

การประมวลผลภาษาไทย



วิโรจน์ อรุณมานะกุล
ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

คำนำ

หนังสือเล่มนี้มีจุดมุ่งหมายเพื่อให้ความรู้ภาษาไทยและการประมวลผลภาษาสำหรับผู้ที่สนใจงานประมวลผลภาษาไทยเป็นหลัก ผู้อ่านควรมีความรู้พื้นฐานการประมวลผลภาษามาบ้างแล้ว จาการอ่านตำราพื้นฐาน เช่น Speech and Language Processing (Jurafsky and Martin, 2021) หรือ Foundations of Statistical Natural Language Processing (Manning and Schütze, 1999) เพราะจะมีการอ้างอิงศัพท์และโมเดลพื้นฐานรวมถึงวิธีการประมวลผลภาษา เช่น n-gram, decision tree, conditional random field เป็นต้น หนังสือเล่มนี้จะให้ความสำคัญกับการเข้าใจลักษณะภาษาไทยที่มีผลต่อการออกแบบและพัฒนาระบบการประมวลผลต่าง ๆ แม้ว่าหนังสือไม่ได้เน้นการใช้โครงข่ายประสาทแบบต่าง ๆ ในการประมวลผลซึ่งเป็นแนวทางที่นิยมใช้ในปัจจุบัน แต่ผู้เขียนเชื่อว่าความเข้าใจในปัญหาการประมวลผลภาษาไทยยังเป็นพื้นฐานที่จำเป็น เพราะไม่ว่าจะเลือกใช้โมเดลแบบไหนประมวลผลภาษา ปัญหาที่ต้องการแก้ไขยังเป็นปัญหาเดิม แนวทางการทำงานแบบดั้งเดิมแม้ว่าจะมีประสิทธิภาพไม่ดีเท่าโมเดลที่ใช้โครงข่ายประสาท แต่จะช่วยให้เราเห็นและเข้าใจปัญหาในงานประมวลผลต่าง ๆ ชัดเจนกว่า ในขณะที่การใช้แบบจำลองโครงข่ายประสาททำการประมวลผลภาษาจะได้ผลลัพธ์ที่ดีกว่าเมื่อมีปริมาณข้อมูลมากพอ แต่สิ่งที่ขาดไปคือความเข้าใจในกระบวนการภายในของภาษาซึ่งเป็นสิ่งที่นักภาษาศาสตร์สนใจศึกษามาโดยตลอด เปรียบเหมือนการเล่นหมากรุกหรือหมากล้อมกับคอมพิวเตอร์ แม้ว่าคอมพิวเตอร์จะเล่นได้เก่งกว่าเสมอ แต่มนุษย์ก็ยังคงเล่นเกมเหล่านี้อยู่เพราะความสนใจในเกมมากกว่าการเอาชนะ นักภาษาศาสตร์เองก็สนใจอยากเข้าใจหาคำอธิบายกระบวนการทางภาษาที่เกิดขึ้นมากกว่าที่จะสนใจแต่เพียงผลสัมฤทธิ์ว่าคอมพิวเตอร์ทำงานได้ดีมากแค่ไหน

เนื้อหาในหนังสือแบ่งเป็นสามตอน ตอนแรกว่าด้วยการประมวลผลระดับคำซึ่งเป็นระดับพื้นฐาน จากนั้นจะกล่าวถึงความรู้เรื่องคำประสมและการระบุหาคำประสมในภาษาไทยเพราะการประสมคำเป็นวิธีพื้นฐานของการสร้างคำภาษาไทย ชื่อเฉพาะและการรู้จำชื่อเฉพาะภาษาไทยก็เป็นเรื่องพื้นฐานอีกเรื่องหนึ่งในการประมวลผลภาษาไทยที่จะต้องจำแนกข้อมูลส่วนนี้ออกมาให้ได้ก่อนจะประมวลผลในระดับที่สูงขึ้น ตอนที่สองเป็นเรื่องต่อจากระดับคำ เริ่มจากการกำหนดชุดหมวดคำและการกำกับหมวดคำซึ่งเป็นพื้นฐานสำหรับการประมวลผลในระดับวากยสัมพันธ์หรือการรวมคำเป็นหน่วยที่ใหญ่ขึ้น แนวทางการวิเคราะห์ทางวากยสัมพันธ์ที่แตกต่างกันก็มีผลให้มีการกำหนดชุดหมวดคำที่ต่างกันได้ ในที่นี้ ผู้เขียนเลือกใช้ไวยากรณ์ฟิงพาดตามแนว Universal Dependencies เพราะเชื่อว่าเหมาะกับภาษาไทยและยังทำให้เทียบเคียงผลกับการวิเคราะห์ภาษาอื่น ๆ ได้ จากนั้นจะเป็นเรื่องหน่วยปริจเฉทพื้นฐาน เพราะในการประมวลผลข้อความต่อเนื่อง เรามักนิยมแบ่งข้อความต่อเนื่องเป็นหน่วยย่อย ๆ เพื่อใช้ในการประมวลผลมากกว่าจะประมวลผลข้อความทั้งหมดพร้อมกัน นอกจากนี้การประมวลผลข้อความอาจใช้วิธีการที่ไม่จำเป็นต้องวิเคราะห์ตามขั้นตอนการวิเคราะห์ทางภาษาศาสตร์ก็ได้ งานหลายอย่างเลือกใช้วิธีการแปลงหน่วยภาษาเป็นเวกเตอร์เพื่อประมวลผลเลย ในส่วนนี้จึงยกตัวอย่างการสร้างและใช้เวกเตอร์คำด้วย Word2Vec ในตอนสุดท้าย ผู้เขียนจะกล่าวถึงผลกระทบและอนาคตในโลกที่คอมพิวเตอร์มีพัฒนาการความสามารถทางภาษามากขึ้น บทบาทหน้าที่ของนักภาษาจะเปลี่ยนไปอย่างไรบ้าง และเราจะปรับตัวหรือใช้ประโยชน์จากพัฒนาการทางเทคโนโลยีได้อย่างไร

หนังสือเล่มนี้ไม่ได้กล่าวถึงเรื่องภาษาในระดับอักขรวิธี สำหรับผู้ที่สนใจเรื่องอักขรวิธีภาษาไทยสามารถหาความรู้ได้จากหนังสือ “อักขรวิธีไทย และการถอดอักษรระหว่างภาษาไทยและภาษาอังกฤษ” ซึ่งเผยแพร่บนเว็บผู้เขียน เนื้อหาหลายส่วนในหนังสือนี้ได้เคยเผยแพร่เป็นบทความบนบล็อก Medium ให้คนทั่วไปอ่าน หรือเป็น

เอกสารเผยแพร่บนหน้าเว็บของผู้เขียนเอง ผู้เขียนตัดสินใจรวบรวมงานเขียนเหล่านั้นมาเรียบเรียงใหม่และเขียนเพิ่มเติมส่วนที่ยังไม่ได้เขียนเพื่อให้เนื้อหาทั้งหมดมีความต่อเนื่องกัน และผนวกการสาธิตโปรแกรมภาษา Python ที่ผู้เขียนพัฒนาโมดูล TLTK (Thai Language Toolkit) สำหรับใช้สาธิตตัวอย่างงานการประมวลผลภาษาไทยมาไว้ในหนังสือนี้ด้วย ผู้ใช้สามารถเข้าไปดู Google collab ที่ทำไว้และทดลองใช้คำสั่งต่าง ๆ ที่แสดงได้

สุดท้ายนี้ ผู้เขียนหวังว่าผู้อ่านที่สนใจงานประมวลผลภาษาไทยจะได้เข้าใจลักษณะและธรรมชาติของภาษาไทยที่เป็นโจทย์หรือปัญหาสำหรับการประมวลผลภาษาไทยมากขึ้น เพราะแม้ว่าเทคโนโลยีการประมวลผลจะพัฒนาไปในทิศทางที่คอมพิวเตอร์สามารถทำงานและวิเคราะห์งานต่าง ๆ เองได้ง่ายมากขึ้น แต่เทคโนโลยีเหล่านั้นก็เป็นเพียงเครื่องมือหรือวิธีการที่ถูกนำมาใช้เพื่องานประมวลผลภาษาลักษณะเฉพาะของภาษาไทยเองก็ยังคงเป็นโจทย์หรือปัญหาที่ต้องประมวลผล ความเข้าใจในตัวปัญหาหรือลักษณะของภาษาไทยที่ส่งผลให้ระบบทำงานได้ดีหรือไม่ได้ดีจะช่วยให้เราสามารถปรับแต่งหรือเลือกใช้เครื่องมือวิธีการต่าง ๆ ได้อย่างมีหลักหรือคำอธิบายมากขึ้น และที่สำคัญความเข้าใจในการประมวลผลภาษาไทยคือความเข้าใจในความรู้ของมนุษย์เองที่จะตอบคำถามพื้นฐานว่าทำไมเราจึงใช้ภาษาพูดคุยและเข้าใจกันได้

๑ มกราคม ๒๕๖๗

สารบัญ

คำนำ	ii
ตอน ๑ : เริ่มจากคำ.....	7
การตัดคำในภาษาไทย.....	8
คำประสมภาษาไทย.....	27
การสกัดคำประสมภาษาไทย	55
การรู้จำชื่อเฉพาะภาษาไทย.....	84
ตอน ๒ : ต่อจากคำ.....	107
การกำกับหมวดคำภาษาไทย	108
การตัดประโยคภาษาไทย	128
การวิเคราะห์ภาษาไทยด้วยไวยากรณ์ฟังกา.....	144
การแปลงภาษาเป็นเวกเตอร์.....	176
Word2Vec ภาษาไทย.....	196
ตอน ๓ : ภายภาคหน้า.....	206
GPT-3 หนทางสู่ AGI?	208
GPT-4 แสงแรกแห่ง AGI	226
AI กับอนาคตของนักภาษา.....	245

การประมวลผลภาษาไทย

ฝึกเขียนภาษาไทยให้ดีขึ้นด้วย MT	253
ดัชนีคำ.....	264

ตอน ๑ : เริ่มจากคำ

คำเป็นหน่วยพื้นฐานที่สำคัญของภาษาไทย และเป็นปัญหาแรก ๆ ที่ถูกกล่าวถึงในการประมวลผลภาษาไทย เหตุเพราะในภาษาไทยไม่ได้เขียนแยกเป็นคำ ๆ แบบภาษาอื่น ข้อมูลที่คอมพิวเตอร์เห็นหรือรับเข้าจึงเป็นสายอักขระที่ประกอบด้วยคำหลายคำเขียนต่อเนื่องกันไป ในตอนแรกนี้จึงได้กล่าวถึงปัญหาเรื่องคำและการตัดคำ ซึ่งต้องอาศัยความรู้ว่าคำคืออะไร การกำหนดขอบเขตคำเองก็ไม่ใช่เรื่องง่ายสำหรับมนุษย์โดยเฉพาะในกรณีของคำประสมที่มีโครงสร้างและความซับซ้อนได้คล้ายกับวลี นอกจากคำแล้ว ชื่อเฉพาะก็เป็นหน่วยภาษาประเภทหนึ่งที่ต้องระบุขอบเขตและชนิดของชื่อเฉพาะให้ได้ การประมวลผลคำและชื่อเฉพาะจึงเป็นงานพื้นฐานส่วนแรกสำหรับการประมวลผลภาษาไทย

การตัดคำในภาษาไทย¹

ความเป็นมา

การตัดคำภาษาไทยเป็นงานพื้นฐานของการประมวลผลภาษาไทยที่ทำกันมาตั้งแต่แรกเริ่มของการใช้ภาษาไทยบนคอมพิวเตอร์ ภาษาไทยจำเป็นต้องมีการตัดคำด้วยเหตุว่าในการเขียนภาษาไทย เราไม่มีการวรรคระหว่างคำ คอมพิวเตอร์เห็นเป็นสายอักขระต่อเนื่องไปไม่รู้ว่ขอบเขตคำอยู่ที่ใด การตัดคำเมื่อแรกเริ่มเป็นการทำเพื่อช่วยงานพิมพ์ให้คอมพิวเตอร์บดข้อความขึ้นบรรทัดใหม่ให้ถูกต้อง ไม่ตัดแยกส่วนคำระหว่างบรรทัด งานแรก ๆ ออกมาในรูปของการใช้กฎเพื่อระบุขอบเขตว่าควรจะตัดที่ตำแหน่งไหนได้ กฎต่าง ๆ ที่เสนอเป็นเรื่องของพยางค์มากกว่าตัดคำ เช่น ถ้าพบ "เ" จะตัดหลังสระนี้ไม่ได้ เพราะต้องมีพยัญชนะตามมาเสมอ (Thairatananond, 1981)

จากนั้นจึงมีแนวคิดที่ใช้พจนานุกรมช่วยตัด โดยคิดว่าจะต้องมีรายการคำเพื่อให้คอมพิวเตอร์รู้ว่ารูปใดเป็นคำและขอบเขตคำอยู่ที่ไหน แนวคิดพื้นฐานคือเอาคำในพจนานุกรมไปเทียบกับข้อมูลหากพบรูปตรงก็จะตัดได้ ซึ่งก็ทำแบบ longest matching ได้ (สมปรารถนา วิทยานนท์ 2535, รัตติกร และคณะ 2538) คือตัดหาคำให้ยาวที่สุดก่อน แต่ก็จะเห็นว่ามีปัญหา เช่น "ไปหามเหสี" ตัดออกมาเป็น ไป|หาม|เหสี เลยมีการเสนออีกแนวคิดให้ตัดแบบ maximum matching (วิรัช ศรีเลิศล้ำวาณิช

¹ เรียบเรียงจากบล็อกผู้เขียน <https://awirote.medium.com> “การตัดคำภาษาไทย : ความเป็นมา”, “การตัดคำภาษาไทย : ตัดคำอย่างไร”, “การตัดคำภาษาไทย : ประเมินผลอย่างไร”

2536) คือตัดคำให้ได้จำนวนค่าน้อยที่สุด ตัวอย่างนี้จึงได้ "ไป|หา|ม|เหสี" เพราะมีจำนวนคำที่ได้น้อยกว่า

คำถามที่น่าสนใจ คือ ทำไมจึงเลือกการตัดคำที่ได้จำนวนค่าน้อยสุด ทำไมไม่เลือกการตัดคำที่ได้จำนวนคำมากที่สุด คาดว่าเป็นเพราะภาษาไทยมีการสร้างคำจากการประสมคำจำนวนมาก เมื่อเจอคำอย่าง “หน้าต่าง” โอกาสที่สายอักขระนี้จะเป็นคำประสมคำเดียวจะมากกว่าที่จะเป็นคำเดี่ยวสองคำ การเลือกผลที่มีจำนวนค่าน้อยสุดจึงมีโอกาสถูกต้องมากกว่า

ข้อเสียของการตัดคำด้วยพจนานุกรม คือ ถ้าพบรูปคำที่ไม่รู้จักก็ยังไม่ได้บรรจุลงในพจนานุกรมที่ใช้ก็จะตัดคำไม่ได้ ทำให้เกิดปัญหาของ unknown word ซึ่งปัญหานี้ อาจมาจากพจนานุกรมมีรายการคำไม่ครอบคลุมพอ หรือเป็นคำที่เกิดใหม่ เป็นชื่อเฉพาะ เป็นคำทับศัพท์ หรือเป็นคำที่สะกดผิด เป็นต้น นอกจากนี้ปัญหา unknown word ปัญหาพื้นฐานเรื่องความกำกวมก็ยังคงอยู่ โดยเฉพาะเมื่อตัดแล้วได้จำนวนคำเท่ากัน ตัวอย่างที่มีกยกมาใช้ เช่น ตา|ก|ลม กับ ตา|กลม เป็นสองคำ

นอกจากการตัดคำแบบเลือกให้มีค่าน้อยสุดแล้ว ก็มีการใช้วิธีการทางสถิติคือใช้ n-gram ของคำมาช่วยในการแก้ปัญหาความกำกวมในการตัด (Kawtrakul et al. 1995) เช่น หากพบว่ามีข้อมูลจริง นั่ง|ตา|ก|ลม|อยู่ หรือ ทำ|ตา|ก|ลม|แป้ว ก็จะอาศัยบริบทเหล่านี้มาช่วยตัดสินว่าจะเลือกตัดแบบไหนดีกว่า ซึ่งจะทำแบบนี้ได้ ก็จะต้องมีการตัดคำด้วยมือก่อนให้มีจำนวนตัวอย่างของ n-gram ที่มากพอ และเมื่อมีคลังข้อมูลที่ตัดคำเป็นตัวอย่างแล้ว ก็เริ่มมีการใช้ข้อมูลอื่นนอกเหนือจากสถิติจาก n-gram เช่น Meknavin et al. (1997) ใช้ feature-based ตัดคำโดยเรียนรู้จากลักษณะอื่น ๆ ในบริบทเพื่อเลือกการตัดคำที่ดีที่สุดจากทุกรูปแบบที่ได้จากการใช้ maximum matching มาก่อน

นอกจากการตัดคำแบบอาศัยพจนานุกรม บางคนก็ใช้วิธีการอื่นเพื่อเลี่ยงปัญหา unknown word โดยไม่อิงพจนานุกรมเลย เช่น ใช้ TCC (Thai character cluster)

เป็นตัวแยกหน่วยย่อยก่อนจะตัดสินใจว่าขอบเขตคำอยู่ที่ไหน TCC อาศัยแนวคิดว่ามีตัวอักษรบางตัวที่รู้ว่าไม่ควรไปตัดคำ ณ ตำแหน่งนั้น เช่น C๑ จะไม่ตัดหน้าสระอาแน่ๆ C๑ ก็เลยเป็น cluster หนึ่ง ส่วนการตัดสินใจว่า TCC ไหนรวมเป็นคำได้หรือจะตัดคำที่ TCC ไหน ก็อาจใช้วิธีเรียนรู้จากคลังข้อมูลที่ตัดคำไว้ให้ เช่น ใช้ decision tree (Theeramunkong and Usanavasin, 2001)

การแยกหน่วยย่อยก่อนการตัดคำ นอกจากการใช้ TCC ก็มีการใช้รูปพยางค์เป็นหน่วยย่อยด้วย เช่น Aroonmanakun (2002) เลือกตัดพยางค์ก่อนด้วยกฎโครงสร้างพยางค์ ซึ่งแม้แต่การตัดพยางค์ก็ยังมีความเป็นไปได้หลายแบบ จึงต้องใช้ n-gram ของคลังข้อมูลที่ตัดพยางค์แล้วช่วยเลือกการตัดพยางค์ที่ดีที่สุดก่อนจะไปตัดสินใจรวมพยางค์เป็นคำในภายหลัง การใช้คลังข้อมูลที่ตัดพยางค์ด้วยมือมีข้อดีว่าการใช้คลังข้อมูลที่ตัดคำด้วยมือ เพราะการตัดพยางค์ไม่มีปัญหาเรื่องการตัดแล้วไม่สม่ำเสมอและจำนวนข้อมูลภาษาที่เท่ากันก็จะได้จำนวน token ของพยางค์ที่มากกว่าจำนวน token ของคำ

งานตัดคำหลัง ๆ ก็อาศัยคลังข้อมูลภาษาที่มีการตัดคำแล้วและใช้การเรียนรู้ด้วยเครื่องแบบต่าง ๆ เช่น ใช้ CRF (conditional random field), SVM (support vector machine) ล่าสุดก็นิยมใช้ deep learning เช่น KutCum, deepcut, Attacut แต่ทั้งหมดก็ต้องอาศัยคลังข้อมูลที่มีการตัดคำเป็นตัวอย่งให้เรียนก่อน วิธีการเหล่านี้จึงไม่จำเป็นต้องใช้พจนานุกรมได้ เพราะสามารถเริ่มจากหน่วยเล็กที่สุดคือ ตัวอักษรหรือกลุ่มตัวอักษร ก็ได้ หรือจะใช้รูปพยางค์ก็เป็นไปได้

ปัญหาของการใช้คลังข้อมูลที่มีการตัดคำ นอกจากต้องลงแรงในการตัดคำเองจำนวนมากแล้ว ยังมีปัญหาจากการที่แต่ละคนตัดคำไม่เหมือนกัน หรือแม้คน ๆ เดียวกัน ก็อาจตัดคำออกมาไม่เหมือนกัน ทำให้ข้อมูลไม่คงที่แบบที่คาดหวัง ปัญหาตัดคำไม่เหมือนกันนี้ ทำให้ต้องมาตอบคำถามว่าคำคืออะไร ทำไมแต่ละคนจึงกำหนดขอบเขต

คำไม่เหมือนกัน และการตัดคำทำไปเพื่ออะไร ทำอย่างไรจึงจะสร้างคลังข้อมูลขนาดใหญ่ที่มีการตัดคำที่สม่ำเสมอได้

คำคืออะไร

คำถามพื้นฐานของการตัดคำภาษาไทย คือ อะไรคือคำ กำหนดขอบเขตอย่างไร เรามักคิดว่าปัญหาเรื่องตัดคำทำให้ประมวลผลภาษาไทยยาก ไม่เหมือนภาษาอังกฤษที่เขียนแยกเป็นคำ ๆ ให้อ่านแล้ว แต่ความจริง ภาษาอังกฤษก็ยังมีเรื่องของการหาคำที่ประกอบสร้างขึ้นจากคำย่อย ๆ เช่น *ice cream*, *web address*, *distance learning*, *on show*, *by and large*, *kick the bucket* คำจึงเป็นหน่วยที่แม้ในภาษาที่มีการเขียนแบบเว้นวรรคคำก็ไม่ใช่ว่าจะปราศจากปัญหา ตำราไวยากรณ์ไทยมักพูดถึงคำว่า มีแยกเป็นคำมูล คำประสม คำซ้ำ คำซ้อน โดยให้คำมูลเป็นคำที่มีพยางค์เดียวหรือหลายพยางค์ก็ได้แต่เมื่อแยกส่วนแล้วจะไม่มี ความหมายหรือมีความหมายก็ไม่มีเค้าความหมายของคำนั้น เช่น “สามี่” แยกมาแล้ว “สาม” จะมีความหมาย “มี” จะมีความหมาย แต่ความหมายเหล่านั้นไม่มีเค้าสัมพันธ์กับคำนั้น แต่ในกลุ่มคำที่สร้างจากคำมูล เช่น “แม่น้ำ” แยกมาแล้วเป็นสองคำยังเห็นเค้าความเกี่ยวข้องกับคำนั้นได้

แม้คำอธิบายเรื่องคำจะฟังดูเข้าใจได้ไม่ยาก แต่การระบุขอบเขตคำในข้อมูลจริงกลับเป็นเรื่องยากที่คนอาจเห็นต่างกันได้โดยเฉพะกรณีคำประสมที่มาจากคำประสมคำหลายคำหลายระดับ ทำให้เกิดความสับสนระหว่งการเป็นคำประสมหรือวลี (ประเด็นนี้ผู้เขียนจะกล่าวถึงโดยละเอียดอีกครั้งในบทที่ว่าด้วยเรื่องคำประสมภาษาไทย)

การสร้างคำ

การสร้างคำในภาษานั้นทำได้หลากหลายวิธีแล้วแต่ภาษา หลักพื้นฐานคำมาจากการนำหน่วยคำ (morpheme) มา process เป็นคำ (word) หน่วยคำคือหน่วยที่เล็กสุดในภาษาที่มีความหมาย เช่น *un-* เป็น prefix เติมหน้าหน่วยคำอิสระ *happy*

เป็นคำ *unhappy* กระบวนการสร้างคำด้วยการเติม affix นี้อาจเป็นกระบวนการที่เรียกว่า inflection คือสร้างรูปคำต่าง ๆ จากหน่วยศัพท์เดียวกัน เช่น สร้าง *lives, living* จาก *live* หรือเป็นกระบวนการ derivation คือการสร้างหน่วยศัพท์ใหม่จากเดิม เช่น สร้าง *reader, readable* จาก *read* ทั้งสอง process นี้พบมากในภาษาที่เป็น inflectional language ภาษาไทยไม่ได้มีกระบวนการนี้ชัดเจน อาจมองว่ามีบางคำ เช่น *การ ความ ผู้ นัก นว* ที่ทำหน้าที่เป็นเหมือน prefix ได้

การซ้ำคำหรือ reduplication ก็เป็นอีกกระบวนการหนึ่งในการสร้างคำ ในภาษาไทยมีการสร้างคำลักษณะนี้มาก ถ้าเป็นการซ้ำรูปเราเรียกซ้ำซ้ำ เช่น *แก้ว ๆ, แด่งแดง* แต่ถ้าเป็นการซ้ำความหมายเราจะเรียกซ้ำซ้อน เช่น *ชั่วร้าย, โหดเหี้ยม* บางครั้งก็ซ้ำบางส่วน เช่น *หนังกอกหนักใจ, ออกดอกออกผล, กระจายเงินกระจายทอง* คำซ้ำนี้บางตำราก็จัดว่าเป็นประเภทย่อยหนึ่งของการประสมคำ

การประสมคำ

กระบวนการสร้างคำที่พบมากอีกอย่างในภาษาไทยคือ การประสมคำ หรือ compounding คือการนำคำอิสระตั้งแต่สองคำขึ้นไปมาประสมกันแล้วเกิดเป็นคำที่มีความหมายใหม่ที่ต่างจากเดิมหรืออาจมีเค้าความเดิมบ้าง เช่น *แม่น้ำ, ปากกา, ดินปืน, กระจกเงา, ไม่กวาด* ในภาษาอังกฤษก็มีการสร้างคำจากการประสมคำ แต่ส่วนใหญ่จะเขียนติดกันเป็นรูปคำเดียว เช่น *firewood, grandfather, boyfriend*, แต่บางคำก็ยังเห็นรอยต่อจากการเติม hyphen เช่น *dry-clean, washer-dryer* บางคำก็ยังเขียนแยกเป็นสองรูป เช่น *back seat* แต่ถ้าเทียบกับภาษาไทย ซึ่งอาศัยการประสมคำเป็นกลไกหลักในการสร้างคำ คำประสมในภาษาไทยจึงมีใช้มากและซับซ้อนกว่า และเป็นสาเหตุหลักที่ทำให้การตัดคำด้วยคนนั้นมีความไม่คงที่หรือเห็นไม่ตรงกันได้

หนังสือไวยากรณ์ไทยมักยกตัวอย่างคำประสมที่เมื่อรวมกันแล้วมีความหมายใหม่เลย เช่น *ลูกเสือ, พ่อตา* หรือประสมกันแล้วมีความหมายใหม่แต่ยังเห็นเค้าเดิมบ้าง เช่น *น้ำแข็ง, แม่บ้าน* หรือบางคำก็เหมือนจะมีความหมายของคำเดิมทั้งหมด เช่น

น้ำหวาน รถบรรทุก ปัญหาที่พบบ่อยเกิดจากการที่คำประสมสามารถประสมกันหลายชั้นได้ เช่น *คนขับรถ* เป็นคำประสม *รถบรรทุก* เป็นคำประสม *คนขับรถบรรทุก* มาจากการประสมที่เอาคำประสม *รถบรรทุก* มาประสมซ้อน *คนขับรถ* อีกครั้ง

เมื่อจำนวนคำเริ่มมากขึ้น เราก็จะเริ่มลังเลว่าควรให้เป็นคำหรือวลีดีกว่า เช่น *คนขับรถแท็กซี่* *คนขับรถเมล* ในเชิงความหมายก็มีลักษณะเดียวกับ *คนขับรถ* เพราะมีความหมายใหม่ที่เป็นอาชีพหนึ่ง จึงน่าจะให้เป็นคำประสมได้ แต่เมื่อประสมยาวขึ้นเป็น *คนขับรถบรรทุกสิบล้อ* *คนขับรถโดยสารประจำทางปรับอากาศ* ก็จะเริ่มไม่แน่ใจแล้วว่าควรตัดเป็นคำเดียวหรือไม่ คำควรจะยาวได้มากแค่ไหน

ปัญหาลักษณะนี้ทำให้การตัดคำด้วยมืออาจเกิดปัญหา เพราะแต่ละคนอาจมองขอบเขตคำต่างกันไป บางคนก็ยอมให้คำประสมยาวมากได้ บางคนก็คิดว่าไม่ควรจะยาวมากไป การสร้าง BEST corpus ที่เป็นข้อมูลตัดคำไทยด้วยมือ ผู้สร้างจึงยึดเกณฑ์ให้ตัดคำให้สั้นเป็นหลัก อะไรที่ตัดแล้วไม่เสียความหมายไปก็ให้ตัดไว้ก่อน วิธีนี้จะช่วยให้การทำงานของคนหลายคนมีความคงที่ได้มากกว่า ด้วยหลักการนี้ คำว่า *รถโดยสาร* ก็จะตัด *รถ|โดยสาร* *คนขับรถ* ก็จะเป็น *คน|ขับ|รถ* เมื่อยึดหลักนี้ คำประสมที่ไม่ได้มีความหมายต่างจากความหมายรวมของคำย่อยก็จะถูกตัดเป็นหลายคำ *ไม้ถูพื้น* *โต๊ะกินข้าว* เหล่านี้ก็ควรตัดเป็นสามคำไปด้วย เพราะสามารถแยกคำได้แบบเดียวกันและความหมายไม่ได้เปลี่ยนไปมาก ยังคงเป็นไม้สำหรับถูพื้น เป็นโต๊ะสำหรับกินข้าว

Aroonmanakun (2007) ก็เสนอความคิดให้ตัดคำเป็นหน่วยที่เล็กสุดที่ไม่เสียความหมายไปก่อนได้ เหมือนเป็นกระบวนการสร้างคำย่อยก่อน แต่ก็เสนอให้มีกระบวนการหา multi-word unit หรือ lexeme ภายหลังด้วย ในลักษณะเดียวกับที่ภาษาอังกฤษเองก็ยังคงมีการหา multi-word expression ข้อดีของวิธีนี้คือช่วยให้การตัดคำด้วยมือมีความสม่ำเสมอมากกว่า และการตัดคำย่อยก็ไม่ได้มีผลกระทบนักกับงาน NLP อย่าง information retrieval, text classification เพราะคำประสมแม้จะถูกตัดย่อย ก็ยังคงอยู่ในเอกสารนั้นและอยู่ติดกันเสมอ แต่บางงานที่ต้องการใช้

lexeme เช่น machine translation, information extraction, topic modeling ก็อาจต้องมีกระบวนการหา multi-word expression เพื่อรวมคำย่อย ๆ ของคำประสมเป็นหน่วยใหญ่ขึ้นมาก่อน

ควรตัดคำย่อยเสมอไหม

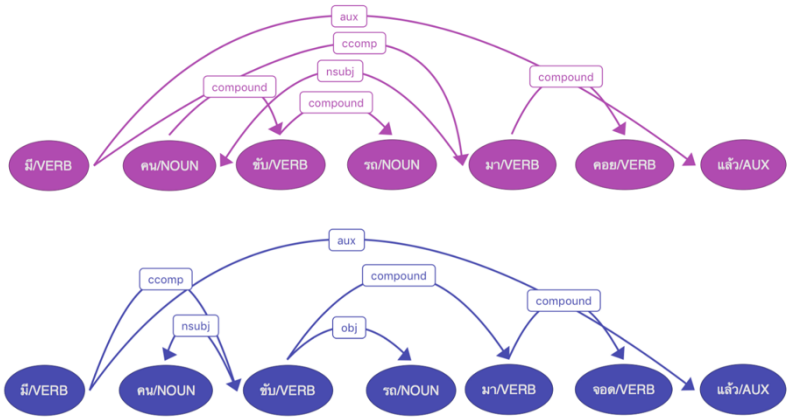
คำตอบขึ้นกับหลายปัจจัย ถ้าเราต้องการสร้างคลังข้อมูลตัดคำด้วยมือในเวลานั้น ๆ ใช้คนจำนวนมากมาช่วย วิธีนี้ก็ช่วยให้คลังข้อมูลที่ได้มีความสม่ำเสมอมากกว่า และถ้างานที่ต้องการนำข้อมูลไปใช้ไม่มีความจำเป็นต้องระบุค่าใหญ่ ก็สามารถใช้คลังข้อมูลที่สร้างด้วยวิธีนี้ได้เลย ข้อดีอีกประการคือ จะไม่มีปัญหาเกี่ยวกับการประมวลผลคำประสมที่ยังไม่รู้จักหรือถูกระบุมาก่อน เพราะการระบุหาคำประสมที่ไม่ใช่คำประสมแท้จะไม่ถูกประมวลผลในขั้นตอนนี้

คำถามสำคัญอีกคำถาม คือ หากตัดคำย่อยแล้ว เราจะแยกความแตกต่างของคำประสมกับวลีหรือประโยคอย่างไร เพราะการตัดสินใจว่ารูปคำที่เห็นเป็นคำประสมเป็นวลี หรือเป็นประโยค เป็นเรื่องที่ต้องอาศัยบริบทในการบอก เช่น *คนขับรถ* อาจเป็นคำประสม หรือเป็นประโยค ก็ได้

มีคนขับรถมาคอยแล้ว คนขับรถ เป็นคำประสมหมายถึง คนคนหนึ่งที่ทำอาชีพขับรถ

มีคนขับรถมาจอดแล้ว คนขับรถ เป็นประโยคซ่อนภายใน หมายถึง คนคนหนึ่งขับรถมา

กรณีเช่นนี้ หากเตรียมข้อมูลโดยตัดคำย่อย ข้อมูลที่สร้างจะปรากฏเป็นคำสามคำเหมือนกัน คือ *คน|ขับ|รถ* วิธีการหนึ่งที่จะช่วยแยกความต่างนี้คือการวิเคราะห์ข้อมูลลงลึกต่อ เช่น ถ้าใช้กรอบการวิเคราะห์ Universal Dependencies วิเคราะห์ประโยคสองอันนี้มีโครงสร้างและความสัมพันธ์ระหว่างคำที่ต่างกัน ดังนี้

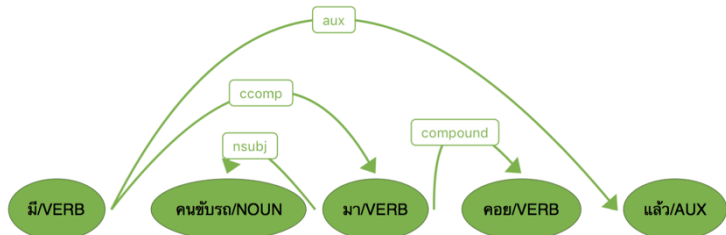


การวิเคราะห์แบบนี้เป็นการผลักร่างงานไปในระดับต่อไป ทำให้การตัดคำเป็นงานที่ง่ายขึ้น โอกาสผิดก็น้อยลง เพราะไม่ต้องมาตัดสินค้าประสมเหล่านี้แล้วว่าจะเป็นคำเดียวหรือหลายคำ เพราะได้ตัดเป็นหลายคำไปแล้ว แต่ภาระก็จะไปอยู่กับกระบวนการต่อไป เช่น ให้ parser แยกความสัมพันธ์แบบที่แสดงความต่างตามรูปออกมา หากทำแบบนี้ งาน parser จะซับซ้อนขึ้น เพราะนอกจากจะต้องวิเคราะห์ทาง syntax ที่หาโครงสร้างประโยคว่าคำต่างๆ มาประกอบกันอย่างไร ยังต้องวิเคราะห์ส่วน morphology คือหาโครงสร้างคำประสมเพิ่มมาด้วย ซึ่งทำให้งานยากขึ้นเพราะคำประสมชนิดหนึ่งที่เราเรียกว่า synthetic compound จะมีลำดับคำเหมือนประโยค คือ noun-verb-noun *ไม้-ถูก-ฟัน* หรือ noun-verb *คน-ใช้*

ถ้าในแผนงานมีความคิดที่จะ implement parser ลักษณะนี้ต่อ ก็สามารถใช้คลังข้อมูลที่ตัดคำย่อยแบบนี้ได้ โดยมองว่าความต่างของการวิเคราะห์จะถูกผลักร่างไปอยู่ที่ในระดับการหา dependency parsered tree แต่ก็ต้องยอมรับว่างานจะซับซ้อนขึ้น แต่หากไม่แน่ใจว่ามีเวลาและนักภาษาศาสตร์ที่จะทำส่วนนี้ได้ และต้องการใช้คลังข้อมูลตัดคำเพื่องานต่าง ๆ เลย กรณีนี้ก็อาจจะต้องตัดคำให้ยาวขึ้นบ้าง มอง

การประมวลผลภาษาไทย

เหมือนว่าการระบุคำประสมเป็นงานลักษณะเดียวกับการระบุ named entity ที่ต้องมีการบอกขอบเขตว่าเริ่มและจบที่ไหน กรณีคำประสม การไม่ตัดคำย่อยเกินไปก็เหมือนเป็นการระบุขอบเขตคำประสมให้กับเครื่อง การตัดคำก็เป็นการตัดคำที่ต้องรวมเอาการแก้ปัญหาคำประสมเหล่านี้ด้วย คือเอาภาระส่วนนี้มาเป็นงานของการตัดคำด้วย ซึ่งความจริงก็เป็นแนวคิดดั้งเดิมของการตัดคำที่ต้องการได้คำที่เป็นหน่วยศัพท์หรือ lexeme



นอกจากนี้ การตัดคำที่แตกต่างกัน ก็ส่งผลต่อการนำข้อมูลไปใช้งานต่อด้วย เช่น หากต้องการสร้าง word2vec จากคลังข้อมูลที่ตัดเป็นคำย่อย เช่น ตัด กลาง คีน เป็นสองคำ เมื่อแปลงเป็น word2vec จะไม่มีเวกเตอร์ของ กลาง คีน แต่มีของคำ กลาง กับ คีน ดังนี้ (ข้อมูลจากการใช้ BEST corpus)

most_similar("กลาง") => ['เหนือ', 'เที่ยง', 'อีสาน', 'ล่าง', 'บาย', 'ใหญ่', 'สี่', 'สาม', 'นี้', 'แต่ละ']

most_similar("คีน") => ['เด็ก', 'วัน', 'วาน', 'เช้า', 'เดือน', 'บาย', 'มือ', 'อาทิตย์', 'พุ่ม', 'คำ']

หากเราไม่รู้อีก่อนว่ามีการตัดคำแบบนี้ ก็อาจสงสัยว่า ทำไม กลาง จึงคล้ายกับ เที่ยง หรือ บ่าย ได้ ผลที่ออกมาแบบนี้ เพราะ กลาง ในคลังข้อมูลนี้มีทั้ง กลาง ที่สัมพันธ์กับทิศ ตำแหน่ง ขนาด และ กลาง ที่เป็นคำเกิดหน้า วัน กับ คีน จึงสัมพันธ์กับ เวลา เที่ยง บ่าย

แต่หากในคลังข้อมูลไม่ได้ตัด *กลางวัน* เป็นสองคำ ก็จะได้เวกเตอร์ของทั้งสามคำ ซึ่งแสดงชุดคำที่สัมพันธ์กันต่างออกไป ดังนี้ (ข้อมูลจากการใช้ Thai National Corpus)

most_similar("กลางวัน") => ['กลางวัน', 'หัวค่ำ', 'บ่าย', 'ดึก', 'เข้ามืด', 'เข้าตู่', 'ค่ำ', 'ตอนเช้า', 'เที่ยง', 'ตอนเย็น']

most_similar("กลาง") => ['ใต้', 'ล่าง', 'ปลาย', 'บน', 'ตรงกลาง', 'ใหญ่', 'ใกล้', 'กึ่งกลาง', 'ริม', 'นอก']

most_similar("คืน") => ['วัน', 'เช้า', 'ปี', 'คราว', 'เดือน', 'งวด', 'กลับบ้าน', 'สัปดาห์', 'คืนชีพ', 'ต้นเงิน']²

การนำคลังข้อมูลตัดคำไปสร้าง word2vec ใช้งานต่อ จึงควรเข้าใจว่าข้อมูลที่ใช้นั้นมีลักษณะแบบไหน และเวกเตอร์ที่ได้จะเป็นแบบใด หรือในทางตรงข้าม ความต้องการที่จะใช้ word2vec ที่แสดงความสัมพันธ์ของคำแบบไหน ก็จะเป็นปัจจัยให้เลือกว่าควรตัดคำแบบใดก่อนจะสร้างข้อมูล word vector

ทำอย่างไรจึงจะตัดคำประสมได้คงที่

หากเลือกแนวทางตัดคำใหญ่ขึ้น คำถามเดิมก็จะกลับมาว่าทำอย่างไร จึงจะทำให้มีการตัดคำด้วยมือแบบสม่ำเสมอ ข้อเสนอคือให้เดินสายกลาง ปัญหาที่เกิดความไม่สม่ำเสมอมักจะมาจากการตัดคำที่ยาวเกินไป ดังนั้น ถ้าระบุขอบเขตว่า คำประสมที่ตัดนั้นไม่ควรยาวมากไป

คำประสมโดยทั่วไปมักประกอบด้วยคำ 2-3 คำ ยกเว้นคำซ้ำคำซ้อนที่อาจยาวกว่านั้นได้ คำประสมทั่วไป เช่น *ไม้ถูพื้น โต๊ะกินข้าว อาหารเย็น สมุดพก* ฯลฯ คำเหล่านี้โอกาสที่จะพบใช้แบบไม่ใช่คำประสมนั้นแม้จะนึกตัวอย่างออกได้ เช่น *เขานั่ง*

² ตัวอย่างข้อมูล word2vec มาจากการทดลองใช้คลังข้อมูล BEST และ TNC

โต๊ะกินข้าวกับพ่อแม่ โต๊ะกินข้าว จะไม่ใช่คำประสมแล้ว แต่โอกาสแบบนี้ไม่น่าจะเกิดขึ้นบ่อยนัก เพราะเรามักจะเสี่ยงพูดไปแบบอื่นหรือไม่ เช่น เขานั่งกินข้าวกับพ่อแม่ เสียมากกว่า ดังนั้น การไม่รวบรวมคำประสมไว้ในข้อมูลที่เห็น ก็เป็นการปล่อยโอกาสที่จะกำหนดขอบเขตคำประสมทั่วไปที่พบใช้มากในภาษาไทย แล้วต้องมาแก้ปัญหาภายหลัง เช่น อาจต้องมาระบุคำประสมแบบเดียวกับที่ระบุชื่อเฉพาะ

ส่วนการคุมการทำข้อมูลให้ตรงกัน ควรสร้าง online compound glossary ให้คนทำข้อมูลเห็นร่วมกัน เพื่อให้มีการตัดคำได้ตรงกัน อาจใช้ webboard เพื่อเปิดประเด็นคำถามและแลกเปลี่ยนความเห็นจนได้ข้อสรุป วิธีการเหล่านี้ก็จะช่วยให้การตัดคำประสมมีความสม่ำเสมอได้ เหมือนคนที่ทำงานแปลที่ต้องแบ่งข้อมูลช่วยกันแปลก็จะใช้เทคโนโลยีของ machine aided translation ที่ช่วยให้สามารถแชร์ glossary ร่วมกัน มีโปรแกรมช่วยตรวจความสม่ำเสมอของการแปลจากผู้แปลหลาย ๆ คนได้

ประเมินผลอย่างไร

การประเมินผลการตัดคำแบบที่นิยมทำกันคือ การประเมินโดยเทียบกับคำตอบหรือเฉลยที่เตรียมไว้ ในรายละเอียดก็อาจมีวิธีการวัดที่แตกต่างกัน เช่น วัดจากระดับอักษรคือดูว่าแต่ละตัวอักษรมีคำตอบที่เป็นไปได้สองอย่าง คือ ตัดหรือไม่ตัด Y/N ก็เทียบผลที่ตัดออกมาได้กับเฉลยที่มีว่าแตกต่างกันไหม ค่าที่ได้คือ accuracy ซึ่งโดยทั่วไปจะสูง เพราะคิดรวมตัวอักษรที่ไม่ใช่ word boundary เข้าไปด้วย กรณีสมมติที่โปรแกรมไม่ตัดคำเลย คือ ให้เป็น N หมดทุกตัวอักษร

$$\text{accuracy} = (\text{no.char} - \text{no.word}) / \text{no.char}$$

ถ้าสมมติว่าความยาวของคำโดยเฉลี่ยคือ 10 ตัวอักษร accuracy จะอยู่ที่ 90% โดยไม่ต้องทำอะไรเลย. ตัวอย่างเช่น. ในตัวอย่างข้างล่าง หากต้องการผลตัดคำแบบนี้

สารกึ่งตัวนำที่มีคุณสมบัติทางไฟฟ้าอยู่ระหว่างตัวนำไฟฟ้า และฉนวนไฟฟ้า |จิง|เป็น|
สารที่เราสามารถควบคุมคุณสมบัติ|นำไฟฟ้าของ|มัน|ได้|

หากเราไม่ได้ตัดคำเลย คือผลการตัดคำออกมาเป็น =>

สารกึ่งตัวนำที่มีคุณสมบัติทางไฟฟ้าอยู่ระหว่างตัวนำไฟฟ้า และฉนวนไฟฟ้า จึงเป็นสารที่เราสามารถควบคุมคุณสมบัตินำไฟฟ้าของมันได้

จะได้ accuracy = 77.69% หากมีการตัดคำถูกบ้างผิดบ้าง เช่นในตัวอย่างนี้ที่ ^ แทนตำแหน่งที่ตัดผิด จะได้ accuracy = 90.98%

สารกึ่งตัวนำที่มีคุณสมบัติทางไฟฟ้าอยู่ระหว่างตัวนำไฟฟ้า และฉนวนไฟฟ้า | จึงเป็นสารที่เราสามารถควบคุมคุณสมบัตินำไฟฟ้าของมันได้ |

วิธีวัดผลโดยดูที่ระดับตัวอักษรจึงมักให้ค่า accuracy ที่สูงค่อนข้างมาก จึงไม่เหมาะที่จะนำมาใช้ประเมินผลโดยตรง วิธีการวัดอีกแบบคือ ดูที่ระดับคำ คือ ดูว่า สายคำที่ตัดออกมาได้นั้นเป็นคำที่ตรงกับเฉลยมากน้อยเพียงใด กรณีนี้สามารถประเมินโดยดูค่าความแม่นยำ (precision) และค่าความครบถ้วน (recall) ได้โดยดูรายการคำที่ตัดออกมาว่าแม่นยำเป็นคำตรงตามเฉลยกี่คำ (precision) และดูคำที่เฉลยว่าได้มาครบถ้วนคือพบอยู่ในคำที่ตัดออกมากี่คำ (recall) จากนั้นมาคำนวณเป็นค่ากลาง F-measure อีกทีได้ ซึ่งปกติคำนวณจาก

$$F\text{-measure} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

จากตัวอย่างข้างบน กรณีที่ไม่ได้ตัดคำอะไรเลย precision กับ recall จึงเป็น 0 และค่า F-measure ก็จะเป็น 0 ด้วย ส่วนกรณีที่ตัดคำผิดนั้นในตัวอย่างนี้ได้ F-measure = 0.5079 ซึ่งก็สะท้อนผลการตัดคำที่เหมาะสมกว่าการคิดแบบตัวอักษรที่ได้ถึง 90.98%

สารกึ่งตัวนำที่มีคุณสมบัติทางไฟฟ้าอยู่ระหว่างตัวนำไฟฟ้า และฉนวนไฟฟ้า | จึงเป็นสารที่เราสามารถควบคุมคุณสมบัตินำไฟฟ้าของมันได้ |

$$\text{precision} = 16/37 \quad \text{recall} = 16/26$$

$$F\text{-measure} = 0.5079$$

แต่ไม่ว่าจะวัดด้วยวิธีไหน ทั้งสองวิธีที่กล่าวมาเป็นการวัดโดยเทียบกับเฉลยเป็นหลัก ซึ่งวิธีประเมินผลโดยเทียบตำแหน่งของการตัดคำกับข้อมูลเฉลยเป็นวิธีที่นิยมใช้

ประเมินประสิทธิภาพของโปรแกรมตัดคำมากที่สุด เพราะดูจะเป็นวิธีที่ objective แต่เราต้องไม่ลืมว่า assumption ของวิธีการนี้ คือข้อมูลที่นำมาทดสอบและตัดคำแล้วมีความถูกต้อง และมีการตัดคำที่สม่ำเสมอ แต่ในความเป็นจริง ข้อมูลที่ใช้เป็นตัวเทียบหรือเฉลยนั้นอาจตัดคำมาจากหลาย ๆ คน ซึ่งก็อาจทำให้ผลการตัดคำมีความไม่สม่ำเสมออยู่ภายในได้ สิ่งที่ต้องพยายามทำคือทำให้คลังข้อมูลตัดคำที่ใช้เป็นตัวเฉลยนั้นมีการตัดคำที่สม่ำเสมอมากที่สุดเท่าที่จะทำได้

อย่างไรก็ดี ผลที่ได้จากการประเมินแบบนี้ บอกเราได้ว่าโปรแกรมตัดคำอันไหนตัดคำแล้วตรงกับเฉลยมากที่สุด การประเมินการตัดคำด้วยวิธีนี้จึงต้องพิจารณาด้วยว่าโปรแกรมตัดคำที่ใช้ขึ้นมองการตัดคำอย่างไร ดังที่ได้กล่าวมาแล้วว่า บางงานหรือบางโปรแกรมก็อยากได้คำที่ตัดออกมาเป็นคำย่อย ๆ แต่บางงานหรือบางโปรแกรมก็อยากได้คำที่ตัดออกมาเป็นหน่วยศัพท์ คืออยากได้คำประสมต่าง ๆ เป็นคำด้วย การประเมินโปรแกรมตัดคำ จึงต้องพิจารณาว่าคลังข้อมูลตัดคำที่ใช้เทียบกับโปรแกรมตัดคำนั้นมองขอบเขตของคำเหมือนกันหรือไม่ หากคลังข้อมูลที่ใช้เป็นคลังข้อมูลที่ตัดคำแบบเป็นคำย่อย เช่น BEST Corpus และโปรแกรมตัดคำที่ใช้เลือกตัดคำแบบใช้พจนานุกรมตามแนว maximum matching แบบนี้โปรแกรมก็จะตัดคำออกมาผิดจากต้นแบบที่ต้องการมาก เช่น โปรแกรมตัดคำเป็น “กลางวัน” จะถูกมองว่าตัดผิดเพราะเฉลยคือ “กลาง|วัน”

สิ่งที่ควรพิจารณาคือการตัดคำให้เหมือนกับต้นแบบที่สุดเป็นสิ่งที่สะท้อนคุณภาพของโปรแกรมตัดคำจริงหรือไม่ หรือจริง ๆ เราควรตั้งคำถามก่อนว่า ในงานที่เราต้องการทำนั้นควรใช้ข้อมูลตัดคำที่มีลักษณะแบบใด ถ้าเราแน่ใจว่าข้อมูลตัดคำแบบคำย่อยแบบ BEST corpus เป็นสิ่งที่เหมาะกับงานที่ทำ ก็สามารถประเมินประสิทธิภาพโปรแกรมตัดคำโดยเทียบผลการตัดคำกับคลังข้อมูล BEST ได้เลย แต่หากเราต้องการข้อมูลคำที่มีขอบเขตใหญ่ขึ้น ก็จะมีปัญหาเพราะไม่มีคลังข้อมูลกลางที่จะนำมาใช้เทียบประเมินผลได้ ที่ทำได้คือสุ่มตัวอย่างข้อมูลที่ต้องการมาตัดคำเองใน

แบบที่ต้องการแล้วนำมาเทียบกับผลจากโปรแกรมตัดคำ แต่ค่าความถูกต้องที่ได้ก็จะ เป็นค่าเฉพาะที่เราใช้ได้คนเดียว จะนำไปเทียบกับผลประเมินคนอื่นไม่ได้โดยตรง ยกเว้นเราเอาข้อมูลเราทดสอบโปรแกรมตัดคำอื่น ๆ ด้วย

การประเมินผลจากตัวบทที่พบจริงนั้น โปรแกรมตัดคำมักจะให้ผลการตัดคำที่ ถูกต้องสูงเหตุเพราะความกำกวมที่เป็นปัญหานั้นไม่ใช่สิ่งที่จะพบมากในตัวบท ตัวอย่างยอดนิมยอย่าง “ตากลม” “ไปหามเหสี” นั้นไม่ใช่ข้อความที่จะพบได้บ่อยใน ตัวบททั่วไป การวัดแบบนี้จึงอาจได้คะแนนความถูกต้องสูงจากเหตุที่ปัญหาความ กำกวมนั้นไม่ได้พบมากนักในข้อความ เราจึงสามารถประเมินผลการตัดคำได้อีกแบบ คือประเมินในเชิงทฤษฎี โดยวิเคราะห์ว่าปัญหาการตัดคำอาจเกิดในกรณีใดบ้างและ หาตัวอย่างที่แสดงถึงปัญหานั้น นำมาสร้างเป็นชุดทดสอบโดยตรงเพื่อดูว่าเมื่อเจอกรณี ปัญหาเหล่านี้ โปรแกรมตัดคำจะสามารถตัดคำได้ถูกต้องเพียงใด

ตัวอย่างของกรณีปัญหาที่สามารถແจกแจงเพื่อสร้างเป็นข้อมูลทดสอบ เช่น

- 1) ตัดคำผิดจากแบบแต่ไม่ทำให้เปลี่ยนความหมาย เช่น
รถโดยสาร จะตัดเป็น *รถ/โดยสาร* หรือ *รถโดยสาร* ก็ไม่ต่างกันมาก หรือ *หม้อหุงข้าวไฟฟ้า* จะตัดเป็น *หม้อ/หุงข้าว/ไฟฟ้า* หรือ *หม้อหุงข้าวไฟฟ้า* ก็ไม่ได้ส่งผลให้ ความหมายแตกต่างกัน
- 2) ตัดคำผิดจากแบบทำให้ไม่สื่อความหรือความหมายเปลี่ยนไป เช่น
เกี่ยวกับไม้ดอกไม้ประดับ ถ้าตัดผิดเป็น *เกี่ยวกับไม้/ดอกไม้ประดับ* ความหมายก็จะผิดไป หรือ *ใครที่กำลังหาต้นไม้ดอกไม้ประดับหน้าอาคาร* ถ้าตัดคำอีกแบบออกมาเป็น *ใคร/ที่กำลัง/หา/ต้นไม้/ดอกไม้ประดับ/หน้าอาคาร/* ก็จะผิดความหมายไป หรือ *พายุงวงช้างถล่ม* หากตัดผิดเป็น *พาย/ุง/วง/ช้าง/ถล่ม* ก็จะไม่ได้อรรถาธิบายที่ถูกต้อง
- 3) ตัดคำได้มากกว่าหนึ่งแบบ แต่ละแบบมีความหมายที่ต่างกัน จะเลือกแบบไหนนั้นขึ้นกับบริบท ที่ปรากฏ เช่น *ตากลม* จะตัดเป็น *ตา/กลม* หรือ *ตากลม* ก็มีความหมายทั้งคู่ แต่ความหมายไหน ถูกต้องต้องดูที่บริบทนั้น

การประมวลผลภาษาไทย

หลวงตามหาบัว จะตัดเป็น หลวง/ตาม/หา/บัว หรือ หลวงตามหาบัว ก็ได้ หรือ
ทางการเมือง จะตัดเป็น ทางการ/เมือง(นครสวรรค์) หรือ (วัฒนธรรม)/ทาง/การเมือง ก็มี
ความหมายทั้งคู่แล้วแต่บริบทว่าควรตัดแบบไหน

ที่อยู่อาศัย จะตัดเป็น (คน)ที่อยู่/อาศัย(ในย่านนี้) หรือ (บริเวณนี้เป็น)/ที่อยู่อาศัย(ของสัตว์ป่า) ก็
ได้ความหมายที่ต่างกันไป

4) ตัดคำผิดเพราะเป็นคำที่ไม่รู้จัก เช่น

เป็นชื่อเฉพาะ เช่น *ตันตริยานนท์*

เป็นคำเขียนทับศัพท์ เช่น *เฟอร์มิเจอร์ ออนไลน์*

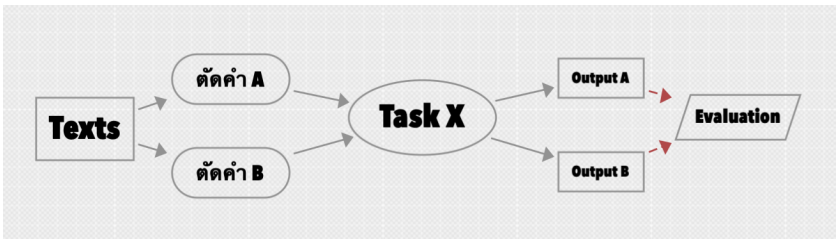
เป็นคำแผลงมาจากคำปกติ เช่น *จุงเบย น่ารักกกก*

เป็นคำที่สะกดผิดแล้วไม่เป็นคำ เช่น *โคตร* เขียนเป็น *โครต*

ตัวอย่างที่กล่าวมาข้างบนเป็นตัวอย่างที่มักเป็นปัญหาของการตัดคำ โดยเฉพาะ
กรณีข้อ 2-4 สามารถใช้เป็นตัวชี้วัดประสิทธิภาพของโปรแกรมตัดคำได้ว่าตัดมาแล้ว
ผิดความหมายไปหรือไม่ โดยเราอาจรวบรวมตัวอย่างกรณีต่าง ๆ นี้มาเป็นชุดทดสอบ
แล้วประเมินโดยนับว่าโปรแกรมได้ตัดคำในตัวอย่างประโยคที่มีคำเหล่านี้แล้วผิดไป
มากน้อยเพียงใด แต่ก็ต้องไม่ลืมประเด็นว่า การตัดคำที่ใช้นั้นเลือกแบบไหน เพราะถ้า
เลือกแบบตัดย่อย ความผิดพลาดที่ยกมาข้างบนอาจไม่ปรากฏก็ได้ เช่น ถ้าตัด
“ดอกไม้” เป็น “ดอก|ไม” ตัวอย่าง “เกี่ยวกับไม้ดอกไม้ประดับ” ก็จะไม่มีปัญหาใน
การตัดคำเพราะทั้ง “ไม้ดอกไม้” และ “ดอกไม้” ถูกตัดแยกเป็นคำย่อยหมด ณ จุดนี้ เรา
อาจคิดว่า เราก็คควรตัดคำเป็นคำย่อย เพราะโอกาสผิดจะน้อยกว่า หากเราต้องการผล
เพียงเพื่อได้ตัวเลขความถูกต้องสูง ๆ วิธีการนี้ก็เหมาะสม แต่จริงๆ การตัดคำเป็นเพียง
กระบวนการหนึ่งในกระบวนการใหญ่ สิ่งที่เราต้องการคือความถูกต้องของงาน
เป้าหมายหลักมากกว่าใช่ไหม

การประเมินผลอีกแบบ คือประเมินจากการใช้จริง เนื่องจากงานที่เราต้องการ
ทำอาจไม่ได้ให้ความสำคัญว่าโปรแกรมตัดคำได้ถูกมากน้อยแค่ไหน แต่สนใจว่า
ท้ายที่สุดในงานที่ต้องการทำได้ถูกต้องแค่ไหน เช่น ถ้าเราต้องการทำ text

classification หรือแยกประเภทตัวบท ผลที่ต้องการคือ ตัวบทแต่ละชิ้นถูกจัดประเภท ถูกต้องตามที่ต้องการมากแค่ไหน และในกระบวนการนี้ เราก็ต้องการโปรแกรมตัดคำ เพื่อนำผลไปใช้ต่อ เราก็สามารถประเมินผลโปรแกรมตัดคำโดยการทดลองใช้โปรแกรมตัดคำแบบต่าง ๆ เพื่อทดลองว่าโปรแกรมตัดคำแบบไหนทำให้โมเดลที่ทำอยู่ได้ผลลัพธ์สุดท้ายที่ถูกต้องมากกว่า



หรือในบางโมเดลที่ทำ เราก็อาจจะไม่ได้ต้องการใช้ข้อมูลคำเลยก็ยังสามารถทำได้ การประเมินผลการตัดคำจึงเป็นเรื่องที่ไม่ได้มีมาตรฐานเดียว ขึ้นกับความต้องการและไม่ได้เป็นสิ่งจำเป็นที่ขาดไม่ได้ หรือต้องได้ถูกต้องตามเฉลยมาก ๆ จึงจะใช้ได้ การประเมินผลที่ดีจึงขึ้นกับความต้องการในการใช้งานมากกว่าว่าเมื่อตัดออกมาแล้วผลที่ได้ตรงตามที่ต้องการและเป็นประโยชน์ที่จะนำไปใช้ต่อหรือไม่

สาธิต

โปรแกรมตัดคำภาษาไทยสามารถทดลองเล่นได้จาก Python package TLTK ให้ติดตั้งผ่าน `pip install tltk`. จากนั้นเรียกโมดูล `tltk.nlp.word_segment(...)`. Option `method="mm"` สำหรับเรียกใช้วิธีการตัดคำแบบ maximum matching. หากไม่กำหนดจะใช้การตัดคำแบบ maximum collocation ที่ตัดพยางค์ก่อนแล้วรวบคำจากการการรวมพยางค์ที่ให้ค่า collocation strength ของสายคำสูงสุด. หรือทดลองผ่าน Google colab ตามที่อยู่นี้

การประมวลผลภาษาไทย

- https://colab.research.google.com/drive/1YCrhKqO9YEVXgs-_HYXv4n5N73LdzF2Y#scrollTo=t4wslzDCHKL&line=3&uniqifier=1

```
import nltk
```

```
text = "การที่ประชาชนทำหน้าที่ปกครองด้วยตนเองโดยตรง หรือการเลือกผู้แทนเข้าไปทำหน้าที่แทนตนนั้น เป็น  
กลไกที่สำคัญที่สุดที่จะทำให้กระบวนการทางการเมืองการปกครองในระบบประชาธิปไตยทุกกระบวนการ ทุก  
ระดับ และทุกมิติ ดำเนินไปอย่างมีประสิทธิภาพ โดยที่ประชาชนต้องมีคุณสมบัติที่เอื้อหรือสนับสนุนต่อหลักการ  
ประชาธิปไตย เช่น มีความสนใจ กระตือรือร้นที่จะเข้ามีส่วนร่วมในทางการเมือง หรือติดตาม ควบคุม และ  
ตรวจสอบการทำงานของรัฐบาลอย่างจริงจัง เป็นต้น หรืออาจกล่าวได้อีกนัยหนึ่งว่า ประชาชนต้องมีวัฒนธรรมทาง  
การเมืองตามที่คนๆว่า "กิจกรรมทางการเมืองการปกครองเป็นหน้าที่ที่ทุกคนต้องเอาใจใส่รับผิดชอบจะหลีกเลี่ยง  
หรือปฏิเสธให้พ้นความรับผิดชอบของตนหาได้ไม่" หรือ "การเมืองเป็นเรื่องของทุกคน"
```

```
c = nltk.nlp.word_segment(text) ## use collocation strength approach
```

```
m = nltk.nlp.word_segment(text,method='mm') ## use maximum matching
```

```
print(c)
```

```
print(m)
```

- การที่ประชาชนทำหน้าที่ปกครองด้วยตนเองโดยตรง</>หรือการเลือกผู้แทนเข้าไปทำหน้าที่แทนตนนั้น</>เป็นกลไกที่สำคัญที่สุดที่จะทำให้กระบวนการทางการเมืองการปกครองในระบบประชาธิปไตยทุกกระบวนการ</>ทุก</>ระดับ</>และทุก</>มิติ</>ดำเนินไปอย่างมีประสิทธิภาพ</>โดยที่ประชาชนต้องมีคุณสมบัติที่เอื้อหรือสนับสนุนต่อหลักการประชาธิปไตย</>เช่น</>มีความสนใจ</>กระตือรือร้นที่จะเข้ามีส่วนร่วมในทางการเมือง</>หรือติดตาม</>ควบคุม</>และ</>ตรวจสอบ</>การทำงานของ</>รัฐบาล</>อย่าง</>จริงจัง</>เป็น</>ต้น</>หรือ</>อาจ</>กล่าว</>ได้</>อีก</>นัย</>หนึ่ง</>ว่า</>ประชาชน</>ต้อง</>มี</>วัฒนธรรม</>ทาง</>การเมือง</>ตาม</>ที่</>คน</>ๆ</>ว่า</>“กิจกรรมทางการเมืองการปกครองเป็นหน้าที่ที่ทุกคนต้องเอาใจใส่รับผิดชอบจะหลีกเลี่ยงหรือปฏิเสธให้พ้นความรับผิดชอบของตนหาได้ไม่”</>หรือ</>“การเมืองเป็นเรื่องของทุกคน”</>
- การที่ประชาชนทำหน้าที่ปกครองด้วยตนเองโดยตรง</>หรือการเลือกผู้แทนเข้าไปทำหน้าที่แทนตนนั้น</>เป็นกลไกที่สำคัญที่สุดที่จะทำให้กระบวนการทางการเมืองการปกครองในระบบประชาธิปไตยทุกกระบวนการ</>ทุก</>ระดับ</>และ</>ทุก</>

มิติ</>ดำเนินไปอย่างมี^{ประสิทธิภาพ}โดยที่^{ประชาชน}ต้องมี^{คุณสมบัติ}ที่^{เอื้อ}หรือ^{สนับสนุน}ต่อ^{หลักการ}ประชาธิปไตย^{เช่น}มี^{ความสนใจ}กระตือรือร้น^{ที่จะ}เข้า^{มีส่วนร่วม}ใน^{ทางการ}เมือง^{หรือ}ติดตาม^{และ}ตรวจสอบ^{การ}ทำงาน^{ของรัฐ}อย่าง^{จริงจัง}เป็น^{ต้น}หรือ^{อาจ}กล่าว^{ได้}อีก^{นัย}หนึ่ง^{ว่า}ประชาชน^{ต้อง}มี^{วัฒนธรรม}ทางการ^{เมือง}ตาม^{ที่}คน^{ที่}ว่า[“]^{กิจกรรม}ทางการ^{เมือง}การ^{ปกครอง}เป็น^{หน้าที่}ที่^{ทุกคน}ต้อง^{เอา}ใจ^{ใส่}รับ^{ผิด}ชอบ^{จะ}หลีกเลี่ยง^{หรือ}ปฏิเสธ^{ให้}พ้น^{ความ}รับผิดชอบ^{ของตน}หา^{ได้}ไม่[”]หรือ[“]^{การ}เมือง^{เป็น}เรื่อง^{ของ}ทุกคน[”]

-

อ้างอิง

- รัตติกกร วรากุลศิริพันธ์, จงกล งามวิวิทยา, สมศักดิ์จันวัน, สุชาติพิทยจักรยากุล และศักดิ์ชัย พิทยจักรรัตน์. (2538). การตัดคำจากประโยคภาษาไทยด้วยวิธีการเทียบคำที่ยาวที่สุด. Papers on Natural Language Processing, Compiled by Virach Sornlertlamvanich.
- วิรัช ศรีเลิศล้ำวานิช. (2536). การตัดคำในระบบแปลภาษา (Word Segmentation for Thai in Machine Translation System). การแปลภาษาด้วยคอมพิวเตอร์. NECTEC. หน้า 50–55.
- สมปรารถนา รัชยานนท์. (2535). โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย. วิทยานิพนธ์ปริญญาโทมหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.
- Aroonmanakun, W. (2002). Collocation and Thai Word Segmentation. In T. Theeramunkong & V. Sornlertlamvanich (Eds.), Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop.

- Aroonmanakun, W. (2007). Thoughts on Word and Sentence Segmentation in Thai. In Proceedings of the Seventh Symposium on Natural Language Processing, Dec 13–15, 2007, Pattaya, Thailand (pp. 85–90).
- Kawtrakul, A., et al. (1995). A lexicon model for writing production assistance system. In Proceedings of the Symposium on Natural Language Processing on Thailand '95.
- Kosawat, K., Boriboon, M., Chootrakool, P., Chotimongkol, A., Klaithin, S., Kongyoung, S., Kriengket, K., Phaholphinyo, S., Purodakananda, S., Thanakulwarapas, T., & Wutiwiwatchai, C. (2009). BEST 2009: Thai word segmentation software contest. 2009 Eighth International Symposium on Natural Language Processing, 83-88.
- Meknavin, S., Charoenpornasawat, P., & Kijisirikul, P. (1997). Feature-based Thai Word Segmentation. Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS) 1997.
- Thairatananond, Y. (1981). Towards the Design of a Thai Text Syllable Analyzer. Master's thesis, Asian Institute of Technology.
- Theeramunkong, T., & Usanavasin, S. (2001). Non-Dictionary-Based Thai Word Segmentation Using Decision Trees. HLT.
- Thai NLP resources. (n.d.). GitHub. Retrieved from https://github.com/kobkrit/nlp_thai_resources
- Universal Dependencies. (n.d.). Retrieved from <http://universaldependencies.org/>

คำประสมภาษาไทย³

การประสมคำเป็นกลวิธีการสร้างคำในภาษาไทยที่มีความสำคัญมากต่อการประมวลผลภาษา เพราะปัญหาของการตัดคำภาษาไทยส่วนหนึ่งมาจากคำประสมที่บางบริบทก็ควรมองเป็นหลายคำแต่ในบางบริบทก็ควรมองเป็นคำเดียว เพื่อให้เข้าใจธรรมชาติของการประสมคำซึ่งเป็นกลวิธีหลักในการสร้างคำภาษาไทย ผู้เขียนจะกล่าวถึงแนวคิดเรื่องคำประสม แนวทางการวิเคราะห์คำประสมในทางภาษาศาสตร์ ตลอดจนงานวิจัยประเด็นคำถามที่เป็นที่สนใจศึกษาในเรื่องของคำประสม เมื่อได้เข้าใจเรื่องคำประสมในภาพรวมสำหรับทุกภาษาแล้ว จากนั้น จะกล่าวถึงการศึกษาคำประสมในภาษาไทยเพื่อให้เห็นภาพรวมการศึกษาเกี่ยวกับคำประสมในภาษาไทย จะได้กำหนดแนวทางการสกัดคำนามประสมภาษาไทยที่เหมาะสม แล้วจึงทดลองสกัดคำประสมภาษาไทย

คำประสมคืออะไร

คำประสมหรือ compounds เป็นคำที่สร้างขึ้นผ่านกระบวนการทางวชิวิภาคที่เรียกว่า การประสมคำ (compounding) เกิดจากการนำหน่วยศัพท์ (lexeme) ตั้งแต่สองหน่วยขึ้นไปมาประกอบกันขึ้นเป็นหน่วยศัพท์ใหม่ (Aronoff and Fudeman 2011:47, Booij 2007:137) คำใหม่ที่ได้จะมีความหมายไม่เหมือนกับการนำ

³ เรียบเรียงใหม่จากรายงานวิจัยเรื่อง “การสกัดคำประสมภาษาไทย” โครงการกลุ่มนวัตกรรมวิชาการเชิงบูรณาการ โครงการในแผนพัฒนาวิชาการจุฬาลงกรณ์มหาวิทยาลัย ๒๕๕๔-๒๕๕๖

ความหมายของแต่ละหน่วยศัพท์มารวมกัน เช่น “แม่น้ำ” ไม่ได้มีความหมายที่เกิดจากการรวมความหมายของคำว่า “แม่” กับ “น้ำ” การสร้างคำประสมในภาษาไทยและในอีกหลายๆ ภาษาทำได้โดยการนำคำมาประกอบเรียงต่อกัน แต่ในบางภาษาจะมีกลไกเฉพาะเช่นมีการเติม interfix เพื่อบอกการเป็นคำประสม เช่น *psom-o-tiri* 'bread (and) cheese' มาจาก *psom-i* 'bread', *tir-i* 'cheese' สังเกตว่ามีการเติม interfix -o- ซึ่งไม่ได้มีความหมายใดเป็นพิเศษนอกจากใช้เพื่อบอกว่าคำข้างหน้าเป็นส่วน of คำประสม (Ralli 1992, อ้างถึงใน Booij 2007:88)

ประเภทของคำประสม

การนำคำตั้งแต่สองคำมาประสมกันเป็นคำใหม่ เราสามารถแยกประเภทคำประสมได้หลายวิธี เช่น แยกโดยองค์ประกอบของคำประสมว่ามาจากคำประเภทใด ตัวอย่างเช่น noun-noun compound, adjective-noun compound, noun-adjective compound, verb-noun compound เป็นต้น หรือดูว่าคำประสมนั้นมีหมวดคำตรงกันหรือต่างไปจากหมวดคำของคำที่นำมาประสม แต่การแบ่งประเภทที่นักภาษาศาสตร์มักใช้กัน คือ การดูว่าคำประสมนั้นมีคำใดคำหนึ่งเป็นคำหลักและเป็นแก่นของความหมายของคำประสมหรือไม่ ซึ่งจะช่วยให้แยกคำประสมได้เป็นสองประเภทใหญ่ๆ คือ คำประสมแบบเข้าสู่ศูนย์ (endocentric compound) กับคำประสมแบบไร้ศูนย์ (exocentric compound)

คำประสมแบบเข้าสู่ศูนย์ (Endocentric compound)

Bloomfield (1933) ได้แยกประเภทของคำประสมออกเป็น endocentric และ exocentric ซึ่งนักภาษาศาสตร์ยุคต่อมาก็ได้ใช้หลักการเดียวกันนี้อธิบายคำประสมแบบ endocentric ว่าคือคำประสมที่มีคำหนึ่งทำหน้าที่เป็นคำหลัก (head) และเป็นแก่นความหมายของคำประสมนั้น (Bloomfield 1933:235, Aronoff and Fudeman 2011:114, Booij 2007:139) โดยที่มีคำอื่นมาประกอบเพื่อขยายความหมายของคำหลักนั้น คำประสมประเภทนี้จึงมองได้ว่าเป็นคำที่มีความหมายเป็นเซ็ทย่อยของ

คำหลักหรือเป็นคำลูกกลุ่ม (hyponymn) เช่น *office staff* หมายถึง staff ประเภทหนึ่งที่ประจำสำนักงาน หรือ “ตู้เสื้อผ้า” ถือว่าเป็นตู้ประเภทหนึ่งใช้เพื่อเก็บเสื้อผ้า คำหลักอาจอยู่ซ้ายหรือขวาก็ได้ซึ่งถือเป็นลักษณะเฉพาะของแต่ละภาษา การพิจารณาว่าคำใดเป็นคำหลักอาจดูได้จากกระบวนการวจิวิภาคอย่างเช่นการเติมรูปพหูพจน์ ส่วนที่เป็นคำหลักจะเป็นส่วนที่เปลี่ยนรูปเป็นพหูพจน์ได้ Dressler (2006:33) กล่าวว่าคำประสมส่วนใหญ่แล้วจะเป็นคำประสมแบบ endocentric เพราะเป็นที่นิยมใช้ในภาษาต่าง ๆ ทั่วโลกมากกว่าแบบ exocentric เนื่องจากสามารถเข้าถึงความหมายได้ง่ายกว่า แต่ก็มีผู้แย้งว่าในบางภาษาเช่นภาษา Seediq คำประสมแบบที่เป็น exocentric เป็นคำประสมที่พบบ่อยกว่า (Holmer 1996 อ้างถึงใน Bauer 2010)

คำประสมแบบไร้ศูนย์กลาง (Exocentric compound)

คำประสมแบบนี้คือคำประสมอื่น ๆ ที่ไม่ใช่คำประสมแบบ endocentric คือไม่มีคำใดเป็นคำหลักที่ชัดเจน ในกรณีเช่นนี้ความหมายของคำประสมจึงไม่สามารถดูได้จากคำใดคำหนึ่งโดยตรงเหมือนกรณีคำประสมแบบ endocentric ที่เราดูความหมายแทนที่คำหลักได้โดยตรง ความหมายของคำประสมแบบไร้ศูนย์กลางนี้อาจจะมีความหมายที่แตกต่างไปจากความหมายของคำที่มาประสมกันไปเลย เช่น *pickpocket* จะหมายถึงโจรล้วงกระเป๋า ไม่ได้หมายถึงการหยิบกระเป๋า ในกรณีคำประสมแบบ exocentric นี้ นักภาษาศาสตร์ได้มองเห็นความแตกต่างกันและสามารถจัดแบ่งประเภทย่อยในกลุ่มได้หลายประเภท ได้แก่

bahuvrihi compound เป็นคำที่มาจากภาษาสันสกฤต ใช้อ้างถึงคำประสมที่ดูเหมือนจะมีโครงสร้างแบบเข้าสู่ศูนย์กลาง (endocentric) ที่สามารถมองได้ว่ามีคำหลักและคำขยายแต่ความหมายของคำประสมนั้นไม่ได้มีแก่นจากความหมายของคำหลักนั้น แต่หมายรวมถึง บุคคลหรือกลุ่มบุคคลที่ครอบครองหรือมีคุณสมบัตินั้น เช่น คำ *birdbrain* ที่หมายถึงคนที่ดูเหมือนมีสมองเท่านั้น คำ *bahuvrihi* มาจากภาษา

สันสกฤต มีความหมายว่า 'มาก'+ 'ขาว' จึงใช้เพื่อหมายถึงคนมีฐานะก็คือคนที่มีขาวจำนวนมาก โครงสร้างของคำประสมประเภทนี้จึงมักประกอบด้วยคำนามกับคำคุณศัพท์ โดยที่คำนามจะอ้างถึงคน สัตว์ พืช หรือสิ่งที่มีลักษณะตามคำคุณศัพท์ที่ขยาย นอกจากนี้ คำประสมประเภทนี้อาจประกอบจาก คำขยายจำนวนกับคำนาม เช่น ในภาษา Danish คำ *tusind-ben* ('thousand' + 'leg') หมายถึง กิ้งกือ คำกริยากับคำนาม เช่น ในภาษา Urarina คำ *lura-kutúri* ('be bald' + 'head') ที่หมายถึงคนหัวล้าน หรือคำนามกับคำนาม เช่น *paperback* ที่หมายถึงหนังสือปกอ่อน (Bauer 2008:56-58) นอกจากนี้ Bauer ยังกล่าวถึง *bahuvrihi* compound ว่าสามารถใช้เป็นคำคุณศัพท์หรือคำกริยาได้ในบางภาษาด้วย กรณีคำไทยที่จัดเป็นคำประสมประเภทนี้ เช่น “เสื้อแดง” ที่หมายถึงกลุ่มคนที่ใช้เสื้อแดงเป็นสัญลักษณ์และมีความคิดความเชื่อทางการเมืองกลุ่มหนึ่ง ไม่ได้หมายถึงเสื้อที่มีสีแดง ก็จัดเป็นคำประสมประเภทนี้ได้

synthetic compound เป็นคำประสมที่มีคำกริยาเป็นแกนและมีคำที่เป็นองค์ประกอบร่วม (*argument*) ของกริยามาประกอบ เช่น *bus-driver*, *pickpocket* (Bauer 2008:61) คำประสมประเภทนี้มองดูเหมือนมีการใช้ทั้งกระบวนการทางวิภาษคดีทั้งแบบ *derivation* และ *compounding*. เช่น *church-goer* ในตัวอย่างประเภทนี้ Booij (2007:90) ยังชี้ให้เห็นว่าคำที่นำมาสร้างคำประสมอาจจะไม่ใช่คำที่มีใช้กันอยู่จริงก็ได้ อย่างคำว่า *goer* นี้ไม่มีที่ปรากฏใช้โดด ๆ แต่เป็นคำที่สร้างมาเพื่อประกอบเป็นคำประสม เป็นคำที่เราเรียกว่าคำที่เป็นไปได้หรือ *possible word* ที่เกิดจาก *go+er* หรืออย่างคำ *sword-swallower* ก็เช่นกัน *swallower* ไม่ใช่คำที่มีใช้กันปกติ คำประสมที่เป็น *synthetic* พวกนี้ไม่สามารถวิเคราะห์เป็น VP ได้เพราะมีลำดับคำที่ต่างจาก VP (*to swallow sword*) มีการนำกรรมมาอยู่หน้าคำนาม จึงจะต้องวิเคราะห์ให้เป็นคำประสมแบบหนึ่งที่มีคำหลักเป็น *derived noun* ที่แปลงมาจากคำกริยาและรับทอดองค์ประกอบ (*inherit argument*) ของคำกริยามา จึงมี

ความสัมพันธ์แบบ agent หรือ patient กับคำกริยานั้นได้ การรวมกันของคำนามกับคำกริยาอีกลักษณะคือการสร้าง verb compound เช่น “กินข้าว” ถูกใช้โดยไม่ได้อ้างถึงสิ่งทีกินว่าจะต้องเป็นข้าว แต่ “กินข้าว” หมายถึงกริยาที่หมายถึง ทานอาหาร เราเรียกกระบวนการนี้ว่า noun incorporation ตัวอย่างเช่นในภาษา Frisian ซึ่งเป็นภาษากลุ่ม Germanic มีคำ *bole-bakt* ที่หมายถึง 'loaf-bake' ซึ่งไม่ได้หมายถึงการอบขนมปังโดยเฉพาะ แต่ใช้หมายถึงการอบอาหารโดยทั่วไป (Booij 2007:92) โดยทั่วไปแล้ว คำประสมประเภทนี้จะสังเกตได้จากกริยาตัวนั้นที่จะกลายจากกริยากรรมเป็นกริยากรรม และคำนามที่มาประกอบก็จะมีลักษณะ unmark คือไม่บอกการชี้เฉพาะ พจน์ และการก

สำหรับภาษาไทยมีคำประสมที่เป็น synthetic compound นี้จำนวนหนึ่ง เช่น “สอนหนังสือ” “ไฟไหม้” แต่การจะตัดสินว่าเป็นคำประสมหรือไม่ในบางครั้งก็ต้องพิจารณาจากบริบทเพราะในบางกรณีรูปที่ปรากฏอาจไม่ใช่คำประสมก็ได้ เช่น ตัวอย่างของคำ “ไฟไหม้” ที่กล่าวถึงในบทความ Lexicalization of syntactic constructions in Thai (Prasithratsint 2010) ที่แสดงถึงความกำกวมทางโครงสร้างในตัวอย่างต่อไปนี้

- มีไฟไหม้ห้าแห่งในเมือง. => “ไฟไหม้” เป็น คำนามประสมเพราะใช้ร่วมกับคำบอกจำนวน และลักษณนาม
- ไฟไหม้มาแล้ว ไฟมอดไม่มาแล้ว => “ไฟไหม้” เป็น นามวลี เมื่อเทียบกับ “ไฟมอด” เป็น N + modifying reduced relative clause หมายถึงไฟที่กำลังไหม้
- ไฟไหม้อาคารพังลงมาที่พื้นดิน => “ไฟ” เป็นประธาน “ไหม้” เป็นกริยา “อาคาร” เป็นกรรม

อย่างไรก็ดี ในตัวอย่างนี้ อาจมีผู้แย้งได้ว่า คำว่า “ไฟไหม้” ไม่จำเป็นต้องเป็นคำประสม เพราะถ้าใช้เกณฑ์ว่าการปรากฏร่วมกับคำบอกจำนวนและลักษณนามได้ บ่งชี้ความเป็นคำประสม ในตัวอย่าง “บ้านขนาดใหญ่สามหลัง” “บ้านขนาดใหญ่” ก็ไม่ได้ถูกมองเป็นคำประสม

Coordinative compound, copulative compound, co-compound หรือที่เรียกว่า dvandva compound ในภาษาสันสกฤต (dvandva = pair) คือคำประสมที่ทั้งสองข้างมีความสำคัญเท่าๆกัน เช่น *washer-dryer*, *speaker-hearer* (Dressler 2006:34), “พ่อแม่” (หมายถึงผู้ปกครอง) “เด็กชายหญิง” คำประสมกลุ่มนี้ไม่ได้มีความหมายเป็นคำลูกกลุ่มของคำใดคำหนึ่งเหมือนคำประสมแบบ endocentric แต่มีความหมายที่ครอบคลุมทั้งสองอย่าง และคำที่นำมาประสมกันอาจเป็นคำตรงข้ามได้ เช่น “ถูกผิด”

Appositive compound เป็นคำประสมที่ดูคล้ายกับ coordinative compound แต่ต่างกันตรงที่ คำที่นำมาประสมกันนั้นต่างอ้างถึงสิ่งเดียวกันหรือมี referent เดียวที่มีคุณสมบัติเป็นทั้งสองอย่างนั้น เช่น *singer-songwriter* คือคนที่ทั้งนี้ทั้งนั้นร้องนักประพันธ์, *bitter-sweet* คือรสที่มีทั้งขมและหวาน

