บทที่ 2
การแปลภาษาด้วยเครื่องคอมพิวเตอร์

การทำวิจัยและพัฒนาการแปลภาษาด้วยเครื่องนั้นเป็นงานแขนงหนึ่งในศาสตร์แห่งการประมวลผลภาษาธรรมชาติ (Natural Language Processing หรือ NLP) เพื่อให้เข้าใจถึงความเป็นมาของระบบแปลภาษาด้วยเครื่องแบบต่างๆ ในบทนี้จึงขออธิบายถึงระบบแปลภาษาด้วยเครื่องแบบต่างๆ รวมถึงข้อดีและข้อเสียของแต่ละระบบ ในท้ายบทจะอธิบายถึง “ภาษิต” ซึ่งเป็นระบบแปลภาษาด้วยเครื่องสำหรับภาษาอังกฤษ-ไทย ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

อย่างไรก็ดี ในปัจจุบันการแปลภาษาด้วยเครื่องนั้นไม่สามารถทำการแปลข้อความหรือบทความให้เกิดประโยคที่มีความไพเราะและสละสลวยได้เหมือนกับการแปลของของมนุษย์ทั่วไป เช่นการแปลกาพย์ โคลงกลอนต่างๆ เนื่องจากการแปลภาษาด้วยเครื่อง ยังมีข้อจำกัดในการแปลมาก ไม่สามารถที่จะทำให้เครื่องมีความรู้ได้เท่ากับมนุษย์จริงๆ ทำให้การแปลภาษาด้วยเครื่องไม่สามารถทดแทนนักแปลได้ แต่การแปลภาษาด้วยเครื่องนั้นจะสามารถช่วยแปลข้อความหรือบทความที่ไม่ต้องการความไพเราะและสละสลวยได้เป็นอย่างดี และจะช่วยทุ่นแรงนักแปลได้มาก

2.1 ระบบการแปลโดยตรง (Direct Machine Translation Strategy)

ระบบการแปลโดยตรงคือระบบการแปลที่ได้รับการออกแบบให้ใช้กับการแปลคู่ภาษาเฉพาะคู่ใดคู่หนึ่ง เป็นลักษณะการแปลแบบง่ายๆ ไม่มีการใช้ทฤษฎีทางภาษาศาสตร์หรือหลักการทางวจีวิภาค (POS : Parts-of-Speech) ระบบนี้จะขึ้นอยู่กับการพัฒนาพจนานุกรมที่สมบูรณ์ที่สุด การวิเคราะห์หน่วยคำ และโปรแกรมการผลิตข้อความเพื่อแปลความหมายแบบคำต่อคำ วลีต่อวลีจากภาษาต้นทางสู่ภาษาปลายทางอย่างสมเหตุสมผล

ระบบการแปลโดยตรงนั้นมีข้อเสียอยู่ที่เป็นระบบที่แปลภาษาได้ทีละคู่ของภาษา กล่าวคือ ถ้าต้องการแปลภาษาแต่ละคู่จะต้องมีการวิเคราะห์คู่ภาษานั้นๆ ทุกครั้ง ลักษณะการแปลแบบนี้ทำให้กระบวนการแปลมีความยุ่งยากมาก สมมติว่าถ้าต้องการแปลภาษา 10 คู่ ภาษาก็ต้องทำการวิเคราะห์และสังเคราะห์คู่ภาษานั้น 10 ครั้ง ทั้งที่บางครั้งภาษาต้นทางนั้น อาจเป็นภาษาเดียวกันก็ได้ ตัวอย่างของระบบนี้คือ ระบบซิสทราน (Systran) ใช้แปลเอกสารจากภาษารัสเซียเป็นภาษาอังกฤษ

2.2 ระบบการแปลแบบเปลี่ยน (Transfer Machine Translation Strategy)

เป็นระบบที่ถือว่าตัวแทนแสดงความหมายของไวยากรณ์ของภาษาต้นทางและภาษาปลายทางนั้นมีลักษณะที่แตกต่างกัน จะต้องมีช่วงเชื่อมต่อที่เทียบตัวแทนแสดงความหมายที่เป็นลักษณะเฉพาะของภาษาหนึ่ง เรียกว่าการเปลี่ยน จากนั้นก็สร้างภาษาปลายทางขึ้น ตัวอย่างเช่น ลักษณะประโยคภาษาอังกฤษ “It is a please to be here.” จะต้องเปลี่ยนเป็นประโยคโครงสร้างของภาษาไทยคือ “ยินดีที่ได้มาอยู่ที่นี่” ซึ่งถ้าเราใช้ระบบการแปลโดยตรงจะได้ประโยคว่า “มันเป็นความยินดีมาอยู่ที่นี่” [6] ระบบนี้มีขั้นตอนการทำงาน 3 ขั้นตอน คือ

2.2.1 การวิเคราะห์ภาษาต้นทาง (Source Language Analysis)

เป็นการวิเคราะห์ประโยคในภาษาต้นทางโดยใช้หลักไวยากรณ์โครงสร้างของภาษา และพจนานุกรมของภาษาต้นทาง (Source Language Dictionary)

2.2.2 การเปลี่ยน (Transfer)

