# การรู้จำชื่อเฉพาะภาษาไทย: การใช้แบบจำลองคอนดิชันนอลแรนดอมฟิลด์ส

นางสาว นัชชา ถิระสาโรช



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

# THAI NAMED ENTITY RECOGNITION: THE APPLICATION OF CONDITIONAL RANDOM FIELDS MODELS

Miss Nutcha Tirasaroj



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 2010

| | |
|---|---|
| หัวข้อวิทยานิพนธ์ | การรู้จำชื่อเฉพาะภาษาไทย: การใช้แบบจำลองคอนดิชันนอลแรนดอมฟิลด์ส |
| โดย | นางสาว นัชชา ถิระสาโรช |
| สาขาวิชา | ภาษาศาสตร์ |
| อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก | รองศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล |

---

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....ประพจน์ อัศววิรุฬหการ..... คณบดีคณะอักษรศาสตร์

(ผู้ช่วยศาสตราจารย์ ดร. ประพจน์ อัศววิรุฬหการ)

คณะกรรมการสอบวิทยานิพนธ์

.....สุดา รังกุพันธุ์..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. สุดา รังกุพันธุ์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล)

.....เทพชัย ทรัพย์นิธิ..... กรรมการภายนอกมหาวิทยาลัย

(ดร. เทพชัย ทรัพย์นิธิ)

นัชชา ถิระสาโรช : การรู้จำชื่อเฉพาะภาษาไทย: การใช้แบบจำลองคอนดิชันนอลแรนดอมฟิลด์ส. (THAI NAMED ENTITY RECOGNITION: THE APPLICATION OF CONDITIONAL RANDOM FIELDS MODELS) อ. ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร. วิโรจน์ อรุณมานะกุล, 147 หน้า.

วิทยานิพนธ์ฉบับนี้มีวัตถุประสงค์เพื่อพัฒนาระบบการรู้จำชื่อเฉพาะภาษาไทยโดยใช้แบบจำลองคอนดิชันนอลแรนดอมฟิลด์สโมเดล (CRFs) และศึกษาเปรียบเทียบประสิทธิภาพของระบบการรู้จำชื่อเฉพาะภาษาไทยระหว่างแบบจำลองที่รับข้อมูลเข้าเป็นพยางค์กับที่รับข้อมูลเข้าเป็นคำ

งานวิจัยนี้ใช้คลังข้อมูลข่าวขนาด 367,673 คำ ประกอบด้วยชื่อเฉพาะทั้งหมด 16,179 ชื่อ แบบจำลองที่ใช้คือ CRF++ เวอร์ชัน 0.53 ทั้งระบบที่รับข้อมูลเข้าเป็นคำและพยางค์ใช้คุณสมบัติแบบเดียวกัน ได้แก่ คุณสมบัติรายการชื่อเฉพาะ คุณสมบัติคำย่อ คุณสมบัติคำบริบท คุณสมบัติคำทั่วไป คุณสมบัติค่าทางสถิติ และคุณสมบัติ unigram และ bigram การเรียนรู้ของระบบเป็นแบบ supervised learning คือมีการให้คำตอบในคลังข้อมูลสำหรับฝึกฝน คำตอบที่ใช้มีทั้งหมด 5 แบบ โดยแบบที่ 1 มีข้อมูลขอบเขตของชื่อเฉพาะน้อยที่สุดและแบบที่ 5 มีข้อมูลขอบเขตของชื่อเฉพาะมากที่สุด พบว่าแบบคำตอบที่ให้ข้อมูลมากช่วยให้ประสิทธิภาพของทั้งสองระบบดีกว่าแบบคำตอบที่ให้ข้อมูลน้อย จากผลการทดสอบระบบ พบว่า ประสิทธิภาพของระบบที่ใช้ข้อมูลตัดคำและตัดพยางค์ไม่ต่างกัน โดยมีค่าความถูกต้อง (F-measure) เท่ากัน คือ 81.30% จากคุณสมบัติทั้งหมด พบว่า คุณสมบัติ unigram และ bigram สนับสนุนระบบที่ใช้ข้อมูลตัดพยางค์มากที่สุด และคุณสมบัติรายการชื่อเฉพาะสนับสนุนระบบที่ใช้ข้อมูลตัดคำมากที่สุด เมื่อนำข้อมูลมาผ่านกระบวนการประมวลผลภายหลังแล้ว ช่วยให้ค่าความครบถ้วนของทั้งสองระบบมากขึ้นจากเดิม 77.64% เป็น 80.15% และ 80.06% ในข้อมูลตัดคำและตัดพยางค์ตามลำดับ

ภาควิชา ภาษาศาสตร์ ลายมือชื่อนิสิต นัช-สาโรช

สาขาวิชา ภาษาศาสตร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก

ปีการศึกษา 2553

## 5080158022 : MAJOR LINGUISTICS

KEYWORDS : THAI NAMED ENTITY / NAMED ENTITY RECOGNITION / CONDITIONAL RANDOM FIELDS / CRFS

NUTCHA TIRASAROJ : THAI NAMED ENTITY RECOGNITION: THE APPLICATION OF CONDITIONAL RANDOM FIELDS MODELS. ADVISOR : ASSOC. PROF. WIROTE AROONMANAKUN, Ph.D., 147 pp.

The main purpose of this study is to develop Thai named entity recognition system using Conditional Random Fields Models (CRFs) as well as comparing the performance of syllable-based system to that of word-based system.

This study uses the news corpus of 367,673 words with 16,179 proper names. CRFs model applied in this research is CRF++ 0.53. Both word-based and syllable-based systems use the same set of features, including gazetteer lists, abbreviation, context clues, general words, statistics, and unigram and bigram. Supervised learning is applied to train CRFs. There are 5 patterns of answer given to the systems, the first pattern having the least information of the named entities' boundaries and the last one having the most information. The results show that the patterns containing more information tend to improve the systems' performances than those having less information. The testing results show that the performances of word-based and syllable-based systems are not different from each other. The recognition rates (F-measure) of these two systems are 81.30%. From all of the features used, the unigram and bigram support the syllable-based system the most, while the gazetteer lists support the word-based system the most. After post-processing, the recalls of the two systems increase from 77.64% to 80.15% and 80.06% in word-based and syllable-based models respectively.

Department : Linguistics Student's Signature N. Tirasaroj

Field of Study : Linguistics Advisor's Signature

Academic Year : 2010

# กิตติกรรมประกาศ

ผู้วิจัยต้องขอขอบพระคุณ รองศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์เป็นอย่างสูง ที่ได้ให้คำแนะนำและความช่วยเหลือในการทำวิจัย ตลอดจนปรับแก้วิทยานิพนธ์ฉบับนี้จนสำเร็จลุล่วงไปด้วยดี และผู้วิจัยขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. สุดา รังกุพันธุ์ และ ดร. เทพชัย ทรัพย์นิธิ กรรมการสอบวิทยานิพนธ์ที่ได้ให้ข้อชี้แนะและเสียสละเวลาในการตรวจแก้วิทยานิพนธ์ฉบับนี้ให้มีความสมบูรณ์มากยิ่งขึ้น

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. สุดาพร ลักษณียนาวิน ที่ให้โอกาสผู้วิจัยได้ช่วยทำงานวิจัยด้านภาษาศาสตร์อันเป็นการเพิ่มพูนประสบการณ์ความรู้ รวมถึงคณาจารย์ภาควิชาภาษาศาสตร์ทุกท่านที่คอยให้คำแนะนำและประสิทธิ์ประสาทความรู้ด้านภาษาศาสตร์แก่ผู้วิจัย

ขอขอบคุณโครงการทุนวิจัยมหาบัณฑิต สกว. ด้านมนุษย์ศาสตร์-สังคมศาสตร์ ที่สนับสนุนทุนการวิจัยต่าง ๆ ศูนย์ความเป็นเลิศทางวิชาการด้านภาษา ภาษาศาสตร์ และวรรณคดี คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่สนับสนุนค่าใช้จ่ายในการกำกับข้อมูล ศูนย์วิจัยการประมวลผลภาษาและวัจนะ จุฬาลงกรณ์มหาวิทยาลัย ที่ให้ความอนุเคราะห์ข้อมูลรายการซื้อเฉพาะประเภทต่าง ๆ และศูนย์วิจัยเทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ที่ให้การสนับสนุนด้านคลังข้อมูลและให้ความอนุเคราะห์ข้อมูล “นามสงเคราะห์ส่วนราชการไทย” อันเป็นประโยชน์ต่อการทำวิจัยครั้งนี้

สุดท้ายนี้ ผู้วิจัยต้องขอขอบพระคุณบิดา มารดา คุณกฤติน และคุณธมน ถิระสาโรช ที่ให้การสนับสนุน ห่วงใยและเป็นกำลังใจที่ดีตลอดมา คุณสมบัษร ถิระสาโรช ที่ให้โอกาสผู้วิจัยได้ศึกษาต่อและสนับสนุนค่าใช้จ่ายตลอดระยะเวลาการศึกษา คุณณัฐดาพร เลิศชีวะ คุณศศิวิมล กาลันสีมา และคุณธารทอง แจ่มไพบูลย์ เพื่อนร่วมวิชาเอกที่ให้กำลังใจและเป็นที่ปรึกษาที่ดีมาโดยตลอด รวมทั้งขอขอบคุณเพื่อน ๆ พี่ ๆ น้อง ๆ และเจ้าหน้าที่ภาควิชาภาษาศาสตร์สำหรับการสนับสนุนและช่วยเหลือในด้านต่าง ๆ เสมอมา

# สารบัญ

บทที่ หน้า

บทที่ หน้า

# สารบัญตาราง

ตารางที่ หน้า

# สารบัญภาพ

# บทที่ 1

# บทนำ

## 1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันคอมพิวเตอร์เข้ามามีบทบาทในชีวิตประจำวันของเราอย่างมาก โดยเฉพาะเรื่องการใช้อินเตอร์เน็ต ที่มีผู้ใช้งานเป็นจำนวนมาก และมีแนวโน้มว่าจะมากขึ้นเรื่อย ๆ สิ่งหนึ่งที่คนนิยมใช้งานอินเตอร์เน็ต คือ การค้นหาข้อมูล เพราะมีความสะดวก และรวดเร็ว อย่างไรก็ตามความสำคัญของการใช้ประโยชน์จากคอมพิวเตอร์ไม่ได้จำกัดอยู่แค่การค้นหาข้อมูลในอินเตอร์เน็ต แต่ยังครอบคลุมไปถึงการใช้งานตามห้องสมุดต่าง ๆ เช่น หากเราต้องการค้นหาหนังสือ ก็จะมีระบบสำหรับการสืบค้นเพื่ออำนวยความสะดวก ซึ่งทั้งการค้นหาข้อมูลบนอินเตอร์เน็ต หรือการค้นหาหนังสือในห้องสมุด เราต่างต้องใส่คำค้นเข้าไป เพื่อให้ระบบสามารถค้นหาและดึงข้อมูลหรือเอกสารที่เกี่ยวข้องกับสิ่งที่เราต้องการออกมาได้ ระบบที่ใช้ดึงข้อมูลหรือเอกสารนี้ เรียกว่า การค้นคืนสารสนเทศ (Information Retrieval) โดยส่วนใหญ่ผลการค้นที่ได้จะต้องมีเอกสารที่ไม่ตรงตามความต้องการของเรารวมอยู่ด้วยไม่มากก็น้อย เช่น หากคำค้นของเราคือ “ร้านเพื่อนช่วยเพื่อน” ซึ่งถือได้ว่าเป็นชื่อเฉพาะ (named entity) ประเภทหนึ่ง ผลการค้นที่ได้อาจจะปนกันทั้งชื่อเฉพาะและคำนามทั่วไป (common noun) ว่า ร้านเพื่อน-ช่วยเพื่อน เช่น “เป็นร้านเพื่อน ช่วยเพื่อนแนะนำ” เป็นต้น ดังนั้นจึงเห็นได้ว่าปัญหาของการค้นคืนสารสนเทศอย่างหนึ่ง คือ ไม่สามารถแยกความแตกต่างระหว่างชื่อเฉพาะและคำนามทั่วไปได้

นอกจากการค้นคืนสารสนเทศที่มีปัญหาเรื่องการแยกความแตกต่างระหว่างชื่อเฉพาะและคำนามทั่วไปแล้ว ยังมีการแปลภาษาด้วยเครื่องคอมพิวเตอร์ (Machine Translation) ที่มีปัญหาเรื่องนี้เช่นกัน เพราะมีผลต่อการแปลข้อมูลของเครื่อง โดยในการแปลภาษาจะเกี่ยวข้องกับคำ ซึ่งหากไม่สามารถแยกแยะได้ว่าคำใดเป็นชื่อเฉพาะ คำใดเป็นคำนามทั่วไป ก็จะมีปัญหาในการแปลอย่างมาก เพราะทำให้แปลผิด เช่น “Tropical Storm Fay” ที่ถูกคือ “พายุโซนร้อนเฟย์” แต่หากคอมพิวเตอร์ไม่รู้ว่า “Fay” คือ ชื่อเฉพาะก็จะแปลออกมาว่า “พายุโซนร้อนนางฟ้า” เพราะในความหมายทั่วไป fay หมายถึง นางฟ้า เป็นต้น

ด้วยเหตุนี้ จึงเห็นได้ว่า การสกัดชื่อเฉพาะออกจากคำนามทั่วไปนั้น เป็นงานที่สำคัญงานหนึ่งของการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ที่มีการทำวิจัยกันอย่างต่อเนื่องในหมู่นักวิศวกรคอมพิวเตอร์ นักวิทยาศาสตร์คอมพิวเตอร์ และนักภาษาศาสตร์คอมพิวเตอร์ เพื่อหาแนวทางและวิธีการที่มีประสิทธิภาพมากที่สุดในการสกัดชื่อเฉพาะ

สำหรับงานวิจัยเกี่ยวกับการรู้จำชื่อเฉพาะในภาษาต่างประเทศได้มีการทำการวิจัยกันมานานแล้วสำหรับภาษาตะวันตก โดยเฉพาะภาษาอังกฤษ ทำให้ค่าความถูกต้องในการรู้จำชื่อเฉพาะของภาษาอังกฤษค่อนข้างสูง ในขณะที่ภาษาในแถบเอเชียเพิ่งเริ่มมีการศึกษาเมื่อไม่นานมานี้ เช่น ภาษาญี่ปุ่น ภาษาจีน สำหรับปริมาณงานวิจัยเรื่องการรู้จำชื่อเฉพาะของภาษาไทยในปัจจุบันนั้น พบว่ายังมีไม่มาก และประสิทธิภาพยังไม่ดีเท่าที่ควรที่จะนำไปประยุกต์ใช้ได้จริง ดังนั้นผู้วิจัยจึงมีความสนใจที่จะศึกษาและพัฒนาระบบวิธีการที่ช่วยในการสกัดชื่อเฉพาะภาษาไทย เพื่อเป็นประโยชน์ต่อการพัฒนาระบบงานที่เกี่ยวข้องกับการประมวลผลภาษาธรรมชาติด้านอื่น ๆ ต่อไป

สำหรับงานวิจัยนี้ ผู้วิจัยมีความสนใจที่จะใช้โมเดลทางสถิติคอนดิชันนอลแรนดอมฟิลด์ส (Conditional Random Fields : CRFs) ในการรู้จำชื่อเฉพาะ ซึ่งจากงานวิจัยเกี่ยวกับการประมวลผลภาษาธรรมชาติของภาษาไทยที่ผ่านมา ได้มีการนำโมเดล CRFs ไปใช้ในการวิเคราะห์หาหน่วยคำไทย เพื่อนำไปใช้ในการตัดคำและการกำกับชนิดของคำ (Kruengkrai, Sornlertlamvanich, and Isahara, 2006) โดยเมื่อศึกษาเปรียบเทียบกับวิธีการอื่น ได้แก่ Longest Matching และ Maximum Matching พบว่า ผลที่ได้จาก CRFs ดีที่สุด นอกจากนี้ Haruechaiyasak, Kongyoung, and Dailey (2008) ได้ทำการศึกษาเปรียบเทียบประสิทธิภาพของระบบการตัดคำจากแบบจำลอง 4 แบบ ได้แก่ Naive Bayes, decision tree, Support Vector Machine, และ CRFs ซึ่งผลที่ออกมาปรากฏว่า CRFs ใช้กับการตัดคำภาษาไทยได้ดีกว่าแบบจำลองอื่น สำหรับงานด้านการรู้จำชื่อเฉพาะในภาษาไทยนั้นยังไม่มีการใช้ CRFs เป็นที่แพร่หลายมากนัก มี อัศนีย์ ก่อตระกูล (2550) ใช้ CRFs ในการรู้จำชื่อเฉพาะกับข้อมูลที่ผ่านการตัดคำแล้ว แต่จากงานวิจัยในภาษาอื่น เช่น ภาษาจีน พบว่าได้มีการนำ CRFs มาใช้ในงานวิจัยอย่างแพร่หลาย อีกทั้ง CRFs ยังใช้ได้ดีกับงานประมวลผลภาษาธรรมชาติด้านต่าง ๆ เช่น การตัดคำ (Wu et al., 2008; Zhao and Kit, 2006) และการรู้จำชื่อเฉพาะ (Feng, Huang, and Sun, 2008; He and Wang, 2008; Mao et al., 2008; Zhou et al., 2006) เป็นต้น

ในงานวิจัยในเรื่องการรู้จำชื่อเฉพาะในภาษาไทยที่ผ่านมา พบว่าส่วนใหญ่จะใช้ข้อมูลเข้าแบบเป็นคำ เช่น Charoenpornsawat, Kijsirikul, and Meknavin (1998), Chanlekha et al. (2002), อัศนีย์ ก่อตระกูล (2550) เป็นต้น แต่ผู้วิจัยมีความเห็นว่าข้อมูลที่ใช้ไม่จำเป็นต้องเป็นแบบคำเสมอไป สามารถเป็นแบบพยางค์ก็ได้ ทั้งนี้เพราะจากงานวิจัยของภาษาจีนซึ่งมีลักษณะคล้ายกับภาษาไทย คือ ไม่มีการแบ่งช่องว่างระหว่างคำ และไม่มีการใช้สัญลักษณ์บ่งบอกชื่อเฉพาะ เช่น อักษรตัวพิมพ์ใหญ่ในภาษาอังกฤษ งานวิจัยของภาษาจีนจะมีงานวิจัยทั้งที่ใช้ข้อมูลเข้าแบบคำและแบบเป็นตัวอักษร (character) หรือเทียบได้กับพยางค์ โดยข้อมูลเข้าแบบคำมีข้อดี คือ สามารถนำคุณสมบัติ (features) ที่สำคัญซึ่งอยู่ในระดับคำมาใช้ได้ เช่น รายการคำบ่งชี้ (Yang,

Zhao, and Zou, 2008; Zhou et al., 2006) เป็นต้น แต่ทั้งนี้ ในส่วนของการตัดคำ หากตัดคำผิดก็จะส่งผลกระทบต่อการรู้จำชื่อเฉพาะได้ด้วยเช่นกัน ดังนั้นเพื่อหลีกเลี่ยงปัญหานี้ จึงมีผู้วิจัยหลายท่าน (He and Wang, 2008; Wu, Jan et al., 2006; Wu, Yang and Lin, 2006) เลือกใช้ข้อมูลเข้าแบบเป็นตัวอักษรแทน แต่ทั้งนี้การใช้ข้อมูลเข้าแบบตัวอักษรก็มีข้อเสียเช่นกัน คือ ขนาดหน้าต่างของข้อความมีขนาดเล็กทำให้ยากต่อการรู้จำชื่อเฉพาะที่ประกอบด้วยตัวอักษรหลายตัว เช่น ชื่อองค์กร (Jing et al., 2003 อ้างถึงใน Yu et al., 2006) เป็นต้น อย่างไรก็ตาม จากงานวิจัยหลายงาน (He and Wang, 2008; Jing et al., 2003 อ้างถึงใน Yu et al., 2006) พบว่า ระบบการรู้จำชื่อเฉพาะที่ใช้ข้อมูลเข้าแบบเป็นตัวอักษรให้ผลดีกว่าข้อมูลเข้าเป็นคำ ดังนั้นผู้วิจัยจึงสนใจศึกษาเปรียบเทียบประสิทธิภาพของระบบการรู้จำชื่อเฉพาะของภาษาไทยที่รับข้อมูลเข้าเป็นพยางค์กับที่รับข้อมูลเข้าเป็นคำว่าจะให้ผลที่แตกต่างกันหรือไม่

## 1.2 วัตถุประสงค์

1. พัฒนาระบบการรู้จำชื่อเฉพาะของภาษาไทยโดยใช้คอนดิชันนอลแรนดอมฟิลด์สโมเดล
2. เปรียบเทียบประสิทธิภาพของระบบการรู้จำชื่อเฉพาะภาษาไทยระหว่างแบบจำลองที่รับข้อมูลเข้าเป็นพยางค์กับที่รับข้อมูลเข้าเป็นคำ
3. วิเคราะห์ลักษณะทางภาษาศาสตร์ที่มีผลต่อประสิทธิภาพของแบบจำลองทั้งสอง

## 1.3 สมมติฐาน

แบบจำลองที่รับข้อมูลเข้าเป็นคำสามารถรู้จำชื่อเฉพาะได้ดีกว่าแบบจำลองที่รับข้อมูลเข้าเป็นพยางค์ เนื่องจากชื่อเฉพาะในภาษาไทยส่วนใหญ่เกิดจากการประสมคำเป็นหลัก

## 1.4 ขอบเขตของการวิจัย

ศึกษาและพัฒนาระบบการรู้จำชื่อเฉพาะประเภทชื่อบุคคล ชื่อองค์กร และชื่อสถานที่จากคลังข้อมูลข่าวในรูปตัวเขียนภาษาไทยที่มีชื่อเฉพาะทั้งหมดไม่ต่ำกว่า 5,000 ชื่อโดยชื่อเฉพาะแต่ละประเภทมีไม่ต่ำกว่า 1,000 ชื่อ

## 1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. เป็นแนวทางให้กับงานวิจัยที่ศึกษาเกี่ยวกับการประมวลผลภาษาธรรมชาติด้านอื่น ๆ ที่ต้องการใช้โมเดลทางสถิติคอนดิชันนอลแรนดอมฟิลด์ส
2. เป็นแนวทางในการศึกษาการรู้จำชื่อเฉพาะในภาษาไทย

## 1.6 วิธีดำเนินการวิจัย

1. ทบทวนวรรณกรรมที่เกี่ยวข้องกับความหมายของชื่อเฉพาะและงานวิจัยที่เกี่ยวกับระบบการรู้จำชื่อเฉพาะในภาษาไทยและภาษาที่มีลักษณะคล้ายคลึงกับภาษาไทย
2. เก็บรวบรวมข้อมูลและสร้างคลังข้อมูล โดยแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูล 90% ใช้ในการฝึกฝน และข้อมูล 10% ใช้ในการทดสอบ ซึ่งจะทดสอบทั้งหมด 10 ครั้งโดยจะแบ่งให้ข้อมูลทุกส่วนได้ใช้ในการทดสอบ
3. กำกับส่วนที่เป็นชื่อเฉพาะในคลังข้อมูล
4. ทำข้อมูลขึ้นเป็นสองชุด โดยชุดหนึ่งเป็นข้อมูลที่ผ่านการตัดพยางค์และอีกชุดหนึ่งเป็นข้อมูลที่ผ่านการตัดคำ
5. กำหนดคุณสมบัติที่จะใช้ในแบบจำลองสองแบบ คือ แบบที่รับข้อมูลเข้าเป็นพยางค์และที่รับข้อมูลเข้าเป็นคำ
6. พัฒนาระบบการรู้จำชื่อเฉพาะด้วยแบบจำลองทางสถิติคอนดิชันนอลแรนดอมฟิลด์ส
7. ทดสอบระบบการรู้จำชื่อเฉพาะด้วยแบบจำลองทางสถิติคอนดิชันนอลแรนดอมฟิลด์ส
8. นำผลที่ได้จากการทดสอบระบบมาผ่านขั้นตอนประมวลผลภายหลัง
9. ประเมินผล วิเคราะห์และสรุปผลการวิจัย

## 1.7 เครื่องมือที่ใช้ในการวิจัย

1. โปรแกรมภาษา Perl ของบริษัท Active Perl
2. แบบจำลองสถิติ CRF++ 0.53 ของ Taku Kudo จากเว็บไซต์ http://crfpp.sourceforge.net/
3. โปรแกรม Thaiseg version 2.01 ของภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

# บทที่ 2

# ทบทวนวรรณกรรม

ในส่วนนี้จะกล่าวถึงความหมายของชื่อเฉพาะจากแนวคิดต่าง ๆ เพื่อให้เข้าใจว่าชื่อเฉพาะจะหมายถึงสิ่งใดได้บ้าง และสิ่งใดไม่ใช่ชื่อเฉพาะ เพราะในภาษามีชื่ออยู่หลายชนิด เช่น ชื่อของบุคคล ชื่อของต้นไม้ ชื่อของสัตว์ เป็นต้น นอกจากนี้จะกล่าวถึงประเภทของชื่อเฉพาะ และแนวทางในการวิจัยการรู้จำชื่อเฉพาะในงานประมวลผลภาษาธรรมชาติ ปัญหาของภาษาไทยที่ทำให้ยากต่อการรู้จำชื่อเฉพาะ รวมถึงงานวิจัยที่ผ่านมาที่เกี่ยวกับระบบการรู้จำชื่อเฉพาะทั้งของภาษาไทยและภาษาต่างประเทศ และแบบจำลองทางสถิติคอนดิชันนอลแรนดอมฟิลด์สเพื่อเป็นแนวทางในการศึกษาการรู้จำชื่อเฉพาะสำหรับงานวิจัยครั้งนี้

## 2.1 ความหมายของชื่อเฉพาะ

### 2.1.1 ความหมายของชื่อเฉพาะในทางภาษาศาสตร์

ชื่อเฉพาะในทางภาษาศาสตร์ (Hanks, 2006: 134-135) คือคำที่ต่างจากคำทั่วไป โดยคำเป็นสิ่งที่ใช้แทนสิ่งของหรือเหตุการณ์ต่าง ๆ บนโลกในขณะที่ชื่อ (Names) เป็นสิ่งที่ตั้งขึ้นเพื่ออ้างถึงสิ่งใดสิ่งหนึ่งเท่านั้น แต่ไม่มีความหมายและไม่ได้แสดงถึงคุณสมบัติของสิ่งที่ใช้ชื่อนั้น ๆ เช่น ‘She is Jane’ Jane เป็นชื่อที่อ้างถึงบุคคลและไม่ได้บ่งบอกลักษณะใด ๆ เกี่ยวกับบุคคลที่ใช้ชื่อนี้ สำหรับโครงสร้างทางวากยสัมพันธ์เกี่ยวกับชื่อเฉพาะนั้นจะแตกต่างกันไปในแต่ละภาษา เช่น ในภาษาอังกฤษ ชื่อเฉพาะจะขึ้นต้นด้วยตัวพิมพ์ใหญ่ หรือในภาษาอังกฤษ ฝรั่งเศส และเยอรมันจะไม่ใช้คำนำหน้านามกับชื่อเมือง ภูเขา ทะเลสาบ แต่จะใช้คำนำหน้านามชี้เฉพาะกับชื่อทะเล แม่น้ำ และมหาสมุทร เป็นต้น

### 2.1.2 ความหมายของชื่อเฉพาะในทางปรัชญา

ชื่อเฉพาะในแง่มุมทางปรัชญา (Reimer, 2006: 137-139) เป็นการศึกษาประเด็นหลักสองประเด็น คือ ความหมายของชื่อเฉพาะคืออะไร และชื่อเฉพาะอ้างถึงสิ่งต่าง ๆ ได้อย่างไร ตามทฤษฎีของ Mill ชื่อไม่มีความหมายหรือคุณลักษณะใด ๆ แต่ความหมายของชื่อคือสิ่งแทนตัวผู้ใช้ชื่อนั้น ๆ และต้องมีตัวตนอยู่จริงในโลก หากเป็นชื่อของตัวละครหรือสิ่งที่ไม่มีอยู่จริงจะถือว่าไม่

มีความหมายในขณะที่ Description Theories กล่าวว่า ความหมายของชื่อเฉพาะมีลักษณะเป็นเชิงพรรณนา กล่าวคือ ความหมายของชื่อจะอยู่ที่คุณสมบัติ หรือลักษณะต่าง ๆ ของสิ่งที่ใช้ชื่อนั้น ๆ ดังนั้นความหมายของชื่อจึงจะไม่ตายตัว และอาจแตกต่างกันไป ขึ้นอยู่กับมุมมองและประสบการณ์ของแต่ละบุคคลที่มีต่อชื่อนั้น ๆ เช่น เมื่อกล่าวถึงชื่อ “พัชรศรี เบญจมาศ” บางคนอาจนึกถึงพิธีกรรายการ “ผู้หญิงถึงผู้หญิง” บางคนอาจนึกถึงพิธีกรช่วง“เก็บตก”หรือบางคนอาจนึกถึงเพื่อนบ้านชื่อกาละแมร์ เป็นต้น สำหรับ Description Theories สิ่งที่ถูกอ้างถึงไม่จำเป็นต้องมีอยู่จริงบนโลกเสมอไป เช่น ชื่อ Sherlock Holmes ใน Description Theories ถือว่ามีความหมายเพราะ Sherlock Holmes คือ นักสืบที่มีชื่อเสียงในนวนิยายของโคนัน ดอยล์ แต่หากเป็นทฤษฎีของ Mill จะถือว่าไม่มีความหมายเพราะเป็นเพียงตัวละครไม่มีอยู่จริงบนโลก

นอกจากสองทฤษฎีที่กล่าวไปแล้ว ทางปรัชญายังมี Causal Theories ซึ่งจะอธิบายถึงความสัมพันธ์ระหว่างชื่อและสิ่งที่ใช้ชื่อนั้น ๆ โดยประกอบด้วย 2 ทฤษฎีย่อย คือ theory of reference ‘fixing’ ที่จะใช้อธิบายว่า มีการกำหนดชื่อให้กับสิ่งนั้น ๆ ได้อย่างไร และ theory of reference ‘borrowing’ ใช้อธิบายว่ามีการนำชื่อมาใช้อ้างถึงสิ่งนั้น ๆ ในเวลาต่อมาได้อย่างไร เช่น เมื่อเราซื้อสุนัขสีน้ำตาลตัวหนึ่งมา แล้วตั้งชื่อให้ว่า ‘โกโก้’ จากนั้นเมื่อเราเรียก ‘โกโก้’ แล้วสุนัขตัวนั้นก็จะวิ่งมาหาเพราะรู้ว่าเรียกชื่อมัน เช่นนี้ถือว่าเป็น theory of reference ‘fixing’ คือ การกำหนดชื่อให้กับสิ่งที่ต้องการใช้อ้างถึงหรือสุนัข จากนั้นเมื่อต้องการกล่าวถึงสุนัขตัวนั้นในภายหลังกับเพื่อนของเรา เราจะใช้ชื่อ ‘โกโก้’ เรียกแทนและเพื่อนก็จะรู้ว่า ‘โกโก้’ คือชื่อของสุนัขเราโดยที่อาจจะยังไม่เคยเห็น ‘โกโก้’ มาก่อน แล้วจากนั้นเพื่อนก็สามารถไปพูดต่อกับคนอื่นถึง ‘โกโก้’ ได้โดยใช้ชื่อ ‘โกโก้’ แทนภาพของสุนัขที่มีเราเป็นเจ้าของ ซึ่งกรณีนี้ถือว่าเป็น theory of reference ‘borrowing’ คือ การยืมชื่อของสิ่ง ๆ หนึ่งไปใช้อ้างถึงมโนภาพของสิ่งนั้น ๆ ในใจของแต่ละบุคคล เนื่องจากยังไม่เคยได้เห็นสิ่งนั้นจริง ๆ อย่างไรก็ตาม ในงานวิจัยนี้จะไม่สนใจความหมายในทางปรัชญานี้

### 2.1.3 ความหมายของชื่อเฉพาะในทางอรรถศาสตร์

ในทางอรรถศาสตร์ (Lehrer, 2006: 141-143) มองว่าในการสร้างชื่อนั้น บางครั้งเมื่อมองแบบผิวเผินดูเหมือนชื่อไม่มีความหมาย แต่จริง ๆ อาจมีความหมายแฝงอยู่แต่ผู้พูดไม่ได้สังเกต เช่น ชื่อที่ตั้งจากคำศัพท์ทั่วไป ความหมายของคำนั้นอาจจะบ่งบอกถึงผู้ที่ถูกตั้งชื่อได้ เช่น Kofi หมายถึง เกิดในวันศุกร์ ซึ่ง Kofi Annan ก็เกิดในวันศุกร์จริง ๆ นอกจากนี้ชื่อยังสามารถบ่งบอกเพศได้ด้วยเช่นกัน เพราะชื่อบางชื่อจะตั้งให้กับเฉพาะผู้ชายหรือผู้หญิงเท่านั้น เช่น Paul เป็นชื่อของผู้ชาย แต่ถ้าเป็น Paula จะเป็นชื่อของผู้หญิง เป็นต้น หรือการตั้งชื่อเล่นก็อาจจะตั้งจาก

ลักษณะเด่นของคนผู้นั้น เช่น Red อาจตั้งให้กับคนที่มีผมสีแดง หรือตั้งแบบสัมพันธ์กับชื่อจริง เช่น Robert อาจตั้งเป็น Rob หรือ Bob ก็ได้

ในบางกรณีชื่อเฉพาะอาจกลายเป็นคำนามทั่วไปได้ด้วยเช่นกัน เมื่อใช้กล่าวถึงสิ่งที่มีส่วนสัมพันธ์กับบุคคลที่ใช้ชื่อนั้น ๆ เช่น สิ่งประดิษฐ์ที่สร้างขึ้นจากบุคคลนั้น ๆ เช่น แซนด์วิช (sandwich) ตั้งขึ้นจาก Lord Sandwich หรือ Beethoven ใช้กล่าวถึงบุคคลที่มีพรสวรรค์ทางด้านดนตรี เป็นต้น

