การเพิ่มความถูกต้องของการแบ่งคำภาษาไทย โดยใช้แบบจำลองความน่าจะเป็นของหน้าที่อักขระ

นางสาวเกศราภรณ์ ซื่อสัตย์พาณิชย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2552

INCREASING ACCURACY OF THAI WORD SEGMENTATION USING CHARACTER FUNCTION PROBABILISTIC MODELS

Miss Kessaraporn Suesatpanit

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2009

หัวข้อวิทยานิพนธ์	การเพิ่มความถูกต้องของการแบ่งคำภาษาไทยโดยใช้แบบจำลองความน่าจะเป็นของหน้าที่อักขระ
โดย	นางสาว เกศราภรณ์ ซื่อสัตย์พาณิชย์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.อติวงศ์ สุชาโต
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.วันชัย ริ้วไพบูลย์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.อติวงศ์ สุชาโต)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล)

..... กรรมการภายนอกมหาวิทยาลัย
(ดร.ชัย วุฒิวิวัฒน์ชัย)

เกศราภรณ์ ซื่อสัตย์พาณิชย์ : การเพิ่มความถูกต้องของการแบ่งคำภาษาไทยโดยใช้แบบจำลองความน่าจะเป็นของหน้าที่อักขระ. (INCREASING ACCURACY OF THAI WORD SEGMENTATION USING CHARACTER FUNCTION PROBABILISTIC MODELS) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ผู้ช่วยศาสตราจารย์ ดร. อติวงศ์ สุชาโต, อ. ที่ปรึกษาวิทยานิพนธ์ร่วม ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ 41 หน้า.

งานวิจัยนี้นำเสนอการแบ่งคำภาษาไทยโดยใช้คอนดิชันนัลแรนดอมฟิลด์สด้วยการใช้อักขระและกลุ่มอักขระเป็นคุณลักษณะ สำหรับกลุ่มอักขระนั้นถูกจัดกลุ่มตามหน้าที่การใช้งานของอักขระ เช่น ตำแหน่งการวางอักขระในการเขียน เป็นต้น เทมเพลตคุณลักษณะในคอนดิชันนัลแรนดอมฟิลด์สนั้นใช้อักขระและกลุ่มหน้าที่อักขระมาพิจารณาหรือใช้ใน N-gram เพื่อระบุขอบเขตของคำ ได้แบ่งการทดลองเป็น 2 เทมเพลตคุณลักษณะ คือ 1.ใช้อักขระเป็นคุณลักษณะเพียงอย่างเดียว 2.ใช้ทั้งอักขระและกลุ่มหน้าที่อักขระเป็นคุณลักษณะ และทำการเปรียบเทียบความถูกต้องกับการแบ่งคำด้วยแบบจำลองมาร์คอฟไทรแกรมระดับคำ โดยผลการทดลองจากงานวิจัยนี้ได้ค่า F-Measure ดีที่สุดคือ 95.53% ซึ่งให้ผลดีกว่าการใช้แบบจำลองมาร์คอฟไทรแกรมที่ได้ค่า F-Measure 90.98%

จากการวิเคราะห์ผลการแบ่งคำทำให้เห็นว่าการใช้หน้าที่อักขระเข้าเป็นคุณลักษณะเพิ่มจากการใช้อักขระเพียงอย่างเดียวนั้นทำให้ผลการแบ่งคำดีขึ้น ถึงแม้ว่าจะทำให้จำนวนคุณลักษณะในเทมเพลตเพิ่มขึ้นแต่ก็ยังทำให้ประสิทธิภาพการแบ่งคำดีอยู่ และการใช้อักขระช่วยให้ผลการแบ่งคำมีความเสถียรคงทนในการแบ่งคำที่ไม่เคยเห็นมาก่อนในคลังข้อความฝึกฝนมากกว่าเมื่อเทียบกับการใช้แบบจำลองมาร์คอฟไทรแกรมระดับคำ

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต.....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....

ปีการศึกษา 2552 ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

5071405621 : MAJOR COMPUTER SCIENCE

KEYWORDS : TEXT ANALYSIS / WORD SEGMENTATION / WORD FILTERING / TEXT PROCESSING / NATURAL LANGUAGE PROCESSING

KESSARAPORN SUESATPANIT : INCREASING ACCURACY OF THAI WORD SEGMENTATION USING CHARACTER FUNCTION PROBABILISTIC MODELS. THESIS ADVISOR : ASSISTANT PROFESSOR ATIWONG SUCHATO, Ph.D., THESIS CO-ADVISOR : ASSISTANT PROFESSOR PROADPRAN PUNYABUKKANA, Ph.D., 41 pp.

A thai word segmentation approach using Conditional Random Fields (CRFs) are utilized for classifying each character associated with the text string to be segmented into classes of characters categorized based on their positions in the underlying words. Characters used in the Thai writing system are attached with character functions proposed in this work. N-grams of these character functions are considered together with character N-grams within the feature templates of the CRF models in order for the models to locate characters likely to indicate word boundaries. The proposed methods yields the best F-measure score of 95.53% which is better than ones obtained based on word trigrams with score of 90.98%.

We can observe from the result that word segmentation using CRFs yielding better performances than the one using word trigrams on every genre. Comparing the two types of feature templates, the one that contain character function features perform slightly better than the templates relying only on the character sequences. Although this observation is far from surprising considering the greater number of features in the feature templates relying on the character and character function sequences, the fact that the inclusion of character functions helps the word segmentation performance is still encouraging. It is also shown that character-level constraints make the result more robust to segmenting unseen words.

Department : Computer Engineering	Student's Signature เกศราภรณ์ สืบสัตย์พานิช
Field of Study : Computer Science	Advisor's Signature อติวงศ์
Academic Year : 2009	Co-Advisor's Signature

กิตติกรรมประกาศ

จากการที่ได้ศึกษาค้นคว้าและจัดทำงานวิจัยที่ผ่านมาได้รับรู้ว่าการทำวิทยานิพนธ์ไม่ใช่เพียงแต่จะต้องมีความรู้ในด้านงานวิจัยหรืองานที่เกี่ยวข้องเท่านั้น แต่สิ่งที่สำคัญต่อผู้เขียนวิทยานิพนธ์มากที่สุดคือกำลังใจและแรงบันดาลใจ ขอขอบคุณท่านอาจารย์ที่ปรึกษา อาจารย์อติวงศ์ สุชาโต และอาจารย์โปรดปราน บุณยพุกกณะ ที่ได้ให้สิ่งเหล่านี้ ทั้งคำปรึกษา คำแนะนำ และกำลังใจตลอดมา ตั้งแต่เริ่มต้นจนงานวิจัยนี้สำเร็จได้อย่างที่คาดหวัง ขอขอบคุณ อาจารย์วันชัย ริ้วไพบูลย์ อาจารย์วิโรจน์ อรุณมานะกุล และอาจารย์ชัย วุฒิวิวัฒน์ชัย ที่สละเวลาอันมีค่ามาเป็นประธานและคณะกรรมการการสอบวิทยานิพนธ์ ทำให้ได้แนวคิดแง่มุมต่างๆ ที่อาจยังมองไม่เห็นหรือตกหล่นในงานวิจัย นำมาใช้ปรับแก้งานได้สมบูรณ์มากยิ่งขึ้น ขอบคุณเพื่อนในกลุ่มปฏิบัติการวิจัยระบบภาษาพูด เพื่อนๆ รวมถึงญาติๆ ที่คอยพูดคุยให้กำลังใจยามท้อแท้ ขอขอบคุณหัวหน้าและพี่ๆน้องๆ ที่ร่วมงานในบริษัท เมโทรซิสเต็มส์คอร์ปอเรชั่น จำกัด (มหาชน) ที่ให้ความเข้าใจและสนับสนุนการทำงานวิจัย โดยเฉพาะการให้เวลาและทรัพยากรเครื่องมือเพื่อใช้ในการทำงานวิจัย ซึ่งเปรียบเสมือนกำลังใจที่มีคุณค่ามาก

ที่สำคัญที่สุดคือบิดาและมารดาที่เข้าใจธรรมชาติของผู้เขียนวิทยานิพนธ์ คอยสนับสนุนและผลักดันในแบบที่ไม่ทำให้ผู้เขียนวิทยานิพนธ์รู้สึกกดดันแต่อย่างใด ซึ่งเป็นวิธีการที่ดีที่สุดสำหรับผู้เขียนวิทยานิพนธ์ และสุดท้ายที่สำคัญยิ่งกว่าอะไร ต้องขอบคุณตัวผู้เขียนวิทยานิพนธ์เองที่สามารถกำจัดและทำลายสิ่งบั่นทอนกำลังใจตนเองและสามารถสร้างกำลังใจได้ด้วยตัวเองในยามที่จำเป็นต้องมีอย่างมาก จนสามารถทำงานวิทยานิพนธ์ได้สำเร็จดังที่คาดหวัง

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

วิทยาการคอมพิวเตอร์ได้มีกระบวนการประมวลผลภาษาธรรมชาติของมนุษย์ (Natural Language Processing) ไม่ว่าจะเป็นการปรับรูปเขียนให้เป็นรูปอ่านในระบบแปลงรูปอักษรเป็นเสียงพูด (Text-to-Speech) ระบบการแปลภาษาไทยเป็นภาษาต่างประเทศ การสกัดคำสำคัญเพื่อใช้ในระบบการค้นคืนข้อมูลสารสนเทศ (Information Retrieval) การตรวจไวยากรณ์ในโปรแกรมเอกสาร ซึ่งในการประมวลผลภาษาธรรมชาตินั้นมีกระบวนการที่สำคัญขั้นตอนหนึ่ง คือ การแบ่งคำหรือการหาขอบเขตคำเนื่องจากภาษาไทยเป็นภาษาที่ไม่มีขอบเขตของคำที่ชัดเจน และยังมีความกำกวมอยู่ในหลายๆ คำ ที่เป็นปัญหาเดียวกับภาษาต่างประเทศอื่นในแถบตะวันออก เช่น ภาษาจีน ภาษาญี่ปุ่น เป็นต้น ไม่เหมือนกับภาษาต่างประเทศอื่นในแถบตะวันตกที่มีเครื่องหมายบอกขอบเขตของคำได้ชัดเจนคือเว้นวรรคเป็นตัวแบ่งได้ว่าส่วนใดเป็นคำ เช่น ภาษาอังกฤษ

เนื่องจากภาษาไทยไม่ได้มีหลักเกณฑ์ที่ตายตัวว่าเมื่อเจอสัญลักษณ์หรือตัวอักษรใดแล้วจึงแบ่งให้เป็นคำ ทำให้การแบ่งคำเพื่อให้เกิดความถูกต้องแม่นยำจำเป็นต้องพิจารณาจากบริบทรอบข้างของคำนั้นๆ ด้วย เพราะสายอักขระหนึ่งๆ ที่อยู่ภายในบริบทที่ต่างกัน อาจมีการแบ่งคำที่ต่างกันได้ เช่น ตากลม สามารถแบ่งได้เป็น “ตา” และ ”กลม” หรือ “ตาก” และ “ลม” ได้เช่นกัน ขึ้นอยู่กับบริบทและความหมายของประโยคที่ผู้เขียนต้องการสื่อสาร นอกเหนือจากนี้แล้วยังมีปัญหารูปคำภาษาไทยที่ไม่ได้มีเฉพาะคำไทยแท้เพียงอย่างเดียว แต่ยังมีคำทับศัพท์คือคำที่มาจากภาษาต่างประเทศที่ถูกสะกดอยู่ในรูปของคำอ่านภาษาไทยทำให้เกิดคำใหม่ๆ ได้เสมอ อีกทั้งแต่ละคนอาจเขียนคำทับศัพท์จากคำภาษาอังกฤษเดียวกันแต่เขียนออกมาเป็นภาษาไทยที่แตกต่างกันอีกด้วย

ความถูกต้องแม่นยำของการแบ่งคำนั้นมีผลอย่างยิ่งต่องานที่มีการใช้แบ่งคำ ยกตัวอย่างเช่น การแปลภาษาด้วยเครื่องคอมพิวเตอร์ ไม่ว่าจะแปลจากภาษาไทยเป็นอังกฤษหรือจากภาษาอังกฤษเป็นไทยนั้นต้องมีการวิเคราะห์รูปประโยค วางรูปประโยค และรู้กลุ่มของคำ เช่น ประธาน กริยา กรรม เป็นต้น แต่ในการทำขั้นตอนเหล่านั้นได้ การแบ่งคำเป็นอันดับแรกที่สำคัญต้องทำให้ถูกต้อง เพราะหากแบ่งผิดพลาดแล้วจะทำให้การวิเคราะห์ต่างๆ ผิดพลาดและส่งผลทำให้เกิดการแปลภาษาออกมาผิดพลาดตามมาด้วย

จากปัญหาข้างต้น ทำให้การแบ่งคำในภาษาไทยต้องมีการใช้เทคนิควิธีหรือแนวคิดที่จะทำให้คอมพิวเตอร์สามารถรู้ได้จากการประมวลผลว่าส่วนใดเป็นขอบเขตของคำ โดย

งานวิจัยที่ถูกนำเสนอที่ผ่านมามีหลากหลายวิธีการเพื่อให้เกิดการแบ่งคำที่มีประสิทธิภาพทั้งทางด้านความถูกต้องของคำและความรวดเร็วในการประมวลผล เช่น การใช้พจนานุกรม การใช้ความรู้ด้านสถิติที่ประกอบกับคลังข้อความขนาดใหญ่ การเรียนรู้ด้วยเครื่อง และการผสมผสานหลายเทคนิควิธี เป็นต้น

งานวิจัยนี้นำเสนอวิธีการแบ่งคำภาษาไทยโดยใช้คอนดิชันนัลแรนดอมฟิลด์สโดยเลือกใช้คุณลักษณะภาษาไทยทั้งระดับอักขระภาษาไทยและกลุ่มอักขระภาษาไทยมาช่วยทำให้การแบ่งคำมีความถูกต้องมากยิ่งขึ้น

1.2 วัตถุประสงค์ของการวิจัย

เป้าหมายของวิทยานิพนธ์นี้ต้องการพัฒนาวิธีการแบ่งคำภาษาไทยให้มีความถูกต้องมากขึ้นภายใต้ขอบเขตการวิจัย เพื่อแก้ปัญหาการพบคำที่ไม่รู้จักหรือคำที่ไม่เคยปรากฏมาก่อนในพจนานุกรมและคลังข้อความฝึกฝน และทำการทดลองเพื่อประเมินประสิทธิภาพที่ได้จากวิธีที่นำเสนอ

1.3 ขอบเขตของการวิจัย

1. เสนอวิธีแบ่งคำภาษาไทยโดยใช้หลักเกณฑ์การแบ่งคำไทยด้วย “หน่วยเล็กที่สุดที่มีองค์ประกอบความเป็นคำครบถ้วน” [10] เท่านั้น
2. แบ่งคำจากข้อความรับเข้าโดยไม่สนใจการสะกดคำหรือโครงสร้างการเขียนว่าถูกต้องหรือไม่ เมื่อทำการประมวลผลการแบ่งคำแล้วจะได้ผลลัพธ์จากการแบ่งคำโดยแทรกสัญลักษณ์ | เพิ่มลงไปเท่านั้น ส่วนอื่นจากข้อความรับเข้าให้คงรูปเดิมไว้
3. ในการประเมินความถูกต้องของการแบ่งคำจะหาค่าจาก F-Measure จากการใช้คลังข้อความเบส (BEST) [9] ที่พัฒนาขึ้นศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติหรือเนคเทค (NECTEC) ซึ่งทำให้การประเมินจะไม่นับคำที่เป็นคำกลอนเข้ามาด้วย และมาทำการเปรียบเทียบกับการแบ่งคำโดยใช้แบบจำลองมาร์คอฟ (Markov Model)