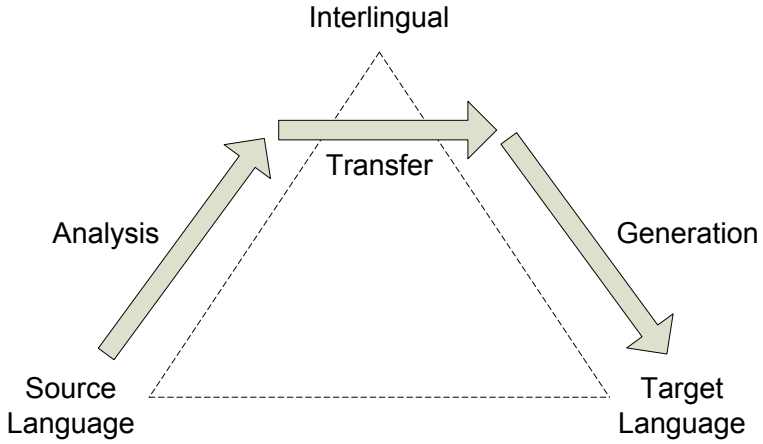
เป็นการเปลี่ยนคลังคำศัพท์ (Lexicon) และโครงสร้างของภาษาต้นทางให้สอดคล้องกับคลังศัพท์ และโครงสร้างของภาษาปลายทาง โดยมีการใช้พจนานุกรมทวิภาษา (Bilingual Transfer Dictionary)

2.2.3 การสร้างภาษาปลายทาง (Target Language Generation)

ในขั้นนี้จะใช้พจนานุกรมของภาษาปลายทาง (Target Language Dictionary) เพื่อเป็นการสร้างภาษาปลายทางให้มีคุณสมบัติทางโครงสร้าง และความหมายของภาษาปลายทางอย่างแท้จริง ซึ่งอาจจะรวมถึงขั้นตอนการจัดลำดับคำ ตัวอย่างของระบบนี้คือระบบ เกต้า (GETA) หรือระบบ ซูซี่ (SUSY)

2.3 ระบบการแปลภาษาแบบการใช้ภาษากลาง (Interlingual Machine Translation strategy)

เป็นระบบการแปลที่พัฒนาจากระบบการแปลแบบการเปลี่ยนเพื่อให้มีลักษณะเป็นสากลมากขึ้นจนเป็นตัวแทนภาษาที่เป็นอิสระ (Language Independent Representation) ระบบการแปลภาษาแบบนี้จะแบ่งเป็น 2 ด้านคือด้านการวิเคราะห์ (Analysis) เป็นการวิเคราะห์ภาษาต้นทางสู่ภาษาที่เราสร้างขึ้นใหม่เรียกว่าภาษากลาง (Interlingua) และด้านการก่อกำเนิด (Generation) ในขั้นตอนนี้เราจะสร้างภาษาปลายทางจากภาษากลาง ระบบนี้เป็นระบบหลากภาษา (Multilingual) ซึ่งแทนระบบทวิภาษาในระบบการแปลแบบเปลี่ยน (transfer)



ภาพที่ 2.1 ระบบการแปลภาษาแบบการใช้ภาษากลาง

ที่กล่าวถึงไปแล้วเป็นแนวทางสำหรับการแปลภาษาในอดีต สำหรับแนวทางการพัฒนาระบบเครื่องแปลภาษาในปัจจุบันนั้นจะสรุปการแปลภาษาเป็น 2 แนวทาง คือแนวทางที่หนึ่งเป็นการพัฒนาปรับปรุงระบบซึ่งอาศัยกฎไวยากรณ์ (Rule-Based MT) ซึ่งมีการค้นคว้าวิจัยมานานแล้วให้ดียิ่งขึ้น กับอีกแนวทางหนึ่งเป็นการพัฒนาระบบซึ่งอาศัยฐานคลังข้อความ (Corpus-Based MT) มาช่วยในการแปล

2.4 การใช้กฎไวยากรณ์ช่วยในการแปล (Rule-Based MT)

2.4.1 Transfer-Based MT

เป็นวิธีที่ถูกใช้ในโครงการแปลภาษาที่สำคัญๆ ในยุคแรกๆ จากที่กล่าวไปข้างต้น โดยวิธีนี้มองว่า กระบวนการแปลประกอบไปด้วย 3 ขั้นตอนคือ การวิเคราะห์ไปเป็นรูปแสดงแทนของภาษาต้นทาง (Abstract Source Language Representation), การย้ายข้าง (Transfer) ไปเป็นรูปแสดงแทนของภาษาปลายทาง (Abstract Target Language Representation) และการผลิตหรือสังเคราะห์ไปเป็นข้อความของภาษาปลายทางถึงแม้ว่าทั้งสองโครงการข้างต้นจะปิดฉากลงไปเรียบร้อยแล้ว แต่ก็ได้มีโครงการใหม่ซึ่งได้พัฒนาต่อไปอีก ซึ่งมีเป้าหมายจะพัฒนาระบบช่วยเหลือของ MT สำหรับนักแปลที่มีลักษณะเป็น User-friendly

2.4.2 Interlingua-Based MT

วิธีนี้ต่างจาก Transfer-Based MT โดยมองกระบวนการการแปลว่าประกอบด้วย 2 ขั้นตอนคือ การวิเคราะห์ไปเป็นรูปแสดงแทนซึ่งไม่ขึ้นกับภาษา และการผลิตจากรูปแสดงแทนนั้นไปเป็นข้อความของภาษาปลายทาง โครงการของญี่ปุ่นและอเมริกาจำนวนมากก็ได้ใช้วิธีนี้ ข้อดีของวิธีนี้คือความง่ายในการขยายเพิ่มประเภทของภาษาต้นทาง และภาษาปลายทางเข้ากับระบบการแปลเมื่อเทียบกับ Transfer-Based MT แต่การกำหนดภาษากลางให้ครอบคลุมทุก

