บทที่ 5

สรุปผลการวิจัย และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

การวิจัยนี้เป็นการวิจัยที่มีจุดประสงค์เพื่อพัฒนาระบบการรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทยโดยใช้แนวทางแบบลูกผสม (hybrid approach) โดยแนวทางดังกล่าวจะใช้วิธีทางสถิติเพื่อคัดกรองกลุ่มพยางค์ที่อาจจะเป็นชื่อเฉพาะ ก่อนจะเข้าสู่วิธีที่ใช้กฎซึ่งจะตัดสินความเป็นชื่อเฉพาะและจำแนกประเภทของชื่อเฉพาะต่อไป โดยผู้วิจัยคาดหวังว่าระบบที่พัฒนานี้จะสามารถรู้จำและจำแนกประเภทชื่อเฉพาะได้ด้วยอัตราการรู้จำมากกว่า 90%

และเมื่อแบ่งงานวิจัยออกเป็นสองส่วน ได้แก่ส่วนที่เป็นการวิเคราะห์ค่าทางสถิติเพื่อใช้ในการหากลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ และส่วนที่เป็นการใช้วิธีกฎซึ่งจะใช้ในการตัดสินและจำแนกประเภทชื่อเฉพาะที่ได้แล้ว

จากวิธีทางสถิติทั้ง 5 วิธี ได้แก่ การใช้ Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI3) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation ร่วมกับการใช้ Localmax algorithm ในการรู้จำชื่อเฉพาะนั้น เมื่อเปรียบเทียบประสิทธิภาพของระบบการรู้จำชื่อเฉพาะภาษาไทยเมื่อใช้วิธีการทางสถิติแบบต่างๆ แล้ว พบว่าวิธีที่ใช้ค่า Mutual Expectation ร่วมกับการใช้ Localmax algorithm ในการรู้จำชื่อเฉพาะนั้นให้ผลอัตราการรู้จำ (recognition rate) ที่วัดด้วยค่า F ได้ดีที่สุด แต่วิธีดังกล่าวก็มีข้อเสียตรงที่ใช้เวลาในการรู้จำ (recognition time) ที่นานเกินไป ทำให้ในงานวิจัยนี้จะใช้วิธีทางสถิติที่ให้ผลอัตราการรู้จำที่ดีรองลงมา นั่นคือ การใช้ค่า Mutual Information ร่วมกับการใช้ Localmax algorithm จากนั้นชื่อเฉพาะที่เลือกมาด้วยวิธีการสถิตินี้ จะถูกนำเข้าสู่ส่วนที่ใช้วิธีกฎ

จากคลังข้อมูล พบว่าชื่อเฉพาะส่วนใหญ่มักมีหลักฐานภายใน (internal evidence) ได้แก่ คำนำหน้าชื่อ ปรากฏร่วมอยู่ด้วย ดังนั้นการเขียนกฎจึงใช้คำนำหน้าชื่อดังกล่าวในการกำหนดจุดเริ่มต้นของขอบเขตชื่อเฉพาะ อีกทั้ง คำนำหน้าชื่อยังสามารถทำให้จำแนกประเภทของชื่อเฉพาะได้ด้วย และนอกจากหลักฐานภายในแล้ว จากคลังข้อมูลยังพบว่ามีหลักฐานภายนอก (external evidence) ซึ่งสามารถหาได้จากบริบทข้างเคียงที่ปรากฏร่วมกับชื่อเฉพาะ โดยเราสามารถใช้หลักฐานภายนอกดังกล่าวเพื่อช่วยในการระบุจุดเริ่มต้นของชื่อเฉพาะในกรณีที่ชื่อเฉพาะไม่มีคำนำหน้าชื่อ เช่นในกรณีของชื่อเฉพาะประเภทชื่อสถานที่ หลักฐานภายนอกได้แก่ คำว่า "ย่าน" , "ฝั่ง" , "เยือน" , "บริเวณ" , "ท้อง" , "ลาน" ถูกนำมาใช้เพื่อกำหนดจุดเริ่มต้นของ

