# การรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทย

นางสาว สุฤดี ฉัตรไตรมงคล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต

สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2548

ISBN 974-53-2979-7

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

# NAMED ENTITY RECOGNITION AND CLASSIFICATION IN THAI

Miss Surudee Chattrimongkol

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Arts Program in Linguistics

Department of Linguistics

Faculty of Arts

Chulalongkorn University

Academic Year 2005

ISBN 974-53-2979-7

| | |
|---|---|
| หัวข้อวิทยานิพนธ์ | การรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทย |
| โดย | นางสาว สุฤดี ฉัตรไตรมงคล |
| สาขาวิชา | ภาษาศาสตร์ |
| อาจารย์ที่ปรึกษา | ผู้ช่วยศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล |

---

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

…….. คณบดีคณะอักษรศาสตร์
(ศาสตราจารย์ ดร.ธีระพันธ์ เหลืองทองคำ)

คณะกรรมการสอบวิทยานิพนธ์

…… ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุดาพร ลักษณียนาวิน)

…….. อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล)

…… กรรมการ
(รองศาสตราจารย์ ดร.กิ่งกาญจน์ เทพกาญจนา)

สุฤดี ฉัตรไตรมงคล : การรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทย. (NAMED ENTITY RECOGNITION AND CLASSIFICATION IN THAI) อาจารย์ที่ปรึกษา : ผศ. ดร. วิโรจน์ อรุณมานะกุล, 75 หน้า. ISBN 974-53-2979-7.

งานวิจัยนี้มีจุดประสงค์เพื่อพัฒนาระบบการรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทยโดยใช้แนวทางแบบลูกผสม (hybrid approach) โดยแนวทางดังกล่าวจะแบ่งออกเป็นสองส่วนคือส่วนที่เป็นระบบทางสถิติและส่วนที่เป็นระบบกฎ

สำหรับส่วนของระบบทางสถิตินั้นจะใช้วิธีทางสถิติร่วมกับโลคอลแมกซ์อัลกอริทึมเพื่อคัดเลือกกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะออกมา ซึ่งวิธีการทางสถิติที่ใช้ในการวัดความสัมพันธ์ระหว่างพยางค์ในที่นี้มี 5 วิธี ได้แก่ การใช้ค่ามิวชวลอินฟอร์เมชัน ค่าไคกำลังสอง ค่าคิวบิกแอสโซซิเอชันเรโช ค่าล็อกไลค์ลิฮูด และค่ามิวชวลเอ็กซ์เป็กเตชันนั้น ผลพบว่าวิธีที่ใช้ค่ามิวชวลเอ็กซ์เป็กเตชันร่วมกับการใช้โลคอลแมกซ์ อัลกอริทึม ในการรู้จำชื่อเฉพาะนั้นให้อัตราการรู้จำได้ผลดีที่สุด แต่วิธีดังกล่าวก็มีข้อเสียตรงที่ใช้เวลาในการประมวลผลที่นานเกินไป ทำให้ในงานวิจัยนี้จะใช้วิธีทางสถิติที่ให้ผลอัตราการรู้จำดีรองลงมา นั่นคือ การใช้ค่ามิวชวลอินฟอร์เมชันร่วมกับการใช้โลคอลแมกซ์ อัลกอริทึม จากนั้นเมื่อได้ชื่อเฉพาะที่เลือกมาด้วยวิธีการทางสถิติแล้ว จะเข้าสู่ส่วนที่เป็นระบบกฎ ซึ่งระบบถูกเขียนขึ้นโดยอิงกับหลักฐานที่ได้จากบริบทภายใน เช่น คำนำหน้าชื่อ และใช้บริบทข้างเคียง เช่น คำปรากฏร่วม เพื่อช่วยในการรู้จำและจำแนกประเภทของชื่อเฉพาะ และจากการทดสอบพบว่าระบบกฎที่สร้างขึ้นสามารถจำแนกประเภทของชื่อเฉพาะโดยให้อัตราการรู้จำ (ค่า F) สำหรับชื่อเฉพาะประเภทชื่อคน 69.15% ชื่อองค์กร 62.95% และชื่อสถานที่ 38.87% ตามลำดับ โดยมีค่าความแม่นยำและค่าความครบถ้วนสำหรับชื่อเฉพาะประเภทชื่อคน 54.00% และ 96.12% ชื่อองค์กร 47.60% และ 92.93% ชื่อสถานที่ 31.67% และ 50.32% ตามลำดับ

| | | |
|---|---|---|
| ภาควิชา | ภาษาศาสตร์ | ลายมือชื่อนิสิต ..... |
| สาขาวิชา | ภาษาศาสตร์ | ลายมือชื่ออาจารย์ที่ปรึกษา..... |
| ปีการศึกษา | 2548 | |

##4580247422 : MAJOR LINGUISTICS
KEY WORD: NAMED ENTITY / RECOGNITION / CLASSIFICATION / MUTUAL EXPECTATION / LOCALMAX ALGORITHM

SURUDEE CHATTRIMONGKOL : NAMED ENTITY RECOGNITION AND CLASSIFICATION IN THAI. THESIS ADVISOR : ASST. PROF. WIROTE AROONMANAKUN, Ph.D., 75 pp. ISBN 974-53-2979-7.

This study aims to develop a Thai named entity recognition and classification system using a hybrid approach. The system is composed of two parts, which are statistical part and rule part.

Statistical part is used for extracting named entity candidates. Localmaxs algorithm and the statistical method are used for measuring associations between syllables. Five statistical methods namely Mutual Expectation, Mutual Information, Chi-square, Cubic Association ratio and Loglikelihood are tested in this part. Mutual Expectation combined with Localmaxs algorithm yields the best result, but this method uses much more times than other methods. Therefore, Mutual Information, which is the second best statistical method, combined with Localmaxs algorithm is used for extracting a chunk of syllables as a candidate of named entity. On the second part, named entity candidates will be recognized and classified by linguistic rules which are manually crafted. Internal evidence, i.e. title names, and external evidence, i.e. collocate words, are used in these rules. The system can recognize and classify named entities with the recognition rate (F-measure) , precision and recall rates at 69.15% , 54.00% and 96.12% for person names, 62.95% , 47.60% and 92.93% for organization names, 38.87% , 31.67% and 50.32% for location names.

| | | |
|---|---|---|
| Department | LINGUISTICS | Student's signature…… |
| Field of study | LINGUISTICS | Advisor's signature……..... |
| Academic year | 2005 | |

# กิตติกรรมประกาศ

ผู้วิจัยต้องขอขอบพระคุณ ผศ.ดร.วิโรจน์ อรุณมานะกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์เป็นอย่างสูง ที่ได้ให้ความช่วยเหลือและคำปรึกษาเกี่ยวกับวิทยานิพนธ์ฉบับนี้มาโดยตลอด หากไม่ได้คำแนะนำจากอาจารย์แล้ว วิทยานิพนธ์ฉบับนี้คงจะสำเร็จสมบูรณ์ไปไม่ได้เลย และผู้วิจัยขอขอบพระคุณ ผศ.ดร.สุดาพร ลักษณียนาวิน และรศ.ดร. กิ่งกาญจน์ เทพกาญจนา กรรมการสอบวิทยานิพนธ์ที่ได้ให้คำปรึกษาและเสียสละเวลาเพื่อตรวจสอบวิทยานิพนธ์ฉบับนี้

สุดท้ายนี้ ผู้วิจัยขอขอบคุณคณาจารย์ภาควิชาภาษาศาสตร์ทุกท่านที่ได้ประสิทธิ์ประสาทความรู้ด้านภาษาศาสตร์ให้แก่ผู้วิจัย รวมทั้งขอขอบคุณเจ้าหน้าที่และเพื่อนพี่น้องภาควิชาภาษาศาสตร์ทุกคนที่ช่วยเหลือและให้กำลังใจผู้วิจัย และขอขอบพระคุณ คุณพ่อสงวน คุณแม่สุภาภรณ์ ฉัตรไตรมงคลสำหรับโอกาสและกำลังใจอันมากเหลือที่ได้รับ รวมทั้งขอขอบคุณพี่ชายคุณมงคล ฉัตรไตรมงคลที่คอยสนับสนุนและให้กำลังใจมาโดยตลอด

# สารบัญ


| | หน้า |
|---|---|
| บทคัดย่อภาษาไทย | ง |
| บทคัดย่อภาษาอังกฤษ | จ |
| กิตติกรรมประกาศ | ฉ |
| สารบัญ | ช |
| สารบัญตาราง | ฌ |
| บทที่ | |
| 1 บทนำ | 1 |
| 1.1 ความเป็นมาและความสำคัญของปัญหา | 1 |
| 1.2 วัตถุประสงค์ของการวิจัย | 2 |
| 1.3 สมมติฐาน | 3 |
| 1.4 ขอบเขตของการวิจัย | 3 |
| 1.5 ประโยชน์ที่คาดว่าจะได้รับ | 4 |
| 1.6 วิธีดำเนินการวิจัย | 4 |
| 1.7 โครงสร้างวิทยานิพนธ์ | 5 |
| 1.8 ศัพท์เฉพาะที่ใช้ในงานวิจัย | 5 |
| 2 ระบบการรู้จำชื่อเฉพาะ | 7 |
| 2.1 ความหมาย รูปแบบต่างๆ และประเภทของชื่อเฉพาะ | 7 |
| 2.2 ระบบที่ใช้ในการรู้จำและจำแนกประเภทของชื่อเฉพาะ | 8 |
| 2.2.1 ระบบที่ใช้วิธีกฎ | 10 |
| 2.2.2 ระบบที่ใช้วิธีการทางสถิติเป็นหลัก | 12 |
| 2.2.2.1 วิธีทางสถิติแบบที่มองความสัมพันธ์การปรากฏร่วมกันระหว่างหน่วยย่อยๆ ภายใน | 12 |
| 2.2.2.1.1 การหาค่า Mutual Information | 12 |
| 2.2.2.1.2 การหาค่า Cubic association ratio ($MI^3$) | 13 |

| บทที่ | | หน้า |
|---|---|---|
| | 2.2.2.1.3 การหาค่า Pearson's Chi-square ..... | 13 |
| | 2.2.2.1.4 การหาค่า Dunning's log likelihood ..... | 13 |
| | 2.2.2.2 วิธีการทางสถิติแบบที่มองการสูญเสียหน่วยย่อยใดๆ ภายในว่ายอมรับได้มากหรือไม่ ..... | 15 |
| | 2.2.2.2.1 การหาค่า Mutual Expectation ..... | 15 |
| | 2.2.3 ระบบแบบลูกผสม (hybrid) ..... | 17 |
| | 2.3 คลังข้อมูลที่ใช้ ..... | 18 |
| | 2.4 ระบบการรู้จำชื่อเฉพาะภาษาไทย ..... | 21 |
| 3 | การหาขอบเขตของชื่อเฉพาะ ..... | 23 |
| | 3.1 วิธีทางสถิติที่ประยุกต์ใช้ ..... | 23 |
| | 3.2 การกำหนดขอบเขตของชื่อเฉพาะ ..... | 26 |
| | 3.3 เปรียบเทียบวิธีทางสถิติที่ใช้ในการหาขอบเขตชื่อเฉพาะ ..... | 28 |
| | 3.4 อภิปรายผล ..... | 32 |
| 4 | การรู้จำชื่อเฉพาะ ..... | 33 |
| | 4.1 การจำแนกประเภทชื่อเฉพาะ ..... | 33 |
| | 4.1.1 หลักฐานภายใน (internal evidence) ..... | 33 |
| | 4.1.2 หลักฐานภายนอก (external evidence) ..... | 39 |
| | 4.2 การสร้างกฎการจำแนกประเภทชื่อเฉพาะ ..... | 39 |
| | 4.3 ผลการตัดสินและจำแนกประเภทชื่อเฉพาะ ..... | 45 |
| | 4.4 อภิปรายผล ..... | 47 |
| 5 | สรุปผลการวิจัยและข้อเสนอแนะ..... | 50 |
| | 5.1 สรุปผลการวิจัย..... | 50 |
| | 5.2 ข้อเสนอแนะ..... | 51 |
| รายการอ้างอิง..... | | 54 |
| ภาคผนวก ..... | | 58 |
| ประวัติผู้เขียนวิทยานิพนธ์ ..... | | 75 |

# สารบัญตาราง


| ตารางที่ | | หน้า |
|---|---|---|
| 1 | (n-1) grams and missing words ..... | 16 |
| 2 | ผลการทดสอบค่าความแม่นยำ ค่าความครบถ้วนและค่า F ที่ได้จากวิธีทางสถิติ 5 วิธี ..... | 30 |
| 3 | ผลการทดสอบเวลาที่ใช้ในการรู้จำที่ได้จากวิธีทางสถิติ 5 วิธี ..... | 31 |
| 4 | การแบ่งประเภทกลุ่มคำนำหน้าชื่อ ..... | 35 |
| 5 | ผลการทดสอบค่าความแม่นยำ ค่าความครบถ้วน และค่า F เมื่อใช้โปรแกรมกฎ ..... | 45 |
| 6 | ผลการจำแนกประเภทของโปรแกรมกฎที่เขียนขึ้น ..... | 46 |

# 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

# 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 รูปแบบต่าง ๆ 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 ได้แก่ภาษาอังกฤษ 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 ด้วยทำให้ใช้ได้กับภาษาที่กำลังในคลังข้อมูลที่กำลังพิจารณาอยู่เท่านั้น

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 ซึ่งได้แก่การใช้ค่า 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 ($P(w_1w_2)$) 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 **วิธีการทางสถิติแบบที่มองการสูญเสียหน่วยย่อยใด ๆ 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 $[w_1p_{12}w_2p_{13}w_3\ldots p_{1i}w_i\ldots p_{1n}w_n]$ 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

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

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

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

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

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

ค่า Fair Point of Expectation นี้นำมาใช้ในการหาค่า Normalised Expectation ดังแสดงในสมการ d

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

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

ค่า $FPE([w_1 p_{12} w_2 \ldots p_{1i} w_i \ldots p_{1n} w_n])$ ก็จะหาได้จากสมการ c

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

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

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

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

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” ซึ่งเป็นชื่อองค์กรเพราะคำนำหน้าชื่อ “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) ซึ่งได้แก่รายการของคำนำหน้าชื่อ เช่น “นาย” , “นาง” , “นางสาว” และหลักฐานภายนอก (external evidence) ซึ่งได้มาจากการพิจารณาลักษณะของบริบทข้างเคียงของชื่อเฉพาะที่ได้มาจากคลังข้อมูล โดยจะเป็นรายการของคำที่มักจะนำหน้าและตามหลังชื่อเฉพาะแต่ละประเภททั้ง 3 ประเภท เช่น คำว่า “ปราศรัย” จะตามหลังชื่อเฉพาะประเภทชื่อคนมากกว่าที่จะพบว่าตามหลังชื่อเฉพาะประเภทอื่น เป็นต้น ซึ่งรายละเอียดจะกล่าวในบทที่ 4 ต่อไป

จากผลที่ได้จากวิธีกฎจะถูกนำมาประเมินประสิทธิภาพในการรู้จำและจำแนกประเภทชื่อเฉพาะด้วยอัตราการรู้จำ ซึ่งการคำนวณหาค่า F ก็จะเป็นไปตามที่กล่าวมาแล้ว

# บทที่ 3

# การหาขอบเขตของชื่อเฉพาะ

ในบทนี้จะกล่าวถึงการทำงานส่วนแรกของระบบ คือการใช้วิธีการทางสถิติเพื่อหาขอบเขตของชื่อเฉพาะ หรือคือการหารายการกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ และเนื่องจากวิธีการทางสถิติที่ใช้ เป็นการคำนวณความสัมพันธ์ระหว่างสองหน่วยหรือสองพยางค์ ในตอนแรกของบทนี้จึงจะกล่าวถึงการปรับวิธีการทางสถิติเพื่อใช้คำนวณความสัมพันธ์ระหว่างหลายพยางค์ได้ ซึ่งจะกล่าวถึงวิธีทางสถิติที่นำมาปรับเพื่อใช้เลือกกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ (candidate) และนำเสนอผลวิธีการทางสถิติที่เหมาะสมที่สุดในการดึงกลุ่มพยางค์ที่น่าจะเป็นชื่อเฉพาะออกมา และในสองส่วนสุดท้ายจะกล่าวถึงผลและการอภิปรายผลของวิธีทางสถิติที่ใช้ในงานวิจัยนี้

## 3.1 วิธีทางสถิติที่ประยุกต์ใช้

จากวิธีการทางสถิติที่กล่าวไปในบทที่แล้ว จะพบว่าวิธีการใช้ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio ($MI^3$) และค่า Dunning's Log Likelihood นั้น ปกติจะใช้คำนวณหาค่าความสัมพันธ์ของ 2 คำ ซึ่งชื่อเฉพาะหลายชื่อมักมีความยาวมากกว่า 2 คำ อีกทั้งชื่อสถานที่หรือชื่อองค์กรก็มักมีความยาวมากกว่า 2 คำขึ้นไป ทำให้วิธีการทางสถิติเหล่านี้มีข้อจำกัด ซึ่งอาจแก้ไขโดยใช้วิธีการขยายหน้าต่าง (Lopes และคณะ ,1999) กล่าวคือ เมื่อหาค่าความสัมพันธ์ระหว่างคำ 2 คำได้แล้ว ก็จะมอง 2 คำนั้นให้เป็นหน่วยเดียวกัน แล้วมองคำถัดไปให้เป็นอีกหน่วยหนึ่ง แล้วหาความสัมพันธ์ระหว่างหน่วยนั้นๆ กับคำถัดไป จากนั้น หน้าต่างก็จะขยายต่อไปเรื่อยๆ หลักการนี้เป็นการมองเอ็นแกรมใดๆ ในรูปของไบแกรมเทียม (pseudo-bigram) เช่น มองเอ็นแกรม ($w_1$, $w_2$, …$w_m$) เป็นไบแกรมของ $w_1$, $w_2$, … $w_i$ กับ $w_{i+1}$ … $w_n$ เป็นต้น ดังนั้น เมื่อต้องการคำนวณค่าอย่างเช่นค่า Mutual Information (MI) ซึ่งปกติจะเท่ากับ log P(x,y)/P(x)P(y) ก็จะคำนวณเป็น log P( $w_1$, $w_2$,… $w_n$)/P( $w_1$, $w_2$,… $w_i$)P( $w_{i+1}$,… $w_n$) และเนื่องจากเราสามารถแบ่งครึ่งเอ็นแกรมใดๆ ที่ตำแหน่งระหว่าง $w_1$- $w_2$, $w_2$- $w_3$, … หรือ $w_{n-1}$- $w_n$ ก็ได้ (Lopes และคณะ ,1999) จึงต้องหาค่าเฉลี่ยของไบแกรมเทียมทั้ง n-1 แบบนี้

ในงานวิจัยนี้ ผู้วิจัยก็ใช้หลักการเดียวกันนี้ เพียงแต่เปลี่ยนจากการมองความสัมพันธ์ระหว่างคำเป็นความสัมพันธ์ระหว่างพยางค์แทน ซึ่งในงานวิจัยนี้ การปรับขยายหน้าต่างจะขยายถึง 8-gram เพื่อให้ครอบคลุมชื่อเฉพาะที่มีความยาวมากๆ

ซึ่งจากการปรับสมการจะทำให้ได้สมการของการหาค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio ($MI^3$) และค่า Dunning's Log Likelihood ออกมาเป็นดังนี้

- การขยายการคำนวณแบบ pseudo bigram เพื่อหาค่า Mutual Information

จากสมการ $I(w_1, w_2) = \log \frac{p(w_1 w_2)}{p(w_1)\,p(w_2)}$

เมื่อแก้สมการ ทำการขยายหน้าต่างจะได้

$$I(w_1 \ldots w_n) = \log \frac{p(w_1 \ldots w_n)}{Avp}$$

โดยที่ $Avp = \frac{1}{(n-1)} * \sum_{i=1}^{i=n-1} p(w_1 \ldots w_i) * p(w_{i+1} \ldots w_n)$

