สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง

การวิเคราะห์การตัดคำภาษาไทย

Thai Word Segmentation Analysis

T055120

น.ส. วราภรณ์ สกาวรัตนากูล
น.ส. สุพัตรา ตรีเกษตรศาสตร์

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต
ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2546

เลขหมู่ 55120
เลขทะเบียน
วัน,เดือน,ปี 8 เม.ย. 2548

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การวิเคราะห์การตัดคำภาษาไทย

Thai Word Segmentation Analysis

โดย

น.ส. วราภรณ์ สภาวรัตนากุล

น.ส. สุพัตรา ตรีเกษตรศาสตร์

อาจารย์ที่ปรึกษา

ดร. วัชระ ฉัตรวิริยะ

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2546

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ปริญญานิพนธ์ ปีการศึกษา2546

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง	การวิเคราะห์การตัดคำภาษาไทย
	Thai Word Segmentation Analysis
คณะผู้จัดทำ	น.ส. วราภรณ์ สกาวรัตนากุล รหัส 43010375
	น.ส. สุพัตรา ตรีเกษตรศาสตร์ รหัส 43010488

..... อาจารย์ที่ปรึกษา

(ดร.วัชระ ฉัตรวิริยะ)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การวิเคราะห์การตัดคำภาษาไทย

น.ส. วราภรณ์ สกาวรัตนากุล 43010375
น.ส. สุพัตรา ตรีเกษตรศาสตร์ 43010488
ดร. วัชระ ฉัตรวิริยะ อาจารย์ที่ปรึกษา
ปีการศึกษา 2546

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์และเปรียบเทียบประสิทธิภาพของหลักการตัดคำภาษาไทยที่มีอยู่ โดยพิจารณาประสิทธิภาพของโปรแกรมในหลาย ๆ ด้านร่วมกัน เช่น ความเร็ว ความถูกต้อง ทรัพยากรที่ใช้เพื่อให้ผู้ใช้สามารถนำรูปแบบการตัดคำที่เหมาะสมมาประยุกต์ใช้กับงานที่ต้องการ ในการทำงานวิจัยนี้ ผู้วิจัยได้ศึกษาการตัดคำและคลังข้อความด้วยวิธีการต่าง ๆ แล้วนำมาสร้างเป็นโปรแกรม และทำการทดสอบประสิทธิภาพในสภาวะที่มีความใกล้เคียงกันมากที่สุด หลังจากนั้นจึงทำการสรุปวิเคราะห์ประสิทธิภาพในด้านต่าง ๆ

ทั้งนี้ผู้วิจัยได้ทำการเขียนโปรแกรมที่ใช้ในการทดลองประสิทธิภาพโดย MATLAB version 6.5 และข้อมูลและข้อมูลที่ได้ทำการทดลองได้แก่ คลังข้อความ และพจนานุกรม ได้นำมาจากศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC)

จากผลการวัดประสิทธิภาพพบว่า การตัดคำโดยใช้พจนานุกรมมีความถูกต้องมากที่สุด การตัดคำโดยกฎมีความถูกต้อง, การใช้ทรัพยากรและเวลาในการตัดคำน้อยที่สุด การตัดคำโดยใช้คลังข้อความมีการใช้ทรัพยากรมากที่สุดและใช้เวลาในการตัดคำนานที่สุด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Thai Word Segmentation Analysis

Waraporn Sakaorattanakul 43010375
Supattra Trikasetsart 43010488
Dr. Watchara Chatwiriya Adviser
Acadamic Year 2004.

Abstract

The purpose of this thesis is to analyze and compare the performance of the Thai word segmentation programs in three ways ,speed ,accuracy and resource. Several method of Thai word segmentation was studied and implemented in the same environment using based on one of the analyze a corpus name Orchid to make sure that the comparison are merit.

The performance testing programs are implemented using MATLAB version 6.5 .Data and testing data ,analyzed sentence from various source, 25,000 sentences ,and dictionary of 30,000 are provided from NECTEC.

Experimental result show that the dictionary method (longest pattern matching), gives the most accurate rate output. While the rule pattern requires minimum memory using but a poor accurate rate. The Corpus segmentation occupy the large memory and also require extensively computing power.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

ปริญญานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างดี ด้วยคำแนะนำ คำปรึกษาจากอาจารย์ที่ปรึกษา ดร.วัชระ ฉัตรวิริยะ ซึ่งต้องขอขอบพระคุณเป็นอย่างสูง ขอขอบคุณ คุณ ไพศาล เจริญพรสวัสดิ์ ที่ห้องปฏิบัติการวิจัยภาษาและวิทยาการความรู้(linguistics and knowledge science laboratory: LINKS) ของศูนย์วิจัยเทคโนโลยีอิเล็คทรอนิคส์และคอมพิวเตอร์แห่งชาติ(National Electronics and computer technology center:Nectec) ที่ให้คำแนะนำที่เป็นประโยชน์รวมทั้งคลังข้อความและพจนานุกรมที่ใช้ในงานวิจัยนี้ ขอขอบคุณภาควิชาวิศวกรรมคอมพิวเตอร์ที่มีอินเทอร์เน็ตความเร็วสูงให้บริการ สำหรับการค้นคว้าหาความรู้ต่างๆ ซึ่งท้ายที่สุดแล้วก็ประกอบกันเป็นส่วนหนึ่งของโครงงานนี้ขอขอบคุณพี่ๆ เพื่อนๆ ในห้องปฏิบัติการเน็ตเวิร์กที่คอยให้ความช่วยเหลือและกำลังใจเสมอมา

และสุดท้ายต้องขอขอบคุณบุคคลที่สำคัญที่สุดในชีวิตที่ทำให้ข้าพเจ้ามีวันนี้ นั่นคือ บิดา มารดาและบุคคลในครอบครัว อันเป็นที่เคารพรัก ซึ่งพร้อมให้โอกาสในการศึกษาอย่างเต็มที่ และยังให้กำลังใจ ความรักเสมอมา ข้าพเจ้าขอกราบพระคุณมา ณ ที่นี้ด้วย

วราภรณ์ สกาวรัตนากุล

สุพัตรา ตรีเกษตรศาสตร์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทนำ

หลักการตัดคำภาษาไทยได้มีการพัฒนาติดต่อกันมานานแล้ว และได้มีการพัฒนาวิธีการต่างๆ เพื่อให้เหมาะสมกับงานแต่ละงาน โดยในระบบต่างๆ ที่นำการตัดคำไปใช้นั้นก็มีความต้องการประสิทธิภาพในแต่ละด้านของการตัดคำไม่เท่ากัน เช่น ในงานบางอย่างอาจจะต้องการความถูกต้องในการตัดคำอย่างมาก มิฉะนั้นจะส่งผลให้ระบบไม่สามารถทำงานได้ถูกต้อง หรือบางงานอาจจะไม่ต้องการความถูกต้องมากนัก แต่ต้องการความรวดเร็วในการตัดคำมากกว่า ส่งผลทำให้มีการพัฒนาการตัดคำในรูปแบบต่างๆ ขึ้น

1.1 ความเป็นมาของปัญหา

ปัจจุบันมีการนำคอมพิวเตอร์เข้ามาใช้ในงานด้านต่าง ๆ อย่างแพร่หลายรวมถึงการนำไปใช้งานด้านการประมวลผลภาษาธรรมชาติ(Natural Language Processing) การประมวลผลภาษาธรรมชาติคือกระบวนการที่ทำให้คอมพิวเตอร์สามารถที่จะเข้าใจภาษามนุษย์ได้

แต่เนื่องจากการเขียนคำในภาษาไทยมีรูปแบบการเขียนที่เรียงติดต่อกัน จะมีการแบ่งวรรคตอน ระยะ ๆ เพื่อให้ผู้อ่านได้หยุดพัก ซึ่งต่างจากภาษาอื่น ๆ เช่น ภาษาอังกฤษ ภาษาฝรั่งเศส ที่มีวรรคตอนเป็นตัวแบ่งคำอย่างชัดเจน ดังนั้นจึงทำให้การประมวลผลภาษาไทยด้วยคอมพิวเตอร์จึงมีความแตกต่างออกไป โดยจำเป็นที่จะต้องมีวิธีการในการหาขอบเขตของคำ เพื่อให้ได้คำที่ถูกต้องตามหลักภาษาศาสตร์

ดังนั้นการตัดคำหรือการแยกคำออกจากประโยคจึงเป็นสิ่งที่มีความจำเป็นขั้นต้นต่อการประมวลผลทางภาษาธรรมชาติ และตัวอย่างงานประมวลผลทางภาษาธรรมชาติที่ต้องการการตัดคำนั้นเช่น

- การวิเคราะห์ไวยากรณ์ (syntax analysis)
- การแปลเอกสาร (machine translation)
- การสังเคราะห์เสียงพูด (speech synthesis)
- การทำดัชนีสำหรับเอกสาร (document indexing)
- การจัดรูปแบบเอกสารในงานประมวลผลคำ (word processing)

ในงานวิจัยของ ไพศาล เจริญพรสวัสดิ์ [4]ได้กล่าวถึงงานในด้านการประมวลผลภาษาธรรมชาติไว้ว่า ในงานด้านนี้นอกจากที่จะต้องรู้ขอบเขตของคำแล้วบางงานยังมีความจำเป็นที่ต้องการทราบหน้าที่คำ (Part of Speed) หรือความหมาย (Semantic) ของคำด้วยเพื่อที่จะสามารถนำไปใช้ในการประมวลผลให้มีประสิทธิภาพมากยิ่งขึ้น เช่น ในการแปลภาษา การที่จะแปลให้ถูกต้องนั้น นอก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ถูกต้องนั้น นอกจากจะต้องทราบขอบเขตของคำแล้ว การทราบหน้าที่ของคำจะเพิ่มความถูกต้องในการแปลด้วย เช่นคำว่า "เกาะ" อาจจะแปลเป็นภาษาอังกฤษได้เป็น "To attach" หรือ "Island" ซึ่งทั้ง 2 คำมีหน้าที่คำต่างกัน ดังนั้นหากเราทราบถึงหน้าที่คำ เช่นถ้าต้องการแปลคำว่า "เกาะ" ที่มีหน้าที่คำเป็นคำนามเราก็จะแปลเป็น "Island" ดังนั้นการทราบหน้าที่คำจึงส่งผลทำให้การแปลภาษามีความถูกต้องมากยิ่งขึ้น

งานวิจัยของ ดร.รัตติกร วรากูลศิริพันธุ์และทีมงาน [13] ได้กล่าวถึงปัญหาของการประมวลผลภาษาไทยด้วยคอมพิวเตอร์ไว้ว่า ลักษณะพื้นฐานของภาษาไทยประกอบด้วยหน่วยคำ (morpheme) ที่เขียนติดกันโดยไม่มีเครื่องหมายหรือช่องว่างบอกคำจบ ซึ่งทำให้เกิดความกำกวมเมื่อทำการประมวลผลประโยคภาษาไทยด้วยคอมพิวเตอร์จึงต้องอาศัยเทคนิคหรืออัลกอริทึมที่สามารถแยกหน่วยคำออกจากประโยคให้ได้ทั้งความถูกต้องและปราศจากความกำกวมไม่ว่าจะเป็นทางด้านไวยกรณ์หรือความหมาย

ในงานวิจัยหลายงานวิจัยนั้นได้ทำคิดค้นวิธีการตัดคำ และ ได้กล่าวถึงประสิทธิภาพของโปรแกรมการตัดคำไว้แต่สังเกตได้ว่าข้อมูลที่ผู้วิจัยนั้น ๆ นำมาทดสอบประสิทธิภาพมีความแตกต่างกัน จึงทำให้การเปรียบเทียบประสิทธิภาพไม่มีมาตรฐานเพียงพอที่จะสรุปได้ว่าโปรแกรมการตัดคำในแบบต่าง ๆ ดีกว่ากันอย่างไร ความต้องการที่จะเปรียบเทียบประสิทธิภาพในแง่มุมต่าง ๆ ของขั้นตอนและวิธีการตัดคำแบบต่าง ๆ จึงจำเป็นในการนำไปสู่การพัฒนาต่อไป และยังช่วยให้เลือกวิธีการตัดคำให้เหมาะสมกับงานที่ต้องการได้

1.2 วัตถุประสงค์ของวิทยานิพนธ์

1.2.1 เพื่อศึกษาแนวทางที่ใช้ในการตัดคำแบบต่าง ๆ ที่มีอยู่ เพื่อหาแนวทางในการเปรียบเทียบประสิทธิภาพของการตัดคำภาษาไทย

1.2.2 เปรียบเทียบข้อดี ข้อเด่น ของประสิทธิภาพในการตัดคำ โดยใช้Framework ที่สร้างขึ้น เช่นประสิทธิภาพด้านความถูกต้อง ด้านความเร็ว และด้านการใช้ทรัพยากร โดยใช้ข้อมูลเดียวกันเพื่อความเป็นมาตรฐาน

1.2.3 ทำการวิเคราะห์คำที่ได้จากโปรแกรมการตัดคำ เช่น สาเหตุที่ทำให้เกิดการตัดคำที่ผิดพลาดขึ้นเพื่อใช้ในการพัฒนาโปรแกรมการตัดคำต่อไป

1.3 ขอบเขตของวิทยานิพนธ์

1.3.1 ใช้ข้อมูลที่ได้จากเนคเทค (NECTEC) ซึ่งมีประโยคประมาณ 25000 ประโยคและได้ทำการตัดคำและกำกับหน้าที่ของคำไว้เรียบร้อยแล้วโดยนักภาษาศาสตร์

1.3.2 ทำการทดลองโดยใช้วิธีการตัดคำด้วยกฎ การตัดคำแบบlongest และการตัดคำโดยใช้คลังข้อความ ในการเปรียบเทียบประสิทธิภาพในด้านต่าง ๆ

1.3.3 สร้าง framework เพื่อใช้ในการทดลอง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.4 ขั้นตอนการวิจัย

1.4.1 กำหนดขอบเขตงานวิจัย

1.4.2 รวบรวบงานวิจัยทางด้านการตัดคำและงานที่เกี่ยวข้องกับการตัดคำที่ผ่านมา

1.4.3 ศึกษาการตัดคำวิธีการตัดคำแบบต่างๆที่ผ่านมา

1.4.4 สร้าง framework เพื่อใช้ในการทดลองวัดประสิทธิภาพและเปรียบเทียบผลของการตัดคำในรูปแบบต่างๆ

1.4.5 ทำการทดลองเพื่อวัดประสิทธิภาพและเปรียบเทียบผลของการตัดคำในแบบต่าง ๆ

1.4.6 สรุปและวิเคราะห์ปัญหา

1.4.7 จัดทำเอกสาร

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ได้ข้อมูลประสิทธิภาพของโปรแกรมการตัดคำภาษาไทย ในแง่มุมต่าง ๆ เพื่อใช้ให้เหมาะสมกับงาน

1.5.2 เป็นเครื่องมือในการวัดประสิทธิภาพของอัลกอริทึมที่ต้องการ

1.5.3 เป็นแนวทางนำไปสู่การพัฒนาการตัดคำภาษาไทยต่อไป

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
แนวคิดและทฤษฎีที่เกี่ยวข้อง

2.1 การตัดคำ

การตัดคำคือการแบ่งสายอักขระ (String) เพื่อหาขอบเขตของแต่ละหน่วยคำ (Morpheme) เนื่องจากโดยปกติแล้ว ภาษาไทยมีการเขียนในลักษณะที่ติดต่อกันโดยไม่มีการใช้เครื่องหมายวรรคตอนใด ๆ ยกเว้นแต่มีการเว้นวรรคเป็นระยะ ๆ ให้ผู้อ่านได้หยุดพัก และทำความเข้าใจความหมายเป็นตอน ๆ ไปเท่านั้น แม้ว่าการมีเว้นวรรคในการเขียนบทความไม่ได้มีกฎเกณฑ์ที่ชัดเจน แต่ก็ช่วยลดความคลุมเครือของคำหรือประโยคได้ ในการตัดคำหรือการหาขอบเขตของหน่วยคำที่ต่อเนื่อง ถ้าหากเรานำเก็บคำทุกคำที่มีอยู่ในภาษานั้น ๆ ลงในพจนานุกรมทั้งหมดแล้วจากนั้นก็ค้นหาและเปรียบเทียบคำศัพท์นั้น ๆ ว่ามีอยู่ในพจนานุกรมหรือไม่ เท่านั้นก็จะสามารถหาขอบเขตของคำได้ทั้งหมด แต่ในความเป็นจริง คำบางคำเป็นคำไม่ได้บรรจุในพจนานุกรม เช่นชื่อ เฉพาะ หรือคำที่เกิดจากการใช้คำใหม่

2.2 เครื่องหมายและการแบ่งวรรคตอนที่ถูกต้อง

เครื่องหมายวรรคตอนตามเอกสารเรื่อง "หลักเกณฑ์การใช้เครื่องหมายวรรคตอน และเครื่องหมายอื่น ๆ หลักเกณฑ์การเว้นวรรคหลักเกณฑ์การเขียนคำย่อ" ของราชบัณฑิตยสถาน พ.ศ.2530 ดังตารางที่ 2.1

ชื่อภาษาไทย	ชื่อภาษาอังกฤษ	เครื่องหมายวรรคตอน
มหัพภาค	full stop ,period	.
จุลภาค	Comma	,
อัฒภาค	Semicolon	;
ทวิภาค	Colon	:
วิภัชภาค	-	:-
ยัติภังค์	Hyphen	-
นขลิขิต (วงเล็บ)	Parentheses	()
วงเล็บเหลี่ยม	Square brackets	[]
วงเล็บปีกกา	Braces	{}
ปรัศนี	Question mark	?
อัศเจรีย์	Exclamation mark	!
อัญประกาศ	Double quotation marks	“ ”

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

อัญประกาศเดี่ยว	Single quotation marks	‘ ’
ไม้ยมก หรือ ยมก	-	ๆ
ไปยาลน้อย หรือ เปยยาลน้อย	-	ฯ
ไปยาลใหญ่ หรือ เปยยาลใหญ่	-	ฯลฯ
ไข่ปลา หรือ จุดไข่ปลา	Elipsis, dotted line	…
เส้นประ	Dashed line	– – –
เสมอภาค หรือ เท่ากับ	Equals	=
สัญประกาศ	Underline	ขีดเส้นใต้
บุพสัญญา	Ditto mark	‖
มหัตถสัญญา	-	(ย่อหน้าขึ้นบรรทัดใหม่)
ทับ	Virgule, slant, slash	/

ตารางที่ 2-1 เครื่องหมายวรรคตอนภาษาไทย

2.3 การพัฒนาและขั้นตอนวิธีการตัดคำภาษาไทยแบบต่างๆ

เนื่องจากการตัดคำได้มีการพัฒนาติดต่อการเป็นเวลานาน ทำให้มีงานวิจัยการตัดคำเกิดขึ้นมากมาย ซึ่งในช่วงแรกนั้น ได้มีการพัฒนาการตัดพยางค์ขึ้นมาก่อน หลังจากนั้นค่อยมีการพัฒนาด้านการตัดคำขึ้นตามมา ในบทนี้จะกล่าวถึงงานวิจัยที่เกี่ยวข้องกับการตัดคำที่ผ่านมาโดยแบ่งเป็น 3 ยุค คือ ยุคการใช้กฎ ยุคการใช้พจนานุกรม และยุคการใช้คลังข้อความ

2.3.1 กฎการหาขอบเขตหน้า-ขอบเขตหลัง (สุรินทร์ จรรยาพรพงษ์ [14])

สมัยก่อนคอมพิวเตอร์ไม่มีความสามารถในการประมวลผลสูงมากนัก ประกอบกับหน่วยความจำในเครื่องคอมพิวเตอร์มีขนาดเล็ก จึงต้องพัฒนาการตัดพยางค์ขึ้นมาก่อน เนื่องจากพยางค์มีรูปแบบกฎเกณฑ์ที่ชัดเจนแน่นอนมากกว่าคำ

สุรินทร์ จรรยาพรพงษ์ [14] ได้ทำการวิจัยเกี่ยวกับการตัดคำภาษาไทยโดยใช้พยางค์ โดยกฎที่นำมาใช้นั้นได้นำมาจากหลักไวยากรณ์ภาษาไทย และได้ทำการวิเคราะห์ลักษณะต่างๆของพยางค์ภาษาไทยออกมาได้ 200 กว่ากฎ (ภาคผนวก ข) โดยลักษณะของกฎที่ได้นี้สามารถแบ่งได้เป็น 2 ชนิดคือ กฎการหาขอบเขตหน้า(Front boundary recognition) และกฎการหาขอบเขตหลัง (Tail boundary recognition) และในแต่ละกฎยังแบ่งออกเป็น 2 กลุ่มย่อย ๆ คือ แบ่งตามคุณสมบัติของตัวอักษร โดยกฎที่ได้ออกมานี้จะจัดให้อยู่ในกลุ่มเอ (Group A) และแบ่งตามคุณสมบัติของรูปแบบการใช้สระ ซึ่งจัดกลุ่มนี้เป็นกลุ่มบี (Group B)

นอกจากนี้ยังมีการสร้างกฎสำหรับจัดการกับพยางค์ที่ไม่ใช่ลักษณะของพยางค์ภาษาไทย ซึ่งอาจประกอบไปด้วย อักขระพิเศษ ตัวเลข คำย่อ หรือ อักขระจากภาษาต่างชาติ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ในงานวิจัยของ ควงแก้ว สวามิภักดิ์ [5]ได้กล่าวถึงข้อด้อยของวิธีนี้ว่า การวิเคราะห์อาศัยตัวสระเป็นเกณฑ์ เท่านั้น ไม่ได้นำตัวอักษรที่สามารถเป็นสระมาวิเคราะห์ด้วย เช่นตัวอักษร "อ" และ "ว" ดังนั้น หากนำไปใช้กับข้อความที่ไม่มีรูปสระเลยก็จะตัดไม่ได้ เช่น "ยอดตลกหลวง" จากตัวอย่างนี้จะแสดงให้เห็นว่าการวิเคราะห์ด้วยรูปสระเพียงอย่างเดียวย่อมจะให้ผลที่สมบูรณ์แบบไม่ได้ สมควรจะต้องวิเคราะห์ถึงกฎเกณฑ์ที่ภาษาสามารถใช้ตัวอักษรสร้างสระด้วย

ในงานวิจัยนี้ได้ทดลองกับเอกสารต่าง ๆ จำนวน 100 เล่มโดยเอกสารนั้นได้นำมาจากเอกสารชนิดต่าง ๆ จำนวน 10 ชนิด เช่น หนังสือพิมพ์ ปรัชญา วิทยาศาสตร์ สาสนา ภาษาศาสตร์ ฯลฯ และจากการทดสอบ ปรากฏว่าสามารถตัดพยางค์ได้ถูกต้องถึง 96%

