การระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคม

นางสาวนัทธมน โมกข์ณรงค์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556

AN IDENTIFICATION OF ROMANIZED TOKENS FOUND IN SOCIAL MEDIA TO READ ALOUD IN THAI

Miss Nutthamon Moknarong

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering Program in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2013

หัวข้อวิทยานิพนธ์	การระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคม
โดย	นางสาวนัทธมน โมกข์ณรงค์
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร. อติวงศ์ สุชาโต
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร. โปรดปราน บุณยพุกกณะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร. บัณฑิต เอื้ออาภรณ์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. เศรษฐา ปานงาม)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รองศาสตราจารย์ ดร. อติวงศ์ สุชาโต)

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

.....กรรมการภายนอกมหาวิทยาลัย
(อาจารย์ ดร. รุ่งภัทร เริงพิทยา)

นัทธมน โมกข์ณรงค์ : การระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคม. (AN IDENTIFICATION OF ROMANIZED TOKENS FOUND IN SOCIAL MEDIA TO READ ALOUD IN THAI) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร. อติวงศ์ สุชาโต, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ผศ. ดร. โปรดปราน บุณยพุกกณะ, 48 หน้า.

ข้อความภายในเครือข่ายสังคมถูกสร้างโดยผู้ใช้งานหรือผู้เขียนจำนวนมาก นอกจากนั้นแต่ละคนยังมีรูปแบบการเขียนเฉพาะตัวที่ขึ้นอยู่กับความคิดสร้างสรรค์หรือทัศนคติของแต่ละบุคคล ในบางครั้งข้อความของคนไทยที่พบทั่วไปในเครือข่ายสังคมออนไลน์มีการเขียนคำไทยในลักษณะของการถอดอักษรแบบโรมัน ดังนั้นระบบแปลงภาษาเขียนเป็นภาษาพูดไม่สามารถทำการอ่านโทเค็นด้วยอักษรเหล่านั้นได้ถูกต้อง งานวิจัยนี้ได้นำเสนอและประเมินวิธีทางสถิติสำหรับการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคม โดยนำเสนอลักษณะสำคัญที่มีการพิจารณาโทเค็นที่ขึ้นกับบริบทรอบข้างและโทเค็นที่ปราศจากบริบท ข้อมูลจริงที่ได้จากเครือข่ายสังคมถูกนำมาใช้ในการสร้างชุดข้อมูลฝึกสอนและข้อมูลทดสอบ ผลการทดลองแสดงว่าผู้เข้าร่วมวิจัยระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคมโดยไม่พิจารณาบริบทมีความแม่นยำโดยมีค่าเฉลี่ย 91.60% ในขณะที่เมื่อพิจารณาคำที่มีบริบทมีค่าความแม่นยำ 99.41% จากลักษณะสำคัญที่นำเสนอทำให้การจำแนกด้วยต้นไม้ตัดสินใจและแบบจำลองเอ็นแกรมมีความแม่นยำในการจำแนกเท่ากับ 87.94% และ 79.30% ตามลำดับ กรณีถัดมาเมื่อพิจารณาการปรากฏของโทเค็นที่สนใจในพจนานุกรมอังกฤษ วิธีนี้มีค่าความแม่นยำเพิ่มขึ้นเป็น 82.28% โดยผลการรวมกันของ 2 วิธีในการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคมมีความแม่นยำเป็น 90.49%

ภาควิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต .....
สาขาวิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....
ปีการศึกษา	2556	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

5570491821 : MAJOR COMPUTER ENGINEERING
KEYWORDS: LANGUAGE IDENTIFICATION / STATISTICAL MODEL / SOCIAL MEDIA / FACEBOOK

NUTTHAMON MOKNARONG: AN IDENTIFICATION OF ROMANIZED TOKENS FOUND IN SOCIAL MEDIA TO READ ALOUD IN THAI. ADVISOR: ASSOC. PROF. ATIWONG SUCHATO, Ph.D.,ASST. PROF. PROADPRAN PUNYABUKKANA, Ph.D., 48 pp.

Social media contents were created by a large number of users or writers. Additionally, each person has one's own writing style, which depends on one's own creative thinking or attitudes. As commonly found in online social networks of Thai users, typed texts sometimes include Thai words that were transliterated with Roman letters. Therefore, text-to-speech systems cannot pronounce these transliterated tokens correctly. In this work, we propose and evaluate statistical methods for detecting Romanized Thai tokens. Both context-dependent and context-free classification features are proposed. Real social network texts are used for constructing the training set and the test set. The result reveals that human subjects can detect Thai Romanized tokens at 91.60% accuracy on average when adjacent contexts are hidden, while the accuracy is at 99.41% with contexts. With the proposed features, a decision tree-based classifier and an N-gram-based classifier yield 87.94% and 79.30% accuracy, respectively. In the latter case, the accuracy increases to 82.28% when the tokens' existence in English dictionaries is considered. Combining the two methods results in a detection accuracy of 90.49%.

Department:	Computer Engineering	Student's Signature .....
Field of Study:	Computer Engineering	Advisor's Signature .....
Academic Year:	2013	Co-Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์ของ รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก และ ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ซึ่งท่านทั้งสองได้ให้ความรู้ คำปรึกษา การสนับสนุน และกำลังใจตลอดระยะเวลาที่ศึกษาและดำเนินการวิจัย

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.เศรษฐา ปานงาม และอาจารย์ ดร.รุ่งภัทร เริงพิทยา กรรมการสอบวิทยานิพนธ์ในครั้งนี้ที่ได้ให้ความรู้ คำชี้แนะ และแนวทางต่างๆ เพื่อนำไปใช้ในการปรับปรุงวิทยานิพนธ์ฉบับนี้ให้ดียิ่งขึ้น

นอกจากนี้ผู้วิจัยขอขอบคุณเพื่อนร่วมงานทุกท่านในห้องปฏิบัติการระบบภาษาพูดและห้องปฏิบัติการเทคโนโลยีช่วยเหลือผู้พิการ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ให้คำแนะนำ มิตรภาพและกำลังใจในการดำเนินงานวิจัยนี้เสมอมา

สุดท้ายนี้ขอขอบคุณครอบครัวที่เข้าใจและสนับสนุนการทำวิจัยในครั้งนี้ตลอดระยะเวลาที่ได้ทำการศึกษา ซึ่งทำให้การจัดทำวิทยานิพนธ์ในครั้งนี้ประสบความสำเร็จได้ด้วยดี

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญตาราง ..... ฌ

สารบัญรูปภาพ ..... ฎ

บทที่ 1 บทนำ 1

1.1. ความเป็นมาและความสำคัญของปัญหา 1

1.2. วัตถุประสงค์ของการวิจัย 2

1.3. ขอบเขตของการวิจัย 2

1.4. ลำดับขั้นตอนในการเสนอผลการวิจัย 3

1.5. ประโยชน์ที่คาดว่าจะได้รับ 3

1.6. ผลงานตีพิมพ์จากวิทยานิพนธ์ 4

1.7. ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์ 4

บทที่ 2 ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง 5

2.1. ทฤษฎีที่เกี่ยวข้อง 5

2.2. วรรณกรรมที่เกี่ยวข้อง 9

บทที่ 3 ขั้นตอนการดำเนินงานวิจัย 12

3.1. ขั้นตอนการสร้างระบบการระบุคำให้อ่านแบบไทย 12

บทที่ 4 การทดลอง และวิธีการวัดผลการทดลอง 28

4.1. ข้อมูลที่ใช้ในการทดลอง 28

4.2. การวัดผลทดลอง 29

4.3. การทดลอง 29

บทที่ 5 ผลการทดลองและวิเคราะห์ผลการทดลอง 35

5.1. ประเมินผลการหาระบบอ้างอิง 35

5.2. ประเมินผลการระบุคำให้อ่านแบบไทยโดยพิจารณาจากการตัดสินของคน 36

5.3. ประเมินผลการการเปรียบเทียบลักษณะสำคัญด้วยต้นไม้ตัดสินใจ 37

หน้า

5.4. ประเมินประสิทธิภาพของแนวทางที่นำเสนอ 38
5.5. วิเคราะห์ข้อผิดพลาดของการระบุคำให้อ่านแบบไทยจากแนวทางที่นำเสนอ 40
บทที่ 6 สรุปผลการวิจัย และข้อเสนอแนะ 44
6.1. สรุปผลการวิจัย 44
6.2. อภิปรายผลการวิจัย 44
6.3. ข้อเสนอแนะ 45
รายการอ้างอิง 46
ประวัติผู้เขียนวิทยานิพนธ์ 48

สารบัญตาราง

หน้า

ตารางที่ 3-1 ตารางแสดงประเภทของการแท็กโทเค็นแบบต่างๆ 14
ตารางที่ 3-2 ตัวอย่างขั้นตอนการเตรียมข้อมูล 15
ตารางที่ 3-3 จำนวนโทเค็นในพจนานุกรม 15
ตารางที่ 3-4 ความถี่ของลักษณะสำคัญ Capital, Allcapital, Thaiprev และ Thainext เทียบกับแท็ก POS และ NEG 16
ตารางที่ 3-5 สัญลักษณ์ 10 อันดับแรกจากข้อมูลฝึกสอน 17
ตารางที่ 3-6 ความถี่และเปอร์เซนต์ความแตกต่างของสัญลักษณ์ที่ปรากฏก่อนและหลังโทเค็นที่สนใจ 17
ตารางที่ 3-7 ผลการเปรียบเทียบเปอร์เซนต์ความแตกต่างของสัญลักษณ์ 18
ตารางที่ 3-8 ผลการเปรียบเทียบจำนวนโทเค็นของลักษณะสำคัญ 19
ตารางที่ 3-9 ลักษณะสำคัญจากสัญลักษณ์ 19
ตารางที่ 3-10 ลักษณะสำคัญ 20
ตารางที่ 3-11 ตัวอย่างข้อมูลที่ทำการสกัดลักษณะสำคัญ 25
ตารางที่ 4-1 จำนวนข้อความในงานวิจัย 28
ตารางที่ 4-2 จำนวนของโทเค็นในงานวิจัย 29
ตารางที่ 4-3 จำนวนของโทเค็นแยกตามประเภทการแท็ก 29
ตารางที่ 4-4 ตัวอย่างการทดลองที่พิจารณาเฉพาะโทเค็นที่สนใจ 31
ตารางที่ 4-5 ตัวอย่างการทดลองที่พิจารณาโทเค็นที่สนใจโดยดูจากบริบทรอบข้างร่วมด้วย 31
ตารางที่ 5-1 ค่าเปอร์เซนต์ความแม่นยำของระบบอ้างอิง 36
ตารางที่ 5-2 ผลการจำแนกภาษาจากการตัดสินใจของคนโดยพิจารณาเฉพาะโทเค็น 36
ตารางที่ 5-3 ผลการจำแนกภาษาจากการตัดสินใจของคนโดยพิจารณาบริบทของโทเค็นร่วมด้วย 37
ตารางที่ 5-4 เปอร์เซนต์ความแม่นยำในการจำแนกด้วยต้นไม้ตัดสินใจ 38
ตารางที่ 5-5 ผลการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคม 39
ตารางที่ 5-6 ผลการระบุคำให้อ่านแบบไทยด้วยชุดข้อมูลฝึกสอน 40
ตารางที่ 5-7 จำนวนโทเค็นและเปอร์เซนต์ของความผิดพลาดจากการระบุคำให้อ่านแบบไทย 40
ตารางที่ 5-8 จำนวนโทเค็นและเปอร์เซนต์ของความผิดพลาดแยกตามลักษณะสำคัญย่อย 41

หน้า

ตารางที่ 5-9 จำนวนโทเค็นและเปอร์เซนต์ของโทเค็นที่ให้อ่านแบบไทยแต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบอื่นๆแยกตามกลุ่มลักษณะสำคัญที่ทำนายผิด..... 42
ตารางที่ 5-10 จำนวนโทเค็นและเปอร์เซนต์ของโทเค็นที่ให้อ่านแบบอื่นๆแต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบไทยแยกตามกลุ่มลักษณะสำคัญที่ทำนายผิด..... 43

สารบัญรูปภาพ

หน้า

ภาพที่ 3-1 แผนภาพแสดงขั้นตอนการสร้างระบบระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมัน .. 12
ภาพที่ 3-2 ตัวอย่างข้อมูลสำหรับ CMU-Cambridge Statistical Language Modeling Toolkit. 23
ภาพที่ 3-3 ตัวอย่างผลการคำนวณความน่าจะเป็นด้วย CMU-Cambridge Statistical Language Modeling Toolkit 23
ภาพที่ 3-4 โครงสร้างของกระบวนการการกำหนดค่าลักษณะสำคัญเอ็นแกรม 24
ภาพที่ 3-5 ตัวอย่างรูปแบบไฟล์ .arff 26
ภาพที่ 3-6 ตัวอย่างต้นไม้ของแบบจำลองการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมัน 26
ภาพที่ 4-1 โครงสร้างของกระบวนการการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันด้วยพจนานุกรม 30
ภาพที่ 4-2 โครงสร้างของการทดลองที่ 4 33

บทที่ 1
บทนำ

1.1. ความเป็นมาและความสำคัญของปัญหา

เทคโนโลยีแปลงภาษาเขียนเป็นภาษาพูด (text-to-speech) ถูกนำมาใช้กับผู้บกพร่องทางสายตาเพื่อเพิ่มประสิทธิภาพในการเรียนรู้ ซึ่งเทคโนโลยีนี้จะสามารถช่วยให้ผู้บกพร่องทางสายตาได้รับโอกาสทางการศึกษาที่มากขึ้น เนื่องจากการอ่านหนังสือเป็นช่องทางหนึ่งที่ใช้ในการศึกษาหาความรู้ ซึ่งเป็นช่องทางที่ง่ายและสะดวกที่สุด บุคคลผู้ที่มีความรู้ย่อมมีโอกาสทางสังคมและเป็นกำลังสำคัญในการพัฒนาสิ่งต่างๆมากกว่าบุคคลอื่น แต่ในสังคมยังมีบุคคลที่ขาดโอกาสทางด้านการศึกษาเป็นจำนวนมาก จึงทำให้บุคคลเหล่านี้มีข้อจำกัดทางด้านการเรียนรู้

การแปลงภาษาเขียนให้เป็นภาษาพูดจึงมีความสำคัญต่อบุคคลที่มีความบกพร่องทางสายตา ทั้งยังเป็นนวัตกรรมหนึ่งที่เข้ามามีบทบาทสำคัญต่อวงการธุรกิจในสังคมไทยเป็นเวลานาน เนื่องจากการแปลงภาษาเขียนให้เป็นภาษาพูดของมนุษย์สามารถนำไปประยุกต์ให้เกิดเป็นเทคโนโลยีที่ก่อให้เกิดประโยชน์ต่างๆได้ ดังเช่น นำมาประยุกต์ใช้เพื่อแปลงเนื้อหาของข่าวเป็นเสียงพูด นำไปใช้สำหรับการเรียนการสอนในผู้ป่วยที่มีปัญหาเรื่องการออกเสียงอันเนื่องมาจากการใส่ท่อช่วยหายใจ หรือนำไปใช้เพื่อแปลงข้อความในมือถือเป็นเสียงพูด [1] เป็นต้น

นอกจากประโยชน์ดังที่กล่าวมาข้างต้นแล้วเรายังสามารถนำเอาเทคโนโลยีนี้ไปประยุกต์ใช้งานกับการอ่านข้อความบนเครือข่ายสังคม โดยลักษณะการเขียนเพื่อสื่อสารของคนไทยแต่ละคนนั้นอาจแตกต่างกันออกไปตามทัศนะส่วนบุคคล โดยการเขียนมีทั้งการเขียนแบบภาษาเดียว หรือการเขียนรวมกันมากกว่าหนึ่งภาษา นอกจากนี้ยังมีการเขียนคำภาษาไทยด้วยอักษรโรมัน เนื่องจากคำไทยนั้นพิมพ์ยากและช้ากว่า [2] ปัญหานี้จะพบเห็นได้บ่อยในเครือข่ายสังคม นักวิจัยได้มีการพัฒนางานทางด้านนี้มาเป็นเวลานานหลายทศวรรษ แต่เนื่องจากรูปแบบการเขียนที่แตกต่างกันออกไปในแต่ละคนทำให้การนำข้อความเหล่านั้นมาประยุกต์เข้ากับเทคโนโลยีแปลงภาษาเขียนเป็นภาษาพูด เช่น Vaja 6.0 เพื่อเพิ่มความสะดวกให้กับผู้ใช้งานกลายเป็นงานหนึ่งที่มีความท้าทาย เนื่องจากในบางครั้งระบบไม่สามารถทำการอ่านข้อความที่เขียนด้วยอักษรโรมันที่อยู่นอกเหนือพจนานุกรมได้หรืออ่านผิดจากที่ควรเป็น

ทั้งนี้พื้นฐานของกระบวนการสังเคราะห์เสียงภาษาไทยเริ่มจากการรับข้อความเข้ามาแล้วประมวลผล ข้อความจะถูกแปลงเป็นหน่วยเสียงคำอ่านด้วยระบบการแปลงรูปอักขระเป็นหน่วยเสียงคำอ่าน (grapheme-to-phoneme) แล้วนำคำอ่านที่เป็นผลลัพธ์ที่ได้จากการแปลงนั้นมาสังเคราะห์เสียงตามหน่วยเสียงของคำอ่านเหล่านั้น ซึ่งระบบการแปลงรูปอักขระเป็นหน่วยเสียงคำอ่านภาษาไทยจะรองรับข้อความที่เป็นคำภาษาไทย ดังนั้นหากข้อความที่รับเข้ามาเป็นข้อความภาษาอังกฤษ ระบบการแปลงรูปอักขระเป็นหน่วยเสียงคำอ่านภาษาไทยจะไม่สามารถทำการแปลงคำภาษาอังกฤษเป็นหน่วยเสียงคำอ่านได้ ทำให้ไม่สามารถสังเคราะห์เสียงได้ถูกต้อง วิธีการแก้แบบหนึ่ง คือ การนำกระบวนการสังเคราะห์เสียงภาษาอังกฤษเข้ามาใช้ขนานกันกับการสังเคราะห์เสียงภาษาไทย [1] นั่น

คือ เมื่อข้อความที่รับเข้ามาเป็นภาษาไทย ข้อความจะถูกแปลงเป็นคำอ่านด้วยระบบสังเคราะห์เสียงภาษาไทย ขณะที่ข้อความที่รับเข้ามาเป็นภาษาอังกฤษจะถูกแปลงเป็นคำอ่านด้วยระบบสังเคราะห์เสียงภาษาอังกฤษ ซึ่งปัญหาที่เกิดขึ้นคือ ข้อความที่รับเข้ามาเป็นภาษาอังกฤษ แต่ไม่ใช่คำให้อ่านแบบภาษาอังกฤษ เช่น naka (นะคะ), Chula (จุฬา), Dao (ดาว) เป็นต้น ซึ่งคำเหล่านี้เป็นคำไทยในรูปอักษรโรมัน ทำให้ระบบการแปลงรูปอักขระเป็นหน่วยเสียงคำอ่านจะไม่สามารถแปลงเป็นคำอ่านได้ถูกต้องและยังส่งผลให้เสียงที่ได้จากระบบสังเคราะห์เสียงผิดจากความจริง ตัวอย่างประโยคที่พบว่ามีปัญหาหลังจากใช้เทคโนโลยีแปลงภาษาเขียนเป็นภาษาพูด Vaja 6.0 เช่น I have breakfast with dao na โดยคำว่า “dao” และ “na” Vaja 6.0 ไม่สามารถอ่านได้ถูกต้อง เพื่อแก้ไขปัญหาดังกล่าวงานด้านการระบุภาษาจึงเป็นอีกหนึ่งแนวทางที่นำมาใช้เพื่อแยกความต่างของภาษาก่อนที่จะนำข้อความไปแปลงเป็นหน่วยเสียงคำอ่านให้ตรงตามกับคำอ่านของข้อความนั้นจริง เพื่อให้เลือกระบบสังเคราะห์เสียงได้ถูกต้อง ดังนั้นถ้าสามารถทำการระบุภาษาของคำได้ จะช่วยส่งผลให้ระบบแปลงภาษาเขียนให้เป็นภาษาพูดมีประสิทธิภาพมากขึ้น

ด้วยเหตุนี้ผู้วิจัยจึงมีแนวคิดที่จะนำเสนอแนวทางในการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันโดยคำนึงถึงแบบจำลองทางสถิติ เพื่อนำผลลัพท์ที่ได้ไปประยุกต์ใช้ในการเตรียมข้อความให้เป็นคำอ่านให้ถูกต้องก่อนทำการสังเคราะห์เสียงของเทคโนโลยีการแปลงภาษาเขียนเป็นภาษาพูด โดยงานวิจัยนี้ทำการพิจารณาการรวมกันของต้นไม้ตัดสินใจและแบบจำลองเอ็นแกรมในการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคม โดยระบุว่าคำในประโยคที่ทำการป้อนเข้ามานั้นเป็นคำที่ให้อ่านในลักษณะของภาษาไทยหรือภาษาอื่นๆ

1.2. วัตถุประสงค์ของการวิจัย

เพื่อทำการเสนอวิธีในการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคม โดยคำนึงถึงวิธีทางสถิติ

