ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์	วิศวกรรมคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง การจับคู่การแปลระดับประโยคภาษาไทย – อังกฤษ โดยใช้ฟังก์ชั่นค่าใช้จ่ายแบบถ่วงน้ำหนัก

Thai – English Sentence Alignment using Weighted Cost Functions

นามผู้วิจัย นายชานน อ่อนมั่น

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(รองศาสตราจารย์อัศนีย์ ก่อตระกูล, D.Eng.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(อาจารย์หัชทัย ชาญเลขา, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(อาจารย์เทพชัย ทรัพย์นิธิ, Ph.D.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์ภุชงค์ อุทโยภาศ, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ ..... **เดือน** ..... **พ.ศ.** .....

วิทยานิพนธ์

เรื่อง

การจับคู่การแปลระดับประโยคภาษาไทย – อังกฤษ โดยใช้ฟังก์ชั่นค่าใช้จ่ายแบบถ่วงน้ำหนัก

Thai – English Sentence Alignment using Weighted Cost Functions

โดย

นายชานน อ่อนมั่น

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2555

ชานน อ่อนมั่น 2555: การจับคู่การแปลระดับประโยคภาษาไทย – อังกฤษ
โดยใช้ฟังก์ชั่นค่าใช้จ่ายแบบถ่วงน้ำหนัก ปริญญาวิศวกรรมศาสตรมหาบัณฑิต
(วิศวกรรมคอมพิวเตอร์) สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: รองศาสตราจารย์อัศนีย์ ก่อตระกูล, D.Eng. 81 หน้า

การแปลภาษาด้วยเครื่องเป็นการแปลข้อความจากภาษาหนึ่งไปเป็นอีกภาษาหนึ่งโดยอัตโนมัติ ปัจจุบันการแปลภาษาด้วยคอมพิวเตอร์มีบทบาทมากขึ้นเนื่องจากมีความต้องการแปลเอกสารจำนวนมากขึ้นภายใต้เวลาและบุคลากรที่จำกัด เทคนิคที่ใช้ในการแปลภาษาด้วยเครื่องมี 2 เทคนิค ได้แก่ การแปลด้วยกฎ และการแปลด้วยคลังประโยค การแปลด้วยกฎจะรับคำศัพท์และกฎการแปลจากนักภาษาศาสตร์คำนวณโดยตรง ซึ่งจะแตกต่างจากการแปลด้วยคลังประโยคที่จะสกัดความรู้สำหรับการแปลจากตัวอย่างการแปลในคลังประโยคคู่ขนานเองโดยอัตโนมัติ การแปลด้วยคลังประโยคจึงสามารถขยายผลไปสู่คู่ภาษาหรือโดเมนอื่นได้ง่ายกว่าการแปลด้วยกฎ

การสร้างคู่ตัวอย่างการแปลทีละประโยคด้วยมือต้องใช้ค่าใช้จ่ายสูง ดังนั้นจึงมีงานวิจัยที่พัฒนาเทคนิคการจับคู่การแปลระดับประโยคจากเอกสารขนานโดยอัตโนมัติ เทคนิคการจับคู่การแปลระดับประโยคในงานที่มีมาแล้วมักคำนวณค่าใช้จ่ายการจับคู่การแปลจากคู่คำแปลหรือความน่าจะเป็นของคู่คำแปลที่สกัดจากคลังประโยคคู่ขนานตั้งต้น จากทรัพยากรทางภาษาที่มีอยู่แล้ว เช่น พจนานุกรม เวิร์ดเน็ต เป็นต้น หรือจากลักษณะเด่นร่วมกันของคู่ภาษา เช่น เครื่องหมายวรรคตอน เป็นต้น จากงานวิจัยที่มีมาแล้วพบว่าการจับคู่การแปลระดับประโยคมักได้ความถูกต้องต่ำสำหรับเอกสารที่แปลโดยมุ่งเน้นการสื่อความหมายและไม่มุ่งเน้นความครบถ้วนของเอกสารภาษาต้นทางหรือแปลโดยอรรถ ซึ่งทำให้เกิดการจับคู่การแปลแบบแทรกและแบบลบปริมาณมาก นอกจากนี้คู่เอกสารภาษาไทย – อังกฤษมีลักษณะเด่นร่วมกันของคู่ภาษาและทรัพยากรทางภาษาน้อยเมื่อเทียบกับคู่ภาษาอื่นจึงไม่สามารถประยุกต์ใช้วิธีการที่มีมาแล้วได้โดยตรง ดังนั้นวิทยานิพนธ์นี้จึงมีวัตถุประสงค์เพื่อศึกษาและพัฒนาเทคนิคสำหรับการจับคู่การแปลระดับประโยคภาษาไทย – อังกฤษ วิธีที่เสนอเป็นการใช้ฟังก์ชั่นค่าใช้จ่ายแบบถ่วงน้ำหนักซึ่งคำนวณจากข้อสนเทศ 3 กลุ่ม ได้แก่ รูปแบบการจับคู่ ความยาวประโยค และความน่าจะเป็นการแปล โดยตัวแปรและค่าถ่วงน้ำหนักของแต่ละฟังก์ชั่นค่าใช้จ่ายจะคำนวณจากคลังประโยคคู่ขนานตั้งต้น

งานวิจัยนี้ทำการทดลองไขว้แบบ 5 ทบเพื่อวัดประสิทธิภาพของโปรแกรมจับคู่คำแปลระดับประโยคที่พัฒนาขึ้น กับเอกสารที่แปลโดยอรรถเกี่ยวกับการท่องเที่ยวจำนวน 115 ย่อหน้าซึ่งประกอบด้วย 941 คู่ประโยค ผลการทดลองพบว่ามีค่าความแม่นยำ ค่าความครอบคลุม และค่าคะแนนเอฟ ที่ดีที่สุดเป็น 0.728, 0.752 และ 0.740 ตามลำดับ นอกจากนี้ยังทดลองกับเอกสารที่แปลแบบทุกข้อความหรือที่เรียกว่าแปลโดยพยัญชนะในโดเมนกฎหมายจำนวน 1,632 คู่ประโยค พบว่าผลความแม่นยำ ความครอบคลุมและค่าคะแนนเอฟมีค่าเป็น 0.998

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Chanon Onman 2012: Thai – English Sentence Alignment using Weighted Cost Functions.
Master of Engineering (Computer Engineering),
Major Field: Computer Engineering, Department of Computer Engineering.
Thesis Advisor: Associate Professor Asanee Kawtrakul, D.Eng. 81 pages.

Machine translation plays an important role these days, since translating a large quantity of text in limited time is required. There are 2 approaches in machine translation which are rule based and corpus based approach. Rule based approach is to translate text by using rules and vocabularies provided by computational linguists. In the other hand, corpus based approach is to automatically extract translation knowledge from parallel corpus. Accordingly, corpus based approach is easier to extend translation capability and apply to new language pairs or domains than rule based approach.

However, manually creating parallel corpus is still laborious work, so developments of automatic sentence alignment tools are required. In previous works, sentence alignment tools utilized language information from initial parallel corpus, existing resources such as dictionary, Wordnet, or common linguistic features of specific language pairs such as punctuation markers, cognate. Previous sentence alignment tools have been proved to yield high accuracy with literal translated document. But in case of free translated document, there are some modifications in translation that subsequently cause insertion and deletion alignments and make the alignment task has more complexity and yields lower accuracy. Moreover, free translated documents tend to contain more loose translation examples than literal translated document and such examples are required for training machine translation in order to generate more naturally translation. The objective of this work is to study and develop a technique for aligning Thai – English parallel sentence in free translated documents. Since Thai – English language pair shares a few common linguistic features and have specific linguistic phenomena which causes more complexity in alignment such as using unknown words, word-description translation pairs, accordingly, a new alignment technique is needed. In this work, the proposed technique uses weighted cost functions which are analyzed from 3 features consisting of alignment type, sentence length, and translation probability. All of variables and weighting values are computed from an initial parallel corpus.

In this work, a 5-fold cross validation was performed with a set of free translated documents which consists of 115 paragraphs consisting of 941 sentence pairs. The proposed technique yields 0.728 precision, 0.752 recall, and 0.740 F-score. This technique is also used to evaluate with a literal document pair which is in legal domain. The experiment result yields 0.998 in precision and recall.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณร.ศ.ดร.อัศนีย์ ก่อตระกูล ดร.หัชทัย ชาญเลขาและดร.เทพชัย ทรัพย์นิธิ ที่ช่วยให้คำปรึกษา แนวคิดและแนวทางการทำวิจัย

ขอขอบคุณสมาชิกศูนย์ความรู้เฉพาะด้านวิศวกรรมความรู้และวิศวกรรมภาษา (Center of Excellence for Unified Knowledge and Language Engineering, U-Know CoE) และห้องปฏิบัติการวิจัยเทคโนโลยีภาษาธรรมชาติและความหมาย (Language and Semantic Technology Laboratory, LST) ที่ช่วยตรวจสอบและให้ข้อมูลที่มีประโยชน์ต่องานวิจัย

วิทยานิพนธ์นี้ได้รับทุนสนับสนุนการทำวิจัยจากสำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ (สวทช.) ในโครงการทุนสถาบันบัณฑิตวิทยาศาสตร์และเทคโนโลยีไทย (Thailand Graduate Institute of Science and Technology, TGIST) เลขที่สัญญา TG-44-11-53-066M

ชานน อ่อนมั่น

มิถุนายน 2555

สารบัญ

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

การจับคู่การแปลระดับประโยคภาษาไทย – อังกฤษ โดยใช้ฟังก์ชั่นค่าใช้จ่ายแบบถ่วงน้ำหนัก

Thai – English Sentence Alignment using Weighted Cost Functions

คำนำ

การแปลภาษาด้วยเครื่อง (Machine translation, MT) เป็นการแปลข้อความจากภาษาหนึ่งไปเป็นอีกภาษาหนึ่งโดยอัตโนมัติ ปัจจุบันการแปลภาษาด้วยเครื่องมีบทบาทมากขึ้นเนื่องจากมีความต้องการแปลเอกสารจำนวนมากขึ้นภายใต้เวลาและบุคลากรที่จำกัด เทคนิคที่ใช้ในการแปลภาษาด้วยเครื่องมี 3 เทคนิค ได้แก่ การแปลโดยใช้กฎ (Rule based machine translation, RBMT) (Arnold, 1986; Teerapong *et al.*, 2005; Systran, 2012) การแปลโดยใช้แบบจำลองสถิติ (Statistical machine translation, SMT) (Brown *et al.*, 1993; Yamada and Knight, 2002; Koehn, 2005; Ortiz-Martínez *et al.*, 2005; Aaron B., 2011) และการแปลโดยใช้ตัวอย่างการแปล (Example based machine translation, EBMT) (Sato and Nagao, 1990; Al-Adhaileh and Enya Kong, 1999; Kritsuthikul, 2006; Aaron B., 2011) การแปลโดยใช้กฎเป็นเทคนิคแรกที่ถูกนำมาใช้ในการแปลแปลภาษาด้วยเครื่อง ซึ่งเป็นการแปลผ่านทางการทำความเข้าใจประโยคที่ต้องการแปล โดยใช้กฎ และพจนานุกรมที่สร้างขึ้นโดยนักภาษาศาสตร์คำนวณ ทำให้การแปลโดยใช้กฎจะสามารถปรับแต่งและคาดเดาผลการแปลได้ง่ายกว่าแนวทางอื่น แต่อย่างไรก็ตามการขยายขอบเขตความสามารถการแปลโดยการสร้างกฎการแปลให้ครอบคลุมสถานการณ์ทั้งหมดเป็นเรื่องยาก นอกจากนี้การขยายผลไปสู่ภาษาหรือโดเมนอื่นจำเป็นต้องใช้เวลาและค่าใช้จ่ายในการสร้างพจนานุกรมและชุดกฎ จึงมีแนวคิดที่จะสกัดความรู้สำหรับการแปลแบบอัตโนมัติโดยเรียนรู้จากตัวอย่างคู่ประโยคแทน ด้วยเหตุนี้จึงมีการเสนอวิธีการแปลโดยใช้แบบจำลองทางสถิติและการแปลโดยใช้ตัวอย่างการแปลซึ่งอาศัยความสามารถของคอมพิวเตอร์ที่มีมากขึ้น การแปลโดยใช้ตัวอย่างไม่จำเป็นต้องใช้กฎการแปลที่กำหนดด้วยมือหรือโดยคนแต่โปรแกรมจะสร้างแบบจำลองการแปลเชิงสถิติ หรือสกัดตัวอย่างการแปลจากคลังข้อความสองภาษา (Bitext) และใช้ความรู้ที่สกัดได้นี้ไปแปลข้อความอื่น ทำให้ในระยะหลังเทคนิคนี้ได้รับความนิยมมากขึ้น (Hutchins, 2006)

คลังข้อความสองภาษา (Bitex) คือชุดของข้อความหรือเอกสารต่างภาษาที่สัมพันธ์กันในลักษณะใดลักษณะหนึ่งซึ่งในบางกรณีอาจมีมากกว่าสองภาษา เช่น ข่าวต่างภาษาในช่วงเวลาใกล้เคียงกันซึ่งไม่จำเป็นต้องเป็นข้อความที่เป็นคู่การแปลกันทั้งหมด สำหรับการแปลภาษาด้วยเครื่องคลังข้อความสองภาษาจะหมายถึงข้อความที่เป็นคู่การแปลกัน (Translation equivalence) คลัง

ข้อความสองภาษาสำหรับการแปลภาษาด้วยเครื่องแบ่งออกได้เป็น 2 ประเภท (Bowker and Pearson, 2002) ได้แก่

1.) คลังข้อความสองภาษาแบบเทียบได้ (Comparable text) หมายถึง คู่ของชุดเอกสารที่ไม่ใช่คู่แปลแต่กล่าวถึงเรื่องเดียวกัน เช่น บทความวิจัยต่างภาษาที่มีเนื้อหาเกี่ยวกับหัวข้อเดียวกัน

2.) คลังข้อความสองภาษาแบบขนาน (Parallel text) หมายถึงคู่ของชุดเอกสารที่มีคู่การแปล เช่น คู่มือการใช้งานอุปกรณ์ไฟฟ้า

สำหรับการสร้างระบบแปลภาษาอัตโนมัติมักนิยมใช้คลังข้อความสองภาษาแบบขนานมากกว่าการใช้คลังข้อความแบบเทียบได้ เนื่องจากคลังข้อความสองภาษาแบบเทียบได้นั้นมีวัตถุประสงค์เพื่อให้ผู้อ่านสามารถทำความเข้าใจข้อความภาษาปลายทางร่วมกับบริบทอื่นของข้อความภาษาต้นทางจึงทำให้ความเป็นคู่กันลดลงไป แต่อย่างไรก็ตามการสร้างคลังข้อความสองภาษาด้วยมือหรือโดยคนต้องใช้ทรัพยากรสูงทั้งในด้านเวลาและบุคคลที่มีความรู้สองภาษา ดังนั้นจึงมีงานวิจัยที่พัฒนาเทคนิคเพื่อการจับคู่การแปลข้อความสองภาษาแบบอัตโนมัติ (Bitext alignment) จากคลังเอกสารสองภาษาเพื่อให้ได้คลังประโยคคู่ขนาน (Parallel sentence)

การจับคู่การแปลข้อความสองภาษามี 3 ระดับ ได้แก่ ระดับเอกสาร ระดับประโยค และระดับวลีหรือคำ (Tiedemann, 2011) การจับคู่ระดับเอกสารมักทำกับคลังข้อความสองภาษาแบบเทียบได้เพื่อระบุคู่เอกสารที่มีเนื้อหาใกล้เคียงกันที่สุดและใช้เป็นข้อมูลนำเข้าสำหรับการจับคู่ระดับอื่นต่อไป (Fung and Cheung, 2004; Munteanu and Marcu, 2005) การจับคู่การแปลระดับวลีหรือคำเป็นส่วนสำคัญสำหรับการฝึกฝนโปรแกรมแปลภาษาเพื่อสร้างคู่คำแปลหรือตัวอย่างการแปลจากประโยคคู่ขนาน (Parallel sentence) เนื่องจากคุณภาพของการจับคู่ระดับวลีหรือคำขึ้นกับคุณภาพของการจับคู่ระดับประโยค ดังนั้นงานวิจัยนี้จึงเน้นการจับคู่ระดับประโยค

ลักษณะการแปลของข้อความในสองภาษามีอยู่ 2 ลักษณะ ได้แก่ การแปลตามพยัญชนะหรือการแปลครบทุกคำ (Literal translation) และการแปลโดยอรรถหรือการแปลเอาความ (Free translation) การแปลครบทุกคำเป็นการแปลตรงตัวทั้งถ้อยคำและเนื้อความ เช่น เอกสารทางการหรือบันทึกการประชุม แต่การแปลโดยอรรถจะเป็นการแปลเอาความโดยมุ่งเน้นการสื่อความหมายเป็นสำคัญและไม่มุ่งเน้นรักษาความถูกต้องแม่นยำของต้นฉบับทุกถ้อยคำ เช่น ข่าวหรือนิตยสารสองภาษา (พรรณา, 1991) ซึ่งส่วนที่ถูกเปลี่ยนแปลงนี้จะถูกกำกับให้เป็นการจับคู่แบบลบหรือแบบแทรกตามนิยามของ Brown (1991) (ดูรายละเอียดเพิ่มเติมในหัวข้อตรวจเอกสาร) ดังนั้นการพัฒนาเทคนิคเพื่อจับคู่การแปลระดับประโยคกับเอกสารที่แปลโดยอรรถทำได้ยากกว่าเอกสารที่แปลครบทุกคำ อย่างไรก็ตามเอกสารสองภาษาที่แปลโดยอรรถมีปริมาณมากกว่าเอกสารที่มีการแปลครบทุกคำ

ดังนั้นงานวิจัยนี้จึงมุ่งเน้นการพัฒนาเทคนิคเพื่อจับคู่การแปลระดับประโยคจากเอกสารแบบขนานที่แปลโดยอรรถ

เทคนิคที่ใช้ในการจับคู่การแปลระดับประโยคในช่วงแรกมักใช้ข้อสนเทศจากความยาวประโยคเพียงอย่างเดียว (Gale and Church, 1991; Brown *et al.*, 1991) เพราะเป็นคุณสมบัติที่ดีและสามารถใช้ได้กับทุกคู่ภาษา แต่การใช้ข้อสนเทศจากความยาวประโยคเพียงอย่างเดียวอาจได้ความถูกต้องต่ำ จึงมีการใช้ร่วมกับข้อสนเทศทางภาษาอื่น เช่น การใช้เครื่องหมายวรรคตอน ชื่อเฉพาะ ตัวเลข รากศัพท์ หรือคู่คำแปลจากพจนานุกรมสองภาษาที่มีอยู่แล้ว (Wu, 1994; Tannin *et al.*, 1998; Melamed, 1999; Al-Adhaileh *et al.*, 2001; Ma, 2006; Uchiyama and Isahara, 2007; Moe, 2008; Mamitimin and Hou, 2009; Li *et al.*, 2010) และแม้ว่าการใช้ข้อสนเทศทางภาษาเหล่านี้จะให้ความถูกต้องสูงขึ้น แต่จำเป็นต้องมีทรัพยากรทางภาษาภายนอกมาก่อน หรือคู่ภาษาต้องมีลักษณะเด่นร่วมกัน แนวทางนี้จึงขึ้นอยู่กับภาษาหรือโดเมนที่ทำการจับคู่การแปลระดับประโยค อีกแนวทางหนึ่งในการจับคู่การแปลระดับประโยคคือการสกัดเอาข้อสนเทศทางภาษาจากคลังประโยคคู่ขนานตั้งต้น และใช้ข้อสนเทศที่สกัดได้นี้ไปใช้ในการจับคู่การแปลระดับประโยคจากเอกสารที่ไม่เคยเห็น ทำให้ข้อสนเทศที่สกัดได้ไม่ขึ้นกับคู่ภาษาหรือโดเมน แต่จำเป็นต้องมีคลังประโยคสำหรับฝึกฝนบางส่วน (Moore, 2002; Munteanu and Marcu, 2005; Németh *et al.*, 2005; Slayden et al,, 2010) นอกจากนี้การใช้ข้อสนเทศแบบผสมจากหลายแหล่งมาคิดรวมกันอาจทำให้ผลกระทบของแต่ละข้อสนเทศไม่เท่ากัน ซึ่งจากงานวิจัยที่ผ่านมาจะใช้วิธีใส่ค่าถ่วงน้ำหนักให้กับแต่ละข้อสนเทศ โดยค่าถ่วงน้ำหนักนี้จำเป็นต้องมีการปรับแต่งสำหรับคลังประโยคนั้นๆ (Németh *et al.*, 2005; Ma, 2006; Li *et al.*, 2010) ทำให้ค่าถ่วงน้ำหนักนี้ไม่สามารถหาได้อย่างอัตโนมัติ และยังอาจขึ้นกับคลังเอกสารนั้นๆอีกด้วย

สำหรับงานวิจัยนี้จะเน้นการจับคู่การแปลระดับประโยคของคู่ภาษาไทย – อังกฤษ ซึ่งมีความแตกต่างจากคู่ภาษาอื่น เพราะมีความแตกต่างกันในระดับไวยากรณ์สูงและมีจำนวนทรัพยากรทางภาษาน้อยเมื่อเทียบกับคู่ภาษาอื่น นอกจากนี้ภาษาไทยไม่มีจุดที่จะใช้บ่งบอกขอบเขตประโยคได้อย่างชัดเจน ทำให้การสกัดคู่ประโยคจากเอกสารคู่ภาษาไทย-อังกฤษต้องคำนึงถึงการกำกับขอบเขตประโยคภาษาไทยไปพร้อมกันด้วย ในงานวิจัยที่มีมาแล้ว สามารถแยกแนวทางการจับคู่การแปลระดับประโยคสำหรับคู่เอกสารภาษาไทย-อังกฤษได้ 2 แนวทาง ระหว่างแนวทางการแยกการพิจารณาการกำกับขอบเขตประโยคจากการจับคู่ประโยคออกจากกัน และแนวทางการแบ่งขอบเขตประโยคภาษาไทยไปพร้อมกับการจับคู่การแปลโดยอาศัยจุดที่อาจเป็นจุดแบ่งขอบเขตประโยค แนวทางแรกทำให้ได้ขอบเขตประโยคที่แท้จริงซึ่งทำให้โปรแกรมจับคู่การแปลระดับประโยคทำงานง่ายขึ้น แต่จำเป็นต้องมีคลังประโยคภายนอกสำหรับการฝึกฝนโปรแกรมแบ่งขอบเขตประโยค

ภาษาไทย (Slayden *et al.*, 2010) แนวทางที่สองจะทำให้เกิดส่วนของข้อความภาษาไทยจำนวนมากซึ่งจะทำให้การจับคู่การแปลทำได้ยากขึ้น แต่ก็ไม่จำเป็นต้องใช้คลังประโยคภายนอก (Tannin *et al.*, 1998; Moe, 2008)

งานวิจัยนี้มีจุดมุ่งหมายเพื่อศึกษาและพัฒนาเทคนิคการจับคู่การแปลระดับประโยคกับเอกสารที่เป็นคู่การแปลภาษาไทย-อังกฤษ ในแนวทางการแบ่งส่วนภาษาไทยแบบที่สองร่วมกับการใช้ความน่าจะเป็นการแปลซึ่งยังไม่มีการทดลองกับคู่ภาษาไทย – อังกฤษ ขอบเขตของงานวิจัยนี้จะรับข้อมูลป้อนเข้าที่กำกับจุดที่อาจเป็นจุดแบ่งขอบเขตของประโยคภาษาไทยด้วยมือ งานวิจัยนี้ทดลองเปรียบเทียบผลการจับคู่การแปลระดับประโยคระหว่างเอกสารที่แปลโดยอรรถสองภาษาคือ เอกสารเกี่ยวกับการท่องเที่ยว และเอกสารที่แปลตามตัวอักษรเป็นรัฐธรรมนูญแห่งราชอาณาจักรไทยฉบับปี พ.ศ.2550 เอกสารทั้งสองนี้มีจุดแตกต่างกันคือ ข้อความในรัฐธรรมนูญมักจะถูกแปลอย่างครบถ้วนสมบูรณ์มากกว่าข้อความจากนิตยสารการท่องเที่ยว ทำให้การจับคู่การแปลบนรัฐธรรมนูญมีแนวโน้มที่จะทำได้อย่างตรงไปตรงมามากกว่า วิธีการที่ใช้ในงานวิจัยนี้เป็นการใช้ฟังก์ชั่นค่าใช้จ่ายแบบถ่วงน้ำหนักซึ่งคำนวณจากข้อสนเทศผสมจากรูปแบบการจับคู่การแปล (ไทย-อังกฤษ) ซึ่งมี 4 กลุ่มได้แก่ แบบลบ (1-0) แบบแทรก (0-1) แบบแทนที่ (1-1) และแบบตัดทอน (m-1) จากอัตราส่วนความยาวประโยคในหน่วยจำนวนอักขระของประโยคภาษาอังกฤษต่อภาษาไทย และจากค่าความน่าจะเป็นการแปล โดยตัวแปรและค่าถ่วงน้ำหนักของแต่ละฟังก์ชั่นค่าใช้จ่ายจะถูกคำนวณจากคลังประโยคคู่ขนานตั้งต้นโดยอัตโนมัติ และเพิ่มการคิดค่าใช้จ่ายแบบฮิวริสติกสำหรับการจับคู่การแปลแบบแทรกและแบบลบโดยเฉพาะ

วัตถุประสงค์

งานวิจัยนี้ศึกษาและพัฒนาฟังก์ชั่นค่าใช้จ่ายแบบถ่วงน้ำหนักสำหรับการจับคู่การแปลระดับประโยคจากคู่เอกสารภาษาไทย-อังกฤษที่แปลโดยอรรถ โดยใช้

1. ฟังก์ชั่นค่าใช้จ่ายที่คำนวณจากรูปแบบการจับคู่การแปล จากความยาวประโยค และจากความน่าจะเป็นการแปล

2. ค่าถ่วงน้ำหนักของแต่ละฟังก์ชั่นค่าใช้จ่ายที่ปรับโดยอัตโนมัติ โดยพิจารณาจากคลังประโยคคู่ขนานตั้งต้น

ขอบเขตของงานวิจัย

1. ศึกษา พัฒนา และวัดผลเปรียบเทียบเทคนิคการจับคู่การแปลระดับประโยคสำหรับคู่เอกสารภาษาไทย – อังกฤษ

2. ใช้เอกสารที่เป็นตัวอย่างของการแปลโดยอรรถจากบทความแปลไทย – อังกฤษในนิตยสารเกี่ยวกับการท่องเที่ยวจำนวน 115 ย่อหน้า 941 คู่ประโยค

3. ใช้เอกสารที่เป็นตัวอย่างของการแปลตามตัวอักษรจากรัฐธรรมนูญแห่งราชอาณาจักรไทย ฉบับปี พ.ศ.2550 จำนวน 1,632 คู่ประโยค

4. เอกสารตั้งต้นมีการกำกับขอบเขตประโยค หรือจุดที่อาจเป็นจุดบ่งบอกขอบเขตประโยคภาษาไทยมาแล้วด้วยมือ

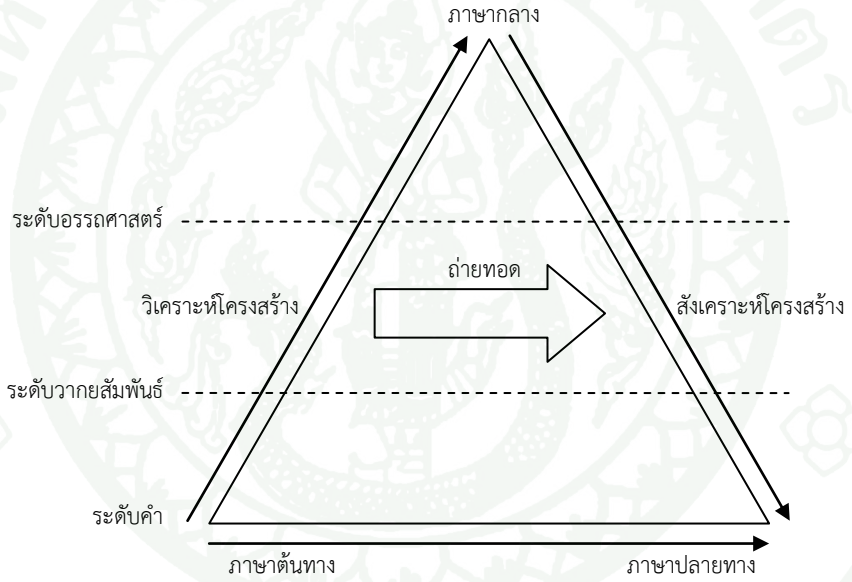
การตรวจเอกสาร

ความรู้พื้นฐาน

การแปลภาษาด้วยเครื่อง

การแปลภาษาด้วยเครื่องหรือโปรแกรมแปลภาษา (Machine translation, MT) เป็นส่วนหนึ่งของการประมวลผลภาษาธรรมชาติ โดยเป็นการแปลข้อความจากภาษาหนึ่งไปเป็นอีกภาษาหนึ่งโดยอัตโนมัติ งานวิจัยเกี่ยวกับการแปลภาษาด้วยเครื่องเริ่มได้รับความสนใจอย่างมากในสหรัฐอเมริกาเมื่อมากกว่า 50 ปีมาแล้วนับตั้งแต่ช่วงสงครามเย็น ในช่วงต้นนั้นจะเน้นไปที่การแปลบทความภาษาอังกฤษ – รัสเซียเกี่ยวกับทางด้านวิทยาศาสตร์ และการทหาร โดยประยุกต์มาจากแนวคิดการเข้าและถอดรหัสสัญญาณวิทยุที่ใช้ในช่วงสงครามโลกครั้งที่ 2 กับการเขียนกฎขึ้นมาวิเคราะห์เชิงวากยสัมพันธ์ และอรรถศาสตร์ แต่คอมพิวเตอร์ในช่วงเวลานั้นมีประสิทธิภาพต่ำมากเมื่อเทียบกับคอมพิวเตอร์ในยุคปัจจุบัน ประกอบกับการขาดทรัพยากรทางภาษาสำหรับคอมพิวเตอร์ จึงทำให้งานวิจัยการแปลภาษาด้วยเครื่องในสมัยแรกไม่สามารถแปลบทความในปริมาณมาก หรือแปลข้อความได้ในขอบเขตที่จำกัด ส่งผลให้งานวิจัยการแปลภาษาด้วยเครื่องในสหรัฐอเมริกาได้รับความสนใจน้อยลง โดยเฉพาะอย่างยิ่งหลังจากมีรายงานประเมินว่าการแปลภาษาด้วยเครื่องมีต้นทุนที่สูงกว่าแต่ได้คุณภาพการแปลที่ต่ำกว่าการแปลโดยคนของคณะกรรมการ Automatic Language Processing Advisory Committee (ALPAC) ปีค.ศ. 1966 แต่ในบางประเทศยังคงวิจัยด้านการแปลภาษาด้วยเครื่องต่อไป เช่น แคนาดา ญี่ปุ่น และเครือสหภาพยุโรป แต่จะเน้นไปที่การแปลบันทึกการประชุมรัฐสภา หรือบทความทางธุรกิจมากขึ้น เช่น บทความเกี่ยวกับธุรกรรม หรือคู่มือการใช้สินค้า เป็นต้น การแปลภาษาด้วยเครื่องได้รับความนิยมมากขึ้นในช่วงทศวรรษ 1990 หลังจากที่มีการเสนอแนวทางการแปลภาษาด้วยเครื่องแบบใหม่ด้วยการเรียนรู้จากตัวอย่างการแปล และการใช้แบบจำลองทางสถิติ ประกอบกับคอมพิวเตอร์มีความสามารถมากขึ้นจึงสามารถประมวลผลและแปลบทความในปริมาณที่มากขึ้นได้ อีกทั้งทรัพยากรทางภาษาก็มีมากขึ้น จึงทำให้งานวิจัยเกี่ยวกับการแปลภาษาด้วยเครื่องได้รับความนิยมมากขึ้นเรื่อยๆ จนกระทั่งปัจจุบัน (Hutchins, 2006)

