การระบุคำไทยและคำทับศัพท์ด้วยแบบจำลองเอ็นแกรม

นายอัครพล เอกวงศ์อนันต์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2548

ISBN 974-53-2360-8

IDENTIFICATION OF THAI AND TRANSLITERATED WORDS BY N-GRAM MODELS

Mr. Akarapol Ekwonganan

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Arts in Linguistics
Department of Linguistics
Faculty of Arts
Chulalongkorn University
Academic Year 2005

ISBN 974-53-2360-8

หัวข้อวิทยานิพนธ์	การระบุคำไทยและคำทับศัพท์ด้วยแบบจำลองเอ็นแกรม
โดย	นายอัครพล เอกวงศ์อนันต์
สาขาวิชา	ภาษาศาสตร์
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะอักษรศาสตร์
(ศาสตราจารย์ ดร.ธีระพันธ์ เหลืองทองคำ)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุดาพร ลักษณียนาวิน)

.....อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล)

.....กรรมการ
(อาจารย์ ดร. ณัฐกร ทับทอง)

อัครพล เอกวงศ์อนันต์ : การระบุคำไทยและคำทับศัพท์ด้วยแบบจำลองเอ็นแกรม. (IDENTIFICATION OF THAI AND TRANSLITERATED WORDS BY N-GRAM MODELS) อาจารย์ที่ปรึกษา : ผศ. ดร. วิโรจน์ อรุณมานะกุล, จำนวน 270 หน้า. ISBN 974-53-2360-8.

วัตถุประสงค์ของการวิจัยครั้งนี้ เพื่อต้องหาสายอักขระเฉพาะสำหรับใช้ในการระบุภาษาของคำโดยใช้คลังข้อมูลคำไทย คำทับศัพท์ภาษาอังกฤษ ภาษาญี่ปุ่นและภาษาฝรั่งเศส และพัฒนาระบบการระบุภาษาของคำไทยและคำทับศัพท์ภาษาต่างประเทศโดยใช้สายอักขระเฉพาะและใช้แบบจำลองเอ็นแกรมขนาด 1-5 แกรม

คลังข้อมูลที่ใช้ในงานวิจัยนี้ คือ คลังข้อมูลคำไทย คำทับศัพท์ภาษาอังกฤษ ภาษาญี่ปุ่น ภาษาละ 10,000 คำ และคำทับศัพท์ภาษาฝรั่งเศส 1,000 คำ โดยเก็บจากข้อมูลที่พบในภาษาธรรมชาติซึ่งอาจจะไม่ได้ทับศัพท์ถูกต้องตามเกณฑ์ของราชบัณฑิตยสถานก็ได้ 80% ของคลังข้อมูลถูกนำมาใช้เพื่อหาสายอักขระเฉพาะและสร้างแบบจำลองเอ็นแกรมของแต่ละภาษา ในขณะที่อีก 20% ถูกใช้เพื่อการทดสอบระบบแบบต่างๆ

สายอักขระเฉพาะที่พบสะท้อนให้เห็นถึงลักษณะเฉพาะของแต่ละภาษาได้ในระดับหนึ่ง จึงมีผลให้ระบบที่ใช้สายอักขระเฉพาะในการระบุภาษาสามารถตัดสินภาษาได้ถูกต้อง 50.58% 48.71% 54.09% และ 20.40% สำหรับคำไทย คำทับศัพท์ภาษาอังกฤษ ภาษาญี่ปุ่น และฝรั่งเศส ตามลำดับ

เมื่อใช้แบบจำลองเอ็นแกรมในการระบุภาษา ระบบสามารถระบุภาษาของคำไทย คำทับศัพท์ภาษาอังกฤษ และญี่ปุ่นได้ถูกต้องกว่า 90% แต่ได้เพียงประมาณ 60% สำหรับคำทับศัพท์ฝรั่งเศส ผลที่ได้ยืนยันว่าขนาดของข้อมูลการฝึกมีผลต่อการทำงานของระบบการระบุภาษาทั้งสองระบบ นอกจากนี้ จากผลที่พบว่าระบบที่ใช้แบบจำลอง 3-แกรมให้ผลดีกว่าระบบที่ใช้ขนาดแกรมอื่นๆ ทำให้สรุปได้ว่า ขนาดของเอ็นแกรมมีผลต่อการทำงานของระบบการระบุภาษา

ภาควิชา	ภาษาศาสตร์	ลายมือชื่อนิสิต .....
สาขาวิชา	ภาษาศาสตร์	ลายมือชื่ออาจารย์ที่ปรึกษา.....
ปีการศึกษา	2548	

##4580259022 : MAJOR LINGUISTICS
KEY WORD: LANGUAGE IDENTIFICATION

AKARAPOL EKWONGANAN : IDENTIFICATION OF THAI AND TRANSLITERATED WORDS BY N-GRAM MODELS. THESIS ADVISOR : ASST. PROF. WIROTE AROONMANAKUN, Ph.D., 270 pp. ISBN 974-53-2360-8.

This research aims to find the unique character sequences of Thai and transliterated words (English, Japanese, and French), and implement language identification systems using unique character sequences and n-gram models (1-5 gram).

The corpora in this research consist of 10,000 Thai words, 10,000 English transliterated words, 10,000 Japanese transliterated words, and 1,000 French transliterated words. Transliterated words are collected from naturally occurring texts, even some of them are not conformed to the Royal Institute guidelines of transliteration. 80% of the corpus is used to extract unique character sequences and to build an n-gram language model of each language, while the other 20% is used for testing the systems.

The unique character sequences reflect some characteristics of the languages. As a result, the system using unique character sequence can identify languages correctly 50.58%, 48.71%, 54.09%, and 20.40% for Thai words, English, Japanese, and French transliterated words respectively.

When an n-gram language model is used, the system can identify languages correctly more than 90% for Thai, English and Japanese transliterated word, but only about 60% for French transliterated words. This confirms that the size of training corpus affects the performances of both systems. The results also show that the system using 3-gram model performs better than other n-gram models. Therefore, we can conclude that the size of n-gram does affect the performance of the language identification system.

Department	LINGUISTICS	Student's signature……
Field of study	LINGUISTICS	Advisor's signature…….....
Academic year	2005	

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลงได้ด้วยความช่วยเหลืออย่างดียิ่งของผู้ช่วยศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำปรึกษา แนะนำ ช่วยตรวจทาน แก้ไข งานวิจัยด้วยความเอาใจใส่เป็นอย่างดี และจุดประกายให้ผู้วิจัยทำงานวิจัยนี้

ผู้วิจัยขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร. สุดาพร ลักษณียนาวิน และ ดร. ณัฐกร ทับทอง ที่ได้กรุณาให้คำแนะนำและข้อคิดเห็นในการทำวิทยานิพนธ์ฉบับนี้ด้วยความเอาใจใส่ ทั้งยังช่วยตรวจแก้ไขให้สำเร็จสมบูรณ์ยิ่งขึ้น

ผู้วิจัยขอขอบพระคุณคณาจารย์ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ให้ความรู้ความเมตตาสั่งสอนทั้งทางด้านวิชาการและการทำงานวิจัย และผู้วิจัยขอกราบขอบพระคุณบิดา คุณน้าและคุณป้าที่ให้ความรัก ความเข้าใจ ความเอาใจใส่ ให้โอกาสและสนับสนุนทุกด้าน รวมทั้งเป็นกำลังใจให้แก่ผู้วิจัยตลอดมา สุดท้ายขอขอบคุณพี่ชายและเพื่อน ๆ ที่คอยให้คำแนะนำ ปรึกษา ในการทำวิจัย รวมทั้งเป็นกำลังใจให้งานวิจัยนี้สำเร็จด้วยดี

สารบัญ

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

ตารางที่ หน้า

สารบัญภาพและแผนภูมิ

ภาพและแผนภูมิ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เนื่องจากในปัจจุบัน ประเทศไทยมีการติดต่อกับต่างประเทศมากขึ้น จึงได้รับอิทธิพลทั้งทางวัฒนธรรมและเทคโนโลยีใหม่ๆ ภาษาที่ใช้ในปัจจุบัน จึงไม่เพียงพอต่อความต้องการ โดยเฉพาะศัพท์วิชาการและศัพท์เทคนิค มีการแพร่หลายอย่างรวดเร็ว จึงมีการสร้างคำหรือศัพท์ใหม่ๆ ขึ้นมาใช้ในภาษาไทย เพื่อเรียกชื่อสิ่งที่เกิดขึ้นใหม่ๆ จากต่างประเทศ ในรูปของการยืมคำ การแปล การบัญญัติศัพท์ใหม่ ฯลฯ ในการบัญญัติศัพท์ใหม่ๆนั้น ทางราชบัณฑิตยสถานได้วางหลักเกณฑ์ไว้ 2 ข้อ คือ 1. สร้างคำใหม่จากคำศัพท์ที่เป็นคำไทยแท้หรือจากบาลีสันสกฤต 2. หากไม่สามารถบัญญัติศัพท์ใหม่ได้ จะด้วยเหตุผลใดๆ ก็ตามก็ให้ใช้การทับศัพท์แทน การทับศัพท์จึงเป็นวิธีหนึ่งของการสร้างคำใหม่

คำทับศัพท์ หมายถึง คำภาษาต่างประเทศที่เขียนด้วยตัวอักษรไทย อาจเป็นคำสามานยนาม เช่น คอมพิวเตอร์ เทคโนโลยี ฟุตบอล หรือคำวิสามานยนาม เช่น แปซิฟิก เมดิเตอร์เรเนียน ไอบีเอ็ม ยูเนสโก ฯลฯ (ราชบัณฑิตยสถาน, 2538)

คำทับศัพท์เริ่มใช้กันอย่างแพร่หลาย และสามารถทับศัพท์จากภาษาต่างประเทศต่าง ๆ ได้มากมาย เห็นได้จากเกณฑ์ทางราชบัณฑิตยสถาน (2535) มีการกำหนดเกณฑ์การทับศัพท์ไว้ทั้งหมด 9 ภาษา คือ ภาษาอังกฤษ ภาษาฝรั่งเศส ภาษาญี่ปุ่น ภาษาเยอรมัน ภาษาอิตาลี ภาษาสเปน ภาษารัสเซีย ภาษาอาหรับ และภาษามลายู ตัวอย่างคำทับศัพท์จากราชบัณฑิตยสถานได้แก่

คำศัพท์ภาษาอังกฤษ	ราชบัณฑิตยสถานบัญญัติว่าเป็น	ทับศัพท์ว่าเป็น
computer	เครื่องจักรสมองกล	*คอมพิวเตอร์*
plastics	-	*พลาสติก*
bacteria	-	*บัคเตรี*

คำทับศัพท์เหล่านี้เขียนโดยใช้ตัวอักษรภาษาไทยและเขียนปะปนกับคำไทยทั่วไป ดังจะเห็นได้จากคำที่ขีดเส้นใต้ในตัวอย่างข้อความจากหนังสือเทคโนโลยีคอมพิวเตอร์และสารสนเทศ (ศรีไพร, 2544) ซึ่งจะแสดงให้เห็นการใช้ศัพท์เฉพาะทางดังนี้

หน่วยความจำหลัก (MAIN MEMORY UNIT) เป็นวงจรรวมหรือชิปที่ใช้บันทึกโปรแกรมและข้อมูล หน่วยความจำหลักจะบรรจุอยู่บนเมนบอร์ดหรือแผงวงจรหลัก หน่วยความจำบางประเภทก็ถูกออกแบบให้อยู่ในชิปพีซียูเลย หน่วยของข้อมูลที่จัดเก็บในหน่วยความจำเรียกว่าไบต์ (byte)ซึ่ง1ไบต์ จะประกอบไปด้วย8บิต นอกจากนี้ยังมีหน่วยเป็นกิโลไบต์ (kilobyteหรือKB)

การแยกว่าคำใดเป็นคำทับศัพท์มาหรือเป็นคำไทยอาจเป็นเรื่องง่ายสำหรับมนุษย์ แต่สำหรับคอมพิวเตอร์แล้ว การสามารถระบุแยกว่าสายอักขระที่พบเป็นคำไทยหรือคำทับศัพท์มาและทับศัพท์จากภาษาใด ถือเป็นงานหนึ่งที่ท้าทายความสามารถของคอมพิวเตอร์ และเป็นพื้นฐานที่จำเป็นในการพัฒนาระบบประมวลผลภาษาไทย (Natural Language Processing) ต่างๆ

ตัวอย่างเช่น ในการพัฒนาโปรแกรมทางด้านการค้นคืนสารสนเทศข้ามภาษา (Cross-Language Information Retrieval) จะมีโปรแกรมส่วนหนึ่งที่ทำหน้าที่ถอดอักษรคำทับศัพท์กลับจากภาษาไทยเป็นภาษาต้นฉบับ (backward transliteration) เพื่อจะได้รู้ว่าคำเดิมในภาษานั้นคืออะไร การจะทำเช่นนี้ได้ จำเป็นต้องมีการระบุภาษาของคำในข้อความ ก่อนที่จะนำไปใช้กฎการถอดอักษร (transliteration) ในแต่ละภาษา

จากตัวอย่างข้างบน จะเห็นได้ว่าไม่มีคำหรือสัญลักษณ์ใดๆ ชี้ให้เห็นว่าคำที่ขีดเส้นใต้เป็นคำทับศัพท์ภาษาอังกฤษ เพราะในภาษาไทย มีการใช้คำไทยแท้และคำทับศัพท์จากภาษาอื่นๆ กลมกลืนกันภายใต้การใช้ตัวเขียนภาษาไทย ถึงแม้ว่าคนทั่วไปอาจระบุได้ว่าคำใดเป็นคำไทย หรือ เป็นคำทับศัพท์จากภาษาอังกฤษ ฝรั่งเศส หรือญี่ปุ่น แต่คอมพิวเตอร์ไม่สามารถแยกความแตกต่างของภาษาเหล่านี้ได้เอง หากเราจะนำข้อความนี้ไปใช้กับโปรแกรมถอดตัวอักษรจากภาษาไทยเป็นภาษาอังกฤษ เราจะต้องมีการระบุว่าคำใดเป็นคำไทย หรือคำทับศัพท์ภาษาอังกฤษก่อน เพื่อให้คอมพิวเตอร์สามารถเลือกถอดอักษรเฉพาะคำที่ไม่ใช่คำทับศัพท์มิฉะนั้นข้อความทั้งหมดจะถูกนำไปถอดตัวอักษร ผู้วิจัยจึงต้องการจะพัฒนาโปรแกรมที่ใช้ในการระบุภาษาของคำ เพื่อนำไปใช้ในการตัดสินใจประมวลผลให้ถูกต้อง กล่าวคือ เมื่อให้คำที่ต้องการมาโปรแกรมจะระบุว่าทับศัพท์มาจากภาษาใด เช่น

แบดมินตัน = คำทับศัพท์ภาษาอังกฤษ ⟹ นำไปถอดตัวอักษรเป็น Badminton
ฮาตาริ = คำทับศัพท์ภาษาญี่ปุ่น ⟹ นำไปถอดตัวอักษรเป็นคำว่า Hatari
หนังสือ = คำไทย ⟹ นำไปแปลด้วยข้อมูลด้วย พจนานุกรม เป็นคำว่า Book

การพัฒนาโปรแกรมที่ใช้ในการระบุภาษาของคำ อาศัยการรวบรวมคลังข้อมูลคำทับศัพท์ภาษาต่างๆ และคำไทย ซึ่งคำไทยในที่นี้รวมไปถึงคำที่มาจากบาลีสันสกฤตด้วย เพราะผู้วิจัยเห็นว่าการระบุที่มาของคำ ไม่ว่าจะเป็นคำไทยแท้หรือเป็นบาลีสันสกฤตนั้น ไม่มีความจำเป็นสำหรับการประมวลผลภาษาไทย เช่น ในการถอดตัวอักษรกลับจากภาษาไทยเป็นภาษาอังกฤษ ทั้งคำไทยและคำบาลีสันสกฤตจะไม่ถูกนำไปถอดอักษร ในการแปลจากภาษาไทยเป็นภาษาอื่นๆ ทั้งคำไทยและคำบาลีสันสกฤตจะถูกนำไปหาคำแปลจากพจนานุกรมสองภาษาเหมือนกัน ในขณะที่ถ้าเป็นคำทับศัพท์ภาษาอื่นจะถูกถอดกลับเป็นภาษาเดิม หลังจากที่ได้คลังข้อมูลคำทับศัพท์แล้ว จึงให้คอมพิวเตอร์ได้เรียนรู้ข้อมูลทางภาษาและสายอักขระเฉพาะ (unique character sequence) ของแต่ละภาษาในเชิงสถิติ จากคลังข้อมูลนี้และในคลังข้อมูลคำทับศัพท์ที่ใช้ในการทดลองนี้ ผู้วิจัยเลือกคำทับศัพท์ภาษาอังกฤษ ญี่ปุ่น และฝรั่งเศส เนื่องจากคำทับศัพท์ภาษาอังกฤษและภาษาญี่ปุ่น มีขนาดของข้อมูลมากและเป็นตัวอย่างเพื่อทดลองเปรียบเทียบระหว่างภาษาแถบตะวันตก และภาษาแถบเอเชีย อีกทั้งยังเลือกทดลองกับคำทับศัพท์ภาษาฝรั่งเศสซึ่งเป็นภาษาที่มีตัวอย่างคำทับศัพท์น้อย เพราะต้องการทดสอบประสิทธิภาพของระบบว่าสามารถระบุที่มาของภาษาเมื่อเรียนรู้จากคลังข้อมูลขนาดเล็กได้หรือไม่

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 หาสายอักขระเฉพาะ (unique character sequence) ในการระบุภาษาของคำโดยใช้คลังข้อมูลคำไทย คำทับศัพท์ภาษาอังกฤษ ญี่ปุ่นและภาษาฝรั่งเศส

1.2.2 พัฒนาระบบการระบุภาษาของคำไทยและคำทับศัพท์ภาษาต่างประเทศโดยใช้แบบจำลองเอ็นแกรม

1.2.3 ประเมินประสิทธิภาพของแบบจำลองเอ็นแกรมขนาดต่างๆ ในการระบุภาษาของคำ

1.3 สมมติฐานการวิจัย

1.3.1 คำไทยและคำทับศัพท์จากภาษาต่าง ๆ จะมีสายอักขระเฉพาะที่ใช้ระบุภาษาของคำนั้นได้

1.3.2 ระบบการระบุภาษาของคำสามารถใช้ระบุภาษาของคำได้ถูกต้องมากกว่า 90%

1.3.3 ขนาดของแบบจำลองเอ็นแกรมมีความสำคัญต่อประสิทธิภาพระบบการระบุภาษา

1.4 ขอบเขตของการวิจัย

1.4.1 คำทับศัพท์ที่ใช้เป็นคำทับศัพท์จากภาษาอังกฤษ ภาษาญี่ปุ่น และภาษาฝรั่งเศสเท่านั้น

1.4.2 คำไทย คำทับศัพท์ภาษาอังกฤษและคำทับศัพท์ภาษาญี่ปุ่น จะรวบรวมจากพจนานุกรมราชบัณฑิตยสถาน และหนังสือต่างๆ จำนวนไม่ต่ำกว่าภาษาละ 10,000 คำ ส่วนคำทับศัพท์ภาษาฝรั่งเศสจะรวบรวมจากหนังสือต่างศัพท์เฉพาะๆ จำนวนไม่ต่ำกว่า 1,000 คำ อย่างไรก็ตาม คำทับศัพท์ที่พบในภาษาธรรมชาติมีทั้งที่ทับศัพท์ตามเกณฑ์ของราชบัณฑิตยสถานและที่ไม่ใช้เกณฑ์ตามราชบัณฑิตยสถาน ในที่นี้ผู้วิจัยรวบรวมข้อมูลคำทับศัพท์โดยไม่สนใจว่าเป็นการทับศัพท์ตามเกณฑ์ของราชบัณฑิตยสถานหรือไม่ เพราะในการทำงานจริงโปรแกรมถอดอักษรกลับ (backward transliteration) ของแต่ละภาษาจะต้องถอดคำทับศัพท์ต่างๆ นั้นกลับเป็นภาษาเดิมให้ได้ โดยไม่จำเป็นต้องรู้ว่าข้อมูลนั้นทับศัพท์ตามเกณฑ์หรือไม่

1.4.3 แบบจำลองเอ็นแกรมที่ใช้จะทดลองใช้ 1-5 แกรม เพื่อทดสอบหาประสิทธิภาพของระบบเอ็นแกรมที่ดีที่สุด และเปรียบเทียบผลการทดลองเมื่อใช้เอ็นแกรมขนาดต่างๆ กัน

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ได้ระบบการระบุภาษาของคำไทยและคำทับศัพท์ภาษาต่างๆ

1.5.2 ระบบนี้สามารถนำไปประยุกต์ใช้ในการค้นคืนสารสนเทศข้ามภาษา (cross-language information retrieval) และการถอดอักษรคำยืมทับศัพท์กลับเป็นภาษาต้นฉบับ (backward transliteration)

1.6 โครงสร้างวิทยานิพนธ์

ผู้วิจัยได้จัดทำทั้งหมด 7 บท มีดังนี้

บทที่ 1 บทนำ โดยนำเสนอภาพโดยรวมของความเป็นมาและความสำคัญของปัญหาการระบุภาษาของคำ จุดประสงค์ของการวิจัย สมมติฐานการวิจัย ขอบเขตของงานวิจัย ประโยชน์ที่คาดว่าจะได้รับ และศัพท์เฉพาะที่ใช้ในงานวิจัย

บทที่ 2 การระบุภาษาของคำและข้อความ โดยนำเสนอภาพรวมของงานการระบุภาษาของคำที่ต่างกับการระบุภาษาของข้อความ แนวทางการใช้แบบจำลองภาษาต่างๆ เพื่อแก้ปัญหา วิธีการและแบบจำลองที่เลือกใช้แก้ปัญหารวมไปถึงเหตุผลที่เลือกใช้

บทที่ 3 คลังข้อมูลคำทับศัพท์ โดยนำเสนอวิธีการสร้างคลังข้อมูล เกณฑ์ในการคัดเลือกข้อมูลและ ปัญหาในการสร้างคลังข้อมูล สำหรับทำข้อมูลการฝึกและข้อมูลทดสอบ เพื่อให้ระบบมีประสิทธิภาพมากที่สุด

บทที่ 4 การระบุภาษาของคำด้วยสายอักขระเฉพาะ โดยนำเสนอภาพรวมของระบบระบุภาษาของคำด้วยสายอักขระเฉพาะ 1-5 ตัว ผลการทดสอบระบบ การอภิปรายผล รวมทั้งนำสายอักขระเฉพาะที่ได้ไปวิเคราะห์ลักษณะเฉพาะของภาษา โดยหลังจากรวบรวมคลังข้อมูลและสร้างชุดอักขระเฉพาะในแต่ละภาษาแล้ว เราจะทราบถึงลักษณะการรวมกันของพยัญชนะเด่นๆ ในแต่ละภาษา ซึ่งจะมีรูปแบบตายตัวไม่ซ้ำซ้อนกัน เป็นลักษณะเฉพาะของแต่ละภาษา ซึ่งสายอักขระเฉพาะเหล่านี้น่าจะสะท้อนให้เห็นถึงลักษณะเฉพาะทางด้านเสียงที่เกิดขึ้นของแต่ละภาษาด้วย ดังนั้นผู้วิจัยจึงสรุปลักษณะของเสียงเด่นๆ ที่เกิดขึ้นจากชุดอักขระเฉพาะเหล่านั้น และการปรากฏใช้ของอักขระเด่นๆ ภายในภาษานั้นๆ

บทที่ 5 การระบุภาษาของคำด้วยแบบจำลองภาษา โดยนำเสนอภาพรวมของระบบระบุภาษาของคำด้วยแบบจำลองภาษาตั้งแต่ 2-5 แกรม ผลการทดสอบระบบ การอภิปรายผลและผลการเปรียบเทียบการใช้ขนาดของเอ็นแกรมต่างๆกัน

บทที่ 6 โปรแกรมระบุภาษาของคำ โดยนำเสนอภาพรวมของระบบระบุภาษาของคำที่ใช้ทั้ง 2 วิธีร่วมกัน ผลการทดสอบระบบโดยรวมและการอภิปรายผล

บทที่ 7 สรุป อภิปรายผลและข้อเสนอแนะ โดยสรุปผลการทดสอบระบบทั้งหมดและสรุปอภิปรายผล รวมทั้งเสนอแนวทางในการแก้ไขระบบ และแนวทางในการพัฒนาระบบต่อไป

1.7 ศัพท์เฉพาะที่ใช้ในงานวิจัย

สายคำ (word sequence) คือ คำที่เขียนเรียงกันมากกว่า 1 คำ เช่น สายคำว่า “....to be positive” จะเห็นได้ว่าสายคำนี้ประกอบจากคำ 3 คำคือ คำว่า “to” “be” และ “positive”

สายอักขระหรือชุดอักขระ (string) คือ ตัวอักษรหรืออักขระที่เขียนประกอบเรียงกันเป็นคำหรือเป็นสายอักขระ ในงานวิจัยนี้หมายถึงตัวอักษรภาษาไทย เช่น คำว่า โมดูล โดยจะนับตัวอักษร 1 ตัวเท่ากับ 1 อักขระ ดังนี้ “โ” “ม” “ด” “ู” และ “ล”

ชุดอักขระเฉพาะ (unique character sequence) คือ ตัวอักษรที่เกิดขึ้นเรียงกันในภาษาใดภาษาหนึ่งเท่านั้น โดยไม่ปรากฏในภาษาอื่น ในงานวิจัยนี้กำหนดให้ใช้สายอักขระเฉพาะจากแบบจำลองตั้งแต่ 1-5 แกรม เท่านั้น

การระบุภาษาของข้อความ (Language Identification) คือ งานสำหรับระบุข้อความว่าเป็นภาษาใด ซึ่งส่วนใหญ่ภาษาที่ต้องการจะระบุมักเป็นภาษาที่ใช้อักษรโรมันเหมือนกัน เช่น ภาษาอังกฤษ ภาษาเยอรมัน ภาษาฝรั่งเศส

แบบจำลองภาษา (language model) คือ แบบจำลองที่ใช้ข้อมูลภาษาเชิงสถิติที่ได้จากข้อมูลการฝึกเป็นตัวกำหนดว่าแต่ละข้อความน่าจะเป็นภาษานั้นมากน้อยเพียงใด โดยเก็บข้อมูลหน่วยภาษา ตั้งแต่ 1-5 แกรม

ตัวจำแนกภาษา (language classifier) คือ โมดูลสำหรับใช้ระบุภาษาของข้อความหรือของคำ

ข้อมูลการฝึก (training data) คือ ข้อมูลภาษาที่ใช้สำหรับสร้างแบบจำลองทางสถิติ

ชุดของคำที่ใช้ทดสอบระบบ (test set) คือ คำสำหรับใช้ในการทดลองในแต่ละโมดูล โดยในขั้นตอนทดสอบของงานวิจัยนี้ จะแบ่งเป็น 2 ส่วน คือคำที่ใช้เป็นข้อมูลการฝึก 80% และคำที่ใช้เป็นข้อมูลทดสอบอีก 20%

บทที่ 2

การระบุภาษาของคำและข้อความ

การระบุภาษาของคำในวิทยานิพนธ์นี้มีความคล้ายคลึงกับการระบุภาษาของข้อความ (Language Identification) ในงานวิจัยอื่นๆ คือ มุ่งไปที่จะระบุคำหรือข้อความนั้นว่าเป็นภาษาใด โดยทำให้คอมพิวเตอร์ได้เรียนรู้ลักษณะสำคัญของภาษาจนสามารถระบุภาษาของคำหรือข้อความที่ให้ได้ ปัญหาที่พบเหมือนกัน คือจำเป็นต้องสร้างระบบที่มีประสิทธิภาพ ทำให้สามารถระบุภาษาได้ครอบคลุมทุกๆ ภาษาในโลก แม้การระบุภาษาของคำทับศัพท์จะไม่มีการเสนอแนวทางแก้ปัญหาไว้อย่างชัดเจน แต่ในเรื่องการระบุภาษาของข้อความได้มีผู้ศึกษาไว้เป็นจำนวนมาก ผู้วิจัยจึงได้ทบทบวนวรรณกรรมงานดังกล่าวไว้ในวิทยานิพนธ์นี้ เพื่อนำไปประยุกต์ใช้ต่อไป

การระบุภาษาของข้อความ หรือ Language Identification ถือว่าเป็นงานวิจัยที่สำคัญอย่างหนึ่ง และมีการพัฒนาจนเป็นระบบใช้กันเรื่อยมา เช่น ระบุว่าข้อความที่พบเป็นข้อความภาษาอังกฤษ ภาษาเยอรมัน ภาษาฝรั่งเศส ฯลฯ ซึ่งภาษาที่จะระบุส่วนมากเป็นภาษาที่มีความกำกวมของสายอักขระหรือตัวอักษร (string) เนื่องจากใช้ตัวอักษรชุดเดียวกัน เช่น ตัวอักษรโรมัน จึงต้องหาลักษณะที่สำคัญของสายอักขระ ที่จะบ่งบอกว่าสายอักขระหรือข้อความนั้นเป็นของภาษาใด โดยในการทดลองส่วนใหญ่นี้ใช้วิธีรับข้อมูลเข้า เป็นข้อความที่เขียนขึ้นด้วยภาษาใดภาษาหนึ่งเท่านั้น แล้วใช้ แบบจำลองภาษา (language model) ที่หาจากสายคำ (word sequence) เป็นระบบจำแนก (classifier) ภาษา ผลที่ได้ ดังตัวอย่างตารางที่ 2.1

ตารางที่ 2.1 ตัวอย่างการจำแนกภาษาของข้อความ

TEXT	LANGUAGE
InterPlas Conference will be the meeting place for delegates to be amidst top-notched…..	English
Il faut faire demi-tour. Traversez le carrefour et lorsque vous voyez la station ….	French

ข้อความ (text) ในตารางใช้สายอักขระหรือตัวอักษรชนิดเดียวกัน คือตัวอักษรโรมัน แต่จริงๆ แล้วเป็นคนละภาษากัน คือภาษาอังกฤษและภาษาฝรั่งเศส

การระบุภาษาของข้อความนี้มีส่วนคล้ายคลึงกับงานที่ผู้วิจัยมุ่งจะทำ คือ เป็นการระบุที่มาของภาษาเหมือนกัน ทั้งสองงานนี้เป็นการจำแนกประเภทของภาษาจากสายอักขระที่ประกอบขึ้นจากตัวอักษรชุดเดียวกัน แต่ต่างกันตรงที่การระบุภาษาของข้อความเป็นการระบุว่าข้อความทั้งหมดนั้นเป็นภาษาอะไร ส่วนการระบุที่มาของภาษาของคำเป็นการระบุว่าคำที่พิจารณาเป็นคำไทยหรือทับศัพท์มาจากภาษาอะไร ดังแสดงในตัวอย่างตารางที่ 2.2

ตารางที่ 2.2 ตัวอย่างการระบุภาษาของคำ

คำ	ที่มา
ชาวนา	คำไทย
ปาปัว	คำทับศัพท์ภาษาฝรั่งเศส
ฟิวส์	คำทับศัพท์ภาษาอังกฤษ

การระบุภาษาของข้อความสามารถทำได้โดยการหาลักษณะที่สำคัญของภาษานั้นก่อน เช่น ลักษณะของคำศัพท์ (vocabulary term), ความยาวโดยเฉลี่ยของคำ (word average length), คุณสมบัติทางโครงสร้างและทางความหมายของคำ (syntactic และ semantic properties) ลักษณะเหล่านี้ของแต่ละภาษาจะไม่เหมือนกัน ทำให้สามารถระบุได้ว่าข้อความนั้นๆ เป็นภาษาอะไร ในทำนองเดียวกันหากเป็นคำไทยหรือคำทับศัพท์ ซึ่งมีที่มาจากภาษาต่างๆ กัน แม้ว่าจะอยู่ในรูปภาษาไทยเหมือนกัน แต่ก็น่าจะมีลักษณะสำคัญของภาษาที่ต่างกันด้วย ดังนั้นวิธีที่ใช้ในการระบุภาษาของข้อความก็น่าจะสามารถนำมาประยุกต์ใช้กับการระบุที่มาของภาษาในที่นี่ด้วย

จากปัญหาของการระบุภาษาของข้อความที่ผ่านมา แม้ว่ามีงานวิจัยที่ใช้วิธีทางภาษาศาสตร์ เช่น การระบุภาษาโดยใช้สายอักขระเฉพาะ (Unique character string identification) เช่น ในงานของ Churcher (1993) และ Churcher at al. (1994) และ การระบุภาษาของข้อความโดยใช้คำที่เกิดขึ้นบ่อยๆ ในแต่ละภาษาเป็นตัวตัดสิน (Frequent word recognition) เช่น ในงานของ Hayes (1993), Johnson (1993) และ Churcher et al. (1994) แต่ผู้วิจัยเห็นว่าวิธีการใช้สายอักขระเฉพาะมีข้อจำกัดคือไม่สามารถระบุภาษาของข้อความได้ทุกๆ ภาษา และเห็นว่าวิธีที่ถูกนำมาใช้มากที่สุดคือ วิธีทางสถิติ เนื่องจากเป็นวิธีที่ง่าย ไม่จำเป็นต้องใช้การวิเคราะห์ทางภาษาศาสตร์อย่างลึกซึ้ง เพียงแค่ใช้ข้อมูลการฝึกเพียงอย่างเดียว ทำให้ใช้ต้นทุนในการวิเคราะห์ต่ำกว่า มีประสิทธิภาพสูง และยังแก้ปัญหาข้อจำกัดของวิธีทางภาษาศาสตร์ได้ คือ ทำให้สามารถระบุภาษาได้ทุกๆภาษา เพียงแค่เตรียมข้อมูลสำหรับใช้ในการฝึกสำหรับภาษานั้นเท่านั้น

ระบบวิธีทางสถิติที่ใช้ในการระบุภาษาของข้อความส่วนใหญ่ เป็นการสร้างแบบจำลองทางภาษา (language model) ที่เก็บข้อมูลทางสถิติสำหรับภาษาต่างๆ จากคลังข้อมูลการฝึก (training corpus) เพื่อสร้างแบบจำลองภาษา (language model) หรือ แบบลักษณ์ (profile) ของแต่ละภาษา แล้วนำแบบจำลองเหล่านี้มาใช้ทดสอบกับข้อความที่ต้องการจะระบุ หลังจากนั้นจึงจะประเมินผลเพื่อดูว่าข้อความที่ทดสอบนั้นใกล้เคียงกับแบบจำลองของภาษาใดมากที่สุด ดังนั้นระบบ (system) จะมี 2 ระบบ คือ ระบบสำหรับสร้างแบบจำลองภาษา เพื่อใช้เป็นตัวจำแนกภาษา (language classifier) และระบบสำหรับหาความน่าจะเป็นของข้อความ ที่จะเป็นภาษาใดภาษาหนึ่ง(เมื่อเทียบกับแบบจำลองภาษาที่มี)

เป้าหมายของการสร้างแบบจำลองภาษาก็เพื่อใช้ในการคาดเดาความเป็นไปได้ของสายคำ ถ้าสายคำเกิดขึ้นบ่อยๆ ในภาษา A ก็จะมีความเป็นไปได้สูงว่าเป็นข้อความในภาษา A นั้น

ลักษณะของแบบจำลองภาษา คือ

1. ยิ่งใช้ข้อมูลในการฝึกในแต่ละภาษามีจำนวนมาก แบบจำลองภาษาก็ยิ่งให้ผลดีขึ้นด้วย
2. ค่าที่ใช้คือ ค่าความถี่ และค่าความน่าจะเป็นของการเกิดคำร่วมกัน
3. เนื่องจากว่าคำในชุดของคำที่ใช้ทดสอบระบบ (test set) อาจไม่มีอยู่ในข้อมูลการฝึก ดังนั้นค่าความน่าจะเป็นของภาษาที่ต้องการจะระบุที่ออกมาจะเท่ากับ 0 จึงต้องมีการปรับแต่งค่า (smoothing) ระบบในการทดลอง

อย่างไรก็ตามแบบจำลองภาษา ที่ใช้ในการระบุภาษาของข้อความนั้นหาได้จากค่าทางสถิติของสายคำ (word sequence) คือดูคำที่รวมกันจนเป็นข้อความนั้น แต่แบบจำลองที่ใช้ในการระบุที่มาภาษาของคำนั้นจะหาจากค่าทางสถิติของสายอักขระ (character sequence) แทน คือ คาดเดาความเป็นไปได้ของสายอักขระหรือตัวอักษรที่เกิดขึ้นว่าตรงกับภาษาใด โดยในงานวิจัยนี้ทดลองด้วยการรับข้อมูลเข้าทีละคำ

เพื่อที่จะให้เข้าใจวิธีทางสถิติสำหรับระบุภาษาของคำ ผู้วิจัยจะทบทวนวรรณกรรมที่เกี่ยวข้องกับการระบุภาษาของข้อความก่อนว่ามีการทำอย่างไรบ้าง และวิธีใดเหมาะสำหรับนำมาประยุกต์ใช้กับงานการระบุภาษาของคำ

2.1 แนวทางในการสร้างแบบจำลองภาษา (Language model) หรือ ตัวจำแนกภาษาของข้อความ (text classifiers)

