การตัดคำภาษาไทยสำหรับข้อความในพิพิธภัณฑ์ปลาน้ำจืด

Thai Word Segmentation For Freshwater Fish Museum

สุรศักดิ์ ตั้งสกุล[1]* และ ฐาปนี เฮงสนั่นกูล[2]

[1]ศูนย์วิจัยและฝึกอบรมปัญญาประดิษฐ์และเทคโนโลยีสารสนเทศ มหาวิทยาลัยขอนแก่น วิทยาเขตหนองคาย

[2]คณะวิทยาศาสตร์และวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตเฉลิมพระเกียรติ จังหวัดสกลนคร

Surasak Tangsakul[1]* and Thapani Hengsanankun[2]

[1]Artificial Intelligence and Information Technology Research and Training Center, Khonkaen University, Nongkhai Campus

[2] Faculty of Science and Engineering Kasetsart University, Chalermphrakiat Sakon Nakhon Province Campus

บทคัดย่อ

งานวิจัยนี้ได้นำเสนอการตัดคำภาษาไทยโดยใช้การเทียบสายอักษรและใช้วิธีการสร้างกราฟเพื่อหาการต่อและทับกันของคำที่ปรากฏในพจนานุกรม โดยใช้ข้อมูลจากกราฟเพื่อหาขอบเขตของคำที่ปรากฏในพจนานุกรมที่ทับกัน ส่วนคำที่ไม่ปรากฏในพจนานุกรมจะใช้วิธีการระบุชนิดของคำและกฎในการหาขอบเขต ซึ่งจากการวัดประสิทธิภาพในส่วนของการตัดคำที่ไม่ปรากฏในพจนานุกรมพบว่ามีค่าความแม่นยำเท่ากับร้อยละ 75.87 ค่าความครบถ้วนร้อยละ 73.67 และค่าความถูกต้องของการตัดคำเท่ากับร้อยละ 74.75 ส่วนการวัดประสิทธิภาพในส่วนของการตัดคำในระดับพยางค์และระดับคำในเชิงความหมายพบว่ามีค่าความแม่นยำ ค่าความครบถ้วนและค่าความถูกต้องของการตัดคำโดยเฉลี่ยเท่ากับร้อยละ 73.13 67.21 และ 70.33 ตามลำดับ และเมื่อพิจารณาค่าความถูกต้องของการตัดคำที่ไม่ปรากฏในพจนานุกรมและการตัดคำในระดับพยางค์และระดับคำในเชิงความหมายพบว่ามีค่าความถูกต้องของการตัดคำทั้งหมดเฉลี่ยเท่ากับร้อยละ 72.54

คำสำคัญ : การตัดคำ การเทียบสายอักษร โครงสร้างพจนานุกรมแบบทรัย

Abstract

This paper presents the segmentation Thai words using string matching and graphs to find overlapping words that occur in Thai dictionaries. Unknown words that do not occur in Thai dictionaries had been classified according to the parts of speech and the boundary rule of each word. To measure the efficiency in segmenting unknown words has been found that the figures of precision, recall, and F-measure were as follows: 75.87%, 73.67%, and 74.75% respectively. Regarding average the efficiency of segmentation of words in terms of syllables and semantics, it was found that the figures of precision, recall, and F-measure were as follows: 73.13%, 67.21%, and 70.33%. Considering the overall segmentation of both unknown and known words, the accurate figure of the research paper was equal to 72.54%

Keywords : word segmentation, string matching, Trie dictionary structure

***Corresponding author.** E-mail: mr_tangsakul@hotmail.com

บทนำ

ธรรมชาติของประโยคภาษาไทยประกอบไปด้วยการนำคำหลายๆ คำมาเชื่อมต่อกันในลักษณะที่ติดกันโดยไม่มีการใช้เครื่องหมายหรือสัญลักษณ์ใดเพื่อบ่งบอกขอบเขตของคำแต่ละคำ แต่อาจมีการเว้นวรรคเป็นระยะๆ เพื่อแยกระหว่างประโยค ซึ่งต่างจากประโยคภาษาอังกฤษที่ทุกคำจะแยกกันอย่างชัดเจน เมื่อเราต้องนำประโยคภาษาไทยไปประมวลผลในงานประยุกต์บางประเภท เช่น ในระบบการสืบค้นข้อมูล (Information Retrieval) หรือในระบบของการแปลภาษาไทยเป็นภาษาอังกฤษ (Thai–English Machine Translation) เป็นต้น จำเป็นต้องมีกระบวนการตัดคำที่ถูกต้องและแม่นยำจึงจะทำให้ระบบดังกล่าวสามารถทำงานได้อย่างมีประสิทธิภาพ ดังนั้นการตัดคำภาษาไทยจึงถือได้ว่าเป็นปัจจัยพื้นฐานที่สำคัญของระบบที่ต้องเกี่ยวข้องกับการประมวลผลภาษาธรรมชาติ (Natural Language Processing)

ปัจจุบันได้มีการนำคำที่มาจากภาษาต่างประเทศมาใช้ร่วมกับคำภาษาไทยเป็นจำนวนมาก จึงทำให้เกิดปัญหาในเรื่องชื่อเฉพาะ คำจากภาษาต่างประเทศ คำทับศัพท์ เป็นต้น วิธีการหนึ่งที่ใช้ในการแก้ไขปัญหาดังกล่าวคือ วิธีการตัดคำโดยใช้พจนานุกรม โดยโครงสร้างของพจนานุกรมที่นิยมใช้กันมากได้แก่ โครงสร้างข้อมูลแบบบีทรี (B-Tree) โครงสร้างข้อมูลแบบ ไบนารีทรี (Binary Tree Structure) โครงสร้างข้อมูลดับเบิ้ลอาร์เรย์ (Double-Array Structure) และโครงสร้างข้อมูลแบบทรัย (Trie Structure) ซึ่งนักวิจัยหลายท่านได้พัฒนาวิธีการตัดคำเพื่อใช้กับพจนานุกรมขึ้นมา วิรัช ศรเลิศล้ำวาณิช (2536) เสนอการตัดคำโดยใช้การเทียบคำที่ยาวที่สุด (Longest Matching) และการตัดคำโดยเลือกแบบเหมือนมากที่สุด (Maximal Matching) ชิดชนก เหลือสินทรัพย์ (2545) เสนอวิธีการตัดคำภาษาไทยโดยใช้การเทียบสายอักษร (String Matching) แต่ผลลัพธ์ที่ได้ยังไม่ดีเท่าที่ควรเนื่องมาจากสาเหตุหลัก 2 ประการคือ ประการที่หนึ่งเกิดจากข้อความกำกวมโดยการตัดคำสามารถตัดคำได้หลายแบบ ทำให้เกิดความสับสนขึ้นว่าแบบไหนจะเป็นแบบที่ถูกต้องที่สุด และประการที่สองเกิดจากคำที่ไม่ปรากฏในพจนานุกรม นอกจากจะตัดคำที่ไม่ปรากฏในพจนานุกรมผิดแล้ว อาจจะส่งผลทำให้คำรอบข้างมีการตัดคำผิดด้วย

งานวิจัยนี้มุ่งเน้นการพัฒนาวิธีการตัดคำจากพจนานุกรม และพัฒนาวิธีการค้นหาขอบเขตของคำที่ไม่ปรากฏในพจนานุกรม โดยมีเป้าหมายเป็นกลุ่มคำที่ปรากฏในพิพิธภัณฑ์ปลาน้ำจืดบริเวณลุ่มน้ำโขงในประเทศไทยเป็นหลัก ทั้งนี้ เนื่องจากต้องการให้กระบวนการตัดคำมีประสิทธิภาพสูงขึ้นและสามารถใช้เป็นพื้นฐานของระบบงานสืบค้นหรือในกระบวนการของระบบอัตโนมัติอื่นๆ ที่เกี่ยวข้องกับพิพิธภัณฑ์ปลาน้ำจืดของไทย