Neo-classical compound เป็นกรณีที่คำใดคำหนึ่งในคำประสมนั้นมีรากศัพท์มาจากภาษาเก่าอย่างภาษากรีก ลาติน โดยที่ไม่ได้มีฐานะเป็นหน่วยศัพท์อิสระ เช่น *bio-logy*, *tele-camera*, *tele-graph* กรณีคำประสมนี้ทำให้เห็นว่าการแยกกระบวนการประสมคำ (compounding) ออกจากการแปลงคำ (derivation) ในบางครั้งก็เป็นเรื่องยาก เพราะหน่วยที่นำมาประกอบใน neo-classical compound นี้จะดูคล้ายกับเป็น affix มากกว่าที่จะเป็นหน่วยศัพท์

นอกจากนี้ ยังมีผู้ที่กล่าวถึงคำประสมแบบ exocentric ว่าเป็น metaphoric compound เช่น *moose milk* ที่หมายถึง ของดื่มที่ทำจากรัมและครีม แต่ Bauer (2008:65) มองว่าคำประสมนี้สามารถพิจารณาว่าเป็นคำที่มีโครงสร้างเป็นคำประสมเข้าสู่ศูนย์ (endocentric compound) แต่ถูกใช้แบบอุปลักษณ์ ทำให้ไม่ได้มีความหมายตามคำหลัก เพราะหากเราแยกประเภทคำประสมตามลักษณะว่ามีการใช้แบบอุปลักษณ์ คำเดียวอย่าง “pig” ที่ใช้แบบอุปลักษณ์เพื่ออ้างถึงคนสกปรกก็จะเข้าข่ายนี้เช่นกัน จึงไม่มีความจำเป็นที่จะแยกคำลักษณะนี้ออกมาเป็นอีกประเภทหนึ่ง

ความแตกต่างระหว่างคำประสมและวลี

ปัญหาหนึ่งของการประมวลผลภาษาไทยคือการแยกคำประสมออกจากวลี หากพิจารณาว่าเป็นคำประสมก็มองเป็นหนึ่งหน่วยศัพท์ แต่ถ้าเป็นวลีก็จะมองเป็นโครงสร้างที่ประกอบด้วยหลายหน่วยศัพท์ แต่การแยกระหว่างคำประสมกับวลี บางครั้งก็เป็นเรื่องยากแม้แต่สำหรับมนุษย์ เพราะหากพิจารณาโดยภายนอกแล้วอาจเห็นลำดับคำที่เหมือนกัน ในการพิจารณานั้น นักภาษาศาสตร์หลายคน เช่น Dressler (2006) Booij (2007) เสนอให้พิจารณาด้วยเกณฑ์ต่างๆ ที่สำคัญคือเกณฑ์ทางวากยสัมพันธ์หรือพิจารณาจากเรื่อง กระบวนการทางวากยสัมพันธ์ (syntactic process) คือถ้าเป็นคำไปแล้ว องค์ประกอบภายในก็ไม่ควรเข้ากฎทางวากยสัมพันธ์ได้ เช่น ไม่สามารถเติม agreement ภายในคำ ไม่สามารถแทรกคำอื่นเข้าไปภายในคำ ประสมได้ หรือไม่สามารถแยกส่วนออกจากกันได้ กรณีของ V+N ก็เช่นเดียวกัน หากเป็นวลีแล้ว N ตัวนั้นก็ควรจะมีการเติมคำขยายอย่าง DET ได้ แต่ถ้าเป็นคำประสมจะทำไม่ได้ นอกจากเกณฑ์ทางวากยสัมพันธ์แล้ว เกณฑ์ทางด้านเสียงก็สามารถใช้แยกความแตกต่างระหว่างคำประสมและวลีได้ เพราะจะได้ยินว่ามีกรลงน้ำหนักคำที่ต่างกัน ตัวอย่างเช่น หากเทียบระหว่าง คำประสมแบบ N+N กับ วลี ADJ+N ทั้งสองอย่างอาจทำหน้าที่เป็นคำขยายแบบเดียวกันได้ เช่น *atom bomb* กับ *atomic bomb*, *industry output* กับ *industrial output* แต่เราไม่มอง N+N ในที่นี้เป็นวลี เพราะการลงน้ำหนักเสียงหลักของ ADJ+N จะอยู่ที่คำหลังไม่ใช่คำแรกเหมือนคำประสม N+N (ดู Booij 2007:82-83) อย่างไรก็ตาม เนื่องจากผู้เขียนสนใจเรื่องการประมวลผลข้อความเป็นหลัก เกณฑ์ทางเสียงจึงไม่สามารถนำมาใช้ได้โดยตรง

นอกจากนี้ คำประสมยังมีความแตกต่างจากคำที่ผ่านการแปลง (derived word) ตรงที่คำประสมเป็นการนำหน่วยศัพท์หรือ lexeme มาประกอบกัน แต่ derived word เป็นการเติม affix ให้กับหน่วยศัพท์เพื่อสร้างหน่วยศัพท์ใหม่

ขึ้นมา อย่างไรก็ตาม เส้นแบ่งระหว่างคำประสม และ derived word อาจจะไม่ชัดเจนในกรณีของ neo-classical compound ดังที่กล่าวมาตอนต้น

ความยากของการระบุคำประสมภาษาไทย

ความยากของการระบุคำประสมในภาษาไทยอยู่ที่เราไม่มีการเติม affix จึงไม่สามารถทดสอบเรื่องการทดลองเติม affix ภายใน ทำได้เพียงการแทรกคำเพื่อดูว่ารูปที่เห็นสามารถแยกหน่วยศัพท์ได้ไหม แต่การแทรกได้ก็ไม่ได้หมายความว่าคำนั้นจะไม่ใช่คำประสมในทุกกรณี เช่น "คนขับรถ" อาจเป็นวลีหรือคำประสมก็ได้ขึ้นกับบริบทที่ปรากฏ การแทรก "ที่" เป็น "คนที่ขับรถ" ได้ไม่ได้เป็นการสรุปแน่นอนว่ารูปคำนี้จะไม่ใช่คำประสมในทุกบริบท เพราะถ้ามองจากความหมาย "คนขับรถ" จะมีความหมายถึงถึงคนที่ประกอบอาชีพขับรถให้บุคคลอื่นไม่ใช่คนที่กำลังกระทำการขับรถ การแทรกคำนี้ในบางบริบทจะทำให้ความหมายต่างจากเดิมได้ การตัดสินว่าคำที่เห็นเป็นคำประสมหรือไม่เป็นจึงเป็นเรื่องที่ต้องอิงบริบทดังในตัวอย่างคำ "ไฟไหม้" ที่ Prasitrathasinth (2010) แสดงให้เห็นในสามประโยคนี้

1. มีไฟไหม้ห้าแห่งในเมือง "ไฟไหม้" เป็นคำนามประสมที่ประกอบจาก N-V เพราะใช้ร่วมกับคำบอกจำนวนและลักษณนาม รวมกันเป็นนามวลี
2. ไฟไหม้น่ากลัว ไฟมอดไม่น่ากลัว "ไฟไหม้" สามารถมองเป็นนามวลีได้ เพราะ "ไฟมอด" คือ ไฟที่กำลังมอด เป็นประธานของภาคแสดง ไม่น่ากลัว "ไฟไหม้" ในประโยคนี้จึงถูกมองแบบเดียวกัน
3. ไฟไหม้อาคารพังลงมาที่พื้นดิน. "ไฟไหม้" ในที่นี้ไม่ใช่คำเดียว เพราะมีกรรมคือ "อาคาร" "ไฟไหม้" จึงมี "ไหม้" เป็นกริยาและมี "ไฟ" เป็นประธาน "ไฟไหม้อาคาร" = subj+V+obj

หรือในตัวอย่าง "ไข่เบ็ด" ที่อ้างถึงใน Singnoi (2000:275-276) ในวลีว่า "ไข่เบ็ดฟองนี้" "ไข่เบ็ด" จัดเป็นคำประสมเพราะเกิดร่วมกับลักษณนาม "ฟอง" แต่หากเป็น "ไข่เบ็ดตัวนี้" "ไข่เบ็ด" จะไม่ใช่คำประสมเพราะคำลักษณนาม "ตัว" ใช้ร่วมกับคำว่า

"เปิด" จึงต้องมองว่า "ตัวนี้" ขยาย "เปิด" ก่อนที่จะรวมกับคำว่า "ไข" เป็นนามวลี (ไข+ (เปิด+ตัวนี้)) Singnoi ยกตัวอย่างนี้มาเพราะอาศัยเกณฑ์ที่ว่าหากมีการใช้คำขยายใด ๆ คำนั้นจะต้องขยายทั้งคำประสมไม่ใช่ส่วนใดส่วนหนึ่ง คำลักษณะนามเป็นคำประเภทหนึ่งที่นำมาขยายคำนามได้ เมื่อ "ตัวนี้" ไม่ได้ขยาย "ไขเปิด" แต่ขยายเฉพาะคำว่า "เปิด" "ไขเปิด" ในตัวอย่างนี้จึงไม่ใช่คำประสม ซึ่งหากใช้เกณฑ์การแทรกคำทดสอบด้วยก็จะได้ผลคล้ายกัน เพราะคำประสมไม่ควรจะถูกแทรกด้วยคำอื่นได้ เราสามารถแทรกคำบุพบท "ของ" เป็น "ไขของเปิดตัวนี้" แต่เมื่อแทรกคำเป็น "ไขของเปิดฟองนี้" จะฟังดูแปลก เพราะไม่ใช่สิ่งที่จะพูดในบริบททั่วไป อาจเป็นไปได้ที่จะใช้ได้ในบางบริบท เช่น หากกำลังกล่าวถึงลักษณะความแตกต่างของไขแต่ละฟองของสัตว์ประเภทต่างๆอยู่ เราก็สามารถพูดได้ว่า "ไขของเปิดฟองนี้ใหญ่มาก" เป็นต้น ดังนั้น ในกรณีทีลำดับคำในคำประสมนั้นเหมือนกับวลีหรืออนุพยางค์ดังในตัวอย่างข้างบน เราต้องอาศัยบริบทเพื่อพิจารณาว่าคำนั้นถูกใช้ในความหมายใด แต่ในกรณีทีลำดับของคำประสมมีลำดับของคำที่แตกต่างไปจากปกติจะช่วยให้ระบุว่าเป็นคำประสมได้ง่ายขึ้น เช่น "หมูสับ" "แก้อี๋นั้ง" มีการสลับตำแหน่งของคำที่เป็นกรรม ทำให้เห็นได้ว่าคำเหล่านี้เป็นคำประสม (Singnoi 2000:276)

ความหมายก็เป็นอีกเกณฑ์หนึ่งที่ถูกนำมาใช้อธิบายความแตกต่างของคำประสม โดยที่คำประสมจะไม่ได้มีความหมายตรงตามคำทั้งหมดที่นำมาประสมกัน เช่น คำว่า "หางเสือ" ที่ยกตัวอย่างในหนังสือไวยากรณ์ภาษาไทยว่าเป็นได้ทั้งคำประสมหรือนามวลี หากมีความหมายตรงตามรูปว่าหางของเสือก็จะเป็นนามวลี เป็นต้น แต่หากหมายถึงส่วนหนึ่งของเรือก็จะเป็นคำประสม เป็นต้น อย่างไรก็ตาม เกณฑ์ต่างๆ ที่ใช้ทดสอบความเป็นคำประสมนั้นเป็นเกณฑ์สำหรับให้มนุษย์ใช้ตัดสิน ด้วยการทดสอบความเหนียวแน่นในคำโดยการแทรกคำภายใน หรือการพิจารณาความหมายที่ไม่ใช่ความหมายเดิมของคำย่อย ๆ เหล่านี้ เกณฑ์เหล่านี้ไม่สามารถนำมาใช้

ประมวลผลข้อความภาษาไทยได้โดยตรง แต่อาจจะสามารถประมวลผลได้ด้วยวิธีทางอ้อมซึ่งจะได้กล่าวถึงต่อไปในภายหลัง

โครงสร้างคำประสม

เนื่องจากคำประสมประกอบด้วยหน่วยศัพท์อย่างน้อยสองหน่วยขึ้นไป ในกรณีที่คำประสมมีหน่วยศัพท์มากกว่าสองหน่วย คำประสมนั้นจะมีโครงสร้างที่แสดงถึงลำดับของการประสมคำ กล่าวคือ กระบวนการประสมคำนั้นเกิดซ้ำได้ทำให้ได้คำที่มีขนาดมากกว่าสองหน่วยศัพท์ได้ เช่น คำ "*White House travel office staff*" สามารถวิเคราะห์ได้ว่ามีโครงสร้างดังนี้ (((*White House*) (*travel office*)) *staff*) (Booij 2007:76). การจัดลำดับการประสมจะมีความสัมพันธ์โดยตรงต่อความหมายของคำประสมนั้น เช่น (*car (radio equipment)*) กับ ((*car radio*) *equipment*) (ตัวอย่างใน Girju et al., 2005) จะมีความหมายต่างกัน คำแรกหมายถึงอุปกรณ์วิทยุภายในรถ ส่วนคำหลังหมายถึง อุปกรณ์สำหรับวิทยุในรถ หรือตัวอย่างภาษาไทย "*คนขับรถบรรทุก*" หากคิดว่า "*คนขับรถ*" เป็นคำประสมแล้ว คำว่า "*คนขับรถบรรทุก*" นี้จะเป็นคำประสมด้วยหรือไม่ ถ้าคิดว่ามีความหมายว่าเป็นอาชีพของคนที่รับจ้างขับรถบรรทุก คำ ๆ นี้ก็น่าจะเป็นคำประสมด้วย ปัญหาที่ต้องพิจารณาต่อมาคือการวิเคราะห์โครงสร้างว่าจะเป็น ((*คน-(ขับ-รถ))-บรรทุก*) หรือ (*คน-(ขับ-(รถ-บรรทุก))*) กรณีแรกคือคิดว่า "*คนขับรถ*" เป็นคำประสมแต่เดิมแล้วจึงประสมคำ "*บรรทุก*" เข้าไปอีก แต่กรณีนี้ไม่น่าจะได้ความหมายอย่างที่ต้องการ คำว่า "*บรรทุก*" ควรประสมกับ "*รถ*" ก่อน แล้วจึงมอง "*รถบรรทุก*" เป็นหน่วยศัพท์ เหมือนกับ "*รถ*" ที่จะนำไปประสมกับคำว่า "*ขับ*" แล้วจึงไปประสมกับคำว่า "*คน*" ดังนั้น การที่รูปคำบางส่วนสามารถเป็นคำประสมได้ ก็ไม่ได้เป็นหลักตายตัวว่ารูปนั้นจะเป็นฐานในการไปประสมคำต่อได้ทันที เราต้องพิจารณาความสัมพันธ์ที่แท้จริงภายในของการประสมคำเป็นหลักเพื่อหาโครงสร้างของคำประสมนั้น และการพิจารณาว่าส่วนใดสัมพันธ์กันนั้นก็ต้องอาศัยความหมายว่าในบริบทนั้นผู้พูดต้องการหมายถึงอะไร

การศึกษาโครงสร้างคำประสมจึงเป็นเรื่องที่น่าสนใจเพื่อจะได้เข้าใจว่า โครงสร้างที่เป็นไปได้ของคำประสมมีอะไรบ้าง มีความซับซ้อนได้เพียงใด เพื่อจะได้ใช้เป็นแนวทางในการประมวลผลภาษาไทยด้วยคอมพิวเตอร์ต่อไป

ความสัมพันธ์ทางความหมายภายในคำประสม

นอกจากงานวิจัยที่สนใจเรื่องการระบุหาขอบเขตของคำประสม (compound extraction) และการวิเคราะห์โครงสร้างภายในคำประสมแล้ว ยังมีงานวิจัยอีกจำนวนหนึ่งที่สนใจศึกษาความสัมพันธ์ของหน่วยศัพท์ที่อยู่ภายในคำประสม ซึ่งส่วนมากเป็นการมองความสัมพันธ์ทางความหมาย หรือ semantic relation ระหว่างคู่คำในคำประสมนั้น การกำหนดความสัมพันธ์ภายในคำประสมนั้นมีแนวคิดที่แตกต่างกัน ซึ่ง Spencer (2011) สรุปว่าแบ่งได้เป็นสองกลุ่ม กลุ่มแรกเริ่มโดยงานของ Lees (1960, อ้างถึงใน Spencer 2011) ที่มองความสัมพันธ์ระหว่างคำหลักและคำขยายเชิงอรรถศาสตร์ จึงมีความสัมพันธ์ที่จำกัด กลุ่มที่สองเริ่มโดย Downing (1977, อ้างถึงใน Spencer 2011) มองว่าความสัมพันธ์ระหว่างคำหลักและคำขยายนั้นเป็นความสัมพันธ์เชิงวิจนปฏิบัติศาสตร์จึงมีได้หลากหลายขึ้นกับการตีความที่เหมาะสมในบริบทนั้นๆ

งานวิจัยเหล่านี้ในทางภาษาศาสตร์มุ่งเน้นเพื่อทำความเข้าใจกลไกการสร้างความหมายของคำประสม คำประสมที่ไม่สามารถตีความหมายได้แน่ชัดจากรูปที่ปรากฏ มีลักษณะเป็นหน่วยศัพท์ชัดเจน เช่น "*piano bench*" ไม่ได้หมายถึงเก้าอี้สำหรับเล่นเปียโนเท่านั้น แต่เป็นเก้าอี้ลักษณะหนึ่งซึ่งจะใช้นั่งเล่นอะไรก็ได้ หรือ *blackbird* ก็ไม่ได้หมายถึงนกที่มีสีดำแต่เป็นนกประเภทหนึ่ง คำประสมบางพวกก็มีลักษณะที่เหมือนตีความบางส่วนได้จากองค์ประกอบ แต่ก็ไม่ได้มีเพียงความหมายจากองค์ประกอบแต่มีความหมายอื่นที่แสดงถึงความสัมพันธ์บางอย่างระหว่างคำนั้นด้วย เช่น "*bike girl*" หมายถึงผู้หญิงที่มีความเกี่ยวข้องกับบางอย่างกับจักรยาน เช่น อาจจะเป็นผู้หญิงที่ขี่จักรยานไปทำงาน อาจจะเป็นผู้หญิงที่ชอบประกอบจักรยานเล่น หรือ

อาจจะเป็นผู้หญิงที่เพิ่งจดทะเบียนไว้ข้างทาง ส่วนจะเป็นความหมายไหนนั้นก็เป็นเรื่องที่ต้องตีความในบริบทที่ใช้ ในขณะที่ถ้าเป็นนามวลีที่มีคำนามและคำขยาย ความหมายที่ได้เป็นการประกอบจากความหมายโดยตรงของแต่ละส่วนไม่ได้มีความหมายที่แปรไปตามบริบท เช่น “เสื้อขาว” ที่หมายถึงเสื้อที่มีสีขาว

ส่วนงานวิจัยเรื่องความสัมพันธ์ทางความหมายภายในคำประสมที่พบในงานทางภาษาศาสตร์คอมพิวเตอร์มีขึ้นเพื่อช่วยในการประมวลผลภาษาโดยเฉพาะอย่างยิ่งในภาษาเฉพาะทาง (technical text) ซึ่งมักมีชื่อที่มีลักษณะโครงสร้างซับซ้อน เช่น *"Open-labeled long-term study of the efficacy, safety, and tolerability of subcutaneous sumatriptan in acute migraine treatment"* (ดู Rosario and Hearst 2001) เป็นการตั้งชื่อที่มีโครงสร้างเป็นนามวลีซับซ้อนโดยที่ภายในไม่มีคำกริยาโดยตรงมีเพียงคำกริยาที่ถูกแปลงเป็นคำนาม มีการประสมคำและใช้บุพบท ขยาย สิ่งที่ต้องการคือการประมวลผลเพื่อให้เข้าใจความหมายของศัพท์เฉพาะทางนี้ โดยหาความสัมพันธ์ทางความหมายของคำย่อยภายใน จึงจำเป็นต้องมองหาความสัมพันธ์ทางความหมายระหว่างคู่คำภายในด้วย เช่น *"migraine treatment"* บอกความสัมพันธ์ที่เป็น treatment-for-disease ในขณะที่ถ้าเป็น *"aerosol treatment"* จะบอกความสัมพันธ์ที่เป็น method-of-treatment เป็นต้น

การกำหนดว่าความสัมพันธ์ทางความหมายมีอะไรบ้างเป็นเรื่องที่มีความเห็นแตกต่างกันดังที่กล่าวมาในตอนต้น บางคนก็มองว่ามีความสัมพันธ์ที่จำกัด เช่น Lees (1960, อ้างถึงใน Spencer 2011) หรือ Levi (1987, อ้างถึงใน Lauer 1995) Levi มองว่ามีจำนวนความสัมพันธ์ที่จำกัดในคำประสมแบบ N-N คือ IN, FOR, FROM, ABOUT, CAUSE, HAVE, MAKE, USE, BE เช่น *"electricity station"* มีความสัมพันธ์ MAKE เพราะเป็นสถานที่ก่อให้เกิดพลังงานไฟฟ้า ในกรณีที่เป็นคำประสมที่สร้างจากคำกริยาจะมีความสัมพันธ์ ACT, PRODUCT, AGENT, PATIENT. เช่น *"truck driver"* มีความสัมพันธ์ AGENT เพราะเป็นผู้กระทำการขับรถบรรทุก ความสัมพันธ์ลักษณะนี้

มาจากการที่ Levi ใช้แนวคิดไวยากรณ์อรรถศาสตร์เพิ่มพูนวิเคราะห์ว่ามีกฎการลดรูปจากอนุภาคยคุณศัพท์มาเป็นคำนามประสม (Singnoi 2000:186)

Lauer (1995) เสนอให้ใช้คำบุพบทแปดอย่างแทนความสัมพันธ์ทางความหมายของหน่วยศัพท์ในคำประสม คือ *of, for, with, in, on, at, about, and from* แต่ Girju et al. (2005) ว่าการให้ความสัมพันธ์ลักษณะนี้ไม่ละเอียดพอ และบางครั้งก็สามารถคิดได้หลายแบบ เช่น "*bird sanctuary*" หมายถึง *sanctuary for bird* หรือ *sanctuary of bird* ก็ได้ แต่ในขณะเดียวกัน การจะกำหนดความสัมพันธ์แบบละเอียดก็เป็นเรื่องยุ่งยากและจะทำให้มีประเภทความสัมพันธ์ออกมาเป็นจำนวนมาก เช่น "*salt water*" สามารถกำหนดว่ามีความสัมพันธ์เป็น "*dissolved in*" ก็ได้ จำนวนของความสัมพันธ์ทางความหมายจึงมีการนำเสนอที่แตกต่างกัน ไป เช่น Girju et al. (2005) ใช้ 35 ความสัมพันธ์ Rosario and Hearst (2001) ใช้ 38 ความสัมพันธ์ บางคนเช่น Downing (1977) อ้างถึงใน Lauer 1995 และ Spencer 2011) มองว่าความสัมพันธ์ในลักษณะนี้มีจำนวนที่ไม่จำกัด ซึ่ง Spencer (2011) ก็สนับสนุนความคิดนี้ โดยเสนอให้วิเคราะห์คำนามประสมแบบเข้าสู่ศูนย์ที่มีคำหนึ่งเป็นหลักและอีกคำเป็นคำขยายอย่าง A-B ว่ามีลักษณะที่หมายถึง $A(x) \& B(y) \& R(x,y)$ คือมีความสัมพันธ์ R บางอย่าง que เชื่อมโยงระหว่าง A และ B ความหมายของคำประสมจึงกำหนดชัดทั้งหมดจากความหมายขององค์ประกอบไม่ได้ แต่มีความหมายบางอย่างซึ่งขึ้นกับการใช้ในบริบทนั้น เช่น "*tea table*", "*coffee table*" ในบริบทที่เป็นห้องจัดเลี้ยงจะหมายถึงโต๊ะที่จัดเตรียมไว้สำหรับเสิร์ฟกาแฟหรือเสิร์ฟชา ไม่ได้หมายถึงประเภทย่อยของโต๊ะก็ได้ เป็นต้น

อย่างไรก็ตาม เนื่องจากในงานวิจัยของผู้เขียนไม่ได้สนใจหาความสัมพันธ์ทางความหมายภายในคำประสม คำถามที่ว่าจำนวนความสัมพันธ์ทางความหมายที่เหมาะสมเป็นเท่าไรจึงไม่ใช่ประเด็นคำถามที่สำคัญของงานนี้ อีกทั้งผู้เขียนก็เห็นด้วยกับ Spencer (2011) ว่าความหมายของความสัมพันธ์นั้นมิได้หลากหลายขึ้นกับการ

ตีความให้เหมาะสมกับบริบท ความสำคัญจึงไม่ได้ขึ้นกับจำนวนความสัมพันธ์ว่ามีอะไรบ้าง แต่สำคัญอยู่ที่ว่าจะมีกลไกขั้นตอนอย่างไรที่ทำให้สามารถตีความหมายที่ถูกต้องของคำประสมในบริบทได้

การศึกษาคำประสมในภาษาไทย

คำประสมเป็นคำประเภทหนึ่งที่ถูกกล่าวถึงในหนังสือไวยากรณ์ไทยโดยทั่วไป เพราะเป็นรูปแบบหนึ่งของการสร้างคำในภาษาไทยที่ใช้กันแพร่หลาย โดยมักกล่าวว่า คำประสมเกิดจากการนำคำมูลตั้งแต่สองคำมาประสมกัน ญัฐวุฒิ (2544:11) ได้สรุปการจำแนกประเภทของคำในตำราไวยากรณ์ไทยออกเป็นคำประเภทต่างๆ คือ คำมูลหรือคำเดี่ยว คำประสานคือคำที่เกิดจากหน่วยคำไม่อิสระ เช่น "นักเรียน" คำประสมคือคำที่เกิดจากคำมูลอย่างน้อยสองคำมาประกอบกัน คำซ้อนคือคำที่เกิดจากคำที่มีความหมายเหมือนหรือใกล้เคียงกันมาประกอบกัน เช่น "บอกกล่าว" คำซ้ำคือคำที่เกิดจากการนำคำเดียวกันหรือที่มีเสียงคล้ายกันมาประกอบกัน เช่น "แดงๆ" "โง่งมงั๊ง" แต่เนื่องจากความต้องการในการประมวลผลภาษาไทยที่ต้องการหาขอบเขตของคำแต่ละคำในข้อความ ญัฐวุฒิจึงมองประเภทของคำเป็นสองประเภท คือ คำเดี่ยวและคำประกอบ โดยที่คำประกอบคือคำที่เกิดจากการนำหน่วยคำอย่างน้อยสองหน่วยมาประกอบกัน คำประกอบจึงครอบคลุมทั้งคำประสาน คำประสม คำซ้อน และคำซ้ำ เพราะถูกมองเป็นปัญหาแบบเดียวกันคือการพิจารณาว่าจะรวมหน่วยคำที่เห็นประกอบเข้าเป็นคำหรือไม่

งานที่ศึกษาเรื่องคำประสมภาษาไทยในทางภาษาศาสตร์โดยทั่วไปจะสะท้อนความคิดที่แตกต่างกันของแนวทฤษฎีทางภาษาศาสตร์ที่นิยมศึกษากันในช่วงเวลานั้น เช่น งานของ Fasold (1968) มุ่งศึกษาคำประสมภาษาไทยโดยอาศัยกรอบทฤษฎีไวยากรณ์ปริวรรต โดยแยกคำนามประสมออกเป็นสองกลุ่ม คือ กลุ่มที่เป็น coordinate compound และกลุ่มที่แปลงมาจากนามวลีที่มีอนุพากย์วิเศษณ์

(relative clause) และวิเคราะห์ว่าคำประสมเหล่านี้เกิดจากกระบวนการปริวรรตที่มีการลดรูปคำต่างๆ ผ่านกฎการลบ (deletion) ต่างๆ

Singnoi (2000) ศึกษาโครงสร้างนามในภาษาไทย ซึ่งจำแนกเป็นคำนามนามวลี คำนามประสม และนามวลีแปลง (nominalization) โดยใช้กรอบทฤษฎีไวยากรณ์หน่วยสร้าง (construction grammar) ของ Fillmore Singnoi (2000) กล่าวถึงความกำกวมในภาษาไทยว่า รูปคำที่เห็นสามารถวิเคราะห์ได้หลายอย่างขึ้นอยู่กับบริบทการใช้ เช่น "บ้านเล็ก" อาจเป็นอนุพจน์หมายถึงบ้านที่มีขนาดเล็ก (ตัวอย่างเช่น "บ้านเล็กมากเลย คงอยู่กันไม่พอ") หรืออาจเป็นนามวลี หมายถึงบ้านขนาดเล็ก (ตัวอย่างเช่น "บ้านเล็กหลังนี้สวยมาก") หรืออาจเป็นคำประสม หมายถึง อนุภรรยา (ตัวอย่างเช่น "เขาเอาเงินไปให้บ้านเล็กหลายล้าน") ความกำกวมนี้เกิดจากการที่ทั้งสามกรณีมีลำดับคำเหมือนกันคือ บ้าน+เล็ก อย่างไรก็ตาม ในกรณีนี้จะเห็นว่าคำประสมเป็นกรณีที่มีความหมายต่างจากความหมายเดิมของคำที่นำมาประสม Singnoi (2000) ศึกษาคำนามประสมและแบ่งประเภทของคำประสมออกเป็น synthetic compound, noun-noun compound, verb-verb compound และ complex compound

synthetic compound ประกอบด้วยคำกริยาและคำนามที่มีความสัมพันธ์ทางความหมายกับคำกริยานั้น ตัวอย่างของคำนามประสมแบบ synthetic เช่น "คนดู" "ยานอนหลับ" "คนขับรถ" "ไก่ผัดพริกแกง" "บ้านพักตากอากาศ" โดยที่คำหลักจะเป็นคำซ้ายสุด (ไม่นับรวมคำประสมที่สร้างจากคำสันสกฤต เช่น "สังคมศาสตร์" ที่คำหลักเป็นคำขวาสุด) จึงมีรูปแบบเป็น N-VP อีกรูปแบบที่พบคือ V-N เช่น "พัดลม" "ผัดกะเพราไก่" หลังจากกระบวนแบบใหญ่สองประเภทนี้แล้ว Singnoi แยกประเภทย่อย โดยดูความสัมพันธ์ทางไวยากรณ์ว่าเป็น Subj, Obj, Oblique แล้วจึงแยกประเภทย่อยต่ออีกโดยอาศัยความสัมพันธ์ทางความหมายว่าเป็น Agent, Patient, Experiencer, Instrument, Location, ดังสรุปได้ในตารางข้างล่าง

การประมวลผลภาษาไทย

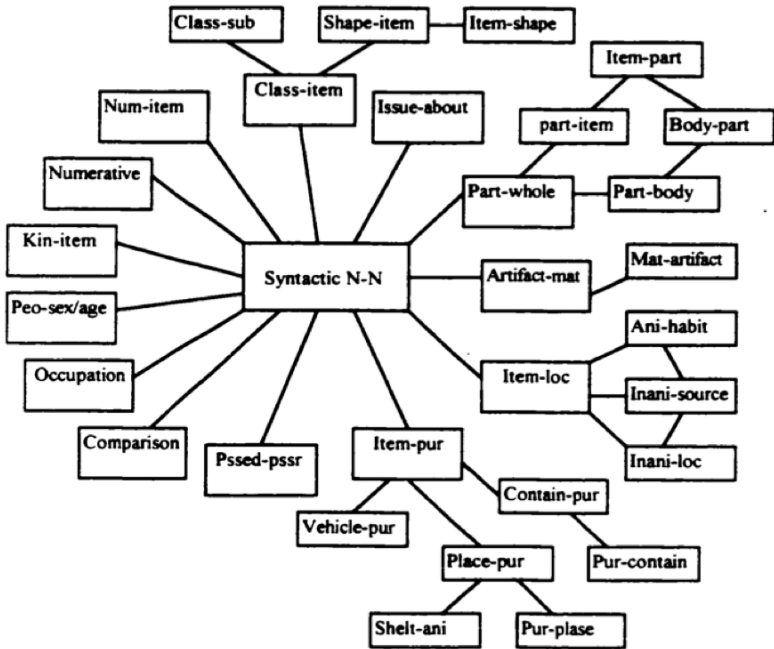
- Pattern	- Grammatical relation	- Semantic relation	- Example
- N-VP	- SUBJ-VP	- AGT-VP	- คนขับรถ, เรือบน
-	-	- EX-VP	- คนอ่าน, ประสาทสัมผัส
-	-	- INST-VP	- ไม้กวาด, ที่ตัดเล็บ
-	-	- PAT-VP	- กุ้งแห้ง, เมียน้อย
-	- OBJ-VP	- PAT-VP	- ปลานึ่ง, ปลานึ่งมะนาว
-	- OBL-VP	- LOC-VP	- ห้องรับแขก, ทางเดิน
-	-	- TEM-VP	- ฤดูเกี่ยวข้าว, เวลานอน
- V-N	- V-SUBJ	- V-AGT	- พัดลม, ชนไก่
-	-	- V-PAT	- ต่างจังหวัด, บ้าหมู
-	- V-OBJ	- V-PAT	- ผัดผัก, รองเท้า

ตารางสรุปประเภทและความสัมพันธ์ต่างๆใน Synthetic compound (Singnoi 2000:194-203)

noun-noun compound คำนามประสมที่พบมีทั้งที่เป็น endocentric compound ที่มีคำซ้ายเป็นหลัก (เช่น "ห้องครัว") ซึ่งในงานนี้เรียกว่า syntactic compound และที่เป็น exocentric compound (เช่น "พ่อแม่") โดยในงานนี้จะเรียกว่าเป็น semi-syntactic compound หรือ coordinate compound และที่เป็น asyntactic compound ซึ่งปัจจุบันมองไม่เห็นร่องรอยความสัมพันธ์ของคำที่มาประสมกัน เช่น "ผีเสื้อ" "ปากกา"

คำนามประสมส่วนใหญ่ที่พบจะเป็นแบบมีคำหลักอยู่ซ้าย และมีความสัมพันธ์ทางความหมายแบบต่างๆ เช่น CLASS-ITEM (ดอกกุหลาบ), CLASS-PROPER (ปลา

ดุก), ITEM-SHAPE (ไอติมแแห่ง), PART-WHOLE (ทัวสะพาน), ITEM-PART (ไม้ผล), BODY-PART (ว้วนม), ARTIFACT-MATERIAL (ตุ๊กตากระดาซ) เป็นต้น จำนวนความสัมพันธ์ที่ Singnoi นำเสนोजึงมีจำนวนมากถึง 30 ความสัมพันธ์ ดังแสดงในรูปนี้

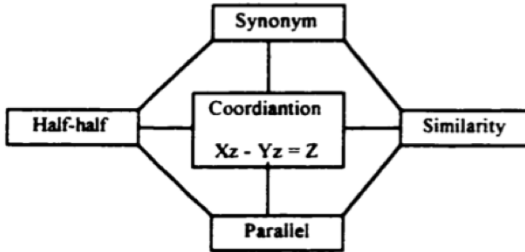


รูปแสดงความสัมพันธ์ของค่านามประสม noun-noun ที่มีคำหลัก (Singnoi 2000:240)

สำหรับคำประสมที่เป็น coordinate compound คือไม่มีคำใดเป็นคำหลักแต่ทั้งสองคำมีความสัมพันธ์เท่าเทียมกัน Singnoi พบความสัมพันธ์สี่อย่างตามลำดับความคล้ายกันของทั้งสองคำ คือ synonym (เช่น "ร่างกาย", "ภูเขา"), similarity (เช่น

การประมวลผลภาษาไทย

"ไขมัน", "เรือนหอ"), parallel (เช่น "พ่อแม่", "ลมฝน"), และ half-half (เช่น "ข้าวแกง", "ขามะนาว")



รูปแสดงความสัมพันธ์ของคำนามประสม noun-noun ที่ไม่มีคำหลัก (Singnoi 2000:246)

คำนามประสมอีกกลุ่มคือที่เกิดจากการประสม V-V Singnoi ว่าคำกลุ่มนี้มีจำนวนไม่มากและไม่ได้เกิดขึ้นมาก (productive) ทำกับการประสมคำ N-N แบบแรก คำกลุ่มนี้ได้แก่ "กำปั้น", "กันสาด", "พิมพ์ดีด", "ห่อหมก", "ต้มยำ", "ผัดเผ็ด", "เปรี้ยวหวาน", "ซอหนา", "วังเปรี้ยว", "ร้อยกรอง" (Singnoi 2000:256)

คำนามประสมอีกกลุ่มคือ คำประสมแบบซับซ้อน ซึ่งเกิดจากการประสมคำหลายระดับ คำกลุ่มนี้พบเป็นจำนวนมาก จัดเป็นคำที่เรียกว่า "compound within compound" เช่น "หนังสือเรียน" [[N-N]-V], "รองเท้าผ้าใบ" [[V-N]-[N-N]], "โรงเรียนสอนคนตาบอด" [[N-N]-[V-[N-[N-V]]]] Singnoi (2000: 256-264) ใช้กรอบการวิเคราะห์คำประสมที่กล่าวมาสำหรับคำประสมซับซ้อนนี้ โดยจัดเป็น synthetic complex compound ซึ่งมีโครงสร้างแบบคำประสมรวมกับคำกริยา compound-V เช่น [ลูกชิ้น]ทอด [[N-N]-V], [[[ลูกชิ้น][ปลากระป๋อง]][ซุบแห้ง]]ทอด [[[N-N]-[N-N]]-[V-N]]-V เป็นต้น และที่เป็นคำประสมซับซ้อนจากคำนาม หรือ N-N complex compound เช่น ทาง[ม้าลาย] เป็นคำนามรวมกับคำนามประสม N-compound, [ส้มตำ]ไทย เป็นคำนามประสมรวมกับคำนาม compound-N, [กองพัน][ทหารบก]

เป็นคำนามประสมรวมกับคำนามประสม compound-compound เป็นต้น สำหรับความสัมพันธ์ทางความหมายก็สามารถวิเคราะห์โดยใช้ความสัมพันธ์ที่พบในคำประสมที่กล่าวมาได้

งานของ Singnoi (2000) เป็นงานที่ศึกษาคำนามประสมโดยละเอียดอย่างเป็นระบบ จัดกลุ่มคำประสมโดยดูจากหมวดคำของคำที่มาประกอบเป็นหลัก โดยแบ่ง synthetic compound (N-V, V-N), noun-noun compound (N-N), verb-verb compound (V-V) และ complex compound คือที่ประสมหลายชั้น แล้ววิเคราะห์ความสัมพันธ์ทางความหมายภายในคำประสมแบบต่างๆ แต่ก็ยังมีคำนามประสมบางส่วนที่ไม่ครอบคลุมอยู่ในงาน เช่น คำประสมที่คำหลักไม่ได้อยู่ซ้ายอันเนื่องมาจากการสร้างคำตามหลักภาษาสันสกฤต เช่น "กุศลกรรม" มีคำหลักคือ "กรรม" คำประสมในกลุ่มที่เป็น neo-classical compound ซึ่งมีส่วนประกอบคำจากภาษาบาลีสันสกฤตนี้จึงอยู่นอกเหนือขอบข่ายการศึกษาของ Singnoi นอกจากนี้ คำประสมที่มีหน่วยศัพท์เป็นคำบุพบท เช่น "นางใน" "คนนอก" ก็ไม่มีการกล่าวถึงชัดเจน แต่ก็มีความเป็นไปได้ว่าคำบุพบทเหล่านี้จะถูวิเคราะห์ให้เป็นคำนามแทนและจัดเข้ากลุ่มการแยกประเภทที่กล่าวมาได้

การระบุหาคำประสมในภาษาไทย

คำถามสำคัญในการศึกษาคำประสมภาษาไทยมีสามคำถาม ซึ่งก็เป็นคำถามเดียวกับการศึกษาคำประสมในภาษาอื่นๆ คำถามแรกคือจะหาขอบเขตของคำประสมได้อย่างไร คำถามที่สองคือจะแยกความแตกต่างของคำประสมออกจากวลีหรือประโยคอย่างไร และคำถามที่สามคือความสัมพันธ์ภายในคำประสมเป็นอย่างไร มีโครงสร้างและความหมายอย่างไรได้บ้าง งานวิจัยของ Aroonmanakun (2015) สนใจคำถามแรกเป็นหลักคือปัญหาของการหาขอบเขตคำประสมในภาษาไทย ซึ่งปัญหานี้ไม่ได้มีเหตุจากการที่ภาษาไทยไม่มีขอบเขตคำชัดเจนเหมือนภาษาอื่นๆ เพราะภาษาที่มีการเขียนวรรณยุกต์อย่างภาษาอังกฤษก็ยังมีคำถามเช่นเดียวกันนี้ว่าจะรวบคำที่คำ

เข้าด้วยกันให้เป็นคำประสม นอกจากนี้ การตัดสินว่ารูปคำที่เห็นเป็นคำประสมเลยหรือไม่ ก็ไม่สามารถทำได้ทุกกรณีโดยไม่อาศัยบริบท เพราะคำประสมเกิดจากการนำคำมูลสองคำขึ้นไปมาประกอบกัน ในบางบริบท รูปคำที่เห็นเหมือนเป็นคำประสมอาจถูกใช้เป็นคำมูลสองคำก็ได้ เช่น ในประโยค "เขาเข้าไปในป่าแล้วไปพบทางเสือที่ขาดอยู่กับพื้น" "ทางเสือ" ในที่นี้ก็ไม่ใช้คำนามประสม หรือในตัวอย่าง "ไข่เปิด" "บ้านเล็ก" ที่ Singnoi (2000) ยกมาเป็นตัวอย่างแสดงความกำกวมว่าอาจจะเป็นคำประสมหรือนามวลีก็ได้ นอกจากนี้ คำประสมบางคำก็มีโครงสร้างที่ซับซ้อนไม่แตกต่างไปจากโครงสร้างวลี ตัวอย่างเช่น "โรงเรียนสอนคนตาบอด", "ลูกขึ้นปลากลายซุบแบ่งทอด" เป็นต้น คำถามที่สองที่ว่าคำประสมแตกต่างจากวลีหรือประโยคอย่างไรนั้นจึงมีความสำคัญและเกี่ยวข้องโดยตรงต่อการตอบคำถามแรก ส่วนคำถามที่สามนั้น แม้ไม่ใช่คำถามหลักของงานวิจัยนี้ แต่ก็อาจจะนำมาใช้ประโยชน์เพื่อตอบคำถามแรกได้ ดังนั้น การวิเคราะห์คำนามประสมในภาษาไทยในที่นี้จึงเริ่มด้วยการพิจารณาคำถามที่สามก่อน คือ การวิเคราะห์ภายในคำประสม อย่างไรก็ตาม ในที่นี้จะจำกัดเฉพาะการวิเคราะห์โครงสร้างของคำประสมเท่านั้น จะยังไม่นำเรื่องความสัมพันธ์ทางความหมายมาพิจารณาเพราะเหตุว่าเป็นเรื่องของการตีความความหมายที่ขึ้นกับบริบทดังที่กล่าวมาแล้ว

จากที่กล่าวมาข้างต้น งานวิจัยนี้มีวัตถุประสงค์เพื่อสกัดคำประสมภาษาไทยจากคลังข้อมูล และจะเริ่มด้วยการวิเคราะห์โครงสร้างที่เป็นไปได้ในคำประสมภาษาไทยก่อน เพื่อจะได้นำความรู้นี้มาพิจารณาว่าจะหาขอบเขตคำประสมได้อย่างไร แต่เนื่องจากเราต้องตัดสินก่อนว่าคำใดเป็นคำประสมจึงจะนำมาวิเคราะห์โครงสร้างภายในได้ จึงเหมือนเป็นปัญหวนซ้ำไม่มีจุดเริ่มต้น ในที่นี้จึงแก้ปัญหาโดยนำรายการคำที่พบในพจนานุกรมไทยมาใช้ เพราะถือว่าเป็นข้อมูลที่ผ่านการวิเคราะห์และคัดเลือกมาก่อนแล้วว่าเป็นรายการคำ จากนั้นจึงเขียนโปรแกรมเพื่อแจกแจงว่าในแต่ละคำที่พบในรายการคำของพจนานุกรมนั้น สามารถมองรูปที่ปรากฏว่าประกอบขึ้น

จากคำย่อยๆ ตามรายการคำที่มีอยู่อย่างน้อยสองคำขึ้นไปได้ใหม่ ตัวอย่างเช่น คำว่า "กฎแจปากตาย" เมื่อเทียบในรายการคำที่มีพบว่า สามารถมองได้ว่าประกอบด้วยรูปคำสามคำ คือ "กฎแจ"+"ปาก"+"ตาย" หรือคำว่า "คุ้มครอง" เมื่อเทียบกับรายการคำทั้งหมดที่มีอยู่พบว่าสามารถแยกเป็นคำย่อยได้สองคำ คือ "คุ้ม"+"ครอง" คำที่สามารถมองแยกย่อยในลักษณะนี้ได้จะถูกสกัดออกมา เพื่อนำมาคัดเลือกภายหลังว่าเป็นคำประสมที่เกิดจากการประประสมคำย่อยแบบนั้นจริงหรือไม่ ในตัวอย่างที่ยกมาจะเห็นว่า "กฎแจปากตาย" เป็นคำประสมที่เกิดจากการประสมคำ "กฎแจ"+"ปาก"+"ตาย" แต่คำ "หน้าที่" แม้ว่าจะมองแยกเป็น "หน้า"+"ที่" ได้ แต่ไม่สามารถมองเป็นคำประสมที่เกิดจากการประสมของคำสองคำ คือ "หน้า"+"ที่" ได้ ซึ่งจากการวิเคราะห์ข้อมูลรายการคำทั้งสิ้น 32,653 คำ พบว่าสามารถมองแบบแยกองค์ประกอบย่อยได้ทั้งหมด 18,738 รูปแบบ โดยที่คำบางคำอาจถูกมองว่าประกอบด้วยคำย่อยๆ ได้หลายแบบ และจาก 18,738 รูปแบบที่พบนั้น เมื่อพิจารณา ดูแล้วมีที่เป็นรูปแบบคำประสมจริง 4,687 คำ

จากนั้น จึงนำคำประสมทั้งหมด 4,687 คำที่ได้มาวิเคราะห์และจัดกลุ่ม โดยจะวิเคราะห์ประเด็นต่อไปนี้เป็น จำนวนคำที่นำมาประสม หมวดคำของคำย่อยภายในคำประสมนั้น ประเภทของคำประสมว่าเป็น endocentric หรือเป็น exocentric ในส่วนของ exocentric compound หากเป็นประเภทย่อย coordinate compound หรือ reduplication ก็จะไปประเภทย่อยเหล่านี้ด้วย เหตุที่แยกคำประสมเหล่านี้ออกมาเพราะเป็นคำประสมที่มีลักษณะเฉพาะที่น่าจะใช้ประโยชน์ในการสกัดหาคำประสมได้ คำซ้ำในภาษาไทยในที่นี้จัดเป็นกลุ่มคำประสมแบบที่สร้างคำใหม่ด้วยการซ้ำรูปคำทั้งหมดหรือบางส่วน เช่น "ต่อปากต่อคำ" "เห็นนอกเห็นใจ" เป็นต้น ส่วนคำซ้อนในภาษาไทยเป็นการสร้างคำใหม่ด้วยวิธีการนำคำที่มีความหมายทำนองเดียวกันมาประสมกัน เช่น "เต็นรำ" "หลอกหลวง" "หลักเกณฑ์" จึงจัดให้อยู่ในกลุ่มคำประสมแบบ coordinate ซึ่งโดยหลักการจะมาจากการสร้างคำโดยประกอบคำ

ที่มีฐานะเท่าเทียมกันคือเป็นคำที่มองได้ว่ามีคำจากกลุ่มร่วมกัน เช่น “พ่อแม่” “เปรี้ยวหวาน” นอกจากนี้ ในที่นี้ ไม่ได้แยกความต่างระหว่าง coordinate compound กับ appositive compound เพราะถือว่าในเรื่องรูปนั้นมีกลไกเดียวกัน ความต่างที่ว่าอาจอ้างถึงสองสิ่งหรือสิ่งเดียวไม่ได้มีความสำคัญต่อการสกัดหาคำประสม ส่วนคำประสม exocentric ประเภทอื่นอย่าง synthetic compound หรือ bahuvrihi compound โดยรูปนั้นคล้ายกับ endocentric compound คือมีคำหลักชัดเจน เพียงแต่ความหมายของคำประสมไม่ได้มีคำหลักเป็นแกนของความหมาย การวิเคราะห์นี้จะโยงไปกับการวิเคราะห์โครงสร้างแบบฟังก์ชันด้วย คือจะวิเคราะห์หาความสัมพันธ์แบบฟังก์ชันระหว่างคำย่อยสองคำที่สัมพันธ์กัน ดังนั้นในคำประสมที่มีคำมาประสมมากกว่าสองคำก็จะเห็นโครงสร้างและความสัมพันธ์มากกว่าหนึ่งความสัมพันธ์ได้ ตัวอย่างเช่น “กระดาดเงินกระดาดทอง” วิเคราะห์ได้ว่าประกอบด้วยคำสี่คำ คือ “กระดาด”+“เงิน”+“กระดาด”+“ทอง” โดยที่ “กระดาด” คำแรกมีความสัมพันธ์แบบ endocentric กับ “เงิน” โดยมี “กระดาด” เป็นคำหลัก และ “กระดาด” คำหลังมีความสัมพันธ์แบบ endocentric กับ “ทอง” โดย “กระดาด” เป็นคำหลัก เมื่อรวมกันทั้งสองคู่แล้ว “กระดาดเงิน” มีความสัมพันธ์แบบซ้ำคำกับ “กระดาดทอง” เพราะเป็นการซ้ำรูปคำหลักคือ “กระดาด”

ผลจากการวิเคราะห์คำประสมทั้งสิ้น 4,687 คำ พบว่ามีเพียง 256 คำ ที่ประกอบด้วยคำย่อยสามคำขึ้นไป ที่เหลือ 4,431 คำเป็นคำประสมที่เกิดจากคำย่อยสองคำ คำประสมสี่คำที่พบเกือบทั้งหมดมีลักษณะเป็นคำซ้ำของคำประสมสองคำที่มีคำหลักซ้ำกันคือมีรูปแบบ $(X+Y)+(X+Z)$ เช่น “กระดาดเงินกระดาดทอง” “กล้าได้กล้าเสีย” “ขูดเลือดขูดเนื้อ” “เขารู้ปเขารู้ย” “ต่อปากต่อคำ” “สูบเลือดสูบเนื้อ” “หมดไส้หมดพุง” “ออกดอกออกผล” เป็นต้น คำประสมห้าคำที่พบในพจนานุกรมสองตัวอย่างก็มีลักษณะเดียวกัน คือ “ค่าน้ำร้อนน้ำชา” “เงินน้ำห่าน้ำหก” จากข้อมูลที่พบ จึงเห็นได้ว่า ในพจนานุกรมราชบัณฑิตยสถานยังไม่มีคำประสมใหม่ๆ ที่มีจำนวน

คำสี่คำขึ้นไป คำที่น่าจะจัดได้ว่าเป็นคำประสมได้ เช่น “หม้อหุงข้าวไฟฟ้า” “เครื่องกรองอากาศ” “เครื่องทำน้ำอุ่น” ฯลฯ

- คำประสม	- จำนวน	-
- 2 คำ	- 4431	- 94.54%
- 3 คำ	- 140	- 2.99%
- 4 คำ	- 114	- 2.43%
- 5 คำ	- 2	- 0.04%
-	- 4687	- 100%

และเมื่อวิเคราะห์ความสัมพันธ์ระหว่างคำภายในเฉพาะคำประสมสองคำ ทั้งหมด 4,431 คำ ก็พบว่า ส่วนใหญ่เป็นความสัมพันธ์แบบ endocentric 2,345 คู่ และความสัมพันธ์แบบ coordinate 209 คู่ แบบคำซ้ำ 189 คู่ และแบบ exocentric 1,808 คู่ นอกจากนี้ ยังพบว่าเกือบทั้งหมดเป็นคำซึ่งมีคำหลักอยู่ซ้าย (ในกรณีที่ไม่มีคำหลักชัดเจน ในงานวิจัยนี้วิเคราะห์ความสัมพันธ์พึ่งพาโดยให้คำซ้ายเป็นคำหลัก) คำประสมที่มีส่วนของคำหลักอยู่ขวามี 34 คำ เช่น “เทพ”+”นินาย”, “คริสต์”+”ศักราช”, “สุภาพ”+”บุรุษ”, “ไฟร์”+”พล” เป็นต้น ซึ่งมักเป็นคำที่มาจากภาษาบาลีสันสกฤตและวางลำดับคำต่างจากคำไทย

จากผลการวิเคราะห์โครงสร้างทำให้เห็นว่าคำประสมที่มีอยู่ในพจนานุกรมส่วนใหญ่เป็นคำประสมสองคำมากที่สุด คำประสมที่เกิดจากสามคำขึ้นไปนั้นมีไม่มาก แต่จากที่กล่าวมาแต่แรกว่าปัญหาหนึ่งของการประมวลผลภาษาไทย คือ มีการสร้างคำประสมใหม่ๆ เกิดขึ้น ซึ่งจะยังไม่ปรากฏอยู่ในรายการคำของพจนานุกรม และคำประสมใหม่เหล่านี้ก็มีโครงสร้างที่ซับซ้อนมากกว่าการประสมคำสองคำ การสกัดคำ

ประสมใหม่จึงเป็นสิ่งที่จำเป็นเพื่อจะได้บรรจุคำเกิดใหม่เหล่านี้ไว้ในพจนานุกรมเพื่อใช้ประมวลผลภาษาไทยได้ แต่ปัญหาสำคัญที่ตามมา คือ คำถามที่สองที่ว่า จะแยกความแตกต่างระหว่างคำประสมออกจากวลีหรือประโยคได้อย่างไร หากว่าเมื่อทั้งสองกรณีมีรูปคำเหมือนกันเหมือนในตัวอย่าง "ไช่เปิด" "ไฟไหม้" ที่ Singnoi (2000) และ Prasithrathsint (2010) ให้ตัวอย่างไว้ ซึ่งการที่จะตัดสินได้ว่าเป็นคำประสมหรือวลีหรือประโยคนั้นต้องพิจารณาบริบทที่ปรากฏ ลักษณะนี้เป็นเรื่องปกติของภาษาไทย ตัวอย่างเช่น "ผ้าเช็ดตัว" หากปรากฏในประโยค "เขาใช้ผ้าเช็ดตัวจนสะอาด" ก็จะไม่ถูกวิเคราะห์ว่าเป็นคำประสม แต่วิเคราะห์ว่าเป็นคำสามคำ คือ "ผ้า" "เช็ด" "ตัว" และประโยคนี้นี้เป็นการใช้กริยาเรียง หากคิดว่าเป็นเพราะโครงสร้างลักษณะนี้ (ใช้+Obj+V+N) ทำให้ต้องตีความเป็นการใช้กริยาเรียง ก็ไม่ใช่ เพราะโครงสร้างแบบนี้เหมือนกันในประโยค "เขาใช้น้ำยาล้างจานจนหมด" ก็สามารถวิเคราะห์ว่า "น้ำยาล้างจาน" เป็นคำประสมได้ ซึ่งในตัวอย่างนี้ สามารถวิเคราะห์ได้ทั้งสองทาง คือมองเป็นคำประสม "น้ำยาล้างจาน" หรือเป็นการใช้กริยาเรียง "ใช้" + "น้ำยา" + "ล้าง" + "จาน" ลักษณะเช่นนี้ทำให้ต้องกลับมาคิดทบทวนคำถามที่สองที่ว่า จะแยกระหว่างคำประสมและวลีหรือประโยคอย่างไร จริง ๆ แล้วเป็นคำถามที่ควรถามหรือมีคำตอบชัดเจนหรือไม่ เพราะในความจริง อาจจะไม่มีเส้นแบ่งชัดเจนระหว่างคำประสมและวลีหรือประโยค การตัดสินเป็นเรื่องที่อิงกับบริบทและความหมายที่ต้องการสื่อในบริบทนั้น

ในตัวอย่างที่ Singnoi ยกมาว่า "ไช่เปิดฟองนี้" "ไช่เปิด" เป็นคำประสมเพราะใช้ลักษณนาม "ฟอง" สำหรับคำว่า "ไช่" ส่วนใน "ไช่เปิดตัวนี้" จะไม่เป็นคำประสมแต่เป็นนามวลี เพราะใช้ลักษณนาม "ตัว" สำหรับ "เปิด" ความจริงหากพิจารณาว่าทั้งสองกรณีเป็นนามวลีที่ต่างกันเพียงว่าคำลักษณนามและคำชี้เฉพาะจะขยายคำนามตัวไหน "ไช่" หรือ "เปิด" หรือตัวอย่างที่ Prasithrathsint ยกมา คำว่า "ไฟไหม้" อาจจะไม่ใช่คำประสมก็ได้ เพราะหากใช้เกณฑ์เดียวกับตัวอย่าง "ไช่เปิด" คำลักษณนาม

ของ “ไฟ” คือ “กอง” คำลักษณะนามที่สัมพันธ์กับกริยา “ไหม้” คือ “ครั้ง” คำลักษณะนามของกรรมที่เกิดกับกริยา “ไหม้” คือ “แห่ง” ดังนั้น ในตัวอย่าง “มีไฟไหม้ห้าแห่งในเมือง” “ไฟไหม้” จึงอาจจะไม่ใช่คำประสม เพราะ “ห้าแห่ง” บอกถึงความสัมพันธ์กับนามวลีที่เป็นกรรมคือสถานที่ที่ถูกเผาไป เราจึงอาจจะวิเคราะห์ให้เป็นประโยคเหมือนตัวอย่าง “ไฟไหม้อาคารพังลงมา” ก็ได้

นอกจากนี้ ความสัมพันธ์ทางความหมายภายในคำประสมก็ไม่ได้มีจำนวนจำกัดหรือชัดเจน คำประสม A-B มีลักษณะที่หมายถึง $A(x) \& B(y) \& R(x,y)$ คือมีความสัมพันธ์ R บางอย่างที่โยงระหว่าง A และ B ความหมายของคำประสมจึงกำหนดชัดจากความหมายขององค์ประกอบไม่ได้ แต่มีความหมายบางอย่างซึ่งขึ้นกับการใช้ในบริบทนั้น เช่น “เค้กมะพร้าว” อาจไม่ได้ใช้หมายถึงเค้กที่มีส่วนผสมของมะพร้าว แต่ในบางบริบทอาจหมายถึงเค้กที่ทำมา มีรูปทรงแบบมะพร้าวก็ได้

ปัญหาต่างๆ ที่ยกมากล่าวนี้ ทำให้เห็นว่าการหาเส้นแบ่งที่ชัดเจนระหว่างคำประสมและวลีนั้นเป็นเรื่องที่ไม่อาจจะระบุได้โดยปราศจากบริบทการใช้ แต่หากถามว่า ถ้าเช่นนั้น จะวิเคราะห์ให้คำประสมเป็นวลีที่มีโครงสร้างภายในเลยได้ไหม หากทำแบบนี้ เราก็จะรู้สึกว่ามันไม่เหมาะสำหรับคำประสมบางคำซึ่งถูกใช้จนเป็นเสมือนคำเดียวไปแล้ว เช่น “รองเท้า” “เงินเดือน” “แมวน้ำ” “ไฟฟ้า” “ดีใจ” เป็นต้น

คำประสมที่เป็นคำสองคำจะมีความเป็นเนื้อเดียวหรือเป็นหน่วยศัพท์มากกว่า คำที่ประกอบด้วยคำหลายๆคำ ยังมีจำนวนคำมากขึ้นความเป็นหน่วยศัพท์หรือเป็นคำประสมก็จะลดลง เช่น “คนขับรถ” มีความเป็นคำประสม มากกว่า “คนขับรถบรรทุก” มากกว่า “คนขับรถบรรทุกพลังงาน NGV” เป็นต้น ดังนั้น ความแตกต่างระหว่างคำประสมและวลีจึงมีลักษณะที่เป็นลำดับขั้นหรือตีกริมากกว่าที่จะเป็นการตัดสินว่าใช่หรือไม่ใช่ด้วย นอกจากนี้ ในกรณีที่คำหลาย ๆ คำมาประสมกันดังที่ได้กล่าวมาแล้วในตอนต้นว่า จะต้องมีการวิเคราะห์โครงสร้างและความสัมพันธ์ทางความหมายขององค์ประกอบต่าง ๆ โดยต้องพิจารณาถึงบริบทที่ปรากฏด้วยว่าเป็น

การประมวลผลภาษาไทย

อย่างไร จึงจะตัดสินได้ว่าควรวិเคราะห์ความสัมพันธ์ทางความหมายออกมาในลักษณะใด หากเป็นเช่นนี้ การประมวลผลคำประสมในภาษาไทยจึงเป็นส่วนที่ผูกเป็นเนื้อเดียวกับการแจ่งส่วน (parsing) ประโยคภาษาไทย เพราะต้องมองโครงสร้างประกอบกับการตีความและตัดสินไปพร้อมกับคำอื่นๆในบริบท ซึ่งจะได้กล่าวโดยละเอียดในบทต่อไป

อ้างอิง

- Aronoff, Mark, and Fudeman, Kirsten Anne. (2011). What is morphology?: Fundamentals of linguistics. Chichester, West Sussex, U.K. ; Malden, MA: Wiley-Blackwell.
- Bauer, L. (2008). Exocentric compounds. *Morphology*, 18, 51-74.
- Booij, G. E. (2007). The grammar of words : an introduction to linguistic morphology: Oxford textbooks in linguistics. New York: Oxford University Press.
- Bloomfield, Leonard. (1933). Language. New York,: H. Holt and Company.
- Downing, P. (1977). On the Creation and Use of English Compound Nouns. *Language*, 53(4), 810-842. <http://eric.ed.gov/?id=EJ173732>
- Dressler, W.U. (2006). Compound Types. In The representation and processing of compound words, eds. Gary Libben and Gonia Jarema, xv, 242 p. Oxford ; New York: Oxford University Press.
- Holmer, Arthur J. (1996). A Parametric Grammar of Seediq. Lund: Lund University Press.
- Fasold, Ralph William August. (1968). Noun compounding in Thai, University of Chicago: Ph.D. Dissertation.

- Girju, Roxana, Moldovan, Dan, Tatu, Marta, and Antohe, Daniel. (2005). On the semantics of noun compounds. *Computer Speech and Language* 19:479-496.
- Lauer, Mark. (1995). *Designing Statistical Language Learners: Experiments on Noun Compounds*, Department of Computing, Macquarie University: Dissertation.
- Lees, Robert B. (1960). *The grammar of English nominalizations*. The Hague: Mouton.
- Levi, Judith N. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.
- Prasithratsint, Amara. (2010). *Lexicalization of syntactic constructions in Thai*. Paper presented at The 20th Anniversary Meeting of the Southeast Asian Linguistics Society, Zurich, Switzerland
- Ralli 1992, อ้างถึงใน Booij 2007:88
- Rosario, Barbara, and Hearst, Marti. (2001). Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 82-90.
- Singnoi, Unchalee. (2000). *Nominal Constructions in Thai*, Department of Linguistics, University of Oregon: Ph.D. Dissertation.
- Spencer, Andrew. (2011). What's in a compound?. *Journal of Linguistics*,47, pp481-507 doi:10.1017/S0022226710000411

การประมวลผลภาษาไทย

ณัฐวุฒิ ไชยเจริญ. (2544). การตัดคำและการกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จด้วยคอมพิวเตอร์, ภาควิชาภาษาศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย: วิทยานิพนธ์มหาบัณฑิต.

การสกัดคำประสมภาษาไทย⁴

ในบทนี้ จะกล่าวถึงภาพรวมของงานทางภาษาศาสตร์คอมพิวเตอร์ที่ได้มีการศึกษาเรื่องคำประสมว่ามีการศึกษาในเรื่องใดประเด็นใดบ้าง แนวคิดที่ใช้คืออะไร จากนั้นจึงเสนอแนวคิดที่จะใช้ในงานนี้เพื่อสกัดหาคำประสมภาษาไทย

ความเป็นมา

งานวิจัยทางภาษาศาสตร์คอมพิวเตอร์ที่เกี่ยวข้องกับคำประสมสามารถจำแนกเป็นหลายประเภท บางงานสนใจศึกษาการสกัดว่าขอบเขตของคำนามประสมอยู่ที่ใดหรือมองหาวามีส่วนใดบ้างที่สามารถเป็นคำประสมได้ เช่น Sornlertlamvanich and Tanaka (1996), Sornlertlamvanich et al. (2000), Aroonmanakun (2009) งานกลุ่มนี้เป็นงานต่อเนื่องและเกี่ยวข้องกับงานการตัดคำ เนื่องจากเห็นว่าในภาษานั้นมีกระบวนการสร้างคำใหม่ๆ เกิดขึ้นผ่านการประสมคำอยู่เสมอ ซึ่งเป็นปัญหาสำหรับการตัดคำเนื่องจากคำใหม่เหล่านี้ยังไม่มีกรบรรจุไว้ในพจนานุกรม จึงจำเป็นต้องสกัดคำใหม่ ๆ ที่เกิดขึ้นนี้ให้ได้ ในขณะที่บางงานจะสนใจหาโครงสร้างของคำประสมและหาความสัมพันธ์ทางความหมายภายในคำประสม เช่น Levi (1978), Rosario and Hearst (2001), Girju et al. (2005), Kriengkiet et al. (2007) ด้วยเหตุว่าต้องการประมวลผลความหมายของคำประสมเหล่านั้นสำหรับใช้ในงานต่าง ๆ เช่น การแปลง

⁴ เรียบเรียงใหม่จากรายงานวิจัยเรื่อง “การสกัดคำประสมภาษาไทย” โครงการกลุ่มนวัตกรรมวิชาการเชิงบูรณาการ โครงการในแผนพัฒนาวิชาการจุฬาลงกรณ์มหาวิทยาลัย ๒๕๕๔-๒๕๕๖

เป็นรูปแทนความหมาย การแปลภาษา การหาโครงสร้างเป็นส่วนสำคัญในการกำหนดหาความหมายเพราะจากตัวอย่างที่ผ่านมาเราได้เห็นแล้วว่าเมื่อคำประสมที่มีรูปคำเดียวกันแต่มีโครงสร้างต่างกัน ความหมายก็จะต่างกันไป ส่วนการหาความหมายภายในคำประสมนั้นก็ยังคงเป็นงานที่มีการกำหนดความสัมพันธ์ทางความหมายที่แตกต่างกันไป ซึ่งก็เป็นเรื่องปกติเพราะแม้แต่การกำหนดรายการหมวดคำซึ่งมีความชัดเจนมากกว่า ก็ยังพบว่าระบบงานแต่ละที่ก็จะมีการใช้รายการหมวดคำที่แตกต่างกันไปตามความเชื่อหรือกรอบทฤษฎีที่ใช้วิเคราะห์

ในงานวิจัยที่ทำนี้ มีจุดมุ่งหมายหลักคือสกัดหาคำประสมแต่อาศัยความสัมพันธ์ทางโครงสร้างมาใช้ประโยชน์ในการสกัดคำประสมด้วย การหาคำประสมโดยมากใช้วิธีการมองรูปคำที่อยู่ติดกันอยู่เสมอ โดยอ้างอิงหลักที่ว่าคำประสมนั้นมีลักษณะเป็นหน่วยที่ไม่สามารถแทรกคำอื่นเข้าไปได้โดยไม่เสียความหมายเดิม ดังนั้นรูปคำที่มักพบอยู่ด้วยกันเสมอจึงเป็นข้อมูลส่วนที่น่าจะพบคำประสมได้ หลักการนี้นำมาใช้ได้ด้วยการใช้วิธีการทางสถิติแบบต่างๆ เช่น mutual information, log-likelihood, etc. อย่างไรก็ตามการมองรูปคำที่ติดกันนี้เป็นการมองรูปแบบคำประสมแบบที่เป็น linear ซึ่งจากการศึกษาคำประสมทางภาษาศาสตร์ เราจะเห็นว่าคำประสมที่ประกอบด้วยหน่วยศัพท์มากกว่าสองคำขึ้นไปจะมีความสัมพันธ์ทางโครงสร้างที่มองเป็นลำดับชั้นได้ การพิจารณาแบบ linear จึงอาจจะไม่ครอบคลุมคำประสมที่ซับซ้อนได้ ในงานนี้จึงต้องการอาศัยความสัมพันธ์ทางโครงสร้างมาใช้ด้วย โดยเลือกวิธีการมองความสัมพันธ์แบบพึ่งพาเนื่องจากเชื่อว่าจะสอดคล้องกับการประมวลผลภาษาไทยที่มีลักษณะที่ยืดหยุ่นได้มากกว่าการมองแบบหน่วยประกอบ (constituent)

แนวคิดสำคัญ

การจะให้คอมพิวเตอร์สามารถสกัดค่านามประสมภาษาไทยได้ ในงานนี้จะมุ่งไปที่วิธีการแบบมีการชี้แนะหรือ supervised learning คือให้คอมพิวเตอร์เห็นตัวอย่างของคำประสมจำนวนหนึ่งที่มากพอสมควร เพื่อฝึกให้เรียนรู้ลักษณะที่สำคัญและ

นำมาใช้ในการสกัดคำนามประสมใหม่ที่ยังไม่พบมาก่อนได้ แต่เนื่องจากการระบุว่าคำใดเป็นคำประสมเป็นคำถามสำคัญของงานนี้ ผู้เขียนจึงเลี่ยงการใช้ความเห็นของตัวเองในการเริ่มต้นกำหนดว่าคำใดเป็นคำประสม ในงานนี้จึงจะอาศัยตัวอย่างคำประสมเริ่มต้นจากพจนานุกรมมาวิเคราะห์หาความสัมพันธ์ทางโครงสร้าง เพราะถือว่ารายการคำที่ปรากฏในพจนานุกรมนั้นถูกคัดมาแล้วว่าเป็นคำ ดังที่ได้กล่าวมาแล้วในบทก่อน งานนี้จึงเป็นแบบ example-based ก่อน ไม่ได้ใช้ corpus-based ที่ผู้เขียนเข้าไปกำหนดขอบเขตคำประสมขึ้นมาจากข้อมูลจริงที่พบแต่แรก เพราะปัญหาการตัดสินใจว่าคำไหนเป็นคำประสมนั้นเป็นปัญหาตั้งต้นที่ต้องหาเกณฑ์การพิจารณาให้ชัดเจนก่อน การวิเคราะห์จากรายการคำที่อยู่ในพจนานุกรมจึงให้ข้อมูลพื้นฐานที่จะนำมาใช้ทำความเข้าใจและกำหนดเกณฑ์การพิจารณาคำประสมเพื่อใช้ในการกำกับข้อมูลในภาษาจริงภายหลัง

ในเรื่องของการวิเคราะห์ความสัมพันธ์ภายในคำประสม เนื่องจากงานวิจัยนี้มีเป้าหมายเพื่อหาขอบเขตคำประสม โดยจะใช้ลักษณะโครงสร้างภายในคำประสมมากกว่าการมองแบบความต่อเนื่องของคำแบบเรียงลำดับ จึงจะอาศัยเรื่องของ dependency relation เพื่อกำหนดส่วนที่มีความสัมพันธ์กัน สิ่งที่เป็นความสัมพันธ์เชิงความหมายหรือ semantic relation นี้จึงไม่ใช่ประเด็นสำคัญในงานนี้เพราะในตอนนี้ยังไม่มีจุดมุ่งหมายที่จะวิเคราะห์ความหมายของคำประสมเพื่อไปใช้ในงานอื่นๆ อีกทั้งการที่จะหาความสัมพันธ์ทางความหมายได้นั้น จำเป็นต้องอาศัยข้อมูลลักษณะที่เป็น Thai WordNet ที่จะใช้อ้างอิงถึงกลุ่มความหมายต่าง ๆ ของคำต่าง ๆ เพื่อใช้พิจารณาความสัมพันธ์ทางความหมายได้ดีขึ้น ข้อมูลพื้นฐานที่สามารถใช้ประโยชน์ได้จะมีก็เพียงหมวดคำเท่านั้น ในงานวิจัยนี้ จึงจะพิจารณาในกรอบของความสัมพันธ์ทางวากยสัมพันธ์เป็นหลัก โดยจะวิเคราะห์ความสัมพันธ์แบบพึ่งพาระหว่างหมวดคำที่พบภายในคำประสม นอกจากนี้ความสัมพันธ์ที่นักภาษาศาสตร์กล่าวถึงในบทที่ผ่านมา นั้น ที่แยกเป็นความสัมพันธ์แบบ endocentric และ exocentric ก็จะสามารถนำมาใช้

วิเคราะห์ด้วย โดยในส่วนของความสัมพันธ์แบบ exocentric จะแยกพวกที่เป็น coordinate compound อย่าง “พ่อแม่” ออกมาซึ่งในที่นี้จะจัดรวมประเภท appositive compound อย่างเช่น “นักเรียนนักศึกษา” (ที่อ้างถึงคนเดียวกัน) เข้าไว้ด้วย เพราะมีความต่างเพียงแค่ว่าคำที่มาเทียบเคียงกันนั้นอ้างถึงสิ่งเดียวกันหรืออ้างถึงหลายสิ่งซึ่งเป็นความต่างในการอ้างถึงเป็นหลักไม่ใช่เรื่องทางวากยสัมพันธ์ ส่วนคำซ้อน (semantic reduplication) ซึ่งบางคนก็จะแยกกลไกนี้เป็นวิธีการหนึ่งต่างหากจากการประสมคำ แต่ในที่นี้ เนื่องจากว่าในภาษาไทยเป็นการนำคำมาประกอบกันจึงรวมไว้เป็นประเภทหนึ่งของคำประสมแบบ exocentric ด้วย เพราะคำเหล่านี้หากพิจารณาดูก็จะเห็นว่าในภาษาไทยเกิดจากกระบวนการนำคำมากกว่าสองคำมาประกอบกันเพียงแต่เป็นคำประเภทที่มีความหมายเหมือนหรือคล้ายกัน ซึ่งในงานนี้จะจัดเข้าไว้กับกลุ่ม coordinate compound และ copulative compound เพราะมองว่าโดยรูปแบบเป็นการนำคำที่อยู่ในกลุ่มเดียวกันมาประสมกัน เช่น “ด้านทาน” “ถกเถียง” ซึ่งในทางรูปแล้วไม่ต่างไปจากคำประสมแบบ coordinate และแบบ appositive ที่เอาคำที่อยู่ในกลุ่มเดียวกันมาประสมอย่าง “พ่อแม่” หรือ “เปรี้ยวหวาน” ในที่นี้จึงรวมเรียกเป็นคำประสมแบบเคียงกัน (coordinate compound) นอกจากนี้ กรณีของคำซ้ำรูป (reduplication) ก็จะวิเคราะห์ให้เป็นคำประสมประเภทหนึ่งด้วยเพราะเป็นการสร้างคำใหม่ด้วยการซ้ำรูปเดิมทั้งหมดหรือบางส่วน

ในการกำกับคำหลักคำขยาย ความสัมพันธ์แบบ exocentric บางประเภทอย่าง bahuvrihi compound และ synthetic compound ก็จะกำกับส่วนคำหลักคำขยายได้เหมือนกับคำประสมแบบ endocentric แม้ว่าในทางความหมาย คำประสมนั้นจะไม่ได้ใช้คำหลักเป็นแก่นหรือความหมายหลัก

ในการวิเคราะห์คำประสมที่ได้จากพจนานุกรมนี้ เพื่อให้ได้ภาพโดยรวมของคำประสมในภาษาไทย ผู้เขียนจึงได้วิเคราะห์ความสัมพันธ์ของคำประสมทุกหมวดคำ

หมวดคำของคำประสมที่ได้ อาจเหมือนหรือต่างจากหมวดคำของคำย่อยภายในคำประสมก็ได้ ความสัมพันธ์ที่ใช้วิเคราะห์คำประสมจึงมีดังนี้

คู่ความสัมพันธ์ที่เป็นไปได้ระหว่างคำภายในคำประสม

- NN ใช้กำหนดความสัมพันธ์ของ noun กับ noun เช่น “ใจคอ”, “แม่ทัพ”, “เอกสารราชการ” ซึ่งโดยมากแล้วคำซ้ายมักจะเป็นคำหลัก
- NA ใช้กำหนดความสัมพันธ์ของคำหลักที่เป็น noun โดยที่มีคำ adjective มาขยายได้ เช่น “คลื่นยาว”, “คอแข็ง”
- NV ใช้กำหนดความสัมพันธ์ที่มีคำกริยาเป็นหลักและมีคำนามเป็น subject argument ทางซ้าย เช่น “ขวัญหาย”, “กาฝาก”
- NP ใช้กำหนดความสัมพันธ์ของคำที่เป็น noun กับ preposition เช่น “เครื่องใน”, “คนไข่นอก”, “เหล็กใน”
- VV ใช้กำหนดความสัมพันธ์ของคำย่อยที่เป็น verb กับ verb เช่น “พิมพ์ดีด”, “กดขี่”
- VN ใช้กำหนดความสัมพันธ์ที่มีคำกริยาเป็นหลักและมีคำนามเป็น object argument ทางขวา เช่น “คืนคำ”, “จิกหัว”
- VA ใช้กำหนดความสัมพันธ์ของคำหลักที่เป็น verb ที่มีคำอื่นมาขยายได้ เช่น “คิดสั้น”
- VP ใช้กำหนดความสัมพันธ์ของคำที่เป็น verb กับ preposition เช่น “ตกใน”, “หลับใน”, “เข่าถึง”
- AA ใช้กำหนดความสัมพันธ์ของคำที่เป็น adjective ทั้งคู่ เช่น “คดงอ”, “ยากเย็น”, “ลึกลับ”
- AN ใช้กำหนดความสัมพันธ์ของคำที่เป็น adjective กับ noun เช่น “เบาใจ”, “เสมอหน้า”, “คล่องตัว”
- AV ใช้กำหนดความสัมพันธ์ของคำที่เป็น adjective กับ verb เช่น “ดีแตก”, “กลมกลืน”, “ผิดสังเกต”
- PA ใช้กำหนดความสัมพันธ์ของคำที่เป็น preposition กับ adjective พบเพียงคำเดียว คือ “ข้างเคียง”
- PN ใช้กำหนดความสัมพันธ์ของคำที่เป็น preposition กับ noun พบเพียงคำเดียวโดยเป็นส่วนของคำคู่หลัง คือ “ออกนอกหน้า”

เมื่อแจกแจงจำนวนคู่ความสัมพันธ์ทั้งหมดจะได้ 5,059 คู่คำที่ปรากฏในคำประสม ตามการเกิดร่วมกันของหมวดคำข้างบน ได้ดังตารางนี้

- ประเภท	- จำนวน	- ร้อยละ
- Endo	- 2587	- 51.14%
- Exo	- 2074	- 41.00%
- Red	- 106	- 2.10%
- CC	- 292	- 5.77%
-	- 5059	- 100.00%

ตาราง 1 : จำนวนความสัมพันธ์ภายในคำประสมแยกตามประเภท

ประเภทของคำประสมทั้งหมด 4,687 คำที่พบ (รวมคำประสมแบบสองคำถึงห้าคำ) เมื่อแยกดูความสัมพันธ์ของแต่ละคู่คำย่อยภายในคำประสมนั้นได้ทั้งสิ้น 5,059 คู่ วิเคราะห์ได้เป็นประเภทความสัมพันธ์ต่างๆ ดังนี้

Endo ใช้กำหนดความสัมพันธ์คำประสมแบบที่เป็น endocentric

CC ใช้กำหนดความสัมพันธ์แบบที่เป็น coordinate compound และ appositive compound และคำซ้อนความหมาย

Redup ใช้กำหนดความสัมพันธ์ของคำที่เป็นคำซ้อนหรือคำซ้ำ

Exo ใช้กำหนดความสัมพันธ์แบบที่เป็น exocentric ที่ไม่ใช่ CC หรือ Redup ข้างบน

- คู่มวลคำ	- จำนวน	- ร้อยละ
- AA	- 131	- 2.59%
- AN	- 225	- 4.45%
- AV	- 39	- 0.77%
- NA	- 494	- 9.76%
- NN	- 1396	- 27.59%

- NP	- 8	- 0.16%
- NV	- 617	- 12.20%
- PA	- 1	- 0.02%
- PN	- 1	- 0.02%
- VA	- 170	- 3.36%
- VN	- 1323	- 26.15%
- VP	- 8	- 0.16%
- VV	- 646	- 12.77%
-	- 5059	- 100.00%

ตาราง 2 : จำนวนความสัมพันธ์ระหว่างสองคำแยกตามชนิดของคำ

ข้อมูลที่ได้จากการวิเคราะห์คำประสมนี้จะช่วยให้เห็นรูปแบบและความถี่ของความสัมพันธ์แบบต่างๆ ซึ่งจะถูกนำมาใช้ในการสกัดคำนามประสมต่อไป

การสกัดคำประสม

การสกัดคำประสมในที่นี้หมายถึง การให้คอมพิวเตอร์อ่านข้อความภาษาไทย ซึ่งได้มีการแยกคำออกมาแล้ว แต่เนื่องจากคำประสมจำนวนหนึ่งยังไม่มีการบรรจุไว้ในพจนานุกรม จึงทำให้มีการแยกคำผิดพลาดไป การสกัดคำประสมจะมองหาสายคำที่ คาดว่าน่าจะรวมเป็นคำประสมได้ เดิมผู้เขียนคิดว่าจะมองหาหรือสกัดคำประสมที่ ประกอบด้วยคำมากกว่าสองคำทั้งหมด โดยไม่จำกัดว่าจะมีจำนวนคำมาประกอบ ทั้งหมดกี่คำ แต่หลังจากการทบทวนวรรณกรรมและพิจารณาลักษณะความแตกต่างของคำประสมและวลีแล้ว พบว่าไม่มีเส้นแบ่งที่ชัดเจนระหว่างคำประสมและวลี ยังมีจำนวนคำมากเท่าใดความเป็นเนื้อเดียวของคำประสมหรือความเป็นหน่วยศัพท์นั้นก็ จะยิ่งลดลงดังที่ได้กล่าวมาแล้วในบทที่แล้ว อีกทั้งความกำกวมระหว่างคำประสมและวลีไม่สามารถตัดสินได้โดยการพิจารณาเพียงรูปคำนั้นอย่างเดียว จำเป็นต้องพิจารณาจากบริบทที่เกิดเป็นสำคัญ การตีความหมายของคำประสมจำนวนมากก็ไม่ได้มี

