การแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง : กรณีศึกษาการแปลรายงานข่าวตลาดหุ้นจากภาษาไทยเป็นภาษาอังกฤษ

นายธเนศ เรืองรจิตปกรณ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2549

AN EXAMPLE-BASED MACHINE TRANSLATION : A CASE STUDY OF TRANSLATING STOCK REPORTS FROM THAI TO ENGLISH

Mr. Taneth Ruangrajitpakorn

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Arts Program in Linguistics

Department of Linguistics

Faculty of Arts

Chulalongkorn University

Academic Year 2006

หัวข้อวิทยานิพนธ์	การแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง : กรณีศึกษาการแปลรายงานข่าวตลาดหุ้นจากภาษาไทยเป็นภาษาอังกฤษ
โดย	นายธเนศ เรืองรจิตปกรณ์
สาขาวิชา	ภาษาศาสตร์
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

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

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.กิ่งกาญจน์ เทพกาญจนา)

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

.....กรรมการ
(อาจารย์ ดร.สุดา รังกุพันธุ์)

ธเนศ เรืองรจิตปกรณ์ : การแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง : กรณีศึกษาการแปลรายงานข่าวตลาดหุ้นจากภาษาไทยเป็นภาษาอังกฤษ (AN EXAMPLE-BASED MACHINE TRANSLATION : A CASE STUDY OF TRANSLATING STOCK REPORTS FROM THAI TO ENGLISH) อ. ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล, 127 หน้า.

แนวทางการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างเป็นแนวทางที่น่าสนใจและยังใหม่ในวงการแปลภาษาด้วยเครื่องของภาษาไทย วิทยานิพนธ์ฉบับนี้จึงทดลองพัฒนาระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างโดยอาศัยแนวทางสกัดแม่แบบการแปล ซึ่งเป็นแนวทางที่ต้องอาศัยตัวอย่างการแปลจากคลังข้อมูลเทียบบท ผู้วิจัยจึงสร้างคลังข้อมูลเทียบบทโดยคัดเลือกรายงานข่าวตลาดหุ้น เพราะเป็นข้อมูลที่มีการปรากฏซ้ำๆ ซึ่งจะเป็นกรณีศึกษาที่ดีต่อการแปลแบบอิงตัวอย่าง

ระบบที่พัฒนานี้มีการทำงานแบ่งออกเป็น 2 ส่วนคือ (1) ระบบสกัดแม่แบบการแปล ซึ่งระบบจะสร้างต้นไม้การปรากฏร่วมของแต่ละภาษาจากข้อมูลภาษาในคลังข้อมูลเทียบบทมาทำการเปรียบเทียบกิ่งที่มีหมายเลขระบุบรรทัดตรงกันและจับคู่คำแปลของส่วนซ้ำภายในบรรทัดนั้นมาสร้างเป็นแม่แบบการแปลส่วนคงที่และจับคู่คำแปลของส่วนไม่ซ้ำเป็นแม่แบบการแปลส่วนผันแปร (2) ระบบรวมคำแปลใหม่ ซึ่งนำแม่แบบการแปลที่สกัดได้มาเทียบข้อมูลรับเข้าและเลือกแม่แบบการแปลที่ใช้แปลข้อความได้มาเทียบแปลข้อความจากส่วนที่ยาวที่สุดก่อนจนครบทั้งข้อความ จึงจะได้ผลการแปลที่สมบูรณ์

ผลการทดลองสกัดแม่แบบการแปลจากคลังข้อมูลโดยตรงพบว่า ระบบสามารถสกัดแม่แบบการแปลได้ถูกต้องเพียงร้อยละ 9.85 ดังนั้นผู้วิจัยจึงทดลองต่อโดยช่วยจัดกลุ่มข้อมูลที่คล้ายกันก่อนที่จะสกัดแม่แบบการแปลจากคลังข้อมูล จึงได้แม่แบบการแปลที่ถูกต้องทั้งหมดเพื่อที่จะนำผลมาทดสอบส่วนการแปลข้อความต่อไป และจากผลการทดลองแปลข้อความสรุปได้ว่า ระบบสามารถแปลโดยอาศัยแม่แบบที่สกัดจากคลังข้อมูลโดยตรงได้ถูกต้องร้อยละ 3.70 ส่วนผลการแปลที่สกัดจากแม่แบบการแปลที่ผู้วิจัยจัดกลุ่มข้อมูลตามความคล้ายคลึงให้ความถูกต้องร้อยละ 67.68

จากการวิเคราะห์ปัญหาพบว่า คลังข้อมูลเทียบบทที่ใช้แม้จะมีการปรากฏซ้ำของข้อความแปลแต่เป็นตัวอย่างการแปลแบบเน้นเจตนา ทำให้มีการละข้อความบางส่วนส่งผลให้เกิดปัญหาต่อการจับคู่ข้อความแปลให้ถูกต้อง นอกจากนี้ยังพบปัญหาความไม่เท่ากันระหว่างภาษาไทยและภาษาอังกฤษที่เป็นปัญหาต่อการแปลด้วยระบบนี้ ได้แก่ การที่ภาษาอังกฤษแสดงกาล การณ์ลักษณะ ทัศนภาวะ ด้วยวิภัตติปัจจัยและคำช่วยหน้ากริยา ในขณะที่ภาษาไทยมีการละคำหรือไม่ก็แสดงด้วยคำช่วยหน้าหรือหลังกริยา การที่ภาษาไทยมีการใช้กริยาเรียงในขณะที่ภาษาอังกฤษเลือกแสดงในโครงสร้างลักษณะอื่น ปัญหาเหล่านี้เป็นอุปสรรคต่อการจับคู่ข้อความเพื่อสร้างแม่แบบการแปลให้ถูกต้อง ดังนั้น คลังข้อมูลเทียบบทที่ใช้สำหรับการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างนี้จึงควรเป็นข้อมูลที่แปลแบบคำต่อคำ และควรมีการวิเคราะห์วจีวิภาคเพื่อแก้ปัญหาความไม่เท่ากันระหว่างภาษาในระดับหนึ่งก่อน

ภาควิชา.... ภาษาศาสตร์..... ลายมือชื่อนิสิต .....

สาขาวิชา.... ภาษาศาสตร์..... ลายมือชื่ออาจารย์ที่ปรึกษา .....

ปีการศึกษา.... 2549.....

4680140022 : MAJOR COMPUTATIONAL LINGUISTICS

KEYWORD: EXAMPLE-BASED MACHINE TRANSLATION / TEMPLATE EXTRACTION TECHNIQUE / COLLOACTION TREE / RECOMBINATION TECHNIQUE / PARALLEL CORPORA.

TANETH RUANGRAJITPAKORN : AN EXAMPLE-BASED MACHINE TRANSLATION : A CASE STUDY OF TRANSLATING STOCK REPORTS FROM THAI TO ENGLISH. THESIS ADVISOR : ASSISTANT PROFESSOR WIROTE AROONMANAKUN, Ph.D., 127 pp.

Example-based approach is novel in the field of Thai-English machine translation. This thesis presents an implementation of an example-based machine translation system using templates automatically extracted from a parallel corpus of stock exchange trade news. This corpus is selected because of the repetition of texts and translation patterns in the corpus. These pattern-based co-occurrences are believed to be a good case study for an English-Thai example-based translation.

The system is divided into two modules: template extraction module and translation recombination module. The template extraction module extracts translation templates from a given corpus by means of collocation tree comparison between those of the source and target languages to determine and to align translation variables and invariant fragments. The translation recombination module matches a given input sentence with extracted templates, and recursively translates the unmatched parts until accomplishment.

Experiments were conducted to elucidate the effects of corpus preparation methods on accuracy of template extraction and translation. Two versions of corpus — a raw version, and the other version which its sentences are clustered into groups regarding to pattern similarity — were used to extract translation templates. The former version yields out extraction accuracy for 9.85%, whereas the latter one yields out for 100%. All extracted templates were used to translate the prepared test set. The former version yields out translation accuracy for 3.70%, whereas the latter one yields out for 67.68%.

The case study reveals several significant issues of automatic Thai-English translation. First, omission both in Thai and in English, affecting translation accuracy in resolving missing parts, fairly occurs in the corpus because of communicative translation. This issue deteriorates the accuracy of template matching and template extraction. Second and finally, linguistic inequalities between Thai and English — i.e. omissible usage of auxiliaries in Thai versus inflection to express tenses, aspects, and mood in English; and verb serialisation in Thai versus usage of subordinate and coordinate structures in English — affects the accuracy of template extraction. Parallel corpora appropriate for Thai-English example-based translation should, in conclusion, be well-aligned in word level and annotated with Thai-English-equalised parts of speech to resolve the linguistic inequalities preliminarily.

Department ....Linguistics……. Student's signature.....

Field of study…. Linguistics……. Advisor's signature.....

Accademic year.....2006.....

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณ ผศ.ดร.วิโรจน์ อรุณมานะกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์เป็นอย่างสูงที่ได้ช่วยเหลือและให้คำปรึกษาแนะนำแนวทางและขัดเกลาวิทยานิพนธ์ฉบับนี้จนสำเร็จลุล่วงลงได้ซึ่งหากขาดความอนุเคราะห์จากอาจารย์แล้ววิทยานิพนธ์ฉบับนี้คงมิอาจสำเร็จลงได้เลย และผู้วิจัยขอขอบพระคุณ รศ.ดร.กิ่งกาญจน์ เทพกาญจนา และอาจารย์ ดร.สุดา รังกุพันธุ์ กรรมการสอบวิทยานิพนธ์ที่ได้ให้คำปรึกษาและเสียสละเวลาเพื่อตรวจสอบแก้ไขวิทยานิพนธ์ฉบับนี้และขอขอบคุณคณาจารย์ภาควิชาภาษาศาสตร์ทุกท่านที่ได้ประสิทธิ์ประสาทความรู้ด้านภาษาศาสตร์พร้อมทั้งช่วยเหลือให้คำปรึกษาเรื่องต่างๆให้แก่ผู้วิจัย

นอกจากนี้ผู้วิจัยขอขอบคุณ ดร.เทพชัย ทรัพย์นิธิ และคุณปรัชญา บุญขวัญแห่งหน่วยปฏิบัติการวิทยาการมนุษยภาษา ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติสำหรับคำชี้แนะแนวทางและปัญหาต่างๆในการทำวิทยานิพนธ์เล่มนี้ และอนุเคราะห์ข้อมูลงานวิจัยที่เกี่ยวข้องและเครื่องมือในการทำวิจัย และขอขอบคุณคุณณัฐพล กฤษสุทธิกุล สำหรับเครื่องมือตัดแบ่งคำ และดร.กฤษณ์ โกสวัสดิ์และพี่มณฑิกา บริบูรณ์แห่งหน่วยปฏิบัติการวิทยาการมนุษยภาษาที่ช่วยอำนวยความสะดวกเรื่องเวลาสำหรับทำวิทยานิพนธ์และเร่งรัดผู้วิจัยในการทำวิทยานิพนธ์ฉบับนี้เสมอมา

ผู้วิจัยขอขอบคุณเจ้าหน้าที่เพื่อนๆ พี่ๆ น้องๆ ทุกคนจากภาควิชาภาษาศาสตร์และศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติที่คอยช่วยเหลืออำนวยความสะดวกและย้ำเตือนผู้วิจัยให้เร่งทำวิทยานิพนธ์ฉบับนี้เสมอมาพร้อมทั้งคอยช่วยเหลือและตักเตือนเรื่องอื่นๆ

สุดท้ายนี้ผู้วิจัยขอขอบพระคุณบิดามารดาสำหรับการสนับสนุนค่าใช้จ่ายระหว่างการศึกษาและเปิดโอกาสให้ได้เรียนในสาขาวิชาที่เลือกด้วยตนเองรวมทั้งอำนวยความสะดวกในทุกๆ ด้านพร้อมทั้งเป็นกำลังใจในการทำวิทยานิพนธ์อย่างดียิ่งตลอดมา

สารบัญ

หน้า

หน้า

หน้า

สารบัญภาพ

หน้า

หน้า

สารบัญตาราง

หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การแปลภาษาด้วยเครื่อง (machine translation หรือ MT) คือ การนำระบบอัตโนมัติที่ใช้บนเครื่องคอมพิวเตอร์มาใช้แปลข้อความจำนวนมากๆ จากภาษาต้นทางไปเป็นภาษาปลายทาง โดยสามารถป้อนข้อมูลข้อความภาษาต้นทาง และจะได้ผลลัพธ์เป็นข้อความภาษาปลายทาง ซึ่งจะช่วยให้สามารถแปลข้อความได้เป็นจำนวนมากและรวดเร็ว การทำวิจัยและพัฒนาการแปลภาษาด้วยเครื่องนั้นนับเป็นงานสำคัญแขนงหนึ่งในศาสตร์แห่งการประมวลผลภาษาธรรมชาติ (Natural Language Processing หรือ NLP) งานวิจัยด้านการแปลภาษาด้วยเครื่องชิ้นแรกที่เผยแพร่สู่สาธารณะถูกผลิตขึ้นในปี พ.ศ. 2497 โดยมหาวิทยาลัยจอร์จทาวน์ (Georgetown University) แปลข้อความจากภาษารัสเซียจำนวน 49 ประโยคเป็นภาษาอังกฤษโดยใช้วงคำศัพท์ 250 คำและกฎไวยากรณ์ 6 กฎ (Hutchin et al., 1994)

การแปลภาษาด้วยเครื่องนั้นคือการพยายามที่จะนำคอมพิวเตอร์เข้ามาใช้ และช่วยในการแปลภาษา การทำให้เครื่องคอมพิวเตอร์สามารถแปลภาษาต่างๆ ที่มีอยู่ในโลกได้อย่างอัตโนมัติ มีความถูกต้องแม่นยำ โดยมีจุดประสงค์เพื่อช่วยในงานแปลที่มีปริมาณมาก เพื่อที่จะช่วยประหยัดปริมาณทรัพยากรบุคคลและทุนทรัพย์ในการแปลภาษา อย่างไรก็ตามในปัจจุบันการแปลภาษาด้วยเครื่องยังไม่บรรลุเป้าหมายนั้น เนื่องจากข้อจำกัดในด้านภาษาเช่นความแตกต่างทางด้านอรรถศาสตร์และวากยสัมพันธ์ ทำให้มีผลต่อการใช้งานทำให้ไม่เกิดประสิทธิภาพ และไม่ถูกต้องแม่นยำ

โดยทั่วไประบบการแปลภาษาด้วยเครื่องที่พัฒนาและใช้กันส่วนใหญ่ มักจะใช้กฎไวยากรณ์ของภาษาเป็นฐานความรู้ในการแปล และต้องใช้ข้อมูลและคลังคำศัพท์จากพจนานุกรมสองภาษาของทั้งภาษาต้นฉบับและภาษาเป้าหมาย ซึ่งมีข้อมูลทางด้านอรรถศาสตร์และวากยสัมพันธ์ของคำนั้นๆ ในบางกรณีเกิดความกำกวมในขั้นตอนการวิเคราะห์ข้อมูลทางด้านอรรถศาสตร์และวากยสัมพันธ์ของภาษาต้นฉบับ และในขั้นตอนของการสังเคราะห์ภาษาเป้าหมาย การนำพจนานุกรมสองภาษาเป็นฐานความรู้ในการแปลก็ไม่อาจแก้ไขและตัดสินปัญหาความกำกวมได้โดยอัตโนมัติ จึงจำเป็นจะต้องหาวิธีทางแก้ปัญหาความกำกวมโดยผู้พัฒนาระบบการแปลภาษาด้วยเครื่องเอง ซึ่งจะใช้การสร้างข้อยกเว้น (exception) หรือกฎไวยากรณ์เฉพาะสำหรับ

ประโยคที่มีความกำกวม ถึงกระนั้นก็ต้องสร้างข้อยกเว้น หรือกฎไวยากรณ์เฉพาะทุกครั้งที่พบปัญหาความกำกวมเพื่อแก้ไขและตัดสินปัญหาความกำกวมนั้นๆ ซึ่งเป็นเรื่องยุ่งยากและซับซ้อนที่ต้องแก้หรือสร้างกฎใหม่ทุกครั้งที่เกิดปัญหาความกำกวมขึ้น

การแก้ปัญหาความกำกวมนับเป็นปัญหาใหญ่ของการแปลภาษาด้วยเครื่อง มีวิธีการอีกอย่างหนึ่งที่สามารถช่วยการแก้ปัญหาความกำกวมได้ คือการใช้คลังข้อมูลเทียบบท (parallel corpus) โดยระบบคอมพิวเตอร์สามารถนำคลังข้อมูลเทียบบทมาเป็นฐานความรู้และเรียนรู้จากคลังข้อมูลเทียบบทมาช่วยในการแก้ปัญหาความกำกวมได้ และยังช่วยประหยัดเวลาในการแก้ไขปรับปรุงให้ใช้เวลาน้อยลง เนื่องจากไม่จำเป็นต้องแก้ไขปัญหาด้วยตนเองทุกครั้งที่เกิดปัญหาเพราะการวางระบบที่จะนำมาใช้กับคลังข้อมูลเทียบบทนั้น จะใช้เทคนิคการเรียนรู้ด้วยเครื่อง (machine learning) เพื่อให้เครื่องคอมพิวเตอร์เรียนรู้จากคลังข้อมูลเทียบบทด้วยตนเอง ดังนั้นการวางระบบในการดึงข้อมูลมาใช้ในการแปลจึงวางเพียงครั้งเดียว หากข้อมูลในคลังข้อมูลเทียบบทมีข้อมูลตัวอย่างภาษาเพียงพอและถูกต้อง ผลการแปลก็จะมีประสิทธิภาพและมีความแม่นยำสูง และนอกจากนี้เมื่อทำการวางระบบออกมาได้เสร็จสมบูรณ์ ก็จะสามารถนำมาประยุกต์ใช้กับคลังข้อมูลเทียบบทอื่นๆ เพื่อใช้ในงานแปลด้านอื่นได้อีกด้วย

ในปัจจุบันวงการการแปลภาษาด้วยเครื่องของภาษาไทยยังมีอยู่น้อยมากเมื่อเทียบกับวงการการแปลภาษาด้วยเครื่องในอเมริกาและยุโรป และส่วนมากจะเป็นการใช้กฎไวยากรณ์ของภาษาในการวิเคราะห์และสังเคราะห์ข้อมูลในการแปลเป็นหลัก เช่น งานการแปลภาษาด้วยเครื่องที่ชื่อว่า “ภาษิต[1]” ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (Sornlertlamvanich et al., 2000) ที่ใช้แนวทางการแปลโดยการวิเคราะห์และสังเคราะห์ข้อมูลจากกฎไวยากรณ์และความหมายจากพจนานุกรม และเปิดให้ใช้บริการอยู่ในอินเทอร์เน็ตให้ใช้งานได้โดยไม่เสียค่าใช้จ่าย งานการแปลภาษาด้วยเครื่องจากภาษาอังกฤษเป็นภาษาไทยที่ชื่อว่า “แปลไทย[2]” ที่ใช้แนวทางการแปลโดยการวิเคราะห์และสังเคราะห์ข้อมูลจากกฎไวยากรณ์และความหมายเช่นกัน และงานการแปลภาษาด้วยเครื่องจากภาษาอังกฤษเป็นภาษาไทยที่ชื่อว่า “Translation[3]” ของบริษัทไทยซอร์ฟแวร์เอ็นเทอร์ไพรส์จำกัด เป็นต้น

งานการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างโดยการสกัดแม่แบบการแปลจากคลังข้อมูลเทียบบทของภาษาไทยแบบในงานวิจัยชิ้นนี้ยังไม่เคยปรากฏ หรือเผยแพร่มาก่อน ดังนั้น

---

[1] http://www.suparsit.com/

[2] http://www.palthai.com/

[3] http://www.thaisoftware.co.th/index.php?tpid=pro:TSE-0111

แนวทางนี้จึงนับว่าเป็นแนวทางใหม่สำหรับการศึกษาและวิจัยในศาสตร์แห่งการประมวลผลภาษาธรรมชาติ

อย่างไรก็ตามงานแปลภาษาด้วยเครื่องแบบอิงตัวอย่างจากคลังข้อมูลเทียบบทจะมีประสิทธิภาพสูงได้นั้น จำเป็นต้องมีคลังข้อมูลเทียบบทที่มีข้อมูลปริมาณมาก และข้อมูลนั้นควรมีลักษณะการปรากฏกันซ้ำๆ แต่ปัจจุบันยังไม่มีคลังข้อมูลเทียบบทของภาษาไทยที่เป็นภาษาธรรมชาติที่มีปริมาณเพียงพอสำหรับใช้สำหรับใช้ในการทำวิจัยและทดลองแปลภาษาด้วยเครื่องแบบอิงตัวอย่างจากคลังข้อมูลเทียบบท ดังนั้นในการศึกษาวิจัยชิ้นนี้ จึงต้องทดลองกับภาษาเฉพาะทางที่มีวงคำศัพท์น้อยกว่าภาษาธรรมชาติ ทำให้คลังข้อมูลเทียบบทของภาษาเฉพาะทางที่จะนำมาใช้เป็นตัวอย่างในการทดลองวิจัยไม่จำเป็นต้องมีขนาดและปริมาณมากเท่ากับที่ต้องใช้เหมือนกับการประมวลผลภาษาธรรมชาติทั่วๆไป

ภาษาในตลาดหลักทรัพย์เป็นภาษาเฉพาะทางที่น่าศึกษาวิจัย และรายงานตลาดหลักทรัพย์เป็นรายงานประเภทหนึ่งที่มีข้อมูลงานแปลปริมาณมากและมีลักษณะการปรากฏซ้ำๆ ดังนั้นรายงานตลาดหลักทรัพย์จึงน่าสนใจที่จะนำมาเป็นกรณีศึกษาในงานแปลภาษาด้วยเครื่องแบบอิงตัวอย่างจากคลังข้อมูลเทียบบทชิ้นนี้

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 พัฒนาระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างเพื่อแปลรายงานตลาดหุ้นจากไทยเป็นอังกฤษ

1.2.2 วิเคราะห์ปัญหาที่เกิดจากการแปลแบบอิงตัวอย่างโดยใช้วิธีการทางภาษาศาสตร์

1.3 สมมติฐาน

1.3.1 ภาษาในรายงานตลาดหุ้นเป็นภาษาเฉพาะที่มีรูปแบบซ้ำๆ สามารถดึงตัวอย่างมาสร้างเป็นแม่แบบของการแปลในระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างได้

1.3.2 ระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างที่พัฒนานี้สามารถแปลรายงานตลาดหุ้นได้ถูกต้องได้ไม่ต่ำกว่าร้อยละ 85

1.3.3 การแปลแบบอิงตัวอย่างไม่สามารถแปลหน่วยภาษาในระดับปริจเฉทได้

1.4 เครื่องมือที่ใช้ในการวิจัย

1.4.1 เครื่องไมโครคอมพิวเตอร์และเครื่องคอมพิวเตอร์แบบพกพาส่วนบุคคล

1.4.2 โปรแกรมภาษา Python รุ่น 2.5 ของบริษัท Active State

1.5 ขอบเขตงานวิจัย

ใช้รายงานตลาดหุ้นแบบรายวัน สามประเภท ได้แก่ (1) รายงานประเภทการรับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียน และการรับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติม (2) รายงานประเภทตลาดหลักทรัพย์เพิ่มสินค้า และ (3) รายงานประเภทการขึ้นเครื่องหมายภายในตลาดหลักทรัพย์ เนื่องจากรายงานตลาดหุ้นแบบรายวันทั้งสามประเภทนี้มีลักษณะการใช้คำในวงจำกัดและมีรูปแบบการปรากฏของข้อความซ้ำๆ กันอย่างเห็นได้ชัด

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.6.1 ได้แนวทางในการพัฒนาระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่าสำหรับภาษาเฉพาะทางอื่นๆ

1.6.2 เป็นต้นแบบเพื่อนำไปพัฒนาระบบแปลภาษาด้วยเครื่องในเชิงพาณิชย์

1.7 วิธีดำเนินการวิจัย

1.7.1 เก็บตัวอย่างรายงานตลาดหุ้นทั้งที่เป็นภาษาไทยและที่แปลเป็นอังกฤษเพื่อจัดทำเป็นคลังข้อมูลเทียบบท

1.7.2 นำคลังข้อมูลภาษาไทยเข้าระบบตัดคำอัตโนมัติ

1.7.3 จับคู่ข้อความในคลังข้อมูลเพื่อสร้างเป็นคลังข้อมูลเทียบบท

1.7.4 พัฒนาระบบเพื่อค้นหาเอนทิตีระบุนามเพื่อให้ระบบรู้จำว่าเป็นคำเดียว

1.7.5 พัฒนาระบบเพื่อสร้างต้นไม้การปรากฏร่วมของแต่ละภาษา

1.7.6 พัฒนาระบบเพื่อจับคู่โหนดใบของต้นไม้การปรากฏร่วมสำหรับการสร้างแม่แบบการแปล

1.7.7 สร้างฐานข้อมูลเพื่อเก็บแม่แบบการแปลที่สกัดได้

1.7.8 พัฒนาระบบคัดเลือกแม่แบบการแปลที่ใช้แปลข้อวามรับเข้าได้

1.7.9 พัฒนาระบบรวมคำแปลใหม่เพื่อรวมคำแปลจากแม่แบบการแปลให้เป็นคำแปลที่สมบูรณ์

1.7.10 ทดสอบความถูกต้องของระบบสกัดแม่แบบการแปล

1.7.11 ตรวจสอบผลการสกัดแม่แบบการแปล

1.7.12 วิเคราะห์หาสาเหตุของผลการสกัดแม่แบบการแปลที่ผิดพลาดและประเมินผล

1.7.13 ทดสอบความถูกต้องของระบบรวมคำแปลใหม่ในการแปลข้อความรับเข้า

1.7.14 ตรวจสอบผลการแปล

1.7.15 วิเคราะห์หาสาเหตุของผลการแปลที่ผิดพลาดและประเมินผล

1.7.16 วิเคราะห์ข้อจำกัดของระบบและข้อเสนอแนะในการพัฒนาระบบ

1.8 โครงร่างของบทต่างๆ ของวิทยานิพนธ์

ในบทที่ 2 ผู้วิจัยได้ศึกษาทบทวนแนวคิดและทฤษฎีที่เกี่ยวข้องกับงานวิจัยชิ้นนี้ โดยแบ่งออกเป็นเรื่องแนวคิดและทฤษฎีการแปลภาษา (หัวข้อ 2.1) เรื่องทฤษฎีและแนวทางต่างๆการแปลภาษาด้วยเครื่อง (หัวข้อ 2.2) เรื่องลักษณะภาษาเฉพาะทางซึ่งเป็นภาษาที่จะใช้เก็บสะสมรวบรวมเป็นคลังข้อมูล (หัวข้อ 2.3) และศึกษาทบทวนเรื่องแนวคิดการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างโดยใช้แนวทางสกัดแม่แบบการแปล (หัวข้อ 2.4) บทที่ 3 จะนำเสนอวิธีการสร้างระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างโดยใช้แนวทางสกัดแม่แบบการแปลสำหรับแปลภาษาไทยเป็นภาษาอังกฤษ อันได้แก่การเก็บรวบรวมและลักษณะของคลังข้อมูลภาษา (หัวข้อ 3.1) การสร้างระบบการสกัดแม่แบบการแปล (หัวข้อ 3.2) และระบบการรวมคำแปลใหม่ (หัวข้อ 3.3) บทที่ 4 จะนำเสนอผลการทดลองของระบบ โดยแบ่งออกเป็น 2 ส่วน คือผลการทดลองสกัดแม่แบบการแปล (หัวข้อ 4.1) และผลการทดลองแปลข้อความ (หัวข้อ 4.2) บทที่ 5 จะนำเสนอผลการวิเคราะห์ปัญหาการทำงานของระบบซึ่งแยกเป็น 3 ปัญหาคือ ปัญหาของแนวทางสกัดแม่แบบการแปล (หัวข้อ 5.1) ปัญหาของคู่ภาษาภายในคลังข้อมูลเทียบบท (หัวข้อ 5.2) และปัญหาของการแปลภาษาตามหลักวิชาภาษาศาสตร์ (หัวข้อ 5.3) และสุดท้าย บทที่ 6 นำเสนอผลการทำงานของระบบเปรียบเทียบกับสมมติฐานที่ตั้งไว้ (หัวข้อ 6.1) และนำเสนอแนวทางในการพัฒนาระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างแนวทางสกัดแม่แบบการแปลต่อไป (หัวข้อ 6.2)

บทที่ 2

ทบทวนวรรณกรรม

ในบทนี้ จะกล่าวถึงงานที่เกี่ยวข้องกับการแปลภาษาด้วยเครื่อง ซึ่งผู้วิจัยได้จำแนกไว้เป็น 3 ส่วน คือ (2.1) ทฤษฎีการแปลภาษา ซึ่งเป็นหลักเกณฑ์ที่มนุษย์จะนำไปใช้ในการแปลภาษาและนำมาประเมินคุณภาพของการแปล (2.2) แนวทางการแปลภาษาด้วยเครื่อง ได้แก่ แนวทางในการแปลภาษาด้วยเครื่องแบบต่างๆ และจุดเด่น จุดด้อย ของแต่ละแนวทาง รวมทั้งการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างที่เป็นรากฐานในงานวิจัยนี้ (2.3) ภาษาเฉพาะทาง ซึ่งเป็นขอบเขตข้อมูลที่มักใช้ในการแปลภาษาด้วยเครื่อง โดยจะอธิบายถึงเรื่องคุณลักษณะของภาษาเฉพาะทางโดยเฉพาะอย่างยิ่งภาษาตลาดหุ้น ซึ่งเป็นฐานข้อมูลตัวอย่าง รายงานตลาดหุ้นซึ่งเป็นตัวอย่างข้อมูลที่จะนำมาแปลด้วยเครื่องของงานวิจัย

2.1 ทฤษฎีการแปลภาษา

สังคมปัจจุบันเป็นสังคมแห่งการสื่อสารข้อมูล การแปลทำให้มนุษย์สามารถสื่อสารความคิด ความต้องการข้ามภาษากันได้ วรรณา แสงอร่ามเรือง (2545) ได้ให้คำนิยามการแปลไว้ว่า “การแปลไม่ใช่การแปลงรหัสโดยผู้แปลทำหน้าที่เป็นผู้เปลี่ยนรหัสเท่านั้น เราไม่ได้แปลเนื้อความที่เรียงร้อยเป็นท่อนๆ แต่เราแปลตัวบท เราจะต้องมองตัวบททั้งหมดเป็นภาพรวม เนื่องจากภาษาไม่ได้เกิดขึ้นมาลอยๆ แต่เกิดจากสถานการณ์บางอย่างที่ทำให้ต้องมีการสื่อสาร มีกรอบวัฒนธรรมเข้ามาเกี่ยวข้องด้วย” (วรรณา, 2545: 7)

ในวงการการแปลได้มีการพัฒนาแนวคิดทฤษฎีมาอย่างต่อเนื่อง วรรณาได้กล่าวไว้ว่า “การศึกษาเกี่ยวกับการแปลในเชิงทฤษฎีมีมานานแล้ว” (วรรณา, 2545: 1) ในช่วงแรกเริ่มที่การแปลเริ่มได้รับความสนใจมากขึ้นนั้น การแปลส่วนมากมักจะเป็นการแปลวรรณกรรมและคัมภีร์ทางศาสนาเป็นภาษาต่างๆ เพื่อเผยแพร่ เช่นการแปลพระคัมภีร์ไบเบิ้ลจากภาษาลาตินเป็นภาษาอังกฤษของจอห์น ไวคลิฟไฟต์ (John Wycliffite) ในปีพ.ศ. 1927 ต่อมาการแปลได้รับการพัฒนาขึ้นมาเป็นการแปลเพื่อการสื่อสารในชีวิตประจำวันทั่วไป โดยทางทฤษฎีได้นิยามว่าการแปลเป็นการสื่อสารและจะอ้างถึงลักษณะของการสื่อสารอย่างใดอย่างหนึ่ง ดังเช่น ทฤษฎีของนิวมาร์ค (Newmark, 1988: 20) เชื่อมโยงการแปลกับการสื่อสาร โดยแยกทฤษฎีการแปลออกเป็น 2 สาขาคือ การแปลแบบตามความหมาย และการแปลเพื่อการสื่อสาร ส่วนทฤษฎีของสไตเนอร์ (Steiner, 1992: 47) ที่กล่าวว่าการแปลนั้นเกี่ยวข้องกับการสื่อสารหรือสื่อความในบริบทของความสัมพันธ์ระหว่างบุคคล และยังกล่าวอีกว่าแบบจำลองของกระบวนการแปลคือแบบจำลองของการสื่อสาร

ดังนั้น จะเห็นได้ว่าการแปลและการสื่อสาร มีความสัมพันธ์กันอย่างชัดเจน เนื่องจากเป็นพฤติกรรมระหว่างมนุษย์ในการแสดงออกซึ่งความหมายความรู้ทางด้านการสื่อสาร และจะช่วยให้ผู้แปลมีความเข้าใจเรื่องการแปลได้ดีขึ้นอีกด้วย (มณีรัตน์, 2548: 5)

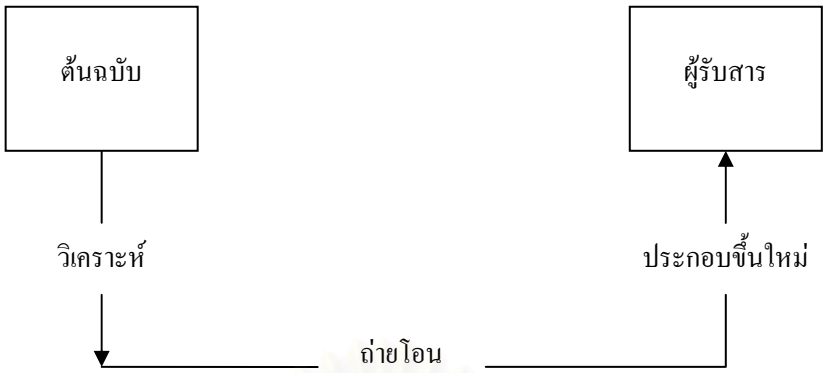
ในช่วงปลายศตวรรษที่ 15 หลังจากวงการแปลเริ่มแพร่หลายมีนักแปลเกิดขึ้นจำนวนมาก เอเตียน โดเลต์ (Etianne Dolet) นักปรัชญาชาวฝรั่งเศสได้ตีพิมพ์หลักการแปล La Manière de Bien Traduire d'une Langue en Autre ที่ถือเป็นหลักการแปลในรุ่นบุกเบิก (Basnett, 1991: 55; Snell-Hornby, 1995: 12) มีหลักในการแปลอยู่ 5 ประการ คือ (1) ผู้แปลต้องทำความเข้าใจอย่างถ่องแท้ ทั้งความรู้สึก และการการสื่อความหมายของเจ้าของต้นฉบับ ถึงแม้ว่ามีอิสระสามารถแปลเพิ่มเติมให้ชัดเจนกว่านี้ได้ (2) ผู้แปลควรมีความรู้ดีเยี่ยมทั้งภาษาต้นฉบับและภาษาเป้าหมาย (3) ผู้แปลควรหลีกเลี่ยงการแปลคำต่อคำ (4) ผู้แปลควรใช้รูปแบบภาษาที่นิยมใช้กันทั่วไป (5) ผู้แปลควรเลือกและเรียงคำให้เหมาะสมทำให้เกิดน้ำเสียงที่ถูกต้อง หลักการแปลของโดเลต์ ได้รับการยอมรับถูกนำไปอ้างอิง และถือเป็นแบบอย่างของนักแปลในสมัยนั้น

หลังจากที่หลักในการแปลของโดเลต์ได้เผยแพร่และเป็นที่ยอมรับในวงการการแปล จึงก่อให้เกิดความตื่นตัวทางด้านทฤษฎีการแปล และมีนักทฤษฎีได้สร้างทฤษฎีออกมา โดยมี 2 แนวทางหลัก คือ หลักการแปลโดยยึดแนวทางจากต้นฉบับ และหลักการแปลที่เน้นวัฒนธรรมปลายทาง

2.1.1 หลักการแปลโดยยึดแนวทางจากต้นฉบับ

โดยทั่วไปการแปลมักจะยึดหลักการอธิบายและประเมินโดยการเปรียบเทียบกับต้นฉบับ นักแปลจำนวนมากจะวิเคราะห์ความสัมพันธ์ระหว่างต้นฉบับกับงานแปล เช่น การหาคำแปลที่ตรงกัน เท่ากัน หรือ เทียบเคียงกัน (equivalence/correspondence) ไม่ว่าจะแปลคำหรือแปลความก็จะวนกลับไปหาต้นฉบับ โดยถือความซื่อสัตย์ต่อเนื้อความในต้นฉบับเป็นหลักหรือวิธีการแปล

ไนด้า (Nida, 1975: 79-97) ได้กล่าวถึงแนวทางในการแปลโดยยึดแนวทางจากต้นฉบับ และยึดแนวทางการตีความจากต้นฉบับเพื่อให้ผู้รับสารเข้าใจโดยมีแนวทางในการแปลคือในแต่ละภาษามีลักษณะที่แตกต่างกันและมีลักษณะเฉพาะตัว สิ่งที่พูดได้ในภาษาหนึ่งจะสามารถนำมาพูดได้ในอีกภาษาหนึ่ง แต่จะมีข้อจำกัดทางด้านรูปแบบ รูปแบบอาจจะต้องเปลี่ยนไป และการรักษาเจตนาของสารเอาไว้สำคัญกว่าการรักษารูปแบบ ถึงแม้ว่าการแปลจะเน้นที่เนื้อความของสาร แต่การเขียนขึ้นใหม่จะต้องไม่มากหรือน้อยกว่าต้นฉบับ ผู้แปลห้ามตีความหรือขยายความเกินกว่าจากความหมายที่แฝงไว้ในต้นฉบับ ขั้นตอนการแปลของไนด้าจะแสดงไว้ในรูปที่ 1



รูปที่ 1 แสดงแผนผังการแปลของไนด้า

ขั้นตอนการแปลของไนด้า (Nida and Taber, 1969: 106-107; Nida, 1975: 102-130) จะเริ่มที่การวิเคราะห์ไวยากรณ์ โดยใช้แนวความคิดเรื่องโครงสร้างประโยคผิว-ลึกของชอมสกี้ (Chomsky, 1965) มาใช้วิเคราะห์ความหมายในแง่ความหมายอ้างถึง (referential meaning) และความหมายแฝง (connotative meaning) หลังจากนั้นจะพิจารณาเรื่องการถ่ายโอนจากภาษาต้นฉบับไปยังภาษาเป้าหมายซึ่งจะต้องผ่านขั้นตอนการปรับโครงสร้างโดยเฉพาะการแปลสำนวนและคำอุปมา แล้วจึงเขียนสารขึ้นมาใหม่ในภาษาของผู้รับสาร โดยการเขียนสารขึ้นใหม่ต้องพิจารณาจาก 3 แง่มุมคือ (1) ความหลากหลายของภาษา หรือลีลาของภาษา (2) องค์ประกอบและลักษณะของลีลา และ (3) เทคนิคที่ใช้ในการผลิตชนิดของลีลา

ดังนั้นในมุมมองของไนด้า การแปลที่ดี จะอยู่ที่การคงสภาพความหมายไว้ (dynamic equivalence) โดยสามารถปรับเปลี่ยนรูปแบบหรือโครงสร้างของภาษาได้ (formal correspondence) การเรียบเรียงประโยคเขียนขึ้นใหม่ (paraphrase) แต่ไม่สามารถเพิ่มเติม ตัดทอน หรือบิดเบือนความหมายให้ไม่ตรงกับต้นฉบับได้

2.1.2 หลักการแปลที่เน้นวัฒนธรรมปลายทาง

เอช. แฟร์เมียร์ (H. Vermeer) (1998) นักปราชญ์ชาวเยอรมัน ได้นำเสนอทฤษฎีการแปลที่เน้นวัฒนธรรมปลายทาง นั่นคือ ทฤษฎีสโกปอส (Skopos Theory หรือ Skopostheorie ในภาษาเยอรมัน) (Reiß and Vermeer, 1984) ซึ่งคำว่า “skopos” เป็นคำในภาษากรีกแปลว่า จุดประสงค์ หรือ เป้าหมาย ทฤษฎีดังกล่าวว่าด้วยการแปลที่เน้นความสำคัญและหน้าที่ของการแปล รวมถึงเจตนา จุดประสงค์ และเป้าหมายของการแปล แก่นความคิดของทฤษฎีนี้ได้นิยามการแปลว่าเป็นการถ่ายทอดข้อเท็จจริงของเนื้อความตามวัฒนธรรมเป้าหมาย (Translations as facts of target culture) จากทฤษฎีดังกล่าว ทัวรี (Toury) (1995) ได้นำเสนอเพิ่มเติมว่า การแปลเกิดขึ้นภายใน

สิ่งแวดล้อมของวัฒนธรรมผู้รับ และเป็นไปตามความต้องการบางอย่างของวัฒนธรรมปลายทาง ดังนั้นผู้แปลจึงมีจุดมุ่งหมายแรกที่สำคัญที่สุด คือการอ้างอิงกับวัฒนธรรมปลายทางของผู้รับสาร

ทฤษฎีการแปลที่เน้นวัฒนธรรมปลายทางอธิบายว่า เนื่องจากไม่ว่าจะมีหน้าที่หรือลักษณะอย่างไร งานแปลก็เป็นส่วนหนึ่งของวัฒนธรรมที่งานแปลเข้าไปอยู่ และสะท้อนให้เห็นถึงข้อเท็จจริงของวัฒนธรรมนั้นๆ ซึ่งจะดูที่ความสัมพันธ์ของหน้าที่ (function) กระบวนการ (process) และงานแปลกับต้นฉบับ (product) ที่เกิดขึ้นในวัฒนธรรมเป้าหมาย หลักสำคัญในการแปลจึงต้องหาระเบียบวิธี แบบแผน หรือ กฎ (norm/law) ที่เป็นตัวชี้หรือกำหนดขอบเขต (parameter) ของความสัมพันธ์เหล่านั้น ตั้งแต่ตำแหน่งและหน้าที่ในวัฒนธรรมที่การแปลอยู่ ซึ่งเป็นตัวกำหนดกระบวนการทำให้เปลี่ยนจากต้นฉบับมาเป็นงานแปลที่เน้นวัฒนธรรมปลายทาง

การให้ความสำคัญกับการแปลที่เน้นวัฒนธรรมปลายทางทำให้มีแนวคิดว่างานแปลทำให้เกิดการเปลี่ยนแปลงในวัฒนธรรมเป้าหมาย (Toury, 1995: 28) กล่าวคือ เมื่ออ่านงานแปลจะได้รับวัฒนธรรมอื่นที่แฝงอยู่ในงานแปลและทำให้เกิดการเปลี่ยนแปลงในวัฒนธรรมผู้รับด้วย ซึ่งเกิดจากข้อเท็จจริงที่ว่า ในขณะที่งานแปลมีจุดมุ่งหมายที่จะตอบสนองความต้องการของวัฒนธรรมผู้รับ งานแปลเองก็เบี่ยงเบนออกจากแบบแผนวัฒนธรรมนั้นๆ ด้วย ไม่ใช่เพราะความซื่อสัตย์ต่อต้นฉบับหรือการที่จะรักษาลักษณะของต้นฉบับไว้ แต่เป็นหลักของการแปลที่ยึดแนวคิดการปรับให้เข้ากับวัฒนธรรม (cultural license) ที่จะเป็นลักษณะบ่งบอกว่าเป็นงานที่แปลมาและบ่งชี้ให้เห็นความแตกต่างของงานแปลกับงานที่ไม่ใช่งานแปล

2.1.3 ปัญหาของการแปล

ในมุมมองของทั้งโดเลต์ และไนด้า การแปลที่ดีต้องสามารถรักษาความหมายของต้นฉบับไว้ได้ ไม่ควรเป็นการแปลแบบคำต่อคำ และสามารถปรับเปลี่ยนโครงสร้างหรือรูปแบบให้เหมาะสมได้ การแปลที่ยึดแนวทางจากต้นฉบับส่วนมากจะเป็นงานแปลประเภทคัมภีร์ทางศาสนา วรรณกรรม บทความทางวิชาการ เป็นต้น ส่วนการแปลที่ยึดแนวทางที่เน้นวัฒนธรรมปลายทางของแฟเมียร์และทัวรีมีมุมมองที่จะให้ประโยชน์แก่ผู้รับสารที่จะเข้าใจงานแปลนั้นได้ง่ายที่สุด เช่น ป้ายประกาศในรถไฟในประเทศต่างๆ ที่เขียนไว้ในรถไฟเป็นภาษาอังกฤษ ที่ประกาศไว้ตรงสัญญาณหยุดรถฉุกเฉิน ในแต่ละประเทศจะแตกต่างกันตามวัฒนธรรมของประเทศนั้น (มณีรัตน์, 2548: 94) เช่นในประเทศอังกฤษ เขียนว่า ‘Alarm signal/To stop train pull handle/Penalty £ 50 for improper use’ ส่วนที่ประเทศเยอรมัน จะเขียนไว้ว่า ‘Emergency break/Pull brake only in/case of emergency/Any misuse will be punished’ ในแคนาดาเขียนว่า ‘Conductor’s valve/emergency only’ ในอิตาลีเขียนว่า ‘Alarm/Pull the handle/in case of danger/penalties for

improper use' ส่วนในประเทศไทยเขียนว่า 'EMERGENCY DOOR RELEASE/PULL TO OPERATE -/THEN SLIDE DOORS OPEN/PENALTY FOR MISUSE' เป็นต้น ซึ่งจะเห็นได้ว่าการเขียนป้ายประกาศเพื่อจุดประสงค์แบบเดียวกันแต่ก็เขียนต่างกันตามลีลาของวัฒนธรรมของแต่ละประเทศ

ปัญหาที่มักจะพบในการแปลภาษาคือการที่ไม่สามารถหาคำศัพท์ที่ตรงกันเท่ากัน หรือ เทียบเคียงกันได้อย่างชัดเจน หรือในกรณีที่เป็นสำนวนและคำอุปมาอุปไมย การตีความหรือวิเคราะห์ก็จะถูกจำกัดอยู่ที่ความสามารถและความรู้ของผู้แปล เช่น ประโยคภาษาอังกฤษว่า 'We have put Man into the Red Planet' ประโยคนี้หากแปลแบบทั่วไปก็จะแปลได้ว่า 'เราส่งมนุษย์ลงไปยังดาวสีแดงแล้ว' ซึ่งในความเป็นจริงแล้ว 'Red Planet' นั้นหมายถึงดาวอังคาร นอกจากนี้ยังมีเรื่องของลีลาในการเขียน เช่นประโยคทางวรรณกรรม (ภควดี, 2540: 10) 'There are no doors that forbid access to the scriptorium from the kitchen and the refectory, or to the library from the scriptorium.' คำว่า 'doors' โดยทั่วไปแล้ว จะแปลว่าประตู แต่การแปลด้วยคำตรงตัวเช่นนั้นจะทำให้เสียลีลาของบทความได้ ประโยคนี้ภควดี วีระภาสพงศ์ได้แปลไว้ว่า 'ไม่มีทวารใด ขวางกั้นทางเข้าสู่หออาลักษณ์จากโรงครัวและโรงอาหาร หรือทางเข้าสู่หอสมุดจากหออาลักษณ์นี้'

ดังนั้นความสามารถในการวิเคราะห์ของผู้แปลจึงเป็นปัจจัยสำคัญในการแปลที่จะสามารถสื่อความหมายได้สอดคล้องกับสารต้นฉบับได้ อย่างไรก็ตามการใช้ทฤษฎีทางภาษาศาสตร์แขนงต่างๆ เข้ามาช่วยในการแปล จะทำงานแปลมีประสิทธิภาพมากขึ้น ทฤษฎีทางภาษาศาสตร์ที่จะนำมาใช้ ได้แก่ อรรถศาสตร์ที่จะเข้ามาช่วยในการหาความสัมพันธ์ทางความหมายของคำศัพท์ในประโยค วากยสัมพันธ์ที่จะเข้ามาช่วยในการหาความสัมพันธ์ของประโยค กาล และวจนะที่จะช่วยสื่อความหมายของประโยคได้อย่างถูกต้อง วัจนปฏิบัติศาสตร์ที่จะเข้ามาช่วยในการหาจุดประสงค์หรือความหมายแฝงของผู้ส่งสารที่ซ่อนอยู่ภายในปริเฉท ซึ่งจะช่วยให้สารที่ได้รับการแปลมีเนื้อหาตามที่ผู้ส่งสารเจตนาจะส่งอย่างถูกต้องไม่ผิดเพี้ยนไป

ในการแปลนั้นอาจกล่าวได้ว่า สามารถแบ่งออกได้เป็น 2 ลักษณะคือ งานแปลที่ไม่ต้องการความสละสลวยมีลักษณะซ้ำๆ และมีปริมาณมาก เช่น คู่มือการใช้งาน ป้ายประกาศ คำเตือน ฯลฯ และงานแปลที่ต้องการความสละสลวย มีลีลา และมีความซับซ้อน เช่น งานวรรณกรรม บทความทางวิชาการ ฯลฯ จะเห็นได้ว่าความสามารถในการแปลภาษาของมนุษย์นั้นสามารถแปลงานที่สลับซับซ้อนที่ต้องการความสละสลวย มีลีลาได้ แต่นักแปลจำเป็นต้องใช้เวลาเรียนรู้สิ่งต่างๆ มากมายเพื่อที่จะมาเป็นนักแปลที่มีความสามารถ การทำให้คอมพิวเตอร์มีความสามารถดังกล่าวเทียบเท่ามนุษย์เพื่อมาช่วยในการแปลงานที่สลับซับซ้อนเช่นนั้นจึงเป็นเรื่องยากและมีข้อจำกัด และถึงแม้ว่าจะเป็นนักแปลที่มีความสามารถก็ยังต้องใช้เวลาในการแปลสารที่มีปริมาณมาก แต่การ

แปลภาษาด้วยเครื่อง (Machine Translation/MT) นั้นถึงแม้จะไม่สามารถแปลไปถึงระดับการแปลของมนุษย์ที่ละเอียดอ่อนได้ แต่ก็สามารถช่วยผ่อนแรงของนักแปลให้ทำงานได้ง่ายและสะดวกขึ้น หากนำไปใช้ในงานแปลที่ไม่ต้องการความสละสลวย มีลักษณะซ้ำๆ และมีปริมาณมาก

2.2 การแปลภาษาด้วยเครื่อง

การแปลภาษาด้วยเครื่องคือ การนำเครื่องคอมพิวเตอร์มาใช้สำหรับแปลข้อความ จากภาษาหนึ่งไปเป็นอีกภาษาหนึ่ง โดยเมื่อป้อนข้อมูลเข้าภาษาหนึ่งเข้าไป เครื่องจะวิเคราะห์ข้อมูลภาษาที่เข้าไป และตัดสินใจเลือกคำแปลและสร้างข้อความแปลออกมา

การแปลภาษาด้วยเครื่องนั้นไม่สามารถทำการแปลข้อความหรือบทความให้เกิดประโยคที่มีความไพเราะและสละสลวยได้เหมือนกับการแปลของนักแปลทั่วไป เช่นการแปลกาพย์โคลงกลอนต่างๆ เนื่องจากการแปลภาษาด้วยเครื่อง มีข้อจำกัดในการแปลมาก ทั้งความรู้เกี่ยวกับโลก หรือความทันต่อเหตุการณ์ปัจจุบันที่ไม่สามารถที่จะทำให้เครื่องมีความรู้ได้เท่ากับมนุษย์จริงๆ ทำให้การแปลภาษาด้วยเครื่องไม่สามารถทดแทนนักแปลได้ แต่การแปลภาษาด้วยเครื่องนั้นจะสามารถช่วยแปลข้อความ หรือบทความที่ไม่ต้องการความไพเราะและสละสลวย และจะช่วยทุ่นแรงนักแปลได้มาก ยิ่งไปกว่านั้นจะช่วยในการประหยัดเวลาและค่าใช้จ่าย หากต้องมีการแปลข้อมูลที่มีเนื้อหาใกล้เคียงกัน เช่นคู่มือการใช้เครื่องอิเล็กทรอนิกส์ต่างๆ หรือเป็นงานแปลซ้ำซากที่มีปริมาณมาก เกิดขึ้นทุกวัน เพราะการจ้างนักแปลมาแปลข้อมูลเหล่านี้จะต้องใช้ค่าใช้จ่ายและกำลังคนสูง แต่ถ้าสามารถใช้เครื่องที่จะแปลภาษา แล้วนำผลที่ได้มาตรวจสอบโดยนักแปล ก็จะช่วยประหยัดเวลาและค่าใช้จ่ายได้

อาร์โนลด์และคณะ (Arnold et al., 2001)ได้แบ่งการแปลภาษาด้วยเครื่องตามลักษณะการทำงานที่แตกต่างกันของระบบออกเป็น 3 กลุ่มใหญ่ได้แก่ (1) การแปลภาษาด้วยเครื่องแบบใช้กฎ (Rule-based Machine Translation/RBMT) (2) การแปลภาษาด้วยเครื่องแบบใช้สถิติ (Statistical Machine Translation/SMT) และ (3) การแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง (Example-Based Machine Translation/EBMT) ซึ่งแนวทางต่างๆ ของงานการแปลภาษาด้วยเครื่อง มีความแตกต่างกันตามลักษณะเด่นของวิธีการวิเคราะห์ข้อมูลของงานการแปลภาษาด้วยเครื่องนั้น โดยในอดีตแนวทางแรกที่ทดลองวิจัยคือการสร้างการแปลภาษาด้วยเครื่องแบบใช้กฎ แล้วจึงพัฒนาต่อมาเป็นแนวทางใหม่ๆ เป็นการแปลภาษาด้วยเครื่องแบบใช้สถิติและการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง

2.2.1 การแปลภาษาด้วยเครื่องแบบใช้กฎ

อาร์โนลด์และคณะ (Arnold et al., 2001)ได้อธิบายไว้ว่า การแปลภาษาด้วยเครื่องแบบใช้กฎ (Rule-based Machine Translation/RBMT) เป็นแนวทางที่สามารถเรียกได้อีกอย่างว่าการแปลภาษาด้วยเครื่องโดยใช้ความรู้ทางภาษาศาสตร์ (Linguistic Approach Machine Translation) เนื่องจากเป็นแนวทางที่ต้องใช้ความรู้ทางภาษาศาสตร์เข้ามาใช้ในการวางระบบให้คอมพิวเตอร์ทำงาน เช่น วากยสัมพันธ์ อรรถศาสตร์ และวัจนปฏิบัติศาสตร์ และในแนวทางนี้ต้องใช้ผลต่างๆที่ได้จากการทำการแยกคุณลักษณะทางภาษาศาสตร์ ส่งต่อข้อมูลที่ผ่านการวิเคราะห์ข้อมูลภาษาต้นฉบับทางภาษาศาสตร์แล้ว ไปสู่ส่วนการวิเคราะห์ข้อมูลภาษาเป้าหมาย จึงมีการเรียกงานการแปลภาษาด้วยเครื่องตามแนวทางนี้ อีกอย่างว่าการแปลภาษาด้วยเครื่องแบบส่งผ่าน (Transfer Machine Translation) ซึ่งสามารถแบ่งย่อยไปตามระดับการวิเคราะห์ทางภาษาศาสตร์ที่ใช้ โดยอาร์โนลด์และคณะ (Arnold et al., 2001) ได้แยก การแปลภาษาด้วยเครื่องแบบใช้กฎออกเป็น 3 กลุ่มได้แก่ (1) การแปลภาษาด้วยเครื่องแบบส่งผ่านทางวากยสัมพันธ์ (Syntactic Transfer Machine Translation) (2) การแปลภาษาด้วยเครื่องแบบส่งผ่านทางอรรถศาสตร์ (Semantics Transfer Machine Translation) และ (3) การแปลภาษาด้วยเครื่องแบบส่งผ่านทางภาษากลาง (Interlingual Machine Translation)

นอกจากนี้ บีเวน (Beaven, 1992) ได้เสนอแนวทางที่แตกต่างในการการแปลภาษาด้วยเครื่องแบบใช้กฎไว้อีกแนวทางหนึ่ง คือ การแปลภาษาด้วยเครื่องแบบส่งผ่านระดับคำ (Lexicalist Machine Translation)

แนวทางทั้ง 4 อย่างนี้ ผู้พัฒนาต้องมีความรู้ทางภาษาศาสตร์มากในการพัฒนาโปรแกรมให้มีประสิทธิภาพและความแม่นยำสูง แต่อย่างไรก็ตามก็ไม่สามารถพัฒนาไปจนถึงระดับการแปลของมนุษย์ที่มีสามารถแปลความตามปริจเฉทหรือแปลสำนวนที่มีความหมายซับซ้อนได้ ซึ่งเป็นผลเนื่องมาจากความสามารถในการวิเคราะห์ภาษาของมนุษย์ที่ลึกซึ้ง และความซับซ้อนของภาษาที่ใช้กันอยู่ในปัจจุบัน

2.2.1.1 การแปลภาษาด้วยเครื่องแบบส่งผ่านทางวากยสัมพันธ์

เวย์ (Way, 2001) ได้อธิบายไว้ว่าการแปลภาษาด้วยเครื่องตามการแปลภาษาด้วยเครื่องแบบส่งผ่านทางวากยสัมพันธ์ ผู้พัฒนาโปรแกรมจะต้องใช้ความรู้และทฤษฎีภาษาศาสตร์ทางด้านวากยสัมพันธ์เป็นหลักในการพัฒนาระบบการแปล โดยที่เริ่มต้นผู้พัฒนาระบบจะต้องศึกษาความสอดคล้องของรูปแบบทางวากยสัมพันธ์ในภาษาต้นฉบับ (source language หรือ SL) และภาษาเป้าหมาย (target language หรือ TL) กล่าวคือต้องค้นหาว่าในภาษาทั้ง 2 ภาษามีการ

เรียงตัวทางด้านวากยสัมพันธ์ออกมาเป็นโครงสร้างต้นไม้ (tree structure) อย่างไร เช่นเป็นภาษาที่มีการเรียงตัวแบบ ประธาน กริยา กรรม หรือ ประธาน กรรม กริยา เป็นต้น แล้วต้องวิเคราะห์หาว่าการเรียงตัวของชนิดของคำ (part of speech/POS) มีความสัมพันธ์กันแบบใด จากนั้นจึงสร้างโครงสร้างทางภาษาเหล่านั้นออกมาเป็นโครงสร้างต้นไม้ (tree structure) แล้วนำโครงสร้างต้นไม้ในแบบต่างๆทั้งหมดนั้น มาสร้างเป็นกฎให้เครื่องคอมพิวเตอร์เรียนรู้ จะได้ประมวลผลได้ว่าข้อมูลที่จะทำการเปรียบเทียบหาคำแปลเป็นโครงสร้างต้นไม้แบบใด แล้วจึงส่งผ่านข้อมูลโครงสร้างต้นไม้ที่ได้ไปแปลงเป็นโครงสร้างต้นไม้ในรูปแบบของภาษาเป้าหมาย

เมื่อทำการถ่ายโอนข้อมูลโครงสร้างต้นไม้และแปลงโครงสร้างต้นไม้ของภาษาต้นฉบับเป็นโครงสร้างต้นไม้ของภาษาเป้าหมายแล้ว ก็จะนำคำศัพท์แต่ละคำมาทำการเปรียบเทียบหาคำแปลโดยใช้พจนานุกรมสองภาษา

อย่างไรก็ตามแนวทางนี้มักพบปัญหาเรื่องการแปลคำศัพท์ เนื่องจากมักจะมีความกำกวมเกิดขึ้นในการแปลคำศัพท์ ในกรณีที่คำศัพท์นั้นแปลได้หลายแบบและเครื่องไม่สามารถตัดสินใจได้เองว่าจะเลือกแปลเป็นคำใด และถึงแม้ว่าส่วนใหญ่ผู้พัฒนาโปรแกรมจะสร้างพจนานุกรมสองภาษาเองเพื่อป้องกันการผิดพลาดในการดึงคำศัพท์แต่ปัญหาความกำกวมนี้ก็เป็นปัญหาอย่างมากสำหรับแนวทางนี้

และปัญหาอีกอย่างที่พบได้บ่อยครั้งคือความกำกวมของโครงสร้างทางวากยสัมพันธ์ของภาษาที่ซับซ้อนและสามารถแบ่งได้มากมาย เช่นปัญหาความกำกวมที่เกิดจากการใช้บุพบทวลีเป็นส่วนขยาย (prepositional phrase attachment) เมื่อนามวลีมีบุพบทจำนวนมากอยู่ภายใน ทำให้เกิดความกำกวมว่าบุพบทภายในนามวลีนั้นขยายคำนามตัวใด ทำให้เกิดความกำกวมในการแจกแจงโครงสร้างทางวากยสัมพันธ์ ผลกระทบที่ตามมาคือ เครื่องอาจจะไม่สามารถเลือกโครงสร้างที่ถูกต้องได้ ทำให้ผลลัพธ์การแปลผิดพลาดได้

นอกจากนี้การแจกแจงรายละเอียดโครงสร้างทางวากยสัมพันธ์ของภาษาต้นฉบับและภาษาเป้าหมายก็ยังพบปัญหาสำคัญอีกอย่างคือ ทั้งภาษาต้นฉบับและภาษาเป้าหมายมีลักษณะวากยสัมพันธ์ที่แตกต่างกันมาก เช่นการแปลระหว่างภาษาที่แยกและไม่แยกพจน์ของคำนาม คู่ภาษาที่มีและไม่มีลักษณะนาม คู่ภาษาที่ใช้และไม่ใช้ สรรพนามไร้รูป (zero anaphora) คู่ภาษาที่มีรูปแบบการเรียงตัวของกริยา (verb serialization) ไม่เหมือนกัน ฯลฯ ซึ่งเป็นการยากที่จะหาวิธีการแปลงและเติมเต็มความแตกต่างระหว่างโครงสร้างต้นไม้ของภาษาต้นฉบับกับโครงสร้างต้นไม้ของภาษาเป้าหมาย

ปัญหาสำคัญที่พบในการแปลภาษาด้วยเครื่องแบบใช้กฎทางวากยสัมพันธ์

- โครงสร้างทางวากยสัมพันธ์ของภาษามักมีความกำกวม
- โครงสร้างทางวากยสัมพันธ์และโครงสร้างต้นไม้ของภาษามีมากจนไม่สามารถแจกแจงได้หมด
- โครงสร้างทางวากยสัมพันธ์ของภาษาต้นฉบับและภาษาเป้าหมายไม่สามารถหากฎมาส่งต่อข้อมูลได้
- ประโยคในภาษาต้นทางอาจไม่สามารถวิเคราะห์ส่วนประชิดได้

2.2.1.2 การแปลภาษาด้วยเครื่องแบบส่งผ่านทางอรรถศาสตร์

อัลเลน (Allen, 1995) ได้อธิบายไว้ว่าการแปลภาษาด้วยเครื่องแบบส่งผ่านทางอรรถศาสตร์ ผู้พัฒนาโปรแกรมจะต้องใช้ความรู้และทฤษฎีภาษาศาสตร์ทางด้านอรรถศาสตร์เป็นหลัก และต้องเชี่ยวชาญในการใช้ตรรกรูป (logical Form) ซึ่งเป็นลักษณะหนึ่งของรูปแทนความหมาย (semantics representation) โดยตรรกรูปนี้สามารถเลือกใช้ได้ตามถนัดของผู้พัฒนาโปรแกรม แต่อย่างไรก็ตามผู้พัฒนาโปรแกรมส่วนมากมักนิยมที่จะใช้ ตรรกรูปเสมือน (Quasi-logical Form/QLF) ซึ่งจะช่วยลดความหลากหลายในการสร้างโครงสร้างต้นไม้เพราะตรรกรูปจะช่วยบ่งชี้ว่าคำใดมีหน้าที่ใดในประโยค เช่นคำนามนี้ เป็นผู้กระทำ (agent) หรือ ผู้ได้รับผลของการกระทำ (patience) เป็นต้น นอกจากนั้นยังมีการกำหนดอรรถลักษณ์ (semantic feature) กำกับคำศัพท์ซึ่งต้องทำพจนานุกรมของทั้งสองภาษาที่มีการกำกับอรรถลักษณ์ไว้ก่อน เพื่อที่จะสามารถดึงคำศัพท์หลังจากที่ส่งผ่านโครงสร้างต้นไม้ของภาษาต้นฉบับไปแปลงเป็นโครงสร้างต้นไม้ของภาษาเป้าหมาย โดยใช้การดึงคำศัพท์ที่มีอรรถลักษณ์เดียวกันออกมาจากพจนานุกรมภาษาเดี่ยว (monolingual dictionary) มาแทนที่คำศัพท์เพื่อเป็นการแปลคำศัพท์

ข้อดีของการใช้ตรรกรูปคือสามารถที่จะได้โครงสร้างต้นไม้ของภาษาต้นฉบับโดยที่ไม่ต้องเสียเวลาวิเคราะห์โครงสร้างทางวากยสัมพันธ์

ปัญหาสำคัญที่พบในแนวทางการแปลภาษาด้วยเครื่องแบบส่งผ่านทางอรรถศาสตร์

- ประโยคที่เป็นการเปรียบเทียบเปรียบเปรยไม่สามารถแปลงเป็นตรรกรูปได้อย่างสมบูรณ์ และไม่สามารถหากฎมารองรับประโยคที่เป็นการเปรียบเทียบ

- ไม่สามารถใช้ได้กับภาษาพูด ประเภทคำอุทานได้ เนื่องจากไม่สามารถเขียนออกมาเป็นตรรกรูปได้
- ไม่เหมาะกับภาษาที่มีโครงสร้างทางวากยสัมพันธ์ที่ซับซ้อน เช่น ภาษาที่มีโครงสร้างของกริยาเรียง (serial verb construction) เพราะมีภาคแสดงหลายตัวสัมพันธ์กันอยู่

2.2.1.3 การแปลภาษาด้วยเครื่องแบบส่งผ่านภาษากลาง

"การแปลภาษาด้วยเครื่องแบบส่งผ่านภาษากลาง (Interlingual Machine Translation) คือแนวทางที่ว่าทุกภาษามีรูปแบบพื้นฐานทางโครงสร้างต้นไม้ทางอรรถศาสตร์เหมือนกัน และการแปลภาษาด้วยเครื่องก็นำแนวความคิดนี้มาใช้เป็นแนวทางหนึ่งในงานวิจัยด้วย โดยที่จะเริ่มจากการหาโครงสร้างที่เป็นกลางและสากลที่สุดที่ทุกภาษาสามารถมีได้ร่วมกันก่อน แล้วจึงไปวิเคราะห์หน่วยคำ (morphological analysis) เพื่อให้ได้โครงสร้างของคำ (word structure) แล้วจึงวิเคราะห์วากยสัมพันธ์ (syntactic analysis) เพื่อให้ได้โครงสร้างทางวากยสัมพันธ์ หลังจากนั้นจึงนำผลที่ได้มาวิเคราะห์ทางอรรถศาสตร์ (semantic analysis) หารูปแบบโครงสร้างทางอรรถศาสตร์ (semantic structure) ของภาษาต้นฉบับและภาษาเป้าหมายที่สอดคล้องกับโครงสร้างทางอรรถศาสตร์ที่เป็นกลางและสากล โดยใช้โครงสร้างที่เป็นกลางนั้นเป็นศูนย์กลางเพื่อส่งผ่านข้อมูลที่ได้ไปสังเคราะห์สร้างโครงสร้างทางอรรถศาสตร์ วากยสัมพันธ์ และคำของภาษาเป้าหมายตามลำดับ" (Dorr et al., 2004) กรรมวิธีการแปลภาษาด้วยเครื่องแบบส่งผ่านภาษากลางจะแสดงไว้ในรูปที่ 2

รูปที่ 2 แสดงการทำงานของระบบการแปลภาษาด้วยเครื่องแบบใช้กฎโดยส่งผ่านภาษากลาง

ข้อดีของการแปลภาษาด้วยเครื่องแบบใช้กฎโดยส่งผ่านภาษากลางคือการที่สามารถเขียนกฎต่างๆ ขึ้นเพียงครั้งเดียว ก็สามารถนำไปใช้ได้กับภาษาหลากหลาย เนื่องจากมีโครงสร้างทางอรรถศาสตร์ของภาษากลาง (Interlingua) เป็นศูนย์กลางของทุกๆ ภาษา กล่าวคือหากมีโครงสร้างทางอรรถศาสตร์และผลการของวิเคราะห์หน่วยศัพท์และวิเคราะห์ศัพท์ของภาษาไทย และภาษาอังกฤษอยู่ และต่อมาโครงสร้างต้นไม้ทางอรรถศาสตร์และผลการของวิเคราะห์หน่วยศัพท์และวิเคราะห์ศัพท์ของภาษาญี่ปุ่นและจีน ก็จะสามารถแปลข้อความ จากภาษาไทยเป็น

ภาษาญี่ปุ่น ภาษาญี่ปุ่นเป็นภาษาจีน ภาษาอังกฤษเป็นภาษาไทยได้โดยไม่ต้องเขียนระบบขึ้นใหม่ ทำให้ประหยัดมากกว่าเมื่อต้องการแปลหลายๆ ภาษา

ปัญหาที่พบคือ สามารถกำหนดภาษากลาง (Interlingua) ได้จริงหรือ ทุกภาษามีรูปแบบพื้นฐานทางอรรถศาสตร์ที่เป็นกลางและสากลจริงหรือ เป็นไปได้ว่ากำหนดโครงสร้างของภาษากลางขึ้นมาใช้กับภาษาคู่หนึ่งอาจเป็นเรื่องที่สามารถทำได้ แต่เมื่อนำมาใช้กับภาษาอื่นๆ อาจจะมีโครงสร้างบางอย่างที่มีลักษณะและรูปแบบที่ต่างออกไป ทำให้จริงๆ แล้วภาษากลางอาจไม่มีทางสมบูรณ์แบบได้ ดังนั้นการที่จะสร้างภาษากลางที่จะเหมาะสมกับการแปลภาษาทุกภาษาด้วยเครื่องจึงทำได้ยาก และพบว่ามักจะก่อให้เกิดความผิดพลาดได้ง่าย เนื่องจากกฎที่สร้างขึ้นมามีความขัดแย้งกันเองซึ่งเป็นผลมาจากความกำกวมและซับซ้อนในโครงสร้างภาษาต่างๆ ที่มีลักษณะและโครงสร้างที่แตกต่างกัน หรือกฎและโครงสร้างทางอรรถศาสตร์ที่สร้างขึ้นไม่ครอบคลุมลักษณะโครงสร้างภาษาทั้งหมด และในการพัฒนา แก้ไขปรับปรุงก็ทำได้ยากลำบากมาก เนื่องจากกฎและโครงสร้างทางอรรถศาสตร์ต่างๆ ที่กำหนดขึ้น เมื่อต้องมีการแก้ไขหรือเพิ่มเติมก็อาจจะต้องเขียนขึ้นมาใหม่หรือแก้ไขที่โครงสร้างพื้นฐานที่อาจจะก่อให้เกิดผลกระทบกับโครงสร้างอื่นๆ อีกเป็นจำนวนมาก

นอกจากนั้นยังโดนกล่าวแย้งข้อดีที่ว่าจะประหยัดกว่าเมื่อใช้แปลหลายๆ ภาษาว่า “การแปลภาษาด้วยเครื่องแบบส่งผ่านภาษากลางไม่ใช่แนวทางเดียวที่จะช่วยประหยัดเวลาในการสร้างระบบ เนื่องจากการแปลภาษาด้วยเครื่องแบบใช้กฎระบบอื่นๆ ก็สามารถใช้ภาษาอังกฤษ เป็นภาษาหลักในการแปลหลายๆ ภาษาได้เช่นเดียวกัน กล่าวคือ การแปลภาษาไทยเป็นภาษาอังกฤษก่อนแล้วจึงแปลภาษาอังกฤษเป็นภาษาฝรั่งเศส ก็มีผลเท่ากับการแปลภาษาไทยเป็นภาษาฝรั่งเศสเหมือนกัน” (Way, 2001)

2.2.1.4 การแปลภาษาด้วยเครื่องแบบส่งผ่านระดับคำ

บีเวน (Beaven, 1992) ได้กล่าวไว้ว่าการแปลภาษาด้วยเครื่องแบบส่งผ่านระดับคำเป็นการแปลภาษาด้วยเครื่องที่ไม่ใช้ความรู้ทางด้านโครงสร้างต้นไม้ทางวากยสัมพันธ์มาช่วยในการแยกองค์ประกอบการแปล แต่ใช้ประโยชน์จากคลังข้อมูลเทียบบท (paralleled corpus) ในการหาค่าความสัมพันธ์ของคำศัพท์ ซึ่งในคลังข้อมูลคู่ภาษานี้ต้องมีการจับคู่หน่วยศัพท์ (lexeme alignment) โดยที่ต้องมีการกำกับหมวดคำ (POS marked-up) เพื่อที่จะบอกคุณสมบัติไว้แล้วว่าคู่คำใดเป็นคำแปลของคำใด โดยที่ต้องมีการวิเคราะห์หน่วยคำ (morphological analysis) ของแต่ละคำไว้ด้วย

การดึงคำศัพท์ต่างๆ มาใช้จากคลังข้อมูลเทียบบทต้องมีการทำดัชนีบ่งชี้ (index) เพื่อที่จะชี้บอกว่าหน่วยศัพท์ใดมีหมวดคำหรือชนิดของคำ (POS) ตรงกับหน่วยศัพท์ใด แล้วจึงทำการจัดผังคำ (word mapping) เพื่อหาความสัมพันธ์ภายในประโยค

เมื่อทำการจัดผังคำเสร็จแล้วจึงทำการส่งต่อข้อมูลกฎการเรียงตัวของผังคำนั้นๆ โดยหากพบว่ามีหน่วยศัพท์ใดที่ใช้ในรูปแบบซ้ำๆ กัน ก็จะสร้างกฎขึ้นมารองรับซึ่งกฎนี้จะถูกใช้เป็นกฎควบคุมส่วนซ้ำซ้อน (redundancy rules) เพื่อให้คอมพิวเตอร์จดจำกฎไปใช้ในกรณีที่พบรูปแบบการเรียงตัวหรือโครงสร้างประโยคที่เหมือนกันและช่วยลดปริมาณผังคำที่มีลักษณะเหมือนกัน

หลังจากนั้นจะนำผลที่ได้ซึ่งเป็นผังคำย่อยๆ มาพัฒนาต่อให้เป็นประโยค เนื่องจากการแปลภาษาด้วยเครื่องแบบใช้กฎส่งผ่านระดับคำจะใช้การวิเคราะห์แยกหน่วย (unit) ต่างๆ ออกมาเป็นหน่วยคำ (lexeme) จึงต้องมีการเรียบเรียงประโยคขึ้นใหม่ (rewrite) เพื่อประกอบเป็นประโยคที่ถูกต้อง ในส่วนการเรียบเรียงประโยคขึ้นใหม่นี้ หลักการทำงานคือการเลื่อนคำทีละตัวจากตัวแรกสุดทางด้านซ้ายไปเรื่อยๆ ให้ตรงกับผังคำโดยใช้การตรวจสอบความถูกต้องจากผังคำที่ได้จากคลังข้อมูล จนกว่าจะได้ประโยคที่ตรงกับโครงสร้างผังคำทั้งหมด

การแปลภาษาด้วยเครื่องแบบส่งผ่านระดับคำเป็นแนวทางที่มีค่าความถูกต้องสูง (Beaven, 1992) และสามารถนำมาใช้ได้ผลดีกับการแปลภาษาในตระกูลภาษาเดียวกัน และไม่ค่อยพบว่าเกิดปัญหาในการแปลจากระบบมากนัก แต่ปัญหาส่วนใหญ่มักเกิดมาจากคลังข้อมูลเทียบบทที่มีตัวอย่างประโยคหรือคำศัพท์ไม่ครอบคลุมพอ ทั้งในภาษาต้นฉบับและภาษาเป้าหมาย ซึ่งก็จะเป็นปัญหาสำคัญในการแปล

2.2.2 การแปลภาษาด้วยเครื่องแบบใช้สถิติ

ในช่วงปลายศตวรรษที่ 90 ได้มีการตื่นตัวในวงการการแปลภาษาด้วยเครื่องแบบใช้สถิติเป็นอย่างมาก เนื่องจากในโครงการวิจัยของบริษัทไอบีเอ็มได้พัฒนาการใช้วิธีการทางสถิติเพียงอย่างเดียวในการวิเคราะห์และการผลิตระบบแปลภาษาด้วยเครื่องโดยทดลองกับ คลังข้อมูลของบันทึกการประชุมรัฐสภาแคนาดา (Canadian Hansard) ซึ่งเป็นบันทึกการอภิปรายในสภาโดยจัดเก็บเป็นภาษาอังกฤษและฝรั่งเศส ผลการทดลองก่อให้เกิดการตื่นตัวในวงการระบบแปลภาษาด้วยเครื่องอย่างมาก เพราะสามารถแปลได้ดีกว่าที่นักวิจัยทั่วไปคาดไว้มาก (Brown et al., 1993)

แนวทางของการแปลภาษาด้วยเครื่องแบบใช้สถิติคือการนำวิธีการทางสถิติมาใช้ช่วยแปล โดยต้องมีคลังข้อมูลเทียบบทที่มีการจับคู่ประโยค (Alignment Paralleled Corpus) เพื่อเป็นฐานความรู้ให้เครื่องเรียนรู้ และใช้ค่าสถิติ เอ็นแกรม (N-gram) (ณัฐพล, 2549)โดยจะสามารถคำนวณค่าการเกิดขึ้นร่วมกันของคู่คำแปล ซึ่งจะช่วยเมื่อภาษาต้นฉบับสามารถมีการแปลได้หลายแบบ โดยสามารถเลือกใช้เอ็นแกรมได้ตั้งแต่ 2 คำ (bigram) 3 คำ (trigram) เป็นต้น โดยมีทฤษฎีหลัก (Al-Onizan et al., 1999) คือ การคำนวณค่าเอ็นแกรม ถ้าคู่ชุดคำชุดใดมีค่าสถิติเอ็นแกรมคือ ค่าความน่าจะเป็น (probability) สูง จะแสดงว่าคำชุดนั้นมักจะปรากฏขึ้นร่วมกันบ่อยครั้ง จากการใช้ค่าสถิติ เอ็นแกรมจะทำให้ได้คู่คำแปลตามค่าความน่าจะเป็น จากคลังข้อมูลเทียบบท และสามารถนำไปเทียบแปลข้อความได้

ข้อดีของวิธีการนี้คือ ไม่มีการใช้กฎไวยากรณ์ ทำให้ไม่เกิดปัญหาเชิงภาษาศาสตร์ อาทิเช่น ปัญหากฎไม่ครอบคลุม ปัญหาการเพิ่มกฎ ปัญหาการแจงส่วนวายกสัมพันธ์ในประโยค (Syntax parsing) และปัญหาการแปลสำนวน เป็นต้น

ปัญหาสำคัญที่พบในแนวทางการแปลภาษาด้วยเครื่องแบบใช้สถิติคือจำเป็นต้องใช้คลังข้อความคู่ประโยคเทียบบทที่มีการจับคู่ระหว่างประโยค (Alignment Paralleled Corpus) จำนวนมากในการที่จะสร้างตัวแบบสถิติ (Statistical-Model)

2.2.3 การแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง

แนวทางนี้เป็นการแปลภาษาด้วยเครื่องโดยใช้ตัวอย่างภาษาจากคลังข้อมูลเทียบบทจึงได้ชื่อว่าการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างหรือการแปลภาษาด้วยเครื่องแบบอิงคลังข้อมูล (Corpus-Based Machine Translation) เป็นแนวทางที่ไม่ใช้ความรู้ทางภาษาศาสตร์ในการกำหนดสร้างกฎ หรือ แบ่งแยกองค์ประกอบโครงสร้างไวยากรณ์ของภาษา แต่จะใช้สิ่งที่พบจากค่าทางสถิติต่างๆ เป็นตัวแปรสำคัญในการหาส่วนสำคัญในภาษา ซึ่งการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างต้องใช้คลังข้อมูลเทียบบทเป็นฐานข้อมูลสำคัญเพื่อที่จะเป็นตัวอย่างจากภาษาที่ใช้กันจริงในปัจจุบัน โดยคลังข้อมูลที่ใช้ ต้องมีปริมาณประโยคและคำเป็นจำนวนมาก เพื่อที่จะนำไปคำนวณหาค่าทางสถิติว่า ประโยคที่แปลควรจะแปลออกมาเป็นประโยคในรูปแบบใด ขนาดของคลังข้อมูลเทียบบทเป็นสิ่งสำคัญที่จะทำให้งานการแปลภาษาด้วยเครื่องจะเกิดประสิทธิภาพและเกิดความถูกต้องแม่นยำ แม็คเทท (McTait, 2001) ได้กล่าวไว้ว่า “เมื่อเพิ่มขนาดคลังข้อมูลเทียบบทจำนวนคำที่จะพบในสองประโยคขึ้นไปก็เพิ่มขึ้นด้วย รวมทั้งจำนวนของการจับคู่ความสัมพันธ์ของคำที่ปรากฏด้วยกัน แต่อย่างไรก็ตามการจับคู่ความสัมพันธ์ของคำที่ปรากฏด้วยกันสามารถเพิ่มจนถึงจุดหนึ่ง ซึ่งถึงแม้ว่าขนาดของคลังข้อมูลเทียบบทจะเพิ่มมากขึ้น แต่จำนวนของการจับคู่

ความสัมพันธ์ของคำที่ปรากฏด้วยกันนี้ก็จะไม่เพิ่มตาม ซึ่งถือเป็นข้อดีของ แนวทางการแปลภาษาด้วยเครื่องนี้ คือสร้างคลังข้อมูลเทียบบทให้พอถึงขนาดหนึ่งก็เพียงพอที่จะนำมาใช้งานได้อย่างมีประสิทธิภาพ" เมื่อได้คลังข้อมูลเทียบบทที่ทำการจับคู่ประโยคและคำในคลังข้อมูลเทียบบทแล้วจึงวางโปรแกรมให้ทำการสกัดข้อมูล (extract) ที่ได้จับไว้เป็นคู่ๆเพื่อสร้างเป็นแม่แบบการแปล (template) ซึ่งจะใช้เป็นต้นแบบพื้นฐานในการเปรียบเทียบกับตัวภาษาที่จะทำการแปล เมื่อได้ทำการเปรียบเทียบแปลแล้วจึงทำการรวมคำแปลใหม่ (recombine) ออกมาเป็นภาษาที่แปลเสร็จสมบูรณ์ การทำงานของการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างมี อยู่ 2 ส่วนหลักคือ การสกัดแม่แบบ (template extraction) จากคลังข้อมูลเทียบบทเพื่อที่จะเป็นแม่แบบในการเปรียบเทียบแปลและการรวมคำแปลใหม่ (recombination) ที่ได้จากการเปรียบเทียบแปลให้เป็นประโยค

ข้อดีของEBMT

- ไม่ใช้กฎโครงสร้างทางวากยสัมพันธ์และพจนานุกรมสองภาษาเป็นหลักในการวิเคราะห์ภาษา ดังนั้นจึงไม่เกิดความกำกวมทางโครงสร้างวากยสัมพันธ์และอรรถศาสตร์
- ประหยัดเวลาในการแก้ไขปรับปรุง เนื่องจากใช้การวางระบบในการดึงข้อมูลมาใช้ในการแปลแค่ครั้งเดียว แล้วคอมพิวเตอร์จะเรียนรู้จากคลังข้อมูลเทียบบทด้วยตนเอง
- เมื่อทำการวางระบบออกมาได้เสร็จสมบูรณ์ ก็สามารถนำมาประยุกต์ใช้กับคลังข้อมูลเทียบบทอื่นๆได้

ข้อเสียของEBMT

- ต้องใช้ทรัพยากรสร้างคลังข้อมูลเทียบบทที่มีปริมาณคำและประโยคตัวอย่างเป็นจำนวนมากเพื่อที่จะให้เครื่องคอมพิวเตอร์เรียนรู้
- ต้องมีการทำการจับคู่ประโยคและคำในคลังข้อมูลเทียบบท ด้วยตนเองให้มีความถูกต้องแม่นยำ เพราะจะส่งผลต่อความแม่นยำในการแปลโดยตรง

จุดมุ่งหมายของการพัฒนาการแปลภาษาด้วยเครื่องไม่ว่าแนวทางใดก็ตาม คือสามารถนำมาใช้ในการแปลภาษาธรรมชาติได้อย่างมีประสิทธิภาพ อย่างไรก็ตามผลการแปลภาษาด้วยเครื่องในปัจจุบันยังนับได้ว่ามีค่าความถูกต้องน้อยในการแปลภาษาธรรมชาติทั่วไปในการใช้สื่อสารในชีวิตประจำวัน แต่หากนำระบบการแปลภาษาด้วยเครื่องไปใช้กับงานที่มีลักษณะเฉพาะแล้ว ผลการแปลที่ได้จะมีความถูกต้องมากกว่า เช่นงานการแปลภาษาเฉพาะทาง (specialised

language) ต่างๆ ดังนั้นการวิจัยและพัฒนาระบบการแปลภาษาด้วยเครื่องส่วนมากจึงนิยมนำมาใช้ทดลองแปลกับภาษาเฉพาะทาง เพราะสามารถหาจุดดี จุดด้อย และข้อผิดพลาดได้ง่ายกว่าการวิจัยและพัฒนาการแปลภาษาด้วยเครื่องกับการแปลภาษาธรรมชาติที่กว้างและซับซ้อนมาก

2.3 ภาษาเฉพาะทาง

ภาษาเฉพาะทางหมายถึงภาษาที่มีลักษณะการใช้ภาษาที่มีจุดมุ่งหมายเฉพาะ มีลักษณะเฉพาะของภาษาที่ใช้ และเป็นภาษาที่ใช้ในคนเฉพาะกลุ่ม เช่น รายงานพยากรณ์อากาศ รายงานตลาดหุ้น การสนทนาทางด้านเวชภัณฑ์เฉพาะด้าน ภาษาของนักวิศวกรรมการบิน เป็นต้น อาร์โนลด์และแซดเลอร์ (Arnold and Sadler, 1989) ได้ให้นิยามของภาษาเฉพาะทางไว้ว่า เป็นภาษาที่แม้มีการใช้คำศัพท์ที่ใช้กันทั่วไปแต่จะมีความหมายเฉพาะที่เป็นที่เข้าใจกันภายในกลุ่มผู้เชี่ยวชาญและผู้รู้ภาษาเฉพาะทาง ซึ่งถือว่าเป็นเอกลักษณ์อย่างหนึ่งของภาษาเฉพาะทาง นอกจากนี้ภาษาเฉพาะทางยังอาจจะมีรูปแบบการเรียงตัวของลักษณะทางไวยากรณ์ ที่เป็นเอกลักษณ์และแตกต่างภาษาธรรมชาติทั่วไปอีกด้วย

อย่างไรก็ตาม อาร์โนลด์และแซดเลอร์ (Arnold and Sadler, 1989) ยังได้กล่าวไว้อีกว่าในงานการแปลภาษาด้วยเครื่องนั้น มีการอ้างอิงถึงภาษาเฉพาะทางเป็นจำนวนมาก แต่ภาษาเฉพาะทางเหล่านั้นบางครั้งจะไม่ใช่ภาษาเฉพาะทางที่มีลักษณะตรงกันกับภาษาเฉพาะทางตามหลักที่กล่าวมามากนัก หากแต่เป็นภาษาที่มีการใช้ในเอกสารเฉพาะทางต่างๆ หรือบทสนทนาเฉพาะทาง เช่น คู่มือการใช้ รายงานวินิจฉัยโรค การให้คำแนะนำจากผู้เชี่ยวชาญแก่บุคคลทั่วไป ในที่นี้จึงใช้ภาษาเฉพาะทางในความหมายที่กว้างคือ เป็นภาษาที่พบในกลุ่มข้อมูลประเภทใดประเภทหนึ่งโดยเฉพาะ

จากลักษณะภาษาของภาษาเฉพาะทาง ทำให้เป็นข้อมูลที่เหมาะสมกับการแปลภาษาด้วยเครื่อง เนื่องจากงานการแปลภาษาด้วยเครื่องมักจะถูกออกแบบมาใช้งานกับข้อมูลที่วงศัพท์จำกัด กำหนดความหมายต่างๆได้ชัดเจน และมีรูปแบบทางวากยสัมพันธ์ที่เจาะจง ดังนั้นการแปลภาษาเฉพาะทางจึงเป็นงานที่เหมาะสมแก่การแปลภาษาด้วยเครื่อง

ถึงแม้ว่าระบบการแปลภาษาด้วยเครื่องนั้นโดยทั่วไปมักจะถูกนำมาใช้เพื่อการแปลภาษาเฉพาะทางเป็นหลัก แต่ก็ไม่สามารถที่จะนำระบบการแปลภาษาด้วยเครื่องที่ออกแบบเพื่อแปลภาษาเฉพาะทางหนึ่งไปใช้กับการแปลภาษาเฉพาะทางอื่นๆได้ เนื่องจากลักษณะทางไวยากรณ์และวงศัพท์ของภาษาเฉพาะทางมีลักษณะเป็นเอกลักษณ์และแตกต่างกันไป ดังนั้นจึงอาจกล่าวได้ว่าการพัฒนาระบบการแปลภาษาด้วยเครื่องโดยทั่วไปจะถูกออกแบบมาโดยเฉพาะสำหรับงานแต่ละงาน แต่การแปลภาษาด้วยเครื่องแบบอิงตัวอย่างมีลักษณะเด่นอยู่ที่การใช้ตัวอย่างจากคลังข้อมูลเทียบบทเป็นต้นแบบในการเปรียบเทียบหาคำแปล กล่าวคือการแปลภาษาด้วยเครื่องแบบอิง

ตัวอย่างสามารถพัฒนาระบบเพียงครั้งเดียว ก็จะสามารถนำไปเรียนรู้และแปลภาษาเฉพาะทางได้หลากหลาย หากมีคลังข้อมูลเทียบบทของภาษาเฉพาะทางนั้น ทำให้ช่วยผ่อนแรงและสะดวกต่อการแปลภาษาเฉพาะทางที่หลากหลายได้

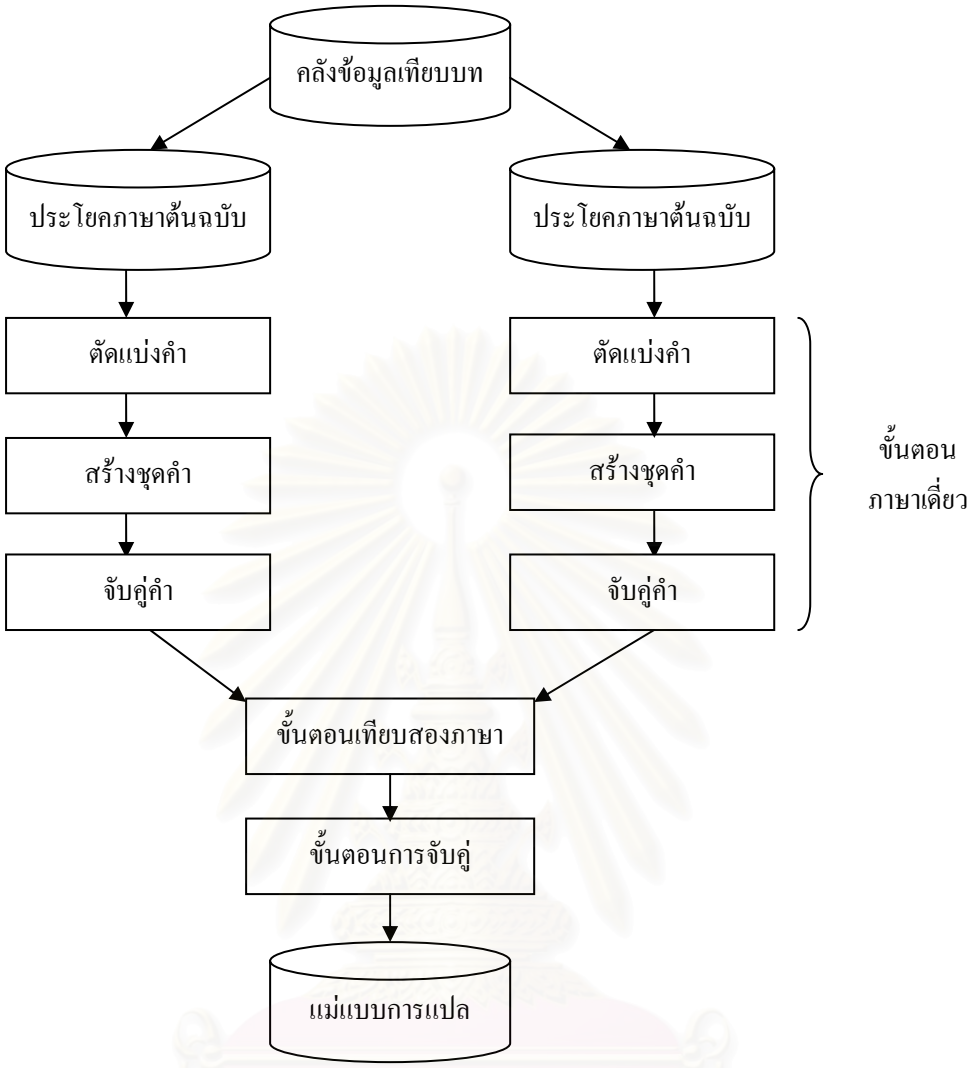
2.4 ระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง

เพื่อที่จะให้เข้าใจการทำงานของระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างซึ่งพัฒนาขึ้นในงานวิจัยนี้ ผู้วิจัยจะกล่าวได้ถึงระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง โดยจะยึดงานของแม็คเทท (McTait, 2001) เป็นหลักในการอธิบาย แม็คเทท (McTait, 2001) ได้เสนอแนวทางการสกัดแม่แบบและการรวมคำแปลใหม่สำหรับการแปลภาษาด้วยเครื่องโดยใช้คลังข้อมูลเทียบบทเป็นตัวอย่างในการแปลไว้ ซึ่งเป็นวิธีการที่น่าสนใจ และนำไปใช้ได้จริง งานการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างของแม็คเทท ได้ใช้แปลภาษาโดยอาศัยคลังข้อมูลเทียบบทเป็นตัวอย่างในการแปล จากภาษาอังกฤษเป็นภาษาฝรั่งเศส ซึ่งเป็นภาษาในตระกูลภาษาเดียวกัน และมีลักษณะรูปแบบการใช้เครื่องหมายต่างๆ[4] เหมือนกันและมีการเว้นวรรคระหว่างคำ

ในงานการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างของแม็คเททมีสมมติฐานหลักอยู่ว่าข้อมูลในคลังข้อมูลเทียบบทต้องมีการปรากฏของโครงสร้างการเรียงตัวของคำศัพท์และมีคำศัพท์ซ้ำๆ กัน และประโยคภายในคลังข้อมูลต้องมีส่วนซ้ำและส่วนที่ไม่ซ้ำกันภายในประโยคนั้น จึงจะสามารถสกัดส่วนคงที่ (invariant) ของแม่แบบการแปลและ ตัวแปร (variable) ของแม่แบบการแปลได้

องค์ประกอบที่สำคัญในการทำงานของงานการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างของแม็คเททได้แก่ ขั้นตอนการสกัดแม่แบบการแปลออกคลังข้อมูลเทียบบทซึ่งแยกย่อยออกมาเป็น 3 ส่วน คือขั้นตอนภาษาเดี่ยว (monolingual phase) ขั้นตอนสองภาษา (bilingual phase) และขั้นตอนการจับคู่ (alignment phase) ดังแสดงไว้ในรูปที่ 3 ซึ่งหลังจาก 3 ขั้นตอนนี้แล้วจะได้แม่แบบการแปล

[4] เครื่องหมายคำถาม เครื่องหมายตกใจ เป็นต้น



รูปที่ 3 แสดงระบบการทำงานของการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างของแม็คเทท

หลังจากที่ได้แม่แบบการแปลจึงเป็นขั้นตอนการรวมคำแปลใหม่จากการเทียบแปลแม่แบบการแปล

2.4.1 การสกัดแม่แบบการแปล

การสกัดแม่แบบการแปลคือการค้นหาส่วนซ้ำและส่วนที่ไม่ซ้ำกันภายในคู่ประโยคในคลังข้อมูลเพื่อสร้างแม่แบบการแปล โดยการจัดกลุ่มคู่ประโยคและระบุส่วนคงที่ (invariant) และส่วนที่แปรผัน (variable) ภายในคู่ประโยคเหล่านั้น การสกัดแม่แบบการแปลต้องใช้คลังข้อมูลเทียบบทที่มีการจับคู่ และต้องตรวจสอบการจับคู่ให้ถูกต้อง แต่ไม่ต้องมีการจับคู่คำและการกำกับข้อมูลคำ[5] ดังนั้นการจัดเตรียมคลังข้อมูลเทียบบทจึงต้องมีการตรวจสอบด้วยตนเองอย่าง

[5] พจน์ (number) เพศ (gender) บุรุษ (person) เป็นต้น

ละเอียดถี่ถ้วน ซึ่งการจับคู่ประโยคในคลังข้อมูลเทียบบทต้องเป็นลักษณะ 1 ประโยคต่อ 1 ประโยคเท่านั้น ปริมาณความถี่ในการเกิดประโยคซ้ำๆ กันในคลังข้อมูลเทียบบท จะส่งผลต่อความแม่นยำในการสกัดแม่แบบการแปล ดังนั้นจึงต้องสร้างคลังข้อมูลเทียบบทที่มีปริมาณคำและประโยคตัวอย่างเป็นจำนวนมาก

การสกัดแม่แบบการแปลมี 3 ขั้นตอนคือ ขั้นตอนแรกเป็นขั้นตอนภาษาเดี่ยว (monolingual phase) ซึ่งจะแบ่งย่อยออกเป็นอีก 3 ส่วนคือ ส่วนการตัดแบ่งคำ (tokenisation) ส่วนการสร้างชุดคำ (word list construction) และส่วนข้อมูลการปรากฏร่วมจำเพาะ (collocation formation) หลังจากนั้นจึงมาเป็นส่วนขั้นตอนที่สองคือ ขั้นตอนเทียบสองภาษา (bilingual phase) และขั้นตอนสุดท้ายคือ ขั้นตอนการจับคู่ (alignment phase)

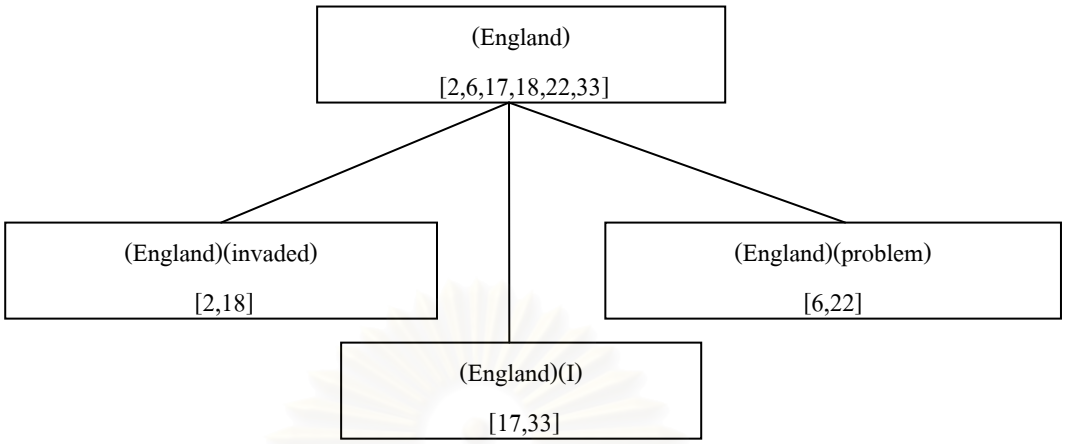
2.4.1.1 ขั้นตอนภาษาเดี่ยว

สิ่งสำคัญสิ่งแรกที่จะต้องทำในขั้นตอนนี้คือการตัดแบ่งคำในคลังข้อมูลเทียบบท อย่างไรก็ตามภาษาอังกฤษจะไม่มีปัญหาเรื่องการตัดคำเนื่องจากมีการเว้นวรรคและเครื่องหมายต่างๆที่แบ่งคำออกจากกันอย่าง อย่างไรก็ตามในภาษาอื่นเช่นภาษาไทยอาจจำเป็นต้องตัดแบ่งคำในขั้นตอนนี้หลังจากการตัดคำ จะตรวจสอบว่าคำใดที่มีการปรากฏ 2 ครั้งขึ้นไปก็จะให้ระบบเก็บข้อมูลไว้เพื่อสร้างชุดคำโดยจะเก็บดัชนีบ่งชี้ของประโยคที่พบคำนั้นด้วย

แต่ในกรณีที่เป็นกลุ่มคำที่ต่อเนื่องหรือเป็นกลุ่มคำที่มีคำมาแทรกกลางก็จะใช้โครงสร้างต้นไม้ (tree) ที่มีลักษณะเป็น โหนดราก (Root node) และ โหนดลูก (Daughter node) เพื่อที่จะแสดงเชื่อมโยงข้อมูลกลุ่มคำที่มีการปรากฏซ้ำๆ กันหลายครั้ง และจะนำมาพัฒนาต่อไปให้เป็นต้นไม้ของการปรากฏร่วม (collocation tree) ที่แตกตัวออกเพื่อที่จะให้เครื่องทำการเรียนรู้และจดจำรูปแบบของการปรากฏร่วมกันของคำทั้งหมดได้ ดังเช่นประโยคตัวอย่างด้านล่าง

ประโยคที่	ข้อความ
2	England invaded France in 1842
6	England encountered flood problem
17	I go to England
18	England invaded Argentina
22	England has traffic problem
33	I visit England

ก็จะได้ผลดังรูปที่ 4



รูปที่ 4 แสดงต้นไม้ของการปรากฏร่วมส่วนหนึ่งที่มีโหนดรากเป็นคำว่า “England”

ผลลัพธ์ของขั้นตอนภาษาเดี่ยวนี้จะได้ต้นไม้ของการปรากฏร่วมที่มีการเรียงกลุ่มคำศัพท์เข้าไว้ด้วยกันเพื่อจะทำให้ทราบว่าคำใดมีการปรากฏร่วมกับคำใดได้บ้าง และเกิดการเรียงตัวกันของคำที่เป็นไปได้มากที่สุดเท่าไหร่ สำหรับให้เครื่องเรียนรู้และจดจำรูปแบบของต้นไม้ของการปรากฏร่วม

2.4.1.2 ขั้นตอนเทียบสองภาษา

ขั้นตอนเทียบสองภาษาคือการนำสิ่งที่ได้จากขั้นตอนภาษาเดี่ยวมาตรวจสอบต้นไม้ของการปรากฏร่วมของทั้งภาษาต้นฉบับกับภาษาเป้าหมายว่าเป็นคำแปลของอีกภาษาหรือไม่ โดยขั้นตอนการตรวจจะดูจากค่าความถี่ของคำหรือกลุ่มคำที่อยู่ปลายกิ่งของต้นไม้การปรากฏร่วม กล่าวคือถ้ามีการปรากฏซ้ำๆ กันในปริมาณที่มาก ความน่าจะเป็นที่จะเป็นคำแปลของกันและกันก็จะสูงขึ้นด้วย และจะทำการเก็บข้อมูลส่วนนั้นโดยจะเก็บข้อมูลที่พบซ้ำกันมากกว่า 2 ครั้ง โดยหากมีคำใดคำหนึ่งหรือมากกว่าที่มีการปรากฏซ้ำ และในประโยคที่จับคู่เทียบบทกันก็มีการปรากฏซ้ำ เครื่องก็จะเก็บข้อมูลและรู้ว่าคำคู่นั้นเป็นคำแปลของกันและกัน เช่นหากพบว่า ในทุกประโยคที่มีคำว่า ‘gave up’ พบว่าในคู่ประโยคในคลังข้อมูลเทียบบทมีคำว่า ‘abandonna’ เครื่องก็จะเก็บข้อมูลว่า ‘abandonna’ เป็นคำแปลของ ‘gave up’ และสกัดส่วนนั้นเป็นส่วนคงที่ (invariant) ของแม่แบบการแปล เช่น

ตย.ประโยค(1) The commission gave the plan up ↔ La commission abandonna le plan

ตย.ประโยค(2) Our government gave all laws up ↔ Notre gouvernement abandonna toutes les lois

จากคู่ประโยค 2 คู่นี้ที่ปรากฏว่าทั้ง 2 คู่ประโยคนี้มีคำที่ปรากฏซ้ำกันคือ ‘gave’ ‘up’ ในภาษาต้นฉบับและ ‘abandonna’ ในภาษาเป้าหมาย ดังนั้นเครื่องก็จะเก็บข้อมูลและสกัด ‘gave’ ‘up’ ว่าเป็นคู่คำแปลกับ ‘abandonna’

และเพื่อป้องกันความผิดพลาดในการจับคู่ในกรณีที่คำที่ปรากฏไม่เกิดขึ้นติดกัน จึงต้องใช้การเกิดขึ้นร่วมกันที่ยาวที่สุดซึ่งเป็นกลุ่มคำที่ปลายกิ่ง เช่น

ถูก	ผิด
(gave)(up) ↔ (abandonna)	(gave) ↔ (abandonna)

หลังจากสกัดแม่แบบการแปลแล้ว จากนั้นจะตรวจสอบส่วนแปรผัน (variable) ที่เหลือจากส่วนคงที่ (invariant) ที่สกัดไปแล้ว ซึ่งโดยหลักแล้วส่วนแปรผันที่เหลือก็ควรจะเป็นคำแปลของกันและกัน

The commission <X1> the plan <X2> ↔ La commission <Y1> le plan

Our government <X1> all laws <X2> ↔ Notre gouvernement <Y1> toutes les lois

2.4.1.3 ขั้นตอนการจับคู่

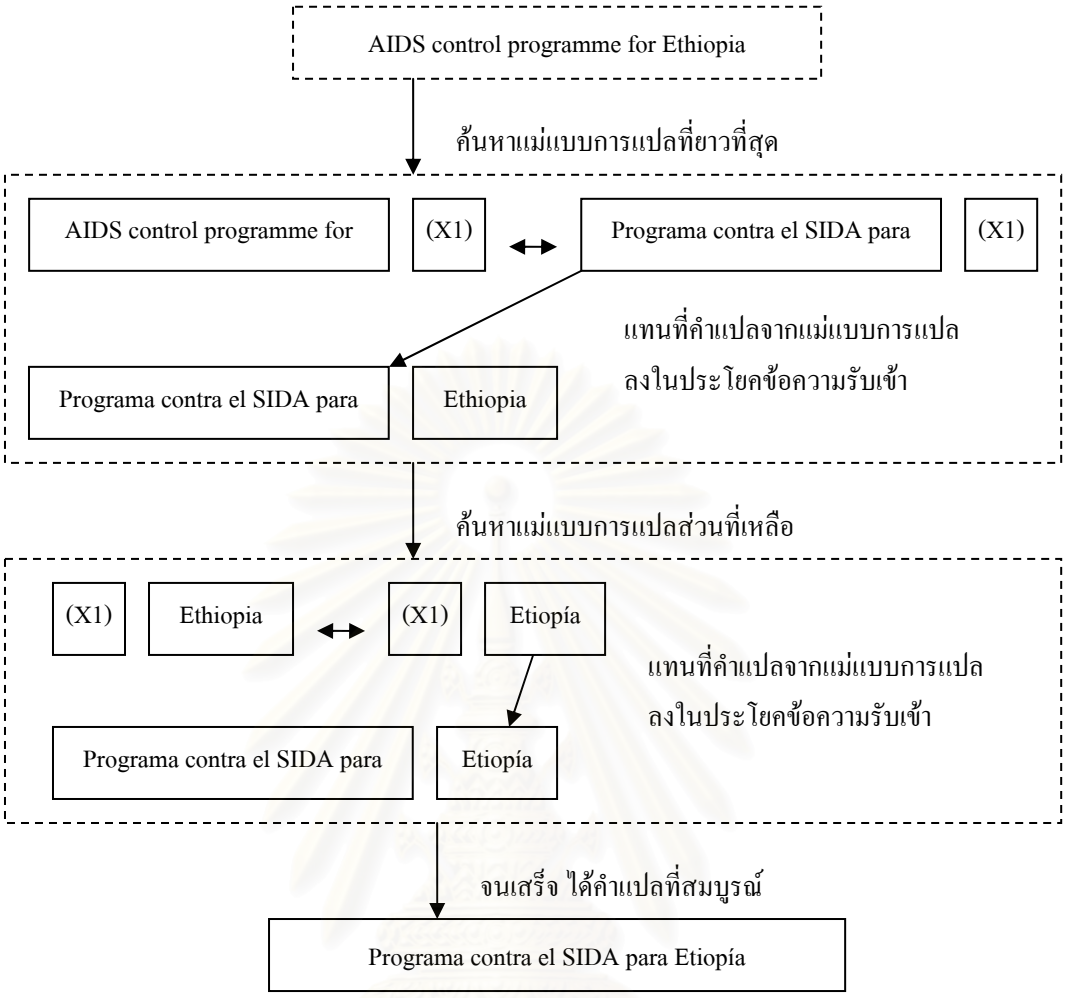
ขั้นตอนการจับคู่คือการตรวจสอบตัวแปรที่เหลือจากการสกัดส่วนคงที่ว่าเป็นคำแปลของกันและกันในตำแหน่งที่ตรงกันหรือไม่ ดังเช่นในกรณี ‘The commission’ กับ ‘La commission’ และ ‘the plan’ กับ ‘le plan’ ทั้ง 2 คู่คำนี้ไม่เป็นปัญหาในการจับคู่ (alignment) เพราะเครื่องจะนำตัวแปรที่เหลือจากการสกัดส่วนคงที่มาทำการจับคู่โดยอัตโนมัติซึ่งจะใช้การพิจารณาจากรากศัพท์เนื่องจากภาษาที่ได้นำมาวิจัยนั้นเป็นภาษาร่วมเชื้อสาย (cognate) หากเป็นคำร่วมเชื้อสายก็มีความเป็นไปได้ว่าเป็นคำแปลของกันและกัน ซึ่งส่วนภาษาร่วมเชื้อสายนี้จะไม่คิดคำนวณความถี่และสามารถนำไปเก็บเป็นแม่แบบการแปลในการเปรียบเทียบแปลต่อไป แต่ถ้าตำแหน่งที่เหลือจากการสกัดไม่ตรงกันและไม่สามารถจับคู่ได้ก็จะไม่นำข้อมูลส่วนนั้นไปเก็บเป็นแม่แบบการแปล ซึ่งตามปกติแล้วหากมีข้อมูลคู่ประโยคในคลังคู่ภาษาเทียบบทมากและครอบคลุมข้อมูลคู่ประโยคก็จะถูกนำไปสร้างต้นไม้ของการปรากฏร่วมทั้งหมดและข้อมูลส่วนที่จะเหลือมาถึงส่วนนี้ก็จะเป็นส่วนที่เป็นชื่อเฉพาะ (proper name) หรือ นิพจน์ระบุนาม (name entity) เท่านั้น

