ระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่าง โดยใช้ตัวแบบเอ็นแกรม

ENGLISH-THAI EXAMPLE-BASED MACHINE TRANSLATION USING N-GRAM MODEL

ณัฐพล กฤษสุทธิกุล

NATTAPOL KRITSUTHIKUL

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2549

ENGLISH-THAI EXAMPLE-BASED MACHINE TRANSLATION USING N-GRAM MODEL

NATTAPOL KRITSUTHIKUL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION TECHONOLOGY

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2006

หัวข้อวิทยานิพนธ์	**ระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่าง โดยใช้ตัวแบบเอ็นแกรม**
นักศึกษา	**นายณัฐพล กฤษสุทธิกุล**
รหัสนักศึกษา	44067499
ปริญญา	**วิทยาศาสตรมหาบัณฑิต**
สาขาวิชา	**เทคโนโลยีสารสนเทศ**
พ.ศ.	2549
อาจารย์ที่ปรึกษาวิทยานิพนธ์	**รศ.ดร. อาริต ธรรมโน**

บทคัดย่อ

การแลกเปลี่ยนข้อมูลสารสนเทศระหว่างประเทศเป็นภารกิจอันสำคัญของสังคมแห่งข้อมูลข่าวสาร การแปลภาษาด้วยเครื่องคอมพิวเตอร์ถูกนำมาใช้เพื่อลดข้อจำกัดทางกำแพงภาษาออกไป เนื่องด้วยความสามารถในการประมวลผลของคอมพิวเตอร์ที่ทำงานได้รวดเร็วและมีประสิทธิภาพมากขึ้น ทำให้เทคโนโลยีการประมวลผลภาษาธรรมชาติที่ใช้คลังข้อมูลขนาดใหญ่กลายเป็นแนวความคิดพื้นฐานของการพัฒนาซอฟต์แวร์ที่ใช้ข้อมูลปริมาณมหาศาล วิทยานิพนธ์ฉบับนี้นำเสนอระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรม และนำเสนอทั้งข้อดีและข้อเสียของวิธีการดังกล่าวที่ได้ถูกนำมาอภิปราย

วิทยานิพนธ์ฉบับนี้นำเสนอระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรม โดยระบบดังกล่าวมีส่วนประกอบสำคัญสองส่วนคือส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรม และส่วนโปรแกรมการก่อกำเนิดแบบเอ็นแกรม

ตัวแบบเอ็นแกรมสามารถแก้ปัญหาการเรียงลำดับคำที่ไม่เหมือนกันของภาษาต้นฉบับและภาษาเป้าหมาย สามารถแก้ปัญหาการเลือกคำที่เหมาะสมตามบริบท อีกทั้งยังสามารถแก้ปัญหาการแปลวลีได้

จากผลการทดลองสรุปได้ว่าวิธีการดังกล่าวสามารถแก้ปัญหาพื้นฐานของระบบแปลภาษาด้วยเครื่อง ดังนี้ การเข้าคู่แบบแม่นตรง 2% แก้ปัญหาการเรียงลำดับคำได้ถูกต้องโดยเฉลี่ย 27% แก้ปัญหาการเลือกใช้คำให้เหมาะสมตามบริบทได้ถูกต้องโดยเฉลี่ย 57% และสามารถแก้ปัญหาการแปลวลีได้ถูกต้องโดยเฉลี่ย 51% ซึ่งผลลัพธ์ดังกล่าวดีกว่าระบบแปลภาษา "ภาษิต" ในกรณีต่อไปนี้ การเข้าคู่แบบแม่นตรง ปัญหาการเลือกใช้คำให้เหมาะสมตามบริบท และปัญหาการแปลวลี แต่ยังด้อยกว่าในการแก้ปัญหาการเรียงลำดับคำ

Thesis Title	English-Thai Example-Based Machine Translation using *n*-gram Model
Student	Nattapol KRITSUTHIKUL
Student ID.	44067499
Degree	Master of Science
Program	Information Technology
Year	2006
Thesis Advisor	Assoc.Prof. Arit THAMMANO

ABSTRACT

The necessity on exchanging information among countries becomes a major task in information based society. Machine translation is an application that enables users communicate each other without language barrier. With the great support on computer's efficiency, corpus based technology becomes a fundamental concept for developing software based on a large amount of data. We introduce the first example based English to Thai machine translation using *n*-gram model and implement the system. Some advantages and disadvantages of this method are discussed.

This thesis presents an English to Thai example-based machine translation using *n*-gram model system. This system has 2 important parts: *n*-gram analysis and *n*-gram generation modules.

The *n*-gram model used in this thesis is not only able to solve the different word-ordering problem in source and target language, but also able to select the word form the translatable word list that suits its context. This model also has a potential to translate phrases.

From the experiment, the system correctly solves the exact matching by approximately 2%, word-ordering problem by approximately 27%, select appropriate translation with respect to its content by approximately 57%, and correctly translates phrases by approximately 51%. The result of exact matching, word selection of its content, and phrases translation are better than Parsit's translation result, but Parsit can do better in ordering word.

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงด้วยดีหากปราศจากแรงผลักดัน และคำแนะนำที่มีประโยชน์ของ รศ.ดร. อาริต ธรรมโน อาจารย์ผู้ควบคุมวิทยานิพนธ์ ข้าพเจ้าขอกราบขอบพระคุณอย่างสูง

ข้าพเจ้าขอขอบคุณสำหรับกำลังใจ คำแนะนำ และประสบการณ์ที่ดีจากพี่ๆ และเพื่อนๆ ชาว NECTEC ประกอบไปด้วย พี่นุ้ย พี่บอมบ์ พี่เก่ง โป้ง อาร์ม โบนัส ที่ให้ให้คำแนะนำในการทำวิทยานิพนธ์ ดร.สรรพฤทธิ์ มฤคทัต (ปืน) ที่ให้คำแนะนำในเรื่องสมการต่างๆ โดยเฉพาะคนที่ข้าพเจ้าต้องขอขอบคุณเป็นอย่างยิ่งคือ ดร.เทพชัย ทรัพย์นิธิหรือพี่เล็ก ที่ให้การช่วยเหลือและเป็นที่ปรึกษาให้ข้าพเจ้า

สุดท้ายนี้คุณค่าและประโยชน์อันพึงมีจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับผู้มีพระคุณทุกท่าน หากวิทยานิพนธ์ฉบับนี้มีข้อผิดพลาดประการใดข้าพเจ้าขอน้อมรับไว้เพียงผู้เดียว

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญภาพ

ภาพที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ภาษาอังกฤษถือได้ว่าเป็นภาษากลางที่ใช้ติดต่อสื่อสารทั่วโลก ทำให้เอกสารจำนวนมากที่เผยแพร่โดยทั่วไปใช้ภาษาอังกฤษ ขณะที่มีคนไทยจำนวนไม่น้อยที่ขาดความเข้าใจภาษาอังกฤษ จึงขาดโอกาสรับข้อมูลข่าวสารทำให้ขาดโอกาสในการเรียนรู้ไปด้วย เพื่อลดช่องว่างดังกล่าวให้ได้มากที่สุด ทำให้เกิดแนวความคิดที่จะทำให้คอมพิวเตอร์มีความสามารถในการสื่อสารด้วยภาษามนุษย์หรือก็คือภาษาธรรมชาติ (Natural Language) ซึ่งเป็นเป้าหมายหลักสำคัญเป้าหมายหนึ่งของวิชาการสาขาปัญญาประดิษฐ์ (Artificial Intelligence) "การประมวลผลภาษาธรรมชาติ" (Natural Language Processing) หรือเรียกโดยย่อว่า NLP

เครื่องคอมพิวเตอร์ในปัจจุบันจะคอยรับคำสั่งเพียงฝ่ายเดียว (One-way Communication) ไม่สามารถโต้ตอบกับมนุษย์ได้ ทำให้เกิดข้อจำกัดมากในการใช้งานคอมพิวเตอร์ ดังนั้นจำเป็นที่ต้องทำให้เครื่องคอมพิวเตอร์มีปัญญา (Intelligence) และความรู้ (Knowledge) เพื่อให้สามารถโต้ตอบกับมนุษย์ไม่ว่าจะเป็นทางตรงหรือทางอ้อม เพื่อให้บรรลุจุดประสงค์ดังกล่าว เครื่องคอมพิวเตอร์จำเป็นที่จะต้องเข้าใจภาษาที่มนุษย์ใช้ หรือที่เรียกกันแบบทางการว่าการเข้าใจภาษาธรรมชาติ (Natural Language Understanding) เครื่องคอมพิวเตอร์จำเป็นที่จะต้องมีฐานความรู้ (Knowledge) ซึ่งมีทั้งความรู้ทางภาษาที่เกี่ยวกับความหมายของคำ (Word Meaning) การแสดงความหมายจากกลุ่มคำและความรู้ที่เป็นภูมิหลัง (Background Knowledge) ในเนื้อหา (Context) สถานการณ์ (Situation) รวมถึงความเป็นมาของเนื้อหานั้นด้วย จนถึงปัจจุบันได้มีการนำคอมพิวเตอร์เข้ามาประยุกต์ใช้ในงานด้านต่างๆ อย่างแพร่หลาย เช่น นำไปใช้ด้านการคำนวณและคาดการณ์สภาวะที่ซับซ้อน (High Performance Computing) นำไปใช้เพื่อจำลองแบบเสมือนจริง (Simulation) นำไปใช้เพื่อการจัดเก็บข้อมูลและเรียกค้นคืนข้อมูล รวมถึงการนำไปใช้งานด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ซึ่งคือกระบวนการที่จะทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ได้ ตัวอย่างเช่น การแปลภาษาด้วยเครื่องคอมพิวเตอร์ (Machine Translation) [1,2,3], การผูกหรือการสร้างประโยคอัตโนมัติ (Text Generation) [4,5] การทำสรุปใจความสำคัญ การสังเคราะห์เสียงภาษาไทย (Thai Speech Synthesis) หรือการสืบค้นข้อความทั้งเอกสาร (Full Text Search) เป็นต้น

เมื่อเครื่องคอมพิวเตอร์มีความรู้ สามารถเข้าใจและโต้ตอบด้วยภาษามนุษย์ได้แล้วก็จะทำให้เราสามารถใช้ประโยชน์จากเครื่องคอมพิวเตอร์ได้โดยตรงมากยิ่งขึ้น โดยอาศัยคุณสมบัติเฉพาะตัวของเครื่องคอมพิวเตอร์ในด้านการประมวลผลข้อมูลได้ในปริมาณครั้งละมากๆ ด้วย

ความเร็วสูงและข้อมูลที่ใช้นั้นก็จะอยู่ในรูปของอิเล็กทรอนิกส์ซึ่งเป็นตัวกลางที่จัดการได้โดยง่ายในระบบต่างๆ ในปัจจุบัน

ระบบแปลภาษาเป็นงานวิจัยแขนงหนึ่งของการประมวลผลภาษาธรรมชาติ (NLP: Natural Language Processing) ซึ่งเป็นการผสมผสานระหว่างปัญญาประดิษฐ์ (Artificial Intelligent) และภาษาศาสตร์คำนวณ (Computation Linguistic) ระบบแปลภาษาจากภาษาอังกฤษเป็นภาษาไทยเป็นงานที่ยาก ซับซ้อน และมีความสำคัญอย่างยิ่ง ปัญหาพื้นฐานที่สำคัญในการแปลภาษาอังกฤษเป็นภาษาไทยมี 3 ประการได้แก่ ปัญหาการเรียงลำดับคำ (Word Ordering Problem) ปัญหาการเลือกคำที่เหมาะสมกับบริบท (Word Selection Problem) ปัญหาการแปลวลี (Phrasal Translation Problem) โดยปัญหาแต่ละประเภทมีความเป็นอิสระไม่ขึ้นต่อกัน

ปัญหาการเรียงลำดับคำ (Word Ordering Problem) คือ ลำดับของคำแปลที่ได้จะต้องถูกต้องเหมาะสมตามหลักไวยากรณ์และสามารถสื่อความหมายได้ตรงกับประโยคต้นฉบับ ตัวอย่างเช่น "I go to school" ควรแปลเป็น "ฉัน/ไป/โรงเรียน/" ไม่ใช่ "ฉัน/โรงเรียน/ไป/" (ผิดหลักไวยากรณ์) หรือ "โรงเรียน/ฉัน/ไป/" (ผิดความหมาย) หรือ "ไป/โรงเรียน/ฉัน/" (ผิดความหมาย) เป็นต้น

ปัญหาการเลือกคำที่เหมาะสมกับบริบท (Word Selection Problem) คือ คำแปลที่เหมาะสมของคำในสถานการณ์ที่แตกต่างกัน เช่น "general have gun" อาจแปลแบบคำต่อคำได้เป็น "ทั่วไป/มี/ ปืน/" ซึ่งคำว่า "general" ในที่นี้ควรแปลเป็น "นายพล/มี/ ปืน/" จึงจะเหมาะสมกว่า จากตัวอย่างนี้เห็นได้ชัดว่าคำแปลมีการเรียงลำดับคำถูกต้องแต่ไม่สามารถเลือกคำแปลที่เหมาะสมกับบริบทได้

ปัญหาการแปลวลี (Phrasal Translation Problem) คือ ความสามารถในการแปลวลีให้มีความหมายที่ถูกต้อง เช่น "he go back home" ควรแปลว่า "เขา/กลับไป/บ้าน/" ไม่ใช่ "เขา/ไป/กลับ/บ้าน/" จากตัวอย่างเห็นได้ชัดว่าปัญหานี้เกิดจากการไม่รู้จักวลี "go back" ไม่ได้เกี่ยวข้องกับปัญหาการเรียงลำดับของคำและไม่ได้เกี่ยวข้องกับปัญหาการเลือกคำที่เหมาะสมกับบริบท

วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรม กระบวนการแปลประกอบด้วย 2 ขั้นตอน ได้แก่ ส่วนวิเคราะห์เอ็นแกรมและส่วนการก่อกำเนิดเอ็นแกรม ตัวแบบเอ็นแกรมจะช่วยให้สามารถตรวจสอบการเรียงลำดับคำ การเลือกคำที่เหมาะสมกับบริบท และการแปลวลีของภาษาอังกฤษ

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

เพื่อลดข้อจำกัดและทลายกำแพงทางภาษา เปิดโอกาสให้คนไทยได้มีโอกาสรับรู้ข่าวสารที่เป็นภาษาต่างประเทศซึ่งมีอยู่เป็นจำนวนมาก อีกทั้งยังใช้เป็นบรรทัดฐานสำหรับการแปลภาษาด้วยเครื่องแบบอิงตัวอย่างสืบต่อไป

1.3 ขอบเขตของการศึกษา

1. นำเสนอต้นแบบระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรม (ขอเรียกโดยย่อว่า "ระบบ") โดยขอบเขต (domain) จะขึ้นกับคลังข้อความแบบคู่และคลังข้อความแบบเดี่ยวของแผนกเทคโนโลยีประมวลผลข้อความ ฝ่ายวิจัยและพัฒนาเทคโนโลยีสารสนเทศ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ประเทศไทย ซึ่งมีคลังข้อความดังกล่าว รายละเอียดโดยสังเขปดังนี้

- คลังข้อความแบบเดี่ยวสำหรับภาษาต้นทาง (ภาษาอังกฤษ) จำนวน 104,893 ประโยค หรือ 561,387 คำ
- คลังข้อความแบบเดี่ยวสำหรับภาษาปลายทาง (ภาษาไทย) จำนวน 152,570 ประโยค หรือ 10,164,366 คำ
- คลังข้อความแบบคู่สำหรับภาษาต้นทางและปลายทาง (ภาษาอังกฤษและภาษาไทย) จำนวน 100,804 คู่ประโยค
- พจนานุกรมอังกฤษ-ไทย Lexitron จำนวน 97,791 คำ
- พจนานุกรมไทย-อังกฤษ Lexitron จำนวน 104,892 คำ

2. ระบบจะเน้นแก้ปัญหาการแปลพื้นฐาน 3 ประการของในกรณีเป็นประโยคบอกเล่า อันได้แก่ ปัญหาการเรียงลำดับคำ ปัญหาการเลือกคำที่เหมาะสมกับบริบท และปัญหาการแปลวลี สำหรับปัญหาในกรณีอื่นไม่ขอกล่าวถึง

3. ระบบสามารถแปลได้เฉพาะคำที่รู้จักจากคลังข้อความเท่านั้น ไม่สามารถรู้จำและไม่สามารถแปลคำระบุชื่อเฉพาะ (Name Entity) ได้

4. ระบบจะรับข้อมูลจากผู้ใช้ผ่านทางแป้นพิมพ์ และแสดงผลออกทางจอภาพ

5. ระบบจะทำงานแบบ off-line

6. ระบบจะถูกพัฒนาโดยใช้ Java 2 Platform, Standard Edition (J2SE) 5.0

1.4 ขั้นตอนของการศึกษา

1. ศึกษางานวิจัยที่เกี่ยวกับระบบแปลภาษาที่มีอยู่แล้ว
2. ศึกษาอัลกอริทึมการเรียนรู้ในแบบต่างๆ
3. เตรียมข้อมูลประโยคและคลังข้อความที่ต้องการ
4. แปลงข้อมูลให้อยู่ในรูปที่พร้อมนำเข้าเรียนรู้สำหรับแต่ละอัลกอริทึม
5. ทดสอบผลการแปลโดยเปรียบเทียบกับ ระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่อง "ภาษิต" โดยทดสอบด้วยวิธีการดังนี้

- ทดสอบจำนวนที่สามารถเข้าคู่แบบแม่นตรง (Exact Matching)

- ทดสอบการประเมินค่าผลการแปลแบบอัตโนมัติของ BLEU-4
- ทดสอบความถูกต้องของการเรียงลำดับคำ
- ทดสอบความถูกต้องของการเลือกคำให้เหมาะสมตามบริบท
- ทดสอบความถูกต้องของการแปลวลี

6. สรุปผลการทดลองพร้อมจัดทำบทความตีพิมพ์ และวิทยานิพนธ์

1.5 รายละเอียดในแต่ละบท

ในวิทยานิพนธ์ฉบับนี้แบ่งเนื้อหาการนำเสนอออกเป็น 5 บท ดังนี้

บทที่ 1 กล่าวถึงความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์และขอบเขตของงานวิจัย

บทที่ 2 กล่าวถึงระบบการแปลภาษาด้วยเครื่อง จุดแตกต่างเมื่อเทียบกับแนวคิดอื่น งานวิจัยที่เกี่ยวข้อง

บทที่ 3 กล่าวถึงระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรม

บทที่ 4 การประเมินค่าความถูกต้องของผลการแปล

บทที่ 5 กล่าวถึงการสรุป วิเคราะห์ผลการทดลอง รวมทั้งข้อเสนอแนะ และแนวทางการทำวิจัยต่อ

บทที่ 2
การแปลภาษาด้วยเครื่องคอมพิวเตอร์

การทำวิจัยและพัฒนาการแปลภาษาด้วยเครื่องนั้นเป็นงานแขนงหนึ่งในศาสตร์แห่งการประมวลผลภาษาธรรมชาติ (Natural Language Processing หรือ NLP) เพื่อให้เข้าใจถึงความเป็นมาของระบบแปลภาษาด้วยเครื่องแบบต่างๆ ในบทนี้จึงขออธิบายถึงระบบแปลภาษาด้วยเครื่องแบบต่างๆ รวมถึงข้อดีและข้อเสียของแต่ละระบบ ในท้ายบทจะอธิบายถึง "ภาษิต" ซึ่งเป็นระบบแปลภาษาด้วยเครื่องสำหรับภาษาอังกฤษ-ไทย ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

