การเปรียบเทียบวิธีการแบ่งแยกคำภาษาไทยด้วยโครงสร้างการเขียนกับโครงสร้างพยางค์

The Comparison of Thai Word Segmentation with Thai Writing Structures and Syllable Structures

วุฒิชัย วิเชียรไชย[1]

Vuttichai Vichianchai[1]

Received: 9 October 2013; Accepted: 24 December 2013

บทคัดย่อ

บทความวิจัยนี้ได้เสนอวิธีการแบ่งแยกคำภาษาไทยโดยเทียบกับโครงสร้างการเขียนของภาษาไทยและอัลกอริทึมการแบ่งแยกคำภาษาไทยโดยโครงสร้างพยางค์ เพื่อศึกษาและเปรียบเทียบวิธีการประมวลผลของการแบ่งแยกคำภาษาไทยและประสิทธิภาพความถูกต้องของอัลกอริทึม เอกสารที่ใช้ในการทดสอบประสิทธิภาพของการแบ่งแยกคำจะใช้ข้อมูลที่ไม่มีรูปภาพ ตัวเลข สมการทางคณิตศาสตร์และอักขระพิเศษนอกเหนืออักขระในภาษาไทย จากเอกสารที่นำมาทดสอบได้นำมาจากข่าวและบทความ จากเอกสาร 20 เอกสาร พบว่าประสิทธิภาพการแบ่งแยกคำภาษาไทยโดยเทียบกับโครงสร้างการเขียนของภาษาไทยมีความถูกต้องร้อยละ 98.75 และประสิทธิภาพการแบ่งแยกคำภาษาไทยโดยโครงสร้างพยางค์มีความถูกต้องร้อยละ 87.64

คำสำคัญ: การเปรียบเทียบการแบ่งแยกคำ การแบ่งแยกคำ โครงสร้างการเขียนภาษาไทย โครงสร้างพยางค์

Abstract

This research paper presents the comparison of Thai word segmentation between two different methods, Thai segmentation with writing structures and syllable structures. The objective is to study and compare the process of Thai word segmentation of the two methods. Documents used to test the performance of the divisions will use the information, no mathematical equations, pictures, numbers, and special characters other than the characters in Thai of 20 documents. Performance evaluation of the accuracy showed that the accuracy of Thai word segmentation with Thai writing structures and syllable structures is 98.75% and 87.64% respectively.

Keywords: comparison of Thai segmentation, word segmentation, Thai writing structures, syllable structures

บทนำ

การเขียนข้อความภาษาไทยเป็นการเขียนติดต่อกันโดยไม่มีเครื่องหมายวรรคตอน แต่ก็มีการแบ่งวรรคตอนเพื่อเป็นการแบ่งแยกประโยคหรือวลีด้วยช่องว่างบ้างเป็นบางคราว ซึ่งมีความแตกต่างกับประโยคภาษาอังกฤษที่มีเครื่องหมายวรรคตอนและการแบ่งแยกคำแต่ละคำด้วยช่องว่างอย่างชัดเจน ลักษณะเช่นนี้ทำให้การประมวลผลภาษาไทยด้วยคอมพิวเตอร์มีความจำเป็นที่จะหาวิธีการให้คอมพิวเตอร์มีความสามารถที่จะเรียนรู้ขอบเขตของคำ เพื่อแก้ปัญหาเรื่องการแบ่งแยกคำในการประมวลผลคำภาษาไทยให้ถูกต้องตามหลักภาษาศาสตร์ ซึ่งจากการศึกษางานวิจัยเกี่ยวกับการแบ่งแยกคำในภาษาไทยได้มีการพัฒนาติดต่อกันมาเป็นเวลานาน โดยสามารถแบ่งงานวิจัยในการแบ่งแยกคำภาษาไทยได้เป็นดังนี้ คือ วิธีการใช้กฎ (Rule base approach) วิธีการใช้อัลกอริทึม (Algorithm approach) วิธีการใช้พจนานุกรม (Dictionary base approach) และวิธีการใช้คลังข้อความ (Corpus based approach) ซึ่งทั้ง 4 วิธีนี้เป็นวิธีที่ให้ประสิทธิภาพการแบ่งแยกคำที่ค่อนข้างสูง แต่ก็ยังมีข้อบกพร่อง โดยวิธีการใช้กฎและใช้อัลกอริทึมทำให้

[1] *ผู้ช่วยศาสตราจารย์, คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยมหาสารคาม ตำบลขามเรียง อำเภอกันทรวิชัย จังหวัดมหาสารคาม 44150*

[1] *Assist. Prof., Faculty of Informatics Mahasarakham University, Khamriang Sub-District, Kantharawichai District, Mahasarakham Province 44150, Thailand.*

* *Corresponding author; vuttichai.v@gmail.com*

เสียเวลาในการเปรียบเทียบ วิธีการเทียบกับพจนานุกรมทำให้สิ้นเปลืองพื้นที่ในการจัดเก็บ และวิธีการใช้คลังข้อความทำให้ใช้เวลานานในการประมวลผล

