บทที่ 1

บทนำ

1.1 ความเป็นมา

ปัจจุบันได้มีการนำคอมพิวเตอร์เข้ามาใช้ในงานด้านต่างๆ อย่างแพร่หลาย ไม่ว่าจะเป็นงานทางด้านการคำนวณ ด้านกราฟฟิก การจัดเก็บฐานข้อมูล รวมถึงการนำไปใช้งานด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing) การประมวลผลภาษาธรรมชาติคือกระบวนการที่จะทำให้คอมพิวเตอร์สามารถที่จะเข้าใจภาษามนุษย์ได้ ตัวอย่างเช่น การแปลภาษาไทย-อังกฤษ (Thai – English Machine Translation) การสังเคราะห์เสียงภาษาไทย (Thai Speech Synthesis) หรือ การสืบค้นหาข้อความทั้งเอกสาร (Full Text Search) เป็นต้น สำหรับภาษาที่ไม่มีการเว้นวรรคระหว่างคำ เช่น ภาษาจีน ภาษาญี่ปุ่น และภาษาอื่นๆ รวมทั้งภาษาไทยด้วย การหาขอบเขตของคำหรือการตัดคำจะเป็นสิ่งที่จำเป็นที่จะต้องทำเป็นอันดับแรกสำหรับงานด้านการประมวลผลภาษาธรรมชาติ และประสิทธิภาพของการตัดคำก็จะส่งผลถึงความถูกต้องของการประมวลผลภาษาธรรมชาติในระบบงานต่างๆ ด้วย ดังนั้นจะเห็นได้ว่าการตัดคำเป็นสิ่งที่สำคัญอย่างยิ่ง ในงานด้านการประมวลผลภาษาธรรมชาติ

งานหลายๆ งานในด้านการประมวลผลภาษาธรรมชาติ นอกจากที่จะต้องรู้ขอบเขตของคำแล้ว บางงานยังมีความจำเป็นต้องการทราบหน้าที่คำ (Part of Speech) หรือความหมาย (Semantic) ของคำด้วย เพื่อที่จะสามารถนำไปใช้ในการประมวลผลให้มีประสิทธิภาพมากยิ่งขึ้น ดังเช่นในการแปลภาษา การที่จะแปลให้ถูกต้องนั้น นอกจากจะต้องทราบขอบเขตของคำแล้ว การทราบหน้าที่คำจะช่วยเพิ่มความถูกต้องในการแปลด้วย เช่นคำว่า "เกาะ" อาจจะแปลเป็นภาษาอังกฤษได้เป็น "To attach" หรือ "Island" ซึ่งทั้ง 2 คำมีหน้าที่คำต่างกัน ดังนั้นถ้าทราบถึงหน้าที่ของคำ เช่นถ้าต้องการแปลคำว่า "เกาะ" ที่มีหน้าที่คำเป็นคำนาม เราก็จะแปลเป็น Island ดังนั้นการทราบหน้าที่คำจะส่งผลทำให้การแปลภาษามีความถูกต้องมากยิ่งขึ้น หรือในระบบแก้ไขคำผิด การทราบหน้าที่คำหรือความหมายของคำ ก็สามารถทำให้ระบบแก้ไขคำผิดเลือกคำที่ถูกต้องอย่างมีประสิทธิภาพมากยิ่งขึ้น หรือในโปรแกรมตรวจสอบความถูกต้องของไวยากรณ์ การทราบหน้าที่ของคำนั้นก็จะมีความจำเป็นอย่างมาก

ดังนั้นจะเห็นได้ว่าการตัดคำและการหารายละเอียดต่างๆ ของคำนั้นจะเป็นกระบวนการพื้นฐานสำหรับการประมวลผลภาษาธรรมชาติ และจะส่งผลถึงประสิทธิภาพของระบบต่างๆ ที่นำไปใช้ด้วย การ

ตัดคำนั้นได้มีการพัฒนาต่อเนื่องมานาน และได้มีการพัฒนาวิธีการต่างๆ เพื่อให้เหมาะสมกับงานแต่ละงาน โดยในระบบต่างๆ ที่นำการตัดคำไปใช้นั้นก็มีความต้องการประสิทธิภาพของการตัดคำไม่เท่ากัน เช่นในงานบางอย่างอาจจะต้องการความถูกต้องในการตัดคำอย่างมากมิฉะนั้นจะส่งผลให้ระบบไม่สามารถทำงานได้ถูกต้อง หรือบางงานอาจจะไม่ต้องการความถูกต้องมากนัก แต่ต้องการความรวดเร็วในการตัดคำมากกว่า ส่งผลทำให้มีการพัฒนาการตัดคำในแบบต่างๆ ส่วนในการหารายละเอียดต่างๆ ของคำ เพิ่งจะเริ่มมีการพัฒนาขึ้นมาไม่นานสำหรับภาษาไทย โดยรายละเอียดของคำที่เริ่มมีการพัฒนาก็คือ การหาหน้าที่ของคำ (Part-of-Speech Tagging) หรือ การหาความหมายของคำ (Semantic Tagging)