- การขยายการคำนวณแบบ pseudo bigram เพื่อหาค่า Pearson's Chi-square

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

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 $E_{i,j}$ คือค่าความถี่ของแต่ละเซลล์ในตารางเมื่อคำสองคำเกิดร่วมกันโดยบังเอิญ ซึ่งค่า $E_{i,j}$ ของแต่ละเซลล์ในตารางจะหาได้จาก ผลรวมของแต่ละแถวคูณกับผลรวมของแต่ละสดมภ์หารด้วยผลรวมของค่าความถี่ทั้งหมด

เมื่อทำการแทนค่า $O_{i,j}$ และ $E_{i,j}$ แล้ว สมการจะได้ดังนี้

$$\chi^2 = \frac{[f(w_1, w_2) * N - (f(w_1) * f(w_2))]^2}{f(w_1) * f(w_2) * (N - f(w_1)) * (N - f(w_2))}$$

โดย $f(w_1)$ , $f(w_2)$ = ค่าความถี่ของคำ ($w_1$) และ ($w_2$) ตามลำดับ

$N$ = จำนวนคำทั้งหมดในคลังข้อมูล

เมื่อคำนวณด้วยการขยายการคำนวณแบบ pseudo bigram จะได้

$$\chi^2(w_1 \ldots w_n) = \frac{[f(w_1 \ldots w_n) * N - Avp]^2}{Avp * (N - Avx) * (N - Avy)}$$

โดยที่ $Avx = \frac{1}{(n-1)} * \sum_{i=1}^{i=n-1} f(w_1 ... w_i)$

$Avy = \frac{1}{(n-1)} * \sum_{i=2}^{i=n} f(w_i ... w_n)$

$Avp = \frac{1}{(n-1)} * \sum_{i=1}^{i=n-1} p(w_1 ... w_i) * p(w_{i+1} ... w_n)$

- การขยายการคำนวณแบบ pseudo bigram เพื่อหาค่า Cubic association ratio ($MI^3$)

จากสมการ

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

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 ($w_1 - \sim w_2$)

เมื่อแก้สมการจะได้

$$MI^3 = log_2 \frac{(f(w_1, w_2))^3 * N}{[f(w_1, w_2) + (f(w_2) - f(w_1, w_2))] * [f(w_1, w_2) + (f(w_1) - f(w_1, w_2))]} \text{ ------z}$$

จากสมการ z จะได้เป็น

$$MI^3 = log_2 \frac{(f(w_1, w_2))^3 * N}{f(w_2) * f(w_1)}$$

ดังนั้น เมื่อทำการขยายการคำนวณแบบ pseudo bigram จะได้

$$MI^3 (w_1 \ldots w_n) = log_2 \frac{(f(w_1 ... w_n))^3 * N}{Avy * Avx}$$

โดยที่ $Avx = \frac{1}{(n-1)} * \sum_{i=1}^{i=n-1} f(w_1 ... w_i)$

$Avy = \frac{1}{(n-1)} * \sum_{i=2}^{i=n} f(w_i ... w_n)$

- การขยายการคำนวณแบบ pseudo bigram เพื่อหาค่า Dunning's log likelihood

จากสมการ

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

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

เมื่อทำการขยายการคำนวณแบบ pseudo bigram จะได้

$Loglike\ (w_1 \ldots w_n) = 2 * (\log l(pf1,kf1,nf1) + \log l(pf2,kf2,nf2) - \log l(pf,kf1,nf1) - \log l(pf,kf2,nf2))$

โดยที่ $kf1 = f(w_1 \ldots w_n)$

$kf2 = Avx - kf1$

$pf = (kf1 + kf2)/N = Avx$

$nf1 = Avy$

$nf2 = N - nf1$

$pf1 = kf1/nf1$

$pf2 = kf2/nf2$

$$Avx = \frac{1}{(n-1)} * \sum_{i=1}^{i=n-1} f(w_1 \ldots w_i)$$

$$Avy = \frac{1}{(n-1)} * \sum_{i=2}^{i=n} f(w_i \ldots w_n)$$

### 3.2 การกำหนดขอบเขตของชื่อเฉพาะ

ข้อมูลใดใดที่ส่งเข้ามาจะถูกแยกพยางค์และมองเป็น n-gram ขนาดต่างๆ พร้อมทั้งคำนวณค่าทางสถิติเพื่อบอกความสัมพันธ์ภายใน n-gram นั้น เมื่อได้ค่าความสัมพันธ์ของ n-gram จากแต่ละวิธีที่กล่าวไปข้างต้นได้แก่ ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio ($MI^3$) และค่า Dunning's Log Likelihood (ที่ผ่านการขยายหน้าต่างเพื่อให้หาความสัมพันธ์ระหว่าง n-grams ได้กว้างขึ้นแล้ว) รวมทั้งค่า Mutual Expectation (ตามที่กล่าวถึงแล้วในบทที่ 2) ก็จะมาทำการเลือก n-gram หรือกลุ่มพยางค์ที่อาจจะเป็นชื่อเฉพาะ (candidate) ซึ่งโดยทั่วไปอาจใช้ระดับหรือ threshold ในการตัดสินว่าขอบเขตของตำแหน่งใดที่น่าจะเป็นจุดเริ่มต้นและจุดสิ้นสุดของชื่อเฉพาะ แล้วทำการคัดให้เป็น กลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ (candidate) แต่วิธีการนี้จะมีข้อเสียตรงที่ ระดับที่ใช้คัดกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ (candidate) นั้นจะได้จากคลังข้อมูลใช้ฝึก (training corpus) ซึ่งทำให้ระดับจะขึ้นอยู่กับขนาดของ

คลังข้อมูลใช้ฝึก (training corpus) ด้วย ดังนั้น จึงมีอีกแนวคิดหนึ่งที่ช่วยแก้ปัญหานี้ นั่นก็คือ การใช้ LocalMaxs algorithm ในการเลือก n-gram ที่มีแนวโน้มที่จะเป็นหน่วยเดียวกัน

การใช้ LocalMaxs algorithm เป็นวิธีการที่ Dias และคณะ (2000) นำมาใช้เพื่อระบุตำแหน่งเริ่มต้นและสิ้นสุดของศัพท์เฉพาะทาง (term) ซึ่งมักจะประกอบด้วยคำย่อยๆ หลายคำซึ่ง Localmax algorithm จะเลือกตัดสินว่ากลุ่มคำนั้นเป็นศัพท์เฉพาะทางหรือไม่ โดยพิจารณาจากคุณสมบัติ 2 ข้อ

1. ค่าที่ได้จากการวัดค่าความสัมพันธ์ยิ่งมีค่ามากจะบอกถึงความสัมพันธ์ระหว่างกลุ่มคำที่มีมากตามไปด้วย
2. ศัพท์เฉพาะทางมักเป็นกลุ่มคำที่ปรากฏร่วมกันและมีตำแหน่งที่แน่นอน

ดังนั้นเราจะอนุมานเอ็นแกรมของคำนี้ให้เป็นศัพท์เดียวกันก็ต่อเมื่อค่าที่ได้จากการวัดความสัมพันธ์สูงกว่าหรือเท่ากันกับค่าที่ได้จากการวัดความสัมพันธ์ของส่วนย่อย (sub-group) ทั้งหมดภายใน (n-1) gram และก็สูงกว่าค่าที่ได้จากการวัดความสัมพันธ์ของ (n+1) gram ด้วย

เมื่อกำหนดให้ *assoc* เป็นค่าที่ได้จากการวัดความสัมพันธ์

*W* เป็น n-gram

$\Omega_{n-1}$ เป็นเซตของ (n-1) gram ทุกตัวที่อยู่ใน *W*

$\Omega_{n+1}$ เป็นเซตของ (n+1) gram ทุกตัวที่อยู่ใน *W*

*sizeof* จะบอกจำนวนของคำใน n-gram

ซึ่ง Localmax algorithm จะนิยามได้เป็น ดังนี้

$\forall x \in \Omega_{n-1}, \forall y \in \Omega_{n+1},$

*W* จะเป็น term เดียวกันก็ต่อเมื่อ

$$(sizeof(W) = 2 \wedge assoc(W) > assoc(y)$$
$$\vee$$
$$(sizeof(W) \neq 2 \wedge assoc(W) \geq assoc(x) \wedge assoc(W) > assoc(y))$$

ตัวอย่างการใช้ Localmax algorithm จะเลือกศัพท์โดยพิจารณาบริบทที่อยู่ข้างเคียงแล้วเลือกกลุ่มคำที่มีความสัมพันธ์ระหว่างกันสูงๆ เช่น เมื่อพิจารณา “Operating System” และ “Operating System Windows” จะพบว่า “Operating System” มีความสัมพันธ์ระหว่างกันสูงกว่าค่าความสัมพันธ์ระหว่าง “Operating System Windows” ซึ่งเป็นเพราะว่ามีคำอื่นที่มีความสัมพันธ์กับ “Operating System” มากกว่า “Windows”

ดังนั้น Localmax algorithm ก็จะเลือกให้ “Operating System” ว่ามีโอกาสที่จะเป็นศัพท์ได้มากกว่า“Operating System Windows”

แต่ในวิทยานิพนธ์นี้ ผู้วิจัยได้นำ Localmax algorithm มาประยุกต์ใช้ในการหาขอบเขตของชื่อเฉพาะ โดยมองหากลุ่มพยางค์ที่มีค่าความสัมพันธ์ภายในสูงกว่าค่าความสัมพันธ์ภายในของกลุ่มพยางค์ที่ถูกลดจำนวนลงหรือถูกเพิ่มจำนวนขึ้น โดยคำนวณค่าความสัมพันธ์ภายในด้วยวิธีการสถิติแบบต่างๆ อันได้แก่ ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio ($MI^3$) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation ในที่นี้จะเรียกกลุ่มพยางค์ที่ถูกเลือกมานี้ว่า "ชื่อเฉพาะที่เลือกมา" เมื่อได้ชื่อเฉพาะที่เลือกมาจากวิธีทางสถิติเหล่านี้แล้ว จากนั้นจะมีการตัดสินว่าชื่อเฉพาะที่เลือกมาจากวิธีทางสถิติวิธีไหนที่ให้ผลดีที่สุด โดยจะตัดสินจาก อัตราการรู้จำ (recognition rate) และ เวลาที่ใช้ในการรู้จำ (recognition time) โดยอัตราการรู้จำในที่นี้ คือค่า F ซึ่งค่า F สามารถหาได้จาก ค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall) ซึ่งจะกล่าวถึงในตอนต่อไป

เมื่อเปรียบเทียบค่าอัตราการรู้จำ (recognition rate) และ เวลาที่ใช้ในการรู้จำ (recognition time) แล้ว จะทำให้ได้คำตอบว่าวิธีการสถิติแบบไหนที่เลือกชื่อเฉพาะได้ดีที่สุด

### 3.3 เปรียบเทียบวิธีทางสถิติที่ใช้ในการหาขอบเขตชื่อเฉพาะ

การทดสอบจะทำโดยการนำรายการชื่อเฉพาะที่เลือกมาด้วยวิธีทางสถิติ 5 วิธีได้แก่ ค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio ($MI^3$) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation มาตรวจสอบว่าเป็นชื่อเฉพาะจริงหรือไม่ ทั้งนี้เพื่อจะหาคำตอบว่าวิธีการสถิติแบบใดที่เหมาะสมที่สุดที่จะใช้ร่วมกับ Localmax algorithm ในการหาขอบเขตของชื่อเฉพาะ ซึ่งชื่อเฉพาะที่เลือกมานี้ ในงานวิจัยนี้จะจำกัดขนาดมากที่สุด 8 พยางค์ ชื่อเฉพาะที่เลือกมาจึงมีตั้งแต่ 2-8 พยางค์

การตรวจสอบจะเปรียบเทียบรายการของชื่อเฉพาะที่เลือกมากับรายการของชื่อเฉพาะที่ถูกต้อง โดยจะเปรียบเทียบทั้งตำแหน่งที่เป็นจุดเริ่มและจำนวนพยางค์ด้วย ซึ่งถ้าหากตำแหน่งที่เป็นจุดเริ่มต้นของชื่อเฉพาะที่ถูกต้องและจุดเริ่มต้นของชื่อเฉพาะที่เลือกมาไม่ตรงกัน ชื่อเฉพาะที่เลือกมานั้นจะถูกคัดทิ้ง และถ้าจำนวนพยางค์ของคำในชื่อเฉพาะที่เลือกมามีมากหรือน้อยกว่าคำจากรายการชื่อเฉพาะที่ถูกต้อง คำนั้นๆ ก็จะถูกคัดทิ้งเช่นกัน

ตัวอย่าง ถ้าชื่อเฉพาะที่เลือกมามี 2 พยางค์ เช่น "สิ – ริ" แต่คำจากรายการชื่อเฉพาะที่ถูกต้องคือ "สิ-ริ-กร" ดังนั้น ชื่อเฉพาะที่เลือกมาตัวนี้ก็จะถูกคัดทิ้งไป

ส่วนในอีกกรณีหนึ่ง ที่ชื่อเฉพาะที่เลือกมามีส่วนของคำนำหน้าชื่อติดมาด้วย เช่น "นาง-สิ-ริ-กร" ชื่อเฉพาะที่เลือกมานี้ก็จะถูกคัดทิ้งไปด้วย

จากการตรวจสอบรายการของชื่อเฉพาะที่เลือกมาจากวิธีทางสถิติทุกแบบแล้ว จะได้รายการของชื่อเฉพาะที่เลือกมาและตรงกับรายการชื่อเฉพาะที่ถูกต้องออกมา ซึ่งจำนวนของรายการของชื่อ

เฉพาะที่เลือกมาจากวิธีทางสถิติแต่ละวิธีนั้น จะถูกนำมาตัดสินจาก อัตราการรู้จำ (recognition rate) และ เวลาที่ใช้ในการรู้จำ (recognition time) โดยอัตราการรู้จำในที่นี้ คือค่า F ซึ่งค่า F สามารถหาได้จาก ค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall)

ซึ่งค่าเหล่านี้จะหาได้จากสูตร

ค่าความแม่นยำ(P) = (จำนวนชื่อเฉพาะที่เลือกมาแล้วถูกต้อง *100)/ จำนวนชื่อเฉพาะที่เลือกออกมาทั้งหมด

ค่าความครบถ้วน (R) = (จำนวนชื่อเฉพาะที่เลือกมาแล้วถูกต้อง *100) / จำนวนชื่อเฉพาะจริงทั้งหมดในข้อมูล

อัตราการรู้จำ หรือ ค่า F คือค่าที่เฉลี่ยให้ความสำคัญกับค่าความแม่นยำและความครบถ้วนเท่าๆกัน = (2 *P *R) / (P+R)

เวลาที่ใช้ในการรู้จำ (recognition time) สามารถคำนวณได้จากจำนวนชื่อเฉพาะที่เลือกออกมาได้ต่อ 1 นาที

ซึ่งผลการทดสอบรายการของชื่อเฉพาะที่เลือกมาจากวิธีทางสถิติอันได้แก่การหาค่า Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio ($MI^3$) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation ร่วมกับการใช้ Localmax algorithm นั้น เมื่อคำนวณค่าความแม่นยำ (Precision) , ค่าความครบถ้วน (Recall) และค่า F (อัตราการรู้จำ) แล้วจะได้ผลดังแสดงอยู่ในตารางที่ 2 โดย จำนวน candidate ทั้งหมด จะเป็น จำนวนรายการของชื่อเฉพาะที่เลือกมาด้วยวิธีทางสถิติแต่ละแบบ

ตารางที่ 2 ผลการทดสอบค่าความแม่นยำ ค่าความครบถ้วนและอัตราการรู้จำที่ได้จากวิธีทางสถิติ 5 วิธี

| วิธีทางสถิติ | ชื่อเฉพาะที่เลือกมาแล้วถูกต้อง | ชื่อเฉพาะที่เลือกมาทั้งหมด | ชื่อเฉพาะที่ถูกต้องทั้งหมด | ค่าความแม่นยำ (%) | ค่าความครบถ้วน (%) | ค่า F |
|---|---|---|---|---|---|---|
| Mutual Information | 8746 | 534999 | 11683 | 1.635 | 74.861 | 3.200 |
| Pearson's Chi-square | 5800 | 542241 | 11683 | 1.070 | 49.645 | 2.095 |
| Cubic Association Ratio | 7406 | 735421 | 11683 | 1.007 | 63.391 | 1.983 |
| Dunning's Log Likelihood | 7396 | 738974 | 11683 | 1.001 | 63.306 | 1.971 |
| Mutual Expectation | 698 | 23270 | 988 | 3.000 | 70.648 | 5.756 |

จากตารางที่ 2 จะสังเกตได้ว่า วิธี Mutual Expectation ให้ผลออกมาต่างจากวิธีอื่นๆ โดยจำนวนชื่อเฉพาะทั้งหมดมีแค่เพียง 988 ชื่อเท่านั้น ทั้งนี้เป็นเพราะในการประมวลผลของระบบที่ใช้วิธีทางสถิติวิธีนี้ ใช้เวลานานมากกว่าจะประมวลผลสำเร็จ ดังนั้นจึงมีการตัดทอนข้อมูลลงเพื่อให้การประมวลผลเป็นไปได้ ซึ่งคลังข้อมูลใช้ฝึก (training corpus) ที่ใช้ประมวลผลกับระบบนี้นั้นจึงมีประมาณ 10% ของคลังข้อมูลใช้ฝึกทั้งหมดที่ใช้กับวิธีทางสถิติวิธีอื่น

ตารางที่ 3 ผลการทดสอบเวลาที่ใช้ในการรู้จำที่ได้จากวิธีทางสถิติ 5 วิธี

| วิธีทางสถิติ | เวลาที่ใช้ (นาที) | ชื่อเฉพาะที่เลือกมาทั้งหมด | เวลาที่ใช้ในการรู้จำ (จำนวนชื่อ/นาที) |
|---|---|---|---|
| Mutual Information | 1 | 534999 | 53499.9 |
| Pearson's Chi-square | 1 | 542241 | 54224.1 |
| Cubic Association Ratio | 1 | 735421 | 73542.1 |
| Dunning's Log Likelihood | 1 | 738974 | 73897.4 |
| Mutual Expectation | 1 | 23270 | 0.86 |

และจากตารางที่ 2 ผลการวัดค่าความแม่นยำ ค่าความครบถ้วน และ คำนวณหาค่าอัตราการรู้จำ (ค่า F) ดังกล่าว จะพบว่า วิธี Mutual Expectation ให้ค่าอัตราการรู้จำ (ค่า F) ออกมามากที่สุด ซึ่งสามารถสรุปได้ว่าจากวิธีทางสถิติทั้ง 5 วิธี อันได้แก่วิธี Mutual Information , ค่า Pearson's Chi-square , ค่า Cubic association ratio ($MI^3$) , ค่า Dunning's Log Likelihood และค่า Mutual Expectation นั้น วิธีที่ใช้ค่า Mutual Expectation สามารถดึงเอาชื่อเฉพาะออกมาแล้วมีความถูกต้องมากที่สุด แต่จากตารางที่ 3 พบว่าวิธีที่ใช้ค่า Mutual Expectation นี้ใช้เวลาในการประมวลผลนานมากกว่าการใช้วิธีการสถิติแบบอื่นๆ ทำให้เวลาที่ใช้ในการรู้จำหรือ recognition time นั้นมีค่าน้อยที่สุด โดยที่วิธีนี้ใช้เวลาในการดึงกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะด้วยความเร็ว 0.86 ชื่อ/ 1 นาที ในขณะที่เมื่อใช้วิธี Mutual Information จะดึงกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะด้วยความเร็ว 53499.9 ชื่อ/ 1 นาที ดังนั้น เมื่อเปรียบเทียบอัตราในการรู้จำแล้ว ผู้วิจัยจึงเลือกใช้ค่าทางสถิติที่ได้มาจากวิธี Mutual Information แทน ซึ่งวิธีนี้เป็นวิธีที่ให้ผลการทดสอบออกมาดีเป็นที่สุดเมื่อเทียบกับอีก 3 วิธีที่เหลือ ดังนั้น วิธี Mutual Information จึงเป็นวิธีทางสถิติที่เหมาะสมที่สุดที่จะนำมาใช้ในการคำนวณค่าความสัมพันธ์ภายในของกลุ่มพยางค์ โดยจะนำมาใช้ร่วมกับ Localmax algorithm ในการหาขอบเขตของชื่อเฉพาะในงานวิจัยนี้