จากความหมายของชื่อเฉพาะที่กล่าวไปแล้วข้างต้น แม้จะมองจากแนวทางที่ต่างกัน แต่สิ่งหนึ่งที่เหมือนกันคือ ชื่อเฉพาะเป็นสิ่งที่ตั้งขึ้นเพื่อใช้อ้างถึงสิ่งใดสิ่งหนึ่งโดยเฉพาะ โดยเมื่อวิเคราะห์จากภาษาที่ใช้จริงในสังคม จะพบว่า ชื่อเฉพาะมีทั้งที่มีความหมายและไม่มีความหมายเกี่ยวข้องกับสิ่งที่ใช้อ้างถึง เช่น ในภาษาไทยหากเด็กผู้หญิงเกิดตรงกับวันวิสาขบูชา พ่อแม่มักตั้งชื่อให้ว่า “วันวิสา” ในกรณีนี้ถือว่าชื่อมีความหมายเกี่ยวข้องกับผู้ที่ถูกอ้างถึง แต่หาก “ชมพู่” ซึ่งเป็นชื่อผลไม้ชนิดหนึ่งถูกนำมาตั้งเป็นชื่อบุคคล เช่นนี้ถือว่าชื่อเป็นสิ่งที่ใช้อ้างถึงเท่านั้น โดยไม่มีความหมายเกี่ยวข้องกับผู้ที่ถูกอ้างถึง

### 2.1.4 ความหมายของชื่อเฉพาะในภาษาไทย

สำหรับชื่อเฉพาะในภาษาไทยพระยาอุปกิตศิลปสาร (2546: 71-72) และกำชัย ทองหล่อ (2550: 197) ได้จัดให้ชื่อเฉพาะเป็นคำนามประเภทวิสามานยนาม หมายถึง คำนามที่เป็นชื่อเฉพาะที่สมมติตั้งขึ้นมาเพื่อใช้เรียก คน สัตว์ และสิ่งของบางอย่าง โดยวิสามานยนามต้องตั้งขึ้นเพื่อใช้เรียกคนคนเดียว สัตว์ตัวเดียว และของสิ่งเดียว ถึงจะเป็นชื่อของหมู่คณะก็ต้องเป็นหมู่หรือคณะเดียว เช่น ชาติไทย หมายถึง ไทยชาติเดียว หรือ ***สมศรี***เป็นหญิงสาวที่เด่นในสังคม เป็นต้น

นววรรณ พันธุเมธา (2549: 6) จัดให้ชื่อเฉพาะเป็นคำนามชนิดหนึ่ง โดยความหมายของคำนาม คือ เป็นคำที่หมายถึงสิ่งต่าง ๆ ที่เป็นรูปธรรมและนามธรรม แบ่งได้เป็น 2 ชนิด ได้แก่ คำนามสามัญ คือ คำที่หมายถึงสิ่งต่าง ๆ โดยทั่วไป เช่น การต่อสู้ ความรัก คน เป็นต้น และคำนามวิสามัญ คือคำที่หมายถึงสิ่งใดสิ่งหนึ่งโดยเฉพาะ เช่น กาญจนบุรี เป็นต้น

## 2.2 ประเภทของชื่อเฉพาะในงานประมวลผลภาษาธรรมชาติ

ในการแบ่งประเภทของชื่อเฉพาะเพื่อใช้ในการศึกษาวิจัยการรู้จำชื่อเฉพาะนั้น ส่วนใหญ่จะใช้ตามหลักการแบ่งจากการประชุมวิชาการ Message Understanding Conference (MUC) ซึ่งแบ่งงานของการรู้จำออกเป็น 3 ประเภท (Chinchor, 1998) ดังนี้

1. Entity names ได้แก่ ชื่อบุคคล (persons) องค์กร (organizations) และสถานที่ (locations)
2. Temporal expressions ได้แก่ วันที่ (dates) และเวลา (times)
3. Number expressions ได้แก่ จำนวนหรือค่าเงิน (monetary values) และเปอร์เซ็นต์ (percentages)

ใน 3 ประเภทนี้ พบว่าผู้ทำวิจัยส่วนใหญ่มักทำวิจัยเกี่ยวกับชื่อเฉพาะประเภทแรกมากที่สุด ทั้งนี้เพราะเมื่อเทียบกันทั้งสามประเภทแล้ว การสกัดชื่อเฉพาะประเภทแรกถือว่าเป็นงานที่ยากที่สุด (Palmer, 1997 อ้างถึงใน Ye, Chua, and Jimin, 2002) เพราะมีรูปแบบและลักษณะการเกิดที่คลุมเครือกว่า และมีผลต่อการประมวลผลภาษามากกว่าสองกลุ่มหลัง (อัศนีย์ ก่อตระกูล, 2549)

## 2.3 แนวทางการศึกษาชื่อเฉพาะ

โดยทั่วไป งานที่เกี่ยวกับการสกัดชื่อเฉพาะแบ่งออกได้เป็น 2 ส่วนหลัก คือ การหาตำแหน่งและขอบเขตของชื่อเฉพาะ และอีกส่วนคือ การระบุประเภทของชื่อเฉพาะ สำหรับแนวทางการศึกษาชื่อเฉพาะที่ผ่านมาสามารถแบ่งออกได้เป็น 3 แนวทาง ดังนี้

**2.3.1 แนวทางการใช้กฎ (rule-based)** คือ การใช้กฎต่าง ๆ ของภาษามาสกัดชื่อเฉพาะ ส่วนใหญ่มักเป็นผู้เชี่ยวชาญ หรือนักภาษาศาสตร์เป็นผู้วิเคราะห์กฎต่าง ๆ ของภาษาออกมา ตัวอย่างระบบที่ใช้กฎ เช่น ระบบ FACILE (Black, Rinaldi, and Mowatt, 1998) ที่กำหนดคุณสมบัติทางภาษา เช่น ลักษณะตัวพิมพ์ใหญ่พิมพ์เล็ก หน้าที่และความหมายของคำให้กับหน่วยคำ (token) แล้วนำมาผ่านกฎที่เขียนขึ้นเอง ตัวอย่างกฎ เช่น

A => B\C/D

โดยที่ A : คำตอบที่แสดงในรูปของชุดคุณสมบัติรวมถึงคะแนนที่ได้

C: หน่วยคำที่นำมาวิเคราะห์

B และ D: บริบทด้านซ้ายและขวาของ C ซึ่งสามารถละได้

เช่น [syn=NP, sem=ORG] (0.9) =>

\ [norm="university"],

[token="of"],

[sem=REGION|COUNTRY|CITY] /;

โดยที่ "university" เป็นรูปปกติของ "University" เป็นบริบทซ้ายของ "of"

"of" คือหน่วยคำที่นำมาวิเคราะห์

บริบทด้านขวามีคุณสมบัติเป็นชื่อเขต ชื่อประเทศ ชื่อเมือง

หากวลีใดตรงกับกฎนี้ จะได้คำตอบเป็นชุดคุณสมบัติที่มีค่าทางวากยสัมพันธ์ (syntactic tag) เป็นนามวลี (NP) และค่าทางอรรถศาสตร์ (semantic tag) เป็นองค์กร (ORG) เป็นต้น

สำหรับค่าความถูกต้องของระบบ FACILE ค่อนข้างสูง โดยผลการทดลองที่ดีที่สุด ได้ค่าความครบถ้วน (recall) 92% และค่าความแม่นยำ (precision) 93% อย่างไรก็ตาม แม้ว่าประสิทธิภาพของการรู้จำชื่อเฉพาะโดยใช้กฎจะค่อนข้างสูง แต่ต้องใช้เวลานานในการพัฒนาระบบและต้องอาศัยความรู้เฉพาะทางในการเขียนกฎ อีกทั้งกฎที่เขียนขึ้นมักใช้ได้กับภาษาที่ทำการทดลองนั้น ๆ เท่านั้นจึงยากจะนำไปปรับใช้กับภาษาอื่น ๆ ได้

2.3.2 **แนวทางการใช้แบบจำลองทางสถิติ (machine learning)** เป็นวิธีการให้เครื่องคอมพิวเตอร์เรียนรู้กฎ โดยอาจให้เครื่องเรียนรู้ลักษณะการปรากฏร่วมของคำและลำดับหมวดคำ ซึ่งเครื่องจะสามารถรู้จำชื่อเฉพาะใหม่ที่มีลักษณะคล้ายกับชื่อเฉพาะแบบเดิมได้ วิธีนี้จะต้องมีการกำหนดคุณสมบัติ (feature) ต่าง ๆ เพื่อช่วยให้แบบจำลองได้เรียนรู้ วิธีนี้จะรวดเร็วกว่าวิธีแรกและไม่ต้องจ้างผู้เชี่ยวชาญในการวิเคราะห์กฎ แต่ต้องใช้คลังข้อมูลขนาดใหญ่ในการฝึกฝนแบบจำลองทางสถิติที่ใช้ เช่น Support Vector Machines (SVMs), Decision Tree, Hidden Markov Models (HMMs), Maximum Entropy (MaxEnt, ME), Conditional Random Fields (CRFs) ฯลฯ

ตัวอย่างงานวิจัยที่ใช้แบบจำลองทางสถิติ เช่น งานของ Chieu and Ng (2002) ใช้แบบจำลอง Maximum Entropy ในการรู้จำชื่อเฉพาะภาษาอังกฤษ โดยใช้คุณสมบัติภายใน (local features) และภายนอก (global features) คุณสมบัติภายในได้จากหน่วยคำ (token) และบริบทรอบข้างของหน่วยคำ เช่น หน่วยคำเป็นอักษรพิมพ์ใหญ่และลงท้ายด้วยจุดหรือไม่ ประกอบด้วยตัวเลขหรือไม่ เป็นต้น ส่วนคุณสมบัติภายนอกได้จากข้อมูลทั้งเอกสารโดยคุณสมบัติภายนอกใช้เพื่อช่วยในการตัดสินหน่วยคำว่าเป็นชื่อเฉพาะหรือไม่ หรือเป็นชื่อเฉพาะชนิดใด ผล

การทดลองพบว่าการใช้คุณสมบัติภายนอกช่วยให้ประสิทธิภาพของระบบดีกว่าการใช้เพียงคุณสมบัติภายในอย่างเดียว โดยเมื่อทดสอบกับคลังข้อมูลของ MUC-6 ได้ค่า F-measure 93.27% จากเดิมที่ไม่ได้ใช้คุณสมบัติภายนอกได้ 90.75% และคลังข้อมูล MUC-7 ได้ 85.22% จาก 87.24%

McCallum and Li (2003) ใช้แบบจำลอง Conditional Random Fields รู้จำชื่อเฉพาะภาษาอังกฤษและภาษาเยอรมัน โดยใช้วิธี WebListing ในการดึงข้อมูลจากเว็บไซต์มาสร้างเป็นคลังศัพท์ ผลที่ออกมาปรากฏว่าระบบสามารถรู้จำชื่อเฉพาะภาษาอังกฤษได้ดีกว่าภาษาเยอรมัน เหตุผลหลักมาจากคุณสมบัติที่ใช้กับภาษาอังกฤษมีมากกว่าภาษาเยอรมัน เนื่องจาก GoogleSet ที่ใช้ช่วยดึงข้อมูลไม่ค่อยสนับสนุนภาษาอื่นนอกจากภาษาอังกฤษทำให้ได้คลังคำศัพท์ของภาษาเยอรมันน้อย แม้ว่าจะเพิ่มคุณสมบัติ bigram และ trigram ของตัวอักษรเข้าช่วยในภาษาเยอรมันแล้วก็ตาม

สำหรับภาษาในแถบเอเชีย เช่น ภาษาญี่ปุ่น Sekine, Grishman, and Shinnou (1998) ได้ใช้แบบจำลอง Decision Tree ในการรู้จำชื่อเฉพาะโดยใช้คุณสมบัติ 3 อย่าง คือ หมวดคำ (part of speech) รายการชื่อเฉพาะ และชนิดตัวอักษร เช่น คันจิ คาตาคานะ อักษรภาษาอังกฤษ ตัวเลข เป็นต้น ในการเรียนรู้ให้เรียนรู้จากคุณสมบัติของหน่วยคำและบริบทซ้ายขวาเป็นหลัก คลังข้อมูลที่ใช้ในการทดลองนี้มี 2 คลังด้วยกัน คลังแรกเป็นข้อมูลรายงานอุบัติเหตุของยานพาหนะได้ค่า F-measure 85% คลังที่สองเป็นข้อมูลเกี่ยวกับเหตุการณ์ความสำเร็จของผู้บริหาร โดยก่อนที่จะทำการทดลอง คณะผู้วิจัยได้เพิ่มชื่อตำแหน่งเข้าไปในรายการชื่อเฉพาะรวมถึงปรับเปลี่ยนระบบให้เข้ากับข้อมูล โดยได้ค่า F-measure 82% งานวิจัยนี้น่าสนใจที่แสดงให้เห็นว่าสามารถนำระบบไปปรับใช้กับข้อมูลประเภทอื่นในภาษาเดียวกันได้โดยใช้เวลาไม่มาก แต่อย่างไรก็ตามก็มีข้อจำกัดว่าข้อมูลที่ใช้จะต้องมีการกำกับไปในทิศทางเดียวกันมิฉะนั้นก็ใช้ไม่ได้หรือต้องปรับเปลี่ยนการกำกับของคลังข้อมูลฝึกฝนใหม่ทั้งหมด

สำหรับภาษาจีนมีงานวิจัยหลายงานใช้ Conditional Random Fields (Feng, Sun, and Lv, 2006; Mao et al., 2008; Wu, Yang et al., 2006; Zhou et al., 2006) และมีการเปรียบเทียบคุณสมบัติระหว่างการใช้ความรู้ภายนอกร่วมด้วย (open track) คือสามารถนำข้อมูลอื่นนอกเหนือจากที่ได้จากคลังข้อมูลฝึกฝนเท่านั้นมาใช้ กับคุณสมบัติที่ใช้เฉพาะข้อมูลที่ดึงจากคลังข้อมูลฝึกฝนเพียงอย่างเดียว (close track) ซึ่งผลการวิจัยต่าง ๆ ออกมาในทิศทางเดียวกันคือระบบที่ใช้ความรู้ภายนอกมาเป็นคุณสมบัติร่วมด้วยให้ผลดีกว่าอย่างชัดเจน (Feng et al., 2006; Wu, Yang et al., 2006; Zhou et al., 2006) เพราะการใช้ความรู้ภายนอกร่วมด้วยทำให้คุณสมบัติที่ใช้เพิ่มมากขึ้นในขณะที่คุณสมบัติที่ใช้กับคลังข้อมูล close track มีจำกัด

การใช้แบบจำลองทางสถิตินั้น สิ่งสำคัญที่สุดคือคุณสมบัติที่ใช้เพราะมีผลต่อประสิทธิภาพของระบบโดยตรง ซึ่งหากออกแบบคุณสมบัติได้ตรงตามลักษณะของข้อมูลหรือชื่อเฉพาะแล้วก็มีแนวโน้มว่าระบบจะมีประสิทธิภาพในการรู้จำได้ดี ดังนั้นจึงมีงานวิจัยบางงานออกแบบคุณสมบัติให้ชื่อเฉพาะแต่ละประเภทแตกต่างกันออกไป (Zhang et al., 2006) หรือแยกโมเดลสำหรับชื่อเฉพาะแต่ละประเภท (Zhang et al., 2006; Zhou et al., 2006)

สิ่งที่น่าสนใจที่ได้จากงานวิจัยของภาษาจีน คือ จำนวนเครื่องหมายกำกับขอบเขตของชื่อมีผลต่อประสิทธิภาพของระบบที่เป็นแบบ supervised learning คือมีการให้คำตอบในคลังข้อมูลฝึกฝน เมื่อเปรียบเทียบงานวิจัยของปี 2006 กับ 2008 พบว่าจำนวนเครื่องหมายกำกับของปี 2008 มีมากขึ้นจากเดิม เครื่องหมายที่ใช้ในปี 2006 เช่น BIO = B: เริ่มต้นชื่อ, I: ภายในชื่อ, O: ไม่ใช่ชื่อ (Feng et al., 2006) ในปี 2008 Mao et al. (2008) เลือกใช้เครื่องหมาย BIOE หรืองานของ Yu et al. (2008) เลือกใช้เครื่องหมาย BIOES แทน BIO โดยเพิ่ม E ซึ่งเป็นจุดสิ้นสุดชื่อเข้าไปและ S สำหรับชื่อเฉพาะที่เป็นอักษรตัวเดียว สาเหตุหลักที่ใช้เครื่องหมายจำนวนมากขึ้น เพื่อเพิ่มข้อมูลให้กับแบบจำลอง และจากผลการทดลองพบว่าจำนวนเครื่องหมายที่มากขึ้นช่วยให้ประสิทธิภาพของระบบดีขึ้น แต่ทั้งนี้การใช้จำนวนเครื่องหมายมากก็มีข้อด้อยเช่นกันโดยเฉพาะถ้าใช้กับ CRFs เพราะทำให้ประมวลผลช้า (Zhao and Kit, 2008)

เนื่องจากแบบจำลองทางสถิติที่ใช้ในงานรู้จำชื่อเฉพาะมีมากมายจึงมีงานวิจัยบางงานที่ทำการเปรียบเทียบแบบจำลองต่าง ๆ เช่น งานของ Feng, Sun, and Zhang (2005) เปรียบเทียบแบบจำลอง CRFs, HMM และ ME ในการรู้จำชื่อเฉพาะภาษาจีนโดยแบบจำลองทั้งสามใช้คุณสมบัติเดียวกันคือ หมวดคำ (POS: Part Of Speech) text (TXT) และทั้งสองอย่างรวมกัน (POSTXT) แต่แบบจำลอง CRFs และ ME จะมีอีกคุณสมบัติหนึ่งเพิ่มเข้าไปคือ ALL ซึ่งเป็นรายการตัวอักษรที่ปรากฏบ่อยและคำขึ้นต้นของชื่อเฉพาะ สาเหตุที่ไม่ใช้ ALL กับ HMM เนื่องจากปัญหาข้อมูลเกิดการกระจาย (data sparse problem) ค่อนข้างมาก ผลการทดลองพบว่าระบบที่ใช้ CRFs มีประสิทธิภาพในการรู้จำมากที่สุด อย่างไรก็ตามเป็นธรรมดาที่ CRFs จะมีประสิทธิภาพมากกว่า HMM และ ME เพราะ CRFs พัฒนามาจาก HMMs และ MEMMs (Maximum Entropy Markov Models) MEMMs เป็นแบบจำลองที่รวมคุณสมบัติของ HMMs และ ME ไว้ด้วยกัน ดังนั้น CRFs จึงสามารถแก้จุดบกพร่องของ HMMs และ MEMMs ได้ (Lafferty, McCallum, and Pereira, 2001) และเป็นเหตุให้สามารถรู้จำชื่อเฉพาะได้มากกว่า

ในภาษาฮินดีได้มีการเปรียบเทียบแบบจำลอง CRFs กับ SVMs (Krishnarao et al., 2009) โดยแบบจำลองทั้งสองใช้คุณสมบัติอย่างเดียวกัน ผลการทดสอบพบว่า CRFs มี

ประสิทธิภาพมากกว่า เนื่องจาก SVMsมีปัญหาในการหาความสัมพันธ์แบบ state-to-state และ feature-to-state ดังนั้นประสิทธิภาพของ SVMs จึงไม่ดีนัก

2.3.3 **แนวทางแบบผสม (hybrid)** เป็นการรวมกันระหว่างวิธีการใช้กฎและวิธีการทางสถิติ หรืออาจเป็นวิธีการใช้สถิติมากกว่าหนึ่งวิธีก็ได้ ทั้งนี้เพื่อเป็นการลดข้อจำกัดของทั้งสองวิธี เช่น หากกฎไม่สามารถรู้จำชื่อเฉพาะใหม่ได้ ก็อาจใช้วิธีการทางสถิติ โดยดูจากความถี่ที่ปรากฏแทน

ตัวอย่างงานที่ใช้กฎร่วมกับวิธีการทางสถิติ เช่น งานของ Fang and Sheng (2002) ที่ศึกษาการรู้จำชื่อเฉพาะในภาษาจีน โดยนำข้อมูลที่ผ่านการตัดคำและกำกับหมวดคำไว้แล้วมาผ่านกฎทางภาษาซึ่งมีลักษณะเป็น Finite-State Cascades (FSC) คือ มีการแบ่งเป็นหลายระดับชั้น โดยข้อมูลจะผ่านกฎที่อยู่ระดับล่างขึ้นสู่ด้านบน สำหรับวิธีการทางสถิติที่ใช้ คือ bootstrapping algorithm ใช้เพื่อช่วยดึงบริบทที่สัมพันธ์กับชื่อเฉพาะและตรงกับรูปแบบที่กำหนดไว้ออกมา เช่น ชื่อบุคคลกับคำกริยา ชื่อบุคคลกับคำนำหน้าชื่อ เป็นต้น คำบริบทใดที่ปรากฏร่วมกับชื่อเฉพาะบ่อยครั้งจะนำมาสร้างเป็นรายการคำ โดยก่อนนำรายการคำไปใช้จะให้คนตรวจสอบก่อน แล้วค่อยนำไปดึงประโยคที่มีคำเหล่านี้ออกมา ประโยคที่ดึงออกมานี้จะถูกนำมาปรับเป็นกฎทางภาษา กฎที่คล้าย ๆ กันจะยุบรวมเข้าด้วยกัน จากนั้นจึงทดสอบกฎก่อนนำไปใช้สกัดชื่อเฉพาะ สำหรับกฎใหม่ที่ได้จากวิธีการทางสถิติจะนำไปเพิ่มในส่วนของกฎทางภาษา และชื่อเฉพาะที่สกัดได้จะนำไปเพิ่มให้กับคลังคำศัพท์ในส่วนของการกำกับหมวดคำ การสกัดชื่อเฉพาะโดยใช้กฎใหม่ร่วมด้วยทำให้ค่าความถูกต้องเพิ่มขึ้น 14.3% อย่างไรก็ตามงานวิจัยนี้ไม่ได้ยกตัวอย่างกฎที่ใช้รวมถึงไม่ได้บอกค่าความถูกต้องตามจริงจึงไม่สามารถวัดได้ว่าระบบมีประสิทธิภาพมากน้อยเพียงใด

ตัวอย่างงานวิจัยที่ใช้วิธีการทางสถิติมากกว่าหนึ่งวิธี เช่น งานของ Chiong (2008) ที่ศึกษาการรู้จำชื่อเฉพาะภาษาอังกฤษ ใช้แบบจำลอง Maximum Entropy Model (MEM) ร่วมกับ Hidden Markov Model (HMM) เพื่อช่วยลดปัญหาเรื่องขนาดคลังข้อมูลฝึกฝนที่ถ้าหากมีไม่เพียงพอจะมีผลต่อประสิทธิภาพของ HMM ดังนั้นในการทดสอบจะใช้ MEM ระบุตำแหน่งของชื่อเฉพาะชั่วคราวไว้ก่อน แล้วจึงให้ HMM ระบุตำแหน่งของชื่อเฉพาะอีกครั้ง สำหรับ HMM จะเน้นไปที่การตรวจสอบบริบทของชื่อเฉพาะที่ปรากฏหลายครั้งในเอกสารเดียวกัน เพื่อช่วยลดปัญหาความกำกวมของชื่อเฉพาะ โดยในขั้นตอนนี้ตำแหน่งของชื่อเฉพาะที่ MEM ระบุไว้ในตอนแรกจะนำมาใช้อ้างอิงในการตรวจสอบความผิดพลาดและปรับแก้คำตอบ จากผลการวิจัยพบว่า ระบบที่ใช้แบบจำลอง MEM ร่วมกับ HMM มีประสิทธิภาพมากกว่าระบบที่ใช้แบบจำลองอย่างใดอย่าง

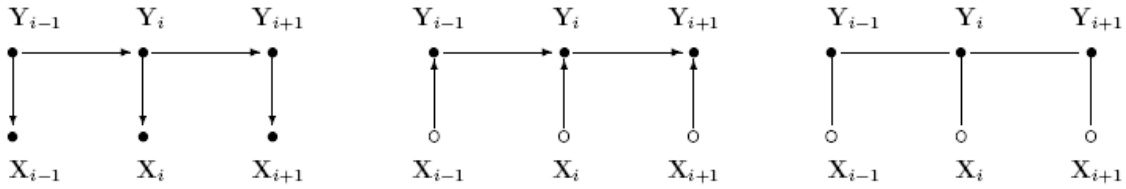
หนึ่งเพียงอย่างเดียวโดยสามารถรู้จำชื่อบุคคลและชื่อสถานที่ในข้อมูลที่เป็นเรื่องทั่วไปได้มากกว่า 90% และชื่อองค์กรมากกว่า 80% โดยไม่มีปัญหาเรื่องขนาดคลังข้อมูลฝึกฝนมีไม่เพียงพอ

จากงานวิจัยข้างต้นเป็นการใช้แนวทางผสมกับชื่อเฉพาะทุกประเภท กล่าวคือชื่อเฉพาะทุกประเภทใช้วิธีการหรือแบบจำลองเดียวกัน ยังมีบางงานวิจัยที่ใช้แนวทางผสมแต่จะแยกตามประเภทของชื่อเฉพาะ เนื่องจากเห็นว่าชื่อเฉพาะแต่ละประเภทมีลักษณะแตกต่างกัน ดังเช่นงานของ Yang et al. (2008) ใช้ language model ในการรู้จำชื่อบุคคล และใช้แบบจำลอง CRFs ในการรู้จำชื่อองค์กรและชื่อสถานที่ อีกทั้งชื่อบุคคลจะใช้ลักษณะของ character-based เพื่อหลีกเลี่ยงปัญหาที่เกิดจากการตัดคำผิด แต่ชื่อองค์กรและชื่อสถานที่จะใช้ลักษณะ word-based ผลการทดลองที่ออกมาค่อนข้างน่าพอใจ โดยค่า F-measure ของชื่อเฉพาะแต่ละประเภทได้มากกว่า 90%

## 2.4 แบบจำลองทางสถิติคอนดิชันนอลแรนดอมฟิลด์ส

แบบจำลองทางสถิติคอนดิชันนอลแรนดอมฟิลด์ส (CRFs) เป็นแบบจำลองที่ยอมรับกันในปัจจุบันว่ามีประสิทธิภาพมากกว่าแบบจำลอง Hidden Markov Models (HMMs) ซึ่งมีลักษณะเป็น Generative models ที่อาศัยค่าความน่าจะเป็นร่วม (joint probability) ระหว่างข้อมูลเข้ากับผลหรือเลเบลที่ออกมา จึงมีปัญหาว่าไม่สามารถจับความสัมพันธ์ระหว่างคุณสมบัติต่าง ๆ ที่เกี่ยวข้องกันในข้อมูลเข้าได้เพราะคุณสมบัติต่าง ๆ เป็นอิสระต่อกัน แบบจำลองที่ลดปัญหาดังกล่าวได้คือ แบบจำลอง Maximum Entropy Markov Models (MEMMs) ที่มีลักษณะเป็น Discriminative models ที่อาศัยค่าความน่าจะเป็นแบบเงื่อนไข (conditional probability) ของผลหรือสายของเลเบล (label sequence) แบบต่าง ๆ เมื่อพบสายข้อมูลเข้า (observation sequence) ซึ่งทำให้จับความสัมพันธ์ระหว่างคุณสมบัติต่าง ๆ ในข้อมูลที่พบได้แต่แบบจำลอง MEMMs ก็ยังพบปัญหาที่เรียกว่า label biased เพราะการตัดสินผล ณ สภาวะใดขึ้นกับสภาวะปัจจุบันและสายข้อมูลที่พบ (observation sequence) เท่านั้น สภาวะอื่น ๆ ทั้งหมดในแบบจำลองไม่มีผลต่อการคำนวณค่าความน่าจะเป็น แบบจำลอง CRFs ที่เสนอโดย Lafferty et al. (2001) เป็นแบบจำลองที่ลดปัญหาที่ว่านี้

เมื่อนำแบบจำลองทั้งสามมาวาดเป็นกราฟเปรียบเทียบกันจะได้ดังภาพที่ 2.1




ภาพที่ 2.1 รูปกราฟแสดงการเปรียบเทียบแบบจำลอง HMMs, MEMMs และ CRFs ของ Lafferty et al. (2001)

จากภาพ X คือ สายข้อมูลที่พบ (observation sequence) และ Y คือ สายของผลหรือเลเบล (label sequence) HMMs และ MEMMs มีลักษณะเป็นกราฟระบุทิศทาง (directed graphical model) โดยที่ HMMs มีลักษณะเป็น Generative model คือแสดงการแจกแจงความน่าจะเป็นร่วม (joint probability distribution) P(X, Y) สายข้อมูลที่พบของ HMMs ไม่ได้มีส่วนเกี่ยวข้องกับสายของเลเบลหรือผลลัพธ์ ดังนั้นผลลัพธ์ Y เป็นสิ่งที่มาก่อนหรือทำให้เกิดลำดับเหตุการณ์ X (Sutton and McCallum, 2007) สำหรับ MEMMs และ CRFs มีลักษณะเป็น Discriminative models สายข้อมูลที่พบของ MEMMs และ CRFs เป็นสิ่งที่ไม่ได้เกิดจากการประมวลผลของแบบจำลองแต่เป็นเงื่อนไขในการกำหนดสายของเลเบล มีลักษณะการแจกแจงความน่าจะเป็นแบบมีเงื่อนไข P(Y|X) นั่นคือสายข้อมูลที่พบ X เกิดขึ้นก่อนสายของเลเบล Y แบบจำลอง CRFs มีลักษณะเป็นแบบจำลองกราฟไม่ระบุทิศทาง (undirected graphical model) จึงต่างจาก MEMMs ที่ CRFs หาค่าความน่าจะเป็นของเลเบลถัดไปโดยนำเลเบลก่อนหน้าทั้งหมดที่มีลำดับเหตุการณ์เป็นเงื่อนไขมาคำนวณด้วย น้ำหนักของคุณสมบัติต่าง ๆ จากสภาวะที่ต่างกันจึงมีการปรับสมดุลให้ค่าไม่เอนเอียงไปสภาวะใดสภาวะหนึ่ง

เมื่ออธิบายแบบจำลอง CRFs ในรูปแบบของกราฟ (Kruengkrai et al., 2006) กำหนดให้ $\boldsymbol{y}$ เป็นกราฟห่วงโซ่ตรง (linear-chain graph)ประกอบด้วยจุด (node) $y_1$ ,...., $y_y$ ที่เชื่อมต่อกัน มีกราฟย่อย (clique) $C_i \in C$ เป็นส่วนประกอบของจุดต่าง ๆ $y_{C_i}$ ซึ่งสามารถหาค่าพารามิเตอร์ได้จากฟังก์ชันศักยภาพ (potential function) $\psi_{C_i}$ ดังนั้นการแจกแจงความน่าจะเป็น (probability distribution)ของกราฟ $\boldsymbol{y}$ จึงได้จากค่าฟังก์ชันศักยภาพของกราฟย่อยทั้งหมดดังนี้

$$p(\boldsymbol{y}) = \frac{1}{Z}\prod_{C_i \in C} \psi_{C_i}(y_{C_i})$$

โดยที่ Z คือการปรับข้อมูลให้เข้ากับบรรทัดฐาน (normalization) คำนวณได้จากสมการ

$$Z = \sum_{y} \prod_{C_i \in C} \psi_{C_i}(y_{C_i})$$

ศักยภาพกราฟย่อยสามารถเขียนในรูปสมการฟังก์ชันคุณสมบัติ (feature function) โดยใช้แนวคิดของ log-linear models ได้ดังนี้

$$\psi_{C_i}(y_{C_i}) = \prod_{k} \exp\{\lambda_k f_k(y_{C_i})\} = \exp\{\sum_{k=1}^{K} \lambda_k f_k(y_{C_i})\}$$

โดยที่ K คือจำนวนของคุณสมบัติทั้งหมด และ $\lambda_k$คือค่าน้ำหนักของฟังก์ชันคุณสมบัติย่อย $f_k$ เมื่อนำมาเขียนเป็นสมการแจกแจงความน่าจะเป็นแบบมีเงื่อนไขของแบบจำลอง CRFs แบบห่วงโซ่ตรง (linear-chain) จึงได้ดังนี้

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t)\right)$$

โดยที่ $y$ คือ ลำดับของแท็กที่เป็นผลลัพธ์

$x$ คือ ลำดับของข้อมูลเข้าหรือเหตุการณ์

$\lambda_k$ คือ ค่าน้ำหนักของฟังก์ชันคุณสมบัติ $f_k(y_{t-1}, y_t, x, t)$

$f_k(y_{t-1}, y_t, x, t)$ คือ ฟังก์ชันคุณสมบัติที่ใช้

$T$ คือ ตำแหน่งของลำดับของสภาวะที่ต่อเนื่องตั้งแต่ $t_1, \ldots, t_T$

$K$ คือ จำนวนฟังก์ชันคุณสมบัติที่นำมาหาค่าน้ำหนักในตำแหน่งของเหตุการณ์นั้น ๆ ตั้งแต่ $k_1, \ldots, k_K$

$Z(x)$ คือการปรับข้อมูลให้เข้ากับบรรทัดฐาน (normalization) คำนวณได้จากสูตร

$$Z(x) = \sum_{y} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t)\right)$$

ค่า $Z(x)$ จะขึ้นอยู่กับ $x$และค่าน้ำหนัก $\lambda$

### 2.4.1 ฟังก์ชันคุณสมบัติ

ฟังก์ชันคุณสมบัติถือเป็นองค์ประกอบสำคัญของแบบจำลอง CRF โดยมีรูปแบบดังนี้ $f_k(y_{t-1}, y_t, x, t)$ โดยที่ $y_{t-1}, y_t$ แสดงแท็กก่อนหน้าและสภาวะปัจจุบันของเหตุการณ์ $x$ คือข้อมูลเข้า และ $T$ คือ ตำแหน่งของข้อมูลเข้า ฟังก์ชันคุณสมบัติจะให้ค่าเป็นจำนวนจริง

ตัวอย่างของการกำหนดค่าของฟังก์ชันคุณสมบัติที่ให้ค่าแบบมีสองค่า คือ 0 กับ 1 เท่านั้น เช่น กำหนดให้ค่าเป็น 1 เมื่อคำปัจจุบัน ($x$) เป็น John และแท็กปัจจุบัน $y_t$ เป็นตัวกำกับชนิดของชื่อเฉพาะบุคคล (PERSON) (Zhu, 2008)

$$f_1(y_{t-1}, y_t, x, t) = \begin{cases} 1 \text{ if } y_t = \text{PERSON and } x = \text{John} \\ 0 \text{ otherwise} \end{cases}$$

(Zhu, 2008)

ค่าน้ำหนัก $\lambda_1$ ของฟังก์ชันคุณสมบัติข้างต้นจะเป็นดังนี้

หากค่าของ $\lambda_1 > 0$ เมื่อพบคำว่า John และเรากำหนดให้แท็ก $Y_t$ เท่ากับ PERSON เช่นนี้จะเป็นการเพิ่มค่าความน่าจะเป็นให้กับ $Y$ ดังนั้น CRFs จึงมีความน่าจะเป็นว่าจะกำหนดแท็ก PERSON ให้กับคำว่า John แต่ในกรณีที่ $\lambda_1 < 0$ CRFs จะหลีกเลี่ยงการใช้แท็ก PERSON กับคำว่า John และหาก $\lambda_1 = 0$ หมายถึง คุณสมบัตินี้ไม่มีผลต่อค่าความน่าจะเป็น

