รายงานวิจัยฉบับสมบูรณ์

การตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่

รายนามคณะผู้วิจัย

1. ผศ. ดร. กานดา รุณนะพงศา
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
มหาวิทยาลัยขอนแก่น อ. เมือง ขอนแก่น 40002
2. นางสาวปโยธร อุราธรรมกุล
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
มหาวิทยาลัยขอนแก่น อ. เมือง ขอนแก่น 40002

วันเริ่มต้นรับทุนวิจัย 1 พฤษภาคม 2548
วันสิ้นสุดรับทุนวิจัย 30 เมษายน 2549

กิตติกรรมประกาศ

โครงการวิจัยนี้ได้รับการสนับสนุนจาก

คณะวิศวกรรมศาสตร์

มหาวิทยาลัยขอนแก่น

ปีงบประมาณที่ได้รับทุน ปีงบประมาณ 2548

บทคัดย่อ

การตัดคำภาษาไทยคือการแยกแต่ละคำในประโยคในเอกสารภาษาไทยซึ่งมีการเขียนคำติดกัน การตัดคำนั้นจำเป็นจะต้องมีเพื่อนำไปใช้ประโยชน์ทางด้านประมวลผลภาษาธรรมชาติ เช่น การสังเคราะห์เสียงพูด และการแปลภาษา เป็นต้น วิธีการตัดคำที่นิยมใช้กันมานานได้แก่วิธีการตัดคำโดยใช้กฎ วิธีการตัดคำโดยใช้พจนานุกรม และวิธีการตัดคำโดยใช้คลังข้อความ แต่เนื่องจากเอกสารในภาษาไทยปัจจุบันมักจะมีคำจากภาษาต่างประเทศซึ่งมักจะมีคำอ่านออกเสียงอยู่ในรูปแบบของภาษาไทย จึงทำให้การตัดคำด้วยวิธีการปัจจุบันไม่สามารถตัดคำได้อย่างถูกต้องกับเอกสารในปัจจุบัน งานวิจัยนี้จึงทำการเสนอวิธีการตัดคำโดยการใช้กฎร่วมกับการใช้พจนานุกรมเพื่อเพิ่มความถูกต้องของการตัดคำ จากผลการทดลองพบว่าเมื่อเปรียบเทียบกับวิธีการตัดคำวิธีอื่น วิธีการตัดคำที่นำเสนอได้ช่วยเพิ่มความถูกต้องของการตัดคำถูกต้องในเอกสารภาษาไทยหลายประเภท

คำสำคัญ: การตัดคำภาษาไทย การตัดคำโดยการใช้กฎ พจนานุกรม

Abstract

Thai word segmentation is the process of separating words in a sentence in documents in Thai language which words are adjacent to one another. Thai word segmentation is necessary for natural language processing such as speech synthesis and language translation. Traditional techniques in Thai word segmentation include rule-based segmentation, dictionary-based segmentation, and segmentation using corpus. However, currently Thai documents often consist of words from foreign languages which are spelt in the form of Thai language. Therefore, current Thai word segmentation technique cannot separate Thai words effectively in Thai documents nowadays. This research work proposes Thai word segmentation by using rules with dictionary to enhance the correctness of Thai word segmentation. Based on experimental results, compared with other techniques, the proposed technique can increase the correctness of word segmentation in various types of Thai documents

Keywords: Thai word segmentation, Rule-based segmentation, Dictionary

สารบัญ

	หน้า
กิตติกรรมประกาศ	ก
บทคัดย่อภาษาไทย	ข
บทคัดย่อภาษาอังกฤษ	ค
สารบัญภาพ	ฉ
สารบัญตาราง	ช
บทที่ 1 บทนำ	1
1. ความเป็นมาและความสำคัญของปัญหา	1
2. วัตถุประสงค์ของงานวิจัย	2
3. ขอบเขตของการวิจัย	2
5. สถานที่ทำงานวิจัย	2
6. ขั้นตอนและวิธีการดำเนินงาน	2
7. ประโยชน์ที่คาดว่าจะได้รับ	2
บทที่ 2 วรรณกรรมและงานวิจัยที่เกี่ยวข้อง	3
1. ลักษณะของภาษาไทย	3
2. วิธีที่ใช้ตัดคำ	4
3. เทคนิคที่ใช้ในการตัดคำ	5
4. งานวิจัยที่เกี่ยวข้อง	6
บทที่ 3 การตัดคำด้วยกฎที่ปรับปรุงและพจนานุกรมแบบใหม่	15
1. การตัดอนุประโยคโดยอาศัยช่องว่าง และ อักขระพิเศษ	15
2. การตัดคำโดยอาศัยกฎการผสมอักษรในภาษาไทย	15
3. วิธีการตัดคำที่นำเสนอ	16
4. การสร้างพจนานุกรมแบบใหม่	18
5. การวิเคราะห์คำที่ไม่มีอยู่ในพจนานุกรมโดยเทียบกฎความเป็นไปได้ที่จะเป็นภาษาต่างประเทศหรือคำเฉพาะ	19
6. การตัดคำในส่วนสุดท้ายของอนุประโยค	20
7. ตัวอย่างวิธีการตัดคำด้วยกฎปรับปรุงและพจนานุกรมแบบใหม่	21

สารบัญ (ต่อ)

		หน้า
บทที่ 4	โครงสร้างของระบบตัดคำภาษาไทยด้วยกฎที่ปรับปรุงและพจนานุกรมแบบใหม่	25
1.	ส่วนของกฎ	25
2.	ส่วนของคำ	27
3.	ส่วนของคำที่ไม่อยู่ในพจนานุกรม	27
4.	ส่วนของพจนานุกรม	27
5.	การออกแบบโปรแกรม	28
บทที่ 5	ผลการตัดคำ	29
1.	การเตรียมการทดลอง	29
2.	การทดลอง	30
3.	การวัดผลการตัดคำภาษไทย	30
4.	ผลการทดลอง	30
บทที่ 6	บทสรุปและข้อเสนอแนะ	32
1.	ประสิทธิภาพของการตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่	32
2.	ข้อเสนอแนะ	32
บรรณานุกรม		33
ภาคผนวก		35
ภาคผนวก ก บทความของผู้ทำวิจัย		35
ภาคผนวก ข ตัวอย่างของเอกสารที่ใช้ในการตัดคำและผลการตัดคำ		49

สารบัญภาพ

	หน้า
ภาพที่ 1 การทำงานของเล็กซ์	12
ภาพที่ 2 การทำงานของระบบการตัดคำภาษาไทย [1]	12
ภาพที่ 3 ภาพรวมของขั้นตอนวิธีการตัดคำภาษาไทย ด้วยวิธีปรับปรุงกฎและพจนานุกรมแบบใหม่	21
ภาพที่ 4 การเก็บคำศัพท์เข้าไว้ในพจนานุกรม	28
ภาพที่ 5 ความถูกต้องของการตัดคำระดับพยางค์และระดับคำ โดยเปรียบเทียบที่ขนาดของข้อมูลแตกต่างกัน	31
ภาพที่ 6 เวลาที่ใช้ในการตัดคำภาษาไทยโดยเปรียบเทียบที่ขนาดของข้อมูลแตกต่างกัน	31

สารบัญตาราง

		หน้า
ตารางที่ 1	แสดงตัวอย่างของการสะกดคำตามอักขรวิธี	3
ตารางที่ 2	แสดงความถี่ของคำที่พบได้บ่อยในเอกสารภาษาไทย	8
ตารางที่ 3	แสดงผลการตัดคำด้วยการใช้กฎที่ปรับปรุงและพจนานุกรมแบบใหม่ร่วมกัน	30

บทที่ 1
บทนำ

1. ความเป็นมาและความสำคัญของปัญหา

ภาษาไทยเป็นภาษาที่มีการเขียนติดกันไปทั้งประโยคโดยไม่มีการเว้นช่องว่างระหว่างคำอย่างในภาษาอังกฤษหรือสามารถเขียนให้อยู่ในรูปของอักษรตัวเดียวอย่างในภาษาจีน ดังนั้นการประมวลผลทางภาษาที่เรียกว่าการตัดคำ (Word segmentation หรือ Word separation) จึงเป็นการแบ่งคำแต่ละคำในประโยคออกจากกันเพื่อวัตถุประสงค์ในการใช้ประโยชน์ในด้านต่างๆ เช่น การจัดรูปแบบเอกสารในการประมวลผลคำ (Word processing) การตรวจสอบตัวสะกดภาษาไทย (Spelling check) การวิเคราะห์ไวยากรณ์ (Syntax analysis) การแปลภาษาด้วยเครื่องจักร (Machine translation) การทำดรรชนีสำหรับเอกสาร (Document indexing) การเชื่อมโยงความหมายของคำ (Thesaurus) การประมวลผลภาษาธรรมชาติ (Natural language processing) การสังเคราะห์เสียงพูด (Speech synthesis) การวิเคราะห์กฎเกณฑ์ในการสร้างประโยค (Syntactic rules analysis) เป็นต้น

การตัดคำภาษาไทยถูกพัฒนาในรูปแบบต่างๆ โดยวิธีหลักที่ใช้มีอยู่ 3 วิธีหลักๆ คือ การใช้กฎ (Rule-based) การใช้พจนานุกรม (Dictionary) และการใช้คลังข้อความ (Corpus) ข้อดีของการใช้กฎคือมีความเร็วสูง มีความถูกต้องหลังการตัดในระดับพยางค์ค่อนข้างสูงแต่ในระดับคำทำได้ไม่ดีนัก เนื่องจากการใช้กฎเป็นการใช้หลักการประสมกันของสระ พยัญชนะและวรรณยุกต์จึงทำได้เพียงคำสั้นๆ ไม่สามารถตัดคำที่ประกอบด้วยคำหลายพยางค์ได้ เช่น คำประสม เป็นต้น ส่วนวิธีการใช้พจนานุกรมและการใช้คลังข้อความเมื่อพิจารณาแล้วยังคงมีการใช้กฎร่วมด้วยเป็นส่วนใหญ่เพื่อจัดการกับคำที่ไม่รู้จักและคำที่ไม่มีอยู่ในพจนานุกรม เช่น คำเฉพาะบางคำ คำที่มาจากภาษาต่างประเทศเป็นต้น ซึ่งในปัจจุบันนี้จะพบคำที่มาจากภาษาต่างประเทศมากขึ้นและมีการถอดความเป็นภาษาไทยไม่ตรงกับหลักไวยากรณ์ไทยตามหลักการถอดความเป็นภาษาไทยของบัณฑิตยสถาน เช่น เลาน์จ (Lounge) ดังนั้นการพัฒนาการตัดคำด้วยกฎให้มีประสิทธิภาพยิ่งขึ้นจึงเป็นการสนับสนุนการตัดคำด้วยวิธีอื่นๆ ไม่ว่าจะเป็นการใช้พจนานุกรมหรือคลังข้อความ

2. วัตถุประสงค์ของงานวิจัย

1. เพื่อศึกษาและพัฒนาอัลกอริทึม (Algorithm) ในการตัดคำเอกสารภาษาไทยโดยการใช้กฎ (Rule-based) ให้มีประสิทธิภาพมากขึ้น
2. เพื่อพัฒนาขั้นตอนวิธีในการตัดคำภาษาไทยโดยการใช้กฎ (Rule-based) ให้มีความถูกต้องสูงขึ้น

3. ขอบเขตของการวิจัย

ศึกษาและพัฒนาขั้นตอนวิธีการตัดคำในภาษาไทยโดยใช้วิธีที่เรียกว่าการใช้กฎ (Rule-based) เพื่อแก้ไขปัญหาคำเฉพาะและคำที่มาจากภาษาต่างประเทศโดยเน้นที่คำมาจากภาษาต่างประเทศร่วมกับการใช้พจนานุกรมแบบใหม่

4. สถานที่วิจัย

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น

5. ขั้นตอนและวิธีการดำเนินงาน

1. ค้นคว้าและศึกษาการทำงานพื้นฐานของการตัดคำ
2. ศึกษาเทคนิคการตัดคำและการรวมคำมูลเข้าด้วยกัน
3. วางขอบเขตและกำหนดเป้าหมายของงานวิจัย
4. ออกแบบและพัฒนาอัลกอริทึม
5. ออกแบบการทดลอง
6. พัฒนาโปรแกรมและทำการทดลอง
7. สรุปผลการทดลองเขียนรายงานสรุปผล
8. เขียนรายงาน
9. นำเสนอรายงาน

6. ประโยชน์ที่คาดว่าจะได้รับจากโครงการวิจัยนี้

ได้ขั้นตอนวิธีในการตัดคำภาษาไทยโดยใช้กฎ (Rule-based) ที่ตัดคำจากเอกสารภาษาไทยได้ถูกต้องยิ่งขึ้นโดยเฉพาะในเอกสารที่มีคำมาจากภาษาต่างประเทศ

บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

การตัดคำภาษาไทย (Thai words segmentation) ได้รับการพัฒนาขึ้นมาโดยใช้วิธีการต่างๆ ที่ต่างกัน เนื่องจากการตัดคำเป็นกระบวนการพื้นฐานของการประมวลผลภาษาธรรมชาติ เช่น การวิเคราะห์เสียงพูด การตัดคำภาษาไทยเองก็เช่นกัน ได้มีผู้คิดค้นวิธีที่จะแยกคำแต่ละคำออกจากประโยคซึ่งมีการเขียนติดกันไปอย่างต่อเนื่องทั้งประโยค ในงานวิจัยนี้จะกล่าวถึงการตัดคำโดยอาศัยอักขรวิธี เป็นหลักการพื้นฐานการประสมคำ

1. ลักษณะของภาษาไทย

ภาษาไทยมีลักษณะแตกต่างจากภาษาอังกฤษ หรือภาษาจีน เนื่องจากในภาษาไทยมีการเขียนติดกันไปทั้งประโยค อีกทั้งคำไทยคำหนึ่งอาจประกอบไปด้วยสระที่เป็นสระประกอบ คือมาจากสระอื่นอีกหลายตัวประกอบกัน เช่น สระเอียะ สระเอือะ เป็นต้น และพยัญชนะบางตัวยังสามารถทำหน้าที่เป็นได้ทั้งตัวสะกด หรือสระด้วยก็ได้ ดังนั้นการแยกแยะในหน่วยย่อยของคำสามารถนำหลักเกณฑ์ที่เรียกว่าอักขรวิธีมาใช้

1.1 อักขรวิธี

คำในภาษาไทยเกิดจากส่วนต่างๆ ของอักษรไทยประกอบกันอย่างน้อยสามส่วน ได้แก่ ส่วนพยัญชนะ สระ และวรรณยุกต์ พยัญชนะของไทยถูกแบ่งออกเป็นอักษรสามหมู่ที่เรียกว่าไตรยางค์ ได้แก่ อักษรสูง อักษรกลาง และอักษรต่ำ สระก็ถูกจัดเป็นประเภท สระเดี่ยว สระประสม วรรณยุกต์ก็มี 4 รูป 5 เสียง การจะทำให้เกิดเสียงและความหมายต้องเกิดจากกฎเกณฑ์ที่มีอยู่

ตารางที่ 1 แสดงตัวอย่างของการสะกดคำตามอักขรวิธี

ส่วน	ลักษณะ	ตัวอย่าง
3 ส่วน	พยัญชนะ + สระ + วรรณยุกต์	หู สา ที
4 ส่วน	พยัญชนะ + สระ + วรรณยุกต์ + ตัวสะกด	คน กิน ข้าว
5 ส่วน	พยัญชนะ + สระ + วรรณยุกต์ + ตัวสะกด + การันต์	แพทย์ สิทธิ์ ฤทธิ์

2. วิธีที่ใช้ตัดคำ

วิธีที่ใช้ตัดคำแบ่งออกเป็น 3 ประเภทใหญ่ๆ ได้แก่ การใช้กฎ การใช้พจนานุกรมและการใช้คลังข้อความ

2.1 การใช้กฎ

การตัดคำโดยการตรวจสอบกฎเกณฑ์ทางอักขรวิธีที่กำหนดลักษณะการประสมอักษร ลักษณะการเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการกำหนดขอบเขตของคำ วิธีการนี้จะมีข้อจำกัดในการทำงาน คือ ความถูกต้องของการตัดคำในระดับพยางค์สูงแต่ความถูกต้องของการตัดคำค่อนข้างต่ำ แต่ข้อดีของวิธีนี้คือมีความรวดเร็วในการทำงานและใช้ทรัพยากรน้อย

2.2 การใช้พจนานุกรม

การตัดคำโดยพจนานุกรมเป็นการตัดคำโดยใช้สายอักขระ (String) มาเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม ซึ่งวิธีนี้จะต้องทำการจัดเก็บคำไว้ในพจนานุกรม วิธีนี้ทำให้ได้ความถูกต้องในการตัดคำสูงกว่าการใช้กฎแต่จะใช้เวลามากกว่า

2.3 การใช้คลังข้อความ

การตัดคำโดยใช้คลังข้อมูลเป็นการตัดคำโดยนำวิธีการทางสถิติเข้ามาใช้ในการประมวลภาษา โดยใช้คลังข้อมูลทางภาษาเป็นฐานความรู้เก็บค่าความถี่ที่ใช้ในการตัดคำ ซึ่งการตัดคำโดยใช้คลังข้อมูลแบ่งออกเป็น 2 วิธี คือการตัดคำโดยอาศัยความน่าจะเป็น (Probabilistic word segmentation) และวิธีการตัดคำโดยอาศัยคุณลักษณะของคำ (Feature-based word segmentation)

วิธีการตัดคำโดยอาศัยค่าความน่าจะเป็นจะเป็นการตัดคำโดยใช้แบบจำลองเอนแกรม (Word n-gram model) ในการหารูปแบบของการตัดคำและลำดับคำที่เป็นไปได้มากที่สุด โดยวิธีการนี้จะต้องมีการใช้คลังข้อมูลที่มีการตัดคำและกำกับหมวดคำที่เตรียมเอาไว้แล้ว ซึ่งวิธีการนี้ผลลัพธ์ที่ได้จะเป็นการเลือกรูปแบบการตัดคำที่มีความน่าจะเป็นมากที่สุด

ตัวอย่างของแบบจำลองไตรแกรม

“การพัฒนาระบบถาม-ตอบ” จะได้ว่า

การ/ ารพ/รพั /พัฒ/ ัฒน/ ฒนา/ นาร/ ารระ/ระบ/บบถ/บถา/ถาม/ ามต/มตอ/ตอบ

หลังจากนั้นจะทำการเลือกคำที่เป็นไปได้เพื่อทำการประมวลผลต่อไปอย่างไร

วิธีการตัดคำโดยอาศัยคุณลักษณะของคำ จะเป็นการแก้ข้อผิดพลาดของการตัดคำโดยอาศัยค่าความน่าจะเป็นของการจำกัดหมวดคำที่จะเป็นแบบจำลองในการตัดคำ ซึ่งวิธีการตัดคำโดยอาศัยคุณลักษณะของคำจะเป็นวิธีการแบบผสม (Hybrid approach)

3. เทคนิคที่ช่วยในการตัดคำ

เทคนิคที่ใช้ในการตัดคำที่นิยมใช้กันทั่วไปคือ วิธีการเทียบคำที่ยาวที่สุด วิธีการเทียบคำที่สั้นที่สุด วิธีการตัดคำที่ใช้ความถี่ของคำและวิธีการย้อนรอยกลับ

3.1 วิธีการเทียบคำที่ยาวที่สุด (Longest word pattern matching)

วิธีนี้จะทำการตรวจสอบสายอักขระ (String) ที่นำเข้ามาจากซ้ายไปขวา จากนั้นนำไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม หากตรวจสอบพบว่าพบพยางค์มากกว่า 1 พยางค์ในพจนานุกรม จะทำการเลือกพยางค์ที่ยาวที่สุดแล้วทำต่อไปเรื่อยๆ จนจบสายอักขระ

ตัวอย่างคำว่า “กองกลาง”

การตัดคำโดยวิธีนี้จะนำสายอักขระไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรมจะพบคำว่า ก , กอ และคำว่า กอง ส่วนคำว่า กองก ไม่พบอยู่ในพจนานุกรม ดังนั้นจึงได้คำว่า กอง ซึ่งเป็นคำที่ยาวที่สุดที่หาพบ ส่วนที่เหลือคือ กลาง เมื่อนำไปค้นในพจนานุกรมจะได้ว่า

ก , กล , กลาง ดังนั้นจึงเลือกคำว่า กลาง

คำที่ได้จากการตัดคำโดยวิธีการนี้จึงเป็น กอง กลาง

วิธีการนี้ให้ความถูกต้องหลังการตัดคำสูงกว่าวิธีการอื่นโดยเฉพาะเมื่อใช้ร่วมกับวิธีย้อนรอยกลับ

3.2 วิธีการเทียบคำที่สั้นที่สุด (Shortest word pattern matching)

วิธีการนี้คล้ายกับวิธีการเทียบคำที่ยาวที่สุด เพียงแต่จะเลือกคำที่สั้นที่สุดที่พบก่อน แต่วิธีนี้พบว่าได้จำนวนคำมากที่สุดแต่ความถูกต้องของคำหลังทำการตัดคำน้อยกว่าการใช้วิธีเทียบคำที่ยาวที่สุด