### 3.4 อภิปรายผล

จากตารางที่ 2 การที่ค่าความแม่นยำของผลที่ได้มาจากวิธีทางสถิติต่างๆ เหล่านี้มีค่าต่ำมาก อาจจะเป็นผลมาจากการที่คลังข้อมูลที่ใช้ในงานนี้เป็นคลังข้อมูลที่ผ่านการแยกพยางค์จึงทำให้จำนวนชื่อเฉพาะที่เลือกออกมาได้มีปริมาณมาก และส่งผลให้ค่าความแม่นยำมีค่าต่ำมากด้วย ซึ่งทั้งนี้อาจเป็นเพราะจากการที่สาย n-grams ซึ่งเป็นพยางค์ต่อๆ กันนี้เข้าสู่โปรแกรมที่ใช้วิธีทางสถิติต่างๆ อันได้แก่ วิธี Mutual Information , วิธี Dunning's Log Likelihood , วิธี Pearson's Chi-square , วิธี Cubic association ratio ($MI^3$) และ วิธี Mutual Expectation ซึ่งวิธีทางสถิติเหล่านี้มีหลักการร่วมกันคือการคำนวณหาค่าความสัมพันธ์ระหว่างหน่วยย่อยโดยหน่วยย่อยซึ่งในที่นี้คือพยางค์ที่มีความสัมพันธ์ระหว่างกันสูงหรือปรากฏร่วมกันบ่อยๆ ในคลังข้อมูลจะถูกดึงออกมา ซึ่งกลุ่มพยางค์ที่มีค่าความสัมพันธ์ระหว่างกันมากๆ อาจจะไม่ได้เป็นชื่อเฉพาะก็ได้ เช่นคำว่า "ปรึกษา" จะถูกแบ่งออกเป็นสองพยางค์ จากนั้นเมื่อผ่านเข้าสู่กระบวนการประมวลผลทางสถิติ จะพบว่า "ปรึกษา" มีความสัมพันธ์ระหว่างกันค่อนข้างสูง เพราะพบว่าคำว่า "ปรึก" กับ "ษา" มักอยู่ติดกัน โดยคำว่า "ปรึก" จะไม่นำหน้าคำอื่นๆ ได้อีกเลย แม้ว่าคำว่า "ษา" อาจจะตามหลังคำอื่นอย่างคำว่า "รัก" ได้ก็ตาม จึงทำให้คำนี้ถูกเลือกออกมาด้วยวิธีการสถิติด้วย

แต่จากผลการคำนวณค่าความครบถ้วนจะพบว่าวิธีทางสถิติทั้ง 4 วิธีให้ผลไม่ต่ำกว่า 60% ซึ่งนับว่าค่อนข้างดี ยกเว้นวิธีไคกำลังสอง (Pearson's Chi-square) ซึ่งให้ผล 49.645% แสดงให้เห็นว่าชื่อเฉพาะก็เป็นกลุ่มคำที่มีความสัมพันธ์กันค่อนข้างสูง ชื่อเฉพาะส่วนหนึ่งที่ไม่สามารถเลือกออกมาได้ด้วยวิธีการสถิติ เพราะวิธีทางสถิติเหล่านี้ก็ยังมีข้อจำกัดประการหนึ่งคือ กรณีที่ชื่อเฉพาะเป็นคำที่มีเพียงพยางค์เดียว เช่น นาย<person>ชวน</person> หรือ ประเทศ<location>ไทย</location> จะไม่สามารถผ่านเงื่อนไขของวิธีทางสถิติที่พยายามหาค่าความสัมพันธ์ระหว่างพยางค์ เพราะชื่อเฉพาะที่จะผ่านเงื่อนไขไปได้ต้องมีไม่ต่ำกว่า 2 พยางค์ขึ้นไป โดยชื่อเฉพาะที่มีเพียงพยางค์เดียวจะมีจำนวน 1028 จากชื่อเฉพาะจำนวน 11683 ชื่อ โดยคิดเป็น 8.799% ของชื่อเฉพาะทั้งหมด

และจากผลการคำนวณอัตราการรู้จำหรือค่า F แสดงให้เห็นว่าวิธีที่ใช้ค่า Mutual Expectation นั้นให้ผลดีที่สุด ทั้งนี้อาจเป็นเพราะแนวคิดของวิธีนี้เป็นการหาความสัมพันธ์ระหว่างคำโดยการคำนวณความเป็นไปได้ที่คำต่างๆ ในสาย n-grams จะหายไป ซึ่งทำให้ได้ค่าความแม่นยำสูงที่สุดในวิธีทางสถิติทั้งหมด แต่วิธี Mutual Expectation นี้ก็ใช้เวลาในการประมวลผลมาก จนทำให้ผู้วิจัยเลือกใช้วิธีทางสถิติที่ให้ผลดีรองลงมา นั่นคือวิธี Mutual Information

# บทที่ 4

# การรู้จำชื่อเฉพาะ

ในบทนี้จะกล่าวถึงการรู้จำชื่อเฉพาะ โดยเป็นการใช้กฎที่เขียนขึ้นเพื่อพิจารณารายการชื่อเฉพาะที่เลือกมาด้วยวิธีการทางสถิติว่าเป็นชื่อเฉพาะจริงหรือไม่และเป็นชื่อเฉพาะประเภทใด ซึ่งบทนี้จะแบ่งเนื้อหาออกเป็น 4 หัวข้อ โดยหัวข้อแรกจะกล่าวถึงการสร้างกฎจากหลักฐานภายใน (internal evidence) รวมถึงหลักฐานจากบริบทข้างเคียง (external evidence) และในหัวข้อส่วนที่สองจะกล่าวถึงกฎที่ถูกสร้างขึ้น ส่วนที่สามและสี่จะกล่าวถึงผลของการใช้กฎที่สร้างขึ้นเพื่อตัดสินการจำแนกประเภทของชื่อเฉพาะว่ามีความถูกต้องมากน้อยเพียงใดและการอภิปรายผลการทดสอบที่ได้

## 4.1 การจำแนกประเภทชื่อเฉพาะ

หลังจากที่ได้เลือกใช้วิธีการทางสถิติแบบ Mutual Information ร่วมกับการใช้ Localmax algorithm เพื่อหาขอบเขตและเลือกชื่อเฉพาะออกมาแล้ว ขั้นตอนต่อไป คือ การตัดสินว่าแต่ละหน่วยที่ดึงมานั้นเป็นชื่อเฉพาะจริงหรือไม่และเป็นชื่อเฉพาะประเภทใด ขั้นตอนนี้จะอาศัยกฎที่สร้างขึ้นเพื่อใช้รู้จำและจำแนกประเภทของชื่อเฉพาะ โดยใช้หลักฐานจากบริบทภายใน (internal evidence) และบริบทข้างเคียง (external evidence) ซึ่งจะมีส่วนช่วยในการรู้จำตำแหน่งและจำแนกประเภทของชื่อเฉพาะ

### 4.1.1 หลักฐานภายใน

หลักฐานภายใน (internal evidence) หมายถึงคำที่มักจะพบว่าปรากฏร่วมกับชื่อเฉพาะบ่อยๆ และกลุ่มคำที่เป็นหลักฐานภายในนี้ก็จะมีลักษณะเป็นกลุ่มคำแบบปิด หรือ closed class กล่าวคือสามารถคาดเดาจำนวนของคำทั้งหมดที่เป็นหลักฐานภายในได้ หากมีคลังข้อมูลที่ใหญ่เพียงพอ ยกตัวอย่างเช่น หลักฐานภายในที่แสดงอรรถลักษณ์เป็น <ปัจเจกบุคคล> และ <บอกอายุ เพศ และสถานภาพการสมรส> ซึ่งจัดอยู่ในประเภทชื่อคนนั้น ก็จะมีกลุ่มคำนำหน้าชื่อ ได้แก่ “นาย” , “นาง” , “นางสาว” เป็นต้น

หลักฐานภายใน ซึ่งได้แก่คำนำหน้าและคำที่ตามหลังชื่อเฉพาะนั้นจะสามารถหาได้จากการพิจารณาจากคลังข้อมูลใช้ฝึก ซึ่งหลักฐานภายในเหล่านี้จะช่วยในการเขียนกฎเพื่อกำหนดขอบเขตของชื่อเฉพาะโดยอาจจะใช้บ่งบอกถึงจุดเริ่มต้นของชื่อเฉพาะ รวมทั้งจุดสิ้นสุดของชื่อ

เฉพาะได้ โดยหลักฐานภายในจะถูกแบ่งออกเป็น 3 ประเภทตามลักษณะการปรากฏของมันกับชื่อเฉพาะ ดังนี้

1. หลักฐานภายในที่ปรากฏอยู่หน้าชื่อเฉพาะ ได้แก่คำนำหน้าชื่อเฉพาะต่างๆ ซึ่งรายการคำนำหน้าชื่อที่เก็บได้มาจากคลังข้อมูลจะแสดงอยู่ด้านล่าง

   "อ." , "นาย" , "นาง" , "ดร." , "พล.ต." , "พล.ท." , "พ.ญ." , "พ.อ." , "ส.ก." , "ร.ต." , "ร.อ." , "ภ.ญ." , "ร.ท." , "แม่อุ้ย" , "จอมพล" , "น.ส." , "นางสาว" , "น.ต." , "อาจารย์" , "หัวหน้า" , "พล.อ." , "น.พ." , "ม.ล." , "คุณหญิง" ,"หลวงพ่อ" , "ร.อ.หญิง" , "พ.ต.ท." , "พ.ต.หญิง" , "ส.ต.อ." , "พล.ร.ท." , "ร.ต.อ." , "พล.ต.ท." , "ม.ร.ว." , "พล.อ.อ." , "พล.ต.ต." , "พล.ต.อ." , "ส.ต.ต." , "พ.ต.ต." , "พ.ต.อ." , "พล.ต.ท." , "ร.ต.ต" , "จ.ส.ต." , "ร.ต.ท." , "พ.ญ.คุณหญิง" , "ว่าที่ร.ต." , "พ.ต.อ.น.พ." , "กรม" , "พรรค" , "องค์การ" , "กระทรวง" , "สำนักงาน" , "สำนัก" ,"กองทัพ" , "บริษัท" , "ธนาคาร" , "องค์กร" , "กอง" , "โรงเรียน" , "โรงพยาบาล" , "วิทยาลัย" , "บรรษัท" , "สหภาพ" , "สภา" , "มูลนิธิ" , "ศาล" , "สถาบัน" , "ศูนย์" , "ทวีป" , "ร้านกาแฟ" , "ร้านอาหาร" , "สวนสัตว์" , "ค่าย" , "สถานทูต" ,"สนามกอล์ฟ" , "ท่าอากาศยาน" , "สุเหร่า" , "บึง" , "คุก" , "มัสยิด" , "สวนสาธารณะ" , "ร.ร." , "พระธาตุ" , "เกาะ" , "สโมสร" , "สมาคม" , "สาธารณรัฐ" , "ราชอาณาจักร" , "จังหวัด" , "แม่น้ำ" , "ลุ่มน้ำ" , "เขา" , "อ่างเก็บน้ำ" , "ท่าน้ำ" , "ท่าเรือ" , "ท่าเรือน้ำลึก" , "ปราสาท" , "ปราสาทหิน" , "คลอง" , "วัด" , "วัง" , "รพ." , "อนุสาวรีย์" , "พระบรมราชานุสาวรีย์" , "เขื่อน" , "อนุสรณ์" , "มณฑล" , "นคร" , "ต." , "อาคาร" , "ตึก" , "โรงแรม" , "สนามบิน" , "ประเทศ" , "แขวง" , "หมู่บ้าน" , "สนามม้า" , "แฟลต" , "ซอย" , "ร้าน" , "เคหะชุมชน" , "โรงงาน" , "สวน" , "บ้าน" , "ธนาคาร" , "ศาลเจ้า" , "สน." , "อุทยาน" , "ถ." , "เขตเทศบาลตำบล" , "ศูนย์ประชุม" , "หอประชุมอำเภอ" , "ทางด่วนสาย" , "ถนนสาย" , "สำนักงานสาขา" , "หอประชุม" , "ที่ว่าการ" , "มูลนิธิ" , "ภูมิภาค" , "สถานี" , "ถนน" , "ที่ทำการสาขา" , "ห้อง" , "สถาบัน" , "อำเภอ" , "หอ" , "ที่ทำการ" , "ภาค" , "ตลาดสดเทศบาล" , "ตลาดสด" , "ตลาด" , "สนามกีฬาเทศบาล" , "สนามกีฬากลาง" , "สนามกีฬา" , "สนาม" , "ศาลา" , "จ." , "เมือง" , "กรุง" , "เขต"
2. หลักฐานภายในที่ปรากฏอยู่ด้านหน้าและหลังชื่อเฉพาะ ได้แก่ บริษัท...จำกัด
3. หลักฐานภายในที่อาจปรากฏอยู่ด้านหน้าหรือด้านหลังชื่อเฉพาะก็ได้ ได้แก่ มหาวิทยาลัย เช่น มหาวิทยาลัยธรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

จากการพิจารณาหลักฐานภายในเหล่านี้ พบว่าชื่อเฉพาะประเภทชื่อคนจะมีหลักฐานชนิดนี้มากที่สุด โดยจะมีคำนำหน้าชื่อถึง 5408 จากจำนวนชื่อคน 5431 ชื่อโดยคิดเป็น 99.577% ชื่อองค์กรมีคำนำหน้าชื่อ 3140 จากจำนวนชื่อองค์กร 3222 ชื่อโดยคิดเป็น 97.455% และชื่อสถานที่มีคำนำหน้าชื่ออยู่ 2158 จากจำนวนชื่อสถานที่ 3030 ชื่อโดยคิดเป็น 71.221% ดังนั้น จากการพบหลักฐานภายในเหล่านี้ การสร้างกฎจึงใช้คำนำหน้าชื่อนี้เป็นหลักในการระบุขอบเขตของชื่อเฉพาะ โดยจะเป็นการกำหนดจุดเริ่มต้นของชื่อเฉพาะนั่นเอง นอกจากนั้นแล้ว คำนำหน้าชื่อยังสามารถบ่งบอกประเภทของชื่อเฉพาะได้อีกด้วย ซึ่งการแบ่งประเภทของชื่อเฉพาะตามคำนำหน้าชื่อนั้น อาจทำได้โดยการศึกษาความหมายของคำนำหน้าชื่อแต่ละคำ ด้วยการวิเคราะห์องค์ประกอบของอรรถลักษณ์ (componential analysis) ซึ่งเป็นวิธีที่นำความหมายของคำนำหน้าชื่อแต่ละคำมาแจกแจงเป็นส่วนความหมายย่อยที่เรียกว่าอรรถลักษณ์ (semantic feature) (อมรทิพย์, 2546) ซึ่งพบว่าจากอรรถลักษณ์ร่วมของกลุ่มคำ จะทำให้แบ่งประเภทของคำนำหน้าชื่อเฉพาะได้ดังตารางที่ 4 ด้านล่าง โดยจะแบ่งออกเป็น 3 ประเภทคือ ชื่อคน ชื่อองค์กร และชื่อสถานที่

ตารางที่ 4 การแบ่งประเภทกลุ่มคำนำหน้าชื่อ

| ประเภทของชื่อเฉพาะ | อรรถลักษณ์ร่วมของกลุ่มคำนำหน้าชื่อ | คำนำหน้าชื่อ |
|---|---|---|
| ชื่อคน | <ปัจเจกบุคคล><br><แสดงฐานะหรือลำดับขั้น> | "พล.ต." , "พล.ท." , "พ.อ." , "พล.อ." , "ส.ก." , "ร.ต." , "ร.อ." , "ร.ท." , "พล.อ.อ." , "พล.ต.ต." , "พล.ต.อ." , "ส.ต.ต." , "พ.ต.ต." , "พ.ต.อ." , "พล.ต.ท." , "ร.ต.ต" , "จ.ส.ต." , "ร.ต.ท." , "ร.อ.หญิง" , "พ.ต.ท." , "พ.ต.หญิง" , "ส.ต.อ." , "พล.ร.ท." , "ร.ต.อ." , "พล.ต.ท." , "จอมพล" , "น.ต." , "ว่าที่ร.ต." , "พ.ต.อ.น.พ." , "คุณหญิง" |
| | <ปัจเจกบุคคล><br><แสดงระดับการศึกษา> | "ดร." |

| | | |
|---|---|---|
| | <ปัจเจกบุคคล><br><แสดงฐานันดร> | "ม.ร.ว." , "ม.ล." |
| | <ปัจเจกบุคคล><br><แสดงสถานะทางศาสนา> | "หลวงพ่อ" |
| | <ปัจเจกบุคคล><br><คำเรียกทางวิชาชีพ> | "พ.ญ." , "ภ.ญ." , "น.พ.", "อาจารย์", "พ.ญ.คุณหญิง" |
| | <ปัจเจกบุคคล><br><บอกอายุ เพศ หรือ/และ สถานภาพการสมรส> | "นาย" , "นาง" , "น.ส." , "นางสาว" , "แม่อุ้ย" |
| ชื่อองค์กร | <กลุ่มบุคคล><br><มีการรวมตัวกัน><br><ประกอบขึ้นเป็นหน่วยเดียว><br><มีสภาพเป็นนิติบุคคล><br><เพื่องานราชการ> | "กรม"* , "องค์การ" , "กระทรวง"* , "กองทัพ" , "กอง" , "ศาล" |
| | <กลุ่มบุคคล><br><มีการรวมตัวกัน><br><ประกอบขึ้นเป็นหน่วยเดียว><br><มีสภาพเป็นนิติบุคคล><br><เพื่อทำธุรกิจ> | "บริษัท"* , "ธนาคาร"* , "บรรษัท" , “บริษัท...จำกัด” |
| | <กลุ่มบุคคล><br><มีการรวมตัวกัน><br><ประกอบขึ้นเป็นหน่วยเดียว><br><มีสภาพเป็นนิติบุคคล><br><เพื่อทำกิจการตามวัตถุประสงค์ใดวัตถุประสงค์หนึ่ง> | "พรรค" , "สำนักงาน"* , "สำนัก" , "องค์กร" , "สหภาพ" , "สภา" , "มูลนิธิ"* , "สถาบัน" , "ศูนย์" |
| ชื่อสถานที่ | <พื้นที่หรือบริเวณ> | "ร้านกาแฟ" , "ร้านอาหาร" , |

