การศึกษาภาษาไทยเพื่องานทางภาษาศาสตร์คอมพิวเตอร์

วิโรจน์ อรุณมานะกุล

เอกสารประกอบการบรรยาย
ในการประชุมวิชาการเรื่องภาษาศาสตร์ภาษาไท-ไทย 5-7 ก.ย. 2545
ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
เพื่อเป็นเกียรติแด่ รองศาสตราจารย์ ดร. ปราณี กุลละวณิชย์ เนื่องในโอกาสเกษียณอายุราชการ

งานทางภาษาศาสตร์คอมพิวเตอร์คืออะไร

ถ้ากล่าวโดยสรุป งานทางภาษาศาสตร์คอมพิวเตอร์คืองานที่คอมพิวเตอร์ทำที่จะต้องเกี่ยวข้องกับการประมวลผลภาษา (Natural Language Processing หรือ NLP) ซึ่งถ้ามองในลักษณะนี้ ก็จะมีคำถามต่อว่า มีความเกี่ยวข้องกับการประมวลผลภาษาในระดับมากน้อยเพียงใด โปรแกรม Word Processing เกี่ยวข้องไหม ก็เกี่ยวข้องในระดับหนึ่งถ้าเรามองว่า ในโปรแกรม Word ภาษาไทยจะมีการปัดคำลงบรรทัดต่อไปเวลาที่พิมพ์ข้อความเกินขอบขวา หรืออาจมีการช่วยตรวจสอบตัวสะกด (spell check) หรือไวยากรณ์ (grammar check) ส่วนงานที่เกี่ยวข้องกับการประมวลภาษาในระดับมาก เช่น การแปลภาษาด้วยเครื่องคอมพิวเตอร์ (Machine Translation) เพราะการแปลภาษาหนึ่งไปเป็นอีกภาษาซึ่งต้องอาศัยการประมวลผลภาษาอย่างมาก มีขั้นตอนที่ซับซ้อนเพื่อบรรลุผลที่ต้องการ งานด้านการค้นคืนสารสนเทศ (Information Retrieval) ถ้าใช้วิธีการเพียงแค่เทียบคำให้ตรงกับคำค้น (keyword) ก็ไม่ต้องใช้ความรู้อะไรทางภาษา แต่ถ้าต้องการค้นโดยใช้คำที่มีความหมายพ้อง เสียงคล้าย หรือเป็นเรื่องที่เกี่ยวข้องกัน แบบนี้ก็ต้องมีกระบวนการในการประมวลผลภาษาในระดับหนึ่ง

ทำไมจึงต้องการความรู้ภาษาไทย

คำตอบง่ายๆ ก็คือ เพราะงาน NLP นั้นขึ้นกับภาษา งานอย่างเช่น การแปลภาษาด้วยเครื่อง (MT), การค้นคืนสารสนเทศ (IR) ต้องอาศัยการวิเคราะห์ภาษาไทย ซึ่งอย่างน้อยก็ต้องมีการแยกคำในข้อมูลได้ มองเห็นความสัมพันธ์ระดับต่างๆของคำได้ เราจึงไม่สามารถนำระบบ NLP ของภาษาอื่นมาใช้ได้โดยตรง เราไม่สามารถซื้อระบบของภาษาอังกฤษหรือภาษาญี่ปุ่นมาแล้วดัดแปลงใช้ได้ ถ้าทำเพียงแค่ดัดแปลงให้รับข้อมูลเข้า (input) หรือแสดงผลลัพธ์ (output) เป็นภาษาไทยก็อาจทำได้ แบบที่เคยทำกันมา แต่ถ้าต้องการให้มีการประมวลผลภาษาไทย เราก็ต้องพัฒนาระบบขึ้นมาเองหรือปรับระบบให้เข้ากับภาษาไทย ซึ่งต้องทำให้ดูเสมือนว่าคอมพิวเตอร์นั้นสามารถรู้จักคำไทย ไวยากรณ์ไทย และสามารถตีความประโยคภาษาไทย

ความรู้ภาษาไทยที่ต้องการมีอะไรบ้าง เป็นลักษณะใด

ถ้าตอบแบบง่ายที่สุดก็คือทุกอย่างที่เกี่ยวข้องกับภาษา แต่ในสภาพปัจจุบัน การศึกษาทางภาษาศาสตร์นั้นค่อนข้างมีความหลากหลายมาก เรื่องบางเรื่องก็อาจจะเป็นสิ่งที่ยังห่างจากความต้องการของการพัฒนาระบบ NLP ในปัจจุบัน อย่างน้อยก็ไม่ใช่ในขั้นตอนเริ่มแรกนี้ เช่น เรื่องทางภาษาศาสตร์สังคม เรื่องทางภาษาศาสตร์เชิงประวัติ หรือแม้แต่งานด้านวัจนปฏิบัติศาสตร์ (pragmatics) เอง ส่วนใหญ่ก็ยังคงเป็นสิ่งที่ไม่สามารถนำมาใช้ประโยชน์ในทางคอมพิวเตอร์ได้ในปัจจุบัน ในขั้นแรกนี้ความต้องการของงานทาง NLP ยัง

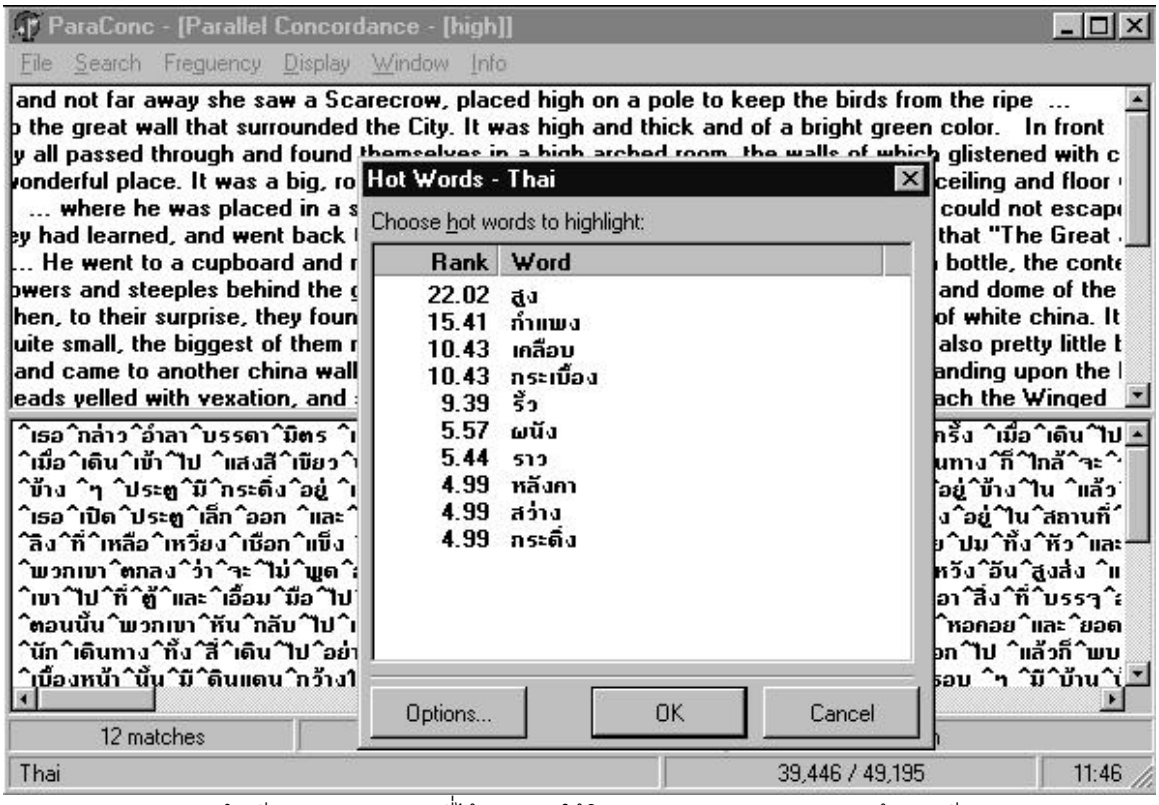
เป็นเพียงการประมวลผลข้อความ (process text) แบบพื้นฐาน ซึ่งเพียงแค่ต้องการให้คอมพิวเตอร์เข้าใจว่าในข้อความที่ได้รับมามี คำ วลี ประโยค อะไรบ้าง มีหมายความว่าอย่างไร ก็ไม่ใช่เรื่องง่ายแล้ว

