Cover Page

1. *Paper title* การตัดคำภาษาไทยด้วยวิธีปรับปรุงกฎและพจนานุกรมแบบใหม่

   Improved Rule-Based and New Dictionary for Thai Word Segmentation

2. *List of authors* ปโยธร อุราธรรมกุล และ กานดา รุณนะพงศา

   Payothorn Urathamakun and Kanda Runapongsa

3. *Email:* payothorn@gmail.com , krunapon@kku.ac.th

4. *Postal address* 123 Mittraphab Rd. Computer Engineering Department Faculty of Engineering

   Khon Kaen University Khon Kaen 40002

5. *Country* Thailand

6. *Tel.* 043-362160

7. *Fax.* 043-362160

8. *Paper keywords* Word segmentation, Rule-based segmentation, Dictionary

9. *Paper category* Natural Language

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: 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 การตัดคำ, กฎการตัดคำ, พจนานุกรม

1. บทนำ

การตัดคำได้รับการพัฒนาขึ้นมาโดยใช้วิธีการต่างๆ ที่ต่างกัน เนื่องจากการตัดคำเป็นกระบวนการพื้นฐานของการประมวลผลภาษาธรรมชาติ เช่น การวิเคราะห์เสียงพูด การตัดคำภาษาไทยเองก็เช่นกัน ได้มีผู้คิดค้นวิธีที่จะแยกคำแต่ละคำออกจากประโยคซึ่งมีการเขียนติดกันไปอย่างต่อเนื่องทั้ง

ประโยค ในงานวิจัยนี้จะกล่าวถึงการตัดคำโดยอาศัยอักขระวิธีเป็นหลักการพื้นฐานการประสมคำ

1.1 ลักษณะของภาษาไทย

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

1.2 อักขระวิธี

คำในภาษาไทยเกิดจากส่วนต่างๆ ของอักษรไทยประกอบกันอย่างน้อยสามส่วน ได้แก่ ส่วนพยัญชนะ สระ และวรรณยุกต์ ยกตัวอย่างคำว่า กา เกิดจากพยัญชนะ ก ประสมกับสระ อา ไม่มีรูปวรรณยุกต์แต่มีเสียงสามัญ พยัญชนะของไทยถูกแบ่งออกเป็นอักษรสามหมู่ที่เรียกว่าไตรยางค์ ได้แก่ อักษรสูง อักษรกลาง และอักษรต่ำ สระก็ถูกจัดเป็นประเภท สระเดี่ยว สระประสม วรรณยุกต์ก็มี 4 รูป 5 เสียง การจะทำให้เกิดเสียงและความหมายต้องเกิดจากกฎเกณฑ์ที่มีอยู่

งานวิจัยนี้เสนอการตัดคำภาษาไทยโดยใช้การปรับปรุงกฎเพื่อเพิ่มความยืดหยุ่นให้กับการสะกดคำและการพัฒนาพจนานุกรมเพื่อเพิ่มประสิทธิภาพสำหรับการตัดคำ ในส่วนที่ 2 และ 3 จะกล่าวถึงวิธีและขั้นตอนต่างๆ ที่ใช้สำหรับตัดคำ ส่วนที่ 4 จะกล่าวถึงขั้นตอนการตัดคำที่นำเสนอตั้งแต่เริ่มต้น 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 กลุ่มย่อยได้แก่กฎการหาขอบเขตหน้า (Front boundary recognition rule) การหาขอบเขตหลัง (Tail boundary recognition rule) และแต่ละกฎยังแบ่งย่อยออกเป็นสองกลุ่มย่อย 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 และตัวสะกดต่างจากอักขระวิธีของไทย ซึ่งจะทำให้เกิดข้อผิดพลาดในการตัดคำในเอกสารที่มีทั้งภาษาไทยและภาษาอังกฤษปนกัน โดยเฉพาะคำภาษาไทยที่มาจากการสะกดคำอ่านภาษาอังกฤษด้วยภาษาไทย อย่างเช่น เว็บเซอร์วิสเป็นโปรแกรมที่สามารถติดต่อได้โดยตรงกับอีกโปรแกรมหนึ่ง