1.3. ขอบเขตของการวิจัย

1. ข้อมูลสำหรับการพิจารณานำมาจากเว็บไซต์เฟซบุ๊ก โดยที่ในประโยคจะต้องมีอักษรอังกฤษประกอบอยู่ในประโยค ข้อมูลที่พิจารณาจะอยู่ในรูปของคำในอักษรโรมันเท่านั้น โดยอักษรพิเศษที่นอกเหนือจากอักษรภาษาอังกฤษรวมไปถึงชื่อเว็บไซต์และอีเมลจะไม่ถูกระบุภาษา เช่น Chloé, @, www.google.com, abc@gmail.com เป็นต้น
2. ข้อมูลฝึกสอนทั้งหมดจำนวน 9,360 ประโยค ข้อมูลทดสอบจำนวน 1,167 ประโยค
3. ลักษณะของคำให้อ่านแบบไทยในงานวิจัยนี้ คือ คำไทยที่เขียนด้วยอักษรโรมันโดยที่คำนั้นไม่ใช่คำทับศัพท์ต่างประเทศ และไม่ใช่ศัพท์ต่างประเทศที่คนไทยนำมาใช้เป็นชื่อเฉพาะ
4. ประสิทธิภาพในการประมวลผลจะไม่รวมถึงประสิทธิภาพเชิงเวลา

5. ผลลัพธ์ของงานวิจัยนี้ คือ ระบุว่าคำใดในประโยคให้อ่านแบบไทย
    ก. ตัวอย่างของคำให้อ่านแบบไทย ka, naja, krub, petcharut, nichakorn, krab
    ข. ตัวอย่างของคำที่ไม่ให้อ่านแบบไทย sushi, dinner, central, bank, iphone, cookie

1.4. ลำดับขั้นตอนในการเสนอผลการวิจัย

1. ขั้นตอนการศึกษาเบื้องต้น
    ก. ศึกษาวรรณกรรมที่เกี่ยวข้องกับการแปลงอักษรเขียนเป็นเสียงพูด
    ข. ศึกษาวรรณกรรมที่เกี่ยวข้องกับการการระบุภาษา
    ค. ศึกษาเทคนิคการเรียนรู้ด้วยเครื่องจักร (Machine Learning) ที่ใช้สำหรับการจำแนกประเภท
    ง. ศึกษาแบบจำลองเอ็นแกรม
    จ. ศึกษาเครื่องมือที่ใช้ในงานวิจัย เช่น เครื่องมือสังเคราะห์เสียงพูดภาษาไทย (text to-speech system), เครื่องมือในการจำแนกประเภท (Classification Tool) และเครื่องมือคำนวณค่าเอ็นแกรม (N-gram Tool)
2. ขั้นตอนการออกแบบระบบและทำการทดลอง
    ก. ออกแบบการทำงานของระบบ
    ข. เก็บข้อความจากเว็บไซต์เฟซบุ๊กและจัดประเภทของคำ
    ค. วิเคราะห์ข้อมูลเพื่อหาลักษณะเด่นที่เหมาะสม
    ง. ออกแบบการทดลอง
    จ. สร้างแบบจำลองในการจำแนกประเภท
    ฉ. ทดสอบการทำงานและบันทึกผลการทดลอง
    ช. วัดผลความแม่นยำและวิเคราะห์ผลการทดลองของระบบในการระบุคำให้อ่านแบบไทย
3. สรุปผลและเรียบเรียงวิทยานิพนธ์

1.5. ประโยชน์ที่คาดว่าจะได้รับ

1. ได้แนวทางในการระบุคำให้อ่านแบบไทยจากข้อความในเครือข่ายสังคม
2. สามารถนำแนวทางนี้ไปประยุกต์ใช้กับเทคโนโลยีการแปลงภาษาเขียนเป็นภาษาพูด เพื่อให้สามารถเลือกโมดูลในการแปลงภาษาได้อย่างเหมาะสม

1.6. ผลงานตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้ตีพิมพ์เป็นบทความทางวิชาการในหัวข้อเรื่อง *"Detecting Romanized Thai Tokens in Social Media Texts"* จัดทำโดย "Nutthamon Moknarong, Atiwong Suchato and Proadpran Punyabukkana" โดยนำเสนอในงานประชุมวิชาการ "The 2013 International Computer Science and Engineering Conference: ICSEC'2013" จัดโดย มหาวิทยาลัยศิลปากร ประเทศไทย ในวันที่ 4-6 กันยายน 2556

1.7. ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์

ในวิทยานิพนธ์นี้ได้แบ่งเนื้อหาออกเป็น 6 บทดังต่อไปนี้ บทที่ 1 เป็นบทนำซึ่งกล่าวถึง ความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ของการวิจัย ขอบเขตของการวิจัย ลำดับขั้นตอนในการเสนอผลการวิจัย ประโยชน์ที่คาดว่าจะได้รับ และผลงานตีพิมพ์จากวิทยานิพนธ์ บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง กล่าวถึง แนวคิดและทฤษฎี ประกอบด้วย ทฤษฎีที่เกี่ยวกับการประมวลผลภาษาธรรมชาติ ซึ่งได้แก่ การระบุภาษา ทฤษฎีที่เกี่ยวกับแนวทางในการวิจัย ได้แก่ การเรียนรู้ของเครื่องจักร และแบบจำลองเอ็นแกรม ส่วนสุดท้าย คือ วรรณกรรมที่เกี่ยวข้องกับการระบุภาษา ในบทที่ 3 แสดงขั้นตอนการดำเนินงานวิจัย บทที่ 4 กล่าวถึงข้อมูลที่ใช้ในการทดลอง วิธีการวัดผลการทดลองและขั้นตอนการทดลอง บทที่ 5 อธิบายผลการทดลองและการวิเคราะห์ผลการทดลอง บทที่ 6 บทสรุปผลการวิจัยและข้อเสนอแนะ

บทที่ 2
ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและวรรณกรรมที่เกี่ยวข้อง ซึ่งแบ่งออกเป็น 2 ส่วน โดยส่วนแรกจะกล่าวถึง ทฤษฎีที่เกี่ยวข้อง ซึ่งได้แก่ ทฤษฎีที่เกี่ยวกับการประมวลผลภาษาธรรมชาติ ได้แก่ การระบุภาษา ทฤษฎีที่เกี่ยวกับการเรียนรู้ของเครื่องจักรที่เกี่ยวข้องกับวรรณกรรมนี้ ได้แก่ การเรียนรู้แบบต้นไม้ตัดสินใจ และทฤษฎีของแบบจำลองเอ็นแกรม ในส่วนที่ 2 จะกล่าวถึงวรรณกรรมที่เกี่ยวข้องกับวิทยานิพนธ์นี้ ซึ่งได้แก่ วรรณกรรมที่เกี่ยวข้องกับการระบุภาษา (Language Identification)

2.1.ทฤษฎีที่เกี่ยวข้อง

2.1.1. ทฤษฎีเกี่ยวกับการประมวลผลภาษาธรรมชาติ

2.1.1.1. การระบุภาษา

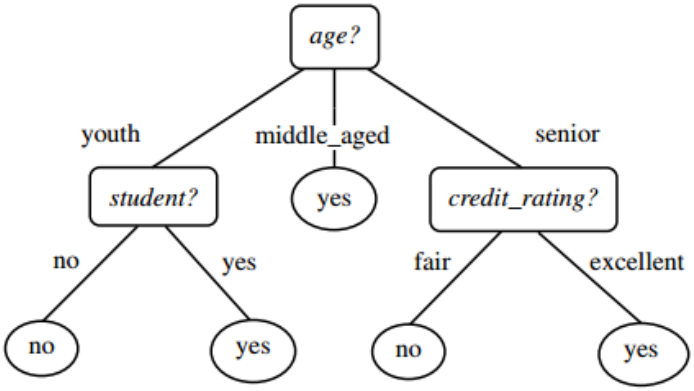
การระบุภาษา คือ กระบวนการในการจำแนกภาษาของเอกสารหรือข้อความที่พิจารณา โดยเป็นกระบวนการหนึ่งที่สำคัญสำหรับระบบประมวลผลภาษาธรรมชาติ (Natural Language Processing) [3] และระบบการแปลงรูปอักขระเป็นหน่วยเสียงคำอ่านให้ถูกต้อง [4] โดยภาษาที่ทำการระบุนั้นส่วนมากจะเป็นภาษาที่มีชุดอักขระเหมือนกัน เนื่องจากอักขระที่เหมือนกันจะสร้างความกำกวมต่อผู้ใช้งาน เช่น ชุดของตัวอักษรโรมัน เป็นต้น กระบวนการของการระบุภาษาสามารถแบ่งได้เป็น 2 ขั้นตอน คือ การฝึกสอน และ การระบุภาษา [5] แบบจำลองสำหรับการระบุภาษาจะถูกฝึกสำหรับภาษาที่เป็นภาษาเป้าหมาย โดยขั้นตอนการฝึกของแต่ละแบบจำลองจะแตกต่างกันไปตามขั้นตอนวิธีที่นำมาใช้ฝึกสอน งานที่นิยมนำการระบุภาษาไปประยุกต์ใช้เช่น การทับศัพท์, ระบบค้นคืนสารสนเทศ หรือการแปลงภาษาด้วยเครื่อง เป็นต้น

2.1.2. การเรียนรู้ของเครื่องจักรที่เกี่ยวข้องกับวรรณกรรม

2.1.2.1. การเรียนรู้แบบต้นไม้ตัดสินใจ

การสร้างต้นไม้ตัดสินใจนั้นเป็นการเรียนรู้ของต้นไม้ตัดสินใจ (Decision Tree Learning) จากลักษณะสำคัญ (feature) ของข้อมูลสอนที่มีการติดสลาก (label) ไว้กับข้อมูลสอน [6] โดยโครงสร้างของต้นไม้ตัดสินใจประกอบด้วยโหนด (node) และเส้นเชื่อม (edge) ซึ่งโหนดจะถูกแทนด้วยลักษณะสำคัญของข้อมูลแบ่งออกเป็นโหนดราก (root node) ซึ่งเป็นส่วนบนสุดของต้นไม้และเป็นลักษณะสำคัญของข้อมูลที่ถูกพิจารณาเป็นประจำในการจำแนกกลุ่มของข้อมูล และโหนดใบ (leaf node) เป็นโหนดที่อยู่ล่างสุดของต้นไม้หรือโหนดที่ไม่มีลูกแทนผลลัพธ์ที่เป็นคำตอบ ส่วนเส้นเชื่อมจะถูกแทนค่าด้วยค่าที่เป็นไปได้ของลักษณะสำคัญ [7]

ตัวอย่างของต้นไม้ตัดสินใจของกลุ่มลูกค้าในการซื้อคอมพิวเตอร์แสดงดังภาพ 2-1 เช่น ในกรณีที่ข้อมูลนำเข้าเป็นเด็กวัยรุ่นที่ยังเป็นนักเรียนอยู่ การทำนายผลเริ่มจากการพิจารณากลุ่มของอายุลูกค้าซึ่งมีค่าเป็นวัยรุ่น ถัดมาจึงต้องพิจารณาว่าเป็นนักเรียนหรือเปล่า ซึ่งยังเป็นนักเรียนอยู่จึงได้ ผลลัพธ์การทำนายว่าเป็นกลุ่มลูกค้าที่ซื้อคอมพิวเตอร์



ภาพที่ 2-1 ต้นไม้ตัดสินใจสำหรับการทำนายว่าเป็นกลุ่มของลูกค้าที่ซื้อคอมพิวเตอร์หรือไม่ [6]

แนวทางหลักในการสร้างต้นไม้ตัดสินใจ คือ การเลือกลักษณะสำคัญจากลักษณะสำคัญทั้งหมดของกลุ่มข้อมูล โดยเมื่อทำการแบ่งกลุ่มของข้อมูลตามลักษณะสำคัญต่างๆที่ได้เลือกไว้แล้ว ทำให้ข้อมูลยังจับกลุ่มโดยไม่แยกออกจากกัน [8]

การเลือกลักษณะสำคัญสำหรับใช้ในการตัดสินใจของแต่ละโหนดมีหลายแนวทางด้วยกัน แต่แนวทางที่นิยมใช้ คือ ค่าเกน (Information Gain) โดยค่าเกนนั้นแสดงถึงความสามารถในการแยกข้อมูลออกจากกันตามลักษณะสำคัญ [7] ค่าเกนที่ดี คือ ค่าเกนที่มีผลลัพธ์สูง แต่ค่าเกนที่มีผลลัพธ์ต่ำจะไม่สามารถแยกคำตอบของข้อมูลออกมาได้ ค่าเกนของลักษณะสำคัญ X คำนวณจากค่าเอนโทรปี (Entropy) ทั้งหมดของชุดข้อมูลลบด้วยค่าเอนโทรปีหลังจากเลือกลักษณะสำคัญ X เพื่อแบ่งข้อมูลออกเป็นกลุ่ม ค่าเอนโทรปีหลังจากเลือกลักษณะสำคัญ X คือ ผลรวมของผลคูณระหว่างค่าเอนโทรปีของแต่ละโหนดกับอัตราส่วนของตัวอย่างในแต่ละกิ่งต่อตัวอย่างทั้งหมดในโหนดนั้นๆ ตัวอย่างเช่น กำหนดให้ชุดข้อมูลสอน คือ T ลักษณะสำคัญที่โหนด คือ X ที่มีค่าที่สามารถเป็นไปได้ทั้งหมด n ค่า โหนดปัจจุบันจะแบ่งตัวอย่างของ T ออกเป็น $\{t_1,t_2,t_3,\ldots,t_n\}$ ค่าเอนโทรปีหลังจากแบ่งชุดข้อมูลเป็นกลุ่มตามลักษณะสำคัญ X เป็นดังสมการ 2.1 และค่าเกนของลักษณะสำคัญ X คำนวณได้จากสมการ 2.2

$$I_x(T) = \sum_{i=1}^{n} \left|\frac{t_i}{T}\right| I(t_i) \tag{2.1}$$

$$Gain(X) = I(T) - I_x(T) \tag{2.2}$$

ค่าเอนโทรปีของชุดข้อมูล T ที่ประกอบด้วยค่าผลลัพธ์ที่สามารถเป็นไปได้ $\{m_1,m_2,m_3,\ldots,m_n\}$ คำนวณได้ด้วยสมการ 2.3

$$I(m) = \sum_{i=1}^{n} -P(m_i)\log_2 P(m_2) \tag{2.3}$$

ตัวอย่างของขั้นตอนวิธีที่ใช้สำหรับการแบ่งกลุ่มเพื่อสร้างต้นไม้ตัดสินใจ เช่น ID3 (Iterative Dichotomiser), C4.5 และ CART (Classification and Regression Trees) เป็นต้น โดยทั้ง 3 ขั้นตอนวิธีนี้เป็นการทำงานแบบละโมภ (Greedy) คือ ทำงานจากบนลงล่าง [6]

ข้อดีของแบบจำลองต้นไม้ตัดสินใจ คือ ผลลัพธ์ที่ได้จากแบบจำลองง่ายต่อความเข้าใจและการตีความ รองรับข้อมูลได้หลากหลายประเภททั้งข้อมูลแบบต่อเนื่องและไม่ต่อเนื่อง ส่วนข้อเสียของแบบจำลองต้นไม้ตัดสินใจนี้ คือ ไม่เหมาะกับระบบที่มีจำนวนลักษณะสำคัญจำนวนมาก [7]

2.1.3. ทฤษฎีแบบจำลองเอ็นแกรม

2.1.3.1. แบบจำลองเอ็นแกรม

แบบจำลองเอ็นแกรม (N-gram) คือ แบบจำลองทางสถิติที่เกิดจากการคำนวณค่าความน่าจะเป็นของการเกิดประโยคที่เกิดจากการเรียงกันของคำ หรือเป็นการประมาณค่าการปรากฏของตัวอักษรตัวหนึ่งที่ขึ้นกับตัวอักษรก่อนหน้าเพียง n-1 ตัว โดยค่าความน่าจะเป็นของคำหรือประโยคประมาณได้จากคลังข้อมูลที่สร้างไว้ [9]

สำหรับแกรมที่ใช้ในแบบจำลองเป็นไปได้ทั้งเสียง อักขระ คำ หรือประโยค ซึ่งแกรมสามารถมีได้หลายขนาด ตั้งแต่ 1 ถึง n ตัว การคำนวณความน่าจะเป็นของการเกิดชุดอักขระที่รวมกันเป็นคำแสดงดังสมการ 2.4

$$P(c_1 \ldots c_n) = P(c_1)\prod_{i=2}^{n} P(c_i|c_{i-1}) \tag{2.4}$$

โดย $\mathrm{P}(c_n|c_{n-1}) = \frac{C(c_{n-1},c_n)}{C(c_{n-1})}$

ตัวอย่างการประมาณค่าความน่าจะเป็นของคำว่า “hello”

กรณีที่ใช้ไบแกรม P(h)P(e|h)P(l|e)P(l|l)P(o|l)

$$\frac{C(h)}{N} * \frac{C(he)}{C(h)} * \frac{C(el)}{C(e)} * \frac{C(ll)}{C(l)} * \frac{C(lo)}{C(l)}$$

กรณีที่ใช้ไตรแกรม P(h)P(e|h)P(l|he)P(l|el)P(o|ll)

$$\frac{C(h)}{N} * \frac{C(he)}{C(h)} * \frac{C(hel)}{C(he)} * \frac{C(ell)}{C(el)} * \frac{C(llo)}{C(ll)}$$

กำหนดให้

P คือ ค่าความน่าจะเป็น

c คือ อักขระ

($c_1c_2c_3...c_n$) คือ ชุดอักขระที่ประกอบด้วยอักขระตั้งแต่ 3 ตัวขึ้นไปจนถึง n ตัว

C (c_{n-1}, c_n) คือ จำนวนครั้งในชุดข้อมูลสอนที่เกิด c_{n-1} คู่กับ c_n

C (c_n) คือ จำนวนครั้งที่เกิด c_n ในชุดข้อมูลสอน

เนื่องจากความง่ายในการดำเนินงานและมีความแม่นยำในการทำนายลำดับที่เป็นไปได้สูง จึงทำให้แบบจำลองเอ็นแกรมกลายเป็นอีกวิธีหนึ่งทางสถิติที่ได้รับความนิยมในงานประมวลผลภาษาธรรมชาติ โดยแนวคิดหลักของการใช้แบบจำลองเอ็นแกรมสำหรับการระบุภาษา คือ ทุกภาษามีเอกลักษณ์เฉพาะของตัวเอง [3]

อย่างไรก็ตามถึงแม้ว่าวิธีแบบจำลองเอ็นแกรมจะให้ผลลัพธ์ที่ดี แต่อาจเกิดกรณีที่ทำการคำนวณความน่าจะเป็นแล้วได้ผลลัพธ์เป็นศูนย์ เนื่องมาจากไม่มีชุดข้อมูลนั้นในคลังข้อมูลจำนวนชุดข้อมูลที่พบจึงเป็นศูนย์ และถ้านำค่านั้นไปคำนวณจะทำให้สมการนั้นมีผลลัพธ์เป็นศูนย์ไปด้วย ซึ่งเป็นปัญหาสำคัญสำหรับแบบจำลองเอ็นแกรม สำหรับการแก้ปัญหานี้สามารถทำได้โดยการปรับเรียบ (Smoothing) [10] ซึ่งมีหลายวิธีดังตัวอย่าง

1. การทำให้ราบเรียบด้วยวิธีการเพิ่ม 1 (Add-one Smoothing) เป็นการทำให้ราบเรียบ โดยทำการเพิ่มหน่วยนับ 1 ให้กับทุกหน่วย ก่อนทำการปรับค่าบรรทัดฐานให้เป็นความน่าจะเป็น วิธีนี้เป็นวิธีที่ง่ายแต่ให้ผลไม่ดี
2. การทำให้ราบเรียบย้อน (Back-off Smoothing) การทำให้ราบเรียบแบบย้อนเป็นอีกหนึ่งวิธีที่ถูกนำมาใช้ในการปรับค่า ตัวอย่างของกรณีที่ไม่มีชุดข้อมูลของบางไตรแกรม $w_{n-2}w_{n-1}w_n$ ในการคำนวณค่า $P(w_n|w_{n-1}w_{n-2})$ จะทำการประมาณด้วยค่าความน่าจะเป็นของไบแกรม $P(w_n|w_{n-1})$ และเมื่อยังหาค่าไม่ได้ก็จะทำการประมาณค่าด้วยยูนิแกรม $P(w_n)$ ดังนั้นการประมาณค่าความน่าเป็นของแบบจำลองไตรแกรมเป็นได้ดังสมการ 2.5

$$P(w_i|w_{i-2}w_{i-1}) = \begin{cases} P(w_i|w_{i-2}w_{i-1}), if C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha_1 P(w_i|w_{i-1}) \quad , if C(w_{i-2}w_{i-1}w_i) = 0 \; and \; C(w_{i-1}w_i) > 0 \\ \alpha_2 P(w) \qquad\qquad , other \end{cases} \quad (2.5)$$

α_1 และ α_2 คือค่าน้ำหนักย้อน ซึ่งขึ้นกับอัลกอริทึมที่เลือกใช้ในการทำให้ราบเรียบย้อน