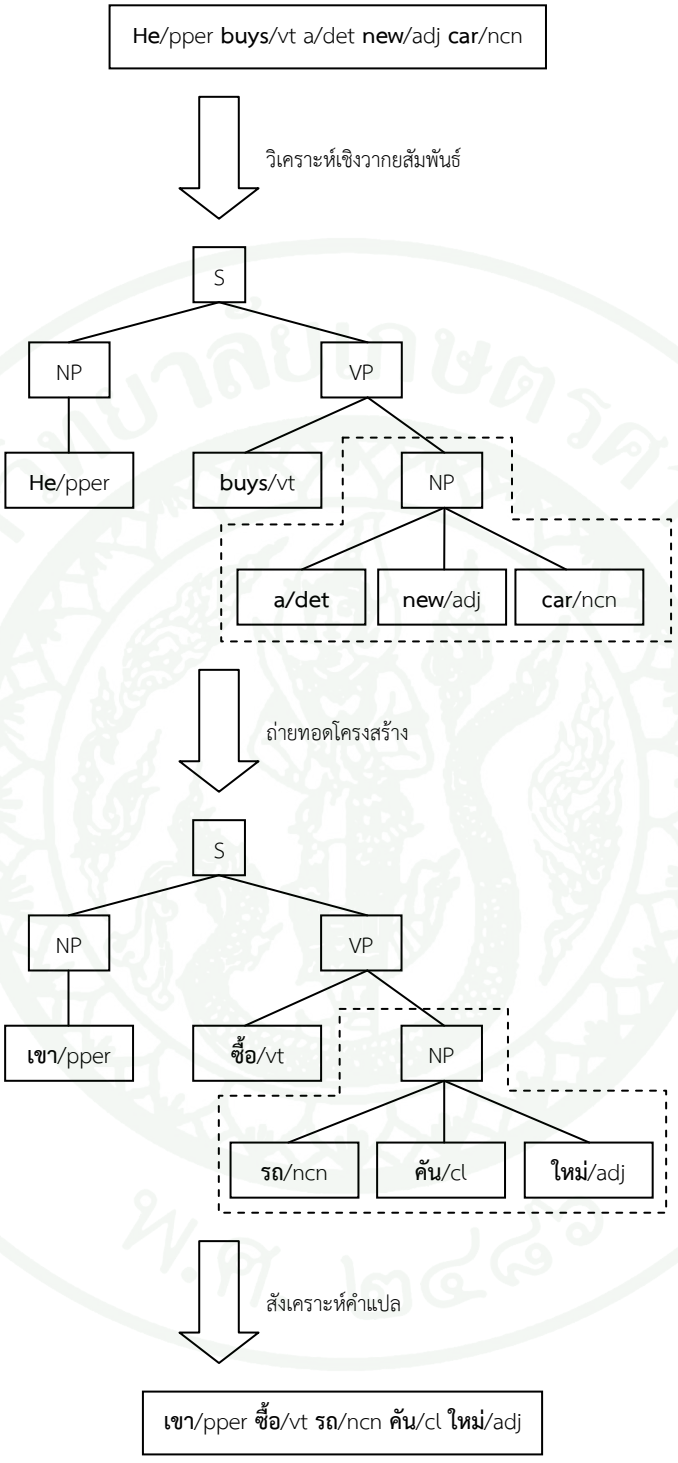
การแปลข้อความจากภาษาหนึ่งไปเป็นอีกภาษาหนึ่งสามารถทำได้ตั้งแต่การแทนที่คำแปลโดยใช้พจนานุกรมสองภาษา (Word substitution) หรือการใช้คำแปลที่จดจำไว้อยู่แล้ว (Translation memory, TM) แต่การแทนที่ด้วยคำแปลเพียงอย่างเดียวอาจยังไม่เพียงพอที่จะสร้างประโยคในภาษาปลายทางที่ดีได้ หรือคำแปลที่จดจำไว้แล้วอาจไม่ครอบคลุม ดังนั้นจึงจำเป็นต้องมีการวิเคราะห์ (Analysis) ประโยคในระดับที่สูงขึ้น ได้แก่ ระดับวากยสัมพันธ์ (Syntactic) ระดับอรรถศาสตร์ (Semantic) และภาษากลาง (Interlingual) จากนั้นจึงถ่ายทอด (Transfer) โครงสร้างที่วิเคราะห์ได้ไปเป็นโครงสร้างของภาษาปลายทาง และสังเคราะห์ (Synthesis) ข้อความในภาษาปลายทางจากโครงสร้างนี้อีกทีหนึ่ง ซึ่งสามารถเขียนเป็นแผนภาพพีระมิดของโวกัวส์ (Vauquois's pyramid) ดังภาพที่ 1



ภาพที่ 1 แผนภาพการแปลภาษาด้วยเครื่อง

ที่มา: http://en.wikipedia.org/wiki/Machine_translation

การวิเคราะห์เชิงวากยสัมพันธ์ (Syntactic analysis) เป็นการวิเคราะห์ความสัมพันธ์ของคำในประโยคเพื่อสร้างต้นไม้ไวยกรณ์ที่เป็นตัวแทนของประโยคที่ต้องการแปล จากนั้นโปรแกรมแปลภาษาจะแปลคำและคำนวณโครงสร้างต้นไม้ไวยกรณ์ที่ถูกต้องตามหลักไวยกรณ์ของภาษาปลายทางเพื่อสร้างเป็นคำแปลผลลัพธ์ ดังตัวอย่างที่แสดงในภาพที่ 2



ภาพที่ 2 ตัวอย่างการวิเคราะห์และถ่ายทอดในระดับวากยสัมพันธ์

เนื่องจากประโยคที่มีคำศัพท์หรือโครงสร้างต้นไม้ไวยกรณ์ที่ต่างกันอาจสามารถสื่อความหมายอย่างเดียวกัน ดังเช่นตัวอย่างที่ 1 ประโยคทั้งสี่ใช้คำศัพท์และโครงสร้างประโยคที่แตกต่างกันแต่ทุกประโยคมีความหมายตรงกัน ดังนั้นถ้ามีการวิเคราะห์เชิงวากยสัมพันธ์เพียงอย่างเดียว โปรแกรมแปลภาษาอาจจำเป็นต้องมีกฎสำหรับการสังเคราะห์คำแปลให้ครอบคลุมทั้งสี่กรณี

ตัวอย่างที่ 1

"She gave a book to John."

"She gave John a book."

"John is given a book by her."

"A book is given to John by her"

การวิเคราะห์เชิงอรรถศาสตร์ (Semantic analysis) เป็นการวิเคราะห์ประโยคเพื่อหาความหมายหรือมโนทัศน์ (Concept) ของสิ่งที่ต้องการสื่อสาร ในการประมวลผลภาษาด้วยคอมพิวเตอร์จะแทนความหมายของประโยคด้วยเพรดิเคตลอจิก (Predicate logic) เช่น ประโยคทั้งสี่ในตัวอย่างที่ 1 สามารถแทนได้ด้วยประพจน์ "give(she, book, John)"

การแทนประโยคด้วยภาษากลาง (Interlingua) เป็นการสร้างตัวแทนของข้อความด้วยโครงข่ายความหมาย (Semantic network) ซึ่งเกิดจากความสัมพันธ์ทางความหมายและทรัพยากรความรู้ เช่น ออนโทโลยี (Ontology) ตัวแทนในภาษากลางจะอ้างอิงเนื้อหาที่สื่อสารด้วยโหนดมโนภาพ (Concept node) ซึ่งไม่ขึ้นกับภาษาใด ทำให้การแทนประโยคด้วยภาษากลางมีข้อดีคือมีการวิเคราะห์ประโยคเพียงครั้งเดียว จากนั้นจะสามารถใช้ตัวแทนที่วิเคราะห์ได้แปลไปเป็นภาษาใดก็ได้ ตัวอย่างตัวแทนแบบภาษากลางที่นิยมคือ Universal Networking Language (UNL)

เทคนิคการแปลภาษาด้วยเครื่อง

การแปลด้วยกฎหรือการแปลโดยใช้ความรู้ (Rule based machine translation หรือ Knowledge based translation) เป็นเทคนิคแรกที่ถูกนำมาใช้ในการแปลภาษาด้วยเครื่อง การแปลโดยใช้แนวทางนี้จะเน้นไปที่การวิเคราะห์ (Analysis) เพื่อทำความเข้าใจภาษา การแปลโดยใช้แนวทางนี้คอมพิวเตอร์จะสร้างตัวแทน (Representation) ในระดับวากยสัมพันธ์ อรรถศาสตร์ หรือภาษากลาง เพื่อแทนความหมายของข้อความที่ต้องการแปลโดยใช้กฎทางภาษาศาสตร์ จากนั้นจึงใช้กฎอีกชุดหนึ่งในการสังเคราะห์คำแปลในภาษาปลายทาง (Generating) ด้วยเหตุนี้การแปลโดยใช้กฎจึงสามารถคาดเดาผลการแปลได้ และสามารถปรับแต่งผลการแปลได้ง่ายกว่าแนวทางอื่น เนื่อง

ความรู้ในการแปลทั้งหมดถูกกำหนดด้วยมือหรือโดยคน แต่อย่างไรก็ตามการขยายขอบเขตความสามารถการแปลโดยการสร้างกฎการแปลให้ครอบคลุมสถานการณ์ทั้งหมดเป็นเรื่องยาก นอกจากนี้การขยายผลไปสู่ภาษาหรือโดเมนอื่นจำเป็นต้องใช้เวลาและค่าใช้จ่ายในการสร้างพจนานุกรมและชุดกฎ ตัวอย่างโปรแกรมแปลภาษาที่ใช้แนวทางการแปลด้วยกฎ ได้แก่ Eurotra ภาษิต (Parsit) Systran เป็นต้น (Arnold, 1986; Teerapong *et al.*, 2005; Systran, 2012)

ในช่วงต้นทศวรรษ 1990 ได้มีจุดเปลี่ยนแปลงของงานวิจัยการแปลภาษาด้วยเครื่องคือ การเสนอวิธีการแปลโดยใช้คลังประโยคคู่ขนาน (Corpus based machine translation) วิธีการแปลแบบนี้แบ่งออกเป็น 2 วิธีการที่แตกต่างกัน วิธีการแรกคือการแปลด้วยสถิติ (Statistical based machine translation) และอีกวิธีการหนึ่งคือการแปลโดยใช้ตัวอย่าง (Example based machine translation)

การแปลด้วยสถิติใช้แบบจำลองทางสถิติ 2 แบบจำลอง ได้แก่ แบบจำลองการแปล (Translation model) และแบบจำลองภาษา (Language model) แบบจำลองการแปลจะระบุว่าคำหรือวลีในภาษาต้นทาง (w_{src}) จะมีโอกาสถูกแปลไปเป็นคำใดในภาษาปลายทาง (w_{target}) ด้วยความน่าจะเป็นเท่าใด ($P(w_{target} \mid w_{src})$) ซึ่งค่าตัวแปรเหล่านี้จะถูกคำนวณขึ้นจากตัวอย่างการแปลหรือคลังประโยคคู่ขนาน (ดูในหัวข้อประโยคคู่ขนาน) ตารางที่ 1 แสดงตัวอย่างความน่าจะเป็นการแปลภาษาไทย – อังกฤษที่คำนวณจากโปรแกรม GIZA++ (Och and Ney, 2003) ซึ่งจะเห็นได้ว่าคำหนึ่งคำสามารถแปลไปเป็นคำในอีกภาษาหนึ่งได้มากกว่าหนึ่งตัวเลือก แต่คำที่เป็นคู่การแปลกัน เช่น beach – หาด/ชายหาด มักมีความน่าจะเป็นที่สูงกว่าคำที่ไม่ได้เป็นคู่การแปลกัน เช่น beach – ระยะทาง

ตารางที่ 1 ตัวอย่างความน่าจะเป็นการแปล

w_{src}	w_{target}	$P(w_{target} \mid w_{src})$
beach	ชายหาด	0.31
beach	หาด	0.62
beach	ระยะทาง	2.17×10^{-6}
ชายหาด	beach	0.99

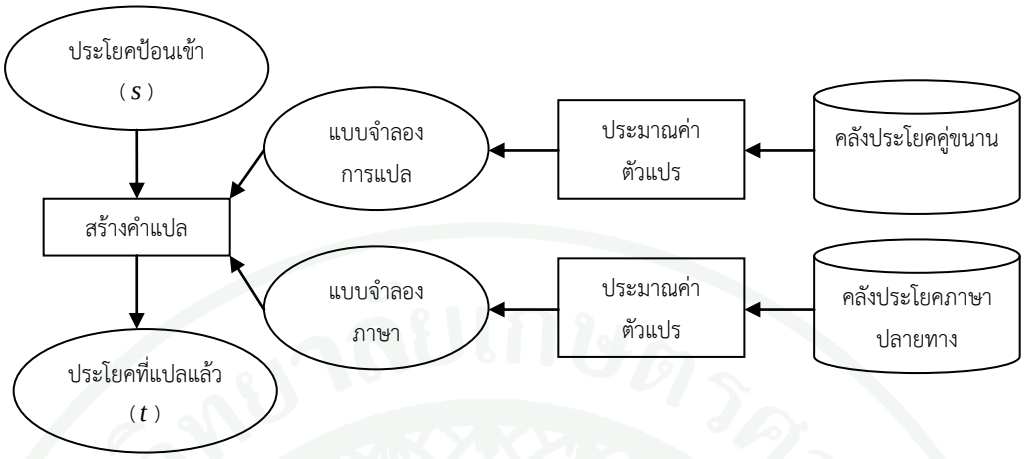
แบบจำลองภาษา (Language model) จะระบุว่าประโยคในภาษาหนึ่งมีโอกาสเกิดขึ้นเท่าใด โดยค่าตัวแปรสำหรับแบบจำลองภาษาสามารถคำนวนจากคลังข้อความภาษาเดียว (Monolingual text) ในการแปลภาษาด้วยเครื่องจะคำนวณความน่าจะเป็นการเกิดประโยคบนสมมติฐานว่าการเกิดคำที่ i ในประโยคจะขึ้นอยู่กับ n คำก่อนหน้า (n-gram model) ดังนั้นความน่าจะเป็นในการเกิดประโยค s ที่มี m คำ $p(s)$ จะสามารถคำนวณได้จากสมการต่อไปนี้

$$s = w_1 w_2 \ldots w_m \tag{1}$$

$$p(s) = p(w_1)\,p(w_2 \mid w_1)\,p(w_2 \mid w_2, w_1)\ldots p(w_i \mid w_{i-1}, w_{i-2}, \ldots, w_{i-n})\ldots p(w_m \mid w_{m-1}, w_{m-2}, \ldots, w_{m-n}) \tag{2}$$

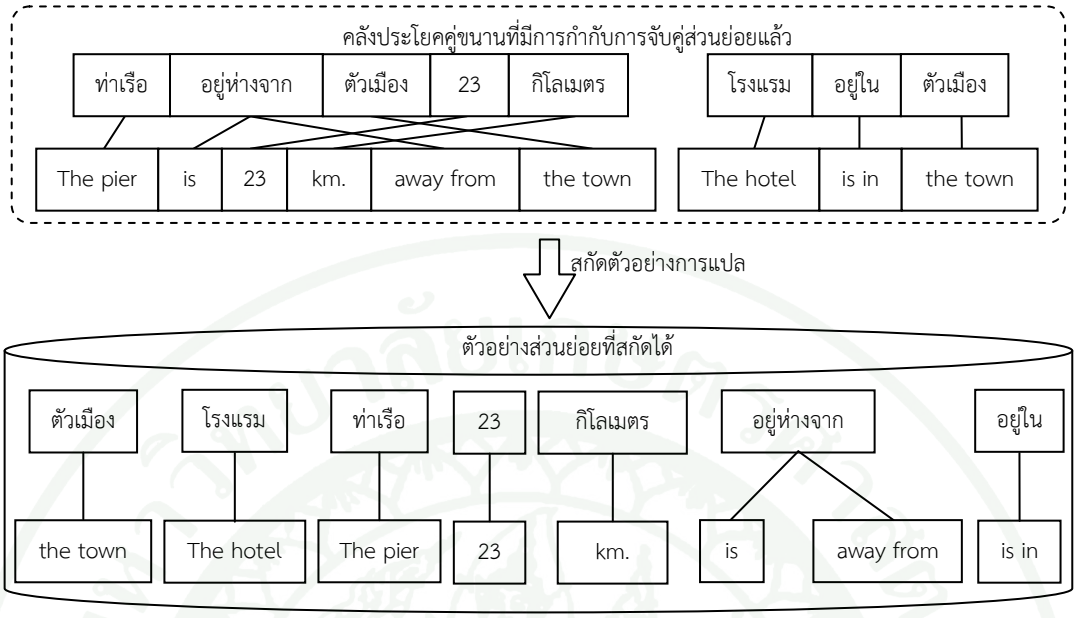
โปรแกรมแปลภาษาที่แปลด้วยสถิติ (Statistical based machine translation) จะใช้แบบจำลองการแปลและแบบจำลองภาษาในการคำนวณหาคำแปลที่มีความน่าจะเป็นสูงที่สุดเมื่อพิจารณาจากแบบจำลองทั้งสองนี้พร้อมกันดังที่แสดงในภาพที่ 3 และสมการที่ 3 ถ้าสมมติให้ s เป็นประโยคป้อนเข้าที่ต้องการแปลและให้ t เป็นคำแปล โปรแกรมจะเริ่มจากการหาคำแปล $\hat{t}$ ที่สอดคล้องกับ s หรือมีค่าความน่าจะเป็นในการแปล $P(\hat{t} \mid s)$ สูง แต่ทั้งนี้นอกจากโปรแกรมจะหาคำแปลที่สอดคล้องกับประโยคป้อนเข้าแล้ว คำแปลที่โปรแกรมจะสร้างขึ้นจะต้องมีความเป็นธรรมชาติ นั่นคือคำแปลที่จะสร้างขึ้นนี้ต้องมีความน่าจะเป็นที่จะพบในคลังประโยคภาษาปลายทาง $P(\hat{t})$ ด้วยเช่นกัน (Brown *et al.*, 1993) และเนื่องจากการแปลโดยสถิติไม่จำเป็นต้องอาศัยความรู้ทางภาษาศาสตร์จากคนนี้เอง จึงทำให้คาดคะเนหรือปรับแต่งผลการแปลทำได้ยากกว่าการแปลโดยใช้กฎ

$$t = \arg\max_{\hat{t}} P(\hat{t})P(\hat{t} \mid s) \tag{3}$$

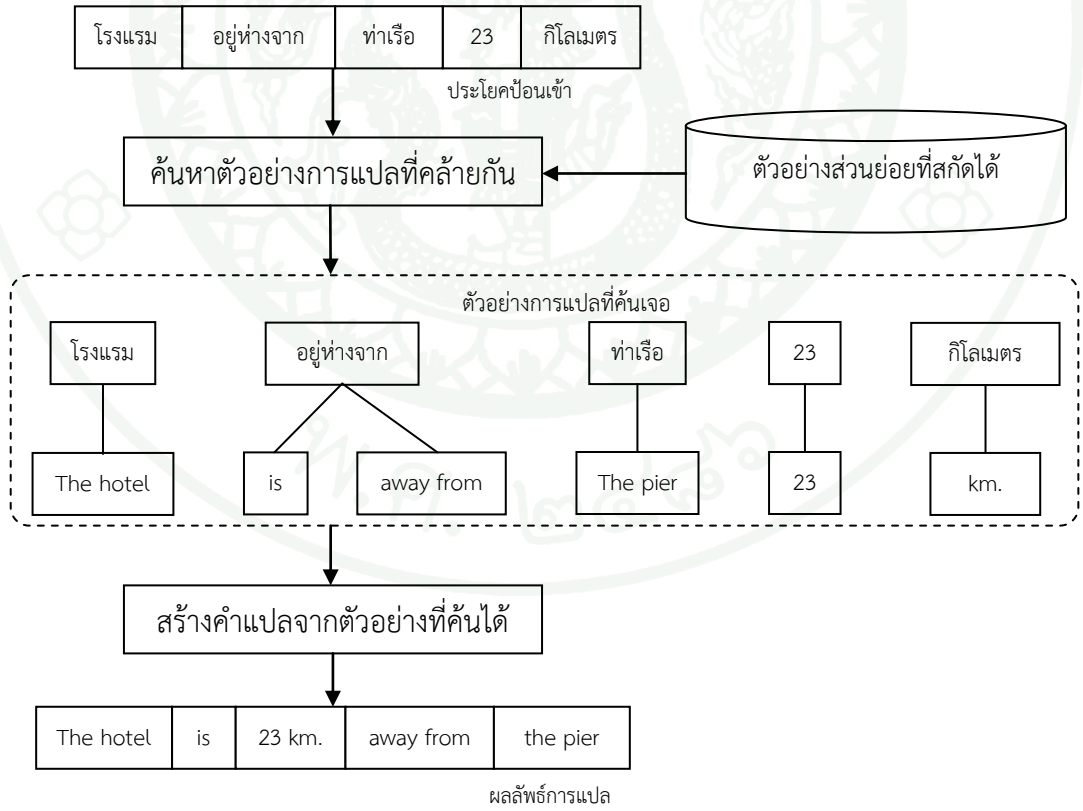


ภาพที่ 3 ภาพรวมการแปลโดยใช้สถิติ

การแปลโดยใช้ตัวอย่าง (Example based machine translation) จะเป็นการแปลโดยการสร้างคำแปลจากตัวอย่างที่คล้ายกัน การแปลโดยใช้ตัวอย่างมีแนวคิดมาจากการแปลโดยใช้คำแปลที่จดจำไว้ก่อน (Translation memory) และแนวคิดที่ว่าการแปลประโยคหนึ่งๆ ผู้แปลอาจไม่จำเป็นต้องวิเคราะห์ประโยคที่จะแปลมากนั้น เพียงแต่หาตัวอย่างคำหรือส่วนของข้อความที่คล้ายกันแล้วใช้คำแปลเดิมที่มีอยู่แล้วมาประกอบกันขึ้นเป็นคำแปลใหม่ โปรแกรมแปลภาษาโดยใช้ตัวอย่างจะแบ่งการทำงานเป็น 2 ส่วน ได้แก่ ส่วนของการสกัดตัวอย่างการแปลย่อยจากตัวอย่างคู่ประโยคที่มีการจับคู่ส่วนการแปลย่อยไว้แล้ว และส่วนค้นหาประโยคที่คล้ายกันเพื่อสร้างคำแปลใหม่ ดังที่แสดงในภาพที่ 5 และภาพที่ 4 โดยที่คำแปลใหม่ที่ได้นี้อาจเกิดจากการผสมกันระหว่างประโยคหลายๆ ประโยคที่ค้นเจอก็ได้ ผลการแปลด้วยวิธีนี้จะขึ้นกับคู่ประโยคตัวอย่างเป็นหลัก ซึ่งคู่ประโยคนี้จำเป็นต้องมีการจับคู่การแปลในระดับวลีหรือคำระหว่างประโยคในสองภาษา ซึ่งใช้ความรู้ทางภาษาศาสตร์น้อยกว่าการสร้างกฎขึ้นมาวิเคราะห์ประโยคในเชิงลึก แต่การแทนที่ด้วยตัวอย่างการแปลทีละส่วนก็ยังอาจทำให้คำแปลที่ได้ไม่เป็นธรรมชาติ (Sato and Nagao, 1990; Al-Adhaileh and Enya Kong, 1999; Aaron B., 2011) ตารางที่ 2 แสดงข้อเปรียบเทียบของเทคนิคการแปลแต่ละแบบ



ภาพที่ 5 ตัวอย่างการสกัดตัวอย่างการแปลส่วนย่อย



ภาพที่ 4 ตัวอย่างการสร้างคำแปลจากตัวอย่างการแปลที่มีอยู่แล้ว

ตารางที่ 2 สรุปคุณลักษณะเทคนิคการแปลภาษาด้วยคอมพิวเตอร์

เทคนิค	ข้อดี	ข้อด้อย
การแปลโดยใช้กฎ (Rule based machine translation)	1. ไม่จำเป็นต้องใช้คลังประโยคคู่ขนานในการสร้างโปรแกรมแปลภาษา 2. ปรับผลการแปลได้ง่าย 3. ใช้ทรัพยากรในการคำนวณน้อย 4. แปลข้อความนอกโดเมนและในโดเมนได้คุณภาพใกล้เคียงกัน 5. สามารถย้อนดูได้ว่าคำแปลสร้างมาจากกฎใดบ้าง	1. เพิ่มข้อยกเว้น (Exception) ยาก เพราะอาจขัดแย้งกับกฎที่มีอยู่แล้ว 2. คำแปลไม่เป็นธรรมชาติ 3. มีค่าใช้จ่ายในการสร้างสูง 4. เปลี่ยนโดเมนหรือคู่ภาษาได้ยาก
การแปลโดยใช้สถิติ (Statistical based machine translation)	1. ไม่จำเป็นต้องสร้างกฎการแปล 2. คำแปลเป็นธรรมชาติ 3. มีค่าใช้จ่ายในการสร้างต่ำ 4. เปลี่ยนโดเมนหรือคู่ภาษาได้ง่าย	1. ต้องใช้คลังประโยคคู่ขนานในการสร้างโปรแกรมแปลภาษา 2. ใช้ทรัพยากรในการคำนวณมาก 3. ปรับกฎเกณฑ์การแปลยาก 4. แปลข้อความนอกโดเมนได้คุณภาพต่ำ 5. ไม่สามารถย้อนดูได้ว่าคำแปลคำนวณมาจากตัวแปรใดบ้าง
การแปลโดยใช้ตัวอย่าง (Example based machine translation)	1. ไม่จำเป็นต้องสร้างกฎการแปล 2. ปรับกฎเกณฑ์การแปลโดยการแก้ไขตัวอย่างการแปลที่มีอยู่เดิม 3. มีค่าใช้จ่ายในการสร้างต่ำกว่าแบบใช้กฎ แต่สูงกว่าแบบใช้สถิติ เพราะต้องกำกับส่วนย่อยด้วย 4. เปลี่ยนโดเมนหรือคู่ภาษาได้ง่าย 5. สามารถย้อนดูได้ว่าคำแปลสร้างมาจากตัวอย่างใดบ้าง	1. ต้องใช้คลังประโยคคู่ขนานในการสร้างโปรแกรมแปลภาษา 2. ต้องจับคู่ส่วนการแปลย่อย 3. ใช้ทรัพยากรในการคำนวณมาก 4. คำแปลไม่เป็นธรรมชาติ 5. แปลข้อความนอกโดเมนได้คุณภาพต่ำ

วิธีการแปล

วิธีการแปลเอกสารสามารถแบ่งออกได้เป็น 2 วิธี ได้แก่ การแปลตามพยัญชนะ (Literal translation style) และการแปลโดยอรรถ (Free translation style) การแปลตามพยัญชนะจะเป็นการแปลตรงตัวทุกคำ และมักไม่มีการเปลี่ยนแปลง เพิ่ม ลดทอน หรือโยกย้ายรายละเอียดใดๆ ตัวอย่างเอกสารที่เป็นการตามพยัญชนะ ได้แก่ กฎหมาย หรือเอกสารทางการ เป็นต้น ตัวอย่างที่ 2 แสดงข้อความภาษาไทยและภาอังกฤษซึ่งเป็นส่วนหนึ่งของรัฐธรรมนูญแห่งราชอาณาจักรไทยฉบับปี พ.ศ.2550 และสามารถแสดงการจับคู่ข้อความในระดับอนุพากย์ (Unit) ได้ดังตารางที่ 3

ตัวอย่างที่ 2
ข้อความภาษาไทย : “เมื่อรัฐสภาให้ความเห็นชอบกับร่างพระราชบัญญัติประกอบรัฐธรรมนูญแล้ว ก่อนนำขึ้นทูลเกล้าทูลกระหม่อมถวายเพื่อทรงลงพระปรมาภิไธย ให้ส่งศาลรัฐธรรมนูญพิจารณาความชอบด้วยรัฐธรรมนูญ ซึ่งต้องกระทำให้แล้วเสร็จภายในสามสิบวันนับแต่วันที่ได้รับเรื่อง”
ข้อความภาษาอังกฤษ : “Upon its approval by the National Assembly, an organic law bill shall be, prior to its presentation to the King for signature, referred to the Constitutional Court for determining its constitutionality, provided that such determination shall be completed within thirty days as from the date of its receipt.”