* คำนำหน้าชื่อที่สามารถจัดให้ชื่อเฉพาะอยู่ในประเภทที่ต่างกัน 2 ประเภท

| | | |
|---|---|---|
| | <สร้างหรือกำหนดขึ้น><br><สำหรับใช้ประโยชน์หรือทำกิจกรรม> | "สวนสัตว์" , "ค่าย" , "สถานทูต" ,"สนามกอล์ฟ" , "ท่าอากาศยาน" , "สุเหร่า" , "คุก" , "มัสยิด" , “โรงเรียน” , "สวนสาธารณะ" , "ร.ร." , "พระธาตุ" , "สโมสร" , "สมาคม" , "อ่างเก็บน้ำ" , "ท่าน้ำ" , "ท่าเรือ" , "ท่าเรือน้ำลึก" , "ปราสาท" , "ปราสาทหิน" , "วัด" , "วัง" , "รพ." , "อนุสาวรีย์" , "พระบรมราชานุสาวรีย์" , "เขื่อน" , "อนุสรณ์" , "อาคาร" , "ตึก" , "โรงแรม" , "สนามบิน" , "สนามม้า" , "แฟลต" , "ซอย" , "ร้าน" , "เคหะชุมชน" , "โรงงาน" , "สวน" , "ศาลเจ้า" , "สน." , "ถ." , "ศูนย์ประชุม" , "หอประชุมอำเภอ" , "ทางด่วนสาย" , "ถนนสาย" , "สถานี" , "สำนักงานสาขา" , "หอประชุม" , "ที่ว่าการ" , "ถนน" , "ที่ทำการสาขา" , "ห้อง" , "หอ" , "ศาลา" , "ที่ทำการ" , "ตลาดสดเทศบาล" , "ตลาดสด" , "ตลาด" , "สนามกีฬาเทศบาล" , "สนามกีฬากลาง" , "สนามกีฬา" , "สนาม" , “มหาวิทยาลัย" , “โรงพยาบาล" |

| | | |
|---|---|---|
| | <พื้นที่หรือบริเวณ><br><สร้างหรือกำหนดขึ้น><br><เป็นเขตปกครอง> | "ทวีป" , "สาธารณรัฐ" , "ราชอาณาจักร" , "จังหวัด" , "มณฑล" , "นคร" , "ต." , "ประเทศ" , "แขวง" , "หมู่บ้าน" , "บ้าน" , "เขตเทศบาลตำบล" , "อำเภอ" , "ภาค", "จ." , "เมือง" , "กรุง" , "เขต" ,"อ." |
| | <พื้นที่หรือบริเวณ><br><สร้างหรือกำหนดขึ้น><br><ระบุตำแหน่งหรือบริเวณ> | "ภูมิภาค" |
| | <พื้นที่หรือบริเวณ><br><เกิดขึ้นตามธรรมชาติ><br><ส่วนพื้นน้ำ> | "บึง" , "แม่น้ำ" , "ลุ่มน้ำ" , "คลอง" |
| | <พื้นที่หรือบริเวณ><br><เกิดขึ้นตามธรรมชาติ><br><ส่วนพื้นดิน> | "เกาะ" , "เขา" , "อุทยาน" |

จากตารางที่ 4 จะพบว่ามีคำนำหน้าชื่อเฉพาะที่สามารถจัดประเภทชื่อเฉพาะให้อยู่ในสองประเภทได้ โดยอาจจะกำกวมอยู่ระหว่างชื่อเฉพาะประเภทชื่อองค์กรและชื่อสถานที่ ซึ่งคำนำหน้าชื่อดังกล่าวได้แก่ “กรม” , “กระทรวง” , “บริษัท” , “ธนาคาร” , “มูลนิธิ” , “สำนักงาน”

จากคำนำหน้าชื่อทั้ง 6 คำนี้ หากพบว่ามีหลักฐานภายนอก (external evidence) ที่ได้มาจากบริบทข้างเคียง เช่น “พื้นที่” , “บริเวณ” , “ที่” , “รอบ” เป็นต้น ปรากฏอยู่ด้านหน้าจะทำให้สามารถนำมาใช้ในความหมายอ้างอิงถึงสถานที่ ดังตัวอย่าง

“เวลา 09.00น. วันเดียวกัน ที่ กรม<location>ประชาสัมพันธ์</location>”

“ไปไหว้สิ่งศักดิ์สิทธิ์ ที่ กระทรวง<location>มหาดไทย</location>ต่อไป”

“พื้นที่ บริษัท<location>ไทยเมล่อน</location>”

“บริเวณ รอบ บริษัท<location>ไทยเมล่อน</location>เป็นพื้นที่ของนักการเมือง”

ดังนั้นหลักฐานภายนอกที่ได้มาจากบริบทข้างเคียงจึงเป็นส่วนสำคัญอีกส่วนหนึ่งที่จะนำมาใช้ประกอบการเขียนกฎเพื่อให้สามารถระบุประเภทของชื่อเฉพาะได้อย่างแม่นยำ

### 4.1.2 หลักฐานภายนอก

หลักฐานภายนอก (external evidence) เป็นหลักฐานที่สามารถหาได้จากบริบทข้างเคียงที่มีชื่อเฉพาะนั้นๆ ปรากฏอยู่ด้วย ซึ่งจะสามารถหาได้จากการพิจารณาลักษณะของรูปภาษาในคลังข้อมูลและดึงหลักฐานที่ได้มาใช้เขียนกฎ ซึ่งจากคลังข้อมูลพบว่าหลักฐานภายนอกนี้มีความหลากหลาย อีกทั้ง หลักฐานแบบนี้จะไม่สามารถคาดเดาจำนวนของกลุ่มคำได้ (opened class)

หลักฐานภายนอกนี้ยังสามารถจำแนกและลดความกำกวมในการจำแนกประเภทของชื่อเฉพาะได้ ดังตัวอย่าง

“ไปไหว้สิ่งศักดิ์สิทธิ์ ที่ กระทรวง<location>มหาดไทย</location>ต่อไป”

ซึ่ง “กระทรวง” เป็นคำนำหน้าชื่อที่บ่งบอกประเภทของชื่อเฉพาะว่าเป็นชื่อองค์กร แต่ในบริบทดังตัวอย่าง คำว่า “ที่” จะช่วยบ่งชี้ได้ว่า “กระทรวงมหาดไทย” ในบริบทนี้ เป็นชื่อสถานที่มิใช่ชื่อองค์กร คำว่า “ที่” นี้จึงจัดเป็นหลักฐานภายนอก

นอกจากนั้น หลักฐานภายนอกยังเป็นส่วนสำคัญในการกำหนดจุดเริ่มต้นและจุดสิ้นสุดของขอบเขตชื่อเฉพาะได้อีกด้วย โดยสำหรับกรณีที่ชื่อเฉพาะนั้นๆ ไม่มีคำนำหน้าชื่อเป็นหลักฐานภายในเพื่อบ่งบอกจุดเริ่มต้นของขอบเขตชื่อเฉพาะ หลักฐานภายนอกก็จะมีส่วนช่วยในการแก้ปัญหานี้ได้ เช่นตัวอย่าง

“ที่ดินย่าน<location>เกียกกาย</location>”

จะเห็นได้ว่า “เกียกกาย” ซึ่งเป็นชื่อเฉพาะประเภทสถานที่ ไม่มีคำนำหน้าชื่อเป็นหลักฐานภายในเพื่อบ่งบอกจุดเริ่มต้นของขอบเขตชื่อเฉพาะ ดังนั้น การใช้หลักฐานภายนอกที่ได้จากบริบทข้างเคียงอย่างคำว่า “ย่าน” จะมีส่วนช่วยในการระบุตำแหน่งและจำแนกประเภทของชื่อเฉพาะ

## 4.2 การสร้างกฎการจำแนกประเภทชื่อเฉพาะ

ในส่วนนี้จะกล่าวถึงการสร้างกฎเพื่อใช้ในการตัดสินว่ารายการชื่อที่ได้จากวิธีการสถิตินั้นรายการไหนเป็นชื่อเฉพาะและเป็นชื่อเฉพาะประเภทใด โดยอาศัยข้อมูลจากหลักฐานภายในและหลักฐานภายนอกดังที่กล่าวมาแล้ว ซึ่งจากการพิจารณาลักษณะของรูปภาษาในคลังข้อมูลพบว่าการระบุตำแหน่งเพื่อหาขอบเขตของชื่อเฉพาะภาษาไทยเป็นไปได้ยาก เพราะภาษาไทยไม่มีสัญลักษณ์หรือลักษณะพิเศษ เช่น ตัวพิมพ์ใหญ่ (capitalization) ที่สามารถใช้ระบุจุดเริ่มต้นของชื่อเฉพาะได้เหมือนในภาษาอังกฤษ อีกทั้งยังไม่มีการเว้นวรรคระหว่างคำเพื่อให้การระบุจุดสิ้นสุดของขอบเขตชื่อเฉพาะเป็นไปได้ง่ายด้วย อย่างไรก็ตาม จากคลังข้อมูล พบว่าชื่อเฉพาะจะมีหลักฐานที่ได้มาจากหลักฐานภายในหรือคำนำหน้าชื่อนำอยู่ด้านหน้า โดยพบว่าชื่อเฉพาะประเภทชื่อคนจะมีหลักฐานชนิดนี้มากที่สุด โดยจะมีคำนำหน้าชื่อถึง 5408 จากจำนวนชื่อคน 5431 ชื่อ

โดยคิดเป็น 99.577% ชื่อองค์กรมีคำนำหน้าชื่อ 3140 จากจำนวนชื่อองค์กร 3222 ชื่อโดยคิดเป็น 97.455% และชื่อสถานที่มีคำนำหน้าชื่ออยู่ 2158 จากจำนวนชื่อสถานที่ 3030 ชื่อโดยคิดเป็น 71.221% จากหลักฐานภายในที่เห็นได้จากคำนำหน้าชื่อนี้เองทำให้รูปแบบของกฎจะเริ่มด้วยการใช้คำนำหน้าชื่อเป็นหลักในการกำหนดจุดเริ่มต้นของขอบเขตของชื่อเฉพาะ โดยกฎจะมีลักษณะเป็นไปดังนี้

"หากพบ [คำนำหน้าชื่อ] คำที่ตามหลังก็จะเป็น [ชื่อเฉพาะ]"

อย่างไรก็ตาม จากการพิจารณาคลังข้อมูลใช้ฝึก พบว่า รายการของคำนำหน้าชื่อบางส่วนก็มิได้นำหน้าชื่อเฉพาะเพียงอย่างเดียว รายการของคำดังกล่าวยังนำหน้าคำทั่วไปด้วย เช่น "พรรค" อาจจะนำหน้าคำว่า "ขนาด" ก็ได้ โดย "ขนาด" มิได้เป็นชื่อเฉพาะแต่อย่างใด ดังตัวอย่าง

"จนจะกลายเป็นพรรคขนาดเล็ก"

ดังนั้น กฎที่ใช้อิงกับคำนำหน้าชื่อนี้จึงเพิ่มกฎการตรวจสอบคำที่อาจตามหลังคำนำหน้าชื่อแต่ไม่ได้เป็นชื่อเฉพาะ เพื่อคัดเอารายการของกลุ่มพยางค์ที่ไม่น่าจะใช่ชื่อเฉพาะออกไป

ซึ่ง [รายการของคำยกเว้นหลัง] อาจประกอบด้วยคำหลายประเภท ตามการกำกับหมวดคำในแนวพระยาอุปกิตศิลปสาร (อุปกิตศิลปสาร, 2541) ได้แก่

คำนำหน้ากริยา เช่น "คง" , "จะ" , "อย่า"

คำกริยา เช่น "สั่ง" , "ประกาศ" , "กล่าว" , "เห็น"

คำวิเศษณ์ เช่น "ใหญ่" , "เล็ก" , "เหม็น"

คำประพันธสรรพนาม เช่น "ที่" , "ซึ่ง" , "อัน"

คำบุพบท เช่น "ทั้ง"

คำสันธาน เช่น "กับ" , "และ" , "เพราะ" , "ก็"

เหตุที่เป็นเช่นนี้ เพราะคำนำหน้าชื่อ เช่น "พรรค" นอกจากจะใช้นำหน้าชื่อเฉพาะแล้ว ยังเป็นคำนามด้วย ซึ่งลักษณะแบบนี้ทำให้คำที่จะตามหลังคำนำหน้าชื่อเฉพาะนี้ มีโอกาสที่จะเป็นส่วนขยายของคำนาม เช่น "พรรคเล็กๆ" หรือเป็นคำในหมวดอื่นๆ ก็ได้ เช่น คำกริยา คำสันธาน คำบุพบท เป็นต้น (ดังที่แสดงไว้ด้านบน) เพราะถือเป็นการเริ่มต้นวลีใหม่ที่ไม่ใช่นามวลีเดิม

ดังนั้น กฎจึงจำเป็นต้องมีส่วนที่จะคัดกรอง [รายการของคำยกเว้นหลัง] เหล่านี้ออก เพื่อให้ความแม่นยำและความครบถ้วนในการรู้จำและจำแนกประเภทชื่อเฉพาะเป็นไปได้ดีขึ้น ผู้วิจัยจึงได้เลือกรายการคำยกเว้นเหล่านี้จากคำที่เป็นคำที่ใช้ทั่วไปและไม่พบเป็นส่วนเริ่มต้นของชื่อเฉพาะ (แต่ปัญหาของวิธีนี้ก็จะทำให้ชื่อใหม่ๆ ที่อาจเกิดขึ้นและนำหน้าด้วยคำเหล่านี้ เช่น "พรรคใหญ่เหนือฟ้า" ถูกกันออกไปด้วย) ทำให้รูปแบบของกฎจะเป็นดังนี้

"พบ [คำนำหน้าชื่อ] และคำที่ตามหลังมิได้อยู่ใน [รายการของคำยกเว้นหลัง] คำที่ตามหลังก็จะเป็น [ชื่อเฉพาะ]"

ซึ่ง [รายการของคำยกเว้นหลัง] เช่นคำว่า "ขนาด" จะแสดงไว้ในภาคผนวก

นอกจากกรณีที่กล่าวไปแล้วนั้น ในคลังข้อมูลยังพบว่ามีกรณีที่พบคำนำหน้าชื่อซึ่งเป็นหลักฐานภายในแล้ว และคำที่ตามหลังก็อาจมองว่าเป็นชื่อเฉพาะได้ แต่ก็ไม่สามารถจัดให้คำตามหลังเหล่านั้นเป็นชื่อเฉพาะได้ เช่น

หัวหน้า "พรรค" ประชาธิปัตย์

ในที่นี้จะไม่ถือว่า "ประชาธิปัตย์" ในที่นี้เป็นชื่อองค์กร เนื่องจากคำที่นำหน้า "พรรค" นั้น คือ คำว่า "หัวหน้า" ซึ่งเป็นหลักฐานภายนอกที่บ่งบอกให้ทราบว่า "หัวหน้าพรรคประชาธิปัตย์" นี้เป็นตำแหน่งทางการเมืองที่อาจอ้างถึงบุคคล มิได้ใช้เป็นชื่อองค์กร ซึ่งจากรูปภาษาดังกล่าวทำให้การพิจารณา "หัวหน้าพรรคประชาธิปัตย์" นั้นจะไม่ได้มองว่า "พรรค" เป็นส่วนหลักของวลี แต่จะมองที่ "หัวหน้า" ซึ่งเป็นส่วนหลัก (head) ของวลีแทน ซึ่งจากลักษณะดังกล่าวนี้ ยังมีกลุ่มคำที่จัดอยู่ในลักษณะเดียวกันนี้ ได้แก่ กลุ่มคำในรายการของคำยกเว้นหน้าของชื่อเฉพาะประเภทองค์กร ดังแสดงในภาคผนวก

และจากรูปแบบภาษาดังกล่าว จะทำให้กฎมีการเปลี่ยนแปลงรูปแบบ โดยจะเป็นดังที่แสดงด้านล่าง

"พบ [คำนำหน้าชื่อ] และคำที่นำหน้ามิได้อยู่ใน [รายการของคำยกเว้นหน้า] และคำที่ตามหลังมิได้อยู่ใน [รายการของคำยกเว้นหลัง] คำที่ตามหลังก็จะเป็น [ชื่อเฉพาะ]"

จากกฎข้างต้นจะทำให้สามารถระบุได้เพียงตำแหน่งเริ่มต้นของชื่อเฉพาะเท่านั้น แต่ในการหาขอบเขตสิ้นสุดของชื่อเฉพาะนั้น ในภาษาไทยซึ่งไม่มีสัญลักษณ์หรือลักษณะพิเศษเช่น การเว้นวรรค เช่นในภาษาอังกฤษ จึงทำให้การระบุหาขอบเขตสิ้นสุดของชื่อเฉพาะเป็นไปได้ยาก แต่อย่างไรก็ตาม จากการพิจารณาคลังข้อมูล พบว่าสามารถใช้หลักฐานภายนอกที่ได้มาจากบริบทข้างเคียงเป็นการบอกจุดสิ้นสุดของขอบเขตของชื่อเฉพาะได้ แต่รายการของหลักฐานภายนอกดังกล่าว ก็จะมีขนาดใหญ่ เพราะไม่อาจจะคาดเดาได้ว่าจะมีคำจากกลุ่มคำประเภทใดบ้างที่จะเป็นจุดสิ้นสุดของชื่อเฉพาะ ทำให้รายการของหลักฐานภายนอกที่สามารถระบุจุดสิ้นสุดของขอบเขตชื่อเฉพาะซึ่งได้มาจากการพิจารณาคลังข้อมูลใช้ฝึกนี้ จะสามารถใช้กับงานวิจัยนี้เท่านั้น หากมีการเพิ่มเติมขนาดของคลังข้อมูล หรือเปลี่ยนแปลงรูปแบบของข้อมูลในคลังข้อมูล อาจทำให้รายการของคำที่สามารถระบุจุดสิ้นสุดของขอบเขตชื่อเฉพาะนี้ใช้ไม่ได้ผลเต็มที่

ซึ่งในงานวิจัยนี้ พบว่ารายการของคำที่ได้มาจากหลักฐานภายนอกจากบริบทข้างเคียงเป็นตัวกำหนด ดังเช่นรายการของกลุ่มคำทั่วไปที่หาได้จากคลังข้อมูลใช้ฝึก เช่น คำว่า "และ"

“อย่า” เป็นต้น ซึ่งคำเหล่านี้แสดงว่าได้มีการขึ้นต้นวลีใหม่จึงใช้บอกจุดสิ้นสุดได้ หรืออาจเป็นกลุ่มของคำนำหน้าชื่อที่ตามมาด้านหลังเอง ดังตัวอย่าง “พ.ต.ท. ทักษิณ ชินวัตร นาง สิริกร” ซึ่งจุดเริ่มของชื่อจะเริ่มจาก “ทักษิณ” และจบที่ “ชินวัตร” และเมื่อพบคำนำหน้าชื่อว่า “นาง” อีกคำก็จะรู้ได้ว่าชื่อแรกได้จบแล้ว จึงสามารถนำมาใช้เพื่อเป็นจุดสิ้นสุดของขอบเขตของชื่อเฉพาะได้

ซึ่งกลุ่มคำที่ถูกใช้เป็นจุดสิ้นสุดของขอบเขตชื่อเฉพาะนี้ นอกจากกลุ่มคำนำหน้าชื่อแล้ว ยังอาจแบ่งกลุ่มคำออกเป็น 7 ประเภท ตามการกำกับหมวดคำในแนวพระยาอุปกิตศิลปสาร (อุปกิตศิลปสาร, 2541) ได้แก่

1. คำนาม เช่น “คน” , “บ้าน” , “บริษัท” , “ตัว”
2. คำนำหน้ากริยา เช่น “คง” , “จะ” , “อย่า”
3. คำกริยา เช่น "กล่าว" , "เปิด" , "เห็น" , "อยาก" , "สู้" , "เป็น" , “เหมือน” , “คล้าย” , "ทราบ" , "ตอบ" , "ยืน" , “ประกาศ”
4. คำวิเศษณ์ เช่น “ใกล้” , “ไกล” , “มาก” , “เล็ก”
5. คำประพันธสรรพนาม เช่น “ที่” , “ซึ่ง” , “อัน”
6. คำบุพบท เช่น “ที่” , “ทั้ง” , “ประมาณ” , “ตลอด”
7. คำสันธาน เช่น “กับ” , “และ” , “เพราะ” , “ก็” , ”ส่วน”

ซึ่งสำหรับกรณีที่คำที่ใช้ในการบอกขอบเขตของชื่อเฉพาะจำนวนหนึ่งเป็นคำนาม ยกตัวอย่างเช่น

“บ้านเมืองไม่ใช่ของพ.ต.ท.<person>ทักษิณชินวัตร</person>คนเดียว”

ซึ่งในกรณีนี้ “คน” เป็นลักษณะนามของ “พ.ต.ท.ทักษิณ ชินวัตร”

สำหรับกรณีที่คำที่ใช้ในการบอกขอบเขตสิ้นสุดของชื่อเฉพาะจำนวนหนึ่งเป็นคำกริยา ทั้งนี้เป็นเพราะชื่อเฉพาะเมื่ออยู่ในประโยคอาจจะทำหน้าที่เป็นประธานของประโยคทำให้เมื่อพบคำกริยาจึงทำให้ทราบว่าพบจุดสิ้นสุดของขอบเขตชื่อเฉพาะแล้ว เช่น

“พ.ต.ท.<person>ทักษิณชินวัตร</person>ประกาศนโยบายของพรรคชุดใหม่”

