

บทที่ 9

บทสรุปและข้อเสนอแนะ

ในบทนี้จะได้ทำการสรุปและวิจารณ์ผลการวิจัย รวมไปถึงปัญหาต่าง ๆ และข้อเสนอแนะที่จะเป็นประโยชน์สำหรับการวิจัยเพิ่มเติมต่อไป

9.1 ผลของความยาวของเอกสารต่อประสิทธิภาพของโปรแกรมตัดคำ

จากผลการวิจัยในบทที่ 8 พบว่าเอกสารภาษาไทยทั่วไปมีลักษณะการใช้คำที่ซ้ำกันมากพอสมควร โดยพบว่าที่จุดจุดหนึ่งจะมีคำใหม่เพิ่มขึ้นน้อยมากหรือไม่มีเพิ่มขึ้นเลย นั่นทำให้การเลือกใช้อรรถาภิธานเป็นสำหรับการวัดประสิทธิภาพการตัดคำ ไม่จำเป็นต้องใช้ข้อมูลจำนวนมากมายมหาศาล แต่จะต้องเลือกข้อมูลให้ครอบคลุมถึงจุดที่จำนวนคำใหม่ปรากฏน้อยหรือไม่มีเลย

9.2 เอกสารประเภทต่าง ๆ กับผลการเปรียบเทียบประสิทธิภาพการตัดคำ

ข้อมูลภาษาไทยหลายประเภทที่เลือกมาสำหรับการวัดประสิทธิภาพนี้ มีแนวโน้มที่จะได้ผลการเปรียบเทียบประสิทธิภาพที่สอดคล้องในแนวทางเดียวกัน จากการทดลองในหัวข้อ 8.11 เมื่อเปรียบเทียบผลการทดลองกับผลการทดลองที่ได้จากฐานข้อมูลภาษาไทยดาต้าเบสพบว่าให้ผลการเปรียบเทียบออกมาไม่แตกต่างกันอย่างเห็นได้ชัดแต่อย่างไร ซึ่งทำให้ได้สมมติฐานว่าประเภทของข้อมูลที่นำมาใช้มีผลน้อยต่อการวัดประสิทธิภาพ ในขณะที่ปริมาณข้อมูลที่เหมาะสมกลับเป็นสิ่งจำเป็นมากกว่า

9.3 สรุปประสิทธิภาพของโปรแกรมตัดคำ

ผลการเปรียบเทียบประสิทธิภาพในแง่มุมต่าง ๆ สามารถสรุปได้ดังนี้

9.3.1 โปรแกรมตัดคำภาษาไทยแบบเปรียบเทียบคำที่ยาวที่สุด (Longest Word Pattern Matching) จะตัดได้จำนวนคำที่ถูกต้องออกมามากที่สุดขณะที่แบบเปรียบเทียบคำที่สั้นที่สุด (Shortest Word Pattern Matching) จะตัดออกมาได้คำถูกน้อยที่สุด

9.3.2 โปรแกรมตัดคำแบบย้อนกลับ (Back Tracking) จะตัดได้คำผิดน้อยที่สุดขณะที่แบบเปรียบเทียบคำที่สั้นที่สุด (Shortest Word Pattern Matching) จะตัดออกมาได้คำผิดมากที่สุด

9.3.3 แบบอาศัยความถี่ของการใช้คำ (Word Usage Frequency) ตัดคำได้สัดส่วนความถูกต้องต่อจำนวนคำในพจนานุกรมสูงสุดขณะที่แบบเปรียบเทียบคำที่สั้นที่สุด (Shortest Word Pattern Matching) ตัดออกมาได้อัตราความถูกต้องต่อจำนวนคำในพจนานุกรมต่ำสุด

9.3.4 แบบเปรียบเทียบคำที่สั้นที่สุด (Shortest Word Pattern Matching) ตัดออกมาได้จำนวนคำสูงสุดขณะที่แบบการแก้ไขย้อนกลับ (Back Tracking) ตัดได้จำนวนคำน้อยที่สุด