ตารางที่ 3 ตัวอย่างข้อความที่แปลตามพยัญชนะ

ข้อความภาษาไทย	ข้อความภาษาอังกฤษ
เมื่อรัฐสภาให้ความเห็นชอบกับร่างพระราชบัญญัติประกอบรัฐธรรมนูญแล้ว	Upon its approval by the National Assembly, an organic law bill shall be,
ก่อนนำขึ้นทูลเกล้าทูลกระหม่อมถวายเพื่อทรงลงพระปรมาภิไธย	prior to its presentation to the King for signature,
ให้ส่งศาลรัฐธรรมนูญพิจารณาความชอบด้วยรัฐธรรมนูญ	referred to the Constitutional Court for determining its constitutionality,
ซึ่งต้องกระทำให้แล้วเสร็จภายในสามสิบวันนับแต่วันที่ได้รับเรื่อง	provided that such determination shall be completed within thirty days as from the date of its receipt.

การแปลโดยอรรถจะเป็นการแปลเอาความโดยจะมุ่งเน้นไปที่การสื่อความหมายเป็นสำคัญ และไม่มุ่งเน้นรักษาความถูกต้องแม่นยำของต้นฉบับทุกถ้อยคำ ทำให้เนื้อความของเอกสารที่แปลโดยอรรถอาจมีการเปลี่ยนแปลงตามความเหมาะสม ตัวอย่างเอกสารที่เป็นการแปลโดยอรรถ ได้แก่ นวนิยาย เรื่องสั้น นิทาน เป็นต้น (พรรณา, 1991) ตัวอย่างที่ 3 แสดงข้อความส่วนหนึ่งจากบทคัดย่อโครงงานวิศวกรรมซึ่งมี 2 ภาษา และแสดงการจับคู่ข้อความได้ดังที่แสดงในตารางที่ 4

ตัวอย่างที่ 3

ข้อความภาษาไทย : “ปัญหาการจราจรติดขัดของรถยนต์ในเมืองมีความรุนแรงเพิ่มขึ้นเรื่อยๆ เนื่องจากจำนวนรถยนต์บนท้องถนนที่เพิ่มขึ้นเรื่อยๆ และอีกส่วนหนึ่งเกิดขึ้นจากการผู้ที่ใช้รถใช้ถนนไม่ได้มีการวางแผนการเดินทางที่ดี”

ข้อความภาษาอังกฤษ : “Today's Traffic problem is becoming more and more savoir in most major city. This is not only due to the amount of vehicles on the roads that have continuously increased but also cause by most drivers who didn’t plan their route before departing.”

ตารางที่ 4 ตัวอย่างข้อความที่แปลโดยอรรถ

ข้อความภาษาไทย	ข้อความภาษาอังกฤษ
ปัญหาการจราจรติดขัดของรถยนต์ในเมืองมีความรุนแรง เพิ่มขึ้นเรื่อยๆ	Today's Traffic problem is becoming more and more savoir in most major city.
เนื่องจากจำนวนรถยนต์บนท้องถนนที่เพิ่มขึ้นเรื่อยๆ	-
-	This is not only due to the amount of vehicles on the roads that have continuously increased
และอีกส่วนหนึ่งเกิดขึ้นจากการผู้ที่ใช้รถใช้ถนนไม่ได้มีการวางแผนการเดินทางที่ดี	but also cause by most drivers who didn’t plan their route before departing.

คลังประโยคคู่ขนาน

ประโยคคู่ขนาน (Parallel sentence) คือ คู่ประโยคที่จับคู่กันระหว่างประโยคในภาษาต้นทาง และคำแปลของประโยคนั้นในภาษาปลายทาง ตัวอย่างประโยคคู่ขนานไทย-อังกฤษแสดงในตัวอย่างที่ 4 ซึ่งเป็นส่วนหนึ่งของข้อความจากนวนิยายแปล

ตัวอย่างที่ 4
ข้อความภาษาไทย : “เขามักจะถูกพูดถึงว่าเป็นคน ‘ตัวเล็ก ผิวขาวซีด สายตาสั้น’ รวมทั้ง ‘มองโลกในแง่ร้าย เสแสร้ง และฉลาดหลักแหลม’”
ข้อความภาษาอังกฤษ : “He was variously described as ‘small, pale, myopic’ and ‘cynical, pretentious and brilliant’.”

ในปัจจุบันมีคลังประโยคคู่ขนานขนาดใหญ่ที่มีการเปิดให้ใช้หลายคู่ภาษาดังที่แสดงไว้ในตารางที่ 5 และคลังประโยคคู่ขนานภาษาไทยดังที่แสดงไว้ในตารางที่ 6 คลังประโยคคู่ขนานมีความสำคัญในขั้นตอนการฝึกฝนโปรแกรมแปลภาษา เนื่องจากคลังประโยคคู่ขนานจะถูกใช้ในการคำนวณแบบจำลองการแปล การสกัดตัวอย่างการแปล และการกำหนดกฎการแปล นอกจากนี้คลังประโยคคู่ขนานจะถูกใช้เป็นตัวอย่างเปรียบเทียบผลลัพธ์การแปลระหว่างผลการแปลด้วยเครื่องและผลการแปลด้วยมือหรือโดยคนในขั้นตอนการวัดประสิทธิภาพของโปรแกรมแปลภาษา เช่น การคำนวณค่าคะแนน BLEU (Papineni *et al.*, 2002) เป็นต้น

ตารางที่ 5 ตัวอย่างคลังประโยคคู่ขนานที่เปิดให้ใช้แล้ว

คลังประโยค (ปีที่เปิดให้ใช้)	คู่ภาษา	ชนิดของเอกสารที่รวบรวม	ปริมาณประโยค (คู่ประโยค)
Canadian hansard (1995)	อังกฤษ – ฝรั่งเศส	บันทึกการประชุมรัฐสภา	ประมาณ 3.9 ล้าน
Europarl (2005)	11 ภาษาในเครือสหภาพยุโรป	บันทึกการประชุมสหภาพยุโรป	ประมาณ 11.5 ล้าน
Hongkong Parallel Text (2004)	จีน – อังกฤษ	กฎหมาย และบันทึกการประชุมรัฐสภา	ประมาณ 2.6 ล้าน
NTCIR-6 (2007)	ญี่ปุ่น – อังกฤษ	สิทธิบัตร	ประมาณ 2 ล้าน
Hunglish Corpus (2005)	ฮังการี – อังกฤษ	เอกสารทางกฎหมาย นิยาย และคู่มือการใช้ซอฟแวร์	ประมาณ 4 ล้าน
SEA lang library (2005)	อังกฤษ และ 14 ภาษาที่ใช้ในกลุ่มประเทศเอเชียตะวันออกเฉียง	นิยาย และบทความสำหรับการเรียนรู้ภาษาอังกฤษ	ประมาณ 6.5 ล้าน
English-Norwegian Parallel Corpus (1997)	อังกฤษ – นอร์เวย์	นิยาย	ไม่มีข้อมูล
English-Swedish Parallel Corpus (2001)	อังกฤษ – สวีเดน	นิยาย	ไม่มีข้อมูล

หมายเหตุ คู่ประโยคใน Europarl แต่ละประโยคมี 10 คำแปลในภาษาอื่นๆ
คู่ประโยคใน SEA lang library แต่ละประโยคอาจมีคำแปลในภาษาอื่นเพียงภาษาเดียว

ตารางที่ 6 ตัวอย่างคลังประโยคคู่ขนานภาษาไทยที่ยังไม่ได้เปิดให้ใช้จากภายนอก

คลังประโยค (หน่วยงานที่พัฒนา)	**คู่ภาษา**	**ชนิดของเอกสารที่รวบรวม**	**ปริมาณประโยค (คู่ประโยค)**
คลังประโยคคู่ขนาน (มหาวิทยาลัยเกษตรศาสตร์)	ไทย – อังกฤษ	นิตยสาร และบทความการท่องเที่ยว	ประมาณ 5,000
คลังประโยคคู่ขนานไทย – อังกฤษ (NECTEC)	ไทย – อังกฤษ	ตัวอย่างข้อความสำหรับการเรียนรู้ภาษาอังกฤษ เอกสารประกอบของซอฟท์แวร์เสรี และบทบรรยายภาพยนตร์	ประมาณ 400,000
London Olimpic (NECTEC)	ไทย – จีน – อังกฤษ	บทความเกี่ยวกับกีฬา คมนาคม การค้าขาย และการท่องเที่ยว	ประมาณ 60,000
BText (NECTEC)	ไทย – จีน – อังกฤษ	บทสนทนาที่ใช้ในเชิงการท่องเที่ยว	ประมาณ 28,000

คลังข้อความสองภาษา

คลังข้อความสองภาษา (Bitex) คือชุดของข้อความหรือเอกสารต่างภาษาที่สัมพันธ์กันในลักษณะใดลักษณะหนึ่งซึ่งในบางกรณีอาจมีมากกว่าสองภาษา เช่น ข่าวต่างภาษาในช่วงเวลาใกล้เคียงกันซึ่งไม่จำเป็นต้องเป็นข้อความที่เป็นคู่การแปลกันทั้งหมด สำหรับการแปลภาษาด้วยเครื่องคลังข้อความสองภาษาจะหมายถึงข้อความที่เป็นคู่การแปลกัน คลังข้อความสองภาษาสามารถแบ่งออกได้เป็น 2 ประเภท ได้แก่ คลังข้อความสองภาษาแบบขนาน (Parallel text) และคลังข้อความสองภาษาแบบเทียบได้ (Comparable text) (Bowker and Pearson, 2002) นิยามของระดับเอกสารคู่ขนานแต่ละประเภทเป็นดังนี้

1. คลังข้อความสองภาษาแบบเทียบได้ (Comparable text) หมายถึง คู่ของชุดเอกสารที่ไม่ใช่คู่แปลแต่กล่าวถึงเรื่องเดียวกัน เช่น บทความวิจัยต่างภาษาที่มีเนื้อหาเกี่ยวกับหัวข้อเดียวกัน หรือข่าวจากต่างสำนักข่าวที่กล่าวถึงเรื่องเดียวกัน เป็นต้น ภาพที่ 6 แสดงตัวอย่างของเอกสารแบบเทียบได้ซึ่งเป็นข่าวจาก 3 สำนักข่าว ซึ่งเห็นได้ว่าข่าวทั้งสามนี้กล่าวถึงเรื่องบริษัทเอกชนให้การสนับสนุนนักกีฬาเยาวชนเหมือนกัน แต่เนื้อความของข่าวจะไม่เหมือนหรือเป็นคู่การแปลกันทั้งหมด

2. คลังข้อความสองภาษาแบบขนาน (Parallel text) หมายถึงคู่ของชุดเอกสารที่มีคู่การแปล เช่น กฎหมาย หรือคู่มือการใช้งานอุปกรณ์ไฟฟ้า เป็นต้น ภาพที่ 7 แสดงตัวอย่างเอกสารแบบขนานซึ่งเป็นคู่มือการใช้งานโทรศัพท์มือ

เอสซีจีหนุน2นักกอล์ฟยช.สาวทีมชาติไทยก้าวสู่อาชีพลุยแอลพีจีเอ

วันที่ 08 กุมภาพันธ์ พ.ศ. 2553 เวลา 15:05:47 น.

นายกานต์ ตระกูลฮุน กรรมการผู้จัดการใหญ่เครือซิเมนต์ไทย(เอสซีจี) แถลงข่าววันที่ 8 กุมภาพันธ์ให้การสนับสนุน น้องโม โมรียา และ น้องเม เอรียา จุฑานุกาล 2 นักกอล์ฟเยาวชนทีมชาติไทย ด้วยสัญญาปีต่อปี โดยจะให้การสนับสนุนเรื่องค่าใช้จ่ายต่าง ๆ ในการแข่งขันทั้งในและต่างประเทศ

รวมถึงการฝึกซ้อมและการจัดผู้เชี่ยวชาญด้านวิทยาศาสตร์การกีฬามาพัฒนาสมรรถภาพทางด้านร่างกาย การดูแลด้านจิตวิทยาและโภชนาการ เพื่อความฝันของทั้ง 2 คน ที่ก้าวสู่นักกอล์ฟอาชีพของแอลพีจีเอ ทัวร์ ในอนาคต

สำหรับในปีแรก เอสซีจีจะให้การสนับสนุนในวงเงิน 4.2 ล้านบาท โดย น้องโม และน้องเม ถือเป็นนักกอล์ฟ 2 คนแรกของไทยที่ได้รับการสนับสนุนจากเอสซีจี หลังจากก่อนหน้าได้ให้การสนับสนุนวงการแบดมินตันของไทยมากว่า 30 ปี

SCG tees off sponsorship for junior golfers

By The Nation
Published on February 10, 2010

Siam Cement Group has decided to support two Thai junior golfers to compete in local and international tournaments with the goal of establishing their place in the professional LPGA Tour.

The rising stars of the national team, sisters Moriya Jutanugarn nicknamed "Mo" and Ariya, nicknamed "May" will receive financial support, top grade scientific training in fitness and mental aspects besides professional golf skills to prepare for the tough competition against the world's best golfers.

"Supporting young sports talents is one of our social contributions," said SCG president Kan Trakulhoon. "This full sponsorship covers the expenses before and during local and international tournaments. We will arrange for the best sports science experts to help them improve their golf skills, nutrition and mental strength to make sure they will have all that it takes to fulfil their dreams. With our sponsorship we believe they can reach their full potential of becoming world class women golfers and a pride of the nation.

เอสซีจีทุ่มงบหนุน2พี่น้องสาวไทยลุยกอล์ฟอาชีพ

ไทยรัฐออนไลน์

โดย ไทยรัฐออนไลน์
9 กุมภาพันธ์ 2553, 04:00 น.

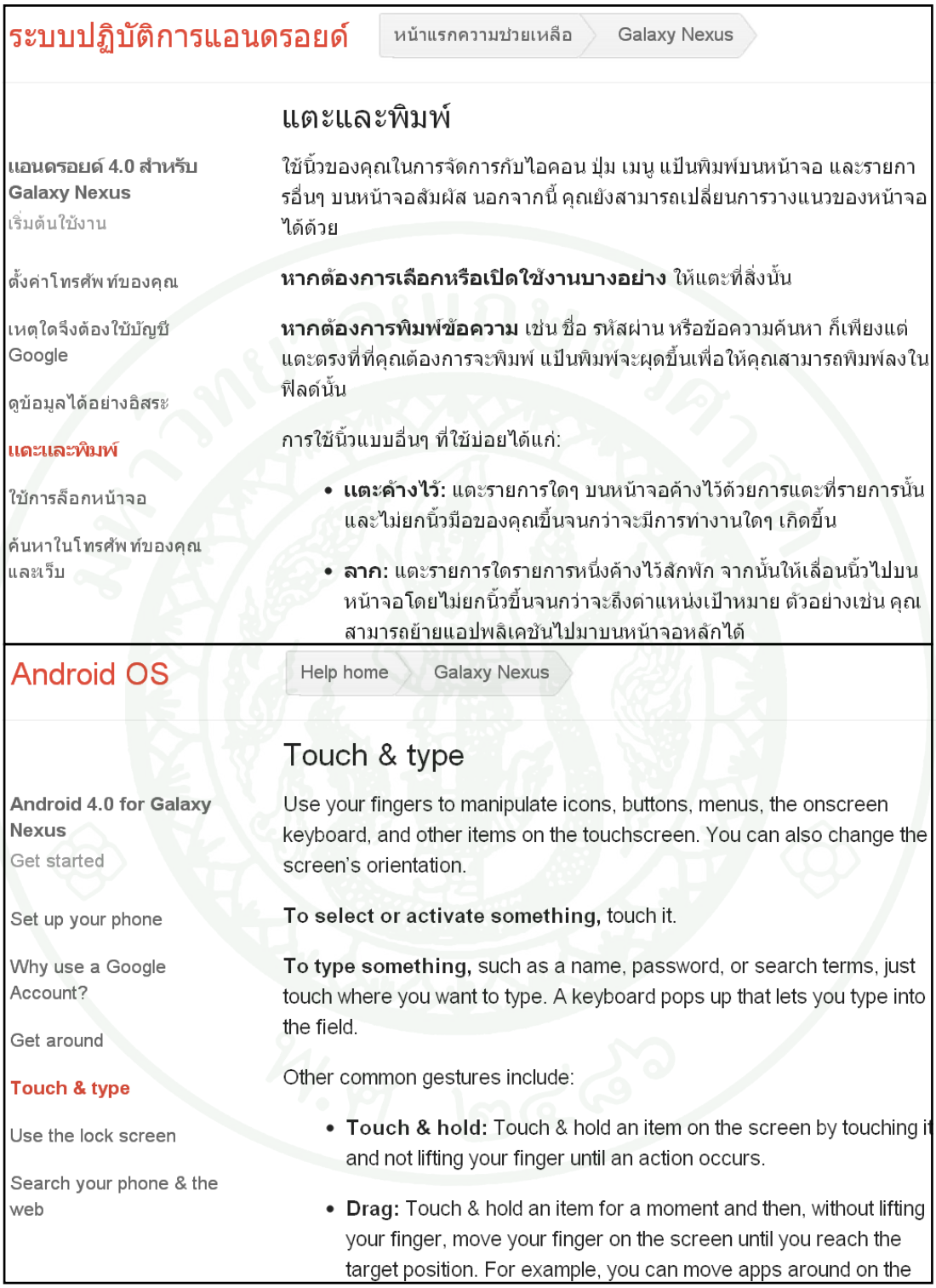
เอสซีจี ทุ่มงบ 4.2 ล้านบาท หนุน โมรียา - เอรียา จุฑานุกาล ในการแข่งขันกอล์ฟ ทั้งในและต่างประเทศ รวมถึงการฝึกซ้อมและหาผู้เชี่ยวชาญมาให้...

เมื่อวันที่ 8 ก.พ. ที่สนามกอล์ฟนวธานี นายกานต์ ตระกูลฮุน กรรมการผู้จัดการใหญ่ เอสซีจี เป็นประธานแถลงข่าวพร้อมร่วมกับสองพี่น้องนักกอล์ฟเยาวชนสาวไทย โมรียา-เอรียา จุฑานุกาล ในการสนับสนุนของ เอสซีจี โดย นายกานต์ เปิดเผยว่า ทางเอสซีจี มุ่งมั่นส่งเสริมศักยภาพของเยาวชน ด้านกีฬามาอย่างต่อเนื่อง ล่าสุด ได้สนับสนุน 2 นักกีฬากอล์ฟเยาวชนทีมชาติ คือ โมรียา จุฑานุกาล หรือ น้องโม และ เอรียา จุฑานุกาล หรือ น้องเม โดยจะดูแลค่าใช้จ่ายเป็นเงินจำนวน 4.2 ล้านบาท

สำหรับการแข่งขันทัวร์นาเมนท์ต่างๆ ทั้งในประเทศ และต่างประเทศ การฝึกซ้อมรวมไปถึงการจัดผู้เชี่ยวชาญ เพื่อพัฒนาศักยภาพตามแนวทางวิทยาศาสตร์การกีฬาอย่างเต็มรูปแบบ ซึ่งประกอบด้วย การพัฒนาทักษะการเล่น การดูแลด้านจิตวิทยา และโภชนา ทำให้มั่นใจว่า ด้วยพรสวรรค์ และฝีมือ บวกกับวิทยาศาสตร์การกีฬา จะทำให้น้องทั้งสองคนสามารถก้าวไปถึงความใฝ่ฝันได้อย่างแน่นอน นอกจากนี้ยังสร้างความภาคภูมิใจ และทำให้ความฝันของคนไทยเป็นจริงที่มีนักกอล์ฟหญิงระดับโลกเป็นคนไทย

ภาพที่ 6 ตัวอย่างเอกสารแบบเทียบได้จาก 3 สำนักข่าว

ที่มา: http://www.thairath.co.th/content/sport/63870
http://www.matichon.co.th/news_detail.php?newsid=1265616385&grpid=03&catid=03
http://www.nationmultimedia.com/home/2010/02/10/sports/SCG-tees-off-sponsorship-for-junior-golfers-30122208.html



ระบบปฏิบัติการแอนดรอยด์ | หน้าแรกความช่วยเหลือ | Galaxy Nexus

แอนดรอยด์ 4.0 สำหรับ Galaxy Nexus
เริ่มต้นใช้งาน
ตั้งค่าโทรศัพท์ของคุณ
เหตุใดจึงต้องใช้บัญชี Google
ดูข้อมูลได้อย่างอิสระ
แตะและพิมพ์
ใช้การล็อกหน้าจอ
ค้นหาในโทรศัพท์ของคุณและเว็บ

แตะและพิมพ์

ใช้นิ้วของคุณในการจัดการกับไอคอน ปุ่ม เมนู แป้นพิมพ์บนหน้าจอ และรายการอื่นๆ บนหน้าจอสัมผัส นอกจากนี้ คุณยังสามารถเปลี่ยนการวางแนวของหน้าจอได้ด้วย

หากต้องการเลือกหรือเปิดใช้งานบางอย่าง ให้แตะที่สิ่งนั้น

หากต้องการพิมพ์ข้อความ เช่น ชื่อ รหัสผ่าน หรือข้อความค้นหา ก็เพียงแต่แตะตรงที่ที่คุณต้องการจะพิมพ์ แป้นพิมพ์จะผุดขึ้นเพื่อให้คุณสามารถพิมพ์ลงในฟิลด์นั้น

การใช้นิ้วแบบอื่นๆ ที่ใช้บ่อยได้แก่:

- **แตะค้างไว้:** แตะรายการใดๆ บนหน้าจอค้างไว้ด้วยการแตะที่รายการนั้นและไม่ยกนิ้วมือของคุณขึ้นจนกว่าจะมีการทำงานใดๆ เกิดขึ้น
- **ลาก:** แตะรายการใดรายการหนึ่งค้างไว้สักพัก จากนั้นให้เลื่อนนิ้วไปบนหน้าจอโดยไม่ยกนิ้วขึ้นจนกว่าจะถึงตำแหน่งเป้าหมาย ตัวอย่างเช่น คุณสามารถย้ายแอปพลิเคชันไปมาบนหน้าจอหลักได้

Android OS | Help home | Galaxy Nexus

Android 4.0 for Galaxy Nexus
Get started
Set up your phone
Why use a Google Account?
Get around
Touch & type
Use the lock screen
Search your phone & the web

Touch & type

Use your fingers to manipulate icons, buttons, menus, the onscreen keyboard, and other items on the touchscreen. You can also change the screen's orientation.

To select or activate something, touch it.

To type something, such as a name, password, or search terms, just touch where you want to type. A keyboard pops up that lets you type into the field.

Other common gestures include:

- **Touch & hold:** Touch & hold an item on the screen by touching it and not lifting your finger until an action occurs.
- **Drag:** Touch & hold an item for a moment and then, without lifting your finger, move your finger on the screen until you reach the target position. For example, you can move apps around on the

ภาพที่ 7 ตัวอย่างเอกสารแบบขนานจากคู่มือการใช้งานโทรศัพท์มือถือ

ที่มา: http://support.google.com/ics/nexus/?hl=th

http://support.google.com/ics/nexus/?hl=en

อย่างไรก็ตามคู่เอกสารที่ถึงแม้จะเป็นคู่การแปลกันแต่บางประโยคอาจไม่ได้ถูกแปลหรือเป็นการแปลโดยอรรถ (ดูเพิ่มเติมที่หัวข้อวิธีการแปล) และคู่เอกสารต่างภาษาที่ไม่ได้กล่าวถึงเรื่องเดียวกันก็อาจยังมีประโยคคู่ขนานอยู่ จึงได้มีการเสนอระดับของเอกสารคู่ขนานอีก 2 ระดับ ได้แก่ คู่เอกสารแปลแบบขนานบางส่วน (Noissy-parallel document) และคู่เอกสารแบบเสมือนเทียบได้ (Quasi-comparable document) (Fung and Cheung, 2004) นิยามของระดับเอกสารคู่ขนานของทั้ง 2 ประเภทเป็นดังนี้

1. คู่เอกสารแปลแบบขนานบางส่วนเป็นคู่เอกสารที่เกิดจากการแปลโดยวิธีการแปลตามตัวอักษรไม่ทุกคู่ประโยค บางประโยคอาจแปลโดยอรรถ เช่น นวนิยายแปล ข่าวสองภาษา บทความแปลทั่วไป เป็นต้น

2. คู่เอกสารแบบเสมือนเทียบได้เป็นคู่เอกสารที่ไม่ได้เป็นคู่การแปลกัน และไม่ได้กล่าวถึงเรื่องหรือช่วงเวลาเดียวกัน แต่เอกสารเหล่านี้อาจยังมีประโยคคู่ขนานอยู่ก็ได้ แต่คุณภาพของประโยคคู่ขนานที่สกัดได้อาจน้อยกว่าที่สกัดได้จากเอกสารคู่ขนานแบบอื่น

การจับคู่การแปลข้อความสองภาษาแบบอัตโนมัติ

การจับคู่การแปลข้อความสองภาษาโดยอัตโนมัติ (Bitext alignment) เป็นการระบุส่วนที่สัมพันธ์กันในคลังข้อความสองภาษา การจับคู่สำหรับการสร้างคลังประโยคคู่ขนานเพื่อฝึกฝนโปรแกรมแปลภาษาจะเป็นการระบุส่วนที่เป็นคู่การแปลกัน (Translation equivalence) การจับคู่การแปลมี 3 ระดับ ได้แก่ ระดับเอกสาร ระดับประโยค และระดับวลีหรือคำ ขอบเขตและจุดประสงค์ของการจับคู่แต่ละระดับเป็นดังนี้ (Tiedemann, 2011)

1. การจับคู่การแปลระดับเอกสารเป็นระดับการจับคู่การแปลระดับแรกสุดและมักจะเป็นการประมวลผลกับข้อมูลขนาดใหญ่ เช่น การจับคู่ข่าวจากต่างสำนักข่าว การจับคู่ระดับนี้จะเป็นการระบุว่าคู่เอกสารไหนเป็นคู่การแปลกันหรืออาจมีส่วนที่เกี่ยวข้องกันอยู่ภายในจากคลังเอกสารแบบเทียบได้เพื่อสร้างผลลัพธ์เป็นคลังเอกสารขนานสำหรับการจับคู่ระดับอื่นต่อไป

2. การจับคู่การแปลระดับประโยคเป็นการระบุว่าส่วนใดในคู่เอกสารขนานเป็นส่วนที่แปลกันเพื่อสกัดประโยคคู่ขนาน ส่วนที่แปลกันนี้แม้จะมาจากเอกสารที่เป็นคู่การแปลกัน การจับคู่การแปลระดับประโยคอาจไม่สามารถทำอย่างตรงไปตรงมาได้ เพราะลำดับของข้อความแปลอาจมีการเปลี่ยนแปลงได้ (ดูหัวข้อวิธีการแปล) การจับคูในระดับนี้บางครั้งจะมีการจับคู่การแปลระดับย่อหน้า

พร้อมกันไปด้วย (Brown *et al.*, 1991; Mamitimin and Hou, 2009) การจับคู่การแปลระดับประโยคมี 2 แบบคือ การจับคู่การแปลแบบมีลำดับ และการจับคู่การแปลแบบไม่มีลำดับ (ดูเพิ่มเติมในหัวข้อการจับคู่การแปลระดับประโยค)

3. การจับคู่การแปลระดับวลีหรือคำเป็นส่วนสำคัญของการสร้างโปรแกรมแปลภาษาเพื่อสร้างคู่คำแปลหรือสกัดตัวอย่างการแปลจากประโยคคู่ขนาน (ดูหัวข้อการแปลภาษาด้วยเครื่อง)

การจับคู่การแปลระดับประโยค

การจับคู่การแปลระดับประโยค (Sentence alignment) รับข้อมูลป้อนเข้าเป็นเอกสารแบบขนาน จุดประสงค์ของการจับคู่การแปลระดับนี้เป็นการระบุว่าส่วนใดในเอกสารเป็นส่วนที่แปลกันซึ่งมักสกัดในระดับประโยคเพื่อสร้างเป็นคลังประโยคคู่ขนาน แม้ว่าประโยคจะมาจากเอกสารที่เป็นคู่การแปลกัน การจับคู่การแปลก็อาจไม่สามารถทำอย่างตรงไปตรงมาได้เพราะลำดับข้อความที่แปลหรือเนื้อความอาจมีการเปลี่ยนแปลงได้ (ดูหัวข้อวิธีการแปล) การจับคู่การแปลระดับประโยคมี 2 แบบคือ การจับคู่การแปลแบบมีลำดับ (Brown *et al.*, 1991; Wu, 1994; Melamed, 1999; Chuang and Yeh, 2005; Mamitimin and Hou, 2009; Chen, 1993; Tannin *et al.*, 1998; Al-Adhaileh *et al.*, 2001; Moore, 2002; Németh *et al.*, 2005; Uchiyama and Isahara, 2007; Moe, 2008; Ma, 2006; Li *et al.*, 2010; Slayden *et al.*, 2010) และการจับคู่การแปลแบบไม่มีลำดับ (Fung and Cheung, 2004; Munteanu and Marcu, 2005)

