บทที่2

ความรู้ทั่วไปของระบบทาเปสตรี

2.1 ส่วนประกอบของระบบทาเปสตรี

ระบบทาเปสตรีเป็นเครื่องมือ (Tools) ที่ใช้พัฒนางานทางด้านการประมวลผลภาษาธรรมชาติ (Natural language processing) [3] พัฒนาขึ้นที่สถาบันวิทยาศาสตร์ระบบ (Institute of System Science) ประเทศสิงคโปร์ ระบบนี้ออกแบบมาเพื่อพัฒนาโปรแกรมประยุกต์ทางด้านการประมวลผลภาษาธรรมชาติ ได้แก่ งานด้านการแบ่งแยกข้อความเป็นส่วน ๆ (Text categorisation) การวางเส้นทางเดินเอกสาร (Message routing) การสรุปใจความสำคัญของข้อความ (Text summarization) การแยกสารสนเทศออกจากข้อมูล (Information extraction) รวมถึงการแปลภาษาด้วยเครื่อง (Machine translation)

ระบบทาเปสตรีสร้างขึ้นโดย Dr. Tong Loong Cheong เขียนด้วยภาษาลิปส์ (Lisp) ที่เป็นเสมือนคอมไพเลอร์ (Compiler) ที่เป็นตัวเรียกพจนานุกรมและกฎไวยากรณ์ต่าง ๆ มาใช้ในการทำงาน ซึ่งประกอบด้วยโมดูล (Module) 8 โมดูล คือ

1. โมดูลจัดการข้อมูลก่อนการประมวลผลให้อยู่ในรูปภาษาที่ต้องการ (Text Document Specification Language, TDSL) ทำหน้าที่ก่อนการประมวลผลข้อความสำหรับการแบ่งข้อความออกเป็นเขตข้อมูล (field) ต่าง ๆ เช่น ในใบรับฝากข้อความ (Memo) สามารถแยกได้เป็นส่วนชื่อผู้ส่ง (Sender) ส่วนชื่อผู้รับ (Receiver) ส่วนหัวเรื่อง (Subject) และส่วนเนื้อความ (Body) เป็นต้น

2. โมดูลแปลงข้อมูลให้อยู่ในลำดับของหน่วยคำ (Morphological Analyser, MORPHO) ทำหน้าที่วิเคราะห์ลักษณะของคำซึ่งจะแปลงรูปข้อความ (text) ให้อยู่ในรูปลำดับของหน่วยคำ

3. โมดูลตรวจสอบรูปแบบของคำหรือวลี (Lexical Pattern Matcher, LEXPAM) ทำหน้าที่ตรวจสอบคำที่เป็นสำนวน คำเฉพาะ หรือวลีที่ได้มาจากการแปลงข้อมูลนั้น

4. โมดูลโครงสร้างไวยากรณ์ภาษา (Tree Grammar Language, TGL) ทำหน้าที่จัดข้อมูลให้อยู่ในรูปโครงสร้างต้นไม้โดยอาศัยกฎไวยากรณ์ของภาษาต้นแบบ แล้วเปลี่ยนโครงสร้างต้นไม้ของภาษาต้นแบบให้เป็นโครงสร้างต้นไม้ของภาษาเป้าหมาย (Tree-to-tree transducer)

5. โมดูลตรวจสอบหน่วยคำกับพจนานุกรม (Lexical Lookup Dictionary Processer, LEXL) ทำหน้าที่แปลความหมายของคำจากภาษาหนึ่งไปเป็นอีกภาษาหนึ่งโดยใช้พจนานุกรม 2 ภาษา

6. โมดูลสังเคราะห์ประโยค (Text Generator, TEXTGEN) ทำหน้าที่สร้างประโยคโดยการเรียงคำให้เป็นประโยคของภาษาเป้าหมาย

7. โมดูลประมวลผลกราฟ (Graph Manipulation Language, GML) ทำหน้าที่ประมวลผลโครงสร้างของกราฟ เพื่อจัดการเกี่ยวกับความถูกต้องในเชิงไวยากรณ์และความหมายของประโยค

8. โมดูลแยกสารสนเทศออกจากข้อความ (Frame for Extracting Information from Messages, FEIM) ทำงานร่วมกับโมดูลประมวลผลกราฟเพื่อใช้หาโครงสร้างของข้อความที่สำคัญและแยกออกมาจากข้อความนั้น