2.3.2 การใช้พจนานุกรม

ในยุคนี้ถือว่าเป็นยุคเริ่มแรกในการตัดคำเนื่องจากคอมพิวเตอร์มีการพัฒนาเพิ่มขึ้น และหน่วยความจำมากขึ้น ทำให้มีการคิดค้นวิธีการแบ่งคำโดยเอาพจนานุกรมเข้ามาใช้ในการตัดคำ

2.3.2.1 Longest Matching (ยืน ภู่วรวรรณและ วิวรรธน์ อิมอารมณ์ [8])

ในงานวิจัยนี้ได้ทำการจัดเก็บพยางค์ต่างๆ ไว้ในพจนานุกรม และมีการนำไวยากรณ์ต่างๆจำนวน 18 กฎเข้ามาช่วยในกรณีที่ไม่พบพยางค์ในพจนานุกรม ได้ออกแบบการแยกพยางค์ โดยใช้ดิกชันนารี โดยยึดหลักว่าดิกชันนารีมีขนาดเล็กพอที่จะบรรจุไว้ในหน่วยความจำหลัก

หลักการทำงานคือ จะทำการตรวจสอบสายอักขระ (String) ที่เข้ามาจากซ้ายไปขวากับพยางค์ที่เก็บไว้ในพจนานุกรมเพื่อทำเครื่องหมายกับทุกสายอักขระที่สามารถเป็นหนึ่งคำให้เป็นจุดเลือก สายอักขระที่ยาวที่สุดเป็นตัวเลือกสำหรับคำแรก ถ้าตัวเลือกนี้สามารถทำให้อัลกอริทึมค้นหาคำที่เหลือได้สมบูรณ์ ตัวเลือกนี้ก็จะเป็นคำแรกจริง แต่ถ้ากรณีที่เลือกพยางค์ที่ยาวที่สุดไปแล้ว ทำให้เกิดพยางค์ที่ไม่ปรากฏในพจนานุกรม ก็ยอมให้มีการย้อนรอย (Back Tracking) แล้วไปเลือกพยางค์ที่มีความยาวรองลงมาแทน และทำการค้นหาคำที่เหลือต่อไปเป็นเช่นนี้ไปเรื่อยๆ จนจบสายอักขระ วิธีการนี้เป็นที่รู้จักกันในชื่อการตัดคำ(พยางค์)แบบเลือกคำ(พยางค์) ยาวที่สุด

ในงานวิจัยนี้ได้เปรียบเทียบความรวดเร็วในการแบ่งพยางค์ ซึ่งสรุปผลได้ว่าเมื่อนำพจนานุกรมมาแบ่งพยางค์จะสามารถตัดพยางค์ได้รวดเร็วกว่าการใช้กฎ โดยที่มีความถูกต้องมากกว่า 99 % แต่วิธีการนี้ยังมีข้อด้อยคือ ต้องเสียเนื้อที่ในการเก็บพจนานุกรมในหน่วยความจำเป็นหลักเป็นจำนวน 50 กิโลไบต์ แต่ก็ยังสามารถเก็บข้อมูลพจนานุกรมไว้ในเครื่องคอมพิวเตอร์สมัยนั้นได้

ต่อมา สัมพันธ์ ระรื่นรมย์ [10] ได้นำวิธีการนี้มาพัฒนา โดยมีเป้าหมายเน้นที่ประสิทธิภาพในด้านความรวดเร็วของขั้นตอนวิธีในการตัดคำและการลดขนาดพจนานุกรม เนื่องจากเมื่อนำพจนานุกรมเข้ามาใช้ในการตัดคำแล้วจะทำให้ความถูกต้องเพิ่มมากขึ้นกว่าการตัดคำโดยใช้กฎอย่างเดียว ดังนั้นในงานวิจัยจึงไม่ได้เน้นการเพิ่มประสิทธิภาพในด้านความถูกต้องมากนัก เพราะถือว่าการตัดคำโดยใช้พจนานุกรมจะให้ค่าความถูกต้องสูงอยู่แล้ว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดยรายละเอียดของขั้นตอนวิธีการตัดคำนั้นจะมีวิธีการคล้ายกับงานวิจัยการแบ่งพยางค์โดยใช้ดิกชันนารี (ยืน ภู่สุวรรณ และ วิวรรธน์ อึ่มอารมณ์ [8]) แต่จะทำการจัดเก็บคำลงในพจนานุกรมแทนพยางค์

2.3.2.2 Maximal Matching(วิรัช ศรเลิศล้ำวานิช [9])

ในงานวิจัยนี้ได้มีการพัฒนาการตัดคำโดยเรียกว่า "การตัดคำโดยเลือกแบบเหมือนกันมากที่สุด (Maximal Matching)" สามารถแก้ไขความบกพร่องของการตัดคำแบบเลือกคำยาวที่สุดได้ เพราะการเลือกคำที่ยาวที่สุดเมื่อเจอข้อความโดยไม่พิจารณาถึงข้อความถัดไป พิจารณาเฉพาะบริเวณใกล้ๆเท่านั้น ทำให้เลือกคำที่ยาวเกินไปตั้งแต่แรก ข้อความที่ตามมาจึงเกิดข้อผิดพลาดได้ แต่วิธีการเลือกคำแบบเหมือนมากที่สุดจะพิจารณาข้อความทั้งหมดแทน หลักการของการตัดคำโดยเลือกแบบเหมือนมากที่สุดคือ ขั้นตอนแรกคือจะทำการตัดคำที่เป็นไปได้ทุกๆ แบบก่อนแล้วหลังจากนั้นจึงเลือกประโยคที่มีจำนวนคำน้อยที่สุด สำหรับในกรณีที่ตัดคำแล้วเกิดได้จำนวนคำที่เท่ากันก็ให้ทำการตัดคำแบบเลือกคำยาวที่สุดเข้ามาช่วยพิจารณา

แม้ว่าการค้นหาการแบ่งคำทุกทางที่เป็นไปได้อาจจะทำให้ต้องเสียค่าใช้จ่าย (cost) มาก โดยทั่วไปวิธีนี้จะให้ความถูกต้องที่สูงกว่าวิธีตัดคำให้ยาวที่สุด แต่ก็มีประโยคจำนวนมากที่ไม่สามารถตัดได้ถูกต้องโดยวิธีนี้ นอกจากนี้ได้มีผู้พัฒนาการตัดคำ ของวิรัช เพิ่มเติมโดยแทนที่จะเปรียบเทียบคำที่ยาวสุดเสมอ จะมีการเปิดตารางค่าความยาวสูงสุดที่พบในพจนานุกรม ของข้อความที่ค้นหา โดยใช้ตัวอักษร 2 ตัวแรกเป็นตัวเปรียบเทียบ ถ้ามีค่าน้อยกว่าความยาวของข้อความก็ให้เริ่มแบ่งคำตั้งแต่ค่าที่น้อยที่สุด เช่น "ฉันมารอกราบ" มีความยาวข้อความเท่ากับ 11 และค่าที่ได้จากการเปิดตารางของ "ฉั" พบว่ามีค่าสูงสุดคือ 5 ฉะนั้นการแบ่งคำจึงเริ่มจากตัวอักษรที่ 5 ลงมา

ในงานวิจัยของคุณ ไพศาล เจริญพรสวัสดิ์ [4] ได้กล่าวถึงข้อดีของวิธีการนี้ว่า วิธีการนี้สามารถช่วยแก้ไขข้อบกพร่องของการตัดคำแบบเลือกคำยาวที่สุดเมื่อเจอข้อความที่กำกวมก่อน โดยไม่มีการพิจารณาถึงข้อความถัดไปซึ่งมีลักษณะเหมือนการใช้ขั้นตอนวิธีแบบโลภ (Greedy Algorithm) ที่พิจารณาเฉพาะบริเวณที่ใกล้ ๆ เท่านั้น แต่วิธีการตัดคำแบบเหมือนกันมากที่สุดนั้นจะเป็นการใช้ขั้นตอนวิธีแบบโลภโดยพิจารณาข้อความทั้งหมดแทน แต่วิธีนี้ก็ยังไม่สามารถตัดคำได้ถูกต้องทั้งหมด จำเป็นต้องมีการนำโครงสร้างทางไวยกรณ์หรือความสัมพันธ์ทางความหมายเข้ามาใช้ประกอบการพิจารณาด้วย

2.3.3 การตัดคำโดยใช้การใช้คลังข้อความ (บุญเสริม กิจศิริกุล [6])

ในยุคนี้ ได้มีการพัฒนาคลังข้อความ (Corpus) ขึ้นจำนวนมาก ทำให้มีการนำความรู้ต่างๆจากคลังข้อความเข้ามาประยุกต์ใช้ด้วย ตัวอย่างความรู้ที่ได้จากคลังข้อความเช่น ค่าสถิติการใช้คำภายในคลังข้อความและลักษณะไวยากรณ์ที่ใช้ในคลังข้อความ เป็นต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

วิธีการนี้เป็นการตัดคำโดยใช้ข้อมูลทางสถิติที่ได้จากคลังข้อความที่เตรียมไว้ โดยใช้ค่าสถิติที่เกิดขึ้นจากลำดับของหน้าที่คำหรือประเภทย่อยของคำ หรืออาจกล่าวได้ว่าเป็นการนำเอาส่วนหนึ่งของไวยากรณ์มาใช้ ซึ่งคลังข้อความนี้เราได้ให้นักภาษาศาสตร์ตัดคำและกำกับหมวดคำที่เหมาะสมไว้แล้ว ส่วนปัญหาการตัดคำของภาษาไทยเราสามารถเขียนแทนด้วยสมการดังนี้

$$\arg\max_{W_{1,n}} P(W_{1,n}/C_{1,m}) = \arg\max_{W_{1,n}} P(W_{1,n})P(C_{1,m}/W_{1,n})/P(C_{1,m}) \quad (2\text{-}1)$$

โดยที่

$C_{1,m} = c_1,c_2,c_3, \ldots,c_m$ คือ ค่าสายอักขระที่เข้ามาเป็นอินพุตของเรา

$W_{1,n} = w_1,w_2,\ldots,w_n$ คือคำที่ที่สามารถตัดได้

สมการที่ 2-1 จึงมีหมายความว่า เรามีสายอักขระ $C_{1,m}$ ที่เป็นอินพุตเข้ามา เราต้องการแบ่งสายอักขระนี้เป็นคำ $w_1,w_2,\ldots,w_n$ $(W_{1,n})$ และเนื่องจาก $P(C_{1,m}/W_{1,n})$ มีค่าเท่ากับ1 และค่า $P(C_{1,m})$ ก็เป็นค่าคงที่ในทุก ๆ ครั้ง ดังนั้นทำให้เราสามารถเขียนสมการที่ 2-1 ในรูปสมการใหม่ได้เป็น

$$\arg\max_{W_{1,n}} P(W_{1,n}/C_{1,m}) = \arg\max_{W_{1,n}} P(W_{1,n}) \quad (2\text{-}2)$$

เมื่อพิจารณาถึงภาษาไทย การที่เราจะกำหนดขอบเขตของคำกำกวมได้นั้นเราต้องอาศัยสถิติของหน้าที่คำ N-gram และ สถิติการเกิดคำ ๆ นั้น จากนั้นก็หาค่าความน่าจะเป็นสูงสุดของคำ/หน้าที่คำ ในแต่ละทั้งประโยค เนื่องจากมีหน้าที่คำเข้ามาเกี่ยวข้องด้วยทำให้สมการที่ 2-2 เปลี่ยนรูปแบบไปเป็นสมการที่ 2-3

$$\arg\max_{W_{1,n}} P(W_{1,n}/C_{1,m}) = \arg\max_{W_{1,n}} \sum_{T_i} P(W_{1,n},T_{1,n}) \quad (2\text{-}3)$$

โดยที่ $T_{1,n} = t_1,t_2,\ldots,t_n$ คือหน้าที่คำของ $w_1,w_2,\ldots,w_n$ ตามลำดับ

และเพื่อให้ได้ค่าของ $P(W_i/C_{1,m})$ สูงที่สุด ในหลักการของไตรแกรมโมเดล เราจะตั้งสมมติฐานดังนี้

1) ความน่าจะเป็นของคำหนึ่ง ๆ จะปรากฏ ณ ตำแหน่งใด ๆ ในประโยคไม่ขึ้นกับสิ่งอื่น ๆ
2) หมวดคำหนึ่ง ๆ จะขึ้นอยู่กับหมวดคำก่อนหน้า 2 หมวดคำเท่านั้น

จากสมมติฐานทั้ง 2 ข้อนั้นจะทำให้สมการที่ 2-3 เปลี่ยนรูปแบบเป็นสมการที่ 2-4

$$\arg\max_{W_{1,n}} P(W_{1,n}/C_{1,m}) = \arg\max_{W_{1,n}} \sum_{T_i} \prod_{i} P(W_i,T_i)*P(T_i/T_{i-1},T_{i-2}) \quad (2\text{-}4)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดยค่าของ $P(W_i,T_j)$ และ $P(T_i/T_{i-1},T_{i-2})$ สามารถคำนวนจากข้อมูลในคลังข้อความที่มีอยู่ เรา สามารถหาค่าของ $P(W_i,T_j)$ ได้โดยนับจำนวนของหมวดคำ t_i ที่เป็นหมวดคำของคำ w_i หารด้วยจำนวนของ t_iที่เป็นหมวดคำของคำใดๆ (จำนวนของหมวดคำ t_iทั้งหมด) ส่วน $P(T_i/T_{i-1},T_{i-2})$ หาโดยนับจำนวนหมวดคำ t_i ที่มีหมวดคำ t_{i-1}และ t_{i-2} ตามหลังที่ปรากฏในคลังข้อความหารด้วยจำนวนของหมวดคำ t_{i-1}และ t_{i-2} ที่ปรากฏทั้งหมด จากค่าเหล่านี้ที่คำนวณได้เราจะนำมาใช้ในการตัดคำภาษาไทยและได้เขียนซอฟท์แวร์เพื่อทดลองผล โดยวิธีการทำงานจะกล่าวถึงในบทที่ 5 หัวข้อที่ 5.3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
มาตรวัดประสิทธิภาพการตัดคำแบบต่าง ๆ

งานวิจัยฉบับนี้มีจุดประสงค์เพื่อวิเคราะห์การทำงานของโปรแกรมตัดคำภาษาไทย การที่เราจะวิเคราะห์ได้ว่าโปรแกรมทำงานมีประสิทธิภาพมากน้อยเพียงใดจำเป็นต้องกำหนดหลักเกณฑ์ที่ใช้ในการวัดประสิทธิภาพ การกำหนดมาตราวัดประสิทธิภาพที่เหมาะสมจะทำให้การเปรียบเทียบความสามารถของโปรแกรมมีความถูกต้องมากยิ่งขึ้น มาตราวัดประสิทธิภาพที่ใช้ในงานวิจัยนี้มี 3 หัวข้อ คือ ความถูกต้อง ความเร็ว และการใช้ทรัพยากร

3.1 ความถูกต้องในตัดคำ

การประเมินความถูกต้องทำได้โดยการตรวจสอบผลลัพธ์ที่ได้จากโปรแกรมตัดคำกับคำตอบซึ่งทำการตัดไว้แล้วโดยนักภาษาศาสตร์ในคลังข้อความ โดยถือว่าเป็นคำตอบที่มีความถูกต้อง 100 %

วิธีในการประเมินความถูกต้องมีหลายรูปแบบ ในงานวิจัยฉบับนี้จะใช้มาตรวัดความถูกต้องโดยวิธีการวัดค่า F-Measure (Van Rijbergen,1979 cited in Manning and Zuchuze[15]) ซึ่งเป็นวิธีพื้นฐานที่ใช้กันโดยทั่วไปสำหรับการวัดความแม่นยำที่อ้างอิงความถูกต้องจากเอกสารที่ผ่านการกำกับจากผู้เชี่ยวชาญ โดยทำการพิจารณาจากค่าความแม่นยำและความครบถ้วน คำนวณได้ดังสมการที่ 3-1

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \qquad (3\text{-}1)$$

ค่าความครบถ้วน (Recall) มีการคำนวณดังนี้

$$R = \frac{\text{จำนวนคำที่โปรแกรมตัดคำได้ถูกต้อง}}{\text{จำนวนคำทั้งหมดที่ตัดไว้ในคลังข้อความ}}$$

ค่าความแม่นยำ (Precision) มีการคำนวณดังนี้

$$P = \frac{\text{จำนวนคำที่โปรแกรมตัดคำได้ถูกต้อง}}{\text{จำนวนคำทั้งหมดที่โปรแกมตัดได้}}$$

ซึ่งในการวัดประสิทธิภาพนี้พิจารณาให้คาทั้งสองมีน้ำหนักเทาๆกัน ($\alpha = 0.5$) ทำให้สามารถแปลงสมการในการวัดคาใหอยูในรูปที่งายขึ้นได้ ดังสมการที่ 3-2 คือ

$$F = \frac{2 \times P \times R}{P + R} \qquad (3\text{-}2)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2 มาตรวัดประสิทธิภาพเชิงความเร็ว

ในการนำโปรแกรมการตัดคำภาษาไทยไปประยุกต์ใช้กับการใช้งานจริง ปัจจัยที่ต้องนำมาร่วมพิจารณาด้วยเสมอก็คือ ความเร็วในการทำงาน เนื่องจากงานในแต่ละรูปแบบจะมีความต้องการในเรื่องความเร็วที่แตกต่างกัน โดยมาตรวัดประสิทธิภาพในเรื่องความเร็วนี้จะคำนวณออกมาในรูปความเร็วเฉลี่ยต่อคำ (sec/word)

3.3 มาตรวัดประสิทธิภาพการใช้ทรัพยากร

การใช้งานทรัพยากรก็เป็นส่วนหนึ่งที่เราต้องคำนึงถึง ในงานบางอย่างที่มีหน่วยความจำมีความจำกัดต้องการอัลกอลิธึมการตัดคำที่มีการใช้พื้นที่ในหน่วยความจำน้อย โดยหน่วยวัดของมาตรวัดประสิทธิภาพนี้จะเป็น ขนาดของโค๊ด(code)และ ข้อมูลที่เกี่ยวข้องในการตัดคำ ซึ่งมีหน่วยเป็น ไบท์ (byte)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
ข้อมูลที่ใช้ในการเปรียบเทียบประสิทธิภาพและการตัดคำ

ข้อมูลที่นำมาใช้ทั้งหมดนี้ ได้รับมาจากห้องวิจัยลิงค์ (LINKS) ที่เนคเทค (NECTEC) ซึ่งสามารถดาวน์โหลดได้ที่ http://www.links.nectec.or.th/www-new/download.php โดยคลังข้อความที่ใช้นั้นมีชื่อว่า Orchid Corpus ซึ่งมีประโยคประมาณ 25,000 ประโยค และ พจนานุกรม มีชื่อว่า RIWord ซึ่งเป็นพจนานุกรมฉบับราชบัณฑิตยสถานมีคำประมาณ 32,575 คำ

4.1 คลังข้อความ

คลังข้อความที่นำมานี้จะใช้ในการทดสอบและการหาค่าสถิติที่ใช้อ้างอิงในการตัดคำ โดยแบ่งลักษณะของคลังข้อความเป็น 2 ส่วน โดยส่วนแรกจะเป็นส่วนที่ใช้ในการทดสอบประมาณ 20 เปอร์เซ็นต์ เพื่อวัดประสิทธิภาพการตัดคำของแต่ละโปรแกรม ส่วนอีก 80 เปอร์เซ็นต์ ใช้หาค่าสถิติที่ในการตัดคำ ลักษณะของคลังข้อความเป็นดังรูปที่ 4-1

```
%TTitle: การพัฒนาต้นแบบคอมพิวเตอร์ถ่ายภาพตัดขวางอวัยวะ (ระยะที่ 2)
%ETitle: COMPUTER X-RAY TOMOGRAPHY (PHASE 2)
%TAuthor: เหรียญชัย พงษ์, อัครินทร์ คุณกิตติ, ยุทธพงษ์ รังสรรค์เสรี, สมชาย เกียงอารีกุล
%EAuthor:
%TInbook: การประชุมทางวิชาการ ครั้งที่ 5, โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์, ปีงบประมาณ 2535
%EInbook: The 5th Annual Conference, Electronics and Computer Research and Development Project, Fiscal Year 1992
%TPublisher: ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม
%EPublisher: National Electronics and Computer Technology Center, Ministry of Science, Technology and Environment
%Page:
%Year: 1993
%File:
#P1
#1
การพัฒนาต้นแบบคอมพิวเตอร์ถ่ายภาพตัดขวางอวัยวะ (ระยะที่ 2)//
การ/FIXN
พัฒนา/VACT
ต้นแบบ/NCMN
คอมพิวเตอร์/NCMN
ถ่าย/VACT
ภาพตัดขวาง/NCMN
อวัยวะ/NCMN
<space>/PUNC
<left_parenthesis>/PUNC
ระยะ/NCMN
ที่ 2/DONM
<right_parenthesis>/PUNC
//
```

