บทที่ 2

ระบบการรู้จำชื่อเฉพาะ

ในบทนี้จะแบ่งออกเป็น 4 ส่วน โดยในส่วนแรกจะกล่าวถึงความหมาย รูปแบบต่างๆ รวมถึงลักษณะของชื่อเฉพาะและการแบ่งประเภทของชื่อเฉพาะ จากนั้นในส่วนที่ 2 จะกล่าวถึงระบบการรู้จำตำแหน่งและจำแนกประเภทของชื่อเฉพาะ ซึ่งโดยทั่วไปแบ่งออกได้เป็น 3 ระบบคือ ระบบที่ใช้กฎ ระบบที่ใช้วิธีทางสถิติ และระบบแบบลูกผสมซึ่งรวมวิธีที่ใช้กฎและสถิติเข้าไว้ด้วยกัน ในส่วนที่ 3 จะกล่าวถึงคลังข้อมูลที่ใช้ในการวิจัยครั้งนี้ การวิเคราะห์ชื่อเฉพาะที่พบในคลังข้อมูล และในส่วนสุดท้ายจะกล่าวถึงระบบการรู้จำชื่อเฉพาะที่ใช้ในงานวิจัยนี้และการประเมินผลการทำงานของระบบ

2.1 ความหมาย รูปแบบต่างๆ และประเภทของชื่อเฉพาะ

ชื่อเฉพาะหรือวิสามานยนามตามความหมายจากพจนานุกรม ฉบับราชบัณฑิตยสถาน (2542) หมายถึงคำนามที่เป็นชื่อเฉพาะตั้งขึ้นไว้สำหรับเรียกคน สัตว์ สิ่งของ และสถานที่เพื่อให้รู้ชัดว่าเป็นใครหรืออะไร เช่น สมชาย (ชื่อคน) เป็นต้น

ชื่อเฉพาะนี้อาจมีคำเรียกเป็นภาษาอังกฤษได้หลายแบบ ได้แก่ proper name, named entity และ specific name เป็นต้น ซึ่งแต่ละแบบอาจใช้เป็นคำเรียกโดยคนต่างกลุ่มกันและอาจต้องการเน้นย้ำสิ่งที่ต่างกัน เช่น นักวิจัยทางด้านภาษาศาสตร์คอมพิวเตอร์มักจะใช้คำว่า named entity ในขณะที่นักภาษาศาสตร์มักจะใช้คำว่า proper name แต่โดยพื้นฐานแล้ว คำเรียกภาษาอังกฤษเหล่านี้มีลักษณะเหมือนกันคือ มีการสร้างคำขึ้นเพื่อแทนสิ่งที่ต้องการอ้างถึงโดยเฉพาะเจาะจง ซึ่งในวิทยานิพนธ์นี้จะใช้คำไทยเรียกคำกลุ่มนี้ว่า "ชื่อเฉพาะ"

สำหรับการแบ่งประเภทของชื่อเฉพาะนั้น มีงานวิจัยหลายงาน (Su and Zhou, 2002 ; Grishman, 1998) ที่มีการแบ่งประเภทของชื่อเฉพาะตามการประชุม MUCs หรือ Message Understanding Conferences โดยใน MUC-7(1998) ซึ่งเน้นที่การระบุหาชื่อเฉพาะ (identify named entity) เขาจะแบ่งประเภทชื่อเฉพาะหรือ named entity ออกเป็น 3 ประเภทคือ

1. entity names ได้แก่ ชื่อองค์กร ชื่อคน และ ชื่อสถานที่
2. temporal expression ได้แก่ วันที่ และ เวลาของวัน เช่น Christmas day
3. number expression ได้แก่ % และ จำนวนเงิน หรือ หน่วยเงิน

จะเห็นได้ว่าการแบ่งประเภทของ MUC-7 นั้นให้ความสำคัญกับ number expression ด้วย ทั้งนี้เป็นเพราะเอกสารหรือบทความที่เป็นคลังข้อมูลของระบบการระบุหาชื่อเฉพาะของ MUC-7 เป็นบทความด้านเศรษฐศาสตร์ ดังนั้นจึงให้ความสำคัญกับ number expression ซึ่งเป็น % และจำนวนเงินหรือหน่วยเงินด้วย อย่างไรก็ตาม งานวิจัยอื่นๆ (Chanlekha , 2002 ; Gaizauskas , 2000) อาจไม่แบ่งประเภทชื่อเฉพาะตาม MUC-7 ก็ได้โดยจะขึ้นอยู่กับประเภทของบทความหรือเอกสารที่นำมาทำเป็นคลังข้อมูล หรืออาจขึ้นกับจุดประสงค์ของผู้วิจัยว่าต้องการรู้จำตำแหน่งของชื่อเฉพาะใดออกจากบทความ เช่น บทความทางวิทยาศาสตร์ ดังในงานของ Demetriou, G. และคณะ (2000) ให้ความสนใจกับชื่อโปรตีน ดังนั้นในงานวิจัยนี้จึงเน้นการรู้จำตำแหน่งของชื่อเฉพาะที่เป็นชื่อโปรตีน เป็นต้น สำหรับงานวิทยานิพนธ์นี้นั้นจะให้ความสนใจที่ชื่อเฉพาะ 3 ประเภทคือ ชื่อคน ชื่อสถานที่และชื่อองค์กร ซึ่งเป็นชื่อเฉพาะประเภทหลักๆ ที่มักพบในบทความข่าวทั่วไปซึ่งเป็นคลังข้อมูลที่ใช้ในงานวิทยานิพนธ์นี้

สำหรับการแยกชื่อเฉพาะประเภทชื่อคน ชื่อองค์กรและชื่อสถานที่นั้น การแยกชื่อคนจะไม่มีปัญหามากเพราะสามารถระบุได้ไม่ยาก โดยอาจอาศัยคำนำหน้าชื่อเฉพาะ ยศ ตำแหน่งเป็นตัวบ่งชี้ได้ เช่น คำว่า "นาย" , "นางสาว" , "ร.ต.อ." เป็นต้น แต่ในการจำแนกประเภทของชื่อองค์กรและชื่อสถานที่นั้นอาจมีความกำกวมได้ เพราะอาจใช้คำนำหน้าชื่อเฉพาะร่วมกันได้ เช่น "มหาวิทยาลัย" จะใช้นำหน้าชื่อเฉพาะที่เป็นชื่อองค์กร เช่น "มหาวิทยาลัยธรรมศาสตร์" แต่อย่างไรก็ตาม ในบางครั้ง "มหาวิทยาลัยธรรมศาสตร์" อาจอ้างถึงสถานที่ก็ได้ เช่น นัดพบกันที่ "มหาวิทยาลัยธรรมศาสตร์" เพราะฉะนั้น ในการจำแนกประเภทชื่อเฉพาะประเภทชื่อองค์กรและชื่อสถานที่ จึงจำเป็นต้องอาศัยบริบทข้างเคียงซึ่งเป็นหลักฐานภายนอก (external evidence) มาช่วยในการตัดสิน

2.2 ระบบที่ใช้ในการรู้จำและจำแนกประเภทของชื่อเฉพาะ

ในส่วนนี้ ผู้วิจัยจะนำเสนอภาพโดยรวมของระบบการรู้จำแบบต่าง ๆ ที่ได้มีการพัฒนามาเพื่อให้เห็นภาพโดยรวมว่ามีระบบต่างๆ อย่างไรบ้าง โดยผู้วิจัยจะนำเสนอระบบที่ใช้ในงานวิจัยนี้ในตอนสุดท้ายของบทนี้

โดยทั่วไประบบการรู้จำและจำแนกประเภทชื่อเฉพาะส่วนใหญ่ อาจมีรายการของชื่อเฉพาะ (gazetteer) หรือ รายการของชื่อที่เป็นที่รู้จักกันดี (known name list) มาช่วยซึ่งเราอาจใช้วิธีการเปรียบเทียบรูปคำ (pattern matching) เพื่อระบุตำแหน่งและจำแนกประเภทชื่อเฉพาะได้

แต่วิธีการนี้ ก็ยังไม่เพียงพอต่อการระบุตำแหน่งและจำแนกประเภทของชื่อเฉพาะออกมาเพราะยังคงมีปัญหาดังต่อไปนี้