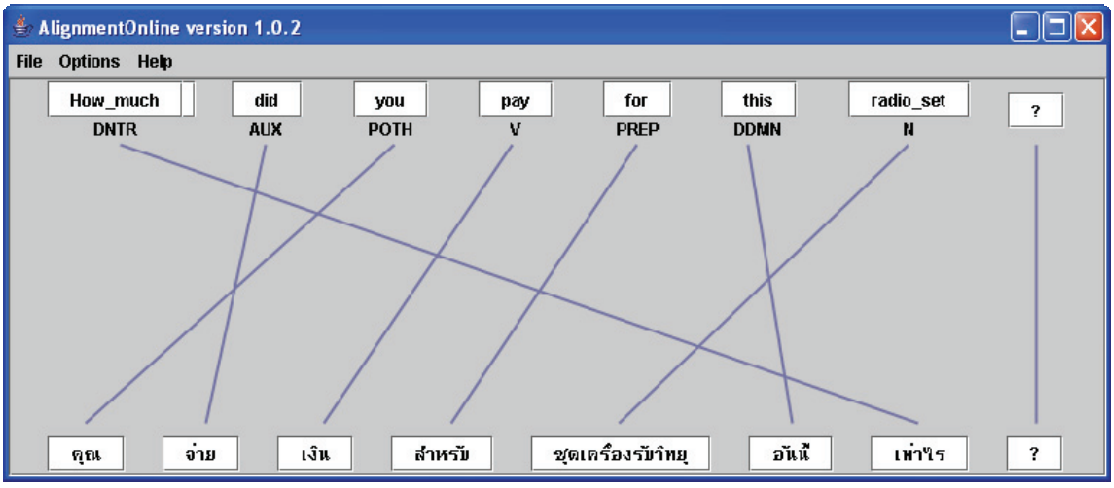
รูปแบบการใช้ภาษาของหลายๆ ภาษาก็เป็นปัญหาที่ยากสำหรับวิธีนี้ ในระยะหลังแนวโน้มของวิธีนี้จะเป็นการใช้ฐานความรู้เข้ามาช่วยในการแปล โดยมีความเชื่อว่าการแปลจะต้องใช้ความรู้มากกว่าความรู้ทางภาษาอย่างเดียว กล่าวคือ ต้องมีฐานความรู้เพื่อทำความเข้าใจ (Understanding) บริบทให้ได้ด้วย ซึ่งในส่วนนี้เองที่หากสามารถแก้ไขปัญหาพื้นฐานบางอย่าง เช่น การแก้ไขปัญหาความกำกวมของคำ (Word Sense Disambiguation) จะช่วยให้การแปลดีขึ้น

2.5 การใช้ฐานบทความช่วยในการแปล (Corpus-Based MT)

2.5.1 ระบบแปลภาษาด้วยเครื่องแบบอิงสถิติ (Statistical-Based MT)

การพัฒนาที่เด่นชัดที่สุดในช่วงปลายศตวรรษที่ 90 เห็นจะได้แก่ การนำวิธีการทางสถิติของ MT ในโครงการวิจัยของบริษัทไอบีเอ็ม ลักษณะที่สำคัญก็คือ การใช้วิธีการทางสถิติเพียงอย่างเดียวในการวิเคราะห์และการผลิต โดยทดลองกับ Corpus ขนาดใหญ่ของ The Canadian Hansard ซึ่งเป็นบันทึกการอภิปรายในสภาโดยจัดเก็บเป็นภาษาอังกฤษและฝรั่งเศส ผลการทดลองก่อให้เกิดการตื่นตัวในวงการระบบแปลภาษาด้วยเครื่องอย่างมาก เนื่องจากสามารถแปลได้ดีกว่าที่นักวิจัยทั่วไปคาดไว้มาก ทำให้นักวิจัยทางระบบแปลภาษาด้วยเครื่องทั้งหลายต้องย้อนกลับไปดูวิธีการทางสถิติ

ภาพที่ 2.2 แสดงตัวอย่างของคู่ประโยคแบบขนานที่มีการวางแนวซึ่งจะเห็นได้ว่าจำเป็นต้องใช้มนุษย์มากำกับ (tag) ความสัมพันธ์ระหว่างประโยคต้นทางและประโยคปลายทางเป็นเหตุให้การเตรียมคลังข้อความของวิธีการนี้เป็นไปอย่างยากลำบากและใช้เวลามาก



ภาพที่ 2.2 คู่ประโยคแบบขนานที่มีการวางแนว (Alignment Paralleled Sentence)

จุดเด่นของวิธีการนี้คือ ไม่มีการใช้กฎไวยากรณ์ ทำไม่เกิดปัญหาเชิงภาษาศาสตร์ อาทิเช่น ปัญหากฎไม่ครอบคลุม ปัญหาการเพิ่มกฎ ปัญหาการแจงประโยควายกสัมพันธ์ (Syntax parsing) และปัญหาการแปลสำนวน เป็นต้น

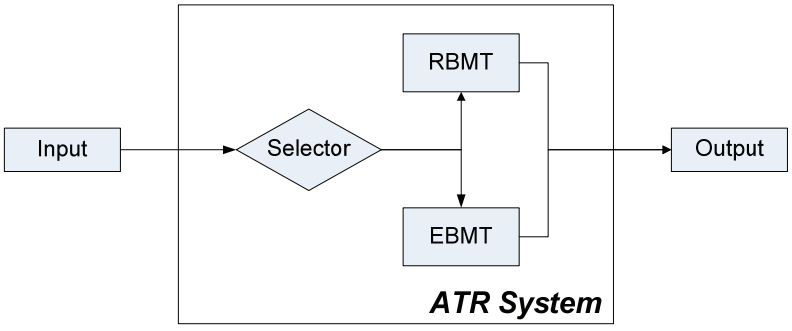
จุดด้อยของวิธีการนี้คือ จำเป็นต้องใช้คลังข้อความคู่ประโยคแบบขนานที่มีการวางแนว (Alignment Paralleled Corpus) จำนวนมากในการที่จะสร้างตัวแบบสถิติ (Statistical-Model)

