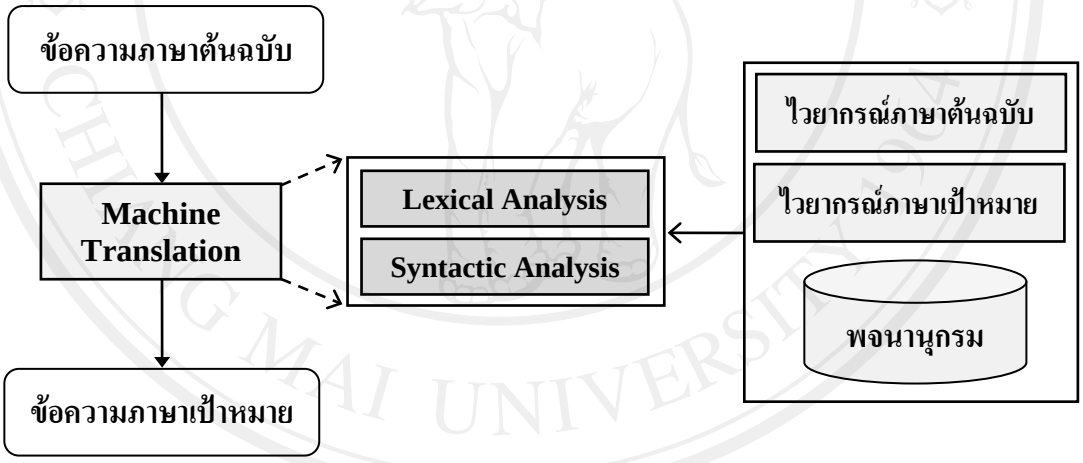
บทที่ 2
การแปลภาษาด้วยเครื่อง

การแปลภาษาด้วยเครื่อง คือ การนำระบบอัตโนมัติที่ทำงานบนเครื่องคอมพิวเตอร์มาใช้ในการแปลข้อความจากภาษา หนึ่งไปเป็นอีกภาษาหนึ่ง โดยเมื่อป้อนข้อ ความภาษาต้นฉบับเข้าไปในระบบ ระบบ จะทำการวิเคราะห์ข้อมูลภาษาต้นฉบับ เลือกคำแปล และสร้างข้อความของภาษาเป้าหมายออกมา

2.1 โครงสร้างของการแปลภาษาด้วยเครื่อง

โครงสร้างของการแปลภาษาด้วยเครื่องประกอบด้วย 2 ส่วนที่สำคัญ ดังรูปที่ 2.1



รูปที่ 2.1 โครงสร้างการแปลภาษาด้วยเครื่อง

2.1.1 การวิเคราะห์เกี่ยวกับคำศัพท์ (Lexical analysis)

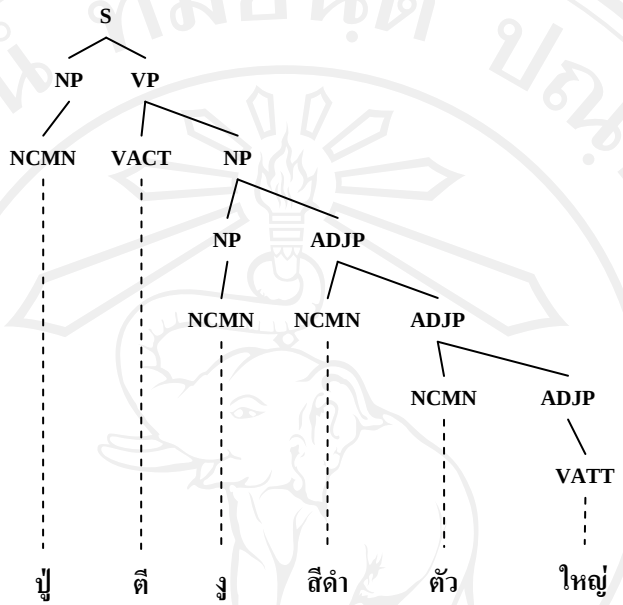
การวิเคราะห์เกี่ยวกับคำศัพท์สามารถประกอบด้วยการทำงานกี่ขั้นตอนก็ได้ แต่จุดประสงค์หลัก คือ การวิเคราะห์คำแล้วสามารถจำแนกและกำหนดประเภทของคำได้ จากนั้นสามารถนำค่าต่างๆ ที่กำหนดไว้ ไปใช้ในขั้นตอนต่อไปได้อย่างครบถ้วน

2.1.2 การวิเคราะห์รูปแบบกฎเกณฑ์ (Syntactic analysis)