อีกตัวอย่าง เช่น

$$f_2(y_{t-1}, y_t, x, t) = \begin{cases} 1 \text{ if } y_t = \text{PERSON and } x_{t+1} = \text{said} \\ 0 \text{ otherwise} \end{cases}$$

(Zhu, 2008)

คุณสมบัตินี้จะให้ค่าก็ต่อเมื่อแท็กของตำแหน่งข้อมูลปัจจุบันกำกับเป็น PERSON และหน่วยคำถัดไปตรงกับคำว่า 'said' จากฟังก์ชันคุณสมบัติที่ 1 และ 2 ข้างต้น สามารถนำมาใช้กับประโยค เช่น "John said so." ได้ทั้งสองฟังก์ชัน จึงเป็นลักษณะของคุณสมบัติที่ทับซ้อนกัน (overlapping features) จึงเป็นการเพิ่มค่าความน่าจะเป็นที่จะกำหนดแท็ก $y_1$ หรือ John เป็น PERSON ให้แก่ $\lambda_1 + \lambda_2$ ลักษณะเช่นนี้เป็นสิ่งที่ HMMs ไม่สามารถทำได้ เพราะ HMMs ไม่

สามารถดูคำถัดไปได้รวมถึงไม่สามารถใช้คุณสมบัติที่มีลักษณะทับซ้อนกันได้ แต่คุณสมบัติของ CRF สามารถใช้ส่วนใดก็ได้จากข้อมูลเข้า ($x$) ทั้งหมด (Zhu, 2008)

คุณสมบัติสำหรับ CRF ไม่ได้จำกัดไว้เพียงแค่ค่าแบบ binary เท่านั้น สามารถกำหนดค่าเป็นแบบอื่นได้เช่นกัน

### 2.4.2 การฝึกฝนแบบจำลอง

การฝึกฝนแบบจำลอง คือการหาชุดค่าน้ำหนักพารามิเตอร์ $\Lambda = \{\lambda_1,\ldots,\lambda_k\}$ จากชุดข้อมูลฝึกฝนที่มีสายข้อมูลที่พบและสายของผล $\mathcal{D}=\{(x^{(1)}, y^{(1)}),\ldots(x^{(n)}, y^{(n)})\}$ โดยใช้วิธีการคาดประมาณค่าความน่าจะเป็นสูงสุด (Maximum Likelihood Estimation: MLE) ของ log-likelihood *L* และใช้ Gaussian prior ในการช่วยปรับสมดุลของข้อมูลในคลังข้อมูลฝึกฝน (smoothing) ดังนี้

$$L_\Lambda = \sum_{i=1}^{\mathrm{N}} \log p\left(y^{(i)}\middle|x^{(i)}\right) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2}$$

โดยที่ $\sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2}$ คือGaussian prior คำนวณจากค่าน้ำหนัก $\lambda_k$ และค่าความแปรปรวน (variance) $\sigma^2$ ใช้เพื่อลดปัญหา overfitting กล่าวคือ โมเดลมีความแม่นยำเฉพาะข้อมูลที่ได้ในคลังข้อมูลฝึกฝนเท่านั้นแต่เมื่อนำไปทดสอบกับข้อมูลที่ไม่เคยพบมาก่อนจะทำให้ค่าความแม่นยำลดลง

นอกจากนี้เพื่อแก้ปัญหา convex optimization จึงใช้ limited memory quasi-Newton method (L-BFGS) เป็นอัลกอริทึมช่วยในการฝึกฝน CRF สามารถเขียนเป็นสมการได้ว่า

$$\frac{\delta L}{\delta \lambda_k} = (\sum_{i=1}^{N} C_k\left(y^{(i)}, x^{(i)}\right) - (\sum_{i=1}^{N}\sum_{y} P_\lambda\left(y\middle|x^{(i)}\right) C_k(y, x^{(i)})) - \frac{\lambda_k}{\sigma^2}$$

โดยที่ $C_k(y,x)$ คือผลรวมของคุณสมบัติ $f_k$ เมื่อได้รับ $y$ และ $x$ จึงมีค่าเท่ากับ $\sum_{t=1}^{T} f_k(y_{t-1}, y_t, x, t)$

จากสมการสองส่วนแรกคือหาความต่างระหว่างค่าคาดหวังของคุณสมบัติ $f_k$ ที่ได้จากการคำนวณจริง (empirical expected value) กับค่าคาดหวังของคุณสมบัติ $f_k$ ของแบบจำลอง (model's expected value) และส่วนที่สามคืออนุพันธ์ของ Gaussian prior

## 2.5 ลักษณะของภาษาไทยที่ทำให้ยากต่อการรู้จำชื่อเฉพาะ

โครงสร้างของภาษาไทยมีลักษณะหลายประการที่ทำให้ยากต่อการรู้จำชื่อเฉพาะ (Chanlekha and Kawtrakul, 2004) ดังนี้

1. ภาษาไทยไม่มีข้อมูลบ่งบอกถึงชื่อเฉพาะ เช่นในภาษาอังกฤษที่ใช้อักษรตัวพิมพ์ใหญ่เสมอเมื่อกล่าวถึงชื่อเฉพาะ ทำให้ยากต่อการแยกแยะระหว่างชื่อเฉพาะและคำทั่วไป เช่น ฉันอยาก*เลย*ไปที่**เลย** "เลย" คำแรกเป็นคำกริยาหมายถึงเกินจุดที่กำหนด ในขณะที่ "เลย" คำที่สองเป็นชื่อของจังหวัด หรือ **สมชาย** นี่ช่าง*สมชาย*สมชื่อจริง ๆ "สมชาย" คำแรกเป็นชื่อบุคคล ในขณะที่ "สมชาย" ที่สองเป็นวลีมาจาก "สม" ซึ่งเป็นคำวิเศษณ์หมายถึง เหมาะ และ "ชาย" เป็นคำนามหมายถึง ผู้ชาย เป็นต้น นอกจากนี้ ภาษาไทยไม่มีการใช้อักขระพิเศษสำหรับชื่อเฉพาะที่ถ่ายถอดเสียงมาจากภาษาต่างประเทศ เช่น จอห์น ไมเคิล เป็นชื่อที่ถ่ายถอดเสียงโดยใช้ตัวอักษรภาษาไทยปกติ ซึ่งต่างจากภาษาญี่ปุ่นที่มีการใช้อักษรคาตาคานะ (Katakana) เพื่อบ่งบอกว่าเป็นชื่อเฉพาะที่ถ่ายถอดเสียงมา

2. ภาษาไทยไม่มีการเว้นวรรคหรือใช้อักษรพิเศษในการแบ่งคำ ทำให้มีปัญหาในการตัดแบ่งคำ ซึ่งถ้าหากตัดแบ่งคำผิดก็จะส่งผลถึงการรู้จำชื่อเฉพาะด้วยเช่น ประโยค "คุณตูนบอก" ที่ตัดคำถูกคือ คุณ-ตูน-บอก โดย "ตูน" เป็นชื่อเฉพาะประเภทบุคคล แต่หากตัดคำเป็น คุณ-ตู-นบ-อก ชื่อเฉพาะจะกลายเป็น "ตู" ซึ่งผิด

3. ลักษณะการสร้างชื่อเฉพาะไม่มีหลักเกณฑ์ที่แน่นอน สามารถสร้างขึ้นใหม่ด้วยคำใดก็ได้ ทำให้ยากต่อการสร้างกฎเช่น บริษัท กระเรียนทอง จำกัด "กระเรียนทอง" เป็นชื่อของนกชนิดหนึ่ง และได้นำมาใช้เป็นชื่อขององค์กรหรือ ชื่อสิ่งต่าง ๆ เช่น ส้ม นุ่น ฟิล์ม ก็สามารถนำมาตั้งเป็นชื่อบุคคลได้เช่นกัน นอกจากนี้ชื่อเฉพาะแต่ละประเภทยังอาจซ้ำกันได้ทำให้ยากต่อการระบุชนิดของชื่อเฉพาะ เช่น "อ่างทอง" เป็นได้ทั้งชื่อจังหวัดและชื่อบุคคล เป็นต้น

4. ลักษณะงานเขียนของภาษาไทยที่จะกล่าวถึงชื่อเต็มของชื่อเฉพาะในครั้งแรก แล้วจากนั้นเมื่อจะกล่าวถึงชื่อเฉพาะนั้นอีกจะใช้ชื่อย่อ หรือกล่าวโดยไม่มีคำบ่งชี้ ทำให้เกิดความกำกวมระหว่างชื่อเฉพาะและคำนามทั่วไปได้เช่น "แหล่งข่าวจาก**บริษัท ห้างสรรพสินค้าโรบินสัน จำกัด (มหาชน)** เปิดเผยกับ "ฐานเศรษฐกิจ" ว่า **ห้างสรรพสินค้าโรบินสัน** ได้ตัดสินใจ... สำหรับ**โรบินสัน** สาขาสีลม เปิดให้บริการมากว่า 24 ปี" เป็นต้น

## 2.6 งานวิจัยที่เกี่ยวกับการรู้จำชื่อเฉพาะในภาษาไทย

การวิจัยเกี่ยวกับการรู้จำชื่อเฉพาะในภาษาไทยเริ่มมีมาเมื่อไม่นานมานี้ โดย Charoenpornsawat et al. (1998) ใช้แนวทางการพิจารณาคุณสมบัติ (Feature-based approach) ได้แก่ คำในบริบทใกล้เคียง (context word) และการปรากฏร่วมของคำ (collocations) และใช้กฎฮิวริสติก (heuristic rule) ในการดึงชื่อเฉพาะที่เป็นไปได้ออกมาจากคลังข้อมูลที่ผ่านการตัดคำและกำกับหมวดคำแล้ว กฎฮิวริสติกแรกคือเมื่อพบคำที่ไม่มีในพจนานุกรมให้กำหนดว่าคำนั้นอาจเป็นชื่อเฉพาะโดยจะรวมคำใกล้เคียงเข้าไปด้วย อีกกฎคือใช้ความสัมพันธ์ของคำและหมวดคำ หากค่า threshold ของคำและหมวดคำต่ำกว่าที่กำหนดไว้ให้กำหนดเป็นชื่อเฉพาะที่เป็นไปได้ โดยดูคำข้างเคียงประกอบเพราะค่าของ threshold ที่น้อยอาจเกิดจากคำข้างเคียงจากนั้นนำชื่อเฉพาะที่เป็นไปได้เหล่านั้นไปรวมกับประโยคเดิมแล้วกำกับหมวดคำใหม่ แล้วจึงใช้ Winnow algorithm ในการระบุชื่อเฉพาะ ซึ่งผลที่ได้มีความถูกต้องค่อนข้างสูง คือ 92.17% อย่างไรก็ตามงานวิจัยนี้ไม่ได้กล่าวถึงปัญหาที่พบในการรู้จำชื่อเฉพาะทำให้ยากต่อการนำไปทำวิจัยต่อยอดได้

Chanlekha et al. (2002) ใช้ Statistical and heuristic rule-based model กับข้อมูลที่ผ่านการกำกับชนิดของคำ ข้อมูลฮิวริสติก (Heuristic information) จะยึดตามบริบทภายนอกและภายในของชื่อเฉพาะ เช่น คำสำคัญ หรือคำที่อยู่ใกล้กับชื่อเฉพาะ และใช้คลังชื่อเฉพาะ (NE lexicon) ในการสกัดชื่อเฉพาะ ซึ่งผลที่ออกมาปรากฏว่าใช้ได้ดีกับหนังสือประเภทนิตยสาร ในขณะที่ถ้าเป็นหนังสือพิมพ์จะให้ผลได้ไม่ค่อยดี ทั้งนี้เนื่องจากลักษณะการเขียนในหนังสือนิตยสารมีรูปแบบที่ชัดเจนกว่าในหนังสือพิมพ์ ทำให้เขียนกฎในการสกัดชื่อได้ง่าย

Chanlekha and Kawtrakul (2004) ได้ทดลองใช้ Maximum Entropy Model (ME) ร่วมกับ กฎฮิวริสติก และพจนานุกรมคำศัพท์ในการสกัดชื่อเฉพาะ คลังข้อมูลที่ใช้ผ่านการตัดคำแล้ว ในการระบุขอบเขตของชื่อเฉพาะหลายพยางค์ที่เป็นไปได้ใช้กฎฮิวริสติก พจนานุกรม และสถิติการปรากฏร่วมของคำ แล้วจึงสกัดชื่อเฉพาะออกมาโดยใช้ ME จากนั้นสกัดชื่อเฉพาะที่เหลือด้วยการนำชื่อเฉพาะที่สกัดมาแล้วไปเปรียบเทียบกับคำที่เหลือในเอกสาร ผลที่ออกมาแสดงให้เห็นว่าใช้ได้ดีกับชื่อเฉพาะประเภทบุคคล สำหรับชื่อองค์กรหากใช้ช่วงบริบทกว้างจะได้รับผลกระทบจากปัญหาข้อมูลเกิดการกระจาย (data sparseness problem) เพราะทำให้ฟังก์ชันคุณสมบัติมีมากขึ้นส่งผลให้ประสิทธิภาพในการคำนวณค่าน้ำหนักของฟังก์ชันคุณสมบัติลดลง ดังจะเห็นได้จากเมื่อใช้บริบทช่วง +/-2 คำได้ผลเพียง 77.76% เมื่อปรับให้เหลือช่วง +/-1 คำค่าความถูกต้องเพิ่มขึ้นเป็น 89.87% ส่วนชื่อสถานที่ได้ค่าความถูกต้องน้อยที่สุดเนื่องจากปรากฏในบริบทที่กำกวมกว่าชื่อเฉพาะประเภทอื่น

สุฤดี (2548) ใช้วิธีการทางสถิติ Mutual Information ร่วมกับ Localmax algorithm ในการคัดเลือกกลุ่มพยางค์ที่คาดว่าจะเป็นชื่อเฉพาะออกมา แต่การใช้วิธีการทางสถิติทำให้จำนวนกลุ่มพยางค์ที่คัดเลือกออกมามีจำนวนมาก เนื่องจากหลักทางสถิติคือการหาค่าความสัมพันธ์ระหว่างหน่วยหรือพยางค์ หากพยางค์ที่ปรากฏร่วมกันบ่อยจะมีความสัมพันธ์ระหว่างกันสูง และกลุ่มพยางค์เหล่านี้จะถูกดึงออกมา แต่ในความเป็นจริงพยางค์ที่ปรากฏร่วมกันไม่จำเป็นต้องเป็นชื่อเฉพาะเสมอไป เช่น “ปรึก-ษา” ดังนั้นค่าความแม่นยำ (precision) จึงต่ำ ในการคัดและแยกประเภทของชื่อเฉพาะใช้กฎที่สร้างขึ้นจากหลักฐานภายใน เช่น คำนำหน้าชื่อ เป็นต้น รวมถึงหลักฐานจากบริบทข้างเคียง เมื่อวัดประสิทธิภาพของกฎแล้ว พบว่ามีอัตราการรู้จำต่ำสาเหตุเพราะมีชื่อเฉพาะที่ไม่ถูกต้องผ่านกฎเข้ามาเป็นจำนวนมาก เนื่องจากความผิดพลาดในการระบุขอบเขตสิ้นสุดของชื่อเฉพาะ

# บทที่ 3

# คลังข้อมูลและการกำกับข้อมูล

คลังข้อมูลที่ใช้ในงานวิจัยเป็นคลังข้อความภาษาไทย “BEST 2009” ซึ่งเป็นชุดทดสอบกลางที่ใช้ในการทดสอบการแข่งขันวัดเปรียบเทียบสมรรถนะของซอฟต์แวร์แบ่งคำภาษาไทย จัดทำโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ร่วมกับจุฬาลงกรณ์มหาวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์ และสถาบันเทคโนโลยีนานาชาติสิรินธร คลังข้อความประกอบด้วยงานเขียน 3 ประเภท ได้แก่ ข้อความจากหนังสือประเภทนวนิยาย ข้อความจากสารานุกรมสำหรับเยาวชนไทย และข้อความจากหนังสือพิมพ์บนอินเตอร์เน็ต

สำหรับในงานวิจัย ผู้วิจัยเลือกใช้เฉพาะคลังข้อความข่าวหนังสือพิมพ์เท่านั้น เนื่องจากในบทความข่าวมีรูปแบบการเขียนที่พบได้ในชีวิตประจำวัน และมีชื่อเฉพาะครบทั้ง 3 ประเภท ได้แก่ ชื่อบุคคล ชื่อองค์กร และชื่อสถานที่ โดยชื่อเฉพาะแต่ละประเภทต้องมีไม่ต่ำกว่า 1,000 ชื่อ และเมื่อรวมทั้งหมดแล้วต้องมีไม่ต่ำกว่า 5,000 ชื่อ

แต่ทั้งนี้ เนื่องจากคลังข้อความภาษาไทย “BEST 2009” ใช้เป็นชุดทดสอบกลางในการทดสอบสมรรถนะของซอฟท์แวร์แบ่งคำ คลังข้อความจึงมีการแบ่งคำไว้แล้ว รวมถึงการใส่เครื่องหมายกำหนดขอบเขตอักษรย่อและชื่อเฉพาะไว้ด้วย เพื่อไม่ให้มีการแบ่งคำในส่วนของอักษรย่อและชื่อเฉพาะ สำหรับงานวิจัยนี้ต้องใช้ข้อมูลจำนวน 2 ชุด คือ ชุดข้อมูลที่ผ่านการตัดคำและชุดข้อมูลที่ผ่านการตัดพยางค์ ผู้วิจัยจึงนำเครื่องหมายกำหนดขอบเขตคำที่อยู่ในคลังข้อความ “BEST 2009” ออก แต่ยังคงเครื่องหมายกำหนดขอบเขตอักษรย่อและชื่อเฉพาะไว้ จากนั้นผู้วิจัยจะใช้โปรแกรมที่สามารถตัดแบ่งได้ทั้งคำและพยางค์ตัดแบ่งข้อมูล โปรแกรมที่ใช้คือ Thaiseg version 2.01 ของภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย โดยการตัดพยางค์ยึดค่าสถิติไตรแกรมเป็นหลัก และการตัดคำใช้หลักการปรากฏร่วมกันของพยางค์ (Aroonmanakun, 2002) ในการตัดพยางค์จะใช้พยางค์ของภาษาเขียนเป็นหลัก ซึ่งจะต่างจากพยางค์ในแง่ของเสียง กล่าวคือ พยางค์ที่มีเสียงอะกึ่งเสียง ไม่ปรากฏรูปของสระ จะไม่ตัดแยกออกมาเป็น 1 พยางค์แต่จะนำไปรวมเข้ากับพยางค์ถัดไป เพราะหากตัดตามแบบการนับเสียงจะได้พยางค์รูปตัวเขียนที่เหลือเพียงแค่ตัวอักษรเท่านั้น เช่น “สหกรณ์” หากเป็นในแง่ของเสียงจะมี 3 พยางค์ คือ ส-ห-กรณ์ (สะ-หะ-กอน) แต่หากตัดในรูปตัวเขียนจะเหลือเพียง 2 พยางค์เท่านั้น คือ สห-กรณ์ เป็นต้น นอกจากนี้คำบางคำยังใช้ตัวอักษรตัวเดียวกันเป็นทั้งเสียงตัวสะกดและเสียงพยัญชนะต้นด้วย ซึ่งหากใช้หลักการตัดพยางค์แบบเสียงจะมีปัญหาว่าต้องเพิ่มพยางค์เข้ามาอีก

ทั้งจะทำให้รูปตัวเขียนผิดไปจากข้อมูลต้นฉบับด้วย เช่น “ปัทมา” (ปัด-ทะ-มา) เมื่อตัดพยางค์รูปตัวเขียนแล้วจะได้ 2 พยางค์เท่านั้น คือ ปัท-มา

เนื่องจากคลังข้อมูลที่จะใช้ในงานวิจัยนี้เป็นคลังข้อมูลแบบ Supervised Learning คือ ให้โมเดลทางสถิติได้เรียนรู้จากข้อมูลที่มีการให้คำตอบไว้แล้ว ผู้วิจัยจึงต้องกำกับชนิดของชื่อเฉพาะเพิ่มเข้าไปในคลังข้อมูล โดยหลักเกณฑ์การจัดประเภทของชื่อเฉพาะ ผู้วิจัยได้ใช้ความรู้ภาษาแม่ของผู้วิจัยและ Simple Named Entity Guidelines Version 6.4-Thai (Phanarangsan et al., 2006) เป็นแนวทางประกอบ สำหรับเครื่องหมายกำหนดขอบเขตและชนิดของชื่อเฉพาะที่ผู้วิจัยใช้มี 5 ชนิด ได้แก่

1. <persName>…..</persName> ใช้กำกับชื่อเฉพาะประเภทบุคคล ได้แก่ ชื่อ นามสกุล ชื่อเล่น ฉายา นามปากกา นามแฝง รวมถึงชื่อของตัวละครในนวนิยาย บทละคร และภาพยนตร์ ทั้งนี้ภายในเครื่องหมายจะรวมคำนำหน้าชื่อต่าง ๆ ทั้งจากตำแหน่งหน้าที่การงาน การศึกษาหรืออาชีพ เช่น นาย นางสาว ด.ญ. ม.ล. นายแพทย์ ดร. รศ. พล.ต.ต. ฯลฯ คำเรียกญาติ เช่น พี่ น้า ป้า เป็นต้น เอาไว้ด้วย โดยจะกำกับชื่อและนามสกุลรวมเข้าไว้ด้วยกัน เช่น

ตามที่<persName>*นางประนอม ทองจันทร์*</persName> กับ <persName><abb>*ด.ช.*</abb>*กิตติพงษ์ แหลมผักแว่น*</persName> และ <persName><abb>*ด.ญ.*</abb>*กาญจนา กรองแก้ว*</persName> ป่วยสงสัยติดเชื้อไข้

หากชื่อที่ปรากฏเป็นชื่อเล่นตามด้วยชื่อและนามสกุล จะกำกับชื่อเล่นและชื่อนามสกุลแยกกัน เช่น

นำโดย <persName>*นิโคล*</persName>และ<persName>*แมว*</persName> <persName>*จีระศักดิ์ ปานพุ่ม*</persName>

กรณีที่ชื่อของชาวต่างประเทศสะกดเป็นภาษาอื่น เช่น ภาษาอังกฤษ จะไม่กำกับให้เป็นชื่อเฉพาะ เนื่องจากงานวิจัยนี้จำกัดขอบเขตไว้ที่การรู้จำชื่อเฉพาะภาษาไทยเท่านั้น แต่หากมีการถ่ายถอดเสียงชื่อออกมาเป็นภาษาไทยจะกำกับให้เป็นชื่อเฉพาะ เช่น

<persName><abb>*ศ.*</abb>*เดวิท สเวิร์น*</persName> ผู้เชี่ยวชาญไข้หวัดนกจาก<placeName>สหรัฐอเมริกา</placeName>

กรณีที่ชื่อบุคคลไปปรากฏเป็นส่วนหนึ่งของชื่อเฉพาะประเภทอื่นจะไม่กำกับชื่อบุคคลนั้นแยกต่างหาก จะกำกับเพียงชื่อเฉพาะหลักเท่านั้น เช่น *สำนักงานกฎหมายบุญมาและเพื่อน*

“บุญมา” เป็นชื่อเฉพาะประเภทบุคคลที่เป็นส่วนหนึ่งของชื่อองค์กร กรณีนี้จะไม่กำกับ “บุญมา” แต่จะกำกับทั้งวลี “สำนักงานกฎหมายบุญมาและเพื่อน” ให้เป็นชื่อเฉพาะประเภทองค์กร

2. <orgName>.....</orgName> ใช้กำกับชื่อเฉพาะประเภทองค์กร ชื่อเฉพาะประเภทองค์กร หมายถึง ชื่อของนิติบุคคล กลุ่มบุคคล หรือกลุ่มที่มีการรวมตัวกันขึ้น โดยมีวัตถุประสงค์หรือเป้าหมายต่าง ๆ ร่วมกัน มีระบบการบริหาร การจัดการ และการดำเนินกิจกรรมต่าง ๆ ทั้งภายในและภายนอกองค์กร เช่น สามารถรับสมัครพนักงาน ออกประกาศหรือทำนิติกรรมต่าง ๆ ได้ ทั้งนี้ชื่อเฉพาะประเภทองค์กรยังรวมถึงชื่อของอาคาร สิ่งปลูกสร้าง ที่สามารถดำเนินการหรือทำกิจกรรมต่าง ๆ ได้เหมือนองค์กร เช่น โรงเรียน มหาวิทยาลัย เป็นต้น โดยจะกำกับรวมส่วนที่เป็นคำบ่งชี้ชื่อองค์กร เช่น บริษัท...จำกัด สมาคม กรม เป็นต้น ไว้ในเครื่องหมายด้วย เช่น

มีการประชุมหารือกันระหว่าง<orgName>*กรมปศุสัตว์*</orgName>กับ <orgName>*คณะสัตวแพทย์*</orgName> จาก<orgName>*จุฬาลงกรณ์มหาวิทยาลัย* </orgName>

กรณีที่ชื่อเฉพาะเป็นชื่อเต็มขององค์กรและตามด้วยชื่อย่ออยู่ในวงเล็บให้กำกับแยกกัน แต่หากชื่อเฉพาะนั้นเป็นภาษาอื่น ไม่ได้ถ่ายถอดเสียงออกมาเป็นภาษาไทยจะไม่กำกับให้เป็นชื่อเฉพาะเช่นเดียวกับชื่อบุคคล เช่น

<orgName>*กระทรวงสาธารณสุข*</orgName> (<orgName><abb>*สธ.*</abb> </orgName>)

<orgName>*องค์การสุขภาพสัตว์โลก*</orgName> (<abb>OIE</abb>) และ <orgName>*องค์การอนามัยโลก*</orgName> (<abb>WHO</abb>)

ตรวจสอบการทำงานของ<orgName><abb>*ดีเอสไอ*</abb></orgName>

บางครั้งชื่อองค์กรมักมีคำขยายชี้เฉพาะเจาะจงสถานที่ตั้งขององค์กรต่อท้ายชื่อ เนื่องจากองค์กรนั้นมีสาขาย่อยหลายสาขา เช่น มหาวิทยาลัยเกษตรศาสตร์มีหลายวิทยาเขต หากต้องการระบุสถานที่ที่เจาะจงก็จะมีชื่อวิทยาเขตต่อท้าย เช่น มหาวิทยาลัยเกษตรศาสตร์กำแพงแสน มหาวิทยาลัยเกษตรศาสตร์บางเขน เป็นต้น ชื่อเฉพาะในรูปแบบนี้จะกำกับชื่อองค์กรตามด้วยชื่อสถานที่ เช่น

<orgName>*มหาวิทยาลัยเกษตรศาสตร์*</orgName><placeName>*กำแพงแสน*</placeName>

กรณีที่ชื่อเฉพาะเขียนย่อด้วยการใช้เครื่องหมายไปยาลน้อย (ฯ) ให้กำกับรวมเครื่องหมายไปยาลน้อยไว้ด้วย โดยไม่ถือว่าชื่อเฉพาะนั้นเป็นชื่อย่อ เช่น <orgName>*กระทรวงเกษตรฯ*</orgName> เป็นต้น นอกจากนี้ กรณีที่ชื่อเฉพาะไปปรากฏเป็นส่วนหนึ่งของชื่อตำแหน่งงาน เช่น รัฐมนตรีว่าการกระทรวงการคลัง อธิบดีกรมควบคุมโรค เป็นต้น ผู้วิจัยจะกำกับให้ชื่อองค์กรเหล่านี้เป็นชื่อเฉพาะด้วย เนื่องจากเมื่อวิเคราะห์โครงสร้างของชื่อตำแหน่งงานแล้วจะเห็นว่าประกอบด้วยตำแหน่ง+หน่วยงานที่สังกัด ดังนั้นจึงกำกับให้ส่วนที่เป็นหน่วยงานที่สังกัดเป็นชื่อองค์กร เช่น

รัฐมนตรีว่าการ<orgName>*กระทรวงการคลัง*</orgName>
อธิบดี<orgName>*กรมควบคุมโรค*</orgName>

3. <placeName>..... </placeName> ใช้กำกับชื่อสถานที่ ชื่อเฉพาะประเภทสถานที่หมายถึง ชื่อของบริเวณหรือพื้นที่ทางภูมิศาสตร์ที่แน่นอน เช่น เมือง จังหวัด ถนน สิ่งที่เกิดขึ้นเองตามธรรมชาติ เช่น แม่น้ำ ภูเขา เกาะ ทะเล หรือแม้แต่สิ่งที่มนุษย์สร้างขึ้นโดยไม่ได้มีจุดมุ่งหมายเพื่อใช้ในการดำเนินงาน เช่น สวนสาธารณะ อนุสาวรีย์ เป็นต้น โดยการกำกับจะรวมส่วนที่เป็นคำบ่งชี้ เช่น ตำบล จังหวัด แม่น้ำ มหาสมุทร อ. เป็นต้น เอาไว้ด้วย แต่กรณีที่ชื่อสถานที่เป็นส่วนหนึ่งของชื่อเหตุการณ์ ชื่องาน ชื่อมหกรรมอื่น ๆ จะไม่กำกับให้เป็นชื่อสถานที่ เช่น งานมหกรรมกินไก่ไทย เป็นต้น ตัวอย่างของการกำกับชื่อเฉพาะสถานที่ เช่น

ที่ <placeName><abb>จ.</abb>*นครปฐม*</placeName>

นอกจากชื่อเฉพาะสถานที่โดยทั่วไปแล้ว ยังรวมถึงฉายาของสถานที่ที่เป็นที่รู้จักโดยทั่วไปด้วย เช่น ฉายาของประเทศต่าง ๆ เช่น “แดนปลาดิบ” หมายถึงประเทศญี่ปุ่น หรือ “เมืองกระทิงดุ” หมายถึงประเทศสเปน เป็นต้น

กรณีที่ชื่อเฉพาะนำไปใช้เป็นคำขยายคำนามหรือคำอุปสรรคที่หมายถึง คน กลุ่มคน เช่นคำว่า “คน” หรือ “ชาว-” ในคำว่า คนไทย ชาวลพบุรี เป็นต้น เช่นนี้ให้กำกับคำขยายนี้เป็นชื่อเฉพาะด้วย เพราะเมื่อวิเคราะห์ทางด้านความหมายแล้ว คำขยายนี้เป็นคำบอกแหล่งที่มาหรือสถานที่ที่คนหรือกลุ่มคนเหล่านั้นอยู่ เช่น ชาวลพบุรี หมายถึง คนหรือกลุ่มคนที่อาศัยอยู่ในจังหวัดลพบุรีหรือมาจากจังหวัดลพบุรี เป็นต้น ตัวอย่างการกำกับคำเหล่านี้ เช่น

คน<placeName>*ไทย*</placeName>

สำหรับอาการผู้ป่วย ชาว<placeName>*จังหวัดนครราชสีมา*</placeName>.....

4. <orgName ref="loc">.....</orgName> ใช้กำกับชื่อเฉพาะประเภทองค์กรที่นำมาใช้อ้างข้ามประเภทเป็นชื่อสถานที่ ชื่อเฉพาะประเภทนี้หมายถึง ชื่อองค์กรที่เมื่อปรากฏอยู่ในข้อความแล้วไม่ได้มีความหมายถึง กลุ่มบุคคล หรือนิติบุคคล แต่หมายถึงสถานที่ตั้งขององค์กรนั้น ๆ โดยปกติสิ่งที่มีบทบาทสำคัญในการทำให้ชื่อเฉพาะเกิดการอ้างข้ามประเภทคือบริบทโดยรอบ เช่น

เข้ามาชันสูตรโรคและรักษาโรค**ที่**<orgName ref="loc">*โรงพยาบาลสัตว์*</orgName><orgName>*มหาวิทยาลัยเกษตรศาสตร์*</orgName><placeName>*บางเขน*</placeName>

จากตัวอย่าง จริง ๆ แล้วโรงพยาบาลสัตว์มีหน้าที่เป็นองค์กรแต่เมื่ออยู่ในบริบทนี้ทำหน้าที่เป็นสถานที่ เพราะเมื่อพิจารณาจากความหมาย "โรงพยาบาลสัตว์" ในที่นี้ไม่ได้แสดงถึงกิจกรรมหรือทำหน้าที่ที่เกี่ยวกับองค์กร อีกทั้งเป็นชื่อที่อยู่ตามหลังคำบุพบท "ที่" ส่วนใหญ่ชื่อที่ตามหลัง "ที่" มักเป็นสถานที่ ดังนั้นในบริบทนี้ "โรงพยาบาลสัตว์" จึงกำกับให้เป็นชื่อองค์กรที่อ้างถึงสถานที่

5. <placeName ref="org">..... </placeName> ใช้กำกับชื่อเฉพาะประเภทสถานที่ที่นำมาใช้อ้างข้ามประเภทเป็นชื่อองค์กร บางครั้งใช้เพื่อหมายถึงองค์กรที่ตั้งขึ้นในบริเวณหรือพื้นที่นั้น ๆ กล่าวคือในบางบริบทมีการนำคำกริยาแสดงอาการหรือการกระทำมาใช้กับสถานที่ โดยปกติคำกริยาเหล่านี้จะใช้ร่วมกับบุคคลหรือองค์กร เพราะบุคคลหรือองค์กรสามารถแสดงอาการหรือดำเนินการต่าง ๆ ได้ แต่เมื่อนำมาใช้กับสถานที่ จึงหมายความว่า สถานที่นั้น ๆ มีลักษณะเหมือนเป็นองค์กร ๆ หนึ่ง เพราะสามารถทำการต่าง ๆ ได้เช่นเดียวกับบุคคลหรือองค์กร เช่น

<placeName ref="org">*ประเทศไทย*</placeName>ได้แจ้งไปยังองค์การระหว่างประเทศเรียบร้อยแล้ว

จากตัวอย่าง จะเห็นว่ามีการใช้คำกริยา "แจ้ง" กับ "ประเทศไทย" ซึ่งเป็นสถานที่ ดังนั้นในบริบทนี้ จึงกำกับให้ "ประเทศไทย" เป็นชื่อสถานที่ที่อ้างข้ามประเภทเป็นชื่อองค์กร เพราะสามารถดำเนินการหรือกระทำการต่าง ๆ ได้เช่นเดียวกับองค์กร

นอกจากเครื่องหมายกำหนดขอบเขตและชนิดของชื่อเฉพาะแล้ว ผู้วิจัยยังได้กำหนดเครื่องหมายกำกับส่วนที่เป็นอักษรย่อ คำย่อต่าง ๆ คือ <abb>.....</abb> โดยใช้กำกับส่วนที่เป็นอักษรย่อ และคำย่อทั้งหมด รวมถึงชื่อย่อของชื่อเฉพาะด้วย โดยในการกำกับชื่อเฉพาะนั้น ให้กำกับซ้อนไว้ด้านในของเครื่องหมายกำหนดขอบเขตและชนิดของชื่อเฉพาะ เช่น

ช่วงต้นเดือน <abb>*ก.พ.*</abb> (ก.พ. = กุมภาพันธ์)

รัฐบาลได้มอบให้ <orgName><abb>*สธ.*</abb></orgName>เป็นตัวหลัก (สธ. = กระทรวงสาธารณสุข)

<persName><abb>*ศ.*</abb><abb>*ดร.*</abb>ภักดี โพธิศิริ</persName>

ในส่วนของชื่อเฉพาะ แม้ชื่อนั้นจะสะกดเป็นภาษาต่างประเทศก็ยังคงกำกับด้วยเครื่องหมายกำกับอักษรย่อเพียงแต่จะไม่กำกับชนิดของชื่อเฉพาะนั้น ซึ่งจะต่างจากชื่อที่มีการถ่ายถอดเสียงออกมาเป็นภาษาไทยที่จะกำกับทั้งชนิดและอักษรย่อ เช่น

หลังจากนักวิทยาศาสตร์ของ <abb>*CSIRO*</abb> แถลงความสำเร็จ.....