ขอบเขตชื่อเฉพาะได้ นอกจากนี้ หลักฐานภายนอกยังสามารถใช้เพื่อระบุจุดสิ้นสุดของขอบเขตชื่อเฉพาะ และเป็นหลักฐานเพิ่มเติมในการจำแนกประเภทของชื่อเฉพาะ อีกทั้งยังช่วยลดความกำกวมในกรณีที่คำนำหน้าชื่อเกิดความกำกวมในการจำแนกประเภทชื่อเฉพาะระหว่างชื่อองค์กรและชื่อสถานที่ ซึ่งคำนำหน้าชื่อเหล่านั้น ได้แก่ "กรม" , "กระทรวง" , "บริษัท" , "ธนาคาร" , "มูลนิธิ" , "สำนักงาน" และจากการทดสอบพบว่าระบบกฎที่สร้างขึ้นสามารถรู้จำและจำแนกประเภทของชื่อเฉพาะโดยให้ค่าอัตราการรู้จำหรือ ค่า F สำหรับชื่อเฉพาะประเภทชื่อคน 69.150% ชื่อองค์กร 62.953% ชื่อสถานที่ 38.869% ตามลำดับ

จากอัตราการรู้จำ พบว่ามีค่าน้อยกว่า 90% ทำให้ระบบมีประสิทธิภาพที่ไม่ค่อยดีนัก แต่หากพิจารณาจากค่าความครบถ้วน (recall rate) แล้ว จะพบว่าสำหรับชื่อเฉพาะประเภทชื่อคนจะมีค่าความครบถ้วนถึง 96.117% ชื่อองค์กร 92.934% ชื่อสถานที่ 50.317% ตามลำดับ ซึ่งการที่ระบบมีประสิทธิภาพน้อยกว่า 90% นี้เป็นเพราะการคำนวณอัตราการรู้จำจะเป็นการคำนวณหาค่าเฉลี่ยระหว่างค่าความแม่นยำและค่าความครบถ้วน เพราะในงานนี้เราจะให้ความสำคัญกับค่าความแม่นยำและค่าความครบถ้วนเท่าๆ กัน จึงใช้ค่า F เป็นอัตราในการรู้จำเพื่อวัดประสิทธิภาพของระบบ

และจากค่าความแม่นยำ จะพบว่าในงานวิจัยนี้ ระบบให้ค่าความแม่นยำไม่ดีนัก ทั้งนี้เป็นเพราะ จำนวนชื่อเฉพาะที่ไม่ถูกต้องซึ่งผ่านเกณฑ์การคัดเลือกจากระบบมีจำนวนมาก ซึ่งเป็นผลมาจากความผิดพลาดในการหาขอบเขตสิ้นสุดของชื่อเฉพาะ ดังในกรณีตัวอย่าง

"กระทรวง <candidate>การคลังก็รับ</candidate>ว่า"

จะสังเกตได้ว่า จากคำนำหน้าชื่อซึ่งเป็นจุดเริ่มต้นคือคำว่า "กระทรวง" ผ่านเงื่อนไขของโปรแกรมเรียบร้อย แต่มาพบปัญหาอยู่ที่จุดสิ้นสุดของชื่อเฉพาะทำให้ได้ชื่อเฉพาะที่ไม่ถูกต้องปนเข้ามาอีกตัวหนึ่ง ยังผลให้ค่าความแม่นยำลดลงไปอีก

5.2 ข้อเสนอแนะ

งานวิจัยนี้เป็นงานวิจัยที่มีจุดประสงค์เพื่อพัฒนาระบบการรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทยโดยใช้แนวทางแบบลูกผสม (hybrid approach) โดยแนวทางดังกล่าวจะใช้วิธีทางสถิติเพื่อคัดกรองกลุ่มพยางค์ที่อาจจะเป็นชื่อเฉพาะ (candidate) ก่อนจะเข้าสู่วิธีที่ใช้กฎซึ่งจะจำแนกประเภทของชื่อเฉพาะต่อไป

ดังนั้นในส่วนแรกของงานหรือระบบที่ใช้วิธีทางสถิตินั้น ในงานนี้มีการใช้คลังข้อมูลที่ผ่านการตัดพยางค์ ซึ่งทำให้ชื่อเฉพาะที่เลือกจากกลุ่มพยางค์ด้วยวิธีการทางสถิติที่ได้มีจำนวนมาก ทั้งนี้เป็นเพราะวิธีทางสถิติที่ใช้งานนี้ 5 วิธี 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 ในการรู้จำชื่อเฉพาะนั้น มีหลักการร่วมกันคือการหาค่าความสัมพันธ์ที่มีระหว่างหน่วยย่อย หากมีค่ามาก โอกาสที่จะเป็นหน่วยเดียวกันก็จะมีสูงตามไปด้วย แต่เนื่องจากข้อมูลที่ใช้ในงานวิจัยนี้เป็นข้อมูลที่มีการตัดพยางค์ จึงทำให้ชื่อเฉพาะที่เลือกมาได้มาจากการรวบพยางค์ที่มีความสัมพันธ์ระหว่างกัน ดังนั้นจำนวนชื่อเฉพาะที่เลือกมาได้จึงมีจำนวนมากตามไปด้วย ทั้งนี้เป็นเพราะกลุ่มพยางค์ที่เลือกมาได้อาจเป็นเพียงคำทั่วไปที่ไม่ใช่ชื่อเฉพาะก็ได้ เช่น คำว่า "สะดวก" ในคลังข้อมูลที่จะใช้ในวิธีทางสถิติคำนี้จะถูกตัดพยางค์ออกเป็น "สะ-ดวก" ทำให้ได้พยางค์ 2 พยางค์ซึ่งพบว่ามีค่าความสัมพันธ์ระหว่างหน่วยย่อยทั้งสองมีมาก ทำให้ในจำนวนของคำตอบ (response) ที่วิธีการทางสถิติคัดเลือกออกมา จึงมีคำว่า "สะดวก" ติดมาด้วย ซึ่งคำดังกล่าวก็มิได้เป็นชื่อเฉพาะแต่อย่างใด จึงทำให้ค่าความแม่นยำที่คำนวณจึงมีค่าต่ำ

ดังนั้น เพื่อการประมวลผลที่ดียิ่งขึ้น จึงอาจทดลองใช้ คลังข้อมูลที่มีการตัดคำ ซึ่งน่าจะให้จำนวนชื่อเฉพาะที่เลือกมามีจำนวนน้อยลงและมีโอกาสเป็นชื่อเฉพาะได้มากกว่า และจะทำให้ค่าความแม่นยำมีค่าสูงขึ้นด้วย อีกทั้งอาจช่วยย่นระยะเวลาในการประมวลผลจากโปรแกรมทางสถิติเหล่านั้นด้วย

ในส่วนที่สองซึ่งเป็นส่วนของวิธีที่ใช้กฎนั้น กฎที่ใช้เป็นกฎที่สร้างขึ้นจากข้อมูลที่ได้จากคลังข้อมูลนี้ทำให้การนำไปใช้กับคลังข้อมูลแบบอื่นอาจต้องมีการปรับกฎเพื่อใช้ในงานอื่นต่อไป เช่น อาจต้องเปลี่ยนรายการของคำยกเว้นหน้าหรือยกเว้นหลัง รวมทั้งอาจมีการเพิ่มรายการคำนำหน้าชื่อ เป็นต้น

อย่างไรก็ตาม การเปลี่ยนแปลงกฎทุกครั้งที่มีการเปลี่ยนแปลงข้อมูลในคลังข้อมูลก็ทำให้การประมวลผลต้องใช้เวลานานและทำให้ไม่สะดวกแก่ผู้ใช้ ดังนั้นจึงอาจจะทดลองเพิ่มขนาดของคลังข้อมูลเพื่อดูความแปรปรวนของข้อมูลและปรับปรุงกฎเพิ่มเติม จากนั้นจึงทดสอบประสิทธิภาพของระบบกฎที่พัฒนาขึ้น จนถึงจุดที่การเพิ่มขนาดของคลังข้อมูลไม่มีผลต่อกฎในระบบอีกต่อไป ก็จะทำให้ได้กฎที่สามารถใช้กับข้อมูลหลากหลายประเภทมากขึ้น

นอกจากนี้ กฎที่เขียนขึ้นนั้นยังมีข้อจำกัดเรื่องการหาขอบเขตของชื่อเฉพาะ ทั้งนี้เพราะภาษาไทยไม่มีสัญลักษณ์หรือลักษณะพิเศษ เช่น ตัวพิมพ์ใหญ่ ที่จะใช้บ่งบอกความแตกต่างระหว่างชื่อเฉพาะกับคำทั่วไปอย่างในภาษาอังกฤษ รวมทั้งไม่มีการแยกคำออกจากกันด้วยการเว้นวรรค อย่างไรก็ตาม การระบุจุดเริ่มต้นของขอบเขตชื่อเฉพาะนั้นอาจทำได้โดยการใช้หลักฐานภายในอย่างคำนำหน้าชื่อเป็นตัวระบุ แต่สำหรับการหาจุดสิ้นสุดของขอบเขตชื่อเฉพาะนั้น จากงานวิจัยนี้จะเห็นได้ว่าชื่อเฉพาะที่ถูกเลือกมา (candidate) จะมีกรณีที่วิธีกฎที่เขียนขึ้นไม่อาจครอบคลุมได้ อย่างเช่น กรณีที่เราอาจจะกำหนดให้คำว่า "ว่า" เป็นหนึ่งในรายการของคำที่เป็นจุดสิ้นสุดของขอบเขตของชื่อเฉพาะ ซึ่งจะทำให้สามารถระบุตำแหน่งและจำแนกประเภทชื่อเฉพาะได้ดังตัวอย่าง

พรรค<organization>ไทยรักไทย</organization>ว่า

แต่ก็อาจจะเกิดกรณีที่นอกเหนือจากนั้น ซึ่งทำให้กฎนี้ใช้ไม่ได้ผลและทำให้มีกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะเพิ่มจำนวนขึ้นมาอีก เช่น

พรรค<candidate>ไทยรักไทยเปิดเผย</candidate>ว่า

จะสังเกตได้ว่า จากคำนำหน้าชื่อซึ่งเป็นจุดเริ่มต้นคือคำว่า "พรรค" ผ่านเงื่อนไขของโปรแกรมเรียบร้อย แต่มาพบปัญหาอยู่ที่จุดสิ้นสุดของชื่อเฉพาะทำให้ได้ชื่อเฉพาะที่ไม่ถูกต้องปนเข้ามาอีกตัวหนึ่ง ยังผลให้ค่าความแม่นยำลดลงไปอีก

ซึ่งปัญหาดังกล่าวอาจแก้ไขได้โดยการเพิ่มโปรแกรมการเรียนรู้ด้วยเครื่อง (machine learning) ซึ่งเขียนขึ้นเพื่อให้สามารถรู้จำชื่อเฉพาะที่ถูกต้องเอาไว้ได้ หรืออาจกำหนดระดับ (threshold) หรือน้ำหนักของชื่อเฉพาะที่เลือกมา (candidate) เพื่อเอาไว้เปรียบเทียบและคัดกรองให้เหลือชื่อเฉพาะที่ถูกต้องให้ได้มากที่สุด เพื่อให้ระบบกฎมีอัตราการรู้จำหรือค่า F สูงกว่าเดิม