ขั้นตอนนี้จะนำ สัญลักษณ์ที่แสดงประเภทของคำมาตรวจสอบว่าประโยคที่รับเข้ามาตรงตามกฎไวยากรณ์ที่มีอยู่หรือไม่ เราเรียกการตรวจสอบนี้ว่า การแจงประโยค (Parsing) และสามารถแบ่งรูปแบบในการแจงประโยคออกเป็น 2 ชนิด คือ

การแจงประโยคแบบบนลงล่าง (Top-down parsing)

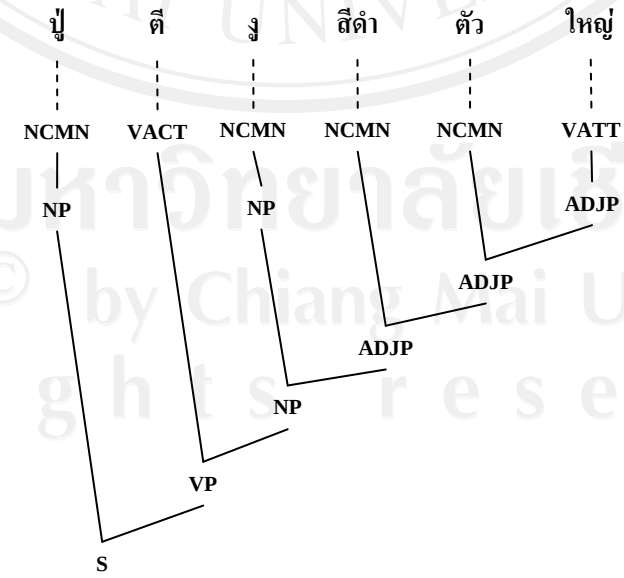
การแจงประโยคแบบบนลงล่างนี้เป็นการแจงประโยคตาม โครงสร้างข้อมูลแบบลำดับชั้น โดยเริ่มจากราก (Root node) และเรียงลำดับแบบ แวะผ่านก่อนลำดับ (เป็นวิธีการท่องไปในโครงสร้างข้อมูลแบบ Root→Left→Right)



รูปที่ 2.2 การแจงประโยคแบบบนลงล่าง (Top-down parsing)

การแจงประโยคแบบล่างขึ้นบน (Bottom-up parsing)

เป็นการแจงประโยคตาม โครงสร้างข้อมูลแบบลำดับชั้น โดยจะเริ่มต้นจากใบ (Leaf node) ซึ่งพิจารณาจากขวาไปซ้าย



รูปที่ 2.3 การแจงประโยคแบบล่างขึ้นบน (Bottom-up parsing)

สำหรับประเภทของคำในภาษาไทยในงานวิจัยนี้ จะใช้ข้อมูลจาก Orchid Corpus ซึ่งเป็นคลังคำศัพท์ภาษาไทย ที่เกิดจากการรวบรวมบทความและเอกสารทางวิชาการ หรือบทความและเอกสารที่เชื่อถือได้ แล้วนำมาตัดคำและใส่หน้าที่ของคำแต่ละคำในประโยค

ตารางที่ 2.1 POS ทั้งหมดใน Orchid Corpus

หมายเลข	POS	คำอธิบาย	ตัวอย่าง
1	NPRP	Proper noun	วินโดวส์ 95, โคโรน่า, โค้ก, พระอาทิตย์
2	NCNM	Cardinal number	หนึ่ง, สอง, สาม, 1, 2, 3
3	NONM	Ordinal number	ที่หนึ่ง, ที่สอง, ที่สาม, ที่1, ที่2, ที่3
4	NLBL	Label noun	1, 2, 3, 4, ก, ข, a, b
5	NCMN	Common noun	หนังสือ, อาหาร, อาคาร, คน
6	NTTL	Title noun	ดร., พลเอก
7	PPRS	Personal pronoun	คุณ, เขา, ฉัน
8	PDMN	Demonstrative pronoun	นี่, นั่น, ที่นั่น, ที่นี่
9	PNTR	Interrogative pronoun	ใคร, อะไร, อย่างไร
10	PREL	Relative pronoun	ที่, ซึ่ง, อัน, ผู้
11	VACT	Active verb	ทำงาน, ร้องเพลง, กิน
12	VSTA	Stative verb	เห็น, รู้, คือ
13	VATT	Attributive verb	อ้วน, ดี, สวย
14	XVBM	Pre-verb auxiliary, before negator "ไม่"	เกิด, เกือบ, กำลัง
15	XVAM	Pre-verb auxiliary, after negator "ไม่"	ค่อย, น่า, ได้
16	XVMM	Pre-verb, before or after negator "ไม่"	ควร, เคย, ต้อง
17	XVBB	Pre-verb auxiliary, in imperative mood	กรุณา, จง, เชิญ, อย่า, ห้าม
18	XVAE	Post-verb auxiliary	ไป, มา, ขึ้น
19	DDAN	Definite determiner, after noun without classifier in between	นี่, นั่น, โน่น, ทั้งหมด
20	DDAC	Definite determiner, allowing classifier in between	นี้, นั้น, โน้น, นู้น
21	DDBQ	Definite determiner, between noun and classifier or preceding quantitative expression	ทั้ง, อีก, เพียง
22	DDAQ	Definite determiner, following quantitative expression	พอดี, ถ้วน
23	DIAC	Indefinite determiner, following noun; allowing classifier in between	ไหน, อื่น, ต่างๆ

