

บทที่ 5

สรุปการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

วิทยานิพนธ์ฉบับนี้ได้นำเสนอระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่าง โดยใช้ตัวแบบเอ็นแกรม ระบบนี้เป็นระบบแรกสำหรับระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่าง ผลการแปลโดยเฉลี่ยดีกว่าเมื่อเปรียบเทียบกับระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงกฎไวยากรณ์ “ภายิต”

การใช้ตัวแบบเอ็นแกรมสามารถแก้ปัญหาการเลือกใช้คำให้เหมาะสมตามบริบทระดับประโยคและการแปลวลีได้เป็นอย่างดี ในขณะที่การประยุกต์ใช้ตัวแบบเอ็นแกรมกับการแก้ไขปัญหาการเรียงลำดับคำยังคงต้องการปริมาณข้อมูลเชิงสถิติอีกเป็นอย่างมาก ทั้งนี้เป็นเพราะต้องตรวจสอบความเป็นส่วนประชิดซึ่งมีความหลากหลายในการสร้าง (diversity of formation) เป็นอย่างมาก

เปรียบเทียบกับระบบแปลภาษาด้วยเครื่องแบบอิงกฎไวยากรณ์ (Rule-based MT) ระบบจะสามารถแปลกลุ่มประโยคที่เป็นสำนวนที่มีลักษณะคำที่ต่อเนื่องกันได้ดี

เปรียบเทียบกับระบบแปลภาษาด้วยเครื่องแบบอิงสถิติ (Statistical-based MT) ระบบจะมีความต้องการใช้ข้อมูลของคลังข้อความแบบคู่สั้นน้อยกว่า เนื่องจากระบบมีการใช้คลังข้อความแบบเดี่ยวสำหรับภาษาต้นทางและคลังข้อความแบบเดี่ยวสำหรับภาษาปลายทาง ซึ่งสามารถรวบรวมได้โดยง่าย

5.2 ข้อเสนอแนะ

ส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรม (n -gram analysis Component) ยังไม่มีการจำกัดขนาดสูงสุดของ n -gram ทำให้ใช้เวลามากในการเข้าสู่ของข้อมูลในคลังข้อความคู่ภาษาสำหรับประโยคต้นทางที่ยาว หากเราสามารถกำหนดขนาดสูงสุดของ n -gram ได้จะทำให้การเข้าสู่ทำได้เร็วขึ้น อย่างไรก็ตามก็ยังคงมีขนาดสูงสุดที่เหมาะสมของ n -gram สำหรับภาษาต้นทางยังต้องมีการทำวิจัยต่อไป

ส่วนโปรแกรมการก่อกำเนิดแบบเอ็นแกรม (n -gram generation Component) ใช้วิธีการแบบละโมภ (Greedy Algorithm) มาตรวจจับส่วนของประโยคที่ดีที่สุดที่สามารถนำมาต่อกันได้ ควรนำวิทยาการศึกษานี้ (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 ควรมีการเพิ่มเติมในส่วนนี้เพื่อเพิ่มคุณสมบัติของระบบ

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