ทฤษฎีและแนวคิดที่เกี่ยวข้อง

1. ชนิดของคำภาษาไทย

พระยาอุปกิตศิลปะสาร (2514) ได้จำแนกชนิดของคำภาษาไทยไว้ 7 ชนิด ดังนี้

1.1 คำนาม เป็นคำที่ใช้บอกชื่อคน สัตว์ สิ่งของ เป็นต้น ซึ่งสามารถแบ่งออกได้ 5 ชนิด ดังนี้

1) สามานยนาม คือคำนามที่เป็นชื่อทั่วไป เช่น คน บ้าน เมือง เวลา เป็นต้น

2) วิสามานยนาม คือคำนามที่เป็นชื่อเฉพาะที่สมมติตั้งขึ้นสำหรับเรียก คน สัตว์ และสิ่งของบางอย่าง เพื่อให้รู้ชัดเจนสิ่งที่ต้องการกล่าวถึง

3) สมุหนาม คือคำนามที่เป็นชื่อ คน สัตว์ และสิ่งของที่รวมอยู่มากด้วยกัน เช่น ภิกษุตั้งแต่ 4 รูปขึ้นไป เรียกว่า “สงฆ์” ช้างหลายตัวรวมกัน เรียกว่า “โขลง” และคำที่หมายถึงจำนวนมากอื่นๆ เช่น หมู่ คณะ ฝูง

4) ลักษณนาม คือคำที่ใช้บอกลักษณะของสามานยนามอีกทีหนึ่ง เช่น คำเรียกพระว่า “รูป” เรียกสัตว์ว่า “ตัว” หรือเรียกเรือว่า “ลำ” เป็นต้น

5) อาการนาม คือคำนามที่เป็นชื่อกริยาอาการ ซึ่งเกิดเนื่องมาจากคำกริยา หรือคำวิเศษณ์ เช่น การกิน การอยู่ ความสวย ความงาม เป็นต้น

1.2 คำสรรพนาม เป็นคำที่ใช้แทนชื่อต่างๆ ซึ่งสามารถแบ่งออกได้ 5 ชนิด ดังนี้

1) ประพันธสรรพนาม คือคำสรรพนามที่ใช้แทนนามที่ติดต่อกัน เช่น “ฉันชอบคนที่ขยัน” ซึ่งคำประพันธสรรพนามคือ “ที่” โดยคำประพันธสรรพนามที่ใช้กันมากนั้น คือ คำว่า “ที่ ผู้ที่ ซึ่ง ผู้ซึ่ง” เป็นต้น

2) วิภาคสรรพนาม คือคำสรรพนามที่ใช้แทนนามข้างหน้า เพื่อจำแนกนามนั้นออกเป็นส่วนๆ ได้แก่ “ต่าง บ้าง กัน” เช่น “ชาวนาบ้างไถนาบ้างทำนา”

3) นิยมสรรพนาม คือคำสรรพนามที่ใช้แทนนามเพื่อให้รู้แน่ชัดว่าเป็นบุคคลใด ได้แก่คำว่า “นี่ นั่น โน่น” เช่น “นี่ของใคร”

4) อนิยมสรรพนาม คือคำสรรพนามที่ใช้แทนนามที่ไม่ระบุแน่ชัดว่าเป็นบุคคลใด ได้แก่คำว่า “ใคร อะไร ไหน ผู้ใดผู้อื่น

ใดๆ อื่นๆ” เป็นต้น เช่น “ฉันไม่เห็นใครแล้ว”

5) ปฤจฉาสรรพนาม คือคำสรรพนามที่ใช้เป็นคำถาม ได้แก่ ใคร อะไร ไหน เช่น “ใครมา?”

1.3 คำกริยา เป็นคำที่แสดงอาการของคำนามหรือคำสรรพนาม เพื่อให้รู้ว่าคำนามหรือคำสรรพนามนั้นๆ ทำอะไร หรือเป็นอย่างไร เช่น “นกบิน” คำว่า “บิน” เป็นคำกริยาแสดงอาการของคำนามคือ “นก” แบ่งได้ 4 ชนิด ดังนี้

1) อกรรมกริยา คือคำกริยาที่มีใจความครบบริบูรณ์โดยไม่ต้องมีคำที่เป็นกรรมมารองรับ เช่น “คนนอน เรามา เขาไป” คำว่า “นอน มา ไป” เป็นคำกริยาที่มีใจความครบบริบูรณ์

2) สกรรมกริยา คือคำกริยาที่มีใจความไม่ครบบริบูรณ์โดยต้องมีคำที่เป็นกรรมมารองรับ เช่น “เขาเห็นคน เขาอยากนอน” เป็นต้น

3) วิกตรรถกริยา คือคำกริยาชนิดนี้ไม่มีเนื้อความในตัวต้องอาศัยเนื้อความของคำศัพท์ เช่น “เขาคล้ายกับฉัน เขาเป็นหมอ” ซึ่งใจความสำคัญนี้คือ “ฉัน หมอ”

4) กริยานุเคราะห์ คือคำกริยาที่ช่วยกิริยาอื่นให้ได้ความหมายครบบริบูรณ์ ได้แก่ “จะ คง ถูก อย่า” เช่น “เขาจะตีฉัน”

1.4 คำวิเศษณ์ เป็นคำที่ใช้ประกอบคำอื่นให้มีความหมายต่างออกไป ได้แก่ ดี ชั่ว ขาว ดำ เป็นต้น ซึ่งสามารถแบ่งได้ 10 ชนิด ดังนี้

1) ลักษณวิเศษณ์ คือคำวิเศษณ์ที่ใช้ประกอบกับคำนาม หรือคำสรรพนามเป็นพื้น ซึ่งคำวิเศษณ์ชนิดนี้จะนำเอาคำชนิดอื่นมาใช้ก็ได้ เช่น ช้างป่า ซึ่งคำว่า “ป่า” เป็นคำนาม แต่เอามาเป็นคำวิเศษณ์ประกอบกับคำว่า “ช้าง” โดยบอกลักษณะว่าเป็นช้างอย่างไร เป็นต้น

2) กาลวิเศษณ์ คือคำวิเศษณ์ที่แสดงเวลาภายหน้า ภายหลัง หรือเวลาปัจจุบัน เช่น คนโบราณ เวลาเช้า

3) สถานวิเศษณ์ คือคำวิเศษณ์ที่แสดงที่อยู่หรือระยะทาง เช่น ไกล ใกล้ เหนือ

4) ประมาณวิเศษณ์ คือคำวิเศษณ์ที่แสดงถึงจำนวน เช่น มาก น้อย ครบ

5) นิยมวิเศษณ์ คือคำวิเศษณ์ที่บอกกำหนดเขตของความหมายชัดเจน ว่าเป็นสิ่งนี้สิ่งนั้นอย่างเดียว เช่น นี้ นั้น ตัวอย่าง คนเหล่านี้ คนนี้ เป็นต้น

6) อนิยมวิเศษณ์ คือคำวิเศษณ์แสดงความไม่กำหนดแน่นอนลงไป เช่น อื่นๆ คนอื่นๆ ตัวอย่าง

7) ปฤจฉาวิเศษณ์ คือคำวิเศษณ์ที่แสดงความสงสัยหรือใช้ในคำถาม

8) ประติชญาวิเศษณ์ คือคำวิเศษณ์ที่แสดงการรับรองในเรื่องเรียกขานและโต้ตอบ เช่น ขอรับ จ๋า