จากปัญหาดังกล่าวผู้วิจัยจึงได้เสนอวิธีการแบ่งแยกคำภาษาไทยโดยใช้โครงสร้างการเขียนภาษาไทย เพื่อแก้ไขลดพื้นที่ในการจัดเก็บคำศัพท์ในพจนานุกรม และวิธีการแบ่งแยกคำภาษาไทยด้วยโครงสร้างพยางค์เพื่อลดการสิ้นเปลืองพื้นที่ในการจัดเก็บพจนานุกรม

โครงสร้างการเขียนภาษาไทย คือ ลำดับการเขียนจะต้องเขียนพยัญชนะหรือ ฤ หรือ ฦ หรือ สระอะ หรือ สระอา หรือ สระอำ หรือ สระเอ หรือ สระแอ หรือ สระโอ หรือ สระไอ หรือ สระใอ หรือ ไม้ยะมุก หรือ รอเก้า ก่อนเสมอ จากนั้นไม้หันอากาศ หรือ สระอิ หรือ สระอี หรือ สระอึ หรือ สระอือ หรือ ไม้ไต่คู้ หรือ หยาดน้ำค้าง และ สระอุ หรือ อู หรือ พินทุ และวรรณยุกต์ จะถูกเขียนเป็นลำดับถัดมา ตามลำดับ (ถ้ามี) โครงสร้างพยางค์ คือ ลำดับการเขียนพยางค์ในภาษาไทยจะต้องขึ้นต้นด้วยพยัญชนะและตามด้วยสระ รวมทั้งอาจจะมีพยัญชนะเป็นตัวสะกด

ยกตัวอย่างการแบ่งแยกคำและพยางค์ของคำว่า “ประเทศไทย” จะสามารถแบ่งแยกคำได้เป็น “ประเทศไทย” และแบ่งพยางค์ได้เป็น “ประ|เทศ|ไทย”

ดังนั้น บทความวิจัยนี้จึงได้ทำการศึกษาและเปรียบเทียบงานวิจัยทั้งสองวิธีที่ได้กล่าวมานี้ เพื่อให้เห็นข้อแตกต่างของกระบวนการและประสิทธิภาพในการแบ่งแยกคำและพยางค์

เอกสารและงานวิจัยที่เกี่ยวข้อง

- งานวิจัย[1] โดยการใช้กฎที่สร้างขึ้นจากหลักไวยากรณ์ภาษาไทยและมีการจัดเก็บพยางค์ต่าง ๆที่เป็นข้อยกเว้นไว้ในแฟ้มข้อมูล เนื่องจากมีบางพยางค์ไม่เป็นไปตามกฎที่สร้างขึ้น ลักษณะของกฎได้ทำการพิจารณาจากลักษณะของอักษรที่ปรากฏในพยางค์หรือคำ ซึ่งทำให้สามารถจัดและแบ่งหมวดหมู่ตัวอักษรภาษาไทยได้เป็นห้ากลุ่มใหญ่ ๆ คือ กลุ่มพยัญชนะ (Consonant) กลุ่มสระ (Vowel) กลุ่มวรรณยุกต์ (Tone mark) กลุ่มตัวเลข (Numeral) และกลุ่มอักขระพิเศษ (Special character)

- งานวิจัย[2] การแบ่งแยกคำไทยด้วยพจนานุกรม โดยมีเป้าหมายเพื่อที่จะเพิ่มประสิทธิภาพในด้านความเร็วของขั้นตอนวิธีในการแบ่งแยกคำและการลดขนาดของพจนานุกรม เนื่องจากเมื่อนำพจนานุกรมเข้ามาใช้ในการแบ่งแยกคำแล้วจะทำให้ความถูกต้องในการแบ่งแยกคำเพิ่มมากกว่าการแบ่งแยกคำโดยใช้กฎอย่างเดียว ดังนั้นงานวิจัยนี้จึงไม่ได้เน้นการเพิ่มประสิทธิภาพในด้านความถูกต้องมากนักเพราะถือว่าการแบ่งแยกคำ โดยใช้พจนานุกรมแบบเทียบคำที่ยาวที่สุด (Longest matching)[3] ให้ค่าความถูกต้องสูง

- งานวิจัย[4] ในงานวิจัยนี้ได้นำเรื่องสถิติเข้ามาใช้ในการแก้ปัญหาการแบ่งแยกคำ และการกำหนดหน้าที่ของคำหรือประเภทย่อยของคำ โดยมีการนำแบบจำลองไตรแกรมเข้ามาช่วยในการแบ่งแยกคำ

