

การประมวลผลภาษามนุษย์ ด้วยคอมพิวเตอร์

อัสนีย์ ก่อตระกูล¹
และคณาจารย์ ภาควิชาภาษาศาสตร์²

ทุกครั้งที่มีการกล่าวถึงคอมพิวเตอร์ เราจะนึกถึงภาพการนำคอมพิวเตอร์มาช่วยพิมพ์เอกสาร คำนวณข้อมูลตัวเลข เรียกค้นข้อมูลจากฐานข้อมูล หรือค้นข้อมูลทั่วโลกผ่านอินเทอร์เน็ต ในบทความนี้จะนำเสนออีกภาพหนึ่งของคอมพิวเตอร์ ที่เราพยายามสร้างขึ้นมาให้มีความสามารถประมวลผลหรือเข้าใจภาษาของเราเอง

1. เป้าหมายของการประมวลผลภาษามนุษย์ด้วยคอมพิวเตอร์

การประมวลผลภาษามนุษย์ด้วยคอมพิวเตอร์หรือการประมวลผลภาษาธรรมชาติ (Natural Language Processing) เป็นส่วนหนึ่งของงานวิจัยทางด้านปัญญาประดิษฐ์ (Artificial Intelligence) ที่พยายามจำลองความฉลาดของมนุษย์ให้กับคอมพิวเตอร์ อันได้แก่ การแก้ปัญหาโดยใช้เหตุผล การเรียนรู้วิธีการแก้ปัญหาด้วยตนเอง รวมทั้งการสื่อสารด้วยภาษาของมนุษย์

เป้าหมายของการประมวลผลภาษาธรรมชาติ คือ การทำให้คอมพิวเตอร์สามารถเข้าใจ และโต้ตอบกับผู้ใช้ด้วยภาษาของผู้ใช้เอง ภาษาธรรมชาติมี 2 รูป คือ รูปเสียงหรือภาษาพูด (speech) และรูปอักษรหรือภาษาเขียน (text) ในที่นี้จะกล่าวถึงภาษาเขียนเท่านั้น

¹ ดร.อัสนีย์ ก่อตระกูล ผู้ช่วยศาสตราจารย์ ภาควิชาวิศวกรรมคอมพิวเตอร์

² กมลดา นาคะศิริ รองศาสตราจารย์, วิสมัย มโนมัยพิบูลย์ รองศาสตราจารย์, ศิริพร แดงเที่ยง รองศาสตราจารย์, วิภากร วงศ์ไทย ผู้ช่วยศาสตราจารย์ และ ดร.ทัศนาลัย บุรพาชีพ ผู้ช่วยศาสตราจารย์ ภาควิชาภาษาศาสตร์

ตัวอย่างการประยุกต์ใช้งานของระบบประมวลผลภาษาธรรมชาติ คือ การแปลภาษาอัตโนมัติ (Machine Translation) การย่อความอัตโนมัติ (Text summarization) การสั่งงานคอมพิวเตอร์ด้วยภาษาของผู้ใช้ (Natural Language front-end) การตรวจเอกสารอัตโนมัติ (อัสนีย์และคณะ, 2541) ฯลฯ

2. ความยากของการประมวลผลภาษาธรรมชาติ

2.1 ภาษาประดิษฐ์ กับ ภาษาธรรมชาติ

ภาษาประดิษฐ์ (Artificial Language) เป็นภาษาที่เราสร้างขึ้นมาเพื่อใช้งานเฉพาะด้านและอยู่ในวงจำกัด เช่น ภาษาลอจิก ภาษาคณิตศาสตร์ ภาษาโทรเลข ภาษาไฟ ภาษาโปรแกรม เป็นต้น นอกจากนี้ภาษาประดิษฐ์ยังเป็นภาษาที่เราสามารถควบคุมได้อย่างสมบูรณ์ทั้งในด้านคำศัพท์ (vocabulary) ด้านวากยสัมพันธ์ (syntax) และทางด้านความหมาย (semantic)

