การประมวลผลภาษาธรรมชาติ

Natural Language Processing

ทัศนวรรณ ศูนย์กลาง

ภาควิชาคอมพิวเตอร์ คณะวิทยาสตร์ มหาวิทยาลัยศิลปากร

หัวข้อในวันนี้

- การประมวลผลภาษาธรรมชาติ
- ระบบภาษาธรรมชาติ
- การประยุกต์ใช้งาน
- การประมวลผลภาษาไทย
- แนวทางการวิจัย

การประมวลผลภาษาธรรมชาติ

ภาษาธรรมชาติ

- สาขาคอมพิวเตอร์ คำว่า "ภาษา" หมายถึง ภาษาคอมพิวเตอร์ต่างๆ เช่น ภาษาซี ภาษาฟอร์แทรน ภาษาโคบอล เป็นต้น
- ภาษาที่มนุษย์ใช้ในการติดต่อสื่อสารกัน ตัวอย่าง เช่น ภาษาไทย ภาษาอังกฤษ ภาษาญี่ปุ่น เป็นต้น
- การนำคอมพิวเตอร์มาใช้ในการประมวลภาษามนุษย์ จึงใช้คำว่า "ภาษาธรรมชาติ" เพื่อให้แตกต่างไปจากคำว่า "ภาษา" ซึ่งจะหมายถึง "ภาษาคอมพิวเตอร์"