ตัวอย่างคำว่า “โคลงเรือ”

การตัดคำโดยวิธีนี้จะเลือกเอาคำแรกที่ค้นหาเจอจากพจนานุกรม ดังนั้นจะได้ว่า

โค ลง เรือ (โดยไม่เลือกคำว่า “โคลง” ที่จะพบต่อไปภายหลังหากทำการค้นหาต่อ)

วิธีนี้ใช้เวลาน้อยกว่าการเทียบคำยาวที่สุด แต่ความถูกต้องที่ได้การตัดคำแบบเทียบคำยาวที่สุดจะมากกว่า

3.3 วิธีการตัดคำที่ใช้ความถี่ของคำ (Word usage frequency)

วิธีการนี้เป็นแนวทางหนึ่งในการแก้ปัญหาคำกำกวมของประโยคภาษาไทยโดยการวิเคราะห์ความถี่ของการใช้คำในชีวิตประจำวันโดยจัดเรียงคำในพจนานุกรมตามความถี่ที่พบ และใช้วิธีการตัดคำแบบเดียวกับข้อ 3.1 และ 3.2

ตัวอย่างคำว่า “ก็อต”

ในกรณีนี้หากใช้ความถี่ของคำจะได้ว่า

ก็ อด (เนื่องจาก ก็ มีความถี่สูงกว่าคำว่า ก็อด ความถี่ของคำสามารถดูได้ที่ตารางที่ 2 หน้า 8)

3.4 วิธีการย้อนรอยกลับ (Back tracking)

เมื่อทำการเปรียบเทียบคำที่นำมาตัดคำกับคำที่มีอยู่ในพจนานุกรม อาจพบกรณีที่คำที่พบมีมากกว่า 1 คำแล้วทำการเลือกคำที่ยาวที่สุดทำให้สายอักขระที่ตามมาจากคำนั้นไม่สามารถตัดคำได้ เนื่องจากไม่พบตามพจนานุกรม กรณีนี้จะทำการย้อนไปอีกคำที่ไม่ถูกเลือกแล้วทำการตัดคำต่อไป

ตัวอย่างเช่นคำว่า “เมื่อยามนี้” การเปรียบเทียบกับพจนานุกรมจะได้ว่า

เมื่อ , เมื่อย ดังนั้นจึงเลือกคำที่ยาวที่สุดจะได้คำว่า เมื่อย

ส่วนที่เหลือคือ –ามนี้ ซึ่งไม่พบอยู่ในพจนานุกรม ดังนั้นจะทำการย้อนกลับไปเพื่อเลือกอีกคำหนึ่งคือ เมื่อ จะได้เป็น

เมื่อ ยาม นี้ (โดยคำว่า ยามเกิดจากการเลือกคำที่ยาวที่สุดระหว่าง ยา และ ยาม)

4. งานวิจัยที่เกี่ยวข้อง

4.1 งานของสุรินทร์ จรรยาพรพงษ์ [8]

เป็นงานวิจัยที่ตัดคำโดยใช้กฎ โดยกฎที่นำมาใช้นั้นได้มาจากไวยากรณ์ไทยและวิเคราะห์ลักษณะของการใช้ โดยลักษณะของกฎแบ่งได้เป็น 2 ชนิด คือ กฎการหาขอบเขตหน้า (Front boundary recognition rule) และกฎการหาขอบเขตหลัง (Tail boundary recognition rule) และในแต่ละกลุ่มยังแบ่งออกเป็น 2 กลุ่มย่อยๆ คือแบ่งตามคุณสมบัติของตัวอักษรโดยกฎที่ได้ออกมานี้จะจัดให้อยู่ในกลุ่มเอ (Group A) และแบ่งตามคุณสมบัติของรูปแบบการใช้สระแต่ละตัวซึ่งกฎที่ได้จะจัดไว้ในกลุ่มบี (Group B)

4.1.1 กฎที่ได้จากคุณสมบัติของอักษรในการหาขอบเขตหน้าของพยางค์

กฎ A-1F : สระต่างๆ เหล่านี้ อะ อา อิ อี อึ อื อุ อู ไต่คู้ อำ ไม้หันอากาศ และวรรณยุกต์ทั้งหมด จะต้องมีพยัญชนะอยู่ข้างหน้าอย่างน้อย 1 ตัวอักษรเสมอ

กฎ A-2F : สระ เ แ ไ โ จะเป็นตัวอักษรแรกของพยางค์เสมอยกเว้นมีพยัญชนะนำหน้า

กฎ A-3F : สระ ใ จะเป็นตัวอักษรแรกของพยางค์เสมอ โดยไม่มีข้อยกเว้น

กฎ A-4F : พยัญชนะ ฉ ผ ฝ ฮ จะต้องเป็นพยัญชนะต้นเสมอยกเว้นมีสระเหล่านี้นำหน้า เ แ ไ โ ใ นำหน้า

กฎ A-5F : มีเพียง 9 คำในภาษาไทยเท่านั้นที่ใช้ ฤ เป็นพยัญชนะต้น เช่น ฤทัย ฤดี นอกนั้นจะพบ ฤ เดี่ยวๆ ยกเว้นคำว่า อมฤต

กฎ A-6F : ห มักพบเป็นคำนำหน้าพยางค์ ยกเว้นบางคำเช่น สห มหา คหบดี มหกรรม หมรสพ คหกรรม เป็นต้น

กฎ B-1F : จากกฎ A1-F ถ้ามีคำที่นำด้วยพยัญชนะสองตัว จะทำการตรวจสอบว่าบางตัวอาจทำหน้าที่เป็นสระ

4.1.2 กฎที่ได้จากคุณสมบัติของอักษรในการหาขอบเขตหลังของพยางค์

กฎ A-1T : พยัญชนะต่อไปนี้ ศ ณ ญ ษ ฐ ฎ ฏ ฒ พ ฌ จะเป็นตัวสะกดเสมอ ยกเว้นพยางค์ต่อไปนี้ ศก ศร ญวน ศตวรรษ และพยางค์อื่นๆ อีกแต่สามารถจัดการโดย A-1F

กฎ A-2T : สระ อึ จะต้องมีตัวสะกดหนึ่งตัวเสมอ ยกเว้น รึ หึ ฮึ

กฎ A-3T : ไม้หันอากาศ จะต้องมีตัวสะกดอย่างน้อยหนึ่งตัวเสมอ

กฎ A-4T : สระ อ และสระ อำ จะต้องตามพยัญชนะเสมอ

กฎ A-5T : การันต์ มักจะปรากฏเป็นตัวสุดท้ายของพยางค์ยกเว้นบางคำเช่น ปาล์ม ฟาร์ม

กฎ A-6T : ไม้หันอากาศจะต้องปรากฏระหว่างพยัญชนะสองตัว เช่น คัน ขัน

กฎ A-7T : เมื่อพยัญชนะต้นตามด้วย ว จะมีตัวสะกดอีก 1 ตัวในกรณีที่ ว ทำหน้าที่เป็นสระ

4.1.3 ตัวอย่างกฎที่ได้จากคุณสมบัติการใช้สระในการหาขอบเขตหลังของพยางค์

กฎ B-1T : สระเหล่านี้ ไม้หันอากาศ อิ อึ อี ถ้ามีวรรณยุกต์แล้วจะต้องมีตัวสะกด 1 ตัวเสมอ

กฎ B-2T : สระ อี ที่ปรากฏวรรณยุกต์นอกจากไม้ตรี ส่วนใหญ่จะไม่มีตัวสะกด

กฎ B-3T : สระ ◌ื เ◌็◌ แ◌็◌ ◌็อ◌ และ เ◌ิ◌ ต้องมีตัวสะกด 1 ตัว

กฎ B-4T : สระในรูปแบบดังต่อไปนี้ อัว อัย อัวะ อือ เอา เอาะ เอะ แอะ โอะ เอียะ เออะ ไม่ต้องการตัวสะกด ยกเว้นบางพยางค์ในรูปแบบดังนี้ เอา- เออ- โดยเครื่องหมาย – แทนพยัญชนะ 1 ตัวอักษร

กฎ B-5T : มีเพียง 20 คำที่ใช้ ใ

กฎ B-6T : สระจากคำ ไ- ที่มีวรรณยุกต์ จะครบพยางค์ เนื่องจาก ไ ไม่มีตัวสะกด มีเพียงบางคำที่นอกเหนือจากนี้ เช่น ไฮ่ย์

งานวิจัยนี้ยังสามารถตัดคำให้อยู่ในรูปแบบพยางค์หรือคำสั้นๆ เพียงแค่นั้น นอกจากนี้กฎที่มีอยู่ยังไม่สามารถรองรับคำที่มีการสะกดแตกต่างออกไป เช่น คำที่มาจากภาษาต่างประเทศ หรือคำเฉพาะ เป็นต้น

4.2 งานของยืน ภู่วรวรรณและวิวรรธน์ อิ่มอารมณ์ [7]

ได้ทำการวิเคราะห์ข้อมูลคำไทยโดยได้สุ่มตัวอย่างหนังสือและเอกสารต่างๆ ซึ่งได้แก่ หนังสือพิมพ์ วารสาร นิตยสาร พ็อกเก็ตบุค รายงาน จดหมายราชการ หนังสืออ่านทั่วไป โดยยกเว้น หนังสือประเภทวรรณคดีและตำราวิชาการที่แปลมาจากต่างประเทศ โดยทำการป้อนข้อมูลเพื่อหาความถี่ของคำที่ใช้ในชีวิตประจำวันดังตารางที่ 2

ตารางที่ 2 แสดงความถี่ของคำที่พบได้บ่อยในเอกสารภาษาไทย

คำ	เปอร์เซ็นต์ความถี่
ที่	2.60
การ	2.04
เป็น	1.55
ได้	1.44
จะ	1.40
ใน	1.32
มี	1.29
ไม่	1.17
ก็	1.16
ของ	1.15
ให้	1.11
ว่า	1.10
ไป	1.07
และ	1.06
มา	0.93
ความ	0.86
ประ	0.80
นี้	0.79
ทำ	0.64
คน	0.64
ผู้	0.63
กัน	0.61

คำ	เปอร์เซ็นต์ความถี่
แล้ว	0.61
แต่	0.61
จาก	0.58
อย่าง	0.58
นั้น	0.57
อยู่	0.55
กับ	0.51
ต้อง	0.49
ทาง	0.45
หรือ	0.42
งาน	0.41
ด้วย	0.41
ใจ	0.39
ขึ้น	0.37
ถึง	0.37
ต่อ	0.36
เข้า	0.35
รับ	0.34

สำหรับรูปแบบคำไทยนั้นได้ทำการแบ่งส่วนประกอบออกเป็น 5 กลุ่มได้แก่

C หมายถึง พยัญชนะต้นรวมทั้งการควบกล้ำ

V : สระ

S : ตัวสะกด

T : วรรณยุกต์

G : การันต์

ตัวอย่างเช่น

สิ้น = CTVS

ศิลป์ = CVSSG

กวาง = CCVS

เมื่อไม่พบคำในพจนานุกรม จะทำการนำไปตรวจสอบกับกฎ โดยงานวิจัยนี้มีกฎอยู่ 18 กฎด้วยกัน โดยพิจารณาตัวอักษรเริ่มต้นหรือตัวสิ้นสุดพยางค์เป็นตัวตัดสิน โดยจะเป็นกฎอย่างง่ายและไม่ครอบคลุมการสะกดคำไทยที่มีอยู่ได้ทั้งหมด ตัวอย่างเช่น

<u>กฎข้อที่ 1</u> กฎ เ แ ไ ใ โ

ให้ตัดหน้า เ แ ไ ใ โ โดยมีข้อยกเว้นคำต่อไปนี้ มโน , ชโล , นโย , วโร , สโม , อโห , จเร , ขโมย , พเน , อเนก , อโศก

<u>กฎข้อที่ 2</u> กฎตัว ฉ ผ ฝ ฮ

ตัดหน้าอักษรเหล่านั้นเป็นสระ เ แ โ ไ และ ใ

...

<u>กฎข้อ 18</u> ทัณฑฆาต

ตัดข้างหลัง เช่น ศักดิ์ พงษ์

ยกเว้นคำจากต่างประเทศ เช่น ปาล์ม บาล์ม เป็นต้น

งานวิจัยนี้ได้นำพจนานุกรมร่วมกับการใช้กฎเพียง 18 กฎซึ่งไม่เพียงพอต่อการตัดคำที่มีอยู่โดยเฉพาะคำที่มีการแปลมาจากคำอ่านภาษาต่างประเทศ ตัวอย่างเช่น เลาน์จ เพาน์ด อีกทั้งเอกสารที่นำมาทำการตัดคำเป็นเอกสารที่มีการสะกดด้วยภาษาไทยเพียงอย่างเดียวโดยไม่มีภาษาต่างประเทศปนอยู่เลย

4.3 งานของดวงแก้ว สวามิภักดิ์ [1]

งานวิจัยนี้ได้สร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกซ์ เป็นงานวิจัยการตัดคำภาษาไทย โดยใช้กฎที่สร้างขึ้นเองจำนวน 43 กฎ ตังตารางที่ 3 และมีการนำพจนานุกรมเข้ามาร่วมด้วย โดยสาเหตุที่นำทั้งกฎไวยากรณ์และพจนานุกรมเข้ามาช่วยในการตัดคำนั้นก็เพื่อจะแก้ไขปัญหาการตัดคำโดยใช้พจนานุกรมเพียงอย่างเดียว ซึ่งไม่สามารถตัดคำได้อย่างถูกต้องในกรณีคำนั้นไม่มีอยู่ในพจนานุกรม

งานวิจัยนี้ได้ทำภายใต้ระบบปฏิบัติการยูนิกซ์และได้มีการนำโปรแกรมเล็กซ์ (Lex) เข้ามาร่วมด้วย โดยกฎที่ได้มานี้จะไม่มีการรวมตัวสะกดเข้าไปในกฎด้วยยกเว้นบางกรณี เนื่องจากโปรแกรมเล็กซ์จะพยายามสร้างกลุ่มอักษร (Token) ที่ยาวที่สุดก่อน ดังนั้นหากมีการนำตัวสะกดเข้ามาใช้จะเป็นสาเหตุให้มีการรวมเอาอักษรตัวหน้าของคำถัดไปมาเป็นตัวสะกดได้ ซึ่งเมื่อได้มีการวิเคราะห์ด้วยกฎแล้ว ขั้นตอนต่อไปจะมีการรวมกลุ่มตัวอักษรเข้าด้วยกัน โดยทำการตรวจสอบจากพจนานุกรม ส่วนโครงสร้างของพจนานุกรมที่นำมาใช้เป็นฐานข้อมูลแบบรีเลชัน (Relation DBMS) ซึ่งใช้คำเป็นดรรชนี (Index) และไฟล์ดรรชนีได้พัฒนาขึ้นโดยใช้โครงสร้างข้อมูลแบบบี-ทรี (B-Tree)

4.3.1 กฎที่ใช้ในการตัดพยางค์มีอยู่ทั้งหมด 43 กฎ โดยใช้สัญลักษณ์ดังนี้

c (Consonant) : พยัญชนะปรกติ

v (Vowel) : สระ

t (Tonal Mark) : วรรณยุกต์

s (Speller) : ตัวสะกด

[…]? : ทางเลือก อาจมีหรือไม่ก็ได้

[a1 a2 a3 … an] : ทางเลือก

4.3.2 กฎ 43 ข้อสำหรับตัดคำภาษาไทยของดวงแก้ว สวามิภักดิ์

1. [c][t]?[ะ ำ า] เช่น ระ ม้า คำ
2. [c][ิ ี ื ุ ู][t]? เช่น ยี รู้
3. [c][ื][t]?[s] เช่น ลืม ลื่น
4. [c][t]?[ุ ู] เช่น รู้ ผุ
5. [c] ั [t]?[s] เช่น ค้น คน
6. [เ แ โ ใ ไ][c][t]? เช่น แท้ ใย
7. [เ แ] [c] ็ [s] เช่น เจ็บ เยน
8. [เ แ โ][c][t]? ะ เช่น แกะ เก๊ะ
9. [เ แ โ][ก ข ค ต ท บ ป พ ฟ จ ซ ศ ส] ร[t]? ะ เช่น แครระ
10. [เ แ โ][ก ข ค บ ป ผ พ]ล[t]?ะ เช่น แกละ เผละ

11. เ[ก ข ค ต ท บ ป พ ฟ จ ซ ศ ส]ร[า ิ ี ึ ื ุ ู][t]? เช่น เครา เครื่
12. เ[ก ข ค บ ป พ ฟ]ล[า ิ ี ึ ื ุ ู][t]? เช่น เคลื่
13. [เ แ โ][ก ข ค]ว[t]?ะ เช่น เก๊ะ แคะ
14. เ[c] ี [t]?ย เช่น เสีย เรี่ย
15. เ[c][t]?าะ เช่น เผาะ เก้าะ
16. เ[c][t]?[าะ] เช่น เตะ เว้า
17. เ[c][t]? เช่น เร เร่
18. [โ ใ ไ]ห[ง ญ น ม ย ร ล ว][t]? เช่น ใหล่ โหล
19. เ[c] ื [t]?อ เช่น เผื่อ เสือ
20. [c] ื [t]?อ เช่น มือ มื้อ
21. [c] ์ เช่น ก์ ส์
22. [็ ่ ้ ๊ ๋ + ะ ั า ำ ิ ี ึ ื ุ ู] เช่น ะ ื
23. ฯ
24. เ[จ ส]ริญ ได้แก่ เจริญ เสริญ
25. ห ร[t]?[ะ า ำ ุ ู] เช่น หรู่ หระ
26. ห ร[ิ ี ึ ื][t]? เช่น หรี่ หริ
27. เ[ก ข ค บ ป ผ พ ฟ]ลื[t]?อ เช่น เกลือ เกลื้อ
28. เ[ก ข ค ต ท บ ป พ ฟ]รื[t]?อ เช่น เครือ เครื่อ
29. เ[ก ข ค บ ป ผ พ ฟ]ลิ[t]?[s] เช่น เพลิน เคลิ้ม
30. เ[ก ข ค ต ท บ ป พ ฟ]ริ[t]?[s] เช่น เกริ่น เกริก
31. แ[ก ข ค ต ท บ ป พ ฟ]ร[t]?[ก ง ว น บ ม ท ต ด] เช่น แทรก แคร้ง
32. แ[ก ข ค บ ป ผ พ ฟ]ล[t]?[ก ง ว น บ ม ท ต ด] เช่น แปลน แกล้ง
33. เห[ง ญ น ม ย ร ล ว][t]?า เช่น เหล้า เหลา
34. [A-Z a-z เครื่องหมายพิเศษต่างๆ]* เช่น A1a thisis
35. ฯลฯ
36. ๆ
37. เ[ก ข ค]วี[t]?ย เช่น เกวีย
38. เ[ก ข ค ต ท บ ป พ ฟ]รี[t]?ย เช่น เตรีย
39. เ[ก ข ค ง บ ผ พ ฟ]ลี[t]?ย เช่น เกลี่ย
40. ช่องว่าง

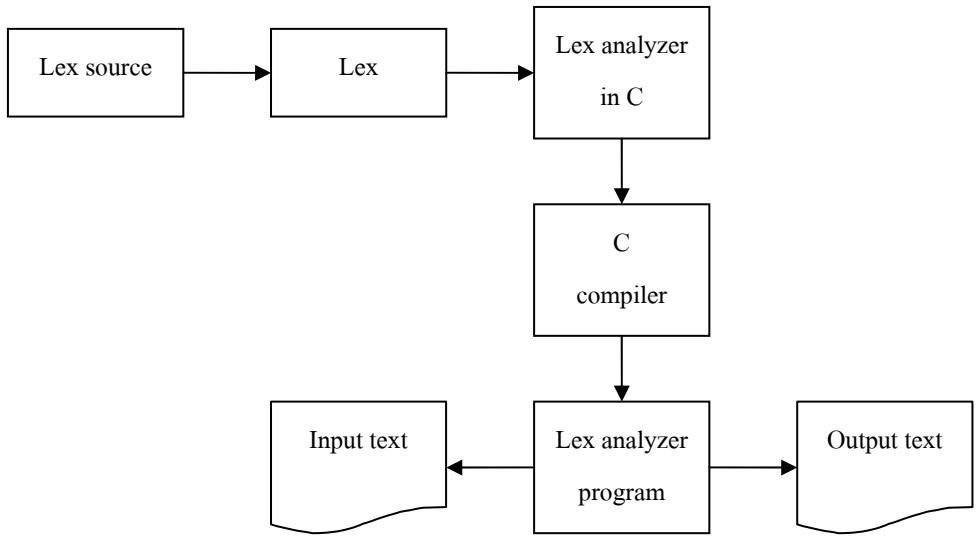
41. [0-9]* เช่น 015789645 1234

42. “\n”