ในงานการระบุภาษาของข้อความนี้ สามารถทำได้หลายวิธี แต่ในที่นี้จะกล่าวถึง วิธีที่ใช้แบบจำลอง*มาร์คอฟ* (Makov Model) หรือแบบจำลองเอ็นแกรม (N-Gram model) โดยละเอียด เพราะเป็นวิธีการที่ใช้ในการวิจัยนี้

2.1.1 แบบจำลอง*มาร์คอฟ* หรือ แบบจำลองเอ็นแกรม

แบบจำลองเอ็นแกรม คือ แบบจำลองที่ใช้คำนวณค่าความน่าจะเป็นของสายอักขระ (character sequence) ที่เกิดขึ้นร่วมกันเป็นคำ หรือค่าความน่าจะเป็นของสายคำ (word sequence) ที่เกิดขึ้นร่วมกันเป็นประโยค โดยค่าความน่าจะเป็นของสายอักขระหรือคำ ประมาณได้จากคลังข้อมูลที่สร้างไว้

แกรม คือ หน่วยที่ใช้ในการสร้างแบบจำลอง อาจจะเป็นเสียง คำ หรือ อักขระก็ได้ และแกรมมีได้หลายขนาดแล้วแต่จะกำหนด ตั้งแต่ 1 จนถึง n

ในแบบจำลองเอ็นแกรมนี้ใช้ความยาวของสายอักขระและสายคำแตกต่างกัน ได้แก่ 2-แกรม 3-แกรม 4-แกรม ฯลฯ ถ้าจะประมาณค่าความน่าจะเป็นของสายคำหรือสายอักขระจากคลังข้อมูลโดยการใช้วิธีเอ็นแกรม ผลที่ได้มีดังนี้

การประมาณค่าด้วย 2-แกรม (Probability bigram) คือ การประมาณค่าความน่าจะเป็นของสายอักขระ (สายคำ) ที่เกิดขึ้นร่วมกันว่ามีค่าเท่ากับผลคูณของความน่าจะเป็นที่จะพบอักขระ (คำ) ทีละ 2 ตัว (คำ) ติดกันในสายอักขระ (สายคำ) นั้น

การประมาณค่าด้วย 3-แกรม (Probability trigram) คือ การประมาณค่าความน่าจะเป็นของสายอักขระ (สายคำ) ที่เกิดขึ้นร่วมกันว่ามีค่าเท่ากับผลคูณของความน่าจะเป็นที่จะพบอักขระ (คำ) ทีละ 3 ตัว (คำ) ติดกันในสายอักขระ (สายคำ) นั้น

การประมาณค่าด้วย 4-แกรม (Probability quadigram) คือ การประมาณค่าความน่าจะเป็นของสายอักขระ (สายคำ) ที่เกิดขึ้นร่วมกันว่ามีค่าเท่ากับผลคูณของความน่าจะเป็นที่จะพบอักขระ (คำ) ทีละ 4 ตัว (คำ) ติดกันในสายอักขระ (สายคำ) นั้น

หรืออาจประมาณค่าความน่าจะเป็นจากความยาวของเอ็นแกรมมากกว่า 4-แกรม ก็ได้ ขึ้นอยู่กับความจำเป็นในการทดลอง แต่ระบบของเอ็นแกรมก็ยิ่งซับซ้อนมากขึ้นตามลำดับ

การประมาณค่าความน่าจะเป็นของสายอักขระ โดยการใช้เอ็นแกรมดังที่กล่าวมา คือ การใช้สมมติฐานของมาร์คอฟ (Markov assumption) ว่า การปรากฏของตัวอักษรตัวหนึ่งขึ้นกับตัวอักษรก่อนหน้าเพียง n-1 ตัว เราจึงเรียกแบบจำลองลักษณะนี้ว่าเป็นแบบจำลองเอ็นแกรม (n-gram model) หรือแบบจำลอง*มาร์คอฟ* ซึ่งวิธีนี้มักนิยมใช้ในงานระบุภาษาของข้อความกันมาก เห็นได้จากงานของ Cavnar และ Trenkle (1994), Dunning (1994), Combrick และ Botha (1995), Sibun และ Reynar (1996), Piotrowski (2000), Peng et al. (2003) เนื่องจากสามารถใช้เพื่อระบุภาษาได้อย่างมีประสิทธิภาพและเรียบง่ายกว่า โดยสามารถประมาณได้ดังนี้

ไบแกรม $P(c_1c_2c_3\ldots..c_n) = P(c_1)\ P(c_2|c_1)\ P(c_3|c_2)\ \ldots P(c_n|c_{n-1})$

ไตรแกรม $P(c_1c_2c_3\ldots..c_n) = P(c_1)\ P(c_2|c_1)\ P(c_3|c_1c_2)\ldots P(c_n|c_{n-2}c_{n-1})$

ควอดิแกรม $P(c_1c_2c_3\ldots..c_n) = P(c_1)\ P(c_2|c_1)\ P(c_3|c_1c_2)\ P(c_4|c_1c_2c_3)\ \ldots P(c_n|c_{n-3}c_{n-2}c_{n-1})$

หมายเหตุ: P แทน ค่าความน่าจะเป็น, c แทน อักขระหรือตัวอักษร และ $(c_1c_2c_3\ldots..c_n)$ แทน สายอักขระที่ประกอบด้วยอักขระตั้งแต่ 3 ตัวขึ้นไปจนถึง n ตัว

ส่วนความน่าจะเป็นของสายคำที่รวมกันเป็นประโยค $w_1w_2w_3\ldots.w_n$ หากประมาณค่าด้วยเอ็นแกรมต่างๆ ผลที่ได้ดังนี้ โดย w แทน คำ n แทน จำนวนนับต่อไป P แทน ค่าความน่าจะเป็น และ $(w_1w_2w_3\ldots..w_n)$ แทน สายคำที่ประกอบด้วยคำมากกว่า 3 คำขึ้นไป

- ความน่าจะเป็นของประโยคโดยใช้วิธี 2-แกรม คือ

$P(w_1w_2w_3\ldots..w_n) = P(w_1)\ P(w_2|w_1)\ P(w_3|w_2)\ \ldots\ P(w_n|w_{n-1})$

- ความน่าจะเป็นของประโยคโดยใช้วิธี คือ 3-แกรม คือ

$P(w_1w_2w_3\ldots w_n) = P(w_1)\ P(w_2|w_1)\ P(w_3|w_1w_2)\ldots P(w_n|w_{n-2}w_{n-1})$

- ความน่าจะเป็นของประโยคโดยใช้วิธี 4-แกรม คือ

$P(w_1w_2w_3\ \ldots\ w_n) = P(w_1)\ P(w_2|w_1)\ P(w_3|w_1w_2)\ P(w_4|w_1w_2w_3)\ldots\ P(w_n|w_{n-3}w_{n-2}w_{n-1})$

<u>ตัวอย่าง</u> การประมาณค่าความน่าจะเป็นของประโยค “He like to eat”

ผู้วิจัยใช้สมการประมาณค่าความน่าจะเป็นพื้นฐานดังนี้

$$P\left(w_1w_2\,\mathrm{K}\ w_T\right) = \prod_{i=1}^{T} P\left(w_i|w_1\,\mathrm{K}\ w_{i-1}\right)$$

โดยที่ P คือ ค่าความน่าจะเป็น (Probability) ซึ่งประมาณได้จากคลังข้อมูล

T คือ จำนวนของคำ

i คือ ลำดับของคำโดยเริ่มต้นที่ลำดับที่ 1

$P(w_i \mid w_1 \ldots w_{i-1})$ คือ ความน่าจะเป็นของคำ w_i หลังจากเกิดคำ $w_1 w_2 \ldots w_{i-1}$ ก่อนหน้านี้แล้ว

ความน่าจะเป็นของประโยคนี้ $P(w_1 w_2 \ldots . w_T)$ สามารถประมาณได้ โดยถือว่าการปรากฏของคำ w_i นั้นขึ้นอยู่กับจำนวนคำข้างหน้า n-1 ตัวเท่านั้นหรือขึ้นอยู่กับขนาดของเอ็นแกรม ดังนั้นถ้าหากประมาณค่าความน่าจะเป็นของประโยคนี้ โดยใช้ 2-แกรม จะปรับเปลี่ยนสมการดังนี้

$$P\left(w_1 w_2 \mathrm{K}\ w_T\right) = P\left(w_1 \mid < s >\right) P\left(w_2 \mid w_1\right) \mathrm{K}\ P\left(w_i \mid w_{i-1}\right)$$

$P(w_1|<s>)$ หมายถึง ความน่าจะเป็นของคำที่หนึ่งเมื่อเกิดเป็นคำแรกของประโยค ซึ่งในที่นี้คือช่องว่าง

$P(w_2|w_1)$ หมายถึง ความน่าจะเป็นของคำ w_2 หลังจากเกิดคำ w_1

$P(w_i \mid w_{i-1})$ หมายถึง ความน่าจะเป็นของคำ w_i หลังจากเกิดคำ w_{i-1}

ดังนั้น จากสูตรประมาณค่าความน่าจะเป็นด้วย 2-แกรม หากต้องการหาความน่าจะเป็นของประโยค “He like to eat“ โดยใช้เอ็นแกรมในระดับของคำ ผลออกมาคือ $P(\mathrm{He} \mid <s>)$ $P(\mathrm{like} \mid \mathrm{He})$ $P(\mathrm{to} \mid \mathrm{like})$ $P(\mathrm{eat} \mid \mathrm{to})$

ส่วนค่าความน่าจะเป็นจะหาได้จากคลังข้อมูล เช่น

$$P(\mathrm{like} \mid \mathrm{He}) = \frac{c(\mathrm{He\text{-}like})}{c(\mathrm{He})}$$ โดยที่ c คือจำนวนนับ

จากสูตรนี้หมายถึง ค่าความน่าจะเป็นของ like เมื่อเกิดร่วมกับคำว่า He คำนวณได้จากนำจำนวนนับของ He ที่เกิดร่วมกับคำว่า like หารด้วยจำนวนนับของการเกิด He เดี่ยวๆ

วิธีการใช้แบบจำลองเอ็นแกรมนี้เป็นวิธีทางสถิติที่นิยมใช้กันมากที่สุด เพราะเป็นวิธีที่เรียบง่าย มีประสิทธิภาพสูง และเหมาะสำหรับวิเคราะห์ภาษา สามารถใช้ระบุภาษาได้ดีกว่าวิธีอื่นๆ โดยเมื่อดูจากทบทวนวรรณกรรมต่อไปจะแสดงให้เห็นประสิทธิภาพของแบบจำลองเอ็นแกรม วิธีการใช้เอ็นแกรมในงานระบุภาษาและผลการทดลอง

2.1.2 ตัวอย่างงานวิจัยของการระบุภาษาของข้อความที่ใช้แบบจำลองเอ็นแกรม

แบบจำลองเอ็นแกรมถูกนำมาใช้ในงานต่างๆ มากมาย เช่น

Dunning (1994) ใช้ แบบจำลองมาร์คอฟ (Markov model) ในระดับตัวอักษรเพื่อระบุภาษาของข้อความที่เขียนด้วยอักษรโรมัน เหตุผลที่เขาเลือกแบบจำลองนี้เพราะ คิดว่าการระบุ

ภาษาของข้อความเป็นเรื่องง่าย ไม่จำเป็นต้องใช้แบบจำลองอื่นๆ ที่ยุ่งยากกว่านี้ เขากล่าวไว้ว่าสามารถระบุภาษาได้เลยแม้ใช้ข้อมูลทดสอบเพียงแค่ 10 ตัวอักษรเท่านั้นแต่จะให้ได้ผลดีมากที่สุดหากใช้ตัวอักษร 50 ตัวหรือมากกว่านั้นก็ได้ ส่วนจำนวนข้อมูลการฝึกก็ใช้น้อยเพียงแค่ข้อความที่ประกอบด้วย 2-3 พันคำก็เพียงพอ ในการทดลองเขาใช้ข้อมูลการฝึกเป็นเอกสารข้อความที่มีขนาดตั้งแต่ 1-50 กิโลไบท์ และใช้ข้อมูลทดสอบเป็นข้อความจำนวน 50 ข้อความโดยขนาดตั้งแต่ 10-500 ไบท์ ลักษณะของแบบจำลองของเขา คือ การดูความน่าจะเป็นของตัวอักษรหรืออักขระที่เกิดขึ้นภายในข้อความ โดยจะดูต่อเนื่องกันทั้งข้อความ ในผลการทดลองเขาก็เปรียบเทียบให้ดูว่า หากใช้ข้อมูลการฝึก และข้อมูลทดสอบยิ่งมาก ค่าความแม่นยำก็ยิ่งสูงตามด้วย โดยอาจสูงถึง 99.9% แต่ก็พบปัญหาคือ เมื่อทดสอบกับภาษาอังกฤษ ฝรั่งเศส เยอรมันและภาษาสเปน อาจระบุภาษาผิดพลาดได้เนื่องจากแต่ละภาษามีรากฐานคล้ายกัน

Sibun and Reynar (1996) ใช้วิธีแบบจำลองเอ็นแกรม เพื่อระบุภาษาที่เขียนด้วยอักษรโรมันซึ่งมีทั้งหมด 27 ภาษา เช่น ภาษาดัชต์ อังกฤษ อิตาลี ลาติน สเปน เป็นต้น เขาใช้แบบจำลองเอ็นแกรมระดับตัวอักษรตั้งแต่ 1-3 แกรม เพื่อสร้าง probability distribution หรือแบบจำลองภาษาจากข้อมูลการฝึก วิธีการสร้างคลังข้อมูลการฝึกของเขาจะมีลักษณะเหมือนกับงานของ Dunning คือเป็นการดูทีละเอกสารข้อความต่อเนื่องกัน โดยในเอกสารจะต้องมีการตัดสัญลักษณ์และเครื่องหมายต่างๆ (markup, accented character and symbol) ออกก่อนเพื่อรับเฉพาะตัวอักษรเท่านั้น และข้อมูลทดสอบก็เป็นเอกสารข้อความเช่นเดียวกัน ในการทดลองเขาใช้วิธีเอ็นโทรปีสัมพัทธ์ (relative entropy) สำหรับวัดค่าความสัมพันธ์ระหว่างข้อมูลทดสอบ (test set) กับชุดข้อมูลการฝึก (training) หรืออาจกล่าวได้ว่า เป็นการวัดค่าความน่าจะเป็นของชุดข้อมูลทดสอบว่าตรงกับชุดข้อมูลการฝึกของภาษาใดภาษาหนึ่ง ผลของการทดลองระบุภาษาทั้งหมด 27 ภาษาโดยใช้ 1-แกรม และ 2-แกรม ค่าความแม่นยำโดยเฉลี่ยมากกว่า 90%

Piotrowski (2000) ระบุภาษาของข้อความที่เขียนด้วยตัวโรมันเพื่อการค้นคืนสารสนเทศว่าเป็นภาษาเยอรมัน อังกฤษ ฝรั่งเศส หรืออิตาลี เขาดูความน่าจะเป็นในระดับของสายอักขระหรือตัวอักษรที่ประกอบขึ้นเป็นข้อความ โดยวิธีที่เขาใช้คือ 2-แกรม ร่วมกับ มาร์คอฟ เชน (Markov Chain) เพื่อสร้างแบบจำลองภาษากับทุกภาษาที่ต้องการระบุ ในแบบจำลองภาษาจะเป็นที่รวมข้อมูลความน่าจะเป็นของสายอักขระที่เกิดขึ้นร่วมกัน ผลการทดลองเขาได้เปรียบเทียบการใช้ขนาดของข้อมูลให้ดูว่า หากใช้ขนาดของข้อมูลการฝึกมาก ค่าความแม่นยำที่ได้ยิ่งสูงตามด้วย เช่น ขนาดของข้อมูลการฝึกประมาณ 20 ไบท์ (bytes) ให้ผลการทดลองค่าความแม่นยำเฉลี่ย 98.73% แต่ถ้าใช้ขนาดของข้อมูลการฝึกประมาณ 50 ไบท์ ให้ผลการทดลองค่าความแม่นยำสูงถึง 99.69%

Combrinck และ Botha (1995) เสนอวิธีสำหรับระบุภาษาของข้อความ โดยทดสอบกับภาษาทั้งหมด 12 ภาษาทั้งภาษาแถบอัฟริกันและยูโรเปียน ระบบของเขาจะเป็นการสร้างแบบจำลองภาษาจากข้อมูลการฝึกด้วยเอ็นแกรม แต่เขาเรียกวิธีนี้ว่าทรานซิชัน-เวกเตอร์ (transition vector) โดยหลักการคือใช้เพื่อดูสายอักขระ หรือดูการรวมของสายอักขระตั้งแต่ 2-3 ตัวในข้อความแต่ละภาษาแล้วจึงสร้างแบบจำลองของภาษานั้นขึ้นมา ระบบของพวกเขาประกอบไปด้วย 3 ขั้นตอน คือ

1. pre-processor เพื่อตัดส่วนที่ไม่ใช่ตัวอักษรออกไป เช่น สัญลักษณ์
2. training module เพื่อสร้างแบบจำลองในแต่ละชุดข้อมูลการฝึก โดยในแบบจำลองจะอยู่ในรูปของทรานซิชัน-เวกเตอร์ ซึ่งจะให้ภาพของการรวมกันของตัวอักษร 2 และ 3 ตัว เช่นเดียวกับเอ็นแกรม 1-2 แกรม
3. recognition module เป็นส่วนที่ใช้ระบุภาษาของข้อความ จาก ทรานซิชัน-เวกเตอร์ที่เด่นๆ ในแต่ละภาษา ซึ่งในที่นี้คือ ทรานซิชัน-เวกเตอร์ ที่มีความถี่ของการรวมกันของตัวอักษรสูงกว่าภาษาอื่นๆ

Combrinck และ Botha (1995) ยังเสนอว่า ระบบของเขาสร้างขึ้นเพื่อรองรับความยาวของสายอักขระมากกว่า 3- แกรม แต่ทรานซิชัน เวกเตอร์ ที่ให้ผลที่สุด คือ 3-แกรม เขาอ้างถึงผลการทดลองว่าตัวจำแนกภาษาของเขาอาจสามารถระบุภาษาได้ถูกต้อง 100% โดยผลที่ดีที่สุดคือ เมื่อใช้กับข้อมูลทดสอบประมาณ 200-300 ตัวอักษรเท่านั้น แต่ระบบของเขายังมีปัญหาคือขาดสเถียรภาพในการใช้งานจริง เขายังเปรียบเทียบให้เห็นว่าหากใช้ข้อมูลขนาดใหญ่เกินไปทั้งข้อมูลการฝึกและข้อมูลทดสอบ ค่าการวัดผลทดลอง (Relative Performance Measure) ที่ออกมาจะต่ำลงตามลำดับ

Peng et al. (2003) เสนอวิธีสำหรับระบุประเภทของข้อความ (classification or categorization) ทั้งหมด 24 ประเภทในภาษาจีนจาก TREC และภาษาญี่ปุ่นจาก NTCIR ในการทดลองเขาใช้ข้อมูลการฝึกทั้งหมด 310,355 ข้อความ แต่ละประเภทมีประมาณ 1,747 - 83,668 ข้อความ และข้อมูลทดสอบมีทั้งหมด 10,000 ข้อความ ซึ่งข้อความในแต่ละประเภทจะมีประมาณ 56 - 2,656 ข้อความ

ลักษณะของแบบจำลองภาษาของเขา คือ การคาดเดาความน่าจะเป็นของสายอักขระที่เกิดขึ้นร่วมกันในข้อความแต่ละประเภท โดยดูทีละข้อความต่อเนื่องกัน ในแบบจำลองเอ็นแกรมที่ใช้มีตั้งแต่ 1-8 แกรม หลังจากที่ได้แบบจำลองภาษาแต่ละประเภทแล้ว เขาใช้แบบจำลองภาษาสำหรับเป็นตัวจำแนกภาษาและประเภทของข้อความ (text classifier) ในตัวจำแนกภาษาและประเภทของข้อความนี้จะดูลักษณะที่สำคัญของภาษาเช่น คำศัพท์ (vocabulary term),

ความยาวเฉลี่ยของคำ (word average length), ลักษณะทางวากยสัมพันธ์และอรรถศาสตร์ (syntactic and semantic properties) สำหรับใช้เพิ่มการตรวจสอบที่จะระบุประเภท (categories) ของข้อความ เนื่องจากข้อความในแต่ละประเภทจะมีลักษณะทางภาษาที่ต่างกัน หากใช้แบบจำลองไบแกรมจะช่วยให้คอมพิวเตอร์สามารถพบลักษณะทางภาษาที่ต่างกันนี้และใช้ระบุประเภทของข้อความเหล่านั้นได้ ตัวอย่างสมมติ เช่น เมื่อใช้ 2-แกรม เราอาจจะพบสายอักขระ “ic” และ “er“ ซึ่งปรากฏเฉพาะในข้อความประเภทข่าวธุรกิจเป็นส่วนใหญ่ ดังนั้นหากนำสายอักขระเหล่านี้มาตรวจสอบ ก็จะสามารถระบุประเภทได้ถูกต้องมากยิ่งขึ้น ในการทดลองเขาทดสอบกับข้อความที่ต้องการระบุประเภทด้วยเอ็นแกรมเหมือนกับที่ใช้ในการฝึก ผลการทดลองประมาณสูงกว่า 80% และเขายังเปรียบเทียบผลการใช้ขนาดเอ็นแกรมต่างๆ ด้วย โดยขนาดของเอ็นแกรม 3-แกรมให้ผลดีที่สุด นอกจากนี้ Pang et al. ยังวิจารณ์ไว้ว่าการสร้างแบบจำลองภาษาที่ใช้สายอักขระ (character level) นี้จะสามารถใช้กับงานการระบุประเภทของข้อความได้ดีกับภาษาที่ไม่มีการเขียนแยกคำ โดยเฉพาะในภาษาแถบเอเชีย มากกว่าแบบจำลองที่ใช้สายคำ (word level)

ในงานทบทวนวรรณกรรมเหล่านี้ใช้วิธีเอ็นแกรมในระดับตัวอักษร เพื่อระบุภาษาของข้อความ แต่ยังมีบางงานที่เสนอการใช้เอ็นแกรมในระดับของคำ เช่น

งานของ Cavnar และ Trenkle (1994) ซึ่งเป็นการระบุหาประเภท (categories) ของข้อความในแต่ละภาษา เช่น ระบุว่าเอกสารนั้นเป็นเอกสารประเภทใด เช่น ด้าน security ด้าน AI ด้าน compilers ด้าน compression และด้านกราฟฟิค งานของเขาจะคล้ายคลึงกับงานที่ทบทวนมาก่อนหน้านี้คือ เป็นการดูทีละข้อความต่อเนื่องกัน ทั้งในข้อมูลการฝึกและข้อมูลทดสอบ แบบจำลองที่เขาใช้คือ แบบจำลองเอ็นแกรมของสายคำ โดยเขาเลือกใช้ตั้งแต่ 1-5 แกรม เพื่อสร้างแบบลักษณ์หรือแบบจำลองภาษาของข้อความแต่ละประเภทภายในแบบจำลองภาษาจะแสดงค่าความถี่และค่าการเกิดร่วมกันของเอ็นแกรมทั้งหมด หลังจากได้ข้อมูลการฝึกของแต่ละประเภทแล้ว เขาก็จะเปรียบเทียบข้อมูลทดสอบกับข้อมูลการฝึกแต่ละประเภทว่า มีค่าความน่าจะเป็นที่จะตรงกันมากน้อยเพียงใด โดยใช้วิธี out-of-place measure หรือค่าความห่าง (distance measure) แล้วเลือกประเภทที่มีค่าความห่างกับข้อมูลทดสอบน้อยที่สุดเป็นคำตอบ ผลจากการทดลองระบุภาษาได้ถูกต้องมากกว่า 90% โดยผลจะต่างกันขึ้นอยู่กับขนาดของข้อมูลการฝึกในแต่ละประเภทและข้อมูลทดสอบ ยิ่งข้อมูลการฝึกและข้อมูลทดสอบมาก ค่าความแม่นยำก็สูงขึ้นด้วย ในงานของเขาจะเป็นการใช้เอ็นแกรมต่างกับงานที่ผู้วิจัยจะทำคือ เป็นการระบุหาประเภทแต่ผู้วิจัยเสนองานการระบุภาษา

จากบททวนวรรณกรรมแนวทางแก้ปัญหาการระบุภาษาและประเภทของข้อความที่ผ่านมา แบบจำลองที่ผู้วิจัยเลือกใช้คือ แบบจำลองเอ็นแกรม (n-gram model) เพราะ จากทบทวนวรรณกรรมที่ผ่านมาเห็นว่า ผลจากการทดลองที่ได้จากงานการระบุประเภทของข้อความด้วยแบบจำลองนี้สูงมากกว่า 90% โดยใช้เพียงแค่ 2-แกรมเท่านั้น และเมื่อเปรียบเทียบแบบจำลองภาษาเอ็นแกรม (LM) กับวิธีการอื่นๆ เช่น Naïve Bays classifier (NB), adhoc n-gram (OOP)[1] และ support vector machine classifier (SVM) พบว่า แบบจำลองภาษา ให้ผลดีที่สุด เห็นได้จากตารางที่ 2.3 ซึ่ง Peng ได้แสดงผลการทดลองเปรียบเทียบค่าความน่าจะเป็นของการระบุประเภทของข้อความประเภทหนึ่งด้วยการใช้แบบจำลองภาษานาอีฟ เบย์ และวิธีการอื่นๆ โดยอ้างถึงผลการทดลองระบุประเภทของข้อความของ Aizawa (2001) ด้วย แสดงให้เห็นว่าเมื่อระบุประเภทในข้อความภาษาจีนด้วยแบบจำลองภาษา จะให้ค่าความถูกต้องสูงที่สุด

ตารางที่ 2.3 ตัวอย่างการเปรียบเทียบความน่าจะเป็นของการระบุประเภทของข้อความ (Peng, et al. 2003 อ้างถึงใน Aizawa (2001))

LM	NB	OOP	SVM
Chinese Character Level (ระบุข้อความภาษาจีนในระดับตัวอักษร)			
0.868	0.856	0.8087	0.817
Japanese Byte Level (ระบุข้อความภาษาญี่ปุ่นในระดับตัวอักษร)			
0.84	0.66	0.4990	85%

นอกจากนี้ ผู้วิจัยเลือกใช้แบบจำลองเอ็นแกรมเพื่อพัฒนาระบบการระบุภาษาของคำ โดยมีเหตุผลดังนี้

1. เอ็นแกรมเป็นวิธีทางสถิติที่เป็นพื้นฐานและเรียบง่ายมากที่สุด ซึ่งจะช่วยลดต้นทุนในการทำระบบการระบุภาษา

2. สามารถนำวิธีเอ็นแกรมมาใช้ร่วมกับวิธีอื่นๆได้ เนื่องจากจะช่วยเพิ่มประสิทธิภาพของระบบให้ดีมากยิ่งขึ้น

---

[1] เป็นวิธีประมาณค่าความน่าจะเป็นโดยใช้แบบจำลองเอ็นแกรมเหมือนกับ LM แต่ต่างกับ LM ตรงที่วิธี OOP จะเลือกข้อความที่ทดสอบว่าตรงกับประเภทใดจากค่าความห่าง (distance measure) โดยที่เลือกใช้คือ out-out-place (OOP) หรือ distance (d,c) ส่วน วิธี LM เลือกข้อความที่ทดสอบว่าตรงกับประเภทใดจากค่า likelihood หรือ Pr(d|c)

3. วิธีเอ็นแกรมสามารถใช้ระบุภาษาได้ทุกภาษา ในกรณีระบุภาษาที่ใช้สายอักขระที่เป็นตัวอักษรประเภทเดียวกัน แต่ต่างภาษากัน เช่น อักษรโรมันที่ใช้เขียนภาษาอังกฤษ เยอรมัน อิตาเลียน หรืออักษรไทย ที่ใช้เขียนคำทับศัพท์ภาษาอังกฤษ ญี่ปุ่นและฝรั่งเศส เป็นต้น

4. เนื่องจากรูปแบบของข้อมูลนำเข้าเพื่อทดสอบและข้อมูลการฝึกควรจะเรียบง่ายมากที่สุด ไม่มีการประมวลผลเบื้องต้น (pre-process) เช่น การ encoding ข้อมูลหรือการใส่ tag ลงบนข้อมูล ดังนั้นระบบเอ็นแกรมจึงเป็นวิธีที่ดีที่สุดที่จะลดความยุ่งยากในการพัฒนาระบบ

5. ในระบบเอ็นแกรม เพียงใช้คลังข้อมูลจำนวนน้อยก็สามารถระบุภาษาของคำได้ถูกต้อง

จากการทบทวนวรรณกรรม ผู้วิจัยพบว่าวิธีรวมสายอักขระซึ่งจะกล่าวต่อไปนี้ เป็นอีกวิธีหนึ่งที่มีผู้วิจัยเคยทำไว้และใช้ในงานระบุภาษาของข้อความ จึงเป็นที่น่าสนใจว่าจะสามารถใช้ในงานระบุภาษาของคำในภาษาไทยได้หรือไม่ ผู้วิจัยจึงต้องการทบทวนวรรณกรรมแนวทางในการสร้างวิธีรวมสายอักขระในส่วนต่อไป เพื่อศึกษาวิธีการและการนำวิธีรวมสายอักขระมาประยุกต์ใช้เพื่อเพิ่มประสิทธิภาพของงานระบุภาษาของคำ

2.2 แนวทางในการสร้างวิธีรวมสายอักขระเฉพาะ (Unique Character Combinations)

วิธีรวมสายอักขระคือ การดูความเป็นไปได้ของจำนวนสายอักขระ (character sequence) ที่เกิดขึ้นตั้งแต่ 3 ตัวขึ้นไป ซึ่งจะแสดงลักษณะเฉพาะของแต่ละภาษาให้เห็นชัดเจน คือเป็นสายอักขระสั้นๆ ที่เกิดขึ้นในภาษาใดภาษาหนึ่งเท่านั้น โดยไม่เกิดขึ้นซ้ำกับภาษาอื่นๆ เลย สายอักขระเฉพาะเหล่านี้พบได้จากคลังข้อมูลภาษา โดยเราจะได้ข้อมูลการเกิดร่วมกันของตัวอักษร พยัญชนะนำ-ตาม สระหรือตัวสะกดที่เป็นลักษณะเด่นของภาษา เห็นได้จากตารางที่ 2.4

ตารางที่ 2.4 ตัวอย่างการใช้วิธีรวมสายอักขระเฉพาะ (Dunning, 1994)

Language	Unique Character
Dutch	“vnd”
English	“ery”
French	“eux”
Italian	“cchi”
Serbo-croat	“lj”

ในอักขระเฉพาะที่กล่าวไว้ในตาราง เป็นสายอักขระเฉพาะที่ประกอบขึ้นจาก 3 ตัวอักษร ที่เกิดขึ้นในท้ายคำของภาษานั้นๆ โดยจะไม่ปรากฏซ้ำกับภาษาอื่นๆ เลย นอกจากนี้ในแต่ละภาษาสามารถมีสายอักขระเฉพาะมากกว่าหนึ่งชุด และมีสายอักขระมากกว่า 3 ตัวติดกันก็ได้

Dunning ยังแสดงให้เห็นปัญหาของวิธีนี้ว่า ถ้าในข้อความนั้นมีคำภาษาอังกฤษอย่าง Pinocchio แล้วจะใช้สายอักขระเฉพาะ cchi สรุปว่าเป็นภาษาอิตาเลียนก็อาจผิดได้ กรณีแบบนี้เป็นเพราะภาษาเหล่านี้ใช้สายอักขระชุดเดียวกัน และยืมคำข้ามภาษามาโดยไม่ต้องทับศัพท์ด้วยตัวเขียนอีกภาษา เช่นจากตัวอย่าง เมื่อทดลองด้วยข้อมูลการฝึกในภาษาอิตาเลียน อาจทำให้ได้ cchi เป็นสายอักขระเฉพาะ เพราะบังเอิญทดลองฝึกในข้อมูลภาษาอังกฤษแล้วไม่มีคำว่า Pinocchio อยู่ แต่พอเอาไปใช้จริง ในข้อความภาษาอังกฤษมีคำว่า pinocchio อยู่ ระบบจึงผิดพลาดได้ เพราะมีข้อมูลว่า cchi เป็นสายอักขระเฉพาะของภาษาอิตาลีซึ่งก็จริง แต่บังเอิญข้อความที่ทดสอบยืมคำจากภาษาอิตาลีมาใช้และทั้งสองภาษาใช้ตัวอักษรโรมันเหมือนกัน จึงเป็นปัญหาที่เกิดขึ้น แต่ในภาษาไทย เราทับศัพท์ภาษาต่างประเทศ คำยืมจึงเขียนด้วยตัวอักษรไทย จึงไม่มีปัญหาที่เกิดจากการยืมคำและคงรูปเดิมไว้

ลักษณะของงานแบบนี้มีมากมาย เช่น จากงานของ Churcher et al. (1994) พวกเขาใช้วิธีสังเกตจากสายอักขระเฉพาะที่ประกอบขึ้นจากตัวอักษรเพื่อระบุภาษาของข้อความในภาษาเขียนด้วยอักษรโรมันถึง 9 ภาษา ในการทดลองเขาสามารถหาสายอักขระเฉพาะ 2 ตัวอักษรติดกันในแต่ละภาษาได้ แต่พวกเขาวิจารณ์ถึงผลการทดลองของวิธีนี้ว่า ไม่มีประสิทธิภาพเพียงพอ เพราะสายอักขระเฉพาะที่ใช้เพียงแค่ 2-แกรม อาจยาวไม่เพียงพอสำหรับกำหนดให้เป็นสายอักขระเฉพาะ และหากใช้ทดสอบกับคำโดยใช้แค่ 2-แกรม ในคำที่ทดสอบคำๆ หนึ่งจะมีสายอักขระที่ตรงกับสายอักขระเฉพาะในหลายๆ ภาษาได้

จากการศึกษาวิธีการใช้สายอักขระเฉพาะ ผู้วิจัยสามารถสรุปปัญหาของวิธีนี้ได้ดังนี้

1. สายอักขระเฉพาะส่วนใหญ่ประกอบด้วยอักขระแค่ 2 หรือ 3 ตัวกันเท่านั้น สายอักขระเฉพาะยาวๆ มีแนวโน้มที่จะเกิดขึ้นยาก เพราะยิ่งต้องการสายอักขระเฉพาะยาวๆ ก็จำเป็นต้องใช้คลังข้อมูลขนาดใหญ่มากขึ้น

2. สายอักขระเฉพาะมีจำกัดคือ ไม่สามารถใช้ระบุภาษาได้ทุกๆข้อความหรือทุกคำ เพราะบางข้อความหรือบางคำอาจไม่มีคำที่ตรงกับสายอักขระเฉพาะเลย จึงต้องนำวิธีการระบุภาษาอื่นๆ เข้ามาช่วย ในที่นี้ผู้วิจัยแก้ปัญหาโดยใช้วิธีทางสถิติร่วมกับการสังเกตจากสายอักขระเฉพาะ

ดังนั้นจากการทบทวนวรรณกรรมทั้งแบบจำลองเอ็นแกรมและวิธีรวมสายอักขระที่ผ่านมา ผู้วิจัยจึงเลือกใช้วิธีเอ็นแกรมร่วมกับวิธีการสังเกตจากสายอักขระเฉพาะ ดังนั้นผู้วิจัยจึงมุ่งที่จะ

นำเสนอแนวทางในการใช้สายอักขระเฉพาะและการสร้างแบบจำลองภาษา เพื่อใช้ในงานการพัฒนาระบบการระบุภาษาของคำนี้

2.3 แนวทางในการจัดทำระบบการระบุภาษาของคำ

จากการศึกษาแนวทางต่างๆของการระบุภาษาของข้อความ ในการพัฒนาระบบการระบุภาษาของคำในวิทยานิพนธ์นี้ ผู้วิจัยจึงประยุกต์วิธีที่เป็นไปได้และเหมาะสมสำหรับภาษาไทย เพื่อกำหนดแนวทางในการพัฒนาระบบการระบุภาษาของคำไทยและคำทับศัพท์ ดังนี้

1. ใช้วิธีการสังเกตจากสายอักขระเฉพาะ (Unique character sequence) เนื่องจากเห็นว่าคำทับศัพท์หรือคำไทยมักมีรูปแบบการเกิดที่ตายตัว เป็นไปได้ที่จะมีลักษณะเฉพาะและมีตัวอักษรหรือพยัญชนะที่ให้ลักษณะเด่นของภาษาได้ชัดเจน เราจึงสามารถระบุภาษาของคำทับศัพท์จากสายอักขระ (character sequence) ที่เป็นตัวอักษรภาษาไทย ตัวอย่างเช่น ในคำทับศัพท์ภาษาอังกฤษ อาจจะมี “ตอร์” เป็นชุดอักขระเฉพาะ ในคำว่า คอมพิวเตอร์ มิเตอร์ ฟิลเตอร์ “ล์ม” ในคำว่า ฟิล์ม และ ”อล” เป็นชุดอักขระเฉพาะ ในคำว่า ดิจิตอล ฟุตบอล โดยในสายอักขระเหล่านี้จะปรากฏเฉพาะในคำทับศัพท์ภาษาใดภาษาหนึ่งเท่านั้น จุดประสงค์หลักในการทำสายอักขระเฉพาะ คือต้องการลดขั้นตอนของระบบ โดยไม่จำเป็นต้องนำไปเทียบกับแบบจำลองภาษา และเป็นการเพิ่มประสิทธิภาพของระบบให้ถูกต้องมากยิ่งขึ้น