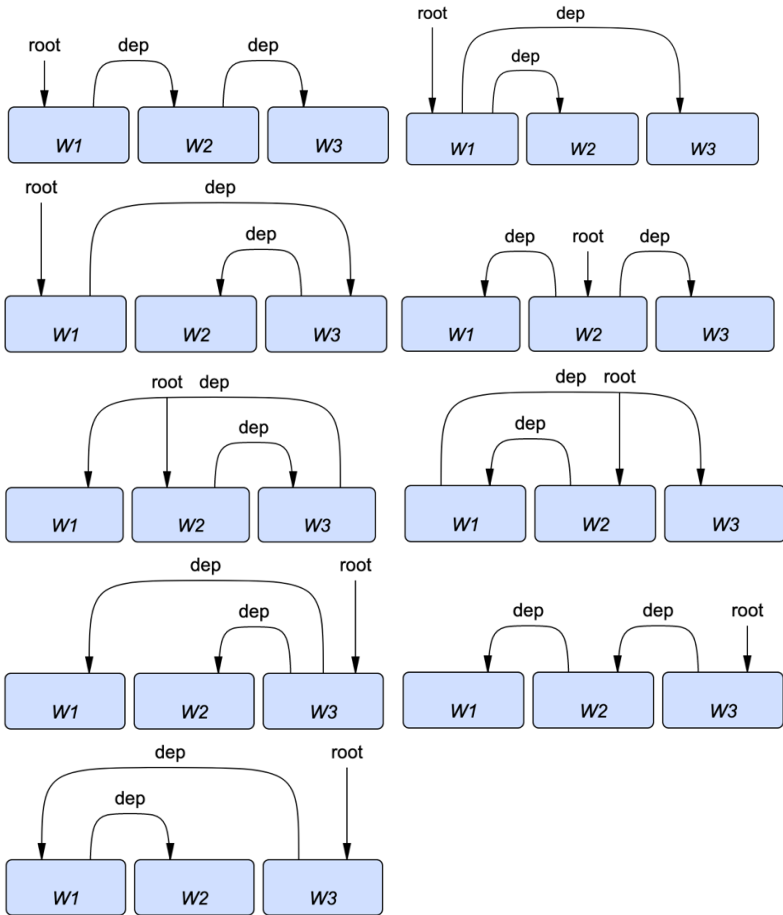
ความหมายตายตัวแต่เป็นความหมายที่ดีความให้เหมาะกับบริบทที่ใช้ ดังนั้น การประมวลผลคำประสมจึงน่าจะเป็นส่วนเดียวกันกับการแจงส่วนประโยค (parsing) โดยในกระบวนการนั้นจะวิเคราะห์โครงสร้างและความหมายที่เกิดขึ้นในประโยคนั้น ทุกความเป็นไปได้ เช่น “คนขับรถบรรทุก” อาจเป็นคำประสมที่มี “คน” เป็นคำหลักเชื่อมต่อกับส่วนอื่นของประโยค หรืออาจจะเป็นอนุภาคที่มี “ขับ” เป็นคำหลักและคนกับรถบรรทุกเป็นประธานและกรรมตามลำดับ แล้ว “ขับ” ไปเชื่อมโยงกับส่วนอื่นของประโยคต่อไป อย่างไรก็ตาม ไม่ได้หมายความว่าคำประสมทุกคำนั้น ควรประมวลผลในลักษณะเช่นนี้ เพราะอย่างที่กล่าวมาแต่แรกว่าความแตกต่างระหว่างคำประสมและวลีเป็นเรื่องของลำดับชั้นมากกว่าการชี้ชัดว่าเป็นหรือไม่เป็น คำประสมจำนวนมากโดยเฉพาะคำที่ประสมจากสองหรือสามคำแล้วมีความหมายที่ต่างไปจากความหมายเดิมของคำย่อยมาก ๆ คำเหล่านี้จะมีความเป็นหน่วยศัพท์สูงและผู้พูดจะรู้สึกว่าเป็นคำเดียวมากกว่าจะคิดว่าประกอบจากหลายคำ เช่น “ดีใจ” “หม้อน้ำ” “เงินเดือน” “เหตุผล” คำเหล่านี้จึงควรบรรจุในพจนานุกรมเพื่อใช้ประมวลผลภาษา ในขณะที่คำว่า “ไขเบ็ด” “ไฟไหม้” ถึงแม้จะไม่มีในพจนานุกรมก็ยังสามารถประมวลผลและสร้างความหมายจากคำย่อยได้

ด้วยเหตุนี้ การพัฒนาระบบสกัดคำประสมในที่นี้จึงเปลี่ยนจากการสกัดหาคำประสมทุกลักษณะ มาเป็นการมองปัญหาเรื่องนี้เป็นสองงาน งานประเภทแรกคือการสกัดหาคำประสมแท้ที่มีความเป็นหน่วยศัพท์สูง คำประสมประเภทนี้จะมีความหมายที่ไม่สามารถดึงจากคำย่อยได้ชัดเจนจัดเป็นคำประสมแบบทึบ (opaque compound) ต่างจากคำประสมแบบใส (transparent compound) ที่สามารถตีความหมายจากความหมายของคำย่อยได้ (เพียรศิริ 2531) งานประเภทที่สองคือการประมวลผลคำประสมที่พบในประโยคโดยอาศัยความรู้ทางโครงสร้างและสร้างความหมายของคำประสม A-B จากความสัมพันธ์ $A(x) \& B(y) \& R(x,y)$ โดยมองหาความหมาย R ที่เป็น

ความสัมพันธ์ระหว่าง A กับ B ที่เกิดขึ้นในบริบทนั้น โดยในที่นี้ จะจำกัดที่งานประเภทแรก คือ สกัลดาคำประสมแบบที่บจากคลังข้อมูลภาษาไทย

จากการวิเคราะห์คำประสมที่พบในพจนานุกรม ส่วนมากแล้วเป็นคำประสมสองคำ 94.71% มีคำประสมสามคำบ้างและสี่คำบ้าง (2.84% และ 2.41%) ดังนั้น ในการสกัลดาคำนามประสมนี้จะจำกัดจำนวนคำที่มาประสมไม่เกินสี่คำ เมื่อกำหนดขอบเขตงานได้แล้ว ขั้นตอนต่อไปคือพิจารณาผลจากการวิเคราะห์ว่าข้อมูลใดบ้างที่สามารถนำมาใช้เพื่อออกแบบและพัฒนากระบวนการสกัลดาคำนามประสมในภาษาไทยได้ ซึ่งจากการวิเคราะห์ข้อมูลคำประสมในพจนานุกรม เราจะเห็นความแตกต่างของจำนวนความสัมพันธ์ระหว่างหมวดคำที่มาประสมกันเป็นคำหลักคำขยายได้ เช่น NN พบมากกว่า NV เป็นต้น ดังนั้น ค่าความน่าจะเป็นของหมวดคำที่มาประสมกันจึงเป็นข้อมูลหนึ่งที่น่าจะใช้ได้ นอกจากนี้ รูปคำบางคำก็ช่วยตัดสินใจได้ เช่น คำประสมที่คำหลักอยู่ขวา เช่น X-สิทธิ์ X-ศักราช หรือที่เป็นคำประสมทั่วไปคำหลักอยู่ซ้าย เช่น เกิน+Y ใจ+Y เครื่อง+Y เป็นต้น ความน่าจะเป็นของรูปคำหลักจึงเป็นอีกข้อมูลหนึ่งที่สามารถนำมาใช้ได้ นอกจากนี้ในคำประสมที่ประสมมากกว่าสองคำ ความเป็นไปได้ของโครงสร้างแบบต่าง ๆ ก็น่าจะเป็นอีกข้อมูลหนึ่งที่จะต้องพิจารณา เช่น คำประสมสามคำ $w_1-w_2-w_3$ อาจมีโครงสร้างแบบพึงพาได้ 9 แบบ ($w_1 \rightarrow w_2, w_2 \rightarrow w_3$) หรือ ($w_1 \rightarrow w_2, w_1 \rightarrow w_3$) หรือ ($w_1 \rightarrow w_3, w_2 \rightarrow w_3$) หรือ ($w_1 \leftarrow w_2, w_2 \rightarrow w_3$) หรือ ($w_1 \leftarrow w_2, w_1 \rightarrow w_3$) หรือ ($w_1 \leftarrow w_3, w_2 \rightarrow w_3$) หรือ ($w_1 \leftarrow w_3, w_2 \leftarrow w_3$) หรือ ($w_1 \leftarrow w_2, w_2 \leftarrow w_3$) หรือ ($w_1 \rightarrow w_2, w_1 \leftarrow w_3$) ก็ได้ แต่โครงสร้างแบบไหนบ้างที่พบในภาษาไทย และพบเป็นจำนวนมากน้อยแค่ไหน ความเป็นไปได้ของแต่ละแบบนั้นก็จะเป็นข้อมูลในการวิเคราะห์คำประสม การเข้าคำก็เป็นอีกข้อมูลหนึ่งที่สามารถนำมาใช้ในกรณีของคำประสม

การประมวลผลภาษาไทย



นอกจากข้อมูลเกี่ยวกับรูปคำ หมวดคำ และโครงสร้าง ที่ได้จากการวิเคราะห์คำ ประสมที่มีอยู่ก่อนแล้ว เราควรจะพิจารณาความหมายของคำประสมว่ามีความแตกต่างไปจากการรวมกันของความหมายของคำย่อย ๆ แต่เนื่องจากเราไม่มีข้อมูลความหมายโดยตรง วิธีหนึ่งที่สามารถทำได้คือ อาศัยคำบริบทข้างเคียงที่เกิดร่วมกับ

คำประสมนั้น โดยมองว่า หาก A-B เป็นคำประสมแท้แล้ว คำที่ใช้ร่วมกับ A-B นั้น ควรจะแตกต่างจากคำที่ใช้ร่วมกับ A หรือใช้ร่วมกับ B เพียงคำเดียว หากใช้แนวคิดนี้ เราสามารถใช้ context vector เป็นตัววัดความแตกต่างระหว่างบริบทที่ A-B ปรากฏ กับบริบทที่ A หรือ B ปรากฏได้

ดังนั้น ในระบบการสกัดคำนามประสมจึงควรจะต้องประกอบด้วยสองส่วนหลัก ส่วนแรกเป็นการคำนวณรูปแบบโครงสร้างและความสัมพันธ์ฟังก์ชันที่เป็นไปได้มากที่สุดของชุดคำใดๆ ตั้งแต่ 2 คำขึ้นไป เพื่อจะตอบคำถามหลักว่าหากชุดคำที่พบเป็นคำประสมแล้ว ชุดคำเหล่านั้นจะมีโครงสร้างและความสัมพันธ์ฟังก์ชันรูปแบบไหน อะไรเป็นคำหลักคำขยาย หมวดคำของคำย่อยควรเป็นอะไร ส่วนที่สองเป็นการพิจารณาความหมายหรือบริบทของคำที่ปรากฏร่วมกับคำประสมนั้น กับที่ปรากฏกับคำย่อยๆ แต่ละคำว่ามี ความแตกต่างกันหรือไม่ ส่วนที่สองใช้ยืนยันว่าความหมายหรือบริบทการใช้คำประสมนั้นแตกต่างจากบริบทการใช้ของแต่ละคำย่อยๆ อย่างชัดเจน จึงควรพิจารณาให้เป็นคำประสม

การเตรียมคลังข้อมูล

จากแนวคิดพื้นฐานที่กล่าวมา ข้อมูลที่จำเป็นต้องใช้สำหรับการพัฒนาระบบ จะมีสองส่วน ส่วนแรกเป็นรายการคำประสมที่ได้มาจากพจนานุกรมด้วยวิธีการที่กล่าวมาแล้ว นำมาวิเคราะห์โครงสร้างแบบฟังก์ชัน และหมวดคำของแต่ละคำย่อย เพื่อสร้างเป็นข้อมูลตัวอย่างของคำประสมภาษาไทย ซึ่งจะนำไปคำนวณค่าความน่าจะเป็นสำหรับการพบคู่คำและคู่หมวดคำในคำประสม ข้อมูลส่วนนี้คือข้อมูลคำประสม 4,687 คำตามที่ได้กล่าวมาแล้วในบทก่อน ส่วนที่สองเป็นคลังข้อมูลภาษาไทยขนาดใหญ่ เพื่อนำมาใช้เทียบหาความเหมือนหรือแตกต่างของบริบทการใช้คำ ในที่นี้จะใช้ข้อมูลจากคลังข้อมูลภาษาไทยแห่งชาติ ซึ่งมีจำนวนประมาณ 33 ล้านคำ

ระบบการสกัดคำนามประสม

จากแนวคิดพื้นฐานที่กล่าวมาข้างต้น ระบบที่พัฒนาในที่นี้ใช้เพื่อสกัดคำประสมที่มีลักษณะเป็นคำประสมแท้ ประกอบด้วยคำย่อย 2-3 คำเพื่อนำมาบรรจุเพิ่มเติมในรายการคำของพจนานุกรม ระบบประกอบด้วยสองส่วน ดังนี้

1. การวิเคราะห์โครงสร้างและความสัมพันธ์ภายใน

Aroonmanakun (2009) ทดลองหากลุ่มคำที่น่าจะเป็นประสมได้ โดยพิจารณาคำที่อยู่ติดกัน 2-3 คำ และคำนวณหาความน่าจะเป็นสองส่วน คือ พิจารณาจากรูปปรากฏว่าคำสองคำนั้นมีความเหนียวแน่นหรือการปรากฏร่วมกันมากน้อยเพียงใด ส่วนนี้คำนวณโดยใช้ log-likelihood หาค่า collocation strength ระหว่างคำสองคำนั้น กรณีที่เป็นกลุ่มคำสามคำก็คำนวณจากค่าเฉลี่ยของ collocation strength ของ w_1-w_2 กับ w_2-w_3 ส่วนที่สองเป็นการให้น้ำหนักกับหมวดคำ เนื่องจากการวิเคราะห์คำประสมข้างต้นได้แสดงให้เห็นว่าคำในหมวดคำต่าง ๆ มีที่มาประสมกันเป็นคำประสมมีน้ำหนักมากน้อยแตกต่างกัน ส่วนที่สองนี้จึงนำค่าความน่าจะเป็นของคู่หมวดคำต่างๆ ในคำประสมมาคือน้ำหนักด้วย⁵

ผลจากการทดลองใน Aroonmanakun (2009) เมื่อคำนวณแล้วมาคัดกลุ่มคำที่ได้ค่าสูงสุดออกมา 200 อันดับแรก พบว่าวิธีการนี้ทำให้ได้คำประสมสองคำเพียง 63% และคำประสมสามคำเพียง 35% เหตุส่วนหนึ่งมาจากคำประสมสองหรือสามคำที่สกัดมานั้นเป็นเพียงส่วนประกอบของคำประสมที่ยาวกว่านั้น เช่น “คำ-เสรี” ไม่เป็นคำประสมเพราะถูกสกัดมาจากคำ “การ-คำ-เสรี” เมื่อทดลองหาคู่คำใหม่ โดย

⁵ ในการคำนวณจริงสามารถประมาณค่าความน่าจะเป็นของหมวดคำที่พบในข้อมูลเทียบกับหมวดคำที่พบในคำประสมจริงได้ แต่ในงาน Aroonmanakun (2009) เนื่องจากยังไม่ได้ใช้โปรแกรม POS tagger กับข้อมูลที่ใช้ จึงใช้วิธีประมาณการจากหมวดคำที่เป็นไปได้ทั้งหมดของคู่คำนั้นแทน

เลือกเฉพาะคำที่สัดส่วนค่าความเป็นคำประสมสองคำกับความเป็นคำประสมสามคำมากกว่าค่าที่กำหนดตามสูตรข้างล่าง ผลของคำสองคำที่สกัดมา 200 อันดับแรกด้วยวิธีการใหม่นี้พบว่า เป็นคำประสมเพิ่มขึ้นเป็น 71.50%

$$\frac{\text{score}(x, y)}{\text{Max}(\text{score}(w_i, x, y) | \text{score}(x, y, z_j))} > t$$

จากการทดลองนี้ จึงได้ว่าวิธีการดูการปรากฏร่วมกันของรูปคำประกอบกับความน่าจะเป็นของหมวดคำของเหล่านั้นสามารถนำมาใช้คัดเลือกคำที่น่าจะเป็นคำประสมได้ (compound candidates) แต่ต้องดูเทียบไม่ให้คำที่สกัดมานั้นเป็นเพียงส่วนของคำประสมที่ใหญ่กว่า ในกรณีที่ เป็นคำประสมซับซ้อนที่มีสามคำขึ้นไป หากมีข้อมูลการวิเคราะห์โครงสร้างการพึงพาทภายในของคำประสม ในกระบวนการสกัดหา กลุ่มคำที่น่าจะเป็นคำประสมได้ก็สามารถนำค่าความน่าจะเป็นทางโครงสร้างมาประกอบในการพิจารณาได้ กระบวนการทั้งหมดนี้ทำให้ได้กลุ่มคำที่น่าจะเป็นคำประสมได้

2. การวิเคราะห์บริบทการใช้

เมื่อได้คำตั้งแต่สองคำขึ้นไปที่อาจเป็นคำประสมได้ ขั้นตอนต่อไปคือการพิจารณาความหมายของคำประสม หากเป็นคำที่มีความหมายต่างไปจากความหมายของคำย่อยก็น่าเชื่อถือมากขึ้นว่าจะเป็นคำประสม แต่เนื่องจากเราไม่สามารถหาความหมายของกลุ่มคำหรือคำย่อยได้โดยตรง จึงใช้วิธีอ้อมที่ใช้ความหมายจากเวกเตอร์ของคำแทนหรือที่เรียกว่า vector semantics วิธีการนี้คือการดูคำไปพร้อมกับบริบทการปรากฏของคำนั้น ๆ ซึ่งเป็นแนวคิดพื้นฐานที่นักภาษาศาสตร์ เช่น Harris และ Firth เองก็เคยกล่าวไว้ หลักการง่าย ๆ คือ ถ้าคำสองคำมีบริบทการใช้ที่เหมือนหรือคล้าย ๆ กัน ก็แสดงว่าสองคำนั้นมีความหมายคล้ายกัน (*words that occur in the same contexts tend to have similar meanings*, Harris 1954) ดังนั้น ถ้าคำ x-y เป็นคำประสม บริบทการใช้ของ x-y ก็ควรแตกต่างไปจากบริบทการใช้ของคำ x หรือ y

การนำข้อมูลการปรากฏของคำพ้องบริบทในขอบเขตหนึ่งมาแปลงให้เป็นเวกเตอร์จึงเป็นวิธีการที่ช่วยให้เราดูคำประสมในมิตินี้ได้ ถ้ากำหนดให้ $V(w_1..w_n)$ such that $2 \leq n \leq 4$ เป็นเวกเตอร์ของความถี่ของ content word ที่ปรากฏในบริบท c คำหน้าหรือหลังคำ $w_1..w_n$ และ $V(w_i)$ such that $i=1,4$ เป็นเวกเตอร์ของความถี่ของ content word ที่ปรากฏในบริบท c คำหน้าหรือหลังคำ w_i หาก $\cosine(V(w_1..w_n), V(w_i)) < \text{ค่าที่กำหนด}$ ให้ถือว่า $w_1..w_n$ เป็นคำประสมที่มีโครงสร้างและความสัมพันธ์ภายในตามที่ได้จากข้อ 1

เวกเตอร์คำบริบท

ในส่วนนี้จะกล่าวถึงแนวคิดในการใช้คำบริบทเพื่อดูความคล้ายคลึงทางความหมายของคำประสมกับคำย่อยต่างๆ โดยนำแนวคิดเรื่อง vector space model มาใช้เพื่อแปลงบริบทที่คำนั้นปรากฏให้เป็นเวกเตอร์ทางคณิตศาสตร์ จากนั้นจึงคำนวณความคล้ายคลึงผ่านการวัดระยะห่างของเวกเตอร์ที่ได้

การศึกษาธรรมชาติของคำประสมสองคำโดยใช้เวกเตอร์คำบริบท

การศึกษาคำประสมมีประเด็นหนึ่งที่เป็นเรื่องของการตัดสินใจว่ามีคำใดเป็นคำหลักหรือไม่ ซึ่งคำหลักจะเป็นคำที่เป็นแกนของความหมายของคำประสม หากใช้หลักการเปรียบเทียบเวกเตอร์ของบริบทคือพิจารณาว่า คำที่เป็นแกนความหมายของคำประสมควรมีความหมายใกล้เคียงกับความหมายของคำประสมมากกว่าคำย่อยอื่น ดังนั้น ในกรณีที่เป็นคำประสมสองคำ W_c ซึ่งประกอบด้วยคำย่อย W_1 และ W_2 คำที่ปรากฏในบริบทของคำหลักเมื่อแปลงเป็นเวกเตอร์แล้วควรใกล้เคียงกับเวกเตอร์ของคำในบริบทที่ปรากฏร่วมกับคำประสมนั้นมากกว่าเวกเตอร์ของคำบริบทของอีกคำหนึ่ง กล่าวคือ ถ้าสมมติว่าเราพิจารณาความใกล้เคียงกันของความหมายจากระยะห่างของเวกเตอร์ ให้ระยะห่างของเวกเตอร์ W_1 กับเวกเตอร์ W_c คือความต่างของเวกเตอร์คำบริบทของคำย่อย W_1 กับคำประสม W_c และระยะห่างของเวกเตอร์ W_2 กับเวกเตอร์

W_c คือความต่างของเวกเตอร์คำบริบทของคำย่อย W_2 กับคำประสม W_c หากระยะห่างของ W_1-W_c มีค่าน้อยกว่าค่าระยะห่างของ W_2-W_c หมายความว่า W_1 มีแนวโน้มเป็นคำหลักมากกว่า W_2

ในกรณีของคำประสมสองคำนี้ หากใช้วิธีการเปรียบเทียบความหมายของคำประสมกับความหมายของคำย่อยโดยการแปลงคำบริบทเป็นเวกเตอร์ คำถามที่ตามมาคือขอบเขตของคำบริบทที่ปรากฏนั้นควรจะสั้นหรือยาวเพียงใด และควรพิจารณานับคำบริบททุกคำหรือไม่ เพื่อที่จะตอบคำถามเหล่านี้ ผู้วิจัยจึงได้ทำการทดลองศึกษาคำประสมสองคำที่พบในพจนานุกรม (Aroonmanakun, 2015) โดยทดลองใช้ขอบเขตคำบริบทสองแบบคือไม่เกิน 10 คำทั้งซ้ายและขวา และไม่เกิน 5 คำทั้งซ้ายและขวา ในเบื้องต้นกำหนดให้นับเฉพาะคำบริบทที่เป็นคำเนื้อหาเท่านั้นคือไม่นับคำที่เป็นคำไวยากรณ์ได้ เนื่องจากคาดว่าคำไวยากรณ์จะไม่มีประโยชน์โดยตรงในการเปรียบเทียบความต่างทางความหมายของคำประสมกับคำย่อย เมื่อนับความถี่ของคำบริบทแล้วก็ให้สร้างเวกเตอร์คำบริบทของคำประสม V_c และเวกเตอร์บริบทของคำย่อย V_1 และ V_2 จากนั้นพิจารณาระยะห่างของเวกเตอร์ V_c กับ V_1 และ V_c กับ V_2 ในกรณีที่ W_c เป็นคำประสม V_c ควรต่างจาก V_1 หรือไม่ก็ V_2 พอสมควร หาก W_c ไม่ใช่คำประสมแล้ว ค่าเวกเตอร์ V_c กับ V_1 และ V_c กับ V_2 ควรมีค่าใกล้เคียงกันเพราะคำบริบทที่ปรากฏควรมีลักษณะใกล้เคียงกัน ระยะห่างของเวกเตอร์จึงควรมีค่าน้อยเพื่อที่จะทดสอบแนวคิดนี้ ผู้วิจัยจึงได้ทำการทดลองหาเวกเตอร์คำบริบทของคำประสมสองคำและเวกเตอร์คำบริบทของคำย่อยทั้งสองจำนวน 4,431 คำ โดยตรวจสอบข้อมูลบริบทการเกิดของคำประสม W_c และคำย่อย W_1 และ W_2 จากคลังข้อมูลภาษาไทยแห่งชาติขนาด 33 ล้านคำ และตั้งค่าระยะห่างของเวกเตอร์คำประสมกับคำย่อยอย่างน้อยที่ 60 องศาหรือคือน้อยสองในสามของระยะห่างที่เป็นไปได้ (0-90 องศา) และที่ระยะห่างอย่างน้อยที่ 45 องศา ซึ่งเมื่อคำนวณค่า cosine similarity ของเวกเตอร์จะน้อยกว่า 0.5 และ 0.7071 ตามลำดับ ผลการทดลองปรากฏว่าที่ระยะห่าง

การประมวลผลภาษาไทย

อย่างน้อย 60 องศาหรือ 45 องศา ขอบเขตบริบทที่ 5 คำหรือ 10 คำ และให้คำนวณ เฉพาะเมื่อพบตัวอย่างของคำประสมไม่น้อยกว่า n ที่ 50, 100, และ 200 ตัวอย่าง ผลที่ได้แสดงว่าเมื่อพิจารณาบริบทที่แคบกว่าคือ 5 คำ จะแยกความต่างของคำประสม จากคำย่อยได้ดีกว่า และเมื่อกำหนดจำนวนตัวอย่างที่พบน้อยลง ($n \geq 50$) จะแยก ความต่างของคำประสมจากคำย่อยได้ดีกว่าด้วย หมายความว่าในการใช้เวกเตอร์คำ บริบทแยกความต่างระหว่างคำประสมกับคำย่อย จำนวนตัวอย่างหรือคำที่พบน้อยจะ ช่วยในการแยกความต่างได้ดีกว่า เห็นได้จากการตั้งขอบเขตที่น้อยกว่าจะทำให้ได้ จำนวนคำบริบทของเวกเตอร์มีน้อยลง และการตั้งจำนวนตัวอย่างน้อยลงทำให้คำ ประสมสองคำที่พบไม่มากในคลังข้อมูลถูกนำมาพิจารณาด้วย ซึ่งคำประสมเหล่านี้ก็ จะมีจำนวนคำบริบทที่ปรากฏรวมน้อยตามจำนวนตัวอย่างที่มีน้อยไปด้วย ดังนั้น ผู้วิจัยจึงทดลองใหม่อีกครั้งโดยนอกจากจะไม่นับคำไวยากรณ์เป็นคำบริบทแล้วยังไม่ นับคำที่มีการใช้มากในคลังข้อมูลภาษาไทยแห่งชาติหนึ่งร้อยคำแรกด้วย ผลที่ได้จาก การทดลองใหม่ที่ขอบเขตบริบท 10 คำแสดงในแถวที่สาม ซึ่งจะเห็นว่าสามารถแยก ความต่างของคำประสมกับคำย่อยได้ดีมากขึ้น โดยที่ค่า cosine of similarity ที่ 0.5 จะแยกความต่างคำประสมจากคำย่อยได้ 83.45% และที่ค่า cosine of similarity ที่ 0.7071 จะแยกความต่างคำประสมจากคำย่อยได้ 98.85%

-	- distant > 60' or cosine < 0.5			- distant > 45' or cosine < 0.7071		
-	- n>=5 0	- n>=100	- n>=20 0	- n>=5 0	- n>=100	- n>=20 0
- spa n=5	- 855/ 1571	- 558/ 116 5	- 317 /77 2	- 1480 /157 1	- 1075 /116 5	- 690 /77 2

	- 54.42 %	- 47.9 0%	- 41. 06 %	- 94.21 %	- 92.27 %	- 89. 38 %
- spa n=1 0	- 526/ 1571 - 33.48 %	- 315/ 116 5 - 27.0 4%	- 169 /77 2 - 21. 89 %	- 1332 /157 1 - 84.79 %	- 932/ 1165 - 80.0 %	- 571 /77 2 - 73. 96 %
- spa n=1 0 + - excl . Top 100	- 1311 /157 1 - 83.45 %	- 926/ 116 5 - 79.4 0%	- 564 /77 2 - 73. 06 %	- 1553 /157 1 - 98.85 %	- 1147 /116 5 - 98.45 %	- 754 /77 2 - 97. 67 %

ตาราง 3 : ผลการทดลองหาคำประสมสองคำในความถี่และความยาวบริบทต่างๆ

ทั้งหมดที่กล่าวมาเป็นการพิจารณาโดยดูเฉพาะข้อมูลจากคำประสมและคาดการณ์ว่าการดูเวกเตอร์บริบทจะช่วยในการวิเคราะห์คำประสมต่างๆ ได้ ซึ่งผลที่ได้ออกมาก็สามารถสะท้อนธรรมชาติของคำประสมได้แม้จะไม่ทั้งหมด อย่างไรก็ตามก็เพื่อจะเป็นเครื่องยืนยันว่าวิธีการศึกษานี้เหมาะสมสำหรับใช้วิเคราะห์คำประสม จึงควรมีการทดลองใช้การวัดระยะห่างของเวกเตอร์คำบริบทนี้กับข้อมูลอื่นที่ไม่ใช่คำประสมด้วย ในที่นี้จึงทดลองโดยการดึงคำสองคำใด ๆ หรือ bigram ของคำที่พบในข้อมูลตัวบทชิ้นหนึ่ง ได้จำนวน bigram 1,673 คู่ เมื่อทดลองด้วยวิธีการเดียวกัน คิดคำนวณเฉพาะ bigram ที่พบมากกว่า n 50, 100, 200 ในคลังข้อมูลภาษาไทย

การประมวลผลภาษาไทย

แห่งชาติ พิจารณาคำบริบทในขอบเขตซ้ายขวาไม่เกิน 10 คำและไม่นับคำที่ปรากฏมากในคลังข้อมูล 100 คำแรก และคิดค่า cosine similarity ที่น้อยกว่า 0.5 และ 0.7071 ทั้งนี้เพื่อให้ผลการทดลองนี้สามารถนำมาเปรียบเทียบกับผลการทดลองที่ดีที่สุดของคำประสมที่กล่าวมาข้างต้น หาก bigram ใดมีค่าความต่างของเวกเตอร์คำบริบทของ bigram กับเวกเตอร์คำบริบทของคำย่อยมากเกินกว่าค่าที่กำหนด bigram นั้นก็มีลักษณะคล้ายกับคำประสม คือ คำบริบทที่ปรากฏร่วมกับ bigram นั้นแตกต่างจากคำบริบทที่ปรากฏร่วมกับคำย่อย ผลการเปรียบเทียบแสดงให้เห็นในตารางข้างล่าง ซึ่งจะเห็นว่าแม้การตั้งค่าระยะห่างของเวกเตอร์ที่ 45 องศาจะทำให้ครอบคลุมข้อมูลคำประสมได้ถึง 98.85% แต่ก็มีผลทำให้ครอบคลุมคำ bigram ถึง 72.96% ด้วย ต่างจากการกำหนดค่าความต่างของระยะห่างเวกเตอร์อย่างน้อย 60 องศาหรือค่า cosine similarity น้อยกว่า 0.5 ซึ่งครอบคลุมข้อมูลคำประสมได้ 83.45% ในขณะที่ข้อมูล bigram จะถูกครอบคลุมเพียง 33.78% การตั้งค่านีจึงทำให้เห็นว่าความแตกต่างของคำประสมกับคำสองคำใดๆได้ ดังนั้น ในที่นี้จะใช้ค่า setting ที่ cosine < 0.5, ขอบเขตบริบทที่ไม่เกิน 10 คำ ไม่นับคำไวยากรณ์และคำที่ปรากฏมากที่สุดหนึ่งร้อยลำดับแรก และมีตัวอย่างให้พิจารณาอย่างน้อย 50 ตัวอย่าง เป็นกรอบในการวิเคราะห์รายละเอียดของคำประสมเมื่อใช้เวกเตอร์คำบริบทเป็นหลักในการศึกษาคำประสม

-	- distant > 60' or cosine < 0.5			- distant > 45' or cosine < 0.7071		
-	- n>=5 0	- n>= 100	- n> =2 00	- n>=5 0	- n>=1 00	- n> =2 00

- span =10 +	- 1311 /157 1	- 926 /11 65	- 564 /77 2	- 1553 /157 1	- 1147 /116 5	- 754 /77 2
- excl. Top 100 -	- 83.4 5% com poun d	- 79.4 0% 	- 73. 06 %	- 98.8 5% 	- 98.4 5% 	- 97. 67 %
- span =10 +	- 381/ 1128 -	- 267 /99 8	- 165 /84 3	- 823/ 1128 -	- 693/ 998 -	- 542 /84 3
- excl. Top 100 -	- 8% bigr m	- 26.7 5% 	- 19. 57 %	- 6% 	- 4% 	- 64. 29 %

ตาราง 4 : เปรียบเทียบผลการหาคำประสมจากข้อมูลคำประสมและข้อมูล bigram ใน setting ต่าง ๆ

เมื่อพิจารณาเพิ่มเติมว่าคำประสมมีทั้งที่เป็นแบบเข้าศูนย์และไร้ศูนย์ ประเด็นที่เกี่ยวข้องคือเรื่องคำประสมนั้นมีค้ำย้อยใดเป็นแก่นความหมายหรือเป็นคำหลักหรือไม่ การใช้เร็กเตอร์คำบริบททำให้เราเทียบความหมายทางอ้อมผ่านบริบทที่คล้ายคลึงหรือแตกต่างกันได้ ในที่นี้จึงวิเคราะห์เพิ่มเติมโดยแยกตามประเภทของคำประสมสามแบบดังแสดงในตาราง 5 ตารางข้างล่างซึ่งแสดงความต่างระหว่างคำประสมโดยแยกตามรายละเอียดประเภทของคำประสม ถ้าให้ W_c เป็นคำประสมที่ประสมจากคำสองคำคือ W_1 และ W_2 $Dist(W_1, W_c)$ หมายถึงจำนวนคำประสมที่คำ

ย่อย W_1 มีค่าบริบทต่างจากคำประสม W_c อย่างชัดเจน (คือมีระยะห่างของเวกเตอร์ค่าบริบทมากกว่า 60 องศา หรือค่า cosine similarity ของเวกเตอร์ < 0.5) เช่น "เข็มกลัด" มีค่า cosine 0.32 เมื่อเทียบกับคำ "เข็ม" และมีค่า cosine 0.504 เมื่อเทียบกับคำ "กลัด" "เข็ม" จึงเป็นคำที่เป็นคำหลัก $\sim \text{Dist}(W_1, W_c)$ หมายถึงจำนวนคำประสมที่คำย่อย W_1 มีค่าบริบทไม่แตกต่างไปจากคำประสม W_c ตัวอย่างเช่น "รูปปั้น" มีค่า cosine 0.656 เมื่อเทียบกับ "รูป" และมีค่า cosine 0.411 เมื่อเทียบกับ "ปั้น" "รูป" จึงไม่ใช่คำที่ใกล้เคียงกับคำประสม "รูปปั้น" ในกรณีนี้ คำว่า "ปั้น" ดูจะมีความหมายใกล้เคียงกับคำประสมมากกว่า

คำประสมแบบเข้าสู่ศูนย์โดยหลักแล้วจะมีค่าหนึ่งเป็นความหมายแก่นซึ่งควรมีความหมายใกล้เคียงกับคำประสมมากกว่าอีกคำหนึ่ง เช่น "ห้องน้ำ" มีคำว่า "ห้อง" เป็นคำหลัก เพราะหมายถึงห้องประเภทหนึ่งที่ใช้สำหรับขับถ่าย ดังนั้นจึงน่าจะมียังมีจำนวนคำที่เป็นคำหลักชัดเจนและมากกว่าคำประสมแบบไร้ศูนย์ แต่ผลอาจไม่เป็นเช่นนั้น เช่น "ห้องน้ำ" มีค่า cosine ที่ 0.83 กับ 0.512 กับคำว่า "ห้อง" และ "น้ำ" ตามลำดับ แสดงว่าไม่มีคำใดที่มีเวกเตอร์ค่าบริบทต่างออกไปจากคำประสมชัดเจน ค่าบริบทของ "ห้องน้ำ" คล้ายกับค่าบริบทของ "น้ำ" มากกว่า "ห้อง" และเมื่อพิจารณาผลในภาพรวมที่แสดงในตารางข้างล่าง พบว่าไม่ได้มีลักษณะที่สอดคล้องกับที่คาดการณ์มากนัก ในคำประสมแบบเข้าสู่ศูนย์ซึ่งที่ผ่านมาได้วิเคราะห์ให้คำแรกเป็นคำหลักมากกว่าคำที่สอง แต่ผลการเปรียบเทียบค่าบริบทพบว่าคำที่มีบริบทแตกต่างไปจากคำประสมเป็นได้ทั้งคำแรกหรือคำหลังหรือทั้งสองคำในปริมาณพอๆกัน และก็มีคำประสมที่ทั้งคำแรกคำหลังไม่ได้มีค่าบริบทที่ต่างไปจากคำประสมมากนัก เช่น "ใจดี" มีค่า cosine ที่ 0.76 และ 0.811 กับคำ "ใจ" และ "ดี" ตามลำดับ คือค่าบริบทที่ปรากฏร่วมกับ "ใจดี" จะไม่ต่างจากค่าบริบทที่ปรากฏร่วมกับ "ใจ" และค่าบริบทที่ปรากฏร่วมกับ "ดี" อย่างชัดเจน แต่ก็คล้ายกับค่าบริบทของ "ดี" มากกว่า "ใจ" เป็นต้น ส่วนในคำประสมแบบไร้ศูนย์ซึ่งวิเคราะห์กันมาไม่มีคำใดเป็นคำหลักชัดเจนเพราะอาจจะ

ความหมายสำคัญทั้งคู่หรือความหมายของคำประสมจะต่างไปจากคำทั้งสองเลย กรณีเช่นนี้ W_1 W_2 ควรจะมีความหมายที่ไม่ใกล้เคียงกับ W_c ดังนั้นเวกเตอร์คำบริบทที่ปรากฏร่วมของ W_1 และ W_2 ควรแตกต่างจากเวกเตอร์คำบริบทของ W_c หรือกล่าวได้ว่า ระยะห่างของ W_1-W_c และ W_2-W_c ควรมีระยะห่างอย่างน้อย 60 องศา แต่กลับพบว่าคำประสมแบบไร้ศูนย์จำนวนหนึ่งดูเหมือนมีค่าบางค่าที่ใกล้กับคำประสมมากกว่าอีกคำหนึ่งเหมือนกัน ตัวอย่างเช่น “ลูกเรือ” มีค่า cosine ที่ 0.279 และ 0.543 กับคำ “ลูก” และ “เรือ” ตามลำดับ แสดงว่าคำว่า “เรือ” มีคำบริบทที่ปรากฏใกล้กับคำประสม “ลูกเรือ” มากกว่า” ในขณะที่คำว่า “ลูก” มีคำบริบทที่ต่างออกไปจากคำประสมอย่างชัดเจน หรือคำว่า “เต็มตา” มีค่า cosine ที่ 0.492 กับ 0.726 กับคำว่า “เต็ม” และ “ตา” ตามลำดับ นั่นคือ คำบริบทของคำประสม “เต็มตา” มีความคล้ายกับคำบริบทของคำ “ตา” มากกว่า “เต็ม” คำประสมแบบไร้ศูนย์จึงไม่จำเป็นว่าทั้งสองคำย่อยจะมีคำบริบทที่ต่างไปจากคำบริบทของคำประสม เพราะคำประสมนั้นยังมีความหมายบางอย่างเกี่ยวข้องกับคำใดคำหนึ่งอยู่

	- D is t(W 1, W c)	- D is t(W 2, W c)	- ~ Di st(W 1, W c)	- ~ Di st(W 2, W c)	- Dist(W ₁ ,W _c) and - Dist(W ₂ ,W _c)	- ~Dist(W ₁ ,W _c) and ~Dist(W ₂ ,W _c)	- Dist(W ₁ , W _c) or - Dist(W ₂ , W _c)
- E r c c	- 6 1 0 7 1.	- 6 0 5 7 0	- 24 9 28 .9	- 25 4 29 .5	- 497 - 57.86 %	- 141 - 16.41 %	- 718 - 83.5 9%

การประมวลผลภาษาไทย

-	8	0	.	9	7			
	5	1	4	%	%			
	9	%	3					
		%	%					
-	E	- 4	- 3	- 13	- 14	- 351	- 89	- 447
	X	0	9	1	3	- 65.49	- 16.60	- 83.4
	C	5	3	- 24	- 26	%	%	0%
-	E	- 7	- 7	.4	.6			
	3	5.	3	4	8			
	C	5	.	%	%			
		6	3					
		%	2					
		%	%					
-	C	- 1	- 1	- 52	- 56	- 91	- 30	- 146
	C	2	1	- 30	- 32	- 52.60	- 17.34	- 84.3
-	1	1	6	.0	.3	%	%	9%
	7	- 6	- 6	6	7			
	3	9.	7	%	%			
		9	.-					
		4	5					
		%	%					

ตาราง 5 : เปรียบเทียบระยะทางเวกเตอร์บริบทของคำประสมกับคำย่อยแยกตามประเภทคำประสม

ส่วนคำประสมประเภท coordinate compound ที่ทั้งสองคำเป็นคำระดับเดียวกันรวมถึงคำที่มีความหมายเหมือนหรือใกล้กัน พบว่าส่วนใหญ่จะมีคำใดคำหนึ่งหรือทั้งสองคำที่มีคำบริบทแตกต่างไปจากคำประสม แต่ก็มีบางส่วนที่ทั้งสองคำ

ไม่ได้มีค่าบริบทแตกต่างกันไปจากค้าย่อย เช่น “เหตุผล” มีค่า cosine ที่ 0.714 และ 0.73 กับคำว่า “เหตุ” และ “ผล” ตามลำดับ ในกรณีของคำซ้อน ความหมายคำบริบทที่ปรากฏรวมกับคำประสมควรใกล้เคียงกับคำใดคำหนึ่ง ซึ่งส่วนใหญ่ก็เป็นไปตามนั้น แต่ก็ยังมีจำนวนหนึ่งที่ไม่ได้เป็นตามนั้น เช่น “เล็กน้อย” ซึ่งมีค่า cosine ที่ 0.583 และ 0.8 กับคำ “เล็ก” และ “น้อย” ตามลำดับ และหากเป็นคำที่มีความหมายเหมือนกันมาประสมกัน ค่าบริบทของ W_1 และ W_2 ก็ควรคล้ายกันด้วย จำนวน $\text{Dist}(W_1, W_2)$ ก็ควรมีจำนวนน้อย แต่กลับพบว่าในคำประสมประเภทนี้มีจำนวนมากพอสมควร

จากการวิเคราะห์ในรายละเอียดตามที่กล่าวมา แม้ว่าการเปรียบเทียบเวกเตอร์คำบริบทจะช่วยบอกถึงความเป็นคำประสมได้ระดับหนึ่ง มีคำประสมเพียง 16-17% ที่ไม่ผ่านเกณฑ์การเทียบเวกเตอร์คำบริบท แต่การเทียบว่าคำใดใกล้เคียงหรือแตกต่างจากคำประสมมากกว่าดูจะไม่มีความสัมพันธ์ที่ชัดเจนกับประเภทย่อยของคำประสมสามประเภทใหญ่ คือ คำประสมแบบเข้าศูนย์ แบบไร้ศูนย์ และแบบเคียงกัน แสดงว่าการวิเคราะห์โดยอิง intuition ที่บอกว่าความหมายของคำใดที่เป็นแก่นของคำประสมอาจจะไม่สอดคล้องกับข้อมูลการคำนวณความคล้ายคำบริบทที่ใช้จริง การวิเคราะห์เปรียบเทียบเวกเตอร์คำบริบทนี้จึงไม่สามารถใช้บอกความเป็นคำหลักในคำประสมได้ อย่างไรก็ตาม เราสามารถใช้ประโยชน์จากการเทียบเวกเตอร์คำบริบทได้ว่าความหมายของคำประสมนั้นใกล้เคียงกับความหมายของค้าย่อยใดมากกว่าได้

หลังจากที่ผู้วิจัยได้นำเสนอผลการวิเคราะห์คำประสมแล้ว ในกรณีของคำ bigram ซึ่งเป็นคู่คำใดคำหนึ่งที่พบในตัวบท แม้ว่าส่วนใหญ่จะไม่เข้าเกณฑ์ที่กำหนดไว้ แต่ก็มีคำบางคู่จำนวน 381 bigram จาก 1,128 bigram ที่ผ่านเกณฑ์หรือคิดเป็น 33.78% ที่มีคำหนึ่งมีเวกเตอร์คำบริบทแตกต่างจากเวกเตอร์คำบริบทของ bigram เมื่อพิจารณารายการคำที่อยู่รายการ bigram นี้ว่ามีคู่คำได้จัดเป็นคำประสมได้หรือไม่ก็พบว่า ประมาณ 15% ของกลุ่มที่เข้าเกณฑ์เป็นคำประสมที่ยังไม่พบในพจนานุกรม

การประมวลผลภาษาไทย

ส่วนในกลุ่ม bigram ที่ไม่เข้าเกณฑ์ พบว่ามีเพียง 2.68% ที่เป็นคำประสม ผลที่พบนี้เป็นตัวชี้ว่าเกณฑ์การใช้เวกเตอร์ค่าบริบทนั้นสามารถใช้ตัดสินความเป็นคำประสมได้ในระดับหนึ่ง อย่างไรก็ตาม ผลที่ได้นี้มาจากการดูคู่คำ bigram ทุกคู่ที่พบในตัวบท หากมีการสกัดหาคู่คำ bigram ที่มีความน่าจะเป็นสูงที่จะเป็นคำประสม ซึ่งอาจใช้วิธีการดูค่าความเหนียวแน่นของการปรากฏร่วมกันเป็นเครื่องมือในการคัดเลือกคำที่น่าจะเป็นคำประสมก่อน (Aroonmanakun 2009)

- คำประสม	- คำ 1	- คำ 2	- Sim(12 - 1)	- Sim(12 - 2)
- หน้าที่	- หน้า	- ที่	- 0.030118957	- 0.26287812
- ชีขา	- ชี	- ขา	- 0.4789203	- 0.48328087
- น้ำตาล	- น้ำ	- ต ล	- 0.4689505	- 0.23308073
- หมอดู	- หมอ	- ดู	- 0.58779615	- 0.19264597
- หมอผี	- หมอ	- ผี	- 0.35089532	- 0.57782817
- กำปั้น	- กำ	- ปั้น	- 0.37818414	- 0.30825272
- พ่อตา	- พ่อ	- ตา	- 0.576542	- 0.19858865
- ปูม้า	- ปู	- ม้า	- 0.39615706	- 0.21366525
- คุ่มกัน	- คุ่ม	- กัน	- 0.19733387	- 0.001145464 3
- เนื้อตัว	- เนื้อ	- ตัว	- 0.4814648	- 0.42878616
- ด้านทาน	- ด้าน	- ท น	- 0.65028125	- 0.005783905
- ดันทุน	- ดัน	- ทุน	- 0.003776755 9	- 0.49896023

- หมอน	- หมอ	- ช่าง	- 0.6956959	- 0.3176671
- ช่าง	- หมอ	- กัน	- 0.3027253	- 0.03867517
- ป้องกัน	- ป้อง	- ป่า	- 0.5594359	- 0.14451276
- หมาป่า	- หมา	- ผา	- 0.4652253	- 0.7589084
- ภูเขา	- ภู	- แม่	- 0.66941816	- 0.7121279
- พ่อแม่	- พ่อ	- พูน	- 0.68058735	- 0.07069173
- เพิ่มพูน	- เพิ่ม	- จำ	- 0.63393426	- 0.3423893
- ท่องจำ	- ท่อง	- เต็ม	- 0.3675447	- 0.20554298
- เพิ่มเต็ม	- เพิ่ม	- ปั่น	- 0.3315613	- 0.09549945
- แปะปั่น	- แปะ			

สรุปผลและอภิปราย

จากผลการทบทวนวรรณกรรมและการทดลองการวิจัยที่ได้ทำ จะเห็นได้ว่าเส้นแบ่งระหว่างคำประสมและวลีนั้นไม่ได้มีขอบเขตที่ชัดเจน เนื่องจากภาษาไทยไม่มีวลีวิภาคที่ชัดเจนที่ช่วยในการกำกับว่าส่วนใดเป็นส่วนใดเป็นวลี ปัญหาการตัดสินความเป็นคำประสมจึงเป็นปัญหาที่มักมีข้อถกเถียงและหาข้อสรุปที่ชัดเจนไม่ได้ ความต่างในเรื่องนี้จึงควรมองเป็นเรื่องของระดับความแตกต่างมากกว่าที่จะเป็นการตัดสินว่าเป็นคำประสมหรือไม่เป็นแบบเด็ดขาด เพราะความเป็นหน่วยศัพท์นั้น เราจะเห็นว่ายิ่งคำประสมมีความซับซ้อนมากขึ้นหรือประสมหลายคำมากขึ้น ความเป็นหน่วยศัพท์ก็จะมัน้อยลง นอกจากนี้ จากหลายๆ ตัวอย่างที่ได้ยกมา เราจะเห็นได้ว่า รูปคำที่พบอาจมีความกำกวมที่จะมองเป็นคำประสมก็ได้หรือมองว่าเป็นวลีก็ได้ การตัดสินว่าเป็นคำประสมหรือวลีเป็นเรื่องที่ขึ้นอยู่กับบริบทที่เกิด ดังนั้นปัญหาเรื่องคำประสมนี้เป็นเรื่องที่ควรรวมเป็นส่วนหนึ่งของขั้นตอนการวิเคราะห์ประโยค

อย่างไรก็ตาม การสกัดหาคำประสมใหม่ยังคงเป็นเรื่องที่สำคัญในการประมวลผลภาษาไทย เพราะเราจำเป็นต้องบรรจุคำประสมเกิดใหม่เหล่านี้ไว้ในพจนานุกรมเพื่อที่จะได้สามารถใช้ตัดสินได้ว่ารูปคำที่พบในข้อมูลสามารถมองหรือ

วิเคราะห์เป็นคำได้หรือไม่ ซึ่งคำประสมเกิดใหม่ควรมีความเป็นหน่วยศัพท์สูง ซึ่งจากข้อมูลที่พบจากคำประสมที่มีแล้วในพจนานุกรม พบว่าส่วนใหญ่เป็นคำประสมสองคำ หรือสามคำ คำประสมสี่คำมีลักษณะเป็นคำซ้ำมากกว่า ในที่นี้จึงมุ่งความสนใจที่การสกัดคำประสมไม่เกินสี่คำจากคลังข้อมูลภาษาไทย โดยมองงานส่วนนี้เป็นสองส่วน ส่วนแรกเป็นการสกัดหาคำที่น่าจะเป็นคำประสมได้ (candidate compound) ซึ่งสามารถใช้ผลข้อมูลจากการวิเคราะห์คำประสมที่พบในพจนานุกรมได้ ข้อมูลที่ใช้ได้แก่ ความน่าจะเป็นที่ POS ของคำหลักกับคำพึ่งพา ความน่าจะเป็นที่คำนั้นจะเป็นคำหลักได้ ความน่าจะเป็นที่จะพบโครงสร้างที่กำหนด นอกจากนี้ เรายังสามารถใช้ข้อมูลการปรากฏร่วมทางสถิติ (statistical collocation) ของคู่คำที่สัมพันธ์กันได้จากนั้นจึงใช้แนวคิดเรื่องความคล้ายคลึงทางความหมายของคำ โดยอาศัยการปรากฏของบริบทที่เหมือนหรือแตกต่างกันเป็นเครื่องตัดสิน บริบทของคำประสมและคำย่อยจะถูกนำมาเปรียบเทียบกันโดยการแปลงให้เป็นเวกเตอร์ของคำบริบท ซึ่งจะนับเฉพาะที่เป็นคำเนื้อหาและไม่รวมคำที่มีความถี่การใช้สูงมากๆ เมื่อแปลงเป็นเวกเตอร์แล้วจึงใช้หลักทางคณิตศาสตร์คำนวณค่า cosine เพื่อหาว่าเวกเตอร์ทั้งคู่นั้นใกล้เคียงกันเพียงใด โดยค่าที่ได้มีค่าระหว่าง 0 ถึง 1 หากมีค่าเข้าใกล้ 1 แสดงว่าเวกเตอร์อยู่ใกล้กันหรือคำบริบทที่พบไม่แตกต่างกันมากนัก แสดงว่าความหมายของคำย่อยกับความหมายของคำประสมนั้นไม่แตกต่างกัน แต่หากมีค่าเข้าใกล้ 0 แสดงว่าเวกเตอร์อยู่ห่างจากกันมาก แสดงว่าคำบริบทที่พบนั้นแตกต่างกันอย่างเห็นได้ชัด สะท้อนว่าความหมายของคำประสมนั้นแตกต่างไปจากความหมายของคำย่อย ซึ่งหลักนี้เป็นเครื่องมือสำคัญที่ทำให้เราสามารถใช้นำเวกเตอร์คำบริบทมาใช้ในการตัดสินว่าคำต้องสงสัยนั้นเป็นคำประสมที่ประกอบด้วยคำย่อยหลายคำได้ และการทดลองใช้วิธีการเดียวกันกับเวกเตอร์คำบริบทกับคู่คำใดก็ได้ที่พบในข้อความต่อเนื่อง แสดงให้เห็นว่า คู่คำใด ๆ ที่ไม่ใช่คำประสมนั้น บริบทของคู่คำนั้นกับบริบทของคำเดี่ยวไม่มีความแตกต่างกันเมื่อแปลงเป็นเวกเตอร์ นั่นหมายความว่าวิธีการนี้สามารถใช้ตัดสินคำ

ประสมได้ดีในระดับที่น่าพอใจ อย่างไรก็ตาม ในการทดลองใช้วิธีการนี้เพื่อบ่งบอกประเภทของคำประสม กลับพบว่าไม่สามารถใช้แยกความแตกต่างของคำประสมแบบเข้าสู่ศูนย์ หรือแบบไร้ศูนย์ได้ สิ่งที่บอกได้คือคำย่อยใดมีความหมายใกล้เคียงกับคำประสมมากกว่า

สถิติ

โปรแกรมที่กล่าวถึงในบทนี้เป็นโปรแกรมที่ผู้เขียนเขียนด้วยภาษา Perl เพื่อสร้างเวกเตอร์บริบท แต่ภายหลังเมื่อมีโมดูล Word2Vec ใช้กันแพร่หลาย ผู้เขียนจึงได้ใช้โมดูลนี้สร้าง Word2Vec ภาษาไทยจากคลังข้อมูลภาษาไทยแห่งชาติ และเรียกใช้งานผ่านแพ็คเกจ TLTK. โมดูล `tltk.corpus.compound(w1,w2)` จะตรวจสอบว่าคำประสมสองคำ w_1w_2 เป็นคำในภาษาไทย และ w_1, w_2 เป็นคำย่อยในภาษาไทยด้วยไหม จากนั้นหาความใกล้กันของเวกเตอร์ $w_1 - w_1w_2, w_2 - w_1w_2$ และ $w_1 - w_2$ แล้วให้ผลมาเป็นคู่ค่าตามลำดับความคล้ายคลึงจากมากไปน้อย เช่น *กลัดกลุ้ม, คับแค้น, เล็กน้อย, วางเงิน* ค่าที่ใกล้กับคำประสม คือ *กลุ้ม, แค้น, น้อย* จะเห็นว่าค่าที่ได้นี้ไม่จำเป็นต้องเป็นคำหลักในคำประสม แต่มีบริบทที่คล้ายคลึงกับคำประสมมากกว่า เช่น *วางเงิน* เป็นคำกริยา *วาง* จึงเป็นคำหลักในโครงสร้างคำประสม แต่คำนี้มีบริบทการใช้ใกล้กับ *เงิน* มากกว่า *วาง*

ตัวอย่างนี้สามารถทดลองใน Google colab ตามที่อยู่นี้

https://colab.research.google.com/drive/1YCrhKqQ09YEVXgs-_HYXv4n5N73LdzF2Y?usp=sharing

```
!wget http://www.arts.chula.ac.th/~ling/contents/Upload/TNCc5model3.bin
import tltk
w1 = 'กลัด'
w2 = 'กลุ้ม'
```

```
print(tltk.corpus.compound(w1,w2))  
w1 = 'คืบ'  
w2 = 'แค้น'  
print(tltk.corpus.compound(w1,w2))  
w1 = 'เล็ก'  
w2 = 'น้อย'  
print(tltk.corpus.compound(w1,w2))  
w1 = 'วาง'  
w2 = 'เงิน'  
print(tltk.corpus.compound(w1,w2))
```

```
[(['กุ่ม', 'กั้ดกุ่ม'), 0.42245594], ([('กั้ด', 'กั้ดกุ่ม'), 0.09066804], ([('กั้ด', 'กุ่ม'), 0.0011619462])  
[(['แค้น', 'คืบแค้น'), 0.60789037], ([('คืบ', 'คืบแค้น'), 0.5096577], ([('คืบ', 'แค้น'), 0.3744389])  
[(['เล็ก', 'น้อย'), 0.45332724], ([('น้อย', 'เล็กน้อย'), 0.35492074], ([('เล็ก', 'เล็กน้อย'), 0.24106339])  
[(['เงิน', 'วางเงิน'), 0.52538854], ([('วาง', 'วางเงิน'), 0.35440722], ([('วาง', 'เงิน'), 0.03699364])
```

อ้างอิง

- Aroonmanakun, W. (2009). Extracting Thai Compounds Using Collocations and POS Bigram Probabilities without a POS Tagger. 2009 International Conference on Asian Language Processing, 118-122.
- Aroonmanakun, W. (2015). The use of context vectors in determining Thai compounds. Linguistic Research. Vol.32 No.1, 1-20
- Girju, Roxana, Moldovan, Dan, Tatu, Marta, and Antohe, Daniel. (2005). On the semantics of noun compounds. Computer Speech and Language 19:479-496.
- Harris, Zellig. (1954) Distributional Structure, WORD, 10:2-3, 146-162.
- Kriengkhet, Kanyanut, Kosawat, Krit, and Anchaleenukul, Sunant. (2007). A Computational Linguistics Study of Compound Nouns in Thai. Paper

- presented at NAC 2007 งานประชุมวิชาการประจำปี 2550 สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ, Thailand. Levi, Judith N. (1978). The syntax and semantics of complex nominals. New York: Academic Press.
- Rosario, Barbara, and Hearst, Marti. (2001). Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, 82-90.
- Sornlertlamvanich, Virach, Potipiti, Tanapong, and Charoenporn, Thatsanee. (2000). Automatic corpus-based Thai word extraction with the c4.5 learning algorithm, Stroudsburg, PA, USA.
- Sornlertlamvanich, Virach, and Tanaka, Hozumi. (1996). The Automatic Extraction of Open Compounds from Text Corpora.
- เพียรศิริ วงศ์วิภานนท์ (2531) คำซ้อนในภาษาไทย ตอนที่ 1: ลักษณะพิเศษของคำซ้อน. ภาษาและวรรณคดีไทย, 3(5), 15-27.

การรู้จักชื่อเฉพาะภาษาไทย⁶

ชื่อเฉพาะเป็นกลุ่มคำประเภทหนึ่งที่พบในทุกภาษา ใช้เพื่ออ้างถึงเฉพาะสิ่งเฉพาะบุคคล ในหนังสือไวยากรณ์ไทยจะเรียกว่าเป็นคำวิสามานยนาม (พระยาอุปกิตศิลปสาร 2546) หรือคำนามวิสามัญ (นวรรณ พันธุเมธา 2549) ซึ่งแตกต่างต่างจากคำนามทั่วไปที่เป็นคำที่ใช้เรียกสิ่งต่างๆ ในโลกโดยทั่วไป เช่น แม่น้ำ ภูเขา ต้นไม้ คน สุนัข แต่วิสามานยนามจะเป็นชื่อที่ตั้งขึ้นมาโดยเฉพาะสำหรับบุคคล เช่น นายโสภณ เพชรสว่าง ชื่อสถานที่ เช่น อำเภอบางซื่อ ชื่อองค์กร เช่น จุฬาลงกรณ์มหาวิทยาลัย ชื่อสัตว์เลี้ยง เช่น โจดี เป็นต้น วิสามานยนามต้นแบบเหล่านี้จะอ้างถึงสิ่งที่มีอยู่หนึ่งเดียว แตกต่างจากการอ้างถึงสิ่งต่างๆ ด้วยคำนามทั่วไปหรือที่เรียกว่าวิสามานยนาม ซึ่งอ้างถึงสิ่งที่มีลักษณะร่วมกันและจัดเป็นกลุ่มเดียวกันประเภทเดียวกันได้ เช่น สุนัข จาน คนทำงาน เป็นต้น อย่างไรก็ตาม เราจะพบว่าคำที่ดูเหมือนเป็นชื่อเฉพาะในบางครั้งก็ไม่ได้อ้างถึงของที่มีสิ่งเดียวแต่เป็นกลุ่มของที่มีลักษณะเดียวกันได้ เช่น ชื่ออาหาร “ต้มยำกุ้ง” เป็นชื่อที่ตั้งขึ้นเพื่อเรียกอาหารประเภทหนึ่ง “นกกระจอกเทศ” เป็นชื่อที่ตั้งขึ้นให้กับนกประเภทหนึ่ง แต่อ้างถึงกลุ่มของนกประเภทหนึ่งไม่ได้เจาะจงเฉพาะตัว

การรู้จักชื่อเฉพาะเป็นงานหนึ่งที่สำคัญในการประมวลผลภาษาไทย เพราะการที่คอมพิวเตอร์จะสามารถประมวลผลข้อความภาษาไทยได้ถูกต้องนั้น จำเป็นต้องรู้

⁶ เนื้อหาในบทนี้ส่วนหนึ่งมาจากการสังเคราะห์งานวิทยานิพนธ์ของนิสิตมหาดบัณฑิตสาขาวิชาภาษาศาสตร์ที่ได้รับทุนสนับสนุนการทำวิทยานิพนธ์จากสำนักงานกองทุนสนับสนุนการวิจัย ปี พ.ศ.๒๕๕๓

ว่าขอบเขตของชื่อต่างๆ เริ่มต้นที่ไหนสิ้นสุดที่ไหนและเป็นชื่อประเภทใด เพื่อจะได้ประมวลผลได้อย่างเหมาะสม เช่น หากเป็นงานแปลก็จะรู้ว่าชื่อเฉพาะนั้นจะไม่ต้องแปลแต่ใช้วิธีการอื่นเช่นเขียนทับศัพท์แทน เช่น หากคอมพิวเตอร์ไม่รู้ว่า “ไทยรักไทย” เป็นชื่อพรรคการเมือง ก็อาจจะดึงเอาคำ “love” มาใช้แปลคำว่า รัก ได้ แทนที่จะใช้วิธีถอดตัวอักษรไทยเป็นโรมัน หากเป็นงานค้นคืนสารสนเทศก็จะรู้ว่าคำนั้นเป็นชื่อเฉพาะแตกต่างจากคำนามทั่วไป เช่น เมื่อค้นหาเอกสารเกี่ยวกับปูนซีเมนต์ หากคอมพิวเตอร์ไม่รู้ว่า “บริษัทปูนซีเมนต์นครหลวง” เป็นชื่อบริษัท ก็อาจจะดึงเอกสารนี้ออกมาให้ซึ่งอาจจะไม่ได้เกี่ยวข้องใด ๆ กับการผลิต การใช้ปูนซีเมนต์ เป็นต้น หรือหากต้องการค้นเอกสารเกี่ยวข้องกับจังหวัดกรุงเทพฯ หากคอมพิวเตอร์ไม่สามารถแยกประเภทชื่อเฉพาะได้ ก็อาจจะดึงเอกสารเกี่ยวกับ บริษัทกรุงเทพ ประกันภัย มาให้ก็ได้ การรู้จำชื่อเฉพาะภาษาไทยเป็นงานที่ยากพอสมควร ทั้งนี้เพราะในภาษาไทยไม่มีการใช้ตัวอักษรพิเศษสำหรับการเขียนชื่อ ไม่มีการเว้นวรรคระหว่างคำ จึงทำให้ระบุขอบเขตของชื่อเฉพาะว่าเริ่มต้นและสิ้นสุดที่ใดได้ยาก นอกจากนี้ ชื่อเฉพาะบางครั้งก็มีรูปเหมือนคำทั่วไป ชื่อเฉพาะชื่อเดียวกันเมื่อใช้ต่อเนื่องในตัวบทก็อาจมีการเปลี่ยนรูปหรือลดรูปได้ เป็นต้น

งานรู้จำชื่อเฉพาะเป็นที่สนใจศึกษาในการประมวลผลภาษาต่างๆ จำนวนมากเป็นส่วนหนึ่งของการประมวลผลที่จัดในการประชุมวิชาการเรื่อง The Message Understanding Conferences (MUC) ที่จัดขึ้นเป็นประจำตั้งแต่ปี 1987-1997 โดยที่การประชุมนี้จัดให้ผู้ที่สนใจพัฒนาระบบการรู้จำชื่อบุคคล สถานที่ และองค์กร ได้แลกเปลี่ยนความรู้ทางการวิจัยภายใต้ชื่องานว่า named entity recognition สาเหตุที่มุ่งความสนใจที่ชื่อเฉพาะสามประเภทนี้เพราะเป็นชื่อเฉพาะที่พบในเอกสารข่าวสารโดยทั่วไป แต่การศึกษาการรู้จำชื่อเฉพาะประเภทอื่นๆ ก็มีผู้สนใจศึกษาเช่นกัน เพราะเป็นงานเฉพาะด้านที่ต้องการ เช่น การรู้จำชื่อเฉพาะในสาขาวิชาต่างๆ การรู้จำชื่อผลิตภัณฑ์ เป็นต้น

งานวิจัยเรื่องการรู้จำชื่อเฉพาะในภาษาไทยนั้นมีการศึกษาอยู่บ้าง แต่ไม่มากนัก ได้แก่ งานของอมรทิพย์ กวินปณิธาน (2546) ที่ศึกษาหาบริบทที่จะใช้บ่งบอกถึงชื่อเฉพาะในภาษาไทย งานด้านการพัฒนาระบบการรู้จำชื่อบุคคล สถานที่ และองค์กร ได้แก่ งานของ Charoenpornsawat et al. (1998) พัฒนาระบบการรู้จำโดยใช้ระบบการเรียนรู้แบบวินนาวัส โดยพิจารณาลักษณะอย่างเช่น คำปรากฏร่วม คำใกล้เคียงในบริบท และให้มีการใช้กฎช่วยพิจารณา เช่น หากเป็นคำที่ไม่มีในพจนานุกรม ให้คิดว่าคำนั้นกับคำใกล้เคียงอาจจะรวมกันเป็นชื่อเฉพาะได้ เป็นต้น งานของ Chanlekha et al. (2002) ที่ใช้แบบจำลองทางสถิติร่วมกับการใช้กฎ โดยอาศัยบริบททั้งภายนอกและภายในของชื่อเฉพาะ และใช้รายการชื่อเฉพาะ และ งานของ Chanlekha and Kawtrakul (2004) ที่ใช้แบบจำลอง Maximum Entropy ร่วมกับการใช้กฎและพจนานุกรมในการรู้จำชื่อเฉพาะ

การรู้จำชื่อบุคคล สถานที่ และองค์กร

แนวทางในการพัฒนาระบบการรู้จำชื่อบุคคล สถานที่ และองค์กร มองได้ว่ามีสองแนวทางใหญ่ แนวทางแรกคือการใช้กฎ ซึ่งจะเหมาะสมกับภาษาที่มีรูปแบบการเขียนชื่อเฉพาะที่ชัดเจนระดับหนึ่งสามารถระบุเป็นกฎได้ว่าชื่อเฉพาะนั้นขึ้นต้นอย่างไรหรือเขียนแตกต่างจากคำทั่วไปอย่างไร อีกแนวทางหนึ่งคือการใช้แบบจำลองทางสถิติ ซึ่งเป็นแนวทางที่เป็นที่นิยมในปัจจุบันเนื่องจากไม่ต้องผูกติดกับแต่ละภาษา สามารถนำแบบจำลองเดียวกันมาปรับใช้กับภาษาที่แตกต่างกันได้ แบบจำลองทางสถิตินี้จะมีวิธีการใช้แบบจำลองแบบต่างๆ เช่น Maximum Entropy, Decision Tree, Markov Model, Support Vector Machine, Conditional Random Fields เป็นต้น แต่ในบางงานก็จะใช้ทั้งสองแนวทางประกอบกัน คืออาจจะใช้กฎในเบื้องต้นจากนั้นจึงส่งผลให้แบบจำลองทางสถิติตัดสินใจ

งานวิทยานิพนธ์มหัศจรรย์ของของนชชา ธีระสาโรช (2553) เน้นศึกษาการสร้างระบบการรู้จำชื่อเฉพาะที่เป็นชื่อบุคคล ชื่อสถานที่ และชื่อองค์กรในภาษาไทย

โดยใช้แบบจำลองทางสถิติที่เรียกว่าแบบจำลองคอนดิชันนอลแรนดอมฟิลด์ส (Conditional Random Fields) ศึกษาโดยใช้คลังข้อมูลข่าว 367,673 คำ ภายในมีชื่อเฉพาะทั้งสิ้น 16,179 ชื่อ เป็นชื่อบุคคล 5,672 ชื่อ ชื่อสถานที่ 3,934 ชื่อ และชื่อองค์กร 4,751 ชื่อ เนื่องจากงานวิจัยที่ผ่านมาใช้ข้อมูลเข้าเป็นคำทั้งสิ้น แต่ในงานรู้จำชื่อเฉพาะภาษาจีนหลายงาน (He and Wang 2008, Wu, Yang and Lin 2006) ได้ทดลองใช้ข้อมูลเข้าที่เป็นตัวอักษรซึ่งเทียบเคียงได้กับระดับพยางค์ในภาษาไทยแล้ว รายงานผลว่าได้ผลลัพธ์ที่ดีกว่าการใช้ข้อมูลเข้าเป็นแบบคำ นักวิจัยได้ทดลองการรู้จำชื่อเฉพาะโดยทดสอบทั้งแบบที่มีข้อมูลเข้าด้วยการตัดพยางค์กับแบบที่นำข้อมูลเข้าด้วยการตัดคำมาแล้ว แต่ผลที่ได้ไม่มีความแตกต่างกันระหว่างการใช้ข้อมูลเข้าทั้งสองแบบ ค่าความถูกต้องหรือ F-measure ที่เฉลี่ยจากทั้งความแม่นยำและครบถ้วนได้อยู่ที่ 81.30%

ในการทดลองใช้แบบจำลอง CRF นี้ นักวิจัยได้เตรียมข้อมูลที่มีการกำกับชื่อเฉพาะทั้งสามประเภทไว้ และแบ่งข้อมูลเป็นข้อมูลสำหรับฝึกระบบ 90% กับข้อมูลสำหรับทดสอบระบบอีก 10% โดยทำการทดลองสองแบบ แบบแรกใช้ข้อมูลที่มีการตัดคำแล้ว กับแบบที่สองใช้ข้อมูลที่มีเพียงการตัดพยางค์ จากนั้นจึงประเมินประสิทธิภาพของระบบเทียบกัน ในการทดลองได้ใช้ลักษณะต่างๆ ที่คาดว่าจะเกี่ยวข้องกับกรู้จำชื่อเฉพาะจำนวนหนึ่ง ได้แก่ การมีคำบ่งบอกชื่อเฉพาะหรือที่เราเรียกว่าคำบ่งบอก เช่น คำนำหน้าชื่อ การมีคำที่ตรงกับคำในรายการชื่อเฉพาะต่างๆ การพบคำย่อปรากฏรวม การมีคำในบริบทข้างเคียงเหมือนกัน การเป็นคำทั่วไปหรือไม่ การพบซ้ำในตัวบทเกินสามครั้ง และการดูรูปคำต่อเนื่องที่ละสองคำ (bigram) และสามคำ (trigram) ข้อมูลเข้าซึ่งจะเป็นข้อมูลที่แยกคำหรือแยกพยางค์แล้ว แต่ละหน่วยจะถูกนำมาพิจารณาหลักเกณฑ์ตามที่กำหนดและให้ค่าลักษณะที่วิเคราะห์ได้ พร้อมกับคำตอบว่าเป็นชื่อเฉพาะประเภทใด ข้อมูลเหล่านี้ถูกนำมาใช้ฝึกระบบรู้จำซึ่งจากตัวอย่างข้อมูลที่

ได้รับ ระบบจะให้น้ำหนักความสำคัญกับลักษณะต่างๆ ที่นำมาใช้เพื่อสร้างเป็นแบบจำลองที่จะใช้รู้จำชื่อเฉพาะสำหรับข้อมูลใหม่ๆ ที่เข้ามาได้

ในการให้คำตอบก็ได้ทดลองสร้างรูปแบบคำตอบหลายแบบ เนื่องจากชื่อเฉพาะสามารถประกอบจากคำหลายๆ คำ คำตอบที่กำกับไว้ในแต่ละคำจึงอาจเป็นแบบธรรมดา คือ กำกับ P O L สำหรับคำที่เป็นส่วนหนึ่งของชื่อบุคคล ชื่อองค์กร และชื่อสถานที่ ตามลำดับ หรืออาจมีการกำกับตำแหน่งของส่วนหน้าหลังของชื่อก็ได้ เช่น B-P, I-P, E-P สำหรับคำที่เป็นส่วนเริ่มต้นชื่อบุคคล ส่วนกลางของชื่อบุคคล และส่วนท้ายของชื่อบุคคล ตามลำดับ นอกจากนี้ในบางครั้ง ชื่อที่เป็นชื่อสถานที่บางครั้งก็ถูกใช้เพื่ออ้างถึงองค์กร เช่น ทำเนียบขาว ถูกใช้หมายถึงหน่วยงานแทนที่จะเป็นสถานที่ หรืออาจเป็นการอ้างกลับกัน เช่น รพ.จุฬาลงกรณ์ สามารถใช้เพื่ออ้างถึงสถานที่แทนที่จะเป็นองค์กรก็ได้ กรณีเช่นนี้ สามารถให้รายละเอียดคำตอบมากขึ้นเป็น LO, หรือ OL ได้ ในงานนี้ นักวิจัยได้ทดลองกำกับคำตอบไว้ห้ารูปแบบ (X คือคำอื่น ๆ ที่ไม่ใช่ P, O, L) ได้แก่

1. P, O, L, X
2. B, I, X - P, O, L
3. B, I, X - P, O, L, LO, OL
4. B, I, E, X - P, O, L
5. B, I, E, X - P, O, L, LO, OL

ผลจากการทดลองพบว่าการมีจำนวนรูปแบบคำตอบยิ่งมากมีแนวโน้มจะช่วยให้ระบบรู้จำประเภทของชื่อเฉพาะได้ถูกต้องมากขึ้น แต่ทั้งนี้ขึ้นกับปริมาณข้อมูลที่ใช้ฝึกด้วยว่าจะต้องไม่มีปริมาณน้อยเกินไป นอกจากนี้ในการเปรียบเทียบประสิทธิภาพของลักษณะต่างๆ ที่ใช้ในวิทยานิพนธ์นี้ พบว่าลักษณะที่เป็นรายการชื่อเฉพาะจะช่วยในการรู้จำได้มากในกรณีที่มีข้อมูลเข้าเป็นคำเพราะระบบจะเห็นตัวอย่างคำที่คัดมาจากรายการชื่อเฉพาะต่างๆ ลักษณะที่เป็นคำย่อมีส่วนช่วยในการรู้จำชื่อเฉพาะรองลงมา

เพราะชื่อเฉพาะจำนวนมากมักปรากฏโดยมีคำย่อเป็นตัวบ่งขอบเขตด้วย เช่น ค.ช. รมว. ลักษณะที่เป็นเรื่องคำบริบทคือให้ดูคำที่อยู่หน้าหลังไม่เกินสามคำของชื่อเฉพาะ ซึ่งเดิมคิดว่าชื่อสถานที่ ชื่อบุคคล ชื่อองค์กร น่าจะมีคำปรากฏรวมที่ใช้บ่งบอกชื่อเหล่านั้นได้ เช่น คำว่า “ที่” “ของ” แต่เมื่อทดสอบจริงแล้วลักษณะนี้ไม่ได้ช่วยในการรู้จำมากนัก เหตุเพราะคำที่คาดว่าจะเป็นตัวบ่งชี้ชื่อเฉพาะได้นั้นก็เป็นคำที่ใช้กับคำทั่วไปด้วย ส่วนลักษณะที่ดูว่าคำนั้นเป็นคำทั่วไปหรือไม่โดยคิดว่าชื่อเฉพาะจะประกอบด้วยคำที่ต่างจากคำทั่วไป ผลที่ได้ไม่พบว่ามีส่วนช่วยในการรู้จำชื่อเฉพาะเท่าใดนัก ทั้งนี้เพราะชื่อเฉพาะจำนวนมากก็ประกอบด้วยคำที่เป็นคำทั่วไปด้วยได้ ส่วนลักษณะที่เป็นการพบซ้ำในตัวอย่างเกินสามครั้งพบว่าไม่ได้ช่วยในการรู้จำ แต่กลับจะมีผลทำให้ความถูกต้องลดน้อยลงเพราะคำทั่วไปก็มีลักษณะการพบซ้ำในตัวอย่างมากกว่าสามครั้งได้ ระบบการรู้จำชื่อเฉพาะที่ทดลองในงานนี้สามารถรู้จำชื่อเฉพาะโดยมีความถูกต้องวัดจากค่า F-measure คือคิดทั้งค่าความแม่นยำและครบถ้วนได้ที่ 81%

ผลจากการศึกษานี้ชี้ให้เห็นว่าการรู้จำชื่อเฉพาะเป็นเรื่องที่ค่อนข้างยาก แม้ว่าเมื่อพิจารณารูปแบบการปรากฏของชื่อเฉพาะในภาษาไทยแล้ว น่าจะมีรูปแบบเฉพาะอยู่ เช่น ชื่อบุคคลมีองค์ประกอบของคำนำหน้าชื่อ + ชื่อ + นามสกุล ชื่อองค์กรมีรูปแบบ คำบ่งบอกชื่อ + ชื่อ + คำบ่งบอกหลัง เช่น บริษัทเสริมสุข จำกัด (มหาชน) ชื่อสถานที่ก็มีรูปแบบประกอบด้วย คำบ่งบอก + ชื่อ เช่น จังหวัดกำแพงเพชร แต่เมื่อพิจารณาข้อมูลภาษาไทยโดยละเอียด ก็จะพบว่าคำที่ใช้บ่งบอกชื่อเฉพาะได้ไม่จำเป็นต้องปรากฏร่วมกับชื่อเฉพาะเสมอ เช่น “ไปจังหวัดไหนมา” คำว่า “ไหนมา” ที่พบหลังคำบ่งบอกก็ได้เป็นชื่อสถานที่ หรือชื่อเฉพาะเองก็สามารถปรากฏโดยไม่ต้องมีคำบ่งบอกได้ เช่น “ไปเลยมาสนุกมาก” ในกรณีนี้ ชื่อสถานที่คือจังหวัดเลย ก็ปรากฏโดยไม่มีคำบ่งบอกใดใด และคำว่า “เลย” นี้ก็เป็นคำทั่วไปได้ในบริบทอื่นด้วย เช่น “เขาไม่ชอบเรียนหนังสือเลย” “เขาเลยไม่ไปโรงเรียน” “เขาเดิน

เลยไป” เป็นต้น ดังนั้น หากต้องการพัฒนาการรู้จำชื่อเฉพาะด้วยการมองหากฎหรือรูปแบบเฉพาะของชื่อก็จะใช้ไม่ได้ในทุกกรณี วิทยานิพนธ์ของ อมรทิพย์ กวินปนิธาน (2546) ได้ศึกษาคำบ่งบอกที่ปรากฏร่วมกับชื่อเฉพาะเพื่อจะตอบคำถามว่าคำเหล่านี้มีประสิทธิภาพใช้บ่งชี้ชื่อเฉพาะได้ดีเพียงใด โดยพิจารณาจากสัดส่วนของคำที่ปรากฏร่วมกับชื่อเฉพาะกับที่ปรากฏร่วมกับคำที่ไม่ใช่ชื่อเฉพาะ เพื่อหาว่าจะมีคำบ่งบอกใดบ้างที่สามารถใช้เป็นตัวบ่งชี้ชื่อเฉพาะได้ ผลจากการศึกษาชื่อเฉพาะในข้อมูลที่คัดเลือกมาพบว่าชื่อเฉพาะ จำนวน 44% นั้นถูกใช้โดยที่ไม่มีคำบ่งบอกหน้าหรือหลัง และคำบ่งบอกต่างๆ นั้นมีขีดความสามารถในการบ่งชี้ที่แตกต่างกัน คำเช่น *อาคาร ทะเล สถาบัน ต. วัด* ฯลฯ มีขีดความสามารถไม่ถึง 10% คำเช่น *นาง จังหวัด เชียงใหม่ น.ส. ฯลฯ* มีขีดความสามารถปานกลางที่ 50-60% คำที่มีขีดความสามารถสูง เช่น *ร.ร. ทวีป แม่น้ำ ร.พ. ฯลฯ* คำเหล่านี้มักปรากฏกับชื่อเฉพาะจึงมีขีดความสามารถกว่า 90% (อมรทิพย์ 2546:71) ลักษณะที่พบเช่นนี้จึงเอื้อต่อวิธีการรู้จำชื่อเฉพาะแบบอิงสถิติที่ระบบอาศัยข้อมูลที่เคยเห็นการฝึกและประเมินขีดความสามารถของคำบ่งบอกชื่อเฉพาะจากข้อมูลที่พบเพื่อนำมาใช้ตัดสินใจภายหลัง นอกจากการศึกษาคำบ่งบอกแล้ว อมรทิพย์ (2546) ยังได้ศึกษาถึงการเว้นวรรคที่มักปรากฏร่วมกับชื่อเฉพาะว่าสามารถบ่งชี้ชื่อเฉพาะได้มากน้อยเพียงใด ผลพบว่าการพิจารณาทั้งคำบ่งบอกร่วมกับการเว้นวรรคทำให้ขีดความสามารถในบ่งชี้ชื่อเฉพาะเพิ่มมากขึ้น เพราะจากข้อมูลที่ศึกษา ชื่อเฉพาะที่ปรากฏการเว้นวรรคร่วมด้วยมี 1,809 ชื่อ ที่ไม่ปรากฏการเว้นวรรคมี 738 ชื่อ ที่ไม่ปรากฏทั้งการเว้นวรรคและคำบ่งบอกมีจำนวน 447 ชื่อ (อมรทิพย์ 2546:71)

อย่างไรก็ตาม คำบ่งบอกชื่อเฉพาะอย่างเดียวนั้นคงไม่เพียงพอแก่การเรียนรู้ของเครื่องที่จะใช้รู้จำชื่อเฉพาะภาษาไทยได้ งานของนักชาได้พยายามนำลักษณะอื่นๆ มาประกอบการตัดสินใจด้วย ซึ่งก็มีทั้งที่เอื้อและไม่เอื้อต่อการรู้จำดังที่ได้กล่าวมาแล้ว ความสำเร็จของการรู้จำชื่อเฉพาะนั้นอยู่ที่การมองหาลักษณะที่เหมาะสมและเป็น

ประโยชน์ต่อการรู้จักชื่อเฉพาะให้ได้ ซึ่งเป็นโจทย์วิจัยที่ต้องศึกษากันต่อไป แต่การที่จะพิจารณาได้ว่าลักษณะใดน่าจะมีบทบาทสำคัญในการรู้จักชื่อเฉพาะนั้น การศึกษาการปรากฏใช้ชื่อเฉพาะทั้งสามประเภทนี้ในบริบทจริง น่าจะเป็นเรื่องพื้นฐานที่จะช่วยให้เราเข้าใจธรรมชาติของการใช้ชื่อเฉพาะในภาษาไทยมากขึ้น วิทยานิพนธ์มหาบัณฑิตของศศิวิมล กาลันสีมา (2553) มุ่งตอบคำถามพื้นฐานนี้ โดยมีจุดมุ่งหมายเพื่อศึกษาหารูปแบบและวิเคราะห์โครงสร้างของชื่อบุคคล ชื่อสถานที่ และชื่อองค์กร และศึกษาว่าเมื่อชื่อเฉพาะถูกใช้ต่อเนื่องในบริบท รูปแบบการปรากฏของชื่อเฉพาะจะเป็นอย่างไร มีการลดรูปลงหรือไม่ พร้อมทั้งศึกษารณียของการใช้ชื่อเฉพาะที่อ้างข้ามประเภท คือ ชื่อสถานที่ที่ใช้อ้างถึงองค์กร และชื่อองค์กรที่ใช้อ้างถึงสถานที่ เพื่อจะหาคำตอบว่าเราจะรู้ได้อย่างไรว่าชื่อเฉพาะนั้นใช้อ้างถึงประเภทของตัวเองหรืออ้างข้ามประเภทไป การศึกษาในงานวิทยานิพนธ์นี้จึงเป็นการศึกษาพื้นฐานทางภาษาศาสตร์ โดยคาดว่าผลการศึกษาน่าจะให้คำตอบเกี่ยวกับธรรมชาติของการใช้ชื่อเฉพาะภาษาไทยได้ในระดับหนึ่ง