1. ชื่อเฉพาะมีจำนวนมากและมีการสร้างขึ้นใหม่ได้เรื่อยๆ จึงไม่สามารถเก็บรายการชื่อเฉพาะทั้งหมดไว้ในพจนานุกรมหรือจัดทำเป็น gazetteer ได้ครบถ้วน
2. ชื่อเฉพาะหนึ่งๆ สามารถปรากฏอยู่ในหลายรูปแบบโดยอาจอยู่ในรูปย่อ เช่น เอ็น.อี.ซี. เป็นชื่อย่อของบริษัทแห่งหนึ่ง หรือ ในบทความหรือเอกสารอาจมีการเรียกย่อๆ เมื่อพูดถึงคนๆ เดิมในครั้งต่อมา เช่น ชัชวาล คงอุดมศาสตร์ อาจเรียกย่อๆ ว่า ชัชวาล เป็นต้น ซึ่งทำให้ไม่อาจจะบรรจุเก็บไว้ในรายการของชื่อเฉพาะ (gazetteer) ได้ทั้งหมด

นอกจากนี้ จากการที่ลักษณะเฉพาะของภาษาแถบยุโรปเช่น ภาษาอังกฤษ และภาษาในแถบเอเชียอย่างเช่น ภาษาจีน ภาษาญี่ปุ่น และภาษาไทย ต่างกันทำให้ระบบที่ใช้อาจต่างออกไปด้วย โดยระบบที่ใช้จัดการกับภาษาจีน ญี่ปุ่น และภาษาไทยนั้นจะต้องจัดการกับปัญหาสำคัญอย่างหนึ่งที่ไม่มีในภาษาอังกฤษคือ ภาษาในแถบเอเชียนี้ไม่มีเครื่องหมายพิเศษหรือการใช้ช่องว่างเพื่อแบ่งคำออกจากกัน ทำให้การระบุขอบเขตของคำเป็นไปได้ยากกว่าในภาษาอังกฤษ ซึ่งทำให้เกิดปัญหาในการระบุขอบเขตของชื่อเฉพาะอีกด้วย ตัวอย่างเช่น ในภาษาไทยซึ่งเป็นภาษาที่มีหลายพยางค์ทำให้เกิดความกำกวมว่าจะตัดให้เป็นคำที่ตำแหน่งใด หากเกิดการตัดคำผิดที่ความหมายจากการตัดคำที่ได้ก็อาจจะต่างจากความหมายที่ต้องการจะสื่อถึง อย่างเช่น "ตากลม" ในประโยค "เด็กคนนั้นกำลังยืนตากลม" การตัดทำได้ 2 แบบ คือตัดแล้วได้คำว่า "ตาก-ลม" ออกมาซึ่งเป็นตำแหน่งการตัดที่ถูกต้อง คือ ทำให้ประโยคมีความหมาย ส่วนการตัดอีกแบบจะได้ "ตา-กลม" ซึ่งทำให้ประโยคมีความหมายผิดไปหากมีการตัดคำแบบนี้ อีกทั้งภาษาไทยยังไม่มีข้อมูลจากลักษณะอักษรที่ช่วยในการหาตำแหน่ง รวมทั้งการระบุขอบเขตของคำและไม่มีการระบุด้วยเครื่องหมายวรรคตอนหรืออื่นๆ ในกรณีของชื่อเฉพาะ เช่นในภาษาทางตะวันตก ได้แก่ ภาษาอังกฤษ โดยในภาษาไทยไม่มีการใช้ตัวพิมพ์ใหญ่หรือ capitalization ซึ่งเป็นปัจจัยที่สำคัญอีกอย่างหนึ่งในการระบุตำแหน่งของชื่อเหมือนในภาษาอังกฤษที่ชื่อมักขึ้นต้นด้วยตัวพิมพ์ใหญ่อีกด้วย นอกจากนี้ ภาษาไทยยังมีปัญหาเรื่อง ตัวย่อ (abbreviation) กล่าวคือถึงแม้ว่าชื่อเฉพาะหลายตัวโดยเฉพาะ ชื่อองค์กร อาจปรากฏอยู่ในรูปคำย่อได้ แต่ก็ไม่ใช่ทุกตัวย่อของคำที่พบในบทความแล้วจะเป็นชื่อเฉพาะ (named entity) หรือถ้าเป็นก็ไม่แน่ว่าจะอยู่ในประเภทของชื่อองค์กรหรือไม่ เช่น ด.ช. ไม่ได้เป็นชื่อองค์กรแต่เป็นคำนำหน้าชื่อที่ย่อมาจากเด็กชาย เป็นต้น

สำหรับระบบที่ใช้แก้ปัญหาในการระบุหรือรู้จำตำแหน่งและจำแนกประเภทชื่อเฉพาะนั้น สามารถแบ่งจากการทบทวนวรรณกรรมออกเป็น 3 ระบบด้วยกัน คือ

2.2.1 **ระบบที่ใช้วิธีกฎ** เป็นระบบที่รู้จำตำแหน่งและจำแนกประเภทของชื่อเฉพาะโดยอาศัยกฎที่เขียนขึ้นมา ซึ่งอาจใช้นักภาษาศาสตร์ที่เชี่ยวชาญภาษานั้นเป็นอย่างดีมาช่วยสร้างกฎให้ และมักใช้หลักฐาน 2 อย่างคือ บริบทภายใน (internal evidence) และ บริบทข้างเคียง (external evidence) ในการระบุตำแหน่งของชื่อเฉพาะในเอกสารรวมทั้งการแยกประเภทของชื่อเฉพาะด้วย (Zhou and Su, 2002) โดยอาจจะใช้รายการของชื่อที่เป็นที่รู้จักกันดีหรืออาจมีรายการของคำสำคัญที่บ่งบอกตำแหน่งของชื่อได้อย่างคำนำหน้าชื่อ เช่น Mr. หรือ Mrs. ในภาษาอังกฤษ นาย , นางหรือนางสาว ในภาษาไทย ซึ่งคำนำหน้าชื่อเหล่านี้เป็นบริบทภายใน (internal evidence) อีกแบบหนึ่งซึ่งบ่งบอกว่าคำที่ตามมาจะมีแนวโน้มที่จะเป็นชื่อคน ร่วมกับการสร้างกฎที่พิจารณาบริบทข้างเคียง (external evidence) ที่เป็นบริบทซึ่งอยู่ข้างเคียงกับชื่อเฉพาะหนึ่งๆ เพื่อดูว่าบริบทแบบใดที่ชื่อเฉพาะปรากฏร่วมอยู่ด้วย เช่น ในภาษาอังกฤษ คำว่า arrive มักจะปรากฏตามหลังชื่อเฉพาะประเภทชื่อคน หรือในภาษาไทย เช่น คำว่า กิน ที่จะปรากฏตามหลังชื่อเฉพาะประเภทชื่อคนมากกว่าที่จะปรากฏร่วมกับชื่อเฉพาะประเภทอื่น เป็นต้น

ถึงแม้ว่าระบบที่ใช้จะมีประสิทธิภาพที่ดี แต่ก็ยังมีข้อจำกัดบางข้อ ได้แก่ ข้อมูลอาจจะไม่ทันสมัยเพราะชื่อเฉพาะเกิดขึ้นใหม่อยู่ตลอดเวลาทำให้การสร้างกฎอาจไม่ทันสมัยพอ ที่จะจัดการกับชื่อเฉพาะรูปแบบใหม่ๆ อีกทั้ง การเขียนกฎต้องอาศัยการเขียนกฎด้วยมือของนักภาษาหรือนักภาษาศาสตร์ของภาษานั้นๆ ที่มีประสบการณ์ และอาจต้องใช้เวลานานในการเขียน นอกจากนี้กฎยังขึ้นกับลักษณะเฉพาะพิเศษของภาษา ประเภทและรูปแบบของ บทความหรือเอกสาร ด้วย ทำให้ใช้ได้กับภาษาที่กำลังในคลังข้อมูลที่กำลังพิจารณาอยู่เท่านั้น

ซึ่งแนวทางในการแก้ไขข้อจำกัดของระบบที่ใช้กฎอาจใช้วิธีการเรียนรู้ด้วยเครื่อง (machine learning) เข้าช่วย โดยอาจเรียนรู้ลักษณะการปรากฏร่วมของลำดับของคำและลำดับของหมวดคำ เช่น คำนาม คำกริยา ฯลฯ รอบๆ ชื่อเฉพาะจากคลังข้อมูล (corpus) ทำให้เครื่องมือที่ใช้ในการกำกับหมวดคำ (Part-of-Speech tagger) เป็นเครื่องมือที่สำคัญอีกเครื่องมือหนึ่งในการนี้ และจากนั้นเมื่อระบบการเรียนรู้ด้วยเครื่องพบชื่อเฉพาะใหม่ที่มีลักษณะใกล้เคียงกับชื่อเฉพาะที่ผ่านการเรียนรู้ไปแล้วก็จะจัดการให้รู้จำชื่อเฉพาะแบบใหม่นี้ได้ โดยอาศัยลักษณะต่างๆ ที่ได้จากการเรียนรู้จากชื่อเฉพาะเดิมที่มีการรู้จำไปแล้ว หรือเครื่องอาจทำการเรียนรู้แล้วสร้างเป็นกฎได้จากคลังข้อมูลเลยโดยไม่ต้องอาศัยนักภาษาศาสตร์ในการสร้างกฎต่างๆ ทำให้การเรียนรู้ด้วยเครื่อง

แล้วสร้างกฎจากคลังข้อมูลนี้จะแก้ปัญหาชื่อเฉพาะที่เพิ่มเข้ามาใหม่เรื่อยๆ เมื่อเวลาผ่านไป ซึ่งข้อดีของระบบการเรียนรู้ด้วยเครื่องจะมีดังนี้

1. ไม่จำเป็นต้องใช้การวิเคราะห์จากนักภาษาศาสตร์ซึ่งมีค่าใช้จ่ายสูง
2. ไม่ต้องใช้คนช่วยอีกหากต้องการเปลี่ยนไปใช้กับตัวบทในอีกเนื้อหา (domain) หนึ่ง
3. ง่ายต่อการปรับใช้กับอีกภาษาหนึ่ง

ซึ่งการเรียนรู้ด้วยเครื่องนี้สามารถทำได้หลายวิธี วิธีหนึ่งที่เป็นที่นิยมคือ วิธีดีซิชั่นทรี* (decision tree) ซึ่งมีงานวิจัยหลายงานได้นำมาประยุกต์เพื่อใช้กับภาษาต่างๆ ดังที่มีการใช้เพื่อแก้ปัญหาข้อจำกัดของระบบกฎที่ไม่ครอบคลุมชื่อเฉพาะแบบใหม่ๆ ของภาษากรีกและภาษาฝรั่งเศส (Karkaletsis และคณะ, 2001) หรือการปรับใช้กับภาษาญี่ปุ่น ที่ได้มีการปรับกฎให้ทันสมัยโดยใช้วิธีดีซิชั่นทรี (Isozaki, 2001) หรือในงานของ Sassano และ Utsuro (2000) ที่ได้มีการนำเอาวิธีดีซิชั่นทรีมาใช้เพื่อทำการเดาขอบเขตของชื่อเฉพาะภาษาญี่ปุ่น โดยเขาจะทำรายการของกฎการตัดสินใจ (decision list) เพื่อทำการเดาขอบเขตของชื่อเฉพาะ หรือในงานของ Grishman และคณะ (1998) ที่มีการใช้ วิธีดีซิชั่นทรี ในการเรียนรู้ขอบเขตของชื่อเฉพาะ และคำบริบทที่อยู่ด้านหน้าและด้านหลังของชื่อเฉพาะ ในข้อมูลที่ใช้ฝึก เพื่อทำการรู้จำและจำแนกประเภทของชื่อเฉพาะ เป็นต้น

นอกจากวิธีดีซิชั่นทรีแล้ว วิธีอื่นๆ ที่มีผู้นำมาประยุกต์ใช้ในการรู้จำชื่อเฉพาะได้แก่ การใช้ EM-style bootstrapping algorithm ดังในงานของ Cucerzan และ Yarowsky (1999) ที่ใช้กับภาษาโรมาเนีย, การใช้ maximum entropy model ดังในงานของ Bender และคณะ (2003) ที่ใช้กับ ภาษาอังกฤษและเยอรมัน , การใช้ memory-based algorithm ดังในงานของ Erik F. Tjong และ Kim Sang (2002) ที่มีการปรับใช้กับภาษาเสปนและภาษาดัตช์ , การใช้วิธี support vector ดังในงานของ Isozaki และ Kazawa (2002) ซึ่งปรับใช้กับภาษาญี่ปุ่น ส่วนงานวิจัยที่ทำกับภาษาไทยนั้นก็มีการนำเอา winnow algorithm ที่เป็นระบบการเรียนรู้ด้วยเครื่อง (learning algorithm) อีกแบบหนึ่งมาใช้ในการรู้จำและจำแนกประเภทชื่อเฉพาะภาษาไทย ดังเช่นในงานของ Charoenpornsawat, P. และคณะ (1998,1999) เป็นต้น

ถึงแม้จะพบว่าระบบการเรียนรู้ด้วยเครื่องแต่ละวิธีที่กล่าวมาข้างต้นนั้น ให้ประสิทธิภาพที่ดีกับระบบการรู้จำและจำแนกประเภทชื่อเฉพาะของภาษานั้นๆ และแก้ปัญหาที่เกิดจากข้อจำกัดของระบบที่ใช้กฎได้ แต่สำหรับในงานวิจัยนี้ไม่สามารถนำเอาระบบการเรียนรู้ด้วยเครื่องเหล่านี้มา

---

* วิธีดีซิชั่นทรี (decision tree) คือวิธีที่มีการใช้รายการของคุณสมบัติ หรือ feature ต่างๆ เป็นตัวกำหนดการตัดสินใจว่าคำที่เข้ามาในระบบซึ่งใช้วิธีนี้นั้นเป็นชื่อเฉพาะหรือไม่ เงื่อนไขการตัดสินใจต่างๆ จะเรียงต่อเนื่องกันเป็นเหมือนโครงสร้างต้นไม้ไล่จากบนลงล่าง หากคำดังกล่าวสามารถผ่านเงื่อนไขของคุณสมบัติเหล่านั้นมาได้ก็จะสามารถรู้จำได้ว่าคำคำนั้นเป็นชื่อเฉพาะ

ปรับใช้กับคลังข้อมูลภาษาไทยได้ ทั้งนี้เป็นเพราะวิธีการต่างๆ นี้อาศัยเครื่องมือที่ใช้ในการกำกับหมวดคำ (Part-of-Speech tagger) ซึ่งโปรแกรมกำกับหมวดคำภาษาไทยนั้นยังเป็นหัวข้อวิจัยที่สามารถพัฒนาต่อไปได้ นอกจากนี้ ผู้วิจัยยังเห็นว่า ในภาษาไทยนั้น การรู้จำชื่อเฉพาะเป็นกระบวนการที่อาจจะต้องประมวลไปพร้อมๆ กับการแยกคำและกำกับหมวดคำ จึงจะไม่นำเครื่องมือที่ใช้ในการกำกับหมวดคำ (Part-of-Speech tagger) มาใช้ร่วมในงานวิจัยนี้ด้วย

2.2.2. **ระบบที่ใช้วิธีการทางสถิติเป็นหลัก** : ระบบแบบนี้มักจะเน้นที่ค่าความถี่และการปรากฏร่วมของคำที่มีความสัมพันธ์กัน โดยไม่มีหลักการทางภาษาศาสตร์เข้ามาเกี่ยวข้อง แนวคิดนี้ใช้กันมากในการหาศัพท์เฉพาะทาง (term) ที่ประกอบด้วยคำย่อยๆ หลายคำ (mulltiword) แต่ในที่นี้ ผู้วิจัยนำมาเพื่อใช้กับการหาชื่อเฉพาะด้วยเพราะคิดว่าใช้หลักการเดียวกันได้ คือ อาจใช้วิธีการพิจารณาค่าความน่าจะเป็นของลำดับของคำที่เรียงต่อกันโดยนับการปรากฏร่วมของคำ 2 คำ ซึ่งจะถือว่าคำที่มีความถี่สูงๆ มีแนวโน้มที่จะเป็นชื่อเฉพาะ หรือในกรณีที่คำ 2 คำหรือเป็นกลุ่มคำที่มากกว่า 2 ปรากฏร่วมกันบ่อยๆ เช่น "สิริ" กับ "กานดา" ซึ่งเป็นคำ 2 คำที่มีความหมายและหากพบว่าปรากฏร่วมกันบ่อยๆ ในบทความ กลุ่มคำนั้นก็อาจจะมีความสัมพันธ์ระหว่างกันมาก ก็มีแนวโน้มที่จะเป็นชื่อคน คือ "สิริกานดา" เช่นกัน ซึ่งวิธีทางสถิติที่ใช้พิจารณาการปรากฏร่วมของคำเป็นหน่วยเดียว สามารถจำแนกเป็นสองประเภท คือ วิธีทางสถิติแบบที่มองความสัมพันธ์การปรากฏร่วมกันระหว่างหน่วยย่อยๆ ภายในว่าพบมากเกินปกติหรือไม่ ซึ่งได้แก่ การใช้ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio (MI^3) , ค่า Dunning's Log Likelihood และวิธีการทางสถิติแบบที่มองการสูญเสียหน่วยใดๆ ภายในว่ายอมรับได้มากหรือไม่ คือ ค่า mutual Expectation ซึ่งวิธีการทางสถิติเหล่านี้จะเป็นไปดังนี้คือ

2.2.2.1 **วิธีทางสถิติแบบที่มองความสัมพันธ์การปรากฏร่วมกันระหว่างหน่วยย่อย ๆ ภายใน** ได้แก่

2.2.2.1.1 **การหาค่า Mutual Information** (Lopes and Silva, 1999) ซึ่งเป็นการวัดความสัมพันธ์ระหว่างคำ 2 คำ โดยถ้าค่า Mutual Information มากกว่า 0 มากๆ ก็แสดงว่าคำ 2 คำมันมีความสัมพันธ์กัน ดังแสดงในสมการที่ 1 คือถ้า คำที่ 1 (w_1) และคำที่ 2 (w_2) มีความสัมพันธ์กัน ค่าความน่าจะเป็นที่จะพบ w_1 และ w_2 อยู่ด้วยกัน ($P(w_1,w_2)$) ก็จะมีค่าสูงกว่าค่าความน่าจะเป็นที่จะพบ w_1 หรือ w_2 เดี่ยวๆ ($P(w_1)$ และ $P(w_2)$)

$$I(w_1, w_2) = \log \frac{p(w_1 w_2)}{p(w_1)\,p(w_2)} \qquad \text{---------- 1}$$

อย่างไรก็ตามวิธีที่ใช้ในการหาค่า Mutual Information (MI) นี้ก็มีข้อเสียคือจะให้ความสำคัญกับคำที่มีความถี่น้อย คือคำที่มีความถี่น้อย ค่า MI จะมีค่าสูง ซึ่งอาจแก้โดยใช้การหาค่า Cubic association ratio (MI^3) เพื่อให้น้ำหนักกับเหตุการณ์ที่เกิดบ่อยมากขึ้น

2.2.2.1.2 **การหาค่า Cubic association ratio (MI^3)** (Lopes and Silva, 1999) เป็นไปดังสมการด้านล่าง

$$MI^3 = log_2 \frac{a^3 N}{(a+b)(a+c)}$$

จากสูตร *a* คือ ค่าความถี่ของ bigram ของคำที่1 และ 2 ($w_1 - w_2$)

b คือ ค่าความถี่ของ bigram ของคำที่ไม่ใช่คำที่ 1 และ คำที่ 2 ($\sim w_1 - w_2$)

c คือ ค่าความถี่ของ bigram ของคำที่1 และ คำที่ไม่ใช่คำที่ 2 ($w_1 - \sim w_2$)

2.2.2.1.3 **การหาค่า Pearson's Chi-square** (Lopes and Silva, 1999) ซึ่งเป็นการทดสอบสมมติฐานที่ว่าคำ w_1 และ w_2 ปรากฏร่วมกันโดยบังเอิญหรือไม่ โดยการคำนวณตามสูตร

$$\chi^2 = \frac{\sum_{i,j}(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

โดยที่ $O_{i,j}$ คือ ค่าความถี่จากการสังเกตของค่า 4 ค่า ดังตาราง

O_{11} = ความถี่ของการปรากฏร่วมระหว่างคำ w_1 และ $\_w_2$	O_{12} = ความถี่ของการเกิดระหว่างคำที่ไม่ใช่ w_1 ที่ปรากฏร่วมกับ w_2
O_{21} = ความถี่ของการเกิดระหว่างคำ w_1 แต่ไม่ได้ตามด้วยคำ w_2	O_{22} = ความถี่ของการปรากฏร่วมระหว่างคำที่ไม่ใช่ทั้ง w_1 และ w_2

ส่วนค่า $E_{i,j}$ คือค่าความถี่ของแต่ละเซลล์ในตารางเมื่อคำ w_1 และ w_2 ที่เกิดร่วมกันโดยบังเอิญ ซึ่งค่า $E_{i,j}$ ของแต่ละเซลล์ในตารางจะหาได้จาก ผลรวมของแต่ละแถวคูณกับผลรวมของแต่ละสดมภ์หารด้วยผลรวมของค่าความถี่ทั้งหมด

ซึ่งถ้าหากค่า Chi-square ที่หามาได้มีค่าสูงมากเท่าใด ก็แสดงว่าระหว่างคำ w_1 และ w_2 ยิ่งมีความสัมพันธ์กันมากเท่านั้น

2.2.2.1.4 **การหาค่า Dunning's log likelihood** (Lopes and Silva, 1999) ซึ่งจะคำนวณได้ดังนี้

$Loglike\ (w_1,w_2) = 2 * (log\ l(p1,k1,n1) + log\ l(p2,k2,n2) - log\ l(p,k1,n1) - log\ l(p,k2,n2))$

เมื่อ $log\ l(P,K,M) = K * ln(P) + (M - K) * ln(1 - P)$

โดยที่

$k1$ คือ ความถี่ของ bigram w_1- $w_2 = f(w_1, w_2)$

$k2 = f(w_1) - k1$

$n1 = f(w_2)$

$n2 = N - n1$

$p1 = k1/n1 = f(w_1, w_2) / f(w_2)$

$p2 = k2/n2 = (f(w_1) - f(w_1, w_2)) / (N - f(w_2))$

$p = (k1 + k2)/N = f(w_1)/N$

N คือ จำนวนคำทั้งหมดในคลังข้อมูล

จากนั้นคูณค่า log likelihood ratio ด้วย -2 แล้วดูค่าจากตาราง chi-square เมื่อดูค่านัยสำคัญที่ 0.005 (df = 1) ต้องมากกว่า 7.88 จึงจะล้มสมมติฐานต้นที่ว่าคำสองคำนั้นเป็นอิสระจากกันได้

ซึ่งการอาศัยแค่ข้อมูลความถี่ ฯลฯ นี้ ทำให้ระบบที่ใช้แค่ข้อมูลสถิติแบบนี้เกิดข้อเสียที่เห็นได้ชัด คือ วิธีการนี้จะมองว่าชื่อเฉพาะเป็นข้อมูลที่สำคัญในบทความหรือเอกสาร จึงควรจะถูกกล่าวถึงบ่อยๆ ดังนั้นคำอื่นๆ ที่มีความถี่ต่ำๆ จึงไม่น่าจะใช่ชื่อเฉพาะ แต่ในความเป็นจริงแล้วชื่อเฉพาะ ก็อาจมีความถี่น้อยเช่น อาจปรากฏเพียง 1 หรือ 2 ครั้งก็ได้ ทำให้การตัดชื่อเฉพาะที่มีความถี่น้อยกว่า 2 ออกจากกลุ่มของกลุ่มคำที่อาจเป็นชื่อเฉพาะ (candidate) ในระบบจะทำให้เกิดความผิดพลาดขึ้นได้

การที่ชื่อเฉพาะมีความถี่ในข้อมูลใช้ฝึกต่ำนั้นอาจเกิดจากขนาดข้อมูลที่อยู่ในคลังข้อมูลใช้ฝึก (training corpus) นั้นไม่เพียงพอหรือเกิดจาก data sparseness คือภายในคลังข้อมูลใช้ฝึกเองก็อาจมีตัวอย่างของชื่อเฉพาะ และบริบทของมันไม่เพียงพอ จึงเป็นไปได้ที่จะพบชื่อเฉพาะบางอย่างในคลังข้อมูลใช้ทดสอบ (test corpus) หรือ ในบทความหรือเอกสารจริงๆ ที่ใช้กันอยู่ แต่ไม่พบในคลังข้อมูลใช้ฝึก ทั้งนี้เป็นเพราะถ้าหากจะเก็บตัวอย่างของชื่อเฉพาะ และบริบทของมันไว้ได้หมดในคลังข้อมูลใช้ฝึก ก็คงต้องใช้พื้นที่ในการเก็บเป็นจำนวนมากซึ่งเป็นการเปลืองเนื้อที่ในการเก็บ อีกทั้งชื่อเฉพาะที่เกิดขึ้นใหม่อาจมีบริบทที่ต่างไปจากเดิมทำให้ข้อมูลมีมากจนเก็บในคลังข้อมูลใช้ฝึกได้ไม่หมด

2.2.2.2 **วิธีการทางสถิติแบบที่มองการสูญเสียหน่วยย่อยใดๆ ภายในว่ายอมรับได้มากหรือไม่** ได้แก่

2.2.2.2.1 **การหาค่า** Mutual Expectation (Dias และคณะ, 2000)

แนวคิดอีกแนวคิดหนึ่งเป็นการประเมินค่าระดับความสัมพันธ์ที่มีอยู่ระหว่างกลุ่มคำ โดยจะใช้ค่า Mutual Expectation ในการประเมินระดับความแข็งแรงของความสัมพันธ์ที่เชื่อมระหว่างคำที่อยู่ใน n-gram ซึ่งใช้เพื่อหาคำประกอบ (multiword lexical unit) วิธีนี้เป็นวิธีที่น่าสนใจอีกวิธีหนึ่ง ทั้งนี้เป็นเพราะไม่จำเป็นต้องใช้เครื่องมือในการกำกับหมวดคำ ทำให้วิธีนี้เป็นวิธีที่น่าจะนำมาปรับใช้กับระบบการระบุและจำแนกประเภทชื่อเฉพาะภาษาไทยของเราได้และวิธีการนี้สามารถใช้กับการหาความสัมพันธ์ของกลุ่มคำที่มีมากกว่า 2 คำขึ้นไป

ซึ่งค่า Mutual Expectation จะหาได้จากสมการค่า Normalised Expectation (Dias และคณะ, 2000) ดังมีรายละเอียดต่อไปนี้

กระบวนการขั้นแรกที่จะต้องทำเมื่อรับข้อความหรือเอกสารเข้ามาแล้วคือ การปรับให้เอกสารเป็นเซตของ n-gram ดังในงานของ Dias และคณะ (2000) ที่ทดลองกับคลังข้อมูลภาษาสโลวาเนียและงานของ Bassano และคณะ (2000) ที่ทดลองกับภาษาโปรตุเกส โดย n-gram ที่ใช้แสดง แทนคำจำนวน n คำ ซึ่งแต่ละคำจะถูกบ่งชี้ด้วยระยะทางระหว่างมันกับคำแกนกลาง โดยกำหนดให้คำแกนกลางเป็นคำแรกของ n-gram ซึ่งจะสามารถแทน n-gram ได้เป็น $[w_1p_{12}w_2p_{13}w_3...p_{1i}w_i...p_{1n}w_n]$ โดยที่ p_{1i} (เมื่อ i = 2 ถึง n) แสดงระยะห่างระหว่างคำ w_i และคำแกนกลาง w_1

จากนั้นก็จะหาค่า Normalised Expectation โดยนิยามค่า Normalised Expectation ที่มีอยู่ระหว่าง n คำ ว่าเป็นค่าเฉลี่ยการคาดหวังว่ามีคำๆ นี้แล้วจะเกิดคำนั้นๆ ถัดไป ซึ่งแนวคิดพื้นฐานของค่า Normalised Expectation คือจะประเมินค่าความสัมพันธ์ โดยจะพิจารณาดูว่าจะเสียคำใดคำหนึ่งใน n-gram ได้หรือไม่ ซึ่งถ้าไม่สามารถยอมรับการสูญเสียคำๆ นั้นไปก็แสดงว่า n-gram ณ ที่นี้ก็มีโอกาสที่จะเป็นศัพท์ (term) ที่ต่อเนื่องกัน โดยอาจเป็นชื่อเฉพาะที่มีขนาดยาวๆ ก็ได้ และยิ่งค่า Normalised Expectation สูงมากเท่าใด การไม่ยอมรับการสูญเสียคำนี้ก็ยิ่งสูงตามไปด้วย ยกตัวอย่างเช่น ค่า Normalised Expectation (NE) สำหรับคำ 3 คำคือ Linux Operating System โดยที่ Linux เป็นคำแกนกลาง ซึ่งค่า NE จะหาโดยวัดเป็นค่าการสูญเสีย 1 ใน 3 คำเดี่ยวครั้งละ 1 คำ ดังนั้นค่าความคาดหวังเฉลี่ยของ 3-gram จะเป็นการคาดหวังว่าจะพบคำ System หลังจากพบคำว่า Linux Operating และยังคาดว่า Operating จะเชื่อมระหว่าง Linux และ System ด้วย อีกทั้งยังคาดหวังว่าจะพบคำว่า Linux ก่อน Operating System โดยจะดูว่าจะยอมรับการสูญเสียคำเหล่านี้ไปได้หรือไม่

หลักการของการหาค่า Normalised Expectation คือจะวัดความคาดหวังของการเกิดของเหตุการณ์ $X = x$ เมื่อรู้เหตุการณ์ $Y = y$ ดังสมการด้านล่าง

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

โดย $p(X = x\ ,\ Y = y)$ เป็นค่าความน่าจะเป็นที่จะพบเหตุการณ์ x,y

$p(Y = y)$ เป็นค่าความน่าจะเป็นที่จะเกิดเหตุการณ์ y

ดังนั้นเมื่อพิจารณา n-gram จะพบว่าการสูญเสียคำในแต่ละครั้งไปจะแสดงดังตารางที่ 1 ซึ่งเส้นจะหมายถึงคำที่หายไป เมื่อรู้ n-1 gram

ตารางที่ 1 : (n-1)-grams and missing words.

(n-1)-gram	missing word
[_____$p_{12}w_2\,p_{23}w_3 \ldots p_{2i}w_i \ldots p_{2n}w_n$]	w_1
[w_1_____$p_{13}w_3 \ldots p_{1i}w_i \ldots p_{1n}w_n$]	w_2
...	...
[$w_1p_{12}w_2p_{13}w_3 \ldots p_{1(i-1)}w_{(i-1)}$_____$p_{1(i+1)}w_{(i+1)} \ldots p_{1n}w_n$]	w_i
...	...
[$w_1p_{12}w_2p_{13}w_3 \ldots p_{1i}w_i \ldots\ p_{1(i-1)}w_{(i-1)}$_____]	w_n

ซึ่งค่าความน่าจะเป็นเมื่อตัดคำแกนกลาง (w_1) ของ n-grams ออกจะเป็นดังสมการ a

$$P(w_1 \mid [w_2 \ldots p_{2i}w_i \ldots p_{2n}w_n]) = \frac{p([w_1p_{12}w_2 \ldots p_{2i}w_i \ldots p_{2n}w_n])}{p([w_2 \ldots p_{2i}w_i \ldots p_{2n}w_n])} \quad \text{a}$$

สมการ b จะเป็นการประเมินค่าของการสูญเสียคำอื่นๆ ใน n-grams

$$\forall_{i,i} = 2..n,$$

$$P(w_i \mid [w_1 \ldots p_{1(i-1)}w_{(i-1)}p_{1(i+1)}w_{(i+1)} \ldots p_{1n}w_n]) = \frac{p([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n])}{p([w_1 \ldots p_{1(i-1)}w_{(i-1)}p_{1(i+1)}w_{(i+1)} \ldots p_{1n}w_n])} \quad \text{b}$$

จากสมการ a และ b พบว่าตัวเศษจะไม่เปลี่ยนค่ามีแค่ส่วนเท่านั้นที่เปลี่ยนไป ดังนั้นจึงประเมินค่าจุดศูนย์กลางของส่วน ซึ่งอาจเรียกว่า Fair Point of Expectation (FPE) ซึ่งค่า FPE จะแสดงได้ดังสมการ c

$$FPE([w_1p_{12}w_2 \ldots p_{1i}w_i \ldots p_{1n}w_n])$$
$$= 1/n * \left(p([w_2 \ldots p_{2i}w_i \ldots p_{2n}w_n]) + \sum_{i=2}^{n} p([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n]) \right) \quad \text{c}$$

โดย $p([w_2 \ldots p_{2i}w_i \ldots p_{2n}w_n])$ เมื่อ i มีค่าตั้งแต่ 3 ถึง n เป็นค่าความน่าจะเป็นของการเกิด n-1 gram ($[w_2 \ldots p_{2i}w_i \ldots p_{2n}w_n]$) ซึ่งเป็นผลจากการสูญเสีย w_1 จาก n-grams ทั้งสาย และ

$p([w_1...p_{1i}w_i...p_{1n}w_n])$ เป็นค่าความน่าจะเป็นของการเกิด n-1 gram ที่มีคำแรก w_1 ส่วนคำ $p_{1i}w_i$ จะหมายถึงคำนั้นเป็นคำที่หายไป

ค่า Fair Point of Expectation นี้นำมาใช้ในการหาค่า Normalised Expectation ดังแสดงในสมการ d

$$NE([w_1...p_{1i}w_i...p_{1n}w_n]) = \frac{p([w_1...p_{1i}w_i...p_{1n}w_n])}{FPE([w_1p_{12}w_2...p_{1i}w_i...p_{1n}w_n])} \quad \text{_____ d}$$

ซึ่ง $p([w_1...p_{1i}w_i...p_{1n}w_n])$ เป็นค่าความถี่ของ $[w_1...p_{1i}w_i...p_{1n}w_n]$ และ

ค่า $FPE([w_1p_{12}w_2...p_{1i}w_i...p_{1n}w_n])$ ก็จะหาได้จากสมการ c

เพราะฉะนั้น จากค่า $NE([w_1...p_{1i}w_i...p_{1n}w_n])$ ที่ได้จากสมการ d เราก็จะหาค่า Mutual Expectation (ME) ระหว่างคำ n คำ ได้ดังสมการที่ e

$ME([w_1...p_{1i}w_i...p_{1n}w_n]) = p([w_1...p_{1i}w_i...p_{1n}w_n]) \times NE([w_1...p_{1i}w_i...p_{1n}w_n])$ ___e

โดย $p([w_1...p_{1i}w_i...p_{1n}w_n])$ เป็นค่าความน่าจะเป็นของ n-grams $([w_1...p_{1i}w_i...p_{1n}w_n])$

$NE([w_1...p_{1i}w_i...p_{1n}w_n])$ เป็นค่า Normalised Expectation ของ n-grams นั้น $([w_1...p_{1i}w_i...p_{1n}w_n])$

2.2.3 **ระบบแบบลูกผสม (hybrid)** : ระบบแบบลูกผสมนี้จะเป็นระบบที่รวมทั้งวิธีที่ใช้กฎและวิธีที่ใช้ข้อมูลทางสถิติไว้ร่วมกัน ซึ่งจะช่วยลดข้อจำกัดของทั้งระบบแบบที่ใช้วิธีกฎและระบบที่ใช้วิธีการทางสถิติได้ โดยถ้าหาก ระบบสามารถใช้ข้อมูลทางสถิติในการหากลุ่มคำที่น่าจะเป็นชื่อเฉพาะ ก็จะไม่ติดปัญหาเรื่องชื่อเฉพาะแบบใหม่ๆ ที่กฎไม่อาจเข้าไปจัดการได้เพราะไม่ทันสมัยเพียงพอ ทั้งนี้เป็นเพราะชื่อเฉพาะที่ถึงแม้ว่าจะเป็นชื่อเฉพาะใหม่แต่ก็ยังมีคุณสมบัติเดียวกับชื่อเฉพาะเดิมคือ เป็นข้อมูลที่สำคัญในบทความ ดังนั้นจึงน่าจะมีการกล่าวถึงบ่อยๆ ทำให้สามารถสังเกตจากค่าความถี่ และดูการปรากฏร่วมของคำเพื่อประกอบการพิจารณาในกรณีที่ชื่อเฉพาะประกอบจากคำหลายคำ และมีการใช้กฎเพื่อช่วยการตัดสินและจำแนกประเภทของชื่อเฉพาะตามรูปแบบที่เป็นที่รู้จัก แต่กระบวนการทำงานของระบบแบบนี้ ก็ยังแยกประเภทได้อีกว่าจะเริ่มจากการใช้กฎก่อนแล้วจึงใช้ข้อมูลทางสถิติมาประกอบการตัดสินใจ หรือใช้ข้อมูลทางสถิติเพื่อหากลุ่มคำที่อาจเป็นชื่อเฉพาะ (candidate) แล้วจึงใช้กฎในการตัดสินว่ากลุ่มคำที่อาจเป็นชื่อเฉพาะตัวใดที่เป็นชื่อเฉพาะและอยู่ในประเภทใด

สำหรับระบบแบบลูกผสมที่ใช้วิธีกฎก่อนการใช้วิธีทางสถิตินั้น โดยส่วนใหญ่ระบบแบบนี้จะใช้กับภาษาซึ่งมีสัญลักษณ์พิเศษ หรือลักษณะพิเศษที่สามารถนำมาเขียนเป็นกฎเพื่อระบุตำแหน่งของชื่อเฉพาะได้ ดังเช่นภาษาอังกฤษซึ่งมีตัวพิมพ์ใหญ่ หรือ capitalization เป็นจุดที่สามารถนำมาระบุตำแหน่งเริ่มต้นของชื่อเฉพาะได้ ดังเช่นในงานของ Gallippi (1996) ที่ระบุ

ตำแหน่งและจำแนกประเภทของชื่อเฉพาะภาษาอังกฤษ โดยอาศัยกฎที่สร้างขึ้นเองเพื่อจำกัดรูปแบบ (template) ของชื่อเฉพาะแบบต่างๆ แล้วจึงนำค่าทางสถิติเข้ามาร่วมตัดสินว่าส่วนไหนของข้อความที่เป็นชื่อเฉพาะและอยู่ในประเภทใด สำหรับงานวิจัยที่ทำกับภาษาไทยนั้น แม้ว่าจะไม่มีลักษณะพิเศษหรือสัญลักษณ์พิเศษมาช่วยระบุตำแหน่งชื่อเฉพาะและเพื่อช่วยให้การเขียนกฎเป็นไปได้ง่ายขึ้น แต่อย่างไรก็ตาม ก็มีงานของ Kawtrakul et al., 1997 และของ Chanlekha et al., 2002 ที่ใช้ระบบแบบนี้ โดยงานของบุคคลทั้งสองกลุ่ม จะใช้ข้อมูลที่ผ่านการกำกับหมวดของคำแล้ว (มีการ tag part-of-speech) มาเป็นฐานข้อมูล ซึ่งงานของ Chanleka et al., 2002 นั้นจะใช้คลังคำศัพท์ชื่อเฉพาะที่รวบรวมมา เช่น ชื่อประเทศต่างๆ เป็นต้น ร่วมกับการใช้กฎที่เขียนขึ้นเพื่อแยกชื่อเฉพาะออกจากคำอื่นๆ และใช้ค่าทางสถิติในการตัดสินคำอื่นๆ ออกจากชื่อเฉพาะที่อยู่นอกเหนือคลังศัพท์ด้วย

แต่สำหรับภาษาที่ไม่มีสัญลักษณ์พิเศษหรือไม่มีลักษณะพิเศษเพื่อช่วยในการแยกและระบุตำแหน่งของชื่อเฉพาะนั้น การใช้วิธีกฎก่อนการใช้วิธีทางสถิติในการระบุตำแหน่งของชื่อเฉพาะนั้นจะเป็นไปได้ยาก เพราะกฎจะไม่อาจครอบคลุมลักษณะของภาษาในคลังข้อมูลได้ทั้งหมด และจะเกิดความยุ่งยากในการเขียนกฎอีกด้วย อีกทั้งหากกฎที่ใช้ไม่ครอบคลุมพอก็จะทำให้ระบบมองข้ามชื่อเฉพาะที่ควรจะหาออกไป

การใช้วิธีทางสถิติก่อนการใช้กฎจึงเป็นอีกทางเลือกหนึ่งของระบบแบบลูกผสม โดยการใช้วิธีการทางสถิติในการมองหาคำหรือกลุ่มคำที่น่าจะเป็นชื่อเฉพาะได้ จากนั้นจึงใช้กฎมาตัดสินว่าเป็นชื่อเฉพาะหรือไม่และเป็นชื่อประเภทใด ซึ่งเป็นแนวทางที่ผู้วิจัยจะเลือกใช้ในงานวิจัยนี้

2.3 คลังข้อมูลที่ใช้

คลังข้อมูลภาษาที่ผู้วิจัยเลือกจัดเก็บนั้นเป็นข้อมูลภาษาเขียนจากหนังสือพิมพ์ โดยผู้วิจัยเลือกเก็บข่าวจากหนังสือพิมพ์เนื่องจากเป็นภาษาที่เกิดขึ้นจริง สามารถพบเห็นและใช้กันทั่วไปในชีวิตประจำวัน และบทความข่าวมักจะมีรูปแบบการเขียนที่มีแบบแผนทำให้ง่ายต่อการจัดการแบบแผนนั้นๆ และมีชื่อเฉพาะที่สำคัญหลักๆ ในบทความครบทั้ง 3 ประเภท ได้แก่ ชื่อคน ชื่อสถานที่และชื่อองค์กร ตรงตามวัตถุประสงค์ของงานวิจัยนี้ทั่วไป โดยการเก็บรวบรวมจะเก็บบทความข่าวจนกว่าจะได้ชื่อคน ชื่อองค์กร และชื่อสถานที่เกิน 3,000 ชื่อขึ้นไป และจะไม่มีการกำกับหมวดคำให้กับข้อมูลที่เก็บมาได้ แต่จะมีการแยกพยางค์ของคำแต่ละคำออกมา ทั้งนี้เป็นเพราะโปรแกรมการแยกคำนั้นยังไม่สมบูรณ์ เนื่องจากข้อตกลงที่จะใช้ร่วมกันนั้นยังไม่สมบูรณ์ จึงทำให้การตัดพยางค์เหมาะสมที่จะใช้กับงานนี้มากกว่า โดยการแยกพยางค์จะทำด้วยโปรแกรม

แยกพยางค์ที่ดาวน์โหลดจากเว็บไซต์ http://www.arts.chula.ac.th/~ling/wordseg/ (Aroonmanakun, 2002)

จากการเก็บรวบรวมบทความข่าวซึ่งเป็นข้อความภาษาไทยที่เผยแพร่ไว้ในเว็บไซต์ของหนังสือพิมพ์ไทยรัฐ (http://www.thairath.co.th) โดยข้อมูลของหนังสือพิมพ์ไทยรัฐจะเป็นข้อมูลจำนวน 1 ปี 25 วัน คือฉบับวันที่ 3 กรกฎาคม พ.ศ. 2547 ถึงวันที่ 27 กรกฎาคม พ.ศ. 2548 ซึ่งชื่อเฉพาะที่รวบรวมมาได้ทั้งหมดมีจำนวน 11683 ชื่อ โดยแบ่งออกเป็น 3 ประเภท ได้แก่ ชื่อเฉพาะประเภทชื่อคนจำนวน 5,431 ชื่อ ชื่อองค์กรจำนวน 3,222 ชื่อ และชื่อสถานที่จำนวน 3,030 ชื่อ

ในการกำกับประเภทของชื่อเฉพาะนั้น ผู้วิจัยกำกับโดยจัดให้ tag ประเภทต่างๆ 3 ประเภทคร่อมอยู่ด้านหน้าและหลังของชื่อเฉพาะ ดังตัวอย่าง

นาย<person>จักรภพเพ็ญแข</person>

กระทรวง<organization>ศึกษาธิการ</organization>

ประเทศ<location>ออสเตรเลีย</location> เป็นต้น

สำหรับการระบุขอบเขตและจัดประเภทของชื่อเฉพาะนั้นจะจัดแบ่งตามเกณฑ์ความหมายของรูปภาษาโดยอาศัยความรู้ภาษาแม่ (native intuition) ของผู้วิจัย ซึ่งบางรูปภาษาผู้วิจัยสามารถตัดสินได้จากรูปภาษา ดังเช่นประโยค “โดยหลังจากกลับจาก<location>ออสเตรเลีย</location>” โดยผู้วิจัยตัดสินให้ “ออสเตรเลีย” เป็นชื่อเฉพาะประเภทชื่อสถานที่ตามความรู้ภาษาแม่ของผู้วิจัยเอง

แต่ในบางรูปภาษา ข้อความที่ปรากฏอยู่รอบข้างรูปภาษานั้นๆ มีส่วนช่วยให้ผู้วิจัยสามารถพิจารณาตัดสินได้ เช่น จ. เชียงใหม่ ที่ผู้วิจัยมี “จ.” ซึ่งปรากฏอยู่ด้านหน้าของชื่อเฉพาะและถือเป็นหลักฐานภายใน (internal evidence) ที่ช่วยให้ผู้วิจัยพิจารณาตัดสินให้ “เชียงใหม่” เป็นชื่อเฉพาะประเภทสถานที่ ซึ่งหลักฐานภายในสามารถปรากฏอยู่หน้าหรือหลัง หรือทั้งด้านหน้าและหลังของชื่อเฉพาะได้

จากการพิจารณาหลักฐานที่ได้มาจากหลักฐานภายในเหล่านี้ พบว่าชื่อเฉพาะประเภทชื่อคนจะมีหลักฐานชนิดนี้มากที่สุด โดยชื่อเฉพาะจะมีคำนำหน้าชื่อถึง 5408 ชื่อ จากจำนวนชื่อคน 5431 ชื่อ โดยคิดเป็น 99.577% ชื่อองค์กรมีคำนำหน้าชื่อ 3140 ชื่อ จากจำนวนชื่อองค์กร 3222 ชื่อ โดยคิดเป็น 97.455% และชื่อสถานที่มีคำนำหน้าชื่ออยู่ 2158 ชื่อ จากจำนวนชื่อสถานที่ 3030 ชื่อ โดยคิดเป็น 71.221%

ซึ่งคำนำหน้าชื่อที่พบเหล่านี้อาจพบอรรถลักษณ์ร่วมของกลุ่มคำนำหน้าชื่อซึ่งทำให้สามารถแยกประเภทของชื่อเฉพาะออกจากกันได้ เช่น กลุ่มคำนำหน้าชื่อประเภทชื่อคน จะมีอรรถลักษณ์ร่วมกัน คือเป็น <ปัจเจกบุคคล> กลุ่มคำนำหน้าชื่อประเภทชื่อองค์กร จะมีอรรถลักษณ์

ร่วมกัน คือเป็น <กลุ่มบุคคล> <มีการรวมตัวกัน> <ประกอบขึ้นเป็นหน่วยเดียว> และ <มีสภาพเป็นนิติบุคคล> และกลุ่มคำนำหน้าชื่อประเภทชื่อสถานที่ จะมีอรรถลักษณ์ร่วมกัน คือเป็น <พื้นที่หรือบริเวณ> เป็นต้น

โดยในกลุ่มคำนำหน้าชื่อทั้ง 3 ประเภทดังกล่าว แต่ละกลุ่มนั้น อาจมีอรรถลักษณ์ที่ต่างกัน ดังเช่นกรณีของกลุ่มคำนำหน้าชื่อประเภทชื่อคน เช่น "พล.ต.ต." และ "นางสาว" ซึ่งมีอรรถลักษณ์ร่วมกันคือเป็น <ปัจเจกบุคคล> แต่อรรถลักษณ์อีกอย่างจะต่างกัน กล่าวคือ "พล.ต.ต." จะแสดงอรรถลักษณ์ว่าเป็นการ <แสดงฐานะหรือลำดับขั้น> แต่ "นางสาว" จะแสดงอรรถลักษณ์ว่าเป็นการ <บอกอายุ เพศ หรือ/และสถานภาพการสมรส> ซึ่งรายละเอียดจะกล่าวอยู่ในบทที่ 4

จากคลังข้อมูล พบว่าชื่อเฉพาะประเภทชื่อคน อาจพบได้ 2 ลักษณะ คือ ชื่อคนที่มีทั้งชื่อและนามสกุล และชื่อคนที่พบเพียงชื่ออย่างเดียว สำหรับความชัดเจนในการรู้จำและการจำแนกประเภทชื่อเฉพาะนั้น เนื่องจากเราพบหลักฐานภายใน (internal evidence) หรือคำนำหน้าชื่อเป็นจำนวนมาก ทำให้ชื่อเฉพาะประเภทชื่อคนจึงสามารถใช้คำนำหน้าชื่อในการระบุตำแหน่งและจำแนกประเภทของชื่อเฉพาะภาษาไทยได้ แต่สำหรับชื่อเฉพาะประเภทชื่อองค์กรและสถานที่นั้น คำนำหน้าชื่อดังกล่าว อาจทำให้เกิดความกำกวมได้ เมื่อจะใช้คำนำหน้าชื่อจำแนกประเภทระหว่างชื่อองค์กรและชื่อสถานที่ ซึ่งความกำกวมนี้สามารถแก้ไขได้โดยการพิจารณาลักษณะของบริบทข้างเคียงหรือหลักฐานภายนอก (external evidence) ที่พบ โดยชื่อองค์กร เมื่อถูกพบในบริบทหนึ่งอาจถูกจำแนกให้เป็นชื่อสถานที่ได้ ดังเช่น "กระทรวงมหาดไทย" ซึ่งเป็นชื่อองค์กร เพราะคำนำหน้าชื่อ "กระทรวง" มีอรรถลักษณ์คือ เป็น <กลุ่มบุคคล> <มีการรวมตัวกัน> <ประกอบขึ้นเป็นหน่วยเดียว> <มีสภาพเป็นนิติบุคคล> <เพื่องานราชการ>

และหากพบในบริบท "กระทรวงมหาดไทยเป็นหน่วยงานที่ใหญ่ขึ้น"

"กระทรวงมหาดไทย" ก็จะถูกจัดให้อยู่ในกลุ่มชื่อเฉพาะประเภทชื่อองค์กร ตามอรรถลักษณ์ที่กล่าวไปแล้ว

แต่เมื่อพบในบริบท "ไปไหว้สิ่งศักดิ์สิทธิ์ ที่ กระทรวง<location>มหาดไทย</location> ต่อไป"

ดังนี้แล้ว "กระทรวงมหาดไทย" จะถูกจัดให้อยู่ในชื่อเฉพาะประเภทชื่อสถานที่

อย่างไรก็ตาม จากคลังข้อมูลที่สร้างขึ้นพบว่า กลุ่มคำนำหน้าชื่อประเภทชื่อองค์กรที่มีความกำกวมว่าอาจจะจัดให้ชื่อเฉพาะที่ตามหลังคำนำหน้าชื่อประเภทนี้นั้นให้เป็นชื่อเฉพาะประเภทชื่อสถานที่ได้ มีอยู่ 6 คำ ได้แก่ "กรม" , "กระทรวง" , "บริษัท" , "ธนาคาร" , "มูลนิธิ" , "สำนักงาน"

คลังข้อมูลที่สร้างขึ้นมานี้จะถูกใช้เพื่อ 2 วัตถุประสงค์คือ เป็นทั้งคลังข้อมูลใช้ฝึก (training corpus) และคลังข้อมูลใช้ทดสอบ (testing corpus) ซึ่งคลังข้อมูลใช้ฝึกจะถูกใช้เพื่อนำมาทดสอบกับวิธีทางสถิติแล้ว ยังใช้เพื่อพิจารณาลักษณะของรูปภาษาที่พบซึ่งจะทำให้เขียนกฎที่จะนำมาใช้ได้ อีกทั้งยังใช้ปรับปรุงระบบที่กำลังพัฒนาเพื่อให้ระบบที่ได้มีประสิทธิภาพสูงที่สุด สำหรับในการทดสอบ คลังข้อมูลจะถูกใช้เพื่อทำการทดสอบระบบที่พัฒนาขึ้นว่ามีประสิทธิภาพมากน้อยเพียงใด และสามารถนำไปใช้กับข้อมูลได้จริงหรือไม่

2.4 ระบบการรู้จำชื่อเฉพาะภาษาไทย

ในส่วนนี้ ผู้วิจัยจะนำเสนอระบบการรู้จำชื่อเฉพาะภาษาไทยที่ใช้ในงานวิจัยนี้ โดยผู้วิจัยเลือกใช้ระบบแบบผสมซึ่งจะเริ่มจากการใช้วิธีการทางสถิติเพื่อหารายการของกลุ่มพยางค์ที่น่าจะเป็นชื่อเฉพาะ ก่อนที่จะใช้กฎเพื่อตัดสินรายการชื่อที่ได้มาว่าเป็นชื่อเฉพาะหรือไม่และจัดอยู่ในประเภทใด

ในขั้นตอนแรกซึ่งเป็นวิธีการทางสถิติ คลังข้อมูลที่ใช้จะถูกตัดแบ่งออกเป็นพยางค์ ทั้งนี้เป็นเพราะโปรแกรมการแยกคำนั้นยังไม่สมบูรณ์ เพราะแม้แต่การแยกคำด้วยคน แต่ละคนก็อาจแยกคำต่างกันได้ ในขณะที่การแยกพยางค์จะเป็นที่ชัดเจนมากกว่าการแยกคำ (Aroonmanakun, 2002) จึงทำให้การตัดพยางค์เหมาะสมที่จะใช้กับงานนี้มากกว่า เมื่อได้คลังข้อมูลที่จะใช้แล้ว จะมีการปรับวิธีการทางสถิติที่ใช้ 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 เนื่องจากวิธีการทางสถิติ 4 วิธีแรกนั้นใช้วัดความสัมพันธ์ระหว่างสองหน่วยซึ่งในที่นี้คือพยางค์แต่ชื่อเฉพาะอาจจะมีความยาวมากกว่า 2 พยางค์ ทำให้ต้องมีการปรับวิธีการคำนวณให้สามารถคำนวณความสัมพันธ์ของพยางค์ที่มีมากกว่า 2 พยางค์ให้ได้ โดยให้มีการมองพยางค์หลายๆ พยางค์เป็นเหมือน pseudo bigram ซึ่งจะช่วยให้สามารถรู้จำชื่อเฉพาะที่มีความยาวมากกว่า 2 พยางค์ได้

นอกจากการใช้วิธีการทางสถิติต่างๆ วัดความสัมพันธ์ของกลุ่มพยางค์แล้ว ในการตัดสินว่ากลุ่มพยางค์ใดน่าจะเป็นชื่อเฉพาะได้ ก็ต้องอาศัย Localmax algorithm เพื่อคัดเลือกรายการของกลุ่มพยางค์ที่น่าจะเป็นชื่อเฉพาะออกมา ซึ่งรายละเอียดของการปรับขยายการคำนวณแบบ pseudo bigram และการใช้ Localmax algorithm จะกล่าวโดยละเอียดในบทที่ 3 ต่อไป ซึ่งผลที่ได้จากการใช้วิธีการทางสถิติคำนวณความสัมพันธ์ของกลุ่มพยางค์ และการใช้ Localmax algorithm ในการตัดสิน จะเป็นรายการของกลุ่มพยางค์ที่คาดว่าน่าจะเป็นชื่อเฉพาะได้

จากนั้น ก็จะนำผลที่ได้มาทำการประเมิน โดยการประเมินผลวิธีทางสถิติจะเป็นการประเมินเพื่อเลือกวิธีทางสถิติที่ดีที่สุดที่จะใช้ร่วมกับ Localmax algorithm ในการคัดเลือกกลุ่มพยางค์ที่น่าจะเป็นชื่อเฉพาะออกมาและทำให้สามารถดำเนินการวิจัยขั้นต่อไปได้ ซึ่งการประเมินจะใช้อัตราการรู้จำที่วัดด้วยค่า F และเวลาที่ใช้ในการรู้จำ

ซึ่ง การที่ใช้ค่า F เป็นอัตราในการรู้จำที่ใช้ในงานวิจัยนี้ เพราะในงานนี้ให้ความสำคัญทั้งค่าความแม่นยำและค่าความครบถ้วน ซึ่งค่า F เป็นค่าที่เฉลี่ยโดยจะให้ความสำคัญกับค่าความแม่นยำและความครบถ้วนเท่าๆ กันจึงเหมาะที่จะใช้เป็นอัตราในการรู้จำมากที่สุด ซึ่งค่า F สามารถคำนวณได้จากสูตร

ค่า F = (2 *P *R) / (P+R)

โดย ค่า precision หรือค่าความแม่นยำ จะคำนวณได้จากสูตร

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

ค่า recall หรือค่าความครบถ้วน จะคำนวณได้จากสูตร

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

นอกจากอัตราในการรู้จำแล้ว ยังใช้เวลาที่ใช้ในการรู้จำในการประเมินประสิทธิภาพด้วย ซึ่งเวลาที่ใช้ในการรู้จำ คือ เวลาที่ใช้ในการประมวลผลจนได้ชื่อเฉพาะที่เลือกมา (candidate) โดยมีหน่วยเป็น จำนวนชื่อเฉพาะต่อ 1 นาที

หลังจากที่ได้รายการของกลุ่มพยางค์ที่น่าจะเป็นชื่อเฉพาะที่มาจากวิธีทางสถิติที่ให้ผลอัตราการรู้จำที่ดีที่สุด และจะนำมาสู่ขั้นตอนถัดมา คือการใช้วิธีกฎ

กฎที่เขียนขึ้นจะใช้เพื่อตัดสินว่าชื่อที่เลือกมานั้น ตัวไหนจัดเป็นชื่อเฉพาะจริงและเป็นชื่อเฉพาะประเภทใดใน 3 ประเภทคือ ชื่อคน ชื่อสถานที่และชื่อองค์กร ซึ่งกฎที่ใช้ในการจำแนกประเภทของชื่อเฉพาะนั้นจะสร้างมาจากการใช้หลักฐานภายใน (internal evidence) ซึ่งได้แก่ รายการของคำนำหน้าชื่อ เช่น "นาย" , "นาง" , "นางสาว" และหลักฐานภายนอก (external evidence) ซึ่งได้มาจากการพิจารณาลักษณะของบริบทข้างเคียงของชื่อเฉพาะที่ได้มาจากคลังข้อมูล โดยจะเป็นรายการของคำที่มักจะนำหน้าและตามหลังชื่อเฉพาะแต่ละประเภททั้ง 3 ประเภท เช่น คำว่า "ปราศรัย" จะตามหลังชื่อเฉพาะประเภทชื่อคนมากกว่าที่จะพบว่าตามหลังชื่อเฉพาะประเภทอื่น เป็นต้น ซึ่งรายละเอียดจะกล่าวในบทที่ 4 ต่อไป

จากผลที่ได้จากวิธีกฎจะถูกนำมาประเมินประสิทธิภาพในการรู้จำและจำแนกประเภทชื่อเฉพาะด้วยอัตราการรู้จำ ซึ่งการคำนวณหาค่า F ก็จะเป็นไปตามที่กล่าวมาแล้ว