แต่ทว่าในการใช้สายอักขระเฉพาะไม่สามารถแก้ปัญหาระบุคำไทยและคำทับศัพท์ได้หมด เพราะอาจเกิดปัญหาคำบางคำอาจไม่มีสายอักขระตรงกับสายอักขระเฉพาะของภาษาใดๆ เลย นอกจากนี้ คำ 1 คำอาจมีสายอักขระที่ไปตรงกับสายอักขระเฉพาะมากกว่าหนึ่งภาษา ตัวอย่างสมมติ คำว่า คอมพายเลอร์ ตรงกับสายอักขระเฉพาะของคำทับศัพท์ภาษาอังกฤษว่า “คอม” และตรงกับสายอักขระเฉพาะของคำทับศัพท์ภาษาฝรั่งเศส “อร์” ด้วย จึงทำให้ตัดสินจากการเทียบสายอักขระเฉพาะไม่ได้ว่าควรเป็นภาษาใด ในกรณีนี้ระบบที่พัฒนาจะนำคำไปทดสอบต่อไปในแบบจำลองภาษาของคำทับศัพท์ที่เป็นไปได้ของทั้ง 2 ภาษานี้ ในที่นี้จึงนำไปทดสอบกับคำทับศัพท์ภาษาอังกฤษและคำทับศัพท์ภาษาฝรั่งเศส

ในงานวิจัยนี้ หาสายอักขระเฉพาะด้วย 1-5 แกรม คือพิจารณาสายอักขระที่เป็นอักษร 1-5 ตัวติดกัน เพราะวิธีเหล่านี้น่าจะให้ลักษณะเด่นของอักขระ (character) ในแต่ละภาษาที่ชัดเจนมากกว่าเพียงแค่ 2-แกรม ซึ่งเป็นการพิจารณาอักษรเพียงสองตัวติดกัน ส่วนข้อมูลการฝึก (training data) สำหรับสร้างสายอักขระเฉพาะ ใช้ข้อมูลเดียวกันกับสำหรับฝึก (training) ทำแบบจำลองภาษา

2. ใช้แบบจำลองทางภาษา คือ แบบจำลองเอ็นแกรม ในที่นี้จะทดลองโดยใช้ 2-5 แกรม นอกจากนี้วิธีเอ็นแกรมที่เลือกใช้จะต่างกับงานทั่วๆไปคือ เปลี่ยนจากการดูทีละคำ (w) เป็นการดูเอ็นแกรมทีละอักขระหรือตัวอักษร (c) เพราะในจุดประสงค์ของการทดลองผู้วิจัยมุ่งจะระบุภาษาของคำทับศัพท์ในภาษาไทยเป็นคำๆ ดังนั้นในข้อมูลทั้งข้อมูลการฝึกและข้อมูลสำหรับทดสอบจึงใช้ทีละคำ และหากจะใช้วิธีเอ็นแกรมสำหรับหาลักษณะเด่นของคำในแต่ละภาษา จะต้องใช้เอ็นแกรมที่ละเอียดกว่าระดับคำ คือ ระดับสายอักขระเท่านั้น จึงจะได้ผลดีที่สุด โดยกำหนดให้ $w = c_1\, c_2\, c_3\, \ldots .\, c_n$ และ c คือ สายอักขระหรือตัวอักษร ในสายอักขระผู้วิจัยจะนับทั้งพยัญชนะ วรรณยุกต์ และสระ เป็นอักขระตัวหนึ่ง เช่น คำว่า “แอกเซส” จะแยกเป็น แ อ ก เ ซ และ ส แล้วจึงนับความถี่และเก็บค่าความน่าจะเป็นตามเอ็นแกรม โดยเก็บเป็น 3-แกรม ตัวอย่างสมมติดังนี้

ข้อมูลไตรแกรม

ตัวอักษร	**ความถี่**
แอก	180
อกเ	954
กเซ	123
เซส	1180

โดยผู้วิจัยจะเก็บข้อมูลลักษณะแบบนี้จากทุกๆ คำในคลังข้อมูล เป็นค่าทางสถิติ

ในการประมาณค่าความน่าจะเป็นของคำ ผู้วิจัยจะประมาณโดยใช้เอ็นแกรมด้วย เช่นในกรณีที่ใช้ 3-แกรม จากคำว่า “แอกเซส” สามารถประมาณค่าความน่าจะเป็นของคำด้วย 3-แกรม ได้ดังนี้

$$P(\text{แ})\, P(\text{อ}|\text{แ})\, P(\text{ก}|\text{แ อ})\, P(\text{เ}|\text{อ ก})\, P(\text{ซ}|\text{ก เ})\, P(\text{ส}|\text{เซ})$$

โดยที่ $P(c_3|c_1\, c_2)$ คำนวณได้จาก $= \dfrac{C(c_1\text{-}c_2\text{-}c_3)}{C(c_1\text{-}c_2)}$

ดังนั้น จากคำว่า แอกเซส เมื่อคำนวณโดยใช้สมการนี้ จะได้ดังนี้ โดยที่ ANS คือ คำตอบ

P(แ) = $\dfrac{\text{จำนวนสายอักขระ (ก)}}{\text{จำนวนสายอักขระทั้งหมด}}$ = ANS1

P(อ|แ) = จำนวนสายอักขระ (แ อ) = ANS2
จำนวนสายอักขระ (แ)

P(ก|แ อ) = จำนวนสายอักขระ (แ อ ก) = ANS3
จำนวนสายอักขระ (แ อ)

P(เ|อ ก) = จำนวนสายอักขระ (อ ก เ) = ANS4
จำนวนสายอักขระ (อ ก)

P(ซ|ก เ) = จำนวนสายอักขระ (ก เ ซ) = ANS5
จำนวนสายอักขระ (ก เ)

P(ส|เ ซ) = จำนวนสายอักขระ (เ ซ ส) = ANS6
จำนวนสายอักขระ (เ ซ)

หลังจากที่คำนวณได้แล้ว จึงนำแต่ละคำตอบมาคูณกัน ผลที่ได้คือ ค่าความน่าจะเป็นของคำว่า แอกเซส ทั้งคำเมื่อประมาณโดยใช้ 3-แกรม โดยค่าความน่าจะเป็นยิ่งใกล้ ค่า 1 แสดงว่าเป็นคำที่พบบ่อยในคลังข้อมูล

อย่างไรก็ตาม จากการคำนวณนี้ อาจเกิดกรณีที่คำนวณแล้วค่าที่ได้เป็นศูนย์ขึ้นได้ เนื่องจากสายอักขระบางสายไม่พบในคลังข้อมูล จำนวนสายอักขระที่พบจึงเป็นศูนย์ หากนำค่าศูนย์ไปคำนวณในสมการนี้ จะไม่สามารถคำนวณผลออกมาได้ หรือ ผลที่ออกมาจะเท่ากับศูนย์ เช่น จากการคำนวณคำว่า แอกเซส

P(ก|แ อ) P(เ|อ ก) P(ซ|ก เ) P(ส|เซ)

0.052 * 0.032 * 0 * 0.045 * 000007 = 0

จึงต้องมีการปรับค่าไม่ให้เป็นศูนย์ ในที่นี้เลือกใช้วิธีแบบง่าย คือ เพิ่มค่า 0.00000001[2] ในการคำนวณความน่าจะเป็นของสายอักขระ P(เ|อ ก) ดังนี้

P(เ|อ ก) = จำนวนสายอักขระ(อกเ) + 0.00000001
จำนวนสายอักขระ(อก) + 0.00000001

โดยกำหนดเงื่อนไขไว้ว่า ในกรณีที่พบว่า จำนวนสายอักขระ (อกเ) หรือ จำนวนสายอักขระ (อ ก) มีค่าเท่ากับศูนย์หรือไม่พบค่าในคลังข้อมูลการฝึกเท่านั้น

[2] ในงานนี้ไม่ได้ใช้วิธีการปรับค่าที่ใช้กันทั่วไป เช่น Witten-Bell, Good Turing, deleted Interpolation, etc. เพราะคิดว่าวิธีที่ใช้นี้ก็เพียงพอกับงานแล้ว

ก่อนจะนำเสนอบทต่อไป ผู้วิจัยจะทบทวนวรรณกรรมหลักเกณฑ์การยืมคำจากภาษาต่างประเทศด้วยการทับศัพท์ก่อน โดยยึดหลักเกณฑ์ทางราชบัณฑิตยสถาน เพื่อแสดงเหตุผลว่าทำไมลักษณะเฉพาะของแต่ละภาษาสามารถใช้บอกที่มาของภาษาได้ และแสดงเหตุผลที่ผู้วิจัยนำสายอักขระเฉพาะมาวิเคราะห์เพื่อหาลักษณะเฉพาะของแต่ละภาษา

2.4 หลักเกณฑ์การทับศัพท์ภาษาต่างประเทศ

จากนิยามคำว่า “คำทับศัพท์” ในพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2525 ทำให้สรุปได้ว่าการทับศัพท์ คือ การถ่ายเสียงศัพท์ภาษาต่างประเทศด้วยตัวอักษรไทยทีละตัว หรือ การถ่ายคำศัพท์ภาษาต่างประเทศให้เป็นอักษรไทย ตัวอย่างการทับศัพท์ตามเกณฑ์ทางราชบัณฑิตยสถาน เช่น

คำทับศัพท์ภาษาอังกฤษ คำว่า	Calculus เขียนทับศัพท์ว่า แคลคูลัส
	Technique เขียนทับศัพท์ว่า เทคนิค
คำทับศัพท์ภาษาฝรั่งเศส คำว่า	Bastia เขียนทับศัพท์ว่า บาสตียา
	Technique เขียนทับศัพท์ว่า เตกนีก

เมื่อดูจากเกณฑ์การทับศัพท์ตามราชบัณฑิตยสถาน จะเห็นได้ว่าการทับศัพท์ภาษาต่างประเทศไม่ได้ใช้ตัวอักษรไทยทุกตัว เช่น ฆ ฏ ฏ ฐ ศ ษ ฒ เป็นต้น เพราะฉะนั้นจึงมีสายอักขระเฉพาะอย่างน้อยที่สุดก็ในคำไทย และการคงรูปตัวอักษรแทนเสียงควบในคำภาษาต่างประเทศ อาจทำให้มีสายอักขระที่ไม่พบในไทย เช่นจากตัวอย่างในคำทับศัพท์ภาษาอังกฤษ เราอาจพบสายอักขระเฉพาะ “เทค” เพราะในคำไทยไม่มีการใช้อักษร ค เป็นตัวสะกดคำ ส่วนในคำทับศัพท์ภาษาฝรั่งเศสเราอาจพบสายอักขระเฉพาะ “เตก” เพราะในคำไทยรูปสระ เอะ + ตัวสะกด จำเป็นต้องมีไม้ไต่คู้กำกับเพื่อให้ออกเสียงสระ เอ็- ได้ เช่นคำว่า เล็ก เป็นต้น จากเหตุผลนี้จึงอธิบายได้ว่าสายอักขระเฉพาะสามารถแสดงลักษณะเฉพาะของภาษาได้ ดังนั้นหากเรานำสายอักขระเฉพาะมาประยุกต์ใช้ในโปรแกรมระบุภาษาของคำ เราก็สามารถใช้สายอักขระเฉพาะบอกที่มาของภาษาได้

แม้ว่าในการทับศัพท์จริงจะแตกต่างจากเกณฑ์ทางราชบัณฑิตยสถานได้ เช่น คำว่า lock หากทับศัพท์ตามเกณฑ์ทางราชบัณฑิตยสถานจะใช้คำว่า “ล็อก” เพราะเสียงพยัญชนะท้าย /k/ ให้แทนด้วยตัวอักษร ก และทางราชบัณฑิตยสถานกำหนดให้ใช้ ไม้ไต่คู้เพื่อแสดงความ

แตกต่างจากคำไทย (ลอก) ด้วย แต่ในการใช้จริงอาจปรากฏการทับศัพท์แบบอื่นได้ เช่น ล็อค ล้อค เพราะผู้ใช้อาจออกเสียงคำทับศัพท์ตามความนิยม ความถนัด หรือออกเสียงให้เข้ากับระบบเสียงของภาษาต่างประเทศที่ใช้ แต่ว่าหากเราพิจารณาเปรียบเทียบหลักเกณฑ์การทับศัพท์ของราชบัณฑิตยสถาน ที่ใช้สำหรับภาษาต่างๆ ก็แสดงในเบื้องต้นได้ว่า มีความแตกต่างระหว่างคำทับศัพท์ภาษาต่างๆ อยู่ เช่น ตัวอักษรที่ใช้ในสำหรับแต่ละภาษาจะไม่เหมือนกัน โดยในคำไทยมีการใช้ตัวอักษร ค ฆ เพื่อออกเสียง ค ใช้ ส ษ ศ เพื่อออกเสียง ส แต่ในภาษาอังกฤษใช้ตัวอักษร ค และ ส เท่านั้น และเสียงที่ถ่ายออกมาเวลาทับศัพท์แต่ละภาษาก็แตกต่างกัน ได้แก่

คำทับศัพท์ภาษาอังกฤษ คำว่า Europe เขียนทับศัพท์ว่า ยุโรป
คำทับศัพท์ภาษาฝรั่งเศส คำว่า Europe เขียนทับศัพท์ว่า เออรอป

จะเห็นได้ว่าคำๆ เดียวกันแต่ออกเสียงต่างกัน ในการถ่ายเสียงเวลาทับศัพท์จึงออกมาต่างกัน ซึ่งสิ่งเหล่านี้ทำให้สะท้อนความแตกต่างออกมาในรูปของสายอักขระที่ปรากฏได้ เช่น จากตัวอย่างคำทับศัพท์ภาษาอังกฤษ เราอาจได้สายอักขระเฉพาะ “โรป” และคำทับศัพท์ภาษาฝรั่งเศส เราอาจได้สายอักขระเฉพาะ “รอป” ดังนั้น เราจึงสามารถนำสายอักขระเฉพาะเหล่านี้มาใช้เพื่อตัดสินภาษาของคำทับศัพท์ได

บทที่ 3

คลังข้อมูลคำทับศัพท์

ในบทนี้ผู้วิจัยจะนำเสนอวิธีการสร้างคลังข้อมูลและเกณฑ์ในการสร้างคลังข้อมูลคำทับศัพท์ เพื่อให้เข้าใจภาพรวมของคลังข้อมูลที่ใช้ในการทดลอง และเหตุผลที่เลือกใช้คลังข้อมูลคำทับศัพท์ภาษาอังกฤษ ญี่ปุ่นและฝรั่งเศส

3.1 การสร้างคลังข้อมูล

คลังข้อมูลที่ใช้ในงานวิจัยนี้ประกอบด้วย คลังข้อมูลคำไทย คำทับศัพท์ภาษาอังกฤษ คำทับศัพท์ภาษาญี่ปุ่น และคำทับศัพท์ภาษาฝรั่งเศส สาเหตุที่เลือกศึกษาคำทับศัพท์ภาษาอังกฤษและภาษาญี่ปุ่นเพราะเป็นคำทับศัพท์ที่พบมากในภาษาไทยและเป็นภาษาที่แตกต่างกันมาก จึงเลือกมาเป็นตัวอย่างของภาษาที่ใช้ทดสอบว่าแบบจำลองเอ็นแกรมในที่นี้จะสามารถระบุที่มาภาษาที่แตกต่างกันนี้ได้หรือไม่ นอกจากนี้ที่ผู้วิจัยเลือกทดลองกับภาษาฝรั่งเศสด้วย เพราะว่าคำทับศัพท์ภาษาฝรั่งเศสนี้จะพบน้อยกว่าภาษาอังกฤษและภาษาญี่ปุ่น จึงถือเป็นการทดสอบระบบว่าจะยังคงมีความทนทานของระบบ (robustness) ว่าจะยังคงดีหรือไม่หากใช้คลังข้อมูลการฝึกจำนวนน้อย

คลังข้อมูลการฝึก (training corpus) ทั้งสำหรับทำแบบจำลองภาษา และชุดอักขระเฉพาะนำมาจาก<u>พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2525 (2538)</u> <u>หลักเกณฑ์การทับศัพท์ภาษาอังกฤษ (2538)</u> <u>หลักเกณฑ์การทับศัพท์ภาษาฝรั่งเศส เยอรมัน อิตาลี สเปน รัสเซีย ญี่ปุ่น อาหรับ และมลายู (2535)</u> <u>ศัพท์เทคนิควิศวกรรมไฟฟ้าสื่อสาร (2541)</u> <u>รวมศัพท์ญี่ปุ่น (2544)</u> และหนังสืออื่นๆ เช่น หนังสือการท่องเที่ยวในต่างประเทศต่างๆ และหนังสือที่แสดงศัพท์เฉพาะทาง (technical term) ต่างๆ นอกจากนี้ยังรวบรวมจากบทความในอินเทอร์เน็ตด้วย ซึ่งจะมีทั้งคำไทย และคำทับศัพท์ภาษาอังกฤษ ญี่ปุ่นและฝรั่งเศส เหตุผลที่เลือกเก็บข้อมูลคำทับศัพท์จากแหล่งข้อมูลหลายแหล่ง เพราะต้องการความหลากหลายของคำทับศัพท์ โดยคำที่ปรากฏในแหล่งข้อมูล ผู้วิจัยเป็นผู้ตัดสินเองว่าเป็นคำไทยหรือคำทับศัพท์ภาษาใด โดยใช้ความรู้เรื่องภาษาและคำทับศัพท์ของผู้วิจัยเองและใช้การสังเกตจากคำบางคำที่ปรากฏทั้งรูปเดิมและรูปทับศัพท์ด้วย หากมีข้อสงสัยก็จะตรวจสอบจากพจนานุกรมภาษานั้นๆ ว่ามีคำนั้นใช้ในความหมายที่พบหรือไม่ จากคลังข้อมูลที่รวบรวมมาเหล่านี้ เราสามารถนำมาใช้ในการฝึก โดย

การนำแต่ละคำแยกเป็นคำไทย และคำทับศัพท์ของแต่ละภาษาก่อน ขนาดของคลังข้อมูลมีทั้ง 4 ประเภท เป็นดังนี้

- คำไทย รวบรวมทั้งหมด 10,000 คำ คำไทยในที่นี้รวมไปถึงคำบาลีและสันสกฤตด้วย เพราะระบบประมวลผลภาษาโดยทั่วไปไม่จำเป็นต้องแยกความต่างระหว่างคำไทยแท้ กับคำไทยที่ยืมมาจากภาษาบาลีหรือสันสกฤต
- คำทับศัพท์ภาษาอังกฤษ รวบรวมทั้งหมด 10,000 คำ คำทับศัพท์ภาษาอังกฤษเป็นคำทับศัพท์ที่แพร่หลายมากที่สุดทั้งในหนังสือวิชาการ บทความ นิตยสาร หนังสือพิมพ์และศัพท์วิชาการส่วนใหญ่ใช้คำทับศัพท์ภาษาอังกฤษ เช่น ศัพท์ทางวิทยาศาสตร์ คอมพิวเตอร์ วิศวกรรมศาสตร์ เป็นต้น
- คำทับศัพท์ภาษาญี่ปุ่น รวบรวมทั้งหมด 10,000 คำ คำทับศัพท์ภาษาญี่ปุ่นส่วนใหญ่รวบรวมจากหนังสือเกี่ยวกับวัฒนธรรม ประเพณี การแต่งกาย อาหารและประวัติศาสตร์ในประเทศญี่ปุ่น
- คำทับศัพท์ภาษาฝรั่งเศส จะรวบรวมเพียงแค่ 1,000 คำ คำทับศัพท์ภาษาฝรั่งเศสจะรวบรวมจากหนังสือเฉพาะทางที่เกี่ยวกับประเทศฝรั่งเศส วัฒนธรรม สถาปัตยกรรมเป็นส่วนใหญ่

3.2 เกณฑ์ในการคัดเลือกข้อมูล

เนื่องจากในปัจจุบันมีการใช้คำทับศัพท์กันอย่างแพร่หลาย และการเขียนทับศัพท์เป็นภาษาไทยของแต่ละภาษาไม่มีการกำหนดกฎเกณฑ์ที่แน่นอน จึงทำให้เกิดคำทับศัพท์หลายรูปแบบ เช่น ตัวอย่างจากข้อความที่พบ

“นายเรจินัลด์ จี แม็คการิตี้ ประกอบอาชญากรรมเครดิตการ์ดระบบออนไลน์ ณ เมืองเลควิลล์ ขณะที่ลูกค้าขอเปิดแอลซีไปยังประเทศฝรั่งเศส”

“อย่าเวิร์กฮาร์ดนักเลย ไปเอ็นจอยไลฟ์ซะบ้าง”

“ผู้รู้ภาษาอย่างดีกับผู้ไม่รู้จะออกเสียงต่างกัน เช่น เซนทรัล เซ็นทรั่น เซ็นทรัล หรือ สเปเชี่ยน สเปเชี่ยว ต่างก็มาจากคำทับศัพท์คำเดียวกัน”

จะพบว่าภายในข้อความเหล่านี้มีการใช้คำทับศัพท์ภาษาอังกฤษมากมายและต่างกัน เช่น คำทับศัพท์ที่เขียนเรียงกันทั้งๆที่เป็นคนละคำ (เวิร์กฮาร์ด เครดิตการ์ด เลควิลล์) ชื่อเฉพาะ (เรจินัลด์ จี แม็คการิตี้) คำย่อ (แอลซี) และคำทับศัพท์ที่เขียนต่างกัน ทั้งๆที่ทับศัพท์จากคำๆเดียวกัน (เซนทรัล เซ็นทรั่น เซ็นทรัล) เป็นต้น หากไม่ได้กำหนดเกณฑ์ใดๆ ในการเก็บรวบรวมข้อมูลคำทับศัพท์เหล่านี้ เพียงแค่พบคำทับศัพท์ก็เก็บแล้ว อาจทำให้ข้อมูลที่รวบรวมไม่สม่ำเสมอ

และเกิดความผิดพลาดของระบบระบุภาษาได้ ในที่นี้ จึงได้กำหนดเกณฑ์ในการเก็บรวบรวมข้อมูลคำทับศัพท์ ดังนี้

1. กรณีที่เป็นคำทับศัพท์ที่มาจากรัสพจน์ (Acronym) และคำย่อ (Abbreviation) เห็นได้จากตัวอักษรพิมพ์ใหญ่ภายในวงเล็บต่อท้ายจากคำทับศัพท์ต่างๆ โดยคำทับศัพท์ทั้ง 2 ประเภทนี้ เมื่อเขียนทับศัพท์เป็นไทยแล้วจะอ่านต่างกัน คือ คำรัสพจน์จะอ่านเป็นคำ ส่วนคำย่อจะอ่านแบบคำย่อ ตัวอย่างเช่น

<u>คำย่อ</u> television มีคำย่อว่า TV อ่านทับศัพท์ ทีวี

Non-performing loan มีคำย่อว่า NPL อ่านทับศัพท์ว่า เอ็นพีแอล

The school of Germological Sciences มีคำย่อว่า SGS มีคำย่อว่า เอสจีเอส

และอีกมากมาย เช่น ซีบีเอส ซีเอดีเอฟ ซีเอฟเออาร์ ซีไอเอ ทีเอ็นที บีทียู บีทีเอส บีทูเอส พีซีโอ พีอาร์ ยูพีเอส ยูวี ยูอาร์แอล ยูเอสทีอาร์ วิดีโอ วีเอชเอส อาร์ซี

<u>รัสพจน์</u> signal plus noise plus distortion to noise plus distortion ratio (SINAD) อ่านทับศัพท์ ไซแนด ไม่ใช่ เอสไอเอ็นเอดี

UNESCO อ่านทับศัพท์ว่า ยูเนสโก

ASEAN อ่านทับศัพท์ว่า อาเซียน

และอีกมากมาย เช่น เอเปก โอเปค เอแบค เฮฟต้า อาฟต้า

ในที่นี้ผู้วิจัยเลือกเก็บข้อมูลคำทับศัพท์ภาษาอังกฤษที่เป็นอ่านเป็นคำ หรือ คำรัสพจน์เท่านั้น เพราะคำย่อเป็นเพียงตัวคำอ่านจากอักษรเท่านั้น ไม่ได้ให้ลักษณะสำคัญของภาษามากเท่ากับสายอักขระที่เกิดร่วมกันเป็นคำ และคำย่อไม่เป็นปัญหาในระบบการระบุภาษาของคำทับศัพท์ เนื่องจากคำย่อมีจำนวนจำกัด เช่น เอ บี ซี ดี จนถึง แซด เท่านั้นจึงไม่นำมารวมอยู่ในปัญหาการระบุภาษาของคำของงานนี้

ส่วนในกรณีที่มีการใช้ผสมระหว่างคำทับศัพท์จากรัสพจน์และคำย่อ เช่น เอเพ็กซ์ (Apex) เอ็มลิ้งค์ (M-link) เอ็มเว็บ (M-web) ไอทีซิตี้ (IT City) อีเมล์ (E-mail) เอ็นเกจ (N-gage) อีคอมเมิร์ซ (E-commerce) เอ็นเกจ (N-gage) ในที่นี้ผู้วิจัยเลือกเก็บทั้งคำโดยไม่สนใจว่าเป็นคำย่อหรือไม่ เพราะเขียนติดกันจนถือว่าเป็นคำๆ เดียวกัน ยกเว้นในกรณีที่เขียนทับศัพท์แล้วมีเครื่องหมายจุดคั่นหรือมีการเว้นช่องไฟ เช่น จี.เอช.ดี. โคลด์ (G.H.D. Cold) ผู้วิจัยจะเลือกเก็บแต่เฉพาะคำทับศัพท์คำว่า โคลด์ เท่านั้น

2. ในกรณีที่เป็นคำทับศัพท์ที่มาจากชื่อเฉพาะ ผู้วิจัยก็จะเลือกเก็บทั้งหมด โดยไม่สนใจว่าเป็นคำทับศัพท์ที่มาจากชื่อเฉพาะประเภทใด เช่น คำว่า พอลล่า ซีเมนต์ โมโตโรล่า พานาโซนิค แต่ชื่อเฉพาะที่ปรากฏในภาษาไทย มีทั้งชื่อเฉพาะที่เขียนติดกัน และชื่อเฉพาะที่มีการเขียนทับศัพท์ผสมระหว่างคำไทยและคำทับศัพท์ ดังนั้นจึงมีเกณฑ์ ดังนี้

2.1 ชื่อเฉพาะที่มีการเขียนทับศัพท์ติดกันไป ไม่มีการแยกคำตามภาษาเดิมหรือมีเครื่องหมายใดๆ มาคั่น กรณีนี้ให้เก็บทั้งคำโดยไม่แยก เพราะถือว่าเป็นคำๆ หนึ่ง[3] เช่น ยูไนเต็ดฟลาวมิลล์ ร้อกเวิธ ทิบโก้ฟู้ดส์ เดทตอล เอปสัน คาร์เนชั่น เป็นต้น

2.2 ชื่อเฉพาะที่มีการเขียนทับศัพท์ผสมระหว่างคำไทยและคำทับศัพท์ กรณีนี้ให้เก็บเฉพาะคำทับศัพท์อย่างเดียวแม้จะเขียนทับศัพท์ติดกันจนเป็นคำๆ เดียวกัน เช่น

“อมรินทร์พริ้นติ้ง แอนด์ พับลิชชิ่ง” เก็บคำว่า พริ้นติ้ง แอนด์ และพับลิชชิ่ง

“ห้องเย็นเอเชี่ยน ซีฟู้ด” เก็บคำว่า เอเชี่ยน และซีฟู้ด

“คอนโดวิภาวดีสวีท” เก็บคำว่า คอนโด และสวีท

จะเห็นได้ว่าคำที่ขีดเส้นใต้ไม่ใช่ทับศัพท์จึงต้องตัดทิ้ง และแยกเก็บเฉพาะคำที่เป็นคำทับศัพท์ภาษาอังกฤษเป็นคำๆ แยกกัน เนื่องจากผู้วิจัยต้องการให้ระบบสามารถระบุภาษาของคำทับศัพท์ เฉพาะในส่วนของคำที่มีที่มาจากคำทับศัพท์ภาษาต่างประเทศจริงๆ เท่านั้น

3. เก็บรวบรวมรูปแบบคำทับศัพท์ทั้งหมดแม้ว่าจะเป็นคำทับศัพท์ภาษาต่างประเทศที่เขียนจากคำศัพท์คำๆ เดียวกันแต่เขียนในรูปต่างกัน เนื่องจากผู้ใช้คำทับศัพท์อาจเขียนคำทับศัพท์ผิดแปลกไปจากเกณฑ์ของราชบัณฑิตยสถาน อาจมีเหตุผลคือไม่ค่อยเป็นที่นิยม ไม่มีความรู้เพียงพอในการเขียนคำทับศัพท์ให้ถูกต้อง หรือผิดพลาดในการตรวจแก้ไขคำทับศัพท์ ดังนั้นคำทับศัพท์ที่ปรากฏตามหนังสือหรือนิตยสารจึงแตกต่างกัน แม้ว่าจะเขียนทับศัพท์จากคำศัพท์ภาษาต่างประเทศตัวเดียวกัน ตัวอย่างเห็นได้จากตารางที่ 3.1 ดังนี้

---

[3] เหตุผลที่ใช้เกณฑ์เช่นนี้ เพราะระบบจะไม่รู้ว่าคำทับศัพท์ที่เห็นเขียนติดกันอยู่นั้นประกอบด้วยคำต่างประเทศกี่คำ โปรแกรมระบุภาษาของคำที่พัฒนานี้ทำได้เพียงตัดสินว่า สายอักขระนั้นมาจากภาษาใด ซึ่งเมื่อรู้คำตอบแล้ว จึงส่งให้โปรแกรมถอดอักษรกลับ (backward transliteration) หาเอาเองว่าเป็นคำต่างประเทศคำใด ซึ่งอาจประกอบด้วยคำมากกว่าหนึ่งคำก็ได้ ดังนั้น ข้อมูลการฝึกจึงเก็บทั้งสายอักขระโดยไม่สนใจว่าประกอบจากคำต่างประเทศมากกว่าหนึ่งคำหรือไม่

ตารางที่ 3.1 การเขียนทับศัพท์ที่ต่างจากเกณฑ์ทางราชบัณฑิตยสถาน

ศัพท์	ทับศัพท์ตามหลักราชบัณฑิตยสถาน	ทับศัพท์แบบอื่น
Lock	ล็อก	ล้อค ล็อค ล๊อค ล้อก
Tent	เต็นท์	เต้นท์
Knot	นอต	น๊อต น็อต
Pyramid	พีระมิด	ปิรามิด ปีรามิด พีรามิด

ในที่นี้ผู้วิจัยเลือกเก็บทั้งหมดโดยไม่สนใจว่าทับศัพท์ตามเกณฑ์ทางราชบัณฑิตยสถานหรือไม่ เนื่องจากผู้วิจัยคาดหวังว่าระบบมีประสิทธิภาพเพียงพอที่จะสามารถระบุภาษาของคำที่รับเข้ามาทดสอบได้ทั้งหมดโดยคำที่รับเข้ามาอาจจะตรงหรือไม่ตรงตามเกณฑ์ของราชบัณฑิตยสถานก็ได้ ดังนั้นข้อมูลการฝึกจึงจำเป็นต้องมีทั้งคำที่ตรงและไม่ตรงตามเกณฑ์ของราชบัณฑิตยสถาน เพื่อให้ตรงกับข้อมูลจริงที่พบใช้กันทั่วไป

4. ในกรณีคำทับศัพท์ที่สร้างจากคำภาษาต่างประเทศที่มีเครื่องหมายยัติภังค์ (hyphen) คั่น จะเก็บรวบรวมทั้งคำ เนื่องจากในคำทับศัพท์ภาษาไทยจะเขียนทับศัพท์ติดต่อกันไปโดยไม่มีเครื่องหมายยัติภังค์แยกคำตามภาษาเดิม เช่น คำทับศัพท์ภาษาอังกฤษ

ครอสสติตช์ (cross-stitch) เก็บคำว่า ครอสสติตช์

แคลเซียมคาร์บอเนต (calcium-carbonate) เก็บคำว่า แคลเซียมคาร์บอเนต

ไวร์คัท (wire-cut) เก็บคำว่า ไวร์คัท

แอนตี้สลิป (Anti-Slip) เก็บคำว่า แอนตี้สลิป

แต่ในกรณีคำทับศัพท์ที่เขียนจากคำศัพท์ที่มีเครื่องหมายยัติภังค์คั่น โดยเขียนทับศัพท์คั่นด้วยเครื่องหมายยัติภังค์เช่นเดียวกับภาษาเดิม ผู้วิจัยจะแยกเก็บเป็นคำๆ ตามภาษาเดิมด้วย เช่น คำว่า McGraw-Hill เขียนทับศัพท์ว่า แมกกรอว์-ฮิลล์ ในกรณีนี้ผู้วิจัยเลือกยกเว้นคำทับศัพท์คำนี้ แต่แยกเก็บเป็นคำๆ แทนคือเก็บคำว่า แมกกรอว์ และฮิลล์

5. เก็บรวบรวมคำทับศัพท์ที่สร้างจากคำภาษาต่างประเทศที่เขียนแยกกัน แม้ว่าภาษาเดิมจะมีการเขียนแยกกันแต่เมื่อทับศัพท์แล้วจะเขียนติดกันไป ไม่ต้องแยกตามภาษาเดิม เช่น คำทับศัพท์ภาษาอังกฤษ

ฟังก์ชั่นไฮเพอร์โบลิก (Hyperbolic function) เก็บคำว่า ฟังก์ชั่นไฮเพอร์โบลิก เป็น 1 คำ

อาร์กติกเซอร์เคิล (Arctic Circle) เก็บคำว่า อาร์กติกเซอร์เคิล เป็น 1 คำ และมีอีกมากมาย เช่น กรีนพีซ (green piece) กรีนไวล์ด (green wild) โกลด์โคสท์ (gold coast) คอมมอนเซนต์ (common sense) เป็นต้น

ลักษณะของคำทับศัพท์แบบนี้จะเกิดขึ้นบ่อยที่สุด จนบางครั้งอาจใช้จนผู้ใช้ไม่รู้ว่าศัพท์คำนั้นประกอบไปด้วยคำทับศัพท์ภาษาอังกฤษกี่คำ บางครั้งอาจคิดว่าเป็นคำทับศัพท์ที่มาจากภาษาอังกฤษคำเดียวกัน ดังนั้นผู้วิจัยจึงเลือกเก็บทั้งคำ โดยไม่ต้องแยกเป็นคำๆ ตามคำศัพท์ภาษาเดิม

อย่างไรก็ตาม มีข้อยกเว้นบางคำที่ภาษาเดิมมีการเขียนแยกกัน เมื่อมีการเขียนทับศัพท์โดยเขียนแยกกันตามภาษาเดิม ในกรณีให้เก็บแยกเป็นคำๆ เช่น คำทับศัพท์ภาษาอังกฤษ

“แฟนซีวู๊ด อินดัสตรีส์” เก็บคำว่า แฟนซีวู๊ด และ อินดัสตรีส์

“ไดอาน่า ดีพาร์ทเมนท์สโตร์” เก็บคำว่า ไดอาน่า และดีพาร์ทเมนท์สโตร์

“ลากูน่า รีสอร์ท” เก็บคำว่า ลากูน่า และรีสอร์ท

นอกจากนี้ในกรณีคำทับศัพท์ที่เขียนติดกันแต่มีคำคุณศัพท์ประกอบกับคำทับศัพท์คำนั้น เช่น คำว่า “เป็นของ” “แบบ” “อย่าง” “ชนิด” “ระบบ” ฯลฯ ในกรณีนี้ผู้วิจัยจะแยกเก็บเป็นทีละคำ เพราะคำเหล่านี้ไม่ใช่มีที่มาจากคำภาษาต่างประเทศ จะต้องเก็บคำที่เป็นคำภาษาต่างประเทศจริงๆ เท่านั้นเพื่อไม่ให้เกิดปัญหาในการใช้งานระบุภาษาของคำทับศัพท์จริงๆ ตัวอย่างคำทับศัพท์ภาษาอังกฤษที่เป็นปัญหาเช่น

Eulerian function จะเขียนทับศัพท์ว่า ฟังก์ชันแบบออยเลอร์

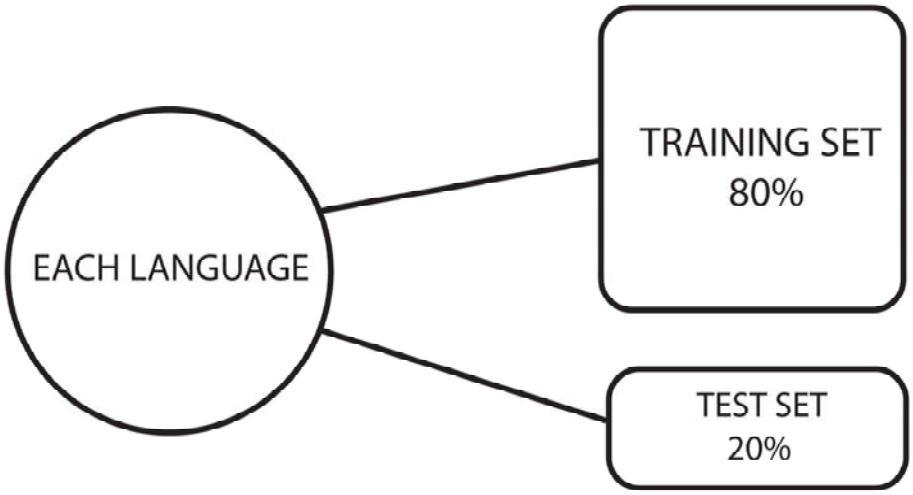
Thermosetting plastic จะเขียนทับศัพท์ว่า พลาสติกชนิดเทอร์โมเซตติง