9) ประติเษธวิเศษณ์ คือคำวิเศษณ์ที่แสดงการบอกความไม่รับรอง หรือห้าม

10) ประพันธวิเศษณ์ คือคำประพันธสรรพนามที่เอามาใช้เป็นคำวิเศษณ์ ได้แก่ ที่ ซึ่ง อัน และคำประสมที่เกี่ยวกับคำพวกนี้ เช่น “เป็นเวลาอันนานซึ่งประมาณไม่ได้”

1.5 คำบุพบท เป็นคำสำหรับนำหน้าคำนามและคำสรรพนาม

1.6 คำสันธาน เป็นคำเชื่อมคำหรือข้อความให้ติดกัน ซึ่งสามารถแบ่งได้ 7 ชนิด ดังนี้

1) สันธานเชื่อมความคล้อยตามกัน เช่น ก็...จึง เมื่อ...ก็

2) สันธานเชื่อมความขัดแย้ง เช่น แต่

3) สันธานเชื่อมความต่างตอนกัน เช่น ฝ่าย ส่วน อีกประการหนึ่ง

4) สันธานเชื่อมความเป็นเหตุเป็นผลกัน เช่น จึง เพราะฉะนั้น

5) สันธานเชื่อมความที่เลือกเอา เช่น หรือ มิฉะนั้น หรือมิฉะนั้น

6) สันธานเชื่อมความแบ่งรับแบ่งสู้ เช่น ถ้า...ก็

7) สันธานเชื่อมความเพื่อความสละสลวย เพื่อช่วยเนื้อหาให้สมบูรณ์มากขึ้น เช่น กับ อันว่า

1.7 คำอุทาน เป็นคำบอกเสียงต่างๆ สามารถแบ่งได้ 2 ชนิด

1) อุทานบอกอาการ คือคำอุทานที่ผู้พูดเปล่งออกมาเพื่อให้รู้จัดอาการต่างๆ ของผู้พูด

2) อุทานเสริมบท คือคำอุทานที่ผู้พูดเสริมขึ้นมักนิยมใช้แต่ในภาษาไทย โดยผู้กล่าวไม่ต้องการเนื้อความของประโยค

2. ประเภทของคำที่ไม่ปรากฏในพจนานุกรม

คำที่ไม่ปรากฏในพจนานุกรมสามารถแบ่งได้เป็น 2 ประเภท ได้แก่

2.1 คำที่ไม่ปรากฏในพจนานุกรมแบบชัดเจน (Explicit Unknown Word) เป็นคำศัพท์ที่ไม่ปรากฏในพจนานุกรม โดยภายในคำนั้นจะไม่มีส่วนใดของคำนั้นที่เป็นคำที่ปรากฏในพจนานุกรม ตัวอย่างเช่น “ฟิล์ม” “โลตัล” และ “สุนีย์” เป็นต้น

2.2 คำที่ไม่ปรากฏในพจนานุกรมแบบซ่อนเร้น (Hidden Unknown Word) เป็นคำที่ไม่ปรากฏในพจนานุกรม

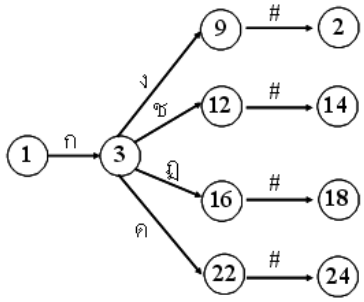
ซึ่งสามารถแบ่งได้เป็น 2 แบบ

1) คำที่ไม่ปรากฏในพจนานุกรมแบบซ่อนเร้นบางส่วน (Partially Hidden Unknown Word) เป็นคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นที่เกิดจากการประกอบกันระหว่างคำที่ปรากฏในพจนานุกรมและคำที่ไม่ปรากฏในพจนานุกรม ตัวอย่างเช่น “สุมานี” และ “ลุงแซม” เป็นต้น

2) คำที่ไม่ปรากฏในพจนานุกรมแบบซ่อนเร้นทั้งหมด (Fully Hidden Unknown Word) เป็นคำศัพท์ที่ไม่ปรากฏในพจนานุกรมอย่างซ่อนเร้นที่เกิดจากการประกอบด้วยคำที่ปรากฏในพจนานุกรมทั้งหมด หรือเป็นคำที่สร้างขึ้นใหม่โดยมีการนำคำศัพท์ต่างๆ มาประกอบกัน เช่น “สิงโตน้ำ เงินคราม” ประกอบด้วย “สิงโต” “น้ำเงิน” และ “คราม” เป็นต้น

3. โครงสร้างพจนานุกรมแบบทรัย (Trie dictionary structure)

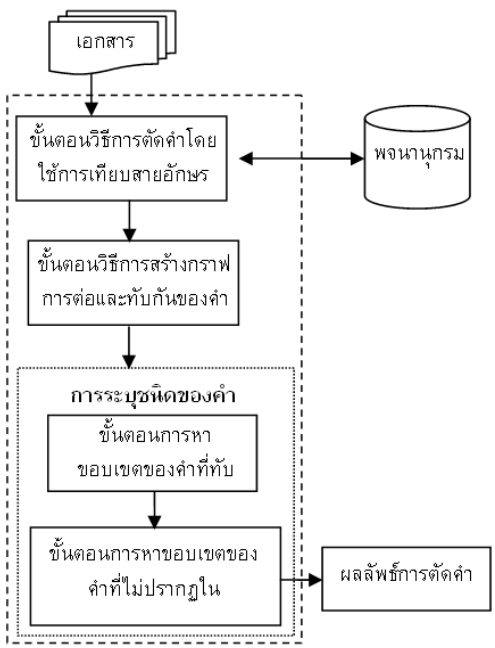
ไพศาล เจริญพรสวัสดิ์ (2541), ได้เสนอโครงสร้างข้อมูลแบบทรัยซึ่งเป็นโครงสร้างพจนานุกรมชนิดหนึ่งที่นิยมใช้ในปัจจุบัน และพีรเดช บางเจริญทรัพย์ (2552) ได้ใช้โครงการนี้ในการเก็บคลังคำศัพท์และความถี่ที่ปรากฏในคลังข้อความ โดยโครงสร้างจะมีลักษณะคล้ายกับโครงสร้างข้อมูลแบบต้นไม้ โดยที่โครงสร้างข้อมูลแบบทรัยนี้แต่ละโหนดจะจัดเก็บตัวอักษรของคำ ดังตัวอย่างที่แสดงในภาพที่ 1 เป็นโครงสร้างข้อมูลแบบทรัยที่จัดเก็บคำว่า กง กช กฎ กด จะประกอบไปด้วยโหนดต่างๆ โดยที่ข้อมูลภายใน 1 โหนด จะประกอบไปด้วยดัชนีที่ชี้ไปยังโหนดของตัวอักษรถัดไป ซึ่งมีจำนวนดัชนีเท่ากับจำนวนตัวอักขระที่จะอนุญาตให้มีได้ในพจนานุกรมบวกกับอักขระที่ใช้ระบุเป็นตัวจบคำศัพท์ (Terminator) อีก 1 ตัวอักขระ ซึ่งสัญลักษณ์ที่ใช้ในที่นี้คือตัวอักษร “#” สำหรับการสืบค้นในโครงสร้างข้อมูลแบบทรัยนี้จะทำโดย เริ่มต้นที่โหนด 1 ถ้าต้องการค้นหาคำศัพท์ก็ให้นำอักษรทีละตัวจากคำศัพท์ที่ต้องการว่าภายในโหนด 1 นั้นมีดัชนีของตัวอักษรที่ต้องการไปชี้โหนดอื่นหรือไม่ ถ้าไม่มีแสดงว่าคำนั้นไม่มีอยู่ในพจนานุกรม แต่ถ้ามีดัชนีที่ชี้ไปโหนดถัดไปก็ให้เลื่อนดัชนีไปยังโหนดที่ดัชนีนั้นชี้ไป แล้วนำตัวอักษรตัวถัดไปมาทำตามขั้นตอนแบบเดิมจนหมด เมื่อนำตัวอักษรทั้งหมดจากดัชนีมาสืบค้นคำศัพท์แล้ว ให้เดินด้วยอักษร “#” แล้วดูว่าดัชนีมีค่าเท่ากับค่าว่าง (null) หรือไม่ ถ้าเท่าแสดงว่าไม่มีคำศัพท์นั้นในพจนานุกรม แต่ถ้าไม่เท่าก็แสดงว่ามีคำศัพท์นั้นอยู่ในพจนานุกรม