2.2.วรรณกรรมที่เกี่ยวข้อง

เนื่องจากรูปแบบการเขียนของคนไทยสำหรับการสื่อสารมีทั้งการเขียนเพียงภาษาเดียว หรือการเขียนรวมกันมากกว่าหนึ่งภาษา นอกจากนี้ยังมีการเขียนคำภาษาไทยด้วยอักษรโรมันอยู่ด้วย ดังจะเห็นได้จากสื่อเครือข่ายสังคมออนไลน์ ถึงแม้ว่าคำเหล่านั้นจะเขียนด้วยอักษรโรมันเหมือนกันแต่ภาษาของทั้งคู่นั้นมีความต่างกันอยู่ ดังนั้นผู้วิจัยจึงได้ทำการศึกษาวรรณกรรมที่เกี่ยวกับการระบุภาษาดังแสดงในหัวข้อถัดไป

2.2.1. วรรณกรรมที่พิจารณาการระบุภาษา

วรรณกรรมทางด้านการระบุภาษาสามารถแบ่งได้เป็น 2 แบบคือ พิจารณาเพื่อระบุภาษาโดยพิจารณาภาพรวมของเอกสารหรือเว็บ หรือพิจารณาเฉพาะคำเพื่อทำการระบุภาษา แต่เนื่องจากงานวิจัยนี้มีความเกี่ยวข้องกับการพิจารณาภาษาของคำ ดังนั้นผู้วิจัยจึงได้ทำการศึกษาเทคนิคต่างๆที่นำมาใช้เพื่อระบุภาษาของคำซึ่งเกี่ยวข้องโดยตรงกับงานวิจัย

Juha Hakkinen และ Jilei Tian [5] นำเสนอแนวทางในการระบุภาษาของชื่อเฉพาะที่แตกต่างกัน 2 แนวทาง คือ การปรับปรุงแบบจำลองเอ็นแกรม และการใช้ต้นไม้ตัดสินใจ โดยได้เลือกใช้แบบจำลองเอ็นแกรมระดับอักขระ สำหรับการปรับปรุงแบบจำลองเอ็นแกรมทำโดยแยกคำออกเป็น 3 ส่วน สำหรับการฝึกสอน สำหรับการสร้างต้นไม้ตัดสินใจทำการฝึกสอนการระบุภาษาโดยพิจารณาอักขระรอบข้างของอักขระที่กำลังพิจารณา การทดสอบทำโดยเปรียบเทียบผลของแบบจำลองเอ็นแกรม (ไบแกรม และไตรแกรม) แบบจำลองเอ็นแกรมที่ได้รับการปรับปรุง (ไบแกรม และไตรแกรม) และต้นไม้ตัดสินใจ ซึ่งต้นไม้ตัดสินใจมีความสามารถในการเรียนรู้ที่ดีกว่าแบบจำลองเอ็นแกรมทั้ง 2 แบบ แต่เมื่อนำมาทดสอบกับข้อมูลชุดอื่นพบว่าแบบจำลองเอ็นแกรมให้ประสิทธิภาพที่ดีกว่า โดยไตรแกรมให้ประสิทธิภาพดีกว่าไบแกรม แต่ไตรแกรมใช้หน่วยความจำจำนวนมากซึ่งเป็นปัญหาต่อระบบขนาดเล็ก หลังจากที่เปรียบเทียบวิธีแบบต่างๆแล้วจึงได้นำต้นไม้ตัดสินใจและการปรับปรุงแบบจำลองไบแกรมมาใช้ในงานรู้จำเสียงพูด เมื่อเปรียบเทียบประสิทธิภาพกับการใช้คลังข้อมูลปรากฏว่าการใช้คลังข้อมูลให้ผลที่ดีกว่าประมาณ 2 เปอร์เซนต์ ถึงแม้ว่าการใช้คลังข้อมูลจะให้ผลที่ดีกว่าแต่จะต้องเตรียมข้อมูลจำนวนมากเพื่อให้ครอบคลุมคำทั้งหมด ซึ่งในความเป็นจริงแล้วมีคำใหม่ๆเกิดขึ้นมาอยู่ตลอดเวลา การใช้วิธีทางสถิติจึงเป็นอีกแนวทางหนึ่งที่ดีสำหรับการนำมาประยุกต์ใช้ ในงานวิจัยถัดมา Shiho Nobesawa และ Ikuo Tahara ได้นำเสนอการระบุภาษาของชื่อทั้ง 9 ภาษาใน 12 ประเทศที่เขียนด้วยอักษรโรมัน

โดยภาษาไทยเป็นหนึ่งภาษาที่ถูกนำมาพิจารณา แนวทางที่ได้นำเสนอ คือ การพิจารณาความยาวของคำ การใช้แบบจำลองเอ็นแกรม (ยูนิแกรม ไบแกรม ไตรแกรม และอินเทอร์โพเลตไตรแกรม) และการใช้ทั้งอินเทอร์โพเลตไตรแกรมร่วมกับการพิจารณาความยาวของคำ เมื่อเปรียบเทียบวิธีการทั้ง 3 แล้วปรากฏว่าการใช้อินเทอร์โพเลตไตรแกรมพร้อมทั้งพิจารณาความยาวของคำให้ประสิทธิภาพที่ดีที่สุด ทั้งนี้ถ้าทำการพิจารณาเพียงความยาวของคำอย่างเดียวไม่สามารถพิจารณาลักษณะพิเศษของแต่ละภาษาได้ครอบคลุมเท่าที่ควร จึงทำให้ประสิทธิภาพที่ได้ไม่ดีเมื่อเทียบกับวิธีเอ็นแกรม การนำความยาวของคำมาเป็นลักษณะสำคัญอีกหนึ่งอย่างถือเป็นสิ่งที่น่าสนใจเนื่องจาก [11] ได้ระบุว่าชื่อภาษาไทย รัสเซีย กรีกเป็นชื่อที่มีความยาวที่สุดเมื่อเทียบกับภาษาอื่นๆในงานวิจัย เราจึงจะนำความยาวของคำมาเป็นลักษณะสำคัญอีกอย่างหนึ่งในงานวิจัย

Yining Chen และคณะ ได้มีการนำแบบจำลองเอ็นแกรมระดับอักขระมาใช้ และได้นำเสนอการจัดกลุ่มของอักขระด้วยพยางค์ (Syllable-Based Letter Cluster: SBLC) [4] โดยชื่อที่รับเข้ามาในแต่ละภาษาจะมีลักษณะการออกเสียงของแต่ละชื่อด้วย ถ้าชื่อนั้นปรากฏมากกว่า 1 ภาษาจะถูกตัดออกเนื่องจากมีความกำกวมและถ้าชื่อใดไม่ได้อยู่ในอักษรอังกฤษจะถูกแปลงเป็นอักษรอังกฤษก่อนนำมาใช้ ในการสร้าง SBLC ทำโดยการจับคู่อักขระเข้ากับหน่วยเสียงเพื่อแยกคำออกเป็นพยางค์ ในกรณีที่คำที่รับเข้ามาไม่มีอยู่ใน SBLC list ก็จะนำเอากฎเข้ามาเพื่อช่วยแยกพยางค์ ในส่วนของการระบุภาษาได้นำไตรแกรมเข้ามาคำนวณหาความน่าจะเป็นของภาษาที่มีต่อคำนั้น และคำใหม่ที่ถูกกำหนดภาษาเข้าไปใหม่จะมีคะแนนสูงสุด หลังจากที่เปรียบเทียบประสิทธิภาพระหว่างทั้ง 2 วิธีแล้วพบว่าทั้งคู่มีประสิทธิภาพที่ดีในภาษาที่แตกต่างกันและมีความผิดพลาดร่วมกันจึงได้นำเอาเอดาบูซมาใช้เพื่อปรับค่าน้ำหนักของข้อมูลสอนที่จำแนกประเภทได้ยากเพื่อลดอัตราความผิดพลาด

นอกจากการคำนวณค่าความน่าจะเป็นในแบบจำลองเอ็นแกรมแล้ว Samuel Thomas และ Ashish Verma [12] ได้นำเสนอการใช้แบบจำลองเอ็นแกรมระดับอักขระ โดยพิจารณาจากการให้ค่าน้ำหนักของคำในคลาส (Class Frequency-Inverse Overall Frequency: CF-IOF) แต่ยังคงทำการเปรียบเทียบประสิทธิภาพกับแบบจำลองเอ็นแกรมที่คำนวณค่าความน่าจะเป็นแบบปกติและต้นไม้ตัดสินใจที่นำมาใช้สำหรับระบุขนาดของอักขระเพื่อลดจำนวนลักษณะสำคัญที่ถูกนำมาเปรียบเทียบ นอกจากนี้แล้ว Samuel Thomas และ Ashish Verma ได้ทำการแบ่งวิธีสำหรับการประเมินประสิทธิภาพออกเป็น 2 แบบ แบบแรกคือ ใช้แบบจำลองแบบกำหนดแกรมของอักขระที่พิจารณา (Fixed Length) แบบที่สองให้แกรมมีค่ายืดหยุ่น (Varying Length) หลังจากที่ได้ทำการทดสอบกับภาษาจำนวน 3 ภาษา การพิจารณาการให้ค่าน้ำหนัก CF-IOF ให้ความแม่นยำที่ดีกว่าแบบคำนวณค่าความน่าจะเป็นในทุกวิธี โดยผลลัพธ์ของแบบจำลองเอ็นแกรมที่ให้แกรมมีค่ายืดหยุ่นให้ผลใกล้เคียงกับการใช้ต้นไม้ตัดสินใจ แต่ต้นไม้ตัดสินใจใช้แกรมขนาดน้อยกว่าก็สามารถระบุภาษาได้ดีเทียบเท่าเอ็นแกรมที่ใช้แกรมมากกว่า นี่จึงเป็นข้อดีอีกอย่างหนึ่งของต้นไม้ตัดสินใจ

งานวิจัยของ Aditya Bhargava และ Grzegorz Kondrak [13] ได้นำเอ็นแกรมและความยาวของชื่อมาเป็นลักษณะสำคัญอย่างหนึ่งของวิธีซัพพอร์ตเวกเตอร์แมชชีน (Support

Vector Machine: SVM) เพื่อใช้สำหรับระบุภาษาของชื่อบุคคล โดยใช้เคอร์เนลฟังก์ชันแบบต่างๆได้แก่ ลิเนียร์ (Linear) ซิกมอยด์ (Sigmoid) และเรเดียลเบสิส (Radial Basis Function: RBF) จากการทดสอบได้ทำการเปรียบเทียบวิธีซัพพอร์ตเวกเตอร์แมชชีนด้วยเคอร์เนลแบบต่างๆ กับแบบจำลองภาษา โดยการใช้เคอร์เนลฟังก์ชันแบบลิเนียร์ให้ผลดีที่สุด และเมื่อเทียบเคอร์เนลทั้ง 3 ด้วยกันเอง เคอร์เนลฟังก์ชันแบบลิเนียร์ยังมีการประมวลผลที่เร็วกว่า 2 แบบที่เหลือ นอกจากนี้แล้วยังได้นำไปทดสอบกับการระบุภาษาอินเดียก่อนจะนำคำเหล่านั้นไปทำการถอดอักษรโรมันแบบถ่ายตัวอักษร

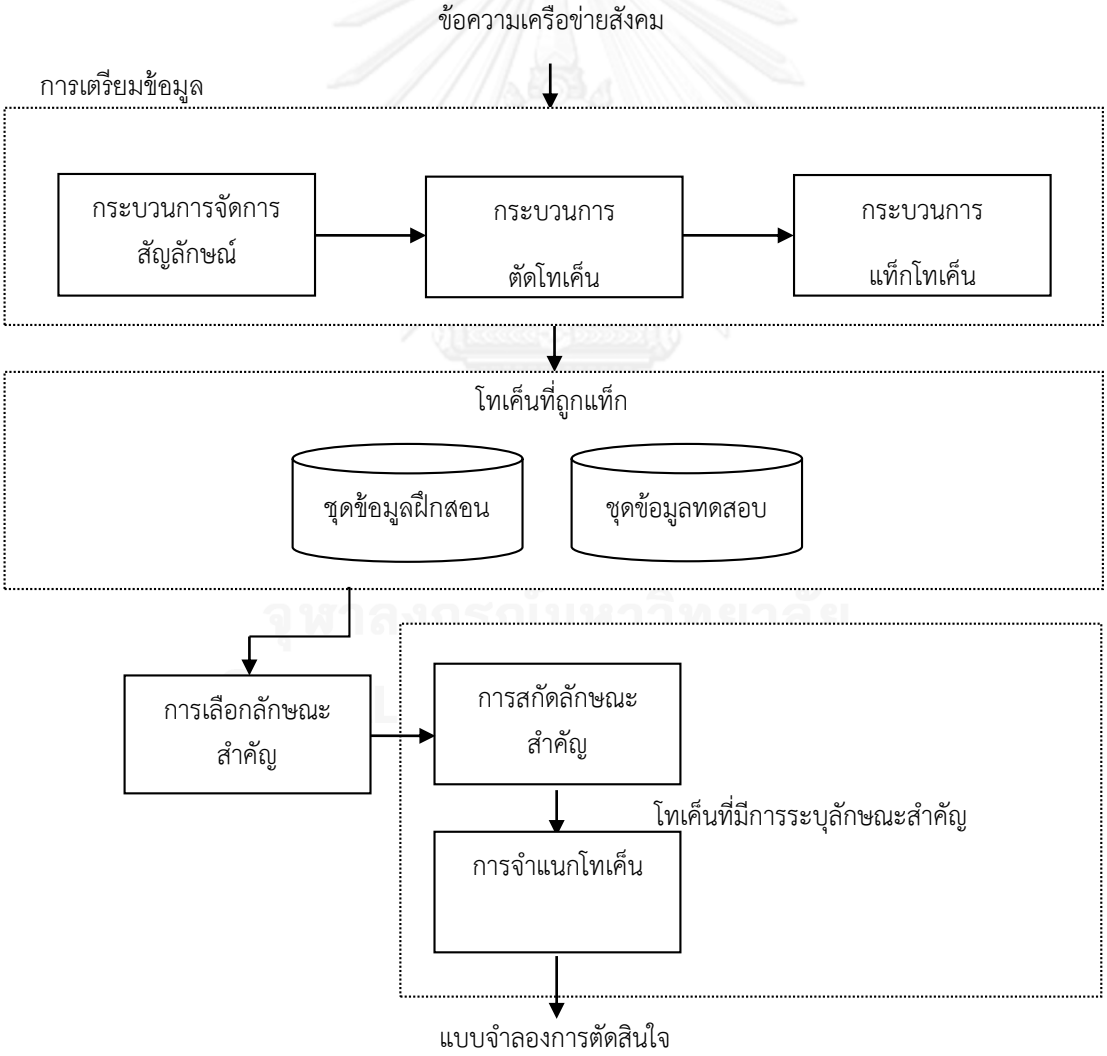
จากการศึกษางานวิจัยพบว่าการใช้แบบจำลองทางสถิติสำหรับการระบุภาษาเป็นวิธีที่ได้รับความนิยมมากที่สุด เช่น การใช้แบบจำลองเอ็นแกรม [4, 5, 11, 12] การใช้ต้นไม้ตัดสินใจ [5, 12] การใช้ซัพพอร์ตเวกเตอร์แมชชีน [13] เป็นต้น นอกจากนี้แล้วการใช้แบบจำลองทางสถิติยังไม่จำเป็นต้องมีความรู้ทางภาษาศาสตร์สูงก็สามารถทำการระบุภาษาได้ โดยแบบจำลองเอ็นแกรมเป็นแบบจำลองที่ได้รับความนิยมจากอดีตจนถึงปัจจุบัน เนื่องจากมีความแม่นยำในการทำนายลำดับของอักขระหรือคำถัดไปที่เป็นไปได้สูง ในขณะที่วิธีต้นไม้ตัดสินใจนั้นมีความสามารถในการเรียนรู้ข้อมูลสอน ง่ายต่อการทำความเข้าใจและตีความ ดังนั้นการนำทั้งสองวิธีนี้มาใช้ร่วมกันอาจจะส่งผลให้ระบบสามารถเพิ่มประสิทธิภาพในการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันได้สูงขึ้น และในงานวิจัยนี้ได้นำการปรับเรียบย้อน [10] มาใช้ในกรณีที่ไม่มีข้อมูลที่สนใจในคลังข้อมูลเนื่องจากเป็นวิธีที่ง่ายและให้ประสิทธิภาพที่ดี

บทที่ 3
ขั้นตอนการดำเนินงานวิจัย

ในบทนี้ผู้วิจัยจะกล่าวถึงภาพรวมของระบบการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคม ซึ่งประกอบด้วย ขั้นตอนการเตรียมข้อมูล การเลือกลักษณะสำคัญ การจำแนกโทเค็น และการสร้างแบบจำลองการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมัน

3.1.ขั้นตอนการสร้างระบบการระบุคำให้อ่านแบบไทย

องค์ประกอบหลักของโครงสร้างสำหรับการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมัน แสดงดังภาพ 3-1



ภาพที่ 3-1 แผนภาพแสดงขั้นตอนการสร้างระบบระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมัน

3.1.1. การเตรียมข้อมูล

การเตรียมข้อมูลเป็นขั้นตอนที่ระบบจะทำการนำข้อความที่ได้รวบรวมจากเว็บไซต์เครือข่ายสังคมมาพิจารณาทีละข้อความ โดยข้อความที่นำเข้ามานั้นต้องมีตัวอักษรภาษาอังกฤษประกอบอยู่ในประโยค ซึ่งประกอบด้วย 3 ขั้นตอนดังแสดงในภาพที่ 3-1

3.1.1.1. กระบวนการจัดการสัญลักษณ์ ขั้นตอนนี้จะทำการพิจารณาหาสัญลักษณ์ในข้อความที่รับเข้ามา ถ้าพบสัญลักษณ์พิเศษในข้อความระบบจะทำการเปลี่ยนสัญลัษณ์พิเศษเหล่านั้นให้เป็นคำที่สื่อถึงสัญลักษณ์พิเศษนั้นๆ เพื่อให้เกิดความสะดวกและเป็นระบบในการจัดเก็บและพิจารณา ตัวอย่างเช่น เมื่อพบสัญลักษณ์ “-“ ระบบจะทำการแทนที่ด้วย “<hyphen>” หรือเมื่อพบสัญลักษณ์ “_” ระบบจะแทนที่สัญลักษณ์นี้ด้วย “<underscore>” เป็นต้น

3.1.1.2. กระบวนการตัดโทเค็น กระบวนการนี้จะนำข้อความมาตัดแบ่งเป็นโทเค็น โดยโทเค็น คือ อักขระ หรือ การประกอบกันของอักขระเป็นคำ วลี หรือประโยคทั้งที่มีความหมายและไม่มีความหมาย ซึ่งได้จากการพิจารณาจากการเว้นวรรคและสัญลักษณ์พิเศษในการแบ่งออกเป็นแต่ละโทเค็น สาเหตุที่เลือกตัดแบ่งโทเค็นตามการเว้นวรรคและการใช้สัญลักษณ์พิเศษ เนื่องจากในการแทรกข้อความภาษาอังกฤษและข้อความภาษาไทยโดยปกติแล้วจะมีการเว้นวรรคเล็กระหว่างคำและสัญลักษณ์พิเศษเหล่านี้ นอกจากนี้ถ้าในกรณีที่ระหว่างอักษรไทยและอังกฤษไม่มีการเว้นวรรคระบบจะทำการตัดอักษรไทยและอักษรอังกฤษเหล่านั้นออกจากกัน เพื่อทำการแบ่งเป็นคนละโทเค็นสำหรับการพิจารณาภาษา

3.1.1.3. การระบุประเภทของโทเค็น หลังจากที่แบ่งข้อความออกเป็นโทเค็นได้แล้ว จะทำการพิจารณาแต่ละโทเค็นในข้อความ เพื่อทำการดูภาษาและประเภทของโทเค็นก่อนที่จะทำการแท็กภาษาให้กับโทเค็นเหล่านั้น สามารถแบ่งประเภทของการแท็กได้ดังตารางที่ 3-1

ตารางที่ 3-1 ตารางแสดงประเภทของการแท็กโทเค็นแบบต่างๆ

THAI	POS	NEG	OTH
THAI (สวัสดี)	TL (naka)	EL (book,happy)	NUM (1)
	TP (phuket)	EP (ipad,happyy)	SYM (@)
			OTH (www.google.com)