เป็นการยืนยันถ้อยแถลง<orgName><abb>*เอฟเอโอ*</abb></orgName>

กรณีที่ชื่อเฉพาะเป็นคำ acronym คือ คำที่เกิดจากการนำอักษรต้นตัวแรกหรือสองสามตัวแรกในคำมารวมกันเป็นคำเดียว แล้วอ่านออกเสียงเป็นคำใหม่ เช่น นาโต้ มาจาก NATO (The North Atlantic Treaty Organization) ซึ่งจะต่างกับชื่อย่อที่จะอ่านออกเสียงเรียงตามตัวอักษร ดังนั้นคำ acronym เหล่านี้จึงไม่จัดให้เป็นชื่อย่อ เพราะเปรียบได้กับเป็นคำใหม่คำหนึ่ง ตัวอย่างการกำกับคำ acronym เช่น

คำแถลงของเจ้าหน้าที่<orgName>*องค์การอนามัยโลก*</orgName>(<orgName>*ฮู*</orgName>)

จากตัวอย่างข้างต้น “ฮู” มาจาก WHO (The World Health Organization) จึงกำกับให้เป็นชื่อเฉพาะอีกชื่อหนึ่ง

เมื่อนำข้อมูลทั้งหมดที่กำกับชื่อเฉพาะแล้วผ่านโปรแกรมนับจำนวนชื่อและโปรแกรมตัดคำและพยางค์ ได้ผลดังนี้ คลังข้อมูลมีจำนวนคำทั้งหมด 367,673 คำ และจำนวนพยางค์ทั้งหมด 487,364 พยางค์ มีชื่อเฉพาะทั้งหมด 16,179 ชื่อ ประกอบด้วยชื่อบุคคล 5,672 ชื่อ ชื่อองค์กร

4,751 ชื่อ ชื่อสถานที่ 3,934 ชื่อ ชื่อองค์กรที่ใช้อ้างถึงสถานที่ 417 ชื่อ และชื่อสถานที่ที่ใช้อ้างถึงองค์กร 1,405 ชื่อ

# บทที่ 4

# ระบบการรู้จำชื่อเฉพาะภาษาไทย

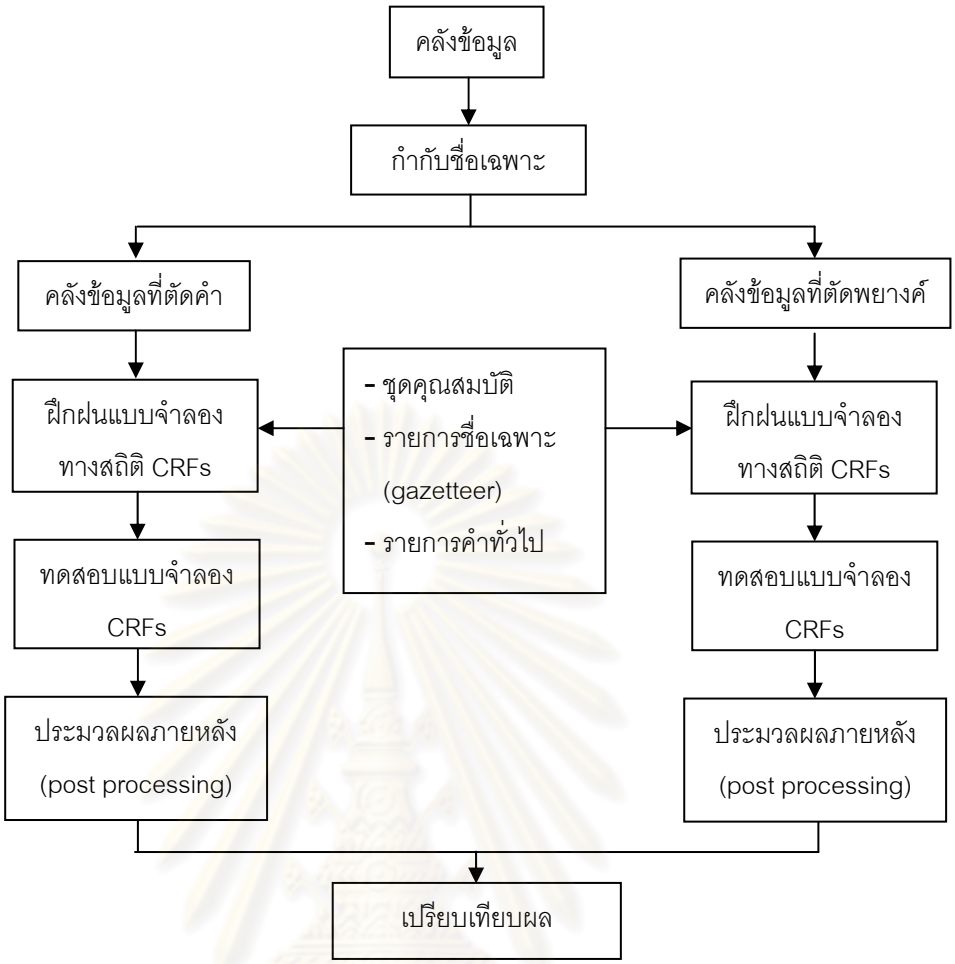
ในส่วนนี้ ผู้วิจัยจะกล่าวถึงขั้นตอนต่าง ๆ ของระบบการรู้จำชื่อเฉพาะในงานวิจัยนี้ ได้แก่ การนำข้อมูลมาฝึกฝนและทดสอบด้วยแบบจำลองคอนดิชันนอลแรนดอมฟิลด์ส และการประมวลผลภายหลังจากนั้นจะเปรียบเทียบประสิทธิภาพของระบบการรู้จำชื่อเฉพาะระหว่างแบบจำลองที่รับข้อมูลเข้าเป็นพยางค์กับที่รับข้อมูลเข้าเป็นคำ รวมถึงวิเคราะห์ลักษณะทางภาษาศาสตร์ที่มีผลต่อประสิทธิภาพของระบบทั้งสอง

## 4.1 ระบบการรู้จำชื่อเฉพาะ

ระบบการรู้จำชื่อเฉพาะ คือ ระบบที่ใช้สกัดส่วนที่เป็นชื่อเฉพาะออกจากข้อมูล โดยชื่อเฉพาะในงานวิจัยนี้หมายถึงชื่อบุคคล ชื่อองค์กร และชื่อสถานที่และใช้วิธีการทางสถิติหรือแบบจำลองในการสกัดชื่อเฉพาะ ในการเรียนรู้ของแบบจำลองจำเป็นต้องใช้คลังข้อมูลสำหรับฝึกฝน ซึ่งประกอบด้วยหน่วยคำหรือพยางค์ (token) หรือกล่าวอีกนัยหนึ่งคือข้อมูลที่ผ่านโปรแกรมตัดคำและพยางค์แล้ว คุณสมบัติต่าง ๆ ซึ่งใช้เพื่อช่วยเพิ่มข้อมูลเกี่ยวกับลักษณะของชื่อเฉพาะแก่ระบบ และคำตอบที่บอกว่าหน่วยคำหรือพยางค์ใดเป็นชื่อเฉพาะและเป็นชื่อเฉพาะชนิดใด สำหรับการทดสอบจะใช้คลังข้อมูลที่มีลักษณะคล้ายกับคลังข้อมูลฝึกฝนแต่จะไม่มีคำตอบให้ และผลลัพธ์ที่ได้จากระบบคือคลังข้อมูลทดสอบที่มีคำตอบเพิ่มเข้ามา อยู่ในคอลัมน์สุดท้าย (ดูภาคผนวก ก ข และ ค ซึ่งแสดงตัวอย่างข้อมูลสำหรับฝึกฝน ข้อมูลสำหรับทดสอบ และผลลัพธ์ที่ได้จากแบบจำลอง) ดังนั้นสิ่งที่จะต้องเตรียมให้แก่ระบบคือ คลังข้อมูลสำหรับฝึกฝนและทดสอบ กำหนดคุณสมบัติที่จะใช้ และรูปแบบของคำตอบที่ต้องการ สำหรับแบบจำลองที่ใช้ในงานวิจัยนี้คือ CRF++ เวอร์ชัน 0.53 พัฒนาขึ้นโดย Taku Kudo สามารถดาวน์โหลดมาใช้เพื่องานวิจัยโดยไม่เสียค่าใช้จ่ายได้จากเว็บไซต์ http://crfpp.sourceforge.net/

ขั้นตอนของระบบการรู้จำชื่อเฉพาะในงานวิจัยนี้ แบ่งออกเป็น 3 ขั้นตอนหลักด้วยกัน คือ

1. เตรียมคลังข้อมูลสำหรับฝึกฝนและทดสอบแบบจำลองคอนดิชันนอลแรนดอมฟิลด์ส
2. ฝึกฝนและทดสอบแบบจำลองคอนดิชันนอลแรนดอมฟิลด์ส
3. นำข้อมูลที่ผ่านการประมวลผลแล้วมาผ่านกฎที่เขียนขึ้นเอง (post processing)




ภาพที่ 4.1 กระบวนการรู้จำชื่อเฉพาะ

คลังข้อมูลที่นำมาใช้ในการฝึกฝนและทดสอบระบบการรู้จำชื่อเฉพาะแบ่งออกเป็น 2 ส่วน คือ ข้อมูล 90% ใช้ในการฝึกฝน และข้อมูล 10% ใช้ในการทดสอบซึ่งจะทดสอบทั้งหมด 10 ครั้ง โดยจะแบ่งให้ข้อมูลทุกส่วนได้ใช้ในการทดสอบ

ในการฝึกฝนให้แบบจำลองสามารถเรียนรู้เกี่ยวกับชื่อเฉพาะได้ จำเป็นต้องกำหนดคุณสมบัติ (feature) ต่าง ๆ ในคลังข้อมูลสำหรับฝึกเพื่อช่วยให้แบบจำลองได้เรียนรู้ว่าลักษณะแบบใดจึงจะเป็นชื่อเฉพาะได้ ดังนั้นประสิทธิภาพของแบบจำลองจึงขึ้นอยู่กับการกำหนดคุณสมบัติเป็นหลัก

คุณสมบัติที่กล่าวถึงนั้นคือข้อมูลต่าง ๆ ที่เพิ่มให้แก่แบบจำลอง เป็นข้อมูลชนิดใดก็ได้ ไม่มีลักษณะตายตัว แต่สามารถช่วยให้แบบจำลองแยกชื่อเฉพาะออกจากคำทั่วไป หรือแยกประเภทของชื่อเฉพาะชนิดต่าง ๆ ได้ เพราะการใช้แค่แบบจำลองอย่างเดียวไม่เพียงพอต่อการนำไปใช้สกัดชื่อเฉพาะ ตัวอย่างข้อมูลที่นำมาใช้เป็นคุณสมบัติ เช่น ลักษณะการปรากฏของชื่อเฉพาะ เช่น ชื่อบุคคลมักปรากฏร่วมกับคำนำหน้าชื่อ และส่วนใหญ่คำนำหน้าชื่อมักไม่ซ้ำกับคำทั่วไป จึงถือเป็น

ชุดคำที่มีเอกลักษณ์ สามารถช่วยในการระบุขอบเขตและประเภทของชื่อเฉพาะได้ดังนั้นคำนำหน้าชื่อจึงสามารถนำมาใช้เป็นคุณสมบัติหนึ่งของแบบจำลองได้ หรือในภาษาอังกฤษ ชื่อเฉพาะมักขึ้นต้นด้วยตัวพิมพ์ใหญ่ ลักษณะเช่นนี้ก็สามารถนำมาเป็นคุณสมบัติได้เช่นกัน

การใช้คุณสมบัติในแบบจำลอง CRFs นั้นจะใช้หลังจากที่ข้อมูลผ่านการตัดคำหรือพยางค์แล้ว โดยหน่วยคำหรือพยางค์นี้จะเรียกแทนว่า token ในแบบจำลอง คุณสมบัติที่จะใช้จำเป็นต้องมีการกำหนดค่าก่อนว่าจะเป็นแบบมีสองค่า (binary) เช่น เป็น “1” หรือ “Y”เมื่อ token นั้นมีคุณสมบัติตรงตามที่ต้องการ และเป็น “0” หรือ “N” เมื่อไม่มีคุณสมบัติตามที่ต้องการหรือเป็นแบบให้ข้อมูลที่มีหลายค่า เช่น หมวดคำ (part of speech) เป็นต้น ในการเทียบคุณสมบัติจะนำแต่ละ token มาพิจารณาว่าตรงกับข้อกำหนดของคุณสมบัติหรือไม่ เช่น กำหนดให้คุณสมบัติคำนำหน้าชื่อบุคคลเป็นแบบมีสองค่า กล่าวคือ ถ้า token ที่นำมาเทียบพบในรายการคำนำหน้าชื่อบุคคลจะมีค่าเป็น “Y” แต่ถ้าไม่พบจะมีค่าเป็น “N” ดังนั้นหาก token นั้นคือ “นาย” จะมีค่าเป็น “Y” เป็นต้น

### 4.1.1 ประเภทของคุณสมบัติ

แม้ว่าระบบการรู้จำชื่อเฉพาะในงานวิจัยนี้จะแยกเป็นระบบที่ใช้ข้อมูลแบบตัดคำและระบบที่ใช้ข้อมูลแบบตัดพยางค์ แต่คุณสมบัติที่ใช้กับทั้งสองระบบนี้ผู้วิจัยจะควบคุมไม่ให้แตกต่างกัน เพราะหากใช้คุณสมบัติที่แตกต่างกันหรือจำนวนไม่เท่ากันแล้วจะทำให้ไม่สามารถเปรียบเทียบได้ว่าระบบใดมีประสิทธิภาพมากกว่ากัน คุณสมบัติต่าง ๆ ที่ใช้มีดังนี้

**4.1.1.1 รายการชื่อเฉพาะประเภทต่าง ๆ (gazetteer)** ถือเป็นฐานข้อมูลสำคัญในงานรู้จำชื่อเฉพาะ เพราะมีส่วนช่วยในการระบุขอบเขตและประเภทของชื่อเฉพาะได้ค่อนข้างมาก เช่น กรณีที่ชื่อเฉพาะปรากฏโดยไม่มีคำบ่งชี้ เช่น ชื่อองค์กรปรากฏโดยไม่มีคำว่า “บริษัท” ด้านหน้า หากชื่อองค์กรนั้นมีอยู่ในรายการชื่อเฉพาะ คุณสมบัตินี้จะมีส่วนช่วยให้ระบบรู้จำชื่อเฉพาะนั้น ๆ ได้ รายการชื่อเฉพาะทั้งหมดที่ใช้เป็นคุณสมบัติได้รับความอนุเคราะห์จากศูนย์วิจัยการประมวลผลภาษาและวัจนะ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย มีจำนวนทั้งสิ้น 39,477 ชื่อ รายละเอียดของชื่อเฉพาะประเภทต่าง ๆ ได้แสดงไว้ในตารางที่ 4.1

ตารางที่ 4.1 รายการชื่อเฉพาะประเภทต่าง ๆ ที่ใช้เป็นคุณสมบัติสำหรับแบบจำลองคอนดิชันนอลแรนดอมฟิลด์ส

| รายการชื่อเฉพาะ | จำนวนชื่อ |
|---|---|
| *ชื่อเฉพาะประเภทบุคคล* | |
| • ชื่อบุคคล | 14,683 |
| • นามสกุล | 18,578 |
| • คำนำหน้าชื่อเช่น ชื่อยศ ตำแหน่ง คำเรียกญาติ เป็นต้น | 117 |
| *ชื่อเฉพาะประเภทองค์กร* | |
| • ชื่อองค์กร | 3,150 |
| • ชื่อย่อองค์กร | 297 |
| *ชื่อเฉพาะประเภทสถานที่* | |
| • ชื่อสถานที่ | 2,635 |
| • คำบ่งชี้ชื่อสถานที่ เช่น คำว่า “ประเทศ”“เมือง” เป็นต้น | 17 |

โดยปกติ รายการชื่อเฉพาะมักนำมาใช้เพื่อช่วยสกัดชื่อเฉพาะ กรณีที่คลังข้อมูลสำหรับฝึกฝนมีไม่เพียงพอเพราะหากคลังข้อมูลมีน้อย ระบบก็จะเรียนรู้ข้อมูลเกี่ยวกับชื่อเฉพาะได้น้อยตามไปด้วย และจะมีผลต่อประสิทธิภาพของระบบ นอกจากนี้รายการชื่อเฉพาะยังสามารถช่วยสกัดชื่อเฉพาะที่ปรากฏไม่บ่อยครั้งหรือเกิดในบริบทที่กำกวมในคลังข้อมูลได้ด้วยเช่นกัน

ในการใช้รายการชื่อเฉพาะ อาจนำรายการชื่อเฉพาะมาเทียบกับสายของ token (sequence of token) โดยตรง (Chanlekha and Kawtrakul, 2002) โดยดูว่าชื่อเฉพาะทั้งชื่อตรงกับสายของ token หรือไม่ สำหรับงานวิจัยนี้จะนำรายการชื่อเฉพาะทั้งหมดยกเว้นรายการคำนำหน้าชื่อ และชื่อย่อองค์กรมาผ่านโปรแกรมตัดคำและพยางค์ เพื่อแยกชื่อเฉพาะออกเป็นส่วนต้น ส่วนกลาง และส่วนท้าย และนำส่วนต่าง ๆ ที่ได้นี้มาสร้างเป็นรายการส่วนของชื่อ ดังนั้นชื่อเฉพาะทั้ง 3 ชนิด เมื่อแยกส่วนของชื่อแล้วจะได้รายการใหม่ของส่วนต่าง ๆ ของชื่อรวม 9 รายการ สาเหตุที่แยกส่วนเพราะชื่อเฉพาะที่อยู่ในรายการส่วนใหญ่มักเป็นชื่อเต็ม ผู้วิจัยจึงเห็นว่าหากมีการแยกส่วนของชื่อ อาจช่วยได้ในเรื่องของชื่อที่ลดรูป เช่น ในข้อมูลพบชื่อองค์กร“บริษัท

เชลล์แห่งประเทศไทย” แล้วลดรูปเหลือเพียง “เชลล์” กรณีนี้หากไม่มีการแยกส่วนของชื่อ “เชลล์” ก็จะไม่พบว่าเป็นชื่ออยู่ในรายการชื่อองค์กร เป็นต้น

ในการกำหนดคุณสมบัติ จะนำแต่ละหน่วยคำหรือพยางค์ (token) ไปเปรียบเทียบกับรายการส่วนของชื่อ 9 รายการ รายการคำนำหน้าชื่อ และรายการคำย่อชื่อองค์กร ค่าของคุณสมบัตินี้ทั้ง 11 รายการนี้กำหนดให้เป็นแบบมีสองค่า (Y, N) โดยหากว่า token นั้นพบอยู่ในรายการค่าคุณสมบัติจะเป็น Y ตัวอย่างการกำหนดค่าคุณสมบัติรายการชื่อเฉพาะให้แก่ token เป็นดังนี้

| token | คำนำหน้าชื่อ | ส่วนต้นชื่อบุคคล | ส่วนกลางชื่อบุคคล | ส่วนท้ายชื่อบุคคล | ชื่อย่อองค์กร | ส่วนต้นชื่อบุคคล | ส่วนกลางชื่อบุคคล | ส่วนท้ายชื่อบุคคล | ส่วนต้นชื่อสถานที่ | ส่วนกลางชื่อสถานที่ | ส่วนท้ายชื่อสถานที่ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| น.พ. | Y | N | N | N | N | N | N | N | N | N | N |
| จรัล | N | Y | N | N | N | N | N | N | N | N | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N |

จากตัวอย่าง “น.พ.” พบอยู่ในรายการคำนำหน้าชื่อเท่านั้น จึงมีค่าเป็น Y แต่ไม่พบอยู่ในรายการอื่นส่วน “จรัล” พบอยู่ในส่วนต้นชื่อบุคคลเท่านั้น และ <s> (ช่องว่าง) ไม่พบอยู่ในรายการใด

4.1.1.2 **คุณสมบัติคำย่อ** เนื่องจากในบทความทั่วไปหรือข่าวส่วนใหญ่ ผู้วิจัยสังเกตว่าชื่อเฉพาะมักเกิดคู่กับคำบ่งชี้ที่เป็นคำย่อ เช่น ชื่อเฉพาะประเภทบุคคลปรากฏร่วมกับคำนำหน้าชื่อ เช่น น.ส. ชื่อยศตำแหน่งทางอาชีพหรือตำแหน่งทางวิชาการเช่น พ.ต.อ., ศ., พญ. เป็นต้น ชื่อเฉพาะประเภทองค์กรบางชื่อมีคำบ่งชี้ที่เป็นคำย่อ เช่น บจ. (บริษัทจำกัด), ร.ร. (โรงเรียน), ม. (มหาวิทยาลัย) เป็นต้น และชื่อเฉพาะประเภทสถานที่ เช่น ชื่อถนน ตำบล อำเภอ จังหวัด มักเกิดร่วมกับคำบ่งชี้ที่เป็นคำย่อ เช่น ถ., ต., อ., จ. เป็นต้น ดังนั้น ในแบบจำลองนี้จึงกำหนดให้คำย่อเป็นคุณสมบัติหนึ่งของระบบเพื่อใช้ช่วยระบุขอบเขตและประเภทของชื่อเฉพาะ

ในการกำหนดค่าของคุณสมบัติคำย่อนี้ กำหนดให้เป็นแบบมีสองค่า สำหรับระบบที่ใช้ข้อมูลแบบตัดคำ ค่าคุณสมบัตินี้จะกำหนดให้เป็น Y เมื่อ token ปัจจุบันและ token ก่อนหน้า 2 token มี token ใด token หนึ่งเป็นคำย่อ สำหรับระบบที่ใช้ข้อมูลแบบตัดพยางค์ จะเพิ่มจำนวน token ก่อนหน้าจาก 2 token เป็น 3 token เนื่องจากเมื่อหาค่าเฉลี่ยจำนวนพยางค์ต่อจำนวนคำในข้อมูลทั้งหมดได้ 1 คำมีค่าเท่ากับ 1.33 พยางค์ โดยในข้อมูลมีจำนวน

พยางค์ทั้งหมดเท่ากับ 487,364 พยางค์ และจำนวนคำทั้งหมดเท่ากับ 367,673 คำ ตัวอย่างการกำหนดค่าของคุณสมบัติคำย่อในข้อมูลตัดคำ เช่น

| token | คุณสมบัติคำย่อ |
|---|---|
| น.พ. | Y |
| จรัล | Y |
| <s> | Y |
| ตฤณ | N |

จากตัวอย่าง “น.พ.” เป็นคำย่ออยู่แล้วส่วน “จรัล” และ “<s>” พบคำย่ออยู่ในช่วง 2 คำก่อนหน้าจึงมีค่าเป็น Y สำหรับ “ตฤณ” ไม่พบคำย่ออยู่ในช่วง 2 คำก่อนหน้าจึงมีค่าเป็น N

4.1.1.3 **คุณสมบัติคำบริบท** (context clue) คำบริบทนำมาเป็นคุณสมบัติข้อหนึ่งเนื่องจากบริบทรอบข้างของชื่อเฉพาะมีส่วนสำคัญในการกำหนดชนิดของชื่อเฉพาะโดยเฉพาะชื่อองค์กรและชื่อสถานที่ ที่เมื่อไปปรากฏในบางบริบทจะต้องเปลี่ยนชนิดของชื่อเฉพาะนั้น เพราะชื่อองค์กรสามารถนำมาใช้หมายถึงชื่อของสถานที่ตั้งขององค์กรนั้น ๆ ได้ เช่น “เหตุเกิดบริเวณหน้า*โรงแรมเจ.บี.*”*โรงแรมเจ.บี.* เดิมเป็นชื่อองค์กร เพราะโรงแรมมีการบริหาร การจัดการ การจ้างงาน และอื่น ๆ ซึ่งเป็นลักษณะขององค์กร แต่เมื่อ*โรงแรมเจ.บี.*ปรากฏอยู่ในบริบทนี้ *โรงแรมเจ.บี.* ไม่ได้หมายถึงองค์กรแต่หมายถึงสถานที่ตั้งขององค์กรว่ามีเหตุการณ์บางอย่างเกิดขึ้นตรงบริเวณนั้นดังนั้น*โรงแรมเจ.บี.*จึงเป็นชื่อองค์กรที่ใช้อ้างถึงสถานที่ คำบริบทเช่น “หน้า” ในตัวอย่างนี้ก็น่าจะมีประโยชน์ต่อการรู้จำของระบบ สำหรับชื่อสถานที่ก็เช่นกันที่เมื่อนำไปใช้ในบางบริบทแล้วจะใช้หมายถึงองค์กร เช่น ชื่อประเทศ โดยปกติจะหมายถึงชื่อสถานที่ตามภูมิศาสตร์แต่บางครั้งมีการนำชื่อประเทศมาใช้เพื่อหมายถึงประชากรหรือหน่วยงานต่าง ๆ ภายในประเทศนั้น เช่น สำหรับคดียิงอุปทูตทาง*ประเทศซาอุดีอาระเบีย*ไม่ติดใจ เป็นต้น

นอกจากบริบทด้านหน้าแล้วบางครั้งบริบทด้านหลังชื่อเฉพาะก็สามารถนำมาใช้เป็นสิ่งบ่งบอกชนิดของชื่อเฉพาะได้เช่นกัน เช่น “*ไทย*สั่งซื้อ 6 ลำ” เป็นต้นโดยปกติ “ไทย” เป็นชื่อสถานที่ แต่กรณีนี้ “ไทย” เป็นชื่อสถานที่อ้างถึงองค์กร เพราะคำบริบทด้านหลังซึ่งคือคำว่า “สั่งซื้อ” เป็นตัวกำหนดว่า “ไทย” มีอำนาจกระทำการได้เช่นเดียวกับองค์กร

ช่วงคำบริบทของชื่อเฉพาะที่นำมาพิจารณา ได้จากคลังข้อมูลสำหรับฝึกฝนเท่านั้น โดยในข้อมูลแบบตัดคำใช้ช่วง 3 token ก่อนหน้าและต่อท้ายชื่อเฉพาะแต่สำหรับข้อมูลแบบตัดพยางค์ จะเพิ่มเป็น 4 token ตัวอย่างช่วงคำบริบทของข้อมูลแบบตัดคำ เช่น

|จาก|การ|ติดตาม|ล่าสุด|<orgName>|กระทรวง|เกษตร|ฯ|</orgName>|ได้|เข้า|ไป|ดู|ใน|จุด|ที่|

3 token ก่อนหน้าชื่อ — 3 token ต่อท้ายชื่อ

token คำบริบทที่สกัดออกมาเหล่านี้จะนำมาสร้างเป็นรายการคำบริบทจากนั้นนำแต่ละ token มาเทียบดูว่าพบอยู่ในรายการคำบริบทหรือไม่ หากพบจะกำหนดให้ token นั้นมีค่าคุณสมบัติเป็น Y ซึ่งคุณสมบัตินี้กำหนดให้เป็นแบบมีสองค่า เช่น

| token | คุณสมบัติคำบริบท |
|---|---|
| ติดตาม | Y |
| หนี้ | N |

จากตัวอย่าง “ติดตามหนี้” เป็นวลีทั่วไป เมื่อแยกเป็น token แล้วพบว่า “ติดตาม” อยู่ในรายการคำบริบท ดังจะเห็นได้จากตัวอย่างก่อนหน้าที่ “ติดตาม” อยู่ในช่วง 3 token ก่อนหน้าชื่อเฉพาะ ดังนั้น “ติดตาม” จึงมีค่าเป็น Y สำหรับคลังข้อมูลทดสอบจะใช้รายการคำบริบทเดียวกันกับที่ใช้ในคลังข้อมูลฝึกฝน เพราะในการทดสอบไม่สามารถนำข้อมูลจากคลังสำหรับทดสอบมาใช้ได้

4.1.1.4 **คุณสมบัติคำทั่วไป (general words)** โดยทั่วไปเมื่อตัดแบ่งคำที่นำมาประกอบเป็นชื่อแล้วจะเห็นว่าชื่ออาจเกิดจากการนำเอาคำหรือสายอักขระที่อาจจะเป็นคำที่เป็นที่รู้จัก หรือมีอยู่ในพจนานุกรมมารวมเข้ากับคำหรือสายอักขระที่ไม่มีอยู่ในพจนานุกรม อาจเป็นคำมีอยู่ในพจนานุกรมทั้งหมดหรืออาจจะไม่มีเลยก็ได้ (Charoenpornsawat et al., 1998) เช่น ชื่อที่ถ่ายถอดเสียงมาจากภาษาต่างประเทศ มักเป็นคำที่ไม่มีในพจนานุกรม เช่น ฮู (WHO: องค์การอนามัยโลก) ชื่อบุคคลที่บางครั้งเกิดจากการนำเอาคำบาลี สันสกฤตหรือภาษาเขมร มาสมาสหรือสนธิกัน หรือบางครั้งชื่อเป็นคำแปลกที่สร้างขึ้นมาใหม่เพื่อสร้างเอกลักษณ์เฉพาะตัวให้แก่ตนเองหรือองค์กร เช่นนี้คำที่มาสร้างชื่อจึงไม่จำกัดเฉพาะคำทั่วไปหรือคำที่มีอยู่ในพจนานุกรมเท่านั้น

รายการคำทั่วไปจึงน่าจะนำมาเป็นคุณสมบัติหนึ่งที่นำมาใช้เพื่อการรู้จำชื่อเฉพาะได้ โดยผู้วิจัยคาดว่าคำที่ไม่ใช่คำทั่วไปหรือไม่พบในพจนานุกรมมีแนวโน้มว่าอาจเป็นส่วนหนึ่งของชื่อได้ การใช้คุณสมบัตินี้กำหนดให้เป็นแบบมีสองค่า ค่าของ token จะเป็น Y ก็ต่อเมื่อไม่พบ token นั้นอยู่ในรายการคำทั่วไป เช่น

| token | คุณสมบัติคำทั่วไป |
|---|---|
| นาง | N |
| สุดารัตน์ | Y |

จากตัวอย่าง “นาง” พบอยู่ในรายการคำทั่วไปจึงมีค่าเป็น N ในขณะที่ “สุดารัตน์” ไม่พบอยู่ในรายการคำทั่วไปจึงมีค่าเป็น Y เป็นต้น

4.1.1.5 **คุณสมบัติค่าทางสถิติ** เนื่องจากในบางข่าวจะมีชื่อเฉพาะชื่อเดิมปรากฏมากกว่าหนึ่งครั้ง ดังนั้นคำที่อยู่ติดกันและปรากฏร่วมกันหลายครั้งจึงมีแนวโน้มว่าอาจเป็นชื่อเฉพาะได้ผู้วิจัยจึงได้กำหนดคุณสมบัติค่าทางสถิตินี้ขึ้นมา โดยกำหนดให้เป็นแบบมีสองค่าโดยหากคำหรือพยางค์ 3 หน่วยที่อยู่ติดกัน คือ หน่วยก่อนหน้า token | token | หน่วยตามหลัง token ปรากฏอยู่ในข่าวนั้น ๆ มากกว่า 3 ครั้ง token นั้นจะมีค่าเป็น Y เช่น

| token | คุณสมบัติค่าทางสถิติ |
|---|---|
| จาก | N |
| จังหวัด | N |
| กาญจนบุรี | Y |
| <s> | N |

จากตัวอย่างจะเห็นว่า “กาญจนบุรี” มีค่าเป็น Y นั่นหมายความว่า “จังหวัด|กาญจนบุรี|<s>” พบมากกว่า 3 ครั้งในข่าวนั้น

4.1.1.6 **Unigram และ Bigram** เป็น template ที่ต้องใช้ในการฝึกฝนแบบจำลอง CRF++-0.53 เป็นค่าความน่าจะเป็นของชุดคำหรือพยางค์ที่ปรากฏอยู่ติดกัน

นอกจาก unigram และ bigram แล้ว ภายใน template ยังต้องกำหนดว่ามีคุณสมบัติใดบ้างที่จะใช้ในการฝึกฝนแบบจำลองตัวอย่าง template ที่ใช้ในงานวิจัยนี้ได้แสดงไว้ดังภาพที่ 4.2