การจับคู่การแปลระดับประโยคแบบมีลำดับจะเป็นการจับคู่โดยมีสมมติฐานว่าลำดับการแปลมีลักษณะไปทางเดียวกัน (Monotonic) หรือไม่มีการจับคู่ที่ไขว้กัน ผลลัพธ์ของการจับคู่การแปลระดับประโยคแบบมีลำดับจะเป็นลำดับของการจับคู่การแปลย่อย (Bead) ซึ่งสามารถแทนด้วยคู่ลำดับของจำนวนประโยคที่จับคู่กันจากแต่ละเอกสาร (Brown *et al.*, 1991) รูปแบบการจับคู่การแปลย่อยสามารถจัดกลุ่มได้เป็น 6 ประเภท ดังแสดงในตารางที่ 7 (Gale and Churh, 1991)

ตารางที่ 7 รูปแบบการจับคู่การแปลย่อย

รูปแบบการจับคู่การแปลย่อย	**จำนวนประโยคที่จับคู่กัน**	
	ภาษาต้นทาง	**ภาษาปลายทาง**
แบบลบ	1	0
แบบแทรก	0	1
แบบแทนที่	1	1
แบบขยาย	1	มากกว่า 1
แบบตัดทอน	มากกว่า 1	1
แบบรวม	มากกว่า 1	มากกว่า 1

ตารางที่ 8 แสดงตัวอย่างผลลัพธ์การจับคู่การแปลระดับประโยคแบบมีลำดับที่มีประโยคภาษาไทยและภาษาอังกฤษอย่างละ 3 ประโยค จะได้รูปแบบการแปลผลลัพธ์คือลำดับ (1, 1) (1, 0) (0, 1) และ (1, 1)

ตารางที่ 8 ตัวอย่างผลลัพธ์การจับคู่การแปลระดับประโยคแบบมีลำดับ

ข้อความภาษาไทย	**ข้อความภาษาอังกฤษ**	**รูปแบบการจับคู่ย่อย**
T1: ปัญหาการจราจรติดขัดของรถยนต์ในเมืองมีความรุนแรง เพิ่มขึ้นเรื่อยๆ	**E1**: Today's Traffic problem is becoming more and more savoir in most major city.	แทนที่ (1, 1)
T2: เนื่องจากจำนวนรถยนต์บนท้องถนนที่เพิ่มขึ้นเรื่อยๆ	-	ลบ (1, 0)
-	**E2**: This is not only due to the amount of vehicles on the roads that have continuously increased	แทรก (0, 1)
T3: และอีกส่วนหนึ่งเกิดขึ้นจากการผู้ที่ใช้รถใช้ถนนไม่ได้มีการวางแผนการเดินทางที่ดี	**E3**: but also cause by most drivers who didn’t plan their route before departing.	แทนที่ (1, 1)

การจับคู่การแปลระดับประโยคแบบไม่มีลำดับจะเป็นการจับคู่ประโยคบนคู่เอกสารที่ไม่มีการรักษาลำดับการแปลหรือยอมให้มีการจับคู่แบบไขว้กันได้ การจับคู่แบบนี้สามารถแปลงให้เป็นปัญหาการคัดแยกสิ่งของ (Classification) ได้ แต่ทำให้การจับคู่แบบไม่มีลำดับจะจับคู่ได้เฉพาะแบบแทนที่ (1-1) เท่านั้น เพราะถ้าจะคำนวณเพื่อจับคู่ให้ได้ทุกรูปแบบการจับคู่ย่อยแล้วจะทำให้ปริภูมิค้นหามีขนาดใหญ่มาก การจับคู่แบบไม่มีลำดับมักทำกับคู่เอกสารที่ไม่ได้เป็นคู่การแปลกัน (Fung and Cheung, 2004; Munteanu and Marcu, 2005) ตัวอย่างเช่น จากภาพที่ 6 จะได้ผลการจับคู่แบบไม่มีลำดับ ได้แก่ ข้อความในย่อหน้าที่ 3 ของข่าวที่ 2 และ 3 เป็นต้น

งานวิจัยที่เกี่ยวข้อง

แนวทางการใช้ข้อสนเทศสำหรับการจับคู่การแปลระดับประโยคแบ่งได้ 3 แนวทางด้วยกัน ได้แก่ ข้อสนเทศจากความยาวประโยค (Purely length based alignment) ข้อสนเทศผสมระหว่างความยาวประโยคและข้อสนเทศทางภาษา (Hybrid length-lexicon based alignment) และข้อสนเทศผสมระหว่างความยาวประโยคและข้อสนเทศการแปล (Hybrid length-translation based alignment)

การจับคู่โดยใช้ข้อสนเทศจากความยาวประโยค

Gale and Church (1991) ได้เสนอวิธีการสกัดประโยคคู่ขนานจากเอกสารรายงานทางการเงินที่มี 3 ภาษา ได้แก่ อังกฤษ ฝรั่งเศส และเยอรมัน โดยกระบวนการสกัดนี้พัฒนาขึ้นจากสมมติฐานที่ว่า “ประโยคภาษาต้นทางที่ยาวมักจะถูกแปลไปเป็นประโยคภาษาปลายทางที่ยาว และประโยคต้นทางที่สั้นก็มักจะถูกแปลไปเป็นประโยคปลายทางที่สั้นเช่นเดียวกัน” วิธีที่ถูกเสนอนี้จึงใช้ความยาวประโยคในหน่วยอักขระร่วมกับวิธีทางสถิติในการคำนวณหาความน่าจะเป็นที่ประโยคจากคนละภาษาจะจับคู่กันด้วยอัตราส่วนความยาวของคู่ประโยค ในงานวิจัยนี้ได้ทำการทดลองกับคลังประโยคขนาด 1,312 คู่ประโยค ซึ่งมีการจับคู่แบบแทนที่คิดเป็น 89% แบบลบหรือแบบแทรกรวมกันประมาณ 1% แบบขยายหรือแบบตัดทอนรวมกันได้ประมาณ 8.9% และการจับคู่แบบรวมมีประมาณ 1.1% พบว่าได้ความถูกต้อง 95.8% เนื่องจากวิธีการที่เสนอนี้มีความซับซ้อนในการคำนวณต่ำและไม่จำเป็นต้องมีการวิเคราะห์ทางภาษาเพิ่มเติมจึงเป็นวิธีที่ไม่ขึ้นกับคู่ภาษา ในภายหลังจึงได้มีการนำเอาวิธีการนี้ไปสกัดประโยคคู่ขนานจากคลังประโยคคู่ขนานขนาดใหญ่ เช่น บันทึกการประชุมของกลุ่มประเทศในสหภาพยุโรป (Europarl) (Koehn, 2005)

Brown *et al.* (1991) เสนอวิธีการสกัดประโยคคู่ขนานโดยใช้อัตราส่วนความยาวประโยคในหน่วยของจำนวนคำระหว่างประโยคภาษาอังกฤษต่อประโยคในภาษาฝรั่งเศส ในบันทึกการอภิปรายรัฐสภาประเทศแคนาดา (Canadian hansard) แต่เนื่องจากคลังประโยคนี้มีขนาดใหญ่ จึงได้มีการเสนอการใช้จุดตรึง (Anchoring point) เพื่อลดเวลาในการคำนวณลง ในงานวิจัยนี้ได้วัดผลโดยการสุ่มเลือกตรวจการจับคู่แบบแทนที่จำนวน 1000 คู่ประโยค จาก 2.8 ล้านคู่ประโยคพบว่ามีความผิดพลาดเพียง 0.9%

การจับคู่โดยใช้ข้อสนเทศผสมระหว่างความยาวประโยคและข้อสนเทศทางภาษา

Wu (1994) เสนอวิธีการสกัดประโยคคู่ขนานจากคลังประโยคคู่ขนานภาษาจีน – อังกฤษที่รวบรวมจากบันทึกการประชุมสภานิติบัญญัติของฮ่องกง แต่เนื่องจากความยาวประโยคภาษาอังกฤษต่อประโยคภาษาจีนอาจมีความแตกต่างกันมาก การใช้ข้อสนเทศจากความยาวประโยคเพียงอย่างเดียวอาจไม่เพียงพอ งานวิจัยนี้จึงได้เสนอวิธีการนำเอาข้อสนเทศจากคลังศัพท์ (Lexicon) มาพิจารณาร่วมด้วย งานวิจัยนี้ได้นำวิธีของ Gale and Church (1991) มาพัฒนาต่อด้วยการเพิ่มการคำนวณความแตกต่างของจำนวนคู่คำระบุนัย (Lexicon cue) ที่พบในแต่ละประโยคมาร่วมด้วย นั่นคือถ้าคู่ประโยคใดมีความแตกต่างของคู่คำระบุนัยต่ำก็จะมีโอกาสจับคู่กันมากขึ้น จากผลการทดลองขนาด 505 ประโยคภาอังกฤษ และ 506 ประโยคภาษาจีน พบว่าวิธีการที่เสนอนี้สามารถเพิ่มผลความถูกต้องจาก 86.4% เป็น 92.1% แต่อย่างไรก็ตามวิธีการนี้จำเป็นต้องมีการกำหนดคลังศัพท์ระบุนัยเบื้องต้นไว้ก่อนโดยคน และคลังศัพท์นี้อาจมีความจำเพาะกับคลังประโยคก็ได้

Melamard (1999) ได้นำวิธีการรู้จำแบบ (Pattern recognition) มาประยุกต์ใช้กับปัญหาการจับคู่การแปลระดับประโยค โดยได้เสนอวิธีที่แบ่งได้เป็น 3 ขั้นตอน คือ ขั้นตอนการสร้างสัญญาณ (Signal generation) ขั้นตอนการกรองสัญญาณรบกวน (Noise filtering) และขั้นตอนการค้นหา (Search) งานวิจัยนี้จะสร้างจุดที่เกี่ยวข้องกัน (Point of correspondence) ในปริภูมิการจับคู่การแปล (Bitext space) จากคำที่อาจเป็นคู่คำแปลกันโดยพิจารณาจากรากศัพท์การสะกดคำที่คล้ายกัน (Orthographic cognate) และจะสร้างผลลัพธ์ซึ่งเป็นรูปแบบการจับคู่การแปลโดยพิจารณาจากจุดเหล่านี้ จากผลการทดลองกับเอกสารบันทึกการประชุมรัฐสภาแคนาดา (Canadian handsard) พบว่าสามารถลดความผิดพลาดเมื่อเทียบกับวิธีของ Gale and Church (1991) จาก 3% เหลือ 1.8% แม้ว่าวิธีของ Melamard (1999) จะให้ผลความถูกต้องที่สูงและไม่จำเป็นต้องตัดประโยคก่อนเพราะสามารถพิจารณาจากตำแหน่งคำที่เกี่ยวข้องกันได้ทันที แต่วิธีการนี้ยังมีข้อจำกัดเนื่องจากการคำนวณความคล้ายของคำจะใช้วิธีการคำนวณค่า edit distance ซึ่งจะสามารถทำได้เฉพาะคู่ภาษาที่มีการใช้ตัวอักษรร่วมกัน เช่น คู่ภาษาอังกฤษ – ฝรั่งเศส และรูปแบบการแปลของเอกสารต้องเป็นการแปลที่ตรงไปตรงมาสูง

Chuang and Yeh (2005) พบว่าวิธีของ Gale and Church (1991) ได้ผลดีกับคู่ภาษาที่มีความคล้ายกัน เช่น อังกฤษ – ฝรั่งเศส – เยอรมัน แต่สำหรับคู่ภาษาที่มีความแตกต่างกันสูง เช่น จีน – อังกฤษ หรือคู่เอกสารที่แปลโดยอรรถ เช่น นิตยสาร จะได้ผลความถูกต้องที่ต่ำกว่า งานวิจัยนี้ได้สำรวจคลังประโยคภาษาจีน – อังกฤษ และพบว่าเครื่องหมายวรรคตอนส่วนใหญ่มักถูกแปลได้อย่างตรงไปตรงมา หรือมีคู่การแปลของแต่ละเครื่องหมายอย่างชัดเจน งานวิจัยนี้จึงเสนอการพิจารณาเครื่องหมายวรรคตอนร่วมกับความยาวประโยคในการสกัดประโยคคู่ขนานโดยปรับปรุงจากวิธีของ Gale and Church (1991) และทำการทดลองกับนิตยสารสองภาษาพบว่าได้ความถูกต้องมากกว่าการใช้ความยาวประโยคเพียงอย่างเดียว โดยได้ผลความถูกต้องมากกว่า 93%

Mamitimin and Hou (2009) พบว่าการจับคู่การแปลระดับประโยคสำหรับบางคู่ภาษา เช่น คู่ภาษาอุยกูร์-จีน การใช้ความยาวทั้งหน่วยของคำและหน่วยของอักขระจะให้ผลดีกว่า เพราะคำในภาษาอุยกูร์มักเกิดจากการรวมกันหลายๆอักขระ (Alphabetic language) แต่คำภาษาจีนมักจะเป็นหนึ่งอักขระต่อหนึ่งคำ (Non-alphabetic language) ดังนั้นความสัมพันธ์ระหว่างความยาวประโยคภาษาอุยกูร์ในหน่วยของคำ และความยาวประโยคภาษาจีนในหน่วยอักขระจึงสูงกว่าเปรียบเทียบกันในหน่วยอักขระหรือคำเพียงอย่างเดียว งานวิจัยนี้เสนอวิธีการคำนวณหาประโยคที่มีความสัมพันธ์กันสูงมาใช้เป็นจุดตรึง (Anchor point) ในขั้นตอนการจับคู่คำแปลระดับประโยค และใช้เทคนิคการจับคู่การแปลระดับประโยคจากอัตราส่วนความยาวประโยคมาจับคู่ประโยคระหว่างจุดตรึงเล่านี้อีกทีหนึ่ง การคำนวณจุดตรึงในงานวิจัยนี้จะพิจารณาจากข้อสนเทศหลายๆอย่าง เช่น ชื่อเฉพาะ ตัวเลข เครื่องหมายวรรคตอน ความยาวประโยค เป็นต้น การทดลองในงานวิจัยนี้ได้ทดลองเปรียบเทียบกับวิธีของ Gale และ Church (1991) และพบว่าการใช้จุดตรึงสามารถเพิ่มความแม่นยำขึ้นจาก 92.7% เป็น 94.6% และความครอบคลุมจาก 92.8% เป็น 94.8%

การจับคู่โดยใช้ข้อสนเทศผสมระหว่างความยาวประโยคและข้อสนเทศการแปล

Chen (1993) เสนอวิธีการจับคู่การแปลระดับประโยคโดยใช้ข้อสนเทศจากคลังศัพท์ร่วม ความยาวประโยคในหน่วยของจำนวนคำแต่ละภาษา วิธีการที่ใช้ในงานวิจัยนี้จะทำการจับคู่การแปลระดับคำไปพร้อมกับการจับคู่ระดับประโยคโดยอาศัยความน่าจะเป็นร่วม (Joint probability) ในการเกิดคู่คำนั้นๆ และคำนวณการจับคู่ระดับประโยคด้วยแบบจำลองทางสถิติแบบปัวซอง (Poisson distribution) จากคลังประโยคคู่ขนานตั้งต้นบางส่วน โดยจะกำหนดตัวแปรสำหรับการจับคู่แบบแทรกและแบบลบให้เป็นค่าคงที่ งานวิจัยนี้ได้ทดลองเปรียบเทียบกับวิธีของ Gale และ Church (1991) และวิธีของ Brown (1991) กับบันทึกการประชุมรัฐสภาแคนาดา พบว่าสามารถลดความผิดพลาดได้ 0.2% โดยไม่ตัดส่วนใดส่วนหนึ่งของเอกสารออกเลย

Thannin *et al.* (1998) เสนอวิธีการจับคู่การแปลระดับประโยคสำหรับคู่ภาษาไทย – อังกฤษ ไปพร้อมกับการตัดประโยคภาษาไทยโดยพิจารณาจากคำเนื้อหา (Content word) ช่องว่างในภาษาไทย และคู่คำแปลที่ป้อนเข้าไปโดยผู้ใช้ งานวิจัยนี้ได้ทดลองกับเอกสารคู่ภาษาไทย – อังกฤษประมาณ 2000 ประโยค และรายงานว่าได้ความถูกต้องมากกว่า 80% แต่อย่างไรก็ตามวิธีการที่เสนอยังมีข้อจำกัดเพราะการใช้พจนานุกรมจับคู่คำแปลคำต่อคำจะมีปัญหากับการจับคู่คำแปลของคำแปลที่เป็นคำประสมมากกว่าหนึ่งคำ

Al-Adhaileh *et al.* (2001) ได้ประยุกต์ใช้วิธีของ Melamard (1999) ในการจับคู่การแปลระดับประโยคคู่ภาษามาเลย์ – อังกฤษ โดยใช้คำแปลในพจนานุกรมสองภาษา Kamus Inggeris Melayu Dewan (KIMD) ซึ่งมีรายการคำแปลประมาณ 20,000 รายการที่เป็นการแปลคำแบบหนึ่งต่อหนึ่ง ร่วมกับการใช้รากศัพท์และเครื่องหมายวรรคตอนเพื่อสร้างจุดที่เกี่ยวข้องกันตามวิธีของ Melamard (1999) งานวิจัยนี้ทำการทดลองเปรียบเทียบผลการจับคู่การแปลระดับประโยคกับเอกสารคู่ภาษามาเลย์ – อังกฤษ จากผลการทดลองในงานวิจัยนี้พบว่าการจับคู่การแปลระดับประโยคกับเอกสารที่เป็นคู่มือหรือหนังสือเรียน ซึ่งมักเป็นการแปลตามตัวอักษร จะได้ความถูกต้องสูงกว่าหนังสือแปลทั่วไปซึ่งอาจมีการแปลโดยอรรถ

Moore (2002) เสนอวิธีการจับคู่การแปลระดับประโยคโดยใช้ข้อสนเทศจากความยาวประโยคคู่กับข้อสนเทศการแปล วิธีการที่ใช้ในงานวิจัยนี้เป็นการประมวลผล 2 รอบ (2 pass process) การประมวลผลรอบแรกจะเป็นการจับคู่ประโยคโดยประยุกต์จากวิธีการของ Brown (1991) จากนั้นจะใช้ผลลัพธ์ของการจับคู่ที่มีความน่าจะเป็นที่จะจับคู่กันสูงมาใช้เป็นตัวอย่างการแปลเพื่อคำนวณค่าความน่าจะเป็นการแปลคำ การประมวลผลรอบที่สองจะเป็นการจับคู่ประโยคโดยใช้

ความยาวประโยคร่วมกับความน่าจะเป็นการแปลที่คำนวณได้จากรอบแรกในการคำนวณความน่าจะเป็นในการแปลจากประโยคภาษาต้นทางไปเป็นประโยคภาษาปลายทาง งานวิจัยนี้ทำการทดลองการจับคู่การแปลระดับประโยคกับคู่มือคอมพิวเตอร์ภาษาอังกฤษ – สเปน จำนวน 27,118 คู่ประโยค และวัดผลกับเฉพาะการจับคู่ประโยคที่จับคู่กันแบบแทนที่ (1-1) จากผลการทดลองนี้พบว่าการใช้ความน่าจะเป็นการแปลร่วมกับความยาวประโยคสามารถลดความผิดพลาดจากการใช้ความยาวประโยคเพียงอย่างเดียว โดยคิดเป็นความผิดพลาดจากความไม่แม่นยำ (Precision error) จาก 0.104% เป็น 0.006% และความผิดพลาดจากความไม่ครอบคลุม (Recall error) จาก 0.409% เป็น 0.340%

Fung and Cheung (2004) เสนอวิธีการสกัดประโยคคู่ขนานจากคลังประโยคแบบเสมือนเปรียบเทียบกันได้ (Quasi-comparable corpus, ดูรายละเอียดที่หัวข้อระดับของเอกสารคู่ขนาน) วิธีที่ใช้ในงานวิจัยนี้จะทำการจับคู่สองระดับได้แก่ ระดับเอกสารและระดับประโยค ในงานวิจัยนี้เอกสารภาษาจีนจะถูกแปลไปเป็นภาษาอังกฤษในขั้นตอนการประมวลผลขั้นต้นโดยใช้พจนานุกรมสองภาษาที่มีอยู่แล้ว เพื่อที่จะสามารถเปรียบเทียบเอกสารจากสองภาษาโดยใช้ค่าส่วนกลับของความถี่ของเอกสาร (Inverted document frequency) และเปรียบเทียบความคล้ายระหว่างประโยคโดยใช้ค่าความคล้ายแบบโคไซน์ได้ นอกจากนี้ยังได้เสนอวิธีการสกัดประโยคคู่ขนานแบบบูทสแตรป (Bootstrapping) โดยการสร้างคู่คำแปลขึ้นมาใหม่จากคู่ประโยคที่สกัดได้และนำคู่คำแปลที่สกัดได้นี้ไปใช้ในการจับคู่เอกสารในรอบต่อๆไป งานวิจัยนี้ได้ทำการทดลองกับคลังประโยค TDT3 (Topic Detection and Tracking 3) ซึ่งเกิดจากการรวบรวมบทพูดทางวิทยุ และมีประโยคภาษาจีนจำนวน 110,000 ประโยค และประโยคภาษาอังกฤษจำนวน 290,000 ประโยค และวัดผลกับคู่ประโยคที่มีค่าความคล้ายมากที่สุดจำนวน 2,500 คู่ประโยค พบว่าการสกัดประโยคคู่ขนานแบบบูทสแตปมีความถูกต้อง 65.7% และการจับคู่โดยไม่ใช้วิธีบูทสแตปมีความถูกต้อง 42.8%

Németh *et al.* (2005) รายงานการรวบรวมและจับคู่การแปลระดับประโยคคู่สำหรับคลังประโยคภาษาฮังการี – อังกฤษ จากเอกสารหลายๆรูปแบบ ได้แก่ วรรณกรรมแปลจากโครงการกูเตนเบิร์ก (Gutenberg project) คัมภีร์ทางศาสนา กฎหมายนานาชาติ คำบรรยายภาพยนตร์ คู่มือการใช้ซอฟท์แวร์ นิตยสารแปล และรายงานผลประกอบการของบริษัท วิธีการจับคู่การแปลระดับประโยคที่ใช้ในงานวิจัยนี้เป็นการจับคู่แบบมีลำดับโดยใช้ข้อสนเทศจากความยาวประโยคและคู่คำแปลจากพจนานุกรมที่มีอยู่ก่อนแล้ว ในขั้นเริ่มต้นข้อความในภาษาต้นทางจะถูกแปลไปเป็นคำในภาษาปลายทางโดยใช้ความถี่การเกิดคู่กันในพจนานุกรม จากนั้นจะจับคู่ประโยคโดยคิดค่าคะแนนในการจับคู่มาจาก 2 ส่วนด้วยกัน ได้แก่ คะแนนการจับคู่จากอัตราส่วนความยาวประโยค คะแนนการจับคู่จากจำนวนคำที่เป็นคู่คำแปลกันระหว่างทั้งสองประโยค ค่าคะแนนทั้งสองนี้จะถูกคิดรวมกันแบบ

ถ่วงน้ำหนักที่ได้มาจากการปรับแต่งให้มีความแม่นยำสูงสุดบนชุดข้อมูลฝึกฝน งานวิจัยนี้ได้เสนอวิธีการแบบบูทสแตรปสำหรับกรณีที่ไม่มีพจนานุกรมสองภาษาด้วยใช้เพียงอัตราส่วนความยาวประโยคในการสกัดประโยคคู่ขนานในรอบแรก จากนั้นจึงสกัดคู่คำแปลจากคู่ประโยคที่สกัดได้โดยการพิจารณาจากความถี่การเกิดคู่กัน และใช้คู่คำแปลที่ได้นี้ในการสกัดประโยคคู่ขนานรอบถัดไป งานวิจัยนี้รายงานว่าการสกัดประโยคคู่ขนานโดยใช้คู่คำแปลมีความถูกต้องสูงขึ้นกว่าการใช้ความยาวประโยคเพียงอย่างเดียว โดยได้ความแม่นยำเพิ่มขึ้นจาก 97.58% เป็น 99.34% สำหรับกรณีที่มีพจนานุกรมอยู่แล้วและ 99.12% สำหรับกรณีที่ใช้คู่คำแปลจากการทำบูทสแตรป และได้ความครอบคลุมเพิ่มขึ้นจาก 97.55% เป็น 99.34% สำหรับกรณีที่มีพจนานุกรมอยู่แล้วและ 99.18% สำหรับกรณีที่ใช้คู่คำแปลจากการทำบูทสแตรป

เนื่องจากโปรแกรมแปลภาษามักจะทำงานได้ดีกับสิ่งที่โปรแกรมได้ถูกฝึกมา นั่นคือโปรแกรมจะแปลได้ดีกับโดเมนของคลังประโยคที่ใช้ฝึกฝน และคุณภาพการแปลจะต่ำลงถ้านำไปแปลข้อความที่ไม่อยู่ในโดเมน แต่เนื่องจากคลังประโยคคู่ขนานที่ตรงกับโดเมนที่ต้องการแปลมีปริมาณน้อยกว่าคลังเอกสารที่ไม่อยู่ในเมนอย่างมาก Munteanu and Marcu (2005) จึงได้เสนอวิธีการเพิ่มประสิทธิภาพของโปรแกรมแปลภาษาที่ฝึกจากคลังเอกสารที่ไม่อยู่ในโดเมนและมีปริมาณน้อย ด้วยการสกัดประโยคคู่ขนานจากคลังเอกสารที่ตรงโดเมนที่มีปริมาณเอกสารมากแต่ยังไม่ได้มีการจับคู่ระดับประโยค หรือจับคู่ประโยคไม่ได้โดยตรงเพราะเป็นเอกสารแบบเปรียบเทียบกันได้ งานวิจัยนี้ได้ทำการทดลองโดยการฝึกโปรแกรมแปลภาษาด้วยคลังประโยคคู่ขนานที่รวบรวมจากบันทึกการประชุมสหประชาชาติ (UN parliamentary proceeding) และสกัดประโยคคู่ขนานจากข่าวจากสำนักข่าวซินหัว (Xinhua news agency) ซึ่งมี 3 ภาษา ได้แก่ ภาษาจีน ภาษาอาหรับ และภาษาอังกฤษ วิธีการสกัดประโยคคู่ขนานในงานวิจัยนี้จะเป็นการจับคู่ประโยคแบบไม่มีลำดับ โดยใช้แบบจำลองคัดแยกแบบเอนโทรปีสูงสุด (Maximum entropy classifier) ที่ฝึกฝนจากคลังประโยคคู่ขนานตั้งต้นบางส่วน และอาศัยคู่คำแปลจากพจนานุกรมที่มีอยู่แล้วในการกรองคู่ประโยคที่มีแนวโน้มว่าจะไม่ใช่คู่การแปลจริงในขั้นต้น จากนั้นได้วัดประสิทธิภาพของโปรแกรมแปลภาษาจากการวัดค่าคะแนน BLEU ซึ่งพบว่าการใช้ประโยคคู่ขนานจากภายนอกสามารถเพิ่มคุณภาพการแปลได้ นอกจากนี้งานวิจัยนี้ยังได้เสนอวิธีการแบบบูทสแตปในกรณีที่ไม่มีพจนานุกรม จากผลการทดลองพบว่าสามารถเพิ่มค่าคะแนน BLEU ได้ 4.5 คะแนน ถ้าใช้คลังประโยคตั้งต้นขนาด 100,000 คำ และเพิ่มขึ้น 1 คะแนนสำหรับกรณีที่ใช้คู่ประโยคตั้งต้นจำนวน 95 ล้านคำ

Uchiyama and Isahara (2007) รายงานวิธีการสร้างคลังประโยคคู่ขนานภาษาญี่ปุ่น – อังกฤษ จากสิทธิบัตรที่รวบรวมจากโครงการ NII Test Collection for IR Systems (NTCIR) ซึ่งพบว่าข้อความในคู่สิทธิบัตรมักมีการแปลแบบครบทุกคำ งานวิจัยนี้ได้เสนอวิธีการจับคู่การแปลระดับ

ประโยคโดยใช้ข้อสนเทศ 3 ชนิด ได้แก่ คู่คำแปลจากพจนานุกรมของ Japan Electronic Dictionary Research Institute (EDR dictionary) ชนิดของคำ (Path-Of-Speech) และการวิเคราะห์หน่วยคำหลัก (Lemma) จาก Wordnet การคำนวณค่าคะแนนการจับคู่ระดับประโยคคำนวณจากจำนวนคำที่เป็นคู่การแปลกัน ค่าคะแนนเฉลี่ยของการจับคู่ประโยคอื่นในเอกสารเดียวกัน และค่าอัตราส่วนจำนวนประโยคที่จำคู่กัน งานวิจัยนี้ทำการทดลองการจับคู่ระดับประโยคกับคู่เอกสารสองภาษาจำนวน 149,603 คู่เอกสาร และเลือกเอาเฉพาะคู่ประโยคที่มีคะแนนการจับคู่สูงประมาณ 2 ล้านคู่ประโยค จากนั้นจึงสุ่มตรวจจำนวน 1,000 คู่ประโยค พบว่าคู่ประโยคที่มีเนื้อหาตรงกันทั้งหมดมี 899 คู่ประโยค คู่ประโยคที่มีเนื้อหาส่วนมาก (มากกว่าร้อยละ 80) ตรงกัน 72 คู่ประโยค คู่ประโยคที่มีเนื้อหาบางส่วน (น้อยกว่าร้อยละ 80) ตรงกัน 26 คู่ประโยค และจับคู่ผิดจำนวน 3 คู่ประโยค