หมายเลข	POS	คำอธิบาย	ตัวอย่าง
24	DIBQ	Indefinite determiner, between noun and classifier or preceding quantitative expression	บาง, ประมาณ, เกือบ
25	DIAQ	Indefinite determiner, following quantitative expression	กว่า, เศษ
26	DCNM	Determiner, cardinal number expression	หนึ่งคน, เสือ 2 ตัว
27	DONM	Determiner, ordinal number expression	ที่หนึ่ง, ที่สอง, ที่สุดท้าย
28	ADVN	Adverb with normal form	เก่ง, เร็ว, ช้า, สม่ำเสมอ
29	ADVI	Adverb with iterative form	เร็วๆ, เสมอๆ, ช้าๆ
30	ADVP	Adverb with prefixed form	โดยเร็ว
31	ADVS	Sentential adverb	โดยปกติ, ธรรมดา
32	CNIT	Unit classifier	ตัว, คน, เล่ม
33	CLTV	Collective classifier	คู่, กลุ่ม, ฝูง, เชิง, ทาง, ด้าน, แบบ, รุ่น
34	CMTR	Measurement classifier	กิโลกรัม, แก้ว, ชั่วโมง
35	CFQC	Frequency classifier	ครั้ง, เที่ยว
36	CVBL	Verbal classifier	ม้วน, มัด
37	JCRG	Coordinating conjunction	และ, หรือ, แต่
38	JCMP	Comparative conjunction	กว่า, เหมือนกับ, เท่ากับ
39	JSBR	Subordinating conjunction	เพราะว่า , เนื่องจาก , ที่, แม้ว่า, ถ้า
40	RPRE	Preposition	จาก, ละ, ของ, ใต้, บน
41	INT	Interjection	โอ้ย,โอ้, เออ, เฮ้, อ้อ
42	FIXN	Nominal prefix	การทำงาน, ความสนุกสนาน
43	FIXV	Adverbial prefix	อย่างเร็ว
44	EAFF	Ending for affirmative sentence	จ๊ะ, จ้ะ, ค่ะ, ครับ, นะ, น่า, เถอะ
45	EITT	Ending for interrogative sentence	หรือ, เหรอ, ไหม, มั้ย
46	NEG	Negator	ไม่, มิได้, ไม่ได้, มิ
47	PUNC	Punctuation	(,), ", ,, ;

2.2 ทฤษฎีการแปลภาษาด้วยเครื่อง [16]

การแปลภาษาด้วยเครื่องสามารถแบ่งตามลักษณะการทำงานของระบบออกเป็น 3 กลุ่ม คือ

2.2.1 การแปลภาษาด้วยเครื่องแบบใช้ฐานกฎ (Rule-based Machine Translation)

เป็นการแปลภาษาด้วยเครื่องโดยใช้ความรู้ทางด้านภาษาศาสตร์มาใช้ในการกำหนดกฎเกณฑ์ของระบบ ซึ่งจะต้องมีการแยกคุณลักษณะและข้อมูลทางภาษาศาสตร์ของภาษาต้นฉบับ จากนั้นจึงทำการวิเคราะห์ตามกฎไวยากรณ์ของภาษาต้นฉบับ แล้วส่งผ่านข้อมูลที่ได้จากการวิเคราะห์นั้นไปยังกระบวนการวิเคราะห์คุณลักษณะและข้อมูลทางภาษาศาสตร์ของภาษาเป้าหมาย แล้วจึงทำการแปลจากพจนานุกรมคู่ภาษาและสร้างรูปประโยคของภาษาเป้าหมายออกมา

ข้อจำกัดของการแปลภาษาด้วยเครื่องกลุ่มนี้คือ ผู้พัฒนาต้องมีความรู้ทางด้านภาษาศาสตร์ของทั้งสองภาษาเป็นอย่างมาก จึงจะทำให้ได้การแปลที่มีประสิทธิภาพ

2.2.2 การแปลภาษาด้วยเครื่องแบบใช้สถิติ (Statistical Machine Translation)