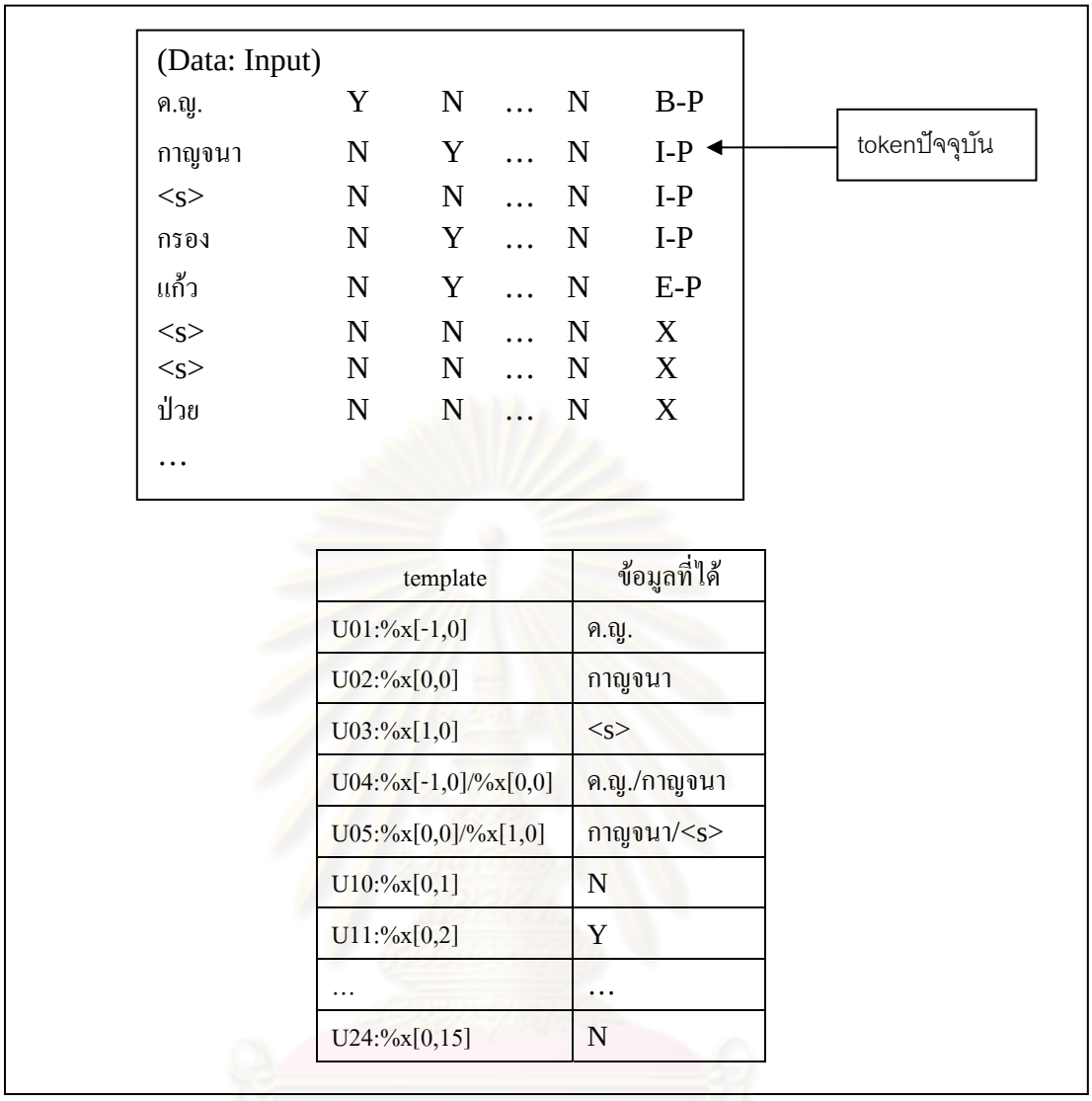
```
# Unigram
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[-1,0]/%x[0,0]
U05:%x[0,0]/%x[1,0]

U10:%x[0,1]
U11:%x[0,2]
U12:%x[0,3]
…
…
U24:%x[0,15]

# Bigram
B
```

ภาพที่ 4.2 ตัวอย่าง template

จากภาพ %x[แถว,คอลัมน์] ใช้กำหนดตำแหน่งต่าง ๆ ในข้อมูล U01-U03 คือคุณสมบัติ unigram U04-U05 คือคุณสมบัติ bigram และ U10-U24 คือ คุณสมบัติต่าง ๆ ที่ใช้ โดย U10-U20 คือ คุณสมบัติรายการชื่อเฉพาะ U21-U24 คือ คุณสมบัติคำบริบท คุณสมบัติคำย่อ คุณสมบัติคำทั่วไป และคุณสมบัติค่าทางสถิติตามลำดับ เหตุที่คุณสมบัติรายการชื่อเฉพาะมีหลายคอลัมน์เพราะได้มีการแบ่งส่วนของชื่อเฉพาะแต่ละประเภทเป็นส่วนต้น ส่วนกลาง และส่วนท้ายของชื่อและได้เพิ่มรายการคำนำหน้าชื่อบุคคลและรายการชื่อย่อองค์กรเข้าไปด้วย จึงมีทั้งหมด 11 คอลัมน์ด้วยกัน สำหรับตำแหน่งต่าง ๆ ของข้อมูลในคลังข้อมูลฝึกฝนจะต้องเรียงให้ตรงกับตำแหน่งใน template ด้วย ตัวอย่างของคลังข้อมูลฝึกฝนและการแทนค่าของข้อมูลใน template ได้แสดงไว้ในภาพที่ 4.3




| template | ข้อมูลที่ได้ |
|---|---|
| U01:%x[-1,0] | ด.ญ. |
| U02:%x[0,0] | กาญจนา |
| U03:%x[1,0] | <s> |
| U04:%x[-1,0]/%x[0,0] | ด.ญ./กาญจนา |
| U05:%x[0,0]/%x[1,0] | กาญจนา/<s> |
| U10:%x[0,1] | N |
| U11:%x[0,2] | Y |
| … | … |
| U24:%x[0,15] | N |

ภาพที่ 4.3 ตัวอย่างคลังข้อมูลฝึกฝนแบบตัดคำ และการแทนค่าของข้อมูลใน template

ลำดับของข้อมูลจะปรับไปตามตำแหน่งของ token ปัจจุบัน โดยที่ token ปัจจุบันจะอยู่ที่ตำแหน่ง [0,0] คือแถวที่ 0 และคอลัมน์ที่ 0 เสมอ สำหรับคลังข้อมูลฝึกฝนและทดสอบจะคล้ายกัน คือ คอลัมน์แรกเป็น token คอลัมน์ถัดไปเป็นคุณสมบัติต่าง ๆ แต่ละคอลัมน์จะแทนแต่ละคุณสมบัติโดยที่คอลัมน์ของแต่ละ token ต้องเท่ากัน และคอลัมน์สุดท้ายในคลังข้อมูลฝึกฝนจะเป็นคำตอบที่ให้แก่แบบจำลองสำหรับเรียนรู้ซึ่งในคลังข้อมูลทดสอบจะไม่มี เพราะคำตอบของคลังข้อมูลทดสอบจะได้จากแบบจำลอง โดยรูปแบบของคำตอบที่ได้จะเหมือนกับรูปแบบของคำตอบที่แบบจำลองได้เรียนรู้จากคลังข้อมูลฝึกฝนจากภาพตัวอย่างคลังข้อมูล (Data:Input) คอลัมน์ที่ 1 คือคำหรือ token คอลัมน์ที่ 2 และ 3 คือคุณสมบัติรายการชื่อเฉพาะ โดยคอลัมน์ที่ 2 เป็นคำนำหน้าชื่อ และคอลัมน์ที่ 3 เป็นส่วนต้นของชื่อบุคคล คอลัมน์รองสุดท้ายคือ คุณสมบัติค่าทางสถิติ และคอลัมน์สุดท้ายคือคำตอบที่ให้แก่แบบจำลอง สำหรับ

รูปแบบของคำตอบที่ใช้จะกล่าวถึงรายละเอียดในหัวข้อถัดไป และตัวอย่างการเพิ่มคุณสมบัติทั้งหมดให้แก่ token สามารถดูได้จากภาคผนวก ก และ ข

### 4.1.2 รูปแบบของคำตอบที่ใช้ในการฝึกฝนแบบจำลอง

รูปแบบของคำตอบที่ใช้ในการฝึกฝนแบบจำลองมีทั้งหมด 5 แบบด้วยกัน เนื่องจากผู้วิจัยพบว่าในงานวิจัยต่าง ๆ มีการให้ข้อมูลเพิ่มเติมนอกเหนือจากชนิดของชื่อเฉพาะในการฝึกฝนแบบจำลองเช่น งานวิจัยการรู้จำชื่อเฉพาะในภาษาจีน งานของ Feng, Sun, and Lv (2006) ให้ข้อมูลขอบเขตของชื่อเฉพาะในรูปแบบ BOI tags โดยที่ B คือจุดเริ่มต้นของชื่อเฉพาะ I คือส่วนที่เป็นชื่อเฉพาะ และ O คือส่วนอื่น ๆ ที่ไม่ใช่ชื่อเฉพาะ ในขณะเดียวกันงานของ Zhou et al. (2006) กำหนดให้มีเครื่องหมายกำหนดขอบเขตของคำ 4 แบบคือ BIFS โดยที่ B เป็นจุดเริ่มต้นของคำ I เป็นส่วนที่อยู่ภายในคำ F คือจุดสิ้นสุดของคำ และ S สำหรับคำที่มีอักษรเพียงตัวเดียว นอกจากนี้ในงานวิจัยการรู้จำชื่อเฉพาะในภาษาจีนในปี 2008 ปรากฏว่าในบางงานได้มีการเพิ่มจำนวน tag เช่น งานของ Yu et al. (2008) ใช้ tag BIOES แทน BIO โดยเพิ่ม tag E และ S เข้ามาโดยที่ E ใช้สำหรับจุดสิ้นสุดของชื่อเฉพาะ และ S สำหรับชื่อเฉพาะที่เป็นอักษรตัวเดียวซึ่งสาเหตุหลักของการเพิ่มจำนวน tag ก็เพื่อเพิ่มข้อมูลให้แก่แบบจำลอง ซึ่งผลการวิจัยพบว่าการใช้ tag ที่มากขึ้นให้ผลได้ดีกว่าการใช้จำนวน tag น้อย ดังนั้นผู้วิจัยจึงต้องการเปรียบเทียบประสิทธิภาพของแบบจำลองว่าหากให้ข้อมูลขอบเขตของชื่อเฉพาะมากขึ้นแล้วจะมีผลทำให้ประสิทธิภาพของแบบจำลองดีขึ้นดังเช่นในงานวิจัยของภาษาจีนหรือไม่ โดยใช้คุณสมบัติเหมือนกันทุกอย่างในการฝึกฝนแบบจำลอง

รูปแบบของคำตอบที่ใช้ในการฝึกฝนทั้งหมดมีดังนี้

1) P, O, L, X เป็นรูปแบบของคำตอบแบบธรรมดา ให้เฉพาะข้อมูลชนิดของชื่อเฉพาะ ดังนี้

P : ชื่อเฉพาะประเภทบุคคล

O : ชื่อเฉพาะประเภทองค์กรรวมถึงชื่อเฉพาะสถานที่อ้างถึงองค์กร

L : ชื่อเฉพาะประเภทสถานที่รวมถึงชื่อเฉพาะองค์กรอ้างถึงสถานที่

X :อื่น ๆ

2) B, I, X – PER, ORG, LOC ให้ข้อมูลขอบเขตของชื่อเฉพาะและชนิดของชื่อเฉพาะ ดังนี้

ตารางที่ 4.2 รายละเอียดรูปแบบคำตอบ B, I, X – PER, ORG, LOC

| | จุดเริ่มต้น | ภายในชื่อ |
|---|---|---|
| ชื่อบุคคล | B – PER | I – PER |
| ชื่อองค์กร<br>ชื่อสถานที่อ้างถึงองค์กร | B – ORG | I – ORG |
| ชื่อสถานที่<br>ชื่อองค์กรอ้างถึงสถานที่ | B – LOC | I – LOC |

หมายเหตุ: X–อื่น ๆ

3) B, I, X – P, O, L, LO, OLให้ข้อมูลขอบเขตของชื่อเฉพาะและชนิดของชื่อเฉพาะ ดังนี้

ตารางที่ 4.3 รายละเอียดรูปแบบคำตอบ B, I, X – P, O, L, LO, OL

| | จุดเริ่มต้น | ภายในชื่อ |
|---|---|---|
| ชื่อบุคคล | B – P | I – P |
| ชื่อองค์กร | B – O | I – O |
| ชื่อสถานที่ | B – L | I – L |
| ชื่อสถานที่อ้างถึงองค์กร | B – LO | I – LO |
| ชื่อองค์กรอ้างถึงสถานที่ | B – OL | I – OL |

หมายเหตุ: X–อื่น ๆ

4) B, I, E, X – PER, ORG, LOC ให้ข้อมูลขอบเขตของชื่อเฉพาะและชนิดของชื่อเฉพาะ ดังนี้

ตารางที่ 4.4 รายละเอียดรูปแบบคำตอบ B, I, E, X – PER, ORG, LOC

| | จุดเริ่มต้น | ภายในชื่อ | จุดสิ้นสุด |
|---|---|---|---|
| ชื่อบุคคล | B – PER | I – PER | E – PER |
| ชื่อองค์กร<br>ชื่อสถานที่อ้างถึงองค์กร | B – ORG | I – ORG | E – ORG |
| ชื่อสถานที่<br>ชื่อองค์กรอ้างถึงสถานที่ | B – LOC | I – LOC | E – LOC |

หมายเหตุ: X– อื่น ๆ

5) B, I, E, X – P, O, L, LO, OLให้ข้อมูลขอบเขตของชื่อเฉพาะและชนิดของชื่อเฉพาะ ดังนี้

ตารางที่ 4.5 รายละเอียดรูปแบบคำตอบ B, I, E, X – P, O, L, LO, OL

| | จุดเริ่มต้น | ภายในชื่อ | จุดสิ้นสุด |
|---|---|---|---|
| ชื่อบุคคล | B – P | I – P | E – P |
| ชื่อองค์กร | B – O | I – O | E – O |
| ชื่อสถานที่ | B – L | I – L | E – L |
| ชื่อสถานที่อ้างถึงองค์กร | B – LO | I – LO | E –LO |
| ชื่อองค์กรอ้างถึงสถานที่ | B – OL | I – OL | E – OL |

หมายเหตุ: X– อื่น ๆ

## 4.2 การประเมินประสิทธิภาพของแบบจำลอง

การวัดประสิทธิภาพของแบบจำลองจะวัดจากค่าความแม่นยำ (Precision) ค่าความครบถ้วน (Recall) และค่า F-measure

ค่าความแม่นยำ คือ ค่าที่แสดงให้เห็นว่าระบบสามารถรู้จำชื่อเฉพาะได้แม่นยำมากน้อยขนาดไหนเมื่อเทียบจากจำนวนชื่อเฉพาะทั้งหมดที่แบบจำลองสกัดออกมา สามารถคำนวณได้จากสูตรดังนี้

$$P = \frac{\text{จำนวนชื่อเฉพาะที่รู้จำได้ถูกต้อง} * 100}{\text{จำนวนชื่อเฉพาะทั้งหมดที่สกัดออกมา}}$$

ค่าความครบถ้วน คือ ค่าที่แสดงให้เห็นว่าแบบจำลองสามารถรู้จำชื่อเฉพาะได้ครบถ้วนเพียงใด เมื่อเทียบกับจำนวนชื่อเฉพาะทั้งหมดในเอกสารนั้น ๆ สามารถคำนวณได้จากสูตรดังนี้

$$R = \frac{\text{จำนวนชื่อเฉพาะที่รู้จำได้ถูกต้อง} * 100}{\text{จำนวนชื่อเฉพาะทั้งหมดในเอกสาร}}$$

ค่า F-measure คือ ค่าความถูกต้องโดยรวม เป็นค่าเฉลี่ยของค่าความแม่นยำและค่าความครบถ้วน สามารถคำนวณได้จากสูตรดังนี้

$$F = \frac{2 * P * R}{P + R}$$

นอกจากการวัดค่าความถูกต้องจากจำนวนชื่อเฉพาะที่ระบบรู้จำได้ถูกต้องจากที่กล่าวไปข้างต้นแล้ว ผู้วิจัยยังได้วัดค่าความถูกต้องโดยดูจาก token ที่ตรงกับคำตอบด้วย เหตุที่วัดจากจำนวน token ด้วยนั้น เนื่องจากมีบางชื่อที่ระบบอาจระบุขอบเขตผิด จึงรู้จำได้เพียงบางส่วนของชื่อเท่านั้น เช่น ชื่อบุคคลที่ระบบรู้จำส่วนของชื่อได้แต่ไม่รู้จำส่วนของนามสกุล เป็นต้น ดังนั้นผู้วิจัยจึงจะหาค่าความแม่นยำ ค่าความครบถ้วน และค่า F-measure ทั้งจากจำนวนชื่อและจำนวน token ของชื่อที่ระบบรู้จำได้ถูกต้อง โดยการหาค่าความแม่นยำและค่าความครบถ้วนของจำนวน token จะคำนวณได้จากสูตรต่อไปนี้

$$P_{token} = \frac{\text{จำนวน } token \text{ ของชื่อเฉพาะที่รู้จำได้ถูกต้อง} * 100}{\text{จำนวน } token \text{ ของชื่อเฉพาะทั้งหมดที่สกัดออกมา}}$$

$$R_{token} = \frac{\text{จำนวน } token \text{ ของชื่อเฉพาะที่รู้จำได้ถูกต้อง} * 100}{\text{จำนวน } token \text{ ของชื่อเฉพาะทั้งหมดในเอกสาร}}$$

ในการทดสอบ เพื่อไม่ให้ผลขึ้นอยู่กับข้อมูลเพียงส่วนใดส่วนหนึ่ง ผู้วิจัยจึงกำหนดให้ข้อมูลทุกส่วนได้ใช้ในการทดสอบโดยการแบ่งข้อมูลออกเป็น 10 ส่วน ข้อมูล 9 ส่วนใช้ในการฝึกฝนและอีก 1 ส่วนใช้ในการทดสอบ จากนั้นจะมีการสลับข้อมูลระหว่างข้อมูลที่ใช้ฝึกฝนและทดสอบรวมทั้งหมด 10 ครั้ง แล้วจึงหาค่าเฉลี่ยของค่าความแม่นยำ ค่าความครบถ้วน และค่า F-measure ของผลการทดสอบทั้งหมด

### 4.3 ผลการทดสอบ

หลังจากฝึกฝนและทดสอบข้อมูลทั้งแบบตัดคำและตัดพยางค์ทั้งหมด 10 ครั้งแล้ว (ดูภาคผนวก ง ซึ่งแสดงผลการทดสอบทั้ง 10 ครั้ง) ได้ค่าเฉลี่ยของค่าความแม่นยำ ค่าความครบถ้วน และค่า F-measure ของแบบจำลองแต่ละแบบที่มีข้อมูลขอบเขตและชนิดของชื่อเฉพาะแตกต่างกัน ดังแสดงในตารางต่อไปนี้

ตารางที่ 4.6 ประสิทธิภาพของแบบจำลองที่ได้รับคำตอบแบบที่ 1 (P, O, L, X)

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG* | SSG** | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 85.69 | 86.95 | 83.93 | 83.04 | 84.77 | 84.90 |
| ชื่อองค์กร | 77.28 | 76.50 | 70.55 | 71.31 | 73.75 | 73.80 |
| ชื่อสถานที่ | 76.82 | 75.55 | 70.57 | 69.06 | 73.52 | 72.10 |
| ทั้งหมด | 80.39 | 80.25 | 75.26 | 74.87 | 77.73 | 77.45 |

ตารางที่ 4.7 ประสิทธิภาพของแบบจำลองที่ได้รับคำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC)

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 90.59 | 90.85 | 86.37 | 84.94 | 88.41 | 87.74 |
| ชื่อองค์กร | 82.68 | 82.06 | 74.61 | 74.96 | 78.42 | 78.33 |
| ชื่อสถานที่ | 80.72 | 80.60 | 70.71 | 69.70 | 75.31 | 74.66 |
| ทั้งหมด | 85.24 | 85.09 | 77.67 | 77.14 | 81.26 | 80.89 |

ตารางที่ 4.8 ประสิทธิภาพของแบบจำลองที่ได้รับคำตอบแบบที่ 3 (B, I, X – P, O, L, LO, OL)

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 90.14 | 90.76 | 86.17 | 85.07 | 88.08 | 87.77 |
| ชื่อองค์กร | 82.57 | 83.00 | 75.15 | 75.73 | 78.65 | 79.15 |
| ชื่อสถานที่ | 80.44 | 81.07 | 73.26 | 72.94 | 76.55 | 76.71 |
| ชื่อสถานที่อ้างถึงองค์กร | 81.43 | 78.64 | 67.32 | 67.70 | 73.43 | 72.64 |
| ชื่อองค์กรอ้างถึงสถานที่ | 75.39 | 65.99 | 43.71 | 41.89 | 54.89 | 50.67 |
| ทั้งหมด | 84.82 | 84.94 | 77.09 | 76.84 | 80.75 | 80.66 |

* ระบบที่ใช้ข้อมูลตัดคำ

** ระบบที่ใช้ข้อมูลตัดพยางค์

ตารางที่ 4.9 ประสิทธิภาพของแบบจำลองที่ได้รับคำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC)

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 92.05 | 92.55 | 86.50 | 85.83 | 89.16 | 89.01 |
| ชื่อองค์กร | 82.19 | 81.84 | 74.23 | 74.58 | 77.99 | 78.02 |
| ชื่อสถานที่ | 79.92 | 79.69 | 70.77 | 70.99 | 74.98 | 74.98 |
| ทั้งหมด | 85.37 | 85.37 | 77.64 | 77.64 | 81.30 | 81.30 |

ตารางที่ 4.10 ประสิทธิภาพของแบบจำลองที่ได้รับคำตอบแบบที่ 5 (B, I, E, X – P, O, L, LO, OL)

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 91.52 | 92.26 | 86.47 | 85.64 | 88.89 | 88.77 |
| ชื่อองค์กร | 82.67 | 82.80 | 75.38 | 76.07 | 78.83 | 79.26 |
| ชื่อสถานที่ | 80.04 | 81.12 | 73.35 | 73.49 | 76.45 | 77.00 |
| ชื่อสถานที่อ้างถึงองค์กร | 81.04 | 78.12 | 68.00 | 67.12 | 73.72 | 72.10 |
| ชื่อองค์กรอ้างถึงสถานที่ | 77.35 | 69.29 | 43.57 | 43.18 | 55.35 | 52.61 |
| ทั้งหมด | 85.29 | 85.46 | 77.37 | 77.25 | 81.12 | 81.12 |

จากตารางที่ 4.6-4.10 แม้ว่าผลลัพธ์ที่ได้จากการให้คำตอบแบบที่ 2-5 จะให้ผลเกือบไม่ต่างกัน โดยต่างกันไม่ถึง 1 เปอร์เซ็นต์ แต่ก็แสดงให้เห็นว่าการให้ข้อมูลขอบเขตของชื่อเฉพาะมากขึ้น มีแนวโน้มว่าจะส่งผลให้ระบบสามารถรู้จำชื่อเฉพาะได้ดีขึ้นตามไปด้วย โดยเมื่อพิจารณาค่า F-measure ของชื่อเฉพาะทั้งหมดในข้อมูลแต่ละแบบ พบว่าการให้คำตอบแบบที่ 4 ให้ผลดีที่สุด รองลงมาคือคำตอบแบบที่ 2 ในข้อมูลแบบตัดคำ และคำตอบแบบที่ 5 ในข้อมูลแบบตัดพยางค์ซึ่งเมื่อนำไปเปรียบเทียบกับผลของคำตอบในแบบที่ 1 ที่ให้เฉพาะข้อมูลชนิดของชื่อเฉพาะเท่านั้น จะเห็นว่ามีความต่างกันมาก

แม้ว่าระบบที่ได้รับคำตอบแบบที่ 5 ซึ่งมีข้อมูลขอบเขตและชนิดของชื่อเฉพาะมากที่สุด จะไม่ใช่ระบบที่ให้ค่า F-measure สูงที่สุด แต่เมื่อเทียบกับระบบที่ได้รับคำตอบแบบที่ 4 แล้วจะเห็นว่าต่างกันไม่ถึง 0.5 เปอร์เซ็นต์ จึงมีแนวโน้มว่าหากขนาดของคลังข้อมูลที่ใช้สำหรับฝึกฝนใหญ่กว่านี้ ระบบที่ได้รับคำตอบแบบที่ 5 อาจมีประสิทธิภาพกว่าระบบอื่น ๆ ได้ เพราะการที่ระบบได้รับข้อมูลมากขึ้นจำเป็นต้องใช้คลังข้อมูลสำหรับฝึกฝนมากขึ้นตามไปด้วย เพราะหากคลังข้อมูลที่ใช้

ในการฝึกฝนมีไม่มากพอจะทำให้แบบจำลองไม่สามารถรู้จำชื่อเฉพาะในบางบริบทได้ดังจะเห็นได้จากประสิทธิภาพการรู้จำชื่อองค์กรอ้างถึงสถานที่ในคำตอบแบบที่ 3 และ 5 ที่มีค่าความครบถ้วนต่ำมากเมื่อเทียบกับของชื่อเฉพาะประเภทอื่น นั่นเป็นเพราะตัวอย่างของชื่อที่มีในคลังข้อมูลมีเพียง 417 ชื่อเท่านั้น ดังนั้นขนาดของคลังข้อมูลที่ใช้ในการฝึกฝนจึงมีผลต่อประสิทธิภาพของแบบจำลองค่อนข้างมาก

เมื่อประเมินค่าความถูกต้องโดยดูจาก token ได้ผลดังแสดงในตารางที่ 4.11-4.15 ดังนี้

ตารางที่ 4.11 ประสิทธิภาพของแบบจำลองที่ได้รับคำตอบแบบที่ 1 (P, O, L, X) เมื่อวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 93.41 | 94.99 | 92.95 | 91.85 | 93.16 | 93.36 |
| ชื่อองค์กร | 81.39 | 82.26 | 73.09 | 77.10 | 76.97 | 79.57 |
| ชื่อสถานที่ | 85.23 | 83.83 | 76.96 | 74.84 | 80.78 | 78.99 |
| ทั้งหมด | 88.16 | 88.20 | 83.10 | 82.71 | 85.54 | 85.35 |

ตารางที่ 4.12 ประสิทธิภาพของแบบจำลองที่ได้รับคำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) เมื่อวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 94.01 | 95.21 | 93.45 | 92.57 | 93.69 | 93.84 |
| ชื่อองค์กร | 81.46 | 82.82 | 75.39 | 78.54 | 78.28 | 80.61 |
| ชื่อสถานที่ | 85.85 | 86.30 | 75.21 | 74.06 | 80.08 | 79.61 |
| ทั้งหมด | 88.61 | 89.05 | 83.60 | 83.36 | 86.02 | 86.10 |

ตารางที่ 4.13 ประสิทธิภาพของแบบจำลองที่ได้รับคำตอบแบบที่ 3 (B, I, X – P, O, L, LO, OL) เมื่อวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 93.84 | 94.94 | 93.27 | 92.64 | 93.52 | 93.74 |
| ชื่อองค์กร | 80.22 | 82.76 | 75.56 | 78.21 | 77.75 | 80.39 |
| ชื่อสถานที่ | 85.87 | 86.92 | 78.34 | 77.86 | 81.75 | 82.04 |
| ชื่อสถานที่อ้างถึงองค์กร | 81.57 | 80.00 | 64.04 | 68.41 | 71.39 | 73.48 |
| ชื่อองค์กรอ้างถึงสถานที่ | 73.79 | 65.64 | 44.36 | 44.40 | 54.49 | 51.86 |
| ทั้งหมด | 88.03 | 88.38 | 82.66 | 82.49 | 85.25 | 85.32 |

ตารางที่ 4.14 ประสิทธิภาพของแบบจำลองที่ได้รับคำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) เมื่อวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 94.56 | 95.78 | 93.10 | 92.82 | 93.78 | 94.23 |
| ชื่อองค์กร | 82.20 | 84.44 | 75.24 | 78.71 | 78.54 | 81.46 |
| ชื่อสถานที่ | 86.20 | 86.22 | 76.27 | 75.48 | 80.80 | 80.36 |
| ทั้งหมด | 89.14 | 89.91 | 83.61 | 83.82 | 86.27 | 86.75 |

ตารางที่ 4.15 ประสิทธิภาพของแบบจำลองที่ได้รับคำตอบแบบที่ 5 (B, I, E, X – P, O, L, LO, OL) เมื่อวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 94.17 | 95.53 | 93.23 | 92.79 | 93.65 | 94.10 |
| ชื่อองค์กร | 81.55 | 83.44 | 76.36 | 79.22 | 78.81 | 81.24 |
| ชื่อสถานที่ | 86.63 | 88.00 | 79.04 | 78.47 | 82.53 | 82.81 |
| ชื่อสถานที่อ้างถึงองค์กร | 82.87 | 80.58 | 64.76 | 68.78 | 72.43 | 74.00 |
| ชื่อองค์กรอ้างถึงสถานที่ | 76.87 | 69.30 | 44.50 | 44.64 | 55.99 | 53.41 |
| ทั้งหมด | 88.85 | 89.14 | 83.01 | 82.90 | 85.82 | 85.89 |

เมื่อพิจารณาจากการประเมินประสิทธิภาพโดยใช้จำนวน token พบว่าผลที่ได้เป็นไปในทิศทางเดียวกับการประเมินโดยใช้จำนวนชื่อ กล่าวคือ คำตอบแบบที่ 4 ได้ค่า F-measure สูงที่สุดทั้งในระบบที่ใช้ข้อมูลตัดคำและตัดพยางค์ รองลงมาคือคำตอบแบบที่ 2 และลำดับที่สามคือคำตอบแบบที่ 5

เมื่อดูจากค่า F-measure ของคำตอบทุกแบบ พบว่าเพิ่มขึ้นจากการวัดโดยใช้จำนวนชื่อมากกว่า 4 เปอร์เซ็นต์ โดยเฉพาะคำตอบแบบที่ 1 ที่เพิ่มขึ้นถึงประมาณ 8 เปอร์เซ็นต์ ทำให้ค่า F-measure ของคำตอบแบบที่ 1 ใกล้เคียงกับคำตอบแบบอื่น ๆ ซึ่งจะต่างจากการวัดโดยใช้จำนวนชื่อที่ค่า F-measure ของคำตอบแบบที่ 1 จะต่ำกว่าคำตอบแบบอื่น ๆ มาก นั่นหมายความว่าระบบที่ใช้คำตอบแบบที่ 1 มีเปอร์เซ็นต์ในการระบุขอบเขตของชื่อเฉพาะผิดมากที่สุด โดยเมื่อดูจากชื่อเฉพาะที่ระบบสกัดออกมาได้พบว่าหากชื่อเฉพาะประเภทเดียวกันอยู่ใกล้กัน ระบบมีแนวโน้มจะสกัดให้ชื่อเฉพาะเหล่านั้นเป็นชื่อเพียงชื่อเดียว เช่น “กรมวิทยาศาสตร์การแพทย์<s>สธ.” “จอย<s>รินลณี” “ตะกั่วทุ่ง<s>ตะกั่วป่า<s>กะปง” เป็นต้น จากตัวอย่างที่กล่าวมา ชื่อเฉพาะแต่ละชื่อจะมีช่องว่างคั่น แต่ระบบที่ใช้คำตอบแบบที่ 1 จะสกัดชื่อเฉพาะเหล่านั้นออกมาเป็นชื่อเฉพาะเพียงชื่อเดียวโดยให้ช่องว่าง (<s>) เป็นส่วนหนึ่งของชื่อ ดังนั้นจึงอาจกล่าวได้ว่าการให้ข้อมูลขอบเขตของชื่อเฉพาะในคำตอบด้วยช่วยให้ประสิทธิภาพในการรู้จำชื่อเฉพาะของระบบดีขึ้น

อย่างไรก็ตามสำหรับการพัฒนาระบบในขั้นต่อไปผู้วิจัยจะใช้เพียงข้อมูลที่ให้คำตอบแบบที่ 4 เท่านั้น เนื่องจากมีค่า F-measure สูงที่สุดและเพื่อประหยัดเวลาในขั้นตอนการฝึกฝนซึ่งใช้เวลาค่อนข้างมาก

คำถามที่น่าสนใจต่อมาคือ คุณสมบัติต่างๆ ที่ใช้นั้นคุณสมบัติใดมีผลต่อประสิทธิภาพของแบบจำลองมากที่สุดในการตอบคำถามนี้ ผู้วิจัยได้เริ่มจากคุณสมบัติ unigram และ bigram ก่อน เนื่องจากคุณสมบัตินี้เป็นคุณสมบัติพื้นฐานหรือเป็น template ของระบบ ซึ่งจำเป็นต้องใช้ในการประมวลผลระบบทุกครั้ง ดังนั้น ผลที่ได้จากการใช้เพียงคุณสมบัติ unigram และ bigram โดยไม่มีคุณสมบัติอื่นร่วมด้วยจะเป็นข้อมูลขั้นต่ำสำหรับการเปรียบเทียบกับผลการทำงานเมื่อมีการเพิ่มคุณสมบัติต่าง ๆ เข้าไปทีละคุณสมบัติในระบบ ในการทดสอบโดยใช้เพียง unigram และ bigram จากการประมวลผลข้อมูลทั้งหมด 10 ครั้ง (ดูภาคผนวก จ) ทั้งจากจำนวนชื่อเฉพาะและจำนวน token ได้ค่าเฉลี่ยของผลการทดสอบดังแสดงไว้ในตารางที่ 4.16 และ 4.17 ดังนี้

ตารางที่ 4.16 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติ unigram และ bigram เท่านั้น

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 91.22 | 93.01 | 80.13 | 84.01 | 85.25 | 88.22 |
| ชื่อองค์กร | 87.51 | 84.03 | 64.53 | 68.61 | 74.21 | 75.50 |
| ชื่อสถานที่ | 82.79 | 80.15 | 64.45 | 67.88 | 72.41 | 73.37 |
| ทั้งหมด | 87.84 | 86.57 | 70.06 | 73.91 | 77.93 | 79.72 |

ตารางที่ 4.17 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติ unigram และ bigram เท่านั้น เมื่อวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 95.49 | 96.15 | 88.91 | 91.69 | 92.01 | 93.83 |
| ชื่อองค์กร | 89.00 | 87.54 | 66.52 | 74.94 | 76.07 | 80.74 |
| ชื่อสถานที่ | 89.09 | 86.93 | 70.54 | 72.27 | 78.67 | 78.84 |
| ทั้งหมด | 92.43 | 91.44 | 77.73 | 81.29 | 84.41 | 86.05 |

จากผลการทดสอบ เมื่อดูจากค่า F-measure แบบนับจำนวนชื่อในตารางที่ 4.16 จะเห็นว่าเฉพาะคุณสมบัติ unigram และ bigram ช่วยให้ระบบมีประสิทธิภาพเกือบถึง 80% ทั้งระบบที่ใช้ข้อมูลแบบตัดคำและตัดพยางค์ และมากกว่า 80% เมื่อวัดประสิทธิภาพจากจำนวน token แต่เมื่อเปรียบเทียบทั้งสองระบบแล้วจะเห็นว่าคุณสมบัติ unigram และ bigram ช่วยให้ประสิทธิภาพของระบบที่ใช้ข้อมูลแบบตัดพยางค์ดีกว่า เพราะเมื่อพิจารณาจากค่าความครบถ้วนแล้วจะเห็นว่าค่าครบถ้วนของชื่อเฉพาะแต่ละชนิดของข้อมูลแบบตัดพยางค์สูงกว่าของข้อมูลแบบตัดคำทุกค่า