ประเภทของการแท็กโทเค็นแบ่งได้เป็น 4 กลุ่มใหญ่ คือ

1) THAI แทนด้วย โทเค็นที่เขียนด้วยอักษรไทยทั้งหมด
   1.1. THAI แทนด้วย โทเค็นที่เขียนด้วยอักษรไทยทั้งหมด
2) POS แทนด้วย โทเค็นที่ให้อ่านแบบไทยที่เขียนด้วยตัวอักษรโรมัน
   2.1. TL แทนด้วย โทเค็นที่ให้อ่านแบบไทยที่ถอดเสียงโดยใช้อักษรโรมัน
   2.2. TP แทนด้วย โทเค็นที่ให้อ่านแบบไทยที่เป็นชื่อเฉพาะ เช่น ชื่อสถานที่ เป็นต้น
3) NEG แทนด้วย โทเค็นที่ให้อ่านแบบต่างประเทศ
   3.1. EL แทนด้วย โทเค็นที่มีปรากฏอยู่ในพจนานุกรม LEXiTRON [14] และ Carnegie Mellon University Pronouncing (CMU) [15]
   3.2. EP แทนด้วย โทเค็นที่ไม่ปรากฏในพจนานุกรม LEXiTRON และ CMU เช่น ชื่อเฉพาะที่ให้อ่านแบบต่างประเทศ โทเค็นที่สะกดผิดหรือตั้งใจเขียนแต่ไม่มีปรากฏในพจนานุกรมทั้ง 2 แบบ เป็นต้น
4) OTH แทนด้วยโทเค็นประเภทอื่นๆนอกเหนือจากที่กล่าว
   4.1. NUM แทนด้วย โทเค็นที่เป็นตัวเลข
   4.2. SYM แทนด้วย โทเค็นที่แทนสัญลักษณ์พิเศษ
   4.3. OTH แทนด้วย โทเค็นประเภทอื่นๆนอกเหนือจากที่กล่าวในกลุ่มนี้ เช่น ชื่อเว็บไซต์ อีเมล เป็นต้น

ตัวอย่างขั้นตอนการดำเนินงานในกระบวนการเตรียมข้อมูลสำหรับนำมาใช้ในงานวิจัย แสดงดังตาราง 3-2

ตารางที่ 3-2 ตัวอย่างขั้นตอนการเตรียมข้อมูล

ข้อความนำเข้า	ลำดับ	ขั้นตอนการดำเนินงาน
HBD na ja n’giftขอให้มีความสุขมากๆ	1	HBD na ja n <apos> giftขอให้มีความสุขมากๆ
	2	|HBD|na|ja|n|<apos>|giftขอให้มีความสุขมากๆ
	3	|HBD|na|ja|n|<apos>|gift|ขอให้มีความสุขมากๆ
	4	|NEG,EP,HBD|POS,TL,na|POS,TL,ja|POS,TL,n| OTH,SYM,<apos>|NEG,EL,gift|THAI,THAI,ขอให้มีความสุขมากๆ
Oh oh na n' wit. Seaw len	1	|Oh oh na n <apos> wit <dot> Seaw len
	2	|Oh|oh|na|n|<apos>|wit|<dot>|Seaw|len
	3	|POS,TL,Oh|POS,TL,oh|POS,TL,na|POS,TL,n| OTH,SYM,<apos>|POS,TP,wit|OTH,SYM,<dot>| POS,TL,Seaw|POS,TL,len

จากประโยคทั้ง 2 เราจะสังเกตุเห็นว่า “n” ถูกแท็กเป็นโทเค็นที่ให้อ่านแบบไทย ที่เป็นเช่นนั้นเนื่องจาก “n” ในประโยคทั้ง 2 มาจากคำว่า “น้อง” จึงทำให้โทเค็นนี้กลายเป็นโทเค็นให้อ่านแบบไทยแทนที่จะเป็นคำศัพท์ต่างประเทศที่ปรากฏในพจนานุกรม

เนื่องจากในงานนี้ได้มีการนำพจนานุกรมมาใช้เป็นส่วนหนึ่งของการแท็กประเภทของโทเค็น และยังได้นำมาใช้เป็นส่วนหนึ่งของการพิจารณาลักษณะสำคัญซึ่งจะอธิบายในหัวข้อ 3.1.2 ตาราง 3-3 จึงแสดงจำนวนโทเค็นของพจนานุกรมแต่ละชนิด คือ LEXiTRON และ CMU ที่ใช้ในงานวิจัยนี้

ตารางที่ 3-3 จำนวนโทเค็นในพจนานุกรม

พจนานุกรม	**จำนวนโทเค็น**
LEXiTRON	52,811
CMU	133,247

3.1.2. การเลือกลักษณะสำคัญ

แนวทางในการเลือกลักษณะสำคัญที่นำมาใช้ในงานวิจัยแสดงดังต่อไปนี้

3.1.2.1. จำนวนตัวอักษร (Count) ลักษณะสำคัญนี้มีแนวคิดมาจากลักษณะสำคัญที่ Shiho Nobesawa และ Ikuo Tahara [11] ได้นำมาใช้ทดสอบในงานวิจัยการระบุภาษาของชื่อบุคคล

3.1.2.2. การปรากฏของโทเค็นในพจนานุกรม (Meaning) เนื่องจากโทเค็นที่พบในพจนานุกรมโดยพื้นฐานจะเป็นโทเค็นต่างประเทศเป็นส่วนมาก ดังนั้นถ้าสามารถกรองโทเค็นบางส่วนออกไปได้ จะช่วยส่งผลให้ระบบมีประสิทธิภาพดีขึ้น

3.1.2.3. ตัวอักษรตัวแรกของโทเค็นที่พิจารณาเป็นตัวพิมพ์ใหญ่ (Capital) ตัวอักษรของโทเค็นที่พิจารณาทั้งหมดเป็นตัวพิมพ์ใหญ่ (Allcapital) อักษรตัวสุดท้ายของโทเค็นก่อนหน้าโทเค็นที่สนใจเป็นอักษรไทย (Thaiprev) และอักษรตัวแรกของโทเค็นหลังจากโทเค็นที่สนใจเป็นอักษรไทย (Thainext) ลักษณะสำคัญเหล่านี้พิจารณามาจากความถี่ที่ปรากฏในชุดข้อมูลสอน โดยเปรียบเทียบกับโทเค็นที่ให้อ่านแบบไทย (POS) และให้อ่านแบบอื่นๆ (NEG) จำนวนโทเค็นที่มีลักษณะสำคัญดังกล่าวแสดงดังตารางที่ 3-4

ตารางที่ 3-4 ความถี่ของลักษณะสำคัญ Capital, Allcapital, Thaiprev และ Thainext เทียบกับแท็ก POS และ NEG

ลักษณะสำคัญ	**POS**	**NEG**
Capital	3,398	2,735
Allcapital	213	830
Thaiprev	237	914
Thainext	511	1,088

3.1.2.4. ลักษณะสำคัญที่ได้จากสัญลักษณ์รอบข้างของโทเค็นที่สนใจ พิจารณามาจากการปรากฏของสัญลักษณ์ในชุดข้อมูลสอน ในชุดข้อมูลสอนมีสัญลักษณ์ปรากฏทั้งหมด 32 สัญลักษณ์ ผู้วิจัยเลือกสัญลักษณ์ 10 อันดับแรกดังแสดงในตารางที่ 3-5 โดยพิจารณาการปรากฏของสัญลักษณ์ในชุดข้อมูลสอนจากสูงไป

ต่ำ หลังจากนั้นทำการพิจารณาสัญลักษณ์ที่เลือกทั้งหมดเพื่อหาลักษณะสำคัญ โดยพิจารณาการปรากฏของสัญลักษณ์นั้นทั้งก่อนหน้าและหลังจากโทเค็นที่สนใจ ความถี่ของสัญลักษณ์นั้นที่ปรากฏก่อนและหลังโทเค็นที่ให้อ่านแบบไทย (POS) และโทเค็นที่ให้อ่านแบบอื่นๆ (NEG) แสดงดังตารางที่ 3-6

ตารางที่ 3-5 สัญลักษณ์ 10 อันดับแรกจากข้อมูลฝึกสอน

ลำดับ	สัญลักษณ์		ลำดับ	สัญลักษณ์	
1	จุด	.	6	ทวิภาค	:
2	อัศเจรีย์	!	7	อะพอสทรอฟี	’
3	แคเรต	^	8	ยัติภังค์	-
4	นขลิขิต	)	9	จุลภาค	,
5	ปรัศนี	?	10	สัญประกาศ	_

ตารางที่ 3-6 ความถี่และเปอร์เซนต์ความแตกต่างของสัญลักษณ์ที่ปรากฏก่อนและหลังโทเค็นที่สนใจ

สัญลักษณ์ก่อนโทเค็นที่สนใจ	POS	NEG	เปอร์เซนต์	สัญลักษณ์หลังโทเค็นที่สนใจ	POS	NEG	เปอร์เซนต์
จุด	71	143	1.14%	จุด	314	261	0.84%
อัศเจรีย์	19	48	0.46%	อัศเจรีย์	137	146	0.14%
แคเรต	8	19	0.17%	แคเรต	131	43	1.39%
นขลิขิต	10	17	0.11%	นขลิขิต	6	15	0.14%
ปรัศนี	13	20	0.11%	ปรัศนี	60	62	0.03%
ทวิภาค	12	74	0.98%	ทวิภาค	234	76	2.51%
อะพอสทรอฟี	157	65	1.46%	อะพอสทรอฟี	251	59	3.05%
ยัติภังค์	44	29	0.07%	ยัติภังค์	48	37	0.17%
จุลภาค	30	69	0.62%	จุลภาค	64	65	0.01%
สัญประกาศ	3	18	0.23%	สัญประกาศ	6	15	0.14%

ในตารางที่ 3-6 คอลัมน์เปอร์เซนต์แสดงการเปรียบเทียบความแตกต่างของการปรากฏของสัญลักษณ์ก่อนและหลังโทเค็นที่ให้อ่านแบบไทยและโทเค็นที่ให้อ่านแบบ

อื่นๆ การที่เปอร์เซนต์ความต่างระหว่างโทเค็นที่ให้อ่านแบบไทยและโทเค็นที่ให้อ่านแบบอื่นๆมีค่าสูงนั้นหมายความว่า ลักษณะสำคัญนั้นมีความสำคัญต่อภาษาใดภาษาหนึ่งมากกว่าอีกภาษาหนึ่ง ซึ่งจะส่งผลให้สามารถทำการแยกโทเค็นของทั้งคู่ออกจากกันได้ง่ายกว่าการที่มีเปอร์เซนต์ต่ำที่มีอัตราการเกิดของโทเค็นที่ใกล้เคียงกัน โดยสามารถเรียงลำดับเปอร์เซนต์จากสูงสุดไปต่ำสุดได้ดังตารางที่ 3-7

ตารางที่ 3-7 ผลการเปรียบเทียบเปอร์เซนต์ความแตกต่างของสัญลักษณ์

ลำดับ	สัญลักษณ์	ก่อน	หลัง	เปอร์เซนต์	ลำดับ	สัญลักษณ์	ก่อน	หลัง	เปอร์เซนต์
1	อะพอสทรอฟี		x	3.05%	11	แคเรต	x		0.17%
2	ทวิภาค		x	2.51%	12	ยัติภังค์		x	0.17%
3	อะพอสทรอฟี	x		1.46%	13	อัศเจรีย์		x	0.14%
4	แคเรต		x	1.39%	14	นขลิขิต		x	0.14%
5	จุด	x		1.14%	15	สัญประกาศ		x	0.14%
6	ทวิภาค	x		0.98%	16	นขลิขิต	x		0.11%
7	จุด		x	0.84%	17	ปรัศนี	x		0.11%
8	จุลภาค	x		0.62%	18	ยัติภังค์	x		0.07%
9	อัศเจรีย์	x		0.46%	19	ปรัศนี		x	0.03%
10	สัญประกาศ	x		0.23%	20	จุลภาค		x	0.01%

จากตารางที่ 3-7 ผู้วิจัยได้ทำการเลือกสัญลักษณ์ที่มีเปอร์เซนต์ความแตกต่างระหว่าง โทเค็นที่ให้อ่านแบบไทยและโทเค็นที่ให้อ่านแบบอื่นๆ 10 อันดับแรกออกมา เพื่อพิจารณาหาลักษณะสำคัญ จากตารางเปอร์เซนต์ของทั้ง 10 อันดับมีค่าใกล้เคียงกัน ดังนั้นเพื่อหาลักษณะสำคัญที่จะนำมาใช้ในงานวิจัย จึงทำการพิจารณาความถี่ของสัญลักษณ์ที่ได้จากตารางที่ 3-7 ในข้อมูลฝึกสอนด้วย สาเหตุที่ต้องพิจารณาความถี่ร่วมกับเปอร์เซนต์ความต่างของโทเค็นที่ให้อ่านแบบไทยและโทเค็นที่ให้อ่านแบบอื่นๆ เนื่องจากถึงแม้ว่าเปอร์เซนต์ความต่างของโทเค็นที่ให้อ่านแบบไทยและโทเค็นที่ให้อ่านแบบอื่นๆจะมีค่าสูง แต่ถ้าอัตราการปรากฏของสัญลักษณ์เหล่านั้นมีต่ำ ระบบจะไม่สามารถแยกภาษาของโทเค็นได้ดีเท่ากับมีความถี่และเปอร์เซนต์ของสัญลักษณ์ที่มีค่าสูง โดยสามารถเรียงลำดับการปรากฏของโทเค็นได้ดังตารางที่ 3-8 ลักษณะสำคัญที่ใช้จะเป็น 5 อันดับแรกที่ได้จากตารางที่ 3-8 ส่วนตารางที่ 3-9 แสดงลักษณะสำคัญและชื่อที่ใช้เรียกลักษณะสำคัญเหล่านั้นในงานวิจัย

ตารางที่ 3-8 ผลการเปรียบเทียบจำนวนโทเค็นของลักษณะสำคัญ

ลำดับ	สัญลักษณ์	ก่อน	หลัง	POS	NEG	รวม	เปอร์เซนต์
1	จุด		x	314	261	575	0.84%
2	อะพอสทรอฟี		x	251	59	310	3.05%
3	ทวิภาค		x	234	76	310	2.51%
4	อะพอสทรอฟี	x		157	65	222	1.46%
5	จุด	x		71	143	214	1.14%
6	แคเรต		x	131	43	174	1.39%
7	จุลภาค	x		30	69	99	0.62%
8	ทวิภาค	x		12	74	86	0.98%
9	อัศเจรีย์	x		19	48	67	0.46%
10	สัญประกาศ	x		3	18	21	0.23%

ตารางที่ 3-9 ลักษณะสำคัญจากสัญลักษณ์

ลำดับ	สัญลักษณ์	ก่อน	หลัง	ลักษณะสำคัญ
1	อะพอสทรอฟี	x		Aposprev
2	อะพอสทรอฟี		x	Aposnext
3	ทวิภาค		x	Colonnext
4	จุด	x		Dotprev
5	จุด		x	Dotnext

นอกจากนี้ผู้วิจัยได้นำลักษณะสำคัญไปทดสอบด้วยข้อมูลฝึกสอนทั้งหมด เพื่อพิจารณาลักษณะสำคัญที่เครื่องมือเวก้าเลือกมาใช้ในการสร้างแบบจำลองต้นไม้ตัดสินใจ ซึ่งลักษณะสำคัญที่เครื่องมือเวก้าเลือกมาใช้แสดงดังตารางที่ 3-10

ตารางที่ 3-10 ลักษณะสำคัญ

ลำดับ	ลักษณะสำคัญ	ลำดับ	ลักษณะสำคัญ
1	Count	7	Thainext
2	Meaning : LEXiTRON	8	Aposprev
3	Meaning : CMU	9	Aposnext
4	Allcapital	10	Colonnext
5	Capital	11	Dotprev
6	Thaiprev	12	Dotnext

3.1.3. การจำแนกโทเค็น

ลักษณะสำคัญที่ใช้ในงานวิจัยนี้ประกอบด้วย 13 ค่า ซึ่งสามารถแบ่งออกเป็น 2 กลุ่มหลักดังนี้

3.1.3.1. การสกัดลักษณะสำคัญของโทเค็น

ลักษณะสำคัญที่นำมาใช้สำหรับงานวิจัยนี้แบ่งออกเป็น 2 กลุ่ม คือ กลุ่มที่พิจารณาเฉพาะโทเค็น และกลุ่มที่พิจารณาจากบริบทรอบข้างของโทเค็นที่สนใจ

1) กลุ่มที่พิจารณาเฉพาะโทเค็น

1.1 จำนวนตัวอักษร (Count) ลักษณะสำคัญนี้แสดงค่าจำนวนอักษรทั้งหมดของโทเค็นที่พิจารณา เช่น คำว่า Hello จะมี Count เท่ากับ 5 ตัวอักษร

1.2 ผลการค้นหาคำในพจนานุกรม ลักษณะสำคัญนี้จะทำการแบ่งเป็น 2 ลักษณะสำคัญย่อย คือ

1.2.1 โทเค็นที่พิจารณามีปรากฏอยู่ในพจนานุกรม LEXiTRON (Meaning: LEXiTRON) ลักษณะสำคัญนี้แสดงค่าการตรวจสอบโทเค็นที่พิจารณากับพจนานุกรม LEXiTRON ถ้ามีโทเค็นเหล่านั้นปรากฏอยู่จะถูกกำหนดค่าเป็น 1 ถ้าไม่มีจะถูกกำหนดค่าเป็น 0

1.2.2 โทเค็นที่พิจารณามีปรากฏอยู่ในพจนานุกรม CMU (Meaning: CMU) ลักษณะสำคัญนี้แสดงค่าการตรวจสอบโทเค็นที่พิจารณากับพจนานุกรม CMU ถ้ามีโทเค็นเหล่านั้นปรากฏอยู่จะถูกกำหนดค่าเป็น 1 ถ้าไม่มีจะถูกกำหนดค่าเป็น 0

1.3 ตัวอักษรตัวแรกของโทเค็นที่พิจารณาเป็นตัวพิมพ์ใหญ่ (Capital) ลักษณะสำคัญนี้แสดงค่าการตรวจสอบตัวอักษรตัวแรกของโทเค็นที่พิจารณา ถ้าพบว่าเป็นตัวพิมพ์ใหญ่ จะถูกกำหนดค่าเป็น 1 ถ้าไม่ใช่จะถูกกำหนดค่าเป็น 0

1.4 ตัวอักษรของโทเค็นที่พิจารณาทั้งหมดเป็นตัวพิมพ์ใหญ่ (Allcapital) ลักษณะสำคัญนี้แสดงค่าการตรวจสอบตัวอักษรทุกตัวในโทเค็น ถ้าพบว่าเป็นตัวพิมพ์ใหญ่ จะถูกกำหนดค่าเป็น 1 ถ้าไม่ใช่ตัวพิมพ์ใหญ่ทั้งหมดจะถูกกำหนดค่าเป็น 0

2) กลุ่มที่พิจารณาจากบริบทรอบข้างของโทเค็นที่สนใจ

โดยการพิจารณาบริบทของโทเค็นจะพิจารณาเฉพาะโทเค็นที่ปรากฏก่อนหน้าและหลังของโทเค็นที่สนใจหนึ่งโทเค็นเท่านั้น เช่น ข้อความ “hbd na ja apple” เมื่อตัวอย่างของโทเค็นที่สนใจคือ na บริบทของโทเค็นที่พิจารณาก่อนและหลังคือ “hbd” และ “na” ตามลำดับ

2.1 โทเค็นก่อนหน้าเป็นอักษรไทย (Thaiprev) ลักษณะสำคัญนี้แสดงค่าการพิจารณาอักษรของโทเค็นก่อนหน้าโทเค็นที่สนใจว่าเป็นอักษรไทยหรือเปล่า ถ้าเป็นอักษรไทยลักษณะสำคัญนี้จะถูกกำหนดค่าเป็น 1 ถ้าไม่ใช่อักษรไทยจะถูกกำหนดค่าเป็น 0

2.2 โทเค็นที่ตามหลังเป็นอักษรไทย (Thainext) ลักษณะสำคัญนี้แสดงค่าการพิจารณาอักษรของโทเค็นที่ตามหลังโทเค็นที่สนใจว่าเป็นอักษรไทยหรือเปล่า ถ้าเป็นอักษรไทยลักษณะสำคัญนี้จะถูกกำหนดค่าเป็น 1 ถ้าไม่ใช่อักษรไทยจะถูกกำหนดค่าเป็น 0

2.3 โทเค็นก่อนหน้าเป็นอะพอสทรอฟี (Aposprev) ลักษณะสำคัญนี้แสดงค่าการพิจารณาโทเค็นก่อนหน้าโทเค็นที่สนใจว่าเป็นอะพอสทรอฟีหรือเปล่า ถ้าเป็นอะพอสทรอฟีลักษณะสำคัญนี้จะถูกกำหนดค่าเป็น 1 ถ้าไม่ใช่อะพอสทรอฟีจะถูกกำหนดค่าเป็น 0

2.4 โทเค็นที่ตามหลังเป็นอะพอสทรอฟี (Aposnext) ลักษณะสำคัญนี้แสดงค่าการพิจารณาโทเค็นที่ตามหลังโทเค็นที่สนใจว่าเป็นอะพอสทรอฟีหรือเปล่า ถ้าเป็นอะพอสทรอฟีลักษณะสำคัญนี้จะถูกกำหนดค่าเป็น 1 ถ้าไม่ใช่อะพอสทรอฟีจะถูกกำหนดค่าเป็น 0

2.5 โทเค็นที่ตามหลังเป็นทวิภาค (Colonnext) ลักษณะสำคัญนี้แสดงค่าการพิจารณาโทเค็นที่ตามหลังโทเค็นที่สนใจว่าเป็นทวิภาคหรือเปล่า ถ้าเป็นทวิภาค ลักษณะสำคัญนี้จะถูกกำหนดค่าเป็น 1 ถ้าไม่ใช่ทวิภาคจะถูกกำหนดค่าเป็น 0

