บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ภาษาธรรมชาติเป็นสิ่งที่มนุษย์เรียนรู้มาตั้งแต่เกิด ไม่ได้มีการสร้างกฎเกณฑ์ต่างๆขึ้นมาก่อนเพื่อกำหนดภาษา ความเข้าใจในภาษาของมนุษย์เป็นกลไกที่ซับซ้อนยากแก่การเข้าใจ การประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) (ยืน ภู่วรวรรณ และชัยยงค์ วงศ์ชัยสุวัฒน์, 2535; Krulee, 1991) ก็เป็นศาสตร์สาขาหนึ่งของ ปัญญาประดิษฐ์ (Artificial Intelligence) ที่พยายามศึกษาหาทาง ให้คอมพิวเตอร์ทำความเข้าใจภาษาธรรมชาติ ในรูปแบบต่าง ๆ เช่น ภาษาเขียน ภาษาพูด ฯลฯ

การที่จะทำให้คอมพิวเตอร์เข้าใจภาษาธรรมชาติได้นั้น เราจะต้องมีฐานความรู้ในด้านต่าง ๆ ให้คอมพิวเตอร์ดึงมาใช้ในการวิเคราะห์ หรือสังเคราะห์ภาษา ฐานความรู้เกี่ยวกับภาษาธรรมชาติ อาจแบ่งออกเป็นความรู้ในด้านต่าง ๆ กันหลายกลุ่ม (Allen, 1987) เช่น

1. ความรู้เกี่ยวกับการอ่าน การออกเสียงคำ (Phonetic and phonological knowledge) เป็นความรู้ในด้านของการหาความสัมพันธ์ ระหว่างคำ กับการออกเสียง

2. ความรู้เกี่ยวกับหน่วยคำ (Morphological knowledge) เป็นความรู้ในการศึกษาองค์ประกอบย่อยของแต่ละหน่วยคำ ว่าประกอบด้วยหน่วยความหมายพื้นฐานอะไร เช่น คำที่ขึ้นต้น ด้วย " นัก " ตามด้วยกริยาบางอย่าง เมื่อนำหน่วยคำย่อยทั้ง 2 ส่วนมารวมกัน จะหมายถึง บุคคลที่กระทำกริยานั้นเป็นหลัก เป็นประจำ ตัวอย่างเช่น นักเรียน คือผู้ที่เรียนเป็นหลัก นักเดินทาง คือผู้ที่เดินทางเป็นประจำ

3. ความรู้เกี่ยวกับไวยากรณ์ของภาษา (Syntactic knowledge) เป็นความรู้ทางภาษาศาสตร์ ในด้านโครงสร้าง หรือองค์ประกอบทางไวยากรณ์ของภาษา

ยกตัวอย่างไวยากรณ์ทางภาษาแบบง่าย ๆ เช่น

ประโยค ประกอบด้วย นามวลี + กริยาวลี

นามวลี อาจประกอบด้วย คำนามโดด ๆ หรือ คำนามรวมกับส่วนขยาย

กริยาวลี ประกอบด้วย คำกริยา + คำนาม เป็นต้น

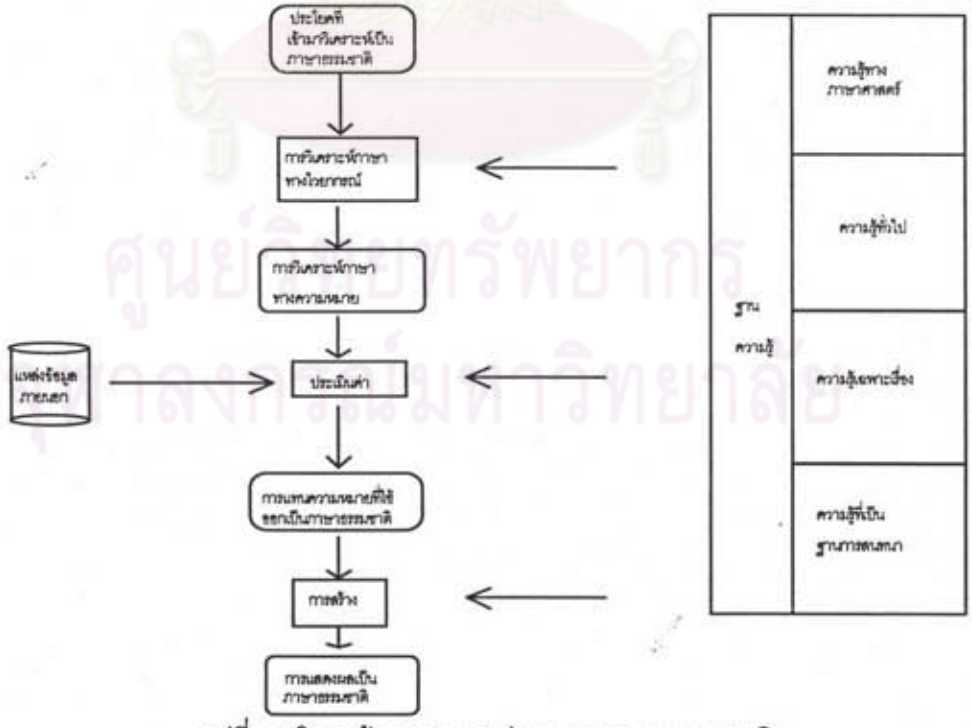
4. ความรู้เกี่ยวกับความหมายของภาษา (Semantic knowledge) เป็นความรู้ทางภาษา ในด้านของความหมาย การทำความเข้าใจภาษาธรรมชาตินั้น นอกจากจะพิจารณาความถูกต้องของไวยากรณ์แล้ว ยังต้องพิจารณาถึงความหมายของภาษาด้วยว่าสามารถเกิดขึ้นได้จริงหรือไม่ เช่น

ประโยค "ตู้เย็นรินไฟมาก" จะเห็นว่าเป็นประโยคที่ถูกต้องตามหลักไวยากรณ์ แต่เมื่อพิจารณาความหมายแล้ว จะเห็นว่าเป็นประโยคที่ผิดความหมาย เพราะกริยา "ริน" ไม่ควรใช้กับตู้เย็น ประโยคนี้คงจะเกิดการพิมพ์ผิดจากคำ "กิน" เป็น "ริน" มากกว่า ประโยคที่ถูกต้องควรเป็น "ตู้เย็นกินไฟมาก"

5. ความรู้เกี่ยวกับความต่อเนื่อง ความเกี่ยวพันกันของประโยค หรือ ความสัมพันธ์กับเหตุการณ์ในขณะนั้น (Pragmatic knowledge) ในบางโอกาส การจะทราบความหมายที่แท้จริงของประโยค เราจำเป็นต้องอาศัยการพิจารณาความหมายจากประโยครอบข้างของบทสนทนา หรือสถานการณ์ในขณะนั้นด้วย จึงจะทราบความหมายแท้จริง ที่ผู้ส่งสารต้องการส่งได้

6. ความรู้เกี่ยวกับเรื่องทั่วไป (World knowledge) เป็นความรู้พื้นฐานทั่วไป ที่ผู้ส่งสารเข้าใจว่าเป็นเรื่องที่ผู้รับสารหรือคนทั่วไปเข้าใจดีอยู่แล้ว เช่น ศัพท์เฉพาะที่ใช้กันในกลุ่มผู้ส่งสาร ความเข้าใจในภาษาหรือวัฒนธรรมท้องถิ่น ฯลฯ

กล่าวโดยสรุปแล้ว ระบบประมวลผลภาษาธรรมชาติ ประกอบด้วยส่วนสำคัญหลายส่วน โครงสร้างโดยทั่วไปของระบบ สามารถแสดงได้ ดังรูปที่ 1 (ยืน ภู่วรวรรณ และชัยยงค์ วงศ์ชัยสุวัฒน์ , 2535)



รูปที่ 1.1 โครงสร้างของระบบประมวลผลภาษาธรรมชาติ

โดยในส่วนของวิทยานิพนธ์นี้ จะทำการศึกษาเกี่ยวกับการวิเคราะห์ประโยคทางไวยากรณ์ หรือที่เรียกว่าการแจงส่วนประโยค (parsing) ซึ่งก็คือการบอกความสัมพันธ์ของคำในประโยคนั่นเอง (ยืน ภู่วรวรรณ และชัยยงค์ วงศ์ชัยสุวัฒน์, 2535)

ปัญหาที่พบในการนำระบบประมวลผลภาษาธรรมชาติ ไปใช้ในการใช้งานจริง ก็คือ ประโยคที่ส่งเข้ามาประมวลผล มีโอกาสสูงที่จะเป็นประโยคที่ผิดรูปแบบไปจากไวยากรณ์ที่กำหนด โดยสาเหตุอาจเกิดมาจากหลายกรณี เช่น

- การผิดพลาดอันเนื่องมาจากระบบกำหนดไวยากรณ์ไว้แคบเกินไป
- การจำแนกคำผิดพลาดของพจนานุกรม (อาจเนื่องมาจาก จำนวนคำในภาษาธรรมชาติมีจำนวนมาก และเกิดใหม่ได้ตลอดเวลา)
- การใช้ภาษาตามความเคยชินของผู้ใช้ อันเป็นภาษาเฉพาะกลุ่ม
- การพิมพ์ผิดพลาดของตัวข้อมูล ทำให้เกิดคำที่ไม่มีในพจนานุกรม หรือมีความหมายผิดไปจากเดิม
- การผิดพลาดอันเนื่องมาจากอุปกรณ์ รับ-ส่งข้อมูลผิดพลาด ในยุคของการสื่อสารของข้อมูลตัวข้อมูลอาจเกิดการผิดพลาดระหว่าง การรับ-ส่งข้อมูลได้

ประโยคที่ผิดรูปแบบเหล่านี้ไม่สามารถนำไปประมวลผลได้ด้วยวิธีการธรรมดา ดังนั้นการศึกษาวิธีการเพื่อให้ระบบประมวลผลภาษาธรรมชาติ สามารถวิเคราะห์ประโยคที่ผิดรูปแบบได้ จึงเป็นสิ่งจำเป็นมาก งานวิจัยในปัจจุบันได้มีการศึกษาการแก้ไขการผิดไวยากรณ์ในหลายวิธีการ (Lesmo and Torasso, 1991; Meknavin 1992) ปัญหาที่พบในการแจงส่วนประโยคภาษาธรรมชาตินั้น นอกจากปัญหาการผิดไวยากรณ์อันเนื่องมาจากตัวข้อมูลเองแล้ว พัฒนาการของภาษา หรือการใช้งานเฉพาะแบบของแต่ละกลุ่มการใช้ภาษา ก็ทำให้การกำหนดไวยากรณ์เพื่อครอบคลุมการใช้งานทั้งหมดเป็นไปได้ยาก หรือถ้าเป็นไปได้ ตัวไวยากรณ์ที่ใช้ก็จะมีขนาดใหญ่และซับซ้อนมาก ทำให้ต้องเสียเนื้อที่หน่วยความจำและเวลาในการประมวลผลนานมาก

1.2 ขอบเขตการทำวิจัย

งานที่ทำในวิทยานิพนธ์เล่มนี้ ได้เสนอแนวคิดที่จะใช้ในการแก้ไขปัญหาการแจงส่วนประโยคผิดรูปแบบไวยากรณ์ใน 2 แนวทางควบคู่กัน คือ

1. <u>การปรับปรุงตัวแจงส่วนประโยคให้สามารถจัดการกับข้อมูลที่ผิดรูปแบบไวยากรณ์ได้</u> การตรวจพบการผิดไวยากรณ์เป็นสิ่งที่ตัวแจงส่วน (parser) ทั่วไปสามารถกระทำได้ แต่การหาจุดที่เป็นสาเหตุแท้จริงของการผิดรูปแบบไวยากรณ์เป็นสิ่งที่กระทำได้ยากมาก ถ้าจะใช้เพียงฐานความรู้ในทางไวยากรณ์อย่างเดียว การกระจายการแก้ปัญหาไปในทุกทางเลือกที่เป็นไปได้ก็ไม่เหมาะสมในทางปฏิบัติ เพราะจะมีทางเลือกที่เป็นไปได้จำนวนมากมาย ดังนั้นการศึกษาการแก้ไขปัญหาการผิดไวยากรณ์ในงานวิจัยนี้ จึงได้เสนอการปรับปรุงตัวแจง

ส่วนแบบจีแอลอาร์ (ตัวแจงส่วนแบบจีแอลอาร์เป็นการขยายการแจงส่วนแบบแอลอาร์รูปแบบหนึ่ง ให้ประมวลผลข้อมูลภาษาธรรมชาติได้) ใน 3 วิธีด้วยกัน แยกตามการพิจารณาหาจุดเริ่มต้นในการแก้ไขประโยค คือ การแก้ไขการผิดไวยากรณ์ ณ จุดที่มีการตรวจพบการผิดไวยากรณ์เพียงอย่างเดียว การแก้ไขการผิดไวยากรณ์โดยการแจงส่วนแบบจีแอลอาร์-อินเวิร์ทจีแอลอาร์ (GLR-Inverted GLR or GLR-IGLR Parsing) และ การแก้ไขการผิดไวยากรณ์โดยการแจงส่วนแบบย้อนรอยกองซ้อนโครงสร้างกราฟ (Graph-Structured Stack or GSS Backtracking Parsing) ซึ่งแต่ละวิธีก็มีข้อดีข้อเสียต่างกันไป โดยจะกล่าวโดยละเอียดในบทต่อๆ ไป

2. การใช้ข้อมูลทางสถิติมาช่วยในการปรับปรุงไวยากรณ์ที่ใช้ แทนที่จะจัดการกับไวยากรณ์ทั้งหมดของภาษา ซึ่งจะมีขนาดใหญ่และซับซ้อนมาก ไวยากรณ์เริ่มต้นที่เราใช้เป็นฐานความรู้ของตัวแจงส่วนในการแจงส่วนประโยคนั้น เราจะพิจารณาเฉพาะไวยากรณ์ที่เป็นแกนหลักของภาษาเท่านั้น ส่วนข้อมูลที่พบว่าเบี่ยงเบนไปจากไวยากรณ์ที่กำหนดจะถูกจัดการ โดยส่วนจัดการข้อมูลผิดรูปแบบไวยากรณ์ในข้อหัวข้อที่ 1 และการเบี่ยงเบนทางไวยากรณ์เหล่านี้จะถูกเก็บไว้เป็นค่าสถิติของข้อมูลที่ผิดรูปแบบไวยากรณ์ ซึ่งข้อมูลที่ได้มาเหล่านี้สามารถที่จะบอกเราได้ว่า ไวยากรณ์ที่ใช้เป็นฐานความรู้ของตัวแจงส่วนในปัจจุบันนั้น เหมาะสมกับงานที่ใช้อยู่หรือไม่ และยังสามารถนำกฎต่างๆ ที่ได้จากการเรียนรู้จากประโยคที่ผิดรูปแบบไวยากรณ์ มาใช้ในการปรับปรุงไวยากรณ์ที่มีอยู่ เพื่อให้การทำงานของตัวแจงส่วนมีประสิทธิภาพ เหมาะสมกับการทำงานมากขึ้น

ในการศึกษาการแก้ไขประโยคที่ผิดรูปแบบไวยากรณ์ด้วยตัวแจงส่วนแบบจีแอลอาร์นี้ โดยลักษณะการทำงานของตัวแจงส่วนแล้ว การทำงานต่างๆ จะขึ้นอยู่กับตารางแอลอาร์เป็นสำคัญไม่ขึ้นกับภาษาที่ใช้ สามารถนำไปใช้งานกับภาษาใดก็ได้ ที่ตัวไวยากรณ์ของภาษาสามารถเขียนให้อยู่ในรูปแบบของไวยากรณ์ไม่พึ่งบริบทได้