

การแลกเปลี่ยนข้อมูลสารสนเทศระหว่างประเทศเป็นภารกิจอันสำคัญของสังคมแห่งข้อมูลข่าวสาร การแปลภาษาด้วยเครื่องคอมพิวเตอร์ถูกนำมาใช้เพื่อลดข้อจำกัดทางกำแพงภาษาออกไป เนื่องด้วยความสามารถในการประมวลผลของคอมพิวเตอร์ที่ทำงานได้รวดเร็วและมีประสิทธิภาพมากขึ้น ทำให้เทคโนโลยีการประมวลผลภาษาธรรมชาติที่ใช้คลังข้อมูลขนาดใหญ่ กลายเป็นแนวความคิดพื้นฐานของการพัฒนาซอฟต์แวร์ที่ใช้ข้อมูลปริมาณมหาศาล วิทยานิพนธ์ฉบับนี้นำเสนอระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่าง โดยใช้ตัวแบบเอ็นแกรม และนำเสนอทั้งข้อดีและข้อเสียของวิธีการดังกล่าวที่ได้ถูกนำมาอภิปราย

วิทยานิพนธ์ฉบับนี้นำเสนอระบบแปลภาษาอังกฤษ-ไทยด้วยเครื่องแบบอิงตัวอย่าง โดยใช้ตัวแบบเอ็นแกรม โดยระบบดังกล่าวมีส่วนประกอบสำคัญสองส่วนคือส่วนโปรแกรมการวิเคราะห์แบบเอ็นแกรม และส่วนโปรแกรมการถอดกำเนิดแบบเอ็นแกรม

ตัวแบบเอ็นแกรมสามารถแก้ปัญหาการเรียงลำดับคำที่ไม่เหมือนกันของภาษาต้นฉบับ และภาษาเป้าหมาย สามารถแก้ปัญหาการเลือกคำที่เหมาะสมตามบริบท อีกทั้งยังสามารถแก้ปัญหาการแปลวลีได้

จากผลการทดลองสรุปได้ว่าวิธีการดังกล่าวสามารถแก้ปัญหาพื้นฐานของระบบแปลภาษาด้วยเครื่อง ดังนี้ การเข้าสู่แบบแมนตรง 2% แก้ปัญหาการเรียงลำดับคำได้ถูกต้องโดยเฉลี่ย 27% แก้ปัญหาการเลือกใช้คำให้เหมาะสมตามบริบทได้ถูกต้องโดยเฉลี่ย 57% และสามารถแก้ปัญหาการแปลวลีได้ถูกต้องโดยเฉลี่ย 51% ซึ่งผลลัพธ์ดังกล่าวดีกว่าระบบแปลภาษา “ภายิต” ในกรณีต่อไป นี้ การเข้าสู่แบบแมนตรง ปัญหาการเลือกใช้คำให้เหมาะสมตามบริบท และปัญหาการแปลวลี แต่ยังคงน้อยกว่าในการแก้ปัญหาการเรียงลำดับคำ

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.