การเรียงลำดับคำในประโยคภาษาไทย โดยอาศัยแบบจำลองความน่าจะเป็นของประเภทคำกระตุ้นการพูด

นางสาววรษา ลิมปนดุษฎี

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทคัดย่อและแฟ้มข้อมูลฉบับเต็มของวิทยานิพนธ์ตั้งแต่ปีการศึกษา 2554 ที่ให้บริการในคลังปัญญาจุฬาฯ (CUIR)
เป็นแฟ้มข้อมูลของนิสิตเจ้าของวิทยานิพนธ์ ที่ส่งผ่านทางบัณฑิตวิทยาลัย
The abstract and full text of theses from the academic year 2011 in Chulalongkorn University Intellectual Repository (CUIR)
are the thesis authors' files submitted through the University Graduate School.

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2557
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

Ordering Words in Thai Sentences using Probabilistic Models of Verbalization Stimulating Word Categories

Miss Worasa Limpanadusadee

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 2014

Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	การเรียงลำดับคำในประโยคภาษาไทย โดยอาศัยแบบจำลองความน่าจะเป็นของประเภทคำกระตุ้นการพูด
โดย	นางสาววรษา ลิมปนดุษฎี
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต ดร.อรอินทรา ภู่ประเสริฐ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

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

คณะกรรมการสอบวิทยานิพนธ์

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

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ดร.อรอินทรา ภู่ประเสริฐ)

.....กรรมการภายนอกมหาวิทยาลัย
(ดร.ประกาศิต กายะสิทธิ์)

วรษา ลิมปนดุษฎี : การเรียงลำดับคำในประโยคภาษาไทย โดยอาศัยแบบจำลองความน่าจะเป็นของประเภทคำกระตุ้นการพูด (Ordering Words in Thai Sentences using Probabilistic Models of Verbalization Stimulating Word Categories) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.โปรดปราน บุณยพุกกณะ, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: รศ. ดร. อติวงศ์ สุชาโต, ดร.อรอินทรา ภู่ประเสริฐ, 100 หน้า.

ภาษาไทยเป็นภาษาที่ไม่มีการแบ่งวรรคตอนอย่างเป็นทางการ เป็นภาษาที่เป็นคำโดด คำในภาษาไทยจะไม่มีการเปลี่ยนแปลงรูปไม่ว่าจะอยู่ในกาล (Tense) การก (Case) มาลา (Mood) หรือวาจก (Voice) ดังนั้นการเรียงลำดับคำในประโยคภาษาไทยมีผลอย่างยิ่งต่อความหมายของประโยค ดังนั้น สำหรับผู้ที่มีปัญหาด้านการจัดเรียงคำในประโยคภาษาไทย เช่น บุคคลออทิสติก (Autistic) ที่มีความผิดปกติด้านภาษาและการสื่อสาร หรือผู้ที่มีความบกพร่องทางการเรียนรู้ (Learning disabilities) บางประเภทที่มีปัญหาด้านการเรียงลำดับข้อมูล จึงต้องการความช่วยเหลือในการจัดเรียงคำ ในงานวิจัยนี้จึงมุ่งเน้นที่จะเรียงลำดับคำในประโยคภาษาไทยได้อย่างอัตโนมัติ โดยนำเสนออัลกอริทึมที่อาศัยแบบจำลองความน่าจะเป็น (Probabilistic Models) ร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (Weighted Finite-State Transducer) และใช้ประเภทคำกระตุ้นการพูดในการพัฒนาอัลกอริทึมนี้ด้วย สำหรับการทดสอบเพื่อวัดประสิทธิภาพของอัลกอริทึมนี้ จะทดสอบโดยวัดความถูกต้องของการเรียงคำในประโยคตามหลักไวยากรณ์ภาษาไทยเปรียบเทียบกับประโยคเป้าหมายที่ถูกเลือกโดยผู้เชี่ยวชาญ มีประโยคทดสอบทั้งสิ้น 440 ประโยค ผลการทดสอบเมื่อป้อนคำเข้าสู่ระบบ 3 และ 4 คำ ระบบสามารถเลือกประโยคคำตอบได้ตรงตามประโยคเป้าหมาย 100% และ 80% ตามลำดับ อย่างไรก็ตามค่าเฉลี่ยในการเลือกประโยคคำตอบของระบบที่นำเสนออยู่ที่ 74% เมื่อมีการเพิ่มจำนวนคำที่ป้อนเข้าสู่ระบบตั้งแต่ 3-6 คำ

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2557 ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

5570507221 : MAJOR COMPUTER SCIENCE
KEYWORDS: NATURAL LANGUAGE GENERATION / LEARNING DISABILITIES / N-GRAM / WEIGHTED FINITE-STATE TRANSDUCER / PROBABILISTIC MODEL / AUTISTIC CHILDREN

WORASA LIMPANADUSADEE: Ordering Words in Thai Sentences using Probabilistic Models of Verbalization Stimulating Word Categories. ADVISOR: ASST. PROF. PROADPRAN PUNYABUKKANA, Ph.D., CO-ADVISOR: ASSOC. PROF. ATIWONG SUCHATO, Ph.D., ONINTRA POOBRASERT, Ph.D., 100 pp.

Word order is considered a challenge in Thai language as it lacks punctuation. Written Thai does not require sentence-end nor inter-word punctuations. However, spaces are added intuitively to separate sentences while words are written continuously without any mark. Moreover, Thai is an isolating language that most words are monosyllabic. Thai verb maintains its form and does not vary according to tense, case, mood, or voice. Consequently, the order of words has significant impact on the meaning. As a result, the nature of Thai language poses issues for some people with learning disabilities or people with autism who have difficulty in sequencing to appropriately arrange words into a sentence with an intent meaning. This research proposes an algorithm to arrange Thai words in simple Thai sentence. The algorithm was developed by employing the Probabilistic Model and Weighted Finite-State Transducer in conjunction with verbalization stimulating word categories. The performance of this algorithm was tested by measuring the accuracy in word sequencing according to Thai grammar comparing to the target sentence chosen by expert. Evaluated upon test sets of total 195 sentences, each sentence composes of 3-6 words with a specific Part-Of-Speech (POS) combination. The sentence generation accuracies were 100% and 80% for 3-word and 4-word sentences, respectively. The average accuracy was at 74% when longer sentences were also included.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Department:	Computer Engineering	Student's Signature .....
Field of Study:	Computer Science	Advisor's Signature .....
Academic Year:	2014	Co-Advisor's Signature .....
		Co-Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์ของ ผู้ช่วยศาสตราจารย์ ดร. โปรดปราน บุณยพุกกณะ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต และ ดร.อรอินทรา ภู่ประเสริฐ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ซึ่งท่านเหล่านี้ได้ให้ความรู้ คำปรึกษา การสนับสนุนและโอกาสที่ดีต่าง ๆ ในระหว่างระยะเวลาที่ศึกษาและดำเนินการวิจัย

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.เศรษฐา ปานงาม และดร.ประกาศิต กายะสิทธิ์ กรรมการสอบวิทยานิพนธ์ที่ให้คำแนะนำเพื่อใช้ในการปรับปรุงแก้ไขในการทำวิทยานิพนธ์ฉบับนี้

ขอขอบพระคุณ ดร.วันทนีย์ พันธชาติ ที่ให้คำแนะนำเทคนิคและวิธีการแบ่งหมวดหมู่ของคำ ซึ่งเป็นหนึ่งในวิธีที่ใช้ในงานวิจัยนี้

ขอขอบคุณเพื่อนร่วมงานทุกคนในห้องปฏิบัติการระบบภาษาพูดและห้องปฏิบัติการเทคโนโลยีช่วยเหลือผู้พิการ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ให้คำแนะนำและกำลังใจในการดำเนินงานเสมอมา

ขอขอบคุณครอบครัว ญาติพี่น้อง และเพื่อนทุกท่าน ที่คอยให้กำลังใจ เข้าใจและสนับสนุนการทำวิจัยในครั้งนี้มาโดยตลอด ทำให้การจัดทำวิทยานิพนธ์ในครั้งนี้ประสบความสำเร็จได้ด้วยดี

นอกจากนี้ขอขอบคุณโครงการทุนสถาบันบัณฑิตวิทยาศาสตร์และเทคโนโลยีไทย (TGIST) สำหรับเงินทุนในการทำวิจัยตลอดระยะเวลาที่ศึกษาและดำเนินการวิจัย

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง
บทคัดย่อภาษาอังกฤษ ..... จ
กิตติกรรมประกาศ ..... ฉ
สารบัญ ..... ช
สารบัญตาราง ..... ฎ
สารบัญภาพ ..... ฐ
บทที่ 1 บทนำ 1
1.1 ความเป็นมาและความสำคัญของปัญหา 1
1.2 วัตถุประสงค์ของการวิจัย 2
1.3 ขอบเขตของการวิจัย 2
1.4 ลำดับขั้นตอนในการทำวิจัย 3
1.5 ประโยชน์ที่คาดว่าจะได้รับ 4
1.6 ผลงานตีพิมพ์จากวิทยานิพนธ์ 4
1.7 ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์ 5
บทที่ 2 ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง 6
2.1 ทฤษฎีที่เกี่ยวข้อง 6
2.1.1 ทฤษฎีเกี่ยวกับการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) 6
2.1.2 ทฤษฎีเกี่ยวกับประเภทของคำ ส่วนประกอบของประโยคและประโยคภาษาไทย 7
2.1.3 ทฤษฎีแบบจำลองเอ็นแกรม (N-gram Model) 13
2.1.4 ทฤษฎีการปรับเรียบ (Smoothing) 14
2.1.5 ทฤษฎีตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (Weighted Finite-State Transduces: WFST) 15

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

หน้า

2.2 งานวิจัยที่เกี่ยวข้อง 17
บทที่ 3 ขั้นตอนการดำเนินงานวิจัยและการออกแบบระบบที่นำเสนอ 21
3.1 ขั้นตอนการดำเนินงานวิจัย 21
3.2 การออกแบบระบบที่นำเสนอ 21
3.2.1 ระบบการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอโดยอาศัยหลักการและลักษณะสำคัญเอ็นแกรม 22
3.2.2 ระบบการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอโดยอาศัยหลักการและลักษณะสำคัญเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (Weighted Finite-State Transducer: WFST) 23
บทที่ 4 การออกแบบและสร้างคลังข้อความ 24
4.1 การเตรียมและเลือกใช้คลังข้อความ 24
4.2 การออกแบบและสร้างคลังข้อความ TELL-S 25
4.3 การระบุหน้าที่ของคำในคลังข้อความ 26
4.3.1 การระบุหน้าที่ของคำตามแนวคิดการสร้างคลังข้อความ Royal-Institute-Based Thai-Part-of-Speech (RIPOS) 26
4.3.2 การระบุหน้าที่ของคำตามแนวคิดที่นำเสนอ 27
บทที่ 5 ขั้นตอนการสร้างระบบและอัลกอริทึมที่นำเสนอ 30
5.1 ขั้นตอนการสร้างระบบที่นำเสนอ 30
5.2 การสร้างแบบจำลองของหลักไวยากรณ์ภาษาไทยตามแนวคิดที่นำเสนอ 31
5.3 การสร้างแบบจำลองของภาษา 31
5.3.1 การสร้างแบบจำลองของคำ 32
5.3.2 การสร้างแบบจำลองของหมวดหมู่ย่อย 36
5.3.3 การสร้างแบบจำลองของหมวดหมู่หลัก 37
5.4 การสร้างระบบคำนวณคะแนนและอัลกอริทึมในการเลือกประโยคคำตอบที่นำเสนอ 38

5.4.1 ระบบคำนวณคะแนนและอัลกอริทึมในการเลือกประโยคคำตอบโดยอาศัยหลักการและลักษณะสำคัญของเอ็นแกรม 38

5.4.2 ระบบคำนวณคะแนนและอัลกอริทึมในการเลือกประโยคคำตอบโดยอาศัยหลักการและลักษณะสำคัญของเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำแบบแบบถ่วงน้ำหนัก (WFST) 41

บทที่ 6 การทดลองและวิธีการวัดผลการทดลอง 53

6.1 ข้อมูลที่ใช้ในการทดสอบระบบ 53

6.2 การกำหนดหมวดหมู่หลักและหมวดหมู่ย่อยของคำที่ใช้ในการทดสอบระบบตามระบบที่นำเสนอ 53

6.3 การกำหนดและระบุหน้าที่ของคำที่ใช้ในการทดสอบระบบตามระบบอ้างอิง (RIPOS) 62

6.4 การสร้างชุดทดสอบและประโยคทดสอบ 63

6.5 การวัดผลทดลอง 68

6.6 การทดลอง 69

บทที่ 7 ผลการทดลองและวิเคราะห์ผลการทดลอง 73

7.1 การประเมินผลการการทดลองระบบที่นำเสนอโดยอาศัยหลักการเอ็นแกรม 73

7.2 การประเมินผลการทดลองหาลักษณะสำคัญของการระบุหมวดหมู่ของคำในคลังข้อความ 82

7.3 การประเมินผลประสิทธิภาพของระบบอ้างอิงโดยอาศัยหลักการเอ็นแกรมร่วมกับวิธีระบุหน้าที่ของคำแบบ RIPOS 86

7.4 การประเมินผลประสิทธิภาพของแนวทางที่นำเสนอโดยอาศัยหลักการเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST) 89

บทที่ 8 สรุปผลการวิจัยและข้อเสนอแนะ 92

8.1 สรุปผลการวิจัย 92

8.2 อภิปรายผลการวิจัย 92

8.3 ข้อเสนอแนะ 94

หน้า

รายการอ้างอิง 95

ประวัติผู้เขียนวิทยานิพนธ์ 100

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญตาราง

ตารางที่		หน้า
4-1	จำนวนประโยคในคลังข้อความ TELL-S โดยแบ่งตามความยาวของประโยค.....	25
4-2	การระบุหน้าที่ของคำ (POS) ตามแนวคิดการสร้างคลังข้อความ RIPOS	27
4-3	หมวดหมู่หลักและหมวดหมู่ย่อยที่ใช้ระบุหน้าที่ของคำในประโยค	28
4-4	ตัวอย่างการระบุหน้าที่ของคำตามหมวดหมู่ต่าง ๆ	29
6-1	การกำหนดหมวดหมู่หลัก หมวดหมู่ย่อยของคำ และจำนวนคำในแต่ละหมวดหมู่ที่อยู่ในคลังข้อความ TELL-S (กลุ่มที่ 1) และคลังข้อความ BEST2010 (กลุ่มที่ 2).	54
6-2	การกำหนดหน้าที่ของคำ (POS) ที่อยู่ในคลังข้อความ TELL-S (กลุ่มที่ 1) และคลังข้อความ BEST2010 (กลุ่มที่ 2).....	57
6-3	การกำหนดหมวดหมู่หลัก หมวดหมู่ย่อยของคำ และจำนวนคำในแต่ละหมวดหมู่ของคำที่ปรากฏอยู่ทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 (กลุ่มที่ 3)	59
6-4	การกำหนดหน้าที่ของคำ (POS) ที่ปรากฏอยู่ทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 (กลุ่มที่ 3).....	61
6-5	การกำหนดหน้าที่ของคำตามระบบอ้างอิง (RIPOS)	63
6-6	ชุดทดสอบที่แบ่งตามจำนวนคำที่ป้อนเข้าสู่ระบบตั้งแต่ 3-6 คำ.....	63
6-7	จำนวนประโยคทดสอบที่ถูกเลือก	66
7-1	จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ TELL-S และสร้างประโยคทดสอบจากคำที่มาจากคลังข้อความ TELL-S (WS1)	73
7-2	จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ TELL-S และสร้างประโยคทดสอบจากคำที่มาจากคลังข้อความ BEST2010 (WS2).....	75

ตารางที่ หน้า

7-3 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ BEST2010และสร้างประโยคทดสอบจากคำที่มาจากคลังข้อความ TELL-S (WS1) 76

7-4 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ BEST2010 และสร้างประโยคทดสอบจากคำที่มาจากคลังข้อความ BEST2010 (WS2)..... 77

7-5 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ TELL-S และสร้างประโยคทดสอบจากคำที่ปรากฏทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 (WS3)..... 78

7-6 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ BEST2010 และสร้างประโยคทดสอบจากคำที่ปรากฏทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 (WS3)..... 80

7-7 ลักษณะประโยคคำตอบที่ระบบเลือกเมื่อพิจารณาลักษณะสำคัญของหมวดหมู่ต่าง ๆ และเลือกใช้ลักษณะสำคัญในกรณีที่แตกต่างกันบนคลังข้อมูลที่แตกต่างกัน..... 83

7-8 ผลการเปรียบเทียบการเลือกประโยคคำตอบของระบบเมื่อใช้วิธีการระบุคำตามที่งานวิจัยนี้เสนอและวิธีการระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS..... 87

7-9 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายโดยอาศัยเอ็นแกรมร่วมกับหลักการตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST)..... 89

7-10 การประเมินผลการเรียงลำดับคำในประโยคภาษาไทยโดยเปรียบเทียบระหว่างระบบที่อาศัยหลักการทำงานเอ็นแกรม (N-gram) เพียงอย่างเดียวกับระบบที่อาศัยหลักการทำงานของเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST)..... 91

สารบัญภาพ

ภาพที่		หน้า
2-1	ตัวอย่างของตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก	15
2-2	ตัวอย่างการใช้ฟังก์ชันประกอบของ (a),(b) ตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก และ (c) ตัวอย่างผลลัพธ์ของตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักที่ได้จากการใช้ฟังก์ชันประกอบ	16
3-1	โครงสร้างการทำงานของระบบที่นำเสนอเมื่ออาศัยหลักการและลักษณะสำคัญเอ็นแกรม.....	22
3-2	โครงสร้างการทำงานของระบบที่นำเสนอเมื่ออาศัยหลักการและลักษณะสำคัญเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก.....	23
4-1	ตัวอย่างข้อมูลในคลังข้อความ TELL-S ที่แปลงเป็นคำทับศัพท์ภาษาอังกฤษ.....	26
5-1	ลำดับขั้นตอนการสร้างระบบ	30
5-2	ตัวอย่างข้อมูลของคำสำหรับใช้ในโปรแกรม CMU Toolkit V.2	32
5-3	ตัวอย่างความถี่ของคำปรากฏอยู่ในคลังข้อความในรูปแบบไฟล์.wfreq.....	33
5-4	ตัวอย่างคำที่ปรากฏอยู่ในคลังข้อความแบบเรียงตามตัวอักษรในรูปแบบไฟล์.vocab	34
5-5	ตัวอย่างค่าความน่าจะเป็นของคำตั้งแต่ 1-แกรม 2-แกรม และ 3-แกรมตามลำดับ	35
5-6	ตัวอย่างข้อมูลของหมวดหมู่ย่อยสำหรับใช้ในโปรแกรม CMU Toolkit V.2	36
5-7	ตัวอย่างข้อมูลของหมวดหมู่หลักสำหรับใช้ในโปรแกรม CMU Toolkit V.2.....	37
5-8	ตัวอย่างค่าความน่าจะเป็นของหมวดหมู่หลัก 2-แกรม ในรูปแบบไฟล์.arpa	38
5-9	โครงสร้างการทำงานของระบบเมื่ออาศัยหลักการและลักษณะสำคัญเอ็นแกรม ..	39
5-10	ตัวอย่างไฟล์ค่าความน่าจะเป็นเอ็นแกรมในรูปแบบ.arpa (a) เมื่อป้อนเข้าสู่โปรแกรม OpenGrm NGram และไฟล์ผลลัพธ์ที่ได้จากโปรแกรม (b) และ (c).....	42
5-11	ตัวอย่างการจัดข้อมูลในไฟล์เข้ารหัสตัวเปลี่ยนแปรสถานะจำกัด (a) และสร้างไฟล์สัญลักษณ์เอาต์พุต (b) เพื่อให้เหมาะสมและพร้อมใช้งานในโปรแกรม OpenFst	43

ภาพที่ หน้า

5-12 ขั้นตอนการสร้างตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักแบบไบนารีไฟล์และแสดงผลออกมาในลักษณะของโครงข่าย 45

5-13 ตัวอย่างโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของคำ..... 46

5-14 ตัวอย่างโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของหมวดหมู่ย่อย..... 46

5-15 ตัวอย่างโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของหมวดหมู่หลัก 46

5-16 ตัวอย่างโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของ POS..... 46

5-17 ตัวอย่างการทำฟังก์ชันประกอบระหว่างโครงข่าย W (a) กับโครงข่าย S (b) และโครงข่าย WS (c) คือผลของการทำฟังก์ชันประกอบของโครงข่าย W และ S..... 48

5-18 การใช้ฟังก์ชันประกอบระหว่างโครงข่ายต่าง ๆ และขั้นตอนการแปลงอินพุตและเอาต์พุตของโครงข่าย WSCP..... 49

5-19 ตัวอย่างการจัดเก็บอินพุต เอาต์พุต และค่าน้ำหนักจากโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก WSCP ไว้ในตารางแฮช..... 50

5-20 โครงสร้างการทำงานของระบบที่นำเสนอ 51

6-1 ตัวอย่างประโยคทดสอบที่ถูกเลือกในชุดทดสอบที่ 3 ที่สร้างแบบพบกันหมดจากชุดคำที่มาจากคลังข้อความ BEST2010 (กลุ่มที่ 2)..... 65

6-2 ขั้นตอนการให้คะแนนประโยคคำตอบ 72

7-1 ผลการทดสอบการเรียงลำดับคำในประโยคภาษาไทยโดยอาศัยหลักการเอ็นแกรมร่วมกับวิธีระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความที่แตกต่างกันและทดสอบบนชุดข้อมูลที่แตกต่างกัน 81

7-2 การประเมินผลความถูกต้องเมื่อเปรียบเทียบประโยคคำตอบที่ระบบเลือกกับประโยคคำตอบที่ผู้เชี่ยวชาญ 3 คนเลือก..... 82

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ภาพที่ หน้า

7-3 คะแนนของลักษณะสำคัญที่ใช้ระบุหมวดหมู่ของคำบนคลังข้อความ เมื่อพิจารณาจำนวนคำที่ป้อนเข้าสู่ระบบตั้งแต่ 3-6 คำ..... 86

7-4 คะแนนรวมของการเลือกประโยคคำตอบโดยเปรียบเทียบระหว่างกรณีการระบุหมวดหมู่และหน้าที่ของคำตามที่งานวิจัยนี้นำเสนอ (กรณีที่ 1) และการระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS (กรณีที่ 2)..... 88

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ภาษาไทยเป็นภาษาที่ไม่มีการแบ่งวรรคตอนอย่างเป็นทางการ เป็นภาษาที่เป็นคำโดด คำในภาษาไทยจะไม่มีการเปลี่ยนแปลงรูปไม่ว่าจะอยู่ในกาล (Tense) การก (Case) มาลา (Mood) หรือวาจก (Voice) [1] สำหรับการลำดับคำในประโยคการเรียงคำในภาษาไทยนั้นมีความสำคัญอย่างยิ่งต่อความหมายของประโยค เพราะการเรียงลำดับคำผิดตำแหน่งความหมายก็จะเปลี่ยนไปด้วย โดยปกติภาษาไทยนิยมเรียงคำจากประธาน กริยา และกรรม ตัวอย่างเช่น หมากัดแมว และฉันตีเขา เป็นต้น [2] เมื่อสลับตำแหน่งคำของผู้กระทำและผู้ถูกกระทำ ความหมายจะแตกต่างกันโดยสิ้นเชิง ส่งผลให้ผู้ฟังหรือผู้อ่านเกิดความเข้าใจผิดได้ [3]

สำหรับบุคคลออทิสติกที่มีความผิดปกติด้านภาษาและการสื่อสารนั้น 50 เปอร์เซ็นต์ของคนกลุ่มโรคนี้ไม่สามารถใช้ภาษาพูดเพื่อการสื่อสารได้ [4] และบุคคลที่มีความบกพร่องทางการเรียนรู้ที่มีปัญหาด้านการเรียงลำดับข้อมูล โดยบุคคลกลุ่มนี้มีปัญหาในเรื่องของสามารถเรียงลำดับข้อมูลที่ตนเองต้องการสื่อสาร [5] ทำให้บุคคลทั้งสองกลุ่มนี้ มีปัญหาด้านการจัดเรียงคำ ข้อความ และข้อมูลต่าง ๆ ที่ต้องการสื่อสาร รวมไปถึงปัญหาด้านการเรียงลำดับคำในประโยคภาษาไทยด้วย

จากการศึกษาแนวคิดและเทคนิคการสอนบุคคลออทิสติกและบุคคลที่มีความบกพร่องทางการเรียนรู้พบว่า การสอนบุคคลออทิสติกโดยมีการนำคำหรือรูปภาพมากระตุ้นความคิด จะเป็นตัวช่วยทำให้บุคคลออทิสติกเกิดความเข้าใจความหมายของสิ่งที่ต้องการสื่อความหมายได้ดียิ่งขึ้น [6, 7] สำหรับเทคนิคเฉพาะในการสอนบุคคลที่มีความบกพร่องทางการเรียนรู้นั้น ควรฝึกฝนทักษะการจัดหมวดหมู่ และทักษะการเรียงลำดับ ซึ่งการใช้คำกระตุ้นการคิดที่มีการแบ่งประเภทคำนั้นเป็นหมวดหมู่ จะช่วยทำให้บุคคลที่มีความบกพร่องทางการเรียนรู้สามารถเรียนรู้ได้เร็วยิ่งขึ้น [5]

ในงานวิจัยของ Bell [8] กล่าวถึงแนวคิดการใช้คำกระตุ้นการพูดและการคิดในการสอน เพื่อฝึกให้เด็กคิดวิเคราะห์สร้างเรื่องราวจากรูปภาพที่เห็น โดยการผ่านการตั้งคำถามจากคำกระตุ้นการพูด ซึ่งจะนำไปสู่การเล่าเรื่องจากรูปภาพของเด็ก

ผู้วิจัยได้เล็งเห็นว่า การพัฒนาอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยโดยใช้คำกระตุ้นการพูดตามแนวคิดของ Bell เพื่อนำไปประยุกต์ใช้กับซอฟต์แวร์เพื่อช่วยบุคคลที่มีปัญหาในการพูด และต้องการสื่อสารด้วยวิธีการอื่นแทนการพูด (Augmentative and Alternative Communication: AAC) [9] ในอนาคต ซึ่งในปัจจุบันซอฟต์แวร์นี้ยังมีข้อจำกัดในเรื่องการเรียงลำดับ

คำหรือรูปภาพแบบอัตโนมัติ เพื่อจะเป็นเครื่องมือสนับสนุนการเรียนการสอนของบุคคลออทิสติกและบุคคลที่มีความบกพร่องทางการเรียนรู้ ในการฝึกทักษะการเรียงลำดับคำในประโยคภาษาไทย ทำให้คนกลุ่มนี้เข้าใจหลักไวยากรณ์ภาษาไทยมากขึ้น รวมถึงสามารถใช้เป็นเครื่องมือสนับสนุนการเรียนการสอนของบุคคลออทิสติกและบุคคลที่มีความบกพร่องทางการเรียนรู้ได้

ดังนั้น งานวิจัยนี้จึงนำเสนอแนวคิดการพัฒนาอัลกอริทึมที่มีความสามารถในด้านการเรียงลำดับคำในประโยคภาษาไทยที่จัดวางอย่างไม่เป็นลำดับ ไม่ถูกต้องตามความหมาย และไม่ถูกต้องตามหลักไวยากรณ์ภาษาไทย โดยอาศัยแบบจำลองความน่าจะเป็น (Probabilistic Models) ร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (Weighted Finite-State Transducer) มาใช้ในการวิเคราะห์ข้อมูลประกอบการพัฒนาอัลกอริทึมในการเรียงลำดับคำในประโยคภาษาไทย เพื่อเรียงเป็นประโยคความเดียวภาษาไทยได้ตามที่ผู้ใช้งานต้องการสื่อสาร นอกจากนี้ยังสามารถนำไปประยุกต์ใช้ร่วมกับการพัฒนาซอฟต์แวร์ AAC ต่อไปในอนาคต ทำให้มีประโยชน์ต่อกลุ่มผู้ใช้งานที่หลากหลายมากขึ้น กล่าวคือ นอกจากบุคคลที่มีความบกพร่องทางการพูดและภาษาจะใช้งานได้แล้ว บุคคลออทิสติก บุคคลที่มีความบกพร่องทางการเรียนรู้ และบุคคลที่ต้องการพัฒนาความรู้ทางภาษาไทยก็สามารถใช้งานซอฟต์แวร์นี้ได้

1.2 วัตถุประสงค์ของการวิจัย

เพื่อสร้างอัลกอริทึมในการเรียงลำดับคำในประโยคความเดียวภาษาไทย โดยอาศัยแบบจำลองความน่าจะเป็นของประเภทคำกระตุ้นการพูด

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1.3 ขอบเขตของการวิจัย

1. อัลกอริทึมที่พัฒนาครอบคลุมเรียงลำดับคำในประโยคความเดียวภาษาไทยเท่านั้น

2. อัลกอริทึมที่พัฒนาสามารถเรียงลำดับคำในประโยค โดยไม่ต้องคำนึงถึงลำดับก่อนหลังของการเลือกคำ

3. อัลกอริทึมที่พัฒนาครอบคลุมเรียงลำดับคำในประโยคทีละประโยค

4. คำที่ใช้ในงานวิจัยเป็นคำที่ได้มาจากรูปภาพที่ได้มาจาก Chula AAC [10] และคำที่ปรากฏในหนังสือเรียนภาษาไทย ที่มีชื่อว่า ภาษาพาที สำหรับนักเรียนระดับชั้นประถมศึกษาปีที่ 1 [11] และระดับชั้นประถมศึกษาปีที่ 2 [12]

1.4 ลำดับขั้นตอนในการทำวิจัย

1. ขั้นตอนการศึกษาเบื้องต้น

1.1 ศึกษาวรรณกรรมที่เกี่ยวข้องกับเครื่องมือหรือซอฟต์แวร์เพื่อช่วยคนที่มีปัญหาในการสื่อสาร

1.2 ศึกษาวรรณกรรมที่เกี่ยวข้องกับการเรียงลำดับคำในประโยคแบบอัตโนมัติ

1.3 ศึกษาโครงสร้างประโยคภาษาไทย และคำในภาษาไทย

1.4 ศึกษาแบบจำลองเอ็นแกรม

1.5 ศึกษาวรรณกรรมที่เกี่ยวข้องกับการประยุกต์ใช้ Weighted Finite-State Transducers (WFSTs)

1.6 ศึกษาเครื่องมือที่ใช้ในงานวิจัย เช่น เครื่องมือคำนวณค่าเอ็นแกรม (N-gram Tool) และเครื่องมือการสร้าง Finite-State Transducers ที่มีชื่อว่า OpenFST [13] เป็นต้น

2. ขั้นตอนการออกแบบระบบและทำการทดลอง

2.1 ออกแบบการทำงานของระบบที่นำเสนอ

2.2 เลือกใช้ข้อมูลที่อยู่ในคลังข้อความ BEST2010

2.3 สร้างข้อมูลประโยคความเดียวจากหนังสือเรียนภาษาไทยระดับประถมศึกษาปีที่ 1 และระดับประถมศึกษาปีที่ 2 เพื่อใช้เป็นคลังข้อความ

2.4 วิเคราะห์ข้อมูลเพื่อทำการแบ่งหมวดหมู่ของคำและประเภทของคำจากคลังข้อความ

2.5 เลือกใช้เครื่องมือในการวิจัยที่เหมาะสม

2.6 ออกแบบการทดลอง

2.7 สร้างประโยคเป้าหมายเพื่อใช้เป็นประโยคคำตอบในการทำการทดลองจากคำที่ถูกเลือกทั้งหมดในคลังข้อความ

2.8 ทดสอบการทำงานและบันทึกผลการทดลอง

2.9 วัดผลความแม่นยำ เปรียบเทียบผลการทดลอง และวิเคราะห์ผลการทดลอง

3. สรุปผลและเรียบเรียงวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้นำคำกระตุ้นการพูด มาใช้ในการพัฒนาอัลกอริทึมการเรียงลำดับคำในประโยคความเดียวภาษาไทยแบบอัตโนมัติ โดยอาศัยแบบจำลองความน่าจะเป็นของประเภทคำกระตุ้นการพูด

2. สามารถนำอัลกอริทึมที่พัฒนาไปประยุกต์ใช้ร่วมกับ AAC ซอฟต์แวร์ ต่อไปในอนาคต เพื่อเป็นอีกหนึ่งทางเลือกที่ช่วยในการสื่อสารและอาจใช้เป็นเครื่องมือฝึกฝนความสามารถในการคิดและการจัดการสำหรับผู้ที่ต้องการฝึกฝน

1.6 ผลงานตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้ตีพิมพ์เป็นบทความวิจัยในหนังสือต่าง ๆ ดังต่อไปนี้

1. หนังสือสรุปประชุมวิชาการระดับนานาชาติ ในหัวข้อเรื่อง “An Automatic Natural Language Sentence Generation from Images” จัดทำโดย “Worasa Limpanadusadee, Proadpran Punyabukkana and Atiwong Suchato” โดยนำเสนอผลงานวิจัยในการประชุมวิชาการระดับนานาชาติ “The 7th International Convention for Rehabilitation Engineering & Assistive Technology (i-CREATe 2013)” สถานที่จัดงาน KINTEX, Gyeonggi ประเทศเกาหลีใต้ เมื่อวันที่ 29-31 สิงหาคม 2556

2. หนังสือสรุปประชุมวิชาการระดับนานาชาติ ในหัวข้อเรื่อง “Text Corpus for Natural Language Story-telling Sentence Generation: A Design and Evaluation” จัดทำโดย “Worasa Limpanadusadee, Proadpran Punyabukkana, Atiwong Suchato and Onintra Poobrasert” โดยนำเสนอผลงานวิจัยในการประชุมวิชาการระดับนานาชาติ “The 11th International Joint Conference on Computer Science and Software Engineering (JCSSE 2014)” สถานที่จัดงาน พัทยา ประเทศไทย เมื่อวันที่ 14-16 พฤษภาคม 2557

3. หนังสือสรุปการประชุมเชิงปฏิบัติการในหัวข้อเรื่อง “An Automatic Natural Language Sentence Generation Using Weighted Finite-State Transducer” จัดทำโดย “Worasa Limpanadusadee, Proadpran Punyabukkana, Atiwong Suchato and Shuichi Ueno” โดยนำเสนอผลงานวิจัยในการประชุมเชิงปฏิบัติการ “The 6th Multidisciplinary International Student Workshop (MISW 2014)” สถานที่จัดงาน Tokyo Institute of Technology เมืองโตเกียว ประเทศญี่ปุ่น เมื่อวันที่ 7-8 สิงหาคม 2557

1.7 ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้ได้แบ่งเนื้อหาออกเป็น 8 บทดังต่อไปนี้

บทที่ 1 บทนำ กล่าวถึง ความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ของการวิจัย ขอบเขตของการวิจัย ลำดับขั้นตอนในการเสนอผลการวิจัย ประโยชน์ที่คาดว่าจะได้รับ และผลงานตีพิมพ์จากวิทยานิพนธ์

บทที่ 2 ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง กล่าวถึง แนวคิดและทฤษฎี ซึ่งประกอบด้วย ทฤษฎีที่เกี่ยวกับการประมวลผลภาษาธรรมชาติ ทฤษฎีเกี่ยวกับประเภทของคำ ส่วนประกอบของประโยคและประโยคภาษาไทย ทฤษฎีแบบจำลองเอ็นแกรม ทฤษฎีการปรับเรียบ ทฤษฎีตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก และวรรณกรรมที่เกี่ยวข้องกับงานวิจัยนี้

บทที่ 3 ขั้นตอนการดำเนินงานวิจัยและการออกแบบระบบที่นำเสนอ กล่าวถึง ระบบการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอโดยอาศัยหลักการและลักษณะสำคัญเอ็นแกรม และระบบการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอโดยอาศัยหลักการและลักษณะสำคัญเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก

บทที่ 4 การออกแบบและสร้างคลังข้อความ กล่าวถึงคลังข้อความที่ใช้ในงานวิจัย รวมไปถึงการระบุหมวดหมู่หลัก หมวดหมู่ย่อย และหน้าที่ของคำ โดยแบ่งการระบุหน้าที่ของคำออกเป็น 2 แบบ ได้แก่ระบุหน้าที่ของคำตามวิธีที่ใช้อ้างอิง และระบุตามวิธีที่งานวิจัยนี้นำเสนอ

บทที่ 5 ขั้นตอนการสร้างระบบและอัลกอริทึมที่นำเสนอ กล่าวถึง ขั้นตอนการสร้างระบบแบบจำลองต่าง ๆ ที่ต้องใช้ในระบบ และอัลกอริทึมที่นำเสนอในงานวิจัยนี้ รวมถึงคำนวณคะแนนและเลือกประโยคคำตอบของระบบ

บทที่ 6 การทดลอง และวิธีการวัดผลการทดลอง กล่าวถึง ข้อมูลที่ใช้ในการทดสอบระบบ การกำหนดหมวดหมู่หลักและหมวดหมู่ย่อยของคำที่ใช้ในการทดสอบระบบ การสร้างชุดทดสอบและประโยคทดสอบ การวัดผลทดลอง และการทดลอง

บทที่ 7 ผลการทดลองและวิเคราะห์ผลการทดลอง กล่าวถึง การประเมินผลการการทดลองระบบที่นำเสนอโดยอาศัยหลักการเอ็นแกรม การประเมินผลการทดลองหาลักษณะสำคัญของการระบุหมวดหมู่ของคำในคลังข้อความ การประเมินผลประสิทธิภาพของระบบอ้างอิง และการประเมินผลประสิทธิภาพของแนวทางที่นำเสนอโดยอาศัยหลักการเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก

บทที่ 8 สรุปผลการวิจัยและข้อเสนอแนะ

บทที่ 2
ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

ในบทนี้กล่าวถึงทฤษฎีและวรรณกรรมที่เกี่ยวข้อง โดยแบ่งออกเป็น 2 ส่วนหลัก ส่วนแรกกล่าวถึง ทฤษฎีที่เกี่ยวข้องกับงานวิจัยนี้ และในส่วนที่สอง จะกล่าวถึงวรรณกรรมทั้งหมดที่เกี่ยวข้องกับวิทยานิพนธ์นี้

2.1 ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้องในงานวิจัยนี้ ได้แก่ ทฤษฎีเกี่ยวกับการประมวลผลภาษาธรรมชาติ ทฤษฎีเกี่ยวกับประเภทของคำ ส่วนประกอบของประโยค และประโยคภาษาไทย ทฤษฎีแบบจำลองเอ็นแกรม ทฤษฎีการปรับเรียบ และ ทฤษฎีตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก ซึ่งมีรายละเอียดดังต่อไปนี้

2.1.1 ทฤษฎีเกี่ยวกับการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP)

ทฤษฎีเกี่ยวกับการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) เป็นกระบวนการที่จะทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ โดยรับข้อมูลนำเข้าเป็นข้อความในภาษาหนึ่ง ๆ แล้วทำความเข้าใจว่าผู้ป้อนข้อความต้องการกล่าวถึงสิ่งใด งานวิจัยทางด้านการประมวลผลภาษาธรรมชาติ สามารถแบ่งได้เป็น 2 ระบบ คือ การทำความเข้าใจภาษาธรรมชาติ Natural Language Understanding (NLU) และการสร้างภาษาธรรมชาติ Natural Language Generation (NLG) โดยทั้ง 2 ระบบจะมีลักษณะที่สมมาตรกัน แต่กลับทิศกัน และต้องอาศัยฐานความรู้เช่นเดียวกันทั้งทางด้านภาษาและฐานความรู้ทั้งในเชิงบริบท รวมทั้งความรู้เฉพาะ และความรู้ทั่วไป การประมวลผลภาษามีได้หลายระดับไม่ว่าจะเป็น ในระดับคำ ในระดับประโยค รวมถึงในระดับแปลความหมายของประโยคก็สามารถทำได้ [14]

1. ขั้นตอนในการเข้าใจภาษาธรรมชาติ

การเข้าใจภาษาธรรมชาติ มีขั้นตอนอ้างอิงจาก [14] ดังต่อไปนี้

1.1 การวิเคราะห์ทางองค์ประกอบ (Morphological Analysis)

เป็นการวิเคราะห์การแยกหน่วยย่อยของแต่ละหน่วยคำ

1.2 การวิเคราะห์ทางวากยสัมพันธ์ (Syntactic Analysis)

เป็นการวิเคราะห์ทางไวยากรณ์โดยมีจุดมุ่งหมายเพื่อต้องการศึกษาโครงสร้างเชิงวากยสัมพันธ์ของประโยคที่มีหลายคำต่อกัน กล่าวคือ เพื่อศึกษาดูว่าคำแต่ละคำทำหน้าที่อะไรในประโยคหรือส่วนใดของประโยคเป็นวลี

1.3 การวิเคราะห์ทางความหมาย (Semantic Analysis)

เป็นการวิเคราะห์ความหมาย ต่อจากการวิเคราะห์ทางวากยสัมพันธ์ โดยมีการกำหนดความหมายของคำแต่ละคำ

1.4 บูรณาการทางวจนิพนธ์ (Discourse Integration)

เป็นการพิจารณาความหมายของประโยคโดยดูจากประโยคข้างเคียง เนื่องจากคำบางคำในประโยคหนึ่ง ๆ จะเข้าใจความหมายได้ต้องดูประโยคก่อนหน้าหรือประโยคที่ตามหลัง

1.5 การวิเคราะห์ทางปฏิบัติ (Pragmatic Analysis)

เป็นการแปลความหมายของประโยคอีกครั้งว่าในความเป็นจริงอะไรคือสิ่งที่ผู้พูดตั้งใจจะสื่อความหมาย

โดยในงานวิจัยนี้ได้นำทฤษฎีการประมวลผลภาษาธรรมชาติ ทั้งในส่วนของการทำความเข้าใจและการสร้างภาษาธรรมชาติ มาใช้ในการเรียงลำดับคำในประโยคความเดียวภาษาไทยแบบอัตโนมัติ โดยเริ่มจากการวิเคราะห์หน้าที่ของคำหรือองค์ประกอบของคำในประโยคว่าประโยคหนึ่ง ๆ สามารถแบ่งออกเป็นหน่วยย่อยได้อย่างไรบ้าง จากนั้นก็ทำการวิเคราะห์ความหมายของประโยคที่ผู้สร้างประโยคต้องการจะสื่อ เพื่อใช้ประโยชน์ในการนำคำจากแต่ละประโยคมาสร้างประโยคใหม่ต่อไป

2.1.2 ทฤษฎีเกี่ยวกับประเภทของคำ ส่วนประกอบของประโยคและประโยคภาษาไทย

2.1.2.1 ประเภทของคำ

คำในภาษาไทย ซึ่งอ้างอิงและคัดลอกเนื้อหาบางส่วนมาจาก [15, 16] แบ่งออกเป็น 7 ชนิด ดังนี้

1. คำนาม คือ คำที่ใช้เรียกชื่อคน สัตว์ สิ่งของ และสถานที่ แบ่งออกเป็น 5 ชนิด ดังนี้

นามทั่วไป หรือสามานยนาม เป็นคำนามที่ไม่ชี้เฉพาะบุคคลใดบุคคลหนึ่ง เช่น คน นักเรียน บุรุษไปรษณีย์ กีฬา เป็นต้น

นามชื่อเฉพาะ หรือวิสามานยนาม เป็นชื่อเฉพาะของคน บุคคล หรือสถานที่ เช่น สมศักดิ์ สุดา มีนบุรี ฉะเชิงเทรา เป็นต้น

นามบอกลักษณะ หรือลักษณนาม เป็นคำนามบอกลักษณะ เช่น ผล ตัว อัน แท่ง ใบ เป็นต้น

นามบอกหมวดหมู่ หรือสมุหนาม เป็นคำนามที่บอกหมวดหมู่ เช่น โครง ฝูง หมู่ เหล่า เป็นต้น

นามบอกอาการ หรืออาการนาม เป็นคำที่เกิดจากคำกริยาหรือคำวิเศษณ์ที่มีคำว่า การ หรือ ความ นำหน้า เช่น การวิ่ง ความดี ความจริง เป็นต้น

2. คำสรรพนาม คือ คำที่ใช้แทนคำนาม เพื่อไม่ต้องการกล่าวคำนามนั้นซ้ำๆ แบ่งออกเป็น 7 ชนิด ดังนี้

บุรุษสรรพนาม เป็นสรรพนามที่ใช้แทนคำนาม มี 3 ชนิด ดังนี้

สรรพนามบุรุษที่ 1 ใช้เรียกแทนผู้พูด เช่น ข้าพเจ้า ดิฉัน ผม เป็นต้น

สรรพนามบุรุษที่ 2 ใช้เรียกแทนผู้ฟัง เช่น เธอ ท่าน คุณ เป็นต้น

สรรพนามบุรุษที่ 3 ใช้แทนผู้ถูกกล่าวถึง เช่น มัน เขา ท่าน เป็นต้น

สรรพนามชี้ระยะ เป็นสรรพนามที่กำหนดให้เรารู้ว่าเป็น คน สัตว์ สิ่งของ หรือสถานที่ ที่กล่าวถึงนั้นอยู่ในระยะใกล้หรือไกลเพียงใด

นี่ ใช้กับ สิ่งที่อยู่ใกล้ที่สุด

นั่น ใช้กับ สิ่งที่อยู่ไกลหรือห่างออกไป

สรรพนามใช้ถาม ใช้แทนคำนามที่ผู้ถามต้องการคำตอบ ได้แก่ อันไหน ใคร ที่ไหน เช่น

เธอชอบ<u>อันไหน</u>

คุณเลือก<u>สิ่งใด</u>ในตู้นี้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สรรพนามบอกความไม่เจาะจง ใช้แทนคำนามที่กล่าวถึง โดยไม่ต้องการคำตอบ ได้แก่ ใคร อะไร สิ่งใด หรืออาจจะใช้คำซ้ำ เช่น

ฉันไม่เห็นใครขยันอย่างเขา

รู้สิ่งใดหรือจะสู้รู้วิชา

สรรพนามบอกความชี้ซ้ำ ใช้แทนคำนามที่กล่าวมาก่อนแล้ว และต้องการกล่าวซ้ำอีกโดยไม่ต้องเอ่ยคำนามนั้นอีกครั้ง ได้แก่ บ้าง ต่าง กัน เช่น

ประชาชนต่างไปลงคะแนนเสียงเลือกตั้งผู้แทนราษฎร

(ความหมายแยกแต่ทำในสิ่งเดียวกัน)

สรรพนามเชื่อมประโยค ใช้แทนคำนามที่อยู่ข้างหน้า และเมื่อต้องการกล่าวซ้ำ ทำหน้าที่เชื่อมประโยค 2 ประโยค เข้าด้วยกัน ได้แก่ ที่ ซึ่ง อัน ผู้ เช่น

คนที่เดินมาเป็นน้องฉัน

เมื่อแยกประโยคจะได้ ดังนี้ คนเดินมา และ น้องฉัน ส่วนคำว่า ที่ แทนคำว่า คน

สรรพนามใช้เน้นนามตามความรู้สึกของผู้พูด ใช้หลังคำนามเพื่อบอกความรู้สึกที่มีต่อบุคคลที่กล่าวถึง เช่น

เด็ก ๆ แกหัวเราะเสียงดัง (บอกความรู้สึกเอ็นดู)

คุณแอมเธอดีกับทุกคน (บอกความรู้สึกยกย่อง)

3. คำกริยา คือ คำที่แสดงอาการ หรือสภาพ หรือการกระทำในประโยค บางคำมีความหมายสมบูรณ์แต่บางคำต้องอาศัยคำอื่นมาประกอบ หรือใช้คำประกอบอื่น เพื่อให้ความหมายชัดขึ้น แบ่งออกเป็น 5 ชนิด ดังนี้

กริยาที่มีความหมายสมบูรณ์ เรียกว่า อกรรมกริยา เช่น เดิน วิ่ง ร้องไห้

กริยาที่ต้องอาศัยกรรมมาทำให้สมบูรณ์ เรียกว่า สกรรมกริยา เช่น ทำ ซื้อ กิน

กริยาที่ช่วยใช้กริยาอื่น มีความหมายชัดเจนขึ้น เรียกว่า กริยานุเคราะห์ เช่น คง จะ น่า แล้ว อาจ นะ ต้อง

กริยาที่ต้องอาศัยส่วนเติมเต็มเพื่อให้มีความหมายสมบูรณ์ ส่วนเติมเต็มนี้ไม่ใช่กรรม คำกริยาดังกล่าว ได้แก่ เป็น เหมือน คล้าย เท่า คือ

กริยาที่ทำหน้าที่คล้ายนาม อาจเป็นประธาน กรรม หรือบทขยายประโยค คำกริยาชนิดนี้มักจะมีคำว่า “การ” เช่น พูด อย่างเดียวไม่ทำให้งานสำเร็จหรอก

4. คำวิเศษณ์ คือ คำที่ช่วยขยายคำอื่น ให้มีเนื้อความชัดเจนยิ่งขึ้น หน้าที่ของคำวิเศษณ์ มีดังนี้

ประกอบคำนาม เช่น คนงาม ชายหนุ่ม มะพร้าวอ่อน ผ้าบาง

ประกอบคำสรรพนาม เช่น เขาสูง เธอสวย มันดุ

ประกอบคำกริยา เช่น วิ่งเร็ว อยู่ไกล นอนมาก

ประกอบคำวิเศษณ์ เช่น อากาศร้อนมาก เธอดื่มน้ำเย็นจัด

5. คำบุพบท คือ คำที่ทำหน้าที่เชื่อมโยงคำหนึ่ง หรือกลุ่มคำหนึ่งให้สัมพันธ์กับคำอื่นหรือกลุ่มคำอื่น เพื่อบอกสถานที่ เวลา แสดงอาการ หรือแสดงความเป็นเจ้าของ เช่น คำว่า กับ แก่ แต่ ต่อ จาก เพื่อ โดย จน บน ใต้ กลาง ของ สำหรับ ระหว่าง เฉพาะ เช่น

แม่ทำ<u>เพื่อ</u>ลูก “เพื่อ” เป็นคำบุพบทแสดงอาการ

เขามา<u>เมื่อ</u>เช้านี้ “เมื่อ” เป็นบุพบทบอกเวลา

ลูกชายนั่งอยู่<u>ข้าง</u>พ่อ “ข้าง” เป็นบุพบทบอกสถานที่

ฉันเดิน<u>กับ</u>เขา “กับ” เป็นบุพบทบอกความเกี่ยวเนื่องกัน

6. คำสันธาน คือ คำที่ใช้เชื่อมประโยคกับประโยค ข้อความที่มีคำสันธานเชื่อมอยู่นั้นมักจะแยกออกได้เป็นประโยคมากกว่าหนึ่งประโยค เช่น

แม่เต่าเดินเร็ว<u>แต่</u>ลูกเต่าเดินช้า

ติ๊ก<u>และ</u>ต๊อกเป็นเด็กน้ำใจ

ครูให้อภัยเขา<u>เพราะ</u>เขาสำนึกผิด

บ้าน<u>และ</u>โรงเรียนควรร่วมมือกันแก้ปัญหาเรื่องเด็ก

7. คำอุทาน คือ คำที่เปล่งออกมาเพื่อแสดงอารมณ์ หรือความรู้สึกของผู้พูด ส่วนมากจะไม่มีความหมายตรงตามถ้อยคำ แต่จะมีความหมายทางการเน้นความรู้สึกและอารมณ์ของผู้พูดเป็นสำคัญ แบ่งออกเป็น 3 ชนิด คือ

คำอุทานบอกอาการ ใช้บอกอาการในการพูดจากัน เช่น

โธ่! ไม่น่าเลย

โอ๊ย! น่ากลัวจริง

เอ๊ะ! ทำอย่างนี้ได้อย่างไร

ตายแล้ว! ทำไมเป็นอย่างนี้

คำอุทานบอกอาการจะมีเครื่องหมายอัศเจรีย์ (!) กำกับอยู่ คำอุทานบอกอาการนี้ จะพูดออกมาเพื่อแสดงอารมณ์ของผู้พูด เช่น ดีใจ เสียใจ ตกใจ แปลกใจ ฯลฯ

คำอุทานในบทร้อยกรอง คำอุทานเหล่านี้จะปรากฏอยู่ในบทร้อยกรอง เช่น อ้า โอ้ เอย เอ่ย แล แฮ นา นอ ฯลฯ ใช้แทรกลงในถ้อยคำเพื่อความสละสลวย เช่น

โอ้ดวงเดือนก่ำฟ้า ชวนฝัน ยิ่งเอย
ในพฤกษ์จับแสงจันทร์ ป่ากว้าง
หมู่แมลงเพรียกหากัน ดุจเพื่อน รักนา
ป่าปลอบผู้อ้างว้าง ปลดเศร้าอุราหมอง

จตุภูมิ วงษ์แก้ว

อุทานเสริมบท ใช้กล่าวเป็นการเสริมคำในการพูดเพื่อเน้นความหมายให้ชัดเจนขึ้น ทำให้สนุกในการออกเสียงให้น่าฟังขึ้น เช่น

เธอไม่สบายต้องกินหยูกกินยานะ
ช่วยหยิบถ้วยโถโอชามให้ด้วย
เข้ามาดื่มน้ำดื่มท่าเสียก่อน
ทำไมบ้านช่องถึงสกปรกอย่างนี้

2.1.2.2 ส่วนประกอบของประโยค

ประโยคหนึ่ง ๆ จะต้องประกอบไปด้วยภาคประธานและภาคแสดงเป็นหลัก อ้างอิงจาก [15, 16]

ภาคประธานในประโยค คือ คำหรือกลุ่มคำที่ทำหน้าที่เป็นผู้กระทำ ซึ่งเป็นส่วนสำคัญของประโยค อาจมีบทขยายซึ่งเป็นคำหรือกลุ่มคำมาประกอบ เพื่อทำให้มีใจความชัดเจนยิ่งขึ้น

ภาคแสดงในประโยค คือ คำหรือกลุ่มคำที่ประกอบด้วยบทกริยา บทกรรม และส่วนเติมเต็ม บทกริยาทำหน้าที่เป็นตัวกระทำหรือตัวแสดงของประธาน ส่วนบทกรรมทำหน้าที่เป็นผู้ถูกกระทำ และส่วนเติมเต็มทำหน้าที่เสริมใจความของประโยคให้สมบูรณ์

2.1.2.3 ประโยคในภาษาไทย

การสื่อสารระหว่างกันจำเป็นต้องใช้ข้อความที่เป็นประโยคมาใช้ในการสื่อสารเพื่อความเข้าใจและกระจ่างชัดตรงกัน โดยประโยคนั้นเกิดจากการนำคำหลาย ๆ คำ หรือวลีมาเรียงต่อกันอย่างมีระเบียบให้แต่ละคำมีความสัมพันธ์กัน มีใจความสมบูรณ์ เพื่อแสดงให้รู้ว่าใคร ทำอะไร ที่ไหน อย่างไร อาจมีขนาดสั้นหรือยาวก็ได้ขึ้นอยู่กับความต้องการของผู้ส่งสาร

ประโยคในภาษาไทยแบ่งออกเป็น 3 ประเภท อ้างอิงจาก [17] คือ

1. ประโยคความเดียว คือ ประโยคที่มีใจความเดียว เป็นประโยคที่มีภาคประธานเพียงบทเดียว และมีภาคแสดงหรือกริยาสำคัญเพียงบทเดียว

2. ประโยคความรวม คือ ประโยคที่รวมเอาประโยคความเดียวตั้งแต่ 2 ประโยคขึ้นไปมารวมกัน โดยใช้สันธานเป็นตัวเชื่อม แต่ก็สามารถแยกออกเป็นประโยคความเดียวที่มีใจความสมบูรณ์ได้เหมือนเดิมโดยไม่ต้องเพิ่มส่วนใดส่วนหนึ่งในประโยค

3. ประโยคความซ้อน คือ ประโยคที่ประกอบด้วยประโยคหลักและประโยคย่อย มารวมเป็นประโยคเดียวกันโดยมีบุพบทเป็นบทเชื่อม

ประโยคหลัก คือ ประโยคที่เป็นใจความสำคัญที่ต้องการสื่อสาร

ประโยคย่อย คือ ประโยคที่ทำหน้าที่ขยายความประโยคหลักให้สมบูรณ์ยิ่งขึ้น

ในงานวิจัยนี้นั้นได้พัฒนาอัลกอริทึมการเรียงลำดับคำในประโยคความเดียวภาษาไทย ซึ่งประโยคความเดียวภาษาไทยที่ได้จากการเรียงลำดับคำในงานวิจัยนี้ ประกอบไปด้วยภาคประธานเพียงบทเดียว และมีภาคแสดงหรือกริยาสำคัญเพียงบทเดียว สามารถมีส่วนขยายส่วนต่าง ๆ ของประโยคได้ โดยแต่ละประโยคจะมีความยาวได้ตั้งแต่ 3 ถึง 6 คำ เนื่องจากประโยคความเดียว เป็นประโยคที่เด็กที่มีความบกพร่องทางการเรียนรู้ และเด็กทั่วไป ใช้เล่าเรื่องราวมาถึง 70 เปอร์เซ็นต์ ตามงานวิจัยของคณะอักษรศาสตร์ มหาวิทยาลัยศิลปากร [18]

2.1.3 ทฤษฎีแบบจำลองเอ็นแกรม (N-gram Model)

ทฤษฎีแบบจำลองเอ็นแกรม [19] คือ แบบจำลองที่ใช้คำนวณค่าความน่าจะเป็นของชุดคำหรือแกรม ที่เกิดขึ้นร่วมกันเป็นประโยค โดยค่าความน่าจะเป็นของคำสามารถประมาณได้จากคลังข้อความที่สร้างไว้ สำหรับแกรมที่ใช้ในแบบจำลองเป็นไปได้ทั้งเสียง อักขระ คำ หรือประโยค ซึ่งแกรมสามารถมีได้หลายขนาด ตั้งแต่ 1 ถึง n ตัว การคำนวณความน่าจะเป็นของการเกิดชุดอักขระที่รวมกันเป็นคำ แสดงดังสมการที่ 2.1

$$P(w_1 \dots w_n) = P(w_1) \prod_{i=2}^{n} P(w_i | w_{i-1}) \tag{2.1}$$

กำหนดให้

P คือ ค่าความน่าจะเป็น

w คือ อักขระ

$(w_1 \dots w_n)$ คือ ชุดอักขระที่ประกอบด้วยอักขระตั้งแต่ 2 ตัวขึ้นไปจนถึง n ตัว

เมื่อมีประโยค $W = w_1,\dots,w_n$ มาให้แบบจำลองเอ็นแกรม จะคำนวณค่าความน่าจะเป็นของการเกิดคำใด ๆ โดยพิจารณาจาก $N-1$คำก่อนหน้า อาทิเช่น $N = 2$ ซึ่งเรียกว่าไบแกรม (Bigram) หรือ 2-แกรม นั้น จะให้ค่าความน่าจะเป็นของ คำใดๆ โดยดูจาก คำก่อนหน้าเพียง 1 คำ โดยมักจะเขียนค่าความน่าจะเป็นนี้ในรูปของ $P(w_2 \mid w_1)$ ซึ่งหมายถึงความน่าจะเป็นที่คำ w_2 จะเกิดตามหลังคำ w_1 อาทิเช่น

$P(\text{ไป}|\text{จะ}) = 0.8$ $P(\text{จะ}|\text{ไป}) = 0.01$ $P(\text{จะ}|\text{ผม}) = 0.7$

$P(\text{ตลาด}|\text{ไป}) = 0.5$ $P(\text{โรงเรียน}|\text{ไป}) = 0.6$ $P(\text{ผม}|\text{ไป}) = 0.02$

เมื่อมีประโยคมีจำนวนคำมากขึ้น $W = w_1,\dots,w_n$ เช่น "ผม จะ ไป โรงเรียน" ค่าความน่าจะเป็นที่จะเกิดประโยคดังกล่าวก็สามารถคำนวณได้โดยการคูณต่อเนื่องกันไปดังนี้

$P(\text{ผม จะ ไป โรงเรียน}) = P(\text{จะ}|\text{ผม}) \times P(\text{ไป}|\text{จะ}) \times P(\text{โรงเรียน}|\text{ไป})$

$= (0.7)(0.8)(0.6)$

$= 0.336$

$P(\text{จะ ไป ผม จะ}) = P(\text{ไป}|\text{จะ}) \times P(\text{ผม}|\text{ไป}) \times P(\text{จะ}|\text{ผม})$

$= (0.8)(0.02)(0.7)$

$= 0.0112$

จะเห็นว่า “จะ ไป ผม จะ” มีโอกาสเกิดน้อยมากเมื่อเทียบกับ “ผม จะ ไป โรงเรียน” ในทำนองเดียวกัน เราสามารถสร้าง 3-แกรม (ไตรแกรม) โดยที่กำหนดค่าความน่าจะเป็นของแต่ละคำโดยพิจารณา 2 คำก่อนหน้าได้เช่นกัน

2.1.4 ทฤษฎีการปรับเรียบ (Smoothing)

ในการคำนวณค่าความน่าจะเป็นอาจเกิดกรณีที่ทำการคำนวณความน่าจะเป็นได้ผลลัพธ์เป็นศูนย์ เนื่องมาจากไม่มีชุดข้อมูลนั้นในคลังข้อความ ดังนั้นค่าความน่าจะเป็นของชุดข้อมูลนั้นจึงเป็นศูนย์ หากนำค่าความน่าจะเป็นนั้นไปคำนวณจะทำให้ผลลัพธ์จากสมการที่ต้องการคำนวณเป็นศูนย์ไปด้วย ซึ่งเป็นปัญหาหนึ่งที่พบเมื่อเลือกใช้แบบจำลองเอ็นแกรม การที่ไม่พบข้อมูลในคลังข้อความ ไม่ใช่ว่าชุดคำนั้นจะไม่มีโอกาสเกิดขึ้นเลย ดังนั้นจึงมีวิธีเพิ่มค่าความน่าจะเป็นสำหรับข้อมูลที่ไม่มีในคลังข้อความ การแก้ปัญหานี้สามารถทำได้โดยใช้ทฤษฎีการปรับเรียบ (Smoothing) ซึ่งปัจจุบันมีหลายวิธี ในที่นี้จะขอยกตัวอย่างการทำให้เรียบแบบย้อนกลับ ซึ่งเป็นวิธีที่ใช้ในงานวิจัยนี้

2.1.4.1 การทำให้ราบเรียบย้อน (Back-off Smoothing)

การทำให้ราบเรียบย้อน [20] เป็นอีกหนึ่งวิธีที่ถูกนำมาใช้ในการปรับค่า ตัวอย่างของกรณีที่ไม่มีชุดข้อมูลของบางไตรแกรม $w_{n-2}w_{n-1}w_n$ ในการคำนวณค่า $P(w_n \mid w_{n-1}w_{n-2})$ จะทำการประมาณด้วยค่าความน่าจะเป็นของไบแกรม $P(w_n \mid w_{n-1})$ และเมื่อยังหาค่าไม่ได้ก็จะทำการประมาณค่าด้วยยูนิแกรม $P(w_n)$ ดังนั้นการประมาณค่าความน่าเป็นของแบบจำลองไตรแกรมเป็นได้ดังสมการ 2.2

$$P(w_n|w_{n-2}w_{n-1}) = \begin{cases} P(w_n|w_{n-2}w_{n-1}), if C(w_{n-2}w_{n-1}w_n) > 0 \\ \alpha_1 P(w_n|w_{n-1}) \ , if C(w_{n-2}w_{n-1}w_n) = 0 \ and \ C(w_{n-1}w_n) \ > 0 \\ \alpha_2 P(w) \qquad\qquad , other \end{cases} \tag{2.2}$$

กำหนดให้

α_1 และ α_2 คือค่าน้ำหนักย้อน ซึ่งขึ้นกับอัลกอริทึมที่เลือกใช้ในการทำให้ราบเรียบย้อน

2.1.5 ทฤษฎีตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (Weighted Finite-State Transduces: WFST)

ตัวเปลี่ยนสถานะจำกัดแบบถ่วงน้ำหนัก [21, 22] จะทำหน้าที่ตรวจสอบสิ่งที่ป้อนเข้ามาสู่ระบบหรืออินพุต (Input) ว่าสามารถยอมรับได้หรือไม่ จากนั้นจะแสดงผลเป็นเอาต์พุต (Output) ออกมาในลักษณะของสายอักขระ โดยกำหนดให้ตัวเปลี่ยนสถานะจำกัดแบบถ่วงน้ำหนัก (WFST) $T = (\Sigma, \Omega, Q, E, i, F, \lambda, \rho)$ และ $\mathbb{K}$ เป็นเซมิริง (Semiring) ซึ่งประกอบด้วย

Σ แทนตัวอักษรที่เป็นอินพุต

Ω แทนตัวอักษรที่เป็นเอาต์พุต

Q แทนเซตของสถานะ

E แทนเซตของการส่งผ่าน (Transitions)

โดยที่ $E \subseteq Q \times (\Sigma \cup \{\varepsilon\}) \times \Omega \cup \{\varepsilon\}) \times \mathbb{K} \times Q$

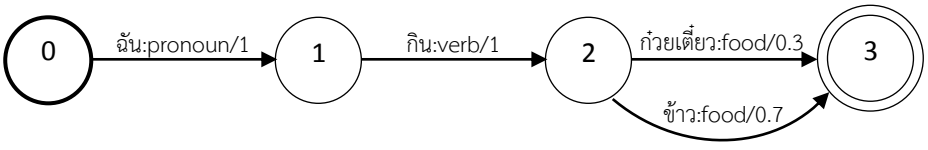
i แทนสถานะเริ่มต้น โดยที่ $i \in Q$

F แทนเซตของสถานะสิ้นสุด $F \in Q$

λ แทนค่าน้ำหนักเริ่มต้น

ρ แทนฟังก์ชันของค่าน้ำหนักสุดท้าย

การจะยอมรับลำดับของสัญลักษณ์ x นั้น จะต้องมีเส้นทาง π เริ่มตั้งแต่สถานะเริ่มต้น i จนถึงสถานะปลายทาง f โดยที่ $f \subseteq F$



ภาพที่ 2-1 ตัวอย่างของตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก

จากภาพ 2-1 แสดงตัวอย่างของตัวเปลี่ยนแปรสถานะแบบถ่วงน้ำหนักของประโยค “ฉันกินข้าว” และ “ฉันกินบะหมี่” โดยที่สัญลักษณ์ที่อยู่บนเส้นเชื่อมแต่ละสถานะนั้น จะแทนด้วย

เลเบลของอินพุต : เลเบลของเอาต์พุต / ค่าน้ำหนัก

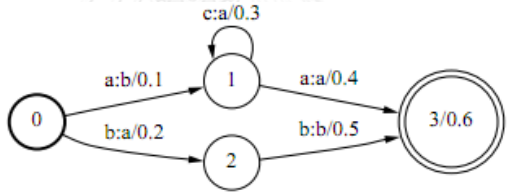
ตัวอย่างเช่น ในสถานะที่สองไปยังสถานะที่สาม (เส้นล่าง) จะมีสัญลักษณ์บนเส้นเชื่อม คือ ข้าว:food/0.7 หมายถึง เมื่อมีเลเบลอินพุตเป็น ข้าว จะให้เลเบลเอาต์พุตเป็น food และมีค่าน้ำหนัก คือ 0.7 โดยค่าน้ำหนักบนเส้นเชื่อมหมายถึงค่าความน่าจะเป็นในการเปลี่ยนสถานะที่เส้นเชื่อมนั้นเชื่อมอยู่ และค่าความน่าจะเป็นของเส้นทางแต่ละเส้นจะมาจากผลคูณของค่าน้ำหนักภายในเส้นเชื่อมแต่ละเส้นในเส้นทางนั้น

2.1.5.1 ฟังก์ชันการประกอบ (Composition Function)

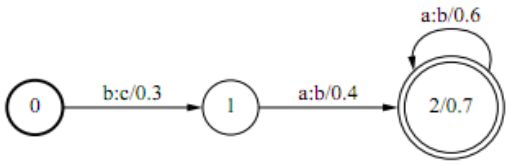
การประกอบกันของตัวเปลี่ยนแปรสถานะจำกัดใดๆ [22] จำนวน 2 ตัว สามารถเขียนแทนด้วย $T_3 = T_1 \circ T_2$ หมายถึง การนำตัวเปลี่ยนแปร 2 ตัว คือ T_1และ T_2 มาประกอบกัน และได้ผลลัพธ์คือตัวเปลี่ยนแปรสถานะ T_3 ดังสมการที่ 2.3

$$[\![T_1 \circ T_2]\!](x,y) = \bigoplus [\![T_1]\!](x,z) \otimes [\![T_2]\!](z,y) \quad (2.3)$$

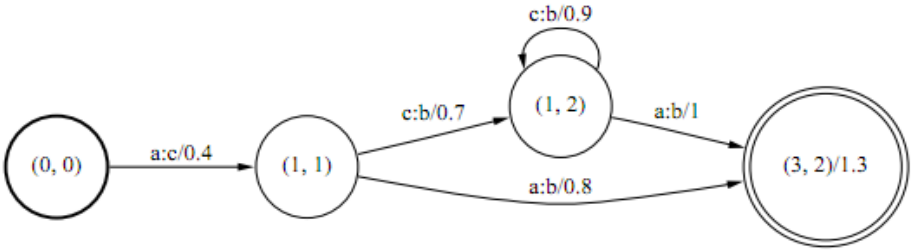
ตัวอย่างการใช้ฟังก์ชันประกอบในการรวมกันของการประกอบกันของตัวเปลี่ยนแปรสถานะจำกัด ตัวใดๆ แสดงดังภาพที่ 2-2



(a)



(b)



(c)

ภาพที่ 2-2 ตัวอย่างการใช้ฟังก์ชันประกอบของ (a),(b) ตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก และ (c) ตัวอย่างผลลัพธ์ของตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักที่ได้จากการใช้ฟังก์ชันประกอบ

จากภาพที่ 2-2 จะการประกอบกันของแทนตัวเปลี่ยนแปรสถานะจำกัดใดๆ 2 ตัว โดยตัวเปลี่ยนแปรสถานะจำกัดภาพที่ 2-2(a) สามารถใช้ฟังก์ชันประกอบกับตัวเปลี่ยนแปรสถานะจำกัดภาพที่ 2-2(b) และได้ผลลัพธ์คือตัวเปลี่ยนแปรสถานะจำกัดใหม่ดังภาพที่ 2-2(c) นั้นสามารถทำได้ตามเงื่อนไขดังนี้

1. สถานะเริ่มต้นจะเป็นคู่ของสถานะเริ่มต้นของตัวเปลี่ยนแปรสถานะทั้งสอง (a),(b)
2. สถานะสิ้นสุดจะเป็นคู่ของสถานะเริ่มต้นของตัวเปลี่ยนแปรสถานะทั้งสอง (a),(b)
3. ตัวเปลี่ยนแปรสถานะจะสามารถประกอบกันได้และส่งผ่านไปยังตัวเปลี่ยนแปรสถานะ (c) นั้น เอาต์พุตของคู่อันดับในตัวเปลี่ยนแปรสถานะ (a) จะต้องมีค่าเท่ากับอินพุตของคู่อันดับในตัวเปลี่ยนแปรสถานะ (b) ที่อยู่ในสถานะเดียวกัน ดังภาพที่ 2-2(a) a:b/0.1 สามารถประกอบกับภาพที่ 2-2(b) b:c/0.3 และส่งผ่านค่าไปยังภาพที่ 2-2(c) a:c/0.4
4. ค่าน้ำหนักที่ส่งผ่านไปยังตัวเปลี่ยนแปรสถานะตัว (c) นั้น เกิดจากผลรวมของค่าน้ำหนักของตัวเปลี่ยนแปรสถานะ (a) และ (b) เมื่อตัวเปลี่ยนแปรสถานะทั้งสองสามารถประกอบกันได้

2.2 งานวิจัยที่เกี่ยวข้อง

งานวิจัยและซอฟต์แวร์ AAC ที่พัฒนาออกมาในปัจจุบันนั้น มีความหลายหลายและมีกลุ่มเป้าหมายที่แตกต่างกัน และมีข้อจำกัดในด้านต่าง ๆ แตกต่างกันไป สำหรับตัวอย่างงานวิจัยและผลิตภัณฑ์ที่เกี่ยวข้องกับ AAC นั้น ได้แก่ Minspeak [23] เป็นซอฟต์แวร์ AAC ที่ผู้ใช้สามารถสร้างประโยคได้หลากหลายจากรูปภาพที่จำกัด เนื่องจากรูปภาพรูปหนึ่งสามารถมีได้หลายความหมาย แต่ผู้ใช้จะต้องเรียงลำดับรูปภาพ และเป็นผู้ระบุหน้าที่ของคำในรูปนั้น ๆ ด้วยตัวของผู้ใช้เอง จึงจะสามารถเรียบเรียงประโยคที่ต้องการและเหมาะสมตามหลักไวยากรณ์ได้ ต่อมา Copestake [24] สร้าง AAC สำหรับ Non-speaking users และได้นำหลัก Natural Language Processing (NLP) มาใช้ร่วมกับการพยากรณ์คำ (Word prediction) ภาษาอังกฤษ และนำ Parts of speech bigram มาใช้หาคำที่แทนที่กันได้ในประโยค แต่ประโยคที่เกิดขึ้นได้นั้น จะต้องมาจากการเลือกคำของผู้ใช้งานที่เรียงกันตามลำดับอย่างถูกต้องและเหมาะสมก่อน ซอฟต์แวร์นี้จึงมีข้อจำกัดในเรื่องของผู้ใช้ที่ต้องมีความสามารถในการเรียงลำดับคำในประโยคได้ดีจึงจะสามารถใช้งานซอฟต์แวร์ตัวนี้ได้ และ How was school today [25] ซอฟต์แวร์สำหรับผู้ที่มีความบกพร่องด้านภาษา (Language-Impaired) ได้นำหลักการทางด้านการสร้างภาษาธรรมชาติมาใช้ใน AAC เพื่อสร้างประโยค รายงานกิจกรรมที่เด็กทำในโรงเรียนให้พ่อแม่ที่บ้านทราบ ผ่านหน้าจอแท็บเล็ตที่เด็กสามารถบอกอารมณ์และสถานที่ทำกิจกรรม ร่วมกันกับตัวเซ็นเซอร์ที่ติดไว้กับรถเข็นของเด็กและสถานที่ต่าง ๆ ในโรงเรียน แต่ซอฟต์แวร์

นี้สามารถแต่งประโยคได้อย่างจำกัด เนื่องจากมีการแทนที่คำเฉพาะในโครงสร้างต้นแบบที่กำหนดไว้เท่านั้น (เติมคำในช่องว่าง) และสามารถใช้ได้ในสถานที่และขอบเขตที่จำกัดคือโรงเรียน หรือสถานที่ที่มีตัวเซ็นเซอร์ติดตั้งไว้เท่านั้น ต่อมาเพื่อความสะดวกสบายในการใช้งานมีการพัฒนา Multilingual AAC on Android [26] ซึ่งเป็น AAC ที่พัฒนาบนระบบปฏิบัติการแอนดรอยด์ ผู้ใช้งานสามารถพกพาได้ง่าย มีหลายภาษา และมีระบบสังเคราะห์เสียงใช้แทนเสียงพูดสำหรับผู้ที่ไม่สามารถสื่อสารได้ แต่ซอฟต์แวร์นี้ยังมีข้อจำกัดในเรื่องของลำดับการเรียงรูปภาพเช่นกัน คือผู้ใช้งานจะต้องเป็นผู้เรียงลำดับรูปภาพด้วยตนเอง จึงสามารถใช้ได้เฉพาะผู้ใช้ที่มีความสามารถในการใช้ภาษาเท่านั้น

จากงานวิจัย AAC ที่กล่าวมาทั้งหมดข้างต้น AAC เป็นเครื่องมือที่ช่วยในการสื่อสารที่ผู้ใช้จำเป็นต้องมีความสามารถในการจัดเรียงลำดับรูปภาพหรือคำ เนื่องจากตัว AAC ไม่สามารถจัดเรียงลำดับแบบอัตโนมัติให้แก่ผู้ใช้งานได้ ดังนั้นผู้วิจัยจึงได้เล็งเห็นถึงความสำคัญของข้อจำกัดในเรื่องของลำดับการเรียงรูปภาพหรือสิ่งที่ต้องการป้อนเข้าไปนี้ เพื่อขยายกลุ่มของผู้ใช้งาน กล่าวคือนอกจากผู้ที่มีปัญหาทางด้านการสื่อสารแล้ว ผู้ที่มีปัญหาทางการกระบวนการจัดการหรือเรียงภาษา ยังสามารถใช้งาน AAC นี้ได้มีประสิทธิภาพมากขึ้นด้วย ดังนั้นผู้วิจัยจึงได้คิดและพัฒนาอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทย โดยใช้เทคนิคแบบจำลองความน่าจะเป็น เพื่อนำไปประยุกต์ใช้ในงานทางด้าน AAC

ผู้วิจัยได้ศึกษางานวิจัยทางด้านการจัดเรียงคำเพื่อสร้างภาษาธรรมชาติ โดยมีงานวิจัยของ Slayden, Hwang และ Schwartz [27, 28] ที่ทำวิจัยเกี่ยวกับการแปลภาษาระหว่างภาษาไทยและภาษาอังกฤษ โดยงานวิจัยทั้งสองนี้ ได้นำหลักการแบบจำลองภาษาและแบบจำลองค่าความน่าจะเป็นเอ็นแกรมมาใช้ เพื่อหาจุดสิ้นสุดของประโยคภาษาไทย การตัดคำในประโยคภาษาไทย และนำสิ่งเหล่านี้มาใช้เพื่อทำการแปลงคำในประโยคภาษาไทยไปเป็นคำภาษาอังกฤษ จากนั้นจะทำการเรียงประโยคใหม่โดยอาศัยกฎการแปลงตามความหมายและหลักไวยากรณ์ทางภาษา

นอกจากนี้ยังมีหลายงานวิจัยที่ได้นำหลักของแบบจำลองเอ็นแกรมและแบบจำลองค่าความน่าจะเป็นมาใช้ในงานวิจัย อาทิเช่น งานวิจัยของ Ratllaparkhi [29] ซึ่งได้นำหลักการทำงานของแบบจำลองเอ็นแกรมและแบบจำลองค่าความน่าจะเป็นมาใช้สร้างรายงานข้อมูลเกี่ยวกับการเดินทางบนอากาศ โดยงานวิจัยนี้จะนำเข้าข้อมูลเป็นกลุ่มของข้อความและแสดงผลออกมาเป็นข้อความแสดงรายงานการเดินทางของสายการบินต่าง ๆ ต่อมาในงานวิจัยที่มีชื่อว่า Non-Syntactic Word Prediction for AAC [30] ได้นำหลักการทำงานของแบบจำลองทางความหมาย (Semantic Grams) เพื่อเสนอคำที่มีความสัมพันธ์กับคำที่อยู่ใกล้กันจากข้อมูลที่มีจำกัด และงานวิจัยของ Liu [31] ได้นำหลักการทำงานของแบบจำลองเอ็นแกรมมาใช้ร่วมกับงานทางการการแปลงภาษา โดยขอบเขตของงานวิจัยนี้คือแปลงภาษาจีนเป็นภาษาอังกฤษ นอกจากนี้ในงานเกี่ยวกับการสร้างแบบจำลอง

English-Chinese translation [32] หรือแบบจำลองการแปลงภาษาจีนเป็นภาษาอังกฤษ ซึ่งได้นำหลักการของตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (Weighted Finite-State Transducer) มาใช้เพื่อเพิ่มประสิทธิภาพของระบบการแปลงภาษา และงานวิจัยของ Povey, Hannemann และ et al. [33] ได้นำหลักการตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักมาใช้มาคำนวณหาเส้นทางคำตอบที่ดีที่สุดเพื่อเลือกเป็นคำตอบของงานวิจัย

จากงานวิจัยที่กล่าวมาข้างต้นนี้ ผู้วิจัยได้นำหลักการคำนวณทางสถิติโดยอาศัยแบบจำลองภาษาและแบบจำลองค่าความน่าจะเป็นเอ็นแกรมมาใช้ในงานวิจัย เพื่อเรียงคำในประโยคภาษาไทยร่วมกับกฎการเรียงลำดับคำในประโยคตามหลักไวยากรณ์ทางภาษาไทย โดยได้นำทฤษฎีแบบจำลองเอ็นแกรมมาสร้างแบบจำลองของคำ แบบจำลองของหมวดหมู่หลัก แบบจำลองของหมวดหมู่ย่อย และแบบจำลองของหน้าที่ของคำ ซึ่งช่วยในการวิเคราะห์ชุดข้อมูล ร่วมกับหลักการตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักเพื่อหาประโยคคำตอบที่ดีที่สุดจากอัลกอริทึมและระบบที่นำเสนอ

จากการศึกษางานวิจัยที่เกี่ยวกับการสร้างคลังข้อความ พบว่ามีแนวทางการสร้างคลังข้อความด้วยกัน 2 ลักษณะด้วยกัน คือ การนำข้อมูลมาจากแหล่งข้อมูลอื่นหลาย ๆ แหล่ง แล้วผ่านกระบวนการเลือกข้อมูลด้วยวิธีการต่างๆ เพื่อสร้างคลังข้อความขึ้นมาใหม่ เช่น เลือกตามคำหรือประโยคที่ต้องการนำมาใช้ ตัวอย่าง คลังข้อความลักษณะนี้ได้แก่ ORCHID [34], LOTUS [35] และ Methods of Sentences Selection for Read-Speech Corpus Design [36] การสร้างคลังข้อความอีกลักษณะหนึ่ง คือการสร้างประโยคขึ้นมาใหม่จากคำที่มีจำกัด เพื่อให้ได้คลังข้อความที่หลากหลายและเพิ่มขนาดของคลังข้อความได้ เช่นงานวิจัย On the Generation of English Sentences [37] โดยอาศัยกฎทางไวยากรณ์มาใช้สร้างประโยคในคลังข้อความ โดยในงานวิจัยนี้ได้ผู้วิจัยได้สร้างคลังข้อความขึ้นมาใหม่โดยอาศัยหลักการสร้างคลังข้อความทั้ง 2 ลักษณะ กล่าวคือ ทั้งมีการนำข้อความจากแหล่งข้อความที่มีอยู่ และสร้างข้อความขึ้นมาใหม่ มารวมกันเป็นคลังข้อความใหม่เพื่อให้เหมาะสมกับงานวิจัยนี้

การระบุหน้าที่ของคำในคลังข้อความ งานวิจัยเกี่ยวกับการระบุหน้าที่ของคำตามแนวคิดการสร้างคลังข้อความ Royal-Institute-Based Thai-Part-of-Speech (RIPOS) [38] ได้อ้างอิงการระบุหน้าที่ของคำตามงานวิจัยที่สร้างคลังข้อความที่มีชื่อว่า ORCHID [34] โดยในงานวิจัยนี้ได้จัดหมวดหมู่ตามหน้าที่ของคำในประโยคขึ้นมาใหม่ ซึ่งแบบเป็น 17 แบบตามหน้าที่ของคำในประโยค โดยในงานวิจัยนี้ผู้พัฒนาได้ทำการแบ่งหน้าที่ของคำในประโยคขึ้นมาใหม่โดยใช้การแบ่งหน้าที่ของคำแบบ RIPOS เป็นระบบอ้างอิง สำหรับการแบ่งหมวดหมู่ของคำ มีหลายงานวิจัยที่ได้เสนอแนวทางการแบ่งหมวดหมู่ของคำ เช่นงานวิจัย [39, 40] สำหรับงานวิจัยนี้ได้แบ่งประเภทหมวดหมู่ของคำเป็นหมวดหมู่หลักทั้งหมด 12 หมวดหมู่ ตามแนวคิดการใช้คำกระตุ้นการพูดและการคิดของ Bell [8] ซึ่ง

เธอเชื่อคำกล่าวของ Albert Einstein ที่ว่า “If I can’t picture it, I can’t understand it.” และได้นำแนวคิดนี้ไปช่วยสอนเด็กเพื่อฝึกให้เด็กคิดวิเคราะห์สร้างเรื่องราวจากรูปภาพที่เด็กเห็น โดยการตั้งคำถาม ตัวอย่างคำถามที่ผู้สอนถามเด็ก เช่น “What should I picture for the boy’s pants? Does he have on blue pants or red pants? Are they long pants or short pants?” ซึ่งสอดคล้องกับทางคณะกรรมการการประถมศึกษาแห่งชาติ และพฤติกรรมการสอนทางวิทยาศาสตร์ [41, 42] ที่ได้กล่าวถึงแนวคิดในการใช้คำถามสำหรับเด็กปฐมวัยว่า คำถามเป็นเครื่องมือที่สำคัญยิ่งในการจัดกิจกรรมเตรียมความพร้อมที่ยึดเด็กเป็นศูนย์กลาง เพราะคำถามจะเป็นเครื่องกระตุ้นให้เด็กเกิดการคิด และสนใจต่อสื่อและสิ่งต่าง ๆ ที่อยู่รอบ ๆ ตัว คำถามซึ่งเกิดจากคำกระตุ้นความคิดเหล่านี้แบ่งออกเป็น 12 หมวดหมู่ ได้แก่ อะไร (What), ขนาด (Size), สี (Color), จำนวน (Number), รูปร่าง (Shape), สถานที่ (Where), เวลา (When), บรรยากาศ (Background), การกระทำ (Movement), อารมณ์ (Mood), ทัศนียภาพ (Perspective) และเสียง (Sound) โดยในงานวิจัยนี้ได้นำคำกระตุ้นการตั้งคำถามทั้ง 12 หมวดหมู่นี้ มาประยุกต์ใช้กับการแบ่งประเภทหมวดหมู่ของคำ เพื่อใช้เป็นแนวทางในการเรียงลำดับคำในประโยคความเดียวภาษาไทย

และในงานวิจัยการศึกษาโครงสร้างวลี ประโยค และสัมพันธสารของเด็กปกติและเด็กออทิสติก [18] พบว่า มีประโยคความเดียว หรือประโยคสามัญ ปรากฏในเรื่องเล่าของเด็กปกติ คิดเป็นร้อยละ 71.80 จากจำนวนประโยคทั้งหมด 468 ประโยค และมีประโยคความเดียว หรือประโยคสามัญ ปรากฏในเรื่องเล่าของเด็กออทิสติก คิดเป็นร้อยละ 84.02 จากจำนวนประโยคทั้งหมด 388 ประโยค จากผลการศึกษาข้างต้นสอดคล้องกับพฤติกรรมของเด็กออทิสติกและเด็กที่มีความบกพร่องทางการเรียนรู้บางกลุ่ม เช่น เด็กที่มีปัญหาทางด้านการจัดเรียง เป็นต้น โดยผู้วิจัยได้ทำการศึกษาสอบถาม และค้นคว้าถึงความสามารถในการแต่งประโยคและเล่าเรื่องราวของเด็กที่มีความบกพร่องทางการเรียนรู้ พบว่าเด็กเหล่านี้มักมีปัญหาการเรียงลำดับของคำในประโยค ทำให้ไม่สามารถแต่งประโยคภาษาไทยที่มีความยาวมากได้ กล่าวคือ มักจะสร้างประโยคและเล่าเรื่องราวเป็นประโยคความเดียว ในงานวิจัยนี้จึงได้กำหนดขอบเขตของงานวิจัยที่จะพัฒนาอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยเฉพาะประโยคความเดียวเท่านั้น

จากงานวิจัยที่กล่าวมาทั้งหมดข้างต้น ผู้วิจัยจึงได้นำแนวคิดของงานวิจัยทางด้านต่าง ๆ มาประกอบกัน เพื่อพัฒนาอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทย โดยอาศัยแบบจำลองความน่าจะเป็นของประเภทคำกระตุ้นการพูด เพื่อช่วยให้เด็กออทิสติกและเด็กที่มาความบกพร่องทางการเรียนรู้บางกลุ่มสามารถเรียงลำดับคำในประโยค และสร้างเรื่องราวที่ต้องการสื่อสารได้ง่ายขึ้น โดยอาศัยแบบจำลองความน่าจะเป็นของประเภทคำกระตุ้นการพูดตามแนวคิดของ Bell

บทที่ 3
ขั้นตอนการดำเนินงานวิจัยและการออกแบบระบบที่นำเสนอ

ในบทนี้จะกล่าวถึงขั้นตอนการดำเนินงานวิจัยและการออกแบบระบบที่นำเสนอ ซึ่งประกอบไปด้วย ระบบการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอโดยอาศัยหลักการและลักษณะสำคัญเอ็นแกรม และระบบการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอโดยอาศัยหลักการและลักษณะสำคัญเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก

3.1 ขั้นตอนการดำเนินงานวิจัย

สำหรับขั้นตอนการพัฒนาอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยในงานวิจัยนี้ เริ่มจากการศึกษาความรู้เบื้องต้นที่เกี่ยวข้องกับงานวิจัยนี้ ซึ่งกล่าวไว้ในบทก่อนหน้า จากนั้นจะต้องศึกษาเครื่องมือที่จะนำมาใช้ในงานวิจัยนี้ และเริ่มออกแบบระบบที่จะนำเสนอ โดยในบทนี้จะกล่าวถึงขั้นตอนการออกแบบระบบ ส่วนรายละเอียดเรื่องอื่น ๆ ส่วนการทดลอง การวัดผลและวิเคราะห์ผลทดลอง และสรุปผลการทดลองจะกล่าวไว้ในบทต่อ ๆ ไป

3.2 การออกแบบระบบที่นำเสนอ

การออกแบบระบบของงานวิจัยนี้ ได้มาจากการศึกษางานวิจัยและทฤษฎีที่เกี่ยวข้องกับงานวิจัยนี้ซึ่งได้กล่าวไว้ในบทที่ 2 จากนั้นทำการศึกษาเครื่องมือ และวิเคราะห์องค์ความรู้ที่ได้จากการศึกษาเบื้องต้น โดยระบบที่ได้ออกแบบในงานวิจัยนี้นั้นจะแบ่งออกเป็น 4 ส่วนหลัก ได้แก่

1. ชุดคำที่ป้อนเข้าสู่ระบบ

สำหรับชุดคำที่ป้อนเข้าสู่ระบบนั้น จะเป็นชุดคำที่มีจำนวนคำตั้งแต่ 3-6 คำ และแต่ละคำจะมีการระบุหมวดหมู่หลัก หมวดหมู่ย่อย และหน้าที่ของคำนั้น ๆ โดย 1 คำ จะมีเพียง 1 ความหมายเท่านั้น

2. แบบจำลองของหลักไวยากรณ์ภาษาไทย

แบบจำลองของหลักไวยากรณ์ภาษาไทยนี้จะทำการคัดกรองประโยคที่ได้จากการนำคำทั้งหมดมาจัดเรียงเป็นประโยคและคัดเลือกเฉพาะประโยคที่มีการเรียงลำดับหน้าที่ของคำถูกต้องตามหลักไวยากรณ์ภาษาไทยให้ผ่านเข้าสู่ระบบการคำนวณคะแนนและเลือกประโยคคำตอบต่อไป

3. ระบบคำนวณคะแนนและเลือกประโยคที่นำเสนอ

สำหรับระบบคำนวณคะแนนและเลือกประโยคคำตอบนั้นจะพิจารณาประโยคที่เป็นไปได้ที่ผ่านแบบจำลองของหลักไวยากรณ์ภาษาไทย และผ่านอัลกอริทึมที่งานวิจัยนี้นำเสนอ เพื่อหาประโยคที่เหมาะสมที่สุดมาเป็นประโยคคำตอบ

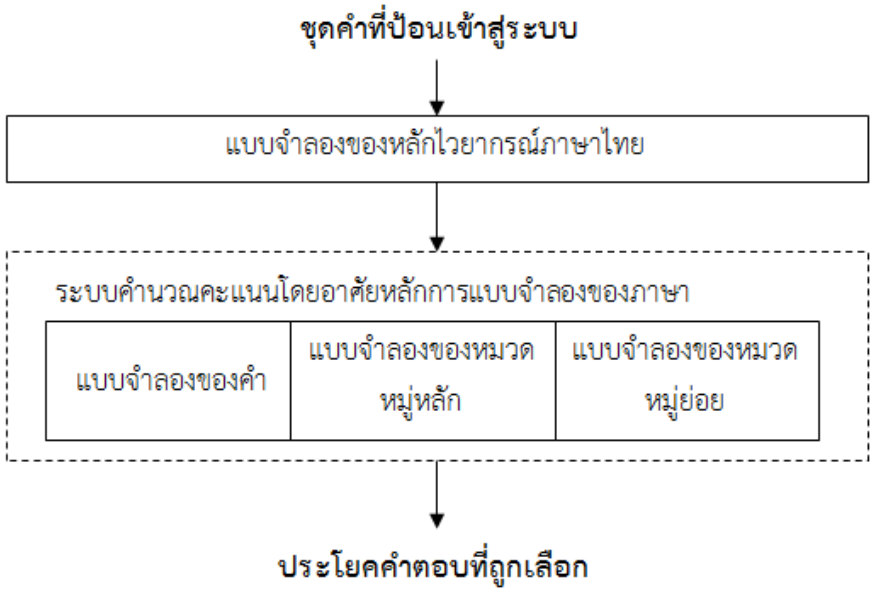
4. ประโยคคำตอบที่ถูกเลือก

ประโยคคำตอบที่ได้จากระบบนี้ จะเป็นประโยคความเดียวภาษาไทยที่มีความยาวเท่ากับจำนวนคำที่ป้อนเข้าสู่ระบบ เป็นประโยคความเดียวภาษาไทยที่ถูกต้องตามหลักไวยากรณ์ภาษาไทย และเป็นประโยคที่ระบบคิดว่ามีความหมายเหมาะสมและดีที่สุด

โดยในงานวิจัยนี้ได้ออกแบบระบบที่นำเสนอออกเป็น 2 ระบบ ได้แก่

3.2.1 ระบบการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอโดยอาศัยหลักการและลักษณะสำคัญเอ็นแกรม

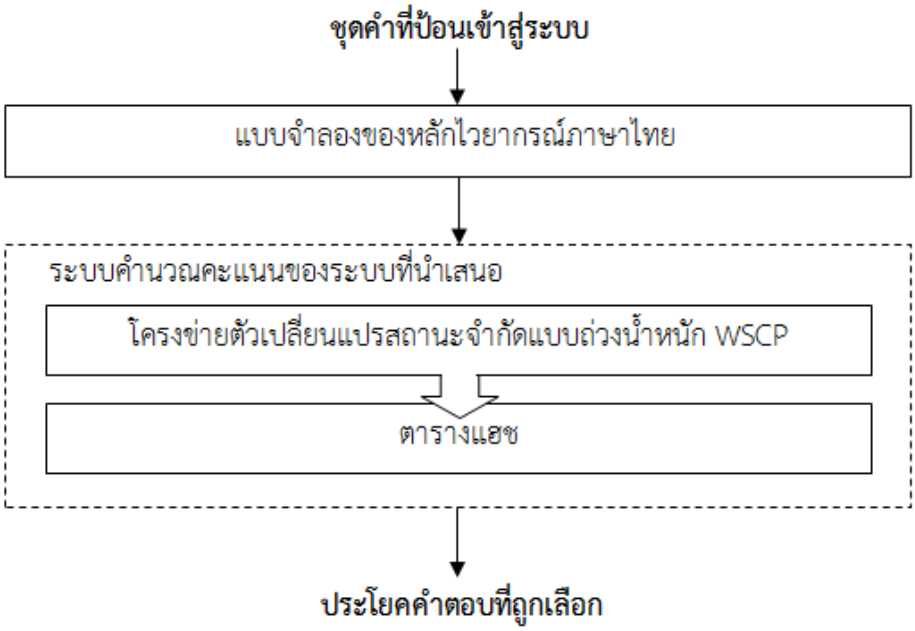
การออกแบบระบบการเรียงลำดับคำในประโยคภาษาไทยโดยอาศัยหลักการการเอ็นแกรมสามารถนำเสนอได้ดังภาพที่ 3-1



ภาพที่ 3-1 โครงสร้างการทำงานของระบบที่นำเสนอเมื่ออาศัยหลักการและลักษณะสำคัญเอ็นแกรม

3.2.2 ระบบการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอโดยอาศัยหลักการและลักษณะสำคัญเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (Weighted Finite-State Transducer: WFST)

การออกแบบระบบการเรียงลำดับคำในประโยคภาษาไทยโดยอาศัยหลักการการเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก สามารถนำเสนอได้ดังภาพที่ 3-2



ภาพที่ 3-2 โครงสร้างการทำงานของระบบที่นำเสนอเมื่ออาศัยหลักการและลักษณะสำคัญเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก

บทที่ 4
การออกแบบและสร้างคลังข้อความ

ในบทนี้กล่าวถึงการออกแบบและสร้างคลังข้อความที่ใช้ในงานวิจัยและระบบที่นำเสนอ รวมไปถึงการระบุหมวดหมู่หลัก หมวดหมู่ย่อย และหน้าที่ของคำ โดยแบ่งการระบุหน้าที่ของคำออกเป็น 2 แบบ ได้แก่ระบุหน้าที่ของคำตามวิธีที่ใช้อ้างอิง และระบุตามวิธีที่งานวิจัยนี้นำเสนอ

4.1 การเตรียมและเลือกใช้คลังข้อความ

ในงานวิจัยนี้จำเป็นต้องใช้คลังข้อความในการฝึกฝนหาค่าความน่าจะเป็นของเกิดร่วมกันของคำในประโยคเพื่อนำค่าความน่าจะเป็นที่ได้มาใช้ในการสร้างประโยค โดยคลังข้อความที่จะนำมาใช้ในการฝึกฝนข้อมูลในงานวิจัยนี้นั้น ต้องมีข้อจำกัดและมีลักษณะดังนี้

1. คลังข้อความที่จะนำมาใช้ต้องบรรจุข้อมูลประเภทข้อความภาษาไทย
2. ข้อความในคลังข้อความต้องมีการตัดคำเรียบร้อยแล้ว
3. ลักษณะของข้อความในคลังข้อความต้องเป็นประโยค
4. เนื้อหาภายในคลังข้อความควรเป็นข้อความที่ใช้ทั่วไปใช้ชีวิตประจำวัน
5. คำในคลังข้อความทุกคำต้องมีการระบุหมวดหมู่หลัก หมวดหมู่ย่อย และหน้าที่ของคำตามแนวคิดที่นำเสนอ

จุฬาลงกรณ์มหาวิทยาลัย

โดยในเบื้องต้นงานวิจัยนี้ได้เลือกใช้คลังข้อความที่มีอยู่แล้ว มีการตัดคำในคลังข้อความ และเปิดให้องค์กรทางการศึกษาสามารถใช้งานได้ คลังข้อความดังกล่าวมีชื่อว่า Benchmark for Enhancing the Standard for Thai language Processing (BEST2010) [43, 44] ซึ่งเป็นคลังข้อความภาษาไทยที่มีการกำกับขอบเขตของคำ ขนาด 5 ล้านคำ ประกอบด้วย 4 หมวด คือ บทความวิชาการ (Article) สารานุกรม (Encyclopedia) ข่าว (News) และ นวนิยาย (Novel) โดยในงานวิจัยนี้ใช้คลังข้อความ BEST2010 ในหมวดหมู่นวนิยายเท่านั้น เนื่องจากเป็นข้อความในหมวดหมู่นี้ เป็นข้อความที่คล้ายกับข้อความที่ใช้สื่อสารจริงในชีวิตประจำวัน ซึ่งประกอบไปด้วยคำจำนวน 116,873 คำ แต่คำในคลังข้อความ BEST2010 นั้นไม่ได้มีการระบุหมวดหมู่และหน้าที่ของคำ และประโยคที่อยู่ในคลังข้อความ BEST2010 นี้ ไม่ได้มีการแบ่งประโยคที่ชัดเจน กล่าวคือข้อความในคลังข้อความนี้จะเรียงต่อกันเป็นเรื่องราว ไม่ได้แบ่งเป็นประโยคทีละประโยค และประโยคที่ปรากฏในคลังข้อความนี้มีหลายประเภท ไม่ได้เป็นประโยคความเดียวทั้งหมด

ด้วยเหตุนี้ในงานวิจัยนี้จึงมีแนวคิดที่จะออกแบบและการสร้างคลังข้อมูลใหม่ที่มีความสอดคล้องและเหมาะสมกับการใช้งานบนซอฟต์แวร์ AAC สำหรับเด็กออทิสติกและเด็กที่มีความบกพร่องทางการเรียนรู้กลุ่มที่เป็นเป้าหมายหมายของงานวิจัยนี้ เพื่อให้อัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยสามารถทำงานได้อย่างมีประสิทธิภาพมากยิ่งขึ้น โดยให้ชื่อคลังข้อความใหม่นี้ว่า Thai Elementary-Level story-telling with Simple sentences หรือ คลังข้อความ TELL-S

4.2 การออกแบบและสร้างคลังข้อความ TELL-S

คลังข้อความ TELL-S [45] เป็นคลังข้อความภาษาไทยที่ประกอบด้วยประโยคความเดียวทั้งหมดที่ได้มาจากการนำประโยคที่ปรากฏอยู่ในหนังสือเรียนภาษาไทยในระดับประถมศึกษาปีที่ 1 และระดับประถมศึกษาปีที่ 2 ทั้งหมดมาแยกเป็นประโยคความเดียวภาษาไทย ซึ่งประกอบไปด้วยคำจำนวน 3,289 คำ และมีจำนวนประโยค 1,023 ประโยค ดังตารางที่ 4-1

ตารางที่ 4-1 จำนวนประโยคในคลังข้อความ TELL-S โดยแบ่งตามความยาวของประโยค

จำนวนคำในประโยค	**จำนวนประโยค**	**เปอร์เซ็นต์ (%)**
2	147	14.37
3	587	57.38
4	238	23.26
5	38	3.81
6	12	1.17
รวม	**1,023**	**100.00**

โดยในการออกแบบและสร้างคลังข้อความ TELL-S ได้ออกแบบให้คลังข้อความนี้ มีลักษณะตามข้อจำกัดที่จำเป็นก่อนจะทำการฝึกฝนข้อมูลในคลังข้อความ ได้แก่ ต้องเป็นคลังข้อความที่บรรจุข้อมูลประเภทข้อความภาษาไทย ข้อความในคลังข้อความต้องมีการตัดคำเรียบร้อยแล้ว ลักษณะของข้อความในคลังข้อความต้องเป็นประโยคที่ใช้ทั่วไปใช้ชีวิตประจำวันทั่วไป และคำในคลังข้อความทุกคำมีการระบุหมวดหมู่หลัก หมวดหมู่ย่อย และหน้าที่ของคำ ตามแนวคิดที่นำเสนอ

และในงานวิจัยนี้ได้สร้างคลังข้อความ TELL-S 2 ลักษณะด้วยกัน ได้แก่ คลังข้อความ TELL-S ที่เก็บข้อความประโยคความเดียวภาษาไทย และคลังข้อความ TELL-S ที่มีการแปลงคำ

ภาษาไทยทุกคำ เป็นคำภาษาอังกฤษที่ทับศัพท์คำภาษาไทย เนื่องจากเครื่องมือที่จำเป็นต้องใช้ในงานวิจัยซึ่งจะกล่าวถึงในบทถัดไปไม่สามารถรองรับข้อความภาษาไทยได้ ดังนั้นคลังข้อความ TELL-S ที่ใช้ในระบบที่นำเสนอนี้จะเป็นคลังข้อความภาษาอังกฤษที่ทับศัพท์ภาษาไทย ดังภาพที่ 4-1

```
1  wairoon thunggroup thing khaya bon tanon suaysuay khongsuansat
2  pheunpheun len football gub baibok took wanyhood tornbai
3  pheunpheun len football gub baibua took wanyhood tornbai
4  baibua len football thee rimtaling took wanyhood tornbai
5  baibok len football thee rimtaling took wanyhood tornbai
6  phooyai eao khong long jark lhang khongpangcone
7  chaw wannee tookkon ma thee baan khongpeesompoi
8  phor khongnamsai lhaing aisacream thee raan klaiklai
9  saingchang rong earkeark akak dung lun lamtarn
10 tornchao thee suansad me kon mai mak
11 tookkon yon hueng thee nhaa baan
12 wanrungkheun loongwan shong dekdek thee baan
13 phor khongdek thungsong yib but khang
14 phor khongphupha kui gub phor khongnamsai
15 phor gub phupha kee koe khongplaimapin
16 tonmai soong yhai me bai kwang
17 chaw wannee chaokhongchang prachum thee lanwud
18 chaw wannee loongkamnun prachum thee lanwud
19 baibok gin nam thee ang leklek
20 baibua gin nam thee ang leklek
21 phor jub pla nai beung lhaytue
22 tornsai khabaunchang mathueng joodnadphob thee mhoobarnchaipa
```

ภาพที่ 4-1 ตัวอย่างข้อมูลในคลังข้อความ TELL-S ที่แปลงเป็นคำทับศัพท์ภาษาอังกฤษ

4.3 การระบุหน้าที่ของคำในคลังข้อความ

วิธีการระบุหน้าที่ของคำในคลังข้อความ ในงานวิจัยนี้จะมีการระบุหน้าที่ของคำ 2 แบบ คือ แบบที่หนึ่งเป็นการระบุหน้าที่ของคำตามระบบอ้างอิง และแบบที่สองเป็นการระบุคำและหน้าที่ของคำตามแนวคิดที่งานวิจัยนี้นำเสนอ เพื่อเปรียบหาว่าวิธีการระบุหน้าที่ของคำแบบใดที่เหมาะสมกับงานวิจัยนี้มากที่สุด

4.3.1 การระบุหน้าที่ของคำตามแนวคิดการสร้างคลังข้อความ Royal-Institute-Based Thai-Part-of-Speech (RIPOS)

คลังข้อความ RIPOS [38] สร้างขึ้นโดยอ้างอิงการระบุหน้าที่ของคำตามงานวิจัยที่สร้างคลังข้อความที่มีชื่อว่า ORCHID [34] โดยในงานวิจัยนี้ได้จัดหมวดหมู่ตามหน้าที่ของคำในประโยคขึ้นมาใหม่ ดังตารางที่ 4-2

ตารางที่ 4-2 การระบุหน้าที่ของคำ (POS) ตามแนวคิดการสร้างคลังข้อความ RIPOS

คำที่ใช้ระบุ	คำอธิบาย	ตัวอย่าง
NCM	Common noun	บ้านฉันอยู่ใกล้โรงเรียน
NNN	Number	รถสองคัน
NNC	Classifier	รถสองคัน
NNP	Proper noun	กรุงเทพเป็นเมืองหลวง
NNT	Title noun	หม่อมราชวงศ์พุฒิภัทร
PRN	Pronoun	ฉันทานข้าว
PREP	Preposition	เพื่อนนั่งในแถวหน้า
CONJ	Conjunction	ฉันและเธอเป็นนักเรียน
INJ	Interjection	ไชโย สอบผ่านแล้ว
VBG	General Verb	ฉันทานข้าว
VBX	Auxiliary Verb	ฉันจะดูหนัง
WSG	General wisate	ฉันดื่มน้ำเย็น
WSQ	Questioning wisate	ราคาเท่าไหร่
WSD	Determining wisate	รถนี่แพง
WSN	Negation wisate	ผมไม่ดื่มเหล้า
WSE	Ending wisate	เรียนสนุกดีครับ
PUNC	Punctuation	ทำไม?

4.3.2 การระบุหน้าที่ของคำตามแนวคิดที่นำเสนอ

การระบุหน้าที่ของคำตามแนวคิดที่นำเสนอนี้จะมีการระบุคำเป็นหมวดหมู่หลัก (Class) หมวดหมู่ย่อย (Subclass) และหน้าที่ของคำในประโยค (Part of speech) ในการระบุคำตามหมวดหมู่หลักนั้น อ้างอิงจากการแบ่งหมวดหมู่ตาม 12 คำถามเพื่อสร้างรูปภาพในงานวิจัยของ Bell ดังนั้นคำ 1 คำ จะมีการระบุหมวดหมู่หลัก หมวดหมู่ย่อย และหน้าที่ของคำในประโยค โดยมีเงื่อนไขที่ว่าคำหนึ่ง ๆ จะประกอบด้วย 1 หมวดหมู่หลัก 1 หมวดหมู่ย่อย 1 หน้าที่ของคำในประโยค และมีเพียง 1 ความหมายเท่านั้น รายละเอียดหมวดหมู่หลักและหมวดหมู่ย่อยดังตารางที่ 4-3 และตัวอย่างการระบุหน้าที่ของคำในคลังข้อความแสดงดังตารางที่ 4-4

ตารางที่ 4-3 หมวดหมู่หลักและหมวดหมู่ย่อยที่ใช้ระบุหน้าที่ของคำในประโยค

ลำดับที่	หมวดหมู่หลัก	หมวดหมู่ย่อย
1	WHAT	Body
		Household
		Family
		Food
		Head
		Cloth
		Sport
		Stationary
		Pronoun
		Day
		Month
		Color
		Transportation
2	SIZE	Quantity (ขยายนาม)
		Speed (ขยายนาม)
		Size (ขยายนาม)
		Quantity (ขยายกริยา)
		Speed (ขยายกริยา)
		Size (ขยายกริยา)
3	COLOR	Color
4	NUMBER	Number
5	SHAPE	Shape
6	WHERE	Place
		Position (ขยายนาม)
		Position (ขยายกริยา)
7	WHEN	Day (ขยายนาม)
		Month (ขยายนาม)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ลำดับที่	หมวดหมู่หลัก	หมวดหมู่ย่อย
		Time (ขยายนาม)
		Day (ขยายกริยา)
		Month (ขยายกริยา)
		Time (ขยายกริยา)
8	BACKGOUND	Weather
9	MOVEMENT	Verb1 (กริยาที่กระทำตั้งแต่เอวลงไป)
		Verb2 (กริยาที่กระทำตั้งแต่ลำคอลงไปถึงเอว)
		Verb3 (กริยาที่กระทำบริเวณใบหน้า)
10	MOOD	Emotion
11	PERSPECTIVE	Perspective (ขยายนาม)
		Perspective (ขยายกริยา)
12	SOUND	Sound (ขยายนาม)
		Sound (ขยายกริยา)

จุฬาลงกรณ์มหาวิทยาลัย

ตารางที่ 4-4 ตัวอย่างการระบุหน้าที่ของคำตามหมวดหมู่ต่าง ๆ

ตัวอย่างคำ	ระบุหมวดหมู่หลัก	ระบุหมวดหมู่ย่อย	หน้าที่ของคำ (Part of Speech)
ฉัน, คุณ	WHAT	PRONOUN	N
โบสถ์, โรงเรียน	WHAT	PLACE	N
หนึ่ง, สอง	SIZE	TIME	ADJ or ADV
ข้างบน, ข้างใน	WHERE	POSITION	ADJ or ADV
โกรธ, มีความสุข	MOOD	EMOTION	V

บทที่ 5
ขั้นตอนการสร้างระบบและอัลกอริทึมที่นำเสนอ

การจะสร้างระบบการเรียงลำดับคำในประโยคภาษาไทยให้มีประสิทธิภาพนั้น จำเป็นที่จะต้องมีการออกแบบระบบและเตรียมชุดข้อมูลฝึกฝนที่ดีและเหมาะสมกับระบบเสียก่อน ดังที่ได้กล่าวไว้ในบทก่อนหน้า โดยในบทนี้จะกล่าวถึงขั้นตอนการสร้างระบบ แบบจำลองต่าง ๆ ที่ต้องใช้ในระบบ และอัลกอริทึมที่นำเสนอในงานวิจัยนี้ รวมถึงคำนวณคะแนนและเลือกประโยคคำตอบของระบบ

5.1 ขั้นตอนการสร้างระบบที่นำเสนอ

ระบบการเรียงลำดับคำในประโยคภาษาไทยมีขั้นตอนการสร้างระบบดังภาพที่ 5-1

ภาพที่ 5-1 ลำดับขั้นตอนการสร้างระบบ

จากภาพที่ 5-1 จะกล่าวถึงขั้นตอนการสร้างระบบ ซึ่งในขั้นตอนการออกแบบระบบ การสร้างคลังข้อความและการระบุหน้าที่ของคำในคลังข้อความนั้น ได้กล่าวไว้ในบทที่ 3 และบทที่ 4 ตามลำดับ ในบทนี้จะกล่าวถึงวิธีการสร้างแบบจำลองของหลักไวยากรณ์ภาษาไทยตามแนวคิดที่นำเสนอ การสร้างแบบจำลองของภาษา อัลกอริทึมที่นำเสนอ และวิธีการสร้างระบบคำนวณคะแนนและการเลือกประโยคคำตอบของระบบ

5.2 การสร้างแบบจำลองของหลักไวยากรณ์ภาษาไทยตามแนวคิดที่นำเสนอ

แบบจำลองของหลักไวยากรณ์ภาษาไทย เป็นแบบจำลองที่ทำหน้าที่ตรวจสอบประโยคที่ระบบสร้างขึ้นว่ามีประโยคความเดียวใดบ้างที่เรียงลำดับคำได้ถูกต้องตามหลักไวยากรณ์ภาษาไทย สำหรับประโยคความเดียวที่ถูกต้องตามหลักไวยากรณ์ภาษาไทยในงานวิจัยนี้จะเป็นประโยคที่มีภาคประธานเพียงบทเดียว และมีภาคแสดงหรือกริยาสำคัญเพียงบทเดียว กรณีการตรวจสอบหลักไวยากรณ์ภาษาไทยสามารถแบ่งได้เป็น 2 กรณี โดยเขียนตามหลักการของ Backus Normal Form [46] ได้ดังนี้

1. กรณีประโยคที่ไม่มีกรรม

<ประโยคที่ไม่มีกรรม> ::= <ภาคประธาน> <ภาคแสดง>

<ภาคประธาน> ::= <คำนาม> <ส่วนขยายนาม>

<ภาคแสดง> ::= <คำกริยา> | <คำกริยา>
<ส่วนขยายกริยา> | <คำกริยา>
<ส่วนขยายกริยา> <ส่วนขยายส่วนขยาย>

2. กรณีประโยคที่มีกรรม

<ประโยคที่ไม่มีกรรม> ::= <ภาคประธาน> <ภาคแสดง>

<ภาคประธาน> ::= <คำนาม> <ส่วนขยายนาม>

<ภาคแสดง> ::= <คำกริยา> <คำนาม> |
<คำกริยา> <คำนาม> <ส่วนขยายกริยา> |
<คำกริยา> <คำนาม> <ส่วนขยายกริยา>
<ส่วนขยายส่วนขยาย>

5.3 การสร้างแบบจำลองของภาษา

ในงานวิจัยนี้ได้แบ่งแบบจำลองของภาษาออกเป็น 3 แบบจำลอง ได้แก่ แบบจำลองของคำ แบบจำลองของหมวดหมู่หลัก และแบบจำลองของหมวดหมู่ย่อย ซึ่งแบ่งตามลักษณะการระบุหมวดหมู่และหน้าที่ของคำในคลังข้อความ ตามที่ได้กล่าวไว้ในบทที่ 4 โดยมีขั้นตอนและวิธีการสร้างแบบจำลองต่าง ๆ ดังต่อไปนี้

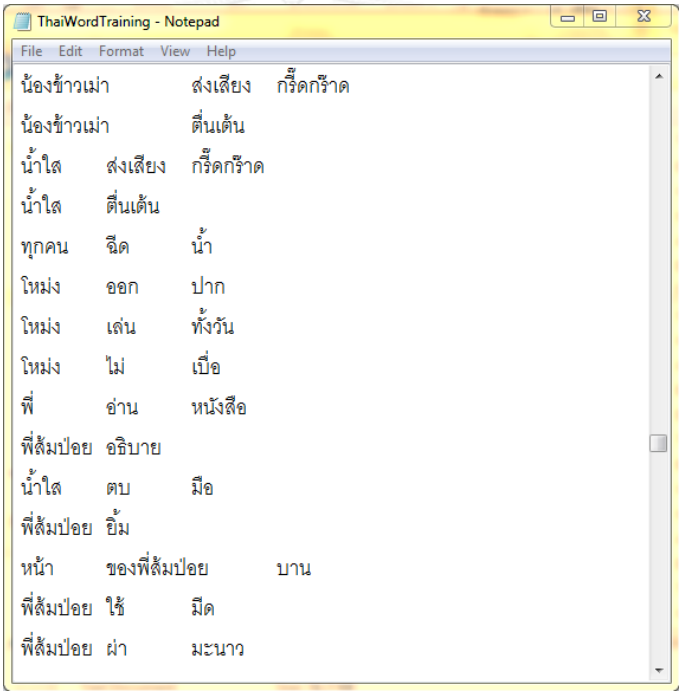
จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

5.3.1 การสร้างแบบจำลองของคำ

การสร้างแบบจำลองของคำนั้น เริ่มจากการพิจารณาลักษณะสำคัญเอ็นแกรมของคำ โดยการคำนวณค่าความน่าจะเป็นด้วยแบบจำลองเอ็นแกรม สำหรับเครื่องมือที่นำมาใช้ในการคำนวณค่าความน่าจะเป็นในงานวิจัยนี้ คือ CMU-Cambridge Statistical Language Modeling Toolkit Version 2 (CMU Toolkit V.2) [47] ซึ่งจะทำการพิจารณาเอ็นแกรมในระดับคำ ในงานวิจัยนี้ใช้การสร้างแบบจำลองเอ็นแกรมแบบไบแกรม เพื่อนำมาเป็นลักษณะสำคัญ

ขั้นตอนการคำนวณค่าความน่าจะเป็นด้วยเครื่องมือ CMU Toolkit V.2 มีดังนี้

1. เตรียมคลังข้อความภาษาไทย BEST2010 และ TELL-S โดยต้องทำการเว้นวรรคคำในคลังข้อความดังภาพ 5-2



ภาพที่ 5-2 ตัวอย่างข้อมูลของคำสำหรับใช้ในโปรแกรม CMU Toolkit V.2

2. เรียกใช้โปรแกรม CMU Toolkit V.2 สำหรับคำนวณหาค่าความน่าจะเป็นของแบบจำลองเอ็นแกรมผ่านคอมมานด์ไลน์ (Command Line) กำหนดให้ไฟล์คลังข้อความที่จะใช้นำเข้าสู่โปรแกรม CMU Toolkit V.2 มีชื่อว่า “ThaiWordTraining.txt” โดยใช้ชุดคำสั่งดังต่อไปนี้

ชุดคำสั่งที่ 1 text2wfreq.exe < ThaiWordTraining.txt > ThaiWordTraining.wfreq

ชุดคำสั่งนี้ใช้เพื่อคำนวณหาความถี่ของคำแต่ละคำที่ปรากฏอยู่ในคลังข้อความ ดังภาพ 5-3

```
5 1
7 2
อย่างรวดเร็ว 1
รูป 1
รู้ 1
ปาก 8
น้ำแข็ง 5
ของพระราชทาน 2
พยักหน้า 4
ผึ้ง 1
พลังงานก๊าซจากพืช 2
เชิญ 1
สับปะรด 2
อากาศ 3
เล็กๆ 2
รับประทาน 2
ความยาวช้าง 4
โกรธ 3
ว่า 1
น้องข้าวเม่า 7
ชีวิต 3
จุดพักช้าง 2
ของส่วนรวม 1
โป๊งเป๊ง 1
หอม 3
อย่างดีใจ 1
ทำงาน 7
```

ภาพที่ 5-3 ตัวอย่างความถี่ของคำปรากฏอยู่ในคลังข้อความในรูปแบบไฟล์.wfreq

ชุดคำสั่งที่ 2 wfreq2vocab.exe < ThaiWordTraining.wfreq > ThaiWordTraining.vocab

ชุดคำสั่งนี้ใช้เพื่อแสดงจำนวนคำทั้งหมดที่ปรากฏอยู่ในคลังข้อความ และเรียงคำทั้งหมดในคลังข้อความตามตัวอักษร ดังภาพ 5-4

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

```
## Vocab generated by v2 of the CMU-Cambridge Statistcal
## Language Modeling toolkit.
##
## Includes 886 words ##
5
7
กด
กตัญญู
กติกา
กรง
กระดิ่ง
กระทง
กระทบ
กระพรวน
กระพือ
กระเป๋า
กระเป๋าสตางค์
กริ่ด
กริ่ดกร้าด
กรุบกรอบ
กรุ่งกริ่ง
กลับ
กลัว
กลางแจ้ง
กลายเป็น
กลิ้น
```

ภาพที่ 5-4 ตัวอย่างคำที่ปรากฏอยู่ในคลังข้อความแบบเรียงตามตัวอักษรในรูปแบบไฟล์.vocab

ชุดคำสั่งที่ 3 text2idngram.exe -vocab ThaiWordTraining.vocab -idngram ThaiWordTraining.idngram < ThaiWordTraining.txt

ชุดคำสั่งที่ 4 idngram2lm.exe -idngram ThaiWordTraining.idngram -vocab ThaiWordTraining.vocab -arpa ThaiWordTraining.arpa

สำหรับชุดคำสั่งที่ 3 และ 4 จะคำนวณหาค่าความน่าจะเป็นของคำตั้งแต่ 1-แกรม 2-แกรม และ 3-แกรม ตามลำดับ ดังภาพที่ 5-5

```
\1-grams:
-3.5188 <UNK>    0.0000
-3.5188 5        -0.1615
-3.2156 7        -0.1832
-3.0395 กด       -0.1755
-3.5188 กตัญญู   -0.1438
-3.5188 กติกา    -0.1615
-3.0395 กรง      -0.1547
-2.9146 กระดิ่ง  -0.1703
-3.5188 กระทง    -0.1605
-3.5188 กระทบ    -0.1613
-3.0395 กระพรวน  -0.1751
-3.5188 กระพือ   -0.1615
-3.2156 กระเป๋า  -0.1514
-2.7385 กระเป๋าสตางค์ -0.1623
-3.5188 กรี๊ด    -0.1502
-3.2156 กรี๊ดกร๊าด -0.1492
-3.5188 กรุบกรอบ -0.1615
-3.2156 กรุ๋งกริ๋ง -0.1403
-2.9146 กลับ     -0.0921
-3.5188 กลัว     -0.1597
-3.5188 กลางแจ้ง -0.1581
-2.9146 กลายเป็น -0.1597
-3.5188 กลิ้น    -0.1612

\2-grams:
-0.5077 5 แท่ง 0.1269
-0.4625 7 สี 0.0661
-0.6386 กด ปุ่ม 0.1269
-0.9848 กด แน่น 0.0128
-0.5077 กตัญญู ภูผา -0.0321
-0.5077 กติกา ของกลุ่มเพื่อน 0.1269
-0.9848 กรง ชะนี 0.1269
-0.9848 กรง นกขุนทอง 0.1489
-0.9848 กรง เด็กๆ -0.0319
-1.1097 กระดิ่ง กระดิ่ง 0.0475
-0.7635 กระดิ่ง ดัง -0.0313
-1.1097 กระดิ่ง แกว่ง -0.0068
-0.5077 กระทง บางคน 0.3913
-0.5077 กระทบ แสงแดด 0.0387
-0.6386 กระพรวน ดัง -0.0313
-0.9848 กระพรวน ส่าย 0.1269
-0.5077 กระพือ ปีก 0.1269
-0.8087 กระเป๋า ข้างน้อย 0.0005
-0.8087 กระเป๋า ใบบัว -0.0328
-1.2858 กระเป๋าสตางค์ กระเป๋าสตางค์ -0.0116
-1.2858 กระเป๋าสตางค์ น้องข้าวเม่า -0.0149
-1.2858 กระเป๋าสตางค์ พีส้มปอย -0.0294
-1.2858 กระเป๋าสตางค์ หล่น 0.1269

\3-grams:
-1.1151 5 แท่ง ภูผา
-1.4162 7 สี น้องข้าวเม่า
-1.4162 7 สี สวยงาม
-1.4162 กด ปุ่ม ทุกคน
-1.4162 กด ปุ่ม พ่อ
-1.1151 กด แน่น น้ำแข็งใส
-1.1151 กตัญญู ภูผา เดิน
-1.1151 กติกา ของกลุ่มเพื่อน โจ๋
-1.1151 กรง ชะนี โหน
-1.1151 กรง นกขุนทอง ไม่
-1.1151 กรง เด็กๆ ทั้งหมด
-1.1151 กระดิ่ง กระดิ่ง ดัง
-0.7926 กระดิ่ง ดัง โป๊กเป๊ก
-1.1151 กระดิ่ง แกว่ง กระพรวน
-1.1151 กระทง บางคน เกี๊ยะ
-1.1151 กระทบ แสงแดด ละอองน้ำ
-0.7926 กระพรวน ดัง กรุ๋งกริ๋ง
-1.1151 กระพรวน ส่าย ภูผา
-1.1151 กระพือ ปีก พืชพับ
-1.1151 กระเป๋า ข้างน้อย โยก
-1.1151 กระเป๋า ใบบัว ร้อง
-1.1151 กระเป๋าสตางค์ กระเป๋าสตางค์ หล่น
-1.1151 กระเป๋าสตางค์ น้องข้าวเม่า พบ
```

ภาพที่ 5-5 ตัวอย่างค่าความน่าจะเป็นของคำตั้งแต่ 1-แกรม 2-แกรม และ 3-แกรม ตามลำดับในรูปแบบไฟล์.arpa

ค่าความน่าจะเป็นที่ได้จากโปรแกรม CMU Toolkit V.2 จะบอกถึงความน่าจะเป็นในการเรียงลำดับของคำในประโยคที่ปรากฏอยู่ในคลังข้อความ สามารถคำนวณได้จากสมการทางคณิตศาสตร์ [48] ต่อไปนี้

ค่าความน่าจะเป็น $P(W_1,...,W_m)$ ที่พิจารณาจากประโยค $W_1,...,W_m$ สามารถคำนวณได้ ดังสมการ 5.1

$$P(W_1,...,W_m) = \prod_{i=1}^{m} P(W_i \mid W_1,...,W_{i-1}) \approx \prod_{i=1}^{m} P(W_i \mid W_{i-(n-1)},...,W_{i-1}) \quad (5.1)$$

โดยในที่นี้กำหนดให้พิจารณาจากค่าความน่าจะเป็นตำแหน่งที่ i^{th} หรือคำตำแหน่งที่ W_i ในบริบทก่อนตำแหน่งคำที่ $i-1$

ค่าความน่าจะเป็นแบบมีเงื่อนไขสามารถคำนวณได้โดยการนับความถี่ของเอ็นแกรม ดังสมการ 5.2

$$P(W_i \mid W_{i-(n-1)},..., W_{i-1}) = \frac{count(W_{i-(n-1)},..., W_{i-1}, W_i)}{count(W_{i-(n-1)},..., W_{i-1})} \quad (5.2)$$

โดย $count(W_{i-(n-1)},...,W_{i-1},W_i)$ คือ จำนวนการพบ $W_{i-(n-1)},...,W_{i-1},W_i$ ในประโยค

$count(W_{i-(n-1)},...,W_{i-1})$ คือ จำนวนการพบ $W_{i-(n-1)},...,W_{i-1}$ ในประโยค

ในงานวิจัยนี้กำหนดให้ใช้ไบแกรมของคำ ซึ่งมีค่า n = 2 คือ ค่าความน่าจะเป็นของคำที่ปรากฏอยู่ในคลังข้อความ เมื่อพิจารณาเฉพาะคำ 2 คำที่อยู่ติดกัน

5.3.2 การสร้างแบบจำลองของหมวดหมู่ย่อย

การสร้างแบบจำลองของหมวดหมู่ย่อย เริ่มจากการพิจารณาลักษณะสำคัญเอ็นแกรมของคำที่ถูกระบุหมวดหมู่ย่อย โดยการคำนวณค่าความน่าจะเป็นด้วยแบบจำลองเอ็นแกรม สำหรับเครื่องมือที่นำมาใช้ในการคำนวณค่าความน่าจะเป็น คือโปรแกรม CMU Toolkit V.2 เช่นเดียวกับการสร้างแบบจำลองของคำและแบบจำลองของหมวดหมู่หลัก วิธีและกระบวนการสร้างเหมือนกับการสร้างแบบจำลองของคำและแบบจำลองของหมวดหมู่หลักทุกประการ สิ่งที่แตกต่างคือชุดข้อมูลที่จะนำเข้าสู่โปรแกรม CMU Toolkit V.2 จะเป็นชุดข้อมูลที่เปลี่ยนจะคำในประโยคเป็นคำที่ถูกระบุหมวดหมู่ย่อยของคำนั้น ๆ ดังภาพที่ 5-6

```
PRONOUN VERB3 POSITION
ANIMAL VERB3 POSITION
ANIMAL VERB3 POSITION
ANIMAL VERB2 HOUSEHOLD
ANIMAL VERB2 HOUSEHOLD VERB1 SOUND
ANIMAL VERB2 CLOTH
ANIMAL VERB1 SOUND
ANIMAL VERB3 PLACE
ANIMAL VERB3 PLACE
ANIMAL VERB2 HOUSEHOLD
ANIMAL VERB2 CLOTH
ANIMAL VERB2 BODY
PRONOUN VERB2 BODY
ANIMAL VERB3
ANIMAL VERB0 PLACE
```

ภาพที่ 5-6 ตัวอย่างข้อมูลของหมวดหมู่ย่อยสำหรับใช้ในโปรแกรม CMU Toolkit V.2

5.3.3 การสร้างแบบจำลองของหมวดหมู่หลัก

การสร้างแบบจำลองของหมวดหมู่หลักนั้น เริ่มจากการพิจารณาลักษณะสำคัญเอ็นแกรมของหมวดหมู่หลัก โดยการคำนวณค่าความน่าจะเป็นด้วยแบบจำลองเอ็นแกรม สำหรับเครื่องมือที่นำมาใช้ในการคำนวณค่าความน่าจะเป็น คือโปรแกรม CMU Toolkit V.2 เช่นเดียวกับการสร้างแบบจำลองของคำ วิธีและกระบวนการสร้างเหมือนกับการสร้างแบบจำลองของคำทุกประการ สิ่งที่แตกต่างคือชุดข้อมูลที่จะนำเข้าสู่โปรแกรม CMU Toolkit V.2 จะเป็นชุดข้อมูลที่เปลี่ยนจะคำในประโยคเป็นคำที่ถูกระบุหมวดหมู่หลักของคำนั้น ๆ ดังภาพที่ 5-7

```
WHAT    MOVEMENT        WHAT    SIZE
WHAT    MOVEMENT        SOUND
WHAT    MOVEMENT        SOUND
WHAT    MOVEMENT        WHAT
WHAT    MOVEMENT        PERSPECTIVE
WHAT    MOVEMENT        WHAT
WHAT    MOVEMENT        SIZE
WHAT    MOVEMENT        WHAT    COLOR
WHAT    MOOD
WHAT    MOVEMENT        WHAT
WHAT    MOVEMENT        WHAT
WHAT    MOVEMENT        WHAT
WHAT    MOVEMENT        WHAT
```

ภาพที่ 5-7 ตัวอย่างข้อมูลของหมวดหมู่หลักสำหรับใช้ในโปรแกรม CMU Toolkit V.2

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

งานวิจัยในส่วนนี้ใช้ 2-แกรม หรือไบแกรมของคำที่ถูกระบุตามหมวดหมู่หลัก คือค่าความน่าจะเป็นของคำที่ถูกระบุตามหมวดหมู่หลักที่ปรากฏอยู่ในคลังข้อความ เมื่อพิจารณาเฉพาะ 2 คำที่อยู่ติดกัน ดังภาพที่ 5-8

```
\2-grams:
-1.0669 BACKGROUND BACKGROUND -0.7719
-99.9990 BACKGROUND COLOR -0.2600
-99.9990 BACKGROUND GROUP -0.2812
-0.4649 BACKGROUND MOVEMENT 0.3738
-1.7202 BACKGROUND NUMBER -0.0917
-99.9990 BACKGROUND PERSPECTIVE 0.0710
-1.1761 BACKGROUND PREPOSITION 0.2688
-1.1181 BACKGROUND SIZE -0.3839
-0.5160 BACKGROUND WHAT -0.0321
-1.3222 BACKGROUND WHEN 0.1745
-1.5441 BACKGROUND WHERE 0.6608
-0.5441 COLOR MOVEMENT -0.3592
-99.9990 COLOR NUMBER -0.2098
-99.9990 COLOR SIZE -0.3543
-0.3680 COLOR WHAT -0.0898
-99.9990 MOOD BACKGROUND -0.2238
-99.9990 MOOD NUMBER -0.1664
-1.6385 MOOD PERSPECTIVE 0.2335
-1.6385 MOOD SIZE 0.2977
-0.0419 MOOD WHAT -0.4627
-99.9990 MOOD WHERE -0.3935
-1.1583 MOVEMENT BACKGROUND 0.3195
```

ภาพที่ 5-8 ตัวอย่างค่าความน่าจะเป็นของหมวดหมู่หลัก 2-แกรม ในรูปแบบไฟล์.arpa

พิจารณาจากภาพที่ 5-8 ค่าที่ปรากฏในหลักแรกด้านซ้ายมือ คือ ค่าความน่าจะเป็นของการเกิดคำที่ปรากฏคลังข้อความในหลักที่ 2 และในหลักที่ 3 ด้านขวามือ คือค่าปรับเรียบแบบย้อน (Back-off Smoothing) ของไบแกรม

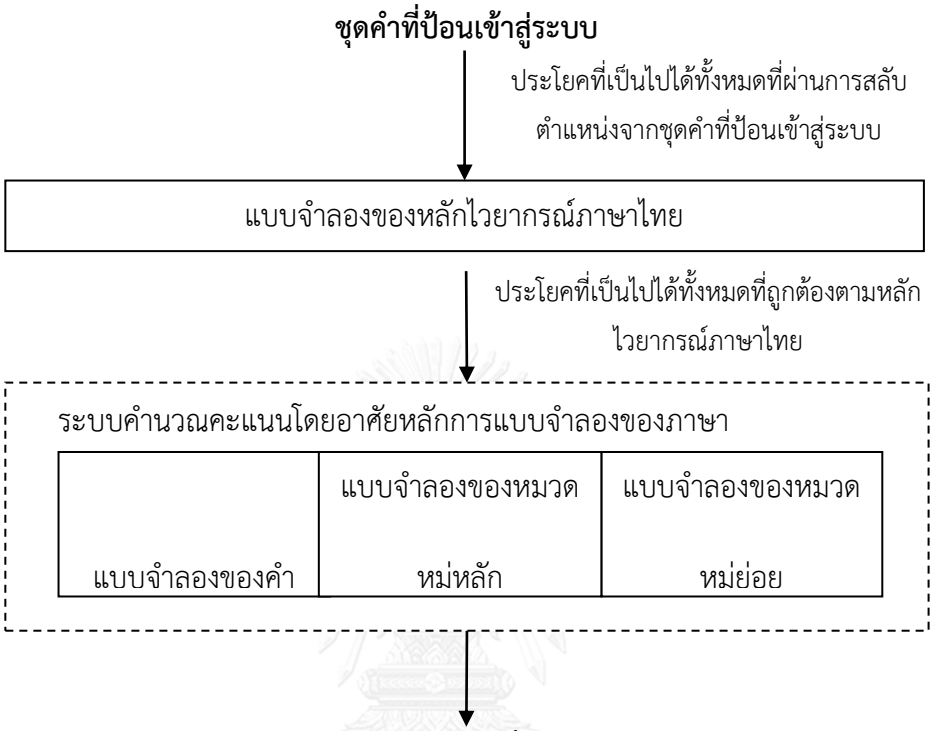
5.4 การสร้างระบบคำนวณคะแนนและอัลกอริทึมในการเลือกประโยคคำตอบที่นำเสนอ

สำหรับระบบคำนวณคะแนนและอัลกอริทึมในการเลือกประโยคคำตอบที่นำเสนอในงานวิจัยนี้ แบ่งออกเป็น 2 ระบบ ได้แก่ ระบบคำนวณคะแนนและอัลกอริทึมในการเลือกประโยคคำตอบโดยอาศัยหลักการและลักษณะสำคัญของเอ็นแกรม และระบบคำนวณคะแนนและอัลกอริทึมในการเลือกประโยคคำตอบโดยอาศัยหลักการและลักษณะสำคัญของเอ็นแกรมร่วมกับการตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก

5.4.1 ระบบคำนวณคะแนนและอัลกอริทึมในการเลือกประโยคคำตอบโดยอาศัยหลักการและลักษณะสำคัญของเอ็นแกรม

สำหรับระบบคำนวณคะแนนและอัลกอริทึมในการเลือกประโยคคำตอบโดยอาศัยหลักการและลักษณะสำคัญของเอ็นแกรมนั้นจะเริ่มจากการพิจารณาชุดของคำที่ป้อนเข้าสู่ระบบและคำนวณคะแนน เพื่อทำการเรียงประโยคแบบอัตโนมัติและหาประโยคที่เหมาะสมมาเป็นประโยคคำตอบ โดยพิจารณาผ่านแบบจำลองของหลักไวยากรณ์ภาษาไทย

แบบจำลองของคำ แบบจำลองของหมวดหมู่หลัก และแบบจำลองของหมวดหมู่ย่อย ดังภาพที่ 5-9



ภาพที่ 5-9 โครงสร้างการทำงานของระบบเมื่ออาศัยหลักการและลักษณะสำคัญของเอ็นแกรม

ขั้นตอนในการคำนวณคะแนนและคัดเลือกประโยคคำตอบของระบบนี้มีดังต่อไปนี้

1. ผู้ใช้ป้อนชุดคำเข้าสู่ระบบ
2. ระบบจะทำการหาประโยคที่เป็นไปได้ทั้งหมดจากชุดคำที่ป้อนเข้าสู่ระบบโดยการสลับตำแหน่งของคำทั้งหมด
3. ระบบนำประโยคทั้งหมดที่ได้ผ่านการสลับตำแหน่งของคำเข้าสู่แบบจำลองของหลักไวยากรณ์ภาษาไทย เพื่อพิจารณาหาประโยคที่ถูกต้องตามหลักภาษาไทย ประโยคที่ผ่านแบบจำลองของหลักไวยากรณ์ภาษาไทย จะมีการเรียงลำดับของคำในประโยคที่ถูกต้องตามหลักไวยากรณ์ภาษาไทยและมีโอกาสที่จะถูกเลือกเป็นประโยคคำตอบ
4. ระบบนำประโยคที่ผ่านการคัดเลือกจากแบบจำลองของหลักไวยากรณ์ภาษาไทย เข้าสู่ระบบคำนวณคะแนนดังภาพที่ 5-9 โดยอาศัยหลักการแบบจำลองของภาษา ซึ่ง

ประกอบไปด้วยแบบจำลองของคำ แบบจำลองของหมวดหมู่หลัก และแบบจำลองของหมวดหมู่ย่อย เพื่อคำนวณหาคะแนนไบแกรมของหมวดหมู่หลัก และไบแกรมของหมวดหมู่ย่อย ซึ่งคำนวณผ่านโปรแกรม CMU Toolkit V.2 ดังสมการ 5.3 5.4 และ 5.5 ตามลำดับ

$$ScoreW_i = \log P(W_i \mid W_{i-1}) \quad (5.3)$$

$$ScoreC_i = \log P(C_i \mid C_{i-1}) \quad (5.4)$$

$$ScoreS_i = \log P(S_i \mid S_{i-1}) \quad (5.5)$$

กำหนดให้

$ScoreW_i$ คือ คะแนนที่ได้จากการคำนวณหาความน่าจะเป็นในการหาไบแกรมของคำ

$ScoreC_i$ คือ คะแนนที่ได้จากการคำนวณหาความน่าจะเป็นในการหาไบแกรมของหมวดหมู่หลัก

$ScoreS_i$ คือ คะแนนที่ได้จากการคำนวณหาความน่าจะเป็นในการหาไบแกรมของหมวดหมู่ย่อย

จากนั้นจะทำการรวมคะแนนของไบแกรมของคำ ไบแกรมของหมวดหมู่หลัก และไบแกรมของหมวดหมู่ย่อยในแต่ละประโยค ดังสมการที่ 5.6 5.7 และ 5.8 ตามลำดับ

$$TotalScoreW = |ScoreW_i| + |ScoreW_{i+1}| + \ldots + |ScoreW_n| \quad (5.6)$$

$$TotalScoreC = |ScoreC_i| + |ScoreC_{i+1}| + \ldots + |ScoreC_n| \quad (5.7)$$

$$TotalScoreS = |ScoreS_i| + |ScoreS_{i+1}| + \ldots + |ScoreS_n| \quad (5.8)$$

เมื่อ $TotalScoreW$ คือ คะแนนรวมของประโยคเมื่อพิจารณาค่าความน่าจะเป็นจากไบแกรมของคำ

$TotalScoreC$ คือ คะแนนรวมของประโยคเมื่อพิจารณาค่าความน่าจะเป็นจากไบแกรมหมวดหมู่หลัก

$TotalScoreS$ คือ คะแนนรวมของประโยคเมื่อพิจารณาค่าความน่าจะเป็นจากไบแกรมหมวดหมู่ย่อย

ในกรณีที่ไม่พบไบแกรมของคำ หรือไบแกรมของหมวดหมู่หลัก หรือไบแกรมของหมวดหมู่ย่อย ในคลังข้อมูลที่ถูกเรียนรู้ ระบบจะมีค่าทำโทษให้กับ $ScoreW_i$ หรือ $ScoreC_i$ หรือ $ScoreS_i$ แต่ละตัวเท่ากับ +10 เนื่องจากค่าความน่าจะเป็นที่นำมาคำนวณเป็นค่า log ของค่าความน่าจะเป็นที่ติดลบ นั่นหมายถึงคะแนนรวมของประโยคยิ่งน้อยยิ่งดี

จากนั้นนำคะแนนของทั้งสามแบบจำลองมารวมกัน เพื่อหาประโยคที่มีคะแนนรวมดีที่สุดมาเป็นประโยคคำตอบโดยกำหนดให้ $TotalSentenceScore$ คือ คะแนนรวมของค่าความน่าจะเป็นในการเกิดประโยคแต่ละประโยค

$$TotalSentenceScore = TotalScoreW + TotalScoreC + TotalScoreS \quad (5.9)$$

เมื่อได้คะแนนรวมของประโยคมาครบทุกประโยคแล้ว ประโยคที่ถูกเลือกคือประโยคที่มีคะแนนรวมของประโยคน้อยที่สุด ดังสมการที่ 5.10

$$Sentence* = \min(TotalSentenceScore) \quad (5.10)$$

เมื่อ $Sentence*$ คือ ประโยคที่ถูกเลือก

5. แสดงประโยคคำตอบที่ถูกเลือก

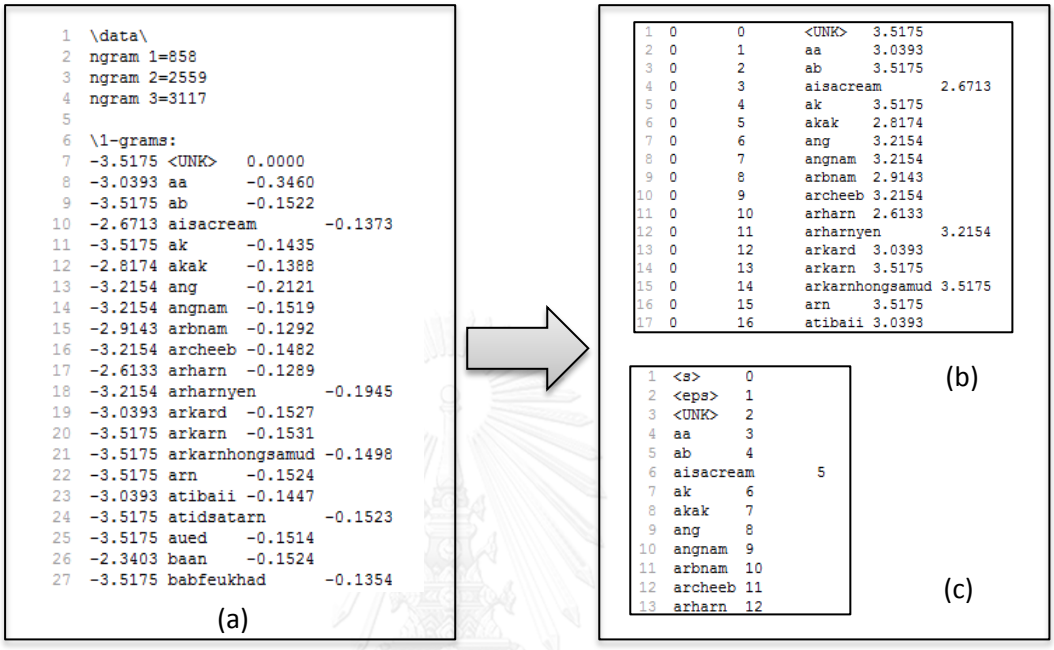
5.4.2 ระบบคำนวณคะแนนและอัลกอริทึมในการเลือกประโยคคำตอบโดยอาศัยหลักการและลักษณะสำคัญของเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำแบบแบบถ่วงน้ำหนัก (WFST)

หลักการตัวเปลี่ยนแปรสถานะจำแบบแบบถ่วงน้ำหนักนั้นเป็นการสร้างโครงข่ายของข้อมูลที่เราต้องการค้นหา ซึ่งเป็นโครงข่ายที่สามารถค้นหาเส้นทางคำตอบที่ดีที่สุดได้อย่างรวดเร็ว การจะสร้างโครงข่ายของคำ โครงข่ายของหมวดหมู่หลัก โครงข่ายของหมวดหมู่ย่อย และโครงข่ายของหน้าที่ของคำ (POS) นั้น มีขั้นตอนการสร้างดังต่อไปนี้

5.4.2.1 ขั้นตอนการเข้ารหัสตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักของคำ หมวดหมู่หลัก หมวดหมู่ย่อย และ POS ด้วยโปรแกรม OpenGrm NGram

การสร้าง WFST นั้น จำเป็นต้องผ่านขั้นตอนการเข้ารหัส WFST เสียก่อน ซึ่งในงานวิจัยนี้ได้ใช้โปรแกรม OpenGrm NGram Version 1.2.1 [49] เพื่อทำการเข้ารหัส โดย

การจะเข้ารหัสได้นั้น ต้องนำไฟล์ค่าความน่าจะเป็นเอ็นแกรมที่อยู่ในรูปแบบ .arpa ที่ได้จากการขั้นตอนสร้างแบบจำลองของคำ หมวดหมู่หลัก หมวดหมู่ย่อย และหน้าที่ของคำมาใช้ ดังตัวอย่างในภาพที่ 5-10



ภาพที่ 5-10 ตัวอย่างไฟล์ค่าความน่าจะเป็นเอ็นแกรมในรูปแบบ.arpa (a) เมื่อป้อนเข้าสู่โปรแกรม OpenGrm NGram และไฟล์ผลลัพธ์ที่ได้จากโปรแกรม (b) และ (c)

จากภาพที่ 5-10 เมื่อป้อนไฟล์เอ็นแกรมที่อยู่ในรูปแบบ .arpa ดังภาพที่ 5-10(a) เข้าสู่โปรแกรม OpenGrm NGram และใช้คำสั่งบนระบบปฏิบัติการ Linux ดังนี้

$>./ngram2fsm file.arpa(in) debug12 symbolfile(out) fsmfile(out)

จะได้ผลลัพธ์จากการใช้คำสั่งนี้ คือ

1. ไฟล์เข้ารหัสตัวเปลี่ยนแปรสถานะจำกัด ดังภาพที่ 5-10(b)
2. ไฟล์สัญลักษณ์อินพุต ดังภาพที่ 5-10(c) ที่จะเก็บลำดับของคำทั้งหมดที่ปรากฏอยู่ในไฟล์เข้ารหัส

เมื่อได้ไฟล์เอาต์พุตทั้งสองมาแล้ว จะต้องทำการแปลงไฟล์เข้ารหัสตัวเปลี่ยนแปรสถานะจำกัด เพื่อให้สามารถใช้งานกับโปรแกรม OpenFst ในขั้นตอนต่อไปได้ โดยต้องแปลงข้อมูลในไฟล์เข้ารหัสตัวเปลี่ยนแปรสถานะจำกัดให้อยู่ในรูปแบบดังภาพที่ 5-11

```
0       0       <UNK>           <UNK>   3.5175
0       1       aa              VERB0   3.0393
0       2       ab              VERB0   3.5175
0       3       aisacream       FOOD    2.6713
0       4       ak              SOUND   3.5175
0       5       akak            SOUND   2.8174
0       6       ang             PLACE   3.2154
0       7       angnam          PLACE   3.2154
```

(a)

```
1  <eps> 0
2  <UNK> 1
3  ADJECTIVE 2
4  ANIMAL 3
5  BODY 4
6  CLOTH 5
7  COLOR 6
8  CONJUNCTION 7
9  DAY 8
10 EMOTION 9
```

(b)

ภาพที่ 5-11ตัวอย่างการจัดข้อมูลในไฟล์เข้ารหัสตัวเปลี่ยนแปรสถานะจำกัด (a) และสร้างไฟล์สัญลักษณ์เอาต์พุต (b) เพื่อให้เหมาะสมและพร้อมใช้งานในโปรแกรม OpenFst

จากภาพที่ 5-11(a) จะแสดงไฟล์เข้ารหัสตัวเปลี่ยนแปรสถานะจำกัดที่มีการเพิ่มข้อมูลในหลักที่ 4 จากเดิมที่มีหลักทั้งสิ้น 4 หลัก ในขั้นตอนนี้จะเพิ่มเป็น 5 หลัก โดยข้อมูลในแต่ละหลักมีดังนี้

หลักที่ 1 คือสถานะเริ่มต้น

หลักที่ 2 คือสถานะสิ้นสุด

หลักที่ 3 คือเอ็นแกรมของสิ่งที่ต้องการให้ทำหน้าที่เป็นอินพุต

หลักที่ 4 คือเอ็นแกรมของสิ่งที่ต้องการให้ทำหน้าที่เป็นเอาต์พุต

หลักที่ 5 คือค่าความน่าจะเป็นของเอ็นแกรมที่ทำหน้าที่เป็นอินพุต

นอกจากนี้ยังจำเป็นต้องสร้างไฟล์สัญลักษณ์เอาต์พุตเพิ่ม เพื่อเก็บลำดับของคำที่ทำหน้าที่เป็นเอาต์พุตทั้งหมดที่ปรากฏอยู่ในไฟล์เข้ารหัสดังภาพที่ 5-11(b)

ในงานวิจัยนี้ต้องการสร้างตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักของคำหมวดหมู่หลัก หมวดหมู่ย่อย และ POS จึงจำเป็นต้องทำการเข้ารหัสตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักทั้งสิ้น 4 ครั้งด้วยกัน ดังนี้

1. การเข้ารหัสตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักของคำ ที่ประกอบด้วยอินพุต คือ คำ และเอาต์พุต คือ หมวดหมู่ย่อย

2. การเข้ารหัสตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักของหมวดหมู่ย่อย ที่ประกอบด้วย อินพุต คือ หมวดหมู่ย่อย และเอาต์พุต คือ หมวดหมู่หลัก
3. การเข้ารหัสตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักของหมวดหมู่หลัก ที่ประกอบด้วย อินพุต คือ หมวดหมู่หลัก และเอาต์พุต คือ POS
4. การเข้ารหัสตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักของ POS ที่ประกอบด้วย อินพุต คือ POS และเอาต์พุต คือ POS

5.4.2.2 ขั้นตอนการสร้างโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (Weighted Finite-State Transducer: WFST) ของคำ หมวดหมู่หลัก หมวดหมู่ย่อย และ POS ด้วยโปรแกรม OpenFst

การสร้างตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักหรือ WFST ในงานวิจัยนี้ ใช้โปรแกรม OpenFst Version 1.4.1 [13] โดยโปรแกรม OpenFst นี้สามารถสร้าง รวม และค้นหาตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักหรือ WFST ได้ แต่ก่อนที่จะทำการสร้างตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักนั้นต้องมีการเลเบลค่าอินพุต ค่าเอาต์พุต และค่าน้ำหนักเสียก่อน โดยในงานวิจัยนี้ได้จัดการเลเบลค่าและเตรียมไฟล์สำหรับเป็นไฟล์อินพุตเข้าสู่โปรแกรม OpenFst เรียบร้อยแล้วในขั้นตอนที่ 5.4.2.1

การจะสร้างตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักหรือ WFST นั้นสามารถทำได้โดยใช้คำสั่งผ่านตัวโปรแกรม OpenFst ดังนี้

$> fstcompile --isymbols=isyms(in).txt --osymbols=osyms(in).txt text(in).fst binary(out).fst

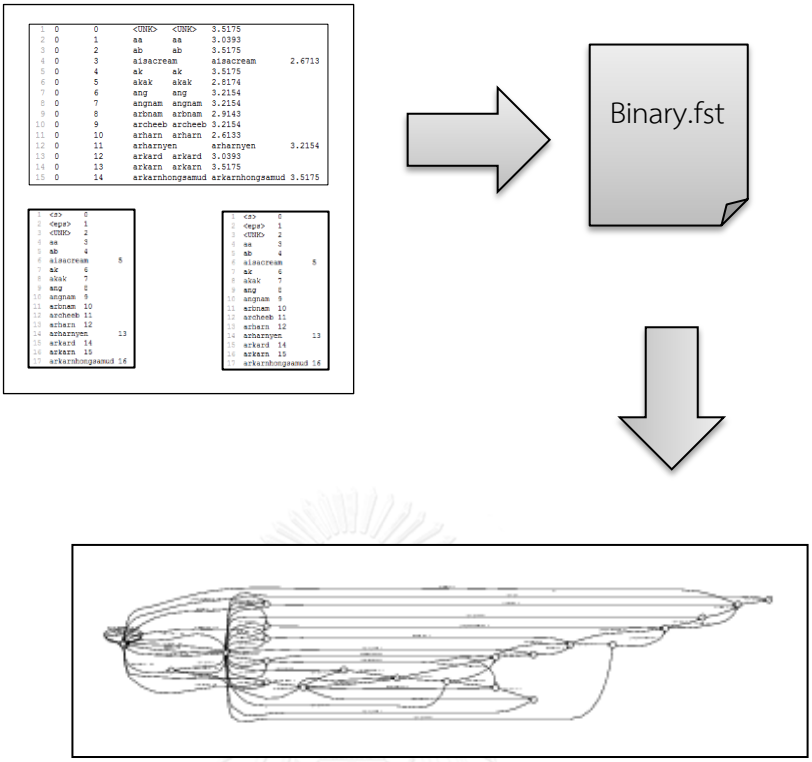
เมื่อ isyms(in).txt คือไฟล์สัญลักษณ์อินพุต

osyms (in).txt คือไฟล์สัญลักษณ์เอาต์พุต

text(in).fst คือไฟล์เข้ารหัสตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก

และ binary(out).fst คือไฟล์ผลลัพธ์จากโปรแกรมที่แปลงไฟล์เข้ารหัสเป็นไฟล์เข้ารหัสในรูปแบบของไบนารี

เมื่อเรียกใช้คำสั่งนี้ผ่านตัวโปรแกรม OpenFst โปรแกรมจะเรียกคำสั่ง fstcompile จากนั้นโปรแกรมจะแปลงไฟล์ isyms(in).txt osyms(in).txt และ text(in).fst แล้วแสดงผลออกมาในไฟล์ binary(out).fst ซึ่งเป็นไฟล์ที่ถูกเข้ารหัสในแบบไบนารี



ภาพที่ 5-12 ขั้นตอนการสร้างตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักแบบไบนารีไฟล์และแสดงผลออกมาในลักษณะของโครงข่าย

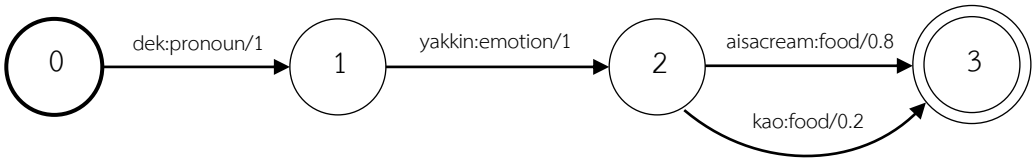
เมื่อสร้างไฟล์เข้ารหัสแบบไบนารีเรียบร้อยแล้ว สามารถแปลงไฟล์เข้ารหัสไบนารีเป็นไฟล์รูปภาพเพื่อให้แสดงผลออกมาในลักษณะของโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก ดังภาพที่ 5-12 ผ่านคำสั่งต่อไปนี้

```
$> fstdraw --isymbols=isyms.txt --osymbols=osyms.txt binary.fst
binary.dot

$> dot -Tps binary.dot >binary.ps
```

โดยในงานวิจัยนี้ต้องการสร้างโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักของคำ หมวดหมู่หลัก หมวดหมู่ย่อย และ POS ดังนี้

1. โครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของคำ ซึ่งประกอบด้วย อินพุต คือ คำ และเอาต์พุต คือ หมวดหมู่ย่อย

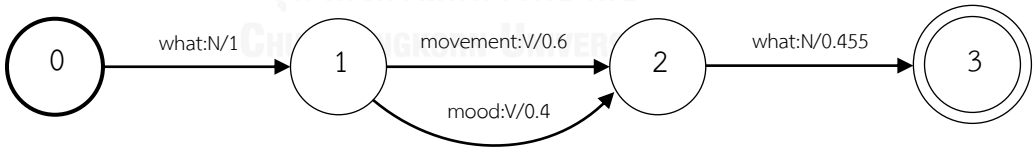


ภาพที่ 5-13 แสดงตัวอย่างโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของคำ

2. โครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของหมวดหมู่ย่อย ซึ่งประกอบด้วย อินพุต คือ หมวดหมู่ย่อย และเอาต์พุต คือ หมวดหมู่หลัก

ภาพที่ 5-14 ตัวอย่างโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของหมวดหมู่ย่อย

3. โครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของหมวดหมู่หลัก ที่ประกอบด้วย อินพุต คือ หมวดหมู่หลัก และเอาต์พุต คือ POS



ภาพที่ 5-15 ตัวอย่างโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของหมวดหมู่หลัก

4. โครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของ POS ที่ประกอบด้วย อินพุต คือ POS และเอาต์พุต คือ POS

ภาพที่ 5-16 ตัวอย่างโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักจากแบบจำลองของ POS

5.4.2.3 ขั้นตอนการนำ WFST ของคำ หมวดหมู่หลัก หมวดหมู่ย่อย และ POS มาประกอบกัน

ในงานวิจัยนี้ได้สร้างโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก $WSCP$ โดยนำโครงข่ายของแบบจำลองของคำ (W) โครงข่ายของแบบจำลองของหมวดหมู่ย่อย (S) โครงข่ายของแบบจำลองของหมวดหมู่หลัก (C) และโครงข่ายของแบบจำลองของ POS (P) มาประกอบกันผ่านฟังก์ชันประกอบ (Composition Function) ดังสมการที่ 5.11

$$WSCP = W \circ S \circ C \circ P \tag{5.11}$$

โดยกำหนดให้อินพุตและเอาต์พุตของแต่ละโครงข่ายมีรายละเอียดดังนี้

1. โครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักของแบบจำลองของคำ (W)

   อินพุต คือ คำ

   เอาต์พุต คือ หมวดหมู่ย่อย

2. โครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักของแบบจำลองของหมวดหมู่ย่อย (S)

   อินพุต คือ หมวดหมู่ย่อย

   เอาต์พุต คือ หมวดหมู่หลัก

3. โครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักของแบบจำลองของหมวดหมู่หลัก (C)

   อินพุต คือ หมวดหมู่หลัก

   เอาต์พุต คือ POS

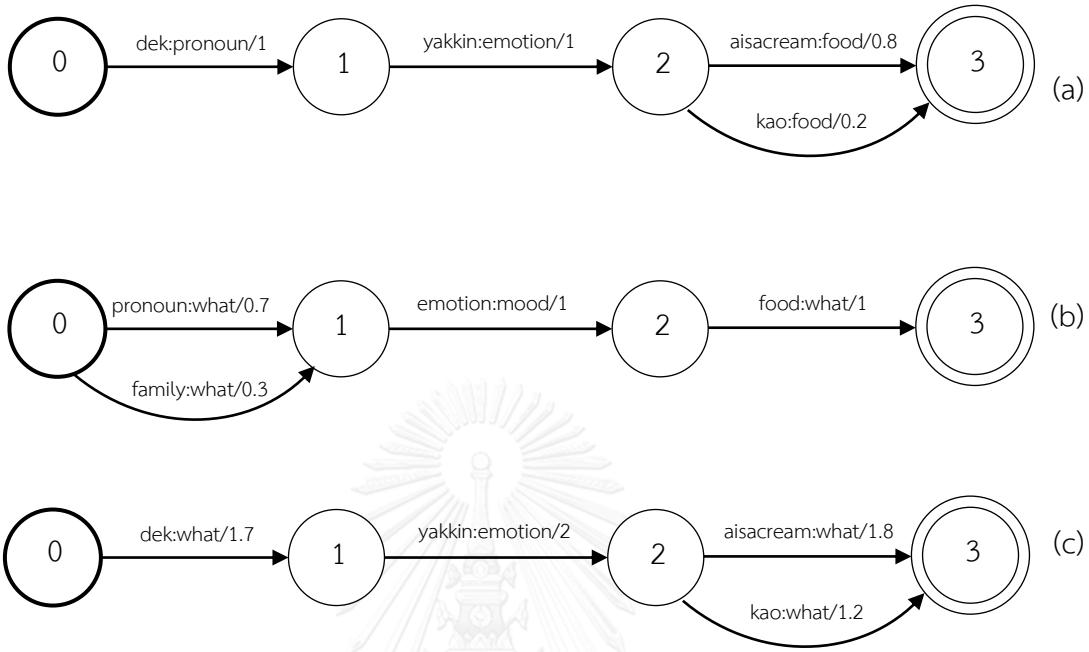
4. โครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักของแบบจำลองของ POS (P)

   อินพุต คือ POS

   เอาต์พุต คือ POS

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

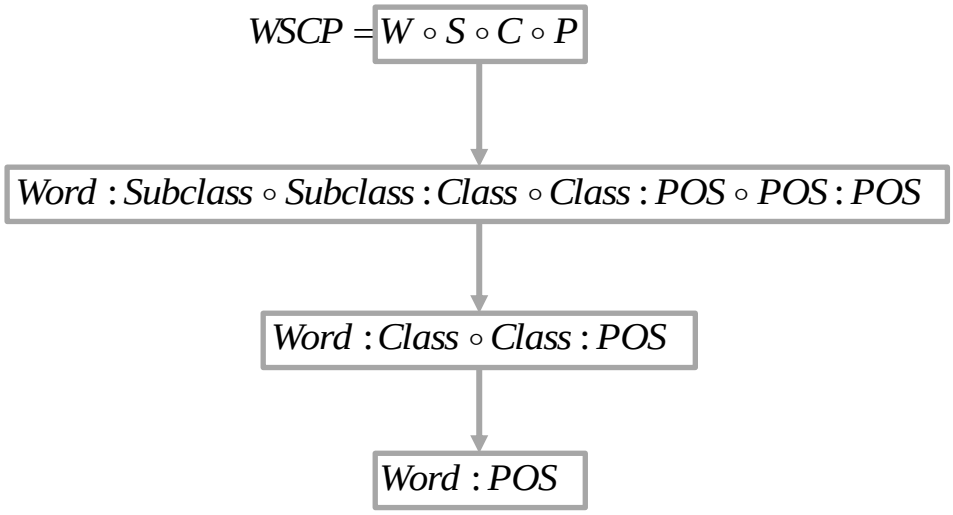
ในการทำฟังก์ชันการประกอบกันระหว่างโครงข่าย W กับโครงข่าย S สามารถแสดงได้ดังตัวอย่างในภาพที่ 5-17



ภาพที่ 5-17 ตัวอย่างการทำฟังก์ชันประกอบระหว่างโครงข่าย W (a) กับโครงข่าย S (b) และโครงข่าย WS (c) คือผลของการทำฟังก์ชันประกอบของโครงข่าย W และ S

จากภาพที่ 5-17(a) เป็นตัวอย่างโครงข่ายแบบจำลองของคำ ซึ่งบอกลักษณะลำดับของคำด้วยค่าความน่าจะเป็นต่าง ๆ โดยสัญลักษณ์บนเส้นเชื่อมประกอบไปด้วย อินพุต: เอาต์พุต/ค่าความน่าจะเป็น โดยค่าความน่าจะเป็นหรือค่าน้ำหนักนั้น หมายถึง ค่าความน่าจะเป็นจากการเปลี่ยนสถานะจากสถานะต้นไปยังสถานะปลาย ตัวอย่างเช่นจากภาพที่ 5-17(b) สัญลักษณ์บนเส้นเชื่อม dek:what/1.7 หมายถึง เมื่อได้รับอินพุตเป็นคำว่า dek จะให้เอาต์พุตเป็น what โดยมีค่าความน่าจะเป็นในการย้ายสถานะจากสถานะต้นเส้นเชื่อมไปยังสถานะปลายเส้นเชื่อมเป็น 1.7

เมื่อพิจารณาอินพุตและเอาต์พุตของโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก $WSCP$ จะเห็นว่าอินพุตของโครงข่าย $WSCP$ คือ คำ และเอาต์พุตของโครงข่าย $WSCP$ คือ POS ซึ่งมีที่มาจากการประกอบกันของโครงข่ายต่าง ๆ ดังนี้



ภาพที่ 5-18 การใช้ฟังก์ชันประกอบระหว่างโครงข่ายต่าง ๆ และขั้นตอนการแปลงอินพุตและเอาต์พุตของโครงข่าย $WSCP$

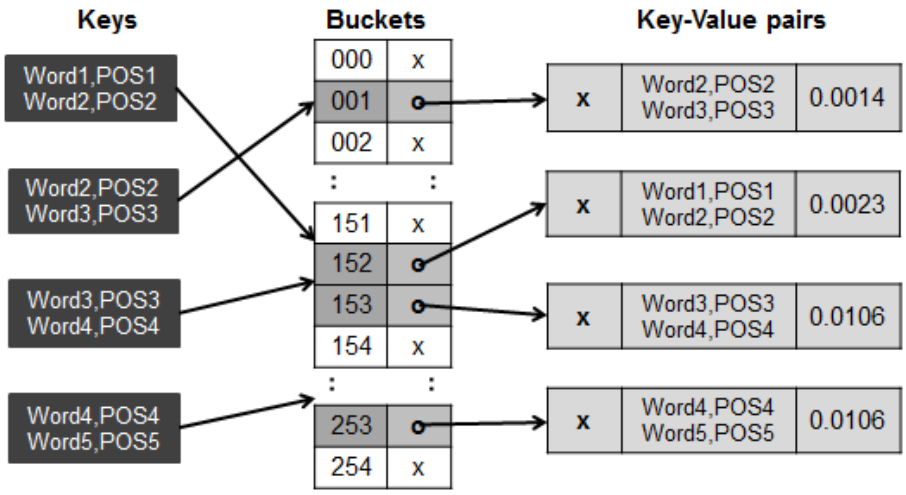
ผลจากการใช้ฟังก์ชันประกอบระหว่างโครงข่ายทั้ง 4 ได้แก่ โครงข่ายของแบบจำลองของคำ (W) โครงข่ายของแบบจำลองของหมวดหมู่ย่อย (S) โครงข่ายของแบบจำลองของหมวดหมู่หลัก (C) และโครงข่ายของแบบจำลองของ POS (P) ทำให้เกิดโครงข่าย $WSCP$ และมีการแปลงอินพุตและเอาต์พุตของโครงข่าย $WSCP$ ดังภาพที่ 5-18 โดยเมื่อเปรียบเทียบในระดับสถานะเดียวกัน แต่ละโครงข่ายจะสามารถประกอบกันได้ ก็ต่อเมื่อเอาต์พุตของโครงข่ายแรกกับอินพุตของโครงข่ายที่สองต้องมีค่าเหมือนกัน และค่าน้ำหนักหรือค่าความน่าจะเป็นบนเส้นเชื่อมของโครงข่าย $WSCP$ นั้น เกิดจากผลรวมค่าน้ำหนักของเส้นเชื่อมแต่ละบนโครงข่าย W โครงข่าย S โครงข่าย C และโครงข่าย P ที่สามารถประกอบกันได้

5.4.2.4 ขั้นตอนการนำข้อมูลจาก WFST เข้าสู่ตารางแฮช (Hash table)

ตารางแฮช เป็นตารางขนาดใหญ่ มักใช้แถวลำดับในการจัดเก็บข้อมูลจำนวนมาก โดยมีลักษณะการเก็บแบบดัชนี (Indexing) [50] สำหรับวิธีการจัดเก็บนั้นจะนำข้อมูลที่จะนำมาเก็บผ่านฟังก์ชันฟังก์ชันหนึ่ง ซึ่งจะเรียกข้อมูลที่ผ่านฟังก์ชันนั้นว่า key ซึ่งเรียกว่าฟังก์ชันแฮช จากนั้นจะได้เลขซึ่งจำเพาะกับข้อมูลนั้น กล่าวคือ ข้อมูลแต่ละตัวเมื่อผ่าน

ฟังก์ชันแฮชแล้ว จะได้เลขที่แตกต่างกัน แล้วจึงนำข้อมูลไปเก็บไว้ในตารางหรือแถวลำดับที่กำหนดไว้

ในงานวิจัยนี้ได้ทำการนำข้อมูลที่อยู่ในโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก WSCP มาจัดเก็บไว้ในตารางแฮช เนื่องจากโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักนั้นมีขนาดใหญ่และใช้หน่วยความจำของระบบมาก นอกจากนี้ยังนำไปประยุกต์บนซอฟต์แวร์ AAC ได้ยาก เนื่องจากโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักนั้นต้องสร้างบนระบบปฏิบัติการ Linux ดังนั้นเพื่อลดขนาดของหน่วยความจำของระบบ เพิ่มความเร็วในการประมวลผล และง่ายต่อการคำนวณหาคะแนนของประโยคคำตอบ งานวิจัยนี้จึงได้เลือกเก็บอินพุต เอาต์พุต และค่าน้ำหนักจากโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก WSCP ไว้ในตารางแฮช ดังภาพที่ 5-19

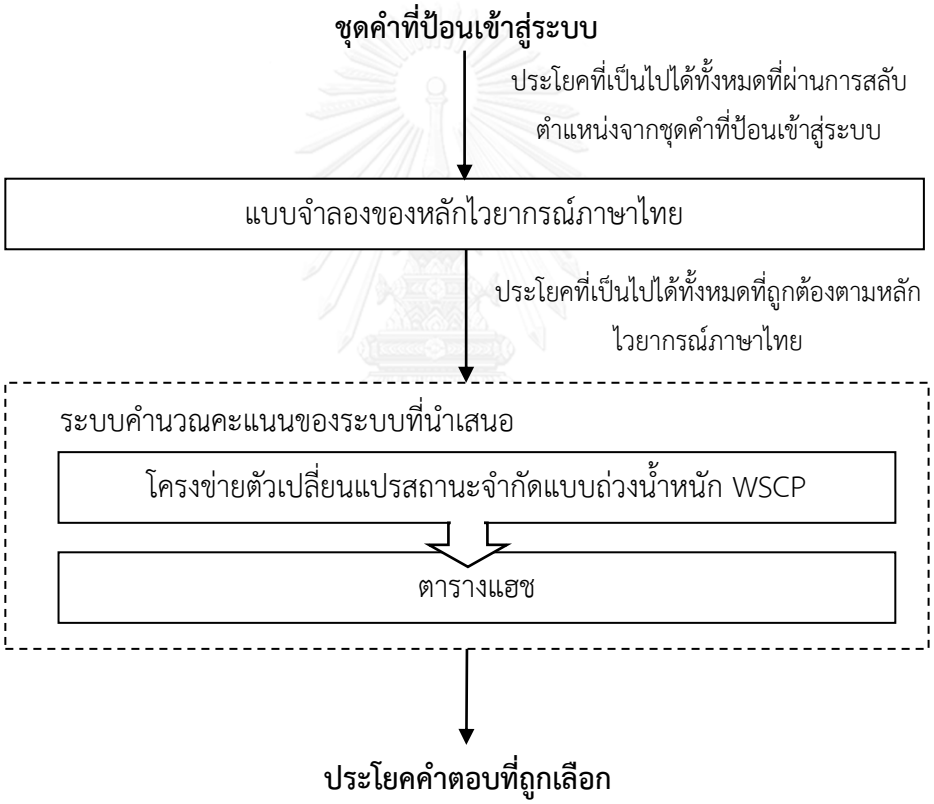


ภาพที่ 5-19 ตัวอย่างการจัดเก็บอินพุต เอาต์พุต และค่าน้ำหนักจากโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก *WSCP* ไว้ในตารางแฮช

โดยกำหนดให้ Key ในตารางแฮชนี้ คือคำสองคำที่อยู่ติดกัน พร้อมระบุ POS ของคำนั้น ๆ หรือค่าอินพุตและเอาต์พุตจากโครงข่ายตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก *WSCP* และ Value ในตารางแฮชนี้จะเก็บค่าผลคูณของค่าน้ำหนักของคำสองคำที่อยู่ติดกันหรือค่าผลคูณค่าน้ำหนักของ Key นั่นเอง

5.4.2.5 ขั้นตอนการสร้างระบบคำนวณคะแนนและเลือกประโยคคำตอบ

สำหรับระบบคำนวณคะแนนและเลือกประโยคคำตอบในระบบที่นำเสนอนี้ จะมีการพิจารณาชุดของคำที่ป้อนเข้าสู่ระบบและคำนวณคะแนน เพื่อทำการเรียงประโยคแบบอัตโนมัติและหาประโยคที่เหมาะสมมาเป็นประโยคคำตอบ โดยพิจารณาผ่านแบบจำลองของหลักไวยากรณ์ภาษาไทยเพื่อหาประโยคที่เป็นไปได้ทั้งหมดที่ถูกต้องตามหลักไวยากรณ์ภาษาไทย และส่งต่อสู่ระบบคำนวณคะแนนของระบบที่นำเสนอ เพื่อหาประโยคคำตอบที่เหมาะสมที่สุด ดังภาพที่ 5-20



ภาพที่ 5-20 โครงสร้างการทำงานของระบบที่นำเสนอ

ขั้นตอนในการคำนวณคะแนนและคัดเลือกประโยคคำตอบของระบบที่นำเสนอมีดังต่อไปนี้

1. ผู้ใช้ป้อนชุดคำที่มีการระบุหมวดหมู่หลัก หมวดหมู่ย่อย และหน้าที่ของคำ POS เข้าสู่ระบบ
2. ระบบจะทำการหาประโยคที่เป็นไปได้ทั้งหมดจากชุดคำที่ป้อนเข้าสู่ระบบโดยการสลับตำแหน่งของคำทั้งหมด
3. ระบบนำประโยคทั้งหมดที่ผ่านการสลับตำแหน่งของคำเข้าสู่แบบจำลองของหลักไวยากรณ์ภาษาไทย เพื่อพิจารณาหาประโยคที่ถูกต้องตามหลักภาษาไทย ประโยคที่ผ่านแบบจำลองของหลักไวยากรณ์ภาษาไทย จะมีการเรียงลำดับของคำในประโยคที่ถูกต้องตามหลักไวยากรณ์ภาษาไทยและมีโอกาสที่จะถูกเลือกเป็นประโยคคำตอบ
4. ระบบนำประโยคที่ผ่านการคัดเลือกเข้าสู่ระบบคำนวณคะแนนที่นำเสนอ โดยจะพิจารณาคำทีละสองคำที่อยู่ติดกันตั้งแต่สองคำแรกจนถึงสองคำสุดท้ายในแต่ละประโยค เพื่อหาคะแนนรวมของประโยคดังสมการ 5.12

$$TS = ScoreW_{i,i+1} + ScoreW_{i+1,i+2} + \ldots + ScoreW_{n-1,n} \qquad (5.12)$$

เมื่อ $ScoreW_{i,i+1}$ คือ คะแนนของคำที่ 1 และ 2 ที่อยู่ติดกันในประโยค โดยกำหนดให้ประโยคมีความยาวตั้งแต่ i ถึง n

$ScoreW_{i+1,i+2}$ คือ คะแนนของคำที่ 2 และ 3 ที่อยู่ติดกันในประโยค โดยกำหนดให้ประโยคมีความยาวตั้งแต่ i ถึง n

$ScoreW_{n-1,n}$ คือ คะแนนของคำก่อนสุดท้ายและคำสุดท้ายที่อยู่ติดกันในประโยค โดยกำหนดให้ประโยคมีความยาวตั้งแต่ i ถึง n

และ TS คือ คะแนนรวมของประโยค

5. เปรียบเทียบคะแนนของแต่ละประโยคและเลือกประโยคที่มีคะแนนสูงสุดเป็นประโยคคำตอบ
6. แสดงประโยคคำตอบที่ถูกเลือก

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 6
การทดลองและวิธีการวัดผลการทดลอง

เนื้อหาภายในบทนี้จะกล่าวถึงข้อมูลที่ใช้ในการทดสอบระบบ การกำหนดหมวดหมู่หลักและหมวดหมู่ย่อยของคำที่ใช้ในการทดสอบระบบ การสร้างชุดทดสอบและประโยคทดสอบ การทดลองระบบที่นำเสนอแบบต่างๆ และวิธีการประเมินประสิทธิภาพของระบบที่นำเสนอ

6.1 ข้อมูลที่ใช้ในการทดสอบระบบ

ข้อมูลที่นำมาใช้ในการทดสอบระบบในงานวิจัยนี้มาจากข้อมูลในคลังข้อความ TELL-S และข้อมูลในคลังข้อความ BEST2010 โดยแบ่งข้อมูลออกเป็น 2 ชุดดังต่อไปนี้

6.1.1 ชุดข้อมูล TELL-S คือคำและประโยคทั้งหมดที่ปรากฏอยู่ในคลังข้อมูล TELL-S โดยคำที่ปรากฏอยู่ในคลังข้อมูลชุดนี้จะมีการระบุหมวดหมู่หลัก การระบุหมวดหมู่ย่อย และการระบุหน้าที่ของคำ ประกอบไปด้วยคำทั้งสิ้น 3,412 คำ

6.1.2 ชุดข้อมูล BEST คือคำและประโยคที่ปรากฏอยู่ในคลังข้อมูล BEST2010 ในหมวดหมู่นวนิยาย (Novel) โดยคำที่ปรากฏอยู่ในคลังข้อมูลนี้จะมีการระบุหมวดหมู่หลัก การระบุหมวดหมู่ย่อย และการระบุหน้าที่ของคำ เฉพาะคำที่ปรากฏอยู่ในชุดคำของซอฟต์แวร์ AAC ที่มีชื่อว่า Chula AAC [10]

6.2 การกำหนดหมวดหมู่หลักและหมวดหมู่ย่อยของคำที่ใช้ในการทดสอบระบบตามระบบที่นำเสนอ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

การกำหนดหมวดหมู่หลักและหมวดหมู่ย่อยของคำที่ใช้ในการทดสอบระบบแบ่งออกเป็น 3 กลุ่ม ได้แก่

กลุ่มที่ 1 คำที่มาจากคลังข้อความ TELL-S

กลุ่มที่ 2 คำที่มาจากคลังข้อความ BEST2010 และ

กลุ่มที่ 3 คำที่ปรากฏอยู่ทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010

โดยมีรายละเอียดการแบ่งหมวดหมู่ การระบุหน้าที่ของคำ และจำนวนของคำในแต่ละหมวดหมู่ รวมถึงการกำหนด POS ของคำในหมวดหมู่ต่าง ๆ ดังตารางที่ 6-1

ตารางที่ 6-1 การกำหนดหมวดหมู่หลัก หมวดหมู่ย่อยของคำ และจำนวนคำในแต่ละหมวดหมู่ที่อยู่ในคลังข้อความ TELL-S (กลุ่มที่ 1) และคลังข้อความ BEST2010 (กลุ่มที่ 2)

ลำดับที่	**หมวดหมู่**	**คำทั้งหมดใน กลุ่มที่ 1**	**คำทั้งหมดใน กลุ่มที่ 2**
What			
1	Body	47	11
2	Household	35	6
3	Family	15	8
4	Food	40	11
5	Animal	0	9
6	Cloth	55	4
7	Sport	19	2
8	Stationary	18	7
9	Pronoun	23	15
10	Shape	15	1
11	Day	9	2
12	Month	12	0
13	Color	18	6
14	Transportation	18	1
15	Natural	12	9
16	Place	21	14
17	Possessive	23	7
รวม		380	113
Size			
1	Size (ADJ)	14	5
2	Size (ADV)	14	4

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ลำดับที่	หมวดหมู่	คำทั้งหมดใน กลุ่มที่ 1	คำทั้งหมดใน กลุ่มที่ 2
3	Quantity (ADJ)	2	1
4	Quantity (ADV)	2	1
5	Speed (ADJ)	2	0
6	Speed (ADV)	2	2
รวม		36	13
Color			
1	Color	18	6
รวม		18	6
Number			
1	Number	31	7
รวม		31	7
Shape			
1	Shape	15	1
รวม		15	1
Where			
1	Position (ADJ)	17	5
2	Position (ADV)	17	8
รวม		34	13
When			
6	Day (ADJ)	9	2
7	Month (ADJ)	12	0
8	Time (ADJ)	67	4
5	Day (ADV)	9	0
6	Month (ADV)	12	0

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ลำดับที่	หมวดหมู่	คำทั้งหมดใน กลุ่มที่ 1	คำทั้งหมดใน กลุ่มที่ 2
7	Time (ADV)	67	6
รวม		176	12
Background			
1	Natural	12	9
รวม		12	9
Movement			
1	Verb1 (กริยาที่กระทำตั้งแต่เอวลงไป)	23	10
2	Verb2 (กริยาที่กระทำตั้งแต่ลำคอลงไปถึงเอว)	17	17
3	Verb3 (กริยาที่กระทำบริเวณใบหน้า)	12	14
4	Verb0 (การกระทำอื่น ๆ)	5	9
รวม		57	50
Mood			
1	Emotion	12	11
รวม		12	11
Perspective			
1	Perspective (ADJ)	12	7
2	Perspective (ADV)	12	9
รวม		24	16
Sound			
1	Sound (ADJ)	2	3
2	Sound (ADV)	2	8

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ลำดับที่	หมวดหมู่	คำทั้งหมดใน กลุ่มที่ 1	คำทั้งหมดใน กลุ่มที่ 2
รวม		28	27
รวมทั้งสิ้น		**823**	**278**

ตารางที่ 6-2 การกำหนดหน้าที่ของคำ (POS) ที่อยู่ในคลังข้อความ TELL-S (กลุ่มที่ 1) และคลังข้อความ BEST2010 (กลุ่มที่ 2)

ลำดับที่	หมวดหมู่	คำทั้งหมดใน กลุ่มที่ 1	คำทั้งหมดใน กลุ่มที่ 2
คำนาม (N)			
1	Body	47	11
2	Household	35	6
3	Family	15	8
4	Food	40	11
5	Animal	0	9
6	Cloth	55	4
7	Sport	19	2
8	Stationary	18	7
9	Pronoun	23	15
10	Shape	15	1
11	Day	9	2
12	Month	12	0
13	Color	18	6
14	Transportation	18	1
15	Natural	12	9
16	Place	21	14

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ลำดับที่	หมวดหมู่	คำทั้งหมดใน กลุ่มที่ 1	คำทั้งหมดใน กลุ่มที่ 2
รวม		357	106
คำกริยา (V)			
1	Verb1 (กริยาที่กระทำตั้งแต่เอวลงไป)	23	10
2	Verb2 (กริยาที่กระทำตั้งแต่ลำคอลงไปถึงเอว)	17	17
3	Verb3 (กริยาที่กระทำบริเวณใบหน้า)	12	14
4	Emotion	12	11
5	Verb0 (การกระทำอื่นๆ)	5	9
รวม		69	61
ส่วนขยายนาม หรือ ส่วนขยายส่วนขยาย (ADJ)			
1	Quantity	2	1
2	Speed	2	0
3	Size	14	5
4	Color	18	6
5	Position	17	5
6	Day	9	2
7	Month	12	0
8	Time	67	4
9	Perspective	12	7
10	Sound	2	3
11	Number	31	7
12	Possessive	23	7

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ลำดับที่	หมวดหมู่	คำทั้งหมดใน กลุ่มที่ 1	คำทั้งหมดใน กลุ่มที่ 2
รวม		209	47
ส่วนขยายกริยา (ADV)			
1	Quantity	2	1
2	Speed	2	2
3	Size	14	4
4	Position	17	8
5	Day	9	0
6	Month	12	0
7	Time	67	6
8	Perspective	12	9
9	Sound	2	8
รวม		137	38
รวมทั้งหมด		**772**	**252**

จุฬาลงกรณ์มหาวิทยาลัย

ตารางที่ 6-3 การกำหนดหมวดหมู่หลัก หมวดหมู่ย่อยของคำ และจำนวนคำในแต่ละหมวดหมู่ของคำที่ปรากฏอยู่ทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 (กลุ่มที่ 3)

ลำดับที่	หมวดหมู่	คำทั้งหมด
What		
1	Body	6
2	Family	4
3	Food	3
4	Household	5
5	Pronoun	2
6	Sport	1
7	Animal	1

ลำดับที่	หมวดหมู่	คำทั้งหมด
8	Place	2
รวม		24
Size		
1	Size (ขยายนาม)	2
2	Size (ขยายกริยา)	2
รวม		4
Color		
1	Color	4
รวม		4
Number		
รวม		0
Shape		
1	Shape	1
รวม		1
Where		
1	Position (ขยายนาม)	2
2	Position (ขยายกริยา)	1
รวม		3
When		
รวม		0
Background		
รวม		0
Movement		
1	Verb0 (กริยาทั่ว ๆ ไป)	2
2	Verb1 (กริยาที่กระทำบริเวณใบหน้า)	6
3	Verb2 (กริยาที่กระทำตั้งแต่ลำคอลงไปถึงเอว)	5
4	Verb3 (กริยาที่กระทำตั้งแต่เอวลงไป)	9
รวม		22
Mood		

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ลำดับที่	หมวดหมู่	คำทั้งหมด
1	Emotion	3
รวม		3
Perspective		
1	Perspective (ขยายนาม)	2
2	Perspective (ขยายกริยา)	1
รวม		3
Sound		
1	Sound (ขยายนาม)	1
2	Sound (ขยายกริยา)	1
รวม		2
รวมทั้งสิ้น		**66**

ตารางที่ 6-4 การกำหนดหน้าที่ของคำ (POS) ที่ปรากฏอยู่ทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 (กลุ่มที่ 3)

ลำดับที่	หมวดหมู่	คำทั้งหมด
คำนาม (N)		
1	Body	6
2	Family	4
3	Food	3
4	Household	5
5	Pronoun	2
6	Sport	1
7	Animal	1
8	Place	2
รวม		24
คำกริยา (V)		
1	Verb1 (กริยาทั่ว ๆ ไป)	2
2	Verb1 (กริยาที่กระทำตั้งแต่ลำคอลงไปถึงเอว)	6

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ลำดับที่	หมวดหมู่	คำทั้งหมด
3	Verb2 (กริยาที่กระทำบริเวณใบหน้า)	5
4	Verb3 (กริยาที่กระทำตั้งแต่เอวลงไป)	9
5	Emotion	3
รวม		25
ส่วนขยายนาม หรือ ส่วนขยายส่วนขยาย (ADJ)		
1	Size	2
2	Color	4
3	Position	2
4	Possessive	2
5	Perspective	1
6	Sound	1
รวม		12
ส่วนขยายกริยา (ADV)		
1	Size	2
2	Position	1
3	Perspective	1
4	Sound	1
รวม		5
รวมทั้งหมด		**66**

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

6.3 การกำหนดและระบุหน้าที่ของคำที่ใช้ในการทดสอบระบบตามระบบอ้างอิง (RIPOS)

การระบุหน้าที่ของคำที่จะใช้ในการทดสอบระบบตามแนวคิดการสร้างคลังข้อความ RIPOS จะใช้คำจากชุดคำกลุ่มที่ 3 หรือคำที่ปรากฏอยู่ทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 โดยมีรายละเอียดการระบุหน้าที่ของคำ และจำนวนของคำในแต่ละคำที่ระบุ ดังตารางที่ 6-5

ตารางที่ 6-5 การกำหนดหน้าที่ของคำตามระบบอ้างอิง (RIPOS)

คำที่ใช้ระบุ	คำอธิบาย	จำนวนคำทั้งหมด
NCM	Common noun	22
PRN	Pronoun	2
VBG	General Verb	25
WSG	General wisate	17
รวม		**66**

การระบุหน้าที่ของคำตามแนวคิดการสร้างคลังข้อความ RIPOS นั้นจริง ๆ แล้วมีการแบ่งคำที่ใช้ระบุออกเป็น 17 กลุ่มด้วยกันตามตารางที่ 4-2 ในบทที่ 4 แต่ในการระบุคำกลุ่มที่ 3 นี้ จะมีการระบุเพียง 4 กลุ่มเท่านั้น เนื่องจากคำในกลุ่มที่ 3 นี้มีจำนวนไม่มากและเป็นคำมีมาจากคลังข้อความที่ใช้ในการฝึกฝนและทดสอบระบบเป็นประโยคความเดียวเท่านั้น และเป็นประโยคที่ไม่ซับซ้อน

6.4 การสร้างชุดทดสอบและประโยคทดสอบ

การทดสอบอัลกอริทึมของระบบในงานวิจัยนี้ ได้ทำการทดสอบความถูกต้องในการสร้างประโยค โดยสร้างชุดทดสอบแบ่งตามจำนวนคำที่ป้อนเข้าสู่ระบบ ตั้งแต่ 3 – 6 คำ และพิจารณาสมาชิกที่เป็นไปได้ในแต่ละชุดทดสอบตามหน้าที่ของคำที่เหมาะสมกับหลักไวยากรณ์ภาษาไทย รวมทั้งสิ้น 15 ชุดทดสอบ ดังตารางที่ 6-6

ตารางที่ 6-6 ชุดทดสอบที่แบ่งตามจำนวนคำที่ป้อนเข้าสู่ระบบตั้งแต่ 3-6 คำ

ชุดทดสอบ	จำนวนคำ	สมาชิก
1	3	$\{N_1, ADJ_1, V_1\}$
2	3	$\{N_1, V_1, ADV_1\}$
3	3	$\{N_1, N_2, V_1\}$
4	4	$\{N_1, ADJ_1, ADJ_2, V_1\}$
5	4	$\{N_1, ADJ_1, V_1, ADV_1\}$
6	4	$\{N_1, N_2, ADJ_1, V_1\}$
7	4	$\{N_1, N_2, V_1, ADV_1\}$
8	5	$\{N_1, ADJ_1, ADJ_2, ADJ_3, V_1\}$

9	5	{N_1, ADJ_1, ADJ_2, V_1, ADV_1}
10	5	{N_1, N_2, ADJ_1, ADJ_2, V_1}
11	5	{N_1, N_2, ADJ_1, V_1, ADV_1}
12	6	{N_1, ADJ_1, ADJ_2, ADJ_3, ADJ_4, V_1}
13	6	{N_1, ADJ_1, ADJ_2, ADJ_3, V_1, ADV_1}
14	6	{N_1, N_2, ADJ_1, ADJ_2, ADJ_3, V_1}
15	6	{N_1, N_2, ADJ_1, ADJ_2, V_1, ADV_1}

การได้มาซึ่งประโยคทดสอบจากทั้ง 15 ชุดทดสอบ ได้มาจากการสร้างประโยคที่เป็นไปได้ทั้งหมดจากคำที่มาจากการแบ่งกลุ่มตามแหล่งที่มาของคำ โดยมีทั้งหมด 3 กลุ่มด้วยกัน ตามหัวข้อที่ 4.2 ได้แก่ คำที่มาจากคลังข้อความ TELL-S (กลุ่มที่ 1) คำที่มาจากคลังข้อความ BEST2010 (กลุ่มที่ 2) และคำที่ปรากฏอยู่ทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 (กลุ่มที่ 3) ซึ่งแบ่งคำตาม POS ที่ถูกระบุไว้ จากนั้นนำคำเหล่านั้นมาสร้างประโยคที่เป็นไปได้ทั้งหมดแบบพบเจอกันหมดแล้วจะทำการสุ่มประโยคจากทั้งหมดที่ถูกสร้าง มาทำการเลือกประโยคที่ถูกต้องตามหลักไวยากรณ์ภาษาไทย และมีความถูกต้องเหมาะสมที่จะเป็นประโยคภาษาไทยธรรมชาติที่จะใช้ทดสอบด้วยผู้เชี่ยวชาญที่พูดและสื่อสารภาษาไทยได้อย่างคล่องแคล่ว ตัวอย่างประโยคทดสอบที่ถูกเลือกแสดงในภาพที่ 6-1

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ภาพที่ 6-1 ตัวอย่างประโยคทดสอบที่ถูกเลือกในชุดทดสอบที่ 3 ที่สร้างแบบพบกันหมดจากชุดคำที่มาจากคลังข้อความ BEST2010 (กลุ่มที่ 2)

สำหรับประโยคทดสอบที่ถูกเลือกมานั้นจะพิจารณาตามกลุ่มของคำทั้ง 3 กลุ่ม ร่วมกับชุดทดสอบในตารางที่ 6-6 โดยมีรายละเอียดของจำนวนประโยคทดสอบทั้งหมดดังตารางที่ 6-7

ตารางที่ 6-7 จำนวนประโยคทดสอบที่ถูกเลือก

กลุ่มคำที่นำมาสร้างประโยคทดสอบ	ชุดทดสอบ	จำนวนประโยคทดสอบที่ถูกเลือก
คำจากกลุ่มที่ 1	ชุดทดสอบที่ 3	15
	ชุดทดสอบที่ 4	15
	ชุดทดสอบที่ 5	15
	ชุดทดสอบที่ 6	15
	ชุดทดสอบที่ 7	15
	ชุดทดสอบที่ 8	15
	ชุดทดสอบที่ 9	15
	ชุดทดสอบที่ 10	15
	ชุดทดสอบที่ 11	15
	ชุดทดสอบที่ 12	15
	ชุดทดสอบที่ 13	15
	ชุดทดสอบที่ 14	15
	ชุดทดสอบที่ 15	15
รวม		195
คำจากกลุ่มที่ 2	ชุดทดสอบที่ 3	15
	ชุดทดสอบที่ 4	15
	ชุดทดสอบที่ 5	15
	ชุดทดสอบที่ 6	15
	ชุดทดสอบที่ 7	15
	ชุดทดสอบที่ 8	15
	ชุดทดสอบที่ 9	15
	ชุดทดสอบที่ 10	15
	ชุดทดสอบที่ 11	15
	ชุดทดสอบที่ 12	15
	ชุดทดสอบที่ 13	15
	ชุดทดสอบที่ 14	15
	ชุดทดสอบที่ 15	15

กลุ่มคำที่นำมาสร้างประโยคทดสอบ	ชุดทดสอบ	จำนวนประโยคทดสอบที่ถูกเลือก
รวม		195
คำจากกลุ่มที่ 3	ชุดทดสอบที่ 3	5
	ชุดทดสอบที่ 4	5
	ชุดทดสอบที่ 5	5
	ชุดทดสอบที่ 6	5
	ชุดทดสอบที่ 7	5
	ชุดทดสอบที่ 9	5
	ชุดทดสอบที่ 10	5
	ชุดทดสอบที่ 11	5
	ชุดทดสอบที่ 14	5
	ชุดทดสอบที่ 15	5
รวม		50
รวมทั้งสิ้น		**440**

การสร้างประโยคคำตอบจากกลุ่มคำที่มาจากคลังข้อความที่แตกต่างกันตามตารางที่ 6-7 พบว่า การสร้างประโยคคำตอบจากชุดคำกลุ่มที่ 1 ร่วมกับชุดทดสอบ 13 ชุดทดสอบ มีประโยคทดสอบที่ถูกเลือกทั้งสิ้น 195 ประโยค เช่นเดียวกับการสร้างประโยคคำตอบจากชุดคำกลุ่มที่ 2 ร่วมกับชุดทดสอบ 13 ชุดทดสอบ มีประโยคทดสอบที่ถูกเลือกทั้งสิ้น 195 ประโยคเช่นกัน สาเหตุที่ไม่นำชุดทดสอบที่ 1 และชุดทดสอบที่ 2 มาใช้ เนื่องจากเป็นชุดทดสอบที่ไม่มีความกำกวมจึงมีความถูกต้องในการสร้างประโยค 100 เปอร์เซ็นต์ สำหรับการสร้างประโยคคำตอบจากชุดคำกลุ่มที่ 3 ร่วมกับชุดทดสอบ 10 ชุดทดสอบ จะมีประโยคทดสอบที่ถูกเลือกทั้งสิ้น 50 ประโยค เนื่องจากไม่มีความจำเป็นต้องพิจารณากรณีชุดทดสอบที่ไม่มีความคลุมเครือ เช่น ชุดทดสอบที่ 1 และชุดทดสอบที่ 2 และไม่พิจารณาชุดทดสอบที่ 8 ชุดทดสอบที่ 12 และชุดทดสอบที่ 13 เนื่องจากประโยคที่สร้างได้นั้นไม่เกิดขึ้นจริง นอกจากนี้ได้ลดจำนวนประโยคที่ถูกเลือกในชุดคำจากกลุ่มที่ 3 ลง เหลือชุดทดสอบละ 5 ประโยค เนื่องจากคำในกลุ่มที่ 3 นั้นมีจำนวนน้อย เพราะคำในกลุ่มนี้เลือกเฉพาะคำที่ปรากฏในคำกลุ่มที่ 1 และกลุ่มที่ 2 เท่านั้น

6.5 การวัดผลทดลอง

การวัดประสิทธิภาพของอัลกอริทึมที่นำเสนอในการเรียงลำดับคำในประโยคภาษาไทย จะพิจารณาจากเปอร์เซ็นต์ความถูกต้องของประโยคที่ระบบสร้างเปรียบเทียบกับประโยคคำตอบที่ถูกเลือกโดยผู้เชี่ยวชาญ โดยวัดเปอร์เซ็นต์ความถูกต้องได้ตามสมการที่ 6.1

$$\text{เปอร์เซ็นต์ความถูกต้อง} = \frac{\text{จำนวนประโยคที่ได้คะแนน}}{\text{จำนวนประโยคทั้งหมด}} \times 100 \qquad (6.1)$$

โดยกำหนดให้ ประโยคที่ได้คะแนน หมายถึง ประโยคที่ระบบสร้างตรงกับประโยคที่ผู้เชี่ยวชาญเลือก โดยถ้าระบบสร้างประโยคตรงกับประโยคทดสอบที่ถูกเลือกจะให้คะแนนเท่ากับ 1 แต่ถ้าตอบไม่ตรงจะให้ 0 คะแนน

การวัดผลการทดลองในงานวิจัยนี้ แบ่งออกเป็น 2 ส่วน ดังนี้

1. การวัดผลโดยการเปรียบเทียบประโยคที่ระบบสร้างกับประโยคทดสอบที่ถูกเลือกโดยผู้เชี่ยวชาญ 1 คนจากคำในกลุ่มที่ 1-3

   กำหนดให้ประโยคที่ถูกเลือกมาเป็นประโยคทดสอบโดยผู้เชี่ยวชาญตามขั้นตอนการเลือกประโยคในหัวข้อที่ 6.4 คือประโยคคำตอบหรือประโยคเป้าหมาย ซึ่งจะใช้เปรียบเทียบกับประโยคที่ระบบสร้างขึ้น หากตรงกันจะถือว่าประโยคที่ระบบสร้างขึ้นมานั้นได้คะแนน

2. การวัดผลโดยการเปรียบเทียบประโยคที่ระบบสร้างกับประโยคที่ถูกสร้างโดยผู้เชี่ยวชาญ 3 คนจากคำในกลุ่มที่ 3

   กำหนดให้ประโยคคำตอบถูกสร้างโดยผู้เชี่ยวชาญ 3 คน คนที่ 1 เพศหญิงอายุ 25 ปี อาชีพนักศึกษา วุฒิการศึกษาระดับปริญญาโท คนที่ 2 เพศชายอายุ 40 ปี อาชีพนักธุรกิจ และคนที่ 3 เพศหญิงอายุ 55 ปี อาชีพครู โดยผู้เชี่ยวชาญทุกคนจะได้ชุดของคำที่จะใช้แต่งประโยคคำตอบ เพื่อเปรียบเทียบประโยคที่สร้างโดยคน กับประโยคที่สร้างโดยระบบ หากประโยคที่ระบบสร้างตรงกับประโยคที่ผู้เชี่ยวชาญสร้าง จะถือว่าประโยคที่ระบบสร้างขึ้นมานั้นได้คะแนน

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

6.6 การทดลอง

6.6.1 การทดลองที่ 1: การเรียงลำดับคำในประโยคภาษาไทย โดยอาศัยหลักการเอ็นแกรมร่วมกับวิธีระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ BEST2010

การทดลองนี้เป็นการทดลองเพื่อวัดประสิทธิภาพของอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอ โดยอาศัยหลักการเอ็นแกรมร่วมกับวิธีระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอ บนคลังข้อความ BEST2010 โดยในการทดลองนี้จะแบ่งการทดลองออกเป็น 3 กลุ่มย่อย ตามกลุ่มของคำที่นำมาใช้สร้างประโยคทดสอบ ดังนี้

กลุ่มที่ 1 หรือ WS1 คือ คำที่มาจากคลังข้อความ TELL-S

กลุ่มที่ 2 หรือ WS2 คือ คำที่มาจากคลังข้อความ BEST2010 และ

กลุ่มที่ 3 หรือ WS3 คือ คำที่ปรากฏอยู่ทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010

โดยใช้ชุดทดสอบ 15 ชุดทดสอบและจำนวนประโยคตามตารางที่ 6-7

6.6.2 การทดลองที่ 2: การเรียงลำดับคำในประโยคภาษาไทย โดยอาศัยหลักการเอ็นแกรมร่วมกับวิธีระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ TELL-S

การทดลองนี้เป็นการทดลองเพื่อวัดประสิทธิภาพของอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอ โดยอาศัยหลักการเอ็นแกรมร่วมกับวิธีระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอ บนคลังข้อความ TELL-S โดยในการทดลองนี้จะแบ่งการทดลองออกเป็น 3 กลุ่มย่อย ตามกลุ่มของคำที่นำมาใช้สร้างประโยคทดสอบ ดังนี้

กลุ่มที่ 1 หรือ WS1 คือ คำที่มาจากคลังข้อความ TELL-S

กลุ่มที่ 2 หรือ WS2 คือ คำที่มาจากคลังข้อความ BEST2010 และ

กลุ่มที่ 3 หรือ WS3 คือ คำที่ปรากฏอยู่ทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010

โดยใช้ชุดทดสอบทั้ง 15 ชุดทดสอบและมีจำนวนประโยคตามตารางที่ 6-7

ในการทำการทดลองที่ 1 และการทดลองที่ 2 ข้างต้นนี้ มีวัตถุประสงค์ในการทำการทดลอง 2 ประการ ดังนี้

ประการที่หนึ่ง เพื่อพิจารณาเปรียบเทียบว่าคลังข้อความใดเหมาะสมจะใช้เป็นคลังข้อความฝึกฝน เพื่อนำคะแนนจากการฝึกฝนมาใช้ในอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทย

ประการที่สอง เพื่อพิจารณาว่ากลุ่มของคำกลุ่มใดเหมาะสมที่จะนำมาใช้เป็นชุดคำในประโยคทดสอบ

โดยในการทดลองข้างต้นนี้ ผู้วิจัยได้ตั้งสมมติฐานว่าการใช้ชุดข้อมูลฝึกฝนบนคลังข้อความ TELL-S น่าจะให้ผลการทดลองที่ดีกว่าการใช้ชุดข้อมูลฝึกฝนบนคลังข้อความ BEST2010 และกลุ่มของคำกลุ่มที่เหมาะสมจะนำมาใช้เป็นชุดคำในประโยคทดสอบคือ ชุดคำในกลุ่มที่ 3 (WS3)

6.6.3 การทดลองที่ 3: การหาลักษณะสำคัญของการระบุหมวดหมู่ของคำในคลังข้อความ

การทดลองนี้มีวัตถุประสงค์เพื่อวัดประสิทธิภาพของอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอ และเพื่อหาลักษณะสำคัญของการระบุหมวดหมู่ของคำในคลังข้อความ BEST2010 และ TELL-S โดยอาศัยหลักการเอ็นแกรม และใช้คำจากกลุ่มที่ 3 หรือ WS3 ในการสร้างประโยคทดสอบ โดยแบ่งการทดลองออกเป็น 8 กรณี ดังนี้

1. กรณีใช้คะแนนไบแกรมของคำจากคลังข้อมูล TELL-S เพียงอย่างเดียว
2. กรณีใช้คะแนนไบแกรมของคำจากคลังข้อมูล BEST2010 เพียงอย่างเดียว
3. กรณีใช้คะแนนไบแกรมของคำและคะแนนไบแกรมของหมวดหมู่หลักจากคลังข้อมูล TELL-S
4. กรณีใช้คะแนนไบแกรมของคำและคะแนนไบแกรมของหมวดหมู่หลักจากคลังข้อมูล BEST2010
5. กรณีใช้คะแนนไบแกรมของคำและคะแนนไบแกรมของหมวดหมู่ย่อยจากคลังข้อมูล TELL-S
6. กรณีใช้คะแนนไบแกรมของคำและคะแนนไบแกรมของหมวดหมู่ย่อยจากคลังข้อมูล BEST2010
7. กรณีใช้คะแนนไบแกรมของคำ คะแนนไบแกรมของหมวดหมู่หลัก และคะแนนไบแกรมของหมวดหมู่ย่อย (การระบุหมวดหมู่ที่งานวิจัยนี้นำเสนอ) จากคลังข้อมูล TELL-S

8. กรณีใช้คะแนนไบแกรมของคำ คะแนนไบแกรมของหมวดหมู่หลัก และคะแนนไบแกรมของหมวดหมู่ย่อย (การระบุหมวดหมู่ที่งานวิจัยนี้นำเสนอ) จากคลังข้อมูล BEST2010

โดยในการทดลองนี้ได้ตั้งสมมติฐานว่ากรณีใช้คะแนนไบแกรมของคำ คะแนนไบแกรมของหมวดหมู่หลัก และคะแนนไบแกรมของหมวดหมู่ย่อย (การระบุหมวดหมู่ที่งานวิจัยนี้นำเสนอ) จากคลังข้อมูล TELL-S จะมีผลการทดลองที่ดีที่สุด และคะแนนไบแกรมของหมวดหมู่หลักจะเป็นลักษณะสำคัญในการคำนวณหาประโยคคำตอบของระบบมากกว่าคะแนนไบแกรมของหมวดหมู่ย่อย

6.6.4 การทดลองที่ 4: การเรียงลำดับคำในประโยคภาษาไทย โดยอาศัยหลักการเอ็นแกรมร่วมกับวิธีระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS บนคลังข้อความ TELL-S

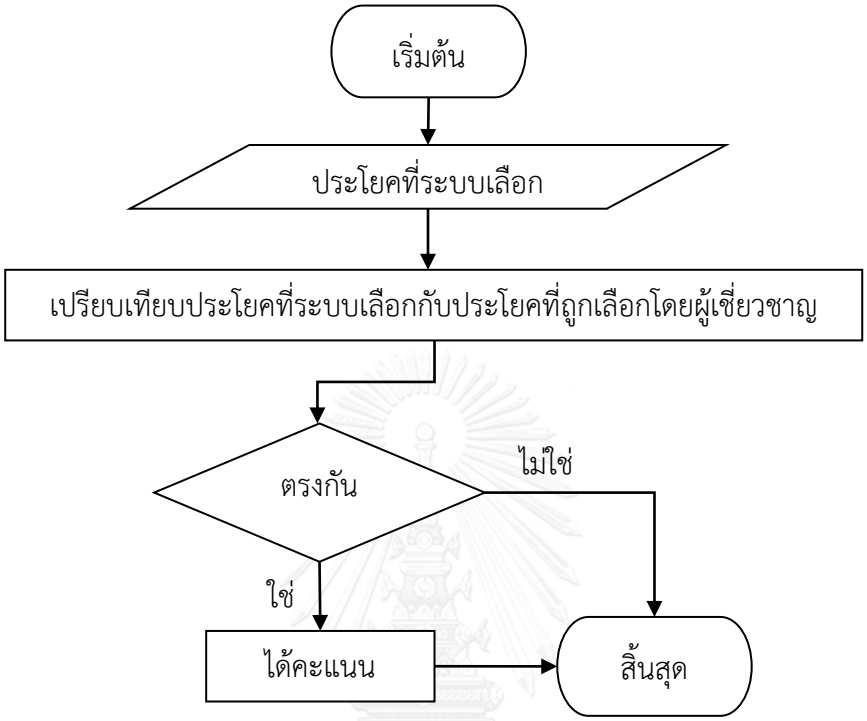
การทดลองนี้มีวัตถุประสงค์เพื่อวัดประสิทธิภาพของอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทย เมื่อใช้วิธีการระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS บนคลังข้อความ TELL-S โดยอาศัยหลักการเอ็นแกรมในการเลือกประโยคคำตอบ และใช้คำจากกลุ่มที่ 3 หรือ WS3 ในการสร้างประโยคทดสอบ สำหรับการวัดผลการทดลองจะพิจารณาประโยคที่มีความยาวตั้งแต่ 3-6 คำ

เมื่อเปรียบเทียบผลการทดลองที่ 3 ในกรณีที่ 7 คือ การระบุหมวดหมู่และหน้าที่ของคำตามแนวคิดที่นำเสนอ กับการทดลองที่ 4 นี้ ผู้วิจัยได้ตั้งสมมติฐานว่า วิธีการระบุหมวดหมู่และหน้าที่ของคำตามแนวคิดที่นำเสนอ จะมีประสิทธิภาพในการเลือกประโยคคำตอบได้ถูกต้องตามประโยคเป้าหมายสูงกว่าการระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS

6.6.5 การทดลองที่ 5: การเรียงลำดับคำในประโยคภาษาไทย โดยอาศัยหลักการเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST) บนคลังข้อความ TELL-S

การทดลองนี้มีวัตถุประสงค์เพื่อวัดประสิทธิภาพของอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอ เมื่อใช้หลักการเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST) ว่ามีประสิทธิภาพสูงกว่าการใช้หลักการเอ็นแกรมเพียงอย่างเดียวหรือไม่ โดยผู้วิจัยตั้งสมมติฐานว่า อัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอเมื่อใช้หลักการเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST) จะมีประสิทธิภาพในการแต่งประโยคคำตอบได้ตรงตามประโยคเป้าหมายสูงที่สุด

ในการทดลองนี้ได้ใช้ชุดข้อมูลฝึกฝนบนคลังข้อความ TELL-S และใช้คำจากกลุ่มที่ 3 หรือ WS3 ในการสร้างประโยคทดสอบ สำหรับการวัดผลการทดลองจะพิจารณาประโยคที่มีความยาวตั้งแต่ 3-6 คำ และมีขั้นตอนการให้คะแนนตามภาพที่ 6-2



ภาพที่ 6-2 ขั้นตอนการให้คะแนนประโยคคำตอบ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 7
ผลการทดลองและวิเคราะห์ผลการทดลอง

ในบทนี้ได้นำเสนอผลการทดลองและการวิเคราะห์การทดลองที่ได้จากบทที่ 6 โดยมีรายละเอียดผลการทดลอง ได้แก่ การประเมินผลประสิทธิภาพของระบบที่นำเสนอโดยอาศัยหลักการเอ็นแกรม การประเมินผลการทดลองหาลักษณะสำคัญของการระบุหมวดหมู่ของคำในคลังข้อความ การประเมินผลประสิทธิภาพของระบบอ้างอิงโดยอาศัยหลักการเอ็นแกรมร่วมกับวิธีระบุหน้าที่ของคำแบบ RIPOS และการประเมินผลประสิทธิภาพของแนวทางที่นำเสนอโดยอาศัยหลักการเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST)

7.1 การประเมินผลการการทดลองระบบที่นำเสนอโดยอาศัยหลักการเอ็นแกรม

การประเมินผลการทดลองระบบที่นำเสนอโดยอาศัยหลักการเอ็นแกรมเป็นประเมินผลการทดลองเพื่อวัดประสิทธิภาพของอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอ โดยอาศัยหลักการเอ็นแกรมร่วมกับวิธีระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอ บนคลังข้อความต่างที่แตกต่างกัน ได้แก่ คลังข้อความ TELL-S และคลังข้อความ BEST2010 และทดสอบกับชุดทดสอบที่แตกต่างกัน ได้แก่ ชุดทดสอบที่ประกอบด้วยประโยคทดสอบจากคำที่มาจากคลังข้อความ TELL-S (WS1) ชุดทดสอบที่ประกอบด้วยประโยคทดสอบจากคำที่มาจากคลังข้อความ BEST2010 (WS2) และชุดทดสอบที่ประกอบด้วยประโยคทดสอบจากคำที่ปรากฏทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 (WS3) โดยในการประเมินผลการทดสอบระบบที่นำเสนอนี้ ได้นำเสนอตามตารางต่อไปนี้

Chulalongkorn University

ตารางที่ 7-1 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ TELL-S และสร้างประโยคทดสอบจากคำที่มาจากคลังข้อความ TELL-S (WS1)

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
3 คำ			
n=2, v=1	15	13	86.67
รวม	**15**	**13**	**86.67**

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
4 คำ			
n=1, adj=2, v=1	15	12	80.00
n=1, adj=1, v=1, adv=1	15	11	73.33
n=2, adj=1, v=1	15	11	73.33
n=2, v=1, adv=1	15	11	73.33
รวม	**60**	**45**	**75.00**
5 รูป			
n=1, adj=3, v=1	15	8	53.33
n=1, adj=2, v=1, adv=1	15	9	60.00
n=2, adj=2, v=1	15	3	20.00
n=2, adj=1, v=1, adv=1	15	6	40.00
รวม	**60**	**26**	**43.33**
6 รูป			
n=1, adj=4, v=1	15	7	46.67
n=1, adj=3, v=1, adv=1	15	6	40.00
n=2, adj=3, v=1	15	3	20.00
n=2, adj=2, v=1, adv=1	15	4	26.67
รวม	**60**	**20**	**33.33**
รวมทั้งหมด	**195**	**104**	**53.33**

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตารางที่ 7-2 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ TELL-S และสร้างประโยคทดสอบจากคำที่มาจากคลังข้อความ BEST2010 (WS2)

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
3 คำ			
n=2, v=1	15	11	73.33
รวม	**15**	**11**	**73.33**
4 คำ			
n=1, adj=2, v=1	15	9	60.00
n=1, adj=1, v=1, adv=1	15	13	86.67
n=2, adj=1, v=1	15	10	66.67
n=2, v=1, adv=1	15	9	60.00
รวม	**60**	**41**	**68.33**
5 รูป			
n=1, adj=3, v=1	15	7	46.67
n=1, adj=2, v=1, adv=1	15	3	20.00
n=2, adj=2, v=1	15	5	33.33
n=2, adj=1, v=1, adv=1	15	7	46.67
รวม	**60**	**22**	**36.67**
6 รูป			
n=1, adj=4, v=1	15	9	60.00
n=1, adj=3, v=1, adv=1	15	2	13.33
n=2, adj=3, v=1	15	2	13.33
n=2, adj=2, v=1, adv=1	15	6	40.00

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
รวม	**60**	**19**	**31.67**
รวมทั้งหมด	**195**	**93**	**47.69**

ตารางที่ 7-3 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ BEST2010 และสร้างประโยคทดสอบจากคำที่มาจากคลังข้อความ TELL-S (WS1)

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
3 คำ			
n=2, v=1	15	10	66.67
รวม	**15**	**10**	**66.67**
4 คำ			
n=1, adj=2, v=1	15	8	53.33
n=1, adj=1, v=1, adv=1	15	10	66.67
n=2, adj=1, v=1	15	3	20.00
n=2, v=1, adv=1	15	8	53.33
รวม	**60**	**29**	**48.33**
5 รูป			
n=1, adj=3, v=1	15	4	26.67
n=1, adj=2, v=1, adv=1	15	4	26.67
n=2, adj=2, v=1	15	4	26.67
n=2, adj=1, v=1, adv=1	15	4	26.67

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
รวม	**60**	**16**	**26.67**
6 รูป			
n=1, adj=4, v=1	15	4	26.67
n=1, adj=3, v=1, adv=1	15	5	33.33
n=2, adj=3, v=1	15	2	13.33
n=2, adj=2, v=1, adv=1	15	3	20.00
รวม	**60**	**14**	**23.33**
รวมทั้งหมด	**195**	**69**	**35.38**

ตารางที่ 7-4 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ BEST2010 และสร้างประโยคทดสอบจากคำที่มาจากคลังข้อความ BEST2010 (WS2)

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
3 คำ			
n=2, v=1	15	11	73.33
รวม	**15**	**11**	**73.33**
4 คำ			
n=1, adj=2, v=1	15	7	46.67
n=1, adj=1, v=1, adv=1	15	11	73.33
n=2, adj=1, v=1	15	11	73.33
n=2, v=1, adv=1	15	10	66.67

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
รวม	**60**	**39**	**65.00**
5 รูป			
n=1, adj=3, v=1	15	8	53.33
n=1, adj=2, v=1, adv=1	15	3	20.00
n=2, adj=2, v=1	15	6	40.00
n=2, adj=1, v=1, adv=1	15	7	46.67
รวม	**60**	**24**	**40.00**
6 รูป			
n=1, adj=4, v=1	15	11	73.33
n=1, adj=3, v=1, adv=1	15	3	20.00
n=2, adj=3, v=1	15	3	20.00
n=2, adj=2, v=1, adv=1	15	7	46.67
รวม	**60**	**24**	**40.00**
รวมทั้งหมด	**195**	**98**	**50.26**

ตารางที่ 7-5 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ TELL-S และสร้างประโยคทดสอบจากคำที่ปรากฏทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 (WS3)

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
3 คำ			
n=2, v=1	5	5	100.00

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
รวม	**5**	**5**	**100.00**
4 คำ			
n=1, adj=2, v=1	5	4	80.00
n=1, adj=1, v=1, adv=1	5	2	40.00
n=2, adj=1, v=1	5	2	80.00
n=2, v=1, adv=1	5	4	80.00
รวม	**20**	**14**	**70.00**
5 รูป			
n=1, adj=2, v=1, adv=1	5	2	40.00
n=2, adj=2, v=1	5	2	40.00
n=2, adj=1, v=1, adv=1	5	2	40.00
รวม	**15**	**6**	**40.00**
6 รูป			
n=2, adj=3, v=1	5	2	40.00
n=2, adj=2, v=1, adv=1	5	2	40.00
รวม	**10**	**4**	**40.00**
รวมทั้งหมด	**50**	**29**	**58.00**

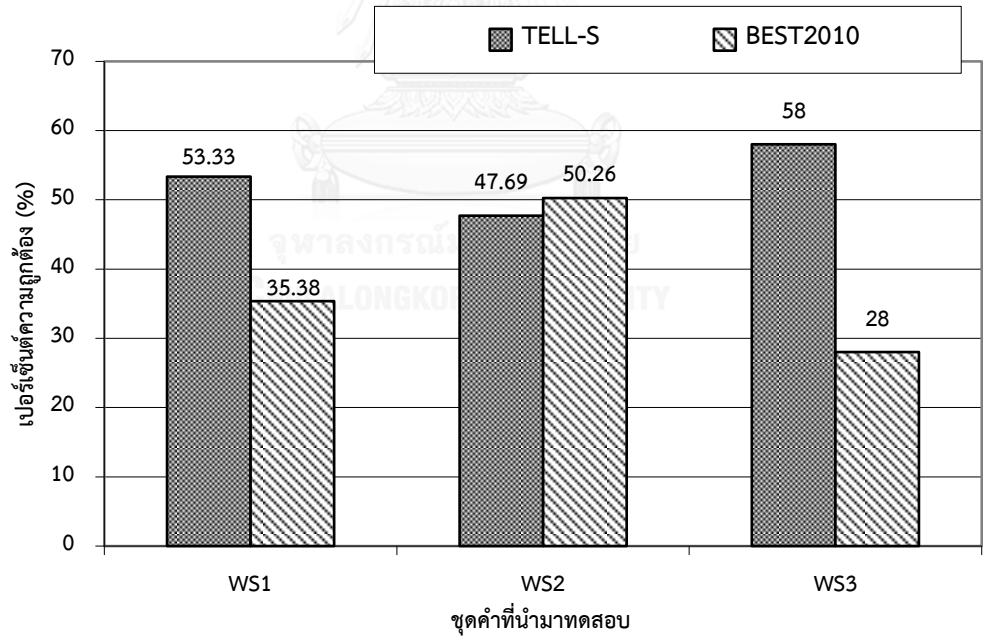
จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตารางที่ 7-6 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายเมื่อระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ BEST2010 และสร้างประโยคทดสอบจากคำที่ปรากฏทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 (WS3)

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
3 คำ			
n=2, v=1	5	3	60.00
รวม	**5**	**3**	**60.00**
4 คำ			
n=1, adj=2, v=1	5	4	80.00
n=1, adj=1, v=1, adv=1	5	1	20.00
n=2, adj=1, v=1	5	1	20.00
n=2, v=1, adv=1	5	2	40.00
รวม	**20**	**8**	**40.00**
5 รูป			
n=1, adj=2, v=1, adv=1	5	2	40.00
n=2, adj=2, v=1	5	0	0.00
n=2, adj=1, v=1, adv=1	5	0	0.00
รวม	**15**	**2**	**13.33**
6 รูป			
n=2, adj=3, v=1	5	0	0.00
n=2, adj=2, v=1, adv=1	5	1	20.00
รวม	**10**	**1**	**10.00**
รวมทั้งหมด	**50**	**14**	**28.00**

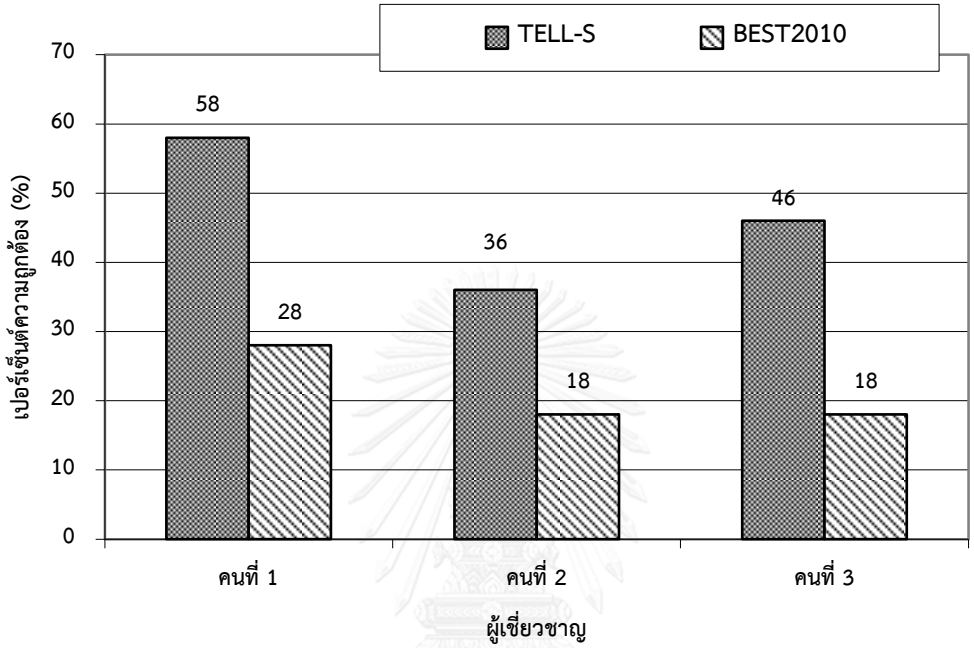
จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จากผลการทดสอบในตารางที่ 7-1 ถึงตารางที่ 7-4 และภาพที่ 7-1 พบว่าผลการทดลองในตารางที่ 7-1 เมื่อระบุหมวดหมู่และหน้าที่ของคำตามแนวคิดที่นำเสนอบนคลังข้อความ TELL-S และใช้คำจาก WS1 มาสร้างประโยคทดสอบ จะมีเปอร์เซ็นต์ความถูกต้องเฉลี่ยสูงสุด คือ 53.33 เปอร์เซ็นต์ รองลงมา คือ ผลการทดลองในตารางที่ 7-4 มีเปอร์เซ็นต์ความถูกต้องในการเลือกประโยคคำตอบเฉลี่ยอยู่ที่ 50.26 เปอร์เซ็นต์ นอกจากนี้ทดลองนี้ยังชี้ให้เห็นว่า เมื่อใช้คลังข้อความเดียวกันทั้งในการฝึกสอนการระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอและสร้างประโยคทดสอบ จะส่งผลให้เปอร์เซ็นต์ความถูกต้องในการเลือกประโยคคำตอบสูงกว่าการใช้คลังข้อมูลฝึกสอนและสร้างประโยคทดสอบที่แตกต่างกัน ดังนั้นเพื่อความชัดเจนในการวัดผล ในงานวิจัยนี้จึงได้สร้างชุดคำที่ใช้สำหรับการทดสอบระบบที่มีความเท่าเทียมกันเพื่อวัดผลการทดลองระหว่างการฝึกสอนการระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความ TELL-S และคลังข้อความ BSET 2010 สำหรับชุดคำกลุ่มนี้ คือชุดคำกลุ่มที่ 3 หรือ WS3 ซึ่งเป็นชุดคำสำหรับสร้างประโยคทดสอบ โดยจะเป็นคำที่ปรากฏอยู่ทั้งในคลังข้อความ TELL-S และคลังข้อความ BEST2010 ผลการทดสอบดังตารางที่ 7-5 ตารางที่ 7-6 และภาพที่ 7-1



ภาพที่ 7-1 ผลการทดสอบการเรียงลำดับคำในประโยคภาษาไทยโดยอาศัยหลักการเอ็นแกรมร่วมกับวิธีระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอบนคลังข้อความที่แตกต่างกันและทดสอบบนชุดข้อมูลที่แตกต่างกัน

จากตารางแสดงผลการทดลองข้างต้น เป็นการวัดผลโดยเปรียบเทียบประโยคที่ระบบเลือกกับประโยคที่ผู้เชี่ยวชาญเพียงคนเดียวเลือกเท่านั้น เพื่อยืนยันความถูกต้องการวัดผลการทำงานของระบบ ในงานวิจัยนี้จะให้ผู้เชี่ยวชาญอีก 2 คน มาร่วมพิจารณาและเลือกประโยคคำตอบที่สร้างจากชุดคำกลุ่มที่ 3 คนละ 50 ประโยค ผลการทดสอบเป็นไปดังภาพที่ 7-2



ภาพที่ 7-2 การประเมินผลความถูกต้องเมื่อเปรียบเทียบประโยคคำตอบที่ระบบเลือกกับประโยคคำตอบที่ผู้เชี่ยวชาญ 3 คนเลือก

จากภาพที่ 7-2 แสดงให้เห็นว่าการเรียงลำดับคำในประโยคภาษาไทย เมื่อทำการฝึกสอนแบบจำลองการระบุหมวดหมู่ของคำและระบุหน้าที่ของคำตามระบบที่นำเสนอบนคลังข้อความ TELL-S จะให้ผลการเลือกประโยคคำตอบที่ถูกต้องมากกว่าการฝึกสอนการระบุหมวดหมู่ของคำและระบุหน้าที่ของคำบนคลังข้อความ BEST2010 เมื่อเปรียบเทียบกับประโยคคำตอบที่ได้จากผู้เชี่ยวชาญ 3 คน

7.2 การประเมินผลการทดลองหาลักษณะสำคัญของการระบุหมวดหมู่ของคำในคลังข้อความ

ผลการทดลองจากการประเมินหาลักษณะสำคัญของการระบุหมวดหมู่ของคำในคลังข้อความนี้ จะแบ่งกรณีการพิจารณาหาลักษณะสำคัญดังนี้

กรณีที่ 1 ใช้เฉพาะคะแนนจากไบแกรมที่ได้จากแบบจำลองของคำที่ฝึกสอนบนคลังข้อความ TELL-S ในการคำนวณหาประโยคคำตอบ

กรณีที่ 2 ใช้เฉพาะคะแนนจากไบแกรมที่ได้จากแบบจำลองของคำที่ฝึกสอนบนคลังข้อความ BEST2010 ในการคำนวณหาประโยคคำตอบ

กรณีที่ 3 ใช้คะแนนจากไบแกรมที่ได้จากแบบจำลองของคำร่วมกับคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่หลักที่ฝึกสอนบนคลังข้อความ TELL-S ในการคำนวณหาประโยคคำตอบ

กรณีที่ 4 ใช้คะแนนจากไบแกรมที่ได้จากแบบจำลองของคำร่วมกับคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่หลักที่ฝึกสอนบนคลังข้อความ BEST2010 ในการคำนวณหาประโยคคำตอบ

กรณีที่ 5 ใช้คะแนนจากไบแกรมที่ได้จากแบบจำลองของคำร่วมกับคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่ย่อยที่ฝึกสอนบนคลังข้อความ TELL-S ในการคำนวณหาประโยคคำตอบ

กรณีที่ 6 ใช้คะแนนจากไบแกรมที่ได้จากแบบจำลองของคำร่วมกับคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่ย่อยที่ฝึกสอนบนคลังข้อความ BEST2010 ในการคำนวณหาประโยคคำตอบ

กรณีที่ 7 ใช้คะแนนจากไบแกรมที่ได้จากแบบจำลองของคำร่วมกับคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่หลักและคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่ย่อยที่ฝึกสอนบนคลังข้อความ TELL-S ในการคำนวณหาประโยคคำตอบ

กรณีที่ 8 ใช้คะแนนจากไบแกรมที่ได้จากแบบจำลองของคำร่วมกับคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่หลักและคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่ย่อยที่ฝึกสอนบนคลังข้อความ BEST2010 ในการคำนวณหาประโยคคำตอบ

และชุดทดสอบที่ใช้ในการทดลองนี้คือชุดคำที่มาจากคำกลุ่มที่ 3 หรือ WS3 โดยมีผลการทดลองดังตารางที่ 7-7

จุฬาลงกรณ์มหาวิทยาลัย

ตารางที่ 7-7 ลักษณะประโยคคำตอบที่ระบบเลือกเมื่อพิจารณาลักษณะสำคัญของหมวดหมู่ต่าง ๆ และเลือกใช้ลักษณะสำคัญในกรณีที่แตกต่างกันบนคลังข้อมูลที่แตกต่างกัน

การทดลอง	**ประโยคคำตอบที่ถูกเลือกตรงตามประโยคเป้าหมาย (%)**	**ไม่สามารถเลือกประโยคคำตอบได้แต่กรองประโยคได้บางส่วน (%)**	**ไม่สามารถเลือกประโยคคำตอบได้ (%)**	**ประโยคคำตอบที่ถูกเลือกไม่ตรงตามประโยคเป้าหมาย (%)**	**คะแนนรวม**
กรณีที่ 1					
3 คำ	100.00	0.00	0.00	0.00	100.00
4 คำ	15.00	10.00	75.00	0.00	20.00

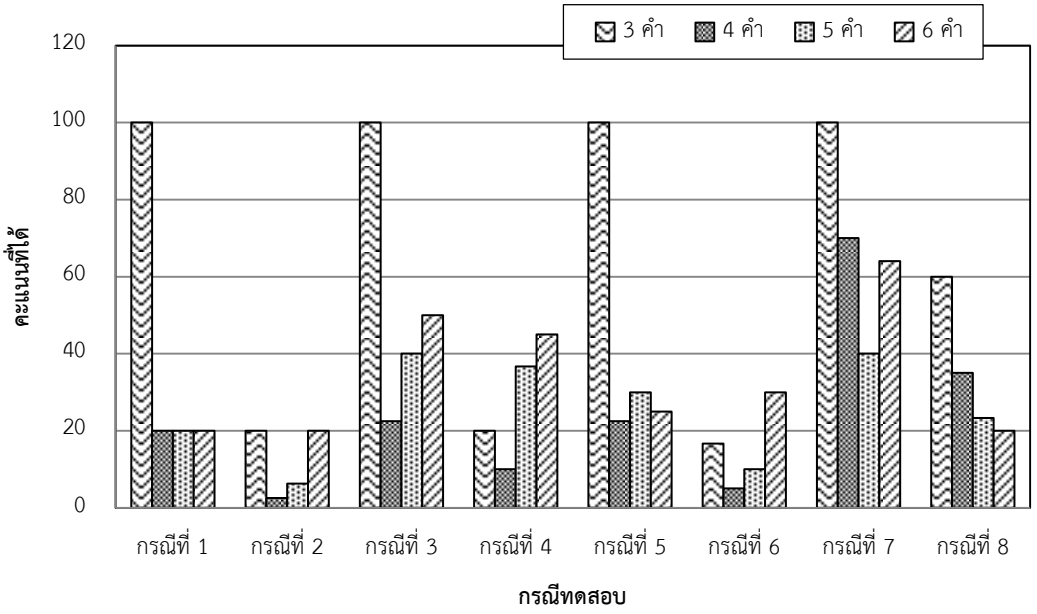
การทดลอง	ประโยคคำตอบที่ถูกเลือกตรงตามประโยคเป้าหมาย (%)	ไม่สามารถเลือกประโยคคำตอบได้แต่กรองประโยคได้บางส่วน (%)	ไม่สามารถเลือกประโยคคำตอบได้ (%)	ประโยคคำตอบที่ถูกเลือกไม่ตรงตามประโยคเป้าหมาย (%)	คะแนนรวม
5 คำ	6.67	26.67	66.67	0.00	20.00
6 คำ	0.00	40.00	60.00	0.00	20.00
กรณีที่ 2					
3 คำ	20.00	0.00	80.00	0.00	20.00
4 คำ	0.00	5.00	95.00	0.00	2.50
5 คำ	0.00	12.50	87.50	0.00	6.25
6 คำ	0.00	40.00	60.00	0.00	20.00
กรณีที่ 3					
3 คำ	100.00	0.00	0.00	0.00	100.00
4 คำ	15.00	15.00	60.00	10.00	22.50
5 คำ	13.33	53.33	33.33	0.00	40.00
6 คำ	0.00	100.00	0.00	0.00	50.00
กรณีที่ 4					
3 คำ	20.00	0.00	80.00	0.00	20.00
4 คำ	10.00	0.00	70.00	20.00	10.00
5 คำ	0.00	73.33	26.67	0.00	36.67
6 คำ	0.00	90.00	10.00	0.00	45.00
กรณีที่ 5					
3 คำ	100.00	0.00	0.00	0.00	100.00
4 คำ	20.00	5.00	75.00	0.00	22.50
5 คำ	26.67	6.67	66.67	0.00	30.00

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

การทดลอง	ประโยคคำตอบที่ถูกเลือกตรงตามประโยคเป้าหมาย (%)	ไม่สามารถเลือกประโยคคำตอบได้แต่กรองประโยคได้บางส่วน (%)	ไม่สามารถเลือกประโยคคำตอบได้ (%)	ประโยคคำตอบที่ถูกเลือกไม่ตรงตามประโยคเป้าหมาย (%)	คะแนนรวม
6 คำ	0.00	50.00	50.00	0.00	25.00
กรณีที่ 6					
3 คำ	16.67	0.00	83.33	0.00	16.67
4 คำ	5.00	0.00	95.00	0.00	5.00
5 คำ	0.00	20.00	80.00	0.00	10.00
6 คำ	10.00	40.00	50.00	0.00	30.00
กรณีที่ 7					
3 คำ	100.00	0.00	0.00	0.00	100.00
4 คำ	70.00	0.00	0.00	30.00	70.00
5 คำ	40.00	0.00	0.00	60.00	40.00
6 คำ	40.00	50.00	0.00	10.00	65.00
กรณีที่ 8					
3 คำ	60.00	0.00	0.00	40.00	60.00
4 คำ	30.00	10.00	5.00	55.00	35.00
5 คำ	13.33	20.00	0.00	66.67	23.33
6 คำ	10.00	20.00	40.00	30.00	20.00

จากตารางที่ 7-7 คะแนนที่ปรากฏในแต่ละแถวนั้น ได้มาจากการคำนวณคะแนน โดยกำหนดให้ แต่ละเปอร์เซ็นต์ของประโยคคำตอบที่ถูกเลือกตรงตามประโยคเป้าหมาย เท่ากับ 1 คะแนน และแต่ละเปอร์เซ็นต์ของการที่ไม่สามารถเลือกประโยคคำตอบได้แต่กรองประโยคได้บางส่วน เท่ากับ 0.5 คะแนน และสามารถสรุปคะแนนแต่ละกรณีเป็นกราฟได้ดังภาพที่ 7-3

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



ภาพที่ 7-3 คะแนนของลักษณะสำคัญที่ใช้ระบุหมวดหมู่ของคำบนคลังข้อความ เมื่อพิจารณาจำนวนคำที่ป้อนเข้าสู่ระบบตั้งแต่ 3-6 คำ

จากตารางที่ 7-7 และภาพที่ 7-3 พบว่าการเลือกใช้คะแนนไบแกรมจากลักษณะสำคัญการระบุหมวดหมู่ที่แตกต่างกัน มีผลต่อการเลือกประโยคคำตอบของระบบ ดังภาพที่ 7-3 จะเห็นว่าในกรณีที่ 3 ที่ใช้คะแนนจากไบแกรมที่ได้จากแบบจำลองของคำร่วมกับคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่หลักที่ฝึกสอนบนคลังข้อความ TELL-S ในการคำนวณหาประโยคคำตอบจะให้ผลดีกว่าการใช้เฉพาะคะแนนจากไบแกรมที่ได้จากแบบจำลองของคำที่ฝึกสอนบนคลังข้อความเพียงอย่างเดียว และดีกว่าใช้คะแนนจากไบแกรมที่ได้จากแบบจำลองของคำร่วมกับคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่ย่อย แต่อย่างไรก็ตามการใช้ลักษณะสำคัญการระบุหมวดหมู่หลักและหมวดหมู่ย่อยร่วมกัน กล่าวคือใช้คะแนนจากไบแกรมที่ได้จากแบบจำลองของคำร่วมกับคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่หลักและคะแนนจากไบแกรมจากแบบจำลองของหมวดหมู่ย่อย และฝึกสอนบนคลังข้อความ TELL-S ย่อมให้ผลการเลือกประโยคคำตอบที่ดีที่สุด ดังกรณีที่ 7 ในภาพที่ 7-3

7.3 การประเมินผลประสิทธิภาพของระบบอ้างอิงโดยอาศัยหลักการเอ็นแกรมร่วมกับวิธีระบุหน้าที่ของคำแบบ RIPOS

สำหรับผลการทดลองเพื่อวัดประสิทธิภาพของอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทย เมื่อใช้วิธีการระบุคำตามที่งานวิจัยนี้เสนอเปรียบเทียบกับวิธีการระบุหน้าที่ของคำตาม

ระบบอ้างอิง RIPOS บนคลังข้อความ TELL-S โดยอาศัยหลักการเอ็นแกรมในการเลือกประโยคคำตอบ และใช้คำจากกลุ่มที่ 3 หรือ WS3 ในการสร้างประโยคทดสอบ ในการทดลองนี้จะแบ่งลักษณะประโยคคำตอบที่ระบบเลือกออกเป็น 4 แบบด้วยกัน คือ