ภาพที่ 1 ตัวอย่างโครงสร้างข้อมูลแบบทรัย

วิธีดำเนินการวิจัย

การวิจัยนี้มุ่งเน้นพัฒนาวิธีการตัดคำภาษาไทยสำหรับข้อความในพิพิธภัณฑ์ปลาน้ำจืด โดยใช้การเทียบสายอักษร และการระบุชนิดของคำในการค้นหาขอบเขตการทับของคำที่ปรากฏในพจนานุกรมและการค้นหาคำที่ไม่ปรากฏในพจนานุกรม ซึ่งกระบวนการตัดคำที่ใช้ในงานวิจัยนี้ ประกอบด้วย 4 ขั้นตอนหลัก คือ ขั้นตอนวิธีการตัดคำโดยใช้การเทียบสายอักษร ขั้นตอนวิธีการสร้างกราฟการต่อและทับกันของคำ ขั้นตอนการหาขอบเขตของคำที่ทับกัน และขั้นตอนการหาขอบเขตของคำที่ไม่ปรากฏในพจนานุกรม ขั้นตอนทั้งหมดแสดงในภาพที่ 2 และในส่วนของพจนานุกรมที่ใช้ในงานวิจัยนี้เป็นของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) (2547) โดยข้อมูลจะประกอบด้วย คำศัพท์ ชนิดของคำ และตัวอย่างการใช้คำศัพท์



ภาพที่ 2 กระบวนการตัดคำที่ใช้ในงานวิจัย

และเอกสารที่นำมาใช้ในการทดลองได้นำเอกสารข้อความในพิพิธภัณฑ์ปลาน้ำจืด จำนวน 27 เอกสาร แต่ละเอกสารมีลักษณะเป็นแฟ้มข้อมูลข้อความ (Text file) ที่มีจำนวนคำไม่เกิน 500 คำ ตัวอย่างข้อมูลแสดงดังนี้

ชื่อไทย *เค้าขาว*
ชื่อสามัญ *GREAT WHITE SHEATFISH*
ชื่อวิทยาศาสตร์ *Wallago attu*
ถิ่นอาศัย *พบอยู่ตามแหล่งน้ำขนาดใหญ่ เช่น แม่น้ำเจ้าพระยา บึงบอระเพ็ด แม่น้ำโขง*
ลักษณะทั่วไป *เป็นปลาใหญ่ ไม่มีเกล็ด อยู่ในวงศ์ปลาเนื้ออ่อน ส่วนหัวและจะงอยปากยื่น แหลม ปากกว้างมาก มีฟันแหลมคมอยู่บนขากรรไกรทั้งสองข้าง มุมปากยาวเกินหลังตา ตาเล็ก มีหนวดที่ริมฝีปากยาวถึงบริเวณครีบก้น หัวและลำตัวตอนหน้าแบนข้างเล็กน้อย แต่ตอนท้ายแบนข้างมาก ครีบหลังเล็ก มีปลายแหลม ครีบก้นเป็นแผ่นยาว ครีบอกใหญ่ ครีบหางเว้าตื้น สีของลำตัวเป็นสีเงินวาวอมเขียวอ่อนที่ด้านหลัง ด้านข้างลำตัวในปลาบางตัวมีแถบสีคล้ำตามแนวยาว ท้องสีจาง ครีบต่างๆ สีเหลืองอ่อน เป็นปลาที่แข็งแรงบางคราวพุ่งตัวขึ้นเหนือน้ำและปล่อยตัวให้ตกลงมาทำให้เกิดเสียงดัง*
การสืบพันธุ์ *เพาะพันธุ์โดยการฉีดฮอร์โมนแล้วผสมเทียมโดยรีดไข่และน้ำเชื้อ ไข่จะฟัก ออกเป็นตัวประมาณ 23 ชั่วโมง*
อาหารธรรมชาติ *กินปลาขนาดเล็ก เช่น ปลาสร้อย*

1. ขั้นตอนวิธีการตัดคำโดยใช้การเทียบสายอักษร (String Matching Algorithm)

การเทียบสายอักษรเป็นขั้นตอนวิธีที่ใช้ในการค้นหาคำโดยจะทำการเปรียบเทียบประโยคหรือข้อความกับคำที่มีในพจนานุกรม วิฑูรย์ (2540) ซึ่งกำหนดให้

T คือ ประโยคหรือข้อความที่ต้องการแบ่งคำ

T_i คือ ตัวอักขระของ T ที่ตำแหน่ง i

$T_{i,j}$ คือ ส่วนหนึ่งของประโยคหรือข้อความ T ที่ตำแหน่ง i ถึง j

w_i คือ คำในพจนานุกรมที่มีตำแหน่งเริ่มต้นเท่ากับตำแหน่งของอักขระที่ i

โดย i = 1,2,...,n และ n = ตำแหน่งสุดท้ายของประโยคหรือข้อความ

ขั้นตอนการตัดคำทำโดยการค้นหาคำ w_i ที่พบในพจนานุกรมโดยพิจารณาเงื่อนไขต่อไปนี้

1.1) ถ้า w_i เป็นคำที่ยาวที่สุดที่มีตำแหน่งเริ่มต้นเท่ากับ T_i กำหนดให้ i' = i - 1 + (ความยาวของ w_i) ดังนั้น $w_i = T_{i,i'}$

1.2) w_i ต้องไม่เป็นส่วนหนึ่งของทุกๆ w_j

เมื่อ $j < i$

ตัวอย่าง T = "ปลาบึกอาศัยอยู่แตกต่างกันตามสภาพในแม่น้ำโขงและแคว้นยูนาน" จะได้ผลการตัดคำดังแสดงในตารางที่ 1

ตารางที่ 1 ตัวอย่างการตัดคำโดยใช้วิธีการเทียบสายอักษร

i	T_i	w_i
0	ปลาบึกอาศัยอยู่แตกต่างกันตามสภาพในแม่น้ำโขงและแคว้นยูนาน	ปลา
3	บึกอาศัยอยู่แตกต่างกันตามสภาพในแม่น้ำโขงและแคว้นยูนาน	บึก
6	อาศัยอยู่แตกต่างกันตามสภาพในแม่น้ำโขงและแคว้นยูนาน	อาศัย
11	อยู่แตกต่างกันตามสภาพในแม่น้ำโขงและแคว้นยูนาน	อยู่
15	แตกต่างกันตามสภาพในแม่น้ำโขงและแคว้นยูนาน	แตกต่าง
18	ต่างกันตามสภาพในแม่น้ำโขงและแคว้นยูนาน	ต่างกัน
25	ตามสภาพในแม่น้ำโขงและแคว้นยูนาน	ตาม
28	สภาพในแม่น้ำโขงและแคว้นยูนาน	สภาพ
32	ในแม่น้ำโขงและแคว้นยูนาน	ใน
34	แม่น้ำโขงและแคว้นยูนาน	แม่น้ำ
40	โขงและแคว้นยูนาน	โขง
43	และแคว้นยูนาน	และ
46	แคว้นยูนาน	แคว้น
53	นาน	นาน