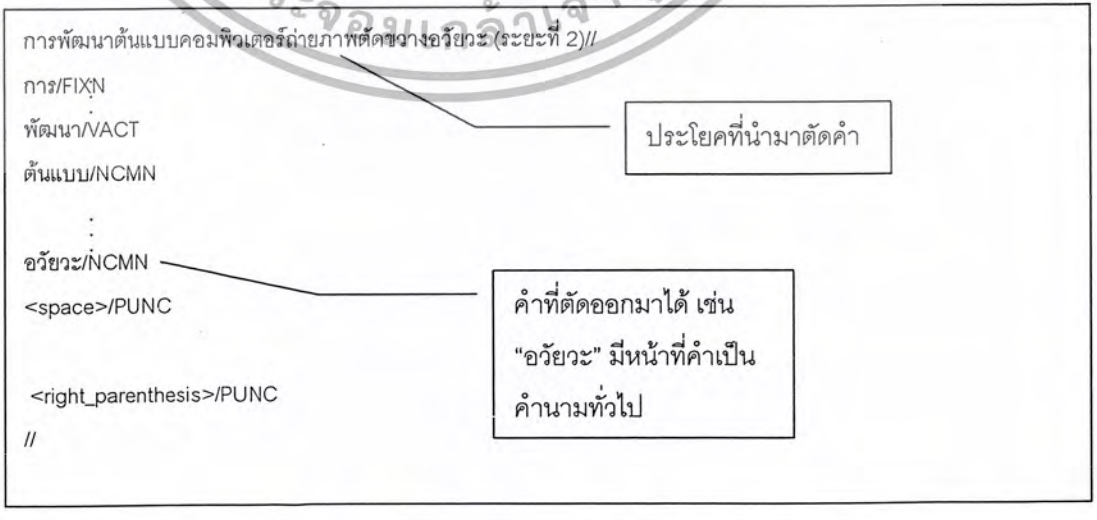
รูปที่ 4-1 แสดงลักษณะของคลังข้อความ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สังเกตได้ว่าคลังข้อความนั้นจะมีการบอกรายละเอียดที่มาของข้อมูล เช่น มาจากที่ใด ใครเป็นผู้แต่ง แต่ส่วนที่เราใช้นั้นจะเป็นส่วนของ ประโยค คำในประโยคที่ทำการตัดไว้ และหน้าที่ของคำเหล่านั้น ซึ่งทั้งหมดทำโดยนักภาษาศาสตร์ สำหรับสัญลักษณ์และอักษรย่อที่ใช้บอกหน้าที่ของคำสามารถดูรายละเอียดได้ในภาคผนวก ก

```
%TTitle: การพัฒนาต้นแบบคอมพิวเตอร์ถ่ายภาพตัดขวางอวัยวะ (ระยะที่ 2)
%ETitle: COMPUTER X-RAY TOMOGRAPHY (PHASE 2)
%TAuthor: ไพรัช ธัชยพงษ์, อัครินทร์ คุณกิตติ, ยุทธพงษ์ รังสรรค์เสรี, สมชาย เกรียงอารีกุล
%EAuthor:
%TInbook: การประชุมทางวิชาการ ครั้งที่ 5, โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์, ปีงบประมาณ 2535
%EInbook: The 5th Annual Conference, Electronics and Computer Research and Development Project, Fiscal
Year 1992
%TPublisher: ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, กระทรวงวิทยาศาสตร์ เทคโนโลยีและสิ่งแวดล้อม
%EPublisher: National Electronics and Computer Technology Center, Ministry of Science, Technology and
Environment
%Page:
%Year: 1993
%File:
#P1
#1
```

รูปที่ 4-2 ส่วนของรายละเอียดของประโยคที่นำมาตัดคำ



รูป 4-3 ส่วนของประโยคและคำที่ตัดไว้เรียบร้อยแล้ว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ทั้งนี้ในการทดลองได้ทำการแก้ไขข้อความบางส่วนที่ผิดพลาด เช่น คำที่ตัดได้กับประโยคที่นำมาตัดมีตัวอักษรบางตัวขาดหายไป แต่มีการกำกับคำได้ถูกต้องก็จะทำการเพิ่มตัวอักษรให้เหมือนกัน เป็นต้น

4.2 พจนานุกรม

พจนานุกรมที่ใช้นั้นเป็นพจนานุกรมฉบับราชบัณฑิตยสถานมีจำนวนคำทั้งหมด 32575 คำ มีขนาด 248 KB เนื่องจากในพจนานุกรมนั้นมีคำซ้ำอยู่บ้างจึงได้ตัดออกไปบางส่วน และได้ตัดอักษรตัวเดียวออก ยกเว้น “ฌ” “ฐ” เนื่องจากอักษรอื่นที่นอกเหนือจากนี้ไม่มีการใช้งานเมื่ออยู่ตัวเดียว ลักษณะของพจนานุกรมเป็นดังรูปที่ 4-4

รูป 4-4 ลักษณะของพจนานุกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
โปรแกรมและอัลกอลิธึมในการตัดคำภาษาไทยที่นำมาทดสอบ

5.1 การตัดคำด้วยกฎ

ในงานวิจัยนี้ได้อ้างอิงกฎที่ใช้ในการตัดคำจากงานวิจัยของคุณ สุรินทร์ จรรยาพรพงษ์ [14] ซึ่งทำการวิเคราะห์รูปแบบการตัดพยางค์โดยพิจารณาจากไวยกรณ์ของสระเป็นหลัก ได้ทำการวิเคราะห์ออกมาเป็นกฎต่าง ๆ กว่า 200 กฎ(ภาคผนวก ข) โดยแบ่งออกเป็นการตัดพยางค์ข้างหน้าสระ(front boundary) และการตัดพยางค์ข้างหลังสระ (tail boundary) โดยใช้รูปสระอย่างเดี่ยวทั้งหมด 17 รูป ได้แก่

ะ , ิ , ี , ุ , ู , ึ , ื , ็ , ั , ่ , ๎ , ฤ , เ , แ , โ , ไ , ใ และรูปสระผสมอีก 13 รูป ได้แก่

เ า, เ าะ, เ อ, เ อะ, เ ว, เ ย, เ ีย, เ ียะ, เ ะ, เ ิ , เ ื , เ ็ , ็ อ นอกจากนี้ยังประกอบด้วยกฎเกณฑ์เกี่ยวกับพยัญชนะอีกบางกฎ เช่น

ตัดหน้าอักษร ฉ ผ ฝ ฮ ยกเว้น จะมีสระ เ แ โ หรือ ใ นำหน้าซึ่งจะทำให้ตัดหน้าสระแทน และยังทำการแบ่งตัวอักษรออกเป็น 5 ประเภทคือ อักษรที่เป็นได้ทั้งตัวพยัญชนะและตัวสะกด อักษรที่เป็นได้ทั้งสระและตัวอักษร อักษรที่ส่วนมากเป็นตัวสะกด อักษรที่ส่วนมากเป็นพยัญชนะ และอักษรที่เป็นพยัญชนะเสมอ

ในแต่ละกฎจะมีการระบุคุณสมบัติของกฎนั้นไว้ด้วย โดยคุณสมบัติของกฎมี 3 อย่าง คือ

Rigid pattern = รูปแบบกฎที่ตายตัวสามารถตัดได้เลยโดยไม่ต้องพิจารณาสระข้างเคียงร่วม

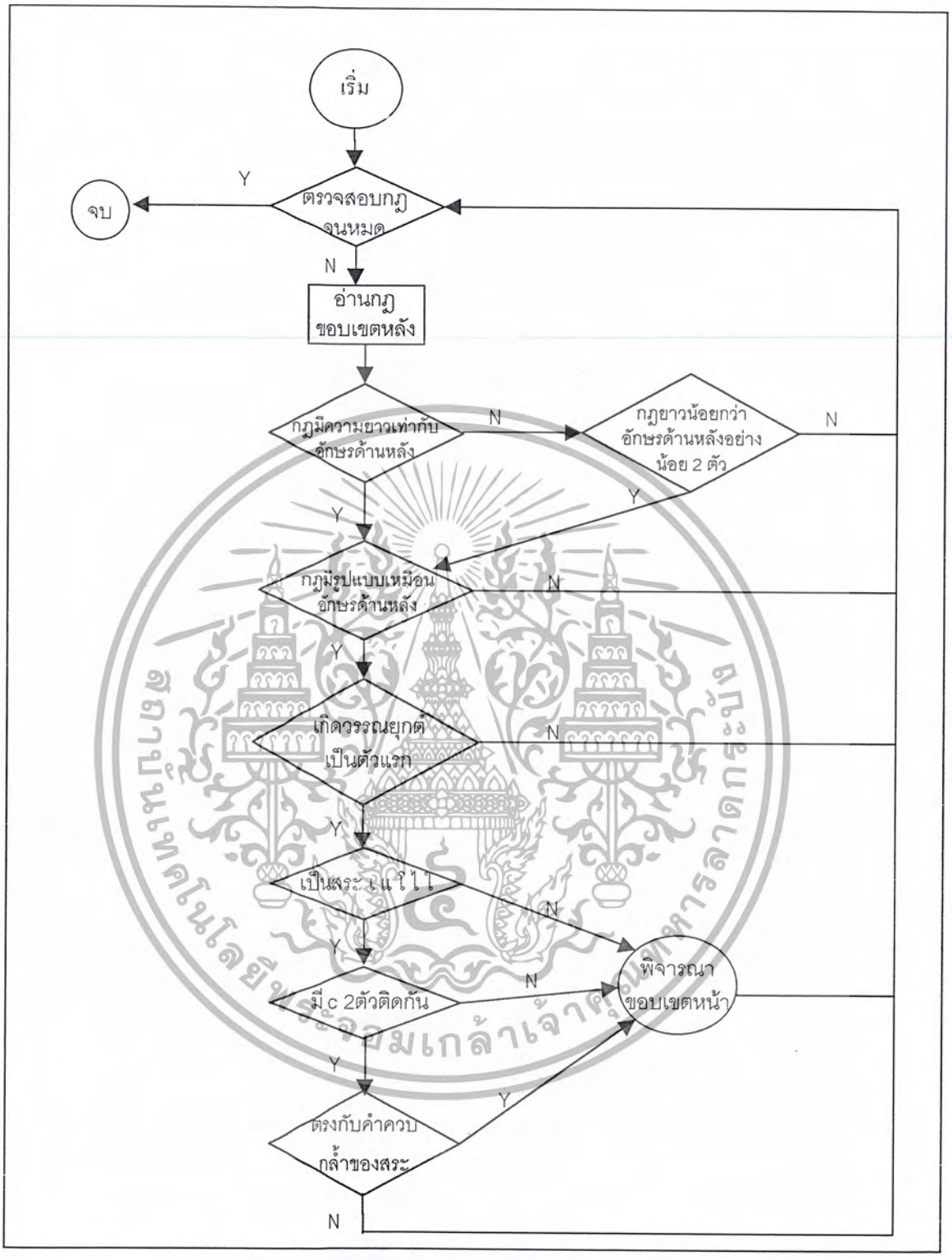
Non – rigid = รูปแบบกฎที่อาจตัวได้ถูกหรือไม่ถูกต้องก็ได้ต้องใช้สระข้างเคียงมาช่วยพิจารณา

Exceptional pattern = กรณียกเว้นพบเพียงไม่กี่พยางค์หรือคำ ซึ่งจะถูกเก็บไว้ในพจนานุกรม

อัลกอลิทึมที่ใช้ในการตัดคำโดยกฎ

เริ่มจากการหาตำแหน่งสระในข้อความที่ต้องการพิจารณา แล้วดูว่าตรงกับกับกรณีพิเศษหรือไม่ ซึ่งกรณีพิเศษเหล่านี้ได้แก่ คำที่เกิดการรวมสระเข้าด้วยกัน (สระผสม) เช่น เ – อ ะ และคำที่มีรูปแบบตายตัวไม่สามารถตัดเป็นรูปแบบอื่นได้ เช่น ใ- จะพบเพียง 20 คำเท่านั้น หากไม่ใช่กรณีพิเศษให้พิจารณาจากหลักการตัดโดยทั่วๆ ไปตามแผนภาพที่ 5-1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

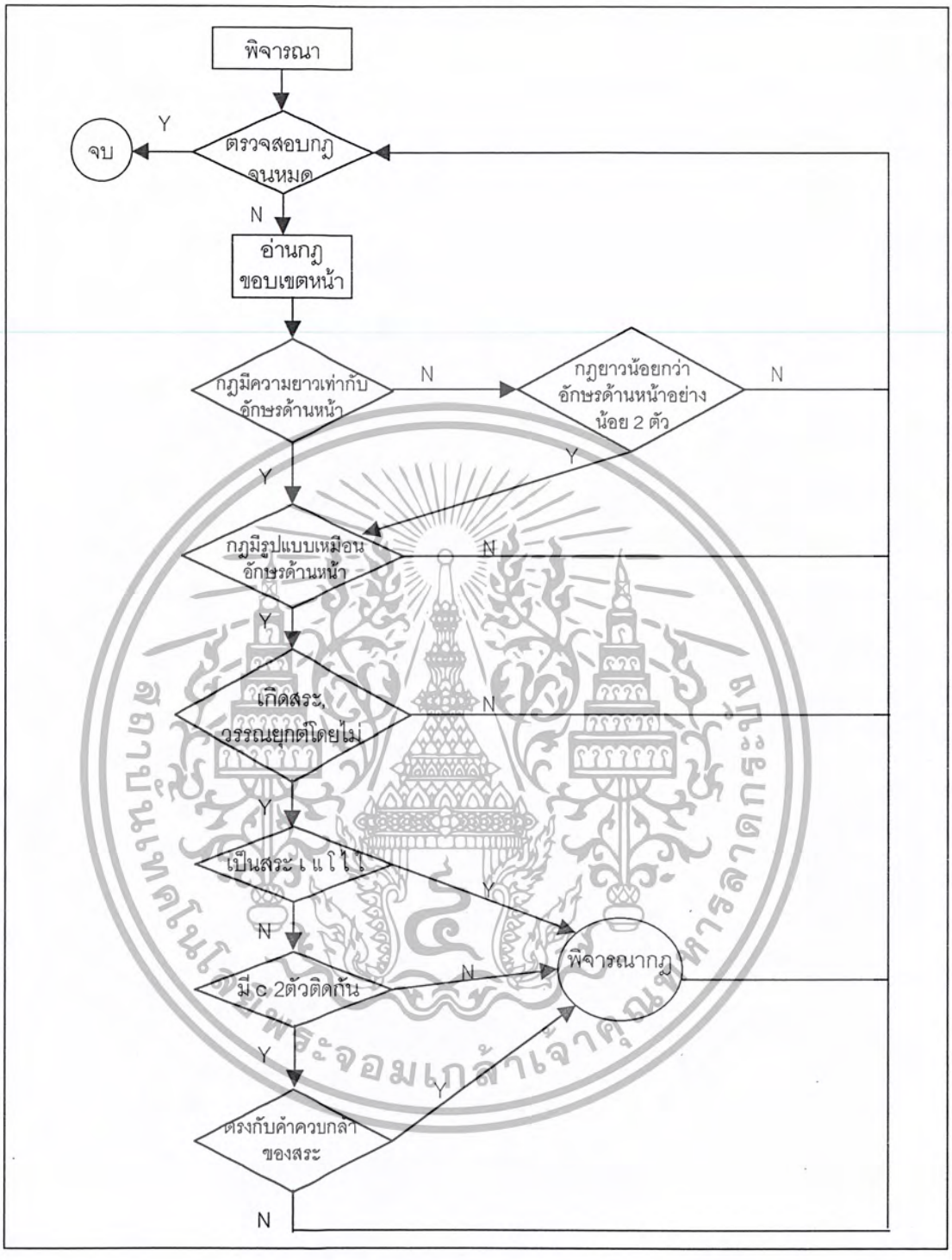


รูปที่ 5-1 แสดงการตัดคำแบบทั่วๆไป ในส่วนการหาขอบเขตหลัง

การพิจารณาขอบเขตหลังนอกจากจะพิจารณาให้ตรงกับกฎแล้ว ยังห้ามตัดแล้วเหลืออักษรตัวเดียว ห้ามตัดหน้าวรรณยุกต์ สำหรับสระ เ , แ , โ , ใ , ไ หากมีพยัญชนะติดกัน 2 ตัว พยัญชนะเหล่านั้นต้องตรงตามหลักอักษรควบกล้ำ หลังจากนั้นจึงทำการหาขอบเขตหน้าของสระตัวหลัง ดังรูปที่ 5-3 หากกฎนั้นไม่เป็นกฎตายตัวจะต้องพิจารณาขอบเขตหน้าของสระตัวหลังประกอบด้วย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง



รูปที่ 5-2 ส่วนของการหาขอบเขตหน้าของสระตัวหลัง

ในการพิจารณาขอบเขตหน้าก็มีหลักการคล้ายคลึงกัน โดยกฎที่ตัดได้ต้องไม่ทำให้เหลืออักษรตัวเดียว ไม่ทำให้เกิดวรรณยุกต์โดยไม่มีตัวอักษรต่อท้าย หากกฎมีพยัญชนะ 2 ตัวติดกัน เฉพาะสระ เ, แ, โ, ใ, ไ เท่านั้นที่ไม่ต้องพิจารณาตัวควบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้
55120

ในการเขียนโปรแกรมได้ใช้หลักการตัดพยางค์ของคุณ สุรินทร์ จรรยาพรพงษ์ เพียงบางส่วนเท่านั้น เนื่องจาก มีเอกสารบางส่วนขาดหายไป เช่น รายละเอียดเกี่ยวกับตัวอักษรแต่ละประเภท เอกสารเกี่ยวกับคำที่ยกเว้นซึ่งเก็บอยู่ในพจนานุกรม ซึ่งอาจส่งผลกระทบต่อประสิทธิภาพความถูกต้องของการตัดพยางค์ได้

5.2 การตัดคำโดยใช้พจนานุกรม

การตัดคำโดยพจนานุกรมเป็นการใช้พจนานุกรมช่วยในการแบ่งอักขระที่เรียงติดต่อกันออกมาเป็นคำๆ โดยนำอักขระที่เรียงติดกันไปเปรียบเทียบกับพจนานุกรม ถ้าหากพบในพจนานุกรมก็แสดงว่าอักขระที่เรียงติดกันนั้นเป็นคำ ๆ หนึ่ง แต่เนื่องจากพจนานุกรมมีคำจำนวนมากและมีขนาดใหญ่ ดังนั้นการหาคำในพจนานุกรมจะทำการหาแบบไล่ดูทีละตัว (Sequential) จึงไม่สามารถทำได้ เพราะจะใช้เวลาในการหานานมาก ดังนั้นจึงต้องมีการสร้างโครงสร้างพจนานุกรมขึ้นเพื่อใช้ในการค้นหาคำในพจนานุกรมมาเปรียบเทียบกับกลุ่มอักขระที่พิจารณาอยู่

5.2.1 โครงสร้างพจนานุกรม

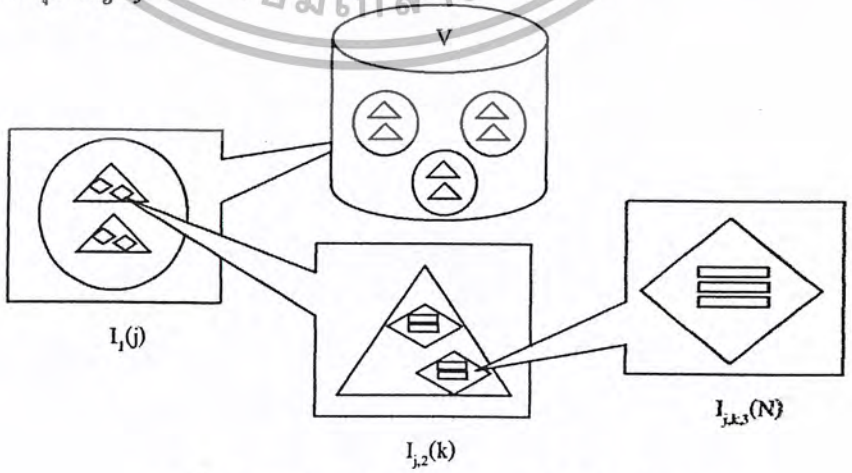
เราจะใช้ดรรชนีในการจัดของกลุ่มคำศัพท์ในพจนานุกรมภาษาไทยทั้งหมด โดยจะทำการเก็บบันทึกคำศัพท์ตามค่าดรรชนี 3 ตัว ที่ได้จากค่าแอสกีนัมเบอร์ (ASCII Number)ของอักขระตัวที่ 1 ค่าแอสกีนัมเบอร์ของอักขระตัวที่ 2 และจำนวนอักขระทั้งหมดของคำศัพท์นั้น ๆ ซึ่งค่าแอสกีนัมเบอร์ของตัวอักขระตัวแรกของคำศัพท์เป็นดรรชนีที่แบ่งคำศัพท์แบ่งคำศัพท์ออกเป็นกลุ่มใหญ่ และให้ค่าแอสกีนัมเบอร์ของอักขระตัวที่ 2 ของคำศัพท์นั้นเป็นดรรชนีที่สองที่แบ่งคำศัพท์ภายในกลุ่มใหญ่ ให้เป็นกลุ่มย่อย และจำนวนของอักขระของคำศัพท์จะถูกใช้เป็นดรรชนีในการแบ่งคำศัพท์ในกลุ่มย่อยอีกที

กำหนดให้ V เป็นคำศัพท์ทั้งหมดที่บรรจุอยู่ในพจนานุกรม

$I_1(j)$ เป็นดรรชนีตัวแรกที่แสดงกลุ่มคำศัพท์ใหญ่กลุ่มที่ j (j = 1,2,3,...)

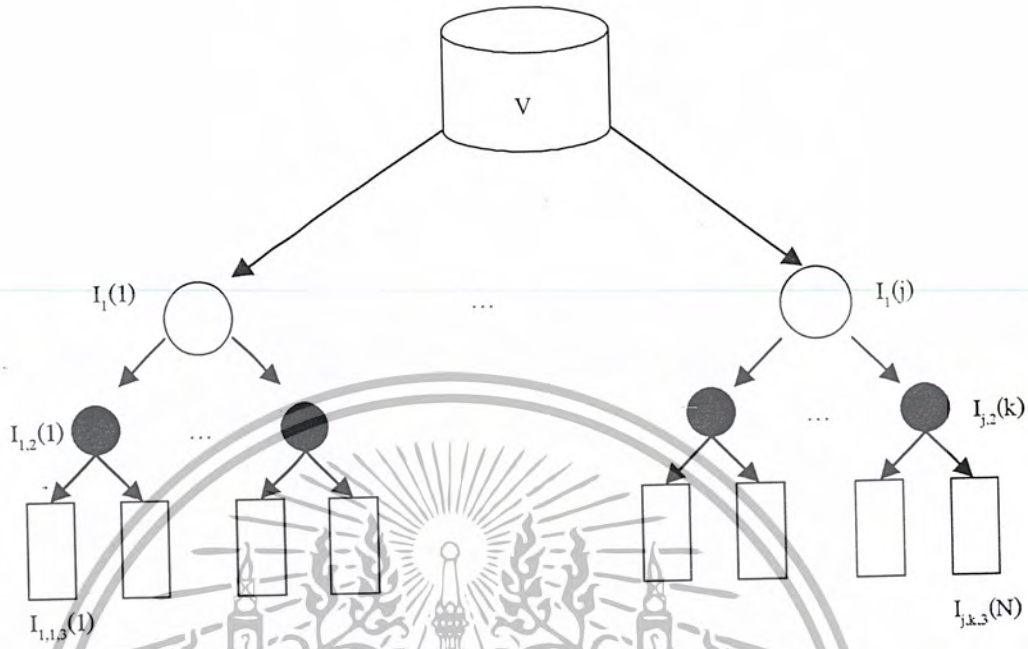
$I_{j,2}(k)$ เป็นดรรชนีตัวที่ 2 ที่แสดงกลุ่มคำศัพท์ย่อยกลุ่มที่ k ภายในกลุ่มใหญ่ที่ j (k = 1,2,3,...)