การแบ่งแยกคำโดยใช้แบบจำลองไตรแกรม คือ การแบ่งแยกคำโดยมีการนำเอาค่าสถิติ ซึ่งพิจารณาจากความต่อเนื่องของหน้าที่คำหรือประเภทย่อยของคำ ส่วนวิธีการเลือกแบบการแบ่งแยกคำที่ดีที่สุดนั้นทำโดยหาประโยคที่มีความน่าจะเป็นมากที่สุด โดยการหาความน่าจะเป็นของแต่ละประโยค สามารถคำนวณตามสมการที่ 1 ดังนี้

$$P(W) = \prod_{i=1}^{n} P(w_i, n) = \prod_{i=1}^{n} P(w_i \mid w_{i-1}, w_{i-2}) \quad (1)$$

จากสมการที่ 1 คือการคำนวณหาค่าความน่าจะเป็นของแต่ละประโยค โดย W เป็นประโยคที่ได้แบ่งแยกคำแล้ว และประโยค W ประกอบไปด้วยคำต่าง ๆซึ่ง $W = w_1 w_2 .. w_n$ โดย w_i เป็นคำศัพท์ และการคำนวณค่าความน่าจะเป็นของแต่ละประโยคจะมีข้อกำหนดว่า ความน่าจะเป็นของ w_i จะขึ้นกับ w_{i-1} และ w_{i-2} เท่านั้น แต่เนื่องจากการคำนวณนั้นจะต้องใช้คลังข้อความ (Corpus) ขนาดใหญ่มากและควรมีขนาดมากกว่า n^3 คำ โดยที่ n เป็นจำนวนคำที่เป็นไปได้ทั้งหมด สาเหตุที่ต้องใช้คลังข้อความขนาดใหญ่กว่า n^3 คำ เนื่องจากวิธีนี้จะต้องมีการนำค่าสถิติการเกิดของคำสามคำที่ติดกันมาใช้ในการคำนวณ ดังนั้นเพื่อให้มีสถิติของการเกิดของคำสามคำติดกันทุก ๆแบบ อย่างน้อยที่สุดจะต้องใช้จำนวนคำเท่ากับ n^3 คำ ซึ่งในความเป็นจริงเราไม่สามารถหาคลังข้อความขนาดดังกล่าวได้ จึงทำให้มีการประมาณจากสมการที่ 1 เป็นสมการที่ 2 แทนดังนี้

$$\prod_{i=1}^{n} P(w_i \mid w_{i-1}, w_{i-2}) = \prod_{i=1}^{n} (\lambda_1 * P(w_i) + \lambda_2 * P(w_i \mid w_{i-1}) + \lambda_3 * P(w_i \mid w_{i-1}, w_{i-2})) \quad (2)$$

จากสมการที่ 2 นี้จะเป็นการแก้ปัญหาเรื่องจำนวนข้อมูลที่นำมาใช้ไม่เพียงพอ โดยจะมีการนำค่าความน่าจะเป็น

ของไบแกรม (Bigram) และยูนิแกรม (Unigram) เข้ามาช่วยในการคำนวณด้วย และค่า $\gamma_1 + \gamma_2 + \gamma_3$ จะต้องมีค่าเท่ากับหนึ่ง ซึ่งการคำนวณนั้นก็จะใช้คลังข้อความเช่นกัน แต่ถ้าในคลังนั้นไม่มีการเก็บค่าสถิติการเกิดของคำสองคำติดกัน ก็จะกำหนดค่าของ γ_2 เท่ากับ 0 และกรณีที่ไม่มีการเก็บค่าสถิติการเกิดของคำสามคำติดกันก็จะกำหนดค่าของ γ_3 เท่ากับ 0 เช่นกัน โดยหลักการทำงานของการวิจัยนี้จะทำการแบ่งแยกคำจากประโยคที่เป็นข้อมูลเข้าที่เป็นไปได้หลายๆแบบแล้วทำการคำนวณหาค่าความน่าจะเป็นจากแบบผลลัพธ์ต่างๆทั้งหมด และนำแบบผลลัพธ์ที่มีค่าความน่าจะเป็นที่มากที่สุดมาเป็นผลลัพธ์ในการแบ่งแยกคำของประโยคนั้น

สรุป วิธีการนี้จะช่วยในการแบ่งแยกคำได้ดี แต่จะใช้เวลาในการประมวลผลนาน เนื่องจากต้องสร้างแบบผลลัพธ์การแบ่งแยกคำที่เป็นไปได้หลายๆแบบ แล้วคำนวณหาค่าความน่าจะเป็น ที่มีค่ามากที่สุดจากแบบผลลัพธ์ทั้งหมดของประโยคที่เป็นข้อมูลเข้า