2.4.2 การรวมคำแปลใหม่

การแปลในระบบนี้คือการนำการแปลส่วนต่างๆ มาเรียงเป็นประโยคขึ้นมา ในการทำงานรวมคำแปลใหม่คือส่วนหลังจากการใส่ข้อความรับเข้า (input) เข้าไปแล้วให้เครื่องนำข้อความรับเข้านั้นไปค้นหาแม่แบบการแปลที่ตรงกับข้อความรับเข้ามาทำการเปรียบเทียบหาคำแปล ซึ่งจะดึงส่วนที่ยาวที่สุดที่ตรงกันมาแล้วนำส่วนที่เหลือมาทำการดึงคำออกมาค้นหาคำที่ตรงกันมาประกอบต่อกัน

เช่น ถ้าประโยคข้อความรับเข้าเป็น ‘AIDS control programme for Ethiopia’ และมีแม่แบบการแปลที่เครื่องได้เรียนรู้เก็บไว้แล้ว คือ ‘AIDS control programme for <X1>’ ↔ ‘Programa contra el SIDA para <Y1>’ ที่สกัดได้จากคู่ประโยคในคลังข้อมูลเทียบบท ‘AIDS control programme for England’ ↔ ‘Programa contra el SIDA para Inglaterra’ และ ‘AIDS control programme for China’ ↔ ‘Programa contra el SIDA para China’ และแม่แบบการแปล ‘<X1> Ethiopia’ ↔ ‘<Y1> Etiopía’ ที่สกัดได้จากคู่ประโยคในคลังข้อมูลเทียบบท ‘I love Ethiopia’ ↔ ‘Amo Etiopía’ และ ‘you went to Ethiopia’ ↔ ‘usted fue a Etiopía’

โดยขั้นตอนการดึงแม่แบบการแปล จะทำการค้นหาแม่แบบการแปลที่มีลักษณะเดียวกันกับประโยคข้อความรับเข้า และพบแม่แบบการแปล ‘AIDS control programme for <X1>’ ↔ ‘Programa contra el SIDA para <Y1>’ จึงทำการเปรียบเทียบแปล แล้วจึงไปค้นหา ‘<X1> Ethiopia’ ต่อ และพบ ‘<X1> Ethiopia’ ↔ ‘<Y1> Etiopía’ จึงได้ประโยคแปลที่สมบูรณ์ คือ ‘Programa contra el SIDA para Etiopía’ รูปที่ 5 ด้านล่างแสดงตัวอย่างกระบวนการแปล



รูปที่ 5 แสดงขั้นตอนการรวมคำแปลใหม่

ดังนั้น ประโยคข้อความรับเข้า ‘AIDS control programme for Ethiopia’ จะถูกแปลเป็น ‘Programa contra el SIDA para Etiopía’

แม่แบบการแปลที่ถูกสกัดไว้ก่อนแล้วจะทำให้สามารถดึงแม่แบบการแปลที่ได้ทำการวิเคราะห์โครงสร้างต้นไม้ไว้แล้ว และเมื่อมีการเปรียบเทียบแปลก็สามารถดึงออกมาใช้ได้ทันที ซึ่งวิธีนี้จะได้ผลดีและแม่นยำที่สุดในกรณีที่มีการจับคู่แบบ 1 ส่วน ต่อ 1 ส่วน ดังเช่นในตัวอย่างด้านบน ที่สำคัญในการนำแม่แบบการแปลมาใช้เทียบหาคำแปลต้องมีจำนวนคำเท่ากันหรือน้อยกว่าจำนวนคำของประโยคข้อความรับเข้าที่ใส่เข้าไปเท่านั้น และแม่แบบการแปลจำเป็นต้องมีคำศัพท์ตรงกับในประโยคข้อความรับเข้าทุกคำ เช่น ประโยคข้อความรับเข้าคือ ‘A B C D E’ ดังนั้นแม่แบบการแปลที่1คือ ‘A B <X1> E’ จึงสามารถนำไปใช้แปลประโยคข้อความรับเข้านี้ได้ แต่แม่แบบการแปลที่2 คือ ‘A B <X1> E F G H’ ไม่สามารถนำไปใช้แปลประโยคข้อความรับเข้านี้ได้ เพราะแม่แบบการแปลที่2 มีจำนวนคำมากกว่าประโยคข้อความรับเข้าคือ ‘F G H’ และแม่แบบการ

แปลที่3 คือ ‘A B <X1> B’ ก็ไม่สามารถนำไปใช้แปลประโยคข้อความรับเข้านี้ได้เพราะแม่แบบการแปลที่3 มีคำศัพท์ไม่ตรงกับในประโยคข้อความรับเข้าทุกคำคือ มี ‘B’ ซ้ำอยู่และไม่ปรากฏในประโยคข้อความรับเข้า ดังนั้นแม่แบบการแปลที่1 จึงเป็นแม่แบบการแปลเดียวที่สามารถใช้แปลประโยคข้อความรับเข้าคือ ‘A B C D E’ ได้

ในกรณีที่พบว่ามีแม่แบบการแปลที่สามารถนำมาใช้แปลได้มากกว่า 1 แม่แบบ เครื่องก็จะเลือกแม่แบบการแปลที่ยาวที่สุดมาใช้ในการแปลเป็นอันดับแรก แต่หากความยาวของแม่แบบการแปลนั้นเท่ากัน เครื่องก็จะเลือกแม่แบบการแปลที่มีความถี่ในการปรากฏในคลังข้อมูลเทียบบทมากกว่ามาใช้ทำการเปรียบเทียบหาตำแปล

อย่างไรก็ตาม ในกรณีที่เครื่องพบประโยคข้อความรับเข้าที่ไม่พบจากแม่แบบการแปล หรือไม่สามารถหาแม่แบบการแปลที่ตรงกันได้ เครื่องก็จะใช้เกณฑ์ในการแปลโดยดึงข้อมูลที่ได้จากขั้นตอนเทียบสองภาษาและขั้นตอนการจับคู่ออกมาทีละคู่และเรียงคำแปลตามข้อความจากภาษาต้นฉบับ ซึ่งโดยส่วนมากจะแปลออกมาไม่ถูกต้องและบางครั้งก็ไม่สามารถแปลได้ และในบางกรณีที่บางส่วนของประโยคข้อความรับเข้าไม่สามารถหาแม่แบบการแปลใดๆ มาเทียบแปลได้ ระบบก็จะละส่วนนั้นโดยแทนที่ด้วยเครื่องหมายปรัศนี (question mark)

แนวทางของแม็คเททนี้ เป็นแนวทางใหม่ในวงการการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างที่น่าสนใจและให้ความสำคัญกับคลังข้อมูลมาก โดยมองว่าคลังข้อมูลคือแหล่งความรู้ที่จะให้เครื่องเรียนรู้และสกัดหาแม่แบบการแปล คล้ายกับการเรียนรู้ภาษาของมนุษย์ที่เรียนรู้คำศัพท์จากการใช้ภาษาจริงในชีวิตประจำวัน ดังนั้นผู้วิจัยจึงพยายามที่จะนำแนวทางนี้มาประยุกต์ใช้กับงานวิจัยการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง : กรณีศึกษาการแปลรายงานข่าวตลาดหุ้นจากภาษาไทยเป็นภาษาอังกฤษ ซึ่งจะกล่าวในบทต่อไป

บทที่ 3

ขั้นตอนการทำงานของระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง

ภายในบทนี้จะอธิบายถึงการเตรียมข้อมูลและขั้นตอนการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างซึ่งจะแบ่งเป็น 2 ส่วนหลักคือส่วนการเตรียมคลังข้อมูลเทียบบทเพื่อใช้ในการฝึกระบบ และส่วนระบบการทำงานของการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างจากภาษาไทยเป็นภาษาอังกฤษ

3.1 คลังข้อมูลที่ใช้ในงานวิจัย

คลังข้อมูลภาษานับเป็นทรัพยากรที่สำคัญสำหรับการวิเคราะห์ภาษาและการศึกษาทางภาษาศาสตร์และการประมวลผลภาษาธรรมชาติ และถูกนำมาใช้อย่างแพร่หลายเพื่อเป็นฐานความรู้ด้านต่างๆ ให้กับระบบ โดยเฉพาะอย่างยิ่ง การประมวลผลภาษาธรรมชาติในปัจจุบันมักประยุกต์วิธีการทางสถิติเข้ามาใช้ คลังข้อมูลจึงกลายเป็นแหล่งข้อมูลทางภาษาขนาดใหญ่สำหรับสำรวจและคำนวณข้อมูลทางสถิติ เพื่อช่วยในการประมวลผลให้มีความถูกต้องแม่นยำและมีประสิทธิภาพสูงขึ้น หรือแม้แต่งานที่อาศัยกฎในปัจจุบันซึ่งได้รับการพัฒนาแก้ไขขึ้นมาใหม่ เช่น เทคนิคการกำกับหมวดคำโดยอัตโนมัติของบริล (Brill, 1992) ก็ยังให้ระบบเรียนรู้และสรุปกฎจากคลังข้อมูลเช่นกัน

คลังข้อมูลที่นำมาใช้ในการประมวลผลภาษาธรรมชาติได้รับการพัฒนารูปแบบและวิธีการเรื่อยมา ทำให้มีลักษณะแตกต่างกันไปทั้งในด้านรูปแบบและเนื้อหา มีทั้งที่เป็นคลังข้อมูลล้วน (plain corpus) ซึ่งประกอบด้วยข้อความอย่างเดียว และคลังข้อมูลที่มีการกำกับข้อมูล (annotated corpus) ประเภท คลังข้อมูลที่กำกับความหมาย และคลังข้อมูลที่กำกับหมวดคำ เช่น ‘ORCHID Corpus’ ของศูนย์คอมพิวเตอร์และอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) (Sornlertlamvanich, 1997) หรือบางงานต้องการข้อมูลที่เป็นตัวแทนที่ครอบคลุมทั้งภาษา (language coverage) ในขณะที่บางงานต้องการข้อมูลภาษาเฉพาะทาง เฉพาะเรื่องเท่านั้น ทั้งนี้ขึ้นอยู่กับจุดประสงค์ วิธีการ และลักษณะของงานที่จะนำคลังข้อมูลไปใช้

ส่วนนี้จะกล่าวถึงการจัดทำคลังข้อมูลเพื่อใช้เป็นฐานความรู้สำหรับการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง เนื่องจากเครื่องจำเป็นต้องใช้คลังข้อมูลคู่ภาษาเพื่อใช้อิงตัวอย่างในการเทียบแปล โดยแบ่งเป็นคลังข้อมูลภาษาต้นฉบับ (source language/SL) และคลังข้อมูลภาษาเป้าหมาย (target language/TL) ดังนั้นคลังข้อมูลที่ใช้ในวิทยานิพนธ์ฉบับนี้ จึงต้องใช้คลังข้อมูลคู่ภาษาที่มีการตัดแบ่งข้อมูลภายในคลังข้อมูลเป็นข้อความ (clause) เพราะภาษาต้นฉบับคือภาษาไทยนั้นเขียน

แบบต่อเนื่องโดยไม่มีการแบ่งประโยคหรือข้อความชัดเจน บางครั้ง 1 ประโยค (sentence) จึงอาจสามารถยาวได้ถึง 1 ย่อหน้า (paragraph) แต่คู่ประโยคในภาษาอังกฤษกลับสั้นและแบ่งออกเป็นประโยคและข้อความได้อย่างชัดเจน เนื่องจากผู้วิจัยตัดแบ่งส่วนของเนื้อความออกเป็น 1 ส่วนต่อ 1 บรรทัด ผู้วิจัยจึงเลือกตัดตามข้อความเพื่อเนื้อความภายใน 1 บรรทัดจะได้ไม่ยาวเกินไป โดยจัดแบ่ง 1 ข้อความต่อ 1 บรรทัดและจับคู่ข้อความในคลังข้อมูลที่เป็นคำแปลของกันและกันในคลังข้อมูลภาษาต้นฉบับกับคลังข้อมูลภาษาเป้าหมายด้วยตนเอง

การจัดทำคลังข้อมูลนี้มีการจับคู่ประโยคที่ตรงกันของคลังข้อมูลภาษาต้นฉบับและคลังข้อมูลภาษาเป้าหมาย โดยใช้หมายเลขบรรทัดหมายเลขเดียวกันในการกำกับและบ่งชี้ว่าประโยคคู่ใดเป็นคำแปลของกันและกัน

3.1.1 การคัดเลือกข้อมูลเพื่อสร้างคลังข้อมูลเทียบบท

ผู้วิจัยได้เก็บรวบรวมรายงานข่าวตลาดหุ้นซึ่งเป็นข้อมูลอิเล็กทรอนิกส์ที่เผยแพร่ไว้ในเว็บไซต์ตลาดหลักทรัพย์แห่งประเทศไทย (http://www.set.or.th) เพื่อนำมาใช้เป็นคลังข้อมูลคู่ภาษา เนื่องจากภายในเว็บไซต์ตลาดหลักทรัพย์แห่งประเทศไทยประกาศรายงานตลาดหุ้นทั้งภาษาไทยและภาษาอังกฤษ ผู้วิจัยนำคู่รายงานข่าวตลาดหุ้นจำนวน 503 คู่รายงานมาเก็บรวบรวมเป็นคลังข้อมูล โดยคัดเลือกเฉพาะรายงานตลาดหุ้นแบบรายวันประเภทการรับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียน และการรับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติม รายงานตลาดหุ้นแบบรายวันประเภทตลาดหลักทรัพย์เพิ่มสินค้า และ รายงานตลาดหุ้นแบบรายวัน ประเภทการขึ้นเครื่องหมายภายในตลาดหลักทรัพย์ คลังข้อมูลที่รวบรวมได้นี้มีขนาดคำรวม 2 ภาษาจำนวน 24,781 คำ โดยแบ่งเป็นภาษาไทย 13,464 คำ และภาษาอังกฤษ 11,317 คำ และมีคู่ข้อความจำนวน 1,607 คู่

ผู้วิจัยได้ใช้โปรแกรม Httrack รุ่น 3.40-2 ซึ่งเป็นฟรีแวร์ของซาเวียร์ โรช (Xavier Roche) ที่สามารถดาวน์โหลดเว็บไซต์เพื่อมาสร้างฐานข้อมูลแบบออฟไลน์ได้โดยอัตโนมัติในการเก็บรายงานตลาดหุ้นทั้งหมดจากเว็บไซต์ตลาดหลักทรัพย์แห่งประเทศไทย และได้นำข้อมูลทั้งหมดนั้นมาคัดแยกเฉพาะรายงานตลาดหุ้นแบบรายวันประเภทการรับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียน และการรับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติม รายงานตลาดหุ้นแบบรายวันประเภทตลาดหลักทรัพย์เพิ่มสินค้า และรายงานตลาดหุ้นแบบรายวัน ประเภทการขึ้นเครื่องหมายภายในตลาดหลักทรัพย์ทั้งส่วนที่เป็นทั้งภาษาไทยและภาษาอังกฤษด้วยตนเอง

จากการสืบค้นข้อมูลทำให้ทราบว่ารายงานตลาดหุ้นแบบรายวันแบ่งประเภทการขึ้นเครื่องหมายภายในตลาดหลักทรัพย์ออกเป็น 7 ประเภทได้แก่

(1) การขึ้นเครื่องหมาย ‘H’ (trading halt) เป็นเครื่องหมายแสดงการห้ามซื้อขายหลักทรัพย์จดทะเบียนเป็นการชั่วคราวโดยแต่ละครั้ง มีขั้นตอนเวลาไม่เกินกว่าหนึ่งรอบการซื้อขาย

(2) การขึ้นเครื่องหมาย ‘SP’ (trading suspension) เป็นเครื่องหมายแสดงการห้ามซื้อขายหลักทรัพย์จดทะเบียนเป็นการชั่วคราว โดยแต่ละครั้ง มีขั้นตอนเวลาเกินกว่าหนึ่งรอบการซื้อขาย

(3) การขึ้นเครื่องหมาย ‘NP’ (notice pending) เป็นเครื่องหมายแสดงถึงบริษัทจดทะเบียนมีข้อมูลที่ต้องรายงานและตลาดหลักทรัพย์อยู่ระหว่างรอข้อมูลจากบริษัท

(4) การขึ้นเครื่องหมาย ‘NR’ (notice recieved) เป็นเครื่องหมายแสดงถึงตลาดหลักทรัพย์ได้รับการชี้แจงข้อมูลจากบริษัทจดทะเบียนที่อยู่ในระหว่างพิจารณา (Pending) หรือได้รับการขึ้นเครื่องหมาย ‘NP’ ไว้แล้วและจะขึ้นเครื่องหมาย ‘NR’ เป็นเวลา 1 วัน

(5) การขึ้นเครื่องหมาย ‘NC’ (non-compliance) เป็นเครื่องหมายแสดงถึงหลักทรัพย์ของบริษัทจดทะเบียนที่เข้าข่ายอาจถูกเพิกถอน

(6) การขึ้นเครื่องหมาย ‘CM’ (call market) เป็นเครื่องหมายแสดงถึงบริษัทจดทะเบียนที่มีการกระจายการถือหุ้นไม่ครบถ้วนตามข้อกำหนดของตลาดหลักทรัพย์ว่าด้วย การดำรงสถานะเป็นบริษัทจดทะเบียนในตลาดหลักทรัพย์ และตลาดหลักทรัพย์กำหนดให้หุ้นสามัญของบริษัทดังกล่าวทำการซื้อขายด้วยวิธีภายใต้ระบบการจับคู่ในช่วงเวลา[6] (call market)

(7) การขึ้นเครื่องหมาย ‘ST’ (stabilization) เป็นเครื่องหมายแสดงถึงหุ้นของบริษัทจดทะเบียนที่มีการซื้อหุ้นเพื่อส่งมอบหุ้นที่จัดสรรเกิน

---

[6] ระบบการจับคู่ในช่วงเวลาคือการซื้อขายหลักทรัพย์ที่มีการกระจายการถือหุ้นไม่เป็นไปตามเกณฑ์ที่กำหนดเพื่อให้การซื้อขายหลักทรัพย์จดทะเบียนมีประสิทธิภาพและเหมาะสมกับการกระจายการถือหุ้นของบริษัท ตลาดหลักทรัพย์ได้กำหนดหลักเกณฑ์การซื้อขายหลักทรัพย์จดทะเบียนที่มีการกระจายการถือหุ้นไม่เป็นไป ตามเกณฑ์ที่ตลาดหลักทรัพย์กำหนด โดยภายใต้ระบบการจับคู่ในช่วงเวลา (call market) ระบบจะจับ คู่คำสั่งซื้อขายโดยอัตโนมัติในคราวเดียว ณ ราคาเดียว (single price auction) และเป็นราคาที่ทำให้ปริมาณการซื้อขายมากที่สุด ซึ่งจะช่วยลดความผันผวนของราคาและเพิ่มสภาพคล่องในการซื้อขายหลักทรัพย์

อย่างไรก็ตามจากการเก็บรวบรวมข้อมูลคู่ภาษาจากรายงานตลาดหลักทรัพย์ ผู้วิจัยพบรายงานประเภทรายงานตลาดหุ้นแบบรายวันประเภทการขึ้นเครื่องหมายเพียง 4 ประเภท ได้แก่ การขึ้นเครื่องหมาย ‘H’ การขึ้นเครื่องหมาย ‘SP’ การขึ้นเครื่องหมาย ‘NP’ และการขึ้นเครื่องหมาย ‘NR’ เท่านั้น จึงจะใช้รายงานตลาดหุ้นแบบรายวันเพียง 4 ประเภทนี้

3.1.2 การสร้างคลังข้อมูลเทียบบท

ข้อมูลคู่ภาษาที่เก็บรวบรวมมานั้น เดิมเป็นข้อความที่อยู่ในรูปแบบภาษา HTML เนื่องจากเป็นข้อมูลที่ประกาศอยู่บนเว็บไซต์ ดังนั้นข้อมูลคู่ภาษาที่ได้จึงมีป้ายระบุ HTML (HTML tag) อยู่ดังรูปที่ 6 ทางผู้วิจัยต้องแปลงให้อยู่ในรูปแบบข้อความล้วน (plain text) โดยการลบส่วนป้ายระบุ HTML ออกแต่ยังคงข้อความต้นฉบับไว้

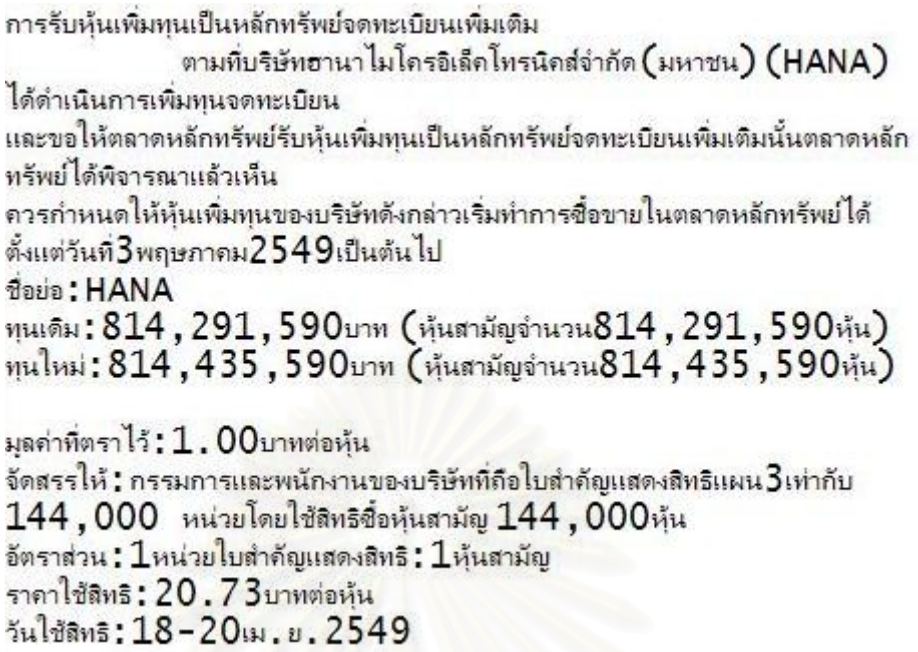
<h1>การรับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติม</h1>
ตามที่บริษัทฮานา ไมโครอิเล็คโทรนิคส์จำกัด (มหาชน) (HANA)
ได้ดำเนินการเพิ่มทุนจดทะเบียน
และขอให้ตลาดหลักทรัพย์รับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติมนั้นตลาดหลักทรัพย์ได้พิจารณาแล้วเห็น
ควรกำหนดให้หุ้นเพิ่มทุนของบริษัทดังกล่าวเริ่มทำการซื้อขายในตลาดหลักทรัพย์ได้
ตั้งแต่วันที่3พฤษภาคม2549เป็นต้นไป<p>
ชื่อย่อ:HANA
ทุนเดิม:814,291,590บาท (หุ้นสามัญจำนวน814,291,590หุ้น)

ทุนใหม่:814,435,590บาท (หุ้นสามัญจำนวน814,435,590หุ้น)

มูลค่าที่ตราไว้:1.00บาทต่อหุ้น
จัดสรรให้:กรรมการและพนักงานของบริษัทที่ถือใบสำคัญแสดงสิทธิแผน3เท่ากับ
144,000 หน่วยโดยใช้สิทธิซื้อหุ้นสามัญ 144,000หุ้น
อัตราส่วน:1หน่วยใบสำคัญแสดงสิทธิ:1หุ้นสามัญ
ราคาใช้สิทธิ:20.73บาทต่อหุ้น
วันใช้สิทธิ:18-20เม.ย.2549

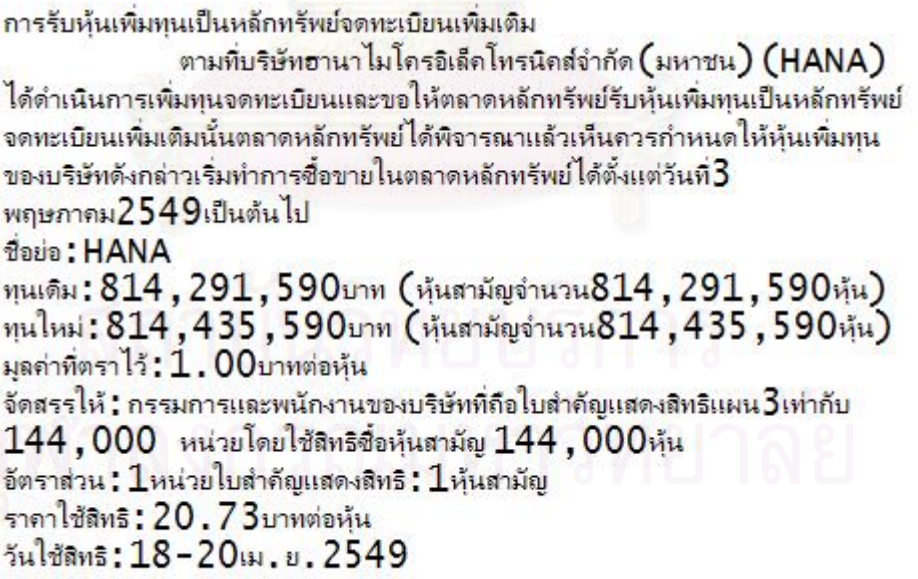
รูปที่ 6 แสดงตัวอย่างข้อความที่มีป้ายระบุ HTML

อย่างไรก็ตามเนื่องจากรูปแบบภาษา HTML มีเครื่องหมายและป้ายระบุแทนที่การเว้นวรรค การขึ้นบรรทัดใหม่และมีการแบ่งข้อความลงตารางทำให้ข้อความขาดตอนถึงแม้ว่าจะยังไม่จบประโยคก็ตาม ดังนั้นเมื่อแปลงเป็นข้อความล้วนแล้ว รูปแบบการเว้นบรรทัดและการเว้นวรรคจะยังคงเหลืออยู่ทำให้ข้อความไม่ต่อเนื่อง และยากต่อการเตรียมข้อมูลเพื่อทำการวิจัย ดังรูปที่



รูปที่ 7 แสดงตัวอย่างข้อความล้วนที่นำป้ายระบุ HTML ออกแล้วแต่ข้อความไม่ต่อเนื่อง

ดังนั้นผู้วิจัยจึงต้องปรับแต่งการแบ่งวรรคตอนให้มีความต่อเนื่องเสมือนกับภาษาที่ใช้ทั่วไปด้วยตนเอง ทั้งนี้เพื่อการสะดวกต่อการตัดแบ่งข้อความและมีรูปแบบเสมือนภาษาจริงให้มากที่สุด ดังรูปที่ 8



รูปที่ 8 แสดงตัวอย่างข้อความล้วนที่ปรับแต่งการแบ่งวรรคตอนแล้ว

การจับคู่บรรทัดภายในคลังข้อมูลเทียบบทโดยกำหนดให้ภาษาต้นฉบับอยู่ทางซ้ายและภาษาเป้าหมายอยู่ทางขวาโดยใช้อักขระกั้นระยะ (tab) เป็นตัวกั้นดังรูปที่ 9 และกำหนดให้ 1 คู่ข้อความต่อ 1 บรรทัดแล้วจึงจัดเก็บลงแฟ้มข้อมูล

1 ตลาดหลักทรัพย์เพิ่มสินค้า : AKR SET adds new listed securities : AKR
2 ตลาดหลักทรัพย์แห่งประเทศไทยขอแจ้งว่าคณะกรรมการตลาดหลักทรัพย์แห่งประเทศไทยได้สั่งให้รับหุ้นสามัญของบริษัท เอกรัฐวิศวกรรม จากัด (มหาชน) เป็นหลักทรัพย์จดทะเบียนในตลาดหลักทรัพย์ ตั้งแต่วันที่ 7 สิงหาคม 2549เป็นต้นไป The Stock Exchange of Thailand (SET) reported that the SET's board has granted a listing of common shares of Ekarat Engineering Public Company Limited from 7 August 2006 onwards.
3 ตลาดหลักทรัพย์ จึงเห็นควรกาหนดให้หุ้นสามัญของบริษัท เอกรัฐวิศวกรรม จากัด (มหาชน) จานวน 790,173,640 หุ้น มูลค่าที่ตราไว้หุ้นละ 1 บาท รวม 790,173,640 บาท ทาการซื้อขายในตลาดหลักทรัพย์ได้โดยจัดอยู่ในกลุ่มอุตสาหกรรมทรัพยากร หมวดธุรกิจพลังงานและสาธารณูปโภค และใช้ชื่อในการซื้อขายหลักทรัพย์ว่า "AKR" ทั้งนี้กาหนดให้เริ่มซื้อขายได้ตั้งแต่วันที่ 7 สิงหาคม 2549 เป็นต้นไป The SET has set common shares of Ekarat Engineering Public Company Limited, amounting to 790,173,640 shares with a par value of Baht 1 per share totaling of Baht 790,173,640 to be traded on the SET under Industry group of Resources, Energy and Utilities sector and using the trading name of "AKR" commencing from 7 August 2006 onwards.
4 หมายเหตุ : ผู้ลงทุนสามารถศึกษาข้อมูลบริษัทได้จากสรุปข้อสนเทศของ AKR ได้จากระบบบริการข้อมูลตลาดหลักทรัพย์ (SETSMART) และข้อมูลของ AKR ที่ http://www.ekarat-transformer.com Note : Please see the Information Memorandum of AKR disseminated on the SET Market Analysis and Reporting Tool (SETSMART) and AKR's information from website : http://www.ekarat-transformer.com

รูปที่ 9 แสดงตัวอย่างข้อความที่ปรับแต่งข้อมูลและจับคู่บรรทัดแล้ว

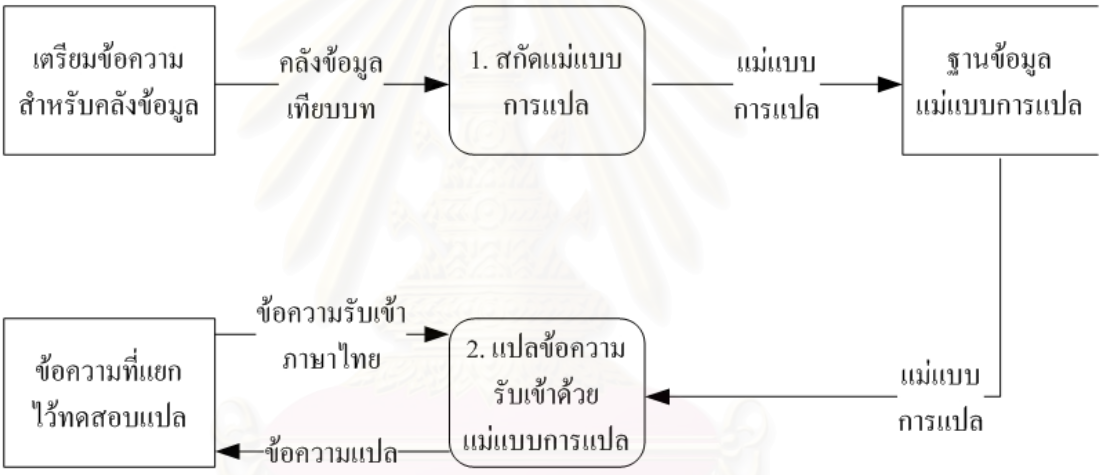
3.1.3 ลักษณะทางภาษาของคลังข้อมูลเทียบบท

เมื่อได้ข้อมูลคู่ภาษาที่เป็นข้อความล้วนและทำการปรับแต่งการแบ่งวรรคตอนแล้ว จึงนำมารวบรวมไว้ด้วยกันเป็นคลังข้อมูลและจับคู่เทียบบทข้อความให้ตรงกันเพื่อให้เป็นคลังข้อมูลเทียบบท ผู้วิจัยใช้เกณฑ์การตัดข้อความจากลักษณะของรูปแบบที่คล้ายกันแทนที่การตัดประโยคและการตัดย่อหน้า เนื่องจากในภาษาไทยยังมีความกำกวมสำหรับเกณฑ์การตัดประโยคอยู่และไม่มีเครื่องหมายจบประโยคที่ชัดเจนดังเช่นภาษาอังกฤษที่มีเครื่องหมายมหัพภาค (full-stop/‘.’) และจากข้อมูลคู่ภาษาที่เก็บสะสมมามีความหลากหลายของลักษณะการเว้นบรรทัดและเว้นวรรค ผู้วิจัยจึงไม่สามารถตรวจสอบได้อย่างชัดเจนว่าข้อมูลคู่ภาษานั้นมีย่อหน้าเริ่มต้นและสิ้นสุดลงตรงไหน ทำให้ไม่สามารถที่จะจับคู่เทียบบทได้ ดังนั้นจึงตัดแบ่งข้อความและตรวจสอบความถูกต้องของการสะกดคำด้วยตนเอง เพื่อความถูกต้องแม่นยำและสะดวกต่อการเป็นตัวอย่างให้เครื่องเรียนรู้

จากการวิเคราะห์ข้อมูลในคลังข้อมูลเทียบบทที่เก็บสะสมมา พบว่ามีการปรากฏของเอนทิตีระบุนาม (named entity) อยู่เป็นจำนวนมาก ซึ่งเอนทิตีระบุนามนั้นหมายรวมถึงคำนามชื่อเฉพาะ (proper noun) ประเภทชื่อคน ชื่อบริษัท ชื่อสถานที่ เช่น “ตลาดหลักทรัพย์”

“บริษัท เอส.อี.ซี. ออโต้เซลส์ แอนด์ เซอร์วิส จำกัด (มหาชน)” ฯลฯ และเอนทิตีระบุเวลา (temporal entity) ประเภทวันที่ เช่น “วันที่ 16 พฤษภาคม 2549” “18 May 2006” ฯลฯ และกลุ่มตัวเลขที่เป็นการนับเช่น “256,322” ต้องเป็น “สองแสนห้าหมื่นหกพันสามร้อยยี่สิบสอง” ไม่ควรเป็น “สองห้าหกสามสองสอง” การกำกับเอนทิตีระบุนามให้เป็นคำเดียวเป็นเรื่องที่มีความสำคัญต่อการประมวลภาษาธรรมชาติ เนื่องจากเอนทิตีระบุนามเป็นส่วนที่ควรรวมเป็นคำเดียวและจัดการแยกส่วนกับคำอื่นๆ ทั่วไป เอนทิตีระบุนามจะมีคำแปลหรือความหมายต่างจากคำอื่นๆ ถึงแม้ว่าจะสะกดเหมือนกัน บางครั้งเอนทิตีระบุนามก็ไม่อาจใช้การแปลได้ แต่ต้องใช้ถอดอักษรไทยเป็นโรมันแทน ดังนั้นผู้วิจัยจึงแยกกลุ่มเอนทิตีระบุนามไว้ต่างหากจากคำทั่วไปโดยกำหนดให้เครื่องรู้จำเฉพาะส่วนเอนทิตีระบุนาม โดยพบเอนทิตีระบุนามที่เป็นชื่อบริษัทมีการปรากฏในแต่ละข้อความมากที่สุดจำนวน 1,436 ข้อความจากจำนวนข้อความทั้งหมด 1,607 ข้อความ คิดเป็นร้อยละ 89.36 และรองลงมาคือเอนทิตีระบุนามที่เป็นวันที่จำนวน 1,016 ข้อความ คิดเป็นร้อยละ 63.22 ส่วนเอนทิตีระบุนามที่ปรากฏน้อยที่สุดคือชื่อบุคคลจำนวน 5 ข้อความ คิดเป็นร้อยละ 0.31

ระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างที่จะนำมาใช้ในงานวิจัยนี้ ได้นำแนวคิดและวิธีการจากงานของแม็คเทท (Mctait, 2002) มาเป็นแนวทางหลักในงานวิจัยชิ้นนี้ อย่างไรก็ตามงานของแม็คเททเป็นการทดลองสกัดหาแม่แบบการแปลและการรวมคำแปลใหม่เพื่อแปลภาษาจากภาษาอังกฤษเป็นภาษาฝรั่งเศสซึ่งเป็นภาษาในตระกูลเดียวกันและมีคำศัพท์จำนวนมากที่มาจากคำร่วมเชื้อสายและรากศัพท์เดียวกัน นอกจากนี้ยังมีลักษณะการเรียงตัวทางวากยสัมพันธ์และการเว้นวรรคระหว่างคำคล้ายคลึงกัน ดังนั้นระบบการเทียบแปลคำศัพท์โดยสืบค้นไปยังรากคำศัพท์และระบบการตัดคำโดยดูจากการเว้นวรรคจึงไม่สามารถนำมาใช้กับงานวิจัยชิ้นนี้ได้ ผู้วิจัยจึงคัดเฉพาะส่วนที่สามารถนำมาประยุกต์ใช้ได้กับการแปลภาษาจากภาษาไทยเป็นภาษาอังกฤษ อย่างไรก็ดีการนำมาประยุกต์ใช้ให้เหมาะสมและได้ผลลัพธ์การแปลที่ดีจึงต้องมีการปรับเปลี่ยนแนวคิดบางส่วนเพื่อให้เหมาะกับการนำมาใช้เป็นระบบแปลภาษาจากภาษาไทยเป็นภาษาอังกฤษซึ่งจะกล่าวต่อไป



รูปที่ 10 แสดงการทำงานหลักของระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง

การทำงานของระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างมี 2 ขั้นตอนหลักซึ่งจะเริ่มที่ (1) ระบบการสกัดแม่แบบการแปล ซึ่งเป็นขั้นตอนการสกัดแม่แบบการแปลจากคลังข้อมูลเทียบบทที่ผู้วิจัยได้จัดเตรียมไว้ดังที่กล่าวมาก่อนหน้า แล้วนำแม่แบบการแปลที่สกัดได้ไปเก็บลงฐานข้อมูล (database) หลังจากนั้นจึงนำแม่แบบการแปลที่เก็บไว้มาแปลข้อมูลหรือเอกสารรับเข้าและระบบจะทำงานส่วน (2) ระบบการรวมคำแปลใหม่ ซึ่งเป็นในขั้นตอนการเรียบเรียงคำแปลเพื่อให้ผลการแปลที่ได้สมบูรณ์และมีการเรียงตัวของคำเป็นภาษาธรรมชาติดังรูปที่ 10

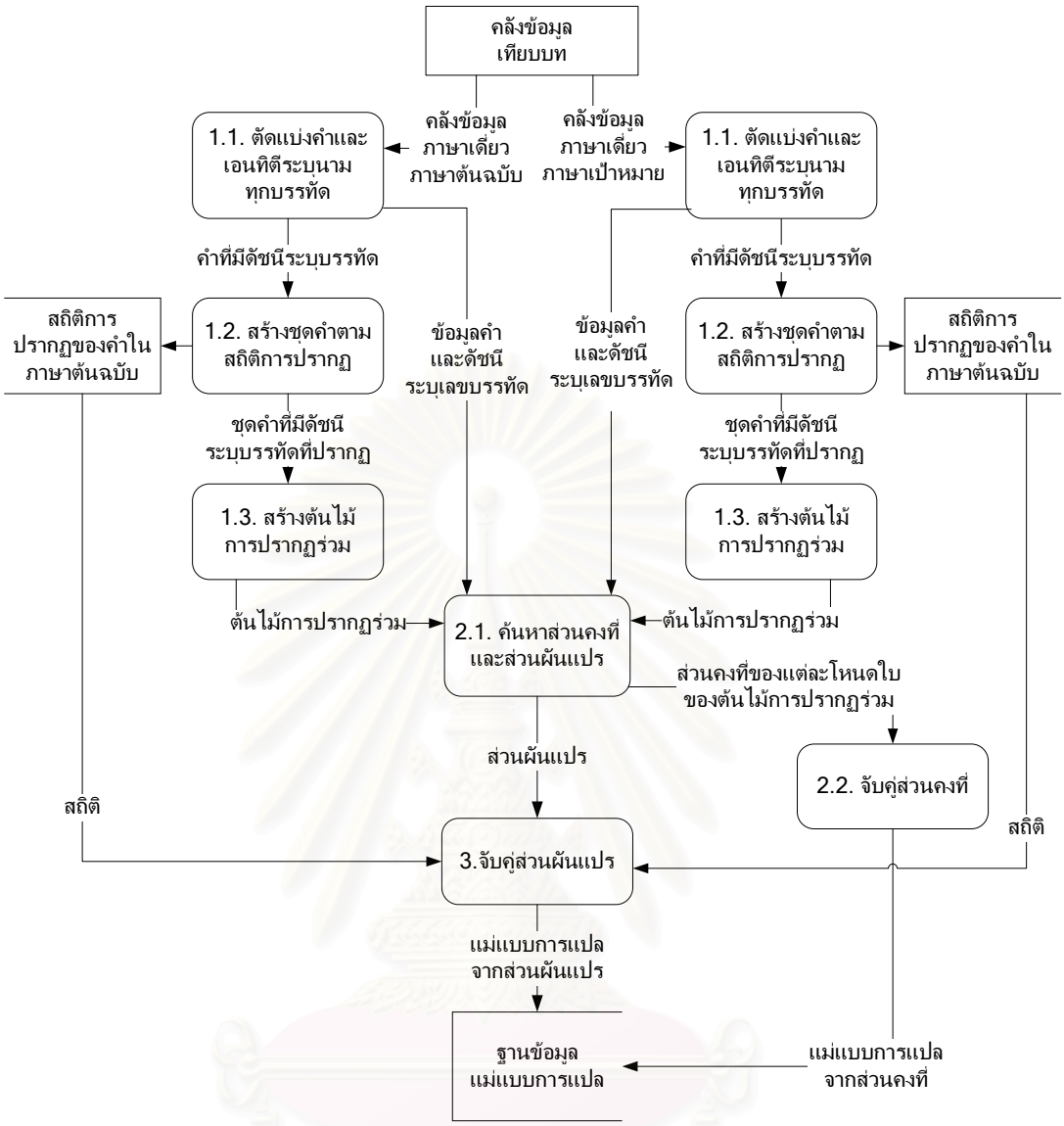
3.2 ระบบการสกัดแม่แบบการแปล

ระบบสกัดหาแม่แบบการแปลคือระบบที่จะสกัดหาส่วนคงที่และส่วนผันแปรจากส่วนซ้ำและส่วนไม่ซ้ำของข้อความภายในคลังข้อมูลเทียบบทมาสร้างเป็นแม่แบบการแปล แบ่งออกเป็นขั้นตอนใหญ่ได้ 3 ขั้นตอนดังรูปที่ 11 ด้านล่าง คือ

(1) ขั้นตอนภาษาเดี่ยวซึ่งจะทำงานแยกแต่ละภาษาเพื่อจะค้นหากลุ่มคำที่เป็นส่วนซ้ำและส่วนไม่ซ้ำภายในคลังข้อมูล ภายในขั้นตอนภาษาเดี่ยวสามารถแบ่งได้เป็น 3 ขั้นตอนย่อยคือ (1.1) ขั้นตอนการตัดแบ่งคำคือส่วนที่จะบอกว่าตัวอักษรกลุ่มใดเป็น 1 คำ (1.2) ขั้นตอนการสร้างชุดคำคือส่วนค้นหาว่าคำใดมีการปรากฏอยู่ภายในข้อความใดบ้างโดยใช้เป็นเลขบรรทัดตัวบ่งชี้ (1.3) ขั้นตอนข้อมูลการปรากฏร่วมคือส่วนค้นหาคำที่ปรากฏร่วมกันในแต่ละข้อความและจะสร้างโครงสร้างต้นไม้การปรากฏร่วมของคำ

(2) ขั้นตอนเทียบสองภาษาซึ่งจะค้นหาแม่แบบการแปลส่วนคงที่จากส่วนซ้ำโดยการเปรียบเทียบคลังข้อมูลทั้ง 2 ภาษา โดยแบ่งเป็นขั้นตอนย่อยได้ 2 ขั้นตอนคือ (2.1) ขั้นตอนการค้นหาส่วนคงที่และส่วนผันแปรจากส่วนซ้ำและส่วนไม่ซ้ำของข้อความ และ (2.2) ขั้นตอนจับคู่ส่วนคงที่เพื่อสร้างแม่แบบการแปลจากส่วนคงที่

(3) ขั้นตอนการจับคู่ซึ่งจะค้นหาแม่แบบการแปลจากส่วนผันแปรที่เหลือ โดยมีกระบวนการย่อยอยู่ 2 กระบวนการคือ (3.1) กระบวนการคำนวณเมตริกความคล้ายคลึงของคู่ภาษา และ (3.2) อัลกอริธึมการเปรียบเทียบลำดับ



รูปที่ 11 แสดงขั้นตอนการทำงานของระบบสกัดหาแม่แบบการแปล

โดยงานวิจัยนี้ ผู้วิจัยได้นิยามแบบจำลองทางคณิตศาสตร์ของแม่แบบการแปลเป็นลำดับ 4 ส่วน (quadruple) (S, T, A_f, A_v) เมื่อ

S แทนลำดับของส่วนข้อความในภาษาต้นฉบับซึ่งแทนด้วย F_k^S คั่นด้วยตัวแปร V_k^S ซึ่งสามารถแทนค่าได้ด้วยส่วนข้อความใดๆ ในภาษาต้นฉบับ เมื่อ $k \in \mathbb{N}$ เช่น $S = \langle F_1^S, V_1^S, F_2^S, V_2^S \rangle$

T แทนลำดับของส่วนข้อความในภาษาเป้าหมายทางซึ่งแทนด้วย F_k^T คั่นด้วยตัวแปร V_k^T ซึ่งสามารถแทนที่ได้ด้วยส่วนข้อความใดๆ ในภาษาเป้าหมายทาง เมื่อ $k \in \mathbb{N}$ เช่น $T = \langle V_1^T, F_1^T, V_2^T, F_2^T \rangle$

$A_f \subseteq \Pi_F^S \times \Pi_F^T$ แทนเซ็ตการจับคู่ของเซ็ตของส่วนข้อความในภาษาต้นฉบับและในภาษาเป้าหมายทาง เมื่อ Π_F^S เป็นการแบ่งส่วน (Partition) ของเซ็ตส่วนข้อความ $F_k^S \in S$ และ Π_F^T เป็นการแบ่งส่วนของเซ็ตส่วนข้อความ $F_k^T \in T$

$A_v \subseteq \Pi_V^S \times \Pi_V^T$ แทนเซ็ตการจับคู่ของเซ็ตของตัวแปรในภาษาต้นฉบับและในภาษาเป้าหมายทาง เมื่อ Π_V^S เป็นการแบ่งส่วนของเซ็ตตัวแปร $V_k^S \in S$ และ Π_V^T เป็นการแบ่งส่วนของเซ็ตตัวแปร $V_k^T \in T$

ทั้งนี้ ส่วนข้อความ (Text fragment) คือลำดับต่อเนื่องของคำซึ่งเป็นส่วนหนึ่งของบรรทัด ภายในลำดับ S จะมีส่วนข้อความ F_p^S จำนวนเท่าไรก็ได้ $p > 0$ ตัว แต่จะต้องมีตัวแปร V_p^S จำนวน p, $p+1$ หรือ $p-1$ ตัวเท่านั้น เช่นเดียวกัน ภายในลำดับ T จะมีส่วนข้อความ F_q^T จำนวนเท่าไรก็ได้ $q > 0$ ตัว โดยจะมีตัวแปร V_q^T จำนวน q, $q+1$ หรือ $q-1$ ตัวเท่านั้น

ตัวอย่าง แม่แบบการแปล

X_1 recently gave X_2 up $\Leftrightarrow$ Y_1 ได้ ยกเลิก Y_2 ไป เมื่อ ไม่นานนี้ : $\{(X_1 \Leftrightarrow Y_1),(X_2 \Leftrightarrow Y_2)\}$

สามารถแทนได้ด้วยลำดับ 4 ส่วนได้เป็น (S,T,A_f,A_v) เมื่อ

$S = \langle V_1^S, F_1^S, V_2^S, F_2^S \rangle$ เมื่อ F_1^S = <recently, gave>, F_2^S = <up>, $V_1^S = X_1$ และ $V_2^S = X_2$

$T = \langle V_1^T, F_1^T, V_2^T, F_2^T \rangle$ เมื่อ F_1^T = <ได้, ยกเลิก>, F_2^T = <ไป, เมื่อ, ไม่นานนี้>, $V_1^T = Y_1$ และ $V_2^T = Y_2$

$A_f = \{(\{F_1^S\},\{F_2^S\},\{F_1^T\},\{F_2^T\})\}$

$A_v = \{(\{X_1\},\{Y_1\}),(\{X_2\},\{Y_2\})\}$

เป็นต้น

จากการนิยามนี้ แม่แบบการแปลจึงไม่จำเป็นต้องเป็นกลุ่มคำที่เรียงต่อกัน ดังเช่น X_0 gave X_1 up $\leftrightarrow$ Y_0 ยกเลิก Y_1 : $\{(X_0 \Leftrightarrow Y_0),(X_1 \Leftrightarrow Y_1)\}$ และสามารถหาคำที่เป็นส่วนคงที่และส่วนผันแปรได้ในแต่ละบรรทัด ทำให้การสกัดแม่แบบการแปลมีความยืดหยุ่นและสามารถสกัดแม่แบบการแปลจากภาษาที่ซับซ้อนและมีสำนวนได้

3.2.1 ขั้นตอนภาษาเดี่ยว

ขั้นตอนภาษาเดี่ยวเป็นการวิเคราะห์สถิติการเกิดขึ้นของคำภายในคลังข้อมูลภาษาเดี่ยวซึ่งแยกกันทำในแต่ละภาษา ซึ่งมีการทำงาน 3 ขั้นตอนคือ (1) ขั้นตอนการตัดแบ่งคำ (2) ขั้นตอนการสร้างชุดคำ และ (3) ขั้นตอนข้อมูลการปรากฏร่วม

3.2.1.1 ขั้นตอนการตัดแบ่งคำ

ขั้นตอนการตัดแบ่งคำคือส่วนที่จะค้นหาขอบเขตของคำภายในคลังข้อมูลซึ่งจะทำการค้นหาและแบ่งขอบเขตของคำทุกบรรทัด ระบบการทำงานในขั้นตอนนี้ผู้วิจัยกำหนดไว้ว่าอักขระว่างเป็นขอบเขตของคำแต่ละคำ ซึ่งอักขระว่างที่เป็นขอบเขตของคำเป็นผลการทำงานของโปรแกรมตัดคำอัตโนมัติ ‘SWATH’ (ไพศาล, 2541) ของศูนย์คอมพิวเตอร์และอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ซึ่งเมื่อได้ผลการตัดคำมาแล้ว ผู้วิจัยได้ตรวจสอบความถูกต้องและแก้ไขด้วยตนเองอีกครั้งหนึ่ง เช่น “ตลาดหลักทรัพย์แห่งประเทศไทย ขอ แจ้ง ว่า คณะกรรมการตลาดหลักทรัพย์แห่งประเทศไทย ได้ สั่ง ให้ รับ หุ้นสามัญ” ก็จะได้คำ 10 คำคือ [“ตลาดหลักทรัพย์แห่งประเทศไทย” “ขอ” “แจ้ง” “ว่า” “คณะกรรมการตลาดหลักทรัพย์แห่งประเทศไทย” “ได้” “สั่ง” “ให้” “รับ” และ“หุ้นสามัญ”]

หลังจากนั้นเนื้อความในคลังข้อมูลสองภาษาจะถูกจับคู่เพื่อสร้างคลังคู่ข้อความสองภาษา การจับคู่เนื้อความจะทำในระดับย่อหน้า โดยพิจารณาจากจำนวนย่อหน้าทั้งในเอกสารฝั่งไทยและเอกสารฝั่งภาษาอังกฤษ เกณฑ์การจับคู่เนื้อความที่ผู้วิจัยพบภายในคลังข้อมูลสองภาษามีดังต่อไปนี้

(1) ถ้าย่อหน้าภายในเอกสารทั้งสองภาษาสามารถจับคู่ได้แบบ 1 ต่อ 1 ย่อหน้าดังกล่าวจะถูกจับคู่เป็น 1 คู่