2.6 โทเค็นก่อนหน้าเป็นจุด (Dotprev) ลักษณะสำคัญนี้แสดงค่าการพิจารณาโทเค็นก่อนหน้าโทเค็นที่สนใจว่าเป็นจุดหรือเปล่า ถ้าเป็นจุดลักษณะสำคัญนี้จะถูกกำหนดค่าเป็น 1 ถ้าไม่ใช่จุดจะถูกกำหนดค่าเป็น 0

2.7 โทเค็นที่ตามหลังเป็นจุด (Dotnext) ลักษณะสำคัญนี้แสดงค่าการพิจารณาโทเค็นที่ตามหลังโทเค็นที่สนใจว่าเป็นจุดหรือเปล่า ถ้าเป็นจุดลักษณะสำคัญนี้จะถูกกำหนดค่าเป็น 1 ถ้าไม่ใช่จุดจะถูกกำหนดค่าเป็น 0

3.1.3.2. การหาลักษณะสำคัญเอ็นแกรม

การคำนวณค่าความน่าจะเป็นด้วยแบบจำลองเอ็นแกรม เพื่อนำมาเป็นลักษณะสำคัญอีกอย่างหนึ่งของแบบจำลองการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมัน เครื่องมือที่นำมาใช้ในการคำนวณค่าความน่าจะเป็นในงานวิจัยนี้ คือ CMU-Cambridge Statistical Language Modeling Toolkit [16] ทำการพิจารณาเอ็นแกรมในระดับอักขระเพื่อทำนายภาษา [5, 11, 12] ซึ่งในงานวิจัยนี้ใช้การสร้างแบบจำลองเอ็นแกรม แบบไตรแกรม (N = 3 อักขระ) เพื่อนำมาเป็นลักษณะสำคัญและทำการแก้ปัญหาการขาดแคลนของข้อมูลฝึกสอนของไตรแกรมด้วยการใช้การปรับเรียบแบบย้อน [8]

ขั้นตอนการคำนวณค่าความน่าจะเป็นด้วยเครื่องมือ CMU-Cambridge Statistical Language Modeling Toolkit

1) เตรียมข้อมูลที่ใช้สำหรับฝึกสอน ทำการแบ่งกลุ่มของโทเค็นออกเป็น 2 กลุ่ม แยกเป็น 2 ไฟล์สำหรับการฝึกสอน คือ กลุ่มของโทเค็นที่ให้อ่านแบบไทย และกลุ่มของโทเค็นที่ให้อ่านแบบอื่นๆ
2) ทำการเว้นวรรคอักษรแต่ละตัวในโทเค็นที่ต้องการฝึกสอนดังภาพที่ 3-2
3) พิมพ์คำสั่งดังแสดงด้านล่างในคอมมานด์ไลน์ เพื่อเรียกเครื่องมือสำหรับคำนวณหาค่าความน่าจะเป็นของแบบจำลองเอ็นแกรม

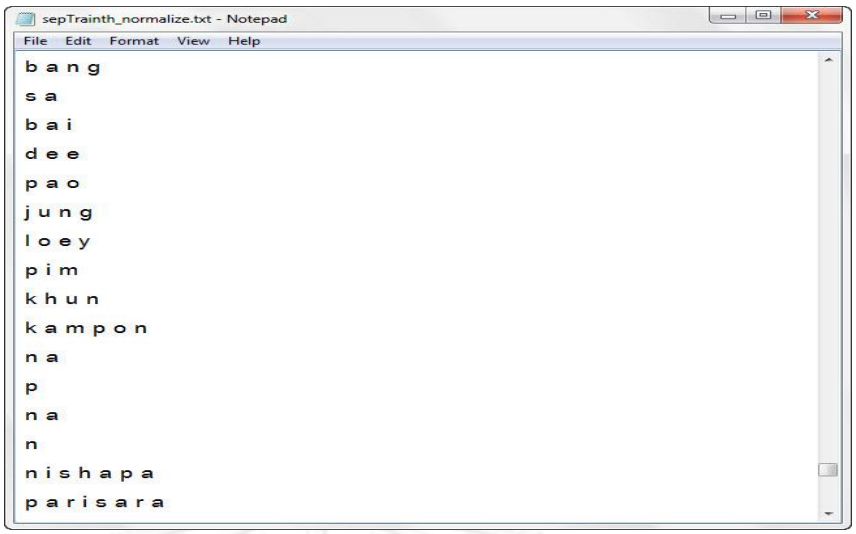
3.1 text2wfreq.exe <input.txt >result.wfreq

3.2 wfreq2vocab.exe < result.wfreq > result.vocab

3.3 text2idngram.exe -vocab result.vocab -idngram result.idngram < input.txt

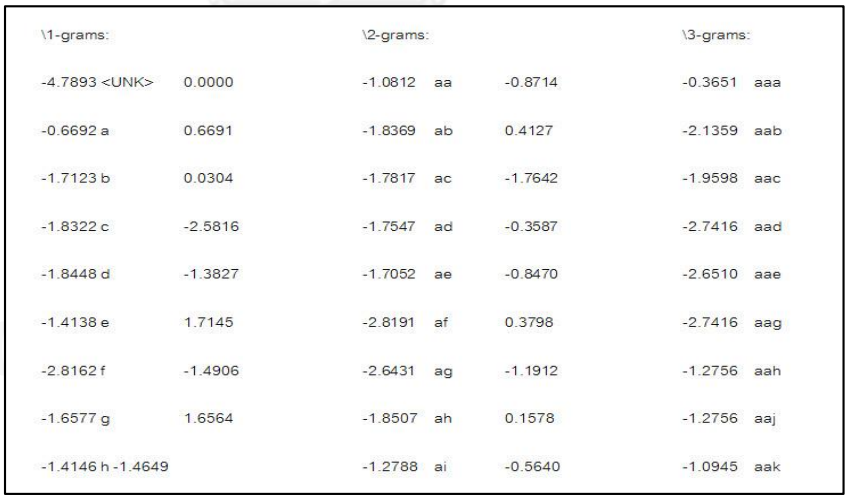
3.4 idngram2lm.exe -idngram result.idngram -vocab result.vocab -arpa result.arpa

เครื่องมือนี้จะทำการอ่านไฟล์ข้อมูลเพื่อคำนวณความถี่ของอักขระที่เกิดขึ้นในชุดข้อมูลสอนทั้งหมดก่อน หลังจากนั้นจึงทำการคำนวณหาค่าความน่าจะเป็นของอักขระ ตั้งแต่ 1 แกรม, 2 แกรม และ 3 แกรมตามลำดับ



ภาพที่ 3-2 ตัวอย่างข้อมูลสำหรับ CMU-Cambridge Statistical Language Modeling Toolkit

ตัวอย่างผลการคำนวณความน่าจะเป็นด้วย CMU-Cambridge Statistical Language Modeling Toolkit ดังภาพที่ 3-3



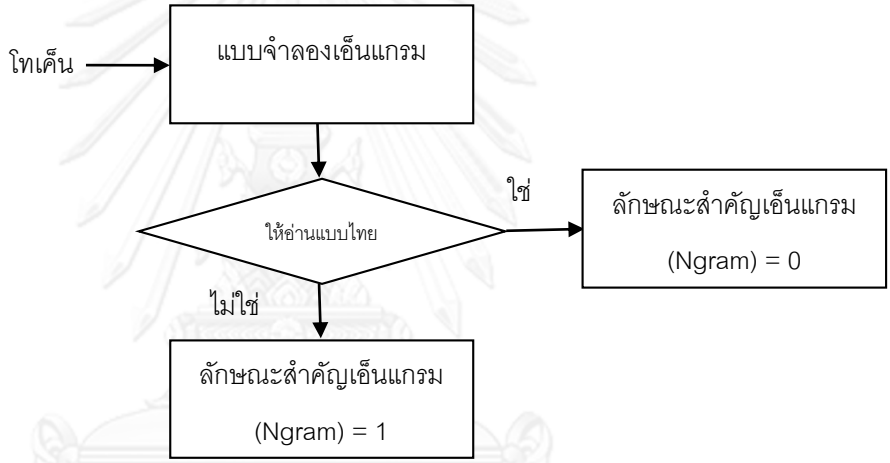
\1-grams:		\2-grams:		\3-grams:	
-4.7893 <UNK>	0.0000	-1.0812 aa	-0.8714	-0.3651 aaa	
-0.6692 a	0.6691	-1.8369 ab	0.4127	-2.1359 aab	
-1.7123 b	0.0304	-1.7817 ac	-1.7642	-1.9598 aac	
-1.8322 c	-2.5816	-1.7547 ad	-0.3587	-2.7416 aad	
-1.8448 d	-1.3827	-1.7052 ae	-0.8470	-2.6510 aae	
-1.4138 e	1.7145	-2.8191 af	0.3798	-2.7416 aag	
-2.8162 f	-1.4906	-2.6431 ag	-1.1912	-1.2756 aah	
-1.6577 g	1.6564	-1.8507 ah	0.1578	-1.2756 aaj	
-1.4146 h -1.4649		-1.2788 ai	-0.5640	-1.0945 aak	

ภาพที่ 3-3 ตัวอย่างผลการคำนวณความน่าจะเป็นด้วย CMU-Cambridge Statistical Language Modeling Toolkit

ค่าที่ปรากฏด้านซ้ายมือของยูนิแกรม (N=1) และ ไบแกรม (N=2) คือ ความน่าจะเป็นของการเกิดคำที่ปรากฏในแถวที่ 2 ในขณะที่ค่าในแถวที่ 3 เป็นค่าปรับเรียบแบบย้อนของไบแกรม และไตรแกรม (N=3) ตามลำดับ ซึ่งใช้คำนวณในกรณีที่ไม่พบคำที่เราสนใจใน

ข้อมูลสอน จากภาพข้างต้นจะเห็นได้ว่าไตรแกรมไม่มีค่าในแถวที่ 3 เนื่องจากไม่มีการคำนวณค่าความน่าจะเป็นในระดับควอดแกรม (N=4) จึงไม่มีค่าปรับเรียบแบบย้อนของควอดแกรมนั่นเอง หลังจากที่เราได้ค่าความน่าจะเป็นในการเกิดอักขระของทั้ง 2 แบบจำลองแล้ว การหาค่าความเป็นไปได้ในการเกิดของโทเค็นในแต่ละภาษาจะเป็นดังสมการ 2.5 ที่ได้นำเสนอในบทที่ 2 ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

หลังจากที่ได้ผลการทำนายภาษาของแต่ละโทเค็นแล้ว จึงนำผลการทำนายเหล่านั้นมาเป็นลักษณะสำคัญ โดยถ้าผลที่ได้จากแบบจำลองเอ็นแกรมทำนายว่าเป็นโทเค็นที่ให้อ่านแบบไทย ระบบจะทำการกำหนดค่าลักษณะสำคัญเอ็นแกรมเป็น 0 แต่ถ้าทำนายเป็นโทเค็นที่ให้อ่านแบบอื่นๆ ลักษณะสำคัญเอ็นแกรมจะถูกกำหนดค่าเป็น 1 แสดงแนวทางการกำหนดค่าลักษณะสำคัญเอ็นแกรมดังภาพที่ 3-4



ภาพที่ 3-4 โครงสร้างของกระบวนการการกำหนดค่าลักษณะสำคัญเอ็นแกรม

3.1.4. การสร้างแบบจำลองสำหรับการระบุคำให้อ่านแบบไทย

การสร้างแบบจำลองสำหรับการทำนายผลการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันในงานวิจัยนี้ ทำโดยใช้แบบจำลองเอ็นแกรมร่วมกับการใช้ต้นไม้ตัดสินใจ โดยนำเครื่องมือเวก้า [17] ที่พัฒนาด้วยมหาวิทยาลัย Waikato มาใช้ในการสร้างต้นไม้ตัดสินใจ

3.1.4.1. ลักษณะสำคัญของชุดข้อมูล

ผู้วิจัยได้นำผลการเปรียบเทียบภาษาที่ได้จากการคำนวณความน่าจะเป็นจากแบบจำลองเอ็นแกรมมาเป็นลักษณะสำคัญ (Ngram) อีกอย่างหนึ่งร่วมกับลักษณะสำคัญที่สกัดได้ทั้ง 2 กลุ่ม คือ กลุ่มที่พิจารณาเฉพาะโทเค็น และกลุ่มที่พิจารณาจากบริบทรอบข้างของโทเค็นที่สนใจโดยทุกโทเค็นที่พิจารณาจะถูกสกัดลักษณะสำคัญทั้ง 13 ค่าดังที่กล่าวออกมา เพื่อใช้เป็นข้อมูลให้เครื่องมือเวก้า ในการสร้างแบบจำลองการระบุคำให้อ่านแบบไทย

ตัวอย่างข้อมูลที่ทำการสกัดลักษณะสำคัญของประโยค “HBD na ja n’giftขอให้มีความสุขมากๆ” แสดงดังตารางที่ 3-11

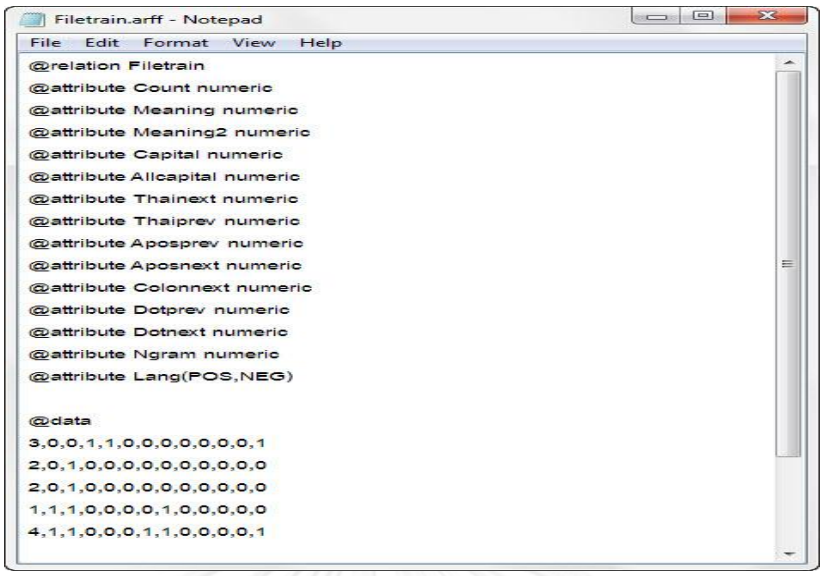
ตารางที่ 3-11 ตัวอย่างข้อมูลที่ทำการสกัดลักษณะสำคัญ

Prev Token	Token	Next Token	Count	Meaning: LEXiTRON	Meaning: CMU	Capital	Allcapital	Thaiprev	Thainext
	HBD	na	3	0	0	1	1	0	0
HBD	na	ja	2	0	1	0	0	0	0
na	ja	n	2	0	1	0	0	0	0
ja	n	<apos>	1	1	1	0	0	0	0
<apos>	gift	ขอให้มีความสุขมากๆ	4	1	1	0	0	0	1
Prev Token	**Token**	**Next Token**	**Aposprev**	**Aposnext**	**Colonnext**	**Dotprev**	**Dotnext**	**Ngram**	
	HBD	na	0	0	0	0	0	1	
HBD	na	ja	0	0	0	0	0	0	
na	ja	n	0	0	0	0	0	0	
ja	n	<apos>	0	1	0	0	0	0	
<apos>	gift	ขอให้มีความสุขมากๆ	1	0	0	0	0	1	

3.1.4.2. ลักษณะชุดข้อมูลที่ใช้สำหรับเครื่องมือเวก้า

หลังจากที่ได้ลักษณะสำคัญของชุดข้อมูลฝึกสอนแล้ว ผู้วิจัยจะนำชุดข้อมูลฝึกสอนมาแปลงให้อยู่ในลักษณะของไฟล์ที่นำมาใช้กับเครื่องมือเวก้า คือ ไฟล์ที่ลงท้ายด้วย *.arff ซึ่งมีโครงสร้างดังนี้

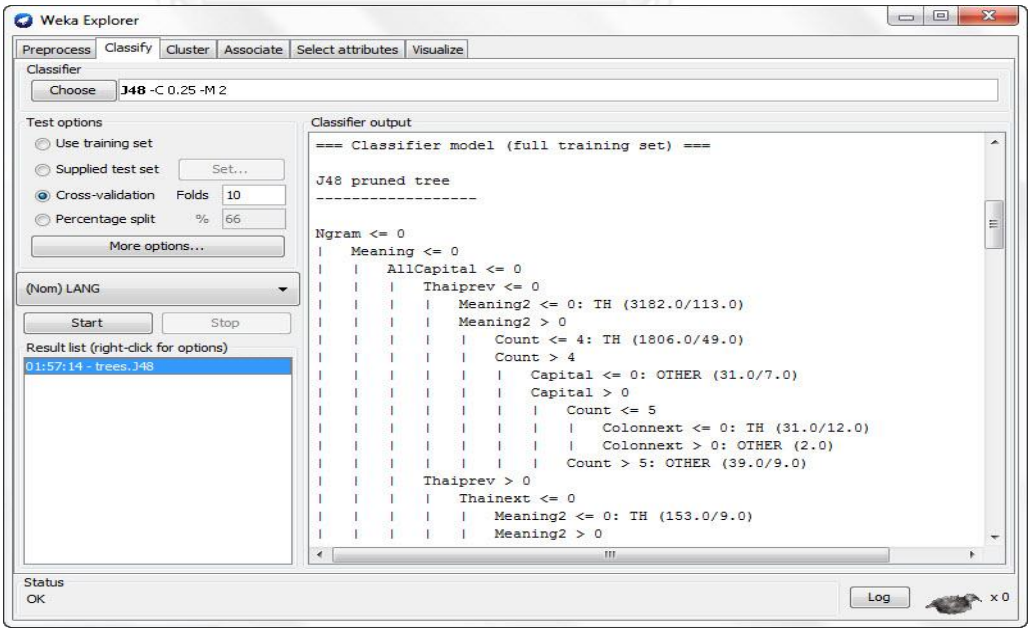
1) บรรทัดแรกของไฟล์ คือ @relation Filetrain ข้อความนี้แสดงชื่อของชุดข้อมูลที่นำมาใช้สร้างแบบจำลอง ดังภาพที่ 3-5
2) บรรทัดถัดมาแสดงชื่อของลักษณะสำคัญที่นำมาใช้ในการสร้างแบบจำลอง @attribute attributename หลังจาก attributename แล้วจะตามด้วยชนิดของ attribute ที่เราใช้ในชุดข้อมูล เช่น numeric, nominal เป็นต้น ถ้าเป็น nominal จะแสดงค่าที่เป็นไปได้ทั้งหมดในวงเล็บ ดัง Attribute Lang ในภาพที่ 3-5
3) attribute สุดท้ายจะเป็นคลาสของโทเค็น โดยจะระบุว่าเป็นคลาสอะไร
4) @data จะเป็นส่วนเริ่มต้นของชุดข้อมูล โดยจะแบ่งแต่ละ attribute ด้วยจุลภาค



ภาพที่ 3-5 ตัวอย่างรูปแบบไฟล์ .arff

3.1.4.3. การกำหนดค่าในเครื่องมือเวก้า

หลังจากที่ได้ไฟล์ที่ใช้กับเครื่องมือเวก้าแล้ว เราจะนำไฟล์นั้นมาสร้างแบบจำลองสำหรับการระบุคำให้อ่านแบบไทยด้วยอัลกอริทึม C4.5 [18] หรือในเครื่องมือเวก้า เรียกว่า j48 โดยได้ทำการกำหนดค่า Cross-validation เป็นจำนวน 10 Folds และให้พารามิเตอร์ต่างๆสำหรับอัลกอริทึมที่ใช้จำแนกมีค่าเป็นค่าตั้งต้นของเครื่องมือ ตัวอย่างของต้นไม้ของแบบจำลองการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันแสดงดังภาพที่ 3-6



ภาพที่ 3-6 ตัวอย่างต้นไม้ของแบบจำลองการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมัน

ขั้นตอนการดำเนินงานวิจัยที่ได้นำเสนอภายในบทนี้ประกอบไปด้วย การเตรียมข้อมูล แนวทางการเลือกลักษณะสำคัญ วิธีการจำแนกโทเค็นจากการสกัดลักษณะสำคัญต่างๆ และการสร้างแบบจำลองการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมัน โดยขั้นตอนการดำเนินงานเหล่านี้จะถูกนำไปใช้กับข้อมูลฝึกสอนและข้อมูลทดสอบสำหรับการทดลองต่างๆในงานวิจัย ซึ่งจะอธิบายในบทถัดไป

บทที่ 4
การทดลอง และวิธีการวัดผลการทดลอง

ในบทนี้จะกล่าวถึงข้อมูลที่นำมาใช้งานวิจัย วิธีการวัดผลการทดลองและการทดลองแบบต่างๆ โดยทำการทดลองกับชุดข้อมูลซึ่งรวบรวมจากเว็บไซต์เครือข่ายสังคม โดยเนื้อหาภายในบทนี้ประกอบด้วย 1. ข้อมูลที่ใช้ในการทดลอง 2. การวัดผลการทดลอง 3. การทดลองเพื่อหาระบบอ้างอิง 4. การระบุคำให้อ่านแบบไทยโดยพิจารณาจากการตัดสินของคน 5. การเปรียบเทียบลักษณะสำคัญด้วยต้นไม้ตัดสินใจ 6. การประเมินประสิทธิภาพของแนวทางที่นำเสนอ