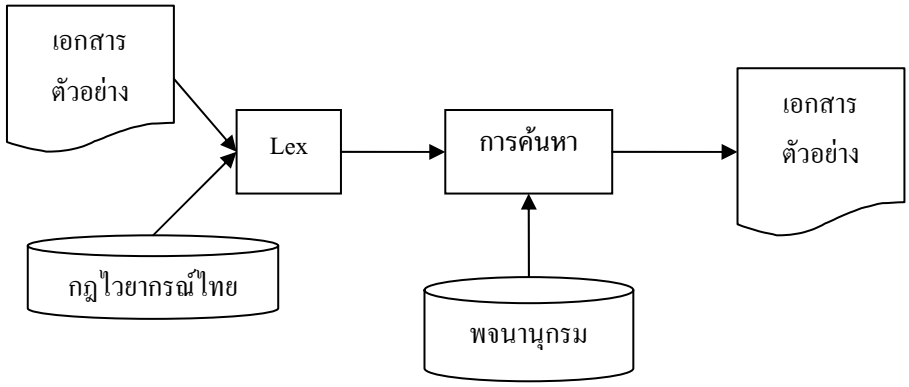
43. ทุกตัวอักษรที่ไม่อยู่ในกฎ 1-42

4.3.3 เล็กซ์

เล็กซ์เป็นซอฟต์แวร์ที่ทำหน้าที่สร้างส่วนโปรแกรมภาษาซี ที่ทำหน้าที่วิเคราะห์เล็กซิเคิล (Lexical analysis) โดยผู้ใช้ระบุกฎในการวิเคราะห์พร้อมทั้งสิ่งที่ต้องการกระทำถ้าเจอข้อความที่ถูกต้องตามกฎ ส่วนกฎเกณฑ์ต่างๆ นี้จะถูกส่งให้กับเล็กซ์ โดยที่เล็กซ์จะทำการสร้างฟังก์ชันภาษาซี ซึ่งสามารถคอมไพล์ด้วยคอมไพเลอร์ภาษาซีได้ และฟังก์ชันเหล่านี้ก็สามารถถูกเรียกใช้ได้จากโปรแกรมทั่วๆ ไป โค้ดที่ได้รับจากการคอมไพล์จะทำงานได้โดยผู้ใช้ส่งเอกสารที่ต้องการวิเคราะห์เข้าไปเสมือนหนึ่งเป็นข้อมูลนำเข้า (Input) ตามปกติ ซึ่งฟังก์ชันที่เล็กซ์สร้างมานี้ก็จะแยกเอกสารเหล่านี้ออกเป็นกลุ่มอักษร ขั้นตอนต่างๆ ที่กล่าวมานี้ได้แสดงไว้ในภาพที่ 1



ภาพที่ 1 การทำงานของเล็กซ์



ภาพที่ 2 การทำงานของระบบการตัดคำภาษาไทย [1]

เนื่องจากงานวิจัยนี้ใช้เล็กซ์ เป็นเครื่องมือที่จะทำการวิเคราะห์โดยยึดหลักการให้ได้กลุ่มอักษรที่ยาวที่สุดและไม่มีการย้อนรอยกลับ ทำให้เกิดกรณีที่คำที่ต้องการนำมาตัดนั้นไม่ตรงกับไวยากรณ์และไม่สามารถจับคู่กับไวยากรณ์ที่เหมาะสม เช่นคำว่า เคราน์ช (เช่นกรณีเป็นชื่อต่างประเทศ) อีกทั้งไวยากรณ์ที่ได้พัฒนาขึ้นทั้ง 43 กฎ ยังไม่สามารถครอบคลุมการสะกดของภาษาไทยที่อยู่นอกเหนือจากอักขรวิธี เช่นคำที่มาจากภาษาต่างประเทศหรือชื่อเฉพาะบางคำซึ่งจะถูกกำหนดไว้ให้เป็นอื่นๆ ตามกฎข้อสุดท้าย

งานวิจัยนี้มีจุดเด่นคือการตัดคำในระดับพยางค์และคำค่อนข้างทำได้สูง แต่เนื่องจากเล็กซ์ไม่สนับสนุนการตัดคำแบบย้อนรอยทำให้การตัดคำบางคำไม่ถูกต้อง โดยกลุ่มของคำที่ตัดไม่ถูกต้องจะมีลักษณะอยู่นอกเหนือจากลักษณะการสะกดตามหลักไวยากรณ์ไทยหรือเป็นคำเฉพาะหรือคำที่มาจากภาษาต่างประเทศ

4.4 งานของไพศาล เจริญพรสวัสดิ์ [5]

งานวิจัยนี้แบ่งปัญหาในการตัดคำออกเป็น 2 ชนิดด้วยกันคือ

4.4.1 ปัญหาความกำกวม

4.4.2 ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม

การแก้ปัญหาการตัดคำได้นำคุณลักษณะโดยการใช้การเรียนรู้ของเครื่อง 2 แบบ คือริปเปอร์และวินโนว์ ซึ่งนำคุณสมบัติและบริบทของคำที่อยู่รอบๆ มาใช้ในการแก้ปัญหาการตัดคำโดยใช้สถิติเข้ามาช่วย

$$\tau = \max \arg PROVB(C_1,...,C_t \mid w_1,...,w_t$$

โดยที่ τ คือ $C_1,...C_t$ ที่ให้ค่าความน่าจะเป็นมีค่ามากที่สุด

C_i คือหน้าที่คำจองคำ W_i

W_i คือลำดับของประโยคหนึ่งๆ

เนื่องจากการตัดคำโดยวิธีนี้ต้องอาศัยหน้าที่ของคำ แต่เนื่องจากคำบางคำที่ไม่ปรากฏในพจนานุกรมหรือคำทียังไม่รู้จักแล้วทำให้การตัดคำนี้ยังทำไม่ได้ทันที ต้องทำการกำหนดหน้าที่คำเสียก่อน

4.5 งานวิจัย Automatic Thai Unknown Word Recogniton [11]

งานวิจัยเน้นที่การแก้ปัญหาคำกำกวม และปัญหาการสะกดผิดโดยการกำหนดหน้าที่คำ (Part of Speech Tagging Ambiguity) ในงานวิจัยได้นำเรื่องสถิติเข้ามาใช้แก้ปัญหาการตัดและกำหนดหน้าที่คำโดยใช้โมเดลไตรแกรม (Tri-gram) และทำการคำนวณความน่าจะเป็นของประโยคดังสมการ ที่ 1

$$P(w) = \prod_{i=1}^{n} P(Wi, n)$$

$$= \prod P(Wi \mid W_{i-1}, W_{i-2}) \qquad (1)$$

โดยทำการหาความน่าจะเป็นของประโยค W และคำต่างๆ w_i เนื่องจากตามสมการจะต้องใช้คลังข้อความใหญ่มากทำให้ในความเป็นจริงอาจจะไม่สามารถหาคลังข้อความขนาดดังกล่าวได้

บทที่ 3
การตัดคำด้วยกฎที่ปรับปรุงและพจนานุกรมแบบใหม่

วิธีการในการตัดคำในเอกสารภาษาไทยมีขั้นตอนดังนี้คือ การตัดอนุประโยคโดยอาศัยช่องว่างและอักขระพิเศษ การตัดคำโดยอาศัยกฎการผสมอักษรในภาษาไทย การแบ่งประเภทอนุประโยค การวิเคราะห์คำที่มีอยู่ในพจนานุกรมและคำที่ไม่มีอยู่ในพจนานุกรมที่มีความเป็นไปได้ที่จะเป็นภาษาต่างประเทศหรือคำเฉพาะ

1. การตัดอนุประโยคโดยอาศัยช่องว่าง และ อักขระพิเศษ

วิธีการนี้จะพิจารณาส่วนที่ติดกันของตัวอักษรในภาษาไทย หากมีช่องว่างระหว่างคำหรือปรากฏตัวอักษรอื่นที่ไม่ใช่อักษรไทยให้ถือว่าเป็นคนละข้อความซึ่งไม่มีความเกี่ยวข้องกันในระดับคำซึ่งเอกสารหนึ่งๆ (D) จะประกอบด้วยหลายประโยคหรืออนุประโยค (S_i) โดยที่ N คือจำนวนของอนุประโยคในเอกสาร

$$D = S_1 + S_2 + S_3 + \ldots + S_N$$

2. การตัดคำโดยอาศัยกฎการผสมอักษรในภาษาไทย

สามารถแบ่งประเภทของพยัญชนะไทย 44 ตัว ออกเป็น 3 หมู่เรียกว่าไตรยางค์ ได้แก่ อักษรสูง อักษรกลาง และอักษรต่ำ และเมื่อพิจารณาถึงการนำไปผสมเป็นระดับพยางค์หรือคำแบ่งเป็น 3 ส่วน 4 ส่วน และ 5 ส่วนดังนี้

3 ส่วน ได้แก่ พยัญชนะ+สระ+วรรณยุกต์ เช่น ตา ดี ไป นา

4 ส่วน ได้แก่ พยัญชนะ+สระ+วรรณยุกต์ +ตัวสะกด เช่น คน

5 ส่วน ได้แก่ พยัญชนะ+สระ+วรรณยุกต์ +ตัวสะกด+ตัวการันต์ เช่น แพทย์ สิทธิ์ ฤทธิ์

โดยกฎการแบ่งส่วนของพยางค์จะได้ว่าสามารถตัดประโยคหรืออนุประโยคที่ประกอบด้วยอักษรน้อยกว่า 4 ตัวอักษรให้เป็น 1 คำหรือพยางค์ได้ทันทีโดยไม่ต้องเปรียบเทียบกับกฎหรือพจนานุกรม เนื่องจากพยางค์ที่สั้นที่สุดคือ 3 ส่วน หากในกรณีที่วรรณยุกต์อยู่ในรูปสามัญจะประกอบด้วยอักษร 2 ตัว (กรณีนี้ไม่รวมถึง ธ ณ ฤ ที่จะมีช่องว่างอยู่หน้าและหลังเสมอ) นั่นหมายถึงหากจะเป็น 2 คำหรือ 2 พยางค์ขึ้นไปต้องประกอบไปด้วย 4 ตัวอักษรขึ้นไป

$$S_i = C_{i1} + C_{i2} + C_{i3} + \ldots + C_{iL}$$

เมื่อ L คือจำนวนตัวอักษรทั้งหมดของอนุประโยค S_i

C คือตัวอักษรในประโยคหรืออนุประโยค S

$S_i = W_{i1} + W_{i2} + W_{i3} + \ldots + W_{iN}$

เมื่อ W_{iN} คือคำที่ประกอบขึ้นเป็นอนุประโยค S_i และ $N \leq L$

นอกจากคำไทยแท้แล้ว ภาษาไทยได้มีการรับภาษาต่างประเทศเข้ามาใช้ เช่น บาลี สันสกฤต อังกฤษ เป็นต้น ทำให้เกิดคำที่ไม่ตรงกับหลักการผสมคำอยู่มาก เช่น พรหม การ์ด มาร์ค เลาน์จ ซึ่งในปัจจุบันนี้จะพบคำที่มาจากภาษาต่างประเทศมากขึ้นและมีการถอดความเป็นภาษาไทยไม่ตรงกับหลักไวยากรณ์ไทย

3. วิธีการตัดคำที่นำเสนอ

การตัดคำเริ่มจากเอกสารนำเข้า แยกออกเป็นอนุประโยคย่อย S_i โดยใช้ช่องว่างเป็นตัวแบ่งและพิจารณาว่ามีอนุประโยคใดบ้างสามารถเป็นคำได้ทันที โดยให้

$S_i = C_{i1} + C_{i2} + C_{i3} + \ldots + C_{iL}$

เมื่อ L คือจำนวนตัวอักษรทั้งหมดของอนุประโยค S_i

C คือตัวอักษรในประโยคหรืออนุประโยค S

$S_i = W_{i1} + W_{i2} + W_{i3} + \ldots + W_{iN}$

เมื่อ W_{iN} คือคำที่ประกอบขึ้นเป็นอนุประโยค S_i และ $N \leq L$

หาก $L < 4$ จะได้ว่า

$W_{i1} = C_{i1} + C_{i2} + C_{i3} + \ldots + C_{iL}$

นั่นคือ $W_{i1} = S_i$

ยกตัวอย่างเช่น ฯลฯ , ฯพณฯ , ธ , ฤ , กิน , ถุง , ค้า เป็นต้น

หลังจากนั้นอนุประโยคอื่นที่เหลือจะนำไปวิเคราะห์ต่อไป การใช้วิธีการนี้ทำให้ลดเวลาในการตัดคำและค้นหาคำศัพท์ในพจนานุกรมเนื่องจากสามารถสรุปว่าเป็นคำได้ทันที

3.1 การแบ่งประเภทของอนุประโยค

นำอนุประโยคมาทำการตัดคำขั้นแรกโดยแยกประโยคเป็น 3 ประเภทด้วยกันตามส่วนประกอบของอนุประโยค

<u>ประเภทที่ 1</u> ประกอบด้วยอักษรไทย หรืออักษรไทยซึ่งอยู่ติดกับอักษรแบ่งวรรคอื่นๆ เช่น (,) , “ , ‘ เป็นต้น ยกเว้น - , / เนื่องจากหากเขียนติดกับอักษรไทยแล้วจะถือเป็นคำ หรือรหัส เช่น 3-ก ก/2 เป็นต้น

การตัดคำประโยคประเภทนี้จะทำการแยกอักษรไทยกับอักษรแบ่งวรรคออกจากกัน เช่น

“ระบอบการปกครอง:2475” จะแยกได้ว่า

“ ระบอบการปกครอง : 2475 ” เนื่องจาก 2475 ไม่ได้เขียนอยู่ติดกับอักษรไทยโดยตรงอนุประโยคนี้จึงถือเป็นประเภทที่ 1

<u>ประเภทที่ 2</u> ประกอบด้วยอักษรไทยซึ่งอยู่ติดกับตัวเลข หรืออักษรแบ่งวรรค - , / หรือ อักษรต่างประเทศ หรืออักษรพิเศษอื่นๆ จะทำการแยกอักษรแบ่งวรรคออกจากกัน ยกเว้นตัวเลข เครื่องหมาย – และ / จะยังคงเขียนติดกับอักษรไทยไว้เช่นนั้น และถือเป็นคำ 1 คำทันที เช่น “ก-2547” และ “23/2ก” จะแยกได้ว่า

“ ก-2547 ” และ “ 23/2ก ”

<u>ประเภทที่ 3</u> อักษรต่างประเทศหรืออักษรแบ่งวรรคยกเว้นเครื่องหมาย - และ / ประเภทที่ 3 นี้ไม่มีอักษรไทยอยู่ในประโยคเลย อนุประโยคประเภทนี้จะทำการแยกอักษรต่างประเทศและอักษรแบ่งวรรคออกจากกัน และนำอนุประโยคที่ได้มาทำการแปลงเป็นคำอ่านภาษาไทยและเก็บไว้สำหรับตัดคำในเอกสารที่ตรงกันหรือมีความใกล้เคียง

เช่น (wonderful) จะได้เป็นคำอ่านเป็นภาษาไทยดังนี้

won – {วอน , วัน , วอน์ , โวน์ , โวน}

der – {เดอ , เดอร์ , เดร์}

ful – { ฟุล , ฟูล , ฟุล์ , ฟูล์ , ฟัล}

จากนั้นจะทำการเก็บคำอ่านคำนี้ไว้เพื่อใช้เปรียบเทียบกับคำในเอกสารที่มีความเป็นไปได้ที่จะตรงกับคำนี้

3.2 การวิเคราะห์หาคำที่มีอยู่ในพจนานุกรมและหาคำที่ใช้อักษรไทยสะกดคำอ่านภาษาต่างประเทศ

คำไทยประเภทที่ 2 และ 3 ที่ผ่านขั้นตอน 4.1 จะนำมาตัดคำโดยอิงกับพจนานุกรม และคำต่างประเทศที่พบในเอกสาร จากนั้นจะได้เอกสารมา 2 ส่วน

ส่วนแรก พบคำตามพจนานุกรม หากพบคำในพจนานุกรมมากกว่าสองครั้งจะถือเอาคำที่ยาวที่สุดเป็นหลัก (Longest matching) คำที่เก็บอยู่ในพจนานุกรมนี้เป็นคำที่พบได้ทั่วไปหรือคำที่มีความยาวหลายพยางค์หรือคำที่มีการสะกดตรงตามอักขรวิธี

ส่วนที่สอง ไม่พบตามพจนานุกรม โดยปรกติส่วนนี้หากเป็นคำเดี่ยวๆ จะนำไปพิจารณากับคำรอบข้างซึ่งมีความเป็นไปได้ที่จะเป็นคำเดียวกันโดยอาศัยกฎการวิเคราะห์ความเป็นไปได้ที่จะเป็นภาษาต่างประเทศหรือคำเฉพาะ คำเฉพาะบางส่วนจะถูกเก็บอยู่ในพจนานุกรมคำเฉพาะและสามารถปรับปรุงได้เพื่อความเหมาะสมกับเอกสารแต่ละประเภท คำเฉพาะที่เหมาะสมได้แก่คำที่มาจากบาลี สันสกฤต ที่นำตัวอักษรเหล่านี้มาใช้ ฆ , ญ , ฐ , ฑ , ฒ , ผ , ภ , ฌ , ศ , ษ ซึ่งมักไม่ค่อยพบอยู่กับคำที่มาจากภาษาต่างประเทศ เป็นต้น เช่นชื่อคน ชื่อสถานที่

4. การสร้างพจนานุกรมแบบใหม่

คำที่ไม่พบในพจนานุกรมทั่วไปมักเป็นคำเฉพาะเช่นชื่อคนหรือสถานที่ คำใหม่ และคำที่มาจากภาษาต่างประเทศ ในที่นี้หากมีภาษาต่างประเทศปนอยู่ในเอกสารจะทำการแปลงเป็นคำสะกดภาษาไทยเพื่อเปรียบเทียบกับภาษาไทยที่อยู่ใกล้เคียงกับคำนั้น ในที่นี้จะเน้นไปที่ภาษาอังกฤษเท่านั้น

ตัวอย่างเช่น

"ซัมไคน์ออฟวันเดอฟูล (Some kind of wonderful)" หากตัดคำตามกฎและพจนานุกรม [1]

จะได้ว่า

ซัม-ไคน์-ออ-ฟ-วัน-เดอ-ฟู-ล เนื่องจากคำว่า ออ วัน และฟู ปรากฏอยู่ในพจนานุกรมทำให้การตัดคำไม่ถูกต้อง หากตัดคำโดยการใช้การแปลงคำจากภาษาอังกฤษเป็นคำอ่านภาษาไทยจะได้ว่า

ซัม-ไคน์-ออฟ-วันเดอฟูล ซึ่งทำให้ได้คำที่ถูกต้อง

พจนานุกรมที่เก็บคำศัพท์นั้นจะเก็บคำที่ซ้ำๆ กันเอาไว้ เมื่อคำหนึ่งคำ(S) ประกอบด้วยคำย่อย (S_i) ซึ่งคำย่อยก็เป็นคำในพจนานุกรม จะได้ว่า

$S_i = S_{i1} + S_{i2} + S_{i3} + \ldots + S_{iN}$

ยกตัวอย่างเช่น ระบอบประชาธิปไตย จะจัดเก็บดังนี้

$W_i = A_1 + A_2 + A_3 + \ldots + A_N$ เมื่อ $A \in \{ W , C\}$

W คือคำที่ตัดได้จากประโยคหรืออนุประโยค S

A คือคำย่อยที่มีอยู่ในพจนานุกรมที่ประกอบเป็นคำ W หรือ ตัวอักษรที่ไม่มีอยู่ในพจนานุกรม

C คือตัวอักษรในประโยคหรืออนุประโยค S

เช่น ระบอบ คือ [code1] , ประชา คือ [code2]

code1 คือรหัสแทนคำว่า "ระบอบ"

code2 คือรหัสแทนคำว่า "ประชา"

ประชาธิปไตย คือ [code2]+ธิปไตย

ระบอบประชาธิปไตย คือ [code1]+[code2]+ธิปไตย การจัดเก็บเป็นรหัสแทนเพื่อลดขนาดของพจนานุกรมและเพิ่มประสิทธิภาพสำหรับการตัดคำ

พจนานุกรมแบบใหม่มีข้อดีคือมีการจัดเก็บเฉพาะหน่วยคำที่เล็กที่สุดไว้ในพจนานุกรม กรณีที่คำหลายพยางค์ที่ประกอบด้วยคำย่อยหลายคำ คำย่อยเหล่านั้นจะถูกเก็บเป็นรหัสแทนทำให้มีการจัดเก็บคำนั้นเพียงครั้งเดียวและลดขนาดของพจนานุกรม แต่ข้อเสียคือมีความยุ่งยากในการจัดเก็บ

4.1 ตัวอย่างคำศัพท์ในพจนานุกรมแบบใหม่

คำ	การจัดเก็บ
กระ	*[กระ]*
กระจายเสียง	[กระ]จาย[เสียง]
กระวนกระวาย	[กระ][วน][กระ]วาย
กระเบียดกระเสียร	[กระ][เบียด][กระ]เสียร
เบียด	*[เบียด]*
เบียดเสียดเยียดยัด	[เบียด][เสียด]เยียด[ยัด]
ยัดเยียด	[ยัด]เยียด
ตกกระ	[ตก][กระ]*

ส่วนที่อยู่ใน [] คือรหัสของคำที่มีอยู่แล้วในพจนานุกรม ส่วนเครื่องหมาย * หมายถึงมีคำที่ใช้คำนี้อยู่ในพจนานุกรมอีก เช่นคำว่า กระ มีการนำไปใช้ เช่น