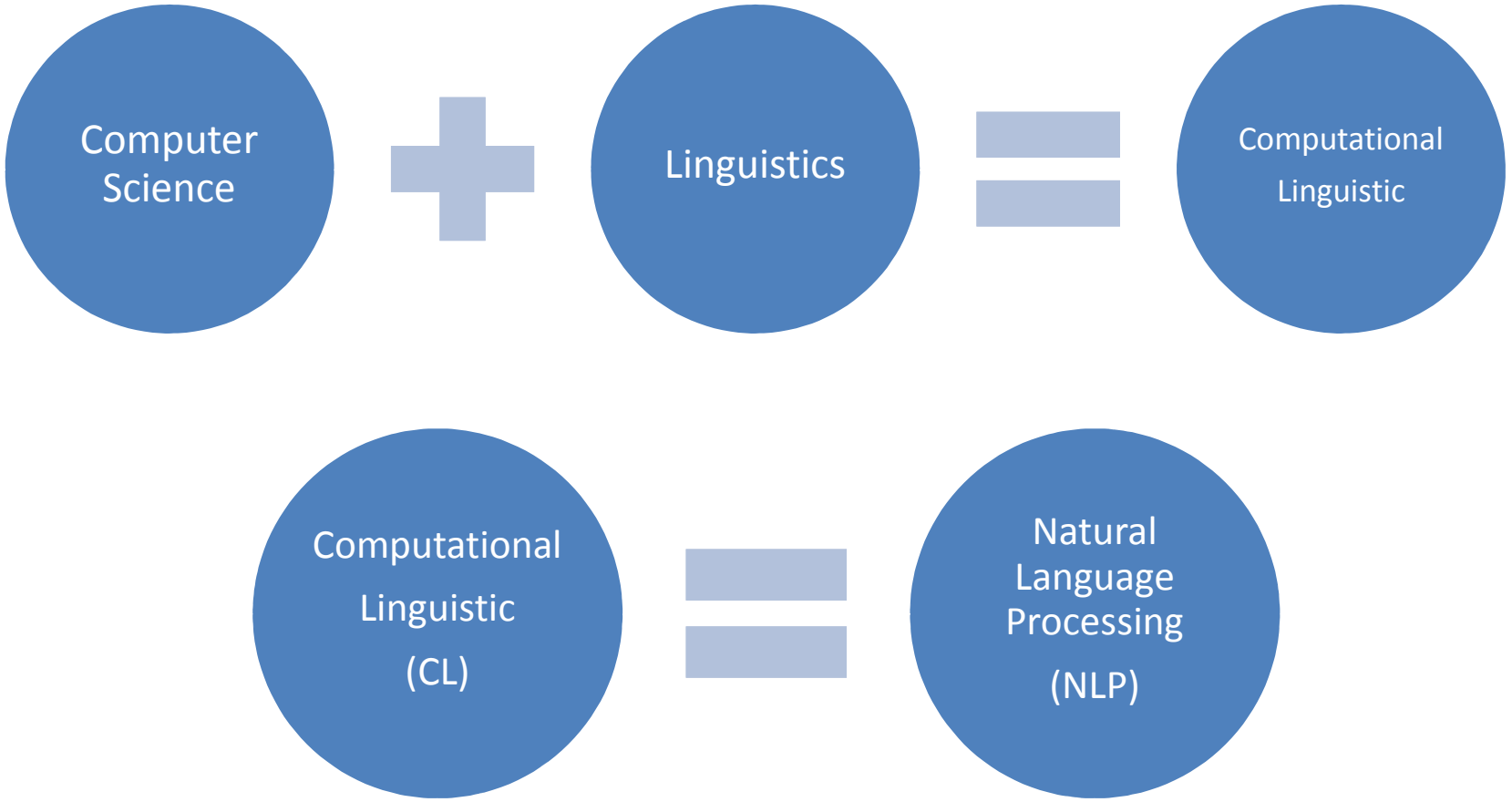
การประมวลผลภาษาธรรมชาติ

- ความหมาย
  - การประมวลผลข้อสนเทศที่อยู่ในรูปภาษาธรรมชาติหรือภาษามนุษย์ โดยพยายามจำลองความฉลาดของมนุษย์ให้กับคอมพิวเตอร์
  - การประมวลผลและใช้งานภาษาธรรมชาติ การทำความเข้าใจภาษาธรรมชาติเพื่อให้คอมพิวเตอร์สามารถเข้าใจภาษามนุษย์ได้
  - การที่เครื่องคอมพิวเตอร์สามารถทำการประมวลผลได้โดยใช้ภาษาธรรมชาติสั่งการให้คอมพิวเตอร์ปฏิบัติตามได้
  - **การประมวลผลที่ทำให้คอมพิวเตอร์เข้าใจและโต้ตอบกับคำสั่ง หรือข้อความที่เป็นภาษา “ธรรมชาติ” ของมนุษย์ได้**