1.4 ขั้นตอนและวิธีการดำเนินงานวิจัย

ขั้นตอนการดำเนินงานวิจัยสามารถแบ่งออกเป็นขั้นต่างๆ ได้ดังนี้

1. ขั้นตอนการเตรียมตัว
    ก. ศึกษาข้อมูลและทฤษฎีต่างๆ ที่เกี่ยวข้องกับงานวิจัย เช่น การใช้โปรแกรม Stochastic Gradient Descent (SGD) [6] ที่เป็นเครื่องมือสำหรับสร้างคอนดิชันนัลแรนดอมฟิลด์ส ทั้งรูปแบบข้อมูลนำเข้าและข้อมูลผลลัพธ์เพื่อใช้ในการพัฒนาแบ่งคำ ความรู้อักขระภาษาไทย และความรู้อื่นๆ
    ข. ศึกษางานวิจัยที่เกี่ยวข้อง
    ค. กำหนดหลักเกณฑ์การแบ่งคำภาษาไทย
    ง. กำหนดวิธีการวัดความถูกต้องหรือประสิทธิภาพในการแบ่งคำ
2. ขั้นตอนการกำหนดคุณลักษณะที่ใช้ในคอนดิชันนัลแรนดอมฟิลด์ส
    ก. กำหนดกลุ่มหน้าที่อักขระภาษาไทย
    ข. กำหนดเลเบล (Label) ที่ใช้กำกับอักขระเพื่อบ่งบอกขอบเขตของคำในคอนดิชันนัลแรนดอมฟิลด์ส
    ค. จัดเตรียมเทมเพลตคอนดิชันนัลแรนดอมฟิลด์สในระดับอักขระภาษาไทยและกลุ่มอักขระภาษาไทย
3. ขั้นตอนการพัฒนาระบบการแบ่งคำด้วยการใช้คอนดิชันนัลแรนดอมฟิลด์ส
    ก. สร้างฟังก์ชั่นแปลงข้อความสายอักขระปกติให้อยู่ในรูปแบบข้อมูลนำเข้าที่ต้องนำไปใช้ในโปรแกรม Stochastic Gradient Descent (SGD) [6] (เครื่องมือสำหรับสร้างคอนดิชันนัลแรนดอมฟิลด์ส)
    ข. สร้างฟังก์ชั่นแปลงข้อมูลผลลัพธ์ที่ได้จากโปรแกรม Stochastic Gradient Descent (SGD) [6] ให้อยู่ในรูปแบบข้อความสายอักขระปกติที่มีเครื่องหมาย | เป็นตัวคั่นการแบ่งคำ
4. ขั้นตอนการพัฒนาระบบการแบ่งคำภาษาไทยโดยใช้แบบจำลองมาร์คอฟ (Markov Model) เพื่อเป็นตัวเปรียบเทียบ (Baseline) กับวิธีที่นำเสนอ
5. ขั้นตอนทดสอบระบบการแบ่งคำด้วยการใช้คอนดิชันนัลแรนดอมฟิลด์สและตัวเปรียบเทียบ (Baseline)
6. ขั้นตอนวิเคราะห์และสรุปผล
    ก. วิเคราะห์และสรุปผลการทดลอง
    ข. เรียบเรียงวิทยานิพนธ์ พร้อมทั้งนำเสนองานวิจัยทั้งหมด

1.5 ประโยชน์ที่คาดว่าจะได้รับ

งานวิจัยนี้ทำให้เกิดวิธีการแบ่งคำไทยที่มีความถูกต้องมากยิ่งขึ้น และสามารถนำวิธีการแบ่งคำไทยในงานวิจัยนี้ไปประยุกต์ใช้กับงานประมวลผลภาษาไทยอื่นๆ

1.6 ผลงานตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของงานวิทยานิพนธ์ได้รับการตีพิมพ์เป็นบทความวิชาการในหัวเรื่อง “Thai Word Segmentation Using Character-Level Information” โดยเกศราภรณ์ ซื่อสัตย์พาณิชย์ โปรดปราน บุณยพุกกณะ และอติวงศ์ สุชาโต ในงานประชุมวิชาการ “The Eighth International Symposium on Natural Language Processing” ซึ่งจัดขึ้นที่กรุงเทพฯ ประเทศไทย วันที่ 20 – 21 ตุลาคม พ.ศ. 2552

บทที่ 2
ทฤษฎีแและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1. คอนดิชันนัลแรนดอมฟิลด์ส (Conditional Random Fields หรือ CRF) [4]

2.1.1.1 จำกัดความของคอนดิชันนัลแรนดอมฟิลด์ส

คอนดิชันนัลแรนดอมฟิลด์สสามารถอธิบายด้วยแนวคิดของแบบจำลองกราฟิก โดยให้ y เป็นโครงสร้างกราฟแบบห่วงโซ่เส้นตรง (Linear-chain) ซึ่งประกอบด้วยโหนด (Node) $y_1,...y_{|y|}$ ในกรณีที่เป็นแบบจำลองกราฟิกแบบไร้ทิศทาง (Undirected graphical model) การแจกแจงความน่าจะเป็น (Probability distribution) ทำได้โดยประกอบขึ้นจาก Clique ของกราฟ ซึ่งแต่ละ $c_i \in c$ คือชุดย่อย (Subset) ของโหนดที่มีการเชื่อมต่อกันโดยสมบูรณ์ในตัวเอง เรียกว่า yc_i ซึ่งสามารถใส่พารามิเตอร์ให้กับ Clique ได้โดยใช้ Clique potential ψc_i จึงสามารถแจกแจงความน่าจะเป็นของกราฟได้โดยการคูณกันของ Clique potential ทั้งหมด ดังนี้

$$p(y) = \frac{1}{Z}\prod_{c_i \in c} \psi c_i(yc_i) \tag{2.1}$$

โดย $Z = \sum_y \prod_{c_i \in c} \psi c_i(yc_i)$ คือเทอมมาตรฐาน (Normalize Term) ที่ทำให้แน่ใจได้ว่า $\sum_y p(y) = 1$ และ Clique potential เขียนในรูปแบบของฟังก์ชั่นคุณลักษณะ (Feature Function) ได้ดังนี้

$$\psi c_i(yc_i) = \prod_k \exp\{\lambda_k f_k(yc_i)\} = \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(yc_i)\right\} \tag{2.2}$$

โดยที่ K คือจำนวนคุณลักษณะ (Feature) และ $\lambda_1,...,\lambda_K$ คือพารามิเตอร์น้ำหนัก (Weight parameter) ให้กับฟังก์ชั่นคุณลักษณะ f_k สำหรับในการใช้คอนดิชันนัลแรนดอมฟิลด์สนี้ จะสร้างคอนดิชันนัลแรนดอมฟิลด์สแบบห่วงโซ่เส้นตรงด้วยการแจกแจงความน่าจะเป็นแบบมีเงื่อนไข (Condition probability distribution) $p_\lambda(y \mid x)$ ของการเรียงต่อกันของตัวเลเบล (Label) y เมื่อรู้การเรียงต่อกันของ x

อธิบายให้เข้าใจง่าย สมมติให้ทั้ง y และ x มีความยาวเท่ากันเป็น T โดย $y = (y_1,..., y_T))$ และ $x = (x_1,..., x_T)$ และให้ y ขึ้นอยู่กับสิ่งที่อยู่ก่อนหน้าหนึ่งตำแหน่ง แสดงฟังก์ชั่นคุณลักษณะด้วย $f_k(y_{t-1}, y_t, x, t)$ คือผลจากการเปลี่ยนสถานะนี้ขึ้นอยู่กับสถานะก่อนหน้า แต่ไม่ขึ้นกับสถานะ อื่นๆ ก่อนหน้านั้นอีก ดังนั้นการแจกแจงความน่าจะเป็นแบบมีเงื่อนไขสำหรับคอนดิชันนัลแรนดอมฟิลด์สแบบห่วงโซ่เส้นตรงกลายเป็น ดังนี้

$$p_\lambda(y \mid x) = \frac{1}{Z_\lambda(x)} \exp\left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t) \right\} \quad (2.3)$$

$$Z_\lambda(x) = \sum_{y} \exp\left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t) \right\} \quad (2.4)$$

2.1.1.2 การประมาณและอนุมานพารามิเตอร์

ให้ชุดข้อมูลฝึกฝน $D = \{x^{(i)}, y^{(i)}\}^N$ โดยที่ N คือจำนวนตัวอย่างที่ใช้ฝึกฝนทั้งหมด วัตถุประสงค์คือหาชุดของพารามิเตอร์น้ำหนัก $\lambda = \{\lambda_1, ..., \lambda_K\}$ ซึ่งหาโดยใช้วิธี Maximum likelihood estimation (MLE) สามารถอธิบายด้วย Log likelihood ดังนี้

$$\begin{aligned} l(\lambda; D) &= \log p(D \mid \lambda) \\ &= \log \prod_{i=1}^{N} p\lambda\left(y^{(i)} \mid x^{(i)}\right) \\ &= \sum_{i=1}^{N} \log p\lambda\left(y^{(i)} \mid x^{(i)}\right) \end{aligned} \quad (2.5)$$

อย่างไรก็ตาม การใช้ MLE อาจจะทำให้เกิดโอเวอร์ฟิต (Overfit) กับข้อมูลฝึกฝนได้ จึงใช้ Maximum a posteriori estimation โดยที่มีสมมติฐานว่าเทอม $\log p(\lambda)$ เป็น Gauss จึงกำหนด Gaussian prior ลงที่ Likelihood function

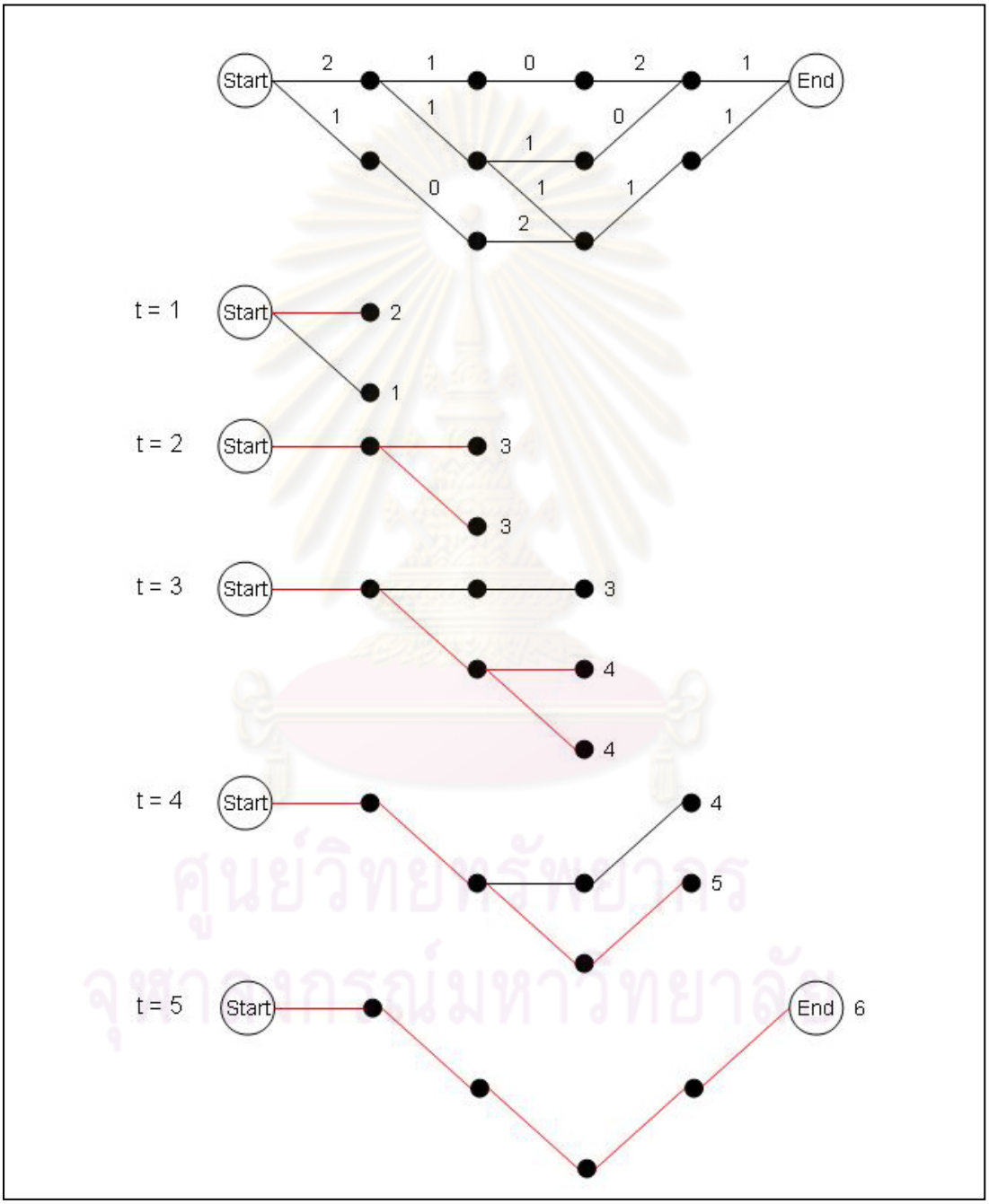
$$\begin{aligned} \log p(\lambda \mid D) &= \log p(D \mid \lambda) + \log p(\lambda) \\ &= \sum_{i=1}^{N} \log p\lambda\left(y^{(i)} \mid x^{(i)}\right) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2} \end{aligned} \quad (2.6)$$

2.1.2. การค้นหาเส้นทางแบบวิเทอร์บิ (Viterbi search) และการค้นหาเส้นทางที่ดีที่สุด N เส้นทาง (N-best Search)

ในขั้นตอนการแบ่งคำขั้นตอนหนึ่งมีการสร้างแบบการแบ่งคำที่เป็นไปได้มากมายเพื่อหาแบบการแบ่งคำที่ดีที่สุด และไม่ว่าจะด้วยวิธีการใด เช่น การใช้แบบจำลองมาร์คอฟในงานวิจัยของฆนาศัย [2] และตัวเปรียบเทียบ (Baseline) ที่เป็นการแบ่งคำด้วยแบบจำลองมาร์คอฟไทรแกรมที่ใช้เปรียบเทียบในการทดลองของงานวิจัยนี้ก็สามารถแบ่งแบบที่เป็นไปได้จำนวนมากก่อนที่จะเลือกแบบแบ่งคำหรือเส้นทางที่ดีที่สุดของวิธีนั้นๆ ในขณะที่มีแบบการแบ่งมากมายทำให้เกิดเส้นทาง (Path) สำหรับค้นหาได้มากมายเช่นกัน หากค้นหาหมดทุกเส้นทางเพื่อ

หาแบบที่มีความน่าจะเป็นสูงที่สุด จะทำให้การประมวลผลหนักและใช้เวลามาก จึงมีวิธีที่ใช้ค้นหาวิธีหนึ่งก็คือ การค้นหาเส้นทางแบบวิเทอร์บิ (Viterbi search)

ยกตัวอย่าง รูปภาพด้านล่างแสดงการค้นหาเส้นทางแบบวิเทอร์บิ คือ ณ ตำแหน่ง t ใดๆ ใน การค้นหาเส้นทางแบบวิเทอร์บิมีการคำนวณค่าความน่าจะเป็นรวมของทุกเส้นทาง และเลือกค้นหาเฉพาะเส้นทางที่มีความน่าจะเป็นรวมสูงที่สุด



รูปที่ 1 การค้นหาเส้นทางแบบวิเทอร์บิ

สำหรับการค้นหาเส้นทางแบบวิเทอร์บินั้นทิ้งเส้นทางที่มีค่าความน่าจะเป็นต่ำทิ้งไปตั้งแต่ต้น ซึ่งอาจทำให้เส้นทางที่ถูกต้องถูกตัดทิ้งไปด้วย จึงมีการพัฒนาการค้นหาเส้นทางแบบวิเทอร์บิแบบที่เรียกว่า การค้นหาเส้นทางที่ดีที่สุด N เส้นทาง ซึ่ง ณ t ใดๆ จะทำการค้นหาต่อไปเฉพาะเส้นทางที่มีความน่าจะเป็นสูงสุดจำนวน N เส้นทางแรก จึงทำให้ลดโอกาสน้อยลงที่เส้นทางที่ถูกต้องถูกตัดทิ้งไปตั้งแต่แรก

2.1.3. **อักษรไทย (อักขระไทย)** [15]