จากปัญหาที่เกิดขึ้นกับงานวิจัยที่ได้กล่าวข้างต้น ผู้วิจัยจึงได้เสนอวิธีการแบ่งแยกคำภาษาไทยโดยใช้โครงสร้างการเขียนภาษาไทย เพื่อแก้ไขลดพื้นที่ในการจัดเก็บคำศัพท์ในพจนานุกรม และวิธีการแบ่งแยกคำภาษาไทยด้วยโครงสร้างพยางค์เพื่อลดการสิ้นเปลืองพื้นที่ในการจัดเก็บพจนานุกรม ซึ่งความรู้ใหม่ที่เกิดขึ้นในงานวิจัยทั้งสองคือการรู้จักโครงสร้างการเขียนภาษาไทยและโครงสร้างพยางค์

วัสดุอุปกรณ์และวิธีการศึกษา

ศึกษางานวิจัยวิธีการแบ่งแยกคำภาษาไทยด้วยวิธีเทียบกับโครงสร้างการเขียนภาษาไทย

- งานวิจัย[5] ได้เสนอวิธีการแบ่งแยกคำโดยใช้รูปแบบโครงสร้างการเขียนคำภาษาไทยที่ได้จากคำศัพท์ในพจนานุกรม ฉบับราชบัณฑิตยสถาน พุทธศักราช 2542 เพื่อลดคำศัพท์ในพจนานุกรม ทำให้ลดพื้นที่ในการจัดเก็บและการลดเวลาในการประมวลผลในการแบ่งแยกคำ ซึ่งประสิทธิภาพการแบ่งแยกคำมีความถูกต้องร้อยละ 99 โดยใช้ข้อความจากเอกสารข่าวหนังสือพิมพ์ บทความ พุทธศาสนา สารานุกรม กฎหมายและนิยาย การทดสอบประสิทธิภาพ สามารถแบ่งการทำงานออกเป็น 2 ขั้นตอน ดังนี้

(1) การหารูปแบบโครงสร้างการเขียนของภาษาไทย เป็นขั้นตอนที่ทำการแปลงคำจากพจนานุกรมเป็นรูปแบบตัวเลขตามระดับโครงสร้างการเขียนภาษาไทย 4 ระดับ[6,7] โดยได้นิยามไว้ในบทนำ ซึ่งลำดับการเขียนจะต้องเขียนระดับที่ 3 ก่อนเสมอ ส่วนระดับที่ 2 หรือระดับที่ 4 จะถูกเขียนเป็นลำดับถัดมา (ถ้ามี) และระดับที่ 1 จะถูกเขียนเป็นลำดับสุดท้าย (ถ้ามี) ยกตัวอย่างเช่น

Figure 1 Thai writing structures, the Thai word is “ลูกพี่”

Figure 1 คำว่า “ลูกพี่” จะถูกแปลงเป็นรูปแบบโครงสร้างการเขียนภาษาไทยเป็นตัวเลขของระดับการเขียน คือ 3-4-3-3-1 และตัดรูปแบบที่เหมือนกันออกจะทำให้จำนวนรูปแบบลดลงและนำไปใช้ในการเทียบเพื่อแบ่งแยกคำ

(2) การแบ่งแยกคำโดยเทียบกับรูปแบบ เมื่อได้รูปแบบการเขียนภาษาไทยแล้วก็จะนำรูปแบบเหล่านั้นมาใช้ในการแบ่งแยกคำ ดังขั้นตอนใน Figure 2

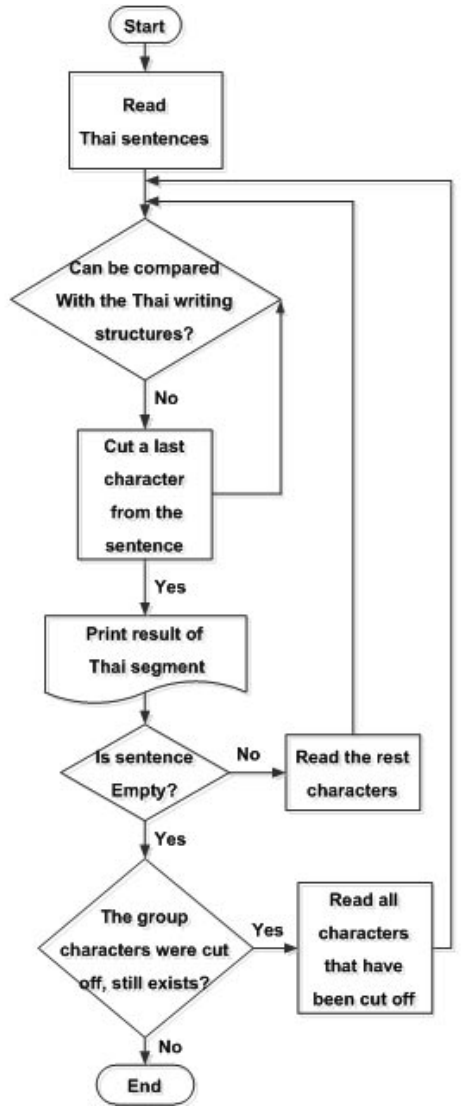

Figure 2 Thai word segmentation method with comparison Thai writing structures

จาก Figure 2 เป็นขั้นตอนการทำงานการแบ่งแยกคำภาษาไทยโดยเปรียบเทียบโครงสร้างการเขียนภาษาไทย และเพื่อเป็นการอธิบายให้เกิดความเข้าใจมากยิ่งขึ้น ผู้วิจัยขอยกตัวอย่างการแบ่งแยกคำของข้อความ “เขาไปไร่” โดยขั้นตอนดังกล่าว ซึ่งสามารถแสดงใน Table 1 ดังนี้