ก่อนที่เราจะเข้าใจว่าความรู้ภาษาไทยที่นักภาษาศาสตร์คอมพิวเตอร์ต้องการนั้นมีลักษณะแบบใด เนื่องจากว่าสาขานี้เป็นสาขาที่เกี่ยวข้องทั้งทางภาษาศาสตร์และทางคอมพิวเตอร์ ทางภาษาศาสตร์เรียกว่า "" ภาษาศาสตร์คอมพิวเตอร์" (Computational Linguistics) ส่วนทางคอมพิวเตอร์เรียกว่า "การประมวลผลภาษาธรรมชาติ" (Natural Language Processing) หรือปัจจุบันมีการใช้คำว่า "วิศวกรรมภาษา" (Language Engineering) โดยทางคอมพิวเตอร์มีคำเรียกผู้ที่ทำงานด้านการประมวลผลภาษาโดยเฉพาะว่านักวิศวกรภาษา (language engineer) ดังนั้น เราลองมาดูก่อนว่าปัจจุบัน ระบบ NLP ที่พัฒนาขึ้นมาโดยนักวิศวกรภาษานั้นมีลักษณะอย่างไร

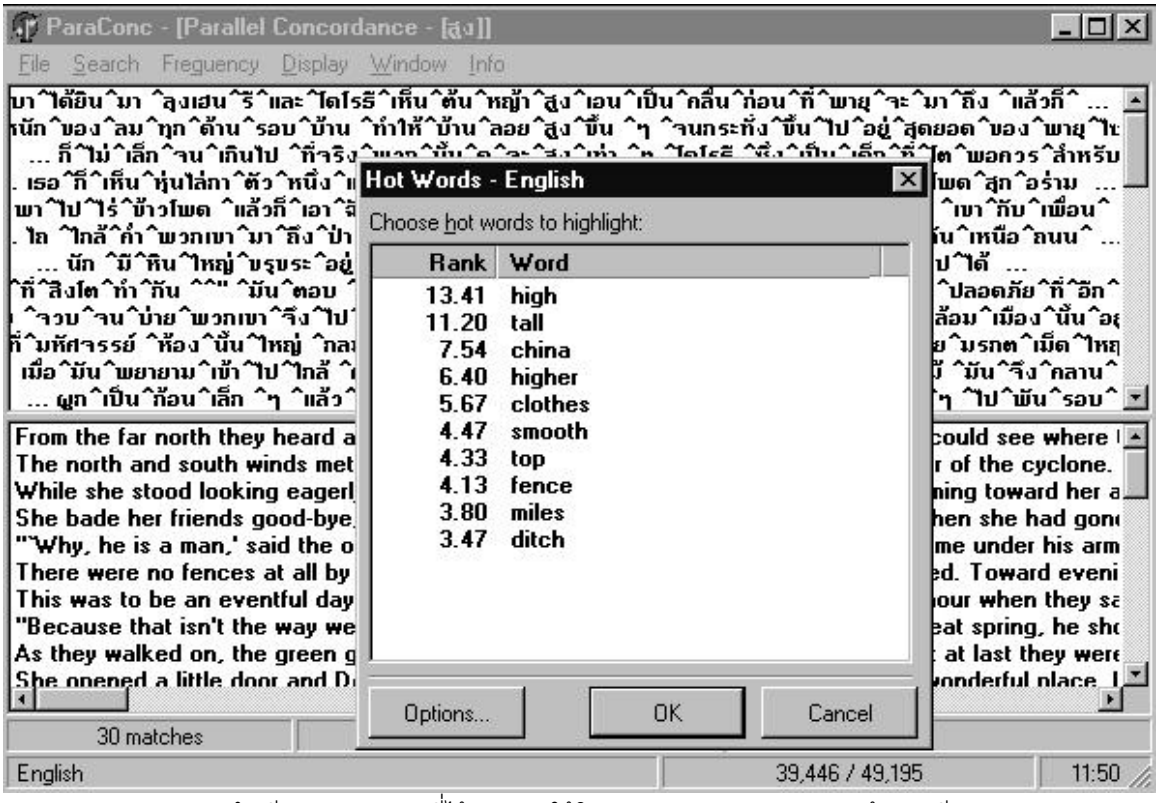
ระบบ NLP ที่พัฒนาโดยนักวิศวกรภาษาเป็นอย่างไร

ตัวอย่างงานที่นักวิศวกรภาษาศึกษา เช่น เรื่องของการตัดคำ การกำกับหมวดคำ การหาคำเทียบเท่าจากคลังข้อมูลเทียบบท (parallel corpora) การหาคำสำคัญในเอกสาร (keyword extraction) ในเรื่องของการประมวลผลข้อความ (text processing) นักวิศวกรภาษาจะมองข้อความเป็นสายอักขระ (string sequence) ในลักษณะที่เป็นตัวอักษรเรียงติดต่อกันไป และใช้วิธีการทางสถิติเป็นสำคัญในการทำงาน ถ้าเราพิจารณาแต่ภาพที่เห็น เราอาจประหลาดใจว่าคอมพิวเตอร์สามารถรู้และทำงานนั้นได้อย่างไร ตัวอย่างเช่น การให้คอมพิวเตอร์หาคำเทียบเท่าหรือคำแปลจากคลังข้อมูลเทียบบทในตัวอย่างนี้[1] เมื่อหาคำว่า "high" คอมพิวเตอร์ก็จะแสดงรายการคำเทียบเท่าซึ่งมีคำไทยว่า "สูง" เป็นตัวเลือกแรก หรือถ้าทำในทางตรงข้าม ถ้าหาคำไทยว่า "สูง" คอมพิวเตอร์ก็จะแสดงรายการคำเทียบเท่าซึ่งมีทั้งคำว่า "high" และ "tall" ในลำดับแรกเช่นกัน หรือในการให้คอมพิวเตอร์ถ่ายเสียงภาษาไทย อย่างเช่น ประโยคว่า "เป็นมนุษย์สุดประเสริฐเลิศคุณค่ากว่าบรรดาฝูงสัตว์เดรัจฉานจงฝ่าฟันพัฒนาวิชาการอย่าล้างผลาญฤาเข่นฆ่าบีฑาใคร" ระบบที่ทำเมื่อใช้เฉพาะกฎทางอักขรวิธีภาษาไทยจะแสดงผลการอ่านถึง 1,728 แบบ ซึ่งลักษณะความกำกวมที่เกิดขึ้นนี้จะพบได้เสมอ และแสดงให้เห็นถึงปัญหาที่เกิดขึ้นจากการประยุกต์ใช้กฎทางภาษากับคอมพิวเตอร์ แต่เมื่อใช้วิธีการทางสถิติอย่างแบบจำลองไตรแกรม ซึ่งอาศัยหลักง่ายๆ คือพิจารณาความน่าจะเป็นที่จะพบพยางค์ต่างๆ ทีละสามพยางค์ไปเรื่อยๆ เข้ามาช่วยในการตัดสินใจ ว่าในบรรดาตัวเลือก 1,728 ตัวเลือกในที่นี้นั้น ตัวเลือกใดเป็นตัวเลือกที่มีโอกาสพบได้จริงมากที่สุด ระบบก็สามารถเลือกคำอ่านที่เราต้องการออกมาได้