สาเหตุที่คุณสมบัติ unigram และ bigram สนับสนุนข้อมูลแบบตัดพยางค์มากกว่าตัดคำนั้น เนื่องจาก unigram และ bigram เป็นการหาค่าความน่าจะเป็นของคำหรือพยางค์ที่อยู่ติดกัน กล่าวคือ คำหรือพยางค์ที่อยู่ติดกันและปรากฏหลายครั้งในข้อมูลมีความน่าจะเป็นชื่อเฉพาะได้ ปัญหาจึงเกิดกับข้อมูลแบบตัดคำ เมื่อชื่อเฉพาะที่เป็นคำ ๆ เดียวเกิดในบริบทที่ไม่มีคำบ่งชี้ เช่น ชื่อจังหวัดปรากฏโดยไม่มี “จ.” หรือคำว่า “จังหวัด” นำหน้า คุณสมบัติ unigram และ bigram จึงไม่สามารถหาค่าความน่าจะเป็นของชื่อเฉพาะนั้นกับบริบทข้างเคียงได้ ในขณะที่ถ้าเป็นข้อมูลแบบตัดพยางค์ บางครั้งคำหนึ่งคำสามารถแยกย่อยได้เป็นหลายพยางค์ เช่น อยุธยา = อยุธ-ยา

และกาญจนบุรี = กาญ-จน-บุ-รี เป็นต้น ดังนั้นระบบจึงสามารถสกัดชื่อเฉพาะที่เป็นคำ ๆ เดียวเหล่านี้ได้แม้ว่าจะเกิดในบริบทที่ไม่มีคำบ่งชี้ก็ตาม เพราะเป็นชื่อที่ประกอบด้วยหลายพยางค์และเกิดร่วมกันหลายครั้งในข้อมูล จึงมีค่าความน่าจะเป็นที่พยางค์เหล่านี้จะเป็นชื่อเฉพาะสูง

แต่ทั้งนี้ข้อมูลแบบตัดพยางค์ก็มีปัญหาในการรู้จำชื่อเฉพาะเช่นกัน เนื่องจากการตัดพยางค์มีการแยกส่วนของคำ บางครั้งส่วนของคำตรงกับชื่อเฉพาะพอดี เช่น ขนมจีน = ขนม-จีน ระบบจึงไปกำหนดให้ “จีน” ในคำว่า “ขนมจีน” เป็นชื่อเฉพาะสถานที่ ในขณะที่ถ้าเป็นระบบที่ใช้ข้อมูลแบบตัดคำจะไม่เกิดปัญหานี้

แต่ปัญหาที่ทั้งสองระบบพบ คือ ไม่สามารถรู้จำชื่อเฉพาะที่เป็นชื่อย่อองค์กรได้หากชื่อย่อนั้นไม่มีคำปรากฏร่วมที่เกิดขึ้นร่วมกันบ่อยครั้ง เนื่องจากชื่อย่อองค์กรมักเป็น token เดียว หากไม่มีคำปรากฏร่วมที่ช่วยบ่งชี้ก็ยากที่ระบบจะสามารถสกัดออกมาได้ เช่น ระบบไม่สามารถรู้จำชื่อ “อคส.” ว่าเป็นชื่อองค์กรได้เมื่อ “อคส.” เกิดในบริบทที่ไม่มีคำบ่ง เช่น “ข้าวของ-อคส.-<s>” ในระบบแบบตัดคำหรือ “ดู-แล-อคส.” ในระบบแบบตัดพยางค์ แต่ระบบจะสามารถรู้จำ “อคส.” ได้ว่าเป็นชื่อองค์กร หาก “อคส.” เกิดตามหลัง “ผอ.” “เจ้าหน้าที่” หรือ “หัวหน้า”นั่นเพราะชื่อองค์กรมักปรากฏตามหลังชื่อตำแหน่งซึ่งระบบได้เรียนรู้จากคลังข้อมูลฝึกฝน ดังนั้นระบบจึงสามารถสกัดชื่อองค์กรเหล่านี้ได้ แต่ทั้งนี้ชื่อย่อองค์กรที่จะตามหลังชื่อตำแหน่งก็มีเป็นส่วนน้อยในข้อมูลทดสอบทำให้ระบบรู้จำชื่อย่อองค์กรเหล่านี้ได้ไม่มากนัก

นอกจากนี้แม้ว่าข้อมูลที่ใช้ในระบบแบบตัดคำและตัดพยางค์จะเป็นชุดเดียวกัน แต่เมื่อผ่านการตัดแยกย่อยเป็นคำและพยางค์แล้วทำให้จำนวน token ที่ออกมาไม่เท่ากัน โดย token ของข้อมูลแบบตัดพยางค์มีทั้งหมด 487,364 พยางค์ ในขณะที่ข้อมูลแบบตัดคำมีทั้งหมด 367,673 คำ ดังนั้นจึงอาจกล่าวได้ว่าระบบที่ใช้ข้อมูลแบบตัดพยางค์มีข้อมูลที่ใช้ในการฝึกฝนมากกว่าข้อมูลแบบตัดคำ

สำหรับคุณสมบัติอื่น ๆ ได้ทดสอบโดยการใช้คุณสมบัติที่ต้องการทดสอบควบคู่ไปกับคุณสมบัติ unigram และ bigram ซึ่งเป็น template ของระบบ เมื่อประมวลผลทั้งหมด 10 ครั้ง (ดูภาคผนวก จ) ได้ค่าเฉลี่ยดังแสดงไว้ในตารางที่ 4.18-4.22 ดังนี้

ตารางที่ 4.18 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติรายการชื่อเฉพาะ

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 92.32 | 93.03 | 84.08 | 83.99 | 87.97 | 88.21 |
| ชื่อองค์กร | 85.83 | 83.14 | 69.42 | 70.60 | 76.71 | 76.33 |
| ชื่อสถานที่ | 81.21 | 79.88 | 69.66 | 69.44 | 74.90 | 74.19 |
| ทั้งหมด | 87.18 | 86.10 | 74.66 | 75.07 | 80.41 | 80.18 |

ตารางที่ 4.19 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติคำย่อ

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 91.29 | 93.28 | 81.00 | 84.13 | 85.78 | 88.40 |
| ชื่อองค์กร | 86.55 | 83.70 | 67.14 | 70.76 | 75.57 | 76.65 |
| ชื่อสถานที่ | 82.48 | 80.58 | 64.14 | 67.31 | 72.09 | 73.22 |
| ทั้งหมด | 87.46 | 86.60 | 71.28 | 74.63 | 78.52 | 80.14 |

ตารางที่ 4.20 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติคำบริบท

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 89.29 | 91.67 | 82.76 | 84.79 | 85.86 | 88.04 |
| ชื่อองค์กร | 84.63 | 82.11 | 67.61 | 70.45 | 75.12 | 75.79 |
| ชื่อสถานที่ | 80.91 | 79.38 | 65.85 | 68.01 | 72.55 | 73.15 |
| ทั้งหมด | 85.62 | 85.20 | 72.54 | 74.98 | 78.52 | 79.74 |

ตารางที่ 4.21 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติคำทั่วไป

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 91.02 | 92.91 | 80.47 | 84.03 | 85.36 | 88.18 |
| ชื่อองค์กร | 86.71 | 83.57 | 66.30 | 69.60 | 75.07 | 75.92 |
| ชื่อสถานที่ | 82.26 | 80.56 | 64.04 | 67.07 | 71.96 | 73.09 |
| ทั้งหมด | 87.33 | 86.48 | 70.75 | 74.12 | 78.15 | 79.80 |

ตารางที่ 4.22 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติค่าทางสถิติ

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 90.78 | 92.86 | 79.90 | 83.82 | 84.93 | 88.05 |
| ชื่อองค์กร | 87.04 | 83.70 | 64.74 | 69.00 | 74.18 | 75.61 |
| ชื่อสถานที่ | 82.79 | 80.69 | 64.17 | 67.52 | 72.24 | 73.39 |
| ทั้งหมด | 87.53 | 86.53 | 70.00 | 73.89 | 77.77 | 79.68 |

จากตารางที่ 4.18-4.22 แสดงให้เห็นว่าคุณสมบัติรายการชื่อเฉพาะช่วยให้ระบบการรู้จำชื่อเฉพาะมีประสิทธิภาพมากขึ้นกว่าคุณสมบัติอื่นในทั้งสองระบบเห็นได้จากค่าความครบถ้วนที่เพิ่มขึ้นจากระบบที่ใช้เฉพาะ templateเพียงอย่างเดียวจาก 70.06% เป็น 74.66% ในข้อมูลแบบตัดคำ และ 73.91% เป็น 75.07% ในข้อมูลแบบตัดพยางค์ นอกจากนี้ค่า F-measure ที่ได้ก็สูงกว่าระบบที่ใช้คุณสมบัติอื่นในขณะที่คุณสมบัติค่าทางสถิติทำให้ประสิทธิภาพการรู้จำของระบบลดลงโดยค่า F-measure ลดจาก 77.93% เหลือ 77.77% ในข้อมูลแบบตัดคำ และ 79.72% เหลือ 79.68% ในข้อมูลแบบตัดพยางค์ สำหรับคุณสมบัติอื่น ๆ ช่วยให้ประสิทธิภาพของระบบดีขึ้นเล็กน้อย

เมื่อพิจารณาจากค่า F-measure จะเห็นว่าในข้อมูลแบบตัดคำค่า F-measure ของระบบที่ใช้คุณสมบัติรายการชื่อเฉพาะเพิ่มขึ้นจากระบบที่ใช้เฉพาะ template เพียงอย่างเดียวถึง 2.48% ในขณะที่ระบบที่ใช้ข้อมูลแบบตัดพยางค์เพิ่มขึ้นเพียง 0.46% ทั้งนี้เนื่องมาจากลักษณะการกำหนดคุณสมบัติที่เอารายการชื่อเฉพาะไปตัดคำและพยางค์ก่อนแล้วนำมาสร้างเป็นรายการชื่อใหม่ จากนั้นจึงนำแต่ละ token มาเทียบดูว่าตรงกับส่วนใดของชื่อหรือไม่ โดยไม่ได้นำเอา token ข้างเคียงมาพิจารณาร่วมด้วย ลักษณะการทำเช่นนี้เป็นผลเสียกับข้อมูลแบบตัดพยางค์ เพราะการตัดที่แยกย่อยทำให้โอกาสที่แต่ละ token จะไปเป็นส่วนหนึ่งของชื่อมีค่อนข้างสูง เช่น คำว่า “ประชุม” ในข้อมูลแบบตัดคำจะถือว่าเป็นคำทั่วไปไม่ใช่ชื่อเฉพาะ แต่เมื่อแยกเป็นพยางค์แล้ว “ประ” สามารถเป็นส่วนหนึ่งของชื่อ “ประภา” ได้และ “ชุม” สามารถเป็นส่วนหนึ่งของชื่อ “ชุมพล” ได้เช่นกันดังนั้นจึงมีผลทำให้คุณสมบัติรายการชื่อเฉพาะช่วยให้ระบบที่ใช้ข้อมูลแบบตัดพยางค์รู้จำชื่อเฉพาะได้น้อยกว่าแบบตัดคำ

สำหรับสาเหตุที่คุณสมบัติค่าทางสถิติทำให้ประสิทธิภาพของระบบลดลงนั้น มาจากลักษณะการกำหนดคุณสมบัติที่ว่าหากหน่วยคำหรือพยางค์ 3 หน่วยปรากฏร่วมกันเกิน 3 ครั้งในข่าวจะกำหนดให้ token นั้นเป็น Y คือมีแนวโน้มว่า token นั้นอาจเป็นชื่อเฉพาะหรือส่วนของชื่อเฉพาะได้ แต่ปรากฏว่า token ที่มีค่าเป็น Y นั้น มักเป็น token ที่ไม่ใช่ชื่อเฉพาะ เนื่องจากคำทั่วไป

สามารถเกิดร่วมกันและปรากฏบ่อยครั้งในข่าวได้เช่นกัน เช่น “เป็น-ไข้หวัด-นก”, “เสีย-ชีวิต-<s>”, “งาน-ที่-เกี่ยว” เป็นต้น ดังนั้นระบบจึงนำส่วนที่ไม่ใช่ชื่อเฉพาะมาประมวลผลด้วยทำให้ประสิทธิภาพของระบบลดลง

สำหรับคุณสมบัติอื่น เมื่อพิจารณาจากค่า F-measure แล้วจะเห็นว่าคุณสมบัติที่ช่วยให้ประสิทธิภาพของระบบดีขึ้นรองลงมา คือ คุณสมบัติคำย่อ แต่ค่าความครบถ้วนของคุณสมบัติคำย่อกลับต่ำกว่าคุณสมบัติคำบริบท นั่นเพราะคุณสมบัติคำย่อส่วนใหญ่จะช่วยระบุขอบเขตและชนิดของชื่อองค์กร เช่น ชื่อย่อของชื่อองค์กรเอง คำย่อของชื่อตำแหน่ง เช่น รมว. รมช. เป็นต้น ดังนั้นจึงทำให้ค่าความครบถ้วนของชื่อองค์กรเพิ่มจากการใช้เพียงคุณสมบัติ unigram และ bigram อย่างเดียวจาก 64.53% เป็น 67.14% ในข้อมูลแบบตัดคำและจาก 68.61% เป็น 70.76% ในข้อมูลแบบตัดพยางค์ สำหรับชื่อบุคคลจะมีชื่อจำนวนหนึ่งที่มีคำนำหน้าชื่อเป็นคำย่อ เช่น ด.ญ. ด.ช. ดร. พ.ต.ท. เป็นต้น แต่ก็มีชื่ออีกจำนวนไม่น้อยที่ใช้คำนำหน้าชื่อเป็นคำอื่น เช่น นาย นาง นายแพทย์ คำเรียกญาติ เช่น พี่ น้อง ตา เป็นต้น ทำให้ค่าความครบถ้วนของชื่อบุคคลเพิ่มขึ้นเพียงเล็กน้อยเท่านั้น ในขณะที่ชื่อสถานที่ส่วนใหญ่มักปรากฏโดยไม่มีคำย่อบ่งชี้ เช่น ชื่อประเทศ ชื่อเมือง ชื่อจังหวัด ทำให้คุณสมบัติคำย่อไม่ช่วยให้ประสิทธิภาพการรู้จำชื่อสถานที่ดีขึ้น

สำหรับคุณสมบัติคำบริบทนั้น เนื่องจากคำบริบทที่ใช้คือช่วงก่อนหน้าและหลัง token 3 หน่วยในข้อมูลตัดคำและ 4 หน่วยในข้อมูลแบบตัดพยางค์ในคลังข้อมูลฝึกฝน เมื่อช่วงข้อมูลกว้างทำให้ปริมาณคำหรือพยางค์ที่ใช้มีจำนวนมากอีกทั้งคำบริบทที่คาดว่าจะช่วยระบุขอบเขตหรือประเภทของชื่อเฉพาะ เช่น “ที่” หรือ “ของ” ซึ่งใช้ช่วยในการระบุประเภทของชื่อเฉพาะประเภทองค์กรและสถานที่ เมื่อเทียบสัดส่วนของคำบริบทเหล่านี้ที่เกิดร่วมกับชื่อเฉพาะกับเกิดร่วมกับคำอื่น จะพบว่าปริมาณครั้งที่เกิดร่วมกับชื่อเฉพาะมีน้อยมาก เช่น ในข้อมูลตัดคำสำหรับฝึกฝนชุดแรก พบว่า “ที่” เกิดกับชื่อเฉพาะในช่วงบริบท 3 คำก่อนหน้าและหลังชื่อเฉพาะทั้งหมด 1,123 ครั้ง แต่เกิดในบริบทอื่นถึง 5,199 ครั้ง เป็นต้น ดังนั้นเมื่อนำเอารายการคำหรือพยางค์ที่ได้มาเทียบกับแต่ละ token ในคลังข้อมูลทดสอบ ทำให้ token ส่วนใหญ่มีค่าเป็น Y แต่ทั้งนี้เนื่องจากคำบริบทมักเป็นคำทั่วไปทำให้ส่วนที่เป็นชื่อเฉพาะในข้อมูลตัดคำส่วนใหญ่มีค่าเป็น N เมื่อเป็นเช่นนี้จึงมีแนวโน้มว่าจะช่วยให้ระบบสามารถสกัดชื่อเฉพาะบางชื่อออกมาได้ แต่สำหรับข้อมูลแบบตัดพยางค์เนื่องจากการตัดพยางค์จะย่อยกว่า ทำให้ส่วนของชื่อเฉพาะที่กำหนดเป็น N ในข้อมูลตัดคำกลายเป็น Y ในข้อมูลตัดพยางค์ เช่น กรมประชาสัมพันธ์ ในข้อมูลตัดคำคือคำ ๆ เดียวแต่ในข้อมูลตัดพยางค์แยกย่อยเป็น กรม-ประ-ชา-สัม-พันธ์ ทำให้ทุกพยางค์มีค่าเป็น Y เนื่องจากแต่ละพยางค์นี้อาจจะไปตรงกับพยางค์ของคำอื่น ดังนั้นจึงมีผลทำให้ค่าความครบถ้วนของระบบที่ใช้ข้อมูลตัดพยางค์เพิ่มขึ้นเพียง 1% ในขณะที่ระบบที่ใช้ข้อมูลตัดคำเพิ่มขึ้นประมาณ 2.5%

กรณีคุณสมบัติคำทั่วไปจะเห็นว่ามีผลต่อระบบทั้งแบบตัดคำและตัดพยางค์เพียงเล็กน้อย เท่านั้นในระบบตัดคำคุณสมบัตินี้จะใช้ไม่ได้กับคำที่มีลักษณะเป็นวลี เช่น นอกจากนี้ วันนี้ การศึกษา คณะกรรมการ เป็นต้น เพราะวลีเหล่านี้ไม่มีอยู่ในรายการคำทั่วไป วลีเหล่านี้จึงมีค่าเป็น Y เช่นเดียวกับส่วนหรือคำที่เป็นชื่อเฉพาะ ในขณะที่ในระบบตัดพยางค์ คำเหล่านี้จะไม่ค่อยเป็นปัญหา เนื่องจากการตัดแบ่งเป็นพยางค์ย่อยหลายพยางค์ทำให้โอกาสที่พยางค์เหล่านี้จะตรงกับรายการคำมีสูง จึงช่วยให้สามารถตัดคำเหล่านี้ออกไปได้ แต่ระบบตัดพยางค์จะมีปัญหาค่อนข้างมากกรณีการตัดแบ่งชื่อเฉพาะออกเป็นพยางค์ย่อย ๆ ซึ่งพยางค์เหล่านี้จะไปตรงกับคำทั่วไป เช่น กองทัพบก จัดเป็นคำ ๆ เดียวในข้อมูลตัดคำจึงมีค่าเป็น Y เพราะไม่พบในรายการคำทั่วไป แต่ในข้อมูลตัดพยางค์จะกลายเป็น 3 พยางค์ คือ กอง-ทัพ-บก ซึ่งทั้ง 3 พยางค์นี้มีค่าเป็น N ดังนั้นจากตัวอย่างดังกล่าว จะเห็นได้ว่าระบบที่ใช้ข้อมูลตัดพยางค์สามารถนำส่วนที่เป็นชื่อเฉพาะจริง ๆ มาคำนวณในฟังก์ชันคุณสมบัติได้น้อยกว่าระบบที่ใช้ข้อมูลตัดคำ

จากตารางที่ 4.18-4.22 เมื่อพิจารณาจากค่า F-measure จะเห็นว่าค่า F-measure ของแบบจำลองที่ใช้ข้อมูลแบบตัดพยางค์สูงกว่าแบบที่ใช้ข้อมูลแบบตัดคำเกือบทุกแบบ ยกเว้นแบบจำลองที่ใช้คุณสมบัติรายการชื่อเฉพาะเท่านั้น ทั้งนี้ไม่ได้หมายความว่าคุณสมบัติเหล่านี้ใช้กับข้อมูลแบบตัดพยางค์ได้ดีกว่าตามเหตุผลข้างต้นที่กล่าวมา สาเหตุที่ค่า F-measure ของแบบจำลองที่ใช้ข้อมูลแบบตัดพยางค์สูงกว่านั้นมาจาก คุณสมบัติ unigram และ bigram ซึ่งเป็น template ที่ใช้ประมวลผลควบคู่กับคุณสมบัติต่าง ๆ ดังแสดงในตารางที่ 4.16 ดังนั้นหากคุณสมบัติที่นำมาใช้ประมวลผลร่วมด้วยนั้นไม่ได้ช่วยระบบใดระบบหนึ่งอย่างชัดเจนดังเช่นคุณสมบัติรายการชื่อเฉพาะ ค่า F-measure ของระบบที่ใช้ข้อมูลแบบตัดพยางค์จึงจะยังคงสูงกว่าแบบตัดคำเสมอ แต่เมื่อนำเอาคุณสมบัติทั้งหมดมาประมวลผลเข้าด้วยกันจะได้ผลดังแสดงในตารางที่ 4.9 คือ ประสิทธิภาพของแบบจำลองที่ใช้ข้อมูลแบบตัดคำและตัดพยางค์ไม่ต่างกัน กล่าวคือ คุณสมบัติ unigram และ bigram สนับสนุนข้อมูลแบบตัดพยางค์ ในขณะที่คุณสมบัติรายการชื่อเฉพาะสนับสนุนข้อมูลแบบตัดคำ

เมื่อประเมินประสิทธิภาพของระบบโดยใช้จำนวน token ได้ผลออกมาดังแสดงไว้ในตารางที่ 4.23-4.27

ตารางที่ 4.23 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติรายการชื่อเฉพาะโดยวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 96.58 | 96.88 | 91.04 | 91.71 | 93.68 | 94.19 |
| ชื่อองค์กร | 87.94 | 87.15 | 70.03 | 75.67 | 77.93 | 80.98 |
| ชื่อสถานที่ | 87.52 | 86.29 | 74.52 | 73.35 | 80.39 | 79.21 |
| ทั้งหมด | 92.23 | 91.41 | 80.65 | 81.80 | 86.03 | 86.32 |

ตารางที่ 4.24 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติคำย่อโดยวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 95.40 | 96.39 | 89.57 | 91.73 | 92.32 | 93.96 |
| ชื่อองค์กร | 88.25 | 86.91 | 67.58 | 75.94 | 76.48 | 81.03 |
| ชื่อสถานที่ | 88.81 | 87.21 | 70.24 | 71.62 | 78.37 | 78.55 |
| ทั้งหมด | 92.11 | 91.34 | 78.27 | 81.52 | 84.60 | 86.13 |

ตารางที่ 4.25 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติคำบริบทโดยวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 91.67 | 94.36 | 91.34 | 92.39 | 91.43 | 93.31 |
| ชื่อองค์กร | 84.71 | 84.93 | 69.86 | 76.15 | 76.53 | 80.29 |
| ชื่อสถานที่ | 86.51 | 86.26 | 72.54 | 73.05 | 78.82 | 79.00 |
| ทั้งหมด | 88.75 | 89.60 | 80.36 | 82.22 | 84.32 | 85.73 |

ตารางที่ 4.26 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติคำทั่วไปโดยวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 95.17 | 95.57 | 89.03 | 91.74 | 91.91 | 93.56 |
| ชื่อองค์กร | 88.35 | 86.96 | 67.29 | 75.35 | 76.34 | 80.72 |
| ชื่อสถานที่ | 88.69 | 87.18 | 70.09 | 71.33 | 78.24 | 78.39 |
| ทั้งหมด | 91.98 | 91.07 | 77.93 | 81.28 | 84.34 | 85.88 |

ตารางที่ 4.27 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติค่าทางสถิติโดยวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 94.44 | 95.43 | 88.87 | 91.63 | 91.48 | 93.43 |
| ชื่อองค์กร | 87.95 | 86.59 | 67.64 | 75.30 | 76.40 | 80.53 |
| ชื่อสถานที่ | 89.10 | 87.30 | 70.04 | 71.62 | 78.34 | 78.58 |
| ทั้งหมด | 91.58 | 90.89 | 77.92 | 81.25 | 84.17 | 85.78 |

เมื่อวัดประสิทธิภาพของระบบโดยใช้ token ได้ผลออกมาต่างจากการใช้จำนวนชื่อเล็กน้อย กล่าวคือ หากมีการนับรวมชื่อเฉพาะที่ถูกเพียงบางส่วนเข้าไปด้วย จะมีเพียงคุณสมบัติรายการชื่อเฉพาะและคุณสมบัติคำย่อเท่านั้นที่ช่วยให้ค่า F-measure ของระบบมากขึ้น ในขณะที่คุณสมบัติอื่นที่เหลือทำให้ค่า F-measure ของระบบลดลงเล็กน้อยเมื่อเทียบกับการใช้คุณสมบัติ unigram และ bigram เพียงอย่างเดียวในตารางที่ 4.17 แต่สิ่งที่เหมือนกับการวัดประสิทธิภาพโดยใช้จำนวนชื่อ คือ คุณสมบัติรายการชื่อเฉพาะสนับสนุนระบบที่ใช้ข้อมูลแบบตัดคำ แม้ว่าเมื่อดูจากค่า F-measure ระบบที่ใช้ข้อมูลแบบตัดคำจะน้อยกว่าแบบตัดพยางค์ดังแสดงในตารางที่ 4.23 แต่เมื่อเทียบกับระบบที่ใช้เฉพาะคุณสมบัติ unigram และ bigram แล้วพบว่าค่า F-measure ของระบบที่ใช้ข้อมูลตัดคำเพิ่มขึ้น 1.62% ในขณะที่ระบบที่ใช้ข้อมูลตัดพยางค์เพิ่มขึ้นเพียง 0.27% เท่านั้น

เมื่อพิจารณาชื่อเฉพาะที่ระบบไม่สามารถสกัดออกมาได้นั้น สาเหตุส่วนหนึ่งมาจากชื่อนั้นเกิดในบริบทที่ไม่ชัดเจนหรือไม่มีคำบ่งชี้ เช่น กรณีชื่อบุคคลเกิดโดยไม่มีคำนำหน้าชื่อ เช่น ระบบสามารถรู้จำชื่อ “นายสมคิด” ได้ แต่จะไม่รู้จำว่า “สมคิด” คือชื่อเฉพาะเช่นกัน เนื่องจาก “สมคิด”

เกิดโดยไม่มีคำบ่งชื่อเฉพาะ ดังนั้นระบบจึงเข้าใจว่าเป็นคำทั่วไป หรือชื่อย่อองค์กรที่มักปรากฏโดยไม่มีคำบ่งชี้ ทำให้ระบบมีปัญหาในการรู้จำชื่อย่อองค์กรค่อนข้างมาก แต่ทั้งนี้หากชื่อย่อนั้นเกิดในวงเล็บตามหลังชื่อเต็มขององค์กรระบบจะสามารถรู้จำชื่อย่อนั้นได้มากกว่า จากตัวอย่างที่กล่าวมาจะเห็นได้ว่าชื่อเฉพาะเดียวกันแต่เกิดในบริบทต่างกัน ระบบจะสามารถรู้จำชื่อเฉพาะนั้นได้ในบางบริบทเท่านั้น ซึ่งลักษณะเช่นนี้ ผู้วิจัยเห็นว่าเราสามารถเขียนกฎเพื่อนำชื่อเฉพาะที่ระบบสกัดได้ มาใช้ช่วยในการสกัดชื่อเฉพาะส่วนที่เหลือออกมา ดังนั้นผู้วิจัยจึงเพิ่มส่วนของการประมวลผลภายหลัง (post processing) เพื่อใช้ช่วยสกัดชื่อเฉพาะที่ระบบไม่สามารถสกัดได้ รายละเอียดขั้นตอนการประมวลผลภายหลังจะกล่าวในหัวข้อถัดไป

## 4.4 ขั้นตอนประมวลผลภายหลัง

ขั้นตอนประมวลผลภายหลังระบบการรู้จำชื่อเฉพาะใช้เพื่อช่วยสกัดชื่อเฉพาะที่ระบบไม่สามารถรู้จำได้ เป็นการช่วยเพิ่มประสิทธิภาพของระบบโดยขั้นตอนประมวลผลภายหลังของระบบที่ใช้ข้อมูลตัดคำและตัดพยางค์จะเหมือนกัน

### 4.4.1 ชื่อเฉพาะบุคคล

1) กรณีระบบรู้จำส่วนที่ประกอบด้วยตัวเลขหรือช่องว่างหลายช่องเป็นชื่อบุคคล

โดยปกติชื่อบุคคลมักไม่มีตัวเลขเป็นส่วนหนึ่งของชื่อ แต่เนื่องจากในการเขียนตัวเลขเรามักจะใช้ช่องว่างในการคั่นหน้าและหลังตัวเลข ซึ่งลักษณะเช่นนี้จะคล้ายกับชื่อบุคคลเพราะชื่อบุคคลบางครั้งจะมีช่องว่างคั่นระหว่างคำนำหน้าชื่อกับชื่อ และชื่อกับนามสกุล ดังนั้นในบางบริบทระบบจะรู้จำให้ส่วนที่เป็นตัวเลขเป็นชื่อบุคคล

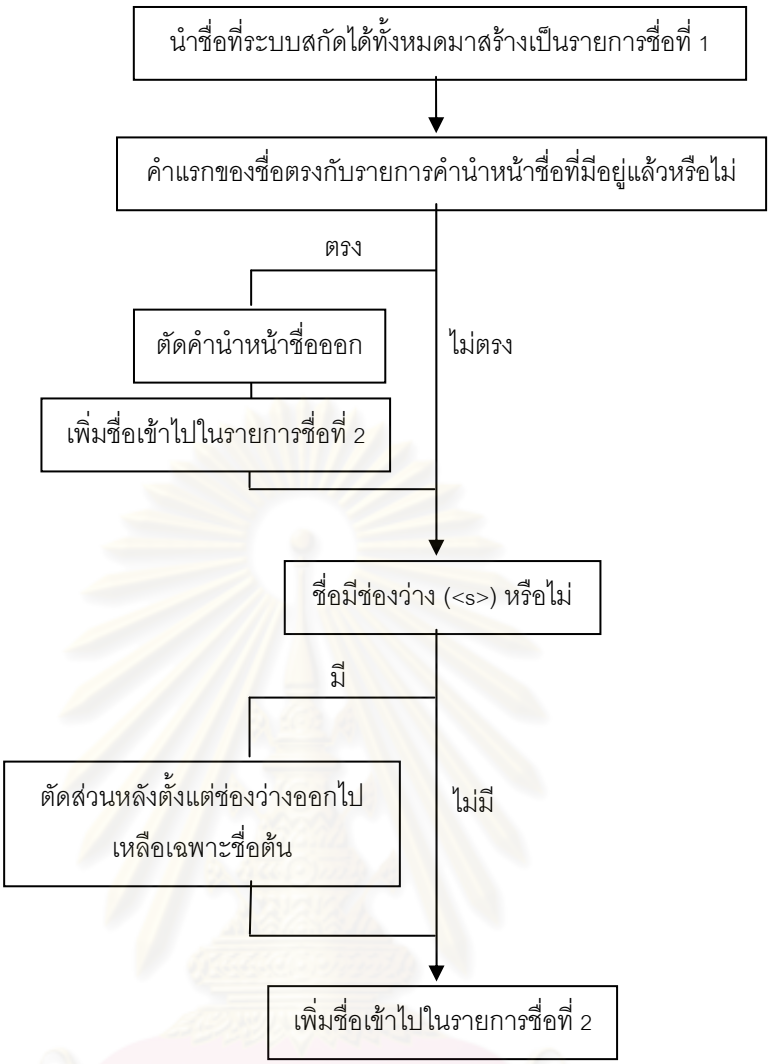
นอกจากนี้หากพิจารณาชื่อบุคคลจากในคลังข้อมูลจะพบว่าสามารถมีช่องว่างได้ประมาณ 4 ช่อง คือ ช่องว่างระหว่างคำนำหน้าชื่อกับชื่อ ชื่อกับนามสกุล และภายในนามสกุลสามารถมีช่องว่างได้อีก 2 ช่อง เช่น “คุณ วิเชษฐ์ กุญชร ณ อยุธยา” เป็นต้น แต่ในจำนวนชื่อที่ระบบสกัดออกมานั้น มีหลายชื่อที่ประกอบด้วยช่องว่างมากกว่า 4 ช่อง ซึ่งส่วนใหญ่มักเป็นข้อความที่มีการเว้นวรรค ตัวอย่างชื่อที่ระบบสกัดออกมาผิด เช่น “พื้นเมือง<s>63,091,577<s>ตัว<s>เป็ดไข่<s>8,878,593<s>ตัว<s>เป็ดเนื้อ<s>877,348<s>ตัว” เป็นต้น

ดังนั้นในขั้นตอนนี้จึงจะไปตรวจสอบชื่อเฉพาะบุคคลที่ระบบสกัดออกมาแล้วว่าชื่อนั้นประกอบด้วยตัวเลขหรือไม่ และช่องว่างที่อยู่ภายในชื่อมากกว่า 4 ช่องหรือไม่ ซึ่งหากว่าชื่อที่สกัดออกมานั้นตรงกับเงื่อนไขใดเงื่อนไขหนึ่งจะกำหนดให้ชื่อนั้นไม่ใช่ชื่อเฉพาะ