จากตัวอย่างนี้ ผู้วิจัยจะเก็บทีละคำคือ ฟังก์ชัน ออยเลอร์ พลาสติก และเทอร์โมเซตติง

3.3 การสร้างข้อมูลการฝึกและข้อมูลทดสอบสำหรับใช้ในการทดลอง

เนื่องจากภายในระบบระบุภาษาจากคำจำเป็นต้องใช้ข้อมูลการฝึกสำหรับสร้างแบบจำลองภาษาและใช้ข้อมูลสำหรับการทดลอง ดังนั้นจากการรวบรวมคลังข้อมูลคำไทยและคำทับศัพท์ภาษาต่างประเทศ ซึ่งมีคำไทย 10,000 คำ คำทับศัพท์ภาษาอังกฤษ 10,000 คำ คำทับศัพท์ภาษาญี่ปุ่น 10,000 คำ และคำทับศัพท์ภาษาฝรั่งเศส 1,000 คำ ผู้วิจัยได้แบ่งส่วนข้อมูลเหล่านี้

โดยใช้หลักการ 5-fold cross validation approach ซึ่งแยกเป็นข้อมูลการฝึก (training set) ภาษาละ 80% จากคลังข้อมูลทั้งหมดในแต่ละภาษา ในการทดลองนี้ ใช้คำไทย 8,000 คำ คำทับศัพท์ภาษาอังกฤษ 8,000 คำ และคำทับศัพท์ภาษาญี่ปุ่น 8,000 คำ และคำทับศัพท์ภาษาฝรั่งเศส 800 คำ และข้อมูลทดสอบ (test set) ภาษาละ 20% จากคลังข้อมูลทั้งหมดในแต่ละภาษา ในการทดลองนี้ ใช้คำไทย 2,000 คำ คำทับศัพท์ภาษาอังกฤษ 2,000 คำ และคำทับศัพท์ภาษาญี่ปุ่น 2,000 คำ และคำทับศัพท์ภาษาฝรั่งเศส 200 คำ โดยได้แบ่งด้วยวิธีการสุ่มเลือกจากคลังข้อมูลโดยใช้โปรแกรมภาษาเพิร์ลเป็นเครื่องมือในการแบ่งส่วน เห็นได้จากภาพที่ 3.1



ภาพที่ 3.1 การแบ่งส่วนคลังข้อมูลในแต่ละภาษา

และเพื่อให้ได้ผลที่ไม่บังเอิญขึ้นกับข้อมูลที่ใช้ฝึกและทดสอบ จะทำการทดลอง 5 ครั้งโดยแต่ละครั้งจะใช้ข้อมูลทดสอบ 20% ที่ต่างกัน ดังนั้น ผลที่ได้ในแต่ละภาษาจะมีชุดข้อมูล 5 ชุด แล้วจึงนำผลแต่ละชุดมาสรุปรวมกันเพื่อแสดงผลการทดสอบการระบุคำของระบบ ในคำไทย คำทับศัพท์ภาษาอังกฤษ และญี่ปุ่น แสดงผลคำทดสอบภาษาละ 10,000 คำ ส่วนคำทับศัพท์ภาษาฝรั่งเศส 1,000 คำ

บทที่ 4

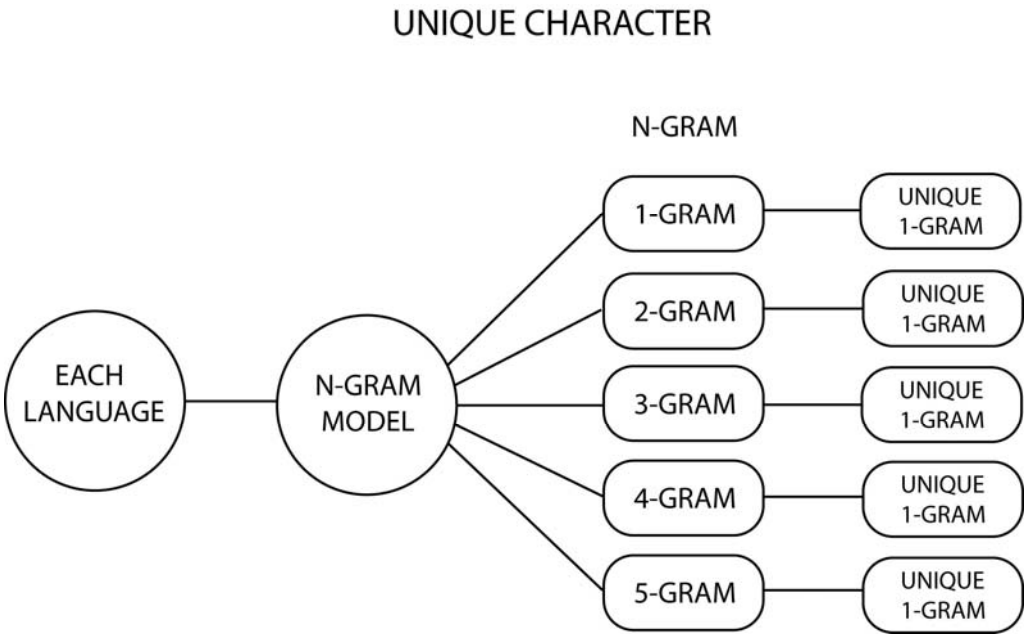
การระบุภาษาของคำด้วยสายอักขระเฉพาะ

ในบทนี้ผู้วิจัยจะกล่าวถึงระบบที่ผู้วิจัยพัฒนา คือ การระบุภาษาของคำด้วยสายอักขระเฉพาะก่อน โดยแสดงถึงภาพรวมของระบบการระบุภาษาโดยใช้สายอักขระเฉพาะ ผลการทดลอง สายอักขระเฉพาะที่พบในคลังข้อมูล สายอักขระเฉพาะที่ใช้จริงในการทดสอบระบบ รวมทั้งนำสายอักขระเฉพาะที่ได้จากคลังข้อมูลไปวิเคราะห์เพื่อหาลักษณะเฉพาะของแต่ละภาษา

4.1 ภาพโดยรวมของวิธีการใช้สายอักขระเฉพาะ (unique characters)

วิธีการใช้สายอักขระเฉพาะในงานวิจัยนี้ เป็นการนำคลังข้อมูลแต่ละภาษาทั้งหมดมาหาสายอักขระเฉพาะของแต่ละภาษาโดยใช้วิธีเอ็นแกรมของสายอักขระตั้งแต่ 1-5 แกรม เช่น ในการสร้างสายอักขระเฉพาะด้วย 3-แกรม เราอาจพบสายอักขระเฉพาะ “ทร์” ในคำไทย และอาจพบสายอักขระเฉพาะ “ตร์” ในคำทับศัพท์ภาษาอังกฤษ เป็นต้น ในการสร้างสายอักขระเฉพาะของแต่ละภาษา จะใช้วิธีนำคลังข้อมูลที่รวบรวมเป็นคำมาแยกเป็นสายอักขระตั้งแต่ 1-5 แกรม แล้วนำมาเปรียบเทียบกันในแต่ละแกรม โดยไม่ขึ้นอยู่กับค่าความถี่ของสายอักขระ แต่ใช้วิธีสังเกตจากการปรากฏของสายอักขระ โดยหากสายอักขระใดที่ปรากฏแล้วไม่ซ้ำกับสายอักขระในภาษาอื่นๆที่เหลือ จะถือว่าเป็นสายอักขระเฉพาะของภาษานั้น

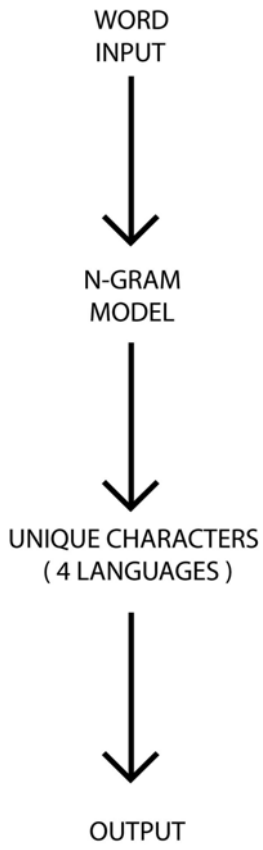
ผลลัพธ์ที่ได้จากการสร้างสายอักขระเฉพาะจะได้ สายอักขระเฉพาะทั้งหมด 4 ภาษาคือ คำไทย คำทับศัพท์ภาษาอังกฤษ ภาษาญี่ปุ่น และภาษาฝรั่งเศส โดยในแต่ละภาษาจะประกอบไปด้วย สายอักขระเฉพาะ 1-5 ตัว เห็นได้จากภาพที่ 4.1



ภาพที่ 4.1 ลักษณะการสร้างสายอักขระเฉพาะ 1-5 แกรม

4.2 โปรแกรมระบุภาษาของคำด้วยสายอักขระเฉพาะ

โปรแกรมระบุภาษาของคำด้วยสายอักขระเฉพาะ แบ่งออกเป็น 2 ขั้นตอน ได้แก่ 1. ขั้นตอนรับคำเข้า 2. ขั้นตอนเทียบกับสายอักขระเฉพาะ เห็นได้จากภาพที่ 4.2



ภาพที่ 4.2 ขั้นตอนการทำงานของโปรแกรมระบุภาษาของคำด้วยสายอักขระเฉพาะ

1. <u>ขั้นตอนรับคำเข้า</u> เป็นขั้นตอนที่รับคำเข้ามาแล้วแยกคำๆ นั้น ออกเป็นสายอักขระด้วยเอ็นแกรมทั้ง 1-5 แกรม เช่น คำที่รับเข้ามาคำว่า “อักขระ” เมื่อผ่านระบบเอ็นแกรม จะแยกเป็นแต่ละเอ็นแกรม เห็นได้จากตารางที่ 4.1 ดังนี้

ตารางที่ 4.1 การแยกสายอักขระตามเอ็นแกรม 1-5 แกรม

1-แกรม	**2-แกรม**	**3-แกรม**	**4-แกรม**	**5-แกรม**
อ	อั	อัก	อักข	อักขร
ั	-ก	-กข	-กขร	-กขระ
ก	กข	กขร	กขระ	
ข	ขร	ขระ		
ร	ระ			
ะ				

2. ขั้นตอนเทียบกับสายอักขระเฉพาะ หลังจากที่แยกสายอักขระเป็นเอ็นแกรมต่างๆ แล้ว ขั้นตอนต่อไปเป็นการนำสายอักขระที่แยกด้วยเอ็นแกรม มาเทียบกับสายอักขระเฉพาะทั้ง 4 ภาษา โดยแยกเทียบกันทั้ง 1-5-แกรม ตามแต่ละเอ็นแกรมว่าตรงกันหรือไม่ หากสายอักขระที่นำมาทดสอบตรงกับสายอักขระเฉพาะของภาษาใดภาษาหนึ่งเท่านั้น ก็ถือว่าคำนั้นเป็นภาษาตามสายอักขระเฉพาะนั้นทันที แต่หากไม่มีสายอักขระที่ตรงกับสายอักขระเฉพาะของภาษาใดๆ เลย หรือ หากพบว่าคำนั้นมีสายอักขระตรงกับสายอักขระเฉพาะของภาษาต่างๆ มากกว่า 1 ภาษา ทำให้ยังตัดสินไม่ได้ว่าเป็นคำจากภาษาใด ตัวอย่างเช่น ในคำว่า “อักขระ” เมื่อทดลองด้วย 4-แกรม อาจจะพบ สายอักขระ “กขระ” ตรงกับสายอักขระเฉพาะภาษาญี่ปุ่น และพบสายอักขระ “อักข“ ตรงกับสายอักขระเฉพาะคำไทย ทำให้ตัดสินไม่ได้ว่าเป็นคำไทยหรือมาจากภาษาญี่ปุ่น ระบบจึงไม่ระบุว่าคำนั้นเป็นภาษาอะไร และข้ามไประบุคำถัดไป ดังนั้น ผลการทดลองจะได้เฉพาะคำที่สามารถระบุภาษาได้ด้วยสายอักขระเฉพาะเท่านั้น ซึ่งในที่นี้ผู้วิจัยเป็นผู้ตรวจสอบผลการระบุภาษาของคำเองว่า ถูกต้องหรือไม่

ในขั้นตอนการเปรียบเทียบกับสายอักขระเฉพาะนี้ ระบบที่ผู้วิจัยพัฒนาจะทดสอบเพียงครั้งเดียว คือ หลังจากที่รับคำที่ทดสอบแล้วแยกเป็นสายอักขระด้วยเอ็นแกรม 1-5-แกรม แล้วระบบจะนำสายอักขระที่ทดสอบ 1-แกรม มาเทียบกับสายอักขระเฉพาะ 1-แกรมในแต่ละภาษาก่อน ถ้ายังไม่พบว่าตรงกัน ระบบก็จะนำสายอักขระที่ทดสอบ 2-แกรม มาเทียบกับสายอักขระเฉพาะ 2-แกรมในแต่ละภาษา หากเปรียบเทียบกันแล้ว ยังไม่ตรงกันอีกก็จะทำงานแบบนี้ต่อไปใน 3 4 และ 5-แกรม ตามลำดับ และหากว่าระบบนำสายอักขระที่ทดสอบเปรียบเทียบกับสายอักขระเฉพาะจนถึง 5-แกรมแล้วยังไม่พบว่าตรงกัน ระบบก็จะข้ามไปทดสอบคำต่อไปเลย

4.3 ผลการทดลอง

จากการทดสอบโปรแกรมระบุภาษาของคำไทยและคำทับศัพท์แสดงให้เห็นผลของระบบระบุภาษาของคำด้วยสายอักขระเฉพาะ 1-5 ตัว ผลที่ได้มี 3 ผลการทดลองคือ ผลการตัดสินภาษาได้ถูกต้อง ผลการตัดสินภาษาผิดพลาด และผลการตัดสินภาษาไม่ได้ โดยเห็นได้จาก ตารางที่ 4.2

ผลการตัดสินภาษาได้ถูกต้อง หมายถึงระบบสามารถระบุภาษาของคำได้ถูกต้อง
ผลการตัดสินภาษาผิดพลาด หมายถึงระบบระบุภาษาของคำผิดพลาด
ผลการตัดสินภาษาไม่ได้ หมายถึงระบบไม่สามารถตัดสินภาษาของคำนั้นได้

ตารางที่ 4.2 ผลการตัดสินภาษาได้ถูกต้อง ผลการตัดสินภาษาผิดพลาด และผลการตัดสินภาษาไม่ได้ ด้วยสายอักขระเฉพาะ 1-5 ตัว โดยผลในตาราง คือ จำนวนคำ ส่วนที่อยู่ในเครื่องหมายวงเล็บคือจำนวนเปอร์เซ็นต์

จำนวนการตัดสินภาษาด้วยสายอักระเฉพาะ 1-5 ตัวอักษร คำ (เปอร์เซ็นต์)				
ภาษา		**ตัดสินภาษาถูกต้อง**	**ตัดสินภาษาผิดพลาด**	**ตัดสินภาษาไม่ได้**
คำไทย	1	1414 (14.14%)	0	
	2	756 (7.56%)	2 (0.02%)	
	3	2104 (21.04%)	274 (2.74%)	
	4	651 (6.51%)	164 (1.64%)	
	5	133 (1.33%)	23 (0.23%)	
รวม		5058 (50.58%)	463 (4.63%)	4479 (44.79%)
คำทับศัพท์ภาษาอังกฤษ	1	0	2 (0.02%)	
	2	122 (1.22%)	14 (0.14%)	
	3	2018 (20.18%)	264 (2.64%)	
	4	1769 (17.69%)	277 (2.77%)	
	5	962 (9.62%)	81 (0.81%)	
รวม		4871 (48.71%)	638 (6.38%)	4491 (44.91%)
คำทับศัพท์ภาษาญี่ปุ่น	1	0	0	
	2	160 (1.6%)	7 (0.07%)	
	3	3392 (33.92%)	186 (1.86%)	
	4	1683 (16.83%)	168 (1.68%)	
	5	174 (1.74%)	38 (0.38%)	
รวม		5409 (54.09%)	399 (3.99%)	4192 (41.92%)
คำทับศัพท์ภาษาฝรั่งเศส	1	0	0	
	2	0	1 (0.10%)	
	3	57 (5.70%)	122 (12.20%)	
	4	101 (10.10%)	99 (9.90%)	
	5	46 (4.60%)	34 (3.40%)	
รวม		204 (20.40%)	256 (25.60%)	540 (54%)

ตารางที่ 4.2 แสดงผลดังนี้ คือ ผลการตัดสินภาษาได้ถูกต้องด้วยสายอักขระเฉพาะ 1-5 ตัว ผลการตัดสินภาษาผิดพลาด และผลการตัดสินภาษาไม่ได้

จากการดูผลการตัดสินภาษาได้ถูกต้องด้วยสายอักขระเฉพาะในแต่ละภาษา เห็นว่าเมื่อระบุภาษาของคำไทย คำทับศัพท์ภาษาอังกฤษ และญี่ปุ่น ด้วยสายอักขระเฉพาะ 3-ตัว ให้ผลการระบุภาษาที่ถูกต้องสูงที่สุด คือ คำไทย 21.04% คำทับศัพท์ภาษาอังกฤษ 20.18% และคำทับศัพท์ภาษาญี่ปุ่น 33.92% แต่ในการระบุภาษาของคำทับศัพท์ภาษาฝรั่งเศส ด้วยสายอักขระเฉพาะ 4-ตัว ให้ผลดีที่สุด คือ 10.1% ส่วนการระบุภาษาด้วยสายอักขระเฉพาะ 1-ตัว สามารถใช้ระบุภาษาได้ถูกต้องได้เฉพาะในคำไทยเท่านั้น

นอกจากนี้ จากตารางที่ 4.2 ยังแสดงให้เห็นผลโดยรวมของระบบการใช้สายอักขระได้ โดยจะต้องดูในช่อง ”รวม” เนื่องจากในการทำงานของระบบการใช้สายอักขระเฉพาะจะเป็นการทำงานทีละตัว โดยเริ่มจากอักขระเฉพาะ 1 ตัว ก่อน พอไม่ผ่านจึงมาที่ 2 3 4 และ 5-ตัว ตามลำดับ ดังนั้นหากคิดผลการทดลองของทั้งระบบการใช้สายอักขระเฉพาะ จึงต้องคิดรวมผลทั้งหมด 1-5 ตัว เช่น จากตารางที่ 4.2 เมื่อต้องการดูผลการระบุของคำไทยจะแสดงผลทีละตัว และนำมารวมกันดังนี้ 14.14 + 7.56+ 21.04 + 6.51 + 1.33 = 50.58% ผลลัพธ์ที่ได้ก็คือ ค่าความถูกต้องจากการตัดสินภาษาของคำไทยด้วยการใช้ระบบสายอักขระเฉพาะตั้งแต่ 1-5 ตัว โดยผลในช่องรวม แสดงให้เห็นว่า วิธีการใช้สายอักขระเฉพาะมีผลต่อโปรแกรมระบุคำประมาณ 50% โดยสายอักขระเฉพาะมีผลต่อการระบุคำทับศัพท์ภาษาญี่ปุ่นมากที่สุด คือ 54.09% รองลงมาคือ คำไทย 50.58% และคำทับศัพท์ภาษาอังกฤษ 48.71% โดยอาจเป็นเพราะภาษาญี่ปุ่นเป็นภาษาที่พบสายอักขระเฉพาะมากที่สุด หรือ เป็นภาษาใช้สายอักขระซ้ำๆ กันมาก ไม่มีการผสมสระมากเท่าภาษาอื่นๆ ซึ่งจะอภิปรายเหตุผลที่แต่ละภาษาได้ผลการทดสอบแตกต่างกันต่อไป ในส่วนวิเคราะห์ผลการทดสอบในหัวข้อ 4.4

จากการดูผลการตัดสินภาษาผิดพลาด พบว่าคำทับศัพท์ภาษาอังกฤษผิดพลาดมากที่สุดถึง 6.38% ซึ่งในส่วนของผลการตัดสินภาษาของคำที่ผิดพลาดนั้นจะกล่าวอย่างละเอียดในหัวข้อ 4.5

และเมื่อดูผลการตัดสินภาษาของคำไม่ได้ พบว่า หลังจากใช้สายอักขระเฉพาะแล้ว ยังเหลือคำที่ระบบยังตัดสินภาษาไม่ได้อีกประมาณ 41-45% คำไทยเหลือ 44.79% คำทับศัพท์ภาษาอังกฤษเหลือ 44.91% คำทับศัพท์ภาษาญี่ปุ่นเหลือ 41.92% และคำทับศัพท์ภาษาฝรั่งเศสเหลือ 54% คำที่ตัดสินภาษาไม่ได้เหล่านี้ เป็นคำที่ไม่มีสายอักขระใดๆ ตรงกับสายอักขระเฉพาะ 1-5 ตัวเลย ดังนั้น คำที่เหลือเหล่านี้จึงต้องนำไปตัดสินภาษาด้วยวิธีอื่นต่อไป โดยในที่นี้นำไปตัดสินภาษาต่อไปด้วยแบบจำลองภาษา 2-5 แกรม (ซึ่งกล่าวถึงในบทที่ 6 การระบุภาษาของคำด้วยสายอักขระเฉพาะ 1-5 ตัวร่วมกับแบบจำลองภาษา 2-5 แกรม)

4.4 วิเคราะห์ผลการทดสอบ

จากการทดสอบด้วยสายอักขระเฉพาะแต่ละภาษา ผู้วิจัยพบว่าทุกภาษาที่ทดสอบ เมื่อระบุด้วยสายอักขระเฉพาะ 3-ตัว จะให้ผลดีที่สุด แต่เราไม่สามารถตัดสินได้ว่าการใช้สายอักขระเฉพาะตัวใดมีประสิทธิภาพดีที่สุด เพราะระบบการทำงานของสายอักขระเฉพาะจะเริ่มจาก 1-ตัวก่อนเสมอและทำงานต่อไป 2 3 4 และ 5 ตามลำดับ

4.4.1 ผลการตัดสินภาษาจากสายอักขระเฉพาะคำไทย

ผลจากการระบุคำไทยด้วยสายอักขระเฉพาะ เมื่อระบุคำไทยด้วยสายอักขระเฉพาะ 3-ตัวให้ผลถูกต้องดีที่สุดคือ 21.04% และผลจะต่ำลงเมื่อใช้สายอักขระ 4-ตัว (6.51%) และ 5-ตัว (1.33%) ตามลำดับ และเมื่อเปรียบเทียบผลระหว่างภาษาพบว่า สายอักขระเฉพาะ 1 ตัว มีผลเฉพาะระบุคำไทยเท่านั้น เพราะ คำไทยมีการใช้พยัญชนะที่ภาษาอื่นไม่ใช้ คือพยัญชนะที่มาจากภาษาบาลี สันสกฤต ได้แก่ ฆ ญ ฌ ฎ ฏ ฒ ณ ธ ภ ศ ษ ฤ ซึ่งในภาษาอื่นใช้พยัญชนะตัวอื่นแทนเสียงเหล่านี้ ทำให้พยัญชนะเหล่านี้ไม่ปรากฏในภาษาอื่น และเมื่อระบุด้วยสายอักขระเฉพาะ 2-ตัว ให้ผลถูกต้องมากกว่าภาษาอื่นๆ คือ 7.56%

4.4.2 ผลการตัดสินภาษาจากสายอักขระเฉพาะคำทับศัพท์ภาษาอังกฤษ

ผลจากการระบุคำไทยด้วยสายอักขระเฉพาะ เมื่อระบุคำทับศัพท์ภาษาอังกฤษด้วยสายอักขระเฉพาะ 3 ตัว ให้ผลถูกต้องดีที่สุดคือ 20.18% และผลจะต่ำลงเมื่อระบุด้วยสายอักขระเฉพาะ 4-ตัว และ 5-ตัว ตามลำดับ และเมื่อเปรียบเทียบผลระหว่างภาษาพบว่า เมื่อระบุภาษาด้วยสายอักขระเฉพาะ 4 และ 5-ตัว จะให้ผลถูกต้องมากกว่าภาษาอื่น คือ 17.69% และ 9.62% และเมื่อระบุด้วยสายอักขระเฉพาะ 2 และ 3-ตัว ให้ผลต่ำกว่าภาษาอื่นมาก คือ 1.22% และ 20.18% อาจเป็นเพราะว่า ภาษาอังกฤษเป็นภาษาที่เขียนขึ้นด้วยอักขระยาวๆ เห็นได้จากมีการใช้สายอักขระเฉพาะยาวๆ ในระบบมาก เช่นสายอักขระเฉพาะ เตอร์ เบอร์ เลอร์ เป็นต้น

4.4.3 ผลการตัดสินภาษาจากสายอักขระเฉพาะคำทับศัพท์ภาษาญี่ปุ่น

ผลจากการระบุคำไทยด้วยสายอักขระเฉพาะ เมื่อระบุคำทับศัพท์ภาษาญี่ปุ่นด้วยสายอักขระเฉพาะ 3-ตัว ให้ผลถูกต้องดีที่สุดคือ 33.92% และผลจะต่ำลงเมื่อระบุด้วยสายอักขระเฉพาะ 4-ตัว และ 5-ตัว ตามลำดับ และเมื่อเปรียบเทียบผลระหว่างภาษาพบว่า เมื่อระบุภาษาด้วยสายอักขระเฉพาะ 3-ตัว ให้ผลถูกต้องมากกว่าภาษาอื่น อาจเป็นเพราะภาษาญี่ปุ่นเป็นภาษาที่เขียนด้วยคำเดี่ยวๆรวมกัน ได้แก่ อะ อิ อุ เอะ โอะ เช่น คำว่า อิคิยะ อิยุ คิโยะ เป็นต้น

4.4.4 ผลการตัดสินภาษาจากสายอักขระเฉพาะคำทับศัพท์ภาษาฝรั่งเศส

ผลจากการระบุคำไทยด้วยสายอักขระเฉพาะ พบว่าสายอักขระเฉพาะ 2-ตัว ไม่มีผลต่อการระบุคำทับศัพท์ภาษาฝรั่งเศส และเมื่อระบุคำทับศัพท์ภาษาฝรั่งเศสด้วยสายอักขระเฉพาะ 4-ตัว ให้ผลถูกต้องดีที่สุดคือ คือ 10.1% โดยผลการทดลองระบุคำทับศัพท์ภาษาฝรั่งเศสอาจไม่ให้ผลที่แน่นอนเนื่องจากใช้คลังข้อมูลน้อย แต่อาจกล่าวได้ว่า หากใช้ขนาดของคลังข้อมูลน้อย จำนวนของสายอักขระเฉพาะที่พบและผลการใช้สายอักขระเฉพาะในการระบุภาษาของคำก็จะน้อยด้วย

4.5 ปัญหาการตัดสินคำผิดพลาดที่เกิดขึ้นจากการใช้สายอักขระเฉพาะ 1-5 ตัว

จากผลการทดสอบระบบ พบว่ามีการระบุภาษาผิดพลาดจากการใช้สายอักขระเฉพาะ 1-5 ตัว โดยที่คำภาษาหนึ่งถูกตัดสินผิดพลาดให้เป็นอีกภาษาหนึ่ง เช่น คำไทย “ตะนอย” ถูกตัดสินผิดพลาดเป็นคำทับศัพท์ภาษาอังกฤษ ข้อผิดพลาดในส่วนนี้แม้จะคิดเป็นเปอร์เซ็นต์ไม่มาก แต่ก็น่าสนใจศึกษาว่า โปรแกรมระบุภาษาคำโดยวิธีการใช้สายอักขระเฉพาะนี้ได้ตัดสินผิดพลาดจากภาษาใดเป็นภาษาใดบ้าง ซึ่งแสดงไว้ในตาราง 4.3 ดังนี้ โดยแสดงผลที่ตัดสินผิดพลาด เป็น จำนวนคำ และ เปอร์เซ็นต์จากคำทดสอบทั้งหมด

ตารางที่ 4.3 ผลการตัดสินภาษาของคำผิดพลาดจากการใช้สายอักขระเฉพาะ 1-5 ตัว

วิธี	คำที่ทดสอบ	ตัดสินผิดเป็น				
		ไทย	อังกฤษ	ญี่ปุ่น	ฝรั่งเศส	รวม
สายอักขระ 1 ตัวอักษร	ไทย	0	0	0	0	
	อังกฤษ	2 (0.02%)	0	0	0	2 (0.02%)
	ญี่ปุ่น	0	0	0	0	
	ฝรั่งเศส	0	0	0	0	
สายอักขระ 2 ตัวอักษร	ไทย	0	2 (0.02%)	0	0	2 (0.02%)
	อังกฤษ	10 (0.1%)	0	4 (0.04%)	0	14 (0.14%)
	ญี่ปุ่น	4 (0.04%)	3 (0.03%)	0	0	7 (0.07%)
	ฝรั่งเศส	1 (0.10%)	0	0	0	1 (0.10%)
สายอักขระ 3 ตัวอักษร	ไทย	0	154 (1.54%)	106 (1.06%)	14 (0.14%)	274 (2.74)%
	อังกฤษ	139 (1.39%)	0	94 (0.94%)	31 (0.31%)	264 (2.64%)
	ญี่ปุ่น	88 (0.88%)	92 (0.92%)	0	6 (0.06%)	186 (1.86%)
	ฝรั่งเศส	22 (2.20%)	90 (9%)	10 (1%)	0	122 (12.2%)
สายอักขระ 4 ตัวอักษร	ไทย	0	87 (0.87%)	63 (0.63%)	14 (0.14%)	164 (1.64%)
	อังกฤษ	86 (0.86%)	0	116 (1.16%)	75 (0.75%)	277 (2.77%)
	ญี่ปุ่น	61 (0.61%)	102 (1.02%)	0	5 (0.05%)	168 (1.68%)
	ฝรั่งเศส	12 (1.20%)	76 (7.60%)	11 (1.10%)	0	99 (9.90%)
สายอักขระ 5 ตัวอักษร	ไทย	0	12 (0.12%)	10 (0.10%)	1 (0.01%)	23 (0.23%)
	อังกฤษ	11 (0.11%)	0	25 (0.25%)	45 (0.45%)	81 (0.81%)
	ญี่ปุ่น	10 (0.10%)	28 (0.28%)	0	0	38 (0.38%)
	ฝรั่งเศส	0	34 (3.40%)	0	0	34 (3.40%)
รวมสายอักขระ 1-5 ตัวอักษร	ไทย	0	255 (2.55%)	179 (1.79%)	29 (0.29%)	463 (4.63%)
	อังกฤษ	248 (2.48%)	0	239 (2.39%)	151(1.51%)	638 (6.38%)
	ญี่ปุ่น	163 (1.63%)	225 (2.25%)	0	11 (0.11%)	399 (3.99%)
	ฝรั่งเศส	35 (3.50%)	200 (20%)	21 (2.10%)	0	256(25.60%)

จากตารางที่ 4.3 แสดงให้เห็นข้อผิดพลาดในการตัดสินภาษา ซึ่งผู้วิจัยได้วิเคราะห์ข้อผิดพลาดในแต่ละภาษาไว้ดังนี้

4.5.1 ข้อผิดพลาดจากการระบุคำไทย

จากข้อมูลการระบุภาษาของคำไทยที่ผิดพลาดจากการใช้สายอักขระเฉพาะ ส่วนใหญ่โปรแกรมตัดสินภาษาผิดพลาดว่าเป็นคำภาษาอังกฤษและภาษาญี่ปุ่น โดยตัดสินผิดว่าเป็นภาษาอังกฤษมากที่สุด (2.55%) เช่น คำว่า ล้อเลียน ถูกตัดสินว่าเป็นคำภาษาอังกฤษจากสายอักขระเฉพาะ “อเลีย” และ ตะนอย ถูกตัดสินผิดว่าเป็นคำภาษาอังกฤษจากสายอักขระเฉพาะ “นอย” เป็นเพราะบังเอิญในคลังข้อมูลการฝึกภาษาอังกฤษมีคำว่า กรอเลียร์ เอเลียน อิลลินอยส์ และในคลังข้อมูลการฝึกภาษาไทยไม่มีคำที่มีอักขระนี้อยู่

4.5.2 ข้อผิดพลาดจากการระบุคำทับศัพท์ภาษาอังกฤษ

จากข้อมูลการระบุภาษาของคำไทยที่ผิดพลาดด้วยสายอักขระเฉพาะ ส่วนใหญ่โปรแกรมตัดสินภาษาผิดพลาดว่าเป็นคำไทย คำภาษาญี่ปุ่นและภาษาฝรั่งเศส โดยตัดสินผิดว่าเป็นคำไทยมากที่สุด (2.48%) เช่น คำว่า เกียร์ ถูกตัดสินว่าเป็นคำไทยจากสายอักขระเฉพาะ “กียร” และ คำว่า ลอการิทึม ถูกตัดสินว่าเป็นคำไทยจากสายอักขระเฉพาะ “อการ” เป็นเพราะบังเอิญในคลังข้อมูลการฝึกคำไทยมีคำว่า เกียรติคุณ เกียรติประวัติ หอการค้าและสื่อการศึกษา เป็นต้น แต่ในคลังข้อมูลการฝึกภาษาอังกฤษไม่มีสายอักขระนี้

4.5.3 ข้อผิดพลาดจากการระบุคำทับศัพท์ภาษาญี่ปุ่น

จากข้อมูลการระบุภาษาของคำไทยที่ผิดพลาดด้วยสายอักขระเฉพาะ ส่วนใหญ่โปรแกรมตัดสินภาษาผิดพลาดว่าเป็นคำภาษาอังกฤษ และคำไทย โดยตัดสินผิดว่าเป็นคำภาษาอังกฤษมากที่สุด (2.25%) เช่นคำว่า เซนทะ ถูกตัดสินว่าเป็นคำภาษาอังกฤษจากสายอักขระเฉพาะ “ซนท” เป็นเพราะบังเอิญในคลังข้อมูลการฝึกภาษาอังกฤษมีคำว่า เซนทรอยด์ และเซนท์ เป็นต้น แต่ในคลังข้อมูลการฝึกภาษาญี่ปุ่นไม่มีสายอักขระนี้

4.5.4 ข้อผิดพลาดจากการระบุคำทับศัพท์ภาษาฝรั่งเศส

จากข้อมูลการระบุภาษาของคำไทยที่ผิดพลาดด้วยสายอักขระเฉพาะ ส่วนใหญ่โปรแกรมตัดสินภาษาผิดพลาดว่าเป็นคำไทย คำภาษาอังกฤษและภาษาญี่ปุ่น โดยตัดสินผิดว่าเป็นคำภาษาอังกฤษมากที่สุด (20%) เช่นคำว่า อาเนต์ ถูกตัดสินว่าเป็นคำภาษาอังกฤษจากสายอักขระเฉพาะ “เนต” และคำว่า เฟรียา ถูกตัดสินว่าเป็นคำภาษาอังกฤษจากสายอักขระเฉพาะ “เฟร” เป็นเพราะบังเอิญในคลังข้อมูลการฝึกภาษาอังกฤษมีคำว่า ลามิเนต คาร์บอเนต เดลี่เฟรช และ เฟรชชี่ เป็นต้น แต่ในคลังข้อมูลการฝึกภาษาฝรั่งเศสไม่มีสายอักขระนี้

เหตุผลที่โปรแกรมตัดสินภาษาผิดพลาดหรือตัดสินภาษาไม่ได้ เป็นเพราะ คลังข้อมูลการฝึกอาจมีขนาดไม่มากพอที่จะทำให้พบสายอักขระเฉพาะที่ครอบคลุมทั้งภาษา ดังจะเห็นได้ชัดเจนในกรณีของคำทับศัพท์ภาษาฝรั่งเศสที่ตัดสินผิดพลาดถึง 25.6% นอกจากนี้ จากผลข้อผิดพลาดที่แสดงในตารางที่ 4.3 เราอาจกล่าวได้ว่า คำทับศัพท์ภาษาญี่ปุ่นมีลักษณะเฉพาะของสายอักขระที่ชัดเจนมากกว่าคำไทย และคำทับศัพท์ภาษาอังกฤษ เพราะมีข้อผิดพลาดเพียง 3.99% ในขณะที่คำทับศัพท์ภาษาอังกฤษดูจะมีลักษณะเฉพาะของสายอักขระที่น้อยกว่า ดังจะเห็นได้จากข้อผิดที่พบ 6.38% และจากการที่คำทับศัพท์ภาษาญี่ปุ่นและคำไทยก็ถูกตัดสินผิดเป็นคำทับศัพท์ภาษาอังกฤษมากที่สุด

นอกจากข้อผิดพลาดดังกล่าว ก็มีประเด็นที่น่าสนใจว่าสายอักขระเฉพาะทุกตัวที่พบในแต่ละภาษา สามารถนำมาใช้จริงได้หรือไม่ และแต่ละภาษามีการใช้สายอักขระเฉพาะใดมากเป็นพิเศษ ดังนั้นผู้วิจัยจึงได้นำเสนอสายอักขระเฉพาะที่พบว่าใช้จริงในระบบและจำนวนครั้งที่ใช้