2.5.2 ระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง (Example-Based MT)

แนวคิดของระบบการแปลแบบอิงตัวอย่าง (EBMT: Example-based MT) เป็นแนวคิดที่พยายามเลียนแบบพฤติกรรมการแปลภาษาของมนุษย์ ซึ่งแนวคิดพื้นฐานมาจากการศึกษาการแปลของคนเราซึ่งมักจะหาลักษณะประโยคคล้ายๆ กันที่เคยแปลมาก่อนมาเทียบในการแปลเป็นประโยคภาษาปลายทาง วิธีการนี้จะใช้คลังข้อความสองภาษา (Bi-Lingual Corpus) ของวลีหรือประโยค ตัวอย่างซึ่งได้จากคลังข้อความขนาดใหญ่

ในปี 1981 มีบุคคลแรกได้นำเสนอแนวคิดซึ่งถือได้ว่าเป็นต้นแบบของระบบแปลภาษาแบบอิงตัวอย่างคือ Nagao [7] ซึ่งเขาได้สังเกตกระบวนการแปลของมนุษย์ว่าเวลาเราเจอประโยคต้นทางใหม่ เราจะพยายามรู้จำความคล้าย (recognizing the similarity) ของประโยคต้นทางที่พบใหม่กับส่วนของประโยคที่เรารู้จักที่มีอยู่ในความทรงจำ (selecting identical phrases available in the translation memory) เว้นแต่ความคล้ายที่รู้จำนั้นเป็นความคล้ายระดับคำ (except for a similar content word) แม้ว่า ณ เวลานั้น คำว่า “การแปลแบบอิงตัวอย่าง” จะยังไม่ได้ถูกกำหนดขึ้นมา แต่แนวคิดของ Nagao ถือได้ว่าเป็นแนวคิดของการแปลแบบอิงตัวอย่าง ทำให้ Nagao ได้รับการยกย่องว่าเป็นผู้ให้กำเนิดการแปลภาษาด้วยแนวคิดนี้ในปี 1984

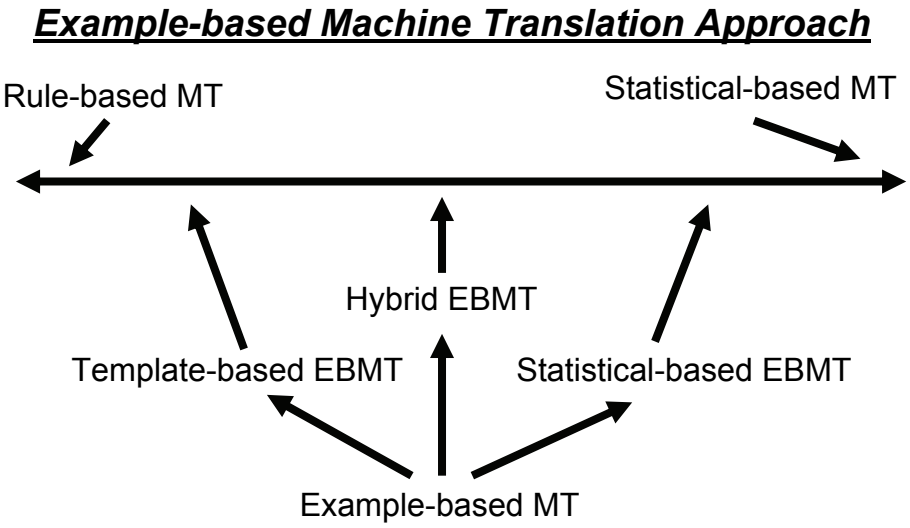
ในปี 1990 เริ่มมีความพยายามนำแนวคิดของ Nagao มาประยุกต์ใช้เพื่อเพิ่มประสิทธิภาพในการแปลของระบบแปลภาษาแบบอิงไวยากรณ์ (RBMT: Rule-based MT) โดยในงานของ Sato และ Nagao [8] จะใช้การแปลแบบอิงตัวอย่างเมื่อระบบการแปลแบบอิงกฎไวยากรณ์ไม่สามารถเข้าคู่การแจงประโยควายกสัมพันธ์ระหว่างภาษาต้นทางและภาษาปลายทางได้โดยตรง วิธีการนี้ถูกนำไปใช้ในระบบ ATR



ภาพที่ 2.3 ระบบ ATR

ระบบ ATR [9,10,11] เป็นระบบแปลภาษาญี่ปุ่น-อังกฤษด้วยเครื่องแบบอิงกฎไวยากรณ์ด้วยเสียง (Spoken Japanese-English MT) ระบบ ATR ถูกจัดว่าเป็นระบบแปลภาษาแบบลูกผสม (Hybrid MT)โดยในระบบจะมีกลไกตัวเลือก (Selector) เพื่อจัดการกับเงื่อนไขพิเศษแบบต่างๆ เพื่อส่งไปยังระบบแปลภาษาที่สามารถจัดการกับเงื่อนไขที่ดีที่สุด ปัญหาหลักของระบบนี้อยู่ที่ความสามารถในการจัดการเงื่อนไขของกลไกตัวเลือก

”แนวความคิดระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างเริ่มมีข้อสรุปที่ชัดเจนว่าแนวความคิดดังกล่าวเป็นแนวคิดที่อยู่กลางระหว่างระบบแปลภาษาด้วยเครื่องแบบอิงสถิติ (Statistical-based Machine Translation) และระบบแปลภาษาแบบอิงกฎไวยากรณ์ (Rule-based Machine Translation)” [12] ดังแสดงในภาพที่ 2.4



ภาพที่ 2.4 แนวความคิดระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่าง

เนื่องจากแนวความคิดระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างเป็นแนวคิดที่อยู่ระหว่างกลางระหว่างทั้งสองแนวคิด หากใช้วิธีที่มีความเอนเอียงไปทางระบบแปลภาษาแบบอิงกฎไวยากรณ์มากกว่า จะเรียกว่า ระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างโดยใช้แม่แบบการแปล (Template-based EBMT) หากใช้วิธีที่มีความเอนเอียงไปทางระบบแปลภาษาแบบอิงสถิติมากกว่า จะเรียกว่า ระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างโดยใช้ค่าสถิติ (Statistical-based EBMT) อย่างไรก็ดี หากมีการนำแนวความคิดของทั้งสองแนวทาง คือ ระบบแปลภาษาแบบอิงกฎไวยากรณ์ และระบบแปลภาษาแบบอิงสถิติมาใช้ร่วมกันเราจะเรียกว่า ระบบแปลภาษาด้วยเครื่องแบบอิงตัวอย่างแบบลูกผสม (Hybrid EBMT)

2.6 ระบบแปลภาษาอังกฤษ-ไทยภาษิต

ในปัจจุบันการแปลภาษาด้วยเครื่องคอมพิวเตอร์ที่สนับสนุนภาษาไทย มีการพัฒนามาแล้วหลายปี โดยที่เราเริ่มต้นในปี 2524 โดยทบวงมหาวิทยาลัยได้มีคำสั่งแต่งตั้งคณะอนุกรรมการโครงการวิจัยการแปลภาษาอังกฤษเป็นภาษาไทยด้วยเครื่องคอมพิวเตอร์ โดยใช้ชื่อว่าระบบอาเรียน (ARIANE) และหลังจากนั้นก็มีงานวิจัยออกมามากมาย จนมาถึงปัจจุบันระบบที่มีความถูกต้องมากระบบหนึ่งก็คือ โครงการพัฒนาระบบเครื่องแปลภาษาสำหรับภาษาในเอเชียซึ่งทางศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) [13,14] ได้ร่วมมือกับรัฐบาลญี่ปุ่น พร้อมด้วยนักวิจัยจากอีก 3 ประเทศคือ จีน อินโดนิเซีย และมาเลเซีย เรียกว่าภาษิต (Parsit) ซึ่งสามารถทดลองระบบการแปลได้ที่ http://www.suparsit.com/ โดยตัวภาษิตนั้นพัฒนามาจากระบบการแปลภาษาอังกฤษเป็นญี่ปุ่น (English-to-Japanese) โดยที่ไม่ได้ทำการแปลแบบคำต่อคำแต่ถ้าสามารถแปลแบบประโยคต่อประโยคได้โดยที่มีการใช้ฐานความรู้วากยสัมพันธ์ (syntax) และทางด้านอรรถศาสตร์ (semantics) การใช้กฎทางภาษาและรวมถึงการใช้พจนานุกรมด้วย โดยขั้นตอนพื้นฐานการทำงานนั้นแบ่งออกเป็น 2 ส่วน คือ

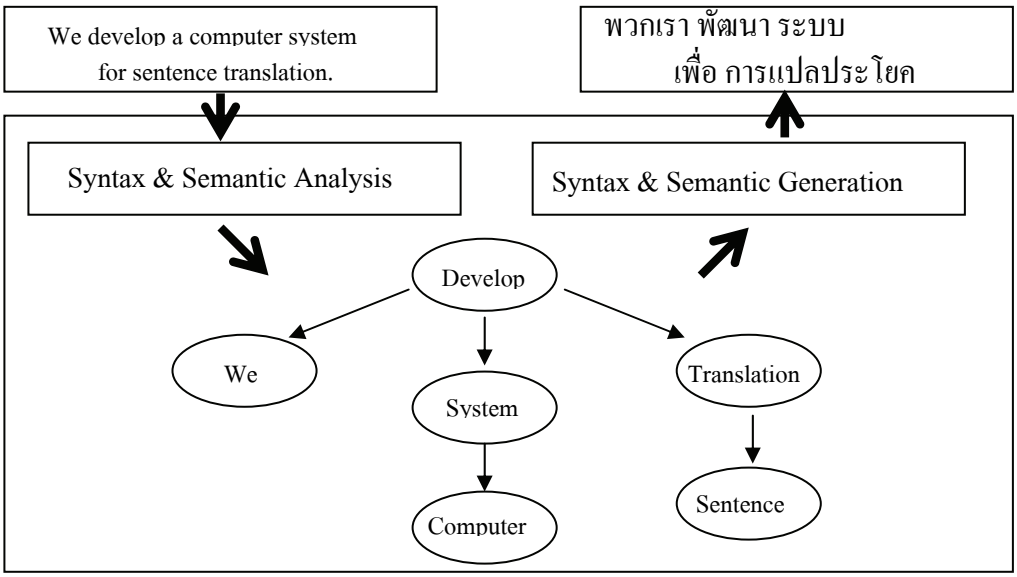
การวิเคราะห์ประโยคภาษาอังกฤษ

- วิเคราะห์หน่วยคำก่อนเข้าสู่กระบวนการหาความสัมพันธ์ทางไวยากรณ์ เช่น She (Dummy) hit (hit + past tense) a ball (a ball + singular)
- พิจารณาความสัมพันธ์ทางไวยากรณ์ของส่วนต่างๆ ของข้อความ She[subject] | hit[predicate] | a ball [object]
- พิจารณาความสัมพันธ์ทางความหมาย (Case relation) ของส่วนต่างๆ She <AGT> hit <OBJ> a ball
- สร้างรูปแทนกลาง