Table 1 Thai segmentation with comparison of Thai writing structures, the Thai sentence is “แบ่งพยางค์”

Sentence / Result	Thai Writing Structures format	Can be compared	(Cutting the last Character)
แบ่งพยางค์	แ-3่-3-3-3-3-า-3-1	No	แบ่งพยาง / (ค์)
แบ่งพยาง	แ-3่-3-3-3-3-า-3	No	แบ่งพยา / (งค์)
แบ่งพยา	แ-3่-3-3-3-3-า	No	แบ่งพ/ (ยางค์)
แบ่งพ	แ-3่-3-3	No	แบ่ง/ (พยางค์)
แบ่ง	แ-3่-3-3	Yes	พยางค์
พยางค์	3-3-3-า-3-1	Yes	

*** หมายเหตุ** เนื่องจากเป็นงานวิจัยในการแบ่งแยกคำภาษาไทย ดังนั้นรายละเอียดที่ยกตัวอย่างในตารางจึงมีความจำเป็นต้องใช้คำและข้อความภาษาไทย

ศึกษางานวิจัยอัลกอริทึมการแบ่งแยกคำภาษาไทยโดยโครงสร้างพยางค์

งานวิจัยการแบ่งแยกคำภาษาไทยด้วยโครงสร้างพยางค์ สามารถเขียนเป็นแผนภาพแสดงขั้นตอนการทำงานได้ดัง Figure 3

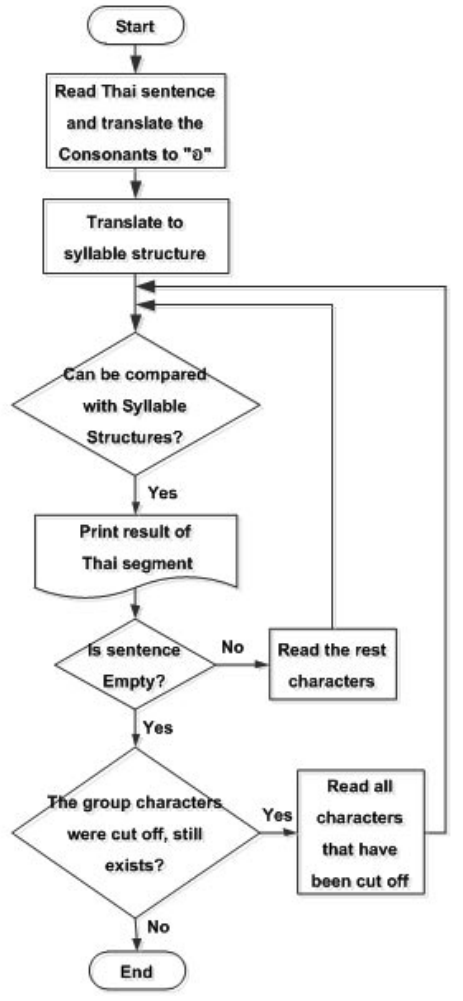

Figure 3 Thai word segmentation method with comparison syllable structures

ใน Figure 3 เป็นการอธิบายขั้นตอนการทำงานการแบ่งแยกพยางค์ ซึ่งจะแบ่งขั้นตอนการทำงานออกเป็น 2 ส่วน ดังนี้

(1) การเปลี่ยนรูปประโยคให้อยู่ในรูปพยางค์ มีขั้นตอนดังต่อไปนี้

- การนำเอาวรรณยุกต์และตัวการันต์ออก
- แทนที่พยัญชนะ “ก” ถึง “ฮ” ด้วยพยัญชนะ “อ” ยกเว้นพยัญชนะ “ย” ที่เขียนต่อจากสระอี และ “อ”

จากขั้นตอนทั้ง 2 ขั้นตอน สามารถอธิบายโดยใช้การยกตัวอย่างประกอบ เช่น ประโยคที่เป็นข้อมูลเข้า “การแบ่งพยางค์” ขั้นแรกจะนำเอาวรรณยุกต์และตัวการันต์ออก ได้เป็น “การแบงพยางค” จากนั้นทำการแทนที่พยัญชนะ “ก” ถึง “ฮ” ด้วยพยัญชนะ “อ” ยกเว้นพยัญชนะ “ย” ที่เขียนต่อจากสระอี จะได้เป็น “อาอแอออออาออ”