4.6 สายอักขระเฉพาะที่ใช้จริงในระบบ

จากสายอักขระเฉพาะ 1-5 ตัวในแต่ละภาษาที่พบจากคลังข้อมูลการฝึก ผู้วิจัยพบว่าสายอักขระเฉพาะเหล่านี้ไม่ได้ถูกนำมาใช้จริงในระบบทุกตัว และจำนวนการใช้สายอักขระเฉพาะแต่ละตัวจะไม่เท่ากัน ดังนั้น ผู้วิจัยจึงเสนอตารางการเปรียบเทียบจำนวนสายอักขระเฉพาะที่พบในคลังข้อมูลการฝึกและจำนวนของสายอักขระเฉพาะที่ใช้จริงในแต่ละภาษา เห็นได้จากตารางที่ 4.4 เพื่อแสดงประสิทธิภาพของสายอักขระเฉพาะ เช่น อักขระเฉพาะที่พบจำนวนมาก แต่ถูกนำมาใช้น้อย หรือในทางกลับกัน อักขระเฉพาะที่พบจำนวนน้อย แต่สามารถใช้ระบุภาษาได้มาก

โดยในคอลัมน์ที่สองจะแสดงจำนวนสายอักขระเฉพาะที่พบในคลังข้อมูลการฝึก ตัวเลขในวงเล็บคือ จำนวนเปอร์เซ็นต์จากสายอักขระที่พบในคลังข้อมูลการฝึกทั้งหมด 1-5 ตัว และในคอลัมน์ที่สามแสดงจำนวนสายอักขระเฉพาะที่ใช้จริงในระบบ ในวงเล็บคือจำนวนเปอร์เซ็นต์คิดจากสายอักขระเฉพาะที่พบในคลังข้อมูลการฝึก

ตาราง 4.4 การเปรียบเทียบจำนวนสายอักขระเฉพาะที่พบในคลังข้อมูลการฝึกและจำนวนของสายอักขระเฉพาะที่ใช้จริงในแต่ละภาษา

สายอักขระเฉพาะ (ตัวอักษร)		จำนวนสายอักขระเฉพาะที่พบในคลังข้อมูลการฝึก	จำนวนสายอักขระเฉพาะที่ใช้จริงในระบบ
คำไทย	1	16 (0.13%)	13 (81.25%)
	2	542 (4.39%)	173 (31.92%)
	3	5736 (46.5%)	872 (15.13%)
	4	5522 (44.76%)	389 (7.05%)
	5	520 (4.22%)	70 (13.46%)
คำทับศัพท์ภาษาอังกฤษ	1	0	0
	2	119 (0.76%)	22 (18.49%)
	3	5247 (33.5%)	781 (14.88%)
	4	8944 (57.10%)	903 (10.10%)
	5	1353 (8.64%)	335 (24.76%)
คำทับศัพท์ภาษาญี่ปุ่น	1	0	0
	2	105 (0.95%)	32 (30.48%)
	3	3537 (31.82%)	781 (22.08%)
	4	6726 (60.51%)	864 (12.85%)
	5	747 (6.72%)	107 (14.32%)
คำทับศัพท์ภาษาฝรั่งเศส	1	0	0
	2	14 (0.49%)	0
	3	679 (23.91%)	36 (5.3%)
	4	1785 (62.85%)	68 (3.81%)
	5	362 (12.75%)	28 (7.73%)

จากตารางที่ 4.4 พบว่าสายอักขระที่ใช้จริงในระบบทุกภาษาส่วนใหญ่ เป็นสายอักขระเฉพาะที่พบมากในคลังข้อมูลการฝึก แต่นำไปใช้ได้จริงน้อย คือไม่ถึง 50% ของสายอักขระเฉพาะในคลังข้อมูลการฝึกทั้งหมด เช่น ในสายอักขระเฉพาะคำทับศัพท์ภาษาอังกฤษ ถูกนำไปใช้จริง

ในระบบสูงสุดแค่ 24.76% ในสายอักขระเฉพาะ 5-ตัว ในสายอักขระเฉพาะคำทับศัพท์ภาษาญี่ปุ่น ถูกนำไปใช้จริงในระบบสูงสุดแค่ 30.48% ในสายอักขระเฉพาะ 2-ตัว ยกเว้นกรณีของสายอักขระเฉพาะคำไทย ซึ่งพบข้อมูลสายอักขระเฉพาะ 1-ตัวในคลังข้อมูลการฝึกเพียงแค่ 16 ตัว แต่นำไปใช้จริงในระบบถึง 13 ตัวหรือ 81.25% ได้แก่ ษ ผ ฐ ฤ ใ ณ ฑ ภ ฏ ฏ ฆ พ ฒ ซึ่งสายอักขระเฉพาะเหล่านี้เป็นอักขระที่มาจากคำยืมภาษาบาลีและสันสกฤต อาจเป็นเพราะภาษาไทยมีการยืมคำจากบาลีและสันสกฤตมาใช้เป็นจำนวนมาก ดังนั้นคำส่วนใหญ่จึงใช้สายอักขระเฉพาะเหล่านี้ ส่วนกรณีของคำทับศัพท์ภาษาฝรั่งเศส สายอักขระเฉพาะถูกนำไปใช้จริงสูงสุดแค่ 7.73% ในสายอักขระเฉพาะ 5-ตัว ซึ่งน้อยที่สุดเมื่อเทียบกับภาษาอื่น เป็นเพราะใช้ขนาดของคลังข้อมูลน้อยที่สุด

นอกจากนี้จากการวิเคราะห์สายอักขระเฉพาะที่ใช้จริงในระบบ ผู้วิจัยพบว่ามีประเด็นที่น่าสนใจ เช่น สายอักขระใดที่ใช้บอกภาษาได้ดีที่สุด ดังนั้นผู้วิจัยจึงแสดงตัวอย่างสายอักขระที่ใช้จริงในระบบที่น่าสนใจไว้บางส่วน ส่วนรายการสายอักขระเฉพาะที่ใช้จริงทั้งหมดจะกล่าวไว้ในภาคผนวก

สายอักขระคำไทยที่ถูกนำมาใช้ในระบบการใช้สายอักขระเฉพาะ โดยที่ผลในตารางคือจำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง ส่วนตัวเลขในวงเล็บคือจำนวนเปอร์เซ็นต์เมื่อคิดจากจำนวนรวมของสายอักขระ 1-5 ตัวที่ใช้ระบุภาษาถูกต้องทั้งหมด คือ 5058 คำ

สายอักขระเฉพาะคำไทย 1-ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 1414 ครั้ง)

สายอักขระ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
ษ	298 (5.89%)
ผ	176 (3.48%)
ฐ	127 (2.51%)
ฤ	125 (2.47%)
ใ	118 (2.33%)
ณ	110 (2.17%)
ฑ	104 (2.06%)
ภ	93 (1.84%)
ฏ	92 (1.82%)
ฎ	57 (1.13%)
ฆ	65 (1.29%)
พ	32 (0.63%)
ฒ	17 (0.34%)

จากสายอักขระเฉพาะ 1-ตัว เห็นได้ว่า สายอักขระเฉพาะ 1-ตัวใช้ตัดสินภาษาได้มากเมื่อเทียบกับสายอักขระเฉพาะขนาดอื่นๆ โดยเฉพาะ ษ ใช้ตัดสินภาษาได้มากที่สุดถึง 5.89%

สายอักขระเฉพาะคำไทย 2 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 756 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
ส้-	35 (0.69%)
ภ-า	32 (0.63%)

ื-ด	23 (0.45%)
ภ้-	22 (0.43%)
า-ณ	21 (0.42%)
อ่-	21 (0.42%)
ศ-า	21 (0.42%)
ภ-ู	20 (0.4%)

ร-ภ	19 (0.38%)
ข-อ	19 (0.38%)
ร-ศ	16 (0.32%)
ศ-ร	15 (0.3%)
น-ื	13 (0.26%)
เ-ภ	13 (0.26%)

ม-ื	12 (0.24%)
ณ-ะ	11 (0.22%)
ุ-ภ	11 (0.22%)
ถ-ุ	10 (0.2%)
ถ-ว	10 (0.2%)
ช-ฌ	10 (0.2%)

จากสายอักขระเฉพาะ 2-ตัว สังเกตเห็นว่า ไม่มีลักษณะใดเด่นชัด แต่พบว่ามีสายอักขระเฉพาะที่ถูกนำไปใช้มากกว่าสายอักขระเฉพาะอื่นๆ คือ ส้- (0.69%) ภ-า (0.63%)

สายอักขระเฉพาะคำไทย 3 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 2104 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
บ-ร-ร	39 (0.77%)
่-า-ง	34 (0.67%)
จ้--า	31 (0.61%)
ุ-ท-ธ	27 (0.53%)
ก-ล้-	20 (0.4%)
า-ร-ย	19 (0.38%)
้-า-ม	18 (0.36%)
้-ต-ร	18 (0.36%)

น-า-ย	17 (0.34%)
อ-ร-ร	15 (0.3%)
่-ย-ง	15 (0.3%)
ห-ล-ว	14 (0.28%)
ร่--อ	14 (0.28%)
้-ญ-ช	14 (0.28%)
ห-ล้-	12 (0.24%)
ว-ร-ร	10 (0.2%)
ล้--ย	10 (0.2%)
พ-จ-น	10 (0.2%)
เ-ห-ล	10 (0.2%)
ุ้--ม	10 (0.2%)

้-ว-ห	10 (0.2%)
้-อ-ย	10 (0.2%)
ส้--ม	9 (0.18%)
ป-ร-ะ	9 (0.18%)
เ-ฟ-ื	9 (0.18%)
ะ-พ-ร	8 (0.16%)
ห้--ว	8 (0.16%)
ค-ร่-	8 (0.16%)
้-า-ย	8 (0.16%)
า-ธ-ร	7 (0.14%)
ะ-ท้-	7 (0.14%)
ล้--	7 (0.14%)

ม-ต้-	7 (0.14%)
บ-ป-ร	7 (0.14%)
น-ห-า	7 (0.14%)
ช้--า	7 (0.14%)
ค-น-ธ	7 (0.14%)
โ-ห-ม	7 (0.14%)
่-า-ว	7 (0.14%)
้-ว-ป	7 (0.14%)
้-ด-ต	7 (0.14%)
า-ย-ต	6 (0.12%)
า-ค-ม	6 (0.12%)
ะ-ก-ร	6 (0.12%)

จากสายอักขระเฉพาะที่ใช้จริง 3-ตัว พบข้อสังเกตคือ มีการใช้สายอักขระเฉพาะที่ประกอบด้วย ตัวอักษรใดก็ตาม + ร หัน หรือ _รร เช่นสายอักขระเฉพาะ บรร อรร วรร ซึ่งลักษณะแบบนี้แสดงถึงการใช้ภาษาไทย

สายอักขระเฉพาะคำไทย 4 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 651 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
ก-ร-ะ-โ	16 (0.32%)
น-ท-ร์-	12 (0.24%)
ก-ร-ะ-ด	12 (0.24%)
ั-ต-น์-	11 (0.22%)
ก-ร-ะ-ส	9 (0.18%)

เ-ต้-า	9 (0.18%)
ั-ม-พ้ั	9 (0.18%)
อ-า-ร้ั	7 (0.14%)
เ-ส-น-า	5 (0.1%)
เ-พ-ล-ิ	5 (0.1%)
ิ-ม-พ์-	5 (0.1%)
ส-ม้ั-ย	4 (0.08%)
ร-า-ล้ั	4 (0.08%)
พ-ิ-น-ท	4 (0.08%)

ต-ะ-ก-ร	4 (0.08%)
ง-ก-ร-า	4 (0.08%)
ค-ร-อ-ง	4 (0.08%)
กั้-น-แ	4 (0.08%)
ก-ร-ะ-ต	4 (0.08%)
ก-ร-ะ-ซ	4 (0.08%)
า-ม้ั-ย	3 (0.06%)
อ-น-ี-ก	3 (0.06%)
อ-ด-อ-อ	3 (0.06%)

ว-า-ม-เ	3 (0.06%)
ว-ร-า-ช	3 (0.06%)
ร-ะ-ต-ิ	3 (0.06%)
ร-ะ-เ-ห	3 (0.06%)
ร-สุ-้-ม	3 (0.06%)
ม-ู-ล-ิ	3 (0.06%)
พ-ิ-ร-ุ	3 (0.06%)
พ-ิ-ต-ร	3 (0.06%)
พั้-น-ธ	3 (0.06%)

จากสายอักขระเฉพาะคำไทย 4-ตัว พบข้อสังเกต คือ มีการใช้ กระ+ตัวอักษรใดก็ตาม หรือ กระ_ ในการตัดสินภาษาได้มาก เช่น กระโ กระด กระส กระต เป็นต้น เป็นเพราะคำไทยมีการใช้ ร ควบกล้ำร่วมกับ สระอะ บ่อย

ส่วนสายอักขระเฉพาะคำไทย 5 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 133 ครั้ง) ซึ่งไม่ค่อยมีอะไรเด่น เนื่องจากแต่ละสายอักขระเฉพาะมีการใช้จริงแค่ 1-7 ครั้งเท่านั้น

สายอักขระคำทับศัพท์ภาษาอังกฤษที่ถูกนำมาใช้ในระบบการใช้สายอักขระเฉพาะ โดยที่ผลในตารางคือ จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง ส่วนตัวเลขในวงเล็บคือจำนวนเปอร์เซ็นต์เมื่อคิดจากจำนวนรวมของสายอักขระ 1-5 ตัวที่ใช้ระบุภาษาถูกต้องทั้งหมด คือ 4871 คำ

สายอักขระเฉพาะภาษาอังกฤษ 2 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 122 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
์- ไ	23 (0.47%)
็-ค	22 (0.45%)

ฟ-ต	11 (0.23%)
็-ป	10 (0.21%)
ฮ-ฟ	7 (0.14%)
็-ค	7 (0.14%)
ฟ-ด	6 (0.12%)

็-ป	5 (0.1%)
ฟ –์	4 (0.08%)
พ-ฟ	2 (0.04%)
ธ-ั	2 (0.04%)
็-ส	2 (0.04%)

จากสายอักขระเฉพาะ 2-ตัว พบข้อสังเกตคือ มีการใช้ ็- ตามด้วยตัวอักษรใดก็ตาม เช่น ็-ค ็-ป และ ็- ตามด้วยตัวอักษร เช่น ็-ค ็-ป ็-ส เป็นต้น

สายอักขระเฉพาะภาษาอังกฤษ 3 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 2018 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
น-ด์-	113 (2.32%)
อ-ล-ล	31 (0.64%)
ิ้-ล	28 (0.57%)
์-เ-ว	28 (0.57%)
แ-อ-น	20 (0.41%)
อ-ล์-	18 (0.37%)
ล-ลี-	18 (0.37%)

็-อ-ค	18 (0.37%)
ส-เ-ป	17 (0.35%)
ิ-ส-ซ	14 (0.27%)
ส-ส์-	13 (0.27%)
บ-ล-ล	13 (0.27%)
เ-จ-อ	13 (0.27%)
ฟ-ฟ์-	12 (0.25%)
บ-อ-ล	12 (0.25%)
ค-ว-อ	12 (0.25%)
์-โ-ม	12 (0.25%)
ู้-ด	11 (0.23%)

อ-ล-ต	10 (0.21%)
ส-ล์-	10 (0.21%)
ว-ท์-	10 (0.21%)
ม-มี-	10 (0.21%)
ิ-ส-ค	10 (0.21%)
ิ-ล-ด	10 (0.21%)
์-บ-า	10 (0.21%)
ว-ย์-	9 (0.19%)
ล-ล์-	9 (0.19%)
ซ-ซ้-	9 (0.19%)
ไ-ล-เ	9 (0.19%)

โ-ฟ-ล	9 (0.19%)
เ-อ-ล	9 (0.19%)
็-อ-ป	9 (0.19%)
็-ค-ท	9 (0.19%)
ท-รู-	8 (0.16%)
ท-ติ-	8 (0.16%)
ท-ช์-	8 (0.16%)
ค-อ-ร	8 (0.16%)
็-อ-ต	8 (0.16%)
ว-ส-ต	7 (0.14%)
พ-ล-ย	7 (0.14%)

จากสายอักขระเฉพาะ 3-ตัว จะสังเกตเห็นรูปแบบของภาษา คือ มีการใช้ _ ์ มาก เช่น นด์ อล์ สส์ ฟฟ์ วย์ ลล์ เป็นต้น โดยเฉพาะ นด์ ใช้ตัดสินภาษาได้ 2.32% ซึ่งสูงที่สุดเมื่อเทียบกับสายอักขระเฉพาะขนาดอื่นๆ และพบสายอักขระเฉพาะ ็-อ_ เช่น ็-อ-ป ็-อ-ค ็-อ-ต เป็นต้น

สายอักขระเฉพาะภาษาอังกฤษ 4 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 1769 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
เ-อ็--ก	33 (0.68%)
ท-อ-ร์-	31 (0.64%)
อ-ร์--ก	16 (0.33%)
เ-น็--ต	16 (0.33%)
อ-ร์--ส	12 (0.25%)
เ-ม-ต-ร	11 (0.23%)
ส-ต-า-ร	10 (0.21%)

์-ตี-้	10 (0.21%)
อ-ม-มิ-	9 (0.18%)
ิ-ล-ล์-	9 (0.18%)
ิ-ร์--น	9 (0.18%)
ลี--ย์-	8 (0.16%)
มิ--ล-ล	8 (0.16%)
ค-เ-ต-อ	8 (0.16%)
ิ-เ-ต-อ	8 (0.16%)
์-เ-บิ-	8 (0.16%)
า-เ-นี-	7 (0.14%)
อ-ริ--ส	7 (0.14%)

อ-เ-ร-ส	7 (0.14%)
พ-า-ร์-	7 (0.14%)
พ-รี--เ	7 (0.14%)
บ-ร-อ-ด	7 (0.14%)
อ-ป-ปี-	6 (0.12%)
ส-โ-ต-น	6 (0.12%)
ลิ--เ-ว	6 (0.12%)
มี--ย-ม	6 (0.12%)
ม-น-เ-ด	6 (0.12%)
บ-อ-ดี-	6 (0.12%)
น-นิ-่	6 (0.12%)

น-เ-ดี-	6 (0.12%)
ค-อ-ม-เ	6 (0.12%)
ค-ลี--น	6 (0.12%)
เ-ว-อ-ล	6 (0.12%)
เ-ว-โ-ล	6 (0.12%)
เ-ม-ติ-	6 (0.12%)
เ-ด้-น	6 (0.12%)
ิ-เ-อ-ร	6 (0.12%)
ั-น-ส์-	6 (0.12%)
ฮ-า-ร์-	5 (0.1%)
อ-รั--ส	5 (0.1%)

จากสายอักขระเฉพาะ 4-ตัว ไม่มีลักษณะเด่นของภาษาให้เห็นชัดเจน แต่สังเกตได้ว่า มีสายอักขระเฉพาะบางตัวที่ใช้ได้จริงมากกว่าสายอักขระอื่นๆ อย่างเห็นได้ชัด คือ เ-อ็--ก (0.68%) และ ท-อ-ร์- (0.64%)

สายอักขระเฉพาะภาษาอังกฤษ 5 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 962 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
ส-เ-ต-อ-ร	30 (0.62%)
เ-ต-อ-ร์	27 (0.55%)
เ-ล-อ-ร์	21 (0.43%)
ล-เ-ล-อ-ร	21 (0.43%)
เ-บ-อ-ร์	20 (0.41%)
ฟ-อ-ร์--ด	16 (0.33%)
ฟ-อ-ร์--เ	14 (0.29%)
น-เ-ซ-อ-ร	14 (0.29%)

ก-เ-ต-อ-ร	13 (0.27%)
เ-ซ-อ-ร์	13 (0.27%)
ฟ-อ-ร์--ม	12 (0.25%)
เ-ซ-อ-ร์	12 (0.25%)
ว-อ-ร์--เ	11 (0.23%)
บ-อ-ร์--เ	11 (0.23%)
เ-ป-อ-ร์	11 (0.23%)
อ-ร์--เ-ก	10 (0.21%)
ว-อ-ร์--ซ	10 (0.21%)
ม-เ-ม-อ-ร	10 (0.21%)
ค-อ-ร์--เ	10 (0.21%)
ั-น-เ-ด-อ	10 (0.21%)

ต-อ-ร์--เ	9 (0.18%)
ว-อ-ร์--ด	8 (0.16%)
ฟ-อ-ร์--แ	8 (0.16%)
ซ-อ-ร์--เ	8 (0.16%)
ค-า-ร์--เ	8 (0.16%)
อ-ร์--เ-ม	7 (0.14%)
อ-ร์--เ-ซ	7 (0.14%)
อ-เ-ว-อ-ร	7 (0.14%)
ล็--ก-ซ์-	7 (0.14%)
ค-า-ร์--โ	7 (0.14%)
เ-รี--ย-ม	7 (0.14%)
อ-อ-ร์--แ	6 (0.12%)

ด-ิ-ั-ง	6 (0.12%)
ก-เ-ก็--ต	6 (0.12%)
อิ--น-ดั	5 (0.1%)
อ-ร์--ริ-	5 (0.1%)
อ-ร์--ติ-	5 (0.1%)
อ-ร์--แ-ม	5 (0.1%)
ว-เ-ซ-อ-ร	5 (0.1%)
ม-อ-ร์--เ	5 (0.1%)
ม-เ-ป-อ-ร	5 (0.1%)
ป-อ-ร์--ม	5 (0.1%)
ป-อ-ร์--เ	5 (0.1%)
บ-อ-ร์--ร	5 (0.1%)

จากสายอักขระเฉพาะ 5-ตัว จะสังเกตเห็นรูปแบบของภาษา คือการใช้ เ_อร์ เช่น เตอร์ เบอร์ เลอร์ เซอร์ เชอร์ เป็นต้น การใช้_อร์_ เช่น ฟอร์ด วอร์เ บอร์เ วอร์ซ ตอร์เ เป็นต้น และ _เ_อร เช่น สเตอร นเซอร กเตอร นเซอร มเปอร เป็นต้น

สายอักขระคำทับศัพท์ภาษาญี่ปุ่นที่ถูกนำมาใช้ในระบบการใช้สายอักขระเฉพาะ โดยที่ผลในตารางคือ จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง ส่วนตัวเลขในวงเล็บคือจำนวนเปอร์เซ็นต์เมื่อคิดจากจำนวนของสายอักขระ 1-5 ตัวที่ใช้ระบุภาษาถูกต้องทั้งหมด คือ 5409 คำ

สายอักขระเฉพาะภาษาญี่ปุ่น 2 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 160 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
ฮ-ะ	58 (1.07%)
ิ-ฮ	16 (0.3%)

ุ-ฉ	12 (0.22%)
ึ-ท	8 (0.15%)
ึ-จ	6 (0.11%)
ะ-ฌ	5 (0.09%)
ด-ะ	5 (0.09%)
ช-จ	5 (0.09%)

ึ-ย	5 (0.09%)
ู-ฮ	4 (0.07%)
ง็-	3 (0.06%)
ไ-ง	3 (0.06%)
ึ-ฮ	3 (0.06%)
ึ-ช	3 (0.06%)

ท-ฮ	2 (0.04%)
ช-ห	2 (0.04%)
ง-ญ	2 (0.04%)
โ-ฌ	2 (0.04%)
ั-อ	2 (0.04%)
ึ-ท	2 (0.04%)

จากสายอักขระเฉพาะ 2-ตัว จะสังเกตเห็นรูปแบบของภาษา คือการใช้ เ_ะ เช่น ฮะ ดะ โดยที่ ฮะ ใช้ตัดสินภาษา และ ึ- _ เช่น ึ-จ ึ-ย ึ-ฮ ึ-ช

สายอักขระเฉพาะภาษาญี่ปุ่น 3 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 3392 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
มุ--ช	7 (0.13%)
ท-สุ-	7 (0.13%)
ซ-คิ-	7 (0.13%)
ชิ--เ	7 (0.13%)
จิ--ช	7 (0.13%)
งิ--ว	7 (0.13%)
คุ--ช	7 (0.13%)
โ-ค-ย	7 (0.13%)
เ-ร-ะ	7 (0.13%)
ุ-ท-ะ	7 (0.13%)
ุ-ต-ะ	7 (0.13%)
ุ-ไ-ก	7 (0.13%)
ุ-โ-ง	7 (0.13%)
ั-ย-ร	7 (0.13%)
ฮิ--ง	6 (0.11%)
อิ--ว	6 (0.11%)
อ-ะ-อ	6 (0.11%)
ย-จิ-	6 (0.11%)
ม-ะ-ส	6 (0.11%)
น-ะ-ช	6 (0.11%)
ท-อิ-	6 (0.11%)
ญิ--ค	6 (0.11%)
ซุ--ม	6 (0.11%)
ง-ญิ-	6 (0.11%)
ง-โ-ง	6 (0.11%)
คิ--จ	6 (0.11%)
ค-า-อ	6 (0.11%)
ค-ะ-ส	6 (0.11%)
ก-า-ค	6 (0.11%)
ก-ะ-ม	6 (0.11%)
ไ-ด-จ	6 (0.11%)
โ-ฮ-ก	6 (0.11%)
โ-ม-จ	6 (0.11%)
ู-ง-	6 (0.11%)
ึ-มิ-	6 (0.11%)
ึ-บุ-	6 (0.11%)
ึ-คิ-	6 (0.11%)
ิ-ด-ะ	6 (0.11%)
ั-ง-ญ	6 (0.11%)
า-ก-	5 (0.09%)
ะ-ฉิ-	5 (0.09%)
อุ--ร	5 (0.09%)
อุ--ค	5 (0.09%)
สุ--ซ	5 (0.09%)
สุ--โ	5 (0.09%)
ส-ะ-ค	5 (0.09%)
ว-โ-ด	5 (0.09%)
น-ะ-ค	5 (0.09%)
ท-ะ-ช	5 (0.09%)
ท-ค-ะ	5 (0.09%)
ต-โ-ฮ	5 (0.09%)
ช-ะ-ว	5 (0.09%)
ช-ย-	5 (0.09%)
ง-ุ-โ	5 (0.09%)
ง-ง-ะ	5 (0.09%)
ง-โ-ญ	5 (0.09%)
ค-ุ--ย	5 (0.09%)
ค-ะ-อ	5 (0.09%)
ก-า-โ	5 (0.09%)
โ-ย-ญ	5 (0.09%)
โ-ค-ก	5 (0.09%)
เ-น-ง	5 (0.09%)
เ-ท-ะ	5 (0.09%)
เ-ช-ะ	5 (0.09%)
เ-ก-โ	5 (0.09%)
ุ-น-ย	5 (0.09%)
ุ-คิ-	5 (0.09%)
ุ-โ-จ	5 (0.09%)
ุ-เ-อ	5 (0.09%)
ิ-ญี-	5 (0.09%)
ิ-ง-ง	5 (0.09%)
ิ-โ-บ	5 (0.09%)

จากสายอักขระเฉพาะ 3-ตัว พบว่ามีการใช้ภาษาที่ส่วนใหญ่ประกอบด้วยสระเสียงสั้น เห็นได้จาก สายอักขระเฉพาะที่ใช้เสียงเอะ หรือ เ_ะ เช่น เระ เทะ เซะ เป็นต้น สายอักขระเฉพาะที่ใช้เสียง อุโอ หรือ _ุโ_ เช่น สุโ งุโ _ุโง _ุโจ เป็นต้น และ สายอักขระเฉพาะที่ใช้เสียง อุอะ หรือ _ุ_ะ เช่น _ุทะ _ุตะ เป็นต้น

สายอักขระเฉพาะภาษาญี่ปุ่น 4 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 1683 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
เ-ค-อิ-	33 (0.61%)
ชิ--น-โ	19 (0.35%)
มู--ร-ะ	16 (0.3%)
คิ--โ-ย	14 (0.26%)
กุ--ร-ะ	14 (0.26%)
อ-ะ-ริ-	13 (0.24%)
ชิ--คิ-	13 (0.24%)
กุ--ชิ-	13 (0.24%)
อุ--จิ-	11 (0.2%)
เ-ส-ง-เ	11 (0.2%)
ฮ-า-ร-า	9 (0.17%)
ชิ--มิ-	9 (0.17%)
ฮิ--ร-ะ	8 (0.15%)
ท-า-ก-า	8 (0.15%)
ชิ--ต-ะ	8 (0.15%)
ั-น-จิ-	8 (0.15%)
ะ-ชิ--โ	7 (0.13%)
ริ--โ-น	7 (0.13%)
ร-ั-น-โ	7 (0.13%)
ร-ั-ค-ค	7 (0.13%)
ร-ะ-ค-	7 (0.13%)
ชิ--กิ-	7 (0.13%)
โ-อ-โ-ม	7 (0.13%)
อุ--ร-า	6 (0.11%)
อิ--ว-า	6 (0.11%)

อิ--โ-ม	6 (0.11%)
อ-า-ชิ-	6 (0.11%)
สุ--เ-ก	6 (0.11%)
น้--น-เ	6 (0.11%)
คู--นิ-	6 (0.11%)
คุ--เ-ร	6 (0.11%)
เ-ค-โ-อ	6 (0.11%)
ิ-ม-บุ-	6 (0.11%)
ะ-โ-ค-โ	5 (0.09%)
ะ-เ-ค-ะ	5 (0.09%)
อิ--จิ-	5 (0.09%)
ริ--คิ-	5 (0.09%)
ม-ะ-เ-น	5 (0.09%)

บ้--ค-ค	5 (0.09%)
น-ะ-ก-ะ	5 (0.09%)
น-เ-ง็-	5 (0.09%)
ด-า-ชิ-	5 (0.09%)
ซ-า-กุ-	5 (0.09%)
ซ-า-โ-ย	5 (0.09%)
จิ--ริ-	5 (0.09%)
คิ--ว-โ	5 (0.09%)
ค-า-โ-ต	5 (0.09%)
เ-ร-คิ-	5 (0.09%)
ิ-ชิ--ม	5 (0.09%)
้-ต-โ-ต	5 (0.09%)
า-ม-า-ก	4 (0.07%)

ะ-อิ--ช	4 (0.07%)
ะ-อิ--โ	4 (0.07%)
ะ-คิ--ส	4 (0.07%)
ะ-คิ--น	4 (0.07%)
อิ--ก-ะ	4 (0.07%)
อิ--โ-ย	4 (0.07%)
อิ--โ-ต	4 (0.07%)
อิ--โ-ซ	4 (0.07%)
อ-ซ-า-ก	4 (0.07%)
สุ--ร-ะ	4 (0.07%)
สุ--ชิ-	4 (0.07%)
ว-า-ก-า	4 (0.07%)
รุ--ชิ-	4 (0.07%)

ร-ะ-กุ-	4 (0.07%)
มู--ร-า	4 (0.07%)
มิ--ก-า	4 (0.07%)
ม-า-ซุ-	4 (0.07%)
ม-า-เ-อ	4 (0.07%)
ม-บุ--น	4 (0.07%)
ม-บุ--ง	4 (0.07%)
น-เ-คี-	4 (0.07%)
ช--น-ค	4 (0.07%)
ชิ--โ-ก	4 (0.07%)
ง้--ก-ก	4 (0.07%)
ค--ริ-	4 (0.07%)

จากสายอักขระเฉพาะ 4-ตัว พบว่ามีการใช้ภาษาที่ประกอบด้วยสระสียงสั้นเห็นได้จากสายอักขระเฉพาะที่ใช้เสียง ออิ เช่น ชิคิ ชิมิ ชิกิ อิจิ ริคิ จิริ เป็นต้น เสียง อิโอ เช่น คิโย อิโม อิโย อิโต อิโซ ชิโก เป็นต้น เสียง ออะ เช่น ฮิระ ชิตะ อิกะ เป็นต้น และเสียง อุอิ เช่น กุชิ อุจิ สุชิ คุริ เป็นต้น นอกจากนี้ยังพบว่ามีสายอักขระเฉพาะที่ใช้ตัดสินภาษาได้ดีกว่าสายอักขระเฉพาะอื่น คือ เคอิ (0.61%)

สายอักขระเฉพาะภาษาญี่ปุ่น 5 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 174 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
เ-รี--ย-ะ	10 (0.18%)
คี--ย-ว-เ	6 (0.11%)
้น-เ-ซ-น	5 (0.09%)
มิ--น-โ-ต	4 (0.07%)
ี-อิ--น-ไ	4 (0.07%)

อ-ะ-มิ--ด	3 (0.06%)
รี--ย-ว-ร	3 (0.06%)
รี--ย-ว-เ	3 (0.06%)
กี--ย-ว-ไ	3 (0.06%)
เ-มี--ย-ว	3 (0.06%)
เ-ค-น-เ-ซ	3 (0.06%)
า-น-า-มิ-	3 (0.06%)
า-ต-า-ริ-	2 (0.04%)
ฮั้-น-โ-ด	2 (0.04%)

อิ--ร้--น	2 (0.04%)
อิ--เ-ซ-อ	2 (0.04%)
อ-ะ-โ-ร-ง	2 (0.04%)
ว-า-น-า-ม	2 (0.04%)
มิ--น-โ-น	2 (0.04%)
ม-ะ-โ-ร-น	2 (0.04%)
บ-า-ย-า-ช	2 (0.04%)
นี--ย-ว-ไ	2 (0.04%)
นิ--โ-อ-เ	2 (0.04%)

น-า-ก-า-ช	2 (0.04%)
น-า-ก-า-โ	2 (0.04%)
น้--น-ไ-ด	2 (0.04%)
น-บุ--ร-ุ	2 (0.04%)
ชิ--ร-า-เ	2 (0.04%)
ชิ--น-เ-ซ	2 (0.04%)
คี--ย-ว-ซ	2 (0.04%)
คี--ย-ว-ไ	2 (0.04%)
คิ--น-เ-ท	2 (0.04%)

จากสายอักขระเฉพาะ 5-ตัว สังเกตเห็นว่ามีการใช้ภาษาที่ใช้เสียง เอียว หรือ ี_ยว_มาก เช่น คียวเ รียวร เรียวเ เมียว นียวไ คียวไ เป็นต้น

สายอักขระคำทับศัพท์ภาษาฝรั่งเศสที่ถูกนำมาใช้ในระบบการใช้สายอักขระเฉพาะ โดยที่ผลในตารางคือ จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง ส่วนตัวเลขในวงเล็บคือจำนวนเปอร์เซ็นต์เมื่อคิดจากจำนวนรวมของสายอักขระ 1-5 ตัวที่ใช้ระบุภาษาถูกต้องทั้งหมด คือ 204 คำ เนื่องจากสายอักขระเฉพาะในคำทับศัพท์ภาษาฝรั่งเศสที่ถูกนำไปใช้จริงมีแค่สายอักขระเฉพาะ 3-5 ตัว และแต่ละตัวก็ถูกนำไปใช้จำนวนน้อย เป็นเพราะคลังข้อมูลน้อย ทำให้ไม่ค่อยพบสายอักขระที่ใช้จริงในระบบ ดังนั้นผู้วิจัยจึงวิเคราะห์สายอักขระเฉพาะ 3-5 ตัวเพียงคร่าวๆ เท่านั้น อาจไม่สามารถนำไปวิเคราะห์เพื่อหาลักษณะของการใช้ภาษาได้

สายอักขระเฉพาะ 3 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 57 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
ง-ป-า	5 (2.45%)
◌ี-แ-ย	4 (1.96%)
◌์-แ-ต	4 (1.96%)
ง-แ-ฟ	3 (1.47%)
◌ู-แ-ต	3 (1.47%)
อ-แ-ต	2 (0.98%)
ตู--ร	2 (0.98%)
ด◌ุ--ช	2 (0.98%)
ด-อ-โ	2 (0.98%)
แ-ต-เ	2 (0.98%)
◌ู-ตู-	2 (0.98%)
◌์-ด-ง	2 (0.98%)

จากสายอักขระเฉพาะ 3-ตัว สังเกตได้ว่า สายอักขระที่ใช้จริงส่วนใหญ่ประกอบด้วย “แ” เป็นส่วนประกอบในสายอักขระเฉพาะ

สายอักขระเฉพาะ 4 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 101 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
อ-ง-เ-บ	4 (1.96%)
ว-า-ล-อ	3 (1.47%)
ติ--เ-ย	3 (1.47%)
ซ็--ง-โ	3 (1.47%)
ก-า-บ-า	3 (1.47%)
เ-บ-ร-อ	3 (1.47%)
า-แ-ค-เ	2 (0.98%)
อ-ง-เ-ซ	2 (0.98%)
ว-า-ลี-	2 (0.98%)
ล-อ-ง-ช	2 (0.98%)
ม-อ-ง-ต	2 (0.98%)
ม-อ-ง-ด	2 (0.98%)
ฟ-อ-ง-แ	2 (0.98%)
ต-า-ส์-	2 (0.98%)
ซ-เ-ว-อ	2 (0.98%)
ง-ช-า-ร	2 (0.98%)
ก-า-ส-ซ	2 (0.98%)
ก็--อ-ง	2 (0.98%)
ก-รั--ว	2 (0.98%)
โ-ร-ต์-	2 (0.98%)
เ-ล-อ-เ	2 (0.98%)
เ-ต-ส-ก	2 (0.98%)
เ-ต-ร-อ	2 (0.98%)
เ-ซ-ลิ-	2 (0.98%)

จากสายอักขระเฉพาะ 4-ตัว สังเกตได้ว่ามีการใช้ภาษาที่ประกอบด้วยเสียง ออง มาก เช่น ลองช มองต มองด ฟองแ และเสียง เออ เช่น เบรอ ซเวอ เลอเ เตรอ เป็นต้น