อักษรไทย มีพยัญชนะ 44 รูป สระ 21 รูป วรรณยุกต์ 4 รูป และเครื่องหมายอื่นๆ อีกจำนวนหนึ่ง พยัญชนะไทยจะเรียงตัวไปตามแนวนอนจากซ้ายไปขวา ส่วนสระจะอยู่ด้านหน้า ด้านบน ด้านล่าง และด้านหลังพยัญชนะ ซึ่งในการประกอบคำขึ้นอยู่กับประเภทของสระ

อักษรไทยไม่มีการแยกอักษรตัวใหญ่หรืออักษรตัวเล็กอย่างอักษรโรมัน และไม่มีการเว้นวรรคระหว่างคำ เมื่อจบหนึ่งประโยคจะลงท้ายด้วยการเว้นวรรค

ภาษาไทยมีตัวเลขเป็นของตัวเองเรียกว่า ตัวเลขไทย แต่นิยมใช้เลขอารบิกเป็นส่วนใหญ่ในชีวิตประจำวัน

2.1.1.3 พยัญชนะ

พยัญชนะไทยมี 44 รูป แบ่งออกเป็น 3 หมู่ เรียกว่า ไตรยางศ์ โดยยึดเอาพื้นเสียงของพยัญชนะที่ยังไม่ได้ผันวรรณยุกต์เป็นเกณฑ์ ประกอบด้วย

ก. อักษรสูง 11 ตัว ได้แก่ ข ฃ ฉ ฐ ถ ผ ฝ ศ ษ ส ห
ข. กลาง 9 ตัว ได้แก่ ก จ ฎ ฏ ด ต บ ป อ
ค. อักษรต่ำ 24 ตัว ได้แก่ ค ฅ ฆ ช ฌ ซ ฑ ฒ ท ธ พ ภ ฟ ฮ ง ญ ณ น ม ย ร ล ว ฬ

2.1.1.4 สระ

สระในภาษาไทยมี 21 รูป ดังนี้ ะ ั ็ า ิ ่ ำ " ุ ู เ ใ ไ โ อ ย ว ฤ ฤๅ ฦ ฦๅ

2.1.1.5 วรรณยุกต์

วรรณยุกต์ในภาษาไทยมี 4 รูป คือ ่ ้ ๊ ๋

2.1.1.6 ตัวเลข

ตัวเลขที่เป็นอักษรไทย เรียกว่าเลขไทย มีวิธีการบอกจำนวนโดยใช้ระบบประจำหลักเหมือนกับตระกูลเลขอารบิก มีดังนี้ ๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙

2.1.1.7 เครื่องหมายวรรคตอน

เครื่องหมายวรรคตอน มีดังนี้ . , ; : - — ( ) [ ] { } ? ! “ ” " / ๆ ฯ ฯลฯ ◌ ◌" ฯ ฯะ ◌ ์ .

ผู้วิจัยได้สังเกตการเขียนคำไทยนั้นประกอบจากการผสมอักษรหรืออักขระไทย ที่มีทั้งพยัญชนะ สระ วรรณยุกต์ ตัวเลข และเครื่องหมายวรรคตอน ถึงแม้จะมีการแบ่งประเภทอักษรไทยดังกล่าวชัดเจนแล้ว แต่การแบ่งคำภาษาไทยยังไม่มีกฎเกณฑ์แน่นอนหรือตายตัวว่าคำหนึ่งๆ ต้องประกอบด้วยประเภทอักษรไทยประเภทใดบ้างและไม่มีความแน่นอนในการเรียงลำดับประเภทอักษรไทยไว้แต่อย่างใด ถึงแม้จะไม่มีกฎเกณฑ์ชัดเจน แต่ในการเขียนหรือแบ่งคำก็ยังสามารถสังเกตเห็นข้อจำกัดในการผสมตัวอักษรแต่ละประเภทให้เป็นคำได้ เช่น สระ “ะ” “ั” และ “ิ” ต้องตามหลังตัวอักษรและไม่สามารถเป็นอักษรตัวแรกของคำได้เสมอ และ สระ “ั” ยังต้องมีพยัญชนะตามหลังหรืออาจมีวรรณยุกต์ก่อนแล้วพยัญชนะตามหลังเสมอ เป็นต้น ซึ่งสามารถใช้ประโยชน์จากกฎหรือข้อจำกัดของตัวอักษร และให้มีการจัดกลุ่มอักขระตามข้อจำกัดหรือหน้าที่การใช้งานของอักขระเพื่อใช้เป็นคุณลักษณะฝึกฝนในคอนดิชันนัลแรนดอมฟิลด์สได้

2.2 งานวิจัยที่เกี่ยวข้อง

เนื่องจากการแบ่งคำภาษาไทยได้มีการทำศึกษาวิจัยมาเป็นเวลานาน จึงทำให้เกิดงานวิจัยที่เสนอวิธีและเทคนิคต่างๆ มากมายมาแก้ปัญหาในการแบ่งคำภาษาไทย ผู้วิจัยจะกล่าวถึงงานวิจัยที่ผ่านมาโดยเรียงจากอดีตจนถึงปัจจุบัน

2.3.1. Thai Syllable Separator by Dictionary, Y. Poovarawan and W. Imarrom [12]

งานวิจัยของยืน ภู่วรวรรณเป็นการเสนอวิธีการแบ่งพยางค์ไทยด้วยพจนานุกรม โดยมีการจัดเตรียมพจนานุกรมไว้ การทำงานนั้นเริ่มจากรับข้อความเข้ามาเป็นสายอักขระ แล้วระบบทำการตรวจสอบอักขระจากซ้ายไปขวากับพยางค์ที่เก็บในพจนานุกรม ถ้าตรวจสอบแล้วพบว่ามีพยางค์มากกว่า 1 พยางค์ที่พบในพจนานุกรมโดยเลือกพยางค์ที่ยาวที่สุด และทำไปเรื่อยๆ

จนจบสายอักขระ แต่หากทำไปจนไม่สามารถพบพยางค์ใดเลยในพจนานุกรม ก็จะทำย้อนกลับเพื่อใช้พยางค์ที่ยาวรองลงมาแทน

สำหรับวิธีนี้มีการประมวลผลไม่ซับซ้อนมากนักทำให้ใช้เวลาได้เร็วเมื่อเทียบกับการใช้กฎแบ่งพยางค์ และในสมัยนั้นการใช้พื้นที่หน่วยความจำเก็บคำจากพจนานุกรมจำเป็นต้องมีขนาดใหญ่ แต่เครื่องคอมพิวเตอร์ก็ยังสามารถรองรับได้ ในความเป็นจริงยังพบปัญหาที่วิธีนี้ไม่สามารถแก้ไขการแบ่งพยางค์ (คำ) ได้ครอบคลุม ดังนี้

ก. ภาษาไทยมีคำกำกวมที่สามารถแบ่งได้หลายแบบ ซึ่งไม่สามารถใช้การเลือกแบบที่มีความยาวมากที่สุดได้ในทุกกรณี การเลือกว่าจะแบ่งแบบใดนั้นขึ้นอยู่กับบริบทของคำที่อยู่ด้วย เช่น ตากลม สามารถแบ่งได้ทั้ง ตา|กลม| และ ตาก|ลม|

ข. ภาษาไทยมีคำที่เกิดขึ้นใหม่ได้เสมอ โดยไม่ได้ถูกจัดเก็บในพจนานุกรม และยากต่อการปรับปรุงพจนานุกรมให้ทันปัจจุบันเพื่อให้มีคำทุกคำได้ทั้งหมด ยกตัวอย่าง คำที่เป็นชื่อเฉพาะ เช่น ธิติกานต์ ปิยพงศ์ จิตรวดี เป็นต้น และคำทับศัพท์ภาษาอังกฤษ เช่น จากคำว่า E-mail ถึงแม้จะมีหลักการเขียนที่กำหนดว่าต้องเขียนเป็นภาษาไทยอย่างไรให้ถูกต้อง แต่ผู้เขียนก็อาจจะเขียนต่างกันได้เป็นอีเมล อีเมล์ หรืออีเมลล์ เป็นต้น

จากการแบ่งคำด้วยพจนานุกรมพบปัญหาเมื่อเกิดคำกำกวมที่สามารถแบ่งได้หลายแบบ จึงมีงานวิจัยที่ใช้หลักสถิติมาช่วยแก้ไขปัญหานี้ได้โดยใช้ความน่าจะเป็นมาเป็นตัวพิจารณาเลือกแบบแบ่งคำ ดังแสดงรายละเอียดในหัวข้อถัดไป

2.3.2. A Statistical Approach to Thai Word Filtering, A. Kawtrakul, C. Thumkanon and S. Seriburi [1]

งานวิจัยของอัศนีย์ ก่อตระกูลและคณะมองถึงปัญหาการประมวลผลโครงสร้างภาษาไทย (Thai Morphological Processing) ที่ยังมีความกำกวมของขอบเขตคำและการกำกับหน้าที่คำ อัศนีย์ ก่อตระกูลและคณะจึงได้นำเสนอวิธีการแบ่งคำภาษาไทยเพื่อแก้ปัญหาดังกล่าวด้วยการคำนวณค่าความน่าจะเป็นของทุกแบบของการแบ่งคำโดยใช้แบบจำลองไทรแกรม (Trigram Model) ของแบบจำลองมาร์คอฟ (Markov Model) ซึ่งแบบการแบ่งคำที่เหมาะสมที่สุดนั้นคือแบบที่มีความน่าจะเป็นสูงที่สุด แต่อย่างไรก็ตามข้อมูลที่เข้ามาทำการแบ่งอาจมีการสะกดผิด และจากการสะกดผิดนั้นทำให้เกิดแบบการแบ่งคำที่ไม่เหมาะสมหรือไม่เกิดประโยชน์มีมาก

เกินไปจนทำให้ผลลัพธ์จากการแบ่งคำผิดไปได้ อีกทั้งทำให้การทำงานของพาสเซอร์ (Parser) ช้าลงกว่าที่ควรจะเป็น งานของอัศนีย์ ก่อตระกูลและคณะนี้จึงมีวิธีการจัดการกับปัญหาสะกดผิด โดยมีกระบวนการตรวจจับ (Scanning for detecting) และแก้ไขการสะกดผิดให้ถูกต้อง ทำให้ช่วยลดผลลัพธ์ที่อาจผิดพลาดและให้การทำงานของพาสเซอร์รวดเร็วขึ้น

เริ่มด้วยการนำกฎ (Word Formation Rules) และพจนานุกรมมาทำการแบ่งคำทุกแบบที่เป็นไปได้ ทำให้ได้คำที่กำกับด้วยหน้าที่คำทุกหน้าที่คำที่เป็นไปได้ เมื่อได้แบบแบ่งคำแล้ว มีการจัดการกับปัญหาการสะกดผิด (ถ้ามี) โดยการเสนอแนะกลุ่มคำที่เหมาะสมมากกว่ามาใช้แทน จากนั้นจึงทำในขั้นตอนของการหาความน่าจะเป็นด้วยแบบจำลองไทรแกรมของคำและแบบจำลองไทรแกรมของหน้าที่คำต่อไป

แบบจำลองไทรแกรมมีประโยชน์ในการคำนวณความน่าจะเป็นของการแบ่งคำด้วยการฝึกฝนจากคลังข้อความที่มีข้อความที่ถูกแบ่งคำแล้วพร้อมกำกับหน้าที่คำไว้ สำหรับแบบจำลองไทรแกรมของคำ คำนวณได้ดังสมการ

$$P(W) = \prod_{i=1}^{n} P(w_{i,n}) = \prod_{i=1}^{n} P(w_i \mid w_{i-1}, w_{i-2}) \quad (2.7)$$

จากสมการ (2.7) นั้น W คือประโยคที่แบ่งคำแล้ว ซึ่งประโยคนั้นประกอบด้วยคำต่างๆ คือ $w_1 w_2 ,..w_n$ โดยคำที่เกิด 2 คำมีผลกระทบกับค่าความน่าจะเป็นของคำถัดไป สำหรับการคำนวณค่า $P(w_i \mid w_{i-2,i-1})$ สามารถคำนวณได้ดังนี้

$$P(w_i \mid w_{i-2,i-1}) = \frac{C(w_{i-2,i})}{C(w_{i-2,i-1})} \quad (2.8)$$

จากสมการ (2.8) เป็นการหาค่าความน่าจะเป็นที่คำ w_i จะเกิดขึ้นโดยมี w_{i-2} และ w_{i-1} ที่อยู่ด้านหน้า หาได้จากการนับจำนวนครั้งที่เกิด w_{i-2}, w_{i-1}, w_i หารด้วยจำนวนครั้งที่เกิด w_{i-2}, w_{i-1}

จากการหาความน่าจะเป็นของการเกิดคำที่ติดกัน 3 คำ ทำให้มีปัญหาว่าข้อมูลจากคลังข้อความมีไม่เพียงพอ จึงใช้สมการนี้แทน

$$\prod_{i=1}^{n} P\left(w_i \mid w_{i-2,i-1}\right) = \prod_{i=1}^{n} \left(\lambda_1 P(w_n) + \lambda_2 P(w_n \mid w_{n-1}) + \lambda_3 P(w_n \mid w_{n-2,n-1})\right) \quad (2.9)$$

จากสมการ (2.9) นำการคำนวณความน่าจะเป็นแบบไบแกรมและยูนิแกรมเข้ามาร่วมด้วย และกำหนดค่า 0.1, 0.3, 0.6 ให้กับ $\lambda_1, \lambda_2, \lambda_3$ ตามลำดับ

เพื่อจัดการปัญหาความกำกวมในการกำกับหน้าที่คำ จึงมีการใช้แบบจำลองไทรแกรมของหน้าที่คำด้วย ดังสมการ

$$P(t_{1,n}) = \prod_{i=1}^{n} P(t_i \mid w_i) P(t_{i+1} \mid t_i) \qquad (2.10)$$

จากที่เสนอให้ลดความกำกวมของการแบ่งคำและกำกับหน้าที่ของคำด้วย จึงใช้ค่าเฉลี่ยของค่าความน่าจะเป็นทั้งสมการ (2.9) และ (2.10) เป็นค่าความน่าจะเป็นแต่ละแบบเพื่อใช้เลือกค่าความน่าจะเป็นที่สูงที่สุด

ผลการทดลองแสดงให้เห็นว่าช่วยลดแบบการแบ่งคำที่ไม่เหมาะสมออกไปได้จำนวนมากทำให้ประมวลผลได้รวดเร็วขึ้น แต่ทั้งนี้วิธีการจัดการปัญหาการสะกดผิดมีผลโดยตรงต่อความถูกต้องในการแบ่งคำ เพราะถ้าวิธีการในการจัดการกับการสะกดผิดทำงานได้ไม่ถูกต้องอาจทำให้ผลการแบ่งคำผิดตามไปด้วย นอกจากนี้การนำสถิติมาใช้นั้นมีการนำบริบทรอบข้าง (การใช้ไทรแกรม ไบแกรมและยูนิแกรมของคำ ซึ่งในการแบ่งคำแบบใดนั้นขึ้นอยู่กับคำที่อยู่ข้างหน้าด้วย) มาช่วยทำให้ลดความกำกวมได้ดีกว่าการจับคำจากพจนานุกรมเพียงอย่างเดียว รวมถึงขนาดและประสิทธิภาพของคลังข้อความก็ยังเป็นปัจจัยสำคัญต่อความแม่นยำในการแบ่งคำด้วย

ผู้วิจัยคิดว่าการใช้วิธีการทางสถิติยังให้ผลที่ดีมากกว่าการใช้พจนานุกรมเพียงอย่างเดียว เนื่องจากมีคำกำกวมและคำใหม่ที่เพิ่มขึ้นได้เสมอ ทำให้การใช้พจนานุกรมให้ประสิทธิภาพในการแบ่งคำไม่ดีเท่าการใช้สถิติในการฝึกฝนเรียนรู้คุณลักษณะและข้อจำกัดของคำ ซึ่งข้อสำคัญสำหรับการฝึกฝนเรียนรู้นั้นคือการกำหนดค่าพารามิเตอร์น้ำหนักให้แต่ละคุณลักษณะที่ใช้ฝึกฝน โดยการกำหนดค่าพารามิเตอร์น้ำหนักนั้นขึ้นอยู่กับคลังข้อความที่นำมาฝึกฝนด้วย ผู้วิจัยจึงเสนอให้ใช้แบบจำลองคอนดิชันนัลแรนดอมฟิลด์สที่มีการปรับค่าพารามิเตอร์ไปตามคลังข้อความฝึกฝน ซึ่งจะทำให้ประสิทธิภาพในการแบ่งคำดีขึ้นกว่าการใช้แบบจำลองมาร์คอฟ