4.1.ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่นำมาใช้ในการทดลองนี้มาจากข้อความที่สื่อสารในเว็บไซต์เฟซบุ๊ก โดยข้อความที่พิจารณามาจากผู้ใช้งานคนไทยจำนวน 980 คน ที่ทำการประกาศข้อความ โดยข้อความที่พิจารณาคือ ข้อความที่มีอักษรอังกฤษปรากฏอยู่แต่ไม่มีอักษรที่แสดงเสียง อักษรพิเศษที่นอกเหนือจากอักษรภาษาอังกฤษรวมไปถึงชื่อเว็บไซต์และอีเมลจะไม่ถูกระบุภาษา จากเงื่อนไขดังกล่าวทำให้จำนวนข้อความทั้งหมดที่พิจารณาจาก 36,497 เป็น 10,527 ข้อความ ข้อมูลทั้งหมดแบ่งออกเป็น 2 ชุด คือ ข้อมูลฝึกสอน (training data) จำนวน 9,360 ข้อความและข้อมูลทดสอบ (test data) จำนวน 1,167 ข้อความ จำนวนข้อมูลในงานวิจัยนี้แสดงดังตารางที่ 4-1

ตารางที่ 4-1 จำนวนข้อความในงานวิจัย

จำนวนของ	ชุดข้อมูลฝึกสอน	ชุดข้อมูลทดสอบ	รวม
ข้อความทั้งหมด	32,594	3,903	36,497
ข้อความที่พิจารณา	9,360	1,167	10,527

จำนวนโทเค็นที่นำมาใช้ในชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบแสดงดังตารางที่ 4-2 และตารางที่ 4-3 แสดงจำนวนโทเค็นที่แยกตามประเภทการแท็กข้อมูล จากตาราง 4-2 แสดงให้เห็นว่าจำนวนโทเค็นส่วนมากที่ใช้ในเครือข่ายสังคม คือ โทเค็นที่ให้อ่านแบบอื่นๆซึ่งมีมากกว่า 30,000 โทเค็น โดยอัตราส่วนระหว่างจำนวนโทเค็นที่ให้อ่านแบบไทยและโทเค็นที่ให้อ่านแบบอื่นๆในชุดข้อมูลฝึกสอนมีอัตราส่วนต่างกันถึง 5 เท่า ดังนั้นเพื่อป้องกันไม่ให้เกิดความลำเอียงในการฝึกสอนสำหรับการสร้างแบบจำลองในการทำนายผล จึงให้จำนวนโทเค็นระหว่างโทเค็นที่ให้อ่านแบบไทยและโทเค็นที่ให้อ่านแบบอื่นๆมีจำนวนเท่ากัน คือ 6,286 โทเค็น

เนื่องจากในการใช้งานจริงในเครือข่ายสังคม โทเค็นที่ให้อ่านแบบอื่นๆมีปรากฏมากกว่าโทเค็นที่ให้อ่านแบบไทย ดังจะเห็นได้จากจำนวนโทเค็นที่ปรากฏทั้งในชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ เราไม่สามารถกำหนดให้ข้อความในเครือข่ายสังคมมีจำนวนโทเค็นที่ให้อ่านแบบไทยเท่ากับ

โทเค็นที่ให้อ่านแบบอื่นๆได้ ดังนั้นจำนวนโทเค็นทดสอบในงานวิจัยนี้จึงเลือกที่จะใช้จำนวนโทเค็นตามจริงที่ปรากฏในข้อความทดสอบ คือ 5,324 โทเค็น

ตารางที่ 4-2 จำนวนของโทเค็นในงานวิจัย

จำนวนของ	**ชุดข้อมูลฝึกสอน**			**ชุดข้อมูลทดสอบ**		
	POS	NEG	รวม	POS	NEG	รวม
โทเค็นทั้งหมด	6,286	31,733	38,019	996	4,328	5,324
โทเค็นที่พิจารณา	6,286	6,286	12,572			

ตารางที่ 4-3 จำนวนของโทเค็นแยกตามประเภทการแท็ก

จำนวนของ	**POS**			**NEG**		
	TL	TP	รวม	EL	EP	รวม
ชุดข้อมูลฝึกสอน	3,205	3,081	6,286	5,241	1,045	6,286
ชุดข้อมูลทดสอบ	534	462	996	3,772	556	4,328

4.2.การวัดผลทดลอง

การวัดประสิทธิภาพในการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคม ทำโดยพิจารณาจากเปอร์เซนต์ค่าความแม่นยำจากสมการที่ 4.1 โดยภาษาที่ถูกของแต่ละโทเค็นได้จากผู้ประเมินจำนวน 2 คน ในกรณีที่ภาษาที่ได้จากผู้ประเมินทั้ง 2 คน ไม่เหมือนกัน จะให้ผู้ประเมินคนที่ 3 ทำการระบุภาษาของโทเค็นนั้น หลังจากที่ระบบทำการระบุภาษาของโทเค็นแล้วจะนำเอาภาษาของแต่ละโทเค็นที่ได้มาเปรียบเทียบกับภาษาของโทเค็นที่ได้จากผู้ประเมิน โดยถ้าระบบทำนายได้ตรงกับภาษาที่ถูกประเมินจะนับเป็นโทเค็นที่ระบบทำนายได้ถูกต้อง

$$\text{เปอร์เซนต์ความแม่นยำ (Accuracy)} = \frac{\text{จำนวนโทเค็นที่ระบบทำนายได้ถูกต้อง}}{\text{จำนวนโทเค็นทั้งหมด}} \times 100 \qquad (4.1)$$

4.3.การทดลอง

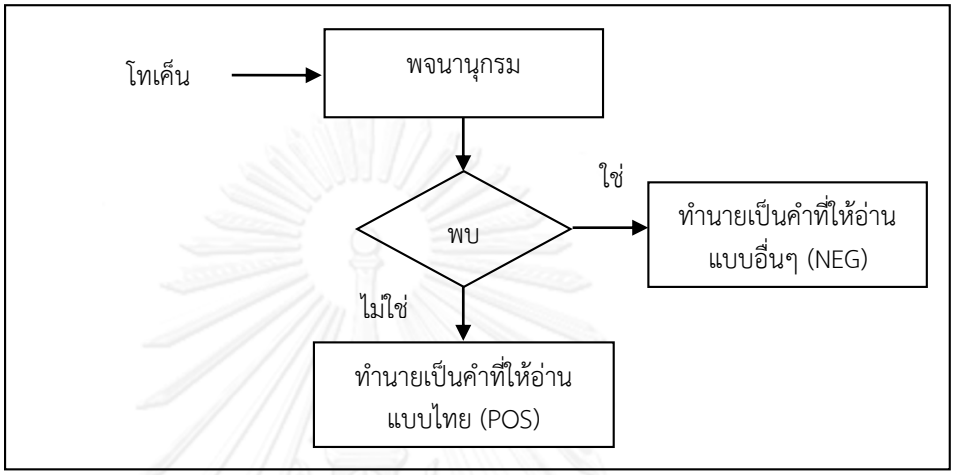
4.1.1. การทดลองที่ 1: การหาระบบอ้างอิง

การทดลองนี้เป็นการทดสอบประสิทธิภาพการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันด้วยวิธีการใช้พจนานุกรม เพื่อทำการหาระบบอ้างอิงที่จะนำมาใช้วัดประสิทธิภาพกับวิธีที่นำเสนอในงานวิจัยนี้ โดยทำการเปรียบเทียบประสิทธิภาพจากพจนานุกรม 3 แบบดังนี้

4.1.1.1. การใช้พจนานุกรม LEXiTRON

4.1.1.2. การใช้พจนานุกรม CMU

4.1.1.3. การใช้พจนานุกรม LEXiTRON ร่วมกับ CMU



ภาพที่ 4-1 โครงสร้างของกระบวนการการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันด้วยพจนานุกรม

โครงสร้างของขั้นตอนการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมัน แสดงดังภาพที่ 4-1 โดยทำการพิจารณาโทเค็นที่สนใจจากพจนานุกรมทีละแบบทั้ง 3 แบบ เพื่อทำนายผลของแต่ละโทเค็นที่สนใจ ถ้าโทเค็นที่สนใจถูกพบในพจนานุกรมให้ทำนายว่าโทเค็นนั้นเป็นโทเค็นที่ให้อ่านแบบอื่นๆ ถ้าไม่พบให้ทำนายว่าเป็นโทเค็นที่ให้อ่านแบบไทย สาเหตุที่ผู้วิจัยให้โทเค็นที่พบในพจนานุกรมเป็นโทเค็นที่ให้อ่านแบบอื่นๆ เนื่องจากในพจนานุกรมจะปรากฏคำศัพท์ต่างประเทศมากกว่าคำไทย ดังนั้นโอกาสที่จะทำนายถูกมีสูงกว่าการให้โทเค็นที่พบในพจนานุกรมเป็นโทเค็นที่ให้อ่านแบบไทย ข้อมูลที่นำมาใช้เพื่อหาระบบอ้างอิงเป็นข้อมูลชุดทดสอบมีจำนวน 5,324 โทเค็น

4.1.2. การทดลองที่ 2: การระบุคำให้อ่านแบบไทยโดยพิจารณาจากการตัดสินของคน

การทดลองที่ 2 นี้เป็นการทดลองเพื่อเปรียบเทียบประสิทธิภาพการระบุคำให้อ่านแบบไทยจากการตัดสินใจของคน ระหว่างการพิจารณาเฉพาะโทเค็น และการพิจารณาบริบทของโทเค็นร่วมด้วย ดังนั้นการทดลองนี้แบ่งการทดสอบการระบุคำให้อ่านแบบไทยจากการพิจารณาโทเค็นที่สนใจเป็น 2 แบบ คือ

4.1.2.1. การพิจารณาเฉพาะโทเค็นที่สนใจโดยปราศจากบริบทร่วม คือ การพิจารณาทีละโทเค็นของข้อความ

4.1.2.2. การพิจารณาโทเค็นที่สนใจโดยดูจากบริบทร่วมด้วย คือ การพิจารณาบริบทของโทเค็นโดยจะพิจารณาเฉพาะโทเค็นที่ปรากฏก่อนหน้าและหลังของโทเค็นที่สนใจหนึ่งโทเค็นเท่านั้น

ในการทดลองนี้ให้ผู้ทดสอบคนไทยตัดสินใจระบุคำให้อ่านแบบไทย

1) จำนวนของผู้ทดสอบมีทั้งหมด 6 คน แบ่งเป็นเพศชาย 3 คนและเพศหญิง 3 คน
2) จำนวนโทเค็นที่ใช้ในการทดสอบมีจำนวนทั้งหมด 200 โทเค็น จาก 175 ข้อความที่สุ่มจากข้อความในข้อมูลทดสอบทั้งหมด โดยแบ่งเป็นโทเค็นที่ให้อ่านแบบไทยจำนวน 100 โทเค็น และโทเค็นที่ให้อ่านแบบอื่นๆจำนวน 100 โทเค็น ให้ผู้ทดสอบทำการพิจารณาได้แบบไม่จำกัดเวลา

โดยทั้ง 2 การทดสอบจะใช้โทเค็นชุดเดียวกันแต่เปลี่ยนจากการพิจารณาเฉพาะโทเค็นเป็นการพิจารณาบริบทร่วมด้วย ตัวอย่างของการระบุคำให้อ่านแบบไทยของประโยค “HBD na ja n’giftขอให้มีความสุขมากๆ” ทั้ง 2 การทดสอบ แสดงดังตารางที่ 4-4 และ 4-5 ตามลำดับ

ตารางที่ 4-4 ตัวอย่างการทดลองที่พิจารณาเฉพาะโทเค็นที่สนใจ

โทเค็น	**POS/NEG**
HBD	NEG
na	POS
ja	POS
n	POS
gift	NEG

ตารางที่ 4-5 ตัวอย่างการทดลองที่พิจารณาโทเค็นที่สนใจโดยดูจากบริบทรอบข้างร่วมด้วย

โทเค็นก่อนหน้า	**โทเค็น**	**โทเค็นถัดไป**	**POS/NEG**
	HBD	na	NEG
HBD	na	ja	POS
na	ja	n	POS
ja	n	<apos>	POS
<apos>	gift	ขอให้มีความสุขมากๆ	NEG

4.1.3. การทดลองที่ 3: การเปรียบเทียบลักษณะสำคัญด้วยต้นไม้ตัดสินใจ

การทดลองที่ 3 นี้เป็นการทดลองเพื่อเปรียบเทียบประสิทธิภาพของลักษณะสำคัญที่พิจารณาเฉพาะโทเค็นและลักษณะสำคัญที่ได้จากการพิจารณาเฉพาะโทเค็นและบริบทรอบข้างของโทเค็น โดยการระบุคำให้อ่านแบบไทยใช้แบบจำลองต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 เพื่อสร้างแบบจำลองในการทำนายผล ซึ่งลักษณะสำคัญในการฝึกสอน และทดสอบด้วยต้นไม้ตัดสินใจแตกต่างกัน จึงแบ่งออกเป็น 2 ส่วน คือ

4.1.3.1. การระบุคำให้อ่านแบบไทยโดยใช้ลักษณะสำคัญที่พิจารณาเพียงโทเค็นอย่างเดียว

4.1.3.2. การระบุคำให้อ่านแบบไทยโดยใช้ลักษณะสำคัญของโทเค็นที่สนใจและลักษณะสำคัญของบริบทของโทเค็นที่สนใจ โดยไม่รวมลักษณะสำคัญเอ็นแกรม

หลังจากที่ได้แบบจำลองต้นไม้ตัดสินใจสำหรับการทำนายผลแล้ว จึงนำเอาข้อมูลทดสอบมาทำการทดสอบ แล้วเปรียบเทียบประสิทธิภาพการทำนายผลระหว่างลักษณะสำคัญ 2 กลุ่ม ข้อมูลที่ใช้ในการทดสอบในการทดลองนี้เป็นข้อมูลชุดเดียวกับการทดลองที่ 1 คือ มีโทเค็นจำนวน 5,324 โทเค็น

4.1.4. การทดลองที่ 4: การประเมินประสิทธิภาพของแนวทางที่นำเสนอ

การทดลองที่ 4 เป็นการทดลองเพื่อประเมินประสิทธิภาพของแนวทางที่นำเสนอ คือ การระบุคำให้อ่านแบบไทยโดยใช้ต้นไม้ตัดสินใจจากลักษณะสำคัญเอ็นแกรม ดังที่กล่าวในหัวข้อ 3.1.3.2 ร่วมกับ ลักษณะสำคัญที่ได้จากต้นไม้ตัดสินใจทั้ง 2 กลุ่ม ได้แก่ ลักษณะสำคัญที่ได้จากการพิจารณาเฉพาะบริบท และ ลักษณะสำคัญที่ได้จากบริบทรอบข้างของโทเค็นที่สนใจ

นอกจากนี้ในการทดลองนี้ยังใช้การระบุคำให้อ่านแบบไทยโดยใช้แบบจำลองเอ็นแกรมโดยใช้ค่าความน่าจะเป็นในการเกิดอักขระของทั้ง 2 แบบจำลอง (กลุ่มของโทเค็นที่ให้อ่านแบบไทย และกลุ่มของโทเค็นที่ให้อ่านแบบอื่นๆ) เพื่อตัดสินใจ โดยผลรวมค่าความน่าจะเป็นของแบบจำลองใดที่มีค่าใกล้ศูนย์จะเป็นคำตอบของการจำแนก

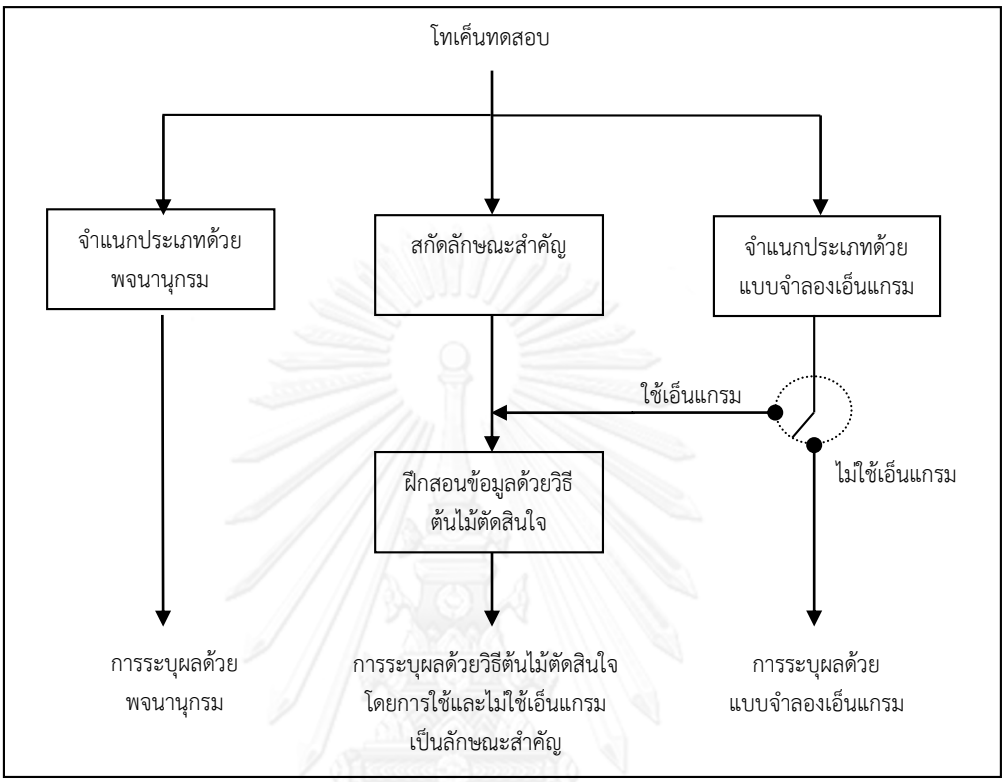
ในการทดลองนี้ทำการเปรียบเทียบประสิทธิภาพของแนวทางที่นำเสนอดังต่อไปนี้

4.1.4.1. การระบุคำให้อ่านแบบไทยด้วยระบบอ้างอิง โดยใช้พจนานุกรม

4.1.4.2. การระบุคำให้อ่านแบบไทยโดยใช้แบบจำลองต้นไม้ตัดสินใจ โดยใช้ลักษณะสำคัญของโทเค็นที่สนใจและลักษณะสำคัญของบริบทของโทเค็นที่สนใจ โดยไม่รวมลักษณะสำคัญเอ็นแกรม

4.1.4.3. การระบุคำให้อ่านแบบไทยโดยใช้แบบจำลองเอ็นแกรม โดยใช้ค่าความน่าจะเป็นที่ได้จากแบบจำลองเอ็นแกรม

4.1.4.4. การระบุคำให้อ่านแบบไทยโดยใช้แบบจำลองต้นไม้ตัดสินใจร่วมกับลักษณะสำคัญที่ได้จากแบบจำลองเอ็นแกรม



ภาพที่ 4-2 โครงสร้างของการทดลองที่ 4

จากภาพที่ 4-2 แสดงโครงสร้างของแนวทางในการดำเนินงานของการทดลองที่ 4 โดยการทดลองนี้ใช้ข้อมูลทดสอบจำนวน 5,324 โทเค็น เช่นเดียวกับการทดลองที่ 1 และ 3

ดังนั้นในงานวิจัยนี้จึงได้แบ่งการทดลองออกเป็น 4 แบบ ซึ่งมีวัตถุประสงค์ที่แตกต่างกัน คือ

1) เพื่อหาระบบอ้างอิงโดยพิจารณาจากการใช้พจนานุกรม
2) เพื่อเปรียบเทียบประสิทธิภาพการระบุคำให้อ่านแบบไทยจากการตัดสินใจของคนระหว่างการพิจารณาเฉพาะโทเค็น และการพิจารณาบริบทของโทเค็นร่วมด้วย
3) เพื่อเปรียบเทียบประสิทธิภาพของลักษณะสำคัญที่พิจารณาเฉพาะโทเค็นและลักษณะสำคัญที่ได้จากการพิจารณาเฉพาะโทเค็นและบริบทรอบข้างของโทเค็น โดยใช้แบบจำลองต้นไม้ตัดสินใจ
4) เพื่อประเมินประสิทธิภาพของแนวทางที่นำเสนอ คือ การใช้ลักษณะสำคัญเอ็นแกรมร่วมกับลักษณะสำคัญที่ได้จากการพิจารณาเฉพาะบริบท และ ลักษณะสำคัญที่ได้จากบริบทรอบข้างของโทเค็นที่สนใจ

โดยข้อมูลที่ใช้ในงานวิจัยแบ่งออกเป็น 2 ส่วน คือ ข้อมูลสำหรับฝึกสอนและข้อมูลสำหรับทดสอบ และการวัดผลในงานวิจัยนี้จะพิจารณาจากความถูกต้องของภาษาที่ได้จากระบบเปรียบเทียบกับภาษาที่ได้จากผู้ประเมิน

บทที่ 5
ผลการทดลองและวิเคราะห์ผลการทดลอง

ในบทนี้ได้นำเสนอผลการทดลองและการวิเคราะห์การทดลองทั้ง 4 แบบ ที่ได้จากบทที่ 4 ซึ่งมีเนื้อหาดังต่อไปนี้