สำหรับกรณีที่คำที่ใช้ในการบอกขอบเขตของชื่อเฉพาะจำนวนหนึ่งเป็นคำวิเศษณ์นั้น ทั้งนี้เป็นเพราะชื่อเฉพาะที่มีลักษณะเหมือนคำนามที่ปรากฏหลังคำกริยาจึงอาจตามด้วยคำวิเศษณ์ที่ขยายคำกริยานั้นได้ เช่น

“ทำให้พรรค<organization>ประชาธิปัตย์</organization>เล็กลง”

สำหรับกรณีที่คำที่ใช้ในการบอกขอบเขตของชื่อเฉพาะจำนวนหนึ่งเป็นคำบุพบทนั้น ทั้งนี้เป็นเพราะชื่อเฉพาะอาจทำหน้าที่เป็นกรรมของประโยคได้ ดังนั้นคำบุพบททำหน้าที่เป็นคำเชื่อม

ระหว่างความที่อาจจะมาขยายกริยาของประโยค จึงทำให้สามารถมองเห็นจุดสิ้นสุดของขอบเขตชื่อเฉพาะได้ เช่น

"บิดเรื่องของตนกับเรื่องของพรรค<organization>ไทยรักไทย</organization> ตลอดเวลา"

ซึ่งในที่นี้ "ตลอด" เป็นบุพบทบอกเวลา

สำหรับกรณีที่คำที่ใช้ในการบอกขอบเขตของชื่อเฉพาะจำนวนหนึ่งเป็นคำสันธานนั้น ทั้งนี้เป็นเพราะคำสันธานจะทำหน้าที่เชื่อมความให้ต่อกัน ทำให้นอกจากจะทราบว่าประโยคสิ้นสุดลงแล้ว ชื่อเฉพาะที่มีตำแหน่งอยู่ท้ายประโยคก็จะพบจุดสิ้นสุดของขอบเขตชื่อเฉพาะด้วย เช่น

"พรรค<organization>ประชาธิปัตย์</organization>ส่วนกรณีอื่น"

จากการพิจารณาคลังข้อมูลและพบว่ากลุ่มคำซึ่งถูกใช้เป็นจุดสิ้นสุดของขอบเขตชื่อเฉพาะนี้มีความหลากหลายมาก โดยอาจจัดอยู่ในประเภทต่างๆ ตามที่ได้กล่าวไปแล้ว อีกทั้งยังมีจำนวนมากซึ่งไม่อาจคาดเดาจำนวนทั้งหมดได้ ทั้งนี้เป็นเพราะกลุ่มคำซึ่งถูกใช้เป็นจุดสิ้นสุดของขอบเขตชื่อเฉพาะนี้ อาจเป็นคำขยายคำนาม ซึ่งชื่อเฉพาะก็จัดว่าเป็นคำนามประเภทหนึ่ง หรืออาจเป็นการสิ้นสุดนามวลี แล้วเริ่มต้นวลีอื่นๆ ดังนั้นกลุ่มคำซึ่งถูกใช้เป็นจุดสิ้นสุดของขอบเขตชื่อเฉพาะนี้จึงมีได้มากมาย

เมื่อพิจารณาคลังข้อมูลใช้ฝึก จนได้รายการของกลุ่มคำบอกจุดสิ้นสุดมาแล้ว ดังนั้นรูปแบบของกฎจะเพิ่มส่วนของการหาจุดสิ้นสุดของชื่อเฉพาะเป็นดังนี้

"พบ [คำนำหน้าชื่อ] และคำที่นำหน้ามิได้อยู่ใน [รายการของคำยกเว้นหน้า] และคำที่ตามหลังมิได้อยู่ใน [รายการของคำยกเว้นหลัง] คำที่ตามหลังก็จะเป็น [ชื่อเฉพาะ] และมีจุดสิ้นสุดของขอบเขตชื่อเฉพาะเมื่อพบ [รายการของคำบอกจุดสิ้นสุดหรือรายการของคำนำหน้าชื่อ]"

จากกฎที่กล่าวไปข้างต้น เมื่อนำมาใช้กับชื่อเฉพาะประเภทชื่อคนนั้น กฎจะถูกลดให้เหลือเพียง

**กฎสำหรับชื่อบุคคล** : "พบ [คำนำหน้าชื่อ] คำที่ตามหลังก็จะเป็น [ชื่อเฉพาะ] และมีจุดสิ้นสุดของขอบเขตชื่อเฉพาะเมื่อพบ [รายการของคำบอกจุดสิ้นสุดหรือรายการของคำนำหน้าชื่อ]"

ทั้งนี้เป็นเพราะ [คำนำหน้าชื่อ] ของชื่อเฉพาะประเภทคนนั้นไม่พบว่าปรากฏนำหน้าคำทั่วไปอื่นๆ เช่น "นาย" จะไม่นำหน้าคำทั่วไปอย่าง "ฝ่าย" เพราะฉะนั้นจึงทำให้กฎเป็นไปตามที่กล่าวมาแล้ว

ส่วนชื่อองค์กรและชื่อสถานที่จะเป็นไปตามกฎนี้

**กฎสำหรับชื่อองค์กร / ชื่อสถานที่** : “พบ [คำนำหน้าชื่อ] และคำที่นำหน้ามิได้อยู่ใน [รายการของคำยกเว้นหน้า] และคำที่ตามหลังมิได้อยู่ใน [รายการของคำยกเว้นหลัง] คำที่ตามหลังก็จะเป็น [ชื่อเฉพาะ] และมีจุดสิ้นสุดของขอบเขตชื่อเฉพาะเมื่อพบ [รายการของคำบอกจุดสิ้นสุดหรือรายการของคำนำหน้าชื่อ]”

อย่างไรก็ตาม กฎสำหรับชื่อองค์กรและชื่อสถานที่อาจปรับไปตามสิ่งที่พบจากคลังข้อมูลใช้ฝึก โดยอาจมีหรือไม่มีส่วนของ [รายการของคำยกเว้นหน้า] ซึ่งนำหน้าคำนำหน้าชื่อ หรืออาจมีหรือไม่มีส่วนของ [รายการของคำยกเว้นหลัง] ที่ตามหลังคำนำหน้าชื่อ โดย [รายการของคำยกเว้นหน้า] และ [รายการของคำยกเว้นหลัง] ก็จะแตกต่างไปตามคำนำหน้าแต่ละคำที่ใช้ด้วย

ซึ่งรายการของคำยกเว้นหน้า คำยกเว้นหลัง และคำบอกจุดสิ้นสุดดังกล่าวจะแสดงอยู่ในภาคผนวก

จากกฎที่กล่าวไปข้างต้น จะช่วยให้ระบุตำแหน่งของชื่อเฉพาะได้ และจากคำนำหน้าชื่อทั้ง 3 ประเภทก็จะทำให้จำแนกประเภทของชื่อเฉพาะดังกล่าวได้ อย่างไรก็ตาม จากคลังข้อมูลใช้ฝึกยังพบว่ามีกรณีที่ชื่อเฉพาะนั้นๆ ไม่มีคำนำหน้าชื่อ ซึ่งจากการพิจารณาคลังข้อมูลใช้ฝึก พบว่า ชื่อเฉพาะประเภทชื่อสถานที่จะมีกรณีที่ชื่อเฉพาะไม่มีคำนำหน้าชื่อมากที่สุดจากชื่อเฉพาะทั้ง 3 ประเภท กล่าวคือกรณีที่ชื่อสถานที่ไม่มีคำนำหน้าชื่อจะมีถึง 28.779% โดยที่ชื่อคนจะไม่มีคำนำหน้าชื่อเพียง 0.423% และชื่อองค์กรจะไม่มีคำนำหน้าชื่อเพียง 2.545% ดังนั้นเพื่อเป็นการเพิ่มประสิทธิภาพของระบบที่ใช้กฎซึ่งอิงคำนำหน้าชื่อเป็นหลัก ผู้วิจัยจึงพิจารณาหาหลักฐานภายนอกซึ่งได้มาจากบริบทข้างเคียงเพื่อช่วยระบุจุดเริ่มต้นของขอบเขตของชื่อเฉพาะในกรณีที่ชื่อเฉพาะนั้นๆ ไม่มีคำนำหน้าชื่อเป็นหลักฐานภายในเพื่อบ่งบอกขอบเขต

ดังตัวอย่าง “ที่ดินย่าน<location>เกียกกาย</location>”

จะเห็นได้ว่า “เกียกกาย” ซึ่งเป็นชื่อเฉพาะประเภทสถานที่ไม่มีคำนำหน้าชื่อเป็นหลักฐานภายในเพื่อบ่งบอกจุดเริ่มต้นของชื่อเฉพาะ ดังนั้น การใช้หลักฐานภายนอกที่ได้จากบริบทข้างเคียงอย่างคำว่า “ย่าน” จะมีส่วนช่วยในการระบุตำแหน่งและจำแนกประเภทของชื่อเฉพาะ ซึ่งบริบทข้างเคียงดังกล่าวสามารถหาได้จากการใช้โปรแกรม Collocation Extract 3.06 ซึ่งสามารถดาวน์โหลดจากเว็บไซต์ http://pioneer.chula.ac.th/~awirote/colloc/index.html (Aroonmanakun, 2004) โดยวิธีทางสถิติที่เลือกใช้ในโปรแกรมคือวิธีการหาค่า Mutual Information (MI) เช่นเดียวกับวิธีทางสถิติที่เลือกใช้ในงานวิจัยนี้ เพื่อหาคำที่ปรากฏร่วมกับชื่อสถานที่มากเป็นพิเศษและนำมาประกอบการใช้กฎเพื่อทำให้ประสิทธิภาพของกฎเป็นไปได้ดีมากขึ้น และรายการคำที่หาได้จะมาจากบริบทข้างเคียงหรือ external evidence เพื่อนำไปใช้เป็นตัวกำหนดขอบเขตเริ่มต้นของชื่อเฉพาะประเภทสถานที่ ได้แก่คำว่า “ย่าน” , “ฝั่ง” , “เยือน” , “บริเวณ” , “ท้อง” , “ลาน” ซึ่ง

หลักฐานภายนอกที่ได้มาจากบริบทข้างเคียงเหล่านี้นอกจากจะช่วยลดจำนวนรายการของกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะ (candidate) ลงแล้ว ยังช่วยจำแนกประเภทของชื่อเฉพาะให้มีความแม่นยำมากขึ้นด้วย

ดังนั้น รูปแบบของกฎจะเป็นดังนี้

**กฎสำหรับชื่อสถานที่** : "พบ [คำนำหน้าชื่อหรือคำ 6 คำที่ได้จากหลักฐานภายนอก] และคำที่นำหน้ามิได้อยู่ใน [รายการของคำยกเว้นหน้า] และคำที่ตามหลังมิได้อยู่ใน [รายการของคำยกเว้นหลัง] คำที่ตามหลังก็จะเป็น [ชื่อเฉพาะ] และมีจุดสิ้นสุดของขอบเขตชื่อเฉพาะเมื่อพบ [รายการของคำบอกจุดสิ้นสุดหรือรายการของคำนำหน้าชื่อ]"

## 4.3 ผลการตัดสินและจำแนกประเภทชื่อเฉพาะ

ชื่อเฉพาะที่เลือกมาด้วยวิธีการทางสถิติดังที่ได้กล่าวไว้แล้วในบทที่ 3 จะถูกนำมาตัดสินด้วยโปรแกรมกฎที่เขียนขึ้นเพื่อระบุและจำแนกประเภทของชื่อเฉพาะตามที่ได้อธิบายในตอนที่แล้ว ผลทดสอบที่ได้จะแสดงในตารางที่ 5 โดยคอลัมน์ที่ 3 จะแสดงจำนวนชื่อที่โปรแกรมกฎตัดสินว่าเป็นชื่อเฉพาะ และในจำนวนนี้มีจำนวนหนึ่งที่โปรแกรมตัดสินได้ถูกต้องซึ่งแสดงไว้ในคอลัมน์ที่ 2 ส่วนคอลัมน์ที่ 4 เป็นจำนวนชื่อเฉพาะจริงที่พบในข้อมูล และในตารางจะแสดงผลแยกเป็นชื่อเฉพาะแต่ละประเภท เพราะกฎที่เขียนขึ้นนั้นเป็นกฎที่แยกตามประเภทของชื่อเฉพาะ

ตารางที่ 5 ผลการทดสอบค่าความแม่นยำ ค่าความครบถ้วน และค่า F เมื่อใช้โปรแกรมกฎ

| ประเภทของชื่อเฉพาะ | จำนวนชื่อเฉพาะที่ตัดสินได้ถูกต้อง | จำนวนชื่อเฉพาะทั้งหมดที่เลือกมา | จำนวนชื่อเฉพาะทั้งหมด | ค่าความแม่นยำ (%) | ค่าความครบถ้วน (%) | ค่า F |
|---|---|---|---|---|---|---|
| ชื่อคน | 3786 | 7009 | 3940 | 54.000 | 96.117 | 69.150 |
| ชื่อองค์กร | 2854 | 5996 | 3071 | 47.598 | 92.934 | 62.953 |
| ชื่อสถานที่ | 873 | 2757 | 1735 | 31.665 | 50.317 | 38.869 |

จากตารางที่ 5 พบว่าชื่อเฉพาะประเภทชื่อคนและชื่อองค์กรนั้น แม้ว่าจะมีอัตราการรู้จำหรือค่า F เพียง 69.150% และ 62.953% ตามลำดับ แต่ก็พบว่าการระบุตำแหน่งและการจำแนก

ประเภทของชื่อเฉพาะทั้ง 2 ประเภท ก็มีค่าความครบถ้วนมากกว่า 90% ซึ่งแสดงให้เห็นว่ากฎที่ใช้สามารถจำแนกประเภทของชื่อเฉพาะได้ดี แต่สำหรับชื่อเฉพาะประเภทชื่อสถานที่ที่มีจำนวนชื่อเฉพาะที่ไม่มีคำนำหน้าชื่อเป็นจำนวนมากที่สุดในชื่อเฉพาะ 3 ประเภทนั้น จะให้ค่าความครบถ้วนน้อยที่สุดด้วย แม้ว่าในกฎจะมีการใช้หลักฐานที่ได้จากบริบทข้างเคียงทั้ง 6 คำได้แก่คำว่า “ย่าน” , “ฝั่ง” , “เยือน” , “บริเวณ” , “ท้อง” , “ลาน” มาเสริมกับการใช้คำนำหน้าชื่อก็ทำให้ผลค่าอัตราการรู้จำมีค่าเพียง 38.869%

ตารางที่ 6 ผลการจำแนกประเภทของโปรแกรมกฎที่เขียนขึ้น

| ประเภทของชื่อเฉพาะ | ชื่อคน | ชื่อองค์กร | ชื่อสถานที่ |
|---|---|---|---|
| ชื่อคน | 3786 | 0 | 0 |
| ชื่อองค์กร | 0 | 2854 | 0 |
| ชื่อสถานที่ | 0 | 3 | 873 |

ในการใช้กฎเพื่อรู้จำและจำแนกประเภทชื่อเฉพาะนั้น อาจเป็นไปได้ว่าโปรแกรมกฎอาจตัดสินหรือจำแนกประเภทชื่อเฉพาะผิดไปก็ได้ ดังนั้นผู้วิจัยจึงนับและแสดงผลจำนวนการตัดสินที่ผิดพลาดไว้ในตารางที่ 6

จากข้อมูลในตาราง พบว่ากฎที่ใช้สามารถจำแนกประเภทของชื่อเฉพาะประเภทชื่อคนและชื่อองค์กรได้ถูกต้องโดยไม่พบการรู้จำและจำแนกชื่อเฉพาะผิดประเภท แต่สำหรับชื่อสถานที่นั้น พบว่ามีกรณีที่กฎจำแนกประเภทให้ชื่อสถานที่ผิดประเภทไปเป็นชื่อองค์กร โดยชื่อเฉพาะที่ถูกจำแนกผิดประเภทดังกล่าวพบว่าเป็นกรณีที่มีการใช้คำนำหน้าชื่อเฉพาะเป็นตัวบ่งบอก และคำนำหน้าชื่อนั้นมีความกำกวม คือสามารถนำหน้าชื่อองค์กรหรือชื่อสถานที่ก็ได้ (คำว่า “กระทรวง” และ “กรม”) ทั้งสามกรณีที่ตัดสินผิด ได้แก่

“ผู้สื่อข่าวรายงานจาก กระทรวงเกษตรและสหกรณ์ ว่าเมื่อวันที่13 มิ.ย.”

“เวลา09.00น.วันเดียวกันที่ กรมประชาสัมพันธ์”

“สำหรับสถานศึกษาที่ทดลองปลูกพืชจีเอ็มโอทั้ง 3 แห่งได้แก่ กรมวิชาการเกษตร”

ซึ่งสาเหตุที่โปรแกรมกฎรู้จำและจำแนกประเภทได้ไม่ผิดประเภทนั้น เป็นเพราะกฎที่เขียนขึ้นใช้ในงานนี้นั้น ได้ปรับให้เข้ากับข้อมูลเป็นอย่างดี จึงมีกรณีที่ผิดพลาดเพียงสามตัวอย่างดังที่แสดงไว้

## 4.4 อภิปรายผล

ผลที่ได้จากตารางที่ 5 แสดงให้เห็นว่าชื่อเฉพาะทั้งหมดที่เลือกมาด้วยวิธีการทางสถิติเพื่อให้โปรแกรมกฎตัดสินว่าเป็นชื่อเฉพาะนั้นมีจำนวนค่อนข้างมาก จึงทำให้ค่าความแม่นยำของการตัดสินในส่วนของการใช้กฎมีค่าน้อย ซึ่งทำให้เมื่อคำนวณอัตราการรู้จำหรือค่า F แล้ว ทำให้มีค่าต่ำด้วย ทั้งนี้อาจเป็นเพราะกฎที่เขียนขึ้นนั้นมีข้อจำกัดเรื่องการหาขอบเขตของชื่อเฉพาะ ถึงแม้ว่าการหาจุดเริ่มต้นของชื่อเฉพาะจะอาศัยรายการคำนำหน้าชื่อซึ่งจัดเป็นหลักฐานภายในที่น่าจะให้ผลดี แต่การระบุจุดสิ้นสุดของชื่อเฉพาะนั้นทำได้ไม่ดีนัก ทั้งนี้เพราะชื่อเฉพาะอาจทำหน้าที่เป็นประธานของประโยคหรือเป็นกรรมของประโยคก็ได้ ทำให้รายการของคำบอกจุดสิ้นสุดของขอบเขตชื่อเฉพาะจึงมีความหลากหลายมากตามไปด้วย ซึ่งนอกจากคำบอกจุดสิ้นสุดอาจเป็นคำนำหน้าชื่อแล้ว คำบอกจุดสิ้นสุดยังอาจจัดอยู่ในกลุ่มคำประเภทใดก็ได้ ได้แก่ คำนาม คำนำหน้ากริยา คำกริยา คำประพันธสรรพนาม คำวิเศษณ์ คำบุพบท และคำสันธาน โดยหากชื่อเฉพาะทำหน้าที่เป็นประธานของประโยคก็อาจจะทำให้คำบอกจุดสิ้นสุดของขอบเขตชื่อเฉพาะจัดอยู่ในหมวดคำกริยา หรือกรณีที่ชื่อเฉพาะทำหน้าที่เป็นกรรมของประโยค เมื่อจะกล่าวถึงประโยคถัดไป ก็อาจใช้คำสันธานมาเชื่อมต่อความได้ อีกทั้งชื่อเฉพาะยังอาจมีคำวิเศษณ์มาขยายความได้ด้วย

อีกทั้งการระบุจุดสิ้นสุดของชื่อเฉพาะ ยังอาจมีข้อจำกัด โดยคำบอกจุดสิ้นสุดอาจครอบคลุมกรณีบางกรณี เช่น เราอาจจะกำหนดให้คำว่า “ว่า” เป็นหนึ่งในรายการของคำที่เป็นจุดสิ้นสุดของขอบเขตของชื่อเฉพาะได้ แต่คำว่า “ว่า” จะครอบคลุมเพียงไม่กี่กรณี ซึ่งหากเกิดกรณีนอกเหนือจากนั้น ก็จะทำให้กฎนี้ใช้ไม่ได้ผลและทำให้มีกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะเพิ่มจำนวนขึ้นมาอีก เช่น