ภาษาธรรมชาติ (Natural Language) เป็นเครื่องมือที่มนุษย์ใช้ถ่ายทอดความคิดและสื่อสารในชีวิตประจำวัน ภาษาธรรมชาติเป็นภาษาที่แตกต่างจากภาษาประดิษฐ์อย่างสิ้นเชิง กล่าวคือภาษาธรรมชาติเป็นภาษาที่ไม่สามารถควบคุมให้มีลักษณะจำกัด ยอมให้มีความผิดพลาดได้ มีความกำกวม และเป็นภาษาที่สื่อความหมายได้โดยไม่จำเป็นต้องมีความชัดเจน

2.2 ปัญหาสำคัญที่พบในการประมวลผลภาษาธรรมชาติ

ด้วยคุณสมบัติพิเศษของภาษาธรรมชาติ ปัญหาสำคัญๆ ที่พบและต้องหาวิธีแก้ไขมีดังต่อไปนี้

2.2.1 ความไม่มีขอบเขตจำกัด

ภาษาธรรมชาติเปรียบเหมือนสิ่งมีชีวิต กล่าวคือ มีคำเกิดใหม่ขึ้นตลอดเวลา คำเหล่านั้นได้แก่ ชื่อเฉพาะ ศัพท์เทคนิค คำประสม และคำย่อ

ก้าวแรกที่จะให้คอมพิวเตอร์เข้าใจภาษามนุษย์ ก็คือ การสอนให้รู้จักคำ วิธีที่ง่ายที่สุดคือ การสร้างพจนานุกรมอิเล็กทรอนิกส์ แต่การสร้างพจนานุกรมให้เก็บคำได้ครบเป็นเรื่องที่ทำได้ไม่ง่าย

2.2.2 การยอมให้มีข้อผิดพลาด

มนุษย์สามารถเข้าใจภาษาได้โดยง่าย แม้ว่าภาษาที่ใช้จะมีข้อผิดพลาด

พลาดบ้าง หรือฝ่าฝืนกฎระเบียบทางภาษา ตัวอย่างเช่น

- มันสำปะหลังปลูกมากในภาคตะวันออก
(เมื่อกำหนดว่า “ปลูก” เป็นสกรรมกริยา)
- ชาวชายดีกว่าปีที่แล้ว
(เมื่อกำหนดว่า “ชาย” ต้องการประธานที่เป็นมนุษย์)
- A man sit on the table.
(เมื่อ sit ใช้กับประธานที่เป็นเอกพจน์ ต้องเติม s เป็น sits)

แต่สำหรับคอมพิวเตอร์จะทำงานภายใต้กฎหรือคำสั่งที่มีระเบียบแบบแผนที่แน่นอน และเป็นระบบเท่านั้น ดังนั้น การให้คอมพิวเตอร์ยอมคำนวณภาษาที่มีข้อผิดพลาด (illformed sentence) จึงต้องเป็นระบบที่มีความแกร่ง (robustness) มีความยืดหยุ่น ยอมรับประโยคที่ใช้ไวยากรณ์ไม่ถูกต้อง หรือมีการสะกดคำผิดพลาดบ้าง ทั้งนี้จะต้องอยู่ภายในขอบเขตที่จำกัด

2.2.3 ความกำกวมของภาษาในระดับต่างๆ

ภาษารธรรมชาติมีความกำกวมเกิดขึ้นได้ทุกระดับ :

ระดับคำ

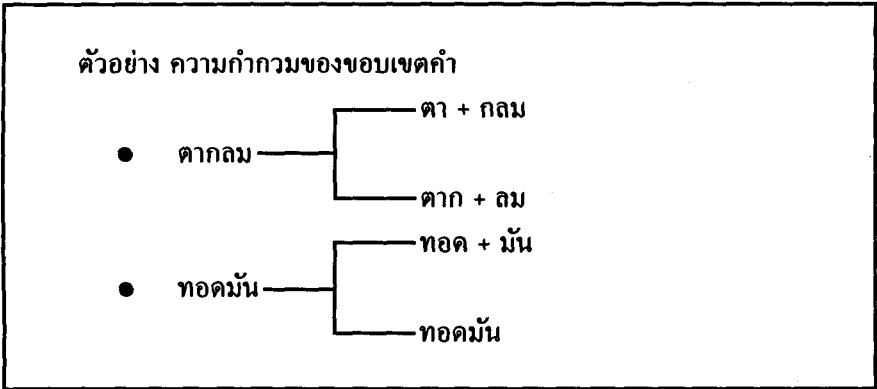
ความกำกวมในระดับคำได้แก่ ความกำกวมของชนิดคำ (part-of-speech ambiguity) ความกำกวมของความหมาย (word sense ambiguity)