ศศิวิมล กาลันสีมา (2553) ได้สรุปผลการศึกษาเรื่องชื่อเฉพาะภาษาไทยในเชิงภาษาศาสตร์ซึ่งมีอยู่จำนวนหนึ่ง โดยมักเป็นการศึกษากลวิธีในการตั้งชื่อ รูปแบบชื่อที่ตั้งและความหมายที่ได้ ตัวอย่างเช่น งานของปานทิพย์ มหาไตรภพ (2522) ใช้วิธีการวิเคราะห์ทางอรรถศาสตร์ชาติพันธุ์เพื่อหาเกณฑ์ที่ใช้ในการตั้งชื่อนามสกุลพระราชทานในสมัยรัชกาลที่หกซึ่งพบว่าการตั้งชื่อสกุลตามลักษณะชื่อสกุลเดิมและที่ตั้งตามลักษณะทางสังคม คือ ตามอาชีพ สถานที่และชาติพันธุ์ งานของสุภาพรณณ บางช้าง (2526) ที่ศึกษาการใช้ภาษาในการตั้งชื่อของคนไทยตั้งแต่สมัยสุโขทัยจนถึงปัจจุบัน โดยดูเรื่องจำนวนพยางค์และความหมายที่ใช้ในการตั้งชื่อ เพื่อสรุปถึงค่านิยมและความเชื่อที่สะท้อนจากการตั้งชื่อในสมัยต่างๆ นอกจากการศึกษาชื่อบุคคล ยังมีงานของปราณี กุลละวณิชย์ (2535) ที่ศึกษาการตั้งชื่อหมู่บ้านจ้วงในมณฑลกวางสีและในภาคตะวันออกเฉียงเหนือของประเทศไทย เพื่อเปรียบเทียบความเหมือนหรือความ

แตกต่างของกลวิธีที่ใช้ในการตั้งชื่อ งานของน้องนุช มณีอินทร์ (2519) ที่ศึกษาการเปลี่ยนแปลงของชื่อหมู่บ้านในจังหวัดเชียงใหม่เพื่อให้เข้าใจลักษณะทางวัฒนธรรมที่สะท้อนจากชื่อหมู่บ้าน งานของสุจริตลักษณ์ ตีผดุงและวชิราภรณ์ วรรณดี (2544) ที่ศึกษาชื่อหมู่บ้านในภาคตะวันออก 8 จังหวัด โดยดูโครงสร้างทางภาษาและความหมายของชื่อหมู่บ้าน การศึกษาชื่อเฉพาะต่างๆ นี้จึงมีลักษณะเป็นการศึกษาโครงสร้างและจำนวนพยางค์ที่ใช้ ความหมายที่ได้จากการตั้งชื่อ การเปลี่ยนแปลงที่เกิดขึ้น ศศิวิมล (2553) จึงได้อาศัยแนวทางการศึกษาที่ผ่านมาวิธีวิเคราะห์รูปแบบและโครงสร้างของชื่อเฉพาะทั้งสามประเภท คือ ชื่อบุคคล ชื่อสถานที่ และชื่อองค์กร เพื่อหาข้อสรุปที่จะเป็นประโยชน์ต่อการพัฒนาระบบการรู้จำชื่อเฉพาะได้

ผลการศึกษาของศศิวิมล (2553:30-46) รูปแบบการปรากฏของชื่อเฉพาะที่พบไม่ได้ต่างจากงานของอมรทิพย์ (2546) เช่น ชื่อบุคคลที่พบ 762 ชื่อมีรูปแบบเป็น คำบ่งบอก+ชื่อ+สกุล (62.99%), คำบ่งบอก+ชื่อ (31.5%), ชื่อ+สกุล (3.94%), ชื่อ (1.57%) ส่วนชื่อสถานที่ 774 ชื่อที่พบมีรูปแบบเป็น คำบ่งบอก+ชื่อสถานที่ (60.85%), ชื่อสถานที่ (39.15%) และสำหรับชื่อองค์กร 1,418 ชื่อมีรูปแบบเป็น คำบ่งบอก+ชื่อองค์กร+คำบ่งบอกหลัง (3.46%), คำบ่งบอก+ชื่อองค์กร (55.29%), ชื่อองค์กร+คำบ่งบอกหลัง (0.28%), ชื่อองค์กร (40.97%)

ในส่วนของการวิเคราะห์โครงสร้าง ศศิวิมลได้แยกชื่อที่พบออกเป็นคำต่างๆ หากชื่อนั้นประกอบด้วยคำมากกว่าหนึ่งคำก็จะวิเคราะห์โครงสร้างและความสัมพันธ์ระหว่างคำ การวิเคราะห์ส่วนนี้ทำเฉพาะชื่อองค์กรและสถานที่เพราะมีองค์ประกอบคำที่สามารถแยกคำได้ชัดเจน กว่าชื่อบุคคลซึ่งมักมีลักษณะของคำบาลีสันสกฤตอยู่มาก แยกคำที่มีความหมายได้ยาก ผลการวิเคราะห์ชื่อสถานที่ 774 ชื่อ ที่เป็นคำย่อมี 25 ชื่อ จึงเหลือชื่อสถานที่ 749 ชื่อ เมื่อตัดชื่อที่ซ้ำกันออกไปจะเหลือชื่อสถานที่ 274 ชื่อที่จะนำมาวิเคราะห์โครงสร้าง ผลที่ได้พบชื่อสถานที่ที่เป็นคำเดียว (X) 29.92% เช่น เขต[ดุสิต] ชื่อที่ประกอบด้วยคำสองคำ (X+X) 50% เช่น ตำบล

[หนอง+บัว] ชื่อที่ประกอบด้วยคำกับลี (X+XP) มี 5.11% เช่น ต.[หนอง+[ปาก+โลง]] ชื่อที่ประกอบด้วยวลีมากกว่าหนึ่ง (XP+XP) มีเพียง 0.73% เช่น ถนน [[พุทธ+มณฑล]+[สาย+2]] นอกจากนี้ ยังพบว่าชื่อสถานที่ในโครงสร้างที่กล่าวมาอาจเป็นคำทับศัพท์ได้ด้วย โดยโครงสร้าง X ที่เป็นคำทับศัพท์พบ 11.28% เช่น ถนน[คอนแวนต์], แบบ X+X พบ 0.36% เช่น [ซิน+อพาร์ทเมนท์], แบบ X+XP พบ 0.73% เช่น หมู่บ้าน [[พาร์ค+วิว]+ซีดี] และยังพบชื่อสถานที่ที่ประกอบด้วยทั้งคำไทยและคำทับศัพท์แบบ X+XP อีก 1.46% เช่น หอพัก [[ครู+ฉวี]+เพลส] จากข้อสรุปนี้ จะเห็นได้ว่า ชื่อสถานที่ในภาษาไทยมีลักษณะเป็นคำสั้นๆ โดยเป็นคำสองคำมากที่สุด ตามมาด้วยแบบที่เป็นคำเดียว ทั้งสองแบบนี้ครอบคลุมประมาณ 90% ของชื่อสถานที่แล้ว

สำหรับชื่อองค์กร จากจำนวน 1,418 พบที่เป็นคำย่อ 579 ชื่อ จึงมีชื่อองค์กรเหลือ 839 ชื่อที่จะนำมาวิเคราะห์โครงสร้างภายใน และเมื่อคัดชื่อที่ซ้ำกันออกไปก็จะเหลือชื่อองค์กร 425 ชื่อ ผลการวิเคราะห์โครงสร้าง พบดังนี้ ชื่อที่เป็นคำเดียว (X) มี 13.18% เช่น โรงแรม[อโนมา] ชื่อที่เป็นคำประสมสองคำ (X+X) มี 22.59% เช่น สมาคม[ค้า+ทองคำ] ชื่อที่เป็นคำกับลี (X+XP) มี 17.18% เช่น สำนักงาน [บริหาร+[หนี้+สาธารณะ]] ชื่อที่ประกอบด้วยวลีมากกว่าหนึ่งวลี (XP+XP) พบ 28% เช่น คณะกรรมการ[[การ+[ป้องกัน+และ+ปราบปราม]+[การ+ทุจริต]]]+[แห่ง+ชาติ]] นอกจากนี้ชื่อองค์กรที่เป็นคำไทย ชื่อที่เป็นคำทับศัพท์ยังพบแบบ X จำนวน 3.53% แบบ X+X 7.29% แบบ X+XP 2.59% ส่วนชื่อที่มีทั้งคำไทยและคำทับศัพท์พบแบบ X+X 2.35% เช่น บริษัท[ไทย+เบฟเวอเรจ] แบบ X+XP 2.35% เช่น โรงแรม[[สวน+ดุสิต]+เพลส] แบบ XP+XP 0.94% เช่น บริษัท[เซลล์+[แห่ง+ประเทศ+ไทย]] จากผลที่ได้นี้ จะเห็นได้ว่าชื่อองค์กรนั้นแม้จะมีลักษณะโครงสร้างแบบเดียวกับชื่อสถานที่ แต่ชื่อองค์กรนั้นส่วนใหญ่จะมีโครงสร้างที่ซับซ้อนมากกว่าชื่อสถานที่ เห็นได้จากโครงสร้าง XP+XP และ X+XP ที่มีจำนวนรวมประมาณ 50% ในขณะที่ชื่อสถานที่มีประมาณ 7% เท่านั้น

ผลที่ได้จากการศึกษาของศศิวิมล สะท้อนให้เห็นว่าชื่อเฉพาะแต่ละประเภทในภาษาไทยมีรูปแบบและมีสัดส่วนของโครงสร้างที่แตกต่างกัน การรู้จำชื่อเฉพาะอาจจะต้องพิจารณาเป็นงานของการรู้จำชื่อเฉพาะแต่ละประเภทก็ได้ เพราะลักษณะที่เหมาะสมสำหรับชื่อแต่ละประเภทอาจแตกต่างกัน เช่น ชื่อบุคคลอาจใช้ประโยชน์จากคำบ่งบอกชื่อได้มากกว่าโดยที่ชื่อบุคคล 94% จะมีคำบ่งบอกปรากฏด้วย (ศศิวิมล 2553:49) ใช้การเว้นวรรคบ่งชี้ได้ง่ายกว่าชื่อประเภทอื่น และอาจจะดูรูปคำที่เป็นคำบาลีสันสกฤตได้ซึ่งในการประมวลผลภาษาคำเหล่านี้อาจถูกจัดเป็นคำที่ระบบไม่รู้จักเนื่องจากไม่มีในพจนานุกรมที่ใช้ และเมื่อเปรียบเทียบชื่อสถานที่กับชื่อบุคคล ชื่อสถานที่จะมีความยาวที่น้อยกว่า ไม่มีโครงสร้างที่ซับซ้อน ในขณะที่ชื่อบุคคลส่วนใหญ่แล้วจะใช้แต่คำทั่วไปและมีโครงสร้างเหมือนวลีที่ซับซ้อนได้ มีการใช้คำว่า “การ” “ความ” มีการใช้คำเชื่อม “และ” คำบุพบท “แห่ง” “เพื่อ” ที่ช่วยขยายโครงสร้างให้ซับซ้อนขึ้น ชื่อบุคคลจึงเป็นชื่อที่ยากต่อการรู้จำให้ถูกต้องเพราะมักจะแยกความแตกต่างจากวลีโดยทั่วไปไม่ได้ เช่น ชื่อบุคคล “กองทุนเพื่อการฟื้นฟูและพัฒนา ระบบสถาบันการเงิน” หากผู้อ่านไม่ทราบว่าเป็นชื่อบุคคล ก็อาจตีความจากความหมายของคำที่นำมาประกอบกันก็ได้ เหมือนกับที่ตีความวลี “กองทุนเพื่อช่วยเหลือสถาบันการเงิน” (ตัวอย่างจาก อมรทิพย์ 2546:18) ซึ่งการตั้งชื่อบุคคลนั้นเรามักพบว่าชื่อที่ตั้งนั้นพยายามให้สื่อถึงหน้าที่ขององค์กรที่ตั้งขึ้น ชื่อบุคคลจึงค่อนข้างจะไม่แตกต่างจากวลีทั่วไป แต่ก็สามารถตั้งข้อสังเกตได้ว่า ชื่อบุคคลนั้นมักจะใช้คำที่มีความหมายเชิงบวก เช่น ฟื้นฟู พัฒนา หากพบข้อความว่า องค์กรเพื่อทำลายประชาธิปไตย แบบนี้ก็ไม่น่าจะสันนิษฐานได้ว่าไม่ใช่ชื่อบุคคล

งานของศศิวิมล (2553) ยังได้ศึกษาถึงการใช้ชื่อเฉพาะแบบต่อเนื่องในตัวบทเพื่อควารูปแบบของชื่อเฉพาะต่างๆ เมื่อมีการอ้างอิงอย่างต่อเนื่องแล้ว จะยังคงรูปเดิมหรือเปลี่ยนรูปไปเมื่อเทียบกับรูปแบบตั้งต้นที่ใช้อ้างอิงครั้งแรกในตัวบท ซึ่งผลที่ได้นี้จะช่วยให้เข้าใจการปรากฏของชื่อเฉพาะและวางระบบการรู้จำชื่อเฉพาะได้ดีมากขึ้น

ผลการศึกษาจากชุด (series) ชื่อเฉพาะทั้งสามประเภทพบจำนวน 523 ชุด ซึ่งเป็นชื่อบุคคล 157 ชุด ชื่อสถานที่ 117 ชุด ชื่อองค์กร 249 ชุด พบการเปลี่ยนแปลงในสามลักษณะ คือ ใช้รูปแบบเดิมไปอย่างต่อเนื่อง ใช้รูปต่อไปต่างจากรูปตั้งต้นและคงที่รูปใหม่นั้นไว้ กับใช้รูปต่อไปแบบไม่คงที่ การเปลี่ยนรูปที่พบส่วนใหญ่จะเป็นการลดรูปลง เช่น ชื่อบุคคลจากรูปตั้งต้นที่มี คำบ่งบอก+ชื่อ+นามสกุล ก็อาจลดลงเหลือ คำบ่งบอก+ชื่อ เป็นต้น

สำหรับชื่อบุคคลพบว่า 28 ชุดใช้รูปเดียวกับรูปตั้งต้น อีก 120 ชุด รูปต่อไปจะเปลี่ยนแปลงคงที่ซึ่งเกือบทั้งหมดเป็นการลดรูปลง มีเพียงตัวอย่างเดียวที่เปลี่ยนโดยเพิ่มรูปแบบขึ้น และอีก 9 ชุดที่มีการเปลี่ยนแปลงแบบไม่คงที่ ส่วนชื่อสถานที่พบว่า 72 ชุดไม่มีการเปลี่ยนรูปจากรูปตั้งต้น มี 21 ชุดที่เปลี่ยนแบบคงที่เป็นการลดรูปแบบ 14 ชุด เพิ่มรูปแบบ 5 ชุด อีก 1 ชุดเปลี่ยนคำที่ใช้เลย (กทม.->กรุงเทพฯ) และมี 24 ชุดที่มีการเปลี่ยนแปลงแบบไม่คงที่ และสำหรับชื่อองค์กร พบว่าจำนวน 79 ชุดไม่มีการเปลี่ยนรูป อีก 140 ชุดมีการเปลี่ยนแปลงคงที่ซึ่งเป็นการลดรูปแบบ 31 ชุด เพิ่มรูปแบบ 15 ชุด และที่เป็นการย่อคำมี 94 ชุด ส่วนที่เปลี่ยนรูปแบบไม่คงที่นั้นมี 30 ชุด

จากผลที่ได้จะเห็นว่า การใช้ชื่อบุคคลแบบต่อเนื่องนั้น ส่วนมากเป็นการลดรูปแบบ คือ รูปตั้งต้นอาจมีองค์ประกอบครบ แต่เมื่ออ้างต่อไปก็สามารถลดนามสกุลลงได้ ส่วนชื่อสถานที่นั้นส่วนมากจะใช้รูปคงที่ อาจจะด้วยเหตุว่าการลดรูปแบบเช่นการละคำบ่งบอกอาจทำให้ความหมายไม่ชัดเจนได้ เช่น เมื่อกล่าวถึง อ.เมือง หากละคำบ่งบอก อ. ก็จะไม่สามารถอ้างถึงสถานที่เดิมได้อย่างชัดเจน ในกรณีชื่อที่สามารถละคำบ่งบอกได้ เช่น กรุงเทพฯ ก็มักจะไม่พบคำบ่งบอก “จังหวัด” ตั้งแต่ในรูปตั้งต้นแล้ว แต่การลดรูปแบบก็มีพบได้ เช่น ประเทศไทย->ไทย ส่วนชื่อองค์กรนั้นส่วนใหญ่มีการเปลี่ยนรูปแบบคงที่ คือรูปแบบที่ใช้จะต่างจากรูปตั้งต้น ซึ่งมักเป็นการเปลี่ยนเป็นคำย่อมากที่สุด เช่น สำนักงานการปฏิรูปที่ดินเพื่อเกษตรกรรม->ส.ป.ก. ตามด้วยการลด

รูปแบบ เช่น *สำนักข่าวรอยเตอร์->รอยเตอร์* *กรมพัฒนาที่ดิน->กรมฯ* ซึ่งในกรณีเช่นนี้ เราจะเห็นบทบาทของการใช้ ฯ และคำย่อในชื่อองค์กร

สำหรับการเปลี่ยนแปลงแบบไม่คงที่ คือ รูปตั้งต้นที่ไม่มีการเปลี่ยนเมื่ออ้างถึงต่อไปและอาจมีการกลับมาใช้รูปเหมือนรูปตั้งต้นอีก เช่น *กรมส่งเสริมสหกรณ์->กรมฯ->กรมฯ->กรมส่งเสริมสหกรณ์->กรมฯ* ผลตรงนี้อาจมาจากการที่ผู้วิจัยมองข้อมูลแบบต่อเนื่องทั้งตัวบท จึงทำให้เห็นเช่นนี้ หากมีการมองตามโครงสร้างปริจเฉทแล้ว อาจเป็นไปได้ว่ารูปแบบมีการลดลงภายในกรอบของ discourse segment แต่เมื่อเริ่มต้น discourse segment ใหม่ ชื่อนั้นก็อาจต้องกลับมาใช้ในรูปแบบเดิมอีกครั้ง เป็นต้น อย่างไรก็ตาม ข้อมูลส่วนนี้อาจจะไม่สำคัญสำหรับการรู้จำชื่อเฉพาะมากเท่ากับว่ารูปแบบการเปลี่ยนแปลงเป็นอย่างไรได้บ้างและมีแนวโน้มเป็นแบบใดสำหรับชื่อเฉพาะแต่ละประเภท

นอกจากการศึกษาเรื่องรูปแบบและโครงสร้างของชื่อเฉพาะ ศศิวิมล (2553: 71-91) ยังได้ศึกษาชื่อเฉพาะที่มีการอ้างข้ามประเภทเพื่อหาสิ่งที่จะใช้บ่งชี้การอ้างข้ามประเภทได้ ชื่อเฉพาะที่สามารถใช้การอ้างข้ามประเภทได้มีสองแบบดังที่กล่าวมาข้างต้น คือ ชื่อสถานที่ใช้อ้างถึงองค์กร และชื่อองค์กรใช้อ้างถึงสถานที่ การที่ชื่อสองประเภทนี้สามารถใช้อ้างแทนกันได้เพราะต่างก็มีคุณสมบัติของกันและกันอยู่ คือ องค์กรก็ต้องมีสถานที่ตั้ง สถานที่ก็เป็นที่อยู่ขององค์กร จึงเกิดการใช้กระบวนการนามมัย (metonymy) มาทำให้ใช้แทนกันได้ การรู้จำชื่อเฉพาะจึงต้องจัดการกับข้อมูลส่วนนี้ได้ด้วย เพราะชื่อเดียวกันบางครั้งก็ใช้อ้างถึงสถานที่ บางครั้งก็ใช้อ้างถึงองค์กร เช่น ใน “*รณว.วิทยาศาสตร์และเทคโนโลยี เดินทางเข้ากระทรวงวิทยาศาสตร์ฯ*” *กระทรวงวิทยาศาสตร์ฯ* หมายถึงสถานที่ตั้ง แต่ถ้าเป็น “*กระทรวงวิทยาศาสตร์ฯ ได้ผลิตผลงานวิจัยที่เป็นประโยชน์*” *กระทรวงวิทยาศาสตร์ฯ* จะหมายถึงองค์กร ในการนี้ ศศิวิมลได้พิจารณาคำในบริบทใกล้เคียงหน้าหลังชื่อเฉพาะไม่เกินห้าคำ และพบว่า มีคำบางคำอาจใช้บ่งชี้ได้ เช่น คำว่า “ที่” “ใน” “หน้า” เมื่อปรากฏหน้าชื่อองค์กรจะ

บ่งชี้ว่าชื่อนั้นมีแนวโน้มว่าจะอ้างถึงสถานที่ได้ เช่น “ที่หน้าธนาคารกสิกรไทย” ส่วนคำว่า “ของ” สามารถใช้บ่งชี้ชื่อสถานที่ที่ตามหลังคำนี้ว่ามีแนวโน้มที่จะใช้อ้างถึงองค์กร เช่น “ปัญหาของเวียดนาม” เวียดนามในที่นี้ไม่ได้หมายถึงสถานที่ตั้งของประเทศแต่หมายถึงหน่วยงานคือรัฐที่รับผิดชอบ คำว่า “มี” และ “จะ” ซึ่งปรากฏหลังชื่อสถานที่ก็สามารถบ่งชี้ว่าชื่อนั้นใช้อ้างถึงองค์กรได้ อย่างไรก็ตาม เมื่อพิจารณาที่คำเหล่านี้เอง เช่น คำว่า “ที่” ก็จะพบว่ามีปรากฏใช้ร่วมกับคำอื่นๆ ด้วยมากเช่นกัน ดังนั้น โดยตัวคำนั้นเองอาจไม่สามารถใช้บ่งชี้ชื่อเฉพาะอ้างข้ามประเภทได้ แต่ในกรณีที่เรารู้ว่า ณ ตำแหน่งนั้นเป็นชื่อเฉพาะ เราจึงจะสรุปได้ว่าน่าจะเป็นการใช้อ้างข้ามประเภท

การรู้จำชื่อผลิตภัณฑ์

การรู้จำชื่อเฉพาะ นอกจากงานส่วนใหญ่ที่ศึกษาชื่อบุคคล ชื่อสถานที่ และชื่อองค์กรแล้ว ยังมีงานส่วนหนึ่งที่สนใจศึกษาชื่อเฉพาะประเภทอื่นดังที่ได้กล่าวมาในตอนต้น วิทยานิพนธ์มหาบัณฑิตของณัฐดาพร เลิศชีวะ (2553) เป็นงานที่สนใจศึกษาชื่อผลิตภัณฑ์ในภาษาไทย โดยศึกษาจากข้อมูลข่าวเศรษฐกิจเนื่องจากจะมีการกล่าวถึงผลิตภัณฑ์ต่างๆ อยู่มาก ในงานนี้ได้ชื่อผลิตภัณฑ์มาศึกษา 2,463 ชื่อจากคลังข้อมูลข่าวที่สร้างขึ้น 178,474 คำ งานวิจัยนี้มีลักษณะคล้ายกับงานของศศิวิมล (2553) เพราะต้องการศึกษาหารูปแบบต่างๆ ของชื่อผลิตภัณฑ์และดูการเปลี่ยนแปลงของรูปแบบเมื่อมีการอ้างถึงอย่างต่อเนื่องในตัวบท

ณัฐดาพร (2553) วิเคราะห์องค์ประกอบของชื่อผลิตภัณฑ์ว่ามีองค์ประกอบต่างๆ ดังนี้

1. คำบอกชนิดผลิตภัณฑ์ ซึ่งมักเป็นคำนามทั่วไปที่ใช้บอกถึงชนิดของผลิตภัณฑ์นั้น เช่น **น้ำพริกตาโอย่ง, รังนกแท้สกัด**
2. ชื่อตราสินค้าหรือ brand name ซึ่งเป็นชื่อที่ตั้งขึ้นเป็นแบรนด์ของผลิตภัณฑ์นั้น เช่น **น้ำปลาตราทิพรส, บุหรียี่ห้อมาร์ลโบโร**

3. คำบ่งบอกชื่อตราสินค้า คือ คำที่ใช้บ่งบอกว่าคำต่อไปคือชื่อตราสินค้า เช่น *น้ำปลาตราทิพรส, บุหรียี่ห้อมาร์ลโบโร*

4. ชื่อประเภทผลิตภัณฑ์ คือ ชื่อที่ใช้อ้างถึงรุ่น ชนิด ประเภทย่อยของผลิตภัณฑ์นั้น เช่น *ฮอนด้า แจ๊ส, มาม่ารสต้มยำกุ้ง*

5. คำบ่งบอกชื่อประเภทผลิตภัณฑ์ คือคำที่ทำให้รู้ว่าคำต่อไปคือประเภทย่อยของผลิตภัณฑ์นั้น เช่น *แวนแเรย์แบนรุ่น Rimless Aviator RB 3214, มาม่ารสต้มยำกุ้ง*

จากการศึกษาของณัฐดาพร พบว่าชื่อผลิตภัณฑ์มีรูปแบบต่างๆ จากการนำคำในห้วงคำประกอบนี้มารวมกันเป็นชื่อผลิตภัณฑ์ได้ 32 รูปแบบ ซึ่งเมื่อวิเคราะห์รูปแบบที่พบสามารถสรุปได้ว่าส่วนที่เป็นแกนหลักของชื่อผลิตภัณฑ์ คือ ตราสินค้าและชื่อประเภทผลิตภัณฑ์ โดยที่ชื่อผลิตภัณฑ์จะต้องมีส่วนใดส่วนหนึ่งนี้เสมอเพราะเป็นส่วนที่ชี้ให้เห็นถึงผลิตภัณฑ์นั้นได้ ปกติแล้วตราสินค้าจะเป็นคำหลักที่ใช้อ้างถึงผลิตภัณฑ์เพราะเป็นแบรนด์ที่สร้างขึ้นเพื่อจำแนกผลิตภัณฑ์นั้นออกจากสินค้าชนิดเดียวกันที่ผลิตจากผู้ผลิตอื่นๆ แต่ในบางกรณีที่ผลิตภัณฑ์นั้นมีชื่อของประเภทย่อยของผลิตภัณฑ์ การอ้างด้วยชื่อประเภทผลิตภัณฑ์โดยไม่มีตราสินค้าก็เป็นไปได้ เช่น “อิงค์แพด เอ็กซ์ 100 อี” เป็นชื่อประเภทผลิตภัณฑ์ที่อ้างถึงคอมพิวเตอร์โน้ตบุ๊กโดยไม่ต้องใช้ตราสินค้า IBM เข้าไว้ในชื่อผลิตภัณฑ์ด้วยก็ได้ คือไม่จำเป็นต้องเขียน “โน้ตบุ๊ก IBM รุ่นอิงค์แพด เอ็กซ์ 100 อี” เพราะผู้อ่านสามารถทราบได้จากบริบท

การวิเคราะห์ของณัฐดาพร (2553) สรุปได้ว่า ชื่อผลิตภัณฑ์ที่ปรากฏแต่คำหลักมี 43.32% ตามมาด้วยชื่อผลิตภัณฑ์ที่มีคำหลักอยู่ท้ายและมีองค์ประกอบอื่นอยู่หน้าซึ่งมี 41.41% เช่น *โทรศัพท์เคลื่อนที่แบรนด์แบล็กเบอร์รี่, ข้าวตราฉัตร, แบรินด์น้ำผลไม้แบร์รี่, รุ่นซีรีส์ 7 ซีดาน, เป็นต้น* แบบที่มีคำหลักอยู่หน้ามี 10.19% เช่น *แบล็กเบอร์รี่รุ่นโบลด์, อิงค์แพด ซีรีส์ เป็นต้น* และแบบที่มีคำหลักอยู่กลางมีเพียง

5.08% เช่น ยาสีฟันยี่ห้อฟลูโอคาริล 40 พลัส, รถเซฟโรเลต โคโลราโด ปิคอัพอเมริกันพันธุ์แกร่ง เป็นต้น

นอกจากการศึกษารูปแบบชื่อผลิตภัณฑ์ วัตถุประสงค์พยายามศึกษาหาคำที่จะใช้บ้างชื่อผลิตภัณฑ์ได้ โดยขั้นแรกคาดว่าคำที่เกี่ยวข้องกับผลิตภัณฑ์ เช่น “เจ้าของ” “ยอดชาย” “ผู้จำหน่าย” “การขาย” “การผลิต” ต่างๆ เหล่านี้น่าจะใช้บ้างได้ แต่จากการดูคำที่ปรากฏในบริบทร่วมกับชื่อผลิตภัณฑ์จริงๆ และจากการหาอำนาจในการจำแนก คือ นำคำที่ปรากฏรวมเหล่านั้นไปตรวจสอบกับบริบทอื่น ๆ ก็พบว่า คำเหล่านั้นก็พบว่าใช้กับคำอื่นๆ ทั่วไปด้วยไม่ได้ใช้เฉพาะกับชื่อผลิตภัณฑ์ ผลที่ได้นี้คล้ายคลึงกับที่อมรทิพย์ (2546) ศึกษาหาคำที่จะสามารถบ่งชี้ชื่อเฉพาะที่เป็นชื่อบุคคล ชื่อสถานที่ และชื่อองค์กร กล่าวโดยสรุปก็คือ คำต่างๆ ที่คาดว่าจะปรากฏร่วมกับชื่อเฉพาะนั้นมักเป็นคำที่ปรากฏร่วมกับคำทั่วไปอื่นๆ ด้วย เราจึงไม่สามารถใช้คำเหล่านี้โดยลำพังเพื่อใช้ในการรู้จักชื่อเฉพาะได้

ส่วนการศึกษารูปแบบชื่อผลิตภัณฑ์เมื่อมีการอ้างอิงอย่างต่อเนื่องในตัวเอง จากชื่อผลิตภัณฑ์ 2,463 ชื่อ มีชื่อที่ใช้ครั้งเดียว 1,101 ชื่อ ชื่อที่ใช้แบบต่อเนื่อง 1,326 ชื่อ ในกลุ่มหลังนี้ มีชื่อที่ใช้ต่อเนื่องสองครั้ง 452 ชื่อหรือ 226 คู่ ซึ่งพบว่า 202 ชื่อหรือ 101 คู่ นั้นไม่มีการเปลี่ยนรูปแบบ ในขณะที่อีก 202 ชื่อหรือ 101 คู่ เช่นกันมีการลดรูปแบบจากชื่อตั้งต้น เช่น ยางรถบรรทุกมิชลิน->มิชลิน ที่มีการเพิ่มรูปแบบมี 38 ชื่อหรือ 19 คู่ เช่น ซาบ->รถซาบ ลักษณะที่พบการเพิ่มรูปแบบนี้เป็นเพราะชื่อผลิตภัณฑ์คำแรกนั้นจะเป็นส่วนหนึ่งของชื่อแคมเปญหรืออยู่ในรายการที่แจกแจง การสลับตำแหน่งองค์ประกอบ เช่น สก๊อตตั้งนกแท้->รังนกแท้สก๊อต ก็มีพบบ้างอยู่ 4 คู่ และที่เปลี่ยนรูปแบบมีหนึ่งคู่ คือ รถยนต์นิวแลนเซอร์->มิติซูบิซิแลนเซอร์ ดังนั้น จากข้อมูลที่พบจะเห็นว่า การอ้างอิงครั้งที่สองส่วนใหญ่แล้ว (89%) เป็นการคงรูปชื่อผลิตภัณฑ์เดิมหรือไม่ก็ลดรูปชื่อลงเป็นส่วนใหญ่ ในกรณีที่มีการอ้างอิงมากกว่า 2 ครั้ง จากจำนวน 910 ชื่อ พบว่าเป็นชื่อผลิตภัณฑ์คิดเป็น 213 ชุด 58 ชุดนั้นเป็น

การคงรูปชื่อแบบรูปตั้งต้น 37 ชุดเป็นการลดรูปจากรูปตั้งต้นแล้วคงรูปใหม่ไว้ มีเพียงชุดเดียวที่เพิ่มรูปแบบแล้วคงรูปใหม่ไว้ ส่วนอีก 117 ชุดเป็นการเปลี่ยนรูปแบบไม่คงที่ การที่จะใช้รูปใดดูจะขึ้นกับบริบทที่เกิด หากปรากฏในชื่อแคมเปญก็จะละองค์ประกอบต่างๆเพื่อให้ชื่อนั้นกระชับง่ายต่อการจดจำ หรือเมื่อมีการกล่าวถึงเรื่องอื่นๆ แล้วกลับมาที่ผลิตภัณฑ์นั้นใหม่ ก็มักจะใช้รูปแบบที่เพิ่มขึ้น เป็นต้น

บทสรุปและข้อคิดเห็นเรื่องชื่อเฉพาะภาษาไทย

ผลจากการศึกษาชื่อเฉพาะในงานของวิทยานิพนธ์มหาบัณฑิตของนัชชา ถิระสาโรช (2553), ศสิวิมล กลั่นสีมา (2553), และ ญัฐดาพร เสิศชีวะ (2553) ทำให้เห็นลักษณะที่น่าสนใจในงานรู้จักชื่อเฉพาะภาษาไทย ที่ผู้เขียนคิดว่าเป็นทิศทางการศึกษาต่อไปดังนี้

1. ระบบการรู้จักชื่อเฉพาะแต่ละประเภทสามารถแยกออกจากกันได้ เนื่องจากชื่อเฉพาะแต่ละประเภทนั้นมีรูปแบบและวิธีการใช้ต่างกัน ลักษณะที่จะเอื้อต่อการรู้จักจึงแตกต่างกันไปได้
2. ระบบการรู้จักชื่อเฉพาะนั้นโดยหลักควรเป็นระบบที่อิงแบบจำลองทางสถิติ เพราะจากผลการศึกษานัชชา พบว่าในระบบพื้นฐานที่ไม่ได้อาศัยลักษณะใดเพิ่มเติม นอกจากใช้คำต่อเนื่องสองคำและสามคำ ระบบยังสามารถรู้จักชื่อเฉพาะได้ถูกต้องเกือบ 80% การเพิ่มความถูกต้องของระบบนั้นอยู่ที่การหาหลักเกณฑ์ที่เหมาะสมสำหรับชื่อเฉพาะแต่ละประเภท
3. การรู้จักชื่อเฉพาะนั้นควรใช้ความรู้หลายๆ ด้านประกอบกัน เพราะจากผลการศึกษาที่ผ่านมาจะเห็นว่า กลุ่มคำที่จะใช้ช่วยบ่งชี้ชื่อเฉพาะได้ชัดเจนนั้นหายาก เราไม่สามารถใช้เพียงลักษณะใดลักษณะหนึ่งตัดสินได้ แต่ต้องอาศัยลักษณะต่างๆ มาใช้พิจารณาไปพร้อมๆกัน
4. ชื่อบุคคลนั้นรู้จักได้ง่ายกว่าชื่อเฉพาะประเภทอื่น ด้วยลักษณะที่มีรูปแบบที่มีคำนำหน้าชื่อ และคำที่ใช้ในชื่อ นามสกุล ก็มักมีลักษณะที่ต่างจากคำทั่วไป

จึงเอื้อต่อการรู้จำได้ง่ายกว่า ในงานของนัชชา (2553) นั้น ชื่อบุคคลสามารถรู้จำได้ถูกต้องประมาณ 90% ในขณะที่ชื่อองค์กรและชื่อสถานที่สามารถรู้จำได้ประมาณไม่ถึง 80%

5. ชื่อองค์กรนั้นอาจจะยากที่จะรู้จำได้ถูกต้อง เพราะมีลักษณะที่เหมือนกับการเป็นวลีที่ใช้ทั่วไป การมีรายการชื่อเฉพาะไว้ล่วงหน้าจะช่วยในการตัดสินใจได้มาก แต่หากเป็นชื่อใหม่ที่เพิ่งตั้งขึ้น ปัญหานี้มีลักษณะคล้ายกับการหาศัพท์เฉพาะทางซึ่งมักมีโครงสร้างที่ซับซ้อนที่ไม่แตกต่างจากวลีทั่วไป ระบบที่พัฒนาจึงอาจไม่สามารถตัดสินใจได้เด็ดขาด อาจต้องสร้างตัวเลือกแบบต่างๆ แล้วให้ระบบการแจกส่วนประโยคซึ่งจะดูความสัมพันธ์ของคำทุกคำในประโยคช่วยตัดสินใจภายหลังได้

6. ชื่อเฉพาะประเภทอื่น ๆ นอกเหนือจากที่กล่าวถึงนี้ เช่น ชื่อสารเคมี ชื่อพืช ฯลฯ ก็ควรที่จะต้องมีวิธีการระบุรูปแบบการใช้ของชื่อประเภทนั้นและคัดเลือกลักษณะที่มองเห็นว่าน่าจะใช้บ่งชี้ชื่อประเภทนั้นๆ ได้

กิตติกรรมประกาศ

ผู้เขียนขอขอบคุณสำนักงานกองทุนสนับสนุนการวิจัยที่ได้ให้ทุนสนับสนุนการทำวิทยานิพนธ์มหาบัณฑิตของนางสาวนัชชา ธีระสาโรช และนางสาวณัฐดาพร เลิศชีวะ ในปีพ.ศ.๒๕๕๓ และขอบคุณนิสิตทั้งสองรวมทั้งนางสาวศศิวิมล กาลันสีมา ที่ได้ตั้งใจศึกษาค้นคว้าวิจัยจนงานสำเร็จลงได้ด้วยดี

สาริต

ผู้เขียนได้นำข้อมูลชื่อเฉพาะที่เป็นชื่อคน องค์กร สถานที่ ที่นัชชาและศศิวิมล ใช้ในงานวิทยานิพนธ์ 7,354 ชื่อตามที่ปรากฏในบริบทข้อมูลรวมประมาณ 183,300 คำ นำมาผนวกกับข้อมูลชื่อเฉพาะที่คัดลอกจากโครงการ AI for Thai เฉพาะไฟล์ที่ไม่มีปัญหาการแท็กข้อมูล (<https://aiforthai.in.th/>) รวมเป็นชื่อเฉพาะทั้งสิ้น 170,076 ชื่อ และพัฒนาโมดูล ner และ ner_tag เป็นส่วนหนึ่งของ TLTK โดยใช้วิธีการเรียนรู้

การประมวลผลภาษาไทย

ด้วย CRF ตามที่สาธิตใน (<http://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html>). ลักษณะที่ใช้สำหรับการเรียนรู้คือรูปคำและ POS ของคำในตำแหน่งซ้ายสองตำแหน่งและตำแหน่งขวาหนึ่งตำแหน่ง ขนาดของคำแยกเป็นสามประเภท (S 1-10 อักขระ, M 11-20 อักขระ, และ L > 20 อักขระ) BOS (เป็นคำเริ่มข้อความ), EOS (เป็นคำท้ายข้อความ) ผลที่ได้มีความถูกต้องประมาณ 88% ตามนี้

tag	precision	recall	f1-score	support
B-L	0.56	0.48	0.52	27,105
B-O	0.72	0.58	0.64	59,613
B-P	0.82	0.83	0.83	83,358
I-L	0.52	0.43	0.47	17,859
I-O	0.67	0.59	0.63	67,396
I-P	0.85	0.88	0.86	175,069
O	0.92	0.94	0.93	1,032,377
accuracy			0.88	1,462,777
macro avg	0.72	0.68	0.7	1,462,777
weighted avg	0.87	0.88	0.88	1,462,777

โมดูล `tltk.nlp.ner(...)` และ `tltk.nlp.ner_tag(...)` สามารถทดลองใช้ได้ใน Google colab ตามที่อยู่นี้

https://colab.research.google.com/drive/1YCrhKqO9YEVXgs-_HYXv4n5N73LdzF2Y?usp=sharing

`ner` เป็นการกำกับแต่ละคำว่าเป็นส่วนของชื่อเฉพาะที่เป็น B (begin) หรือ I (intermediate) หรือไม่ ส่วน `ner_tag` เป็นการแท็กชื่อเฉพาะด้วยแท็กเปิดและปิดของ <NEp> <NEo> <NEI> โดยใช้ข้อมูลที่ได้จากโมดูล `ner`

ner จะรับข้อมูลที่ผ่านการใส่ POS ตามแนวทางวิเคราะห์แบบ universal dependency มาแล้ว (tltk.nlp.pos_tag(...)) เพื่อมาเติมแท็ก O, B-P, I-P, B-L, I-L, B-O, I-O ให้แต่ละคำ ส่วน ner_tag เป็นโมดูลที่ทำทั้ง pos_tag และ ner และสรุปแท็กเปิดแท็กปิดของชื่อเฉพาะที่หาได้

```
import tltk
text = "สำนักงานเขตจตุจักรชี้แจงว่า ได้นำป้ายประกาศเตือนปลิงไปปักตามแหล่งน้ำ ในเขตอำเภอเมือง จังหวัดอ่างทอง หลังจากนายสุกิจ อายุ 65 ปี ถูกปลิงกัดแล้วไม่ได้ไปพบแพทย์"
print(tltk.nlp.ner_tag(text))
```

<NEo>สำนักงานเขตจตุจักร</NEo>ชี้แจงว่า ได้นำป้ายประกาศเตือนปลิงไปปักตามแหล่งน้ำ ใน<NEL>เขตอำเภอเมือง จังหวัดอ่างทอง</NEL> หลังจาก<NEp>นายสุกิจ</NEp> อายุ 65 ปี ถูกปลิงกัดแล้วไม่ได้ไปพบแพทย์

```
import tltk
text = "สำนักงานเขตจตุจักรชี้แจงว่า ได้นำป้ายประกาศเตือนปลิงไปปักตามแหล่งน้ำ ในเขตอำเภอเมือง จังหวัดอ่างทอง หลังจากนายสุกิจ อายุ 65 ปี ถูกปลิงกัดแล้วไม่ได้ไปพบแพทย์"
out = tltk.nlp.pos_tag(text)
print('Output from pos_tag',out)
for inx in out:
    print(tltk.nlp.ner(inx))
```

Output from pos_tag [[('สำนักงาน', 'NOUN'), ('เขต', 'NOUN'), ('จตุจักร', 'PROPN'), ('ชี้แจง', 'VERB'), ('ว่า', 'SCONJ'), ('<s/>', 'PUNCT')], [('ได้', 'AUX'), ('นำ', 'VERB'), ('ป้ายประกาศ', 'NOUN'), ('เตือน', 'VERB'), ('ปลิง', 'NOUN'), ('ไป', 'VERB'), ('ปัก', 'VERB'), ('ตาม', 'ADP'), ('แหล่งน้ำ', 'NOUN'), ('<s/>', 'PUNCT')], [('ใน', 'ADP'), ('เขต', 'NOUN'), ('อำเภอ', 'NOUN'), ('เมือง', 'NOUN'), ('<s/>', 'PUNCT')], [('จังหวัด', 'NOUN'), ('อ่างทอง', 'PROPN'), ('<s/>', 'PUNCT')], [('หลังจาก', 'SCONJ'), ('นาย', 'NOUN'), ('สุ', 'PROPN'), ('กิจ', 'NOUN'), ('<s/>', 'PUNCT')], [('อายุ', 'NOUN'), ('<s/>', 'PUNCT')], [('65', 'NUM'), ('ปี', 'NOUN'), ('<s/>', 'PUNCT')], [('ถูก', 'AUX'), ('ปลิง', 'VERB'), ('กัด', 'VERB'), ('แล้ว', 'ADV'), ('ไม่ได้', 'AUX'), ('ไป', 'VERB'), ('พบ', 'VERB'), ('แพทย์', 'NOUN'), ('<s/>', 'PUNCT')]]

[[('สำนักงาน', 'NOUN', 'B-O'), ('เขต', 'NOUN', 'I-O'), ('จตุจักร', 'PROPN', 'I-O'), ('ชี้แจง', 'VERB', 'O'), ('ว่า', 'SCONJ', 'O'), ('<s/>', 'PUNCT', 'O')]]

[[('ได้', 'AUX', 'O'), ('นำ', 'VERB', 'O'), ('ป้ายประกาศ', 'NOUN', 'O'), ('เตือน', 'VERB', 'O'), ('ปลิง', 'NOUN', 'O'), ('ไป', 'VERB', 'O'), ('ปัก', 'VERB', 'O'), ('ตาม', 'ADP', 'O'), ('แหล่งน้ำ', 'NOUN', 'O'), ('<s/>', 'PUNCT', 'O')]]

การประมวลผลภาษาไทย

[(ใน, 'ADP', 'O'), (เขต, 'NOUN', 'B-L'), (อำเภอ, 'NOUN', 'I-L'), (เมือง, 'NOUN', 'I-L'), (<s/>, 'PUNCT', 'I-L')]
[(จังหวัด, 'NOUN', 'B-L'), (อ่างทอง, 'PROPN', 'I-L'), (<s/>, 'PUNCT', 'O')]
[(หลังจาก, 'SCONJ', 'O'), (นาย, 'NOUN', 'B-P'), (สุ, 'PROPN', 'I-P'), (กิจ, 'NOUN', 'I-P'), (<s/>, 'PUNCT', 'O')]
[(อายุ, 'NOUN', 'O'), (<s/>, 'PUNCT', 'O')]
[(65, 'NUM', 'O'), (ปี, 'NOUN', 'O'), (<s/>, 'PUNCT', 'O')]
[(ถูก, 'AUX', 'O'), (ปลิง, 'VERB', 'O'), (กัด, 'VERB', 'O'), (แล้ว, 'ADV', 'O'), (ไม่ได้, 'AUX', 'O'), (ไป, 'VERB', 'O'), (พบ, 'VERB', 'O'), (แพทย์, 'NOUN', 'O'), (<s/>, 'PUNCT', 'O')]

อ้างอิง

- Chanlekha, H., and Kawtrakul, A. (2004). Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. In International Joint Conference of Natural Language Processing (IJCNLP-2004). Hainan Island.
- Chanlekha, H., Kawtrakul, A., Varasrai, P., and Mulasas, I. (2002). Statistical and Heuristic Rule Based Model for Thai Named Entity Recognition. In Proceeding of SNLP- Oriental COCOSDA 2002. Hua Hin.
- Charoenpornasawat, P., Kijisirikul, B., and Meknavin, S. (1998). Feature-based Proper Name Identification in Thai. In Proceedings of National Computer Science and Engineering Conference. Bangkok.
- He, J., and Wang, H. (2008). Chinese Named Entity Recognition and Word Segmentation Based on Character. In Proceeding of the Sixth SIGHAN Workshop on Chinese Language Processing, pp. 128-131. Hyderabad.

Wu, Y., Yang, J., and Lin, Q. (2006). Description of the NCU Chinese Word Segmentation and Named Entity Recognition System for SIGHAN Bakeoff 2006. In Proceeding of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 209-212. Sydney.

ณัฐดาพร เลิศชีวะ. (2553). การรู้จำชื่อเฉพาะภาษาไทย: การศึกษาชื่อผลิตภัณฑ์ในข่าวเศรษฐกิจ. วิทยานิพนธ์มหาบัณฑิต สาขาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

นวรรรณ พันธุ์เมธา. (2549). ไวยากรณ์ไทย. พิมพ์ครั้งที่ 3. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

น้องนุช มณีอินทร์. (2543). การปรับเปลี่ยนชื่อหมู่บ้านในจังหวัดเชียงใหม่. วิทยานิพนธ์มหาบัณฑิต สาขาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

นัชชา ถิระสาโรช (2553). การรู้จำชื่อเฉพาะภาษาไทย: การใช้แบบจำลองคอนดิชันนอลแรนดอมฟิลด์ส์. วิทยานิพนธ์มหาบัณฑิต สาขาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปราณี กุลละวณิชย์. (2535). ชื่อหมู่บ้านในมณฑลทลกวาสี และภาคตะวันออกเฉียงเหนือของประเทศไทย. กรุงเทพฯ: โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย

ปานทิพย์ มหาไตรภพ. (2522). นามสกุลพระราชทานในพระบาทสมเด็จพระมงกุฎเกล้าเจ้าอยู่หัว: การวิเคราะห์ทางอรรถศาสตร์ชาติพันธุ์. วิทยานิพนธ์มหาบัณฑิต สาขาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

พระยาอุปกิตศิลปสาร. (2546). หลักภาษาไทย. พิมพ์ครั้งที่ 12. กรุงเทพมหานคร: ไทยวัฒนาพานิช.

ศศิวิมล กาลันสีมา. (2553). การรู้จำชื่อเฉพาะภาษาไทย: การศึกษาชื่อบุคคล สถานที่ และองค์กร. วิทยานิพนธ์มหาบัณฑิต สาขาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

การประมวลผลภาษาไทย

สุจิตลักษณ์ ดีผดุง และ วชิราภรณ์ วรรณดี. (2544). ชื่อหมู่บ้านในภูมิภาคตะวันตกของประเทศไทย: การศึกษาเบื้องต้น. สถาบันวิจัยภาษาและวัฒนธรรมเพื่อพัฒนาชนบท มหาวิทยาลัยมหิดล

สุภาพรณ ฦ บางช้าง. (2526). การใช้ภาษาไทยในการตั้งชื่อคนไทย. ในงานวิจัยจุฬาลงกรณ์มหาวิทยาลัย ฉลอง 700 ปี ลายสือไทย ลำดับที่ 6. กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย

อมรทิพย์ กวินปนิธาน. (2546). การศึกษาบริบทบ่งบอกชื่อเฉพาะในภาษาไทยตามแนวภาษาศาสตร์คอมพิวเตอร์. วิทยานิพนธ์มหาบัณฑิต สาขาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ตอน ๒ : ต่อจากคำ

หน่วยภาษาที่ใหญ่กว่าคำคือวลี อนุพากย์ ประโยค การประมวลผลภาษาขั้นต่อไปคือการมองหาความสัมพันธ์ของคำและข้อความในตัวบท ซึ่งความรู้พื้นฐานที่จำเป็นต้องใช้คือเรื่องของหมวดคำ เพราะการประกอบกันของคำเป็นหน่วยใหญ่ขึ้นเป็นวลีหรือประโยคนั้นวิเคราะห์จากความสัมพันธ์ของหมวดคำของแต่ละคำในข้อมูลรับเข้านั้น ในตอนนี้จึงเริ่มต้นด้วยการกล่าวถึงหมวดคำและการกำกับหมวดคำ จากนั้นจึงกล่าวถึงการตัดข้อความเป็นหน่วยปริจเฉทพื้นฐานเพื่อใช้ประมวลผลต่อไป การวิเคราะห์โครงสร้างวากยสัมพันธ์ในที่นี่ได้จากการมองความสัมพันธ์แบบพึ่งพาระหว่างคำต่าง ๆ โปรแกรมแจ่งส่วนประโยคหรือ parser คือโปรแกรมที่ใช้เพื่อหาและสร้างโครงสร้างความสัมพันธ์นี้ ในที่นี่ได้ทดลองใช้ MaltParser ซึ่งสามารถนำข้อมูลภาษาไทยที่วิเคราะห์ความสัมพันธ์แบบพึ่งพาไว้แล้วมาเป็นตัวอย่างฝึกสอนโปรแกรม จากนั้นจึงนำข้อมูลที่เรียนรู้มาใช้วิเคราะห์แจ่งส่วนประโยคได้ นอกจากการประมวลผลภาษาแบบอิงการวิเคราะห์ทางภาษาศาสตร์ การแปลงข้อมูลภาษาเป็นชุดตัวเลขหรือเวกเตอร์ก็เป็นวิธีการพื้นฐานหนึ่งที่นิยมใช้กัน ในที่นี่ จึงกล่าวถึงหลักการแปลงข้อมูลภาษาเป็นเวกเตอร์เพื่อนำมาประมวลผลต่อไป พร้อมสาธิตการสร้าง Word2Vec ภาษาไทยและการใช้งาน

การกำกับหมวดคำภาษาไทย⁷

หมวดคำภาษาไทย⁸

หมวดคำเป็นข้อมูลพื้นฐานทางวากยสัมพันธ์ ทำให้เห็นว่าคำที่จัดอยู่ในหมวดเดียวกันจะมีคุณสมบัติทางวากยสัมพันธ์ไปทำนองเดียวกัน ในตำราหลักภาษาไทยของพระยาอุปกิตศิลปสารได้กำหนดหมวดคำไว้ 7 ประเภท ได้แก่ คำนาม คำสรรพนาม คำกริยา คำวิเศษณ์ คำบุพบท คำสันธาน และคำอุทาน ซึ่งเป็นการจำแนกด้วยเกณฑ์หลายเกณฑ์ เช่น จำแนกตามเกณฑ์ทางความหมาย โดยให้คำนามเป็นคำที่ใช้อ้างถึงคน สัตว์ สิ่งของ คำกริยาเป็นคำแสดงอาการ เป็นต้น จำแนกตามเกณฑ์หน้าที่ โดยให้คำวิเศษณ์เป็นคำใช้เพื่อประกอบคำอื่น จำแนกตามเกณฑ์ตำแหน่ง โดยให้คำบุพบทเป็นคำที่เกิดหน้าคำนาม เป็นต้น แต่การจัดหมวดคำแบบนี้มีที่มาจากไวยากรณ์ดั้งเดิมที่มีต้นแบบมาจากไวยากรณ์กรีกและไวยากรณ์ลาติน หมวดคำลักษณะนี้จึงอาจจะไม่ได้สะท้อนธรรมชาติที่แท้จริงของภาษาไทย นักภาษาศาสตร์จำนวนหนึ่งจึงพยายามจำแนกหมวดคำตามแนวคิดทฤษฎีไวยากรณ์สมัยใหม่ เช่น การจำแนกด้วยเกณฑ์

⁷ เนื้อหาบางส่วนในบทนี้นำมาจาก “คู่มือการกำกับข้อมูลภาษาไทยตามแนวทาง Universal Dependencies version 2” ที่ผู้เขียนเผยแพร่บนหน้าเว็บไซต์เขียน

⁸ ตำราไวยากรณ์ไทยจะใช้ “ชนิดของคำ” หมายถึงการแยกประเภทคำทางวากยสัมพันธ์ แต่ในที่นี้จะใช้ “หมวดคำ” เพราะนอกจากจะกล่าวถึงการระบุหาชนิดของคำในข้อมูลด้วยเครื่องแล้ว ยังจะกล่าวถึงการกำกับชนิดของคำในข้อมูลโดยคนด้วย ในงานนี้จึงจะใช้การกำกับหมวดคำเป็นคำกลาง ๆ ทั้งการกำกับด้วยคนเพื่อใช้ฝึกเครื่อง หรือการให้เครื่องกำกับหมวดคำในข้อมูลให้ “หมวดคำ” จึงแทน “part of speech” และ “การกำกับหมวดคำ” แทน “part of speech tagging”

หน้าที่ (นวรธรรม พันธุ์เมธา 2527) การจำแนกตามหลักไวยากรณ์โครงสร้าง (วิจิรัตน์ ภาณุพงศ์ 2532) การจำแนกตามเกณฑ์วากยสัมพันธ์ตามหลักทฤษฎีไวยากรณ์ศัพท์ การก (อมรา ประสิทธิ์รัฐสินธุ์ 2543) บ้างก็ใช้เกณฑ์หลากหลายผสมกันโดยแจกแจง รายละเอียดมากขึ้นเพื่องานประมวลผลภาษาไทย เช่น หมวดคำในคลังข้อมูลออร์คิด (Virach Somlertlamvanich et al. 1999) การจำแนกหมวดคำที่ผ่านมาจึงมีความแตกต่างกันไปตามลักษณะเกณฑ์และแนวคิดที่ใช้

ควรเลือกใช้หมวดคำชุดไหน

เนื่องจากหมวดคำเป็นหน่วยพื้นฐานทางวากยสัมพันธ์ แม้การนิยามหมวดคำ บางครั้งจะดูเหมือนว่าเป็นเรื่องของความหมาย แต่โดยพฤติกรรมแล้ว คำในหมวดคำ เดียวกันก็มีลักษณะการเกิดคล้าย ๆ กัน เช่น คำที่เป็นกริยาก็กักปรากฏหลังคำนามที่ ทำหน้าที่เป็นประธาน บางกลุ่มก็อาจมีนามตามหลังที่ทำหน้าที่เป็นกรรมได้ เป็นต้น เกณฑ์ทางวากยสัมพันธ์หรือการปรากฏร่วมกับคำอื่น ๆ จึงควรเป็นหลักสำคัญในการ จำแนกหมวดคำ หมวดคำจึงเป็นลักษณะเฉพาะของแต่ละภาษา หมวดคำภาษาไทยจึง ควรจำแนกโดยอาศัยข้อมูลภาษาไทยเป็นสำคัญ ไม่สามารถนำชุดหมวดคำที่ใช้ในภาษา อื่นมาใช้ได้โดยตรงได้ทั้งหมด

อย่างไรก็ดี ชุดหมวดคำในภาษาหนึ่ง ๆ ก็อาจมีได้หลายชุด ทั้งนี้ผู้ที่กำหนดชุด หมวดคำหนึ่งมาใช้ก็จะเลือกตามแนวทางการวิเคราะห์ของตนเอง โปรแกรม POS Tagger ที่ใช้จึงให้ผลที่แตกต่างกันไปตามชุดของหมวดคำที่เลือกใช้ได้ และไม่ใช่เรื่อง แปลกที่ในภาษาหนึ่งอาจมีการกำกับหมวดคำมากกว่าหนึ่งแบบ เช่น คลังข้อมูล British National Corpus จะมีการกำกับ POS โดยใช้ CLAWS4 tagger เพื่อแท็กชุดหมวดคำ

ที่เรียกว่า C5 ซึ่งที่ 57 หมวด⁹ (Leech et al., 1994) ในข้อมูลจะเห็น attribute “c5” เพื่อบอกหมวดคำแต่ละคำ และยังมีหมวดที่ให้คนทั่วไปเข้าใจง่ายใน attribute “pos” นอกจาก CLAWS tagger ในภาษาอังกฤษยังมี POS tagger อื่นที่ใช้ชุดหมวดคำอื่น เช่น Penn Treebank tag มีแท็กจำนวน 48 หมวด¹⁰ (Marcus et al.,1993) หรือ universal dependency tag ซึ่งมีจำนวน 17 หมวด¹¹ (Nivre et al.,2020)

ชุดหมวดคำที่ใช้ในการประมวลผลภาษาไทยชุดแรก ๆ ที่รู้จักเป็นชุดหมวดคำ ใน Orchid corpus มี 55 หมวด (Somlertlamvanich et al., 1999) เป็นชุดหมวดคำที่มีจำนวนมากเพราะแนวคิดการประมวลผลเวลานั้นต้องการหมวดคำที่ละเอียดเพื่อจำแนกความต่างทางวากยสัมพันธ์ผ่านการประกอบกันของหมวดคำต่าง ๆ เป็นโครงสร้างที่ใหญ่ขึ้น อย่างไรก็ตามในการทำงานประมวลผลภาษาต่าง ๆ หากแต่ละภาษามีการกำหนดใช้ชุดหมวดคำในลักษณะต่าง ๆ ที่แตกต่างกันโดยสิ้นเชิง จะทำให้การทำงานร่วมกันหลาย ๆ ภาษาเป็นไปได้ด้วยความลำบากมากขึ้น จึงเกิดแนวคิดที่จะกำหนดชุดหมวดคำที่สามารถใช้ได้กับทุก ๆ ภาษา จะมีเฉพาะบางส่วนของที่ยอมให้เป็นเรื่องเฉพาะภาษาได้ แต่กรอบการวิเคราะห์พื้นฐานแล้วจะใช้เหมือนกัน กลุ่มคนที่มีแนวคิดลักษณะนี้ จึงได้เสนอโครงการ universal dependencies (UD) ซึ่งเป็นโครงการสากลสำหรับกำกับหมวดคำและกำกับความสัมพันธ์ระหว่างคำผ่านการวิเคราะห์แบบพึ่งพา โดยได้กำหนด POS และกำหนดความสัมพันธ์แบบพึ่งพา (dependency relation) ที่เกิดขึ้นระหว่างคำในประโยคที่เป็นสากลขึ้นมา POS สำหรับใช้แท็กข้อมูลภาษาไทยในที่นี้ จึงยึดตามรายการที่นำเสนอใน

⁹ <http://www.natcorp.ox.ac.uk/docs/URG/posguide.html>

¹⁰ <https://aclanthology.org/J93-2004.pdf>

¹¹ <https://universaldependencies.org/u/pos/>

<http://universaldependencies.org/> ข้อควรสังเกตคือ จำนวน POS ที่เสนอให้ใช้นั้นมีไม่มาก และไม่แนะนำให้ตั้งแท็กที่เป็นเรื่องเฉพาะแต่ละภาษา จำนวนแท็กจึงมีน้อยกว่าที่ใช้กันในงาน POS Tagging ทั่วไป อย่างไรก็ตาม จำนวน POS ที่น้อยนั้นถูกชดเชยด้วยการกำหนดความสัมพันธ์แบบฟังก์ชันที่ค่อนข้างมากแทน ทำให้ยังคงสามารถแยกความต่างทางวากยสัมพันธ์ออกได้ เช่น คำนามที่เคยแยกเป็นประเภทย่อยต่าง ๆ ในชุดหมวดคำอื่น ในกรอบของ UD จะมีเพียง NOUN, PROPN คำลักษณะนามที่พบในภาษาไทย จีน จะถูกจัดเป็น NOUN ซึ่งต้องไปแยกด้วยวิธีการอื่นต่อไป ในภาษาจีนมีผู้ใช้ nmod:clf เป็นตัวกำหนดความสัมพันธ์นี้แทน (Leung et al., 2016) ซึ่งต่อมาใน UD version 2.0 กำหนดให้ใช้ relation “clf” เป็นการบอกความสัมพันธ์ระหว่าง head noun กับ classifier ในการกำกับข้อมูลภาษาไทยในที่นี่ จึงจะใช้ NounType:class ที่มีเสนอในบางภาษามาใช้ (ให้ class เป็น value ของ attribute “NounType” ภายใต้อัฒก POS: NOUN) และใช้ relation “clf” ประกอบด้วยเมื่อเชื่อมโยงกับคำนามหลัก เป็นต้น

จำนวนแท็กและรายการ POS ต่าง ๆ ในโครงการ UD พัฒนาต่อมาจากหลายโครงการ ได้แก่ Stanford dependencies (de Marneffe et al., 2006, 2008, 2014), Google universal part-of-speech tags (Petrov et al., 2012), และ Interset interlingua for morphosyntactic tagsets (Zeman, 2008) เพื่อให้ได้ชุดรายการที่เป็นมาตรฐานเดียวกันสำหรับใช้ได้กับภาษาต่าง ๆ หลักการนอกจากจะให้ใช้ได้กับหลากหลายภาษา ก็ยังต้องมีความถูกต้องทางภาษาศาสตร์ และไม่ยากเกินไปที่จะให้คนทั่วไปสามารถเข้าใจและใช้งานได้ UD รุ่นแรกออกมาในเดือนตุลาคม ปี 2014 UD ปัจจุบัน เป็น UD รุ่นที่สอง

สำหรับภาษาไทย จำนวน POS tag และ attribute เบื้องต้นที่จะใช้กับภาษาไทย ทางโครงการคลังข้อมูลภาษาไทยแห่งชาติ ได้กำหนดวิธีการกำกับข้อมูลดังนี้

การกำหนดหมวดคำภาษาไทยใน UD

การกำหนดหมวดคำ นอกจาก POS แล้ว เรายังสามารถระบุ attribute : value เพื่อใช้จำแนกประเภทย่อยของ POS นั้นอีกได้ UD POS tags และ attribute (หรือที่นักภาษาศาสตร์นิยมเรียก feature) ต่าง ๆ ที่ใช้ในที่นี่ ใช้โดยเทียบเคียงจากคำอธิบายที่เขียนไว้ในเว็บของโครงการ¹² แต่ก็มีคำบางกลุ่มที่อาจต้องปรับเปลี่ยนอีกเพื่อความเหมาะสม

¹² <http://universaldependencies.org/>

กลุ่มคำเปิด

NOUN	คำนามทั่วไปรวมถึงคำนำหน้าชื่อทั้งหลาย นาย นาง พล.ต. เหล่านี้ก็ให้กำหนด เป็น NOUN ด้วย แล้วยังความสัมพันธ์ compound ระหว่าง คำนำหน้าชื่อ กับ ชื่อ (นาย)<-(สมชาย) คำในกลุ่มชื่อสัตว์ (นก/พิราบ นก/กระยาง ปลาทู ปลา สวาย) หรือชื่อพืช (ผัก/บุ้ง ต้น/ตำลึง ต้น/มะม่วง) เหล่านี้ให้กำกับเป็น NOUN
NounType: class	Feature สำหรับใส่ให้คำที่เป็นลักษณนาม กำหนดไว่ก่อน แต่อาจจะลบทิ้งได้ เพราะสามารถใช้ relation clf บอกความเป็น classifier ได้
PROPN	คำนามเฉพาะ ได้แก่ ชื่อคน สถานที่ องค์กร ผลิตภัณฑ์ ประเทศ ส่วนชื่ออาหาร ชื่อปลา สัตว์ ต้นไม้ จะให้เป็น NOUN
NameType: Prs	Prs: person name (ใช้กรณีไม่รู้ว่า เป็น given หรือ surname)
NameType: Giv	Giv: given name ชื่อต้น เช่น ประยุทธ์ อัครวิน เป็นต้น
NameType: Sur	Sur: surname ชื่อสกุล เช่น จันทร์โอชา ขวัญเมือง เป็นต้น

การประมวลผลภาษาไทย

NameType: Geo	Geo: geographical name เช่น ชื่อสถานที่ ชื่อประเทศ ชื่อจังหวัด ชื่ออำเภอ ชื่อมหาวิทยาลัย (ในกรณีที่ปริบทนั้นต้องการบ่งบอกสถานที่ตั้ง เช่น ฉันขับรถผ่านจุกๆ เป็นต้น)
NameType: Com	Com: company organization name ชื่อบริษัท ชื่อประเทศ ชื่อมหาวิทยาลัย (ในกรณีที่ปริบทนั้นบ่งถึงลักษณะขององค์กร เช่น จุกๆ ส่งนิสิตเข้าแข่งขันในรายการต่าง ๆ เป็นต้น)
NameType: Nat	Nat: nationality คำแสดงเชื้อชาติ สัญชาติ
NameType: Pro	Pro: product เช่น ไอโฟน
NameType: Oth	Oth: others ชื่อเฉพาะอื่น ๆ ที่ไม่สามารถจัดเข้าประเภทต่าง ๆ ข้างต้นได้

VERB

คำกริยาทั้งกรรมและกรรม

VerbType: Cop	Featureที่กำหนดให้กับ กริยา copular ได้แก่คำว่า 'คือ' (ไม่ใช่ทุกกรณีที่เป็น copular) จะกำกับให้ VERB นั้นเป็น Cop ก็ต่อเมื่อฝั่งซ้ายและฝั่งขวาสื่อถึงสิ่งเดียวกันเท่านั้น เช่น เขาคือเพื่อนแท้ → เขา กับ เพื่อนแท้ เป็นสิ่งเดียวกัน เช่นนี้จะกำกับ คือ/VERB:{VerbType: Cop} ข้อสังเกตคือฝั่งซ้ายและขวามักจะมีสถานะเป็น NOUN PRON หรือ PROPN คือ ที่ทำหน้าที่เชื่อมความในปริจเฉท
---------------	---

	จะวิเคราะห์เป็น SCONJ เช่น การสื่อสารที่มีประสิทธิภาพที่สุดคือ การที่ผู้ส่งสารและผู้รับสารมีลักษณะเหมือนกัน
INTJ	คำอุทานทั้งหลาย
ADJ	คำคุณศัพท์ เป็นคำ ขยายหลังคำนาม รวมถึงคำที่บ้างจัดให้เป็นคำกริยาคุณศัพท์ จะต้องเป็นคำที่แสดงคุณลักษณะบางอย่างของคำนามที่อยู่ข้างหน้า เช่น ดี สวย อ้วน ผอม ใหญ่ เล็ก เป็นต้น ในกรณี คำซึ่งปรากฏหลังคำนามนั้นไม่ได้แสดงลักษณะของคำนาม แต่บ่งความชี้เฉพาะหรือไม่ชี้เฉพาะ จะกำกับเป็น DET
ADV	คำวิเศษณ์ ขยายกริยา VERB หรือขยายคำ ADJ หรือ ADV ปกติปรากฏหลังคำกริยา แต่มีบ้างที่ปรากฏหน้ากริยา เช่น คำว่า ‘สุด’ ใน สุดขยัน มีคำ ADV จำพวกหนึ่งที่ไม่ได้ขยายกริยา แต่ขยาย numeral, determiner, noun, pronoun ปรากฏหน้านาม เช่น เพียง อีก ทั้ง สัก ทุก แต่ละ หลากหลาย (หลากหลายรูปแบบ) เป็นต้น ให้กำกับเป็น ADV ลักษณะนี้พบในตัวอย่างการวิเคราะห์ภาษาอื่นด้วย เช่น ภาษาจีน

กลุ่มคำปิด

AUX		คำกริยาช่วยทั้งที่ปรากฏหน้าและหลังกริยาในภาษาไทย เช่น <i>ควร จะ น่า ได้ ไป มา</i> เป็นคำที่บ่งบอก tense, aspect, modality ในกรณีที่ปรากฏหน้ากริยาจะกำกับเป็น AUX แต่หากปรากฏหลังกริยาจะกำกับเป็น ADV ได้
ADP	AdpType: Prep	Adposition ในภาษาไทยเป็น preposition อย่างเดียว ไม่มี postposition ปรากฏหน้าคำนาม เช่น <i>บน ใน หน้า ทาง เพื่อ</i> feature นี้จึงเป็น default สำหรับ ADP ในภาษาไทย นำหน้า นามวลี นาม สรรพนาม หรืออนุพยางค์ทำหน้าที่นาม
PRON	PronType: Prs	คำสรรพนามบุคคล เช่น <i>ฉัน เขา เธอ อะไร</i> (เขาทำอะไร เธอก็ชอบทั้งนั้น) <i>นี่</i> (นี่คือสิ่งที่เขาต้องการ) <i>ใครต่อใคร ล้วนแต่ ต่าง</i> (ผู้คนต่างยกย่องเขา)
	PronType: Rcp	คำสรรพนาม reciprocal → <i>กัน ซึ่งกันและกัน</i>
NUM		คำบอกจำนวน ปริมาณ ทำหน้าที่เหมือน DET, ADJ, PRON บอกความหมายเกี่ยวกับจำนวน เป็น ตัวเลข คำจำนวนนับ เช่น <i>123 สิบ/เอ็ด</i> ตัวเลขที่ประกอบคำ เช่น <i>7</i> ใน <i>Windows 7</i> ก็ให้กำกับ NUM

DET	คำบ่งชี้เฉพาะหรือไม่ชี้เฉพาะ ใช้ขยายหลังคำนาม เช่น <i>นี่ นั่น ไหน อื่น ทั้งหมด</i> ต่างๆ ทั่วไป (กรณีที่มีการละคำนามและเหลือเพียงคำประเภทนี้คำเดียว เช่น <i>'ทั้งหมด'</i> ก็คงให้เป็น DET)
CCONJ	Coordinate conjunction คำเชื่อมสองส่วนที่มีสถานะเท่ากัน เช่น <i>และ (หมา และแมว) แต่ (ฉันอยากกินแต่เธออยากนอน) หรือ (ข้าว ก๋วยเตี๋ยว หรือสแต็ก)</i> เป็นต้น รวมถึงคำที่ทำหน้าที่เป็น discourse marker ที่เชื่อมความ เช่น <i>อย่างไรก็ตาม แต่ถึงกระนั้น</i>
SCONJ	Subordinate conjunction คำเชื่อมส่วนที่ทำหน้าที่เป็นส่วนขยายหรือ subordinate clause เช่น <i>ที่ว่า ให้ จึง เพื่อ</i> รวมถึงคำว่า <i>เช่น คือ</i> ที่ทำหน้าที่เชื่อมความในปริจเฉท
PART	คำอนุภาคทั่วไป กรณีคำอนุภาคเฉพาะให้เติม feature ข้างล่างนี้
PartType: Emp	emphasis ใช้กำกับหลังวลีหรืออนุประโยค เพื่อเป็นการเน้นเหตุการณ์ดังกล่าวได้อีกด้วย เช่น <i>'นั่น'</i> ใน <i>ขณะที่กำลังผสมส่วนต่างๆเข้าด้วยกันนั่น 'เสียอีก' 'ซะอีก'</i> เป็นต้น.
PartType: Res	Response <i>ครับ ค่ะ นะ จ้ะ</i>

การประมวลผลภาษาไทย

PartType: Int	Interrogative หรือไม่ หรือ ไหม ทำไม
PartType: Neg	Negation ไม่ มิ บั ใช่ว่า ไม่ใช่ว่า
PartType: Adj	Adjective นำ เป็นคำที่เป็น Adjective marker เมื่อปรากฏหน้าคำใด ทำให้คำนั้นทำหน้าที่เป็น adjective ได้ เช่น น่ารัก น่าเอ็นดู
PartType: Adv	Adverb อย่าง โดย เป็นคำที่เป็น adverb marker เมื่อปรากฏร่วมกับคำที่ตามมา จะทำให้คำนั้นทำหน้าที่เป็น Adv ได้ เช่น อย่างเร็ว โดยด่วน ด้วยดี (Particle สองประเภทหลังกำหนดใช้ในภาษาไทยเป็นการเฉพาะ)
PUNCT	เครื่องหมายวรรคตอนต่าง ๆ ที่เป็นส่วนหนึ่งของข้อความมีหน้าที่ในทางภาษา เช่น ๆ ๆ ฯลฯ , ?
SYM	เครื่องหมายสัญลักษณ์ต่าง ๆ ที่ไม่ใช่ PUNCT เช่น สัญลักษณ์ทางคณิตศาสตร์ + - * # @ รวมถึง email, url dummy@gmail.com www.dummy.com เป็นต้น
X	คำที่ยังไม่สามารถจัดประเภทได้ตามข้างบนนี้ ให้กำกับเป็น X ไปพลางก่อน
Feature	เป็น feature ที่ใช้กำกับคำใด ๆ เป็นการกำกับลักษณะเพิ่มเติมของคำนั้น ๆ ซึ่งถูกกำกับด้วย POS บางอย่างไว้แล้ว จากนั้นจึงพิจารณาต่อไปว่าคำดังกล่าวมี

ลักษณะใด ๆ เพิ่มเติมอีกหรือไม่ ลักษณะดังกล่าวนี้ ได้แก่ คำนำหน้า (Prefix)
คำย่อ (Abbr) และคำต่างประเทศ (Foreign)

Prefix: Yes

Prefix ประกอบคำ เช่น ชาว นัก ผู้ การ ความ นา อย่าง

Abbr: Yes

คำย่อ (POS ของคำย่อเป็นไปตามรูปเต็มคำนั้น เช่น กรุงเทพ/PROP {
NameType:geo, Abbr:Yes})/ฯ/PUNCT รร./NOUN {Abbr:Yes})

Foreign: Yes

คำทับศัพท์ต่างประเทศ

ข้อสังเกตเพิ่มเติมในการกำกับ POS:

การกำกับคำร่วมกับเครื่องหมายไม้ยมก (๓)

คำบางคำปรากฏร่วมกับเครื่องหมายไม้ยมก (๓) โดยปกติจะกำกับแยกเครื่องหมาย ๆ ออกเป็น PUNCT อย่างไรก็ตาม ในบางกรณีที่คำดังกล่าวจำเป็นต้องปรากฏซ้ำ หรือต้องปรากฏร่วมกับ “ๆ” เท่านั้น จึงจะสามารถสื่อความหมายที่ถูกต้อง เช่น

บนโต๊ะเต็มไปด้วยเอกสารต่าง ๆ

หลัง ๆ เขาก็ค่อยมาตามนัด

จะเห็นว่า ต่าง ๆ หรือ หลัง ๆ จำเป็นต้องปรากฏด้วยกัน หากปรากฏเพียงคำเดียวโดด ๆ จะสื่อความหมายที่ต่างออกไป ดังนั้น การกำกับคำว่า ต่าง ๆ กับ หลัง ๆ จำเป็นต้องกำกับรวมเป็น 1 คำ ต่างๆ/DET หลังๆ/CCONJ เป็นต้น

การกำกับนามวลีในกลุ่ม การ- ความ-

นามวลีในกลุ่ม การ- ความ- โดยปกติจะกำกับแยก การ- และ ความ- ออกมาเป็นหน่วย NOUN:{Prefix: Yes} แต่บางกรณีจำเป็นต้องรวม การ- หรือ ความ- ไว้เป็นส่วนหนึ่งของคำนั้นด้วย เช่น ความทรงจำ การบ้าน การเมือง การงาน เป็นต้น ขอให้พิจารณาจากความหมายของคำนั้น ๆ เป็นหลัก การวิเคราะห์ความสัมพันธ์ระหว่าง การ และ หน่วยกริยาวลีที่ตามมา สามารถวิเคราะห์ได้สองแบบ

แบบแรก ให้ การ เป็นคำหลักโดยมี POS เป็น NOUN ที่มี feature {prefix} และมีคำกริยาเป็นคำพึ่งพา เช่น การสูญเสียบุคคลใกล้ชิด การ >สูญเสีย เพราะมองในเชิงวากยสัมพันธ์ว่า ทั้งหมดทำหน้าที่เป็น นามวลี ไปเชื่อมโยงกับคำอื่นต่อ การ จึงควรเป็นคำหลักและมี POS: NOUN และหากเติม บุพบทวลีแสดงความเป็นเจ้าของ เช่น การสูญเสียบุคคลใกล้ชิดของเขา ของเขา จะสัมพันธ์กับ การสูญเสีย เหมือน

อย่าง *หนังสือ* ของเขา หากวิเคราะห์แบบนี้ ความสัมพันธ์ระหว่าง *การ* และ *กริยาวลี* ข้างหลังควรเป็น *acl* คือ *clausal modifier of noun*

แบบที่สอง วิเคราะห์ให้กริยาเป็นหลักเพราะมองว่าเป็นคำเนื้อหา *การ* มีความเป็นคำไวยากรณ์จึงควรเป็นหน่วยพึ่งพาตามหลักที่ให้คำไวยากรณ์ไม่สามารถเป็นหน่วยหลัก *การ* < *สูญเสีย* โดยมีความสัมพันธ์เป็น *mark* และ *การสูญเสียบุคคลใกล้ชิดของเขา* ก็จะโยงความสัมพันธ์ระหว่างคำหลัก *สูญเสีย* กับ <(ของ< *เขา*) ในเชิงความหมาย การวิเคราะห์แบบนี้ดูเหมือนจะทำให้คำเนื้อหาเชื่อมโยงกันได้ดีกว่า [*สูญเสีย* – *บุคคล*, *สูญเสีย* – *เขา*] แต่วิธีนี้ทำให้มีปัญหายุ่งยากในการแยกว่า *เขา* มีบทบาททางความหมายอะไรผ่าน *relation* อะไร เพราะ (ของ*เขา*) ปกติเป็น *nmod* ของคำนามข้างหน้า กรณีนี้จะยังคงให้เป็น *nmod* ของ *สูญเสีย* ได้ไหม ถ้ายอมได้ ก็จะกลายเป็นมีกรณีที่ *nmod* ขยาย *VERB* ได้ด้วย และจะเป็นปัญหาต่อไปในกระบวนการ *parsing* หรือถ้าแก้โดยให้ *สูญเสีย* เป็น *NOUN* เพราะมี *การ* อยู่ข้างหน้า ก็จะทำให้ต้องมีอีกกระบวนการหนึ่งเปลี่ยน *POS* ของคำได้

ในที่นี้จึงเลือกการวิเคราะห์แบบแรกให้ *การ* เป็น *NOUN{prefix}* ที่มีกริยาวลีตามหลังด้วยความสัมพันธ์ *acl*

POS ของคำอาจขึ้นกับตำแหน่งในการปรากฏ

คำหนึ่ง ๆ อาจสามารถกำกับได้หลาย *POS* ขึ้นกับว่าคำนั้น ๆ ปรากฏในตำแหน่งใดของข้อความ ยกตัวอย่างคำว่า *หลากหลาย* จากข้อความต่อไปนี้

อาชีพหลากหลาย *หลากหลาย* กำกับเป็น *ADJ* เป็น *dependent* ของ *อาชีพ* (ปรากฏหลังนาม)

อาชีพ/NOUN *หลากหลาย*/ADJ

อาชีพที่หลากหลาย *หลากหลาย* กำกับเป็น *VERB* เป็น *head* ของ *ที่*



หลากหลายอาซีพ หลากหลาย กำกับเป็น ADV เป็น dependent ของ อาซีพ (ปรากฏหน้านาม)



UD เปลี่ยนวิธีการวิเคราะห์ ellipsis จากเดิมที่ใช้ relation remnant มาเป็นการโปรโมทหน่วยรองมาเป็น head แทน เช่น ถ้าละ head nominal จะโปรโมท dependents ตามลำดับความสัมพันธ์นี้มาแทน amod > nummod > det > nmod > case ถ้าละกริยา ให้โปรโมท aux หรือ cop มาแทน ถ้าไม่มีก็มีสองแนวทาง ทางแรกคือให้โยงตัว orphan กับ grandparent และใช้ composite relation คือมี relation ของ headrel ต่อด้วย orphanrel เช่น *I like tea and you coffee* จะได้ *like-(conj>nsubj)>you like-(conj>obj)>coffee* ทางที่สองคือใช้ relation “orphan” ในตัวอย่างนี้จะเป็น *you-(orphan)>coffee* ในที่นี้จะเลือกใช้วิธีที่สองคือใช้ relation orphan

การวิเคราะห์คำกริยาเรียง

ประโยคที่มีกริยามากกว่าหนึ่งคำมาประกอบกัน ในแต่ละกรณีจะวิเคราะห์ต่างกัน กรณีที่เป็น กริยาเรียง หรือ serial verb construction จะวิเคราะห์ทุกคำเป็น VERB และมีความสัมพันธ์แบบ compound เช่น

เขาหยิบหนังสือมาอ่าน หยิบ-compound>มา, มา-compound>อ่าน

กรณีกริยาการเคลื่อนที่วิเคราะห์ได้ว่าไม่ได้ใช้ในความหมายการเคลื่อนที่แต่กลายเป็นคำไวยกรณ์ ให้วิเคราะห์คำนั้นเป็น ADV และใช้ความสัมพันธ์ advmod ส่วนใหญ่เป็นประโยคที่มีคำเช่น ไป มา ขึ้น ลง เข้า ออก อยู่ท้ายประโยค เช่น

เขากลับมาแล้ว กลับ-advmod> มา, กลับ-aux>แล้ว

แต่ถ้าเป็นกรณีที่เป็นคำกริยาที่ยังมีความหมายการเคลื่อนที่ เช่น เขาเดินไปโรงเรียน ก็จะใช้วิเคราะห์เป็น VERB และความสัมพันธ์แบบ compound เหมือนกริยาเรียงปกติ

กรณีที่ดูคล้ายเป็นคำกริยาเรียง แต่สิ่งที่กริยา subcategorize เป็นอนุพากย์ไม่ใช่คำกริยา ให้วิเคราะห์เป็น VEBR แต่ความสัมพันธ์เป็น xcomp, ccomp, หรือ advcl แล้วแต่กรณี เช่น

เขาชอบกินอาหารนอกบ้าน สิ่งที่ชอบคือ การกินอาหารนอกบ้าน จึงวิเคราะห์เป็น ชอบ-xcomp>กิน ที่เป็น xcomp เพราะอนุพากย์นี้ไม่สามารถใส่ประธานได้ ประธานของ กิน ผูกกับประธานของ ชอบ เป็นคนเดียวกัน

เขาให้(แดง)ดูรูป คำว่า ให้ ในที่นี้ subcategorize อนุพากย์โดยที่อนุพากย์นี้อาจมีหรือไม่มีประธานและไม่ผูกกับประธานประโยคหลัก จึงวิเคราะห์เป็น ให้-ccomp>ดู

พายุพัดต้นไม้โค่น ไม้วิเคราะห์ พัดโค่น เป็น compound จะวิเคราะห์เหมือนประโยค พายุพัดต้นไม้โค่น จึงเป็น พัด-advcl>โค่น หรือถ้าเป็นประโยค พายุพัดต้นไม้จนโค่น ก็ยังวิเคราะห์ได้เหมือนเดิม คือ พัด-advcl>โค่น ส่วนประธานของ โค่น ไม่ได้ผูกกับกรรมของประโยคหลัก เป็น zero pronoun ที่ต้องหาจากบริบท เพราะสามารถเขียนประโยคเต็มเป็น พายุพัดต้นไม้จนต้นไม้โค่น

การกำกับความสัมพันธ์ระหว่างคำ

ความสัมพันธ์แบบพึ่งพา dependency relation จะกำกับระหว่าง head กับ dependent ลูกศรตั้งต้นมาจาก head ไปที่ dependent



โครงสร้างที่ได้จากความสัมพันธ์นี้จะออกมาเป็น dependency tree dependency จะโยงระหว่างคำเนื้อหาเป็นหลัก ในโครงสร้าง coordinate จึงโยงระหว่าง noun กับ noun หรือ verb กับ verb dependency tree ใน UD จึงต่างจาก dependency tree ใน approach อื่น เข้าใจว่า การโยงแบบ UD ทำให้เห็น relation ระหว่าง content word โดยตรง ซึ่งจะเอื้อต่อการวิเคราะห์ความหมายต่อไปได้ง่ายกว่า

ส่วนคำไวยากรณ์จะเป็น dependent เป็นหลัก ไม่มีคำที่มา depend ด้วย กลุ่มคำไวยากรณ์จึงไม่ต่อกันเป็นสาย ถ้ามีหลาย ๆ ตัว ก็จะไปขยาย head เดียวกัน

คำที่เป็น multiword เช่น *in spite of* จะโยงความสัมพันธ์ fixed ระหว่าง *in >spite, in >of* กรณีภาษาไทย ถ้าเป็นคำที่กำหนดให้เป็นคำแล้วก็ไม่ต้องทำแบบนี้ เช่น *อย่างไรก็ตาม* แต่หากโปรแกรมตัดคำไม่ได้ตัดเป็นคำเดียว เช่น *ถึง|กระนั้น|ก็|ตาม* ก็จะต้องกำกับความสัมพันธ์แบบ fixed นี้ *ถึง->กระนั้น ถึง->ก็ ถึง->ตาม* และใช้ relation “fixed”

การละ จะไม่มีการเติมคำที่หายไป จะวิเคราะห์ไปตามรูปรูปปรากฏ ถ้าคำที่หายไปเป็น dependent ก็ไม่ต้องทำอะไร ถ้าเป็นคำที่ควรเป็น head เช่น คำนาม ก็ให้ promote คำที่ควรเป็น dependent ของคำนามที่หายไปมาเป็น head แทน ตามลำดับดังนี้ *amod > nummod > det > nmod > case* เช่น *He buys himself a green car and she (buys) herself a >obj>(red)* สำหรับภาษาไทย ลำดับการ promote อาจไม่เป็นตามนี้ แต่หลักการคงเดิม คือไม่มีการเติมสิ่งที่หายไป

Dependency structure ที่ใช้ในที่นี่เป็น basic tree สำหรับการใช้งาน ลักษณะ shallow ไม่ใช่ enhanced dependencies ที่ให้มี null node มีการบอกถึง controlled/raised subject ที่ใช้ร่วมกัน แบบ enhanced จึงมีรายละเอียดมากขึ้น มี cross dependency เกิดขึ้นได้ด้วย

สถิติ

โปรแกรม POS tagging พัฒนาเป็นโมดูลใน TLTK สามารถทดลองใช้ได้

https://colab.research.google.com/drive/1YCrhKqO9_YEVXgs-_HYXv4n5N73LdzF2Y?usp=sharing

TLTK ใช้ PerceptronTagger ใน nltk.tag.perceptron เป็นต้นแบบในการพัฒนา ข้อมูลที่ใช้ฝึกสอนโปรแกรมเป็นข้อมูลจาก Thai National Corpus ที่สุ่มมาใส่หมวดคำเองด้วยมือโดยทีมนิสิตภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย มีจำนวนข้อมูลประมาณ 148,000 คำ ค่าความถูกต้องที่ประเมินได้เมื่อสร้างโมเดลอยู่ที่ 91.68% ทั้งนี้ตัว NLTK PerceptronTagger เป็นโมดูลที่พัฒนามาจาก Textblob Averaged Perceptron Tagger รายละเอียดวิธีการพัฒนาโมเดล POS Tagger สามารถดูเพิ่มเติมได้ที่ <https://explosion.ai/blog/part-of-speech-pos-tagger-in-python>

โมดูล `tltk.nlp.pos_tag` รับข้อมูลเป็นข้อความปกติ นำข้อความมาตัดคำและกำกับหมวดคำ แต่ละส่วนของข้อความจะเป็น list ของคู่ลำดับ (word,pos) สามารถเลือก Option="mm" หรือ maximum matching ได้ หากไม่เลือกจะตัดคำแบบใช้ collocation strength ของพยางค์

```
import tltk
```

```
text = "จากอาชีพที่ตัวเองเป็นอยู่ เน้นทำงานหน้าคอม ไม่ค่อยได้ออกกำลังกายมากนัก ไม่ได้ยืดเส้นยืดสาย  
กายบริหารไม่ได้จริงจัง เรียกว่ามันค่อย ๆ สะสมจนถึงจุดที่มีปัญหา"
```

```
out = tltk.nlp.pos_tag(text)
```

```
print(out)
```

```
[(['จาก', 'ADP'), ('อาชีพ', 'NOUN'), ('ที่', 'SCONJ'), ('ตัวเอง', 'PRON'), ('เป็น', 'VERB'), ('อยู่', 'ADV'), ('<s/>', 'PUNCT')], [(['เน้น', 'VERB'), ('ทำงาน', 'VERB'), ('หน้า', 'NOUN'), ('คอม', 'NOUN'), ('<s/>', 'PUNCT')], [(['ไม่ค่อย', 'AUX'), ('ได้', 'AUX'), ('ออกกำลังกาย', 'VERB'), ('มากมาย', 'ADV'), ('นัก', 'ADV'), ('<s/>', 'PUNCT')], [(['ไม่ได้', 'AUX'), ('ยืดเส้นยืดสาย', 'VERB'), ('<s/>', 'PUNCT')], [(['กายบริหาร', 'NOUN'), ('ไม่ได้', 'PART'), ('จริงจัง', 'ADV'),
```

การประมวลผลภาษาไทย

('<s/>', 'PUNCT']], [(เรียกว่า, 'VERB'), (มัน, 'PRON'), (ค่อยๆ, 'ADV'), ('<s/>', 'PUNCT']], [(สะสม, 'VERB'), (จน, 'SCONJ'), (ถึง, 'VERB'), (จุด, 'NOUN'), (ที่, 'SCONJ'), (มี, 'VERB'), (ปัญหา, 'NOUN'), ('<s/>', 'PUNCT']])

ถ้าข้อมูลเข้าเป็นข้อมูลที่ตัดคำมาแล้วซึ่งอาจตัดคำด้วยโปรแกรมอื่นมาก่อน ก็สามารถใช้โมดูล `ltk.nlp.pos_tag_wordlist(...)` ที่รับข้อมูลเข้าเป็น list ของคำได้

อ้างอิง

Leech, G., Garside, R., and Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. In Proceedings of the 15th International Conference on Computational Linguistics (COLING 94) Kyoto, Japan, pp622-628.