กระทรวง <candidate>การคลังก็รับ</candidate>ว่า

จะสังเกตได้ว่า จากคำนำหน้าชื่อซึ่งเป็นจุดเริ่มต้นคือคำว่า “กระทรวง” ผ่านเงื่อนไขของโปรแกรมเรียบร้อย แต่มาพบปัญหาอยู่ที่จุดสิ้นสุดของชื่อเฉพาะทำให้ได้ชื่อเฉพาะที่ไม่ถูกต้องปนเข้ามาอีกตัวหนึ่ง ยังผลให้ค่าความแม่นยำลดลงไปอีก

สำหรับชื่อเฉพาะประเภทสถานที่นั้น นอกจากกฎที่เขียนขึ้นเพื่อจัดการกับชื่อเฉพาะที่มีหลักฐานที่ได้จากบริบทภายในหรือที่มีรายการของคำนำหน้าชื่อแล้ว ก็ยังมีการเพิ่มกฎเพื่อใช้กับกรณีที่ชื่อเฉพาะปรากฏแบบไม่มีคำนำหน้าชื่อนำอยู่ด้านหน้าด้วย ดังนั้น ในกรณีแบบนี้ ผู้วิจัยใช้โปรแกรม Collocation Extract ในการหารายการของคำที่มีความสัมพันธ์กับชื่อเฉพาะประเภท

สถานที่ซึ่งรายการของคำเหล่านี้จัดเป็นหลักฐานที่หาได้จากบริบทข้างเคียง ซึ่งได้ออกมา 6 คำ ดังนี้คือ ย่าน , ฝั่ง , เยือน , บริเวณ , ท้อง , ลาน ซึ่งผลการทดสอบเมื่อรับเอาคำเหล่านี้เข้าไปใช้ในโปรแกรมกฎ พบว่าค่าความแม่นยำ ความครบถ้วน และอัตราการรู้จำหรือค่า F ก็ยังไม่สูงนัก

ส่วนผลของค่าความครบถ้วนนั้น พบว่ากฎที่เขียนขึ้นทำให้สามารถจำแนกประเภทชื่อเฉพาะประเภทชื่อคนกับชื่อองค์กรได้เกิน 90% แต่ให้ผลไม่ดีนักกับชื่อสถานที่ ทั้งนี้เป็นเพราะ ชื่อสถานที่ที่เข้ามาสู่ระบบกฎนั้น ส่วนใหญ่มักไม่มีคำนำหน้าชื่อ เช่น

“ในเขตที่ไม่มีส.ส.ของพรรคเช่น <location>ยานนาวา</location> <location>ดินแดง</location> <location>ห้วยขวาง</location> <location>บางคอแหลม</location> และ <location>สัมพันธวงศ์</location>”

ทำให้การดึงชื่อสถานที่ที่ถูกต้องออกมาโดยโปรแกรมกฎนั้นเป็นไปได้ยากและทำให้ได้ค่าความครบถ้วนและค่าความแม่นยำน้อย ยังผลให้อัตราในการรู้จำหรือค่า F ต่ำด้วย

สำหรับกรณีที่คำนำหน้าชื่อซึ่งถูกใช้เป็นจุดเริ่มต้นของขอบเขตชื่อเฉพาะมีความกำกวมในการจำแนกประเภทชื่อเฉพาะระหว่างชื่อองค์กรและชื่อสถานที่นั้น จากผลในตารางที่ 6 ทำให้ทราบว่าโปรแกรมกฎจำแนกประเภทชื่อเฉพาะที่มีคำนำหน้าชื่อซึ่งกำกวมระหว่างชื่อองค์กรและชื่อสถานที่ ได้แก่คำว่า “กรม” , “กระทรวง” ผิดประเภท ซึ่งชื่อเฉพาะดังกล่าวควรถูกจัดอยู่ในชื่อเฉพาะประเภทชื่อสถานที่ แต่โปรแกรมกฎจัดให้อยู่ในประเภทชื่อองค์กร โดยกรณีการจำแนกประเภทชื่อสถานที่ผิดเป็นชื่อองค์กรจะแสดงไว้อีกครั้งที่ด้านล่างนี้

“ผู้สื่อข่าวรายงานจาก กระทรวง<organization>เกษตรและสหกรณ์</organization> ว่าเมื่อวันที่13มิ.ย.”

“เวลา09.00น.วันเดียวกันที่ กรม<organization>ประชาสัมพันธ์</organization>”

“สำหรับสถานศึกษาที่ทดลองปลูกพืชจีเอ็มโอทั้ง 3 แห่งได้แก่ กรม<organization> วิชาการเกษตร</organization>”

ซึ่งจากบริบทข้างเคียงที่ปรากฏรอบชื่อเฉพาะนี้ พบว่าไม่มีคำที่สามารถดึงมาเพื่อใช้เป็นหลักฐานภายนอกเพื่อช่วยระบุประเภทของชื่อเฉพาะประเภทชื่อสถานที่ ดังเช่นคำว่า “บริเวณ” , “รอบ” , “พื้นที่” จึงทำให้กฎไม่สามารถรู้จำและจำแนกประเภทชื่อเฉพาะที่มีความกำกวมได้ แต่อย่างไรก็ตาม พบว่ากรณีการจำแนกประเภทผิดพลาดมีเพียง 3 ชื่อเฉพาะเท่านั้น ซึ่งคิดเป็น 0.17% จากชื่อเฉพาะทั้งหมด

และนอกจากนั้น กรณีดังกล่าวก็ยังพบปัญหาเรื่องการหาขอบเขตสิ้นสุดของชื่อเฉพาะเหมือนชื่อเฉพาะอื่นๆ ที่คำนำหน้าชื่อไม่ได้มีความกำกวม ซึ่งยังผลให้เมื่อคำนวณค่าความแม่นยำ

และค่าความครบถ้วนแล้ว ทำให้อัตราในการรู้จำหรือค่า F มีค่าต่ำกว่า 90% ซึ่งทำให้ระบบมีประสิทธิภาพไม่ดีเท่าที่ควร

# 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 หลักฐานภายนอกได้แก่คำว่า “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

# รายการอ้างอิง

ภาษาอังกฤษ


Arnold, D. Balkan, L. Humphreys, R. L. Meijer, S. and Sadler, L. 1994. Machine Translation: an Introductory Guide. London : Blackwells-NCC.

Aroonmanakun, W. 2002. Collocation and Thai Word Segmentation. In Thanaruk Theeramunkong and Virach Sornlertlamvanich, eds. Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop. Pathumthani: Sirindhorn International Institute of Technology. pages 68-75.

Bassano, J. C. Dias, G. Guillore, S. and Lopes, J.G.P. 2000. Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association ? In Proceedings of Recherche d'Informations Assistee par Ordinateur. Paris, France.

Bikel, D. Miller, S. Schwartz, R. and Weischedel, R. 1997. Nymble: a High-Performance Learning Name-finder. In proceedings of ANLP-1997. Washington, DC. pages 195-201.

Bender, O. and Och, F. and Ney, H. 2003. Maximum Entropy Models for Named Entity Recognition. In Proceedings of CoNLL-2003. Edmonton, Canada. pages 148-151.

Chanlekha, H. Kawtrakul, A. Mulasas, I. and Varasrai, P. 2002. Statistical and Heuristics Rule Based Model for Thai Named Entity Recognition. In Proceeding of SNLP-Oriental COCOSDA 2002.

Charoenpornsawat, P. Kijsirikul, B. and Meknavin, S. 1998. Feature-Based proper Name Identification in Thai. In Proceeding of the National Computer Science and Engineering Conference'98 (NCSEC'98). Bangkok, Thailand.

Charoenpornsawat, P. Kijsirikul, B. and Meknavin, S. 1998. Feature-Based Thai Unknown Word Boundary Identification Using Winnow. In Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems (APCCAS'98). Chieng Mai, Thailand.

Charoenpornsawat, P. Kijsirikul, B. and Meknavin, S. 1999. Comparing Winnow and RIPPER in Thai Named Entity Identification. In Proceedings of the Natural Language Processing Pacific Rim Symposium 1999(NLPRS'99). Beijing, China.

Cucerzan, S. and Yarowsky, D. 1999. Language Independent Named Entity Recognition Combining Morphological and contextual Evidence. In Proceedings of 1999 Joint SIGDAT Conference on EMNLP and VLC, University of Maryland, MD.

Dias, G. Guillore, S. Pereira Lopes, G. and Vintar, S. 2000. Normalising the IJS-ELAN Slovene-English Parallel Corpus for the Extraction of Multilingual Terminology. In Monachesi, P. (ed.) Proceedings of the CLIN'99 (Computational Linguistics in the Netherlands).

Demetriou, G. Humphreys, K. and Gaizauskas, R. 2000. Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures. [online]. Available from : http:// helix-web.stanford.edu/ psb00/humphreys.pdf

Erik F. Tjong Kim Sang. 2002. Memory-Based Named Entity Recognition. In Proceedings of CoNLL-2002. Taipei, Taiwan. pages 203-206.

Gaizauskas, R. 2000. Text Processing for Biologists [online]. Available from : bioinf2.leeds.ac.uk/www/presentations/TextAnalysis/TextAnalysis/index.htm

Gaizauskas, R. and Stevenson, M. 2000. Improving Named Entity Recognition using Annotated Corpora. LREC Workshop on "Information Extraction meets Corpus Linguistics". Athens, Greece.

Gallippi, A. 1996. Learning to Recognize Names Across Languages. In Proceedings of the Sixteenth International Conference on Computational Linguistics. Copenhagen, Denmark.

Grishman, R. Sekine, S. And Shinnou, H. 1998. A Decision Tree Method for Finding and Classifying Names in Japanese Texts. In Proceedings of the Sixth Workshop on Very Large Corpora.

Isozaki, H. 2001. Japanese Named Entity Recognition based on a Simple Rule Generator and Decision Tree Learning. Proc. of the 39th Annual Meeting of the Association for Computational Linguistics, pages 306--313.

Isozaki, H. and Kazawa, H. 2002. Efficient Support Vector Classifiers for Named Entity Recognition. In Proceedings of COLING-2002. pages 390-396.

Karkaletsis, V. Paliouras, G. Petasis, G. Spyropoulos, C. Vichot, F. and Wolinski, F. 2001. Using Machine Learning to Maintain Rule-based Named Entity Recognition and Classification systems. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

Kawtrakul, A. ,et.al. 1997. Automatic Thai unknown Word Recognition. NLPRS'97. Thailand.

Lopes, G. and Silva, J. 1999. A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora. In Proceedings of the 6th Meeting on the Mathematics of Language. pages 369-381. Orlando.

McDonald, D. 1996. Internal and External Evidence in the Identification and Semantic Catagorization of Proper names. In B.Boguraev and J.Pustejovsky (eds.), Corpus processing for lexical acquisition. pages 21-39.

Sassano, M. and Utsuro, T. 2000. Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition. In Proceedings of COLING 2000. pages 705-711.

Su, J. and Zhou, G. 2002. Named Entity Recognition using an HMM-based Chunk Tagger. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pages 473-480.

Turney, P. 2000. Learning Algorithms for Keyphrase Extraction. Information Retrieval 2(4): pages 303-336.

ภาษาไทย

อุปกิตศิลปสาร, พระยา. 2541. หลักภาษาไทย. พิมพ์ครั้งที่ 4. กรุงเทพมหานคร: ไทยวัฒนาพานิช.

อมรทิพย์ กวินปณิธาน. 2546. การศึกษาบริบทบ่งบอกชื่อเฉพาะในภาษาไทยตามแนวภาษาศาสตร์คอมพิวเตอร์. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

# ภาคผนวก

# ภาคผนวก

จากกฎ “พบ [คำนำหน้าชื่อ] และคำที่นำหน้ามิได้อยู่ใน [รายการของคำยกเว้นหน้า] และคำที่ตามหลังมิได้อยู่ใน [รายการของคำยกเว้นหลัง] คำที่ตามหลังก็จะเป็น [ชื่อเฉพาะ] และมีจุดสิ้นสุดของขอบเขตชื่อเฉพาะเมื่อพบ [รายการของคำบอกจุดสิ้นสุดหรือรายการของคำนำหน้าชื่อ]”

รายการดังต่อไปนี้จะแสดง [รายการของคำยกเว้นหน้าและคำยกเว้นหลัง] ที่นำหน้าหรือตามหลังคำนำหน้าชื่อแล้วสิ่งที่ตามหลังคำนำหน้าชื่อนั้นจะไม่เป็นชื่อเฉพาะประเภทชื่อองค์กร

| คำนำหน้าชื่อ | รายการของคำยกเว้นหน้า | รายการของคำยกเว้นหลัง |
|---|---|---|
| พรรค | "ผอ." , "ผอง" , "ประธาน" , "หัวหน้า" , "โฆษก" , "งานใหญ่" , "สาขา" , "ทำงาน" , "อธิบดี" , "แกนนำ" , "หัวโจก" , "หัวโต๊ะ" , "สมัชชา" , "รองหัวหน้า" , "ที่ทำการ" , "สำนักงาน" , "ผู้บริหาร" , "ผู้ก่อตั้ง" , "ประสานงาน" , "วิสามัญ" , "ประธานส.ส." , "เหรัญญิก" , "การบริหาร" , "อำนวยการ" , "การเลือกตั้ง" , "ที่ปรึกษา" , "เลขาธิการ" , "ด้านกฎหมาย" | "ว่า" , "การ" , "บ้าง" , "พอ" , "ให้" , "ขนาด" , "ทาง" , "หรือ" , "รวม" , "คิด" , "นำ" , "คง" , "ไหน" , "ใหม่" , "โฮล" , "ใหญ่" , "ซึ่ง" , "นั้น" , "รวบ" , "อื่น" , "และ" , "โดย" , "ได้" , "เป็น" , "ที่" , "พวก" , "ร่วม" , "ยืน" , "ดัง" , "ไว้" , "อย่าง" , "จำ" , "จะ" , "ออก" , "ไม่" , "อีก" , "มา" , "ส่วน" , "ต้อง" , "เข้า" , "ทำ" , "ลา" , "กว่า" , "แล้ว" , "อยู่" , "บาง" , "ตน" , "ขึ้น" , "ตาม" , "หรือ" , "เช่น" , "เดียว" , "ใด" , "ใน" , "นี้" , "มอง" , "ตรวจ" , "เรื่อย" , "เสร็จ" , "ก็" , "หลาย" , "แสดง" , "คือ" , "เหลือ" , "รับ" , "มี" , "ยัง" , "เล็ก" , "วัน" , "เดิม" , "ทุก" , "ทั้ง" , "กำ" , "แห่ง" , "แบ่ง" , "เก่า" , "นาย" , "ไป" , "นอก" , "ต่ำ" , "ยิ่ง" , "หวัง" , "อาจ" , "ต่อ" , "ขอ" , "ทราบ" , "กรรม" , "สัญญา" , "ซ้อน" , "เหมือน" , "ตั้ง" , "ด้วย" , "เตรียม" , "ตาย" , "แต่" , "เพื่อ" , "หา" , "อึด" , "เสมอ" , "ก่อน" , "ของ" , "อย่า" , |

| | | |
|---|---|---|
| | | "จริง" , "มิ" , "อัน" , "ดู" , "เท่า" , "รอง" , "แตก" , "วาง" , "ลำ" , "แทน" , "กับ" , "กล่าว" , "ขณะ" , "อยาก" , "สอบ" , "ต่าง" , "เพิ่ม" , "ย้าย" , "ปัจจุบัน" , "รู้" , "เพราะ" , "ภาย" , "ถ้า" , "ลด" , "ยอม" , "มาก" , "จาก" , "หนึ่ง" , "เลือก" , "ถือ" , "ส่ง" , "พ่าย" , "จึง" , "เมื่อ" , "คาด" , "กลับ" , "ทุจริต" , "กังวล" , "ควร" , "เปลี่ยน" , "พบ" , "พรรค" , "ถูก" , "ผู้" , "สูง" , "กฎ" , "ยึด" , "ตก" , "เคย" , "ณ" , "ราว" , "รอบ" , "ซื่อ" , "หาก" , "มอบ" , "ถึง" , "ผิด" , "เรียก" , "แกน" , "ชัด" , "สำ" , "ตลอด" , "เริ่ม" , "เปิด" , "ติด" , "ดี" , "ผ่าน" , "รัฐ" , "ย่อม" , "สุด" , "พันธุ์" , "เสีย" , "ฉาบ" , "เกิด" , "เฉพาะ" , "เดิน" , "เกี่ยว" , "แน่" , "ลง" , "คน" , "เพียง" , "ตัว" , "น่า" , "แข่ง" , "เรา" , "ชนะ" , "เล่น" , "ฝ่าย" , "ทั่ว" , "ปรับ" , "แก่" , "คู่" , "เอา" , "ชี้" , "ตัด" , "แม้" , "ภาค" , "ช่วย" , "เสื้อ" , "ปลด" , "รูป" , "เน้น" , "หนุน" , "ชุด" , "โดน" , "สี" , "มั่น" , "คอย" , "จัด" , "ชู" , "สรุป" , "ฟัง" , "เห็น" , "ปรากฏ" , "สร้าง" , "แรก" , "แยก" , "ค่อน" , "กลัว" , "ประมาณ" , "ประเมิน" , "ประสบ" , "ประกอบ" , "ต้นสังกัด" , "ส.ส." , "กรณี" , "อาทิ" , "ประกาศ" , "ประชุม" |
| กรม | "อธิบดี" | "เขต" , "ส.ส." , "ต้อง" , "วัด" , "ว่า" , "ไป" , "หลวง" , "ผู้" , "ต่าง" |
| องค์การ | "ผอ." , "ห้องสมุด" | - |

| | | |
|---|---|---|
| กระทรวง | “ตรวจราชการ” , "ปลัด" , "รัฐมนตรี" , "เจ้า" , "ที่" , "รมว." , "โฆษก" , "ข่าว" | "เร่ง" , "ไหน" , "ข้าม" , "ให้" , "ไม่" , "ล่า" , "สั่ง" , "จะ" , "ทบ" , "ต่าง" , "ทั้ง" , "ภาย" , "วาง" , "เป็น" , "ที่" , "ได้" , "และ" , "เชื่อ" , "กรม" , "ตาม" |
| สำนักงาน | "มติ" , "ต่อ" , "เปิด" , "งาน" , "หน้า" , "มาที่" , "ไปที่" , "ประจำ" , "ดูแล" , "เลขาธิการ" , "โฆษก" , "ผู้อำนวยการ" | "ใหญ่" , "และ" , "พรรค" , "ปลัด" , "ส่ง" , "ให้" , "เกษตร" , "เขต" |
| สำนัก | "ประจำ" , "ผอ." , "ปลัด" , "โฆษก" | "งาน" , "ป้องกัน" , "พัฒนา" , "กฎ" , "ออก" , "มี" , "ระบุ" , "ที่" |
| กองทัพ | "โฆษก" , "การ" | "อย่าง" , "ไม่" , "เสีย" , "ดู" , "ไทย" , "เป็น" , "ไป" |
| บริษัท | “พื้นที่” , "เจ้าหน้าที่" , "บริหาร" , "ใหญ่" , "ประธาน" , "การ" , "รอบ" | "อื่น" , "เจ๊ง" , "นี้" , "กับ" , "เลียน" , "ครอบครัว" , "ก็" , "จึง" , "ยื่น" , "เดียว" , "ดัง" , "พรรค" , "ต่าง" , "ขาย" , "เมื่อ" , "เล็ก" , "พ่อ" , "ด้วย" , "ยักษ์" , "ใหญ่", "หนึ่ง" , "ใน" , "ใด", "อย่าง" , "ที่", "เสนอ" , "ร้อง" , "ไหน" , "เกี่ยวกับ" , "เอกชน" , "ทั้ง" , "ตน" , "เครือ" |
| ธนาคาร | "ที่" | "แทน" , "ของ" , "ได้" , "แห่ง" , "หมู่" , "หรือ" , "ทาง" , "ต่าง" , "ไป" |
| องค์กร | "สัมพันธ์" , "ประธาน" , "ประชุม" , "นำ" | "เป็น" , "เอกชน" , "ต่าง" , "และ" , "ปกครอง" , "ส่วน" , "ที่" , "เครือ" , "มาก" , "อีก" , "อิสระ" , "ออก" , "ชุม" , "ภาค" , "ถ้า" , "ตาม" , "แต่" , "ใหญ่" , "ทุก" , "เข้า" , "ของ", "อื่น" , "ทั้ง" , "เหล่า" , "นั้น" , "ก็" , "จะ" , "ฝ่าย" , "มี" |
| กอง | - | "ทัพ" , "เกียรติ" , "ทุน" , "ทอง" , "โดย" , "โชว์" , "เมื่อ" , "เพราะ" , "นั้น" , "บิน" , "รับ" , "เพื่อ" , "แผน" , "มี" , "กำลัง" |

| | | |
|---|---|---|
| โรงเรียน | "การ" , "ผอ." , "ที่" , "เช่น" , "ทุก" , "จาก" | "ต่าง" , "อาจ" , "และ" , "เมือง" , "โครง" , "เดียว" , "ใน" , "ต้อง" , "ไหน" , "ทั้ง" , "กอบ" , "ย่าน" , "แถว" , "ดัง" , "ถูก", "ที่" , "แห่ง" |
| โรงพยาบาล | "ชาว" , "ของ" , "ที่" , "เข้า" , "การ" | "คือ" , "เอกชน" , "ถูก" , "เลย" , "ช้า" |
| วิทยาลัย | - | - |
| บรรษัท | - | - |
| สหภาพ | - | - |
| สภา | - | - |
| มูลนิธิ | - | "พึง" , "โดย" , "และ" , "มี" , "ยัง" , "พัฒนา" |
| ศาล | "ไพ" , "มหา" , "พิ" , "การ" , "ประธาน" , "งาน" , "อำนาจ" , | "เฉพาะ" , "พระภูมิ" , "เป็น" , "ออก" , "ต่อ" , "ดี" , "ราย" , "เลือก" , "พิ" , "สุด" , "อีก" , "ชั้น" , "เท่า" , "ก็" , "ผู้" , "โดย" , "เมื่อ" , "หาก" , "ส่วน" , "ทั้ง" , "สั่ง" , "จะ" , "ว่า" , "วัน" , "เพราะ" , "กล่าว" , "ตัด" , "เก่า" , "เจ้า" , "หลัก" , "การ" |
| สถาบัน | "ผอ." , "รวม" , "กับ" , "การ" , "ละเอียด" , "เป็น" , "ใช่" , "ของ" , "เชิญ" | "เดียว" , "เก่า" , "ครอบครัว" , "ที่" , "พัฒนา" , "ศาล" , "และ" , "จัด" , "พล" |
| ศูนย์ | เยี่ยม | - |
| มหาวิทยาลัย | "อาจารย์" , "ที่" , "อธิบดี" | "เกิด" , "รัฐ" , "หรือ" , "เอกชน" |