กระ อาจมีคำต่อท้าย เป็นกระจายเสียง ได้

ตกกระ อาจมีคำต่อท้ายเช่น ตกกระป๋อง

4.2 การเพิ่มคำศัพท์ลงในพจนานุกรม

เมื่อมีคำศัพท์ที่ไม่มีอยู่ในพจนานุกรม สามารถเพิ่มคำลงไปได้ การเพิ่มคำใหม่ลงในพจนานุกรมทำดังนี้

4.2.1 คำศัพท์พื้นฐาน จะถูกเก็บลงในพจนานุกรมก่อนและทำการจัดเก็บลงในและทำการจำทำรหัสโดยรหัสจะขึ้นต้นด้วยชื่อของไฟล์ที่จัดเก็บอักษรตัวแรกของคำ

4.2.1 คำต่อเนื่อง จทำการจัดเก็บ

4.3 การแก้ไขและลบคำศัพท์ในพจนานุกรม

5. การวิเคราะห์คำที่ไม่มีอยู่ในพจนานุกรมโดยเทียบกฎความเป็นไปได้ที่จะเป็นภาษาต่างประเทศหรือคำเฉพาะ

การวิเคราะห์คำเฉพาะจะทำในขั้นตอนสุดท้ายกรณีที่คำที่นำมาไม่ตรงกับพจนานุกรมหรือไม่มีความเป็นไปได้ในการที่จะเป็นภาษาต่างประเทศ ซึ่งพจนานุกรมศัพท์เฉพาะนี้สามารถเพิ่มเติมได้ตลอดเวลาเพื่อให้เหมาะสมกับเอกสารที่ต้องการนำมาตัดคำ ตัวอย่างคำเฉพาะเช่น ชื่อคน ชื่อสถานที่ โดยมีการวิเคราะห์ดังนี้

5.1 คำที่มาจากภาษาต่างประเทศ

5.1.1 เป็นคำที่ไม่มีอยู่ในพจนานุกรม

5.1.2 เป็นคำที่ไม่มีอักษรไทยต่อไปนี้ ฃ ฅ ฆ ฏ ฐ ฑ ฒ ณ ฤ ฦ ศ ษ ฬ ๅ และเครื่องหมายเว้นวรรคและเลขไทยต่างๆ

5.2 คำเฉพาะ เช่น ชื่อคน ชื่อสถานที่

5.2.1 เป็นคำที่ไม่มีอยู่ในพจนานุกรม

5.2.2 มักใช้อักษรดังต่อไปนี้ในการสะกดคำ ฆ ฉ ฌ ญ ฏ ฐ ฑ ฒ ณ ธ ภ ฤ ศ ษ ฬ ฮ เป็นต้น

นอกจากนี้คำที่ไม่ปรากฏในพจนานุกรมที่ถูกวิเคราะห์ให้เป็นคำที่มาจากภาษาต่างประเทศหรือคำเฉพาะจะถูกนำไปวิเคราะห์ต่อเนื่องไปอีกในกรณีที่คำข้างเคียงอาจเป็นคำเดียวกับคำที่มาจากภาษาต่างประเทศหรือคำเฉพาะนั้น

5.3 พิจารณาคำข้างเคียง

5.3.1 เป็นคำข้างเคียงคำที่วิเคราะห์ให้เป็นคำที่มาจากภาษาต่างประเทศหรือคำเฉพาะ

5.3.2 เป็นคำที่อาจปรากฏอยู่ในพจนานุกรมหรือไม่ก็ได้

5.3.3 ในกรณีเป็นคำข้างเคียงคำภาษาต่างประเทศ ต้องมีลักษณะตรงตาม 5.1.2

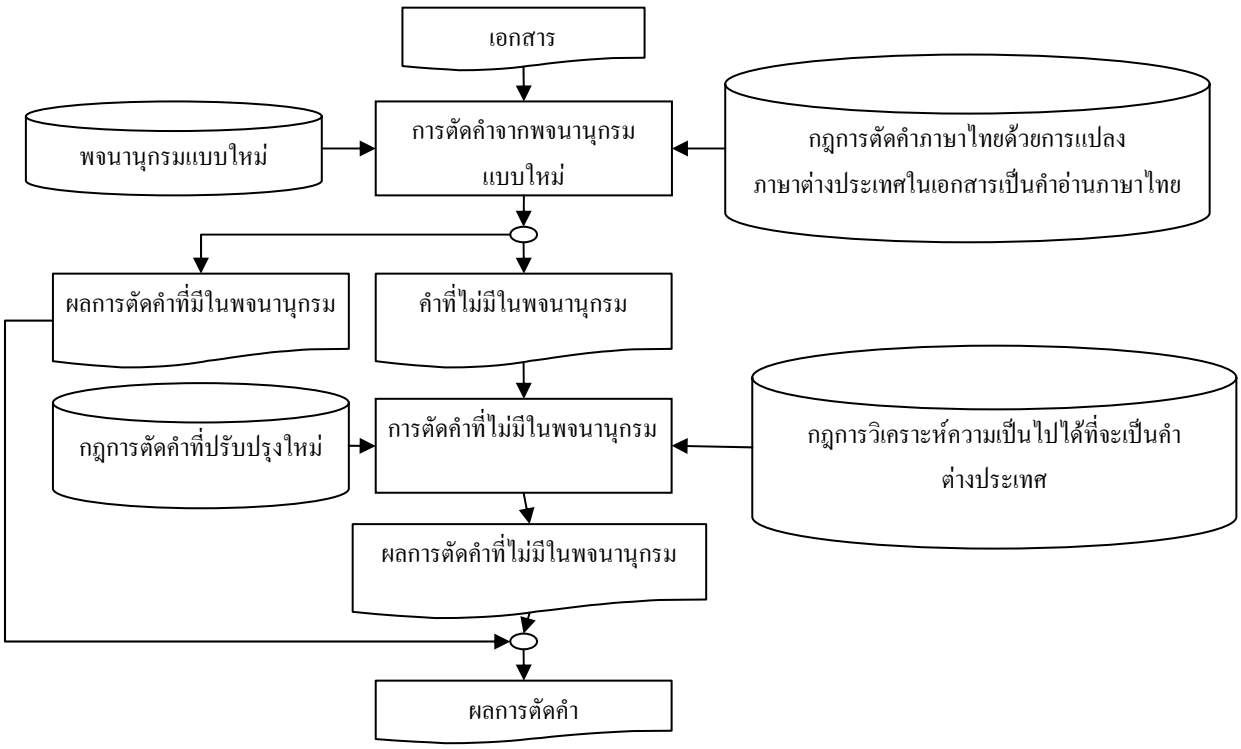
5.3.4 มักจะพบคำข้างเคียงปรากฏคู่กันกับคำที่มาจากภาษาต่างประเทศหรือคำเฉพาะในเอกสาร ตัวอย่างเช่น “แอนนา”

ในกรณีคำว่า “แอน” ถูกวิเคราะห์ให้เป็นคำที่มาจากภาษาต่างประเทศตาม 5.1 และคำว่า “นา” ซึ่งมีความหมายตามพจนานุกรม หากพบคำว่า “นา” ทุกครั้งที่มีการปรากฏคำว่า “แอน” ให้วิเคราะห์ว่าคำว่า “แอนนา” เป็นคำที่มาจากภาษาต่างประเทศตามลักษณะของคำว่า “แอน” ซึ่งถูกวิเคราะห์ให้เป็นคำที่มาจากภาษาต่างประเทศอยู่ก่อน การวิเคราะห์นี้ได้แสดงไว้ในภาพที่ 4

6. การตัดคำในส่วนสุดท้ายของอนุประโยค

ในส่วนสุดท้าย อนุประโยคหรือส่วนของอนุประโยคใดไม่ตรงกับข้อที่กล่าวมาข้างต้นจะทำการตัดคำโดยใช้กฎการตัดพยางค์แทน

จากข้อ 1-6 ที่กล่าวมาถูกแสดงไว้ในภาพรวมของขั้นตอนวิธีการตัดคำภาษาไทยด้วยวิธีปรับปรุงกฎและพจนานุกรมแบบใหม่ดังในภาพที่ 3



ภาพที่ 3 ภาพรวมของขั้นตอนวิธีการตัดคำภาษาไทยด้วยวิธีปรับปรุงกฎและพจนานุกรมแบบใหม่

7. ตัวอย่างวิธีการตัดคำด้วยกฎปรับปรุงและพจนานุกรมแบบใหม่

ตัวอย่างการตัดคำด้วยกฎปรับปรุงและพจนานุกรมแบบใหม่

“เจเป็กแอลเอ็ส (JPEG-LS – Joint Photographic Expert Group Lossless) เป็นมาตรฐานการบีบอัดภาพแบบไม่สูญเสีย สำหรับภาพเทาต่อเนื่อง ขนาดของจุดที่ 2–16 บิตต่อจุด ซึ่งกำหนดมาตรฐานโดยคณะกรรมการเจเป็ก (JPEG - Joint Photographic Expert Group) ในปี 2540 เจเป็กแอลเอ็ส (JPEG-LS) ใช้เทคนิคการบีบอัดภาพแบบการทำนายจุด (Predictive Coding) ร่วมกับการบีบอัดภาพเชิงสถิติ (Probabilistic Coding)”

การแบ่งอนุประโยค ในส่วนนี้จะทำการแบ่งอนุประโยคโดยอาศัยช่องว่าง อักขระพิเศษ และอักษรต่างประเทศ

<u>ประเภทที่ 1</u> ประกอบด้วยอักษรไทย และอักษรไทยที่อยู่ติดกับอักษรแบ่งวรรคอื่นๆ เช่น (,) , “ , ‘ เป็นต้น ยกเว้น - , /

<u>ประเภทที่ 2</u> ประกอบด้วยอักษรไทยซึ่งอยู่ติดกับตัวเลข หรืออักษรแบ่งวรรค - , / หรือ อักษรต่างประเทศ หรืออักษรพิเศษอื่นๆ

ประเภทที่ 3 อักษรต่างประเทศหรืออักษรแบ่งวรรคยกเว้นเครื่องหมาย - และ / ประเภทที่ 3 นี้ไม่มีอักษรไทยอยู่ในประโยคเลย

จากตัวอย่างจะได้ว่า

อนุประโยคที่ได้	ประเภทอนุประโยคที่ได้
เจเป็กแอลเอ็ส	ประเภทที่ 1
(	ตัดคำเรียบร้อย
JPEG-LS	ประเภทที่ 3
–	ตัดคำเรียบร้อย
Joint	ประเภทที่ 3
Photographic	ประเภทที่ 3
Expert	ประเภทที่ 3
Group	ประเภทที่ 3
Lossless	ประเภทที่ 3
)	ตัดคำเรียบร้อย
เป็นมาตรฐานการบีบอัดภาพแบบไม่สูญเสีย	ประเภทที่ 1
สำหรับภาพเทาต่อเนื่อง	ประเภทที่ 1
ขนาดของจุดที่	ประเภทที่ 1
2–16	ประเภทที่ 2
บิตต่อจุด	ประเภทที่ 1
ซึ่งกำหนดมาตรฐานโดย คณะกรรมการเจเป็ก	ประเภทที่ 1
(	ตัดคำเรียบร้อย
JPEG	ประเภทที่ 3
-	ตัดคำเรียบร้อย
Joint	ประเภทที่ 3
Photographic	ประเภทที่ 3
Expert	ประเภทที่ 3
Group	ประเภทที่ 3
ในปี	ประเภทที่ 1
2540	ประเภทที่ 2

อนุประโยคที่ได้	ประเภทอนุประโยคที่ได้
เจเป็กแอลเอ็ส	ประเภทที่ 1
(	ตัดคำเรียบร้อย
JPEG-LS	ประเภทที่ 3
)	ตัดคำเรียบร้อย
ใช้เทคนิคการบีบอัดภาพแบบการทำนายจุด	ประเภทที่ 1
(	ตัดคำเรียบร้อย
Predictive	ประเภทที่ 3
Coding	ประเภทที่ 3
)	ตัดคำเรียบร้อย
ร่วมกับการบีบอัดภาพเชิงสถิติ	ประเภทที่ 1
(	ตัดคำเรียบร้อย
Probabilistic	ประเภทที่ 3
Coding	ประเภทที่ 3
)	ตัดคำเรียบร้อย

ทำการแปลงอนุประโยคประเภทที่ 3 ให้เป็นคำอ่านในภาษาไทย ดังตัวอย่าง

อนุประโยคประเภทที่ 3 ที่ได้มา	คำอ่านภาษาไทยได้
JPEG-LS	{จ, เจ} {เพก์, เปก์, เพก, เปก, เพ็ก, เป็ก} - {ลส, แอลเอส, แอลเอ็ส}

ทำการเก็บผลลัพธ์จากการแปลงไว้สำหรับเปรียบเทียบคำที่พบในเอกสาร

ทำการตัดคำอนุประโยคประเภทที่ 1 ยกตัวอย่าง “เจเป็กแอลเอส”

ทำการตัดคำโดยเทียบกับพจนานุกรม

เจ เป็ก แอล เอ็ส แต่เนื่องจากพบคำว่า “เจเป็ก” และ “แอลเอ็ส” จากอนุประโยคที่ 3 ที่ทำไว้ก่อนหน้านี้ จึงได้ว่า “เจเป็ก แอลเอส”

คำที่ไม่พบตามพจนานุกรมได้แก่ “บิตต่อจุด” โดยคำว่า “บิต” ไม่พบตามพจนานุกรม จึงใช้กฎในการตัดคำ

ผลลัพธ์การตัดคำเป็นดังนี้

เจเป็ก แอลเอ็ส (JPEG-LS – Joint Photographic Expert Group Lossless) เป็น มาตรฐาน การบีบ อัด ภาพ แบบ ไม่ สูญเสีย สำหรับ ภาพ เทา ต่อเนื่อง ขนาด ของ จุด ที่ 2–16 บิต ต่อ จุด ซึ่งกำหนด มาตรฐานโดย คณะกรรมการ เจเป็ก (JPEG -Joint Photographic Expert Group) ใน ปี 2540 เจเป็ก แอล เอ็ส (JPEG-LS) ใช้ เทคนิค การ บีบ อัด ภาพ แบบ การ ทำนาย จุด (Predictive Coding) ร่วม กับ การ บีบ อัด ภาพ เชิง สถิติ (Probabilistic Coding)

บทที่ 4
โครงสร้างของระบบตัดคำภาษาไทยด้วยกฎที่ปรับปรุงและพจนานุกรมแบบใหม่

ในระบบการตัดคำภาษาไทยด้วยกฎที่ปรับปรุงและพจนานุกรมแบบใหม่มีองค์ประกอบหลายส่วนด้วยกัน ดังนี้

1. ส่วนของกฎ

การพัฒนาระบบการตัดคำภาษาไทยด้วยกฎที่ปรับปรุงและพจนานุกรมแบบใหม่นี้ได้รับการพัฒนาบนระบบวินโดวส์เอ็กซ์พี ประกอบด้วยส่วนการทำงาน ส่วน ดังนี้

1. การตัดพยางค์ซึ่งใช้กฎที่ปรับปรุงจาก [1] จะได้ว่า

1.1 กฎที่ใช้ในการตัดพยางค์มีอยู่ทั้งหมดสัญลักษณ์ดังนี้

c (Consonant) : พยัญชนะปรกติ

v (Vowel) : สระ

t (Tonal Mark) : วรรณยุกต์

s (Speller) : ตัวสะกด

[…]? : ทางเลือก อาจมีหรือไม่ก็ได้

[a1 a2 a3 … an] : ทางเลือก

1.2 กฎ สำหรับตัดคำภาษาไทยที่ปรับปรุง

1. [c][t]?[ะ ำ า] เช่น ระ ม้า คำ
2. [c][ิ ี ึ ุ ู][t]? เช่น ยี รู้
3. [c][ึ][t]?[s] เช่น ลืม ลื่น
4. [c][t]?[ุ ู] เช่น รู้ ผุ
5. [c] ั [t]?[s] เช่น ค้น คน
6. [เ แ โ ใ ไ][c][t]? เช่น แท้ ใย
7. [เ แ] [c] ็ [s] เช่น เจ็บ เยน
8. [เ แ โ][c][t]? ะ เช่น แกะ เก๊ะ
9. [เ แ โ][ก ข ค ต ท บ ป พ ฟ จ ซ ศ ส] ร[t]? ะ เช่น แครระ
10. [เ แ โ][ก ข ค บ ป ผ พ]ล[t]?ะ เช่น แกละ เผละ
11. เ[ก ข ค ต ท บ ป พ ฟ จ ซ ศ ส]ร[า ิ ี ึ ื ุ ู][t]? เช่น เครา เครื่

12. เ[ก ข ค บ ป พ ฟ]ล[า ิ ี ึ ื ุ ู][t]? เช่น เคลื่
13. [เ แ โ][ก ข ค]ว[t]?ะ เช่น เก๊ะ แคะ
14. เ[c] ี [t]?ย เช่น เสีย เรี่ย
15. เ[c][t]?าะ เช่น เผาะ เก้าะ
16. เ[c][t]?[าะ] เช่น เตะ เว้า
17. เ[c][t]? เช่น เร เร่
18. [โ ใ ไ]ห[ง ญ น ม ย ร ล ว][t]? เช่น ใหล่ โหล
19. เ[c] ื [t]?อ เช่น เผื่อ เสือ
20. [c] ื [t]?อ เช่น มือ มื้อ
21. [c] ์ เช่น ก์ ส์
22. [๘ ๙ ๗ + ะ ั า ำ ิ ี ึ ื ุ ู] เช่น ะ ื
23. ฯ
24. เ[จ ส]ริญ ได้แก่ เจริญ เสริญ
25. ห ร[t]?[ะ า ำ ุ ู] เช่น หรู่ หระ
26. ห ร[ิ ี ึ ื][t]? เช่น หรี่ หริ
27. เ[ก ข ค บ ป ผ พ ฟ]ลื[t]?อ เช่น เกลือ เกลื้อ
28. เ[ก ข ค ต ท บ ป พ ฟ]รื[t]?อ เช่น เครือ เครื่อ
29. เ[ก ข ค บ ป ผ พ ฟ]ลิ[t]?[s] เช่น เพลิน เคลิ้ม
30. เ[ก ข ค ต ท บ ป พ ฟ]ริ[t]?[s] เช่น เกริ่น เกริก
31. แ[ก ข ค ต ท บ ป พ ฟ]ร[t]?[ก ง ว น บ ม ท ต ด] เช่น แทรก แคร้ง
32. แ[ก ข ค บ ป ผ พ ฟ]ล[t]?[ก ง ว น บ ม ท ต ด] เช่น แปลน แกล้ง
33. เห[ง ญ น ม ย ร ล ว][t]?า เช่น เหล้า เหลา
34. [A-Z a-z เครื่องหมายพิเศษต่างๆ]* เช่น A1a thisis
35. ฯลฯ
36. ๆ
37. เ[ก ข ค]วี[t]?ย เช่น เกวีย
38. เ[ก ข ค ต ท บ ป พ ฟ]รี[t]?ย เช่น เตรีย
39. เ[ก ข ค ง บ ผ พ ฟ]ลี[t]?ย เช่น เกลี่ย
40. ช่องว่าง
41. [0-9]* เช่น 015789645 1234

42. “\n”
43. css สำหรับพยัญชนะใดๆ
44. า์c [s] สำหรับ า ใดๆ
45. า์c [ss] สำหรับ า ใดๆ
46. ส่วนของคำที่มีความยาวน้อยกว่า 4 ตัวอักษร
47. ทุกตัวอักษรที่ไม่อยู่ในกฎ

2. ส่วนของคำ

การตัดคำจะทำการตัดคำโดยค้นหาจากพจนานุกรมโดยที่พจนานุกรมจะอยู่ในรูปของฐานตัวอักษร (Text-based)