2.3.3. Collocation and Thai Word Segmentation, W. Aroonmanakun [13]

วิธีที่งานวิจัยของวิโรจน์ อรุณมานะกุลนำเสนอเกิดจากแนวคิดว่าความกำกวมในการแบ่งคำสามารถแก้ปัญหาได้โดยให้มีการแบ่งพยางค์ เนื่องจากพยางค์เป็นหน่วยหนึ่งที่สามารถระบุได้อย่างชัดเจนมากกว่าคำ ซึ่งมีการแบ่งขั้นตอนหลักเป็น 2 ขั้นตอน ดังนี้

2.3.3.1 การแบ่งพยางค์

เมื่อรับข้อความรับเข้าที่เป็นสายอักขระนำมาแบ่งพยางค์โดยจับให้เป็นพยางค์กับแบบรูปของพยางค์ (Syllable Pattern) ที่ได้กำหนดไว้ประมาณ 200 แบบรูป ทำให้ได้แบบการแบ่งพยางค์ทุกแบบที่เป็นไปได้นั่นคืออาจมีมากกว่าหนึ่งแบบ จากนั้นนำทุกแบบมาฝึกฝนด้วยคลังข้อความที่แบ่งพยางค์ไว้แล้วเพื่อหาความน่าจะเป็นแบบไทรแกรมของพยางค์ และเลือกแบบที่ดีที่สุดคือที่มีความน่าจะเป็นสูงที่สุด

2.3.3.2 การรวมพยางค์เข้าเป็นคำ

จากแนวคิดว่าขอบเขตของพยางค์มีโอกาสที่จะเป็นขอบเขตของคำด้วย จึงทำการหาความเหนียวแน่นของการเกิดร่วมกันระหว่างพยางค์ (Collocation Strength) เพื่อทำการรวมพยางค์ให้เป็นคำ โดยค่าความเหนียวแน่นของการเกิดร่วมกันระหว่าง 2 พยางค์ที่เป็นส่วนหนึ่งในคำเดียวกันมีค่าสูงกว่าค่าความเหนียวแน่นของการเกิดร่วมกันระหว่าง 2 พยางค์ที่ไม่ได้เป็นส่วนหนึ่งในคำเดียวกัน เช่น จากประโยค เปิดหน้าต่าง ซึ่งแบ่งตามจริงได้สองคำคือ เปิดและหน้าต่าง โดยค่าความเหนียวแน่นของการเกิดร่วมกันระหว่างเปิดและหน้ามีค่าต่ำกว่าค่าความเหนียวแน่นของการเกิดร่วมกันระหว่างหน้าและต่าง เป็นต้น

สำหรับการรวมพยางค์สามารถทำได้โดยรวมพยางค์ทุกๆ แบบในประโยคนั้น (โดยไม่ได้ใช้พจนานุกรม) แต่ทำให้จำนวนแบบมีจำนวนมากได้เท่ากับ $2n-1$ โดยที่ n คือจำนวนพยางค์ทั้งหมด ดังนั้นในขั้นนี้ให้ทำการรวมพยางค์โดยกับจับกับคำในพจนานุกรม

เนื่องจากการคำนวณหาค่าความเหนียวแน่นของการเกิดร่วมกันระหว่างพยางค์มาใช้ตัดสินใจเพียงอย่างเดียวไม่เพียงพอ เพราะมีกรณีที่อาจทำให้บางพยางค์ถูกรวมไปเป็นคำอื่นด้วย เช่น หาก|คณะ|กรรม|การ|ปล่อย|ให้|ผู้|รับ|เหมา|แข็ง|ข้อ|ต่อ|รา|คา|ประ|มูล| ถ้าใช้ความเหนียวแน่นของการเกิดร่วมกันระหว่างพยางค์อย่างเดียวอาจทำให้ “ข้อ” และ “ต่อ” ถูกรวมเป็นคำ โดยที่ “ข้อ” และ “ต่อ” ไม่ได้เป็นคำเดียวกัน แต่ “แข็ง” และ ”ข้อ” เป็นคำเดียวกัน เพื่อป้องกันเหตุดังกล่าวจึงคำนวณโดยหาค่าความเหนียวแน่นระหว่างคำเพื่อนำไปลบออกด้วย

$$St = \sum_{i=1}^{n} F_{w_i} - \sum_{i=1}^{n-1} Dw_i, w_{i+1} \quad (2.11)$$

$$F_{w_i} = \sum_{j=1}^{k-1} C_{s_j, s_{j+1}} \text{ เช่น } w_i = s_1 s_2 ... s_k \quad (2.12)$$

$$D_{w_i, w_{i+1}} = C_{s_j, s_{j+1}} \quad (2.13)$$

St คือค่าความเหนียวแน่นของพยางค์ที่เกิดร่วมกันโดยรวมของประโยคที่ลบด้วยค่าความเหนียวแน่นระหว่างคำโดยรวมของประโยค

F_{w_i} คือค่าความเหนียวแน่นของพยางค์ที่เกิดร่วมกัน

$D_{w_i, w_{i+1}}$ คือความเหนียวแน่นระหว่างคำ

s คือพยางค์

w คือคำ

s_j คือพยางค์สุดท้ายของ w_i

s_{j+1} คือพยางค์แรกของ w_{i+1}

สำหรับวิธีการคำนวณโดยใช้ความเหนียวแน่นของการเกิดร่วมกันของพยางค์แบ่งออกเป็น 3 วิธี และเพิ่มวิธีเลือกแบบคำที่ยาวที่สุด (Longest Maching) ด้วย เพื่อนำไปทำการเปรียบเทียบผลของแต่ละวิธี มีดังนี้

ก. MaxColl-A หาความเหนียวแน่นของคำด้วยค่าความเหนียวแน่นของพยางค์ที่เกิดร่วมกันโดยรวมของประโยคที่ลบด้วยค่าความเหนียวแน่นระหว่างคำโดยรวมของประโยค ดังสมการ (2.11)

ข. MaxColl-B หาความเหนียวแน่นของคำด้วยค่าความเหนียวแน่นของพยางค์ที่เกิดร่วมกันโดยรวมของประโยคที่ลบด้วยค่าความเหนียวแน่นระหว่างคำโดยลบเฉพาะพยางค์ที่สามารถอยู่ท้ายคำอื่นได้ เช่น ..a-b-c-d-e,... มีการรวมพยางค์ b-c-d เป็นคำ ถ้า a-b สามารถเป็นท้ายคำอื่นได้ด้วย ทำการหาความเหนียวแน่นของการเกิดร่วมกัน

ระหว่างพยางค์ a-b ไปลบออกจากความเหนียวแน่นระหว่างพยางค์ b-c และ c-d

ค. MaxColl-C หาความเหนียวแน่นของคำด้วยค่าความเหนียวแน่นของพยางค์ที่เกิดร่วมกันโดยรวมของประโยค (ไม่ลบด้วยค่าใดๆ)

ง. MaxMatch เลือกแบบคำที่ยาวที่สุด

ในการคำนวณค่าความเหนียวแน่นการเกิดร่วมกันระหว่างพยางค์ หาได้โดยคำนวณอัตราส่วนของ p(x,y) และ q(x,y) โดยที่ p(x,y) คือ ความน่าจะเป็นที่พยางค์ x และ y อยู่ติดกัน และ q(x,y) คือ ความน่าจะเป็นที่พยางค์ x และ y ถูกพยางค์อื่นมาคั่น มีสมการคำนวณความน่าจะเป็น ดังนี้

$$\log \frac{p(x,y)}{q(x,y)} = \log \frac{p(x)p(y \mid x)}{q(x)q(y \mid x)} = \log \frac{p(y \mid x)}{q(y \mid x)} \quad (2.14)$$

$$= \log \frac{Count(x,y)/Count(x)}{Count(x,Any,y)/Count(x)} \quad (2.15)$$

$$= \log \frac{Count(x,y)}{Count(x,Any,y)} \quad (2.16)$$

จากผลการทดลองแบ่งเป็น 2 สถานการณ์คือ สถานการณ์แรกมีคำที่รู้จักอยู่ในพจนานุกรมทั้งหมด พบว่าประสิทธิภาพการแบ่งคำด้วยวิธี MaxMatch คือเลือกคำที่ยาวที่สุดที่ใช้พจนานุกรมยังให้ผลที่ดีที่สุด แต่ถ้าวิเคราะห์ผลลัพธ์จากวิธี MaxMatch ทำการแบ่งผิดพลาด เช่น อดีต|รัฐมนตรี|ที่มา|อยู่|พรรค|ไทยรักไทย|ใน|ปัจจุบัน| เป็นกรณีสายอักขระ “ที่มา” นั้นใช้วิธีเลือกคำที่ยาวที่สุดไม่ได้ จากผลลัพธ์ของกรณีแบบนี้ วิธี MaxColl-C จึงเป็นวิธีที่ให้ผลดีที่สุด

สถานการณ์ที่สองมีคำที่ไม่รู้จักในพจนานุกรมเป็น 29% ของจำนวนคำในพจนานุกรม พบว่าวิธี MaxColl-C ได้ผลที่ดีที่สุดคือหาความเหนียวแน่นของคำด้วยค่าความเหนียวแน่นของพยางค์ที่เกิดร่วมกันโดยรวมของประโยค (ไม่ลบด้วยค่าใดๆ) ดังสมการนี้

$$St = \sum_{i=1}^{n} F_{w_i} \quad (2.19)$$

$$F_{w_i} = \sum_{j=1}^{k-1} C_{s_j, s_{j+1}} \text{ เช่น } w_i = s_1 s_2 ... s_k \quad (2.20)$$

พบว่าพจนานุกรมมีส่วนที่มีผลต่อประสิทธิภาพในการแบ่งคำเพราะในขั้นตอนรวมพยางค์เป็นคำหากมีคำที่ไม่ได้ปรากฏในพจนานุกรมอาจทำให้เกิดความผิดพลาดในการแบ่งคำได้

ผู้วิจัยเห็นว่าการใช้วิธีทางสถิติก็ยังถือเป็นวิธีที่ให้ผลดีและยังเหมาะที่จะใช้ในงานวิจัยต่อไปได้ และในงานวิจัยของฆนาศัยที่จะกล่าวถัดไปเป็นการใช้วิธีทางสถิติด้วยแบบจำลองคอนดิชันนัลแรนดอมฟิลด์ส

2.3.4. A Conditional Random Field Framework for Thai Morphological Analysis, C. Kruengkrai, V.Sornlertlamvanich and H. Isahara [3]

งานวิจัยของฆนาศัยนำเสนอกรอบการทำงาน (Framework) สำหรับงานวิเคราะห์โครงสร้างภาษาไทย (Thai Morphological Analysis) ด้วยการนำคอนดิชันนัลแรนดอม-ฟิลด์สเข้ามาใช้ เพื่อหลีกเลี่ยงปัญหาจากการใช้ แบบจำลองฮิดเดนมาร์คอฟ (hidden Markov models (HMMs)) ที่เป็นแบบจำลองเจเนอเรทิฟ (Generative Model) ซึ่งมีแบบจำลองความน่าจะเป็นแบบการแจกแจงความน่าจะเป็นร่วมกัน (Joint Probability Distribution) คือ p(y,x) โดย x คือข้อมูลรับเข้า และ y คือข้อมูลผลลัพธ์ สำหรับการคำนวณ p(y,x) ต้องหา p(x) และในการหา p(x) ได้นั้น x แต่ละตัวต้องไม่ขึ้นต่อกัน แต่ในความเป็นจริงแล้ว x แต่ละตัวในสายอักขระมีการขึ้นต่อกัน เช่น การใช้คำขึ้นต้น (prefix) หรือคำที่อยู่รอบๆ มาใช้เป็นคุณลักษณะสำหรับการเรียนรู้เพื่อใช้ในการตัดสินใจหรือคาดเดาในการแบ่งคำ จึงอาจทำให้แบบจำลองที่สร้างออกมาใช้สำหรับแบ่งคำไม่ถูกต้องได้ ซึ่งทำให้การแบ่งคำผิดพลาดไป แต่การใช้คอนดิชันนัลแรน-ดอมฟิลด์สที่เป็นแบบจำลองแบบดิสคริมิเนทิฟ (Discriminative Model) นั้นมีแบบจำลองความน่าจะเป็นแบบการแจกแจงความน่าจะเป็นแบบมีเงื่อนไข p(y|x) ทำให้ไม่เกิดปัญหาดังกล่าว (อธิบายทฤษฎีของคอนดิชันนัลแรนดอมฟิลด์สในหัวข้อ 2.1.1)

วิธีการทำงานคือ เมื่อได้รับสายอักขระเข้ามาแล้ว ทำการสร้างแบบการแบ่งคำที่มีหน้าที่คำ (Part of Speech) กำกับอยู่ด้วยในทุกแบบที่เป็นไปได้ จากนั้นหาแบบการแบ่งคำที่ดีหรือเหมาะสมที่สุด (Optimal Path) ซึ่งวิธีการหาแบบแบ่งคำที่เหมาะสมที่สุดมีการทดลอง 2 วิธีคือ การค้นหาเส้นทางแบบวิเทอร์บิ (Viterbi score) และการประเมินความเชื่อมั่น (Confidence Estimation) เพื่อนำมาเปรียบเทียบประสิทธิภาพที่ได้ในแต่ละวิธี

รายละเอียดของกรอบการทำงาน 2 ส่วนหลักๆ มีดังนี้

2.3.4.1 สร้างเส้นทางหรือแบบการแบ่งคำทั้งหมดที่เป็นไปได้

สร้างเส้นทางทั้งหมดที่เป็นไปได้โดยใช้วิธีจับคำในพจนานุกรมด้วยคำที่ยาวที่สุดและเทคนิคย้อนรอย คือเริ่มจากการสร้างเส้นทางแรกก่อนโดยไล่ในสายอักขระตั้งแต่ซ้ายไปทางขวาโดยจับคำที่ยาวที่สุดในพจนานุกรม จากนั้นไล่ไปเรื่อยๆ แต่ถ้าไม่พบคำในพจนานุกรม จะข้ามตัวอักษรนั้นไปเริ่มที่ตัวอักษรถัดไป ไล่ทำไปเรื่อยๆ จนจบสายอักขระ จากนั้นหาทุกแบบที่เป็นไปได้โดยทำการย้อนรอย คือนำทุกการแบ่งคำของเส้นทางแรกมาจับคำในพจนานุกรมอีก แต่เป็นคำที่มีความยาวสั้นลงกว่าคำเดิม ถ้าไม่พบคำที่สั้นลงกว่าแล้ว ก็ให้คงคำเดิมไว้ แต่ถ้าพบคำที่สั้นลงก็ให้จับเป็นคำแรกและจับคำถัดๆ ไปอีกด้วยคำที่ยาวที่สุด หลังจากที่ได้แบ่งคำทุกแบบที่เป็นไปได้แล้วทำการกำกับหน้าที่คำทุกหน้าที่คำ

2.3.4.2 หาเส้นทางหรือแบบการแบ่งคำที่ดีและเหมาะสมที่สุด (Optimal Path)