Moe (2008) ได้พัฒนาเทคนิคจับคู่การแปลระดับประโยคสำหรับเอกสารคู่ภาษาไทย – อังกฤษ งานวิจัยนี้ไม่ได้กำกับขอบเขตประโยคภาษาอย่างแท้จริง แต่จะกำกับจุดที่อาจเป็นขอบเขตของประโยคภาษาไทยตามการเว้นวรรคที่มีอยู่แล้ว (Preexiting space) ซึ่งทำให้เกิดส่วนของข้อความภาษาไทยสั้นๆจำนวนมาก แต่เพียงพอที่จะทำการหาขอบเขตประโยคที่แท้จริงไปพร้อมกับการจับคู่การแปลระดับประโยคต่อไปได้ การขับคู่การแปลระดับประโยคในงานวิจัยนี้ใช้ข้อสนเทศ 3 ชนิด ได้แก่ ความยาวประโยค คู่คำแปลจากพจนานุกรมสองภาษา และสัมพันธ์เชิงความหมายจาก Wordnet งานวิจัยนี้เสนอการวัดค่าความแม่นยำ (Precision) และความครอบคลุม (Recall) จากจำนวนของคู่ลำดับประโยคที่จับคู่กันถูกต้อง และจำนวนส่วนของข้อความภาษาไทยส่วนแรกของการจับคู่ทำให้คำนวณระดับความถูกต้องของการจับคู่ย่อยได้ ซึ่งแตกต่างจากงานอื่นที่จะวัดความถูกต้องเป็นผิดและถูกเท่านั้น การทดลองในงานวิจัยนี้ทำการทดลองกับบทความและนวนิยายแปลและได้ผลความถูกต้องเป็นค่าความแม่นยำ 0.85 และค่าความครอบคลุม 0.91

Ma (2006) พัฒนาโปรแกรม Champollion ซึ่งเป็นโปรแกรมสำหรับการจับคู่การแปลระดับประโยคสำหรับเอกสารคู่ภาษาจีน – อังกฤษ โดยใช้ข้อสนเทศจากรูปแบบการจับคู่ย่อย ความยาวประโยค และคู่คำแปลจากพจนานุกรมที่มีอยู่แล้ว งานวิจัยนี้ได้เสนอการคำนวณความสำคัญของคำจากค่าน้ำหนักของคำแปลที่คำนวณจากค่าความถี่ของคำในเอกสารนั้น (Term frequency, tf) และค่าส่วนกลับความถี่ของคำที่พบในเอกสาร (Invert document frequency, idf) และใช้ค่าถ่วงน้ำหนักที่ปรับแต่งสำหรับการจับคู่ย่อยแต่ละชนิด งานวิจัยนี้ทำการทดลองกับนิตยสารสองภาษาบันทึกการประชุมสภาฮ่องกง และเอกสารทางการขององค์การสหประชาชาติ ผลการทดลองพบว่าถ้าใช้ขนาดของพจนานุกรมที่ใหญ่ขึ้นจะได้ผลความถูกต้องที่สูงขึ้น โดยผลความถูกต้องที่ดีที่สุดได้ค่าความแม่นยำ 0.970 และค่าความครอบคลุม 0.969

Li *et al.* (2010) ได้เสนอวิธีการปรับปรุงโปรแกรม Champollion เดิมซึ่งให้ความถูกต้องสูง แต่ใช้เวลาในการประมวลมาก โดยใช้จุดตรึงที่พิจารณาจากผลลัพธ์การจับคู่โดยใช้ความยาวประโยค คู่คำแปลจากพจนานุกรมสองภาษา และเสนอการคิดลักษณะเฉพาะของประโยคให้เป็นคำที่ไม่ปรากฏในประโยคข้างเคียง จากผลการทดลองพบว่าโปรแกรมที่ปรับปรุงแล้วมีความถูกต้องใกล้เคียงกับโปรแกรม Champollion เดิม และสามารถเพิ่มความเร็วจากเดิมได้ประมาณ 5 เท่า

Slayden *et al.* (2010) เสนอวิธีการเตรียมคลังประโยคภาษาไทยสำหรับการสร้างคลังประโยคคู่ขนานภาษาไทย – อังกฤษ ที่ใช้ในการฝึกฝนโปรแกรมแปลภาษา งานวิจัยนี้ได้รายงานปัญหาและวิธีแก้ไขเกี่ยวกับการประมวลผลภาษาไทยเบื้องต้น ได้แก่ ความกำกวมในการเข้ารหัสภาษาไทย (Mis-coding) และความกำกวมของลำดับอักขระภาษาไทย งานวิจัยนี้แปลงปัญหาการกำกับขอบเขตภาษาไทยไปเป็นปัญหาการคัดแยกสิ่งของ โดยการใช้ฟังก์ชั่นคุณสมบัติ (Feature function) จากรูปคำและจำนวนคำทั้งก่อนและหลังช่องว่างเพื่อฝึกแบบจำลองคัดแยกแบบเอนโทรปีสูงสุด (Maximum entropy classifier) ในการคัดแยกช่องว่างในข้อความภาษาไทยเป็นจุดแบ่งขอบเขตประโยคหรือไม่ หลังข้อความภาษาไทยที่ได้รับการกำกับขอบเขตประโยคแล้วจะถูกจับคู่การแปลระดับประโยคโดยใช้วิธีการของ Moore (2002) เพื่อสร้างเป็นคลังประโยคคู่ขนาน งานวิจัยนี้ได้ทดลองเปรียบเทียบความถูกต้องของการตัดประโยคกับวิธีอื่นที่เคยมีมาก่อนบนคลังข้อความ Orchid ซึ่งเป็นคลังข้อความภาษาไทยที่กำกับชนิดของคำ (Part-Of-Speech) และขอบเขตประโยค จากผลการทดลองพบว่าการตัดประโยคโดยวิธีที่เสนอในงานวิจัยนี้ได้ความถูกต้องมากการวิธีที่เคยมีมาก่อนเล็กน้อยแต่ไม่จำเป็นต้องใช้ข้อสนเทศจากชนิดของคำ งานวิจัยนี้วัดคุณภาพของคลังประโยคคู่ขนานผลลัพธ์จากค่าคะแนน BLEU ซึ่งพบว่าได้ 0.233 คะแนน สำหรับการแปลไทย – อังกฤษ และได้ 0.194 คะแนนสำหรับการแปลอังกฤษ – ไทย

ตารางที่ 9 เปรียบเทียบงานวิจัยเดิม

งานวิจัย	คู่ภาษา	ชนิดเอกสาร	รูปแบบการจับคู่	ปริมาณข้อมูลทดสอบ (คู่ประโยค)	ข้อสนเทศในการจับคู่ประโยค	ความถูกต้อง	ปริมาณการจับคู่แบบแทรกและลบ
Gale and Church (1991)	อังกฤษ – ฝรั่งเศส – เยอรมัน	บันทึกการเงิน	มีลำดับ	1,316	ความยาวประโยคในหน่วยจำนวนอักขระ	95.8%	1%
Brown *et al*. (1991)	อังกฤษ – ฝรั่งเศส	บันทึกการประชุมรัฐสภา	มีลำดับ/จุดตรึง	1,000	ความยาวประโยคในหน่วยจำนวนคำ	99%	0.7%
Chen (1993)	อังกฤษ – ฝรั่งเศส	บันทึกการประชุมรัฐสภา	มีลำดับ	500	ความยาวประโยค และความน่าจะเป็นการแปล	99.6%	ไม่ได้ระบุ
Wu (1994)	จีน – อังกฤษ	บันทึกการประชุมรัฐสภา	มีลำดับ	479	ความยาวประโยค และคำระบุนัยที่จำเพาะกับคลังข้อความนั้นๆ	92.1%	ไม่ได้ระบุ
Thannin *et al*. (1998)	ไทย – อังกฤษ	ไม่ได้ระบุ	มีลำดับ	ประมาณ 2,000	คู่คำแปลจากพจนานุกรมที่ป้อนโดยผู้ใช้	มากกว่า 80%	ไม่ได้ระบุ
Melamard (1999)	อังกฤษ – ฝรั่งเศส	บันทึกการประชุมรัฐสภา	มีลำดับ	9,816	คู่คำแปลที่มาจากรากศัพท์เดียวกัน	98.74%	ไม่ได้ระบุ
Al-Adhaileh *et al*. (2001)	มาเลย์ – อังกฤษ	ตำราวิชาการและนวนิยายแปล	มีลำดับ	41,419	คู่คำแปลจากพจนานุกรมที่มีอยู่แล้ว	97.88%	ไม่ได้ระบุ

ตารางที่ 9 (ต่อ)

งานวิจัย	คู่ภาษา	ชนิดเอกสาร	รูปแบบการจับคู่	ปริมาณข้อมูลทดสอบ (คู่ประโยค)	ข้อสนเทศในการจับคู่ประโยค	ความถูกต้อง	ปริมาณการจับคู่แบบแทรกและลบ
Moore (2002)	อังกฤษ – สเปน	คู่มือซอฟท์แวร์	มีลำดับ	27,155	ความยาวประโยค และความน่าจะเป็นการแปล	99.98%	ไม่ได้ระบุ
Fung and Cheung (2004)	จีน – อังกฤษ	บทพูดทางวิทยุ	ไม่มีลำดับ/ บูทสแตป	2,500	คู่คำแปลจากพจนานุกรมที่มีอยู่แล้ว	65.7%	ไม่ได้ระบุ
Chuang and Yeh (2005)	จีน – อังกฤษ	นิตยสารสองภาษา	มีลำดับ	ประมาณ 1,500	ความยาวประโยค และเครื่องหมายวรรคตอน	93%	0.56%
Munteanu and Marcu (2005)	จีน – อังกฤษ – อาหรับ	ข่าวสองภาษา	ไม่มีลำดับ	95 ล้านคำ	คู่คำแปลจากพจนานุกรมที่มีอยู่แล้ว	ไม่มีข้อมูล	ไม่ได้ระบุ
Németh et al.(2005)	ฮังการี – อังกฤษ	กฎหมาย บทความทางศาสนา และนวนิยายแปล	มีลำดับ/ จุดตรึง/ บูทสแตป	ประมาณ 6,700	ความยาวประโยค และคู่คำแปลจากพจนานุกรมที่มีอยู่แล้ว	99.34%	ไม่ได้ระบุ

ตารางที่ 9 (ต่อ)

งานวิจัย	คู่ภาษา	ชนิดเอกสาร	รูปแบบการจับคู่	ปริมาณข้อมูลทดสอบ (คู่ประโยค)	ข้อสนเทศในการจับคู่ประโยค	ความถูกต้อง	ปริมาณการจับคู่แบบแทรกและลบ
Uchiyama and Isahara (2007)	ญี่ปุ่น – อังกฤษ	สิทธิบัตร	มีลำดับ	1,000	คู่คำแปลจากพจนานุกรมที่มีอยู่แล้ว ชนิดของคำ (Part-Of-Speech) และหน่วยคำหลักจาก Wordnet	97%	ไม่ได้ระบุ
Moe (2008)	ไทย – อังกฤษ	บทความและนวนิยายแปล	มีลำดับ	1,000	ความยาวประโยค และคู่คำแปลจากพจนานุกรมที่มีอยู่แล้ว	85%	ไม่ได้ระบุ
Mamitimin and Hou (2009)	อุยกูร์ – จีน	กฎหมาย บทสนทนา บทความและนวนิยายแปล	มีลำดับ/จุดตรึง	1,300	ความยาวประโยค และคำระบุนัย	94.6%	ไม่ได้ระบุ
Ma (2006) และ Li et al. (2010)	จีน – อังกฤษ	ไม่ได้ระบุ	มีลำดับ/จุดตรึง/บูทสแตป	ประมาณ 3,700	คู่คำแปลจากพจนานุกรมที่มีอยู่แล้ว และค่า tf-idf	97%	6.4% และ 0.5%
Slayden et al. (2010)	ไทย – อังกฤษ	ไม่ได้ระบุ	มีลำดับ	ไม่มีข้อมูล	ความยาวประโยค และความน่าจะเป็นการแปล	ไม่มีข้อมูล	ไม่ได้ระบุ

บทวิเคราะห์งานวิจัยเดิม

ข้อสนเทศที่ใช้ในการจับคู่การแปลระดับประโยค

จากงานวิจัยที่ผ่านมา แนวทางการใช้ข้อสนเทศสำหรับการจับคู่การแปลระดับประโยคแบ่งได้ 3 แนวทางด้วยกัน ได้แก่ ข้อสนเทศจากส่วนต่างความยาวประโยค (Purely length based alignment) ข้อสนเทศผสมระหว่างความยาวประโยคและคลังศัพท์ (Hybrid length-lexicon based alignment) และข้อสนเทศผสมระหว่างความยาวประโยคและการแปล (Hybrid length-translation based alignment) การจับคู่การแปลระดับประโยคโดยใช้ข้อสนเทศจากความยาวประโยค เป็นแนวทางแรกๆที่ได้รับการศึกษาและพัฒนา การจับคู่การแปลระดับประโยคแบบนี้จะตั้งอยู่บนสมมติฐานที่ว่า ประโยคที่ยาวในเอกสารภาษาต้นทางมักจะถูกแปลไปเป็นประโยคที่ยาวในเอกสารภาษาปลายทาง และเช่นเดียวกันกับกรณีประโยคสั้น การใช้ข้อสนเทศจากความยาวประโยคมีทั้งใช้ความยาวประโยคในหน่วยของคำ (Brown *et al.*, 1991) และหน่วยของอักขระ (Gale and Church, 1993) การใช้ข้อสนเทศจากความยาวประโยคในหน่วยของอักขระจะได้รับความนิยมมากกว่าหน่วยของคำ เพราะไม่จำเป็นต้องวิเคราะห์หน่วยคำก่อน ทำให้พัฒนาโปรแกรมได้ง่ายกว่า แต่การจับคู่การแปลระดับประโยคสำหรับบางคู่ภาษา เช่น คู่ภาษาอุยกูร์-จีน การใช้ความยาวทั้งหน่วยของคำและหน่วยของอักขระจะให้ผลดีกว่า เพราะคำในภาษาอุยกูร์มักเกิดจากการรวมกันหลายๆอักขระ (Alphabetic language) แต่คำภาษาจีนมักจะเป็นหนึ่งอักขระต่อหนึ่งคำ (Non-alphabetic language) ดังนั้นความยาวประโยคภาษาอุยกูร์ในหน่วยของคำ และความยาวประโยคภาษาจีนในหน่วยอักขระจะมีความสัมพันธ์กันสูง (Mamitimin and Hou, 2009) ดังนั้นการใช้ความยาวในหน่วยคำหรืออักขระจำเป็นต้องคำนึงถึงคู่ภาษานั้นๆด้วย

แม้ว่าข้อสนเทศจากความยาวประโยคจะเป็นข้อสนเทศที่มีประโยชน์ และสามารถประยุกต์ใช้ได้กับทุกคู่ภาษา แต่ว่าการใช้ข้อสนเทศจากความยาวประโยคเพียงอย่างเดียวก็ยังอาจเกิดความกำกวมได้ เพราะไม่ได้พิจารณาคำหรือข้อความภายในประโยค ต่อมาจึงได้มีการเสนอการใช้ข้อสนเทศจากคลังศัพท์ (Lexicon information) เข้ามาร่วมพิจารณาในการจับคู่การแปลระดับประโยค ข้อสนเทศจากคลังศัพท์มักจะเป็นลักษณะเฉพาะร่วมระหว่างคู่ภาษา เช่น เครื่องหมายวรรคตอนเมื่อถูกแปลไปเป็นภาษาปลายทางแล้วมักจะมีลักษณะและตำแหน่งคล้ายกับภาษาต้นทาง เป็นต้น จากงานวิจัยที่ผ่านมา ข้อสนเทศจากคลังศัพท์ที่นำมาร่วมพิจารณามาได้จากหลายทาง เช่น คำระบุนัย ตัวเลข ชื่อเฉพาะ หรือเครื่องหมายวรรคตอน เป็นต้น (Wu, 1994; Chuang and Yeh, 2005; Mamitimin and Hou, 2009) แต่การจับคู่การแปลระดับประโยคโดยใช้ข้อสนเทศจากคลังศัพท์

มักจะใช้ได้กับเฉพาะบางคู่ภาษาเท่านั้น เพราะการใช้เครื่องหมายวรรคตอนใช้ได้เฉพาะกับคู่ภาษาที่มีการใช้เครื่องหมายวรรคตอนคล้ายกันเท่านั้น หรือการใช้คำระบุนัยที่เป็นชื่อเฉพาะหรือตัวเลขเป็นการใช้ข้อมูลภายนอกที่จำเพาะกับกลุ่มการใช้นั้นๆ และจำเป็นต้องป้อนเข้าไปโดยคน ทำให้การจับคู่การแปลระดับประโยคโดยใช้ข้อสนเทศจากคลังศัพท์มีข้อจำกัดในการประยุกต์ใช้กับข้อมูลที่ไม่เคยเห็นหรือจากต่างกลุ่มการใช้กัน

อีกแนวทางหนึ่งของการสกัดคู่ประโยค คือการสกัดคู่ประโยคโดยใช้ข้อสนเทศจากการแปล โดยข้อสนเทศการแปลจะเป็นลักษณะคำแปล (Translation word) จากพจนานุกรมสองภาษาที่มีอยู่แล้ว (Thannin *et al.*, 1998; Ma, 2006; Utiyama and Isahara, 2007; Moe, 2008) คำแปลที่สกัดได้จากคลังประโยคคู่ขนานตั้งต้น (Németh *et al.*, 2005) หรือข้อสนเทศการแปลจากแบบจำลองการแปล (Translation model) ที่ประมาณขึ้นจากคลังประโยคคู่ขนานที่สกัดได้บางส่วน (Moore, 2002; Munteanu and Marcu, 2005) การใช้ข้อสนเทศจากการแปลสามารถใช้เป็นจุดเชื่อมโยงระหว่างภาษาได้โดยไม่จำกัดกลุ่มการใช้ แต่จำเป็นต้องคำนึงถึงความครอบคลุมของพจนานุกรมที่ใช้หรือคลังประโยคคู่ขนานที่จะใช้ในการหาแบบจำลองการแปลด้วย

เทคนิคที่ใช้ในการจับคู่การแปลระดับประโยค

การจับคู่การแปลระดับประโยคมี 2 แบบ ได้แก่ การจับคู่แบบไม่มีลำดับ และแบบมีลำดับ การจับคู่การแปลระดับประโยคแบบไม่มีลำดับมักจะถูกแปลงไปเป็นปัญหาการคัดแยกสิ่งของซึ่งในงานวิจัยการจับคู่การแปลระดับประโยคคือการคัดแยกว่าคู่ประโยคใดเป็นคู่การแปลกัน และสามารถจับคู่ประโยคที่สลับลำดับการแปลกันได้ การจับคู่แบบนี้มักทำกับคู่ภาษาที่สามารถตัดประโยคได้อย่างชัดเจน เนื่องจากจะสกัดได้เฉพาะการจับคู่แบบแทนที่เท่านั้น และคุณภาพการแปลจะต่ำกว่าคู่ประโยคที่สกัดจากคลังข้อความสองภาษาแบบขนาน การจับคู่แบบไม่มีลำดับจึงเหมาะสมกับคลังข้อความสองภาษาแบบเทียบได้ซึ่งสามารถชดเชยคุณภาพของคู่ประโยคที่สกัดได้ด้วยปริมาณที่มีมากกว่าคลังข้อความสองภาษาแบบขนาน (Munteanu and Marcu, 2005; Wu and Fung, 2005)

การจับคู่การแปลระดับประโยคแบบมีลำดับมักจะถูกแปลงเป็นปัญหาการหาสายลำดับที่ดีที่สุด (Sequence alignment) ซึ่งแก้โดยการใช้การโปรแกรมเชิงพลวัตร (Dynamic programming) การจับคู่แบบนี้จะตั้งอยู่บนสมมติฐานว่าคู่เอกสารมีการรักษาลำดับการแปลจึงเหมาะสมกับคลังข้อความสองภาษาแบบขนาน แม้ว่าการจับคู่แบบมีลำดับจะสามารถจับคู่ได้ทุกรูปแบบการจับคู่ย่อย

แต่ถ้ามีการย้ายที่ลำดับประโยคที่แปลจะทำให้ต้องกำกับการจับคู่นั้นให้เป็นการจับคู่แบบรวม หรือในบางกรณีอาจต้องกำกับให้เป็นการแปลแบบแทรกหรือแบบลบ

การโปรแกรมเชิงพลวัตรแก้นั้นมีปริภูมิค้นหา (Search space) ขนาดใหญ่ เพราะปริภูมิค้นหาจะโตในลักษณะ $O(mn)$ เมื่อ m เป็นจำนวนประโยคในเอกสารภาษาต้นทาง และ n เป็นจำนวนประโยคในเอกสารภาษาปลายทาง จึงมีแนวคิดที่จะลดขนาดของปริภูมิค้นหาลงโดยการใช้จุดตรึง (Anchor) โดยจุดตรึงที่ใช้นี้อาจจะเป็นย่อหน้า (Brown *et al.*, 1991) หรือประโยค (Mamitimin and Hou, 2009; Li *et al.*, 2010) ที่สามารถจับคู่กันแบบ 1-1 ระหว่างเอกสารภาษาต้นทางและเอกสารภาษาปลายทางได้ชัดเจน จากนั้นจึงจับคู่การแปลระดับประโยคระหว่างจุดตรึงเหล่านี้อีกครั้ง

งานวิจัยที่เกี่ยวข้องกับการจับคู่ประโยคคู่ขนานภาษาไทย – อังกฤษ

เนื่องจากโปรแกรมจับคู่การแปลระดับประโยคจะเริ่มจากการรับข้อมูลป้อนเข้าที่เป็นคู่เอกสารที่มีการกำกับขอบเขตของประโยคมาแล้ว ซึ่งสามารถทำได้ง่ายในภาษาอื่นที่มีจุดบ่งบอกขอบเขตประโยคอย่างชัดเจน เช่น อังกฤษ ฝรั่งเศส จีน ญี่ปุ่น เป็นต้น แต่ภาษาไทยไม่มีจุดที่จะใช้บ่งบอกขอบเขตประโยคได้อย่างชัดเจน ประกอบทั้งการกำกับขอบเขตประโยคด้วยมือบนเอกสารเดียวกัน แต่ให้คนไทยต่างคนกำกับ ก็อาจจะได้ผลลัพธ์ไม่เหมือนกัน (Aroonmanakun, 2007) ทำให้การสกัดคู่ประโยคจากเอกสารคู่ภาษาไทย-อังกฤษต้องคำนึงถึงการกำกับขอบเขตประโยคภาษาไทยไปพร้อมกันด้วย

วิธีการที่เคยเสนอในงานวิจัยที่เคยมีมาแล้ว สามารถแยกแนวทางการสกัดคู่ประโยคสำหรับคู่เอกสารภาษาไทย-อังกฤษได้ 2 แนวทางด้วยกัน แนวทางแรกคือ การแยกพิจารณาการกำกับขอบเขตประโยคและการจับคู่ประโยคอออกจากกัน แนวทางนี้จะกำกับขอบเขตประโยคภาษาไทยในการประมวลผลขั้นต้น (Preprocessing) ด้วยการใช้แบบจำลองคัดแยก (Classifier) จากนั้นจึงสกัดประโยคคู่ขนานโดยใช้เครื่องมือที่เคยมีมาแล้ว (Slayden *et al.*, 2010) อีกแนวทางหนึ่ง คือ พิจารณาการกำกับขอบเขตประโยคกับการจับคู่การแปลไปพร้อมๆกัน โดยกำกับขอบเขตของคำหรือช่องว่างไว้ก่อนจากนั้นจึงรวมคำหรือข้อความระหว่างช่องว่างเหล่านั้นให้เป็นประโยคใหญ่โดยเทียบกับประโยคภาษาอังกฤษ (Tannin *et al.*, 1998; Moe, 2008)

บทสรุปงานวิจัยเดิม และแนวทางการพัฒนา

จากผลการทดลองเบื้องต้นพบว่าค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson's correlation coefficient) ระหว่างความยาวของคู่ประโยคภาษาไทย – อังกฤษ มีค่า 0.81 ซึ่งค่อนข้างสูง แต่ยังน้อยกว่างานวิจัยที่เคยมีมาพอสมควร ตัวอย่างเช่น 0.991 สำหรับคู่ประโยคภาษาอังกฤษ – ฝรั่งเศสในงานวิจัยของ Gale and Church (1991) หรือ 0.977 สำหรับคู่ประโยคภาษาจีน – อุยกูร์ในงานวิจัยของ Mamitimin and Hou (2009) ดังนั้นจึงจำเป็นต้องใช้ข้อสนเทศอย่างอื่นเพิ่มเติม จากการศึกษางานวิจัยที่มีมาพบว่าข้อสนเทศอื่นนอกเหนือจากความยาวประโยคจะมาจากข้อสนเทศการแปล (Translation information) (Thannin *et al.*, 1998; Moore, 2002; Munteanu and Marcu, 2005; Németh *et al.*, 2005; Ma, 2006; Utiyama and Isahara, 2007; Moe, 2008) และข้อสนเทศจากคำศัพท์ (Lexicon information) เช่น คำระบุนัย รากศัพท์ ตัวเลข หรือเครื่องหมายวรรคตอน เป็นต้น (Wu, 1994; Chuang and Yeh, 2005; Mamitimin and Hou, 2009) แต่อย่างไรก็ตามข้อสนเทศจากคำศัพท์มีแนวโน้มที่จะจำเพาะกับคลังข้อความหรือคู่ภาษา ทำให้ไม่สามารถนำข้อสนเทศจากคำศัพท์มาประยุกต์ใช้กับการจับคู่การแปลระดับประโยคภาษาไทย – อังกฤษได้โดยตรง ข้อสนเทศการแปลสามารถใช้เป็นจุดเชื่อมโยงระหว่างภาษาได้โดยไม่จำกัดกลุ่มการใช้ แต่จำเป็นต้องมีการสกัดคู่คำแปลเอาไว้ก่อน จากการศึกษางานที่มีมาข้อสนเทศการแปลจะถูกสกัดมาจากคู่คำแปลในพจนานุกรมที่มีอยู่แล้ว หรือการสกัดคู่คำแปลจากคลังประโยคคู่ขนานตั้งต้น จากการสำรวจพจนานุกรม Lexitron ซึ่งเป็นพจนานุกรมภาษาไทย – อังกฤษมีบางรายการเป็นการแปลจากคำไปเป็นคำอธิบายทำให้ไม่สามารถใช้ได้โดยตรง และการใช้คู่คำแปลจากพจนานุกรมที่มีอยู่แล้วไม่สามารถคำนวณน้ำหนักของค่าคำแปลได้ ตัวอย่างเช่น คำว่า “book” เมื่อปรากฏอยู่ในเอกสารการท่องเที่ยวมักจะมีแนวโน้มแปลว่า “จอง” มากกว่า “หนังสือ” เป็นต้น

การจับคู่การแปลระดับประโยคจำเป็นต้องรับข้อมูลป้อนเข้าที่กำกับขอบเขตประโยคมาแล้ว ซึ่งสามารถทำได้ง่ายกับภาษาที่มีเครื่องหมายจบประโยคอยู่แล้ว เช่น อังกฤษ จีน ญี่ปุ่น เป็นต้น แต่ภาษาไทยไม่มีเครื่องหมายจบประโยคที่ชัดเจนและมีการใช้เครื่องหมายวรรคตอนอื่นน้อยมากเมื่อเทียบกับข้อความภาษาอังกฤษ จึงส่งผลให้การประมวลผลเพื่อกำกับขอบเขตประโยคภาษาไทยทำได้ยากกว่าภาษาอื่น จากการศึกษางานที่มีมาแล้วพบว่ามีงานวิจัยหลายงานพัฒนาเทคนิคสำหรับการกำกับขอบเขตประโยคภาษาไทยโดยอัตโนมัติและได้นำมาใช้กับงานจับคู่การแปลระดับประโยค (Slayden, 2010) แต่ทั้งนี้เทคนิคการกำกับขอบเขตประโยคที่เสนอในงานวิจัยเดิมจำเป็นต้องมีคลังประโยคสำหรับฝึกฝนโปรแกรมกำกับขอบเขตประโยคภาษาไทยต่างหากจากคลังประโยคคู่ขนานสำหรับสกัดข้อสนเทศในการจับคู่การแปลระดับประโยค นอกจากนี้การแปลแบบเอาความหรือแปล

โดยอรรถอาจทำให้ประโยคภาษาไทยไม่ได้ถูกแปลทุกคำ ซึ่งอาจทำให้ประโยคคู่ขนานที่สกัดได้ไม่สามารถจับคู่ส่วนย่อยภายในได้ทั้งหมดและจะกลายเป็นสัญญาณรบกวน (Noise) ในการจับคู่ระดับคำหรือวลีต่อไปได้ งานวิจัยของ Tannin (1998) และ Moe (2008) ได้ใช้แนวทางการกำกับขอบเขตข้อความภาษาไทยก่อนการทำการจับคู่ระดับประโยคที่แตกต่างออกไป โดยใช้วิธีการกำกับจุดที่อาจเป็นจุดแบ่งขอบเขตของประโยคโดยการใช้ขอบเขตของคำหรือช่องว่างที่มีอยู่แล้วแทน แนวทางนี้ทำให้เกิดส่วนของข้อความภาษาไทยสั้นๆจำนวนมาก ซึ่งอาจทำให้การจับคู่การแปลระดับประโยคทำได้ยากขึ้นเพราะมีตัวเลือกในการจับคู่มากขึ้น แต่แนวทางนี้ไม่จำเป็นต้องใช้คลังประโยคภายนอกสำหรับฝึกฝนโปรแกรมกำกับขอบเขตประโยคภาษาไทย และลดปัญหาข้อความภาษาไทยที่แปลไม่ครบทุกคำได้บางส่วน