อย่างไรก็ดี ในปัจจุบันการแปลภาษาด้วยเครื่องนั้นไม่สามารถทำการแปลข้อความหรือบทความให้เกิดประโยคที่มีความไพเราะและสละสลวยได้เหมือนกับการแปลของของมนุษย์ทั่วไป เช่นการแปลกาพย์ โคลงกลอนต่างๆ เนื่องจากการแปลภาษาด้วยเครื่อง ยังมีข้อจำกัดในการแปลมาก ไม่สามารถที่จะทำให้เครื่องมีความรู้ได้เท่ากับมนุษย์จริงๆ ทำให้การแปลภาษาด้วยเครื่องไม่สามารถทดแทนนักแปลได้ แต่การแปลภาษาด้วยเครื่องนั้นจะสามารถช่วยแปลข้อความหรือบทความที่ไม่ต้องการความไพเราะและสละสลวยได้เป็นอย่างดี และจะช่วยทุ่นแรงนักแปลได้มาก

2.1 ระบบการแปลโดยตรง (Direct Machine Translation Strategy)

ระบบการแปลโดยตรงคือระบบการแปลที่ได้รับการออกแบบให้ใช้กับการแปลคู่ภาษาเฉพาะคู่ใดคู่หนึ่ง เป็นลักษณะการแปลแบบง่ายๆ ไม่มีการใช้ทฤษฎีทางภาษาศาสตร์หรือหลักการทางวจีวิภาค (POS : Parts-of-Speech) ระบบนี้จะขึ้นอยู่กับการพัฒนาพจนานุกรมที่สมบูรณ์ที่สุด การวิเคราะห์หน่วยคำ และโปรแกรมการผลิตข้อความเพื่อแปลความหมายแบบคำต่อคำ วลีต่อวลี จากภาษาต้นทางสู่ภาษาปลายทางอย่างสมเหตุสมผล

ระบบการแปลโดยตรงนั้นมีข้อเสียอยู่ที่เป็นระบบที่แปลภาษาได้ทีละคู่ของภาษา กล่าวคือ ถ้าต้องการแปลภาษาแต่ละคู่จะต้องมีการวิเคราะห์คู่ภาษานั้นๆ ทุกครั้ง ลักษณะการแปลแบบนี้ทำให้กระบวนการแปลมีความยุ่งยากมาก สมมติว่าถ้าต้องการแปลภาษา 10 คู่ ภาษาก็ต้องทำการวิเคราะห์และสังเคราะห์คู่ภาษานั้น 10 ครั้ง ทั้งที่บางครั้งภาษาต้นทางนั้น อาจเป็นภาษาเดียวกันก็ได้ ตัวอย่างของระบบนี้คือ ระบบซิสทราน (Systran) ใช้แปลเอกสารจากภาษารัสเซียเป็นภาษาอังกฤษ

2.2 ระบบการแปลแบบเปลี่ยน (Transfer Machine Translation Strategy)

เป็นระบบที่ถือว่าตัวแทนแสดงความหมายของไวยากรณ์ของภาษาต้นทางและภาษาปลายทางนั้นมีลักษณะที่แตกต่างกัน จะต้องมีช่วงเชื่อมต่อที่เทียบตัวแทนแสดงความหมายที่เป็นลักษณะเฉพาะของภาษาหนึ่ง เรียกว่าการเปลี่ยน จากนั้นก็สร้างภาษาปลายทางขึ้น ตัวอย่างเช่น ลักษณะประโยคภาษาอังกฤษ "It is a please to be here." จะต้องเปลี่ยนเป็นประโยคโครงสร้างของภาษาไทยคือ "ยินดีที่ได้มาอยู่ที่นี่" ซึ่งถ้าเราใช้ระบบการแปลโดยตรงจะได้ประโยคว่า "มันเป็นความยินดีมาอยู่ที่นี่" [6] ระบบนี้มีขั้นตอนการทำงาน 3 ขั้นตอน คือ

2.2.1 การวิเคราะห์ภาษาต้นทาง (Source Language Analysis)

เป็นการวิเคราะห์ประโยคในภาษาต้นทางโดยใช้หลักไวยากรณ์โครงสร้างของภาษา และพจนานุกรมของภาษาต้นทาง (Source Language Dictionary)

2.2.2 การเปลี่ยน (Transfer)

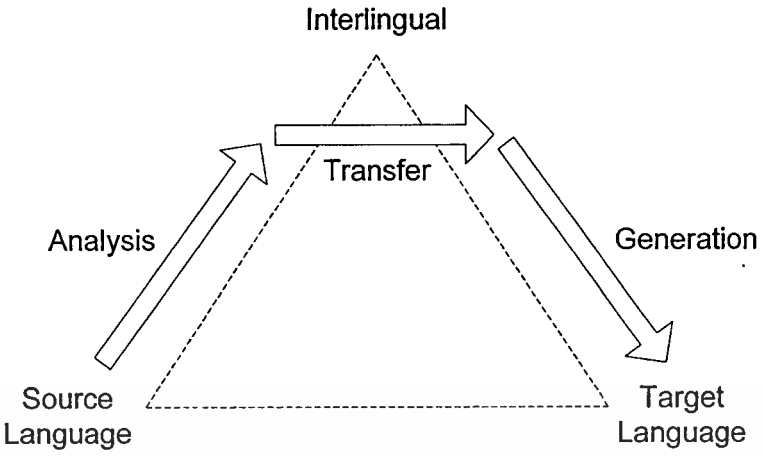
เป็นการเปลี่ยนคลังคำศัพท์ (Lexicon) และโครงสร้างของภาษาต้นทางให้สอดคล้องกับคลังศัพท์ และโครงสร้างของภาษาปลายทาง โดยมีการใช้พจนานุกรมทวิภาษา (Bilingual Transfer Dictionary)

2.2.3 การสร้างภาษาปลายทาง (Target Language Generation)

ในขั้นนี้จะใช้พจนานุกรมของภาษาปลายทาง (Target Language Dictionary) เพื่อเป็นการสร้างภาษาปลายทางให้มีคุณสมบัติทางโครงสร้าง และความหมายของภาษาปลายทางอย่างแท้จริง ซึ่งอาจจะรวมถึงขั้นตอนการจัดลำดับคำ ตัวอย่างของระบบนี้คือระบบ เกต้า (GETA) หรือระบบซูซี่ (SUSY)

2.3 ระบบการแปลภาษาแบบการใช้ภาษากลาง (Interlingual Machine Translation strategy)

เป็นระบบการแปลที่พัฒนาจากระบบการแปลแบบการเปลี่ยนเพื่อให้มีลักษณะเป็นสากลมากขึ้นจนเป็นตัวแทนภาษาที่เป็นอิสระ (Language Independent Representation) ระบบการแปลภาษาแบบนี้จะแบ่งเป็น 2 ด้านคือด้านการวิเคราะห์ (Analysis) เป็นการวิเคราะห์ภาษาต้นทางสู่ภาษาที่เราสร้างขึ้นใหม่เรียกว่าภาษากลาง (Interlingua) และด้านการก่อกำเนิด (Generation) ในขั้นตอนนี้เราจะสร้างภาษาปลายทางจากภาษากลาง ระบบนี้เป็นระบบหลากภาษา (Multilingual) ซึ่งแทนระบบทวิภาษาในระบบการแปลแบบเปลี่ยน (transfer)



ภาพที่ 2.1 ระบบการแปลภาษาแบบการใช้ภาษากลาง

ที่กล่าวถึงไปแล้วเป็นแนวทางสำหรับการแปลภาษาในอดีต สำหรับแนวทางการพัฒนาระบบเครื่องแปลภาษาในปัจจุบันนั้นจะสรุปการแปลภาษาเป็น 2 แนวทาง คือแนวทางที่หนึ่งเป็นการพัฒนาปรับปรุงระบบซึ่งอาศัยกฎไวยากรณ์ (Rule-Based MT) ซึ่งมีการค้นคว้าวิจัยมานานแล้วให้ดียิ่งขึ้น กับอีกแนวทางหนึ่งเป็นการพัฒนาระบบซึ่งอาศัยฐานคลังข้อความ (Corpus-Based MT) มาช่วยในการแปล

2.4 การใช้กฎไวยากรณ์ช่วยในการแปล (Rule-Based MT)

2.4.1 Transfer-Based MT

เป็นวิธีที่ถูกใช้ในโครงการแปลภาษาที่สำคัญๆ ในยุคแรกๆ จากที่กล่าวไปข้างต้น โดยวิธีนี้มองว่า กระบวนการแปลประกอบไปด้วย 3 ขั้นตอนคือ การวิเคราะห์ไปเป็นรูปแสดงแทนของภาษาต้นทาง (Abstract Source Language Representation), การย้ายข้าง (Transfer) ไปเป็นรูปแสดงแทนของภาษาปลายทาง (Abstract Target Language Representation) และการผลิตหรือสังเคราะห์ไปเป็นข้อความของภาษาปลายทางถึงแม้ว่าทั้งสองโครงการข้างต้นจะปิดฉากลงไปเรียบร้อยแล้ว แต่ก็ได้มีโครงการใหม่ซึ่งได้พัฒนาต่อไปอีก ซึ่งมีเป้าหมายจะพัฒนาระบบช่วยเหลือของ MT สำหรับนักแปลที่มีลักษณะเป็น User-friendly

2.4.2 Interlingua-Based MT

วิธีนี้ต่างจาก Transfer-Based MT โดยมองกระบวนการการแปลว่าประกอบด้วย 2 ขั้นตอนคือ การวิเคราะห์ไปเป็นรูปแสดงแทนซึ่งไม่ขึ้นกับภาษา และการผลิตจากรูปแสดงแทนนั้นไปเป็นข้อความของภาษาปลายทาง โครงการของญี่ปุ่นและอเมริกาจำนวนมากก็ได้ใช้วิธีนี้ ข้อดีของวิธีนี้คือความง่ายในการขยายเพิ่มประเภทของภาษาต้นทาง และภาษาปลายทางเข้ากับระบบการแปลเมื่อเทียบกับ Transfer-Based MT แต่การกำหนดภาษากลางให้ครอบคลุมทุก

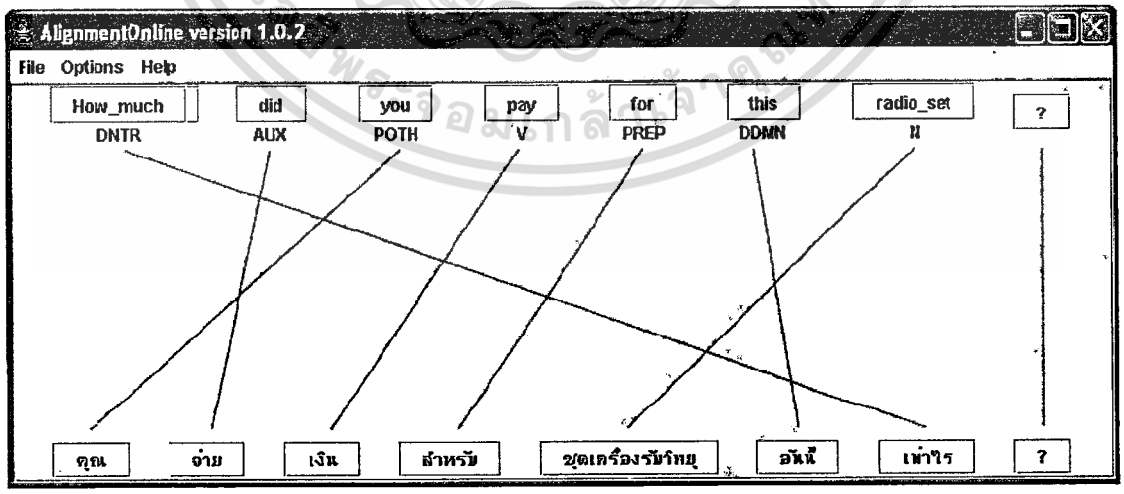
รูปแบบการใช้ภาษาของหลายๆ ภาษาก็เป็นปัญหาที่ยากสำหรับวิธีนี้ ในระยะหลังแนวโน้มของวิธีนี้จะเป็นการใช้ฐานความรู้เข้ามาช่วยในการแปล โดยมีความเชื่อว่าการแปลจะต้องใช้ความรู้มากกว่าความรู้ทางภาษาอย่างเดียว กล่าวคือ ต้องมีฐานความรู้เพื่อทำความเข้าใจ (Understanding) บริบทให้ได้ด้วย ซึ่งในส่วนนี้เองที่หากสามารถแก้ไขปัญหาพื้นฐานบางอย่าง เช่น การแก้ไขปัญหาความกำกวมของคำ (Word Sense Disambiguation) จะช่วยให้การแปลดีขึ้น

2.5 การใช้ฐานบทความช่วยในการแปล (Corpus-Based MT)

2.5.1 ระบบแปลภาษาด้วยเครื่องแบบอิงสถิติ (Statistical-Based MT)

การพัฒนาที่เด่นชัดที่สุดในช่วงปลายศตวรรษที่ 90 เห็นจะได้แก่ การนำวิธีการทางสถิติของ MT ในโครงการวิจัยของบริษัทไอบีเอ็ม ลักษณะที่สำคัญก็คือ การใช้วิธีการทางสถิติเพียงอย่างเดียวในการวิเคราะห์และการผลิต โดยทดลองกับ Corpus ขนาดใหญ่ของ The Canadian Hansard ซึ่งเป็นบันทึกการอภิปรายในสภาโดยจัดเก็บเป็นภาษาอังกฤษและฝรั่งเศส ผลการทดลองก่อให้เกิดการตื่นตัวในวงการระบบแปลภาษาด้วยเครื่องอย่างมาก เนื่องจากสามารถแปลได้ดีกว่าที่นักวิจัยทั่วไปคาดไว้มาก ทำให้นักวิจัยทางระบบแปลภาษาด้วยเครื่องทั้งหลายต้องย้อนกลับไปดูวิธีการทางสถิติ

ภาพที่ 2.2 แสดงตัวอย่างของคู่ประโยคแบบขนานที่มีการวางแนวซึ่งจะเห็นได้ว่าจำเป็นต้องใช้มนุษย์มากำกับ (tag) ความสัมพันธ์ระหว่างประโยคต้นทางและประโยคปลายทางเป็นเหตุให้การเตรียมคลังข้อความของวิธีการนี้เป็นไปอย่างยากลำบากและใช้เวลามาก



ภาพที่ 2.2 คู่ประโยคแบบขนานที่มีการวางแนว (Alignment Paralleled Sentence)

จุดเด่นของวิธีการนี้คือ ไม่มีการใช้กฎไวยากรณ์ ทำไม่เกิดปัญหาเชิงภาษาศาสตร์ อาทิเช่น ปัญหากฎไม่ครอบคลุม ปัญหาการเพิ่มกฎ ปัญหาการแจงประโยควายกสัมพันธ์ (Syntax parsing) และปัญหาการแปลสำนวน เป็นต้น

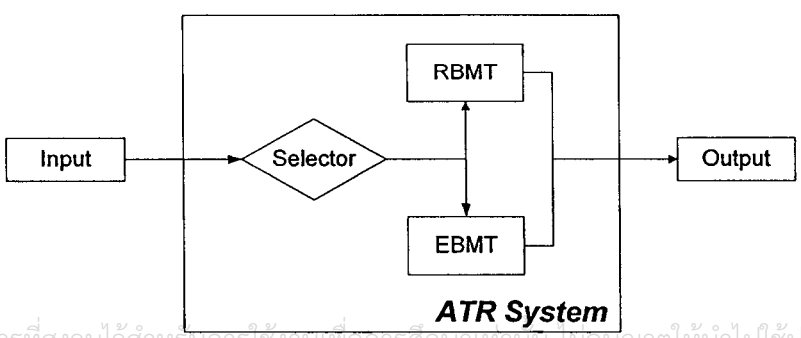
จุดด้อยของวิธีการนี้คือ จำเป็นต้องใช้คลังข้อความคู่ประโยคแบบขนานที่มีการวางแนว (Alignment Paralleled Corpus) จำนวนมากในการที่จะสร้างตัวแบบสถิติ (Statistical-Model)

2.5.2 ระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง (Example-Based MT)

แนวคิดของระบบการแปลแบบอิงตัวอย่าง (EBMT: Example-based MT) เป็นแนวคิดที่พยายามเลียนแบบพฤติกรรมการแปลภาษาของมนุษย์ ซึ่งแนวคิดพื้นฐานมาจากการศึกษาการแปลของคนเราซึ่งมักจะหาลักษณะประโยคคล้ายๆ กันที่เคยแปลมาก่อนมาเทียบในการแปลเป็นประโยคภาษาปลายทาง วิธีการนี้จะใช้คลังข้อความสองภาษา (Bi-Lingual Corpus) ของวลีหรือประโยค ตัวอย่างซึ่งได้จากคลังข้อความขนาดใหญ่

ในปี 1981 มีบุคคลแรกได้นำเสนอแนวคิดซึ่งถือได้ว่าเป็นต้นแบบของระบบแปลภาษาแบบอิงตัวอย่างคือ Nagao [7] ซึ่งเขาได้สังเกตกระบวนการแปลของมนุษย์ว่าเวลาเราเจอประโยคต้นทางใหม่ เราจะพยายามรู้จำความคล้าย (recognizing the similarity) ของประโยคต้นทางที่พบใหม่กับส่วนของประโยคที่เรารู้จักที่มีอยู่ในความทรงจำ (selecting identical phrases available in the translation memory) เว้นแต่ความคล้ายที่รู้จำนั้นเป็นความคล้ายระดับคำ (except for a similar content word) แม้ว่า ณ เวลานั้น คำว่า "การแปลแบบอิงตัวอย่าง" จะยังไม่ได้ถูกกำหนดขึ้นมา แต่แนวคิดของ Nagao ถือได้ว่าเป็นแนวคิดของการแปลแบบอิงตัวอย่าง ทำให้ Nagao ได้รับการยกย่องว่าเป็นผู้ให้กำเนิดการแปลภาษาด้วยแนวคิดนี้ในปี 1984

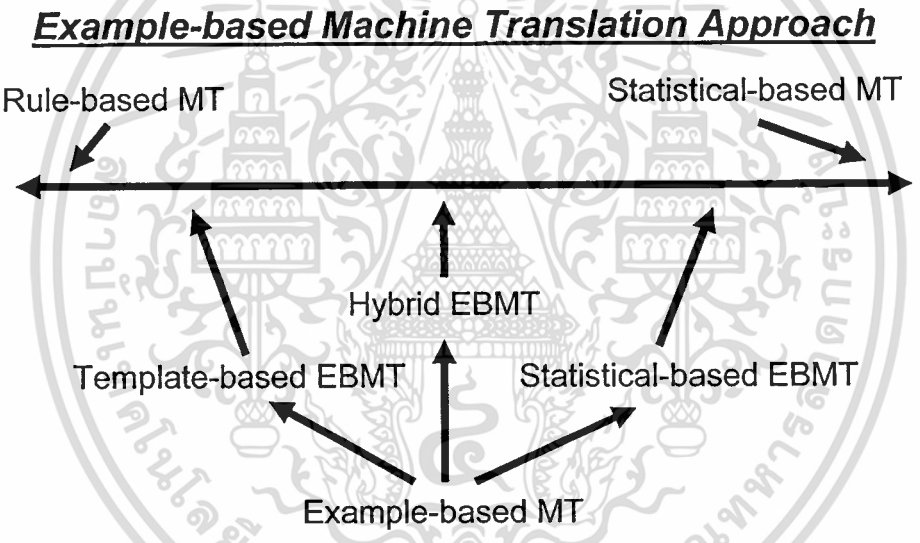
ในปี 1990 เริ่มมีความพยายามนำแนวคิดของ Nagao มาประยุกต์ใช้เพื่อเพิ่มประสิทธิภาพในการแปลของระบบแปลภาษาแบบอิงไวยากรณ์ (RBMT: Rule-based MT) โดยในงานของ Sato และ Nagao [8] จะใช้การแปลแบบอิงตัวอย่างเมื่อระบบการแปลแบบอิงกฎไวยากรณ์ไม่สามารถเข้าคู่การแจงประโยควายกสัมพันธ์ระหว่างภาษาต้นทางและภาษาปลายทางได้โดยตรง วิธีการนี้ถูกนำไปใช้ในระบบ ATR