2) กรณีระบบไม่รู้จำชื่อบุคคลในบริบทต่าง ๆ

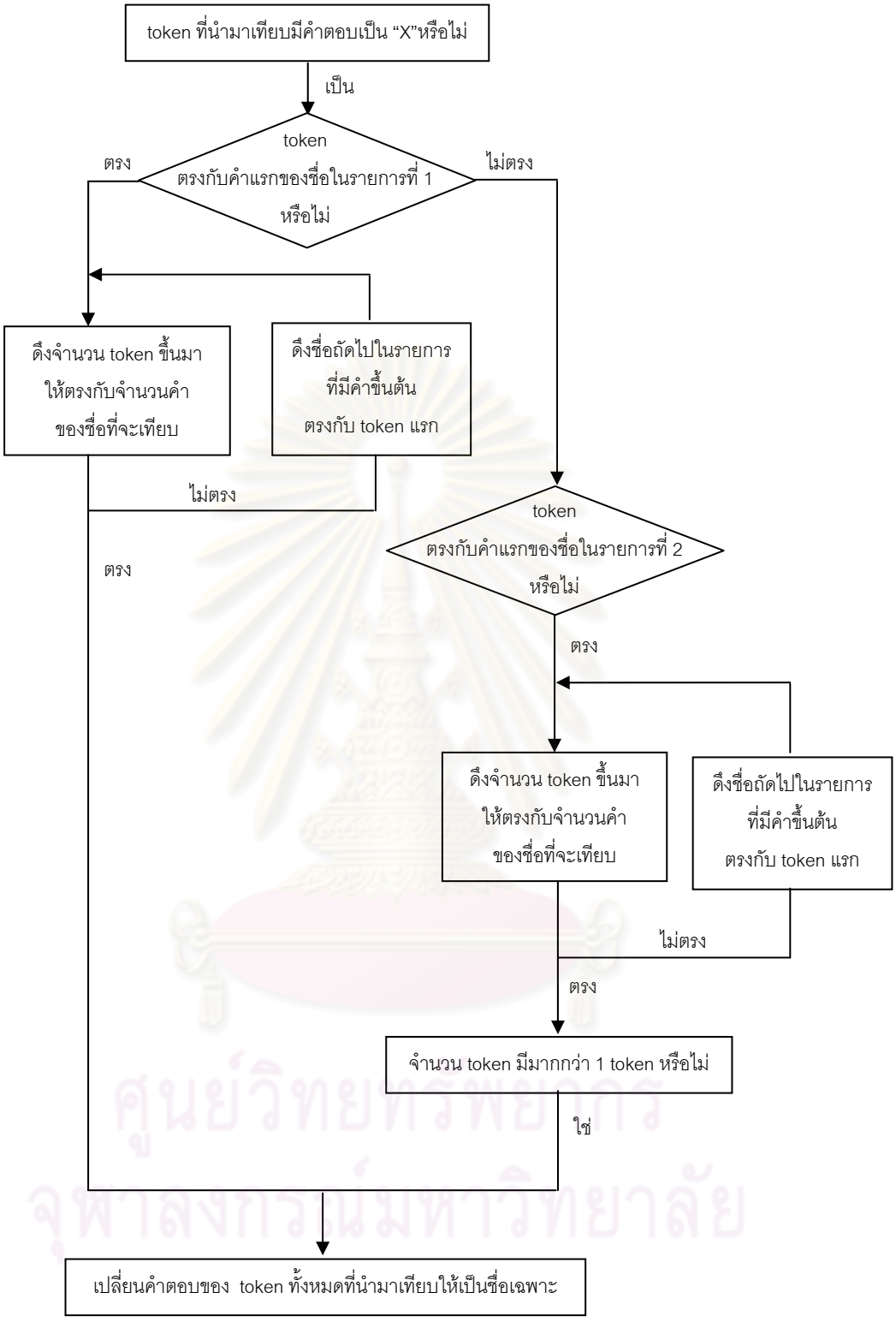
มีหลายกรณีที่ชื่อบุคคลเกิดโดยไม่มีคำนำหน้าชื่อ หรือเมื่อปรากฏครั้งแรกเป็นชื่อและนามสกุลแต่ครั้งต่อไปจะเหลือเพียงชื่อเท่านั้น หรือบางกรณีจะมีการบอกชื่อจริงและชื่อเล่นในตอนแรกและครั้งต่อไปใช้ชื่อเล่นแทน ทำให้บางครั้งระบบไม่สามารถรู้จำชื่อออกมาได้หมด ขั้นตอนนี้จึงเพิ่มขึ้นมาเพื่อช่วยสกัดชื่อที่เกิดในลักษณะดังกล่าว โดยจะนำเอาชื่อที่ระบบสามารถสกัดออกมาได้ ชื่อที่ตัดคำนำหน้าชื่อ และชื่อที่ตัดคำนำหน้าชื่อและนามสกุลมาเปรียบเทียบกับ token ที่ได้รับคำตอบเป็น “X” สำหรับชื่อที่นำมาตัดคำนำหน้าชื่อและนามสกุลนั้นได้จากชื่อที่ระบบสกัดออกมาจากข้อมูล ในขั้นตอนนี้ไม่ได้ใช้ชื่อที่ตัดเฉพาะนามสกุลมาเทียบด้วย เพราะส่วนใหญ่ระบบไม่ค่อยมีปัญหากับการรู้จำชื่อในลักษณะนี้ ดังนั้นชื่อที่ตัดเฉพาะนามสกุลจึงจะรวมอยู่ในชื่อที่ระบบสามารถสกัดออกมาได้อยู่แล้ว

ขั้นตอนที่ใช้ช่วยรู้จำชื่อบุคคลที่ระบบไม่สามารถสกัดออกมาได้แบ่งออกเป็น 2 ขั้นตอน ดังนี้ ขั้นตอนสร้างรายการชื่อเฉพาะที่ระบบสกัดออกมาได้จากข้อมูล และขั้นตอนการรู้จำชื่อเฉพาะดังแสดงในแผนภาพที่ 4.4 และ 4.5 ตามลำดับ




ภาพที่ 4.4 ขั้นตอนการสร้างรายการชื่อเฉพาะ

ในขั้นตอนการสร้างรายการชื่อเฉพาะและการรู้จำชื่อเฉพาะนี้ ข้อมูลแบบตัดพยางค์จะใช้ขั้นตอนเดียวกันกับข้อมูลแบบตัดคำ รายการชื่อเฉพาะที่ได้ในขั้นตอนการสร้างรายการชื่อเฉพาะนี้จะต่างจากรายการชื่อเฉพาะที่นำมาใช้เป็นคุณสมบัติ เพราะเป็นชื่อที่มาจากการสกัดของระบบ รายการชื่อเฉพาะจะมี 2 รายการ โดยรายการแรกคือรายการชื่อที่ได้จากการสกัดจากระบบโดยตรง ส่วนรายการที่สองคือนำชื่อจากรายการแรกมาตัดคำนำหน้าชื่อออก เพื่อเหลือเพียงชื่อและชื่อกับนามสกุล จากนั้นจะสกัดชื่อออกจากนามสกุลอีกครั้งโดยตรวจดูว่าส่วนที่เป็นชื่อมีช่องว่างหรือไม่เพราะโดยปกติแล้วชื่อกับนามสกุลจะมีช่องว่างคั่น หากมีช่องว่างจะตัดส่วนที่เป็นนามสกุลทิ้งไปเหลือเพียงแค่ชื่อเท่านั้น




ภาพที่ 4.5 ขั้นตอนการรู้จำชื่อเฉพาะของระบบตัดคำ

สำหรับขั้นตอนการรู้จำชื่อเฉพาะนั้นจะนำเอา token ที่ได้รับคำตอบเป็น “X” มาเทียบกับรายการชื่อใหม่ที่ได้จากขั้นตอนก่อนหน้านี้ โดยจะเทียบกับรายการชื่อที่ 1 ก่อนว่า token นั้นตรงกับคำแรกหรือพยางค์แรกของชื่อใดหรือไม่ หากตรงให้ดึง token ขึ้นมาให้เท่ากับ

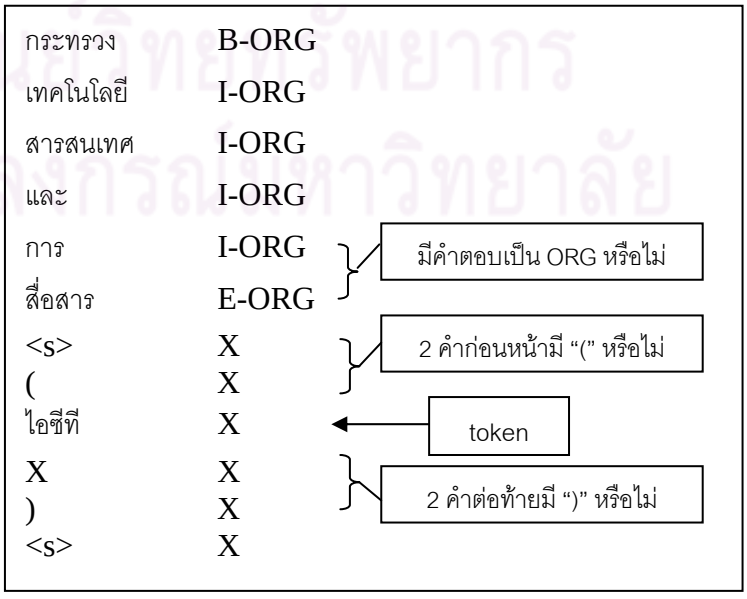
จำนวนคำหรือพยางค์ของชื่อนั้นแล้วนำมาเทียบกัน แต่หากไม่พบในรายการชื่อที่ 1 จะนำมาเทียบกับรายการชื่อที่ 2 ว่ามีหรือไม่หากมีจะทำตามขั้นตอนดังเช่นรายการที่ 1 แต่สำหรับการนำมาเทียบกับรายการชื่อที่ 2 จะเพิ่มขั้นตอนว่าจะไม่ให้รู้จำชื่อที่มีคำเพียงคำเดียวหรือพยางค์ ๆ เดียว เนื่องจากชื่อที่ได้ในรายการที่ 2 คือชื่อที่มีการตัดคำนำหน้าชื่อและนามสกุลออก ดังนั้นหากเป็นคำพยางค์เดียวก็มีแนวโน้มจะไปซ้ำกับคำทั่วไปได้ เช่น นาม กบ เป็นต้น ซึ่งจะมีผลทำให้รู้จำชื่อเฉพาะผิดได้อีกทั้งส่วนใหญ่ชื่อมักปรากฏโดยมีคำนำหน้าชื่ออยู่ด้วย ดังนั้นโอกาสที่ชื่อจะปรากฏเป็นคำ ๆ เดียวหรือพยางค์เดียวจึงมีน้อย แต่สาเหตุที่ในขั้นตอนของรายการชื่อที่ 1 ไม่มีการเทียบจำนวน token นี้ เพราะชื่อที่ได้จากรายการที่ 1 มักเป็นคำที่มากกว่า 1 คำหรือมากกว่า 1 พยางค์ เพราะประกอบด้วยคำนำหน้าชื่อและนามสกุล หากแม้เป็นชื่อคำเดียวก็มีแนวโน้มว่าจะเป็นชื่อเฉพาะที่พบบ่อยในข้อมูล เช่น ทักษิณ เป็นต้น อีกทั้งเมื่อดูจากค่าความแม่นยำแล้วจะเห็นว่าชื่อบุคคลได้ 92.05% และ 92.55% ในข้อมูลแบบตัดคำและตัดพยางค์ตามลำดับ ดังนั้นจึงอนุมานได้ว่าชื่อบุคคลที่ระบบสกัดออกมามีความถูกต้องที่พอเชื่อถือได้

### 4.4.2 ชื่อเฉพาะองค์กร

1) กรณีที่ชื่อย่ออยู่ภายในวงเล็บต่อจากชื่อองค์กร

ส่วนใหญ่ชื่อองค์กรที่ยาวมักมีชื่อย่อขององค์กรนั้น ๆ อยู่ในวงเล็บต่อท้าย ดังนั้นด้วยลักษณะเช่นนี้จึงสามารถนำมาใช้เขียนกฎในการสกัดชื่อองค์กรได้ดังนี้

สำหรับข้อมูลแบบตัดคำ หาก token ใดที่พบ “(” ในช่วง 2 คำก่อนหน้าและคำตอบของ 2 คำก่อนหน้า “(” เป็นองค์กร และพบ “)” ช่วง 2 คำต่อจาก token ให้อนุมานได้ว่า token นั้นเป็นชื่อองค์กร สำหรับข้อมูลแบบตัดพยางค์ได้เพิ่มช่วงของคำจาก 2 คำเป็น 3 พยางค์ ตัวอย่างของข้อมูลที่นำมาใช้พิจารณาเป็นดังตัวอย่างภาพที่ 4.6




ภาพที่ 4.6 ตัวอย่างข้อมูลที่นำมาพิจารณาเมื่อผ่านกฎชื่อย่อในวงเล็บในข้อมูลแบบตัดคำ

ตำแหน่ง token ที่พิจารณา คือ “ไอซีที” จากภาพตัวอย่างเมื่อ “ไอซีที” ผ่านกฎนี้แล้วจะเปลี่ยนจาก “X” เป็น “B-ORG”

2) กรณีที่ระบบไม่ได้สกัดชื่อย่อขององค์กรในบริบทต่าง ๆ

โดยปกติชื่อองค์กรเมื่อกล่าวถึงในครั้งแรกมักเป็นชื่อเต็ม แต่เมื่อกล่าวถึงในครั้งต่อไปมักอ้างถึงโดยใช้ชื่อย่อซึ่งบางครั้งเมื่อชื่อย่อไปปรากฏในบริบทที่คลุมเครือไม่ชัดเจนระบบอาจจะไม่สามารถสกัดชื่อเฉพาะนั้นออกมาได้

กฎที่เขียนขึ้นเพื่อสกัดชื่อย่อองค์กรในรูปแบบนี้คือ นำชื่อย่อองค์กรทั้งหมดที่ระบบสามารถสกัดออกมาได้มาทำเป็นรายการชื่อย่อ จากนั้นนำรายการที่ได้นี้ไปเปรียบเทียบกับแต่ละ token ในข้อมูล หาก token ใดตรงกับชื่อในรายการให้อนุมานว่า token นั้นเป็นชื่อเฉพาะ

### 4.4.3 ชื่อเฉพาะอ้างข้ามประเภท

สำหรับชื่อเฉพาะที่ใช้อ้างข้ามประเภทนี้จะใช้กับชื่อองค์กรและชื่อสถานที่เท่านั้น โดยนำเอาชื่อสถานที่และชื่อองค์กรทั้งหมดที่ระบบดึงออกมาได้มาเก็บเป็นรายการชื่อเฉพาะแยกจากกัน จากนั้นนำมาเทียบกันว่าชื่อใดที่สามารถเป็นได้ทั้งสองประเภท และมีจำนวนครั้งในการเกิดเป็นประเภทใดมากกว่ากันในตัวบทที่ประมวลผล หากชื่อนั้นมีจำนวนครั้งในการเกิดเป็นชื่อองค์กรมากกว่าก็จะเก็บชื่อนั้นอยู่ในรายการของชื่อองค์กร แต่ถ้าเกิดเป็นชื่อสถานที่มากกว่าก็จะเก็บอยู่ในรายการชื่อสถานที่ เนื่องจากชื่อสามารถเป็นได้ทั้งสองประเภท ดังนั้นจำนวนครั้งจะช่วยให้สามารถทำนายได้ว่าชื่อนั้นมีแนวโน้มจะเกิดเป็นประเภทใดมากกว่า จากนั้นนำรายการทั้งสองไปเทียบกับ token ที่ได้รับคำตอบเป็น “X” หาก token นั้นตรงกับรายการชื่อเฉพาะใดก็จะกำกับใหม่ให้เป็นชื่อเฉพาะประเภทนั้น

### 4.4.4 ชื่อเฉพาะสถานที่

กรณีของชื่อเฉพาะสถานที่จะนำเอา token ที่ได้รับคำตอบเป็น “X” มาเทียบกับรายการชื่อประเทศโดยตรง หาก token นั้นตรงกับชื่อประเทศใดจะกำกับใหม่ให้เป็นชื่อสถานที่ โดยขั้นตอนนี้จะทำหลังจากที่ข้อมูลผ่านกฎสำหรับชื่อเฉพาะอ้างข้ามประเภทแล้วเนื่องจากชื่อสถานที่สามารถเป็นชื่อองค์กรได้เช่นกัน

หลังจากที่นำข้อมูลมาผ่านขั้นตอนประมวลผลภายหลัง พบว่าค่าความครบถ้วนของข้อมูลทั้งสองแบบเพิ่มขึ้น แต่ค่า F-measure เพิ่มขึ้นเฉพาะข้อมูลแบบตัดคำเท่านั้นในขณะที่ข้อมูลแบบตัดพยางค์กลับลดลง รายละเอียดของประสิทธิภาพของระบบที่ไม่ผ่านกระบวนการประมวลผล

ภายหลังและระบบที่ผ่านกระบวนการประมวลผลภายหลังแบบประเมินจากจำนวนชื่อได้แสดงไว้ในตารางที่ 4.28 และ 4.29 (ดูรายละเอียดการประมวลผลทั้ง 10 ครั้งได้จากภาคผนวก ฉ)

ตารางที่ 4.28 ประสิทธิภาพของระบบที่ไม่ผ่านกระบวนการประมวลผลภายหลัง

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 92.05 | 92.55 | 86.50 | 85.83 | 89.16 | 89.01 |
| ชื่อองค์กร | 82.19 | 81.84 | 74.23 | 74.58 | 77.99 | 78.02 |
| ชื่อสถานที่ | 79.92 | 79.69 | 70.77 | 70.99 | 74.98 | 74.98 |
| ทั้งหมด | 85.37 | 85.37 | 77.64 | 77.64 | 81.30 | 81.30 |

ตารางที่ 4.29 ประสิทธิภาพของระบบเมื่อผ่านกระบวนการประมวลผลภายหลัง

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 90.00 | 80.10 | 88.50 | 88.15 | 89.20 | 83.80 |
| ชื่อองค์กร | 79.09 | 77.54 | 77.43 | 77.33 | 78.22 | 77.41 |
| ชื่อสถานที่ | 77.77 | 76.53 | 72.72 | 72.78 | 75.08 | 74.55 |
| ทั้งหมด | 82.74 | 78.24 | 80.15 | 80.06 | 81.40 | 79.11 |

ในขั้นตอนประมวลผลภายหลังนี้ เมื่อดูจากค่าความครบถ้วนจะเห็นว่าทั้งสองระบบเพิ่มขึ้นมาประมาณ 2.5% แต่ค่าความแม่นยำกลับลดลงโดยเฉพาะข้อมูลแบบตัดพยางค์ที่ลดลงถึง 7.13% จึงส่งผลให้ค่า F-measure ซึ่งเป็นค่าเฉลี่ยของค่าความแม่นยำและค่าความครบถ้วนลดลงตามไปด้วย

สำหรับผลการประเมินโดยใช้จำนวน token ได้แสดงไว้ในตารางที่ 4.30 และ 4.31 ดังนี้

ตารางที่ 4.30 ประสิทธิภาพของระบบที่ไม่ผ่านกระบวนการประมวลผลภายหลังโดยวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 94.56 | 95.78 | 93.10 | 92.82 | 93.78 | 94.23 |
| ชื่อองค์กร | 82.20 | 84.44 | 75.24 | 78.71 | 78.54 | 81.46 |
| ชื่อสถานที่ | 86.20 | 86.22 | 76.27 | 75.48 | 80.80 | 80.36 |
| ทั้งหมด | 89.14 | 89.91 | 83.61 | 83.82 | 86.27 | 86.75 |

ตารางที่ 4.31 ประสิทธิภาพของระบบเมื่อผ่านกระบวนการประมวลผลภายหลังโดยวัดจากจำนวน token

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 95.18 | 91.14 | 93.49 | 93.77 | 94.30 | 92.43 |
| ชื่อองค์กร | 80.78 | 81.95 | 77.54 | 80.53 | 79.10 | 81.22 |
| ชื่อสถานที่ | 85.14 | 84.08 | 77.42 | 76.84 | 80.97 | 80.19 |
| ทั้งหมด | 88.59 | 86.53 | 84.78 | 85.21 | 86.63 | 85.86 |

เมื่อดูผลจากการประเมินโดยใช้จำนวน token จะเห็นว่าค่าความแม่นยำของระบบที่ใช้ข้อมูลตัดพยางค์ลดลงเช่นกัน สาเหตุที่ทำให้ค่าความแม่นยำลดลงนั้น มาจากขั้นตอนที่เน้นไปที่การสกัดชื่อเฉพาะในส่วนที่ระบบไม่ได้สกัดออกมา โดยขั้นตอนส่วนใหญ่ไม่ได้ไปปรับแก้ชื่อเฉพาะที่ระบบสกัดออกมาได้ก่อนหน้า ดังนั้นเมื่อนำข้อมูลไปผ่านขั้นตอนประมวลผลภายหลังจึงทำให้จำนวนของชื่อเฉพาะเพิ่มขึ้น แต่นั่นไม่ได้หมายความว่าชื่อเฉพาะที่สกัดได้ในขั้นตอนประมวลผลภายหลังจะถูกต้องทั้งหมด เนื่องจากมีหลายขั้นตอนที่มีการนำเอาชื่อเฉพาะที่ระบบสกัดได้มาใช้ ซึ่งหากชื่อที่สกัดมาได้นั้นผิด เมื่อนำเอาไปเปรียบเทียบกับ token ที่เหลือ ก็จะได้จำนวนชื่อที่ผิดเพิ่มขึ้นด้วย

นอกจากนี้โดยตัวโปรแกรมที่ใช้ในการประมวลผลภายหลังก็มีข้อผิดพลาดเช่นกัน เช่น โปรแกรมที่ใช้สกัดชื่อย่อที่อยู่ในวงเล็บต่อจากชื่อองค์กร ซึ่งโปรแกรมนี้จะดูว่าช่วง token ก่อนหน้าวงเล็บเปิดนั้นมีการกำกับเป็นชื่อองค์กรหรือไม่ หากใช่จะกำกับให้ token ที่อยู่ในวงเล็บเป็นชื่อองค์กรตามไปด้วย ตรงจุดนี้โปรแกรมสามารถสกัดชื่อออกมาผิดได้หากชื่อเฉพาะที่ระบบสกัด

ออกมาได้นั้นระบุขอบเขตหรือชนิดของชื่อเฉพาะผิด เช่น ชื่อองค์กร “พรรคคอมมิวนิสต์แห่งประเทศไทย (ใหม่)” ที่ถูกต้องคือต้องให้วงเล็บเป็นส่วนหนึ่งของชื่อ แต่ระบบรู้จำเพียงแค่ “พรรคคอมมิวนิสต์แห่งประเทศไทย” ดังนั้นเมื่อนำข้อมูลมาผ่านโปรแกรมจึงได้คำว่า “ใหม่” เป็นชื่อย่อองค์กรอีกชื่อหนึ่ง จากนั้นเมื่อนำข้อมูลไปผ่านโปรแกรมสำหรับสกัดชื่อย่อองค์กรในบริบทต่าง ๆ จะได้ว่า token ใดที่เป็นคำว่า “ใหม่” จะถูกกำกับเป็นชื่อองค์กรหมด ดังนั้นปริมาณชื่อที่ผิดจึงเพิ่มขึ้น

ในขั้นตอนประมวลผลภายหลังนี้ ลักษณะการตัดแบ่งข้อมูลก็มีผลเช่นกันโดยในขั้นตอนการรู้จำชื่อบุคคลจะมีการกำหนดว่าไม่ให้รู้จำชื่อที่มีคำเดียวหรือพยางค์เดียว เพราะมีแนวโน้มว่าอาจจะเป็นคำทั่วไปได้ สำหรับข้อมูลตัดคำขั้นตอนนี้ช่วยให้ระบบไม่สกัดเอาคำทั่วไปมาเป็นชื่อได้พอสมควร แต่สำหรับข้อมูลตัดพยางค์ยังคงมีปัญหาค่อนข้างมาก เพราะคำหนึ่งคำสามารถเป็นได้หลายพยางค์ และหากจะเปลี่ยนกฎให้โปรแกรมไม่รู้จำชื่อที่มีพยางค์เดียวและ 2 พยางค์ก็จะทำให้ชื่อจำนวนหนึ่งหายไป เพราะชื่อที่มี 2 พยางค์ก็มีอยู่ค่อนข้างมาก และอีกทั้งเพื่อต้องการเปรียบเทียบประสิทธิภาพของระบบใช้ข้อมูลตัดคำและตัดพยางค์ จึงออกแบบให้ใช้กฎเดียวกัน ผลที่ออกมาคือชื่อบุคคลในข้อมูลตัดพยางค์จะมีคำทั่วไปเพิ่มเข้ามาด้วยค่อนข้างมาก เนื่องจากมีบางชื่อที่เมื่อตัดคำนำหน้าชื่อและนามสกุลออกแล้ว ปรากฏว่าชื่อนั้นไปตรงกับคำทั่วไป เช่น “นายสามารถ โชคคณาพิทักษ์” เมื่อตัดคำนำหน้าชื่อและนามสกุลออกจะเหลือเพียง “สามารถ” สำหรับข้อมูลตัดคำจะไม่มีปัญหา เพราะ “สามารถ” คือคำ ๆ เดียว แต่สำหรับข้อมูลตัดพยางค์ “สามารถ” ประกอบด้วยสองพยางค์ คือ “สา-มารถ” ดังนั้นโปรแกรมจึงดึงคำว่า “สามารถ” ที่ไม่ใช่ชื่อเฉพาะออกมาทั้งหมด และเมื่อดูผลที่ได้จากการประเมินโดยใช้จำนวนชื่อ ลักษณะดังกล่าวจึงส่งผลให้ค่าความแม่นยำของชื่อบุคคลในข้อมูลแบบตัดพยางค์ลดลงจาก 92.55% เหลือเพียง 80.10% ซึ่งลดลงมากกว่า 10% ในขณะที่ในข้อมูลแบบตัดคำค่าความแม่นยำของชื่อบุคคลลดลงเพียง 2% เท่านั้น คือ จาก 92.05% เหลือ 90%

## 4.5 เปรียบเทียบประสิทธิภาพของระบบการรู้จำชื่อเฉพาะระหว่างแบบจำลองที่รับข้อมูลเข้าเป็นพยางค์กับที่รับข้อมูลเข้าเป็นคำ

จากการประมวลผลข้อมูลโดยใช้แบบคำตอบทั้ง 5 แบบ ดังได้ผลการทดสอบตามตารางที่ 4.6-4.10 นั้น จะเห็นได้ว่าประสิทธิภาพของระบบที่ใช้ข้อมูลแบบตัดคำและตัดพยางค์ไม่แตกต่างกันโดยดูจากค่า F-measure ของชื่อเฉพาะทั้งหมดในแต่ละตาราง แม้ว่าค่า F-measure ของบางคำตอบในระบบที่ใช้ข้อมูลแบบตัดคำจะสูงกว่าระบบที่ใช้ข้อมูลแบบตัดพยางค์แต่ก็ไม่ได้มากกว่าอย่างมีนัยสำคัญ

สิ่งที่สำคัญและมีผลต่อประสิทธิภาพของระบบคือคุณสมบัติที่เลือกใช้และการออกแบบคุณสมบัติว่าเหมาะสมกับลักษณะของข้อมูลหรือไม่ ดังจะเห็นได้จากคุณสมบัติ unigram และ bigram จะเหมาะสมกับข้อมูลแบบตัดพยางค์มากกว่า เพราะการตัดพยางค์ทำให้คุณสมบัติ unigram และ bigram สามารถหาค่าความสัมพันธ์ภายในคำได้ หากคำเกิดขึ้นโดยไม่มีบริบทบ่งชี้ เนื่องจากคุณสมบัตินี้จะดูความสัมพันธ์ของหน่วยที่อยู่ติดกันไปทีละสองหน่วย เช่น “พิจิตร” สามารถแยกได้เป็น 2 พยางค์ คือ “พิ-จิตร” หากเกิดโดยไม่มี “จ.” ซึ่งเป็นคำบ่งชี้สถานที่ด้านหน้า คุณสมบัติ unigram และ bigram ก็สามารถหาได้ว่าพยางค์ “พิ” และ “จิตร” เกิดร่วมกันอย่างมีนัยสำคัญ ดังนั้นระบบจึงสามารถรู้จำชื่อ “พิจิตร” ได้โดยไม่ต้องมีคำบ่งชี้ได้ ซึ่งจะต่างจากระบบที่ใช้ข้อมูลตัดคำที่หากชื่อเฉพาะเป็นคำ ๆ เดียวและเกิดโดยไม่มีคำบริบทช่วยบ่งชี้ ก็ยากที่ระบบจะรู้จำได้ว่าคำนั้นเป็นชื่อเฉพาะดังจะเห็นได้จากค่าครบถ้วนของระบบที่ใช้ข้อมูลตัดพยางค์ที่มากกว่าระบบที่ใช้ข้อมูลตัดคำ 3.85% ดังแสดงในตารางที่ 4.16

สำหรับคุณสมบัติอื่น ๆ ที่นำมาใช้ร่วมกับคุณสมบัติ unigram และ bigram หากพิจารณาจากลักษณะการกำหนดค่าของคุณสมบัติที่ใช้ ส่วนใหญ่จะใช้การเปรียบเทียบ token กับรายการคำหรือชื่อที่มีอยู่เป็นหลักซึ่งลักษณะดังกล่าวจะเหมาะกับข้อมูลแบบตัดคำมากกว่า เนื่องจากการตัดพยางค์มีการแยกย่อยมากกว่าทำให้โอกาสที่พยางค์จะไปตรงกับส่วนใดส่วนหนึ่งของคำหรือชื่อในรายการย่อมมีสูง และโอกาสผิดพลาดก็จะมีมากขึ้นตามไปด้วยดังจะเห็นได้จากค่า F-measure ของระบบที่ใช้ข้อมูลตัดพยางค์เพิ่มขึ้นน้อยกว่าระบบที่ใช้ข้อมูลตัดคำเมื่อใช้คุณสมบัติอื่นร่วมกับคุณสมบัติ unigram และ bigram ดังแสดงไว้ในตารางที่ 4.32

ตารางที่ 4.32 ผลต่างของค่า F-measure แบบประเมินโดยใช้จำนวนชื่อ ระหว่างการใช้คุณสมบัติ unigram และ bigram อย่างเดียวกับการใช้คุณสมบัติ unigram และ bigram ร่วมกับคุณสมบัติอื่น

| คุณสมบัติ | F (%) | | ผลต่างของ F (%) | |
|---|---|---|---|---|
| | WSG | SSG | WSG | SSG |
| Unigram และ bigram | 77.93 | 79.72 | | |
| Unigram และ bigram + รายการชื่อเฉพาะ | 80.41 | 80.18 | 2.48 | 0.46 |
| Unigram และ bigram + คำย่อ | 78.52 | 80.14 | 0.59 | 0.42 |
| Unigram และ bigram + คำบริบท | 78.52 | 79.74 | 0.59 | 0.02 |
| Unigram และ bigram + คำทั่วไป | 78.15 | 79.80 | 0.22 | 0.08 |
| Unigram และ bigram + สถิติ | 77.77 | 79.68 | -0.16 | -0.04 |

จากตารางเป็นการหาผลต่างของการประเมินประสิทธิภาพของระบบโดยใช้จำนวนชื่อ เมื่อดูจากค่าผลต่างของ F-measure ของการใช้คุณสมบัติ unigram และ bigram อย่างเดียวกับการใช้คุณสมบัติ unigram และ bigram ร่วมกับคุณสมบัติอื่น จะเห็นว่าจากคุณสมบัติทั้งหมดที่ช่วยให้ประสิทธิภาพของระบบดีขึ้น มีเพียงคุณสมบัติคำย่อที่ผลต่างของทั้งสองระบบไม่ต่างกันมาก เพราะการกำหนดค่าคุณสมบัติไม่ได้ใช้การเปรียบเทียบ token กับรายการคำหรือชื่อ ในขณะที่อีก 3 คุณสมบัติที่เหลือใช้การกำหนดคุณสมบัติในลักษณะดังกล่าว แม้ผลต่างของทั้งสองระบบไม่ได้ต่างกันอย่างชัดเจน แต่ก็แสดงให้เห็นว่าคุณสมบัติทั้ง 3 มีแนวโน้มช่วยให้ระบบที่ใช้ข้อมูลตัดคำดีกว่าตัดพยางค์

ในส่วนของการรู้จำชื่อเฉพาะนั้น เมื่อดูจากค่าความแม่นยำ ค่าความครบถ้วน และค่า F-measure ของชื่อเฉพาะแต่ละชนิดแล้ว จะเห็นว่าประสิทธิภาพของระบบทั้งสองไม่ต่างกันดังแสดงไว้ในตารางที่ 4.28 และเมื่อดูจากข้อมูลจะพบว่าจุดที่ผิดของทั้งสองระบบจะคล้ายกัน โดยสิ่งที่ระบบมักรู้จำผิดคือประเภทของชื่อเฉพาะที่สามารถเป็นได้ทั้งชื่อองค์กรและชื่อสถานที่ แต่สิ่งที่ระบบที่ใช้ข้อมูลตัดพยางค์จะรู้จำได้ดีกว่า คือ ชื่อเฉพาะคำเดียวที่สามารถแยกได้เป็นหลายพยางค์และไม่มีคำบ่งชี้ในบริบทข้างเคียง แต่ทั้งนี้จะมีบางบริบทที่กำกวมไม่มีการเว้นวรรคตอนทำให้ระบบที่ใช้ข้อมูลตัดพยางค์ระบุขอบเขตของชื่อผิด เช่น “3ก.พ.นพ.เสรี ตู้จินดา” ระบบจะรู้จำว่า “พ.นพ.เสรี ตู้จินดา” หรือ “20.20น.พ.ญ.สุรภี เรืองสุวรรณ” รู้จำว่า “น.พ.ญ.สุรภี เรืองสุวรรณ” เนื่องจากในการตัดพยางค์จะตัดแยกส่วนที่เป็นคำย่อด้วยหากมีจุดคั่น แต่จากตัวอย่างปัญหาส่วนหนึ่งมาจากการไม่เว้นวรรคตอนของข้อมูลด้วย ซึ่งหากข้อมูลมีการเว้นวรรคตอนถูกต้องก็มีแนวโน้มว่าระบบจะสามารถระบุขอบเขตของชื่อได้ถูกต้องมากขึ้นเช่นกัน

ในส่วนของขั้นตอนประมวลผลภายหลัง แม้ว่าเมื่อดูจากตารางที่ 4.29 แล้วจะเห็นว่าระบบที่ใช้ข้อมูลตัดพยางค์แย่กว่าระบบที่ใช้ข้อมูลตัดคำโดยปัญหาหลักมาจากการที่ระบบที่ใช้ข้อมูลตัดพยางค์รู้จำคำทั่วไปเป็นชื่อเฉพาะ เนื่องจากโปรแกรมใช้การคัดกรองคำทั่วไปเพียงแค่ดูว่าเป็นคำ ๆ เดียวหรือพยางค์เดียวหรือไม่ ดังนั้นจึงเป็นปัญหากับข้อมูลตัดพยางค์ค่อนข้างมาก เพราะคำ 1 คำสามารถเป็นได้หลายพยางค์ ดังนั้นในการคัดกรองของข้อมูลตัดพยางค์อาจต้องใช้วิธีที่ซับซ้อนกว่าตัดคำ คือ อาจใช้การนำเอารายการชื่อที่ได้มาเทียบกับรายการคำทั่วไปก่อนแล้วค่อยนำรายการชื่อที่ตัดคำทั่วไปออกแล้วมาเทียบกับข้อมูล