2.2 การตัดคำ

การตัดคำนิยมใช้พจนานุกรมเข้ามาช่วย ได้แก่งานวิจัยของยืน ภู่วรวรรณและวิวรรธน์ อิ่มอารมณ์ [6] เนื่องจากพจนานุกรมสามารถตัดคำได้ชัดเจนกว่าการตัดพยางค์ เพราะคำอาจเกิดจากการมารวมกันของหลายพยางค์ แต่ขอบเขตของคำยังคงซับซ้อนและกำกวม ในงานวิจัยได้ใช้เทคนิคที่เรียกว่าวิธีการย้อนกลับ (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 หมู่ที่เรียกว่าไตรยางค์ ได้แก่ อักษรสูง อักษรกลาง และและอักษรต่ำ 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 กิน ข้าว

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

ประโยคหรืออนุประโยคที่มีการประสมกันระหว่างอักษรไทยและอักษรอื่นในประโยคหรืออนุประโยคเดียวกันให้ถือเป็นหนึ่งคำ จะพบลักษณะนี้ได้ในรหัสหรือชื่อเฉพาะบางคำ

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

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

L คือจำนวนตัวอักษรทั้งหมดของอนุประโยค S_i โดยที่ $L \geq 1$

N คือจำนวนคำย่อยซึ่งประกอบเป็นอนุประโยค S_i โดยที่ $1 \leq N \leq L$

C คือตัวอักษรในประโยคหรืออนุประโยค S

W คือคำที่ตัดได้จากประโยคหรืออนุประโยค S

ถ้าหาก $C_{ij} \notin T$ เมื่อ T คือชุดของตัวอักษรภาษาไทยให้ถือว่า

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

เช่น ก3 , 2/5ส , ก.3 , 2จ-3 เป็นต้น

3.3 พจนานุกรม

พจนานุกรมที่เก็บคำศัพท์นั้นจะเก็บคำที่ซ้ำๆ กันเอาไว้เ มื่ อคำหนึ่งคำ(S) ประกอบด้วยคำย่อย (S_i) ซึ่งคำย่อยก็เป็นคำในพจนานุกรม จะได้ว่า

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

ยกตัวอย่างเช่น ระบอบประชาธิปไตย จะจัดเก็บดังนี้

$W_i = A_1 + A_2 + A_3 + \ldots + A_N$

$A \in \{ W , C\}$

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

3.4 คำที่ไม่พบในพจนานุกรม

คำที่ไม่พบในพจนานุกรมมักเป็นคำเฉพาะเช่นชื่อคนหรือสถานที่ 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

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

<u>ประเภทที่ 2</u> ประกอบด้วยอักษรไทยซึ่งอยู่ติดกับตัวเลข หรืออักษรแบ่งวรรค - , / หรือ อักษรต่างประเทศ หรืออักษรพิเศษอื่นๆ จะทำการแยกอักษรแบ่งวรรคออกจากกัน ยกเว้นตัวเลขเครื่องหมาย – และ / จะยังคงเขียนติดกับอักษรไทยไว้เช่นนั้นและถือเป็นคำ 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 }

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 หรือคำที่มีความยาวหลายพยางค์หรือคำที่มีการสะกดตรงตามอักขระวิธี