3. ส่วนของคำที่ไม่อยู่ในพจนานุกรม

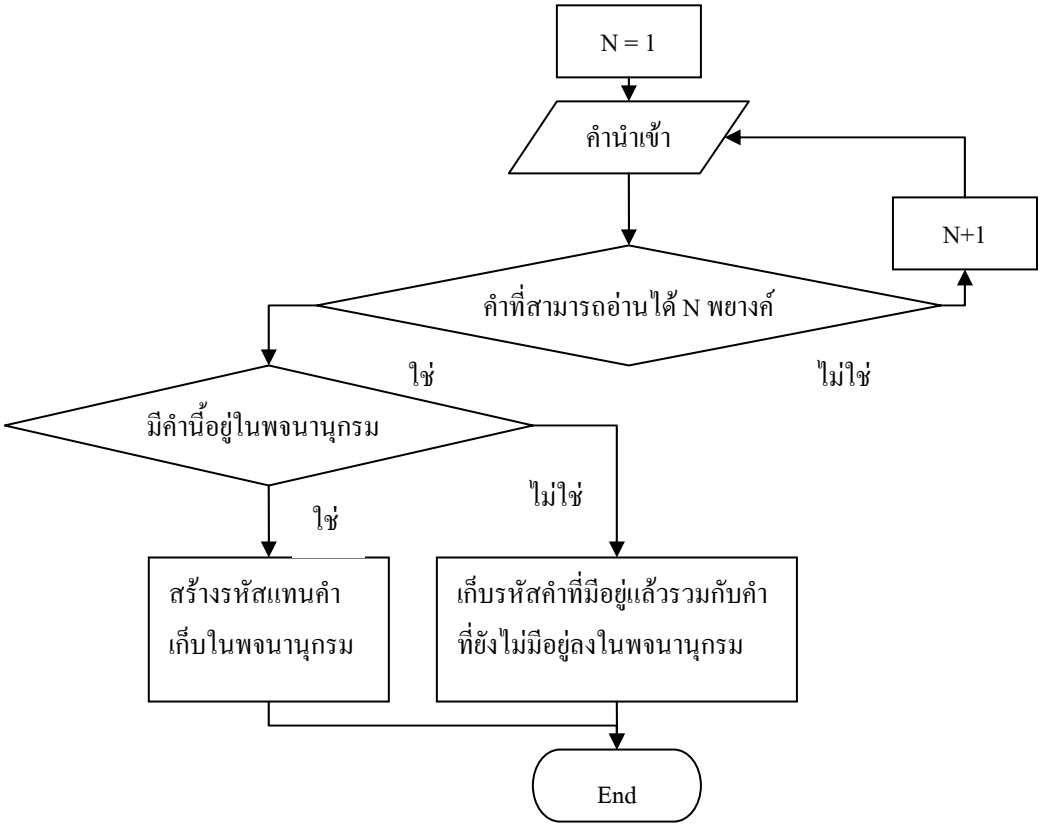
3.1 กฎของคำที่อาจเป็นภาษาต่างประเทศหรือคำเฉพาะ

3.2 กฎของการรวมคำข้างเคียงเข้าด้วยกัน

4. ส่วนของพจนานุกรม

4.1 ส่วนของการเพิ่มคำลงในพจนานุกรมแบบใหม่

การเพิ่มคำลงในพจนานุกรมจะทำการใช้คำที่มีอยู่ในพจนานุกรมแบบเดิม จากนั้นทำการเพิ่มคำโดยแยกส่วนที่เป็นคำมูล หรือคำพยางค์ที่มีความหมายก่อน จากนั้นจึงทำการแยกเก็บคำลงในพจนานุกรมดังภาพที่ 4



ภาพที่ 4 การเก็บคำศัพท์เข้าไว้ในพจนานุกรม

5. การออกแบบโปรแกรม

5.1 ส่วนของพจนานุกรม

การจัดเก็บคำศัพท์จะถูกจัดลงในเท็กซ์ไฟล์ โดยการจัดทำดรรชนีของคำศัพท์ไว้ดังนี้

5.2 ส่วนของการตัดคำ

การตัดคำจำทำการ

บทที่ 5
ผลการตัดคำ

บทนี้จะมีการนำเสนอเกี่ยวกับการทดลองและผลการทดลองการตัดคำภาษาไทยด้วยกฎปรับปรุงและพจนานุกรมแบบใหม่

1. การเตรียมการทดลอง

เอกสารที่นำมาทำการตัดคำภาษาไทย

เอกสารที่นำมาทำการตัดคำภาษาไทยนั้นไม่จำเป็นต้องเป็นเอกสารที่มีภาษาไทยเพียงอย่างเดียวโดยไม่มีอักษรต่างประเทศปนอยู่ เนื่องจากคำที่สะกดเลียนเสียงมาจากภาษาต่างประเทศบางคำไม่มีอยู่ในพจนานุกรม หรือมีอยู่ในพจนานุกรมเพียงบางส่วน ทำให้การตัดคำผิดไป ดังนั้นเอกสารที่นำมาใช้ทดลองทำการตัดคำนี้จึงเป็นเอกสารที่มีทั้งอักษรไทยและอักษรอังกฤษปะปนกันไปและเป็นเอกสารที่พบได้ทั่วไป

ประเภทของเอกสาร

งานวิจัยนี้แบ่งประเภทเอกสารที่นำมาทำการทดลองตัดคำภาษาไทยออกเป็น 5 กลุ่มซึ่งมีความหลากหลายด้านเนื้อหาซึ่งเอกสารแต่ละประเภทจะประกอบไปด้วยทั้งภาษาไทยและภาษาต่างประเทศ คือ

ข่าวเศรษฐกิจ

ข่าวต่างประเทศและกีฬา

ข่าวอื่นๆ

บทความวิชาการ

วารสารทั่วไป

ขนาดของเอกสารที่นำมาทำการตัดคำ

การทดลองทำการตัดเอกสารตั้งแต่ 1 MB , 2 MB , 3 MB , … และ 7 MB ซึ่งเป็นขนาดที่มากพอสำหรับพิจารณาความถูกต้องของการตัดคำ เพื่อทำการวัดความถูกต้องและเวลาที่ใช้ในการตัดคำภาษาไทยเมื่อข้อมูลมีขนาดเพิ่มขึ้น

2. การทดลอง

การเตรียมเอกสารนำเข้า

เอกสารที่อยู่ในรูปของไฟล์จะถูกจัดเตรียมในลักษณะของแฟ้มข้อมูลตัวอักษร โดยอยู่ในรูปแบบ .txt หรือ .dat

เอกสารที่คัดลอกมาสามารถนำมาวางในส่วนของการนำข้อมูลเข้าเพื่อทำการตัดคำได้ทันที

3. การวัดผลการตัดคำภาษไทย

การวัดผลการตัดคำภาษาไทยแบ่งออกเป็น 2 ชนิดด้วยกัน คือ

ความถูกต้องในเชิงของคำ

ความถูกต้องจะนับคำที่ตัดออกมาได้ถูกต้องตามความหมายต่อจำนวนคำที่ตัดออกมาได้ทั้งหมด

ความถูกต้องในเชิงของพยางค์

ความถูกต้องในเชิงพยางค์จะนับคำที่อ่านออกเสียงถูกต้องตามหลักของการอ่านคำนั้นๆ หรือสะกดตรงตามลักษณะการแบ่งพยางค์ของคำนั้นๆ

4. ผลการทดลอง

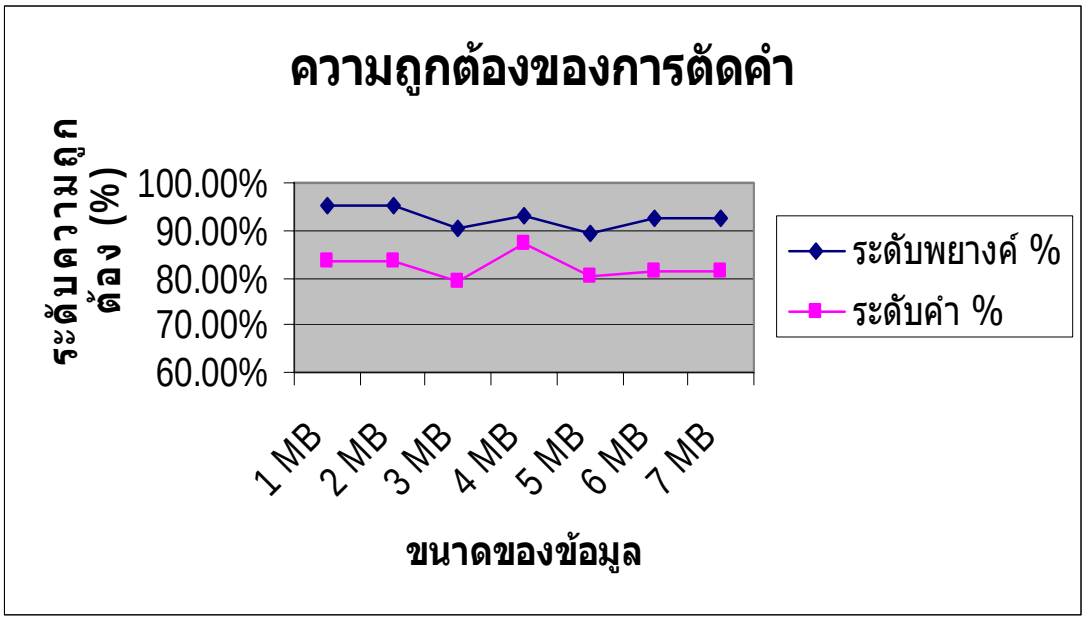
การทดลองการตัดคำภาษาไทยในระดับพยางค์และระดับคำโดยแยกทำการตัดคำกับเอกสารกลุ่มต่างๆ 5 กลุ่ม โดยเปรียบเทียบกับการตัดคำแบบเดิม [1] ได้ผลดังตารางที่ 3

ตารางที่ 3 แสดงผลการตัดคำด้วยการใช้กฎที่ปรับปรุงและพจนานุกรมแบบใหม่ร่วมกัน

เอกสาร	การตัดคำแบบเดิม		การตัดคำที่นำเสนอ		ความถูกต้องที่เพิ่มขึ้น	
	ระดับพยางค์	ระดับคำ	ระดับพยางค์	ระดับคำ	ระดับพยางค์	ระดับคำ
ข่าวเศรษฐกิจ	95.34%	83.50%	94.67%	85.04%	-0.67%	1.54%
ข่าวต่างประเทศและกีฬา	90.14%	79.25%	92.00%	81.07%	1.86%	1.82%
ข่าวอื่นๆ	93.25%	87.01%	97.28%	98.05%	4.03%	11.04%
บทความวิชาการ	89.40%	80.21%	88.25%	87.44%	-1.15%	7.23%
วารสารทั่วไป	92.67%	81.35%	93.52%	93.75%	0.85%	12.40%

จากตารางที่ 3 การตัดคำด้วยการใช้กฎที่ปรับปรุงและพจนานุกรมแบบใหม่ร่วมกันมีการตัดคำในระดับพยางค์โดยเฉลี่ยเพิ่มขึ้น 0.98% และระดับคำเพิ่มขึ้น 6.81% แต่เมื่อพิจารณาตามแต่ละประเภทเอกสาร เอกสารที่ตัดคำได้ถูกต้องเพิ่มขึ้น 4.03% ซึ่งสูงที่สุดในระดับพยางค์ และ11.04% ในระดับคำซึ่งจะมีเนื้อหาทั่วไปไม่เจาะจง และคำที่ไม่พบในพจนานุกรมสามารถแก้ไขด้วยการใช้กฎปรับปรุง ส่วนเอกสารที่มีการตัด

คำได้ถูกต้องน้อยลงในระดับพยางค์ได้แก่ข่าวเศรษฐกิจและบทความทางวิชาการซึ่งจะมีคำที่มีความยาวหลายพยางค์ ทำให้การตัดคำด้วยกฎนั้นตัดได้ไม่ถูกต้องนัก



ภาพที่ 5 ความถูกต้องของการตัดคำระดับพยางค์และระดับคำโดยเปรียบเทียบที่ขนาดของข้อมูลแตกต่างกัน

ภาพที่ 6 เวลาที่ใช้ในการตัดคำภาษาไทยโดยเปรียบเทียบที่ขนาดของข้อมูลแตกต่างกัน

บทที่ 6

บทสรุปและข้อเสนอแนะ

ในบทนี้จะกล่าวถึงการสรุปผลและวิจารณ์ผลการวิจัย รวมถึงปัญหาต่างๆ และข้อเสนอแนะเพื่อประโยชน์สำหรับทำวิจัยต่อไปภายภาคหน้า

1. ประสิทธิภาพของการตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่

การตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่มุ่งเน้นเพื่อแก้ไขปัญหาการตัดคำด้วยกฎที่มีอยู่เดิมซึ่งไม่ครอบคลุมคำในภาษาไทยซึ่งมีคำที่มีลักษณะที่แตกต่างซับซ้อนออกไปจากเดิมโดยเฉพาะคำเฉพาะ และคำที่มีจากภาษาต่างประเทศโดยการเพิ่มกฎเพื่อความยืดหยุ่นมากขึ้น อีกทั้งเพิ่มกฎความเป็นไปได้ที่จะเป็นภาษาต่างประเทศหรือคำเฉพาะเพื่อเพิ่มความถูกต้องของการตัดคำโดยเฉพาะคำที่ไม่มีอยู่ในพจนานุกรม อีกทั้งการตัดคำโดยอาศัยคำภาษาอังกฤษที่ปรากฏในเอกสารสามารถลดความผิดพลาดจากการตัดคำคำอ่านภาษาไทยที่บางส่วนของคำหรือทั้งหมดของคำไม่ปรากฏ

2. ข้อเสนอแนะ

รายงานนี้ได้ปรับปรุงกฎเดิมเพื่อตัดคำที่มาจากภาษาต่างประเทศและคำที่ไม่พบในพจนานุกรม แต่คำเฉพาะบางคำที่มีการสะกดอ่านได้หลายพยางค์ยังไม่สามารถรวมให้เป็นคำเดียวได้นอกจากทำการบันทึกคำนั้นไว้ในพจนานุกรมซึ่งทำได้เพียงบางส่วนจากภาษาอังกฤษที่อยู่ในเอกสารนำมาแปลงให้เป็นคำอ่านในภาษาไทย ส่วนคำกำกวมสามารถแก้ปัญหาโดยการใช้วิธีตัดคำที่ยาวที่สุดร่วมกับวิธีการย้อนกลับแต่ยังไม่สามารถตัดคำได้ถูกต้องทุกครั้งเนื่องจากการเลือกคำที่ยาวที่สุดไม่ใช่กรณีที่ดีที่สุด ดังนั้นควรมีการเพิ่มเติมปรับปรุงสำหรับตัดคำประเภทนี้ด้วยการใช้ค่าสถิติของคำที่พบในเอกสารทั่วไปร่วมกับการใช้ความถี่ของคำที่พบในเอกสารที่นำมาตัดคำ

บรรณานุกรม

[1] ดวงแก้ว สวามิภักดิ์, *การสร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกซ์*: มหาวิทยาลัยธรรมศาสตร์,2533.

[2] บรรจบ พันธุเมธา, *ลักษณะภาษาไทย* กรุงเทพ: สำนักพิมพ์มหาวิทยาลัยรามคำแหง. 1-45. 2540.

[3] ปโยธร อุราธรรมกุล และ กานดา รุณนะพงศา, "การปรับปรุงการตัดคำในเอกสารไทย", NECSEC ครั้งที่ 1 , 2548, หน้า 41-45.

[4] ปโยธร อุราธรรมกุล และ กานดา รุณนะพงศา, "การตัดคำภาษาไทยด้วยวิธีปรับปรุงกฎและพจนานุกรมแบบใหม่", JCSSE 2006 , 2549, หน้า 34-40.

[5] ไพศาล เจริญพรสวัสดิ์, Feature-based thai word segmentation. จุฬาลงกรณ์มหาวิทยาลัย. 2542.

[6] พิสิทธิ์ พรมจันทร์, *การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรมแยกคำภาษาไทย,* จุฬาลงกรณ์มหาวิทยาลัย. 2540.

[7] พระยาอุปกิตศิลปสาร, *หลักภาษาไทย*, กรุงเทพฯ : โรงพิมพ์ไทยวัฒนาพานิช. 18-28. 2539.

[8] ยืน ภู่วรวรรณ และ วิวรรธ อิ่มอารมณ์, "การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี". รายงานการประชุมวิชาการวิศวกรรมไฟฟ้าครั้งที่ 9, 2529.

[9] สุรินทร์ จรรยาพรพงษ์. *A Thai Syllable Seperation Algorithm.* Asian Institute of Technology, 1983.

[10]หัชทัย ชาญเลขา, อัศนีย์ ก่อตระกูล. การสกัดนิพจน์ระบุนามในภาษาไทยโดยใช้แมกซิมัมเอนโทรปีโมเดลและอิงความรู้. หน่วยปฏิบัติการวิจัยเชี่ยวชาญเฉพาะการประมวลผลภาษาธรรมชาติและเทคโนโลยีสารสนเทศอัจฉริยะ. มหาวิทยาลัยเกษตรศาสตร์. 2547.

[11] Kawtrakul, A, Automatic Thai Unknown Word Recogniton. Proceedings of the Natural Language Processing Pacific Rim Symposium. 1997.

[12]B. Kijsirikul. "Comparing Winnow and RIPPER in Thai Named-Entity Identification", Chulalongkorn. 1997.

[13]C. Kooptiwoot. "Segmentation of Ambiguous Thai Words by Inductive Logic Programming". Chulalongkorn. 1999.

[14]D. D. Plamer. "A Trainable Rule-based Algorithm for Word Segmentation". 1995.

[15]P. Charoenpornsawat, B. Kijsirkul, "Feature-based Proper Name Identification in Thai", Chulalongkorn. 1998.

[16]P. Charoenpornsawat, B. Kijsirkul, S. Meknavin. "Feature-based Thai Unknown Word Boundary Identification Using Winnow", Chulalongkorn. 1998.

[17]P.Charoenpornsawat, V. Sornlertlamvanich. Automatic Sensentence Break Disambiguation for Thai. NECTEC. 2000.

[18]T. Pongthai, V. Sornlertlamvanish. “Grapheme to Phoneme for Thai”, NECTEC. 2003.

[19]T. Theeramunkong, V. Sornlertlamvanich, T. tanhermhong, W. Chinnan. "Character Cluster Based Thai Information Retrieval”, National Electronics and Computer Technology Center (NECTEC), National Science and Technology Development Agency (NSTDA). 1999.

[20]V. Sornlertlamvanich, T. Potipiti, T. Charoenporn. “Automatic Corpus-based Thai Word Extraction with the C4.5 Learning Algorithm”. National Electronics and Computer Technology Center (NECTEC).

[21]V. Tesprasit, P. Charoenpornsawat , V. Sornlertlamvanich. “Learning Pharse Break Detection in Thai Text-to-Speech”. EUROSPEECH, 2003.

[22]W. Arronmanakun. “Collocation and Thai Word Segmentation”. 1995.

ภาคผนวก ก

บทความของผู้ทำวิจัย

บทความของผู้วิจัย

1. ปโยธร อุราธรรมกุล และ กานดา รุณนะพงศา, “การปรับปรุงการตัดคำในเอกสารไทย”, NECSEC ครั้งที่ 1 , 2548, หน้า 41-45.

2. ปโยธร อุราธรรมกุล และ กานดา รุณนะพงศา, “การตัดคำภาษาไทยด้วยวิธีปรับปรุงกฎและพจนานุกรมแบบใหม่”, JCSSE 2006 , 2549, หน้า 34-40.