ภาพที่ 2.3 ระบบ ATR

ระบบ ATR [9,10,11] เป็นระบบแปลภาษาญี่ปุ่น-อังกฤษด้วยเครื่องแบบอิงกฎไวยากรณ์ด้วยเสียง (Spoken Japanese-English MT) ระบบ ATR ถูกจัดว่าเป็นระบบแปลภาษาแบบลูกผสม (Hybrid MT)โดยในระบบจะมีกลไกตัวเลือก (Selector) เพื่อจัดการกับเงื่อนไขพิเศษแบบต่างๆ เพื่อส่งไปยังระบบแปลภาษาที่สามารถจัดการกับเงื่อนไขที่ดีที่สุด ปัญหาหลักของระบบนี้อยู่ที่ความสามารถในการจัดการเงื่อนไขของกลไกตัวเลือก

"แนวความคิดระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างเริ่มมีข้อสรุปที่ชัดเจนว่าแนวความคิดดังกล่าวเป็นแนวคิดที่อยู่กลางระหว่างระบบแปลภาษาด้วยเครื่องแบบอิงสถิติ (Statistical-based Machine Translation) และระบบแปลภาษาแบบอิงกฎไวยากรณ์ (Rule-based Machine Translation)" [12] ดังแสดงในภาพที่ 2.4



ภาพที่ 2.4 แนวความคิดระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง

เนื่องจากแนวความคิดระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างเป็นแนวคิดที่อยู่ระหว่างกลางระหว่างทั้งสองแนวคิด หากใช้วิธีที่มีความเอนเอียงไปทางระบบแปลภาษาแบบอิงกฎไวยากรณ์มากกว่า จะเรียกว่า ระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างโดยใช้แม่แบบการแปล (Template-based EBMT) หากใช้วิธีที่มีความเอนเอียงไปทางระบบแปลภาษาแบบอิงสถิติมากกว่า จะเรียกว่า ระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างโดยใช้ค่าสถิติ (Statistical-based EBMT) อย่างไรก็ดี หากมีการนำแนวความคิดของทั้งสองแนวทาง คือ ระบบแปลภาษาแบบอิงกฎไวยากรณ์ และระบบแปลภาษาแบบอิงสถิติมาใช้ร่วมกันเราจะเรียกว่า ระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างแบบลูกผสม (Hybrid EBMT)

2.6 ระบบแปลภาษาอังกฤษ-ไทยภาษิต

ในปัจจุบันการแปลภาษาด้วยเครื่องคอมพิวเตอร์ที่สนับสนุนภาษาไทย มีการพัฒนามาแล้วหลายปี โดยที่เราเริ่มต้นในปี 2524 โดยทบวงมหาวิทยาลัยได้มีคำสั่งแต่งตั้งคณะอนุกรรมการโครงการวิจัยการแปลภาษาอังกฤษเป็นภาษาไทยด้วยเครื่องคอมพิวเตอร์ โดยใช้ชื่อว่าระบบอาเรียน (ARIANE) และหลังจากนั้นก็มีงานวิจัยออกมามากมาย จนมาถึงปัจจุบันระบบที่มีความถูกต้องมากระบบหนึ่งก็คือ โครงการพัฒนาระบบเครื่องแปลภาษาสำหรับภาษาในเอเชียซึ่งทางศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) [13,14] ได้ร่วมมือกับรัฐบาลญี่ปุ่น พร้อมด้วยนักวิจัยจากอีก 3 ประเทศคือ จีน อินโดนิเซีย และมาเลเซีย เรียกว่าภาษิต (Parsit) ซึ่งสามารถทดลองระบบการแปลได้ที่ http://www.suparsit.com/ โดยตัวภาษิตนั้นพัฒนามาจากระบบการแปลภาษาอังกฤษเป็นญี่ปุ่น (English-to-Japanese) โดยที่ไม่ได้ทำการแปลแบบคำต่อคำแต่ถ้าสามารถแปลแบบประโยคต่อประโยคได้โดยที่มีการใช้ฐานความรู้วากยสัมพันธ์ (syntax) และทางด้านอรรถศาสตร์ (semantics) การใช้กฎทางภาษาและรวมถึงการใช้พจนานุกรมด้วย โดยขั้นตอนพื้นฐานการทำงานนั้นแบ่งออกเป็น 2 ส่วน คือ

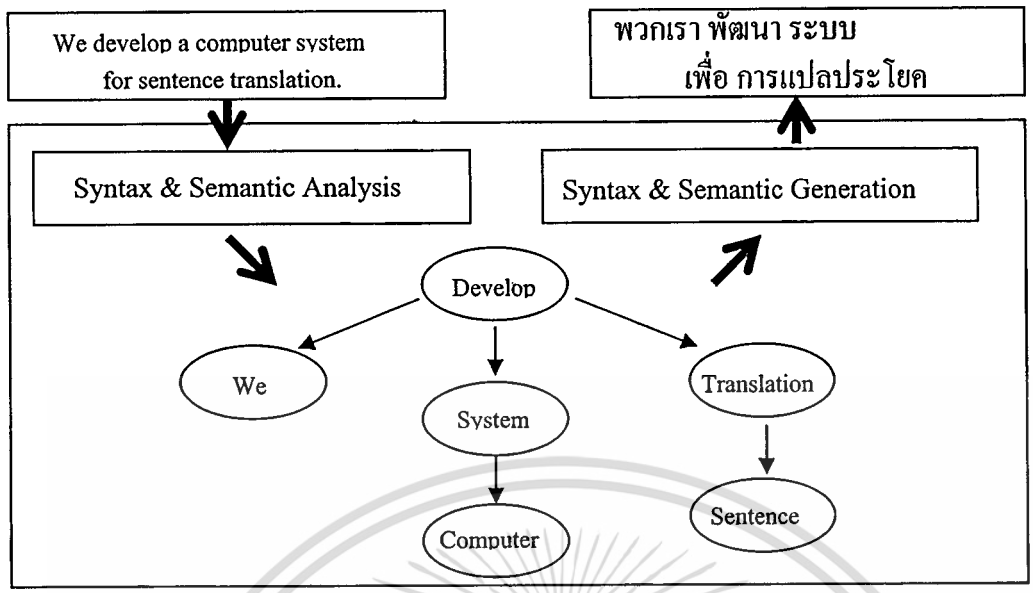
การวิเคราะห์ประโยคภาษาอังกฤษ

- วิเคราะห์หน่วยคำก่อนเข้าสู่กระบวนการหาความสัมพันธ์ทางไวยากรณ์ เช่น She (Dummy) hit (hit + past tense) a ball (a ball + singular)
- พิจารณาความสัมพันธ์ทางไวยากรณ์ของส่วนต่างๆ ของข้อความ She[subject] | hit[predicate] | a ball [object]
- พิจารณาความสัมพันธ์ทางความหมาย (Case relation) ของส่วนต่างๆ She <AGT> hit <OBJ> a ball
- สร้างรูปแทนกลาง

การสังเคราะห์ภาษาไทย

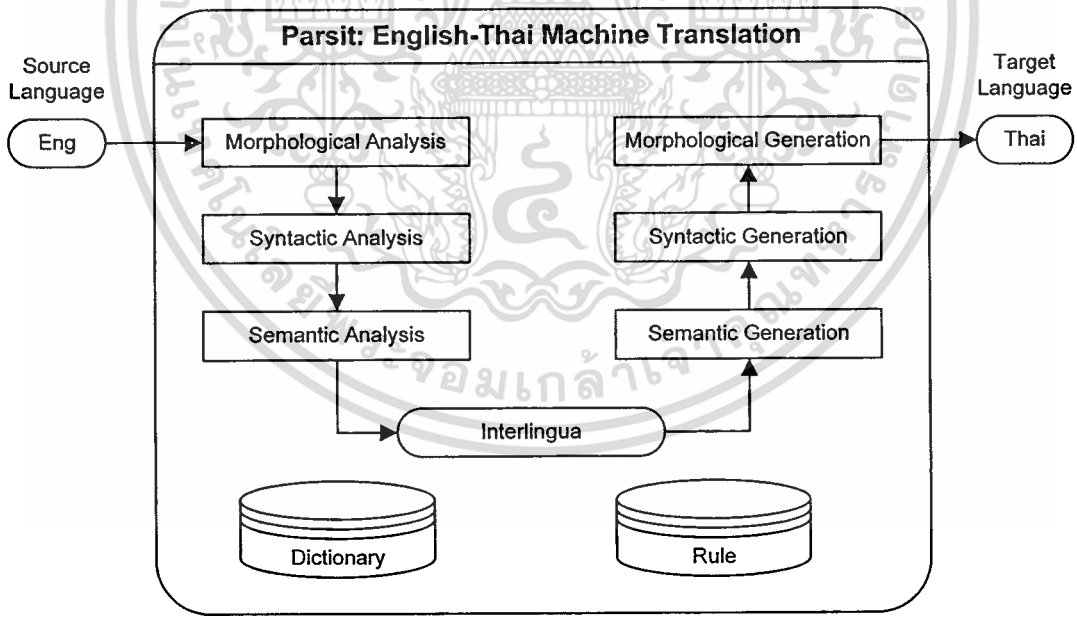
- เปลี่ยนความสัมพันธ์ทางความหมายให้เป็นความสัมพันธ์ทางไวยากรณ์
- นำข้อมูลทางไวยากรณ์ที่ได้รับผ่านรูปแทนกลางมาประกอบการสังเคราะห์ประโยค
- เปลี่ยนรูปคำจากอังกฤษเป็นภาษาไทย
- เรียงลำดับหน่วยคำต่างๆ ตามลักษณะไวยากรณ์ไทย

ภาพที่ 2.5 แสดงขั้นตอนการแปลภาษาของระบบภาษิต สำหรับข้อมูลเชิงเทคนิคของ "ภาษิต" จะกล่าวถึงในส่วนท้ายของบทนี้



ภาพที่ 2.5 แสดงขั้นตอนการแปลภาษาของระบบภาษิต

สถาปัตยกรรมระบบแปลภาษาอังกฤษ-ไทย "ภาษิต" ถูกแสดงไว้ใน ภาพที่ 2.6



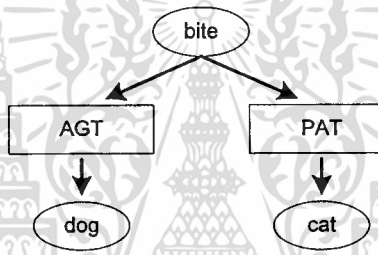
ภาพที่ 2.6 สถาปัตยกรรมระบบแปลภาษาอังกฤษ-ไทย "ภาษิต"

โมดูลวิเคราะห์สัณฐานคำ (Morphological Analysis) คือ การวิเคราะห์เชิงสัณฐาน ทำหน้าที่วิเคราะห์หน่วยคำในภาษาอังกฤษให้เป็นรากศัพท์รวมถึงการเตรียมความพร้อมให้แก่ระบบ เช่น การโหลดพจนานุกรม (Dictionary) เข้าสู่หน่วยความจำ การแก้ปัญหาความกำกวมของหน่วยคำ (morphological disambiguation)

โมดูลวิเคราะห์วากยสัมพันธ์ (Syntactic Analysis) คือ การวิเคราะห์เชิงวากยสัมพันธ์ ทำหน้าที่วิเคราะห์โครงสร้างวากยสัมพันธ์ของข้อความในภาษาอังกฤษ จากนั้นจะนำไปสร้างต้นไม้การแจงวากยสัมพันธ์ (syntactic parse tree) เพื่อนำไปใช้ในส่วนถัดไป

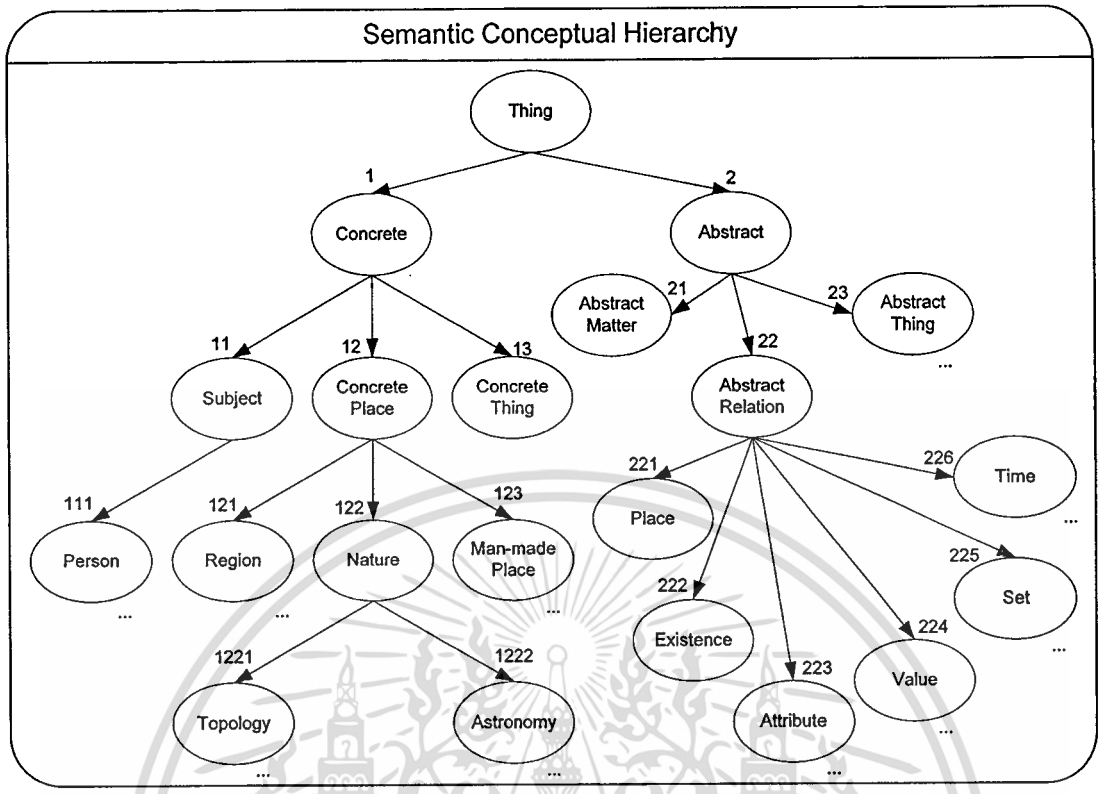
โมดูลวิเคราะห์อรรถศาสตร์ (Semantic Analysis) คือ การวิเคราะห์ความหมายจากโครงสร้างไวยากรณ์ ทำหน้าที่วิเคราะห์ต้นไม้แจงความหมายวากยสัมพันธ์ แล้วสร้างรูปแทนความหมาย (semantic representation) โดยใช้ภาษากลาง (Interlingua) เป็นสื่อในการถ่ายทอดความหมายไปเป็นภาษาไทย

ภาษากลาง (Interlingua) เป็นรูปแทนความหมายที่ไม่ขึ้นต่อภาษาใดๆ ความหมายของประโยคจะถูกแทนด้วยความสัมพันธ์ระหว่างมโนทัศน์เชิงอรรถศาสตร์ (semantic concept) เช่น ประโยค “สุนัขกัดแมว” จะแทนด้วยภาษากลางได้ดังภาพที่ 2.7



ภาพที่ 2.7 ตัวอย่างความสัมพันธ์ระหว่างมโนทัศน์เชิงอรรถศาสตร์

ในที่นี้ มโนทัศน์เชิงอรรถศาสตร์ (dog) กับ (bite) สัมพันธ์กันโดยความสัมพันธ์ Agent (ผู้กระทำกริยา) และ (cat) กับ (bite) สัมพันธ์กันโดยความสัมพันธ์ Patient (ผู้ได้รับผลโดยตรงจากการกระทำกริยา) ทั้งนี้แต่ละมโนทัศน์เชิงอรรถศาสตร์จะมีข้อกำหนด (constraint) ที่ใช้บังคับการสร้างความสัมพันธ์ระหว่างมโนทัศน์เชิงอรรถศาสตร์ เช่น (bite) จะสร้าง [AGT] กับสิ่งมีชีวิต (animate) เท่านั้น และสร้าง [PAT] กับสิ่งที่มีตัวตน (concrete object) เท่านั้น ข้อกำหนดดังกล่าวจำเป็นต้องใช้ลำดับชั้นของมโนทัศน์เชิงอรรถศาสตร์ดังที่แสดงไว้ในภาพที่ 2.8



ภาพที่ 2.8 ลำดับชั้นของมโนทัศน์เชิงอรรถศาสตร์ (Semantic Conceptual Hierarchy)

โมดูลสังเคราะห์เชิงอรรถศาสตร์ (Semantic Generation) คือ การตีความรูปแทนความหมายในภาษากลาง เพื่อนำไปสร้างต้นไม้แจงความหมายวากยสัมพันธ์ (syntactic parse tree) สำหรับภาษาไทย

โมดูลสังเคราะห์วากยสัมพันธ์ (Syntactic Generation) คือ การตีความหมายจากการแจงความหมายวากยสัมพันธ์แล้วนำไปสร้างรูปประโยคดั้งต้นในภาษาไทย ในขั้นตอนนี้โครงสร้างต้นไม้ไวยากรณ์จะถูกวิเคราะห์เพื่อสร้างลำดับการเรียงคำและเพิ่มเติมส่วนขยาย เช่น คุณสมบัติของลักษณนาม (classifier) ที่ถูกต้องตามหลักไวยากรณ์ภาษาไทย

โมดูลสังเคราะห์เชิงสัณฐานคำ (Morphological Generation) คือ การสร้างรูปผิวของคำสำหรับภาษาไทย โดยรับลำดับการเรียงคำและส่วนขยายที่เพิ่มเติมมาสร้างประโยคที่สมบูรณ์ในภาษาไทย ระบบจะเติมส่วนขยาย เช่น ลักษณนาม ลงในประโยคตามกฎการใช้ส่วนขยายนั้นๆ เพื่อให้เป็นประโยคที่สมบูรณ์

พจนานุกรมอิเล็กทรอนิกส์ที่ใช้ใน "ภาษิต" ประกอบด้วยพจนานุกรมหลัก 2 ประเภทคือ พจนานุกรมสำหรับคำหลัก (content words) ประกอบด้วยคำในกลุ่มคำนาม (noun) คำกริยา (verb) คำคุณศัพท์ (adjective) คำวิเศษณ์ (adverb) เป็นหลัก และพจนานุกรมสำหรับคำไวยากรณ์ (function word) ซึ่งประกอบด้วยคำในกลุ่มคำช่วยกริยา คำเชื่อม เป็นต้น นอกจากนี้ระบบยังให้ผู้ใช้เพิ่มพจนานุกรมตัวอื่นๆ ได้อีกตามความต้องการ เช่น พจนานุกรมศัพท์เฉพาะ เพื่อใช้ในกรณีที่ต้องการแปลเอกสารที่มีเนื้อหาเฉพาะทาง