$I_{j,k,3}(N)$ เป็นดรรชนีตัวที่ 3 ที่แสดงกลุ่มคำศัพท์ย่อยกลุ่มที่ N โดยที่คำศัพท์นั้นอยู่ในกลุ่มย่อยที่ k และกลุ่มใหญ่ที่ j



รูปที่ 5-3 แสดงโครงสร้างของพจนานุกรมไทย (ในรูปแบบ chart)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 5.4 แสดงโครงสร้างของพจนานุกรมไทย (ในรูปแบบ tree)

5.2.2 การแยกหน่วยคำออกจากประโยค

อัลกอลิธึมของการแยกหน่วยคำมีหลักการทำงานดังนี้

ตัวแปรที่ใช้ในการตัดคำ

len = จำนวนอักขระที่ใช้เปรียบเทียบ

sen = ประโยคที่นำมาตัดคำ

buck = array ที่เก็บตำแหน่งของคำที่ตัดได้

ตัวแปรที่ใช้ชี้ข้อมูล

x = เป็นตัวชี้ตำแหน่งเริ่มต้นของคำที่จะตัด

y = เป็นตัวอ้างตำแหน่งเดิมของ x (หากตัดคำแล้ว x จะเลื่อนตำแหน่งไป หากไม่เลื่อน (y = x)แสดงว่าภายในคำนั้นไม่มีอยู่ในพจนานุกรม)

ตัวแปรที่ใช้เก็บข้อมูล

len = ตัวแปรที่เก็บความยาวของคำที่จะตัด

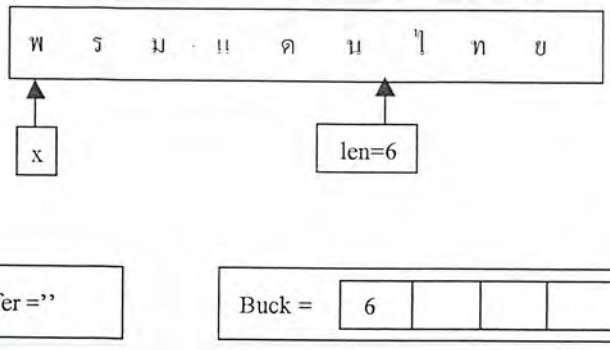
sen = array ที่เก็บประโยคที่จะนำมาตัดคำ

buck = array ที่เก็บตำแหน่งของคำที่ตัดได้

back = ตัวแปรที่บอกว่ามีการย้อนไปตัดคำก่อนหน้าใหม่

buffer = ตัวแปรสำหรับเก็บจำนวนอักขระที่ตัดคำไม่ได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



1 ตรวจสอบว่ามีประโยคเข้ามาหรือไม่ หากมี ให้ตัวแปร sen รับค่าประโยคที่เข้ามาและกำหนดค่าเริ่มต้นให้กับ x ให้เท่ากับ 0 เพื่อชี้ที่ตำแหน่งแรกจากนั้นก็ทำตามขั้นตอนต่อไป หากไม่มีก็ออกจากโปรแกรม

2 ให้ len เท่ากับความยาวที่มากที่สุดของคำศัพท์ที่ขึ้นต้นด้วย sen[x] (ในตอนแรก x = 0 ดังนั้น sen[0] ก็คืออักขระตัวแรกของประโยค)แต่หาก len มากกว่าความยาวทั้งหมดของประโยค ให้ len เท่ากับความยาวทั้งหมดของประโยค

3 ค้นหาคำศัพท์ในพจนานุกรมตั้งแต่อักขระตัวที่ x จนถึง len หากไม่พบในพจนานุกรม จะลด len ลงไปทีละ 1 จนกระทั่งพบคำศัพท์นั้นหรือ len = 0 จึงหยุดการเปรียบเทียบ

4 หากในขั้นตอนที่ 3 สามารถตัดคำได้ ให้นำค่า len+x เก็บลงใน buck แล้วไปขั้นตอนที่ 7 แต่หากตัดคำไม่ได้
(x = y) ให้ไปขั้นตอนต่อไป

5 ตรวจสอบว่าค่า buffer ไม่เท่ากับ 0 (นั่นคืออักขระตัวก่อนหน้าก็ไม่สามารถตัดเป็นคำได้) หรือ มี pause เท่ากับ 1 (ไม่สามารถย้อนกลับไปตัดคำก่อนหน้าได้)หรือไม่ หากตรงกับเงื่อนไขใดเงื่อนไขหนึ่งให้เลื่อน x ไป1 ตำแหน่งและให้ buffer เพิ่มขึ้นอีก1 เป็นการข้ามอักขระตัวนั้น แล้วไปขั้นตอนที่ 7 หากไม่ตรงกับเงื่อนไขใดไปขั้นตอนต่อไป

6 ตรวจสอบว่าคำก่อนหน้านี้สามารถตัดได้สั้นกว่านี้หรือไม่ โดยให้ cur เก็บค่าตำแหน่งเดิมก่อนตัด หากคำก่อนหน้าไม่สามารถตัดได้สั้นกว่านี้ ให้เลื่อน x ไป 1 ตำแหน่ง และให้ buffer เพิ่มค่าไปอีกหนึ่ง และให้ค่า pause = 1 เพื่อเป็นการบอกว่าไม่สามารถย้อนกลับไปได้ แต่หากว่าคำก่อนหน้าสามารถตัดได้เป็นคำสั้น ๆ 2 คำ แต่สุดท้ายก็ยังคงอยู่ที่ตำแหน่งเดิม(x = cur) เช่น คำว่า “ระหว่าง” สามารถแยกได้เป็น “ระ” และ”หว่าง” ก็ให้เลื่อนค่า x ไป1 ตำแหน่งเพื่อข้ามอักขระนั้นไปและ ให้ buck เก็บค่าตำแหน่งเดิม

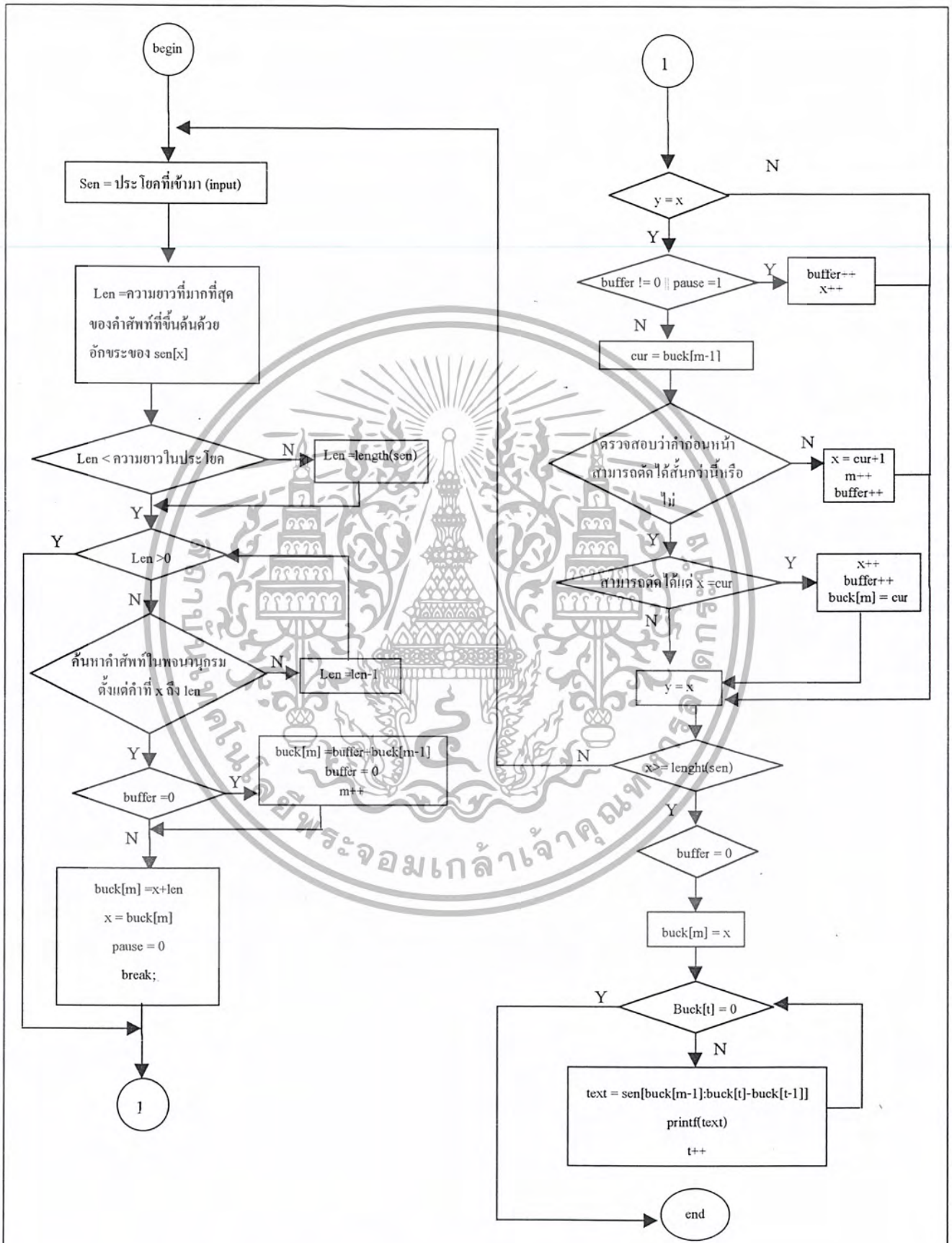
7 ให้ y = x เพื่อใช้ในการตรวจสอบต่อไป

8 ตรวจสอบว่าตำแหน่ง x เป็นตำแหน่งสุดท้ายของประโยคหรือไม่ ถ้าใช่ไปขั้นตอนต่อไป ถ้าไม่ใช่กลับไปขั้นตอนที่ 1ใหม่เพื่อหาคำที่จะตัดต่อไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

9 ตรวจสอบว่า buffer มีค่าอยู่หรือไม่หากมีให้นำตำแหน่งสุดท้ายมาเก็บไว้ใน buck จากนั้น วนลูปนำค่าตำแหน่งต่าง ๆ ที่เก็บไว้ใน buck ขึ้นมาเพื่อแสดงคำที่ตัดได้ จนหมดพร้อมทั้ง clear ค่าใน array จากนั้นไปขั้นตอนที่ 1 ใหม่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

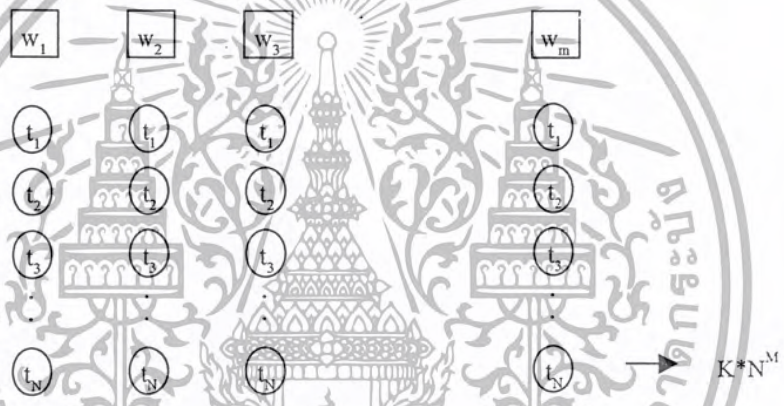
รูปที่ 5-5 โฟลว์ชาร์ตแสดงการทำงานของโปรแกรมการตัดคำด้วยพจนานุกรม

5.3 การตัดคำโดยใช้คลังข้อความ

จากที่ได้กล่าวมาแล้วในหัวข้อที่ 2.3.3.1 จะได้สมการในการตัดคำเป็นดังสมการที่ 2-3 คือ

$$\arg\max_{W_{1,n}} P(W_i/C_{1,m}) = \arg\max_{W_{1,n}} \sum_{T_i} \prod_{i} P(W_i,T_i)*P(T_i/T_{i-1},T_{i-2})$$

หากเราทำการคำนวณตามสมการที่ 2-3 โดยตรง จะทำให้การกำกับหน้าที่คำทำได้ช้ามาก เนื่องจากต้องเสียเวลาในการคำนวณทุกรูปแบบที่เป็นไปได้ เช่น ถ้านำสายอักขระที่มีจำนวนคำทั้งหมด M คำ และมีจำนวนหมวดคำทั้งหมด N หมวด หากคิดในกรณี worst case คือ ทุกคำในสายอักขระสามารถเป็นได้ทั้ง N หมวด ดังนั้นจะต้องเสียเวลาในการคำนวณประมาณ $k*N^M$ ครั้งต่อสายคำเดียว โดย k คือค่าคงที่ (ไพศาล เจริญพรสวัสดิ์ [4])



จะเห็นว่าวิธีการนี้ใช้เวลาค่อนข้างนานมาก ซึ่งเวลาที่ใช้จะขึ้นอยู่กับจำนวนคำและจำนวนหมวดคำที่เป็นไปได้ของแต่ละคำในสายนั้น ซึ่งทำให้เวลาที่ใช้เป็นสัดส่วนแบบเอกซ์โปเนนเชียล (Exponential) ดังนั้นจึงได้นำเทคนิคเรื่องไดนามิคโปรแกรมมิ่งเข้ามาช่วย ซึ่งเทคนิคที่เลือกใช้นี้มีชื่อว่า "ขั้นตอนวิธีวิเทอร์บิ (Viterbi Algorithm)" ซึ่งเป็นเทคนิคที่นิยมนำมาใช้กับแบบจำลองไตรแกรม เมื่อนำมาประยุกต์กับการกำกับคำด้วยไตรแกรมสามารถแสดงได้ดังรูปที่ 5-6

กำหนดให้

$w_1,w_2,\ldots,w_m$	เป็นลำดับในประโยค
M	คือ จำนวนคำในประโยคที่นำมากำกับหน้าที่คำ
$t_1,t_2,\ldots t_n$	เป็นหน้าที่คำทั้งหมดที่เป็นไปได้
N	คือ จำนวนหน้าที่คำที่เป็นไปได้ทั้งหมด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$prob(w_i/t_i)$ คือค่าความน่าจะเป็นของการเกิดคำศัพท์ w_i ที่มีหน้าที่คำเป็น t_i หาได้โดยการนับจำนวนคำ w_i ที่มีหน้าที่คำเป็น t_i หารด้วย จำนวนหน้าที่คำ t_i ทั้งหมดในคลังข้อความ

$prob(t_i/t_{i-1},t_{i-2})$ คือค่าความน่าจะเป็นของการเกิดลำดับของหน้าที่คำเป็น t_{i-2} t_{i-1} และตามด้วย t_i หาได้โดยการนับจำนวนของหน้าที่คำที่มีลำดับเป็น t_{i-2},t_{i-1},t_i ทั้งหมดในคลังข้อความ หารด้วย จำนวนของหน้าที่คำที่มีลำดับเป็น t_{i-2},t_{i-1} ทั้งหมดในคลังข้อความ

จะทำการสร้าง Array ขนาด NxNxM จำนวน 2 Array คือ

- seqscore [i][j][r] ทำการเก็บค่าความน่าจะเป็นที่ดีที่สุดของการกำกับหน้าที่คำของ $w_1,w_2,...,w_r$ ซึ่งคำที่ r –1 และ r มีการกำกับหน้าที่เป็น t_i และ t_j ตามลำดับ
- backptr [i][j][r] จะเก็บหน้าที่คำของคำ r-2 เมื่อคำที่ r-1 และ r มีหน้าที่คำเป็นที่ t_i t_j ตามลำดับ

หลังจากนั้นจะทำการหาลำดับของหน้าที่คำ $C_1,C_2,...,C_m$ ที่เป็นลำดับของคำในประโยคที่มีความน่าจะเป็นมากที่สุด

Initialization step

For i=1 to N do

For j=1 to N do

seqscore[i][j][2] = $prob(w_1/t_i)*prob(t_i/ø)*prob(w_2/t_j)*prob(t_j/t_i,ø)$

backptr[i][j][2] = 0

Iteration step

For r=3 toM do

For j=1 to N do

For k =1 to N do

seqscore[j][k][r] = $max_{i=1,n}$(seqscore[i][j][r-1] *$prob(t_k/t_j,t_i)$)*$prob(w_r/t_k)$

backptr[j][k][r] = ค่า i ที่ทำให้ค่าสมการที่ผ่านมาเป็นค่าที่มากที่สุด

Sequence identification step

C[M] = k and c[M-1] =j โดยที่ j และk นั้นที่ทำให้ Seqscore[j][k][M] มีค่ามากที่สุด

For i= M-2 to 1 do

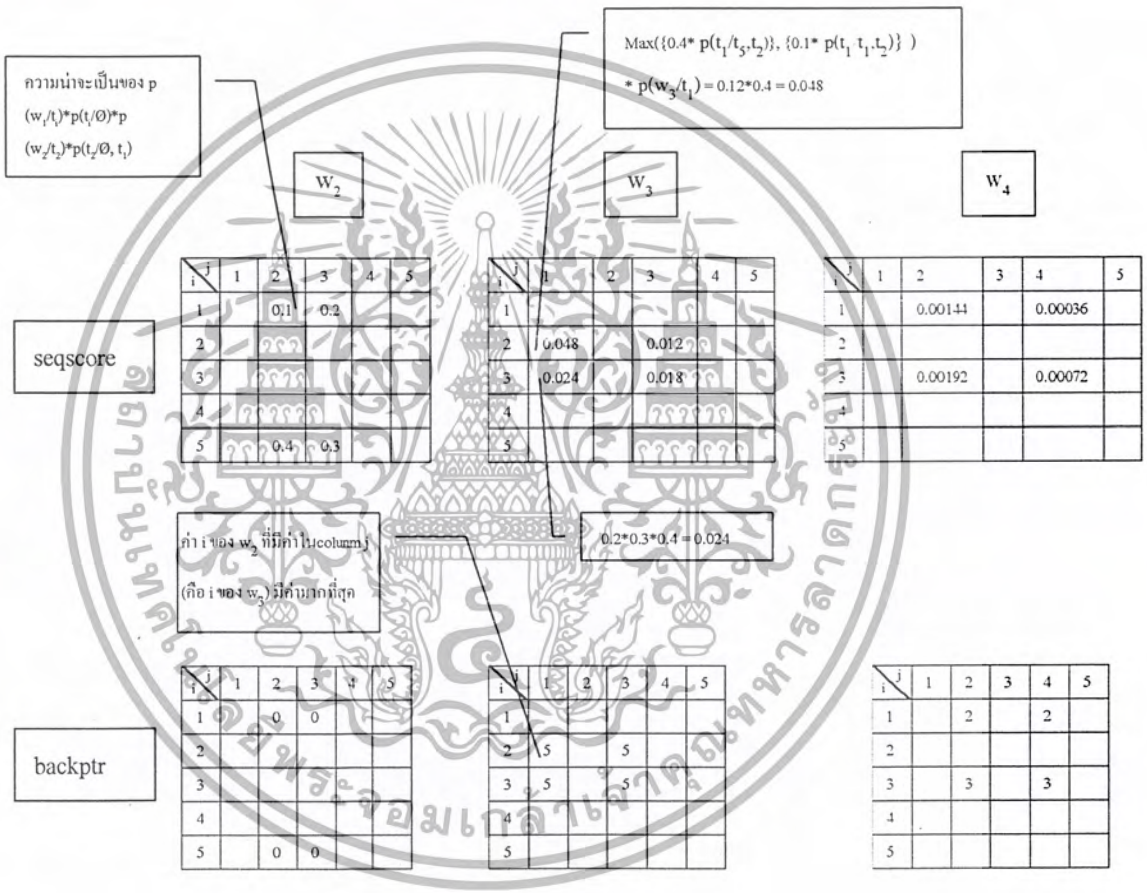
C[i] = backptr [C[i+1]] [C[i+2]] [i+2]

รูปที่ 5-6 ขั้นตอนการนำวิธีวิเทอร์บิมาใช้กับการตัดคำภาษาไทย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวอย่าง สมมติว่ามีหน้าที่คำทั้งหมด 5 หมวดคำ และมีคำในประโยคทั้งหมด 4 คำ ซึ่งแต่ละคำมีการกำกับหมวดคำตามภาพที่แสดงด้านล่าง

W_1	W_2	W_3	W_4	$P(t_1/t_1,t_2) = 0.2$ $P(t_1/t_1,t_3) = 0.3$ $P(t_1/t_5,t_2) = 0.3$ $P(t_1/t_5,t_3) = 0.2$	$P(t_3/t_1,t_2) = 0.6$ $P(t_3/t_1,t_3) = 0.1$ $P(t_3/t_5,t_2) = 0.1$ $P(t_3/t_5,t_3) = 0.2$	$P(t_2/t_2,t_1) = 0.3$ $P(t_2/t_2,t_3) = 0.4$ $P(t_2/t_3,t_1) = 0.1$ $P(t_2/t_3,t_3) = 0.2$	$P(t_4\ t_2,t_1) = 0.4$ $P(t_4\ t_2,t_3) = 0.3$ $P(t_4\ t_3,t_1) = 0.1$ $P(t_4\ t_3,t_3) = 0.2$	$P(w_3/t_1) = 0.4$ $P(w_3/t_3) = 0.3$ $P(w_4\ t_2) = 0.1$ $P(w_4/t_4) = 0.2$
t_1	t_2	t_1	t_2					
t_5	t_3	t_3	t_4					



รูปที่ 5-7 แสดงตัวอย่างการทำวิเทอร์บิร่วมกับการตัดภาษาไทย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Initialization step

1. ทำการกำหนดค่าเริ่มต้นใน array ของ w2 โดยให้มีค่าเท่ากับ

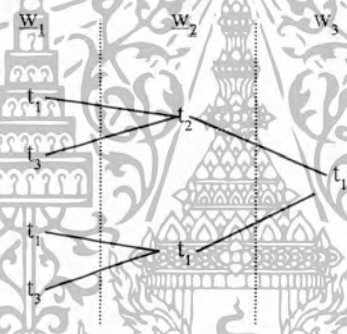
$$seqscore[i][j][2] = prob(w_1/t_i)*prob(t_i/ø)*prob(w_2/t_j)*prob(t_j/t_i,ø)$$

โดย $prob(w_1/t_i)*prob(t_i/ø)$ หมายถึง ค่าความน่าจะเป็นที่คำ w_1 มีหมวดคำเป็น t_i และความน่าจะเป็นที่หมวดคำ t_i อยู่ต้นประโยค และ $prob(w_2/t_j)*prob(t_j/ø,t_i)$ หมายถึงความน่าจะเป็นที่ w_2 มีหมวดคำเป็น t_j และความน่าจะเป็นที่ t_j ตามหลัง t_i ที่อยู่ต้นประโยค

