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 และปัญหาการแปลวลีสำหรับปัญหาในกรณีอื่นไม่ขอกล่าวถึง

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