ข้อมูลต่างๆ ของคำศัพท์แต่ละคำที่เก็บไว้ในพจนานุกรมแบ่ง 3 ส่วนคือ

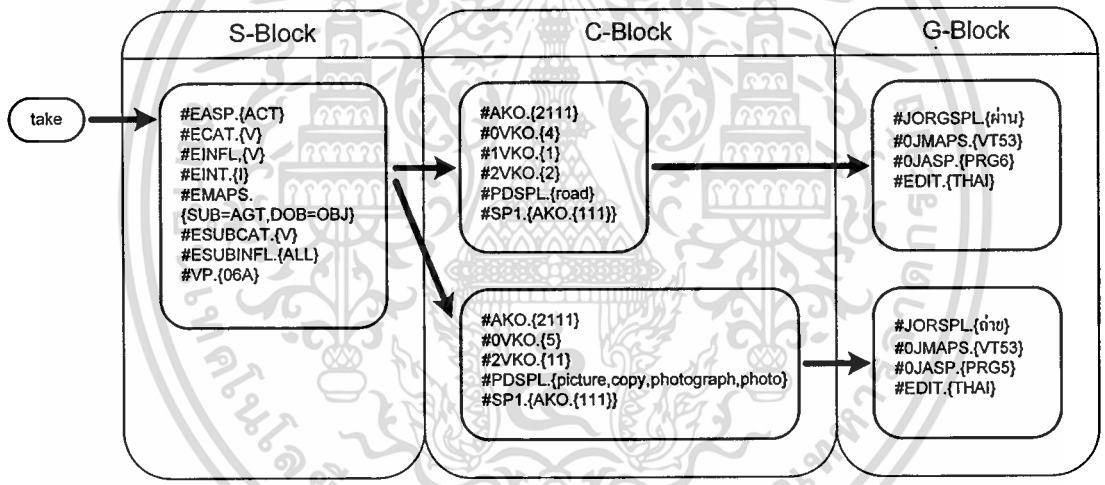
S-BLOCK ประกอบด้วยข้อมูลทางวากยสัมพันธ์ของคำศัพท์ภาษาต้นทาง

C-BLOCK ประกอบด้วยข้อมูลทางความหมาย

G-BLOCK ประกอบด้วยข้อมูลทางวากยสัมพันธ์ของคำศัพท์ปลายทาง

เพื่อเป็นการง่ายในการจัดการโครงสร้างของพจนานุกรมจึงถูกออกแบบไว้เป็นดังนี้คือ คำศัพท์ 1 คำ สามารถมี S-BLOCK มากกว่า 1 แต่ละ S-BLOCK มี C-BLOCK ได้เพียง 1 เท่านั้น แต่สามารถมี G-BLOCK ได้มากกว่า 1

ตัวอย่างข้อมูลที่เก็บในพจนานุกรม เช่น คำว่า take ซึ่งเป็นรูปคำที่มีการใช้ (usage) หลายรูปแบบ take สามารถแปลเป็นภาษาไทยได้หลายคำโดยขึ้นอยู่กับบริบท พจนานุกรมจะเก็บลักษณะทางวากยสัมพันธ์และอรรถศาสตร์ของคำว่า take ไว้เป็นชุดๆ เรียกว่าโหนด (node) ดังแสดงไว้ในภาพที่ 2.9



ภาพที่ 2.9 ตัวอย่างข้อมูลในพจนานุกรมของคำว่า take

ข้อมูลในพจนานุกรมนี้จะใช้ร่วมกับกฎการวิเคราะห์ในการเลือกโหนดที่มี S-BLOCK ที่เหมาะสมที่สุดเพียงโหนดเดียวเพื่อสร้างเป็นรูปแทนกลางซึ่งจะส่งให้ระบบสังเคราะห์ภาษาไทยต่อไป

เมื่อโมดูลสังเคราะห์ภาษาไทยรับข้อมูลซึ่งแสดงอยู่ในรูปของโครงสร้างต้นไม้ ก็จะเปิดพจนานุกรมเพื่อดึงข้อมูลใน G-BLOCK ของคำจากโหนดที่รูปแทนกลางส่งมาเท่านั้น หากมี G-BLOCK มากกว่า 1 กฎการสังเคราะห์ภาษาไทยจะทำการคัดเลือกให้เหลือเพียง 1 G-BLOCK จากนั้นจึงสร้างเป็นข้อความภาษาไทยต่อไป

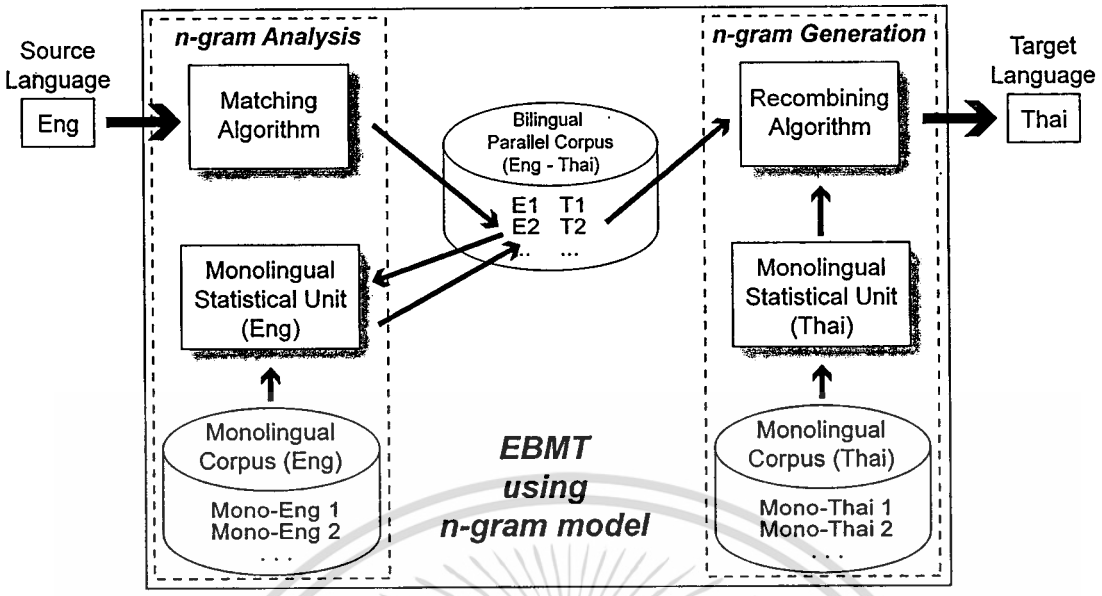
บทที่ 3
ระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่าง โดยใช้ตัวแบบเอ็นแกรม

ในบทนี้จะกล่าวถึงรายละเอียดต่างๆ ของระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรม (ขอเรียกโดยย่อว่า "ระบบ") อันประกอบด้วย สถาปัตยกรรมระบบ (System Architecture) แนวคิดของเอ็นแกรม (The *n*-gram Approach) ส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรม (*n*-gram analysis Component) ส่วนโปรแกรมการก่อกำเนิดแบบเอ็นแกรม (*n*-gram generation Component) และคลังข้อความ (Corpus)

3.1 สถาปัตยกรรมระบบ (System Architecture)

สถาปัตยกรรมของระบบนี้ได้ถูกแสดงไว้ในภาพที่ 3.1 ระบบนี้ประกอบด้วยส่วนสำคัญ 2 ส่วน คือ ส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรม (*n*-gram analysis component) และ ส่วนโปรแกรมการก่อกำเนิดแบบเอ็นแกรม (*n*-gram generation component) ระบบนี้ได้นำแนวคิดแบบตัวแบบเอ็นแกรมเข้าแก้ปัญหาความหลากหลายของแบบอย่าง (pattern) โดยอิงจากประโยคหรือส่วนของประโยค เมื่อระบบได้รับข้อมูลประโยคภาษาต้นทาง จะส่งข้อมูลดังกล่าวเพื่อทำการหาประโยคที่เข้าคู่ (match) กันได้ในคลังข้อความแบบคู่ ทุกประโยคหรือส่วนของประโยคที่เข้าคู่จะถูกนำไปเปรียบเทียบเพื่อหาผลลัพธ์ที่มีความใกล้เคียงกับคลังข้อความแบบเดี่ยวสำหรับภาษาปลายทาง

คลังข้อความแบบเดี่ยวถูกใช้ในขั้นตอนวิเคราะห์และก่อกำเนิดเพื่อหาผลลัพธ์ที่ดีที่สุดสำหรับแต่ละประโยคหรือส่วนของประโยคที่เข้าคู่กัน ขณะที่คลังข้อความแบบคู่จะประกอบด้วยคู่ประโยคหรือส่วนของประโยคของภาษาอังกฤษและภาษาไทยจำนวนมาก โดยข้อมูลในคลังข้อความแบบคู่จะทำหน้าที่เป็นกฎการโอนย้าย (Transfer Rule) อย่างไรก็ดี คลังข้อความทั้งหมดที่ใช้ในวิทยานิพนธ์นี้ถูกเตรียมโดยนักภาษาศาสตร์ของแผนกเทคโนโลยีประมวลผลข้อความ ฝ่ายวิจัยและพัฒนาเทคโนโลยีสารสนเทศ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ประเทศไทย (NECTEC)



ภาพที่ 3.1 ระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรม

3.2 แนวคิดของเอ็นแกรม (The *n*-gram Approach)

ตัวแบบภาษาของเอ็นแกรม (*n*-gram language model) อยู่ภายใต้สมมติฐานต่อไปนี้ คำตำแหน่งที่ n จะมีความสัมพันธ์เฉพาะกับตัวก่อนหน้าของคำนั้นๆ ซึ่งก็คือคำตำแหน่งที่ n-1 ดังนั้น ค่าประมาณความน่าจะเป็นของตัวแบบ $P(w)$ จะสามารถเขียนได้เป็น $P(w_n|w_1,...,w_{n-1})$ สำหรับประโยคที่มีจำนวน N คำ

ให้ $w_1, w_2, ..., w_N$ เป็นคำที่มาจากคลังข้อความแบบคู่ จะสามารถคำนวณความน่าจะเป็นของทั้งประโยคได้ดังสมการต่อไปนี้

$$P(w) = \sum_{i=1}^{N} P(w_i|w_{i-n+1},...,w_{i-1}) \quad (3.1)$$

หากคลังความมีขนาดใหญ่เพียงพอ ความน่าจะเป็นของ $P(w) = \sum_{i=1}^{N} P(w_i|w_{i-n+1},...,w_{i-1})$ จะคำนวณโดยหลักการความควรจะเป็นสูงสุด (Maximum likelihood principle) ได้เป็น

$$P(w_n|w_1,...,w_{n-1}) = \frac{C(w_1,...,w_n)}{C(w_1,...,w_{n-1})} \quad (3.2)$$

เมื่อ $C(w_1,...,w_{n-1})$ และ $C(w_1,...,w_n)$ คือ จำนวนครั้งที่ปรากฏของคำในสายข้อความ $w_1,...,w_{n-1}$ และ $w_1,...,w_n$ ตามลำดับ

3.3 ส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรม (*n*-gram analysis Component)

ส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรมมีจุดมุ่งหมายเพื่อวิเคราะห์ประโยคต้นแบบโดยตัดส่วนของประโยคออกมาในทุกๆ ทางที่เป็นไปได้ และเข้าคู่ตัวเลือกที่เหมาะสมที่สุด การทำงานหลักของส่วนโปรแกรมนี้คือ ขั้นตอนวิธีเข้าคู่ (Matching Algorithm)

ขั้นตอนวิธีเข้าคู่ จะถูกนำมาใช้เพื่อค้นหาประโยคหรือส่วนของประโยคที่ยาวที่สุด ส่วนของประโยคที่ยาวที่สุดจะอนุมานจากการเข้าคู่ของข้อมูลในคลังข้อความคู่ภาษา (คลังข้อความแบบคู่) ในกรณีที่มีทางให้เลือกมากกว่า 1 ทาง ระบบจะเลือกส่วนของประโยคที่สมควรนำมาใช้โดยอิงจากข้อมูลทางสถิติ ที่ได้จากคลังข้อความแบบเดี่ยวสำหรับภาษาต้นทาง

ส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรมมีขั้นตอนการทำงานดังนี้

Step 1: Define *Fragment_Set* = *{SL}* and
Result_Set = *{ }*
Step 2: Generate sub-fragment Sf_i from *Fragment_Set* by segmenting groups of words that $next(w_i) \neq w_{i+1}$
Step 3: For each Sf_i that has more than one elements, find the maximum sub-sentence Max_Sf_i in Sf_i
Step 3.1: Push Max_Sf_i into *Result_Set*
Step 3.2: Delete Max_Sf_i from *Fragment_Set*
Step 4: Repeat Step 1 until each sub-fragment Sf_i has only one element
Step 5: Return Sf_i
Step 6: Return *Result_Set*

จากประโยคตัวอย่าง "Arsenal picked up a big victory in Champions League" เราตั้งสมมติฐานว่า มีส่วนของประโยค {a big victory}, {picked up}, และ {Champions League} อยู่ในคลังข้อความคู่ภาษา ทำให้ Matching Algorithm จะประมวลผลตามลำดับขั้นตอนต่อไปนี้

Step 1: *Fragment_Set = { {Arsenal picked up a big victory in Champions League} }* ,
Result_Set = { }
Step 2: Since $next(w_i) = w_{i+1}$ for all w_i , sub-fragment Sf_i has only one sub-fragment that is
Sf_1 *= {Arsenal picked up a big victory in Champions League}*
Step 3: maximum sub-sentence
Max_sf_1 *= {a big victory}*
Step 3.1: *Result_Set ={ {a big victory} }*
Step 3.2: *Fragment_Set = { {Arsenal picked up},*
{in Champions League} }
Step 4: *Fragment_Set = { {Arsenal picked up},*
{in Champions League} } ,
Result_Set = { {a big victory} }
Step 5: Since $next(w_i) \neq w_{i+1}$ at w_i= up, there are two sub-fragments,
Sf_1 *= {Arsenal picked up} and*
Sf_2 *= {in Champions League}*
Step 6: maximum sub-sentence Max_sf_1 *= {picked up}*
Step 6.1: *Result_Set ={ {picked up}, {a big victory} }*
Step 6.2: *Fragment_Set = { {Arsenal}, {in Champions League} }*
Step 7: maximum sub-sentence Max_sf_2 *= {in Champions League}*

Step 7.1: *Result_Set = { {Champions League}, {picked up}, {a big victory} }*
Step 7.2: *Fragment_Set = { {Arsenal}, {in} }*
Step 8: Since $next(w_i) \neq w_{i+1}$ at w_i= Arsenal, there are two sub-fragments Sf_1 = *{ Arsenal }* and Sf_2 =*{ in }*
Step 9: Return *Result_Set = { {Arsenal}, {in}, {Champions League}, {picked up}, {a big victory} }*

ผลจากส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรมจะถูกแปลเป็นภาษาไทยบนพื้นฐานของเกณฑ์ 2 เกณฑ์ ถ้าหากมีสมาชิกใดๆ ใน Result_Set ที่ไม่ใช่คำโดด ให้ใช้คำแปลจากคลังข้อความคู่ภาษาถ้าหากสมาชิกใน Result_Set เป็นคำโดด ให้ดึงคำแปลมาจากพจนานุกรมแทน ผลลัพธ์การแปลจาก Result_Set จะถูกส่งไปที่ส่วนโปรแกรมการก่อกำเนิดแบบเอ็นแกรม เช่น Result_Set = {{Arsenal}, {in}, {Champions League}, {picked up}, {a big victory}} จะถูกแปลเป็น {{เจ้าปืนใหญ่อาร์เซนอล}, {ใน}, {แชมเปี้ยนลีค}, {ได้}, {ชัยชนะครั้งใหญ่}}

3.4 ส่วนโปรแกรมการก่อกำเนิดแบบเอ็นแกรม (*n*-gram generation Component)

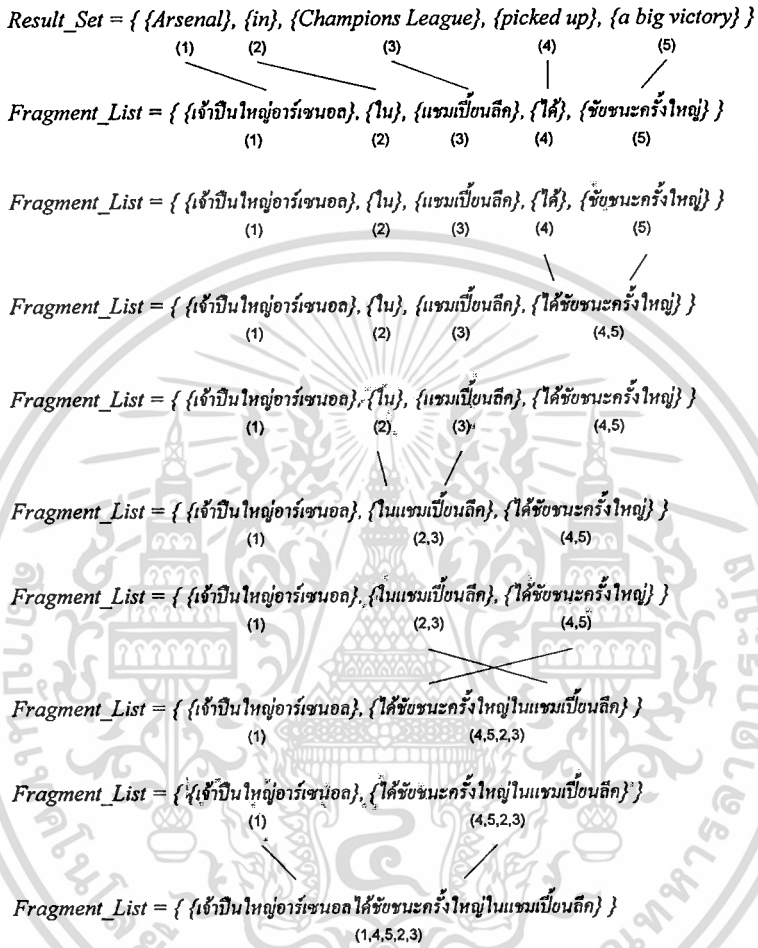
ส่วนโปรแกรมการก่อกำเนิดแบบเอ็นแกรมมีไว้เพื่อสร้างประโยคในภาษาปลายทาง โดยการรวมและเรียงลำดับชิ้นส่วนของประโยคให้เป็นประโยคเต็ม หน้าที่หลักของส่วนนี้คืออัลกอริธึมการ Recombine ในส่วนนี้จะใช้ Greedy Algorithm มาตรวจจับส่วนของประโยคที่ดีที่สุดที่สามารถนำมาต่อกันได้ กระบวนการ Recombine จะนำส่วนของประโยคมารวมกันเป็นประโยคในภาษาปลายทางโดยพิจารณาตามลำดับในประโยค แต่ละส่วนของประโยคจะนำมาต่อกันกับตัวใกล้เคียงซึ่งมีความถี่ที่ปรากฏในคลังข้อความสูงที่สุด กระบวนการ Recombine จะมีขั้นตอนดังต่อไปนี้

Step 1: Define *Fragment_List* = $\{Fr_1, Fr_2, ..., Fr_n\}$
Step 2: [combine the maximum probability of sub-sentence and its neighbor]
For each Fr_a and Fr_b in *Fragment_List* that
$1 \leq a, b \leq n$ and $|a-b| = 1$, Fr_{ab} = max_combine (Fr_a , Fr_b)
Step 3: Substitute Fr_a and Fr_b with Fr_{ab} and
delete Fr_a and Fr_b from *Fragment_List*
Fragment_List = $\{Fr_1, Fr_2, ..., Fr_{ab}, ..., Fr_n\}$
Step 4: Repeat Step 1 until *Fragment_List* has only one element
Step 5: Return *Fragment_List*