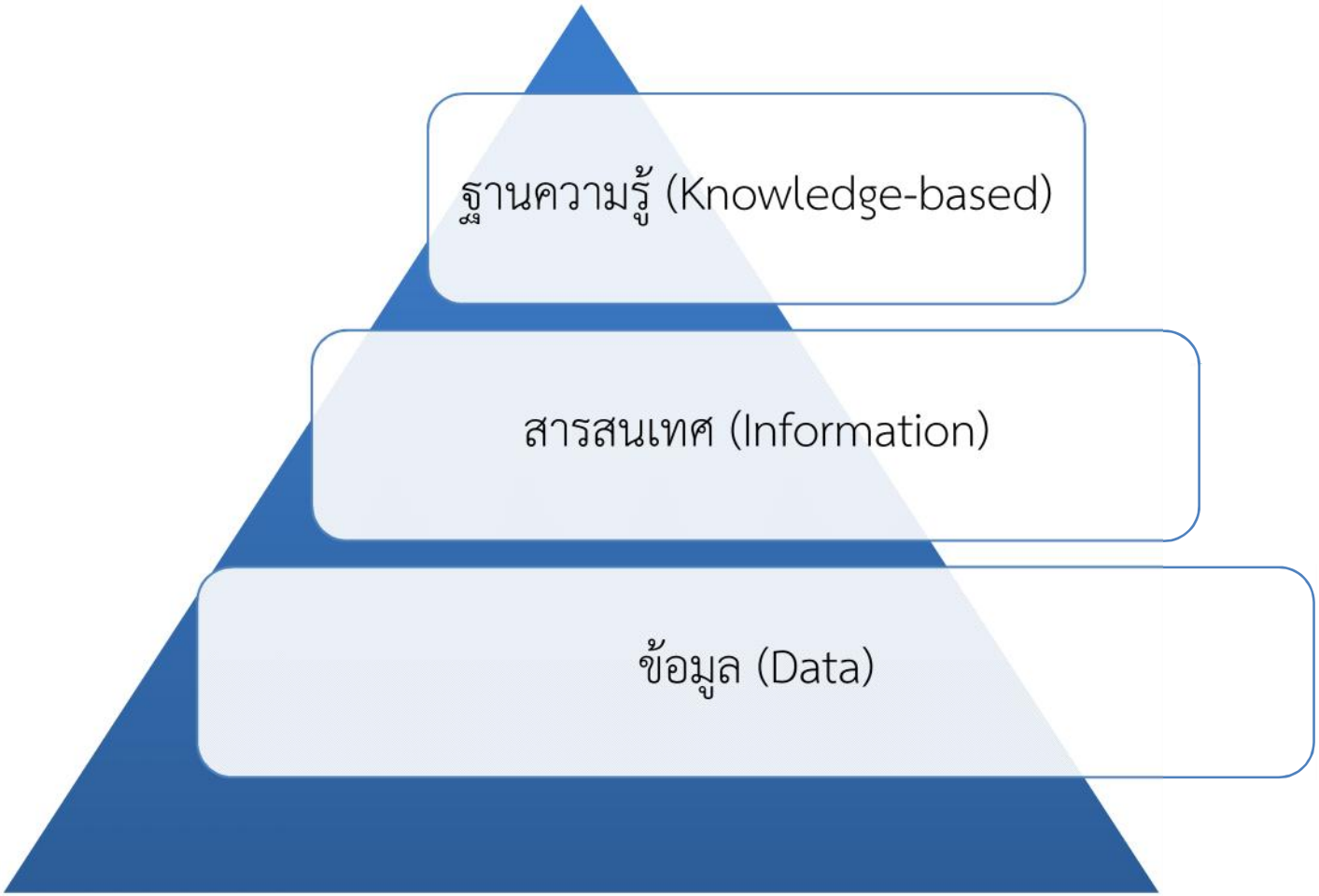
แนวทางการพัฒนา

- ด้านปัญญาประดิษฐ์
  - การแทนความรู้ (knowledge representation)
  - การให้เหตุผล (reasoning)
  - การเรียนรู้ด้วยเครื่อง (machine learning)
- ด้านภาษาศาสตร์
  - ศึกษาและเข้าใจโครงสร้างทางภาษาศาสตร์
  - ทฤษฎีที่เกี่ยวกับรูปแบบ/โครงสร้างของภาษา
  - การวิเคราะห์โครงสร้างและความหมาย

ศาสตร์ที่เกี่ยวข้อง



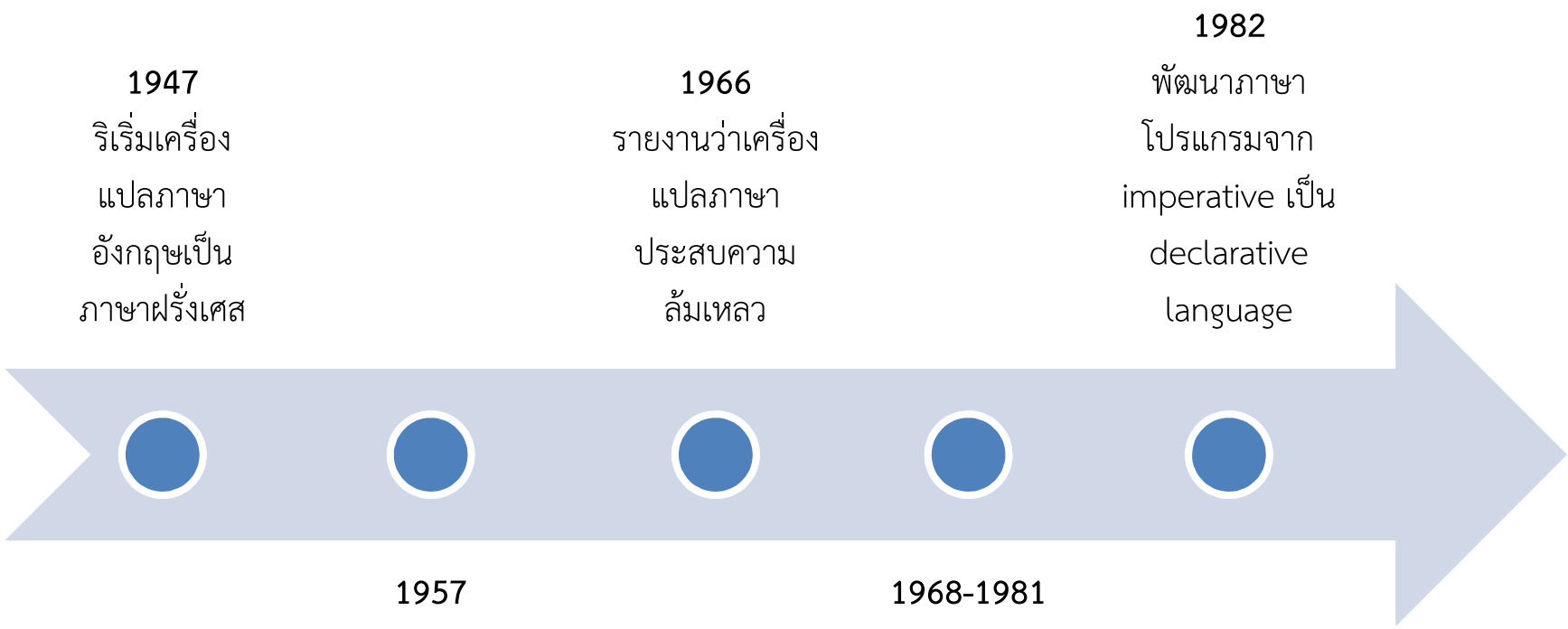
ยุคของการประมวลผล



แรงจูงใจในการวิจัย

- ภาษาธรรมชาติเป็นภาษาที่มนุษย์ใช้ในการติดต่อสื่อสารกัน
- ข้อมูลมหาศาลเก็บอยู่ในรูปของภาษาธรรมชาติ
- คอมพิวเตอร์สามารถเข้าใจข้อมูล และแปลงเป็นสารสนเทศได้
- คอมพิวเตอร์สามารถเข้าใจสารสนเทศ และแปลงเป็นฐานความรู้ได้
- คอมพิวเตอร์ไม่จำเป็นต้องแปลงข้อมูล แต่เข้าใจความหมายและนำไปใช้เป็นฐานความรู้ได้โดยตรง
- คอมพิวเตอร์สามารถเข้าถึงและใช้ประโยชน์จากข้อมูลได้มีประสิทธิภาพ
- ลดช่องว่างระหว่างคอมพิวเตอร์กับมนุษย์

