เจรจากันหมายมั่นความสำเร็จในการเจรจา โดยวรรณกรรมที่เกี่ยวข้องกับมหาอำนาจและการเมืองระหว่างประเทศอาจแบ่งได้เป็น 2 กลุ่ม ดังนี้

Text B:

ปัจจัยภายนอกที่สำคัญอีกประการหนึ่ง นั่นคือบทบาทของประเทศมหาอำนาจ และข้อพิจารณาทางการเมืองระหว่างประเทศต่อการทำความตกลงการค้าเสรี ซึ่งนักวิชาการหลายท่านพบว่าการทำความตกลงการค้าเสรีเป็นเครื่องมือหรือยุทธศาสตร์ในการดำเนินความสัมพันธ์ระหว่างประเทศ ทั้งของมหาอำนาจและประเทศอื่นๆ อาทิเช่น Sisira Jayasuriya และ Gary Magee จากหนังสือชื่อ Negotiating a preferential trading agreement: Issues, constraints and practical options ที่เชื่อว่ารัฐต่างๆ ใช้ความตกลงการค้าเสรีเพื่อทดสอบความสัมพันธ์อันเหนียวแน่นระหว่างกัน ส่งผลให้ผู้นำของประเทศที่เจรจากันหมายมั่นความสำเร็จในการเจรจา โดยวรรณกรรมที่เกี่ยวข้องกับมหาอำนาจและการเมืองระหว่างประเทศอาจแบ่งได้เป็น 2 กลุ่ม ดังต่อไปนี้

เปอร์เซนต์ความเหมือน

Case: 005

Text A:

อย่างไรก็ตาม บุคคลที่ถือว่าเป็นตัวเร่งให้แนวความคิดทฤษฎีปัจเจกชนนิยมของล๊อคให้มีผลอย่างจริงจังในการปฏิวัติฝรั่งเศส ได้แก่ ฌอง ฌาคส์ รุสโซ (Jean Jacques Rousseau) โดยรุสโซเห็นว่า สิทธิเสรีภาพส่วนบุคคล เป็นลักษณะตามธรรมชาติของมนุษย์ ดังนั้นเมื่อมนุษย์ได้สูญเสียเสรีภาพ และความเสมอภาคของตนไปจากการรวมตัวกันเป็นสังคม มนุษย์จึงต้องค้นหารูปแบบของการรวมตัวกันเป็นสังคมที่สามารถให้หลักประกันอย่างเพียงพอแก่สิทธิและเสรีภาพ รวมถึงความเสมอภาคที่เป็นสิทธิตามธรรมชาติของตน เพราะฉะนั้น มนุษย์จึงต้องทำสัญญาประชาคมระหว่างกันขึ้น และ "รัฐ" จึงเป็น สิ่งจำเป็นที่ต้องมี เพื่อทำหน้าที่รับรองและคุ้มครองสิทธิและเสรีภาพ อันเป็นสิทธิตามกฎหมายที่ประชาชนทุกคนสามารถยกขึ้นใช้ยันรัฐได้อย่างเท่าเทียมกัน

Text B:

อย่างไรก็ตาม บุคคลที่ถือได้ว่าเป็นตัวเร่งให้แนวความคิดทฤษฎีปัจเจกชนนิยมของล๊อคให้มีผลอย่างจริงจังในการปฏิวัติฝรั่งเศส อันได้แก่ ฌอง ฌาคส์ รุสโซ (Jean Jacques Rousseau) โดยรุสโซเห็นว่า สิทธิเสรีภาพส่วนบุคคล เป็นลักษณะตามธรรมชาติของมนุษย์ ดังนั้นเมื่อมนุษย์ได้สูญเสียเสรีภาพ และความเสมอภาคของตนไปจากการรวมตัวกันเป็นสังคม มนุษย์จึงต้องคันหารูปแบบของการรวมตัวกันเป็นสังคมที่สามารถที่จะให้หลักประกันอย่างเพียงพอแก่สิทธิและเสรีภาพ รวมถึงความเสมอภาคที่เป็นสิทธิตามธรรมชาติของตน เพราะฉะนั้น มนุษย์จึงต้องทำสัญญาประชาคมระหว่างกันขึ้น และ "รัฐ" จึงเป็น สิ่งจำเป็นที่ต้องมี เพื่อทำหน้าที่รับรองและคุ้มครองสิทธิและเสรีภาพ อันเป็นสิทธิตามกฎหมายที่ประชาชนทุกคนสามารถที่จะยกขึ้นใช้ยันรัฐได้อย่างเท่าเทียมกัน

เปอร์เซนต์ความเหมือน

ประวัติผู้เขียนวิทยานิพนธ์

นายศุภวัจน์ แต่รุ่งเรือง เกิดเมื่อวันพุธที่ 1 มกราคม 2529 ที่กรุงเทพมหานคร สำเร็จการศึกษาศิลปศาสตรบัณฑิต เกียรตินิยมอันดับหนึ่ง สาขาวิชาภาษาไทย จากคณะมนุษยศาสตร์ มหาวิทยาลัยเชียงใหม่ ในปีการศึกษา 2550 จากนั้นได้เข้าทำงานในตำแหน่งผู้ช่วยวิจัยประจำโครงการแบบทดสอบมาตรฐานวัดสมรรถภาพการใช้ภาษาไทยของผู้พูดภาษาไทยเป็นภาษาแม่ ศูนย์ภาษาไทยสิรินธร จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2553 ได้เข้าศึกษาต่อในหลักสูตรอักษรศาสตรมหาบัณฑิต สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย และได้ผ่านการสอบวัดคุณสมบัติขั้นสูง เพื่อเปลี่ยนระดับเข้าสู่หลักสูตรอักษรศาสตรดุษฎีบัณฑิต สาขาวิชาภาษาศาสตร์ ในภาคการศึกษาต้น ปีการศึกษา 2554 ได้รับทุนอุดหนุนการศึกษาเฉพาะค่าเล่าเรียนประเภท 60/40 (ปีการศึกษา 2555-2557) และได้รับสนับสนุนทุนวิจัยจาก “ทุน 90 ปีจุฬาลงกรณ์มหาวิทยาลัย” กองทุนรัชดาภิเษกสมโภช รุ่นที่ 31 (2/2559)