การปรับปรุงกฎสำหรับตัดคำในเอกสารไทย

Improved Rule-Based for Thai Documents

ปโยธร อุราธรรมกุล
นักศึกษาระดับบัณฑิตศึกษา
ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ หาวิทยาลัยขอนแก่น
Email:payothorn@gmail.com

กานดา รุณนะพงศา
ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น
Email: krunapon@kku.ac.th

บทคัดย่อ

การตัดคำไทย (Thai Word Segmentation) คือการแยกแต่ละคำในเอกสารไทยออกจากกันเพื่อนำไปใช้ประโยชน์ในด้านอื่นๆ เช่น การสังเคราะห์เสียงพูด การแปลภาษา เป็นต้น เอกสารที่มีอยู่ ณ ปัจจุบันไม่เพียงแต่จะมีคำไทยแท้เท่านั้น คำบางคำที่มาจากภาษาต่างประเทศที่ถูกสะกดอยู่ในรูปของคำอ่านภาษาไทยบางคำจะมีการผสมอักษรที่แตกต่างนอกเหนือออกไปจากกฎการตัดคำ (Rule-based) แบบเดิมที่มีอยู่ เนื่องจากคำเหล่านี้มีอยู่มากมายและเกิดใหม่อยู่เสมอ บทความนี้นำเสนอการปรับปรุงกฎการตัดคำให้มีความยืดหยุ่นมากขึ้นจะเป็นประโยชน์สำหรับการตัดคำที่ไม่รู้จักหรือไม่มีความหมายอยู่ตามพจนานุกรม จากผลการทดลองพบว่าการตัดคำระดับพยางค์เพิ่มขึ้นถึง 2.14 เปอร์เซ็นต์ และระดับคำเพิ่มขึ้นถึง1.76 เปอร์เซ็นต์จากเทคนิคเดิม

คำสำคัญ การตัดคำ, กฎการตัดคำ

1. บทนำ

ลักษณะของประโยคในภาษาไทยมีการเขียนติดกันไป ทำให้ยากต่อการนำไปใช้งานในบางด้าน เช่น การสังเคราะห์เสียงพูด การแปลภาษา เป็นต้น ได้มีผู้คิดค้นวิธีที่จะแยกคำแต่ละคำออกจากประโยคซึ่งมีการเขียนติดกันไปอย่างต่อเนื่องทั้งประโยค ในงานวิจัยนี้จะกล่าวถึงการตัดคำโดยอาศัยอักขรวิธีเป็นหลักการพื้นฐานการประสมคำซึ่งมีความแตกต่างไปจากภาษาอังกฤษ หรือภาษาจีน เนื่องจากคำไทยหนึ่งคำเกิดการการประสมกันของอักษรไทยหลายตัวเข้าด้วยกัน การเขียนติดกันไปอาจทำให้การแยกแยะคำมีปัญหา ดังนั้นในการแยกแยะระดับย่อยของคำสามารถนำหลักเกณฑ์ที่เรียกว่าอักขรวิธีมาใช้ให้เป็นประโยชน์

ปัจจุบันคำที่มาจากภาษาต่างประเทศที่ถูกนำมาใช้ร่วมกับภาษาไทยมีเป็นจำนวนมากขึ้น และคำเหล่านั้นนอกจากจะไม่ปรากฏในพจนานุกรมแล้ว หลายคำที่พบมีลักษณะการรวมกันของอักษรไทยที่แตกต่างออกไป เช่น เพาน์ด ฟิล์ม การ์ด เลาน์จ เพื่อให้กฎครอบคลุมและตัดคำได้อย่างมีประสิทธิภาพจึงได้เพิ่มกฎที่จัดการในส่วนนี้ไว้ด้วย

ในส่วนเนื้อหาส่วนที่ 2 จะกล่าวถึงลักษณะโครงสร้างส่วนประกอบของคำในภาษาไทย ส่วนที่ 3 คือวิธีการตัดคำภาษาไทยที่ใช้เป็นหลักพร้อมทั้งเปรียบเทียบถึงข้อดีข้อเสียของทั้งสามวิธี เนื้อหารายละเอียดของการตัดคำด้วยกฎที่นำเสนอในงานวิจัยนี้ จะอยู่ภายในส่วนที่ 4 และส่วนที่ 5 โดยวิธีการพัฒนาและการวัดผลการทดลองจะอธิบายไว้ในเนื้อหาส่วนที่ 6 พร้อมทั้งแสดงตัวอย่างขั้นตอนจากการนำกฎที่นำเสนอไปใช้ไว้งานจริงในส่วนที่ 7 ส่วนสุดท้ายคือบทสรุปสำหรับการใช้กฎที่นำเสนอตัดคำในเอกสารไทย

2. ลักษณะของภาษาไทย

ประโยคภาษาไทยประกอบไปด้วยคำ และคำในภาษาไทยก็ประกอบไปด้วยส่วนต่างๆ ซึ่งสามารถแบ่งได้เป็นแบบสามส่วน สี่ส่วน และห้าส่วน สามส่วนได้แก่ พยัญชนะ สระ และวรรณยุกต์ เช่นคำว่า กา ก่า ก้า ก๊า ก๋า เป็นต้น โดยคำว่า กา มีไม่มีรูปวรรณยุกต์ แต่มีเสียงวรรณยุกต์สามัญ ส่วนแบบสี่ส่วนจะเพิ่มเติมตัวสะกดเข้ามา เช่น กาย บิน รวม และสุดท้ายแบบห้าส่วนจะเพิ่มในส่วนของการันต์ได้แก่คำว่า การณ์ จลน์ เป็นต้น หลักการเหล่านี้จะเรียกว่าอักขรวิธี [4] การประสมกันระหว่างตัวอักษรก็มีหลักการอีกมากมาย อันได้แก่ การแบ่งอักษรไทยทั้ง 44 ตัวออกเป็นมาตรา คือ อักษรสูง อักษรกลาง และอักษรต่ำ การวางตำแหน่งของสระในคำ อักษรบางตัวที่ไม่นำไปเป็นตัวสะกด จากหลักการที่มีอยู่นี้ค่อนข้างแน่นอนพอสมควรในการนำไปแยกแยะคำไทยแต่ละคำ แต่ก็ยังไม่เพียงพอ เพื่อความถูกต้องยิ่งขึ้นจึงมีการพัฒนาและปรับปรุงกฎเพื่อให้ได้ความถูกต้องเพิ่มขึ้น

3. การตัดคำ

การใช้อักขรวิธีในการตัดคำสามารถทำได้ในระดับหนึ่งเนื่องจากคำบางคำเป็นคำที่เลียนเสียงจากภาษาต่างประเทศ ดังนั้นอาจมีการประสมคำนอกเหนือไปจากอักขรวิธี จึงมีการพัฒนาวิธีการต่างๆ ในการตัดคำในเอกสารไทย เพื่อให้ได้ความถูกต้องสูงสุด วิธีการหลักสำหรับตัดคำในเอกสารไทยมีดังนี้

3.1 การใช้กฎ

ลักษณะของการใช้กฎเพื่อตัดคำในภาษาไทย จะใช้ไวยากรณ์ทางภาษา โดยภาษาไทยจะแบ่งตัวอักษรเป็นหมวดหมู่ตามลักษณะการใช้งาน ได้แก่ กลุ่มพยัญชนะ กลุ่มสระ กลุ่มวรรณยุกต์ กลุ่มตัวเลขและกลุ่มตัวอักษรพิเศษ ขั้นตอนการตัดพยางค์จะทำจากซ้ายไปขวาเป็นส่วนใหญ่ ส่วนคำที่ไม่เป็นไปตามกฎที่สร้างไว้จะถูกเก็บไว้ในแฟ้มข้อมูล การวิเคราะห์โดยการหาขอบเขตหน้า (Front boundary recognition rule) และกฎการหาขอบเขตหลัง (Tail boundary recognition rule) [7] ได้เสนอกฎที่ได้จากคุณสมบัติการนำไปประสมกับอักษรไว้ในกฎกลุ่ม A และคุณสมบัติการนำสระไปไว้ในกฎกลุ่ม B การใช้กฎในการตัดคำนี้ยังคงประสบปัญหาการหาขอบเขตของคำ เนื่องจากคำหนึ่งคำอาจประกอบไปด้วยพยางค์เดียวหรือหลายพยางค์ จึงต้องมีการนำวิธีการอื่นเข้ามาในการตัดคำนอกเหนือไปการการตัดคำด้วยกฎเพียงอย่างเดียว

3.2 การใช้พจนานุกรม

การนำพจนานุกรมมาใช้ จะทำให้ผลลัพธ์ที่ได้อยู่ในระดับคำ โดยมีหลักการว่าให้ทำการตรวจสอบสายอักขระ (String) ซึ่งเป็นชุดของตัวอักษรที่ได้จากเอกสาร จากนั้นจะนำอักษรแต่ละตัวไปค้นหาคำจากพจนานุกรม หากพบคำในพจนานุกรมที่สามารถเป็นคำในสายอักขระนั้นได้มากกว่าหนึ่งคำ จะทำการเลือกคำที่ยาวที่สุด (Longest matching) หากอักษรตัวต่อมาไม่สามารถพบคำที่ตรงกับที่ในพจนานุกรมมีอยู่จะทำการย้อนกลับไปเลือกคำที่สั้นกว่าแทนเรียกวิธีการนี้ว่าวิธีการย้อนรอย (Back tracking) [6] นอกจากนี้ยังมีการนำเสนอวิธีแยกพยางค์ด้วยกฎก่อนแล้วจึงใช้พจนานุกรมเพื่อตัดและรวบรวมให้เป็นคำ [1]

โดยที่เทคนิคนี้จะใช้โปรแกรมเล็กซ์ (Lex) [1] สร้างกลุ่มตัวอักษร (Token) ที่ยาวที่สุดก่อนโดยไม่รวมตัวสะกด เนื่องจากตัวสะกดที่อยู่ท้ายคำอาจมีโอกาสเป็นพยัญชนะต้นของคำถัดมาได้

จากนั้นจึงวิเคราะห์อีกครั้งโดยพิจารณาตัวสะกดร่วมด้วยพร้อมกับการใช้พจนานุกรม ภาพรวมของระบบการตัดคำที่นำเสนอปรากฏในรูปที่ 1

3.3 การใช้คลังข้อความ

การใช้คลังข้อความ (Corpus) ในการตัดคำในเอกสารไทยเป็นการนำค่าทางสถิติมาร่วมพิจารณา เช่นค่าสถิติการใช้คำ ค่าสถิติหน้าที่ของคำ งานวิจัยที่ใช้คลังข้อความมาใช้ในการตัดคำมีจุดประสงค์เพื่อเพิ่มความถูกต้องในการตัดคำและลดคำกำกวม ยกตัวอย่างเช่น ที่อยู่ ที่ตั้ง (ซึ่งอาจเป็นได้ทั้ง ที่ อยู่, ที่ ตั้ง) จะนำคำที่ผ่านการตัดคำมาระดับหนึ่งแล้วอาจเป็นจากการใช้กฎหรือการใช้พจนานุกรมที่กล่าวมาข้างต้น มาผ่านการวิเคราะห์ด้วยคลังข้อความ ความรู้ภายในคลังข้อความอาจเป็นค่าสถิติหรือลักษณะไวยากรณ์ เป็นต้น

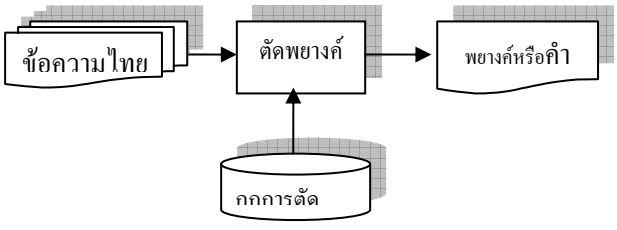
จากวิธีการสามวิธีหลักที่กล่าวมาสามารถสรุปเป็นข้อดีและข้อด้อยดังตารางที่ 1

ตารางที่ 1 ข้อดีข้อเสียของการตัดคำด้วยวิธีต่างๆ

	ความเร็ว	ความถูกต้อง		ขนาดของคลังข้อความ	การตัดคำที่ไม่มีอยู่ในพจนานุกรม
		ระดับพยางค์	ระดับคำ		
กฎ	///	//	/	/	//
พจนานุกรม	//	//	//	//	/
คลังข้อความ	/	//	///	///	/

///มาก // ปานกลาง / น้อย

4. วิธีการตัดคำโดยใช้กฎที่นำเสนอ



รปที่ 1 ภาพรวมของระบบการตัดคำด้วย

ลักษณะประโยคหรือวลีของภาษาไทยเกิดจากการรวมกันของคำ ซึ่งแต่ละคำอาจเป็นคำเดี่ยวหรือเกิดจากการรวมกันของคำคำอื่น เมื่อผ่านขั้นตอนแรกเพื่อคัดแยกข้อมูลจากเอกสารให้เป็นอนุประโยค ดังนั้นหากแทนค่า D เป็นเอกสารที่ต้องการนำมาแยกคำและ s[i] เป็นอนุประโยคหรือประโยคที่ได้จากการแบ่งครั้งแรก ณ ตำแหน่งเว้นวรรคหรืออักขระพิเศษที่กำหนดไว้จากเอกสาร D ดังสมการ (1)

$$Document \quad D = \sum_{i=1}^{N} s[i] \quad (1)$$

คำที่อยู่ในประโยค s[i] ให้แทนด้วย w[i,j] และพยางค์ที่ประกอบขึ้นเป็น w[i,k] แทนด้วย c[i,j] ภายใน s[i] จะประกอบไปด้วย Character c[i,j] ซึ่งเป็นอักขระต่างๆ ใน s[i] ซึ่ง i มีค่าตั้งแต่ 1 ถึง N และ j มีค่าตั้งแต่ 1 ถึง L โดยที่ L คือความยาว String s[i] c[i,j] อาจเป็นอักขระไทยหรือไม่ก็ได้ ดังนั้นการพิจารณาคำไทยจึงต้องตรวจสอบว่า c[i,j] เป็นอักษรไทยที่เป็นส่วนหนึ่งของคำไทยได้หรือไม่ นั่นคือ character in Thai ct[i,j] และWord w[i,k] คือคำไทยที่ทำการตัดออกมาได้จากประโยคที่ i โดยที่ w[i,k] เกิดจากการรวมกันของ ct[i,j] โดยที่ค่า k และ j มีค่าระหว่าง 1 กับ L และ i มีค่าระหว่าง 1 กับ N ซึ่งจะได้ตามดังสมการ(2)-(3)

$$s[i] = \sum_{j=1}^{L} c[i, j] \quad (2)$$

$$w[i, j] = \sum_{j=first}^{end} ct[i, j] \quad (3)$$

โดยที่ $1 \le first \le end \le L$

และ $ct[i, j] \in c[i, j]$

4.1 กฎที่นำเสนอ

เมื่อนำ ct[i,j] มาพิจารณาโดยใช้กฎจะได้ผลลัพธ์ในระดับพยางค์ จากกฎที่มีอยู่เดิม [1][7] ได้เพิ่มกฎสำหรับการตัดคำที่มีความเกี่ยวข้องกับภาษาต่างประเทศและคำที่อยู่นอกเหนือจากการประสมตามอักขรวิธี

กฎการรวมกันของต่างภาษา

$$s[i] = w[i, j]$$

ถ้าหาก $c[i, j] \in w[i, j]$ และ $c[i, j] \neq ct[i, j]$

ยกตัวอย่างเช่น ส-8 3กช6 หรือ ม.6/5

กฎการเพิ่มตัวสะกด คำที่มาจากภาษาต่างประเทศบางคำมีการสะกดที่แตกต่างไปจากอักขรวิธีที่มีอยู่แต่เดิม การมีตัวสะกดมากกว่าหนึ่ง หรือพบว่าสระที่เป็นตัวสุดท้ายของคำเสมอนั้นไม่ใช่เสมอไป เช่น เลาน์จ เพาว์น เป็นต้น

4.2 การค้นหาจากพจนานุกรม

การค้นหาคำศัพท์จากพจนานุกรมจะนำ w[i,j] ที่ได้จากการตัดคำด้วยกฎที่นำเสนอ เพื่อรวมกันให้เป็นคำศัพท์ที่มีอยู่ในพจนานุกรม การค้นหาคำหากพบมากกว่าหนึ่งคำปรากฏอยู่จะใช้เทคนิค Longest Matching เพื่อเลือกคำศัพท์ที่ยาวกว่า

5. วิธีการพัฒนาและวัดผลการทดลอง

การพัฒนาการตัดคำภาษาไทยโดยใช้กฎที่นำเสนออยู่บนมาตรฐาน ANSI C บนระบบปฏิบัติการ Windows ข้อมูลที่ถูกนำมาทำการตัดคำถูกนำมาจากเอกสารที่พบเห็นทั่วไปได้แก่นิตยสารและหนังสือพิมพ์ รวมทั้งบทความทางวิชาการที่มีทั้งคำไทยและคำที่มาจากภาษาต่างประเทศ คำที่ได้จากเอกสารจะถูกนำไปแบ่งครั้งแรกด้วยช่องว่างระหว่างประโยค และนำไปแยกในระดับพยางค์โดยการใช้กฎ ผลที่ได้จากการแบ่งด้วยกฎจะถูกเก็บเป็นสองส่วน คือ ส่วนที่ต้องนำไปค้นหาต่อในพจนานุกรมและไม่ต้องนำไปค้นหาต่อ

คำที่ไม่ต้องนำไปค้นหาต่อได้แก่คำดังนี้ คำปรากฏสัญลักษณ์ - / , . คำที่มีตัวเลขและคำที่มีตัวอักษรต่างประเทศ อยู่ระหว่างประโยคที่ถูกแบ่งครั้งแรกด้วยช่องว่าง ซึ่งให้ถือเป็นคำสมบูรณ์แล้ว นอกจากคำเหล่านี้คำอื่นๆ ที่ได้จากการแบ่งด้วยกฎจะนำไปค้นหาในพจนานุกรมเพื่อทำการแบ่งในระดับคำต่อไป

การวัดผลการทดลองใช้ค่าความถูกต้องของคำ (Word validity) หน่วยวัดประสิทธิภาพคือสัดส่วนของคำที่ตัดได้ถูกต้องต่อจำนวนคำที่ตัดออกมาได้

6. ตัวอย่างการตัดคำจากวิธีที่นำเสนอ

จากการวิธีการตัดคำโดยใช้กฎที่นำเสนอขึ้นซึ่งได้กล่าวไว้ในส่วนที่สี่และห้า ตัดคำจากเอกสารไทยที่มีส่วนผสมระหว่างคำไทยและคำจากต่างประเทศ จากประโยคตัวอย่างจะแสดงขั้นตอนการทำงานของการตัดคำไว้ดังในตารางที่ 2 นี้

ตารางที่ 2 ตัวอย่างจากการใช้กฎที่นำเสนอ

ประโยคตัวอย่าง	“เขานั่งที่เลาน์จเพื่อรอเพื่อน ก่อนจะนั่งรถทะเบียน กข3477 ไปที่สนามบิน”
ขั้นที่ 1	เขานั่งที่เลาน์จเพื่อรอเพื่อน ก่อนจะนั่งรถทะเบียน กข3477 ไปที่สนามบิน
ขั้นที่ 2	เขา นั่ง ที่ เลาน์ จ เพื่อ รอ เพื่อน
ขั้นที่ 3	เขา นั่ง ที่ เลาน์จ เพื่อ รอ เพื่อน
ขั้นที่ 4	ก่อน จะ นั่ง รถ ทะ เบียน
ขั้นที่ 5	ก่อน จะ นั่ง รถ ทะเบียน
ขั้นที่ 6	ไป ที่ ส นาม บิน
ขั้นที่ 7	ไป ที่ ส นาม บิน
ผลลัพธ์	เขา นั่ง ที่ เลาน์จ เพื่อ รอ เพื่อน ก่อน จะ นั่ง รถทะเบียน กข3477 ไป ที่ สนาม บิน

จากตารางที่ 2 จะข้ามคำว่า กข3477 ซึ่งเป็นคำที่มีตัวเลขอยู่ในประโยคที่แยกออกมาครั้งแรกด้วยช่องว่าง คำนี้ให้ถือเป็นคำสมบูรณ์เนื่องจากเป็นคำที่ไม่พบในพจนานุกรมและการเขียนอักษรต่างภาษาอยู่ติดกันให้ถือว่าทั้งคำเป็นคำคำเดียว

7. เปรียบเทียบผลการตัดคำด้วยกฎเดิมและกฎที่นำเสนอ

การตัดคำนี้ได้นำเอกสาร 3 ประเภทได้แก่ บทความทางวิชาการ นิตยสารและหนังสือพิมพ์ ซึ่งแต่ละประเภทมีจำนวน 10 เอกสาร แต่ละเอกสารมีขนาดประมาณ 600-1000 คำ ค่าความถูกต้องของคำจากการใช้กฎเดิมและกฎที่

นำเสนอในการตัดคำระดับพยางค์และระดับคำ ได้ถูกแสดงไว้ ตามตารางที่ 3

ตารางที่ 3 เปรียบเทียบผลการตัดพยางค์และคำที่ถูกต้องที่มาได้จากกฎเดิมและกฎที่นำเสนอ

เอกสารที่ใช้ทดสอบ	ระดับพยางค์ %		ระดับคำ %	
	แบบเดิม	ปรับปรุง	แบบเดิม	ปรับปรุง
บทความ	90.02	92.01	85.60	87.36
นิตยสาร	92.97	95.11	88.45	88.98
หนังสือพิมพ์	90.66	90.76	85.72	85.73

จากตารางที่ 3 จะเห็นได้ว่าการใช้กฎที่นำเสนอจะได้ค่าความถูกต้องของคำที่ได้สูงกว่าการใช้กฎแบบเดิม ซึ่งกฎเดิมและกฎที่นำเสนอใช้ข้อมูลจากพจนานุกรมเดียวกัน เพื่อเข้ามาสนับสนุนการตัดในระดับคำ เนื่องจากการตัดคำในระดับคำมีความซับซ้อนกว่าระดับพยางค์ ค่าความถูกต้องของคำโดยเฉลี่ยจึงมีค่าต่ำกว่าการตัดระดับพยางค์ ถึงแม้ว่าเวลาที่ใช้ในการตัดคำโดยการใช้กฎที่นำเสนอใช้เวลามากกว่าเวลาที่ใช้โดยการใช้กฎเดิมแต่ว่าไม่ได้แตกต่างกันมากนัก

8. สรุป

จากการเปรียบเทียบผลการตัดคำในเอกสารไทยที่ผ่านมาทำให้สรุปได้ว่าภาษาไทยมีลักษณะซับซ้อน การปรับปรุงกฎเพื่อให้มีความยืดหยุ่นของโครงสร้างคำไทย ทำให้การตัดคำด้วยกฎสามารถลดปัญหาคำที่มาจากภาษาต่างประเทศได้ส่วนหนึ่งอีกทั้งการนำพจนานุกรมเข้ามาช่วยเสริมนอกเหนือไปจากการตัดคำด้วยกฎเพียงอย่างเดียวทำให้ได้ผลลัพธ์จากการตัดคำในเอกสารไทยถูกต้องยิ่งขึ้นในระดับพยางค์และคำ ซึ่งในอนาคตการตัดคำโดยการใช้กฎที่นำเสนอนี้จะนำไปใช้กับเอกสารที่มีขนาดใหญ่ขึ้นและประเภทเอกสารที่นำมาทดสอบมีความแตกต่างทางด้านเนื้อหาหลากหลายยิ่งขึ้น

เอกสารอ้างอิง

3. ดวงแก้ว สวามิภักดิ์, *การสร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกซ์*: มหาวิทยาลัยธรรมศาสตร์,2533.

4. บรรจบ พันธุเมธา, *ลักษณะภาษาไทย* กรุงเทพ: สำนักพิมพ์มหาวิทยาลัยรามคำแหง. 1-45. 2540.

5. พิสิทธิ์ พรมจันทร์, *การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรมแยกคำภาษาไทย,* จุฬาลงกรณ์มหาวิทยาลัย. 2540.

6. พระยาอุปกิตศิลปสาร, *หลักภาษาไทย*, กรุงเทพฯ : โรงพิมพ์ไทยวัฒนาพานิช. 18-28. 2539.