โปรแกรมประยุกต์ทางด้านการประมวลผลภาษาธรรมชาติ อาจจะไม่ได้ใช้โมดูลทั้งหมดทั้ง 8 โมดูลในการทำงาน จะใช้โมดูลเพียงบางส่วนเท่านั้น เช่น งานด้านการแบ่งแยกข้อความเป็นส่วน ๆ (Text categorisation) ภายใต้ระบบทาเปสตรี จะใช้เพียง 3 โมดูล คือ

1. โมดูลจัดการข้อมูลก่อนการประมวลผลให้อยู่ในรูปภาษาที่ต้องการ (Text Document Specification Language, TDSL)
2. โมดูลแปลงข้อมูลให้อยู่ในลำดับของหน่วยคำ (Morphological Analyser, MORPHO)

3. โมดูลตรวจสอบรูปแบบของคำหรือวลี (Lexical Pattern Matcher, LEXPAM)

ในขณะที่ การแยกสารสนเทศออกจากข้อมูล (Information extraction) ภายใต้ระบบทาเปลตรี ใช้โมดูล 6 โมดูล คือ

1. โมดูลจัดการข้อมูลก่อนการประมวลผลให้อยู่ในรูปภาษาที่ต้องการ (Text Document Specification Language, TDSL)
2. โมดูลแปลงข้อมูลให้อยู่ในลำดับของหน่วยคำ (Morphological Analyser, MORPHO)
3. โมดูลตรวจสอบรูปแบบของคำหรือวลี (Lexical Pattern Matcher, LEXPAM)
4. โมดูลโครงสร้างไวยากรณ์ภาษา (Tree Grammar Language, TGL)
5. โมดูลประมวลผลกราฟ (Graph Manipulation Language, GML)
6. โมดูลแยกสารสนเทศออกจากข่อความ (Frame for Extracting Information from Messages, FEIM)

สำหรับการแปลภาษาด้วยเครื่องภายใต้ระบบทาเปลตรีนี้จะใช้โมดูลเพียง 5 โมดูลคือ

1. โมดูลแปลงข้อมูลให้อยู่ในลำดับของหน่วยคำ (Morphological Analyser, MORPHO)
2. โมดูลตรวจสอบรูปแบบของคำหรือวลี (Lexical Pattern Matcher, LEXPAM)
3. โมดูลโครงสร้างไวยากรณ์ภาษา (Tree Grammar Language, TGL)
4. โมดูลตรวจสอบหน่วยคำกับพจนานุกรม (Lexical Lookup Dictionary Processer, LEXL)
5. โมดูลสังเคราะห์ประโยค (Text Generator, TEXTGEN)

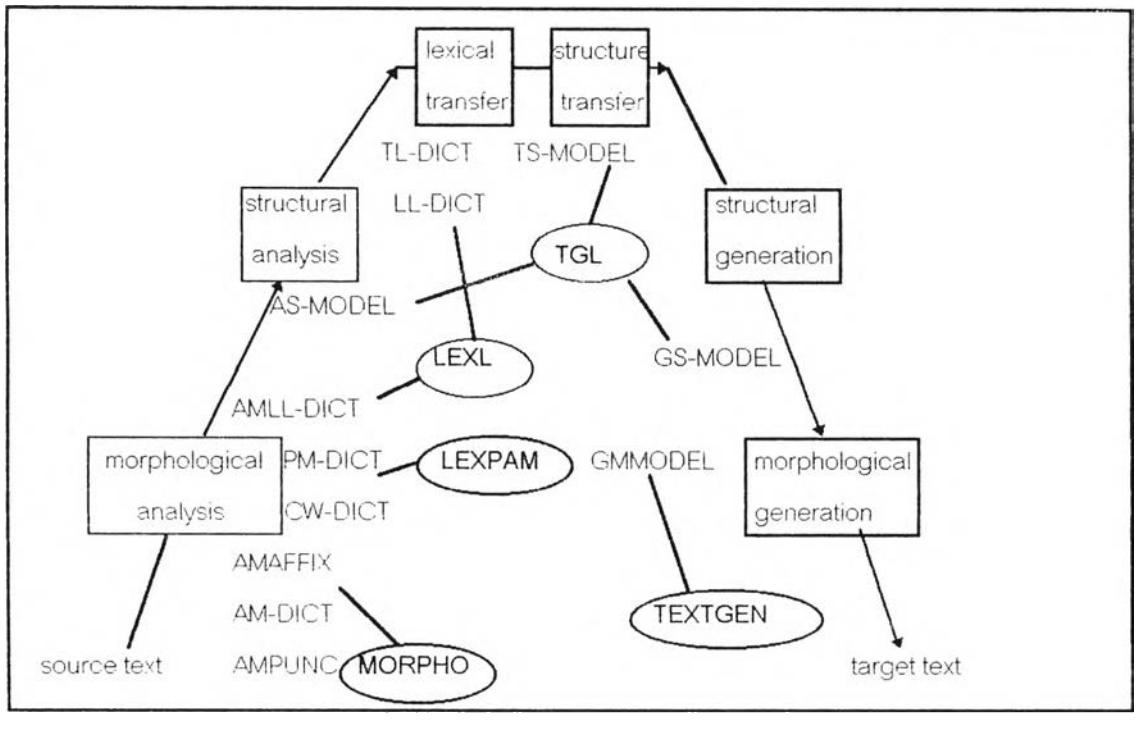
ตัวอย่างของระบบการแปลภาษาจากภาษาหนึ่งไปยังอีกภาษาหนึ่ง ที่พัฒนาภายใต้ระบบทาเปลตรี ได้แก่