2. ขั้นตอนวิธีการสร้างกราฟการต่อและทับกันของคำ (Overlapping Graph)

ฐาปนี เฮงสนั่นกูล และพุธษดี ศิริแสงตระกูล (2548) ได้เสนอการตัดคำภาษาไทยโดยใช้การเทียบสายอักษรโดยค้นหาคำจากพจนานุกรม ซึ่งได้ประยุกต์ใช้การสร้างกราฟโดยวิธีการหาเส้นทางที่สั้นที่สุดเพื่อหาเซตของกลุ่มคำที่ตัดได้จากข้อความนำเข้าและประยุกต์ใช้การระบุชนิดของคำตามหลักภาษาไทย เพื่อสร้างกฎสำหรับการหาขอบเขตของคำที่ทับกันและขอบเขตของคำที่ไม่ปรากฏในพจนานุกรม เนื่องจากผลการตัดคำโดยใช้การเทียบสายอักษรจะได้คำที่ปรากฏในพจนานุกรมทั้งหมด ประกอบด้วย 2 ส่วน คือ ส่วนของคำที่มีการทับกันของตัวอักษรหรือคำ และส่วนของคำที่มีการหายไปบางส่วนของตัวอักษรคือคำที่ไม่ปรากฏในพจนานุกรม ซึ่งในการสร้างกราฟจะนำส่วนคำที่มีการทับกันของตัวอักษรหรือคำมาพิจารณาหาค่าน้ำหนักของเส้นทาง ของกราฟ แต่คำที่มีการหายไปบางส่วนของตัวอักษรจะไม่นำมาพิจารณาในการหาค่าน้ำหนักของเส้นทางของกราฟ ซึ่งกำหนดให้ G (V, E)

เป็นกราฟแบบมีทิศทางและน้ำหนัก

โดยที่ G คือกราฟการต่อและทับกันของคำ

V คือโหนดของคำที่ได้จากการเทียบสายอักษร (w_i) โดยที่ $w_i \neq \phi$ และ $1 \leq i \leq n$

E คือเส้นทางของ (w_i, w_j) โดย w_i มีส่วนต่อกันพอดีหรือทับกันกับ w_j และ $i < j$

k คือตำแหน่งเริ่มต้นที่มีการทับกันของคำ

โดยให้นำเส้นทางของ (w_i , w_j)มาพิจารณาค่าน้ำหนักของเส้นทาง ตามเงื่อนไขดังตารางที่ 2 และพิจารณาตามกรณีต่อไปนี้

กรณีที่ 1 คือ คำสองคำต่อกันพอดี

กรณีที่ 2 คือ คำสองคำทับกันบางส่วน แต่ส่วนที่ไม่ได้ทับกันของทั้งสองคำนั้นต้องเป็นคำที่ปรากฏในพจนานุกรม

กรณีที่ 3 คือ คำสองคำทับกัน แต่ต้องมีอย่างน้อยหนึ่งรูปแบบในการแบ่งคำที่ปรากฏในพจนานุกรม

กรณีที่ 4 คือ กรณีที่มีการทับกันแต่ไม่อยู่ในกรณีที่ 2 และกรณีที่ 3

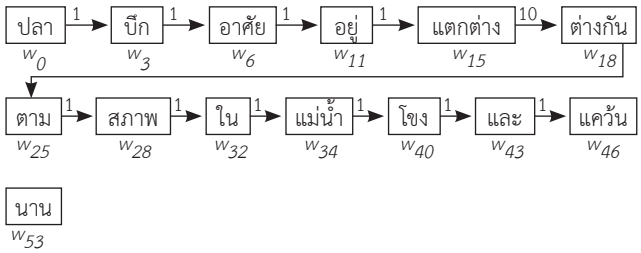
จากนั้นนำแต่ละส่วนของกราฟมาหาเส้นทาง โดยค่าน้ำหนักที่ได้จะนำมาพิจารณาหาเส้นที่มีค่าน้ำหนักน้อยที่สุดจาก ซ้ายสุดไปทางขวาสุดของแต่ละส่วนของกราฟ ตามวิธีการของ John *et al*. (2004) ผลลัพธ์ที่ได้คือเซตของคำที่ตัดได้ ซึ่งกำหนดให้ W คือ เซตของคำที่มีเส้นทางที่สั้นที่สุด

ตัวอย่างการสร้างกราฟการต่อและทับกันของคำจากประโยคตัวอย่างที่แสดงในภาพที่ 3 ซึ่งประกอบด้วย 2 ส่วนคือ ส่วนที่มีลูกศรแสดงค่าน้ำหนัก จะเป็นส่วนของคำที่ปรากฏในพจนานุกรม และส่วนที่ไม่มีลูกศรจะแสดงส่วนของคำที่มีตัวอักษรหายไป จะได้

$$W = \left\{ \begin{matrix} w_0, w_3, w_6, w_{11}, w_{15}, w_{18}, w_{25}, w_{28}, w_{32}, \\ w_{34}, w_{40}, w_{43}, w_{46}, w_{53} \end{matrix} \right\}$$

ตารางที่ 2 การกำหนดค่าน้ำหนักของเส้นทางในกราฟ

กรณี	เงื่อนไข	น้ำหนัก
1	$j = i' + 1$	1
2	ถ้า $j \leq i'$ และ $T_{i,\ j-1}, T_{j,\ i}, T_{i'+1,\ j}$ ทั้งหมดอยู่ในพจนานุกรม	10
3	ถ้า $j \leq i'$ และ k ที่ $j-1 \leq k \leq i'$ โดย $T_{i,\ k}$ และ $T_{k+1,\ j'}$ อยู่ในพจนานุกรมทั้งสองคำ	100
4	การทับกันในกรณีอื่นๆ	1000



ภาพที่ 3 ตัวอย่างการสร้างกราฟการต่อและทับกันของคำ

3. ขั้นตอนการหาขอบเขตของคำที่ทับกัน

นำผลลัพธ์จากเซตของ W ที่ได้จากขั้นตอนที่ 3.2 มาทำการระบุชนิดของคำ โดยพิจารณาชนิดของคำจากทฤษฎีของหลักภาษาไทยตามที่พระยาอุปกิตศิลปะสารได้จำแนกไว้ โดยแบ่งชนิดของคำได้เป็น 7 ชนิด ทั้งหมดแสดงดังตารางที่ 3 ซึ่งประกอบด้วยชนิดของคำและสัญลักษณ์ย่อที่ใช้แทนชนิดของคำ

โดยกำหนดให้

W_A คือ เซตของคำที่มีเส้นทางที่สั้นที่สุดหลังจากระบุชนิดของคำที่มีการทับกันของคำ

ตารางที่ 3 ชนิดของคำและสัญลักษณ์

ชนิดของคำ	สัญลักษณ์
คำนาม	N
คำสรรพนาม	PRON
คำกริยา	V
คำบุพบท	PREP
คำสันธาน	CONJ
คำอุทาน	INT
คำวิเศษณ์	ADV

และจากประโยคตัวอย่างจะได้

$$W_A = \left\{ \begin{matrix} w_0(N), w_3(ADJ), w_6(V), w_{11}(V), w_{15}(V), \\ w_{18}(V), w_{25}(ADV), w_{28}(N), w_{32}\ (PRER), \\ w_{34}(N), w_{40}(N), w_{43}(CONJ), w_{46}(N), \\ w_{53}(ADV) \end{matrix} \right\}$$