สายอักขระเฉพาะ 5 ตัว (จำนวนครั้งที่ระบุภาษาได้ถูกต้องทั้งหมด 46 ครั้ง)

สายอักขระเฉพาะ	จำนวนครั้งที่ใช้ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)
ร์--เ-ล-อ	5 (2.45%)
ู-ร์--เ-น	4 (1.96%)
ก-า-ร์--ก	3 (1.47%)
า-ร์--โ-ต	2 (0.98%)
า-ต-า-ล-อ	2 (0.98%)
ร-า-เ-บ-อ	2 (0.98%)
ป็--เ-อ-ต	2 (0.98%)
บ--ร์--เ	2 (0.98%)
บ-ร์--ก-อ	2 (0.98%)
ท-ริ--อ-า	2 (0.98%)
แ-ว-ร์--ด	2 (0.98%)
ู-เ-อ-ต์-	2 (0.98%)
า-ว-า-ร์-	1 (0.49%)
า-ร์--น-า	1 (0.49%)
อ-ม-แ-บ-ร	1 (0.49%)
อ-ง-ท-ร-า	1 (0.49%)
ล-อ-ง-เ-ก	1 (0.49%)
ม-า-เ-ล-อ	1 (0.49%)
ป็--เ-อ-ม	1 (0.49%)
บ-ริ--ซ-โ	1 (0.49%)
ซ-อ-ร์--บ	1 (0.49%)

จากสายอักขระเฉพาะ 5-ตัว สังเกตได้ว่า มีการใช้ภาษาที่ประกอบด้วย เสียง อา และ เออ มาก เช่น ร์เลอ การ์ก าตาลอ ราเบอ มาเลอ กาบาส เป็นต้น

จากสายอักขระเฉพาะที่ใช้จริงเหล่านี้ สามารถแสดงให้เห็นลักษณะสำคัญของแต่ละภาษาได้ เช่น สายอักขระเฉพาะคำไทย 3-ตัว ที่ใช้บ่อย คือ บรร กรร อรร สรร ทำให้วิเคราะห์ได้ว่า ภาษาไทยมีการใช้ ร หัน บ่อย จนเป็นลักษณะที่เกิดเฉพาะในภาษาไทย ซึ่งลักษณะเฉพาะของแต่ละภาษาเหล่านี้จะแสดงต่อไปใน หัวข้อ 4.7 การวิเคราะห์สายอักขระเฉพาะ

นอกจากนี้ จากการศึกษาสายอักขระเฉพาะที่ใช้จริงยังพบข้อสังเกตบางอย่าง เช่น ถ้าสังเกตในภาษาไทย อักขระตัวเดียวสามารถใช้ระบุภาษาได้ เมื่อเป็นอักขระ 2 ตัวอักษรอย่าง ษ ผ ฐ ฤ ใ จะไม่ปรากฏในสายอักขระแบบ 2 ตัวอักษร แต่บางตัวอักษรอย่าง ณ ภ กลับยังพบในสายอักขระแบบ 2 อักษร แต่ไม่พบใน 3 4 และ 5 ตัวอักษร เพราะว่า โดยหลักแล้วเมื่อตัวอักษรอย่าง ษ ใช้ตัดสินเป็นคำไทยได้แล้ว ก็จะไม่พบใน 2 ตัวอักษรแล้ว ส่วน ณ_ หรือ ณ+ ตัวอะไรก็ตาม ใน 2 อักขระ ก็จะไม่พบใน 3-5 ตัวอักษรแล้ว ซึ่งก็เป็นจริง ทั้งหมดนี้เป็นข้อสังเกต ซึ่งจะกล่าวโดยละเอียดในหัวข้อ 4.7

4.7 การวิเคราะห์สายอักขระเฉพาะ

เนื่องจากในการทดลองทำให้เห็นการปรากฏของสายอักขระเฉพาะในคำไทย คำทับศัพท์ในภาษาอังกฤษ ภาษาญี่ปุ่นและภาษาฝรั่งเศส และสายอักขระเฉพาะสามารถใช้ในการระบุภาษาได้มากกว่า 50% ในบทนี้ผู้วิจัยจึงสนใจที่จะวิเคราะห์สายอักขระเฉพาะเหล่านี้ ทั้งตำแหน่งที่ปรากฏของสายอักขระและด้านการใช้ภาษา โดยผู้วิจัยทำการวิเคราะห์ด้านรูปและเสียง เพื่อเป็นข้อมูลสนับสนุนว่าทำไมสายอักขระเฉพาะเหล่านี้จึงใช้ในการระบุภาษาได้

หลังจากทดลองโปรแกรมใช้คลังข้อมูลแต่ละภาษาเพื่อสร้างสายอักขระเฉพาะ ผลจากการทดลองที่ได้คือ คลังข้อมูลสายอักขระเฉพาะทั้งหมด 4 ภาษา ซึ่งภายในคลังข้อมูลประกอบด้วย สายอักขระเฉพาะที่มีความยาวอักขระหรือตัวอักษรตั้งแต่ 1-5 ตัว จากการใช้เอ็นแกรม 1-5 แกรม และเนื่องจากในการทดลองต้องทดสอบเพื่อหาสายอักขระเฉพาะจากชุดข้อมูล 5 ชุดในแต่ละภาษา ดังนั้นสายอักขระเฉพาะทั้งหมดที่ได้ คือ สายอักขระเฉพาะทั้งหมด 4 ภาษาและภาษาละ 5 ชุดสายอักขระ แต่เพื่อให้เห็นสายอักขระเฉพาะจริงๆ ในแต่ละภาษา สายอักขระเฉพาะที่ผู้วิจัยจะนำมาพิจารณาคือ เอาสายอักขระเฉพาะจากข้อมูลคำรวมกันทั้งหมดมาพิจารณาทีเดียว เพราะสายอักขระเฉพาะที่ได้จากการทดลองชุดข้อมูลที่ 1 อาจไม่ใช่สายอักขระเฉพาะที่แท้จริงในชุด 2 3 4 และ 5 ก็ได้

จากการรวบรวมคำทับศัพท์และหาสายอักขระเฉพาะของแต่ละภาษา ทำให้เห็นลักษณะสำคัญของแต่ละภาษา เช่น ลักษณะทางเสียงของภาษา ที่สะท้อนให้เห็นเมื่อเขียนด้วยอักษรไทย และยังให้ข้อสังเกตของภาษาซึ่งไม่สามารถอธิบายเหตุผลได้ ผู้วิจัยจึงต้องการวิเคราะห์สายอักขระเฉพาะเพื่อแสดงลักษณะสำคัญทางคำไทย คำทับศัพท์ในภาษาอังกฤษ ภาษาญี่ปุ่นและภาษาฝรั่งเศสตามลำดับต่อไป

4.7.1 สายอักขระคำไทย

ในการวิเคราะห์สายอักขระคำไทย ซึ่งผู้วิจัยวิเคราะห์ด้วยสายอักขระเฉพาะขนาด 1-5 แกรม ทำให้พบว่าคำไทยมีการใช้ตัวอักษร อักขรวิธีของการใช้คำไทย และข้อสังเกตของสายอักขระเฉพาะ ซึ่งสิ่งเหล่านี้จะสะท้อนให้เห็นลักษณะของภาษาไทยที่คำภาษาอื่นไม่ปรากฏ มีดังนี้

1. <u>การใช้ตัวอักษร ทั้งพยัญชนะ สระ และวรรณยุกต์</u> การใช้ตัวอักษรก็มีส่วนในการสะท้อนให้เห็นลักษณะของคำไทยได้ โดยตัวอักษรที่ปรากฏเฉพาะในคำไทย ได้แก่ ฆ ฎ ฏ ฐ ฑ ฒ ผ ฤ ษ ฬ ใ –ำ และ วรรณยุกต์จัตวา ซึ่งอักขระหรือตัวอักษรเหล่านี้ไม่ปรากฏในคำทับศัพท์ภาษาอังกฤษ ญี่ปุ่น และฝรั่งเศส เพราะมีการใช้อักขระหรือตัวอักษรอื่นแทนเสียงของอักขระเหล่านี้อยู่แล้ว และถือว่ามีความหมายเดียวกัน เช่น ในการใช้พยัญชนะ อักขระ “ฆ” ในภาษาอื่นใช้ “ค”, อักขระ “ฎ” ในภาษาอื่นใช้ “ด”, อักขระ “ฏ” ในภาษาอื่นใช้ “ต”, อักขระ “ฐ ฑ ฒ” ในภาษาอื่นใช้ “ท ถ”, อักขระ“ผ” ในภาษาอื่นใช้ “พ”, อักขระ“ฤ ฬ” ในภาษาอื่นใช้ “ร” และ อักขระ

“ษ” ในภาษาอื่นใช้ “ส” แต่ในกรณีของคำไทยตัวอักษรเหล่านี้มีพบใช้ในคำไทยโดยเฉพาะคำที่มาจากบาลีสันสกฤต

และการใช้สระในคำไทย มีการใช้เสียงจากสระ “ใ” และสระ “–ำ” ซึ่งไม่ปรากฏในคำทับศัพท์ภาษาใดๆเลย เพราะเสียงจากสระ “ ใ” ในคำภาษาอื่นมักใช้เสียงจาก “ไ” แทน เช่น คำว่า ไมโคร ไนส์ เป็นต้น ส่วนเสียงจากสระ –ำ ในคำภาษาอื่นๆ มักใช้ เสียง อัม มาแทน เช่น คำว่า บัมเปอร์(bumper) ปั๊มป์(pump) เป็นต้น

นอกจากนี้ ในเรื่องของการใช้วรรณยุกต์ ซึ่งจริงๆแล้ว ไม่มีเกณฑ์ใดๆ ที่ให้คำทับศัพท์สามารถใส่วรรณยุกต์ได้ ผู้วิจัยพบว่ารูปสายอักขระ(วรรณยุกต์) ที่ปรากฏเมื่อเขียนทับศัพท์จากภาษาอื่นจะมีทั้งรูปวรรณยุกต์เอก โท ตรี เช่น คำว่า โน้ต เซ็นทรัล จ๊อกกิ้ง ปั๊มป์ เป็นต้น เป็นเพราะเมื่อทับศัพท์แล้วคำที่ทับศัพท์นั้นมีเสียงซ้ำกับคำไทย จึงใช้วรรณยุกต์เพื่อให้ออกเสียงในภาษาไทยได้และเคยใส่วรรณยุกต์กันมานานจนเป็นที่ยอมรับกันทั่วไป แต่พบว่ารูปวรรณยุกต์จัตวานั้นเป็นรูปที่ปรากฏในคำไทยเท่านั้น อาจเป็นเพราะไม่นิยมใส่เครื่องหมายจัตวา

2. <u>การใช้ ร หัน หรือ รร</u> ในภาษาไทย มีการใช้ _รร เพื่อแสดงเสียง –ัน โดยมักพบในคำมาจากภาษาสันสกฤต ซึ่งจะแตกต่างกับคำทับศัพท์ภาษาอื่นๆ ที่ใช้สายอักขระ “–ัน” สายอักขระเดียวเท่านั้น คือใช้รูปสายอักขระ “–ัน” แทนเสียง อัน เช่นคำว่า ธันเดอร์ เป็นต้น สายอักขระเฉพาะ _รร ที่ปรากฏ เช่น กรร ครร จรร ดรร ธรร นรร บรร ภรร มรร วรร สรร หรร อรร เป็นต้น โดยสายอักขระที่ปรากฏบ่อย คือ กรร บรร ธรร วรร อรร แต่ในสายอักขระ _รร นี้ก็มีพบบ้างในคำทับศัพท์ภาษาอังกฤษ (แพรรี)

3. <u>การใช้เครื่องหมายการันต์</u> แม้ว่าในคำภาษาอื่นมีการใช้เครื่องหมายการันต์ด้วย แต่มีเพียงคำในภาษาไทยเท่านั้น ที่มีการใช้เครื่องหมายวรรณยุกต์ร่วมกับอักขระที่เป็นสระ เช่น สายอักขระ ◌็ จาก ดิ์ ติ์ ธิ์ เป็นต้น โดยเสียงของสายอักขระที่ปรากฏเด่นชัด คือ ธิ์ และสายอักขระ ◌์ จาก ตุ์ ถุ์ ทุ์ ธุ์ เป็นต้น

4. <u>การใช้สายอักขระ เ ◌ื</u> เช่น เกื เขื เคื เงื เจื เฉื เตื เถื เทื เนื เพื เมื เรื เลื เศื เสื เหื เอื เป็นต้น โดยสายอักขระที่ปรากฏเด่นชัด คือ เรื และ เสื โดยในสายอักขระนี้มาจากสายอักขระ เ◌ื_อ หรือเสียง เออ ซึ่งเป็นเสียงที่ไม่พบในคำทับศัพท์ภาษาอื่นเลย

5. <u>การใช้สายอักขระ ◌ื</u> ซึ่งจริงๆ แล้วในภาษาอื่นมีการใช้สระ อือ แต่เมื่อสระอือปรากฏตามด้วยรูปวรรณยุกต์ เอก โท ตรี จัตวา เช่น กื้ งื้ จื้ ฉื่ ชื่ ซื่ ปื่ ผื่ พื่ เป็นต้น จะเป็นสายอักขระเฉพาะที่เกิดขึ้นในภาษาไทยเท่านั้น โดยสายอักขระที่ปรากฏเด่นชัด คือ ชื่ ลื่ เช่น ในคำว่า ชื่อ เลื่อน เกื้อ เป็นต้น

6. สายอักขระ _าะ ซึ่งรูปแบบของสายอักขระนี้มาจากเสียง เอาะ ในคำไทย เช่น เกาะ เกราะ เป็นต้น ซึ่งในภาษาอื่นไม่มีเสียงนี้ เพราะภาษาอื่นต้องมีเสียงพยัญชนะสะกด จึงทำให้ไม่มีการใช้รูปสายอักขระ เอาะ แต่ใช้รูป ออ แทนและทำให้เป็นเสียงสั้นด้วยไม้ไต่คู้ เช่น ล็อต เป็นต้น

7. สายอักขระ ี_ยบ ซึ่งในคำไทย สายอักขระนี้มาจากเสียง เ_ียบ เช่น ชียบ ทียบ บียบ พียบ รียบ เป็นต้น แต่สายอักขระนี้ก็มีพบบ้างในคำทับศัพท์ภาษาญี่ปุ่น เช่น เฮียบังโตะ และ เกียบเบ็ทซึ

8. การใช้ ห นำ ในภาษาไทยมีการใช้ ห นำ เพื่อทำให้กลายเป็นเสียงจัตวา และเพื่อเพิ่มความหมาย เช่น หรัด หวัง หงิด หยิก หริ่ หมดเ เป็นต้น โดยที่การทับศัพท์ภาษาต่างๆ ในงานนี้มักไม่เขียนด้วย ห นำ เป็นเพราะในคำทับศัพท์มีการใช้เสียงจัตวาแต่ไม่นิยมใช้ ห นำ มาแทนเสียง จัตวา ส่วนใหญ่ละการใช้รูป ห นำ แต่ก็ยังออกเสียงจัตวาอยู่ เช่น คำว่า แกรนิต – แกรหนิต คลินิก – คลิหนิก คาร์บอเนต – คาร์บอเหนต เป็นต้น แต่มียกเว้นบางกรณี เช่น คำทับศัพท์ภาษาอังกฤษ มีการใช้คำว่า ไต้หวัน อาหรับ เป็นต้น ซึ่งในเกณฑ์ทับศัพท์ภาษาอังกฤษปัจจุบันไม่มีการใช้ ห นำ แต่คำเหล่านี้ทับศัพท์มาแต่เดิมจนเป็นที่ยอมรับกันแล้ว จึงอนุโลมให้ใช้ต่อไป และในกรณีของคำทับศัพท์ภาษาญี่ปุ่น ในคำว่า ฟูหยิคิมะ เป็นต้น ซึ่งเกณฑ์ทับศัพท์ภาษาญี่ปุ่นก็ไม่มีการใช้ ห นำ แต่คำนี้อาจเกิดขึ้นเพราะผู้ใช้อาจเขียนทับศัพท์โดยไม่คำนึงถึงกฎเกณฑ์เช่นเดียวกับในภาษาอังกฤษ เพียงแค่ทับศัพท์ตามการออกเสียงในภาษาไทยเท่านั้น

9. ข้อสังเกตของการใช้สายอักขระคำไทย ในการสร้างคำไทยซึ่งเป็นการใช้สายอักขระร่วมกัน ทำให้พบว่าการใช้สายอักขระบางสายร่วมกันทำให้เกิดสายอักขระเฉพาะ ซึ่งแม้ว่าจะไม่ใช่สายอักขระเฉพาะของภาษาไทย แต่ก็เป็นสายอักขระที่พบมากในภาษาไทย จึงนำมาแสดงไว้ดังต่อไปนี้

- สายอักขระ _า_ก มาจาก เสียง อา ตามด้วยเสียง โอะ ลดรูป ตามด้วยพยัญชนะ ก เช่น ทารก จารก ธารก บาลก ลาดก วาดก ชาลก เป็นต้น ซึ่งเป็นการใช้เสียง อา ร่วมกับเสียง อก แต่ก็พบบ้างในภาษาอื่น เช่น คำภาษาอังกฤษ (อลาสกา) คำภาษาญี่ปุ่น (อายาดกะ) และคำภาษาฝรั่งเศส (ปาสกาล)

- สายอักขระ _นทร์ เช่น วนทร์ พนทร์ เป็นต้น ซึ่งเป็นการใช้สายอักขระที่เป็นพยัญชนะติดกันร่วมกับเครื่องหมายการันต์ สายอักขระนี้แสดงให้เห็นลักษณะเฉพาะของการใช้ตัวการันต์ที่มากกว่าหนึ่งตัวอักษรของคำไทย เช่น ทร์ เป็นต้น

4.7.2 สายอักขระคำทับศัพท์ภาษาอังกฤษ

ในการวิเคราะห์สายอักขระคำทับศัพท์ภาษาอังกฤษ ซึ่งผู้วิจัยวิเคราะห์ด้วยสายอักขระเฉพาะขนาด 2-5 แกรม ทำให้พบว่าคำทับศัพท์ภาษาอังกฤษ มีการใช้ตัวอักษร อักขรวิธีของภาษาซึ่งในที่นี้คือการใช้เครื่องหมายการันต์ และข้อสังเกตของภาษาจากสายอักขระเฉพาะ ซึ่งสิ่งเหล่านี้จะสะท้อนให้เห็นลักษณะของคำทับศัพท์ภาษาอังกฤษ ที่คำภาษาอื่นไม่ปรากฏ มีดังนี้

1. <u>การใช้อักษรที่เป็นตัวสะกด</u> เช่น ตัวสะกด ล ค ส ฟ เป็นตัวสะกดที่พบมากในคำทับศัพท์ภาษาอังกฤษ เนื่องจากในภาษาไทยมีการใช้เสียงพยัญชนะท้ายหรือตัวสะกดเป็นเสียงระเบิดไม่ก้องและเสียงนาสิกเท่านั้น แต่เมื่อถอดอักษรจากอักษรภาษาอังกฤษ l k s และ f จึงปรากฏอักขระ ล ค ส ฟ ในพยัญชนะท้ายในคำทับศัพท์ อาจเป็นเพราะผู้ใช้ต้องการออกเสียงด้วยสำเนียงของภาษาอังกฤษ

โดยจากการใช้อักขระเหล่านี้เมื่อปรากฏร่วมกับเสียงสระอื่นๆเช่น เสียง “แอ” และ “ออ” จะทำให้เกิดสายอักขระเฉพาะของภาษาอังกฤษ ได้แก่

- สายอักขระ แ_ค เช่น แมค แทค แนค แลค แอค เป็นต้น
- สายอักขระ แ_ส เช่น แจส แพส แวส แอส เป็นต้น แต่สายอักขระนี้ยังพบได้บ้างในภาษาฝรั่งเศส เช่น แตแรส นาร์บงแนส มงแตสกิเออ เป็นต้น
- สายอักขระ _อส เช่น กอส คอส ลอส เป็นต้น แต่สายอักขระนี้ยังพบได้บ้างในคำไทยแต่เป็นการมองข้ามพยางค์ (หอสมุด ซอสามสาย) และในภาษาฝรั่งเศส (กาซอส ปลูวิยอส)
- สายอักขระ _อฟ เช่น คอฟ ซอฟ ทอฟ มอฟ เป็นต้น แต่สายอักขระนี้ยังพบได้บ้างในภาษาฝรั่งเศส เช่น กอฟแฟรง ออฟฟิซิเยส์ เป็นต้น

และยังพบตัวสะกดที่พบมากในภาษาอังกฤษ คือ ต เช่น สายอักขระ เ_็ต เช่น เด็ต เน็ต เร็ต เล็ต เป็นต้น แต่ตัวสะกดนี้ก็มีบ้างในภาษาอื่น เช่น คำไทย (ภูเก็ต) ภาษาญี่ปุ่น (เท็ตสึยะ) และในภาษาฝรั่งเศส (กัมเบ็ตตา)

2. <u>การใช้เครื่องหมายการันต์</u> ซึ่งจะต่างกับคำในภาษาอื่นตรงที่ ในภาษาอื่นจะไม่มีการใช้เครื่องหมายการันต์ในลักษณะดังต่อไปนี้

- สายอักขระ เ_น ์ เช่น เชนส์ เรนต์ เรนส์ เป็นต้น สายอักขระนี้เกิดขึ้นเพราะคำเหล่านี้เมื่อถอดอักษรเป็นคำทับศัพท์ภาษาอังกฤษแล้ว จะปรากฏเสียง เ_น_ ซึ่งในคำไทยออกเสียงแค่ เอน เท่านั้น ดังนั้นจึงมีการใช้เครื่องหมายการันต์ในพยัญชนะท้ายที่ไม่ออกเสียงในภาษาไทย เช่น ในภาษาอังกฤษ lens จะถูกทับศัพท์เป็น เลนส จึงมีการใส่การันต์บน ส เพื่อตัดเสียง ส ออกไป

- สายอักขระ เ_อร์ เช่น เชอร์ เคอร์ เจอร์ เบอร์ เปอร์ เลอร์ เป็นต้น สายอักขระนี้เกิดขึ้นเพราะ ในภาษาอังกฤษมีการใช้ er ท้ายพยางค์ และเมื่อถอดอักษรทับศัพท์เป็นไทยแล้วจะเป็น เออร เแต่ในภาษาไทยไม่ออกเสียง ร ดังนั้นจึงต้องใส่เครื่องหมายการันต์ เพื่อตัดเสียง ร ออกไป แต่สายอักขระนี้ยังพบได้บ้างในภาษาฝรั่งเศส เช่น เมอซีเออร์ กูร์มาเยอร์ เป็นต้น

- สายอักขระ ี_ _์ เช่น กีตส์ กียร์ คียร์ จียร์ เป็นต้น คำเหล่านี้เมื่อถอดอักษรเป็นคำทับศัพท์ภาษาอังกฤษแล้วเป็นพยัญชนะตัวสุดท้ายเป็นพยัญชนะที่ไม่ออกเสียง จึงต้องใส่เครื่องหมายการันต์กำกับ โดยสายอักขระนี้พบมากใน เ ี_ _์ เช่น คำว่า เกียร์ (gear) เบียร์ (bear) เป็นต้น

3. <u>ข้อสังเกตของการสร้างคำทับศัพท์ภาษาอังกฤษ</u> ในการสร้างคำทับศัพท์ภาษาอังกฤษ เป็นการใช้สายอักขระร่วมกันทั้งพยัญชนะและสระ ดังนั้น ทำให้พบว่าการใช้สายอักขระบางสายร่วมกันทำให้เกิดสายอักขระ ซึ่งแม้ว่าจะไม่ใช่สายอักขระเฉพาะของคำภาษาอังกฤษ แต่ก็เป็นสายอักขระที่พบมากในคำภาษาอังกฤษในขณะที่ไม่พบในภาษาอื่นหรือพบน้อยกว่ามาก ดังนั้นจึงนำมาแสดงไว้ดังต่อไปนี้

- สายอักขระ _ไ เช่น ค์ไ ซ์ไ ด์ไ ต์ไ เป็นต้น

- สายอักขระ _์_ไ เช่น ร์สไ ร์คไ ท์ซไ ล์ดไ เป็นต้น

- สายอักขระ ู_เ_เ เช่น ู-เมเ ู-เรเ ู-เลเ ู-เอเ เป็นต้น

- สายอักขระ แ_ ี เช่น แกรี แครี แทรี แบรี แพรี แฟรี แมรี เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในภาษาอื่น เช่น คำไทย (กุญแจผี) และภาษาฝรั่งเศส (แดกีย์ บูแตอี)

- สายอักขระ แ_ ู เช่น แมนู แคนู แวลู เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในภาษาฝรั่งเศส เช่น อาแชกูร์ เป็นต้น

- สายอักขระ าเ_เ เช่น -าเชเ -าเดเ เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในภาษาอื่น เช่น ภาษาฝรั่งเศส (กาเปเตียง) และภาษาญี่ปุ่น (คาเคเมะ)

- สายอักขระ าิ_ เช่น -าเซิ -าเทิ -าเบิ -าเยิ -าเวิ เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในคำไทย เช่น ร่าเริง ปลาเงินปลาทอง

- สายอักขระ ิเ_เ เช่น ิ-เคเ ิ-เนเ ิ-เรเ ิ-เลเ ิ-เอเ เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในภาษาญี่ปุ่น เช่น นิเซเอะ จิดซุริเสเรียวกุ เป็นต้น

- สายอักขระ ิเ_ _เ เช่น ิเซนเ ิเอนเ เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในภาษาญี่ปุ่น เช่น เคฮิเอนเงกิ โยบิเซนเคียะ เป็นต้น

- สายอักขระ ูโ_เ เช่น ู-โคเ ู-โทเ ู-โรเ เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในภาษาญี่ปุ่น เช่น เซนชูโชเตน เป็นต้น

- สายอักขระ ◌ิ_ึม เช่น ริซึม ลิซึม ริทึม เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในภาษาญี่ปุ่น เช่น คะมิซึมิยะ โคะอิซึมิ เป็นต้น

- สายอักขระ ◌ี ◌ิ_เ_ เช่น ◌ีซิเอ ◌ีลิเน ◌ีลิเว ◌ีดิเต เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในภาษาฝรั่งเศส เช่น ตาปีซิเยร์ เป็นต้น

- สายอักขระ อ_เ_อ เช่น อกเตอ อนเดอ อนเนอ อมเมอ เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในภาษาอื่น เช่น ภาษาฝรั่งเศส (ชองเตอลู อองเลอองส์) และภาษาญี่ปุ่น (เองเคอิ)

- สายอักขระ ◌ั_เ_ อ เช่น ◌ังเกอ ◌ัตเตอ ◌ันเดอ ◌ันเนอ ◌ัมเปอ ◌ัลเลอ เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างอื่นๆ เช่น ภาษาญี่ปุ่น (ซังเคอิ นันเซอิ) และภาษาฝรั่งเศส (กัทเธอรีน ตัสเตอแว็ง)

4.7.3 สายอักขระคำทับศัพท์ภาษาญี่ปุ่น

ในการวิเคราะห์สายอักขระคำทับศัพท์ภาษาญี่ปุ่น ซึ่งผู้วิจัยวิเคราะห์ด้วยสายอักขระเฉพาะขนาด 2-5 แกรม ทำให้พบว่าคำทับศัพท์ญี่ปุ่น มีการใช้อักขรวิธีของภาษาและข้อสังเกตของภาษาจากสายอักขระเฉพาะ ซึ่งสิ่งเหล่านี้จะสะท้อนให้เห็นลักษณะของคำทับศัพท์ภาษาญี่ปุ่น ดังนี้

1. <u>สายอักขระ ทซึ และ ทสึ</u> สายอักขระนี้สะท้อนให้เห็นการใช้พยัญชนะของภาษาญี่ปุ่นร่วมกับเสียง อึ โดยทับศัพท์มาจาก ttsu และ tsu ซึ่งจากเกณฑ์ทับศัพท์ ttsu และ tsu จะถอดเสียงเป็น ตสึ แต่อาจนิยมถอดเสียงเพี้ยนเป็น ทสึ หรือ ทซึ ได้

2. <u>ข้อสังเกตของการสร้างคำทับศัพท์ภาษาญี่ปุ่น</u> จากการที่ภาษาญี่ปุ่นมีสระเดี่ยวน้อยกว่าคำไทย คือ มี 10 สระ ได้แก่ อะ อา เอะ เอ อิ อี โอะ โอ อุ อู ในขณะที่คำไทยมี 18 สระ ได้แก่ อะ อา อิ อี อึ อือ อุ อู เอะ เอ แอะ แอ โอะ โอ เอาะ ออ เออะ เออ ดังนั้น ทำให้รูปแบบภาษาญี่ปุ่นอาจมีสระเดี่ยว 10 ตัวนี้ต่อเนื่องกันค่อนข้างมากกว่าภาษาอื่น เป็นเหตุให้พบสายอักขระเหล่านี้ในภาษาญี่ปุ่นค่อนข้างมาก ในขณะที่พบในภาษาอื่นน้อยหรืออาจไม่พบเลย โดยสายอักขระเหล่านี้ได้แก่

- สายอักขระ ะโ_เ เช่น ะโปเ ะโรเ ะโอเ เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในคำภาษาอังกฤษ เช่น นโปเลียน อะโรเมติกส์ เป็นต้น

- สายอักขระ ◌ู_า ◌ิ เช่น ◌ูมาริ ◌ูนาชิริ ◌ูราอิ เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในภาษาอังกฤษ เช่น ดูกากิส ยูคาลิปตัส เป็นต้น

- สายอักขระ เ_ะ_ะ เช่น เซะงะ เบะระ เมะคะ เนะงะ เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในคำไทย เช่น ระเกะระกะ เป็นต้น

- สายอักขระ ิทซ และ ิทส เช่น คิทส จิทซ จิทส ชิทซ นิทซ บิทซ มิทซ มิทส ริทโ ริทช ริทซ ริทส อิทโ อิทซ อิทส เป็นต้น ซึ่งเป็นสายอักขระเฉพาะที่ประกอบด้วยสระอิร่วมกับพยัญชนะ 2 ตัวติดกัน แต่สายอักขระนี้ยังพบบ้างใน ภาษาอื่นๆ เช่น คำไทย (สนิทสนม) คำภาษาอังกฤษ (วิทส์เลอร์ ฮิทซ์ เฮิทซ์)

- สายอักขระ ินง เช่น คินง มินง ยินง อินง เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในคำไทย เช่น หินงอก เป็นต้น

- สายอักขระ า_ุ_ะ เช่น าบุระ าอุตะ ากุระ าขุสะ เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในคำไทย เช่น สาธุสะ เป็นต้น

- สายอักขระ ุเ_ี เช่น ุเกี ุเคี ุเมี เป็นต้น แต่ในสายอักขระนี้ยังพบบ้างในคำไทย เช่น ทุเรียน เป็นต้น

นอกจากนี้ยังพบสายอักขระที่พบเฉพาะในคำทับศัพท์ภาษาญี่ปุ่น โดยมีดังนี้

- สายอักขระ าเ_ะ เช่น าเกะ าเคะ เป็นต้น
- สายอักขระ า_ะโ เช่น -าตะโ –ามะโ เป็นต้น
- สายอักขระ า_ูโ เช่น -าซูโ –ารูโ เป็นต้น
- สายอักขระ ึเ เช่น คึเ ซึเ สึเ เป็นต้น
- สายอักขระ ึโ เช่น กึโ ซึโ สึโ เป็นต้น
- สายอักขระ -ิเ_ะ เช่น ิเกะ ิเคะ –ิเทะ เป็นต้น
- สายอักขระ -ิโ_ะ เช่น ิโกะ ิโยะ ิโดะ ิโนะ เป็นต้น
- สายอักขระ ุเ_ะ เช่น ุเกะ ุเคะ เป็นต้น
- สายอักขระ ุเ_โ เช่น ุเคโ ุเรโ ุเอโ เป็นต้น
- สายอักขระ ุโ_เ เช่น –ุโยเ –ุโรเ เป็นต้น
- สายอักขระ ุีเ เช่น ุรีเ ุวีเ เป็นต้น
- สายอักขระ ุูโ เช่น ุรูโ ุยูโ เป็นต้น
- สายอักขระ ุ_ุเ เช่น ุสุเ ุรุเ ุคุเ ุสุเ เป็นต้น
- สายอักขระ ูเ_ะ เช่น ูเกะ ูเคะ ูเนะ เป็นต้น
- สายอักขระ เ_ ะ_ุ เช่น เกะรุ เคะรุ เนะคุ เมะอุ เอะรุ เป็นต้น
- สายอักขระ ิโ_ ุ เช่น ิโมกุ –ิโรกุ –ิโยอุ –ิโรคุ เป็นต้น
- สายอักขระ ิ_ะ_ู เช่น ิระมู –ิตะมู –ิกะรู เป็นต้น

4.7.4 สายอักขระคำทับศัพท์ภาษาฝรั่งเศส

ในการวิเคราะห์สายอักขระคำทับศัพท์ภาษาฝรั่งเศส ซึ่งผู้วิจัยวิเคราะห์ด้วยสายอักขระเฉพาะขนาด 2-5 แกรม ทำให้พบว่าในการเขียนทับศัพท์ภาษาฝรั่งเศส มีการใช้อักขรวิธี และข้อสังเกตของภาษา ซึ่งสามารถสะท้อนให้เห็นลักษณะของคำทับศัพท์ภาษาฝรั่งเศสได้ แต่สิ่งที่สะท้อนให้เห็นในภาษานี้มักปรากฏให้เห็นน้อยกว่าภาษาอื่นๆ เพราะผู้วิจัยใช้คลังข้อมูลคำทับศัพท์ภาษาฝรั่งเศสน้อยกว่าในภาษาอื่น ซึ่งแสดงไว้ดังนี้

1. <u>การใช้สายอักขระที่มีเสียง ออ</u> ซึ่งเสียง ออ (aur, am, an, aon, em) เป็นเสียงที่พบมากในภาษานี้ แต่ก็พบมากในภาษาอื่นด้วย เช่น ในคำไทย (ขอ พ่อ) คำทับศัพท์ภาษาอังกฤษ (เมนทอล) คำทับศัพท์ภาษาญี่ปุ่น (คอนจิ) แต่ในกรณีที่เสียง ออ นี้ ตามหลังด้วยเสียง อุ เห็นได้จาก สายอักขระ _อ_ุ เช่น ชอนุ ฟอรุ ลอปุ เป็นต้น แต่สายอักขระนี้ยังพบบ้างในภาษาอังกฤษ เช่น เดอลุกซ์ ออกุสต์ เป็นต้น

2. <u>การใช้สายอักขระที่มีเสียง อง</u> นำหน้า ซึ่งเสียง อง (om) เป็นเสียงที่พบมากในภาษานี้แต่พบในภาษาอื่นด้วย เช่น คำไทย (กงจักร) คำทับศัพท์ภาษาญี่ปุ่น (บงโอโดริ) แต่เมื่อเสียง อง ตามด้วยเสียง อง ทำให้เกิดสายอักขระเฉพาะ _ง _ง เช่นคำว่า องบง ดาลองซง บรียองซง ลงฟง เป็นต้น แต่สายอักขระนี้ยังพบบ้างในภาษาอื่น เช่น คำไทย (หงองแหงง) คำภาษาอังกฤษ (ฮ่องกง) และภาษาญี่ปุ่น (เตงบง)

3. <u>ข้อสังเกตของการสร้างคำทับศัพท์ฝรั่งเศส</u> จากการสร้างคำทับศัพท์ภาษาฝรั่งเศสมีการใช้สระที่มีเสียง อู ออง แอ และ โอ มากในภาษา จึงทำให้เกิดสายอักขระเฉพาะที่พบมากในคำทับศัพท์ภาษาฝรั่งเศส ในขณะที่ภาษาอื่นพบน้อย ได้แก่

- สายอักขระ โ_ _ อง เช่น โกลอง โตบอง โบมอง โวบอง เป็นต้น
- สายอักขระ -ูร์แ_ เช่น -ูร์แค -ูร์แน เป็นต้น แต่สายอักขระนี้ยังพบบ้างในภาษาอังกฤษ เช่น มูร์แลนด์ เป็นต้น

4.8 สรุปผลการใช้วิธีสายอักขระเฉพาะ

จากผลการตัดสินภาษาได้ถูกต้องด้วยสายอักขระเฉพาะในแต่ละภาษาเมื่อมองแต่ภาษาหลักที่ใช้จำนวนคลังข้อมูลการฝึกเท่ากัน พบว่าวิธีการใช้สายอักขระเฉพาะมีผลต่อโปรแกรมระบุคำประมาณ 50% คือ คำทับศัพท์ภาษาญี่ปุ่น 54.09% คำไทย 50.58% และคำทับศัพท์ภาษาอังกฤษ 48.71% โดยสายอักขระเฉพาะมีผลต่อการระบุคำทับศัพท์ภาษาญี่ปุ่นมากที่สุด

อาจเป็นเพราะภาษาญี่ปุ่นเป็นภาษาที่พบสายอักขระเฉพาะมากที่สุด และตัดสินภาษาผิดพลาดน้อยที่สุด หรือหากมองในเชิงภาษาศาสตร์ ภาษาญี่ปุ่นเป็นภาษาที่ใช้สายอักขระซ้ำๆ กันมาก ไม่มีการผสมสระมากเท่าภาษาอื่นๆ ส่วนวิธีสายอักขระเฉพาะมีผลต่อการระบุคำไทยมากถึง 50% อาจเป็นเพราะคำไทยถูกตัดสินภาษาด้วยสายอักขระเฉพาะ 1-ตัวถึง 14.14% ซึ่งในภาษาอื่นไม่มีการใช้สายอักขระเฉพาะ 1-ตัวนี้