จากตัวอย่างข้างบน ผลลัพธ์ Result_Set ในส่วนวิเคราะห์โครงสร้างประโยคด้วย *n*-gram จะได้เป็น {{Arsenal}, {in}, {Champions League}, {picked up}, {a big victory}} ซึ่งสามารถนำมาแปลเป็นรายคำได้เป็น Fragment_List = { เจ้าปืนใหญ่อาร์เซนอล, ใน, แชมเปี้ยนลีค, ได้, ชัยชนะครั้งใหญ่ }

เมื่อนำกระบวนการ Recombine มาใช้ ส่วนของประโยคใน Fragment_List จะถูกรวมในขั้นตอนที่ 2 โดยพิจารณาจากส่วนของประโยคและส่วนใกล้เคียง ณ ระยะกระจัด 1 ดังได้แสดงไว้

ในภาพที่ 2 ส่วนที่ทาบสีเทาเอาไว้จะเป็นคู่ของคำหรือวลีที่สามารถนำมา combine กันได้ในขั้นตอนต่อไป ในท้ายที่สุดแล้ว จะสังเคราะห์ผลการแปลได้เป็นประโยค “เจ้าปืนใหญ่อาร์เซนอลได้ชัยชนะครั้งใหญ่ในแชมเปี้ยนลีค” ออกมาดังภาพที่ 3.2



ภาพที่ 3.2 ตัวอย่างการทำงานส่วนโปรแกรมการก่อกำเนิดแบบเอ็นแกรม

3.5 คลังข้อความ (Corpus)

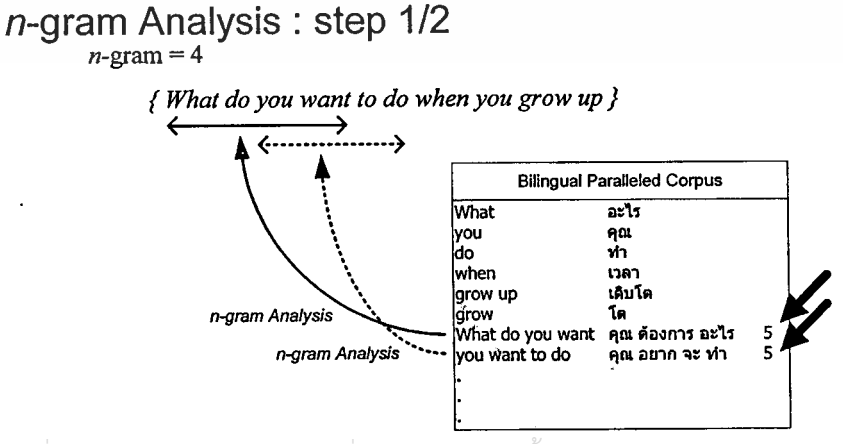
คลังข้อความที่ใช้ในระบบประกอบคลังข้อความหลายประเภท ได้แก่ คลังข้อความแบบเดี่ยว (Monolingual Corpus) และ คลังข้อความแบบคู่ (Bilingual Paralleled Corpus) เนื่องจากระบบนี้เป็นแบบออฟไลน์ ดังนั้นคลังข้อความทุกประเภทจะต้องถูกจัดเตรียมไว้ก่อนแล้ว

คลังข้อความจะถูกจัดเตรียมโดยนำส่วนที่เป็นภาษาไทยไปผ่านกระบวนการตัดคำซึ่งเป็นความสามารถที่มีอยู่แล้วใน Java 2 Platform, Standard Edition (J2SE) 5.0 จาก จากนั้นจึงนำไปเข้าสู่กระบวนการเรียนรู้ค่าสถิติของตัวแบบเอ็นแกรม

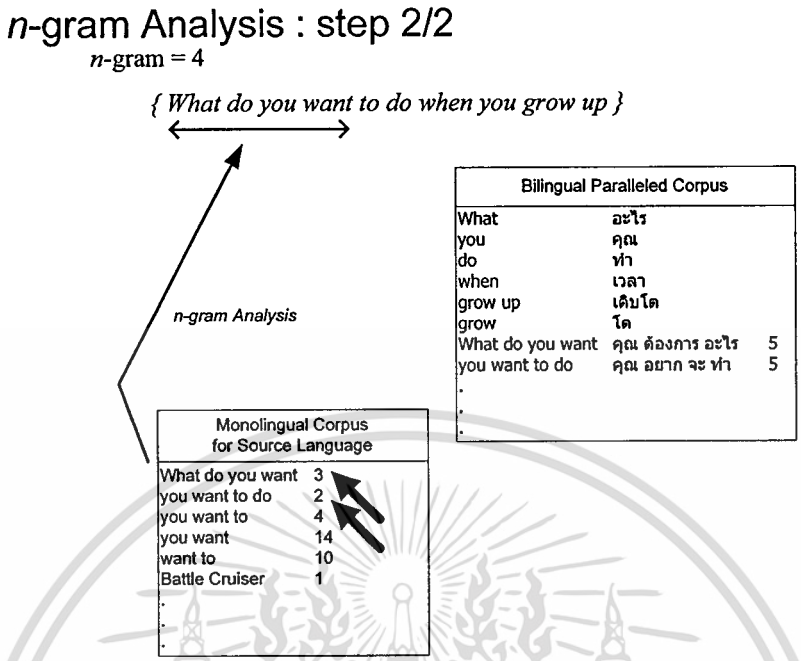
3.5.1 คลังข้อความแบบเดี่ยว (Monolingual Corpus)

คลังข้อความแบบเดี่ยวมี 2 ประเภท คือ คลังข้อความแบบเดี่ยวสำหรับภาษาต้นทาง (Monolingual Corpus for Source Language) และ คลังข้อความแบบเดี่ยวสำหรับภาษาปลายทาง (Monolingual Corpus for Target Language) โดยคลังข้อความทั้งสองประเภทเกิดจากการคำนวณค่าความถี่แบบเอ็นแกรม (*n*-gram frequency) ไว้ล่วงหน้า โดยจะมีการเตรียมความถี่ของ 2-gram, 3-gram และ 4-gram [15] และถูกจัดเก็บไว้ในรูปแบบแฟ้มข้อมูลแบบคั่นด้วยอักขระตั้งระยะ (TSV : Tab-separated value) โดยจะมีทั้งหมด 2 สดมภ์ (column) สดมภ์ที่หนึ่งจะเป็นข้อความที่ถูกแบ่งคำไว้แล้ว โดยจำนวนคำที่ถูกแบ่งจะหมายถึงจำนวนเอ็นแกรมของระเบียน (record) สดมภ์ที่สองคือจำนวนความถี่แบบเอ็นแกรม โดยสดมภ์ที่สองนี้จะมีหรือไม่ก็ได้ หากไม่มีจะถือว่ามีค่าความถี่แบบเอ็นแกรมเป็นหนึ่ง

คลังข้อความแบบเดี่ยวสำหรับภาษาต้นทาง (Source Language) ถูกใช้ในขั้นตอนส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรมเพื่อแก้ปัญหาความกำกวมในการหาส่วนของประโยคสำหรับภาษาต้นทาง โดยความกำกวมในการหาส่วนของประโยคสำหรับภาษาต้นทางจะเกิดขึ้นในขณะที่ส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรมพบว่า มีจำนวนส่วนของประโยคที่สามารถเข้าคู่กันได้และมีความยาวเท่ากันมากกว่า 1 คู่ โดยส่วนของประโยคที่พบดังกล่าวมีการซ้อนทับกัน (overlap) ในกรณีเช่นนี้ ระบบจะเลือกส่วนของประโยคที่มีการซ้อนทับกันซึ่งมีค่าความถี่สูงสุดจากคลังข้อความแบบคู่ แต่ก็เป็นไปได้ที่ความถี่ของส่วนของประโยคที่เข้าคู่และมีการซ้อนทับกันจะมีค่าความถี่เท่ากัน หากเป็นเช่นนี้ ระบบจะนำทุกประโยคต้นทางจากส่วนของประโยคที่เข้าคู่และซ้อนทับกันทั้งหมดที่ตรวจพบ นำไปเปรียบเทียบเพื่อหาส่วนของประโยคที่มีความถี่สูงสุดในคลังข้อความแบบเดี่ยวสำหรับภาษาต้นทาง หากความถี่ในคลังข้อความแบบเดี่ยวสำหรับภาษาต้นทางยังเท่ากันอยู่อีก ระบบจะเลือกส่วนของประโยคที่เข้าคู่และซ้อนทับกันประโยคแรกโดยอิงตำแหน่งเริ่มต้นจากซ้ายไปขวา



ภาพที่ 3.3 การแก้ปัญหาความกำกวม (1/2)



ภาพที่ 3.4 การแก้ปัญหาความกำกวม (2/2)

คลังข้อความแบบเดี่ยวสำหรับภาษาต้นทางจะถูกนำไปคำนวณเป็นสถิติสำหรับคลังข้อความแบบเดี่ยวสำหรับภาษาต้นทาง (Monolingual Statistical Unit for Source Language) โดยจะคำนวณหาค่าความถี่ของเอ็นแกรม (*n*-gram frequency) ตั้งแต่ 2-gram ถึง 4-gram จากข้อความภาษาต้นทาง (ภาษาอังกฤษ) ซึ่งสามารถรวบรวมได้โดยง่ายจาก webpage ทั่วไปที่มีเฉพาะภาษาต้นทางเท่านั้น ในงานวิจัยนี้ได้ทำการรวบรวมค่าความถี่ของเอ็นแกรมจากทั้งหมด 4,000 webpage โดยแบ่งได้เป็น 104,893 ประโยค หรือ 561,387 คำ

What do you want	3
you want to do	2
you want to	4
you want	14
want to	10
.	
.	
.	

ภาพที่ 3.5 ตัวอย่างคลังข้อความแบบเดี่ยวสำหรับภาษาต้นทาง

คลังความแบบเดี่ยวสำหรับภาษาปลายทาง (Target Language) ถูกใช้ในขั้นตอนส่วนโปรแกรมการก่อกำเนิดแบบเอ็นแกรมเพื่อใช้ตัดสินลำดับที่ถูกต้องในกระบวนการ Recombine เพื่อสร้างประโยคผลลัพธ์ในภาษาปลายทางที่ถูกต้อง

เป็น อันตราย	5
เป็น อันตราย ต่อ	2
เป็น อันตราย ถึง	1
ก็ เป็น อันตราย	1
อาจ เป็น อันตราย	1
ซึ่ง อาจ เป็น อันตราย	1
.	
.	
.	

ภาพที่ 3.6 ตัวอย่างคลังข้อความแบบเดี่ยวสำหรับภาษาปลายทาง

ในงานวิจัยชิ้นนี้ คลังข้อความแบบเดี่ยวสำหรับภาษาปลายทางจะถูกนำไปคำนวณเป็นสถิติสำหรับคลังข้อความแบบเดี่ยวสำหรับภาษาปลายทาง (Monolingual Statistical Unit for Target Language) โดยจะคำนวณหาค่าความถี่ของเอ็นแกรม (*n*-gram frequency) ตั้งแต่ 2-gram ถึง 4-gram ไว้ล่วงหน้า จากแหล่งข้อความภาษาปลายทางต่อไปนี้

- คลังข้อความทั่วไปเฉพาะภาษาปลายทางที่ถูกแบ่งประโยคและแบ่งคำไว้แล้ว มีจำนวน 13,758 ประโยค และมีจำนวนคำทั้งหมด 152,396 คำ
- คลังข้อความข่าวหนังสือพิมพ์ไทยรัฐ ย้อนหลัง 10 ปี ซึ่งถูกแบ่งประโยคและแบ่งคำไว้แล้ว มีจำนวน 138,812 ประโยค และมีจำนวนคำทั้งหมด 10,011,970 คำ

3.5.2 คลังข้อความแบบคู่ (Bilingual Paralleled Corpus)

คลังข้อความแบบคู่ ทำหน้าที่เป็นข้อมูลตัวอย่างประโยค ซึ่งเป็นหัวใจหลักสำหรับระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง คลังข้อความแบบคู่นี้ส่งผลโดยตรงต่อประสิทธิภาพการแปลของระบบ หากคลังข้อความแบบคู่มีข้อมูลตัวอย่างมากเพียงพอและถูกต้อง ข้อมูลตัวอย่างควรมีลักษณะเป็นส่วนของประโยคที่ปรากฏซ้ำๆ กันและมีปริมาณมาก คู่ความหมายควรมีความหมายที่เป็นกลางไม่เฉพาะเจาะจงมากนัก ผลการแปลจึงจะมีประสิทธิภาพและมีความแม่นยำสูง

What	อะไร	
you	คุณ	
do	ทำ	
when	เวลา	
grow up	เติบโต	
grow	โต	
What do you want	คุณ ต้องการ อะไร	5
you want to do	คุณ อยาก จะ ทำ	5
.		
.		
.		

ภาพที่ 3.7 ตัวอย่างคลังข้อความแบบคู่

หากคลังข้อความแบบคู่ มีการจัดเก็บคู่ภาษาต้นทางในระดับคำ จะเห็นได้อย่างชัดเจนว่าคลังข้อความแบบคู่จะมีสภาพไม่แตกต่างไปจากพจนานุกรมสองภาษา ดังนั้นพจนานุกรมสองภาษาจึงเสมือนว่าถูกรวมอยู่ในคลังข้อความแบบคู่ด้วย ซึ่งทำให้ส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรมสามารถหาความหมายของคำเมื่อ $n=1$ ได้สำเร็จ

คลังข้อความแบบคู่นี้ถูกจัดเก็บไว้ในรูปแบบแฟ้มข้อมูลแบบคั่นด้วยอักขระตั้งระยะ โดยจะมีทั้งหมด 3 สดมภ์ สดมภ์ที่หนึ่งจะเป็นข้อความสำหรับภาษาต้นทางที่ถูกแบ่งคำไว้แล้ว โดยจำนวนคำที่ถูกแบ่งจะหมายถึงจำนวนเอ็นแกรมของระเบียน สดมภ์ที่สองจะเป็นข้อความสำหรับภาษาปลายที่ถูกแบ่งคำไว้แล้ว โดยข้อมูลในสดมภ์ที่สองนี้จะถูกใช้เป็นคำแปลของภาษาปลายทาง สดมภ์ที่สามคือจำนวนความถี่แบบเอ็นแกรมที่แสดงถึงอัตราการปรากฏขึ้นระหว่างคู่ความหมายของภาษาต้นทางและภาษาปลายทาง ทั้งนี้สดมภ์ที่สามนี้จะมีหรือไม่ก็ได้ หากไม่มีจะถือว่ามีจำนวนความถี่แบบเอ็นแกรมเป็นหนึ่งโดยปริยาย

ในงานวิจัยชิ้นนี้ คลังข้อความแบบคู่จะถูกคำนวณค่าความถี่ของเอ็นแกรม (n-gram frequency) ตั้งแต่ 2-gram ถึง 4-gram ไว้ล่วงหน้า จากแหล่งคลังข้อความต่อไปนี้

- ตัวอย่างประโยคจากหนังสือ English by Examples จำนวน 100,804 ประโยค ซึ่งถูกแบ่งประโยคและแบ่งคำไว้แล้ว
- พจนานุกรมอังกฤษ-ไทย Lexitron จำนวน 97,791 คำ ซึ่งถูกแบ่งคำไว้แล้ว
- พจนานุกรมไทย-อังกฤษ Lexitron จำนวน 104,892 คำ ซึ่งถูกแบ่งคำไว้แล้ว

เนื่องจากคลังข้อความแบบคู่มักหาไม่ได้โดยทั่วไปนัก อีกทั้งยังรวบรวมได้ยากเนื่องจากระบบต้องการคู่ความหมายระดับประโยคหรือส่วนของประโยค แต่แหล่งข้อมูลส่วนมากแม้จะมีลักษณะเป็นคู่ภาษากันอยู่แล้ว แต่ก็ยังไม่ได้แบ่งแบบประโยคต่อประโยค (โดยมากจะพบแบบย่อหน้าต่อย่อหน้า) อย่างไรก็ดีระบบมีความจำเป็นต้องใช้คลังข้อความแบบคู่เป็นปริมาณมากพอเพื่อให้ประสิทธิภาพในการแปลออกมาดี จำเป็นต้องรวบรวมคลังข้อความแบบคู่เพิ่มเติมให้ได้มากที่สุด

บทที่ 4
การประเมินค่าความถูกต้องของผลการแปล

ในบทนี้จะกล่าวถึงขั้นตอนการประเมินค่าการแปล เพื่อประเมินค่าผลการแปลจากระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรมจากที่ได้อธิบายถึงอัลกอริทึมการทำงานในบทก่อนหน้านี้ และอธิบายถึงขั้นตอนการเตรียมข้อมูลเพื่อนำไปประเมินค่าผลการแปลของระบบ ทั้งนี้ผลการแปลที่เคยพบมาก่อนจะสามารถแปลได้อย่างถูกต้องแม่นยำโดยระบบนี้ ดังนั้นการประเมินค่าผลการแปลของระบบจะประเมินเฉพาะผลการแปลจากประโยคที่ไม่เคยพบมาก่อนเท่านั้น

4.1 การประเมินค่าความถูกต้องของผลการแปลสำหรับระบบแปลภาษาด้วยเครื่อง (Machine Translation Evaluation)

การประเมินค่าผลการแปลของระบบแปลภาษาด้วยเครื่อง (ขอเรียกโดยย่อว่า "การประเมินค่า") ทำได้ยาก เนื่องจากสามารถประเมินค่าได้จากหลายมุมมอง เช่น

- การประเมินค่าเชิงไวยากรณ์ (Grammatical Evaluation)
- การประเมินค่าเชิงคุณภาพการแปลอย่างเพียงพอ (Adequacy Evaluation)
- การประเมินค่าเชิงคุณภาพการแปล (Quality Assessment)

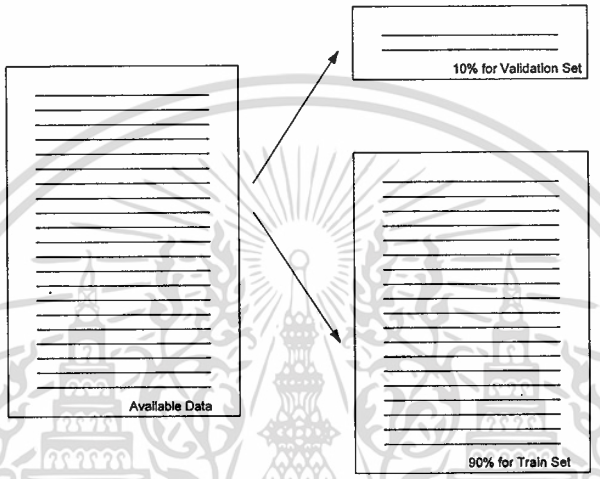
การประเมินค่าจากแต่ละมุมมองย่อมมีข้อโต้แย้งกันอีกมาก เนื่องจากการประเมินค่ามักขึ้นกับประสบการณ์และทักษะของผู้ประเมินค่าซึ่งเป็นความสามารถส่วนบุคคล ทำให้ผลการประเมินค่ามักจะมีความเอนเอียง (bias) เพื่อแก้ปัญหานี้จึงมีวิธีการประเมินค่าแบบอัตโนมัติ (Automatic Metrics for MT Evaluation) เพื่อใช้ประเมินค่าที่เป็นกลางและไม่ยึดติดกับความสามารถของบุคคล อย่างไรก็ตาม จุดด้อยของการประเมินค่าแบบอัตโนมัติคือไม่สามารถประเมินค่าความถูกต้องที่สูงกว่าระดับผิว (Surface Form) ได้ ทั้งนี้ยังไม่มีการประเมินค่าแบบใดดีที่สุด