7. ยืน ภู่วรวรรณ, “การวิเคราะห์ข้อมูลคำไทย “.

8. ยืน ภู่วรวรรณ และ วิวรรธ อิ่มอารมณ์, “การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี”. รายงานการประชุมวิชาการวิศวกรรมไฟฟ้าครั้งที่ 9, 2529.

9. สุรินทร์ จรรยาพรพงษ์. *A Thai Syllable Seperation Algorithm.* Asian Institute of Technology, 1983.

10. หัชทัย ชาญเลขา, อัศนีย์ ก่อตระกูล. การสกัดนิพจน์ระบุนามในภาษาไทยโดยใช้แมกซิมัมเอนโทรปีโมเดลและอิงความรู้. หน่วยปฏิบัติการวิจัยเชี่ยวชาญเฉพาะการประมวลผลภาษาธรรมชาติและเทคโนโลยีสารสนเทศอัจฉริยะ. มหาวิทยาลัยเกษตรศาสตร์.

11. B. Kijsirikul. “Comparing Winnow and RIPPER in Thai Named-Entity Identification”, Chulalongkorn.

12. C. Kooptiwoot. “Segmentation of Ambiguous Thai Words by Inductive Logic Programming”. Chulalongkorn. 1999.

13. D. D. Plamer. “A Trainable Rule-based Algorithm for Word Segmentation”.

14. P. Charoenpornsawat, B. Kijsirkul, “Feature-based Proper Name Identification in Thai”, Chulalongkorn. 1998.

15. P. Charoenpornsawat, B. Kijsirkul, S. Meknavin. “Feature-based Thai Unknown Word Boundary Identification Using Winnow”, Chulalongkorn. 1998.
16. T. Pongthai, V. Sornlertlamvanish. “Grapheme to Phoneme for Thai”, NECTEC.
17. T. Theeramunkong, V. Sornlertlamvanich, T. tanhermhong, W. Chinnan. "Character Cluster Based Thai Information Retrieval”, National Electronics and Computer Technology Center (NECTEC), National Science and Technology Development Agency (NSTDA).
18. V. Sornlertlamvanich, T. Potipiti, T. Charoenporn. “Automatic Corpus-based Thai Word Extraction with the C4.5 Learning Algorithm”. National Electronics and Computer Technology Center (NECTEC).
19. V. Tesprasit, P. Charoenpornsawat , V. Sornlertlamvanich. “Learning Pharse Break Detection in Thai Text-to-Speech”. EUROSPEECH, 2003.
20. W. Arronmanakun. “Collocation and Thai Word Segmentation”.

การตัดคำภาษาไทยด้วยวิธีปรับปรุงกฎและพจนานุกรมแบบใหม่

Improved Rule-Based and New Dictionary for Thai Word Segmentation

ปโยธร อุราธรรมกุล, กานดา รุณนะพงศา

ภาควิชาวิศวกรรมคอมพิวเตอร์, คณะวิศวกรรมศาสตร์, มหาวิทยาลัยขอนแก่น, 40002, ประเทศไทย

E-mail: payothorn@gmail.com , krunapon@kku.ac.th

Abstract

Thai word segmentation is the process of separation Thai words from one another which is in Thai documents in order to use it in other aspects, such as speech synthesis translation. The present document does not contain only Thai words, but also exist some foreign words which are spelt in the form of Thai language. These foreign words are combined characters referring to alphabets differently beyond Rule-based segmentation. Since there are many these special words, this article proposes the improvement of Rule-based segmentation in order to make the segmentation more flexible and adaptable. This technique will be useful for particularly the document with unknown words or words that are not in Thai dictionary.

Keywords: Thai Word segmentation, Rule-based segmentation, Dictionary

บทคัดย่อ

การตัดคำไทย (Thai Word Segmentation) คือการแยกแต่ละคำในเอกสารไทยออกจากกันเพื่อนำไปใช้ประโยชน์ในด้านอื่นๆ เช่น การสังเคราะห์เสียงพูด การแปลภาษา เป็นต้น เอกสารที่มีอยู่ ณ ปัจจุบันไม่เพียงแต่จะมีคำไทยแท้เท่านั้นยังมีคำบางคำที่มาจากภาษาต่างประเทศที่ถูกสะกดอยู่ในรูปของคำอ่านภาษาไทย คำบางคำจะมีการผสมอักษรที่แตกต่างนอกเหนือออกไปจากกฎการตัดคำ (Rule-based) แบบเดิมที่มีอยู่ เนื่องจากคำเหล่านี้มีอยู่มากมายและเกิดใหม่อยู่เสมอ บทความนี้นำเสนอการปรับปรุงกฎการตัดคำให้มีความยืดหยุ่นมากขึ้นจะเป็นประโยชน์สำหรับการตัดคำที่ไม่รู้จักหรือไม่มีความหมายอยู่ตามพจนานุกรม (Dictionary-based)

คำสำคัญ การตัดคำภาษาไทย, กฎการตัดคำ, พจนานุกรม

1. บทนำ

การตัดคำได้รับการพัฒนาขึ้นมาโดยใช้วิธีการต่างๆ ที่ต่างกัน เนื่องจากการตัดคำเป็นกระบวนการพื้นฐานของการประมวลผลภาษาธรรมชาติ เช่น การวิเคราะห์เสียงพูด การตัดคำภาษาไทยเองก็เช่นกัน ได้มีผู้คิดค้นวิธีที่จะแยกคำแต่ละคำออกจากประโยคซึ่งมีการเขียนติดกันไปอย่างต่อเนื่องทั้งประโยค ในงานวิจัยนี้จะกล่าวถึงการตัดคำโดยอาศัยอักขรวิธีเป็นหลักการพื้นฐานการประสมคำ

1.1 ลักษณะของภาษาไทย

ภาษาไทยมีลักษณะแตกต่างจากภาษาอังกฤษ หรือภาษาจีน เนื่องจากในภาษาไทยมีการเขียนติดกันไปทั้งประโยค อีกทั้งคำไทยคำหนึ่งอาจประกอบไปด้วยสระที่เป็นสระประกอบ คือมาจากสระอื่นอีกหลายตัวประกอบกัน เช่น สระเอียะ สระเอือะ เป็นต้น และพยัญชนะบางตัวยังสามารถทำหน้าที่เป็นได้ทั้งตัวสะกด หรือสระด้วยก็ได้ ดังนั้นการแยกแยะ

ในหน่วยย่อยของคำสามารถนำหลักเกณฑ์ที่เรียกว่าอักขรวิธีมาใช้

1.2 อักขรวิธี

คำในภาษาไทยเกิดจากส่วนต่างๆ ของอักษรไทยประกอบกันอย่างน้อยสามส่วน ได้แก่ ส่วนพยัญชนะ สระ และวรรณยุกต์ พยัญชนะของไทยถูกแบ่งออกเป็นอักษรสามหมู่ที่เรียกว่าไตรยางค์ ได้แก่ อักษรสูง อักษรกลาง และอักษรต่ำ สระก็ถูกจัดเป็นประเภท สระเดี่ยว สระประสม วรรณยุกต์ก็มี 4 รูป 5 เสียง การจะทำให้เกิดเสียงและความหมายต้องเกิดจากกฎเกณฑ์ที่มีอยู่

งานวิจัยนี้เสนอการตัดคำภาษาไทยโดยใช้การปรับปรุงกฎเพื่อเพิ่มความยืดหยุ่นให้กับการสะกดคำและการพัฒนาพจนานุกรมเพื่อเพิ่มประสิทธิภาพสำหรับการตัดคำ ในส่วนที่ 2 และ 3 จะกล่าวถึงวิธีและขั้นตอนต่างๆ ทั่วไปที่ใช้สำหรับตัดคำ ส่วนที่ 4 จะกล่าวถึงขั้นตอนการตัดคำที่นำเสนอในงานวิจัยชิ้นนี้ ส่วนที่ 5 ที่เป็นส่วนสุดท้ายจะกล่าวถึงผลสรุปของการตัดคำโดยวิธีที่นำเสนอ

2. งานวิจัยที่เกี่ยวข้อง

การตัดคำได้รับการพัฒนามาเป็นเวลาพอสมควร ทำให้เกิดแนวคิดเกี่ยวกับการตัดคำขึ้นหลากหลาย เทคนิคการตัดคำแบ่งออกเป็นลักษณะใหญ่ๆ ได้ดังนี้

2.1 การตัดพยางค์

การตัดพยางค์เป็นการใช้หลักการของภาษาไทยที่มีกฎเกณฑ์ค่อนข้างตายตัว ยกเว้นบางพยางค์ งานวิจัยที่เกี่ยวข้องกับการตัดพยางค์ได้แก่ งานวิจัยของยุพิน ไทยรัตนานนท์ [12] ซึ่งใช้กฎในการผสมกันของพยางค์ แบ่งตัวอักษรออกเป็น 5 กลุ่มคือ กลุ่มพยัญชนะ สระ วรรณยุกต์ ตัวเลขและอักขระพิเศษ โดยใช้อักษรดังนี้แทนแต่ละกลุ่ม

C แทนพยัญชนะต้น, V แทนสระ, S แทนตัวสะกด
T แทนวรรณยุกต์, G แทนการันต์
ยกตัวอย่างเช่น
สิ้น CTVS, ศิลป์ CVSSG, กวาง CCVS

กฎนี้ไม่ยุ่งยากซับซ้อนมากนักจึงไม่มีความยืดหยุ่นเท่าที่ควร อีกทั้งอักษรไทยบางตัวสามารถเป็นได้ทั้งตัวสะกดและพยัญชนะต้น นั่นคือเป็นได้ทั้ง C และ S ทำให้การตัดคำบางคำเป็นไปอย่างไม่ถูกต้องเท่าที่ควร

ส่วนงานของสุรินทร์ จรรยาพรพงษ์ [10] ได้ใช้เทคนิคที่เรียกว่ากฎการหาขอบเขตหน้า และกฎการหาขอบเขตหลังและในแต่ละกฎยังแบ่งออกเป็น 2 กลุ่มย่อย หากแบ่งตามลักษณะของตัวอักษรจะจัดอยู่ในกลุ่ม A แบ่งตามลักษณะการใช้สระจะแบ่งเป็นกลุ่ม B

ตัวอย่างกฎของสุรินทร์

A-1F ได้แก่สระอะ อา อิ อี อึ อื อุ อู ไต่คู้ อำ ไม้หันอากาศ และรูปวรรณยุกต์ทุกตัว

A-2F คือสระที่มักอยู่หน้าคำ ได้แก่ เอ แอ ไอ โอ

A-3F คือสระที่อยู่หน้าคำเสมอ ได้แก่ ใอ เป็นต้น

แต่การใช้กฎเพียงอย่างเดียวยังคงประสบปัญหาการหาขอบเขตของคำ เนื่องจากคำหนึ่งอาจประกอบไปด้วยพยางค์เดียวหรือหลายพยางค์ จึงต้องมีวิธีการอื่นเข้ามาช่วยอีกทั้งยังไม่มีการใช้พจนานุกรมจึงไม่สามารถตัดคำในระดับคำได้ดีนัก

งานของดวงแก้ว สวามิภักดิ์ [1] ได้สร้างกฎไวยากรณ์ขึ้นมาพร้อมทั้งใช้พจนานุกรมประกอบ โดยกฎที่สร้างขึ้นประกอบด้วย 43 กฎ แต่งานนี้ไม่ครอบคลุมไปถึงวิธีการสะกดคำบางคำ เช่นคำที่มาจากภาษาต่างประเทศที่มีตัวสะกด การันต์ และตัวสะกดต่างจากอักขรวิธีของไทย ซึ่งจะทำให้เกิดข้อผิดพลาดในการตัดคำในเอกสารที่มีทั้งภาษาไทยและภาษาอังกฤษปนกัน โดยเฉพาะคำภาษาไทยที่มาจากการสะกดคำอ่านภาษาอังกฤษด้วยภาษาไทย อย่างเช่น เว็บเซอร์วิสเป็นโปรแกรมที่สามารถติดต่อได้โดยตรงกับอีกโปรแกรมหนึ่ง

2.2 การตัดคำ

การตัดคำนิยมใช้พจนานุกรมเข้ามาช่วย ได้แก่งานวิจัยของยืน ภู่วรวรรณและวิวรรธน์ อิ่มอารมณ์ [6] เนื่องจากพจนานุกรมสามารถตัดคำได้ชัดเจนกว่าการตัดพยางค์ เพราะคำอาจเกิดจากการมารวมกันของหลายพยางค์ แต่ขอบเขตของคำยังคงซับซ้อนและกำกวม ในงานวิจัยได้ใช้เทคนิคที่เรียกว่า

วิธีการย้อนกลับ (Back Tracking) [1] และการเลือกคำที่ยาวที่สุด[1] (Longest Matching)

วิธีการนี้มีข้อดีคือย้อนกลับไปเพื่อเลือกคำอีกครั้งได้แต่ข้อเสียคือเมื่อหากคำนั้นเมื่อย้อนกลับไปแล้วยังไม่พบตามพจนานุกรมทำให้เสียเวลา

3. ขั้นตอนวิธีการตัดคำโดยทั่วไป

การตัดคำเริ่มโดยการตัดที่ส่วนที่เป็นช่องว่างระหว่างประโยค อนุประโยค หรือคำ

3.1 การตัดอนุประโยคโดยอาศัยช่องว่าง และ อักขระพิเศษ

จะพิจารณาส่วนที่ติดกันของตัวอักษรในภาษาไทย หากมีช่องว่างระหว่างคำหรือปรากฏตัวอักษรอื่นที่ไม่ใช่อักษรไทยให้ถือว่าเป็นคนละข้อความซึ่งไม่มีความเกี่ยวข้องกันในระดับคำ ซึ่งเอกสารหนึ่งๆ (D) จะประกอบด้วยหลายประโยคหรืออนุประโยค (S_i)

$D = S_1 + S_2 + S_3 + \ldots + S_N$

3.2 การตัดคำโดยอาศัยกฎการผสมอักษรในภาษาไทย

สามารถแบ่งประเภทของพยัญชนะไทย 44 ตัว ออกเป็น 3 หมู่เรียกว่าไตรยางค์ ได้แก่ อักษรสูง อักษรกลาง และอักษรต่ำ และเมื่อพิจารณาถึงการนำไปผสมเป็นระดับพยางค์หรือคำแบ่งเป็น 3 ส่วน, 4 ส่วน และ 5 ส่วนดังนี้

3 ส่วน ได้แก่ พยัญชนะ+สระ+วรรณยุกต์เช่น ตา ดี ไป นา

4 ส่วน ได้แก่ พยัญชนะ+สระ+วรรณยุกต์ +ตัวสะกด เช่น คน

5 ส่วน ได้แก่ พยัญชนะ+สระ+วรรณยุกต์ +ตัวสะกด+ตัวการันต์ เช่น แพทย์ สิทธิ์ ฤทธิ์

โดยกฎการแบ่งส่วนของพยางค์จะได้ว่าสามารถตัดประโยคหรืออนุประโยคที่ประกอบด้วยอักษรน้อยกว่า 4 ตัวอักษรให้เป็น 1 คำหรือพยางค์ได้ทันทีโดยไม่ต้องเปรียบเทียบกับกฎหรือพจนานุกรม เนื่องจากพยางค์ที่สั้นที่สุดคือ 3 ส่วน หากในกรณีที่วรรณยุกต์อยู่ในรูปสามัญจะประกอบด้วยอักษร 2 ตัว (กรณีนี้ไม่รวมถึง ธ ณ ฤ ที่จะมีช่องว่างอยู่หน้าและหลังเสมอ) นั่นหมายถึงหากจะเป็น 2 คำหรือ 2 พยางค์ขึ้นไปต้องประกอบไปด้วย 4 ตัวอักษรขึ้นไป

$S_i = C_{i1} + C_{i2} + C_{i3} + \ldots + C_{iL}$

$S_i = W_{i1} + W_{i2} + W_{i3} + \ldots + W_{iN}$

$W_{i1} = C_{i1} + C_{i2} + C_{i3} + \ldots + C_{iL}$ เมื่อ $L < 4$

เมื่อ C คือตัวอักษรในประโยคหรืออนุประโยค S

W คือคำที่ตัดได้จากประโยคหรืออนุประโยค S

นอกจากคำไทยแท้แล้ว ภาษาไทยได้มีการรับภาษาต่างประเทศเข้ามาใช้ เช่น บาลี สันสกฤต อังกฤษ เป็นต้น ทำให้เกิดคำที่ไม่ตรงกับหลักการผสมคำอยู่มาก เช่น พรหม การ์ด มาร์ค เลาน์จ ซึ่งในปัจจุบันนี้จะพบคำที่มาจากภาษาต่างประเทศมากขึ้นและมีการถอดความเป็นภาษาไทยไม่ตรงกับหลักไวยากรณ์ไทย

4. วิธีการตัดคำที่นำเสนอ

การตัดคำเริ่มจากเอกสารนำเข้า แยกออกเป็นอนุประโยคย่อย S_i โดยใช้ช่องว่างเป็นตัวแบ่งและพิจารณาว่ามีอนุประโยคใดบ้างสามารถเป็นคำได้ทันที โดยให้

$S_i = C_{i1} + C_{i2} + C_{i3} + \ldots + C_{iL}$

$S_i = W_{i1} + W_{i2} + W_{i3} + \ldots + W_{iN}$

$W_{i1} = C_{i1} + C_{i2} + C_{i3} + \ldots + C_{iL}$ เมื่อ $L < 4$

เมื่อ C คือตัวอักษรในประโยคหรืออนุประโยค S

W คือคำที่ตัดได้จากประโยคหรืออนุประโยค S

ยกตัวอย่างเช่น ฯลฯ , ฯพณฯ , ธ , ฤ , กิน , ถุง , ค้า เป็นต้น

หลังจากนั้นอนุประโยคอื่นที่เหลือจะนำไปวิเคราะห์ต่อไป

4.1 การแบ่งประเภทของอนุประโยค

นำอนุประโยคมาทำการตัดคำขั้นแรกโดยแยกประโยคเป็น 3 ประเภทด้วยกันตามส่วนประกอบของอนุประโยค

<u>ประเภทที่ 1</u> ประกอบด้วยอักษรไทย หรืออักษรไทยซึ่งอยู่ติดกับอักษรแบ่งวรรคอื่นๆ เช่น (,) , “ , ‘ เป็นต้น ยกเว้น - , /

การตัดคำประโยคประเภทนี้จะทำการแยกอักษรไทยกับอักษรแบ่งวรรคออกจากกัน เช่น

“ระบอบการปกครอง:2475” จะแยกได้ว่า

“ ระบอบการปกครอง : 2475 ” เนื่องจาก 2475 ไม่ได้เขียนอยู่ติดกับอักษรไทยโดยตรงอนุประโยคนี้จึงถือเป็นประเภทที่ 1

ประเภทที่ 2 ประกอบด้วยอักษรไทยซึ่งอยู่ติดกับตัวเลข หรืออักษรแบ่งวรรค - , / หรือ อักษรต่างประเทศ หรืออักษรพิเศษอื่นๆ จะทำการแยกอักษรแบ่งวรรคออกจากกัน ยกเว้นตัวเลขเครื่องหมาย – และ / จะยังคงเขียนติดกับอักษรไทยไว้เช่นนั้น และถือเป็นคำ 1 คำทันที เช่น “ก-2547” และ “23/2ก” จะแยกได้ว่า

“ ก-2547 ” และ “ 23/2ก ”

ประเภทที่ 3 อักษรต่างประเทศหรืออักษรแบ่งวรรคยกเว้นเครื่องหมาย - และ / ประเภทที่ 3 นี้ไม่มีอักษรไทยอยู่ในประโยคเลย อนุประโยคประเภทนี้จะทำการแยกอักษรต่างประเทศและอักษรแบ่งวรรคออกจากกัน และนำอนุประโยคที่ได้มาทำการแปลงเป็นคำอ่านภาษาไทยและเก็บไว้สำหรับตัดคำในเอกสารที่ตรงกันหรือมีความใกล้เคียง

เช่น (wonderful) จะได้เป็นคำอ่านเป็นภาษาไทยดังนี้

won – {วอน , วัน , วอน์ , โวน์ , โวน}

der – {เดอ , เดอร์ , เดร์}

ful – { ฟุล , ฟูล , ฟุล์ , ฟูล์ , ฟัล}

จากนั้นจะทำการเก็บคำอ่านคำนี้ไว้เพื่อใช้เปรียบเทียบกับคำในเอกสารที่มีความเป็นไปได้ที่จะตรงกับคำนี้

4.2 การวิเคราะห์หาคำที่มีอยู่ในพจนานุกรมและหาคำที่ใช้อักษรไทยสะกดคำอ่านภาษาต่างประเทศ

คำไทยประเภทที่ 2 และ 3 ที่ผ่านขั้นตอน 4.1 จะนำมาตัดคำโดยอิงกับพจนานุกรม และคำต่างประเทศที่พบในเอกสาร จากนั้นจะได้เอกสารมา 2 ส่วน

ส่วนแรก พบคำตามพจนานุกรม หากพบคำในพจนานุกรมมากกว่าสองครั้งจะถือเอาคำที่ยาวที่สุดเป็นหลัก (Longest matching) คำที่เก็บอยู่ในพจนานุกรมนี้เป็นคำที่พบได้ทั่วไป หรือคำที่มีความยาวหลายพยางค์หรือคำที่มีการสะกดตรงตามอักขรวิธี