1. เม็ตซ์ (METS) เป็นระบบแปลภาษามาเลย์เป็นภาษาอังกฤษ (Malay-English MT)
2. ซาเกะ (SAGE-MT) เป็นระบบแปลภาษาอังกฤษเป็นภาษาจีน (English-Chinese MT)
3. เจมาห์ (JEMAH) เป็นระบบแปลภาษาอังกฤษเป็นภาษามาเลย์ (English-Malay MT)

ระบบการแปลภาษาทั้ง 3 ระบบข้างบนใช้โมดูล 5 โมดูลข้างต้นของระบบทาเปสตรีในการพัฒนาเช่นเดียวกัน และในงานวิจัยนี้จะใช้ไวยากรณ์วิเคราะห์โครงสร้างภาษาอังกฤษและพจนานุกรมคำศัพท์ภาษาอังกฤษ จากระบบเจมาห์ มาเป็นต้นแบบในการวิจัย

2.2 ขั้นตอนการทำงานของการแปลภาษาด้วยเครื่องภายใต้ระบบทาเปสตรี

ขั้นตอนการทำงานของการแปลภาษาด้วยเครื่องภายใต้ระบบทาเปสตรีเป็นไปตามรูปที่ 3 และตาราง 2.1



รูปที่ 3 ไดอะแกรมขั้นตอนการทำงานของระบบการแปลภาษาด้วยเครื่อง

ขั้นตอนการทำงานของการแปลภาษาด้วยเครื่องภายใต้ระบบทาเปสตรีมี 6 ขั้นตอน คือ

1. ขั้นการวิเคราะห์ลักษณะของคำ (Morphological analysis)

เป็นขั้นตอนที่นำเอาภาษาต้นแบบมาวิเคราะห์คำในระดับต่าง ๆ เพื่อแยกความกำกวมของภาษาในระดับต่าง ๆ ให้ชัดเจน ขั้นตอนนี้ประกอบด้วย กฎไวยากรณ์ 2 กฎและพจนานุกรมคำศัพท์ 4 พจนานุกรม คือ

1.1 กฎไวยากรณ์ตรวจสอบเครื่องหมายวรรคตอน (AM-PUNC)

1.2 กฎไวยากรณ์ตรวจสอบอุปสรรคและปัจจัย (AMAFFIX)

1.3 พจนานุกรมเก็บคำศัพท์ที่เป็นคำพื้นฐาน (AM-DICT)

1.4 พจนานุกรมเก็บคำศัพท์ที่เป็นกลุ่มคำ (CW-DICT)

1.5 พจนานุกรมเก็บคำศัพท์ที่เป็นวลี (PM-DICT)

1.6 พจนานุกรมเก็บลักษณะเฉพาะของคำในระดับอรรถศาสตร์ (AMLL-SEM)

2. ขั้นการวิเคราะห์โครงสร้าง (Structural analysis)