ในงานวิจัยชิ้นนี้ได้เลือกใช้วิธีการประเมินค่าแบบอัตโนมัติของ BLEU-4 [16] ซึ่งเป็นวิธีที่ใช้กันเป็นที่แพร่หลายในการประเมินค่าแบบอัตโนมัติสำหรับระบบแปลภาษาด้วยเครื่อง

4.2 วิธีการทดสอบความถูกต้องของการเรียนรู้

4.2.1 การตรวจสอบความสมเหตุสมผลแบบไขว้ (Cross-Validation)

สำหรับวิทยานิพนธ์นี้เป็นการทดสอบเกี่ยวกับการเรียนรู้ของเครื่อง (Machine Learning) ด้วยวิธีซูเปอร์ไวส์เลินนิ่ง (Supervised Learning) ซึ่งเป็นการเรียนรู้ที่จะต้องสอนคำตอบที่ถูกต้องให้ ดังนั้นจึงต้องเตรียมข้อมูลตัวอย่างพร้อมคำตอบที่ถูกต้อง



ภาพที่ 4.1 แสดงการเตรียมชุดข้อมูลเพื่อสอนและทดสอบ

วิธีการหนึ่งที่เป็นที่นิยมใช้กันมากในการทดสอบหาค่าความถูกต้อง ซึ่งมีพื้นฐานทางด้านสถิติก็คือการทำ Cross-Validation โดยนำข้อมูลทั้งหมด (Available Data) นำมาแบ่งออกเป็นส่วนๆ ส่วนแรกเพื่อใช้ในการทดสอบ (Validation Set) และส่วนที่สองใช้ในการสอน (Train Set) วิทยานิพนธ์นี้ได้ใช้เทคนิค Cross-Validation เพื่อหาค่าความถูกต้องของแต่ละอัลกอริทึม โดยการแบ่งข้อมูลออกเป็น 10 ส่วน เรียกว่า 10-fold Cross Validation โดยมีการทำงานคือ ให้ข้อมูลแต่ละส่วนเขียนแทนด้วย Data1, Data2, ..., Data10 ในครั้งแรกให้ Data2 ถึง Data10 เป็นชุดข้อมูลที่ใช้เรียนรู้ และให้ Data1 เป็นชุดข้อมูลทดสอบ ครั้งต่อไปให้ Data1, Data3 ถึง Data10 เป็นข้อมูลสอนและให้ Data2 เป็นชุดข้อมูลทดสอบ ทำอย่างนี้จนครบ 10 ครั้งแล้วนำค่าความถูกต้องของการเรียนรู้ทั้ง 10 ครั้งมาหาค่าเฉลี่ย

4.2.2 การประเมินค่าแบบอัตโนมัติของ BLEU-4

การประเมินค่าอัตโนมัติของ BLUE-4 ตั้งอยู่บนสมมติฐานที่ว่าหากผลลัพธ์การแปลมีส่วนของข้อความที่เกิดร่วมกันกับส่วนของข้อความในผลลัพธ์การแปลอ้างอิงเป็นจำนวนที่มากกว่าก็น่าเชื่อได้ว่าผลลัพธ์การแปลนั้นจะมีความถูกต้องสูง ดังนั้นการประเมินค่าอัตโนมัติของ BLEU –n (เรียกโดยย่อว่า BLEU) คือการคำนวณหาค่าเฉลี่ยเลขคณิตของสัดส่วนของคำที่ซ้อนทับ

กัน (overlapping) ของชุดคำตอบจำนวน n คำติดกันที่ปรากฏในชุดคำตอบอ้างอิง โดย BLEU -4 หมายถึง n มีค่าเท่ากับ 4 (คำนวณหาค่าเฉลี่ยเลขคณิตของสัดส่วนการซ้อนทับกันของคำที่ติดกัน 4 คำในชุดคำตอบกับชุดคำตอบอ้างอิง) การคำนวณค่า BLEU ถูกคำนวณตามสมการต่อไปนี้

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4.1}$$

โดย

BP (Brevity Penalty) คือ ค่าถ่วงน้ำหนักของชุดคำตอบที่สั้นกว่าชุดคำตอบอ้างอิง โดยคำนวณจากสมการต่อไปนี้

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases} \tag{4.2}$$

โดย c คือจำนวนคำของชุดคำตอบ

r คือจำนวนคำของชุดคำตอบอ้างอิง

N จำนวนคำที่เรียงติดกัน

หากต้องการคำนวณค่า BLUE-4 จะทำให้ $N = 4$

w_n คือ ค่าถ่วงน้ำน้ำหนักของคะแนนความแม่นยำ โดย $\sum_{n=1}^{N} w_n = 1$

หากใช้การถ่วงน้ำหนักแบบ Uniform Weight จะทำให้ $w_n = \frac{1}{N}$

p_n คือ คะแนนความแม่นยำ (Precision Score) โดยคำนวณจากสมการต่อไปนี้

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count(n-gram)} \tag{4.3}$$

อย่างไรก็ดี ค่า BLEU ที่ได้มักมีค่าที่ค่อนข้างน้อยมากๆ ดังนั้นจึงต้องเปลี่ยนเป็นหน่วย log เพื่อให้เปรียบเทียบคะแนนได้ง่ายขึ้น

$$\log BLEU = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n \log p_n \tag{4.4}$$

4.3 การเตรียมข้อมูลสำหรับการประเมินค่าการแปล

การทดลองนี้ใช้คลังข้อความแบบคู่ของตัวอย่างประโยคจากหนังสือ English by Examples จำนวน 100,804 ประโยค (ขอเรียกโดยย่อว่า "คลังข้อความแบบคู่") จะถูกใช้เป็นข้อมูลตั้งต้นเพื่อนำไปใช้ทดสอบความถูกต้องของการเรียนรู้ด้วยวิธีการ 10-fold Cross-Validation

การทดลองจะนำแต่ละส่วนที่แบ่งไว้ มาแบ่งเป็นชุดสอนและชุดทดสอบ โดยมีอัตราส่วนชุดสอนต่อชุดทดสอบเป็น 9 ต่อ 1 จากนั้นนำข้อมูลชุดสอนไปเรียนรู้และนำข้อมูลชุดทดสอบมาแบ่งออกเป็นเฉพาะภาษาต้นทางและปลายทาง โดยประโยคต้นทางที่ได้จากชุดทดสอบเรียกว่าชุดต้นทาง (Source Set) ประโยคปลายทางจากชุดทดสอบจะถูกใช้เป็นผลการแปลอ้างอิงหรือชุดคำตอบอ้างอิง (Reference Result Set) จากนั้นนำชุดต้นทางไปแปลให้เป็นภาษาปลายทางด้วยระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรม (jEBMT) ผลลัพธ์ที่แปลออกมาได้เรียกว่าชุดคำตอบ (Result Set) และทำเช่นเดียวกันนี้กับระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่อง "สุภาษิต" (Parsit)

ขั้นตอนสุดท้ายจะนำ ชุดต้นทาง ชุดคำตอบอ้างอิง และชุดคำตอบจากระบบทั้งสอง ไปใช้ทำการทดลองตามกรณีต่างๆ ต่อไป

4.4 ผลการทดลอง

การทดลองสามารถทำได้โดยประเมินค่าความถูกต้องของการแปล โดยจะแบ่งการประเมินออกเป็น 2 กลุ่ม โดยกลุ่มแรกจะประเมินค่าความถูกต้องของการแปลแบบอัตโนมัติ ซึ่งจะถูกแบ่งเป็น การประเมินค่าแบบอิงจำนวนการเข้าคู่แบบแม่นตรง (Exact Matching) และการประเมินค่าอัตโนมัติของ BLUE-4 กลุ่มที่สองจะประเมินค่าความถูกต้องของการแปลโดยมนุษย์ ประกอบด้วย ความถูกต้องของการเรียงลำดับคำ ความถูกต้องของการเลือกคำให้เหมาะสมตามบริบท และความถูกต้องของการแปลวลี

4.4.1 ประเมินค่าความถูกต้องของการแปลแบบอัตโนมัติ

4.4.1.1 ผลการทดลองของการเข้าคู่แบบแม่นตรง (Exact Matching)

การทดลองนี้ใช้เทคนิค 10-fold Cross-Validation เพื่อเทียบผลการแปลชุดคำตอบแต่ละชุดที่ได้จากแต่ละระบบกับชุดคำตอบอ้างอิง จากนั้นจึงนับจำนวนที่สามารถเข้าคู่แบบแม่นตรง (เหมือนกันทั้งหมด) ผลลัพธ์ที่ได้แสดงไว้ในตารางที่ 4.1

ตารางที่ 4.1 ตารางสรุปการเข้าคู่แบบแม่นตรง

ครั้งที่ของการการทำ 10-fold Cross-Validation	จำนวนที่สามารถเข้าคู่แบบแม่นตรง				
	Parsit		jEBMT		จำนวนชุดคำตอบที่นำมาเทียบ
	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	
1	279	2.77%	308	3.06%	10,081
2	239	2.37%	294	2.92%	10,081
3	223	2.21%	289	2.87%	10,081
4	261	2.59%	254	2.52%	10,081
5	237	2.35%	282	2.80%	10,080
6	233	2.31%	296	2.94%	10,080
7	276	2.74%	287	2.85%	10,080
8	247	2.45%	254	2.52%	10,080
9	239	2.37%	283	2.81%	10,080
10	238	2.36%	299	2.97%	10,080
ค่าเฉลี่ย	***2,472***	***2.45%***	***2,846***	***2.82%***	***100,804***

ผลการทดลองแสดงให้เห็นว่า jEBMT ให้ผลการเข้าคู่แบบแม่นตรงโดยเฉลี่ยดีกว่า Parsit ร้อยละ 15 เนื่องจากชุดคำตอบอ้างอิงถูกแปลโดยมนุษย์ทำให้การแปลแบบอิงตัวอย่างให้ผลลัพธ์ที่ใกล้เคียงกับการแปลของมนุษย์มากกว่าการแปลโดยใช้กฎ

4.4.1.2 ผลการทดลองการประเมินค่าอัตโนมัติของ BLUE-4

ในการทดลองนี้ได้ใช้ชุดเครื่องมือ NIST MT Evaluation kit [17] สำหรับประเมินค่าแบบอัตโนมัติของ BLEU-4 ค่าความถูกต้องของการแปลที่ได้จะแสดงไว้ในตารางที่ 4.2

ตารางที่ 4.2 ตารางสรุปผลการประเมินค่าแบบอัตโนมัติของ BLEU-4

ครั้งที่ของการการทำ 10-fold Cross-Validation	ค่าความถูกต้องของการประเมินค่าแบบอัตโนมัติของ BLEU-4	
	Parsit	jEBMT
1	0.0276	**0.0301**
2	0.0234	**0.0268**
3	0.0245	**0.0281**
4	0.0265	**0.0303**
5	0.0255	**0.0310**
6	**0.0267**	0.0266
7	**0.0273**	0.0255
8	0.0264	**0.0303**
9	0.0262	**0.0313**
10	0.0262	**0.0292**
ค่าเฉลี่ย	***0.0260***	***0.0289***

จากการทดลองพบว่า jEBMT สามารถแปลโดยเฉลี่ยได้ดีกว่า Parsit ร้อยละ 11 โดยอิงจากผลการทดลอง ปัญหาหลักที่ทำให้ผลการแปลจาก jEBMT ไม่ดีส่วนใหญ่เกิดจากการเรียงประโยคปลายทางที่ผิดพลาด ปัญหาเหล่านี้สามารถแก้ไขได้โดยการรวบรวมคลังข้อความแบบเดี่ยวสำหรับภาษาปลายทางที่มีเนื้อหาเกี่ยวข้องกับขอบเขตที่แปล

4.4.2 ประเมินค่าความถูกต้องของการแปลโดยมนุษย์

ในงานวิจัยชิ้นนี้ได้ทดสอบปัญหาพื้นฐานของระบบแปลภาษา 3 ชนิด อันได้แก่ ความถูกต้องของการเรียงลำดับคำ ความถูกต้องของการเลือกคำให้เหมาะสมตามบริบท ความถูกต้องของการแปลวลี การประเมินความถูกต้องทั้งหมดนี้กระทำโดยนักภาษาศาสตร์

การทดสอบทั้ง 3 ชนิดกระทำโดยสุ่มเลือกประโยคมา 200 ประโยค จำนวน 5 ชุด โดยแต่ละชุดมีข้อมูลไม่ซ้ำกัน จากนั้นจึงให้นักภาษาศาสตร์คนเดียวกันทำการประเมินโดยวิธีนับจำนวนประโยคที่ยอมรับได้ตามเงื่อนไขของแต่ละกรณี

4.4.2.1 ความถูกต้องของการเรียงลำดับคำ

เกณฑ์ในการทดสอบความถูกต้องของการเรียงลำดับคำ จะพิจารณาเฉพาะตำแหน่งที่เหมาะสมของคำแปล แต่อาจไม่ได้เลือกคำแปลที่มีความหมายเหมาะสมกับบริบท ตัวอย่างเช่น "experience in administration" อาจจะถูกแปลเป็น "ลิ้มลอง/ใน/การบริหาร/" จะเห็นได้ว่า "experience" ถูกแปลเป็น "ลิ้มลอง" ซึ่งไม่เหมาะสมกับบริบท แต่ตำแหน่งคำแปลอยู่ในตำแหน่งที่ถูกต้องเพียงแต่ไม่สามารถเลือกคำแปลที่เหมาะสมกับบริบทได้ ผลการทดลองถูกแสดงไว้ในตารางที่ 4.3

ตารางที่ 4.3 ตารางสรุปผลการทดลองความถูกต้องของการเรียงลำดับคำ

ครั้งที่ทดลอง	ค่าความถูกต้องของการเรียงลำดับคำ				
	Parsit		jEBMT		จำนวนที่ใช้ในการทดลอง
	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	
1	100	50.00%	56	28.00%	200
2	91	45.50%	50	25.00%	200
3	113	56.50%	49	24.50%	200
4	99	49.50%	61	30.50%	200
5	98	49.00%	55	27.50%	200
ค่าเฉลี่ย	***100.2***	***50.10%***	***54.2***	***27.10%***	***200***

ผลการทดลองเห็นได้ชัดว่า Parsit มีความถูกต้องในการแก้ไขปัญหาการเรียงลำดับคำโดยเฉลี่ยสูงกว่า jEBMT อย่างเห็นได้ชัดถึงร้อยละ 84 เหตุที่เป็นเช่นนี้สืบเนื่องจากค่าสถิติของตัวแบบเอ็นแกรม มีความเบาบาง (sparse) เกินกว่าจะนำมาตัดสินความเป็นส่วนประชิด (constituent) ของผลการแปลในภาษาปลายทางได้

4.4.2.2 ความถูกต้องของการเลือกคำให้เหมาะสมตามบริบท

เกณฑ์ในการทดสอบความถูกต้องของการเลือกคำให้เหมาะสมตามบริบทจะพิจารณาเฉพาะคำแปลที่เหมาะสมกับประโยคต้นฉบับเท่านั้น ไม่จำเป็นต้องมีการเรียงลำดับคำที่ถูกต้อง ตัวอย่างเช่น "close an account at the bank in her name" อาจจะถูกแปลเป็น "ปิดบัญชี/ณ/ใน/ชื่อ/ธนาคาร/เธอ/" จะเห็นได้ว่าทุกคำมีการเลือกคำที่มีความหมายถูกต้องแต่การเรียงลำดับของคำไม่ถูกต้อง ผลการทดลองแสดงไว้ในตารางที่ 4.4

ตารางที่ 4.4 ตารางสรุปผลการทดลองความถูกต้องของการเลือกคำให้เหมาะสมตามบริบท

ครั้งที่ทดลอง	ค่าความถูกต้องของการเลือกคำให้เหมาะสมตามบริบท				
	Parsit		jEBMT		จำนวนที่ใช้ในการทดลอง
	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	
1	88	44.00%	124	62.00%	200
2	100	50.00%	112	56.00%	200
3	92	46.00%	121	60.50%	200
4	95	47.50%	106	53.00%	200
5	87	43.50%	111	55.50%	200
ค่าเฉลี่ย	***92.4***	***46.20%***	***114.8***	***57.40%***	***200***

ผลการทดลองแสดงให้เห็นว่า jEBMT มีความถูกต้องในการเลือกใช้คำให้เหมาะสมตามบริบทระดับประโยคมากกว่า Parsit พอสมควรในอัตราร้อยละ 24 โดยเฉลี่ย ทั้งนี้เป็นเพราะค่าสถิติของตัวแบบเอ็นแกรมสามารถนำมาช่วยแก้ไขความกำกวมในการเลือกคำแปลตามบริบทได้แม้ว่าข้อมูลจะมีปริมาณเบาบางก็ตาม

4.4.2.3 ความถูกต้องของการแปลวลี

เกณฑ์ในการทดสอบความถูกต้องของการแปลวลี จะพิจารณาเฉพาะคำแปลของวลีที่ปรากฏในประโยคต้นฉบับเท่านั้น ตัวอย่างเช่น "I have an account with the bank" อาจจะถูกแปลเป็น "ฉัน/มี/บัญชีเงินฝาก ที่ ธนาคาร/" จะเห็นได้ว่า "an account with the bank" ถูกแปลได้อย่างถูกต้องเป็น "บัญชีเงินฝาก ที่ ธนาคาร" ผลการทดลองแสดงไว้ในตารางที่ 4.5

ตารางที่ 4.5 ตารางสรุปผลการทดลองความถูกต้องของการแปลวลี

ครั้งที่ทดลอง	ค่าความถูกต้องของการแปลวลี				
	Parsit		jEBMT		จำนวนที่ใช้ในการทดลอง
	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	จำนวนที่เข้าคู่	ร้อยละการเข้าคู่	
1	56	28.00%	108	54.00%	200
2	60	30.00%	112	56.00%	200
3	59	29.50%	101	50.50%	200
4	49	24.50%	99	49.50%	200
5	54	27.00%	97	48.50%	200
ค่าเฉลี่ย	***55.6***	***27.80%***	***103.4***	***51.70%***	***200***

ผลการทดลองแสดงให้เห็นอย่างชัดเจนว่า jEBMT มีความถูกต้องในการแปลวลีมากกว่า Parsit เป็นอย่างมากถึงร้อยละ 85 โดยเฉลี่ย ทั้งนี้เป็นเพราะ jEBMT สามารถรู้จำ (recognize) และแปลวลีที่มีส่วนประชิดต่อเนื่องกัน (continuous constituent) จากคลังข้อความแบบคู่โดยใช้ค่าสถิติของตัวแบบเอ็นแกรมได้เป็นอย่างดีแม้ว่าจะมีปริมาณข้อมูลเบาบางก็ตาม

4.5 สรุปผลการประเมินค่าการแปล

การใช้ตัวแบบเอ็นแกรมสามารถแก้ปัญหาการเลือกใช้คำให้เหมาะสมตามบริบทระดับประโยคและการแปลวลีได้เป็นอย่างดี ในขณะที่การประยุกต์ใช้ตัวแบบเอ็นแกรมกับการแก้ไขปัญหาการเรียงลำดับคำยังคงต้องการปริมาณข้อมูลเชิงสถิติเพิ่มอีกเป็นจำนวนมาก ทั้งนี้เป็นเพราะต้องตรวจสอบความเป็นส่วนประชิดซึ่งมีความหลากหลายในการสร้าง (diversity of formation) เป็นอย่างมาก