(2) ถ้าหลายย่อหน้าในภาษาไทยสามารถจับคู่ได้กับ 1 ย่อหน้าภาษาอังกฤษ ก็ให้แบ่งย่อหน้าภาษาอังกฤษออกเป็นหลายส่วนให้สอดคล้องเนื้อความของย่อหน้าภาษาไทย แล้วจึงจับคู่แต่ละย่อหน้า

(3) ถ้า 1 ย่อหน้าภาษาไทยสามารถจับคู่ได้กับหลายย่อหน้าภาษาอังกฤษ ก็ให้แบ่งย่อหน้าภาษาไทยออกเป็นหลายส่วนให้สอดคล้องกับเนื้อความของย่อหน้าภาษาอังกฤษ แล้วจึงจับคู่แต่ละย่อหน้า

ทั้งนี้ผู้วิจัยไม่พบกรณีที่หลายย่อหน้าภาษาไทยสามารถจับคู่กับหลายย่อหน้าภาษาอังกฤษ จึงไม่ได้กำหนดเกณฑ์การจับคู่เนื้อความสำหรับกรณีนี้ไว้

อย่างไรก็ตามเอนติทีระบุนามประเภทชื่อบริษัทและวันที่ เช่น “ตลาดหลักทรัพย์แห่งประเทศไทย” หรือ “วันที่ 15 มีนาคม 2549” เป็นต้น มักจะมีอักขระว่างแทรกอยู่ภายใน เพราะเอนทิตีระบุนามเหล่านี้มีคู่คำแปลไม่เหมือนกับคำทั่วไปเช่น “วันที่ 15 มีนาคม 2549” แปลว่า “March 15, 2006” ไม่ได้แปลเป็นคำๆ แยกออกจากกัน เป็น “fifteenth day March 2549” ดังนั้นจึงต้องระบุคำเหล่านี้ออกมาให้ได้ก่อนการตัดคำ เพื่อให้เครื่องเรียนรู้ว่าคำหรือกลุ่มคำใดเป็นเอนติทีระบุนาม และสามารถรู้จำได้ว่าเอนติทีระบุนามเป็นคำเดียวถึงแม้ว่าจะมีอักขระว่างอยู่ภายใน ซึ่งการระบุคำที่เป็นเอนทิตีระบุนามในที่นี้ผู้วิจัยได้ใช้การรู้จำของนิพจน์ปกติ (regular expression) ซึ่งก็คือการใช้เครื่องหมายและสัญลักษณ์พิเศษมาตรวจสอบหรือเทียบค้นหาตัวอักษร (ดูภาคผนวก ก) เช่น การใช้เครื่องหมายปรัศนี (question mark) ‘?’ หมายความว่าแทนที่เครื่องหมายปรัศนีด้วยตัวอักษรอะไรก็ได้ อย่างน้อย 1 ตัว เช่น “ก?ข” หมายถึง จะมีตัวอักษร ‘ก’ 1 ตัว และมีตัวอักษรตัวสุดท้ายเป็นตัวอักษร ‘ข’ 1 ตัว แต่มีตัวอักษรอะไรก็ได้คั่นกลาง ตัวอักษร “ก” และ “ข” 1 ตัว ดังนั้น “ก?ข” จึงตรวจสอบพบ “กกข” และ “ก4ข” เป็นต้น

อย่างไรก็ตามข้อความภาษาไทยและภาษาอังกฤษต้องใช้นิพจน์ปกติสำหรับรู้จำที่แตกต่างกัน ดังเช่นเดียวกับนิพจน์ปกติที่ผู้วิจัยได้แยกใช้ในการค้นหาเอนทิตีระบุนามของแต่ละภาษา เช่น

(1) นิพจน์ปกติสำหรับค้นหาและรู้จำชื่อบริษัทในภาษาอังกฤษ

```
((\()?[A-Z][A-Za-z\-\.]+(\))?\s+)+Public\s+Company\s+Limited
```

การทำงานของนิพจน์ปกตินี้คือการตรวจสอบและค้นหากลุ่มของตัวอักษรภาษาอังกฤษตัวพิมพ์ใหญ่ 1 ตัวที่ตามด้วยตัวอักษรภาษาอังกฤษตัวพิมพ์ใหญ่หรือตัวพิมพ์เล็กหรือเครื่องหมายมหัพภาคกี่ตัวก็ได้อย่างน้อย 1 ตัวตามด้วยอักขระว่างอย่างน้อย 1 ตัวและกลุ่มของตัวอักษรนี้ต้องมีอย่างน้อย1 กลุ่มแล้วตามด้วยคำว่า ‘Public’ ตามด้วยอักขระว่างอย่างน้อย 1 ตัว แล้วตามด้วย คำว่า ‘Company’ ตามด้วยอักขระว่างอย่างน้อย 1 ตัวแล้วตามด้วย คำว่า ‘Limited’ ซึ่งนิพจน์ปกตินี้จะจับกลุ่มคำที่เป็นชื่อบริษัทในภาษาอังกฤษเป็นโทเค็นเดียวกัน เช่น ‘S.E.C. Auto Sales And Services Public Company Limited’ และ ‘Shin Corporation Public Company Limited’

(2) นิพจน์ปกติสำหรับค้นหาและรู้จำชื่อบริษัทในภาษาไทย

```
\xba\xc3\xd4\xc9\xd1\xb7\s*.+\s*\xa8\xd3\xa1\xd1\xb4\s*' \'\(\s*\xc1\xcb\xd2\xaa\xb9\s*\)
```

การทำงานของนิพจน์ปกตินี้คือการตรวจสอบและค้นหา ‘บริษัท’ ที่ตามด้วยตามด้วยอักขระว่างที่มีกี่ตัวก็ได้หรือไม่มีก็ได้และตามด้วยตัวอักษร อักขระพิเศษหรือตัวเลขอะไรก็ได้อย่างน้อย 1 ตัวตามด้วยตามด้วยอักขระว่างที่มีกี่ตัวก็ได้หรือไม่มีก็ได้ และตามด้วย คำว่า ‘(มหาชน)’ ซึ่งนิพจน์ปกตินี้จะจับกลุ่มคำที่เป็นชื่อบริษัทในภาษาไทยเป็นโทเค็นเดียวกัน เช่น บริษัท ไทยฮิทเอ็กซ์เช้นจ์ จำกัด (มหาชน) และ บริษัทอาหารสยามจำกัด (มหาชน)

(3) นิพจน์ปกติสำหรับค้นหาและรู้จำวันที่ในภาษาอังกฤษ

```
[A-Z][a-z]+\s*(\d{1,2}(\s*\,\s*\d\d\d\d)?
```

การทำงานของนิพจน์ปกตินี้คือการตรวจสอบและค้นหากลุ่มของตัวอักษรภาษาอังกฤษตัวพิมพ์ใหญ่ หรือตัวพิมพ์เล็กอย่างน้อย 1 ตัวที่ตามด้วยอักขระว่างที่มีกี่ตัวก็ได้หรือไม่มีก็ได้แล้วจึงตามด้วยตัวเลขจำนวน 1หรือ 2 ตัวตามด้วยกลุ่มของอักขระว่างที่มีกี่ตัวก็ได้หรือไม่มีก็ได้ตามด้วยเครื่องหมายจุลภาค(comma) ‘,’ ตามด้วยอักขระว่างที่มีกี่ตัวก็ได้หรือไม่มีก็ได้แล้วจึงมีตัวเลข 4 ตัว ซึ่งกลุ่มนี้จะมีหรือไม่ก็ได้ แต่ถ้ามี ต้องมีกลุ่มเดียว ซึ่งนิพจน์ปกตินี้จะจับกลุ่มคำและตัวเลขที่เป็นวันที่ต่างๆ เป็นโทเค็นเดียวกัน เช่น ‘September 8, 2006’ และ ‘May 26, 2006’

เมื่อทำการรู้จำเอนทิตีระบุนามและแบ่งขอบเขตคำตามอักขระว่างแล้วจึงสามารถนำข้อมูลเข้าสู่ระบบส่วนต่อไป

3.2.1.2 ขั้นตอนการสร้างชุดคำ

ขั้นตอนการสร้างชุดคำมีวัตถุประสงค์เพื่อค้นหาคำต่างๆ ที่ปรากฏในคลังข้อมูลว่าคำแต่ละคำเกิดขึ้นคำละจำนวนเท่าใดและปรากฏในบรรทัดใดบ้าง เพื่อนำไปใช้ในการค้นหาคำที่ปรากฏร่วมกันในขั้นตอนต่อไป

ภายในระบบขั้นตอนสร้างชุดคำจะค้นหาว่าคำใดมีการปรากฏอยู่ภายในบรรทัดใดบ้างโดยใช้เป็นเลขบรรทัดตัวบ่งชี้โดยการสร้างแฟ้มข้อมูลผกผัน (inverted file) ไว้เป็นไฟล์อ้างอิงเพื่อเก็บดัชนีบ่งชี้ถึงคำแต่ละคำในแต่ละบรรทัดสำหรับการคำนวณค่าทางสถิติว่าคำในคลังข้อมูลมีปริมาณเท่าใดและปรากฏในบรรทัดไหนบ้างซึ่งผลของการทำงานส่วนนี้จะทำให้ได้ชุดคำและดัชนีบ่งชี้ถึงคำแต่ละคำในคลังข้อมูลภาษาเดี่ยว

เช่นหากภายในคลังข้อมูลภาษาเดี่ยวมีข้อความ 5 บรรทัดและมีคำดังต่อไปนี้

บรรทัดที่	ตัวอย่าง
1	เริ่ม ทำ การ ซื้อขาย ใน ตลาดหลักทรัพย์ ได้ ตั้งแต่ วันที่ 15 เมษายน 2549
2	กำหนด ให้ เริ่ม ซื้อขาย ได้ ตั้งแต่ วันที่ 18 พฤษภาคม 2549 เป็นต้นไป
3	สรุป จาก ระบบ บริการ ข้อมูล ตลาดหลักทรัพย์
4	ผู้ลงทุน สามารถ ศึกษา ข้อมูล หลักทรัพย์ ได้ จาก ตลาดหลักทรัพย์
5	การ รับ หุ้น เพิ่ม ทุน เป็น หลักทรัพย์ จดทะเบียน เพิ่มเติม

จากตัวอย่างข้อมูลข้างต้น คำว่า “เริ่ม” ปรากฏเพียงแค่ในบรรทัดที่ 1 และ 2 ดังนั้นภายในแฟ้มดัชนีผกผันจะเก็บเลขบรรทัด 1 และ 2 เป็นค่าดัชนีบ่งชี้ของคำไว้ เป็น “(เริ่ม)[1,2]” ส่วนคำว่า “ทำ” ปรากฏแค่ในบรรทัดที่ 1 เท่านั้น จึงเก็บดัชนีบ่งชี้ของคำเป็น “(ทำ)[1]” ระบบจะทำการเก็บดัชนีบ่งชี้จนครบทุกคำ และได้ชุดคำและดัชนีบ่งชี้ในแฟ้มข้อมูลผกผันดังนี้

(เริ่ม)[1,2] (ทำ)[1] (การ)[1,5] (ซื้อขาย)[1,2] (ใน)[1] (ตลาดหลักทรัพย์)[1,3,4] (ได้)[1,2,4] (ตั้งแต่)[1,2] (วันที่ 15 เมษายน 2549)[1] (กำหนด)[2] (ให้)[2] (วันที่ 18 พฤษภาคม 2549)[2] (เป็นต้นไป)[2] (สรุป)[3] (จาก)[3,4] (ระบบ)[3] (บริการ)[3] (ข้อมูล)[3,4] (ผู้ลงทุน)[4] (สามารถ)[4] (ศึกษา)[4] (รับ)[5] (หุ้น)[5] (เพิ่ม)[5] (ทุน)[5] (เป็น)[5] (จดทะเบียน)[5] (เพิ่มเติม)[5]

ภายในแฟ้มข้อมูลผกผันจะประกอบด้วยรายการคำพร้อมด้วยหมายเลขบรรทัดกำกับบ่งบอกบรรทัดที่คำนั้นปรากฏ ชุดคำและดัชนีบ่งชี้ที่เป็นผลลัพธ์ของขั้นตอนนี้จะถูกเก็บไว้ในแฟ้มข้อมูลผกผันในฐานข้อมูลซึ่งจะถูกนำไปใช้ในขั้นตอนต่อไป

3.2.1.3 ขั้นตอนการสร้างต้นไม้การปรากฏร่วม

ขั้นตอนการสร้างต้นไม้การปรากฏร่วมคือการนำชุดคำและดัชนีบ่งชี้ที่เป็นผลลัพธ์จากขั้นตอนการสร้างชุดคำมาสร้างต้นไม้การปรากฏร่วมซึ่งเป็นขั้นตอนที่สำคัญมากในการสกัดแม่แบบการแปล

ขั้นตอนนี้จะสร้างเป็นการหาส่วนข้อความซ้ำภายในประโยคในคลังข้อมูลภาษาเดี่ยว โดยใช้ดัชนีแสดงความสัมพันธ์ระหว่างกลุ่มคำที่ปรากฏร่วมกันรวมทั้งตำแหน่งที่ปรากฏภายในคลังข้อมูลภาษา หลังจากนั้นจึงรวบรวมกลุ่มคำที่ปรากฏอยู่ในบรรทัด

เดียวกันเพื่อค้นหากลุ่มคำที่ยาวที่สุดที่ปรากฏร่วมกันหลายครั้ง มากำหนดให้เป็นส่วนซ้ำของข้อความเพื่อจะสร้างเป็นส่วนคงที่สำหรับแม่แบบการแปล

ทั้งนี้ระบบจะพยายามค้นหาข้อความที่ปรากฏร่วมกันในคลังข้อมูลสองครั้งขึ้นไป เพื่อแยกความแตกต่างของส่วนข้อความที่ซ้ำออกจากส่วนข้อความที่ไม่ซ้ำ ทั้งนี้กระบวนการหาส่วนข้อความซ้ำจะทำโดยการสร้างต้นไม้การปรากฏร่วมของส่วนข้อความ ยกตัวอย่างเช่น โหนด {(ตลาดหลักทรัพย์)(แจ้ง)[1, 2, 3, 4]} แสดงให้เห็นถึงการปรากฏร่วมของคำ ‘ตลาดหลักทรัพย์’ และ ‘แจ้ง’ ในบรรทัดที่ 1, 2, 3, และ 4

โหนดแม่เป็นโหนดที่สามัญ ในขณะที่โหนดลูกของโหนดแม่นั้นจะเป็นโหนดที่จำเพาะกว่าโหนดแม่ ยกตัวอย่างเช่น โหนด {(ตลาดหลักทรัพย์)(แจ้ง)[1, 2, 3, 4]} เป็นโหนดแม่ของโหนดลูก {(ตลาดหลักทรัพย์)(แจ้ง)(ยกเลิก)[2, 3, 4]}

หลังจากรวบรวมโหนดจนกลายเป็นต้นไม้แล้ว จะได้โหนดที่จำเพาะที่สุดพร้อมด้วยเลขบรรทัดโหนดนั้นปรากฏเป็นโหนดใบ ซึ่งจะเป็นโหนดที่ยาวที่สุด และจะถูกนำไปใช้ในการเทียบสองภาษา สาเหตุที่เลือกใช้โหนดที่ยาวที่สุดเป็นเพราะโหนดที่ยาวที่สุดเป็นกลุ่มของคำที่มักจะถูกนำมาใช้ร่วมกันบ่อยครั้ง ซึ่งมีแนวโน้มที่จะเป็นส่วนคงที่ของแม่แบบการแปล

ดังนั้นจึงกล่าวได้ว่าการสร้างต้นไม้การปรากฏร่วมมีจุดประสงค์หลักอยู่ที่การค้นหาโหนดใบซึ่งเป็นโหนดปลายทางของต้นไม้ซึ่งจะถูกนำมาใช้เทียบเคียงในขั้นตอนสองภาษาเพื่อไปเป็นแม่แบบการแปลต่อไป

กระบวนการสร้างต้นไม้ของการปรากฏร่วมสามารถอธิบายในรายละเอียดได้ดังต่อไปนี้ ซึ่งจะเป็นอัลกอริธึมของวิธีการสร้างต้นไม้การปรากฏร่วม ในบรรทัดแรกของทุกอัลกอริธึมจะระบุชื่อฟังก์ชัน (function) พารามิเตอร์รับเข้า (input parameter) และค่าส่งคืน (return)

อัลกอริธึมที่ 1 แสดงกระบวนการสร้างต้นไม้การปรากฏร่วมที่ระดับบน (McTait, 2001: 55)

```
algorithm CollectCollocs(L : LexicalItems) returns LeafNodeList
  let T be a new empty collocation tree node.

  for each lexicon l in L do
    let c be a collocation node of l
    if not AddToChildrenColloc(T,c) then
      AddChildColloc(T,c)
    end if
  end for

  let Q = CollectLeaves(T) be a list of leaf nodes of T
  Filter(Q)
  OrderLexicalItems(Q)
  return Q
end algorithm
```

อัลกอริธึมนี้แสดงกระบวนการสร้างต้นไม้การปรากฏร่วมที่ระดับบนซึ่งจะรับข้อมูลชุดคำ (LexicalItems) ที่กำกับด้วยดัชนีบ่งชี้เลขบรรทัดแล้วที่เป็นผลลัพธ์ของขั้นตอนการสร้างชุดคำ กระบวนการนี้เริ่มจากกำหนดให้ L เท่ากับรายการคำในข้อมูลชุดคำและ T เท่ากับโหนดต้นไม้การปรากฏร่วมต้นใหม่ที่ยังไม่มีโหนดใดๆ ภายใน โดยเริ่มตรวจสอบว่าคำว่า l อยู่ใน L และกำหนดให้ c เป็นโหนดการปรากฏร่วมของ l

กระบวนการสร้างต้นไม้การปรากฏร่วมที่ระดับบนนี้จะเริ่มจากการตรวจสอบว่า ถ้าเรียกคำสั่ง AddToChildrenColloc() จากอัลกอริธึมที่ 2 มาตรวจสอบเพื่อเพิ่ม c ลงใน T แล้วค่าส่งคืนเป็นเท็จ (false) ก็จะเรียกคำสั่ง AddChildColloc() มาสร้าง T อันใหม่และเพิ่ม c ลงไป

เมื่อเสร็จจากกระบวนการขั้นต้นแล้ว กำหนดให้ Q เท่ากับผลการทำงานของคำสั่ง CollectLeaves() จากอัลกอริธึมที่ 6 กับโหนดต้นไม้การปรากฏร่วมต้นที่ผ่านกระบวนการทำงานขั้นต้นด้านบนมาแล้วซึ่ง Q จะเท่ากับชุดของโหนดใบของ T และจะเริ่มจัดรูปแบบ Q ด้วยการกรองจากการเรียกคำสั่ง Filter() จากอัลกอริธึมที่ 7 และระบบจะทำการจัดเรียงโหนดใบตามลักษณะการปรากฏร่วมของคำภายในคลังข้อมูลภาษาเดี่ยวด้วยคำสั่ง OrderLexicalItems()

ผลลัพธ์ที่ได้จากกระบวนการนี้คือจะได้ข้อมูลต้นไม้การปรากฏร่วมซึ่งแสดงให้เห็นถึงการปรากฏร่วมกันภายในกลุ่มของข้อความในคลังข้อมูลภาษาเดี่ยวซึ่งจะได้ต้นไม้การปรากฏร่วมจำนวนมากเป็นป่า (forest) ต้นไม้การปรากฏร่วม

อัลกอริธึมที่ 2 แสดงกระบวนการเพิ่มโหนดลูกในต้นไม้การปรากฏร่วม (McTait, 2001: 56)

```
algorithm AddToChildrenColloc(Colloc,CollocNode) returns boolean
  for each child C of CollocNode
    Collocate(Colloc,C)
  SplitTree(Colloc,CollocNode)
  if any Colloc added to Children of CollocNode
    Return True
  else Return False
end algorithm
```

อัลกอริธึมนี้แสดงกระบวนการเพิ่มโหนดลูกในต้นไม้การปรากฏร่วม โดยกำหนดให้ Colloc เท่ากับคำที่ปรากฏร่วมพร้อมด้วยดัชนีบ่งชี้เลขบรรทัด และ Collocnode เท่ากับโหนดในต้นไม้การปรากฏร่วม และ C เท่ากับโหนดลูกของ Collocnode ค่าส่งคืนของกระบวนการนี้คือ จริง (true) หรือเท็จ (false)

กระบวนการนี้เริ่มจากการตรวจสอบความเป็นกลุ่มย่อยของการปรากฏร่วมกันของ Colloc และ C กระบวนการทำงานจะเริ่มจากการเพิ่มเติม Colloc ลงในโหนดลูกทุกตัวด้วยคำสั่ง Collocate()จากอัลกอริธึมที่ 3 หลังจากนั้นจะทำกระบวนการแยกโหนดในต้นไม้การปรากฏร่วมด้วยคำสั่ง SplitTree() จากอัลกอริธึมที่ 5

ผลจากกระบวนการนี้ ถ้ามีการเพิ่มคำลงโหนดลูกในต้นไม้การปรากฏร่วม ค่าส่งคืนของกระบวนการนี้คือจริง (true)

อัลกอริธึมที่ 3 แสดงกระบวนการเพิ่มเติมโหนดลูกลงในโหนดลูก (McTait, 2001: 56)

```
algorithm Collocate(Col1,Col2) returns boolean
  if Intersection(IDs(Col1),IDs(Col2)) ≥ 2
    if not AddToChildrenColloc(Col1,Col2) then
      AddChildColloc(CombineCollocs(Col1,Col2),Col2)
  else Return False
end algorithm
```

อัลกอริธึมนี้แสดงกระบวนการเพิ่มเติมโหนดลูกลงในโหนดลูก โดยกำหนดให้ Col1 เท่ากับคำที่ 1 และ Col2 เท่ากับคำที่ 2 ค่าส่งคืนของกระบวนการนี้คือ จริง (true) หรือเท็จ (false)

กระบวนการทำงานของอัลกอริธึมนี้คือการตรวจสอบว่า ถ้า Col1 และ Col2 ปรากฏร่วมกันในหมายเลขบรรทัดเดียวกันอย่างน้อย 2 ครั้ง ก็จะเรียกคำสั่ง AddToChildrenColloc() จากอัลกอริธึมที่ 2 เพื่อเพิ่ม Col1 เป็นโหนดลูกของ Col2 และเรียกซ้ำทำ

คำสั่งนี้ตามเงื่อนไขการปรากฏร่วมกันในหมายเลขบรรทัดเดียวกันอย่างน้อย 2 ครั้ง จนไม่สามารถเพิ่ม Col1 เป็นโหนดลูกของ Col2 ได้อีก หลังจากนั้นก็จะเรียกคำสั่ง AddChildColloc() เพื่อทำการสร้างโหนดลูกชุดใหม่ขึ้นมาโดยการเรียกซ้อนด้วยคำสั่งCombineCollocs() จากอัลกอริธึมที่ 4

ผลจากกระบวนการนี้ ถ้าไม่ตรงตามเงื่อนไขคือการปรากฏร่วมกันในหมายเลขบรรทัดเดียวกันอย่างน้อย 2 ครั้ง ก็จะได้ค่าส่งคืนเป็นเท็จ (false) ทันที

อัลกอริธึมที่ 4 แสดงกระบวนการย่อยของการรวมโหนด (McTait, 2001: 57)

```
algorithm CombineCollocs(Col1,Col2) returns NewColloc
  NewColloc = new Colloc()
  NewColloc.SetIds(Intersection(IDs(Col1),IDs(Col2))
  NewColloc.SetLexicalItems(LexItems(Col1), LexItems(Col2))
  return NewColloc
end algorithm
```

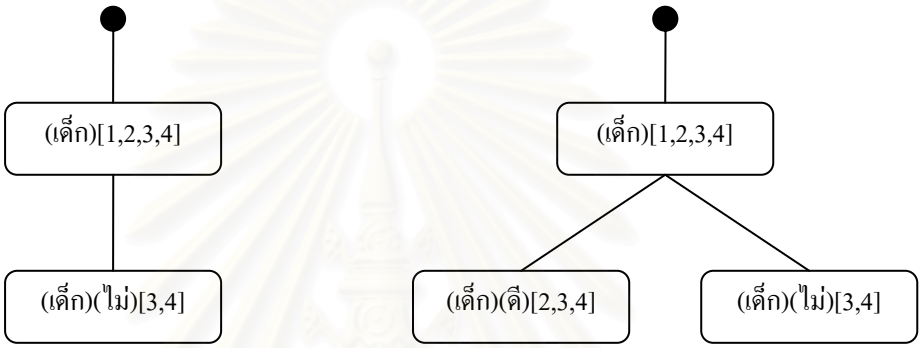
อัลกอริธึมนี้แสดงกระบวนการย่อยของการรวมโหนด Col2 กับโหนด Col1 โดยการสร้างโหนด Colloc ขึ้นใหม่และนำโหนด Col2 และโหนด Col1 มารวมกันอยู่ในโหนด colloc ที่สร้างขึ้นใหม่ เป็นกลุ่มคำใหม่ที่ถูกกำกับด้วยดัชนีบ่งชี้เลขบรรทัดนั้นโดยจะดึงเฉพาะเลขบรรทัดที่เกิดการปรากฏร่วมเท่านั้น เช่น มีชุดคำ (เด็ก)[1,2,3] (ดี)[2,3,4] กระบวนการนี้จะพยายามสร้างโหนดใหม่โดยรวม (เด็ก)(ดี)[2,3] เข้าด้วยกัน โดยค่าส่งคืนคือโหนดของกลุ่มคำที่รวมกัน และจะนำค่าส่งคืนนี้ไปเรียก AddChildColloc() เพื่อนำกลุ่มคำใหม่นี้ไปเป็นโหนดลูกของโหนดคำเดิมด้วย

อัลกอริธึมที่ 5 แสดงกระบวนการแยกโหนดในต้นไม้การปรากฏร่วม (McTait, 2001: 57)

```
algorithm SplitTree(Col1,Col2) returns void
  TempColloc = CombineCollocs(Col1,Col2)
  for each Child C of Col2
    if StrictSubsetP(IDs(C),IDs(TempColloc))
      SplitP = True
      AppendDaughters(TempColloc,C)
      DeleteChild(Col2,C)
  if SplitP == True
    AppendDaughters(Col2,TempColloc)
end algorithm
```

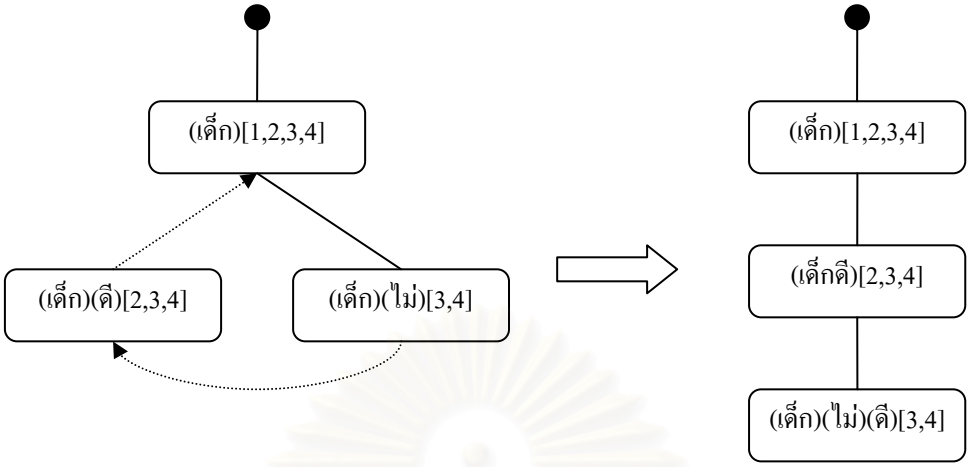
อัลกอริธึมนี้แสดงกระบวนการแยกโหนดในต้นไม้การปรากฏร่วม โดยกำหนดให้ TempColloc เท่ากับผลการทำงานของคำสั่ง CombineCollocs()จากอัลกอริธึมที่ 4 ค่าส่งคืนของกระบวนการนี้คือค่าว่าง (void)

กระบวนการทำงานของอัลกอริธึมนี้คือการปรับโหนดลูกของโหนด Col2 มาเป็นโหนดลูกของโหนด Col1 เพื่อการค้นหาและสกัดการปรากฏร่วมที่ยาวที่สุด (longest possible collocation) เช่น มีชุดคำ (เด็ก)[1,2,3,4] (ไม่)[3,4] (ดี)[2,3,4] จะได้การเรียงตัวของโหนดคือ (เด็ก)[1,2,3,4] เป็นโหนดแม่และไม่[3,4] เป็นโหนดลูกเป็นอันดับแรก ปัญหาที่พบคือระบบจะทำการเพิ่มโหนดลูก (ดี)[2,3,4] เป็นโหนดลูกที่ 2 ของ (เด็ก)[1,2,3,4] แต่ว่า (ไม่)[3,4] ก็เป็นส่วนซ้ำของ (ดี)[2,3,4] เช่นกัน ดังนั้นการทำงานของอัลกอริธึมนี้จะปรับปัญหาดังกล่าวนี้ให้เหลือเป็นการเรียงตัวของโหนดเพียงกิ่งเดียวคือ (เด็ก)[1,2,3,4] (ไม่)[3,4] (ดี)[2,3,4] ดังรูปที่ 12



รูปที่ 12 แสดงภาพต้นไม้การปรากฏร่วมที่ไม่ได้ทำกระบวนการแยกโหนด

กระบวนการแยกโหนดในต้นไม้การปรากฏร่วมจะตรวจสอบและพบว่าเมื่อ (ดี)[2,3,4] เป็น Col1 ที่ถูกเพิ่มเป็นโหนดลูกของ (เด็ก)[1,2,3,4] ซึ่งเป็น Col2 เมื่อการทำงานของกระบวนการ CombineCollocs() เสร็จจะได้ผลลัพธ์คือ (เด็ก)(ดี)[2,3,4] ซึ่งจะถูกเก็บเป็น TempColloc โหนดลูกทุกโหนดของ Col2 หรือในที่นี้คือ (เด็ก)[1,2,3,4] ว่าโหนดลูกใดเป็นเซตย่อย (subset) ของโหนดลูกอื่นหรือไม่ จากตัวอย่างด้านบน เด็กไม่[3,4] ที่เป็นโหนดลูกของโหนดแม่ (เด็ก)[1,2,3,4]นั้นเป็นเซตย่อย (subset) ของ (เด็ก)(ดี)[2,3,4] ที่เป็นโหนดลูกของโหนดแม่ (เด็ก)[1,2,3,4] เช่นกัน ระบบจะทำการลบ (delete) (เด็ก)(ไม่)[3,4] และสร้าง (เด็ก)(ไม่)(ดี)[3,4] เป็นโหนดลูกของ (เด็ก)(ดี)[2,3,4] แทนซึ่งจะทำให้ระบบสามารถสกัดการปรากฏร่วมที่ยาวที่สุดได้ดังรูปที่ 13 ด้านล่าง



รูปที่ 13 แสดงภาพต้นไม้การปรากฏร่วมที่ผ่านกระบวนการแยกโหนด

จากตัวอย่างด้านบนเป็นการค้นหาและสกัดการปรากฏร่วมที่ยาวที่สุดที่สั้นและง่าย ข้อมูลจริงที่นำมาสกัดการปรากฏร่วมที่ยาวที่สุดจะมีความซับซ้อนและมีโหนดลูกหลายโหนด เพราะระบบสามารถสร้างโหนดและเพิ่มโหนดลูกได้มากมายหากตรงตามเงื่อนไขมีการปรากฏร่วมซ้ำกันเกิน 2 ประโยค

อัลกอริธึมที่ 6 แสดงกระบวนการรวบรวมโหนดใบของต้นไม้การปรากฏร่วม (McTait, 2001: 60)

```
algorithm CollectLeaves( CollocList ) returns LeafNodeList
  LeafNodeList = new List()
  for each Colloc C in CollocList
    if HasDaughtersP(C)
      CollectLeaves(GetDaughters(C));
    else
      LeafNodeList.addElement(C);
  return LeafNodeList;
end algorithm
```

อัลกอริธึมนี้แสดงกระบวนการรวบรวมโหนดใบของต้นไม้การปรากฏร่วม โดยกำหนดให้ชุดของโหนดใบ (LeafNodeList) เท่ากับรายการโหนดว่างเปล่า (new List) และค่าส่งคืนคือชุดของโหนดใบ เนื่องจากโหนดใบของต้นไม้การปรากฏร่วมทุกโหนดจะเป็นส่วนที่ยาวที่สุดของการปรากฏร่วมและส่วนที่ยาวที่สุดนี้หากนำไปจับเทียบข้ามภาษาก็จะถูกนำไปใช้เป็นแม่แบบการแปล

กระบวนการนี้จะรวบรวมโหนดใบทุกใบในทุกต้นไม้การปรากฏร่วม โดยเริ่มจากการค้นหาชุดหรือกลุ่มคำที่ปรากฏร่วมที่เป็นโหนดลูกของโหนดราก collocList โดยทุก

ต้นไม้การปรากฏร่วมจะถูกไล่ค้นหาแบบเวียนซ้ำจนกว่าจะพบโหนดใบหรือโหนดปลายทางของทุกโหนดราก

อัลกอริธึมที่ 7 แสดงกระบวนการกรองโหนดใบ (McTait, 2001: 60)

```
algorithm Filter(LeafNodeList) returns FilteredLeafNodeList
  for each Colloc Ci in LeafNodeList
    for each Colloc Ci+1 in LeafNodeList
      if IDs(Ci) = IDs(Ci+1) AND Length(Ci) > Length(Ci+1)
        remove(Ci+1,LeafNodeList)
  return LeafNodeList
end algorithm
```

อัลกอริธึมนี้แสดงกระบวนการกรองโหนดใบเพื่อให้เหลือส่วนที่ยาวที่สุดของการปรากฏร่วมเท่านั้น โดยกำหนดให้ Ci คือกลุ่มคำของการปรากฏร่วมจากโหนดใบ และค่าส่งคืนคือชุดของโหนดใบ

สรุป ขั้นตอน 3.2.1 ผลของขั้นตอนภาษาเดี่ยวที่ได้จากกระบวนการทั้งหมดคือต้นไม้การปรากฏร่วมจากคลังข้อมูลแต่ละภาษา และส่วนที่จะนำไปใช้ต่อในขั้นตอนเทียบสองภาษาคือโหนดใบของกลุ่มคำของการปรากฏร่วมที่ยาวที่สุด ตัวอย่างของต้นไม้การปรากฏร่วมจะแสดงไว้ในรูปที่ 14 ซึ่งในรูปที่ 14 ขะเป็นตัวอย่างที่มีเพียง 2 โหนดรากคือ โหนดราก “ขอแจ้งว่าคณะกรรมการ” และโหนดราก “

```
_ROOT_
(ขอแจ้งว่าคณะกรรมการ)[0,1,2,3,4,5,6]
  (ขอแจ้งว่าคณะกรรมการ)(ตลาดหลักทรัพย์แห่งประเทศไทย)[0,1,2,3,4,5,6]
    (ขอแจ้งว่าคณะกรรมการ)(ตลาดหลักทรัพย์แห่งประเทศไทย)(ตั้งแต่)[0,1,2,3,4,5,6]
      (ขอแจ้งว่าคณะกรรมการ)(ตลาดหลักทรัพย์แห่งประเทศไทย)(ตั้งแต่)(เป็นต้นไป)[0,1,2,3,4,5,6]
        (ขอแจ้งว่าคณะกรรมการ)(ตลาดหลักทรัพย์แห่งประเทศไทย)(ตั้งแต่)(เป็นต้นไป)(เป็นหลักทรัพย์จดทะเบียน)
        [0,1,2,3,4,5,6]
          (ขอแจ้งว่าคณะกรรมการ)(ตลาดหลักทรัพย์แห่งประเทศไทย)(ตั้งแต่)(เป็นต้นไป)(เป็นหลักทรัพย์จดทะเบียน)
          (ได้สั่งรับให้หุ้นสามัญของ)[0,1,2,3,4,5,6]
(จัดสรรให้)[7,8,9,10,11]
  (จัดสรรให้)(หน่วย)[7,8,9,10,11]
    (จัดสรรให้)(หน่วย)(หุ้น)[7,8,9,10,11]
      (จัดสรรให้)(หน่วย)(หุ้น)(แปลงเป็นสามัญได้)[7,8,9,10,11]
        (จัดสรรให้)(หน่วย)(หุ้น)(แปลงเป็นสามัญได้)(ใบสำคัญแสดงสิทธิ)[7,8,9,10,11]
          (จัดสรรให้)(หน่วย)(หุ้น)(แปลงเป็นสามัญได้)(ใบสำคัญแสดงสิทธิ)(กรรมการและพนักงานจำนวน)[7,8,9,10,11]
```

รูปที่ 14 แสดงตัวอย่างส่วนหนึ่งของต้นไม้การปรากฏร่วม

3.2.2 ขั้นตอนการเทียบสองภาษา

ขั้นตอนการเทียบสองภาษาคือกระบวนการเทียบหาคู่คำแปลของกันและกันจากคลังข้อมูลเทียบบท โดยใช้ความสอดคล้องของการปรากฏจากแต่ละบรรทัดในคลังข้อมูลโดยแบ่งเป็น 2 ขั้นตอนคือ (1) ขั้นตอนการค้นหาส่วนคงที่และส่วนผันแปร และ (2) ขั้นตอนการจับคู่ส่วนคงที่เพื่อสร้างแม่แบบการแปล

3.2.2.1 ขั้นตอนการค้นหาส่วนคงที่และส่วนผันแปร

ในขั้นตอนนี้จะนำผลของขั้นตอนภาษาเดี่ยวมาใช้เพื่อทำการค้นหาส่วนซ้ำและส่วนไม่ซ้ำของข้อมูลจากคลังข้อมูลเพื่อตรวจสอบหาส่วนคงที่และส่วนผันแปรของแต่ละบรรทัด โดยกำหนดให้กลุ่มคำที่ปรากฏซ้ำๆ กันในข้อมูลตัวอย่างแต่ละบรรทัดเป็นส่วนคงที่ และส่วนที่ไม่ซ้ำกันเป็นส่วนผันแปร เนื่องจากคำหรือกลุ่มคำใดที่ปรากฏซ้ำๆ กันน่าจะเป็นส่วนที่เป็นแกนหลักสำคัญเพราะจะเป็นกลุ่มคำมีความเป็นไปได้ว่าจะเกิดขึ้นอีกในรายงานประเภทเดียวกันโดยเฉพาะอย่างยิ่งในภาษาเฉพาะทางซึ่งมีวงคำศัพท์ไม่มาก

การค้นหาส่วนคงที่และส่วนผันแปรสามารถตรวจสอบได้จากโหนดใบของต้นไม้การปรากฏร่วมที่เป็นผลลัพธ์ของขั้นตอนภาษาเดี่ยว โดยพิจารณาว่าคำที่ถูกเก็บอยู่ในโหนดใบเป็นส่วนคงที่และคำอื่นเป็นส่วนผันแปร เช่น โหนดใบมีข้อมูลคือ “(ตลาดหลักทรัพย์) (เพิ่ม) (สินค้า) (:) [1, 12]” และเมื่อตรวจสอบกลับไปที่ข้อความต้นฉบับจากดัชนีบ่งชี้เลขบรรทัดและพบ“ตลาดหลักทรัพย์ เพิ่ม สินค้า : AKR [1]” จากบรรทัดที่ 1 และ “ตลาดหลักทรัพย์ เพิ่ม สินค้า : BNT [12]” จากบรรทัดที่ 12 ทำให้สามารถกำหนดได้ว่า “(ตลาดหลักทรัพย์) (เพิ่ม) (สินค้า) (:)” เป็นส่วนคงที่ โดยที่ “(AKR)” และ “(BNT)” เป็นส่วนผันแปร ดังนั้นจึงกล่าวได้ว่าโหนดใบทุกโหนดเป็นส่วนคงที่ และส่วนผันแปรก็จะถูกแทนค่าด้วยตัวแปร ตัวแปรที่ใช้ในภาษาต้นฉบับกำหนดให้ใช้ตัวแปร “<X>” และตัวแปรที่ใช้ในภาษาเป้าหมาย กำหนดให้ใช้ตัวแปร “<Y>”

อย่างไรก็ตามส่วนผันแปรในแต่ละโหนดใบอาจจะมีมากกว่า 1 ตัว ดังนั้นตัวแปรทุกตัวจึงต้องมีหมายเลขกำกับตามตำแหน่งจากทางซ้ายไปขวา โดยหมายเลขกำกับจะเริ่มต้นที่ 0 หากตัวแปรปรากฏในตำแหน่งก่อนส่วนคงที่ แต่หากตัวแปรปรากฏในตำแหน่งหลังส่วนคงที่หมายเลขกำกับจะเริ่มต้นที่ 1 หลังจากนั้นจึงจะไล่ตัวเลขเรียงไปจนครบทุกตัวแปรโดยไม่อนุญาตให้มีตัวเลขซ้ำกัน เช่น “ตลาดหลักทรัพย์แห่งประเทศไทย ขอแจ้งว่าคณะกรรมการ ตลาดหลักทรัพย์แห่งประเทศไทย ได้สั่งให้รับหุ้นสามัญของ <X1> เป็น <X2> ในตลาดหลักทรัพย์ตั้งแต่ <X3> เป็นต้นไป” จากตัวอย่างข้างต้นจะพบว่าตัวแปรเริ่มจาก <X1> เพราะมีส่วนคงที่ “ตลาดหลักทรัพย์แห่งประเทศไทย ขอแจ้งว่าคณะกรรมการ ตลาดหลักทรัพย์แห่งประเทศไทย ได้สั่งให้รับ

หุ้นสามัญของ” ปรากฏอยู่หน้าตัวแปรแรก และจะไล่ตัวเลข <X2> <X3> ตามลำดับเพื่อบ่งบอกตำแหน่งการปรากฏของส่วนผันแปร ส่วนคงที่ที่มีการกำกับหลายเลขตำแหน่งของส่วนผันแปรนี้จะถูกเรียกว่าแม่แบบภาษาเดี่ยวและจะถูกนำไปค้นหาส่วนคงที่ในอีกภาษาเพื่อเป็นคู่คำแปลต่อไป

3.2.2.2 ขั้นตอนการจับคู่ส่วนคงที่เพื่อสร้างแม่แบบการแปล

ขั้นตอนการจับคู่ส่วนคงที่เพื่อสร้างแม่แบบการแปลคือการค้นหาว่าแม่แบบการแปลภาษาเดี่ยวใดเป็นคู่คำแปลของกันและกัน เพื่อจะนำไปสร้างเป็นแม่แบบการแปลต่อไป

ผลลัพธ์จากขั้นตอน 2.1 คือได้แม่แบบภาษาเดี่ยวจากโหนดใบของต้นไม้การปรากฏร่วม การจับคู่แม่แบบภาษาเดี่ยวต้องใช้ความสอดคล้องของการปรากฏจากแต่ละบรรทัดในคลังข้อมูล และทุกครั้งที่แม่แบบภาษาเดี่ยวปรากฏสอดคล้อง กลุ่มของเลขบรรทัดทั้งภาษาต้นทาง (IDs(CollocSL)) และภาษาปลายทาง (IDs(CollocTL)) ต้องตรงกันทั้งหมด ซึ่งมีการคำนวณดังสมการที่ 1

$$IDs(Colloc_{SL}) = IDs(Colloc_{TL}) \quad (1)$$

เช่น “(ตลาดหลักทรัพย์) (เพิ่ม) (สินค้า) (:) <X1> [1, 12, 36, 44]” แม่แบบภาษาเดี่ยวนี้ปรากฏในบรรทัดที่ 1 บรรทัดที่ 12 บรรทัดที่ 36 และบรรทัดที่ 44 จากคลังข้อมูลภาษาไทย และ “(SET) (adds) (new) (listed) (securities) (:) <X1> [1, 12, 36, 44]” แม่แบบภาษาเดี่ยวนี้ปรากฏในบรรทัดที่ 1 บรรทัดที่ 12 บรรทัดที่ 36 และบรรทัดที่ 44 จากคลังข้อมูลภาษาอังกฤษ ดังนั้น กลุ่มคำนี้จะถูกจับเทียบเป็นคู่คำแปลของกันและกัน เป็น “(ตลาดหลักทรัพย์) (เพิ่ม) (สินค้า) (:) <X1>” ↔ “(SET) (adds) (new) (listed) (securities) (:)<X1>”

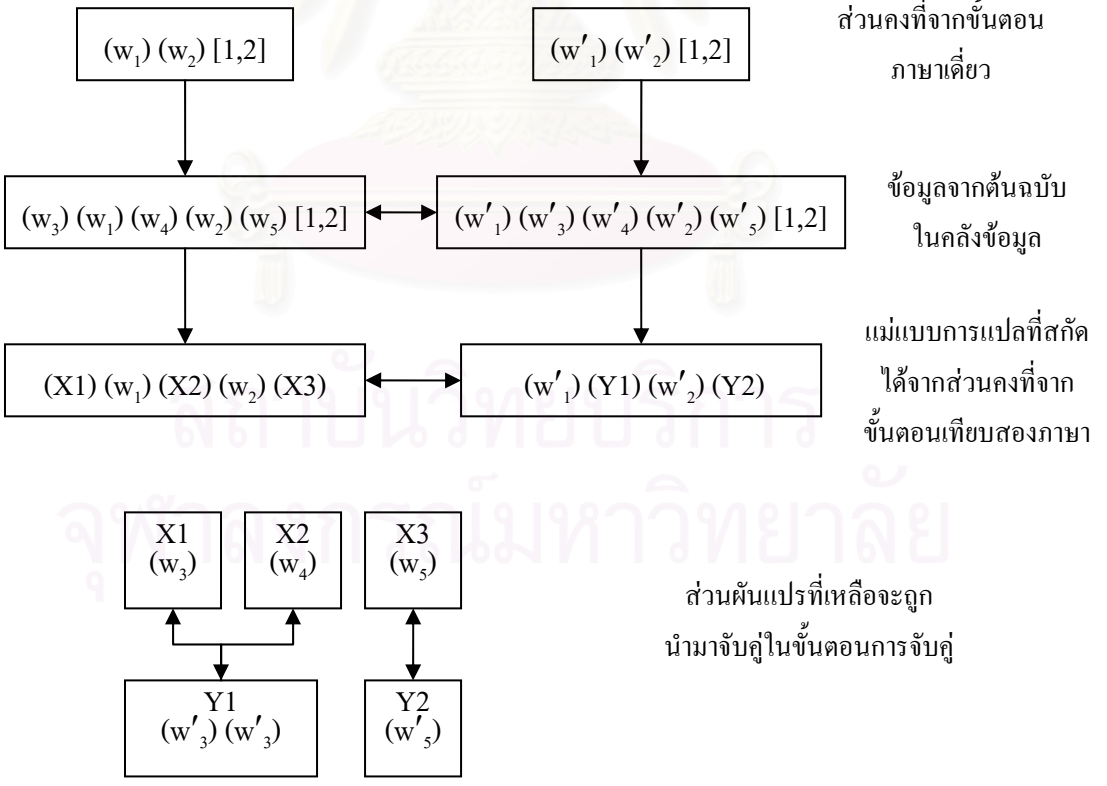
การทำงานในขั้นตอนนี้จะทำให้ได้แม่แบบการแปลของส่วนคงที่ซึ่งจะเป็นส่วนหลักในการเปรียบเทียบหาคำแปลในงานวิจัยชิ้นนี้

อย่างไรก็ตามขั้นตอนเทียบสองภาษาจะสามารถสกัดแม่แบบการแปลออกมาได้แค่เพียงส่วนหนึ่งเท่านั้น และยังไม่เพียงพอต่อการนำไปใช้แปลข้อความรับเข้า ดังนั้นจึงต้องนำส่วนผันแปรไปจับคู่เทียบแปลเพื่อเพิ่มวงคู่คำศัพท์ให้กับระบบซึ่งจะเป็นการทำงานในขั้นตอนถัดไป

3.2.3 ขั้นตอนการจับคู่ส่วนผันแปร

ขั้นตอนนี้คือการนำส่วนผันแปรมาจับคู่เทียบแปลเพื่อเพิ่มวงคู่คำศัพท์ให้กับระบบ โดยจะนำส่วนผันแปรมาจับคู่ที่เป็นคำแปลของกันและกัน อย่างไรก็ตามส่วนผันแปรบางกรณี อาจมีจำนวนไม่เท่ากัน เนื่องจากส่วนผันแปร 1 ส่วนอาจจะมีคู่คำแปลมากกว่า 1 ก็ได้เช่น ส่วนผันแปร 2 ส่วนมีคู่คำแปล 1 ส่วนดังรูปที่ 15

ภายหลังจากที่สามารถระบุตำแหน่งส่วนผันแปรในแม่แบบการแปลได้โดยการใช้การปรากฏร่วมของส่วนข้อความที่สกัดออกมาได้แล้ว ส่วนผันแปรต่างๆ จะถูกจับคู่กันระหว่างสองภาษาเพื่อระบุความสอดคล้องระหว่างกัน ส่วนผันแปรแต่ละตัวจะถูกทดลองจับคู่ในทุกความเป็นไปได้โดยใช้อัลกอริธึมการเปรียบเทียบลำดับ (Sequence Comparison Algorithm) โดยใช้เมตริกความคล้ายคลึงของคู่ภาษา (Bilingual Similarity Metric / BS) ในการตัดสินใจเลือกการจับคู่ส่วนผันแปรที่เหมาะสมที่สุด หลังจากกระบวนการนี้ จะได้แม่แบบการแปลที่ระบุความสอดคล้องระหว่างส่วนผันแปรในคู่ภาษาออกมา โดยส่วนผันแปรจะเป็นตัวแปรที่มีค่าว่าง ซึ่งสามารถนำไปใช้ในกระบวนการแปลโดยใช้แม่แบบการแปลในภายหลังได้



รูปที่ 15 แสดงส่วนที่เหลือจากขั้นตอนเทียบสองภาษาและการจับคู่ส่วนผันแปร

ภายในขั้นตอนการจับคู่ส่วนผันแปรมีกระบวนการทำงานหลักอยู่ 2 กระบวนการคือ (1) กระบวนการคำนวณเมตริกความคล้ายคลึงของคู่ภาษา และ (2) อัลกอริธึมการเปรียบเทียบลำดับ

3.2.3.1 กระบวนการคำนวณเมตริกความคล้ายคลึงของคู่ภาษา

กระบวนการคำนวณเมตริกความคล้ายคลึงของคู่ภาษาคือการคำนวณค่าคะแนนจากความคล้ายคลึงของคำในส่วนผันแปร เพื่อใช้ในการตัดสินใจเลือกการจับคู่ส่วนผันแปรที่เหมาะสมที่สุดในอัลกอริธึมการเปรียบเทียบลำดับ การคำนวณความคล้ายคลึงของคู่ภาษาในงานของแม็คเทท (McTait, 2001:64) ประกอบด้วยปัจจัย 2 ส่วน คือ คะแนนคำร่วมเชื้อสาย (cognate) และคะแนนการกระจายตัวของคำในคู่ภาษา (bilingual lexical distribution/BLD) โดยมีสมการในการคำนวณคือ

$$BS = \frac{BLD + |Cognates|}{1 + |Cognates|} \tag{2}$$

คำร่วมเชื้อสายคือคำที่มาจากรากศัพท์เดียวกันแต่มีการผันแปรไปตามภาษาต่างกันไป ซึ่งจะพบในภาษาตระกูลเดียวกัน เช่น ภาษาตระกูลอินโดยูโรเปียนที่เป็นต้นกำเนิดของภาษาทางฝั่งประเทศยุโรป อาทิ ภาษาอังกฤษ ภาษาฝรั่งเศส ภาษาสเปน เป็นต้น งานของแม็คเททเป็นการแปลภาษาในตระกูลเดียวกันจึงมีคำร่วมเชื้อสายปรากฏอยู่ในคลังข้อมูลเป็นจำนวนมาก ทำให้เมตริกความคล้ายคลึงต้องคำนวณโดยอาศัยคำร่วมเชื้อสายเป็นหลักตามสมการที่ (2)

อย่างไรก็ตามในงานวิจัยชิ้นนี้ที่วิจัยการแปลภาษาไทยเป็นภาษาอังกฤษ ซึ่งคู่ภาษานี้เป็นภาษาต่างตระกูลจึงไม่สามารถนำคะแนนคำร่วมเชื้อสายมาเป็นตัวแปรในการคำนวณคะแนนได้ จึงต้องตัดคะแนนส่วนนี้ออกไปและคงไว้แต่ส่วนคะแนนการกระจายตัวของคำในคู่ภาษาดังนั้นสมการในการคำนวณเมตริกความคล้ายคลึงของคู่ภาษา จึงได้รับการปรับเปลี่ยนดังนี้ เนื่องจากต้องตัดคะแนนคำร่วมเชื้อสาย จึงทำให้ $|Cognates| = 0$ สมการที่ (2) จึงลดรูปเป็นดังสมการที่ (3)

$$\begin{aligned} BS &= \frac{BLD}{1} \\ \therefore BS &= BLD \end{aligned} \tag{3}$$

ดังนั้นในงานวิจัยชิ้นนี้ คะแนนกระบวนการคำนวณเมตริกความคล้ายคลึงของคู่ภาษาจึงเท่ากับคะแนนการกระจายตัวของคำในคู่ภาษา การคำนวณคะแนนการ

กระจายตัวของคำในคู่ภาษาจึงเป็นส่วนที่สำคัญมากในงานวิจัยชิ้นนี้โดยมีสมการสัมประสิทธิ์ของไดซ์ (Dice's co-efficient) (Dice, 1945) คือ

$$WBLD(S,T) = \frac{2(|S \cap T|)}{|S| + |T|} \tag{4}$$

การคำนวณคะแนนการกระจายตัวของคำในคู่ภาษา (WBLD) จะดูจากความสอดคล้องของเลขบรรทัดที่กลุ่มคำดังกล่าวปรากฏ โดยมีสมมติฐานว่า คำที่ปรากฏอยู่ในคู่บรรทัดเดียวกันมีแนวโน้มว่าจะเป็นคู่คำแปลของกันและกันถ้าคำนั้นปรากฏในทุกคู่บรรทัดดังกล่าว สมการสัมประสิทธิ์ของไดซ์มีหลักการคำนวณคือ จะนำจำนวนบรรทัดที่คำคู่ดังกล่าวปรากฏ มาคูณสองและหารด้วยจำนวนบรรทัดที่คำของภาษาต้นฉบับหรือคำของภาษาเป้าหมายปรากฏอยู่ ผลที่ได้คือค่าความน่าจะเป็นของการเป็นคู่คำแปลซึ่งกันและกัน (WBLD)