รายการดังต่อไปนี้จะแสดง [รายการของคำยกเว้นหน้าและคำยกเว้นหลัง] ที่นำหน้าหรือตามหลังคำนำหน้าชื่อแล้วสิ่งที่ตามหลังคำนำหน้าชื่อนั้นจะไม่เป็นชื่อเฉพาะประเภทชื่อสถานที่

| คำนำหน้าชื่อ | รายการของคำยกเว้นหน้า | รายการของคำยกเว้นหลัง |
|---|---|---|
| ทวีป | - | - |
| ร้านกาแฟ | - | - |
| ร้านอาหาร | - | - |
| สวนสัตว์ | - | - |

| | | |
|---|---|---|
| ค่าย | - | "จะ" , "ที่" , "ใน" |
| สถานทูต | - | "เพื่อ" |
| สนามกอล์ฟ | - | - |
| ท่าอากาศยาน | - | - |
| สุเหร่า | - | “ย่าน” |
| บึง | - | - |
| คุก | “จำ” | "กัน" , "คาม" , "แน่" , "ติด" , "แทน" , "ให้" , "เพื่อ" , "ไป" , "ข้าง" , "ตั้ง" |
| มัสยิด | - | "เพื่อ" |
| สวนสาธารณะ | - | "และ" |
| ร.ร. | - | - |
| พระธาตุ | - | "ที่" , "มี" , "วัด" , "ทั้ง" |
| เกาะ | - | “ติด" , "ดัง" |
| สโมสร | - | - |
| สมาคม | "ยก" , "แทน" | - |
| สาธารณรัฐ | "ทูต" , "รัฐมนตรี" , "ประธานาธิบดี" | - |
| ราชอาณาจักร | "แห่ง" , "กัน" , "ให้" | - |
| อ. | "พ.ต." , "พล." , "พ." , "พ.อ." , "ร." , "การ" , "ชาว" , "บริษัท" , "อดีต" | “ต.” |
| จังหวัด | "สภา" , "การ" , "หลาย" , "ทั้ง" , "สาร" , "ค้า" , "ทุก" , "ต่าง" , "เที่ยว" , "คณะ" , "ปลัด" | "ประกอบ" , "นำ" , "ภาค" , "จะ" , "ชายแดน" , "ใกล้" , "และ" , "ที่" , "อื่น" , "เดียว" , "โดย" , "ทำ" , "รวม" , "ไม่" , "เสนอ" , "นั้น" , "มูล" , "ใด" , "หนึ่ง" , "เพื่อ" , "เพราะ" , "ต่าง" , "เพิ่ม" , "เหล่า" , "สุด" , "แก้" , "แล้ว" , "เป้า" , "แรก" , |

| | | |
|---|---|---|
| | | "ใน" , "แบบ" , "ด้วย" , "ใหญ่" , "งาน" , "ผ่าน" , "คือ" , "แต่" , "ล่วง" , "มี" , "จาก" , "เร่ง" , "ต่อ" , "ทุก" , "ทาง" , "นี้" , "ของ" , "แล้ว" , "อีก" , "อย่าง" , "ทำ" , "ได้" , "อยาก" , "ไหน" , "ละ" , "นำ" , "ตาม" , "ผลิต" , "กลับ" , "ขอ" , "เท่า" , "ทั้ง" , "ถูก" , "ท่อง" , "ว่า" , "กำลัง" , "ยาก" , "ต้น" , "ก็" , "ก่อน" , "ล้าน" |
| แม่น้ำ | "ข้าม" , "แก่ง" | "ใน" , "ส่ง" |
| ลุ่มน้ำ | "ที่" , "นา" | "เช่น" , "ใน" |
| เขา | "วัด" , "อ." , "หุบ" | "ไป" , "ก็" , "ด้วย" , "กลัว" , "ปรารถนา" , "เข้า" , "เอง" , "ที่" , "จะ" , "หลัก" , "บอก" , "เป็น" , "คง" , "มา" , "ไว้" , "ลง" , "ต้อง" , "ทำ" , "ไม่" , "เคย" , "ได้" , "อยาก" , "สรุป" , "แถลง" , "ทราบ" , "ก่อน" , "เจ็บ" , "แล้ว" , "ควร" , "รู้" , "ตาม" , "จูง" , "เมื่อ" , "มัน" , "อาจ" , "หรือ" , "เลย" , "ให้" , "พอ" , "แต่" , "ยัง" , "ซึ่ง" , "มี" , "ตก" |
| อ่างเก็บน้ำ | - | "เพื่อ" |
| ท่าน้ำ | - | - |
| ท่าเรือ | - | "กัน" , "และ" , "เป็น" , "แต่" , "น้ำ" |
| ท่าเรือน้ำลึก | - | - |

| | | |
|---|---|---|
| ปราสาท | - | “และ” |
| ปราสาทหิน | - | - |
| คลอง | "ปาก" , "น้ำ" , "เขต" , "เสีย" , "ต." , "ทุจริต" , "เขื่อน" , "เข้า" | "เหม็น" , "ธรรม" , "ด่าน" , "ทุก" , "จึง" |
| วัด | "จังหวัด" , "ตรวจ" , "เจ้าอาวาส" , "เรียน" | "ความ" , "ที่" , "ตลอด" , "ผล" , "ส่วน" , "หลัง" , "ต่าง" , "ละ" , "คือ" , "อื่น" , "เงิน" , "เปิด" |
| วัง | "กลุ่ม" , "ระ" , "อ." , "เขต" , "มี" | "บัว" , "พญา" , "วน" , "นั้น" , "ต้อง" , "นัด" , "ยม" , "ชา" |
| รพ. | "ประจำ" , "ผอ." | - |
| โรงพยาบาล | "การ" | "จำ" , "คือ" , "เอกชน" , "ถูก" , "เลย" , "โดย" , "ช้า" |
| อนุสาวรีย์ | - | - |
| พระบรมราชานุสาวรีย์ | - | - |
| เขื่อน | - | "ป้องกัน" , "ฝั่ง" , "ทด" , "ไว้" , "ที่" , "จะ" , "ใน" |
| มณฑล | "พุทธ" , "นาย" , "ปริ" , "ละ" | "และ" , "มี" |
| นคร | "จ." , "มหา" , "ผวจ." , "กลุ่ม" , "จังหวัด" , "พระ" , "ลง" , "เมือง" , "นคร" , "สกล" , "นาย" | "โดย" , "นคร" , "พรรค" , "เรา" |
| ต. | "พ." , "พล." , "ร." , "พ.ต." , "จ.ส." , "น." | "ท." , "ค." , "อ." |
| อาคาร | - | "โรงเรียน" , "อเนก" , "เทียบ" , "ที่" , "หลัง" , "เด็ก" , "ชั่ว" , "ชั้น" , "เรียน" , "จอด" |
| ตึก | - | "นั้น" |
| โรงแรม | - | "เดิม" , "แต่" , "จาก" , "อย่าง" , "ห้อง" , "และ" , "ทุก" , "เพื่อ" |
| สนามบิน | - | "ที่" , "ได้" , "โดย" , "ก็" , "และ" , "ใคร" |

| | | |
|---|---|---|
| ประเทศ | "แห่ง" , "ต่าง" , "ทั้ง" , "ทั่ว" , "ระหว่าง" , "ระดับ" , "ศรี" , "บาง", "ทุก" , "นานา" , "นำ" , "หลาย" , "สอง" | "ลด" , "เพื่อ" , "ถูก" , "หลาย" , "เรา" , "อยู่" , "อื่น" , "ที่" , "ใน" , "ใหญ่" , "ชาติ" , "อย่าง" , "มหาอำนาจ" , "เพื่อนบ้าน" , "เช่น" , "เป็น" , "ก็" , "จะ" , "ส่วน" , "โดย" , "ต่าง" , "เหล่า" , "นั้น" , "แต่" , "มาก" , "เดียว" , "ไม่" , "มา" , "นี้" , "ก่อน" , "แล้ว" , "เนื่อง" , "พื้น" , "ดัง" , "ยิง" , "และ" , "เข้า" , "เมื่อ" , "ต้อง" , "ฝาก" , "นั่น" , "ควร" , "ยั่ง" , "ได้" , "มี" , "ทำ" , "ตาม" , "เดิน" , "ทั้ง" , "ไป" , "สร้าง" , "อนาคต" , "เพิ่ม" , "ขนาด" , "ต่อ" , "พัฒนา" , "ว่า" , "คือ" , "ผู้" , "ของ" , "ทั่ว" , "เข้ม" , "ดี" , "จุ่ม" , "ให้" |
| แขวง | - | - |
| หมู่บ้าน | "กองทุน" , "ทุก" , "ธนาคาร" | "ถือ" , "ทั่ว" , "ไหน" , "ได้" , "ชุมชน" , "และ" , "ขนาด" , "ที่" , "แบ่ง" , "ภาย" , "ต้น" , "ใน" , "ทุก" , "ต่าง" |
| สนามม้า | - | - |
| แฟลต | - | - |
| ซอย | "ซอก" | "และ" |
| ร้าน | - | "ค้า" , "น้อง" , "ไว้" , "เลี้ยง" , "เต้น" |
| เคหะชุมชน | - | - |
| โรงงาน | - | "ผู้" , "เดียว" , "ผลิต" |
| สวน | "สอบ" , "สืบ" , "ไต่" , "วัน" , | "ทาง" , "ครัว" , "ป่า" , "ส้ม" , |

| | | |
|---|---|---|
| | "เดิน" , "ซอย" , "ชาว" | "ทัน" |
| บ้าน | "ชาว" , "เพื่อน" , "ทุบ" , "ยิง" , "วัด" , "การ" , "แม่" , "แขวง" , "ถู" , "เปิด" , "เช่า" , "ซื้อ" , "สร้าง" , "หน้า" , "ทุก" , "หญ้า" , "ใหญ่" , "หมู่" , "ร." , "แรม" , "ปา" , "คา" , "เผา" , "เรียน" , "จำ" | "แถลง" , "ของ" , "สนาม" , "พิษ" , "เมือง" , "พัก" , "ย่าน" , "ซอย" , "ตน" , "เกิด" , "สวน" , "ถ." , "ริม" , "เดียว" , "นาย" , "ต้อง" , "ให้" , "อยู่" , "แล้ว" , "นาง" , "เดิม" , "คน" , "เพื่อ" , "เรา" , "ไม่" , "เลย" , "ใน" , "ช่วย" , "นอก" , "ได้" , "โดย" , "ตาม" , "ปัญหา" , "กำลัง" |
| โรงเรียน | "ครู" , "ผอ." , "น้อง" , "ทุก" , "สร้าง" , "อาคาร" | "ต่าง" , "อาจ" , "และ" , "โครง" , "เดียว" , "ใน" , "ไหน" , "ทั้ง" , "ย่าน" , "แถว" , "ดัง" , "ถูก" , "แห่ง" |
| ธนาคาร | "ของ" , "การ" , "ผ่าน" , "เรื่อง" | "แทน" , "ของ" , "ได้" , "เพื่อ" , "แห่ง" , "โลก" , "หรือ" , "ทาง" , "ต่าง" , "ไป" |
| ศาลเจ้า | - | - |
| สน. | - | - |
| อุทยาน | - | "กับ" |
| ถ. | - | - |
| เขตเทศบาลตำบล | - | - |
| ศูนย์ประชุม | - | - |
| หอประชุมอำเภอ | - | - |
| ทางด่วนสาย | - | - |
| ถนนสาย | - | - |
| สำนักงานสาขา | - | - |
| หอประชุม | - | - |
| ที่ว่าการ | - | - |
| มูลนิธิ | "อาคาร" , "การ" , "เรียน" | "โดย" , "และ" , "มี" |

| | | |
|---|---|---|
| ภูมิภาค | - | "เดียว" , "ที่" , "ให้" , "เกิด" , "จึง" , "อื่น" |
| สถานี | "บอร์ด" , "การ" | "รถ" , "โทรทัศน์" , "วิทยุ" , "แล้ว" , "ยก" , "ทด" , "ตำรวจ" |
| ถนน | ท้อง | "เข้า" , "เชื่อม" , "ปลอด" , "แค่" , "สี่" , "ช่วง" , "ลาด" , "ที่" , "เส้น" , "การ" , "จะ" , "ซึ่ง" , "สาย" , "คน" , "คู่" , "ผ่าน" , "ใหญ่" , "จาก" , "หรือ" |
| ที่ทำการสาขา | "ป้าย" , "โดย" | - |
| ห้อง | - | "สมุด" , "ชั้น" , "น้ำ" , "ทำ" , "นั้น" , "แอร์" , "สน" , "ผู้" , "เพื่อ" , "รับ" , "พัก" |
| สถาบัน | "และ" , "แยก" , "เป็น" , "การ" , "ผอ." , "หรือ" , "ใน" | "การ" , "เก่า" , "ศาล" , "และ" , "มาก" , "จัด" |
| อำเภอ | "นาย" , "การ" , "ทุก" , "ทั้ง" , "สอง" , "กิ่ง" | "รอบ" , "และ" , "เมื่อ" , "เมือง" , "ใจ" , "เพื่อ" , "ต่าง" , "อื่น" , "ทั้ง" , "ละ" , "โดย |
| หอ | "งาน" , "การ" | "ต้อง" , "พัก" |
| สำนักงาน | "ต่อ" , "กับ" , "มติ" , "กลาง" , "และ" , "การ" , "หมาย" , "คือ" , "ตั้ง" , "ทรัพย์" , "หา" , "เป็น" , "สร้าง" , "ทาง" | "ใหญ่" , "สาขา" , "และ" , "กฎ" , "ตรวจ" , "ปลัด" , "สลาก" , "ที่" , "ส่ง" , "ศาล" , "ให้" |
| ที่ทำการ | "ย้าย" , "หน้า" , "เข้า" , "ว่า" | "สาขา" , "พรรค" |
| ภาค | "ภูธร" , "ภูมิ" , "ละ" , "ทัพ" , "ประธาน" , "ยก" , "มห" , "เสมอ" , "สตรี" | "ภูมิ" , "รวม" , "พื้น" , "ต่าง" , "ที่" , "รัฐ" , "เอกชน" , "อื่น" , "ยก" , "แรก" , "สอง" , "อุต" , "ละ" , "ต้อง" , "สนาม" , "การ" , "เกษตร" , "ถือ" , "ทั้ง" , "เรียน" , "และ" , "รอง" , |

| | | |
|---|---|---|
| | | "ไหน" , "เศรษฐกิจ" , "หาก" |
| ตลาดสดเทศบาล | - | - |
| ตลาดสด | - | - |
| ตลาด | "การ" , "เดิน" , "คลอง" , "ยาง" , "เจาะ" | "โลก" , "ใหญ่" , "หรือ" , "เชิง" , "การ" , "ใหม่" , "ต่าง" , "ด้วย" , "เงิน" , "แต่" , "ธุรกิจ" , "มาก" , "หลัก" , "พร้อม" , "เขต" , "จึง" , "หุ้น" , "ทุน" , "เหล็ก" |
| สนามกีฬาเทศบาล | - | - |
| สนามกีฬากลาง | - | - |
| สนามกีฬา | - | "เมือง" , "หน้า" |
| สนาม | "บ้าน" , "ภาค" | "เลือก , "หน้า" , "ใหญ่" , "การ" , "เสือ" , "รบ" , "ม้า" , "จัด" , "หญ้า" , "ใน" |
| ศาลา | "หน้า" , "ท่า" | "ยา" , "ไทย" |
| จ. | "ส." , "ชาว" | "ส.ต." , "ใน" , "จัด" , "จด" , "แจก" , "คือ" , "ปลาย" , "ที่" |
| เมือง | "คน" , "เต็ม" , "ย่าง" , "คุณ" , "ชน" , "ทอง" , "ดวง" , "หลัก" , "ตลาด" , "การ" , "นา" , "ศรี" , "ชุม" , "บ้าน" , "ชาน" , "อ." , "ตี" , "พื้น" , "พล" , "ผัง" , "เรียน" | "ศูนย์" , "ทั่ว" , "ที่" , "นั้น" , "เซ" , "นาย" , "หลัก" , "ใหญ่" , "หนาว" , "ก็" , "จะ" , "เหมือน" , "ไป" , "อ." , "จ." , "การ" , "น่า" , "เดิน" , "ก่อน" , "ทุก" , "พร้อม" , "หรือ" , "กรุง" , "ถ้า" , "ใน" , "สวย" , "แห่ง" , "ลิง" , "คน" , "จุด" , "ทั้ง" |
| กรุง | - | "แยก" , "ไทย" , "ด้วย" |
| เขต | "บาง" , "ทุก" , "ขอบ" , "ยา" , "หลาย" , "เสถียร" , "ไทย" , "พร" , "กว่า" , "นำ" , "กี่" , "ต." , "ผอ." , "คา" | "ภาค", "การ", "ใน" , "จะ" , "ที่" , "นั้น" , "พื้น" , "ขึ้น" , "เท่า" , "เลือก" , "ต้อง" , "ชาน" , "ซึ่ง" , "ก็", "เทศ" , "เดิม" , "ใคร" , |

| | | |
|---|---|---|
| | | "มัน" , "จาก" , "นาย" , "ของ" , "ส่วน", "เดียว", "ได้" , "ว่า" , "และ" , "แต่" , "แทน" , "พรรค" , "อ." , "จึง" , "หา" , "เช่น" , "ด้วย" , "อยู่" , "เพราะ" , "อื่น" , "ตัว" , "โดย" , "ผู้" , "ไม่" , "ป่า" , "อัน" , "แรก" , "นี้" , "คือ" , "ทำ" , "เข้า" , "ยัง" , "มี" , "ใด" , "ภาย" , "แล้ว" , "มา" , "นำ" , "ดัง" , "อีก" , "ชั้น" , "ต่าง" , "ทั่ว" , "ละ" , "เก็บ" , "ไหน" , "นอก" , "หลัก" , "ตรง" , "หวัง" , "คาด" , "แบ่ง" , "มาก" , "ส่วน" , "ขาด" , "กลาง" , "ขณะ" , "เป็น" , "แบบ" , "เพื่อ" , "จาก" , "เอา" , "ชุมชน" , "ต่อ" , "หนึ่ง" , "รับ" , "ยึด" , "เป้า" , "เสร็จ" , "ล้ม" , "อื่น" , "รวม" , "ขอ" , "ชลประทาน" , "ปริมณฑล" , "แจ้ง" |
| ย่าน | - | "ชุม" , "ตัว" , "ถนน" |
| ฝั่ง | - | "และ" , "ไม่" , "ตรง" , "ที่" , "ไหน" , "จ." |
| เยือน | - | "ภาย" , "ประ" , "คาด" , "ครั้ง" , "ให้" , "มา" , "โดย" , "พื้น" , "แล้ว" |
| บริเวณ | - | "ตลาด" , "หน้า" , "ชุม" , "ลุ่ม" , "เชิง" , "ชั้น" , "ท้อง" , "ดัง" , "เต็นท์" , "รอบ" , "ใกล้" , "โดย" , "ลาน" , |

| | | |
|---|---|---|
| | | "แยก" , "ชาย" , "ที่" , "คลัง" , "สนาม" , "บึง" , "สี่" , "ฝั่ง" , "รัฐ" , "ข้าง" , "ติด" , "ไหน" , "งาน" , "พื้น" |
| ท้อง | - | "ถิ่น" , "ตลาด" , "ถนน" , "ที่" |
| ลาน | - | "จอด" , "หน้า" , "สนาม" |