9.3.5 โครงสร้างข้อมูลสำหรับพจนานุกรมที่ใช้ในโปรแกรมตัดคำมีผลมากในด้านความเร็วโดยที่โครงสร้างข้อมูลแบบทรี (Trie) [14] เป็นที่นิยมเนื่องจากให้ความเร็วสูงสุดในปัจจุบันโดยที่เวลาที่ใช้ในการค้นหาคำจากพจนานุกรมเป็นสัดส่วนกับความยาวของคำ

9.3.6 สำหรับในแง่ของการใช้งานทรัพยากรที่มีประสิทธิภาพนั้นพบว่าทางเลือกคำที่นำมาใช้เป็นพจนานุกรมมีความสำคัญมาก จำนวนคำในพจนานุกรมที่น้อยกว่าจะทำให้สิ้นเปลืองทรัพยากรน้อยกว่า โปรแกรมตัดคำแบบอาศัยความถี่ของการใช้คำเป็นแบบที่ใช้ทรัพยากรที่มีประสิทธิภาพที่สุด เทคนิคต่าง ๆ ในการลดขนาด (compression) ของพจนานุกรมก็มีส่วนช่วยเพิ่มประสิทธิภาพในแง่นี้ได้ ได้สรุปข้อดีข้อด้อยของโปรแกรมตัดคำในแต่ละแบบที่ได้จากการทดลองที่ผ่านมาไว้ดังตารางที่ 9.1

9.4 ปัญหาต่าง ๆ ที่พบในงานวิจัยนี้

9.4.1 การวัดประสิทธิภาพความถูกต้องทางไวยากรณ์ และความหมายของประโยคที่ตัดได้โดยโปรแกรมตัดคำ ได้ใช้วิธีการแบบที่ยังไม่สมบูรณ์เนื่องจากมีปัญหาความซับซ้อนของการพัฒนาขั้นตอนการวัดโดยใช้ระบบอัตโนมัติ การวิจัยเพิ่มเติมเพื่อศึกษาหาวิธีการพัฒนาขั้นตอนการวัดประสิทธิภาพในแง่นี้เป็นสิ่งจำเป็น

9.4.2 งานวิจัยนี้ได้พบปัญหาในการใช้โปรแกรมตัดคำภาษาไทยของบริษัทไมโครซอฟท์มาใช้ร่วมวัดประสิทธิภาพ กล่าวคือจะพบปัญหาโปรแกรมที่เรียกใช้โปรแกรมตัดคำนี้ไม่สามารถทำงานกับข้อมูลที่มีขนาดใหญ่ได้และเนื่องจากข้อจำกัดทางเวลาผู้วิจัยจึงเลือกที่จะไม่รวมเอาผลของการวัดประสิทธิภาพของโปรแกรมนี้นี้ด้วย

9.4.3 ปัญหาความกำกวมของการตัดคำภาษาไทยยังเป็นหัวข้องานวิจัยที่สามารถทำเพิ่มเติมได้อีกมาก จากผลการวิจัยพบว่าโปรแกรมตัดคำที่มีในปัจจุบันแก้ปัญหานี้ได้ในระดับที่ยังไม่ดีพอ โปรแกรมตัดคำที่มีคุณลักษณะดีหรือโดดเด่นน่าจะได้รับการพัฒนาเพิ่มเติมให้มีความสามารถแก้ปัญหานี้ได้ดียิ่งขึ้นยกตัวอย่างเช่นแบบการอาศัยความถี่ของการใช้คำ แบบที่ใช้พจนานุกรมลดความกำกวมช่วยและแบบที่มีความสามารถวิเคราะห์เนื้อความข้างเคียง [13] เป็นต้น นอกจากนี้แล้วการผสมผสานจุดเด่นของแต่ละโปรแกรมก็เป็นสิ่งที่น่าสนใจ