ตัวอย่าง ความกำกวมในระดับคำ

- ไก่ขัน
- เค้นคักน้ำใส่ขัน

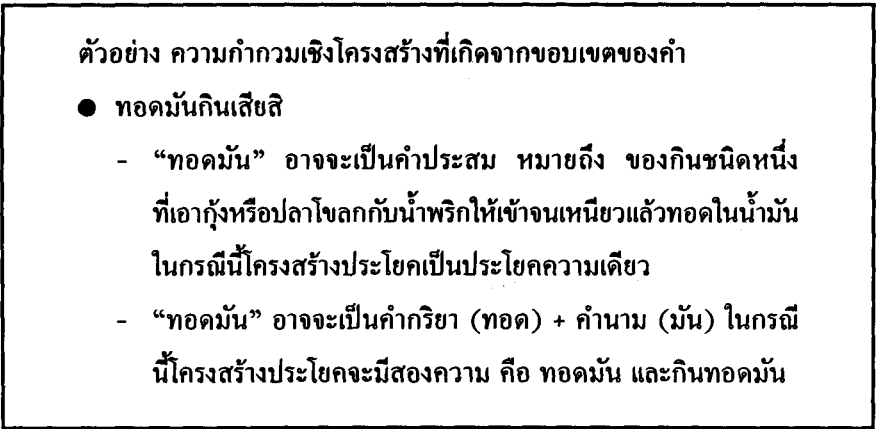
คำว่า “ชั้น” เป็นได้ทั้ง คำนาม และคำกริยา ในขณะเดียวกันก็มีความหมายที่แตกต่างกัน

สำหรับภาษาไทย ความกำกวมระดับคำยังรวมถึง ความกำกวมในการหาขอบเขตของคำ (word boundary ambiguity)

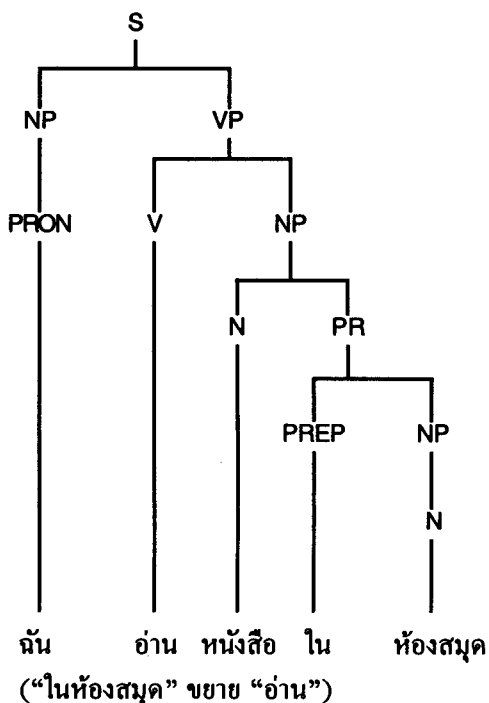
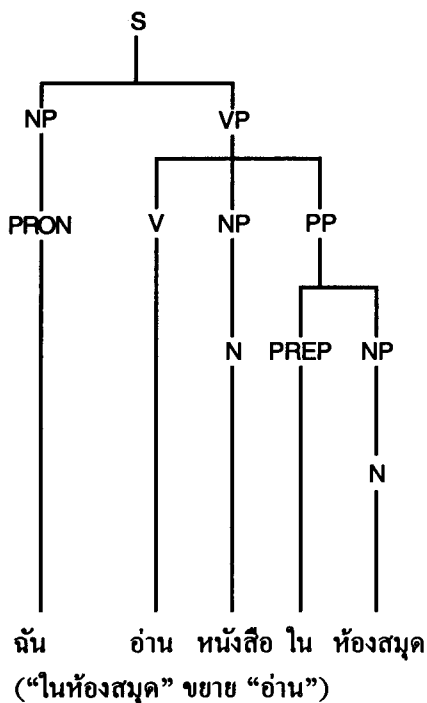


ระดับประโยค

ความกำกวมในระดับประโยค คือ ความกำกวมในเชิงโครงสร้าง (structural ambiguity) ที่มีผลมาจากความคลุมเครือของขอบเขตคำ การละคำ และความคลุมเครือของส่วนขยาย โครงสร้างของประโยคที่ต่างกัน จะให้ความหมายที่ต่างกัน



ตัวอย่าง ความกำกวมทางโครงสร้างที่เกิดจากส่วนขยาย
- ฉันอ่านหนังสือในห้องสมุด



ระดับความหมาย

ความกำกวมในระดับความหมาย คือ การแปลความหมายของประโยคได้มากกว่าหนึ่งอย่าง

ตัวอย่าง ความกำกวมของการแปลความหมายเมื่อใช้คำว่า “ไม่”
พ่อไม่ได้ขับรถไปหัวหิน

- ความหมายที่ 1 ไม่ใช่พ่อแต่เป็นคนอื่นที่ขับรถไปหัวหิน
- ความหมายที่ 2 พ่อไปหัวหินโดยวิธีอื่น เช่น รถไฟ ฯลฯ
- ความหมายที่ 3 พ่อไม่ได้ขับรถไปหัวหิน แต่ไปที่พัทยา

นอกจากลักษณะพิเศษของภาษาธรรมชาติที่กล่าวไปแล้วข้างต้น ภาษายังมีเครื่องมือทางภาษาที่ทำให้ภาษามีความกระชับ ไม่ฟุ่มเฟือย เช่น การใช้คำอ้างอิง (Anaphoric Reference) การละคำในประโยค คอมพิวเตอร์จึงต้องมีเทคนิค และกลไกต่างๆ รวมทั้งการใช้ฐานความรู้นอกเหนือจากความรู้ทางภาษา เพื่อให้บรรลุเป้าหมายสุดท้ายคือการตีความ หรือแปลความหมายได้อย่างสมบูรณ์