2. ใส่ค่า backptr[i][j][2] = 0 เพื่อให้รู้ว่าเป็นจุดเริ่มต้น

Iteration step

จะทำการคำนวณไตรแกรม จากตัวอย่าง ทำการพิจารณาต่อที่ w_3 ซึ่งมีหมวดคำที่เป็นไปได้ 2 หมวดคือ t_1 และ t_3 โดยทำการพิจารณาที่หมวดคำ t_1 ก่อนซึ่งมีรูปแบบของหน้าที่คำมายัง t_1 ได้ 4 รูปแบบคือ



เราไม่รู้ว่าเส้นทางใดดีที่สุดระหว่าง $t_1 t_2 t_1$, $t_1 t_3 t_1$, $t_3 t_2 t_1$ และ $t_3 t_3 t_1$ ดังนั้นเราจึงดูความน่าจะเป็นของหมวดคำก่อนหน้าที่จะมาถึง w_3 ว่าค่าใดมีค่ามากที่สุด โดยที่เรารู้ค่า j ของ array w_2 ซึ่งก็คือหมวดคำก่อนหน้า w_3 (ค่า i ใน array w_3) จึงทำหาค่า i ที่ทำให้ column j ของ array w_2 ที่ทำให้มีค่า $seqscore[i][j][2] * prob(t_k/t_i,t_j)$ มากที่สุด ถ้าเราพิจารณาหมวดคำ t_2 ของ w_2 (ค่า i=2 ใน array ของ w_3 แต่เป็นค่า j=2 ใน array ของ w_2) จะเห็นได้ว่า ในคอลัมน์ที่ 2(ค่า j) ของ array w_2 มีค่า 0.1 ที่ i = 1 และ มีค่า 0.4 ที่ i = 5 ดังนั้นเราจะต้องหาค่า $seqscore[i][2][2] * prob(t_k/t_i,t_2)$ ที่มากที่สุดระหว่างค่า i ทั้งสอง ของ w_3 ดังนั้นสุดท้ายแล้วจะได้ค่าของ Backptr[2][1][3] = 5 เมื่อทำจนจบประโยคแล้วจะทำการ Backtracking โดยใช้ Backptr เช่น Backptr[2][1][3] = 5 (ซึ่งหากค่านี้เป็นเส้นทางที่มีค่าความน่าจะเป็นมากที่สุด) จะหมายถึง คำที่ 3 มีหน้าที่คำเป็น 1 คำที่ 2 มีหน้าที่คำเป็น 2 และ คำที่ 1 มีหน้าที่คำเป็น 5 ซึ่งค่า C จะใช้ในการเก็บเส้นทางที่ได้จากการ Backtrack

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 6
ขั้นตอนในการวัดประสิทธิภาพ

6.1 สร้างค่าของอินพุตและคำตอบในการเปรียบเทียบ

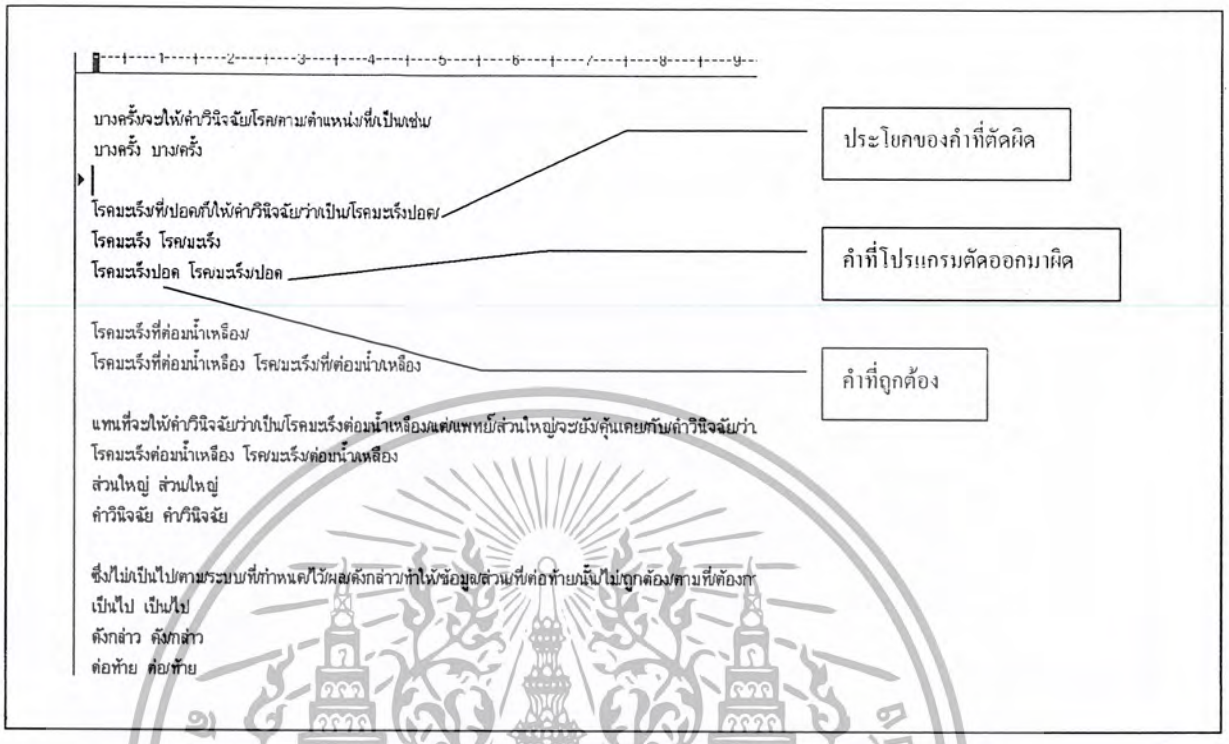
อินพุตของการวัดประสิทธิภาพนั้นเราได้นำมาจาก 20% ของคลังข้อความซึ่งมีประมาณ 4881 ประโยค โดยที่ในแต่ละประโยคจะมีทั้งตัวอักษรภาษาไทย ภาษาอังกฤษ เครื่องหมายพิเศษ เครื่องหมายเว้นวรรค ซึ่งในส่วนนี้จะนำประโยคมาสร้างเป็นไฟล์ที่ใช้เป็นอินพุตในการทดสอบโปรแกรมตัดคำต่าง ๆ ส่วนคำที่ตัดไว้แล้วในคลังข้อความของประโยคนั้นๆ จะนำมาเป็นไฟล์ที่ใช้เป็นคำตอบในการเปรียบเทียบความถูกต้อง โดยจะถือคำตอบในไฟล์นี้เป็นคำตอบที่ถูกต้อง 100 %

6.2 วัดประสิทธิภาพด้วย Framework ที่สร้างขึ้น

การทำงานของ framework สามารถแบ่งได้เป็นขั้นตอนดังนี้

1. รับ path และ ชื่อของโปรแกรมที่ต้องการวัดประสิทธิภาพ นำมาเก็บไว้ใน name.txt
2. อ่านชื่อของ โปรแกรม ใน file name.txt ขึ้นมาเพื่อ run และเริ่มการจับเวลาเพื่อวัดประสิทธิภาพด้านเวลา
3. ทำการหาจำนวน byte ทั้งหมดที่อยู่ในmemory เพื่อวัดประสิทธิภาพด้านการใช้ทรัพยากร
4. นำเอาต์พุตที่ได้จากโปรแกรมการตัดคำไปทำการเปรียบเทียบกับไฟล์ที่ใช้เป็นคำตอบที่กล่าวไว้ด้านบน จากนั้นคำนวณหาค่าความถูกต้องออกมา ตามสมการที่ 3-2 โดยมีการแสดงจำนวนคำที่ถูก จำนวนคำทั้งหมดที่ตัดได้จากโปรแกรม จำนวนคำทั้งหมดที่ตัดไว้ในคลังข้อความ
5. รายละเอียดของประโยคที่ตัดผิดนั้นจะอยู่ใน ไฟล์ detail_ <ชื่อโปรแกรมที่นำมาตัดคำ>.txt

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 6-1 แสดงตัวอย่างของไฟล์ detail_dict.txt

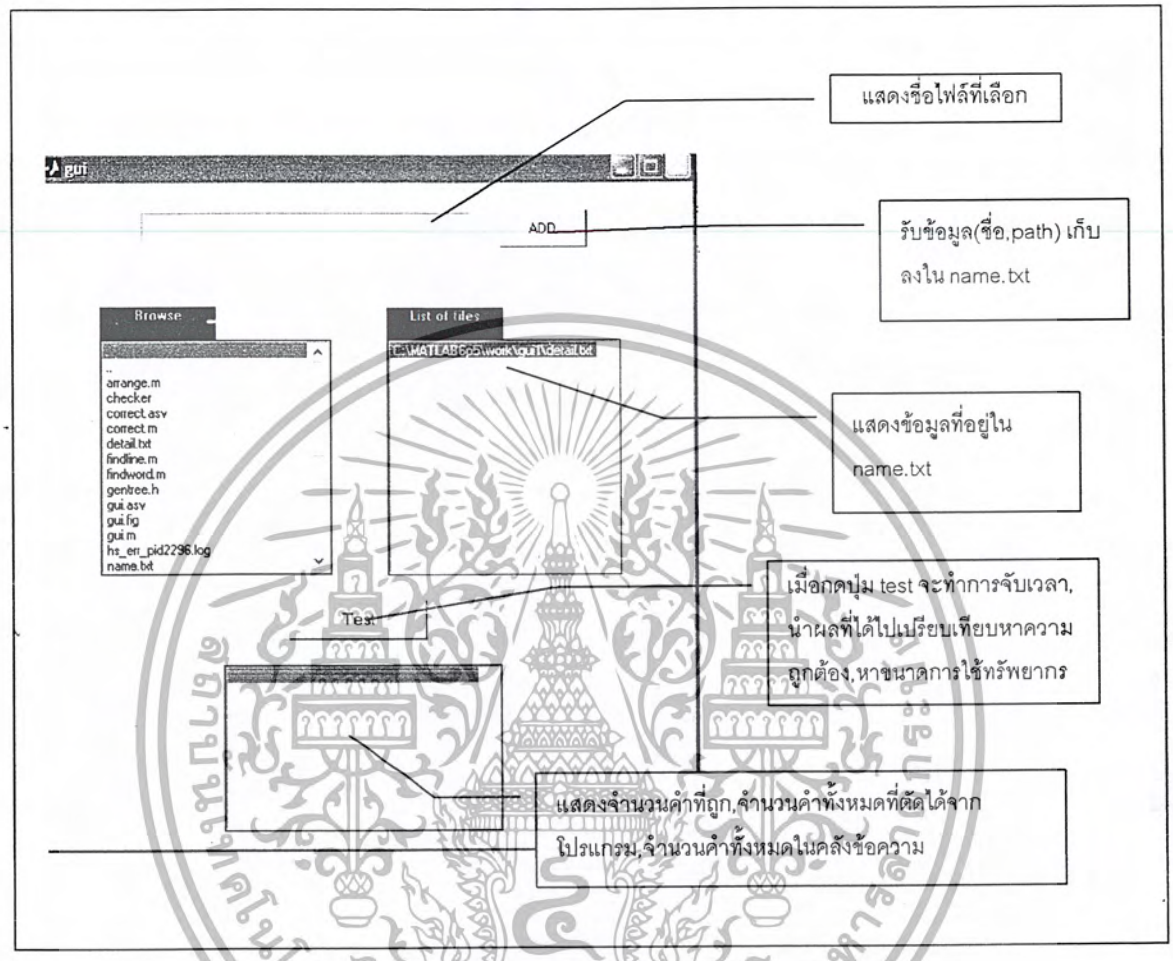
6.2.1 เงื่อนไข

1. ไฟล์ที่เป็นเอาต์พุต (output) ของโปรแกรมที่ตัดคำนั้นของต้องเป็นชื่อ "out_ชื่อไฟล์ที่นำวัดประสิทธิภาพ" เช่น มีไฟล์ dict.dll ซึ่งเป็นโปรแกรมการตัดคำโดยใช้พจนานุกรม จะต้องมีการเขียนผลการตัดคำลงในไฟล์ชื่อ out_dict.txt เท่านั้น
2. รายละเอียดของคำที่ตัดได้ผิดนั้นจะอยู่ในไฟล์ชื่อ "detail_ชื่อไฟล์ที่นำวัดประสิทธิภาพ" เช่น ในตัวอย่างของข้อที่1 เราจะได้ไฟล์ที่เก็บรายละเอียดของคำที่ตัดผิดชื่อ detail_dict.txt
3. ไฟล์ที่จะนำมาเป็นอินพุตในการตัดคำนั้น ให้มีชื่อว่า input.txt

6.2.2 ขั้นตอนการใช้ framework

1. เลือกชื่อไฟล์ที่ต้องการวัดประสิทธิภาพใน listbox ด้านซ้าย ชื่อไฟล์ที่ต้องการจะปรากฏใน edit box ด้านบน
2. กดปุ่ม ADD เพื่อเพิ่มไฟล์ที่ต้องการวัดประสิทธิภาพลงใน name.txt
3. กดปุ่ม Test เพื่อวัดประสิทธิภาพ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 6-2 แสดงการทำงานในส่วนต่าง ๆ ของ framework

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

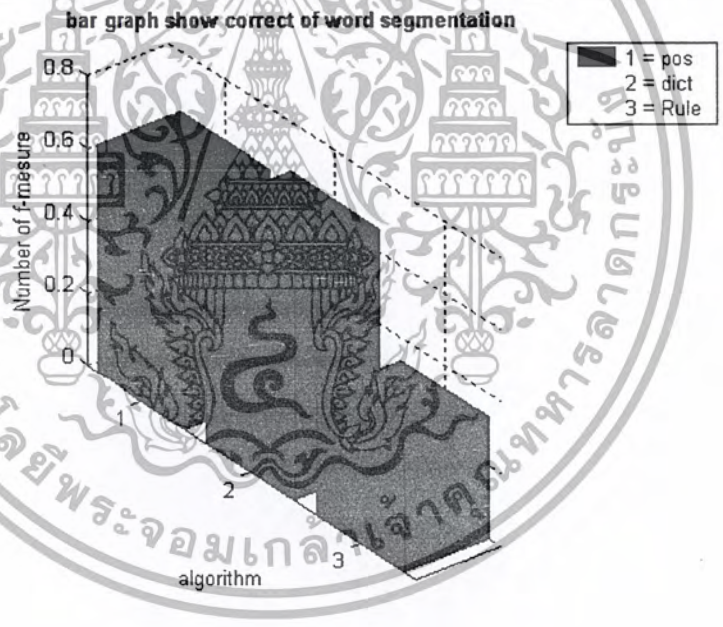
บทที่ 7

ผลการทดลอง

ในบทนี้ได้ทำการรวบรวมและเปรียบเทียบข้อมูลสถิติ ที่เกิดจากการทดลองวัดประสิทธิภาพของโปรแกรมตัดคำภาษาไทยจาก framework ที่ได้สร้างขึ้น โดยที่ framework จะทำการคำนวณและสร้างกราฟต่าง ๆ ให้โดยอัตโนมัติ ซึ่งมีผลการทดลองดังนี้

7.1 เปรียบเทียบด้านความถูกต้องของคำ

จากการทดลองทำให้ได้ผลจำนวนคำที่โปรแกรมตัดคำแบบต่าง ๆ ตัดคำออกมาได้ดังตารางที่ 7-1



รูปที่ 7-1 แสดงการเปรียบเทียบโดยค่า F - Measure

มาตรวัดประสิทธิภาพ / โปรแกรมการตัดคำ	ความถูกต้อง (f-measure)	จำนวนคำที่ตัดได้ถูกต้อง	จำนวนคำที่ตัดได้ทั้งหมด
การตัดคำด้วยคลังข้อความ	0.63064	38859	71084
การตัดคำด้วยพจนานุกรม	0.66231	39823	68103
การตัดคำด้วยกฎ	0.33803	22297	79772

ตารางที่ 7-1 ค่าความถูกต้อง, จำนวนคำที่ตัดได้ถูกต้อง, จำนวนคำที่ตัดได้ออกมาทั้งหมด

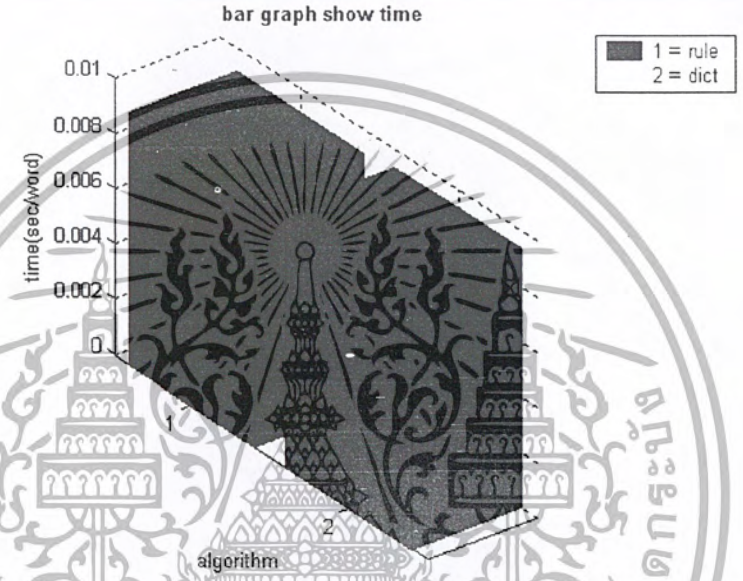
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

7.2 เปรียบเทียบประสิทธิภาพเชิงความเร็ว

จากการทดลองทำให้ได้ผลจำนวนคำที่โปรแกรมตัดคำแบบต่าง ๆ ตัดคำออกมาได้ดังตารางที่ 7-2 เนื่องจากการตัดคำคลังข้อความใช้เวลานานมาก เช่น ในประโยค

"โรคมะเร็งที่ปอดก็ให้คำวินิจฉัยว่าเป็นโรคมะเร็งปอด"

ใช้เวลาตัดประมาณ 3 นาที จึงไม่ขอนำมาเปรียบเทียบ



รูปที่ 7-2 แสดงการเปรียบเทียบความเร็วเฉลี่ย

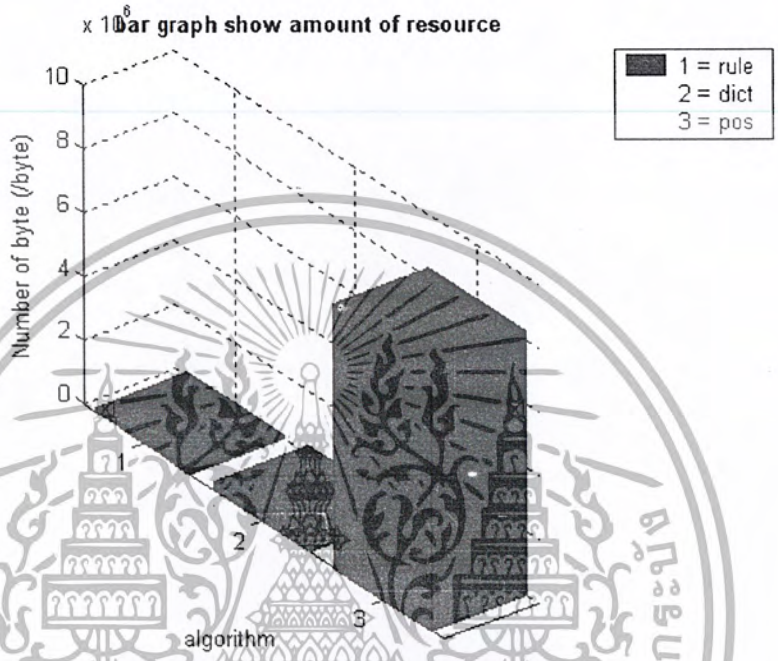
มาตรวัดประสิทธิภาพ / โปรแกรมการตัดคำ	ความเร็ว
การตัดคำด้วยพจนานุกรม	0.0097224 sec/word
การตัดคำด้วยกฎ	0.0077636 sec/word

ตารางที่ 7-2 ค่าความเร็วเฉลี่ยในการตัดคำ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

7.3 เปรียบเทียบประสิทธิภาพการใช้ทรัพยากร

การใช้ทรัพยากรของโปรแกรมตัดคำแบบต่างๆมีผลมาจากขนาดของพจนานุกรม และคลังข้อความที่ใช้งานด้วย ซึ่งขนาดของการใช้ทรัพยากรได้แสดงไว้ดังตารางที่ 7-3



รูปที่ 7-3 แสดงการเปรียบเทียบการใช้ทรัพยากร

มาตรวัดประสิทธิภาพ / โปรแกรมการตัดคำ	ทรัพยากร
การตัดคำด้วยกฎ	123210 byte
การตัดคำด้วยพจนานุกรม	295027 byte
การตัดคำด้วยคลังข้อความ	8317641 byte

ตารางที่ 7-3 จำนวนการใช้ทรัพยากรของโปรแกรมการตัดคำแบบต่างๆ

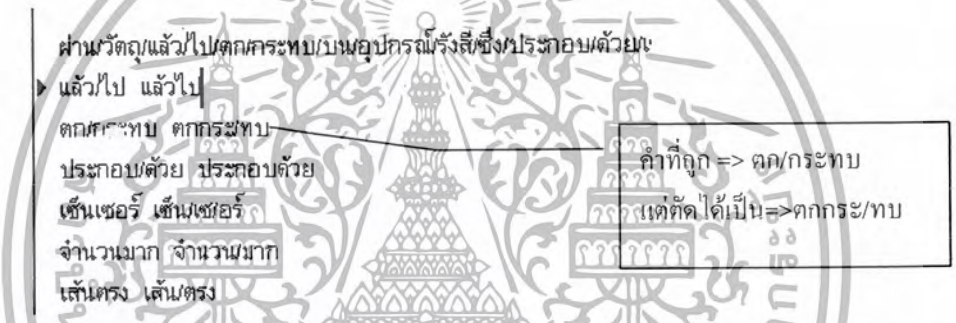
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 8
การวิเคราะห์ผลจากการตัดคำแล้ว

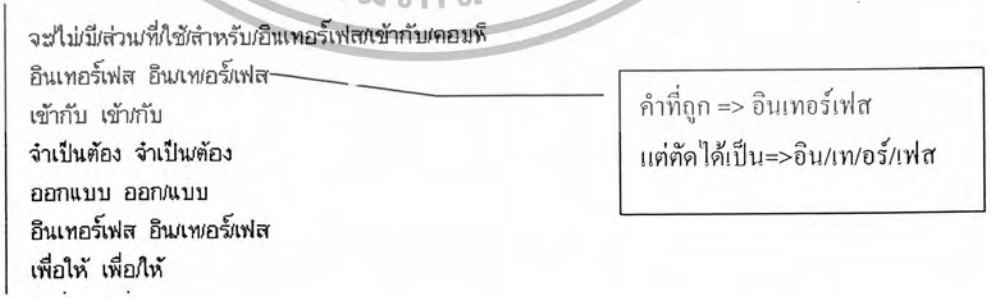
ผลจากการวัดประสิทธิภาพในด้านต่างๆสามารถวิเคราะห์สาเหตุของการตัดคำผิดโดยแยกตามอัลกอลิทึมที่ใช้ ได้ดังต่อไปนี้