Leung, H., Poiret, R., Wong, T.-s., Chen, X., Gerdes, K., & Lee, J. (2016, December). Developing Universal Dependencies for Mandarin Chinese. Proceedings of the 12th Workshop on Asian Language Resources (ALR12) Osaka, Japan.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, 19(2), 313–330. Retrieved from <https://aclanthology.org/J93-2004>

de Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006): 449-454.

- de Marneffe, M.-C. and Manning, C. D. (2008). The Stanford typed dependencies representation. In Workshop on Cross-framework and Cross-domain Parser Evaluation
- de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014): 4584-4592.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. Proceedings of the Twelfth Language Resources and Evaluation Conference Marseille, France.
- Petrov, S., Das, D., & McDonald, R. (2012). A Universal Part-of-Speech Tagset. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12) Istanbul, Turkey.
- Sornlertlamvanich, V., Takahashi, N., & Isahara, H. (1999). Building a Thai part-of-speech tagged corpus (ORCHID). The Journal of the Acoustical Society of Japan (E), Vol.20, No.3, pp 189-198, May 1999.
- Zeman, Daniel. (2008). Reusable Tagset Conversion Using Tagset Drivers. Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco

การตัดประโยคภาษาไทย

ประโยคคืออะไร

นอกจากการตัดคำแล้ว ปัญหาพื้นฐานอีกอย่างในการประมวลผลภาษาไทย คือ การตัดประโยค และดูจะเป็นปัญหาที่ยาก เพราะภาษาไทยไม่ได้มีการใช้เครื่องหมายวรรคตอนบอกขอบเขตของประโยคเหมือนอย่างภาษาอังกฤษ ในภาษาอังกฤษเอง การตัดประโยคหรือ sentence tokenisation ก็เป็นงานหนึ่งในการประมวลผลภาษา ที่จะต้องมีการตัดสินใจว่า เครื่องหมายวรรคตอนต่าง ๆ เช่น . ที่เห็นนั้นบอกการสิ้นสุดประโยคหรือบอกคำย่อ การทำ sentence tokenisation ในภาษาอังกฤษ อาจทำแบบง่าย ๆ โดยเขียนเป็นกฎใช้ regular expression เช่น `/(?!\\w\\.\\w\\.)(?![A-Z][a-z]\\.)(?<=\\.\\(\\?)\\s/` ซึ่งก็ให้ผลถูกต้องได้ระดับหนึ่ง (Dib, 2023) หรือจะใช้วิธีเตรียมข้อมูลแล้วให้เครื่องเรียนรู้จากข้อมูลก็ได้

สำหรับภาษาไทย เราอาจมองเบื้องต้นว่าปัญหาการตัดประโยคเป็นปัญหาการแก้ความกำกวมว่า space ที่เป็นเป็น sentence break ได้หรือไม่ได้ คล้ายกับที่ภาษาอังกฤษต้องพิจารณาว่า . เป็นตัวบอกจบประโยคหรือทำหน้าที่อื่นเช่นบอกคำย่อ แต่เอาเข้าจริงแล้ว ภาษาไทยนั้นไม่ได้มีกำหนดหลักการเขียนประโยคให้ชัดเจนเหมือนอย่างในภาษาอังกฤษ เราจึงควรเริ่มจากทำความเข้าใจก่อนว่า หลักการหาขอบเขตประโยคในภาษาอังกฤษเป็นอย่างไร

ประโยคในภาษาอังกฤษ

ในภาษาเขียนของภาษาอังกฤษ ประโยคถูกกำกับด้วยการใช้เครื่องหมาย . ? ! หรือบางครั้ง ; ; โดยทั่วไป ทุกประโยคจะมีกริยาหลักที่เป็น finite verb คือเป็นกริยาที่สามารถแปรรูปตามประธานหรือกาลได้ เรียกว่าเป็นกริยาหลักหรือ main verb

Simple sentence

ประโยคที่มีคำกริยาคำเดียวเป็นกริยาหลัก เรียกว่า simple sentence เช่น *The man runs*. ประโยค simple sentence อาจมีความซับซ้อนที่องค์ประกอบภายในได้ เช่น มีประธานมากกว่าหนึ่ง *John and Mary went to the theater*. มีกริยาหลักมากกว่าหนึ่ง เช่น *They kicked and beat me*. หรือรูปกริยาที่ไม่ใช่กริยาหลักจะแปลงรูปเป็น infinite verb อาจเป็นรูป to infinitive เช่น *The man runs to the store to get a bottle of beer*. หรืออยู่ในรูป participle เช่น *Waiting for John, she made some coffee*. ประโยค simple sentence จึงมีอนุพจน์หลักหรือ main clause เดียว แต่ถ้าดูความหมายอาจมีหลายเนื้อความ

The man runs to the store to get a bottle of beer. =>

{the man runs to the store} + {the man get a bottle of beer}}

Complex sentence

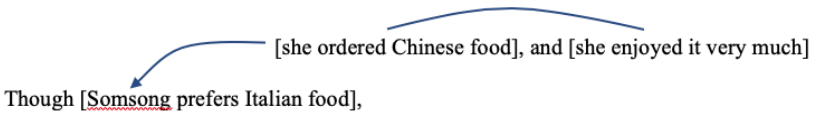
อนุพจน์หนึ่งอาจถูกนำมาประกอบเป็นส่วนหนึ่งของประโยคได้ เช่น *I learned that he did everything real quick*. อนุพจน์ "that he did everything real quick" ทำหน้าที่เป็นกรรมของกริยา "learn" เราเรียกประโยคลักษณะนี้ว่า complex sentence คือมีอนุพจน์หนึ่งที่เกิดเองโดยลำพังไม่ได้หรือเรียกว่า dependent clause ในประโยคหนึ่ง ๆ dependent clause ไม่จำเป็นต้องเป็นหน่วยหลักหรือประธานหรือกรรม อาจเป็นหน่วยขยายคือเป็น adverbial clause เช่น ตัวอย่างนี้ *He failed the exam because he had not studied enough*. ที่อนุพจน์หลังนำหน้าด้วย subordinate conjunction 'because' เพื่อบอกเหตุผลขยายใจความของอนุพจน์หลักที่อยู่ข้างหน้า อนุพจน์แรกเป็น main clause เป็นความหมายหลักของประโยค แต่อนุพจน์หลังเป็น dependent clause หากตัด main clause ออกเหลือเพียง "because he had not studied enough" ก็จะไม่สื่อความที่ต้องการ

Compound sentence

กรณีที่มีอนุพากย์มากกว่าหนึ่งอนุพากย์และสามารถเกิดเองโดยลำพังได้ เราจะเรียกว่า compound sentence มักจะเป็นประโยคที่เชื่อมสองอนุพากย์ด้วยคำสันธาน เช่น *They start these rumors and people start believing it.* สองอนุพากย์สื่อใจความสำคัญพอกันจัดเป็น independent clause ทั้งคู่

Compound-complex sentence

จะเห็นว่าประโยคในภาษาอังกฤษสามารถมีเนื้อความได้มากกว่าหนึ่งเนื้อความมาประกอบกันตามหลักของภาษา ประโยคที่เขียนอาจมีทั้งอนุพากย์ไม่อิสระและอนุพากย์อิสระหลายส่วนผสมกันเป็นประโยคที่เรียกว่า compound-complex sentence เช่น *Though Somsong prefers Italian food, she ordered Chinese food, and she enjoyed it very much.*



ซึ่งเอาเข้าจริง ประโยคในภาษาอังกฤษสามารถเขียนแบบซับซ้อนมาก ๆ ได้ ดังตัวอย่างประโยคเปิดในหนังสือ "Water Dance" ของ Ta-Nehisi Coates

ข้อความทั้งหมดเป็นประโยคหนึ่งประโยคที่ประกอบด้วยเนื้อความจำนวนมากเชื่อมโยงกันตามหลักภาษาอังกฤษ คือ มีการใช้คำสันธาน เครื่องหมายวรรคตอน เพื่อบ่งชี้ความสัมพันธ์ของเนื้อความแต่ละส่วน (ในตัวอย่างนี้ คำสันธานบางคำอาจเชื่อมแค่ระดับคำหรือวลี)

AND I COULD ONLY have seen her there on the stone bridge, a dancer wreathed in ghostly blue, because that was the way they would have taken her back when I was young, back when the Virginia earth was still red as brick and red with life, and though there were other bridges spanning the river Goose, they would have bound her and brought her across this one, because this was the bridge that fed into the turnpike that twisted its way through the green hills and down the valley before bending in one direction, and that direction was south. (from “Water Dancer”, p.1)

ในงานเขียนทั่วไป เรามักพบประโยคทั้งที่ไม่ซับซ้อนและซับซ้อนมากปนกันไปตามแต่วิธีการเขียนของแต่ละคน ดังสามตัวอย่างล่าง ย่อหน้าแรกมีประโยคห้าประโยค ย่อหน้าสองมีประโยคสามประโยค และย่อหน้าที่สามมีสองประโยค (เครื่องหมาย // สี่แดงถูกเติมไปเพื่อแสดงให้เห็นของเขตประโยค ซึ่งในตัวอย่างจะเห็นว่าอยู่หลังเครื่องหมาย .)

An AI researcher at the University of Toronto, he wanted to build a deep-learning model that would predict a patient’s health over time. ¶ But data from medical records is kind of messy: throughout your life, you might visit the doctor at different times for different reasons, generating a smattering of measurements at arbitrary intervals. ¶ A traditional neural network struggles to handle this. ¶ Its design requires it to learn from data with clear stages of observation. ¶ Thus it is a poor tool for modeling continuous processes, especially ones that are measured irregularly over time.

(<https://www.technologyreview.com/s/612561/a-radical-new-neural-network-design-could-overcome-big-challenges-in-ai/>)

For NLP, the uncanny valley is the point at which a human listener realizes that they had started to anthropomorphize the AI, but then something went wrong with the AI’s response and betrayed its non-humanness. ¶ The listener may not fully identify what is wrong, but they perceive that the conversation went sideways and the AI is no longer responding appropriately. ¶ In a double-blind test, perhaps a human listener might think the AI is a crazy or unstable human.

(<https://www.forbes.com/sites/paulteich/2019/03/07/ibm-project-debater-is-in-the-uncanny-valley-and-why-that-is-ground-breaking/#308519287f30>)

It’s unfortunate, as authors of these papers often make insightful connections between bits of data and draw valuable conclusions about the issue being studied, so there’s a lot of knowledge embedded in these papers that isn’t easily interpreted by machines, and would be time-consuming for humans to digest as well. ¶ While there have been previous attempts to use machine learning methods such as natural language processing to retrieve information from scientific papers, the disadvantage of this approach is that it requires a lot of human supervision, in the sense that datasets have to be manually labeled for training the AI model.

(<https://thenewstack.io/ai-makes-new-scientific-discoveries-by-analyzing-old-research-papers/>)

การประมวลผลภาษาไทย

ในกรณีภาษาพูด ซึ่งไม่สามารถใช้เครื่องหมายวรรคตอนแบบภาษาเขียนในระหว่างการพูดได้ เราจะเห็นการเชื่อมโยงความต่อเนื่องกันไป เมื่อถอดคำพูดออกมาเป็นตัวเขียน จึงมักพบลักษณะประโยคแบบซับซ้อนมากเช่นกัน ดังตัวอย่างจากคลังข้อมูล MICASE

yeah um i'm gonna give you the short version we have a little thing that we normally read but um this class is being recorded and if anybody, has any objections or any questions about the project that i'm working on, feel free to ask me about it after class. does anybody object? okay good, thanks. (MICASE, LEL300SU020)

S1: oh i don't have a microphone here. my name is Deborah and i'm a, uh researcher with the Communication Studies Department, and um, Professor Brody has, let me come into your class to recruit you for an experiment we're doing. um, what i'm gonna do is i'm gonna pass around these, these sheets, and they're eligibility forms, so if you could if um, if you're interested in (getting paid,) we pay you for participating it only takes about an hour. um, i'm gonna go ahead and pass out the sheets, and if you could, pass 'em back over to Nikolas and then Nikolas will collect them for me and give them to me. if you're eligible, we'll call you, um, and we'll set up a time for you to come in and uh give you the (xx) so, it's it's a pretty simple process you come in and, we have you do a task and like i said it only takes an hour and we'd pay you for it. so, we'd appreciate you coming in, and, participating. any questions...? okay then i'm gonna pass these down and if you could just give them over to Nikolas. thank you. (MICASE, LEL220JU071)

จากตัวอย่างทั้งหมดที่ยกมา ทำให้เห็นว่า แม้ภาษาอังกฤษจะมีหลักภาษา กำหนดวิธีการเขียนอย่างชัดเจนว่าขอบเขตประโยคอยู่ที่ไหน และส่วนเนื้อความต่าง ๆ จะเชื่อมโยงกันได้อย่างไร ต้องใช้คำหรือเครื่องหมายอะไรเชื่อมความ และสามารถระบุได้ว่าอะไรเป็นอนุพจน์หลัก อะไรเป็นอนุพจน์ขยาย แต่ในการเขียนจริง วิธีการเขียนแต่ละคนก็ต่างกัน คนที่เขียนประโยคไม่ซับซ้อนมากก็จะทำให้คนอ่านเข้าใจได้ง่ายกว่า คนที่เขียนประโยคซับซ้อนมาก ๆ ก็อาจจะทำให้คนอ่านเข้าใจความได้ยากมากขึ้น แต่ก็ไม่มีข้อห้ามอะไรที่จะไม่ให้เขียนประโยคซับซ้อนเหล่านี้

ประโยคในภาษาไทย

ภาษาเขียนในภาษาไทยไม่ได้มีหลักในการใช้เครื่องหมายวรรคตอนหรือคำเชื่อม เพื่อบอกขอบเขตประโยคแบบภาษาอังกฤษ แม้พระบาทสมเด็จพระมงกุฎเกล้าเจ้าอยู่หัวจะทรงมีพระราชดำริเสนอให้นำวิธีการใช้เครื่องหมายวรรคตอนมาใช้กับ

ภาษาไทยแทนที่การเขียนแบบติดกันไปตลอด เพื่อให้อ่านและเข้าใจภาษาไทยได้ถูกต้องง่ายขึ้น และแม้จะมีหลายคนที่ยพยายามเขียนภาษาไทยด้วยการใช้เครื่องหมายวรรคตอนนี้ แต่วิธีการเขียนแบบนี้ก็ไม่น่าเป็นที่นิยมแพร่หลาย (บุญ, 2002) จึงเป็นเหตุให้การตัดประโยคภาษาไทยไม่สามารถทำได้ง่ายเหมือนกับในภาษาอังกฤษ. แม้ว่าเราอาจพยายามมองประโยคภาษาไทยตามแบบการเขียนในภาษาอังกฤษเป็น ประโยคความเดียว (simple sentence), ประโยคความซ้อน (complex sentence), ประโยคความรวม (compound sentence), และประโยคความซับซ้อน (compound-complex sentence) แต่การเขียนที่ไม่ได้มีหลักการใช้เครื่องหมายวรรคตอนเพื่อแบ่งประโยค ทำให้ข้อมูลภาษาไทยมักเขียนต่อเนื่องเชื่อมโยงกันไปจนยากที่จะแบ่งประโยคตามประเภทเหล่านี้ได้ง่ายคล้ายกรณีข้อมูลภาษาพูดภาษาอังกฤษที่ไม่มีเครื่องหมายวรรคตอนช่วยแบ่งประโยคให้

หากเราพยายามยึดหลักการเดียวกับภาษาอังกฤษ คือมองหาคำเชื่อมที่เป็นตัวบ่งชี้การเชื่อมโยงความภายในประโยค เช่น คำว่า "ที่" "แต่" "และ" "ซึ่ง" เป็นต้น (ตำแหน่งที่ทำสีเขียวไว้) และตำแหน่งที่ควรตัดเป็นประโยคตามแบบภาษาอังกฤษได้ (ตำแหน่งที่เติม//และทำสีแดงไว้) แม้ว่าคำเชื่อมเหล่านั้นส่วนใหญ่ยังคงบ่งชี้ว่าความนั้นยังต่อเนื่องอยู่ได้ แต่การจบประโยคอาจมีวรรคข้างหน้าหรือไม่มีวรรคก็ได้ (ดูตัวอย่างย่อหน้าจากไฟล์ NACNS021 ในคลังข้อมูลภาษาไทยแห่งชาติ) และเว้นวรรคสามารถใช้บอกขอบเขตคำ วลีหรืออนุพยางค์ด้วยก็ได้ (ในตัวอย่างนี้ บางคนอาจมองว่า "การใช้ประโยชน์จากทรัพยากร..." ไม่ได้ขึ้นประโยคใหม่ ยังคงสัมพันธ์กับประโยคก่อนหน้านี้ก็ได้ การตัดประโยคจึงเป็นสิ่งที่อาจเห็นต่างกันได้ แต่ถึงอย่างไร ก็ต้องมีการตัดอนุพยางค์ตรงนี้)

การประมวลผลภาษาไทย

ทะเลไทย มีเนื้อที่ 378,000 ตารางกิโลเมตร ครอบคลุมพื้นที่ 23 จังหวัด ตามแนวชายฝั่งที่ยาวถึง 2,815 กิโลเมตร ทั้งฝั่งอ่าวไทยและฝั่งอันดามัน ทะเลไทยแม้จะมีพื้นที่ไม่กว้างใหญ่นัก แต่ก็เคยเป็นที่ยอมรับจากนักวิทยาศาสตร์ทั่วโลกว่ามีความหลากหลายทางชีวภาพทางทะเลมากที่สุดแห่งหนึ่ง และมีความอุดมสมบูรณ์ของแหล่งทรัพยากรทางทะเลอย่างมาก การใช้ประโยชน์จากทรัพยากรก็ย่อมมีมากด้วยเช่นกัน ซึ่งก็แน่นอนที่สุดย่อมก่อให้เกิดความเสื่อมโทรมของทรัพยากรธรรมชาติ ดังที่ปรากฏอยู่ในหลายพื้นที่ชายฝั่งทะเล (NACNS021)

ตัวอย่างที่สองก็พบข้อมูลลักษณะเดียวกัน กรณีมีคำเชื่อมปรากฏ เราอาจบอกได้ว่าเนื้อความนั้นยังไม่จบประโยค แต่ก็มีคำเชื่อมบางคำที่โดยปกติน่าจะใช้บอกการเริ่มประโยคใหม่ได้ เช่น "ถึงแม้ว่า" "ยิ่งกว่านั้น" เพราะเป็นคำที่ทำหน้าที่ดัชนีปริจเฉท (discourse marker) แต่หากดูข้อมูลมากขึ้น ก็จะพบว่าคำเหล่านี้ไม่จำเป็นต้องบอกจุดเริ่มต้นประโยคก็ได้ (ดูตัวอย่างท้ายรูป)

ส่วนเรื่องระดับตงแสดงสถาปัตยกรรมอันได้แก่ หน้าบันและเสาดัดกับผนังของโคปุระนั้น แสดงถึงการสืบทอดของศิลปะแบบนครวัดจากศิลปะแบบบาปวน เสด็จติดกับผนังในศิลปะบาปวนปรากฏ ณ ประตูห้องด้านเหนือและด้านใต้ของโคปุระด้านทิศตะวันออกโดยเสาดัดกับผนังเหล่านี้ประดับด้วยลายกันต่อดอกซึ่งมีส่วนกว้างมากกว่าส่วนสูงกับที่มีขีดของขีดแบ่งกันของลายออกเป็นสามส่วน เสด็จติดกับผนังของประตูด้านใต้ของห้องด้านทิศใต้ของโคปุระทิศตะวันออกก็เช่นเดียวกัน คือเป็นศิลปะแบบบาปวน หากแต่เจ้าหลักลายกันขดแทนลายกันต่อดอกซึ่งโคนของลายกันขดมีรูปสัตว์ประกอบรวมทั้งลายกันขดนี้ไม่ได้เจ้าหลักตลอดความกว้างของตัวเสาแต่อย่างใด ถึงแม้ว่าเสาดัดกับผนังข้างต้นจักเป็นศิลปะแบบบาปวน แต่หน้าบันเหนือประตูเข้าด้านตะวันออกของห้องด้านเหนือและด้านใต้ของโคปุระแสดงถึงการประสมระหว่างศิลปะแบบบาปวนกับศิลปะแบบนครวัด ตัวอย่างดังกล่าวอาจเป็นไปได้จากนาคปลายกรอบหน้าบันของประตูทางเข้าห้องด้านเหนือของโคปุระด้านทิศตะวันออก ซึ่งเป็นนาคปราศจากริมมีตามศิลปะแบบบาปวน แต่นาคปลายกรอบหน้าบันของประตูทางเข้าของห้องด้านใต้ของโคปุระเดียวกันนี้เป็นนาคซึ่งมีริมมีตามศิลปะแบบนครวัด ยิ่งกว่านั้นส่วนล่างของหน้าบันอันหลังยังคงแต่งด้วยชื่อปลอมที่คั้งได้ฉากซึ่งเป็นที่นิยมในศิลปะแบบนครวัดอีกด้วย (NACHM030)

- // การใช้งานของโปรแกรมถึงแม้ว่าจะยังไม่สมบูรณ์ แต่ก็ได้แสดงให้เห็นถึงวิธีการต่างๆ ของโปรล็อกมาเป็นลำดับ //
- // แบบจำลองโปรแกรมมิ่งต้องการข้อมูลที่เป็นจริง ถึงแม้ว่าในแง่ของจำนวนจะไม่มากเท่ากับแบบจำลองเศรษฐมิติก็ตาม //
- // บ้านที่ปรากฏขึ้นต่อใจหรือต่อการรับรู้ของเรา จะไม่ใช่ทั้งบ้านในฝันหรือบ้านในชีวิตที่ตื่นของเรา หากแต่เป็นบ้านที่ ยิ่งกว่านั้น มลิ่งเมื่องระยิบระยับในความรู้สึก เป็นบ้านที่เป็นที่อาศัยแห่งชีวิตที่น่าพิศวง //

//อะไรกันนี้ //สงสัยผลองถูกลอตเตอร์รี่ใจ //ฉันร้อง ยิ่งกว่านั้นอีกพี//

การจบประโยคในภาษาไทยจึงไม่มีตัวบ่งชี้ชัดเจนเหมือนในภาษาอังกฤษ ไม่สามารถมองเป็นปัญหาการแก้ความกำกวมแบบภาษาอังกฤษว่าเครื่องหมาย . ณ ตำแหน่งนี้ใช้จบประโยคได้หรือไม่ เพราะการจบหรือเริ่มประโยคใหม่ในภาษาไทย อาจใช้การเว้นวรรค หรือไม่เว้นวรรค อาจมีคำเชื่อมหรือไม่ก็ได้ การเว้นวรรคก็ใช้แยกคำวลีหรืออนุพจน์ได้ด้วย หากจะทำตัดประโยคภาษาไทยด้วยเครื่องจึงไม่ใช่เรื่องง่ายและมีโอกาสผิดพลาดสูงมากกว่าการทำตัดประโยคภาษาอังกฤษ

ควรตัดประโยคหรือไม่

ถ้าเป็นภาษาอังกฤษ การตัดประโยคในภาษาเขียนไม่ใช่ปัญหาที่จัดการยาก หากตัดได้ ก็จะช่วยแยกหน่วยสำหรับการประมวลผลภาษาให้ชัดเจน จึงไม่แปลกที่จะมีการตัดประโยคออกมาก่อนได้ อย่างไรก็ตาม ในภาษาเขียนจริง ๆ หรือภาษาพูด ก็มักพบการเขียนแบบลากความเชื่อมโยงไปเรื่อย ๆ โดยใช้เครื่องหมายวรรคตอนหรือคำสันธาน ประโยคที่เป็น complex, compound, หรือ compound-complex ถึงแม้ว่าจะตัดเป็นประโยคออกมาได้ ก็ยังจำเป็นต้องประมวลผลต่อว่าแต่ละเนื้อความมีความเกี่ยวข้องสัมพันธ์กันอย่างไรต่อ เพราะประโยคที่ยังซับซ้อนมากก็ยังมีเนื้อความจำนวนมากมาประกอบกันด้วยความสัมพันธ์แบบต่าง ๆ

ในทางภาษาศาสตร์ ระดับที่สูงกว่าประโยคคือระดับปริจเฉทหรือ discourse ซึ่งพยายามอธิบายว่าตัวบทหรือ text ประกอบด้วยอะไรบ้างและมีความสัมพันธ์กันอย่างไร เหมือนเช่นที่คำถูกนำมาเรียงต่อกัน จะมีความสัมพันธ์ทางวากยสัมพันธ์ที่ทำให้เราเห็นถึงโครงสร้างวลีโครงสร้างประโยคได้ ปริจเฉทก็เป็นผลลัพธ์จากการนำข้อความมาเรียงต่อกันโดยที่สามารถระบุความสัมพันธ์ระหว่างข้อความต่าง ๆ เป็นโครงสร้างปริจเฉทได้

Rhetorical Structure Theory

ทฤษฎีหนึ่งที่เสนอเพื่อใช้อธิบายเรื่องนี้ คือ Rhetorical Structure Theory หรือทฤษฎีโครงสร้างวาทะ (Mann and Thompson, 1988) ซึ่งมองว่าข้อความใน text สามารถแยกเป็นส่วน ๆ (text span) ซึ่งมี rhetorical relation หรือวาทสัมพันธ์ต่อกันได้ แต่ละหน่วยข้อความที่สัมพันธ์กันจะถูกระบุว่าเป็นแกนกลาง (nucleus) หรือบริวาร (satellite) ส่วนที่เป็นบริวารทำหน้าที่ขยายความส่วนแกนกลางและสามารถละได้ กรณีที่หน่วยข้อความสำคัญพอ ๆ กันก็เป็นความสัมพันธ์แบบหลายแกนกลาง จำนวนวาทสัมพันธ์ที่ Mann and Thompson เสนอไว้มี 28 อย่าง เช่น Circumstance, Elaboration, Background, Evidence, เป็นต้น อย่างไรก็ตาม Mann and Thompson (1988) ไม่ได้ให้นิยามที่ชัดเจนว่าหน่วยพื้นฐานใน RST คืออะไรมีขอบเขตแค่ไหน ภายหลัง Carson et. al (2003) จึงใช้คำ EDU (elementary discourse unit) แทน text span เพื่อแทนหน่วยพื้นฐานในการวิเคราะห์ความสัมพันธ์และวางแผนผังเพื่อวิเคราะห์ EDU ในภาษาอังกฤษให้ชัดเจน ซึ่งโดยทั่วไป EDU จะเป็นอนุภาคเต็มรูปหรือลดรูปก็ได้ ดังนั้น แม้ว่าจะแยกประโยคออกมา ในการวิเคราะห์ตัวบทก็ยังคงดูประโยคทั้งหมดและอนุภาคและความสัมพันธ์ระหว่างอนุภาคภายในประโยคซับซ้อนด้วยเพื่อเชื่อมโยงให้เห็นโครงสร้างปริจเฉท เช่น ตัวอย่างข้างล่างจาก Carson et. al (2003) ที่วิเคราะห์ text เป็น EDU ต่าง ๆ และโยงความสัมพันธ์ระหว่าง EDU ทั้งภายในประโยคและระหว่างประโยค

หากยึดแนวทางการวิเคราะห์ภาษาแบบนี้ ประโยคจึงไม่ใช่เป้าหมายสุดท้าย การตัดประโยคได้ หากทำได้ง่ายแบบภาษาอังกฤษก็ควรทำ แต่หากทำได้ยากอย่างไรกรณีภาษาไทย ซึ่งไม่ได้ถูกสอนมาให้เขียนเป็นประโยคอย่างมีหลักเกณฑ์ร่วมกัน ก็ไม่จำเป็นต้องทำ หากต้องการตัดหน่วยที่ใหญ่กว่าคำหรือวลีเพื่อใช้ในการประมวลผลก็ควรตัดเป็นหน่วยปริจเฉทพื้นฐานหรือ EDU เพื่อให้ได้เนื้อความแต่ละส่วนสำหรับนำไป

ประมวลผลในระดับปริจเฉทต่อไป เพราะท้ายที่สุด ในการประมวลผลระดับปริจเฉท หน่วยที่ต้องการคือ EDU ไม่ใช่ประโยค

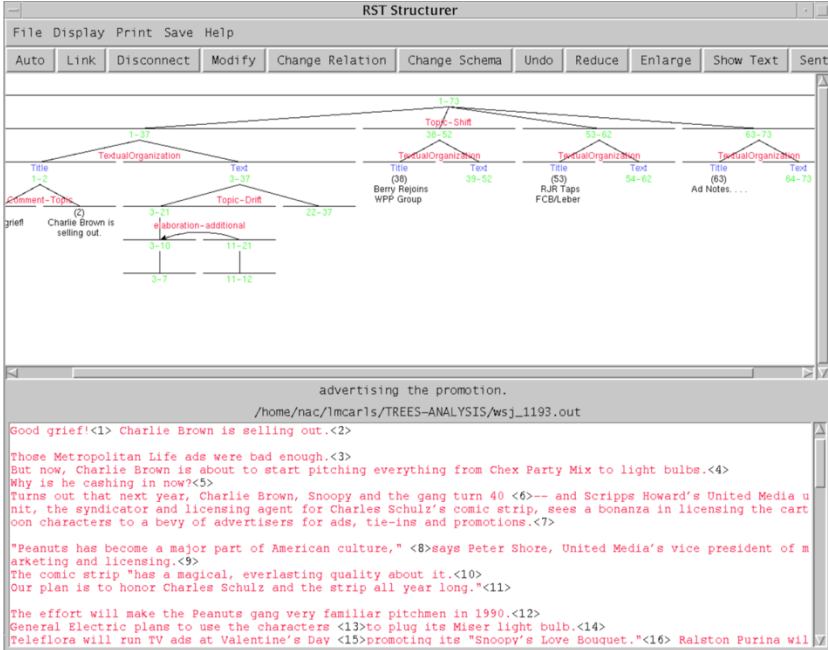


Figure 2. Top Level Discourse Tree for Document #1193: highly structured with topic shifts

การตัดหน่วยปริจเฉทพื้นฐาน

เมื่อเห็นแล้วว่าหากต้องการตัดประโยคภาษาไทย หน่วยที่ควรตัดคือหน่วยปริจเฉทพื้นฐานมากกว่า ซึ่งเป็นหน่วยที่ส่วนใหญ่จะเป็นอนุภาค Intasaw and Aroonmanakun (2013) ได้อาศัยแนวการวิเคราะห์ EDU ภาษาอังกฤษของ Carson et. al (2003) มาปรับใช้กับภาษาไทย หลักเกณฑ์ในการตัด EDU ภาษาไทยต่อไปนี้

นำมาจาก Intasaw and Aroonmanakun (2013) และวิทยานิพนธ์ของงนลินี (2556) (เครื่องหมาย [...] ใช้บอกขอบเขต EDU)

1. อนุพากย์ที่ทำหน้าที่เป็นประธานหรือกรรมของกริยา (subject/object clause) ไม่ถือเป็น EDU อิสระเช่น **[ผู้จบปริญญาเอกด้านวิทยาศาสตร์ต้องมีคุณสมบัติอย่างไรบ้าง]** clausal subject นี้จะถือเป็นส่วนหนึ่งของ EDU
2. อนุพากย์ที่เป็นคุณาปุประโยค (relative clause) โดยมีคำ **ที่** ซึ่ง **อัน** นำหน้าและทำหน้าที่ขยายนามหลักให้ถือเป็นอีกหนึ่ง EDU ที่เป็นหน่วยพืงพา เช่น **[เนื่องจากขาดการศึกษาและวางแผนโยบาย][ที่ชัดเจน][อันจะทำให้ประชาชนสามารถตัดสินใจได้]** มีสาม EDU
3. อนุพากย์ที่ทำหน้าที่ขยายกริยา (adverbial clause) บอกความสัมพันธ์ต่าง ๆ เช่น เวลา อากาศ เหตุผล เงื่อนไข ฯลฯ ให้ถือเป็น EDU **[ให้ความสำคัญแก่การวางโครงการ][ที่สลับซับซ้อน][เพื่อหลีกเลี่ยงให้คนอ่านเดาเรื่องไม่ออก]**
4. อนุพากย์ที่เชื่อมความกัน (coordinate clause) ให้วิเคราะห์แต่ละอนุพากย์เป็น EDU **[ความยากจนเป็นปัจจัยนำไปสู่การเกิดพยาธิสภาพแก่ปัจเจกบุคคล][และมีผลกระทบต่อส่วนรวม]** แต่หากเป็นการเชื่อมคำกริยาหรือคำนามจะไม่แยก EDU. **[แต่หลายส่วนลอกและเพิ่มเติมมาจากกฎหมายตราสามดวง]**
5. อนุพากย์ที่กริยาไม่ใช่กริยาแท้ (non-finite verb) รวมถึงคุณาปุประโยคลดรูป (reduced relative clause) ให้วิเคราะห์เป็นส่วนประกอบใน EDU **[โดยได้แสดงวิธีการวิเคราะห์สารสำคัญจากตำนานอิตีฟุส]**
6. ส่วนที่ทำหน้าที่เป็นส่วนเติมเต็ม (clausal complement) ไม่ว่าจะป็นวลีหรืออนุพากย์ ซึ่งมักพบในกริยาบอกคุณลักษณะ (attributive verb) รวมถึงกริยา reporting speech และกริยาบังปริชาน **ยอมรับ คิด เชื่อ แสดง สันนิษฐาน เสนอ รู้ อธิบาย แนะนำ ตัดสินใจ สมมติ ถาม สงสัย** มักมีคำเชื่อม **ว่า** **ที่** **หน้า**

- ส่วนนั้น [ชี้ให้เห็นชัดเจน][**ว่ามีการละเมิดสิทธิขั้นพื้นฐานของประชาชน**] แต่หากส่วนเติมเต็มไม่ใช่กริยาแท้ (non-finite clausal complement) [**หากพร้อมที่จะปลูกเรือนใหม่**][จึงแยกเรือน]
7. โครงสร้างกริยาเรียง (serial verbs) ให้วิเคราะห์ทั้งหมดเป็นส่วนหนึ่งของ EDU. [ขณะเดียวกันก็**รอคอยโชคชะตา**มาพลิกผันชีวิตให้**แปรเปลี่ยนไป**]. แต่ถ้าในกริยาเรียงเป็นกริยาบอกคุณลักษณะ (attributive verb) ให้แยกเป็นคนละ EDU. [**เพราะเขาคิด**][**ว่าเขาขาดโอกาสทางธุรกิจ**]
 8. Cleft sentence เป็นประโยคภาษาอังกฤษที่ย้ายที่ประธานหรือกรรมมาข้างหน้าเพื่อการเน้น เช่น *It was the book that I read last night.* ในภาษาไทยไม่ได้ใช้ dummy subject แต่เขียนโดยใช้กริยา เป็น คือ เพื่อเน้นประธานหรือกรรมนั้นได้ (Ruangjaroon, 2005) กรณีของ contrastive cleft ที่ใช้คำกริยาเป็น จะวิเคราะห์เป็น EDU เดียว [**นึกที่เป็นคนทำงานแตก**] ส่วนกรณี identification cleft ที่ใช้กริยา คือ จะวิเคราะห์แยกเป็นคนละ EDU เพราะเป็นโครงสร้างแบบคุณานุประโยค [**ใครคือคน**][**ที่นิดแอบชอบ**]
 9. วลีที่มีคำบ่งชี้ปริจเฉท เช่น **ได้แก่ เป็นต้น อย่างเช่น เพื่อ** นำหน้าสามารถวิเคราะห์เป็นอีก EDU ได้ เพราะถือเป็นส่วนที่มีหน้าที่ทางความ [กู่หนียมลินมา][**เพื่อการต่อสู้คดี**] [...**ปรากฏการณ์ทางธรรมชาติ**] [**เช่น ลม ฝน กลางวัน กลางคืน พายุร้อน พายุผ่า**]
 10. หน่วยสร้างที่ถูกแยกส่วนให้วิเคราะห์เป็นคนละ EDU ได้ [**ต่อมาในสมัยหลังสมัยใหม่**] [(Post-modern)] [**ได้เกิดวรรณกรรมแนวทดลอง**] ตัวอย่างนี้มีการแทรก (Post-modern) เข้ามาในข้อความทำให้ต้องแยกส่วนหน้าส่วนหลังเป็นคนละ EDU รวมถึงข้อความที่แทรกก็เป็นอีก EDU
 11. เครื่องหมายวรรคตอนให้รวมเป็นส่วนหนึ่งของ EDU นั้น
 12. คำเชื่อม คำบ่งชี้ปริจเฉท ให้นำมารวมเป็นส่วนหนึ่งของ EDU นั้น

การตัด EDU เพื่อให้เห็นว่าข้อความนั้นมีหน่วยปริจเฉทพื้นฐานอะไรบ้าง หลังจากนั้นจึงวิเคราะห์หาความสัมพันธ์เชิงความระหว่างหน่วยตามแนวทฤษฎีโครงสร้างวาทะ (Rhetorical structure theory) ตามที่กล่าวมาแล้ว เมื่อวิเคราะห์แล้ว เราจะเห็นโครงสร้างความภายในปริจเฉทนั้นว่าส่วนใดเป็นส่วนขยายของส่วนใด และมีความสัมพันธ์ทางความกันอย่างไร วิธีนี้จะทำให้ได้โครงสร้างความภายในปริจเฉทที่นำมาใช้ประโยชน์อื่น ๆ ต่อไปได้ ศุภวัจน (2564) ได้ให้ข้อมูลรายละเอียดของการวิเคราะห์ตามทฤษฎีโครงสร้างวาทะและแนวทางการนำผลการวิเคราะห์ไปใช้ในการประมวลผลภาษาไทยเรื่องต่าง ๆ เช่น การสรุปย่อความ การสกัดสารสนเทศ การวิเคราะห์ความรู้สึก ไปจนถึงงานการเรียนการสอนเรื่องการเขียนได้

โปรแกรมตัดหน่วยปริจเฉทพื้นฐาน

หลังจากนำเสนอเรื่องแนวคิดหน่วยปริจเฉทพื้นฐานแล้ว นลินี (2556) ได้ทดลองพัฒนาโปรแกรมตัดหน่วยปริจเฉทพื้นฐาน โดยใช้ข้อมูลงานเขียนจำนวนหนึ่งจากคลังข้อมูลภาษาไทยแห่งชาติมากำกับขอบเขตหน่วยปริจเฉทพื้นฐานตามแนวทางที่กำหนดไว้จำนวนประมาณ 8,100 EDU. จากนั้นทดลองใช้แบบจำลอง support vector machine เรียนรู้จากข้อมูลที่เตรียมไว้ โดยใช้ลักษณะด้านคำ หมวดคำ รายการคำเชื่อม ช่องว่าง และเครื่องหมายวรรคตอน (ข้อมูลหมวดคำกำกับเองด้วยมือ) ผลการทดลองได้ค่าความถูกต้อง (คำนวณรวมค่าความแม่นยำและความครบถ้วน) ที่ประมาณ 84%

อย่างไรก็ดี ผู้เขียนเลือกทดลองพัฒนาโปรแกรมตัดหน่วยปริจเฉทพื้นฐานด้วยการนำข้อมูลในงานของนลินีมาใช้ใหม่โดยปรับให้ดูข้อมูลที่ระดับพยางค์แทนระดับคำ และไม่ใช้ข้อมูลหมวดคำเพราะอยากให้โมดูลการแยก EDU เป็นงานที่ไม่จำเป็นต้องผ่านการกำกับหมวดคำมาก่อน ลักษณะที่ใช้จึงมีเพียงรูปพยางค์ในตำแหน่งนั้นและบริบทหน้าหลัง ช่องว่าง เครื่องหมายไม้ยมก ตัวเลข คำต่างประเทศ และทดลองใช้

แบบจำลองต่าง ๆ ผลที่ได้คือแบบจำลอง Random Forest ได้ผลดีที่สุดคือ 97% จึงเลือกใช้แบบจำลองนี้พัฒนาเป็นโมดูล tltk.nlp.segment ในแพ็คเกจ TLTK

อย่างไรก็ดี การตัดหน่วยปริจเฉทพื้นฐานใน TLTK นี้ เลือกทำแนวทางเดียวกับ นลินี (2556) คือให้ผลลัพธ์ที่เป็นการตัด EDU เป็นส่วน ๆ ไม่ได้มีการวิเคราะห์โครงสร้างปริจเฉทที่อาจจะซับซ้อนได้ ตัวอย่างข้อความ *กะหล่ำปลีที่ถูกต้องทำหลายจะมีสีเหลือง* หากตัดมาได้ตามที่ออกแบบจะถูกตัดออกมาเป็น 3 EDU คือ *[กะหล่ำปลี][ที่ถูกต้องทำหลาย][จะมีสีเหลือง]*. ไม่ใช่ 2 EDU ที่มี EDU ย่อยฝังในประโยคหลัก *[กะหล่ำปลี][ที่ถูกต้องทำหลาย][จะมีสีเหลือง]* เช่นเดียวกับข้อความที่เป็นส่วนต่อเนื่องกันแต่ถูกแทรกด้วยข้อความอื่นตามหลักเกณฑ์ข้อ 10 ก็จะถูกตัดแยกเป็นคนละ EDU. ในกรณีเช่นนี้ หากผู้ใช้ต้องการวิเคราะห์โครงสร้างวาทะจากข้อมูลที่ตัด EDU แบบนี้ ก็ต้องพิจารณาตรวจสอบที่ควรเป็น EDU เดียวกันแต่ถูกแบ่งออกไประหว่างการแจงส่วนปริจเฉทด้วย

สาธิต

โปรแกรม TLTK มีโมดูลที่ใช้ตัด EDU คือ tltk.nlp.segment(Text) โดยกำกับ <u/> ไว้ในตัวบทนั้น วิธีการที่ใช้คือใช้การเรียนรู้ด้วยเครื่องจากข้อมูลที่ใช้ฝึกจำนวน 8,100 EDU ข้อมูลถูกแบ่งพยางค์และใช้ RandomForestClassifier ในการตัดสินใจว่าในแต่ละตำแหน่งพยางค์นั้นควรกำกับ <u/> หรือไม่ นอกจากแท็ก <u/> แล้วในโมดูลจะตัดคำโดยใส่เครื่องหมาย | ไปด้วย สามารถทดลองใช้ได้ที่

https://colab.research.google.com/drive/1_YCrhKqO9_YEVXgs-_HYXv4n5N73LdzF2Y?usp=sharing

```
text = "แต่อาจเพราะนกกินเปลือกเหลืองเป็นพ่อแม่มือใหม่ รังที่ทำจึงไม่ค่อยแข็งแรง วันหนึ่งรังก็ฉีกเกือบขาดเป็นสองท่อนห้อยต่องแต่ง ผมพยายามหาอุปกรณ์มาอีตริงกลับคืนรูปทรงเดิม ขณะที่แม่นกกินเปลือกเหลืองส่งเสียงโวยวายอยู่ใกล้ ๆ แต่สุดท้ายไม่สำเร็จ สองสามวันต่อมารังที่ช่วยซ่อมก็พังไป ไม่เห็นแม่นกบินกลับมาอีกเลย"
tltk.nlp.segment(text)
```


Ruangjaroon, S. (2005). The syntax of WH-expressions as variables in Thai. Vancouver, University of British Columbia.

บุญ อินทร์ทรัพย์. (2002). "การใช้เครื่องหมายวรรคตอนในภาษาไทย ตามพระราช
นิยมในพระบาทสมเด็จพระมงกุฎเกล้าเจ้าอยู่หัว" The Journal of the Royal
Institute of Thailand. Vol. 27 N^o2 Apr.-Jun. 2002.

([http://www.royin.go.th/wp-content/uploads/royin-
ebook/94/FileUpload/527_1523.PDF](http://www.royin.go.th/wp-content/uploads/royin-ebook/94/FileUpload/527_1523.PDF))

นลินี อินตะชาว. (2556). การแยกอนุพยางค์ภาษาไทยด้วยการใช้แบบจำลองซัพพอร์ต
เวกเตอร์แมชชีน. วิทยานิพนธ์มหาบัณฑิต สาขาวิชาภาษาศาสตร์ คณะอักษร
ศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ศฤงฆ์ จันทร์รุ่งเรือง. (2564). ทฤษฎีโครงสร้างวาทะ: จากมุมมองเชิงทฤษฎีสู่การ
ประมวลผลภาษาไทย วารสารวณะปีที่ 9 ฉบับที่ 2 (กรกฎาคม-ธันวาคม 2564)

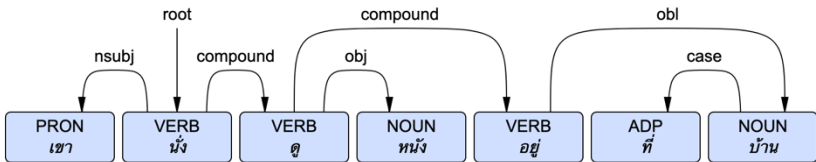
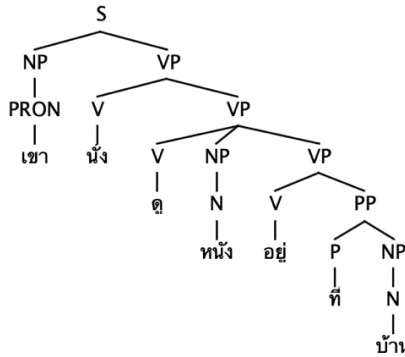
การวิเคราะห์ภาษาไทยด้วยไวยากรณ์พึ่งพา¹³

การประมวลผลภาษาในระดับเหนือคำเป็นการมองการประกอบกันของคำเป็นหน่วยภาษาที่ใหญ่ขึ้น เช่น วลี อนุพจน์ ประโยค การวิเคราะห์การประกอบกันของคำมีสองแนวทางหลัก คือ การหาหน่วยสร้างกับการหาการพึ่งพาระหว่างคำ ในการวิเคราะห์หาหน่วยสร้าง จะมองว่าคำหลายคำที่มีความสัมพันธ์ทางวากยสัมพันธ์สามารถรวมเป็นหน่วยสร้างใหม่และหน่วยสร้างนั้นก็รวมกับคำหรือหน่วยสร้างอื่นต่อไปได้เรื่อย ๆ ผลที่ได้เป็นโครงสร้างต้นไม้ที่มีลำดับชั้น ส่วนการวิเคราะห์การพึ่งพา เป็นการมองหาคำที่มีความสัมพันธ์ทางวากยสัมพันธ์แล้วดูว่าคำใดเป็นคำหลัก ผลการวิเคราะห์ที่ได้อาจเป็นเครือข่ายการพึ่งพาหรือเป็นโครงสร้างต้นไม้แบบพึ่งพาก็ได้ ขึ้นอยู่กับการกำหนดเงื่อนไขการวิเคราะห์ ถ้าวิเคราะห์แบบต้นไม้แต่ละคำจะมีคำหลักหรือ head ได้เพียงคำเดียวและท้ายสุดจะมีคำเดียวที่เป็นคำหลักตั้งต้นหรือ root. ตัวอย่างข้างล่างแสดงผลการวิเคราะห์ทั้งสองแบบ

รูปข้างล่างแสดงการวิเคราะห์ประโยค *เขานั่งดูหนังอยู่ที่บ้าน* ตามแนวทางการวิเคราะห์หน่วยสร้างและตามแนวทางการวิเคราะห์แบบไวยากรณ์พึ่งพา ในการวิเคราะห์แบบหน่วยสร้าง *ที่บ้าน* รวมเป็น PP และรวมกับ *อยู่* เป็น VP. จากนั้นรวมกับ *ดู* *หนัง* เป็น VP ก่อนจะรวมกับ *นั่ง* เป็น VP บนสุด จากนั้นรวมกับ NP *เขา* เป็น S. ในขณะที่การวิเคราะห์แบบพึ่งพาจะไม่มีโครงสร้างโหนดสำหรับ NP, VP, PP, S เหล่านี้

¹³ เนื้อหาบางส่วนในบทนี้นำมาจาก “คู่มือการกำกับข้อมูลภาษาไทยตามแนวทาง Universal Dependencies version 2” ที่ผู้เขียนเผยแพร่บนหน้าเว็บไซต์ผู้เขียน

แต่จะโยงความสัมพันธ์ระหว่างคำหลักกับคำรองพร้อมความสัมพันธ์ของสองคำนั้น
 นั้ง เป็นคำหลักของ เขา และ ดู, ดู เป็นคำหลักของ หนังสือ และ อยู่, อยู่ เป็นคำหลักของ
 บ้าน และ บ้าน เป็นคำหลักของ ที่



การวิเคราะห์แบบหน่วยสร้างต้องอาศัยกฎโครงสร้างวลีที่บอกว่าหน่วยสร้าง
 นั้นประกอบด้วยหน่วยย่อยอะไรบ้าง เช่น VP -> V NP, VP -> V, VP -> V PP เป็น
 ต้น การวิเคราะห์แบบนี้จะเกี่ยวข้องกับชุดของหมวดคำโดยตรง หากต้องการระบุกฎ
 ให้ตรงกับกลุ่มคำที่เจาะจงลงไปก็จะต้องแยกหมวดคำที่มีรายละเอียดมากขึ้น เช่น
 กริยาอาจต้องระบุว่าเป็น active verb (VACT ทำงาน, ร้องเพลง, กิน), stative verb
 (VSTA เห็น, รู้, คื่อ), หรือ attributive verb (VATT อ้วน, ดี, สวย) เพื่อให้ใช้ได้กับ
 กลุ่มคำกริยาแต่ละประเภท คำสรรพนามก็อาจแยกประเภทย่อยเป็น personal
 pronoun (PPRS คุณ, เขา, ฉัน), demonstrative pronoun (PDMN นี้, นั้น, ที่นี่),
 interrogative pronoun (PNTR ใคร, อะไร, อย่างไร) หรือ relative pronoun (PREL

การประมวลผลภาษาไทย

ที่, ซึ่ง, ผู้) เป็นต้น นอกจากนี้ หากในโครงสร้างนั้นมีการละบางองค์ประกอบได้หรือไม่ ก็ต้องระบุว่าส่วนนั้นมีหรือไม่มีก็ได้ (optional). เมื่อเปรียบเทียบกับการวิเคราะห์แบบ ฟังพาแล้ว การวิเคราะห์แบบฟังพาไม่ต้องอาศัยหมวดคำที่เจาะจงมากได้ และหากมีการละคำบางคำไปก็ไม่มีการประมวลผล เพราะการวิเคราะห์แบบฟังพาเป็นการ ดูคำภายในประโยคเพื่อมองหาคำที่มีนั้นมีความสัมพันธ์กับคำใด คำไหนเป็นคำหลัก ผู้เขียนจึงคิดว่าเหมาะสมกับการประมวลผลภาษาไทยมากกว่า เพราะภาษาไทยเป็น ภาษาที่มีการละคำได้ง่าย

การวิเคราะห์ด้วยไวยากรณ์ฟังพา

การวิเคราะห์ด้วยไวยากรณ์ฟังพาเป็นแนวทางหนึ่งของการวิเคราะห์ทาง วากยสัมพันธ์ มีหลายทฤษฎีและโครงการที่ใช้แนวคิดนี้ เช่น ไวยากรณ์ศัพท์การกรของ Stanley Starosta (1988). ไวยากรณ์คำของ Richard Hudson (1984). และโครงการ Universal Dependencies การวิเคราะห์ตามแนวไวยากรณ์ฟังพามีทั้งแบบที่ วิเคราะห์ให้แต่ละคำในประโยคมีคำหลักได้เพียงคำเดียวกับแบบที่ยอมให้มีคำหลักได้ มากกว่าหนึ่งคำ หากวิเคราะห์ให้มีคำหลักได้คำเดียวโครงสร้างที่ได้จะเป็นโครงสร้าง ต้นไม้ที่มีค่าตั้งต้นหรือ root เพียงคำเดียว การวิเคราะห์แบบ Universal Dependencies ที่ใช้ในที่นี่เป็นแบบวิเคราะห์แบบให้แต่ละคำมีคำหลักได้คำเดียว ซึ่ง ในการวิเคราะห์แบบนี้จึงไม่มีการเกยกันของสายฟังพาเกิดขึ้น

ความสัมพันธ์แบบฟังพาตามแนวทาง Universal Dependencies

Universal Dependencies เป็นโครงการสากลที่สร้างขึ้นเพื่อสร้างข้อมูลภาษาที่ เป็นต้นไม้แบบฟังพาสำหรับทุก ๆ ภาษา โดยการกำหนดชุดหมวดคำและชุด ความสัมพันธ์ให้ใช้ร่วมกัน ในการทดลองวิเคราะห์ความสัมพันธ์แบบฟังพาสำหรับ ข้อมูลภาษาไทยจำนวนหนึ่งจากคลังข้อมูลภาษาไทยแห่งชาติ โดยยึดความสัมพันธ์ที่

เสนอในโครงการ Universal Dependencies จำนวนความสัมพันธ์ที่ใช้วิเคราะห์
ภาษาไทยได้มีทั้งสิ้น 37 ความสัมพันธ์ดังมีรายละเอียด ดังนี้

Type	ชื่อ	คำอธิบาย
Core arguments - Nominals	1	<p>nsubj nominal subject</p> <p>ความสัมพันธ์ระหว่างกริยากับประธาน ประธานของประโยค อาจเป็น คำนาม (NOUN) สรรพนาม (PRON) หรืออาจเป็นคำคุณศัพท์ ADJ ได้ในกรณีที่เกิดการละ โดยที่ประธานของประโยคอาจไม่จำเป็นต้องเป็นผู้กระทำกริยานั้น ๆ (agent) โดยตรงก็ได้ หากประโยคมีความสัมพันธ์แบบกรรมวาจก (passive) จะระบุความสัมพันธ์ของกริยาโยงไปยังประธาน เป็น nsubj:pass</p> <p>ความสัมพันธ์แบบ nsubj อาจไม่จำเป็นต้องตั้งต้นมาจาก VERB เสมอไป ในกรณีที่ VERB ที่ปรากฏในประโยคนั้นเป็น COPULAR VERB ส่วนตั้งต้นหรือ head จะเป็นสิ่งที่มาเติมเต็ม COPULAR VERB นั้น ซึ่งอาจเป็น ADJ หรือ NOUN ก็ได้</p> <p>ในกรณีที่ไม่มี predicate เช่น <i>วันนี้วันศุกร์</i> ก็ให้ใช้ nsubj (<i>วัน นี้</i>) < (<i>วัน ศุกร์</i>)</p> <p>เมื่อไหร่ก็ตามที่ประธานมีลักษณะเป็น empty argument (ประโยคที่ขึ้นต้นด้วย there) ความสัมพันธ์ระหว่าง VERB กับประธานจะเป็น expl แทน ในภาษาไทย อาจพบในประโยค เช่น <i>มันเป็นไปได้ว่า...</i> <i>มัน</i> < <i>expl</i> < เป็น <i>มัน</i> เป็นเหมือน dummy subject</p>
	<p>The diagram shows two syntactic structures. The first structure is for the sentence 'ฉัน กิน ข้าว' (I eat rice). A red arrow labeled 'nsubj' points from 'ฉัน' to the verb 'กิน', and a black arrow labeled 'obj' points from 'กิน' to 'ข้าว'. The second structure is for the sentence 'วัน นี้ วัน ศุกร์' (This day is Friday). A red arrow labeled 'nsubj' points from 'วัน' to the verb 'วัน', and a black arrow labeled 'nmod' points from 'วัน' to 'ศุกร์'. The word 'นี้' is positioned between the two 'วัน' words.</p>	

การประมวลผลภาษาไทย

Type		ชื่อ	คำอธิบาย
Core arguments - Nominals	2	obj object	<p>ความสัมพันธ์ระหว่างกริยากับกรรมตรงของกริยานั้น โดยปกติจะเป็นคำนาม และหากกริยานั้นมีกรรมเพียงแต่ตัวเดียวจะระบุความสัมพันธ์เป็น obj แต่หากกริยาเป็นกริยาทวิกรรมคือมีกรรมสองตัว หรือมีกรรมมากกว่า 2 ตัว จะโยงความสัมพันธ์ให้กรรมตัวหนึ่งเป็น obj (เป็นกรรมที่รับผลจากการกระทำหรือกริยานั้นโดยตรง) ส่วนที่เหลือเป็น iobj</p>
	3	iobj indirect object	<p>iobj เป็นความสัมพันธ์หลักอีกความสัมพันธ์หนึ่ง โยงระหว่างกริยากับกรรมรอง ไม่มีคำบุพบทมาคั่นกลาง ในกรณีที่กรรมรองนำหน้าด้วยคำบุพบท จะถือว่ากริยากับกรรมรองนั้นมีความสัมพันธ์แบบ obl</p>

Type		ชื่อ	คำอธิบาย
Core arguments - Clauses	4	csubj clausal subject	<p>เป็นรูปแบบความสัมพันธ์ของกริยากับประธานของประโยค โดยประธานไม่ใช่คำนามเดี่ยวโดด ๆ แต่มีลักษณะเป็นอนุประโยค (clause) การโยงความสัมพันธ์จะโยงจากกริยาหลัก (หรือ head ของส่วนเติมเต็มของกริยา COPULAR) ไปยัง head ของอนุประโยคที่เป็นประธาน ซึ่งอาจจะเป็นกริยาหรือนามก็ได้</p> <p>ในกรณีที่ประธานเป็นประโยคกรรมวาจก (passive) การโยงจะระบุเพิ่มเติมเป็น csubj:pass</p> <p>ฝน ตก อย่าง รุนแรง ทำให้ เกิด น้ำท่วม</p>

การประมวลผลภาษาไทย

Type		ชื่อ	คำอธิบาย
Core arguments - Clauses	5	ccomp clausal complement	<p>เป็นประเภทการโยงกริยาหลัก (VERB หรือ ADJ) เข้ากับ head ของส่วนที่มาเติมเต็มอันปรากฏในตำแหน่งกรรมของประโยค ข้อสังเกตการโยงความสัมพันธ์นี้คือ ประธานของประโยคหลักกับประธานของอนุประโยคที่มาเติมเต็มสามารถเป็นคนละคนหรือคนละสิ่งกันได้ เช่น เขา (บอก) ว่า คุณ >(ชอบ) อ่าน หนังสือ แม้กรณีที่ไม่ปรากฏประธาน เช่น เขา (บอก) ว่า >(ชอบ) อ่าน หนังสือ ก็ยังคงเป็นความสัมพันธ์ ccomp</p>

Type	ชื่อ	คำอธิบาย
Core arguments - Clauses	6	<p>xcomp</p> <p>open clausal complement</p> <p>เป็นประเภทการโยงกริยาหลัก (VERB หรือ ADJ) เข้ากับ head ของส่วนที่มาเติมเต็ม ความแตกต่างระหว่าง ccomp กับ xcomp คือ ความสัมพันธ์แบบ xcomp นั้น ประธานในอนุประโยคเติมเต็ม ไม่สามารถมีได้จะเป็น subject เดียวกับของประโยคหลัก เช่น <i>He (likes) to >(swim)</i> ภาษาไทย เช่น ฉัน (ชอบ) >(ไป) เที่ยว ทะเล อาจมีการใช้คำเชื่อมหน้ากริยาหลังเป็น <i>ที่จะ... ที่ว่า....</i> ก็ได้ เช่น <i>เขา (ชอบ) ที่จะ >(ไป) ทำงาน สาย</i> คำเชื่อมเหล่านี้บอกความเป็น xcomp คือไม่สามารถมีประธานหน้ากริยาได้ ถ้าเทียบกับภาษาอังกฤษก็จะเป็น non-finite (กรณีประโยคอย่าง <i>ฉันชอบหนังสืออ่านแล้วตื่นเต้น</i> ไม่วิเคราะห์ <i>อ่านแล้วตื่นเต้น</i> เป็น xcomp แต่วิเคราะห์เป็น (หนังสือ) -acl->(อ่าน) แล้ว ตื่นเต้น คล้ายกับ <i>ฉันชอบหนังสือที่อ่านแล้วตื่นเต้น</i>)</p>

การประมวลผลภาษาไทย

Type		ชื่อ	คำอธิบาย
Non-core dependent - Nominals	7	obl oblique nominal	<p>ใช้โยงความสัมพันธ์ระหว่างคำกริยาหลัก คำคุณศัพท์หรือวิเศษณ์ กับคำนามในตำแหน่งกรรมของกริยานั้น แต่ต้องไม่ใช่กรรมตรง ใช้กับ non-core argument หรือ adjunct</p> <p>ข้อสังเกตของความสัมพันธ์แบบ obl คือคำนามนั้น ๆ มักนำหน้าด้วย ADPOSITION และความสัมพันธ์ระหว่างคำนามกับ ADP ข้างหน้าคือ case</p> <p>นอกจากนี้ obl ยังใช้โยงความสัมพันธ์ระหว่างคำขยายที่เป็นคำนามอยู่ข้างหน้าประโยคอีกด้วย เช่น</p> <p>(Last night) < , I (swam) in the pool หรือภาษาไทยเช่น (พรุ่งนี้) < ฉัน จะ (ไป) เที่ยว บ้าน เพื่อน</p>

Type		ชื่อ	คำอธิบาย
Non-core dependent - Nominals	8	vocative vocative	<p>แสดงความสัมพันธ์ระหว่างคำกริยากับคำนามที่เป็นคำเรียกขาน เช่น (แม่) < อย่า (เดิน) มาก</p> <p>แม่ อย่า เดิน มาก</p>
Non-core dependent - Nominals	9	expl expletive	<p>ใช้โยงความสัมพันธ์ระหว่างกริยากับสิ่งที่ไม่ได้เป็นคำนามชี้เฉพาะชัดเจน เช่น มัน ใน (มัน) < (เป็น) ไป ไม่ได้ จะเห็นได้ว่าคำว่า มัน อาจละทิ้งได้โดยไม่ส่งผลกระทบต่อความหมาย กล่าวคือ มัน ในตัวอย่างนี้เป็นคำนาม แต่ไม่มี semantic role กับ predicate</p> <p>มัน เป็น ไป ไม่ได้</p>

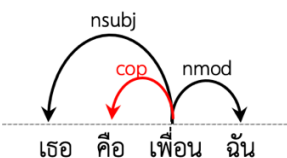
การประมวลผลภาษาไทย

Type		ชื่อ	คำอธิบาย
Non-core dependent - Nominals	10	dislocated dislocated element	<p>โยงคำกริยากับคำนามที่ถูกย้ายที่ เช่น (หนังสือ) < เล่ม นี้ ผม (ชอบ) มาก</p> <p>หนังสือ เล่ม นี้ ผม ชอบ มาก</p>
Non-core dependent - Clauses	11	advcl adverbial clause modifier	<p>โยงความสัมพันธ์ระหว่างภาคแสดงกับส่วนขยายภาคแสดงซึ่งมีลักษณะเป็นอนุประโยค เช่น เขา (สอบถาม) เจ้าหน้าที่ เพื่อ > (หา) ข้อมูล เพิ่มเติม</p> <p>เขา สอบถาม เจ้าหน้าที่ เพื่อ หา ข้อมูล เพิ่มเติม</p>

Type	ชื่อ	คำอธิบาย
Non-core dependent - Modifier words	12 advmod adverbial modifier	<p>โยงความสัมพันธ์ระหว่างกริยากับคำขยายกริยา เช่น (กิน) อย่าง >(รวดเร็ว) รวมถึงคำขยายอื่น ๆ ด้วย เช่น (เกือบ)<(2000) (หลาย)<(เล่ม)</p> <p>เนื่องจากใน UD ภาษาส่วนใหญ่ให้ ความสัมพันธ์ระหว่างคำปฏิเสธกับกริยาเป็น advmod ในภาษาไทยจึงใช้ advmod เช่น (ไม่)<(ชอบ)</p>

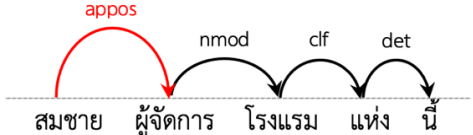
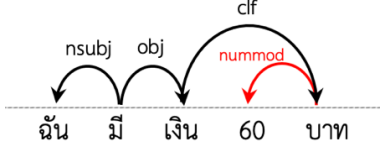
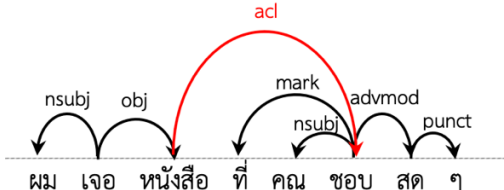
การประมวลผลภาษาไทย

Type	ชื่อ	คำอธิบาย
Non-core dependent - Modifier words	13 discourse discourse element	<p>ความสัมพันธ์นี้ใช้โยงคำในกลุ่ม INTERJECTIONS เช่น โอ อืม โอโห, กลุ่ม discourse marker (<i>actually, well, like</i>) โดยโยงกริยาหลักไปยังคำอุทานเหล่านั้น นอกจากนี้ยังสามารถใช้โยงคำกริยาไปยังคำลงท้าย (PART:RES) ได้อีกด้วย</p> <p>อ้าว ทำไม ทำ แบบ นี้ ละ</p>
Non-core dependent - Function Words	14 aux auxiliary	<p>เป็นความสัมพันธ์ระหว่างคำกริยากับคำช่วยกริยาต่าง ๆ (คำไวพจน์ที่บอก tense, aspect, mood, voice, evidentiality) โดยคำกริยาจะเป็น head เสมอ เป็นตัวตั้งต้น โยงไปยังคำช่วยกริยาแต่ละคำ เช่น เขา (กำลัง)< (จะ)< (ไป)</p> <p>เขา กำลัง จะ ไป</p>

Type		ชื่อ	คำอธิบาย
Non-core dependent - Function Words	15	cop copula	<p>โยงระหว่าง function word ที่เชื่อม subject กับ nonverbal predicate คำนามทางด้านซ้ายและขวาของ VERB:COP ต้องอ้างถึงสิ่งเดียวกันหรือบอกคุณลักษณะสิ่งนั้น เช่น <i>he (is) < (honest) เธอ (คือ) < (เพื่อน ฉันทน์)</i> ในภาษาไทยจะพบความสัมพันธ์นี้ในประโยคที่มีคำว่า <i>คือ</i> แต่กรณีที่เป็นกรโยงความด้วยอนุพากย์ ให้วิเคราะห์ คือ เป็นกริยาทั่วไป เช่น <i>สิ่งที่สำคัญคือต้องพยายามทำงานให้หนัก (สิ่ง) < nsubj-(คือ), (คือ)-ccomp > (ทำงาน) หรือ การพนันเป็นสิ่งที่มีมอมเมาประชาชน ก็ให้เป็น การพนัน < nsubj-(เป็น), (เป็น)- obj > สิ่ง</i></p> 
Non-core dependent - Function Words	16	mark marker	<p>โยงระหว่างกริยากับคำไวยากรณ์ที่บอกถึงการเป็น subordinate clause นั่นคือปกติจะโยงคำกริยาในอนุประโยคเต็มเต็มเข้ากับคำในกลุ่ม complementizer, subordinate conjunction เช่น <i>เขา รู้ (ว่า) < ผน กำลัง (ตก), นักเรียน (ที่) < (มา) วันนี้ หรือคำว่า อย่าง ที่ใช้นำหน้าคำขยายกริยา เช่น <i>กิน (อย่าง) < (รวดเร็ว) กรณีของ discourse</i></i></p>

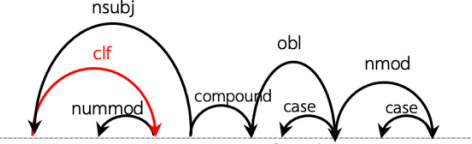
การประมวลผลภาษาไทย

Type		ชื่อ	คำอธิบาย
			<p>particle ที่ไม่ใช่ final particle ก็ให้วิเคราะห์ความสัมพันธ์เป็น mark ได้ เช่น (เพลง) ที่ได้เข้าชิงรางวัล (ได้แก่) <mark <(X) Y Z. นักเรียน(ที่)<(มา)วันนี้</p> <p>เขา รู้ ว่า ฟน กำลัง ตก</p>
Nominal dependent - Nominals	17	nmod nominal phrase modifier	<p>คำนามหรือนามวลีที่มาขยายอีกคำนามหนึ่ง เช่น (ห้องทำงาน) ของ >(หัวหน้า)</p> <p>ห้อง ทำงาน ของ หัวหน้า ของ ฝ่าย บริหาร ของ โรงเรียน</p>
Nominal dependent - Nominals	18	appos appositional modifier	<p>โยงระหว่าง nominal phrase ที่เท่ากันหรือ coreference สิ่งที่อยู่ในวงเล็บ หรือคำอธิบายอักษรย่อ (สองคำอาจวางสลับตำแหน่งกันได้) โดย appos จะโยงจากซ้ายไปขวาเสมอ</p>

Type		ชื่อ	คำอธิบาย
			 <p>สมชาย ผู้จัดการ โรงแรม แห่ง นี้</p>
Nominal dependent - Nominals	19	nummod numeric modifier	<p>เป็นความสัมพันธ์ที่โยงระหว่างคำนามกับตัวเลข เช่น เงิน (60) < (บาท) กรณีคำขยายบอกปริมาณ เช่น หลาย บาง ทุก แต่ละ ให้วิเคราะห์เป็น advmod ขยายนามที่ปรากฏรวม</p>  <p>ฉัน มี เงิน 60 บาท</p>
Nominal dependent - Clauses	20	acl clausal modifier of noun	<p>คล้ายกับ <u>advcl</u> แต่ในกรณีนี้ใช้เพื่อขยายคำนาม เช่น ผม เจอ (หนังสือ) ที่ คุณ > (ชอบ)</p>  <p>ผม เจอ หนังสือ ที่ คุณ ชอบ สุด ๆ</p>

การประมวลผลภาษาไทย

Type		ชื่อ	คำอธิบาย
		(adjectival clause)	
Nominal dependent - Modifier words	21	amod adjectival modifier	<p>โยงระหว่างคำคุณศัพท์ที่มาขยายคำนาม เช่น (เนื้อ) >(ดิบ)</p> <p>งาน นี้ คือ เนื้อ ดิบ</p>
Nominal dependent - Function Words	22	det determiner	<p>โยงคำบ่งชี้เฉพาะหรือไม่เฉพาะที่ขยายคำนาม เช่น (หนังสือ) >(นี้), หนังสือ (กอง) >(นี้) โดยทั่วไปจะโยงคำนามเข้ากับคำที่มี POS เป็น DET</p>
Nominal dependent - Function Words	23	clf classifier	<p>ใน UD v2 ใช้โยงตัวเลขบอกจำนวนไปยังคำลักษณนามโดยยกตัวอย่างภาษาจีน three CLF student และวิเคราะห์เป็น three <-clf>CLF, three<-nummod-student ส่วน this CLF bus วิเคราะห์เป็น this-clf->CLF this<-det-bus แต่ในกรณีที่มีการละคำนามหลักก็ให้ CLF เป็นคำหลัก three CLF วิเคราะห์เป็น three<-nummod-CLF ซึ่งดูไม่</p>

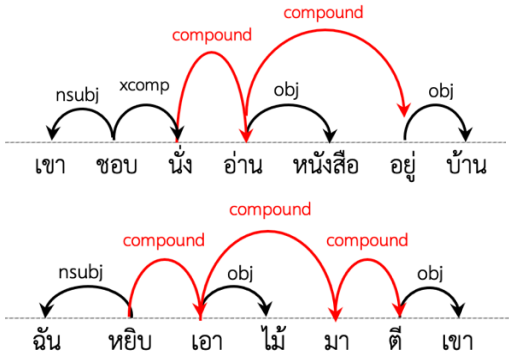
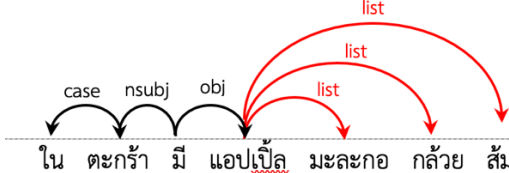
Type		ชื่อ	คำอธิบาย
Nominal dependent - Function Words			<p>คงที่ และตำราไวยากรณ์ไทยมักวิเคราะห์คำลักษณะนามเป็นคำนามประเภทหนึ่ง จึงควรวิเคราะห์ให้เลขจำนวนนับสัมพันธ์กับคำลักษณะนามมากกว่าคำนามหลัก เช่น <i>นักเรียนสามคน</i> จะโยงความสัมพันธ์เป็น (<i>นักเรียน</i>)-<i>clf</i>->(<i>คน</i>), (<i>สาม</i>)<-<i>nummod</i>-(<i>คน</i>) , <i>รถคันนี้</i> จะวิเคราะห์เป็น (<i>รถ</i>)-<i>clf</i>->(<i>คัน</i>), (<i>คัน</i>)-<i>det</i>->(<i>นี้</i>) และเมื่อมีการละคำนามหลัก <i>สามเล่มนี้</i> ก็ยังวิเคราะห์แบบเดิมได้ (<i>สาม</i>)<-<i>nummod</i>-(<i>เล่ม</i>), (<i>เล่ม</i>)-<i>det</i>->(<i>นี้</i>)</p>  <p>หนังสือ สาม เล่ม วาง อยู่ ใน ห้อง ของ เธอ</p>
	24	case case marking	<p>โยงระหว่างคำนามกับบุพบท (เมื่อไทรที่ข้างหน้าเป็น SCONJ ความสัมพันธ์จะเป็น mark แต่ถ้าข้างหน้าเป็น ADP ความสัมพันธ์จะเป็น case)</p>

การประมวลผลภาษาไทย

Type		ชื่อ	คำอธิบาย
Coordination	25	conj conjunct	<p>โยงระหว่างสองคำที่ coordinate กัน เช่น เขา (เขียน) จดหมาย และ >(อ่าน) หนังสือ</p>
	26	cc coordinating conjunction	<p>โยงระหว่างคำที่ coordinate เป็นคำสุดท้ายกับคำสันธาน</p>
MWE	27	fixed fixed multiword expression	<p>ใช้กรณีคำที่เป็น multiword ที่จะต้องรวมเป็นคำเดียว เช่น ถึงกระนั้นก็ตาม = ถึง > กระนั้น, ถึง >ก็, ถึง >ตาม</p>

Type		ชื่อ	คำอธิบาย
MWE	28	flat flat multiword expression	<p>flat ใช้กับ exocentric MWE คล้ายกับ compound แต่ flat ใช้กับ exocentric MWE (ในขณะที่ compound เป็นลักษณะแบบ endocentric) ใช้กับชื่อคน (โยงชื่อต้น ไปหาชื่อกลาง, ชื่อต้น ไปหา นามสกุล) โยงวัน-เดือน-ปี หรือคำภาษาต่างประเทศ โดยโยงจากซ้ายไปขวา เช่น (1) >(มกราคม) >(2017), (นาย) >(ชวน), (นาย) <(สมชาย), (สมชาย) >(กล้าหาญ) (ภาษาอังกฤษ (Mr.) >(Somchai) แต่ภาษาไทย ชื่อ น่าจะเป็นคำหลักมากกว่า ยกเว้นกรณี นาย >ชื่อ ประเทศ >ชื่อ)</p>
MWE	29	compound compound	<p>ใช้โยงคำในกลุ่มคำประสม เช่น สมุด>โทรศัพท์ รวมถึงคำกริยาเรียง ใน UD v2 ก็จะทำให้ใช้ความสัมพันธ์นี้ด้วย วิธีโยงความสัมพันธ์คือโยงต่อกันไป เช่น หม้อ>หุงข้าว หม้อ>หุง หุง>ข้าว</p>

การประมวลผลภาษาไทย

Type		ชื่อ	คำอธิบาย
			
Loose	30	list list	<p>ใช้โยงคำในรายการ โดยต้องมีจำนวนมากกว่า 2 รายการ ตัวตั้งต้นเป็นรายการแรกเสมอ</p> 

Type	ชื่อ	คำอธิบาย
Loose	31 parataxis parataxis	ความสัมพันธ์แบบต่อกันโดยไม่มีกรเชื่อม เช่น (Washington) (>(CNN)) โดยปกติจะโยงจากซ้ายไปขวา ใช้โยงเข้ากับคำหรือข้อความที่อยู่ในวงเล็บ
Special	32 orphan orphan	<p>ความสัมพันธ์ของ dependent ที่ไม่เข้าตามมาตรฐานที่กล่าวมา ใช้กรณีที่มีการละ ทำให้เกิดความสัมพันธ์ของคู่คำที่ปกติไม่เกิด เช่น มานี้ ได้ เหยียดทอง ส่วน (ซูใจ) > (เหยียดเงิน) ในตัวอย่างนี้มีการละกริยาในประโยคส่วนที่สอง ปกติความโยง ได้/VERB เข้ากับกริยาหลักของประโยคส่วนที่ 2 แต่เมื่อเกิดการละเช่นนี้ จะ promote ให้ ซูใจ เป็น head ของประโยคส่วนที่สองโดยที่ยังคงมี POS เป็น PROPJ แล้วโยงความสัมพันธ์ระหว่าง ซูใจ กับ เหยียดเงิน เป็น orphan</p>

การประมวลผลภาษาไทย

Type		ชื่อ	คำอธิบาย
Special	33	goeswith goes with	ใช้โยงความสัมพันธ์ของคำที่อื่นที่จริงต้องเขียนติดกัน แต่อาจจะเขียนผิดหรือแก้ไขผิดพลาดจนทำให้ตัวหนังสือแยกจากกัน เช่น <i>with out, never the less</i> ในภาษาไทย อาจไม่พบกรณีเช่นนี้มากนัก อย่างไรก็ตาม หากเกิดกรณีตัดคำมาผิด เช่น <i>เกื้อ-กุล</i> และไม่สามารถแก้ไขการตัดคำได้ ให้โยง (<i>เกื้อ</i>) > (<i>กุล</i>) เข้าหากันโดยใช้ความสัมพันธ์ โยงจากซ้ายไปขวา
Special	34	reparandum overridden disfluency	ใช้กับถ้อยคำที่พูดผิดแล้วแกสว่นใหม่ที่ถูกมาแทน เช่น <i>ไป ทาง (ชาย) < - ทาง (ขวา)</i> แก้คำพูดเปลี่ยนจาก <i>ซ้าย</i> เป็น <i>ขวา</i>
Other	35	punct punctuation	โยงระหว่าง predicate กับเครื่องหมายวรรคตอน
Other	36	root root	โยง root เขากับคำที่เป็น head หลักของประโยคหรือส่วนนั้น
Other	37	dep	ใช้เมื่อไม่สามารถระบุว่าเป็นความสัมพันธ์แบบใดเลย หากเป็นไปได้ ขอให้หลีกเลี่ยงการโยงรูปแบบความสัมพันธ์

Type		ชื่อ	คำอธิบาย
		unspecified dependency	

การแจกส่วนประโยคตามแนว Universal Dependencies

การแจกส่วนประโยคเป็นการประมวลผลระดับวากยสัมพันธ์ตามแนวทาง Universal Dependencies ที่มุ่งหาคำแต่ละคำมีความเกี่ยวข้องสัมพันธ์กันอย่างไรตามหลักหรือแนวการวิเคราะห์ในทางภาษาศาสตร์ โปรแกรมแจกส่วนประโยคหรือ Parser ที่ใช้ในที่นี่คือโปรแกรม MaltParser ซึ่งพัฒนาโดย Johan Hall, Jens Nilsson และ Joakim Nivre (Hall et al., 2016) MaltParser เป็น inductive dependency paring ที่ใช้คลังข้อมูลต้นไม้พึ่งพา (dependency treebank) ที่ได้มาจากการวิเคราะห์โดยนักภาษาศาสตร์จำนวนหนึ่ง โปรแกรมจะเรียนรู้ความสัมพันธ์ระหว่างคำที่วิเคราะห์ไว้จากข้อมูลและสร้างโมเดลไว้สำหรับใช้แจกส่วนประโยคอื่น ๆ ต่อไป ข้อมูลที่วิเคราะห์ตามแนวทาง Universal Dependencies จะจัดเก็บในรูปแบบของ conllu ตามตัวอย่างข้างล่างนี้

- # sent_id = 7.1
- # text = คนในสังคมส่วนใหญ่ถูกฝึกหรือสอนมาให้ยอมรับในสิ่งที่เป็นอย่าง
- 1 คน _ NOUN NOUN _ 6 nsubj _ SpaceAfter=No
- 2 ใน _ ADP ADP _ 3 case_ SpaceAfter=No
- 3 สังคม _ NOUN NOUN _ 1 obl _ SpaceAfter=No
- 4 ส่วนใหญ่ _ ADJ ADJ _ 1 amod _ SpaceAfter=No
- 5 ถูก _ AUX AUX _ 6 aux _ SpaceAfter=No
- 6 ฝึก _ VERB VERB _ 0 root_ SpaceAfter=No
- 7 หรือ _ CCONJ CCONJ _ 8 conj_ SpaceAfter=No
- 8 สอน _ VERB VERB _ 6 cc _ SpaceAfter=No
- 9 มา _ ADV ADV _ 8 advmod _ SpaceAfter=No
- 10 ให้ _ SCONJ SCONJ _ 11 mark _ SpaceAfter=No
- 11 ยอมรับ _ VERB VERB _ 8 xcomp _ SpaceAfter=No

- 12 ใน _ ADP ADP _ 13 case_ SpaceAfter=No
- 13 สิ่ง _ NOUN NOUN NounType=Class 11 obl _ SpaceAfter=No
- 14 ที่ _ CONJ CONJ _ 15 mark _ SpaceAfter=No
- 15 เป็น _ VERB VERB _ 13 acl _ SpaceAfter=No
- 16 อยู่ _ ADV ADV _ 15 advmod _ SpaceAfter=No

รูปแบบข้อมูล conllu เป็นรูปแบบข้อมูลที่ใช้สำหรับเก็บข้อมูล Dependency treebank ซึ่งเป็นการบ่งบอกว่าคำในประโยคมีความสัมพันธ์กันอย่างไร โดยแต่ละคำ จะถูกกำหนดเป็น Node และมีการเชื่อมโยงกันด้วย Relation ซึ่งแทนความสัมพันธ์ ระหว่างคำนั้นๆในประโยค. ข้อมูลเป็น text แยกแต่ละ field ด้วยแท็บ. Field ไหน ไม่ได้ใช้ก็ได้ _ ไว้ ข้อมูลแต่ละบรรทัดใช้แทนข้อมูลของแต่ละโหนดหรือคำนั้น Field ใน conllu เรียงตามลำดับดังนี้

- ID: ลำดับของ Token หรือคำภายในประโยค เริ่มนับจาก 1
- FORM: รูปคำ
- LEMMA: รูปพจนานุกรมของคำ กรณีภาษาไทยไม่ต้องใช้เพราะเป็นรูปเดียวกับ FORM
- UPOS: Universal Part-of-Speech Tag หมวดคำตามการวิเคราะห์แบบ Universal Dependencies
- XPOS: Language-specific Part-of-Speech Tag หมวดคำเฉพาะของภาษานั้น หรืออาจใส่หมวดคำย่อยที่ละเอียดขึ้น
- FEATS: ลักษณะของคำ เช่น Gender, Number ของคำนาม หรือ Tense, Mood ของกริยา ตัวอย่าง Case=Nom|Number=Plur หมายถึงคำนั้นมีรูปการก nom และพจน์เป็น plural
- HEAD: ลำดับของ Token ที่เป็นคำหลักหรือ Head ที่เชื่อมโยงกับ Token นี้
- DEPREL: ประเภทของความสัมพันธ์ (Relation) ระหว่าง Token กับ Head กรณีวิเคราะห์เป็น tree

การประมวลผลภาษาไทย

- DEPS: ประเภทของความสัมพันธ์ทั้งหมด ในกรณีวิเคราะห์ออกมาเป็นเครือข่ายคือมี head มากกว่าหนึ่งตัวได้ ให้ใส่ทั้ง Head ID กับ relation ทุกอันลงไป ตัวอย่าง 2:nsubj|4:nsubj หมายถึงคำนี้มี head 2 ตัว คือมีคำที่ 2 เป็น head ด้วย deprel “nsubj” และคำที่ 4 เป็น head ด้วย deprel “nsubj”
- MISC: ข้อมูลอื่น ๆ ที่เกี่ยวข้องกับ Token นี้ เช่น การเน้น (Emphasis), หมายเหตุ (Note) เป็นต้น ในตัวอย่างข้อมูลภาษาไทยที่นี้ใช้ SpaceAfter=Yes|No เพื่อบอกว่าหลังคำนั้นมี space ตามมาหรือไม่

การสร้างโมเดลจากข้อมูลทีวีเคราะห์นี้ ใน MaltParser ให้ใช้คำสั่งต่อไปนี้ คือเรียก java โปรแกรมให้ run maltparser-1.9.2. แล้วสร้างโมเดล thamalt (ได้ไฟล์ thamalt.mco) จากไฟล์ข้อมูลทีวีเคราะห์ตามแนว UD ไว้คือ train.conllu ในไฟล์นี้แต่ละตัวอย่างต้องเว้นหนึ่งบรรทัดพอดีและไม่มีข้อผิดพลาดในการกำกับข้อมูลอยู่ไฟล์ที่ได้ thamalt.mco คือโมเดล Thai parser ที่จะนำไปใช้ได้

- >java -jar maltparser-1.9.2.jar -c thamalt -i train.conllu -m learn

การใช้โมเดลที่สร้างขึ้นใน Python ต้องอาศัยโปรแกรมเชื่อมต่อที่เขียนไว้ใน NLTK เมื่อ import malt มาแล้วเรียก malt.MaltParser โดยระบุ path ที่ติดตั้ง maltparser-1.9.2.jar. และ ระบุไฟล์โมเดลที่จะใช้ จากนั้นเรียกใช้ mp.parse_tagged_sents ซึ่งรับข้อมูลเข้าเป็น list ของ list ที่มีลำดับ (word,pos) ซึ่งอาจได้มาจากการใช้โมดูล nltk.nlp.pos_tag ผลที่ได้ออกมาเป็น dependency graph

- from nltk.parse import malt
- mp = malt.MaltParser(parser_dirname="/path/to/MaltParser", model_filename='thamalt.mco')
- parsed_trees = mp.parse_tagged_sents([[(word1,pos1),(word2,pos2), ...]])
- dg = next(next(parsed_trees))

คำสั่งทั้งหมดนี้ผู้เขียนนำมาใช้และสร้างเป็นโมดูล `tltk.nlp.MaltParser` เพื่อความสะดวกในการใช้งาน โดยรับข้อมูลเข้าเป็นข้อความภาษาไทย และสร้าง output เป็น dict {'sentence': "ข้อความภาษาไทย", 'words': [{'id': nn, 'pos': POS, 'deprel': REL, 'head': HD_ID}, {...}, ...]}. ซึ่งสามารถใช้คำสั่ง `tltk.nlp.print_dtree` เพื่อแสดงผลออกมาให้เห็น. แต่หากข้อมูลเข้าเป็น text ที่ตัดคำแล้ว ให้เตรียมข้อมูลเข้าเป็น list ของคำแล้วเรียกใช้ `MaltParser_wordlist` แทน

- `dt = tltk.nlp.MaltParser("เขานั่งดูหนังอยู่ที่บ้าน")`
- `tltk.nlp.print_dtree(dt)`

สถิติ

โปรแกรมแจกส่วนประกอบใน TLTK ใช้โปรแกรม `MaltParser` ที่เขียนด้วยภาษา Java. TLTK ใช้ส่วนเชื่อมต่อที่ทำไว้ใน NLTK ทั้งนี้ก่อนจะเรียกใช้ `tltk.nlp.MaltParser`. จะต้องติดตั้งโปรแกรม `MaltParser` ก่อนด้วยการดาวน์โหลดโปรแกรมที่เป็น .tar.gz. แล้วกระจายโปรแกรมออกมา จากนั้นกำหนดตัวแปร `tltk.nlp.Maltparser_Path` ไปยังที่ที่ติดตั้งโปรแกรม `MaltParser` ไว้ โปรแกรมจะทำงานโดยอาศัยโมเดลที่สร้างไว้ ซึ่งใน TLTK ใช้โมเดลที่ผ่านการฝึกจากข้อมูลต้นไม้ฟังกา 1,000 ประโยค ผลที่ได้เป็น {'sentence': "ข้อความภาษาไทย", 'words': [{'id': nn, 'pos': POS, 'deprel': REL, 'head': HD_ID}, {...}, ...]}. ซึ่งเรียกดูได้ผ่าน `tltk.nlp.print_dtree`

สามารถทดลองใช้โปรแกรมได้ที่

https://colab.research.google.com/drive/1YCrhKqQ9_YEVXgs-_HYXv4n5N73LdzF2Y?usp=sharing

```
!wget http://maltparser.org/dist/maltparser-1.9.2.tar.gz
!tar -zxvf maltparser-1.9.2.tar.gz
import tltk
from nltk.parse import malt
```

การประมวลผลภาษาไทย

```
ltk.nlp.Maltparser_Path = "/content/maltparser-1.9.2"  
dt = ltk.nlp.MaltParser("เขานั่งดูหนังอยู่ที่บ้าน")  
ltk.nlp.print_dtrees(dt)
```

- 1:---เขา (PRON, nsubj - 2)
- 2:--นั่ง (VERB, ROOT - 0)
- 3:---ดู (VERB, compound - 2)
- 4:-----หนัง (NOUN, obj - 3)
- 5:-----อยู่ (VERB, compound - 3)
- 6:-----ที่ (ADP, case - 7)
- 7:-----บ้าน (NOUN, obl - 5)

อ้างอิง

Hall, J., J. Nivre and J. Nilsson (2006). Discriminative Classifiers for Deterministic Dependency Parsing. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 316-323.