(2) การเปลี่ยนรูปประโยคพยางค์เป็นโครงสร้างพยางค์ เสียงพยางค์ไทยสามารถแสดงลักษณะโดยทั่วไปโดยแบ่งเป็น 4 กลุ่ม คือ เสียงพยัญชนะเริ่มต้น เสียงสระ เสียงพยัญชนะสุดท้ายและเสียงสูงเสียงต่ำ เมื่อพิจารณาถึงโครงสร้างพยางค์ไทยจะถูกแบ่งเป็น 4 แบบ คือ

1) แบบ CV เช่น (ปา [pa ;0], รี [ri ;1])

2) แบบ CCV เช่น (ครู [khru ;0])

3) แบบ CVC เช่น (กาก [ka ; k1])

4) แบบ CCVC เช่น (กวาด [kwa ; t1])

ตัวอักษรแรก “C ” แสดงเสียงพยัญชนะเริ่มต้น ขณะที่ “CC”แทนกลุ่มเสียงพยัญชนะ เสียงสระของพยางค์ถูกแสดงโดยตัวอักษร “V” และเสียงพยัญชนะสุดท้ายที่แสดงโดยตัวอักษรสุดท้าย “C” ตัวเลขอาหรับ (0-4) แสดงลักษณะของเสียงสูงเสียงต่ำ[8]

ขั้นตอนการเปลี่ยนรูปประโยคพยางค์เป็นโครงสร้างพยางค์ มีดังนี้

1) การตัดอักขระทางด้านขวาของประโยคออกจนกว่าจะสามารถเปลี่ยนเป็นโครงสร้างของพยางค์ได้ โดยวิธีการลดอักขระ คือ

- ถ้าอักขระตัวสุดท้ายเป็นตัวการันต์และอักขระตัวรองสุดท้ายเป็นสระอิ ให้ทำการตัดอักขระออก 3 ตัว เช่น “สิทธิ์” จะเหลือ “สิท” เป็นต้น

- ถ้าอักขระตัวสุดท้ายเป็นตัวการันต์และอักขระตัวถัดมาเป็น “ทร” หรือ “ทน” ให้ทำการตัดอักขระออก 3 ตัว เช่น “จันทร์” หรือ “จันทน์” จะเหลือ “จัน” เป็นต้น

- ถ้าอักขระตัวสุดท้ายเป็นตัวการันต์และอักขระตัวถัดมาตัวที่ 2 เป็นสระอุ ให้ทำการตัดอักขระออก 3 ตัว เช่น “กาพสินธุ์” จะเหลือ “กาพสิน” เป็นต้น

- ถ้าอักขระตัวสุดท้ายเป็นสระอา หรือ สระอิ หรือ สระอุ หรือ สระอู หรือ สระอะ หรือไม้หันอากาศ ให้ทำการตัดอักขระออก 2 ตัว เช่น “สบายดี” จะเหลือ “สบาย” หรือ “กินยา” จะเหลือ “กิน” หรือ “กากี” จะเหลือ “กา” เป็นต้น

- ถ้าอักขระตัวสุดท้ายเป็นอักขระที่นอกเหนือจากที่กล่าวถึง ให้ทำการตัดอักขระออก 1 ตัว เช่น “ธรรม” จะเหลือ “ธรร” หรือ “ขับรถ” จะเหลือ “ขับร” เป็นต้น

จากขั้นตอนข้างต้นสามารถแสดงกลุ่มของพยางค์เพื่อใช้ในการเปลี่ยนโครงสร้างพยางค์ดังนี้

Table 2 The format of group syllables to syllable structures

The format of group syllables	Syllable Structures
เอียะ / เอือ / เอือะ / เอือ / อัอะ / เออะ / เอะ / แอะ / โอะ / เอาะ / เออะ / เออ /เอ / แอ / ออ / อะ / อุ / อา /โอ / ออำ / อำ / ไอ / ใอ / เอา / อี / อิ / อุ / อู / อ	CV
เออียะ / เออีย / เออือะ / เออือ / ออัอะ / ออัอ / โออะ / เออาะ / เออ / ไออ / เออา /ใออ / ออา / ออู / ออุ / ออะ / ออิ / ออี	CCV
เอือ / อัอ / เอือ / แออ / อ็อ / เอ็อ / แอ็อ / อาอ / อือ / อือ / โออ / ออ / อูอ / อุอ /อิอ / อือ / อัอิ /	CVC
ออัอ / เออ็อ / แออ็อ / ออ็ออ / โออ / ออือ / ออือ / ออูอ / ออาอ / อออ / ออิอ / ออือ / ออุอ / เออิอ / เออออ	CCVC

*** หมายเหตุ** เนื่องจากเป็นงานวิจัยในการแบ่งแยกพยางค์ภาษาไทย ดังนั้นรูปแบบกลุ่มในตารางจึงมีความจำเป็นต้องใช้อักขระภาษาไทย