อย่างไรก็ตามสมการสัมประสิทธิ์ของไดซ์ สามารถหาความสัมพันธ์ได้แค่ในระดับคำเดี่ยว แต่ในภาษาที่พบมีการปรากฏของกลุ่มคำอยู่ จึงต้องนำสมการสัมประสิทธิ์ของไดซ์มาพัฒนาต่อเป็นสมการดังนี้

$$BS(S,T) = BLD(S,T) = \frac{2\sum_{w_S \in S} \arg\max_{w_T \in T} WBLD(w_S, w_T)}{|S| + |T|} \tag{5}$$

สมการนี้จะช่วยหากลุ่มคำและค่าความสัมพันธ์ของความสอดคล้องของเลขบรรทัดซึ่งสามารถเขียนเป็นอัลกอริธึมได้ดังนี้

อัลกอริธึมที่ 8 แสดงกระบวนการคำนวณคะแนนการกระจายตัวของกลุ่มคำในคู่ภาษา

```
algorithm bld(SLFragment, TLFagment)
  for each SL Word W_S ∈ SLFragment
    for each TL Word W_T ∈ TLFragment
      add Dice(W_S,W_T) to DistributionScores list
    add Maximum(DistributionScores) to MaxScores list
return 2*(Σ MaxScores/(Length(SLFragment)+Length(TLFragment)))
end algorithm
```

อัลกอริธึมที่ 8 แสดงวิธีการคำนวณคะแนนการกระจายตัวของกลุ่มคำในคู่ภาษาของส่วนข้อความที่กำหนดให้ (SLFragment และ TLFragment) กระบวนการคำนวณจะเป็นดังนี้ สำหรับคำหนึ่งๆ ในส่วนข้อความของภาษาต้นฉบับ ค่าสัมประสิทธิ์ของไดซ์จะถูกคำนวณเทียบกับคำทุกคำในส่วนข้อความของภาษาเป้าหมาย เพื่อหาค่าสัมประสิทธิ์ที่มีค่ามากที่สุด จากนั้น

จึงนำค่าสัมประสิทธิ์ที่มีค่ามากที่สุดของแต่ละคำในส่วนข้อความของภาษาต้นฉบับมารวมกัน คูณด้วยสอง หารด้วยความยาวรวมของส่วนข้อความทั้งภาษาต้นฉบับและภาษาปลายทาง

เมื่อได้คะแนนการกระจายตัวของกลุ่มคำแล้ว คะแนนนี้จะถูกนำไปใช้ในการตัดสินใจเลือกวิธีการจับคู่ส่วนผันแปรที่ดีที่สุดในอัลกอริธึมเปรียบเทียบลำดับในขั้นตอนต่อไป

3.2.3.2 อัลกอริธึมการเปรียบเทียบลำดับ

กระบวนการนี้มีวัตถุประสงค์เพื่อจับคู่ส่วนผันแปรให้เป็นคู่คำแปลกัน เพื่อสร้างแม่แบบการแปลส่วนผันแปร โดยกระบวนการนี้จะเลือกส่วนผันแปรแต่ละตัว จากหลายๆ ตัวในข้อความของในแต่ละภาษามาจับคู่กัน

จากกระบวนการที่แล้ว ค่าคะแนนการกระจายตัวของกลุ่มคำในคู่ภาษาจะเป็นตัวบ่งบอกว่าคำใดหรือกลุ่มคำใดเป็นคู่คำแปลกับคำใด ซึ่งการจับคู่อาจเกิดแบบ 1 คำต่อ 1 คำ หรือ 1 คำต่อ 2 คำ หรือ 2คำต่อ 1 คำก็ได้ การจับคู่ในส่วนนี้จึงต้องใช้อัลกอริธึมโปรแกรมพลวัติ (Dynamic Programming Algorithm) กับการจับคู่ลำดับ (sequence alignment) เพื่อตรวจสอบและจับคู่ส่วนผันแปร โดยนิยามการจับคู่ลำดับไว้ดังนี้

ถ้าหากมีลำดับของส่วนข้อความ 2 ตัวที่สอดคล้องกัน คือ x และ y ซึ่งมีความยาว m และ n ตามลำดับ ลำดับทั้งสองสามารถจับคู่กันได้โดยใช้อัลกอริธึมโปรแกรมพลวัติ ดังสมการที่ 6

$$D(i,j)=\min\begin{cases} D(i,j-1)+d(\varnothing,\{y_j\}) & \ldots\textbf{(1)}\\ D(i-1,j)+d(\{x_i\},\varnothing) & \ldots\textbf{(2)}\\ D(i-1,j-1)+d(\{x_i\},\{y_j\}) & \ldots\textbf{(3)}\\ D(i-1,j-2)+d(\{x_i\},\{y_{j-1},y_j\}) & \ldots\textbf{(4)}\\ D(i-1,j-3)+d(\{x_i\},\{y_{j-2},y_{j-1},y_i\}) & \ldots\textbf{(5)}\\ D(i-2,j-1)+d(\{x_{i-1},x_i\},\{y_j\}) & \ldots\textbf{(6)}\\ D(i-3,j-1)+d(\{x_{i-2},x_{i-1},x_i\},\{y_j\}) & \ldots\textbf{(7)}\\ D(i-2,j-2)+d(\{x_{i-1}\},\{y_j\})+d(\{x_i\},\{y_{j-1}\}) & \ldots\textbf{(8)} \end{cases} \quad (6)$$

เมื่อ x_i แทนส่วนข้อความหนึ่งในภาษาต้นฉบับ และ y_j แทนส่วนข้อความหนึ่งในภาษาเป้าหมาย ฟังก์ชัน $D(i,j)$ จะเป็นคะแนนการแก้ไขน้อยที่สุด (minimum edit distance) ระหว่างลำดับของส่วนข้อความ $\langle x_1,x_2,x_3,\ldots,x_i\rangle$ และลำดับส่วนข้อความที่เป็นคำแปล $\langle y_1,y_2,y_3,\ldots,y_j\rangle$ เมื่อ $1\le i\le m$ และ $1\le j\le n$ ดังนั้นคะแนนขั้นตอนการแก้ไขสั้นสุดของ

ลำดับ x และ y จึงเป็น $D(m,n)$ ฟังก์ชัน $d(\cdot)$ เป็นฟังก์ชันความห่าง ซึ่งที่จริงแล้วก็คือส่วนกลับของมาตรวัดความคล้ายในคู่ภาษา (bilingual similarity metric) กล่าวคือ $d(x_i, y_j) = \frac{1}{BS(x_i, y_j)}$ ซึ่งการคำนวณคะแนนความคล้ายในคู่ภาษา ได้อธิบายไว้ในหัวข้อ 3.1

ลำดับของตัวเลือกในสมการที่ 6 จะสอดคล้องกับความสัมพันธ์แบบ (1) การแทรก มีความสัมพันธ์แบบ 0 ต่อ 1 (2) การลบ 1 ต่อ 0 (3) การแทนที่ 1 ต่อ 1 (4-5) การขยาย 1 ต่อ 2 และ 1 ต่อ 3 (6-7) การบีบอัด 2 ต่อ 1 และ 3 ต่อ 1 และ (8) การสลับตำแหน่งตัวที่ประชิดกัน (Lowrance และ Wagner, 1975)

นอกจากนี้ผู้วิจัยยังได้กำหนดเพิ่มเติมว่า คู่ของกลุ่มคำใดมีรูปผิว (surface form) ตรงกันทุกประการ คะแนน $d(x_i, y_j)$ จะเป็นศูนย์ ซึ่งหมายความว่าเป็นคู่คำแปลของกันและกันโดยสมบูรณ์ เพื่อลดจำนวนตัวเลือกในการจับคู่ส่วนผันแปรที่มีส่วนข้อความตรงกันทุกประการ ดังนั้นสมการของฟังก์ชันความห่าง จึงเป็นดังสมการที่ 7

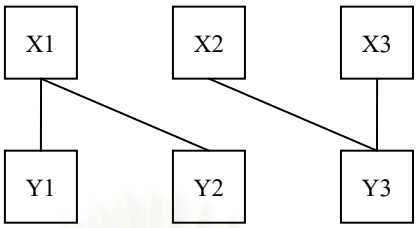
$$d(x_i, y_j) = \begin{cases} \dfrac{1}{BS(x_i, y_j)} & , x_i \neq y_j \\ 0 & , x_i = y_j \end{cases} \tag{7}$$

ภายในอัลกอริธึมโปรแกรมพลวัติ ค่าของ $D(i,j)$ จะบรรจุภายในอยู่ในแมทริกส์ขนาด $m \times n$ โดยที่ขั้นตอนเริ่มต้น (initialization step) แถวและคอลัมน์ที่ 0 ของแมทริกส์ ซึ่งก็คือค่าของ $D(0,j)$ และ $D(i,0)$ ตามลำดับ จะมีค่าเท่ากับผลรวมของคะแนนการแทรกส่วนข้อความ i ตัวแรกของ x และผลรวมของคะแนนการลบส่วนข้อความ j ตัวแรกของ y ตามลำดับ ขั้นตอนเริ่มต้นสามารถเขียนเป็นสูตรได้ดังสมการที่ 8

$$\begin{aligned} D(0,0) &= 0 \\ D(0,j) &= \sum_{k=1}^{j} d(0, y_k), \text{ for } 1 \leq j \leq n \\ D(i,0) &= \sum_{k=1}^{i} d(x_k, 0), \text{ for } 1 \leq i \leq m \end{aligned} \tag{8}$$

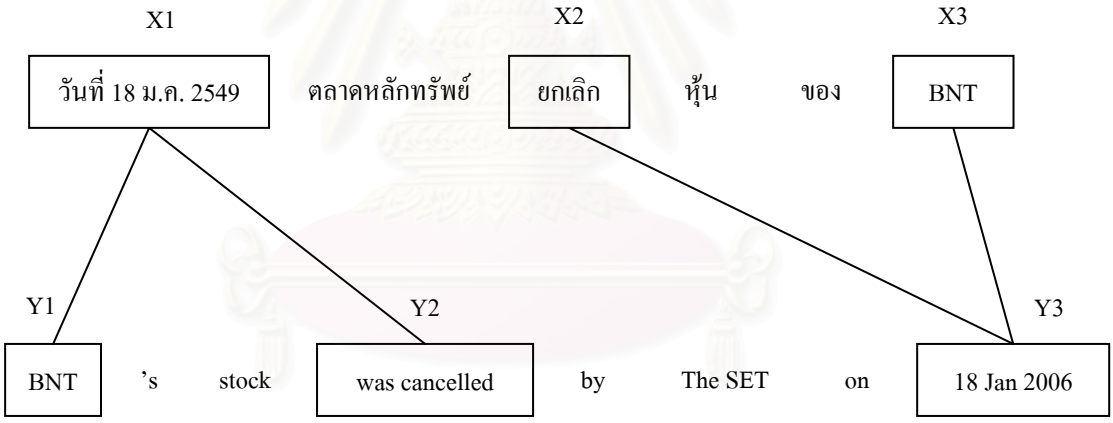
อัลกอริธึมนี้จะคำนวณคะแนนและเลือกจับคู่คำหรือกลุ่มคำที่มีคะแนนสูงสุด อย่างไรก็ตามอัลกอริธึมโปรแกรมพลวัติสามารถคำนวณการจับคู่ตัวแปรแบบประชิด (adjacent variable alignment) ของส่วนผันแปรเท่านั้น เพราะอัลกอริธึมโปรแกรมพลวัติสามารถคำนวณจับคู่คำที่ไม่ซ้อนเหลื่อมเท่านั้น ดังรูปที่ 16 ด้านล่าง ที่เห็นว่า ตัวแปร X1 ไม่สามารถจับคู่

ข้ามไปหา ตัวแปร Y3 ได้ แต่จะสามารถจับคู่ตัวแปรได้เพียงแค่ ตัวแปร Y1 และ Y2 เท่านั้น เพราะผลลัพธ์จากอัลกอริธึมโปรแกรมพลวัติไม่สามารถจับคู่ตัวแปรข้ามทับกันได้



รูปที่ 16 แสดงการจับคู่ตัวแปรแบบประชิดของส่วนผันแปร

ดังนั้นการจับคู่ตัวแปรแบบประชิดของส่วนผันแปรอาจทำให้จับคู่ผิดพลาดดังเช่นรูปที่ 17 กล่าวคือ ตัวแปร X1 จะถูกจับคู่ได้แค่กับ Y1 และ Y2 เท่านั้น โดยที่จะไม่สามารถจับคู่ข้ามไปหาตัวแปร Y3 ซึ่งเป็นตัวแปรที่เป็นคู่คำแปลของกันเพราะข้อจำกัดของอัลกอริธึมโปรแกรมพลวัติ



รูปที่ 17 แสดงปัญหาการจับคู่ตัวแปรแบบประชิดของส่วนผันแปรที่ไม่สามารถซ้อนเหลื่อม

ดังนั้นระบบจึงจำเป็นต้องมีกระบวนการในการจับคู่ตัวแปรแบบไม่ประชิด (non-adjacent variable adjacent) ด้วย โดยใช้อัลกอริธึมที่ 9 เพื่อคำนวณตัวเลือกของการจับคู่ตัวแปรแบบไม่ประชิด

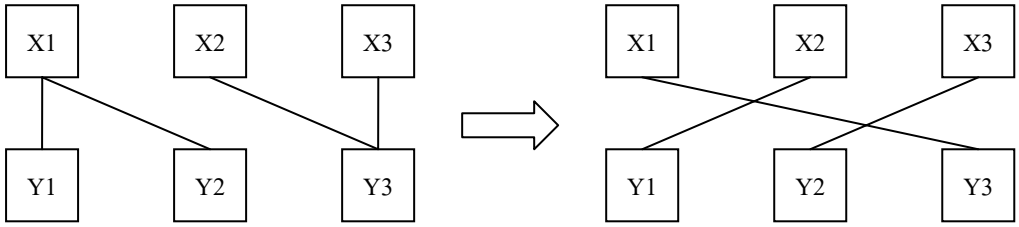
อัลกอริธึมที่ 9 แสดงการคำนวณตัวเลือกของการจับคู่ตัวแปรแบบไม่ประชิด

```
algorithm to compute candidate non-adjacent alignment
  for i = 1 to m
    for j = 1 to n
      if Score(Xi, Yj) > Threshold & abs(i-j) > 1
         add {Xi, Yj, Score(Xi,Yj)} to list of candidate non-
adjacent alignments
end algorithm
```

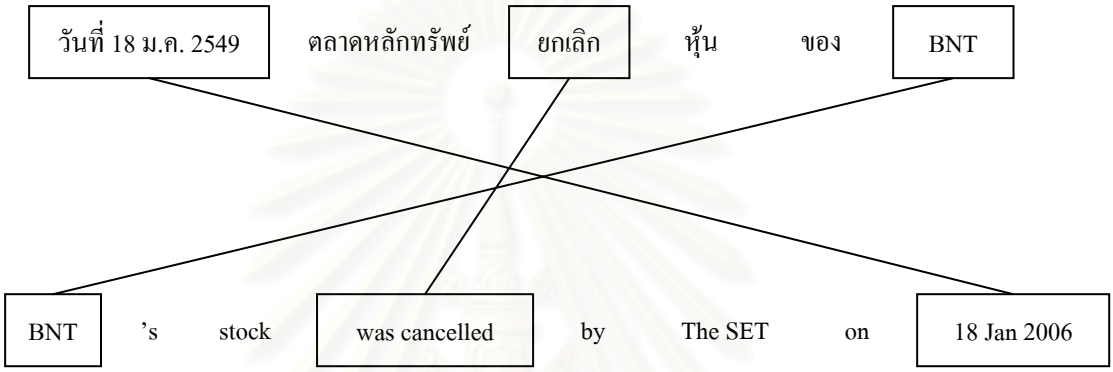
อัลกอริธึมจะค้นหาตัวเลือกที่เป็นไปได้ของการจับคู่ตัวแปรแบบไม่ประชิด โดยกำหนดให้ i แทนดัชนีของตัวแปรในภาษาต้นฉบับ j แทนดัชนีของตัวแปรในภาษาเป้าหมาย m คือจำนวนตัวแปรทั้งหมดของภาษาต้นฉบับ และ n คือจำนวนตัวแปรทั้งหมดของภาษาเป้าหมาย

อัลกอริธึมที่ 9 มีขั้นตอนในการค้นหาตัวแปรแบบไม่ประชิดคือ (1) สร้างรายการของตัวเลือกที่สามารถเป็นคู่คำแปลของกันได้ โดยต้องมีคะแนนเมตริกความคล้ายคลึงของคู่ภาษาตั้งแต่ค่าเริ่มแรก (threshold) ที่กำหนดเป็นต้นไปและระยะทางจาก i และ j ต้องมีตัวแปรอื่นมาคั่นอย่างน้อย 1 ตัว (2) คัดตัวเลือกที่สามารถเป็นคู่คำแปลออกจากรายการด้วยเงื่อนไข 2 ข้อ เงื่อนไขแรกคือเมื่อ X_i เป็นสมาชิกของ A และ Y_j เป็นสมาชิกของ B ถ้าเคยมีการจับคู่กลุ่มตัวแปร $< A, B >$ จากการจับคู่ตัวแปรแบบประชิดไปแล้ว จะไม่มีการจับคู่ $< X_i, Y_j >$ ในส่วนการจับคู่ตัวแปรแบบไม่ประชิดอีก และเงื่อนไขที่สองคือเมื่อ X_i เป็นสมาชิกของ A_p และ Y_j ไม่เป็นสมาชิกของ B_p และ X_i ไม่เป็นสมาชิกของ A_q และ Y_j เป็นสมาชิกของ B_q ถ้าการจับคู่ตัวแปร $< X_i, Y_j >$ ใดๆ ในส่วนการจับคู่ตัวแปรแบบไม่ประชิดต้องมีค่าคะแนนเมตริกความคล้ายคลึงของคู่ภาษาของ $< X_i, Y_j >$ มากกว่าค่าคะแนนเมตริกความคล้ายคลึงของคู่ภาษาของทุกกลุ่มตัวแปร $< A_p, B_p >$ และ $< A_q, B_q >$ อื่นๆ และ (3) ตรวจสอบผลลัพธ์ซ้ำโดยคัดตัวเลือกที่สามารถเป็นคู่คำแปลออกจากรายการด้วยเงื่อนไขคือถ้าเลือก $< X_i, Y_j >$ เป็นคู่คำแปลไปแล้ว $< X_i', Y_j' >$ อื่นๆ ที่มี $X_i = X_i'$ หรือ $Y_j = Y_j'$ จะถูกเลือกอีกไม่ได้

ผลของอัลกอริธึมที่ 9 ทำให้ระบบสามารถจับคู่ตัวแปรที่ไม่ได้ปรากฏประชิดติดกัน ได้ ทำให้สามารถจับคู่ตัวแปรได้หลากหลายขึ้น ดังรูปที่ 18 ที่ระบบสามารถจับคู่ตัวแปร X1 กับตัวแปร Y3 ได้ ทำให้สามารถจับคู่ข้อความที่เป็นส่วนผันแปรได้อย่างถูกต้องดังเช่นรูปที่ 19



รูปที่ 18 แสดงการจับคู่หลังผ่านกระบวนการจับคู่ตัวแปรแบบไม่ประชิดที่หลากหลายขึ้น



รูปที่ 19 แสดงการจับคู่หลังผ่านกระบวนการจับคู่ตัวแปรแบบไม่ประชิด

เมื่อผ่านกระบวนการจับคู่ตัวแปรแบบไม่ประชิดแล้วก็จะนำส่วนที่คัดออกตามเงื่อนไขมาทำการจับคู่ตัวแปรแบบประชิดอีกครั้งเพื่อให้ทุกคำหรือกลุ่มคำได้รับการจับคู่คำแปล ซึ่งผลของการทำกระบวนการจับคู่ทั้งหมดทำให้สามารถจับคู่ได้หลากหลายรูปแบบมากขึ้น และได้แม่แบบการแปลส่วนผันแปรที่มีประสิทธิภาพ

แม่แบบการแปลที่ผ่านการจับคู่ตัวแปรมาแล้ว สามารถนำมาขยายความครอบคลุมฐานข้อมูลแม่แบบการแปลได้ โดยการกลับตัวแปรไปเป็นข้อความส่วนคงที่ และกลับข้อความส่วนคงที่ไปเป็นตัวแปร จากนั้นก็ใช้กรรมวิธีการจับคู่ตัวแปรทั้งแบบประชิดและไม่ประชิดอีกครั้ง ซึ่งจะทำให้ได้แม่แบบการแปลเพิ่มขึ้น

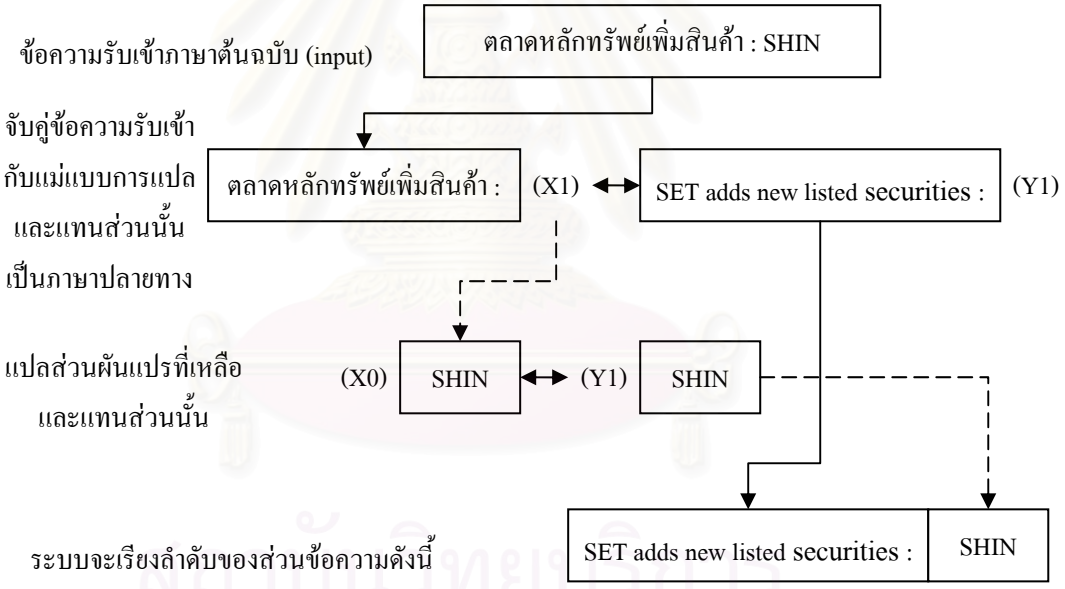
สรุปภายในขั้นตอน 3.2 ทั้งหมดคือจะได้แม่แบบการแปลทั้งจากส่วนคงที่และส่วนผันแปร ซึ่งแม่บบการแปลจะมีลักษณะดังนี้คือ [ตลาดหลักทรัพย์ <X1> สินค้า : <X2> ↔ The SET <Y1> listed securities : <Y2>] (โดยที่ตัวแปร X1 เป็นคู่คำแปลกับตัวแปร Y1 และตัวแปร X2 เป็นคู่คำแปลกับตัวแปร Y2) แม่แบบการแปลที่ได้ก็จะพร้อมนำไปใช้ในการเปรียบเทียบหาคำแปลข้อความรับเข้าต่อไป นอกจากนี้ยังเก็บข้อความทั้งข้อความเป็นแม่แบบการแปลด้วย เพื่อช่วยสำหรับกรณีที่ข้อความรับเข้าเหมือนกับข้อความเดิมที่ได้ผ่านการสกัดแม่แบบการแปลทุกประการก็

จะสามารถดึงข้อความต้นแบบมาเป็นแม่แบบการแปลและเปรียบเทียบแปลได้ทันทีเพื่อเป็นการประหยัดเวลาในการแปลข้อความอีกด้วย

3.3 ระบบการรวมคำแปลใหม่

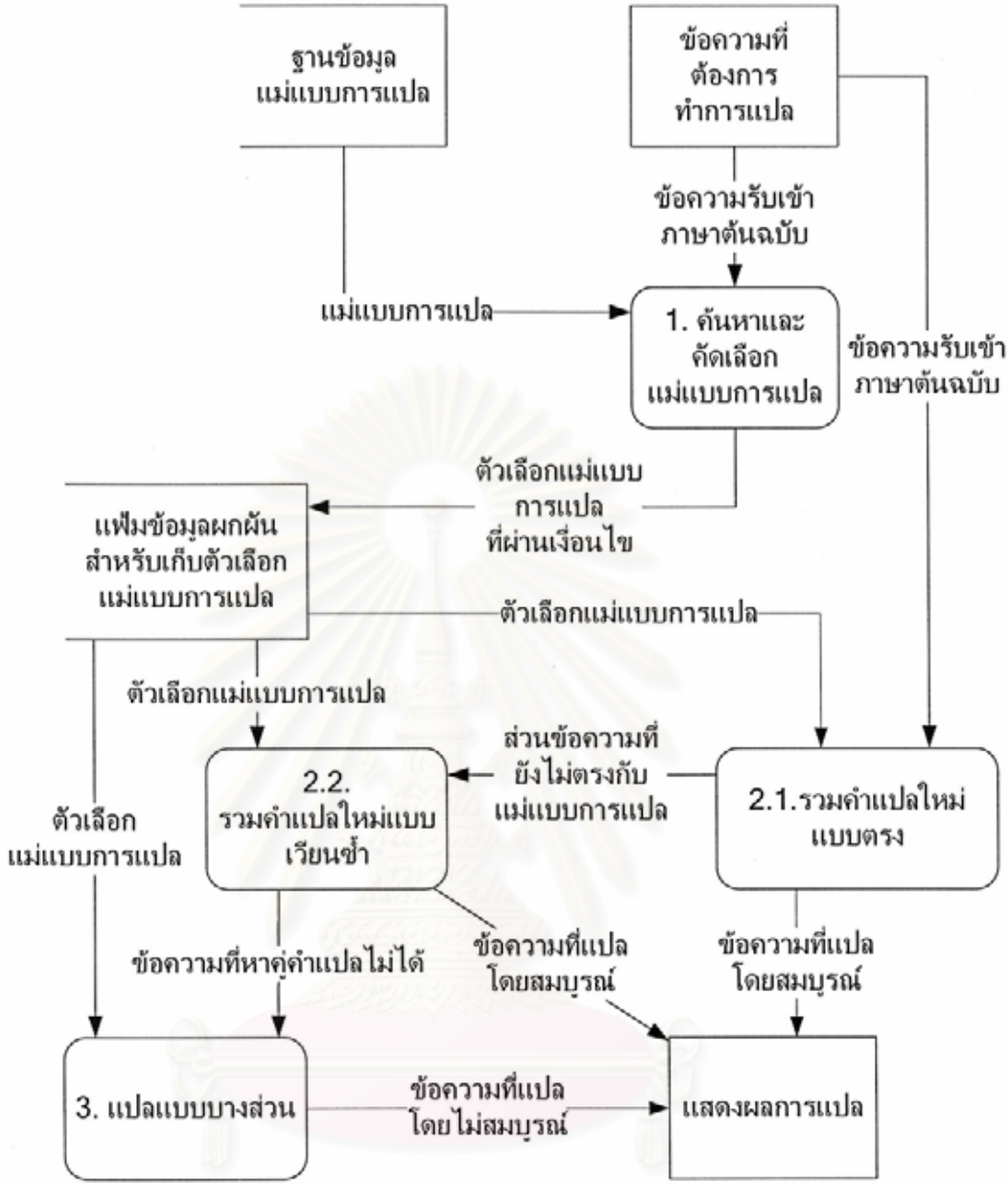
ระบบการรวมคำแปลใหม่จะเป็นขั้นตอนการทำงานหลังจากสกัดแม่แบบการแปลแล้วและนำแม่แบบการแปลนั้นมาเปรียบเทียบหาคำแปล ดังนั้นตัวแปรรับเข้าของระบบดังกล่าวนี้จึงต้องมี 2 ส่วนคือ (1) ข้อความรับเข้า (input) ภาษาต้นฉบับที่ต้องการแปลซึ่งในงานวิจัยนี้คือภาษาไทย และ(2) แม่แบบการแปลที่สกัดได้ (extracted templates) ทั้งหมด

กระบวนการทำงานหลักของระบบคือทำการจับคู่ข้อความรับเข้าของภาษาต้นฉบับกับแม่แบบการแปลตัวใดตัวหนึ่งที่ตรงกันและยาวที่สุด แล้วแปลส่วนผันแปรที่เหลือเรื่อยๆโดยใช้การเรียกซ้ำหาคู่คำแปลที่ตรงกันที่สุด ดังรูปที่ 20



รูปที่ 20 แสดงกระบวนการทำงานหลักของระบบการรวมคำแปลใหม่

ระบบการรวมคำแปลใหม่จะมีขั้นตอนย่อยอยู่ 3 ขั้นตอนคือ (1) ขั้นตอนค้นหาแม่แบบการแปล (pattern retrieval) (2) ขั้นตอนการรวมคำแปลใหม่ (core recombination method) และ (3) ขั้นตอนการแปลแบบบางส่วน (partial translation) ดังรูปที่ 21 ซึ่งภายในรูปที่ 21 จะแสดงรูปฐานข้อมูลในรูปสี่เหลี่ยมเปิดข้าง “⎕” และแสดงรูปข้อมูลรับเข้า(input) และข้อมูลส่งออก(output) ในรูปสี่เหลี่ยม “□” และแสดงรูปกระบวนการทำงานในรูปสี่หลี่ยมมน “▢”



รูปที่ 21 แสดงขั้นตอนการทำงานของระบบการรวมคำแปลใหม่

3.3.1 ขั้นตอนค้นหาแม่แบบการแปล

ภายในขั้นตอนนี้ กระบวนการทำงานจะเริ่มจากระบบจะค้นหาแม่แบบการแปลที่น่าจะเป็นตัวเลือกของคู่คำแปลของข้อความรับเข้าภาษาต้นฉบับซึ่งอาจจะเป็นแม่แบบการแปลที่คลอบคลุมทั้งข้อความรับเข้าหรือบางส่วนของข้อความรับเข้าก็ได้

ระบบการค้นหาตัวเลือกแม่แบบการแปลจะเริ่มจากการค้นหาตัวเลือกแม่แบบการแปลซึ่งจะใช้อัลกอริธึมที่ 10 ด้านล่าง

อัลกอริธึมที่ 10 แสดงกระบวนการค้นหาตัวเลือกแม่แบบการแปล

```
Algorithm RetrieveCandidatePatterns()
For each Word w in SL Input
  For each Translation Pattern P
    If PSL Contains w
      Add P to list of CandidateTranslationPatterns
Return CandidateTranslationPatterns
```

อัลกอริธึมนี้จะทำการค้นหาตัวเลือกแม่แบบการแปล โดยตรวจสอบว่าแม่แบบการแปลใดมีคำที่ข้อความรับเข้ามีบ้างแล้วนำแม่แบบการแปลนั้นเก็บเข้าแฟ้มข้อมูลผกผันโดยมีดัชนีบ่งชี้เป็นหมายเลขกำกับแม่แบบการแปลนั้นๆ ดังรูปที่ 22 ด้านล่าง

(ตลาดหลักทรัพย์) → {P1, P2, P3, P4, ...}

(กำหนด) → {P1, P2, P3, P4, P5, ...}

(จำนวน) → {P1, P2, P3, P4, P5, ...}

(ให้) → {P1, P3, P4, P5 ...}

รูปที่ 22 แสดงลักษณะของแฟ้มข้อมูลผกผันที่มีดัชนีบ่งชี้แม่แบบการแปล

ผลจากกระบวนการค้นหาตัวเลือกแม่แบบการแปลจะทำให้ได้ตัวเลือกแม่แบบการแปลที่มีคำอย่างน้อย 1 คำที่ปรากฏในข้อความรับเข้าอยู่ในแฟ้มข้อมูลผกผันซึ่งจะทำการคัดเลือกแม่แบบการแปลที่จะนำมาเปรียบเทียบแปลต่อโดยตรวจสอบจากเงื่อนไข ซึ่งระบบจะใช้อัลกอริธึมที่ 11 ในการตรวจสอบ

อัลกอริธึมที่ 11 แสดงกระบวนการตรวจสอบเงื่อนไขแม่แบบการแปล

```
Algorithm FilterPatterns(Patterns, SLInput)
For each Translation Pattern P
  If Length(PSL) ≤ Length(SLInput)
    If PSL ContainsOnlyLexicalItemsP(SLInput)
      If WordFrequencyEqualP(PSL, SLInput)
        Add P to list of FilteredPatterns
Return FilteredPatterns
```

ผลลัพธ์ของกระบวนการนี้คือจะคงเหลือตัวเลือกแม่แบบการแปลที่มีความสามารถในการเปรียบเทียบหาคำแปลกับข้อความต้นฉบับจำนวนน้อยลงด้วยการกรองตัวเลือกแม่แบบการแปลด้วยเงื่อนไข 3 ข้อคือ (1) จำนวนคำของแม่แบบการแปลต้องมีเท่ากันหรือน้อยกว่าข้อความรับเข้า (2) ทุกคำของแม่แบบการแปลต้องมีอยู่ในข้อความรับเข้า และ (3) จำนวน

ของแต่ละคำที่มีอยู่ในแม่แบบการแปลส่วนภาษาต้นฉบับต้องมีปริมาณเท่ากับจำนวนของแต่ละคำที่มีอยู่ในข้อความรับเข้า

หลังจากนั้นจึงนำตัวเลือกแม่แบบการแปลที่ผ่านการกรองมาทำการตรวจสอบเงื่อนไขอีกข้อคือ (4) ส่วนผันแปรของแม่แบบการแปลเมื่อแทนแม่แบบการแปลลงในข้อความรับเข้าแล้ว ตำแหน่งการปรากฏส่วนผันแปรของของแม่แบบการแปลต้องจับคู่ตำแหน่งกับการปรากฏส่วนผันแปรของข้อความรับเข้าได้โดยสมบูรณ์

การค้นหาตัวเลือกแม่แบบการแปลที่จะนำมาเปรียบเทียบหาคำแปลต้องผ่านเงื่อนไขทั้ง 4 ข้อ เช่น ข้อความรับเข้า คือ “ตลาดหลักทรัพย์จึงกำหนดให้หุ้นของ ABC จำนวน 10 หุ้น” และภายในฐานข้อมูลแม่แบบการแปลมีแม่แบบการแปลที่เป็นตัวเลือกสำหรับใช้ในการเปรียบเทียบแปลอยู่ 6 แม่แบบคือ

	คำในแม่แบบการแปลภาษาต้นฉบับ
แม่แบบการแปลที่ 1	(ตลาดหลักทรัพย์)(จึง)(เห็นควร)(กำหนด)(ให้)(หุ้น)(ของ) <X1> (จำนวน) <X2> (หุ้น)
แม่แบบการแปลที่ 2	(ตลาดหลักทรัพย์)(จึง)(กำหนด)(จำนวน)(หุ้น)(ของ) <X1> (จำนวน) <X2> (หุ้น)
แม่แบบการแปลที่ 3	(ตลาดหลักทรัพย์)(จึง)(กำหนด)(ให้) <X1> (หุ้น)(ของ)(จำนวน) <X2> (หุ้น)
แม่แบบการแปลที่ 4	(ตลาดหลักทรัพย์)(จึง)(กำหนด)(ให้)(หุ้น)(ของ) <X1> (จำนวน) <X2> (หุ้น)
แม่แบบการแปลที่ 5	<X0> (ABC)
แม่แบบการแปลที่ 6	<X0> (ABC) <X1>

จากตัวอย่างแม่แบบการแปลทั้ง 6 นี้ แม่แบบการแปลที่ 4 และ 6 ผ่านตามเงื่อนไขทั้ง 4 ข้อ แต่แม่แบบการแปลที่ 1 จะไม่ผ่านการกรองเพราะไม่ผ่านเงื่อนไขที่ 2 เพราะแม่แบบการแปลมีคำนอกเหนือไปจากที่ข้อความรับเข้ามีคือ (เห็นควร) และแม่แบบการแปลที่ 2 ถึงแม้ว่าจะผ่านเงื่อนไขที่ 1 และ 2 คือมีจำนวนคำน้อยกว่าและมีทุกคำในข้อความรับเข้าแต่ไม่ผ่านเงื่อนไขที่ 3 คือ มีคำว่า (จำนวน) 2 ตำแหน่งจึงจะไม่ผ่านการกรองเช่นเดียวกันกับแม่แบบการแปลที่ 3 ที่ไม่ผ่านเงื่อนไขที่ 4 คือ ตำแหน่งของส่วนผันแปรของแม่แบบการแปลไม่สามารถจับคู่ตรงกับตำแหน่งของข้อความรับเข้า และแม่แบบการแปลที่ 5 จะไม่ผ่านการกรองเพราะผิดเงื่อนไขที่ 4 ซึ่งผลจากการกรองตัวเลือกแม่แบบการแปลจะถูกเก็บเป็นตัวเลือกในการเปรียบเทียบแปล

เมื่อข้อความรับเข้าได้ตัวเลือกแม่แบบการแปลครบทุกส่วนก็จะถูกนำเข้าระบบทันที แต่ข้อความรับเข้าที่มีตัวเลือกแม่แบบการแปลที่ผ่านการกรองจากขั้นตอนค้นหาแม่แบบการแปลมากกว่า 1 ตัวเลือกก็จำเป็นต้องมีกระบวนการทำงานเพื่อคัดให้เหลือแม่แบบการแปลหลักเพียง

แม่แบบการแปลเดียวโดยการคำนวณค่าความสามารถในการครอบคลุม (coverage) ข้อความรับเข้าของตัวเลือกแม่แบบการแปล กล่าวคือตัวเลือกแม่แบบการแปลใดยาวกว่าและครอบคลุมคำศัพท์ของข้อมูลรับเข้ามากกว่าก็จะได้คะแนนค่าความสามารถในการครอบคลุมสูง โดยใช้สมการที่ 9

$$Cover = \frac{Lenght(Pattern_{sl})}{Lenght(SLInput)} \quad (9)$$

ค่าคะแนนที่เป็นผลจากสมการนี้จะช่วยระบบเลือกในกรณีที่ตัวเลือกแม่แบบการแปลมีมากกว่า 1 ตัวเลือก เช่น ข้อความรับเข้า คือ “ตลาดหลักทรัพย์จึงกำหนดให้หุ้นของ ABC จำนวน 10 หุ้น” และมีตัวเลือกแม่แบบการแปลอยู่ 5 ตัวเลือกคือ

	ตัวเลือกแม่แบบการแปล
แม่แบบการแปลที่ 1	(กำหนด)(ให้)(หุ้น)(ของ) <X1> (จำนวน) <X2> (หุ้น) ↔ (has) (set) (shares) (of) <Y1> (,) (amounting) (to) <Y2> (shares)
แม่แบบการแปลที่ 2	(ตลาดหลักทรัพย์)(จึง)(กำหนด)(ให้)(หุ้น)(ของ) <X1> ↔ (The SET) (has) (set) (shares) (of) <Y1>
แม่แบบการแปลที่ 3	(ตลาดหลักทรัพย์)(จึง)(กำหนด)(ให้)(หุ้น)(ของ) <X1> (จำนวน) <X2> (หุ้น) ↔ (The SET) (has) (set) (shares) (of) <Y1> (,) (amounting) (to) <Y2> (shares)
แม่แบบการแปลที่ 4	<X1> (ABC) <X2> ↔ <Y1> (ABC) <Y2>
แม่แบบการแปลที่ 5	<X1> (10) <X2> ↔ <Y1> (10) <Y2>

ในกรณีนี้ระบบก็จะเลือกแม่แบบการแปลที่ 3 เป็นแม่แบบการแปลหลักของข้อความรับเข้าเพราะแม่แบบการแปลที่ 3 มีความยาวที่สุดและมีคำศัพท์ครอบคลุมข้อมูลรับเข้ามากที่สุดเมื่อเปรียบเทียบกับตัวเลือกแม่แบบอื่นจึงทำให้มีคะแนนค่าความสามารถในการครอบคลุมสูงที่สุด และแม่แบบการแปลที่ 4 และ 5 จะถูกนำเก็บเป็นแม่แบบการแปลส่วนผันแปรทันทีเนื่องจากไม่มีตัวเลือกแม่แบบการแปลอื่น

เมื่อได้แม่แบบการแปลจากตัวเลือกแม่แบบการแปลแล้ว จะถูกนำไปเก็บไว้ในฐานข้อมูลตัวเลือกแม่แบบการแปลเพื่อจะนำไปใช้ในการเปรียบเทียบแปลและใช้ในขั้นตอนการรวมคำแปลใหม่ต่อไป

3.3.2 ขั้นตอนการรวมคำแปลใหม่

ขั้นตอนการรวมคำแปลใหม่คือส่วนที่ระบบจะทำการรวบรวมและเรียบเรียงแม่แบบการแปลส่วนต่างๆ เข้าไว้ด้วยกันเป็นข้อความแปล ระบบจะนำแม่แบบการแปลที่ผ่าน

ขั้นตอนค้นหาแม่แบบการแปลมาใช้ในการเรียบเรียงคำแปลออกมาเป็นข้อความแปล แม่แบบการแปลหลักที่ได้จะถูกนำมาใช้เป็นฐานของการแทนที่เนื้อความของข้อความรับเข้าเพราะเป็นส่วนที่ยาวและมีปริมาณคำซ้ำกับข้อความรับเข้ามากที่สุด หลังจากนั้นจึงนำแม่แบบการแปลอื่นมาเติมส่วนผันแปร ในการเติมส่วนผันแปรลงในแม่แบบการแปลหลักระบบจะใช้อัลกอริธึมที่ 12 ด้านล่างเป็นอัลกอริธึมหลัก

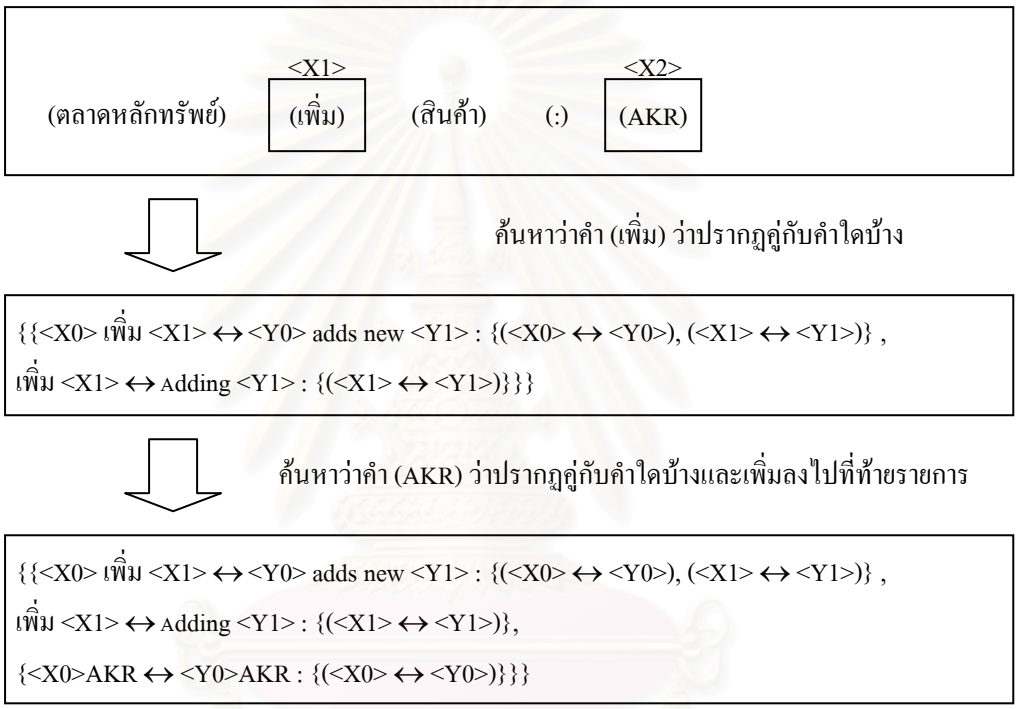
อัลกอริธึมที่ 12 แสดงกระบวนการหลักของการรวมคำแปลใหม่

```
Algorithm recombine_vars(t : matched template,
                         vs : list of instantiated variables in t,
                         D : template database)
returns list of translation results
if length(vs) == 0 then
  return [t]
else
  let result = [], First = vs[0], Rest = vs[1:], exact_result = []
  for subtrans in recombine_exact(t, first, D) do
    for trans in recombine_vars(subtrans, rest, D) do
      append trans to exact_result
    end for
  end for
  if length(exact_result) > 0 then return exact_result
  end if
  let recur_result = []
  for subtrans in recombine_recur(t, first, D) do
    for trans in recombine_vars(subtrans, rest, D) do
      append trans to recur_result
    end for
  end for
  if lenght(recur_result) > 0 then return recur_result
  end if
end if
end algorithm
```

การทำงานของอัลกอริธึมที่ 12 เริ่มจากระบบจะสร้างรายการคำส่วนผันแปรซึ่งเป็นรายการของการปรากฏของคำในแม่แบบการแปลที่ผ่านขั้นตอนค้นหาแม่แบบการแปลสำหรับส่วนผันแปร

เช่นเมื่อข้อความรับเข้าคือ “ตลาดหลักทรัพย์เพิ่มสินค้า : AKR” และได้แม่แบบการแปลหลักคือ [ตลาดหลักทรัพย์ <X1> สินค้า : <X2> ↔ The SET <Y1> listed securities : <Y2>] ซึ่งการจับคู่ส่วนผันแปรคือ <X1> เป็นคู่คำแปลกับ <Y1> และ <X2> เป็นคู่คำแปลกับ <Y2> ระบบจะแทนที่คำ (เพิ่ม) ด้วยตัวแปร <X1> และแทนที่คำ (AKR) ด้วยตัวแปร <X2> หลังจากนั้น

ระบบนำแม่แบบการแปลสำหรับส่วนผันแปรของ (เพิ่ม) และ (AKR) มาสร้างรายการคำส่วนผันแปรทีละตัวแปรตามลำดับ ตัวแปรแรกคือ <X1> (เพิ่ม) ทั้งหมดซึ่งมีอยู่ 2 คู่คำแปลคือ [<X0> เพิ่ม <X1> ↔ <Y0> adds new <Y1>] ซึ่งการจับคู่ส่วนผันแปรคือ <X0> เป็นคู่คำแปลกับ <Y0> และ <X1> เป็นคู่คำแปลกับ <Y1> และ [เพิ่ม <X1> ↔ Adding <Y1>] ซึ่งการจับคู่ส่วนผันแปรคือ <X1> เป็นคู่คำแปลกับ <Y1> โดยที่จะนำคู่คำแปลทั้ง 2 คู่ใส่ลงในรายการคำส่วนผันแปร แล้วจึงนำคู่คำแปลของ (AKR) ซึ่งมีอยู่ 1 คู่คือ [AKR ↔ AKR] ซึ่งการจับคู่ส่วนผันแปรคือ <X0> เป็นคู่คำแปลกับ <Y0> และนำผลไปเพิ่มต่อท้ายรายการคำส่วนผันแปร ดังรูปที่ 23 ด้านล่าง



รูปที่ 23 แสดงขั้นตอนการทำงานของเพิ่มคู่คำแปลลงในรายการคำส่วนผันแปร

เมื่อได้รายการคำส่วนผันแปรแล้วระบบจะทำการตรวจสอบว่าคู่คำแปลในรายการคำส่วนผันแปรสามารถแทนที่ส่วนผันแปรของแม่แบบการแปลหลักได้โดยตรงพอดี (exact) หรือไม่ ถ้าแทนที่ได้โดยตรงพอดี ก็จะใช้ระบบการรวมคำแปลใหม่แบบตรง(exact recombination) ถ้าไม่สามารถแทนที่ได้ก็จะใช้ระบบการรวมคำแปลใหม่แบบเวียนใช้ (recursive recombination)

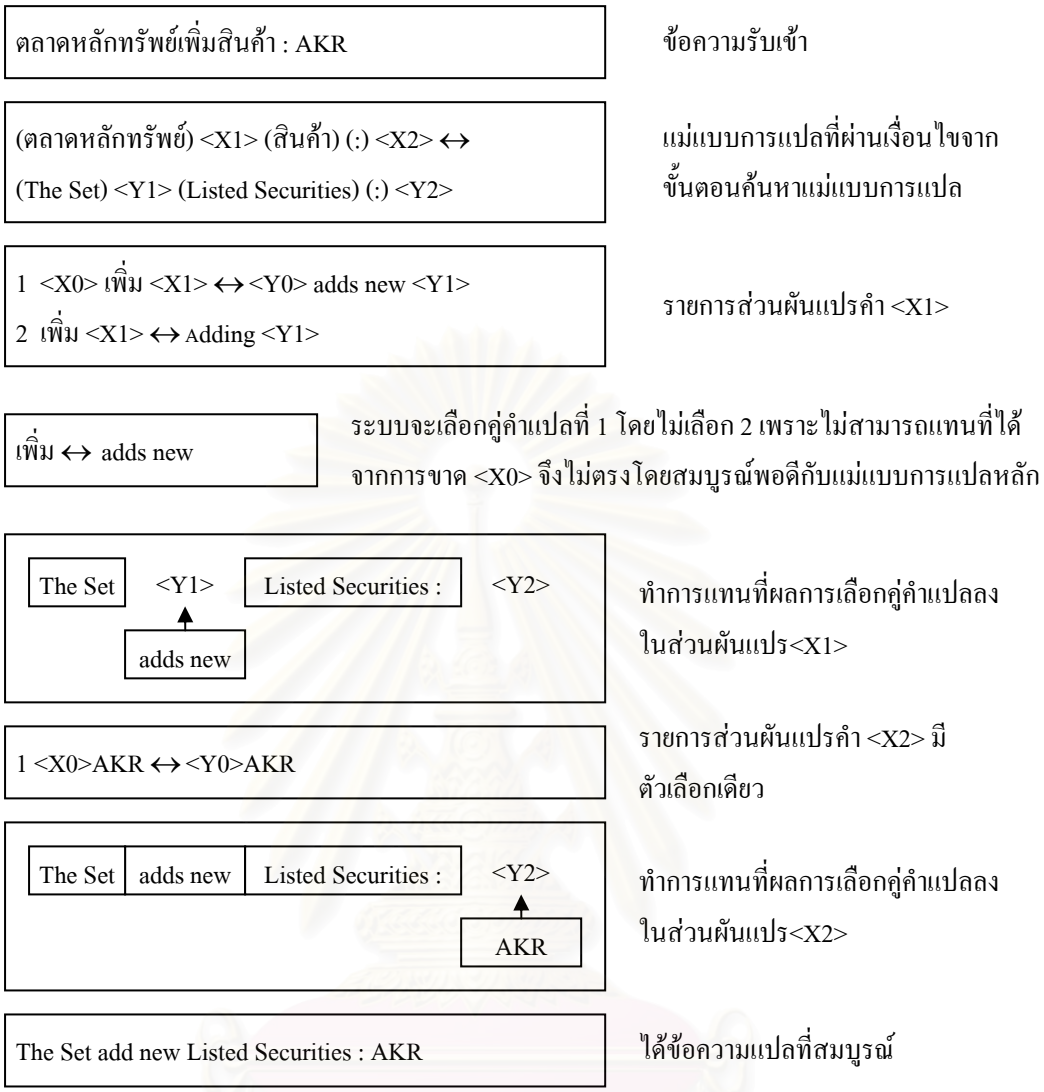
3.3.2.1 ระบบการรวมคำแปลใหม่แบบตรง

ระบบการรวมคำแปลใหม่แบบตรงจะถูกนำมาใช้หากคู่คำแปลในรายการคำส่วนผันแปรสามารถแทนที่ส่วนผันแปรของแม่แบบการแปลหลักได้โดยตรงพอดี โดยใช้อัลกอริธึมที่ 13 ด้านล่าง

อัลกอริธึมที่ 13 แสดงกระบวนการรวมคำแปลใหม่แบบตรง

```
algorithm recombine_exact(t : matched template,
                          v : instantiated variable,
                          D : template database)
returns list of translation results
let result = []
let v_value = value of v
let tpls = get_templates(v_value, D)
if length(tpls) > 0 then
  for each template tpl in tpls do
    let s = substitution of v with tpl in t
    append s to result
  end for
end if
return result
end algorithm
```

อัลกอริธึมกระบวนการรวมคำแปลใหม่แบบตรงจะถูกเรียกจากอัลกอริธึมที่ 12 ซึ่งเป็นอัลกอริธึมหลัก อัลกอริธึมที่ 13 จะเริ่มจากการนำรายการคำส่วนผันแปรและดึงคำในรายการมาแทนที่ส่วนผันแปรในแม่แบบการแปลหลักทีละส่วนผันแปรตามลำดับจนครบทุกส่วนผันแปรเพื่อจะสร้างเป็นข้อความแปลดังรูปที่ 24 ด้านล่าง หากส่วนผันแปรใดมีมากกว่า 1 ตัวเลือกที่สามารถแทนที่ส่วนผันแปรของแม่แบบการแปลหลักได้โดยตรงพอดี ข้อความแปลที่ได้ก็จะมีหลายข้อความตามจำนวนส่วนผันแปรที่แทนที่ได้นั้น หากรายการคำส่วนผันแปรไม่สามารถแทนที่ส่วนผันแปรของแม่แบบการแปลหลักได้โดยตรงพอดี ระบบการก็จะข้ามการทำงานในอัลกอริธึมที่ 13 ไปและเรียกระบบการรวมคำแปลใหม่แบบเวียนใช้แทน



รูปที่ 24 แสดงกระบวนการรวบรวม 3 ตัวแปรภายในสายอักขระเพื่อสร้างข้อความแปล

3.3.2.2 ระบบการรวมคำแปลใหม่แบบเวียนใช้

ระบบการรวมคำแปลใหม่แบบเวียนใช้คือระบบที่จะเรียบเรียงคำแปลจากรายการคำส่วนผันแปรแทนที่ลงในแม่แบบการแปลหลักทีละส่วนแม้จะไม่สามารถแทนที่ได้โดยตรงพอดีแต่คู่คำแปลในรายการคำส่วนผันแปรมีความสามารถในการเติมเต็มส่วนผันแปรได้ เนื่องจากไม่มีแม่แบบการแปลที่สามารถแทนที่ได้โดยตรงพอดีในฐานข้อมูลแม่แบบการแปล ขั้นตอนของระบบการรวมคำแปลใหม่แบบเวียนใช้คือจะนำคู่คำแปลจากรายการคำส่วนผันแปรที่ไม่สามารถแทนที่โดยตรงพอดีมาแทนที่ลงส่วนผันแปรทีละส่วนจนข้อความแปลสมบูรณ์ ระบบการรวมคำแปลใหม่แบบเวียนใช้จะใช้อัลกอริธึมที่ 14 ด้านล่าง