8.1 โปรแกรมการตัดคำด้วยพจนานุกรม

สาเหตุที่เกิดการตัดโดยพจนานุกรมมีความผิดพลาดโดยวิเคราะห์ผลลัพธ์ที่ผิดในการตัดคำ มีดังต่อไปนี้

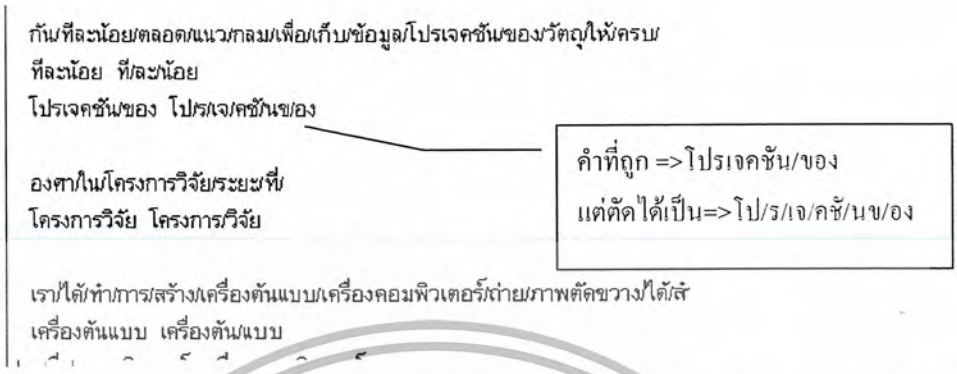


จากตัวอย่างด้านบนการตัดคำว่า "ตกกระทบ" ไม่ถูกต้องนั้นเกิดจากการที่การตัดคำโดยใช้พจนานุกรมพบว่าคำว่า "ตกกระ" เป็นคำที่ยาวที่สุดซึ่งทำให้คำข้างหลังสามารถตัดได้ ดังนั้นวิธีการนี้จึงตัดคำได้เป็น "ตกกระ"-"ทบ" โดยที่ทั้ง 2 คำนี้ไม่มีความหมายเกี่ยวข้องกันทางด้านไวยากรณ์เลย ซึ่งคำกำกวมในลักษณะนี้จะทำให้ประสิทธิภาพการตัดคำประเภทนี้ค่อนข้างต่ำ



การตัดคำโดยใช้พจนานุกรมนั้นหากไม่พบคำในพจนานุกรมจะทำให้ไม่สามารถตัดคำนั้นได้ถูกต้อง ซึ่งคำที่มักจะไม่พบในพจนานุกรมได้แก่ ชื่อคน คำทับศัพท์ภาษาต่างประเทศ ดังตัวอย่างด้านบน

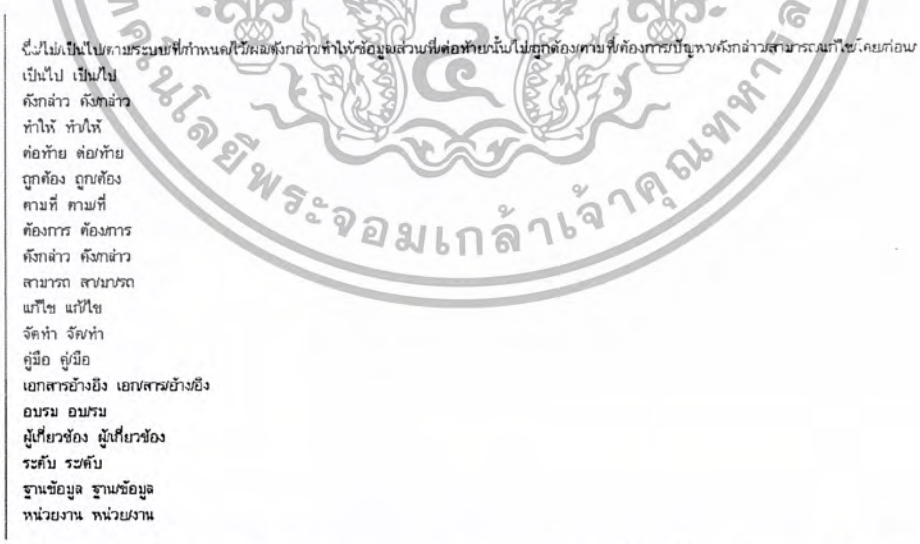
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



จากตัวอย่างด้านบน สาเหตุอีกอย่างที่ทำให้การตัดคำผิดพลาดก็คือ ผลกระทบจากคำที่ไม่พบในพจนานุกรม อาจส่งผลถึงคำใกล้เคียง เช่น โปรเจคชั่น/ของ ตัดได้เป็น โป/ร/เจ/คชั่/นข/อง จะเห็นว่าคำว่าของจะถูกตัดผิดไปด้วย ซึ่งเป็นผลกระทบจากคำข้างหน้า คือคำว่า "โปรเจคชั่น" เนื่องจากคำว่า "โปรเจคชั่น" นั้นเป็นคำที่ไม่มีในพจนานุกรม

8.2 โปรแกรมตัดคำด้วยคลังข้อความ

จากผลการทดลอง สังเกตได้ว่าลักษณะการตัดผิดของการตัดคำโดยใช้คลังข้อความ ส่วนใหญ่จะมีการผิดติดต่อกันเกือบทั้งประโยค ดังภาพที่แสดงไว้ด้านล่าง



สาเหตุที่ทำให้เกิดการตัดคำผิดพลาดอาจเกิดจากการตัดคำที่ผิดตั้งแต่คำแรกๆ ซึ่งจะส่งผล ให้คำด้านหลังเกิดการผิดพลาดไปด้วยเนื่องจากการตัดคำประเภทนี้คำแต่ละคำมีการเกี่ยวข้องกัน ทั้งนี้การตัดคำที่ผิดพลาดอาจมีสาเหตุมาจาก ไม่เคยเจอความสัมพันธ์หมวดคำในรูปแบบนั้น ๆมาก่อน และการกำกับหน้าที่คำที่ผิดพลาด ซึ่งความผิดพลาดนี้อาจส่งผลให้การตัดคำของคำต่อ ๆไปจำนวนมากผิดพลาดได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

8.3โปรแกรมการตัดคำด้วยกฎ

การตัดคำโดยใช้กฎนั้น เหมาะกับการตัดพยางค์มากกว่ามาตัดคำ เนื่องจากคำแต่ละคำไม่ได้มีรูปแบบตายตัวเหมือนเช่นพยางค์ ซึ่งธรรมชาติของภาษาไทยหากคำ 2 คำสามารถรวมเป็นกลุ่มคำได้ ก็มีความเป็นไปได้สูงที่คำที่ถูกต้องจะเป็นกลุ่มคำที่เกิดจากการรวมคำ2 คำนั้น (นัฐวฒิ ไชยเจริญ [12])ทำให้การใช้วิธีนี้ไม่ค่อยมีประสิทธิภาพมากนัก แต่กฎก็สามารถตัดคำที่เกิดจากการทับศัพท์ภาษาต่างประเทศได้ เช่น คำว่า "รังสีเอ็กซ์" ตัดได้เป็น "รัง"-"สี"-"เอ็กซ์"

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 9
บทสรุปและข้อเสนอแนะ

9.1 สรุปประสิทธิภาพของโปรแกรมการตัดคำ

โปรแกรมการตัดคำแบบพจนานุกรม มีความเร็วในการตัดคำมากที่สุด ในขณะที่การตัดคำโดยใช้คลังข้อความมีความเร็วในการตัดคำน้อยที่สุด

ในเรื่องของความถูกต้อง การตัดคำโดยใช้กฎมีความถูกต้องน้อยที่สุด การตัดคำโดยพจนานุกรมสามารถตัดคำได้มีความถูกต้องมากที่สุด

การตัดคำโดยใช้คลังข้อความมีการใช้ทรัพยากรมากกว่าการตัดคำอีก 2 แบบที่เหลือมาก โดยการตัดคำแบบใช้กฎใช้ทรัพยากรน้อยที่สุด

9.2 ความเหมาะสมของการประยุกต์ใช้โปรแกรมตัดคำ

จากผลวิจัยนี้สามารถทำให้เห็นแนวทางในการพิจารณาในการนำโปรแกรมไปใช้กับงานประเภทต่าง ๆ ให้เหมาะสม สำหรับการพิจารณาในเบื้องต้นมีดังนี้

1. โปรแกรมตัดคำโดยคลังข้อความเหมาะสมกับงานที่ต้องการความถูกต้องด้านไวยกรณ์สูง เช่น งานแปลเอกสาร เนื่องจากมีการใช้หน้าที่ของคำมาร่วมพิจารณาในการตัดคำ ทำให้ผลการตัดคำที่ได้มีความถูกต้องทางด้านไวยกรณ์มากกว่ารูปแบบอื่นๆ

2. โปรแกรมการตัดคำโดยใช้พจนานุกรม สามารถนำมาประยุกต์ใช้ได้ในหลายงาน เนื่องจากทั้งความเร็วและความถูกต้องในการตัดคำมากกว่าแบบอื่นๆ จึงอาจนำไปใช้ในงานประเภทจัดรูปแบบเอกสาร และ การตรวจสอบคำสะกด ได้

3. โปรแกรมการตัดคำด้วยกฎ แม้ว่าจะมีความถูกต้องต่ำ แต่มีความสามารถในการตัดคำที่ไม่เคยพบมาก่อน เช่น คำทับศัพท์ภาษาต่างประเทศ ได้ถูกต้องมากกว่าการตัดคำในรูปแบบอื่นๆ ดังนั้นหากนำไปใช้ร่วมกับโปรแกรมการตัดคำแบบอื่นจะสามารถเพิ่มประสิทธิภาพในการตัดคำของโปรแกรมเหล่านั้นเพิ่มมากยิ่งขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

9.3 ประเภทเอกสารกับผลของการเปรียบเทียบประสิทธิภาพการตัดคำ

เนื่องจากเอกสารที่ใช้ทดลองนั้นส่วนมากจะเกี่ยวข้องกับเรื่องของเทคโนโลยี ซึ่งมักจะมีคำที่เกิดจากการทับศัพท์ภาษาต่างประเทศเป็นจำนวนมาก ดังนั้นวิธีการตัดคำแบบต่าง ๆ อาจจะตัดคำออกมาได้ผิดพลาดมาก เนื่องจากเป็นคำที่ไม่รู้จัก ดังนั้นสามารถสรุปได้ว่าประเภทของเอกสารมีผลต่อความถูกต้องของการตัดคำ นั่นคือหากเอกสารมีคำพื้นฐานเป็นส่วนใหญ่ความถูกต้องของการตัดคำก็จะมากขึ้น ในทางตรงกันข้ามหากเอกสารมีคำที่เกิดจากการทับศัพท์ภาษาต่างประเทศ หรือคำที่เฉพาะเจาะจางของแต่ละสาขาเป็นส่วนใหญ่ ความถูกต้องของการตัดคำก็จะลดลง

9.4 ปัญหาต่าง ๆที่พบในงานวิจัย

9.4.1 งานวิจัยได้พบปัญหาเกี่ยวกับการตัดคำโดยใช้คลังข้อความ เนื่องจากเวลาที่ใช้ในการตัดคำนั้นใช้เวลาค่อนข้างนานมาก ดังนั้นผู้วิจัยจึงไม่ได้เปรียบเทียบ ในด้านเวลา

9.4.2 เนื่องจากเอกสารที่นำมา (คลังข้อความ) มีจุดผิดพลาดในการตัดคำค่อนข้างมาก และเนื้อหาที่ไม่หลากหลาย มักจะเกี่ยวข้องกับเทคโนโลยี ซึ่งในการใช้งานจริงเราจะเจอคำศัพท์ประเภทนี้น้อยมาก ทำให้ประสิทธิภาพการตัดคำในการใช้งานจริงอาจไม่ได้ผลตามที่ได้ทำการทดลอง

9.5 ข้อเสนอแนะและงานวิจัยที่สามารถทำเพิ่มเติมต่อจากงานวิจัยนี้

9.5.1 งานในด้านการพัฒนาคลังข้อความก็เป็นสิ่งหนึ่งที่น่าจะพัฒนาเพิ่มเนื่องจากเนื้อสาระที่มีนั้นค่อนข้างที่จะเกี่ยวกับเรื่องของเทคโนโลยีเป็นส่วนใหญ่ ดังนั้นอาจมีการพัฒนาเพิ่มข้อมูลลงในคลังข้อความในเนื้อหาสาระเรื่องอื่น ๆ ก็เป็นสิ่งที่ควรพัฒนาต่อ

9.5.2 ศึกษาการใช้งานการตัดคำแต่ละแบบ เพื่อความเหมาะสมในการเลือกใช้งาน

9.5.3 การพัฒนารูปแบบในการตัดคำใหม่ ๆ เพื่อความถูกต้องมากยิ่งขึ้น

9.5.4 ควรเพิ่มเติมคำย่อ คำเฉพาะที่พบบ่อย ลงในพจนานุกรม เพื่อความถูกต้องมากยิ่งขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก ก

ความหมายของคำย่อและสัญลักษณ์ต่าง ๆ ในคลังข้อความ Orchid

Mark-up	Description
%TTitle:	Title of the document written in Thai.
%ETitle:	Title of the document written in English.
%TAuthor:	Author's name written in Thai.
%EAuthor:	Author's name written in English.
%TInbook:	Title of the book where the document exists, written in Thai.
%EInbook:	Title of the book where the document exists, written in English.
%TPublisher:	Publisher of the book, written in Thai.
%EPublisher:	Publisher of the book, written in English.
%Page:	Page number or the range of pages of the document.
%Year:	Published year (A.D.).
%File:	File number of the document. A long document may be separated into a number of files.

Mark-up	Description
#P[number]	Paragraph number of a text. The number in the bracket is shown in a sequence within a text.
#[number]	Sentence number of a paragraph. The number in the bracket is shown in a sequence within a paragraph.

Mark-up	Description
\\	Line break symbol.
//	Sentence break symbol.
/[POS]	Tag marker for the appropriate POS of a word.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

อักขระพิเศษ

Special characters	Defined strings	Special characters	Defined strings
	<space>	/	<slash>
!	<exclamation>	:	<colon>
"	<quotation>	;	<semi_colon>
#	<number>	<	<less_than>
$	<dollar>	=	<equal>
%	<percent>	>	<greater_than>
&	<ampersand>	?	<question_mark>
'	<apostrophe>	@	<at_mark>
(	<left_parenthesis>	[	<left_square_bracket>
)	<right_parenthesis>	]	<right_square_bracket>
*	<asterisk>	^	<circumflex_accent>
+	<plus>	_	<low_line>
,	<comma>	{	<left_curly_bracket>
-	<minus>	}	<right_curly_bracket>
.	<full_stop>	~	<tilde>

ตารางความหมายของคำย่อของหมวดคำต่าง ๆ

No.	POS	Description	Example
1	NPRP	Proper noun	วินโดวส์ 95, โคโรน่า, โอ๊ค, พระอาทิตย์
2	NCNM	Cardinal number	หนึ่ง, สอง, สาม, 1, 2, 3
3	NONM	Ordinal number	ที่หนึ่ง, ที่สอง, ที่สาม, ที่1, ที่2, ที่3
4	NLBL	Label noun	1, 2, 3, 4, ก, ข, a, b
5	NCMN	Common noun	หนังสือ, อาหาร, อาคาร, คน
6	NTTL	Title noun	ดร., พลเอก
7	PPRS	Personal pronoun	คุณ, เขา, ฉัน
8	PDMN	Demonstrative pronoun	นี่, นั่น, ที่นั่น, ที่นี่
9	PNTR	Interrogative pronoun	ใคร, อะไร, อย่างไร
10	PREL	Relative pronoun	ที่, ซึ่ง, อัน, ผู้
11	VACT	Active verb	ทำงาน, ร้องเพลง, กิน
12	VSTA	Stative verb	เห็น, รู้, คือ
13	VATT	Attributive verb	อ้วน, ดี, สวย
14	XVBM	Pre-verb auxiliary, before negator "ไม่"	เกิด, เกือบ, กำลัง
15	XVAM	Pre-verb auxiliary, after negator "ไม่"	ค่อย, น่า, ได้
16	XVMM	Pre-verb, before or after negator "ไม่"	ควร, เคย, ต้อง
17	XVBB	Pre-verb auxiliary, in imperative mood	กรุณา, จง, เชิญ, อย่า, ห้าม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

18	XVAE	Post-verb auxiliary	ไป, มา, ขึ้น
19	DDAN	Definite determiner, after noun without classifier in between	นี่, นั่น, โน่น, ทั้งหมด
20	DDAC	Definite determiner, allowing classifier in between	นี้, นั้น, โน้น, นู้น
21	DDBQ	Definite determiner, between noun and classifier or preceding quantitative expression	ทั้ง, อีก, เพียง
22	DDAQ	Definite determiner, following quantitative expression	พอดี, ถ้วน
23	DIAC	Indefinite determiner, following noun; allowing classifier in between	ไหน, อื่น, ต่างๆ
24	DIBQ	Indefinite determiner, between noun and classifier or preceding quantitative expression	บาง, ประมาณ, เกือบ
25	DIAQ	Indefinite determiner, following quantitative expression	กว่า, เศษ
26	DCNM	Determiner, cardinal number expression	หนึ่งคน, เสือ 2 ตัว
27	DONM	Determiner, ordinal number expression	ที่หนึ่ง, ที่สอง, ที่สุดท้าย
28	ADVN	Adverb with normal form	เก่ง, เร็ว, ช้า, สม่ำเสมอ
29	ADVI	Adverb with iterative form	เร็วๆ, เสมอๆ, ช้าๆ
30	ADVP	Adverb with prefixed form	โดยเร็ว
31	ADVS	Sentential adverb	โดยปกติ, ธรรมดา
32	CNIT	Unit classifier	ตัว, คน, เล่ม
33	CLTV	Collective classifier	คู่, กลุ่ม, ฝูง, เชิง, ทาง, ด้าน, แบบ, รุ่น
34	CMTR	Measurement classifier	กิโลกรัม, แก้ว, ช้อนโต๊ะ
35	CFQC	Frequency classifier	ครั้ง, ที่ซา
36	CVBL	Verbal classifier	ม้วน, มัด
37	JCRG	Coordinating conjunction	และ, หรือ, แต่
38	JCMP	Comparative conjunction	กว่า, เหมือนกับ, เท่ากับ
39	JSBR	Subordinating conjunction	เพราะว่า, เนื่องจาก, ที่, แม้ว่า, ถ้า
40	RPRE	Preposition	จาก, ณ, ของ, ใต้, บน
41	INT	Interjection	โอ้ย, โอ้, เออ, เฮ้, อ้อ
42	FIXN	Nominal prefix	การทำงาน, ความสนุกสนาน
43	FIXV	Adverbial prefix	อย่างเร็ว
44	EAFF	Ending for affirmative sentence	จ๊ะ, จ้ะ, ค่ะ, ครับ, นะ, น่า, เถอะ
45	EITT	Ending for interrogative sentence	หรือ, เหรอ, ไหม, มั้ย
46	NEG	Negator	ไม่, มิได้, ไม่ได้, มิ
47	PUNC	Punctuation	(,), “, „ ;

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก ข

กฎของคุณสุรินทร์ จรรยาพรพงษ์

+----------+

ความหมายของคำย่อที่ใช้

+----------+

c	คือ อักษรนำ
s	คือ ตัวสะกด
t	คือ วรรณยุกต์
g	คือ อักษรที่มีการันต์ต่อท้ายได้
*	คือ ขอบเขตของพยางค์
..	คือ ตัวอักษรใดใด

+----------+

+----------+

รูปแบบของคุณลักษณะ

+----------+

R	คือ รูปแบบกฎที่ตายตัว
N	คือ รูปแบบกฎที่อาจตัวได้ถูกหรือไม่ถูกต้องก็ได้
X	คือ กรณียกเว้น

+----------+

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-1

รูปแบบสระ - า	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	า s g g ์ *	ศาสตร์ พราหมณ์	R
	า s g ์ *	การณ์ ปราชญ์ กาพย์	R
	า g ์ *	การ์ บาร์	R
	า g ์ s *	การ์ด ฟาร์ม	X
	า s ิ *	ชาติ ญาติ พยาธิ	X
	า s *	ธาตุ	X
	า s s *	มาตร มารถ มารค	X,N
	า s *	การ มาก ปาก	N
	t า s*	ข้าม บ้าน	N
	า*	กา มา ขา	N
	t า*	ข้า ท่า	N
ขอบเขตหน้า	* c า	กา บาร์	N
	* c c า	มหา สถาน	N
	* c t า	น้า ย้าย	N
	* c c t า	หน่าย คล้าย	N

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางแสดงอักษรควบของรูปสระ : - า

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	อ ท ห ส ข ช
ร	ก อ ท ต ส ป ค พ บ ฆ
อ	ส
ง	ห ส ผ
ม	ห ส
ย	อ ห ส พ บ
ว	ค ท ห ส ค ข ผ ถ
ล	ก อ ท ต ห ส ป ค พ ข ผ ถ ฉ
ท	ป
ต	ส
ห	ม ท ส
บ	ส ช
น	ส
ถ	ส
ญ	ห
ฦ	ม ส
ฦ	ช

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-2

รูปแบบสระ ิ	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	s g g ์ *	รินทร์ มินทร์	R
	s g ์ *	สิทธิ์	R
	s g ์ *	สิงห์ ดิษฐ์	R
	s s *	จิตร มิตร	X,N
	s *	คิด จิต ขวิด	N
	*	วิ คิ ปฏิ	N
	t s *	สิ่ง ทิ้ง ทิ่ม	R
ขอบเขตหน้า	* c	คิด จิบ	N
	* c c	หยิ่ง ปฏิ	N
	* c c c	อมรินทร์ สปริง อคติ	X,N

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางแสดงอักษรควบของรูปสระ : -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	อ ว ห ส ช ถ
ร	ก อ ห ป บ จ ค พ ภ
ง	ห
ม	อ ท ห ส
ย	ห
ว	อ ท ห ส ค ข
ล	ก อ ค ห ส ป พ ผ
ต	ส
ห	อ
ช	ว
ถ	ส
ณ	ค พ
ธ	อ
ญ	ห
ภ	อ
ฦ	ป