เป็นการแปลภาษาด้วยเครื่องโดยใช้วิธีการทางสถิติมาช่วนในการแปล ซึ่งการแปลกลุ่มนี้จำเป็นต้องมีฐานข้อมูลคู่ภาษาที่มีการจับคู่ประโยค เพื่อเป็นฐานความรู้ให้ระบบทำการเรียนรู้ และใช้ค่าทางสถิติ เอ็นแกรม (N-Gram Model) ซึ่งเป็นการคำนวณค่าของการที่คำเกิดขึ้นร่วมกัน ถ้าชุดคำชุดใดมีค่าเอ็นแกรมหรือค่าความน่าจะเป็น (Probability) สูง แสดงว่าชุดคำนี้มีโอกาสเกิดขึ้นร่วมกันบ่อยครั้ง การคำนวณหาค่าเอ็นแกรมของชุดคำที่มีอยู่ในฐานข้อมูลคู่ภาษา ทำให้ได้ค่าความน่าจะเป็นของชุดคำต่างๆ เพื่อนำไปเปรียบเทียบและใช้ในการแปลข้อความได้ โดยสามารถเลือกใช้ค่าเอ็นแกรมได้ตั้งแต่ 2 คำ (bigrams), 3 คำ (trigrams) เป็นต้น ยิ่งใช้ค่าเอ็นแกรมมาก ยิ่งทำให้ภาษาต้นฉบับมีการได้หลากหลายและมีความละเอียดมากขึ้น

ข้อดีของวิธีการนี้คือ ผู้พัฒนาไม่จำเป็นต้องมีความรู้ทางด้านภาษาศาสตร์ของทั้งสองภาษา จึงทำให้ไม่เกิดปัญหาเรื่องไวยากรณ์ต่างๆ แต่การแปลด้วยวิธีการนี้จำเป็นต้องมีฐานข้อมูลคู่ภาษาที่มีจำนวนข้อมูลมหาศาลเพื่อหาค่าทางสถิติที่จะนำไปใช้เปรียบเทียบกับประโยคที่จะนำมาแปลได้อย่างครอบคลุม

2.2.3 การแปลภาษาด้วยเครื่องแบบอ้างอิงตัวอย่าง (Example-based Machine Translation)

เป็นการแปลภาษาด้วยเครื่องโดยมีฐานข้อมูลคู่ภาษาเป็นองค์ประกอบสำคัญ ซึ่งทำหน้าที่เก็บคำและประโยคตัวอย่างของคู่ภาษาที่ใช้จริงในชีวิตประจำวัน เพื่อนำไปคำนวณหาว่าประโยคที่รับเข้ามานั้น ควรทำการแปลออกมาเป็นประโยคในรูปแบบใด ระบบการทำงานของวิธีการนี้

แบ่งเป็น 2 ส่วนหลัก คือ การสร้างต้นแบบการแปล และการรวมประโยคคำแปลใหม่ โดยในส่วนแรกจะเป็นการจับคู่คำและประโยคตัวอย่างในฐานข้อมูลคู่ภาษา สร้างเป็นต้นแบบการแปลพื้นฐานเพื่อนำไปใช้เปรียบเทียบกับประโยคที่รับเข้ามา จากนั้นในส่วนที่ 2 จะทำการรวบรวมคำแปลที่ได้จากการเปรียบเทียบ มาสร้างเป็นประโยคผลลัพธ์

ข้อดีของวิธีการนี้คือ ไม่ใช้กฎไวยากรณ์และพจนานุกรมคู่ภาษา จึงไม่เกิดปัญหาเรื่องไวยากรณ์ต่างๆ อีกทั้งยังช่วยประหยัดเวลาในการแก้ไขปรับปรุงระบบ เพราะใช้เวลาในการพัฒนาระบบที่จะทำงานเชื่อมต่อกับฐานข้อมูลคู่ภาษาเพียงครั้งเดียว ถ้าพัฒนาระบบให้ทำงานได้อย่างมีประสิทธิภาพแล้ว สามารถนำไปใช้กับฐานข้อมูลคู่ภาษาอื่นๆ ได้อีกด้วย แต่การแปลด้วยวิธีการนี้จำเป็นต้องมีฐานข้อมูลคู่ภาษาที่มีจำนวนคำและประโยคตัวอย่างเป็นจำนวนมาก และต้องทำการจับคู่คำและประโยคตัวอย่างให้มีความถูกต้องแม่นยำ เพื่อให้ได้การแปลที่มีประสิทธิภาพ

ในงานวิจัยนี้จะเลือกใช้การแปลภาษาด้วยเครื่องแบบใช้ฐานกฎ ซึ่งเป็นวิธีเริ่มต้นที่มักจะใช้ในการแปลคู่ภาษาใหม่ๆ นอกจากนั้นการแปลภาษาด้วยเครื่องแบบใช้ฐานกฎไม่จำเป็นต้องใช้ฐานข้อมูลคู่ภาษาขนาดใหญ่ ซึ่งทำให้ง่ายต่อการพัฒนาเพื่อแปลประโยคพื้นฐานของคู่ภาษานั้นๆ