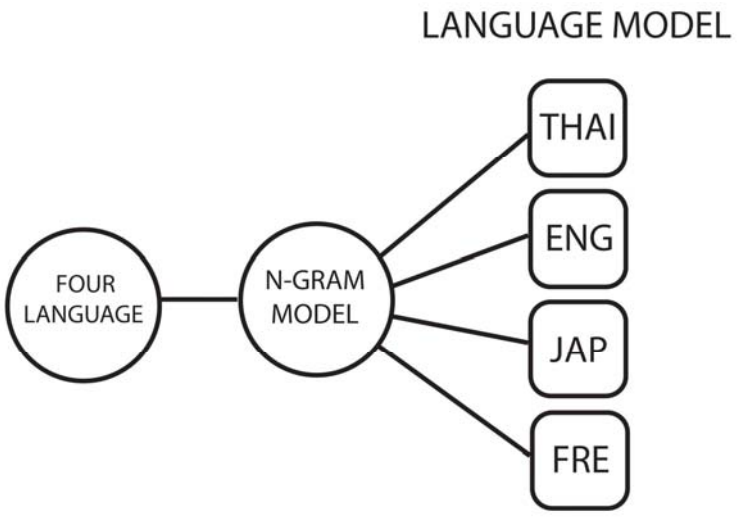
บทที่ 5

การระบุภาษาของคำด้วยแบบจำลองภาษา

ในบทนี้ผู้วิจัยจะกล่าวถึงอีกระบบหนึ่งที่ผู้วิจัยพัฒนาเพื่อใช้ร่วมกับการใช้สายอักขระเฉพาะ คือ การระบุภาษาของคำด้วยแบบจำลองภาษา 2-5 แกรม โดยแสดงถึงภาพรวมของระบบการระบุภาษาโดยใช้แบบจำลองภาษา ผลการทดสอบระบบ ผลเปรียบเทียบการใช้ขนาดของเอ็นแกรมต่างกัน และข้อผิดพลาดที่เกิดขึ้น

5.1 การสร้างแบบจำลองภาษา (Language model)

การสร้างแบบจำลองภาษาในงานวิจัยนี้เป็นเพียงการเก็บค่าทางสถิติของชุดอักขระที่เกิดขึ้นในคลังข้อมูลด้วยวิธีเอ็นแกรมทั้ง 2-5 แกรม โดยจะใช้งานในช่วงเวลาขณะที่ต้องการจะทดสอบระบบ ผลลัพธ์ที่ได้ประกอบด้วยแบบจำลองภาษาทั้งหมด 4 ภาษา คือแบบจำลองคำไทย คำทับศัพท์ภาษาอังกฤษ ภาษาญี่ปุ่นและภาษาฝรั่งเศส เห็นได้จากภาพที่ 5.1 โดยในแบบจำลองแต่ละภาษาก็แยกเป็น 2-แกรม 3-แกรม 4-แกรม และ 5-แกรม เพื่อทดลองเปรียบเทียบว่าวิธีเอ็นแกรมขนาดใดให้ประสิทธิภาพดีกว่า

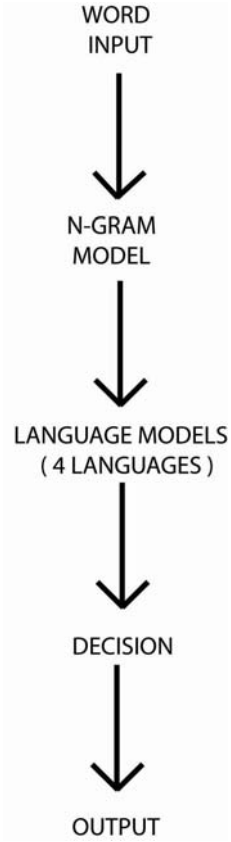


ภาพที่ 5.1 ลักษณะของแบบจำลองภาษาเอ็นแกรม

ต่อไปจะกล่าวถึงโปรแกรมระบุภาษาของคำด้วยแบบจำลองภาษา โดยจะอธิบายโครงสร้างของโปรแกรมระบุภาษาของคำด้วยแบบจำลองภาษาที่พัฒนาและขั้นตอนของโปรแกรมทั้งหมดทีละขั้นตอน

5.2 โปรแกรมระบุภาษาของคำด้วยแบบจำลองภาษา

โปรแกรมระบุภาษาของคำด้วยแบบจำลองภาษา แบ่งออกเป็น 2 ขั้นตอน ได้แก่ 1. ขั้นตอนรับคำเข้า 2. ขั้นตอนระบุภาษาด้วยแบบจำลองภาษา เห็นได้จากภาพที่ 5.2



ภาพที่ 5.2 ขั้นตอนการทำงานของโปรแกรมระบุภาษาของคำด้วยแบบจำลองภาษา

1. ขั้นตอนรับคำเข้า เป็นขั้นตอนที่รับคำเข้ามาแล้วแยกคำๆ นั้น ออกเป็นสายอักขระด้วยเอ็นแกรมทั้ง 2-5 แกรม เช่น คำที่รับเข้ามา คือ คำว่า “อักขระ” เพื่อผ่านระบบเอ็นแกรม จะแยกเป็นแต่ละเอ็นแกรม เห็นได้จากตารางที่ 5.1 ดังนี้

ตารางที่ 5.1 การแยกสายอักขระตามเอ็นแกรม 2-5 แกรม

2-แกรม	3-แกรม	4-แกรม	5-แกรม
อั	อัก	อักข	อักขร
-ก	-กข	-กขร	-กขระ
กข	กขร	กขระ	
ขร	ขระ		
ระ			

2. ขั้นตอนระบุภาษาด้วยแบบจำลองภาษา เป็นขั้นตอนที่นำสายอักขระที่แยกมาคำนวณหาค่าความน่าจะเป็นที่เกิดขึ้นในคลังข้อมูล เช่น เมื่อหาความน่าจะเป็นของคำว่า อักขระ โดยใช้ 4-แกรม เราก็ประมาณค่าโดยนำความน่าจะเป็นของแต่ละสายอักขระมาคูณกันดังนี้ *P*(อ) *P*(◌ั | อ) *P*(ก|อั) *P*(ข | อัก) *P*(ร | ◌ักข) *P*(ะ | กขร) ในขั้นตอนนี้ใช้วิธีเอ็นแกรมตั้งแต่ 2-5 แกรมพร้อมกับวิธีปรับค่าไม่ให้เป็นศูนย์ (smoothing) โดยการเพิ่มค่า 0.00000001 ในกรณีที่ทดลองแล้วไม่พบสายอักขระในข้อมูลการฝึก

ในขั้นตอนระบุภาษาของคำด้วยแบบจำลองภาษา จะทดสอบระบบทีละแกรม คือ คำที่รับเข้ามาต้องตัดสินผ่านภาษาเริ่มจาก 2-แกรม พอตัดสินได้แล้วก็จะนำคำนั้นตัดสินต่อไปใน 3-แกรมอีก ระบบจะทำลักษณะแบบนี้จนถึง 5-แกรม โดยจะทำเช่นนี้กับทุกแบบจำลองภาษา (คำไทย คำทับศัพท์ภาษาอังกฤษ คำทับศัพท์ภาษาญี่ปุ่นและคำทับศัพท์ภาษาฝรั่งเศส)

ตัวอย่างผลการทดลองที่ได้ค่าความน่าจะเป็นจากการใช้แบบจำลอง 4-แกรม เห็นได้จากตารางที่ 5.2 ดังนี้

ตารางที่ 5.2 ตัวอย่างผลการทดลองที่ได้จากแบบจำลองภาษา

ค่าความน่าจะเป็นจาก 4-แกรม				
คำที่ทดสอบจากแบบจำลองภาษา	**คำไทย**	**ภาษาอังกฤษ**	**ภาษาญี่ปุ่น**	**ภาษาฝรั่งเศส**
สินเชื่อ	0.079547	0.00000008	0.0001125	0.0004389
แจ็กเก็ต	0.0004527	0.23293787	0.0046872	0.035714
ทาเกจิ	0.0045454	0.00011531	0.0571847	0.0296052
ซองตราล	0.1150739	0.0244702	0.000025	0.2577250

หลังจากนั้นที่ได้ค่าความน่าจะเป็นของแบบจำลองภาษาแล้วระบบก็จะตัดสินว่าเป็นคำจากภาษาใดโดยเลือกจากแบบจำลองภาษาที่ให้ค่าความน่าจะเป็นสูงที่สุด ซึ่งจากตัวอย่างนี้ คำว่า “สินเชื่อ” จะถูกตัดสินว่าเป็นคำไทย คำว่า “แจ็กเก็ต” ตัดสินว่าเป็นคำทับศัพท์ภาษาอังกฤษ คำว่า “ทาเกจิ” ตัดสินว่าเป็นคำทับศัพท์ภาษาญี่ปุ่น และคำว่า “ซองตราล” ตัดสินว่าเป็นคำทับศัพท์ภาษาฝรั่งเศส โดยหลังจากการทดลองระบุภาษาด้วยโปรแกรมแล้ว ผู้วิจัยเป็นผู้นำคำนั้นตรวจสอบความถูกต้องเองว่าระบบระบุภาษาของคำได้ถูกต้องหรือไม่ เพื่อเป็นตรวจสอบความถูกต้องและทนทานของระบบ

5.3 สรุปผลการทดลอง

จากการทดสอบโปรแกรมระบุภาษาของคำไทยและคำทับศัพท์แสดงให้เห็นผลของระบบระบุภาษาด้วยแบบจำลอง 2-5 แกรม และเนื่องจากระบบนี้จะทดสอบทีละเอ็นแกรม คือ ระบบจะเริ่มตัดสินภาษาคำทุกคำที่แบบจำลอง 2- แกรม ก่อน พอระบบทำงานเสร็จ ก็นำคำทุกคำที่ตัดสินภาษาแล้ว ผ่านมายังแบบจำลองภาษา 3-แกรมเพื่อตัดสินภาษาอีกซึ่งจะตัดสินเป็นภาษาที่ต่างกันได้ โดยทำลักษณะแบบนี้จนถึง 5-แกรม ดังนั้น ผลการทดลองที่แสดงในตาราง 5.3 คือ ค่าความถูกต้อง ที่แยกกันในแต่ละขนาดของเอ็นแกรมนั้น ผลการทดสอบที่ได้จึงให้ทีละแกรมแยกจากกัน เห็นได้จาก ตารางที่ 5.3 โดยผลการทดสอบระบบที่ได้คือ ค่าความถูกต้อง ซึ่งหมายถึง ระบบสามารถระบุภาษาของคำได้ถูกต้อง

ตารางที่ 5.3 ผลการตัดสินภาษาของคำได้ถูกต้องด้วยแบบจำลอง 2-5 แกรม
โดยผลการทดลอง คือ จำนวนคำที่ระบุภาษาได้ถูกต้อง ในวงเล็บคือจำนวนเปอร์เซ็นต์

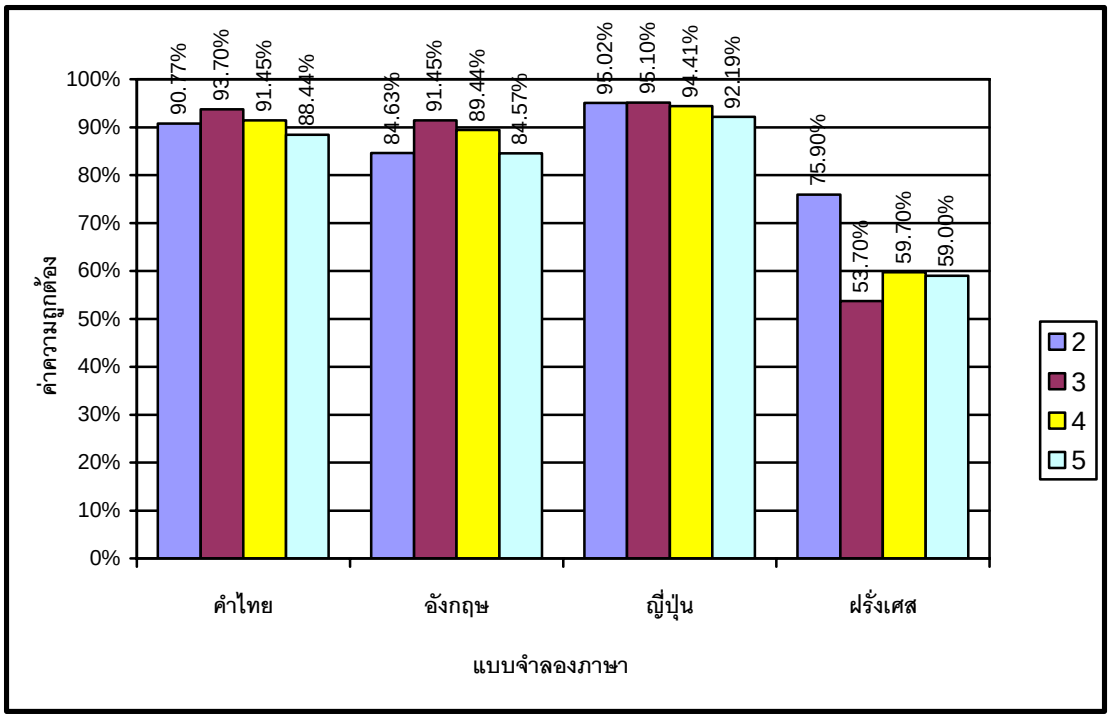
ภาษา / วิธี		จำนวนการระบุภาษาของคำได้ถูกต้อง			
		จำนวนภาษาละ 10000 คำ			1000 คำ
		ไทย	อังกฤษ	ญี่ปุ่น	ฝรั่งเศส
แบบจำลองภาษา 2-5 แกรม	2	9077 (90.77%)	8463 (84.63%)	9502 (95.02%)	759 (75.90%)
	3	9370 (93.70%)	9145 (91.45%)	9510 (95.10%)	537 (53.70%)
	4	9145 (91.45%)	8944 (89.44%)	9441 (94.41%)	597 (59.70%)
	5	8844 (88.44%)	8457 (84.57%)	9219 (92.19%)	590 (59%)

จากตารางที่ 5.3 แสดงให้เห็นว่า เมื่อระบุภาษาของคำแต่ละภาษาด้วยแบบจำลองภาษา 3-แกรม ให้ค่าความถูกต้องดีที่สุด ยกเว้นเมื่อระบุภาษาของคำทับศัพท์ภาษาฝรั่งเศส ส่วนการระบุภาษาด้วย 5-แกรมให้ผลแย่ที่สุด แต่ในทางกลับกันเมื่อระบุคำทับศัพท์ภาษาฝรั่งเศสให้ผลดีที่สุด

นอกจากนี้ ผลการทดลองของแต่ละภาษาที่มีลักษณะคล้ายกัน คือยิ่งใช้แกรมสูงผลจะต่ำลงอาจเป็นเพราะในแต่ละภาษาไทย อังกฤษ ญี่ปุ่น จะเดาตัวอักษรที่สามได้ต้องเห็นมาก่อนสองตัวอักษร ดังนั้นแบบจำลอง 3-แกรม จึงได้ผลมากที่สุดในทุกภาษา แต่เมื่อเพิ่มเป็น 4-แกรม ผลถูกต้องลดลง อาจเป็นเพราะเมื่อขยายขนาดเอ็นแกรมมากขึ้น ทำให้จำนวนปริมาณเอ็นแกรมที่พบในข้อมูลการฝึกไม่มากพอทำให้ผลไม่ดีนัก ซึ่งยืนยันได้จากภาษาฝรั่งเศสซึ่งได้ผลไม่ดีเลยในทุกแกรมเพราะข้อมูลการฝึกมีน้อยไป ส่วนภาษาญี่ปุ่น เห็นได้ว่า ภาษาญี่ปุ่นในแต่ละแกรมให้ผลสูงกว่าภาษาอื่นทั้งหมด โดยเมื่อใช้ 2 และ 3-แกรม จะสูงกว่าภาษาอื่นมาก อาจเป็นเพราะ ลักษณะของภาษาญี่ปุ่น มีการใช้สระเดี่ยวๆ ติดกันมาก และมีเพียงสระ อะ อิ อุ เอะ โอะ เท่านั้น ทำให้เพียงแค่ใช้แบบจำลองเอ็นแกรมสั้นๆ ก็สามารถระบุภาษาได้ดี และสังเกตว่าผลจากแบบจำลอง 2-4 แกรมในภาษาญี่ปุ่นไม่ต่างกันมาก คือ 95.02% 95.10% และ 94.41% แสดงว่าเห็นตัวอักษรมาก่อน 1 ตัวก็เดาได้ง่ายกว่า ส่วนภาษาอังกฤษให้ผลต่ำกว่าภาษาอื่นทั้งหมด อาจเป็นเพราะลักษณะของภาษาอังกฤษมีการใช้ภาษาที่มีการสร้างคำทั้งคำสั้นและคำยาวๆ ไม่เหมือนกับคำไทยที่มีพื้นฐานของภาษาเป็นคำโดดหรือคำสั้นๆ หรือภาษาญี่ปุ่นที่ใช้แต่คำสั้นๆ

นอกจากนี้เรายังสามารถเปรียบเทียบผลของการใช้แบบจำลองภาษาต่างได้ ดังนี้

แผนภูมิที่ 5.3 การเปรียบเทียบการใช้แบบจำลองของแต่ละภาษา



จากแผนภูมิที่ 5.3 แสดงการเปรียบเทียบการใช้แบบจำลองของแต่ละภาษา สังเกตเห็นว่า ระบบระบุภาษาของคำใช้แบบจำลองภาษาขนาด 3-แกรม ได้ค่าความถูกต้องสูงที่สุดและค่าความถูกต้องเริ่มต่ำลงใน 4 และ 5-แกรม ตามลำดับ ยกเว้นการระบุคำทับศัพท์ภาษาฝรั่งเศสจะให้ค่าความถูกต้องสูงที่สุดเมื่อระบุภาษาด้วย 2-แกรม นอกจากนี้ จากการเปรียบเทียบการใช้ขนาดของเอ็นแกรมต่างๆกัน ในกรณีที่ใช้คลังข้อมูลเท่ากัน คือ พิจารณาเฉพาะคำไทย คำทับศัพท์ภาษาอังกฤษและญี่ปุ่น ก็อาจสรุปได้ว่า ขนาดของเอ็นแกรมมีผลต่อการตรวจสอบลักษณะเฉพาะของคำ (sparse data) โดยเมื่อขนาดของเอ็นแกรมมากกว่า 3 โอกาสที่จะเจอลักษณะเฉพาะของคำที่ทดสอบในคลังข้อมูลการฝึกก็ยิ่งน้อยลงทำให้ระบุภาษาผิดพลาด ดังนั้นยิ่งขนาดของเอ็นแกรมมากขึ้นผลก็ยิ่งแย่ลง สังเกตได้จากแผนภูมิแสดงการเปรียบเทียบขนาดของเอ็นแกรม เมื่อขนาดเอ็นแกรมเท่ากับ 5 ค่าความถูกต้องจะน้อยลงอย่างเห็นได้ชัดในการระบุทุกภาษา และในกรณีของคำทับศัพท์ภาษาฝรั่งเศสซึ่งมีจำนวนข้อมูลการฝึกน้อยมาก ผลจากเอ็นแกรมตั้งแต่ 3-5แกรมจะต่ำกว่า 2-แกรม อย่างเห็นได้ชัด

5.4 ปัญหาการตัดสินคำผิดพลาดที่เกิดขึ้นจากแบบจำลองภาษา

จากผลการทดสอบระบบ พบว่ามีการระบุภาษาผิดพลาดจากการใช้แบบจำลองภาษา 2-5 ตัว ซึ่งแสดงไว้ในตาราง 5.4 ดังนี้ โดยแสดงผลที่ตัดสินผิดพลาด เป็นจำนวนคำและเปอร์เซ็นต์เทียบกับข้อมูลทั้งหมด

ตารางที่ 5.4 ผลการตัดสินภาษาของคำผิดพลาดจากการใช้แบบจำลองภาษา 2-5 แกรม

วิธี	คำที่ทดสอบ	ตัดสินผิดเป็น				
		ไทย	อังกฤษ	ญี่ปุ่น	ฝรั่งเศส	รวม
แบบจำลอง 2แกรม	ไทย	0	464 (4.64%)	264 (2.64%)	195 (1.95%)	923 (9.23)
	อังกฤษ	362 (3.62%)	0	239 (2.39%)	936 (9.36%)	1537 (15.37%)
	ญี่ปุ่น	171 (1.71%)	239 (2.39%)	0	88 (0.88%)	498 (4.98%)
	ฝรั่งเศส	33 (3.30%)	196(19.60%)	12 (1.20%)	0	241(24.10%)
รวม		566 (8.63%)	899 (26.63%)	515 (6.23%)	1219 (12.19%)	3199 (53.68%)
แบบจำลอง 3แกรม	ไทย	0	353 (3.53%)	220 (2.20%)	57 (0.57%)	630 (6.30%)
	อังกฤษ	289 (2.89%)	0	218 (2.18%)	348 (3.48%)	855 (8.55%)
	ญี่ปุ่น	158 (1.58%)	294 (2.94%)	0	38 (0.38%)	490 (4.90%)
	ฝรั่งเศส	58 (5.80%)	384(38.40%)	21 (2.1%)	0	463 (46.3%)
รวม		505 (10.27%)	1031 (44.77%)	459 (6.48%)	443 (4.43%)	2438 (66.05%)
แบบจำลอง 4แกรม	ไทย	0	391 (3.91%)	308 (3.08%)	156 (1.56%)	855 (8.55%)
	อังกฤษ	262 (2.62%)	0	340 (3.40%)	454 (4.54%)	1056 (10.56%)
	ญี่ปุ่น	144 (1.44%)	338 (3.38%)	0	77 (0.77%)	559 (5.59%)
	ฝรั่งเศส	46 (4.60%)	315 (31.5%)	42 (4.20%)	0	403(40.30%)
รวม		452 (8.66%)	1044 (38.79%)	690 (10.68%)	687 (6.87%)	2873 (65%)
แบบจำลอง 5แกรม	ไทย	0	473 (4.73%)	409 (4.09%)	274(2.74%)	1156 (11.56%)
	อังกฤษ	348 (3.48%)	0	517 (5.17%)	678 (6.78%)	1543 (15.43%)
	ญี่ปุ่น	218 (2.18%)	432 (4.32%)	0	131 (1.31%)	781 (7.81%)
	ฝรั่งเศส	49 (4.90%)	299(29.90%)	62 (6.20%)	0	410 (41%)
รวม		615 (10.56%)	1204 (38.95%)	988 (15.46%)	1083 (10.83%)	3890 (75.80%)

จากตารางที่ 5.4 เมื่อดูผลรวมการระบุภาษาผิดพลาดจากแบบจำลองภาษาในแต่ละแกรม (แนวตั้ง) ยกเว้นในคำทับศัพท์ภาษาฝรั่งเศสเพราะใช้ข้อมูลการฝึกไม่เท่ากัน พบว่าในการระบุคำทับศัพท์ภาษาอังกฤษด้วยแบบจำลองภาษา 2-5 แกรม ให้ข้อผิดพลาดมากที่สุด คือ 15.37% 8.55% 10.56% และ 15.43% ตามลำดับ และภาษาที่ถูกตัดสินผิดพลาดน้อยที่สุด คือ คำทับศัพท์ภาษาญี่ปุ่น 4.98% 4.90% 5.59% และ 7.81% ตามลำดับ และเมื่อดูผลรวมของภาษาที่ระบุผิดพลาดในแต่ละแกรม (แนวนอน) พบว่าแบบจำลองภาษาในแต่ละแกรมส่วนใหญ่มักระบุผิดว่าเป็นคำทับศัพท์ภาษาอังกฤษมากที่สุด คือ 26.63% 44.77% 38.79% และ 38.95% ตามลำดับ

นอกจากนี้เมื่อเปรียบเทียบกันระหว่างแต่ละแกรม ยกเว้นคำทับศัพท์ภาษาฝรั่งเศส พบว่า เมื่อระบุภาษาทุกภาษาด้วยแบบจำลอง 5-แกรม ให้ผลผิดพลาดมากที่สุด คือ คำไทย (11.56%) คำทับศัพท์ภาษาอังกฤษ (15.43%) และคำทับศัพท์ภาษาญี่ปุ่น (7.81%) จึงสรุปได้ว่ายิ่งใช้ขนาดเอ็นแกรมยิ่งมากผลผิดพลาดก็ยิ่งมากขึ้นด้วย ส่วนการระบุคำทับศัพท์ภาษาฝรั่งเศสด้วย 3-แกรมให้ผลผิดพลาดมากที่สุด คือ 46.3% และเมื่อเปรียบเทียบข้อผิดพลาดกับที่ระบุในภาษาอื่นๆ พบว่าในคำทับศัพท์ภาษาฝรั่งเศสให้ข้อผิดพลาดมากที่สุด อาจเป็นเพราะคลังข้อมูลการฝึกจำนวนน้อยกว่าภาษาอื่น

และตารางที่ 5.4 ในแต่ละภาษายังแสดงให้เห็นข้อผิดพลาดในการตัดสินภาษาผิดเป็นภาษาอื่น ซึ่งจากตารางจะเห็นว่าในสามภาษาหลัก ไทย อังกฤษ และญี่ปุ่น (ไม่พิจารณาฝรั่งเศสเพราะเป็นความต่างที่มาจากขนาดข้อมูลเป็นสำคัญ) คำไทยและภาษาอังกฤษจะตัดสินผิดเป็นภาษาญี่ปุ่นน้อยที่สุดใน 2-3 แกรม และหากเทียบกับภาษาอื่นแล้ว ภาษาญี่ปุ่นถูกตัดสินภาษาผิดน้อยที่สุด แสดงว่ามีความสับสนระหว่างภาษาไทยและภาษาอังกฤษมากกว่า จึงสรุปได้ว่าสายอักขระ 2-3 ตัวของภาษาญี่ปุ่นมีลักษณะที่ต่างจากภาษาไทยและภาษาอังกฤษ แต่พอเป็น 4-5 แกรม เริ่มไม่ชัดเจนแล้ว อาจเป็นเพราะยิ่งแกรมสูงข้อมูลการฝึกไม่เพียงพอ จึงไม่ได้ภาพความสับสนระหว่างภาษาแบบที่พบในแบบจำลอง 2-5แกรม และจากสาเหตุข้อมูลการฝึกไม่เพียงพอ จึงเป็นเหตุผลว่าทำไมคำภาษาอังกฤษถูกตัดสินผิดพลาดว่าเป็นคำไทยใน 2-3 แกรม แต่ใน 4-5 แกรม คำอังกฤษถูกตัดสินผิดว่าเป็นคำญี่ปุ่น

ในส่วนของความสับสนระหว่างภาษาไทยและภาษาอังกฤษ คำไทยมักตัดสินภาษาผิดว่าเป็นภาษาอังกฤษใน 2-5 แกรม เช่น คำว่า ภูเก็ต คอแร้ง จามจุรี น้ำพุ และคำภาษาอังกฤษมักตัดสินผิดว่าเป็นคำไทยใน 2-3 แกรม เช่น คำว่า วิตามิน ละติน แกรนิต เพราะลักษณะของภาษาอาจมีความคล้ายคลึงกัน เนื่องจากการเขียนทับศัพท์ภาษาอังกฤษในปัจจุบัน อาจมีการเขียนคำอังกฤษให้ตรงกับสำเนียงไทยหรือออกเสียงให้ตรงกับระบบเสียงของภาษาไทย เช่น มี

การใช้ตัวอักษร สระ วรรณยุกต์ที่เหมือนกับคำไทย ทำให้คำไทยกับอังกฤษจึงมีการเขียนที่คล้ายกัน ซึ่งจะต่างกับภาษาญี่ปุ่นที่ต่างกันชัดเจน คือมีแต่การใช้คำสั้นๆเท่านั้น

อย่างไรก็ตาม ภาษาญี่ปุ่นถูกตัดสินผิดพลาดน้อยสุด แต่ที่ตัดสินผิดพลาดส่วนใหญ่มักจะตัดสินผิดว่าเป็นคำภาษาอังกฤษใน 2-5 แกรม เช่นคำว่า โคโมง ราเมง เซนเบอิ คานอิชิ อาจเป็นเพราะขนาดของคลังข้อมูลการฝึกไม่เพียงพอต่อการแสดงลักษณะของภาษาญี่ปุ่นให้ชัดเจน

5.5 สรุปการใช้แบบจำลองภาษาเอ็นแกรม

เหตุผลที่ระบบที่ใช้แบบจำลองเอ็นแกรมตัดสินผิดพลาดเป็นเพราะคลังข้อมูลอาจไม่เพียงพอต่อการแสดงลักษณะของภาษา หรือแต่ละภาษามีความคล้ายคลึงกันมากจนอาจต้องใช้คลังข้อมูลมากกว่านี้ อย่างไรก็ตาม จากการทดสอบระบบระบุภาษาของคำด้วยแบบจำลองภาษาเอ็นแกรม ได้ผลการทดสอบมากกว่า 90% จึงสรุปได้ว่า เมื่อเทียบกับระบบสายอักขระเฉพาะ วิธีการใช้แบบจำลองภาษาเอ็นแกรมมีประสิทธิภาพและความทนทานของระบบดีกว่า เพราะระบบสายอักขระเฉพาะได้ผลความถูกต้องน้อยกว่า คือ คำไทย 50.58% คำทับศัพท์ภาษาอังกฤษ 48.71% คำทับศัพท์ภาษาญี่ปุ่น 54.09% และคำทับศัพท์ภาษาฝรั่งเศส 20.40% ในขณะที่ระบบที่ใช้แบบจำลอง 3-แกรม ได้ผลความถูกต้องมากกว่า คือ คำไทย 93.70% คำทับศัพท์ภาษาอังกฤษ 91.45% คำทับศัพท์ภาษาญี่ปุ่น 95.10% และคำทับศัพท์ภาษาฝรั่งเศส 53.70% แต่ระบบเอ็นแกรมนี้ มีข้อเสียคือ ใช้เวลาในการประมาณผลช้ากว่า ผู้วิจัยจึงนำทั้ง 2 ระบบผสมผสานกันเป็นระบบระบุภาษาของคำด้วยการใช้สายอักขระเฉพาะและแบบจำลองภาษาซึ่งจะกล่าวต่อไปในบทที่ 6

บทที่ 6

โปรแกรมระบุภาษาของคำ

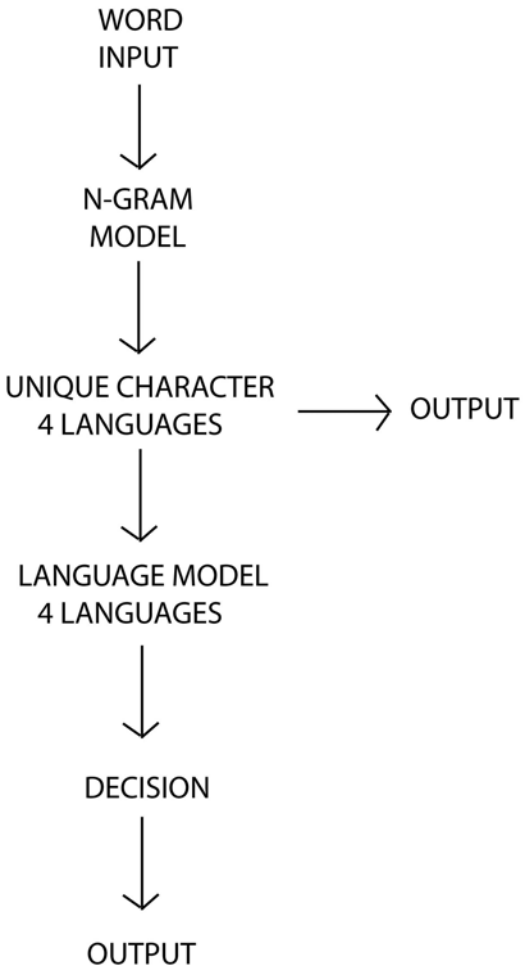
ในบทนี้ผู้วิจัยจะกล่าวถึงระบบที่ผู้วิจัยพัฒนา คือ โปรแกรมระบุภาษาของคำด้วยการใช้สายอักขระเฉพาะ 1-5 ตัวร่วมกับการใช้แบบจำลองภาษาเอ็นแกรม 2-5 แกรม รวมทั้งรายงานและวิเคราะห์ผลการทดสอบระบบที่พัฒนาร่วมกัน โดยจะอธิบายภาพโดยรวมของระบบก่อน เพื่อให้เข้าใจถึงแนวคิดและวิธีการสร้างโปรแกรมระบุภาษาของคำด้วย 2 ระบบนี้

6.1 ภาพโดยรวมของระบบ

โปรแกรมระบุภาษาของคำ คือ โปรแกรมที่ใช้ระบุที่มาภาษาของคำซึ่งพัฒนาโปรแกรมด้วยภาษาเพิร์ล(Perl) ทั้งหมด แนวคิดของการพัฒนาโปรแกรมระบุภาษาของคำนี้ คือ รับข้อมูลคำที่ต้องการตรวจสอบเข้ามาแล้วตรวจหาสายอักขระเฉพาะ โดยนำคำที่ต้องการตรวจสอบเข้ามาแยกเป็นสายอักขระก่อน แล้วจึงนำไปเทียบกับสายอักขระเฉพาะของแต่ละภาษา ถ้าเทียบแล้วตรงกับสายอักขระเฉพาะของภาษาใด ก็จะระบุว่าเป็นคำภาษานั้น แต่ถ้าเทียบแล้วไม่ตรงกัน คือ ระบุคำไม่ได้ว่าเป็นภาษาอะไร จึงนำคำที่ต้องการตรวจสอบนั้นหาความน่าจะเป็นภาษาหนึ่งๆ โดยเทียบกับแบบจำลองภาษาในแต่ละภาษา เหตุผลที่เลือกพัฒนาระบบในลักษณะนี้เพราะ การระบุภาษาด้วยสายอักขระเฉพาะจะทำงานได้เร็วกว่าการใช้แบบจำลองเอ็นแกรมมาก และสามารถระบุภาษาได้ถูกต้องแล้วประมาณ 50% โดยมีข้อผิดพลาดประมาณ 5% (ดูผลการทดลองบทที่ 4) จึงได้กำหนดให้เป็นส่วนแรกของโปรแกรมที่พัฒนา ในกรณีที่ไม่สามารถตัดสินจากสายอักขระเฉพาะได้ว่าเป็นภาษาใด จึงจะใช้แบบจำลองภาษาเป็นตัวตัดสินว่าคำที่รับเข้ามานั้นใกล้เคียงกับภาษาใดมากที่สุด ซึ่งในการทดลองนี้ก็จะทดลองกับแบบจำลองภาษาตั้งแต่ 2-5 แกรมเพื่อยืนยันอีกครั้งว่าควรใช้แบบจำลองภาษาที่เอ็นแกรมใด และสำหรับแนวคิดและภาพรวมของวิธีการหาสายอักขระเฉพาะ และการสร้างแบบจำลองภาษาของแต่ละภาษา ผู้วิจัยได้นำเสนอและอธิบายไว้แล้วในบทที่ 4 และ 5 ดังนั้นผู้วิจัยจึงต้องการแสดงขั้นตอนของระบบต่อไป

6.2 โปรแกรมระบุภาษาของคำ

โปรแกรมระบุภาษาของคำ แบ่งออกเป็น 3 ขั้นตอน ได้แก่ 1. ขั้นตอนรับคำเข้า 2. ขั้นตอนเทียบกับสายอักขระเฉพาะและ 3. ขั้นตอนระบุภาษาด้วยแบบจำลองภาษา เห็นได้จากภาพที่ 6.1



ภาพที่ 6.1 ขั้นตอนการทำงานของโปรแกรมระบุภาษาของคำ

1. ขั้นตอนรับคำเข้า เป็นขั้นตอนที่รับคำเข้ามาแล้วแยกคำๆ นั้น ออกเป็นสายอักขระด้วยเอ็นแกรมทั้ง 1-5 แกรม เช่น คำที่รับเข้ามา คือ คำว่า “อักขระ” เพื่อผ่านระบบเอ็นแกรมจะแยกเป็นแต่ละเอ็นแกรม