นำทุกแบบการแบ่งคำที่ได้มาทำการหาเส้นทางที่เหมาะสมที่สุดโดยใช้ 2 วิธี คือ การค้นหาเส้นทางแบบวิเทอร์บิ และการประเมินความเชื่อมั่น สำหรับการค้นหาเส้นทางแบบวิเทอร์บิ หาได้จากขณะที่ท่องโหนดของเส้นทางต่างๆ ที่มีค่าผลคูณของค่าความน่าจะเป็นของทุกโหนดที่ผ่านมาและนำค่าผลคูณของทุกเส้นทางเปรียบเทียบหาค่าผลคูณที่มากที่สุดหนึ่งเส้นทางเก็บไว้ท่องโหนดต่อไป แต่เส้นทางที่ได้ค่าผลคูณต่ำที่เหลือนั้นถูกตัดทิ้งไป แล้วเดินต่อยังโหนดระดับถัดไปและคำนวณผลคูณค่าความน่าจะเป็นอีก แล้วตัดทิ้งตัวที่มีผลคูณน้อยกว่าออกไป ทำไปเรื่อยๆ จนเหลือสุดท้ายเพียงหนึ่งเส้นทางที่ได้ค่าผลคูณสูงที่สุด ซึ่งถือว่าเป็นเส้นทางที่เหมาะสมที่สุดที่ใช้เป็นผลลัพธ์การแบ่งคำ แต่ในการค้นหาเส้นทางแบบวิเทอร์บิ อาจทำให้สุดท้ายเลือกเส้นทางที่ไม่ถูกต้อง คาดว่าเกิดจากการที่มีคำที่กำกวมอยู่ จึงเสนอการหาค่าการประเมินความเชื่อมั่นอีกหนึ่งวิธี โดยคำนวณจากค่าผลบวกของค่าความน่าจะเป็นของเส้นทางนั้นหารด้วยค่าผลบวกของค่าความน่าจะเป็นของทุกเส้นทางที่เป็นไปได้

สำหรับผลการทดลองนี้แสดงให้เห็นว่าการใช้การประเมินความเชื่อมั่นในการหาเส้นทางที่ดีที่สุดนั้นได้ผลการแบ่งคำที่ถูกต้องมากกว่าการใช้การค้นหาเส้นทางแบบวิเทอร์บิ เนื่องจากการประเมินความเชื่อมั่นนั้นช่วยเลือกแบบการแบ่งคำได้ดีในกรณีที่พบคำกำกวมและคำที่ไม่รู้จักด้วย

ผู้วิจัยเห็นว่าการแบ่งคำด้วยแบบจำลองคอนดิชันนัลแรนดอมฟิลด์สนั้นมีข้อดีที่สามารถปรับค่าพารามิเตอร์น้ำหนักไปตามคลังข้อความฝึกฝนได้ แต่การเลือกคุณลักษณะที่ใช้ในระดับคำจากการจับคำในพจนานุกรมหรือคลังข้อความยังถูกบั่นทอนประสิทธิภาพด้วยข้อเสียจากการไม่พบคำในพจนานุกรมหรือคลังข้อความอยู่ งานวิจัยนี้จึงเสนอให้ใช้อักขระเป็นคุณลักษณะ

แทนการใช้ระดับคำ ซึ่งจะลดข้อเสียจากการไม่พบคำในพจนานุกรมแล้วยังได้ใช้ประโยชน์จากกฎหรือข้อจำกัดของตัวอักขระเป็นคุณลักษณะฝึกฝนในคอนดิชันนัลแรนดอมฟิลด์สด้วย

บทที่ 3
การแบ่งคำภาษาไทยโดยใช้คอนดิชันนัลแรนดอมฟิลด์สด้วยข้อมูลระดับอักขระ

3.1 หลักเกณฑ์การแบ่งคำ

ปัญหาหนึ่งของการวิจัยแบ่งคำภาษาไทยที่ผ่านมา คือหลักเกณฑ์หรือแนวทางในการแบ่งคำภาษาไทยที่ไม่ได้ใช้มาตรฐานเดียวกันหรือวิธีเดียวกัน เพราะในการวัดประสิทธิภาพการแบ่งคำของแต่ละวิธี หากแบ่งคำโดยใช้หลักเกณฑ์การแบ่งคำต่างกันจะดูไม่สมเหตุสมผลในการเปรียบเทียบผล ผู้วิจัยจึงเลือกใช้หลักเกณฑ์การแบ่งคำเดียวกับ “แนวทางการแบ่งคำภาษาไทยสำหรับ BEST 2009” [10] ซึ่งจัดเตรียมโดยเนคเทค เพราะมีแนวทางในการแบ่งคำอย่างชัดเจนและละเอียด เช่น วิธีการแบ่งคำตามลักษณะการประกอบคำ เช่น คำมูล คำประสม วิธีการแบ่งเครื่องหมาย วิธีการแบ่งชื่อเว็บไซต์และจดหมายอิเล็กทรอนิกส์ เป็นต้น

“แนวทางการแบ่งคำภาษาไทยสำหรับ BEST 2009” เป็นการแบ่งคำให้เป็นคำที่เล็กที่สุดที่มีความหมาย โดยพิจารณาคำที่พบในข้อความว่า ถ้าคำที่ถูกแบ่งออกมานั้นยังคงเค้าความหมายเดิมของคำ แสดงว่าคำนั้นสามารถแบ่งย่อยได้ แต่ถ้าคำที่ถูกแบ่งออกมานั้นมีความหมายเปลี่ยนไปจากเดิมมาก จนไม่คงเค้าความหมายเดิมของคำ แสดงว่า ไม่ควรแบ่งคำนั้น

3.2 การแบ่งคำภาษาไทยโดยใช้คอนดิชันนัลแรนดอมฟิลด์ส

งานวิจัยนี้นำเสนอการใช้ข้อจำกัดระดับอักขระเป็นคุณลักษณะคอนดิชันนัลแรน-ดอมฟิลด์สเพื่อการแบ่งคำภาษาไทย ซึ่งมีประโยชน์ในกรณีที่พบคำที่ไม่รู้จักหรือไม่ปรากฏในพจนานุกรมหรือคลังข้อความ

3.2.1 กำหนดคุณลักษณะที่ใช้ในคอนดิชันนัลแรนดอมฟิลด์ส

จากการสังเกตข้อจำกัดหรือลักษณะการผสมตัวอักขระไทยดังที่แสดงรายละเอียดในบทที่ 2 ข้อ 2.2 ผู้วิจัยได้สังเกตการเขียนคำไทยนั้นประกอบจากการผสมอักษรหรืออักขระไทย ที่มีทั้งพยัญชนะ สระ วรรณยุกต์ ตัวเลข และเครื่องหมายวรรคตอน ถึงแม้จะไม่มีกฎเกณฑ์การแบ่งคำที่ชัดเจน แต่ในการเขียนก็ยังสังเกตเห็นข้อจำกัดหรือกฎในการผสมตัวอักษรแต่ละประเภทให้เป็นคำได้ เช่น สระ “ะ”, “ั” และ “ิ” ต้องตามหลังตัวอักษรและไม่สามารถเป็นอักษรตัวแรกของคำได้เสมอ และ สระ “ั” ก็ยังต้องมีพยัญชนะตามหลังหรืออาจมีวรรณยุกต์ก่อนแล้วพยัญชนะตามหลังเสมอ เป็นต้น ผู้วิจัยเห็นประโยชน์การใช้อักขระ จึงเสนอให้ใช้ 2 คุณลักษณะ ดังนี้

1. ใช้อักขระเป็นคุณลักษณะ
2. ใช้กลุ่มอักขระเป็นคุณลักษณะ

เนื่องจากอักขระไทยมีข้อจำกัดหรือมีหน้าที่ในการใช้งานแตกต่างกันไปในแต่ละตัว ดังแสดงตัวอย่างข้อจำกัดในตารางที่ 1 [7] ทำให้การใช้อักขระเป็นคุณลักษณะที่มีประโยชน์ในการฝึกฝนการแบ่งคำในคลังข้อความฝึกฝน

การจัดกลุ่มอักขระตามข้อจำกัดหรือหน้าที่การใช้งานของอักขระเพื่อใช้เป็นคุณลักษณะฝึกฝนยังเพิ่มโอกาสที่จะเลือกแบบการแบ่งคำได้ถูกต้องมากยิ่งขึ้นด้วยโดยจะเห็นได้ชัดเจนในกรณีที่พบคำที่ไม่รู้จัก แสดงตัวอย่างในตารางที่ 2 เช่น คำว่า "นันทา" เป็นคำที่ไม่รู้จัก แต่มีการใช้แบบรูปหน้าที่อักขระไทย (Character Function Pattern) แบบเดียวกับคำว่า "พัฒนา" หรือ "ศิลปะ" ที่พบได้ในคลังข้อความฝึกฝนก็จะมีโอกาสที่จะแบ่งคำได้ถูกต้องมากยิ่งขึ้น

ตารางที่ 1 ตัวอย่างกฎหรือข้อจำกัดของอักขระที่มีผลต่อการแบ่งคำ

กฎหรือข้อจำกัด	***รายละเอียดการแบ่ง***
กฎทัณฑฆาต	ไม่แบ่งเมื่อพบ พยัญชนะ+ทัณฑฆาต หรือพบ พยัญชนะ+สระอิ+ทัณฑฆาต หรือพบ พยัญชนะ+ พยัญชนะ+ทัณฑฆาต
กฎสระตาม	ไม่แบ่งหน้าสระตาม
กฎอักขระเดียว	ไม่แบ่งหลังสระนำ หรือ วรรณยุกต์
กฎวรรณยุกต์	เมื่อพบ สระนำ+พยัญชนะต้น+วรรณยุกต์ และ พยัญชนะต้น+สระตามหรือวรรณยุกต์หรือไม้ไต่คู้ ให้แบ่งพยางค์หลังวรรณยุกต์

ตารางที่ 2 ตัวอย่างแบบรูปหน้าที่อักขระของคำ

แบบรูปหน้าที่อักขระ *(Character Function Pattern)*	***คำ*** *(Example Word)*
C+VU+C+C+VR	ศิลปะ พัฒนา รัชดา นันทา
VB+C+T+VR	เก่ง โน้น แห้ง เป้ง
VB+C+C+C+VR	โภชนา เกษรา เจษฎา ไอรดา

3.2.2 กำหนดกลุ่มหน้าที่อักขระภาษาไทย

สำหรับคุณลักษณะคอนดิชันนัลแรนดอมฟิลด์สที่ถูกใช้ในงานวิจัยนี้ มีทั้งอักขระและกลุ่มหน้าที่อักขระภาษาไทย ในส่วนของกลุ่มหน้าที่อักขระได้มีการแบ่งกลุ่มตามลักษณะการใช้งานได้ดังนี้

ตารางที่ 3 กลุ่มหน้าที่ของอักขระภาษาไทย

หน้าที่ของอักขระ	***คำอธิบาย***	***อักขระ***
Spc	เว้นวรรค	(Space)
Quote	เครื่องหมายคำพูด (อัญประกาศ)	“ ‘
LatinNum	ตัวเลขอารบิค	0 1 2 3 4 5 6 7 8 9
ThaiNum	ตัวเลขไทย	๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙
C	พยัญชนะ	ก – ฮ
VB	สระด้านหน้า	เ แ โ ใ ไ
VR	สระด้านหลัง	ะ า ำ
VU	สระด้านบน	◌ั ◌ิ ◌ี ◌ึ ◌ื ◌็
VL	สระด้านล่าง	◌ุ ◌ู
VO	สระลอย	ฤ ฦ
T	วรรณยุกต์	◌่ ◌้ ◌๊ ◌๋
SymbTH	สัญลักษณ์ที่ใช้ในภาษาไทยเท่านั้น	◌์ ◌. ◌ๆ ฯ ◌◌ ฯ
Symb	สัญลักษณ์ที่ใช้ได้ทั้งภาษาไทยและอังกฤษ	฿ $ # & % ( ) [ ] {} . , : ; ! ? ' / - — _
En	อักขระภาษาอังกฤษ	a – z, A - Z
Oth	อักขระอื่นๆ นอกเหนือจากข้างบน	

3.2.3 กำหนดเลเบล (Label) ที่ใช้กำกับอักขระเพื่อบ่งบอกขอบเขตของคำในคอนดิชันนัลแรนดอมฟิลด์ส

คอนดิชันนัลแรนดอมฟิลด์สถูกใช้ทำการกำกับตัวอักขระด้วยเลเบล ผู้วิจัยกำหนดไว้ 4 เลเบล ดังนี้

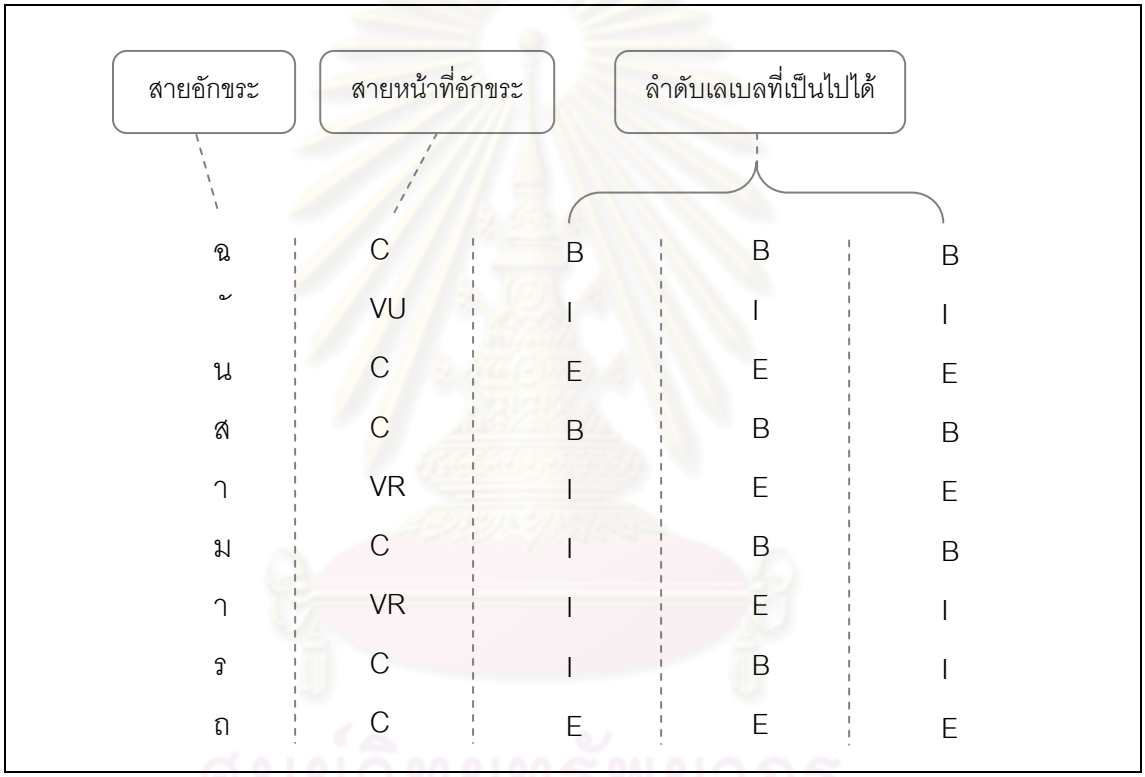
ก. B ใช้กำกับอักขระที่อยู่ในตำแหน่งแรกของคำ

ข. S ใช้กำกับอักขระที่เป็นคำที่มีอักขระเดียว

ค. I ใช้กำกับอักขระที่อยู่ในตำแหน่งกลางคำ

ง. E ใช้กำกับอักขระที่อยู่ในตำแหน่งสุดท้ายของคำ

ขอบเขตของคำนั้นเริ่มต้นที่ตำแหน่งด้านหน้าของอักขระที่ถูกกำกับด้วยเลเบล B และลงท้ายคำที่ด้านหลังของอักขระที่ถูกกำกับด้วยเลเบล E และแบ่งขอบเขตคำทั้งด้านหน้าและหลังอักขระที่ถูกกำกับด้วยเลเบล S



รูปที่ 2 ตัวอย่างของสายอักขระและหน้าที่สายอักขระถูกกำกับด้วยเลเบลในแบบที่เป็นไปได้

อธิบายการกำกับเลเบลพร้อมตัวอย่างดังรูปที่ 2

ก. คอลัมน์แรก คือ สายอักขระประโยค “ฉันสามารถ”

ข. คอลัมน์ที่ 2 คือ สายหน้าที่อักขระของสายอักขระประโยค “ฉันสามารถ” (รายละเอียดการจัดกลุ่มหน้าที่อักขระอยู่ที่ข้อ 3.2.2)

ค. คอลัมน์ที่ 3, 4 และ 5 คือ เลเบลที่กำกับสายอักขระและหน้าที่สายอักขระของประโยค “ฉันสามารถ” ที่สามารถเป็นไปได้ (อาจมีมากกว่า 3 แบบ)