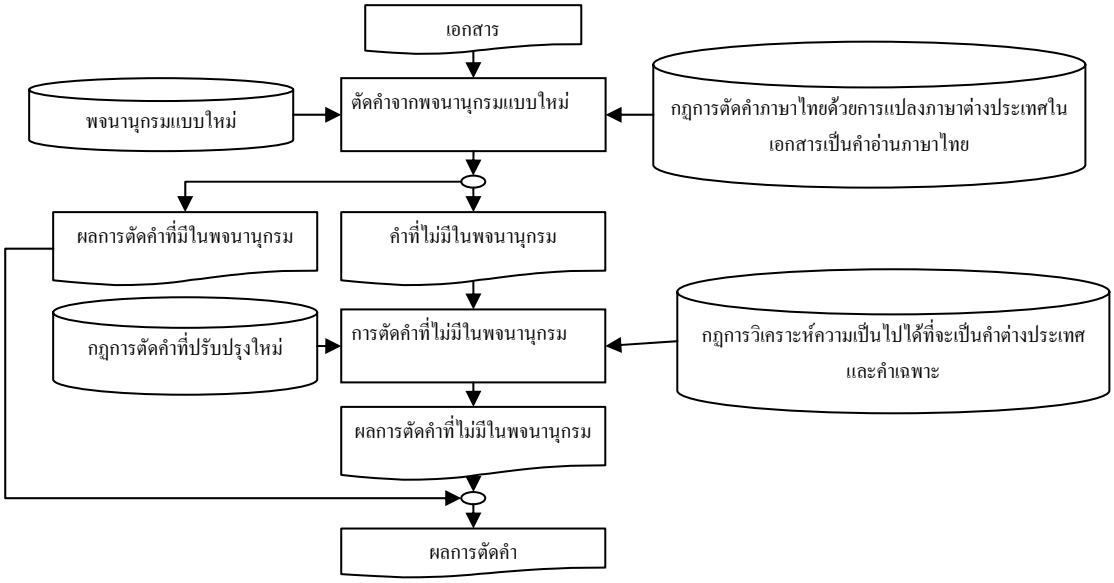
ส่วนที่สอง ไม่พบตามพจนานุกรม โดยปรกติส่วนนี้หากเป็นคำเดี่ยวๆ จะนำไปพิจารณากับคำรอบข้างซึ่งมีความเป็นไปได้ที่จะเป็นคำเดียวกันโดยอาศัยกฎการวิเคราะห์ความเป็นไปได้ที่จะเป็นภาษาต่างประเทศหรือคำเฉพาะ คำเฉพาะบางส่วนจะถูกเก็บอยู่ในพจนานุกรมคำเฉพาะและสามารถปรับปรุงได้เพื่อความเหมาะสมกับเอกสารแต่ละประเภท คำเฉพาะที่เหมาะสมได้แก่คำที่มาจากบาลี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

4.3 การวิเคราะห์คำที่ไม่มีอยู่ในพจนานุกรมโดยเทียบกฎความเป็นไปได้ที่จะเป็นภาษาต่างประเทศหรือคำเฉพาะ

การวิเคราะห์คำเฉพาะจะทำในขั้นตอนสุดท้ายกรณีที่คำที่นำมาไม่ตรงกับพจนานุกรมหรือไม่มีความเป็นไปได้ในการที่จะเป็นภาษาต่างประเทศ ซึ่งพจนานุกรมศัพท์เฉพาะนี้สามารถเพิ่มเติมได้ตลอดเวลาเพื่อให้เหมาะสมกับเอกสารที่ต้องการนำมาตัดคำ ตัวอย่างคำเฉพาะเช่น ชื่อคน ชื่อสถานที่ คำที่เหมาะ

4.4 การตัดคำในส่วนสุดท้าย

ในส่วนสุดท้าย อนุประโยคหรือส่วนของอนุประโยคใดไม่ตรงกับข้อที่กล่าวมาข้างต้นจะทำการตัดคำโดยใช้กฎการตัดพยางค์แทน



รูปที่ 1 ภาพรวมของขั้นตอนวิธีการตัดคำ

5. การทดลองและวิเคราะห์ผล

การตัดคำโดยใช้เอกสารจากแหล่งต่างๆ ได้แก่ หนังสือพิมพ์ บทความทางวิชาการและวารสารซึ่งเอกสารจากหนังสือพิมพ์จะแยกย่อยออกเป็นข่าวเศรษฐกิจ ข่าวต่างประเทศ-กีฬา และข่าวอื่นๆ ขนาดของเอกสารทั้งหมดคือ 1,793 กิโลไบต์ ซึ่งแต่ละประเภทมีขนาดเอกสารไม่ต่างกันมากนัก

โดยการวัดผลจะวัดจากจำนวนคำที่ได้ถูกต้องทั้งหมดต่อจำนวนคำที่ตัดได้ โดยผลการตัดคำจะแสดงไว้ในตารางที่ 1 โดยเปรียบเทียบกับวิธีการตัดคำของ [1] โดยใช้เอกสารสำหรับตัดคำชุดเดียวกัน

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 ในส่วนของข่าวเศรษฐกิจและบทความวิชาการที่ได้เปอร์เซ็นต์ความถูกต้องน้อยกว่าการตัดคำแบบเดิมเนื่องจากทั้งสองเอกสารมีคำเฉพาะเป็นจำนวนมาก การเพิ่มคำเฉพาะลงในพจนานุกรมเพื่อเพิ่มความถูกต้องสามารถทำได้แต่หากคำเฉพาะมีบางส่วนที่ปรากฏอยู่ในพจนานุกรมการตัดคำจะใช้วิธีการเทียบหาคำที่มีความยาวที่สุด การตัดคำในระดับพยางค์จะเน้นที่รูปแบบการสะกดเป็นหลัก แต่ในระดับคำ การตัดคำทำได้ถูกต้องมากขึ้น การตัดคำที่ได้ผลดีที่สุดคือวารสารทั่วไปเนื่องจากวารสารที่นำมาตัดคำมีภาษาอังกฤษและคำอ่านที่สะกดด้วยภาษาไทยปนอยู่หลายคำซึ่งเหมาะสมต่อการใช้การตัดคำโดยใช้วิธีการที่นำเสนอ

6. สรุป

การวิจัยนี้ได้ทำการตัดคำภาษาไทยโดยการปรับปรุงกฎการตัดคำเพื่อมุ่งเน้นแก้ปัญหาความซับซ้อนของคำโดยเฉพาะคำที่มาจากภาษาต่างประเทศซึ่งไม่สอดคล้องกับอักขรวิธีและคำหรือบางส่วนของคำที่ไม่พบในพจนานุกรมที่ทำให้การตัดคำไม่ถูกต้องและเพื่อให้การตัดคำยืดหยุ่นมากขึ้นโดยวิธีการแก้ปัญหาที่นำเสนอจะได้ผลเฉลี่ยการตัดคำภาษาไทยในระดับพยางค์ดีขึ้นเป็น 93.14 เปอร์เซ็นต์และในระดับคำดีขึ้นเป็น 89.07 เปอร์เซ็นต์ แต่การตัดคำพ้องรูปยังทำได้ไม่ดีนักซึ่งการตัดคำควรใช้หน้าที่ของคำและโครงสร้างของประโยคเข้ามาร่วมพิจารณาด้วย งานต่อไปคือการให้โปรแกรมเพิ่มคำเฉพาะที่ไม่พบในพจนานุกรมในตอนแรกโดยอัตโนมัติเพื่อให้การตัดคำในเอกสารแบบเดียวกันเป็นไปได้ดีขึ้น

เอกสารอ้างอิง

[1] ดวงแก้ว สวามิภักดิ์, “การสร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกส์” : มหาวิทยาลัยธรรมศาสตร์,2533.

[2] บรรจบ พันธุเมธา, ลักษณะภาษาไทย, ระบบเสียงภาษาไทย กรุงเทพ, สำนักพิมพ์มหาวิทยาลัยรามคำแหง หน้า 1-45, 2540.

[3] พิสิทธิ์ พรมจันทร์, “การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรมแยกคำภาษาไทย”, จุฬาลงกรณ์มหาวิทยาลัย หน้า 18-28.

[4] ไพศาล เจริญพรสวัสดิ์, “การตัดคำภาษาไทยโดยใช้คุณลักษณะ, จุฬาลงกรณ์มหาวิทยาลัย”, 2541.

[5] ปโยธร อุราธรรมกุล และ กานดา รุณนะพงศา, “การปรับปรุงการตัดคำในเอกสารไทย”, หน้า 41-45 .NECSEC ครั้งที่ 1 ,2005.

[6] ยืน ภู่วรวรรณ และ วิวรรธ อิ่มอารมณ์, “การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี”. รายงานการประชุมวิชาการวิศวกรรมไฟฟ้าครั้งที่ 9, 2529.

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

[8] D. D. Plamer. “A Trainable Rule-based Algorithm for Word Segmentation”.

[9] P. Charoenpornsawat, B. Kijsirkul, S. Meknavin. “Feature-based Thai Unknown Word Boundary Identification Using WinnoW”, Chulalongkorn. 1998.

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

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

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