การจับคู่การแปลแบบลบหรือแบบแทรกในการจับคู่การแปลระดับประโยคจะแตกต่างจากระดับคำหรือวลี เนื่องจากส่วนของข้อความที่ไม่ได้แปลในการจับคู่ระดับประโยคมักจะมาจากการตัดสินใจของผู้แปลเองไม่ได้เกิดจากบริบท (Context) ของข้อความ นั่นคือเป็นไปได้ว่าข้อความอย่างเดียวกันในบริบทเดียวกันอาจมีทั้งที่ถูกแปลและไม่ถูกแปล จากงานวิจัยที่ผ่านมาการจับคู่แบบลบหรือแบบแทรกจะเป็นการคำนวณค่าคะแนนหรือค่าใช้จ่ายในการจับคู่กับประโยคว่าง (Null sentence) ซึ่งอาจทำให้โปรแกรมจับคู่การแปลระดับประโยคพยายามที่จะไม่จับคู่แบบลบหรือแบบแทรก (Undergenerate) เนื่องจากมีค่าใช้จ่ายในการจับคู่ที่สูงมาก หรือในบางงานวิจัยจะใช้วิธีตั้งเป็นค่าตัวแปรที่ปรับจูนเอาไว้สำหรับคลังข้อความนั้นๆ (Ma, 2006; Li *et al.*, 2010) แต่แนวทางนี้ก็อาจทำให้ค่าตัวแปรจะมีความจำเพาะกับคลังข้อความหรือคู่ภาษาได้

งานวิจัยนี้มุ่งเน้นพัฒนาเทคนิคสำหรับการจับคู่การแปลระดับประโยคจากเอกสารแบบขนานภาษาไทย – อังกฤษที่มีการแปลโดยอรรถ การจับคู่ประโยคจะคำนวณค่าใช้จ่ายจากความยาวประโยคร่วมกับการแปลที่สกัดจากคลังประโยคคู่ขนานตั้งต้นบางส่วน และปรับค่าถ่วงน้ำหนักของแต่ละข้อสนเทศโดยอัตโนมัติ ทั้งนี้ยังเพิ่มการปรับแต่งฟังก์ชั่นค่าใช้จ่ายสำหรับการจับคู่แบบลบและแบบแทรกเพื่อทำให้โปรแกรมทำการจับคู่แบบลบและแบบแทรกมากขึ้น เทคนิคที่ใช้ในงานวิจัยนี้แตกต่างจากวิธีของ Moore (2002) หรือ Slayden (2010) คือการคำนวณค่าใช้จ่ายจากความน่าจะเป็นการแปลจะเป็นการคำนวณทั้งสองทิศทาง และการปรับแต่งฟังก์ชั่นค่าใช้จ่ายสำหรับการจับคู่แบบลบและแบบแทรกซึ่งพบว่าการจับคู่ทั้งสองแบบนี้มีในเอกสารที่แปลโดยอรรถมากกว่าเอกสารที่แปลโดยพยัญชนะ

อุปกรณ์และวิธีการ

อุปกรณ์

1. ฮาร์ดแวร์
    1.1 เครื่องคอมพิวเตอร์

2. ซอฟท์แวร์
    2.1 ตัวแปลภาษา Python รุ่น 2.6
    2.2 คอมไพล์เลอร์ภาษา C++
    2.3 ชุดคำสั่ง NLTK
    2.4 ชุดคำสั่ง KUCut
    2.5 ชุดคำสั่ง GIZA++

3. ชุดข้อมูลที่ใช้ในการทดลอง
    3.1 บทความแปลเกี่ยวกับการท่องเที่ยวจังหวัดตราดจำนวน 941 คู่ประโยค
    3.2 รัฐธรรมนูญแห่งราชอาณาจักรไทยฉบับปี พ.ศ. 2550 จำนวน 1632 คู่ประโยค

วิธีการ

การเตรียมข้อมูลทดลอง

คลังข้อความสองภาษาที่ใช้ในการทดลองเป็นเอกสารคู่ภาษาไทย-อังกฤษเกี่ยวกับการท่องเที่ยวจังหวัดตราดที่แปลโดยอรรถจากภาษาไทยไปเป็นภาษาอังกฤษจำนวน 115 ย่อหน้า และรัฐธรรมนูญแห่งราชอาณาจักรไทยฉบับปี พ.ศ. 2550 ซึ่งเป็นเอกสารที่แปลโดยพยัญชนะ เพื่อทดสอบเปรียบเทียบความถูกต้องของการจับคู่การแปลระดับประโยคบนเอกสารที่มีลักษณะการแปลที่ต่างกัน

งานวิจัยนี้กำกับขอบเขตประโยคภาษาอังกฤษตามเครื่องหมายมหัพภาค (.) และในกรณีของบทความท่องเที่ยวจะกำกับขอบเขตข้อความภาษาไทยตามขอบเขตของอนุพากย์ (Unit) ซึ่งอาจเป็นวลี อนุประโยค หรือในบางกรณีอาจเป็นประโยคสั้นๆ (เมธี, 2006) และข้อความภาษาไทยในรัฐธรรมนูญสามารถกำกับขอบเขตของประโยคได้โดยใช้ขอบเขตของย่อหน้าอยู่แล้ว หลังจากกำกับ

ขอบเขตของประโยคหรืออนุพากย์แล้วจะได้จำนวนประโยคหรืออนุพากย์ตามตารางที่ 10 และจำนวนของการจับคู่การแปลย่อยแต่ละแบบตามตารางที่ 11 ซึ่งจากตารางทั้งสองนี้จะพบว่าบทความการท่องเที่ยวมีปริมาณการจับคู่แบบแทรกและแบบลบมากกว่ารัฐธรรมนูญ และมากกว่าปริมาณที่พบในงานที่มีมาก่อน และถึงแม้ว่าวิธีที่ใช้ในงานที่มีมาก่อนจะจับคู่การแปลได้ความถูกต้องในภาพรวมสูง แต่มักจะได้ความถูกต้องสำหรับการจับคู่การแปลแบบแทรกและแบบลบต่ำ แต่เนื่องจากการจับคู่แบบแทรกและแบบลบในงานที่มาก่อนมักมีปริมาณที่น้อยมากเมื่อเทียบกับขนาดคลังข้อความสองภาษาทั้งหมด จึงมีผลกระทบกับความถูกต้องโดยรวมเพียงเล็กน้อย (Gale and Church, 1991; Brown, 1991; Wu, 1994; Chuang and Yeh, 2005; Mamitimin and Hou, 2009; Li *et al.*, 2010)

งานวิจัยนี้กำกับขอบเขตของคำภาษาอังกฤษโดยใช้ชุดคำสั่ง NLTK (Loper and Bird, 2002) และขอบเขตภาษาไทยโดยชุดคำสั่ง KUCut (Sudprasert and Kawtrakul, 2003)

ตารางที่ 10 จำนวนประโยคหรืออนุพากย์ของแต่ละเอกสาร

จำนวนประโยค/อนุพากย์	**รัฐธรรมนูญ**	**บทความท่องเที่ยว**
ภาษาไทย	1,629	1,250
ภาษาอังกฤษ	1,631	756

ตารางที่ 11 จำนวนรูปแบบการจับคู่ย่อย

รูปแบบการจับคู่ย่อย	**รัฐธรรมนูญ**		**บทความท่องเที่ยว**	
(ไทย-อังกฤษ)	**จำนวน**	**ร้อยละ**	**จำนวน**	**ร้อยละ**
1-0	1	0.06	198	21.04
0-1	3	0.18	32	3.40
1-1	1628	99.75	480	51.01
2-1	0	0	154	16.37
3-1	0	0	49	5.21
4-1	0	0	14	1.49
อื่นๆ	0	0	14	1.49
รวม	1632	100	941	100

ตัวแปรเบื้องต้น

กำหนดให้ T และ E เป็นย่อหน้าภาษาไทย และภาษาอังกฤษตามลำดับ ซึ่งย่อหน้าในงานวิจัยนี้จะแทนด้วยลำดับของประโยคหรืออนุพากย์ในกรณีที่เป็นภาษาไทย และประโยคหรืออนุพากย์จะแทนด้วยลำดับของคำ ดังแสดงในสมการที่ 4 และ 5

$$T = t_1 t_2 t_3 \ldots t_{|T|} \tag{4}$$

$$E = e_1 e_2 e_3 \ldots e_{|E|} \tag{5}$$

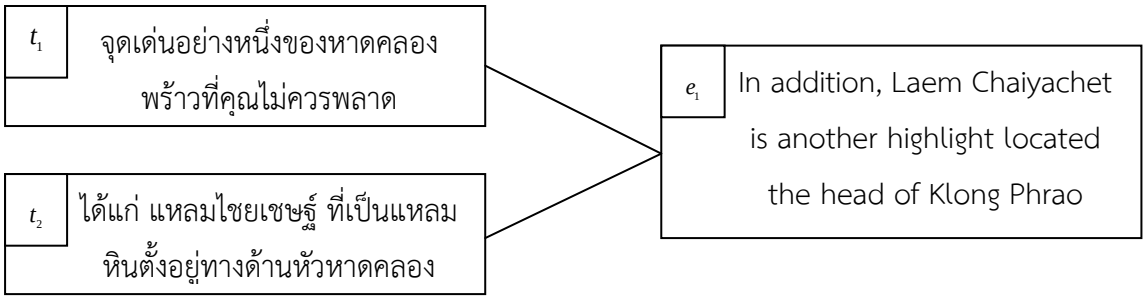
โดย T หมายถึงย่อหน้าภาษาไทย

$t_1 t_2 t_3 \ldots t_{|T|}$ หมายถึงอนุพากย์ภาษาไทยที่ต่อกันเป็นย่อหน้า

E หมายถึงย่อหน้าภาษาอังกฤษ

$e_1 e_2 e_3 \ldots e_{|E|}$ หมายถึงอนุพากย์ภาษาอังกฤษที่ต่อกันเป็นย่อหน้า

กำหนดให้ S เป็นคลังประโยคคู่ขนานตั้งต้น ซึ่งแทนด้วยเซตของ 3 สิ่งลำดับ $((n_t, n_e), t, e)$ ประกอบด้วยการจับคู่การแปลย่อย (n_t, n_e) แทนด้วยคู่ลำดับของจำนวนอนุพากย์ภาษาไทย (n_t) และประโยคภาษาอังกฤษ (n_e) ข้อความที่จับคู่กันแทนด้วยตัวแปร t สำหรับข้อความภาษาไทยที่เกิดจากการนำอนุพากย์ภาษาไทยมาต่อกัน (Concatenate) จำนวน n_t อนุพากย์และ e สำหรับข้อความภาษาอังกฤษที่เกิดจากการนำประโยคภาษาอังกฤษมาต่อกันจำนวน n_e ประโยค ภาพที่ 8 แสดงตัวอย่างคู่ประโยคในคลังประโยคคู่ขนานตั้งต้นซึ่งแทนได้ด้วย 3 สิ่งลำดับ $((2, 1), t_1 t_2, e_1)$ หรือ ((2, 1), “จุดเด่นอย่างหนึ่งของหาดคลองพร้าวที่คุณไม่ควรพลาด ได้แก่ แหลมไชยเชษฐ์ ที่เป็นแหลมหินตั้งอยู่ทางด้านหัวหาดคลอง”, “In addition, Laem Chaiyachet is another highlight located the head of Klong Phrao”)



ภาพที่ 8 ตัวอย่างคู่ประโยคในคลังประโยคคู่ขนานตั้งต้น

กำหนดให้ L_t และ L_e เป็นตัวแปรสุ่มของความยาวอนุพากย์ภาษาไทย และความยาวประโยคภาษาอังกฤษในหน่วยตัวอักขระที่พบในคลังประโยคคู่ขนานตั้งต้นตามลำดับ c และ s เป็นค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานอัตราส่วนความยาวประโยคภาษาอังกฤษต่อความยาวอนุพากย์ภาษาไทยตามลำดับ ดังสมการที่ 6 และ 7

$$c = E\left[L_e / L_t\right] \tag{6}$$

$$s = \sqrt{Var\left[L_e / L_t\right]} \tag{7}$$

โดย L_e หมายถึงตัวแปรสุ่มของความยาวอนุพากย์ภาษาไทย

L_t หมายถึงตัวแปรสุ่มของความยาวประโยคภาษาอังกฤษ

c หมายถึงค่าเฉลี่ยอัตราส่วนความยาวประโยคภาษาอังกฤษต่อภาษาไทยที่จับคู่กัน

s หมายถึงค่าส่วนเบี่ยงเบนมาตรฐานอัตราส่วนความยาวประโยคภาษาอังกฤษต่อภาษาไทยที่จับคู่กัน

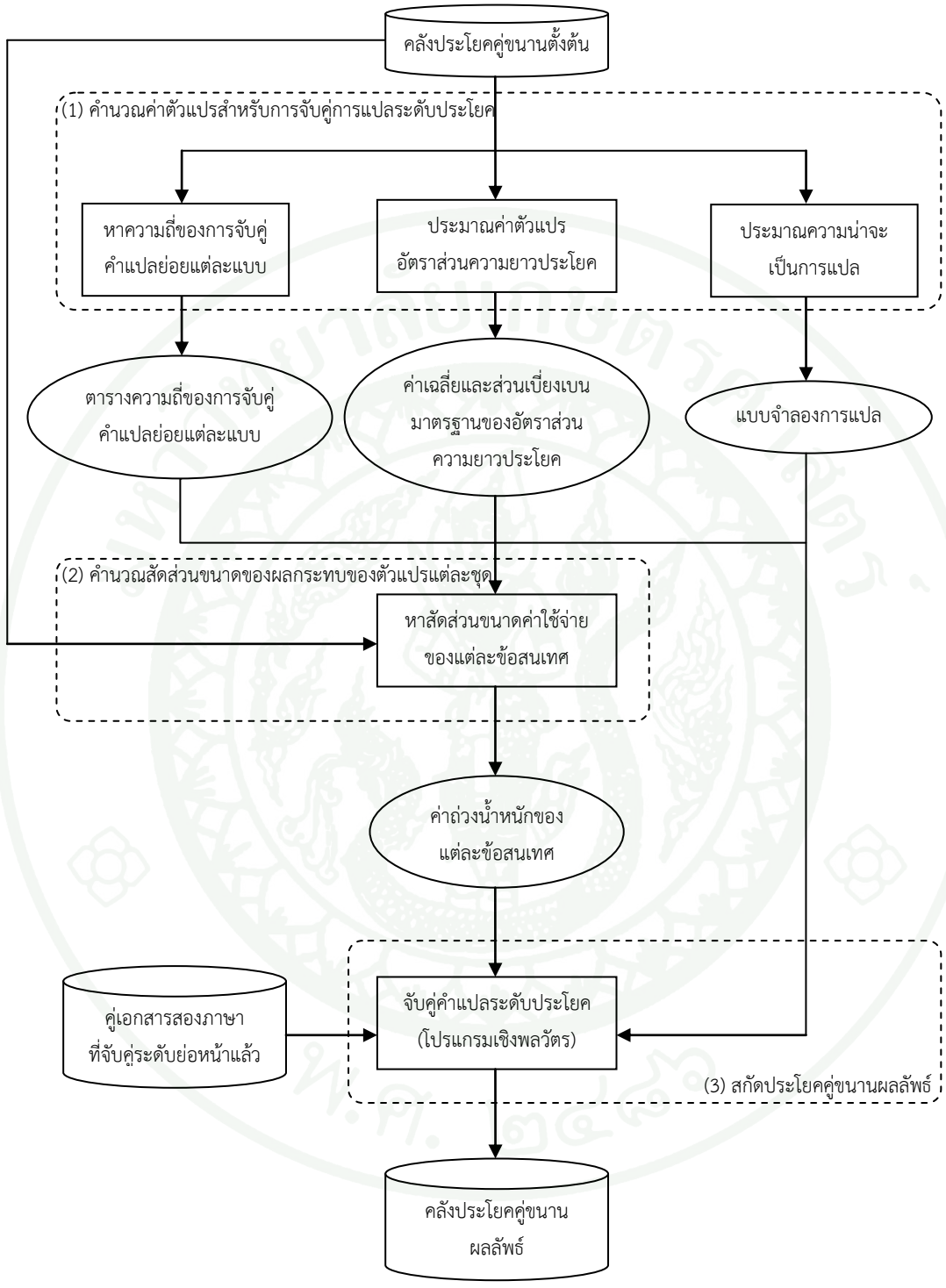
ภาพรวมการจับคู่การแปลระดับประโยค

ภาพรวมการทำงานของโปรแกรมแสดงในภาพที่ 9 โดยการทำงานของโปรแกรมแบ่งเป็น 3 ขั้นตอน ดังนี้

1. ประมาณค่าตัวแปรที่ใช้ในการจับคู่การแปลระดับประโยค ได้แก่
    1.1 ความถี่ของการจับคู่การแปลย่อยแต่ละแบบ
    1.2 ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของอัตราส่วนความยาวประโยค
    1.3 แบบจำลองการแปล

2. คำนวณสัดส่วนขนาดของผลกระทบของตัวแปรแต่ละชุดบนคลังประโยคคู่ขนานตั้งต้นเพื่อกำหนดค่าถ่วงน้ำหนักของตัวแปรแต่ละชุด

3. คำนวณหารูปแบบการจับคู่การแปลที่มีค่าใช้จ่ายน้อยที่สุดบนชุดข้อมูลทดสอบ โดยใช้การโปรแกรมเชิงพลวัตร



ภาพที่ 9 ภาพรวมการทำงานการจับคู่การแปลระดับประโยค

ตัวแปรที่ใช้ในการจับคู่การแปลระดับประโยค

โปรแกรมจะเริ่มจากการคำนวณค่าตัวแปรที่ใช้ในการจับคู่การแปลระดับประโยคจากคลังประโยคคู่ขนานตั้งต้น ตัวแปรที่ใช้มี 3 ชุด ได้แก่ ตารางความถี่สัมพัทธ์ของรูปแบบการจับคู่การแปลย่อยแต่ละแบบ (ดูตารางที่ 11) ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของอัตราส่วนความยาวประโยคภาษาอังกฤษต่อความยาวประโยคภาษาไทย (ดูสมการที่ 4 และ 5) และตารางความน่าจะเป็นการแปล (Translation probability table) ที่ได้จากการคำนวณด้วยชุดคำสั่ง GIZA++ (Och and Ney, 2003) บนชุดข้อมูลฝึกฝน โดยที่การคำนวณค่าความน่าจะเป็นการแปลจะเป็นคำนวณโดยขั้นตอนวิธีการ EM (Expectation Maximization Algorithm) เพื่อหาตัวแปรซ่อนเร้นที่ระบุความน่าจะเป็นการแปล $tr(w_{target} \mid w_{src})$ คำในภาษาต้นทาง (w_{src}) ไปเป็นคำในภาษาปลายทาง (w_{target}) ตารางที่ 12 เป็นตัวอย่างค่าความน่าจะเป็นการแปลที่ได้จากชุดคำสั่ง GIZA++ จากตัวอย่างนี้จะได้ว่าคำที่เป็นคู่การแปลกันมักจะมีค่าความน่าจะเป็นการแปลสูง เช่น beach – หาด/ชายหาด ซึ่งมีค่าความน่าจะเป็นการแปล 0.62 และ 0.31 ตามลำดับ และคำที่ไม่ได้เป็นคู่การแปลกันมักจะมีค่าความน่าจะเป็นการแปลต่ำ เช่น beach – ระยะทาง ซึ่งมีค่าความน่าจะเป็นการแปลเพียง 2.17×10^{-6}

ตารางที่ 12 ตัวอย่างตารางค่าความน่าจะเป็นการแปล

w_{src}	w_{target}	$tr(w_{target} \mid w_{src})$
beach	ชายหาด	0.31
beach	หาด	0.62
beach	ระยะทาง	2.17×10^{-6}
ชายหาด	beach	0.99

การคำนวณค่าใช้จ่ายของการจับคู่ประโยค

การจับคู่การแปลของประโยคภาษาไทย t กับประโยคภาษาอังกฤษ e ด้วยรูปแบบการจับคู่แบบ (n_t, n_e) จะถูกกำหนดด้วยฟังก์ชั่นค่าใช้จ่ายแบบถ่วงน้ำหนัก (c_{total}) ซึ่งคำนวณจากค่าใช้จ่ายจากรูปแบบการจับคู่การแปล (c_b) ค่าใช้จ่ายจากอัตราส่วนความยาวประโยค (c_l) และค่าใช้จ่ายจากความน่าจะเป็นการแปล (c_{tr}) และค่าถ่วงน้ำหนักของแต่ละฟังก์ชั่นค่าใช้จ่าย (w_b , w_l , และ w_{tr}) ดังสมการที่ 8

$$c_{total}((n_t, n_e), t, e) = w_b c_b(n_t, n_e) + w_l c_l(t, e) + w_{tr} c_{tr}(t, e) \quad (8)$$

โดย c_{total} หมายถึงค่าใช้จ่ายรวมสำหรับการจับคู่การแปล

c_b หมายถึงค่าใช้จ่ายจากรูปแบบการจับคู่ย่อย

c_l หมายถึงค่าใช้จ่ายจากส่วนต่างอัตราส่วนความยาวประโยค

c_{tr} หมายถึงค่าใช้จ่ายจากความน่าจะเป็นการแปล

w_b หมายถึงค่าถ่วงน้ำหนักสำหรับค่าใช้จ่ายจากรูปแบบการจับคู่ย่อย

w_l หมายถึงค่าถ่วงน้ำหนักสำหรับค่าใช้จ่ายจากส่วนต่างอัตราส่วนความยาว

w_t หมายถึงค่าถ่วงน้ำหนักสำหรับค่าใช้จ่ายจากความน่าจะเป็นการแปล

n_t หมายถึงจำนวนอนุพากย์ภาษาไทย

n_e หมายถึงจำนวนประโยคภาษาอังกฤษ

t หมายถึงข้อความภาษาไทย

e หมายถึงข้อความภาษาอังกฤษ

ค่าใช้จ่ายจากรูปแบบการจับคู่ (c_b) จะคำนวณจากความน่าจะเป็นของรูปแบบการจับคู่แบบ (n_t, n_e) ที่พบในคลังประโยคคู่ขนานตั้งต้น (S) ดังสมการที่ 9 ฟังก์ชั่นค่าใช้จ่ายจากรูปแบบการจับคู่นี้ทำให้รูปแบบการจับคู่ที่มีโอกาสเกิดขึ้นน้อยในชุดข้อมูลฝึกฝน มีค่าใช้จ่ายสูงกว่าการจับคูโดยใช้รูปแบบการจับคู่ที่เกิดขึ้นบ่อยกว่า ดังตัวอย่างที่แสดงในตารางที่ 13 เช่น รูปแบบการจับคู่แบบแทนที่ (1-1) มีโอกาสพบในคลังประโยคตั้งต้นสูงกว่าการจับคู่แบบอื่นหรือคิดเป็น 51% ของการจับคู่ทั้งหมดที่พบในคลังประโยคตั้งต้น ดังนั้นคู่ประโยคที่จับคู่กันแบบแทนที่จึงมีค่าใช้จ่ายต่ำซึ่งในกรณีนี้คิดเป็น 0.67 และในทางกลับกันการจับคู่แบบแทรก (0-1) มีโอกาสเกิดขึ้นในคลังประโยคตั้งต้นเพียง 3.4% ดังนั้นการจับคู่แบบแทรกจึงมีค่าใช้จ่ายสูงซึ่งคิดเป็น 3.38

$$c_b(n_t, n_e) = -\log \Pr[(n_t, n_e) | S] \quad (9)$$

โดย c_b หมายถึงค่าใช้จ่ายจากรูปแบบการจับคู่ย่อย

(n_t, n_e) หมายถึงรูปแบบการจับคู่ย่อย

$Pr[(n_t, n_e)|S]$ หมายถึงความน่าจะเป็นในการเกิดการจับคู่ย่อยแบบ (n_t, n_e) ใน S

ตารางที่ 13 ตัวอย่างการคำนวณค่าใช้จ่ายจากรูปแบบการจับคู่การแปลย่อย

n_t	n_e	$\Pr[(n_t, n_e) \mid S]$	c_b
0	1	0.034	3.38
1	0	0.210	1.56
1	1	0.510	0.67

ฟังก์ชั่นค่าใช้จ่ายจากอัตราส่วนความยาวประโยค (c_l) จะคำนวณจากอัตราส่วนความยาวประโยคภาษาไทย (l_t) และความยาวประโยคภาษาอังกฤษ (l_e) เทียบกับค่าเฉลี่ย (c) และส่วนเบี่ยงเบนมาตรฐาน (s) ของอัตราส่วนความยาวประโยคที่วัดได้จากชุดข้อมูลฝึกฝน การคำนวณภายในเครื่องหมายค่าสัมบูรณ์ในสมการที่ 10 เป็นการแปลงค่าการเปรียบเทียบอัตรส่วนให้มีการประจายตัวเป็นแบบปกติมาตรฐาน (ค่าเฉลี่ยเท่ากับศูนย์และส่วนเบี่ยงเบนมาตรฐานเท่ากับหนึ่ง) และการคำนวณในสมการที่ 11 เป็นการคำนวณค่าส่วนกลับของความน่าจะเป็นที่ประโยคภาษาไทยที่มีความยาว l_t จะจับคู่กับประโยคภาษาอังกฤษที่มีความยาว l_e ซึ่งทำให้คู่ประโยคที่มีอัตราส่วนความยาวประโยคแตกต่างจากที่วัดได้มากจะมีค่าใช้จ่ายในการจับคู่สูง

$$\delta(l_t, l_e) = \left| \frac{(l_e - l_t c)}{s(l_e / c + l_t)} \right| \tag{10}$$

$$c_l(t, e) = -\log 2(1 - \Pr[\delta(l_t, l_e)]) \tag{11}$$

โดย c_l หมายถึงค่าใช้จ่ายจากส่วนต่างอัตราส่วนความยาวประโยค

l_e หมายถึงความยาวอนุพากย์ภาษาไทย

l_t หมายถึงความยาวประโยคภาษาอังกฤษ

c หมายถึงค่าเฉลี่ยอัตราส่วนความยาวประโยค

s หมายถึงค่าส่วนเบี่ยงเบนมาตรฐานอัตราส่วนความยาวประโยค

δ หมายถึงฟังก์ชั่นคำนวณส่วนต่างอัตราส่วนความยาวที่มีการกระจายตัวแบบปกติมาตรฐาน

$2(1 - \Pr[\delta(l_t, l_e)])$ หมายถึงค่าความน่าจะเป็นที่คู่ประโยคที่มีส่วนต่างอัตราส่วนความยาวเท่ากับ δ จะไม่จับคู่กัน

จากผลการทดลองเบื้องต้นผบว่าค่าอัตราส่วนความประโยคภาษาอังกฤษต่ออนุพากย์ภาษาไทยมีค่าประมาณ 0.83 และได้ค่าส่วนเบี่ยงเบนมาตรฐานประมาณ 0.30 ดังนั้นคู่ประโยคใดมีค่าอัตราส่วนนี้แตกต่างจากค่า 0.8 มากจะมีค่าใช้จ่ายที่สูงขึ้น ดังตัวอย่างที่แสดงในตารางที่ 14 ถ้าอัตราส่วนความยาวข้อความภาษาอังกฤษต่อภาษาไทยมีค่า 0.8 จะทำให้ค่าใช้จ่ายในการจับคู่มีค่าต่ำเข้าใกล้ศูนย์ และค่าใช้จ่ายจะสูงขึ้นเมื่ออัตราส่วนความยาวแตกต่างจากค่า 0.8 มากขึ้น

ตารางที่ 14 ตัวอย่างการคำนวณค่าใช้จ่ายจากอัตราส่วนความยาวประโยค

l_t	l_e	l_e / l_t	c_l
100	60	0.6	0.42
100	80	0.8	0.04
100	100	1.0	0.23
100	120	1.2	0.49

ฟังก์ชั่นค่าใช้จ่ายจากความน่าเป็นการแปล (c_{tr}) จะคำนวณจากส่วนกลับของผลรวมความน่าจะเป็นของการแปลระหว่างคำในประโยคภาษาไทย (ทุก w_t ใน t) และประโยคภาษาปลายทาง (ทุก w_e ใน e) โดยใช้ตารางความน่าจะเป็นการแปลมาคำนวณร่วมด้วย (ดูตารางที่ 12 ประกอบ) การคำนวณความน่าจะเป็นการแปลทั้งสองทิศทางแบบไม่สนใจตำแหน่งของคำ (Bag of words) และตัดพจน์ความน่าจะเป็นเบื้องต้น ($p(t)$ และ $p(e)$) ที่เกิดจากจำนวนคำออก ($|e|\cdot\log|t|+|t|\cdot\log|e|$) เนื่องจากสามารถทดแทนได้ด้วยค่าใช้จ่ายที่เกิดจากความยาวประโยค (c_l) และการใส่พจน์ที่เกิดจากจำนวนคำเข้าไปอาจทำให้ประโยคที่ยาวจะมีค่าใช้จ่ายที่สูงมากเกินไป ดังแสดงในสมการ 12 และจากตัวอย่างในตารางที่ 12 แสดงให้เห็นว่าคำแต่ละคำมีโอกาสแปลไปได้มากกว่าหนึ่งคำแปล เช่น “beach” สามารถแปลได้ทั้ง “หาด” และ “ชายหาด” แต่คำที่เป็นคู่การแปลกันมักจะมีความน่าจะเป็นการแปลสูงกว่าคำที่ไม่ได้เป็นคู่การแปลกัน และเมื่อคำนวณค่าใช้จ่ายแล้วจะทำให้ได้ค่าใช้จ่ายในการจับคู่ต่ำกว่า เช่น ความน่าจะเป็นในการแปลคำว่า “beach” ไปเป็นคำว่า “หาด” และ “ชายหาด” จะสูงกว่าความน่าจะเป็นในการแปลไปเป็นคำว่า “ระยะทาง” เป็นต้น

$$c_{tr}(t,e) = -\log\prod_{w_t\in t}\left[\sum_{w_e\in e} tr(w_t|w_e)\right] - \log\prod_{w_e\in e}\left[\sum_{w_t\in t} tr(w_e|w_t)\right] \qquad (12)$$