2. ขั้นตอนเทียบกับสายอักขระเฉพาะ หลังจากที่แยกสายอักขระเป็นเอ็นแกรมต่างๆ แล้ว ขั้นตอนต่อไปเป็นการนำสายอักขระที่แยกด้วยเอ็นแกรมแล้ว มาเทียบกับสายอักขระเฉพาะทั้ง 4 ภาษา โดยแยกเทียบกันทั้ง 1-5-แกรม ตามแต่ละเอ็นแกรมว่าตรงกันหรือไม่ หากสายอักขระที่นำมาทดสอบตรงกับสายอักขระเฉพาะของภาษาใดภาษาหนึ่งเท่านั้น ก็ถือว่าคำนั้นเป็นภาษาตามสายอักขระเฉพาะนั้นทันที โดยไม่จำเป็นต้องผ่านขั้นตอนเปรียบเทียบกับแบบจำลองภาษา แต่หากไม่มีสายอักขระที่ตรงกับสายอักขระเฉพาะของภาษาใดๆ เลย ก็ให้นำคำนั้นเทียบกับแบบจำลองภาษาทั้ง 4 ภาษาในขั้นตอนต่อไป แต่หากพบว่าคำนั้นมีสายอักขระตรงกับสายอักขระเฉพาะของภาษาต่างๆ มากกว่า 1 ภาษา ทำให้ยังตัดสินไม่ได้ว่าเป็นคำจากภาษาใดก็จะเปรียบเทียบกับแบบจำลองภาษาเหล่านั้น ตัวอย่างเช่น ในคำว่า “อักขระ” เมื่อทดลองด้วย 4-แกรม อาจจะพบ สายอักขระ “กขระ” ตรงกับสายอักขระเฉพาะภาษาญี่ปุ่น และพบสายอักขระ “อักข“ ตรงกับสายอักขระเฉพาะคำไทย ทำให้ตัดสินไม่ได้ว่าเป็นคำไทยหรือมาจากภาษาญี่ปุ่น ดังนั้นระบบจะเปรียบเทียบคำนั้นกับแบบจำลองภาษาญี่ปุ่นและแบบจำลองคำไทยเท่านั้น เพื่อตัดสินว่าน่าจะเป็นคำในแบบจำลองใดมากกว่ากัน

ในขั้นตอนการเปรียบเทียบกับสายอักขระเฉพาะนี้ ระบบที่ผู้วิจัยพัฒนาจะทำงานเพียงครั้งเดียว เช่นเดียวกับระบบระบุภาษาด้วยสายอักขระเฉพาะ 1-5 ตัว ในบทที่ 4 และหากว่าระบบนำสายอักขระที่ทดสอบเปรียบเทียบกับสายอักขระเฉพาะจนถึง 5-แกรมแล้วยังไม่พบว่าตรงกัน ระบบก็จะทดลองต่อไปในขั้นตอนระบุภาษาด้วยแบบจำลองภาษา

3. ขั้นตอนระบุภาษาด้วยแบบจำลองภาษา เป็นขั้นตอนสุดท้ายสำหรับระบุภาษา โดยจะใช้ในกรณีที่สายอักขระที่แยกด้วยเอ็นแกรมไม่ตรงกับสายอักขระเฉพาะใดๆเลย หรือสายอักขระที่ทดลองตรงกับสายอักขระเฉพาะมากกว่า 1 ภาษา ในขั้นตอนนี้จะทดลองใช้วิธีเอ็นแกรม 2-5 แกรม เช่นเดียวกับระบบระบุภาษาด้วยแบบจำลองภาษาในบทที่ 5

หลังจากนั้นที่ได้ค่าความน่าจะเป็นของแบบจำลองภาษาของภาษาต่างๆ แล้วระบบก็จะเลือกผลจากภาษาที่มีค่าความน่าจะเป็นสูงที่สุดเป็นคำตอบที่ถูกต้อง หลังจากการทดลองระบุภาษาด้วยโปรแกรมแล้ว ผู้วิจัยเป็นผู้นำคำนั้นตรวจสอบความถูกต้องเองว่าระบบระบุภาษาของคำได้ถูกต้องหรือไม่ เพื่อเป็นตรวจสอบประสิทธิภาพของระบบ

6.3 สรุปผลการทดลอง

จากการทดลองโปรแกรมระบุภาษาของคำทับศัพท์และคำไทย แสดงให้เห็นผลของระบบระบุภาษาด้วยการสังเกตจากสายอักขระเฉพาะ (1-5 สาย) และแบบจำลองภาษาทุกภาษาด้วยเอ็นแกรม (2-5 แกรม) โดยหลังจากที่ผ่านการระบุภาษาของคำด้วยสายอักขระเฉพาะแล้ว ซึ่งเป็นขั้นตอนที่ 2 ของระบบ พบว่าสายอักขระเฉพาะสามารถแก้ไขปัญหาระบุคำได้ประมาณ 50% ยังมีคำเหลืออีกประมาณ 50% ที่ระบบการใช้สายอักขระเฉพาะไม่สามารถตัดสินภาษาได้ จึงส่งมายังขั้นตอนต่อไป คือการระบุภาษาของคำด้วยแบบจำลองภาษา ในขั้นตอนนี้ระบบจะทำงานแยกกันในแต่ละเอ็นแกรม คือ คำที่ยังไม่ได้ตัดสินอีก 50% ก็จะผ่านมายัง แบบจำลองภาษา 2-แกรมทั้งหมด พอระบบทำงานเสร็จ ก็นำคำที่ยังไม่ได้ติดสินอันเดียวกันนี้ 50% ผ่านมายังแบบจำลองภาษา 3-แกรมอีก โดยทำลักษณะแบบนี้จนถึง 5-แกรม ดังนั้น ผลการทดลองที่แสดงในตารางคือ ค่าความถูกต้องที่แยกกันในแต่ละขนาดของเอ็นแกรม หากต้องการจะดูผลของทั้ง 2 ระบบรวมกัน จะต้องนำผลที่ได้จากสายอักขระเฉพาะในช่อง “รวม” ซึ่งเป็นผลทดสอบที่รวมการใช้สายอักขระเฉพาะทั้งหมด 1-5 ตัว มารวมกับผลจากแบบจำลองภาษาทีละแกรม เพราะในการทำงานของระบบการใช้สายอักขระเฉพาะจะเป็นการทำงานเพียงครั้งเดียว ดังนั้นหากคิดผลการทดลองของทั้งระบบ จะต้องนำผลจากการใช้สายอักขระทั้ง 1-5 ตัว มารวมกันก่อน

ตัวอย่างการดูผลของระบบโดยรวม เช่น หากเราต้องการวิเคราะห์ผลการระบุภาษาของคำไทย เราก็นำ ผลการทดสอบจากสายอักขระเฉพาะรวมตั้งแต่ 1-5 ตัว 50.58% + ผลทดสอบจากแบบจำลองภาษา 2-แกรม 39.83% คำตอบที่ได้ คือ 90.41% (ในช่อง ”รวมถูก”) ซึ่งเป็นผลโดยรวมของระบบระบุหลังจากผ่านสายอักขระเฉพาะและแบบจำลองภาษา 2-แกรม หรือหากต้องการรู้ผลโดยรวมเมื่อใช้แบบจำลองภาษา 3-แกรม เราก็นำผลจากสายอักขระเฉพาะรวมตั้งแต่ 1-5 ตัว 50.58% + ผลทดสอบจากแบบจำลอง 3-แกรม 41.33% ผลที่ได้คือ 91.91% ซึ่งผลโดยรวมทั้งหมดแต่ละเอ็นแกรม ผู้วิจัยได้รวมไว้ในช่อง “รวมถูก” และแยกแต่ละภาษาและเอ็นแกรมไว้ เห็นได้จากตาราง 6.1

ตารางที่ 6.1 ผลการตัดสินภาษาของคำได้ถูกต้องด้วยโปรแกรมระบุคำ (สายอักขระเฉพาะ 1-5 ตัว และแบบจำลอง 2-5 แกรม) โดยผลการทดลอง คือ จำนวนคำที่ระบุภาษาได้ถูกต้อง (เปอร์เซ็นต์)

ภาษา / วิธี		ผลการระบุภาษาของคำได้ถูกต้อง											
		จำนวนคำที่ทดสอบ 10,000 คำ									1,000 คำ		
		คำไทย			ภาษาอังกฤษ			ภาษาญี่ปุ่น			ภาษาฝรั่งเศส		
		ถูก		ผิด	ถูก		ผิด	ถูก		ผิด	ถูก		ผิด
ผลจากสายอักขระเฉพาะ	1	1414 (14.14%)		0	0		2	0		0	0		0
	2	756 (7.56%)		2	122 (1.22%)		14	160 (1.6%)		7	0		3
	3	2104 (21.04%)		274	2018 (20.18%)		264	3392 (33.92%)		186	57(5.7%)		122
	4	651 (6.51%)		164	1769 (17.69%)		277	1683 (16.83%)		168	101 (10.1%)		99
	5	133 (1.33%)		23	962 (9.62%)		81	174 (1.74%)		38	46 (4.60%)		34
รวม		5058 (50.58%)		463	4871 (48.71%)		638	5409 (54.09%)		399	204 (20.40%)		258
		ถูก	รวมถูก	ผิด	ถูก	รวมถูก	ผิด	ถูก	รวมถูก	ผิด	ถูก	รวมถูก	ผิด
ผลจากแบบจำลองภาษา	2	3983	9041 (90.41 %)	496	3728	8599 (85.99 %)	763	3903	9312 (93.12 %)	289	410	614 (61.40 %)	128
	3	4133	9191 (91.91 %)	346	4036	8907 (89.07 %)	455	3898	9307 (93.07 %)	294	286	490 (49%)	252
	4	4012	9070 (90.7 %)	467	3876	8747 (87.47 %)	615	3833	9242 (92.42 %)	359	319	523 (52.30 %)	219
	5	3849	8907 (89.07 %)	630	3569	8440 (84.4 %)	922	3720	9129 (91.29 %)	472	333	537 (53.70 %)	205

จากผลการทดลองเมื่อเปรียบเทียบกันในแต่ละภาษา เห็นว่า ระบบระบุภาษาของคำภาษาญี่ปุ่นให้ผลดีที่สุดในทุกๆแกรม คือ 93.12% 93.07% 92.42% และ 91.29% และระบบระบุภาษาของคำภาษาอังกฤษให้ผลแย่ที่สุด ในทุกๆ แกรม คือ 85.09% 89.07% 87.47% และ 84.4% และเมื่อพิจารณาในแต่ละภาษา จะพบว่าในภาษาไทย อังกฤษและญี่ปุ่น ระบบที่เลือกใช้แบบจำลองภาษา 3-แกรมจะให้ผลดีที่สุด

นอกจากนี้จากผลการทดลองซึ่งเป็นระบบที่ประยุกต์นำวิธีสายอักขระเฉพาะและแบบจำลองภาษาเข้าด้วยกัน เมื่อนำไปเปรียบเทียบกับระบบการใช้สายอักขระเฉพาะ (บทที่4) และระบบการใช้แบบจำลองภาษา (บทที่5) ไว้ในตารางที่ 6.2

ตารางที่ 6.2 ผลการเปรียบเทียบค่าความถูกต้องของระบบสายอักขระเฉพาะ ระบบแบบจำลองภาษา และระบบที่ประยุกต์ทั้ง 2 ระบบ โดยระบบสายอักขระเฉพาะ คือ UN1-5 ระบบแบบจำลองภาษา คือ LM 2-5 และ ระบบที่ประยุกต์ทั้ง 2 ระบบ คือ UN-LM

	จำนวน 10000 คำ									จำนวน 1000 คำ		
ภาษา	คำไทย			ภาษาอังกฤษ			ภาษาญี่ปุ่น			ภาษาฝรั่งเศส		
ระบบ	UN1-5	LM 2-5	UN+LM	UN1-5	LM 2-5	UN+LM	UN1-5	LM 2-5	UN+LM	UN1-5	LM 2-5	UN+LM
ผลการทดลอง	50.58%	90.77%	90.41%	48.71%	84.63%	85.99%	54.09%	95.02%	93.12%	20.40%	75.5%	61.40%
		93.70%	91.91%		91.45%	89.07%		95.10%	93.07%		53.7%	49%
		91.45%	90.70%		89.44%	87.47%		94.41%	92.42%		59.7%	52.30%
		88.44%	89.07%		84.57%	84.40%		92.19%	91.29%		59%	53.70%

จากการเปรียบเทียบพบว่า ระบบที่ประยุกต์ทั้ง 2 ระบบรวมกัน (UN+LM) ดีกว่าระบบการใช้สายอักขระเฉพาะ (UN) อย่างเห็นได้ชัด เพราะ ผลจากระบบสายอักขระเฉพาะประมาณ 50% แต่ ผลจากระบบที่ประยุกต์ทั้ง 2 ระบบ (UN+LM) ให้ผลมากกว่า 90%

ส่วนกรณีเมื่อนำระบบที่ประยุกต์ทั้ง 2 ระบบรวมกัน ไปเปรียบเทียบกับระบบที่ใช้แบบจำลองภาษาเอ็นแกรม (LM) พบว่าระบบที่ใช้แบบจำลองเอ็นแกรมได้ผลดีกว่าในทุกๆ แกรม ยกเว้นแค่ในการระบุคำทับศัพท์ภาษาอังกฤษด้วย 2-แกรม ผลจากระบบประยุกต์ดีกว่าเล็กน้อย

ดังนั้น ผู้วิจัยจึงเสนอ 2 แนวทางในการเลือกใช้ระบบคือ อาจกลับไปใช้เฉพาะวิธีเอ็นแกรมจะดีกว่าระบบที่ประยุกต์ทั้ง 2 ระบบ และปรับปรุงระบบเอ็นแกรมให้ดีขึ้น เพื่อให้ผลการทดลองใกล้เคียง 100% มากที่สุด อาจทำได้โดยการเพิ่มคลังข้อมูลการฝึกให้มากขึ้นหรือใช้ข้อมูลการฝึกที่มีการกระจายข้อมูลมากขึ้น แต่อาจมีข้อเสียคือ การประมาณผลของระบบช้ามากขึ้น

ส่วนทางเลือกที่สอง คือ อาจใช้ระบบประยุกต์ทั้ง 2 ระบบเหมือนเดิม เพียงแค่ปรับปรุงระบบให้ดีขึ้น ในการปรับปรุงส่วนของระบบสายอักขระเฉพาะ คือ สร้างสายอักขระเฉพาะให้ครอบคลุมทั้งภาษา โดยการเพิ่มคลังข้อมูลการฝึกให้มากขึ้น เพื่อให้ผลผิดพลาดที่เกิดขึ้นน้อยที่สุด ส่วนการปรับปรุงระบบแบบจำลองภาษา คือ อาจเพิ่มขนาดของข้อมูลการฝึก ซึ่งจะกล่าวไว้ในส่วนของข้อเสนอแนะ

บทที่ 7

สรุปผลการวิจัย อภิปรายผลและข้อเสนอแนะ

วิทยานิพนธ์นี้ได้นำเสนองานเกี่ยวกับการระบุภาษาของคำไทย คำทับศัพท์ภาษาอังกฤษ คำทับศัพท์ภาษาญี่ปุ่นและคำทับศัพท์ภาษาฝรั่งเศส โดยสรุปผลการวิจัยทั้งหมดไว้ดังนี้

7.1 สรุปผลการวิจัย

เนื่องจากในปัจจุบันมีการใช้คำทับศัพท์ภาษาต่างประเทศปะปนกับคำไทยจำนวนมาก หากเราต้องการพัฒนาโปรแกรมทางด้านการค้นคืนสารสนเทศข้ามภาษา (Cross - Language Information Retrieval) เช่น โปรแกรมถอดอักษรคำทับศัพท์กลับจากภาษาไทยเป็นภาษาต้นฉบับ (backward transliteration) โดยก่อนที่จะถอดอักษรกลับเป็นภาษาต้นฉบับ เราจำเป็นต้องรู้ว่าคำเดิมนั้นเป็นคำในภาษาอะไร จึงต้องมีการระบุภาษาของคำในข้อความ ก่อนที่จะนำไปใช้กฎการถอดอักษร (transliteration) ในแต่ละภาษา นอกจากนี้ ยังพบว่าเคยมีผู้วิจัยทำงานในลักษณะแบบนี้ เช่น การใช้แบบจำลองภาษาเอ็นแกรม และการใช้สายอักขระเฉพาะ โดยผู้วิจัยเห็นว่าแบบจำลองภาษาเอ็นแกรมให้ผลถูกต้องสูงและเรียบง่ายมากที่สุด และวิธีการใช้สายอักขระเฉพาะสามารถประยุกต์ใช้กับงานแบบนี้ได้

ดังนั้น วัตถุประสงค์ที่ต้องการจะทำ คือ ต้องการหาสายอักขระเฉพาะสำหรับใช้ในการระบุภาษาของคำโดยใช้คลังข้อมูลคำไทย คำทับศัพท์ภาษาอังกฤษ ญี่ปุ่นและภาษาฝรั่งเศส และพัฒนาระบบการระบุภาษาของคำไทยและคำทับศัพท์ภาษาต่างประเทศโดยใช้แบบจำลองเอ็นแกรมขนาด 1-5 แกรม เพื่อทดสอบประสิทธิภาพของเอ็นแกรมที่ต่างกัน

ในการดำเนินการวิจัย ผู้วิจัยเริ่มจากเก็บคลังข้อมูลภาษาต่างๆก่อน ในคำไทย คำทับศัพท์ภาษาอังกฤษ และภาษาญี่ปุ่น เก็บภาษาละ 10,000 คำ ส่วนคำทับศัพท์ภาษาฝรั่งเศสเก็บเพียง 1,000 คำ โดยผู้วิจัยได้แบ่งส่วนของคลังข้อมูลสำหรับทำข้อมูลการฝึกภาษาละ 80% ไว้สำหรับใช้ในการสร้างสายอักขระเฉพาะ 1-5 สายในแต่ละภาษา และสำหรับสร้างแบบจำลองภาษาแต่ละภาษา ส่วนอีก 20% แบ่งเป็นคลังข้อมูลทดสอบ เหตุที่แบ่งส่วนข้อมูลเพราะต้องการให้ได้ผลที่ไม่บังเอิญขึ้นกับข้อมูลที่ใช้ในการฝึกและทดสอบ

ในขั้นตอนของการทดลองระบบระบุภาษาของคำ คือ รับข้อมูล (คำ) เข้ามาแล้วตรวจหาสายอักขระเฉพาะในแต่ละภาษาถ้าตรงกันกับสายอักขระเฉพาะใดก็ตัดสินว่าเป็นภาษานั้น แต่

ถ้ายังตัดสินภาษาไม่ได้ จึงหาความน่าจะเป็นของภาษาหนึ่งๆ โดยเทียบกับแบบจำลองภาษาต่างๆ และตัดสินผลจากแบบจำลองภาษาที่ให้ค่าความน่าจะเป็นที่สูงที่สุด

หลังจากการทดลองผู้วิจัยเป็นผู้นำคำที่ถูกตัดสินภาษานั้นมาตรวจสอบความถูกต้องเองว่าระบบระบุภาษาของคำได้ถูกต้องหรือไม่ เพื่อเป็นการตรวจสอบความทนทานของระบบ

7.2 อภิปรายผลการวิจัย

ผลการทดลองของ แบ่งออกเป็น 3 ระบบ คือ การระบุภาษาของคำด้วยสายอักขระเฉพาะ 1-5 ตัว การระบุภาษาของคำด้วยแบบจำลองภาษา 2-5 แกรม และการระบุภาษาของคำด้วยสายอักขระเฉพาะร่วมกับแบบจำลองภาษา ดังนี้

7.2.1 การทดลองระบุภาษาของคำด้วยสายอักขระเฉพาะ 1-5 ตัว

สรุปได้ว่าทุกภาษาสามารถพบสายอักขระเฉพาะได้ ซึ่งสายอักขระเฉพาะ 1-ตัว พบเฉพาะในคำไทยเท่านั้น และในการทดลอง สามารถนำสายอักขระเฉพาะเหล่านี้ใช้ในการระบุภาษาได้ เห็นได้จาก ในการระบุภาษาของคำไทย คำทับศัพท์ภาษาอังกฤษ และญี่ปุ่น ด้วยสายอักขระเฉพาะ 3-ตัว ให้ผลการระบุภาษาที่ถูกต้องสูงที่สุด คือ คำไทย 21.04% คำทับศัพท์ภาษาอังกฤษ 20.18% และคำทับศัพท์ภาษาญี่ปุ่น 33.92% แต่ในการระบุภาษาของคำทับศัพท์ภาษาฝรั่งเศส ด้วยสายอักขระเฉพาะ 4-ตัว ให้ผลดีที่สุด คือ 10.10% ส่วนการระบุภาษาด้วยสายอักขระเฉพาะ 1-ตัว สามารถใช้ระบุภาษาได้ถูกต้องได้เฉพาะในคำไทยเท่านั้น

และเมื่อดูผลรวมของการระบุภาษาของคำที่ถูกต้องจากการใช้สายอักขระเฉพาะทั้ง 1-5 ตัว แสดงให้เห็นว่า วิธีการใช้สายอักขระเฉพาะมีผลต่อการระบุภาษาของคำประมาณ 50% โดยสายอักขระเฉพาะมีผลต่อการระบุคำทับศัพท์ภาษาญี่ปุ่นมากที่สุด คือ 54.09% รองลงมาคือ คำไทย 50.58% คำทับศัพท์ภาษาอังกฤษ 48.71% คำทับศัพท์ภาษาฝรั่งเศส 20.40% อาจเป็นเพราะภาษาญี่ปุ่นเป็นภาษาที่พบสายอักขระเฉพาะมากที่สุด หรือ เป็นภาษาใช้สายอักขระซ้ำๆ กันมาก ไม่มีการผสมสระมากเท่าภาษาอื่นๆ

นอกจากนี้ ยังสังเกตได้ว่า ขนาดของคลังข้อมูลมีผลต่อการใช้สายอักขระเฉพาะด้วย เห็นได้จากการระบุภาษาในคำทับศัพท์ภาษาฝรั่งเศส เมื่อใช้สายอักขระเฉพาะ 1-5 ตัว ให้ผลถูกต้องแค่ 20.40% และผลผิดพลาดถึง 25.60% ซึ่งมากกว่าภาษาอื่นมาก อาจเป็นเพราะ หาก

ใช้จำนวนคลังข้อมูลการฝึกน้อย สายอักขระเฉพาะที่ได้ก็จะน้อย ทำให้สายอักขระเฉพาะที่ได้ไม่ครอบคลุมทั้งภาษา

7.2.2 การทดลองระบุภาษาของคำด้วยแบบจำลองภาษา

เมื่อระบุภาษาของคำแต่ละภาษาด้วยแบบจำลองภาษา 3-แกรม ให้ค่าความถูกต้องดีที่สุด ยกเว้นเมื่อระบุภาษาของคำทับศัพท์ภาษาฝรั่งเศส ส่วนการระบุภาษาด้วย 5-แกรมให้ผลแย่ที่สุด แต่ในทางกลับกันเมื่อระบุคำทับศัพท์ภาษาฝรั่งเศสให้ผลดีที่สุด เหตุผลที่ค่าความถูกต้องแตกต่างกัน อาจเป็นเพราะลักษณะของขนาดข้อมูลการฝึก เห็นได้จากการทดลองระบุคำทับศัพท์ภาษาฝรั่งเศสให้ผลการทดลองน้อยที่สุดเมื่อเทียบกับภาษาอื่น คือ สูงสุดแค่ 75.90% ใน 2-แกรม จึงสรุปได้ว่าขนาดของคลังข้อมูลมีผลต่อระบบเอ็นแกรม

และจากการเปรียบเทียบผลการทดลองจากการใช้แบบจำลองของแต่ละภาษา สังเกตเห็นว่า ระบบระบุภาษาของคำใช้แบบจำลองภาษาขนาด 3-แกรม ได้ค่าความถูกต้องสูงที่สุดและค่าความถูกต้องเริ่มต่ำลงใน 4 และ 5-แกรม ตามลำดับ ยกเว้นการระบุคำทับศัพท์ภาษาฝรั่งเศสจะให้ค่าความถูกต้องสูงที่สุดเมื่อระบุภาษาด้วย 2-แกรม

จากการเปรียบเทียบการใช้ขนาดของเอ็นแกรมต่าง ๆกัน ในกรณีที่ใช้คลังข้อมูลเท่ากัน คือ พิจารณาเฉพาะคำไทย คำทับศัพท์ภาษาอังกฤษและญี่ปุ่น จึงสรุปได้ว่า ขนาดของเอ็นแกรมมีผลต่อการตรวจสอบลักษณะเฉพาะของคำ โดยเมื่อขนาดของเอ็นแกรมมากกว่า 3 โอกาสที่จะเจอลักษณะเฉพาะของคำที่ทดสอบในคลังข้อมูลการฝึกก็ยิ่งน้อยลงทำให้ระบุภาษาผิดพลาด ดังนั้น ยิ่งขนาดของเอ็นแกรมมากขึ้นผลก็ยิ่งแย่ เห็นได้จากผลการทดลอง เมื่อใช้ขนาดเอ็นแกรมเท่ากับ 5 ค่าความถูกต้องจะน้อยลงอย่างเห็นได้ชัด และในกรณีของคำทับศัพท์ภาษาฝรั่งเศสซึ่งมีจำนวนข้อมูลการฝึกน้อยมาก ผลจากเอ็นแกรมตั้งแต่ 3-5 แกรมจะต่ำกว่า 2-แกรมอย่างเห็นได้ชัด

นอกจากนี้ เมื่อนำระบบเอ็นแกรมไปเทียบกับระบบสายอักขระเฉพาะ พบว่าวิธีการใช้แบบจำลองภาษาเอ็นแกรมมีประสิทธิภาพและความทนทานของระบบดีกว่า เพราะระบบเอ็นแกรมให้ผลความถูกต้องประมาณ 90% แต่ระบบสายอักขระเฉพาะได้ผลความถูกต้องประมาณ 50% คือ คำไทย 50.58% คำทับศัพท์ภาษาอังกฤษ 48.71% คำทับศัพท์ภาษาญี่ปุ่น 54.09% และคำทับศัพท์ภาษาฝรั่งเศส 20.40% แต่ระบบเอ็นแกรมนี้ มีข้อเสียคือ ใช้เวลาในการประมาณผลช้ากว่า

7.2.3 การทดลองระบบระบุภาษาของคำด้วยสายอักขระเฉพาะร่วมกับแบบจำลองภาษา

จากการทดลองทำให้สรุปได้ว่า ระบบที่พัฒนาเป็นไปตามเป้าหมาย คือให้ผลค่าความถูกต้องมากกว่า 90% โดยเฉพาะใน การระบุภาษาของคำไทยด้วย 3-แกรม ได้ถูกต้อง 91.91% และในการระบุคำทับศัพท์ภาษาญี่ปุ่นด้วย 3-แกรม ได้ถูกต้อง 93.07% แต่เมื่อเทียบระบบนี้กับระบบที่ใช้สายอักขระเฉพาะ พบว่า ผลจากระบบนี้ดีกว่าระบบสายอักขระเฉพาะมาก เพราะให้ผลการทดลองประมาณ 90% แต่ระบบสายอักขระเฉพาะให้ผลแค่ประมาณ 50% และเมื่อนำระบบนี้ไปเทียบกับระบบเอ็นแกรม พบว่า ผลจากระบบนี้ให้ผลแย่กว่าระบบเอ็นแกรมในทุกๆ แกรม ยกเว้น เมื่อระบุคำทับศัพท์ภาษาอังกฤษใน 2-แกรม ให้ผลดีกว่า

7.3 ข้อเสนอแนะ

หลังจากการวิเคราะห์ข้อผิดพลาดของระบบแล้ว ข้อผิดพลาดเกิดจากทั้งการใช้สายอักขระเฉพาะและแบบจำลองภาษา แม้ว่าข้อผิดพลาดจะเกิดขึ้นน้อยเมื่อเทียบกับคำที่ระบุได้ถูกต้องทั้งหมดในระบบ แต่ผู้วิจัยต้องการให้ระบบมีประสิทธิภาพมากที่สุด ผู้วิจัยจึงเสนอแนวทางในการพัฒนาและแก้ไขปัญหาในทีละระบบ ดังนี้

7.3.1 ระบบสายอักขระเฉพาะ

จากผลการทดลอง เป็นที่น่าสังเกตว่าระบบสายอักขระเฉพาะให้ผลถูกต้องแค่ประมาณ 50% และระบบสายอักขระเฉพาะให้ผลผิดพลาดบ้าง แม้จะเป็นส่วนน้อย ผู้วิจัยจึงเสนอแนวทางแก้ไข โดยการเพิ่มขนาดของคลังข้อมูลการฝึกให้มากขึ้น เพื่อใช้สร้างสายอักขระเฉพาะให้ครอบคลุมทั่วทั้งภาษา อย่างไรก็ตาม ในการเพิ่มขนาดคลังข้อมูลเพื่อหาสายอักขระเฉพาะ ผู้วิจัยก็ยังไม่ได้ศึกษาว่า ขนาดของคลังข้อมูลควรอยู่ในระดับเท่าใด ถึงจะเหมาะสมที่สุดสำหรับใช้ในการหาสายอักขระเฉพาะ จึงเป็นแนวทางที่นำไปพัฒนาระบบต่อได้

และเนื่องจากรูปแบบของภาษา (pattern) ที่กล่าวไว้ในระบบ ผู้วิจัยเป็นผู้วิเคราะห์หารูปแบบเอง ซึ่งอาจจะไม่ครอบคลุมทั่วทั้งภาษา จึงเสนอแนวทางที่จะเพิ่มรูปแบบภาษาให้มากขึ้น หรือหารูปแบบภาษาที่พบบ่อยๆ ในสายอักขระได้ โดยเลือกแทนที่ตัวอักษรบางตัวด้วย

วิธี dummy แล้วตรวจดูว่าตรงกับข้อมูลสายอักขระเฉพาะมากน้อยแค่ไหน เช่น ทำให้เราพบว่าในภาษาอังกฤษมีรูปแบบภาษาที่พบบ่อย คือ แ_ส ดังที่เห็นในตัวอย่าง

7.3.2 ระบบเอ็นแกรม

จากผลการทดลอง เป็นที่น่าสังเกตว่าหากต้องการให้ระบบสามารถระบุภาษาของคำไทยและคำทับศัพท์ภาษาต่างประเทศให้ได้ผลถูกต้องมากขึ้น ก็เป็นสิ่งที่เป็นไปได้เพียงแค่เพิ่มเติมในส่วนของข้อมูลการฝึกให้มากขึ้นเพราะจากผลการทดลอง การใช้ข้อมูลการฝึกมากระบบก็สามารถระบุภาษาได้ถูกต้องมากกว่าข้อมูลการฝึกจำนวนน้อย เห็นได้จากเมื่อเทียบกับกรณีการระบุภาษาของคำทับศัพท์ภาษาฝรั่งเศส แต่การเพิ่มขนาดของข้อมูลการฝึกเมื่อเพิ่มถึงจุดหนึ่งแล้ว อาจจะไม่ทำให้ผลลัพธ์ดีขึ้น ดังนั้นผู้วิจัยจึงเสนอให้ทดลองเพิ่มเติม โดยทดลองใช้คลังข้อมูลหลายๆ ขนาด และเพิ่มขนาดคลังข้อมูลการฝึกมากขึ้น เพื่อหาขนาดของคลังข้อมูลที่เหมาะสมที่สุด

และถึงแม้เป็นที่แน่ชัดแล้วว่าขนาดของเอ็นแกรม 3-แกรม ให้ผลการทดลองดีที่สุด แต่ผู้วิจัยหวังว่าจะสามารถพัฒนาระบบได้อีก โดยใช้ขนาดของเอ็นแกรมมากขึ้นเพื่อดูว่าการเปลี่ยนแปลงผลการทดลอง หากใช้ขนาดของเอ็นแกรมมากขึ้น อาจเป็นไปได้ว่า หากใช้คลังข้อมูลการฝึกมากขึ้นและขนาดเอ็นแกรมสูงๆ ผลการทดลองก็อาจจะดีขึ้นได้

7.3.3 ระบบประยุกต์ที่ใช้ระบบสายอักขระเฉพาะร่วมกับแบบจำลองเอ็นแกรม

เนื่องจากผลจากระบบประยุกต์นี้ให้ผลแย่กว่าระบบที่ใช้แบบจำลองภาษาเอ็นแกรม จึงเสนอว่า หลังจากที่แก้ปัญหาระบบที่ใช้สายอักขระเฉพาะและระบบเอ็นแกรม ให้ผลการทดลองถูกต้องมากขึ้นแล้ว จึงน่าจะนำระบบประยุกต์นี้มาประมวลผลอีกครั้งเพื่อดูว่า ผลของระบบประยุกต์นี้ดีกว่าระบบเอ็นแกรมหรือไม่ และให้ทดลองประสิทธิภาพความเร็วในการทำงานเพื่อหาว่า ระบบแบบเอ็นแกรมอย่างเดียวกับระบบแบบผสมทำงานด้วยความเร็วต่างกันอย่างมีนัยยะสำคัญหรือไม่

นอกจากนี้อาจนำระบบที่พัฒนานี้ไปใช้ระบุภาษาของคำทับศัพท์ภาษาต่างประเทศอื่นๆ ได้อีก เพียงแค่ใช้คลังข้อมูลการฝึกของภาษานั้นเท่านั้น

รายการอ้างอิง

ภาษาไทย

คณะอนุกรรมการปรับปรุงศัพท์เทคนิคทางวิศวกรรมไฟฟ้า. 2541. ศัพท์เทคนิควิศวกรรมไฟฟ้าสื่อสาร. พิมพ์ครั้งที่ 4. กรุงเทพฯ: สมาคมวิศวกรรมสถานแห่งประเทศไทยในพระบรมราชูปถัมภ์,

ราชบัณฑิตยสถาน. 2538. พจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ. 2525. พิมพ์ครั้งที่ 5. กรุงเทพฯ: ราชบัณฑิตยสถาน.

ราชบัณฑิตยสถาน. 2535. หลักเกณฑ์การทับศัพท์ ภาษาฝรั่งเศส ภาษาเยอรมัน ภาษาอิตาลี ภาษาสเปน ภาษารัสเซีย ภาษาญี่ปุ่น ภาษาอาหรับ ภาษามลายู. กรุงเทพฯ: ราชบัณฑิตยสถาน.

ราชบัณฑิตยสถาน. 2538. หลักเกณฑ์การทับศัพท์ภาษาอังกฤษ. พิมพ์ครั้งที่ 6. กรุงเทพฯ: ราชบัณฑิตยสถาน.

ศรีไพร ศักดิ์รุ่งพงศากุล. 2544. เทคโนโลยีคอมพิวเตอร์และสารสนเทศ. กรุงเทพฯ: ซีเอ็ดยูเคชั่น.

สมชาย ชัยธนะตระกูล. 2544. รวมศัพท์ญี่ปุ่น. พิมพ์ครั้งที่ 2. กรุงเทพฯ: สำนักพิมพ์เพจพับ - บลิชชิง.

ภาษาอังกฤษ

Aizawa, A. 2003. Linguistic techniques to Improve the performance of automatic text categorization. In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPR2001), pp.307-314. Cited in Peng, Schuurmans, Wang and Huang. Text classification in Asian languages without word segmentation. In Proceedings of The Sixth International Workshop on Information Retrieval with Asian Languages (IRAL 2003). Association for Computational Linguistics, July 7. Sapporo, Japan.

Cavnar, W. and Trenkle, J. 1994. N-gram-based text categorization. In Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR). (11-13 April 1994): pp.161-175. Las Vegas, USA.

Chen, S.F. and Goodman, J.T. 1998. An Empirical study of smoothing techniques for language modeling. In Proceedings of Thirty-Fourth Annual Meeting of the Association for Computational Linguistics, pp.310-318.

Churcher Gavin. 1994. Distinctive character sequences. Personal communication.

Churcher Gavin, Hayes Judith, Johnson Stephen, and Souter Clive. 1994. Bigraph and trigraph models for language identification and character recognition. In Proceedings of 1994 AISB Work-shop on Computational Linguistics for Speech and Handwriting Recognition, April. University of Leeds.

Combrinck, H.P. and Botha, E.C. 1995. Text-based automatic language identification. In Proceedings of the Sixth Annual South African Workshop on Pattern Recognition, November. Rand Afrikaans University.

Dunning, T. 1994. Statistical identification of language. In Technical report CRL MCCS-94-273, Computing Research Lab, March. New Mexcio State University.

Harbeck, S Ohler, U, Noth, E and Niemann, H. 1999. Information theoretic based segments for language identification. In Proceeding of the Workshop on Text, Speech and dialog (TSD'99), pp.189-202. Berlin.

Hayes Judith. 1993. Language recognition using two-and three-letter clusters. Technical Report. school of computer studies, University of Leeds.

Johnson Stephen. 1993. Solving the problem of language recognition. Technical report, school of computer studies, University of Leeds.

Peng, F. and Schuurmans D. 2003. Combing naïve bayes and n-gram language models for text classification. In F. Sebastiani (Eds.): Advances in Information Retrieval: Proceedings of The 25th European Conference on Information Retrieval Research (ECIR03), LNCS 2633: pp.335-350. Pisa, Italy.

Peng F., Schuurmans D. and Wang S. 2003. Language and task independent text categorization with simple language models. In Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL'03), (May 27 - June 1 2003): pp.110-117. Edmonton, Canada.

Peng F., Schuurmans D. and Wang S. 2004. Augmenting Naive Bayes text classifier with statistical language models. Information Retrieval 7 (3-4), pp.317–345.

Peng, F., Schuurmans D., Wang S. and Huang X. 2003. Text classification in Asian languages without word segmentation. In Proceedings of The Sixth International Workshop on Information Retrieval with Asian Languages (IRAL 2003). Association for Computational Linguistics, July 7. Sapporo, Japan.

Piotrowki Michael. 2000. Statistical language identification for NLP-supported full text retrieval. Master's Thesis, April 28.

Ponte, M and Bruce Croft. 1998. A Language modeling approach to information retrieval. In Proceeding of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp.275-281. Australia.

Sibun Penelope and Reynar Jeffrey C. 1996. Language identification: examining the issues. In 5th symposium on document analysis and information retrieval, pp.125-135. Las Vegas, USA.

Teahan, W. and Harper, D. 2001. Using compression-based language models for text categorization. In Proceedings of Workshop on LMIR.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. Journal of Information Retrieval 1(1/2), pp.69-90.