และจาก W_A ที่ได้ นำมาพิจารณาการหาขอบเขตของคำ ซึ่งแบ่งการทับกันได้เป็น 2 ลักษณะ คือ การทับกันของคำในระดับตัวอักษร และการทับกันในระดับคำ ด้วยสมการ $w_x = w_{x(1)}\ w_{x(2)}$ (A)

โดยที่ x คือ i หรือ j

w_x คือ คำที่ปรากฏในพจนานุกรมตำแหน่งที่ x

$w_{x(1)}$ คือ ส่วนหน้าของคำที่มีการทับกัน

$w_{x(2)}$ คือ ส่วนท้ายของคำที่มีการทับกัน

A คือ ส่วนบ่งบอกชนิดของคำของ w_x

และคำที่ได้จะถูกนำมาพิจารณาตามกรณีต่างๆ ดังต่อไปนี้

3.1 กรณีการทับกันในระดับคำ

การพิจารณาการทับกันของคำในระดับคำ จะพิจารณาเฉพาะกรณีที่ 2 ดังที่แสดงในตารางที่ 2 ซึ่งมีลักษณะของคำ 2 คำที่มีการทับกันของคำบางส่วน ให้พิจารณาคำที่ปรากฏในพจนานุกรมตามกฎต่อไปนี้

กฎที่ 1 กำหนดให้ $(w_{i(1)}\, w_{i(2)}\, (V)\, ,\ _{wj(1)}\, w_{j(2)}\, (N)) = (w_{i(1)}\, w_{i(2)}\, (V)\, , w_{j(2)}\, (V))$

กฎที่ 2 กำหนดให้ $(w_{i(1)}\, w_{i(2)}\, (V)\, ,\ w_{j(1)}\, w_{j(2)}\, (V)) = (w_{i(1)}\, w_{i(2)} (V)\, , w_{j(2)}\, (V))$

กฎที่ 3 กำหนดให้ $(w_{i(1)}\, w_{i(2)}\, (N)\, ,\ w_{j(1)}\, w_{j(2)}\, (N)) = (w_{i(1)}\, w_{i(2)} w_{j(2)}\, (N))$

กฎที่ 4 กำหนดให้ $(w_{i(1)}\, w_{i(2)}\, (N)\, ,\ w_{j(1)}\, w_{j(2)}\, (V)) = (w_{i(1)}\, w_{i(2)} w_{j(2)}\, (N))$

สำหรับกรณีการทับกันของคำในระดับคำที่ไม่เป็นตามกฎข้างต้น ให้นำคู่ตำแหน่งของ $(w_i(A),\ w_j(A))$ ที่ไม่เป็นไปตามกฎนั้นมารวมกันเป็นเป็นคำคำเดียว

3.2 กรณีการทับกันในระดับตัวอักษร

การพิจารณาการทับกันของคำในระดับตัวอักษร จะพิจารณาเฉพาะกรณีที่ 3 ดังที่แสดงในตารางที่ 2 ซึ่งมีการทับกันของคำในระดับตัวอักษร คำที่นำมาพิจารณาต้องเป็นคำที่ไม่มีตำแหน่งหรือส่วนของตัวอักษรที่หายไปในประโยค โดยให้ทำการเลือกคำถัดไปภายในประโยค

4. ขั้นตอนการหาขอบเขตของคำที่ไม่ปรากฏในพจนานุกรม

ในการพิจารณาหาขอบเขตของคำที่ไม่ปรากฏในพจนานุกรมนั้น งานวิจัยนี้ได้พิจารณาเฉพาะส่วนของคำแบบซ่อนเร้นบางส่วนและแบบชัดเจน โดยกำหนดให้

U = เซตของคำที่ไม่ปรากฏในพจนานุกรม

M = เซตของตัวอักษรที่หายไป

m = ตำแหน่งเริ่มต้นของตัวอักษรที่หายไป

พิจารณาสำหรับทุกคู่ตำแหน่ง $(w_i(A),\ w_j(A))$ จากกราฟการต่อและทับกันของคำ ในส่วนของคำไม่มีลูกศรแสดงส่วนของคำที่มีตัวอักษรหายไปดังแสดงในภาพที่ 3

4.1 ขั้นตอนการค้นหากลุ่มตัวอักษรที่หายไป

1) พิจารณา w_i ตัวแรกในเซต W โดย ถ้า $i \neq 1$ แล้วเพิ่ม $T_{1,\, i-1}$ ไว้ใน M

2) พิจารณา (w_i, w_j) ของเซต W โดยพิจารณาในส่วนของคำไม่มีลูกศรซึ่งแสดงส่วนของคำที่มีตัวอักษรหายไป โดย $i < j$ และ $i < m < j$ เพิ่ม $T_{i'+1,\, j-1}$ ไว้ใน M ดังนั้นตัวอย่างในภาพที่ 3 จะได้ $M = \{$ ยู $\}$

4.2 ขั้นตอนการหาคำไม่ปรากฏในพจนานุกรม

พิจารณาสำหรับทุกคู่ตำแหน่ง $(w_i(A),\ w_j(A))$ จากกราฟการต่อและทับกันของคำ ในส่วนของคำไม่มีลูกศรแสดงส่วนของคำที่มีตัวอักษรหายไปดังแสดงในภาพที่ 3 โดยในการหาคำที่ไม่ปรากฏในพจนานุกรม จะพิจารณา $(w_i(A),\ w_j(A))$ ตามชนิดของคำและเซตของ M ซึ่งสามารถแบ่งได้ 5 กรณี ดังนี้

1) กรณีการหายไปของตัวอักษรตำแหน่งแรกของประโยค พิจารณา $w_i(A)$ ตัวแรกของเซต W ถ้า $w_i \neq w_0$ และ w_i มีชนิดของคำเป็นคำสรรพนามหายทั้งคำ แต่ถ้าไม่ใช่ ให้นำมารวมกับ w_i ตัวแรกของเซต W ไว้ใน U

2) กรณีการหายไปของตัวอักษรตำแหน่งส่วนหน้าของคำภายในประโยค พิจารณา $(w_i(A),\ w_j(A))$ ที่อยู่ติดกับตัวอักษรที่หาย โดยพิจารณาชนิดของคำที่ได้มีลักษณะดังแสดงในตารางที่ 4 จะนำตัวอักษรที่หายไป (M) จะนำไปต่อกับตำแหน่งของส่วนหน้าของ w_j ตัวอย่างการหาขอบเขตของคำที่ทับกันแล้วเกิดการหายไปที่ตำแหน่งส่วนหน้าของคำ เช่น “แคว้นยูนาน” จะระบุขอบเขตของคำได้เป็น “แคว้น”,“นาน” ซึ่งมีคำที่หายไปคือ “ยู” เมื่อผ่านการค้นหาคำที่หายไปแล้วจะได้ แคว้น(N), ยู$(Miss)$, นาน(ADV) และเมื่อพิจารณาตามกฎแล้วจะได้เป็น {แคว้น, ยูนาน}

3) กรณีการหายไปของตัวอักษรตำแหน่งส่วนท้ายของคำภายในประโยค โดยพิจารณาชนิดของคำที่ได้มีลักษณะดังแสดงในตารางที่ 5 แล้วนำตัวอักษรที่หายไป (M) จะนำไปต่อกับตำแหน่งของส่วนท้ายของ w_i