วิวัฒนาการของ NLP

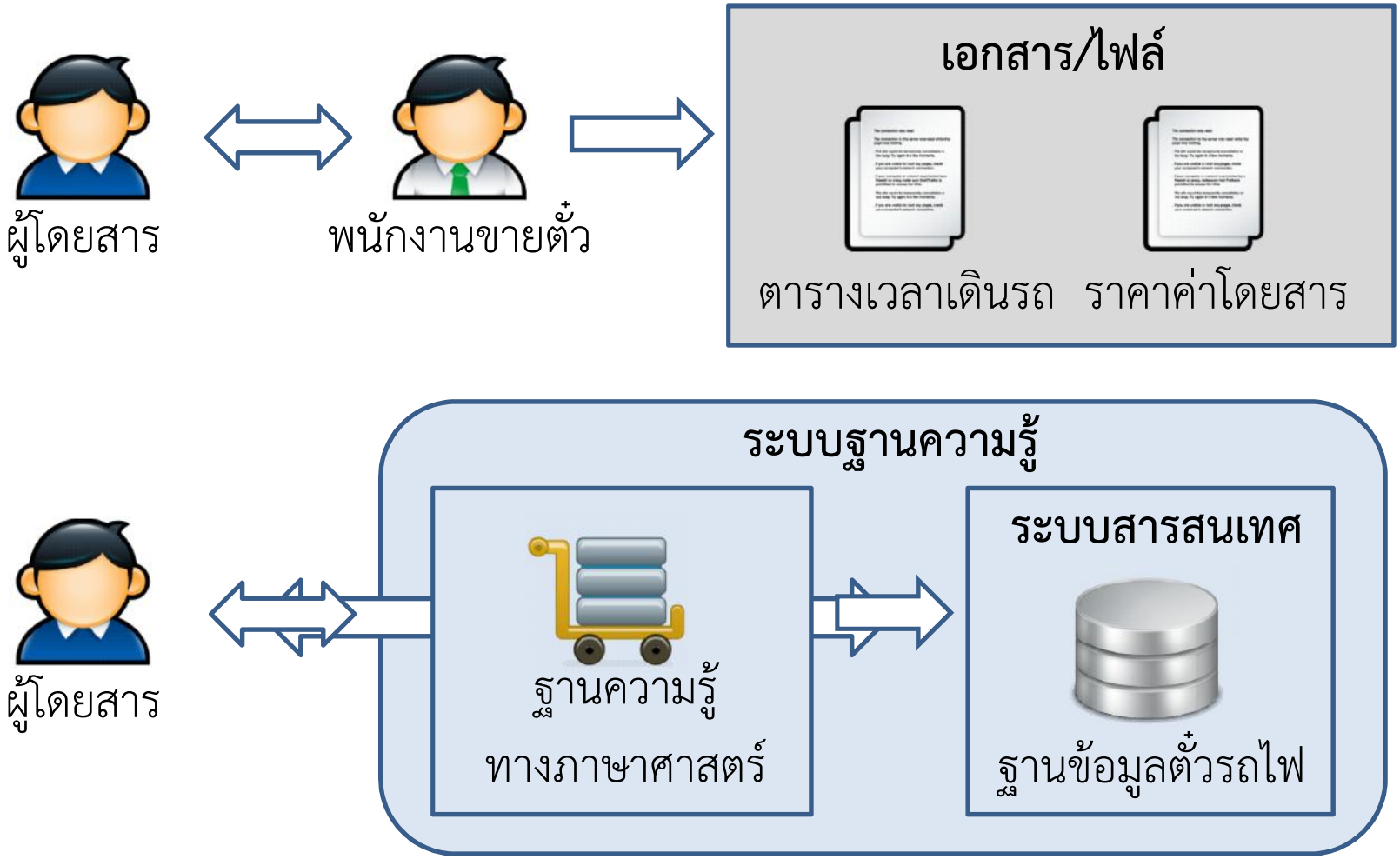


ระบบภาษาธรรมชาติ

ระบบภาษาธรรมชาติ

Input	Process	Output
• ใช้การสั่ง/ติดต่อกับคอมพิวเตอร์ด้วยภาษาธรรมชาติ	• ใช้หลักการพื้นฐานของฐานความรู้เกี่ยวกับไวยากรณ์ ความหมาย และความเข้าใจของภาษาธรรมชาติ	• แสดงผล/ติดต่อกับคอมพิวเตอร์ด้วยภาษาธรรมชาติ

ตัวอย่างระบบสอบถามตั๋วรถไฟ



กระบวนการวิเคราะห์

- รับข้อมูลที่เป็นข้อความต่อเนื่อง
- วิเคราะห์ระดับคำ
- วิเคราะห์ชนิดคำ
- วิเคราะห์โครงสร้างไวยากรณ์
- วิเคราะห์ความหมายของข้อความ
- ใช้ฐานความรู้ เพื่อจัดเก็บเข้าคอมพิวเตอร์
- เรียกใช้ฐานข้อมูลเพื่อแสดงผลเป็นไปตามเป้าหมายที่ต้องการ

องค์ประกอบของระบบ

- การวิเคราะห์ในเชิงโครงสร้าง (Syntactic Analysis)
  - เป็นการตรวจสอบโครงสร้างทางไวยากรณ์เกี่ยวกับการวางตำแหน่งของกลุ่มคำประเภทต่างๆ ที่รวมกันเป็นประโยค
  - ในกรณีที่ประโยคอินพุตที่รับเข้ามาไม่ถูกต้องตามหลักไวยากรณ์ คอมพิวเตอร์ควรจะบอกได้ว่าเป็นประโยคที่ผิด

ประโยค “The man old cried” ← ประโยคที่มีโครงสร้างผิดหลักไวยากรณ์
ลำดับที่ถูกจะต้องเป็น “The old man cried”

องค์ประกอบของระบบ (ต่อ)

- การวิเคราะห์ในเชิงความหมาย (Semantic Analysis)
  - เป็นการตรวจสอบความถูกต้องในเชิงความหมายของประโยค โดยประโยคที่วางกลุ่มคำชนิดต่างๆ ตามโครงสร้างไวยากรณ์จะมีความหมายอย่างใดอย่างหนึ่งเท่านั้น
  - ในบางครั้งประโยคที่กำลังพิจารณาอาจจะเขียนถูกต้องตามหลักไวยากรณ์ แต่มีความหมายกำกวมหรือเป็นความหมายที่เป็นไป ไม่ได้หรือไม่ให้ความหมายอะไรเลย

"The stones eat the boys" โครงสร้างของประโยคถูกต้องตามหลักไวยากรณ์ เมื่อวิเคราะห์ดูความหมาย จะเห็นว่าประโยคนี้มีความหมายที่เป็น ไปไม่ได้ เพราะหินเป็นสิ่งไม่มีชีวิตจึงทำกริยา"กิน"ไม่ได้

องค์ประกอบของระบบ (ต่อ)

- การวิเคราะห์ในเชิงตีความ (Pragmatic Analysis)
  - ประโยคที่เราพูดออกมาบางครั้งก็อาจจะไม่ได้มีความหมายตรงตามข้อความนั้นๆ ซึ่งจะต้องตีความตามสถานการณ์ที่เกิดขึ้นโดยที่ทั้งผู้ส่งข่าวสารและผู้รับข่าวสารจะ ต้องอยู่ในสถานการณ์เดียวกัน

สมมติ ว่าตอนนี้เราอยู่ที่สถานีรถไฟและกำลังกังวล ว่าขณะนี้เวลาเท่าไรเราเลยหันไปถามว่า “Do you have a watch?”
ถ้าเราได้คำตอบว่า “yes” หรือ “no” แสดงว่าคำตอบที่ได้ผิด เพราะคำตอบที่เราต้องการจริง ๆ คือ เวลา ณ ขณะนี้

ปัญหาและข้อจำกัดของ NLP

- ความกำกวมของภาษา (Ambiguity)
  - คำแต่ละคำอาจจะตีความได้แตกต่างกัน ถ้าอยู่ในประโยคที่มีบริบทแตกต่างกัน
  - “ตากลม” อาจเป็น ตา-กลม หรือ ตาก-ลม
  - “I see a man with a telescope” อาจหมายถึง see with a telescope หรือ a man with a telescope
- การใช้คำอ้างอิง/คำสรรพนาม (Reference/Pronoun)
  - คำอ้างอิงหรือคำสรรพนามอาจไม่ชัดเจน ทำให้ตีความได้หลายแบบ
  - “After putting the disk in the cabinet, John sold ***it***” อาจหมายถึง disk หรือ cabinet

ปัญหาและข้อจำกัดของ NLP

- การใช้ประโยคย่อ (Ellipsis)
  - การเขียนประโยคสามารถละคำหรือกลุ่มคำ เพื่อหลีกเลี่ยงการใช้คำซ้ำซ้อน ทำให้ต้องหา “คำที่ละ” ในประโยคว่าหมายถึงอะไร
  - What is the length of river Kwai? of river Khong?
  - สมศรีพาลูกไปงานเลี้ยง ส่วนสมทรงพาหลานไป (ละคำว่า “งานเลี้ยง” ไว้)
- อุปมาอุปมัย (Metaphor and Metonymy)
  - เป็นลักษณะประโยคที่ไม่สามารถตีความหมายตามตัวอักษรได้โดยตรง
- การใช้ภาษาอย่างไม่สมบูรณ์ (Ill-formedness)
  - การใช้ภาษาที่ผิดหลักไวยากรณ์ ทำให้ไม่สามารถตีความหมายได้

การประยุกต์ใช้งาน

Machine Translation

- การแปลภาษาด้วยเครื่องคอมพิวเตอร์
  - เป็นการนำเครื่องคอมพิวเตอร์เข้ามาใช้ในการแปลจากภาษาหนึ่งเป็นอีก ภาษาหนึ่งตามที่ต้องการ
    - แปลระหว่างภาษาอังกฤษกับภาษาฟินนิช http://www.sunda.fi/eng/translator.html
    - Google translation ใช้เครื่องแปลแบบใช้สถิติ (Statistical machine translation; SMT) ใช้ตัวอย่างคู่ภาษาที่แปลแล้วจำนวนมากร่วมกับหลักการทางสถิติมาช่วยในการตัดสินใจ
    - ทดลองทำเล่นๆ ดูได้โดยใช้โปรแกรม Moses และ Giza ร่วมกับโปรแกรมตัดคำไทย http://demo.statmt.org/

Information Retrieval

- การสืบค้นข้อมูลโดยใช้ภาษาธรรมชาติ
  - เป็นลักษณะของการสืบค้นข้อมูลในฐานข้อมูลด้วย ภาษาที่ใช้ในชีวิตประจำวันแทนการใช้ภาษาคอมพิวเตอร์
    - ถ้าต้องการข้อมูลวัดที่มีชื่อเสียงในประเทศไทยจากฐานข้อมูล แทนที่จะต้องมาเขียนคำสั่ง SQL ก็สามารถใช้คำถามที่เป็นภาษา ธรรมชาติได้ดังนี้ What temple is famous in Thailand? โดย จะต้องมีการแปลงจากภาษาธรรมชาติที่ใส่เข้าไปให้อยู่ในรูปของภาษา SQL ที่เครื่องเข้าใจ
    - การนำ ontology มาช่วยให้คอมพิวเตอร์เข้าใจความหมายของคำ เพื่อช่วยในการค้นคืนข้อมูลที่เกี่ยวข้อง

Text Categorization

- การแบ่งประเภทข้อมูล
  - ทำให้สามารถเก็บข้อมูลเอกสารได้อย่างเป็นระเบียบ เพิ่มประสิทธิภาพในการค้นหาข้อมูล
    - การจัดกลุ่มเอกสารโดยอาศัย keywords หรือกลุ่มคำเป็นตัวแทนของเอกสาร โดยเก็บในรูป vector และใช้ vector นี้ในการคำนวณหาความคล้ายกันระหว่างระหว่างเอกสาร ถ้ามีความคล้ายกันก็จัดรวมกลุ่มเอกสารเข้าด้วยกัน ถ้ามีความห่างกันมากก็แยกเป็นคนละกลุ่ม

Text Summarization

- การย่อความ
  - เป็นการสรุปใจความสำคัญจากเอกสารเอาเฉพาะส่วนที่สำคัญ สร้างเป็นเอกสารใหม่ที่สั้นกว่าเดิมแต่มีเนื้อหาข้อมูลที่เหมือน ต้นฉบับโดยใช้วิธีการต่างๆ
    - Text Extraction Technique วิธีนี้จะมีการให้คะแนนประโยคหรือข้อความโดยดูจากความสำคัญ ของประโยคหรือข้อความและนำมาสร้างเป็นเอกสารใหม่โดยคะแนน ที่ให้อาจจะพิจารณาจากความถี่ของคำที่ปรากฏในเอกสาร การเน้น ข้อความหรือตำแหน่งของคำ เป็นต้น

Question & Answering

- ระบบถามตอบอัตโนมัติ
  - เป็นระบบที่ผู้ใช้สามารถถามด้วยภาษาธรรมชาติ และระบบทำการประมวลผลเพื่อหาคำตอบ และตอบกลับด้วยภาษาธรรมชาติเช่นกัน
    - ELIZA โปรแกรมสนทนาโต้ตอบกับมนุษย์ในเรื่องใดๆ ก็ได้ ใช้หลักการเปรียบเทียบคำเฉพาะ (Keyword matching) ไม่เข้าใจประโยคอย่างแท้จริง
    - ABDUL สามารถช่วยตอบคำถามที่ผู้ถามอยากรู้คำตอบแบบโต้ตอบกันทันทีผ่านทาง MSN
    - ศิราณี ศรีสยาม พูดคุย ถามตอบได้ผ่าน MSN

Sign Language Processing

- ระบบประมวลผลภาษามือ
  - คล้ายกับระบบแปลภาษา แต่จัดการกับข้อมูลที่เป็น 3 มิติ
  - แบ่งได้เป็น 2 ขั้นตอนหลัก คือ การแปลงโครงสร้างระหว่างภาษาธรรมชาติกับภาษามือ และ การจัดการกับการรับข้อมูลภาพ
    - การวิเคราะห์ประโยค
    - การวิเคราะห์คำ
    - การเลือกคำสำหรับไวยากรณ์ภาษามือ
    - การเรียงคำในรูปไวยากรณ์ภาษามือ
    - การแสดงผล

Sign Language Processing

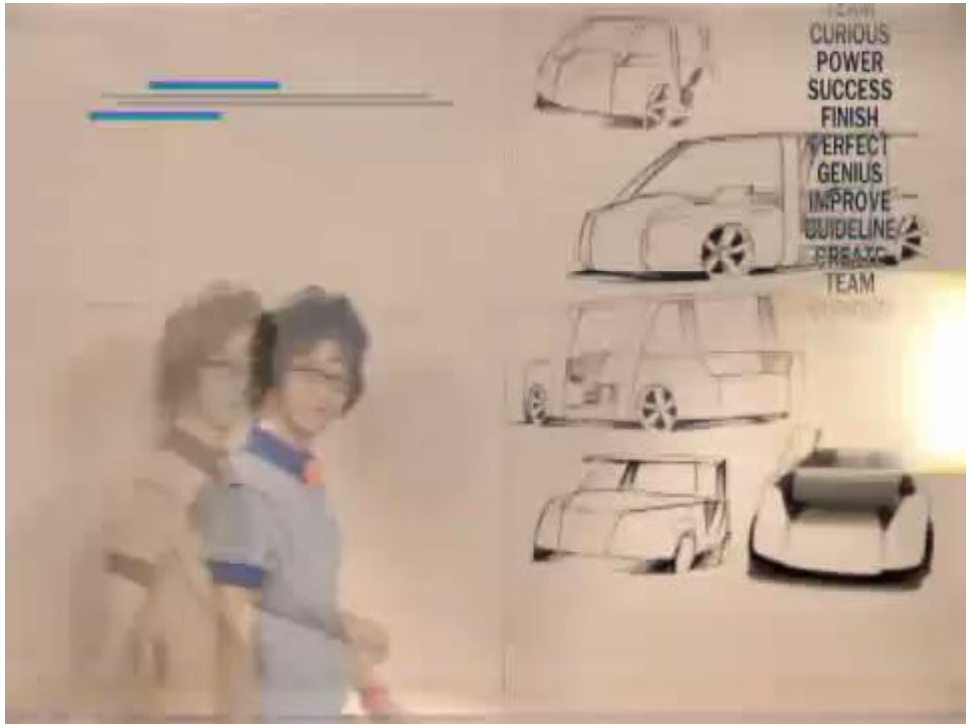
- โปรแกรมช่วยเหลือการสื่อสารผ่านภาษามือ (Sign Language Communication Translator: SLCT) คณะเทคโนโลยีสารสนเทศ สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

  http://www.dailynews.co.th/newstartpage/index.cfm?page=content&categoryID=478&contentID=153300
- โปรแกรมแปลงภาษาไทยเป็นภาษามือไทย 3 มิติ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเชียงใหม่

  http://www.youtube.com/watch?v=07WnWR2Ji9A
- ระบบแปลอัจฉริยะภาษาไทย-ภาษามือไทย เพื่อช่วยในการเรียนรู้ภาษา โดย ดร. ศรีสวคนธ์ แดงสอาด มหาวิทยาลัยราชภัฏธนบุรี

  http://www.il.mahidol.ac.th/km/images/stories/articles/innovation/inno_9_1.pdf

Sign Language Processing



Speech Processing

- ระบบประมวลผลเสียงพูด
  - การสังเคราะห์เสียงพูด speech synthesis
    - การตัดคำ
    - แปลงเป็นสัญลักษณ์คำอ่าน (phoneme sequence)
    - การสังเคราะห์เสียง
  - การรู้จำเสียงพูด speech recognition
  - ระบบสอบถามข้อมูลทางโทรศัพท์อัตโนมัติ (IVR)

การประมวลผลภาษาไทย

- โปรแกรมการเรียงลำดับคำไทย
- โปรแกรมตัดพยางค์/คำภาษาไทย
- โปรแกรมการสืบค้นคำไทย
- โปรแกรมการสืบค้นคำไทยตามเสียงอ่าน
- โปรแกรมแปลภาษา
- โปรแกรมตรวจสอบตัวสะกดและไวยากรณ์
- โปรแกรมสังเคราะห์เสียง

แนวทางการวิจัย

- การประมวลผลคำ (Word processing)
- การประมวลผลข้อความ (Text processing)
- การประมวลผลเสียงพูด (Speech processing)
- การจัดการข้อมูลสารสนเทศ (Information management)
- ระบบสารสนเทศอัจฉริยะ (Intelligent Information Systems)
- การสร้างทรัพยากรและเครื่องมือ (Language resources and tools)

(อ้างอิงจาก http://www.hlt.nectec.or.th/hlt/index.php/about-hlt)

References

- ธนารักษ์ ธีระมั่นคง, “เทคโนโลยี่การประมวลผลภาษาธรรมชาติในปัจจุบัน” http://www.jaist.ac.jp/~ping/paper/NLP_tis.txt
- ยืน ภู่วรวรรณ, “การประมวลผลภาษาธรรมชาติ”
- http://guru.sanook.com/search/natural_language_processing/
- http://ict.siit.tu.ac.th/kindml/thainest/