โดย c_{tr} หมายถึงค่าใช้จ่ายจากความน่าจะเป็นการแปล

t หมายถึงข้อความภาษาไทย

e หมายถึงข้อความภาษาอังกฤษ

w_t หมายถึงคำในข้อความภาษาไทย

w_e หมายถึงคำในข้อความภาษาอังกฤษ

$tr(w_t|w_e)$ หมายถึงความน่าจะเป็นการแปลจากคำภาษาอังกฤษเป็นไทย

$tr(w_e|w_t)$ หมายถึงความน่าจะเป็นการแปลจากคำภาษาไทยเป็นอังกฤษ

การปรับค่าถ่วงน้ำหนักของแต่ละฟังก์ชั่นค่าใช้จ่าย

การใช้ฟังก์ชั่นค่าใช้จ่ายจากข้อสนเทศต่างชนิดกัน ทำให้เป็นไปได้ว่าค่าใช้จ่ายจากบางข้อสนเทศจะมีผลกระทบกับการจับคู่การแปลมากกว่าข้อสนเทศอื่น จึงจำเป็นต้องมีการถ่วงน้ำหนักค่าใช้จ่ายของแต่ละข้อสนเทศเพื่อปรับขนาดค่าใช้จ่ายจากแต่ละฟังก์ชั่นค่าใช้จ่ายมีค่าเทียบเท่ากัน ในงานวิจัยนี้ใช้วิธีปรับค่าถ่วงน้ำหนักของฟังก์ชั่นค่าใช้จ่ายทั้งสามฟังก์ชั่นโดยอัตโนมัติด้วยการคำนวณจากสัดส่วนผลกระทบของแต่ละค่าใช้จ่ายจากคู่ประโยคในคลังประโยคคู่ขนานตั้งต้น โดยแต่ละการจับคู่ $((n_t, n_e), t, e)$ ในคลังประโยคคู่ขนานตั้งต้น (S) โปรแกรมจะคำนวณค่าสัดส่วนค่าใช้จ่ายแต่ละค่าเทียบกับค่าใช้จ่ายรวมแบบไม่มีค่าถ่วงน้ำหนัก (λ) และกำหนดค่าถ่วงน้ำหนักของฟังก์ชั่นค่าใช้จ่ายจากค่าเฉลี่ยสัดส่วนค่าใช้จ่ายรวมของค่าใช้จ่ายอื่น เช่น ค่าใช้จ่ายจากรูปแบบการจับคู่ (w_b) จะคำนวณจากสัดส่วนคค่าใช้จ่ายรวมระหว่างค่าใช้จ่ายจากความยาวประโยค (c_l) และค่าใช้จ่ายจากความน่าจะเป็นการแปล (c_{tr}) การปรับค่าถ่วงน้ำหนักแบบนี้จะส่งผลให้ค่าใช้จ่ายที่มักมีค่ามากจะมีค่าถ่วงน้ำหนักที่น้อยและค่าใช้จ่ายที่มักมีค่าน้อยจะมีค่าถ่วงน้ำหนักมาก ดังแสดงในสมการต่อไปนี้

$$\lambda((n_t, n_e), t, e) = c_b(n_t, n_e) + c_l(t, e) + c_{tr}(t, e) \tag{13}$$

$$w_b = \frac{1}{|S|} \sum_{((n_t, n_e), t, e) \in S} \frac{c_l(t, e) + c_{tr}(t, e)}{\lambda((n_t, n_e), t, e)} \tag{14}$$

$$w_l = \frac{1}{|S|} \sum_{((n_t, n_e), t, e) \in S} \frac{c_b(n_t, n_e) + c_{tr}(t, e)}{\lambda((n_t, n_e), t, e)} \tag{15}$$

$$w_{tr} = \frac{1}{|S|} \sum_{((n_t, n_e), t, e) \in S} \frac{c_b(n_t, n_e) + c_l(t, e)}{\lambda((n_t, n_e), t, e)} \tag{16}$$

โดย λ หมายถึงค่าใช้จ่ายรวมสำหรับการจับคู่การแปลแบบไม่คิดค่าถ่วงน้ำหนัก

c_b หมายถึงค่าใช้จ่ายจากรูปแบบการจับคู่ย่อย

c_l หมายถึงค่าใช้จ่ายจากส่วนต่างอัตราส่วนความยาวประโยค

c_{tr} หมายถึงค่าใช้จ่ายจากความน่าจะเป็นการแปล

w_b หมายถึงค่าถ่วงน้ำหนักสำหรับค่าใช้จ่ายจากรูปแบบการจับคู่ย่อย

w_l หมายถึงค่าถ่วงน้ำหนักสำหรับค่าใช้จ่ายจากส่วนต่างอัตราส่วนความยาว

w_t หมายถึงค่าถ่วงน้ำหนักสำหรับค่าใช้จ่ายจากความน่าจะเป็นการแปล

การปรับฟังก์ชั่นค่าใช้จ่ายสำหรับการจับคู่แบบแทรกและแบบลบ

เนื่องจากปริมาณการจับคู่แบบแทรกและแบบลบในเอกสารที่แปลโดยอรรถมีปริมาณมาก (ดูตารางที่ 11) และมักจะเกิดจากความคิดเห็นของผู้แปลเอง นั่นคือข้อความเดียวกันในบริบทเดียวกันอาจเป็นไปได้ทั้งที่จะถูกแปลและไม่ถูกแปล ดังนั้นจึงไม่สามารถคำนวณค่าใช้จ่ายสำหรับการจับคู่แบบแทรกและแบบลบได้โดยตรงเหมือนการจับคู่ชนิดอื่น และถึงแม้ว่าฟังก์ชั่นค่าใช้จ่ายจากความยาวประโยค (c_l) หรือความน่าจะเป็นการแปล (c_{tr}) ดังที่กล่าวมาแล้วจะสามารถคำนวณค่าใช้จ่ายสำหรับการจับคู่แบบแทรกและแบบลบได้ แต่ค่าใช้จ่ายที่คำนวณได้จะเป็นการจับคู่กับประโยคว่าง (Null sentence) ซึ่งเป็นค่าที่สูงเกินไป และอาจส่งผลให้โปรแกรมไม่พยายามทำการจับคู่แบบแทรกหรือแบบลบ (Undergenerate) งานวิจัยนี้จึงกำหนดค่าใช้จ่ายจากความยาวประโยคให้เป็นค่ากึ่งกลางระหว่างจับและไม่จับคู่หรือ $-\log(0.5)$ และกำหนดค่าใช้จ่ายจากความน่าจะเป็นการแปลให้เป็นค่าใช้จ่ายที่สูงที่สุดที่พบในคลังประโยคคู่ขนานตั้งต้น ($\max_{(\hat{t},\hat{e})\in S} c_{tr}(\hat{t},\hat{e})$) เนื่องจากค่าความน่าจะเป็นในการแปลคู่ประโยคที่เป็นคู่การแปลกันจริงอาจมีค่าต่ำกว่า 0.5 เนื่องจากค่าความน่าจะเป็นการแปลคู่ประโยคจะคำนวณจากการนำค่าความน่าจะเป็นของการแปลคำมาคูณกัน ฟังก์ชั่นค่าใช้จ่ายที่ปรับแต่งสำหรับการจับคู่แบบแทรกและแบบลบแสดงในสมการ 17 และ 18

$$c_l((n_t,n_e),t,e)=\begin{cases}-\log(0.5) & ,n_t=0\vee n_e=0\\ -\log 2(1-\Pr[\delta(l_t,l_e)]) & ,otherwise\end{cases} \tag{17}$$

$$c_{tr}((n_t,n_e),t,e)=\begin{cases}\max_{(\hat{t},\hat{e})\in S} c_{tr}(\hat{t},\hat{e}) & ,n_t=0\vee n_e=0\\ -\log\prod_{w_t\in t}\left[\sum_{w_e\in e} tr(w_t|w_e)\right]-\log\prod_{w_e\in e}\left[\sum_{w_t\in t} tr(w_e|w_t)\right] & ,otherwise\end{cases} \tag{18}$$

โดย c_l หมายถึงค่าใช้จ่ายจากส่วนต่างอัตราส่วนความยาวประโยค

c_{tr} หมายถึงค่าใช้จ่ายจากความน่าจะเป็นการแปล

$\max_{(\hat{t},\hat{e})\in S} c_{tr}(\hat{t},\hat{e})$ หมายถึงค่าใช้จ่ายจากความน่าจะเป็นการแปลที่สูงที่สุดในชุดข้อมูลฝึกฝน

n_t หมายถึงจำนวนอนุพากย์ภาษาไทย

n_e หมายถึงจำนวนประโยคภาษาอังกฤษ

t หมายถึงข้อความภาษาไทย

e หมายถึงข้อความภาษาอังกฤษ

l_e หมายถึงเป็นความยาวอนุพากย์ภาษาไทย

l_t หมายถึงเป็นความยาวประโยคภาษาอังกฤษ

การหารูปแบบการจับคู่การแปลระดับประโยคที่มีค่าใช้จ่ายน้อยที่สุด

การหารูปแบบการจับคู่การแปลระดับประโยคสำหรับคู่เอกสารหนึ่งๆ จะเป็นการหาลำดับของการจับคู่ย่อยที่มีผลรวมค่าใช้จ่ายต่ำที่สุด ถ้ากำหนดให้ s_i^n เป็นข้อความที่เกิดจากนำประโยคหรืออนุพากย์ที่ i จนถึง $i+n-1$ มาต่อกัน (ดูสมการที่ 19) และ $Opt[i, j]$ เป็นค่าใช้จ่ายต่ำที่สุดของการจับคู่การแปลระดับประโยคระหว่างเอกสารภาษาไทยที่มี i อนุพากย์ และเอกสารภาษาอังกฤษที่มี j ประโยค จะสามารถกำหนดสมการเวียนเกิดของ $Opt[i, j]$ ได้ดังสมการที่ 20 และสามารถหาการจับคู่ผลลัพธ์ที่มีค่าใช้จ่ายที่ดีที่สุดได้โดยการโปรแกรมเชิงพลวัตร ดังอธิบายในภาพที่ 10

$$s_i^n = \begin{cases} \varnothing & ,n \le 0 \\ s_i s_{i+1} \ldots s_{i+n-1} & ,n > 0 \end{cases} \tag{19}$$

$$Opt[i, j] = \begin{cases} 0 & ,i = 0 \wedge j = 0 \\ \infty & ,i < 0 \vee j < 0 \\ \min\limits_{(n_t, n_e) \in P} Opt[i - n_t, j - n_e] + c_{total}(n_t, n_e, t_{i-n_t}^{n_t}, e_{j-n_e}^{n_e}) & ,otherwise \end{cases} \tag{20}$$

โดย P หมายถึงเซตของรูปแบบการจับคู่ย่อย

$Opt[i,j]$ หมายถึงค่าใช้จ่ายต่ำที่สุดของการจับคู่การแปลระดับประโยคระหว่างเอกสารภาษาไทยที่มี i อนุพากย์และเอกสารภาษาอังกฤษที่มี j ประโยค

c_{total} หมายถึงค่าใช้จ่ายรวมสำหรับการจับคู่การแปล

Input : คู่เอกสารขนาน (T, E) , เซตรูปแบบการจับคู่ย่อย P

Output : ตารางค่าใช้จ่ายสำหรับการจับคู่การแปลระดับประโยค Opt

```
1  for i ← 0 ...|T| do
2      for j ← 0 ...|E| do
3          cost_min ← ∞                                  { กำหนดค่าใช้จ่ายต่ำสุด }
4          for p ← 1 ...|P| do                           { คำนวณค่าใช้จ่ายของทุกรูปแบบการจับคู่ย่อย }
5              (n_th, n_en) ← P_p
6              c ← c_total((n_t, n_e), T^{n_th}_{i-n_t}, E^{n_en}_{j-n_e})   { คำนวณค่าใช้จ่ายของการจับคู่แบบ P_p }
7              if c < cost_min then cost_min ← c          { ปรับค่าใช้จ่ายต่ำสุดใหม่ }
8          Opt[i, j] ← cost_min                          { กำหนดค่าใช้จ่ายที่ต่ำที่สุดลงที่ตำแหน่ง i , j }
9  return Opt
```

ภาพที่ 10 ขั้นตอนวิธีการคำนวณค่าใช้จ่ายที่ดีที่สุด

การวัดผลความถูกต้อง

ในงานวิจัยนี้จะวัดผลความถูกต้องของโปรแกรม โดยเปรียบเทียบผลการจับคู่การแปลที่โปรแกรมทำได้เทียบกับผลการจับคู่การแปลด้วยมือหรือโดยคน การวัดผลจะนับจำนวนการจับคู่ย่อยที่ถูกต้อง และวัดผลเป็นค่าความแม่นยำ (Precision, P) ค่าความครอบคลุม (Recall, R) และค่าคะแนนเอฟ (F score, F) โดยมีค่าถ่วงน้ำหนักเบต้า (β) เท่ากับ 1 ดังสมการที่ 21 22 และ 23 ตามลำดับ

$$P = \frac{C}{G} \tag{21}$$

$$R = \frac{C}{A} \tag{22}$$

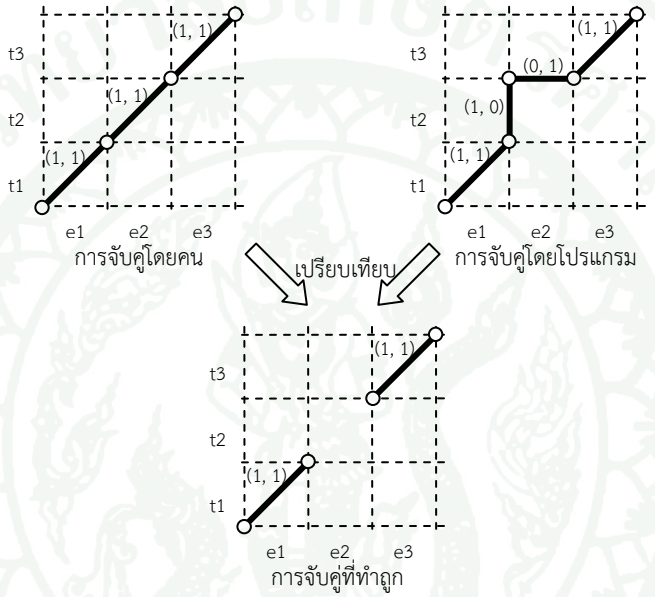
$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} = \frac{2PR}{P + R} \tag{23}$$

โดย C หมายถึงจำนวนคู่ประโยคที่สกัดได้ถูกต้อง

G หมายถึงจำนวนคู่ประโยคทั้งหมดที่โปรแกรมสกัดได้

A หมายถึงจำนวนคู่ประโยคทั้งหมดที่สกัดโดยคน

ภาพที่ 11 แสดงตัวอย่างการเปรียบเทียบผลการจับคู่ระหว่างผลการจับคู่ด้วยมือและการจับคู่โดยเครื่อง จากตัวอย่างนี้จะได้ว่าคนกำกับคู่ประโยคไว้ 3 คู่ ได้แก่ (t_1,e_1), (t_2,e_2), และ (t_3,e_3) โปรแกรมจับคู่ประโยคได้ทั้งหมด 4 คู่ ได้แก่ (t_1,e_1), (t_2,$\emptyset$), ($\emptyset$,e_2), และ (t_3,e_3) แต่ตรงกับที่คนกำกับคำตอบไว้ 2 คู่ประโยค ได้แก่ (t_1,e_1) และ (t_3,e_3) คิดเป็นค่าความแม่นยำ (Precision) 2/4 หรือ 0.50 ค่าความครอบคลุม (Recall) 2/3 หรือ 0.67 และค่าคะแนนเอฟ (F score) เป็น 0.57



ภาพที่ 11 ตัวอย่างการเปรียบเทียบการจับคู่โดยคนและโปรแกรม

ผลและวิจารณ์

การทดลอง

การทดลองที่ทำขึ้นเป็นการทดลองเพื่อเปรียบเทียบความถูกต้องของการสกัดคู่ประโยคจากการใช้ข้อสนเทศจากความยาวประโยค ข้อสนเทศการแปล การปรับค่าถ่วงน้ำหนัก และการปรับแต่งฟังก์ชั่นค่าใช้จ่ายตามชนิดของการจับคู่ การทดลองมีทั้งหมด 6 แบบ โดยพิจารณาจากค่าตัวแปรที่ต่างกัน ได้แก่ ค่าใช้จ่ายจากรูปแบบการจับคู่ ค่าใช้จ่ายจากความยาวประโยค ค่าใช้จ่ายจากความน่าจะเป็นการแปล การปรับค่าถ่วงน้ำหนัก และการปรับแต่งฟังก์ชั่นค่าใช้จ่ายสำหรับการจับคู่แบบแทรกและแบบลบ แต่ละแบบจะทดลองด้วยการทำการทดลองไขว้แบบ 5 ทบ (5-fold cross validation) ซึ่งชุดข้อมูลสำหรับการทดลองจะถูกแบ่งเป็น 2 ส่วนที่ไม่ทับซ้อนกัน ได้แก่ ส่วนสำหรับฝึกฝนโปรแกรม (Training set) และส่วนสำหรับทดสอบ (Test set) รายละเอียดของการทดลองแต่ละแบบได้แสดงไว้ในตารางที่ 15 โดยการทดลองแต่ละแบบจะใช้ข้อสนเทศจากรูปแบบการจับคู่ และความยาวประโยคเป็นข้อสนเทศพื้นฐานในการสกัดคู่ประโยค

ตารางที่ 15 รายละเอียดการทดลองแต่ละแบบ กับเอกสารเกี่ยวกับการท่องเที่ยว

แบบที่	ค่าใช้จ่ายจากรูปแบบการจับคู่	ค่าใช้จ่ายจากความยาวประโยค	ค่าใช้จ่ายจากความน่าจะเป็นการแปล	ปรับค่าถ่วงน้ำหนัก	ปรับแต่งฟังก์ชั่นค่าใช้จ่าย
1	✓	✓	✗	✗	✗
2	✓	✓	✓	✗	✗
3	✓	✓	✓	✓	✗
4	✓	✓	✗	✗	✓
5	✓	✓	✓	✗	✓
6	✓	✓	✓	✓	✓

ผลการทดลอง

ผลการทดลองทั้ง 6 แบบ โดยตารางที่ 16 ถึงตารางที่ 27 เป็นผลลัพธ์การจับคู่การแปลระดับประโยคบนเอกสารการท่องเที่ยวทั้งบนชุดข้อมูลฝึกฝนและข้อมูลทดสอบ และตารางที่ 28 ถึงตารางที่

39 เป็นผลลัพธ์การจับคู่การแปลระดับประโยคบนรัฐธรรมนูญทั้งบนชุดข้อมูลฝึกฝนและข้อมูลทดสอบ ทั้งนี้การวิเคราะห์ผลลัพธ์ให้ดูในบทการวิเคราะห์ผลการทดลอง

ตารางที่ 16 ผลการทดลองแบบที่ 1 บนชุดข้อมูลฝึกฝนของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	128	0	0	-	0
1-0	792	0	0	-	0
1-1	1920	1781	1183	0.6642	0.6161
2-1	616	734	229	0.3120	0.3718
3-1	196	285	44	0.1544	0.2245
4-1	56	224	16	0.0714	0.2857
อื่นๆ	56	0	0	-	0
รวม	3764	3024	1472	0.4868	0.3911

ตารางที่ 17 ผลการทดลองแบบที่ 1 บนชุดข้อมูลทดสอบของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	32	0	0	-	0
1-0	198	0	0	-	0
1-1	480	444	283	0.6374	0.5896
2-1	154	186	54	0.2903	0.3506
3-1	49	70	14	0.2000	0.2857
4-1	14	56	4	0.0714	0.2857
อื่นๆ	14	0	0	-	0
รวม	941	756	355	0.4696	0.3773

ตารางที่ 18 ผลการทดลองแบบที่ 2 บนชุดข้อมูลฝึกฝนของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	128	87	70	0.8046	0.5469
1-0	792	44	42	0.9545	0.0530
1-1	1920	1734	1666	0.9608	0.8677
2-1	616	634	457	0.7208	0.7419
3-1	196	322	147	0.4565	0.7500
4-1	56	247	54	0.2186	0.9643
อื่นๆ	56	0	0	-	0
รวม	3764	3068	2436	0.7940	0.6472

ตารางที่ 19 ผลการทดลองแบบที่ 2 บนชุดข้อมูลทดสอบของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	32	32	12	0.3750	0.3750
1-0	198	3	3	1.0000	0.0152
1-1	480	413	372	0.9007	0.7750
2-1	154	168	96	0.5714	0.6234
3-1	49	74	24	0.3243	0.4898
4-1	14	69	9	0.1304	0.6429
อื่นๆ	14	0	0	-	0
รวม	941	759	516	0.6798	0.5484

ตารางที่ 20 ผลการทดลองแบบที่ 3 บนชุดข้อมูลฝึกฝนของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	128	15	15	1.0000	0.1172
1-0	792	0	0	-	0
1-1	1920	1825	1650	0.9041	0.8594
2-1	616	617	420	0.6841	0.6818
3-1	196	327	142	0.4343	0.7245
4-1	56	240	44	0.1833	0.7245
อื่นๆ	56	0	0	-	0
รวม	3764	3025	2271	0.7510	0.6033

ตารางที่ 21 ผลการทดลองแบบที่ 3 บนชุดข้อมูลทดสอบของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	32	1	1	1.0000	0.1111
1-0	198	0	0	-	0
1-1	480	453	381	0.8411	0.7937
2-1	154	166	88	0.5301	0.5714
3-1	49	79	25	0.3165	0.5102
4-1	14	57	7	0.1228	0.5000
อื่นๆ	14	0	0	-	0
รวม	941	756	502	0.6640	0.5335

ตารางที่ 22 ผลการทดลองแบบที่ 4 บนชุดข้อมูลฝึกฝนของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	128	0	0	-	0
1-0	792	82	28	0.3415	0.0354
1-1	1920	1806	1204	0.6667	0.6271
2-1	616	751	235	0.3129	0.3815
3-1	196	258	44	0.1705	0.2245
4-1	56	209	19	0.0909	0.3393
อื่นๆ	56	0	0	-	0
รวม	3764	3106	1530	0.4926	0.4065

ตารางที่ 23 ผลการทดลองแบบที่ 4 บนชุดข้อมูลทดสอบของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	32	0	0	-	0
1-0	198	21	6	0.2857	0.0303
1-1	480	451	289	0.6408	0.6021
2-1	154	189	53	0.2804	0.3442
3-1	49	64	12	0.1875	0.2449
4-1	14	52	5	0.0962	0.3571
อื่นๆ	14	0	0	-	0
รวม	941	777	365	0.4698	0.3879

ตารางที่ 24 ผลการทดลองแบบที่ 5 บนชุดข้อมูลฝึกฝนของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	128	129	100	0.7752	0.7812
1-0	792	325	312	0.9600	0.3939
1-1	1920	1769	1736	0.9813	0.9042
2-1	616	642	495	0.7710	0.8036
3-1	196	314	170	0.5414	0.8673
4-1	56	170	56	0.3294	1.0000
อื่นๆ	56	0	0	-	0
รวม	3764	3349	2869	0.8567	0.7622

ตารางที่ 25 ผลการทดลองแบบที่ 5 บนชุดข้อมูลทดสอบของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	32	165	19	0.1152	0.5938
1-0	198	269	93	0.3457	0.4697
1-1	480	354	319	0.9011	0.6646
2-1	154	132	72	0.5455	0.4675
3-1	49	57	20	0.3509	0.4082
4-1	14	48	8	0.1667	0.5714
อื่นๆ	14	0	0	-	0
รวม	941	1025	531	0.5180	0.5643

ตารางที่ 26 ผลการทดลองแบบที่ 6 บนชุดข้อมูลฝึกฝนของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	128	53	53	1.0000	0.4141
1-0	792	235	226	0.9617	0.2854
1-1	1920	1843	1727	0.9371	0.8995
2-1	616	641	474	0.7395	0.7695
3-1	196	308	166	0.5390	0.8469
4-1	56	179	51	0.2849	0.9107
อื่นๆ	56	0	0	-	0
รวม	3764	3259	2697	0.8276	0.7165

ตารางที่ 27 ผลการทดลองแบบที่ 6 บนชุดข้อมูลทดสอบของเอกสารเกี่ยวกับการท่องเที่ยว

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	32	20	10	0.5000	0.3125
1-0	198	113	64	0.5664	0.3232
1-1	480	478	404	0.8452	0.8417
2-1	154	152	94	0.6184	0.6104
3-1	49	69	29	0.4203	0.5918
4-1	14	37	8	0.2162	0.5714
อื่นๆ	14	0	0	-	0
รวม	941	869	609	0.7008	0.6472

ตารางที่ 28 ผลการทดลองแบบที่ 1 บนชุดข้อมูลฝึกฝนของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	12	12	4	0.3333	0.3333
1-0	4	4	0	0	0
1-1	6512	6512	6496	0.9975	0.9975
รวม	6528	6528	6500	0.9957	0.9957

ตารางที่ 29 ผลการทดลองแบบที่ 1 บนชุดข้อมูลทดสอบของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	3	3	1	0.3333	0.3333
1-0	1	1	1	1.0000	1.0000
1-1	1628	1628	1626	0.9988	0.9988
รวม	1632	1632	1628	0.9975	0.9975

ตารางที่ 30 ผลการทดลองแบบที่ 2 บนชุดข้อมูลฝึกฝนของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	12	12	12	1.0000	1.0000
1-0	4	0	0	-	0
1-1	6512	6512	6512	1.0000	1.0000
2-1	0	4	0	0	-
รวม	6528	6524	6520	0.9994	0.9988

ตารางที่ 31 ผลการทดลองแบบที่ 2 บนชุดข้อมูลทดสอบของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	3	3	3	1.0000	1.0000
1-0	1	0	0	-	0
1-1	1628	1627	1627	1.0000	0.9994
2-1	0	1	0	0.0000	-
รวม	1632	1631	1630	0.9994	0.9988

ตารางที่ 32 ผลการทดลองแบบที่ 3 บนชุดข้อมูลฝึกฝนของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	12	12	12	1.0000	1.0000
1-0	4	4	4	1.0000	1.0000
1-1	6512	6512	6512	1.0000	1.0000
รวม	6528	6528	6528	1.0000	1.0000

ตารางที่ 33 ผลการทดลองแบบที่ 3 บนชุดข้อมูลทดสอบของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	3	3	3	1.0000	1.0000
1-0	1	0	0	-	0
1-1	1628	1628	1628	1.0000	1.0000
2-1	0	1	0	0	-
รวม	1632	1631	1629	0.9988	0.9982

ตารางที่ 34 ผลการทดลองแบบที่ 4 บนชุดข้อมูลฝึกฝนของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	12	12	12	1.0000	1.0000
1-0	4	4	4	1.0000	1.0000
1-1	6512	6512	6512	1.0000	1.0000
รวม	6528	6528	6528	1.0000	1.0000

ตารางที่ 35 ผลการทดลองแบบที่ 4 บนชุดข้อมูลทดสอบของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	3	3	3	1.0000	1.0000
1-0	1	1	1	1.0000	1.0000
1-1	1628	1628	1628	1.0000	1.0000
รวม	1632	1632	1632	1.0000	1.0000

ตารางที่ 36 ผลการทดลองแบบที่ 5 บนชุดข้อมูลฝึกฝนของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	12	12	12	1.0000	1.0000
1-0	4	0	0	-	0
1-1	6512	6508	6508	1.0000	0.9994
2-1	0	4	0	0	-
รวม	6528	6524	6520	0.9994	0.9988

ตารางที่ 37 ผลการทดลองแบบที่ 5 บนชุดข้อมูลทดสอบของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	3	3	3	1.0000	1.0000
1-0	1	0	0	-	0
1-1	1628	1627	1627	1.0000	0.9994
2-1	0	1	0	0	-
รวม	1632	1631	1630	0.9994	0.9988

ตารางที่ 38 ผลการทดลองแบบที่ 6 บนชุดข้อมูลฝึกฝนของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	12	12	12	1.0000	1.0000
1-0	4	4	4	1.0000	1.0000
1-1	6512	6512	6512	1.0000	1.0000
รวม	6528	6528	6528	1.0000	1.0000

ตารางที่ 39 ผลการทดลองแบบที่ 6 บนชุดข้อมูลทดสอบของรัฐธรรมนูญ

รูปแบบการจับคู่	จำนวนทั้งหมด	จำนวนที่โปรแกรมตอบ	จำนวนที่ตอบถูก	ความแม่นยำ	ความครอบคลุม
0-1	3	3	3	1.0000	1.0000
1-0	1	1	1	1.0000	1.0000
1-1	1628	1628	1628	1.0000	1.0000
รวม	1632	1632	1632	1.0000	1.0000

ตารางที่ 40 และตารางที่ 41 สรุปมาจากตารางที่ 16 ถึงตารางที่ 27 และตารางที่ 28 ถึงตารางที่ 39 โดยพิจารณาจากค่าความแม่นยำและค่าความครอบคลุมของทั้งบนชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบของบทความการท่องเที่ยว และรัฐธรรมนูญตามลำดับ

ตารางที่ 40 ผลการทดลองของแต่ละแบบ บนเอกสารเกี่ยวกับการท่องเที่ยว

แบบที่	ความถูกต้องบนชุดข้อมูลฝึกฝน		ความถูกต้องบนชุดข้อมูลทดสอบ	
	ความแม่นยำ	ความครอบคลุม	ความแม่นยำ	ความครอบคลุม
1	0.4868	0.3911	0.4696	0.3773
2	0.7940	0.6472	0.6798	0.5484
3	0.7510	0.6033	0.6640	0.5335
4	0.4926	0.4065	0.4698	0.3879
5	0.8567	0.7622	0.5180	0.5643
6	0.8276	0.7165	0.7008	0.6472