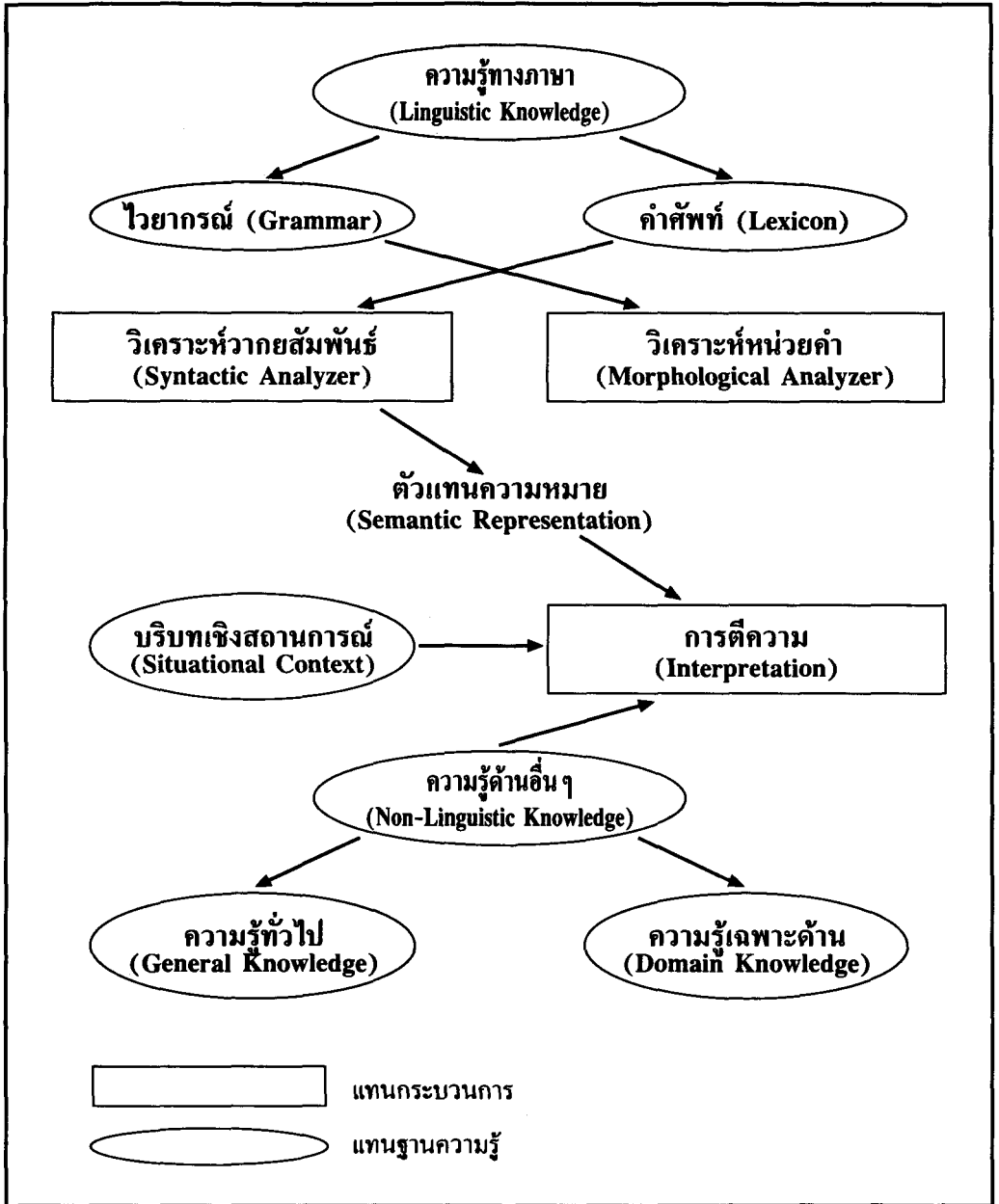
ตัวอย่าง ระบบทำความเข้าใจภาษาที่สมบูรณ์ จะต้องสามารถตอบคำถามว่า “นายบังรอนได้ทานแฮมเบอร์เกอร์หรือไม่” หลังจากทำความเข้าใจบทความต่อไปนี้

“นายบังรอนจองครรถหน้าร้านแมคโดนัล
เขาสั่งแฮมเบอร์เกอร์ชนิดพิเศษ
แต่คนขายทำขนมปังไหม้ เขาโกรธมาก และ
เดินผลุนผลันออกจากร้าน”

3. สถาปัตยกรรมของระบบประมวลผลภาษาธรรมชาติ

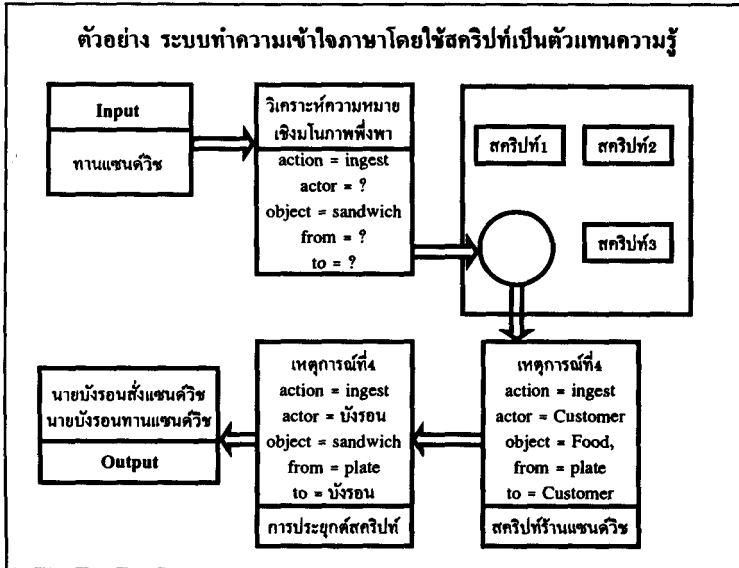
3.1 ฐานความรู้ กับ การตีความประโยค

ในรูปที่ 1 แสดงให้เห็นถึงการใช้ความรู้ทางภาษา ซึ่งประกอบด้วยคำศัพท์และไวยากรณ์ ความรู้ที่เป็นด้านอื่น (Non-linguistic Knowledge) ได้แก่ ความรู้ทั่วไป (World Knowledge) ความรู้เฉพาะด้าน (Domain Knowledge) รวมทั้งบริบทเชิงสถานการณ์ของประโยคเพื่อใช้ในการแปลความหมายที่แท้จริง



รูปที่ 1 ความสัมพันธ์ระหว่างฐานความรู้กับการตีความ

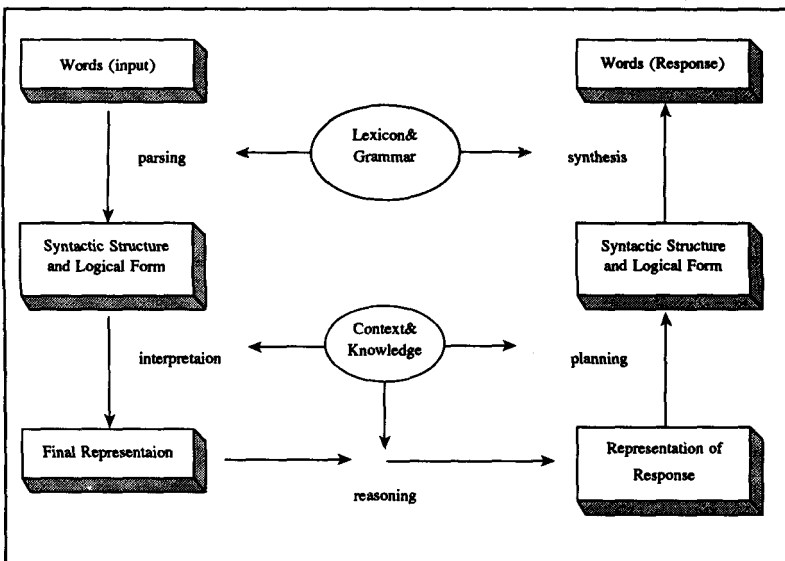
ตัวอย่าง การทำความเข้าใจประโยค “ทานแซนด์วิช” ใน “...นายบ๊รอนเข้าร้านแมคโดนัลทานแซนด์วิช...” ต้องอาศัยสคริปต์เป็นตัวแทนความรู้ทั่วไปเกี่ยวกับการสั่งอาหารในร้านอาหาร เพื่ออนุมานหาความหมายรายละเอียดที่ไม่ได้ปรากฏในบทความ (implicit information) (ดูรูปที่ 2) คือ นายบ๊รอนสั่งแซนด์วิช



รูปที่ 2 ระบบทำความเข้าใจภาษาที่ใช้สคริปต์เป็นตัวแทนความรู้

3.2 สถาปัตยกรรมของระบบประมวลผลภาษาธรรมชาติ

ในรูปที่ 3 แสดงให้เห็นถึงสถาปัตยกรรมของระบบประมวลผลภาษาธรรมชาติที่ประกอบขึ้นด้วย การทำความเข้าใจภาษา (Natural Language Understanding) และการสร้างภาษา (Natural language Generation)



รูปที่ 3 สถาปัตยกรรมของระบบประมวลผลภาษาธรรมชาติ

4. แนวทางหลักในการประมวลผลภาษาธรรมชาติ

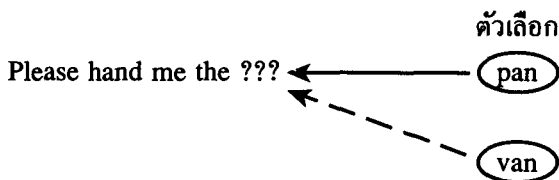
แนวทางหลักในการประมวลผลภาษาธรรมชาติที่ใช้กันมากในปัจจุบันมีอยู่ 2 แนวทาง ได้แก่ แนวทางอิงความรู้ (Knowledge based หรือ AI approach) และแนวทางอิงสถิติ (Statistical based หรือ Corpus based approach) เทคนิคและวัตถุประสงค์ของแต่ละแนวทางได้สรุปไว้ในตารางที่ 1

ตารางที่ 1 เทคนิคและเป้าหมายในการแก้ไขปัญหของแนวทางอิงความรู้ และแนวทางอิงสถิติ

แนวทาง	เทคนิคที่ต้องการ	เป้าหมาย
แนวทางอิงความรู้ (ปัญญาประดิษฐ์)	● การแทนความรู้ (Knowledge Representation) ● การใช้และการจัดการความรู้ (Reasoning and Knowledge Manipulation)	● เพื่อแก้ไขปัญหาคำถามทำความเข้าใจของภาษาทุกระดับ
แนวทางอิงสถิติ	● คลังประโยคขนาดใหญ่ ● เทคนิคทางสถิติและความน่าจะเป็น	● เพื่อแก้ไขปัญหาคำถามครอบคลุม (coverage) ความแข็งแกร่ง (robustness) และความสามารถในการขยายตัว (extensibility) ของระบบเพื่อประมวลผลภาษาที่ใช้จริง

อย่างไรก็ตาม ปัจจุบันได้มีการใช้แนวทางผสม (Hybrid approach) ระหว่างเทคนิคปัญญาประดิษฐ์และเทคนิคทางสถิติเพื่อเพิ่มประสิทธิภาพในการแก้ปัญหาที่พบในระบบประมวลผลภาษาธรรมชาติ รวมทั้งมีการศึกษาระบบภาษาที่เกี่ยวข้องกับพฤติกรรมของมนุษย์ อันได้แก่ การศึกษาว่าคนเราใช้ภาษาอย่างไร มีวิธีเรียนรู้และเข้าใจภาษาอย่างไร แนวทางใหม่ที่อิงพฤติกรรมมนุษย์นี้เรียกว่า จิตภาษาศาสตร์เชิงคำนวณ (Computational Psycholinguistic)

-ตัวอย่าง การประยุกต์แนวทางสถิติกับการรู้จำเสียง (Speech Recognition)



ขั้นตอนในการประมวลผลตัวอย่างข้างต้น คือ ประมวลผลคำที่เสียงไม่ชัดได้ 2 คำ คือ van กับ pan แล้วใช้ทฤษฎีความน่าจะเป็นตัดสินว่า hand กับ pan และ hand กับ van คุใดน่าจะเป็นไปได้มากกว่ากัน

5. ทิศทางของการประมวลผลภาษาธรรมชาติ

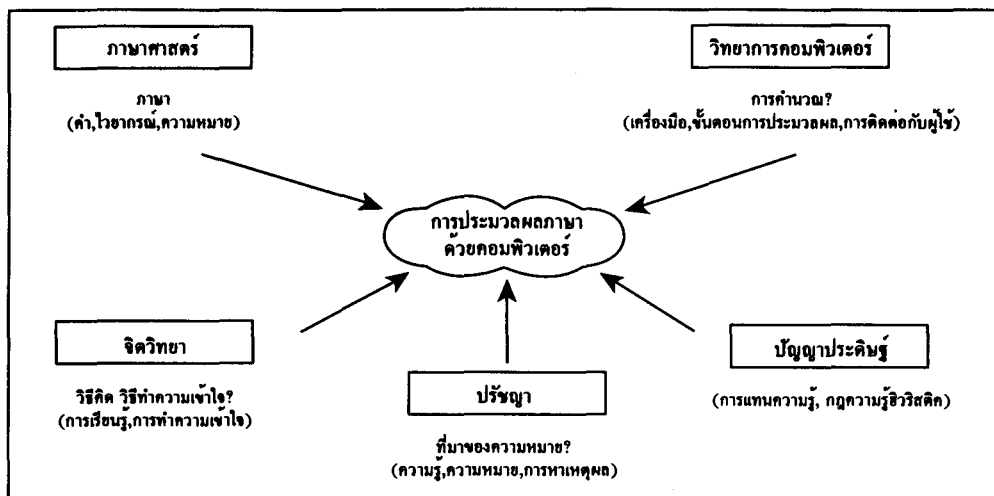
งานวิจัยของการประมวลผลภาษาธรรมชาติมี 2 ทิศทาง :

- มุ่งเน้นการสร้างทฤษฎีเชิงคำนวณ และโมเดลในการประมวลผล เช่น ระบบไวยากรณ์ (grammar formalism) โมเดลการประมวลผลบทความ (Discourse processing Model) การแทนความเชื่อ และเป้าหมาย (Belief and Goal Representation) เป็นต้น

- การมุ่งเน้นการประยุกต์เทคนิคการประมวลผลภาษาธรรมชาติเพื่อสร้างระบบที่ให้ผลเชิงพาณิชย์ เช่น การแปลภาษาอัตโนมัติ การประมวลผลดัชนีอัตโนมัติ การค้นข้อมูลในห้องสมุดอย่างฉลาด การตรวจรู้ฟบทความอัตโนมัติ การแยกบทความหรือเอกสารในรูปอิเล็กทรอนิกส์อย่างอัตโนมัติ เป็นต้น

6. สรุป

ระบบประมวลผลภาษาธรรมชาติ คือ ระบบคอมพิวเตอร์ที่สามารถประมวลผลภาษามนุษย์ ซึ่งการให้คอมพิวเตอร์ทำความเข้าใจและตีความหมายได้อย่างถูกต้องนั้น ต้องอาศัยศาสตร์หลายแขนง ดังรูปที่ 4



รูปที่ 4 ศาสตร์ต่างๆ ที่ใช้ในระบบประมวลผลภาษาธรรมชาติ

อย่างไรก็ตาม การแปลความหมายให้ถูกต้องและสมบูรณ์ยังคงเป็นเรื่องที่นักวิจัยด้านการประมวลผลภาษาธรรมชาติต้องฟันฝ่าอุปสรรคอีกมาก ดังภาพต่อไปนี้



เอกสารอ้างอิง

- Allen, James. 1989. **Natural Language Understanding**. Menlo Park, Calif.: Benjamin Cummings.
- Armstrong, Susan, editor. 1993. **Using Large Corpora**. MIT Press.
- Grishman, Ralph. 1992. **Computational Linguistics : An Introduction**. Cambridge: Cambridge Press.
- Introduction to Natural Language Processing**. 199?. [online]. Available: <http://www.cs.um.edu.mt/~mros/cs305>.
- Pitrat, J. 1988. **An Artificial Intelligence Approach to Understanding Natural Language**. North Oxford Academic.
- Reilly, Ronan G. and Sharkey, Noel E. 1992. **Connectionist Approaches to Natural Language Processing**. Lawrence Erlbaum Associates.
- Sowa, John F. 1984. **Conceptual Structures : Information Processing in Mind and Machine**. Addison-Wesley.
- จินดา งามสุทธิ. 2524. **ภาษาศาสตร์ภาษาไทย**. กรุงเทพฯ: สำนักพิมพ์โอเดียนสโตร์.
- อัศนีชัย ก่อตระกูล และคนอื่นๆ. 2541. **การตรวจสอบไวยากรณ์และรูปแบบประโยคภาษาไทยโดยอัตโนมัติ**. รายงานฉบับสมบูรณ์เสนอสภาวิจัยแห่งชาติ.