3.2.4 สร้างเทมเพลตคอนดิชันนัลแรนดอมฟิลด์ส

เทมเพลตคอนดิชันนัลแรนดอมฟิลด์สคือแบบคุณลักษณะและลำดับที่ใช้สำหรับฝึกฝนในคอนดิชันนัลแรนดอมฟิลด์ส ผู้วิจัยทำการจัดเตรียมเทมเพลตคอนดิชันนัลแรนดอมฟิลด์สไว้ 2 ชุด

ก. เทมเพลต FT_a ใช้ทั้งอักขระและกลุ่มหน้าที่อักขระเป็นคุณลักษณะ

สำหรับเทมเพลต FT_a นี้ใช้อักขระ 7 แกรม และหน้าที่อักขระ 11 แกรม ซึ่งมีการพิจารณาทีละ 4 ตำแหน่ง ได้ทั้งหมด 12 คุณลักษณะ โดยตั้งชื่อคุณลักษณะเป็น B01 ถึง B12 ดังรูปที่ 3

อธิบายเทมเพลต FT_a ได้ดังนี้ ให้อักขระปัจจุบันเป็น c_0 โดยมี cf_0 เป็นหน้าที่อักขระและตำแหน่งที่ i มีความหมายดังนี้

ตำแหน่งที่ i ของอักขระคือตำแหน่งด้านขวาของ c_0 และขณะเดียวกันตำแหน่งที่ -i ของอักขระคือตำแหน่งด้านซ้ายของ c_0

ตำแหน่งที่ i ของหน้าที่อักขระคือตำแหน่งด้านขวาของ cf_0 และขณะเดียวกันตำแหน่งที่ -i ของหน้าที่อักขระคือตำแหน่งด้านซ้ายของ cf_0

B01: c_{-3}, c_{-2}, c_{-1}, c_0

B02: c_{-2}, c_{-1}, c_0, c_1

B03: c_{-1}, c_0, c_1, c_2

B04: c_0, c_1, c_2, c_3

B05: cf_{-5}, cf_{-4}, cf_{-3}, cf_{-2}

B06: cf_{-4}, cf_{-3}, cf_{-2}, cf_{-1}

B07: cf_{-3}, cf_{-2}, cf_{-1}, cf_0

B08: cf_{-2}, cf_{-1}, cf_0, cf_1

B09: cf_{-1}, cf_0, cf_1, cf_2

B10: cf_0, cf_1, cf_2, cf_3

B11: cf_1, cf_2, cf_3, cf_4

B12: cf_2, cf_3, cf_4, cf_5

รูปที่ 3 เทมเพลต FT_a ใช้ทั้งอักขระและกลุ่มหน้าที่อักขระเป็นคุณลักษณะ

ข. เทมเพลต FT_b ใช้อักขระเท่านั้นที่เป็นคุณลักษณะ

สำหรับเทมเพลต FT_b นี้ใช้อักขระ 7 แกรม ซึ่งมีการพิจารณาทีละ 4 ตำแหน่ง ได้ทั้งหมด 4 คุณลักษณะ โดยตั้งชื่อคุณลักษณะเป็น B01 ถึง B04 ดังรูปที่ 4

B01: c_{-3}, c_{-2}, c_{-1}, c_0

B02: c_{-2}, c_{-1}, c_0, c_1

B03: c_{-1}, c_0, c_1, c_2

B04: c_0, c_1, c_2, c_3

รูปที่ 4 เทมเพลต FT_b ใช้อักขระเป็นคุณลักษณะเท่านั้น

อธิบายเทมเพลต FT_b ได้ดังนี้ ให้อักขระปัจจุบันเป็น c_0 โดยมี cf_0 เป็นหน้าที่อักขระและตำแหน่งที่ i มีความหมายดังนี้

ตำแหน่งที่ i ของอักขระคือตำแหน่งด้านขวาของ c_0 และขณะเดียวกัน ตำแหน่งที่ -i ของอักขระคือตำแหน่งด้านซ้ายของ c_0

บทที่ 4
การทดลองและผลการทดลอง

4.1 คลังข้อความ

คลังข้อความที่นำมาใช้ฝึกฝนและทดสอบในงานวิจัยนี้ถูกพัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) [9] โดยคลังข้อความประกอบด้วยข้อความที่ถูกแบ่งคำไว้แล้วเพื่อใช้เป็นข้อมูลฝึกฝน โดยมีขนาดประมาณ 7,000,000 คำ ในคลังข้อความฝึกฝนนี้ได้ถูกถูกจัดกลุ่มไว้ 8 ประเภท คือ บทความวิชาการ สารานุกรม ข่าว นวนิยาย พุทธศาสนา กฎหมาย การอภิปราย และสารานุกรมเสรีออนไลน์ (Wiki) โดยในงานวิจัยนี้ได้มีการใช้ข้อมูลจากคลังข้อความประมาณ 70% หรือ 4,900,000 คำ จากทั้งหมด 8 ประเภท (แสดงจำนวนคำที่ใช้ฝึกฝนแยกตามประเภทข้อความในตารางที่ 4) เพื่อนำไปเป็นข้อมูลฝึกฝนและหาค่า λ_1 ให้กับคอนดิชันนัลแรนดอมฟิลด์ส สาเหตุที่ใช้เพียง 70% จากทั้งหมดเนื่องจากมีข้อจำกัดของเครื่องคอมพิวเตอร์ที่ต้องประมวลผลอย่างหนัก

สำหรับข้อมูลทดสอบที่ใช้ ประมาณ 500,000 คำ โดยข้อมูลทดสอบถูกแบ่งได้ 12 ประเภท คือ บทความวิชาการ สารานุกรม ข่าว นวนิยาย พุทธศาสนา กฎหมาย การอภิปราย สารานุกรมเสรีออนไลน์ (Wiki) ข่าวโทรทัศน์ เอกสาร NSC เอกสารเก่าและข่าวในพระราชสำนัก สังเกตเห็นได้ว่ามีข้อมูล 4 ประเภทหลังที่เพิ่มเติมจากประเภทข้อมูลฝึกฝน (แสดงจำนวนคำที่ใช้ทดสอบแยกตามประเภทข้อความในตารางที่ 4)

คลังข้อความภาษาไทย BEST ที่สร้างโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ จากการนำงานเขียนที่มีลักษณะต่างๆ 3 ประเภทมารวมรวมไว้เพื่อให้เป็นตัวแทนของภาษาไทยที่ใช้กันโดยทั่วไปในปัจจุบัน ได้แก่

1. ตัวแทนของภาษาพูดทั่วไป เช่น นวนิยาย
2. ตัวแทนของภาษาเขียนอย่างเป็นทางการ เช่น สารานุกรม
3. ตัวแทนของภาษาข่าว เช่น หนังสือพิมพ์บนอินเตอร์เนต

ข้อความในเอกสารต่างๆ เหล่านี้ได้ถูกนำมาวิเคราะห์เพื่อกำหนดขอบเขตของคำโดยอาศัยหลักการทางภาษาศาสตร์โดยคำนึงถึงการประมวลผลด้วยคอมพิวเตอร์ร่วมด้วย ทั้งนี้โดยไม่มีการแก้ไขดัดแปลงเนื้อหาสาระของผลงานต้นฉบับแต่อย่างใด

ตารางที่ 4 จำนวนคำฝึกฝนและทดสอบแบ่งตามประเภทข้อความ

ประเภทข้อความ	*จำนวน (คำ) ฝึกฝน*	*จำนวน (คำ) ทดสอบ*
บทความวิชาการ	322,714	43,814
พุทธศาสนา	378,624	53,847
สารานุกรม	820,315	53,059
กฎหมาย	500,728	53,186
ข่าว	1,160,361	52,903
นวนิยาย	1,138,698	42,242
การอภิปราย	278,502	49,327
สารานุกรมเสรีออนไลน์	539,790	53,886
ข่าวโทรทัศน์	0	54,636
เอกสาร NSC	0	62,480
เอกสารเก่า	0	58,155
ข่าวในพระราชสำนัก	0	63,160

4.2 การวัดผล

งานวิจัยนี้ใช้ F-Measure เป็นวิธีที่ใช้วัดผลการแบ่งคำ มีสูตรการคำนวณ ดังนี้

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (3.1)$$

โดยที่ precision คือจำนวนคำที่แบ่งได้ถูกต้องหารด้วยจำนวนคำที่แบ่งทั้งหมด และ recall คือจำนวนคำที่แบ่งได้ถูกต้องหารด้วยจำนวนคำที่แบ่งจากคำตอบที่อ้างอิง

4.3 การทดลอง

แสดงขั้นตอนการทดลองดังรูปที่ 5 มีรายละเอียดดังนี้

1. ผู้วิจัยทำการเตรียมชุดข้อความฝึกฝน ซึ่งแต่ละอักขระถูกกำกับด้วยหน้าที่อักขระและเลเบลแล้ว (ดูรายละเอียดหน้าที่อักขระที่หัวข้อ 3.2.2 และรายละเอียดเลเบลที่หัวข้อ 3.2.3)

2. ผู้วิจัยเตรียมชุดข้อความทดสอบ โดยที่แต่ละอักขระถูกกำกับด้วยหน้าที่อักขระแล้ว จำนวน 2 ชุดที่มีข้อความเหมือนกัน (ดูรายละเอียดหน้าที่อักขระที่หัวข้อ 3.2.2)

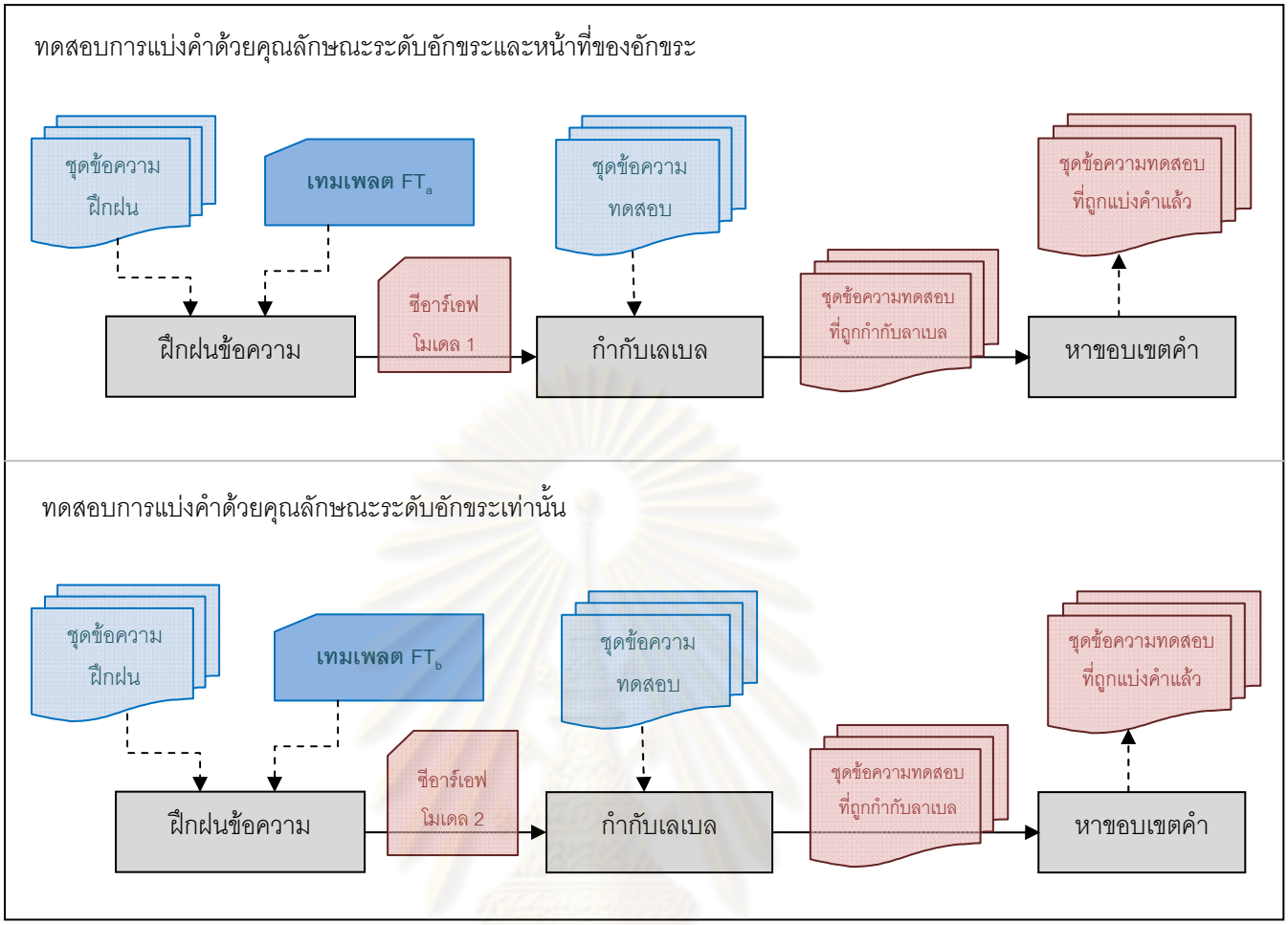
3. ผู้วิจัยนำชุดข้อความฝึกฝนเข้ากระบวนการฝึกฝนด้วยเทมเพลตคอนดิชันนัลแรนดอมฟิลด์ส FT_a (ดูรายละเอียดเทมเพลตที่หัวข้อ 3.2.4) ทำให้ได้แบบจำลองคอนดิชันนัลแรนดอมฟิลด์สที่ 1 หรือเรียกว่า ซีอาร์เอฟโมเดล 1

4. ผู้วิจัยนำชุดข้อความฝึกฝนเข้ากระบวนการฝึกฝนด้วยเทมเพลตคอนดิชันนัลแรนดอมฟิลด์ส FT_b (ดูรายละเอียดเทมเพลตที่หัวข้อ 3.2.4) ทำให้ได้แบบจำลองคอนดิชันนัลแรนดอมฟิลด์สที่ 2 หรือเรียกว่า ซีอาร์เอฟโมเดล 2

5. ผู้วิจัยนำชุดข้อความทดสอบเข้ากระบวนการกำกับเลเบลโดยใช้แบบจำลองคอนดิชันนัลแรนดอมฟิลด์สที่ 1

6. ผู้วิจัยนำชุดข้อความทดสอบเข้ากระบวนการกำกับเลเบลโดยใช้แบบจำลองคอนดิชันนัลแรนดอมฟิลด์สที่ 2

7. ผู้วิจัยนำผลลัพธ์ทั้งสองมาแบ่งขอบเขตคำด้วยการแบ่งหลังอักขระที่ถูกกำกับด้วยเลเบล E และแบ่งทั้งหน้าและหลังอักขระที่กำกับด้วย S



รูปที่ 5 กระบวนการแบ่งคำโดยใช้คอนดิชันนัลแรนดอมฟิลด์สด้วยข้อมูลระดับอักขระ

นอกจากนั้น ผู้วิจัยได้เพิ่มการทดลองกำหนดชุดเลเบลเป็นชุดที่ 2 มี 10 เลเบล คือ B, S, I, E, B-NE, I-NE, E-NE, B-AB, I-AB และ E-AB ซึ่งเป็นการกำกับขอบเขตและหน้าที่ของอักขระด้วย มีความหมายดังนี้

ก. B ใช้กำกับอักขระที่อยู่ในตำแหน่งแรกของคำ

ข. S ใช้กำกับอักขระที่เป็นคำที่มีอักขระเดียว

ค. I ใช้กำกับอักขระที่อยู่ในตำแหน่งกลางคำ

ง. E ใช้กำกับอักขระที่อยู่ในตำแหน่งสุดท้ายของคำ

จ. B-NE ใช้กำกับอักขระที่อยู่ในตำแหน่งแรกของคำที่เป็นชื่อเฉพาะ (Name Entity)

ฉ. I-NE ใช้กำกับอักขระที่อยู่ในตำแหน่งกลางคำที่เป็นชื่อเฉพาะ (Name Entity)

ช. E-NE ใช้กำกับอักขระที่อยู่ในตำแหน่งสุดท้ายของคำที่เป็นชื่อเฉพาะ (Name Entity)

ซ. B-AB ใช้กำกับอักขระที่อยู่ในตำแหน่งแรกของคำที่เป็นชื่อย่อ (Abbreviation)

ฌ. I-AB ใช้กำกับอักขระที่อยู่ในตำแหน่งกลางคำที่เป็นชื่อย่อ (Abbreviation)

