บทที่ 1

บทนำ

1.1 ความเป็นมาของปัญหา

เนื่องด้วยสาเหตุที่คำในประโยคภาษาไทยจะเขียนติดกันเสียเป็นส่วนใหญ่ จะมีการแบ่งวรรคตอนบ้าง เป็นบางคราว โดยใช้เครื่องหมายแบ่งวรรคตอน ซึ่ง จะแตกต่างจากภาษาอื่น ๆ หลายภาษา เช่น ภาษาอังกฤษ ภาษาฝรั่งเศส หรือภาษาเยอรมันเป็นต้น ที่จะมีการเว้นระยะระหว่างคำชัดเจน จึงส่งผลให้การประมวลผลภาษาไทยด้วยคอมพิวเตอร์มีความแตกต่างออกไป

ดังนั้นการตัดคำ หรือการแยกคำ ออกมาจากคำที่เขียนติดกันอยู่ในประโยคภาษาไทย จึงเป็นความจำเป็นขั้นต้น ก่อนที่จะสามารถประมวลผลอื่นต่อไปได้อย่างเช่นภาษาอื่น ซึ่งอาจกล่าวได้ว่า การตัดคำเป็นองค์ประกอบวิกฤติ (critical factor) ของการประมวลผลภาษาไทยเลยก็อาจจะได้ สำหรับงานต่าง ๆที่ต้องการการตัดคำยกตัวอย่างได้ เช่น

1.1.1 การจัดรูปแบบเอกสารในงานประมวลผลคำ (word processing)
1.1.2 การตรวจสอบตัวสะกดภาษาไทย (spelling check)
1.1.3 การวิเคราะห์ไวยากรณ์ (syntax analysis)
1.1.4 การแปลภาษาด้วยเครื่องจักร (machine translation)
1.1.5 การทำดรรชนีสำหรับเอกสาร (document indexing)
1.1.6 การเชื่อมโยงความหมายของคำ (thesaurus)
1.1.7 การประมวลผลภาษาธรรมชาติ (natural language processing)
1.1.8 การสังเคราะห์เสียงพูด (speech synthesis) จากประโยคภาษาไทย
1.1.9 การวิเคราะห์กฎเกณฑ์ในการสร้างประโยค (syntactic rules analysis)

การพัฒนาขั้นตอนวิธีการตัดคำได้ดำเนินการอย่างหลากหลายจากหน่วยงานวิจัยต่าง ๆทั้งภาครัฐบาล และ เอกชน โดยส่วนมากจะเป็นส่วนหนึ่งของโปรแกรมประยุกต์ใด ๆ หรือส่วนจัดการภาษาไทยของโปรแกรมจัดการระบบ (operating system) ซึ่งแต่ละวิธีการตัดคำเหล่านั้นจะมีความแตกต่างกัน ทั้งในด้านความถูกต้องของการตัดคำ ประโยคที่ได้ ความรวดเร็วของการทำงาน ตลอดจนปริมาณการใช้ทรัพยากรต่าง ๆ

งานวิจัยของ ดร.รัตติกร วรากูลศิริพันธุ์และทีมงาน [8] ได้กล่าวถึงปัญหาของการประมวลผลภาษาไทยด้วยคอมพิวเตอร์ไว้ว่า ลักษณะพื้นฐานของภาษาไทยจะประกอบด้วยหน่วยคำ (morpheme) ที่เขียนติดกันโดยไม่มีเครื่องหมายหรือช่องว่างบอกการจบคำ ซึ่งทำให้เกิดความกำกวมเมื่อทำการประมวลผลประโยคภาษาไทยด้วยคอมพิวเตอร์จึงต้องอาศัยเทคนิคหรือ

อัลกอริธึมที่สามารถแยกหน่วยคำออกจากประโยคให้ได้ทั้งความถูกต้องและปราศจากความกำกวมไม่ว่าจะเป็นทางด้านไวยากรณ์หรือความหมาย