หมายเหตุ : [-] ใน ช่องรายการของคำยกเว้นหน้าและคำยกเว้นหลังทั้งสองช่อง หมายความว่า คำนำหน้าชื่อนั้นๆ สามารถใช้บ่งบอกชื่อเฉพาะได้เลยโดยจุดสิ้นสุดก็จะดูจากรายการคำบอกจุดสิ้นสุดที่แสดงด้านล่าง

รายการดังต่อไปนี้จะแสดง [รายการของคำบอกจุดสิ้นสุด] ที่จะเป็นจุดสิ้นสุดของขอบเขตชื่อเฉพาะประเภทชื่อคน

"แล้ว" , "กล่าว" , "เปิด" , "นาย" , "เคย" , "ใน" , "มอบ" , "แจ้ง" , "ว่า" , "ที่" , "แต่" , "และ" , "หรือ" , "จะ" , "นั้น" , "อีก" , "อยาก" , "สู้" , "เป็น" , "ให้" , "ตก" , "คง" , "ถึง" , "ถ้า" , "ตั้ง" , "พร้อม" , "ด้วย" , "ไม่" , "ได้" , "บ้าง" , "เอง" , "ยืน" , "ห่ม" , "โกง" , "เชื่อ" , "ตลอด" , "จึง" , "บอก" , "พูด" , "มี" , "ลง" , "เห็น" , "รวม" , "อยู่" , "ตรวจ" , "เข้า" , "แยก" , "มาก" , "ของ" , "กลุ่ม" , "ใช้" , "เชิญ" , "ออก" , "เมื่อ" , "ตอบ" , "อย่า" , "ตาม" , "ถือ" , "ทำ" , "ห้าม" , "อ้าง" , "ไป" , "หา" , "ขาด" , "เลย" , "เพื่อ" , "ขณะ" , "ยัง" , "ก็" , "นัก" , "ส่ง" , "คน" , "ขอ" , "ทราบ" , "ตัด" , "ส่วน" , "ซึ่ง" , "สั่ง" , "โดย" , "อย่าง" , "ผู้" , "รอง" , "อดีต" , "ยื่น" , "เดิน" , "เรา" , "พ้น" , "ชี้" , "นำ" , "การ" , "ถูก" , "ต้อง" , "ยึด" , "เท่า" , "แถลง" , "ติด" , "เกี่ยว" , "ค่อน" , "เสีย" , "ย้าย" , "บ่าย" , "เรียก" , "รัก" , "ชิง" , "เพราะ" , "โทร" , "มอง" , "จ." , "เขต" , "หัว" , "รมช." , "รมว." , "ผอ." , "ปลัด" , "แสดง" , "เกาะ" , "สนับ" , "นัด" , "ถาม" , "จบ" , "ปั้น" , "ครั้ง" , "ดู" , "คุย" , "รู้" , "กลับ" , "ขึ้น" , "หลาย" , "ย้อน" , "สิ่ง" , "เรื่อง" , "ไว้" , "ด่า" , "ตี" , "ชนะ" , "คลุก" , "ทิ้ง" , "เจอ" , "กลาย" , "ล้วน" , "หาก" , "เอา" , "กับ" , "ร่วม" , "หมด" , "จริง" , "ทาง" , "เสนอ" , "ตน" , "ทั้ง" , "แม้" , "เหลือ" , "เหมือน" , "ปรับ" , "กด" , "ยอม" , "ชาว" , "แต่ง" , "ข้อ" , "ไล่" , "บ้าน" , "มา" , "ส่อ" , "ทัก" , "เคลื่อน" , "จัด" , "ปิ้ง" , "ขับ" , "รับ" , "แห่ง" , "เช่น" , "ทุก" , "รวบ" , "ล็อบ" , "น่า" , "อม" , "แน่" , "เกรง" , "วาง" , "พรรค" , "รีบ" , "นั่ง" , "แฉ" , "กลัว" , "เหลือบ" , "เส้น" , "เจ้า" , "ก่อน" , "เก็บ" , "พัก" , "ร้อง" , "ลด" , "ชก" , "สาด" , "มิ" , "หมิ่น" , "สาป" , "ผิด" , "จุด" , "ลาก" , "หน้า" , "ขนาด" , "จอง" , "แชมป์" , "เพิ่ง" , "โดน" , "วัย" , "บ่น" , "ตอก" , "พัก" , "จเร" , "ผวจ." , "รมต." , "ผบช." , "ครม." , "ผบช." , "หนึ่ง" , "กกต." , "ผบ." , "แพทย์" , "สว." ,

"สวส." , "ผบก." , "ทนาย" , "โหร" , "อดีต" , "บุตรชาย" , "นายก" , "บุตรสาว" , "สำหรับ" , "ช่วยผลัก" , "ประกาศ" , "สัมภาษณ์" , "อาทิ" , "ลูกชาย" , "น้องชาย" , "บุตรเขย" , "หลานชาย" , "สามี" , "ภรรยา" , "ภริยา" , "น้องสาว" , "พยายาม" , "ลำบาก" , "สะดวก" , "ปราศรัย" , "ปฏิเสธ" , "นอกจาก" , "เนื่องจาก" , "ลาออก" , "จากนั้น" , "รายงาน" , "หาเสียง" , "สะท้อน" , "อธิบาย" , "สาเหตุ" , "ช่วยเร่ง" , "พาตน" , "อนุมัติ" , "เพียงแต่" , "ประธาน" , "ส.ส." , "รมช." , "รมว." , "โฆษก" , "เลขา" , "ส.ว." , "แม่ทัพ" , "ส.ก." , "ส.จ." , "ควรจะ" , "แนะนำ" , "กำลัง" , "ยินดี" , "เพียงคน" , "เกิดเพราะ" , "ควรกลับ" , "สุดๆ" , "ประสบ" , "ช่วยสาน" , "ปรากฏ" , "กำชับ" , "ต่อสู้" , "พาไป" , "เกิดขึ้น" , "ประเมิน" , "ยิ่งถ้า" , "ชะลอ" , "กรณี" , "จำนวน" , "ช่วยใน" , "อาจจะ" , "อาจมี" , "วันที่" , "ควรไป" , "ลำดับ" , "เพียงอย่าง" , "อาจยัง" , "เติบโต" , "ยิ้มแย้ม" , "ประตู" , "ปฏิบัติ" , "หลังจาก" , "ดำรง" , "ใจเย็น" , "ลาหยุด" , "ควรทำ" , "สนิทสนม" , "ชักชวน" , "ชอบพอ" , "อายุ" , "มั่นใจ" , "กรรมการ" , "รักษา" , "กกต." , "สมาชิก" , "ตัวแทน" , "กำนัน" , "สารวัตร" , "อธิบดี" , "อธิการ" , "คณบดี" , "พระเอก" , "นางเอก" , "แกนนำ" , "ดารา" , "พิธี" , "ราษฎร" , "อาจารย์"

รายการดังต่อไปนี้จะแสดง [รายการของคำบอกจุดสิ้นสุด] ที่จะเป็นจุดสิ้นสุดของขอบเขตชื่อเฉพาะประเภทชื่อองค์กร

"บริษัท" , "จำกัด" , "ประกาศ" , "ประเทศ" , "ประมูล" , "หมู่บ้าน" , "ระบุ" , "สำนัก" , "โรงเรียน" , "จำนวน" , "สามารถ" , "สำหรับ" , "ตั้งแต่" , "ดำเนิน" , "อาทิ" , "วันนี้" , "ระดม" , "สานต่อ" , "ตำหนิ" , "ร้องเรียน" , "เจรจา" , "ประกอบ" , "กรณี" , "อภิปราย" , "เลขา" , "เคารพ" , "ประกวด" , "หวังผล" , "ประท้วง" , "ทอดทิ้ง" , "ปฏิเสธ" , "ระมัด" , "ทันที" , "สมาชิก" , "ความรอด" , "วานนี้" , "คะแนน" , "พยายาม" , "กำลัง" , "ดาหน้า" , "ปรามาส" , "สนใจ" , "จำเป็น" , "ปรากฏ" , "ปราศรัย" , "วันที่" , "วันเดียว" , "ฉะนั้น" , "ทางภาค" , "กังวล" , "สำคัญ" , "ระวัง" , "บริหาร" , "เหนือกว่า" , "สาขา" , "สัมมนา" , "น.พ." , "จำนวน" , "สนับสนุน" , "นายก" , "ประเมิน" , "แสดง" , "ใน" , "รับ" , "อ." , "โดย" , "รับ" , "นำ" , "และ" , "เสนอ" , "เมื่อ" , "เอง" , "ก็" , "ไป" , "เพิ่ม" , "มี" , "เป็น" , "หรือ" , "ของ" , "ถือ" , "ใช้" , "ที่" , "ยัง" , "ทำ" , "ซึ่ง" , "นั้น" , "อ้าง" , "ส่วน" , "ให้" , "จึง" , "ก่อน" , "ตรวจ" , "แล้ว" , "ว่า" , "เกี่ยว" , "เรื่อง" , "จะ" , "หู" , "เดิน" , "จน" , "สภา" , "คง" , "จัด" , "วิ" , "ได้" , "สั่ง" , "เท่า" , "แม้" , "ไม่" , "แผนก" , "ตี" , "ตัด" , "กลับ" , "จาก" , "ขึ้น" , "เปิด" , "ต่อ" , "เพราะ" , "เหมือน" , "เช่น" , "ตัว" , "เข้า" , "ควร" ,"เห็น" ,"ด้าน" ,"จุด" , "คือ" , "อาจ" , "ต้อง" , "แต่" , "ตาม" , "เพื่อ" , "มา" , "อย่าง" , "อยู่" , "กล่าว" , "หลาย" , "ค่อน" , "รวม" , "ขอ" , "แน่" , "ขณะ" , "ครั้ง" , "ออก" , "หาก" , "การ" , "อีก" , "พรรค" , "เสีย" , "กลาย" , "ด้วย" , "ห้าม" , "เขต" , "คน" , "จริง" , "ตลอด" , "ไว้" , "นอก" , "เกิด" , "คิด" , "ลง" , "ข้า" , "ย้ำ" , "นาน" , "เสีย" , "เหตุ" , "น่า" , "ทั้ง" ,

"นาย" , "พร้อม" , "เริ่ม" , "เตรียม" , "บน" , "ย้าย" , "ดัง" , "กลัว" , "ต่อ" , "ยก" , "มอง", "ทราบ" , "ย่อม" , "ยอม" , "ล้ม", "บวก" , "เกือบ" , "กวาด" , "ขาด", "มุ่ง" , "ช่วง" , "นี้" , "คาด", "บ้าง" , "ทักทาย" , "พา" , "บ่ง", "ข่ม" , "ศาล" , "บอก" , "หลัง" , "พบ" , "คั่น" , "กกต." , "อยาก" , "เหลือ" , "หวั่น" , "สับ" , "เยอะ" , "บาง" , "ซื้อ" , "ตอบ" , "ท่าม" , "เรา" , "เน้น" , "หมาย" , "ลด" , "ข้อ" , "วาง" , "ถึง" , "ล้วน" , "เชื่อ" , "ยิ่ง" , "ดี" , "เล็ก" , "ขน" , "หนี" , "เอา" , "บุตร" , "แจ้ง" , "หา" , "นาง" , "มั่น" , "จับ" , "สมัย" , "ฝ่าย" , "หลาน" , "ดู" , "ชู" , "ตก" , "โต" , "ทิ้ง" , "แยก" , "มาก" , "ปรับ" , "อัน" , "ยึด" , "จ้าง" , "เนื่อง" , "เคย" , "ย้ำ" , "รอ" , "แจก" , "ปลด" , "มอบ" , "ยื่น" , "อ่อน" , "จ่าย" , "สูญ" , "แพ้" , "พูด" , "ร่วม" , "หัน" , "ถูก" , "เด็ด" , "ภาย" , "ตอน" , "ชัด" , "ทุก" , "สู้" , "ล่า" , "ตน" , "ทั่ว" , "สร้าง" , "ตั้ง" , "ครบ" , "เจริญ" , "กลุ่ม" , "จ." , "แทน" , "ถ้า" , "แตก" , "ส่ง" , "ฮั้ว" , "แน่น" , "เว้น" , "ผู้" , "ชนะ" , "เลย" , "หนุน" , "หมด" , "เปลี่ยน" , "กับ" , "ตราบ" , "สมัคร" , "กี่" , "ข่าว" , "เดิม" , "อดีต" , "กว่า" , "ต่าง" , "รู้" , "ลา" , "ยืน" , "แถลง" , "หนึ่ง" , "เรียก" , "ทาบ" , "ยุบ" , "เลือก" , "เลื่อน" , "เล่น" , "ถาม" , "ติด" , "ปั่น" , "ใช่" , "ล้าน" , "ใหญ่" , "ชี้" , "ปลอม" , "แกน"

รายการดังต่อไปนี้จะแสดง [รายการของคำบอกจุดสิ้นสุด] ที่จะเป็นจุดสิ้นสุดของขอบเขตชื่อเฉพาะประเภทชื่อสถานที่

"จ." , "เป็น" , "ล้วน" , "เหมือน" , "มั่น" , "พร้อม" , "ถนน" , "ว่า" , "ทั้ง" , "คน" , "พ." , "นาย" , "เพื่อ" , "ก็" , "และ" , "โดย" , "ยัง" , "เกี่ยว" , "อ." , "ต." , "จาก" , "แม้" , "รวม" , "ไม่" , "ซึ่ง" , "จะ" , "มี" , "เขต" , "เท่า" , "ถึง" , "อย่าง" , "แต่" , "เนื่อง" , "ด้วย" , "ของ" , "เสร็จ" , "จน" , "เรื่อง" , "เลย" , "ล่อง" , "ดี" , "ส่วน" , "เพราะ" , "ช่วย" , "ให้" , "หาย" , "อยู่" , "ถือ" , "ก่อน" , "เที่ยง" , "เช่น" , "ไป" , "ใน" , "จึง" , "เกิด" , "นัก" , "เสนอ" , "ที่" , "พื้น" , "แผน" , "ลง" , "มูล" , "เดิน" , "หลาย" , "ย่าน" , "หรือ" , "ตลอด" , "แล้ว" , "แพทย์" , "อดีต" , "โพล" , "ไหล" , "เอา" , "กว่า" , "เขื่อน" , "นคร" , "กิ่ง" , "สนาม" , "ได้" , "ทำ" , "พรรค" , "นาง" , "แกน" , "ซอย" , "ผู้" , "นั้น" , "ตน" , "ขึ้น" , "กับ" , "สา" , "ออก" , "เข้า" , "ขอ" , "ทัน" , "มาก" , "แห่ง" , "จริง" , "ยาม" , "ใกล้" , "ต่อ" , "ต้อง" , "ดำ" , "ตาม" , "พ้น" , "น่า" , "ทุก" , "ตก" , "วัน" , "หัว" , "นอก" , "คง" , "เปิด" , "ครั้ง" , "การ" , "ภู" , "เมื่อ" , "พบ" , "เมือง" , "หมู่" , "สมาคม" , "แนะ" , "ภาย" , "ถ้า" , "แขวง" , "ช่วง" , "โครง" , "ตั้ง" , "คณะ" , "ถูก" , "คลอง" , "หน้า" , "ตอน" , "ไฮ" , "อีก" , "วาน" , "กล่าว" , "หัน" , "จำ" , "หลัง" , "เพิ่ม" , "เบื่อ" , "ลา" , "สมัย" , "อัน" , "ฝั่ง" , "รับ" , "สนับสนุน" , "คือ" , "แถลง" , "เกือบ" , "จัด" , "ภาค" , "เริ่ม" , "แทน" , "เหลือ" , "สรุป" , "ปัญหา" , "ไข้" , "แบบ" , "ขั้น" , "เขา" , "ณ" , "แทบ" , "ผ่าน" , "รอบ" , "โรค" , "ครบ" , "คาด" , "ดัง" , "ส่ง" , "หาก" , "กี่" , "ไว้" , "ยก" , "บัตร" , "มัน" , "เต็ม" , "อยาก" , "บ้าง" , "กอง" , "ใคร" , "ป้าย" , "ครบ" , "หมด" , "เดือน" , "เอง" , "พอ" , "ตรวจ" , "ดู" , "ทาง" , "ซัก" , "ญาติ" , "ขณะ" , "นำ" , "พูด" ,

"เดิม" , "ขบวน" , "โจมตี" , "ผูก" , "รัฐ" , "กิจการ" , "ยอม" , "อาจ" , "แก้" , "ห้าม" , "ท่าม" , "คิด" , "เห็น" , "กรุง" , "เคย" , "หลัก" , "ประมาณ" , "ระหว่าง" , "สำหรับ" , "ปรากฏ" , "จังหวัด" , "ศาลา" , "สำนัก" , "คุณหญิง" , "จำนวน" , "ประสบ" , "รัฐบาล" , "สามารถ" , "กำลัง" , "ประกาศ" , "ประธาน" , "อาทิ" , "ระยะ" , "โรงแรม" , "ประเทศ" , "ประปา" , "ประตู" , "ดำเนิน" , "มาต้อน" , "กำนัน" , "มารับ" , "กรณี" , "มาถึง" , "พล.ต." , "มาช่วย" , "บรรดา" , "ทำให้" , "มาแล้ว" , "พิจารณา" , "พล.อ." , "วันที่" , "ม.ร.ว." , "สักการะ" , "ปราศรัย" , "มารวม" , "บางส่วน" , "มาทำ" , "บางเขต" , "ประชาชน" , "ขัดขวาง" , "มาตั้ง"

# ประวัติผู้เขียนวิทยานิพนธ์

นางสาวสุฤดี ฉัตรไตรมงคล เกิดที่กรุงเทพมหานคร สำเร็จการศึกษาระดับปริญญาตรี วิทยาศาสตรบัณฑิต สาขาเทคโนโลยีทางอาหาร จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2544 และเข้าศึกษาต่อในหลักสูตร อักษรศาสตรมหาบัณฑิต ภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อปี พ.ศ. 2545