ส่วนที่สอง ไม่พบตามพจนานุกรม โดยปรกติส่วนนี้หากเป็นคำเดี่ยวๆ จะนำไปพิจารณากับคำรอบข้างซึ่งมีความเป็นไปได้ที่จะเป็นคำเดียวกันโดยอาศัยกฎการวิเคราะห์ความเป็นไปได้ที่จะเป็นภาษาต่างประเทศหรือคำเฉพาะ คำเฉพาะบางส่วนจะถูกเก็บอยู่ในพจนานุกรมคำเฉพาะและสามารถปรับปรุงได้เพื่อความเหมาะสมกับเอกสารแต่ละประเภท คำเฉพาะที่เหมาะสมได้แก่คำที่มาจากบาลี สันสกฤต ที่นำตัวอักษรเหล่านี้มาใช้ ฆ , ญ , ฒ , ฎ , ฏ , ฒ , ผ , ภ , ฌ , ศ , ษ ซึ่งมักไม่ค่อยพบอยู่กับคำที่มาจากภาษาต่างประเทศ เป็นต้น เช่นชื่อคน ชื่อสถานที่

4.2.1 การสร้างพจนานุกรมแบบใหม่

คำที่ไม่พบในพจนานุกรมทั่วไปมักเป็นคำเฉพาะเช่นชื่อคนหรือสถานที่ คำใหม่ และคำที่มาจากภาษาต่างประเทศ ในที่นี้หากมีภาษาต่างประเทศประเทศปนอยู่ในเอกสารจะทำการแปลงเป็นคำสะกดภาษาไทยเพื่อเปรียบเทียบกับภาษาไทยที่อยู่ใกล้เคียงกับคำนั้น ในที่นี้จะเน้นไปที่ภาษาอังกฤษเท่านั้น

ตัวอย่างเช่น

“ซัมไคน์ออฟวันเดอฟูล (Some kind of wonderful)” หากตัดคำตามกฎและพจนานุกรม [1] จะได้ว่า

ซัม-ไคน์-ออ-ฟ-วัน-เดอ-ฟู-ล เนื่องจากคำว่า ออ วัน และฟู ปรากฏอยู่ในพจนานุกรมทำให้การตัดคำไม่ถูกต้อง

หากตัดคำโดยการใช้การแปลงคำจากภาษาอังกฤษเป็นคำอ่านภาษาไทยจะได้ว่า

ซัม-ไคน์-ออฟ-วันเดอฟูล ซึ่งทำให้ได้คำที่ถูกต้อง

พจนานุกรมที่เก็บคำศัพท์นั้นจะเก็บคำที่ซ้ำๆ กันเอาไว้เมื่อคำหนึ่งคำ(S) ประกอบด้วยคำย่อย (S_i) ซึ่งคำย่อยก็เป็นคำในพจนานุกรม จะได้ว่า

$$S_i = S_{i1} + S_{i2} + S_{i3} + \ldots + S_{iN}$$

ยกตัวอย่างเช่น ระบอบประชาธิปไตย จะจัดเก็บดังนี้

$$W_i = A_1 + A_2 + A_3 + \ldots + A_N \text{ เมื่อ } A \in \{ W , C\}$$

W คือคำที่ตัดได้จากประโยคหรืออนุประโยค S

A คือคำย่อยที่มีอยู่ในพจนานุกรมที่ประกอบเป็นคำ W หรือตัวอักษรที่ไม่มีอยู่ในพจนานุกรม

C คือตัวอักษรในประโยคหรืออนุประโยค S

เช่น ระบอบ คือ [code1] , ประชา คือ [code2]
code1 คือรหัสแทนคำว่า “ระบอบ”
code2 คือรหัสแทนคำว่า “ประชา”
ประชาธิปไตย คือ [code2]+ธิปไตย
ระบอบประชาธิปไตย คือ [code1] +[code2]+ธิปไตยการจัดเก็บเป็นรหัสแทนเพื่อลดขนาดของพจนานุกรมและเพิ่มประสิทธิภาพสำหรับการตัดคำ

4.3 การวิเคราะห์คำที่ไม่มีอยู่ในพจนานุกรมโดยเทียบกฎความเป็นไปได้ที่จะเป็นภาษาต่างประเทศหรือคำเฉพาะ

การวิเคราะห์คำเฉพาะจะทำในขั้นตอนสุดท้ายกรณีที่คำที่นำมาไม่ตรงกับพจนานุกรมหรือไม่มีความเป็นไปได้ในการที่จะเป็นภาษาต่างประเทศ ซึ่งพจนานุกรมศัพท์เฉพาะนี้สามารถเพิ่มเติมได้ตลอดเวลาเพื่อให้เหมาะสมกับเอกสารที่ต้องการนำมาตัดคำ ตัวอย่างคำเฉพาะเช่น ชื่อคน ชื่อสถานที่ คำที่เหมาะ

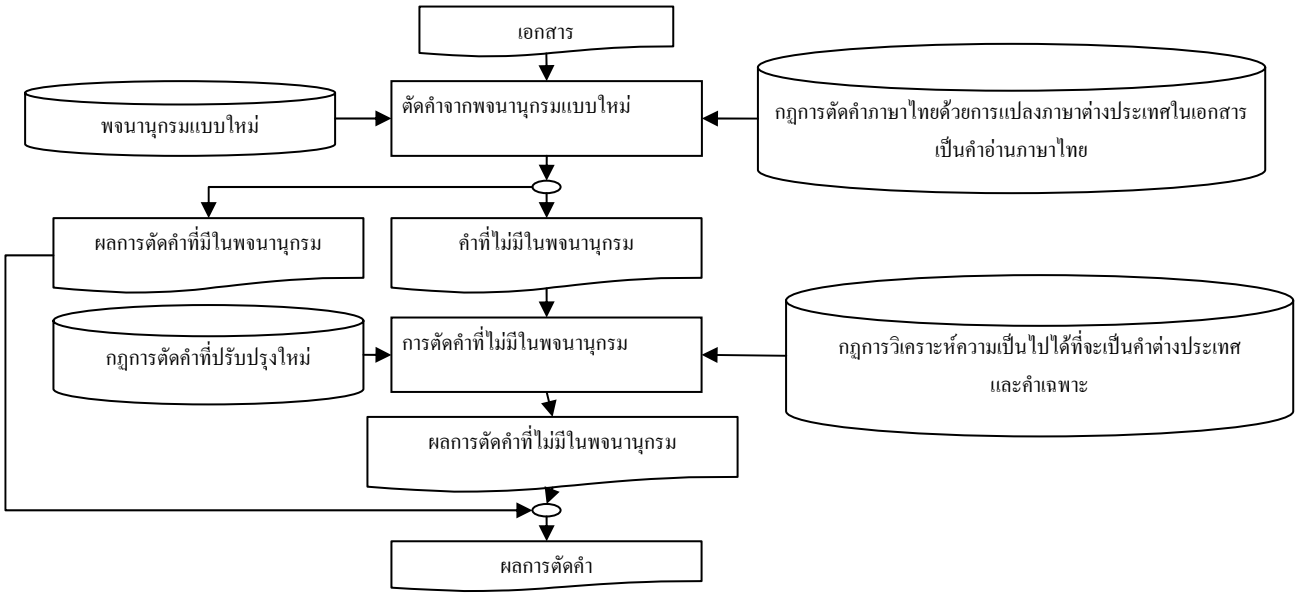
4.4 การตัดคำในส่วนสุดท้าย

ในส่วนสุดท้าย อนุประโยคหรือส่วนของอนุประโยคใดไม่ตรงกับข้อที่กล่าวมาข้างต้นจะทำการตัดคำโดยใช้กฎการตัดพยางค์แทน ภาพรวมของขั้นตอนวิธีการตัดคำภาษาไทยด้วยวิธีปรับปรุงกฎและพจนานุกรมแบบใหม่ได้แสดงไว้ในรูปที่ 1

5. การทดลองและวิเคราะห์ผล

การตัดคำโดยใช้เอกสารจากแหล่งต่างๆ ได้แก่หนังสือพิมพ์ บทความทางวิชาการและวารสารซึ่งเอกสารจากหนังสือพิมพ์จะแยกย่อยออกเป็นข่าวเศรษฐกิจ ข่าวต่างประเทศ-กีฬา และข่าวอื่นๆ ซึ่งแต่ละประเภทมีขนาดเอกสารไม่ต่างกันมากนัก

ต่อจำนวนคำที่ตัดได้ โดยผลการตัดคำจะแสดงไว้ในตารางที่ 1 โดยเปรียบเทียบกับวิธีการตัดคำของ [1] โดยใช้เอกสารสำหรับตัดคำชุดเดียวกัน



รูปที่ 1 ภาพรวมของขั้นตอนวิธีการตัดคำภาษาไทยด้วยวิธีปรับปรุงกฎและพจนานุกรมแบบใหม่

ตารางที่ 1 แสดงผลการตัดคำด้วยการใช้กฎที่ปรับปรุงและพจนานุกรมแบบใหม่ร่วมกัน

เอกสาร	การตัดคำแบบเดิม		การตัดคำที่นำเสนอ	
	ระดับพยางค์	ระดับคำ	ระดับพยางค์	ระดับคำ
ข่าวเศรษฐกิจ	95.34%	83.50%	94.67%	85.04%
ข่าวต่างประเทศและกีฬา	90.14%	79.25%	92.00%	81.07%

ข่าวอื่นๆ	93.25%	87.01%	97.28%	98.05%
บทความวิชาการ	89.40%	80.21%	88.25%	87.44%
วารสารทั่วไป	92.67%	81.35%	93.52%	93.75%

จากตารางที่ 1 จะเห็นได้ว่าการใช้วิธีการตัดคำที่นำเสนอจะได้ค่าผลลัพธ์โดยเฉลี่ยสูงขึ้นกว่าการตัดคำด้วยกฎเดิมโดยที่การตัดคำในข่าวทั่วไปและวารสารในระดับคำให้ผลที่ถูกต้องสูงขึ้นถึง 11.04 และ 12.40 เปอร์เซ็นต์

ส่วนการตัดคำระดับพยางค์นั้นความถูกต้องใกล้เคียงกับกฎเดิมเนื่องจากพยางค์ส่วนใหญ่สะกดตามหลักอักขรวิธีมีเพียงส่วนน้อยเท่านั้น ในส่วนของข่าวเศรษฐกิจและบทความวิชาการที่ได้เปอร์เซ็นต์ความถูกต้องน้อยกว่าการตัดคำแบบเดิม เนื่องจากทั้งสองเอกสารมีคำเฉพาะเป็นจำนวนมาก การเพิ่มคำเฉพาะลงในพจนานุกรมเพื่อเพิ่มความถูกต้องสามารถทำได้แต่หากคำเฉพาะมีบางส่วนที่ปรากฏอยู่ในพจนานุกรมการตัดคำจะใช้วิธีการเทียบหาคำที่มีความยาวที่สุด การตัดคำในระดับพยางค์จะเน้นที่รูปแบบการสะกดเป็นหลัก แต่ในระดับคำ การตัดคำทำได้ถูกต้องมากขึ้น การตัดคำที่ได้ผลดีที่สุดคือวารสารทั่วไปเนื่องจากวารสารที่นำมาตัดคำมีภาษาอังกฤษและคำอ่านที่สะกดด้วยภาษาไทยปนอยู่หลายคำซึ่งเหมาะสมต่อการใช้การตัดคำโดยใช้วิธีการที่นำเสนอ

6. สรุป

การวิจัยนี้ได้ทำการตัดคำภาษาไทยโดยการปรับปรุงกฎการตัดคำเพื่อมุ่งเน้นแก้ปัญหาความซับซ้อนของคำโดยเฉพาะคำที่มาจากภาษาต่างประเทศซึ่งไม่สอดคล้องกับอักขรวิธีและคำหรือบางส่วนของคำที่ไม่พบในพจนานุกรมที่ทำให้การตัดคำไม่ถูกต้องและเพื่อให้การตัดคำยืดหยุ่นมากขึ้น โดยวิธีการแก้ปัญหาที่นำเสนอจะได้ผลเฉลี่ยการตัดคำภาษาไทยในระดับพยางค์ดีขึ้นเป็น 93.14 เปอร์เซ็นต์และในระดับคำดีขึ้นเป็น 89.07 เปอร์เซ็นต์ แต่การตัดคำพ้องรูปยังทำได้ไม่ดีนักซึ่งการตัดคำควรใช้หน้าที่ของคำและโครงสร้างของประโยคเข้ามาร่วมพิจารณาด้วย งานต่อไปคือการให้โปรแกรมเพิ่มคำเฉพาะที่ไม่พบในพจนานุกรมในตอนแรกโดยอัตโนมัติเพื่อให้การตัดคำในเอกสารแบบเดียวกันเป็นไปได้ดีขึ้น

เอกสารอ้างอิง

21. ดวงแก้ว สวามิภักดิ์, “การสร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกส์”, มหาวิทยาลัยธรรมศาสตร์,2533.

22. บรรจบ พันธุเมธา, ลักษณะภาษาไทย, ระบบเสียงภาษาไทย กรุงเทพ , สำนักพิมพ์มหาวิทยาลัยรามคำแหง, 2540, หน้า 1-45.

23. พิสิทธิ์ พรมจันทร์, “การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรมแยกคำภาษาไทย”, จุฬาลงกรณ์มหาวิทยาลัย , 2540,หน้า 18-28.

24. ไพศาล เจริญพรสวัสดิ์, “การตัดคำภาษาไทยโดยใช้คุณลักษณะ, จุฬาลงกรณ์มหาวิทยาลัย”, 2541.

25. ปโยธร อุราธรรมกุล และ กานดา รุณนะพงศา, “การปรับปรุงการตัดคำในเอกสารไทย”, NECSEC ครั้งที่ 1 , 2548, หน้า 41-45.

26. ยืน ภู่วรวรรณ และ วิวรรธ อิ่มอารมณ์, “การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี”, รายงานการประชุมวิชาการวิศวกรรมไฟฟ้าครั้งที่ 9, 2529.

27. C. Kooptiwoot, “Segmentation of Ambiguous Thai Words by Inductive Logic Programming”, Chulalongkorn, 1999.

28. D. D. Plamer, “A Trainable Rule-based Algorithm for Word Segmentation”, http://citeseer.ist.psu.edu/palmer97trainable.html.

29. P. Charoenpornsawat, B. Kijsirkul, S. Meknavin, “Feature-based Thai Unknown Word Boundary Identification Using Winnow”, Chulalongkorn, 1998.

30. S. Charnyapornpong , “A Thai Syllable Separation Algorithm”, Master Thesis, Asian Institute of Technology, 1983.

31. T. Pongthai, V. Sornlertlamvanish, “Grapheme to Phoneme for Thai”, NECTEC, http://www.afnlp.org/nlprs2001/html/toc.html.

32. Y. Thairatananond, “Towords the design of a Thai text syllable analyzer”, Master Thesis, Asian Institute of Technology.

ภาคผนวก ข

ตัวอย่างการตัดคำ

1. ตัวอย่างการตัดคำจากบทความทางวิชาการ

เจเป็ก แอลเอ็ส (JPEG-LS – Joint Photographic Expert Group Lossless) เป็น มาตรฐาน การ **บีบ อัด** ภาพ แบบ ไม่ สูญเสีย สำหรับ **ภาพ เทา** ต่อเนื่อง ขนาด ของ จุด ที่ 2–16 บิต ต่อ จุด ซึ่ง

กำหนด มาตรฐานโดย คณะกรรมการ เจเป็ก (JPEG -Joint Photographic Expert Group) ใน ปี 2540 เจเป็ก แอลเอ็ส (JPEG-LS) ใช้ เทคนิค การ **บีบ อัด** ภาพ แบบ การ ทำนาย จุด (Predictive Coding) ร่วม กับ การ **บีบ อัด** ภาพ เชิง สถิติ (Probabilistic Coding) สำหรับ การ **เข้า รหัส** จุด ทั่วไป และ การ ใช้ เทคนิค การ **เข้า รหัส** แบบ รัน เลนท์ (Run-Length Coding) สำหรับ บริเวณ จุด ที่ มี ข่าวสาร ของ ข้อมูล ต่ำ (Low Entropy) หรือ ข้อมูล บริเวณ นั้น ซ้ำ กัน โดย มี จุดเด่น ทาง ด้าน กระบวนการ ทำงาน ไม่ ซับซ้อน และ ความเร็ว ใน การ ทำงาน สูง เมื่อ เปรียบเทียบ กับ เทคนิค การ บีบ อัด ภาพ ที่ ใช้ การ คำนวณ ทาง คณิตศาสตร์ อื่น ๆ แต่ อัตรา การ บีบ อัด ภาพ ใน ลักษณะ ที่ สร้าง จาก คอมพิวเตอร์ (Syntethic) มี อัตรา การ **บีบ อัด** ที่ ต่ำ กว่า เทคนิค การ **บีบ อัด** ภาพ ที่ ไม่ สนใจ ข่าวสาร ของ ข้อมูล เช่น กิฟ (GIF – Graphic Interchange Format) ที่ ใช้ เทคนิค การ พจนานุกรม ข้อมูล แบบ เคลื่อนที่ (Dynamic Dictionary Technique) เป็นต้น เทคนิค การ **บีบ อัด** ข้อมูล แบบ พจนานุกรม ข้อมูล เป็น กระบวนการ แทน ข้อมูล ที่ กำลัง **เข้า รหัส** ด้วย สัญลักษณ์ อ้างอิง กลับ ไป ยัง ข้อมูล ที่ ซ้ำ กัน ที่ **เข้า รหัส** ผ่าน มา โดย ไม่มี การ คำนวณ ข่าวสาร ของ ข้อมูล และ สามารถ ทำ การ คำนวณ ค่า ก่อน การ ทำงาน (Pre-Processing) เพื่อ ให้ ค่า ที่ กำลัง ทำ การ **เข้า รหัส** มี ความ เหมาะสม กับ การ **บีบ อัด** ข้อมูล

คำที่ตัดไม่ถูกต้อง

บีบอัด

ภาพเทา

เข้ารหัส

แต่พบว่าหากเพิ่มคำเหล่านี้ลงในพจนานุกรมแล้วสามารถตัดคำได้อย่างถูกต้อง ส่วนคำว่า “เจเป็ก แอลเอ็ส” นั้นสามารถตัดคำได้จากภาษาอังกฤษที่ปรากฏอยู่ในเอกสาร หากใช้วิธีของดวงแก้ว[1] จะต้องทำการเพิ่มคำลงในพจนานุกรมเสียก่อน มิฉะนั้นจะตัดคำได้ว่า “เจ เป็ก แอล เอ็ส”

2. ตัวอย่างการตัดคำจากข่าวเศรษฐกิจ

บาง จาก เอา จริง ตั้ง รง. ไบ โอ ดีเซล หวัง ครอง ตลาด

บาง จาก เจรจา พันธมิตร ลุย ตั้ง โรงงาน ไบ โอ ดีเซล เอง หลัง ประเมิน มี ความต้องการ ใช้ สูง พร้อม วางแผน พัฒนา ยกระดับ สถานี บริการ พลังงาน ทดแทน

นาย อนุสรณ์ แสง นิ่มนวล กรรมการ ผู้จัดการ ใหญ่ บริษัท บางจาก ปิโตรเลียม จำกัด (มหาชน) กล่าวว่า บริษัท มี แผนที่ จะ ลงทุน สร้าง โรงงาน ผลิต ไบโอ ดีเซล เพื่อ ใช้ เป็น พลังงาน ทาง เลือก ของ ผู้บริโภค ใน ยุค ที่ น้ำมันเชื้อเพลิง มี ราคา เพิ่ม สูงขึ้น อย่าง ต่อเนื่อง โดย ต้องการ สร้าง เป็น โรงงาน ที่ มี กำลัง การผลิต ประมาณ 3-4 แสน ลิตร ต่อ วัน และ เป็น การ ผลิต แบบ ครบวงจร ขณะ นี้ อยู่ ใน ช่วง การ หา พันธมิตร เข้า ร่วมทุน

" หลังจาก ที่ บริษัท ได้ ทดลอง จำนวน น้ำมัน บี 5 ซึ่ง ผสม ไบโอ ดีเซล 5% ใน สถานี บริการ น้ำมัน 12 แห่ง พบว่า ผู้บริโภค ยัง มี ความต้องการ ใช้ ไบโอดีเซล เป็นจำนวนมาก ขณะ ที่ ไบ โอ ดีเซล ที่ ผลิต ใน ประเทศ ยัง มี ปริมาณ ไม่ มาก นัก ดังนั้น จึง เชื่อ ว่า เมื่อ ผลิต ขึ้น มา เอง แล้ว จะ สามารถ จำหน่าย ได้ หมด อย่าง แน่นอน " นาย อนุสรณ์ กล่าว

คำที่ตัดไม่ถูกต้อง
แผนที่
บางจาก เนื่องจากคำว่า บางจาก และไบโอดีเซล เป็นคำเฉพาะ สามารถแก้ปัญหาได้ด้วย การเพิ่มคำลงในพจนานุกรม