การสังเคราะห์ภาษาไทย

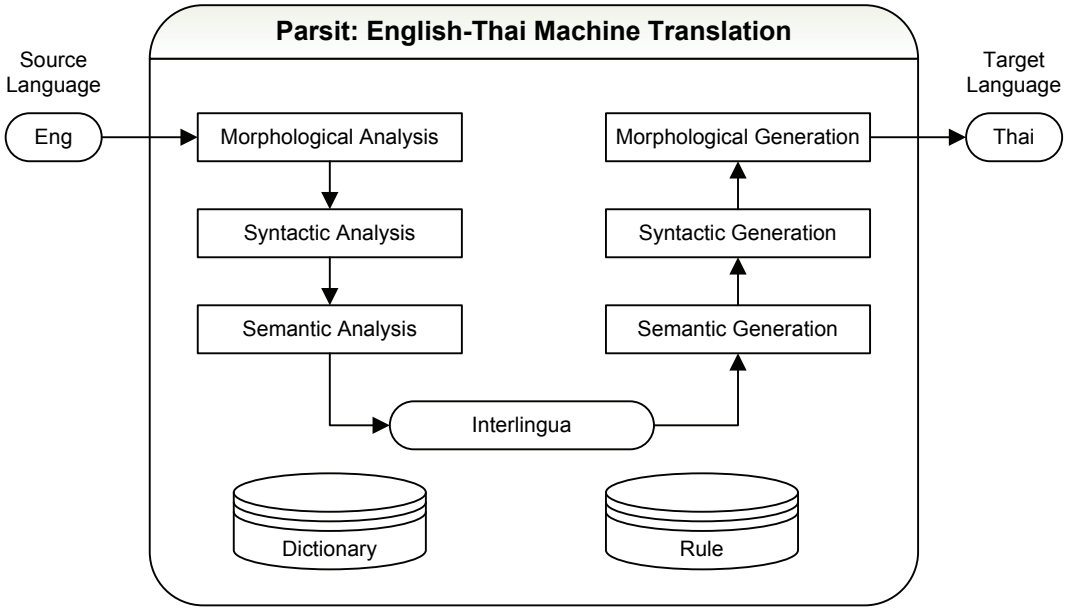
- เปลี่ยนความสัมพันธ์ทางความหมายให้เป็นความสัมพันธ์ทางไวยากรณ์
- นำข้อมูลทางไวยากรณ์ที่ได้รับผ่านรูปแทนกลางมาประกอบการสังเคราะห์ประโยค
- เปลี่ยนรูปคำจากอังกฤษเป็นภาษาไทย
- เรียงลำดับหน่วยคำต่างๆ ตามลักษณะไวยากรณ์ไทย

ภาพที่ 2.5 แสดงขั้นตอนการแปลภาษาของระบบภาษิต สำหรับข้อมูลเชิงเทคนิคของ “ภาษิต” จะกล่าวถึงในส่วนท้ายของบทนี้



ภาพที่ 2.5 แสดงขั้นตอนการแปลภาษาของระบบภาษิต

สถาปัตยกรรมระบบแปลภาษาอังกฤษ-ไทย “ภาษิต” ถูกแสดงไว้ใน ภาพที่ 2.6



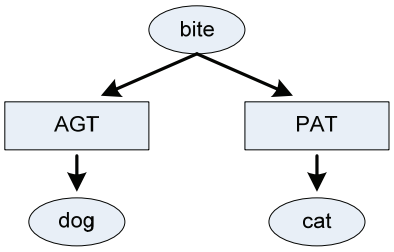
ภาพที่ 2.6 สถาปัตยกรรมระบบแปลภาษาอังกฤษ-ไทย “ภาษิต”

โมดูลวิเคราะห์สัณฐานคำ (Morphological Analysis) คือ การวิเคราะห์เชิงสัณฐาน ทำหน้าที่วิเคราะห์หน่วยคำในภาษาอังกฤษให้เป็นรากศัพท์รวมถึงการเตรียมความพร้อมให้แก่ระบบ เช่น การโหลดพจนานุกรม (Dictionary) เข้าสู่หน่วยความจำ การแก้ปัญหาความกำกวมของหน่วยคำ (morphological disambiguation)

โมดูลวิเคราะห์วากยสัมพันธ์ (Syntactic Analysis) คือ การวิเคราะห์เชิงวากยสัมพันธ์ ทำหน้าที่วิเคราะห์โครงสร้างวากยสัมพันธ์ของข้อความในภาษาอังกฤษ จากนั้นจะนำไปสร้างต้นไม้การแจงวากยสัมพันธ์ (syntactic parse tree) เพื่อนำไปใช้ในส่วนถัดไป