ผลการประเมินค่าโดยรวมสามารถสรุปได้ว่าระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างมีความสามารถแปลได้ในระดับหนึ่ง โดยผลการแปลส่วนใหญ่จะสามารถจับใจความได้ แต่ยังไม่ถูกต้องตามหลักไวยากรณ์ รูปแบบผลการแปลสามารถแบ่งได้เป็นประเภทต่างๆ ดังนี้

1. ไม่สามารถจับใจความได้ (แทนด้วยสัญลักษณ์ ☹)
2. สามารถอ่านจับใจความสำคัญได้ แต่ไวยากรณ์ไม่ถูกต้อง (แทนด้วยสัญลักษณ์ ☺)
3. สามารถอ่านจับใจความสำคัญได้ ไวยากรณ์ถูกต้อง แต่ไม่ตรงกับคำตอบอ้างอิง (แทนด้วยสัญลักษณ์ ☺☺)
4. สามารถอ่านจับใจความสำคัญได้ ไวยากรณ์ถูกต้อง และตรงกับคำตอบอ้างอิง (แทนด้วยสัญลักษณ์ ☺☺☺)

<u>ตัวอย่างการแปลพร้อมบทวิเคราะห์</u>

ข้อความรับเข้า:	for mutual advantage
คำตอบอ้างอิง:	เพื่อ ประโยชน์ของ ทั้งสองฝ่าย
คำตอบจาก Parsit:	สำหรับ ผล ประโยชน์ สอง ฝ่าย
คำตอบจาก jEBMT:	สำหรับ ซึ่ง กันและกัน ความ ได้ เปรียบ
สรุป:	Parsit ☺ jEBMT ☹
วิเคราะห์:	คำตอบจาก jEBMT มีการเรียงลำดับคำของคำแปลที่ผิดพลาด ทำให้ไม่สามารถจะเข้าใจใจความสำคัญของประโยคต้นทางได้ คำตอบถูกต้องควรเป็น สำหรับ ความ ได้ เปรียบ ซึ่ง กันและกัน ที่เป็นเช่นนี้เกิดจากค่าสถิติของการประกอบส่วนประชิดมีความเบาบางและไม่มากพอที่จะทำให้การเรียงลำดับคำถูกต้อง

ข้อความรับเข้า:	a superior in age
คำตอบอ้างอิง:	ผู้ มี วัยวุฒิ สูง
คำตอบจาก Parsit:	กว่า ผู้ใหญ่ ใน อายุ
คำตอบจาก jEBMT:	ผู้ที่ มีอายุ สูง กว่า
สรุป:	Parsit ☹ jEBMT ☺
วิเคราะห์:	คำตอบจาก Parsit มีการสังเคราะห์ประโยคที่ผิดพลาด คำตอบจาก jEBMT มีความถูกต้องเนื่องจากคำเหล่านี้ คือ “มีอายุ” และ

คำว่า “สูงกว่า” เคยมีการปรากฏแบบประชิดร่วมกันในคลังข้อความ ทำให้ค่าสถิติบ่งชี้ให้เกิดการเรียงลำดับที่ถูกต้อง

ข้อความรับเข้า: reach an agreement

คำตอบอ้างอิง: ได้ ข้อตกลง

คำตอบจาก Parsit: เข้าถึง ข้อ ตกลง

คำตอบจาก jEBMT: บรรลุ ข้อตกลง

สรุป: Parsit ☹ jEBMT ☺☺

วิเคราะห์: คำตอบที่ได้จาก Parsit ใช้คำว่า เข้าถึง ซึ่งมีใจความสำคัญไม่ตรงกับประโยคต้นทาง ในขณะที่คำตอบจาก jEBMT สามารถแปลได้ดีกว่าคำตอบอ้างอิง แต่ใช้คำไม่ตรงกับคำตอบอ้างอิง ซึ่งแสดงให้เห็นว่าตัวอย่างภายในคลังข้อความแบบคู่ให้ผลลัพธ์การแปลที่เหมาะกว่าการแปลโดยการใช้กฎของ Parsit

ข้อความรับเข้า: have a high aim in life

คำตอบอ้างอิง: มี เป้าหมาย ที่ สูง ใน ชีวิต

คำตอบจาก Parsit: มี จุด มุ่งหมาย ที่ สูง ใน ชีวิต

คำตอบจาก jEBMT: มี สูง ใน ชีวิต เป้า

สรุป: Parsit ☺☺ jEBMT ☹

วิเคราะห์: คำตอบจาก Parsit สามารถแปลได้ดีกว่าคำตอบอ้างอิง แต่ใช้คำไม่ตรงกับคำตอบอ้างอิงแต่ยังคงความหมายถูกต้อง

คำตอบจาก jEBMT มีการเรียงลำดับคำของคำแปลที่ผิดพลาด ทำให้ไม่สามารถจะเข้าใจความหมายที่แท้จริงของประโยคต้นทางได้ ประโยคที่ถูกควรเป็น มี เป้าสูง ใน ชีวิต ที่เป็นเช่นนี้เกิดจากค่าสถิติของการประกอบส่วนประชิดมีความเบาบางและไม่มากพอที่จะทำให้การเรียงลำดับคำถูกต้อง

ข้อความรับเข้า: the correct answer

คำตอบอ้างอิง: คำตอบ ที่ ถูกต้อง

คำตอบจาก Parsit: คำ ตอบ ที่ ถูก ต้อง

คำตอบจาก jEBMT: ดัด นิสัย ตอบ

สรุป: Parsit ☺☺☺ jEBMT ☹

วิเคราะห์: คำตอบจาก Parsit ถูกต้องและเหมือนคำตอบอ้างอิง

คำตอบจาก jEBMT เลือกคำผิดพลาดจากส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรม โดยแปลคำว่า the correct เป็น ดัด นิสัย เนื่องจากภายในระบบมีการตัดคำบ่งชี้เฉพาะ (article เช่น a, an, the) ทำให้ the correct ถูกตัดทอนเหลือ correct ซึ่งในคลังข้อความมีค่าสถิติของ correct ที่อยู่ต้นประโยคเป็นคำกริยา ทำให้ correct ของข้อความรับเข้าถูกเลือกความหมายเป็น ดัด นิสัย

ข้อความรับเข้า: retire from the army

คำตอบอ้างอิง: ปลดเกษียณ จาก กองทัพ

คำตอบจาก Parsit: ลา ออก กองทัพ

คำตอบจาก jEBMT: ปลดประจำการ จาก ทหารบก

สรุป: Parsit ☹ jEBMT ☹

วิเคราะห์: คำตอบจาก Parsit และ jEBMT ต่างมีใจความสำคัญที่ไม่ถูกต้องซึ่งเป็นปัญหาที่เรียกว่าปัญหาในการเลือกคำที่เหมาะสมกับบริบท เนื่องจาก retire ควรมีความหมายเป็น “เกษียณอายุ” ซึ่งมีความที่แตกต่างจาก ลา ออก ซึ่งเป็นคำตอบจาก Parsit และ ปลดประจำการ ซึ่งเป็นคำตอบจาก jEBMT

จากการวิเคราะห์ผลการแปลพบว่าระบบ jEBMT สามารถแปลวลีและสำนวนได้อย่างถูกต้องแม่นยำมากกว่า Parsit ในขณะที่ Parsit สามารถรองรับการแปลตามกฎไวยากรณ์ได้ดีกว่า jEBMT สาเหตุของผลการแปลดังกล่าวเกิดจากปัจจัยสำคัญ 2 ประการ ได้แก่ สภาพการเกาะกลุ่มของวลีและสำนวน (Clustering of Phrase and Idioms) และความเบาบางของข้อมูลที่ใช้สอนระบบ (Data Sparseness)

สภาพการเกาะกลุ่มของวลีและสำนวนเป็นปัจจัยสำคัญที่ทำให้ส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรมสามารถช่วยปรับปรุงผลการแปลได้ จากการทดลองพบว่า วลีและสำนวนมักอยู่ติดกันภายใน 3 คำตามตำแหน่งของคำ มีเพียงวลีและสำนวนส่วนน้อยที่อยู่แยกจากกันเกิน 3 คำ ทำให้ผลการแปลวลีและสำนวนของ jEBMT จึงดีกว่าของ Parsit

ความเบาบางของข้อมูลเป็นปัจจัยที่ทำให้ผลความถูกต้องของการแปลของ jEBMT ลดต่ำลง การอนุมานกฎไวยากรณ์จากคลังข้อความจำเป็นต้องใช้ข้อมูลจำนวนมาก Keh-Yih Su, Tung-Hui Chiang and Jing-Shin Chang [17] กล่าวว่า จำนวนประโยคตัวอย่างสำหรับสร้างตัวแบบเอ็นแกรมสำหรับภาษาใดๆ ที่มีจำนวนคำในภาษา w คำ จะต้องมีอย่างน้อย $10w^3$ ตัวอย่าง เนื่องจากคำในภาษาอังกฤษและภาษาไทยมีจำนวนมากในระดับมากกว่าแสนคำ คลังข้อความที่ใช้จึงไม่เพียงพอต่อการสร้างตัวแบบเอ็นแกรมของภาษา ขณะที่ Parsit ซึ่งใช้อิงกฎไวยากรณ์ในการแปลที่สร้างจากมนุษย์โดยตรง จึงทำงานได้ดีกว่า jEBMT ในการแปลประโยคโดยใช้กฎ

บทที่ 5
สรุปการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

วิทยานิพนธ์ฉบับนี้ได้นำเสนอระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่างโดยใช้ตัวแบบเอ็นแกรม ระบบนี้เป็นระบบแรกสำหรับระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่าง ผลการแปลโดยเฉลี่ยดีกว่าเมื่อเปรียบเทียบกับระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงกฎไวยากรณ์ "ภาษิต"

การใช้ตัวแบบเอ็นแกรมสามารถแก้ปัญหาการเลือกใช้คำให้เหมาะสมตามบริบทระดับประโยคและการแปลวลีได้เป็นอย่างดี ในขณะที่การประยุกต์ใช้ตัวแบบเอ็นแกรมกับการแก้ไขปัญหาการเรียงลำดับคำยังคงต้องการปริมาณข้อมูลเชิงสถิติอีกเป็นอย่างมาก ทั้งนี้เป็นเพราะต้องตรวจสอบความเป็นส่วนประชิดซึ่งมีความหลากหลายในการสร้าง (diversity of formation) เป็นอย่างมาก

เปรียบเทียบกับระบบแปลภาษาด้วยเครื่องแบบอิงกฎไวยากรณ์ (Rule-based MT) ระบบจะสามารถแปลกลุ่มประโยคที่เป็นสำนวนที่มีลักษณะคำที่ต่อเนื่องกันได้ดี

เปรียบเทียบกับระบบแปลภาษาด้วยเครื่องแบบอิงสถิติ (Statistical-based MT) ระบบจะมีความต้องการใช้ข้อมูลของคลังข้อความแบบคู่น้อยกว่า เนื่องจากระบบมีการใช้คลังข้อความแบบเดี่ยวสำหรับภาษาต้นทางและคลังข้อความแบบเดี่ยวสำหรับภาษาปลายทาง ซึ่งสามารถรวบรวมได้โดยง่าย

5.2 ข้อเสนอแนะ

ส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรม (*n*-gram analysis Component) ยังไม่มีการจำกัดขนาดสูงสุดของ *n*-gram ทำให้ใช้เวลามากในการเข้าคู่ของข้อมูลในคลังข้อความคู่ภาษาสำหรับประโยคต้นทางที่ยาว หากเราสามารถกำหนดขนาดสูงสุดของ n-gram ได้จะทำให้การเข้าคู่ทำได้เร็วขึ้น อย่างไรก็ดีขนาดสูงสุดที่เหมาะสมของ *n*-gram สำหรับภาษาต้นทางยังต้องมีการทำวิจัยต่อไป

ส่วนโปรแกรมการก่อกำเนิดแบบเอ็นแกรม (*n*-gram generation Component) ใช้วิธีการแบบละโมบ (Greedy Algorithm) มาตรวจจับส่วนของประโยคที่ดีที่สุดที่สามารถนำมาต่อกันได้ ควรนำวิทยาการศึกษาสำนึก (heuristics) มาใช้เพิ่มเติม

เนื่องจากแบบจำลองทางสถิติที่ใช้เป็นแบบเอ็นแกรม ทำให้ไม่สามารถประมวลผลส่วนของข้อความที่ไม่ติดกันได้ ทำให้ผลการแปลที่เกิดขึ้นสำหรับกรณีเช่นนี้ไม่ดีนัก แนวทางที่

เหมาะสมสำหรับแก้ปัญหานี้คือควรจะใช้วิธีการแปลแบบอิงแม่แบบ (Template-based Translation) มาช่วยประยุกต์ใช้ในงานนี้

เนื่องจากขอบเขตของระบบนี้สามารถแปลได้เฉพาะคำที่รู้จักจากคลังข้อความเท่านั้น ไม่สามารถรู้จำและแปลคำระบุชื่อเฉพาะ (Name Entity) ได้ ทำให้ผลการแปลคำระบุชื่อเฉพาะที่มีรูปแบบบางประเภทที่พบได้โดยทั่วไปไม่ถูกต้อง อันได้แก่ นิพจน์เชิงตัวเลข (Numeral Expression) เช่น 1,234,567.89 นิพจน์เชิงเวลา (Date/Time Expression) เช่น 23:15, 9.34 a.m., 5-Dec-2006, 13-12-2006, 9/22/2001 นิพจน์เงินตรา (Currency Expression) เช่น $20,000 นิพจน์นับได้ (Countable Expression) เช่น 5 coins นิพจน์เศษส่วน (Fraction Expression) เช่น one-ninth, two ninths, three-ninths เป็นต้น ควรมีการเพิ่มเติมในส่วนนี้เพื่อเพิ่มความสมบูรณ์ของระบบ

ด้วยทรัพยากรทางด้านคลังข้อความสำหรับภาษาไทยที่มีจำกัดและไม่ได้เจาะจงเฉพาะขอบเขต (Domain) ใดขอบเขตหนึ่ง ทำให้เกิดปัญหาที่เรียกว่าข้อมูลเบาบาง (Data Sparseness) ทำให้แบบจำลองทางสถิติที่ใช้เป็นแบบเอ็นแกรมให้ผลที่ไม่ครอบคลุมกับงานแปลในสภาพแวดล้อมที่แท้จริง ดังนั้นหากเราสามารถรวบรวมคลังข้อความที่ความเฉพาะเจาะจงของขอบเขตใดขอบเขตหนึ่งได้น่าเชื่อได้ว่าจะทำให้ผลการแปลดีขึ้นมาก

เอกสารอ้างอิง

[1] Slocum J. "Machine Translation Systems." **Cambridge University Press.** 1998.

[2] Nirenburg S. "Machine Translation: Theoretical and Methodological Issues." **Cambridge University Press.** 1987.

[3] Lenhrberger J. and L.Boubeau. "Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation." Philadelphia : Benjamins. 1988.

[4] Mckeown K.R. and W.R. Swartout .. "Lanugage generation and explanation." Annual Review of Computer Science, Volume 2. 1987

[5] Mcdonald D.D. and L. Bolc.. "Natural Lanugage Generation Systems." Springer-Verlag. 1988.

[6] นิตยา กาญจนะวรรณ. "รายงานการวิจัยเรื่องการศึกษาโครงสร้างภาษาไทยเพื่อใช้ในระบบการแปลภาษาด้วยเครื่องคอมพิวเตอร์." คณะมนุษย์ศาสตร์ มหาวิทยาลัยรามคำแหง. 2529.

[7] Makoto N. "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle." **Artificial and Human Intelligence,** North Holland, 1984. pp. 173-180.

[8] Satoshi S. and Makoto N. "Toward Memory based Translation." **In Proceedings of the 13th COLING,** 1990. pp. 247-252.

[9] Eiichiro S., Hitoshi I. and Hideo K. "Translating with examples: A new approach to Machine Translation." **In Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language,** Linguistics Research Centre, University of Texas at Austin, USA, 1990. p.p. 203-212.

[10] Eiichiro S. and Hitoshi I.. "Experiments and prospects of example-based Machine Translation." **In Proceedings of 29th ACL,** University of California, Berkeley, Ca., 1991. p.p. 185-192.

[11] Konstantin M., Satoshi N., Hiromi N., Hisashi K., Takatoshi J., Zhin-Song Z., Hirofumi Y., Genichiro K. and Seiichi Y. "The ATR Multi-lingual Speech-To-Speech Translation System." **IEEE Transactions on Audio, Speech, and Language Processing,** vol.14, no.2, 2006. pp.365-376.

[12] HUTCHINS J., "Towards a definition of example-based machine translation." **In Proceedings of EBMT Workshop of MT Summit X,** 2005, pp. 63-70.

[13] Sornlertlamvanich V., Chareonpornsawat P. and Boriboon M. "ParSit: Online English-Thai Machine Translation Service, Language Issues." **Digital Publishing.** 2001. 31(3) : 6-7.

[14] Sornlertlamvanich V., Charoenpornsawat P., Boriboon M. and Boonmana L.. "ParSit: English-Thai Machine Translation Services on Internet." **In 12th Annual Conference,** ECIT and New Economy, National Electronics and Computer Technology Center. Bangkok: Thailand. 2000.

[15] Dirix P., Schuurman I., and Vandeghinste V. "METIS-II: Example-based machine translation using monolingual corpora – System description." **In Proceedings of EBMT Workshop of MT Summit X,** 2005. pp. 43-50.

[16] Papineni K., Roukos S., Ward T. and Zhu W.J. "BLEU: a Method for Automatic Evaluation of Machine Translation." **in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002),** Philadelphia, PA, July 2002.

[17] NIST. **"NIST MT evaluation kit."** [Online]. Available: http://www.nist.gov/speech/tests/mt/resources/scoring.htm. 2001.

[18] Su K., Chiang T. and Chang J. "An Overview of Corpus-Based Statistics-Oriented (CBSO) Techniques for Natural Language Processing." **in International Journal of Computational Linguistics & Chinese Language Processing (CLCLP-1996),** vol. 1, no. 1, August 1996. pp. 100-156.

ภาคผนวก

งานวิจัยที่ได้รับการตีพิมพ์

1. N. Kritsuthikul, A. Thammano, and T. Supnithi, **"English-Thai Example-Based Machine Translation using n-gram model,"** 2006 IEEE International Conference on Systems, Man, and Cybernetics (IEEE-SMC-2006), Conference Digest, p. 193 , The Grand Hotel, Taipei, Taiwan, October 8-11, 2006.

English-Thai Example-Based Machine Translation using *n*-gram model

Nattapol KRITSUTHIKUL, Arit THAMMANO, and Thepchai SUPNITHI

***Abstract* — The necessity on exchanging information among countries become a major task in information based society. Machine translation is an application that enables users to communicate each other without language barrier problem. With the great support on computer's efficiency, corpus-based technology becomes a fundamental concept for developing software based on a large amount of data. We introduce the first example-based English to Thai machine translation using *n*-gram model and implemented the system. Some advantages and disadvantages of this method are discussed.**

I. Introduction

THERE are several approaches on Machine Translation (MT) research. The most of distinctive methods are rule-based [1] and corpus-based methods. Research on the corpus-based approach has emphasized on the important of text corpora as fundamental sources of data for linguistic and knowledge database. There have been two major approaches in the corpus-based MT: statistical-based approach [2] and example-based approach [3]. Currently, a lot of English to Thai machine translation software products, such as ParSit [4], AgentDict [5], were launched to end users. All of them use the rule-based approach, whose all knowledge from linguists is externalized as a set of inference rules. This approach has several drawbacks related to time consumption and rule conflict. Then, corpus-based MT becomes much more interesting topics in NLP research field. However, there are very few researches on corpus-based MT in Thai. Some research groups try to apply corpus as a memory for editing the incorrect answer from rule based results [6]. In this paper, we concentrate on corpus based MT in example-based approach. We generate the possible patterns by applying *n*-gram model.