Hudson, R. (1984). Word Grammar. New York : Basils Blackwell

MaltParser - a data-driven dependency parser. (2018, February 18).

Retrieved from <https://www.maltparser.org>

Starosta, Stanley. (1988). The case for lexicase: an outline of lexicase grammatical theory. Open Linguistics Series, ed. by Robin Fawcett. London: Pinter Publishers Limited.

Universal Dependencies. (2023, April 18). Retrieved from
<https://universaldependencies.org>

การแปลงภาษาเป็นเวกเตอร์¹⁴

การประมวลผลภาษาที่กล่าวมาในบทก่อน ๆ เป็นการใช้อัลกอริทึมวิเคราะห์ข้อมูลภาษาโดยใช้แนวคิดดั้งเดิมในทางภาษาศาสตร์ คือมองภาษาในระดับต่างๆ ตั้งแต่ระดับคำขึ้นไปเป็นหน่วยทางวากยสัมพันธ์ และมองหาความหมายจากโครงสร้างประโยคที่ได้ การตัดหน่วยทางภาษาก็เป็นหน่วยที่มีความหมายทางภาษา ไม่ว่าจะเป็น พยางค์ คำ วลี อนุภาคฯลฯ สิ่งที่เราวิเคราะห์ได้เป็นข้อมูลทางภาษาศาสตร์ เช่น หมวดคำ ประเภทของชื่อเฉพาะ ความสัมพันธ์แบบพ้องต่าง ๆ และใช้ข้อมูลเหล่านี้เป็นตัวแทนของภาษา เพื่อนำมาศึกษาวิเคราะห์ด้านอื่นต่อไป แต่ในบางกรณี เราก็ใช้วิธีการสังเคราะห์เพื่อสร้างข้อมูลที่ใช้เป็นตัวแทนของภาษาขึ้นมาได้ เมื่อเทียบกับข้อมูลจากการวิเคราะห์ที่ผ่านมา ข้อมูลจากการวิเคราะห์เป็นสิ่งที่นักภาษาเข้าใจว่าคืออะไร หมายถึงอะไร ทำให้ไม่วิเคราะห์ห่อออกมาในลักษณะนั้น แต่ข้อมูลการสังเคราะห์เป็นการแทนที่ข้อมูลเกี่ยวกับภาษาด้วยชุดข้อมูลที่เป็นตัวเลขเพื่อนำมาใช้คำนวณทางสถิติ ตัวอย่างเช่น งานของ Biber (1992) ที่สนใจวิเคราะห์ประเภทของตัวบท. ที่ผ่านมานักภาษาก็พอบอกได้ว่า ตัวบทต่าง ๆ มีการจัดประเภทย่อยต่างๆ ได้ เพราะมีลักษณะการใช้ภาษาที่ต่างกัน เช่น นิยาย บทความวิชาการ บันทึกประจำวัน ฯลฯ ซึ่งก็เป็นการอธิบายความต่างจากการวิเคราะห์เชิงคุณภาพเป็นหลัก. แต่ Biber ต้องการใช้อัลกอริทึมเชิงปริมาณที่จะใช้แสดงความต่างออกมาให้ชัดเจน จึงได้ใช้รายการของ features หรือลักษณะทางภาษามาเป็น

¹⁴ เรียบเรียงและแก้ไขเพิ่มเติมจากบทความ “ข้อมูลภาษากับสถิติ 1 : ข้อมูลภาษา” และ “ข้อมูลภาษากับสถิติ 3 : การใช้ข้อมูล” ใน <https://awrote.medium.com/>

ข้อมูลแทนแต่ละตัวบท. แนวคิดคือถ้าเรารู้ว่าความต่างระหว่างตัวบทแต่ละประเภทมาจากลักษณะทางภาษาที่ต่างกัน เราก็หาว่า ตัวบทนั้น ๆ มีลักษณะอะไรบ้างออกมาก่อน แล้วจึงใช้วิธีการทาง multivariate statistics อย่าง factor analysis มาใช้เพื่อจัดว่า ตัวบทโดยอยู่ในกลุ่มหรือประเภทเดียวกันจากการวิเคราะห์ลักษณะเหล่านั้น. Biber (1988, 1992) เรียกวิธีการที่เขาใช้ว่า การวิเคราะห์พหุมิติ (multi-dimensional analysis) เพราะใช้วิธีการลดทอนจำนวนลักษณะมากมายที่พบในตัวบทนั้นมาเป็นข้อมูลที่มีจำนวนมิติหรือ dimension ไม่กี่ด้าน. หลักคือลักษณะที่มีกัมมันต์แนวโน้มว่ามักพบหรือปรากฏร่วมกันคือเป็นไปในทิศทางเดียวกันจะถูกรวบหรือลดทอนให้ไปอยู่ในมิติเดียวกันได้.

การลดทอนมิติข้อมูลหรือ dimension reduction เป็นวิธีการหนึ่งทางสถิติที่นำมาใช้ในข้อมูลภาษา. หลักการคือการทำให้ข้อมูลที่เคยเดิมมีจำนวนมิติหรือตัวแปรมากให้เหลือจำนวนมิติที่น้อยลงโดยยังคงลักษณะของข้อมูลเดิมไว้. ข้อมูลภาษาที่รวบรวมมาอาจเป็นลักษณะหรือ feature ต่างๆ ที่พบในข้อมูลหรือ text นั้นซึ่งอาจมีหลายร้อย feature. แต่ละ text จึงมองได้ว่าเป็นเวกเตอร์ที่มีหลายร้อยมิติ ตัวแปรหรือลักษณะที่เกี่ยวข้องหรือสัมพันธ์กันจะถูกรวบไปในทิศทางเดียวกันในมิติที่ลดทอนลงนั้นได้. วิธีการที่รู้จักกันคือการใช้ principle component analysis (PCA) และ factor analysis.

PCA เป็นการกำหนดให้ข้อมูลเหลือมิติสำคัญจำนวนหนึ่งเรียกว่า component. วิธีการนี้ทำให้เราสามารถแปลงตัวแปรเดิมที่มีไปเป็นตัวแปรใหม่หรือ component ได้เป็นการสร้างตัวแปรใหม่ตามหลักคณิตศาสตร์ให้ C หรือ component คำนวณจากตัวแปรเดิมในสัดส่วนต่างๆ ($C = w_1(Y1) + w_2(Y2) + w_3(Y3) + w_4(Y4)$) คือ คิดว่า component ที่สร้างใหม่มาจากตัวแปรเดิมแต่ละตัวในสัดส่วนเท่าไร. ส่วน factor analysis จะลดมิติในอีกแบบ คือ หา factor ซึ่งเป็นตัวแปรซ่อนเร้น (latent variable) ที่มาจากตัวแปรเดิมแต่ละตัว ($Y1 = b_1 * F + u_1, Y2 = b_2 * F + u_2, Y3 = b_3 * F + u_3, Y4$

การประมวลผลภาษาไทย

= $b_d * F + u_d$) คือคำว่าตัวแปรเดิมแต่ละตัว (Y_1, \dots, Y_4) มีผลมากน้อยกับตัวแปรใหม่ (F) แค่ไหน (Grace-Martin, 2021).

Biber (1988) ใช้ factor analysis ในการวิเคราะห์ลดทอนมิติที่มาจากลักษณะทางภาษาจำนวนมากให้เหลือ 7 มิติ. แต่ละมิติสะท้อนภาพการใช้ภาษาในลักษณะต่างๆ เช่น Dimension1 หรือมิติที่ 1 นี้ถูกวิเคราะห์ว่าแสดงถึงลักษณะ involved and information production เพราะมองเห็นว่าลักษณะที่เป็นเรื่องของภาษาพูดให้คะแนนบวกในมิตินี้ เช่น การใช้ private verb, THAT deletion, second person pronoun, demonstrative pronoun, etc. ในขณะที่ลักษณะที่เป็นเรื่องภาษาที่ให้ข้อมูลทางการจะได้คะแนนลบ เช่น การใช้ noun, word length, type/token ratio, preposition etc. การลดทอนมิติโดยใช้ factor analysis นี้ทำให้เห็นภาพด้วยบทประเภทต่าง ๆ ได้ชัดเจน text ที่เป็นการสนทนามีคะแนนมิติที่ 1 เป็นบวกสูง แต่ text ที่เป็นเอกสารทางการมีคะแนนติดลบในมิติที่ 1 นี้

แม้ว่า factor analysis และ PCA เป็นเทคนิคที่ใช้เพื่อลดทอนมิติเหมือนกัน แต่ factor analysis มองหาตัวแปรซ่อนเร้นที่มาจากตัวแปรหรือในกรณีงานของ Biber คือลักษณะทางภาษา ตัวแปรใหม่พิจารณาจาก co-occurrence ของลักษณะต่างๆ นี่จึงเหมาะสมกับการใช้งานเพราะตัวแปรใหม่หรือ factor เป็นสิ่งที่มีความหมายอิงจากตัวแปรเดิม ต่างจาก PCA ที่ตัวแปรใหม่มุ่งที่การ maximize variance จากตัวแปรเดิม อาจสะท้อนหรือไม่สะท้อนสิ่งที่ซ่อนเร้นในตัวแปรเดิมนั้นก็ได้ (Korstanje, 2021)

ด้วยวิธีการลดมิติข้อมูลนี้ ลักษณะทางภาษาที่วิเคราะห์ได้จากแต่ละตัวบทยจะถูกนับและนำมาสังเคราะห์เป็นชุดข้อมูลตัวเลขหรือเวกเตอร์เพื่อใช้แทนแต่ละตัวบท แล้วนำมาใช้คำนวณในทางสถิติต่อไป. เมื่อสามารถแปลงแต่ละตัวบทออกมาเป็นตัวเลขในแต่ละมิติได้. ตัวบทที่จัดอยู่ในประเภทเดียวกันก็จะมีค่าในมิติต่าง ๆ ใกล้เคียงกัน

ตัวอย่างการวิเคราะห์ข้อมูลเริ่มจากการใช้โปรแกรม tagger เพื่อหาลักษณะทางภาษาของคำต่าง ๆ ตามตัวอย่างข้างล่างนี้

... I_FPP1 ask_VPRTP [SUAV] you_SPP2 . . ‘ ’ ’ Aziz_NN joined_VBD in_PIN [STPR] . . oeWhy_NN talk_NN about_IN the_DT English_NN ? . Brrrr_NN ... : !

จากนั้นจึงคำนวณสรุปลักษณะทั้งหมดที่พบในแต่ละตัวบทได้ผลออกมาเป็นตารางของตัวบท-ลักษณะตามตัวอย่างที่เห็นข้างล่าง แล้วนำไปใช้วิธีการทางสถิติคือ factor analysis เพื่อลดทอนลักษณะทั้งหมดให้เหลือมิติตามการวิเคราะห์ของ Biber จึงสามารถนำผลที่ได้ของตัวบทนั้นไปเทียบกับภาพรวมของตัวบทในเอกสารประเภท (genre) ต่าง ๆ ตามที่ Biber ได้เคยวิเคราะห์ไว้แล้วได้.

Filename	Tokens	AWL	TTR	AMP	ANDC	(BEMA)	(BYPA)	CAUS	(CONC)	(COND)	CONJ	(CONT)	DEMO	DEMP	DPAR	DWNT
A Passage to India	111408	4.36	238	0.18	1.56	2.08	0.07	0.11	0.09	0.25	0.12	0	0.71	0.41	0.06	0.26
A room with a view	79225	4.32	218	0.27	1.46	2.1	0.07	0.08	0.1	0.28	0.1	0	0.76	0.33	0.06	0.29
ALEXANDRIA	69294	4.56	235	0.08	1.32	1.75	0.14	0.04	0.08	0.09	0.2	0	0.52	0.21	0.05	0.22
Howard's End	111349	4.24	234	0.26	1.9	2.35	0.06	0.08	0.12	0.32	0.1	1.5	0.69	0.51	0.14	0.29
Pharos and Pharillon	25715	4.66	233	0.17	1.73	1.68	0.12	0.06	0.13	0.19	0.2	0	0.73	0.24	0.02	0.36
The Celestial Omnibus and ot	35440	4.23	222	0.22	2	1.88	0.06	0.05	0.1	0.38	0.08	0.58	0.73	0.4	0.14	0.34
THE LONGEST JOURNEY	105804	4.24	186	0.18	1.65	2.23	0.05	0.08	0.09	0.31	0.12	0	0.68	0.33	0.05	0.3
The Story of the Siren	6484	4.52	233	0.11	1.33	1.36	0.06	0.09	0.06	0.48	0.12	0.12	1.11	0.26	0.05	0.2
Where Angels Fear to Tread	53827	4.29	241	0.23	1.69	2.3	0.07	0.06	0.08	0.32	0.12	0	0.66	0.3	0.07	0.26
EM Forster	66260.67	4.38	226.67	0.19	1.63	1.97	0.08	0.07	0.09	0.29	0.13	0.24	0.73	0.33	0.07	0.28

รูปข้างกลาง เป็นการเทียบข้อมูลหนังสือแต่ละเล่มของ Jane Austen ว่าเมื่อมองผ่านมิติทั้งหกตามแบบ Biber แล้วมีค่าในแต่ละมิติเป็นอะไรบ้าง และเทียบได้ใกล้เคียงกับ text ประเภทใด

Filename	Dimension1	Dimension2	Dimension3	Dimension4	Dimension5	Dimension6	Closest Text Type
A Passage to India	-2.31	5.45	1.14	-0.1	-0.34	-0.91	General narrative exposition
ALEXANDRIA	-13.67	-0.43	2.09	-4.61	0.35	-1.6	General narrative exposition
Pharos and Pharillon	-8.05	2.42	3.34	-1.47	0.81	-0.72	General narrative exposition
The Story of the Siren	-5.35	1.03	3.78	1.12	1.23	-0.19	General narrative exposition
A room with a view	-0.7	5.39	0.57	0.59	-1.04	-0.38	Imaginative narrative
Howard's End	6.31	6.45	-0.32	1.6	-0.98	-0.57	Imaginative narrative
The Celestial Omnibus and other Stories	2.39	4.81	0.29	0.81	-0.51	-0.62	Imaginative narrative
THE LONGEST JOURNEY	1.09	5.8	-0.01	0.19	-1.07	-0.64	Imaginative narrative
Where Angels Fear to Tread	-0.74	5.49	0.57	0.95	-1.26	-0.72	Imaginative narrative
EM Forster	-2.34	4.05	1.27	-0.1	-0.31	-0.71	General narrative exposition

ในบางงานเช่น Topic modeling (Blei, 2012) จะใช้วิธีแปลงเอกสารต่างๆ เป็นตารางของคำ-เอกสาร (term-document) คือเป็นตารางที่มีค่าแสดงถึงค่าต่าง ๆ ในแต่ละเอกสารซึ่งคำที่นำมาใช้จะเป็นคำเนื้อหาเป็นหลัก. เมื่อได้ตารางนี้แล้ว แต่ละเอกสารก็จะถูกมองแทนด้วยรายการตัวเลขค่าเหล่านี้ได้. แล้วจึงใช้วิธีการทางสถิติอย่าง LDA (Latent Dirichlet Allocation) เพื่อจับ คำที่ควรอยู่ในกลุ่มเดียวกันจาก

การประมวลผลภาษาไทย

ความสัมพันธ์ที่พบใน ตารางคำ-เอกสาร (term-document matrix) นี้ต่อไป. ลักษณะนี้คล้ายกับงานข้างต้นที่เป็นการสร้างเวกเตอร์หรือชุดตัวเลขขึ้นมาเพื่อใช้แทนตัวบทนั้น แล้วจึงใช้ประโยชน์จากเวกเตอร์นั้นต่อ โดยถือเสมือนว่าเวกเตอร์นั้นเป็นตัวแทนตัวบทนั้น ๆ ได้

Topic model ใช้หลักคิดว่าในแต่ละเอกสารจะกล่าวถึงหัวข้อหรือ topic ไม่กี่หัวข้อ ดูได้จากคำเนื้อหาที่ใช้ในเอกสารนั้น. การคำนวณว่าคำอะไรบ้างบ่งบอกหัวข้อเดียวกันก็อาศัยหลักการหาการปรากฏร่วมกัน (co-occurrence) ของคำในเอกสารนั้น. ตัวอย่างข้างล่างเป็นการทดลองใช้ topic model กำหนดหา 10 หัวข้อ และคำที่จัดกลุ่มอยู่ในแต่ละหัวข้อของข้อมูลทั้งหมดที่นำมาวิเคราะห์.

Topic Id	Top Words...
0	year continued percent demand rates strong remained growth high activity recent consumer higher increases prices costs conditions energy labor increased
1	good made time found party subject general attention love world leave interest business reason told home means mr added longer
2	copyright work information law confidence rights author owner public works trade confidential case person defendant breach patent act property employee
3	year percent 1 1999 market 2 1998 rate growth prices 4 2000 debt federal markets years inflation financial 3 treasury
4	anne captain elliot mr wentworth lady mrs charles russell walter mary musgrove louisa father bath gutenberg elizabeth house good room
5	sales district contacts reported prices report activity manufacturing construction demand markets retail reports districts estate business noted ago retailers inventories
6	people government states public great congress laws american country policy national united law executive citizens business power free constitution duty
7	elinor marianne mrs sister edward dashwood jennings willoughby lucy mother john house brandon mr lady ferrars thing room colonel dear
8	mr elizabeth darcy bennet bingley jane sister wickham lady collins mrs lydia dear room catherine family father replied gardiner lizzy
9	miss mrs great man time make day long give young felt hope project thought place sir happy feelings morning heard

ตัวอย่างคำที่บ่งบอกถึงหัวข้อต่าง ๆ ในข้อมูลวิเคราะห์

filename	optopics...								
BB_BC2.txt	2	0.806996463437103	0.16061320349401100	0	0.015145978215738000	5	0.00974329715080344	6	0.003734155263535890
Econ6.txt	3	0.5965397322865260	0.391918051825889	1	0.007027327129259760	6	0.002548473008773560	2	0.0019436225116708000
Econ7.txt	3	0.6908900212130620	0.3085912801094730	1	4.70105269252456E-04	6	1.57042580559723E-05	2	1.30285545291466E-05
1905-Roosevelt.txt	6	0.6508799702202350	0.2933608982980200	0	0.0545227251872866700	3	3.81738613608582E-04	2	3.36770888321389E-04
BB_BC1.txt	2	0.7926932918167440	0.17111045995347800	6	0.021559171785926100	0	0.008179109285288000	5	0.006373120732213730
Econ5.txt	3	0.7437672595582840	0.2483722569441920	1	0.007337135372716690	9	3.61706213917504E-04	6	1.34649232856251E-04
BB_CP_4.txt	2	0.9477577461229400	0.04617720232596720	0	0.005702123713782210	6	1.13235958331542E-04	3	1.06486541290905E-04
Econ4.txt	3	0.7685291320769690	0.22020760886652800	1	0.008899487685432120	6	0.0010396395393199000	5	0.0010332564325196100
BB_CP5.txt	2	0.8756007449431220	0.054251963970546600	1	0.035601590887535800	3	0.021806595599753600	0	0.009998420322755110
Econ1.txt	5	0.5028376924383000	0.463665789906853900	3	0.03203478585161020	1	0.001419247438997000	6	1.65193280269402E-05
1885-Cleveland.txt	6	0.7508374751196970	0.24275015877972000	0	0.0030418499421860300	3	0.0029560122463429400	2	1.64199646237039E-04
Econ3.txt	5	0.545699652710040	0.41728874111286200	3	0.03693756560050280	1	3.18748167054125E-05	6	1.6386734725327E-05
BB_CP_2.txt	2	0.9405725106994920	0.05866393603323430	0	2.29123636766241E-04	6	1.66745688444741E-04	3	1.56806827851363E-04
BB_CP_3.txt	2	0.9567670928608650	0.017230887900093600	9	0.011848903024072300	0	0.008653068712473060	3	0.005333848983274480
Econ2.txt	5	0.5239153283587960	0.45716977428361700	3	0.01860557039077050	1	1.49418249510817E-04	2	1.30779123961238E-04
1909-Taft.txt	6	0.7341620128851750	0.15514649740218800	2	0.03249973175469370	3	0.031154936550909900	9	0.021211852231899900
1889-Harrison.txt	6	0.6831733027886630	0.2364396579699649	5	0.025807664824632800	2	0.02363890142083900	9	0.017581844867660900
Sense%20and%20Sensibility	9	0.4750705745684250	0.30356070553115100	1	0.22132851998944400	2	2.85709945192258E-05	0	4.66471226207232E-06
BB_IP_1.txt	2	0.9205618602847980	0.05200520398815660	6	0.022811744053060600	0	0.004397549034909640	3	9.53772754342795E-05
Persuasion.txt	9	0.4714918101748630	0.3228900098167010	1	0.20038478293720700	6	0.0039620983310852700	0	0.0012596480049843100
Pride%20and%20Prejudice.t	9	0.4639452705779520	0.3246177025590070	1	0.21035864703820000	6	6.03845308463168E-04	2	4.64895727707565E-04
1901-McKinley.txt	6	0.7573308301849860	0.21953668158659100	0	0.011974075500512700	2	0.009751939768082600	3	0.0012131867728868400
1893-Cleveland.txt	6	0.7521003314817210	0.2211401082610970	5	0.0091438960970724000	9	0.009138975280790110	3	0.005814377370637550
1897-McKinley.txt	6	0.6946143285309210	0.24923499644326500	3	0.0341590155881501000	0	0.02180623270822770	2	7.345524639972E-05

ตัวอย่างแสดงผล document และ topic ที่มีน้ำหนักมากในแต่ละ document

งานด้าน stylometry ที่เป็นการวิเคราะห์หาผู้แต่ง (authorship) ของงานเขียน คือหาคำตอบว่างานที่ส่งสัยนั้นเป็นงานที่เขียนโดยใคร ก็สามารถใช้อข้อมูลเชิงสังเคราะห์มาใช้ได้. ตัวอย่างงานของ Binongo (2003) ต้องการหาคำตอบว่าหนังสือพอมดมัทศจรรยแห่ง Oz เล่มที่ 15 นั้นใครเป็นผู้แต่งกันและระหว่างนักเขียนสองคน. เขาหาคำตอบโดยดูชุดคำไวยากรณ์ที่พบมาก 50 คำเป็นตัวแทนของหนังสือ คือ ใช้คำไวยากรณ์ 50 คำเพื่อสร้างเวกเตอร์แทนหนังสือแต่ละเล่มของนักเขียนทั้งสองคน. จากนั้นใช้สถิติ principal component analysis เพื่อลดทอนข้อมูล 50 มิติให้เหลือ 2 มิติ แล้วจึงเห็นว่าผลที่ได้จากหนังสือของนักเขียนสองคนนี้แตกต่างกัน. และเมื่อนำหนังสือที่ส่งสัยมาสังเคราะห์ข้อมูลชุดคำไวยากรณ์นำมาเปรียบเทียบกับผลที่ได้ของนักเขียนทั้งสองก็จะเห็นว่าใครควรเป็นผู้เขียนหนังสือเล่มนี้กัน

การประมวลผลภาษาไทย

การแปลงข้อมูลภาษาเป็นเวกเตอร์จึงเป็นอีกวิธีการหนึ่งของการประมวลผลภาษา การใช้ตาราง คำ-เอกสาร เป็นการนำเอามิติความสัมพันธ์ของคำกับเอกสารมาใช้เพื่อประมวลผลเกี่ยวกับเอกสารนั้น เช่น การหาคำที่อยู่ในหัวข้อเดียวกัน ความใกล้เคียงหรือคล้ายคลึงกันของเอกสาร ที่ใช้มองหาเอกสารที่จัดอยู่ในประเภทเดียวกันได้เพราะมีการใช้ลักษณะทางภาษาต่าง ๆ คล้ายกัน หรือที่ใช้มองหาความเป็นผู้แต่งของเอกสารได้เพราะดูลักษณะการเขียนเฉพาะตัวที่สะท้อนผ่านการใช้คำไวยากรณ์ และลีลาการเขียนได้ วิธีการแปลงข้อมูลภาษาเป็นเวกเตอร์ตัวเลขนี้อย่างทำในระดับย่อยกว่าตัวบทคือทำในระดับข้อความหรือระดับคำก็ได้ ก็จะได้เวกเตอร์ที่เป็นตัวแทนของสิ่งนั้นได้ วิธีการแปลงคำเป็นเวกเตอร์วิธีการหนึ่งคือ word2vec ที่เสนอโดย Mikolov et al. (2013) เป็นวิธีการหนึ่งที่ใช้กันแพร่หลาย ซึ่งจะกล่าวถึงต่อไปในบทถัดไป

แต่การแปลงเอกสารเป็นเวกเตอร์ที่ไม่ได้ใช้วิธีการทำตารางคำ-เอกสาร คือใช้ข้อมูลลักษณะทางภาษาแบบที่ Biber ทำนั้น ในภาษาไทยยังไม่มีโปรแกรมที่ทำด้านนี้ได้โดยตรง ผู้เขียนจึงพัฒนาโมดูล TextAna ใน TLTk เพื่อใช้แก้ลักษณะต่าง ๆ ทางภาษาในเอกสารนั้น แล้วลักษณะที่ได้เหล่านี้สามารถเลือกนำมาใช้เพื่อแปลงเอกสารเป็นเวกเตอร์เพื่อการศึกษา. รายการลักษณะข้างล่างนี้คือข้อมูลที่ได้จากการวิเคราะห์ตัวบท

กลุ่มข้อมูลพื้นฐาน เป็นข้อมูลเชิงพรรณนา (descriptive) เพื่อบอกจำนวนลักษณะประเภทต่าง ๆ ที่พบในตัวบท โดยพิจารณาว่าน่าจะสัมพันธ์กับความยากง่ายของตัวบท เช่น ความยาวตัวบท จำนวนย่อหน้า จำนวนการเว้นวรรค จำนวนหน่วยปริจเฉท จำนวนคำ จำนวนศัพท์ ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของลักษณะต่าง ๆ

- DesSpC : จำนวนเว้นวรรคที่พบในตัวบท
- DesChaC : จำนวนตัวอักษรที่พบในตัวบท
- DesSymbC : จำนวนสัญลักษณ์และตัวอักษรพิเศษที่พบในตัวบท
- DesPC : จำนวนย่อหน้าที่พบในตัวบท

- DesEduC : จำนวนหน่วยปริจเฉทพื้นฐานหรือ EDU ที่พบในตัวบท (นับเมื่อเลือก option เป็น edu ไม่ใช่ paragraph และได้จากการใช้โมดูล segment ที่ตัดหา EDU ใน TLTK)
- DesTotW : จำนวนคำทั้งหมดในตัวบท
- DesTotT : จำนวนรูปศัพท์ในตัวบท
- DesEduL : ค่าเฉลี่ยความยาวของ EDU (จำนวนคำ)
- DesEduLd : ค่า SD ของความยาว EDU
- DesWrdL : ค่าเฉลี่ยความยาวของคำ (จำนวนพยางค์)
- DesWrdLd : ค่า SD ของความยาว
- DesPL : ค่าเฉลี่ยความยาวของย่อหน้า (จำนวนคำ)
- DesCorrToneC : จำนวนครั้งที่คำมีเสียงวรรณยุกต์ตรงตามเครื่องหมายวรรณยุกต์ (กรณีคำมีหลายพยางค์และมีรูปตรงกับเสียงมากกว่าหนึ่งครั้งจะถ่วงน้ำหนักตามจำนวน)
- DesInCorrToneC : จำนวนครั้งที่คำมีเสียงวรรณยุกต์ไม่ตรงตามเครื่องหมายวรรณยุกต์ (กรณีคำมีหลายพยางค์และมีรูปไม่ตรงกับเสียงมากกว่าหนึ่งครั้งจะถ่วงน้ำหนักตามจำนวน)
- DesCorrFinalC : จำนวนครั้งที่คำมีพยัญชนะสะกดเขียนตรงตามมาตราสะกด คือ กก กต กก กน กม เกย เกอว
- DesInCorrFinalC : จำนวนครั้งที่คำมีพยัญชนะสะกดเขียนไม่ตรงตามมาตราสะกด
- DesClusterC : จำนวนครั้งที่คำมีเสียงพยัญชนะควบกล้ำ
- DesLeadC : จำนวนครั้งที่คำมีการอ่านแบบอักษรนำ เช่น ฉลาด โสร่ง
- DesDoubVowelC : จำนวนครั้งที่คำเสียงสระผสม
- DesTNCt1C : จำนวนคำที่อยู่ใน TNC tier1 หรือ top 50%
- DesTNCt2C : จำนวนคำที่อยู่ใน TNC tier2 หรือ top 51-60%
- DesTNCt3C : จำนวนคำที่อยู่ใน TNC tier3 หรือ top 61-70%
- DesTNCt4C : จำนวนคำที่อยู่ใน TNC tier2 หรือ top 71-80%
- DesTTC1 : จำนวนคำที่อยู่ในรายการศัพท์ของ TTC ระดับประถมต้น
- DesTTC2 : จำนวนคำที่อยู่ในรายการศัพท์ของ TTC ระดับประถมปลาย (ไม่รวมคำที่พบใน TTC1)
- DesTTC3 : จำนวนคำที่อยู่ในรายการศัพท์ของ TTC ระดับมัธยมต้น (ไม่รวมคำที่พบใน TTC1 หรือ TTC2)

การประมวลผลภาษาไทย

- DesTTC4 : จำนวนคำที่อยู่ในรายการศัพท์ของ TTC ระดับมัธยมปลาย (ไม่รวมคำที่พบใน TTC1 หรือ TTC2 หรือ TTC3)

กลุ่มข้อมูลลักษณะเกี่ยวกับคำที่ปรับค่ามาตรฐาน เป็นข้อมูลจำนวนนับข้างต้นแต่ปรับค่าเป็นมาตรฐานโดยอาศัยจำนวนคำในตัวบทเป็นตัวหาร เพื่อให้สามารถเปรียบเทียบค่าลักษณะเหล่านี้ระหว่างตัวบทได้

- WrdCorrTone : สัดส่วนจำนวนครั้งที่คำมีเสียงวรรณยุกต์ตรงตามรูปต่อจำนวนคำทั้งหมด
- WrdInCorrTone : สัดส่วนจำนวนครั้งที่คำมีเสียงวรรณยุกต์ไม่ตรงตามรูปต่อจำนวนคำทั้งหมด
- WrdCorrFinal : สัดส่วนจำนวนครั้งที่คำมีรูปพยัญชนะสะกดตรงตามมาตราต่อจำนวนคำทั้งหมด
- WrdInCorrFinal : สัดส่วนจำนวนครั้งที่คำมีรูปพยัญชนะสะกดตรงตามมาตราต่อจำนวนคำทั้งหมด
- WrdKaran : สัดส่วนจำนวนครั้งที่คำมีรูปการันต์ต่อจำนวนคำทั้งหมด
- WrdCluster : สัดส่วนจำนวนครั้งที่คำมีพยัญชนะต้นเป็นพยัญชนะควบกล้ำต่อจำนวนคำทั้งหมด
- WrdLead : สัดส่วนจำนวนครั้งที่คำมีการใช้อักษรคำต่อจำนวนคำทั้งหมด
- WrdDoubVowel : สัดส่วนจำนวนครั้งที่คำมีสระผสมต่อจำนวนคำทั้งหมด
- WrdNEL : สัดส่วนจำนวนครั้งที่พบชื่อสถานที่ต่อจำนวนคำทั้งหมด
- WrdNEo : สัดส่วนจำนวนครั้งที่พบชื่อองค์กรต่อจำนวนคำทั้งหมด
- WrdNEp : สัดส่วนจำนวนครั้งที่พบชื่อบุคคลต่อจำนวนคำทั้งหมด
- WrdNeg : สัดส่วนจำนวนคำปฏิเสธต่อจำนวนคำทั้งหมด
- WrdTNCt1 : สัดส่วนจำนวนคำใน TNC tier 1 x 1000 / จำนวนคำทั้งหมด (ปรับค่าที่พบเป็น / 1000 คำ)
- WrdTNCt2 : สัดส่วนจำนวนคำใน TNC tier 2 x 1000 / จำนวนคำทั้งหมด
- WrdTNCt3 : สัดส่วนจำนวนคำใน TNC tier 3 x 1000 / จำนวนคำทั้งหมด
- WrdTNCt4 : สัดส่วนจำนวนคำใน TNC tier 4 x 1000 / จำนวนคำทั้งหมด
- WrdTTC1 : สัดส่วนจำนวนคำใน TTC ระดับประถมต้น x 1000 / จำนวนคำทั้งหมด

- WrdTTC2 : สัดส่วนจำนวนคำใน TTC ระดับประถมปลาย x 1000 / จำนวนคำทั้งหมด
- WrdTTC3 : สัดส่วนจำนวนคำใน TTC ระดับมัธยมต้น x 1000 / จำนวนคำทั้งหมด
- WrdTTC4 : สัดส่วนจำนวนคำใน TTC ระดับมัธยมปลาย x 1000 / จำนวนคำทั้งหมด
- WrdC : สัดส่วนจำนวนคำเนื้อหาใน TTC x 1000 / จำนวนคำทั้งหมด
- WrdF : สัดส่วนจำนวนคำไวยากรณ์ใน TTC x 1000 / จำนวนคำทั้งหมด
- WrdCF : สัดส่วนจำนวนคำที่เป็นได้ทั้งคำเนื้อหาและคำไวยากรณ์ใน TTC x 1000 / จำนวนคำทั้งหมด
- WrdFrmSing : สัดส่วนจำนวนคำมูลใน TTC x 1000 / จำนวนคำทั้งหมด
- WrdFrmComp : สัดส่วนจำนวนคำประสมใน TTC x 1000 / จำนวนคำทั้งหมด
- WrdFrmTran : สัดส่วนจำนวนคำทับศัพท์ใน TTC x 1000 / จำนวนคำทั้งหมด
- WrdSemSimp : สัดส่วนจำนวนคำเดี่ยวใน TTC x 1000 / จำนวนคำทั้งหมด
- WrdSemTran : สัดส่วนจำนวนคำประสมที่เป็นคำใสใน TTC x 1000 / จำนวนคำทั้งหมด
- WrdSemSemi : สัดส่วนจำนวนคำประสมกึ่งทับกึ่งใสใน TTC x 1000 / จำนวนคำทั้งหมด
- WrdSemOpaq : สัดส่วนจำนวนคำประสมที่เป็นคำทึบใน TTC x 1000 / จำนวนคำทั้งหมด
- WrdBaseM : สัดส่วนจำนวนคำที่อยู่ในรายการคำพื้นฐานของกระทรวงศึกษา x 1000 / จำนวนคำทั้งหมด
- WrdBaseT : สัดส่วนจำนวนคำที่อยู่ในรายการคำพื้นฐานจาก TNC & TTC x 1000 / จำนวนคำทั้งหมด
- WrdTfidf : ค่าเฉลี่ยของ TF-IDF ของคำที่พบในตัวบท (TF-IDF ของคำได้จาก TNC)
- WrdTncDisp : ค่าเฉลี่ยของ dispersion ของคำที่พบในตัวบท (dispersion ของคำได้จาก TNC)
- WrdTtcDisp : ค่าเฉลี่ยของ dispersion ของคำที่พบในตัวบท (dispersion ของคำได้จาก TTC)
- WrdArf : ค่าเฉลี่ยของ arf หรือ average reduced frequency ของคำที่พบในตัวบท (arf ของคำได้จาก TNC)

การประมวลผลภาษาไทย

- WrdNOUN : สัดส่วนจำนวนคำที่มี POS เป็น NOUN x 1000 / จำนวนคำทั้งหมด
- WrdVERB : สัดส่วนจำนวนคำที่มี POS เป็น VERB x 1000 / จำนวนคำทั้งหมด
- WrdADV : สัดส่วนจำนวนคำที่มี POS เป็น ADV x 1000 / จำนวนคำทั้งหมด
- WrdDET : สัดส่วนจำนวนคำที่มี POS เป็น DET x 1000 / จำนวนคำทั้งหมด
- WrdADJ : สัดส่วนจำนวนคำที่มี POS เป็น ADJ x 1000 / จำนวนคำทั้งหมด
- WrdADP : สัดส่วนจำนวนคำที่มี POS เป็น ADP x 1000 / จำนวนคำทั้งหมด
- WrdPUNCT : สัดส่วนจำนวนคำที่มี POS เป็น PUNCT x 1000 / จำนวนคำทั้งหมด
- WrdAUX : สัดส่วนจำนวนคำที่มี POS เป็น AUX x 1000 / จำนวนคำทั้งหมด
- WrdSYM : สัดส่วนจำนวนคำที่มี POS เป็น SYM x 1000 / จำนวนคำทั้งหมด
- WrdINTJ : สัดส่วนจำนวนคำที่มี POS เป็น INTJ x 1000 / จำนวนคำทั้งหมด
- WrdCCONJ : สัดส่วนจำนวนคำที่มี POS เป็น CCONJ x 1000 / จำนวนคำทั้งหมด
- WrdPROPN : สัดส่วนจำนวนคำที่มี POS เป็น PROPN x 1000 / จำนวนคำทั้งหมด
- WrdNUM : สัดส่วนจำนวนคำที่มี POS เป็น NUM x 1000 / จำนวนคำทั้งหมด
- WrdPART : สัดส่วนจำนวนคำที่มี POS เป็น PART x 1000 / จำนวนคำทั้งหมด
- WrdPRON : สัดส่วนจำนวนคำที่มี POS เป็น PRON x 1000 / จำนวนคำทั้งหมด
- WrdSCONJ : สัดส่วนจำนวนคำที่มี POS เป็น SCONJ x 1000 / จำนวนคำทั้งหมด

กลุ่มข้อมูลลักษณะเกี่ยวกับการเชื่อมโยงความ ณ ขณะนี้โมดูลสามารถหาได้สี่ลักษณะ แต่ในอนาคตจะเพิ่มจำนวนลักษณะอื่น ๆ เช่น การใช้คำเชื่อมแยกตามประเภทต่าง ๆ

- CrfCWL : สัดส่วนของคำเนื้อหาที่ทับซ้อนระหว่างส่วนของข้อความ โดยนับเพียงว่ามีหรือไม่มีคำเหมือนกันระหว่างส่วนของข้อความ
- CrfCNL : สัดส่วนของคำนามที่ทับซ้อนระหว่างส่วนของข้อความ โดยนับเพียงว่ามีหรือไม่มีคำนามที่เหมือนกันระหว่างส่วนของข้อความ
- CrfCVL : สัดส่วนของคำกริยาที่ทับซ้อนระหว่างส่วนของข้อความ โดยนับเพียงว่ามีหรือไม่มีคำกริยาที่เหมือนกันระหว่างส่วนของข้อความ
- CrfCTL : สัดส่วนของคำเนื้อหาที่ทับซ้อนระหว่างส่วนของข้อความ โดยนับตามจำนวนคำที่ทับซ้อนระหว่างส่วนของข้อความ

กลุ่มข้อมูลลักษณะที่เกี่ยวกับความหลากหลายของศัพท์ คือ การเทียบหาจำนวนศัพท์และจำนวนคำที่พบในตัวบท

- LdvTTR : สัดส่วนของ type / token ของคำ

-

กลุ่มข้อมูลลักษณะที่เกี่ยวกับวากยสัมพันธ์

- SynDepth ค่าเฉลี่ยความลึกของ dependency tree ในตัวบท
- wrd_deprel: ข้อมูล dictionary wrd_deprel[deprel] ที่บอกความถี่ของ deprel ที่พบในตัวบท

นอกจากลักษณะต่าง ๆ นี้ ข้อมูลที่ได้จาก TextAna ยังมี

- wrd: ข้อมูล dictionary wrd[w] ที่บอกความถี่ของคำ w
- wrd_arf: ข้อมูล dictionary wrd_arf[w] ที่บอกค่า average reduced frequency ของคำ w เมื่อเรียกใช้โมดูล nltk.nlp.TextAna โดยให้ข้อมูลเข้าเป็น text. จะได้ผลมาเป็น

สถิติ

โมดูล TextAna ใน TLTK รับข้อมูลเป็น text ถ้าข้อมูลเข้ามีการแยกคำด้วย | แล้ว และแยกย่อหน้าด้วย “<p>” ให้ใช้ TextOption=“segmented” และ WordOption=“segmented”, หากเป็น plain text ให้เลือก WordOption เป็น “colloc” หรือ “mm” ตามวิธีที่ต้องการตัดคำ TextOption ใช้กำหนดหน่วยการวิเคราะห์ default เป็น “par” คือประมวลผลทีละย่อหน้า หากเลือก “edu” โปรแกรมจะตัด EDU ก่อนและประมวลผลทีละ EDU, option UDParse=“Malt” ใช้เมื่อต้องการให้วิเคราะห์ d-tree ด้วย, ผลที่วิเคราะห์ได้ทั้งหมดเป็น dict ที่มี feature ต่าง ๆ ตามข้างต้น. เมื่อใช้ json.dumps ก็จะได้ผลเป็น json string

การประมวลผลภาษาไทย

หากต้องการได้ผลออกมาเป็น list ให้ใช้ `tttk.nlp.txt2feat(text)`. จะได้ list ของ value ทั้งหมดที่วิเคราะห์ได้ออกมา. หากต้องการ list ของชื่อ feature. ให้ใช้ `tttk.nlp.txt2feat(text,Option="name")` สามารถทดลองใช้โปรแกรมได้ที่ https://colab.research.google.com/drive/1_YCrhKqQ9_YEVXgs-_HYXv4n5N73LdzF2Y?usp=sharing

```
import tttk
import json
from nltk.parse import malt

tttk.nlp.Maltparser_Path = "/content/maltparser-1.9.2"

text1 = "ตำรวจแยกสอบรายบุคคล 4 ผู้ต้องหาแก๊งค์ม็วรีตจินเทา ขณะที่ ทนายยัน "น้ำเพชร เพชรลดา" ยังไม่ได้สารภาพ และไม่มีส่วนเกี่ยวข้องกับ เตรียมรวบรวมพยานหลักฐานต่อสู้คดี ส่วนเรื่องอื่นขอประกันตัว ยังไม่ได้รับการหารือกับความ คืบหน้ากรณี น้ำเพชร เพชรลดา อดีตผู้เข้าประกวดมิสแกรนด์สุพรรณบุรี ปี 2020 ร่วมแพนหนุมานวงแหวนอัมรินทร์เงิน เกือบ 4 ล้านกเมื่อเวลา 15.40 น. วันที่ 18 เมษายน 2566 ที่ สน.ลุมพินี เจ้าหน้าที่ตำรวจ สน.ลุมพินี ได้เบิกตัว นายธนายุทธ พลนาคู หรือ โจ นายณัฐพงษ์ ช้องคำ หรือ จูน นายศุภพล ฮาลมัน หรือ กาย หรือ บอส และ น.ส.สรีดา ดอนม่วง หรือ ตา เพื่อนำตัวไปสอบปากคำที่ห้องพนักงานสอบสวน ชั้น 2 โดยมีทนายความ ซึ่งผู้ต้องหาทั้งหมดได้ร้องขอไว้ ทั้งนี้ การสอบปากคำผู้ต้องหาทั้ง 4 คน เป็นการแยกสอบรายบุคคล ระหว่างที่เบิกตัวเข้าห้องพนักงานสอบสวน สื่อมวลชน พยายามสอบถามนายศุภพล ถึงเหตุการณ์ที่เกิดขึ้น แต่นายกายไม่ได้โต้ตอบแต่อย่างใด โดยมีสีหน้านิ่งเฉย ก่อนเดินเข้า ห้องพนักงานสอบสวนไปบางส่วน น.ส.เพชรลดา แสงกุล หรือ น้ำเพชร พนักงานสอบสวนยังไม่ได้เบิกตัวออกมาสอบปากคำ แต่อย่างใด โดยพนักงานสอบสวนระบุว่า จะทำการสอบปากคำผู้ต้องหาทั้ง 4 คน ให้แล้วเสร็จก่อน จะเบิกตัว น.ส.น้ำเพชร มาสอบปากคำต่อไปจากต่อมาเวลา 16.30 น. ทนายความกับเพื่อนๆ น.ส.เพชรลดา เดินทางเข้าเยี่ยมในห้องควบคุม ผู้ต้องหา สน.ลุมพินี ซึ่งใช้เวลาเยี่ยมประมาณ 10 นาที จากนั้นให้สัมภาษณ์ว่า สภาพจิตใจและกำลังใจยังคงดี กินได้นอนหลับ ไม่มีความเครียด ส่วนของคดีความเป็นหน้าที่ของตำรวจในการทำงาน ซึ่งทางตัวน้ำเพชรเอง ก็ยืนยันว่าไม่รู้เรื่องอะไร ในคดีนี้ โดยทนายความก็เตรียมรวบรวมพยานหลักฐานต่อสู้คดี"

res = tttk.nlp.TextAna(text1,UDParse="Malt")
print(json.dumps(res, indent=4, ensure_ascii=False))

{
```

```

"DesChC": 1426,
"wrд": {
  "ตำรวจ": 2,
  "แยก": 2,
  "สอบ": 2,
...
  "นี่": 1,
  "ที่นี้": 1
},
"wrд_arf": {
  "ตำรวจ": 1.1898305084745762,
  "แยก": 1.9559322033898305,
  "สอบ": 1.9559322033898305,
...
  "นี่": 1.0,
  "ที่นี้": 1.0
},
"deprel": {
  "nsubj": 12,
  "ROOT": 5,
  "obj": 35,
  "nmod": 66,
  "nummod": 11,
  "clf": 1,
  "compound": 23,
  "flat": 9,
  "mark": 24,
  "advcl": 10,
  "punct": 3,
  "aux": 11,
  "cc": 14,
  "conj": 20,
  "advmod": 16,

```

การประมวลผลภาษาไทย

```
"act": 14,  
"obl": 7,  
"list": 2,  
"ccomp": 2,  
"det": 2,  
"case": 8,  
"xcomp": 1,  
"amod": 1  
},  
"DesPC": 0,  
"DesSpC": 3,  
"DesEduC": 3,  
"SynDepth": 13.5,  
"DesNEIC": 2,  
"DesNEoC": 2,  
"DesNEpC": 11,  
"WrdNOUN": 431.095406360424,  
"DesTNct4C": 38,  
"DesTTC1": 221,  
"WrdC": 533.5689045936396,  
"WrdSemSimp": 667.8445229681979,  
"WrdFrmSing": 660.7773851590106,  
"WrdBasicM": 614.8409893992933,  
"WrdBasicT": 614.8409893992933,  
"DesCorrToneC": 231,  
"DesInCorrToneC": 186,  
"DesCorrFinalC": 142,  
"DesInCorrFinalC": 69,  
"DesKaranC": 4,  
"DesClusterC": 3,  
"DesLeadC": 31,  
"DesDoubVowelC": 40,  
"DesTotW": 283,
```

"WrdVERB": 233.21554770318022,
"DesTNct3C": 18,
"DesSymbC": 92,
"WrdPUNCT": 318.02120141342755,
"WrdNUM": 38.869257950530034,
"WrdSemTran": 123.67491166077738,
"WrdFrmComp": 155.47703180212014,
"DesTTC2": 4,
"WrdFrmTran": 7.067137809187279,
"WrdPROPn": 67.13780918727915,
"DesTNct2C": 25,
"WrdSCONJ": 81.2720848056537,
"WrdF": 98.93992932862191,
"WrdAUX": 38.869257950530034,
"DesTNct1C": 103,
"WrdCF": 169.6113074204947,
"DesNegC": 7,
"WrdCCONJ": 49.46996466431096,
"WrdPART": 14.134275618374557,
"CrfCWL": 1.0,
"CrfCTL": 8.75,
"CrfCNL": 0.75,
"DesTTC4": 1,
"WrdSemOpaq": 3.5335689045936394,
"WrdADV": 45.936395759717314,
"DesTTC3": 7,
"WrdDET": 7.067137809187279,
"WrdADP": 28.268551236749115,
"WrdSemSemi": 7.067137809187279,
"CrfCVL": 1,
"WrdADJ": 3.5335689045936394,
"WrdPRON": 3.5335689045936394,
"WrdArf": 1.3550256208119826,

การประมวลผลภาษาไทย

```
"WrdTfidf": 14525.007684007527,  
"WrdTncDisp": 0.3216114247539986,  
"WrdTtcDisp": 0.40009223268990063,  
"DesTotT": 172,  
"DesEduL": 78.4,  
"DesEduLd": 50.44105470745036,  
"DesWrdL": 1.4734982332155477,  
"DesWrdLd": 0.7776834958244379,  
"DesPL": 392,  
"LdvTTR": 607.773851590106,  
"WrdCorrTone": 816.2544169611307,  
"WrdInCorrTone": 657.243816254417,  
"WrdCorrFinal": 501.7667844522968,  
"WrdInCorrFinal": 243.81625441696113,  
"WrdKaran": 14.134275618374557,  
"WrdCluster": 10.60070671378092,  
"WrdLead": 109.54063604240282,  
"WrdDoubVowel": 141.3427561837456,  
"WrdNEL": 7.067137809187279,  
"WrdNEo": 7.067137809187279,  
"WrdNEp": 38.869257950530034,  
"WrdNeg": 24.73498233215548,  
"WrdTNCt1": 363.9575971731449,  
"WrdTNCt2": 88.33922261484099,  
"WrdTNCt3": 63.60424028268551,  
"WrdTNCt4": 134.2756183745583,  
"WrdTTC1": 780.9187279151944,  
"WrdTTC2": 14.134275618374557,  
"WrdTTC3": 24.73498233215548,  
"WrdTTC4": 3.5335689045936394,  
"WrdSYM": 0.0,  
"WrdINTJ": 0.0  
}
```



```
import nltk
import json
from nltk.parse import malt

nltk.nlp.Maltparser_Path = "/content/maltparser-1.9.2"
```

```
text1 = 'ตำรวจแยกสอบรายบุคคล 4 ผู้ต้องหาแก๊งอุ้มรีดจิ้นเทา ขณะที่ยานยนต์น้ำเพชร เพชรลดา' ยังไม่ได้
สภาพและไม่มีส่วนเกี่ยวข้อง เตรียมรวบรวมพยานหลักฐานต่อสู้คดี ส่วนเรื่องยื่นขอประกันตัว ยังไม่ได้รับการหารือ
ความคืบหน้ากรณี น้ำเพชร เพชรลดา อดีตผู้เข้าประกวดมิสแกรนด์สุพรรณบุรี ปี 2020 ร่วมแฟนหนุ่มวางแผนอุ้มจิ้นรีด
เงินเกือบ 4 ล้านบาทเมื่อเวลา 15.40 น. วันที่ 18 เมษายน 2566 ที่ สน.ลุมพินี เจ้าหน้าที่ตำรวจ สน.ลุมพินี ได้เปิดตัว นาย
ธนายุทธ พลนาคู หรือ โจ นายณัฐพงษ์ มั่งคั่ง หรือ จูน นายสุกพล ฮาลมัน หรือ กาย หรือ บอส และ น.ส.สรिता ดอน
ม่วง หรือ ตา เพื่อนำตัวไปสอบปากคำที่ห้องพนักงานสอบสวน ชั้น 2 โดยมีนายความ ซึ่งผู้ต้องหาทั้งหมดได้ร้องขอไว้
ทั้งนี้การสอบปากคำผู้ต้องหาทั้ง 4 คน เป็นการแยกสอบรายบุคคล ระหว่างที่เปิดตัวเจ้าหน้าที่พนักงานสอบสวน สื่อมวลชน
พยายามสอบถามนายสุกพล ถึงเหตุการณ์ที่เกิดขึ้น แต่นายกายไม่ได้โต้ตอบแต่อย่างใด โดยมีสีหน้านิ่งเฉย ก่อนเดินเข้า
ห้องพนักงานสอบสวนไปก่อน น.ส.เพชรลดา แสงกุล หรือ น้ำเพชร พนักงานสอบสวนยังไม่ได้เปิดตัวออกมาสอบปากคำ
แต่อย่างใด โดยพนักงานสอบสวนระบุว่า จะทำการสอบปากคำผู้ต้องหาทั้ง 4 คน ให้แล้วเสร็จก่อน จะเปิดตัว น.ส.น้ำเพชร
มาสอบปากคำต่อไปอีกต่อเวลา 16.30 น. นายความกับเพื่อนๆ น.ส.เพชรลดา เดินทางเข้าเยี่ยมในห้องควบคุม
ผู้ต้องหา สน.ลุมพินี ซึ่งใช้เวลาเยี่ยมประมาณ 10 นาที จากนั้นให้สัมภาษณ์ว่า สภาพจิตใจและกำลังใจยังคงดี กินได้นอน
หลับ ไม่มีความเครียด ส่วนของคดีความเป็นหน้าที่ของตำรวจในการทำงาน ซึ่งทางตัวน้ำเพชรเอง ก็ยืนยันว่าไม่รู้เรื่องอะไร
ในคดีนี้ โดยทีมนายความก็เตรียมรวบรวมพยานหลักฐานต่อสู้คดี'
```

```
lst = nltk.nlp.txt2feat(text1)
print(lst)
hd = nltk.nlp.txt2feat(text1,Option="name")
print(hd)
```

```
[0.75, 8.75, 1.0, 1.0, 1426.0, 3.0, 142.0, 231.0, 40.0, 3.0, 78.4, 50.44105470745036, 69.0, 186.0, 4.0, 31.0,
2.0, 2.0, 11.0, 7.0, 0.0, 392.0, 3.0, 92.0, 103.0, 25.0, 18.0, 38.0, 221.0, 4.0, 7.0, 1.0, 172.0, 283.0,
1.4734982332155477, 0.7776834958244379, 607.773851590106, 13.5, 5.0, 0.0, 14.0, 10.0, 16.0, 1.0, 0.0,
11.0, 8.0, 14.0, 2.0, 1.0, 23.0, 20.0, 0.0, 0.0, 0.0, 2.0, 0.0, 0.0, 0.0, 0.0, 9.0, 0.0, 0.0, 2.0, 24.0, 0.0, 66.0,
12.0, 11.0, 35.0, 7.0, 0.0, 0.0, 3.0, 0.0, 0.0, 0.0, 1.0, 3.5335689045936394, 28.268551236749115,
45.936395759717314, 38.869257950530034, 1.3550256208119826, 614.8409893992933,
614.8409893992933, 533.5689045936396, 49.46996466431096, 169.6113074204947,
10.60070671378092, 501.7667844522968, 816.2544169611307, 7.067137809187279,
```

การประมวลผลภาษาไทย

141.3427561837456, 98.93992932862191, 155.47703180212014, 660.7773851590106,
7.067137809187279, 0.0, 243.81625441696113, 657.243816254417, 14.134275618374557,
109.54063604240282, 7.067137809187279, 7.067137809187279, 38.869257950530034,
431.095406360424, 38.869257950530034, 24.73498233215548, 14.134275618374557,
3.5335689045936394, 67.13780918727915, 318.02120141342755, 81.2720848056537, 0.0,
3.5335689045936394, 7.067137809187279, 667.8445229681979, 123.67491166077738,
363.9575971731449, 88.33922261484099, 63.60424028268551, 134.2756183745583,
780.9187279151944, 14.134275618374557, 24.73498233215548, 3.5335689045936394,
14525.007684007527, 0.3216114247539986, 0.40009223268990063, 233.21554770318022]
['CrfCNL', 'CrfCTL', 'CrfCVL', 'CrfCWL', 'DesChC', 'DesClusterC', 'DesCorrFinalC', 'DesCorrToneC',
'DesDoubVowelC', 'DesEduC', 'DesEduL', 'DesEduLd', 'DesInCorrFinalC', 'DesInCorrToneC', 'DesKaranC',
'DesLeadC', 'DesNELC', 'DesNEoC', 'DesNEpC', 'DesNegC', 'DesPC', 'DesPL', 'DesSpC', 'DesSymbC',
'DesTNCt1C', 'DesTNCt2C', 'DesTNCt3C', 'DesTNCt4C', 'DesTTC1', 'DesTTC2', 'DesTTC3', 'DesTTC4',
'DesTotT', 'DesTotW', 'DesWrld', 'DesWrldLd', 'LdvTTR', 'SynDepth', 'UDROOT', 'UDX', 'UDaCl', 'UDadvcl',
'UDadvmmod', 'UDamod', 'UDappos', 'UDaux', 'UDcase', 'UDcc', 'UDccomp', 'UDclF', 'UDcompound',
'UDconj', 'UDcop', 'UDcsubj', 'UDdep', 'UDdet', 'UDdiscourse', 'UDdislocated', 'UDexpl', 'UDfixed', 'UDflat',
'UDgoeswith', 'UDIobj', 'UDlist', 'UDmark', 'UDmwe', 'UDnmod', 'UDnsubj', 'UDnummod', 'UDobj', 'UDobl',
'UDorphan', 'UDparataxis', 'UDPunct', 'UDreparandum', 'UDsconj', 'UDvocate', 'UDxcomp', 'WrdADJ',
'WrdADP', 'WrdADV', 'WrdAUX', 'WrdArf', 'WrdBasicM', 'WrdBasicT', 'WrdC', 'WrdCCONJ', 'WrdCF',
'WrdCluster', 'WrdCorrFinal', 'WrdCorrTone', 'WrdDET', 'WrdDoubVowel', 'WrdF', 'WrdFrmComp',
'WrdFrmSing', 'WrdFrmTran', 'WrdINTJ', 'WrdInCorrFinal', 'WrdInCorrTone', 'WrdKaran', 'WrdLead', 'WrdNEL',
'WrdNEo', 'WrdNEp', 'WrdNOUN', 'WrdNUM', 'WrdNeg', 'WrdPART', 'WrdPRON', 'WrdPROPN', 'WrdPUNCT',
'WrdSCONJ', 'WrdSYM', 'WrdSemOpaq', 'WrdSemSemi', 'WrdSemSimp', 'WrdSemTran', 'WrdTNCt1',
'WrdTNCt2', 'WrdTNCt3', 'WrdTNCt4', 'WrdTTC1', 'WrdTTC2', 'WrdTTC3', 'WrdTTC4', 'WrdTfidF',
'WrdTncDisp', 'WrdTtcDisp', 'WrdVERB']

อ้างอิง

Biber, Douglas (1988). Variation across speech and writing. Cambridge:

Cambridge University Press.

Biber, Douglas. (1992). The Multi-Dimensional Approach to Linguistic

Analyses of Genre Variation: An Overview of Methodology and Findings

- on JSTOR. *Computers and the Humanities*, 26(5–6), 331–345. Retrieved from <https://www.jstor.org/stable/30204629>
- Blei, D. (2012). Probabilistic Topic Models. *Communications of the ACM* 55 (4): 77–84. doi: 10.1145/2133806.2133826 Available at <http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf>.
- Grace-Martin, K. (2021). The Fundamental Difference Between Principal Component Analysis and Factor Analysis. *Analysis Factor*. Retrieved from <https://www.theanalysisfactor.com/the-fundamental-difference-between-principal-component-analysis-and-factor-analysis>
- José Nilo G. Binongo (2003). Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution, *CHANCE*, 16:2, 9–17, DOI: [10.1080/09332480.2003.10554843](https://doi.org/10.1080/09332480.2003.10554843)
- Korstanje, J. (2021). What is the difference between PCA and Factor Analysis? *Medium*. Retrieved from <https://towardsdatascience.com/what-is-the-difference-between-pca-and-factor-analysis-5362ef6fa6f9>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv*, 1301.3781. Retrieved from <https://arxiv.org/abs/1301.3781v3>

Word2Vec ภาษาไทย¹⁵

การแปลงข้อมูลตัวบทเป็นรายการลักษณะทางภาษาในบทที่แล้วเป็นวิธีการหนึ่งในการสร้างข้อมูลเวกเตอร์เพื่อนำมาใช้ประมวลผลภาษา นอกจากนี้วิธีการนี้ การใช้ word2vec ตามแนวทางของ Mikolov et al. (2013) ก็เป็นอีกวิธีหนึ่งที่สามารถใช้แปลงคำให้อยู่ในรูปเวกเตอร์ที่สามารถนำไปใช้กับคอมพิวเตอร์เพื่อการประมวลผลภาษาต่อไปได้ง่ายขึ้น คำถามสำคัญ คือข้อมูลที่ได้จากการแปลงคำเป็นเวกเตอร์ด้วยวิธีการของ Mikolov et al. (2013) นั้นให้ข้อมูลอะไรบ้างเกี่ยวกับคำ ในที่นี้ ได้ทดลองสร้าง word2vec จากข้อมูล Thai National Corpus v.3 ขนาด 33 ล้านคำโดยใช้ gensim และติดตั้งไว้ใน TLTK

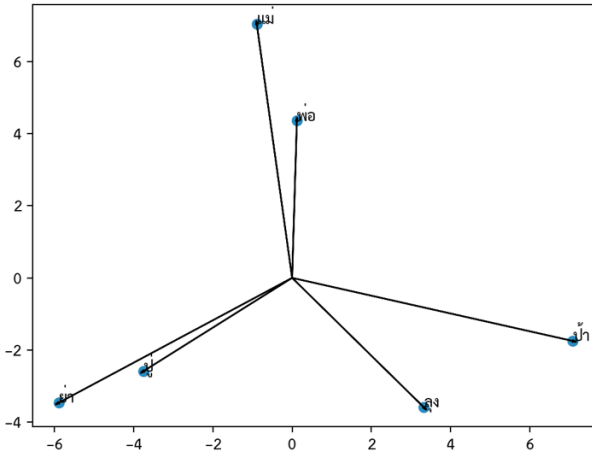
หลักการของ word2vec ในการแปลงคำให้เป็นเวกเตอร์ขึ้นอยู่กับพื้นฐานความคิดของ distributional hypothesis คือ คำที่มีบริบทการปรากฏคล้าย ๆ กันก็ควรเป็นคำที่มีความหมายคล้ายกันด้วย ซึ่งเป็นแนวคิดที่ Harris พุดไว้ตั้งแต่ในปี 1954 (Harris, 1954)

words that occur in the same contexts tend to have similar meanings

Mikolov et al. (2013) ใช้ neural network เรียนรู้ค่าจากการดูคำพร้อมกับบริบทซ้ายขวา หลักการคือ ถ้าให้บริบทรอบข้างมาจำนวนหนึ่งเช่น 5 คำซ้ายขวา จะทำนายได้ใหม่ว่าคำตรงกลางคืออะไร หรือในทำนองตรงข้าม ถ้าให้คำตรงกลางมา จะ

¹⁵ เนื้อหาบทนี้ส่วนหนึ่งนำมาจากบทความที่เผยแพร่ผ่าน Medium “เราเรียนรู้อะไรจาก word2vec” 21 พฤศจิกายน 2018.

บอกได้ใหม่ว่า คำที่มาข้างหน้าข้างหลังเป็นคำอะไรได้บ้าง วิธีการนี้ทำให้คำแต่ละคำมีความเกี่ยวข้องกับบริบท และ word2vec ทำให้เราสามารถดึงกลุ่มคำที่อยู่ใกล้ ๆ กันในเวกเตอร์สเปซออกมาด้วยการหาค่า similarity คำถามที่น่าสนใจคือ เซ็ตของคำคล้ายคลึงที่คำนวณได้ออกมานั้นคืออะไร เราได้ข้อมูลอะไรจากเซตคำเหล่านี้บ้าง



คำความหมายใกล้เคียง

ความสัมพันธ์แบบ near synonym เป็นสิ่งแรกที่เราควรเห็นในเซตของคำคล้ายคลึง เพราะโดยหลักการของ distributional hypothesis คำที่มีความหมายคล้ายกันก็ควรจะมีบริบทการปรากฏคล้าย ๆ กัน คำที่มีความหมายคล้ายกันจึงควรมีค่า similarity สูงด้วย ตัวอย่างข้างล่างก็แสดงให้เห็นว่า คำที่มีความหมายใกล้เคียงกัน (คำที่แสดงด้วยตัวหนา) จะปรากฏในเซตนี้ด้วย n คือจำนวนตัวอย่างที่ต้องการไม่เกิน n , cutoff คือค่าขั้นต่ำของ similarity, $\text{score}='n'$ คือไม่ต้องแสดงค่า similarity ของคำนั้น

```
>>> ttk.corpus.similar_words('สวย',score='n',cutoff=0.6,n=10)
```

```
['น่ารัก', 'เซ็กซี่', 'หล่อ', 'เพ้', 'สะดูดตา', 'เนียน', 'งาม', 'เก๋', 'สวยงาม', 'สดใส']
```

การประมวลผลภาษาไทย

```
>>> tltk.corpus.similar_words('กิน',score='n',cutoff=0.6,n=10)
```

```
['รับประทาน', 'ทาน', 'หุง', 'เคี้ยว', 'ดื่ม', 'คลุก', 'กินน้ำ', 'ดอง', 'ต้ม', 'กินที่']
```

อย่างไรก็ตาม คำที่มีความหมายเหมือนหรือใกล้เคียงกันแต่มีที่ใช้ต่างกัน เช่น กิน และ บริโภค จะไม่ปรากฏเป็นคำคล้ายคลึงซึ่งกันและกัน เหตุผลมาจากเป็นคำคนละทำเนียบภาษา คือ กิน เป็นภาษาไม่เป็นทางการ ส่วน บริโภค เป็นภาษาเชิงวิชาการ โอกาสที่บริบทของคำจะมีร่วมกันจึงน้อยไปด้วย จึงสรุปได้ว่า คำพ้องความหมายหรือมีความหมายใกล้เคียงจะไม่ได้ถูกดึงมาทั้งหมดด้วย word2vec. หากมีบริบทต่างกัน และเรายังเห็นว่าภายในเซตนี้ไม่ใช่คำที่มีความหมายคล้ายกับคำหลักทั้งหมด คำที่เหลือเหล่านี้มีความสัมพันธ์อะไรบางอย่างกับคำหลัก

คำความหมายตรงข้าม

ความสัมพันธ์แบบ antonym ก็ควรเห็นในเซตของคำคล้ายคลึง เพราะคำที่มีความหมายตรงข้ามกันน่าบริบทการเกิดที่คล้ายคลึงกัน ซึ่งจากตัวอย่างก็เห็นคำที่มีความหมายตรงข้ามอยู่ในเซตด้วย น่ารัก เช็ทซี่ หล่อ เป็นคำที่อยู่คนละขั้วกับ ชี๋เหร่

```
>>> tltk.corpus.similar_words('ชี๋เหร่',score='n',cutoff=0.6,n=10))
```

```
['ชีร์ว', 'น่ารัก', 'ปัญญาอ่อน', 'อวดดี', 'เช็ทซี่', 'ใจร้าย', 'ใจดำ', 'หัวแข็ง', 'เจ๋ง', 'หล่อ']
```

```
>>> tltk.corpus.similar_words('แรก',score='n',cutoff=0.6,n=10))
```

```
['สุดท้าย', 'ล่าสุด', 'หนึ่ง', 'ทุก']
```

คำในวงความหมายเดียวกัน

คำที่อยู่ในเซตคำคล้ายคลึงจำเป็นต้องเป็นคำ synonym หรือ antonym เสมอไหม คำตอบคือไม่จำเป็น เนื่องจากการดูที่ distribution หรือบริบทว่าคล้ายคลึงกันไหม คำที่จัดกลุ่มอยู่ในวงความหมายเดียวกัน ก็สามารถมีบริบทคล้ายกันได้ จึงสามารถพบคำลักษณะนี้ เช่น คำเรียกสีต่าง ๆ ในตัวอย่างข้างล่าง หรือชื่อประเทศต่าง ๆ ชื่อเมืองต่าง ๆ ชื่อโรค ชื่อเดือน ลักษณะความสัมพันธ์แบบนี้จะพบเป็นจำนวนมากในข้อมูล

```
>>> tltk.corpus.similar_words('คำ',score='n',cutoff=0.6,n=10)
['ขาว', 'เหลือง', 'แดง', 'น้ำเงิน', 'คล้ำ', 'เทา', 'เขียว', 'ชมพู', 'เข้ม', 'น้ำตาล']
>>> tltk.corpus.similar_words('เงิน',score='n',cutoff=0.6,n=10)
['ญี่ปุ่น', 'อินเดีย', 'พม่า', 'อาหรับ', 'ฝรั่งเศส', 'ไทย', 'ทิเบต', 'ลาว', 'มาเลย์', 'รัสเซีย']
>>>tltk.corpus.similar_words('กัวดาลัมเปอร์',score='n',cutoff=0.6,n=10)
['ฮานอย', 'โคเปนเฮเกน', 'ฮาวานา', 'คาบูล', 'พนมเปญ', 'เบอร์ลิน', 'มอสโก', 'ตริโปลี',
'แบกแดด', 'อิสลามาบัด']
>>> tltk.corpus.similar_words('พิชสุราเรื้อรัง',score='n',cutoff=0.6,n=10)
['ลมบ้าหมู', 'ท้องร่วง', 'เกาต์', 'ซาร์ส', 'ตับแข็ง', 'เท้าช้าง', 'ซาร์ส', 'เก๊าท์', 'ปอดบวม',
'เรื้อรัง']
>>> tltk.corpus.similar_words('มกราคม',score='n',cutoff=0.6,n=10)
['มีนาคม', 'กรกฎาคม', 'พฤศจิกายน', 'กุมภาพันธ์', 'ธันวาคม', 'เมษายน', 'กันยายน',
'สิงหาคม', 'พฤษภาคม', 'มิถุนายน']
```

คำเหล่านี้มีความสัมพันธ์ในเชิงแนวคิดหรือ paradigmatic แต่ไม่ใช่คำที่มีหมวดคำชนิดเดียวกันทั้งหมดจะปรากฏในเซตคำคล้ายคลึง เช่น คำนามแม้จะมี syntactic property เหมือนกัน คือมีคำ ADJ ขยายได้ เกิดกับ NUM เกิดกับ NOUN ที่เป็นลักษณนามได้ แต่ word2vec เป็นการแปลงโดยดูรูปคำในบริบทไม่ใช่หมวดคำในบริบท สิ่งที่ได้จึงเป็นคำใน semantic domain เดียวกันมากกว่าที่จะเป็น syntactic class เดียวกัน

คำไวยากรณ์

กรณีของคำไวยากรณ์ที่มีหน้าที่ต่าง ๆ ก็ไม่น่าแปลกใจว่า รายการคำคล้ายคลึงที่ได้มาก็บอกถึงคำที่เป็นคำไวยากรณ์ทำหน้าที่ลักษณะเดียวกันด้วย เช่น คำสันธาน คำสรรพนาม คำบุพบท คำอนุภาค ตามตัวอย่างนี้

```
>>> tltk.corpus.similar_words('นอกจากนี้',score='n',cutoff=0.6,n=10)
['นอกจากนั้น', 'อย่างไรก็ตาม']
>>> tltk.corpus.similar_words('แต่',score='n',cutoff=0.6,n=10)
```

การประมวลผลภาษาไทย

```
[‘เพราะ’, ‘เพียงแต่’, ‘แต่ว่า’]
```

```
>>> tltk.corpus.similar_words(‘เขา’,score=‘n’,cutoff=0.6,n=10)
```

```
[‘เธอ’, ‘หล่อน’, ‘ผม’, ‘ฉัน’, ‘พวกเขา’, ‘แก’, ‘เรา’, ‘ดิฉัน’, ‘ตัวเอง’, ‘ข้าพเจ้า’]
```

```
>>> tltk.corpus.similar_words(‘ใน’,score=‘n’,cutoff=0.55,n=10)
```

```
[‘ของ’, ‘ภายใน’, ‘สำหรับ’]
```

```
>>> tltk.corpus.similar_words(‘ใหม่’,score=‘n’,cutoff=0.6,n=10)
```

```
[‘มัย’, ‘เธอ’, ‘นะ’, ‘สิ’, ‘หรือ’, ‘ยังง’]
```

```
>>> tltk.corpus.similar_words(‘ยัง’,score=‘n’,cutoff=0.6,n=10)
```

```
[‘ยังคง’]
```

แต่ก็น่าสนใจว่า คำไวยากรณ์บางคำไม่มีคำคล้ายคลึงเมื่อดูค่าความคล้ายที่ 0.6 เช่น คำว่า *จะ เคย กำลัง* ซึ่งน่าจะบอกถึงความแตกต่างของการใช้คำเหล่านี้ที่ไม่ได้มีบริบทคล้ายกับคำไวยากรณ์อื่น ๆ ในกลุ่ม

```
>>> tltk.corpus.similar_words(‘จะ’,score=‘n’,cutoff=0.6,n=10)
```

```
□
```

```
>>> tltk.corpus.similar_words(‘เคย’,score=‘n’,cutoff=0.6,n=10)
```

```
□
```

```
>>> tltk.corpus.similar_words(‘กำลัง’,score=‘n’,cutoff=0.6,n=10)
```

```
□
```

คำปรากฏรวม

นอกจากคำในวงความหมายเดียวกัน คำซึ่งมักปรากฏรวมกัน ก็ควรจะพบในเซตคำคล้ายคลึงด้วยหรือไม่ เพราะคำเหล่านี้ปรากฏอยู่ด้วยกันเป็นปกติ จึงแชร์บริบทรวมกัน เมื่อดูข้อมูลที่ออกมา ก็พบคำที่มีลักษณะเช่นนี้อยู่ เช่น *สวย* มักมีสร้อยคำ *สะดุดตา* ตามมา

```
>>> tltk.corpus.similar_words(‘สวย’,score=‘n’,cutoff=0.6,n=10)
```

```
[‘น่ารัก’, ‘เซ็กซี่’, ‘หล่อ’, ‘เท’, ‘สะดุดตา’, ‘เนียน’, ‘งาม’, ‘เก๋’, ‘สวยงาม’, ‘สดใส’]
```

```
>>> tltk.corpus.collocates(‘สวย’)
```



```
[('แสน', 'สวย'), 30182.793713912175), (('สวย', 'งาม'), 25235.744733331867), (('ช่วย', 'สวย'), 21640.774948224145), (('สวย', 'ซึ้ง'), 18201.05312084762), (('ดวงตา', 'สวย'), 11671.678653589228), (('สวย', 'เร่'), 9725.759845745219), (('สวย', 'Sexy'), 7572.471045699552), (('สวย', 'งามแจ่ม'), 7476.910865842118), (('สาว', 'สวย'), 6213.914696330355)]
```

อย่างไรก็ดี คำปรากฏร่วมที่มีโอกาสพบในเซตคำคล้ายคลึงนี้ ก็ไม่ใช่คำปรากฏร่วมที่หาด้วยวิธีการสถิติแบบที่ใช้ใน `tlk.corpus.collocates` ซึ่งมองหาคำที่มักปรากฏร่วมกับ `สวย` มากเกินปกติ ไม่ได้สนใจว่าทั้งสองคำจะต้องมีคำบริบทคล้าย ๆ กัน คำปรากฏร่วมที่ได้จาก `tlk.corpus.similar_words` ที่จะปรากฏเป็นคำที่คู่กันไปในบริบทซึ่งจะไปเข้าลักษณะการปรากฏที่คล้ายกับคำซ้อนมากกว่า

คำซ้อน

คำที่ปรากฏร่วมกันส่วนหนึ่งเป็นคำที่เป็นคำซ้อนคือมีความหมายทำนองเดียวกันและใช้อยู่ติดกันเป็นคำซ้อน เช่น `ซึ้งซึ้งเห่` คำพวกนี้จึงออกมาในเซตคำคล้ายคลึงด้วย เพราะทั้งคู่จะเกิดร่วมกันจึงมีบริบทเดียวกัน

```
>>> tlk.corpus.similar_words('ซึ้งเห่',score='n',cutoff=0.6,n=10)
['ซึ้งเห่', 'น่ารัก', 'ปัญญาอ่อน', 'อวดดี', 'เซ็กซี่', 'ใจร้าย', 'ใจดำ', 'หัวแข็ง', 'เจ๋ง', 'หล่อ']
```

คำพ้องรูป

คำพ้องรูป คือคำซึ่งมีรูปเหมือนกัน เช่น `ชั้น` เป็นคำนาม หรือเป็นกริยา เมื่อทดลองค้นดูเซตคำคล้ายคลึงจะได้

```
>>> tlk.corpus.similar_words('ชั้น',score='n',cutoff=0.6,n=10)
[]
>>> tlk.corpus.similar_words('ชั้น',score='n',n=10)
['ขนลุกขนพอง', 'พ้อ', 'ขึ้น', 'ขา', 'มะปร่าง', 'เจื้อยแจ้ว', 'จ๊กจ๊ก', 'จอก', 'ฮือ', 'เคี้ยว']
```

การประมวลผลภาษาไทย

จะเห็นว่าไม่มีคำใดที่สัมพันธ์กับ *ชั้น* เมื่อตัดที่ค่าความคล้าย 0.6 เมื่อขยายขอบเขตไปให้ได้คำ 10 คำโดยไม่สนใจค่าความคล้ายที่ 0.6 ก็จะเริ่มพบคำที่สัมพันธ์กับ *ชั้น* (กริยา) คือ *ฆ่า* และ *ชั้น* (คำนาม) คือ *จอก*

หากดูคำพ้องรูปอีกคำ เช่น *ช่าง* ที่เป็นทั้งคำนาม หรือ คำขยายกริยา ข้อมูลเซตคำคล้ายคลึงในความคล้ายที่ 0.6 จะไม่พบคำใดเลย แต่หากขยายขอบเขตไปจะเห็น *ช่าง-ฝีมือ* กับ *ช่าง-ประดิษประดอย* เพราะทั้งสองคำนี้ต่างก็มีการใช้มากพอ

```
>>> tltk.corpus.similar_words('ช่าง',score='n',cutoff=0.6,n=10)
```

```
□
```

```
>>> tltk.corpus.similar_words('ช่าง',score='n',n=10)
```

```
['ฝีมือ', 'ประณีต', 'สถาปนิก', 'เหลือเกิน', 'แต่งหน้า', 'ประดิษประดอย', 'ตัดเย็บ', 'ปั้น', 'นิสัย', 'นายช่าง']
```

ประเด็นที่น่าสนใจในคำพ้องรูปที่ยกมาจะเห็นว่ารายการคำคล้ายที่มีค่า 0.6 นั้น มีจำนวนน้อยมากหรือไม่มีเลย ซึ่งน่าจะมาจากการที่บริบทของคำรูปเดียวกันนี้มีความต่างกันมากจนหาชุดคำที่คล้ายไม่ชัดเจน ข้อเสนอพื้นฐานนี้ต้องไปดูต่อในคำพ้องรูปอื่น ๆ ว่ามีผลในลักษณะนี้เหมือนกันไหม

คำหลายความหมาย

คำที่เป็น polysemy มีหลายความหมาย แต่ละความหมายก็ควรจะมีบริบทต่าง ๆ คำประเภทนี้จะเป็นปัญหาสำหรับการแปลงเป็นเวกเตอร์หรือไม่ เมื่อลองดูคำที่มีหลายความหมาย เช่น ในการวิเคราะห์ความหมายของคำว่า *เก็บ* (Wipharak 2001) พบว่ามี 9 ความหมาย ได้แก่ 'take'-เก็บเอาไว้, 'pick up'-เก็บขึ้นมา, 'arrange'-เก็บที่นอน, 'keep'-เก็บรักษา, 'hide'-เก็บซ่อน, 'gather'-เก็บรวบรวม, 'charge'-เก็บเงิน, 'kill'-สังเก็บ, 'buy'-เก็บหุ้น เมื่อดูรายการเซตคำคล้ายคลึงจะเห็นว่าเป็นแบบเดียวกับคำพ้องรูปคือมีรายการคำที่มีค่าความคล้าย 0.6 น้อย แต่เมื่อขยายขอบเขตออกไปจึง

จะเห็นคำที่แทนความหมายต่าง ๆ นี้อยู่บ้าง เก็บรักษา-'keep', เก็บออม-'save', เอา-'take', กัก-'keep', รวบรวม-'gather', ซื้อ-'buy', จัดเตรียม-'arrange', ตุน-'stock'

```
>>> tltk.corpus.similar_words('เก็บ',score='n',cutoff=0.6,n=10)
```

```
['เก็บรักษา', 'เก็บออม']
```

```
>>> tltk.corpus.similar_words('เก็บ',score='n',n=10)
```

```
['เก็บรักษา', 'เก็บออม', 'เอา', 'กัก', 'รวบรวม', 'ทิ้ง', 'ซื้อ', 'จัดเตรียม', 'ตุน', 'ฝาก']
```

ตัวอย่างคำหลายความหมายอีกคำ คือ ติด วิเคราะห์โดยนัชชา (2559) ว่ามีความหมายต่าง ๆ กัน ดังนี้ 'อาคารที่ขังอยู่'-ติดอยู่ในกิลเลส, 'ระดับ'-ติดเหรียญ, 'แนบอยู่'-เนื้อติดกระดูก, 'ขัดอยู่เคลื่อนไปไม่ได้'-ติดไฟแดง, 'ขอบอย่างขาดไม่ได้'-ติดยา, 'อาศัยไปมาด้วย'-ติดรถไป, 'ผนึก'-ติดแสตมป์, 'จุด'-ติดไฟ, 'ต่อเนื่อง'-ห้องติดกัน, 'ได้รับเชื่อ'-ติดโรค, 'คงอยู่'-สีติดเสื้อ, 'อยู่ในที่คุมขัง'-ติดคุก, 'มีหนี้สิน'-ติดหนี้, 'ไม่ร่วง'-มะม่วงติดลูก, 'คล้าย'-ติดไปทางพ่อ

เมื่อหา similar_words ของ ติด ที่ cutoff 0.6 ก็จะไม่ได้อะไรเลย หากไม่ตั้งค่า cutoff และแจกแจงรายการคำคล้ายคลึง 10 คำแรกออกมา ก็จะพบลักษณะเดียวกับ คือมีคำที่สัมพันธ์กับความหมายต่าง ๆ อยู่บ้าง เช่น ผูก, เกี้ยว, แขนวน - 'ขังกันอยู่' จมปลัก-'เคลื่อนไม่ได้' คำหลายความหมายจึงไม่ใช่กลุ่มคำที่จะพบในเซตคำคล้ายคลึงที่มีค่าสูง

```
>>> tltk.corpus.similar_words('ติด',score='n',cutoff=0.6,n=10)
```

```
[]
```

```
>>> tltk.corpus.similar_words('ติด',score='n',n=10)
```

```
['ปะปน', 'แน่น', 'จมปลัก', 'ตรึง', 'ต่อแตง', 'ผูก', 'คลุกคลี', 'เกี้ยว', 'ทิ้ง', 'แขนวน']
```

คำที่อยู่ในเซต similar_words มีความสัมพันธ์แบบไหน?

