บทที่ 3

มาตรวัดประสิทธิภาพของขั้นตอนวิธีการตัดคำแบบต่าง ๆ

วัตถุประสงค์ของงานวิจัยนี้เพื่อวิเคราะห์พฤติกรรมของโปรแกรมแยกคำภาษาไทย ลักษณะสมบัติสำคัญที่มีผลต่อประสิทธิภาพ และเสาะหามาตรในการวัดลักษณะสมบัติเหล่านั้น ซึ่งจะได้วิเคราะห์ และสังเคราะห์มาตรวัดประสิทธิภาพที่ใช้ทำการวัดนั้น ในบทนี้จะได้แจกแจงมาตรวัดประสิทธิภาพต่าง ๆ ที่ได้สังเคราะห์ขึ้นมา

3.1 ความสามารถที่จะตัดคำได้ (separatability)

เป็นมาตรวัดประสิทธิภาพเบื้องต้นของขั้นตอนวิธีการตัดคำแบบต่าง ๆ ในเรื่องของจำนวนคำที่ตัดออกมาได้ จากประโยคที่ป้อนเข้าไป มาตรวัดประสิทธิภาพตัวนี้มีพื้นฐานมาจากปัญหาเบื้องต้นของประโยคภาษาไทยที่ไม่มีเครื่องหมายแบ่งคำชัดเจน ดังที่งานวิจัยของ ดร.รัตติกร วรากูลศิริพันธุ์และทีมงาน [8] และงานวิจัยของคนอื่น ๆ ที่ได้กล่าวถึงปัญหานี้ไว้ตรงกันในเรื่องที่เกี่ยวกับการประมวลผลภาษาไทยด้วยคอมพิวเตอร์

หน่วยวัดของมาตรวัดประสิทธิภาพตัวนี้จะเป็นจำนวนคำที่ตัดออกมาได้ทั้งหมด

3.2 ความถูกต้องของคำหลังจากผ่านการตัดแล้ว (word validity)

คำที่ได้หลังจากการตัดออกมาจากประโยค จะต้องมีความถูกต้อง ซึ่งอาจสามารถตรวจสอบได้ โดยการหาว่าอย่างน้อยคำนั้น ต้องมีปรากฏอยู่ในพจนานุกรม วิธีการตัดคำแบบต่าง ๆ จะได้ระดับของความถูกต้องของคำหลังการตัดไม่เท่ากัน มาตรวัดประสิทธิภาพนี้จะใช้วัดความสามารถของวิธีการตัดคำนั้น ๆ มาตรวัดประสิทธิภาพตัวนี้มีพื้นฐานมาจากปัญหาที่ว่า โปรแกรมตัดคำจะต้องตัดคำ ให้ได้หน่วยคำที่มีความหมายถูกต้องตามพจนานุกรมภาษาไทย เพื่อจะนำคำเหล่านั้นไปหาโครงสร้างทางไวยากรณ์ต่อไป ดังปรากฏอยู่ในงานวิจัยของ ดร.รัตติกร วรากูลศิริพันธุ์และทีมงาน [5]

หน่วยวัดของมาตรวัดประสิทธิภาพตัวนี้จะเป็นสัดส่วนของจำนวนคำที่ตัดออกมาได้ถูกต้องตามพจนานุกรม ต่อจำนวนคำทั้งหมดที่ตัดออกมาได้

3.3 สัดส่วนความถูกต้องของคำที่ตัดออกมาได้ต่อจำนวนคำที่ใช้เป็นพจนานุกรม

มาตรวัดตัวนี้จะสามารถชี้ให้เห็นความสามารถในการเลือกกลุ่มคำที่จะมาเป็นพจนานุกรมโปรแกรมตัดคำภาษาไทยในปัจจุบันจะใช้พจนานุกรมช่วยเป็นส่วนมาก มีจำนวนคำที่นำมาใช้เป็น

พจนานุกรมนี้แตกต่างกัน สัดส่วนความถูกต้องของการตัดคำต่อคำที่นำมาเป็นพจนานุกรมหนึ่งคำคือมาตรวัดประสิทธิภาพอันนี้ และมาตรวัดนี้สามารถชี้ให้เห็นประสิทธิภาพในการใช้ทรัพยากรได้ด้วย

หน่วยวัดของมาตรวัดประสิทธิภาพตัวนี้จะเป็นสัดส่วนของจำนวนคำที่ตัดออกมาได้ถูกต้อง ต่อจำนวนคำที่ใช้เป็นพจนานุกรม

3.4 ความถูกต้องเชิงไวยากรณ์ของประโยคหลังจากผ่านการตัดแล้ว (syntax validity)

ประโยคที่ได้หลังการตัดคำสำหรับวิธีการตัดคำแต่ละประเภท จะมีความถูกต้องในเชิงไวยากรณ์มากน้อยแค่ไหน มาตรวัดประสิทธิภาพนี้จะสามารถบอกได้ มาตรวัดประสิทธิภาพตัวนี้มีพื้นฐานมาจากการที่โปรแกรมตัดคำจะมีผลลัพธ์ได้มากกว่าหนึ่งประโยคซึ่งจะต้องอาศัยกฎทางไวยากรณ์ (syntax) และความหมาย (semantic) เป็นตัวช่วยตัดสินใจดังที่ปรากฏใน งานวิจัยของ ดร.รัตติกร วรากูลศิริพันธุ์และทีมงาน [5] และงานวิจัยของ สมปรารถนา รัทยานนท์ [1]

หน่วยวัดของมาตรวัดประสิทธิภาพตัวนี้ เป็นจำนวนประโยคที่ตัดออกมาได้และมีความถูกต้องในเชิงไวยากรณ์

3.5 ความถูกต้องเชิงความหมายของประโยคหลังจากตัดแล้ว (semantic validity)

ถึงแม้ว่าประโยคที่ตัดออกมาแล้วจะมีความถูกต้องเชิงไวยากรณ์ แต่ความหมายของประโยคสามารถนำไปใช้งานได้ หรือเหมาะสมกับกาลเทศะหรือไม่ มาตรวัดประสิทธิภาพอันนี้จะใช้สำหรับวัดความสามารถในแง่มุมนี้ ของวิธีการตัดคำแบบต่าง ๆ พื้นฐานของมาตรวัดประสิทธิภาพดังเช่นปัญหาที่ระบุไว้ในข้อ 3.4 จึงได้มาตรวัดประสิทธิภาพตัวนี้ออกมา

หน่วยวัดของมาตรวัดประสิทธิภาพตัวนี้ เป็นจำนวนประโยคที่ตัดออกมาได้และมีความหมายถูกต้อง

3.6 ความสามารถที่จะรู้จักเครื่องหมายแบ่งวรรคตอนภาษาไทย (separating character recognizability)

เครื่องหมายแบ่งวรรคตอนที่กำหนดโดยราชบัณฑิตยสถาน เป็นตำแหน่งแบ่งแยกคำพื้นฐานอยู่แล้ว วิธีการตัดคำที่สมบูรณ์จะต้องรู้จักเครื่องหมายเหล่านั้น

หน่วยวัดของมาตรวัดประสิทธิภาพตัวนี้ เป็นสัดส่วนของเครื่องหมายแบ่งวรรคตอนทั้งหมดที่โปรแกรมตัดคำรู้จัก ต่อเครื่องหมายแบ่งวรรคตอนทั้งหมดที่กำหนดโดยราชบัณฑิตยสถาน

3.7 ความสามารถที่จะปรับสระ และวรรณยุกต์ที่ติดกันอย่างไม่ถูกต้อง (adjacent tone and vowel correctivity)