เป็นขั้นตอนวิเคราะห์โครงสร้างของภาษาต้นแบบ เพื่อเปลี่ยนโครงสร้างของประโยคให้อยู่ในรูปโครงสร้างต้นไม้ โดยใช้หลักการกลไลการควบคุมที่ถูกกำหนดโดยกฎไวยากรณ์ (grammar rule) โดยการตรวจสอบโครงสร้างต้นไม้และความสัมพันธ์ของลักษณะทางภาษาที่ตรงกับเงื่อนไขที่กำหนดไว้ ถ้าตรง แล้วจะเปลี่ยนแปลงโครงสร้างต้นไม้เป็นโครงสร้างต้นไม้แบบใหม่และจะเปลี่ยนแปลงลักษณะทางภาษาหรือ กำหนดลักษณะทางภาษาเพิ่มเติม ถ้าไม่ตรง จะข้ามไปตรวจสอบกับอีกกฎไวยากรณ์อีกกฎหนึ่ง ขั้นตอนนี้ประกอบด้วย กฎไวยากรณ์ 1 กฎ คือ กฎไวยากรณ์วิเคราะห์โครงสร้าง (AS-MODEL)

3. ขั้นการถ่ายทอดหน่วยคำ (Lexical transfer)

การถ่ายทอดความหมายเป็นกระบวนการหลังจากการวิเคราะห์ การถ่ายทอดความหมายนี้จะใช้ข้อมูลของพจนานุกรมแบบสองภาษา เพื่อถ่ายทอดความหมายของคำของภาษาหนึ่งไปยังอีกภาษาหนึ่ง ขั้นตอนนี้ประกอบด้วย พจนานุกรมคำศัพท์ 2 พจนานุกรม คือ

3.1 พจนานุกรมถ่ายทอดความหมายของคำ (TL-DICT)

3.2 พจนานุกรมเก็บลักษณะอื่น ๆ ของคำ (LL-DICT)

4 ขั้นการถ่ายทอดโครงสร้าง (Structural transfer)

การถ่ายทอดโครงสร้างของภาษา เป็นกระบวนการถ่ายทอดโครงสร้างต้นไม้ของภาษาต้นแบบไปเป็นโครงสร้างต้นไม้ของภาษาเป้าหมายโดยอาศัยกฎไวยากรณ์โครงสร้างของ

ทั้งสองภาษา ขั้นตอนนี้ประกอบด้วย กฎไวยากรณ์ 1 กฎ คือ กฎไวยากรณ์ถ่ายทอดโครงสร้าง (TS-MODEL)

5. ขั้นการสังเคราะห์โครงสร้าง (Structural generation)

การสังเคราะห์โครงสร้างเป็นกระบวนการจัดรูปแบบโครงสร้างต้นไม้ให้เป็นไปตามรูปแบบของโครงสร้างประโยคของภาษาเป้าหมาย ขั้นตอนนี้ประกอบด้วย กฎไวยากรณ์ 1 กฎ คือ กฎไวยากรณ์สังเคราะห์โครงสร้าง (GS-MODEL)

6 ขั้นการสังเคราะห์ลักษณะของคำ (Morphological generation)

การสังเคราะห์หน่วยคำเป็นกระบวนการทำประโยคให้ถูกต้องสมบูรณ์โดยการเติมลักษณะพิเศษทางภาษาของภาษาเป้าหมาย ขั้นตอนนี้ประกอบด้วย กฎไวยากรณ์ 1 กฎ คือ กฎไวยากรณ์สังเคราะห์ลักษณะของคำ (GM-MODEL)

ตาราง 2.1 แสดงการเรียกใช้กฎไวยากรณ์ และพจนานุกรมคำศัพท์จากโมดูล 5 โมดูลของขั้นตอน 6 ขั้นตอน

MT phase	grammar / dictionary	module
1. AM - morphological analysis	AM-DICT	MORPHO
	AMAFFIX	MORPHO
	CW-DICT	LEXPAM
	PM-DICT	LEXPAM
	AMALL-DICT	LEXL
2. AS - structural analysis	AS-MODEL	TGL
3. TL - lexical transfer	TL-DICT	LEXL
	LL-DICT	LEXL
4. TS - structural transfer	TS-MODEL	TGL
5. GS - structural generation	GS-MODEL	TGL
6. GM - morphological generation	GM-MODEL	TEXTGEN

ในงานวิจัยนี้ จะทำเพียงขั้นตอนที่ 3 ถึงขั้นตอน 6 ส่วนขั้นตอนที่ 1 และขั้นตอนที่ 2 จะใช้ตามมาตรฐานเดิมที่ระบบเจมาร์ทำไว้