อัลกอริธึมที่ 14 แสดงกระบวนการรวมคำแปลใหม่แบบเวียนใช้

```
algorithm recombine_recur(t : matched template,
                          v : instantiated variable,
                          D : template database)
returns list of translation results
let result = []
let v_value = value of v
let trans = recombine(v_value, D)
if lenght(trans) > 0 then
  for each translation t' in trans do
    let s = substitution of v with t' in t
    append s to result
  end for
end if
return result
end algorithm
```

อัลกอริธึมที่ 14 จะถูกเรียกใช้เมื่ออัลกอริธึมที่ 12 เรียกอัลกอริธึมที่ 13 มาแต่ไม่สามารถสร้างผลลัพธ์เป็นข้อความแปลได้ อัลกอริธึมกระบวนการรวมคำแปลใหม่แบบเวียนใช้จะเริ่มจากการนำรายการคำส่วนผันแปรและดึงคำในรายการมาแทนที่ส่วนผันแปรในแม่แบบการแปลหลักทีละคู่คำแปลและเรียกซ้ำแทนที่ไปเรื่อยๆ จนครบทุกส่วนผันแปรเพื่อจะสร้างเป็นข้อความแปล

เช่น ข้อความรับเข้าคือ “กขคงจ” และมีแม่แบบการแปลคือ “<X0>(ง)(จ) ↔ <Y0>(D)(E)” และ “(ก)<X1> ↔ (A)<Y1>” และ “<X0>(ข)(ค)<X1> ↔ <Y0>(B)(C)<X1>” จากแม่แบบการแปลทั้ง 3 นี้ แม่แบบการแปลที่ผ่านตัวเลือกแม่แบบการแปลหลักคือ “<X0>(ง)(จ) ↔ <Y0>(D)(E)” ดังนั้นภายในขั้นตอนนี้จะทำการเติมแม่แบบการแปลที่เหลือลงแม่แบบการแปลหลักทีละส่วนคือนำ “<X0>(ข)(ค)<X1> ↔ <Y0>(B)(C)<X1>” เติมลงแม่แบบการแปลหลัก จะได้เป็น “<X0>(ข)(ค)(ง)(จ) ↔ <Y0>(B)(C)(D)(E)” และเรียกระบบซ้ำเพื่อเติมแม่แบบการแปลที่เหลือคือ “(ก)<X1> ↔ (A)<Y1>” ก็จะได้เป็นข้อความแปลที่สมบูรณ์คือ “(ก)(ข)(ค)(ง)(จ) ↔ (A)(B)(C)(D)(E)”

ผลการทำงานของขั้นตอนการรวมคำแปลใหม่จะได้ข้อความแปลที่สมบูรณ์หากข้อความรับเข้ามีคำและการเรียงตัวของข้อความเหมือนหรือใกล้เคียงกับแม่แบบการแปลที่สกัดได้จากตัวอย่างภายในคลังข้อมูลเทียบบท อย่างไรก็ตามหากข้อความรับเข้ามีเพียงบางส่วนของเนื้อความใกล้เคียงกับแม่แบบการแปลก็จะสามารถแปลได้เพียงบางส่วนซึ่งต้องใช้ขั้นตอนต่อไปในการแปลข้อความคือขั้นตอนการแปลแบบบางส่วน

3.3.3 ขั้นตอนการแปลแบบบางส่วน

ภายในขั้นตอนการแปลแบบบางส่วนจะทำงานคล้ายกับขั้นตอนการรวมคำแปลใหม่ทุกอย่างเพียงแต่จะทำงานได้เพียงบางส่วนที่มีคำตรงกับแม่แบบการแปลภายในฐานข้อมูลเท่านั้น ส่วนของข้อความใดที่ไม่มีในแม่แบบการแปลจะไม่สามารถถูกแปลได้ ดังนั้นส่วนข้อความที่ไม่สามารถแปลได้ ผู้วิจัยจึงกำหนดให้นำข้อความรับเข้าของส่วนนั้นมาทันทีซึ่งเป็นส่วนที่ไม่เหมือนกับของแม็คเททที่จะให้เปลี่ยนส่วนที่ไม่สามารถแปลได้เป็นเครื่องหมายปรัศนี

ขั้นตอนนี้จะช่วยให้ระบบแปลข้อความได้เท่าที่แม่แบบการแปลจะสามารถแปลได้สูงสุด ซึ่งผู้วิจัยเชื่อว่าข้อความส่วนใหญ่ที่ยังไม่พบในคลังข้อมูลคือข้อความที่เป็นชื่อบริษัทและวันที่เท่านั้น การดึงข้อความต้นฉบับมาใช้ในส่วนข้อความที่แปลไม่ได้จึงอาจช่วยในกรณีที่เป็นชื่อย่อของบริษัทซึ่งเป็นตัวอักษรภาษาอังกฤษอยู่แล้ว เช่น “NYK” “PTT” เป็นต้น และเป็นผลให้การแปลถูกต้องมากขึ้น อย่างไรก็ตามเมื่อไม่มีแม่แบบการแปล ผลการแปลที่ได้ก็จะไม่สมบูรณ์เหมือนกับการแปลที่มีแม่แบบการแปลครบถ้วน ซึ่งจุดนี้เป็นข้อจำกัดของการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง

สรุปบทที่ 3 ระบบของการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างในงานวิจัยชิ้นนี้จะเริ่มจากการสกัดแม่แบบการแปลจากคลังข้อมูลเทียบบท และนำแม่แบบการแปลนั้นมาเทียบแปลข้อความรับเข้า และกล่าวได้ว่าขั้นตอนการสกัดแม่แบบการแปลเป็นขั้นตอนที่สำคัญที่สุด เพราะเมื่อสกัดแม่แบบการแปลไม่ได้ขั้นตอนอื่นๆ ก็จะไม่สามารถทำงานได้อย่างมีประสิทธิภาพหรือไม่สามารถแปลข้อความได้

บทที่ 4

ผลการทดลองแปลภาษาด้วยเครื่อง

ภายในบทที่ 4 จะกล่าวถึงการทดลองของระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง ซึ่งจะแยกออกเป็น 2 ส่วนหลักคือการทดลองสกัดแม่แบบการแปลจากคลังข้อมูล และการทดลองแปลข้อความ

กระบวนการทดลองระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างจะเริ่มจากการนำคลังข้อมูลเทียบบทที่เก็บรวบรวมมาแบ่งแบบสุ่มออกเป็น 2 ส่วนคือ (1) ส่วนที่นำไปสกัดเป็นแม่แบบการแปลจำนวนร้อยละ 90 จำนวน 452 คู่รายงานข่าว นับจำนวนคู่ข้อความได้ 1,310 คู่ และ (2) ส่วนที่ไม่ได้นำไปสกัดเป็นแม่แบบการแปลอีกร้อยละ 10 จำนวน 51 คู่รายงานข่าว นับจำนวนคู่ข้อความได้ 297 คู่ที่จะนำไปเป็นชุดข้อมูลเพื่อทดลองแปลข้อความ โดยผู้วิจัยเน้นไปที่การทดลองสกัดแม่แบบการแปลเนื่องจากพบว่าความถูกต้องของการสกัดแม่แบบการแปลเป็นปัจจัยสำคัญต่อความถูกต้องของระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง เพราะหากแม่แบบการแปลที่ได้มีความถูกต้องและครอบคลุม ผลการแปลก็จะมีความถูกต้องมากขึ้นไปด้วย

ผลที่ได้จากการทดลองสกัดแม่แบบการแปลและการทดลองการแปลจะถูกนำมาวิเคราะห์ความถูกต้อง โดยการทดลองสกัดแม่แบบการแปลใช้วิธีการตรวจสอบประเมินผลด้วยตนเองจากการตรวจสอบว่า ส่วนซ้ำของคู่ข้อความที่ถูกสกัดเป็นส่วนคงที่ของทั้งสองภาษานั้นเป็นคู่คำแปลของกันและกัน และส่วนไม่ซ้ำของข้อความที่ถูกสกัดเป็นส่วนผันแปรของทั้งสองภาษานั้นเป็นคู่คำแปลของกันและกัน ถ้าพิจารณาแล้วว่าเป็นคู่คำแปลของกันได้ก็จะประเมินว่าเป็นการสกัดที่ถูกต้อง ส่วนการทดลองแปลข้อความใช้วิธีการตรวจสอบประเมินผล 2 แบบคือ แบบแรกประเมินจากการเทียบเคียงกับข้อความภาษาอังกฤษที่คู่กับข้อความภาษาไทยนั้นและแบบที่สองประเมินโดยผู้วิจัยว่าข้อความที่แปลได้นั้นยอมรับได้หรือไม่

4.1 ผลการทดลองสกัดแม่แบบการแปลจากคลังข้อมูล

ผู้วิจัยได้นำคลังข้อมูลเทียบบทที่สุ่มแยกออกมาจำนวนร้อยละ 90 ของคลังข้อมูลทั้งหมดหรือคู่รายงานข่าวตลาดหุ้นแบบรายวันจำนวน 452 คู่ จำนวนคู่ข้อความ 1,310 คู่มาทำการทดลองสกัดแม่แบบการแปล โดยนำคลังข้อมูลทั้งหมดมาคัดลอกออกเป็นจำนวน 4 ชุด โดยจะเรียกว่า คลังข้อมูลชุดทดลองที่ 1 2 3 และ 4 ตามลำดับ โดยคลังข้อมูลชุดทดลองทั้ง 4 จะถูกแบ่งเป็น 4 ประเภทการทดลองคือ

(1) คลังข้อมูลชุดทดลองที่ 1 ซึ่งภายในคลังข้อมูลชุดนี้ได้รวบรวมคู่ข้อความทั้งหมดจำนวน 1,310 คู่ข้อความ ลงในแฟ้มข้อมูลเดียวกันและทำการตัดแบ่งคำภาษาไทยในคลังข้อมูลโดยใช้โปรแกรมตัดคำอัตโนมัติ 'SWATH' (ไพศาล, 2541) ของศูนย์คอมพิวเตอร์และอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) แทนการตัดแบ่งคำด้วยตนเองเนื่องจากภาษาไทยยังมีความกำกวมเรื่องเกณฑ์การตัดคำ (Aroonmanakun, 2002) เพื่อให้ผลการตัดคำที่ได้มีความสม่ำเสมอ โดยการแบ่งคำในภาษาไทยผู้วิจัยใช้การเว้นวรรคเป็นขอบเขตของคำแต่ละคำเช่นเดียวกับคำในภาษาอังกฤษที่แยกคำได้จากการเว้นวรรค

(2) คลังข้อมูลชุดทดลองที่ 2 ซึ่งภายในคลังข้อมูลชุดนี้ได้รวบรวมคู่ข้อความทั้งหมดจำนวน 1,310 คู่ข้อความลงในแฟ้มข้อมูลเดียวกัน แต่ไม่ได้ตัดแบ่งคำ

(3) คลังข้อมูลชุดทดลองที่ 3 ซึ่งภายในคลังข้อมูลชุดนี้ได้ทำการตัดแบ่งคำภาษาไทยในคลังข้อมูลและแบ่งแฟ้มข้อมูลออกตามความคล้ายคลึงกันของข้อความที่มีส่วนซ้ำของข้อความตรงกันภายในแต่ละรายงานข่าว กล่าวคือผู้วิจัยได้ทำการจำแนกข้อความที่มีส่วนซ้ำเหมือนกันด้วยตนเองโดยแบ่งออกเป็นกลุ่มแฟ้มข้อมูลจำนวน 184 กลุ่มแฟ้มข้อมูล

(4) คลังข้อมูลชุดทดลองที่ 4 ซึ่งภายในคลังข้อมูลชุดนี้ไม่ได้ทำการตัดแบ่งคำภาษาไทยในคลังข้อมูลและแบ่งแฟ้มข้อมูลออกตามความคล้ายคลึงกันของข้อความที่มีส่วนซ้ำของข้อความตรงกันภายในแต่ละรายงานข่าว จำนวน 184 กลุ่มแฟ้มข้อมูล

สรุปได้ว่าคลังข้อมูลชุดทดลองที่ 1 และ 2 เป็นข้อมูลที่เป็นตัวอย่าง (example) แต่คลังข้อมูลชุดทดลองที่ 3 และ 4 เป็นข้อมูลที่เป็นการตัดแบ่งเนื้อหาออกตามความคล้ายคลึงกันของข้อความที่มีส่วนซ้ำของข้อความตรงกันโดยผู้วิจัยเองซึ่งเป็นตัวอย่างที่ดี (exemplar) สำหรับการทดลอง และคาดว่าผลการสกัดแม่แบบการแปลที่ได้จากคลังข้อมูลชุดทดลองที่ 3 และ 4 จะเป็นแม่แบบการแปลที่ถูกต้องและสามารถใช้ตรวจสอบแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 31 และ 2 ได้

ตารางที่ 1 แสดงลักษณะของคลังข้อมูลชุดทดลอง

คลังข้อมูลชุดทดลองที่	ตัดแบ่งคำ	แยกข้อความเป็นกลุ่มตามความคล้ายคลึง
1	ทำ	ไม่ทำ
2	ไม่ทำ	ไม่ทำ
3	ทำ	ทำ
4	ไม่ทำ	ทำ

4.1.1 ผลการทดลองจากคลังข้อมูลชุดทดลองที่ 1

จากการสกัดแม่แบบการแปลคลังข้อมูลชุดทดลองที่ 1จะได้แม่แบบการแปลทั้งหมดจำนวน 528 แม่แบบ โดยจากการตรวจสอบผลลัพธ์ที่ได้พบว่า แม่แบบการแปลทั้งที่เป็นแม่แบบการแปลที่ผ่านการสกัดส่วนซ้ำของข้อความเป็นส่วนคงที่และสกัดส่วนไม่ซ้ำเป็นส่วนผันแปรมีจำนวนเพียง 52 แม่แบบที่จัดได้ว่าถูกต้อง เพราะแม่แบบที่สกัดมาทั้งของภาษาไทยและอังกฤษนั้นเป็นคู่คำแปลของกันและกัน แม่แบบการแปลเหล่านี้สามารถนำไปใช้ในการเปรียบเทียบแปลได้ ตัวอย่างเช่น ผลการสกัดจากข้อความ

(การ) (รับ) (หุ้น) (เพิ่ม) (ทุน) (เป็น) (หลักทรัพย์) (จดทะเบียน) (เพิ่มเติม) (:) (DELTA) ↔ (LISTED) (SECURITIES) (GRANTED) (BY) (THE SET) (:) (DELTA)
(การ) (รับ) (หุ้น) (เพิ่ม) (ทุน) (เป็น) (หลักทรัพย์) (จดทะเบียน) (เพิ่มเติม) (:) (EASTW) ↔ (LISTED) (SECURITIES) (GRANTED) (BY) (THE SET) (:) (EASTW)
(การ) (รับ) (หุ้น) (เพิ่ม) (ทุน) (เป็น) (หลักทรัพย์) (จดทะเบียน) (เพิ่มเติม) (:) (HANA) ↔ (LISTED) (SECURITIES) (GRANTED) (BY) (THE SET) (:) (HANA)

จะได้แม่แบบการแปล คือ [การ รับ หุ้น เพิ่ม ทุน เป็น หลักทรัพย์ จดทะเบียน เพิ่มเติม : <X1>] ↔ [LISTED SECURITIES GRANTED BY THE SET : <Y1>] โดยจับคู่ส่วนผันแปร <X1> ↔ <Y1> ที่เป็นชื่อย่อของบริษัทซึ่งเป็นเอนทิตีระบุนาม

นอกจากแม่แบบการแปลข้างต้น ผลการสกัดยังได้แม่แบบการแปลส่วนคงที่ที่ถูกต้องอีก 7 แม่แบบได้แก่

1) {มูลค่าที่ตราไว้ : <X1> บาทต่อหุ้น ↔ Par Value : <Y1> Baht per share} โดยจับคู่ {<X1> ↔ <Y1>}
2) {ราคาขายหุ้นละ : <X1> บาท ↔ Offering Price : <Y1> Baht per share} โดยจับคู่ {<X1> ↔ <Y1>}
3) {ราคาการใช้สิทธิ : <X1> บาทต่อหุ้น ↔ Exercise Price : <Y1> Baht per share} โดยจับคู่ {<X1> ↔ <Y1>}
4) {อัตราการจองซื้อ : <X1> หุ้นเดิม : <X2> หุ้นปันผล ↔ Ratio : <Y1> existing shares : <Y2> stock dividends} โดยจับคู่ {<X1> ↔ <Y1> และ <X2> ↔ <Y2>}
5) {วันใช้สิทธิและชำระเงิน : <X1> ↔ Exercise and Payment Date : <Y1>} โดยจับคู่ {<X1> ↔ <Y1>}

6) {ทุนใหม่ : <X1> ↔ New Capital : <Y1>} โดยจับคู่ {<X1> ↔ <Y1>}

7) {ทุนเดิม : <X1> ↔ Old Capital : <Y1>} โดยจับคู่ {<X1> ↔ <Y1>}

จากการที่ระบบสามารถสร้างต้นไม้การปรากฏร่วมที่สามารถจับคู่กันได้อย่างเด่นชัดจะทำให้ระบบสามารถนำคู่โหนดใบของต้นไม้การแปลนั้นไปสร้างแม่แบบการแปลส่วนคงที่ได้อย่างถูกต้องและระบบจะนำคู่ของส่วนไม่ซ้ำไปสร้างแม่แบบการแปลส่วนผันแปร โดยคู่ของส่วนไม่ซ้ำที่ปรากฏจากการทดลองสกัดจะเป็นคู่ของชื่อย่อบริษัท เช่น “ASIAN ↔ ASIAN” “SSEC ↔ SSEC” และคู่ของตัวเลขที่บอกปริมาณและจำนวนเงิน เช่น “1.5 ↔ 1.5” “30 ↔ 30” เป็นต้น รวมเป็นแม่แบบการแปลส่วนผันแปรที่ถูกต้องจำนวน 44 แม่แบบ

ผลการสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 1 ได้แม่แบบการแปลส่วนคงที่ที่ถูกต้องจำนวน 8 แม่แบบคิดเป็นร้อยละ 1.52 และแม่แบบการแปลส่วนผันแปรที่ถูกต้องจำนวน 44 แม่แบบคิดเป็นร้อยละ 8.33 รวมได้แม่แบบการแปลที่ถูกต้องจำนวน 52 แม่แบบคิดเป็นร้อยละ 9.85 และได้แม่แบบการแปลที่ไม่ถูกต้องจำนวน 476 แม่แบบคิดเป็นร้อยละ 90.15

ตารางที่ 2 แสดงผลการสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 1

ชนิดของแม่แบบการแปล		จำนวน	ร้อยละ	จำนวน	ร้อยละ
แม่แบบการแปลที่ถูกต้อง	ส่วนคงที่	8	1.52	52	9.85
	ส่วนผันแปร	44	8.33		
แม่แบบการแปลที่ไม่ถูกต้อง				476	90.15
แม่แบบการแปลทั้งหมดที่สกัดได้				528	100

อย่างไรก็ตาม แม่แบบการแปลจำนวนร้อยละ 90.15 ของผลการทดลองสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 1 จะสกัดได้แม่แบบการแปลที่ไม่ถูกต้อง และไม่สามารถนำไปใช้ในการเปรียบเทียบแปลได้ เพราะจับคู่คำแปลส่วนคงที่และส่วนผันแปรผิด กล่าวคือจับคู่คำที่ไม่ได้คู่เป็นคำแปลของกันและกัน เช่น จากข้อความ

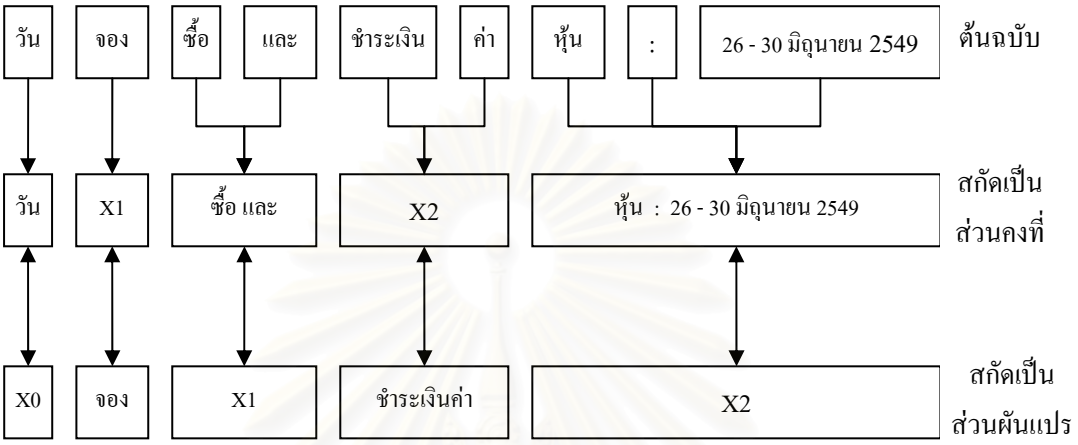
(วัน) (จอง) (ซื้อ) (และ) (ชำระเงิน) (ค่า) (หุ้น) (:) (26 – 30 มิถุนายน 2549) ↔ (Subscription) (and) (Payment) (Date) (:) (June 26-30, 2006)

จะได้แม่แบบการแปลที่สกัดได้ คือ

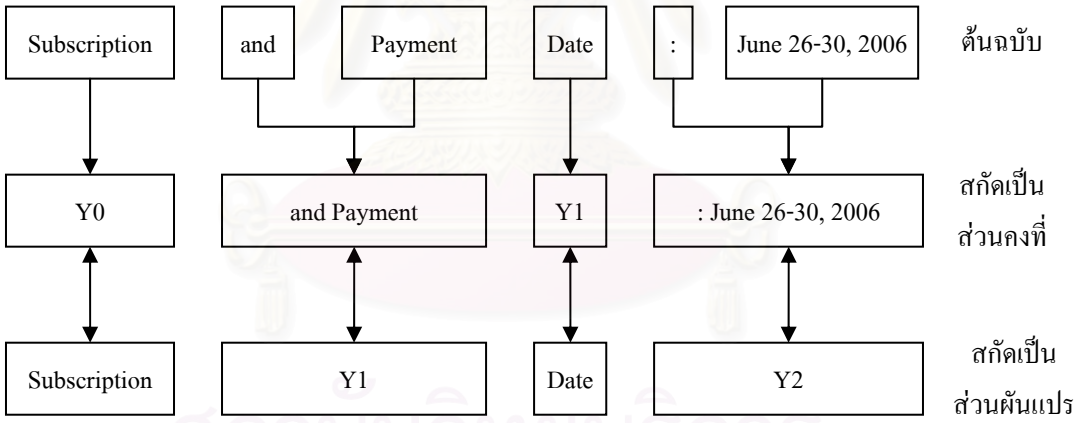
[วัน <X1> ซื้อ และ <X2> หุ้น : 26 – 30 มิถุนายน 2549] ↔ [<Y0> and Payment <Y1> : June 26-30, 2006] และระบบได้กำหนดให้ส่วนผันแปร <X1> ↔ <Y0> และ <X2> ↔ <Y1>

[<X0> จอง <X1> ชำระเงิน ค่า <X2>] ↔ [Subscription <Y1> Date <Y2>] และระบบได้กำหนดให้ส่วนผันแปร <X0> <X1> ↔ <Y1> และ <X2> ↔ <Y2>

ผลการสกัดสามารถจัดลำดับความสัมพันธ์ได้ดังรูปที่ 25 และรูปที่ 26 ด้านล่าง



รูปที่ 25 แสดงความสัมพันธ์ของแม่แบบการแปลภาษาไทยจากคลังข้อมูลชุดทดลองที่ 1



รูปที่ 26 แสดงความสัมพันธ์ของแม่แบบการแปลภาษาอังกฤษจากคลังข้อมูลชุดทดลองที่ 1

ความสัมพันธ์ที่เกิดขึ้นภายในแม่แบบการแปลที่สกัดได้จากคลังข้อมูลชุดทดลองที่ 1 จากรูปที่ 25 และรูปที่ 26 ซึ่งเป็นตัวอย่างของการค้นหาส่วนซ้ำมาสกัดเป็นส่วนคงที่และการค้นหาส่วนไม่ซ้ำมาสกัดเป็นส่วนผันแปร ซึ่งทั้งส่วนคงที่และส่วนผันแปรของทั้ง 2 ภาษาไม่ได้เป็นคู่คำแปลของกันและกัน ดังนั้นเมื่อระบบทำการจับคู่เทียบแม่แบบการแปลระหว่างภาษาจะได้แม่แบบการแปลส่วนคงที่เป็น [วัน <X1> ซื้อ และ <X2> หุ้น : 26 – 30 มิถุนายน 2549] ↔ [<Y0> and Payment <Y1> : June 26-30, 2006] ซึ่งไม่ได้เป็นคู่คำแปลกัน และแม่แบบการแปลส่วน

ผันแปรเป็น [<X0> จอง <X1> ชำระเงิน ค่า <X2>] ↔ [Subscription <Y1> Date <Y2>] ซึ่งก็ไม่ได้เป็นคู่คำแปลกันเช่นเดียวกัน

ดังนั้นจึงสรุปได้อีกว่า หากแม่แบบการแปลใดมีคำที่สกัดส่วนซ้ำและไม่ส่วนไม่ซ้ำผิดแม้เพียงคำเดียวภายในแม่แบบการแปลนั้น คู่แม่แบบการแปลจะผิดทั้งส่วนคงที่และส่วนผันแปรทันที

ผลการทดลองสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 1 ที่มีการตัดคำไม่สามารถสกัดแม่แบบการแปลที่เพียงพอกับการเปรียบเทียบหาคำแปลกับข้อความรับเข้าทั่วไป ผู้วิจัยจึงมีสมมติฐานว่า หากไม่ตัดคำในคลังข้อมูลอาจจะมีผลให้แม่แบบการแปลส่วนคงที่และส่วนผันแปรมีจำนวนลดลงทำให้แม่แบบการแปลมีความถูกต้องมากขึ้นและการทำงานของระบบก็จะรวดเร็วขึ้น ดังนั้นผู้วิจัยจึงได้ทดลองสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 2 ซึ่งเป็นคลังข้อมูลที่ไม่ผ่านการตัดคำ

4.1.2 ผลการทดลองจากคลังข้อมูลชุดทดลองที่ 2

ผลการสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 2 จะได้จำนวนแม่แบบการแปลที่สกัดได้มีจำนวนน้อยกว่าคลังข้อมูลชุดทดลองที่ 1 คือ สกัดได้แม่แบบการแปลจำนวน 489 แม่แบบ ซึ่งเป็นผลจากส่วนของข้อความมีจำนวนน้อยลงเพราะไม่ได้ใช้การตัดคำแต่ให้ระบบตรวจสอบดูที่รูปผิวของแต่ละข้อความเอง ทำให้โอกาสในการรวมกลุ่มคำของกระบวนการรวมกลุ่มคำในขั้นตอนการสร้างต้นไม้การปรากฏร่วมมากขึ้นซึ่งทำให้จำนวนโหนดลูกของต้นไม้การปรากฏร่วมน้อยลงและทำให้ความกำกวมของต้นไม้การปรากฏร่วมน้อยลง แต่ก็ไม่เพียงพอในการสร้างต้นไม้การปรากฏร่วมได้อย่างถูกต้อง ผลลัพธ์ของการสกัดจะพบปัญหาเช่นเดียวกันกับผลการสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 1 คือค้นหาและจับส่วนคงที่และส่วนผันแปรที่ผิดและผลการสร้างต้นไม้การปรากฏร่วมเกิดความกำกวมทำให้ผลการสกัดแม่แบบการแปลส่วนมากไม่สามารถนำมาใช้เปรียบเทียบแปลได้

อย่างไรก็ตามผลการสกัดแม่แบบการแปลที่ถูกต้องและนำไปใช้แปลข้อความได้จากคลังข้อมูลชุดทดลองที่ 2 คือได้ผลการสกัดแม่แบบการแปลส่วนคงที่และส่วนผันแปรเหมือนกับผลลัพธ์การสกัดคลังข้อมูลชุดทดลองที่ 1 ทุกแม่แบบ แต่ได้แม่แบบการแปลส่วนคงที่เพิ่มขึ้นอีก 1 แม่แบบการแปลคือ

{หมายเหตุ : ผู้ลงทุนสามารถศึกษาข้อมูลเกี่ยวกับลักษณะเงื่อนไขและสาระสำคัญของใบสำคัญแสดงสิทธิได้จากสรุปข้อสนเทศของ <X1> ในระบบบริการข้อมูลตลาดหลักทรัพย์ (SETSMART) ↔ Note : Please see the description, condition and major characteristics of <Y1> in SET information Management System (SETSMART)} โดยจับคู่ {<X1> ↔ <Y1>}

ผลการสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 2 ได้แม่แบบการแปลส่วนคงที่ที่ถูกต้องจากการตรวจสอบ จำนวน 9 แม่แบบคิดเป็นร้อยละ 1.84 และแม่แบบการแปลส่วนผันแปรที่ถูกต้องจำนวน 44 แม่แบบคิดเป็นร้อยละ 9 รวมได้แม่แบบการแปลที่ถูกต้องตามหลักจำนวน 53 แม่แบบคิดเป็นร้อยละ 10.84

ตารางที่ 3 แสดงผลการสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 2

ชนิดของแม่แบบการแปล		จำนวน	ร้อยละ	จำนวน	ร้อยละ
แม่แบบการแปลที่ถูกต้อง	ส่วนคงที่	9	1.84	53	10.84
	ส่วนผันแปร	44	9		
แม่แบบการแปลที่ไม่ถูกต้อง				436	89.16
แม่แบบการแปลทั้งหมดที่สกัดได้				489	100

4.1.3 ผลการทดลองจากคลังข้อมูลชุดทดลองที่ 3

เนื่องจากผลการสกัดแม่แบบการแปลจากคลังข้อมูลชุดที่ 1 และ 2 ได้แม่แบบการแปลที่มีปัญหาเพราะไม่สามารถสกัดแม่แบบการแปลที่ถูกต้องตามที่คาดไว้ได้ เพราะปัญหาต่างๆ ของระบบและข้อมูลที่ใช้ทดลองซึ่งปัญหาเหล่านี้จะเอาไปอภิปรายในบทที่ 5 แต่เพื่อให้สามารถทำการทดลองระบบการแปลภาษาต่อได้ ผู้วิจัยจึงทำการจัดกลุ่มข้อมูลเองตามความคล้ายคลึงเพื่อช่วยให้ระบบสกัดแม่แบบการแปลที่จะใช้ได้จริงมากขึ้น ผู้วิจัยจึงทดลองสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 3 ซึ่งมีการแบ่งแฟ้มข้อมูลออกตามความคล้ายคลึงกันของข้อความภายในแต่ละรายงานข่าวเป็น 184 กลุ่ม กล่าวคือผู้วิจัยรวบรวมข้อความที่มีส่วนซ้ำเหมือนกันไว้เป็นแฟ้มข้อมูลเดียวกันและหาแม่แบบจากภายในกลุ่มนั้นเอง ผู้วิจัยจัดข้อความแปลข้างล่างนี้ให้อยู่ในแฟ้มข้อมูลเดียวกัน เช่น

ตลาดหลักทรัพย์เพิ่มสินค้า : AKR ↔ SET adds new listed securities : AKR
ตลาดหลักทรัพย์เพิ่มสินค้า : CCET-W1 ↔ SET adds new listed securities : CCET-W1
ตลาดหลักทรัพย์เพิ่มสินค้า : DSGT ↔ SET adds new listed securities : DSGT
ตลาดหลักทรัพย์เพิ่มสินค้า : FORTH ↔ SET adds new listed securities : FORTH
ตลาดหลักทรัพย์เพิ่มสินค้า : JTS ↔ SET adds new listed securities : JTS
ตลาดหลักทรัพย์เพิ่มสินค้า : MINT-W3 ↔ SET adds new listed securities : MINT-W3
ตลาดหลักทรัพย์เพิ่มสินค้า : RICH ↔ SET adds new listed securities : RICH
ตลาดหลักทรัพย์เพิ่มสินค้า : RRC ↔ SET adds new listed securities : RRC
ตลาดหลักทรัพย์เพิ่มสินค้า : SECC ↔ SET adds new listed securities : SECC
ตลาดหลักทรัพย์เพิ่มสินค้า : STHAI-W1 ↔ SET adds new listed securities : STHAI-W1
ตลาดหลักทรัพย์เพิ่มสินค้า : TOG ↔ SET adds new listed securities : TOG

โดยผลการสกัดจะสามารถสกัดแม่แบบการแปลที่มีความถูกต้องจากการตรวจสอบว่าส่วนซ้ำของคู่ข้อความที่ถูกสกัดเป็นส่วนคงที่และส่วนคงที่นั้นเป็นคู่คำแปลของกันและกัน และส่วนไม่ซ้ำของข้อความที่ถูกสกัดเป็นส่วนผันแปรและส่วนผันแปรนั้นเป็นคู่คำแปลของกันและกันได้อย่างถูกต้อง และแม่แบบการแปลที่ได้ก็สามารถนำไปใช้เปรียบเทียบหาคำแปลได้

ผลการสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 3 ทั้งหมดได้แม่แบบการแปลจำนวน 633 แม่แบบ โดยแบ่งเป็นแม่แบบการแปลส่วนคงที่ที่ถูกต้องจำนวน 165 แม่แบบ คิดเป็นร้อยละ 26.07 และแม่แบบการแปลส่วนผันแปรที่ถูกต้องจำนวน 449 แม่แบบคิดเป็นร้อยละ 70.93 นอกจากนี้ ในการทดลองชุดนี้ยังได้แม่แบบการแปลแบบตัวอย่างซึ่งเป็นแม่แบบการแปลที่สกัดจากข้อความทั้งหมดนั้นเป็นส่วนซ้ำและสามารถนำไปใช้เป็นตัวอย่างในการแปลทั้งข้อความได้ทันทีโดยแม่แบบการแปลแบบตัวอย่างมีจำนวน 19 แม่แบบคิดเป็นร้อยละ 3

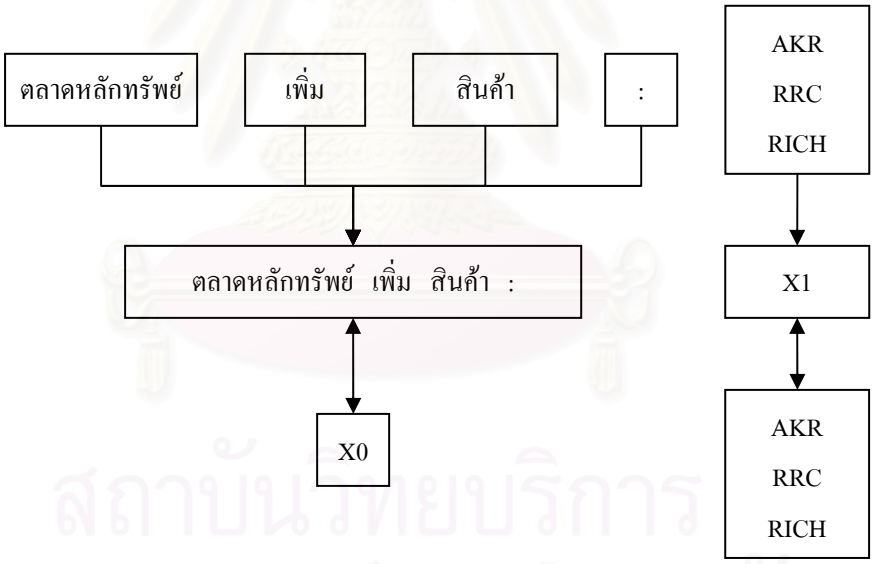
ตารางที่ 4 แสดงผลการสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 3

ชนิดของแม่แบบการแปล		จำนวน	ร้อยละ	จำนวน	ร้อยละ
แม่แบบการแปลที่ถูกต้อง	ส่วนคงที่	165	26.07	633	100
	ส่วนผันแปร	449	70.93		
	แบบตัวอย่าง	19	3		
แม่แบบการแปลที่ไม่ถูกต้อง				0	0
แม่แบบการแปลทั้งหมดที่สกัดได้				633	100

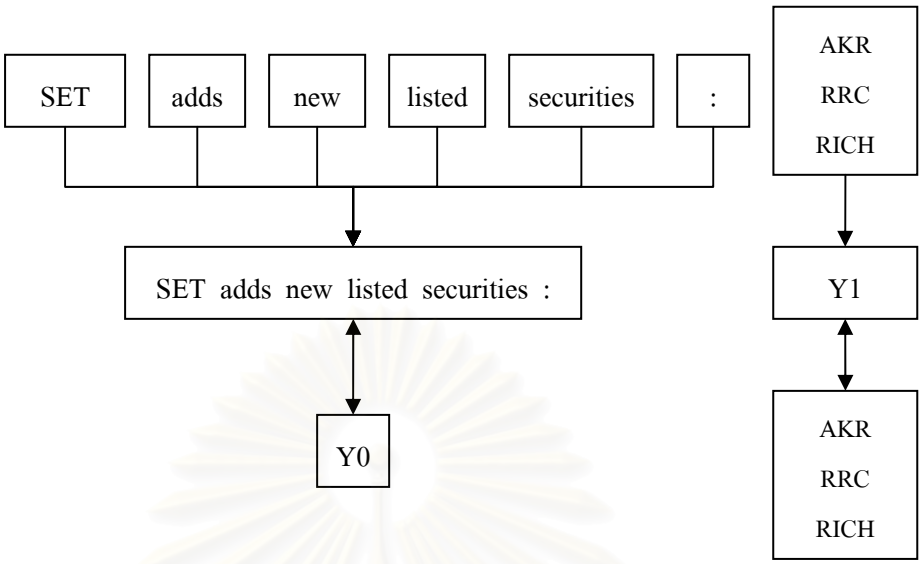
เนื่องจากการแบ่งคลังข้อมูลแบ่งแฟ้มข้อมูลออกตามความคล้ายคลึงกันของข้อความจะสามารถแยกส่วนคงที่และส่วนผันแปรจากส่วนซ้ำและส่วนไม่ซ้ำได้อย่างชัดเจน ทำให้สามารถนำไปสร้างต้นไม้การปรากฏร่วมได้ง่ายและมีความกำกวมน้อยลงมากเมื่อนำไปเปรียบเทียบจากผลการสกัดของคลังข้อมูลชุดทดลองที่ 1 และ2 เช่นการสกัดแม่แบบการแปลจากข้อความตัวอย่าง

ตลาดหลักทรัพย์เพิ่มสินค้า : AKR ↔ SET adds new listed securities : AKR
ตลาดหลักทรัพย์เพิ่มสินค้า : RICH ↔ SET adds new listed securities : RICH
ตลาดหลักทรัพย์เพิ่มสินค้า : RRC ↔ SET adds new listed securities : RRC

จะได้แม่แบบการแปลคือ [(ตลาดหลักทรัพย์) (เพิ่ม) (สินค้า) (:) <X1>] ↔ [(SET) (adds) (new) (listed) (securities) (:) <Y1>] และ [<X0> (AKR)] ↔ [<Y0> (AKR)] และ [<X0> (RICH)] ↔ [<Y0> (RICH)] และ [<X0> (RRC)] ↔ [<Y0> (RRC)] ซึ่งมีความสัมพันธ์ของแม่แบบการแปลดังรูปที่ 27 และรูปที่ 28 ด้านล่าง



รูปที่ 27 แสดงความสัมพันธ์ของแม่แบบการแปลภาษาไทยจากคลังข้อมูลชุดทดลองที่ 3



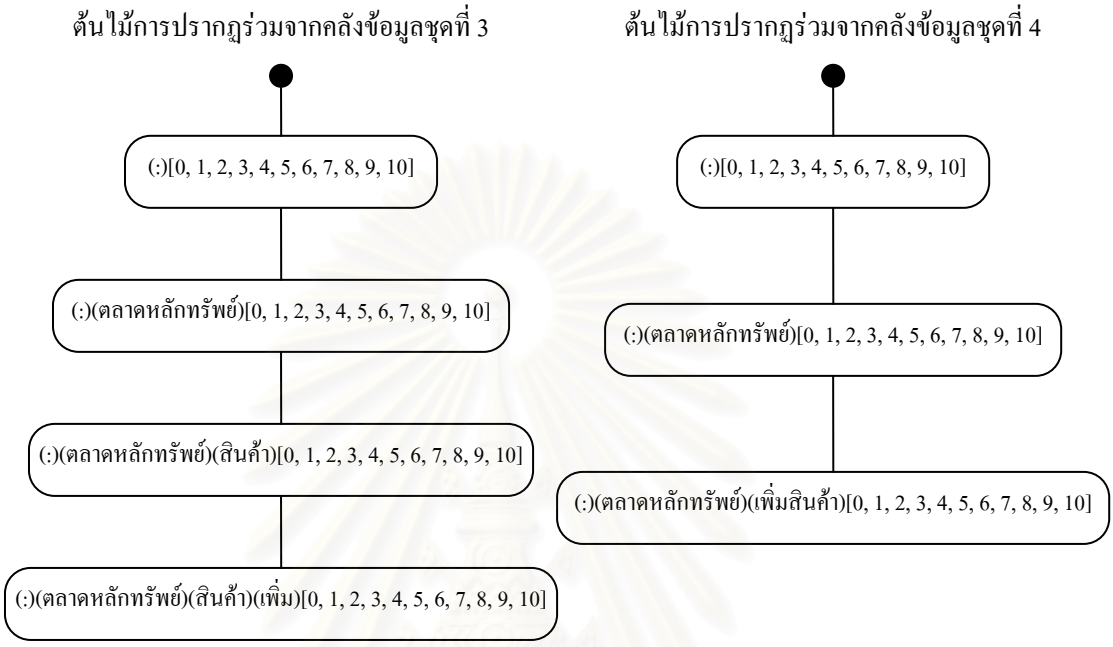
รูปที่ 28 แสดงความสัมพันธ์ของแม่แบบการแปลภาษาอังกฤษจากคลังข้อมูลชุดทดลองที่ 3

ความสัมพันธ์ที่เกิดขึ้นภายในแม่แบบการแปลที่สกัดได้จากคลังข้อมูลชุดทดลองที่ 3 จะเห็นได้ชัดเจนว่าการทำงานของกระบวนการรวมกลุ่มคำที่ปรากฏด้วยกันของส่วนซ้ำสามารถทำได้ถูกต้องและต้นไม้การปรากฏร่วมที่ได้จะมีการเรียงลำดับกันอย่างชัดเจน จึงทำให้ง่ายต่อการค้นหาและจับส่วนคงที่และส่วนผันแปรมาสร้างแม่แบบการแปล เพราะในการส่งเข้าสกัดแบบแยกชุดตามความคล้ายคลึงทุกคำที่ปรากฏจะเกิดซ้ำกันในทุกข้อความและมีความกำกวมในการปรากฏซ้ำซ้อนน้อย และสอดคล้องกับสมมติฐานของแม็คเททมากคือทำให้ส่วนซ้ำและไม่ซ้ำภายในคลังข้อมูลเด่นชัดขึ้น ผลการสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 3 จะได้แม่แบบการแปลที่มีความสมบูรณ์และสามารถนำไปเปรียบเทียบหาคำแปลได้อย่างมีประสิทธิภาพ

4.1.4 ผลการทดลองจากคลังข้อมูลชุดทดลองที่ 4

จากผลการทดลองของคลังข้อมูลชุดทดลองที่ 2 ที่จำนวนของโหนดลูกของต้นไม้การปรากฏร่วมน้อยกว่าผลการทดลองของคลังข้อมูลชุดทดลองที่ 1 ทำให้ผู้วิจัยมีสมมติฐานต่อว่า คลังข้อมูลเทียบบทที่ไม่ได้ตัดคำแต่แบ่งออกเป็นชุดตามความคล้ายคลึงของข้อความซึ่งจะทดลองเป็นคลังข้อมูลชุดทดลองที่ 4 อาจจะให้ผลลัพธ์เหมือนหรือดีกว่าการทดลองกับคลังข้อมูลชุดทดลองที่ 3 จึงทำการทดลองสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 4 และได้ผลการสกัดแม่แบบการแปลที่ได้เหมือนกับการทดลองคลังข้อมูลชุดทดลองที่ 3 แต่ความสัมพันธ์ภายในต้นไม้ของการปรากฏร่วมมีความซับซ้อนน้อยกว่าและทำให้ส่วนซ้ำและไม่ซ้ำภายในคลังข้อมูลเด่นชัดขึ้น โดยดูได้จากความสัมพันธ์ภายในต้นไม้ของการปรากฏร่วมเปรียบเทียบกันระหว่างผลการทดลองคลังข้อมูลชุดทดลองที่ 3 และ 4 ดังเช่นรูปที่ 29 ด้านล่าง ซึ่งจะเห็นได้ว่าการเรียงตัวของ

โหนดลูกของต้นไม้การปรากฏร่วมของคลังข้อมูลชุดทดลองที่ 4 จะน้อยกว่าเพราะปริมาณชุดคำน้อยกว่าชุดคำของคลังข้อมูลชุดทดลองที่ 3 ที่ผ่านการตัดคำ ดังนั้นคลังข้อมูลชุดทดลองที่ 4 จะไม่พบปัญหาการปรากฏซ้ำซ้อนและกำกวมของคำและกลุ่มคำภายในข้อความ



รูปที่ 29 แสดงต้นไม้การปรากฏร่วมเปรียบเทียบระหว่างคลังข้อมูลชุดทดลองที่ 3 และ 4

อย่างไรก็ตาม ถึงแม้ว่าการเรียงตัวของโหนดลูกของต้นไม้การปรากฏร่วมของคลังข้อมูลชุดทดลองที่ 4 จะน้อยกว่า แต่จำนวนแม่แบบการแปลที่สกัดได้ก็เท่ากับจำนวนแม่แบบการแปลที่สกัดได้จากคลังข้อมูลชุดทดลองที่ 3 คือ 633 แม่แบบการแปล โดยแบ่งเป็นแม่แบบการแปลส่วนคงที่ที่ถูกต้องจำนวน 165 แม่แบบและแม่แบบการแปลส่วนผันแปรที่ถูกต้องจำนวน 449 แม่แบบและแม่แบบการแปลแบบตัวอย่างจำนวน 19 แม่แบบ เพราะระบบตรวจสอบส่วนซ้ำและส่วนไม่ซ้ำได้ง่ายจากความคล้ายคลึงของข้อความ ดังนั้นความแตกต่างของคลังข้อมูลชุดทดลองที่ 3 และชุดที่ 4 ที่มีเพียงแค่การตัดคำหรือไม่ตัดคำจึงไม่มีผลต่อปริมาณการสกัดแม่แบบการแปล

ตารางที่ 5 แสดงผลการสกัดแม่แบบการแปลจากคลังข้อมูลชุดทดลองที่ 4

ชนิดของแม่แบบการแปล		จำนวน	ร้อยละ	จำนวน	ร้อยละ
แม่แบบการแปลที่ถูกต้อง	ส่วนคงที่	165	26.07	633	100
	ส่วนผันแปร	449	70.93		
	แบบตัวอย่าง	19	3		
แม่แบบการแปลที่ไม่ถูกต้อง				0	0

แม่แบบการแปลทั้งหมดที่สกัดได้	633	100

จากผลการทดลองสกัดแม่แบบการแปลจากคลังข้อมูลชุดข้อมูลทั้ง 4 ชุดพบว่าผลการสกัดจากคลังข้อมูลชุดที่ 3 และ 4 มีผลลัพธ์เหมือนกัน และเมื่อนำแม่แบบการแปลที่ถูกต้องนี้ไปเปรียบเทียบกับผลการสกัดจากคลังข้อมูลชุดที่ 1 และ 2 พบว่ามีความแตกต่างกันมากทั้งปริมาณแม่แบบการแปลและความถูกต้อง โดยพบว่าผลการสกัดจากคลังข้อมูลชุดที่ 1 ที่นำไปใช้เป็นแปลข้อความได้มีเพียงร้อยละ 9.85 และผลการสกัดจากคลังข้อมูลชุดที่ 2 ที่นำไปใช้เป็นแปลข้อความได้มีเพียงร้อยละ 10.84

จากผลการทดลองสกัดแม่แบบการแปลจากคลังข้อมูล ได้ข้อสรุปว่าการสกัดแม่แบบการแปลจากคลังข้อมูลโดยตรงดังเช่นคลังข้อมูลชุดที่ 1 ให้ผลการสกัดที่ไม่ตรงกับที่คาดไว้และมีความถูกต้องน้อยเพียงร้อยละ 9.85 แต่เมื่อเทียบกับคลังข้อมูลชุดที่ 2 ที่ไม่ได้ตัดคำพบว่าคลังข้อมูลที่ไม่ผ่านระบบตัดคำมีผลลัพธ์ในการค้นหาคำปรากฏร่วมสำหรับสร้างต้นไม้การปรากฏร่วมดีกว่าเล็กน้อยเมื่อผลความถูกต้องของการสกัดเพิ่มขึ้นมาอีกร้อยละ 0.99 เป็นร้อยละ 10.84 เพราะจะช่วยลดขนาดของต้นไม้การปรากฏร่วมลงซึ่งเกิดจากปริมาณโหนดลูกของโหนดรากคำต่างๆ จะลดลง อย่างไรก็ตามผลการสกัดแม่แบบการแปลจากคลังข้อมูลทั้ง 2 ชุดเป็นผลลัพธ์ที่ต่ำกว่าที่คาดไว้ซึ่งอาจเป็นเพราะข้อจำกัดของระบบ ปัญหาของภาษาภายในคลังข้อมูล และพฤติกรรมทางภาษาของภาษาไทยและภาษาอังกฤษที่แตกต่างกัน ดังนั้นจึงต้องนำรายละเอียดดังกล่าวมาวิเคราะห์หาสาเหตุและอภิปรายในบทต่อไป

อย่างไรก็ตาม เมื่อผลการสกัดแม่แบบการแปลจากทั้ง 2 ชุดข้างต้นให้ผลที่ไม่ดีนัก ดังนั้นผู้วิจัยจึงต้องทดลองสกัดแม่แบบการแปลจากคลังข้อมูลชุดที่ 3 และ 4 ซึ่งเป็นคลังข้อมูลชุดที่ผู้วิจัยได้ช่วยระบบโดยการจัดแบ่งข้อมูลไว้โดยเฉพาะ (exemplar) เพื่อทดสอบการทำงานของระบบและจะได้นำแม่แบบการแปลจากการสกัดนี้ไปทดลองในส่วนการทดลองแปลต่อไปได้

อย่างไรก็ตาม สำหรับการทดลองแปลข้อความผู้วิจัยได้เลือกแม่แบบการแปลที่สกัดได้จากคลังข้อมูลชุดทดลองที่ 1 มาเป็นแม่แบบการแปลหลักในงานวิจัยชิ้นนี้เพราะเป็นแม่แบบการแปลที่สกัดได้จากคลังข้อมูลที่ผ่านการตัดแบ่งคำเพียงอย่างเดียวซึ่งเป็นลักษณะของคลังข้อมูลทั่วๆ ไป โดยจะนำแม่แบบการแปลที่ได้นี้ไปทดลองแปล จากนั้นผู้วิจัยจะนำแม่แบบการแปลที่สกัดได้จากคลังข้อมูลชุดทดลองที่ 3 ซึ่งเป็นแม่แบบการแปลที่ควรจะได้ มาทำการทดลองแปลข้อความและนำผลการแปลที่ได้มาเปรียบเทียบกัน

4.2 ผลการทดลองแปลข้อความ

เมื่อสกัดแม่แบบการแปลและเก็บลงฐานข้อมูลเรียบร้อยแล้ว ระบบก็จะสามารถนำแม่แบบการแปลเหล่านั้นมาใช้เป็นตัวอย่างในการเปรียบเทียบแปลข้อความรับเข้าได้ โดยข้อความรับเข้าจะนำมาจากคลังข้อมูลเทียบบทที่สุ่มแยกไว้ 51 คู่รายงานข่าว รวมมีจำนวนข้อความอยู่ 297 ข้อความโดยแบ่งเป็น 1 ข้อความต่อ 1 บรรทัดเช่นเดียวกับข้อมูลที่ใช้ในการสกัดแม่แบบการแปล แต่จะนำเฉพาะส่วนรายงานข่าวภาษาไทยเท่านั้นที่จะใช้ข้อความรับเข้าเพื่อทดลองแปลจากไทยเป็นอังกฤษ และจะประเมินผลแปลที่ได้ใน 2 ลักษณะคือ (1) ตรวจสอบประเมินความถูกต้องของผลแปลว่าเหมือนข้อความภาษาอังกฤษที่มีอยู่ในคลังข้อมูลหรือไม่ และ (2) ตรวจสอบประเมินผลแปลด้วยผู้วิจัยเองว่าผลแปลนั้นยอมรับได้หรือไม่ เพราะเป็นไปได้ว่าระบบอาจแปลไม่เหมือนข้อความภาษาอังกฤษที่มีอยู่แต่ก็เป็นการแปลที่ยอมรับได้ ในการทดลองแปลข้อความรับเข้าทั้งหมดจะใช้แม่แบบการแปล 2 ชุดคือ

(1) แม่แบบการแปลชุดทดลองที่ 1 ซึ่งสกัดได้จากคลังข้อมูลชุดทดลองที่ 1 โดยมีแม่แบบการแปลทั้งหมดจำนวน 528 แม่แบบ แบ่งเป็นแม่แบบการแปลที่ถูกต้องจำนวน 52 แม่แบบ โดยมีแม่แบบการแปลส่วนคงที่จำนวน 8 แม่แบบและแม่แบบการแปลส่วนผันแปรจำนวน 44 แม่แบบ และแม่แบบการแปลที่ไม่ถูกต้องจำนวน 476 แม่แบบ