1. ประเมินผลการหาระบบอ้างอิง
2. ประเมินผลการระบุคำให้อ่านแบบไทยโดยพิจารณาจากการตัดสินของคน
3. ประเมินผลการการเปรียบเทียบลักษณะสำคัญด้วยต้นไม้ตัดสินใจ
4. ประเมินผลประสิทธิภาพของแนวทางที่นำเสนอ

5.1.ประเมินผลการหาระบบอ้างอิง

ผลประเมินความแม่นยำของการระบุคำให้อ่านแบบไทยจากอักษรโรมันที่พบในเครือข่ายสังคม เพื่อทำการหาระบบอ้างอิง โดยใช้พจนานุกรมที่แตกต่างกัน 3 แบบ คือ LEXiTRON, CMU และ การใช้ทั้ง LEXiTRON และ CMU แสดงดังตารางที่ 5-1

จากตารางแสดงให้เห็นว่าการนำพจนานุกรม CMU มาใช้ให้ประสิทธิภาพการทำนายผลสูงกว่าการใช้พจนานุกรม LEXiTRON เนื่องจากพจนานุกรม CMU มีความครอบคลุมโทเค็นของข้อมูลมากกว่าพจนานุกรม LEXiTRON ดังแสดงในตารางที่ 3-3 (แสดงจำนวนโทเค็นของพจนานุกรม LEXiTRON และ CMU) และเมื่อนำพจนานุกรม LEXiTRON และ CMU มาใช้ร่วมกัน ส่งผลให้เปอร์เซนต์เฉลี่ยของความแม่นยำในการทำนายมีเปอร์เซนต์ที่สูงมากขึ้น เพราะเมื่อนำทั้งสองพจนานุกรมมารวมกันทำให้ครอบคลุมโทเค็นของข้อมูลที่ใช้ในการทดสอบมากกว่าการใช้เพียงพจนานุกรมเดียว ดังนั้นการเลือกระบบอ้างอิงจึงควรเลือกจากพจนานุกรมที่ให้เปอร์เซนต์ความแม่นยำในการทำนายคำให้อ่านแบบอื่นๆที่มีค่าสูง ซึ่งหมายความว่าผลเฉลี่ยของเปอร์เซนต์ความแม่นยำในการทำนายจะสูงตามด้วยเช่นกัน ด้วยเหตุนี้ผู้วิจัยจึงเลือกระบบอ้างอิงที่ใช้พจนานุกรม LEXiTRON และ CMU ร่วมกัน ซึ่งมีเปอร์เซนต์การทำนายคำที่ให้อ่านแบบอื่นๆสูงที่สุด คือ 87.15% และนอกจากนี้ยังมีค่าเฉลี่ยเปอร์เซนต์ความแม่นยำสูงที่สุด คือ 82.28%

นอกจากนี้จากการสังเกตลักษณะข้อมูลที่ใช้ในการวิจัยที่ได้จากข้อความเพื่อใช้ประกาศในเครือข่ายสังคมพบว่า คนส่วนมากเขียนข้อความด้วยโทเค็นที่ให้อ่านแบบอื่นๆมากกว่าโทเค็นที่ให้อ่านแบบไทยประมาณ 4-5 เท่า จากอัตราส่วนของโทเค็นที่ได้จากข้อมูลที่ใช้ในงานวิจัยนี้ แสดงให้เห็นว่าข้อมูลที่ได้จากโทเค็นที่ให้อ่านแบบอื่นๆมีจำนวนมากกว่าโทเค็นที่ให้อ่านแบบไทย ดังนั้นจึงเป็นเหตุผลสนับสนุนว่าเราควรจะเลือกใช้พจนานุกรมที่สามารถทำนายโทเค็นที่ให้อ่านแบบอื่นๆได้สูงกว่าโทเค็นที่ให้อ่านแบบไทย นั่นคือ การเลือกใช้พจนานุกรม LEXiTRON และ CMU มาใช้ร่วมกัน

ตารางที่ 5-1 ค่าเปอร์เซนต์ความแม่นยำของระบบอ้างอิง

พจนานุกรม	เปอร์เซนต์ความแม่นยำ		
	POS	NEG	ค่าเฉลี่ย
LEXiTRON	86.84%	75.64%	77.74%
CMU	61.74%	86.43%	81.81%
LEXiTRON/CMU	61.14%	87.15%	82.28%

5.2.ประเมินผลการระบุคำให้อ่านแบบไทยโดยพิจารณาจากการตัดสินของคน

ผลการจำแนกภาษาของโทเค็นของคนไทยจำนวน 6 คน กับโทเค็นจำนวน 200 โทเค็น จาก 175 ข้อความที่สุ่มจากข้อความในข้อมูลทดสอบทั้งหมด เพื่อพิจารณาความสำคัญของการปรากฏและไม่ปรากฏของบริบทรอบข้างของโทเค็นที่สนใจแสดงดังตารางที่ 5-2 และ 5-3

จากตารางพบว่าการจำแนกภาษาโดยพิจารณาบริบทของโทเค็นร่วมด้วยให้เปอร์เซนต์ความแม่นยำที่ดีกว่าการพิจารณาเฉพาะโทเค็นที่สนใจอย่างเดียว โดยเปอร์เซนต์ความแม่นยำเฉลี่ยของทั้ง 2 การทดสอบอยู่ที่ 99.41% และ 91.60% ตามลำดับ

ตารางที่ 5-2 ผลการจำแนกภาษาจากการตัดสินใจของคนโดยพิจารณาเฉพาะโทเค็น

ผู้ทดสอบ	เปอร์เซนต์ความแม่นยำ		
	POS	NEG	ค่าเฉลี่ย
1	78.00%	99.00%	88.50%
2	83.00%	98.00%	90.50%
3	82.00%	97.00%	89.50%
4	90.00%	93.00%	91.50%
5	93.00%	95.00%	95.00%
6	92.00%	92.00%	92.00%
ผลเฉลี่ย	86.63%	95.66%	91.60%

จากการวิเคราะห์ผลการจำแนกภาษาจากการตัดสินใจของคนโดยพิจารณาเฉพาะโทเค็น พบว่ากลุ่มของโทเค็นส่วนมากที่จำแนกผิดเป็นกลุ่มของโทเค็นที่ให้อ่านแบบไทย โดยเปอร์เซนต์ความแม่นยำของการจำแนกภาษามีค่าเท่ากับ 86.63% ในขณะที่เมื่อทำการจำแนกภาษาโดยพิจารณาบริบทรอบข้างของโทเค็นร่วมด้วย ส่งผลให้ประสิทธิภาพของการจำแนกโทเค็นที่ให้อ่านแบบไทยมี

เปอร์เซนต์ความแม่นยำสูงขึ้นเมื่อเปรียบเทียบกับการพิจารณาเฉพาะโทเค็นเพียงอย่างเดียว โดยผู้ทดสอบสามารถทำนายโทเค็นที่ให้อ่านแบบไทยได้สูงขึ้นถึง 12% นอกจากนี้การจำแนกโดยพิจารณาบริบทของโทเค็นร่วมด้วยยังส่งผลให้เปอร์เซ็นต์ความแม่นยำของโทเค็นที่ให้อ่านแบบอื่นๆ และค่าเฉลี่ยในการจำแนกโทเค็นของการทดสอบมีค่าสูงขึ้นอีกด้วย

ตารางที่ 5-3 ผลการจำแนกภาษาจากการตัดสินใจของคนโดยพิจารณาบริบทของโทเค็นร่วมด้วย

ผู้ทดสอบ	**เปอร์เซนต์ความแม่นยำ**		
	POS	NEG	ค่าเฉลี่ย
1	98.00%	100.00%	99.00%
2	99.00%	100.00%	99.50%
3	99.00%	100.00%	99.50%
4	98.00%	100.00%	99.00%
5	99.00%	100.00%	99.50%
6	100.00%	100.00%	100.00%
ผลเฉลี่ย	98.83%	100.00%	99.41%

จากการทดลองนี้แสดงให้เห็นว่าการพิจารณาบริบทส่งผลต่อความแม่นยำในการจำแนกประเภทของโทเค็น ดังนั้นถ้านำเอาการพิจารณาบริบทรอบข้างมาพิจารณาเป็นลักษณะสำคัญร่วมด้วยย่อมให้ประสิทธิภาพที่ดีกว่าการพิจารณาลักษณะสำคัญที่มีเฉพาะโทเค็นที่สนใจเพียงอย่างเดียว

5.3.ประเมินผลการการเปรียบเทียบลักษณะสำคัญด้วยต้นไม้ตัดสินใจ

จากการทดลองที่ 2 ที่ใช้การตัดสินใจจากคนได้แสดงว่าบริบทรอบข้างส่งผลต่อความแม่นยำในการระบุภาษา ดังนั้นในการทดลองนี้ผู้วิจัยจึงทำการทดลองด้วยต้นไม้ตัดสินใจ เพื่อเปรียบเทียบประสิทธิภาพของลักษณะสำคัญที่พิจารณาเฉพาะโทเค็นและลักษณะสำคัญของโทเค็นที่สนใจและลักษณะสำคัญของบริบทของโทเค็นที่สนใจ โดยไม่รวมลักษณะสำคัญเอ็นแกรม

จากผลการทดลองจากการแบ่งการพิจารณาลักษณะสำคัญออกเป็น 2 กลุ่ม คือ พิจารณาลักษณะสำคัญเฉพาะโทเค็นที่สนใจ และพิจารณาลักษณะสำคัญของโทเค็นที่สนใจและลักษณะสำคัญของบริบทของโทเค็นที่สนใจ โดยไม่รวมลักษณะสำคัญเอ็นแกรม ผลการทดลองเป็นดังที่คาดการณ์ไว้ คือ การใช้ลักษณะสำคัญของโทเค็นที่สนใจและลักษณะสำคัญของบริบทของโทเค็นที่สนใจ โดยไม่รวมลักษณะสำคัญเอ็นแกรม ให้ความแม่นยำในการจำแนกที่ดีกว่าการใช้เฉพาะลักษณะสำคัญของโทเค็นที่สนใจโดยไม่พิจารณาบริบทเช่นเดียวกับการพิจารณาของคนในการทดลองที่ 2 โดยเปอร์เซนต์เฉลี่ยของการพิจารณาเฉพาะโทเค็นและเปอร์เซนต์เฉลี่ยของการพิจารณาโทเค็นและบริบทของโทเค็นร่วม

โดยไม่รวมลักษณะสำคัญเอ็นแกรมมีค่าเท่ากับ 86.34% และ 87.94% ตามลำดับ ผลการทดลองแสดงดังตารางที่ 5-4

ถึงแม้ว่าการพิจารณาลักษณะสำคัญเฉพาะโทเค็นที่สนใจจะให้ประสิทธิภาพในการจำแนกโทเค็นที่ให้อ่านแบบอื่นๆได้สูงใกล้เคียงกับการพิจารณาด้วยลักษณะสำคัญของโทเค็นที่สนใจและลักษณะสำคัญของบริบทของโทเค็นที่สนใจ โดยไม่รวมลักษณะสำคัญเอ็นแกรม แต่อย่างไรก็ตามการเพิ่มลักษณะสำคัญที่พิจารณาบริบทรอบข้างของโทเค็นร่วมด้วยช่วยส่งผลให้ระบบสามารถจำแนกประเภทของโทเค็นที่ให้อ่านแบบไทยได้ดีมากขึ้น เมื่อทำการเปรียบเทียบกับการไม่พิจารณาบริบทรอบข้าง ซึ่งแนวทางนี้สามารถลดอัตราความผิดพลาดในการจำแนกได้ประมาณ 31% เมื่อเปรียบเทียบกับการใช้พจนานุกรมที่เป็นระบบอ้างอิง และจากการสังเกตต้นไม้ตัดสินใจที่พิจารณาลักษณะสำคัญของโทเค็นที่สนใจและลักษณะสำคัญของบริบทของโทเค็นที่สนใจ โดยไม่รวมลักษณะสำคัญเอ็นแกรม ลักษณะสำคัญที่เป็นรากโหนด คือ ลักษณะสำคัญที่ได้จาก Meaning: LEXiTRON ซึ่งแสดงให้เห็นว่าค่าลักษณะสำคัญนี้มีความสามารถในการจำแนกได้ดีกว่าค่าลักษณะสำคัญแบบอื่น

ตารางที่ 5-4 เปอร์เซนต์ความแม่นยำในการจำแนกด้วยต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ	**เปอร์เซนต์ความแม่นยำ**		
	POS	NEG	ค่าเฉลี่ย
พิจารณาเฉพาะโทเค็น	83.63%	86.96%	86.34%
พิจารณาโทเค็นและบริบท	87.75%	87.98%	87.94%

5.4.ประเมินประสิทธิภาพของแนวทางที่นำเสนอ

เนื่องจากผลที่ได้จากการทดลองที่ 2 และ 3 ได้ยืนยันว่าบริบทรอบข้างของโทเค็นมีผลต่อการจำแนกภาษาของคำ ดังนั้นแนวทางที่ผู้วิจัยได้นำเสนอจึงใช้ลักษณะสำคัญที่มีการพิจารณาบริบทรอบข้างร่วมกับการพิจารณาลักษณะสำคัญของโทเค็นที่สนใจด้วย อีกทั้งยังใช้ผลการระบุภาษาที่ได้จากแบบจำลองเอ็นแกรมมาเป็นลักษณะสำคัญอีกอย่างหนึ่ง

ผลการเปรียบเทียบความแม่นยำในการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคมด้วยวิธีต่างๆ คือ พจนานุกรม, ต้นไม้ตัดสินใจ, แบบจำลองเอ็นแกรม และแนวทางที่นำเสนอ คือ การใช้แบบจำลองต้นไม้ตัดสินใจร่วมกับลักษณะสำคัญที่ได้จากแบบจำลองเอ็นแกรม แสดงในตารางที่ 5-5

จากผลการทดสอบทั้ง 4 แบบ พบว่าการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันด้วยลักษณะสำคัญที่ได้จากแบบจำลองเอ็นแกรมและต้นไม้ตัดสินใจให้เปอร์เซนต์ความแม่นยำเฉลี่ยที่สูงที่สุด โดยแนวทางที่นำเสนอมีเปอร์เซนต์ความแม่นยำเฉลี่ยเท่ากับ 90.49% ในขณะที่ต้นไม้ตัดสินใจ, ระบบอ้างอิงและแบบจำลองเอ็นแกรมมีเปอร์เซนต์ความแม่นยำเฉลี่ย 87.94%, 82.28% และ 79.30% ตามลำดับ

ตารางที่ 5-5 ผลการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคม

วิธี	เปอร์เซนต์ความแม่นยำ		
	POS	NEG	ค่าเฉลี่ย
พจนานุกรม (ระบบอ้างอิง)	61.14%	87.15%	82.28%
ต้นไม้ตัดสินใจ	87.75%	87.98%	87.94%
แบบจำลองเอ็นแกรม	85.74%	77.81%	79.30%
ต้นไม้ตัดสินใจร่วมกับค่าลักษณะสำคัญเอ็นแกรม	84.63%	91.84%	90.49%

จากผลของตารางที่ 5-5 พบว่าผลการทดลองด้วยต้นไม้ตัดสินใจร่วมกับค่าลักษณะสำคัญเอ็นแกรมมีเปอร์เซนต์ความแม่นยำเฉลี่ยเท่ากับ 90.49% ซึ่งการใช้ต้นไม้ตัดสินใจร่วมกับค่าลักษณะสำคัญเอ็นแกรมสามารถลดอัตราการทำนายผลที่ผิดพลาดได้ 46.34% เมื่อเทียบกับการใช้พจนานุกรมที่เป็นระบบอ้างอิงของงานวิจัยนี้ และการพิจารณาบริบทรอบข้างของโทเค็นที่สนใจช่วยเพิ่มอัตราความแม่นยำให้ระบบได้สูงยิ่งขึ้น ถึงแม้ว่าผลการทดลองจากแบบจำลองเอ็นแกรมให้ประสิทธิภาพในการจำแนกชุดข้อมูลทดสอบได้เปอร์เซนต์เฉลี่ยต่ำที่สุด แต่อย่างไรก็ตามเมื่อนำผลการทำนายจากแบบจำลองเอ็นแกรมไปใช้เป็นลักษณะสำคัญอีกอย่างหนึ่งร่วมกับลักษณะสำคัญทั้ง 2 กลุ่มด้วยต้นไม้ตัดสินใจ ผู้วิจัยพบว่าระบบสามารถทำการจำแนกภาษาของโทเค็นได้สูงที่สุดเมื่อเปรียบเทียบกับแนวทางอื่นๆ และจากการสังเกตต้นไม้ตัดสินใจของแนวทางที่นำเสนอลักษณะสำคัญที่เป็นรากโหนดคือ ลักษณะสำคัญที่ได้จากแบบจำลองเอ็นแกรม ซึ่งแสดงให้เห็นว่าลักษณะสำคัญนี้มีความสามารถในการจำแนกได้ดีกว่าค่าลักษณะสำคัญแบบอื่น

ผู้วิจัยได้พิจารณาความสามารถของตัวจำแนกโดยใช้ค่าลักษณะสำคัญต่างๆบนชุดข้อมูลฝึกสอนพบว่า ผลการระบุคำให้อ่านแบบไทยโดยใช้ค่าลักษณะสำคัญที่ได้จากแบบจำลองเอ็นแกรมมีความแม่นยำสูงที่สุด เมื่อเปรียบเทียบกับลักษณะสำคัญที่ได้จากการพิจารณาด้วยพจนานุกรม LEXiTRON และ CMU ดังแสดงในตารางที่ 5-6 (เนื่องจากค่าลักษณะสำคัญทั้ง 3 ค่านี้สามารถใช้ในการระบุภาษาของโทเค็นได้โดยตรง ซึ่งแตกต่างจากค่าลักษณะสำคัญอื่นๆที่ไม่สามารถบอกคำตอบของภาษาได้) จึงเป็นเหตุผลสนับสนุนว่าลักษณะสำคัญที่ได้จากแบบจำลองเอ็นแกรมมีความสำคัญสูงที่สุดและถูกใช้เป็นรากโหนด

ตารางที่ 5-6 ผลการระบุคำให้อ่านแบบไทยด้วยชุดข้อมูลฝึกสอน

ลักษณะสำคัญ	เปอร์เซนต์ความแม่นยำ		
	POS	NEG	ค่าเฉลี่ย
LEXiTRON	88.49%	73.27%	80.88%
CMU	58.19%	82.80%	70.49%
N-gram	90.85%	81.60%	86.23%

5.5.วิเคราะห์ข้อผิดพลาดของการระบุคำให้อ่านแบบไทยจากแนวทางที่นำเสนอ

ในงานวิจัยนี้เสนอแนวทางการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันโดยใช้แบบจำลองต้นไม้ตัดสินใจร่วมกับลักษณะสำคัญที่ได้จากแบบจำลองเอ็นแกรม โดยมีเปอร์เซนต์ความแม่นยำเฉลี่ย 90.49% ซึ่งจำนวนโทเค็นและเปอร์เซ็นต์ที่จำแนกผิดแสดงดังตารางที่ 5-7

ตารางที่ 5-7 จำนวนโทเค็นและเปอร์เซนต์ของความผิดพลาดจากการระบุคำให้อ่านแบบไทย

จำนวน	POS	NEG	รวม
โทเค็นที่พิจารณา	996	4,328	5,324
โทเค็นที่จำแนกผิด	153	353	506
เปอร์เซนต์ที่จำแนกผิด	15.36%	8.15%	9.50%

จากแนวทางที่นำเสนอสามารถนำไปประยุกต์ใช้เพื่อเพิ่มประสิทธิภาพในการอ่านข้อความจากเว็บไซต์เครือข่ายสังคมให้กับระบบสังเคราะห์เสียงได้ แต่อย่างไรก็ตามเมื่อวิเคราะห์ความผิดพลาดที่เกิดจากการการระบุคำให้อ่านแบบไทย มีเพียง 9.50% ซึ่งความผิดพลาดของผลการระบุสามารถแบ่งออกเป็น 2 กลุ่มใหญ่ดังแสดงในตารางที่ 5-8

1) กลุ่มที่โทเค็นให้อ่านแบบไทย แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบอื่นๆ จำนวน 153 โทเค็น คิดเป็น 15.36% โดยแบ่งเป็นความผิดพลาดที่เกิดจาก 3 กรณีย่อย ได้แก่ ลักษณะสำคัญเอ็นแกรมทำนายผิด พจนานุกรมทำนายผิด และลักษณะสำคัญต่างๆ (ที่ไม่ใช่ค่าลักษณะสำคัญเอ็นแกรม และลักษณะสำคัญที่ใช้พจนานุกรม) ทำนายผิด แต่พบว่าสองกรณีย่อยที่ทำให้เกิดความผิดพลาดสูง ได้แก่ ลักษณะสำคัญเอ็นแกรมทำนายผิด 64 โทเค็นคิดเป็น 41.83% และ พจนานุกรมทำนายผิด 135 โทเค็น คิดเป็น 88.23%
2) กลุ่มที่โทเค็นให้อ่านแบบอื่นๆ แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบไทย จำนวน 353 โทเค็น คิดเป็น 8.15 % โดยแบ่งเป็นความผิดพลาดที่เกิดจาก 3 กรณีย่อย