งานวิจัยของ ดร.รัตติกร วราภูลศิริพันธุ์และทีมงาน [5] ได้กล่าวว่า โปรแกรมตัดคำจะต้องตัดคำให้ได้หน่วยคำที่มีความหมายถูกต้องตามพจนานุกรมภาษาไทยเพื่อจะนำคำเหล่านั้นไปหาโครงสร้างทางไวยากรณ์ต่อไป ซึ่งงานวิจัยดังกล่าวได้ทำการทดลองกับประโยคกว่า 1,000 ประโยคด้วยพจนานุกรมคำศัพท์ประมาณ 14,000 คำ ผลลัพธ์ที่ได้จากการแยกแยะคำในประโยคไม่ปรากฏข้อผิดพลาดเลย แต่ในบางกรณีจะมีผลลัพธ์ได้มากกว่าหนึ่งประโยคซึ่งจะต้องอาศัยกฎทางไวยากรณ์ (Syntax) และความหมาย (Semantic) เป็นตัวช่วยตัดสินใจ โดยจะมีพฤฒิกรรมทำนองเดียวกับงานวิจัยของ สมปรารถนา รัทยานนท์ [1] ที่ได้ประโยคมากกว่าหนึ่งประโยคจากการตัดคำภาษาไทยด้วยคอมพิวเตอร์เช่นกัน

งานวิจัยของ ดร.รัตติกร วราภูลศิริพันธุ์และทีมงาน [6] ได้กล่าวไว้ว่า การสร้างฐานความรู้ (Knowledge Base) เพื่อเลือกประโยคที่ได้จากการตัดคำที่ถูกต้องนั้นโดยพื้นฐานแล้วจะใช้กฎไวยากรณ์ภาษาไทยเป็นหลักซึ่งจะต้องมีกฎไวยากรณ์จำนวนมากทำให้ฐานความรู้มีขนาดใหญ่งานวิจัยนั้นจึงได้เสนอวิธีการที่อาศัยความถี่ของการใช้คำไทยแทนและสรุปได้ว่าความถูกต้องของผลลัพธ์จะขึ้นอยู่กับการหาค่าความถี่ของคำไทยที่ใช้อยู่ในชีวิตประจำวันการสุ่มตัวอย่างประโยคได้มากและครอบคลุมทุกคำศัพท์ที่จำเป็นจะมีผลให้การหาความถี่ของคำไทยและความน่าจะถูกใช้ในภาษาของคำไทยมีความถูกต้องมากขึ้น

งานวิจัยของ ยืน ภู่วรวรรณ และ วิวรรธน์ อึ่มอารมณ์ [3] ได้กล่าวถึงปัญหาการตัดคำโดยใช้พจนานุกรมว่าย่อมเปลืองที่เก็บแต่สามารถลดขนาดโดยใช้เทคนิคของโครงสร้างข้อมูลเพื่อลดความซ้ำซ้อนทำให้ใช้เนื้อที่น้อยลงซึ่งทำให้ยืนยันได้ว่าการประมวลผลภาษาไทยด้วยเครื่องคอมพิวเตอร์เป็นเรื่องเป็นไปได้ซึ่งก่อนหน้านี้ทุกคนจะคิดว่าการตรวจสอบตัวสะกดภาษาไทยด้วยเครื่องคอมพิวเตอร์นั้นเป็นเรื่องเป็นไปไม่ได้

ความต้องการที่จะเปรียบเทียบประสิทธิภาพในแง่มุมต่าง ๆ ของขั้นตอนและวิธีการตัดคำแบบต่าง ๆ จึงจำเป็นมากขึ้นในการที่จะนำไปสู่ความเป็นมาตรฐานของการพัฒนาต่อไป และการใช้งานการตัดคำจากประโยคภาษาไทยด้วยคอมพิวเตอร์ อีกทั้งช่วยให้การเลือกใช้ขั้นตอนวิธีการตัดคำภาษาไทยแบบต่าง ๆ ได้เหมาะสมกับงานที่ต้องการ รวมทั้งอาจใช้ขั้นตอนวิธีการตัดคำนี้ร่วมกัน สำหรับโปรแกรมประยุกต์หลาย ๆโปรแกรม

1.2 วัตถุประสงค์ของวิทยานิพนธ์

1.2.1 เพื่อศึกษา วิเคราะห์การทำงานรวมทั้งวิธีแนวทางที่ใช้ของการตัดคำแบบต่าง ๆที่มีอยู่เพื่อหาแนวทางในการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีการตัดคำนั้น

1.2.2 เพื่อเปรียบเทียบข้อดี ข้อเด่น รวมทั้งคุณสมบัติที่มีผล ต่อประสิทธิภาพของขั้นตอนวิธีการตัดคำภาษาไทย แบบต่าง ๆ ที่มีอยู่