(2) แม่แบบการแปลชุดทดลองที่ 2 ซึ่งสกัดได้จากคลังข้อมูลชุดทดลองที่ 3 โดยมีแม่แบบการแปลทั้งหมดจำนวน 633 แม่แบบการแปลแบ่งเป็นแม่แบบการแปลส่วนคงที่จำนวน 165 แม่แบบและแม่แบบการแปลส่วนผันแปรจำนวน 449 แม่แบบและแม่แบบการแปลแบบตัวอย่างจำนวน 19 แม่แบบ

4.2.1 การประเมินความถูกต้องจากการเทียบเคียงกับคู่ข้อความต้นฉบับ

ผู้วิจัยได้ประเมินความถูกต้องของผลการแปลโดยใช้คู่รายงานข่าวภาษาอังกฤษเป็นแกนอ้างอิงหลัก กล่าวคือ ผลลัพธ์การแปลที่ได้จากระบบจะถูกเปรียบเทียบกับคู่คำแปลในคลังข้อมูลสองภาษา ผลลัพธ์การแปลจากระบบจะถูกเปรียบเทียบกับคู่เอกสารภาษาอังกฤษในระดับรูปผิวของคำเป็นหลัก หากรูปผิวของผลลัพธ์การแปลเหมือนกับคู่เอกสารภาษาอังกฤษทุกประการ จะนับว่าผลการแปลดังกล่าวถูกต้อง แต่ถ้าหากแตกต่างจากคู่เอกสารภาษาอังกฤษ ก็จะนับว่าผลการแปลไม่ถูกต้อง

ในการประเมินความถูกต้องในส่วนนี้ ผู้วิจัยได้รวบรวมชุดข้อความทดสอบจำนวน 279 อนุพากย์ให้เป็นแฟ้มข้อมูลเดียว โดยจัดรูปแบบเป็น 1 อนุพากย์ต่อ 1 บรรทัด ผู้วิจัยได้

นำชุดข้อความทดสอบดังกล่าวไปทดลองแปลด้วยชุดแม่แบบการแปลที่ 1 (แบ่งกลุ่มข้อความก่อนสกัดแม่แบบการแปล) และชุดแม่แบบการแปลที่ 2 (ไม่แบ่งกลุ่มข้อความก่อนสกัดแม่แบบการแปล) ได้ผลการทดลองดังตารางที่ 6

ตารางที่ 6 แสดงผลการแปลที่เหมือนคู่ต้นฉบับเปรียบเทียบระหว่างชุดทดลองที่ 1 และ 2

	แม่แบบการแปลชุดที่ 1		แม่แบบการแปลชุดที่ 2	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
จำนวนข้อความที่สามารถแปลได้เหมือนคู่ข้อความต้นฉบับ	11	3.70	201	67.68
จำนวนข้อความที่ไม่สามารถแปลได้เหมือนคู่ข้อความต้นฉบับ	286	96.30	96	32.32
รวม	297	100	297	100

เมื่อทดลองด้วยชุดแม่แบบการแปลที่ 1 จะได้ผลการแปลที่เหมือนกันทุกประการกับต้นฉบับภาษาอังกฤษ จำนวน 11 ข้อความ คิดเป็นร้อยละ 3.70 และมีข้อความที่แปลได้ไม่เหมือน จำนวน 286 ข้อความ คิดเป็นร้อยละ 96.30 ในขณะที่เมื่อทดลองด้วยชุดแม่แบบการแปลที่ 2 จะได้ผลการแปลที่เหมือนกันทุกประการกับต้นฉบับ จำนวน 201 ข้อความ คิดเป็นร้อยละ 67.68 และมีข้อความที่แปลได้ไม่เหมือน จำนวน 96 ข้อความ คิดเป็นร้อยละ 32.32

จากการทดลอง ผู้วิจัยพบว่า ถ้าข้อความรับเข้ามีข้อความเหมือนกับแม่แบบการแปล ระบบจะสามารถแปลข้อความเหล่านั้นได้อย่างถูกต้อง ไม่ว่าจะเป็นแม่แบบการแปลจากชุดทดลองที่ 1 หรือชุดทดลองที่ 2 ยกตัวอย่างเช่น สมมติว่าข้อความทดสอบเป็นดังนี้

การรับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียน : SSSC

และภายในฐานข้อมูลแม่แบบการแปลมีแม่แบบการแปล

● [การรับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติม : <X1> ↔ LISTED SECURITIES GRANTED BY THE SET : <Y1>]”
● “[<X0>SSSC ↔ <Y0>SSSC]

ระบบจะสามารถแปลข้อความทดสอบเป็นผลการแปลได้ดังนี้

LISTED SECURITIES GRANTED BY THE SET : SSSC

อย่างไรก็ตาม การที่ระบบไม่สามารถแปลข้อความที่เหลือได้ตรงกับคู่ข้อความต้นฉบับ มีสาเหตุ 2 ประการ คือ (1) ไม่มีข้อความเหล่านั้นเป็นตัวอย่างอยู่ภายในคลังข้อมูลเทียบบท

หรือ (2) มีบางส่วนในข้อความรับเข้าที่ไม่สามารถหาแม่แบบการแปลมาเทียบแปลได้ซึ่งจะทำให้ไม่สามารถเติมเต็มข้อความแปลได้จนครบทั้งข้อความ

ตารางที่ 7 แสดงผลการแปลที่ไม่เหมือนคู่ต้นฉบับเปรียบเทียบระหว่างชุดทดลองที่ 1 และ 2

	แม่แบบการแปลชุดที่ 1		แม่แบบการแปลชุดที่ 2	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ
จำนวนข้อความที่ไม่สามารถหาแม่แบบการแปลมาเทียบแปลได้	97	33.92	13	13.54
จำนวนข้อความที่หาแม่แบบการแปลมาเทียบแปลได้แต่แปลไม่เหมือน	189	66.08	83	86.46
รวมข้อความที่แปลไม่เหมือนคู่ต้นฉบับ	286	100	96	100

ในกรณีที่ 1 ที่ไม่มีข้อความเหล่านั้นเป็นตัวอย่างอยู่ภายในคลังข้อมูลเทียบบทและระบบไม่สามารถค้นหาแม่แบบการแปลใดมาแทนที่ส่วนใดของข้อความได้เลย เช่น

> ตามที่บริษัท ได้มีหนังสือแจ้งเรื่องการจำหน่ายเงินลงทุนในบริษัทร่วมตามที่อ้างถึงข้างต้นนั้น เพิ่มเติม พร้อมกันนี้ ขอแจ้งข้อมูลเพิ่มเติมว่าวันเกิดรายการคือ วันที่ 15 ธันวาคม 2548 และการขายหุ้นในครั้งนี้ไม่มีภาระผูกพันใดๆ

ข้อความนี้ไม่เคยปรากฏอยู่ในคลังข้อมูลและเป็นข้อความที่แตกต่างจากข้อความที่เก็บรวบรวมไว้ในคลังข้อมูลเทียบบท ทำให้ระบบไม่สามารถแปลข้อความนี้ได้เนื่องจากไม่มีตัวอย่างสำหรับการแปล

ผลการแปลโดยใช้แม่แบบการแปลชุดทดลองที่ 1 พบว่าจากข้อความแปลที่แปลไม่เหมือนมีจำนวนทั้งหมดอยู่ 286 ข้อความ จาก 286 ข้อความนั้นมีข้อความจำนวน 97 ข้อความที่ระบบไม่ได้เพราะไม่มีข้อความเหล่านั้นเป็นตัวอย่างอยู่ภายในคลังข้อมูลเทียบบท ความคิดเป็นร้อยละ 33.92 ส่วนผลการแปลโดยใช้แม่แบบการแปลชุดทดลองที่ 2 พบว่าจากข้อความแปลที่แปลไม่เหมือนมีจำนวนทั้งหมดอยู่ 96 ข้อความ จาก 96 ข้อความนั้นมีอยู่จำนวน 13 ข้อความที่ระบบแปลได้ไม่เหมือนเพราะไม่มีข้อความเหล่านั้นเป็นตัวอย่างอยู่ภายในคลังข้อมูลเทียบบทคิดเป็นร้อยละ 4.38 ดังปรากฏอยู่ในตารางที่ 7

ในกรณีที่ 2 ที่ระบบไม่สามารถหาแม่แบบการแปลมาเทียบแปลบางส่วนในข้อความรับเข้าได้ ระบบจะไม่สามารถเติมเต็มข้อความแปลได้จนครบทั้งข้อความและผลการแปลที่ได้จะเป็นผลที่ไม่สมบูรณ์ คือมีข้อความภาษาไทยปนอยู่กับภาษาอังกฤษ ดังนั้นเมื่อนำผลการแปลนี้

ไปเทียบเคียงกับคู่ข้อความต้นฉบับก็จะได้ผลว่ารูปผิวของข้อความแปลไม่เหมือนกับคู่ข้อความต้นฉบับ เช่นข้อความรับเข้าคือ

ตามที่ บริษัทศูนย์บริการเหล็กสยามจำกัด (มหาชน) (SSSC) ได้ดำเนินการเพิ่มทุนจดทะเบียนและขอให้ ตลาดหลักทรัพย์ รับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติมนั้น ตลาดหลักทรัพย์ ได้พิจารณาแล้วเห็นควรกำหนดให้หุ้นเพิ่มทุนของบริษัทดังกล่าวเริ่มทำการซื้อขายใน ตลาดหลักทรัพย์ ได้ตั้งแต่ วันที่ 30 พฤษภาคม 2549 เป็นต้นไป

และมีแม่แบบการแปลที่ใช้แทนที่ได้คือ

[ตามที่ <X1> (<X2>) ได้ดำเนินการเพิ่มทุนจดทะเบียนและขอให้ ตลาดหลักทรัพย์ รับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติมนั้น ตลาดหลักทรัพย์ ได้พิจารณาแล้วเห็นควรกำหนดให้หุ้นเพิ่มทุนของบริษัทดังกล่าวเริ่มทำการซื้อขายใน ตลาดหลักทรัพย์ ได้ตั้งแต่<X3>เป็นต้นไป ↔ Starting from <Y1>, the Stock Exchange of Thailand (SET) allowed the securities of <Y2> (<Y3>) to be listed and traded on the SET after finishing capital increase procedures.]

และระบบได้กำหนดให้ส่วนผันแปร <X1> ↔ <Y2> <X2> ↔ <Y3> <X3> ↔ <Y1> เป็นคู่คำแปลกัน แต่ระบบไม่สามารถค้นหาแม่แบบการแปลที่ใช้แทนที่ส่วนผันแปร “บริษัทศูนย์บริการเหล็กสยามจำกัด (มหาชน)” “SSSC” และ “วันที่ 30 พฤษภาคม 2549” ได้ ดังนั้นผลการแปลจึงได้ข้อความแปลคือ

Starting from วันที่ 30 พฤษภาคม 2549, the Stock Exchange of Thailand (SET) allowed the securities of บริษัทศูนย์บริการเหล็กสยามจำกัด (มหาชน) (SSSC) to be listed and traded on the SET after finishing capital increase procedures.

ซึ่งผลการแปลลักษณะนี้ผลการแปลจากแม่แบบการแปลชุดทดลองที่ 1 มีจำนวน 189 ข้อความจากข้อความที่แปลไม่เหมือนทั้งหมดจำนวน 286 ข้อความคิดเป็นร้อยละ 66.08 และผลการแปลจากแม่แบบการแปลชุดทดลองที่ 2 มีจำนวน 83 ข้อความจากข้อความที่แปลไม่เหมือนทั้งหมดจำนวน 96 ข้อความคิดเป็นร้อยละ 86.46 ดังปรากฏอยู่ในตารางที่ 7

ดังนั้นจึงสรุปผลการทดลองแปลได้ว่าข้อความทั้งหมด 297 ข้อความ โดยใช้การประเมินความถูกต้องจากการเทียบเคียงกับคู่ข้อความต้นฉบับ ผลการแปลจากแม่แบบการแปลชุดทดลองที่ 1 จะได้ข้อความแปลที่เหมือนกับคู่ข้อความต้นฉบับ 11 ข้อความคิดเป็นร้อยละ 3.70 และได้ข้อความแปลที่ไม่เหมือนกับคู่ข้อความต้นฉบับ 286 ข้อความคิดเป็นร้อยละ 96.30 และผล

การแปลจากแม่แบบการแปลชุดทดลองที่ 2 จะได้ข้อความแปลที่เหมือนกับคู่ข้อความต้นฉบับ 201 ข้อความคิดเป็นร้อยละ 67.68 และได้ข้อความแปลที่ไม่เหมือนกับคู่ข้อความต้นฉบับ 96 ข้อความคิดเป็นร้อยละ 32.32

4.2.2 การประเมินความถูกต้องของเนื้อความ

ในการแปลข้อความทั่วไป ข้อความภาษาต้นฉบับ 1 ข้อความสามารถแปลเป็นข้อความที่ยอมรับได้ในภาษาเป้าหมายได้หลายแบบ ผู้วิจัยจึงตรวจสอบประเมินความถูกต้องของการแปลโดยนำผลการแปลข้อความทั้งหมดมาตรวจสอบในเชิงความหมาย ว่าข้อความแปลนั้นสามารถสื่อความหมายได้ตรงกับข้อความต้นฉบับหรือไม่ ถ้าสื่อความหมายได้ก็จะนับว่าแปลได้ถูกต้อง ซึ่งแนวทางการประเมินความถูกต้องของผลการแปลนี้คล้ายกับเกณฑ์การเปรียบเทียบของหลักการวัดความถูกต้องของการแปลภาษาอัตโนมัติ BLEU (Bilingual Evaluation Understudy) ซึ่งมีหลักเกณฑ์อยู่ว่า "The closer a machine translation is to a professional human translation, the better it is." "ยิ่งการแปลภาษาด้วยเครื่องทำงานได้ดีใกล้เคียงกับการแปลด้วยนักแปลมืออาชีพมากขึ้นเท่าไหร่ ก็ยิ่งดีเท่านั้น" (Papineni, 2002) นั่นคือผลลัพธ์การแปลจากระบบจะถูกเปรียบเทียบกับคู่เอกสารภาษาอังกฤษในระดับรูปผิวของคำเป็นหลัก หากรูปผิวของผลลัพธ์การแปลแตกต่างจากคู่เอกสารภาษาอังกฤษ คะแนนความถูกต้องจะลดลงเป็นลำดับสัดส่วนลงไป

อย่างไรก็ตาม เกณฑ์การเปรียบเทียบของผู้วิจัยจะแตกต่างจากเกณฑ์ของ BLEU ตรงที่ในการพิจารณาผลนี้ผู้วิจัยกำหนดเกณฑ์ความยืดหยุ่นในการตัดสินความถูกต้องไว้ 4 ประการ คือ

(1) ยอมรับผลการแปลที่ไม่สามารถแปลเอนทิตีระบุนามได้ ทั้งนี้เพราะระบบที่ใช้ในปัจจุบันจะไม่สามารถแปลเอนทิตีระบุนามที่ไม่ปรากฏอยู่ในฐานข้อมูลแม่แบบการแปลได้ แต่ปัญหาเรื่องการแปลเอนทิตีระบุนามสามารถแก้ไขได้โดยง่าย หากมีการกำหนดให้ทำระบบประมวลผลการรู้จำเอนทิตีระบุนามและแปลเอนทิตีระบุนามต่างหากก่อน เช่นการแปลวันที่หรือเวลา เช่นเมื่อเจอ “วันที่ 30 พฤษภาคม 2549” ก็ทำการแปลเป็น “30 May, 2006” ก่อนได้ทันที เนื่องจากการแปลเอนทิตีระบุนามเหล่านี้มีกฎการแปลโดยสมบูรณ์ ส่วนเอนทิตีระบุนามอื่นๆ ประเภทชื่อเฉพาะ อาจใช้การถอดอักษรไทยเป็นโรมันเพื่อแปลงเป็นภาษาเป้าหมายได้ หรืออาจใช้วิธีการสร้างฐานข้อมูลพิเศษสำหรับชื่อเฉพาะประเภทชื่อบริษัททั้งหมดในตลาดหลักทรัพย์เพื่อให้เครื่องทำการรู้จำและสร้างแม่แบบการแปลของเอนทิตีระบุนามต่างหากไว้ก่อนก็ได้ ดังนั้นการแปลเอนทิตีระบุนามไม่ได้ ผู้วิจัยจะกำหนดให้เป็นผลการแปลที่ยอมรับได้ เพราะเป็นปัญหาการแปลจากข้อจำกัดของคลังข้อมูลและสามารถแก้ไขปัญหาได้ง่าย

(2) ยอมรับผลการแปลกาลในส่วนภาษาอังกฤษของข้อความแปลที่ไม่ถูกต้องได้ ทั้งนี้เพราะในบางครั้งข้อความภาษาไทยอาจไม่มีกริยาช่วยที่บ่งบอกกาล ดังนั้นกริยาในข้อความแปลหากผันไม่ตรงกาลถือว่ายอมรับได้ เช่น การแปลโดยใช้ “do” แทนที่จะเป็น “did” หรือการแปลโดยใช้ “do” แทนที่จะเป็น “will do” เป็นต้น

(3) ยอมรับผลการแปลที่การผันตามพจน์ของประธานในภาษาอังกฤษอาจไม่ถูกต้อง เช่น การแปลโดยใช้ “do” แทนที่จะเป็น “does” เป็นต้น เนื่องจากบางกรณีที่แม่แบบการแปลปรากฏเพียงแค่รูปพหูพจน์ แต่ข้อความรับเข้าภาษาไทยที่มีรูปคำเหมือนกันเป็นเอกพจน์ ทำให้ผลการแปลไม่สามารถผันพจน์

(4) ยอมรับผลการแปลที่ใช้คำกำกับนาม (article) ไม่ถูกต้อง เช่น แปล “`ตลาดหลักทรัพย์แห่งประเทศไทย” เป็น “SET” แทนที่จะใช้ “The SET” ก็ถือว่ายังยอมรับได้เพราะไม่ได้เป็นเนื้อความสำคัญภายในข้อความและสามารถอ่านได้ใจความดังเดิม

ผลการทดลองแปลของแม่แบบการแปลชุดทดลองที่ 1 พบว่าจากข้อความรับเข้าทั้งหมดจำนวน 297 ข้อความได้ข้อความแปลที่ยอมรับได้ตามเกณฑ์เงื่อนไขที่กล่าวมาเพิ่มอีก 8 ข้อความซึ่งเป็นข้อความที่ยอมรับได้ตามเกณฑ์เงื่อนไขข้อแรกคือยอมรับผลการแปลที่ไม่สามารถแปลเอนทิตีระบุนามได้ โดยยังคงใช้แม่แบบการแปลส่วนคงที่เดียวกับที่ใช้แปลข้อความที่แปลได้เหมือนที่มีในคลังข้อมูลจากการตรวจสอบวิธีแรก รวมเป็นจำนวน 19 ข้อความคิดเป็นร้อยละ 6.40 นอกจากนั้นเป็นข้อความที่แปลไม่ได้เลยจำนวน 97 ข้อความคิดเป็นร้อยละ 32.66 และข้อความที่แปลแล้วยอมรับไม่ได้จำนวน 181 ข้อความคิดเป็นร้อยละ 60.94

ตารางที่ 8 แสดงผลการแปลจากความถูกต้องของเนื้อความของชุดทดลองที่ 1

ผลการแปล		จำนวน	ร้อยละ	จำนวน	ร้อยละ
แปลเป็นข้อความที่ยอมรับได้	แปลได้ทั้งข้อความ	11	3.70	19	6.40
	แปลได้บางส่วน	8	2.70		
แปลเป็นข้อความที่ไม่สามารถยอมรับได้				181	60.94
ไม่สามารถแปลข้อความได้				97	32.66
รวม				297	100

ในการแปลข้อความที่แปลแล้วยอมรับไม่ได้เกิดจากแม่แบบการแปลที่ผิด และแม่แบบการแปลเหล่านี้จะสามารถใช้แปลได้แค่บางส่วนของข้อความเท่านั้น เช่น เมื่อข้อความรับเข้าคือ

ตลาดหลักทรัพย์จึงขึ้นเครื่องหมาย H หลักทรัพย์ของบริษัทสำหรับการซื้อขายหลักทรัพย์ รอบ เช้า จนกว่าบริษัทจะเผยแพร่ข้อมูลดังกล่าวอย่างครบถ้วนและทั่วถึง

แต่ผลการแปลข้อความนี้จะได้เป็นข้อความแปลดังนี้

Therefore จึงขึ้นเครื่องหมาย H หลักทรัพย์ของบริษัทสำหรับการซื้อขายหลักทรัพย์ effective ของ วันที่ 5 มีนาคม 2549 first บริษัท clarified เผยแพร่ข้อมูล allow อย่างครบถ้วนและ

ผลการแปลนี้เกิดจากการนำแม่แบบการแปลที่มาเป็นตัวอย่างเปรียบเทียบแปลข้อความรับเข้าที่สามารถแปลได้แค่บางส่วนของข้อความคือ

ตลาดหลักทรัพย์ <X1> รอบ เช้า <X2> จนกว่า <X3> จะ <X4> ดังกล่าว <X5> ทั่วถึง ↔ Therefore <Y1> effective <Y2> first <Y3> clarified <Y4> allow <Y5>

แต่ระบบไม่สามารถค้นหาแม่แบบการแปลมาแทนที่แปลข้อความส่วนที่เหลือได้เพราะแม่แบบการแปลอื่นๆ มีคำนอกเหนือจากข้อความ เช่น แม่แบบการแปล

<X0> จึง ขึ้น เครื่องหมาย H หลักทรัพย์ ของ บริษัท สำหรับ การซื้อขาย หลักทรัพย์ <X1> ของ วันที่ 21 มีนาคม 2549 <X2> บริษัท <X3> เผยแพร่ ข้อมูล <X4> อย่างครบถ้วน และ <X5> ↔ <Y0> , the SET has temporarily halted trading of the company 's securities , <Y1> from the <Y2> trading session of March 21, 2006 until the company has <Y3> or disclosed this information to the SET and <Y4> such information to be disseminated to the public.
<X0> จึง ขึ้น เครื่องหมาย H หลักทรัพย์ ของ บริษัท สำหรับ การซื้อขาย หลักทรัพย์ <X1> ของ วันที่ 31 พฤษภาคม 2549 <X2> บริษัท <X3> เผยแพร่ ข้อมูล <X4> อย่างครบถ้วน และ <X5> ↔ <Y0> , the SET has temporarily halted trading of the company 's securities , <Y1> from the <Y2> trading session of May 31, 2006 until the company has <Y3> or disclosed this information to the SET and <Y4> such information to be disseminated to the public.
<X0> จึง ขึ้น เครื่องหมาย H หลักทรัพย์ ของ บริษัท สำหรับ การซื้อขาย หลักทรัพย์ <X1> ของ วันที่ 1 พฤษภาคม 2549 <X2> บริษัท <X3> เผยแพร่ ข้อมูล <X4> อย่างครบถ้วน และ <X5> ↔ <Y0> , the SET has temporarily halted trading of the company 's securities , <Y1> from the <Y2> trading session of May 1, 2006 until the company has <Y3> or disclosed this information to the SET and <Y4> such information to be disseminated to the public.

แม่แบบการแปลข้างต้นทั้ง 3 แม่แบบนี้ ไม่สามารถนำไปแทนที่แปลข้อความที่เหลือได้ เพราะมีวันที่ (“วันที่ 21 มีนาคม 2549” “วันที่ 31 พฤษภาคม 2549” และ “วันที่ 1 พฤษภาคม 2549”) ปรากฏอยู่ในแม่แบบการแปลอยู่ซึ่งวันที่เหล่านี้ไม่ได้ปรากฏอยู่ในข้อความรับเข้า แม่แบบการแปลข้างต้นทั้ง 3 เหล่านี้จึงไม่ผ่านเกณฑ์ของระบบในการคัดเลือกแม่แบบการแปลเพื่อแปลข้อความรับเข้านี้ ถึงแม้ว่าแม่แบบการแปลข้างต้นทั้ง 3 แม่แบบสกัดได้มาข้อความที่มีเนื้อความเหมือนกันทั้งหมด ยกเว้นส่วนเอนทิตีระบุนามที่เป็นวันที่เท่านั้น แต่เนื่องจากระบบค้นหาและตรวจสอบส่วนซ้ำและส่วนไม่ซ้ำของข้อความผิด จึงทำให้แม่แบบการแปลที่สกัดมาผิดและไม่อาจนำมาใช้แปลส่วนที่เหลือของข้อความรับเข้า ผลการแปลลักษณะดังกล่าวนี้เกิดขึ้นเป็นจำนวนมากถึงร้อยละ 60.94 และเป็นผลการแปลที่ยอมรับไม่ได้เพราะไม่สามารถสื่อความได้ นอกจากนั้นยังมีข้อความรับเข้าจำนวนร้อยละ 32.66 ที่ระบบไม่สามารถหาแม่แบบการแปลมาเทียบแปลได้ ดังนั้นผลการแปลโดยใช้แม่แบบการแปลชุดทดลองที่ 1 จึงมีความถูกต้องต่ำเพียงร้อยละ 6.40 เท่านั้น

ผลการทดลองแปลของแม่แบบการแปลชุดทดลองที่ 2 พบว่าจากข้อความรับเข้าทั้งหมดจำนวน 297 ข้อความนอกเหนือจากข้อความที่แปลได้เหมือนที่มีในคลังข้อมูลจากการตรวจสอบวิธีแรกจำนวน 201 ข้อความแล้ว ข้อความที่แปลไม่ตรงกับที่มีในคลังข้อมูลมีจำนวน 83 ข้อความที่ยอมรับได้ตามเกณฑ์เงื่อนไขที่กล่าวมา เมื่อรวมข้อความที่แปลทั้งสองประเภทจึงมีจำนวน 284 ข้อความคิดเป็นร้อยละ 95.62 ส่วนข้อความที่ระบบไม่สามารถแปลได้มีจำนวน 13 ข้อความคิดเป็นร้อยละ 4.38 และไม่มีข้อความใดที่ระบบแปลได้แล้วยอมรับไม่ได้

สาเหตุที่ระบบไม่สามารถแปลข้อความได้จำนวน 13 ข้อความเป็นเพราะว่าข้อความรับเข้าทั้ง 13 ข้อความนี้เป็นข้อความที่ไม่เคยปรากฏในคลังข้อมูลเทียบบท และมีคำไม่ตรงกับแม่แบบการแปลที่สกัดมาได้ ดังนั้นระบบจึงไม่สามารถแปลได้เนื่องจากไม่มีตัวอย่างสำหรับการแปล ซึ่งเป็นข้อจำกัดของระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง

ตารางที่ 9 แสดงผลการแปลจากความถูกต้องของเนื้อความของชุดทดลองที่ 2

ผลการแปล		จำนวน	ร้อยละ	จำนวน	ร้อยละ
แปลเป็นข้อความที่ยอมรับได้	แปลได้ทั้งข้อความ	201	67.68	284	95.62
	แปลได้บางส่วน	83	27.94		
แปลเป็นข้อความที่ไม่สามารถยอมรับได้				0	0
ไม่สามารถแปลข้อความได้				13	4.38
รวม				297	100

จากผลการทดลองแปลข้อความพบว่า ข้อความแปลที่มาจากรายงานตลาดหุ้นแบบรายวัน ระบบสามารถแทนที่ส่วนของข้อความซึ่งเป็นส่วนยาวของข้อความได้ร้อยละ 95.62 เช่นข้อความรับเข้าคือ

ตามที่ตลาดหลักทรัพย์ได้ขึ้นเครื่องหมาย NP (Notice Pending) หลักทรัพย์ของบริษัท คอมพาสส์ อีสต์ อินดัสตรี้ (ประเทศไทย) จำกัด (มหาชน) (CEI) สำหรับการซื้อขายหลักทรัพย์ตั้งแต่วันที่ 16 มิถุนายน 2549

และมีแม่แบบการแปลที่ใช้แทนที่ได้คือ

ตามที่ ตลาดหลักทรัพย์ ได้ ขึ้น เครื่องหมาย NP (Notice Pending) หลักทรัพย์ ของ บริษัท<X1> (<X2>) สำหรับ การ ซื้อขาย หลักทรัพย์ ตั้งแต่<X3> ↔ Previously, the SET has posted the "NP" (Notice pending) sign on <Y1> (<Y2>) effective from <Y3>.

และเมื่อระบบสามารถแทนที่ส่วนยาวของข้อความรับเข้าได้แล้ว ส่วนที่คงเหลืออยู่จึงมีเพียงแค่เอนทิตีระบุนามเท่านั้นซึ่งตามเกณฑ์ความยืดหยุ่นถือว่ายอมรับได้ ดังนั้นผลการแปลร้อยละ 95.62 จึงเป็นผลการแปลที่ยอมรับได้ในส่วนการประเมินความถูกต้องด้วยตนเอง และมีผลการแปลเป็นข้อความที่ไม่สามารถแปลได้อีกร้อยละ 4.38 เพราะไม่มีแม่แบบการแปลมาแทนที่เพื่อใช้แปลข้อความนั้น

เมื่อเปรียบเทียบผลการแปลของแม่แบบการแปลทั้ง 2 ชุดจะพบว่าความถูกต้องของการแปลมีความแตกต่างกันมากดังตารางที่ 11

ตารางที่ 10 แสดงผลการแปลเปรียบเทียบ

	ข้อความแปลที่ยอมรับได้		ข้อความแปลที่ยอมรับไม่ได้		ข้อความที่แปลไม่ได้	
	จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ
แม่แบบการแปลชุดที่ 1	19	3.70	181	60.94	97	32.66
แม่แบบการแปลชุดที่ 2	284	95.62	0	32.32	13	4.38

เมื่อทำการเปรียบเทียบผลการแปลจากแม่แบบการแปลทั้ง 2 ชุดจะพบว่าแม่แบบการแปลชุดทดลองที่ 2 ซึ่งเป็นแม่แบบการแปลที่ระบบสกัดจากชุดตัวอย่างที่จัดไว้เฉพาะและเป็นตัวอย่างที่ดี (exemplar) จะได้ผลการแปลที่ดีกว่ามากถึงร้อยละ 91.92 ดังนั้นจึงสรุปได้ว่าแม่แบบการแปลเป็นปัจจัยสำคัญในการแปลข้อความของระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง หากข้อความรับเข้ามีส่วนของข้อความตรงกับแม่แบบการแปลก็จะแปลได้และข้อความ

แปลก็จะเป็นข้อความที่ยอมรับได้ ดังนั้นหากสกัดแม่แบบการแปลครอบคลุมข้อความรับเข้าก็จะได้ผลการแปลครบถ้วน

ผลการทดลองแปลข้อความบ่งบอกว่าหากระบบสามารถค้นหาแม่แบบการแปลจากฐานข้อมูลได้ ระบบก็จะสามารถแปลได้ดังนั้นการพัฒนาระบบให้มีผลการแปลถูกต้องมากขึ้นจึงขึ้นอยู่กับลักษณะของคลังข้อมูลและความหลากหลายและครอบคลุมของข้อความภายใน ซึ่งจะนำมาสร้างแม่แบบการแปล หากระบบสามารถสกัดแม่แบบการแปลได้ถูกต้องและมีแม่แบบการแปลคลอบคลุมข้อความรับเข้า ระบบการแปลภาษาด้วยเครื่องก็จะสามารถแปลข้อความรับเข้านั้นได้อย่างแน่นอน ดังนั้นสิ่งสำคัญของระบบการแปลนี้จึงตกไปอยู่ที่การสกัดแม่แบบการแปลจากคลังข้อมูลเทียบบท แต่ผลการสกัดแม่แบบการแปลของระบบให้ผลลัพธ์ไม่ดีเท่าที่ควรซึ่งเกิดจากองค์ประกอบหลายส่วน ซึ่งจะทำการวิเคราะห์ปัญหาเหล่านั้นทีละส่วนในบทต่อไป

บทที่ 5

วิเคราะห์ปัญหาของระบบ

ผลการทดลองทำงานพบว่าความถูกต้องของการสกัดแม่แบบการแปลเป็นปัจจัยสำคัญต่อความถูกต้องของระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง กล่าวคือถ้าระบบสามารถสกัดแม่แบบการแปลได้ถูกต้องก็จะสามารถนำไปเป็นตัวอย่างในการเทียบแปลได้ อย่างไรก็ตามปัญหาที่พบของระบบคือไม่สามารถสกัดแม่แบบการแปลที่ใช้งานได้จากคลังข้อมูลเทียบบททั่วไปที่ไม่มีการจัดแบ่งเป็นกรณีพิเศษ เช่นคลังข้อมูลชุดทดลองที่ 1 ซึ่งมีการตัดแบ่งคำเท่านั้นและจับคู่เทียบบทข้อความที่เป็นคำแปลของกันและกันเท่านั้น

จากการตรวจสอบการทำงานของระบบพบว่า ระบบสามารถสร้างต้นไม้การปรากฏร่วมของคู่ภาษาไทยและภาษาอังกฤษจากคลังข้อมูลได้แต่ผลที่ได้ไม่สามารถใช้แปลได้ ซึ่งขั้นตอนการสร้างต้นไม้การปรากฏร่วมเป็นขั้นตอนที่สำคัญที่สุดในการค้นหาส่วนซ้ำและส่วนไม่ซ้ำของคู่ข้อความเพื่อจะสร้างเป็นแม่แบบการแปลซึ่งปัญหาที่ระบบไม่สามารถสกัดแม่แบบการแปลได้แบ่งได้ 3 อย่างคือ (1) ปัญหาของระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง (2) ปัญหาอันเกิดจากลักษณะภาษาของคลังข้อมูลเทียบบท และ (3) ปัญหาของทางไวยากรณ์การแปลภาษาไทยเป็นอังกฤษ

5.1 ปัญหาของระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง

ปัญหาที่ทำให้ระบบไม่สามารถสกัดแม่แบบการแปลจากคลังข้อมูลที่ไม่ได้ทำการตัดแบ่งชุดข้อมูลตามความคล้ายคลึงของข้อความเช่นคลังข้อมูลชุดทดลองที่ 1 และ 2 ได้ถูกต้องเกิดจากสมมติฐานเบื้องต้นคือคลังข้อมูลเทียบบทต้องมีส่วนซ้ำและส่วนไม่ซ้ำอย่างเด่นชัดภายในข้อความที่คล้ายคลึงกัน ระบบสกัดแม่แบบการแปลจะค้นหาส่วนซ้ำและส่วนไม่ซ้ำของข้อความภายในคลังข้อมูลแต่ละภาษาจากการสร้างต้นไม้การปรากฏร่วม โดยกำหนดให้ส่วนซ้ำคือโหนดใบของต้นไม้การปรากฏร่วมซึ่งจะเป็นกลุ่มของคำที่ปรากฏร่วมกันมากที่สุดของข้อความนั้น แล้วจึงนำโหนดใบมาจับคู่เทียบสองภาษาโดยตรวจสอบจากเลขบรรทัดที่ตรงกัน อย่างไรก็ตามเมื่อหมายเลขบรรทัดไม่ตรงกัน ระบบก็จะไม่สามารถจับคู่ข้อความได้ ซึ่งมีความเป็นไปได้ที่จะเกิดหมายเลขบรรทัดไม่ตรงกันในกรณีที่คำในภาษาต้นฉบับสามารถมีคู่คำแปลในภาษาเป้าหมายมากกว่า 1 คำแปล ยกตัวอย่างเช่นคำว่า “ตลาดหลักทรัพย์” ที่มีคู่คำแปลเป็นได้ทั้ง “The SET” และ “Stock Exchange of Thailand” และปรากฏสลับกันไปมาในคู่ข้อความที่เหมือนกัน เช่น “ตลาดหลักทรัพย์เพิ่มสินค้า... ↔ The SET adds new listed securities...[1,4]” และ “ตลาดหลักทรัพย์เพิ่ม

สินค้า... ↔ Stock Exchange of Thailand adds new listed securities...[2,3]” ทำให้มีการปรากฏของกลุ่มคำ “ตลาดหลักทรัพย์เพิ่มสินค้า...” ในบรรทัดหมายเลขที่ 1 2 3 และ 4 แต่การปรากฏของ “The SET adds new listed securities...” มีแค่ในบรรทัดหมายเลขที่ 1 และ 4 และ “Stock Exchange of Thailand adds new listed securities...” มีแค่ในบรรทัดหมายเลขที่ 2 และ 3 ดังนั้นเมื่อกลุ่มของหมายเลขบรรทัดไม่ตรงกันและไม่เท่ากัน ระบบก็จะไม่สามารถจับคู่ข้อความได้

นอกจากนี้การปรากฏของคำบางคำบ่อยครั้งในปริมาณข้อความจำนวนมาก เช่น คำภาษาไทย “ตลาดหลักทรัพย์” ทำให้ต้นไม้การปรากฏร่วมของคำนั้นมีขนาดใหญ่โตและซับซ้อนมาก และเป็นผลให้ส่วนซ้ำของคำอื่นๆ ภายในข้อความไม่เด่นชัดเท่ากับคำที่ปรากฏบ่อยครั้ง ดังนั้นโหนดใบบางโหนดจึงมีส่วนซ้ำอยู่ถึง 2 ถึง 3 ส่วน เมื่อต้องการแปลข้อความที่มีส่วนซ้ำที่ไม่เด่นชัดเท่านั้น ก็จะไม่สามารถแปลได้ เพราะในโหนดใบปรากฏคำที่ปรากฏบ่อยรวมอยู่ด้วย เช่น เมื่อต้องการแปลขอความรับเข้า “บริษัทศูนย์บริการเหล็กสยามจำกัด (มหาชน) เพิ่มสินค้า” ก็จะไม่สามารถแปลได้ เพราะคำ “ตลาดหลักทรัพย์” ปรากฏมากที่สุดจึงถูกกำหนดให้เป็นโหนดราก และกลุ่มคำ “เพิ่มสินค้า” ปรากฏร่วมกับคำว่า “ตลาดหลักทรัพย์” บ่อยครั้ง จนทำให้โหนดใบของกลุ่มคำที่มี “เพิ่มสินค้า” เป็นโหนดลูกของโหนดราก “ตลาดหลักทรัพย์” ทั้งหมด ดังนั้นระบบจึงไม่สามารถสร้างโหนดใบของ “เพิ่มสินค้า” ที่ไม่มีคำว่า “ตลาดหลักทรัพย์” ได้

และโหนดใบของคำที่ปรากฏบ่อยครั้งก็จะพบว่ามีคำอื่นปรากฏร่วมด้วยจำนวนมาก ซึ่งอาจเป็นคำประเภทชื่อเฉพาะเช่นชื่อบริษัทที่ควรเป็นส่วนไม่ซ้ำเพราะสามารถแปรเปลี่ยนเป็นชื่อบริษัทอื่นได้ แต่บังเอิญปรากฏร่วมกับคำนั้นซ้ำกันมากกว่า 1 ครั้ง จึงทำให้ระบบเก็บรวบรวมชื่อบริษัทนั้นเป็นส่วนหนึ่งของส่วนซ้ำ และนำคำอื่นที่ไม่ได้ปรากฏซ้ำในข้อความที่คล้ายคลึงกันเป็นส่วนไม่ซ้ำแทน และระบบจะไม่นำข้อความที่คล้ายคลึงอื่นที่ไม่มีชื่อบริษัทข้างต้นมาสร้างโหนดใบอีกโหนดเพราะมีส่วนซ้ำและส่วนไม่ซ้ำไม่ตรงกัน เช่น คำภาษาไทย “ตลาดหลักทรัพย์” ที่ปรากฏร่วมกับชื่อเฉพาะ “บริษัทปิคนิคคอร์ปอเรชั่นจำกัด(มหาชน)” ถึง 3 ครั้ง ในบรรทัดหมายเลข 557 559 และ 567 นอกจากนั้นยังมีคำอื่นๆ ที่ปรากฏร่วมอีกด้วย เช่น “งบการเงิน” “ฉบับ” “สอบทาน” “สิ้นสุด” และ “เนื่องจาก” เป็นต้นดังเช่นรูปที่ 30 แต่คำอื่นๆ ที่กล่าวมาเหล่านี้จะไม่ถูกนำไปสร้างโหนดใบอื่นๆ อีก ดังนั้นหากต้องการแปลคำที่กล่าวมานี้ ต้องปรากฏร่วมกับ ชื่อเฉพาะ “บริษัทปิคนิคคอร์ปอเรชั่นจำกัด(มหาชน)” เท่านั้น ทำให้ตรงส่วนนี้จะเกิดปัญหา 2 จุดคือ(1) ระบบจะไม่สามารถแปลคำอื่นๆ ที่กล่าวมานี้ได้เพราะไม่มีโหนดใบที่จะแปลงเป็นแม่แบบการแปลที่ไม่มีชื่อเฉพาะอยู่ และ (2) ไม่สามารถแปลชื่อเฉพาะ “บริษัทปิคนิคคอร์ปอเรชั่นจำกัด(มหาชน)” ได้หากไปปรากฏเป็นส่วนไม่ซ้ำในข้อความรับเข้าที่จะแปลได้เพราะชื่อเฉพาะนี้ไม่ได้ถูกจัดเก็บเป็นแม่แบบส่วนผันแปร

(ตลาดหลักทรัพย์)[0,1,2,...]
(2)(ตลาดหลักทรัพย์)[22,28,35,...]
(2)(งบการเงิน)(ตลาดหลักทรัพย์)[453,454,525,...]
(2)(งบการเงิน)(ตลาดหลักทรัพย์)(สิ้นสุด)[453,454,525,...]
(2)(งบการเงิน)(ตลาดหลักทรัพย์)(สิ้นสุด)(ไตรมาส)[453,454,525,...]
(2)(งบการเงิน)(ตลาดหลักทรัพย์)(สิ้นสุด)(เนื่องจาก)(ไตรมาส)[453,454,557,...]
(2)(งบการเงิน)(ตลาดหลักทรัพย์)(วันที่ 30มิถุนายน 2549)(สิ้นสุด)(เนื่องจาก)(ไตรมาส)[453,557,559,...]
(2)(งบการเงิน)(ตลาดหลักทรัพย์)(วันที่ 30มิถุนายน 2549)(สอบทาน)(สิ้นสุด)(เนื่องจาก)(ไตรมาส)[557,559,567,...]
(2)(งบการเงิน)(ฉบับ)(ตลาดหลักทรัพย์)(วันที่ 30มิถุนายน 2549)(สอบทาน)(สิ้นสุด)(เนื่องจาก)(ไตรมาส)[557,559,567,...]
(2)(PICNI)(งบการเงิน)(ฉบับ)(ตลาดหลักทรัพย์)(วันที่ 30มิถุนายน 2549)(สอบทาน)(สิ้นสุด)(เนื่องจาก)(ไตรมาส)[557,559,567]
(2)(PICNI)(งบการเงิน)(ฉบับ)(ตลาดหลักทรัพย์)(บริษัทปิคนิคคอร์ปอเรชั่นจำกัด(มหาชน))(วันที่ 30มิถุนายน 2549)(สอบทาน)(สิ้นสุด)(เนื่องจาก)(ไตรมาส)[557,559,567]
...

รูปที่ 30 แสดงตัวอย่างต้นไม้การปรากฏร่วมของคำ “ตลาดหลักทรัพย์”

นอกจากนี้ข้อความภายในคลังข้อมูลที่ยาวจะมีปริมาณคำอยู่มากทำให้โหนดลูกมีปริมาณมากและส่งผลให้โหนดใบมีคำที่ปรากฏร่วมกันจำนวนมากอยู่ภายใน และระบบจะนำโหนดใบนั้นมาจับคู่เทียบสองภาษาเป็นแม่แบบการแปล ซึ่งจะเป็นปัญหากับการนำไปใช้แปลข้อความรับเข้าที่มีรูปแบบข้อความและคำภายในข้อความคล้ายคลึงกันแต่จำนวนคำอาจจะขาดหายไปหนึ่งคำ แม่แบบการแปลนั้นก็จะใช้ไม่ได้เพราะไม่ผ่านเงื่อนไขในขั้นตอนค้นหาแม่แบบการแปลว่าทุกคำของแม่แบบการแปลต้องมีอยู่ในข้อความรับเข้า และทำให้ระบบไม่สามารถแปลข้อความรับเข้านั้นได้ เช่น หากมีแม่แบบการแปล

[ตามที่ <X1> (<X2>) ได้ ดำเนินการ เพิ่ม ทุน จดทะเบียน และ ขอให้ ตลาดหลักทรัพย์ รับ หุ้น เพิ่ม ทุน เป็น หลักทรัพย์จดทะเบียนเพิ่มเติม นั้น ตลาดหลักทรัพย์ ได้ พิจารณา แล้ว เห็นควร กำหนด ให้ หุ้น เพิ่ม ทุน ของ บริษัท ดังกล่าว เริ่ม ทำ การ ซื้อขาย ใน ตลาดหลักทรัพย์ ได้ ตั้งแต่ <X3> เป็นต้นไป ↔ Starting from <Y1>, the Stock Exchange of Thailand (SET) allowed the securities of <Y2> (<Y3>) to be listed and traded on the SET after finishing capital increase procedures.]

และต้องการแปลข้อความ “ตามที่บริษัทไทยฮีทเอ็กซ์เช้นจ์จำกัด (มหาชน) (THEX) ได้ดำเนินการเพิ่มทุนจดทะเบียนและขอให้ตลาดหลักทรัพย์รับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติมนั้น

ตลาดหลักทรัพย์ได้พิจารณาแล้วเห็นควรกำหนดให้หุ้นเพิ่มทุนของบริษัทดังกล่าวเริ่มซื้อขายในตลาดหลักทรัพย์ได้ตั้งแต่วันที่ 5 สิงหาคม 2549 เป็นต้นไป" จากข้อความรับเข้าดังกล่าว ระบบไม่สามารถแปลข้อความนั้นได้เพราะข้อความรับเข้านี้ไม่สามารถใช้แม่แบบการแปลข้างต้นได้ เนื่องจากข้อความรับเข้านี้ขาดคำไป 2 คำ คือ คำว่า "ทำ" "การ" เมื่อเทียบกับในแม่แบบการแปล [... เห็นควรกำหนดให้หุ้นเพิ่มทุนของบริษัทดังกล่าวเริ่ม**ทำการ**ซื้อขายในตลาดหลักทรัพย์ได้...] ซึ่งทำให้แม่แบบการแปลไม่ถูกคัดเลือกมาใช้ในการแปลข้อความรับเข้านี้ กล่าวได้ว่าปัญหาส่วนนี้เกิดจากการเลือกนำเฉพาะโหนดใบมาใช้เป็นแม่แบบการแปล

นอกจากนี้ยังพบปัญหาการแปลแบบไม่ครบเนื้อความซึ่งเกิดจากบางครั้งคู่ข้อความที่เป็นคนละภาษาอาจจะมีวัฒนธรรมในการสื่อความไม่ตรงกัน เช่น ผู้ใช้ภาษาไทยมักนิยมใช้การบรรยายที่แบบละเอียดและใช้หน่วยกริยาเรียง ส่วนผู้ใช้ภาษาอังกฤษมักนิยมใช้คำกริยาเดียวที่สื่อความหมายครอบคลุมเป็นต้น ซึ่งจะทำให้เกิดคำหรือกลุ่มของคำที่ไม่มีคู่คำแปลและจะทำให้ระบบจับคู่คำแปลกันผิด เช่นจากคู่ข้อความ

ตามที่บริษัทโรงพยาบาลบำรุงราษฎร์จำกัด (มหาชน) (BH) ได้ดำเนินการเพิ่มทุนจดทะเบียนและขอให้ตลาดหลักทรัพย์รับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติมนั้นตลาดหลักทรัพย์ได้พิจารณาแล้วเห็นควรกำหนดให้หุ้นเพิ่มทุนของบริษัทดังกล่าวเริ่มทำการซื้อขายในตลาดหลักทรัพย์ได้ตั้งแต่วันที่ 4 พฤษภาคม 2549 เป็นต้นไป ↔ Starting from 4 May 2006 , the Stock Exchange of Thailand (SET) allowed the securities of Bumrungrad Hospital Public Company Limited (BH) to be listed and traded on the SET after finishing capital increase procedures.

จะเห็นได้ว่าข้อความส่วนภาษาไทยคือ "ตามที่บริษัทโรงพยาบาลบำรุงราษฎร์จำกัด (มหาชน) (BH) ได้ดำเนินการเพิ่มทุนจดทะเบียนและขอให้ตลาดหลักทรัพย์รับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติมนั้น" นั้นไม่มีคู่คำแปลเลยในข้อความส่วนภาษาอังกฤษ และกริยาวลี "ได้พิจารณาแล้วเห็นควรกำหนดให้" มีคำจำนวนมากแต่มีคู่คำแปลแค่เพียง "allowed" เท่านั้น เหตุการณ์เช่นนี้จะทำให้ระบบที่ไม่สามารถจับคู่ข้อความแบบไร้คู่ได้ ต้องจับคู่คำแปลของ "ได้พิจารณาแล้วเห็นควรกำหนดให้" ให้เป็นคู่คำแปลของ "allowed" ซึ่งในบางครั้งการปรากฏของคำส่วนภาษาไทยในข้อความอื่น สามารถสามารถใช้เพียงแค่คำกริยา "กำหนดให้" เท่านั้น ซึ่งจะทำให้ระบบจับคู่คำแปลแค่เพียง "กำหนดให้ ↔ allowed" และ "ได้พิจารณาแล้วเห็นควร" ก็จะไปรวมกับส่วนข้อความอื่นแทน เป็น "ตลาดหลักทรัพย์ได้พิจารณาแล้วเห็นควร <X1> หุ้นเพิ่มทุน ↔ the Stock Exchange of Thailand (SET) <Y1> the securities" โดยที่ <X1> ↔ <Y1> หากพิจารณาจากแม่แบบการแปลนี้แล้วจะเห็นได้ว่ากลุ่มคำ "ได้พิจารณาแล้วเห็นควร" ไม่ควรไปรวมกับ "ตลาดหลักทรัพย์" เพราะใน

คู่คำแปลไม่มีส่วนใดมีความเกี่ยวพันกับกลุ่มคำนี้เลย ดังนั้นการที่ระบบไม่อนุญาตให้เกิดการจับคู่แบบไร้คู่ ก็จะทำให้เกิดปัญหาเหล่านี้ตามมาและจะทำให้ความถูกต้องของการสร้างแม่แบบการแปลและการแปลข้อความลดลงตามลำดับ

ดังนั้นอาจสรุปได้ว่าระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างมีข้อจำกัดคือเหมาะกับคลังข้อมูลที่สามารถแยกส่วนซ้ำและส่วนไม่ซ้ำของข้อความได้อย่างชัดเจนกล่าวคือไม่จำเป็นต้องมีขนาดใหญ่แต่ควรมีความคล้ายคลึงของข้อความมาก และคำทุกคำควรต้องมีคู่คำแปลของกันและกันอย่างชัดเจน นอกจากนั้นระบบนี้ไม่เหมาะกับคลังข้อมูลที่มีคำเดียวกันปรากฏบ่อยครั้งในข้อความจำนวนมากแต่แปลได้หลากหลาย และคลังข้อมูลที่มีข้อความยาวและมีลักษณะซับซ้อน

5.2 ปัญหาอันเกิดจากลักษณะภาษาของคลังข้อมูลเทียบบท

ระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างมีส่วนสำคัญหลักอยู่ที่ตัวอย่างของภาษาที่จะนำมาสร้างแม่แบบการแปล คลังข้อมูลเทียบบทจึงเป็นฐานข้อมูลที่สำคัญสำหรับระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างที่จะขาดไปไม่ได้

อย่างไรก็ตามคลังข้อมูลเทียบบทภาษาเฉพาะทาง ภาษาตลาดหุ้นที่ผู้วิจัยได้เก็บรวบรวมมาจากรายงานตลาดหุ้นแบบรายวันที่นำมาใช้ในงานวิจัยชิ้นนี้มีลักษณะการแปลของคู่ภาษาเป็นการแปลที่เน้นเจตนา จุดประสงค์ และเป้าหมายของเนื้อความจากภาษาอังกฤษเป็นภาษาไทยตามแนวทางหลักการแปลที่เน้นวัฒนธรรมปลายทางที่จะคงเจตนา จุดประสงค์ และเป้าหมายของเนื้อความเท่านั้น ไม่ได้เน้นการแปลแบบรักษาโครงสร้างภาษาให้ใกล้เคียงภาษาต้นฉบับ ทำให้คำศัพท์ในคู่ข้อความแปลไม่สามารถจับคู่กันได้โดยสมบูรณ์ การจับคู่คำแปลจึงมีความหลากหลายและเกิดความกำกวมในการสร้างต้นไม้การปรากฏร่วม

ด้วยเหตุนี้รายงานตลาดหุ้นแบบรายวันที่นำมาใช้ในงานวิจัยชิ้นนี้จึงมีความหลากหลายในการใช้คู่คำแปลที่แตกต่างกัน เช่น ข้อความภาษาไทยคือ “จัดสรรให้ : ใบสำคัญแสดงสิทธิ” มีคู่ข้อความแปลถึง 2 แบบคือ “Allocate to : Warrants” และ “Allocation : Warrants” แต่ในบางครั้ง “Allocation : Warrants” ก็มีคู่ข้อความแปลคือ “การจัดสรร : ใบสำคัญแสดงสิทธิ” ทำให้คู่คำแปลมีความหลากหลายและเกิดความกำกวม

นอกจากนั้นความแตกต่างกันทางวัฒนธรรมของคู่ภาษายังทำให้ไม่สามารถตัดแบ่งข้อความเป็น 1 ประโยคต่อ 1 บรรทัดได้ เนื่องจากบางครั้ง ประโยคภาษาไทย 2-3 ประโยคจะถูกแปลเป็นภาษาอังกฤษที่เป็นประโยคซับซ้อนเพียงประโยคเดียว เช่นข้อความภาษาไทยคือ

ตลาดหลักทรัพย์แห่งประเทศไทยขอแจ้งว่าตลาดหลักทรัพย์ได้สั่งรับใบสำคัญแสดงสิทธิที่จะซื้อหุ้นสามัญของบริษัทซีเอสพี สตีลเซ็นเตอร์ จำกัด (มหาชน) ครั้งที่ 1 เป็นหลักทรัพย์จดทะเบียนในตลาดหลักทรัพย์ตั้งแต่วันที่ 16 มกราคม 5249 เป็นต้นไป และกำหนดให้ใบสำคัญแสดงสิทธิที่จะซื้อหุ้นสามัญของบริษัทซีเอสพี สตีลเซ็นเตอร์ จำกัด (มหาชน) ครั้งที่ 1 จำนวน 12,000,000 หน่วย เริ่มซื้อขายในตลาดหลักทรัพย์ได้ตั้งแต่วันที่ 16 มกราคม 2549 เป็นต้นไป โดยจัดอยู่ในหมวดใบสำคัญแสดงสิทธิในการซื้อหุ้นสามัญและใช้ชื่อย่อในการซื้อขายหลักทรัพย์ว่า "CSP-W1"

มีคู่ข้อความแปลภาษาอังกฤษคือ

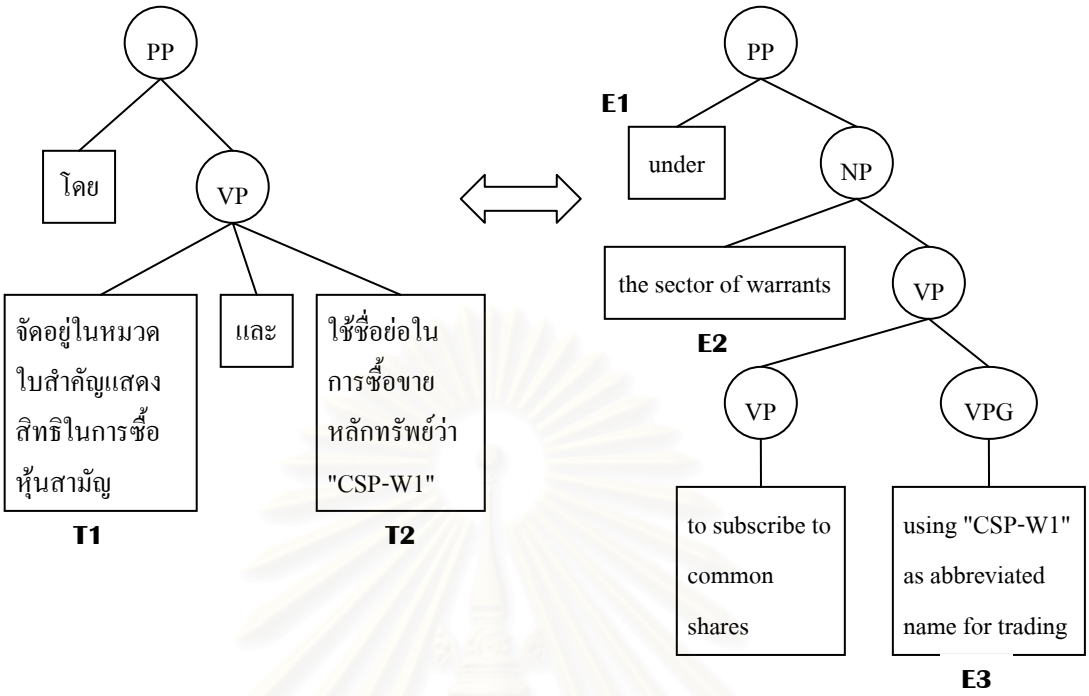
The Stock Exchange of Thailand (SET) has granted the listing of 12,000,000 units of Certificates representing the rights to purchase shares No.1 of CSP Steel Center Public Company Limited (CSP-W1) to be traded on the SET commencing 16 January 2006 under the sector of warrants to subscribe to common shares using "CSP-W1" as abbreviated name for trading.