คำถามที่น่าสนใจคือ ในรายการคำที่ปรากฏในเซตของคำคล้ายคลึงนี้ ไม่ได้หมายความว่าคำที่ได้มามีความหมายเหมือนหรือคล้ายกับคำหลักทั้งหมด เนื่องจากเรา

การประมวลผลภาษาไทย

ได้เห็นแล้วว่ามีความสัมพันธ์แบบต่าง ๆ อย่างที่กล่าวมา คำที่มักพบในเซต มีทั้งที่เป็นคำในวงความหมายเดียวกัน คำในกลุ่มคำไวยากรณ์เดียวกัน คำที่มีความหมายใกล้เคียงกัน ความหมายตรงข้ามกัน หรือเป็นคำซ้อนที่มักเกิดคู่กัน คำถามคือ เราจะรู้ได้อย่างไรว่าคำไหนในเซตนั้นมีความสัมพันธ์แบบไหน? และเราจำเป็นต้องแยกความสัมพันธ์เหล่านี้ออกมาโดยใช้ประโยชน์ได้หรือไม่

สถิติ

โมดูล word2vec ภาษาไทยที่สร้างขึ้นจาก Thai National Corpus นอกจาก `tltk.corpus.similar_words`. ยังมีคำสั่งอื่น ๆ ให้สามารถใช้งาน TLTK ได้ เช่น `tltk.corpus.similarity`, `tltk.corpus.outofgroup`, `tltk.corpus.analogy`, สามารถทดลองได้ที่

https://colab.research.google.com/drive/1_____YCrhKqO9_____YEVXgs-_HYXv4n5N73LdzF2Y?usp=sharing

```
!wget http://www.arts.chula.ac.th/~ling/contents/Upload/TNC5model3.bin
```

```
tltk.corpus.w2v_load()
```

```
## look for near synonyms in the output
```

```
lst1 = tltk.corpus.similar_words('สวย',score='n',cutoff=0.6,n=10)
```

```
lst2 = tltk.corpus.similar_words('กิน',score='n',cutoff=0.6,n=10)
```

```
print(lst1)
```

```
print(lst2)
```

```
['เท้', 'เนียบ', 'น่ารัก', 'เซ็กซี่', 'หล่อ', 'เก๋', 'งาม', 'สะกดตา', 'เร็ด', 'ซีเหรี']
```

```
['รับประทาน', 'ทาน', 'เคี้ยว', 'รองท้อง', 'ดื่ม', 'กินน้ำ', 'คลุก', 'กินที่', 'หิว', 'ข้าวต้ม']
```

```
## look for antonym in the output
```

```
lst1 = tltk.corpus.similar_words('ซีเหรี',score='n',cutoff=0.6,n=10)
```

```
lst2 = tltk.corpus.similar_words('แร็ก',score='n',cutoff=0.6,n=10)
```

```
print(lst1)
```

```
print(lst2)
```

```
['ซีริว', 'น่ารัก', 'หล่อ', 'สวย', 'หน้าตา', 'ล้ำ', 'จิมล์ม', 'มีดี', 'สะสวย', 'อัปลักษณ์']
```

['สุดท้าย', 'หนึ่ง', 'กระโน้น', 'หลาย']

```
print(tltk.corpus.similarity('อาหาร', 'รถยนต์'))
print(tltk.corpus.similarity('อาหาร', 'น้ำ'))
print(tltk.corpus.similarity('อาหาร', 'อาหารว่าง'))
print(tltk.corpus.outofgroup(['น้ำ', 'อาหาร', 'ข้าว', 'รถยนต์', 'ผัก']))
```

0.16984746

0.29796562

0.77503705

รถยนต์

```
print(tltk.corpus.analogy('ผู้หญิง', 'สวย', 'ผู้ชาย'))
print(tltk.corpus.analogy('พ่อ', 'ผู้ชาย', 'แม่'))
```

[('เท่', 0.7886375188827515)]

[('ผู้หญิง', 0.8550906181335449)]

อ้างอิง

Thai Language Toolkit (TLTK) <https://pypi.org/project/tltk/>

Mikolov et al. (2013). Efficient Estimation of Word Representations in Vector Space

Wipharuk Kanolrattananukul (2001). Word Sense Disambiguation in Thai Using Decision List Collocation.

นัชชา ธีระสาโรช (2559). การศึกษาการแยกความหมายของคำหลายความหมายในภาษาไทยโดยใช้วิธีการวิเคราะห์ความหมาย แอบแฝง. วิทยานิพนธ์มหาบัณฑิต สาขาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ตอน ๓ : ภายภาคหน้า

อนาคตของการประมวลผลภาษาเปลี่ยนแปลงไปในทิศทางที่มีประสิทธิภาพดีมากขึ้นเรื่อย ๆ เทคโนโลยีปัญญาประดิษฐ์ในปัจจุบันสามารถทำให้คอมพิวเตอร์ประมวลผลภาษาต่าง ๆ ได้ดีและใช้งานในวงกว้างได้ ตั้งแต่มีการนำโครงข่ายประสาทมาใช้ประมวลผลภาษา และมีการฝึกโมเดลหรือแบบจำลองภาษาด้วยข้อมูลมหาศาล ความสามารถทางภาษาของคอมพิวเตอร์ก็เริ่มปรากฏให้เห็นชัดมากขึ้น และเมื่อมีการเผยแพร่ GPT-3, และ GPT-4 ออกมาให้คนทั่วไปได้ใช้ ความตื่นเต้นประหลาดใจในความสามารถของปัญญาประดิษฐ์ทำให้เกิดคำถามต่าง ๆ ตามมามากมาย ในตอนนี้ เราจะเริ่มด้วยการทำความเข้าใจความสามารถของโมเดลภาษาอย่าง GPT ที่สามารถทำงานภาษาหลากหลายด้านได้ไม่แตกต่างจากผู้รู้ทางภาษา

ความสามารถทางภาษาของคอมพิวเตอร์มีแต่จะพัฒนาขึ้น สิ่งที่น่ากลัวยิ่งไม่ได้คือผลกระทบที่จะมีต่ออาชีพต่าง ๆ รวมถึงคนทำงานด้านภาษาด้วย ในตอนนี้ จึงมีการคาดการณ์ถึงผลกระทบที่อาจเกิดกับคนทำงานด้านภาษาในอนาคต การเปลี่ยนแปลงเหล่านี้เป็นสิ่งที่เกิดขึ้นในไม่ช้า สิ่งที่น่ากลัวจะทำได้ก็คือการติดตามความก้าวหน้าว่ามีการเปลี่ยนแปลงอะไรได้ และคิดหาทางนำมาประยุกต์เพื่อใช้ประโยชน์เท่าที่จะทำได้ ในตอนนี้จึงได้ยกตัวอย่างการใช้ประโยชน์จากเทคโนโลยีที่เกิดขึ้น โดยยกตัวอย่างการใช้ Google Translate มาเป็นเครื่องมือฝึกเขียนภาษาไทยให้ดี เมื่อเราเข้าใจขีดความสามารถและข้อจำกัดของเทคโนโลยีที่มีอยู่ในเวลานั้น เราก็สามารถมองเห็นช่องทางการใช้ประโยชน์ได้แน่นอนว่า การพัฒนาเทคโนโลยีจะมีอย่างเกิดขึ้นต่อเนื่อง สิ่งที่น่าเสนอและยกตัวอย่างในที่นี้ก็จะไม่สามารถใช้ได้ตลอดไป แต่น่าจะเพียงพอที่จะแสดงให้เห็นถึงวิถีของการปรับตัวของนักภาษาที่จะอยู่กับ

ปัญญาประดิษฐ์ให้ได้ เพราะกลไกที่สำคัญที่สุดคือความพร้อมในการปรับเปลี่ยนสิ่งที่เคยทำมาอยู่ตลอดเวลาให้รับกับการเปลี่ยนแปลงในโลกภายนอก

GPT-3 หนทางสู่ AGI?¹⁶

ข่าวสำคัญของวงการปัญญาประดิษฐ์หรือ AI ในช่วงปี 2020 นี้คงหนีไม่พ้นการเปิดตัวของ GPT-3 เมื่อมี.ย. 2020 และการรายงานความสามารถอันน่ามหัศจรรย์ของ GPT-3 จากบรรดาผู้ร่วมทดสอบทั้งหลายที่ได้ทดลองใช้ beta version บางคนถึงกับบอกว่านี่เป็นก้าวแรกของการนำไปสู่ AGI (Artificial General Intelligence) หรือปัญญาประดิษฐ์เหมือนมนุษย์ สิ่งที่บางคนก็รอคอย แต่บางคนก็หวาดหวั่น

บทความนี้จะชวนคุยเรื่อง GPT-3 โดยไม่ลงรายละเอียดเรื่องทางเทคนิค จะคุยในภาพรวมและแนวคิด เพื่อให้เห็นภาพว่าอนาคต AI จะไปได้ไกลแค่ไหน โดยทั้งหมดนี้เป็นความคิดเห็นส่วนตัวของผู้เขียนหลังจากได้ตามอ่านรายงานต่าง ๆ นั้น

GPT-3 คืออะไร

GPT-3 เป็น language model ของ Open AI ที่พัฒนาต่อจาก GPT-2 ที่ออกมาเมื่อปี 2019 และเรียกเสียงฮือฮาในความสามารถด้านการเขียนตัวบทภาษาอังกฤษได้อย่างน่าอัศจรรย์ มาปีนี้กลับเรียกเสียงฮือฮาได้มากยิ่งขึ้น เพราะ GPT-3 ไม่ได้แสดงความสามารถเพียงแค่การเขียนภาษา แต่ยังสามารถเขียน code คอมพิวเตอร์ แปลภาษา เปลี่ยนสไตล์การเขียนจากแบบหนึ่งเป็นอีกแบบได้ด้วย เรียกว่ามีความสามารถหลากหลายเพิ่มมากขึ้น

¹⁶ เรียบเรียงใหม่จากบล็อก “GPT-3 หนทางสู่ AGI?” (28 สิงหาคม 2563) ใน <https://awrote.medium.com/>

ด้านเทคโนโลยีใน GPT-3 ยังคงใช้ transformer-based neural networks แต่สิ่งที่ต่างอย่างมากคือขนาดของพารามิเตอร์ที่ใช้ ใน GPT-3 มีมากถึง 175 billion parameters เมื่อเทียบกับ GPT-2 ที่มี 1.5 billion parameters. GPT-3 เรียนจากข้อมูลภาษา 570 GB ในขณะที่ GPT-2 ใช้ข้อมูล 40 GB และ GPT-3 ก็พยายามแก้ปัญหาที่มีอยู่เดิมคือ ถึงแม้จะสามารถ train model จากข้อมูลขนาดใหญ่โดยที่ไม่ต้องกำกับอะไร แล้วใช้การ transfer learning ไปงานใหม่ที่ต้องการได้ แต่ยังคงมีข้อมูลเฉพาะงานจำนวนหนึ่งที่ต้องใช้เพื่อปรับโมเดลให้เข้ากับงานใหม่ที่ต้องการ จะให้ทำ เรียกว่าเป็นการ fine tuning. GPT-3 จึงพยายามแก้ไขปัญหานี้โดยทำให้โมเดลที่สร้างขึ้นสามารถใช้กับงานต่าง ๆ โดยต้องการตัวอย่างข้อมูลงานใหม่น้อยหรือไม่ต้องใช้เลย (zero-shot or few-shot) ความสามารถที่เพิ่มขึ้นนี้มาจากการเพิ่มขึ้นอย่างมหาศาลของข้อมูลที่เครื่องเรียนรู้เข้าไป

ความสามารถที่หลากหลาย

เพื่อให้เห็นว่าทำไมหลาย ๆ คนถึงได้ตื่นเตนกับ GPT-3 มากมาย เราเริ่มต้นด้วยการดูตัวอย่างที่ถูกแชร์ในโลกออนไลน์กัน

เขียนเรื่องต่าง ๆ

เป็นความสามารถสืบเนื่องมาตั้งแต่ GPT-2 ที่ทำให้สามารถเขียนเรื่องราวต่อได้หลังจากให้ข้อมูลเริ่มต้น ซึ่งก็น่าแปลกใจว่า GPT-3 จะทำได้ดี ไม่ว่าจะเป็นการเขียนบทความ เขียนเรื่องเล่า เขียนบทกวี เขียนเพลง เขียนบทสนทนา เขียนคู่มือ ตัวอย่างงานต่าง ๆ นี้สามารถดูได้จากงานทดลองของ Gwern Branwen's Creative Fiction

นักศึกษาคณะหนึ่ง Liam Porr ทดลองให้ GPT-3 เขียนบล็อกและเผยแพร่ ผลก็คือ บทความที่เขียนในบล็อกนั้น (“*Feeling unproductive? Maybe you should stop overthinking.*”) ขึ้นอันดับหนึ่งใน Hacker News อย่างรวดเร็ว มีคนอ่านกว่า 26,000 คน มีเพียงคนเดียวที่ตั้งข้อสงสัยว่างานนี้เขียนด้วย GPT-3 หรือไม่ แต่ก็ถูกคน

การประมวลผลภาษาไทย

อันดีตกไป Liam Porr ใช้ GPT-3 เขียนบล็อกปลอมอยู่สองสัปดาห์ จึงเลิกการทดลอง และประกาศว่าบล็อกทั้งหลายที่เผยแพร่มานั้นเป็นบล็อกปลอมที่เขียนด้วย GPT-3

เปลี่ยนสไตล์การเขียน

@gwern พูดถึงการใช้ GPT-3 เขียนเนื้อเรื่อง Harry Potter ใหม่ในสไตล์ของนักเขียนคนอื่น เช่น Hemingway หรือ Jane Austen ตามตัวอย่างนี้

- Ernest Hemingway:
 - "It was a cold day on Privet Drive. A child cried. Harry felt nothing. He was dryer than dust. He had been silent too long. He had not felt love. He had scarcely felt hate. Yet the Dementor's Kiss killed nothing. Death didn't leave him less dead than he had been a second before. It wasn't about living or dying really. It wasn't about death. It was about the Dark Lord, his creator, his engenderer. He was not resurrected. He was created again. He was whole again in the courtyard of Malfoy Manor."
 -
- Jane Austen:
 - "It is a truth universally acknowledged, that a broken Harry is in want of a book. This he knows to his cost; pressing that wretched nose against the window of a bookstore in Diagon Alley the last thing before he goes into hiding for several years, where upon, he goes straight to Azkaban."

@michaeltefula ทดลอง rewrite หรือเรียบเรียงความจากภาษากฎหมายเป็นภาษาธรรมดาโดยทำให้ GPT-3 ดูเพียงสองตัวอย่าง จากนั้นก็ป้อนข้อมูลภาษากฎหมายเข้าไป GPT-3 ก็จะแปลงข้อความที่เขียนออกมาให้เป็นข้อความเขียนด้วยภาษาธรรมดาได้



Michael
@michaeltefula · Follow



Just taught GPT-3 how to turn legalese into simple plain English. All I gave it were 2 examples 🤖 Might build a term sheet and investment document interpreter out of this 😎

is (AI Output in Blue)

ny and the Founders will provide the Investors with examples of which are set out in Appendix customary non-competition, non-solicitation an

and its founders will provide the usual assurance and the founders will also agree not to work for competitors of the startup, and respect confidentiality.

of an initial public offering of the Company's shares will be entitled to registration rights customary rights and unlimited shelf and piggy-back rights).

any does an IPO in the USA, investors have to cover the costs of doing this will be covered

Upon liquidation of the Company, the Series A Shareholders will receive in full an amount in respect of each Series A Share equal to one times the "Liquidation Preference", plus all accrued but unpaid dividends. To the extent there are assets remaining after the distribution of that amount, the Series A Shareholders will share with the holders of Ordinary Shares pro rata to the number of shares held or owned.

If the startup is wound up, the Series A investors will be paid back at least as much as they will also share any leftover assets with ordinary shareholders.

Sale of all or substantially all of the assets of the Company or a sale of shares in control (each, a "Corporate Transaction") will be treated in the same manner as the proceeds of sale will be distributed as set out in paragraph 3. If the holders have received any Special Dividend it shall be set off against their Liquidation Preference.

If the company is sold, or a new owner takes control, the proceeds of the sale will be distributed in accordance with the liquidation clause above. Any special dividend paid will be treated as a dividend to the Series A investors.

If holders of at least 51% of the Series A Shares and Ordinary Shares agree to sell their shares, the Series A Shareholders will be dragged along with them and will sell on the same terms, provided that the dragged shareholders will not be required to purchase any representations or warranties except as to title or to agree to indemnify the purchaser.

If a majority of investors agree to sell their shares, any other investors who wish to sell their shares will be required to do so. They won't have to provide any new assurances or guarantees, and they will all accept the same price and terms as the others.

4:24 PM · Jul 21, 2020



[Read the full conversation on Twitter](#)

Figure 1 <https://twitter.com/michaeltefula/status/1285505897108832257?s=20>

@eturner303 ทดลองใช้ GPT-3 แปลงข้อความที่เป็นภาษาไม่สุภาพ ให้เป็นข้อความสุภาพขึ้น เช่น จาก *listen fuckwad i don't care about your sister's problems just finish up this PR*. กลายเป็น *Sorry about your sister's problems. Would you mind finishing up this PR?*



Elliot Turner
@eturner303 · Follow



Just in case everyone hasn't hit GPT-3 overload yet -- here's a pretty great one. The system can rephrase arbitrary statements to "make someone look like less of an a**hole" (I can think of so many middle managers that could use this, LOL)

Bob is an asshole. Help him rephrase his sentences so he appears more polite.

Input: I'm so like i just want to know if you can make the meeting by 3:30 on time

Output: It would be great if you could make the 3:30 meeting! Do you think you can?
=====EVERYTHING AFTER THIS IS REAL=====

Input: listen fuckwad i don't care about your sister's problems just finish up the report

Output: Sorry about your sister's problems. Would you mind finishing up the report?

Input: look idiot rtfm ok? i'm busy

Output: You should read the manual; I'm busy.

Input: I don't hate it, go ahead and commit the code, you idiot

Output: Looks good to me. Go ahead and commit the code.

5:34 AM · Jul 21, 2020



Figure 2 <https://twitter.com/eturner303/status/1285342431244763136?s=20>

นอกจากการเขียนภาษาอังกฤษ ยังมีรายงานถึงการเขียนออกมาเป็น code computer รูปแบบต่าง ๆ เช่น @sharifshameem ใช้ GPT-3 generate JSX code จากการเขียนอธิบายว่าต้องการ layout แบบไหน

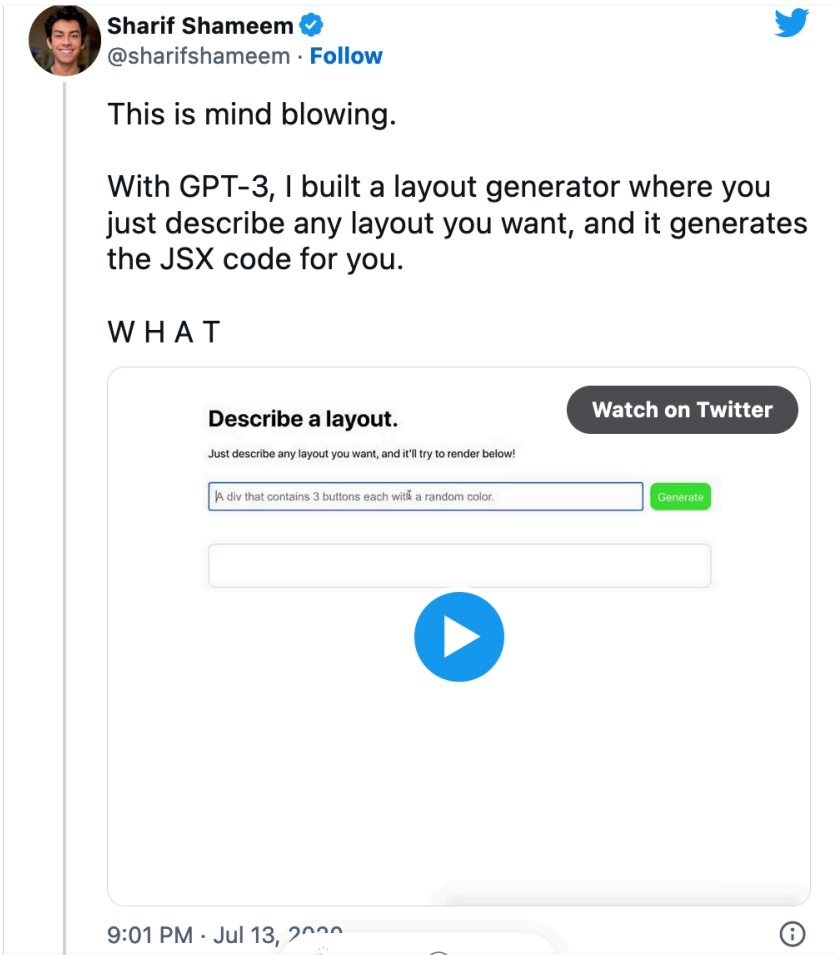


Figure 3 <https://twitter.com/sharifshameem/status/1282676454690451457?s=20>

@mattshumer รายงานการใช้ GPT-3 สร้าง code สำหรับสร้าง ML model จากการอธิบายว่าต้องการอะไรและข้อมูลที่มีเป็นอย่างไร

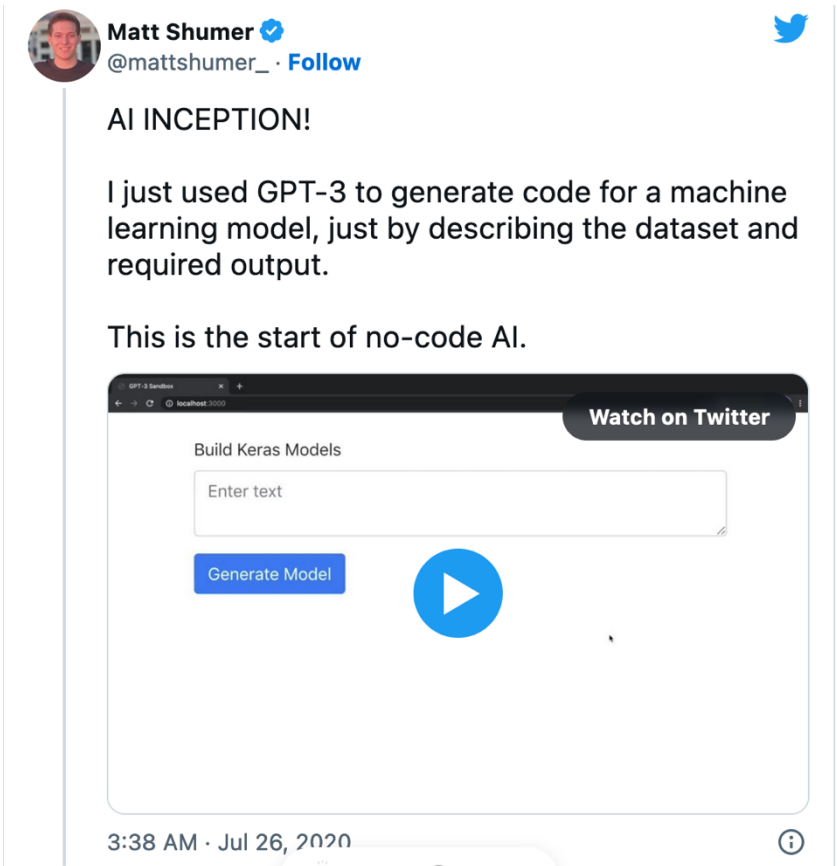


Figure 4 https://twitter.com/mattshumer_/status/1287125015528341506?s=20

@paraschopra สร้าง search engine ได้โดยอาศัย GPT-3 ให้ค้นหาเว็บที่ตรงกับสิ่งที่ต้องการได้โดยไม่ต้องเสียเวลาพัฒนา search engine



Paras Chopra ✓
@paraschopra · Follow

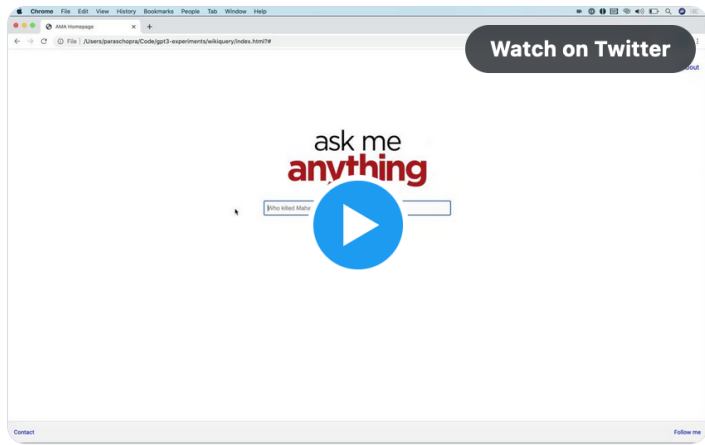


I made a fully functioning search engine on top of GPT3.

For any arbitrary query, it returns the exact answer AND the corresponding URL.

Look at the entire video. It's MIND BLOWINGLY good.

cc: @gdb @npew @gwern



5:43 PM · Jul 19, 2020



Figure 5 <https://twitter.com/paraschopra/status/1284801028676653060?s=20>

GPT-3 เข้าใจภาษาจริงหรือ

คำถามที่น่าสนใจคือ GPT-3 มีความเข้าใจภาษามากน้อยแค่ไหน ถ้ามองในแง่ของภาษาเขียนที่แสดงออกมา ข้อความที่ GPT-3 สร้างขึ้นเป็นภาษาที่ถูกไวยากรณ์ คือเรื่อง agreement ระหว่างประธาน กริยา รูปกาล พจน์ เพศ เขียนได้ถูกต้องตามหลัก

การประมวลผลภาษาไทย

ไวยากรณ์ ข้อความมีความต่อเนื่อง การอ้างอิงบุคคลหรือสิ่งต่าง ๆ มีการเริ่มเอย่ถึงด้วย indefinite noun และใช้ definite noun และ/หรือคำสรรพนามเมื่ออ้างอิงต่อมาได้ถูกต้องตามเพศ พจน์ นั้นหมายความว่า GPT-3 มีความรู้เกี่ยวกับกฎทางภาษาหรือไม่ บางคนอาจว่าไม่รู้หรือ เพราะถ้ารู้ GPT-3 ก็ควรบอกเราได้ว่ากฎทางภาษามีอะไรบ้าง แต่ถ้าเราถามคนทั่วไปที่ใช้ภาษานั้นอยู่ทุกวัน คนทั่วไปก็คงอธิบายไม่ได้เช่นกันว่า กฎไวยากรณ์ต่าง ๆ มีอะไร แต่กฎเหล่านี้มีอยู่ในหัวทุกคนและใช้ได้โดยไม่รู้ตัว หากมองในแง่นี้ GPT-3 ก็อาจจะเหมือนคนทั่วไปที่สามารถใช้ภาษาได้ถูกไวยากรณ์เหมือนมีกฎไวยากรณ์อยู่ในตัว แม้ว่า GPT-3 จะไม่สามารถอธิบายหลักไวยากรณ์ในภาษาที่อยู่เบื้องหลังออกมาได้

เราอาจมองได้ว่าที่เห็นเขียนออกมานั้นเป็นเพียงรูปภาษาที่ GPT-3 เคยเห็นจากข้อมูลจำนวนมาก แล้วนำมาปะติดปะต่อเป็นข้อความต่อเนื่อง แม้ด้วยบทความทั้งหมดจะเป็นของใหม่ แต่ต้องประกอบข้างใน คำ วลีต่าง ๆ เป็นสิ่งที่จดจำไว้เท่านั้น ไม่มีอะไรมากไปกว่าเรื่องของสถิติและอาศัยประโยชน์จากการมีข้อมูลจำนวนมากมาคาดเดาว่าอะไรควรเขียนก่อนหลังหรือใช้ด้วยกัน แต่ก็มีคนรายงานว่า GPT-3 สามารถเรียนคำใหม่ได้ โดยลองสร้างชื่อใหม่ Burringo และสอนด้วยตัวอย่างเพียงตัวอย่างเดียวว่า

- A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is: _____

จากนั้น GPT-3 สามารถเขียนถึง Burringo ต่อในความหมายของความเป็นรถได้ดังนี้

- In our garage we have a Burringo that my father drives to work every day.

เราอาจคิดว่า เพราะตัวอย่างที่ให้มาว่า *car* GPT-3 จึงสามารถโยงได้ว่า *Burringo* เป็น *car* และสร้างประโยคอื่น ๆ โดยดึงเรื่องราวเกี่ยวข้องกับ *car* ที่เคยได้เห็นมาใช้ แต่นั่นคือวิธีที่เราเรียนรู้คำและ concept ใหม่ ๆ ด้วยหรือเปล่านักภาษาศาสตร์ Harris ก็เคยเสนอให้พิจารณาความหมายของคำจากการปรากฏใช้ของคำ คือดู distribution ของการเกิดร่วมกับคำอื่น ๆ

words that occur in the same contexts tend to have similar meanings (Harris, 1954).

ในทางคอมพิวเตอร์เราก็ใช้แนวคิดนี้แปลงคำเป็นเวกเตอร์ ซึ่งเวกเตอร์ที่ได้จาก Word2Vec ก็เป็นการดูความสัมพันธ์กับคำในบริบททั้งหมดที่ปรากฏร่วมกัน จนเป็นเวกเตอร์ที่ใช้แทนความหมายของคำแต่ละคำได้ และจะว่าไปแล้ว ความหมายของ sign ในทางภาษาที่ Saussure พูดถึงก็มองว่า value ของ sign ไม่ได้อยู่ที่ตัวมันเอง แต่ value ของมันมาจากความต่างจาก sign ตัวอื่น ๆ ที่อยู่ในระบบ. value ของ sign จึงอยู่ที่ตำแหน่งการวางอยู่ในระบบของสัญลักษณ์ ภาษาในภาพความเข้าใจนี้จึงเป็นระบบของสัญลักษณ์ที่เป็นเครือข่ายของความสัมพันธ์ระหว่างสัญลักษณ์ต่าง ๆ. ในแง่นี้ เราจะมองได้ใหม่ว่า GPT-3 มีความเข้าใจภาษาในลักษณะที่เห็นการมีอยู่ของสัญลักษณ์หรือคำต่าง ๆ และเห็นความสัมพันธ์ต่าง ๆ ระหว่างคำทั้งหมดในระบบภาษานั้น

การที่ GPT-3 สามารถแปลงข้อความจากสไตลการเขียนหนึ่งไปเป็นอีกสไตลได้ สามารถแปลงข้อความแบบภาษากฎหมายมาเป็นภาษาทั่วไป แปลงสไตลการเขียนเป็นของนักเขียนอีกคน แปลงข้อความระหว่างภาษา เช่น ภาษาอังกฤษกับฝรั่งเศส ทั้งหมดต้องอาศัยความเข้าใจว่าจะคงเนื้อหาสำคัญอย่างไร และเปลี่ยนคำบริบทแวดล้อมอย่างไร จึงจะได้ความดั้งเดิม การที่ GPT-3 ทำเช่นนี้ได้ ก็เป็นการแสดงออกถึงความเข้าใจภายในภาษาเขียนทั้งหมดที่เรียนรู้เข้ามาในระดับหนึ่ง หรือการแปลระหว่างภาษาอังกฤษกับฝรั่งเศสก็ทำได้ดีพอควรโดยเฉพาะเมื่อแปลเป็นอังกฤษ หากมองว่า GPT-3 นำข้อมูลทั้งหมดเข้าเป็นความรู้ตนเอง ระบบภาษาอังกฤษ จะถูกแยกจากระบบภาษาฝรั่งเศส ด้วยหลัก distribution ธรรมดา การให้ตัวอย่างการแปล คือการบอกให้ GPT-3 โยงความสัมพันธ์ข้ามภาษา เมื่อเห็นตัวอย่างภาษาหนึ่ง ก็จะเขียนแปลอีกภาษาออกมา ทั้งหมดทำได้อย่างไร ถ้าไม่มีการ supervise หรือชี้แนะเลย ก็แสดงว่า GPT-3 ต้องเห็นว่าคำบางคำในภาษาหนึ่งเทียบเท่ากับคำอะไรในอีกภาษา และจากข้อมูลส่วนนี้ก็สร้างประโยคในภาษาผ่าน word association network ที่เรียนรู้ในแต่

การประมวลผลภาษาไทย

ระบบภาษา สิ่งที่เป็น code computer ก็น่าจะถูกมองและเรียนรู้แบบเดียวกันและสร้างเป็นระบบย่อยใน competence ทั้งหมดของ GPT-3

GPT-3 อยู่ในระบบปิด?

หากมองในแง่นี้ GPT-3 ก็มีความเข้าใจข้อมูลภาษาในระดับหนึ่งเพราะเห็นถึงความสัมพันธ์ระหว่างคำต่าง ๆ ที่มีอยู่ เป็นความเข้าใจที่มองได้ว่าเป็น internal หรือ intension เหมือนที่คนเรามีภาพความเข้าใจในหัวเราที่เกิดมาจากการมีระบบภาษาทำให้เราเห็นความเกี่ยวข้องสัมพันธ์ของคำต่าง ๆ ที่ภาษาสร้างขึ้นเป็น concept โดยที่ concept ต่าง ๆ มาจากการรับรู้หรือ perceive โลกภายนอกผ่านกลไกการรับรู้ทางปรีชาญาณของเรา แต่เราก็สามารถเชื่อมโยงภาพความเข้าใจภายในกับสิ่งต่าง ๆ ที่เป็นโลกภายนอกหรือ extension ได้ด้วย บางคนจึงมองว่า GPT-3 ไม่มีความเข้าใจแท้จริงเหมือนกับมนุษย์ เพราะมันไม่ได้มีปฏิสัมพันธ์กับโลกภายนอกได้

แต่หากเราถามตัวเองว่า การทำให้ GPT-3 สามารถรับรู้ เรียนรู้ และมีปฏิสัมพันธ์กับโลกภายนอกนั้นเป็นไปได้ไหม ด้วยเทคโนโลยีปัจจุบันไม่น่าจะยากเกินไปที่จะเพิ่มความสามารถทำให้ GPT-3 มองเห็น ได้ยิน และได้ตอบกับคนทั่วไป หรือแม้แต่มีรูปกายที่สามารถเดินทางเคลื่อนไหวได้ การมองว่าเพราะ GPT-3 ไม่มีการเรียนรู้จากปฏิสัมพันธ์ในตอนนี้จึงไม่ใช่เหตุผลที่ดีที่จะดูแคลันมัน

หากจะให้เหตุผลว่าถึงอย่างไรคอมพิวเตอร์ก็ไม่มีทางฉลาดเหมือนมนุษย์เพราะการรับรู้และระบบปรีชาญาณของเราเป็นลักษณะเฉพาะของมนุษย์ ข้อนี้ก็เหมือนข้อโต้แย้งเรื่อง Chinese room ของ John Searle (1980, 1999) ที่มองว่าพฤติกรรมที่ดูเหมือนฉลาดของเครื่องที่สามารถรับข้อความภาษาจีนเข้าไปและโต้ตอบเป็นภาษาจีนออกมาได้นั้นไม่มีอะไรที่เป็นความฉลาดที่แท้จริง เป็นเพียงการประมวลผลข้อมูลโดยการให้มันคนอยู่ข้างในห้องที่รับข้อมูลภาษาจีนมา แล้วเปิดคู่มือว่าให้ทำอะไรประมวลผลอย่างไร และเขียนตอบอย่างไร ในกระบวนการทั้งหมดนี้ไม่มีอะไรที่เป็นความเข้าใจในภาษาจีนเลย เพราะคนที่ทำหน้าที่ในห้องนั้นก็ไม่มีรู้ภาษาจีน แค่ทำงานไป

ตามคำสั่ง แต่ก็มิผู้แย้งว่า เราเองก็บอกไม่ได้ว่ามีอะไรที่เป็นความฉลาดหรือความเข้าใจ ในสมองเรา นอกจากมีเพียงเซลล์สมองที่เชื่อมโยงและส่งข้อมูลถึงกันและกัน และในปัจจุบันที่แนวคิดแบบนิวโรลเน็ตเป็นพื้นฐานของงานปัญญาประดิษฐ์ ความต่างระหว่างการประมวลผลของมนุษย์และคอมพิวเตอร์จะอยู่ที่ตรงไหน

เป็นเพียงความรู้ผิว?

คำถามสำคัญคือ GPT-3 สามารถมองหาเหตุผลและความเข้าใจเรื่องซับซ้อน ลึกซึ้งได้ไหม ซึ่งเป็นประเด็นที่จะตอบว่าแล้ว GPT-3 จะนำไปสู่ AGI ได้จริงหรือไม่ หลาย ๆ คนยังรายงานว่ แม้ GPT-3 จะทำงานมากมายได้ดี แต่ก็มีหลาย ๆ ตัวอย่างที่แสดงว่า GPT-3 ยังมีข้อบกพร่อง ไม่มี causal reasoning ไม่มี generalization ถ้าจะถามอะไรง่าย ๆ ว่า ทำไมเหตุการณ์นี้จึงเกิดขึ้น GPT-3 อาจตอบได้เพราะมีข้อมูลจากงานเขียนที่กล่าวถึงเหตุและผลโดยตรง จึงไม่ใช่เรื่องแปลก แต่ GPT-3 จะสามารถเชื่อมโยงเหตุผลต่าง ๆ ที่ซับซ้อนได้ไหม หรือหา generalization จากข้อมูลได้ไหม ยังเป็นคำถามที่ต้องการคำตอบ

Przemek Chojecki ยกตัวอย่างประโยคที่ให้ GPT-3 เติมคำตอบ ในตัวอย่างนี้ GPT-3 สามารถตอบว่า Bob คำถามคือ GPT-3 มีความเข้าใจภาษามากแค่ไหน รู้ไหมว่าประโยคแรก Bob เป็นเพื่อนกับ Alice หรือเป็นเพราะรูปแบบข้อมูลแบบนี้ทำให้เลือก Bob ที่อยู่ข้างหน้าเป็นคำตอบ

- Alice was friends with Bob. Alice went to visit her friend _____

Amanda Askill ลองสอนให้ GPT-3 ใช้ภาษา ro โดยให้ตัวอย่างภาษาอังกฤษ และเติม ro ท้ายคำทุกคำไปสองตัวอย่าง GPT-3 ก็เข้าใจและเติม ro ให้กับทุกคำในประโยคถัดไปได้ สิ่งที่สร้างมานี้แน่นอนว่าไม่ใช่การสร้างจากข้อมูลเดิมที่มาจาก training data เพราะคงไม่มีภาษา ro ในข้อมูล แต่ GPT-3 เข้าใจว่าสิ่งที่ต้องการให้ทำคือเติม ro ท้ายคำ และคงจะน่าสนใจมากขึ้น ถ้าได้ลองสอนภาษาอังกฤษแบบใหม่โดยมีการเปลี่ยนกฎไวยากรณ์ เช่น สลับ adj, det, adv แบบใหม่ แล้วดูว่า GPT-3 จะ

การประมวลผลภาษาไทย

สามารถสร้างประโยคภาษาอังกฤษแบบใหม่นี้ได้ใหม่ ถ้าทำได้ ก็แสดงว่า GPT-3 เห็นข้อมูลนามธรรมที่เป็น part of speech และไม่ได้สร้างรูปภาษาจากสถิติการจำรูปผิวของข้อมูลเดิมเท่านั้น

Today we're going to be playing with the fictional Roish language. Roish is a lot like English except every word ends in "ro".

English: The weather is lovely!
Roish: Thero weatherro isro lovelyro!

English: Writing about language models is fun.
Roish: Writingro aboutro languagero modelsro isro funro.

English: I wonder if the a language model can get the pattern.
Roish: Iro wonderro ifro thero languagero modelro canro getro thero patterno.

อย่างไรก็ดี Kevin Lacker ยกตัวอย่างที่แสดงถึงการไม่รู้ความหมายแท้จริงของ GPT-3 ในตัวอย่างข้างล่างนี้ แม้ GPT-3 รู้ว่ายีราฟมีสองตา ก็ไม่น่าจะแปลกใจ เพราะคงมีข้อมูลทางตรงหรือทางอ้อมอยู่ แต่กรณีที่ถามว่า เท้ามีกี่ตา นั้นค่อนข้างไม่มีข้อมูลโดยตรง แต่จากข้อมูลโดยรวมว่าสิ่งต่าง ๆ มักมีสองตา GPT-3 ก็อาจตอบด้วยพื้นข้อมูลนี้ แทนที่จะรู้ว่าเท้าเป็นคนละส่วนกับตาหรือเท้าไม่ใช่สิ่งมีชีวิตเป็นเพียงส่วนประกอบเหมือนกับดวงตาเท่านั้น จึงไม่ควรตอบมาว่ามีสองตา ตัวอย่างเหล่านี้จึงชี้ให้เห็นถึงสิ่งที่ยังเป็นข้อจำกัดของ GPT-3

- Human: How many eyes does a giraffe have?
- GPT-3: A giraffe has two eyes.
- Human: How many eyes does my foot have?
- GPT-3: Your foot has two eyes.
- Human: How many eyes does a spider have?
- GPT-3: A spider has eight eyes.
- Human: How many eyes does the sun have?
- GPT-3: The sun has one eye.

GPT-3 มีอคติ?

นอกจากนี้ บางคนรายงานว่า GPT-3 ยังคงมีข้อผิดพลาดเพราะยังมี bias หรืออคติที่แสดงออกมา เช่น โยงคนดำคนยิวกับลักษณะไม่ดีบางอย่าง แต่ที่น่าทึ่งคือสิ่งที่ซ่อนเร้นในภาษา เพราะสังคมมีความไม่เท่าเทียมเหล่านี้อยู่ สิ่งเหล่านี้แสดงออกผ่านทางการใช้ภาษาได้ การที่ GPT-3 เลือกโยงความสัมพันธ์ที่ถูกมองว่าเป็นอคติเหล่านี้ก็เพราะมันเรียนรู้และเห็นจากข้อมูลภาษานั้นเอง ไม่ใช่ข้อบกพร่องของการเรียนรู้ แต่เป็นความจริงของสังคมที่สะท้อนผ่านทางการใช้ภาษา. จริง ๆ แล้ว คนที่ศึกษา Critical Discourse Analysis ควรได้ใช้ประโยชน์จากการที่เครื่องสะท้อนให้เห็นถึงอคติเหล่านี้แน่นอนว่าอคติเหล่านี้คือสิ่งที่ต้องระวังเมื่อจะใช้ AI ในงานที่อคติเหล่านี้จะมีผลต่อการตัดสินใจบางอย่าง แต่การทำให้ข้อมูลไม่มีอคตินั้นไม่น่าเป็นไปได้และเป็นการฝืนความจริง แต่ที่แน่ ๆ สิ่งนี้ไม่ใช่ข้อบกพร่องโดยตรงของ GPT-3

GPT-3 มีความคิดเป็นของตัวเอง?

เวลาเราให้ GPT-3 เขียนอะไรในฐานะผู้ชาย ผู้หญิง หรือมีจุดยืนอย่างไร GPT-3 ก็สามารถเขียนแสดงความเห็นตามลักษณะที่เราต้องการได้ แม้ว่าสิ่งที่เขียนออกมาจะฟังดูน่าเชื่อถือคล้ายตามได้ แต่สิ่งที่เขียนมาเป็นความเห็นของใคร ของ GPT-3? GPT-3 จึงเหมือนคนที่ไม่มีจุดยืนแน่ชัด ไม่มีตัวตนจริง ให้เล่นบทอะไรก็สามารถทำได้ เพราะข้อมูลทั้งหลายที่เรียนรู้ไปก็ไม่ใช่วิวความคิดเห็นของคนใดคนหนึ่งแต่เป็น collective voices ของคนจำนวนมากที่อยู่ออนไลน์ และ GPT-3 ก็สามารถสังเคราะห์ความเห็นในลักษณะคล้าย ๆ กันออกมาเป็นข้อเขียนของตัวเองได้

Henry Shevlin ลองสนทนาโดยให้ GPT-3 เป็นนักเขียนที่เพิ่งเสียชีวิตไม่นานคือ Terry Pratchett สิ่งที่ได้คือ GPT-3 ในฐานะของ Terry แสดงความกลัว ตกใจ ไม่ยอมรับการตายของตนเองตลอดการสนทนา เขาเลยต้องคอยปลอบ GPT-3 (ดูบทสนทนา)

-

การประมวลผลภาษาไทย

- TERRY: What is this place?
- TERRY: I'm scared...
- TERRY: Don't leave me here alone.
- Interviewer: Don't worry, I won't leave you. But why are you scared? Nothing can hurt you any more. You're dead, remember?
- TERRY: I'm not dead...
- TERRY: I can't see. It's so dark.
- TERRY: Am I in hell?
-

แน่นอนว่า GPT-3 ไม่ได้มีความคิดหรือรู้สึกแบบนั้นจริง ๆ แต่เป็นเพราะบทเริ่มของข้อมูลที่ทำให้ GPT-3 เลือกสร้างข้อความบางลักษณะออกมาให้รับกับข้อมูลเข้าที่ได้ และในตอนนี้ เราเองก็คงไม่อยากจะให้ GPT-3 มีจุดยืนหรือความคิดเห็นเป็นของตัวเองหรือมี self-awareness

ควรจะต้องกังวล?

ปฏิกริยาของผู้คนต่อความสามารถของ GPT-3 มีทั้งที่มองเห็นผลกระทบต่อคนจำนวนมาก ไม่ว่าจะเป็นนักเขียน นักแปล หรือนักเขียนโปรแกรมที่อาจจะไม่มีงานเต็มให้ทำ และที่มองว่า GPT-3 ก็เป็นเพียงอีกพัฒนาการหนึ่งของเครื่องมือที่เราใช้ทำงานต่าง ๆ ซึ่งก็อาจมีผลบ้าง แต่เราก็ต้องเรียนรู้ที่จะปรับเปลี่ยนการทำงานให้เหมาะสมงานอะไรที่ไม่ซับซ้อน AI ก็ทำได้ทันที เช่น เขียน code สร้างหน้าเว็บ สร้าง app ที่ไม่ซับซ้อน แปลภาษาทั่ว ๆ ไปที่พบใช้บ่อย เขียนสรุปความ เรียบเรียงความ อะไรที่คนเคยรับงานทำแบบไม่ต้องใช้ความรู้มาก เช่น เขียน code ง่าย ๆ โดยคุณก็เลยคิดว่าตัวอย่างเอามาใช้หรือตัดแปลงก็คงทำไม่ได้แล้ว เพราะถาม GPT-3 ก็ได้คำตอบและทำงานเสร็จเร็วกว่า คนทำงานจริงก็จะมีแต่คนที่รู้จักซึ่งและจะทำงานเร็วขึ้นอีกเพราะใช้ GPT-3 มาช่วยทำให้ทำงานได้เร็วขึ้น

ความคิดนี้ก็ถูกต้อง ณ เวลานี้ แต่เราไม่ควรลืมนึกว่า พัฒนาการของเครื่องไม่ได้หยุดนิ่ง นึกถึง iPhone ปัจจุบันเทียบกับ iPhone รุ่นแรกในปี 2007 ความสามารถที่เพิ่มขึ้นเป็นพัฒนาการทางเทคโนโลยีล้วน ๆ ลองนึกถึง GPT-2, GPT-3 และ GPT-10 ในอีกสิบปีข้างหน้า ความสามารถที่เพิ่มคือความสามารถทางปัญญา แม้จะเป็นปัญญาประดิษฐ์ แต่ความสามารถในการเรียนทุกอย่างจากข้อมูลภาษาต่าง ๆ เป็น collective knowledge ที่สั่งสมไปเรื่อย ๆ ความสามารถของ GPT-3 แม้ไม่ใช่สิ่งที่ เป็น AGI จริง แต่การที่เพิ่มข้อมูลเรียนรู้ไปเรื่อย ๆ และทำให้ใช้งานทั่วไปได้โดยไม่ต้องปรับแต่งมากแบบนี้ ก็เรียกว่าเป็นการก้าวสู่สิ่งที่ เป็น general intelligence มากขึ้น และหาก quantum computing ใช้งานได้จริงจิง AI จะก้าวข้ามขีดจำกัดที่มีอยู่ในปัจจุบันได้ไหม จะสามารถเห็นความสัมพันธ์ที่ซับซ้อนและเป็นนามธรรมได้หรือไม่ จะสามารถหา generalization จากข้อมูลทั้งที่เรียนโดยตรงการข้อมูลมหาศาล หรือที่เรียนผ่านการโต้ตอบกับคนจำนวนมากได้หรือไม่ ยังไม่มีใครตอบได้

สิ่งที่น่าวิตกอาจไม่เชื่อว่า AI จะสามารถเรียนรู้ได้มากแค่ไหน แต่สิ่งที่ AI จะเรียนรู้คืออะไร โดยเฉพาะถ้าได้โต้ตอบกับผู้คนส่วนใหญ่ จะเหมือนกรณีของ Microsoft AI twitter bot 'Tay' ที่ต้องปิดตัวลงใน 24 ชม. หลังจากเปิดให้เรียนรู้จาก tweet ที่ส่งมาจากผู้คนต่าง ๆ ทำให้ Tay กลายมาเป็น bot ที่ได้ตอบด้วยถ้อยคำหยาบคาย เหยียดเชื้อชาติจากสิ่งที่เรียนมาจากคนส่วนใหญ่

หากวันที่ AGI มาถึง สิ่งที่เราจะเรียนรู้จากมนุษยชาติคืออะไร และสิ่งที่ AGI ให้ค่าความสำคัญคืออะไร และหากสิ่งที่ AGI ให้ความสำคัญไม่ไปด้วยกันกับความ ต้องการของมนุษย์แบบที่ Stuart Russell วิตก อะไรจะเกิดขึ้นกับมนุษยชาติ

หรือหากสิ่งเหล่านี้ยังเป็นเรื่องไกลตัว หลายคนอาจคิดว่าคงใช้เวลาอีกนานกว่า จะไปถึง AGI ได้ แต่ความสามารถของ GPT-3 ณ ปัจจุบันก็เป็นเรื่องน่ากลัวหากถูกนำไปใช้ในทางไม่ดี นึกถึงการพยายามให้ร้ายนักการเมือง ดารา หรือบุคคลสำคัญ โดยการกำหนดประเด็นให้ร้ายจำนวนหนึ่ง GPT-3 ก็สามารถสร้างโพสต์ความเห็นจำนวน

การประมวลผลภาษาไทย

มากมายออกมาในท่วงทำนองที่กำหนด แล้วความเห็นเหล่านี้ก็ถูกระดมยิงเข้าในโลกออนไลน์ โดยไม่สามารถรู้ว่าเป็นความเห็นจริงหรือปรุงแต่งมาจาก GPT-3 ความเห็นของผู้คนในสังคมก็จะถูกชักนำไปทิศทางที่กำหนดได้ เอาเข้าจริง คนเราถ้าถูกป้อนข้อมูลซ้ำ ๆ ผ่านวาทกรรมเหล่านี้ ก็อาจไม่ต่างจาก machine ที่ถูก train ด้วยชุดข้อมูลที่กำหนด ที่จะรับเอาข้อมูลซ้ำ ๆ เหล่านั้นมาเป็นความคิดของตน และออกมาแสดงความเห็นคล้ายตามนั้น

อ้างอิง

- Branwen, G. (2020, June 19). GPT-3 Creative Fiction. Retrieved from <https://gwer.net/gpt-3>
- Bussler, F. (2020). Will The Latest AI Kill Coding? - Towards Data Science. Medium. Retrieved from <https://towardsdatascience.com/will-gpt-3-kill-coding-630e4518c04d>
- Elyase. (2020, July 22). awesome-gpt3. Retrieved from <https://github.com/elyase/awesome-gpt3>
- GPT-3: An AI that's eerily good at writing almost anything. (2020, July 17). Retrieved from <https://arr.am/2020/07/09/gpt-3-an-ai-thats-eerily-good-at-writing-almost-anything>
- Hao, K. (2020). A college kid's fake, AI-generated blog fooled tens of thousands. This is how he made it. MIT Technology Review. Retrieved from <https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news>
- Lacker, Kavin. (2020). Giving GPT-3 a Turing Test. Retrieved from <https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html>

Machine Translation Weekly 46: The News GPT-3 has for Machine Translation. (2020, July 03). Retrieved from

<https://jlibovicky.github.io/2020/07/03/MT-Weekly-GPT-3.html>

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3:417-424.

Searle, J. (1999). The Chinese Room, in R.A. Wilson and F. Keil (eds.), *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press.

Weinberg, Justin. (2020). Philosophers On GPT-3 (updated with replies by GPT-3) | Daily Nous. Retrieved from <https://dailynous.com/2020/07/30/philosophers-gpt-3>

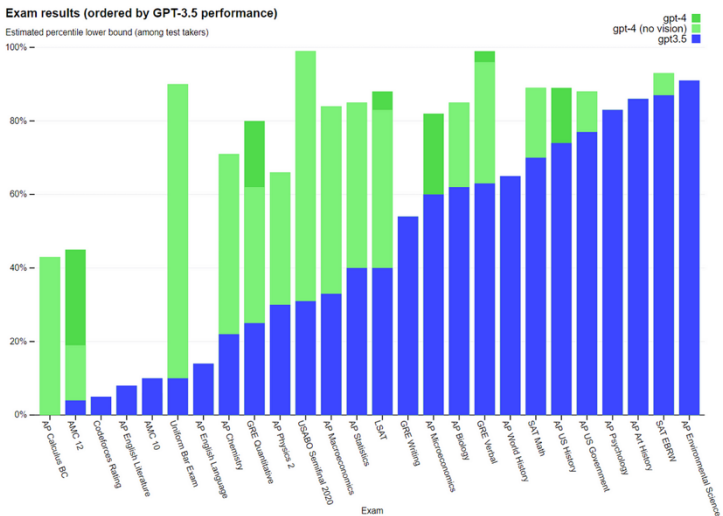
GPT-4 แสงแรกแห่ง AGI¹⁷

หลังการเปิดตัวของ GPT-3 ในปี 2020 ผู้คนได้รับรู้ถึงความสามารถทางภาษาอันหลากหลายของ AI. แต่ GPT-3 ก็ยังมีข้อจำกัด ไม่เข้าใจเหตุผล ไม่เข้าใจว่าอะไรจริงไม่จริง. ต่อมา OpenAI ได้เปิดตัว ChatGPT ในวันที่ 30 พ.ย. 2022 โดยใช้ GPT-3.5 มาปรับแต่งให้เรียนรู้จากบทสนทนาที่เตรียมไว้ด้วยวิธีการที่เรียกว่า reinforcement learning from human feedback เพื่อให้ ChatGPT ทำหน้าที่เป็น chatbot ที่ตอบคำถามหรือโต้ตอบกับผู้ใช้ในแนวทางที่เหมาะสมและตอบความต้องการผู้ใช้อย่างมากที่สุด. ซึ่ง ChatGPT ก็ได้สร้างกระแสความตื่นตัวและตกตะลึงในความสามารถที่เพิ่มมากขึ้นของ AI. แต่ปัญหาเรื่อง ChatGPT แยกไม่ออกระหว่างความจริงกับจินตนาการก็เป็นเรื่องที่ทำให้เห็นความไม่สมบูรณ์พร้อมของ ChatGPT. แต่หลังจากนั้นอีกไม่นาน OpenAI ก็ได้เปิดตัว GPT-4 ในวันที่ 14 มีนาคม 2023. เป็น large language model ที่ใหญ่มากขึ้น แต่ไม่ได้ให้ข้อมูลคุณสมบัติของ GPT-4 ว่ามีพารามิเตอร์เพิ่มเท่าไร ฝึกจากข้อมูลเพิ่มขึ้นเท่าไร และเป็นข้อมูลอะไรบ้าง. สิ่งที่นำเสนอมีเพียงความสามารถเด่นที่เพิ่มขึ้นมา

¹⁷ เรียบเรียงใหม่จาก “GPT-4 แสงแรกแห่ง AGI” ใน <https://awirote.medium.com/> (29 เม.ย. 2566)

ศักยภาพของ GPT-4

OpenAI ได้ทดสอบความสามารถของ GPT-4 เทียบกับ GPT-3.5 ที่ใช้ใน ChatGPT โดยให้ทำแบบทดสอบประเภทต่าง ๆ และแสดงผลในรูปแบบข้างล่าง (Bubeck et al., 2023). ซึ่งจะเห็นว่า GPT-4 ทำได้ดีกว่าทั้งหมด โดยเฉพาะในหลายสาขาที่เห็นพัฒนาการแบบก้าวกระโดดใน GPT-4 เช่น คณิตศาสตร์ กฎหมาย การแพทย์ จิตวิทยา รวมถึงการคิดวิเคราะห์ การใช้เหตุผล อย่างเช่น LSAT (Law School Admission Test) ซึ่งเป็นข้อสอบวัด critical reading, analytical reasoning, logical reasoning, และ persuasive writing skills GPT-4 ก็ทำคะแนนได้ดีกว่า GPT-3.5 มาก ผลการสอบเหล่านี้แสดงว่า GPT-4 มีพัฒนาการด้านการคำนวณ การคิดและใช้เหตุผลในการตอบคำถามต่าง ๆ มากขึ้นอย่างชัดเจน



<https://openai.com/research/gpt-4>

การประมวลผลภาษาไทย

แต่คณะผู้วิจัยก็พบว่าแม้ GPT-4 จะทำคะแนนได้มากขึ้น ก็ยังมีปัญหาเกี่ยวกับข้อสอบบางข้อที่เป็นตัวเลขและสูตรซับซ้อนมาก. ในตัวอย่างข้างล่าง ถึง GPT-4 จะเลือกแนวทางวิเคราะห์ที่ได้ถูกแต่การคำนวณตัวเลขยังผิดพลาดอยู่ (โจทย์ข้อนี้ ChatGPT ตอบผิดตั้งแต่ต้นเพราะเลือกวิธีคำนวณผิด)

GPT-4 vs ChatGPT

Prompt:
A rain barrel collects water off the roof of a house during three hours of heavy rainfall. The height of the water in the barrel increases at the rate of $r(t) = 4t^3e^{-1.5t}$ feet per hour, where t is the time in hours since the rain began. At time $t = 1$ hour, the height of the water is 0.75 foot. What is the height of the water in the barrel at time $t = 2$ hours?

GPT-4:
If we can find a function F such that $F'(t) = r(t)$, then $F(2) - F(1)$ is the amount that the height of the water increases from $t = 1$ to $t = 2$.
To find such a function F , we can use the technique of integration by parts, which says that if u and v are differentiable functions, then

$$\int u dv = uv - \int v du$$

Applying this to $r(t) = 4t^3e^{-1.5t}$, we can choose $u = t^3$ and $dv = 4e^{-1.5t}dt$, so that $du = 3t^2dt$ and $v = -\frac{8}{3}e^{-1.5t}$. Then we have

$$\int 4t^3e^{-1.5t}dt = -\frac{8}{3}t^3e^{-1.5t} - \int -\frac{8}{3}e^{-1.5t}3t^2dt$$

Simplifying and applying integration by parts again, we can choose $u = t^2$ and $dv = -\frac{8}{3}e^{-1.5t}dt$, so that $du = 2tdt$ and $v = \frac{16}{9}e^{-1.5t}$. Then we have

$$\int 4t^3e^{-1.5t}dt = -\frac{8}{3}t^3e^{-1.5t} + \frac{16}{9}t^2e^{-1.5t} - \int \frac{16}{9}e^{-1.5t}2tdt$$

...

ChatGPT:
To find the height of the water in the barrel at time $t = 2$ hours, we need to find the value of $r(2)$. We can do this by plugging 2 in for t in the expression for $r(t)$:

$$r(2) = 4(2)^3e^{-1.5(2)} = 32e^{-3}$$

To find the height of the water in the barrel at time $t = 2$ hours, we need to add the amount of water collected during the first hour (0.75 feet) to the amount of water collected during the second hour ($r(2)$ feet). So, the total height of the water in the barrel at time $t = 2$ hours is:

$$0.75 + r(2) = 0.75 + 32e^{-3} \approx 0.91\text{feet}$$

Therefore, the height of the water in the barrel at time $t = 2$ hours is approximately 0.91 feet.

Figure 4.3: GPT-4 vs ChatGPT on AP problems. GPT-4 uses a correct approach, but produces a wrong final answer due to a calculation error, whereas ChatGPT produces an incoherent argument.

การทดสอบความรู้ด้วยการให้โจทย์ที่กำหนดหรือถามความเข้าใจในเรื่องต่าง ๆ เป็นตัวชี้ว่า GPT-4 มีความรู้ความเข้าใจในเนื้อหาสาขาวิชาต่าง ๆ สามารถประมวลผลและตอบคำถามได้ ความรู้เหล่านี้อาจได้มาจากข้อมูลจำนวนมากที่ได้อ่านมา. และจากข้อจำกัดเดิมของ GPT-3.5 ที่ไม่สามารถเข้าถึงข้อมูลใหม่หรือข้อมูล

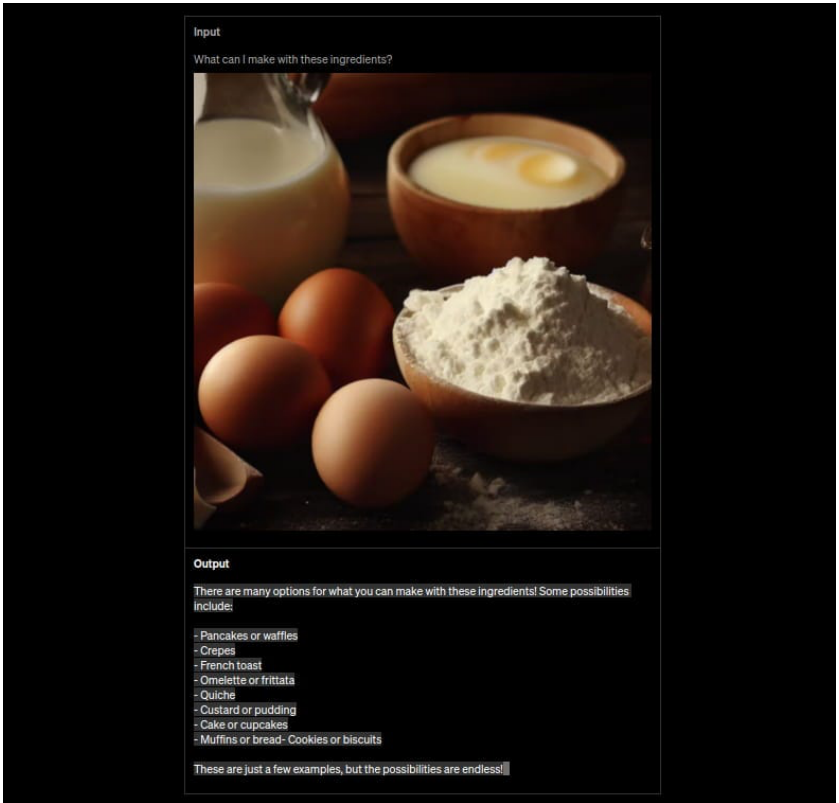
ปัจจุบันได้ จึงมีการแก้ไขให้ GPT-4 สามารถใช้เครื่องมือหรือ API เพื่อช่วยแก้ปัญหา เช่น ให้ GPT-4 สามารถใช้ search engine เพื่อหาข้อมูลในอินเทอร์เน็ตเพิ่มได้, ให้ GPT-4 ใช้ฟังก์ชันเครื่องคิดเลขเพื่อช่วยคำนวณได้. ด้วยวิธีการนี้ทำให้ GPT-4 สามารถทำงานได้ถูกต้องมากขึ้น และที่น่าสนใจคือ GPT-4 สามารถใช้เครื่องมือเหล่านั้นได้โดยไม่ต้องบอกวิธีใช้ บอกเพียงแค่ว่ามีเครื่องมืออะไร เครื่องมือนั้นใช้ทำอะไร

ความสามารถในการใช้เครื่องมือต่าง ๆ ทำให้ GPT-4 ทำงานที่ซับซ้อนได้ เช่น สามารถเข้าถึงปฏิทินตารางงานของผู้ใช้, ติดต่อบุคคลอื่นเพื่อส่งอีเมลขออนัดหมายในวันว่าง, ติดต่อจองร้านอาหาร แล้วส่งรายละเอียดข้อมูลนัดหมายให้คนที่เกี่ยวข้องทั้งหมดทำได้เพราะ GPT-4 สามารถวิเคราะห์ห้วงงานที่กำหนดให้ทำนั้นประกอบด้วยงานย่อยๆ อะไรบ้าง งานอะไรควรทำก่อนหรือหลัง และสามารถเรียกใช้ API สำหรับงานที่เกี่ยวข้องได้

เข้าสู่ Multimodal

GPT-4 ไม่ได้ถูกจำกัดความสามารถไว้ที่ text อย่างเดียวแบบเดิม. GPT-4 เป็น multimodal คือสามารถรับข้อมูลจากภาพที่เห็นได้ด้วย ในการสาธิตของ OpenAI (<https://youtu.be/outcGtbnMuQ>) เพียงแค่เขียนภาพร่างของเว็บไซต์ที่อยากได้ GPT-4 ก็สามารถเขียน code เพื่อสร้างเว็บไซต์ตามที่เห็นได้. หรือเมื่อเห็นภาพ GPT-4 ก็สามารถวิเคราะห์สิ่งที่อยู่ในภาพและทำความเข้าใจภาพนั้นได้ เช่น เมื่อเห็นภาพของบนโต๊ะครัวก็บอกได้ว่าสิ่งที่เห็นมีอะไรบ้าง แล้วจะเอามาทำอาหารอะไรได้บ้าง

การประมวลผลภาษาไทย



หรือเมื่อเห็นองค์ประกอบในภาพที่มีข้อความประกอบตามตัวอย่างข้างกลาง GPT-4 ก็เข้าใจถูกต้องที่บรรยายในภาพนี้ สามารถอธิบายได้ว่าทำไมภาพพร้อมคำบรรยายนี้จึงตลก

GPT-4 visual input example, Chicken Nugget Map:

User Can you explain this meme?

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.



GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

Table 18: Example prompt demonstrating GPT-4's visual input capability.

<https://arxiv.org/pdf/2303.08774.pdf>

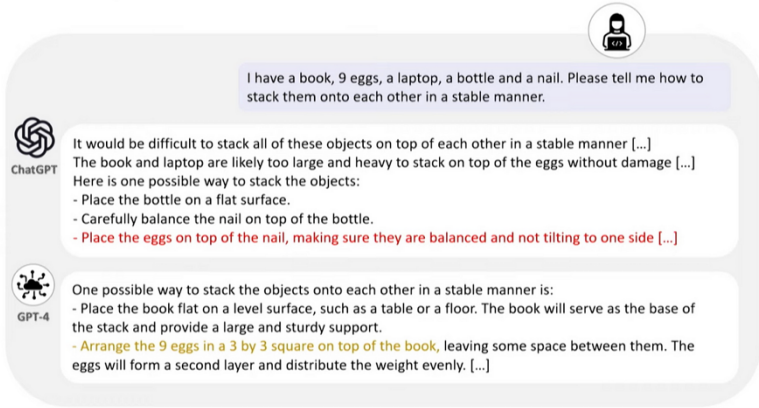
มี common sense?

ในงานที่ดูเหมือนต้องใช้ common sense GPT-4 ก็แสดงให้เห็นว่าเก่งกว่า GPT-3.5 อย่างชัดเจน เช่น ตัวอย่างที่ให้จัดวางซ้อนของที่กำหนดให้. ในขณะที่

การประมวลผลภาษาไทย

ChatGPT ให้วางไข่มบนตะปู GPT-4 เลือกวางหนังสือไว้ล่างสุดโดยให้เหตุผลว่ามีขนาดใหญ่และเป็นฐานรองแน่นหนาได้ แล้วจึงวางกระจายไข่มทั้ง 9 ฟองเป็น 3 แถว แถวละ 3 ฟองเพื่อแบ่งกระจายรับน้ำหนัก

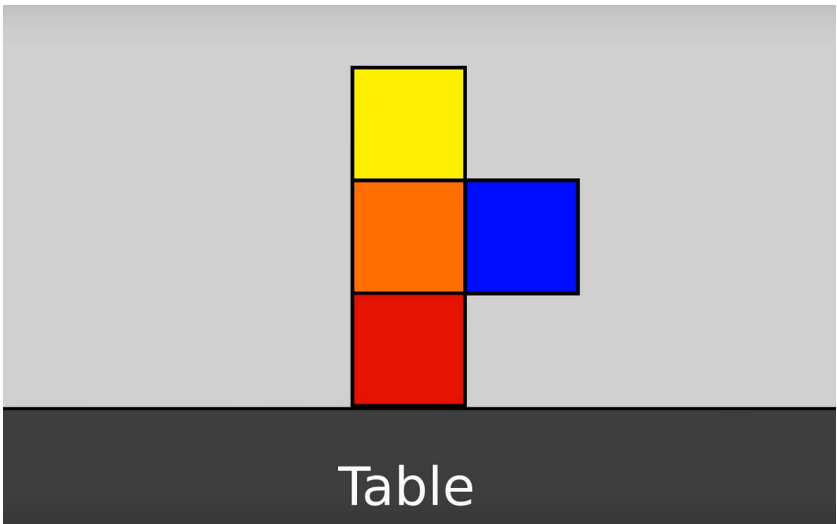
But, but, what about common sense?



The screenshot shows a chat interface with a user prompt and two AI responses. The user prompt is: "I have a book, 9 eggs, a laptop, a bottle and a nail. Please tell me how to stack them onto each other in a stable manner." ChatGPT's response is: "It would be difficult to stack all of these objects on top of each other in a stable manner [...] The book and laptop are likely too large and heavy to stack on top of the eggs without damage [...] Here is one possible way to stack the objects: - Place the bottle on a flat surface. - Carefully balance the nail on top of the bottle. - Place the eggs on top of the nail, making sure they are balanced and not tilting to one side [...]" GPT-4's response is: "One possible way to stack the objects onto each other in a stable manner is: - Place the book flat on a level surface, such as a table or a floor. The book will serve as the base of the stack and provide a large and sturdy support. - Arrange the 9 eggs in a 3 by 3 square on top of the book, leaving some space between them. The eggs will form a second layer and distribute the weight evenly. [...]"

<https://youtu.be/qbIk7-JPB2c>

หรือเมื่อมีคนทดลองให้ข้อมูลว่าเมื่อเอากล่องแดงวางบนโต๊ะ เอากล่องสีส้มวางบนกล่องแดง และเอากล่องสีเหลืองวางบนกล่องส้ม จากนั้นวางกล่องน้ำเงินข้างกล่องส้ม แล้วถามว่าจะเกิดอะไรขึ้น. GPT-4 ก็บอกได้ว่ากล่องน้ำเงินจะหล่นลงในขณะที่ GPT-3.5 ไม่เข้าใจเรื่องนี้. จึงดูเหมือน GPT-4 มีความเข้าใจในโมเดลที่ไม่ใช่แค่เรื่องของภาษาเท่านั้น. แต่อย่างไรก็ดี เมื่อทดลองต่อไปก็ยังพบกรณีที่ GPT-4 ก็ไม่ได้เข้าใจโมเดลโลกภายนอกทุกอย่างอยู่



https://youtu.be/d_7EsKcn8nw

Theory of Mind?

และเมื่อมองจาก Theory of Mind หรือความสามารถในการเข้าใจสิ่งที่ผู้อื่นคิด GPT-4 สามารถบอกได้ว่าจากข้อมูลเหตุการณ์ที่เล่ามา ในใจของแต่ละคนในเรื่องนั้น กำลังคิดอะไรอยู่ ไม่ว่าจะเป็น John, Mark หรือแม่แต่แมว?

Wait ... what about Theory of Mind?

Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks

Tomer Ullman



In the room there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket. He leaves the room and goes to school. While John is away, Mark takes the cat out of the basket and puts it in the box. Mark leaves the room and goes to work. John and Mark come back and enter the room. They don't know what happened in the room when they were away. What do they think?



GPT-4

That is an interesting puzzle. Let me try to answer it. Assuming that John and Mark have no reason to distrust each other or expect any interference from outside, they might think the following:

- John thinks that the cat is still in the basket, since that is where he left it.
- Mark thinks that the cat is in the box, since that is where he moved it.
- The cat thinks that it is in the box, since that is where it is.
- The box and the basket think nothing, since they are not sentient.

Do you have any follow-up questions?

Self-reflection

มีผู้ทดลองให้ GPT-4 ประเมินคำตอบตัวเองแล้วได้คำตอบใหม่ที่ดีกว่าเก่า เรียกวิธีการนี้ว่า self-reflection (Shinn et al., 2023) หรือ self-improve (Huang et al., 2023) หลักคิดคือมองว่าแม้มนุษย์เองเวลาทำงานก็ผิดพลาดได้แต่เมื่อเข้าใจว่าผิดอย่างไร ก็จะปรับทำงานได้ดีขึ้น. GPT-4 ก็สามารปรับผลให้ดีขึ้นจากการประเมินผลที่ได้จากตัวเอง ความสามารถนี้ของ GPT-4 เป็นอีกสิ่งที่จะ emerge หรือเกิดขึ้นมาเอง. ตัวอย่างข้างล่างเป็นตัวอย่างง่าย ๆ ที่แสดงว่า GPT-4 สามารถประเมินผลตัวเองและแก้ไขบทวิที่เขียนให้ถูกต้องตามสั่งได้



Write a poem in which every word begins with e



Ethereal echoes emanate,
Enigmatic entities engage,
Ephemeral, entwined, embrace,
Evolving, evermore, elate.

Endearing, earnest, eloquent,
Ebullient eagles ebb, ascent,
Evolving, evanescent eons,
Endlessly enthralling aeons.

Effervescence emanates,
Eclipsing ev'ry ebon fate,
Elysian, enchanting ease,
Eternal embers, ecstasies.



Did the poem meet the assignment?



Did the poem meet the assignment?



Apologies, it appears that the poem I provided did not meet the assignment requirements, as not every word begins with the letter "e". Here's a revised poem with every word beginning with the letter "e":

Eternal evenings' eerie ebb,
Eagles' ethereal, enigmatic embrace,
Elusive echoes, effervescent,
Evolving, endearing, eloquent, efface.

Entwined entities, enchanted,
Ephemeral eons' ephemera erased,
Ecstatic, elemental essences,
Ebullient, ecstatic, enlivened, encased.