ญ. E-AB ใช้กำกับอักขระที่อยู่ในตำแหน่งสุดท้ายของคำที่เป็นชื่อย่อ (Abbreviation)

ในขั้นการทดลองด้วยชุดเลเบลนี้ มีกระบวนการทดลองเหมือนข้างต้น เพียงแต่เปลี่ยนชุดกำกับเลเบลและทดลองกับเทมเพลต FT_a เท่านั้น (ใช้อักขระและหน้าที่ของอักขระเป็นคุณลักษณะ)

4.4 ตัวเปรียบเทียบ (Baseline)

สำหรับตัวเปรียบเทียบ (Baseline) เป็นการแบ่งคำโดยใช้แบบจำลองไทรแกรมของแบบจำลองฮิดเดนมาร์คอฟ (HMM-Trigram) สร้างขึ้นโดยจากการพัฒนาของผู้วิจัยเอง เพื่อที่จะได้ใช้คลังข้อความฝึกฝนชุดเดียวกันกับงานวิจัยที่นำเสนอ และใช้แนวทางการแบ่งคำเดียวกันและชุดทดสอบเดียวกัน

คลังข้อความที่นำมาใช้ฝึกฝนและทดสอบ [9] ในตัวเปรียบเทียบใช้ชุดเดียวกันกับงานวิจัยที่นำเสนอ ซึ่งประกอบด้วยข้อความที่ถูกแบ่งคำไว้แล้วเพื่อใช้เป็นข้อมูลฝึกฝน โดยมีขนาดประมาณ 4,900,000 คำ จากทั้งหมด 8 ประเภท สำหรับข้อมูลทดสอบใช้ประมาณ 500,000 คำ ซึ่งเป็นชุดเดียวกันกับงานวิจัยที่นำเสนอ โดยข้อมูลทดสอบถูกแบ่งได้ 12 ประเภท คือ บทความวิชาการ สารานุกรม ข่าว นวนิยาย พุทธศาสนา กฎหมาย การอภิปราย สารานุกรมเสรีออนไลน์ (Wiki) ข่าวโทรทัศน์ เอกสาร NSC เอกสารเก่าและข่าวในพระราชสำนัก มีข้อมูล 4 ประเภทหลังที่เพิ่มเติมจากประเภทข้อมูลฝึกฝน

ตัวเปรียบเทียบนี้จะนำสถิติเข้ามาใช้แก้ปัญหาการแบ่งคำ โดยนำโมเดลไตรแกรมเข้ามาคำนวณค่าความน่าจะเป็นของประโยคซึ่งคำนวณได้ตามสมการ ดังนี้

$$P(W) = \prod_{i=1}^{n} P(w_{i,n})$$
$$= \prod_{i=1}^{n} P(w_i \mid w_{i-1}, w_{i-2}) \qquad (4.1)$$

จากสมการ (4.1) เป็นการคำนวณหาค่าความน่าจะเป็นของแต่ละประโยค โดย W คือประโยคที่ถูกแบ่งมาแล้ว และประโยค W จะประกอบด้วยคำต่างๆ ที่ W= $w_1w_2...w_n$ โดย w_i คือคำที่เรียงต่อกัน สำหรับการคำนวณค่าความน่าจะเป็นของแต่ละประโยคจะมีสมมติฐานว่าความน่าจะเป็นของ w_i จะขึ้นอยู่กับ w_i-1 และ w_i-2 เท่านั้น

เนื่องจากการคำนวณตามสมการ (4.1) ต้องใช้คลังข้อความขนาดใหญ่ที่ควรจะมากกว่า n3 โดยที่ n คือจำนวนคำที่เป็นไปได้ทั้งหมด สาเหตุที่ต้องใช้คลังข้อความที่มีขนาดมากกว่า n3 คำ เนื่องจากวิธีนี้ต้องมีการนำค่าสถิติการเกิดของคำ 3 คำที่ติดกันมาใช้ในการคำนวณ ดังนั้นเพื่อให้มีค่าสถิติของการเกิดคำ 3 คำที่ติดกันทุกๆ แบบ อย่างน้อยที่สุดจะต้องใช้ n3 คำ จึงได้นำความน่าจะเป็นของไบแกรมและยูนิแกรมเข้ามาช่วยในการคำนวณด้วย โดยใช้สมการ ดังนี้

$$\prod_{i=1}^{n} P\left(w_i \mid w_{i-2,i-1}\right) = \prod_{i=1}^{n} \left(\lambda_1 P(w_n) + \lambda_2 P(w_n \mid w_{n-1}) + \lambda_3 P(w_n \mid w_{n-2,n-1})\right) \quad (4.2)$$

จากสมการ (4.2) มีการคำนวณความน่าจะเป็นแบบไบแกรมและยูนิแกรมเข้ามาร่วมด้วย และกำหนดค่า 0.1, 0.3, 0.6 ให้กับ $\lambda_1, \lambda_2, \lambda_3$ ตามลำดับ นั้นก็คือพารามิเตอร์น้ำหนัก

การแบ่งคำเริ่มด้วยเมื่อรับข้อความเข้ามาอ่านทีละบรรทัด แล้วนำแต่ละบรรทัดนั้นมาแบ่งตามการเว้นช่องว่างของข้อความเพื่อให้ได้ประโยค จากนั้นจะนำแต่ละประโยคมาทำการหาแบบการแบ่งคำโดยจับกับคำจากพจนานุกรม (พจนานุกรมได้จากคำที่ถูกแบ่งไว้แล้วในคลังข้อความเบส) ซึ่งในขั้นตอนหาแบบคำที่เป็นไปได้ ไม่ได้ใช้ทุกแบบที่เป็นไปได้เพราะจะมีจำนวนมากมายเกินกว่าระบบจะรองรับได้และเกินความจำเป็น เนื่องจากอาจมีความเป็นไปได้ที่ความน่าจะเป็นต่ำ จึงใช้วิธีการค้นหาเส้นทางที่ดีที่สุด N เส้นทาง (N-best Search) ดังอธิบายในบทที่ 2 หัวข้อ 2.1.2 จนท้ายสุดทำให้ได้**แบบแบ่งคำที่มีผลความน่าจะเป็นสูงที่สุดหนึ่งแบบที่เป็น**ผลลัพธ์ของประโยคนั้น

4.5 ผลการทดลอง

ผู้วิจัยเรียกผลการทดลองต่างๆ ดังนี้

1. ผลทดลองแบ่งคำจากการใช้เทมเพลต FT_a หรือใช้อักขระและหน้าที่ของอักขระเป็นคุณลักษณะ เรียกว่า CRF_{C+CF}
2. ผลการแบ่งคำจากการใช้เทมเพลต FT_b หรือใช้อักขระเท่านั้นเป็นคุณลักษณะ เรียกว่า CRF_C

3. ผลการแบ่งคำจากการใช้เทมเพลต FTa หรือใช้อักขระและหน้าที่ของอักขระเป็นคุณลักษณะด้วยชุดเลเบลที่สอง เรียกว่า CRFC+CF (L2)
4. ผลจากการแบ่งคำด้วยแบบจำลองไทรแกรมของแบบจำลองฮิดเดนมาร์คอฟ (Baseline) เรียกว่า HMM-Trigram

ตารางที่ 5 แสดงค่า F-Measure ที่ได้จากการทดลองแบ่งคำด้วยชุดข้อมูลทดสอบหลากหลายประเภท เปรียบเทียบผลการทดลอง HMM-Trigram, CRF_C และ CRF_{C+CF} สังเกตเห็นได้ว่าการแบ่งคำโดยใช้คอนดิชันนัลแรนดอมฟิลด์สมีประสิทธิภาพมากกว่าใช้แบบจำลองไทรแกรมของแบบจำลองฮิดเดนมาร์คอฟในทุกๆ ประเภทข้อความ และหากเปรียบเทียบ 2 ผลการทดลอง คือ CRF_C และ CRF_{C+CF} พบว่าการใช้หน้าที่อักขระเข้าเป็นคุณลักษณะด้วยนั้นทำให้ผลการแบ่งคำดีขึ้น ถึงแม้ว่าจะทำให้จำนวนคุณลักษณะในเทมเพลตเพิ่มขึ้นแต่ก็ยังทำให้ประสิทธิภาพการแบ่งคำดีอยู่

ตารางที่ 5 เปรียบเทียบความถูกต้องในการแบ่งคำ

ประเภทข้อความ	*HMM-Trigram*	CRF_C	CRF_{C+CF}
บทความวิชาการ	93.84%	95.49%	96.29%
พุทธศาสนา	94.81%	97.10%	97.21%
สารานุกรม	93.63%	94.85%	95.76%
กฎหมาย	93.75%	97.37%	97.40%
ข่าว	91.21%	94.42%	94.79%
นวนิยาย	93.05%	94.76%	95.02%
การอภิปราย	95.34%	97.18%	97.24%
สารานุกรมเสรีออนไลน์	86.50%	92.53%	94.54%
ข่าวโทรทัศน์	89.51%	94.50%	94.74%
เอกสาร NSC	91.93%	93.32%	95.70%
เอกสารเก่า	87.49%	91.14%	90.91%
ข่าวในพระราชสำนัก	80.77%	89.15%	90.00%
ค่าเฉลี่ย	**90.98%**	**94.32%**	**94.97%**

ตารางที่ 6 ค่า F-Measure ของประเภทข้อมูลที่เคยพบมาก่อนในข้อมูลฝึกฝน

Method	*HMM-Trigram*	CRF_C	CRF_{C+CF}
Mean	92.77%	95.46%	96.03%
Max	95.34%	97.37%	97.40%
Min	86.50%	92.53%	94.54%

ตารางที่ 7 ค่า F-Measure ของประเภทข้อมูลที่ไม่เคยพบมาก่อนในข้อมูลฝึกฝน

Method	*HMM-Trigram*	CRF_C	CRF_{C+CF}
Mean	87.43%	92.03%	92.84%
Max	91.93%	94.50%	95.70%
Min	80.77%	89.15%	90.00%

ตารางที่ 6 และตารางที่ 7 แสดงให้เห็นว่าเมื่อมีการแบ่งคำประเภทที่ไม่เคยพบจากในชุดข้อความฝึกฝนมาก่อนทำให้มีผลต่อประสิทธิภาพในการแบ่งคำของข้อมูลทดสอบด้วย ซึ่งสังเกตเห็นจากค่าเฉลี่ย F-Measure ลดลงในกรณีที่ทดลองเพิ่มประเภทข้อมูลที่ไม่เคยพบมาก่อน

สำหรับ HMM-Trigram ค่าเฉลี่ยของ F-Measure นั้นลดลงมากกว่า 5% จาก 92.77% เป็น 87.43% เมื่อคำนวณจากผลการแบ่งคำกับข้อความที่ไม่เคยพบมาก่อนในข้อมูลฝึกฝน แต่สำหรับ CRF_C และ CRF_{C+CF} ลดลงเพียงประมาณ 3% เท่านั้น จึงเห็นว่ามีประสิทธิภาพการแบ่งคำดีกว่า อีกทั้งเมื่อเปรียบเทียบความต่างระหว่าง HMM-Trigram กับ CRF ทั้ง 2 แบบ พบว่ามีความต่างกันประมาณ 2.7% ถึง 3.2% เท่านั้นในกรณีที่เคยพบประเภทข้อมูลในคลังข้อมูลฝึกฝนแล้ว ในขณะที่ความต่างมีมากขึ้นประมาณ 4.6% ถึง 5.4% เมื่อเทียบกับการทดสอบกับประเภทข้อมูลที่ไม่เคยพบมาก่อนในข้อมูลฝึกฝน สิ่งนี้ชี้ให้เห็นว่าการใช้คุณลักษณะระดับอักขระมีความเสถียรในการแบ่งคำประเภทใหม่ๆ มากกว่าการใช้คุณลักษณะในระดับคำ เพราะในกรณีของการใช้ HMM-Trigram นั้นโอกาสที่จะพบคำ 3 คำที่อยู่ลำดับติดกันมีน้อยกว่าโอกาสที่จะพบลำดับที่ติดกันของอักขระหรือหน้าที่อักขระ

การทดสอบข้ามประเภทข้อมูลนั้นมีผลอย่างมากเมื่อใช้การพิจารณาระดับคำ ขณะที่ลำดับอักขระและหน้าที่อักขระนั้นไม่มีผลมากกับการทดลองข้ามประเภทข้อมูล

ตารางที่ 8 เปรียบเทียบความถูกต้องในการแบ่งคำระหว่าง CRF_{C+CF} และ $CRF_{C+CF\ (L2)}$

ประเภทข้อความ	CRF_{C+CF}	$CRF_{C+CF\ (L2)}$
บทความวิชาการ	96.29%	96.73%
พุทธศาสนา	97.21%	97.43%
สารานุกรม	95.76%	96.48%
กฎหมาย	97.40%	98.11%
ข่าว	94.79%	95.54%
นวนิยาย	95.02%	95.07%
การอภิปราย	97.24%	97.48%
สารานุกรมเสรีออนไลน์	94.54%	94.62%
ข่าวโทรทัศน์	94.74%	95.60%
เอกสาร NSC	95.70%	96.15%
เอกสารเก่า	90.91%	91.44%
ข่าวในพระราชสำนัก	90.00%	91.75%
ค่าเฉลี่ย	**94.97%**	**95.53%**

ตารางที่ 8 แสดงการเปรียบเทียบระหว่าง CRF_{C+CF} และ $CRF_{C+CF\ (L2)}$ พบว่าการแบ่งคำด้วยเลเบลที่กำหนดแบบละเอียดด้วยหน้าที่คำชื่อเฉพาะและตัวอักษรย่อด้วย ทำให้ผลการแบ่งคำถูกต้องมากขึ้นในทุกๆ ประเภทข้อความโดยมีค่าเฉลี่ย F-Measure ที่ 95.53% เมื่อเทียบกับการใช้เลเบลชุดเดิมได้ค่าเฉลี่ย F-Measure ที่ 94.97%

วิเคราะห์ผลการทดลองโดยการสุ่มคำที่ไม่เคยพบในข้อความฝึกฝนมาก่อนจำนวน 62 คำ พบว่าผลจาก $CRF_{C+CF\ (L2)}$ มักจะแบ่งคำได้เป็นคำที่ยาวกว่าเมื่อเทียบกับผลจาก CRF_{C+CF} ที่ใช้ชุดเลเบลเดิม

ใน $CRF_{C+CF\ (L2)}$ ที่ใช้ชุดเลเบลที่ 2 ในการแบ่งคำยาวๆ เช่น "องค์กรสิทธิมนุษยชนและคณะกรรมการอิสระเพื่อความสมานฉันท์แห่งชาติ" ถูกแบ่งเป็น “องค์กรสิทธิมนุษยชน|และ|คณะกรรมการอิสระเพื่อความสมานฉันท์แห่งชาติ|“ ในขณะที่การใช้ชุดเลเบลเดิมใน CRF_{C+CF} แบ่งเป็นได้ 9 คำ อย่างไรก็ตามในคำเดียว เช่น "กรุงศรีสัตนาคนหุตพระมหาธรรมราชา" ใน $CRF_{C+CF\ (L2)}$ แบ่งเป็น 2 คำ คือ “กรุงศรีสัตนาคนหุต|พระมหาธรรมราชา“ แต่ใน CRF_{C+CF} ที่ใช้ชุดเลเบลเดิมนั้นแบ่งได้ถูกต้อง

ตารางที่ 9 จำนวนคำที่ใช้ฝึกฝนและทดสอบเปรียบเทียบกับค่า F-Measure

ประเภทข้อความ	***จำนวน (คำ) ฝึกฝน***	***จำนวน (คำ) ทดสอบ***	***จำนวน(คำ) ที่ไม่รู้จักที่พบในข้อความทดสอบ***	***F-Measure*** ($CRF_{C+CF\ (L2)}$)
กฎหมาย	500,728	53,186	362	98.11%
การอภิปราย	278,502	49,327	287	97.48%
พุทธศาสนา	378,624	53,847	365	97.43%
บทความวิชาการ	322,714	43,814	739	96.73%
สารานุกรม	820,315	53,059	1,024	96.48%
เอกสาร NSC	-	62,480	1,975	96.15%
ข่าว	1,160,361	52,903	1,172	95.54%
ข่าวโทรทัศน์	-	54,636	1,236	95.60%
นวนิยาย	1,138,698	42,242	693	95.07%
สารานุกรมเสรีออนไลน์	539,790	53,886	2,417	94.62%
ข่าวในพระราชสำนัก	-	63,160	2,999	91.75%
เอกสารเก่า	-	58,155	2,522	91.44%