1.2 ปัญหาการตัดคำ

การตัดคำได้มีการพัฒนาอย่างต่อเนื่องมาเป็นเวลานานกว่า 10 ปี แต่ก็ยังไม่มีวิธีการใดที่สามารถจะตัดคำได้ถูกต้องทั้งหมด และงานทางด้านการประมวลผลภาษาธรรมชาติของภาษาไทยนั้นมีความจำเป็นที่จะต้องมีวิธีการตัดคำที่มีประสิทธิภาพมากที่สุด และจากงานวิจัยการตัดคำที่ผ่านมามีการนำเอากฎพจนานุกรม ค่าสถิติ ไวยากรณ์ เข้ามาช่วยในการตัดคำ แต่ก็ยังไม่สามารถตัดคำได้ถูกต้องทั้งหมด โดยสาเหตุที่ทำให้การตัดคำโดยมีการใช้พจนานุกรมที่ผ่านมานั้น ไม่สามารถตัดคำได้ถูกต้อง เนื่องมาจากสาเหตุดังต่อไปนี้

1.2.1 ข้อความกำกวม ทำให้การตัดคำสามารถตัดคำได้หลายแบบ ทำให้เกิดความสับสนขึ้นว่าแบบไหนจะเป็นแบบที่ถูกต้องที่สุด

1.2.2 คำศัพท์ที่ไม่ปรากฏอยู่ในพจนานุกรม จะเป็นสาเหตุทำให้การตัดคำไม่สามารถทำได้ถูกต้อง ซึ่งนอกจากจะตัดคำที่ไม่มีในพจนานุกรมผิดแล้ว อาจจะส่งผลทำให้คำรอบข้างมีการตัดทำผิดด้วย

1.3 วัตถุประสงค์ของวิทยานิพนธ์

1.3.1 เพิ่มประสิทธิภาพของการตัดคำ โดยนำเอาคุณลักษณะ(Feature) เข้ามาแก้ปัญหาต่อไปนี้

1.3.1.1 แก้ไขปัญหาความกำกวม

1.3.1.2 แก้ไขปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมที่เป็นชื่อเฉพาะ (ชื่อคน, ชื่อองค์กร หรือ ชื่อสถานที่ เท่านั้น)

1.3.2 เปรียบเทียบประสิทธิภาพการเรียนรู้คุณลักษณะต่างๆ ที่จะนำมาแก้ปัญหาการตัดคำ โดยจะทำการเปรียบเทียบประสิทธิภาพของการเรียนรู้ของเครื่องระหว่าง ริปเปอร์ กับ วินโนว์

1.3.3 กำกับหน้าที่คำ

1.4 ขอบเขตของวิทยานิพนธ์

1.4.1 แก้ไขปัญหาที่ทำให้การตัดคำภาษาไทยโดยใช้พจนานุกรมไม่สามารถตัดคำได้ถูกต้อง ซึ่งวิทยานิพนธ์นี้จะทำการแก้ไขปัญหา โดยสามารถแบ่งปัญหาได้เป็น 2 กรณี คือ

1.4.1.1 ในกรณีที่ข้อความที่จะนำมาตัดคำไม่มีคำที่ไม่ปรากฏในพจนานุกรม โดยในกรณีนี้เมื่อทำการตัดคำแล้ว จะสามารถตัดได้หลายแบบ โดยที่ทุกๆ คำในแต่ละแบบจะปรากฏอยู่ในพจนานุกรมทั้งหมด ซึ่งอาจจะมีความหมาย หรือไม่มีความหมายก็ได้ และเป็นสาเหตุทำให้การตัดคำไม่ถูกต้อง ตัวอย่างเช่นข้อความ "ตากลม" สามารถตัดได้เป็น "ตาก ลม" หรือ "ตา กลม" ซึ่งทั้ง 2 แบบจะมีความหมายทั้งคู่ หรืออีกตัวอย่างหนึ่งเช่น "ขนบนอก" สามารถตัดได้เป็น "ขน บน อก" และ "ขนบ นอก" โดยที่แบบแรกเท่านั้นที่มีความหมาย