ซึ่งจะเห็นได้ชัดว่าเป็นคู่ข้อความนี้มีส่วนของข้อความบางส่วนที่ไม่ได้มีการแปลและถูกตัดทิ้งไป เช่น “ตลาดหลักทรัพย์แห่งประเทศไทยขอแจ้งว่า” ที่ไม่ได้ปรากฏในข้อความภาษาอังกฤษ เพราะวัฒนธรรมภาษาไทยมีการเติม “ตลาดหลักทรัพย์แห่งประเทศไทยขอแจ้งว่า” เพื่อให้รายงานข่าวมีลักษณะเป็นทางการ ในขณะที่ภาษาอังกฤษไม่มี ทำให้คู่ข้อความรายงานข่าวนี้ไม่สามารถจับคู่คำศัพท์กันได้อย่างสมบูรณ์

ดังที่กล่าวไว้ข้างต้นว่าคู่รายงานข่าวนี้เป็นการแปลแบบเอาความ ข้อความไม่มีโครงสร้างทางภาษาคล้ายกัน ในบางกรณีคู่ข้อความส่วนภาษาไทยปรากฏอยู่ในลักษณะประโยค แต่ข้อความส่วนภาษาอังกฤษปรากฏอยู่ในลักษณะวลี การจัดวางองค์ประกอบก็แตกต่างกัน ซึ่งหากจะสร้างเป็นต้นไม้ทางไวยากรณ์ของแต่ละภาษา ก็จะได้ต้นไม้ทางไวยากรณ์ที่แตกต่างกันเป็นอย่างมาก เช่นการสร้างต้นไม้ทางไวยากรณ์จากคู่วลีนี้

โดยจัดอยู่ในหมวดใบสำคัญแสดงสิทธิในการซื้อหุ้นสามัญและใช้ชื่อย่อในการซื้อขายหลักทรัพย์ว่า "CSP-W1" $\leftrightarrow$ under the sector of warrants to subscribe to common shares using "CSP-W1" as abbreviated name for trading

จะได้เป็นต้นไม้ทางไวยากรณ์ดังรูปที่ 31

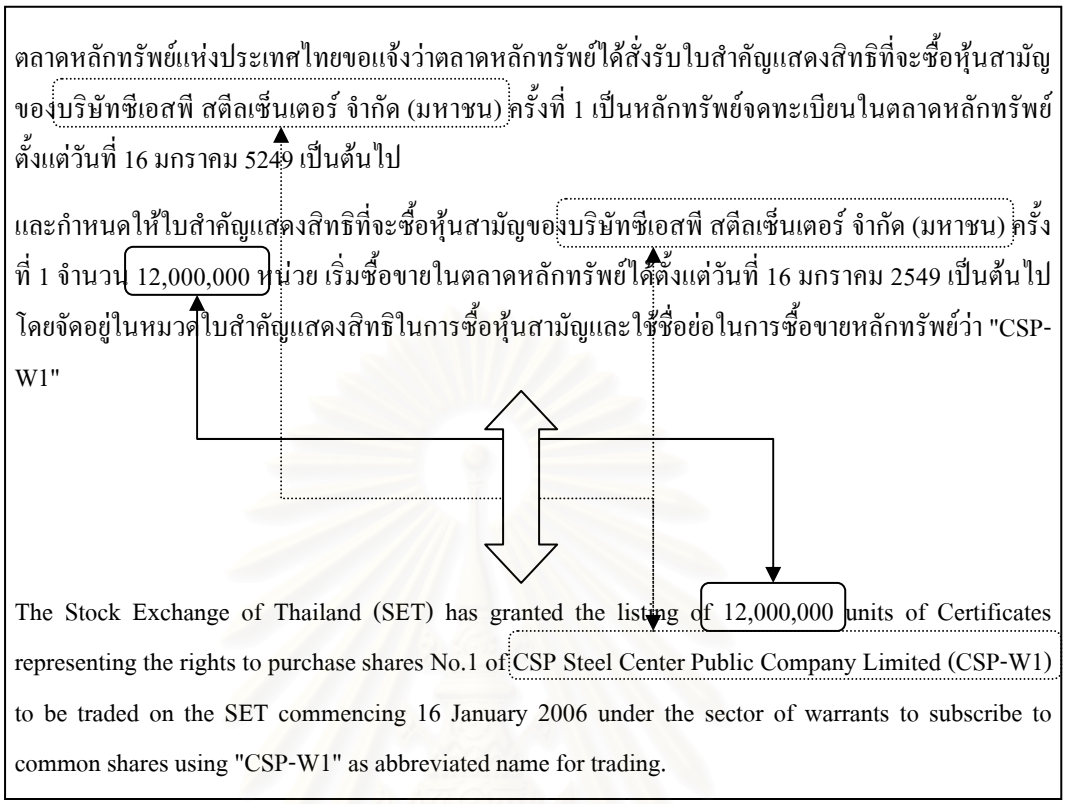


รูปที่ 31 แสดงภาพต้นไม้ทางไวยากรณ์ของคู่วลีที่เป็นคู่คำแปลกัน

จากรูปที่ 31 จะเห็นได้ว่ากริยาวลี (verb phrase–VP) T1 เป็นคู่คำแปลกับนามวลี (noun phrase–NP) E1 และกริยาวลี E2 และกริยาวลี T2 เป็นคู่คำแปลกับวลีรูปกริยาขยาย (gerundial verb phrase–VPG) E3 ซึ่งเป็นคู่คำแปลที่แตกต่างกันทางด้านโครงสร้างไวยากรณ์อย่างเด่นชัด

ความแตกต่างกันของโครงสร้างทางภาษาทำให้คู่ข้อความแปลมีความแตกต่างกันมาก ทั้งจำนวนประโยค จำนวนคำศัพท์ และวาจกของประโยค ความแตกต่างของคู่ข้อความแปลที่กล่าวมานี้เป็นเหตุให้ระบบไม่สามารถค้นหาส่วนซ้ำและส่วนไม่ซ้ำของข้อความได้เพราะจะเกิดความกำกวมเนื่องจากความหลากหลายของคู่ข้อความ

นอกจากนี้ภายในเนื้อความยังมีการปรากฏของคู่คำแปลมากกว่า 1 ครั้งหรือมีคู่คำแปลข้ามประโยคกัน ดังรูปที่ 32 ด้านล่าง ซึ่งจะทำให้เกิดความยากลำบากในการตัดประโยค ทำให้ไม่สามารถกำหนดเกณฑ์การตัดประโยคได้ที่แน่นอนได้ เพราะหากตัดประโยคไปแล้วคำบางคำจะไม่มีคู่คำแปลในประโยค



รูปที่ 32 แสดงความสัมพันธ์ของคู่คำแปลภายในข้อความ

สรุปได้ว่าในงานวิจัยชิ้นนี้ได้พบปัญหาอันเกิดจากลักษณะภาษาของคลังข้อมูลเทียบบทคือ การใช้คู่คำแปลที่แตกต่างกัน การปรากฏของคำที่ไม่ได้มีคู่คำแปล จำนวนประโยคและจำนวนคำศัพท์ที่แตกต่างกัน และการปรากฏของคู่คำแปลข้ามประโยค ดังนั้นคลังข้อมูลที่เหมาะสมกับการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างคือคู่ข้อความที่แปลแบบใกล้เคียงภาษาต้นฉบับประเภทข้อความที่แปลแบบคำต่อคำ และแปลแบบ วลีต่อวลี แต่คู่รายงานตลาดหุ้นแบบรายวันเป็นการแปลแบบเน้นเจตนา ไม่ได้คงความคล้ายของต้นฉบับไว้จึงไม่เหมาะสมสำหรับการนำมาใช้กับระบบนี้

5.3 ปัญหาทางไวยากรณ์ของการแปลภาษาไทยเป็นอังกฤษ

ปัญหานี้เกิดจากการที่ภาษาไทยและภาษาอังกฤษมีลักษณะทางไวยากรณ์ที่แตกต่างบางประการ ซึ่งเป็นอุปสรรคต่อการหาแม่แบบการแปลให้ถูกต้อง ปัญหาที่พบว่าสำคัญ ได้แก่ ปัญหาของทางไวยากรณ์การแปลภาษาจากภาษาไทยเป็นอังกฤษสามารถอธิบายด้วยทฤษฎีทางวากยสัมพันธ์ (syntax) ตามหลักวิชาภาษาศาสตร์ โดยในงานวิจัยชิ้นนี้พบปัญหาใหญ่ 2 ปัญหาคือ (1) ปัญหาที่เกิดจากกริยาช่วย และ (2) ปัญหาที่เกิดจากโครงสร้างกริยาเรียง

5.3.1 ปัญหาที่เกิดจากกริยาช่วย

จากการตรวจสอบคู่ข้อความภาษาไทยและภาษาอังกฤษภายในคลังข้อมูลพบว่าคู่ข้อความมีความแตกต่างกันทางวากยสัมพันธ์มากเพราะภาษาไทยมีการใช้คำกริยาช่วยหรือคำกริยานุเคราะห์ (auxiliary verb) จำนวนมาก “กริยาช่วยมีลักษณะทางความหมายและการเกิดร่วมต่างจากกริยาอื่นๆ อยู่พอสมควร ในด้านความหมายกริยาช่วยจะช่วยเสริมความหมายทางไวยากรณ์ให้กับกริยาหลักในแง่การณ์ลักษณะ (aspect) มาลา (mood) หรือวาจก(voice)ในด้านการปรากฏ โดยทั่วไปกริยาช่วยจะเกิดร่วมกับกริยาหลัก ส่วนใหญ่กริยาช่วยมักเกิดหน้ากริยาหลัก” (นัฐวุฒิ, 2544: 66) กริยาช่วยบางตัวอาจไม่มีคู่คำแปลในภาษาอังกฤษและถึงแม้ว่าไม่มีในข้อความก็ยังสื่อความได้ คำกริยาช่วยเช่น “ได้” ก็สามารถทำหน้าที่เป็นได้ทั้งคำกริยาหลัก คำกริยาช่วยหน้ากริยาหลักและคำกริยาช่วยหลังกริยาหลัก ทำให้ปริมาณการปรากฏเป็นจำนวนมาก ปริมาณการปรากฏเป็นจำนวนมากทำให้ต้นไม้การปรากฏร่วมมีกิ่งจำนวนมาก แต่คำกริยา “ได้” ในบางกรณีจะมีคู่คำแปลและบางกรณีจะไม่มีคู่คำแปลในคู่ข้อความ ดังนั้นเมื่อนำคำหรือกลุ่มคำมาจับเทียบสองภาษา จะเกิดความกำกวมขึ้นกับระบบและสกัดแม่แบบการแปลออกมาผิด

นอกจากนี้ภาษาไทยและภาษาอังกฤษมีความแตกต่างในการใช้กริยาช่วยเพื่อบ่งบอกกาล (tense) การณ์ลักษณะ (aspect) และทัศนภาวะ (modality) ภายในข้อความของทั้งสองภาษา ซึ่งเป็นปัญหากับระบบแปลภาษาแบบอิงตัวอย่างเนื่องจากระบบการนี้จะดูที่รูปผิวของข้อความเป็นหลัก โดยภาษาไทยการแสดงกาล การณ์ลักษณะ และทัศนภาวะจะเติมคำกริยาช่วยลงหน้าหรือหลังกริยาหลักเพียงเท่านั้นโดยไม่เปลี่ยนแปลงรูปผิวของกริยาหลัก ส่วนภาษาอังกฤษมีการแสดงกาล การณ์ลักษณะ และทัศนภาวะทั้งในรูปแบบการผันคำกริยาและเติมคำกริยาช่วยลงหน้ากริยาหลัก เช่น

	ตัวอย่างข้อความไทย	ตัวอย่างข้อความอังกฤษ
แสดงปัจจุบันกาลและการณ์ลักษณะสัจพจน์	เขา ขับ รถ	he drives a car
แสดงอดีตกาล	เขา ขับ รถ แล้ว	he drove a car
แสดงปัจจุบันกาลสมบูรณ์	เขา ได้ ขับ รถ	he has driven a car
แสดงทัศนภาวะ	เขา น่าจะ ขับ รถ	he might drive a car
แสดงทัศนภาวะ	เขา ขับ รถ ได้	he can drive a car

จะเห็นได้ว่าส่วนภาษาไทยรูปผิวของคำกริยา “ขับ” จากตัวอย่างทุกข้อความไม่มีการเปลี่ยนแปลงและจาก 5 ข้อความนี้ ระบบสามารถนำ (เขา) (ขับ) (รถ) เป็นส่วนซ้ำของข้อความได้ แต่ในส่วนภาษาอังกฤษรูปผิวของกริยา “drive” มีการผันแปรตามลักษณะการแสดงการณ์

ลักษณะ และทัศนภาวะทำให้ระบบไม่สามารถหาส่วนซ้ำในส่วนกริยาของข้อความภาษาอังกฤษได้ ซึ่งในกรณี5 ข้อความนี้ ระบบจะสร้างต้นไม้การปรากฏร่วมภาษาไทย โดยมีโหนดใบคือ (ขับ)(รถ)(เขา)[1, 2, 3, 4, 5] และสร้างต้นไม้การปรากฏร่วมภาษาอังกฤษโดยมีโหนดใบคือ (a)(car)(drive)(he)[4, 5] ซึ่งโหนดใบทั้งคู่กลุ่มของเลขบรรทัดไม่เท่ากัน ระบบจึงไม่สามารถจับคู่เทียบสองภาษาได้ และไม่ได้แม่แบบการแปล ซึ่งเป็นผลมาจากการแสดงกาล การณ์ลักษณะ และทัศนภาวะที่แตกต่างกันโดยการเติมกริยาช่วยและการผันกริยาของคู่ภาษาไทย-อังกฤษ

5.3.2 ปัญหาที่เกิดจากโครงสร้างกริยาเรียง

ความเฉพาะตัวของโครงสร้างกริยาเรียง (serial verb construction) ของภาษาไทยซึ่ง รศ.ดร. กิ่งกาญจน์ เทพกาญจนา ได้ให้นิยามไว้ว่า “โครงสร้างกริยาเรียงคือ โครงสร้างทางวากยสัมพันธ์ที่มีกริยา (หรือกริยาวลี) สองตัวขึ้นไป เรียงต่อกันโดยไม่มีตัวเชื่อมความใดๆ” (Thepkanjana, 2006) นอกจากนี้นัฐวุฒิยังได้อธิบายไว้ว่า “เป็นเรื่องในระดับวากยสัมพันธ์ที่คำกริยาหลายคำสามารถปรากฏเรียงต่อเนื่องกันได้” (นัฐวุฒิ, 2544: 65) ก็มีรูปแบบเฉพาะซึ่งจะมีกริยาแก่นอยู่ 1 คำและมีกริยาเสริมอื่นๆ อยู่เรียงกันซึ่งคู่คำแปลในภาษาอังกฤษมักจะปรากฏแค่กริยาหลักเท่านั้นจึงกล่าวได้ว่าข้อความคู่ภาษาไทยและภาษาอังกฤษแม้สื่อความหมายเดียวกันแต่การปรากฏทางวากยสัมพันธ์ต่างกันมาก เช่น คู่ข้อความ

ตามที่บริษัทเอเชียนมารีนเซอร์วิสส์จำกัด (มหาชน) (ASIMAR)ได้ดำเนินการเพิ่มทุนจดทะเบียนและขอให้ตลาดหลักทรัพย์รับหุ้นเพิ่มทุนเป็นหลักทรัพย์จดทะเบียนเพิ่มเติมนั้นตลาดหลักทรัพย์ได้พิจารณาแล้วเห็นควรกำหนดให้หุ้นเพิ่มทุนของบริษัทดังกล่าวเริ่มทำการซื้อขายในตลาดหลักทรัพย์ได้ตั้งแต่วันที่ 28 มีนาคม 2549 เป็นต้นไป ↔ Starting from 28 March 2006, the Stock Exchange of Thailand (SET) allowed the securities of Asian Marine Services Public Company Limited (ASIMAR) to be listed and traded on the SET after finishing capital increase procedures.

ซึ่งเป็นคู่คำแปลของกันและกัน แต่ข้อความภาษาไทยและภาษาอังกฤษไม่สามารถแยกคำแต่ละคำเพื่อมาบอกว่าคำใดเป็นคู่แปลของคำใดได้จนครบทุกคำ เช่นภายในข้อความส่วนภาษาไทยมีคำกริยาปรากฏอยู่หลายคำ และมีกริยาเรียง “ดำเนินการเพิ่ม...ขอให้...รับ...เป็น” ซึ่งไม่มีคู่คำแปลของชุดกริยาเรียงนี้ในส่วนภาษาอังกฤษเลย และชุดกริยาเรียง “พิจารณาแล้วเห็นควรกำหนดให้... เริ่มทำการซื้อขาย” มีคู่คำแปลเป็นภาษาอังกฤษคือ “allowed...to be listed and traded” เป็นต้น ทำให้ระบบเกิดความกำกวมในการจับคู่เทียบต้นไม้การปรากฏร่วมสองภาษา เพราะถ้าคำเหล่านั้นสามารถปรากฏในข้อความอื่นและมีคำอื่นในคู่ภาษาที่ปรากฏร่วมด้วยเช่นเดียวกับในคู่ข้อความนี้ก็จะจับคู่คำเหล่านี้กับคำอื่นแทน เช่นคำ “ทุนจดทะเบียน” และ “ขอให้” สามารถปรากฏในข้อความ

อื่นที่มีคู่เลขบรรทัดตรงกับข้อความที่มีคำ “the Stock Exchange of Thailand” ในเนื้อความ ทำให้ระบบเก็บแม่แบบการแปลเป็น “<X0>ทุนจดทะเบียน<X1>ขอให้ตลาดหลักทรัพย์<X2> ↔ <Y0> the Stock Exchange of Thailand <Y1>” แทน เพราะคู่คำเหล่านี้ปรากฏร่วมกันในคู่ข้อความ ซึ่งเป็นแม่แบบการแปลที่ผิดและจะทำให้การสกัดแม่แบบการแปลจากข้อความกลุ่มนี้ผิดทั้งหมด

ปัญหาที่กล่าวมาทั้งหมดเป็นสาเหตุที่ระบบไม่สามารถสกัดแม่แบบการแปลจากคลังข้อมูลเทียบบททั่วไปเช่นคลังข้อมูลชุดทดลองที่ 1 และ 2 ได้ ซึ่งปัญหาเหล่านี้แบ่งเป็นข้อจำกัดของระบบการสกัดหาแม่แบบการแปล ลักษณะที่เหมาะสมของคลังข้อมูลเทียบบท และความแตกต่างทางวากยสัมพันธ์ของคู่ภาษา ดังนั้นการจะพัฒนาระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างจากภาษาไทยเป็นภาษาอังกฤษด้วยแนวทางการสกัดแม่แบบการแปลในอนาคตจึงต้องนำปัญหาเหล่านี้เข้ามาเป็นองค์ประกอบสำคัญด้วย

บทที่ 6

สรุปและข้อเสนอแนะ

ภายในบทนี้จะนำเสนอสรุปผลการทดลองตรวจสอบกับสมมติฐานที่ตั้งไว้ก่อนหน้านี้ เพื่อดูว่าการศึกษาของวิทยานิพนธ์ฉบับนี้ได้ผลลัพธ์เป็นไปตามที่คาดไว้หรือไม่ อย่างไร และจากการศึกษาครั้งนี้สามารถเห็นแนวทางในการพัฒนาปรับปรุงระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างแนวทางสกัดแม่แบบการแปลอย่างไรบ้าง โดยแบ่งเนื้อหาภายในบทนี้ออกเป็น 2 ส่วนคือ (1) สรุปผลเปรียบเทียบกับสมมติฐาน และ (2) ข้อเสนอแนะแนวทางพัฒนาปรับปรุงระบบ

6.1 สรุปผลเปรียบเทียบกับสมมติฐาน

วิทยานิพนธ์ฉบับนี้ได้นำเสนอวิธีการพัฒนาระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง เนื่องจากระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างต้องใช้ตัวอย่างคู่ภาษา ผู้วิจัยจึงเก็บรวบรวมคู่ข้อความแปลจากรายงานข่าวแบบรายวันจากเว็บไซต์ตลาดหลักทรัพย์แห่งประเทศไทย (http://www.set.or.th) แล้วจึงพัฒนาระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างโดยการใช้ระบบสกัดแม่แบบการแปลเพื่อสกัดหาตัวอย่างจากคลังข้อมูลเทียบบทมาจากการค้นหาส่วนซ้ำและส่วนไม่ซ้ำของรูปแบบข้อความที่ปรากฏในคลังข้อมูล และเก็บส่วนที่ปรากฏซ้ำกันของทั้งคู่ประโยคตัวอย่างเป็นส่วนคงที่เพื่อเป็นแม่แบบการแปลหลัก และเก็บส่วนที่ไม่ซ้ำกันไปเป็นแม่แบบการแลส่วนผันแปรเพื่อเติมเต็มข้อความที่ต้องการแปลให้สมบูรณ์ เมื่อระบบสกัดได้แม่แบบการแปลจากทุกข้อความแล้วจะเก็บแม่แบบการแปลเหล่านั้นลงสู่ฐานข้อมูลแม่แบบการแปลเพื่อเปรียบเทียบแปลข้อความรับเข้าต่อไป เมื่อมีข้อความรับเข้าระบบจะทำการค้นหาแม่แบบการแปลที่ใช้เปรียบเทียบแปลได้และเริ่มทำการแปลข้อความทีละส่วนจากแม่แบบการแปลขนาดใหญ่ที่สุดที่มีรูปแบบเหมือนข้อความรับเข้าตามลำดับจนแปลเสร็จทั้งข้อความ แล้วจะได้ผลลัพธ์เป็นข้อความแปล

ผลการทดลองสกัดแม่แบบการแปลจากคลังข้อมูลโดยตรงพบว่า ระบบสามารถสกัดแม่แบบการแปลได้ถูกต้องเพียงร้อยละ 9.85 เนื่องจากระบบพบปัญหาในการสร้างต้นไม้การปรากฏร่วมจากข้อความภายในรายงานตลาดหุ้นและคู่ข้อความแปลของรายงานตลาดหุ้นที่เป็นการแปลแบบเอาเนื้อความโดยมีความแตกต่างกันมากทางด้านโครงสร้างทางไวยากรณ์ ดังนั้นผู้วิจัยจึงทดลองต่อโดยช่วยจัดกลุ่มข้อมูลเพื่อที่จะสามารถสกัดแม่แบบการแปลที่จะสามารถใช้การได้เพื่อจะนำผลมาทดสอบส่วนการแปลข้อความต่อไป

ดังนั้นสมมติฐานที่ว่าภาษาในรายงานตลาดหุ้นเป็นภาษาเฉพาะที่มีรูปแบบซ้ำๆ สามารถดึงตัวอย่างมาสร้างเป็นแม่แบบของการแปลในระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างได้จึงไม่เป็นไปตามคาด เพราะแม้ว่าภาษาในรายงานตลาดหุ้นจะเป็นภาษาเฉพาะทางที่มีรูปแบบซ้ำๆ กันก็จริง แต่คู่ภาษาในรายงานตลาดหุ้นมีคู่ข้อความแปลที่เป็นการแปลแบบเน้นเจตนาจุดประสงค์ และเป้าหมายของเนื้อความมากจึงทำให้คู่คำแปลของเนื้อความมีความแตกต่างกันมากทั้งจำนวนคำศัพท์ ลักษณะทางไวยากรณ์ และความหมายของคำศัพท์ ทำให้ไม่เหมาะกับการสกัดแม่แบบการแปลแบบทั่วไป จึงไม่สามารถดึงตัวอย่างมาสร้างเป็นแม่แบบของการแปลในระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างได้ ดังนั้นจึงกล่าวได้ว่าข้อมูลรายงานตลาดหุ้นและบทแปลรายงานตลาดหุ้นในที่นี้ ไม่เหมาะกับการนำมาใช้เป็นตัวอย่างในการแปลข้อความของระบบการแปลภาษาด้วยเครื่อง

อาจกล่าวได้ว่า งานวิจัยชิ้นนี้ได้ย้ำเตือนให้เห็นถึงความสำคัญของการวิเคราะห์ลักษณะทางภาษาของคลังข้อมูลเทียบบทไว้อย่างชัดเจน การพัฒนางานแปลภาษาด้วยเครื่องนั้นไม่ใช่แค่พัฒนาระบบเท่านั้น แต่ต้องวิเคราะห์ด้วยว่าระบบการแปลภาษาด้วยเครื่องที่พัฒนามานั้นเหมาะสมในการใช้งานกับคลังข้อมูลเทียบบทหรือคู่ภาษาแบบใด ดังนั้นการวิเคราะห์ลักษณะทางภาษาของคลังข้อมูลเทียบบทจึงเป็นสิ่งสำคัญที่จะต้องมีการวิเคราะห์ลักษณะ และปัญหาความต่างทางภาษาก่อนเพื่อดูว่าข้อมูลนั้นสามารถใช้ได้กับระบบที่เลือกได้มากน้อยเพียงไร

ความถูกต้องในการแปล อันเป็นผลมาจากลักษณะทางภาษาในการแปลเทียบบทของคลังข้อมูลเทียบบทนั้น ขึ้นอยู่กับความสอดคล้องข้ามภาษาในระดับไวยากรณ์ (cross-lingual grammatical correspondence) เป็นสำคัญ ปัจจัยทางไวยากรณ์ต่างๆ ของคู่ภาษาไม่ว่าจะเป็น ระบบการเรียงคำ (word ordering system) ความคล้ายคลึงในเชิงความหมายของคำศัพท์ (lexical-semantic similarity) คำยืมข้ามภาษา (cognates) ระบบการวิวัฒน์คำ (word formation system) โครงสร้างส่วนเติมเต็มของคำ (argument structure) คุณสมบัติเชิงวากยสัมพันธ์ (syntactic features) รูปแบบโครงสร้างการผูกประโยค (sentence formation) ซึ่งบ่งบอกถึงความคล้ายคลึงหรือความแตกต่างของตระกูลภาษา ล้วนมีอิทธิพลต่อความถูกต้องของการแปลภาษาด้วยเครื่องทั้งสิ้น ยิ่งคู่ภาษาใดที่มีความคล้ายคลึงกันในปัจจัยที่ได้กล่าวมานี้ ผลการแปลภาษาด้วยเครื่องของคู่ภาษาดังกล่าวก็ย่อมมีความถูกต้องเพิ่มมากขึ้นตามไปด้วย

ระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างในงานวิจัยชิ้นนี้จะทำงานได้อย่างมีประสิทธิภาพสูงสุดหากคลังข้อมูลเทียบบทมีลักษณะเป็นคู่คำแปลที่คงแบบต้นฉบับไว้ เมื่อนำกลับมาเปรียบเทียบกับการแปลทั่วไป คู่ภาษาในการแปลเอกสารส่วนใหญ่มักเป็นภาษาคนละตระกูล เช่น การแปลระหว่างภาษาอังกฤษ ซึ่งอยู่ในตระกูลมีการผันคำ กับภาษาไทย ซึ่งอยู่ใน

ตระกูลที่เป็นคำโดด ภาษาที่ใช้ในการแปลทั่วไปก็ไม่สามารถจำกัดวงศัพท์ได้ อีกทั้งความหลากหลายของการสร้างอรรถาธิบายในคู่ภาษาก็มีอยู่มากมาย ดังนั้นจึงอาจสรุปได้ว่าการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างนั้นจะสามารถใช้แปลข้อความที่มีลักษณะเป็นคู่คำแปลที่คงแบบต้นฉบับไว้อย่างชัดเจน ไม่เหมาะกับการใช้แปลข้อความทั่วไป

ดังนั้นผลการทดลองแปลข้อความที่ได้จึงไม่เป็นไปตามที่คาดไว้ในสมมติฐานที่คาดไว้ว่าระบบจะสามารถแปลรายงานตลาดหุ้นได้ถูกต้องได้ไม่ต่ำกว่าร้อยละ 85 โดยการแปลที่อาศัยแม่แบบที่สกัดจากคลังข้อมูลโดยตรงให้ความถูกต้องเพียงร้อยละ 3.70 ส่วนผลการแปลที่สกัดจากแม่แบบการแปลที่ผู้วิจัยจัดกลุ่มข้อมูลตามความคล้ายคลึงให้ความถูกต้องร้อยละ 67.68 ซึ่งก็ยังต่ำกว่าสมมติฐาน ซึ่งเป็นผลมาจากปริมาณข้อมูลคู่ภาษาในคลังข้อมูลเทียบบทมีไม่เพียงพอและไม่ครอบคลุมข้อความรับเข้า จึงทำให้ไม่สามารถหาตัวอย่างมาแปลข้อความรับเข้าบางข้อความได้ ส่วนสาเหตุที่ปริมาณข้อมูลคู่ภาษาในคลังข้อมูลเทียบบทมีน้อยเป็นเพราะปริมาณการประกาศรายงานข่าวภายในตลาดหุ้นมีน้อย

สำหรับปัญหาที่ผู้วิจัยคาดว่าระบบการแปลแบบอิงตัวอย่างนี้จะไม่สามารถแปลหน่วยภาษาในระดับปริจเฉทได้นั้น จากการทดลองพบว่าข้อมูลที่นำมาใช้ไม่มีหน่วยทางปริจเฉทมากเพียงพอที่จะนำมาอภิปรายและตอบคำถามได้ แต่อย่างไรก็ตามจากการทดลองด้วยคลังข้อมูลที่ใช้ในงานวิจัยนี้ ระบบจะสามารถแปลหน่วยทางปริจเฉทได้และไม่พบปัญหาใดๆ หากหน่วยทางปริจเฉทนั้นมีการแปลที่คงที่และสม่ำเสมอ แต่ถ้าหน่วยทางปริจเฉทนั้นมีการแปลได้หลายแบบหรือละได้ก็จะเป็นปัญหาเช่นเดียวกับกรณีอื่นๆ ที่กล่าวมาแล้ว เพราะระบบใช้การแปลโดยการแทนที่คำแปลจากแม่แบบการแปลโดยตรง ถึงแม้ว่าระบบจะไม่ตรวจสอบและไม่สามารถทราบได้ว่ารูปแบบของตัวอย่างคู่ภาษามีลักษณะทางไวยากรณ์เป็นอย่างไร ระบบจะจดจำเพียงรูปผิวของภาษาซึ่งก็คือการเรียงตัวของตัวอักษรเท่านั้น

6.2 ข้อเสนอแนะแนวทางพัฒนาปรับปรุงระบบ

จากการทดลองพัฒนาระบบการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างแนวทางสกัดแม่แบบการแปลพบว่าระบบมีจุดด้อยอยู่บางประการ ปัญหาส่วนใหญ่ของระบบจะเกิดขึ้นที่ระบบการสกัดแม่แบบการแปลซึ่งเป็นระบบที่สำคัญที่สุด ดังนั้นการพัฒนาระบบให้มีประสิทธิภาพสูงขึ้นจึงต้องเน้นเพิ่มศักยภาพในการสกัดแม่แบบการแปลให้มีความถูกต้องมากขึ้น

(1) คลังข้อมูลที่ใช้กับระบบสกัดแม่แบบการแปลควรใช้คู่ข้อความที่เป็นคู่คำแปลแบบคล้ายต้นฉบับ กล่าวคือเป็นคู่คำแปลที่คงลักษณะไวยากรณ์และความหมายของคำศัพท์ของกันและกันไว้เช่น “ภาษาศาสตร์ไม่ใช่วิชาที่มุ่งศึกษาทักษะการใช้ภาษาต่างๆ” เป็นคู่ข้อความแปลกับ

“Linguistics is not a subject which aims at developing skills in using languages.” หรืออาจเป็นคู่ภาษาที่เป็นตระกูลเดียวกันและมีความคล้ายคลึงกันทั้งทางด้านวงคำศัพท์และลักษณะไวยากรณ์ เช่นคู่ภาษาไทย-ลาว เป็นต้น ซึ่งภาษาตระกูลเดียวกันจะสามารถใช้การคำนวณค่าคะแนนจากรากศัพท์ของภาษาตระกูลเดียวกันและค่าคะแนนความคล้ายของตัวอักษรมาช่วยในการจับคู่คำแปล ซึ่งจะทำให้การสกัดแม่แบบการแปลมีประสิทธิภาพเพิ่มมากขึ้น

(2) อัลกอริธึมของระบบถูกวางตามแนวความคิดว่าภายในข้อความจะมีส่วนซ้ำเป็นแกนของข้อความและมีส่วนไม่ซ้ำอยู่ในตำแหน่งด้านหน้าหรือด้านหลังหรือคั่นกลางส่วนซ้ำ ตำแหน่งละ 1 ส่วน จากแนวความคิดนี้ทำให้ส่วนไม่ซ้ำที่อยู่ติดกันทั้งหมดรวมเป็นส่วนไม่ซ้ำเดียวกัน 1 ส่วน ซึ่งบางกรณีส่วนไม่ซ้ำที่อยู่ติดกันไม่ควรรวมเป็นส่วนเดียวกันทำให้แม่แบบการแปลส่วนผันแปร (แม่แบบการแปลของส่วนไม่ซ้ำ) และระบบรวมคำแปลใหม่จะไม่สามารถเติมเต็มข้อความได้หากข้อความรับเข้ามีแค่ส่วนใดส่วนหนึ่งของส่วนซ้ำนั้น เช่น คู่ข้อความ “วันที่ 15 ธันวาคม 2548 BNT ได้รับเครื่องหมาย NR ↔ NR signs posted against BNT on 15 December, 2006” จะได้แม่แบบการแปล

[<X0> ได้รับเครื่องหมาย NR ↔ NR signs posted against <Y1> on <Y2>]

[วันที่ 15 ธันวาคม 2548 BNT<X1> ↔ <Y0> BNT <Y1> 15 December, 2006]

ผลการสกัดแม่แบบการแปลชุดนี้ก็จะได้แม่แบบการแปลส่วนผันแปรที่ไม่ครอบคลุมและมีความเป็นไปได้ต่ำที่จะถูกเรียกใช้ไปแปลข้อความรับเข้าเพราะการเกิดของส่วนไม่ซ้ำที่เกิดคู่กัน มีโอกาสน้อยมากที่จะเกิดคู่กันอีกครั้ง และฐานข้อมูลแม่แบบการแปลก็จะไม่มีคำแปลให้กับ “วันที่ 15 ธันวาคม 2548” และ “BNT” ถ้าข้อความรับเข้ามีเพียงคำใดคำหนึ่ง ดังนั้นการพัฒนาระบบให้สามารถแบ่งส่วนไม่ซ้ำออกได้ ก็จะเพิ่มปริมาณแม่แบบการแปลให้ครอบคลุมข้อความรับเข้ามากขึ้น

(3) การสร้างต้นไม้การปรากฏร่วมเพื่อนำโหนดใบซึ่งเป็นตำแหน่งของกลุ่มคำที่ปรากฏซ้ำกันทั้งหมดมาใช้สร้างเป็นแม่แบบการแปลจะทำให้ได้แม่แบบการแปลที่มีกลุ่มคำที่ปรากฏซ้ำกันทั้งหมดแต่บางกรณีที่ข้อความรับเข้าที่ใกล้เคียงกับแม่แบบการแปลนั้นไม่มีคำซ้ำกับแม่แบบการแปลแค่เพียง 1 คำ แม่แบบการแปลนั้นจะไม่สามารถใช้แปลได้ ดังนั้นถ้าระบบสามารถอนุญาตให้นำโหนดอื่นๆ ที่อยู่ก่อนโหนดใบไป 1-2 ระดับ ก็จะช่วยแก้ปัญหาข้อความรับเข้าขาดคำ 1 คำในแม่แบบการแปลได้

(4) เนื่องจากข้อความภาษาไทยมีการปรากฏคำกริยาช่วยที่เป็นคำแสดงกาล การณ์ลักษณะ และทัศนภาวะ และภาษาอังกฤษที่ใช้การผันกริยาหลักและเติมคำกริยาช่วยแสดงกาล การณ์ลักษณะ และทัศนภาวะ ซึ่งเป็นต้นเหตุของความกำกวมเพราะคำเหล่านี้บางกรณีก็สามารถทำหน้าที่เป็นกริยาหลัก ดังนั้นหากมีระบบประมวลก่อน (preprocessor) ที่สามารถค้นหาและแปลงคำกริยาช่วยและการผันเป็นลักษณ์ (feature) แสดงไว้กับกริยาหลักจะช่วยลดความกำกวมเพื่อให้ต้นไม้การปรากฏร่วมได้กลุ่มคำที่เป็นคำหลักของเนื้อความภาษาไทยอย่างแท้จริงเพื่อนำไปเทียบบทจับคู่ และระบบจับคู่คำแปลจะสามารถทำงานได้ง่ายขึ้นเพราะการปรากฏซ้ำของคำมีความเด่นชัดขึ้น เช่น

การแสดงอดีตกาล “เขาขับรถแล้ว ↔ he drove a car” จะถูกแปลงเป็น (เขา) (ขับ{person:3rd/sing, tense:past}) (รถ) ↔ (he) (drive{person:3rd/sing, tense:past}) (a) (car)

การแสดงปัจจุบันกาลสมบูรณ์ “ฉันได้ขับรถ ↔ I has driven a car” จะถูกแปลงเป็น (ฉัน) (ขับ{person:1st/sing, tense:presperf}) (รถ) ↔ (I) (drive{person:1st/sing, tense:persperf}) (a) (car)

การแสดงทัศนภาวะ “คุณขับรถได้ ↔ you can drive a car” จะถูกแปลงเป็น (คุณ) (ขับ{person:2nd/sing, tense:pres, modality:ablity}) (รถ) ↔ (you) (drive{person:2nd/sing, tense:pres, modality:ablity})) (a) (car)

(5) เนื่องจากข้อความภาษาไทยมีการละเป็นจำนวนมาก ดังนั้นหากระบบยอมให้มีการจับคู่ข้อความแบบไร้คู่ (null-alignment) ก็จะช่วยแก้ปัญหาการละที่จะเกิดขึ้นในข้อความภาษาไทยได้ เช่น คู่ข้อความต่อไปนี้ “ผมไม่รู้ ↔ I don’t know” และ “ผมก็ไม่รู้ ↔ I don’t know” การไม่อนุญาตให้มีการจับคู่ข้อความแบบไร้คู่ทำให้คู่ข้อความนี้ไม่สามารถรับ “ผมก็ไม่รู้ ↔ I don’t know” ได้เพราะมีการละคำว่า “ก็” เกิดขึ้น ดังนั้นการมีการจับคู่ข้อความแบบไร้คู่จะช่วยแก้ปัญหาการละของคำบ่างคำภายในข้อความและจะทำให้ได้แม่แบบการแปลที่ใช้งานได้หลากหลายขึ้น

รายการอ้างอิง

ภาษาไทย

ณัฐพล กฤษสุทธิกุล. 2549. **ระบบแปลภาษาอังกฤษ-ไทย ด้วยเครื่องแบบอิงตัวอย่าง โดยใช้ตัวแบบเอ็นแกรม**. วิทยานิพนธ์ปริญญามหาบัณฑิต. ภาควิชาเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

นัฐวุฒิ ไชยเจริญ. 2544. **การตัดคำและการกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จด้วยคอมพิวเตอร์**. วิทยานิพนธ์ปริญญามหาบัณฑิต. ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

นิตยา กาญจนะวรรณ. 2534. **ภาษาศาสตร์คอมพิวเตอร์ 1**. กรุงเทพมหานคร: สำนักพิมพ์มหาวิทยาลัยรามคำแหง.

ไพศาล เจริญพรสวัสดิ์. 2540. **การตัดคำภาษาไทยโดยใช้คุณลักษณะ**. วิทยานิพนธ์ปริญญามหาบัณฑิต. ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยลัย จุฬาลงกรณ์มหาวิทยาลัย.

มณีรัตน์ สวัสดิวัฒน์ ณ อยุธยา. 2548. **การแปล: หลักการและการวิเคราะห์**. กรุงเทพมหานคร: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

วรรณา แสงอร่ามเรือง. 2545. **ทฤษฎีและหลักการแปล**. กรุงเทพมหานคร: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

วิลเลี่ยม วีฟเวอร์. 2540. **สมัญญาแห่งดอกกุหลาบ**. แปลโดย ภควดี วีระภาสพงษ์. กรุงเทพมหานคร: คบไฟ.

ภาษาอังกฤษ

Al-Onaizan, Y., et al. 1999. Statistical Machine Translation. **Proceedings of JHU Workshop**. Baltimore, USA.

Allen, J. 1995. **Natural Language Understanding**. Massachusetts: The Benjamin.

Arnold, D. J., Balkan, L., Meijer, S., Humphreys, R. L., and Sadler, L. 2001. **Machine Translation: An Introductory Guide**. London: NCC Blackwell.

Arnold, D. J., and Sadler, L. 1989. Non-compositionality and Translation. **Recent Developments and Applications of Natural Language Processing**: 23-55.

Aroonmanakun, W. 2002. Collocation and Thai Word Segmentation. **Proceedings of SNLP-Oriental COCOSDA**, pp. 68-75. Hua Hin, Thailand, May 9 to 11, 2002.

Beaven, J. L. 1992. **Lexicalist Unification-based Machine Translation**. Doctoral dissertation. Department of Artificial of Intelligence, University of Edinburgh.

Basnett, S. 1991. **Translation Studies**. London: Routledge.

Boonkwan, P. 2005. **Performance Analysis of SLR-Based Prasing Method on Thai Lexical-Functional Grammar**. Master's Thesis. Department of Computer Engineering, Kasetsart University.

Brill, E. 1992. A Simple Rule-Based Part of Speech Tagger. **Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics**. University of Delaware, Deleware, USA., 1992.

Brown, P. F., Pietra, S. A. D., Pietra, V. D. J., and Mercer, R. L. 1993. The mathematics of Machine Translation: Parameter estimation. **Computational Linguistics** 19(2): 263-312.

Charoenpornsawat, P., Sornlertlamvanich V., and Charoenporn, T. 2002. Improving Translation Quality of Rule-based Machine Translation. **Proceedings of COLING 2002**, Taipei, Taiwan, 2002.

Dice, L. R. 1945. Measures of the Amount of Ecological Association Between Species. **Geology** 26: 297-302.

Dorr, B. J., Hovy, E., and Levin, L. 2004. Machine Translation: Interlingual Methods. **Encyclopedia of Language and Linguistics 2nd edition**: 615-632.

Grishman, R., and Kittredge, R.I. 1986. **Analyzing Language in Restricted Domains: Sublanguage Description and Processing**. New Jersey: Hillsdale.

Gordon, T. D. 1985. **Translation Theory and Methods**. Pennsylvania: Grove City Press.

Hutchins, W. J., and Somers, H. L. 1992. **An Introduction to Machine Translation**. London: Academic Press.

Hutchins, W. J., Plumb, R. K., and Kenny, H. C. 1994. **MT News International** 8 (May 1994): 15-18.

Kocbek, A. 2006. Language and Culture in International Legal Communication. **Managing Global Transitions** 4(3): 231-247.

Kashioka, H., Maruyama, T., and Tanaka, H. 2003. Building a parallel corpus for monologues with clause alignment. **Proceedings of MT Summit IX**. pp. 216-223. New Orleans, USA., Sept. 23 to 27, 2003.

Manning, C., and Schütze, H. 1999. **Foundation of Statistical Natural Language Processing**. London: The MIT Press.

McTait, K. 2001. **Translation Pattern Extraction and Recombination for Example-Based Machine Translation**. Doctoral dissertation. Centre for Computational Linguistics, Department of Language Engineering, University of Manchester Institute of Science and Technology.

Newman, A. 1980. **Mapping Translation Equivalence**. Leuven: ACCO.

Newmark, P. 1988. **A Textbook of Translation**. New York: Prentice.

Nida, E. A. 1975. **Language Structure and Translation**. Stratford: Oxford University Press.

Nida, E. A., and Taber, C. R. 1969. **The Theory and Practice of Translation**. Leiden: Brill.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. 2002. **Bleu: A Method for Automatic Evaluation of Machine Translation: Technical Report RC22176 (W0109-022).** (Unpublished Manuscript)

Reiß, K., and Vermeer, H. J. 1984. **Grundlegung einer allgemeinen Translationstheorie**. Tübingen: Niemeyer.

Schulte R., and Biguenet, J. 1992. **Theories of Translation: An Anthology of Essays from Dryden to Derrida**. Chicago: University of Chicago Press.

Snell-Hornby. 1995. **Translation Studies: An Integrated Approach**. Amsterdam: Benjamins.

Sornlertlamvanich, V., Potipiti, T., Wutiwiwatchai, C., and Mittrapiyanuruk P. 2000. The State of the Art in Thai Language Processing. **Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics**. pp. 597-598. Hong Kong, Oct. 22, 2000.

Sornlertlamvanich, V., Charoenporn, T., and Isahara, H. 1997. **ORCHID: Thai Part-Of-Speech Tagged Corpus: Technical Report: TR-NECTEC-1997-001**. (Unpublished Manuscript)

Sornlertlamvanich, V., and Meknavin, S. 1995. **Thai Generation Rules: Technical Report: 6-CICC-MT50**. (Unpublished Manuscript)

Sornlertlamvanich, V., and Phantachat W. 1995. **Thai Analysis Rules: Technical Report: 6-CICC-MT46**. (Unpublished Manuscript)

Thepkanjana, K. 2006. Properties of events expressed by serial verb constructions in Thai. **Proceedings of 11th Biennial Symposium: Intertheoretical Approaches to Complex Verb Constructions**. Rice University, USA..

Toury, G. 1995. **Descriptive Translation and Beyond**. Amsterdam: John Benjamins.

Vermeer, H. J. 1998. **Didactics of Translation**. New York: Routledge.

Way, A., and Kenny, D. 2001. Teaching Machine Translation & Translation Technology: A Constrastive Study. **Proceedings of the MT Summit Workshop on Teaching Machine Translation**, Santiago de Compostela, Spain, 2001.

ภาคผนวก

นิพจน์ปกติ (regular expression) ซึ่งก็คือการใช้เครื่องหมายและสัญลักษณ์พิเศษมาตรวจสอบหรือเทียบค้นหาตัวอักษร เช่น การใช้โดยมีเครื่องหมายและสัญลักษณ์พิเศษหลักที่ใช้ดังนี้

- เครื่องหมายไปป์ (pipe) '|' ใช้เมื่อต้องการเสนอทางเลือกอย่างใดอย่างหนึ่ง เช่น "การ|ความ" ใช้คำว่า 'การ' หรือ 'ความ' ก็ได้

- เครื่องหมายปรัศนี (question mark) '?' หมายความว่าแทนที่เครื่องหมายปรัศนีด้วยตัวอักษรอะไรก็ได้ อย่างน้อย 1 ตัว เช่น "ก?ข" หมายถึง จะมีตัวอักษร 'ก' 1 ตัว และมีตัวอักษรตัวสุดท้ายเป็นตัวอักษร 'ข' 1 ตัว แต่มีตัวอักษรอะไรก็ได้คั่นกลาง ตัวอักษร "ก" และ "ข" 1 ตัว ดังนั้น "ก?ข" จึงตรวจสอบพบ "กกข" และ "ก4ข"

- เครื่องหมายบวก (plus) '+' หมายความว่าตัวอักษรที่อยู่หน้าเครื่องหมายนี้ ต้องมีปรากฏในคำที่นำมาตรวจสอบ อย่างน้อย 1 ตัว เช่น "ท+" จะตรวจสอบพบ "ท" "ทท" "ททททททททท" เป็นต้น

- เครื่องหมายดอกจัน (asterisk) '*' หมายความว่าตัวอักษรที่อยู่หน้าเครื่องหมายนี้อาจจะมีปรากฏในคำที่นำมาตรวจสอบ หรือไม่ก็ได้ ถ้ามีจะมีกี่ตัวก็ได้ เช่น "ก*ข" หมายถึง อาจจะมีตัวอักษร 'ก' หรือไม่มีก็ได้ ถ้ามีตัวอักษร 'ก' จะมีกี่ตัวก็ได้ แต่ตัวอักษรตัวสุดท้ายต้องมีตัวอักษร 'ข' 1 ตัว

- เครื่องหมายมหัพภาค (period) '.' หมายความถึงใช้แทนตัวอักษร อักขระพิเศษหรือตัวเลขอะไรก็ได้ เช่น"ก.[0-9]" หมายถึง ตัวอักษร 'ก' ตามด้วยตัวอักษรอักขระพิเศษหรือตัวเลขอะไรก็ได้ และต่อด้วยเลขอารบิค 0 ถึง 9 จำนวน 1 ตัวเลข

- เครื่องหมายปีกกา (braces) '{' และ '}' ใช้เมื่อต้องการแสดงจำนวนครั้งที่ซ้ำกัน เช่น "กข{2}" หมายถึงให้มีตัวอักษร 'ข' จำนวน 2 ตัว เช่น 'กขข' "กข{2,}" หมายถึงให้มีตัวอักษร 'ข' อย่างน้อย 2 ตัว เช่น 'กขขขข' และ"กข{3,5}" หมายถึงให้มีตัวอักษร 'ข' จำนวน 3-5 ตัวเท่านั้น คือ 'กขขข' 'กขขขข' และ 'กขขขขข'

- เครื่องหมายนขลิขิต (parentheses) '(' และ ')' ใช้เมื่อต้องการรวมกลุ่มตัวอักษรเข้าด้วยกันเป็นส่วนเดียวกัน เช่น "ก(ขค)*" หมายถึง มีตัวอักษร 'ก' และอาจจะ

ตามด้วยตัวอักษร ‘ขค’ กี่ตัวก็ได้ หรือไม่มีตัวอักษร ‘ขค’ ก็ได้ และ“ก(ขค){1,3}” หมายถึง มีตัวอักษร ‘ก’ แล้วจะตามด้วย ‘ขค’ จำนวน 1-3 ชุด เช่น “กขคขคขค” หรือ “กขคขค” ก็ได้

- เครื่องหมายแบคสแลช (backslash) ‘\’ เป็นเครื่องหมายที่ใช้บ่งบอกว่าตัวอักษรหรือเครื่องหมายที่อยู่ด้านหลังไม่ใช่เป็นเครื่องหมายคำสั่งของนิพจน์ปกติ เช่น ใช้เครื่องหมายแบคสแลชใส่ไว้หน้าเครื่องหมายดอกจัน (*) เป็นการสั่งให้ตรวจหาเครื่องหมายดอกจัน เป็นต้น

- เครื่องหมายแบคสแลชตามด้วยตัวอักษร ‘s’ พิมพ์เล็ก (backslash following with small ‘s’) ‘\s’ หมายถึงอักขระว่าง (white space) 1 ตัว

- เครื่องหมายแบคสแลชตามด้วยตัวอักษร ‘d’ พิมพ์เล็ก (backslash following with small ‘d’) ‘\d’ หมายถึงตัวเลขอารบิกตัวใดก็ได้ (digit) 1 ตัว

ประวัติผู้เขียนวิทยานิพนธ์

นายธเนศ เรืองรจิตปกรณ์ เกิดวันที่ 18 ธันวาคม พ.ศ. 2522 ที่กรุงเทพมหานคร สำเร็จการศึกษาระดับปริญญาตรี ศิลปศาสตรบัณฑิต วิทยาลัยศาสนศึกษา มหาวิทยาลัยมหิดลในปีการศึกษา 2544 และเข้าศึกษาต่อในหลักสูตรอักษรศาสตรมหาบัณฑิตภาควิชาภาษาศาสตร์ที่จุฬาลงกรณ์มหาวิทยาลัยเมื่อปีพ.ศ.2546