1.2.3 งานวิจัยจะทำการวิเคราะห์ และสังเคราะห์ มาตรวัดประสิทธิภาพ (performance metrics) เพื่อใช้เป็นมาตรวัดประสิทธิภาพของ โปรแกรมแยกคำภาษาไทย ที่มีอยู่และที่จะมีการพัฒนาในอนาคต

1.3 ขอบเขตของวิทยานิพนธ์

1.3.1 งานวิจัยจะเปรียบเทียบคุณลักษณะเด่น และคุณลักษณะด้อย ของวิธีการเหล่านั้นในลักษณะ Quantitative

1.3.2 งานวิจัยนี้ได้ทำการวิเคราะห์ และ สังเคราะห์ มาตรวัดประสิทธิภาพ (performance metrics) ขึ้นมา

1.3.3 จะใช้มาตรวัดประสิทธิภาพนี้วัดความสามารถของโปรแกรมตัดคำภาษาไทย ที่ได้มีการพัฒนาและใช้งานอยู่ในปัจจุบัน

1.3.4 ต้นแบบที่ใช้ทำการทดลองจะใช้กับรหัสภาษาไทย ส.ม.อ. เท่านั้น

1.4 ขั้นตอนการวิจัย

1.4.1 ค้นคว้าศึกษางานวิจัยที่เกี่ยวข้อง ผู้วิจัยจะทำการศึกษา ค้นคว้า รวบรวม และวิเคราะห์ การทำงานของขั้นตอนวิธีการตัดคำแบบต่าง ๆ ที่ได้มีการพัฒนา และใช้งานในประเทศไทย เช่น วิธีการใช้กฎเกณฑ์ (rules based) วิธีการใช้พจนานุกรม (dictionary approach) วิธีการเทียบคำที่ยาวที่สุด (longest word mapping) และวิธีการย้อนกลับ (back tracking) เป็นต้น

1.4.2 วิเคราะห์พฤติกรรมที่มีผลต่อประสิทธิภาพของโปรแกรมแยกคำภาษาไทย

1.4.3 สังเคราะห์มาตรวัดประสิทธิภาพที่จะใช้วัด

1.4.4 กำหนดหน่วยวัดของมาตรวัดประสิทธิภาพแต่ละตัว

1.4.5 กำหนดกลุ่มตัวอย่างของข้อมูลที่จะเป็นอินพุท

1.4.6 ใช้มาตรวัดประสิทธิภาพทำการวัดประสิทธิภาพ

1.4.7 สรุปผล วิเคราะห์ และ วิจารณ์ รวมทั้งรวบรวมข้อแนะนำ และมาตรฐานต่าง ๆ ที่จะเป็นประโยชน์ในการพัฒนา หรือประยุกต์ใช้ขั้นตอนวิธีการตัดคำ และเครื่องมือวัดประสิทธิภาพอันนี้ต่อไปในอนาคต

1.5 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1.5.1 ได้มาตรวัดประสิทธิภาพ ที่จะสามารถวัดความสามารถ ของโปรแกรมตัดคำภาษาไทย ด้วยคอมพิวเตอร์ในแง่มุมต่าง ๆ ในการที่จะเลือกใช้ขั้นตอนวิธีการตัดคำให้เหมาะสมกับงานหรือการใช้การตัดคำร่วมกันในงานหลาย ๆ งาน

1.5.2 ได้ข้อแนะนำ แนวทางที่จะนำไปสู่ความเป็นมาตรฐาน ในการพัฒนาโปรแกรมตัดคำภาษาไทยต่อไป หรือการประยุกต์ใช้งานสำหรับขั้นตอนวิธีการตัดคำภาษาไทย และงานประมวลผลภาษาไทยอื่น ๆ อาจรวมไปถึงโปรแกรมจัดการระบบภาษาไทยด้วย

1.5.3 เป็นเครื่องมือพื้นฐานที่จะใช้สำหรับการวิจัย และพัฒนาระบบการประมวลผลภาษาไทยโดยใช้คอมพิวเตอร์ต่อไป

จากปัญหาเบื้องต้นในการประมวลผลภาษาไทยด้วยคอมพิวเตอร์ดังที่ได้กล่าวมาแล้วนั้น ในบทต่อไปจะได้กล่าวในรายละเอียดของแนวคิดและทฤษฎีที่เกี่ยวข้องรวมทั้งความคืบหน้าของงานวิจัยแขนงนี้ในปัจจุบัน