ตารางที่ 4 กฎการหาขอบเขตคำในกรณีการหายของตัวอักษรส่วนหน้าคำ

กฎที่	$(w_i(A),\ w_j(A))$	กฎที่	$(w_i(A),\ w_j(A))$
1	$(w_i\ (V),\ w_j\ (N))$	6	$(w_i\ (CONJ),\ w_j\ (V))$
2	$(w_i\ (V),\ w_j\ (V))$	7	$(w_i\ (CONJ),\ w_j\ (N))$
3	$(w_i\ (V),\ w_j\ (ADJ))$	8	$(w_i\ (PRON),\ w_j\ (N))$
4	$(w_i\ (N),\ w_j\ (ADJ))$	9	$(w_i\ (ADV),\ w_j\ (V))$
5	$(w_i\ (N),\ w_j\ (ADV))$	10	$(w_i\ (ADJ),\ w_j\ (V))$

ตารางที่ 5 กฎการหาขอบเขตคำกรณีการหายของตัวอักษรส่วนท้ายของคำ

กฎที่	$(w_i(A), w_j(A))$	กฎที่	$(w_i(A), w_j(A))$
1	$(w_i (N), w_j (V))$	4	$(w_i (V), w_j (CONJ))$
2	$(w_i (ADJ), w_j (N))$	5	$(w_i (ADV), w_j (CONJ))$
3	$(w_i (N), w_j (CONJ))$	6	$(w_i (N), w_j (N))$

4) กรณีการหายไปของคำที่อยู่ภายในประโยค ให้พิจารณา $(w_i(A), w_j(A))$ ที่มีตัวใดตัวหนึ่งเป็นคำบุพบท (PREP) จะนำมาจัดเก็บใน U ทั้งคำ

5) กรณีการทับกันของคำที่ไม่ปรากฏในพจนานุกรม สำหรับ $(w_i(A), w_j(A))$ ที่มีน้ำหนักเท่ากับ 1000 ในกรณีที่ 4 จากตารางที่ 2 โดยตำแหน่งที่เกิดการทับกันของตัวอักษรจะนำมารวมกันเป็นคำเดียว

ผลการวิจัยและวิจารณ์ผล

1. ผลการทดลองการตัดคำภาษาไทย

การวัดประสิทธิภาพของการตัดคำจะใช้การหาค่าความแม่นยำ (*P=Precision*) ค่าความครบถ้วน (*R=Recall*) และค่าความถูกต้องการตัดคำ (*F=F-Measure*) ซึ่งค่าประสิทธิภาพของการตัดคำสามารถแสดงโดยความสัมพันธ์ระหว่างค่าความแม่นยำและค่าความครบถ้วนโดยใช้ดังสมการ

$$F = \frac{2 \times P \times R}{P + R}$$

โดยประสิทธิภาพของการตัดคำจะวัดจากผลการทดลองการตัดคำ 2 ส่วน ได้แก่ ผลการทดลองการตัดคำที่ไม่ปรากฏในพจนานุกรมโดยใช้การระบุชนิดของคำและกฎ และผลการทดลองการตัดคำในระดับพยางค์และระดับคำในเชิงความหมาย

1.1 ผลการทดลองการตัดคำที่ไม่ปรากฏในพจนานุกรมโดยใช้การระบุชนิดของคำและกฎ

กำหนดให้

ค่าความแม่นยำ คือ ค่าความถูกต้องในการตัดคำที่ไม่ปรากฏในพจนานุกรมเทียบกับค่าความถูกต้องในการตัดคำที่ไม่ปรากฏในพจนานุกรมโดยผู้เชี่ยวชาญ

ค่าความครบถ้วน คือ ค่าความถูกต้องในการตัดคำที่ไม่ปรากฏในพจนานุกรมเทียบกับค่าความถูกต้องในการตัดคำที่ไม่ปรากฏในพจนานุกรมจากฐานข้อมูล

จากการทดลองสามารถคำนวณประสิทธิภาพของการตัดคำที่ไม่ปรากฏในพจนานุกรมโดยใช้การระบุชนิดของคำและกฎได้เฉลี่ยเท่ากับร้อยละ 74.75 รายละเอียดแสดงดังตารางที่ 6

ตารางที่ 6 ประสิทธิภาพของการตัดคำที่ไม่ปรากฏในพจนานุกรมโดยใช้การระบุชนิดของคำและกฎ

ค่าความแม่นยำ (%)	75.87
ค่าความครบถ้วน (%)	73.67
ค่าความถูกต้อง (%)	74.75

1.2 ผลการทดลองการตัดคำในระดับพยางค์และระดับคำในเชิงความหมาย

ผลการทดลองการตัดคำในระดับพยางค์และระดับคำในเชิงความหมายจะพิจารณาจากค่าประสิทธิภาพของการตัดคำ โดยแบ่งเป็น 2 ระดับ คือ ประสิทธิภาพของการตัดคำในระดับพยางค์และประสิทธิภาพของการตัดคำในระดับคำในเชิงความหมาย

กำหนดให้

ค่าความแม่นยำ คือ ค่าความถูกต้องในการตัดคำที่ได้เทียบกับค่าความถูกต้องในการตัดคำโดยผู้เชี่ยวชาญ

ค่าความครบถ้วน คือ ค่าความถูกต้องในการตัดคำที่ได้เทียบกับค่าความถูกต้องในการตัดคำจากฐานข้อมูล

จากการทดลองสามารถคำนวณประสิทธิภาพของการตัดคำในระดับพยางค์และระดับคำในเชิงความหมายได้เท่ากับร้อยละ 74.92 และร้อยละ 65.73 ตามลำดับ รายละเอียดแสดงในตารางที่ 7

ตารางที่ 7 ผลการวัดประสิทธิภาพของการตัดคำในระดับพยางค์และระดับคำในเชิงความหมาย

ระดับ	ระดับพยางค์	ระดับคำเชิงความหมาย	เฉลี่ย
ค่าความแม่นยำ (%)	76.01	70.24	73.13
ค่าความครบถ้วน (%)	72.65	61.76	67.21
ค่าความถูกต้อง (%)	74.92	65.73	70.33

สรุปผลการวิจัย

การทดลองการตัดคำโดยใช้ขั้นตอนวิธีการตัดคำโดยใช้การเทียบสายอักษร ขั้นตอนวิธีการสร้างกราฟการต่อและทับกันของคำ ขั้นตอนการหาขอบเขตของคำที่ทับกัน และขั้นตอนการหา

ขอบเขตของคำที่ไม่ปรากฏในพจนานุกรมโดยใช้การระบุชนิดของคำและกฎ ตัวอย่างเช่น ประโยค “ปลาบึกอาศัยอยู่แตกต่างกันตามสภาพในแม่น้ำโขงและแคว้นยูนาน” สามารถตัดคำได้ดังนี้

T = {ปลา, บึก, อาศัย, อยู่, แตกต่าง, กัน, ตาม, สภาพ, ใน, แม่น้ำ, โขง, และ, แคว้น, ยูนาน}

U = {ยูนาน}

ส่วนคำที่ควรตัดได้ คือ T = {ปลาบึก, อาศัย, อยู่, แตกต่าง, กัน, ตาม, สภาพ, ใน, แม่น้ำโขง, และ, แคว้น, ยูนาน}

ซึ่งสามารถคำนวณค่าประสิทธิภาพในส่วนของการตัดคำที่ไม่ปรากฏในพจนานุกรมโดยมีค่าความแม่นยำเท่ากับร้อยละ 75.87 ค่าความครบถ้วนร้อยละ 73.67 และค่าความถูกต้องของการตัดคำได้เท่ากับร้อยละ 74.75 และในส่วนของการตัดคำในระดับพยางค์และระดับคำในเชิงความหมายสามารถคำนวณค่าประสิทธิภาพโดยมีค่าความแม่นยำเฉลี่ยเท่ากับร้อยละ 73.13 ค่าความครบถ้วนเฉลี่ยร้อยละ 67.21 และค่าความถูกต้องของการตัดคำโดยเฉลี่ยได้เท่ากับร้อยละ 70.33 และเมื่อพิจารณาค่าความถูกต้องของการตัดคำที่ไม่ปรากฏในพจนานุกรมและการตัดคำในระดับพยางค์และระดับคำในเชิงความหมายพบว่ามีค่าความถูกต้องของการตัดคำทั้งหมดเฉลี่ยเท่ากับร้อยละ 72.54