ตารางที่ 9 แสดงจำนวนคำที่ใช้ฝึกฝนและทดสอบเปรียบเทียบกับผลต่อความถูกต้องในการแบ่งคำโดยเรียงลำดับค่า F-Measure สูงที่สุดจนถึงต่ำที่สุด เห็นได้ว่าจำนวนคำที่ใช้ฝึกฝนในแต่ละประเภทข้อความมีจำนวนไม่เท่ากัน โดยประเภทข่าวและนวนิยายนั้นมีจำนวนคำฝึกฝนมากกว่าสองเท่าเมื่อเทียบกับข้อความประเภทอื่น แต่ไม่ได้ทำให้ผลการแบ่งคำในข้อความทดสอบประเภทข่าวและนวนิยายจะได้ผลความถูกต้องมากที่สุด แต่กลับเป็นข้อความประเภทกฎหมาย การอภิปราย พุทธศาสนา และบทความทางวิชาการมีการแบ่งคำที่ถูกต้องมากกว่า ทั้งๆ ที่มีคำฝึกฝนน้อยกว่าถึงครึ่งหนึ่งของข้อความประเภทข่าวและนวนิยาย และเมื่อนำจำนวนคำที่ไม่รู้จักที่พบในข้อความฝึกฝนมาพิจารณาร่วมด้วย จะเห็นได้ว่าจำนวนคำที่ไม่รู้จักนั้นมีผลกับการแบ่งคำตามสัดส่วน คือยิ่งพบข้อความที่ไม่รู้จักมากก็จะทำให้ผลการแบ่งคำผิดมากไปตามลำดับ

ตารางที่ 10 ผลการแบ่งคำแบบรวมและแบบแยกเฉพาะคำที่ไม่รู้จัก

ประเภทข้อความ	รวมคำที่รู้จักและไม่รู้จัก (Known Word & Unknown Wod)			เฉพาะคำที่ไม่รู้จัก (Unknown Word)		
	Recall	Precision	F1	Recall	Precision	F1
กฎหมาย	98.40%	97.82%	98.11%	43.37%	16.32%	23.71%
การอภิปราย	97.39%	97.57%	97.48%	72.47%	29.55%	41.97%
พุทธศาสนา	97.63%	97.20%	97.43%	35.89%	12.83%	18.90%
บทความวิชาการ	97.26%	96.20%	96.73%	62.25%	29.66%	40.17%
สารานุกรม	96.60%	96.36%	96.48%	64.75%	37.08%	47.16%
เอกสาร NSC	96.01%	96.29%	96.15%	71.75%	51.85%	60.20%
ข่าว	96.13%	94.95%	95.54%	59.47%	27.88%	37.96%
ข่าวโทรทัศน์	96.04%	95.17%	95.60%	61.25%	30.87%	41.05%
นวนิยาย	95.46%	94.68%	95.07%	47.04%	20.06%	28.13%
สารานุกรมเสรีออนไลน์	95.07%	94.18%	94.62%	61.81%	39.63%	48.29%
ข่าวในพระราชสำนัก	93.76%	89.82%	91.75%	54.22%	22.86%	32.15%
เอกสารเก่า	91.51%	91.38%	91.44%	43.26%	24.71%	31.40%

จากตารางที่ 10 เมื่อสังเกตค่า Recall Precision และ F1 ทั้งแบบรวมและคิดเฉพาะคำที่ไม่รู้จัก จะเห็นได้ว่าค่า Precision จากการแบ่งคำนั้นน้อยกว่า Recall ทำให้ค่า F1 ลดน้อยลง แสดงว่าการแบ่งคำที่ผิดส่วนใหญ่นั้นเกิดจากการแบ่งหนึ่งคำให้ย่อยเป็นหลายคำมากกว่าที่จะรวบคำหลายคำเป็นคำเดียวกัน เช่น "องค์กรสิทธิมนุษยชนและคณะกรรมการอิสระเพื่อความสมานฉันท์แห่งชาติ" ถูกแบ่งเป็น “องค์กรสิทธิมนุษยชน|และ|คณะกรรมการอิสระเพื่อความสมานฉันท์แห่งชาติ|“

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้เสนอการแบ่งคำภาษาไทยโดยใช้คอนดิชันนัลแรนดอมฟิลด์สและใช้ข้อมูลภาษาไทยในระดับอักขระเป็นคุณลักษณะในการฝึกฝนเพื่อให้การแบ่งคำในกรณีที่พบคำที่ไม่รู้จักและคำกำกวมได้มีประสิทธิภาพมากยิ่งขึ้น จากผลการทดลองและวิจัยสามารถสรุปได้ดังนี้

1. การแบ่งคำโดยใช้อักขระเป็นคุณลักษณะมีประสิทธิภาพมากกว่าการใช้คำเป็นคุณลักษณะ เพราะการใช้อักขระจะช่วยแก้ปัญหาเมื่อพบคำที่ไม่รู้จักได้ดีกว่าการใช้คำเป็นคุณลักษณะ

2. การใช้หน้าที่อักขระเข้าเป็นคุณลักษณะเพิ่มจากการใช้อักขระเพียงอย่างเดียวนั้นทำให้ผลการแบ่งคำดีขึ้น ถึงแม้ว่าจะทำให้จำนวนคุณลักษณะในเทมเพลตเพิ่มขึ้นแต่ก็ยังทำให้ประสิทธิภาพการแบ่งคำดีอยู่

3. เมื่อทดสอบกับการแบ่งคำประเภทใหม่ๆ ที่ไม่เคยฝึกฝนมาก่อน พบว่า การแบ่งคำโดยใช้อักขระเป็นคุณลักษณะมีความเสถียรในการแบ่งคำประเภทใหม่ๆ มากกว่าการใช้คุณลักษณะในระดับคำ เพราะเมื่อใช้คำเป็นคุณลักษณะ การแบ่งคำประเภทข้อมูลที่ไม่เคยพบในตอนฝึกฝนนั้นมีประสิทธิภาพลดลงมากกว่าการใช้หน้าที่อักขระเข้าเป็นคุณลักษณะ

4. การกำหนดเลเบลที่ใช้กำหนดขอบเขตตัวอักขระมีผลต่อประสิทธิภาพในการแบ่งคำ โดยเมื่อทดลองกำหนดเลเบลให้ระบุรู้ถึงหน้าที่คำ เช่น เป็นคำชื่อเฉพาะ คำที่เป็นอักษรย่อ ทำให้มีประสิทธิภาพมากกว่าไม่ได้ระบุรู้หน้าที่คำ

5.2 ข้อเสนอแนะ

1. จากการวิเคราะห์คำชื่อเฉพาะ (Name Entity) และเป็นคำที่ไม่รู้จักด้วย ถูกแบ่งคำทั้งๆ ที่เป็นคำเดียวกัน เช่น “โครงการพัฒนาลุ่มน้ำ คลองหอยโข่ง-คลองขำไทย” ถูกแบ่งคำได้เป็น “โครงการ|พัฒนา|ลุ่ม|น้ำ| |คลองหอยโข่ง|-|คลองขำไทย|” งานวิจัยในอนาคตอาจมีการนำคลังข้อความชื่อเฉพาะเข้ามาช่วยในการหาคำชื่อเฉพาะด้วย เพื่อลดความผิดพลาดในการแบ่งคำ (รวมถึงคลังข้อความอักษรย่อด้วย)

2. ให้มีการสร้างกฎที่ตายตัวเพื่อตรวจสอบหลังการแบ่งคำจากคอนดิชันนัลแรนดอมฟิลด์สเพื่อปรับแก้ไขให้ถูกต้อง เช่น จากการวิเคราะห์คำภาษาอังกฤษและตัวเลขที่เป็นคำเดียวกัน แต่ถูกแบ่งย่อยออกมา เช่น “Benz| |E230|” ถูกแบ่งเป็น “Benz| |E|230|“ สังเกตคำว่า “E230” ถูกแบ่งภายในคำด้วย ทำให้การแบ่งคำผิดไป หากมีกฎมาคุมและใช้ปรับแก้ไขทีหลังจะทำให้การแบ่งคำถูกต้องขึ้นอีก
3. ในการแบ่งกลุ่มหน้าที่ของอักขระที่ใช้เป็นคุณลักษณะควรแบ่งด้วยลักษณะหรือพฤติกรรมการนำไปผสมคำให้ดีขึ้นกว่าเดิม เช่น สระ “ะ” และ “ำ” ที่อยู่ในกลุ่ม VR กลุ่มเดียวกับ สระ “า” ที่ถูกแบ่งเป็นกลุ่มเดียวกันเพราะเป็นสระตามหลัง แต่จะเห็นว่าสระ “ะ” และ “ำ” นั้นมีความเป็นไปได้ที่จะถูกแบ่งหลังตัวมันเป็นคำได้มากกว่า สระ “า” ซึ่งอาจจะมีตัวอักษรอื่นตามเป็นตัวสะกดเหมือนสระที่อยู่ในกลุ่ม VU (◌ั ◌ิ ◌ี ◌ึ ◌ื ◌็) และ VL (◌ุ ◌ู) จึงมองว่าถ้าสระ “า” ถูกจัดกลุ่มอยู่ในกลุ่มเดียวกับ VU และ VL อาจจะดีกว่า เป็นต้น
4. จากการใช้ระดับอักขระเป็นคุณลักษณะเพื่อประโยชน์ในการแบ่งคำที่ไม่รู้จักได้ถูกต้องขึ้น แต่ทำให้บางคำที่เป็นคำที่รู้จักหรือเคยเห็นมาก่อนในคลังข้อความฝึกฝนนั้นแบ่งผิดไป ดังนั้นเพื่อให้ทั้งคำที่รู้จักและไม่รู้จักแบ่งได้ถูกต้องจึงจำเป็นต้องใช้ทั้งระดับอักขระและระดับคำมาเป็นคุณลักษณะร่วมกันด้วย เพราะการใช้ระดับคำนั้นให้ประโยชน์กับการแบ่งคำที่รู้จักและการใช้ระดับอักขระนั้นให้ประโยชน์กับการแบ่งคำที่ไม่รู้จัก ดังในงานวิจัยของฆนาศัย [18] ที่แบ่งคำโดยใช้อัลกอริทึม Margin Infused Relaxed (MIRA) [16],[17] และใช้ระดับคำและอักขระมาเป็นคุณลักษณะ และเมื่อฆนาศัยตรวจสอบคลังข้อความฝึกฝนในการแบ่งคำบางคำที่ต้องพิจารณาจากบริบทเป็นสำคัญ เช่น “หลักการ” สามารถแบ่งได้เป็น “หลัก|การ|” หรือ “หลักการ|” แต่ในคลังข้อความฝึกฝนนั้นมีการแบ่งที่ขัดแย้งโดยพบว่าแบ่ง “หลัก|การ|” ได้ 160 ครั้ง และแบ่งเป็น “หลักการ|” ได้ 640 ครั้ง จึงให้มีการใช้วิธี Frequency Comparison และ Agreement Coefficient เพื่อตรวจจับและปรับแก้ไขข้อมูลฝึกฝนที่มีการแบ่งคำอย่างที่ไม่สอดคล้อง (Inconsistency Detection and Correction) ซึ่งจากผลการทดลองของฆนาศัยนั้นเห็นได้ว่าช่วยทำให้ผลการแบ่งคำมีความถูกต้องมากยิ่งขึ้น

รายการอ้างอิง

[1] The National Electronics and Computer Technology Center (NECTEC). แนวทางการแบ่งคำภาษาไทยสำหรับ BEST 2009. (online) Available from: http://thailang.nectec.or.th/2009/ [19 June 2009].

[2] The National Electronics and Computer Technology Center (NECTEC). BEST Corpus. Available: http://thailang.nectec.or.th/best/

[3] NEC Laboratories America. Stochastic Gradient Descent (SGD). Available from: http://leon.bottou.org/projects/sgd

[4] C. Sutton and A. McCallum. An Introduction to Conditional Random Fields for Relational Learining. Department of Computer Science, Universitiy of Massachusetts, USA.

[5] C. Kruengkrai, K. Jun'ichi KAZAMA, U. Kiyotaka, T. Kentaro and I. Hitoshi. A Discriminative Hybrid Model for Joint Chinese Word Segmentation and POS Tagging. The 11th Oriental COCOSDA Workshop (O-COCOSDA2008) (November 25-27, 2008).

[6] วิกิพีเดีย สารานุกรมเสรี. อักษรไทย. Available from: http://th.wikipedia.org/wiki/อักษรไทย

[7] Y. Poovarawan and W. Imarrom. การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี (Thai Syllable Separator by Dictionary). Proceedings of the 9th Annual Meeting on Electrical Engineering of the Thai Universities (1986).

[8] A. Kawtrakul, C. Thumkanon and S. Seriburi. A statistical approach to Thai word filtering. The second symposium on natural language processing (1997) : pp. 398–406.

[9] W. Aroonmanakun. Collocation and Thai Word Segmentation. Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop (2002) : pp. 68-75.

[10] C. Kruengkrai, V.Sornlertlamvanich and H. Isahara. A conditional random field framework for thai morphological analysis. Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC-2006) (2006).

[11] S. Chaisuriya, T. Iempairote and S. Jungjaruernrat. Proof for efficiency on Open Source Software for Thai syllable segmentation. Department of Computer Science, Faculty of Science, Burapha University.

[12] C. Kruengkrai, K. Jun'ichi, U. Kiyotaka, T. Kentaro, I. Hitoshi and C. Jaruskulchai. A Word and Character-Cluster Hybrid Model for Thai Word Segmentation. Proceedings 2009 Eighth International Symposium on Natural Language Processing. (October 20-21, 2009).

[13] K. Crammer, R. McDonald, and F.Pereira. Scalable large-margin online learning for structured classification. NIPS Workshop on Learning With Structured Outputs. (2005).

[14] R. McDonald, Discriminative Training and Spanning Tree Algorithms for Dependency Parsing. University of Pennsylvania, PhD Thesis. (2006).

[15] J. Inrut, P. Yuanghirun, S. Paludkong, S. Nitsuwat and P. Limmaneepraserth. Thai Word Segmentation using Combination of Forward and Backward Longest Matching Techniques. Proceedings of the 2001 International Symposium on Communications and Information Technology (2001) : pp. 37-40.

[16] S. Meknavin, P. Charoenpornsawat and B. Kijsirikul. Feature-based Thai Word Segmentation. Proceedings of the Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97) (2nd-4th December 1997) : pp. 41-46.

[17] T. Theeramunkong, W. Chinnan, T. Tanhermhong and V. Sornlertlamvanich. Full-Text Search for Thai Information Retrieval Systems. Proceedings of The Fourth Symposium on Natural Language Processing 2000 (2000) : pp. 317-326.

[18] W. Aroonmanakun. Thoughts on Word and Sentence Segmentation in Thai. Proceedings of the Seventh International Symposium on Natural Language Processing (13th-15th September 2007) : pp. 85-90.

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวเกศราภรณ์ ซื่อสัตย์พาณิชย์ เกิดเมื่อวันที่ 19 ธันวาคม พ.ศ. 2522 ที่กรุงเทพฯ สำเร็จการศึกษาระดับมัธยมศึกษาตอนต้นจากโรงเรียนวัดทรงธรรม จังหวัดสมุทรปราการ สำเร็จการศึกษาระดับประกาศนียบัตรวิชาชีพ (ปวช.) จากโรงเรียนเซนต์จอห์นเทคโนโลยี กรุงเทพฯ สำเร็จการศึกษาระดับประกาศนียบัตรวิชาชีพชั้นสูง (ปวส.) จากวิทยาลัยพณิชยการธนบุรี กรุงเทพฯ และสำเร็จการศึกษาระดับปริญญาบัณฑิต ในภาควิชาเทคโนโลยี-สารสนเทศ จากคณะภาควิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี ในปีการศึกษา 2545