1.4.1.2 ในกรณีที่ข้อความที่จะนำมาตัดคำมีคำที่ไม่ปรากฏในพจนานุกรม ในกรณีนี้คำที่ไม่ปรากฏในพจนานุกรมนั้นจะเป็นสาเหตุทำให้ตัดคำผิด ตัวอย่างเช่น "ไมโครซอฟต์" จะตัดคำได้เป็น "ไม โค ร ซอ ฟต์" สำหรับการตัดคำที่ใช้พจนานุกรมเพียงอย่างเดียว ซึ่งถ้าต้องการจะแก้ไขปัญหานี้ควรจะต้องมีการนำข้อมูลอื่นๆ เข้ามาประกอบด้วย โดยในงานวิทยานิพนธ์นี้จะทำการแก้ไขปัญหาประเภทนี้ โดยจะจำกัดเฉพาะคำที่เป็นชื่อคน ชื่อสถานที่ และชื่อองค์กรเท่านั้น

1.4.2 นำเอาการเรียนรู้ของเครื่องเข้ามาช่วยในการดึงคุณลักษณะต่างๆ จากคลังข้อความ เพื่อที่จะนำเอาคุณลักษณะต่างๆ ที่ได้จากการเรียนรู้ของเครื่องเข้ามาใช้ในการตัดคำ และทำการเปรียบเทียบการเรียนรู้ของเครื่องในแบบต่างๆ โดยการเรียนรู้ของเครื่องที่จะนำมาใช้นั้น มีอยู่ 2 วิธีคือ

1.4.2.1 ริปเปอร์ (RIPPER)

1.4.2.2 วินโนว์ (Winnow)

1.5 ขั้นตอนการวิจัย

1.5.1 ศึกษาการตัดคำวิธีการต่างๆ ที่ผ่านมา

1.5.2 ศึกษาการเรียนรู้ของเครื่องเพื่อที่จะนำมาประยุกต์ใช้ในการตัดคำ

1.5.2.1 ศึกษาการเรียนรู้ของเครื่องที่มีชื่อว่า ริปเปอร์

1.5.2.2 ศึกษาการเรียนรู้ของเครื่องที่มีชื่อว่า วินโนว์

1.5.3 พัฒนาโปรแกรมตัดคำและกำหนดหน้าที่ของคำโดยใช้โมเดลไตรแกรม

1.5.4 ออกแบบและพัฒนาระบบการตัดคำแบบใหม่ที่นำเอาการเรียนรู้ของเครื่องเข้ามาใช้

1.5.5 ทำการทดลองเพื่อวัดประสิทธิภาพและเปรียบเทียบผลของการตัดคำในแบบต่างๆ

1.5.6 สรุปผลการวิจัย และ จัดทำรายงานวิทยานิพนธ์

1.6 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1.6.1 รวบรวมงานวิจัยทางด้านการตัดคำและงานที่เกี่ยวข้องที่ผ่านมา

1.6.2 เปรียบเทียบประสิทธิภาพการเรียนรู้ของเครื่องระหว่าง ริปเปอร์ กับ วินโนว์ ในการนำคุณลักษณะต่างๆ เข้ามาแก้ปัญหาการตัดคำ

1.6.3 นำกระบวนการเรียนรู้ของเครื่องมาประยุกต์ใช้ในการแก้ปัญหาการตัดคำ

1.6.4 การตัดคำแบบใหม่ที่มีประสิทธิภาพมากยิ่งขึ้น

1.6.5 สรุปปัญหาและอุปสรรคในการตัดคำ

1.6.6 แนวทางการเพิ่มประสิทธิภาพของการตัดคำ

1.7 สิ่งตีพิมพ์ที่ได้จากงานวิทยานิพนธ์

จากงานวิทยานิพนธ์นี้ได้มีบทความที่ได้รับการตีพิมพ์ทั้งหมดจำนวน 3 บทความคือ

1.7.1 บทความเรื่อง "Feature-based Thai Word Segmentation" ในงานประชุมวิชาการ Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97) (Incorporating SNLP'97)" โดย Surapant Meknavin, Paisarn Charoenpornsawat and Boonserm Kijsirikul . สถานที่จัด จ. ภูเก็ต วันที่ 2-4 ธันวาคม พ.ศ. 2540

1.7.2. บทความเรื่อง "Feature-Based Proper Name Identification in Thai" ในงานประชุมวิชาการ "The National Computer Science and Engineering Conference'98 (NCSEC'98)" โดย Paisarn Charoenpornsawat, Boonserm Kijsirikul and Surapant Meknavin. สถานที่จัด มหาวิทยาลัยเกษตรศาสตร์ วันที่ 19-21 ตุลาคม พ.ศ. 2541

1.7.3 บทความเรื่อง "Feature-based Thai Unknown Word Boundary Identification Using Winnow" ในงานประชุมวิชาการ "1998 IEEE Asia-Pacific Conference on Circuits and Systems (APCCAS'98)" โดย Paisarn Charoenpornsawat, Boonserm Kijsirikul and Surapant Meknavin. สถานที่จัด จ. เชียงใหม่ วันที่ 24-27 พฤศจิกายน พ.ศ. 2541