วิจารณ์ผล

การตัดคำภาษาไทยนั้นไม่สามารถตัดคำได้ถูกต้องทั้งหมด เนื่องจากสาเหตุบางประการที่ส่งผลกระทบต่อการตัดคำภาษาไทย ได้แก่ ปัญหาที่เกิดจากคำที่ไม่ปรากฏในพจนานุกรมแบบซ่อนเร้นทั้งหมด คำกำกวม การทับกันของคำที่ไม่เป็นไปตามกฎ และกระบวนการสร้างรูปแบบลำดับชนิดของคำ ซึ่งปัญหาที่พบในการทดลองทั้งหมดสรุปได้ดังต่อไปนี้

1. ปัญหาที่เกิดจากคำที่ไม่ปรากฏในพจนานุกรมแบบซ่อนเร้นทั้งหมด

เนื่องจากข้อความในพิพิธภัณฑ์ปลาน้ำจืด มีลักษณะเป็นคำไม่ปรากฏในพจนานุกรมแบบซ้อนเร้นทั้งหมด มีส่วนของคำที่ปรากฏในพจนานุกรม และส่วนใหญ่เป็นคำเฉพาะเพราะมีการตั้งชื่อตามลักษณะของปลาน้ำจืด หรือตั้งชื่อตามแหล่งที่อยู่ของปลาน้ำจืด เช่น “กดขาว” เป็นคำที่ปรากฏในพจนานุกรมทั้งหมด สามารถตัดได้เป็น “กด” และ “ขาว” แต่ในเชิงความหมายนั้นเป็นคำเฉพาะ ทำให้ประสิทธิภาพในการตัดคำในเชิงความหมายลดลง

2. ปัญหาที่เกิดจากคำกำกวม

เนื่องจากคำกำกวม เป็นคำที่สามารถตัดคำได้หลายแบบ ทำให้เกิดความสับสนของคำข้าง และเกิดการซ้อนทับกันบางส่วนของตัวอักษรของคำที่ปรากฏในพจนานุกรม จึงทำให้การตัดคำเกิดข้อผิดพลาดสำหรับการตัดคำในเชิงความหมาย

3. ปัญหาการทับกันของคำที่ไม่เป็นไปตามกฎ

สำหรับกฎที่สร้างขึ้นในงานวิจัยนี้ได้พิจารณาจากการทดลองการตัดคำภาษาไทย ดังนั้นทำให้การหาขอบเขตของคำที่ทับกันจึงยังมีกฎไม่ครอบคลุมทุกกรณีทำให้เกิดข้อผิดพลาดขึ้น ในส่วนของการหาขอบเขตของคำที่ไม่ปรากฏในพจนานุกรม ซึ่งส่งผลให้การตัดคำผิดพลาดตามไปด้วย

4. ปัญหากระบวนการสร้างรูปแบบลำดับชนิดของคำ

สาเหตุหนึ่งที่ทำให้การตัดคำภาษาไทยเกิดข้อผิดพลาดนั้นเกิดจากลักษณะของประโยคในเอกสารบางประโยคไม่ครบถ้วนตามหลักภาษาไทย จึงทำให้การสร้างรูปแบบลำดับชนิดของคำของประโยคที่มีคำที่ไม่ปรากฏในพจนานุกรมไม่ชัดเจน ส่งผลให้กระบวนการสร้างรูปแบบเกิดข้อผิดพลาด ซึ่งทำให้การตัดสินใจในการแยกกลุ่มของคำที่ไม่ปรากฏในพจนานุกรมไม่เป็นไปตามกลุ่มที่ถูกต้อง

สำหรับแนวทางการพัฒนาต่อในอนาคตเพื่อให้การตัดคำมีประสิทธิภาพเพิ่มขึ้น ซึ่งวิธีการตัดคำที่ได้นำเสนอนี้ยังไม่สามารถแก้ปัญหาได้คือ ปัญหาคำที่ไม่ปรากฏในพจนานุกรมแบบซ่อนเร้นทั้งหมด อาจแก้ไขได้โดยการเพิ่มคำในพจนานุกรม หรือสร้างกฎเพิ่มเติมในส่วนของโครงสร้างของชื่อพันธุ์ปลา ปัญหาคำกำกวมที่มีพื้นฐานจากการที่คำสามารถตัดได้หลายแบบนั้น สามารถแก้ไขได้โดยเพิ่มกระบวนการในการวิเคราะห์ในเชิงความหมายและความคลุมเครือของคำบางกรณี หรืออาจเพิ่มกระบวนย้อนกลับเพื่อตรวจสอบความหมายของประโยค และสำหรับในส่วนของการสร้างกฎและการสร้างตัวแบบของคำที่ไม่ปรากฏในพจนานุกรมนั้นยังไม่ครอบคลุมทุกกรณีที่เกิดขึ้น โดยอาจพิจารณาเพิ่มกฎในส่วนของประโยคที่ไม่ครบถ้วนตามหลักภาษาไทย

กิตติกรรมประกาศ

ขอขอบคุณมหาวิทยาลัยขอนแก่น วิทยาเขตหนองคายที่ให้ทุนสนับสนุนในการทำงานวิจัย และให้ความสะดวกในเรื่องอุปกรณ์การทดลอง ห้องปฏิบัติการ ที่ทำให้งานวิจัยครั้งนี้สำเร็จด้วยดี

เอกสารอ้างอิง

ชิดชนก เหลือสินทรัพย์. (2545). Analysis & Design of Algorithms. กรุงเทพมหานคร : School & University Media.

ฐาปนี เฮงสนั่นกูล และ พุธษดี ศิริแสงตระกูล. (2548). การตัดคำโดยใช้เทคนิค Fast and Compact Updating Algorithm. The 2nd Joint Conferenceon Computer Science and Software Engineering, 144-150.

พระยาอุปกิตศิลปะสาร. (2514). หลักภาษาไทย. กรุงเทพมหานคร : ไทยวัฒนาพานิช.

พีรเดช บางเจริญทรัพย์. (2552). A Machine-Translation based Approach to Word Boundary Identification: A Projective Analogy of Bilingual Translation. National Software Contest (NSC). Available: http://www.nectec.or.th/nsc.

ไพศาล เจริญพรสวัสดิ์. (2541). การตัดคำภาษาไทยโดยใช้คุณลักษณะ. วิทยานิพนธ์ปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์. บัณฑิตวิทยาลัย. จุฬาลงกรณ์มหาวิทยาลัย.

วิฑูรย์ กัลยาณวัฒน์. (2540). ระบบการค้นคืนข้อความภาษาไทยโดยใช้แฟ้มข้อมูลผกผัน. วิทยานิพนธ์ปริญญา. วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์. บัณฑิตวิทยาลัย. จุฬาลงกรณ์มหาวิทยาลัย.

วิรัช ศรเลิศล้ำวาณิช. (2536). การตัดคำภาษาไทยในระบบแปลภาษา. กรุงเทพมหานคร : ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. 50-55.

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ. (2547). ฐานข้อมูลพจนานุกรมภาษาไทย. http://lexitron.nectec.or.th/dnload_data/lexitron-data.zip.

John R. Hubbard and Anita Hyray. (2004). Data Structures with Java. New Jersey : Pearson Education Inc.