ได้แก่ ลักษณะสำคัญเอ็นแกรมทำนายผิด พจนานุกรมทำนายผิด และลักษณะสำคัญต่างๆ (ที่ไม่ใช่ค่าลักษณะสำคัญเอ็นแกรม และลักษณะสำคัญที่ใช้พจนานุกรม) ทำนายผิด แต่พบว่าสองกรณีย่อยที่ทำให้เกิดความผิดพลาดสูง ได้แก่ ลักษณะสำคัญเอ็นแกรมทำนายผิด 202 โทเค็นคิดเป็น 57.22% และพจนานุกรมทำนายผิด 268 โทเค็น คิดเป็น 75.92%

ตารางที่ 5-8 จำนวนโทเค็นและเปอร์เซนต์ของความผิดพลาดแยกตามลักษณะสำคัญย่อย

ประเภท	ลักษณะสำคัญ	จำนวนโทเค็น	เปอร์เซนต์
โทเค็นที่ให้อ่านแบบไทยแต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบอื่นๆ	ลักษณะสำคัญเอ็นแกรมทำนายผิด	64	41.83%
	พจนานุกรมทำนายผิด	135	88.23%
โทเค็นที่ให้อ่านแบบอื่นๆแต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบไทย	ลักษณะสำคัญเอ็นแกรมทำนายผิด	202	57.22%
	พจนานุกรมทำนายผิด	268	75.92%

จากการพิจารณาความผิดพลาดของกรณีแรก คือ กลุ่มที่โทเค็นให้อ่านแบบไทย แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบอื่นๆ จำนวน 153 โทเค็น สามารถวิเคราะห์ความผิดพลาดโดยแบ่งเป็นกรณีย่อยได้ 4 กรณีดังตารางที่ 5-9 ได้แก่ ความผิดพลาดที่เกิดจากทั้งลักษณะสำคัญ เอ็นแกรม และพจนานุกรมทำนายผิดร่วมกัน 50 โทเค็น คิดเป็น 32.67% ความผิดพลาดที่เกิดจากลักษณะสำคัญเอ็นแกรมและลักษณะสำคัญต่างๆทำนายผิด 14 โทเค็น คิดเป็น 9.15% ความผิดพลาดที่เกิดจากพจนานุกรมและลักษณะสำคัญต่างๆทำนายผิด 85 โทเค็น คิดป็น 55.56% และความผิดพลาดจากลักษณะสำคัญต่างๆ (ลักษณะสำคัญเอ็นแกรมและพจนานุกรมทำนายถูก) ทำนายผิด 4 โทเค็น คิดเป็น 2.61%

ทั้งนี้จากความผิดพลาดของการระบุคำให้อ่านแบบไทยพบว่าความผิดพลาดที่เกิดจากการใช้พจนานุกรมมีความผิดพลาดมากที่สุด และความผิดพลาดจากการใช้ลักษณะสำคัญต่างๆมีความผิดพลาดน้อยกว่าความผิดพลาดของการใช้ลักษณะสำคัญต่างๆร่วมกับลักษณะสำคัญเอ็นแกรม หรือร่วมกับพจนานุกรม ซึ่งความผิดพลาดที่น้อยกว่าของการใช้ลักษณะสำคัญต่างๆแบบไม่ใช้ร่วมกับลักษณะสำคัญเอ็นแกรม หรือไม่ใช้ร่วมกับพจนานุกรม สะท้อนว่าความสำคัญของการใช้ลักษณะสำคัญแบบต่างๆมีต่ำกว่าการใช้ลักษณะสำคัญเอ็นแกรม หรือใช้พจนานุกรม

นอกจากนี้ผู้วิจัยได้สังเกตข้อผิดพลาดของกลุ่มที่โทเค็นให้อ่านแบบไทย แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบอื่นๆ และยกตัวอย่างของความผิดพลาด ดังนี้

1) ตัวอย่างของกลุ่มที่โทเค็นให้อ่านแบบไทย แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบอื่นๆ ที่เกิดจากความผิดพลาดของการใช้พจนานุกรม เพราะ เป็นโทเค็นที่ให้อ่าน

แบบไทยแต่มีปรากฏในพจนานุกรม เช่น pen (จากประโยค pen ngai kub) เป็นต้น

2) ตัวอย่างของโทเค็นที่ได้จากการใช้พจนานุกรมแต่ระบบจำแนกผิด โดยเป็นโทเค็นที่ถูกละอะพอสทรอฟี เช่น P’ (ถูกละเป็น P) ที่อ่านว่า พี่ หรือ N’ (ถูกละเป็น N) ที่อ่านว่า น้อง เนื่องจากโทเค็นเหล่านี้ใช้อะพอสทรอฟีเป็นลักษณะสำคัญเพื่อใช้เป็นตัวย่อในการอ่านบางโทเค็นที่เป็นภาษาไทยในเครือข่ายสังคม ซึ่งเมื่อมีการละอะพอสทรอฟี ทำให้โทเค็นนั้นไปตรงกับอักษรภาษาอังกฤษทำให้ระบบจำแนกผิด

ตารางที่ 5-9 จำนวนโทเค็นและเปอร์เซนต์ของโทเค็นที่ให้อ่านแบบไทยแต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบอื่นๆแยกตามกลุ่มลักษณะสำคัญที่ทำนายผิด

ลักษณะสำคัญ	จำนวนโทเค็น	เปอร์เซนต์
ลักษณะสำคัญเอ็นแกรมและพจนานุกรมทำนายผิด	50	32.67%
ลักษณะสำคัญเอ็นแกรมและลักษณะสำคัญต่างๆทำนายผิด	14	9.15%
พจนานุกรมและลักษณะสำคัญต่างๆทำนายผิด	85	55.56%
ลักษณะสำคัญต่างๆทำนายผิด (ลักษณะสำคัญเอ็นแกรมและพจนานุกรมทำนายถูก)	4	2.61%

จากการพิจารณาความผิดพลาดของกรณีที่สอง คือ กลุ่มที่โทเค็นให้อ่านแบบอื่นๆ แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบไทย จำนวน 353 โทเค็น สามารถวิเคราะห์ความผิดพลาดโดยแบ่งเป็นกรณีย่อยได้ 4 กรณีดังตารางที่ 5-10 ได้แก่ ความผิดพลาดที่เกิดจากทั้งลักษณะสำคัญเอ็นแกรม และพจนานุกรมทำนายผิดร่วมกัน 134 โทเค็น คิดเป็น 37.96% ความผิดพลาดที่เกิดจากลักษณะสำคัญเอ็นแกรมและลักษณะสำคัญต่างๆทำนายผิด 68 โทเค็น คิดเป็น 19.26% ความผิดพลาดที่เกิดจากพจนานุกรมและลักษณะสำคัญต่างๆทำนายผิด 134 โทเค็น คิดปีน 37.96% และความผิดพลาดจากลักษณะสำคัญต่างๆ (ลักษณะสำคัญเอ็นแกรมและพจนานุกรมทำนายถูก) ทำนายผิด 17 โทเค็น คิดเป็น 4.81% ซึ่งเป็นการยืนยันว่า ผลกระทบจากการใช้ค่าลักษณะสำคัญต่างๆ มีความสำคัญในการทำนายต่ำกว่าการใช้ร่วมกับลักษณะสำคัญเอ็นแกรม หรือร่วมกับพจนานุกรม เนื่องจากการใช้ลักษณะสำคัญต่างๆให้ความผิดพลาดน้อยกว่าความผิดพลาดของการใช้ลักษณะสำคัญต่างๆ ร่วมกับลักษณะสำคัญเอ็นแกรม หรือร่วมกับพจนานุกรม

นอกจากนี้ผู้วิจัยได้สังเกตข้อผิดพลาดของกลุ่มที่โทเค็นให้อ่านแบบอื่นๆ แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบไทย และยกตัวอย่างของความผิดพลาด ดังนี้

1) ตัวอย่างของกลุ่มที่โทเค็นให้อ่านแบบอื่นๆ แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบไทยเนื่องจากไม่มีปรากฏในพจนานุกรม เช่น littlebody, cinderalla, boyyy เป็นต้น

2) โทเค็นที่มีอักษรตัวแรกเป็นตัวใหญ่ เช่น Britney, Art เป็นต้น ซึ่งข้อผิดพลาดเหล่านี้อาจจะเกิดขึ้นเนื่องจากข้อมูลที่ใช้ในการฝึกสอนรวบรวมมาจากข้อความบนเครือข่ายสังคมของคนไทย ซึ่งมีชื่อเฉพาะของคนไทยที่แสดงการขึ้นต้นอักษรด้วยตัวใหญ่จำนวนมาก ทำให้ตัวจำแนกเกิดการเรียนรู้ที่ผิด

ตารางที่ 5-10 จำนวนโทเค็นและเปอร์เซนต์ของโทเค็นที่ให้อ่านแบบอื่นๆแต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบไทยแยกตามกลุ่มลักษณะสำคัญที่ทำนายผิด

ลักษณะสำคัญ	**จำนวนโทเค็น**	**เปอร์เซนต์**
ลักษณะสำคัญเอ็นแกรมและพจนานุกรมทำนายผิด	134	37.96%
ลักษณะสำคัญเอ็นแกรมและลักษณะสำคัญต่างๆทำนายผิด	68	19.26%
พจนานุกรมและลักษณะสำคัญต่างๆทำนายผิด	134	37.96%
ลักษณะสำคัญต่างๆทำนายผิด (ลักษณะสำคัญเอ็นแกรมและพจนานุกรมทำนายทำนายถูก)	17	4.81%

บทที่ 6
สรุปผลการวิจัย และข้อเสนอแนะ

6.1.สรุปผลการวิจัย

วิทยานิพนธ์ฉบับนี้ได้นำเสนอแนวทางในการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคมด้วยวิธีการทางสถิติ คือ การใช้แบบจำลองต้นไม้ตัดสินใจร่วมกับลักษณะสำคัญที่ได้จากแบบจำลองเอ็นแกรม โดยแนวทางนี้เป็นแนวทางที่ให้ผลการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันได้ดีที่สุดเมื่อเปรียบเทียบกับวิธีอื่นๆในงานวิจัย นอกจากนี้ยังได้นำลักษณะสำคัญที่พิจารณาเฉพาะโทเค็นที่สนใจและลักษณะสำคัญที่พิจารณาจากบริบทรอบข้างของโทเค็นที่สนใจมาใช้ร่วมกัน โดยเมื่อนำลักษณะสำคัญที่ได้จากการพิจารณาบริบทรอบข้างของโทเค็นที่สนใจมาใช้ ทำให้ระบบสามารถเพิ่มอัตราความแม่นยำได้สูงยิ่งขึ้นเมื่อเทียบกับการพิจารณาลักษณะสำคัญเฉพาะโทเค็นที่สนใจเพียงอย่างเดียว จากแนวทางที่นำเสนอสามารถนำกระบวนการระบุคำให้อ่านแบบไทยจากข้อความอักษรโรมันที่พบในเครือข่ายสังคมไปประยุกต์ใช้เป็นโมดูลย่อยของระบบการแปลงภาษาเขียนเป็นภาษาพูด (Text-to-Speech, TTS) เพื่อให้ได้เสียงอ่านของโทเค็นที่เป็นข้อความอักษรโรมันที่ถูกต้อง ซึ่งจะเป็นประโยชน์ต่อการพัฒนาระบบแปลงภาษาเขียนเป็นภาษาพูดให้มีประสิทธิภาพสูงขึ้นได้

6.2.อภิปรายผลการวิจัย

จากแนวทางที่ได้นำเสนอในงานวิจัยนี้ คือ การใช้แบบจำลองต้นไม้ตัดสินใจร่วมกับลักษณะสำคัญที่ได้จากแบบจำลองเอ็นแกรม โดยเปอร์เซนต์ความแม่นยำในการระบุคำให้อ่านแบบไทยจากข้อความที่พบในเครือข่ายสังคมที่ได้มีค่าเป็น 90.49% ซึ่งมีค่าสูงที่สุดเมื่อเปรียบเทียบกับวิธีอื่นๆ คือ พจนานุกรม, ต้นไม้ตัดสินใจ (ใช้ลักษณะสำคัญของโทเค็นที่สนใจและลักษณะสำคัญของบริบทของโทเค็นที่สนใจ โดยไม่รวมลักษณะสำคัญเอ็นแกรม) และแบบจำลองเอ็นแกรม โดยเมื่อนำผลการทำนายด้วยแบบจำลองเอ็นแกรมมาเป็นลักษณะสำคัญช่วยทำให้ระบบมีประสิทธิภาพในการระบุคำให้อ่านแบบไทยได้สูงยิ่งขึ้นเมื่อเปรียบเทียบกับการใช้ต้นไม้ตัดสินใจเพียงอย่างเดียว ทั้งนี้ลักษณะสำคัญที่ได้จากผลการทำนายด้วยแบบจำลองเอ็นแกรมยังเป็นลักษณะสำคัญที่มีความสำคัญที่สุด สังเกตได้จากลักษณะสำคัญนี้ถูกนำมาใช้เป็นรากโหนดในแบบจำลองการระบุคำให้อ่านแบบไทย นอกจากนี้การนำลักษณะสำคัญที่พิจารณาบริบทรอบข้างมาใช้ร่วมด้วยช่วยทำให้ระบบมีประสิทธิภาพสูงกว่าการใช้ลักษณะสำคัญเฉพาะโทเค็นที่สนใจเพียงอย่างเดียว

เมื่อวิเคราะห์ความผิดพลาดของแนวทางที่นำเสนอ พบว่ามีเปอร์เซนต์ความผิดพลาดเป็น 9.50% โดยสามารถแบ่งได้เป็น 2 กลุ่มใหญ่ คือ กลุ่มที่โทเค็นให้อ่านแบบไทย แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบอื่นๆ และกลุ่มที่โทเค็นให้อ่านแบบอื่นๆ แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบไทย โดยจากทั้ง 2 กลุ่มนี้ พบว่ามี 2 กรณีย่อยที่ทำให้เกิดความผิดพลาดสูง ได้แก่ กรณีที่เกิดจากลักษณะสำคัญเอ็นแกรมทำนายผิด และกรณีที่เกิดจากพจนานุกรมทำนายผิด

นอกจากนี้เมื่อพิจารณาความผิดพลาดร่วมกันของลักษณะสำคัญโดยแบ่งเป็น 4 กรณี คือ ความผิดพลาดที่เกิดจากทั้งลักษณะสำคัญเอ็นแกรมและพจนานุกรมทำนายผิดร่วมกัน ความผิดพลาดที่เกิดจากลักษณะสำคัญเอ็นแกรมและลักษณะสำคัญต่างๆทำนายผิด ความผิดพลาดที่เกิดจากพจนานุกรมและลักษณะสำคัญต่างๆทำนายผิด และความผิดพลาดจากลักษณะสำคัญต่างๆ (ลักษณะสำคัญเอ็นแกรมและพจนานุกรมทำนายถูก) ทำนายผิด พบว่าผลกระทบจากการใช้ค่าลักษณะสำคัญต่างๆ มีความสำคัญในการทำนายต่ำกว่าการใช้ร่วมกับลักษณะสำคัญเอ็นแกรม หรือร่วมกับพจนานุกรม โดยตัวอย่างของข้อผิดพลาดของกลุ่มที่โทเค็นให้อ่านแบบไทย แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบอื่นๆ เช่น โทเค็นที่ให้อ่านแบบไทยแต่มีปรากฏในพจนานุกรม โทเค็นที่ถูกละอะพอสทรอฟี เป็นต้น และตัวอย่างของข้อผิดพลาดของกลุ่มที่โทเค็นให้อ่านแบบอื่นๆ แต่ระบบจำแนกเป็นโทเค็นที่ให้อ่านแบบไทย เช่น โทเค็นที่ไม่มีปรากฏในพจนานุกรม โทเค็นที่มีอักษรตัวแรกเป็นตัวใหญ่ เป็นต้น

6.3.ข้อเสนอแนะ

เนื่องจากรูปแบบการเขียนข้อความบนเครือข่ายสังคมมีความหลากหลาย ดังนั้นการเพิ่มจำนวนข้อมูลในข้อมูลฝึกสอนอาจช่วยสร้างแบบจำลองที่มีความครอบคลุมได้มากขึ้น ดังเช่นในกรณีของความผิดพลาดที่เกิดจากตัวอักษรตัวแรกเป็นตัวใหญ่ มีสาเหตุมากจากชื่อเฉพาะของคนไทยที่แสดงบนเครือข่ายสังคมมีการขึ้นต้นอักษรด้วยตัวใหญ่จำนวนมาก นอกจากนี้การเพิ่มข้อมูลฝึกสอนอาจช่วยแก้ปัญหาจากกรณีของการละอะพอสทรอฟีได้อีกด้วย

สำหรับกรณีของโทเค็นที่มีการซ้ำตัวอักษร อาจจะทำการแก้โดยตัดอักษรที่เป็นตัวซ้ำออกให้เหลือเพียงตัวอักษรที่แท้จริง ซึ่งอาจจะช่วยลดความผิดพลาดจากการระบุภาษาของโทเค็นได้

นอกจากนี้ข้อผิดพลาดที่เกิดจากโทเค็นที่ให้อ่านแบบไทยแต่มีปรากฏในพจนานุกรม อาจจะแก้ข้อผิดพลาดนี้โดยการถอดโทเค็นโรมันเหล่านั้นให้เป็นโทเค็นที่เป็นอักษรไทยแล้วกลับไปตรวจสอบกับพจนานุกรมภาษาไทยอีกครั้ง

รายการอ้างอิง

1. Saychum, S., et al. *A bi-lingual Thai-English TTS system on Android mobile devices*. in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2012 9th International Conference on*. 2012.
2. Sangkavichitr, C., *A Design of Thai-English Transliterated Word Retrieval for Smart Phones*, in *International Conference on Systems and Electronic Engineering (ICSEE'2012)*. 2012.
3. Choong, C.Y., Y. Mikami, and C.A. Marasinghe, *Optimizing n-gram Order of an n-gram Based Language Identification Algorithm for 68 Written Languages.* International Journal on Advances in Ict for Emerging Regions (icter), 2009. **2**(2).
4. Chen, Y., et al., *Identifying Language Origin of Person Names With N-Grams of Different Units*, in *International Conference on Acoustics, Speech, and Signal Processing*. 2006.
5. Hakkinen, J. and J. Tian, *n-gram and decision tree based language identification for written words*, in *IEEE Workshop on Automatic Speech Recognition and Understanding*. 2001.
6. Han, J., M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2 ed. 2006, San Francisco: Morgan Kaufmann.
7. ก่อตระกูล, อ., การประมวลผลภาษามนุษย์ด้วยคอมพิวเตอร์ : เส้นทางสู่การพัฒนาระบบสารสนเทศอัจฉริยะ. 2549: หน่วยปฏิบัติการวิจัยเชี่ยวชาญเฉพาะการประมวลผลภาษาธรรมชาติและเทคโนโลยีสารสนเทศอัจฉริยะ (NAiST) ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์.
8. จีรยุทธ ไชยจารุวณิช, et al., การประยุกต์ใช้ดีซิชันทรีอินดักชันในการศึกษาพันธุศาสตร์ประชากร. NECTEC Technical Journal, November 2003 - October 2004. **4**.
9. เอกวงศ์อนันต์, อ., การระบุคำไทยและคำทับศัพท์ด้วยแบบจำลองเอ็นแกรม, in ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์. 2548, จุฬาลงกรณ์มหาวิทยาลัย: กรุงเทพฯ.
10. กิตติกูล, ช., การถอดคำแบบถ่ายเสียงสำหรับชื่อบุคคลภาษาไทยที่เขียนด้วยอักษรโรมัน, in ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์. 2554, จุฬาลงกรณ์มหาวิทยาลัย.
11. Nobesawa, S. and I. Tahara, *Language Identification for Person Names Based on Statistical Information*, in *Proceedings of the 19th Asia-Pacific Conference on Language, Information and Computation (PACLIC 2005)*.
12. Thomas, S. and A. Verma, *Language identification of person names using CF-IOF based weighing function*, in *Annual Conference of the International Speech Communication Association*. 2007. p. 1769-1772.

13. Bhargava, A. and G. Kondrak, *Language identification of names with SVMs*, in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2010, Association for Computational Linguistics: Los Angeles, California. p. 693-696.
14. NECTEC. *LEXiTRON data.* [cited 2012 14 September]; Available from: http://lexitron.nectec.or.th/2009_1.
15. University, C.M. *The Carnegie Mellon Pronouncing Dictionary.* [cited 2012 14 September]; Available from: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.
16. Clarkson, P. and R. Rosenfeld, *Statistical Language Modeling using the CMU-Cambridge toolkit*, in *Proceedings of the EUROSPEECH* 1997. p. 2707-2710.
17. Hall, M., et al., *The WEKA data mining software: an update.* Sigkdd Explorations, 2009. **11**(1): p. 10-18.
18. Quinlan, J.R., *C4.5: Programs for Machine Learning*. 1993.

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวนัทธมน โมกข์ณรงค์ เกิดเมื่อวันที่ 31 ธันวาคม พ.ศ.2532 ที่จังหวัดกรุงเทพมหานคร สำเร็จการศึกษาระดับปริญญาตรี หลักสูตรวิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยมหิดลในปี 2554 และเข้าศึกษาต่อระดับปริญญาโทในหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย