การขจัดความกำกวมคำลักษณนามของภาษาไทยในกระบวนการจับคู่คำสองภาษาสำหรับการแปลภาษาด้วยเครื่องเชิงสถิติของคู่ภาษา ไทย-อังกฤษ

โดย

นายภูวเมศฐ์ พิมลปัญญารัตน์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
ภาควิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2558

การขจัดความกำกวมคำลักษณนามของภาษาไทยในกระบวนการจับคู่คำสองภาษาสำหรับการแปลภาษาด้วยเครื่องเชิงสถิติของคู่ภาษา ไทย-อังกฤษ

โดย

นายภูวเมศฐ์ พิมลปัญญารัตน์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
ภาควิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2558

THAI CLASSIFIER DISAMBIGUATION IN BILINGUAL ALIGNMENT PROCESS FOR THAI-ENGLISH SMT

BY

MR. PUWAMED PIMONPANYARAD

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE)
DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF SCIENCE AND TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2015

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

วิทยานิพนธ์

ของ

นายภูวเมศฐ์ พิมลปัญญารัตน์

เรื่อง

การขจัดความกำกวมคำลักษณนามของภาษาไทยในกระบวนการ
จับคู่คำสองภาษาสำหรับการแปลภาษาด้วยเครื่องเชิงสถิติ
ของคู่ภาษา ไทย-อังกฤษ

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

เมื่อ วันที่ 14 กรกฎาคม พ.ศ. 2559

ประธานกรรมการสอบวิทยานิพนธ์ _____
(อาจารย์ ดร. ปกป้อง ส่องเมือง)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร. รัชฎา คงคะจันทร์)

กรรมการสอบวิทยานิพนธ์ _____
(อาจารย์ ดร. วสิศ ลิ้มประเสริฐ)

กรรมการสอบวิทยานิพนธ์ _____
(ดร. เทพชัย ทรัพย์นิธิ)

คณบดี _____
(รองศาสตราจารย์ ปกรณ์ เสริมสุข)

หัวข้อวิทยานิพนธ์	การขจัดความกำกวมคำลักษณนามของภาษาไทยในกระบวนการจับคู่คำสองภาษาสำหรับการแปลภาษาด้วยเครื่องเชิงสถิติ ของคู่ภาษา ไทย-อังกฤษ
ชื่อผู้เขียน	นายภูวเมศฐ์ พิมลปัญญารัตน์
ชื่อปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา/คณะ/มหาวิทยาลัย	วิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผู้ช่วยศาสตราจารย์ ดร. รัชฎา คงคะจันทร์
ปีการศึกษา	2558

บทคัดย่อ

งานวิจัยนี้ได้นำเสนอวิธีดำเนินการจัดการกับคำลักษณนามของภาษาไทยในการแปลภาษาด้วยเครื่องเชิงสถิติ ซึ่งคำลักษณนามของภาษาไทยนั้นทำให้เกิดความสับสนกับคำนามในกระบวนการจับคู่คำตั้งแต่ความเหมือนกันของคำ และทำให้เกิดความสับสนในกระบวนการเรียนรู้ โดยวิธีการที่นำเสนอคือการจับรูปแบบของการใช้คำลักษณนาม และไม่ทำการจับคู่คำลักษณนามในตอนแรก ซึ่งเมื่อทำการจับคู่คำโดยไม่มีคำลักษณนามเสร็จสิ้นแล้ว คำลักษณนามนั้นจะถูกบังคับให้จับคู่เข้ากับคำนามหลักที่มีความสัมพันธ์กัน จากการทดสอบพบว่าวิธีการที่นำเสนอสามารถพัฒนาปรับปรุงการจับคู่คำคู่ภาษาไทย-อังกฤษ โดยเฉพาะประโยคที่มีตัวเลขซึ่งมีคำลักษณนามติดอยู่ด้วย จากผลการทดลองพบว่าวิธีการที่นำเสนอสามารถจัดการกับคำลักษณนามที่เป็นรูปแบบหน่วยซึ่งเป็นปัญหาหลักในการจับคู่คำในคู่ภาษาไทย-อังกฤษ

คำสำคัญ: ลักษณนามของภาษาไทย, การตรวจหาด้วยรูปแบบ, การจับคู่คำสองภาษา, การแปลภาษาด้วยเครื่องโดยใช้สถิติ

Thesis Title	THAI CLASSIFIER DISAMBIGUATION IN BILINGUAL ALIGNMENT FOR THAI-ENGLISH SMT
Author	Mr. Puwamed Pimonpanyarad
Degree	Master of Science
Major Field/Faculty/University	Computer Science Faculty of Science and Technology Thammasat University
Thesis Advisor	Assistant Professor Dr. Rachada Kongkachandra
Academic Years	2015

ABSTRACT

This paper presents a method to handle Thai classifier in statistical machine translation implementation. Classifier in Thai normally confuses with noun in alignment process since their surface is identical and causes further ambiguity in learning process. The proposed method is to map a pattern of classifier usage and ignores it in alignment. Once the alignment is complete with left classifier, the classifier is forced to align with its relative core noun. From testing scenario, we found that this method can improve alignment in Thai-English especially on sentences with numeric expression since they naturally contain classifier. From the result, the proposed method can impressively handle unit classifier which is a main issue in Thai-English alignment.

Keywords: Thai classifier, Detection by pattern, Bilingual alignment, SMT

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปด้วยดี เนื่องจากได้รับความกรุณาจาก ผู้ช่วยศาสตราจารย์ ดร. รัชฎา คงคะจันทร์ ซึ่งเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่เป็นผู้ให้คำปรึกษาแนะนำความรู้ แนวคิด ตลอดจนแนวทางในการแก้ไขปัญหาที่เกี่ยวข้องกับวิทยานิพนธ์ และช่วยเหลือสนับสนุนผู้วิจัย ในเรื่องของการนำเสนอผลงานวิจัยในงานประชุมวิชาการ ผู้วิจัยมีความซาบซึ้งในความกรุณาเสมอมา ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้ด้วย

ขอขอบคุณ อาจารย์ ดร. ปกป้อง ส่องเมือง ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร. วสิศ ลิ้มประเสริฐ และ ดร. เทพชัย ทรัพย์นิธิ กรรมการสอบวิทยานิพนธ์ ที่ให้ข้อเสนอแนะในการปรับปรุงให้วิทยานิพนธ์มีความสมบูรณ์มากยิ่งขึ้น

ขอขอบคุณ คุณ กัญยาลักษณ์ โพธิ์ดง, คุณ ธเนศ เรืองรจิตปกรณ์ และ คุณ วิภาส สุตันตยาวลี ที่ชี้แนะให้คำปรึกษาและแนะแนวทางด้านโปรแกรม และการเขียนบทความทางวิชาการ

สุดท้ายนี้ผู้วิจัยขอขอบคุณครอบครัวที่รัก บิดา มารดา ที่คอยดูแลผู้วิจัยเป็นอย่างดี รวมถึงคนรักที่คอยดูแล ห่วงใย และคอยให้กำลังใจผู้วิจัยมาโดยตลอด

นายภูวเมศฐ์ พิมลปัญญารัตน์

สารบัญ

หน้า

บทคัดย่อภาษาไทย (1)

บทคัดย่อภาษาอังกฤษ (2)

กิตติกรรมประกาศ (3)

สารบัญตาราง (6)

สารบัญภาพ (8)

บทที่ 1 บทนำ 1

1.1 ความเป็นมาและความสำคัญของงานวิจัย 1
1.2 วัตถุประสงค์ของงานวิจัย 2
1.3 ขอบเขตของงานวิจัย 3
1.4 ข้อจำกัดของงานวิจัย 3
1.5 ประโยชน์ที่คาดว่าจะได้รับ 3

บทที่ 2 วรรณกรรมและงานวิจัยที่เกี่ยวข้อง 4

2.1 ทฤษฎีที่เกี่ยวข้อง 4
2.1.1 คำลักษณนาม 4
2.1.2 ระบบแปลภาษาเชิงสถิติ (SMT: Statistical Machine Translation) 6
2.1.3 คำลักษณนามในภาษาไทย (Thai Classifier) 9
2.1.4 การแปลภาษาระดับวลี (Phrase-base Translation) 10
2.1.5 GIZA Software 10
2.2 งานวิจัยที่เกี่ยวข้อง 14

บทที่ 3 วิธีการวิจัย 16

3.1 ภาพรวมของระบบ 16
3.2 ขั้นตอนการทำงานของระบบ 17
3.2.1 เตรียมข้อมูลคลังคู่ภาษา (Parallel Corpus) 17
3.2.2 กระบวนการค้นหาลักษณนามในภาษาไทย 17
3.2.3 กระบวนการจัดการความกำกวมของลักษณนามในภาษาไทย 19
3.2.4 จับคู่คำสองภาษาด้วยโปรแกรม GIZA 23
3.2.5 กระบวนการคืนคำลักษณนามในการจับคู่คำ 23
3.3 การวัดผลการทดลอง 25

บทที่ 4 ผลการวิจัยและการอภิปรายผล 27

4.1 ผลลัพธ์ความถูกต้องในการจับคู่คำ 27
4.2 ผลลัพธ์ความถูกต้องในการแปลภาษาที่สืบเนื่องมาจากการจับคู่คำ 28

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ 44

5.1 สรุปผลการดำเนินงานวิจัย 44
5.2 สรุปผลการทดลอง 45
5.3 ข้อเสนอแนะและแนวทางวิจัยต่อไปในอนาคต 45

รายการอ้างอิง 47

ประวัติผู้เขียน 49

สารบัญตาราง

ตารางที่ หน้า

2.1 แสดงเทคนิคที่ใช้ และผลลัพธ์ ในงานวิจัย 14
3.1 ตัวอย่างข้อมูลคลังคู่ภาษาไทย-อังกฤษ 17
3.2 ตัวอย่างประโยคก่อน-หลัง ตัดคำ และ ติดป้ายชนิดของคำ 17
3.3 แสดงประเภทของลักษณนาม และ ตัวอย่างคำลักษณนาม 18
3.4 แสดงรูปแบบของประโยคลักษณนาม 18
3.5 ตัวอย่างผลลัพธ์ของกรณีที่ 1 20
3.6 ตัวอย่างผลลัพธ์ของกรณีที่ 2 (ภาษาไทย) 21
3.7 ตัวอย่างผลลัพธ์ของกรณีที่ 2 (ภาษาอังกฤษ) 21
3.8 ตัวอย่างผลลัพธ์การจับคู่คำที่ได้จากโปรแกรม GIZA 23
3.9 ตัวอย่างก่อน-หลัง กระบวนการคืนลักษณนามในการจับคู่คำของกรณีที่ 1 24
3.10 ตัวอย่างก่อน-หลัง กระบวนการคืนลักษณนามในการจับคู่คำของกรณีที่ 2 24
3.11 แสดง Spec ของ Server ที่ใช้ Run ผลการทดลอง 25
4.1 ผลลัพธ์การจับคู่คำของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม 26
4.2 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 1 29
4.3 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 2 30
4.4 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 3 31
4.5 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 4 32
4.6 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 5 33
4.7 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 6 34
4.8 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 7 35
4.9 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 8 36
4.10 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 9 37

4.11 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 10 38
4.12 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 11 39
4.13 แสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ของประโยคที่มีลักษณนามรูปแบบที่ 12 40
4.14 แสดงค่า BLEU Score ประโยคลักษณนามทั้งหมดของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม 41
4.15 แสดงค่า BLEU Score ประโยคทั้งหมดของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม 42

สารบัญภาพ

ภาพที่		หน้า
1.1	แสดงผลการแปลประโยคที่มีคำลักษณนามจาก Google Translation ประโยคที่ 1	1
1.2	แสดงผลการแปลประโยคที่มีคำลักษณนามจาก Google Translation ประโยคที่ 2	2
2.1	ตัวอย่างประโยคที่ใช้คำลักษณนามร่วมกับคำนามที่มาพร้อมกับตัวเลข	4
2.2	ตัวอย่างประโยคที่ไม่ใช้คำลักษณนาม	4
2.3	ตัวอย่างประโยคที่ใช้คำลักษณนามร่วมกับคำคุณศัพท์	4
2.4	ตัวอย่างประโยคที่ใช้คำลักษณนามร่วมกับคำบ่งชี้	5
2.5	ตัวอย่างประโยคที่ใช้คำลักษณนามร่วมกับคำนามที่ตามด้วยตัวเลข และคำคุณศัพท์	5
2.6	ตัวอย่างประโยคที่ใช้คำลักษณนามร่วมกับคำนามที่ตามด้วยตัวเลข และคำบ่งชี้	5
2.7	ตัวอย่างประโยคที่ใช้คำลักษณนามร่วมกับคำนามที่ตามด้วยตัวเลขคำคุณศัพท์ และคำบ่งชี้	6
2.8	แสดงการจับคู่คำของภาษาต้นทาง (สเปน) และภาษาปลายทาง (อังกฤษ)	7
2.9	แสดงการจับคู่คำแบบ bidirectionally aligned corpora [f->e, e->f] แล้วนำมา intersection of alignment point	8
2.10	แสดงการทำ grow additionally of alignment point	8
2.11	แสดงการแปลภาษาระดับวลี	10
2.12	แสดงตัวอย่างวลีคู่ภาษา (เยอรมัน-อังกฤษ) ตั้งแต่ 2-7 คำ สำหรับการจับคู่คำ	11
2.13	แสดงตัวอย่างแบบการฝึกในการจับคู่คำ	12
2.14	แสดงตัวอย่างการจับคู่คำ	13
3.1	ภาพรวมการทำงานของระบบ	16
3.2	แสดงผังงานขั้นตอนการจัดการความกำกวมของลักษณนาม	22
4.1	กราฟผลลัพธ์การจับคู่คำของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม	28
4.2	กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 1	29
4.3	กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 2	30
4.4	กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 3	31
4.5	กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 4	32
4.6	กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 5	33

4.7 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 6 34
4.8 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 7 35
4.9 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 8 36
4.10 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 9 37
4.11 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 10 38
4.12 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 11 39
4.13 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม ในแต่ละประโยคของรูปแบบที่ 12 40
4.14 กราฟแสดงค่า BLEU Score ประโยคลักษณนามทั้งหมดของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม 41
4.15 กราฟแสดงค่า BLEU Score ประโยคทั้งหมดของระบบที่นำเสนอ เปรียบเทียบกับระบบเดิม 43

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของงานวิจัย

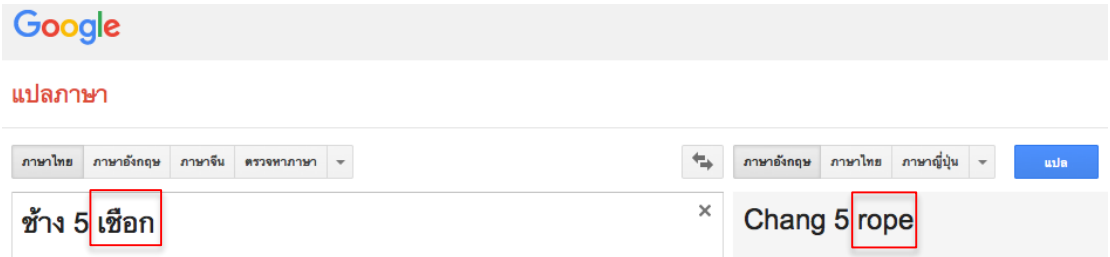
ระบบการแปลภาษาด้วยเครื่องเชิงสถิติ (SMT: Statistical Machine Translation) คือ ระบบแปลภาษาด้วยเครื่องคอมพิวเตอร์ ที่อาศัยข้อมูลจากคลังคู่ภาษา (Parallel Corpus) เพื่อสร้างคำแปลในภาษาปลายทางที่เหมาะสมที่สุดตามข้อมูลที่มีอยู่ในระบบ และการแปลภาษาเชิงสถิตินั้นจะมีความถูกต้องแม่นยำก็ขึ้นอยู่กับข้อมูลที่มีอยู่ในระบบเช่นกัน

การแปลภาษาด้วยเครื่อง จากภาษาไทยเป็นภาษาอังกฤษในปัจจุบัน ยังคงมีความไม่สมบูรณ์ เนื่องมาจากความแตกต่างกันของภาษา ซึ่งมีอยู่หลายสาเหตุ โดยหนึ่งในสาเหตุที่ทำให้การแปลภาษาด้วยเครื่องในคู่ภาษาไทยกับภาษาอังกฤษนั้นมีความไม่สมบูรณ์คือ ลักษณนาม (Classifier)

ลักษณนาม (Classifier) คือคำนามที่ใช้บอกหน่วยนับโดยเป็นคำบ่งบอกลักษณะของคำนามหลักข้างหน้าหรือเพื่อใช้แสดงลักษณะ และชนิดของสิ่งต่างๆให้ชัดเจน ซึ่งในภาษาไทยนั้นมีการใช้ลักษณนามเป็นภาษาธรรมชาติ (Natural Language) โดยลักษณนามมักเขียนอยู่หลังจำนวนนับ เช่น นก 3 *ตัว*, ส้ม 2 *ลูก* เป็นต้น ทำให้เราเข้าใจว่าคำนามนั้นๆต้องการสื่อความหมายอะไร โดยในภาษาไทยหากไม่ใช้คำลักษณนามแล้วก็จะทำให้ภาษาไทยนั้นแปลกไป และทำให้ภาษาไม่มีความเป็นธรรมชาติ เช่น นก 3, ส้ม 2 เป็นต้น และยังส่งผลให้ประโยคนั้นๆ มีความหมายที่กำกวมไม่ชัดเจน แต่ในภาษาอังกฤษนั้นไม่มีการใช้คำลักษณนาม และเมื่อต้องการบ่งบอกจำนวนคำนามหลักก็จะอยู่ข้างหลัง และมีการผันคำตามพจน์ เป็นพหูพจน์ตามหลักไวยากรณ์ เช่น 3 birds., 3 oranges. เป็นต้น

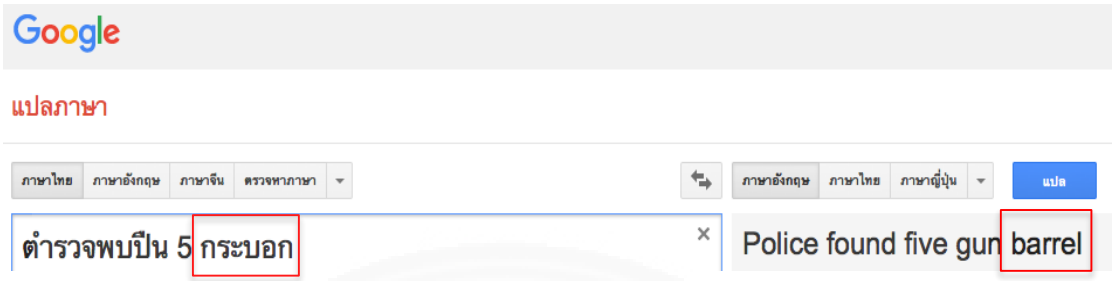
จากความแตกต่างกันของลักษณนามและรูปแบบของทั้งสองภาษา เมื่อนำประโยคที่มีคำลักษณนาม ร่วมอยู่ในประโยคไปทดลองแปลด้วย Google Translation เนื่องจาก Google Translation เป็นเครื่องมือการแปลภาษาด้วยเครื่องที่แพร่หลายและเป็นระบบการแปลภาษาเชิงสถิติที่มีคลังข้อมูลภาษาขนาดใหญ่ และมีแบบจำลองภาษาขนาดใหญ่ (Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och & Jeffrey Dean, 2007, pp.858-867) ก็ยังคงพบความผิดพลาดในการแปล (Miss Translation) ดังนี้

1. ประโยคที่ถูก: ช้าง 5 เชือก = 5 elephants.
   Google แปล: ช้าง 5 เชือก = Chang 5 rope.



ภาพที่ 1.1 แสดงผลการแปลประโยคที่มีคำลักษณนามจาก Google Translation ประโยคที่ 1

2. ประโยคที่ถูก: ตำรวจพบปืน 5 กระบอก = Police found five guns.
Google แปล: ตำรวจพบปืน 5 กระบอก = Police found five gun barrel



ภาพที่ 1.2 แสดงผลการแปลประโยคที่มีคำลักษณนามจาก Google Translation ประโยคที่ 2

สาเหตุของการแปลผิดมาจากเครื่องคอมพิวเตอร์เลือกจับคู่คำลักษณนามในภาษาไทยกับภาษาอังกฤษผิดพลาด ซึ่งจากตัวอย่างข้างต้นจะเห็นว่าแม้จะเป็นประโยคง่ายๆแต่เครื่องก็ยังเลือกจับคู่คำผิดพลาด เพราะ ภาษาอังกฤษนั้นไม่มีคำลักษณนาม เมื่อเครื่องเจอคำลักษณนามในภาษาไทย จึงนำคำในภาษาอังกฤษที่มีความหมายตรงกันจากข้อมูลและสถิติในระบบมาทำการจับคู่ให้กับคำลักษณนามนั้น โดย การจับคู่คำ (Word Alignment) เป็นกระบวนการหนึ่งในระบบการแปลภาษาด้วยเครื่องเชิงสถิติ ซึ่งถ้าหากการจับคู่คำในระหว่างสองภาษา(Bilingual Word Alignment) มีความถูกต้อง ก็จะส่งผลให้การแปลมีความถูกต้อง ในทางกลับกันถ้าการจับคู่คำผิดก็จะส่งผลให้การแปลมีความกำกวม และมีความหมายที่ผิดเพี้ยนไปจากที่ควรจะเป็น

จากปัญหาที่ได้กล่าวมางานวิจัยนี้จึงมีแนวคิดที่จะแก้ปัญหาด้วยการลดความกำกวมที่เกิดจากคำลักษณนามในภาษาไทย (Thai Classifier Disambiguation) โดยมุ่งเน้นการแก้ปัญหาไปที่การจับคู่คำระหว่างคู่ภาษาไทยกับภาษาอังกฤษ เพื่อให้เครื่องคอมพิวเตอร์สามารถจับคู่คำในประโยคที่มีคำลักษณนามได้ถูกต้อง โดยมีสมมติฐานว่า ถ้านำคำลักษณนามออกจากประโยคก่อนนำมาจับคู่คำแล้ว หลังจากที่ทำการจับคู่คำเสร็จเรียบร้อยแล้ว ค่อยนำคำลักษณนามกลับเข้าไปในประโยคทีหลัง จะทำให้การจับคู่คำระหว่างคู่ภาษาไทยกับภาษาอังกฤษมีความถูกต้อง และลดความกำกวมในการจับคู่คำที่เกิดจากคำลักษณนามในภาษาไทยได้

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อแก้ไขปัญหาความกำกวมที่เกิดจากประโยคที่มีคำลักษณนามในภาษาไทย
2. เพื่อทำให้การจับคู่คำในประโยคที่มีคำลักษณนาม ในคู่ภาษาไทย และภาษาอังกฤษ มีความถูกต้องมากขึ้น
3. เพื่อตรวจสอบว่าการจัดการกับความกำกวมที่เกิดจากประโยคลักษณนาม ในกระบวนการจับคู่คำในคู่ภาษาไทย และภาษาอังกฤษ มีผลทำให้การแปลภาษาด้วยเครื่องในคู่ประโยคที่มีคำลักษณนามมีความถูกต้องมากขึ้น

1.3 ขอบเขตของงานวิจัย

1. ข้อมูลการฝึก (Training Data) จะถูกเลือกมาจากคลังข้อมูลคู่ภาษาจำนวน 50,000 ประโยค โดยเลือกเอาเฉพาะประโยคที่มีลักษณนาม
2. กระบวนการจับคู่คำ (Word Alignment) จะดำเนินการโดยอัตโนมัติด้วยโปรแกรม GIZA ซึ่งเป็นซอฟท์แวร์ฟรีของ Statistical Machine Translation (www.statmt.org)
3. การลดความกำกวมจะมุ่งเน้นไปที่ลักษณนาม (Classifier) เท่านั้น
4. การลดความกำกวมของลักษณนาม จะกระทำโดยโปรแกรมเสริม (Pre-Post Processing) โดยไม่ได้ไปแก้ไขโปรแกรม GIZA ที่มีอยู่เดิม
5. การวัดผลการจับคู่คำวัดผลจากวลีที่มีลักษณนามเท่านั้น (Classifier Phrase Only) โดยตรวจสอบจากการจับคู่คำของคำลักษณนามในวลีนั้นๆ
6. การวัดผลการจับคู่คำของประโยคที่มีลักษณนาม จะวัดผลจากความถูกต้องของการจับคู่คำ และผลการแปลที่ได้จากการจับคู่คำ

1.4 ข้อจำกัดของงานวิจัย

1. โปรแกรมเสริม (Pre-Post Processing) สามารถจัดการกับประโยคที่มีความกำกวม ที่เกิดจากการจับคู่คำ ในประโยคที่มีคำลักษณนามได้เท่านั้น
2. เนื่องจากการลดความกำกวมของลักษณนามของงานวิจัยนี้ต้องผ่านกระบวนการตัดคำ (Word Segmentation) และการติดป้ายชนิดของคำ (POS Tagging) โดยโปรแกรม SWATH ซึ่งหาก SWATH ทำการตัดคำ และติดป้ายชนิดของคำลักษณนามผิดพลาด ก็จะส่งผลต่อประโยคที่มีคำลักษณนามประโยคนั้นๆ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อลดความกำกวมของคำลักษณนามที่เกิดจากภาษาไทย
2. เพื่อพัฒนาการจับคู่คำ ในคู่ภาษาไทย และภาษาอังกฤษ ที่มีคำลักษณนามรวมอยู่ในประโยคให้มีความถูกต้องมากขึ้น
3. เพื่อให้การแปลภาษาด้วยเครื่องในคู่ภาษาไทย และภาษาอังกฤษ ที่มีคำลักษณนามรวมอยู่ในประโยคมีความถูกต้องมากขึ้น
4. เพื่อเป็นแนวทางในการพัฒนาการแปลภาษาด้วยเครื่องต่อไป

บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 คำลักษณนาม (Classifier) เป็นคำที่ใช้ร่วมกับคำนามเป็นปกติในภาษาไทย เพื่อให้ประโยคมีการสื่อความหมายได้สมบูรณ์และถ้าหากไม่ใช้จะทำให้ประโยคมีความแปลกประหลาด ผิดไวยากรณ์ ซึ่ง (Pornsiri Singhapreecha, 2001, pp.259-170) ได้มีการจำแนกลักษณะการใช้งานของลักษณนามในภาษาไทย เทียบเคียงกับคำแปลในภาษาอังกฤษ ดังนี้

1. แบบง่าย (Simple Nominal)

1.1 คำลักษณนามที่ใช้เมื่อมีคำนาม ที่มาพร้อมกับตัวเลข โดยจะแสดงในภาพที่ 2.1 และ ถ้าหากไม่ใช้จะทำให้ประโยคในภาษาไทยแปลกประหลาดผิดไวยากรณ์ โดยจะแสดงในภาพที่ 2.2

นก 3 *ตัว* (bird) (three) (*classifier*) three birds.

ภาพที่ 2.1 ตัวอย่างประโยคที่ใช้คำลักษณนาม
ร่วมกับคำนามที่มาพร้อมกับตัวเลข

? นก 3 **Ø** (bird) (three) (~~*classifier*~~) three birds.

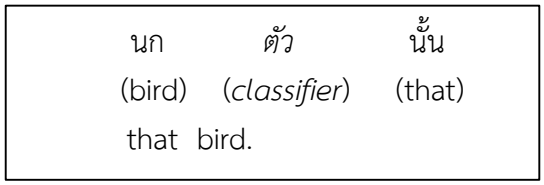
ภาพที่ 2.2 ตัวอย่างประโยคที่ไม่ใช้คำลักษณนาม

1.2 คำลักษณนามที่ใช้ร่วมกับคำคุณศัพท์ (Adjective) โดยจะแสดงในภาพที่ 2.3

นก *ตัว* เล็ก (bird) (*classifier*) (little) little bird.

ภาพที่ 2.3 ตัวอย่างประโยคที่ใช้คำลักษณนาม
ร่วมกับคำคุณศัพท์

1.3 คำลักษณนามที่ใช้ร่วมกับคำบ่งชี้ (Determiner) โดยจะแสดงในภาพที่ 2.4



ภาพที่ 2.4 ตัวอย่างประโยคที่ใช้คำลักษณนาม
ร่วมกับคำบ่งชี้

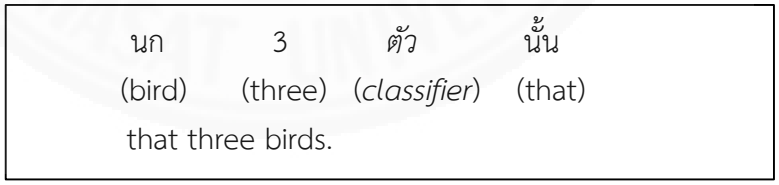
2. แบบซับซ้อน (Complex Nominal)

คำลักษณนาม ที่ใช้ในประโยคร่วมกับทั้ง 3 รูปแบบในแบบง่าย ตั้งแต่ 2 รูปแบบขึ้นไป

2.1 คำลักษณนาม ที่ใช้ร่วมกับ คำนามที่มาพร้อมกับตัวเลข และคำคุณศัพท์ โดยจะแสดงในภาพที่ 2.5

ภาพที่ 2.5 ตัวอย่างประโยคที่ใช้คำลักษณนาม
ร่วมกับคำนามที่ตามด้วยตัวเลข และคำคุณศัพท์

2.2 คำลักษณนาม ที่ใช้ร่วมกับ คำนามที่มาพร้อมกับตัวเลข และคำบ่งชี้ โดยจะแสดงในภาพที่ 2.6



ภาพที่ 2.6 ตัวอย่างประโยคที่ใช้คำลักษณนาม
ร่วมกับคำนามที่ตามด้วยตัวเลข และคำบ่งชี้

2.3 คำลักษณนาม ที่ใช้ร่วมกับ คำนามที่มาพร้อมกับตัวเลข คำคุณศัพท์ และคำบ่งชี้ โดยจะแสดงในภาพที่ 2.7

นก *ตัว* เล็ก 3 *ตัว* นั้น (bird) (*classifier*) (little) (three) (*classifier*) (that) that three little birds.

ภาพที่ 2.7 ตัวอย่างประโยคที่ใช้คำลักษณนาม
ร่วมกับคำนามที่ตามด้วยตัวเลข คำคุณศัพท์ และคำบ่งชี้

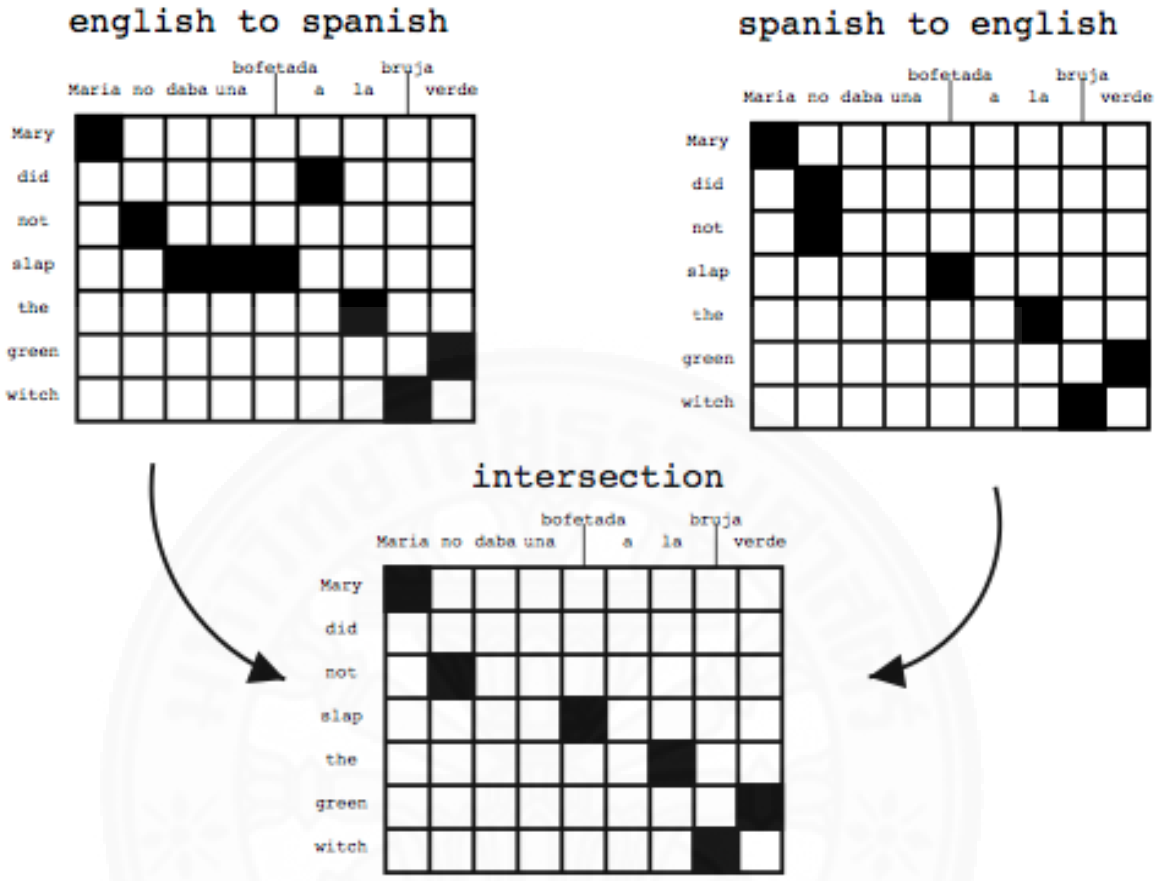
2.1.2 ระบบแปลภาษาเชิงสถิติ (SMT: Statistical Machine Translation) คือ ระบบแปลภาษาด้วยคอมพิวเตอร์ ที่อาศัยข้อมูลจากคลังข้อมูลคู่ภาษา (Parallel Corpus) เพื่อสร้างประโยคในภาษาปลายทางที่เหมาะสมที่สุดตามข้อมูลสถิติที่มีอยู่ในระบบ ซึ่ง (Kevin Knight & Philipp Koehn, 2003) ได้ให้หลักการแปลภาษาเชิงสถิตินั้นมีเป้าหมาย คือ การสร้างประโยคในภาษาอังกฤษ (e) จะได้ P(e) ซึ่งเรียกว่าแบบจำลองภาษา (Language Model) ที่มีค่าความน่าจะเป็นสูงที่สุดที่จะเป็นคำแปลของประโยคภาษาไทย (f) จะได้ P(f | e) ซึ่งเรียกว่า แบบจำลองการแปล (Translation Model) โดยแบบจำลองการแปลจะช่วยคำนวณความน่าจะเป็นของคำในประโยคต้นทาง (f) แล้วนำไปแปลเป็นคำในประโยคปลายทาง (e) ซึ่งแบบจำลองการแปลนี้ถูกสร้างขึ้นโดย IBM Model เมื่อมีแบบจำลองภาษา และ แบบจำลองการแปลแล้ว พอมีประโยคใหม่เข้ามาระบบก็จะทำการถอดรหัสการแปล (Decoder) โดยหาคำแปลในภาษาอังกฤษ (e) ที่มีค่าความน่าจะเป็นที่สูงที่สุดจาก P(e) x P(f | e) ซึ่งเรียกว่าขั้นตอนการถอดรหัส (Decoder Algorithm) ดังสมการ

$e_{best} = \text{argmax}_e P(e) \times P(f \mid e)$

แบบจำลองการแปล (Translation Model) มีขั้นตอนสำคัญขั้นตอนหนึ่งก่อนจะมาถึงขั้นตอนของการถอดรหัส นั่นก็คือการจับคู่คำ (Word Alignment) โดยจะเป็นการจับคู่คำที่มีความหมายตรงกันของภาษาต้นทางและภาษาปลายทาง ซึ่งเป็นส่วนที่งานวิจัยนี้ให้ความสนใจ โดยแบบจำลองการแปลนั้นจะประกอบด้วยคู่ของคำ หรือคู่ของวลี แล้วจัดหาคู่คำในการแปลของคำ หรือวลีที่รับเข้ามา โดยการจับคู่คำมีลักษณะ ดังแสดงในภาพที่ 2.8

ภาพที่ 2.8 แสดงการจับคู่คำของภาษาต้นทาง (สเปน) และ ภาษาปลายทาง (อังกฤษ) (Kevin Knight & Philipp Koehn, 2003)

ต่อมามีการทำจับคู่คำ โดย (Franz Josef Orch & Hermann Ney, 2000) และ (Philipp Koehn, 2003) ให้จับคู่คำทั้งสองทาง คือ จากภาษาต้นทาง (f) ไปยังภาษาปลายทาง (e) และทำจากภาษาปลายทาง (e) ไปยังภาษาต้นทาง (f) [f -> e, e -> f] ซึ่งเรียกว่า bidirectionally aligned corpora เมื่อทำการจับคู่คำทั้งสองทางเรียบร้อยแล้ว ให้นำผลการจับคู่คำที่ได้มาทำ intersection of alignment point ดังแสดงในภาพที่ 2.9 ซึ่งการทำ intersection of alignment point คือการเลือกเอาเฉพาะคู่คำที่จับคู่คำเหมือนกันที่ได้จากการจับคู่คำ จากภาษาต้นทางไปยังภาษาปลายทาง และจากภาษาปลายทางมายังภาษาต้นทาง ซึ่งจะส่งผลให้ precision นั้นมีความแม่นยำที่สูงขึ้น แต่ recall ยังคงต่ำอยู่ จึงได้มีการปรับ grow additional alignment point เพื่อให้ ค่า recall ที่ต่ำนั้นมีค่าสูงขึ้น โดยที่ค่า precision ยังมีความแม่นยำที่สูงเช่นเดิม ดังแสดงในภาพที่ 2.10



ภาพที่ 2.9 แสดงการจับคู่คำแบบ bidirectionally aligned corpora [f -> e, e -> f] แล้วนำมา intersection of alignment point (Kevin Knight & Philipp Koehn, 2003)

ภาพที่ 2.10 แสดงการทำ grow additional of alignment point

2.1.3 คำลักษณนามในภาษาไทย (Thai Classifier) นอกจากจะมีลักษณะการใช้งานที่หลากหลายแล้ว ประเภท และ รูปแบบการใช้ก็ยังมีความหลากหลายด้วย โดย (Virach Sornlertlamvanich, Wantanee Pantachat & Surapant Meknavin, 1994) ได้จำแนกประเภทของลักษณนามไว้ดังนี้

1. Unit Classifier

คือลักษณนามที่ใช้เพื่อระบุความสัมพันธ์กับคำนามเป็นหน่วยของจำนวนนับ เช่น ตัว, ลูก, เล่ม

ตัวอย่าง : นก 3 *ตัว*

2. Collective Classifier

คือลักษณนามที่ใช้เพื่อระบุความสัมพันธ์กับคำนามในลักษณะของการอยู่รวมกันเป็นกลุ่ม เช่น ฝูง, กลุ่ม, กอง

ตัวอย่าง : นก 3 *ฝูง*

3. Metric Classifier

คือลักษณนามที่ใช้เพื่อระบุมาตรวัดที่สามารถวัดได้ที่มีความสัมพันธ์กับคำนามและคำกริยา เช่น กิโลกรัม, เมตร

ตัวอย่าง : หนัก 3 *กิโลกรัม*

4. Frequency Classifier

คือลักษณนามที่ใช้เพื่อระบุถึงความถี่ที่เกิดขึ้นจากเหตุการณ์ต่างๆ เช่น รอบ, ครั้ง

ตัวอย่าง : บิน 3 *รอบ*

5. Verbal Classifier

คือลักษณนามที่ใช้เพื่อระบุถึงความสัมพันธ์กับคำนามในลักษณะที่เกิดจากการกระทำ เช่น ม้วน, ห่อ

ตัวอย่าง : กระดาษ 3 *ม้วน*

และได้จำแนกรูปแบบของลักษณนามไว้ดังนี้

1. Enumeration

Noun/Verb - Cardinal number - Classifier

ตัวอย่าง : นก 3 ตัว

2. Ordinal

Noun - Classifier - /tii/ที่ - Cardinal number

ตัวอย่าง : นกตัวที่ 3

3. Determination

Noun - Determiner - Classifier

ตัวอย่าง : นกตัวนั้น

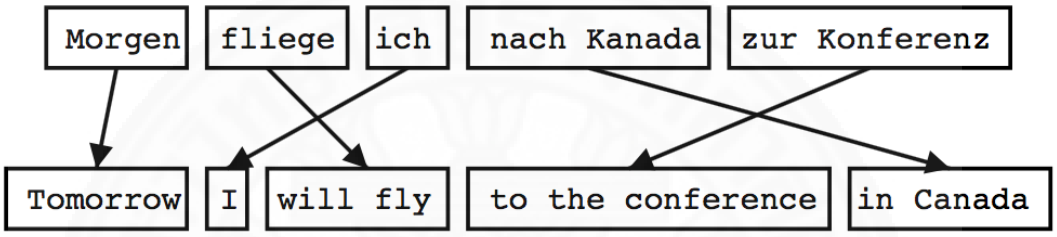
4. Attributive

Noun - Classifier - Attributive

ตัวอย่าง : นกตัวอ้วน

จากนั้นได้นำประโยคที่มีคำลักษณนามที่ได้ทำการตัดคำ (Segmentation) และ ติดป้ายชนิดของคำ (Parts of Speech tagging) มาจำแนกออกตามรูปแบบต่างๆ (Pattern Matching) แล้ว ทำให้คำลักษณนามต่างๆใช้กับคำนามนั้นๆได้ถูกต้องมากขึ้น

2.1.4 การแปลภาษาด้วยเครื่องโดยใช้ระบบแปลภาษาเชิงสถิติ มีหลายระดับ โดย (Philipp Koehn, Franz Josef Och & Daniel Marcu, 2003) ได้เสนอการแปลภาษาในระดับของวลี (Phrase-base translation) คือการแปลโดยใช้การจับคู่คำจากการแปลภาษาในระดับคำเพื่อนำมาสกัดคู่วลีและใช้บริบทของคำในการแปล มีความสัมพันธ์แบบกลุ่มต่อกลุ่ม



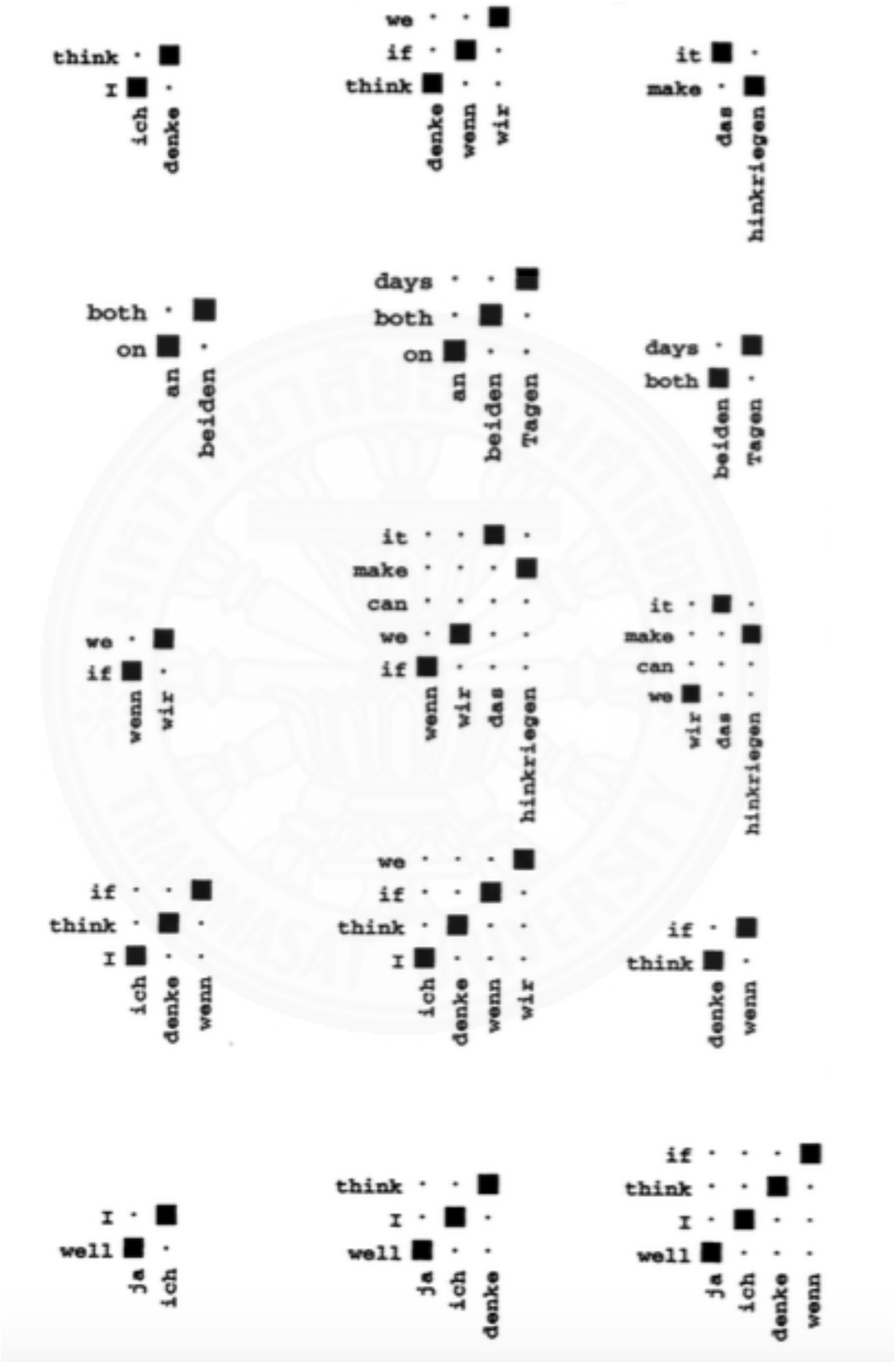
ภาพที่ 2.11 แสดงการแปลภาษาระดับวลี (Phrase-base translation)
(Philipp Koehn, Franz Josef Och & Daniel Marcu, 2003)

ซึ่งการแปลภาษาระดับวลีนั้นให้ผลที่ดีกว่าการแปลภาษาในระดับคำ และถ้ามีข้อมูลที่มากขึ้น ก็ยิ่งสามารถเรียนรู้กลุ่มคำได้ยาวมากขึ้นด้วย

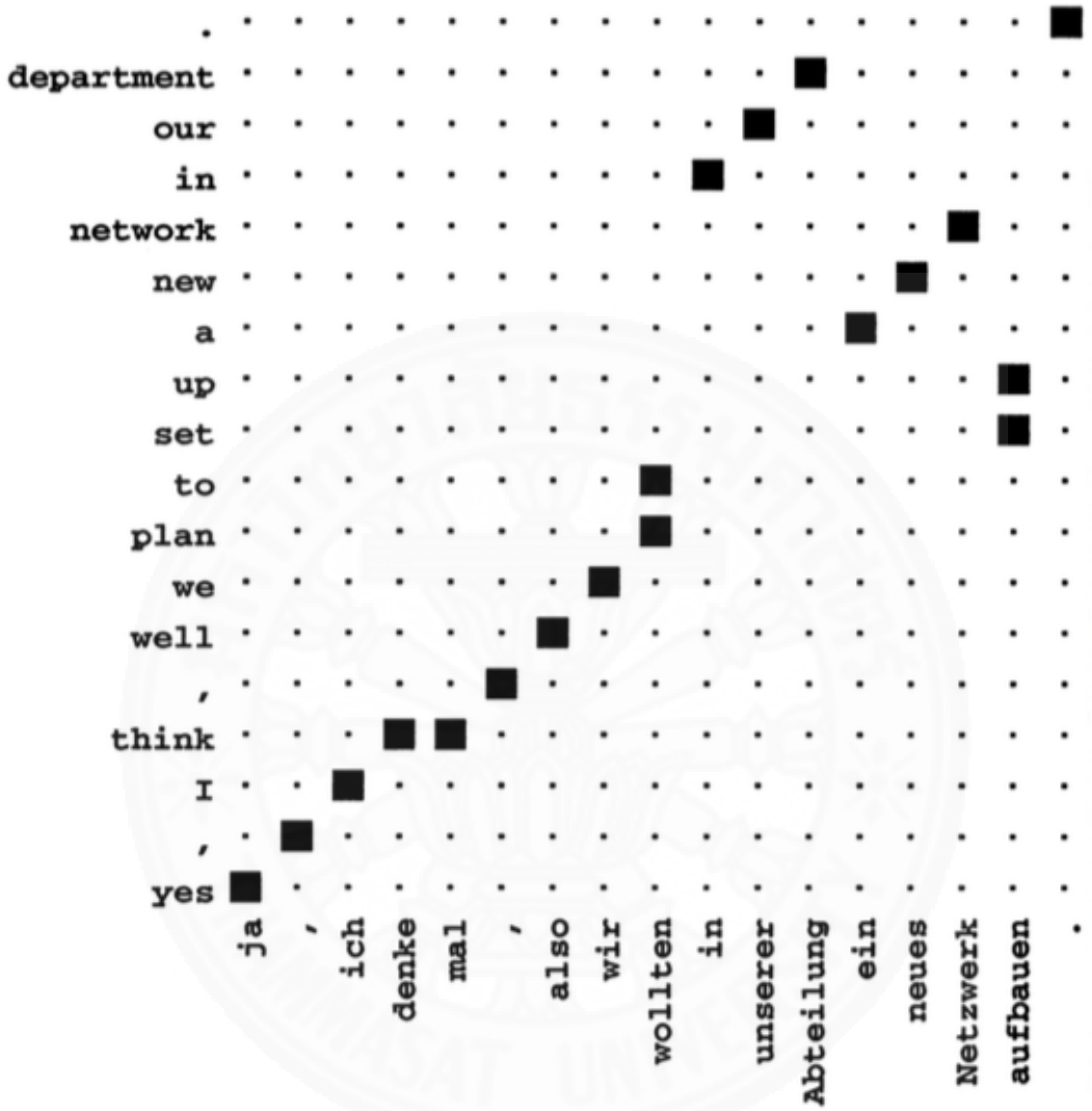
2.1.5 GIZA เป็นหนึ่งในเครื่องมือสำหรับการแปลภาษาด้วยเครื่องเชิงสถิติ (Statistical Machine Translation) โดยเป็นเครื่องมือที่ใช้จับคู่คำ (Word Alignment) ที่ถูกสร้างโดย (Franz Josef Och & Hermann Ney, 2000) โดยใช้ฝึก (Train) จากแบบจำลอง ซึ่ง GIZA นั้นเป็นเครื่องมือที่ถูกนำมาใช้อย่างแพร่หลายสำหรับการจับคู่คำในคลังข้อมูลคู่ภาษา (Bilingual Corpus) โดยมีแบบแผนซึ่ง (Franz Josef Och & Hermann Ney, 2004, pp.417-449) ได้ให้ตัวอย่างไว้ดังนี้

ja ,	yes ,
ja , ich	yes , I
ja , ich denke mal	yes , I think
ja , ich denke mal ,	yes , I think ,
ja , ich denke mal , also	yes , I think , well
, ich	, I
, ich denke mal	, I think
, ich denke mal ,	, I think ,
, ich denke mal , also	, I think , well
, ich denke mal , also wir	, I think , well we
ich denke mal	I think
ich denke mal ,	I think ,
ich denke mal , also	I think , well
ich denke mal , also wir	I think , well we
ich denke mal , also wir wollten	I think , well we plan to
denke mal ,	think ,
denke mal , also	think , well
denke mal , also wir	think , well we
denke mal , also wir wollten	think , well we plan to
, also	, well
, also wir	, well we
, also wir wollten	, well we plan to
also wir	well we
also wir wollten	well we plan to
wir wollten	we plan to
in unserer	in our
in unserer Abteilung	in our department
in unserer Abteilung ein neues Netzwerk	a new network in our department
in unserer Abteilung ein neues Netzwerk aufbauen	set up a new network in our department
unserer Abteilung	our department
ein neues	a new
ein neues Netzwerk	a new network
ein neues Netzwerk aufbauen	set up a new network
neues Netzwerk	new network

ภาพที่ 2.12 แสดงตัวอย่างวลีคู่ภาษา(เยอรมัน - อังกฤษ) ตั้งแต่ 2-7 คำ สำหรับการจับคู่คำ



ภาพที่ 2.13 แสดงตัวอย่างแบบการฝึก(training) ในการจับคู่คำ



ภาพที่ 2.14 แสดงตัวอย่างการจับคู่คำ

ซึ่งในปัจจุบันนั้นการจับคู่คำเป็นพื้นฐานสำหรับการทำแบบจำลองระดับวลี (Phrase-base Model)

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 เนื่องจากคำลักษณนามในภาษาไทยนั้นมีอยู่จำนวนมาก และคำนามแต่ละคำก็ยังใช้คำลักษณนามแตกต่างกันไป โดย (Virach Sornlertlamvanich et al, 1994) ได้เสนอวิธีการเลือกคำลักษณนามให้กับคำนาม โดยใช้คลังข้อมูลภาษาไทยที่มีขนาดใหญ่ นำมาทำการตัดคำ (Word Segmentation) แล้วติดป้ายชนิดของคำ (POS tagging) เพื่อให้ทราบว่าคำไหนเป็นคำลักษณนาม จากนั้นนำมาเทียบกับรูปแบบของคำลักษณนาม (Classifier Pattern Matching) ที่เกิดขึ้นในภาษาไทยรูปแบบต่างๆ ทำให้ทราบว่า คำนามแต่ละคำมีความถี่ที่เกิดร่วมกับคำลักษณนามมากน้อยแค่ไหน ส่งผลให้สามารถเลือกคำลักษณนามต่างๆมาใช้ร่วมกับคำนามนั้นๆได้อย่างถูกต้อง

จะเห็นได้ว่า จากขั้นตอนที่กล่าวมาทำให้เราสามารถระบุ และจำแนกรูปแบบคำลักษณนามต่างๆได้ แต่จากที่กล่าวมาก็ยังไม่มีการนำมาประยุกต์กับการจับคู่คำ (Alignment) ของคู่ภาษาไทยและภาษาอังกฤษ ซึ่งมีวลีที่มีคำลักษณนามร่วมอยู่ในประโยค

2.2.2 ภาษาญี่ปุ่นเป็นหนึ่งในหลายภาษาที่มีการใช้คำลักษณนาม โดย (Francis Bond et al, 1996) ได้นำเสนอวิธีการแปลลักษณนาม โดยจำแนกลักษณนามออกเป็นประเภทต่างๆ ได้แก่ Unit Classifier, Metric Classifier, Group Classifier และ Species Classifier ซึ่งมีลักษณะคล้ายคลึงกับในภาษาไทย แล้วกำหนดกฎ (Rule-Based) มาจัดการกับ Classifier ในแต่ละประเภท ส่งผลให้สามารถแปลวลีที่มีคำลักษณนามได้ถูกต้องมากขึ้น

2.2.3 (Michael Paul, Eiichiro Sumita & Seiichi Yamamoto, 2002) ได้นำเสนอวิธีการแปลลักษณนามที่มีตัวเลขร่วมด้วย โดยวิธีการใช้ Corpus-based กับ Numeral Classifier Phrase Alignment ในคลังข้อมูลคู่ภาษาญี่ปุ่นและภาษาอังกฤษที่มีขนาดใหญ่ซึ่งทำให้การแปลมีผลลัพธ์ที่ถูกต้องมากขึ้น

ตารางที่ 2.1

แสดงเทคนิคที่ใช้และผลลัพธ์ในงานวิจัย

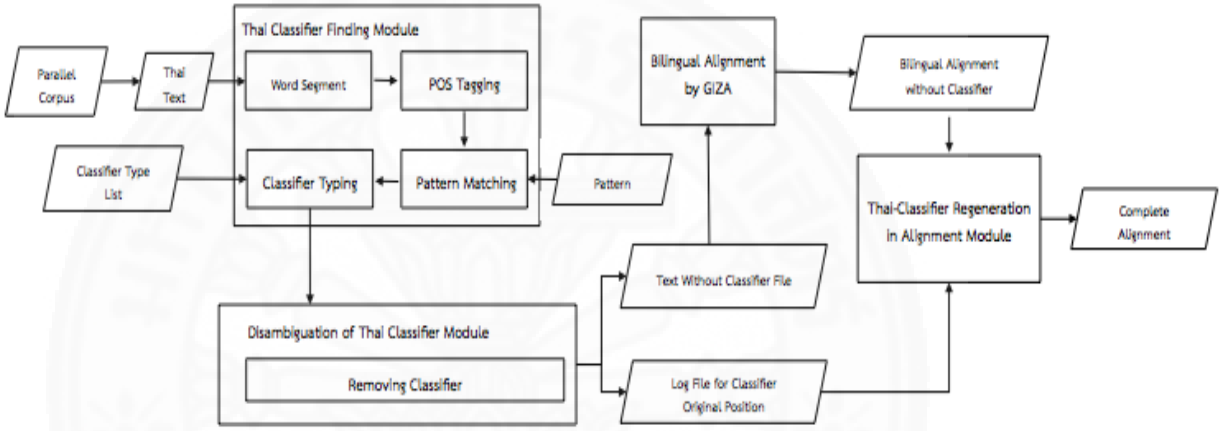
งานวิจัย	เทคนิคที่ใช้	ผลลัพธ์
Virach Sornlertlamvanich et al., 1994	Word Segmentation, POS tagging, Classifier Pattern Matching และใช้ Corpus-Based	สามารถระบุลักษณนามให้กับคำนามได้ถูกต้องมากขึ้น
Francis Bond et al., 1996	แยกชนิดของลักษณนาม(Classifier type) และกำหนดกฎ(Rule-based) มาจัดการกับลักษณนามแต่ละชนิด	สามารถแปลภาษาที่มีลักษณนามได้ถูกต้องมากขึ้น (ภาษาญี่ปุ่น-อังกฤษ)

Michael Paul et al., 2002	ใช้ Numeral Classifier Phrase Alignment โดยดูจากวลีที่ลักษณนามมีตัวเลขร่วมด้วยเป็นหลัก โดยใช้ Corpus-based	สามารถแปลภาษาที่มีลักษณนามได้ถูกต้องมากขึ้น (ภาษาญี่ปุ่น-อังกฤษ)

บทที่ 3
วิธีการวิจัย

3.1 ภาพรวมของระบบ

ระบบการแก้ไขความกำกวมของลักษณนามในภาษาไทยนี้เป็นโปรแกรมเสริม (Plug-in) เพื่อช่วยให้การจับคู่คำนั้นมีความถูกต้องมากขึ้นซึ่งไม่ได้เข้าไปปรับเปลี่ยนหรือแก้ไขโปรแกรมการจับคู่คำที่มีอยู่เดิม โดยภาพรวมของระบบจะแสดงในภาพที่ 3.1



ภาพที่ 3.1 ภาพรวมการทำงานของระบบ

โดยการทำงานของระบบมีดังนี้

1. เริ่มจากคลังข้อมูลคู่ภาษา (Parallel Corpus)
2. เลือกประโยคที่เป็นภาษาไทย (Thai Text)
3. เข้าสู่กระบวนการค้นหาลักษณนาม (Thai Classifier Finding Module)
4. นำผลลัพธ์จากข้อ 3.(Classifier Typing) เข้าสู่กระบวนการจัดการความกำกวมของลักษณนาม (Disambiguation of Thai Classifier Module)
5. นำผลลัพธ์จากข้อ 4. (Thai without Classifier Word) ไปเข้าสู่กระบวนการจับคู่คำโดย GIZA (Bilingual Alignment by GIZA)
6. นำข้อมูลที่บันทึกจากข้อ 4.(Log File for Classifier Original Position และ ผลลัพธ์จากข้อ 5.(Bilingual Alignment without Classifier) เข้าสู่กระบวนการคืนลักษณนามในการจับคู่คำ (Thai-Classifier Regeneration in Alignment Module)
7. การจับคู่คำเสร็จสมบูรณ์ (Complete Alignment)

3.2 ขั้นตอนการทำงานของระบบ

การทำงานของระบบจะมุ่งเน้นไปที่การจัดการกับความกำกวมที่เกิดขึ้นจากลักษณนาม ในภาษาไทยโดยส่วนของโปรแกรมเสริม (Pre-Post Processing) โดยขั้นตอนมีดังนี้

3.2.1 เตรียมข้อมูลจากคลังคู่ภาษา (Parallel Corpus) ดังแสดงตัวอย่างในตารางที่ 3.1

ตารางที่ 3.1

ตัวอย่างข้อมูลคลังคู่ภาษาไทย - อังกฤษ

ภาษาไทย	ภาษาอังกฤษ
นก 3 ตัว	three birds.
นกตัวเล็ก	little birds.
นกตัวเล็ก 3 ตัว	three little birds.
นกตัวนั้น	that birds.
นก 3 ตัวนั้น	that three birds.

3.2.2 กระบวนการค้นหาลักษณนามในภาษาไทย (Thai Classifier Finding Module)

3.2.2.1 นำข้อมูลประโยคภาษาไทยจากข้อ 3.2.1 จากคลังข้อมูลภาษา (Corpus) มาทำการตัดคำ (Word Segmentation) และติดป้ายชนิดของคำ (POS Tagging) ซึ่งเลือกตัดคำโดยใช้โปรแกรม SWATH โดย (Paisarn Charoenpornsawat, 2003) เนื่องจากโปรแกรม SWATH สามารถตัดคำ และติดป้ายชนิดของคำได้โดยอัตโนมัติ โดยการตัดคำและการติดป้ายของคำจาก SWATH นั้นใช้ข้อมูลจาก ORCHID Corpus โดย (Virach Sornlertlamvanich, Thatsanee Chareonporn & Hitoshi Isahara, 1997) และได้ระบุชุดชนิดของคำไว้ทั้งหมด 47 ชนิด ผลลัพธ์ที่ได้จากการตัดคำและติดป้ายชนิดของคำจะแสดงในตารางที่ 3.2

ตารางที่ 3.2

ตัวอย่างประโยค ก่อน-หลัง ตัดคำ และ ติดป้ายชนิดของคำ

ก่อน ตัดคำ และ ติดป้ายชนิดของคำ	หลัง ตัดคำ และ ติดป้ายชนิดของคำ
แม่ซื้อไก่ 4 ตัว	แม่@NCMN|ซื้อ@VACT|ไก่@NCMN 4 ตัว@CNIT
แอปเปิ้ล 3 ลูก วางอยู่บนโต๊ะ	แอปเปิ้ล@NCMN 3 ลูก@CMTR วาง@VACT|อยู่@XVAE| บน@RPRE|โต๊ะ@NCMN

กุหลาบ 2 ช่อนี้ ผมให้คุณ	กุหลาบ@NCMN 2 ช่อ@CLTV|นี้@DDAC ผม@PPRS|ให้@JSBR|คุณ@NTTL
เขากำลังตัดต้นไม้ต้นนั้น	เขา@PPRS|กำลัง@XVBM|ตัด@VACT|ต้นไม้@NCMN|ต้น@CMTR|นั้น@DDAC

โดยรูปแบบของการตัดคำและการติดป้ายชนิดของคำคือ คำ+@+ชนิดของคำ และใช้เครื่องหมาย | ในการแบ่งคำแต่ละคำ

3.2.2.2 หลังจากที่ตัดคำและติดป้ายชนิดของคำเสร็จเรียบร้อย ก็จะนำประโยคมาเทียบกับรูปแบบ และ ประเภทของลักษณนาม ทำให้ทราบว่าประโยคนั้นๆจัดเป็นประโยคที่มีลักษณนามที่อยู่ในประเภท ดังแสดงในตารางที่ 3.3 และ รูปแบบใด ดังแสดงในตารางที่ 3.4 เพื่อนำไปจัดการกับความกำกวมในกระบวนการต่อไป

ตารางที่ 3.3

แสดงประเภทของลักษณะนาม และ ตัวอย่างคำลักษณนาม

ประเภทของลักษณนาม (Classifier Type)	**ตัวอย่างคำลักษณนาม (Classifier List)**
Unit : cl_u	คน, ตัว, เล่ม, แท่ง, ลูก
Collective : cl_c	กลุ่ม, ฝูง, ช่อ
Norminal Metric : cl_m.n	ลิตร, กิโลเมตร
Frequency : cl_f	ครั้ง, รอบ, ที
Verbal : cl_v	ม้วน, ห่อ

ตารางที่ 3.4

แสดงรูปแบบของประโยคลักษณนาม

นิยาม (Expression)	**รูปแบบ (Patterns)**
Enumeration	Noun – Cardinal number - Classifier
Ordinal	Noun – Classifier - /tii/(ที่) – Cardinal number
Determination	Noun – Classifier - Determiner
Attributive	Noun – Classifier – Attributive/Adjective verb

โดยรูปแบบชองประโยคลักษณนามที่พบในคลังข้อมูลคู่ภาษามีทั้งหมด 12 รูปแบบดังนี้

- ประโยคลักษณนามที่ไม่ซับซ้อน มีรูปแบบเพียง 1 รูปแบบ

รูปแบบที่ 1 : Noun – Cardinal number – Classifier
รูปแบบที่ 2 : Noun – Classifier – Attributive/Adjective verb
รูปแบบที่ 3 : Noun – Classifier – Determiner

- ประโยคลักษณนามที่ซับซ้อน มีรูปแบบมากกว่า 1 รูปแบบรวมอยู่ในประโยคลักษณนาม

รูปแบบที่ 4 : Noun – Cardinal number – Classifier
+ Noun – Cardinal number – Classifier
รูปแบบที่ 5 : Noun – Classifier – Attributive/Adjective verb
+ Noun – Classifier – Attributive/Adjective verb
รูปแบบที่ 6 : Noun – Classifier – Determiner
+ Noun – Cardinal number – Classifier
รูปแบบที่ 7 : Noun – Classifier – Attributive/Adjective verb
+ Noun – Cardinal number – Classifier
รูปแบบที่ 8 : Noun – Classifier – Determiner
+ Noun – Classifier – Attributive/Adjective verb
รูปแบบที่ 9 : Noun – Classifier – Determiner
+ Noun – Classifier – Determiner
รูปแบบที่ 10 : Noun – Cardinal number – Classifier
+ Noun – Cardinal number – Classifier
+ Noun – Cardinal number – Classifier
รูปแบบที่ 11 : Noun – Cardinal number – Classifier
+ Noun – Cardinal number – Classifier
+ Noun – Classifier – Attributive/Adjective verb
รูปแบบที่ 12 : Noun – Classifier - /tii/(ที่) – Cardinal number
+ Noun – Classifier – Determiner

3.2.3 กระบวนการจัดการความกำกวมของลักษณนามในภาษาไทย (Disambiguation of Thai Classifier Module)

การจัดการกับความกำกวมของลักษณนามในภาษาไทยนั้น หลังจากที่เราตัดคำ และ ติดป้ายชนิดของคำ จะแบ่งจัดการความกำกวมตามประเภทของลักษณนามตามที่ได้อธิบายไว้ในข้อ 3.2.1.2 โดยแบ่งออกเป็น 3 กรณี และมีขั้นตอนดังนี้

3.2.3.1 กรณีที่ 1 : ในกรณีนี้จะจัดการเฉพาะฝั่งภาษาไทยฝั่งเดียวเท่านั้น

- เมื่อนำประโยคมาเปรียบเทียบกับประเภทของลักษณะนามแล้วพบว่าเป็น Unit Classifier (cl_u) ให้ทำตามขั้นตอนดังนี้

(1) ดึงคำที่เป็นลักษณนามออกมาบันทึกไว้ที่ Log file

- นก 3 (ตัว)
- ตัว → Log file

(2) บันทึกตำแหน่งของคำลักษณนามที่ดึงออกมาไว้ที่ Log file

- ตัว = ตำแหน่งที่ 3 → Log file

(3) ลบคำลักษณนามออกจากประโยค (Removing Classifier)

- นก 3 ตัว → นก 3

หลังจากเสร็จสิ้นกระบวนการ จะได้ผลลัพธ์ออกมา 2 files คือ Text without classifier file และ Log file for classifier original position ดังแสดงตัวอย่างในตารางที่ 3.5

ตารางที่ 3.5

ตัวอย่างผลลัพธ์ของกรณีที่ 1

Text without classifier file	Log file for classifier original position
นก|3	ตัว 2 (word^{ith})
นก|เล็ก|3	ตัว 1 ตัว 4
นก|นั้น	ตัว 1

3.2.3.2 กรณีที่ 2 : ในกรณีนี้จะจัดการทั้งสองภาษา คือ ทั้งภาษาไทย และภาษาอังกฤษ

- เมื่อนำประโยคมาเปรียบเทียบกับประเภทของลักษณนาม แล้วพบว่าเป็น Collective Classifier (cl_c) / Metric Classifier (cl_m.n) หรือ Verbal Classifier (cl_v) ให้ทำตามขั้นตอนดังนี้

(1) ดึงคำที่เป็นลักษณนามออกมาบันทึกไว้ที่ Log file

- นก 3 (ฝูง) / Three (flock of) birds.
- ฝูง / flock of → Log file

(2) บันทึกตำแหน่งของคำลักษณนามที่ดึงออกมาไว้ที่ Log file

- ฝูง = ตำแหน่งที่ 3 / flock of = ตำแหน่งที่ 2, 3 → Log file

(3) ลบคำลักษณนามออกจากประโยค (Removing Classifier)

- นก 3 ฝูง → นก 3
- Three flock of birds. → Three birds.

หลังจากเสร็จสิ้นกระบวนการ จะได้ผลลัพธ์ออกมา 4 files คือ Text without classifier file และ Log file for classifier original position (อย่างละ 2 files เนื่องจากจัดการทั้งภาษาไทย และ ภาษาอังกฤษ) ดังแสดงตัวอย่างของฝั่งภาษาไทยในตารางที่ 3.6 และแสดงตัวอย่างของฝั่งภาษาอังกฤษในตารางที่ 3.7

ตารางที่ 3.6

ตัวอย่างผลลัพธ์ของกรณีที่ 2 (ภาษาไทย)

Text without classifier	Log file for classifier original position
นก|3	ฝูง 2(word^{ith})
กุหลาบ|2	ช่อ 2
น้ำ|4|อยู่|ใน|ขวด	ลิตร 2
ยาย|ซื้อ|ทิชชู่|4	ม้วน 4

ตารางที่ 3.7

ตัวอย่างผลลัพธ์ของกรณีที่ 2 (ภาษาอังกฤษ)

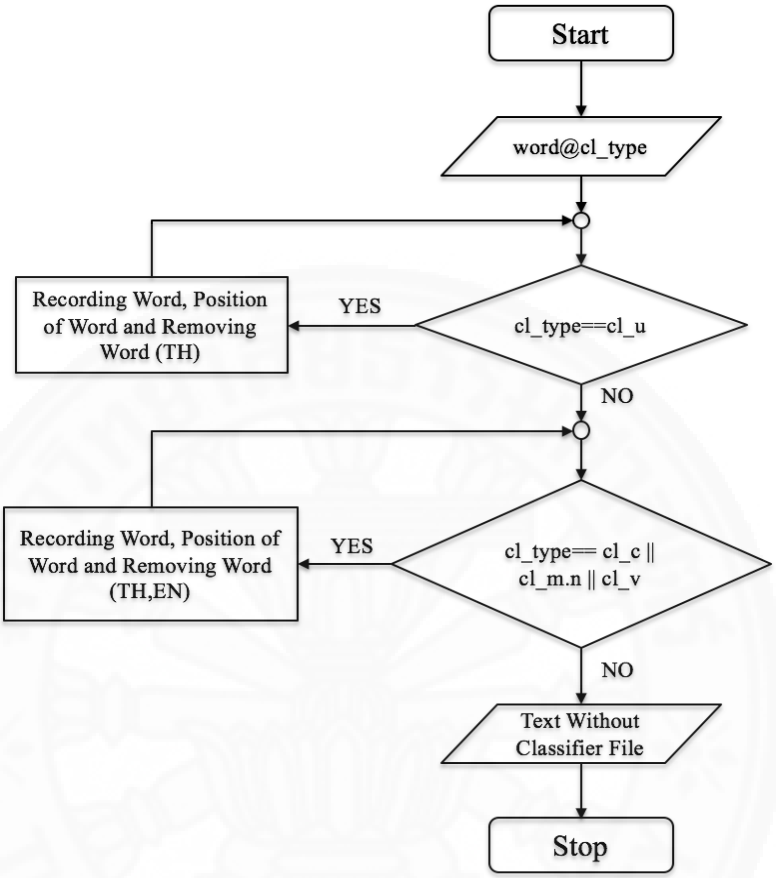
Text without classifier	Log file for classifier original position
three birds.	flocks 1(word^{ith}) of 2(word^{ith})
two roses.	bouquets 1 of 2
4 water is in bottles.	litres 1 of 2
grandmom buys 4 tissue.	rolls 3 of 4

3.2.3.3 กรณีที่ 3 : ในกรณีนี้จะไม่มีการแก้ไขอะไรกับประโยคทั้งสองภาษา

- เมื่อนำประโยคมาเปรียบเทียบกับประเภทของลักษณะนาม แล้วพบว่าเป็น Frequency Classifier (cl_f)

ซึ่งทั้ง 3 กรณีนั้น จะพิจารณาโดยเริ่มจากทางฝั่งภาษาไทยก่อน เพื่อดูว่าประโยคที่มีคำลักษณนามนั้นเป็นประโยคที่มีคำลักษณนามเข้าข่ายกรณีไหน แล้วจึงมาพิจารณาฝั่งภาษาอังกฤษในลำดับถัดมา

โดยผังงาน (Flow chart) ของการจัดการกับความกำกวมของคำลักษณนามตามประเภทต่างๆ จะแสดงดังภาพที่ 3.2



ภาพที่ 3.2 แสดงผังงานขั้นตอนการจัดการความกำกวมของลักษณนาม

โดย 1. เริ่มจากนำประโยคที่ผ่านการตัดคำ และติดป้ายชนิดของคำ มาตรวจสอบว่าตรงกับลักษณนามประเภทใด

2. กรณีที่ 1 :ถ้าประเภทของลักษณนาม (Classifier type) ตรงกับ Unit Classifier (cl_u) ให้ดึงคำลักษณนามและตำแหน่งของคำลักษณนามนั้นจากฝั่งภาษาไทย มาบันทึกไว้ใน Log file จากนั้นทำการลบคำลักษณนามนั้นออกจากประโยค

กรณีที่ 2 :ถ้าประเภทของลักษณนาม (Classifier type) ตรงกับ Collective Classifier (cl_c) หรือ Nominal Metric Classifier (cl_m.n) หรือ Verbal Classifier (cl_v) ให้ดึงคำลักษณนาม และตำแหน่งของคำลักษณนามนั้นจากทั้งสองภาษา (ไทย-อังกฤษ) มาบันทึกไว้ใน Log file จากนั้นทำการลบคำลักษณนามนั้นออกจากประโยค

3. ได้ Text without classifier file ซึ่งเป็น file ที่บันทึกประโยคที่ถูกลบคำลักษณนามออกไปแล้ว เพื่อนำไปทำการจับคู่คำในกระบวนการถัดไป

3.2.4 นำข้อมูลจาก Text without classifier file ซึ่งได้จากกระบวนการในข้อ 3.2.3 ไปทำการจับคู่คำสองภาษาด้วยโปรแกรม GIZA (Bilingual Alignment by GIZA) โดยผลลัพธ์ของการจับคู่คำที่ไม่มีคำลักษณนาม จะแสดงในตารางที่ 3.8

ตารางที่ 3.8

ตัวอย่างผลลัพธ์การจับคู่คำที่ได้จากโปรแกรม GIZA

Sentence pair(1) three little birds. NULL({}) นก({3}) เล็ก({2}) 3({1})
Sentence pair(2) grandmom buy 4 tissue. NULL({}) ยาย({1}) ซื้อ({2}) ทิชชู่({4}) 4({3})

จากตารางที่ 3.8 ตัวเลขที่อยู่ในเครื่องหมาย ({ }) หลังคำในประโยคภาษาไทยแต่ละคำ คือตัวเลขที่แสดงตำแหน่งของคำในภาษาอังกฤษ

จากคู่ประโยคที่ 1 : ตำแหน่งของแต่ละคำในประโยคนี้คือ three=1 little=2 birds=3

จากคู่ประโยคที่ 2 : ตำแหน่งของแต่ละคำในประโยคนี้คือ grandmom=1 buy=2 4=3 tissue=4

3.2.5 กระบวนการคืนคำลักษณนามของภาษาไทยในการจับคู่คำ (Thai Classifier Regeneration in Alignment Module)

กระบวนการนี้จะเริ่มดำเนินการหลังจากที่ได้ผลลัพธ์ของการจับคู่คำจาก GIZA ออกมาเป็นที่เรียบร้อยแล้ว ก็จะทำการคืนคำลักษณนามกลับไปในประโยค โดยการคืนคำลักษณนามกลับไปในประโยคนั้น แบ่งออกเป็น 2 กรณี ดังนี้

3.2.5.1 กรณีที่ 1 : Unit Classifier (cl_u)

ในกรณีนี้ จะทำการคืนคำลักษณนาม เฉพาะฝั่งภาษาไทยฝั่งเดียวเท่านั้น เนื่องจากคำลักษณะนามประเภทนี้ได้ถูกดึงออกมาจากฝั่งภาษาไทยเพียงฝั่งเดียว โดยการคืนคำลักษณนามกลับเข้าไปในประโยคตามตำแหน่งเดิมที่ได้เก็บบันทึกไว้ใน Log file จากกระบวนการลดความกำกวมในข้อ 3.2.3 เพื่อให้จับคู่คำเข้ากับคำนามหลัก โดยจะแสดงในตารางที่ 3.9

ตารางที่ 3.9

ตัวอย่างก่อน-หลัง กระบวนการคืนคำลักษณนามในการจับคู่คำของกรณีที่ 1

ก่อนคืนคำลักษณนาม	หลังคืนคำลักษณนาม
นก (~~ตัว~~) เล็ก 3 (~~ตัว~~) three little birds.	นก ***ตัว*** เล็ก 3 ***ตัว*** three little birds.

3.2.5.2 กรณีที่ 2 **:** Collective Classifier (cl_c) / Metric Classifier (cl_m.n) หรือ Verbal Classifier (cl_v)

ในกรณีนี้ จะทำการคืนคำลักษณนามจากทั้งสองภาษา ทั้งฝั่งภาษาไทยและภาษาอังกฤษ เนื่องจากคำลักษณะนามในกรณีนี้ได้ถูกดึงออกมาจากทั้งสองภาษา โดยเมื่อคืนลักษณนามกลับเข้าไปในประโยคของแต่ละภาษาตามเดิม ที่ได้เก็บบันทึกตำแหน่งไว้ใน Log file จากกระบวนการก่อนหน้านี้ เพื่อให้คำลักษณนามของทั้งสองภาษาจับคู่คำเข้าด้วยกัน โดยจะแสดงในตารางที่ 3.10

ตารางที่ 3.10

ตัวอย่างก่อน-หลัง กระบวนการคืนคำลักษณนามในการจับคู่คำของกรณีที่ 2

ก่อนคืนคำลักษณนาม	หลังคืนคำลักษณนาม
ยาย ซื้อ ทิชชู 4 (~~*ม้วน*~~) grandmom buy 4 (~~*rolls of*~~) tissue	ยาย ซื้อ ทิชชู 4 ***ม้วน*** grandmom buy 4 ***rolls of*** tissue

3.3 การวัดผลการทดลอง

3.3.1 การ Run ผลการทดลอง จะ Run บนเครื่อง Server ที่มี Spec ของ Hardware ดังแสดงในตารางที่ 3.11

ตารางที่ 3.11

แสดง Spec ของ Server ที่ใช้ Run ผลการทดลอง

SERVER SPEC	
CPU	8 core 2.3GHz
RAM	32 GB
Hard Disk	500 GB

3.3.2 การวัดผลการทดลองของงานวิจัยฉบับนี้ จะใช้การวัดผลโดยแยกเป็น 2 ส่วน คือ

1. วัดผลความถูกต้อง ของการจับคู่คำในวลีที่มีลักษณนาม (Classifier Phrase) โดยวัดความถูกต้องจากผู้เชี่ยวชาญทางด้านภาษาศาสตร์ (Expert)

   Percentage of Alignment Result
   = (Correct Alignment Sentences/All Classifier Sentences)*100

2. วัดผลความถูกต้อง ในการแปลที่สืบเนื่องมาจากการจับคู่คำ (Alignment) ด้วยค่า BLEU (Bilingual Evaluation Understudy)

$$BLEU = BP * EXP(\sum_{n=1}^{N}(W_n log P_n))$$

โดยจะทำการวัดผลการทดลองแยกตามรูปแบบที่เกิดขึ้นในประโยคที่มีคำลักษณนามทั้ง 12 รูปแบบ ดังนี้

รูปแบบที่ 1 : Noun – Cardinal number – Classifier
รูปแบบที่ 2 : Noun – Classifier – Attributive/Adjective verb
รูปแบบที่ 3 : Noun – Classifier – Determiner
รูปแบบที่ 4 : Noun – Cardinal number – Classifier
+ Noun – Cardinal number – Classifier
รูปแบบที่ 5 : Noun – Classifier – Attributive/Adjective verb
+ Noun – Classifier – Attributive/Adjective verb
รูปแบบที่ 6 : Noun – Classifier – Determiner
+ Noun – Cardinal number – Classifier
รูปแบบที่ 7 : Noun – Classifier – Attributive/Adjective verb
+ Noun – Cardinal number – Classifier

รูปแบบที่ 8 : Noun – Classifier – Determiner
+ Noun – Classifier – Attributive/Adjective verb
รูปแบบที่ 9 : Noun – Classifier – Determiner
+ Noun – Classifier – Determiner
รูปแบบที่ 10 : Noun – Cardinal number – Classifier
+ Noun – Cardinal number – Classifier
+ Noun – Cardinal number – Classifier
รูปแบบที่ 11 : Noun – Cardinal number – Classifier
+ Noun – Cardinal number – Classifier
+ Noun – Classifier – Attributive/Adjective verb
รูปแบบที่ 12 : Noun – Classifier - /tii/(ที่) – Cardinal number
+ Noun – Classifier – Determiner

เมื่อทำการวัดผลแยกทั้ง 12 รูปแบบเรียบร้อยแล้ว จะนำประโยคที่มีคำลักษณนามทั้งหมดมารวมกันเพื่อทำการทดสอบอีกครั้ง

หลังจากที่ได้ผลการทดลองจากประโยคที่มีคำลักษณนามทั้งหมด ซึ่งเป็นส่วนที่ผู้วิจัยมุ่งเน้นในงานวิจัยฉบับนี้เป็นที่เรียบร้อยแล้ว ผู้วิจัยจะนำประโยคที่มีคำลักษณนามรวมเข้ากับประโยคทั้งหมดเพื่อทดสอบหาความถูกต้องในการแปลภาษา

บทที่ 4
ผลการวิจัยและอภิปรายผล

การทดลองของงานวิจัยฉบับนี้ ได้ทดสอบโดยเปรียบเทียบกับการจับคู่คำ (Word Alignment) จาก GIZA ที่ยังไม่มีการแก้ไข ซึ่งใช้ข้อมูลจากคลังข้อมูลคู่ภาษา (Parallel Corpus) ที่สุ่มมาจาก BTEC Task ซึ่งเป็นคลังข้อมูลคู่ภาษาไทยและภาษาอังกฤษ จำนวน 50,000 ประโยค มาใช้เป็นชุดข้อมูลในการเรียนรู้ (Training Data) จากจำนวน 50,000 ประโยคนี้ พบว่ามี 1,021 ประโยคที่มีคำลักษณนามในภาษาไทย (Thai Classifier) ซึ่งเป็นเป้าหมาย (Target) ที่ทางผู้วิจัยมุ่งเน้น (Focus) ในงานวิจัยฉบับนี้

เนื้อหาในส่วนนี้จะกล่าวถึงการอภิปรายผลการทดลองซึ่งได้มาจากวิธีการดำเนินงานของการวิจัย โดยประกอบด้วย 2 ส่วน คือ

1. ผลลัพธ์ความถูกต้องในการจับคู่คำ (Alignment) ที่ได้จากวิธีดำเนินการของงานวิจัยฉบับนี้ (Proposed Method) เปรียบเทียบกับผลการจับคู่คำ (Alignment) ด้วย GIZA ระบบเดิม (Baseline)
2. ผลลัพธ์ความถูกต้องในการแปลภาษาที่สืบเนื่องมาจากการจับคู่คำจากวิธีดำเนินการของงานวิจัยฉบับนี้ (Proposed Method) เปรียบเทียบกับ ผลการแปลที่สืบเนื่องมาจากการจับคู่คำ (Alignment) ด้วย GIZA ระบบเดิม (Baseline)

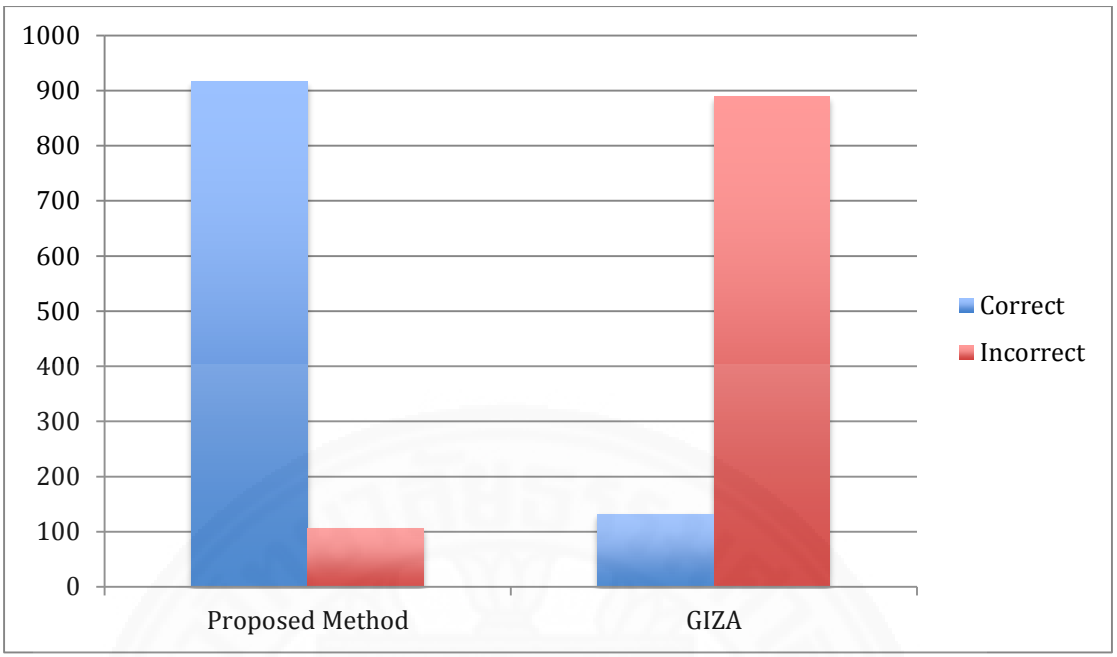
4.1 ผลลัพธ์ความถูกต้องในการจับคู่คำ

ในส่วนของความถูกต้องของการจับคู่คำ (Alignment) ผู้วิจัยจะมุ่งเน้นสังเกตเฉพาะความถูกต้องของวลีที่มีคำลักษณนาม (Classifier Phrase) เท่านั้น โดยตรวจสอบจากการจับคู่ของคำในวลีที่มีคำลักษณนาม (Word Aligning in Classifier Phrase) ซึ่งผลลัพธ์ความถูกต้องจะแสดงในตารางที่ 4.1

ตารางที่ 4.1

ผลลัพธ์การจับคู่คำของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA)

	Proposed Method		GIZA(Baseline)	
	Correct	Incorrect	Correct	Incorrect
Case Amount	916	105	132	889
Percentage(%)	89.72	11.28	12.93	87.07



ภาพที่ 4.1 กราฟผลลัพธ์การจับคู่คำของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA)

จากผลลัพธ์ พบว่า วิธีการดำเนินการของงานวิจัยฉบับนี้สามารถจับคู่คำในวลีที่มีคำลักษณนามได้ถูกต้องเพิ่มมากขึ้นอย่างน่าประทับใจเมื่อเปรียบเทียบกับ GIZA ที่เป็นระบบเดิม ซึ่งส่วนที่จับคู่คำได้ถูกต้องจาก GIZA นั้นก็เป็นส่วนหนึ่งของการจับคู่คำที่ถูกต้องของงานวิจัยฉบับนี้ด้วย

ในส่วนของวลีที่มีคำลักษณนาม 105 ประโยคที่จับคู่คำได้ไม่ถูกต้องจากวิธีการของงานวิจัยฉบับนี้ คือวลีที่ GIZA ก็จับคู่คำผิดเช่นเดียวกัน ซึ่งผู้วิจัยพบว่าสาเหตุที่ทำให้การจับคู่คำผิดนั้น มีผลมาจากกระบวนการของการติดป้ายชนิดของคำ (POS tagging) โดยที่กระบวนการดังกล่าวติดป้ายชนิดของคำได้ไม่ถูกต้อง (Wrong POS tagging) ซึ่งทำให้วลีลักษณนามในประโยคนั้นๆไม่ได้ถูกตรวจพบ ทำให้วลีที่มีคำลักษณนามในประโยคดังกล่าวไม่ได้ถูกจัดการตามกระบวนการของงานวิจัยฉบับนี้ โดยมีทั้งสิ้น 102 ประโยค ส่วนอีก 3 ประโยคนั้นเกิดมาจากสาเหตุที่คำลักษณนามนั้นไปจับคู่คำกับคำนามที่ไม่ถูกต้อง เนื่องจากประโยคเหล่านี้มีคำนาม 2 คำต่อเนื่องกันตัวอย่างเช่น (กล่อง | ฟิล์ม) ซึ่งวิธีดำเนินการของงานวิจัยฉบับนี้ ไม่ได้รวมการตรวจสอบคำนามหลัก และการระบุคำลักษณนามของคำนามแต่ละคำเข้าไปด้วย เป็นเหตุทำให้คำลักษณนามคำนั้นถูกจับคู่คำกับคำนามที่อยู่ใกล้ที่สุด ส่งผลให้การจับคู่คำนั้นเกิดความคลาดเคลื่อน ซึ่งในส่วนนี้สามารถพัฒนาปรับปรุงได้โดยการระบุรายการของคำลักษณนามที่เป็นไปได้ให้กับคำนามหลัก

4.2 ผลลัพธ์ความถูกต้องในการแปลภาษาที่สืบเนื่องมาจากการจับคู่คำ

ในส่วนของความถูกต้องในการแปลภาษานั้น ผู้วิจัยได้ใช้ BLEU Score ซึ่งเป็นการวัดค่าความถูกต้องในการแปลของการแปลภาษาด้วยเครื่อง โดย (Kishore Papineni, Salim Roukos, Todd Ward & Wei-Jing Zhu, 2002, pp.311-318) โดยจะเปรียบเทียบการแปลกับเอกสารอ้างอิง ซึ่ง

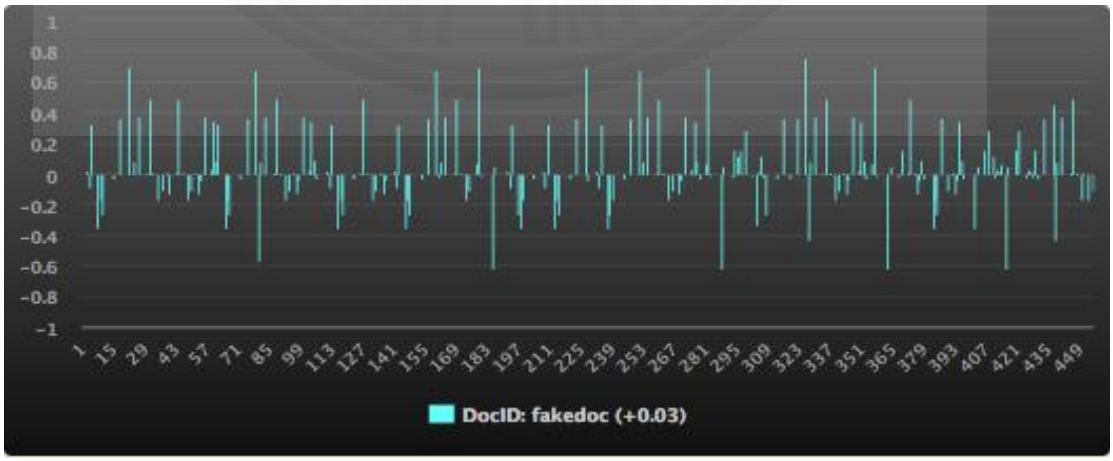
BLEU Score จะประเมินความถูกต้องของการแปลภาษาด้วยเครื่อง ซึ่งถ้า BLEU Score มีค่ามาก นั่นหมายความว่าการแปลภาษาด้วยเครื่องยิ่งมีความถูกต้อง

เพื่อที่จะวัดว่าการจับคู่คำ (Alignment) ของวลีที่มีลักษณนามจากวิธีดำเนินการของงานวิจัยส่งผลต่อความถูกต้องในการแปลภาษาด้วยเครื่อง ผู้วิจัยจึงได้นำผลการจับคู่คำที่ได้จากวิธีดำเนินการของงานวิจัยฉบับนี้ (Proposed Method) แทรกไปแทนผลการจับคู่ของ GIZA ที่เป็นระบบเดิม (Baseline) แล้วนำไปแปลด้วย Moses ที่เป็นเครื่องมือแปลภาษาอัตโนมัติของการแปลภาษาด้วยเครื่อง โดยตัว GIZA เองนั้นก็เป็นขั้นตอนหนึ่งของการแปลภาษาด้วย Moses แล้วนำผลการแปลที่สืบเนื่องจากการจับคู่คำด้วย GIZA ซึ่งเป็นระบบเดิม และการจับคู่คำด้วยวิธีดำเนินการที่งานวิจัยฉบับนี้มาเปรียบเทียบกันด้วย BLEU Score ซึ่งผลลัพธ์มีดังนี้

ตารางที่ 4.2

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 1

Translation Result	BLEU Score
GIZA(Baseline)	31.02
Proposed Method	33.94

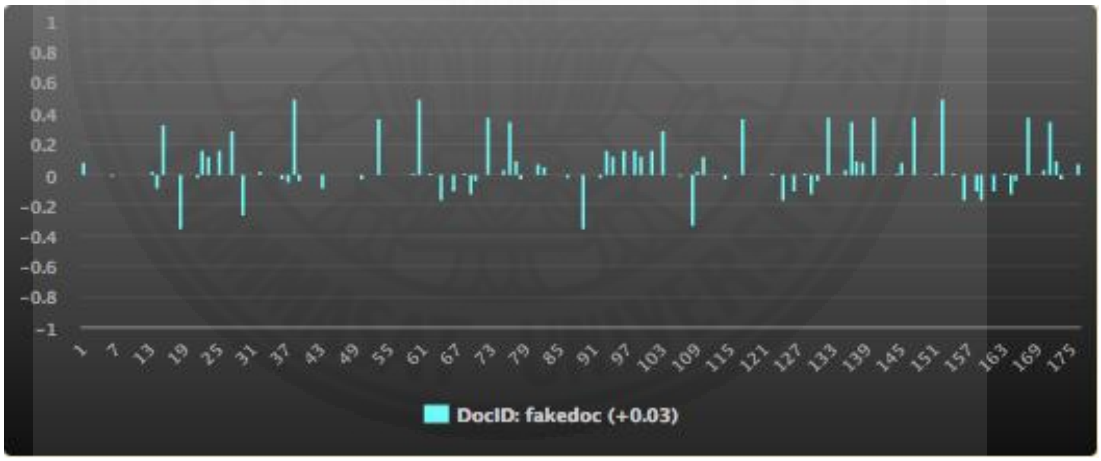


ภาพที่ 4.2 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 1

ตารางที่ 4.3

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 2

Translation Result	BLEU Score
GIZA(Baseline)	30.95
Proposed Method	33.86

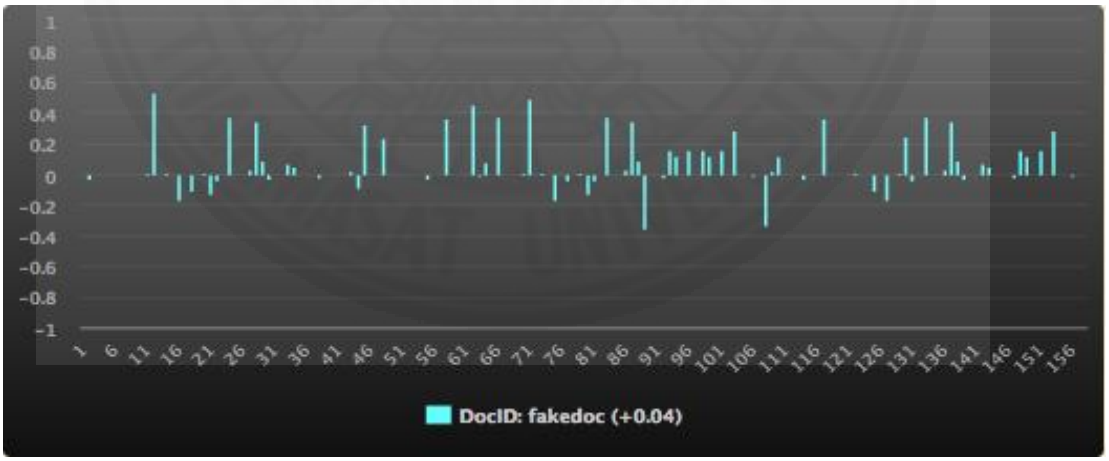


ภาพที่ 4.3 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 2

ตารางที่ 4.4

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 3

Translation Result	BLEU Score
GIZA(Baseline)	28.37
Proposed Method	32.38



ภาพที่ 4.4 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 3

ตารางที่ 4.5

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 4

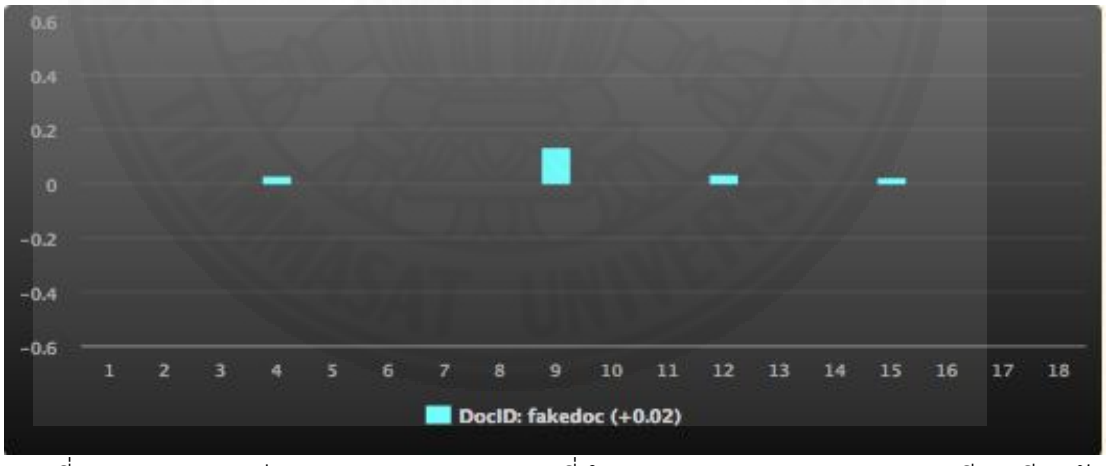
Translation Result	BLEU Score
GIZA(Baseline)	28.93
Proposed Method	32.61

ภาพที่ 4.5 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 4

ตารางที่ 4.6

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 5

Translation Result	BLEU Score
GIZA(Baseline)	41.38
Proposed Method	43.37



ภาพที่ 4.6 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 5

ตารางที่ 4.7

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 6

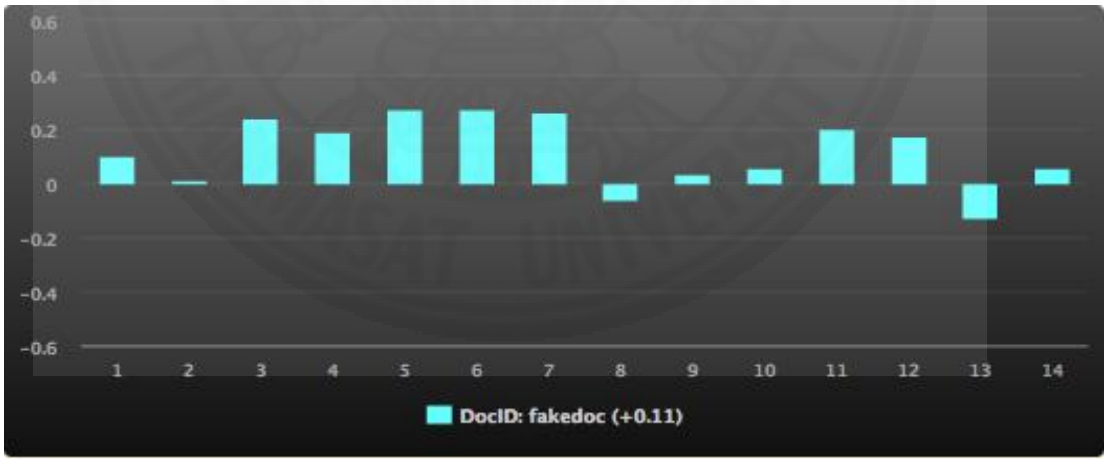
Translation Result	BLEU Score
GIZA(Baseline)	34.58
Proposed Method	35.98

ภาพที่ 4.7 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 6

ตารางที่ 4.8

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 7

Translation Result	BLEU Score
GIZA(Baseline)	20.46
Proposed Method	31.37

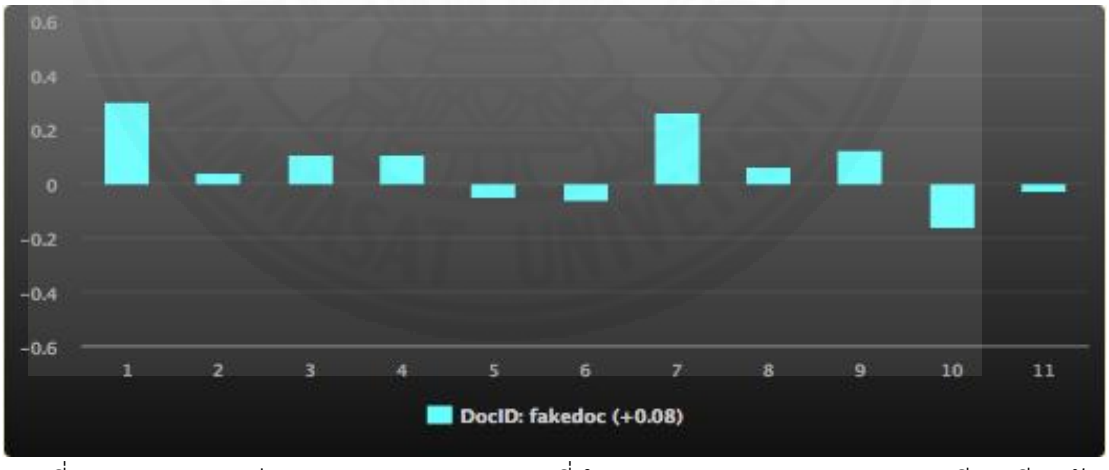


ภาพที่ 4.8 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 7

ตารางที่ 4.9

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 8

Translation Result	BLEU Score
GIZA(Baseline)	33.33
Proposed Method	41.18



ภาพที่ 4.9 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 8

ตารางที่ 4.10

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 9

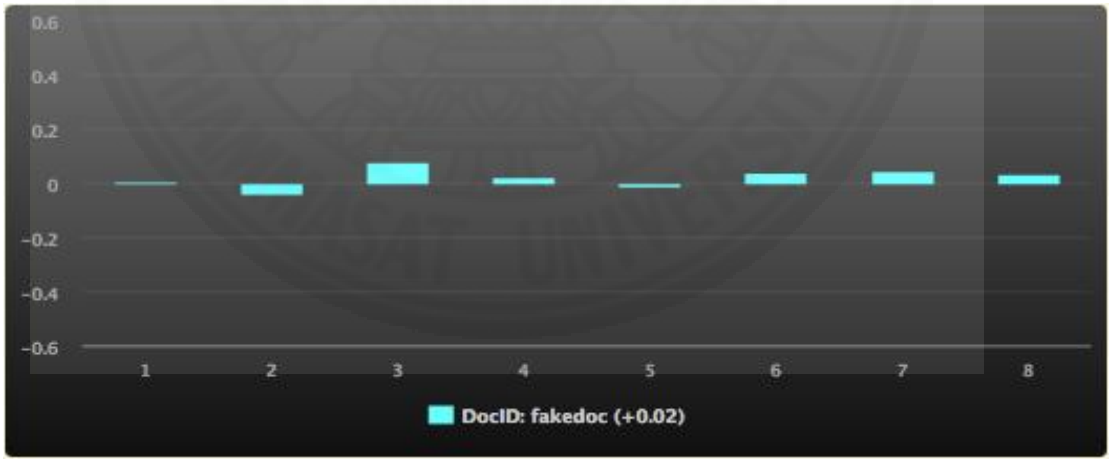
Translation Result	BLEU Score
GIZA(Baseline)	17.39
Proposed Method	23.27

ภาพที่ 4.10 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 9

ตารางที่ 4.11

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 10

Translation Result	BLEU Score
GIZA(Baseline)	11.11
Proposed Method	12.74



ภาพที่ 4.11 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 10

ตารางที่ 4.12

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 11

Translation Result	BLEU Score
GIZA(Baseline)	17.08
Proposed Method	18.74

ภาพที่ 4.12 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 11

ตารางที่ 4.13

แสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ของประโยคที่มีลักษณนามรูปแบบที่ 12

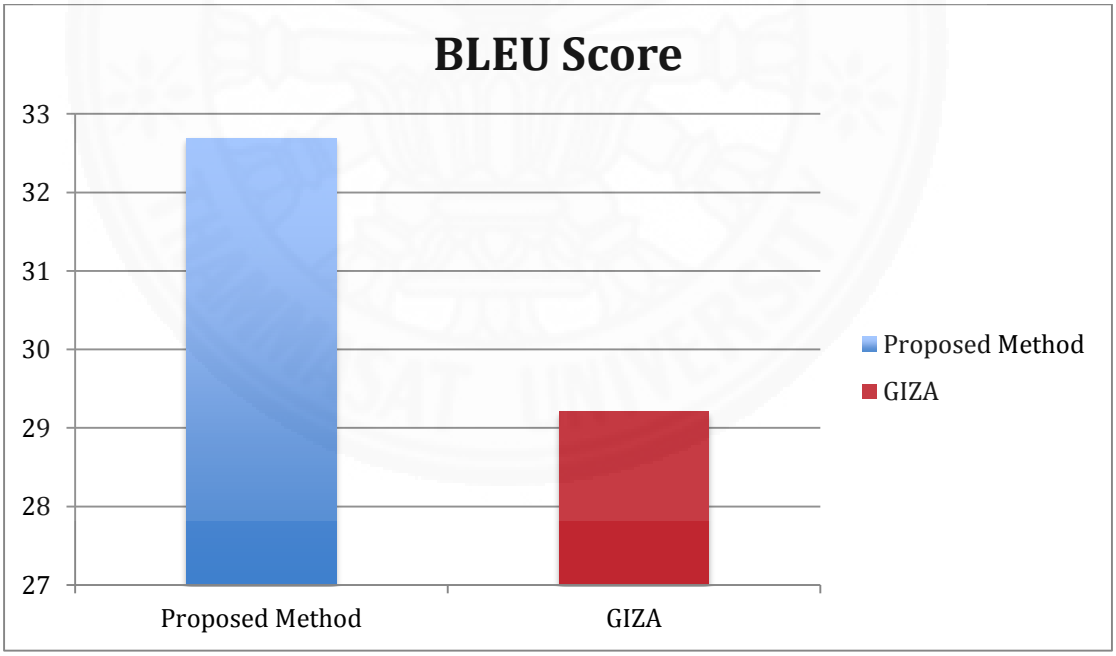
Translation Result	BLEU Score
GIZA(Baseline)	5.93
Proposed Method	6.74

ภาพที่ 4.13 กราฟแสดงค่า BLEU Score ของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA) ในแต่ละประโยคของรูปแบบที่ 12

ตารางที่ 4.14

แสดงค่า BLEU Score ประโยคลักษณนามทั้งหมดของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA)

Translation Result	BLEU Score
GIZA(Baseline)	29.21
Proposed Method	32.69



ภาพที่ 4.14 กราฟแสดงค่า BLEU Score ประโยคลักษณนามทั้งหมดของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA)

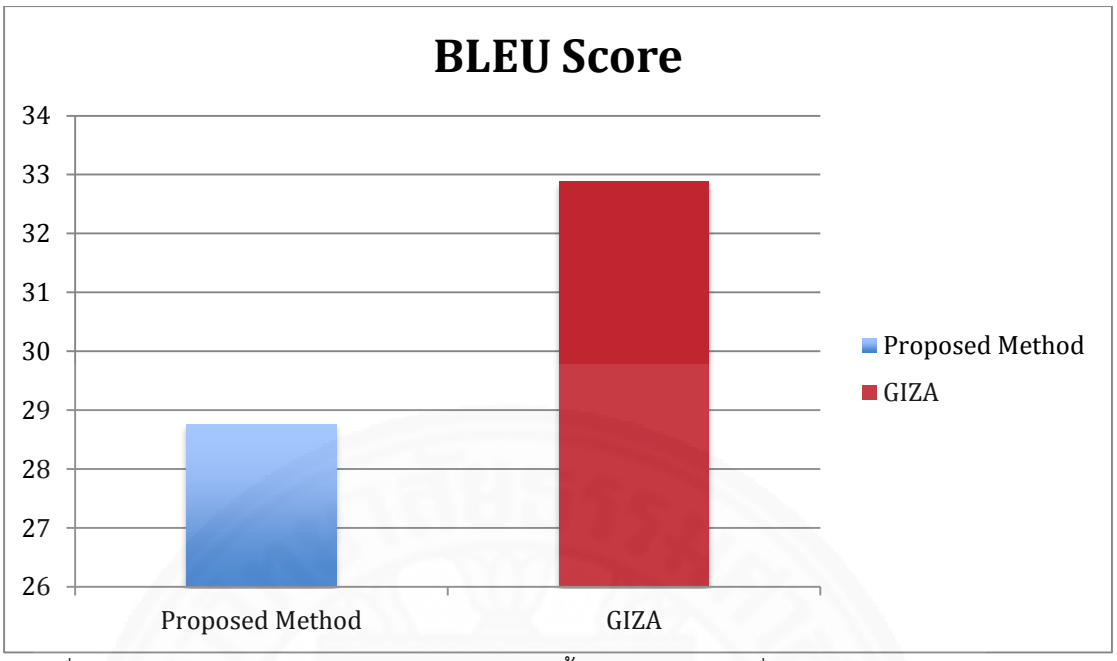
จากผลลัพธ์ แสดงให้เห็นว่าวิธีดำเนินการที่นำเสนอในงานวิจัยฉบับนี้ ทำให้การแปลประโยคซึ่งมีคำลักษณนามอยู่นั้นแปลได้ถูกต้องมากขึ้น โดยจะเห็นได้จากค่า BLEU Score ที่เพิ่มมากขึ้น เพราะเมื่อนำคำลักษณนามในภาษาไทยออกไปในขั้นตอนของกระบวนการลดความกำกวม ส่งผลให้การจับคู่คำอื่นๆนั้น จับคู่กันได้ถูกต้องเพิ่มมากขึ้น ตัวอย่างเช่น ลักษณนามคำว่า “ลูก” จะไม่ถูกจับคู่เข้ากับ คำว่า “**Child**” ในภาษาอังกฤษ และเมื่อนำคำลักษณนามกลับคืนเข้าไปในประโยค ก็จะทำให้คำลักษณนามนั้นถูกจับคู่เข้ากับคำที่ถูกต้อง ซึ่งทำให้ผลการแปลภาษาด้วยเครื่องซึ่งสืบเนื่องมาจากการจับคู่คำมีความถูกต้องมากขึ้น ซึ่งผลการแปลนี้วัดจากประโยคที่มีคำลักษณนามเท่านั้น

เมื่อทำการทดสอบเฉพาะประโยคที่มีลักษณนามซึ่งเป็นส่วนที่ผู้วิจัยมุ่งเน้นในงานวิจัยนี้เสร็จเรียบร้อยแล้ว ทางผู้วิจัยได้ทดลองนำประโยคแบบอื่นๆทั้งหมด ทั้งประโยคที่มีลักษณนาม และประโยคที่ไม่มีลักษณนาม มาทำการทดสอบเพื่อหาผลลัพธ์ความถูกต้องในการแปลภาษา ซึ่งได้ผลลัพธ์ดังนี้

ตารางที่ 4.15

แสดงค่า BLEU Score ประโยคทั้งหมดของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA)

Translation Result	**BLEU Score**
GIZA(Baseline)	32.89
Proposed Method	28.75



ภาพที่ 4.15 กราฟแสดงค่า BLEU Score ประโยคทั้งหมดของระบบที่นำเสนอ (Proposed Method) เปรียบเทียบกับระบบเดิม (GIZA)

จากผลลัพธ์พบว่า เมื่อนำประโยคทั้งหมดมาร่วมทดสอบด้วยนั้น ค่า BLEU Score ของระบบเดิมที่ได้มาจากการแปลที่สืบเนื่องมาจากการจับคู่คำของ GIZA ให้ผลลัพธ์การแปลที่ดีกว่าระบบที่นำเสนอในงานวิจัยฉบับนี้ เนื่องจากระบบที่นำเสนอนั้นมุ่งจัดการปัญหาเฉพาะประโยคที่มีคำลักษณนามเท่านั้น ซึ่งเมื่อนำประโยคอื่นๆมาทดสอบด้วยแล้วทำให้ผลการแปลได้ค่าน้อยลง มีสาเหตุเนื่องมาจากว่า การแก้ไขกระบวนการจับคู่คำในประโยคลักษณนามนั้น ส่งผลกระทบไปยังขั้นตอนอื่นๆในกระบวนการแปลภาษา เช่น การเรียงลำดับของคำ (Reorder) และการสร้างตารางวลี (Phrase Table) ทำให้การแปลภาษาในขั้นตอนของการถอดรหัส (Decode) เพื่อเลือกคำหรือวลีมาแปลภาษาในประโยคอื่นๆที่ไม่ใช่ประโยคที่มีคำลักษณนามนั้นแปลได้ถูกต้องน้อยลง ซึ่งประโยคอื่นๆที่ไม่ใช่ประโยคที่มีคำลักษณนามนั้นก็มีจำนวนประโยคในคลังข้อมูลคู่ภาษามากกว่าประโยคที่มีคำลักษณนามซึ่งเป็นส่วนน้อย ทำให้ผลลัพธ์ในการแปลจากวิธีดำเนินการที่งานวิจัยนี้นำเสนอได้ผลถูกต้องน้อยกว่าการแปลจากระบบเดิม

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

วิทยานิพนธ์ฉบับนี้ ได้นำเสนอวิธีดำเนินการในการขจัดความกำกวมคำลักษณนามของภาษาไทย ในกระบวนการจับคู่คำสองภาษา ด้วยการนำคำลักษณนามออกจากประโยคนั้นๆ ก่อนที่จะนำมาจับคู่คำ และเมื่อจับคู่คำเสร็จสิ้นแล้วจึงนำคำลักษณนามกลับคืนเข้าไปในประโยค และบังคับจับคู่เข้ากับคำนามหลัก และวัดผลการทดลองด้วยการวัดความถูกต้องของการจับคู่คำ และความถูกต้องในการแปลภาษาที่สืบเนื่องมาจากการจับคู่คำ โดยเปรียบเทียบกับผลการจับคู่คำ และความถูกต้องในการแปลภาษาที่สืบเนื่องมาจากการจับคู่คำของระบบเดิมที่ยังไม่มีการจัดการกับคำลักษณนาม ซึ่งในบทนี้จะกล่าวถึงผลสรุปของการดำเนินงานวิจัย สรุปผลการทดลอง และข้อเสนอแนะเพื่อใช้เป็นแนวทางในการศึกษาวิจัยต่อไปในอนาคต

5.1 สรุปผลการดำเนินงานวิจัย

วิทยานิพนธ์ฉบับนี้ต้องการขจัดความกำกวมที่เกิดจากคำลักษณนามของภาษาไทย ที่เกิดขึ้นในกระบวนการจับคู่คำสองภาษา ซึ่งการสรุปผลการวิจัยจะแบ่งออกเป็น การสรุปวัตถุประสงค์ของการวิจัย สรุปวิธีการศึกษาวิจัย ดังนี้

5.1.1 วัตถุประสงค์ของการวิจัย

งานวิจัยฉบับนี้ได้บรรลุวัตถุประสงค์ที่กำหนดขึ้น ดังนี้

1. ขจัดความกำกวมที่เกิดขึ้นจากคำลักษณนามของภาษาไทย
2. การจับคู่คำระหว่างประโยคที่มีคำลักษณนาม ระหว่างคู่ภาษาไทย และภาษาอังกฤษ มีความถูกต้องมากขึ้น
3. การแปลภาษาด้วยเครื่องที่เป็นผลสืบเนื่องมากจากการแก้ไขการจับคู่คำของประโยคที่มีคำลักษณนาม ระหว่างคู่ภาษาไทย และภาษาอังกฤษ มีความถูกต้องมากขึ้น

5.1.2 วิธีการวิจัย

งานวิจัยนี้ได้ใช้คู่ภาษาไทย และภาษาอังกฤษ ที่มีคำลักษณนามรวมอยู่ในประโยคเป็นข้อมูลเริ่มต้น (input) ซึ่งมีวิธีการโดยสรุปดังนี้

1. นำประโยคที่มีคำลักษณนามไปทำการตัดคำ (Word Segmentation) และติดป้ายชนิดของคำ (POS tagging) ด้วย SWATH
2. ทำการเทียบชนิดและรูปแบบของการใช้คำลักษณนาม
3. ทำการนำคำลักษณนามออกจากประโยค
4. นำประโยคที่ไม่มีคำลักษณนามไปจับคู่คำด้วย GIZA
5. นำคำลักษณนามคืนกลับไปยังตำแหน่งเดิม และจับคู่คำลักษณนามเข้ากับคำนามหลัก
6. นำผลการจับคู่คำมาตรวจสอบความถูกต้อง โดยเปรียบเทียบระหว่าง ผลการจับคู่คำด้วยวิธีดำเนินการของงานวิจัยฉบับนี้ กับ การจับคู่คำด้วย GIZA โดยไม่มีการแก้ไขคำลักษณนาม
7. นำผลการจับคู่คำที่ได้ ไปแปลภาษาด้วย Moses ซึ่งเป็นเครื่องมือแปลภาษาด้วยเครื่องอัตโนมัติ แล้วเปรียบเทียบค่าความถูกต้องด้วย BLEU Score

5.2 สรุปผลการทดลอง

จากการทดลองพบว่า วิธีที่นำเสนอในงานวิจัยฉบับนี้สามารถจัดการความกำกวมที่เกิดจากคำลักษณนามของภาษาไทยในกระบวนการจับคู่คำสองภาษา โดยการตรวจจับเพื่อหาชนิด รูปแบบ และความรู้ทางภาษา ถูกนำไปใช้เพื่อเป็นกฎในการแก้ปัญหาการจับคู่คำลักษณนามเหล่านี้ จากวิธีที่ผู้วิจัยได้นำเสนอ โดยการนำคำลักษณนามออกจากประโยคที่มีการจับคู่คำทั้งหมด จากนั้นค่อยนำ คำลักษณนามคืนกลับเข้าไปในประโยคและปรับผลการจับคู่คำ ส่งผลให้การจับคู่คำระหว่างคู่ประโยคภาษาไทยและภาษาอังกฤษที่มีคำลักษณนามรวมอยู่ด้วยดีขึ้น และถูกต้องมากขึ้น ซึ่งจากการทดสอบพบว่าวิธีการที่นำเสนอในงานวิจัยฉบับนี้ให้ผลการจับคู่คำที่ถูกต้องมากกว่าผลการจับคู่คำที่ได้จาก GIZA สูงขึ้นประมาณ 77%

เมื่อนำผลการจับคู่คำที่ได้จากวิธีการที่งานวิจัยนี้นำเสนอแทรกเข้าไปแทนผลการจับคู่คำที่ได้จาก GIZA ในระหว่างขั้นตอนการแปลภาษาด้วยเครื่องอัตโนมัติด้วยโปรแกรม Moses ซึ่งผู้วิจัยขอเรียกว่า การแฮ็ก (Hack) เข้าไประหว่างการทำงานของ Moses เพื่อให้โปรแกรม Moses นั้นใช้ผลจากการจับคู่คำที่ได้จากวิธีการที่นำเสนอของงานวิจัยฉบับนี้ไปใช้ในการแปล โดยจากการทดสอบพบว่าผลการแปลที่ได้จากผลการจับคู่คำจากวิธีการที่งานวิจัยฉบับนี้นำเสนอ สามารถแปลได้ถูกต้องมากกว่าผลการแปลที่ได้จากผลการจับคู่คำที่ได้จาก GIZA ซึ่งในแต่ละรูปแบบของประโยคที่มีคำลักษณนามนั้นสามารถแปลได้ถูกต้องมากขึ้น และเมื่อนำประโยคที่มีคำลักษณนามทุกรูปแบบมารวมกันแล้วแปล ผลปรากฏว่าสามารถแปลประโยคที่มีคำลักษณนามได้ถูกต้องมากขึ้นกว่าระบบเดิม โดยมีค่า BLEU Score สูงขึ้น 3.48 คะแนน

วิธีการที่นำเสนอในงานวิจัยฉบับนี้สามารถจัดการกับความกำกวมที่เกิดขึ้นจากประโยคที่มีคำลักษณนามของภาษาไทยในกระบวนการจับคู่คำสองภาษาสำหรับการแปลภาษาด้วยเครื่องเชิงสถิติของคู่ภาษาไทย และภาษาอังกฤษเท่านั้น โดยที่วลีที่มีลักษณนามเป็นส่วนที่ผู้วิจัยมุ่งเน้นให้ความสนใจในงานวิจัยฉบับนี้ ซึ่งยังไม่รวมถึงการจัดการกับความกำกวมที่เกิดขึ้นจากประโยครูปแบบอื่นๆ

5.3 ข้อเสนอแนะและแนวทางวิจัยต่อไปในอนาคต

จากผลการทดลองที่ผ่านมาแสดงให้เห็นว่า วิธีการที่นำเสนอในงานวิจัยนี้สามารถขจัดความกำกวมที่เกิดจากคำลักษณนามของภาษาไทย ในกระบวนการจับคู่คำสองภาษาได้อย่างมีประสิทธิภาพ โดยแนวทางสำหรับการศึกษาวิจัยในอนาคต มีดังนี้

1. การตัดคำ โดยในงานวิจัยนี้ได้ใช้โปรแกรมตัดคำ SWATH ซึ่งในบางประโยคยังตัดคำได้ไม่ถูกต้อง ตัวอย่างเช่น “ สตรอเบอรี่ ” เมื่อใช้ SWATH ทำการตัดคำ จะได้
   “ สตรอ@NCMN | เบอรี่@NPRP ” ทำให้คำนามหลักถูกแบ่งออกเป็นสองคำ และส่งผลถึงการจับคู่คำระหว่างสองภาษาด้วย โดยถ้าหากมีการปรับปรุงการตัดคำ หรือเปลี่ยนไปใช้เทคนิคการรู้คำ (Word Recognition) ด้วยการเรียนรู้ด้วยเครื่อง (Machine Learning) อาจจะทำให้มีประสิทธิภาพขึ้นได้

2. ในงานวิจัยนี้ เมื่อนำคำลักษณนามกลับคืนเข้าไปในประโยค คำลักษณนามนั้นๆ จะถูกบังคับให้จับคู่เข้ากับคำนามหลักที่ใกล้ที่สุดซึ่งทำให้การจับคู่นั้นเกิดความผิดพลาด ดังนั้นถ้ามีการเพิ่มรายการของคำลักษณนามที่เป็นไปได้สำหรับคำนามคำนั้นๆ เข้าไปในระบบน่าจะทำให้การจับคู่คำในประโยคที่มีคำลักษณนามมีความถูกต้องเพิ่มมากขึ้น

3. หากมีการนำวิธีการที่นำเสนอไปประยุกต์ใช้ เพื่อเป็นตัวช่วยในกระบวนการแปลภาษาสำหรับการแปลภาษาด้วยเครื่องเชิงสถิติ น่าจะทำให้การแปลภาษามีประสิทธิภาพที่ดีมากขึ้น

รายการอ้างอิง

Book

Koehn, P. (2015). Moses: Statistical Machine Translation System User Manual and Code Guide.

Conferences

Singhapreecha, P. (2001). Thai Classifiers and The Structure of Complex Thai Nominals. Proceedings of the 15th Pacific Asia Conference on Language Information and Computation, pp.259-270.

Sornlertlamvanich, V., Pantachat, W., & Meknavin, S. (1994). Classifier Assignment by Corpus-Based Approach. Proceedings of the 15th Conference on Computational Linguistics, PA, USA.

Netjinda, N., Facundes, N., & Sirinaovakul, B. (2009). Toward Statistical Machine Translation for Thai and English. Proceeding of International Science and Technology Conference :ISTEC.

Koehn, P., Hoang, H., Birch, A., Burch, C. C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of The Association for Computational Linguistics(ACL), demonstration session, Prague, Czech Republic.

Bond, F., Ogura, K., & Ikehara, S. (1996). Classifiers in Japanese-to-English Machine Translation. Proceeding of the 16th Conference on Computational Linquistics, pp.125-130, PA, USA.

Paul, M., Sumita, E., & Yamamoto, S. (2002). Corpus-Based Generation of Numeral Classifier Using Phrase Alignment. Proceeding of the 19th International Conference on Computational Linguistics, pp.1-7, PA, USA.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. Proceeding of the 40th Annual Meeting on Association for Computational Linguistics, PA, USA, pp.311-318.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-Based Translation. Proceeding of the 2003 Conference of The North American Chapter of The Association for Computational Linguistics on Human Language Technology-Volume 1, pp.48-54.

Madnani, N. (2011). iBLEU: Interactively Debugging and Scoring Statistical Machine Translation Systems. Proceeding of the 5th IEEE International Conference on Semantic Computing, Palo Alto, CA, pp.213-214.

Journals

Och, F. J., & Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. Journal of Computational Linguistics, pp. 417-449.

Electronic Media

Paisarn Charoenpornsawat. (2003). SWATH (Smart Word Analysis for THai), Retrieved from: http://www.cs.cmu.edu/~paisarn/software.html

Och, F. J., & Ney, H. (2000). GIZA++: Training of Statistical Translation Models. Internal Report, RWTH Aachen University, Retrieved from: http://www.statmt.org/moses/giza/GIZA++.html

Koehn, P., Hoang, H., Birch, A., Burch, C. C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2008). Moses: Open Source Toolkit for Machine Translation. Proceeding of the 45th Annual 1618-1621, Brisbane, Australia, Retrieved from: http://www.statmt.org/moses/

Google Translation, Retrieved from: https://translate.google.co.th/

Websites

Knight, K., & Koehn, P. (2003). What's New in Statistical Machine Translation. Information Sciences Institute University of Southern California, Tutorial Retrieved from:http://people.csail.mit.edu/koehn/publications/tutorial2003.pdf

Casacub, F., & Vidal, E. (2007). System and Tools for Machine Translation GIZA++: Training of Statistical Translation Models. 4 June 2007, Retrieved from: https://www.prhlt.upv.es/~evidal/students/master/sht/transp/giza2p.pdf

ประวัติผู้เขียน

ชื่อ	นายภูวเมศฐ์ พิมลปัญญารัตน์
วันเดือนปีเกิด	11 สิงหาคม พ.ศ.2532
วุฒิการศึกษา	วิทยาศาสตรบัณฑิต สาขา วิทยาการคอมพิวเตอร์
(ระดับปริญญาตรี)	คณะวิทยาศาสตร์และเทคโนโลยี
	มหาวิทยาลัยธรรมศาสตร์
	ปีการศึกษา 2553

ผลงานทางวิชาการ

Puwamed Pimonpanyarad, Kanyalag Phodong, Taneth Ruangrajitpakorn and Rachada Kongkachandra. (2015). Thai Classifier Disambiguation in Bilingual Alignment Process for Thai-English SMT. In Proceeding of 2015 International Symposium on Multimedia and Communication Technology. September 23-25, 2015, pp.247-250.

ประสบการณ์ทำงาน	พ.ศ. 2556 ตำแหน่ง Programmer
	ธนาคารกรุงเทพ