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-3

รูปแบบสระ -	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	g ์ *	คีย์ รีย์ ทรีย์	R
	s *	อีก รีบ มีด	N
	t *	ลี่ นี้ หมี่ ชี้	R
	์ s *	จี๊ด กรี๊ด	R
	*	มี ดี สี ปี	N
ขอบเขตหน้า	* c	ดี ตี ฝี	N
	* c c	วลี หนี ศรี	N
	* c c c	สตรี	X,N

ตารางแสดงอักษรควบของรูปสระ : -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	อ ห
ร	ก ท ต ห ส ป ค ศ ฟ
ก	ส
ม	ห
ย	ห
ว	ก ท ห ส ค
ล	ก ว ห ป ค พ ผ ถ
ด	บ ค
บ	ก
พ	ร
ณ	ม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-4

รูปแบบสระ -ุ	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	s g ิ ์ *	สุทธิ์	R
	s g ์ *	มนุษย์ ศุกร์	R
	s ิ *	วุฒิ พุฒิ	X
	s s *	บุตร ยุทธ อุณห สมุทร	X
	s *	ลุก สมุด สุก	N
	*	สุ อุ ตุ ทุ	N
	t s *	ยุ่ง ปุ๋ย มุ้ง	N
	t *	ตุ๊	N
ขอบเขตหน้า	* c ุ	ขุด จุด หุ้น	N
	* c c ุ	สมุด ปลุก หนุ่ม	N

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางแสดงอักษรควบของรูปสระ : -ุ

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	อ ม ห ส
ร	ก อ ว ส ป ค ผ
ก	อ ม ส ศ
อ	ช
ง	อ
ม	ห ส
ย	อ ห ส พ
ว	ห
ส	ก ห ส ล พ ผ ฟ
ท	ป
ค	ส ผ
ส	อ พ
ถ	ส
ธ	ม

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-5

รูปแบบสระ - ู	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	s g ์ *	ศูนย์ บูรณ์	R
	s s *	สูตร มูตร	X,N
	s *	สูบ ลูก รูป	N
	*	ดู ปู หู	N
	t s *	บู้ต	N
	t *	สู้ อยู่	N
ขอบเขตหน้า	* c ู	สู้ ดู ชู้	N
	* c c ู	หนู อยู่ สบู่	N
	* c c c ู	สกรู	X,N

ตารางแสดงอักษรควบของรูปสระ : - ู

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ม ห ธ
ร	ห ป ค
ม	ห
ย	อ ห
ว	ห
ล	อ ห ป ฟ
ค	ส
ห	พ
บ	ส
ณ	อ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-6

รูปแบบสระ -ั-	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	s g g g ์ *	ลักษมณ์	R
	s g ิ g ์ *	กษัตริย์	R
	s g g ์ *	ยันตร์	R
	s g ิ ์ *	ศักดิ์	R
	s g ์ *	พันธุ์	R
	s g ์ *	ศัพท์ ยันต์	R
	s ิ *	บัติ วัติ ฉัติ	R
	s s *	บัตร มัณฑ สมัคร	X,N
	s *	กัด มัน หัก	N
	t s *	ตั้ง ยั่ง ขั้น	R
	ว *	บัว วัว หัว	R
	ิ ว *	รั้ว ยั่ว มั่ว	R
	ย *	วัย นัย ศัย	R
	t ย *	ไม่พบคำที่สอดคล้อง	R
	ว ั *	ผัวะ พลัวะ	R
ขอบเขตหน้า	* c	คัด จับ รัก	N
	* c c	หยั่ง สงัด หวัด	N
	* c c c	มฆวัน	X

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางแสดงอักษรควบของรูปสระ : -ู

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	อ ห ส ป พ ผ ฉ
ร	ก อ ท ต ห ส ป บ จ ค พ ฝ
ก	ส
ง	ห ส ผ
ม	ห ส
ย	ห ข
ว	ร ต ห ส ค บ ช
ล	ก อ ต ห ส ป ค พ ผ
ท	ห
ค	ส
ห	ร อ ม
ส	อ ว
บ	ฉ
ภ	อ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-7

รูปแบบสระ -	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	ึ s g ์ *	ลึงค์ ดึงส์ ตรึงส์	R
	s *	สึก ลึก ดัง	N
	*	รึ หึ ฮึ	X,N
	ึ t s *	ขึ้น ผึ้ง ทึ่ง	R
	ึ t *	ไม่พบคำที่สอดคล้อง	N
ขอบเขตหน้า	* c	ดึก จึง ถึง	N
	* c c	หนึ่ง สลึง หมึก	N

ตารางแสดงอักษรควบของรูปสระ : -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	อ ห ค ผ
ร	ต ป ค พ
ม	ห
ล	ก อ ส ค ผ ถ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-8

รูปแบบสระ -	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	s *	คืน ยืม ปืน	R
	t *	ไม่พบคำที่สอดคล้อง	R
	t s *	ฟื้น ตื่น คลื่น	R
	อ *	คือ มือ หรือ	R
	t อ *	รื่อ มื้อ สื่อ	R
ขอบเขตหน้า	* c	มืด คืน	N
	* c c	หมื่น กลืน หรือ	N

ตารางแสดงอักษรควบของรูปสระ : -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ห
ร	ห ป ค
ม	ห
ว	ห
ล	ก ห ป ค

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-9

รูปแบบสระ -็ , -็อ-	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	์ อ s *	ช็อค น็อค ล็อค	R
	์ *	ก็ (เป็นคำเดียวที่สอดคล้อง)	R
ขอบเขตหน้า	* c	มืด คืน	N
	* c c	หมื่น กลืน หรือ	N

ตารางแสดงอักษรควบของรูปสระ : -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
ม	ส
ล	พ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-10

รูปแบบสระ - ะ	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	ะ *	กะ อะ พระ ชนะ	R
ขอบเขตหน้า	* c ะ	มะ ระ	N
	* c c ะ	สระ กระ ขยะ	N
	* c t ะ	ค่ะ ล่ะ จ๊ะ	N
	* c c t ะ	ไม่พบคำที่สอดคล้อง	N
	* c c c ะ	สรณะ อมตะ รตนะ	X,N

ตารางแสดงอักษรควบของรูปสระ : - ะ

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ช
ร	ก ท ศ ห ส ป พ ข
ว	ห ส
ล	ห ส ค ผ
ณ	ค ข ษ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-11

รูปแบบสระ -ำ	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	*	รำ ตำ กำ สำ	R
ขอบเขตหน้า	* c ◌ำ	คำ จำ นำ	N
	* c c ◌ำ	หยำ ตรำ	N
	* c t ◌ำ	ค่ำ น้ำ ย้ำ	N
	* c c t ◌ำ	ปล้ำ สม่ำ หน่ำ	N

ตารางแสดงอักษรควบของรูปสระ : -ำ

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ห
ร	ก ต ป ค พ
ม	ส ข
ย	ห
ว	ห ค
ล	ก ห ป ค พ ถ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-12

รูปแบบสระ - ฤ	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	ฤ s g ิ ์ *	ฤทธิ์	X,R
	ฤ s g ์ *	ฤกษ์	R
	ฤ g ์ *	คฤห์	R
	ฤ s *	กฤษ พฤษ มฤค ศฤง	N
	ฤ *	พฤ คฤ นฤ	N
	ฤ . .	ฤดู ฤดี ฤทัย ฤทธี ฤทธิ	X,R
ขอบเขตหน้า	* c ฤ	พฤ สฤงค์ หฤ	R
	* c c ฤ	อมฤต	X,R
	* ฤ	ฤษี ฤาษี ฤทธิ์	X,R
		(พบเพียง 9 คำที่สอดคล้อง)	

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-13

รูปแบบสระ เ -	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	เ c s g ์ *	เกณฑ์ เทศก์ เศษฐ์	R
	เ c t g ์ *	เล่ห์ เท่ห์	R
	เ c g ์ *	เรย์ เนย์	R
	เ c s *	เหตุ เกตุ เมรุ	X
	เ c s s *	เมตร เนตร	N
	เ c s *	เอก เมฆ เวร	N
	เ c *	เท เห เก	N
	เ c t *	เร่ เข้	N
	เ c t s *	เร่ง เข้ม เก่ง	N
	เ c c *	เปล่	N
	เ c c s *	เปลว เกรด อเนก	N
	เ c c s s *	เศรษฐ เกษตร	X,N
	เ c c t *	เตร่	N
	เ c c t s*	เปล่ง เขม้น	N
	เ - (t) า . .	ดู pattern C-13.a	
	เ - (t) อ . .	ดู pattern C-13.b	
	เ - (t) ว . .	ดู pattern C-13.c	
	เ - (t) ย . .	ดู pattern C-13.c	
	เ - ็ . .	ดู pattern C-13.d	
	เ - (t) ะ . .	ดู pattern C-13.e	
	เ - ิ . .	ดู pattern C-13.f	
	เ - ี . .	ดู pattern C-13.g	
	เ - ึ . .	ดู pattern C-13.h	

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขอบเขตหน้า	* เ	พบเป็นส่วนมาก	N
	* c เ	พเนจร อเวจี อเมริกา (พบเพียงไม่กี่คำ)	X,N

ตารางแสดงอักษรควบของรูปสระ : เ -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	อ ห
ร	ก ท ห ค ศ
ง	ห
ย	ห
ว	ห ค ข ศ
ล	ห พ ผ
ย	ก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-13.a

รูปแบบสระ เ–า , เ-าะ	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	เ - า g ์ *	เยาว์ เชาว์ เมาห์	R
	เ - า c *	เคราะห์ เพาะว์	X,N
	เ - า *	เกาะ เคาะ เปราะ	R
	เ - าะ g ์ *	เกณฑ์ เทศก์ เศษณ์	R
	เ - าะ *	เรย์ เนย์	R
ขอบเขตหน้า		ดู pattern C-13	

ตารางแสดงอักษรควบของรูปสระ : เ - า

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
ร	ป ค พ ศ
ง	ห
ม	ห ข
ย	ห ข
ว	ห
ล	ก ห ป ค พ ข

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางแสดงอักษรควบของรูปสระ : เ - า ะ

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
ร	ก ป ค พ
ม	ห
ย	ห
ล	ห ป พ ผ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-13.b

รูปแบบสระ เ–อ ,เ-อะ	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	เ - อ g ์ s *	เทอร์ม เตอร์ก	X,N
	เ - อ g ์ *	เบอร์ เฟอร์ เทอร์	R
	เ - อ s *	เทิด เทอม เทอญ เยอร	X,N
	เ - อ *	เออ เจอ เห่อ เสมอ เกลอ	N
	เ - อ ะ *	เลอะ เกรอะ เปรอะ เทอะ	R
ขอบเขตหน้า	* เ	ดู pattern C-13	

ตารางแสดงอักษรควบของรูปสระ : เ - อ

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ส
ร	ก ห ป
ม	ห ส
ย	ห
ล	ก ห ผ

ตารางแสดงอักษรควบของรูปสระ : เ - อะ

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
ร	ก ป ค

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-13.c

รูปแบบสระ เ–ว ,เ-ย	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	เ - ว ย *	เทอร์ม เตอร์ก	X,N
	เ - ว *	เบอร์ เฟอร์ เทอร์	R
	เ - ย *	เทิด เทอม เทอญ เยอร	X,N
ขอบเขตหน้า	* เ	ดู pattern C-13	

ตารางแสดงอักษรควบของรูปสระ : เ - ว

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
ล	ห ป

ตารางแสดงอักษรควบของรูปสระ : เ - ย

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ข
ว	ห
ล	ป ช ฉ
บ	ส

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-13.d

รูปแบบสระ เ– ย ,เ- ยะ	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	เ - ี ย g ์ *	เกียร์ เชียร์ เคลียร์	R
	เ - ี ย s *	เสี่ยง เรียก เหนียว	N
	เ - ี ย *	เสีย เลีย เมีย	N
	เ - ี t ย *	เสี่ย เมี่ย เมีย	N
	เ - ี t ย s *	เสี่ยง เยี้ยว เหี่ยว	N
	เ - ี ย ะ *	เปรียะ	R
	เ - ี t ย ะ *	เจี๊ยะ เพลี๊ยะ	R
ขอบเขตหน้า	* เ	ดู pattern C-13	

ตารางแสดงอักษรควบของรูปสระ : เ– ย

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ห ส พ
ร	ก ต ห ค พ
ม	ห ส
ย	ห
ว	ก ห ฉ
ล	ก ห ป พ ฉ
บ	ส
ษ	ก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางแสดงอักษรควบของรูปสระ : เ– ย ะ

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
ร	ป
ล	พ ผ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-13.e

รูปแบบสระ เ - ะ	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	เ - ะ *	เกะ เอะ เตะ เต็ะ	R
		เประ เขละ เพละ	N
ขอบเขตหน้า	* เ	ดู pattern C-13	

ตารางแสดงอักษรควบของรูปสระ : เ– ะ

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
ร	ป
ล	บ ผ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-13.f

รูปแบบสระ เ- -	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	เ- s *	เกิน เชิญ เปิด เลิก	R
		เพลิด เสริฐ เผอิญ	R
	เ- t s *	เพิ่ง เนิ่น เพิ่ม	R
	เ- . .	เช็ด เฮิร์ตซ์	X
ขอบเขตหน้า	* เ	ดู pattern C-13	

ตารางแสดงอักษรควบของรูปสระ : เ– -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
ร	ก ส จ พ
อ	ผ
ย	ห ผ
ล	ต ห ค พ ฉ
ด	ผ
ช	ผ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-13.g

รูปแบบสระ เ – ื อ	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	เ - อ s *	เลือก เหมือง เปลือง	N
	เ - อ *	เสือ เกลือ เหลือ	N
	เ - t อ *	เสื้อ เมื่อ เชื่อ	N
	เ - t อ s *	เชื่อง เครื่อง เหนื่อย	N
	เ - (t) อ ะ *	ไม่มีใช้	R
ขอบเขตหน้า	* เ	ดู pattern C-13	

ตารางแสดงอักษรควบของรูปสระ : เ– อ

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ห
ร	ป ค พ
ง	ห
ม	ห ส
ย	ห
ล	ก อ ม ห ป ค ถ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-13.h

รูปแบบสระ เ–็ -	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	เ - s g ์ *	เซ็นต์ เต็นท์	R
	เ - s *	เล็ก เย็น เจ็บ เอ็ด	N
		เสร็จ เมล็ด เตล็ด เสด็จ	N
ขอบเขตหน้า	* เ	ดู pattern C-13	

ตารางแสดงอักษรควบของรูปสระ : เ–็ -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ห
ร	ก ค ท ต พ
ม	ห ส ข
ว	ค จ
ล	ก ป ต ค
ค	ส ผ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-14

รูปแบบสระ แ -	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	แ c c s g ์ *	แสดงปี	R
	แ c s g ์ *	แพทย์	R
	แ c g ์ *	แคร์ แฟร์	R
	แ c s *	แสง แขน แดด	N
	แ c *	แก แห แพ	N
	แ c t *	แก่ แม่ แพ้	N
	แ c t s *	แกง แร้ง แล้ว	N
	แ c c s *	แมลง แสดง แขวง	N
	แ c c *	แคว แพร แตร	N
	แ c c t s *	แคล้ว แกร่ง แสร้ง	N
	แ c c t *	แพร่ แคร่	N
	แ c ะ *	แพะ และ แคะ	R
	แ c c ะ *	แกละ แคระ	R
	แ c t ะ *	แน่ะ	R
	แ c c t ะ *	ไม่พบคำที่สอดคล้อง	R
	แ c s *	แข็ง แล็บ แค็ป	R
	แ c c s *	แผล็บ แขย็ก แขม็บ	R
ขอบเขตหน้า	* แ	พบเป็นส่วนมาก	
	* c แ	แทยง สแน็ค สแล็คส์ (พบเพียงไม่กี่คำ)	

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางแสดงอักษรควบของรูปสระ : แ -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ห พ ผ
ร	ก ท ส ป ค พ
ง	ห
ม	ห ข
ย	ท ห ข
ว	ก ส ค ข
ล	ก ม ห ป ค พ ช ถ ฆ
ค	ส

ตารางแสดงอักษรควบของรูปสระ : แ - ็ -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
ร	ก
ม	ข
ย	ฆ
ล	พ ผ

ตารางแสดงอักษรควบของรูปสระ : แ - ะ

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
ร	ค
ม	ห
ย	ห ส
ว	ห ข
ล	ห พ ผ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-15

รูปแบบสระ โ -	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	โ c s g ์ *	โยชน์ โภชน์ โรจน์	R
	โ c t g ์ *	โล่	R
	โ c g ิ *	โพธิ์	R
	โ c g ์ *	โชว์ โปร์ โบว์	R
	โ c s ิ *	โชติ	X
	โ c s s *	โคตร โอษฐ	X
	โ c s *	โดด โขก โฆษ	N
	โ c *	โต โห โท	N
	โ c t *	โช่ โก้ โง่	N
	โ c t s *	โค้ง โม่ง โก่ง	N
	โ c c *	โปร โหล	N
	โ c c s *	โปรด โกรธ	N
	โ c c s s *	ไม่พบคำที่สอดคล้อง	R
	โ c c t *	โหล่ โหว่ โหล่	N
	โ c c t s *	โปร่ง โหม่ง	N
	โ c ะ *	โละ โปะ	R
	โ c t ะ *	โต๊ะ โป๊ะ	R
ขอบเขตหน้า	* โ	พบเป็นส่วนมาก	R
	* c โ	นโยบาย สโมสร อโศก มโหฬาร (พบเพียงไม่กี่คำ)	X,R

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางแสดงอักษรควบของรูปสระ : โ -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ห ฉ
ร	ท ห ป ค พ ศ
ม	ห ข
ย	ห พ ข
ว	ห
ล	ก ห ค พ ผ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-16

รูปแบบสระ ไ -	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	ไ c c g ์ *	ไฟลท์	R
	ไ c g ์ *	ไมล์ ไลท์ สไลด์ สไตล์	R
	ไ c *	ไป ไท ไย ไฟ	N
	ไ c t *	ไร่ ไข้ ไก่ ไผ่	R
	ไ c c *	ไหน ไหม ไกว	N
	ไ c c t *	ไตร่ ไหม้ ไขว้ ไหว้	N
	ไ c ย *	ไทย ไตย	N
ขอบเขตหน้า	* ไ	พบเป็นส่วนมาก	R
	* c ไ	สไบ ชไมพร สไตล์ มไหศวรรย์ (พบเพียงไม่กี่คำ)	X,R

ตารางแสดงอักษรควบของรูปสระ : ไ -

พยัญชนะตัวที่ 2 (c2)	พยัญชนะตัวแรก (c1)
น	ห ฉ
ร	ก ต ห ป พ ซ
ม	ห
ว	ก ห ข
ล	ก ห ค พ ถ
ท	ผ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PATTERN C-17

รูปแบบสระ ไ -	ขอบเขตของพยางค์	ตัวอย่าง	รูปแบบคุณลักษณะ
ขอบเขตหลัง	ไ c *	ใจ ใบ ใส ใด ใน ใย	R
	ไ c t *	ให้ ใส้ ใช้ ใด้ ใบ้	R
	ไ c c *	ใหล ใคร	N
	ไ c c t *	ใหญ่ ใหม่ ใคร่ ใกล้	N
		(พบคำที่ขึ้นต้นด้วย ไ เพียง 20 ตัว)	
ขอบเขตหน้า	* ไ	พบในทุกคำ	R

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก ค

ตัวอย่างรายละเอียดของคำที่ตัดผิดจากโปรแกรมการตัดคำโดยใช้คลังข้อความ

บางครั้ง/จะ/ให้/คำ/วินิจฉัย/โรค/ตาม/ตำแหน่ง/ที่/เป็น/เช่น/
วินิจฉัย/โรค วินิจฉัยโรค

โรคมะเร็ง/ที่/ปอด/ก็/ให้/คำ/วินิจฉัย/ว่า/เป็น/โรคมะเร็งปอด/
โรคมะเร็ง โรค/มะเร็ง
โรคมะเร็งปอด โรค/มะเร็ง/ปอด

โรคมะเร็งที่ต่อมน้ำเหลือง/
โรคมะเร็งที่ต่อมน้ำเหลือง โรค/มะเร็ง/ที่/ต่อ/ม/น้ำ/เหลือ/ง

แทนที่จะ/ให้/คำ/วินิจฉัย/ว่า/เป็น/โรคมะเร็งต่อมน้ำเหลือง/แต่/แพทย์/ส่วนใหญ่/จะ/ยัง/คุ้นเคย/กับ/คำวินิจฉัย/
ว่า/เป็น/
แทนที่จะ แทน/ที่/จะ
โรคมะเร็งต่อมน้ำเหลือง โรค/มะเร็ง/ต่อ/ม/น้ำ/เหลือ/ง
ส่วนใหญ่ ส่วน/ใหญ่
คำวินิจฉัย คำ/วินิจฉัย

ซึ่ง/ไม่/เป็นไป/ตาม/ระบบ/ที่/กำหนด/ไว้/ผล/ดังกล่าว/ทำให้/ข้อมูล/ส่วน/ที่/ต่อท้าย/นั้น/ไม่/ถูกต้อง/ตามที่/
ต้องการ/ปัญหา/ดังกล่าว/สามารถ/แก้ไข/โดย/ก่อน/การ/นำ/ระบบ/ไป/ใช้/นั้น/จะ/ต้อง/มี/การ/จัดทำ/คู่มือ/
อธิบาย/เพื่อ/ไว้/เป็น/เอกสารอ้างอิง/และ/มี/การ/จัด/อบรม/ให้/ผู้เกี่ยวข้อง/ทุก/ระดับ/ทำ/ความ/เข้าใจ/ใน/ระบบ/
ที่/ได้/พัฒนา/ขึ้น/การ/พัฒนา/ระบบ/ฐานข้อมูล/ของ/หน่วยงาน/ต่างๆ/
เป็นไป เป็น/ไป
ดังกล่าว ดัง/กล่าว
ทำให้ ทำ/ให้
ต่อท้าย ต่อ/ท้าย
ถูกต้อง ถูก/ต้อง
ตามที่ ตาม/ที่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ต้องการ ต้อง/การ
ดังกล่าว ดัง/กล่าว
สามารถ สา/มา/รถ
แก้ไข แก้/ไข
จัดทำ จัด/ทำ
คู่มือ คู่/มือ
เอกสารอ้างอิง เอก/สาร/อ้าง/อิง
อบรม อบ/รม
ผู้เกี่ยวข้อง ผู้/เกี่ยวข้อง
ระดับ ระ/ดับ
ฐานข้อมูล ฐาน/ข้อมูล
หน่วยงาน หน่วย/งาน

