ชเนศ 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

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.