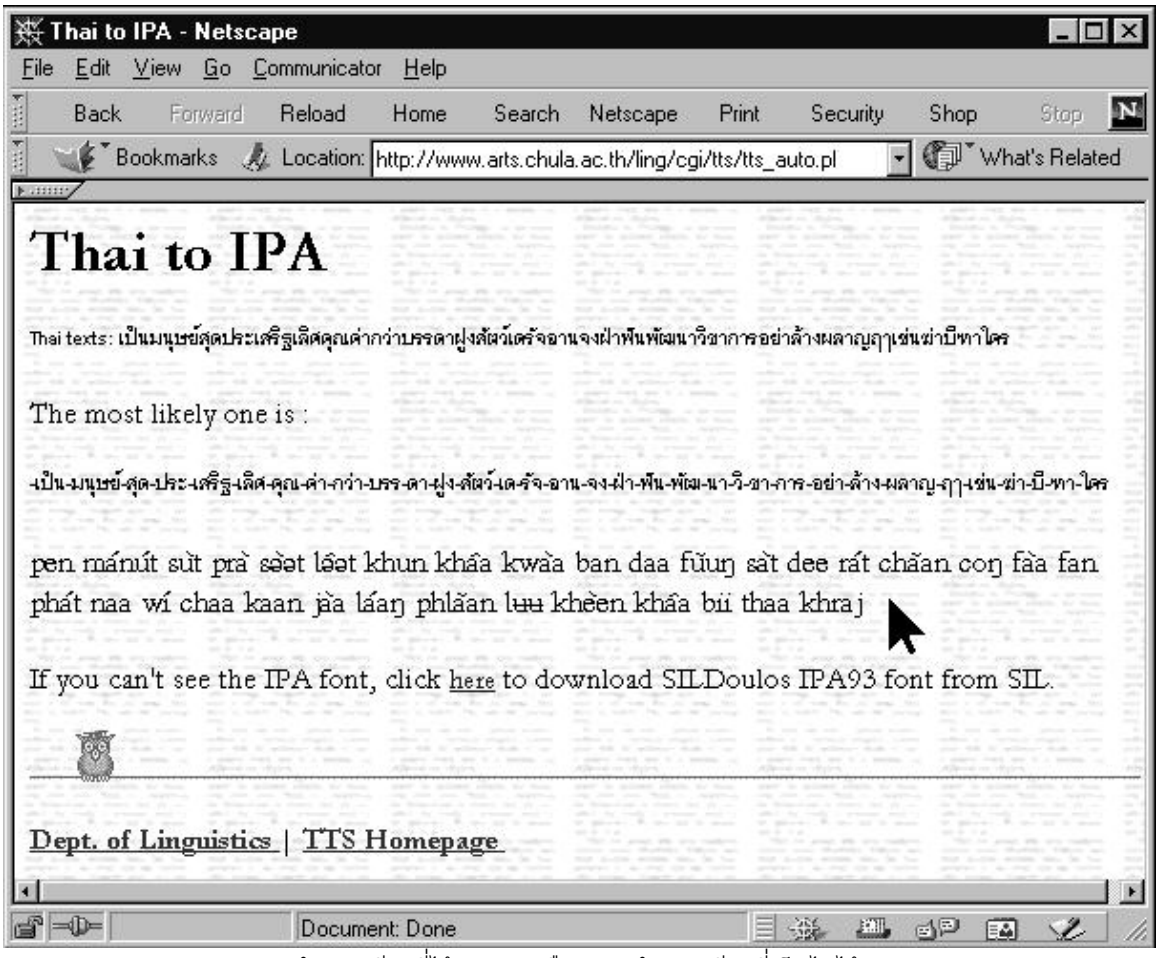
[1] โปรแกรมที่ใช้ในที่นี้คือโปรแกรม ParaConc ของ Michael Barlow (http://athel.com/)



รูป 1 : แสดงคำเทียบเท่าของ high ที่ได้จากการใช้โปรแกรม ParaConc กับคลังข้อมูลเทียบบท



รูป 2 : แสดงคำเทียบเท่าของ สูง ที่ได้จากการใช้โปรแกรม ParaConc กับคลังข้อมูลเทียบบท



รูป 3 : แสดงผลคำถ่ายเสียงที่ได้จากการเลือกจากคำถ่ายเสียงที่เป็นไปได้ 1,728 แบบ

ดังนั้น เบื้องหลังสิ่งที่ดูเสมือนว่าคอมพิวเตอร์นั้นฉลาด แท้จริงอาจไม่ใช่ความฉลาดในแบบที่คนเรามีหรือเป็นก็ได้ เราคิดว่าคอมพิวเตอร์ฉลาดเพราะมันทำในสิ่งที่ถ้ามนุษย์เป็นผู้ทำก็ต้องอาศัยปัญญาความรู้ในการแก้ปัญหานั้น คือเราตัดสินโดยดูเพียงพฤติกรรมที่เห็น แต่เมื่อดูที่รายละเอียดแล้วก็จะเห็นว่าไม่ได้มีอะไรมากไปกว่าการคำนวณค่าทางสถิติเพื่อช่วยในการเดา หลายคนอาจตั้งข้อสงสัยว่า ถ้าใช้แต่การคำนวณค่าความน่าจะเป็นด้วยวิธีการทางสถิติ ก็ไม่เห็นจะเกี่ยวข้องกับภาษาศาสตร์แต่อย่างไร ประเด็นในเรื่องนี้นั้น Abney (1996) ได้เขียนบทความเรื่อง “Statistical methods and linguistics” เพื่อชี้ให้เห็นว่า ความจริงแล้วเรื่องทางสถิติก็มีส่วนเกี่ยวข้องโดยตรงกับการศึกษาทางภาษาศาสตร์ การพยายามมองกฎไวยากรณ์ในลักษณะที่เป็น discrete เพื่อตัดสินว่าประโยคใดถูกไวยากรณ์ (grammatical) หรือไม่ถูกไวยากรณ์ (ungrammatical) อาจไม่ใช่แนวคิดที่เหมาะสม ในปัจจุบัน ก็เริ่มมีผู้พยายามพัฒนาทฤษฎีไวยากรณ์ในลักษณะที่ผนวกเอาเรื่องสถิติของความน่าจะเป็นเข้าไว้ด้วย และแนวคิดของนักภาษาศาสตร์ที่ทำงานโดยอาศัยคลังข้อมูลภาษา ก็มักจะให้ความสำคัญกับเรื่องความถี่ของรูปแบบที่พบ มองว่ามีความสัมพันธ์ลักษณะเป็นกลุ่มก้อน (chunk) หรือกลุ่มรูปแบบที่ปรากฏบ่อยเป็นปกติ

เหตุผลหนึ่งที่วิธีการทางสถิติสามารถนำมาใช้อย่างได้ผลในการประมวลผลภาษา อาจเป็นเพราะในภาษานั้นมี “ข้อมูล” ที่ซ้ำซ้อน (redundant) อยู่เป็นจำนวนมาก เราจะเห็นว่าเวลาที่เราฟังคนพูดในที่เสียงดัง เราอาจไม่ได้ยินคำครบทุกคำ แต่ก็สามารถเดาคำที่ไม่ได้ยินได้โดยอาศัยข้อมูลจากคำข้างเคียง หรือกรณีอ่าน

ข้อความที่พิมพ์ผิด เราก็สามารถใช้ข้อมูลข้างเคียงช่วยในการแก้ไขข้อมูลให้ถูกต้องได้ และด้วยคุณสมบัติที่ภาษามีความซ้ำซ้อน (redundancy) อยู่ทำให้เราสามารถบีบอัดแฟ้มข้อมูล (compress file) ที่พิมพ์ข้อความเพื่อลดขนาดของข้อมูลลงได้ หรือกล่าวอีกนัยหนึ่ง การบีบอัดแฟ้มข้อมูล คือตัวอย่างที่แสดงให้เห็นถึงความซ้ำซ้อนที่มีอยู่ในภาษา เมื่อขจัดความซ้ำซ้อนออกจากข้อความที่พิมพ์เข้าไปโดยที่ยังคงปริมาณสารสนเทศ (information) เหมือนอย่างเดิมไว้[2] ขนาดของแฟ้มข้อมูลก็จะเล็กลง ตัวอย่างเช่นแฟ้มข้อมูลที่เก็บเฉพาะข้อความภาษาไทยที่มีขนาด 409 KBytes เมื่อบีบอัดด้วยโปรแกรม WinZip แล้วจะเหลือขนาดเพียง 166 KBytes หรือลดลงประมาณ 40%

การมีข้อมูลที่ซ้ำซ้อนนั้นมาจากการที่เราสามารถคาดเดาได้ว่าถ้าพบข้อมูลคำนี้คำต่อไปที่ควรจะเป็นน่าจะเป็นคำใดได้บ้าง เช่น หนังสือสอง.... คำที่เราจะนึกถึงคือ เล่ม ตั้ง พัน ร้อย เป็นต้น คือจะมีคำชุดหนึ่งเท่านั้นที่ข้อความนี้กระตุ้นให้เรานึกถึง ลักษณะที่สามารถคาดเดาได้ที่พบในภาษานี้ ทำให้แนวคิดในเรื่องของความน่าจะเป็นหรือการใช้วิธีการทางสถิติสามารถนำมาใช้ช่วยในการพัฒนาระบบ NLP ได้ นักวิศวกรภาษาบางคนก็เชื่อว่า หากเราป้อนข้อมูลภาษาจำนวนมากๆให้กับระบบคอมพิวเตอร์ที่ถูกออกแบบมาให้เรียนรู้ที่จะดึงข้อมูลทางสถิติของภาษา ระบบคอมพิวเตอร์นั้นก็จะสามารถประมวลผลภาษาตามที่ต้องการได้ คำถามคือถ้าเราใช้แต่สถิติก็เพียงพอ นักภาษาศาสตร์จะมีบทบาทอะไร จำเป็นต้องมีทฤษฎีภาษาศาสตร์ หรือจำเป็นต้องมีทฤษฎีไวยากรณ์อีกหรือไม่

บทบาทของนักภาษาศาสตร์

นักวิศวกรภาษาจะสนใจอยู่ที่การหาวิธีการต่างๆเพื่อทำให้ระบบ NLP นั้นสามารถทำงานอย่างที่ต้องการได้ บางครั้งก็หาเหตุผลอธิบายในสิ่งที่ทำไม่ได้ รู้แต่ว่าถ้าใช้วิธีการนี้แล้วจะได้ผลดีกว่าอีกวิธีการหนึ่ง ตัวอย่างเช่น ในเรื่องการตัดคำ เริ่มแรกนั้น ทำโดยการพยายามหากฎทางอักขรวิธีที่จะใช้ตัดสิน เช่น สระอำจะไม่มีตัวสะกดตาม แต่กฎที่หาได้ช่วยในเรื่องการแยกพยางค์เท่านั้น ต่อมาจึงมีความคิดให้เทียบตัวอักษรกับคำในพจนานุกรม ซึ่งก็เริ่มต้นด้วยวิธีเทียบคำให้ยาวที่สุดก่อนโดยทำจากซ้ายไปขวา หรือที่เรียกว่าการตัดคำแบบ longest matching ซึ่งก็ยังมีข้อผิดพลาด เลยมีวิธีเทียบโดยให้ได้จำนวนคำทั้งหมดน้อยที่สุด หรือที่เรียกว่าการตัดคำแบบ maximum matching (วิรัช 2536) ซึ่งก็ยังมีข้อผิดพลาดอยู่ จึงมีผู้เสนอวิธีการใหม่ๆมาทดลอง งานส่วนใหญ่จะเน้นที่วิธีการเทคนิคต่างๆ แต่มักจะไม่ได้พูดถึงเหตุผลเบื้องหลังที่ซ่อนอยู่ของแต่ละวิธี หรืออภิปรายว่าทำไมจึงได้ผล ทำไมจึงไม่ได้ผล เช่น วิธีการตัดคำแบบ maximum matching ถ้าพิจารณาดูจะเห็นข้อสันนิษฐานพื้นฐาน (basic assumption) คือ เชื่อว่าคำมีลักษณะที่เป็นคำประกอบมากกว่าที่จะเป็นคำโดดถามว่าจริงไหม ถ้าดูจากผลการตัดคำด้วยวิธีนี้ที่ได้ผลมากกว่า 80% ก็แสดงว่าข้อสันนิษฐานนี้น่าจะเป็นจริง แต่นักวิศวกรภาษาจะไม่สนใจศึกษาประเด็นเช่นนี้ หากเราจะพิสูจน์ก็ต้องแจงรายการคำทั้งหมดออกมาแล้วดูว่าคำในภาษาไทยมีลักษณะดังกล่าวหรือไม่ ในที่นี้ ผู้เขียนได้ทดลองให้คอมพิวเตอร์ตรวจดูคำแต่ละคำในพจนานุกรมว่าสามารถมองเห็นเป็นคำหลายๆคำประกอบกันได้หรือไม่ ก็พบว่ามีคำเป็นจำนวนมากที่เป็นเช่นนี้ จากคำ 33,046 คำในพจนานุกรม สามารถแจงออกมาได้ว่ามีคำซึ่งความจริงเป็นคำๆเดียว แต่คอมพิวเตอร์สามารถมองว่าประกอบด้วยคำหลายคำได้ 18,738 แบบ (จากรูปคำ 16,783 คำ) ดังนั้น จึงไม่แปลกที่การเลือกตัดคำให้มีจำนวนคำน้อยที่สุดจึงได้ผลในระดับหนึ่ง

---

[2] เพราะเราสามารถคลายการบีบอัด (decompress) แฟ้มข้อมูลกลับมาเป็นข้อความเดิมได้

กรรมเวร	กร รม เวร
กรอบแกรบ	กร อบ แก รบ
มุมตกกระทบ	มุม ตกกระ ทบ
มูลค่า	มูล ค่า
สั่งสอน	สั่ง สอน

ตัวอย่างคำที่สามารถมองแยกออกเป็นคำได้หลายคำ

ในที่นี้จะขอยกตัวอย่างอีกตัวอย่างหนึ่ง เพื่อให้เห็นภาพความสนใจของนักวิศวกรภาษา คือ งานในการดึงคำออกจากตัวบทภาษาไทยโดยอัตโนมัติ ซึ่งใช้วิธีการเรียนรู้จากคลังข้อมูลภาษา ในบทความวิจัยเรื่อง “Automatic Corpus-Based Thai Word Extraction with the C4.5 Learning Algorithm” (Potipiti et al. 2000) ผู้วิจัยใช้เทคนิคต่างๆในการดูสถิติเกี่ยวกับตัวอักษรภาษาไทยเพื่อที่จะให้โปรแกรมคอมพิวเตอร์ดึงคำต่างๆที่พบในเอกสารออกมาให้ ถามว่าคอมพิวเตอร์รู้จักไหมว่าคำคืออะไร คำตอบคือไม่รู้ สิ่งที่คอมพิวเตอร์เห็นคือตัวอักษรภาษาไทยเรียงต่อกันไป ทำอย่างไรจึงจะให้คอมพิวเตอร์บอกได้ว่าคำแต่ละคำเริ่มต้นที่ตัวอักษรใดและจบที่ตัวอักษรใด เทคนิคแรกที่ใช้ในงานวิจัยนี้คือ ใช้ค่าเอ็นโทรปี (entropy)[3] โดยดูจากสถิติของการเกิดร่วมกันของตัวอักษรว่าที่จุดนั้นๆ มีความไม่แน่นอนมากน้อยเพียงใด ค่าเอ็นโทรปีที่สูงแสดงว่ามีความไม่แน่นอนมาก เช่น ใน “ปรากฏ” เอ็นโทรปีของ “ปราก” น่าจะมีค่าต่ำ เพราะความไม่แน่นอนที่จะพบตัวอักษรอื่นๆ ตามหลัง “ปราก” นั้นต่ำ จุดที่จะเป็นจุดสิ้นสุดคำได้จึงควรจะมีค่าเอ็นโทรปีที่สูง เอ็นโทรปีของ “ปรากฏ” จะมีค่าสูง เพราะตัวอักษรต่อไปอาจเป็นอะไรก็ได้มากมาย นอกจากการใช้ค่าเอ็นโทรปี ในงานนี้ยังใช้ค่า mutual information ซึ่งเป็นวิธีการทางสถิติที่บอกความเกี่ยวพันระหว่างสองสิ่งว่ามีมากน้อยเพียงใด เช่น “กปรากฏ” จะมีค่า mutual information ระหว่าง “ก” และ “ปรากฏ” ต่ำ จุดที่จะเป็นขอบเขตคำได้จึงน่าจะมีค่า mutual information นี้ต่ำ