แม้/จะ/มี/ความ/ยุ่งยาก/สลับซับซ้อน/และ/มี/ความ/หลากหลาย/แตกต่าง/กัน/แต่/ก็/ไม่ใช่/เป็น/สิ่ง/ที่/เหนือ/
วิสัย/ความ/สามารถ/ของ/มนุษย์/ปัญหา/สำคัญ/อยู่/ที่/ผู้ใช้/จะ/ยอม/ใช้/ระบบ/ที่/พัฒนา/มา/อย่างดี/แล้ว/หรือไม่/
ต่างหาก/ซึ่ง/โดย/ธรรมชาติ/ของ/มนุษย์/
สลับซับซ้อน สลับ/ซับซ้อน

ไม่ใช่ ไม่/ใช่
สามารถ สา/มา/รถ
ผู้ใช้ ผู้/ใช้
อย่างดี อย่าง/ดี
หรือไม่ หรือ/ไม่
ต่างหาก ต่าง/หา/ก
ธรรมชาติ ธรรม/ชา/ติ

จะ/มี/ความ/ไม่/เป็น/ระบบ/ใน/ตัวเอง/
ตัวเอง ตัว/เอง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ทำ/ตามใจ/ตัวเอง/และ/ไม่/ชอบ/ถูก/บังคับ/และ/สิ่ง/นี้/คือ/จุดสำคัญ/ของ/ผู้ที่/จะ/พัฒนา/ระบบ/ฐานข้อมูล/ต้อง/ทำ/ความ/เข้าใจ/กับ/ผู้ใช้/อย่างถ่องแท้/เพื่อให้/ระบบ/ฐานข้อมูล/ที่/พัฒนา/แล้ว/มี/ผู้ใช้/ตอบสนอง/อย่างคุ้มค่า/สูงสุด/กิตติกรรมประกาศ/คณะ/ผู้วิจัยพัฒนา/ใคร่/ขอขอบพระคุณ/

ตามใจ ตาม/ใจ

ตัวเอง ตัว/เอ/ง

จุดสำคัญ จุด/สำคัญ

ผู้ที่ ผู้/ที่

ฐานข้อมูล ฐาน/ข้อมูล

ผู้ใช้ ผู้/ใช้

อย่างถ่องแท้ อย่าง/ถ่อ/งแท้

เพื่อให้ เพื่อ/ให้

ฐานข้อมูล ฐาน/ข้อมูล

ผู้ใช้ ผู้/ใช้

อย่างคุ้มค่า อย่าง/คุ้ม/ค่า

สูงสุด สูง/สุด

กิตติกรรมประกาศ กิตติ/กร/รม/ป/ระ/กาศ

ผู้วิจัยพัฒนา ผู้/วิจัย/พัฒนา

ขอขอบพระคุณ ขอ/ขอบ/พระ/คุณ

วิวัฒนฑีปะ/

สุทธิพล/วิวัฒนฑีปะ สุทธิพลวิวัฒนฑี/ปะ

ที่/ได้/ให้/ข้อเสนอแนะ/ตลอดจน/ข้อมูล/ต่างๆ/อัน/มีประโยชน์/อย่างมาก/ต่อ/การ/วิจัย/พัฒนา/ใน/ครั้ง/นี้/นอกจากนี้/คณะ/ผู้วิจัยพัฒนา/ใคร่/ขอขอบพระคุณ/สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ/และ/ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ/

ข้อเสนอแนะ ข้อ/เสนอ/แนะ

ตลอดจน ตลอด/จน

มีประโยชน์ มี/ประโยชน์

อย่างมาก อย่าง/มาก

นอกจากนี้ นอก/จาก/นี้

ผู้วิจัยพัฒนา ผู้/วิจัย/พัฒนา

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขอขอบพระคุณ ขอ/ขอบ/พระ/คุณ
สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ สำนัก/งาน/พัฒนา/วิทยาศาสตร์/และ/เทคโนโลยี/แห่ง/ชา/ติ
ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ศูนย์/เทคโนโลยี/อิเล็กทรอนิกส์/และ/คอมพิวเตอร์/แห่ง/ชา/ติ

ที่/ได้/เห็น/ความ/สำคัญ/ของ/โครงการ/นี้/และ/ให้/การ/สนับสนุน/ใน/การ/วิจัย/พัฒนา/ครั้ง/นี้/เอกสารอ้างอิง/กรมการปกครอง/
เอกสารอ้างอิง เอก/สาร/อ้าง/อิง
กรมการปกครอง กรม/การ/ปก/ครอง

เอกสาร/สิ่งพิมพ์/รายชื่อ/จังหวัด/
เอกสาร เอก/สาร
สิ่งพิมพ์ สิ่ง/พิมพ์
รายชื่อ ราย/ชื่อ

ไทย/พร้อมกับ/รหัส/ที่/เกี่ยวข้อง/
พร้อมกับ พร้อม/กับ

เอกสาร/พิมพ์/จาก/ฐานข้อมูล/ใน/เครื่องคอมพิวเตอร์/
เอกสาร เอก/สาร
ฐานข้อมูล ฐาน/ข้อมูล
เครื่องคอมพิวเตอร์ เครื่อง/คอม/พิวเตอร์

กรุงเทพมหานคร/
กรุงเทพมหานคร กรุงเทพ/มหา/น/คร

กระทรวงมหาดไทย/
กระทรวงมหาดไทย กระทรวง/มหา/ด/ไทย

การสื่อสารแห่งประเทศไทย/

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวอย่างรายละเอียดของคำที่ตัดผิดจากโปรแกรมการตัดคำโดยใช้พจนานุกรม

บางครั้ง/จะ/ให้/คำ/วินิจฉัย/โรค/ตาม/ตำแหน่ง/ที่/เป็น/เช่น/
วินิจฉัย/โรค วินิจฉัยโรค

โรคมะเร็ง/ที่/ปอด/ก็/ให้/คำ/วินิจฉัย/ว่า/เป็น/โรคมะเร็งปอด/
โรคมะเร็ง โรค/มะเร็ง
โรคมะเร็งปอด โรค/มะเร็ง/ปอด

โรคมะเร็งที่ต่อมน้ำเหลือง/
โรคมะเร็งที่ต่อมน้ำเหลือง โรค/มะเร็ง/ที่/ต่อ/ม/น้ำ/เหลือ/ง

แทนที่จะ/ให้/คำ/วินิจฉัย/ว่า/เป็น/โรคมะเร็งต่อมน้ำเหลือง/แต่/แพทย์/ส่วนใหญ่/จะ/ยัง/ลุ้นเลย/กับ/คำวินิจฉัย/ว่า/เป็น/
แทนที่จะ แทน/ที่/จะ
โรคมะเร็งต่อมน้ำเหลือง โรค/มะเร็ง/ต่อ/ม/น้ำ/เหลือ/ง
ส่วนใหญ่ ส่วน/ใหญ่
คำวินิจฉัย คำ/วินิจฉัย

ซึ่ง/ไม่/เป็นไป/ตาม/ระบบ/ที่/กำหนด/ไว้/ผล/ดังกล่าว/ทำให้/ข้อมูล/ส่วน/ที่/ต่อท้าย/นั้น/ไม่/ถูกต้อง/ตามที่/ต้องการ/ปัญหา/ดังกล่าว/สามารถ/แก้ไข/โดย/ก่อน/การ/นำ/ระบบ/ไป/ใช้/นั้น/จะ/ต้อง/มี/การ/จัดทำ/คู่มือ/อธิบาย/เพื่อ/ไว้/เป็น/เอกสารอ้างอิง/และ/มี/การ/จัด/อบรม/ให้/ผู้เกี่ยวข้อง/ทุก/ระดับ/ทำ/ความ/เข้าใจ/ใน/ระบบ/ที่/ได้/พัฒนา/ขึ้น/การ/พัฒนา/ระบบ/ฐานข้อมูล/ของ/หน่วยงาน/ต่างๆ/
เป็นไป เป็น/ไป
ดังกล่าว ดัง/กล่าว
ทำให้ ทำ/ให้
ต่อท้าย ต่อ/ท้าย
ถูกต้อง ถูก/ต้อง
ตามที่ ตาม/ที่
ต้องการ ต้อง/การ
ดังกล่าว ดัง/กล่าว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สามารถ สา/มา/รถ
แก้ไข แก้/ไข
จัดทำ จัด/ทำ
คู่มือ คู่/มือ
เอกสารอ้างอิง เอก/สาร/อ้าง/อิง
อบรม อบ/รม
ผู้เกี่ยวข้อง ผู้/เกี่ยวข้อง
ระดับ ระ/ดับ
ฐานข้อมูล ฐาน/ข้อมูล
หน่วยงาน หน่วย/งาน

แม้/จะ/มี/ความ/ยุ่งยาก/สลับซับซ้อน/และ/มี/ความ/หลากหลาย/แตกต่าง/กัน/แต่/ก็/ไม่ใช่/เป็น/สิ่ง/ที่/เหนือ/วิสัย/ความ/สามารถ/ของ/มนุษย์/ปัญหา/สำคัญ/อยู่/ที่/ผู้ใช้/จะ/ยอม/ใช้/ระบบ/ที่/พัฒนา/มา/อย่างดี/แล้ว/หรือไม่/ต่างหาก/ซึ่ง/โดย/ธรรมชาติ/ของ/มนุษย์/
สลับซับซ้อน สลับ/ซับซ้อน

ไม่ใช่ ไม่/ใช่
สามารถ สา/มา/รถ
ผู้ใช้ ผู้/ใช้
อย่างดี อย่าง/ดี
หรือไม่ หรือ/ไม่
ต่างหาก ต่าง/หา/ก
ธรรมชาติ ธรรม/ชา/ติ

จะ/มี/ความ/ไม่/เป็น/ระบบ/ใน/ตัวเอง/
ตัวเอง ตัว/เอง

ทำ/ตามใจ/ตัวเอง/และ/ไม่/ชอบ/ถูก/บังคับ/และ/สิ่ง/นี้/คือ/จุดสำคัญ/ของ/ผู้ที่/จะ/พัฒนา/ระบบ/ฐานข้อมูล/ต้อง/ทำ/ความ/เข้าใจ/กับ/ผู้ใช้/อย่างถ่องแท้/เพื่อให้/ระบบ/ฐานข้อมูล/ที่/พัฒนา/แล้ว/มี/ผู้ใช้/ตอบสนอง/อย่างคุ้มค่า/สูงสุด/กิตติกรรมประกาศ/คณะ/ผู้วิจัยพัฒนา/ใคร่/ขอขอบพระคุณ/
ตามใจ ตาม/ใจ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวเอง ตัว/เอ/ง

จุดสำคัญ จุด/สำคัญ

ผู้ที่ ผู้/ที่

ฐานข้อมูล ฐาน/ข้อมูล

ผู้ใช้ ผู้/ใช้

อย่างถ่องแท้ อย่าง/ท่อ/งแท้

เพื่อให้ เพื่อ/ให้

ฐานข้อมูล ฐาน/ข้อมูล

ผู้ใช้ ผู้/ใช้

อย่างคุ้มค่า อย่าง/คุ้ม/ค่า

สูงสุด สูง/สุด

กิตติกรรมประกาศ กิตติ/กรรม/ป/ระ/กาศ

ผู้วิจัยพัฒนา ผู้/วิจัย/พัฒนา

ขอขอบพระคุณ ขอ/ขอบ/พระ/คุณ

วิวัฒนฑีปะ/

สุทธิพล/วิวัฒนฑีปะ สุทธิพลวิวัฒนฑี/ปะ

ที่/ได้/ให้/ข้อเสนอแนะ/ตลอดจน/ข้อมูล/ต่างๆ/อัน/มีประโยชน์/อย่างมาก/ต่อ/การ/วิจัย/พัฒนา/ใน/ครั้ง/นี้/นอกจากนี้/คณะ/ผู้วิจัยพัฒนา/ใคร่/ขอขอบพระคุณ/สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ/และ/ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ/

ข้อเสนอแนะ ข้อ/เสนอ/แนะ

ตลอดจน ตลอด/จน

มีประโยชน์ มี/ประโยชน์

อย่างมาก อย่าง/มาก

นอกจากนี้ นอก/จาก/นี้

ผู้วิจัยพัฒนา ผู้/วิจัย/พัฒนา

ขอขอบพระคุณ ขอ/ขอบ/พระ/คุณ

สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ สำนัก/งาน/พัฒนา/วิทยาศาสตร์/และ/เทคโนโลยี/แห่ง/ชา/ติ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ศูนย์/เทคโนโลยี/อิเล็กทรอนิกส์/และ/คอมพิวเตอร์/แห่ง/ชา/ติ

ตัวอย่างรายละเอียดของคำที่ตัดผิดจากโปรแกรมการตัดคำโดยใช้กฎ

บางครั้ง/จะ/ให้/คำ/วินิจฉัย/โรค/ตาม/ตำแหน่ง/ที่/เป็น/เช่น/

บางครั้ง บางค/รั้ง

วินิจฉัย วิ/นิจ/ฉัย

ตำแหน่ง ตำ/แหน่ง

โรคมะเร็ง/ที่/ปอด/ก็/ให้/คำ/วินิจฉัย/ว่า/เป็น/โรคมะเร็งปอด/

โรคมะเร็ง โรค/มะ/เร็ง

วินิจฉัย วิ/นิจ/ฉัย

โรคมะเร็งปอด โรค/มะ/เร็ง/ปอด

โรคมะเร็งที่ต่อมน้ำเหลือง/

โรคมะเร็งที่ต่อมน้ำเหลือง โรค/มะ/เร็ง/ที่/ต่อม/น้ำ/เหลือง

ซึ่ง/มี/เซลล์/มะเร็ง/ชื่อ/

มะเร็ง มะ/เร็ง

แทนที่จะ/ให้/คำ/วินิจฉัย/ว่า/เป็น/โรคมะเร็งต่อมน้ำเหลือง/แต่/แพทย์/ส่วนใหญ่/จะ/ยัง/คุ้นเคย/กับ/คำวินิจฉัย/ว่า/เป็น/

แทนที่จะ แทน/ที่/จะ

วินิจฉัย วิ/นิจ/ฉัย

โรคมะเร็งต่อมน้ำเหลือง โรค/มะ/เร็ง/ต่อม/น้ำ/เหลือง

ส่วนใหญ่ ส่วน/ใหญ่

คุ้นเคย คุ้น/เคย

คำวินิจฉัย คำ/วิ/นิจ/ฉัย

ซึ่ง/ไม่/เป็นไป/ตาม/ระบบ/ที่/กำหนด/ไว้/ผล/ดังกล่าว/ทำให้/ข้อมูล/ส่วน/ที่/ต่อท้าย/นั้น/ไม่/ถูกต้อง/ตามที่/ต้องการ/ปัญหา/ดังกล่าว/สามารถ/แก้ไข/โดย/ก่อน/การ/นำ/ระบบ/ไป/ใช้/นั้น/จะ/ต้อง/มี/การ/จัดทำ/คู่มือ/

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

อธิบาย/เพื่อ/ไว้/เป็น/เอกสารอ้างอิง/และ/มี/การ/จัด/อบรม/ให้/ผู้เกี่ยวข้อง/ทุก/ระดับ/ทำ/ความ/เข้าใจ/ใน/ระบบ/ที่/ได้/พัฒนา/ขึ้น/การ/พัฒนา/ระบบ/ฐานข้อมูล/ของ/หน่วยงาน/ต่างๆ/

เป็นไป เป็น/ไป

ระบบ/ที่ ระ/บบ/ที่

กำหนด กำ/หนด

ผล/ดังกล่าว ผลดังก/ล่าว

ทำให้/ข้อมูล ทำ/ให้/ข้อ/มูล

ต่อท้าย ต่อ/ท้าย

ถูกต้อง ถูก/ต้อง

ตามที่ ตาม/ที่

ต้องการ ต้อง/การ

ปัญหา ปัญ/หา

ดังกล่าว ดังก/ล่าว

สามารถ สา/มารถ

แก้ไข แก้/ไข

ระบบ ระ/บบ

จัดทำ จัด/ทำ

คู่มือ คู่/มือ

อธิบาย อธิ/บาย

เอกสารอ้างอิง เอก/สาร/อ้าง/อิง

จัด/อบรม จัดอ/บรม

ผู้เกี่ยวข้อง ผู้/เกี่ยว/ข้อง

ระดับ ระ/ดับ

เข้าใจ เข้า/ใจ

ระบบ/ที่ ระ/บบ/ที่

พัฒนา พัฒ/นา

พัฒนา พัฒ/นา

ระบบ ระ/บบ

ฐานข้อมูล/ของ/หน่วยงาน ฐาน/ข้อ/มูล/ของ/หน่วย/งาน

ต่างๆ ต่าง/ๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

แม้/จะ/มี/ความ/ยุ่งยาก/สลับซับซ้อน/และ/มี/ความ/หลากหลาย/แตกต่าง/กัน/แต่/ก็/ไม่ใช่/เป็น/สิ่ง/ที่/เหนือ/
วิสัย/ความ/สามารถ/ของ/มนุษย์/ปัญหา/สำคัญ/อยู่/ที่/ผู้ใช้/จะ/ยอม/ใช้/ระบบ/ที่/พัฒนา/มา/อย่างดี/แล้ว/หรือ ไม่/
ต่างหาก/ซึ่ง/โดย/ธรรมชาติ/ของ/มนุษย์/
มี/ความ มีค/วาม
ยุ่งยาก/สลับซับซ้อน ยุ่ง/ยากส/ลับ/ซับ/ซ้อน
มี/ความ/หลากหลาย มีค/วามห/ลากห/ลาย
แตกต่าง แตก/ต่าง
ไม่ใช่ ไม่/ใช่
วิสัย วิ/สัย
สามารถ สา/มารถ
ของ/มนุษย์ ของม/นุษย์
ปัญหา ปัญ/หา
สำคัญ/อยู่ สำ/คัญอ/ยู่
ผู้ใช้ ผู้/ใช้
ระบบ/ที่ ระ/บบ/ที่
พัฒนา พัฒ/นา
มา/อย่างดี มาอ/ย่าง/ดี
หรือไม่ หรือ/ไม่
ต่างหาก ต่าง/หาก
โดย/ธรรมชาติ/ของ โดยธ/รรม/ชาติ/ของ

จะ/มี/ความ/ไม่/เป็น/ระบบ/ใน/ตัวเอง/
มี/ความ มีค/วาม
ระบบ ระ/บบ
ตัวเอง ตัว/เอง

ทำ/ตามใจ/ตัวเอง/และ/ไม่/ชอบ/ถูก/บังคับ/และ/สิ่ง/นี้/คือ/จุดสำคัญ/ของ/ผู้ที่/จะ/พัฒนา/ระบบ/ฐานข้อมูล/ต้อง/
ทำ/ความ/เข้าใจ/กับ/ผู้ใช้/อย่างถ่องแท้/เพื่อให้/ระบบ/ฐานข้อมูล/ที่/พัฒนา/แล้ว/มี/ผู้ใช้/ตอบสนอง/อย่างคุ้มค่า/
สูงสุด/กิตติกรรมประกาศ/คณะ/ผู้วิจัยพัฒนา/ใคร่/ขอขอบพระคุณ/
ตามใจ ตาม/ใจ
ตัวเอง ตัว/เอง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บรรณานุกรม

ภาษาไทย

[1] สมปรารถนา รัทยานนท์ "โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็คทรอนิคส์ภาษาไทย" วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์จุฬาลงกรณ์มหาวิทยาลัย พ.ศ. 2535

[2] ยืน ภู่วรวรรณ และ นายวิวรรธน์ อิ่มอารมณ์ "การตรวจสอบตัวสะกดด้วยคอมพิวเตอร์" บทความทางวิชาการ การประชุมทางวิชาการ วิศวกรรมไฟฟ้า 2530

[3] พิสิทธิ์ พรมจันทร์ "การวิเคราะห์แนวทางการเปรียบเทียบสมรรถนะของโปรแกรมแยกคำภาษาไทย" วิทยานิพนธ์ ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย พ.ศ. 2540

[4] ไพศาล เจริญพรสวัสดิ์ "การตัดคำภาษาไทยโดยใช้คุณลักษณะ" วิทยานิพนธ์ ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย พ.ศ. 2541

[5] ควงแก้ว สวามิภักดิ์ "การสร้างซอฟต์แวร์วิเคราะห์ไวยากรณ์ไทยภายใต้ระบบยูนิกซ์" กรุงเทพ : สำนักพิมพ์มหาวิทยาลัยธรรมศาสตร์ พ.ศ.2533

[6] บุญเสริม กิจศิริกุล "การกำกับหมวดคำสำหรับข้อความภาษาไทย" กรุงเทพ: สถาบันวิจัยและพัฒนาคณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย พ.ศ.2540

[7] ชมพูนุช ศุปติวุฒิ "การตัดคำกำกวมในข้อความภาษาไทยด้วยการโปรแกรมตรรกะเชิงอุปนัย" ปริญญามหาบัณฑิต สาขาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย พ.ศ.2542.

[8] ยืน ภู่วรวรรณ และ วิวรรธน์ อิ่มอารมณ์. "การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี." รายงานการประชุมวิชาการวิศวกรรมไฟฟ้า ครั้งที่ 9 พ.ศ.2529.

[9] วิรัช ศรเลิศล้ำวาณิช "การตัดคำภาษาไทยในระบบแปลภาษา" การแปลภาษาด้วยคอมพิวเตอร์. หน้า 50-55. กรุงเทพ : ศูนย์เทคโนโลยีอิเลคทรอนิกส์และคอมพิวเตอร์แห่งชาติ พ.ศ.2536

[10] สัมพันธ์ ระรื่นรมย์ "การแบ่งคำไทยด้วยพจนานุกรม" โครงงานวิศวกรรม ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย พ.ศ. 2534.

[11] สมศักดิ์ จันวัน "ระบบวิเคราะห์โครงสร้างภาษาไทยด้วยคอมพิวเตอร์" วิทยานิพนธ์มหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง พ.ศ. 2534.

[12] นัฐวุฒิ ไชยเจริญ. "การตัดคำและกำกับหมวดคำภาษาไทยแบบเบ็ดเสร็จด้วยคอมพิวเตอร์" วิทยานิพนธ์ปริญญาอักษรศาสตร์มหาบัณฑิต ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย พ.ศ. 2544.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

[13] รัตติกร วรากูลศิริพันธุ์,สง่า คงสุพานิช "การวิเคราะห์เลือกประโยคที่ถูกต้องจากความถี่ของการใช้คำ" Papers on Natural Language Processing , Compiled by Virach Sornlertlamvanich 1995

ภาษาอังกฤษ

[14] Charnyapornpong, S. " A Thai Syllable Separation Algorithm" Master Thesis.Asian Institute of Technology. 1983.

[15] VAN Rijsebergen , C.J. "Information Retrieval".1979 London: Butterworths.cited in Manning ,Chistoper D. And Hinrich Schutze. Foundations of statistical natural language processing .Cambrige:MIT Press. 1999.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้