This paper is organized as follows. Section 2 focuses on our system architecture, all components in the system, and techniques in analyzing and generating sentences. Section 3 mainly explains the corpus used in this system. Section 4 illustrates the implementation design and tool. Finally, section 5 shows a conclusion and future work.

Nattapol KRITSUTHIKUL and Arit THAMMANO, are with the Computational Intelligence Laboratory, Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Chalongkrung Road, Bangkok 10520, Thailand (e-mail: s4467499@kmitl.ac.th and arit@it.kmitl.ac.th)

Nattapol KRITSUTHIKUL and Thepchai SUPNITHI, is with the Research and Development in Information Division, National Electronic and Computer Technology Center, Thailand (e-mail: thepchai@nectec.or.th)

II. System Architecture

The system architecture is shown in Figure 1. Our system composes of two main components: analysis component and generation component. We apply the *n*-gram approach to achieve varieties of patterns based on partial sentences. When a system receives a source sentence, it will be passed to match with data in our prepared bilingual parallel corpus. All appropriate matching alternatives (sentences or partial sentences) will be retrieved and sent to find the most appropriate target sentences. Monolingual corpus will help analyzing and generating an appropriate alternatives in each component. Bilingual corpus composes of parallel sentences and parallel partial sentences. All data in this corpus functions as transfer rules. In the current version, we collect partial sentences manually from linguists.

A. The n-gram approach

The *n*-gram language model is based on the following assumption: the n^{th} word is only related to its preceding *n*-1 words. Therefore the probability estimation of the language model $P(w)$ can be written as $P(w_n|w_1,\ldots,w_{n-1})$. For a sentence with n words, given the candidates $w_1,w_2,\ldots,w_n$ from the bilingual corpus, the probability of the whole sentence is calculated by the equation (1).

$$P(w)=\sum_{i=1}^{N}P(w_i\,|w_{i-n+1},\ldots,w_{i-1}) \qquad (1)$$

For a large amount of text corpora, the probability of $P(w_n|w_1,\ldots,w_{n-1})$ can be estimated from the maximum likelihood principle by the equation (2):

$$P(w_n\,|w_1,\ldots,w_{n-1})=\frac{C(w_1,\ldots,w_n)}{C(w_1,\ldots,w_{n-1})} \qquad (2)$$

where $C(w_1,\ldots,w_{n-1})$ and $C(w_1,\ldots,w_n)$ represent the occurrence number of the word string $w_1,\ldots,w_{n-1}$ and $w_1,\ldots,w_n$ respectively.

B. N-gram analysis component

The *n*-gram analysis component is aimed to analyze a source sentence by breaking it into all possible pieces of sentences and to select the most suitable alternative. The main function in the analysis component is matching algorithm. Matching

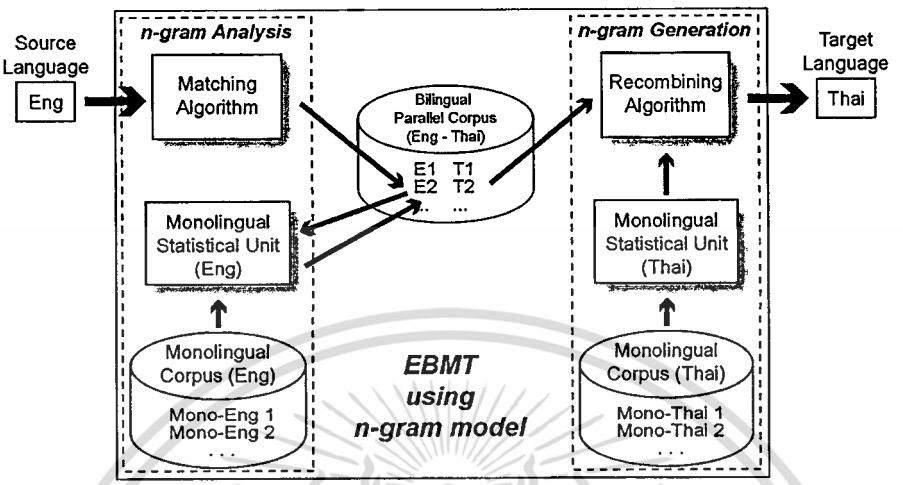


Fig.1. System Overview of EBMT using n-gram model

algorithm is applied to break a sentence into a combination of sub-sentences that composes of the longest sub-sentence. The longest sub-sentence is defined by matching with all data in bilingual corpus. It will lead us to find the least number of fragmentations in the sentence.

The matching algorithm is defined as follows.

Step 1: Define *Fragment_Set* = *{SL}* and *Result_Set* = *{ }*
Step 2: Generate sub-fragment Sf_i from *Fragment_Set* by segmenting groups of words that $next(w_i) \neq w_{i+1}$
Step 3: For each Sf_i that has more than one elements, find the maximum sub-sentence Max_Sf_i in Sf_i
Step 3.1: Push Max_Sf_i into *Result_Set*
Step 3.2: Delete Max_Sf_i from *Fragment_Set*
Step 4: Repeat Step 1 until each sub-fragment Sf_i has only one element
Step 5: Return Sf_i
Step 6: Return *Result_Set*

From the example sentence "*Arsenal picked up a big victory in Champions League*", we assume that there are *{a big victory}*, *{picked up}* and *{Champions League}* in our bilingual corpus. The matching algorithm will be processed as follows.

Step 1: *Fragment_Set* = *{ {Arsenal picked up a big victory in Champions League} }*, *Result_Set* = *{ }*
Step 2: Since $next(w_i) = w_{i+1}$ for all w_i, sub-fragment Sf_1 has only one sub-fragment that is Sf_1 = *{Arsenal picked up a big victory in Champions League}*
Step 3: maximum sub-sentence Max_sf_1 = *{a big victory}*
Step 3.1: *Result_Set* = *{ {a big victory} }*
Step 3.2: *Fragment_Set* = *{ {Arsenal picked up}, {in Champions League} }*
Step 4: *Fragment_Set* = *{ {Arsenal picked up}, {in Champions League} }*, *Result_Set* = *{ {a big victory} }*
Step 5: Since $next(w_i) \neq w_{i+1}$ at w_i= up, there are two sub-fragments, Sf_1 = *{Arsenal picked up}* and Sf_2 = *{in Champions League}*
Step 6: maximum sub-sentence Max_sf_1 = *{picked up}*
Step 6.1: *Result_Set* = *{ {picked up}, {a big victory} }*
Step 6.2: *Fragment_Set* = *{ {Arsenal}, {in Champions League} }*
Step 7: maximum sub-sentence Max_sf_2 = *{in Champions League}*
Step 7.1: *Result_Set* = *{ {Champions League}, {picked up}, {a big victory} }*
Step 7.2: *Fragment_Set* = *{ {Arsenal}, {in} }*
Step 8: Since $next(w_i) \neq w_{i+1}$ at w_i= Arsenal, there are two sub-fragments Sf_1 = *{ Arsenal }* and Sf_2 = *{ in }*
Step 9: Return *Result_Set* = *{ {Arsenal}, {in}, {Champions League}, {picked up}, {a big victory} }*

Results from the *n*-gram analysis will be translated to Thai based on two criteria. If an element in *Result_Set* is not a singleton, translated results in bilingual corpus will be retrieved. Otherwise, the translated results will be retrieved from dictionary. Translated results from *Result_Set* will be sent to the *n*-gram generation. From the example, *Result_Set* = *{ {Arsenal}, {in}, {Champions League}, {picked up}, {a big victory} }* will be translated to

{ {เข้าเป็นใหญ่อาร์เซนอล}.{ใน}, {แชมเปี้ยนลีก}.{ได้}, {ชัยชนะครั้งใหญ่} }

C. *N-gram generation component*

The *n*-gram generation component is aimed to generate a target sentence by merging and ordering pieces of sentences into one sentence. The main function in the generation component is recombining algorithm. We apply Greedy Algorithm to detect the most suitable sub-sentences that should be concatenated. Recombining algorithm merges sub-sentences into a target sentence by considering the word ordering in sentence. The algorithm is explained as follows:

Step 1: Define $Fragment\_List = \{Fr_1, Fr_2, \ldots, Fr_n\}$

Step 2: [combine the maximum probability of sub-sentence and its neighbor]
For each Fr_a and Fr_b in *Fragment_List* that $1 \le a, b \le n$ and $|a-b| = 1$,
Fr_{ab} = max_combine (Fr_a , Fr_b)

Step 3: Substitute Fr_a and Fr_b with Fr_{ab} and delete Fr_a and Fr_b from *Fragment_List*
$Fragment\_List = \{Fr_1, Fr_2, \ldots, Fr_{ab}, \ldots, Fr_n\}$

Step 4: Repeat Step 1 until *Fragment_List* has only one element

Step 5: Return *Fragment_List*

For the above example, the output *Result_Set* in the *n*-gram analysis component is *{ {Arsenal}, {in}, {Champions League}, {picked up}, {a big victory} }* . We obtain the result

Fragment_List = { เข้าเป็นใหญ่อาร์เซนอล .ใน .แชมเปี้ยนลีก, ได้, ชัยชนะครั้งใหญ่ }

When we apply the recombining algorithm, the *Fragment_List* will be merged in step 2 by considering each sub-sentence and its neighbor (at distance 1) as shown step-by-step result in figure 2. The gray highlight identifies a pair of word/phrase that will be combined for the next step. Finally, "เข้าเป็นใหญ่อาร์เซนอลได้ชัยชนะครั้งใหญ่ในแชมเปี้ยนลีก" will be generated as translation result.

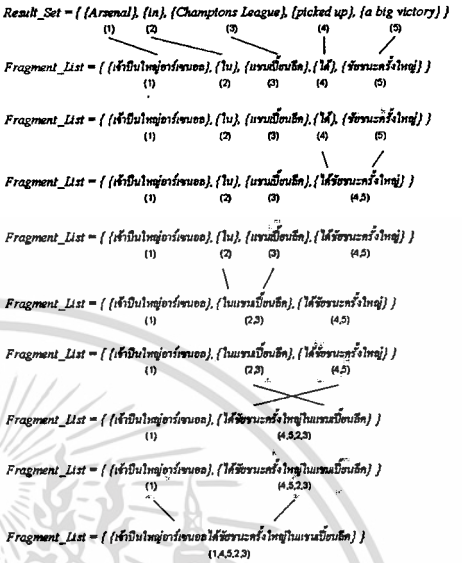


Fig. 2. *n*-gram generation component by example

III. CORPUS IN EBMT SYSTEM

In our EBMT approaches, it is necessary to prepare a large amount of parallel corpus in order to find a good matching. One of the major problems is time consumption for collecting a large amount of parallel corpus. We decide to apply monolingual corpus in each component to compensate the inadequate parallel corpus. Monolingual corpora in English and Thai for analysis phase and generation phase is much easier to collect. Since there is no word boundary in Thai, we segment sentences into words by using word segmentation tools called SWATH [7]. The occurrences of *n*-gram in each corpus are counted as statistic information and applied in the matching algorithm and the recombining algorithm. In bilingual corpus, we collect bilingual corpus based on our previous work in general domain. Sub-sentences pairs are manually constructed to increase the opportunity of matching in the analysis phase.

Currently, there are 104,893 sentences and 561,387 words in our English corpus, 16,749 sentences and 84,292 words in our Thai corpus, respectively. In parallel corpus, we collect 64,990 sentence pairs. There are 218,144 words in English and 182,243 words in Thai. The number of sentence and sub-sentence pairs is 91,068.

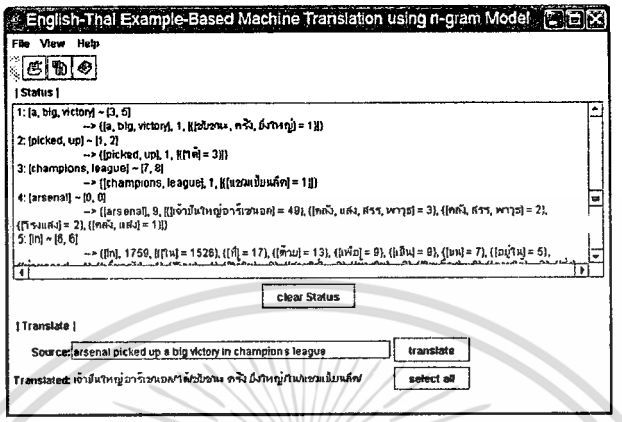


Fig. 3. EBMT using *n*-gram model application main screen

IV. Implementation

Figure 3 shows a screen of our EBMT system. When a user inputs a source sentence in the **Source** textbox and press the **translate** button. Translation process will be started. All translation processes, starting from the analysis phase to the generation phase, will be represented in the **Status** text area. The translated result will be shown in the **Translated** textbox. To increase the speed performance, all necessary information for translation, both from our corpora and our dictionary, are loaded in memory. We develop the *Trie* structure [8] at the word level to represent data in memory. This system is implemented in JAVA 5 platform (current version is jdk1.5.0_07).

V. Discussion

Example-based machine translation is a powerful technique that applies corpus-based approach. To translate from English to Thai, we investigated results and found that our methods have advantages on the complex sentence with exact partial phrase, if they were already collected in our bilingual corpus. Moreover, translation results of some metaphoric sentences and proverbs are also acceptable. However, there are some drawbacks in our method. Firstly, we need to collect partial sentence in parallel corpus as much as possible to avoid misleading results. Secondly, if suitable results cannot be found in the parallel corpus, the system will alternatively retrieve the answer from dictionary. It might lead to incorrect result because of word ambiguity and word omission. For example, a word "*the*" has no suitable translation result in Thai and will be omitted. Unfortunately, it will be translated to "*คำนำนามที่เฉพาะ*" which means "the definite article preceding proper nouns".

VI. Conclusion and Future Work

This paper explains an *n*-gram-based EBMT system for English to Thai machine translation. The translation results can be achieved by applying *n*-gram model technique in combination with information from corpus and dictionary. In the future work, three main issues will be considered. Firstly, since it is difficult to develop sub-sentences manually, we plan to analyze the possible patterns in bilingual paralleled corpus and develop an application for generating sub-sentence automatically. Secondly, If there exist more than one alternatives for each sub-sentence, we will select the most appropriate alternatives by considering the statistical information from the SL monolingual corpus. Thirdly, we will develop all possibilities of target result by applying dynamic programming technique. Finally, translated results of this system should be evaluated based on standard test set.

Acknowledgment

Special thanks to Dr. Krit KOSAWAT for helpful discussions, Mr. Prachya BOONKWAN and Mr. Taneth RAUNGRAJITPAKORN for insightful comments, and Mr. Sitthaa PHAHOLPHINYO and Ms. Monthika BORIBOON for providing the corpora. This research was supported in part by the National Electronics and Computer Technology Center (NECTEC), Thailand.

REFERENCES

[1] D.J. Arnold and Louis des Tombe, *Basic theory and methodology in Eurotra*. In Sergei Nirenberg, editor, Cambridge University Press, Cambridge, 1987, pp. 114-135.

[2] Peter et al Brown. "A statistical approach to language translation," in *Proceedings of the 12th COLING*, 1988, pp. 71-76.

[3] Makoto Nagao, "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle," *Artificial and Human Intelligence*, North Holland, 1984, pp. 173-180.

[4] NECTEC. "ParSit: Online English-Thai Machine Translation Service" [Online]. Available: http://www.suparsit.com

[5] AgentDict. [Online]. Available: http://www.agentdict.net

[6] Sitthaa Phaholphinyo, Teerapong Modhiran, Nattapol Kritsuthikul, and Thepchai Supnithi, "A Practical of Memory-based Approach for Improving Accuracy of MT," in *Proceedings of MT Summit X*, 2005, pp. 41-46.

[7] SWATH. Smart Word Analysis for Thai., http://www.links.nectec.or.th/download.php

[8] Donald R. Morrison, "PATRICIA-Practical Algorithm To Retrieve Information Coded in Alphanumeric," in ACM Journal, Vol. 15, No. 4, October 1968, pp. 514-534

[9] Toni Badia, Gemma Boleda, Maite Melero, and Antoni Oliver, "An *n*-gram approach to exploiting a monolingual corpus for Machine Translation," in *Proceedings EBMT Workshop of MT Summit X*, 2005, pp. 1-7.

[10] Peter Dirix, Ineke Schuurman, and and Vincent Vandeghinste, "METIS-II: Example-based machine translation using monolingual corpora – System description," in *Proceedings of EBMT Workshop of MT Summit X*, 2005, pp. 43-50.

[11] John Fry, "Assembling a parallel corpus from RSS news feeds," in *Proceedings of EBMT Workshop of MT Summit X*, 2005, pp. 59-62.

[12] John HUTCHINS, "Towards a definition of example-based machine translation," in *Proceedings of EBMT Workshop of MT Summit X*, 2005, pp. 63-70.

[13] Stella Markantonatou, Sokratis Sofianopoulos, Vassiliki Spilioti, Yiorgos Tambouratzis, Marina Vassiliou, Olga Yannoutsou, and Nikos Ioannou, "Monolingual Corpus-based MT using Chunks," in *Proceedings of EBMT Workshop of MT Summit X*, 2005, pp. 91-97.

[14] Vincent Vandeghinste, Peter Dirix, and Ineke Schuurman, "Example-based Translation without Parallel Corpora: First experiments on a prototype," in *Proceedings of EBMT Workshop of MT Summit X*, 2005, pp. 135-142.

[15] Satoshi Sato and Makoto Nagao, "Toward Memory based Translation," in *Proceedings of the 13th COLING*, 1990, pp. 247-252.

[16] Eiji Aramaki, Sadao Kurohashi, Satoshi Sato, and Hideo Watanabe, "Finding translation correspondences from parallel parsed corpus for example-based translation," in *Proceedings of MT Summit VIII*, 2001, 27-32.

[17] Taro Watanabe, and Eiichiro Sumita, "Example-based Decoding for Statistical Machine Translation," in *Proceedings of MT Summit IX*, 2003, pp. 410-417.

[18] Taro Watanabe, and Eiichiro Sumita, "Bidirectional decoding for statistical machine translation," in *Proceedings of 19th COLING*, 2002, pp. 1079-417.

ประวัติผู้เขียน

1. ประวัติส่วนตัว

ชื่อ-สกุล	นายณัฐพล กฤษสุทธิกุล
การศึกษา	วิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์)
	คณะวิทยาศาสตร์ ภาควิชาวิทยาการคอมพิวเตอร์
	มหาวิทยาลัยกรุงเทพ

2. ประวัติการทำงาน

พ.ศ. 2543-2545	Web Master
	บริษัท Diethelm & Co Ltd
พ.ศ. 2545-2546	ผู้ช่วยนักวิจัย
	งานซอฟต์แวร์พื้นฐานและทั่วไป
	ฝ่ายวิจัยและพัฒนาสาขาสารสนเทศ
	ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC)
พ.ศ. 2546-2549	ผู้ช่วยนักวิจัย 1
	งานเทคโนโลยีประมวลผลข้อความ
	ฝ่ายวิจัยและพัฒนาสาขาสารสนเทศ
	ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC)