บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันการพัฒนาทางเทคโนโลยีได้ก้าวไปข้างหน้าอย่างรวดเร็ว ข้อมูลข่าวสารต่างๆ ถูกส่งผ่านออกไปได้ในหลายช่องทาง ทั้งทางสิ่งพิมพ์และอินเทอร์เน็ต (internet) การรับข้อมูลของมนุษย์ก็เป็นไปในทางที่สะดวกสบายขึ้น เช่น การค้นหาข้อมูลของสิ่งใดสิ่งหนึ่งทางอินเทอร์เน็ตก็สามารถทำได้โดยง่ายเพียงใช้แค่คำสำคัญ (keyword) คำเดียวเท่านั้น เป็นต้น ซึ่งสิ่งที่ช่วยอำนวยความสะดวกในการรับข้อมูลคือ งานต่างๆ ที่มีการจัดการทางการประมวลผลภาษาธรรมชาติหรือ NLP (Natural Language Processing) ซึ่งใช้จัดการกับข้อมูลในบทความหรือเอกสารเพื่อสนองความต้องการของผู้ใช้ ยกตัวอย่างเช่น การค้นคืนสารสนเทศ (Information Retrieval) การสกัดสารสนเทศ (Information Extraction) และ การแปลภาษาด้วยเครื่อง (Machine Translation) เป็นต้น

การค้นคืนสารสนเทศ (Information Retrieval) (Gaizauskas, 2000) เป็นระบบที่ใช้เพื่อดึงข้อมูลที่ต้องการออกมา โดยผู้ใช้อาจให้คำสำคัญ (keyword) หรือ ป้อนคำถาม (query) เข้าไป ซึ่ง คำสำคัญ (keyword) ที่ใช้จะเป็นหัวข้อหลักๆ ของความสนใจของผู้ที่ต้องการข้อมูล แล้วระบบก็จะให้รายการเอกสารที่เกี่ยวข้องกับคำสำคัญนั้นๆ ออกมาให้

ชื่อเฉพาะ (Named entity หรือ proper name) คือ นิพจน์ที่ใช้เรียกหรือระบุถึงสิ่งใดๆ เช่น ชื่อบุคคล ชื่อองค์กร หรือชื่อสถานที่ นิพจน์เหล่านี้ถือเป็นองค์ประกอบพื้นฐานและเป็นข้อมูลสำคัญอย่างยิ่งในบทความหรือเอกสารอีกชนิดหนึ่ง ซึ่งสามารถใช้เพื่อเป็นคำสำคัญ (keyword) ในการค้นหาเอกสารข้อมูลเกี่ยวข้องที่เราต้องการค้นหาได้ ดังนั้นหากมีการรู้จำชื่อเฉพาะแล้วทำรายการชื่อเฉพาะไว้ก็จะช่วยผู้ใช้ในการค้นหาข้อมูลที่เกี่ยวข้องกับชื่อเฉพาะที่ต้องการได้ง่ายขึ้น คือผู้ใช้สามารถเลือกชื่อเฉพาะจากรายการชื่อเฉพาะตามความต้องการที่จะใช้ข้อมูลต่างๆ ของชื่อนั้นๆ หรือในอีกทางหนึ่งจะทำให้ไม่ดึงข้อมูลของชื่อเฉพาะอื่นที่ไม่ต้องการออกมา อย่างเช่น ผู้ใช้อยากทราบข้อมูลของ "Bill Clinton" แต่จำชื่อต้นคือ "Bill" ไม่ได้ การมีรายการชื่อไว้ก็จะช่วยให้สามารถดึงข้อมูลของ "Bill Clinton" ออกมาได้โดยง่าย เป็นต้น หรือในกรณีที่ผู้ใช้ต้องการค้นคำที่เป็นคำนามทั่วไป แต่บังเอิญรูปไปซ้ำกับชื่อเฉพาะ เช่น ต้องการค้นเรื่อง ปฏิวัติ ซึ่งผู้ใช้ก็คงไม่

ต้องการข้อมูลของคนที่มีชื่อว่า "ปฏิวัติ" เป็นต้น ดังนั้น ความสามารถในการรู้จำชื่อเฉพาะจึงเป็นสิ่งหนึ่งที่จำเป็นสำหรับระบบการค้นคืนสารสนเทศ

การสกัดสารสนเทศ (Information Extraction) (Gaizauskas, 2000 และ Turney, 2000) เป็นระบบซึ่งย่อยข้อมูลสำคัญๆ ออกมาจากบทความหรือเอกสารโดยจะแสดงข้อมูลของบทความหรือเอกสารไว้ในแม่แบบ (template) ซึ่งจะมีข้อมูลที่บอกว่า มีเหตุการณ์อะไรเกิดขึ้น ใครเป็นคนทำ เกิดเหตุการณ์ขึ้นเมื่อใด เหตุการณ์เกิดขึ้นได้อย่างไร ซึ่งการรู้จำชื่อเฉพาะจะช่วยให้ระบบการสกัดสารสนเทศทำงานเร็วขึ้น เพราะจะดึงข้อมูลชื่อคน องค์กรและสถานที่ ที่จะใส่ในแม่แบบบางส่วนออกมาแล้ว

สำหรับงานการแปลภาษาด้วยเครื่อง (Machine Translation) (Arnold และคณะ, 1994) เป็นระบบที่แปลบทความหรือเอกสารจากอีกภาษาหนึ่งไปเป็นอีกภาษาหนึ่ง ปัญหาที่พบในงานนี้ก็ยังรวมถึงการแปลที่ผิดพลาดซึ่งเกิดจากการที่ไม่รู้ว่าคำที่ต้องการแปลเป็นชื่อเฉพาะทำให้แปลชื่อออกมาผิด หากมีการผนวกงานการรู้จำชื่อเฉพาะหรือ NER (Named Entity Recognition) มาช่วยในการจำแนกพวกชื่อเฉพาะเหล่านี้ออกมา ก็จะทำให้ไม่เกิดการผิดพลาดในการแปลชื่อเฉพาะขึ้น อย่างเช่น ชื่อบริษัท "Built to Built" ก็จะไม่ถูกแปลออกมาเป็น บริษัท "สร้างเพื่อสร้าง" แต่จะแปลโดยการทับศัพท์เป็น บริษัท "บิลท์ทูบิลท์"

จากความจำเป็นในการใช้ระบบการรู้จำและจำแนกประเภทชื่อเฉพาะในงานการประมวลผลภาษาธรรมชาติต่างๆ ที่กล่าวมาแล้วนี้เอง ทำให้งานการรู้จำและจำแนกประเภทชื่อเฉพาะ จึงเป็นอีกงานหนึ่งที่น่าสนใจและเป็นพื้นฐานสำหรับงานดังกล่าว

1.2 วัตถุประสงค์ของการวิจัย

1. เปรียบเทียบประสิทธิภาพของระบบการรู้จำชื่อเฉพาะภาษาไทยเมื่อใช้วิธีการทางสถิติแบบต่างๆ ได้แก่ การใช้ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation ร่วมกับการใช้ Localmax algorithm ในการรู้จำชื่อเฉพาะ
2. ทดสอบประสิทธิภาพของกฎทางภาษาศาสตร์ที่ใช้บริบทภายในและบริบทข้างเคียงในการจำแนกประเภทของชื่อเฉพาะภาษาไทย
3. พัฒนาระบบการรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทยโดยใช้แนวทางแบบลูกผสม (hybrid approach) ระหว่างวิธีการทางสถิติในข้อที่ 1 และกฎทางภาษาศาสตร์ในข้อที่ 2

4. ประเมินประสิทธิภาพและวิเคราะห์ปัญหาของระบบการรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทยที่พัฒนาขึ้น

1.3 สมมติฐาน

1. วิธีทางสถิติแบบที่ใช้ค่า Mutual Expectation ร่วมกับการใช้ Localmax algorithm ให้ผลในการรู้จำชื่อเฉพาะภาษาไทยได้ดีกว่าการใช้วิธีทางสถิติแบบอื่น ได้แก่ การใช้ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) และค่า Dunning's Log Likelihood
2. บริบทข้างเคียงที่สามารถนำมาใช้ในการจำแนกประเภทของชื่อเฉพาะภาษาไทยได้ ได้แก่ คำนำหน้าและคำตามหลังชื่อเฉพาะ
3. ระบบการรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทยที่พัฒนาขึ้นจะมีอัตราการรู้จำไม่ต่ำกว่า 90%

1.4 ขอบเขตของการวิจัย

1. ประเภทของข้อมูลในคลังข้อมูลที่จะใช้คือ บทความข่าวการเมืองที่มีปริมาณของชื่อเฉพาะ 3 ประเภท ได้แก่ ชื่อเฉพาะประเภท ชื่อคน ชื่อสถานที่และชื่อองค์กร โดยชื่อเฉพาะแต่ละประเภทจะมีปริมาณไม่ต่ำกว่า 3,000 ชื่อ
2. วิธีทางสถิติที่จะใช้ในการรู้จำตำแหน่งของชื่อเฉพาะ ได้แก่ การใช้ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation ร่วมกับการใช้ Localmax algorithm
3. กฎที่ใช้ในการจำแนกประเภทของชื่อเฉพาะนั้นจะสร้างมาจากหลักฐานจากบริบทภายในซึ่งได้แก่ รายการของคำนำหน้าชื่อเฉพาะ เช่น นาย นาง หรือนางสาว เป็นต้น หรือคำตามหลังชื่อเฉพาะ เช่น จำกัด เป็นต้น และลักษณะของบริบทข้างเคียงของชื่อเฉพาะที่ได้มาจากคลังข้อมูลใช้ฝึกสอน (training corpus) โดยจะเป็นรายการของคำที่มักจะนำหน้าและตามหลังชื่อเฉพาะแต่ละประเภททั้ง 3 ประเภท เช่น คำว่า "กิน" จะตามหลังชื่อเฉพาะประเภทชื่อคนมากกว่าที่จะพบว่าตามหลังชื่อเฉพาะประเภทอื่น เป็นต้น

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ได้ระบบที่ใช้ในการรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทยที่จะนำไปใช้ในระบบการประมวลผลภาษาอื่นๆ เช่น การค้นคืนสารสนเทศ การสกัดสารสนเทศ และ การแปลภาษาด้วยเครื่อง เป็นต้น

1.6 วิธีดำเนินการวิจัย

1. ทบทวนวรรณกรรม เพื่อศึกษาและเลือกวิธีที่จะใช้เพื่อรู้จำตำแหน่งและจำแนกประเภทของชื่อเฉพาะภาษาไทย
2. เก็บข้อมูลและสร้างคลังข้อมูลข่าว เพื่อใช้สำหรับการฝึก (training) และการทดสอบ (testing)
3. เขียนโปรแกรมตามวิธีทางสถิติที่เลือกใช้ซึ่งได้แก่ การใช้ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation ร่วมกับการใช้ Localmax algorithm เพื่อใช้ในการดึงกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ (candidate) ออกมา
4. ทำการทดลองในคลังข้อมูลใช้ทดสอบ
5. วิเคราะห์ผลการทดลองที่ได้และเปรียบเทียบประสิทธิภาพของแต่ละวิธีในการหากลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ (candidate) ว่ามีความถูกต้องมากน้อยเพียงใดโดยใช้การวัดค่าความแม่นยำ (precision) ค่าความครบถ้วน (recall) เพื่อคำนวณหาค่า F-measure
6. เลือกวิธีทางสถิติที่ให้ผลดีที่สุดในการระบุหาชื่อเฉพาะมาใช้กับกฎเพื่อจำแนกประเภทของชื่อเฉพาะและทดสอบประสิทธิภาพของกฎที่ใช้บริบทภายในและบริบทข้างเคียงในการจำแนกประเภทของชื่อเฉพาะภาษาไทย โดยกฎจะมาจากการวิเคราะห์ลักษณะภาษาไทยภายในคลังข้อมูลใช้ฝึก (training corpus)
7. สร้างกฎเพื่อใช้ในการระบุและจำแนกประเภทของชื่อเฉพาะภาษาไทยโดยใช้หลักฐานที่ได้มาจากบริบทภายในและบริบทข้างเคียง
8. ทำการทดลองในคลังข้อมูลใช้ทดสอบ (testing corpus)
9. วิเคราะห์และสรุปผลการทดลอง

1.7 โครงสร้างวิทยานิพนธ์

ในบทที่ 2 จะกล่าวถึงความหมาย รูปแบบต่างๆ ลักษณะและการแบ่งประเภทของชื่อเฉพาะ จากนั้นจะกล่าวถึงระบบการรู้จำตำแหน่งและจำแนกประเภทของชื่อเฉพาะ ซึ่งโดยทั่วไปแบ่งออกได้เป็น 3 ระบบคือ ระบบที่ใช้กฎ ระบบที่ใช้วิธีทางสถิติ และระบบแบบลูกผสมซึ่งรวมวิธีที่ใช้กฎและสถิติเข้าไว้ด้วยกัน ในส่วนถัดมา จะกล่าวถึงคลังข้อมูลที่ใช้ในการวิจัยครั้งนี้ การวิเคราะห์ชื่อเฉพาะที่พบในคลังข้อมูล และในส่วนสุดท้ายจะกล่าวถึงระบบการรู้จำชื่อเฉพาะที่ใช้ในงานวิจัยนี้และการประเมินผลการทำงานของระบบ

ในบทที่ 3 จะกล่าวถึงการใช้วิธีทางสถิติเพื่อหารายการกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ และเนื่องจากวิธีการทางสถิติที่ใช้เป็นการคำนวณความสัมพันธ์ระหว่างสองหน่วยหรือสองพยางค์ ในส่วนแรกของบทนี้จึงจะกล่าวถึงการปรับวิธีการทางสถิติเพื่อใช้คำนวณความสัมพันธ์ระหว่างหลายพยางค์ได้ ซึ่งจะกล่าวถึงวิธีทางสถิติที่นำมาปรับเพื่อใช้เลือกกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ (candidate) และนำเสนอผลวิธีการทางสถิติที่เหมาะสมที่สุดในการดึงกลุ่มพยางค์ที่น่าจะเป็นชื่อเฉพาะออกมา และในสองส่วนถัดมา จะกล่าวถึงผลและการอภิปรายผลของวิธีทางสถิติที่ใช้ในงานวิจัยนี้

ในบทที่ 4 จะกล่าวถึงการรู้จำชื่อเฉพาะ โดยเป็นการใช้กฎที่เขียนขึ้นเพื่อพิจารณารายการชื่อเฉพาะที่เลือกมาด้วยวิธีการทางสถิติว่าเป็นชื่อเฉพาะจริงหรือไม่และเป็นชื่อเฉพาะประเภทใด ซึ่งบทนี้จะแบ่งเนื้อหาออกเป็น 4 ส่วนหลัก โดยในส่วนแรกจะกล่าวถึงการสร้างกฎจากหลักฐานภายใน (internal evidence) รวมถึงหลักฐานจากบริบทข้างเคียง (external evidence) ส่วนถัดมาจะกล่าวถึงกฎที่ถูกสร้างขึ้น สองส่วนสุดท้าย จะกล่าวถึงผลและการอภิปรายผลของการใช้กฎที่สร้างขึ้นเพื่อการจำแนกประเภทของชื่อเฉพาะว่ามีความถูกต้องมากน้อยเพียงใด

และบทที่ 5 ซึ่งเป็นบทสุดท้ายจะเป็นบทที่สรุปกระบวนการศึกษาและผลการศึกษาการรู้จำและจำแนกประเภทของชื่อเฉพาะภาษาไทย และนำเสนอแนวทางในการพัฒนาระบบการรู้จำและจำแนกประเภทของชื่อเฉพาะภาษาไทยเพิ่มเติมต่อไป

1.8 ศัพท์เฉพาะ (term) ที่ใช้ในงานนี้

1. ค่า precision หรือค่าความแม่นยำ เป็นค่าที่จะแสดงให้เห็นว่าระบบหรือวิธีการที่ทำในงานวิจัยมีความแม่นยำมากเพียงใด โดยค่าความแม่นยำจะคำนวณได้จากสูตร

ความแม่นยำ(P) = (จำนวนคำตอบที่ถูกต้อง *100)/ จำนวนคำตอบทั้งหมด

2. ค่า recall หรือค่าความครบถ้วน เป็นค่าที่จะแสดงให้เห็นว่าระบบหรือวิธีการที่ทำในงานวิจัยสามารถดึงคำตอบที่ถูกต้องออกมาได้ครบถ้วนเพียงใด ซึ่งค่าความครบถ้วนจะคำนวณได้จากสูตร

   ค่าความครบถ้วน (R) = (จำนวนคำตอบที่ถูกต้อง *100) / จำนวนคำตอบที่ถูกต้องทั้งหมดในข้อมูล

3. ค่า F คือค่าที่เฉลี่ยให้ความสำคัญกับค่าความแม่นยำและความครบถ้วนเท่าๆ กัน ซึ่งสามารถคำนวณได้จากสูตร

   ค่า F = (2 *P *R) / (P+R)

4. อัตราการรู้จำ (recognition rate) เป็นค่าที่จะแสดงให้เห็นว่าระบบมีอัตราในการรู้จำชื่อเฉพาะมากเพียงใด ซึ่งจะใช้เพื่อหาวิธีทางสถิติที่เหมาะสมที่จะใช้ในงานวิจัยนี้ และทำให้สามารถวัดประสิทธิภาพของวิธีกฎที่ใช้ในการจำแนกประเภทของชื่อเฉพาะด้วย ซึ่งอัตราการรู้จำจะวัดจากค่า F และค่า F จะคำนวณได้จากค่าความแม่นยำ (P) และค่าความครบถ้วน (R) ตามสูตรที่กล่าวไปแล้วในข้อ 3

5. เวลาในที่ใช้การรู้จำ (recognition time) เป็นปัจจัยอีกปัจจัยหนึ่งที่ใช้ในการตัดสินหาวิธีทางสถิติที่เหมาะสมที่สุดที่จะใช้ในงานวิจัยนี้ ซึ่งจะคำนวณจากจำนวนชื่อเฉพาะที่เลือกออกมาได้ต่อ 1 นาที