นอกจากการใช้ค่าของเอ็นโทรปีและ mutual information แล้ว ในงานวิจัยนี้ ยังรวมเอาเทคนิคอื่นๆ มาใช้ประกอบ เช่น กำหนดว่า ภายในคำจะต้องไม่มีส่วนที่เป็นคำไวยากรณ์ (function word) เช่น “จะ” “ก็” ใช้ความถี่ของชุดตัวอักษรที่พบ ใช้ความยาวของชุดตัวอักษรโดยกำหนดให้เลือกชุดที่ยาวก่อนชุดที่สั้น เป็นต้น แต่สิ่งที่นักวิศวกรภาษาไม่ได้อภิปรายในงานก็คือ ทำไมจึงเลือกใช้เทคนิควิธีการเหล่านั้น ทำไมเทคนิคที่ใช้จึงได้ผลหรือไม่ได้ผลอย่างไร มีความจำเป็นในการใช้ทั้งค่าเอ็นโทรปีและ mutual information เพียงใดในเมื่อ mutual information ก็พัฒนามาจากค่าเอ็นโทรปีอีกต่อหนึ่ง ค่าที่คำนวณได้จริงๆบอกขอบเขตของคำหรือขอบเขตของพยางค์ ทำไมการดูสถิติการปรากฏร่วมกันของตัวอักษรจึงใช้บอกขอบเขตคำได้ คำถามเหล่านี้ไม่เป็นที่สนใจของนักวิศวกรภาษาเพราะความสนใจหลักของนักวิศวกรภาษาอยู่ที่ทำอย่างไรจึงจะทำให้ระบบ NLP ทำงานที่ต้องการได้

หากจะเปรียบเทียบกับนักภาษาศาสตร์คอมพิวเตอร์ นักภาษาศาสตร์คอมพิวเตอร์ไม่ได้สนใจเพียงประเด็นการพัฒนาระบบให้ทำงานได้ แต่ยังสนใจศึกษาว่าทำไมวิธีการต่างๆที่ใช้ถึงได้ผลหรือไม่ได้ผล ซึ่งก็เป็นลักษณะเดียวกับนักภาษาศาสตร์ทั่วไป คือ สนใจอยากรู้ว่าทำไมด้วย ตรงนี้ มีคนเคยพูดเล่นไว้ว่า นักวิศวกร

[3] ดู Shannon 1948.

ภาษาเป็นผู้ที่สร้างระบบให้ทำงานได้แต่มักจะไม่รู้ว่าทำงานได้เพราะเหตุใด ส่วนนักภาษาศาสตร์คอมพิวเตอร์เป็นผู้ที่มักจะรู้ว่าทำไมระบบที่สร้างขึ้นมาถึงทำงานไม่ได้ (LE makes things work without knowing why. CL knows why their systems don't work.)

จำเป็นไหมที่ต้องสนใจศึกษาถึงเหตุผลในการทำงานของระบบ โดยส่วนตัวแล้วผู้เขียนคิดว่าจำเป็น เพราะการเข้าใจเหตุผลเบื้องหลังว่าวิธีการใดใช้ได้หรือไม่ได้เพราะเหตุใด ช่วยให้เรากำหนดทิศทางในการพัฒนาระบบได้ว่าจะต้องพิจารณาโดยดูเรื่องใดบ้าง เราไม่สามารถลองผิดลองถูก หรือเลือกเอาเทคนิคที่มีผู้เคยทำในภาษาอื่นมาทดลองได้หมด บุคลากรทางด้านนี้ของไทยมีจำกัดเมื่อเทียบกับคนที่พัฒนาระบบ NLP ภาษาอังกฤษหรือภาษาอื่นๆ ที่มีความพร้อมทั้งในภาครัฐและภาคเอกชน สิ่งที่เราทำกันอยู่ คือนำเสนอว่าเอาเทคนิคของคนนั้นคนนี้มาทดลองแล้วก็สรุปว่าได้ผลดีค่อนข้างมาก เรามักจะเน้นนำเสนองานในส่วนที่บอกว่าประสบความสำเร็จ แต่ส่วนที่สำคัญไม่น้อยคือส่วนที่ยังไม่ประสบความสำเร็จ ทำไมส่วนนั้นจึงไม่ประสบความสำเร็จ เพราะนั่นคือปัญหา การหาคำอธิบายหรือเหตุผลจึงเป็นงานของนักภาษาศาสตร์คอมพิวเตอร์ นักภาษาศาสตร์คอมพิวเตอร์ควรจะต้องหาคำอธิบายว่าทำไมวิธีการต่างๆที่นำมาใช้ถึงได้ผลและในบางกรณีไม่ได้ผล ซึ่งแน่นอนว่าต้องเป็นคำอธิบายที่อ้างอิงความรู้ความเข้าใจในธรรมชาติของภาษา งานทางภาษาศาสตร์คอมพิวเตอร์ควรจะให้ข้อมูลพื้นฐานที่จำเป็นในการพัฒนาระบบ NLP แต่ความรู้ทางภาษานี้แตกต่างจากการศึกษาภาษาศาสตร์ทั่วไปที่มุ่งเน้นที่ทำให้คนเราเข้าใจธรรมชาติของภาษา นักภาษาศาสตร์คอมพิวเตอร์ต้องนำเสนอความรู้ทางภาษาที่สามารถนำไปใช้กับระบบคอมพิวเตอร์ได้ ซึ่งหมายความว่า ผู้ที่ทำงานทางภาษาศาสตร์คอมพิวเตอร์จำเป็นต้องเข้าใจด้วยว่าแนวทางการพัฒนาระบบ NLP มีลักษณะใดบ้าง ความรู้ทางภาษาศาสตร์จะเข้าไปเสริมในจุดใดได้ งานที่ทำนั้น จึงจะเป็นประโยชน์ สามารถนำไปประยุกต์ใช้กับคอมพิวเตอร์ได้

แนวทางใหม่ในการศึกษาภาษาไทย

งานวิจัยทางภาษาศาสตร์ภาษาไทยนั้นมีมานานหลายสิบปีแล้ว ถ้าถามว่างานเหล่านั้นเป็นประโยชน์โดยตรงต่อการนำไปพัฒนาระบบ NLP หรือไม่ คำตอบคือส่วนใหญ่แล้วไม่เป็น ซึ่งนักภาษาศาสตร์ทั่วไปก็จะตอบว่าไม่เป็นไร เพราะภาษาศาสตร์เป็นการศึกษาภาษาเพื่อให้เข้าใจสามัติถิยะภาษา (language competence) ไม่จำเป็นต้องไปประยุกต์ใช้กับศาสตร์อื่นๆ แต่ในปัจจุบัน คำถามในเรื่องของประโยชน์หรือการนำไปประยุกต์ใช้นั้นมีความสำคัญมากขึ้นเรื่อยๆ การศึกษาที่จะเป็นประโยชน์กับการใช้กับคอมพิวเตอร์จึงเป็นแนวทางหนึ่งที่น่าจะทำได้

แล้วงานทางภาษาศาสตร์แบบใดที่จะเป็นประโยชน์ ถ้าดูจากลักษณะของคอมพิวเตอร์ ก็คงต้องเป็นงานทางภาษาศาสตร์ที่มีลักษณะเป็นแบบแผนนิยม (formalism) เป็นการศึกษาหากฎในเรื่องต่างๆ เช่น กฎทางไวยากรณ์ กฎทางอักขรวิธี แต่ดังในตัวอย่างที่กล่าวมาแล้ว การใช้กฎในงานทาง NLP จริงๆ จะมีปัญหาที่ความกำกวมเป็นจำนวนมากที่เกิดจากการใช้กฎเหล่านั้น ซึ่งในปัจจุบันการใช้กฎเพียงอย่างเดียวไม่ใช่วิธีการที่นิยมใช้กันในการพัฒนาระบบ NLP แล้ว แต่ถ้าระบบ NLP อาศัยการใช้สถิติเป็นหลักแล้ว ยังจะเหลืองานอะไรให้นักภาษาศาสตร์ทำที่จะเป็นประโยชน์อีก ในเมื่อนักวิศวกรภาษาหันมาใช้แบบจำลองทางสถิติในการแก้ปัญหา NLP นักวิศวกรภาษาเพียงแค่ตั้งแบบจำลองหนึ่งขึ้นมา แล้วป้อนข้อมูลมากๆให้ระบบเรียนรู้จากค่าสถิติเหล่านั้นเพื่อแก้ปัญหา แต่ถึงจะเป็นเช่นนั้น ผู้เขียนก็เชื่อว่ายังมีประเด็นที่นักภาษาสามารถศึกษาได้ เพราะการสร้างแบบจำลองทางสถิติไม่ได้มีเพียงแบบเดียว การตัดสินใจเลือกใช้แบบจำลองไหนหรือเสนอแบบจำลองใหม่ขึ้นมาต้องอาศัยความเข้าใจในตัวภาษาเป็นพื้นฐาน แบบจำลองที่ทำงานได้ดีจะต้องสอดคล้องกับลักษณะธรรมชาติของตัวภาษานั้น จึงจำเป็นต้องศึกษาตัวภาษานั้นเองอยู่ดี เพียงแต่เราอาจจะต้องเปลี่ยนวิธี

การมองปัญหาทางภาษาศาสตร์ จากเดิมที่มองปัญหาเพื่อตอบคำถามทางภาษาศาสตร์อย่างที่เคยทำกันมา มาเป็นการมองปัญหาเพื่อตอบคำถามให้กับคอมพิวเตอร์ หมายความว่าต้องมองปัญหาจากมุมมองของคอมพิวเตอร์เพื่อหาคำอธิบายให้คอมพิวเตอร์เข้าใจไม่ใช่ให้คนเราเข้าใจดังที่เคยทำกัน ตัวอย่างเช่น ในการศึกษาเรื่องคำ ถ้าเป็นการมองแบบภาษาศาสตร์ทั่วไป เราจะศึกษาเรื่องประเภทของคำว่ามีอะไรบ้าง มีคำมูล คำประสม กลวิธีในการสร้างคำใหม่มีอะไรบ้าง เช่น เขียนทับศัพท์ ใช้คำเดิมแต่สร้างความหมายใหม่ เป็นต้น นั่นเป็นการพยายามอธิบายลักษณะของภาษาเพื่อให้มนุษย์เข้าใจ แต่ถ้ามองจากมุมมองของคอมพิวเตอร์ก็จะสนใจประเด็นอย่างเช่น ทำอย่างไรจึงจะรู้ได้ว่า สิ่งที่มองเหมือนเป็นคำสามคำอย่าง “มุม ตกกระ ทบ” จริงๆเป็นเพียงคำเดียว ลักษณะคำไทยเป็นอย่างไร ทำไมการตัดคำแบบ maximum matching จึงได้ผลดีในบางกรณี ทำไมจึงไม่ได้ผลดีในบางกรณี หรือถ้าต้องการระบุหา (identify) ชื่อเฉพาะ ถ้าศึกษาตามแนวภาษาศาสตร์ก็บอกเพียงว่ามีกลวิธีอะไรบ้าง เช่น มีการใช้คำบ่งชี้นำหน้า เช่น “นาย” “นาง” “พล.ต.” แต่ถ้ามองจากมุมทางคอมพิวเตอร์ ก็ต้องศึกษาด้วยว่า ถ้าพบคำว่า “นาย” จะแน่ใจได้เพียงใดว่าคำที่ตามมานั้นเป็นชื่อเฉพาะ เพราะเราต้องการความรู้สำหรับให้คอมพิวเตอร์เข้าใจไม่ใช่สำหรับให้มนุษย์เราเองเข้าใจ การศึกษาภาษาศาสตร์ในมิติที่ต้องมองจากมุมของคอมพิวเตอร์นี้ เราจึงต้องคำนึงถึงด้วยว่าคอมพิวเตอร์มีแหล่งข้อมูลอะไรที่จะใช้ได้บ้างในการแก้ปัญหานั้นๆ นอกจากนี้ เรายังจำเป็นต้องรู้ขอบข่ายงานที่เป็นที่สนใจในงาน NLP ด้วย เพื่อช่วยในการเลือกเรื่องหรือประเด็นที่น่าสนใจศึกษาได้

ตัวอย่างประเด็นที่น่าสนใจในงาน NLP

เพื่อที่จะช่วยให้เห็นภาพขอบข่ายงานที่เป็นที่สนใจในงาน NLP ผู้เขียนจะขอยกตัวอย่างของงาน NLP ที่เป็นที่สนใจในปัจจุบัน เรื่องแรกคือเรื่องการค้นคืนสารสนเทศ (Information Retrieval) ซึ่งก็ต้องอาศัยความรู้ทางภาษามาช่วย ถ้าต้องการพัฒนาระบบที่ทำได้มากกว่าการเทียบคำที่ค้น (keyword matching) ในการค้นคืนสารสนเทศนี้ มีงานหลายด้านที่เกี่ยวข้องกับภาษา เช่น เรื่องการค้นคืนสารสนเทศข้ามภาษา (cross-language retrieval) นอกจากจะต้องรู้คำแปลที่เทียบเท่าในอีกภาษาแล้ว ในความเป็นจริง มีการเขียนคำข้ามภาษาจากภาษาหนึ่งมาในระบบของอีกภาษา เช่น Michael เขียน ไมเคิล หรือ จุฬา เขียนเป็น chula การค้นข้อมูลจึงต้องอาศัยการถอดอักษรระหว่างภาษา (transliteration) คือถอดคำภาษาต้นแบบเข้ามาในระบบของอีกภาษา และการถอดอักษรแบบย้อนกลับ (backward transliteration) คือ แปลงจากรูปที่ถอดออกมาได้กลับเป็นรูปเดิมในภาษาต้นแบบ ซึ่งแบบหลังนี้จะยากกว่า ตัวอย่างเช่น ในการถอดคำต่างประเทศเป็นคำไทยสามารถเขียนออกมาได้หลายแบบ(แม้ว่าจะมีการวางเกณฑ์ในการทับศัพท์ไว้แล้วก็ตาม) เช่น electronics บางครั้งเขียน “อิเล็คทรอนิกส์” “อิเล็กทรอนิกส์” “อิเล็กโทรนิคส์” เป็นต้น แต่ในการถอดอักษรกลับเป็นคำต้นแบบนั้น จะมีคำตอบที่ต้องการเพียงเดียว “อิเล็คทรอนิกส์” จะต้องสามารถถอดอักษรกลับเป็น “electronics” ไม่สามารถยอมรับคำตอบอื่นได้ นอกจากนี้ในเรื่องการค้นคืนสารสนเทศยังมีประเด็นการศึกษาหาคำสำคัญที่จะเป็นตัวแทนของเอกสารนั้นๆ ซึ่งที่ทำกันมาส่วนมากจะใช้วิธีหาความถี่ของคำ (term frequency หรือ tf) ความถี่ของเอกสารที่พบคำนั้น (document frequency หรือ df) โดยมีข้อสันนิษฐานว่าเอกสารที่ต้องการจะต้องมีคำสำคัญนั้นปรากฏอยู่บ่อย และปรากฏในบางกลุ่มของเอกสารไม่ได้ปรากฏอยู่ทั่วไปในทุกๆเอกสาร ตรงนี้ก็อาจเป็นประเด็นให้ศึกษาได้ว่า การคัดสรรเอกสารที่ตรงกับเรื่องที่ต้องการสามารถดูที่ปัจจัยใดทางภาษาได้อีกบ้างนอกจากการพิจารณาการปรากฏของคำผ่านทางค่า tf, df นี้

อีกเรื่องหนึ่งคือ เรื่องการระบุหาคำสรรพนาม (identify proper name) ซึ่งก็เป็นประโยชน์กับงานค้นคืนสารสนเทศและงานอื่นๆ เช่น การแปลภาษาด้วยเครื่อง เริ่มแรกถูกมองว่าเป็นการเรื่องของคำที่ไม่ปรากฏใน

พจนานุกรมที่คอมพิวเตอร์ใช้ (unknown word) เมื่อเจอสายอักขระที่ไม่สามารถเทียบกับคำในพจนานุกรมได้ ก็ยึดส่วนนั้นไว้แล้วพยายามสร้างคำต่างๆ โดยดึงส่วนข้างเคียงมาประกอบ เช่น “เขาไปเที่ยวสิงคโปร์มา” สายอักขระที่ไม่สามารถเทียบหาจากพจนานุกรมได้คือ “คโปร์” เมื่อลองต่อกับคำข้างเคียงในขอบเขตหนึ่งที่กำหนดก็จะได้สายอักขระอย่างเช่น “สิงคโปร์” “คโปร์มา” “เที่ยวสิงคโปร์” “เที่ยวสิงคโปร์มา” เป็นต้น จากนั้นจึงมาตัดสินทีหลังว่าจะเลือกสายใดเป็นชื่อเฉพาะ แต่ชื่อเฉพาะก็ไม่จำเป็นต้องประกอบขึ้นจากสายอักขระที่ไม่มีในพจนานุกรม ชื่อเฉพาะอาจประกอบจากคำที่มีปรากฏในพจนานุกรมทั้งหมดก็ได้ เช่น “สมชาย” “สมประสงค์” มีส่วนประกอบเป็นคำว่า “สม” “ชาย” และ “สม” “ประสงค์” ต่อมาก็มีผู้พยายามระบุหาชื่อเฉพาะโดยใช้วิธีการทางสถิติดูค่าความน่าจะเป็นของคำและหมวดคำที่คำพวกนี้มาประกอบกัน (Charoenpornsawat et al. 1998) บ้างก็ดูค่าความน่าจะเป็นของคำและกลุ่มความหมายที่คำพวกนี้มาประกอบกัน (Kawtrakul et al. 1997) บ้างก็ใช้วิธีตั้งน้ำหนักของการเชื่อมต่อกันของส่วนต่างๆ ให้ต่างกัน (Kanlayanawat and Prasitjutrakul 1997) ข้อสันนิษฐานของการมองชื่อเฉพาะแบบนี้คืออะไร มีหนทางอื่นอีกไหมที่จะช่วยคอมพิวเตอร์ในการระบุหาชื่อเฉพาะ

เรื่องที่สามคือเรื่องการดึงศัพท์เฉพาะจากคลังข้อมูลภาษาหรือ term extraction ทำอย่างไร จึงจะให้คอมพิวเตอร์ช่วยระบุหาศัพท์เฉพาะด้านออกมาให้เราได้ วิธีการที่ทำกันก็คือใช้เรื่องทางสถิติ เนื่องจากศัพท์เฉพาะมักเป็นคำประกอบคือประกอบด้วยคำหลายๆคำ เช่น water treatment, power supply, intellectual property right, เป็นต้น จึงมีผู้ใช้วิธีการทางสถิติเพื่อดูการปรากฏร่วมกันขององค์ประกอบคำเหล่านั้น (collocation) ว่าปรากฏร่วมกันบ่อยมากจนผิดปกติหรือไม่ ถ้าใช่ ก็น่าจะเป็นศัพท์เฉพาะได้ บางคนก็ดูองค์ประกอบทางโครงสร้างว่าศัพท์เฉพาะสามารถมีโครงสร้างแบบใดได้บ้าง เช่น เป็น N-N-N, N-Prep-N, เป็นต้น เพื่อเพิ่มเติมเกณฑ์ว่าคำย่อยเหล่านั้นสามารถรวมกันเป็นศัพท์เฉพาะได้หรือไม่ นอกจากวิธีการเหล่านี้ เราจะศึกษาในเรื่องนี้อย่างไรได้บ้าง และเมื่อศึกษาภาษาไทย ศัพท์เฉพาะในภาษาอังกฤษเวลาแปลเป็นไทยแล้วเป็นอย่างไร กลวิธีที่ใช้มีอะไรบ้าง ใช้ทับศัพท์หรือแปลคำหรือสร้างคำใหม่ เราจะระบุหาศัพท์เฉพาะภาษาไทยได้อย่างไร ทั้งหมดนี้ก็ต้องอาศัยพื้นฐานการศึกษาเกี่ยวกับศัพท์เฉพาะในภาษาไทย

เรื่องที่สี่คือเรื่องการจับคู่การแปล (alignment) ในคลังข้อมูลเทียบบท (parallel corpus) ซึ่งโดยทั่วไปเป็นการจับคู่ในระดับประโยค ข้อสันนิษฐานพื้นฐานในเรื่องนี้คือ เราสามารถจับคู่การแปลในระดับประโยคได้ ซึ่งอาจเป็นการจับคู่แบบประโยคต่อประโยค (1-1), หนึ่งประโยคต่อหลายประโยค (1-many) หรือหลายประโยคกับหนึ่งประโยค (many-1), หรือต้องจับคู่ทีละหลายๆประโยค (many-many) ก็ได้ และข้อสันนิษฐานต่อมาคือการแปลมีลักษณะลำดับที่สอดคล้องกับต้นฉบับ การจับคู่การแปลเองด้วยมือนั้นเป็นงานที่เสียเวลามาก จึงมีผู้สนใจศึกษาหาวิธีที่จะให้คอมพิวเตอร์ช่วยในการจับคู่ให้โดยอัตโนมัติ คำถามคือ จะใช้วิธีการอะไรได้บ้างในการจับคู่การแปลนี้ เช่น ดูจากคำแปลที่ปรากฏในประโยคทั้งคู่ ดูลำดับคำแปลที่พบในประโยคได้ไหม หรือคำนวณค่าความคล้ายคลึง (similarity) ของประโยคทั้งสองออกมาเป็นในรูปของเว็กเตอร์ (vector) ในทางคณิตศาสตร์ นอกจากนี้ยังมีประเด็นเรื่องจับคู่ในระดับคำ หรือคือการหาคำเทียบเท่า ซึ่งข้อสันนิษฐานในเรื่องนี้คือการแปลมีลักษณะของการแปลแบบคำต่อคำ (word by word translation) ข้อสันนิษฐานนี้เป็นจริงหรือไม่ในการแปลเป็นภาษาไทย

ทั้งหมดนี้เป็นเพียงส่วนหนึ่ง เป็นส่วนเริ่มต้นของการศึกษาในระดับคำเสียเป็นส่วนมาก ผู้เขียนยังไม่ได้กล่าวถึงประเด็นการศึกษาในระดับที่สูงกว่าคำ เช่น การระบุหาขอบเขตประโยค (sentence identify) การ

ดึงใจความสำคัญจากตัวบท (information extraction) การสรุปย่อเนื้อหาตัวบท (text summarization) การจัดประเภทเอกสารโดยอัตโนมัติ (text classification) เป็นต้น ซึ่งทุกเรื่องนั้นจะได้ประโยชน์จากการศึกษาภาษาไทยที่สามารถให้คำอธิบายที่ตรงกับความต้องการ กล่าวโดยสรุป นักภาษาศาสตร์คอมพิวเตอร์จะต้องศึกษาประเด็นต่างๆที่เป็นที่สนใจในงาน NLP และพยายามตอบคำถามว่าวิธีการต่างๆที่ใช้นั้นมีข้อสันนิษฐานเกี่ยวกับภาษาอย่างไร เป็นที่ยอมรับหรือสอดคล้องกับลักษณะธรรมชาติของภาษาไทยได้หรือไม่ หากไม่ได้ ลักษณะทางภาษาไทยที่จะสามารถนำมาประยุกต์ใช้ในเรื่องนั้นๆ น่าจะเป็นไปในทำนองใด โดยจะต้องพิจารณาจากข้อจำกัดต่างๆของการพัฒนาระบบในเรื่องนั้นประกอบด้วย ซึ่งผลสุดท้ายก็จะนำมาสู่การปรับปรุงหรือนำเสนอวิธีการใหม่ในงาน NLP ในที่สุด

อนาคตและความจำเป็นของการศึกษาภาษาไทยตามแนวทางนี้

เมื่อทำงานกับคอมพิวเตอร์ ปัญหาง่ายๆ ไม่ใช่ปัญหาง่ายๆอีกต่อไป คอมพิวเตอร์นั้นไม่รู้อะไรเลย เราจึงต้องศึกษาภาษาไทยจากมุมมองของคอมพิวเตอร์ เพื่ออธิบายให้คอมพิวเตอร์นั้นเข้าใจและสามารถแก้ไขปัญหาภาษาที่ต้องการได้ และถึงแม้ระบบ NLP ปัจจุบันจะมีลักษณะเป็นแบบจำลองทางสถิติ แบบจำลองก็มีอยู่มากมายแตกต่างกัน การเลือกแบบจำลองใดต้องพิจารณาลักษณะทางภาษาด้วย และการอธิบายภาษาในลักษณะที่เป็นแบบจำลองทางสถิตินี้ จริงๆไม่ใช่เรื่องใหม่และไม่ใช่เรื่องที่นอกเหนือจากภาษาศาสตร์ คณิตศาสตร์เป็นพื้นฐานแบบจำลองทางทฤษฎีที่ใช้ในศาสตร์ต่างๆอยู่แล้ว ไม่ว่าจะเป็น ฟิสิกส์ เคมี ชีววิทยา เศรษฐศาสตร์ เป็นต้น ความจริงทฤษฎีไวยากรณ์ของชอมสกี้ (Chomsky) หรือของคนอื่นๆ ก็เป็นแบบจำลองทางคณิตศาสตร์ เพียงแต่ใช้แบบจำลองที่มีลักษณะเป็น discrete mathematics ไม่ใช่แบบที่เป็น stochastic

นอกจากนี้ เราต้องตระหนักถึงบทบาทและอิทธิพลของเทคโนโลยีที่มีต่อภาษาด้วย เราคงไม่สามารถเรียกร้องให้อนุรักษ์ภาษาไทยเพียงเพื่อให้คงอยู่ในโลกใบเก่า แต่ไม่สามารถปรับภาษาให้อยู่กับเทคโนโลยีใหม่ๆได้ เมื่อสมัยแรกๆของการใช้โปรแกรม Word เราอาจจะเคยเห็นความอุตสาหะของผู้ใช้ที่นั่งจัดกั้นหลังของแต่ละบรรทัดเองเพราะคอมพิวเตอร์ปัดคำขึ้นบรรทัดใหม่ไม่ถูกต้อง ทุกวันนี้ เรายังเห็นผู้คนทำแบบนี้อยู่หรือไม่ แทบจะไม่มีใครใส่ใจเรื่องนี้เวลาพิมพ์งานจริงๆ ถามว่า ปัจจุบัน ไม่มีการปัดคำขึ้นบรรทัดใหม่ที่ผิดแล้วหรือ ก็ไม่ใช่ เพราะเรายังพบการตัดคำผิดอย่างน่าเกลียดในหน้าหนังสือพิมพ์ เช่น ตัด “น” ขึ้นบรรทัดใหม่ออกจากคำว่า “บ้าน” แต่เป็นเพราะในโลกสมัยใหม่ที่เวลาเป็นปัจจัยของการผลิตด้วย ผู้คนไม่มีเวลามากนัก ประเด็นก็คือผู้ใช้จะยอมรับสิ่งที่ถูกกำหนดโดยเทคโนโลยีโดยปริยาย หากเราลองนึกเปรียบเทียบว่า ถ้ามีผู้ทำโปรแกรมถอดตัวอักษรอังกฤษเป็นไทยซึ่งอาจได้ผลที่ตรงตามเกณฑ์ของราชบัณฑิตบ้าง ไม่ตรงบ้าง แต่สะดวกต่อการใช้เพียงแค่ปลายนิ้วคลิ้กเดียวก็ได้ผลออกมา กับการที่ผู้ใช้จะต้องถอดตัวอักษรอังกฤษเป็นไทยเองโดยอ่านกฎหรือคู่มือของราชบัณฑิต ซึ่งก็ไม่แน่ว่าจะถอดอักษรออกมาแล้วถูกต้องตามเกณฑ์ ผู้ใช้จะเลือกทางไหน ดังนั้น ผู้ที่จะมีบทบาทกำหนดอนาคตของภาษาไทยอาจจะเป็นนักวิศวกรภาษามากกว่านักภาษา เพื่อที่จะกำหนดการใช้ภาษาไทยให้อยู่ในลักษณะที่ต้องการ นักภาษาจึงจำเป็นต้องนำตัวเองเข้ามามีบทบาทในการประยุกต์ใช้งานด้านนี้ด้วย แม้ว่าจะมองดูเป็นเรื่องยาก แต่เป็นเรื่องที่เป็นไปได้ ดังจะเห็นได้จากที่ ในปัจจุบัน ก็มีนิสิตสายอักษรศาสตร์ที่สนใจศึกษาด้านภาษาศาสตร์คอมพิวเตอร์ และสามารถพัฒนางานวิจัยด้านนี้ออกมาได้

หนังสืออ้างอิง

- วิรัช ศรเลิศล้ำวาณิช. 2536. การตัดคำไทยในระบบแปลภาษา. ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ
- Abney, Steven. 1996. Statistical Methods and Linguistics. In Judith Klavans and Philip Resnik, eds., The Balancing Act. Cambridge, MA.:MIT Press. (http://www.sfs.nphil.uni-tuebingen.de/~abney/)
- Charoenpornsawat, Paisarn, Boonserm Kijsirikul, and Surapant Meknavin. 1998. “Feature-based Proper Name Identification in Thai” In Proc. Of National Computer Science and Engineering Conference: NCSEC’98.
- Kanlayanawat, W. and S. Prasitjutrakul. 1997. "Automatic Indexing for Thai Text with Unknown Words using Trie Structure",Proceeding of the Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97), pp. 115-120, Phuket, Thailand, December 2-4, 1997.
- Kawtrakul, Asanee, et.al., 1997. "Automatic Thai Unknown Word Recognition "NLPRS 97, THAILAND.
- Potipiti, Tanapong, Virach Sornlertlamvanich, and Thatsanee Charoenporn. 2000. Automatic Corpus-Based Thai Word Extraction. In Proceedings of The Fourth Symposium on Natural Language Processing 2000.
- Shannon, C.E. A Mathematical Theory of Communication. 1948. Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.