2) นำประโยคที่ได้จากการตัดอักขระออกทางด้านขวาแล้วไปเปรียบเทียบกับกลุ่มพยางค์เพื่อแปลงเป็นโครงสร้างพยางค์

ถ้าทำการเปลี่ยนเป็นโครงสร้างพยางค์ได้ให้ทำการแบ่งพยางค์ออกมา จากนั้นนำข้อความที่เหลือไปเปรียบกับกลุ่มพยางค์ใน Table 2 อีก ซึ่งถ้าหากเทียบไม่ได้ก็จะทำการลดอักขระจนกว่าจะเทียบได้ โดยจะทำลักษณะเช่นนี้ซ้ำๆ จนกว่าจะแบ่งแยกคำหมด ดังตัวอย่างใน Table 3

Table 3 Result of Thai segmentation with comparison of syllable structures, the Thai sentence is “แบ่งพยางค์”

Sentence	**Translate the Consonants to “อ”**	**Can be compared**	**Result**
แบ่งพยางค์	แอออออาอ	No	
แบ่งพยาง	แอออออาอ	No	
แบ่งพยา	แอออออา	No	
แบ่งพ	แอออ	No	
แบ่ง	แออ	Yes (CVC)	แบ่ง
พยางค์	ออาอ	Yes (CCVC)	แบ่งพยางค์

*** หมายเหตุ** เนื่องจากเป็นงานวิจัยในการแบ่งแยกคำภาษาไทย ดังนั้นรายละเอียดที่ยกตัวอย่างในตารางจึงมีความจำเป็นต้องใช้คำและข้อความภาษาไทย

ผลการศึกษาและสรุปผลการทดลอง

จากการศึกษางานวิจัยทั้ง 2 งานวิจัย สามารถสรุปผลการเปรียบเทียบการแบ่งแยกคำของทั้ง 2 วิธี โดยได้ทำการแบ่งหัวข้อตามขั้นตอนที่ได้จากการศึกษา ซึ่งสามารถแสดงเป็นตารางการเปรียบเทียบใน Table 4 ดังนี้

Table 4 Shown comparison of Thai word Segmentation method between Word segmentation with Thai writing structures and syllable structures

Topics for Comparison	Using Thai Writing Structures	Using Syllable Structures
Use Rules for Word segmentation	No	Yes
Use Dictionary for Word segmentation	Yes	No
Translate from input to other format	Yes	Yes
Use Trie Structure for Word segmentation	Yes	No
Cut the last Character	Yes	Yes
Fast processing	Yes	Yes
Number of format for comparison	1,648	79

วิธีการแบ่งแยกคำภาษาไทยโดยโครงสร้างการเขียนสามารถลดคำศัพท์ในพจนานุกรมจาก 16,749 คำ เหลือ 1,648 รูปแบบโครงสร้างการเขียนและอัลกอริทึมการแบ่งแยกคำภาษาไทยด้วยโครงสร้างพยางค์จะใช้กลุ่มพยางค์ในการแปลงโครงสร้างพยางค์เพื่อใช้ในการแบ่งแยกคำทั้งหมด 79 กลุ่ม จากการทดสอบประสิทธิภาพในการแบ่งแยกคำและพยางค์สามารถสรุปผลโดยใช้ตารางคอนฟิวชั่นเมตริก[9] จากการทดสอบเอกสาร 20 เอกสาร ซึ่งเอกสารที่ใช้ในการทดสอบจะให้ผู้เชี่ยวชาญตรวจสอบและปรับแก้คำให้มีความถูกต้องก่อนนำมาทดสอบเพื่อให้เป็นมาตรฐานเดียวกัน สามารถแสดงผลการทดสอบใน Table 5 และ Table 6 รวมทั้งแสดงการเปรียบเทียบประสิทธิภาพอัลกอริทึมของงานวิจัยทั้ง 2 ใน Table 7 ดังนี้

Table 5 Show performance evaluation by Confusion matrix of Thai word segmentation with Thai writing structures

The actual Target of Thai Word Segmentation	The goal of forecasting		
	Number of words expected to be incorrect	Number of words expected to be correct	Total
Number of words to be incorrect	1,640	328	1,968
Number of words to be correct	984	102,274	103,258
Total	2,624	102,602	105,226

Table 6 Shown performance evaluation by Confusion matrix of Thai word segmentation with syllable structures

The actual Target of Thai Word Segmentation	The goal of forecasting		
	Number of words expected to be incorrect	Number of words expected to be correct	Total
Number of words to be incorrect	1,020	12,238	13,258
Number of words to be correct	765	91,203	91,968
Total	1,785	103,441	105,226

Table 7 Shown performance between Thai word segmentation with Thai writing structures and syllable structures

Performance	Thai word segmentation	
	Using Writing Structures	Using Syllable Structures
Sensitivity	99.05	99.17
Specificity	83.33	7.69
Accuracy	98.75	87.64
False Positive	0.31	11.83
False Negative	37.5	42.86

จาก Table 7 เป็นการแสดงการเปรียบเทียบประสิทธิภาพความถูกต้องการการแบ่งแยกคำภาษาไทยระหว่างการแบ่งแยกคำด้วยโครงสร้างการเขียนภาษาไทยและโครงสร้างพยางค์ พบว่าการแบ่งแยกคำโดยโครงสร้างการเขียนภาษาไทยมีความถูกต้องมากกว่าการแบ่งแยกคำโดยโครงสร้างพยางค์ เนื่องจากการแบ่งแยกคำโดยโครงสร้างการเขียนภาษาไทยนั้นได้รูปแบบเพื่อใช้เปรียบเทียบจากพจนานุกรมซึ่งมีมากกว่ารูปแบบที่ได้จากโครงสร้างพยางค์ ทำให้วิธีการแบ่งแยกคำโดยโครงสร้างพยางค์ไม่สามารถแบ่งแยกคำได้ถูกต้องในบางกลุ่มคำ เช่น กลุ่มคำว่า “ในอนาคต” จะแบ่งแยกคำได้เป็น “ในอ|นาค|ต” หรือกลุ่มคำว่า “ใช้ปฏิบัติ” จะแบ่งแยกคำได้เป็น “ใช้ป|ฏิ|บัติ” เป็นต้น

สรุป

บทความวิจัยนี้มีความสนใจในการศึกษาและเปรียบเทียบวิธีการประมวลผลของการแบ่งแยกคำภาษาไทยและประสิทธิภาพความถูกต้องของอัลกอริทึม ทำการศึกษาโดยการพัฒนาโปรแกรมตามอัลกอริทึม และทำการทดสอบโปรแกรมโดยใช้เอกสารจำนวน 20 เอกสาร ซึ่งจะประเมินผลโดยผู้เชี่ยวชาญและนำผลเข้าตารางคอนฟิวชั่นเมตริกซ์เพื่อวัดประสิทธิภาพของโปรแกรม จากผลลัพธ์ในการแบ่งแยกคำนั้น ยังขาดความถูกต้องในการแบ่งแยกคำ ซึ่งสามารถพัฒนาแนวคิดในการศึกษาและสร้างกฎเพื่อแบ่งแยกคำให้ถูกต้องมากยิ่งขึ้น

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณคณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคามที่ให้ทุนอุดหนุนงานวิจัยนี้

เอกสารอ้างอิง

1. Y. Thairatananond. 1981. Towards the design of a Thai text syllable analyzer. Master Thesis. Asian Institute of Technology.
2. สัมพันธ์ ระรื่นรมย์. 2534. การแบ่งแยกคำไทยด้วยพจนานุกรม. โครงงานวิศวกรรมศาสตร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย.
3. ยืน ภู่วรวรรณ และวิวรรธ์ อิ่มอารมณ์. 2529. การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี. รายงานประชุมวิชาการวิศวกรรมไฟฟ้า ครั้งที่ 9.
4. A. Kawtrakul, C. Thumkano and S. Seriburi. 1995. A Statistical Approach to Thai Word Filtering. In Proceedings of the Symposium on Natural Language Processing in Thailan’95.
5. V. Vichianchai. Thai-Word Segmentation through Thai Writing Structure Matching, IPCSIT vol.10, 2011 International Conference on Information and Digital Engineering, September 16-18, 2011, Singapore, pp.184-188, 2011. (Publisher : IACSIT Press, ISBN 978-981-08-8723-0).
6. วุฒิชัย วิเชียรไชย. การกรองคำหยาบในข้อความภาษาไทยโดยไฟไนท์สเตทแมชชีน ใน: เอกสารประกอบการประชุมวิชาการสถิติประยุกต์ระดับชาติ สถาบันบัณฑิตพัฒนบริหารศาสตร์. กรุงเทพฯ; สิงหาคม 2549. หน้า 241-251.
7. วิจินต์ ภาณุพงศ์. 2524. โครงสร้างภาษาไทย : ระบบไวยากรณ์. กรุงเทพฯ : มหาวิทยาลัยรามคำแหง.
8. ชัชวาลย์ หาญสกุล, บรรเทิง ประดิษฐ์มิตราปิยานุรักษ์, วิรงรอง เทศประสิทธิ์และวิรัช ศรเลิศล้ำ วาณิช. Issues in ThaiText-to-Speech Synthesis:The NECTEC Approach., NECTEC Technical Journal ปีที่ 2 ฉบับที่ 7, 2543.
9. วิชิต หล่อจีระชุณห์กุล และจิราวัลย์ จิตรถเวช . เอกสารประกอบการประชุมวิชาการสถิติประยุกต์ระดับชาติ : การประยุกต์การทำเหมืองข้อมูลกับการประกันรถยนต์. ตุลาคม 2547. กรุงเทพฯ : โรงแรม Merchant Court.