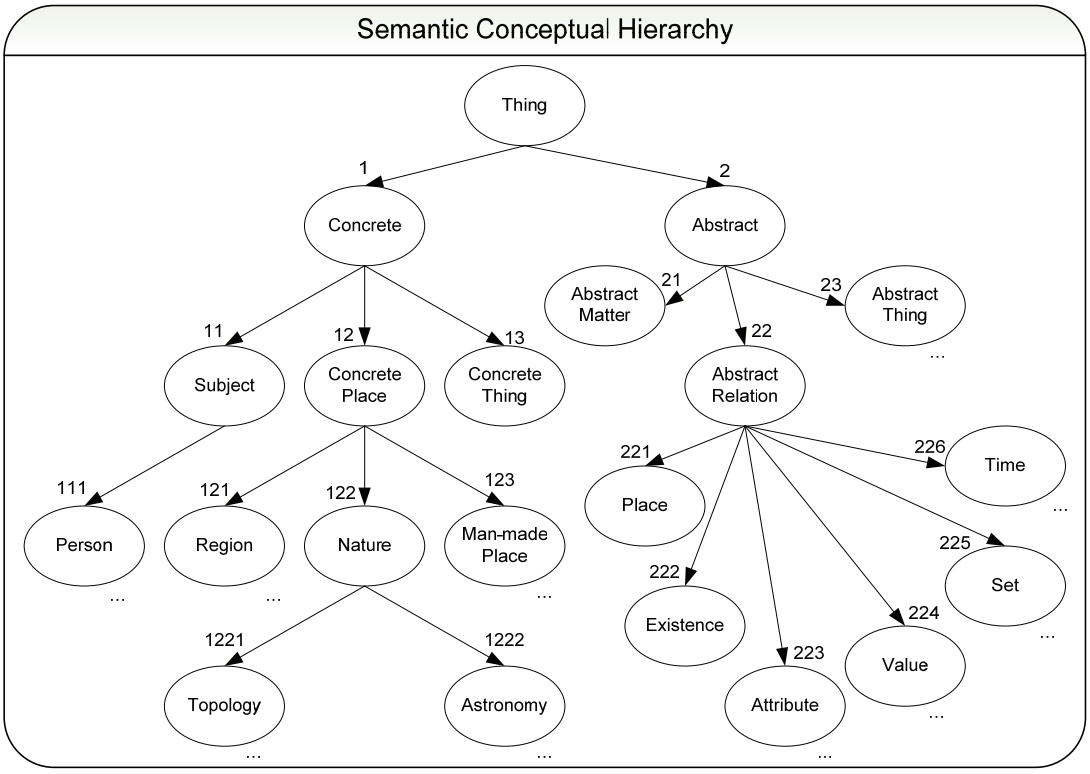
โมดูลวิเคราะห์อรรถศาสตร์ (Semantic Analysis) คือ การวิเคราะห์ความหมายจากโครงสร้างไวยากรณ์ ทำหน้าที่วิเคราะห์ต้นไม้แจงความหมายวากยสัมพันธ์ แล้วสร้างรูปแทนความหมาย (semantic representation) โดยใช้ภาษากลาง (Interlingua) เป็นสื่อในการถ่ายทอดความหมายไปเป็นภาษาไทย

ภาษากลาง (Interlingua) เป็นรูปแทนความหมายที่ไม่ขึ้นต่อภาษาใดๆ ความหมายของประโยคจะถูกแทนด้วยความสัมพันธ์ระหว่างมโนทัศน์เชิงอรรถศาสตร์ (semantic concept) เช่น ประโยค “สุนัขกัดแมว” จะแทนด้วยภาษากลางได้ดังภาพที่ 2.7



ภาพที่ 2.7 ตัวอย่างความสัมพันธ์ระหว่างมโนทัศน์เชิงอรรถศาสตร์

ในที่นี้ มโนทัศน์เชิงอรรถศาสตร์ (dog) กับ (bite) สัมพันธ์กันโดยความสัมพันธ์ Agent (ผู้กระทำกริยา) และ (cat) กับ (bite) สัมพันธ์กันโดยความสัมพันธ์ Patient (ผู้ได้รับผลโดยตรงจากการกระทำกริยา) ทั้งนี้แต่ละมโนทัศน์เชิงอรรถศาสตร์จะมีข้อกำหนด (constraint) ที่ใช้บังคับการสร้างความสัมพันธ์ระหว่างมโนทัศน์เชิงอรรถศาสตร์ เช่น (bite) จะสร้าง [AGT] กับสิ่งมีชีวิต (animate) เท่านั้น และสร้าง [PAT] กับสิ่งที่มีตัวตน (concrete object) เท่านั้น ข้อกำหนดดังกล่าวจำเป็นต้องใช้ลำดับชั้นของมโนทัศน์เชิงอรรถศาสตร์ดังที่แสดงไว้ในภาพที่ 2.8



ภาพที่ 2.8 ลำดับชั้นของมโนทัศน์เชิงอรรถศาสตร์ (Semantic Conceptual Hierarchy)

โมดูลสังเคราะห์เชิงอรรถศาสตร์ (Semantic Generation) คือ การตีความรูปแทนความหมายในภาษากลาง เพื่อนำไปสร้างต้นไม้แจงความหมายวากยสัมพันธ์ (syntactic parse tree) สำหรับภาษาไทย

โมดูลสังเคราะห์วากยสัมพันธ์ (Syntactic Generation) คือ การตีความหมายจากการแจงความหมายวากยสัมพันธ์แล้วนำไปสร้างรูปประโยคตั้งต้นในภาษาไทย ในขั้นตอนนี้โครงสร้างต้นไม้ไวยากรณ์จะถูกวิเคราะห์เพื่อสร้างลำดับการเรียงคำและเพิ่มเติมส่วนขยาย เช่น คุณสมบัติของลักษณนาม (classifier) ที่ถูกต้องตามหลักไวยากรณ์ภาษาไทย

โมดูลสังเคราะห์เชิงสัณฐานคำ (Morphological Generation) คือ การสร้างรูปผิวของคำสำหรับภาษาไทย โดยรับลำดับการเรียงคำและส่วนขยายที่เพิ่มเติมมาสร้างประโยคที่สมบูรณ์ในภาษาไทย ระบบจะเติมส่วนขยาย เช่น ลักษณนาม ลงในประโยคตามกฎการใช้ส่วนขยายนั้นๆ เพื่อให้เป็นประโยคที่สมบูรณ์

พจนานุกรมอิเล็กทรอนิกส์ที่ใช้ใน “ภาษิต” ประกอบด้วยพจนานุกรมหลัก 2 ประเภทคือ พจนานุกรมสำหรับคำหลัก (content words) ประกอบด้วยคำในกลุ่มคำนาม (noun) คำกริยา (verb) คำคุณศัพท์ (adjective) คำวิเศษณ์ (adverb) เป็นหลัก และพจนานุกรมสำหรับคำไวยากรณ์ (function word) ซึ่งประกอบด้วยคำในกลุ่มคำช่วยกริยา คำเชื่อม เป็นต้น นอกจากนี้ระบบยังให้ผู้ใช้เพิ่มพจนานุกรมตัวอื่นๆ ได้อีกตามความต้องการ เช่น พจนานุกรมศัพท์เฉพาะ เพื่อใช้ในกรณีที่ต้องการแปลเอกสารที่มีเนื้อหาเฉพาะทาง

ข้อมูลต่างๆ ของคำศัพท์แต่ละคำที่เก็บไว้ในพจนานุกรมแบ่ง 3 ส่วนคือ

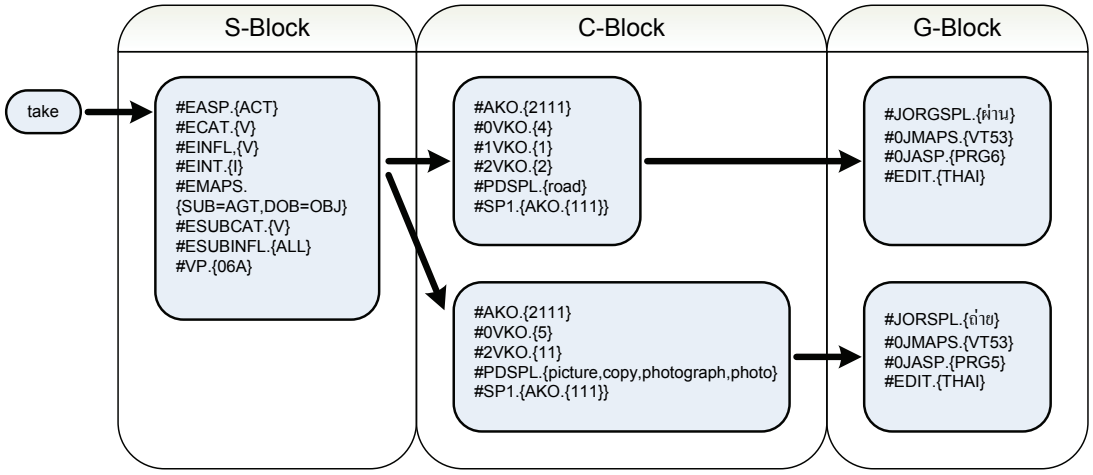
S-BLOCK ประกอบด้วยข้อมูลทางวากยสัมพันธ์ของคำศัพท์ภาษาต้นทาง

C-BLOCK ประกอบด้วยข้อมูลทางความหมาย

G-BLOCK ประกอบด้วยข้อมูลทางวากยสัมพันธ์ของคำศัพท์ปลายทาง

เพื่อเป็นการง่ายในการจัดการโครงสร้างของพจนานุกกรมจึงถูกออกแบบไว้เป็นดังนี้คือ คำศัพท์ 1 คำ สามารถมี S-BLOCK มากกว่า 1 แต่ละ S-BLOCK มี C-BLOCK ได้เพียง 1 เท่านั้น แต่สามารถมี G-BLOCK ได้มากกว่า 1

ตัวอย่างข้อมูลที่เก็บในพจนานุกรม เช่น คำว่า take ซึ่งเป็นรูปคำที่มีการใช้ (usage) หลายรูปแบบ take สามารถแปลเป็นภาษาไทยได้หลายคำโดยขึ้นอยู่กับบริบท พจนานุกรมจะเก็บลักษณะทางวากยสัมพันธ์และอรรถศาสตร์ของคำว่า take ไว้เป็นชุดๆ เรียกว่าโหนด (node) ดังแสดงไว้ในภาพที่ 2.9



ภาพที่ 2.9 ตัวอย่างข้อมูลในพจนานุกรมของคำว่า take

ข้อมูลในพจนานุกรมนี้จะใช้ร่วมกับกฎการวิเคราะห์ในการเลือกโหนดที่มี S-BLOCK ที่เหมาะสมที่สุดเพียงโหนดเดียวเพื่อสร้างเป็นรูปแทนกลางซึ่งจะส่งให้ระบบสังเคราะห์ภาษาไทยต่อไป

เมื่อโมดูลสังเคราะห์ภาษาไทยรับข้อมูลซึ่งแสดงอยู่ในรูปของโครงสร้างต้นไม้ ก็จะเปิดพจนานุกรมเพื่อดึงข้อมูลใน G-BLOCK ของคำจากโหนดที่รูปแทนกลางส่งมาเท่านั้น หากมี G-BLOCK มากกว่า 1 กฎการสังเคราะห์ภาษาไทยจะทำการคัดเลือกให้เหลือเพียง 1 G-BLOCK จากนั้นจึงสร้างเป็นข้อความภาษาไทยต่อไป