ตารางที่ 41 ผลการทดลองของแต่ละแบบ บนรัฐธรรมนูญ

แบบที่	ความถูกต้องบนชุดข้อมูลฝึกฝน		ความถูกต้องบนชุดข้อมูลทดสอบ	
	ความแม่นยำ	ความครอบคลุม	ความแม่นยำ	ความครอบคลุม
1	0.9957	0.9957	0.9975	0.9975
2	0.9994	0.9988	0.9994	0.9988
3	1.0000	1.0000	0.9988	0.9982
4	1.0000	1.0000	1.0000	1.0000
5	0.9994	0.9988	0.9994	0.9988
6	1.0000	1.0000	1.0000	1.0000

วิจารณ์ผลการทดลอง

จากตารางที่ 40 และตารางที่ 41 พบว่าความถูกต้องของการการจับคู่การแปลระดับประโยคกับรัฐธรรมนูญมีความถูกต้องสูงโดยมีค่าความแม่นยำและค่าความครอบคลุมตั้งแต่ 0.9957 ถึง 1 บนชุดข้อมูลฝึกฝน และ 0.9975 ถึง 1 บนชุดข้อมูลทดสอบ แต่ในทางกลับกันความถูกต้องของการจับคู่การแปลระดับประโยคกับเอกสารการท่องเที่ยวซึ่งเป็นเอกสารที่แปลโดยอรรถมีความถูกต้องต่ำกว่ามากโดยได้ค่าความแม่นยำ 0.4868 ถึง 0.8567 บนชุดข้อมูลฝึกฝน และ 0.4696 ถึง 0.7008 บนชุดข้อมูลทดสอบ และได้ค่าความครอบคลุม 0.3911 ถึง 0.7622 บนชุดข้อมูลฝึกฝน และ 0.3773 ถึง 0.6472 บนชุดข้อมูลทดสอบ แสดงว่าการแปลที่ทำให้เนื้อความเปลี่ยนแปลงไปหรือการแปลโดยอรรถมีผลทำให้ความถูกต้องของการจับคู่การแปลระดับประโยคลดลง

ตารางที่ 40 แสดงการเปรียบเทียบความถูกต้องของการจับคู่การแปลระดับประโยคกับเอกสารเกี่ยวกับการท่องเที่ยว จากการทดลองพบว่าการจับคู่การแปลระดับประโยคโดยใช้ข้อสนเทศจากความยาวประโยคเพียงอย่างเดียว (การทดลองแบบที่ 1 และ 4) ให้ผลความถูกต้องบนชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบใกล้เคียงกัน โดยมีความถูกต้องต่างกันไม่เกิน 3% ในขณะที่การทดลองแบบที่พิจารณาข้อสนเทศการแปลด้วย (การทดลองแบบที่ 2, 3, 5 และ 6) จะมีความแตกต่างของค่าความถูกต้องระหว่างบนชุดข้อมูลฝึกฝนและข้อมูลทดสอบมาก โดยมีความแม่นยำแตกต่างกันสูงสุดคิดเป็น 34% และความครอบคลุมแตกต่างกันสูงสุดคิดเป็น 20% (การทดลองแบบที่ 5) แสดงว่าการใช้ข้อสนเทศการแปลมาพิจารณาร่วมด้วยจำเป็นต้องมีข้อมูลฝึกฝนมาก เพื่อที่จะสกัดข้อสนเทศการแปลให้ได้ครบถ้วน แต่อย่างไรก็ตามการนำข้อสนเทศการแปลมาพิจารณาร่วมด้วยจะให้ความถูกต้องที่สูงกว่าตั้งแต่ 5% ถึง 23%

จากผลการทดลองตารางที่ 19 (การทดลองแบบที่ 2) และตารางที่ 25 (การทดลองแบบที่ 5) พบว่าการปรับแต่งฟังก์ชั่นค่าใช้จ่ายทำให้ความถูกต้องลดลง เพราะโปรแกรมจะสกัดคู่ประโยคที่จับคู่กันแบบแทรกและแบบลบ (0-1 และ 1-0) มากเกินไป (ดูตารางที่ 25) โดยโปรแกรมตอบการจับคู่แบบแทรก (0-1) 165 คู่จากที่มีจริง 32 คู่ และตอบการจับคู่แบบลบ (1-0) 269 คู่จากที่มีจริง 198 คู่ ซึ่งถึงแม้ว่าจะทำให้ค่าความครอบคลุมของการจับคู่แบบลบและแทรกสูงขึ้นกว่าการทดลองแบบที่ 2 แต่ความแม่นยำต่ำลงจึงทำให้ความถูกต้องในภาพรวมมีค่าลดลง ในขณะเดียวกันการปรับค่าถ่วงน้ำหนักของแต่ละข้อสนเทศจะทำให้ความถูกต้องลดลงเล็กน้อยคิดเป็นประมาณ 1.5% (ดูที่ผลการทดลองที่ 2 และ 3) และโปรแกรมจะสกัดคู่ประโยคที่จับคู่กันแบบแทรกและแบบลบน้อยเกินไป

โปรแกรมจะสกัดคู่ประโยคได้ผลความถูกต้องมากที่สุดเมื่อปรับค่าถ่วงน้ำหนักพร้อมกับการปรับแต่งฟังก์ชั่นค่าใช้จ่าย (ดูที่ผลการทดลองที่ 6) แต่ความถูกต้องของการสกัดคู่ประโยคก็ยังไม่สูงนัก

ความถูกต้องของการสกัดคู่ประโยคที่จับคู่กันแบบแทนที่ (1-1) จะสูงที่สุด ในขณะที่ความถูกต้องของการสกัดคู่ประโยคแบบตัดทอน (many-1) จะน้อยลงตามจำนวนที่ปรากฏในชุดข้อมูล เนื่องจากประโยคที่จับคู่กันแบบแทนที่เป็นข้อมูลส่วนใหญ่ในชุดข้อมูลทั้งหมด ดังนั้นค่าของตัวแปรสำหรับการจับคู่การแปลระดับประโยคซึ่งได้มาจากการคำนวณขึ้นจากข้อมูลตัวอย่างในชุดข้อมูลฝึกฝน จึงมีแนวโน้มไปในทางที่จะอธิบายการจับคู่แบบแทนที่ได้ดีกว่าแบบอื่น

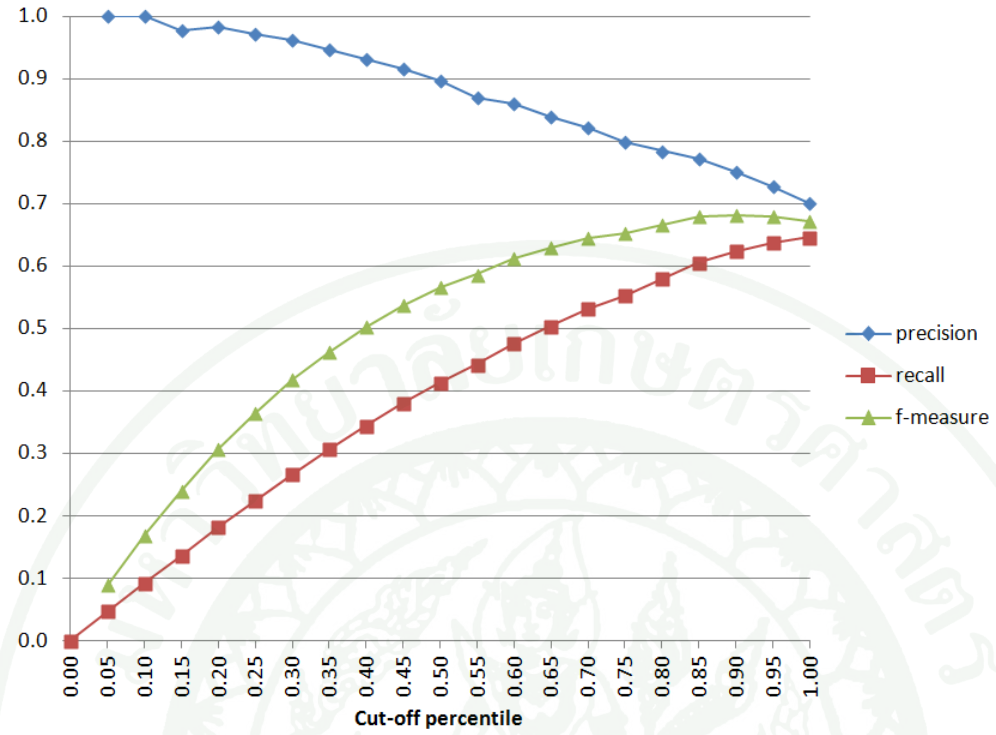
จากการทดลองเพิ่มเติมโดยการกรองคู่ประโยคที่สกัดได้ตามค่าลำดับเปอร์เซ็นไทล์ (Percentile rank) พบว่าสามารถทำให้ความแม่นยำเพิ่มขึ้น แต่ในขณะเดียวกันความครอบคลุมจะลดลง และค่าขีดแบ่งที่ให้ค่าคะแนนเอฟที่ดีที่สุดคือค่าลำดับเปอร์เซ็นไทล์ที่ 0.9 หรือเลือกเอาเฉพาะ 90% ของคำตอบที่ดีที่สุด ซึ่งได้ค่าความแม่นยำ ค่าความครอบคลุม และค่าคะแนนเอฟเป็น 0.751, 0.624 และ 0.681 ตามลำดับ (ดูตารางที่ 42 และแผนภูมิในภาพที่ 12) และถ้าไม่นำการจับคู่ผลลัพธ์ที่เป็นแบบแทรกและแบบลบมาคิดด้วยแล้วจะได้ค่าความแม่นยำ ค่าความครอบคลุม และค่าคะแนนเอฟเป็น 0.778, 0.724 และ 0.750 ตามลำดับ (ดูตารางที่ 43 และแผนภูมิในภาพที่ 13)

ตารางที่ 42 ความถูกต้องหลังการกรองตามค่าใช้จ่าย (คิดแบบแทรกและแบบลบด้วย)

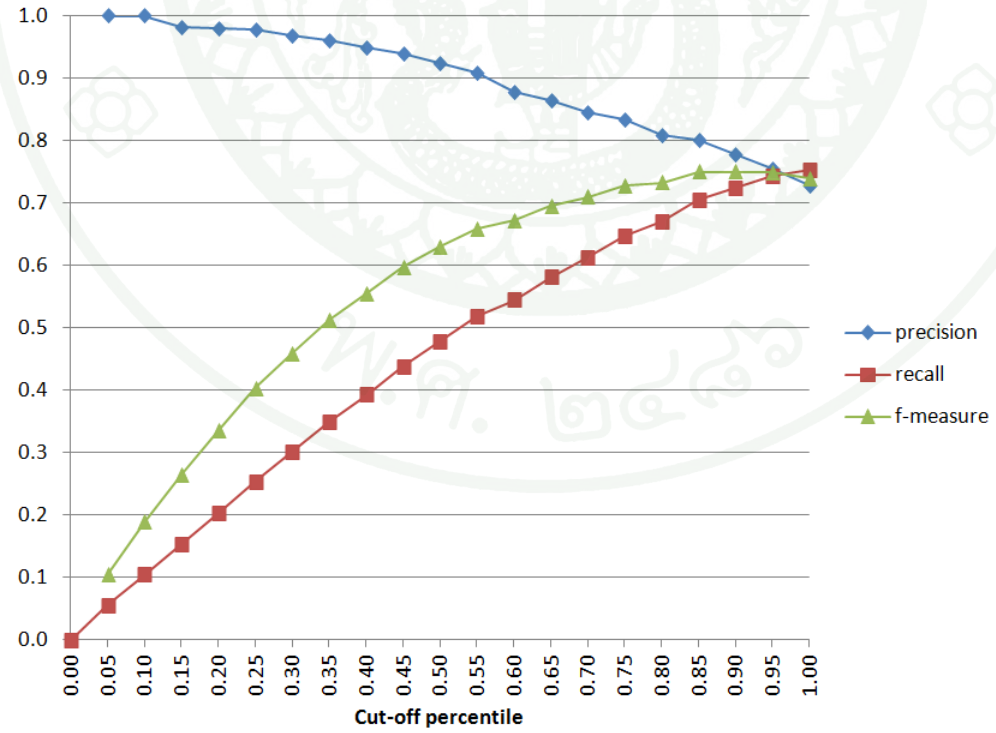
ระดับเปอร์เซ็นไทล์	ค่าความแม่นยำ	ค่าความครอบคลุม	ค่าคะแนนเอฟ
0.05	1.0000	0.0468	0.0893
0.10	1.0000	0.0925	0.1693
0.15	0.9771	0.1360	0.2388
0.20	0.9828	0.1817	0.3067
0.25	0.9724	0.2242	0.3644
0.30	0.9617	0.2667	0.4176
0.35	0.9474	0.3061	0.4627
0.40	0.9310	0.3443	0.5027
0.45	0.9156	0.3804	0.5375
0.50	0.8963	0.4134	0.5658
0.55	0.8703	0.4421	0.5863
0.60	0.8599	0.4761	0.6129
0.65	0.8389	0.5037	0.6295
0.70	0.8224	0.5313	0.6456
0.75	0.7988	0.5526	0.6533
0.80	0.7842	0.5792	0.6663
0.85	0.7724	0.6057	0.6790
0.90	0.7506	0.6238	0.6814
0.95	0.7273	0.6376	0.6795
1.00	0.7005	0.6461	0.6722

ตารางที่ 43 ความถูกต้องหลังการกรองตามค่าใช้จ่าย (ไม่คิดแบบแทรกและแบบลบ)

ระดับเปอร์เซ็นไทล์	ค่าความแม่นยำ	ค่าความครอบคลุม	ค่าคะแนนเอฟ
0.05	1.0000	0.0549	0.1040
0.10	1.0000	0.1041	0.1885
0.15	0.9820	0.1533	0.2652
0.20	0.9796	0.2025	0.3357
0.25	0.9783	0.2532	0.4022
0.30	0.9683	0.3010	0.4592
0.35	0.9612	0.3488	0.5119
0.40	0.9490	0.3924	0.5552
0.45	0.9396	0.4374	0.5969
0.50	0.9239	0.4782	0.6302
0.55	0.9086	0.5176	0.6595
0.60	0.8776	0.5443	0.6719
0.65	0.8640	0.5809	0.6947
0.70	0.8447	0.6118	0.7096
0.75	0.8333	0.6470	0.7284
0.80	0.8095	0.6695	0.7329
0.85	0.8016	0.7046	0.7500
0.90	0.7779	0.7243	0.7502
0.95	0.7554	0.7426	0.7489
1.00	0.7279	0.7525	0.7400

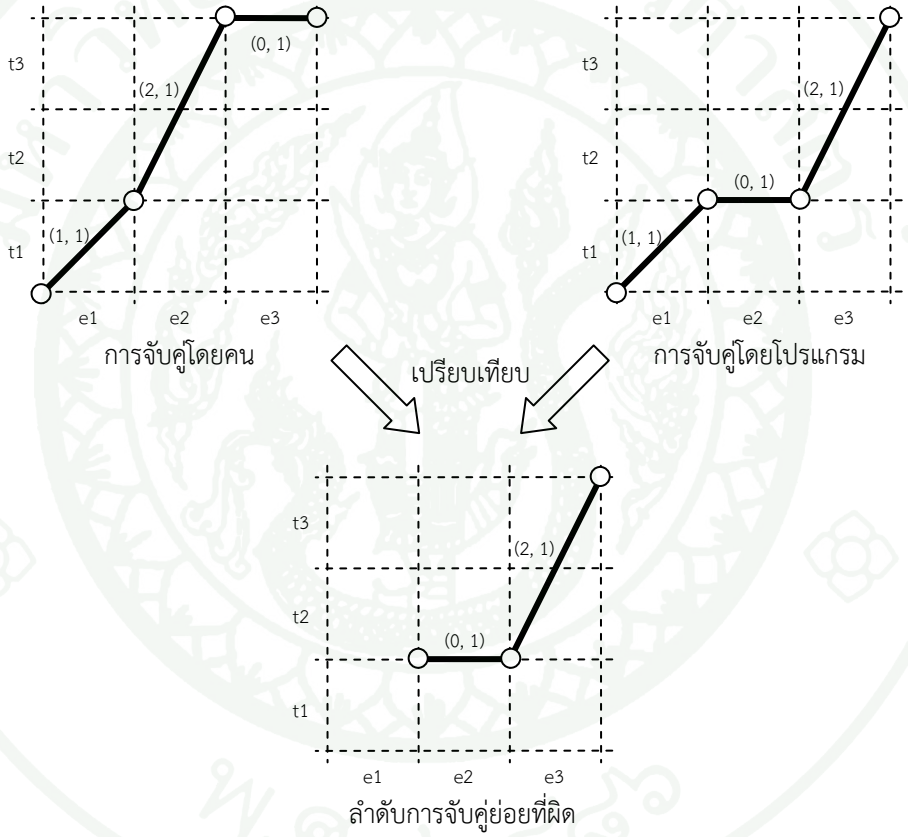


ภาพที่ 12 ความถูกต้องหลังการกรองตามค่าใช้จ่าย (คิดแบบแทรกและแบบลบด้วย)



ภาพที่ 13 ความถูกต้องหลังการกรองตามค่าใช้จ่าย (ไม่คิดแบบแทรกและแบบลบ)

เนื่องจากการจับคู่ที่ผิดหนึ่งจุดจะทำให้การจับคู่ที่ตามมาผิดไปด้วยเสมอ (ดูภาพที่ 14) ดังนั้น งานวิจัยนี้จึงสำรวจผลการจับคู่เพิ่มเติมโดยพิจารณาเป็นลำดับการจับคู่ย่อยที่ผิดพลาด (Sub-solution error) พบว่าลำดับการจับคู่ย่อยที่ผิดพลาดในการทดลองแบบที่ 6 มีจำนวนทั้งหมด 119 ลำดับย่อย ในจำนวนนี้พบว่ามี 83 ลำดับย่อยที่เป็นจุดผิดพลาดเพราะมีรูปแบบการจับคู่แบบลบอยู่ภายใน หรือต้องตัดประโยคภาษาไทยบางประโยคออกถึงจะถูกต้อง และมี 12 ลำดับย่อยที่ผิดเพราะมีรูปแบบการจับคู่แบบแทรกอยู่ภายใน จากการสำรวจนี้แสดงให้เห็นว่าการจับคู่แบบลบและแบบแทรกที่เกิดจากการแปลเอาความยังมีผลต่อความผิดพลาดของการสกัดประโยคคู่ขนานมาก



ภาพที่ 14 ตัวอย่างลำดับการจับคู่ย่อยที่ผิด

สรุปและข้อเสนอแนะ

ในงานวิจัยนี้ให้ความสนใจกับการจับคู่การแปลระดับประโยคจากคู่เอกสารภาษาไทย-อังกฤษ วิธีการที่ใช้ในงานวิจัยนี้เป็นการใช้ฟังก์ชั่นค่าใช้จ่ายแบบถ่วงน้ำหนักซึ่งคำนวณจากข้อสนเทศผสมที่คิดจากรูปแบบการจับคู่การแปล จากอัตราส่วนความยาวประโยคในหน่วยจำนวนอักขระของประโยคภาษาอังกฤษต่อภาษาไทย และจากค่าความน่าจะเป็นการแปล โดยตัวแปรและค่าถ่วงน้ำหนักของแต่ละฟังก์ชั่นค่าใช้จ่ายจะถูกคำนวณจากคลังประโยคคู่ขนานตั้งต้นโดยอัตโนมัติ และเพิ่มการคิดค่าใช้จ่ายแบบฮิวริสติกสำหรับการจับคู่การแปลแบบแทรกและแบบลบโดยเฉพาะ

งานวิจัยนี้ศึกษาและเปรียบเทียบการจับคู่การแปลระดับประโยคระหว่างเอกสารที่แปลโดยอรรถเกี่ยวกับการท่องเที่ยวขนาด 941 คู่ประโยค กับรัฐธรรมนูญฉบับปี พ.ศ.2550 ซึ่งเป็นเอกสารที่ตามตัวอักษรและขนาด 1,632 คู่ประโยค พบว่าเนื้อความในบทความท่องเที่ยวมีโอกาสที่เนื้อความจะถูกตัดทอนหรือเปลี่ยนแปลง ซึ่งเนื้อหาส่วนที่ถูกเปลี่ยนแปลงไปนี้จะทำให้เกิดการจับคู่แบบแทรกและแบบลบจำนวนมากกว่ารัฐธรรมนูญ และจากผลการทดลองพบว่าการจับคู่แบบแทรกและแบบลบนี้มีผลกระทบกับโปรแกรมจับคู่การแปลระดับประโยคมาก

งานวิจัยนี้ทำการทดลองไขว้แบบ 5 ทบ ประโยคคู่ขนานที่สกัดได้จากโปรแกรมจะถือว่าถูกต้องก็ต่อเมื่อตรงกับประโยคคู่ขนานที่ทำโดยคนทั้งตำแหน่งและจำนวนประโยคในแต่ละภาษา จากผลการทดลองพบว่าการจับคู่การแปลระดับประโยคจากรัฐธรรมนูญ ได้ค่าความแม่นยำ (Precision) และค่าความครอบคลุม (Recall) เป็น 0.998 ซึ่งสูงกว่าความถูกต้องของการจับคู่การแปลระดับประโยคจากบทความท่องเที่ยว ซึ่งได้ค่าความแม่นยำ ค่าความครอบคลุม และค่าคะแนนเอฟ (F score) ที่ดีที่สุดเป็น 0.728, 0.752 และ 0.740 ตามลำดับ

ขอบเขตของงานวิจัยนี้ จะเริ่มจากการรับข้อมูลป้อนเข้าที่มีการกำกับขอบเขตประโยค หรือจุดที่อาจเป็นจุดแบ่งขอบเขตประโยคมาแล้ว ทำให้เป็นไปได้ยากที่จะประยุกต์วิธีการที่เสนอนี้กับข้อมูลที่มีขนาดใหญ่ขึ้น ดังนั้นในอนาคตจึงควรมีการศึกษาและทดลองการจับคู่การแปลระดับประโยคกับโปรแกรมกำกับขอบเขตประโยคอัตโนมัติ หรือวิธีการกำกับจุดที่อาจเป็นจุดแบ่งประโยคแบบอื่น

จากการวิเคราะห์ผลการทดลอง พบว่าลำดับการจับคู่การแปลย่อยส่วนใหญ่มีความผิดพลาดเพราะลำดับย่อยนั้นมีการจับคู่แบบแทรกหรือแบบลบอยู่ภายใน สิ่งที่ควรพัฒนาต่อในอนาคตคือการระบุส่วนย่อยที่ไม่ถูกแปล ซึ่งอาจเป็นการประมวลผลภายหลังจากการจับคู่การแปลระดับประโยค

เอกสารและสิ่งอ้างอิง

พรรณา สวาสดิ์. 1991. **วิธีอ่านและแปลข่าวหนังสือพิมพ์ภาษาอังกฤษ (News reading & Translation)**, โอเดียนศโตร์, กรุงเทพฯ.

เมธี วัฒนะเมธานนท์. 2549. **การรู้จำความสัมพันธ์ปริจเฉทในเอกสารภาษาไทยโดยใช้โมเดลการคัดแยกแบบเนอีฟเบย์**. วิทยานิพนธ์ปริญญาโท, มหาเกษตรศาสตร์.

Al-Adhaileh, M. H. and T. Enya Kong 1999. Example-Based Machine Translation Based on the Synchronous SSTC. In **Machine Translation Summit VII**. Singapore.

Al-Adhaileh, M. H., T. Enya Kong, *et al.* 2001. **Malay-English Bitext Mapping and Alignment Using SIMR/GSA Algorithms**. In Selangor Darul Ehsan, Malaysia.

Arnold, D. 1986. Eurotra: A European perspective on MT. In **Proceedings of the IEEE 74(7)**.

Aroonmanakun, W. 2007. Thoughts on Word and Sentence Segmentation in Thai. In **Seventh Symposium on Natural Language Processing**. Pattaya, Thailand.

Bowker and J. Pearson 2002. **Working with Specialized Language: A Practical Guide to Using Corpora**, Taylor & Francis Group.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer 1991. Aligning sentences in parallel corpora. In **Proceedings of the 29th annual meeting on Association for Computational Linguistics**. Berkeley, California.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer 1993. The mathematics of statistical machine translation: parameter estimation. **Comput. Linguist. 19**(2).

Chen, S. F. 1993. Aligning sentences in bilingual corpora using lexical information. In **Proceedings of the 31st annual meeting on Association for Computational Linguistics**. Columbus, Ohio.

Chuang, T. C. and K. C. Yeh. 2005. Aligning Parallel Bilingual Corpora Statistically with Punctuation Criteria. In **Computational Linguistics and Chinese Language Processing**.

Fung, P. and P. Cheung. 2004. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)**. Barcelona, Spain.

Gale, W. A. and K. W. Church. 1993. **A program for aligning sentences in bilingual corpora.** Comput. Linguist. **19**(1).

Hutchins, J. 2006. Machine translation: history of research and use. **Encyclopedia of Languages and Linguistics. 2nd edition**. Elsevier Ltd.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In **Machine Translation Summit X**. Phuket, Thailand.

Nattapol Kritsuthikul, Arit Thammano, and Thepchai Supnithi 2006. English-Thai Example-Based Machine Translation using n-gram model. In **The 2006 IEEE International Conference on Systems Man and Cybernetics 2006 (SMC06)**. Taipei.

Li Peng, Sun Maosong, and Xue Ping 2010. Fast-Champollion: A Fast and Robust Sentence Alignment Algorithm. In **Coling 2010: Poster Volume**. Beijing, China.

Loper, E. and S. Bird. 2002. NLTK: the Natural Language Toolkit. In **Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1**. Philadelphia, Pennsylvania.

Ma, X. 2006. **Champollion: A Robust Parallel Text Sentence Aligner**. In **The fifth international conference on Language Resources and Evaluation (LREC 2006)**. Genoa, Italy.

Mamitimin, S. and M. Hou. 2009. Chinese-Uyghur sentence alignment: an approach based on anchor sentences. In **Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora**. Suntec, Singapore.

Melamed, I. D. 1999. Bitext maps and alignment via pattern recognition. **Comput. Linguist. 25**(1): 107-130.

Modhiran, T., K. Kosawat, *et al.* 2005. PARSIT: Online Thai-English Machine Translation. In **Machine Translation Summit X**. Phuket, Thailand.

Moe, L. 2008. **AUTOMATIC BITEXT ALIGNMENT FOR SOUTHEAST ASIAN LANGUAGES**. M.S. Thesis, Asian Institute of Technology.

Moore, R. C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In **Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users**. Tiburon, CA, USA.

Munteanu, D. S. and D. Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. **Comput. Linguist. 31**(4): 477-504.

Nagao, M. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In **Proc. of the international NATO symposium on Artificial and human intelligence**. Lyon, France.

Och, F. J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. **Comput. Linguist. 29**(1).

Ortiz-martínez , D., I. García-varea, *et al.* 2005. Thot: A toolkit to train phrase-based statistical translation models. In **The Tenth Machine Translation**. Phuket, Thailand.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu 2002. BLEU: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. Philadelphia, Pennsylvania.

Phillips, A. B. 2011. Cunei: open-source machine translation with relevance-based models of each translation instance. **Machine Translation 25**(2): 161-177.

Sato, S. and M. Nagao. 1990. Toward memory-based translation. In **Proceedings of the 13th conference on Computational linguistics - Volume 3**. Helsinki, Finland.

Slayden Glenn, Hwang Mei-Yuh, and Schwartz Lee 2010. Thai Sentence-Breaking for Large-Scale SMT. In **Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP)**. Beijing, China.

Sudprasert, S. and A. Kawtrakul (2003). Thai Word Segmentation based on Global and Local Unsupervised Learning. In **The 7th National Computer Science and Engineering Conference (NCSEC 2003)**. Chonburi, Thailand.

SYSTRAN Software, I. 2012. "SYSTRAN - Online translation software and tools." 2012, from http://www.systransoft.com.

Tannin, N., K. Chancharoen, *et al.* 1998. Alignment for Thai-English sentence. In **The 1998 IEEE Asia-Pacific Conference on Circuits and Systems**. Chiangmai, Thailand.

Tiedemann, J. 2011. **Bitext Alignment**. Morgan &cLaypool publishers.

Utiyama, M. and H. Isahara. 2007. A Japanese-English Patent Parallel Corpus. In **Machine Translation Summit XI**. Copenhagen, Denmark.

Varga Dániel, Németh László, Halácsy Péter, Kornai András, Trón Viktor, and Nagy Viktor 2005. Parallel corpora for medium density languages. In **Recent Advances in Natural Language Processing (RANLP 2005)**. Borovets, Bulgaria.

Wu, D. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In **Proceedings of the 32nd annual meeting on Association for Computational Linguistics**. Las Cruces, New Mexico.

Yamada, K. and K. Knight (2002). A decoder for syntax-based statistical MT. In **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. Philadelphia, Pennsylvania.

ประวัติการศึกษาและการทำงาน

ชื่อ	นายชานน อ่อนมั่น
เกิดวันที่	25 ธันวาคม 2530
สถานที่เกิด	อำเภอเมือง จังหวัดลำปาง
ประวัติการศึกษา	วศ.บ. (คอมพิวเตอร์) มหาวิทยาลัยเกษตรศาสตร์
ทุนการศึกษาที่ได้รับ	โครงการทุนสถาบันบัณฑิตวิทยาศาสตร์และเทคโนโลยีไทย สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ (พ.ศ. 2553)