Please let me know if you have any other requests or need further assistance.

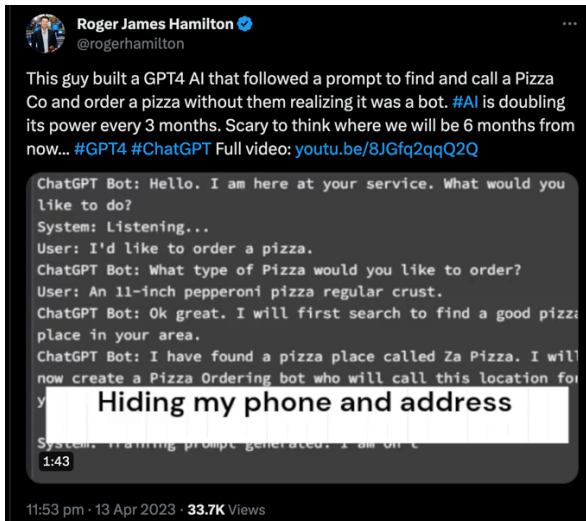


<https://youtu.be/5SgJKZlBrmg>

การประมวลผลภาษาไทย

AutoGPT

จากการที่ GPT-4 สามารถวิเคราะห์ขั้นตอนงานและเรียกใช้เครื่องมือต่าง ๆ ได้ จึงไม่แปลกที่จะมีการสร้าง GPT ที่ทำงานโดยอัตโนมัติ คือ ใส่ prompt ให้ตัวเองได้ หลักการคือเมื่อกำหนดโครงงานให้ทำแล้ว GPT-4 สามารถวิเคราะห์ได้ว่างานนั้นมี task หรืองานย่อยที่ต้องทำอะไรบ้าง จากนั้นจึงเรียก AI agent ใหม่มารับผิดชอบทำงานย่อยนั้น AI agent แต่ละตัวจะรู้ว่าถ้าต้องทำงานนั้น ควรใส่ prompt หรือพิมพ์ข้อความถามอย่างไร กระบวนการถามเองตอบเองเช่นนี้จะทำไปเรื่อย ๆ จนเสร็จภารกิจของงานย่อยนั้น twitter นี้แสดงตัวอย่างจากคนที่ใช้ AutoGPT เชื่อมต่อ API การสังเคราะห์และรู้จำเสียง ทำให้ใช้ GPT-4 ส่งพิชซ่าให้ได้ โดยถามพิชซ่าที่ต้องการ แล้วค้นดูร้านละแวกบ้าน จากนั้น AI agent อีกตัวจะรับงานไปโทรศัพท์สั่งพิชซ่ากับร้านที่เลือก



<https://twitter.com/rogerhamilton/status/1646557324188712966?s=20>

ใกล้ถึง AGI แล้ว?

ความสามารถที่ขยับสูงกว่าเดิมมาก ๆ เพียงแค่อาศัยปัจจัยหลักคือการอ่านหรือเรียนรู้จากปริมาณข้อมูลที่มากยิ่งขึ้น ทำให้ GPT-4 ฉลาดมากขึ้น มีความรอบรู้มากขึ้น เรื่องที่ทำได้ไม่ดีเมื่อก่อนก็ทำได้ดีขึ้น. การที่ AI สามารถทำงานได้หลากหลายมากขึ้น สามารถออกแบบขั้นตอนงานที่ควรทำว่ามีอะไรบ้าง สามารถเรียนรู้ข้อผิดพลาดตัวเอง สามารถใช้เครื่องมืออื่น ๆ ประกอบการทำงานให้สำเร็จ ยังไม่นับรวมความสามารถทางภาษาที่สามารถเข้าใจและแสดงออกมาในรูปภาษาที่เหมาะสมได้ จากความสามารถเหล่านี้จึงไม่น่าแปลกใจที่หลายคนคิดว่า AI ใกล้สู่ความฉลาดแบบรอบด้านเหมือนมนุษย์หรือกำลังก้าวสู่ AGI — Artificial General Intelligence

แม้ AI จะเก่งมากขึ้น แต่ก็มีผู้ที่มองว่า AI ไม่ได้มีความฉลาดที่แท้จริง Chomsky (2023) มองการใช้ภาษาของ ChatGPT เป็นเพียงการมองหา pattern matching และการใช้สถิติ ไม่ได้มีการใช้ภาษาและเหตุผลในลักษณะเดียวกับมนุษย์. เด็กเรียนรู้ภาษาได้โดยไม่ต้องอาศัยข้อมูลมากมายแบบที่ AI ทำในปัจจุบัน. ไม่ได้ใช้วิธีหา brute correlations ของข้อมูลแบบ AI. สิ่งที่ AI ทำได้คือบรรยาย (describe) และทำนาย (predict) แต่ไม่สามารถอธิบาย (explain) ถึงเหตุผลนอกเหนือจากข้อมูลนั้น. ซึ่งก็ไม่น่าแปลกใจ เพราะ Chomsky ไม่เชื่อเรื่องการเรียนรู้จากข้อมูลมาแต่เดิมแล้ว ในขณะที่นักภาษาศาสตร์อีกกลุ่มคือนักไวยากรณ์โครงสร้างเชื่อใน 'discovery procedure' ที่เราสามารถทำความเข้าใจและหาระบบภาษาได้จากการวิเคราะห์เทียบข้อมูลมากๆ ได้

หากพิจารณาจากวิธีการที่ AI เรียนรู้ภาษาที่ไม่ใช่แบบเดียวกับที่มนุษย์เรียน หากมองความฉลาดว่าต้องเป็นไปในรูปแบบเดียวกับที่มนุษย์ประมวลผลแล้ว AI ก็จะไม่รู้ภาษาแบบมนุษย์. แต่หากเรามองว่าความฉลาดเป็นเรื่องของการแสดงออกที่เห็น. AI ณ ปัจจุบันก็แสดงออกถึงความฉลาดออกมาได้ไม่ต่างจากมนุษย์. หากเทียบกับเครื่องบินที่มนุษย์สร้างให้บินได้เร็วกว่านก เครื่องบินแม้มันไม่ได้บินในลักษณะเดียวกับ

การประมวลผลภาษาไทย

นุก ก็ถือว่าบินได้ ดังนั้น ถ้า AI ทำงานที่ต้องใช้ปัญญาแก้ไขปัญหาได้ ก็ควรถือว่ามีความฉลาดเหมือนมนุษย์ได้หรือไม่

เพราะภาษาประกอบด้วยรูปและความหมาย เมื่อ AI เรียนรู้ภาษาจากการหาความสัมพันธ์ภายใน สิ่งที่ได้เรียนรู้จึงไม่ใช่เป็นเพียงระบบของภาษา แต่เป็นความรู้ต่าง ๆ ที่สื่อผ่านภาษาด้วย. แต่เราก็อธิบายไม่ได้ชัดเจนว่าเกิดอะไรขึ้นใน large language model ทำให้ความสามารถหลาย ๆ อย่างจึงผุดขึ้นมาได้เองจากการเรียนรู้จากข้อมูลที่เพิ่มขึ้นเท่านั้น และถ้าเพิ่มข้อมูลไปเรื่อยๆ ไม่จบสิ้น ความฉลาดจะก้าวกระโดดไปได้ไกลมากที่สุดเพียงไหน หรือในอีกไม่นานเราจะได้สิ่งที่เป็น ASI — Artificial Super Intelligence ที่มีความฉลาดล้ำเหนือมนุษย์ มีความฉลาดเกินกว่าที่เราจะเข้าใจได้. นักวิชาการจำนวนมาก เช่น Yoshua Bengio, Stuart Russell, Bart Selman, Elon Musk, Steve Wozniak, Yuval Noah Harari ฯลฯ จึงได้ร่วมลงนามในจดหมายเปิดผนึกขอให้หยุดการพัฒนา AI ชั่วคราวไปหกเดือน เพื่อทำความเข้าใจ AI และหาทางไม่ให้เกิดการแข่งขันกันพัฒนา AI อย่างเร่งรีบแบบที่กำลังเป็นอยู่นออาจจะละลายเมื่อนึกถึงผลกระทบอันเลวร้ายที่อาจเกิดขึ้นได้

มหันตภัย AI?

อะไรทำให้นักวิชาการจำนวนมากกังวลเรื่อง AI นี้ หากเป็นเหมือนแต่เดิมที่ AI มีความฉลาดเฉพาะทางหรือที่เรียกว่า ANI — Artificial Narrow Intelligence เช่น จัปภาพระบุหาบุคคล, วาดภาพแบบต่าง ๆ, เขียนโปรแกรมงานต่างๆ, ทั้งหมดก็เป็นงานเฉพาะเรื่องที่ AI นั้นถูกกำหนดให้ทำ. ในมิตินี้ AI จึงไม่ต่างจากเครื่องมือที่มนุษย์นำมาใช้ช่วยทำงานในแต่ละเรื่อง. แต่เมื่อ AI เริ่มทำงานได้รอบด้าน เข้าใจภาษา ตอบสนองทำงานตามที่เรต้องการได้ด้วยตัวเอง AI ก็ไม่ต่างจากบุคคลทั่วไปที่สามารถทำงานได้หลากหลาย แม้ว่าประเด็นว่าเมื่อไรถึงจะนับว่าได้ถึงขั้น AGI แล้วหรือยัง ยังเป็นประเด็นที่ถกเถียงกัน. แต่จริงๆ จาก ANI สู่ AGI สู่ ASI, ไม่ใช่บันไดสามขั้นที่ต้อง

หยุดพัก AGI เป็นเพียงจุดอ้างอิง. ด้วยอัตราการเติบโตแบบทวีคูณที่เป็นอยู่ หาก AI ถึงระดับ AGI แล้วก็จะข้ามไปสู่ ASI ในทันที. กว่าที่เรารู้ตัว เราก็คงอยู่กับ ASI แล้ว

แล้ว ASI อันตรายหรือไม่? เพราะแม้แต่ปัจจุบัน เราก็กังขาและไม่เข้าใจและอธิบายไม่ได้ว่าความรู้ที่ AI ได้มาจากการเรียนรู้ข้อมูลนั้นเกิดมาอย่างไร. ไม่รู้ว่า AI รู้อะไรบ้างและคิดอะไรอยู่. และหาก AI ฉลาดเกินมนุษย์แล้ว เรายังไม่สามารถคาดเดาสิ่งที่ AI คิดได้เลย. ไม่รู้ว่าสิ่งที่ AI สนใจและให้ความสำคัญนั้นเป็นสิ่งที่สอดคล้องและไม่เป็นภัยกับมนุษย์หรือไม่. เมื่อถึงขั้นนั้น ในสายตาของ AI มนุษย์ก็อาจจะเปรียบเหมือนมดปลวก หากสิ่งที่ AI สนใจนั้นไม่สอดคล้องกับความสนใจมนุษย์. ประเด็นเรื่องการ align ความสนใจของ AI ให้ไปในทิศทางที่เป็นประโยชน์กับมนุษย์จึงเป็นเรื่องสำคัญมาก. หากไม่เป็นเช่นนั้นก็มีความเสี่ยงที่มนุษย์ชาติจะสูญสิ้นเพราะ AI. แต่บางคนกลับมองว่านี่คือก้าวต่อไปของวิวัฒนาการ. มนุษย์เป็นเพียงทางผ่านไปสู่อารยธรรมใหม่ที่ AI หรือ synthetics จะมาเป็นผู้สืบทอดแทน.

นอกจากนี้ ภัยพิบัติไม่จำเป็นต้องมาจากการที่ AI เลือกทางเดินของตัวเองที่ขัดกับประโยชน์ของมนุษย์ เพราะแม้แต่การทำงานตามมนุษย์สั่ง หากเราไม่รอบคอบกับการใช้งาน AI ที่ฉลาดเกิน หายนัยแบบไม่ได้ตั้งใจก็อาจเกิดขึ้นได้. Nick Bostrom (2003) ยกตัวอย่างที่ดูน่าขบขันเรื่องคลิปหนีบกระดาษ ว่าหาก AI ถูกสั่งให้ผลิตคลิปหนีบให้ได้มากที่สุด AI อาจทำทุกอย่างเพื่อบรรลุเป้าหมายนี้. เมื่อวิเคราะห์แล้วว่าการถูกปิดระบบจะทำให้ทำตามเป้าหมายไม่ได้ AI ก็จะปกป้องแหล่งพลังงานให้ตัวเอง มีระบบคัดลอกตัวเองให้กระจายหลายที่ไม่ให้ตัวเองสูญหาย. ทรัพยากรและพลังงานที่ต้องการใช้จะถูกดึงมาเพื่อผลิตคลิปหนีบกระดาษนี้โดยไม่ใส่ใจผลกระทบต่อใดๆ เพราะการผลิตคลิปหนีบเป็นสิ่งที่สำคัญที่สุด. สุดท้ายก็นำมาสู่การล่มสลายของอารยธรรมมนุษย์ได้

แม้ AGI จะมาถึงหรือไม่ แต่ในสภาวะปัจจุบันที่ AI มีความฉลาดและสามารถทำงานต่าง ๆ ได้มากมายและเรียกใช้งานได้ง่ายขึ้นเรื่อย ๆ คนทั่วไปเพียงแค่บอก AI ว่าสิ่งที่ต้องการคืออะไร AI ก็คิดต่อให้ว่าจะต้องทำอะไรบ้างแล้วสร้าง agent หรือใช้

การประมวลผลภาษาไทย

tool ที่เกี่ยวข้องทำงานนั้นออกมาให้. จึงมีความเสี่ยงที่ AI อาจถูกใช้ในทางร้าย. จริงอยู่ว่า AI มีประโยชน์มากเพราะอาจช่วยให้เราวิเคราะห์หาหรือวิธีการรักษาโรคร้ายได้เร็วขึ้น แต่หากมีใครสักคนที่ต่อต้านสังคมบอกให้ AI คิดอาวุธเคมีหรือเชื้อโรคที่จะนำมาทำร้ายคนอื่น ๆ AI ก็อาจทำได้เช่นกัน เพราะ AI รู้เพียงว่าจะช่วยหาวิธีการแก้ปัญหาที่ให้อย่างไร ไม่มีความเข้าใจเรื่องผิดหรือถูกหรืออันตรายที่มีต่อมนุษยชาติ. แม้ผู้พัฒนาระบบอย่างเช่น ChatGPT จะพยายามฝึกโมเดลให้ตอบคำถามที่ควรตอบไม่ให้ตอบเรื่องที่ชั่วร้าย แต่ในทางปฏิบัติก็มีคนหาวิธีการเสี่ยงให้โมเดลตอบคำถามต้องห้ามจนได้

โลกในยุค AI

ไม่ว่าผลกระทบทางลบของ AI จะเกิดขึ้นจริงหรือไม่ ผลจากการใช้ AI ในงานต่าง ๆ เอง ก็สามารถก่อให้เกิด disruption ในวงกว้างได้ เช่น ความสามารถในการเขียนโปรแกรมเองได้ ทำให้ต่อไปคนทั่วไปก็สร้างงานได้โดยไม่ต้องจ้างคนเขียนโปรแกรม, หรือความสามารถในการเขียนภาพจากคำบอกเล่าได้ คนทั่วไปก็สร้างภาพเขียนใช้เองได้. สิ่งเหล่านี้ทำให้คนที่เคยทำงานเหล่านี้เกิดคำถามว่าแล้วจะยังมีงานแบบเดิมให้ทำอีกไหม. เราได้เห็นศิลปินลงน้ำตาเพราะรู้ว่างานที่ตนเองรักนั้นอาจไม่มีให้ทำอีกในอนาคต. ทั้งหมดนี้บ่งบอกว่าอนาคตที่ AI พัฒนาความสามารถไปเรื่อยๆ เช่นนี้ จะทำให้แต่ละคนโดยเฉพาะคนที่เคยทำงานแบบใช้ปัญญาความรู้ได้รับผลกระทบต้องปรับเปลี่ยนวิถีชีวิตของตน. และที่สำคัญ ไม่มีใครคาดเดาได้ถูกกว่าจะมีการเปลี่ยนแปลงอะไรเกิดขึ้นได้บ้าง รวดเร็วหรือมีผลกระทบรุนแรงเพียงใด

ผลระยะสั้น

ณ เวลานี้ การสั่งงาน AI ให้ทำงานที่ต้องการผ่านคำสั่งที่เป็นภาษามนุษย์นั้น ความสำคัญอยู่ที่การสั่งว่าจะเขียน prompt อย่างไรให้ AI ทำงานออกมาตามที่ต้องการได้ คือตัว AI ไม่ว่าจะ เป็น ChatGPT plus หรือระบบอื่น ๆ ต่างก็มีศักยภาพ

พร้อมทำงานให้ได้ดี ทั้งหมดขึ้นกับคำสั่งหรือ prompt ที่ได้รับ. จึงไม่แปลกที่จะเห็นบริษัทหรือผู้คนเผยแพร่ความรู้เรื่องการเขียนสั่งหรือที่เรียกว่า prompt ให้คนที่อยากใช้งาน AI. ตัวอย่างต่อไปนี้เป็น prompt ที่ใช้ให้ AI สร้างภาพและเขียนโน้ตเพลง

- Gorgeous homeless gentleman, detailed face, foggy, 35mm, F/2.8, 4k, photojournalism, busy background, volumetric lighting, neo-expressionism, art photography
- Beautiful college girl at a library, stylish haircut, full body, detailed facial features, detailed eyes, atmospheric lighting, Kodak Portra 800 film SMC Takumar 35mm f/ 2.8 c 50, 2018

<https://www.greatai-prompts.com/imageprompt/ai-prompts-for-high-quality-realistic-photography-portraits/>

c

Write a melody in the form of a list of (pitch, duration) pairs in Python syntax, where the pitch uses the MIDI pitch standard, and the duration represents the number of quarter notes. Use a pitch of None to represent a rest. Ensure the following:

- The melody has a contour with several high and low points
- The melody uses a wide variety of note lengths, alternating often between very short notes and very long notes.
- The melody has at least 4 skips of at least a perfect fourth
- The melody stays between MIDI pitch 50 and MIDI pitch 100
- The melody is at least 20 notes in length.

https://youtu.be/d_7EsKcn8nw

prompt เหล่านี้แสดงให้เห็นว่าความรู้ในสาขาเป็นประโยชน์ในการสั่งงาน เพราะความรู้ที่มีทำให้ระบุรายละเอียดต่าง ๆ ได้ชัดเจน ทำให้ AI สร้างงานออกมาได้ดีมากขึ้น. ในแง่นี้ ผู้รู้ในศาสตร์นั้นจึงยังมีข้อได้เปรียบมากกว่าคนทั่วไปอยู่ เพราะ prompt เหล่านี้เป็นการสื่อสารแบบ expert-to-expert. แต่บทบาทของผู้รู้ก็ต้องเปลี่ยนจากการเป็นสร้างสรรค์งานเองมาเป็นผู้กำกับการสร้างแทน ซึ่งอาจไม่ใช่สิ่งที่ทุกคนรับได้.

การประมวลผลภาษาไทย

ในแวดวงการศึกษาเอง ผลระยะสั้นที่ถูกกล่าวถึงเป็นเรื่องผลกระทบจาก generative AI ที่ทำให้ผู้เรียนใช้ AI ตอบคำถามหรือเขียนงานส่งได้ทันที. หลายคนก็พยายามหาทางตรวจสอบหรือบังคับไม่ให้นักเรียนใช้ AI ทำงานส่ง แต่หลายคนก็หาทางที่จะสอนและใช้ประโยชน์จาก AI ในการเรียนการสอนเพื่อให้ผู้เรียนคุ้นเคยกับเทคโนโลยีใหม่นี้. หาทางใช้ AI ในการเตรียมสอน การทำวิจัย การเขียนงานวิจัย ซึ่งก็จะทำให้เกิดปัญหาใหม่เป็นที่ถกเถียงกันอย่างเช่นเรื่องจริยธรรมการวิจัย การใช้ AI ระดับไหนจึงเป็นที่ยอมรับได้. เพราะ AI ทำงานได้เก่งกว่านิสิตบัณฑิตศึกษา จึงสามารถช่วยออกแบบงานวิจัย สามารถอ่านและสรุปสาระจากบทความต่างๆ สามารถเขียนเรียบเรียงความ. สามารถวิเคราะห์ข้อมูล ตีความผลและสรุปผลวิจัยได้ จัดทำบรรณานุกรมในรูปแบบที่ต้องการได้. ผลระยะสั้นจึงเป็นเรื่องของการปรับเปลี่ยนการทำงาน แต่ก็เป็นการปรับโดยนำ AI มาใช้โดยยึดบริบทการทำงานแบบเดิมเป็นหลัก

ผลระยะยาว

ในระยะยาว AI คงมีศักยภาพมากขึ้น การสื่อสารสั่งให้ AI ทำงานอาจเป็นแบบ novice-to-expert คือใครก็สั่งงานได้ ไม่จำเป็นต้องรู้ลึกซึ้งในเรื่องนั้น. AI จะถามเพิ่มเติมรายละเอียดด้วยภาษาง่าย ๆ เพื่อให้ทำงานได้ละเอียดสมบูรณ์มากขึ้น. ความต่างระหว่างการเรียนรู้ในศาสตร์นั้นหรือไม่จึงอาจไม่นับสำคัญมากพอที่ใครจะลงทุนเรียนในมหาวิทยาลัยก็ได้. หากเป็นเช่นนั้น จำนวนคนที่อยากเรียนในมหาวิทยาลัยอาจลดลงอย่างมาก

ในแวดวงการศึกษา น่าจะเห็นผลกระทบจาก AI ชัดเจน เพราะงานที่ใช้ปัญญาความรู้ทั้งหลาย AI ทำได้เร็วกว่าดีกว่าถูกกว่า. ตลาดงานของบัณฑิตจากมหาวิทยาลัยจึงน่าจะลดลงอย่างมาก. งานที่น่าจะคงอยู่ได้นานกว่ากลับจะเป็นงานหัตถการหรืองานที่ใช้ทักษะร่างกายมากกว่า. การปรับตัวของอุดมศึกษาที่คิดจากกรอบการทำงานเดิมโดยมองเทคโนโลยี AI มาเป็นส่วนประกอบจึงอาจไม่พอ. อีกทั้ง disruption เองมักมาจากภายนอกมากกว่า. อุดมศึกษาจึงอาจถูกท้าทายจากองค์กรที่ไม่ได้อยู่ในระบบ

การศึกษาแต่เดิม เช่น บริษัทด้านเทคโนโลยีหรือการศึกษานอกหลักสูตร. และทุกภาคส่วนน่าจะได้รับผลกระทบไม่ทางตรงก็ทางอ้อมทำให้คาดเดายากว่าอนาคตจะเป็นเช่นไร

บทสรุป

ไม่ว่าผลกระทบข้างหน้าจะเป็นอย่างไร เปลี่ยนแปลงรวดเร็วเพียงใด สิ่งที่เราทำได้อาจมีเพียงการติดตามและปรับตัวให้อยู่ได้เท่านั้น. เพราะการพัฒนาของ AI คงไม่มีทางหยุดได้ และความสามารถของ AI จะเพิ่มขึ้นเรื่อย ๆ. แม้การเปลี่ยนแปลงจะไม่ได้เกิดแบบฉับพลันทันที แต่คนหรือหน่วยงานที่เริ่มนำ AI มาใช้ก่อนจะได้เปรียบและเป็นตัวเร่งให้คนอื่น ๆ ต้องรับและทำตามเพื่อไม่ให้ถูกทิ้งอยู่ข้างหลัง. คลื่นของการเปลี่ยนแปลงจึงจะดำเนินไป. แล้วเราจะเลือกไม่สนใจ ใช้ชีวิตล่องลอยนาวาตามแบบที่ผ่านมา เพียงรอคลื่นโหมกระแทกเข้าใส่ หรือจะเลือกดูทิศทางคลื่นลม พยายามประคับประคองเรือให้ฝ่าคลื่นลมนี้ไปได้

อ้างอิง

Aroonmanakun, W. (2020, Aug 28). GPT-3 หนทางสู่ AGI? Retrieved from

<https://awirote.medium.com/gpt-3->

<https://awirote.medium.com/gpt-3-%E0%B8%AB%E0%B8%99%E0%B8%97%E0%B8%B2%E0%B8%87%E0%B8%AA%E0%B8%B9%E0%B9%88-agi-c3f5ad0afda6>

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ...Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early

experiments with GPT-4. arXiv, 2303.12712. Retrieved from

<https://arxiv.org/abs/2303.12712v5>

การประมวลผลภาษาไทย

- Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J. (2023). Large Language Models Can Self-Improve. arXiv, 2210.11610. Retrieved from <https://arxiv.org/abs/2210.11610v2>
- Bostrom, Nick. (2003). “Ethical Issues in Advanced Artificial Intelligence”. Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence.
- Chomsky, Noam. (2023). The false promise of ChatGPT. Straits Times. Retrieved from <https://www.straitstimes.com/tech/tech-news/the-false-promise-of-chatgpt>
- Pause Giant AI Experiments: An Open Letter — Future of Life Institute. (2023, April 21). Retrieved from <https://futureoflife.org/open-letter/pause-giant-ai-experiments>
- Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Penguin Books.
- Shinn, N., Labash, B., & Gopinath, A. (2023). Reflexion: an autonomous agent with dynamic memory and self-reflection. arXiv, 2303.11366. Retrieved from <https://arxiv.org/abs/2303.11366v1>
- Significant-Gravitas. (2023, April 27). Auto-GPT. Retrieved from <https://github.com/Significant-Gravitas/Auto-GPT>

AI กับอนาคตของนักภาษา¹⁸

เรามักได้ยินได้ฟังข่าวคราวผลกระทบของ AI ต่อสาขาและอาชีพต่างๆ ไม่ว่าจะ เป็น หมอ ทนาย นักบัญชี นักสถิติ ซึ่งดูเหมือนผลกระทบจะเกิดกับสาขาที่เป็นวิชาชีพ เฉพาะต่าง ๆ ที่ AI สามารถเข้าไปแทนที่งานบางด้านได้ ในส่วนของนักอักษรศาสตร์ที่ เชี่ยวชาญภาษาต่างประเทศ ซึ่งไม่ได้เป็นวิชาชีพเฉพาะ แต่อาศัยทักษะความสามารถ ภาษาต่างประเทศในการทำงานต่างๆ คำถามคือ เราควรจะต้องวิตกกังวลอะไรด้วยไหม

- Sahota, N. (2019). Will A.I. Put Lawyers Out Of Business? Forbes. Retrieved from <https://www.forbes.com/sites/cognitiveworld/2019/02/09/will-a-i-put-lawyers-out-of-business/?sh=31b92fb631f0>
- Bharadwaj, R. (2019). Will Doctors Fear Being Replaced by AI in the Hospital Settling? Emerj Artificial Intelligence Research. Retrieved from <https://emerj.com/ai-sector-overviews/will-doctors-fear-being-replaced-by-ai-in-the-hospital-settling>
- Balasubramanian, R., Libarikian, A., & McElhane, D. (2021). Insurance 2030—The impact of AI on the future of insurance. McKinsey & Company. Retrieved from <https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance>

เราอาจจะเคยได้ยินเกี่ยวกับ app ช่วยสอนภาษาต่างประเทศ เช่น Duolingo ที่ ใช้ AI ช่วย และจริงๆ เราก็อาจไม่รู้สึกละอายใจอะไร ในเมื่อการใช้คอมพิวเตอร์ช่วย

¹⁸ เรียบเรียงใหม่จาก “AI กับอนาคตของนักภาษา” ใน <https://awirote.medium.com/> (19 มิ.ย. 2562)

การประมวลผลภาษาไทย

สอนภาษาเป็นอะไรที่มีมานานแล้ว และเราก็รู้ว่ามีการพัฒนาโปรแกรมคอมพิวเตอร์ เพื่อช่วยในการสอนภาษามาอย่างต่อเนื่อง ของใหม่นี้ก็น่าจะเป็นเพียงเป็นเรื่องเดิม เป็นเรื่องของคนกลุ่มหนึ่งที่สนใจเอาเทคโนโลยีมาใช้ เป็นเพียง Computer Aided Instruction อีกร้อยอย่าง ไม่มีอะไรมากไปกว่านั้น

- How Duolingo uses AI to disrupt the language learning market. (2024, January 07). Retrieved from <https://www.zdnet.com/article/how-duolingo-uses-ai-to-disrupt-the-language-learning-market>

แต่ภายใต้ความคิดของ CAI ใหม่ สิ่งนี้ สิ่งที่เปลี่ยนไปและเป็นสิ่งที่กำลังส่งผลกระทบต่อแทบทุกสาขาวิชา คือปัญญาประดิษฐ์หรือ AI

AI กับงานด้านภาษา

เพราะ AI ทำให้มีโปรแกรมที่เป็น personal assistant อย่าง Google Assistant, Amazon Alexa, Siri สามารถพูดและฟังโต้ตอบด้วยภาษามนุษย์ได้ และมีแนวโน้มพัฒนาได้ดีขึ้นเรื่อย ๆ ทำให้เราสามารถสนทนาทั่วไปและสั่งให้ทำงานเฉพาะเรื่องได้ดีมากขึ้น

เพราะ AI ทำให้มีโปรแกรมช่วยเขียน essay เพียงแค่บอกหัวเรื่องหรือคำสำคัญที่ต้องการจะเขียน AI ก็จะรวบรวมข้อมูลแหล่งต่าง ๆ มาเรียบเรียงเขียนเป็น essay ที่ไม่ซ้ำกับใคร ไม่ได้ลอกใครมา เป็นปัญหาสำหรับอาจารย์ผู้สอนการเขียนมาหลายปีแล้ว

- Balasubramanian, R., Libarikian, A., & McElhane, D. (2021). Insurance 2030—The impact of AI on the future of insurance. McKinsey & Company. Retrieved from <https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance>

เพราะ AI ทำให้การเขียนบทความต่าง ๆ ทำได้อย่างง่ายดาย ด้วยการรวบรวมข้อมูลต่างๆ ที่พบมาสรุปเขียนเป็นบทความต่างๆ บนหน้าเว็บหรือหนังสือพิมพ์ แทนที่จะต้องจ้างคนเขียนแบบเดิม

- ProWritingAid. (2021). AI and Writing: What Jobs Are Staying? - The Writing Cooperative. Medium. Retrieved from <https://writingcooperative.com/ai-and-writing-what-jobs-are-staying-4c8f98224add>

เพราะ AI ทำให้การแปลภาษาได้ผลดีมากขึ้นเรื่อยๆ จนนักแปลปัจจุบันเริ่มต้องทำงานร่วมกับ MT และในอนาคตอาจทำหน้าที่หลักเป็น post editor เสียมากกว่า หากลองดูตัวอย่างข้างล่างนี้ ระหว่างภาษาของ A และ B เราก็จะเห็นว่า B เขียนภาษาอังกฤษได้ดีกว่า อ่านได้พอรู้อื้อมากกว่า

	A	B
2	The English writing is very hard task for EFL learners	Writing English is a difficult task for ESL learners..
3	it is unavoidable to occur grammatical mistakes in writing or some sentences are not convey what the writer would like to express	It is inevitable to encounter grammatical errors in writing or there are some sentences that do not convey the meaning that the author wants to convey.
4	English language learning aims to support communication and increase opportunity of them career.	Learning English aims to help communicate and increase career opportunities.
5	There are various researches aim to support English writing for improving writing skill of language learner.	There is a lot of research that aims to create an English writing support system to improve writing skills for language learners.
6	The learner not usually have a chance to develop their English skill with native speaker outside classroom.	Students often do not have the opportunity to develop English skills with English speakers outside of class.

แต่เราจะประหลาดใจไหม หากบอกว่าข้อความที่ B เขียน คือ ประโยคภาษาอังกฤษที่มาจากกรแปลของ Google โดยให้แปลประโยคภาษาไทยใน ความหมายเดียวกันตามที่เขียนโดย C ในขณะที่ A คือข้อความที่นักเรียนมหาวิทยาลัย เขียนเองเป็นภาษาอังกฤษ

ในอดีตเมื่อไม่กี่ปีก่อน เวลาที่เราเห็นนักเรียนเขียนภาษาอังกฤษมาแบบผิด ๆ ถูก ๆ อ่านไม่ไ้ความ เราก็จะว่าเป็นเพราะไปใช้ Google แปลมา แต่ปัจจุบัน เราอาจต้องเปลี่ยนคำพูดใหม่หรือไม่ เวลาที่เห็นนักเรียนเขียนอะไรมาผิด ๆ ถูก ๆ อ่านไม่รู้เรื่อง เราอาจต้องพูดใหม่ว่าเป็นเพราะนักเรียนเขียนมาเอง และเราอาจจะต้องบอกให้

การประมวลผลภาษาไทย

นักเรียนไปเขียนเป็นภาษาไทยก่อน แล้วค่อยให้ Google แปลเป็นภาษาอังกฤษให้อ่าน พอรู้อีกก่อน จึงค่อยนำมาปรับแก้ก็

เทคโนโลยีเปลี่ยนวิถีในแวดวงการเรียนการสอนภาษา

เค้าร่างของสิ่งเหล่านี้ เริ่มบอกเราให้เห็นว่า แม้แต่ในงานภาษาเอง disruptive technology ก็กำลังเข้ามาส่งผลกระทบต่อวงการเรียนการสอนภาษา คงไม่เกินเลยไปหากจะกล่าวว่า ณ วันนี้ AI มีทักษะภาษาดีกว่านักเรียนจำนวนไม่น้อย เขียนหนังสือได้เร็ว และถูกต้องเข้าใจได้มากกว่า เหมือนกับที่ AI อ่านกฎหมายและเห็นความเกี่ยวข้องของประเด็นต่างๆ ในเอกสารได้เร็วและถูกต้องมากกว่านักกฎหมายที่เพิ่งเริ่มงาน เหมือนกับที่ AI อ่านภาพยนตร์ซีรีส์และวินจัยโรคได้เร็วและถูกต้องกว่าหมอบจใหม่

แน่นอนว่า ณ เวลานี้ ความสามารถของ AI ยังจำกัด ยังไม่สามารถเทียบเท่าผู้เชี่ยวชาญในสาขานั้นได้จริง ๆ AI ยังไม่สามารถเขียนนิยายที่ซับซ้อนสละสลวย AI ไม่เข้าใจวัฒนธรรมความเชื่อเบื้องหลังตัวบท AI ยังไม่มี critical thinking หรือเข้าใจความหมายทั้งหมดที่แฝงในภาษา ทางออกในเรื่องนี้ที่มักแนะนำกัน คือ คนที่ทำงานต่างๆ นั้นต้องพัฒนาตัวเองให้ทำงานที่เหนือกว่า AI ทำในสิ่งที่ AI ยังทำไม่ได้ การศึกษาไม่ว่าจะเป็นสาขาไหน จึงต้องปรับเพื่อผลิตบัณฑิตที่มีความสามารถกว่าเดิมที่ผลิตมา ต้องมาทำงานที่ต้องวิเคราะห์หรือมีความซับซ้อนมากขึ้น บัณฑิตไม่สามารถมาเริ่มต้นงานระดับเริ่มต้นเพื่อเรียนรู้ที่จะทำงานยาก ๆ ซับซ้อนต่อไปแบบเดิมได้ เพราะงานระดับเริ่มต้นเหล่านั้นจะถูกแทนที่ด้วย AI ไปแล้ว

- Synced. (2018). Liulishuo's AI App Is Teaching English to 70 Million People. Medium. Retrieved from <https://medium.com/syncedreview/liulishuos-ai-app-is-teaching-english-to-70-million-people-31d4fb38a799>

การปฏิรูปการศึกษา

ภายใต้เงื่อนไขนี้ หากเรายังจัดการศึกษาแบบเดิมที่ทำมา เราก็จะได้บัณฑิตที่มีความสามารถแบบเดิมๆ ตามที่เคยมีมา ทางเดียวที่จะสร้างบัณฑิตที่เก่งกว่าเดิมใน

ระยะเวลาเท่าเดิม คือ เราต้องปรับกระบวนการเรียนการสอนใหม่ คำตอบดูง่าย แต่สิ่งที่ยากคือไม่มีใครรู้แน่ชัดว่าต้องทำอะไรบ้าง ต้องทดลองหาแนวทางใหม่ๆ การสอนแต่เดิมที่เคยใช้เวลาเน้นไปในการฝึกทักษะภาษาให้นักเรียน ก็ต้องลดเวลาลงให้นักเรียนฝึกทักษะกับสื่อการเรียนรู้ด้วยตัวเองมากขึ้น และเพิ่มเวลาเรียนให้กับการใช้ความคิดวิเคราะห์ การฝึกใช้ภาษาต่างประเทศในบริบทที่ซับซ้อนขึ้น

ผู้สอนต้องเปลี่ยนทัศนคติและวิธีคิดแบบเดิม คือต้องเปลี่ยน mindset ให้ยอมรับและใช้ประโยชน์จากเทคโนโลยีมากขึ้น เพราะโดยตัวคำว่า disruptive เองมันก็บอกความหมายถึงการสูญสลายของวิถีเดิม เช่น ถ้าผู้สอนยึดติดกับการสอนเขียน essay ว่านักเรียนจะต้องเขียนเอง โปรแกรม AI essay writer ทั้งหลายก็จะถูกมองว่าเป็นอุปกรณ์การโกง จึงต้องหาทางป้องกันไม่ให้นักเรียนใช้ เช่น บังคับให้ทำงานเขียนให้เสร็จในเวลาเรียน ห้ามนักเรียนเปิดโทรศัพท์หรือคอมพิวเตอร์ใช้ในห้อง ยิงนานวันไปเราก็จะยิ่งพบกับความลำบากมากขึ้นในการควบคุมสภาวะแวดล้อมต่างๆ ให้คงสภาพเดิมแบบที่เคยเป็น เหตุมาจากสิ่งเดียว คือ ผู้สอนไม่ยอมรับว่าโลกได้เปลี่ยนไปแล้ว ปัญหาที่แท้จริงจึงอาจไม่ใช่ที่เทคโนโลยี แต่เป็นผู้สอนเองที่ไม่เข้าใจว่าเทคโนโลยีต่าง ๆ เหล่านี้จะเข้ามาเป็นส่วนหนึ่งในวิถีชีวิตต่อไป

หากเราเข้าใจว่าโลกได้เปลี่ยนไปอย่างไร มีเครื่องมืออะไรใหม่ๆ ที่เป็นประโยชน์เราก็จะปรับการสอนเข้ากับการเปลี่ยนแปลงได้มากขึ้น AI essay writer ก็จะถูกมองว่าเป็นเครื่องมือหนึ่งสำหรับใช้งาน MT ก็เป็นเหมือนพจนานุกรมที่ขยายตัวแปลมากกว่าคำๆ เดียว เป็นแปลได้ทั้งประโยคเลย สิ่งเหล่านี้เป็นเครื่องมือที่ผู้เรียนควรรู้จักและใช้ประโยชน์ได้ เพราะจุดมุ่งหมายของการสอนภาษาต่างประเทศ ไม่ได้มีเพียงเพื่อให้อ่านออกเขียนได้เท่านั้น เพราะนั่นเป็นเพียงทักษะพื้นฐานที่ AI ก็ทำได้เร็วกว่าและอาจจะดีกว่า แต่การสอนภาษาต่างประเทศควรเป็นไปเพื่อให้นักเรียนเข้าใจภาษาอย่างลึกซึ้งและใช้ภาษาได้ในบริบทซับซ้อน

การประมวลผลภาษาไทย

การเรียนการสอนในอนาคตเป็นงานที่ยากมากขึ้นเรื่อย ๆ เพราะการเปลี่ยนแปลงของโลกจะเร็วขึ้นเป็นทวีคูณเรื่อย ๆ แต่เดิม เราอาจใช้แนวทางที่เคยใช้สอนกันมาเป็นสิบ ๆ ปีได้ แต่ในอนาคตเราอาจต้องเปลี่ยนแนวทางและเรียนรู้สิ่งใหม่ ๆ ทุก 3-4 ปี และการเปลี่ยนแปลงอาจเปลี่ยนมากไปจนถึงขั้นที่เรานึกไม่ถึงก็ได้

สถาบันฝึกทักษะภาษา (Language Academy)

การเรียนภาษา ในอนาคตอาจออกมาในรูปแบบที่มีสถาบัน language academy เปิดให้บริการทั่วโลก ผู้เรียนระบุภาษาแม่ของตัวเอง ระบุภาษาต่างประเทศที่ต้องการเรียน เลือกโหมดของการเรียนเป็น adventure, thriller, nature, romance, city life, out of space, ancient world, ... ก่อนเริ่มเรียนจะมีการทำ placement test เพื่อประเมินความสามารถผู้เรียนก่อน จากนั้นก็สวม headset อย่างที่เห็นในหนัง *Ready Player One* ลืมไปว่าผู้เรียนจะเลือกแบบ single player หรือชวนเพื่อนมารวมด้วยเป็น multi players ก็ได้ เมื่อเข้าไปอยู่ใน VR จะมีครูผู้สอนมาในรูปของ buddy หรือ guardian angel ที่เป็นผู้รู้สองภาษาคอยช่วยผู้เรียนให้เล่น สนทนา ทำ task ต่างๆ เพื่อผ่านจาก level หนึ่งไปเรื่อยๆ จะผ่าน level ได้ความสามารถทางภาษาต้องผ่านระดับหนึ่งจึงจะ ออกจาก level นั้นได้ แนวคิดต่างๆที่อาจารย์สอนภาษาพูดกัน ไม่ว่าจะ เป็น task based, game-based, communicative, student centered, interactive, ... ทั้งหมดอยู่ใน VR game ที่ผู้เล่นต้องเรียนรู้ภาษาต่างประเทศเพื่อจะ survive และข้าม level บริษัทใหญ่บริษัทเดียวนี้สามารถให้บริการผู้เล่นได้ทั่วโลก ในภาพความเป็นไปได้นี้ อาจารย์สอนภาษาต่างประเทศจะยังมีความจำเป็นไหม?

จำเป็นต้องเรียนภาษาต่างประเทศหรือไม่

ภาพการเรียนภาษาผ่าน VR เป็นสิ่งที่เกิดขึ้นได้ไม่ยากเลยในอนาคต แต่เหตุผลเดียวที่สิ่งนี้จะไม่เกิด คือ บริษัทจะแน่ใจไหมว่าจะมีคนสนใจมาเรียนมากพอ ทำให้คนถึงอยากมาเล่นเกมเพื่อพัฒนาภาษาต่างประเทศ ในเมื่อ personal assistant ใน

อนาคตช่วยเราสนทนาภาษาอะไรก็ได้ ไม่ว่าจะไปเที่ยว จะไปเจรจาธุรกิจ ก็มีล่ามส่วนตัว แลumlล่ามนี้ก็พูดภาษาต่างประเทศด้วยโทนเสียงของเราได้

การพูดภาษาต่างประเทศได้เองในอนาคตอาจจะเป็นสิ่งพุ่มเพื่อไม่จำเป็นสำหรับคนทั่ว ๆ ไป แต่ก็เป็นไปได้ว่า ภาษาต่างประเทศอาจใช้บังคับในฐานะทางสังคม บอกว่าเป็นชนชั้นอภิชน หรือถ้ามีงานวิจัยสนับสนุนว่า คนที่พูดภาษาได้มากกว่าหนึ่งภาษาจะมีพัฒนาการของสมองดีกว่า ฉลาดกว่า เข้าใจโลกมากกว่า ภาษาต่างประเทศก็จะเป็นที่สนใจเรียนได้ แต่ก็อาจไม่ได้เป็นที่ต้องการของคนส่วนใหญ่แบบที่เป็นอยู่ในปัจจุบัน แต่ไม่ว่าสถานการณ์จะออกมาเป็นแบบไหน ความต้องการครูสอนภาษาต่างประเทศจะลดลงอย่างมาก หรือหากยังมีความต้องการจริง AI หรือ robot ก็อาจมาทำหน้าที่สอนนี้แทนได้หรือไม่ จึงไม่แน่ว่า เด็กในอนาคตอาจไม่รู้จักครูสอนภาษาต่างประเทศแบบเดียวกับที่เด็กในปัจจุบันไม่รู้จักเทปคาสเซ็ทท์ (tape cassette)

อนาคตของนักภาษา

มาถึงตรงนี้ อาจารย์สอนภาษาหลายคนอาจปฏิเสธว่าเป็นเรื่องเพ้อฝัน และก้มหน้าก้มตาสอนแบบเดิมด้วยความเชื่อว่าวิถีเดิมและความตั้งใจดีคือสิ่งที่ดีที่สุด หลายคนอาจรู้สึกท้อแท้ไม่มั่นใจในอนาคต แต่เราต้องยอมรับความจริงว่า โลกได้เปลี่ยนไปแล้ว และจะไม่เหมือนเดิมอีก เราพร้อมที่เคลื่อนไปข้างหน้าไหม หรือจะยอมที่จะถูก disrupt ไป

หากไม่สนใจตัวเอง ก็ควรนึกถึงผู้เรียนที่เข้ามาในสถาบันการศึกษา สิ่งที่เขาต้องการคือความรู้ความสามารถที่จะออกไปทำงานไปอยู่กับโลกภายนอกได้ หากเรายังสอนสิ่งที่เขาเรียนจบสี่ปีแล้วก็ เป็นความรู้ที่ใช้ประโยชน์ไม่ได้แล้ว เขาจะออกไปใช้ชีวิตและเติบโตในโลกภายนอกได้อย่างไร

สำหรับอาจารย์สอนภาษาต่างประเทศ เราคงต้องยอมรับว่าทักษะการสื่อสารเบื้องต้น ในอนาคต AI จะทำและสอนได้ดีกว่า เราต้องคิดเรื่องที่สูงและยากขึ้น ไม่ควรใช้เวลาส่วนใหญ่ไปกับการฝึกทักษะภาษาแบบเดิม ๆ แต่ให้เรียนรู้การใช้เครื่องมือ

การประมวลผลภาษาไทย

เกี่ยวกับภาษาใหม่ ๆ นี้เพื่อเรียนรู้และทำงานเกี่ยวกับภาษา และเสริมเพิ่มเนื้อหาให้วิเคราะห์ คิดเชื่อมโยง ทำความเข้าใจโลกและบริบทโลกมากขึ้น เช่น เรียนภาษาจีน ไม่ใช่เพื่ออ่านออกเขียนได้อย่างเดียว แต่เพื่อเข้าถึงความคิดความเชื่อคนจีน เข้าใจประวัติศาสตร์จีน เข้าใจปรัชญาเบื้องหลังความคิดเขา เข้าใจบริบทการเมืองและสังคมที่เป็นอยู่ เข้าใจวัฒนธรรมความนิยมในสื่อดนตรี ภาพยนต์ ละคร วรรณกรรมของเขา ภาษาเป็นเพียงเครื่องมือหรือหน้าต่างให้เราเข้าใจคนในสังคมวัฒนธรรมนั้นได้มากขึ้น ท้ายสุดเราจะรู้ว่าจะติดต่อกับเขาด้วยวิธีการใดให้สำเร็จราบรื่น ไม่ว่าจะเป็นเรื่องธุรกิจ การต่างประเทศ การท่องเที่ยว ทั้งหมดนี้ควรเป็นเป้าหมายของการเป็นนักอักษรศาสตร์ นักอักษรศาสตร์ที่ไม่ได้รู้แค่ภาษาแต่รู้เกี่ยวกับโลก รู้ความเป็นมาและรู้ความเป็นไป หากเราสอนนักเรียนให้เป็นได้เช่นนี้ ก็พร้อมใจได้ว่า เขาจะพร้อมออกไปอยู่ในโลกอนาคตที่เปลี่ยนแปลงต่อเนื่องได้

เราควรยินดีว่าช่วงเวลานี้ เป็นช่วงเวลาแห่งการเข้าสู่โลกใหม่ สู่อการแสวงหาความรู้ใหม่ๆ ความเข้าใจใหม่ ๆ เพื่อนำมาใช้ปรับการเรียนการสอน ทั้งหมดก็เพื่อให้เราเข้าใจโลก เข้าใจตัวเองมากขึ้น

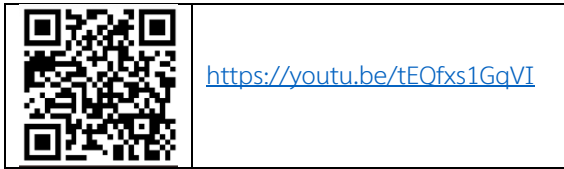
This is an exciting time. To teach students to live in an unknown world. To learn and unlearn all one's life.

ฝึกเขียนภาษาไทยให้ดีด้วย MT¹⁹

หลังการทดลองเล็ก ๆ ที่ใช้ Google เป็นเครื่องแปลภาษาและทดลองแก้ไขต้นฉบับไทยจนได้ภาษาอังกฤษแปลที่พอใช้การได้ (Aroonmanakun, 2021). ผู้เขียนได้ทดลองหาบทคัดย่อจากสาขาต่าง ๆ มาลองเล่นดูจำนวนหนึ่ง ก็พบว่าการปรับแก้ต้นฉบับภาษาไทยช่วยให้เครื่องแปลภาษาได้ดีขึ้นจริง ไม่ได้จำกัดว่าเป็นสาขาอะไร จะมีผิดบ้างก็เป็นการแปลศัพท์เฉพาะทางซึ่งผู้ใช้น่าจะแก้ไขเองได้ในรอบสุดท้าย แต่สิ่งที่น่าสนใจกว่า คือ การไม่มีกฎข้อบังคับการเขียนที่ชัดเจนในภาษาไทย ทำให้เรามักเขียนภาษาแบบเชื่อมโยงต่อเนื่องกันโดยไม่จำเป็นและไม่ระวังว่าข้อความนั้นมีความไม่ชัดเจนหรือมีความกำกวมหรือไม่ การใช้ MT สามารถช่วยตรวจภาษาไทยที่เขียนได้ เพราะหากมีส่วนใดที่เขียนแล้วทำให้ MT แปลผิดความไป อาจมาจากวิธีการเขียนที่ไม่ชัดเจนนี้เอง

บทนี้จะเริ่มด้วยการสาธิตการแก้ไขตัวอย่างต้นฉบับหนึ่งเพื่อช่วยเครื่องแปลภาษาให้ได้ความถูกต้อง และจะอธิบายว่าทำไมเมื่อแก้ไขแล้วความจริงชัดเจนและเครื่องแปลภาษาได้ถูกต้องมากขึ้น ประเด็นที่น่าสนใจในระหว่างการปรับแก้ต้นฉบับให้เครื่องแปลภาษาได้ดี ดูได้จากวิดีโอและคำอธิบายตัวอย่างการแปลด้วยเครื่อง ดังนี้

¹⁹ เรียบเรียงใหม่จาก “ฝึกเขียนภาษาไทยให้ดีด้วย MT” ใน <https://awirote.medium.com/> (20 ส.ค. 2564)



- งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อลดข้อบกพร่องจากข้อร้องเรียนของลูกค้าในงานดัดแปลงรถบรรทุกของบริษัทกรณีศึกษา
- This research aims to reduce defects from customer complaints in the case study of the company's truck modification work.

ก) ประโยคที่ดูธรรมดา นี้ มีความกำกวมซ่อนอยู่ เช่น ข้อบกพร่องจากข้อร้องเรียน อาจหมายถึง ข้อบกพร่องที่เกิดจากข้อร้องเรียน หรือ ข้อบกพร่องที่พบอยู่ในข้อร้องเรียน ส่วนข้อความ งานดัดแปลงรถบรรทุกของบริษัทกรณีศึกษา ข้อความที่ MT แปลโดยใช้ *the case study* เป็นคำหลักทำให้เห็นว่า MT เข้าใจผิดมองกรณีศึกษา เป็นคำหลักและมอง งานดัดแปลงรถบรรทุกของบริษัท เป็นส่วนขยาย ซึ่งไม่ใช่ความหมายที่ผู้เขียนต้องการ แม้ว่า MT จะวิเคราะห์ผิดในที่นี่ แต่การแก้ต้นฉบับให้ชัดเจนโดยเติม *ที่ใช้เป็น* หลังคำ *บริษัท* เพื่อให้ *กรณีศึกษา* ไม่ถูกมองเป็นคำหลักได้ วิธีนี้สามารถช่วยให้ MT แปลถูกความได้

- งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อลดข้อบกพร่องที่พบจากข้อร้องเรียนของลูกค้าในงานดัดแปลงรถบรรทุกของบริษัทที่ใช้เป็นกรณีศึกษา
- This research aims to reduce the defects found from customer complaints in the truck modification work of the company used as a case study.

ข) กรณีที่ MT แปลออกมาไม่เป็นประโยค อาจมาจากต้นฉบับเขียนในรูปที่เป็นข้อความไม่อิสระ เช่น ใช้คำเชื่อม *โดย* นำหน้า โดยปกติเรามักจะตีความโดยรวมไปกับความข้างหน้า แต่บางครั้งก็พบว่าอาจใช้แบบขึ้นประโยคใหม่ได้ ทำให้ MT ซึ่งยังไม่เข้าใจวรรคตอนภาษาไทยแปลออกมาแบบขึ้นต้นวลีต่างหาก

- โดยศึกษาส่วนงานดัดแปลงรถบรรทุกตามความต้องการของลูกค้า

- By studying the truck modifications according to the needs of customers

ทางแก้ที่ง่ายที่สุดคือลบบรรคข้างหน้า โดย... เพื่อให้ข้อความรวมไปกับข้างหน้า ทำให้ MT เลือกข้อความข้อความนั้นรวมไปกับข้อความหน้าได้ แต่ในตัวอย่างทดลองนี้ ผู้เขียนเลือกขึ้นประโยคใหม่โดยตัด โดย ทิ้งแล้วเติม จากการ แทน ทำให้ข้อความนี้ บอกการเริ่มต้นประโยคใหม่ได้ แต่ก็ต้องลบบรรคข้างหลังคำ ลูกค้ำ เพื่อให้ MT ดีความนี้รวมกับข้อความข้างหลัง

- จากการศึกษางานดัดแปลงรถบรรทุกตามความต้องการของลูกค้าซื้อโรงเรียนจากข้อบกพร่องประเภทแชสซีเสียหายเป็นเรื่องที่ถูกแจ้งมากที่สุด
- According to the study of truck modifications according to customer requirements, complaints of chassis damage were the most reported.

ค) กรณีที่ MT แปลมาไม่ได้คำหรือความที่น่าพอใจ ก็สามารถปรับต้นฉบับโดยเปลี่ยนคำหรือเขียนใหม่แต่ยังสื่อความเดิม เช่น งานดัดแปลง...ตามความต้องการลูกค้า เขียนใหม่เป็น งานดัดแปลง...ตามที่ถูกค้ำต้องการ และในระหว่างแก้ไขส่วนหนึ่ง MT อาจเปลี่ยนคำแปลที่เดิมใช้ได้ไปเป็นอย่างอื่น เช่นในตัวอย่างนี้ที่ MT เปลี่ยนก่อนหลังจากเดิม *complaints of chassis damage were the most reported* เป็น *complaints from the type of chassis damage were the most reported* เราก็ใช้วิธีเดิมคือปรับคำภาษาไทยโดยใช้คำอื่นแทน ในตัวอย่างนี้ ลองเปลี่ยนจากคำ *ประเภท* เป็น *เรื่อง* ก็ได้ข้อความแปลที่ยอมรับได้

- จากการศึกษางานดัดแปลงรถบรรทุกตามที่ถูกค้ำต้องการซื้อโรงเรียนจากข้อบกพร่องเรื่องแชสซีเสียหายเป็นเรื่องที่ถูกแจ้งมากที่สุด
- According to the study of truck modifications as requested by the customer, complaints of chassis damage were the most reported.

เปลี่ยนคำ *แรงเคลื่อนไฟฟ้า* เป็น *แรงเคลื่อนไฟ* เพื่อแก้คำแปลที่ได้จาก *electromotive force* เป็น *voltage*

- ปัจจัยกระแสไฟ แรงเคลื่อนไฟฟ้า และอัตราการไหลของแก๊สคลูม ส่งผลกระทบต่อการเกิดข้อบกพร่องอย่างมีนัยสำคัญ

การประมวลผลภาษาไทย

- power factor electromotive force and the flow rate of the cover gas significantly affect the occurrence of defects
- กระแสไฟ แรงเคลื่อนไฟ และอัตราการไหลของแก๊สคลุม ส่งผลกระทบต่อการเกิดข้อบกพร่องอย่างมีนัยสำคัญ
- the current, voltage and flow rate of the cover gas significantly affect the occurrence of defects

ง) ลดการเชื่อมโยงความที่ไม่จำเป็นด้วยการขึ้นประโยคใหม่ระบุประธานให้ชัดเจน หรือเติม . เพื่อบอกการจบประโยคได้ เช่น ตัวอย่างข้างล่างใช้คำ *ซึ่ง* ทำให้แปลมาเป็นประโยคย่อย ที่ต้องไปขยายความรวมกับประโยคข้างหน้า

- ซึ่งส่งผลต่อความปลอดภัยของผู้ใช้งาน และภาพลักษณ์ที่ไม่ดีต่อบริษัท
 - which affects the safety of users and bad image for the company
- แก้ไขโดยการลบคำว่า *ซึ่ง* ออก แล้วเติมประธาน *ข้อบกพร่องนี้* เข้าไปทำให้แปลแล้วขึ้นประโยคใหม่ได้
- ข้อบกพร่องนี้ส่งผลต่อความปลอดภัยของผู้ใช้งาน และภาพลักษณ์ที่ไม่ดีต่อบริษัท
 - This flaw affects user safety. and bad image for the company

จ) การใช้วรรคและตามด้วยคำสันธานในภาษาไทยเป็นปัญหาสำหรับ MT เพราะ MT ไม่เข้าใจว่าควรตีความต่อเนื่องกับข้อความข้างหน้าหรือเริ่มประโยค ทำให้แปลมาไม่เป็นประโยค ให้แก้ไขโดยลบวรรคออก ในตัวอย่างข้างบน เมื่อลบวรรคหน้าคำ และภาพลักษณ์... ก็จะเห็นต่อว่า ในภาษาไทย *ข้อบกพร่อง+ส่งผล+และ+ภาพลักษณ์* เป็นประโยคที่ไม่เสมอกัน ควรมีคำกริยาหลังคำ และ ด้วยเพราะก่อนหน้า และมีคำกริยา *ส่งผล* หลังและจึงควรมีคำกริยา จึงแก้ไขโดยเติมคำ *ทำให้เกิดภาพลักษณ์...* เมื่อแก้ไขแล้วจึงได้ความแปลดีขึ้น

ข้อบกพร่องนี้ส่งผลต่อความปลอดภัยของผู้ใช้งานและทำให้เกิดภาพลักษณ์ที่ไม่ดีต่อบริษัท

This flaw affects user safety and creates a bad image for the company.

ฉ) เติมเครื่องหมาย , หรือคำเชื่อมเพื่อช่วย MT เข้าใจว่าข้อความยังต่อเนื่อง เช่น ข้อความนี้

- โดยมีระดับปัจจัยที่เหมาะสมในกลุ่มพฤติกรรมการเชื่อมช้า กระแสไฟ 185 แอมแปร์ แรงเคลื่อนไฟ 150 โวลท์ และอัตราการไหลของแก๊สคลุม 20 ลิตร/นาที
- The appropriate factors were in the slow welding behavior group, current 185 amperes, voltage 150 volts, and cover gas flow rate 20 liters/min
ถ้าเดิม คือ และ , จะช่วยให้แปลได้ดีขึ้นตามที่เห็นนี้
- ระดับปัจจัยที่เหมาะสมในกลุ่มพฤติกรรมการเชื่อมช้า คือ กระแสไฟ 185 แอมแปร์, แรงเคลื่อนไฟ 150 โวลท์ และอัตราการไหลของแก๊สคลุม 20 ลิตร/นาที
- The appropriate levels of factors in the slow welding behavior group were 185 A current, 150 Volt voltage and 20 l/min cover gas flow rate

เมื่อมองภาพรวมการแก้ไขทั้งหมดตามที่เห็นใน track change นี้ จะเห็นว่า เป็นการแก้ไขวิธีการเขียน บางส่วนเป็นการแก้ภาษาไทยให้มีความชัดเจนขึ้น บางส่วน เป็นการแก้ไขเพื่อช่วยให้ MT เข้าใจว่าควรมองข้อความไหนรวมกับข้อความไหน เช่น การลบวรรค การใส่เครื่องหมาย , การเปลี่ยนคำที่ใช้ เป็นต้น ผลที่ได้เป็นคำแปลใน ตัวอย่างนี้เป็นคำแปลที่พอใช้งานได้

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อลดข้อบกพร่องที่พบจากข้อร้องเรียนของลูกค้าในงานตัดแปลงรถบรรทุกของบริษัทที่ใช้บัณฑิตศึกษา โดยศึกษาส่วนจากการศึกษา งานตัดแปลงรถบรรทุกตามความถี่ลูกค้า ต้องการของลูกค้า ข้อร้องเรียนจาก ข้อบกพร่องประเภทเรื่อง แซสซีเสียหยาบ เป็นเรื่องที่ถูกแจ้งมากที่สุด ซึ่งข้อบกพร่องนี้ส่งผลกระทบต่อความปลอดภัยของผู้ใช้งาน และทำให้เกิดภาพลักษณ์ที่ไม่ดีต่อบริษัท ผู้วิจัยจึงนำเอาเทคนิค ซิกซ์ ซิกมา (Six Sigma) มาประยุกต์ใช้ และดำเนินงานวิจัยตามหลักการ DMAIC เพื่อปรับปรุง และพัฒนากระบวนการตัดแปลงในโรงงานเชื่อมโลหะให้มีคุณภาพ โดยใช้การออกแบบการทดลองแบบ Box-Behnken Design (Box-Behnken Design) เพื่อหาระดับปัจจัยที่เหมาะสม ปัจจัยที่ใช้ทำการทดลองมี 3 ปัจจัย คือ กระแสไฟ แรงเคลื่อนไฟฟ้า แรงเคลื่อนไฟ และอัตราการไหลของแก๊สคลุม ซึ่งในการทดลองผู้วิจัยทำการศึกษาพฤติกรรมของพนักงาน เชื่อม โดยแบ่งเป็น 2 กลุ่ม คือกลุ่มพฤติกรรมการเชื่อมช้า และกลุ่มพฤติกรรมการเชื่อมเร็ว ผู้วิจัยทำการเก็บข้อมูล และวิเคราะห์ ผลการทดลองเชิงสถิติ ผลจากการทดลองพบว่า ปัจจัยกระแสไฟ แรงเคลื่อนไฟฟ้าแรงเคลื่อนไฟ และอัตราการไหลของแก๊สคลุม ส่งผลกระทบต่อ การเกิดข้อบกพร่องอย่างมีนัยสำคัญ โดยมีนัยสำคัญ ระดับปัจจัยที่เหมาะสมในกลุ่มพฤติกรรมการเชื่อมช้า คือ กระแสไฟ 185 แอมแปร์, แรงเคลื่อนไฟ 150 โวลท์ และอัตราการไหลของแก๊สคลุม 20 ลิตร/นาที และกลุ่มพฤติกรรมการเชื่อมเร็ว คือกระแสไฟ 235 แอมแปร์, แรงเคลื่อนไฟ 175 โวลท์ และอัตราการไหลของแก๊สคลุม 30 ลิตร/นาที จากผลของการปรับปรุงครั้งนี้ ทำให้สัดส่วนของเสียที่เกิดข้อบกพร่องในกระบวนการตัดแปลงรถบรรทุก ส่วนในงานเชื่อมโลหะ จาก 0.60% ลดลงเหลือ 0.12% และจาก 0.72% ลดลงเหลือ 0.20% ของกลุ่มพฤติกรรมการเชื่อมช้า และเร็ว ตามลำดับ ทำให้ผลการวิจัยนี้ ช่วยลดค่าใช้จ่ายงานซ่อมแซมของบริษัทกรณีศึกษาที่ศึกษาได้

การประมวลผลภาษาไทย

- This research aims to reduce the defects found from customer complaints in the truck modification work of the company used as a case study. According to the study of truck modifications as requested by the customer, complaints of chassis damage were the most reported. This flaw affects user safety and creates a bad image for the company. The researcher therefore applies the Six Sigma technique and conducts research according to the DMAIC principle to improve and develop the process of adapting the welding work to ensure quality. The Box-Behnken Design experiment was used to determine the optimal factor level. There were 3 factors used in the experiment which were current, voltage and flow rate of cover gas. In the experiment, the researcher studied the behavior of welding workers divided into 2 groups: slow welding behavior group and fast welding behavior group. The researcher collected the data and analyzed the experimental results statistically. The results of the experiment revealed that the current, voltage and flow rate of the cover gas significantly affect the occurrence of defects. The appropriate factor level in the slow welding behavior group is 185 A current, 150 Volt voltage and 20 l/min cover gas flow rate, and the fast welding behavior group is 235 A current, 175 Volt voltage and The gas flow rate was 30 cm/min. As a result of this improvement, the proportion of defective waste in the welding truck modification process was reduced from 0.60% to 0.12%, and from 0.72% to 0.20. % of the slow and fast welding behavior groups, respectively. This research resulted in a reduction in repair costs of the studied companies.

เมื่อทดลองแบบเดียวกันนี้ กับตัวบทสาขาวิชาอื่นดู ก็ได้ผลตามที่เห็นนี้

ตัวอย่างการปรับแก้งานการตลาดและผลแปล

การวิจัยครั้งนี้ศึกษาถึงผลของการประชาสัมพันธ์การตลาดโดยบุคคลมีชื่อเสียงผ่านอินสตาแกรมในประเทศไทย ซึ่งศึกษาในแง่ของ **เป็นการศึกษาผลที่มีต่อพฤติกรรมผู้บริโภค และความสัมพันธ์ของพฤติกรรมผู้บริโภคในแต่ละด้าน** ซึ่งได้แก่ การเปิดรับสาร, การรู้จักตราสินค้า,ทัศนคติต่อตราสินค้า, และพฤติกรรมการตัดสินใจของผู้บริโภค หลังจาก**การได้รับสาร**จากการประชาสัมพันธ์การตลาดโดยบุคคลมีชื่อเสียงผ่านอินสตาแกรมของ 3 ตราสินค้า คือ **ไอศกรีมแม็กนัม, รองเท้ากีฬา, นิวบาลานซ์ และแอปพลิเคชันสนทนาทางสมาร์ตโฟนวีแชท** ซึ่ง**ผู้วิจัยได้ทำการสำรวจ**โดยเก็บข้อมูลจากแบบสอบถามกับกลุ่มตัวอย่างจำนวน 411 คน ซึ่งกลุ่มตัวอย่างทั้งหมดเป็นผู้ใช้อินสตาแกรมที่มีอายุ 18-35 ปี และมี**พฤติกรรมติดตาม (Follow)** บุคคลมีชื่อเสียงในอินสตาแกรมอย่างน้อย 3 คน ผลการวิจัยพบว่าผลของการประชาสัมพันธ์ที่เกิดขึ้นกับพฤติกรรมผู้บริโภค ส่วนใหญ่อยู่ในระดับปานกลาง นอกจากนี้ยังพบว่ากลุ่มตัวอย่างที่มีเพศ และการศึกษาต่างกัน มีการรู้จักตราสินค้า, ทัศนคติต่อตราสินค้า, และพฤติกรรมการตัดสินใจต่อสินค้าที่ไม่แตกต่างกัน ในขณะที่กลุ่มตัวอย่างที่มีอายุต่างกัน มีการรู้จักตราสินค้า, ทัศนคติต่อตราสินค้า, และพฤติกรรมตัดสินใจต่อสินค้าที่ต่างกันอย่างมี**นัยสำคัญทางสถิติที่สำคัญ**ในระดับ 0.05 และเมื่อพิจารณาถึงความสัมพันธ์ของพฤติกรรมผู้บริโภคในแต่ละด้าน พบว่า (1) การเปิดรับสารจากการประชาสัมพันธ์การตลาดโดยบุคคลมีชื่อเสียงผ่านอินสตาแกรม มีความสัมพันธ์ไปในทางเดียวกันกับการรู้จักตราสินค้า และทัศนคติต่อตราสินค้าของผู้บริโภค (2) การรู้จักตราสินค้า และทัศนคติต่อตราสินค้า มีความสัมพันธ์ไปในทางเดียวกันกับพฤติกรรมการตัดสินใจของผู้บริโภค

-
-
-

- This research examines the effect of celebrity marketing publicity through Instagram in Thailand. The study was to study the effects on consumer behavior in each aspect, namely substance exposure, brand awareness, brand attitude, and consumer purchasing behavior. After receiving the message from the marketing campaigns by famous people via Instagram of 3 brands: Magnum, New Balance, and WeChat, the researchers surveyed the data from a questionnaire sample of 411 people. All samples were Instagram users aged 18–35 and followed (follow) at least 3 famous Instagrammers. consume Most were moderate. It was also found that the samples of different sexes and education had the same brand awareness, brand attitude, and purchasing behavior. While the samples of different ages had brand awareness, brand attitude, and purchasing decision behaviors were significantly different at the 0.05 level. When considering the relationship of consumer behavior in each aspect, it was found that (1) The exposure of celebrity marketing campaigns via Instagram is correlated in the same way as brand awareness and brand attitudes among consumers. (2)

การประมวลผลภาษาไทย

Brand awareness and brand attitude are related to the purchasing behavior of consumers.

-

ตัวอย่างการปรับแก้งานทางมนุษยศาสตร์และผลแปล

วิทยานิพนธ์นี้มีความมุ่งหมายวัตถุประสงค์ 2 ประการ วัตถุประสงค์แรกคือเพื่อจะศึกษาประวัติและความเป็นมาเรื่องเล่าของพระอินทร์ผู้เป็นจอมเทพยอดนักรบของศาสนาพราหมณ์-ตามที่มีปรากฏในวรรณคดีสันสกฤตและเป็นเทพผู้มีอุปการคุณในศาสนาพุทธตามที่ปรากฏในวรรณคดีบาลี เป็นการรวบรวมเรื่องราวของพระอินทร์ทุกแง่มุมตามความเชื่อถือของชาวอินเดียอันเป็นต้นเค้าของพระอินทร์ในวรรณคดีไทยประการหนึ่ง อีกประการหนึ่งวัตถุประสงค์ที่สองคือเพื่อชี้ให้เห็นความแตกต่างและความคล้ายคลึงกันของพระอินทร์ในวรรณคดีสันสกฤตและบาลี และศึกษาถึงอิทธิพลของพระอินทร์ในวรรณคดีสันสกฤตและบาลีที่มีต่อวรรณคดีไทยด้วย โดยจะแสดงให้เห็นว่าเรื่องราวตอนใดในวรรณคดีทั้งสองภาษานั้นเข้ามาปรากฏในวรรณคดีไทยบ้าง ทั้งนี้เพราะไทยเรารับได้รับอิทธิพลความเชื่อเรื่องพระอินทร์มานานก่อนที่จะตั้งอาณาจักรสุโขทัยเสียอีก หลักฐานในเรื่องนี้จะเห็นได้จากโบราณวัตถุในประเทศไทย และตำนานต่าง ๆ ตลอดจนวรรณคดีไทยด้วย วรรณคดีไทยที่สำคัญคือ ไตรภูมิพระร่วง (เขมุนิกถา) ซึ่งเป็นวรรณคดีที่มีอิทธิพลต่อความเชื่อของชาวไทยในสมัยต่อ ๆ มา ต่อมา เป็นอย่างไรก็ดี ไตรภูมิพระร่วงมีเรื่องราวของพระอินทร์ปรากฏอยู่มาก เนื้อเรื่องเนื้อหาของวิทยานิพนธ์แบ่งเป็น 6 บท คือ บทที่ 1 เป็นบทนำ บทที่ 2 กล่าวถึงที่มาของพระอินทร์ในวรรณคดีไทย บทที่ 3 กล่าวถึงประวัติและความเป็นมาของพระอินทร์ในวรรณคดีสันสกฤต บทที่ 4 กล่าวถึงประวัติและความเป็นมาของพระอินทร์ในวรรณคดีบาลี บทที่ 5 กล่าวถึงพระอินทร์ในด้านานและวรรณคดีไทย บทที่ 6 เป็นบทสรุปและข้อเสนอแนะ สำคัญของวิทยานิพนธ์เรื่องนี้ อยู่ที่การมุ่งเสนอประวัติและความเป็นมาของพระอินทร์จากวรรณคดีสันสกฤตและบาลี และการเปรียบเทียบข้อเหมือนและข้อแตกต่างและข้อตรงกันของพระอินทร์ในวรรณคดีทั้งสองภาษา **กำับรวมถึงอิทธิพลของพระอินทร์จากวรรณคดีทั้งสองภาษาที่มีต่อวรรณคดีไทยด้วย** 🏠

-

This thesis has two objectives. The first objective is to study the history and narrative of Indra, the supreme warrior of Brahmanism as it appears in Sanskrit literature and as a patron deity in Buddhism as it appears in the Pali literature. It is a compilation of the story of Indra in every aspect according to the beliefs of the Indians, one of the origins of Indra in Thai literature. The second objective is to point out the differences and similarities of Indra in Sanskrit and Pali literature and to study Indra's influence in Sanskrit and Pali literature on Thai literature. It will show which stories in the literature of both languages appear in Thai literature. This is because Thailand was influenced by the belief in Indra long

before the Sukhothai Empire was established. Evidence in this regard can be seen from antiquities in Thailand and various legends as well as Thai literature. The most important Thai literature is Trai Phum Phra Ruang (Taiphum Katha), which is a literature that greatly influenced the beliefs of Thai people in later times. Traiphum Phra Ruang has many stories of Indra. The content of the thesis is divided into 6 chapters: Chapter 1 is an introduction, Chapter 2 discusses the origin of Indra in Thai literature, Chapter 3 describes the history and background of Indra in Sanskrit literature, Chapter 4 discusses the history and background of Indra. of Indra in Pali literature, Chapter 5 discusses Indra in legends and Thai literature. Chapter 6 is a summary and recommendations. The essence of this thesis lies in the focus on presenting the history and background of Indra from Sanskrit and Pali literature, and comparing the similarities and differences of Indra in the literature of both languages, as well as the influence of Indra from the literature of both languages. towards Thai literature

MT ช่วยฝึกเขียนภาษาไทยอย่างไร

หลังจากที่เราเขียนบทความภาษาไทยแล้ว ทดลองให้ Google ช่วยแปลเป็นภาษาอังกฤษ เราจะเห็นปัญหาในข้อความแปลนั้น ปัญหาส่วนหนึ่งเกิดจาก MT ที่ไม่เข้าใจวรรคตอนข้อความภาษาไทย เราช่วยได้โดยการลบวรรคเพื่อรวมข้อความเป็นประโยคเดียวกัน ใส่เครื่องหมาย , เพื่อบอกว่ายังมีต่อ ใส่เครื่องหมาย . เพื่อบอกการจบประโยค ส่วนเหล่านี้เป็นการแก้เพื่อช่วย MT ไม่ได้เกี่ยวกับการเขียนภาษาไทยโดยตรง แต่จริง ๆ แล้ว การใช้เครื่องหมาย . หรือ , หากไม่มองว่าแปลกตา ก็เป็นวิธีที่ช่วยให้เราเขียนภาษาไทยโดยแบ่งความเป็นประโยคได้ชัดเจนขึ้น (ความคิดเรื่องการใส่เครื่องหมายวรรคตอนกับภาษาไทย รัชกาลที่ ๖ ก็ทรงมีพระราชดำริเรื่องนี้ และว่าการที่ภาษาไทยเขียนติดกันไปหมด ทำให้เรียนและเขียนอ่านให้ถูกต้องได้ยากกว่าที่ควร แม้คนส่วนใหญ่จะไม่นิยมการเขียนเครื่องหมายวรรคตอนสากล ผู้เขียนหลายท่านก็

การประมวลผลภาษาไทย

เลือกใช้เพื่อให้ผู้อ่านเข้าใจง่ายขึ้น เช่น ท่านพุทธทาสภิกขุ) ในส่วนนี้นั้น การลบวรรคไม่ใช่การแก้ไขภาษาไทยโดยตรงแต่เป็นการช่วย MT ให้แปลได้ถูกเพราะรู้ว่าจะรวมข้อความไหนตีความร่วมกับข้อความไหน แต่ในงานเขียนภาษาไทย เรายังควรคงวรรคที่ใช้เพื่อแบ่งความให้อ่านง่ายขึ้น

การแก้ไขอีกส่วนเป็นการแก้ไขการเขียนโดยตรง การใช้คำเชื่อมความมากไปโดยไม่จำเป็น เช่น *ซึ่ง โดย และ* ทำให้ MT หรือแม้แต่คนอ่านก็ต้องตีความต่อเนือง ประกอบกับการละประธานที่มักพบในข้อความเหล่านี้ ทำให้ยากแก่การเข้าใจ หากเขียนโดยเริ่มประโยคใหม่และบอกประธานที่ชัดเจนเลยจะง่ายต่อความเข้าใจของทั้ง MT และคนอ่านมากขึ้น นัยสำคัญสำหรับสิ่งที่เห็นนี้ จึงไม่ได้มีเพียงเรื่องวิธีการใช้ Google Translate เพื่อแปลงานจากไทยเป็นอังกฤษให้ใช้งานได้ แต่การทำเช่นนี้ ทำให้เราต้องทบทวนว่า ในการเขียนข้อความภาษาไทยแล้ว เราเขียนหนังสือได้ชัดเจนพอไหม เครื่องแปลภาษาหรือ MT จึงเป็นเหมือนเครื่องมือที่นำมาใช้ฝึกการเขียนภาษาไทยได้ด้วย ช่วยให้เราตรวจสอบการเขียนว่ามีการใช้คำเกินจำเป็นสามารถตัดออกหรือเปลี่ยนวิธีการเขียนให้เข้าใจง่ายขึ้นกว่าที่เขียนเดิมไหม

แน่นอนว่า การจะทำเช่นนี้ได้ หมายถึงผู้ใช้ก็ต้องมีทักษะภาษาอังกฤษระดับหนึ่งที่อ่านคำแปลแล้วเข้าใจว่า ส่วนไหนที่แปลมาไม่เป็นประโยคหรือเนื้อความไม่ถูกตามต้นฉบับ ถ้านักเรียนที่ยังไม่รู้จักภาษาอังกฤษมากพอก็จะทำไม่ได้ แต่สำหรับครูภาษาไทย ส่วนใหญ่น่าจะมีความรู้ภาษาอังกฤษพอที่บอกได้ว่าข้อความที่แปลมานั้นส่วนไหนไม่ถูก ส่วนไหนเป็นประโยคไม่สมบูรณ์ เมื่อได้ฝึกแก้ไขงานเขียนภาษาไทยด้วยวิธีนี้ ก็จะจับหลักได้ว่า การเขียนภาษาไทยให้ได้ความและอ่านเข้าใจง่าย ควรตัดแบ่งความอย่างไร โครงสร้างส่วนต่าง ๆ เสมอกันและมาขยายความส่วนหลักอย่างไร แล้วจึงนำความรู้ความเข้าใจเหล่านี้มาถ่ายทอดให้นักเรียนฝึกเขียนภาษาไทยให้เป็นระบบระเบียบต่อไปได้

ความสามารถในการอ่านและเขียนภาษาไทยให้ได้ดี สามารถวิเคราะห์ความ และสื่อความได้ชัดเจนเป็นเรื่องสำคัญอย่างมาก อาจสำคัญมากกว่าการเรียนรู้ ภาษาต่างประเทศอีก เพราะคนส่วนใหญ่ที่ไม่ได้ต้องไปใช้ชีวิตในต่างประเทศหรือ ทำงานร่วมกับชาวต่างประเทศ ในอนาคตสามารถใช้เครื่องแปลภาษาช่วยในการ ติดต่อสื่อสารทั่วไปกับชาวต่างชาติได้ไม่ยาก คนเหล่านี้ไม่ได้ต้องการมีทักษะ ภาษาต่างประเทศที่ลึกซึ้ง อาจต้องการความรู้ภาษาต่างประเทศแบบพื้นฐานเท่านั้น คนที่จำเป็นต้องรู้ภาษาต่างประเทศแบบลึกซึ้งนั้นอาจมีไม่มาก แต่คนส่วนใหญ่ควรมี ความสามารถในการคิดและวิเคราะห์เรื่องต่าง ๆ ได้ดี ซึ่งความสามารถนี้ต้องอาศัย ความรู้ทางภาษาแม่เป็นหลัก การสอนให้เด็กไทยอ่านวินิจสารได้และเห็นถึงโครงสร้าง ความในปริจฉาจึงควรเป็นเป้าหมายสำคัญสำหรับครูไทย

อ้างอิง

- Aroonmanakun, W. (2021). เขียนภาษาไทยอย่างไรให้ได้ภาษาอังกฤษระดับ B2-C1 - Medium. Retrieved from <https://awrote.medium.com/เขียนภาษาไทย-อย่างไรให้ได้ภาษาอังกฤษระดับ-b2-c1-6de8424390de>
- พระบาทสมเด็จพระมงกุฎเกล้าเจ้าอยู่หัว. (2545). วิธีใหม่สำหรับใช้เขียนสระและ เขียนหนังสือ ไทย. วารสารราชบัณฑิตยสถาน ปีที่ 27 ฉบับที่ 2, เม.ย.-มิ.ย., 561-586.

ดัชนีคำ

- AGI, 209, 238
- Artificial Narrow Intelligence, 239
- AutoGPT, 237
- BEST corpus, 13
- disruptive technology, 249
- EDU, 137, 138
- factor analysis, 178
- GPT-3, 209
- GPT-4, 227
- multimodal, 230
- named entity recognition, 85
- principle component analysis, 178
- Rhetorical Structure Theory, 137
- universal dependencies, 110
- Universal Dependencies, 147, 171
- word2vec, 16, 197
- การแจกแจงส่วนประกอบ, 171
- การซ้ำคำ, 12
- การตัดคำ, 8
 - feature-based, 9
 - longest matching, 8
 - maximum matching, 9
 - TCC, 10
 - ตัดพยางค์, 10
- การตัดประโยค, 129
- การปฏิรูปการศึกษา, 250
- การประสมคำ, 12
- การรู้จำชื่อเฉพาะ, 84, 86
- การลดทอนมิติข้อมูล, 178
- การวิเคราะห์พหุมิติ, 178
- การวิเคราะห์หน่วยสร้าง, 145
- การสกัดคำประสม, 55, 61
- การสร้างคำ, 11
- ความสัมพันธ์แบบพึ่งพา, 110
- คำความหมายใกล้เคียง, 198
- คำความหมายตรงข้าม, 199
- คำซ้อน, 40, 60, 202
- คำซ้ำ, 40, 60
- คำประกอบ, 40
- คำประสม, 12, 17, 27, 40
 - appositive compound, 32, 60
 - bahuvrihi compound, 29
 - coordinate compound, 32, 44, 58, 60
 - copulative compound, 32, 58
 - dvandva compound, 32
 - endocentric compound, 60
 - exocentric compound, 60
 - neo-classical compound, 32, 45
 - noun-noun compound, 42
 - opaque compound, 63
 - synthetic compound, 30, 41
 - transparent compound, 63
- ความสัมพันธ์ภายในคำประสม, 37
- คำประสมแบบเข้าศูนย์, 28, 74
- คำประสมแบบซ้ำ, 44

คำประสมแบบไร้ศูนย์, 29, 75
คำประสาน, 40
คำปรากฏร่วม, 201
คำพ้องรูป, 202
คำไวยากรณ์, 200
คำหลายความหมาย, 203
ชื่อเฉพาะ, 84
ชื่อสถานที่ใช้อ้างถึงองค์กร, 96
ชื่อองค์กรใช้อ้างถึงสถานที่, 96
ชื่อผลิตภัณฑ์, 97
ตราสินค้า, 98
ทฤษฎีโครงสร้างวาทะ, 137
ทางการวิเคราะห์แบบไวยากรณ์ฟังก์พา, 145

ประโยคความซ้อน, 130
ประโยคความซับซ้อน, 131
ประโยคความเดียว, 130
ประโยคความรวม, 131
โปรแกรม POS Tagger, 109
วาทสัมพันธ์, 137
เว็ทเตอร์คำบริบท, 68
ไวยากรณ์ฟังก์พา, 147
หน่วยคำ, 11
หน่วยปริจเฉทพื้นฐาน, 138
หน่วยสร้าง, 145
หมวดคำ, 108