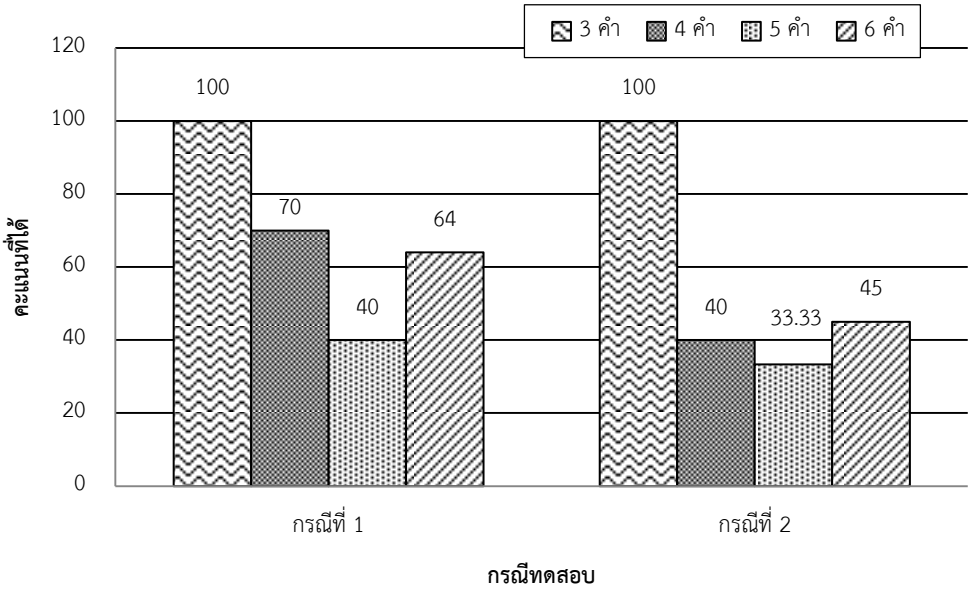
1. ประโยคคำตอบที่ระบบเลือกตรงตามประโยคเป้าหมาย
2. ระบบไม่สามารถเลือกประโยคคำตอบได้แต่กรองประโยคคำตอบได้บางส่วนและมีประโยคคำตอบอยู่ในส่วนที่กรองนั้น
3. ระบบไม่สามารถเลือกประโยคคำตอบได้ เนื่องจากคะแนนของประโยคที่เป็นไปได้เท่ากันหมด
4. ประโยคคำตอบที่ระบบเลือกไม่ตรงตามประโยคเป้าหมาย

มีผลการทดลองดังตารางที่ 7-8

ตารางที่ 7-8 ผลการเปรียบเทียบการเลือกประโยคคำตอบของระบบเมื่อใช้วิธีการระบุคำตามที่งานวิจัยนี้เสนอและวิธีการระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS

การทดลอง	**ประโยคคำตอบที่ถูกเลือกตรงตามประโยคเป้าหมาย (%)**	**ไม่สามารถเลือกประโยคคำตอบได้แต่กรองประโยคได้บางส่วน (%)**	**ไม่สามารถเลือกประโยคคำตอบได้ (%)**	**ประโยคคำตอบที่ถูกเลือกไม่ตรงตามประโยคเป้าหมาย (%)**	**คะแนนรวม**
กรณีที่ 1: การระบุหมวดหมู่และหน้าที่ของคำตามงานวิจัยที่นำเสนอ					
3 คำ	100.00	0.00	0.00	0.00	100.00
4 คำ	70.00	0.00	0.00	30.00	70.00
5 คำ	40.00	0.00	0.00	60.00	40.00
6 คำ	40.00	50.00	0.00	10.00	65.00
กรณีที่ 2: การระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS					
3 คำ	100.00	0.00	0.00	0.00	100.00
4 คำ	35.00	10.00	20.00	35.00	40.00
5 คำ	6.67	53.33	0.00	40.00	33.33
6 คำ	0.00	90.00	10.00	0.00	45.00

จากตารางที่ 7-8 กำหนดให้ กรณีที่ 1 คือการระบุหมวดหมู่และหน้าที่ของคำตามที่งานวิจัยนี้นำเสนอ และกรณีที่ 2 คือการระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS สามารถเปรียบเทียบคะแนนรวมของการเลือกประโยคและสร้างเป็นกราฟได้ดังภาพที่ 7-4



ภาพที่ 7-4 คะแนนรวมของการเลือกประโยคคำตอบโดยเปรียบเทียบระหว่างกรณีการระบุหมวดหมู่และหน้าที่ของคำตามที่งานวิจัยนี้นำเสนอ (กรณีที่ 1) และการระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS (กรณีที่ 2)

จากตารางที่ 7-8 และภาพที่ 7-4 พบว่าการระบุระบุหมวดหมู่และหน้าที่ของคำตามที่งานวิจัยนี้นำเสนอ (กรณีที่ 1) จะช่วยคัดกรองและเลือกประโยคได้ดีกว่าการระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS (กรณีที่ 2) หากพิจารณากรณีที่ป้อนคำเข้าสู่ระบบ 3 คำ จะเห็นว่าสามารถสร้างประโยคคำตอบที่ถูกต้องได้คิดเป็น 100 คะแนนเท่ากัน แต่เมื่อยิ่งเพิ่มจำนวนคำที่ป้อนเข้าสู่ระบบจะยิ่งพบความแตกต่างในการเลือกประโยคคำตอบคือ เมื่อป้อนคำเข้าสู่ระบบ 4 คำ การระบุหมวดหมู่และหน้าที่ของคำตามที่งานวิจัยนี้นำเสนอจะให้คะแนนความถูกต้องถึง 70 คะแนน ในขณะที่การระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS มีคะแนนความถูกต้องเพียงแค่ 40 คะแนน เช่นเดียวกับกรณีการป้อนคำเข้าสู่ระบบ 5 คำ และ 6 คำ ที่การระบุหมวดหมู่และหน้าที่ของคำตามที่งานวิจัยนี้นำเสนอได้คะแนนความถูกต้องในการเลือกประโยคคำตอบสูงกว่าการระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS

7.4 การประเมินผลประสิทธิภาพของแนวทางที่นำเสนอโดยอาศัยหลักการเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST)

การทดสอบประสิทธิภาพของแนวทางที่นำเสนอโดยอาศัยหลักการเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST) เมื่อระบุหมวดหมู่และหน้าที่ของคำในคลังข้อความ TELL-S ตามแนวทางที่นำเสนอ และสร้างประโยคทดสอบจากคำในกลุ่มที่ 3 หรือ WS3 มีผลการทดลองดังตารางที่ 7-9

ตารางที่ 7-9 จำนวนประโยคที่ระบบเลือกถูกต้องตามประโยคเป้าหมายโดยอาศัยเอ็นแกรมร่วมกับหลักการตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST)

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
3 คำ			
n=2, v=1	5	5	100.00
รวม	**5**	**5**	**100.00**
4 คำ			
n=1, adj=2, v=1	5	4	80.00
n=1, adj=1, v=1, adv=1	5	2	80.00
n=2, adj=1, v=1	5	2	80.00
n=2, v=1, adv=1	5	4	80.00
รวม	**20**	**16**	**80.00**
5 รูป			
n=1, adj=2, v=1, adv=1	5	2	60.00
n=2, adj=2, v=1	5	2	60.00
n=2, adj=1, v=1, adv=1	5	2	100.00
รวม	**15**	**11**	**73.33**
6 รูป			

กรณี	จำนวนประโยคทดสอบทั้งหมด	จำนวนประโยคที่ถูกต้องตามประโยคเป้าหมาย	เปอร์เซ็นต์ความถูกต้อง (%)
n=2, adj=3, v=1	5	2	40.00
n=2, adj=2, v=1, adv=1	5	3	60.00
รวม	**10**	**5**	**50.00**
รวมทั้งหมด	**50**	**37**	**74.00**

จากตารางที่ 7-9 ซึ่งแสดงผลการทดสอบระบบที่นำเสนอโดยอาศัยหลักการเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST) เมื่อทำการระบุหมวดหมู่และหน้าที่ของคำบนคลังข้อความ TELL-S และสร้างประโยคทดสอบจากคำใน WS3 พบว่าเมื่อป้อนคำ 3 คำเข้าสู่ระบบ ระบบจะสามารถเลือกประโยคคำตอบที่ถูกต้องตรงตามประโยคเป้าหมาย 100% เมื่อป้อนคำ 3 คำเข้าสู่ระบบ ระบบจะสามารถเลือกประโยคคำตอบที่ถูกต้องตรงตามประโยคเป้าหมาย 80% เมื่อป้อน 3 คำเข้าสู่ระบบ ระบบจะสามารถเลือกประโยคคำตอบที่ถูกต้องตรงตามประโยคเป้าหมาย 73.33% และเมื่อป้อน 6 คำเข้าสู่ระบบ ระบบจะสามารถเลือกประโยคคำตอบที่ถูกต้องตรงตามประโยคเป้าหมาย 50%

เมื่อเปรียบเทียบประสิทธิภาพการเรียงลำดับคำในประโยคภาษาไทยระหว่างระบบที่อาศัยหลักการทำงานเอ็นแกรม (N-gram) เพียงอย่างเดียว กับระบบที่นำเสนอที่อาศัยหลักการทำงานของเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST) จะได้ผลการทดลองดังตารางที่ 7-10

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ตารางที่ 7-10 การประเมินผลการเรียงลำดับคำในประโยคภาษาไทยโดยเปรียบเทียบระหว่างระบบที่อาศัยหลักการทำงานเอ็นแกรม (N-gram) เพียงอย่างเดียวกับระบบที่อาศัยหลักการทำงานของเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST)

กรณี	**เปอร์เซ็นต์ความถูกต้อง (%)**		**เปอร์เซ็นต์ความถูกต้องเพิ่มขึ้น (%)**
	N-gram	**WFST**	
3 คำ	100	100	0
4 คำ	70	80	10
5 คำ	40	73.33	33.33
6 คำ	40	50	10
เฉลี่ย	**58**	**74**	**16**

จากตารางที่ 7-10 พบว่าเมื่อนำหลักการทำงานของตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (WFST) มาใช้ในการคำนวณคะแนนหาประโยคคำตอบจากระบบการเรียงลำดับคำในประโยคภาษาไทย จะมีเปอร์เซ็นต์ในการเลือกประโยคคำตอบที่ถูกต้องเพิ่มขึ้น 16 เปอร์เซ็นต์ เมื่อเปรียบเทียบกับการเลือกประโยคคำตอบโดยอาศัยหลักการทำงานของเอ็นแกรมเพียงอย่างเดียว และเมื่อพิจารณาในกรณีการป้อนคำเข้าสู่ระบบพบว่า กรณีที่ป้อนคำเข้าสู่ระบบ 3 คำ จะไม่มีการเปลี่ยนแปลงเปอร์เซ็นต์ในการเลือกประโยคคำตอบที่ถูกต้อง เนื่องจากทั้ง 2 วิธีสามารถสร้างประโยคจากคำ 3 คำได้ถูกต้อง 100 เปอร์เซ็นต์อยู่แล้ว กรณีที่ป้อนคำเข้าสู่ระบบ 4 คำ, 5 คำ และ 6 คำ การใช้หลักการของ WFST มาช่วยคำนวณคะแนนของประโยคคำตอบ สามารถเพิ่มเปอร์เซ็นต์ความถูกต้องในการเลือกประโยคคำตอบได้ถึง 10 เปอร์เซ็นต์ 33.33 เปอร์เซ็นต์ และ 10 เปอร์เซ็นต์ตามลำดับ

บทที่ 8
สรุปผลการวิจัยและข้อเสนอแนะ

8.1 สรุปผลการวิจัย

วิทยานิพนธ์ฉบับนี้ได้นำเสนออัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยที่จัดเรียงอย่างไม่เป็นลำดับ และไม่ถูกต้องตามหลักไวยากรณ์ภาษาไทย ให้เรียงลำดับคำได้อย่างเป็นลำดับที่ถูกต้องตามหลักไวยากรณ์และมีความหมาย โดยอาศัยแบบจำลองความน่าจะเป็นของประเภทคำกระตุ้นการพูด กล่าวคือ ในงานวิจัยนี้ได้ใช้แบบจำลองเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก เพื่อใช้ในการเรียงลำดับคำในประโยคภาษาไทย นอกจากนี้ยังได้นำเสนอประเภทคำกระตุ้นการพูด ที่ได้อ้างอิงตามแนวคิดของ Nanci Bell ซึ่งเป็นลักษณะสำคัญในการการระบุหมวดหมู่ของคำและการระบุหน้าที่ของคำบนคลังข้อความ BEST2010 และคลังข้อความที่สร้างขึ้นมาใหม่ จากหนังสือเรียนภาษาไทยระดับชั้นประถมศึกษาปีที่ 1 และ 2 โดยมีว่า TELL-S ซึ่งเป็นคลังข้อความที่เหมาะกับบุคคลออทิสติกและบุคคลที่มีปัญหาทางด้านการเรียนรู้ที่มีความบกพร่องทางด้านการจัดเรียงและภาษา จากแนวทางการเรียงลำดับคำในประโยคภาษาไทยตามแนวทางที่นำเสนอ พบว่ามีเปอร์เซ็นต์ความถูกต้องเฉลี่ยอยู่ที่ 74 เปอร์เซ็นต์ โดยจากการทดลองทั้งหมด สามารถสรุปได้ว่า เมื่อใช้ข้อมูลจากคลังข้อความ TELL-S ร่วมกับการเลือกใช้ลักษณะสำคัญในการระบุหมวดหมู่ของคำตามแนวคิดของ Nanci Bell และการระบุหน้าที่ของคำตามแนวคิดที่นำเสนอ โดยอาศัยหลักการของแบบจำลองเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักในการคำนวณคะแนนและเลือกประโยคคำตอบ จะสามารถเลือกประโยคคำตอบซึ่งเป็นประโยคความเดียวที่ตรงกับประโยคเป้าหมายที่ถูกเลือกโดยผู้เชี่ยวชาญได้ดีที่สุด เมื่อเปรียบเทียบกับการเลือกใช้ลักษณะสำคัญและวิธีการระบุหน้าที่ของคำแบบอื่น ๆ ในงานวิจัย

8.2 อภิปรายผลการวิจัย

จากผลการทดลองทั้ง 5 การทดลองในบทที่ 7 สามารถอภิปรายผลได้ดังนี้

ในการทดลองที่ 1 และการทดลองที่ 2 นั้นทำการทดลองเพื่อพิจารณาเปรียบเทียบว่าคลังข้อความใดเหมาะสมจะใช้เป็นคลังข้อความฝึกฝน เพื่อนำคะแนนจากการฝึกฝนมาใช้ในอัลกอริทึมการสร้างประโยคในระบบการเรียงลำดับคำในประโยคภาษาไทย และพิจารณาว่ากลุ่มของคำกลุ่มใดเหมาะสมที่จะนำมาใช้เป็นชุดคำในประโยคทดสอบ โดยผลการทดลองในบทที่ 7 แสดงให้เห็นว่าเมื่อใช้คลังข้อความจากแหล่งเดียวกันทั้งในการฝึกสอนการระบุหมวดหมู่และหน้าที่ของคำที่นำเสนอและสร้างประโยคทดสอบ จะส่งผลให้เปอร์เซ็นต์ความถูกต้องในการเลือกประโยคคำตอบสูงกว่าการใช้

คลังข้อมูลฝึกสอนและสร้างประโยคทดสอบที่แตกต่างกัน แต่เพื่อความเท่าเทียมกันในการวัดผล จึงควรใช้คำจากกลุ่มที่ 3 (WS3) ซึ่งเป็นคำที่ปรากฏอยู่ในคลังข้อความ BEST2010 และคลังข้อความ TELL-S มาเป็นชุดคำที่ใช้สร้างประโยคทดสอบ และเมื่อสร้างประโยคทดสอบจาก WS3 แล้ว เปรียบเทียบคลังข้อความที่ควรใช้เป็นคลังข้อความฝึกฝนพบว่า ในงานวิจัยนี้คลังข้อความ TELL-S เหมาะสมจะเป็นคลังข้อความฝึกฝนมากกว่าคลังข้อความ BEST2010

ในการทดลอง 3 ทำการทดลองเพื่อวัดประสิทธิภาพของอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอ และเพื่อหาลักษณะสำคัญของการระบุหมวดหมู่ของคำในคลังข้อความ BEST2010 และ TELL-S โดยอาศัยหลักการเอ็นแกรม พบว่าการเลือกใช้ลักษณะสำคัญในการระบุหมวดหมู่ของคำตามแนวคิดของ Bell และการระบุหน้าที่ของคำตามแนวคิดที่นำเสนอร่วมกัน มีผลทำให้อัลกอริทึมในการเลือกประโยคคำตอบมีประสิทธิภาพมากที่สุด

ในการทดลองที่ 4 ทำการทดลองเพื่อเปรียบเทียบการระบุหมวดหมู่และหน้าที่ของคำตามแนวคิดที่นำเสนอซึ่งเป็นกรณีที่วัดผลการทดลองแล้วมีประสิทธิภาพสูงสุดในทดลองที่ 3 กับวิธีการระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS พบว่า วิธีการระบุหมวดหมู่และหน้าที่ของคำตามแนวคิดที่นำเสนอ มีประสิทธิภาพในการเลือกประโยคคำตอบได้ถูกต้องตามประโยคเป้าหมายสูงกว่าการระบุหน้าที่ของคำตามระบบอ้างอิง RIPOS

และในการทดลองที่ 5 ทำการทดลองเพื่อวัดประสิทธิภาพของอัลกอริทึมการเรียงลำดับคำในประโยคภาษาไทยที่นำเสนอ โดยอาศัยหลักการของแบบจำลองเอ็นแกรมร่วมกับตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักในการคำนวณคะแนนและเลือกประโยคคำตอบ พบว่าสามารถเลือกประโยคคำตอบซึ่งเป็นประโยคความเดียวที่ตรงกับประโยคเป้าหมายถูกต้องมากที่สุด เมื่อเปรียบเทียบกับการเลือกใช้ลักษณะสำคัญและวิธีการระบุหน้าที่ของคำแบบอื่น ๆ ในงานวิจัย

เมื่อวิเคราะห์ความผิดพลาดของแนวทางที่นำเสนอ พบว่ามีเปอร์เซ็นต์ความผิดพลาดเฉลี่ยอยู่ที่ 26 เปอร์เซ็นต์ และจะเห็นว่ายิ่งป้อนคำเข้าสู่ระบบมากขึ้นความผิดพลาดในการสร้างและเลือกประโยคคำตอบก็จะยิ่งเพิ่มขึ้นด้วย สาเหตุหลักที่ทำให้เกิดข้อผิดพลาดเช่นนี้เนื่องจากในคลังข้อความมีประโยคความเดียวภาษาไทยที่มีความยาวตั้งแต่ 5 คำขึ้นไปมีจำนวนน้อย คิดเป็น 4.98 เปอร์เซ็นต์ของประโยคทั้งหมด และประโยคความเดียวที่มีการระบุหน้าที่ของคำและระบุหมวดหมู่ของคำที่มีอยู่ในคลังข้อความก็มีจำนวนน้อยเช่นกัน อีกทั้งในการทดสอบการเลือกประโยคคำตอบจากระบบที่นำเสนอนี้ มีการสร้างชุดทดสอบที่มีความคลุมเครือมากในส่วนของส่วนขยาย กล่าวคือในความเป็นจริง ประโยคที่พูดกันในชีวิตประจำวันไม่ได้เป็นประโยคที่มีคำขยายมากมายนัก และในคลังข้อความที่

ใช้ฝึกสอนก็ไม่มีประโยคที่มีคำขยายมากเกินความจำเป็นอยู่ในคลังข้อความเช่นกัน ด้วยเหตุนี้จึงทำให้ประโยคคำตอบที่สร้างจากระบบที่ทำการป้อนคำ 5-6 คำ มีเปอร์เซ็นต์ความผิดพลาดสูง

นอกจากนี้เมื่อพิจารณาประโยคภาษาไทยบางประโยคที่สามารถสลับตำแหน่งของคำนามสองคำที่ทำหน้าที่เป็นประธานกับกรรมในประโยคแล้วไม่ได้ทำให้ประโยคนั้นเป็นประโยคที่ผิดหรือไม่ได้ความหมาย เช่น “ฉันรักพ่อ” สามารถสลับตำแหน่งของคำนามในประโยคเป็น “พ่อรักฉัน” ได้เช่นกัน ประโยคในลักษณะนี้ เมื่อป้อนคำ 3 คำ คือ {ฉัน, พ่อ, รัก} เข้าสู่ระบบ ระบบจะเลือกประโยคคำตอบออกมาเพียงประโยคเดียว ซึ่งประโยคคำตอบนั้นอาจไม่ตรงกับประโยคที่คนเลือก ซึ่งส่งผลให้เกิดเป็นคะแนนความผิดพลาดของประโยคเช่นกัน

8.3 ข้อเสนอแนะ

เนื่องจากปัจจุบันคลังข้อมูลที่เก็บข้อความที่แบ่งเป็นประโยคภาษาไทยมีน้อยและส่วนใหญ่เป็นข้อความสั้น ๆ รวมถึงคลังข้อมูลส่วนมากที่เก็บข้อความประโยคภาษาไทยยังไม่มีการแบ่งประโยคที่ชัดเจนและยังไม่มีการระบุหน้าที่ของคำในประโยค ดังนั้นการเพิ่มจำนวนข้อมูลในคลังข้อความให้มีความหลากหลาย เพื่อใช้เป็นข้อมูลฝึกสอนอาจช่วยสร้างแบบจำลองที่มีความครอบคลุมได้มากขึ้น ตัวอย่างเช่นในกรณีความผิดพลาดที่เกิดจากจำนวนคำที่ป้อนเข้าสู่ระบบ ยิ่งป้อนเข้าไปมากก็จะยิ่งมีความผิดพลาดมาก ซึ่งมีสาเหตุมาจากการที่ข้อมูลในคลังข้อความที่ใช้เป็นข้อมูลฝึกสอน มีประโยคความเดียวที่มีความยาว 5 คำขึ้นไปน้อยมาก จนทำให้การฝึกสอนแบบจำลองต่าง ๆ ไม่ดีเท่าที่ควรและอาจช่วยแก้ปัญหาจากกรณีของการสลับตำแหน่งของส่วนขยายได้อีกด้วย

นอกจากนี้อัลกอริทึมและระบบที่นำเสนอนี้ ยังสามารถนำไปประยุกต์ใช้ร่วมกับซอฟต์แวร์ AAC ต่อไปในอนาคต เพื่อเป็นประโยชน์กับบุคคลออทิสติกและบุคคลที่มีความบกพร่องทางการเรียนรู้กลุ่มที่มีความบกพร่องด้านการจัดเรียงและภาษาให้สามารถใช้ซอฟต์แวร์นี้ได้ ซึ่งเป็นอีกหนึ่งทางเลือกที่จะเป็นตัวช่วยในการสื่อสาร และสามารถนำไปใช้เป็นเครื่องมือฝึกฝนความสามารถในการคิดและการจัดการต่อไปได้

รายการอ้างอิง

1. วิกิพีเดีย สารานุกรมเสรี, *ภาษาไทย*. [cited 2014 27 November], http://th.wikipedia.org/wiki/%E0%B8%A0%E0%B8%B2%E0%B8%A9%E0%B8%B2%E0%B9%84%E0%B8%97%E0%B8%A2.
2. ความรู้ภาษาไทย, *ความรู้ทั่วไปเกี่ยวกับภาษาไทย*. 2012 [cited 2014 5 December], http://lru-5214.blogspot.com/.
3. ภาษาไทยเพื่อการพัฒนาการสื่อสาร, *การร้อยเรียงประโยค*. [cited 2014 18 November], http://thai-m5-gurubird.blogspot.com/2010/12/blog-post_4308.html.
4. จอมสุรางค์ โพธิสัตย์, แพทย์หญิง, *ออทิสติก: กลุ่มโรคออทิสติก (Autistic spectrum disorder)*. [cited 2014 18 November], http://haamor.com/th/%E0%B8%AD%E0%B8%97%E0%B8%B4%E0%B8%AA%E0%B8%95%E0%B8%B4%E0%B8%81/.
5. สถาบันราชานุกูล, *เด็กแอลดีคู่มือสำหรับครู*. [cited 2014 18 November], http://apps.qlf.or.th/member/UploadedFiles/prefix-07082557-052708-Lk811G.pdf.
6. วาสนี วิพัทนะพร, *การสื่อสารกับเด็กออทิสติกโดยการใช้รูปภาพ (VISUAL STRATEGIES)*. [cited 2014 18 November], http://www.ser01.com/page/wichakan_autistic11.htm.
7. สถาบันราชานุกูล, *เด็กออทิสติกคู่มือสำหรับครู*. [cited 2014 18 November], http://apps.qlf.or.th/member/UploadedFiles/prefix-07082557-052705-4Py1Bw.pdf.
8. Bell, N., *Imagery and the Language Processing Spectrum*, in *ICDL Clinical Practice Guidelines*. p. 615-638.
9. เบญจา ชลธาร์นนท์, ผศ.ดร., *เทคโนโลยีสิ่งอำนวยความสะดวกสำหรับคนพิการ*. [cited 2013 30 May], http://dusithost.dusit.ac.th/~education/book/nawasarn1/01.pdf.
10. Chula AAC, *Chula AAC*. [cited 2013 30 May], http://delivery.acm.org/10.1145/2510000/2500759/a5-suchato.pdf?ip=161.200.188.23&id=2500759&acc=NO%20RULES&key=E2F11C531D2FCAD3%2E290424E65E6BD228%2E4D4702B0C3E38B35%2E4D4702B0C3E3

8B35&CFID=460976643&CFTOKEN=71905816&__acm__=1418107025_e932e9c7a63c7a9aa14f52a7b424d5e3.

11. สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน, *กระทรวงศึกษาธิการชุดภาษาเพื่อชีวิต ภาษาพาที รายวิชาพื้นฐาน กลุ่มสาระการเรียนรู้ ภาษาไทย ชั้น ประถมศึกษาปีที่ 1*. 2551.
12. สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน, *กระทรวงศึกษาธิการชุดภาษาเพื่อชีวิต ภาษาพาที รายวิชาพื้นฐาน กลุ่มสาระการเรียนรู้ ภาษาไทย ชั้น ประถมศึกษาปีที่ 2*. 2551.
13. Openfst Library, *OpenFst Quick Tour*. 2014 [cited 2014 9 August], http://www.openfst.org/twiki/bin/view/FST/WebHome.
14. บุญเสริม กิจศิริกุล, *ปัญญาประดิษฐ์*, ed. 1.0.2. 2548, http://www.cp.eng.chula.ac.th/~boonserm/.
15. สมศักดิ์ ทองช่วย, *ประโยค*, in เอกสารประกอบการจัดการเรียนการสอน รายวิชา ท๔๐๑๐๕ หลักภาษาไทยในชีวิตประจำวัน. โรงเรียนมหิดลวิทยานุสรณ์.
16. เอกรินทร์ สี่มหาศาล และคณะ, *ภาษาไทย ป.๖.* 2552, อักษรเจริญทัศน์. 270.
17. สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน. ประโยค. [cited 2013 30 May].
18. เกียรติชัย เดชพิทักษ์ศิริกุล, *การศึกษาเปรียบเทียบโครงสร้างวลี ประโยค และสัมพันธสารของเด็กปกติและเด็กออทิสติก*, in คณะอักษรศาสตร์ สาขาวิชาภาษาไทย2551, มหาวิทยาลัยศิลปากร: มหาวิทยาลัยศิลปากร.
19. SAAM Speech Audio & Miltimedia. *หลักการทำงานของ N-gram model*. [cited 2013 30 May], http://tvis.nectec.or.th/speech/index.php?option=com_content&view=article&id=66&Itemid=92&lang=th.
20. Wikipedia. *Katz's back-off model*. [cited 2014 27 October 2014], http://en.wikipedia.org/wiki/Katz's_back-off_model.
21. McDonough, J. *Search and Weighted Finite-State Transducers*, 2009.
22. OpenFst, *An Open-Source, Weighted Finite-State Transducer Library and its Applications to Speech and Language*, in *Part I. Theory and Algorithms*.
23. Baker, B., *Minspeak, a Semantic Compaction System that Makes Self-Expression Easier for Communicatively Disabled Individuals*, 1982, Byte. p. 186-202.

24. Copestake, A. *Augmented and alternative NLP techniques for augmentative and alternative communication*. in *ACL workshop on Natural Language Processing for Communication Aids*. 1997.

25. Reiter, R., et al. *Using NLG to help languageimpaired users tell stories and participate in social dialogues*. in *European Workshop on Natural Language Generation (ENLG*-09). 2009.

26. Suchato, A., et al. *Multilingual AAC on Android*. in *The 5th International Conference on Rehabilitation Engineering and Assistive Technology (i-CREATe* 2011). 2011. Bangkok, Thailand.

27. Slayden, G., M.-Y. Hwang, and L. Schwartz, *Thai Sentence-Breaking for Large-Scale SMT*, in *Tthe 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), The 23rd International Conference on Computational Linguistics (COLING)*2010: Beijing. p. 8–16.

28. Glenn Slayden, M.-Y.H., Lee Schwartz, *Large-Scale Thai Statistical Machine Translation*, MSR-TR-2010-41, Editor 2010: Microsoft Research.

29. Ratnaparkhi, A. *Trainable methods for surface natural language generation*. in *The first conference on North American chapter of the Association for Computational Linguistics*. 2000. Seattle, Washington.

30. Wiegand, K. and P. R. *Non-syntactic word prediction for AAC*. in *The Third Workshop on Speech and Language Processing for Assistive Technologies*. 2012. Montreal, Canada.

31. Liu, H., et al. *Use of statistical N-gram models in natural language generation for machine translation*. in *ICASSP*. 2003. IEEE.

32. Wei, P. and Bom, X. *Chinese-English transliteration using weighted finite-state transducers*. in International Confrence on Audio, Language and Image Processing (ICALIP2008). 2008.

33. Povey, D., et al. *Generating Exact Lattices in the WFST Framework*. in 37th IEEE International Confrence on Acoustics, Speech, and Signal Processing (ICASSP2012). 2012.

34. Sornlertlamvanich, V. , Charoenporn, T. and Isahara, H. *ORCHID: Thai Part-of Speech Tagged Corpus*. National Electronics and Computer Technology Center Technical Report, 1997: p. 5-19.

35. Wutiwatchai, C., et al. *Phonetically Distributed Continuous Speech Corpus for Thai Language.* in Third International Conference on Language Resources and Evaluation (LREC2002). 2012.

36. Vlasta, R. and Vopalka, P. *Methods of Sentences Selection for Read-Speech Corpus Design*. in TSD'99, LNAI 1692, M.V.e. al., Editor 1999: Berlin Heidelberg, Springer-Verlag. p. 165-170.

37. Huber, F. *On the Generation of English Sentence*. IEEE Transactions on Computers, 1976. 25: p. 90-91.

38. Thammasudjarit, R. and C. Pluempitiwiriyawej. *Design and Construction of the Royal-Institute-Based Thai-Part-of-Speech Corpus* in *The Tenth Symposium on Natural Language Processing (SNLP-*2013). 2013. Phuket, Thailand.

39. Hanks, P. and E. Ježek. *Shimmering lexical sets*. in *the XIII EURALEX International Congress*. 2008. Barcelona.

40. Hanks, P. *The organization of the Lexicon: Semantic Types and Lexical Sets*. in *Euralex Proceedings* 2006. 2006.

41. เพิมพันธ์ เดชะคุปต์, *พฤติกรรมการสอนวิทยาศาสตร์*. 2545: กรุงเทพฯ: บริษัท พัฒนาคุณภาพวิชาการ (พว.) จำกัด.

42. GotoKnow. *การใช้คำถามกระตุ้นความคิด*. [cited 2013 10 August], http://www.gotoknow.org/posts/306302.

43. National Electronics and Computer Technology Center. Enhancing the Standard of Thai language processing 2010 (BEST2010).

44. BEST. *Benchmark for Enhancing the Standard of Thai Language Processing*. [cited 2013 30 May], http://thailang.nectec.or.th/best/.

45. Limpanadusadee, W., Punyabukkana, P. and Suchato, A., *Text Corpus for Natural Language Story-telling Sentence Generation: A Design and Evaluation*, in 11*th International Joint Conference on Computer Science and Software Engineering (JCSSE* 2014)2014: Pattaya, Thailand. p. 80-85.

46. Wikipedia. *Backus-Naur Form*. [cited 2013 30 May], http://en.wikipedia.org/wiki/Backus%E2%80%93Naur_Form.

47. Clarkson, P. *Statistical Language Modeling Toolkit*. 1999 7th June 1999 [cited 2013 30 May], http://www.speech.cs.cmu.edu/SLM/toolkit.html.

48. Wikipedia. *Language model*. [cited 2013 30 May], http://en.wikipedia.org/wiki/Language_model.

49. OpenGrm Libraries. *OpenGrm NGram Library*. 2014 [cited 2014 1 August], http://www.opengrm.org/.

50. Wikipedia. *Hash table*. [cited 2014 1 August], http://en.wikipedia.org/wiki/Hash_table.

ประวัติผู้เขียนวิทยานิพนธ์

นางสาววรษา ลิมปนดุษฎี เกิดเมื่อวันที่ 9 สิงหาคม พ.ศ.2532 ที่จังหวัดสงขลา สำเร็จการศึกษาระดับปริญญาตรี หลักสูตรวิทยาศาสตรศาสตรบัณฑิต สาขาการพัฒนาซอฟต์แวร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปี พ.ศ.2555 และเข้าศึกษาต่อระดับปริญญาโทในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย และได้รับทุนสนับสนุนการวิจัยจากโครงการทุนสถาบันบัณฑิตวิทยาศาสตร์และเทคโนโลยีไทย (TGIST) ผลงานทางวิชาการที่ตีพิมพ์เผยแพร่ ได้แก่ หนังสือสรุปประชุมวิชาการระดับนานาชาติ ในหัวข้อเรื่อง “An Automatic Natural Language Sentence Generation from Images”, หนังสือสรุปประชุมวิชาการระดับนานาชาติ ในหัวข้อเรื่อง “Text Corpus for Natural Language Story-telling Sentence Generation: A Design and Evaluation” และหนังสือสรุปการประชุมเชิงปฏิบัติการในหัวข้อเรื่อง “An Automatic Natural Language Sentence Generation Using Weighted Finite-State Transducer”

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University