9.5 ข้อเสนอแนะและงานวิจัยที่สามารถทำเพิ่มเติมต่อจากงานวิจัยนี้ได้ จากผลการวิจัยนี้รวมไปถึงปัญหาต่าง ๆ ที่พบทำให้ได้ข้อเสนอแนะและงานวิจัยที่สามารถทำเพิ่มเติมต่อไปได้คือ

9.5.1 การหาความสัมพันธ์ระหว่างคุณภาพของข้อมูลภาษาไทย กับประสิทธิภาพของการตัดคำเป็นประเด็นปัญหาที่สำคัญมาก เพราะความหลากหลายของข้อมูลภาษาไทยซึ่งไม่มีเอกสารอ้างอิงในทำนองนี้สำหรับภาษาอังกฤษ

9.5.2 การวัดความถูกต้องของการตัด โดยคำนึงถึงเนื้อความข้างเคียง การพัฒนาขั้นตอนการวัดประสิทธิภาพในแง่มุมนี้จะช่วยให้ได้ผลการเปรียบเทียบที่สมบูรณ์มากขึ้น

9.5.3 การวัดความสามารถในการกู้คืนจากการตัดพลาด การผิดพลาดจากการตัดคำครั้งหนึ่งจะส่งผลให้เกิดความผิดพลาดต่อเนื่องกันไป

9.5.4 การพัฒนาฐานข้อมูลภาษาไทย เพื่อเป็นข้อมูลดิบ ในการวัดประสิทธิภาพการตัดคำเพื่อใช้เป็นเครื่องมือวัดโปรแกรมตัดคำใหม่ ๆ และเพื่อศึกษาคุณลักษณะของเอกสารภาษาไทย

9.5.5 การเลือกกลุ่มคำแบบต่าง ๆ ที่มาเป็นพจนานุกรมกับประสิทธิภาพการตัดคำ จำนวนคำและกลุ่มของคำในพจนานุกรมเป็นตัวแปรสำคัญที่มีผลต่อประสิทธิภาพ

9.5.6 ศึกษาการใช้งานการตัดคำในแต่ละแบบ วิธีการตัดคำแต่ละแบบมีความเหมาะสมกับงานต่างประเภทกัน

โปรแกรมตัดค่า	ข้อเด่น	ข้อด้อย	งานที่เหมาะสมที่จะนำไปใช้
แบบย้อนรอยกลับ	ได้จำนวนค่าผิดน้อยที่สุด มีวิธีการฟื้นฟูจากความผิดพลาด	ได้จำนวนค่าทั้งหมดน้อยสุด	ตรวจสอบค่าสะกด
แบบเทียบค่าสายสุด	ได้จำนวนค่าที่ถูกต้องมากที่สุด		จัดรูปแบบเอกสาร
แบบเทียบค่าสั้นสุด	ได้จำนวนค่าทั้งหมดมากที่สุด	ได้จำนวนค่าผิดมากที่สุด ได้จำนวนค่าที่ถูกต้องมาน้อยสุด ได้สัดส่วนความถูกต้องต่อจำนวนค่า ในพจนานุกรมต่ำสุด	ไม่เหมาะกับการนำไป ประยุกต์ใช้
แบบอาศัยความถี่ของค่า	ได้สัดส่วนความถูกต้องต่อจำนวนค่า ในพจนานุกรมสูงสุด		ต้องพัฒนาเพิ่มเติมจึงจะนำ ไปใช้งานได้
แบบเลือกจำนวนค่าน้อยสุด			จัดรูปแบบเอกสาร
แบบใช้พจนานุกรมค่าก้ำกวม	ได้ความผิดพลาดที่เกิดจากความ ก้ำกวมน้อยที่สุด		สังเคราะห์เสียงพูดจากประโยค การแปลภาษาด้วยเครื่อง

ตาราง 9.1 สรุปคุณลักษณะข้อเด่น ข้อด้อย และตัวอย่างการนำไปใช้งานที่เหมาะสม