ในการป้อนข้อมูลภาษาไทยเข้าเครื่องคอมพิวเตอร์นั้น ลำดับของสระ และวรรณยุกต์ที่เขียนติดกัน เช่น วรรณยุกต์เอกและสระอา ในคำ "ว่า" อาจสามารถสลับกันได้ และอาจไม่เหมือนกันสำหรับผู้ใช้งานแต่ละคน ซึ่งทำให้วิธีการตัดคำมองเห็นเป็นคำที่ผิด หรือค้นหาไม่พบในพจนานุกรมได้ทั้ง ๆ ที่เป็นคำที่ถูกต้องคำหนึ่ง ขั้นตอนวิธีการตัดคำ และโปรแกรมประมวลผลภาษาไทยอื่น ควรจะต้องทราบในจุดนี้ พื้นฐานของมาตรวัดประสิทธิภาพตัวนี้จะมาจากปัญหาข้างบนซึ่งพบในงานของ Zuill และทีมงาน [9]

มาตรวัดนี้ไม่จำเป็นต้องนำมาคิดสำหรับการตัดคำภาษาไทย แต่ก่อนทำการตัดคำหรือประมวลผลภาษาไทยจะต้องกรองข้อมูลผ่านโปรแกรมนี้เสียก่อน

3.8 มาตรวัดประสิทธิภาพเชิงความเร็ว (speed metric)

ความรวดเร็วของการทำงานเป็นคุณลักษณะที่จำเป็นสำหรับการตัดคำ เนื่องจากการตัดคำเป็นหนทางวิกฤติของการประมวลผลภาษาไทย มาตรวัดประสิทธิภาพตัวนี้จะใช้เปรียบเทียบความสามารถของวิธีการตัดคำแบบต่าง ๆ ในแง่มุมนี้

3.9 มาตรวัดประสิทธิภาพการใช้ทรัพยากร (resource utilization metric)

การใช้งานทรัพยากรที่มากเกินความจำเป็น อาจส่งผลให้การประมวลผลอื่น ๆ ประสิทธิภาพด้อยลง หรือทำงานไม่ได้เลย เมื่อใช้ร่วมกันกับการตัดคำ มาตรวัดประสิทธิภาพตัวนี้ สำหรับการเปรียบเทียบความสามารถด้านนี้ของขั้นตอนวิธีการตัดคำแบบต่าง ๆ มาตรวัดประสิทธิภาพตัวนี้มีพื้นฐานมาจากปัญหาที่พบในงานวิจัยของ ยืน ภู่วรวรรณ และ วิวรรธน์ อิ่มอารมณ์ [3] ที่ว่าการตัดคำโดยใช้พจนานุกรมย่อมเปลืองที่เก็บ แต่สามารถลดขนาดโดยใช้เทคนิคของโครงสร้างข้อมูลเพื่อลดความซ้ำซ้อนทำให้ใช้เนื้อที่น้อยลง

หน่วยวัดของมาตรวัดประสิทธิภาพตัวนี้ เป็นขนาดของการใช้งานทรัพยากรของโปรแกรมตัดคำ

3.10 มาตรวัดประสิทธิภาพการแก้ไขความผิดพลาดในการตัดคำ (recoverability)

สาเหตุจากความกำกวมในคำภาษาไทยที่เขียนติดกันทำให้โปรแกรมตัดคำภาษาไทยตัดคำออกมาได้ผิดพลาดไป แต่ถ้าโปรแกรมตัดคำนั้นมีความสามารถในการแก้ไขความผิดพลาดในเบื้องต้นได้ความถูกต้องของการตัดคำก็จะสูงขึ้นตามไป

มาตรวัดต่าง ๆ ที่ได้สังเคราะห์ขึ้นมานี้บางตัวก็สามารถนำไปใช้วัดประสิทธิภาพของโปรแกรมตัดคำภาษาไทยได้เลยโดยง่ายเช่น 3.1 3.2 และ 3.3 แต่บางตัวไม่สามารถทำได้โดย

ง่ายเช่น 3.4 และ 3.5 ในงานวิจัยนี้ได้พัฒนาขั้นตอนวิธีวัดและเลือกใช้มาตรวัดหลัก ๆ ที่จะทำให้เห็นการเปรียบเทียบประสิทธิภาพในแง่มุมต่าง ๆ ได้อย่างชัดเจน ในบทต่อ ๆ ไปได้กล่าวถึงสิ่งนี้โดยละเอียด