ดังนั้นจึงอาจสรุปได้ว่าประสิทธิภาพของแบบจำลองทั้งสองไม่ต่างกันมาก ขึ้นอยู่กับคุณสมบัติที่ใช้ แม้ว่าชื่อเฉพาะในภาษาไทยส่วนใหญ่จะเกิดจากการประสมคำแต่ภายในคำก็สามารถหาความเกาะเกี่ยวระหว่างพยางค์ที่อยู่ภายในคำได้ แต่ทั้งนี้ข้อมูลตัดพยางค์จะมีข้อจำกัดมากกว่าข้อมูลตัดคำ เช่น คุณสมบัติที่ใช้ เมื่อดูจากผลการทดสอบจะเห็นว่ามีเพียงคุณสมบัติ unigram และ bigram เท่านั้นที่ช่วยให้ระบบที่ใช้ข้อมูลตัดพยางค์ดีกว่าตัดคำ ในขณะที่คุณสมบัติ

อื่น ๆ ส่วนใหญ่จะสนับสนุนข้อมูลแบบตัดคำมากกว่า และในการให้ข้อมูลกับระบบในระดับที่สูงขึ้น เช่น การกำกับหมวดคำ (part of speech) หรือดูความสัมพันธ์ด้านความหมายก็จะเหมาะกับการดูในระดับคำมากกว่า เพราะในภาษาไทยคำเป็นหน่วยที่เล็กที่สุดที่มีความหมาย

# บทที่ 5

# ลักษณะทางภาษาศาสตร์ที่มีผลต่อประสิทธิภาพของระบบการรู้จำชื่อเฉพาะ

ในงานวิจัยนี้ นอกจากมีวัตถุประสงค์เพื่อพัฒนาระบบการรู้จำชื่อเฉพาะแล้ว วัตถุประสงค์อีกข้อที่สำคัญคือ วิเคราะห์ลักษณะทางภาษาศาสตร์ที่มีผลต่อประสิทธิภาพของแบบจำลองโดยในหัวข้อนี้ ผู้วิจัยจะนำเสนอตามลำดับ ดังนี้

1. ลักษณะทางภาษาศาสตร์ที่พบในชื่อเฉพาะประเภทต่าง ๆ
2. อภิปรายคุณสมบัติที่ใช้ที่มีความเกี่ยวข้องกับลักษณะทางภาษาศาสตร์
3. ลักษณะทางภาษาศาสตร์ที่มีผลต่อประสิทธิภาพของแบบจำลอง

## 5.1 ลักษณะทางภาษาศาสตร์ที่พบในชื่อเฉพาะประเภทต่าง ๆ

**ชื่อบุคคล**

จากข้อมูลทั้งหมด เมื่อวิเคราะห์โครงสร้างของชื่อบุคคลแล้วพบว่ามีลักษณะ ดังนี้

[คำบ่งบอก] + ชื่อ + [นามสกุล]

“คำบ่งบอก” ที่ใช้ในงานวิจัยนี้หมายถึง คำที่มักปรากฏร่วมกับชื่อบุคคล องค์กรและสถานที่ สามารถช่วยจำแนกชนิดของชื่อเฉพาะได้ เนื่องจากคำบ่งบอกของชื่อเฉพาะแต่ละชนิดใช้ชุดคำที่ต่างกัน นอกจากนี้คำบ่งบอกยังให้ข้อมูลเสริมเกี่ยวกับชื่อเฉพาะนั้น ๆ ด้วย เช่น ชื่อบุคคลที่ใช้คำบ่งบอก “น.ส.” หมายความว่าผู้ใช้ชื่อนี้เป็นผู้หญิงและยังไม่ได้จดทะเบียนสมรส “ตา” ใช้กับผู้ชายที่อายุมาก “ธนาคาร” เป็นคำบ่งบอกปรากฏร่วมกับชื่อองค์กรที่เกี่ยวข้องกับด้านการเงิน หรือ “จ.” เป็นคำบ่งบอกของชื่อสถานที่ในระดับภูมิภาค เป็นต้น

สำหรับชื่อบุคคลคำบ่งบอกจะปรากฏนำหน้าชื่อเสมอ ตามด้วยชื่อ และนามสกุลอยู่หลังชื่อ ส่วนที่จะต้องมีเสมอคือ ชื่อ แต่คำบ่งบอกและนามสกุลเป็นส่วนที่ละได้ โดยในบางบริบทอาจมีแค่ชื่อเท่านั้น หรือชื่ออาจเกิดร่วมกับคำบ่งบอกแล้วละนามสกุลไว้ หรือเกิดร่วมกับนามสกุลแต่ละคำบ่งบอก หรืออาจเกิดร่วมกันทั้ง 3 ส่วน

จากข้อมูลที่ศึกษาคำบ่งบอกที่ปรากฏร่วมกับชื่อบุคคลได้แก่

1) คำนำหน้าชื่อทั่วไป เช่น นาย นาง น.ส. ด.ช. ด.ญ. คุณ ยาย ลุง เป็นต้น
2) ตำแหน่งงานหรือตำแหน่งทางการศึกษา เช่น จ.ส.ต. ดร. ผศ. รองนายกฯ อาจารย์ เป็นต้น

3) ยศ เช่น ม.ร.ว. ม.ล. คุณหญิง เป็นต้น

คำบ่งบอกนอกจากจะใช้บ่งบอกว่าชื่อที่ตามมาคือชื่อบุคคลแล้ว ยังช่วยให้ข้อมูลเพิ่มเติมเกี่ยวกับบุคคลนั้น ๆ ด้วย เช่น เพศ อายุ สถานภาพทางสังคม อาชีพ เป็นต้น ในคลังข้อมูลที่ศึกษามีชื่อบุคคลทั้งหมด 5,672 ชื่อ พบชื่อที่ปรากฏร่วมกับคำบ่งบอกทั้งสิ้น 5,153 ชื่อ คิดเป็น 90.85% นั่นแสดงให้เห็นว่าชื่อบุคคลมักปรากฏร่วมกับคำบ่งบอก

ในส่วนของนามสกุล ชื่อที่ปรากฏร่วมกับนามสกุล มีจำนวนทั้งสิ้น 2,523 ชื่อ คิดเป็น 44.48% และชื่อที่ปรากฏโดยไม่มีนามสกุล มีจำนวนทั้งสิ้น 3,149 ชื่อ คิดเป็น 55.52% สาเหตุที่จำนวนของชื่อที่ปรากฏโดยไม่มีนามสกุลมากกว่านั้น เป็นเพราะข่าวโดยปกติมักมีการกล่าวชื่อเต็มในครั้งแรกเพื่อเป็นการแนะนำบุคคลในข่าว จากนั้นหากมีจะต้องกล่าวถึงบุคคลคนเดียวกันในครั้งต่อไปจะละส่วนของนามสกุลไว้เหลือเพียงชื่อเท่านั้น เช่น

> “ตามที่<persName>**นางประนอมทองจันทร์**</persName>กับ
> <persName><abb>**ด.ช.**</abb>**กิตติพงษ์แหลมผักแว่น**</persName>และ
> <persName><abb>**ด.ญ.**</abb>**กาญจนากรองแก้ว**</persName>ป่วย
> สงสัยติดเชื้อไข้ขณะนี้ยังไม่ดีขึ้น
> หลังเข้าเยี่ยมดูอาการผู้ป่วยแล้ว<persName><abb>น.พ.</abb>จรัล
> </persName>ประชุมร่วมกับเจ้าหน้าที่ทุกฝ่ายเพื่อสรุปผลการดำเนินการรวมทั้ง
> สอบสวนโรคก่อนที่ผู้ป่วยจะถูกส่งมารักษาตัวจากนั้นร่วมกันแถลงข่าวโดย
> <persName><abb>น.พ.</abb>จรัล</persName>กล่าวว่าขณะนี้ผู้ป่วยทั้ง 3
> รายอาการยังทรงโดยในรายของ<persName><abb>**ด.ช.**</abb>**กิตติพงษ์**
> </persName>กับ<persName><abb>**ด.ญ.**</abb>**กาญจนา**</persName>
> ปอดหายเป็นปกติแล้วคาดว่าจะกลับบ้านได้ในไม่ช้านี้แต่ในรายของ
> <persName>**นางประนอม**</persName>อาการยังน่าเป็นห่วง”

นอกจากนี้มีข้อสังเกตของโครงสร้างชื่อบุคคลที่เห็นได้ชัดเจนคือมีการใช้ช่องว่างแยกส่วนของชื่อและนามสกุลออกจากกันทุกครั้ง รวมถึงหากพิจารณาลักษณะการปรากฏของชื่อในข่าวจะสังเกตเห็นว่ามักมีการใช้ช่องว่างระบุขอบเขตของชื่อ กล่าวคือ มีช่องว่างอยู่ด้านหน้าคำบ่งบอกและอยู่ด้านหลังชื่อหรือนามสกุลดังตัวอย่างข้างต้น เพื่อแยกส่วนของเนื้อหาและชื่อเฉพาะ ดังนั้นในส่วนของงานระบบก็อาจนำช่องว่างไปเป็นคุณสมบัติหนึ่งในการช่วยระบุขอบเขตหรือประเภทของชื่อเฉพาะได้

สำหรับชื่อบุคคล ส่วนใหญ่มักมาจากภาษาบาลี สันสกฤต มีการใช้คำสมาสและสนธิ นอกจากนี้การตั้งชื่อบุคคลยังอาจมีการปรับเปลี่ยนพยัญชนะเพื่อสร้างความเป็นเอกลักษณ์เฉพาะตัว หรือมีการนำความเชื่อเข้ามาเกี่ยวข้อง เช่น คนเกิดวันจันทร์ห้ามมีสระภายในชื่อ เพราะ

เป็นตัวกาลกิณี เช่น ณฐนนท ไม่มีไม้หันอากาศและทัณฑฆาต แต่อ่านว่า นัด-ถะ-นน เป็นต้น ดังนั้นชื่อบุคคลส่วนใหญ่จึงมักไม่สามารถตัดแบ่งคำได้ หรือหากแบ่งได้ก็มักประกอบด้วยคำที่ไม่มีความหมายเป็นส่วนหนึ่งของชื่อ เช่น ภัสพร สามารถแยกได้เป็น ภัส-พร “พร” มีความหมายในพจนานุกรม แต่ “ภัส” ไม่ใช่คำในพจนานุกรม หรือแม้ว่าแต่ละคำมีความหมาย แต่ความหมายของชื่อก็อาจไม่ได้เกิดจากการนำความหมายของคำแต่ละคำมารวมกัน เช่น สกุล “ดามาพงศ์” แม้จะแบ่งเป็น ดา-มา-พงศ์ ได้และคำแต่ละคำมีความหมายตามพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ.2542 (ราชบัณฑิตยสถาน, 2546) ดังนี้

ดา คำนามหมายถึงชื่อแมลงชนิดหนึ่ง; คำกริยาหมายถึงเรียงหน้ากันเข้าไป เป็นหน้ากระดาน

มา คำกริยาหมายถึงเคลื่อนออกจากที่เข้าหาตัวผู้พูด

พงศ์ คำนามหมายถึงเชื้อสาย เทือกเถา สกุล

แต่เมื่อนำเอาความหมายของคำทั้ง 3 คำมารวมกันแล้ว พบว่าไม่สื่อความดังนั้นจึงยากที่จะนำชื่อบุคคลมาศึกษาโครงสร้างภายในได้ซึ่งจะต่างจากชื่อเฉพาะชนิดอื่นดังจะกล่าวต่อไป

**ชื่อองค์กร**

เมื่อวิเคราะห์ข้อมูลแล้วพบว่าชื่อองค์กรสามารถปรากฏได้ทั้งในรูปแบบเต็มและรูปแบบย่อ สำหรับรูปแบบเต็มมีโครงสร้างดังนี้

[คำบ่งบอกหน้า] + ชื่อ + [คำบ่งบอกหลัง]

จากโครงสร้าง ชื่อองค์กรสามารถละส่วนที่เป็นคำบ่งบอกหน้าและคำบ่งบอกหลังได้ ส่วนของคำบ่งบอกหน้า เช่น บริษัท, โรงเรียน, สมาคม, บงล. เป็นต้น คำบ่งบอกหลัง เช่น จำกัด, จำกัด (มหาชน) เป็นต้น คำบางคำเป็นได้ทั้งคำบ่งบอกหน้าและบ่งบอกหลัง เช่น มหาวิทยาลัย เป็นต้น ส่วนใหญ่ลักษณะของชื่อองค์กรที่สามารถละคำบ่งบอกได้จะต้องเป็นชื่อที่ไม่มีความหมายอื่น หรือพ้องกับคำหรือวลีทั่วไป เช่น ธนาคารกรุงไทย สามารถละ “ธนาคาร” ซึ่งเป็นคำบ่งบอกหน้าได้ แต่ถ้าเป็น ธนาคารกรุงเทพ จะไม่สามารถละคำบ่งบอกหน้าได้ เพราะ “กรุงเทพ” พ้องกับชื่อสถานที่

จากข้อมูลชื่อองค์กรรวมถึงชื่อองค์กรที่ใช้อ้างถึงสถานที่จำนวน 5,168 ชื่อ มีชื่อที่ปรากฏร่วมกับคำบ่งบอกทั้งสิ้น 3,062 ชื่อ คิดเป็น 59.25% สำหรับคำบ่งบอกหน้าชื่อองค์กรเป็นสิ่งที่ให้ข้อมูลเพิ่มเติมเกี่ยวกับองค์กรนั้น ๆ เช่น ลักษณะหรือประเภทของธุรกิจ เป็นองค์กรของรัฐหรือเอกชน เช่น “กระทรวง” “กรม” แสดงว่าเป็นหน่วยงานของรัฐ หรือ “บงล.” แสดงว่าองค์กรนี้ประกอบธุรกิจเกี่ยวกับด้านการเงินและหลักทรัพย์ เป็นต้น สำหรับคำบ่งบอกหลังมักเป็นคำที่ต้องใช้คู่กับคำบ่งบอกหน้า เช่น จำกัด, จำกัด (มหาชน) ดังนั้นโดยทั่วไปคำบ่งบอกหลังจึงมักปรากฏ

ร่วมกับคำบ่งบอกหน้าเสมอ เช่น บริษัท ทศภาค จำกัด, ธ. กรุงเทพ จำกัด (มหาชน) และคำบ่งบอกหลังจะเป็นส่วนที่มีการละมากกว่าคำบ่งบอกหน้า

สำหรับรูปแบบย่อขององค์กร จากข้อมูลสามารถเป็นได้ 3 รูปแบบ ดังนี้

1) รูปอักษรย่อ เช่น รฟท. คือ อักษรย่อของ การรถไฟแห่งประเทศไทย
2) Acronym คือ การนำนำอักษรต้นตัวแรกหรือสองสามตัวแรกในคำมารวมกันเป็นคำเดียว แล้วอ่านออกเสียงเป็นคำใหม่ เช่น ฮู มาจาก WHO ซึ่งเป็น acronym ของชื่อองค์การอนามัยโลก หรือ The World Health Organization เป็นต้น
3) การใช้เครื่องหมายไปยาลน้อย (ฯ) ละส่วนของชื่อที่เป็นที่รู้กันโดยทั่วไป หรือละส่วนของชื่อที่มีการกล่าวถึงไปก่อนหน้า เช่น กระทรวงเกษตรฯ (กระทรวงเกษตรและสหกรณ์) กรมชลฯ (กรมชลประทาน) กระทรวงการพัฒนาสังคมฯ (กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์) เป็นต้น

สำหรับรูปแบบของชื่อองค์กรที่ปรากฏในข่าวนั้น ในครั้งแรกชื่อองค์กรมักปรากฏในรูปแบบเต็ม แต่หากมีการอ้างถึงในครั้งต่อไป มักใช้รูปแบบย่อหรือลดรูป คือ ละคำบ่งบอกหน้าหรือหลัง เช่น หากชื่อองค์กรนั้นมีชื่อย่อ ชื่อย่อมักปรากฏอยู่ในวงเล็บต่อจากชื่อองค์กร และในการอ้างถึงชื่อองค์กรเดียวกันนี้ในครั้งต่อไปจะใช้ชื่อย่อเป็นหลัก เช่น

> “...นโยบายดังกล่าวแรกๆทำท่าว่าจะไปได้ดีเพราะหน่วยงานที่เกี่ยวข้องอย่าง <orgName>กระทรวงเกษตรฯ</orgName>และ<orgName>**กระทรวงสาธารณสุข**</orgName> (<orgName><abb>**สธ**.</abb></orgName>) ได้เร่งดำเนินการให้เป็นรูปธรรมแต่เมื่อเกิดวิกฤตการณ์ไข้หวัดนกนโยบายนี้ถึงกับออกอาการเซ
> อย่างไรก็ตามรัฐบาลก็ยืนยันที่จะดำเนินนโยบายนี้ต่อโดยที่ประชุมคณะรัฐมนตรีเมื่อวันที่ 22 กุมภาพันธ์ที่ผ่านมาได้อนุมัติงบประมาณ 271 ล้านบาทเพื่อใช้ในโครงการอาหารปลอดภัย
> การดำเนินงานในโครงการดังกล่าวรัฐบาลได้มอบให้<orgName><abb>**สธ**.</abb></orgName>เป็นตัวหลักในการประสานแผนยุทธศาสตร์โดยบูรณาการความร่วมมือระหว่างหน่วยงานภาครัฐและเอกชนมีการเสนอแผนงานหรือมาตรการเพื่อพิจารณาตามห่วงโซ่อาหารให้เกิดความชัดเจนโดยหน่วยงานจาก <orgName><abb>**สธ**.</abb></orgName>ที่รับผิดชอบเรื่องนี้มี 3 หน่วยงาน คือ<orgName>สำนักงานอาหารและยา</orgName> (<orgName><abb>อย.</abb></orgName>) <orgName>กรมอนามัย</orgName>และ<orgName>กรมวิทยาศาสตร์การแพทย์</orgName>...”

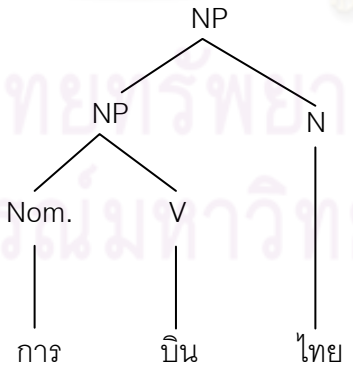
ตัวอย่างการละคำบ่งบอก เช่น

“<persName>นายกนกอภิรดี</persName>กรรมการผู้อำนวยการใหญ่ <orgName>**บริษัทการบินไทยจำกัด(มหาชน)**</orgName>เปิดเผยถึงมาตรการรักษาความปลอดภัยภายในเครื่องบินเพื่อป้องกันเชื้อไข้หวัดนกที่ผู้โดยสารทั้งในประเทศและต่างประเทศวิตกกังวลว่า<orgName>**การบินไทย**</orgName>ได้ฆ่าเชื้อจำนวน 32 จุดทั่วบริเวณของเครื่องบินเพื่อเป็นการป้องกันว่าไม่มีปัญหาติดเชื้อไข้หวัดนกบนเครื่องบินโดยสารทั้งภายในและต่างประเทศแน่นอนเพื่อสร้างความมั่นใจให้แก่ผู้โดยสารขณะเดียวกันก็ได้งดเมนูอาหารที่ทำจากไก่ทั้งหมดบนเครื่องด้วยจนกว่ารัฐบาลจะสามารถควบคุมการแพร่ระบาดของเชื้อไข้หวัดนกได้แล้ว<orgName>**การบินไทย**</orgName>จึงจะนำเมนูอาหารไก่เข้ามาเสิร์ฟตามปกติ”

ส่วนของโครงสร้างภายในชื่อ เมื่อวิเคราะห์แล้วได้ 2 โครงสร้างหลัก ดังนี้

1) ชื่อองค์กรที่เป็นชื่อเฉพาะหรือคำ ๆ เดียว เช่น กระทรวงกลาโหม, กระทรวงสาธารณสุข, บริษัท เคียวคุโย จำกัด เป็นต้น
2) ชื่อองค์กรที่มีโครงสร้างเป็นวลีมีทั้งแบบซับซ้อนและไม่ซับซ้อนเช่น บริษัท การบินไทย จำกัด (มหาชน), มูลนิธิคุ้มครองสัตว์ป่าและพรรณพืชแห่งประเทศไทย เป็นต้น

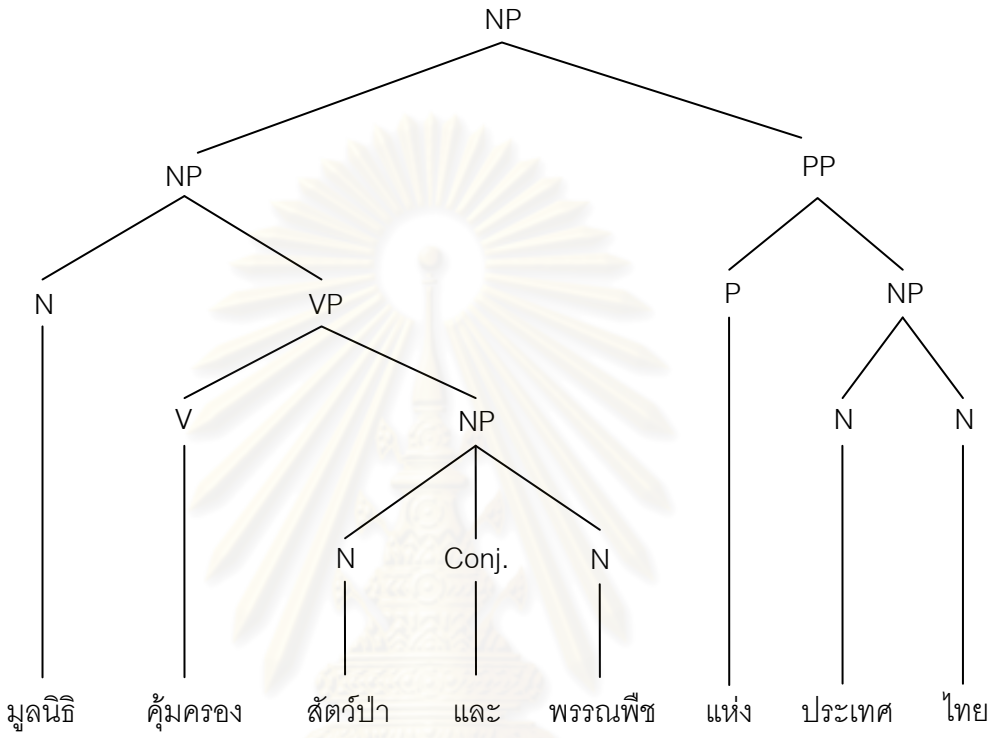
สำหรับโครงสร้างของชื่อองค์กรแบบแรกคือชื่อที่ไม่สามารถแยกองค์ประกอบภายในชื่อได้อีก ในขณะที่ชื่อแบบที่สองสามารถนำมาแยกองค์ประกอบภายในได้ องค์ประกอบในที่นี้หมายถึงหมวดคำ หรือวลี ที่นำมาประกอบรวมเข้าด้วยกันเป็นชื่อจากตัวอย่างข้างต้น บริษัท การบินไทย จำกัด (มหาชน) ชื่อ “การบินไทย” สามารถแยกองค์ประกอบได้ดังนี้




จากตัวอย่างแผนภูมิต้นไม้ เมื่อกำกับหมวดคำให้คำแต่ละคำจะได้ดังนี้ “การ” เป็นคำอุปสรรค (prefix) ทำหน้าที่เปลี่ยนหน่วยคำที่ตามมาให้เป็นคำนาม (Nom.) “บิน” เป็นคำกริยา (V)

และ “ไทย” เป็นคำนาม (N) เมื่อนำคำอุปสรรค “การ” รวมเข้ากับคำกริยา “บิน” ได้เป็นนามวลี (NP) เมื่อเพิ่มคำนาม “ไทย” เข้าไปจะได้นามวลีที่ใหญ่ขึ้นเป็น “การบินไทย”

อีกตัวอย่างคือ มูลนิธิคุ้มครองสัตว์ป่าและพรรณพืชแห่งประเทศไทย สามารถแยกองค์ประกอบได้ดังนี้




จากตัวอย่างจะเห็นว่า “มูลนิธิคุ้มครองสัตว์ป่าและพรรณพืชแห่งประเทศไทย” มีความซับซ้อนมากกว่า “การบินไทย” ดังจะเห็นได้จากแผนภูมิต้นไม้มีหลายชั้น เพราะประกอบด้วยวลีหลายวลีรวมเข้าด้วยกัน ได้แก่ นามวลี (NP) กริยาวลี (VP) และบุพบทวลี (PP) รวมถึงมีการใช้คำเชื่อม (Conj.) “และ” อยู่ภายในชื่อด้วย

ข้อสังเกตหนึ่งของชื่อองค์กร คือ จะมีเฉพาะคำไวยากรณ์กลุ่มหนึ่งที่ปรากฏเป็นส่วนหนึ่งของชื่อ เช่น “เพื่อ” “และ” “แห่ง” “แก่” และหน่วยคำแปลงเป็นคำนาม ได้แก่ “การ” “ความ” เป็นต้น แต่จะไม่พบคำว่า “หรือ” และ “เกี่ยวกับ” เป็นส่วนหนึ่งของชื่อ นั่นเพราะคำว่า “หรือ” จะใช้ในกรณีให้เลือกอย่างใดอย่างหนึ่ง ส่วน “เกี่ยวกับ” จะเป็นคำที่ใช้ในกรณีที่ต้องการอธิบายความ ดังนั้นทั้งสองคำนี้จึงไม่เหมาะแก่การนำไปตั้งชื่อ

นอกจากนี้ ภายในชื่อองค์กรยังสามารถประกอบด้วยช่องว่างและตัวเลข โดยช่องว่างมักใช้กับชื่อที่มีการถ่ายถอดเสียงมาจากภาษาต่างประเทศ เช่น สายการบินยูไนเต็ด แอร์ไลนส์, บริษัท สุมิโตโม มิตซุย คอนสตรัคชั่น จำกัด เป็นต้น หรือใช้กับชื่อที่ประกอบด้วยวงเล็บหรือตัวเลข

ภายในชื่อ เช่น ศาลอุทธรณ์ภาค 1, สำนักผู้ตรวจราชการประจำเขตตรวจราชการที่ 9, องค์การบริหารส่วนตำบล (อบต.) ทุ่งคลอง เป็นต้น

**ชื่อสถานที่**

จากในข้อมูล รูปแบบของชื่อสถานที่สามารถเป็นได้ทั้งรูปแบบเต็มและรูปแบบย่อ สำหรับรูปแบบเต็มสามารถสรุปได้ดังนี้

[คำบ่งบอก] + ชื่อ

ส่วนของคำบ่งบอกชื่อสถานที่จะปรากฏอยู่ด้านหน้าชื่อ และเป็นส่วนที่ละได้ คำบ่งบอกชื่อสถานที่ช่วยให้ข้อมูลด้านภูมิประเทศ ภูมิศาสตร์ และสิ่งปลูกสร้างของสถานที่นั้น ๆ เช่น ประเทศ อำเภอ แม่น้ำ ห้วย สะพาน อนุสาวรีย์ อาคาร เป็นต้น จากจำนวนชื่อสถานที่รวมถึงชื่อสถานที่ที่ใช้อ้างถึงองค์กรทั้งหมด 5,339 ชื่อ มีชื่อที่ปรากฏร่วมกับคำบ่งบอก 3,103 ชื่อ คิดเป็น 58.12%

รูปแบบย่อของชื่อสถานที่พบทั้งหมด 2 รูปแบบด้วยกัน คือ

1) รูปอักษรย่อ จากข้อมูลทั้งหมดพบชื่อสถานที่ที่เขียนเป็นอักษรย่อเพียงชื่อเดียว คือ กทม. (กรุงเทพมหานคร)
2) การใช้เครื่องหมายไปยาลน้อย (ฯ) เช่น กรุงเทพฯ (กรุงเทพมหานคร) สหรัฐฯ (สหรัฐอเมริกา) ซอยจรัญฯ (ซอยจรัญสนิทวงศ์) ถนนรัชดาฯ (ถนนรัชดาภิเษก) เป็นต้น

ลักษณะการปรากฏซ้ำของชื่อสถานที่ในข่าวส่วนใหญ่มักคงรูปเดิม หรือหากลดรูปก็จะละส่วนของคำบ่งบอกไป เนื่องจากชื่อสถานที่มักเป็นชื่อสั้น ๆ อยู่แล้ว เช่น

> “หลังรัฐบาลพยายามปกปิดข้อมูลเรื่องไข้หวัดนกในไก่และสร้างความมั่นใจให้ประชาชนด้วยการกินไก่โชว์แต่ในที่สุดก็ต้องออกมายอมรับว่า<placeName ref=”org”>ไทย</placeName>กำลังถูกโรคไข้หวัดนกคุกคามภายหลังที่มีผู้ป่วยโรคไข้หวัดนกเป็นเด็กที่<placeName><abb>จ.</abb>สุพรรณบุรี</placeName>และ<placeName><abb>**จ**.</abb>**กาญจนบุรี**</placeName>เข้ารับการรักษาใน<abb>รพ.</abb>และผู้ป่วยวัย 6 ขวบจาก<placeName><abb>**จ**.</abb>**กาญจนบุรี**</placeName>ได้เสียชีวิตจากไข้หวัดนกเป็นรายแรกของประเทศ
> โดยผู้เสียชีวิตจากโรคไข้หวัดนกรายแรกของ<placeName ref=”org”>ประเทศไทย</placeName>คือ<persName><abb>ด.ช.</abb>กัปตันบุญมานุช</persName>วัย 6 ขวบที่ถูกส่งตัวจาก<placeName><abb>**จ**.</abb>

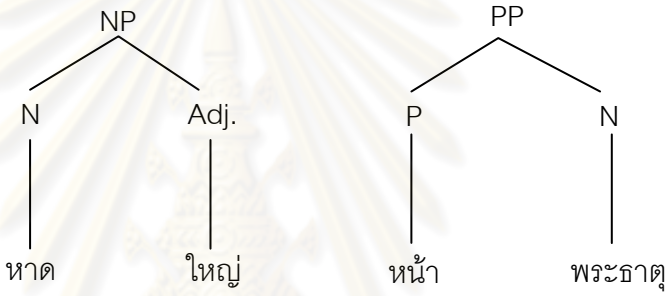
**กาญจนบุรี**</placeName>เข้ามารักษาตัวที่<orgName ref="loc"><abb>รพ.
</abb>ศิริราช</orgName>…"

เมื่อวิเคราะห์ลักษณะโครงสร้างภายในของชื่อสถานที่ พบว่ามี 2 ลักษณะ ดังนี้

1) ชื่อเฉพาะเป็นคำ ๆ เดียวไม่สามารถแยกโครงสร้างภายในได้ เช่น ฮ่องกง ยะลา (แม่น้ำ)ยม ถ.พระจันทร์ เป็นต้น

2) ชื่อเฉพาะมีลักษณะเป็นคำประสมหรือวลี เช่น อ.หาดใหญ่ ถ.หน้าพระธาตุ เป็นต้น

เนื่องจากชื่อสถานที่ส่วนใหญ่มักเป็นชื่อสั้น ๆ ดังนั้นลักษณะโครงสร้างภายในของชื่อจึงไม่ซับซ้อนเท่าชื่อองค์กร หากเป็นวลีก็จะไม่ซ้อนกันหลายชั้น เช่น จากตัวอย่างข้างต้น อ.หาดใหญ่ และ ถ.หน้าพระธาตุ สามารถแยกองค์ประกอบได้ดังนี้




นอกจากนี้ข้อสังเกตหนึ่งคือชื่อสถานที่มีการใช้เครื่องหมายยัติภังค์ (-) และตัวเลขมากกว่าชื่อเฉพาะชนิดอื่น โดยเฉพาะชื่อซอยและชื่อถนน เช่น ถนนสาย 3259, ซอยสุขุมวิท 19, ถนนสายท่าศาลา-นบพิตำ เป็นต้น

**ชื่อเฉพาะอ้างข้ามประเภท**

ชื่อเฉพาะที่สามารถอ้างข้ามประเภทได้มีเพียงชื่อองค์กรและชื่อสถานที่เท่านั้นเหตุที่มีการใช้ชื่อองค์กรอ้างข้ามเป็นชื่อสถานที่ เพราะโดยปกติการจัดตั้งองค์กรจำเป็นต้องมีสถานที่ตั้งขององค์กรเพื่อใช้ในการติดต่อและดำเนินงานต่าง ๆ ดังนั้นชื่อองค์กรจึงสามารถนำไปใช้เพื่อหมายถึงสถานที่ตั้งขององค์กรนั้นได้ ส่วนชื่อสถานที่ใช้อ้างข้ามประเภทเป็นชื่อองค์กรนั้น ใช้เพื่อแทนกลุ่มคน หน่วยงาน หรือองค์กรต่าง ๆ ที่อยู่ในสถานที่นั้น โดยชื่อสถานที่นี้จะสามารถกระทำการต่าง ๆ ได้เสมือนเป็นองค์กร ๆ หนึ่ง

สิ่งที่น่าสังเกต คือ ส่วนใหญ่ชื่อองค์กรจะใช้อ้างข้ามประเภทเป็นชื่อสถานที่ได้ เพราะองค์กรจะต้องมีสถานที่ตั้งเป็นของตนเอง ซึ่งจะต่างจากชื่อสถานที่ที่ใช้อ้างเป็นชื่อองค์กรที่มีข้อจำกัดมากกว่า เพราะชื่อสถานที่ที่สามารถนำมาใช้อ้างข้ามประเภทได้นั้น ต้องเป็นสถานที่ที่มี

หน่วยงานหรือกลุ่มคนอาศัยอยู่ในสถานที่นั้น ๆ ด้วย เพื่อที่ว่าเมื่อนำชื่อมาใช้อ้างข้ามประเภทแล้วทำให้สามารถเข้าใจได้ว่าหมายถึงหน่วยงานหรือกลุ่มคนใด เช่น ชื่อประเทศใช้แทนหน่วยงานของรัฐของประเทศนั้น ชื่อตำบล อำเภอ ใช้แทนหน่วยงานที่บริหารส่วนตำบลหรืออำเภอ เป็นต้น แต่ถ้าเป็นชื่อสถานที่ที่ใช้แทนสภาพทางภูมิประเทศ ภูมิศาสตร์ สิ่งปลูกสร้าง แต่ไม่พบว่ามีหน่วยงานใดตั้งอยู่ในสถานที่นั้น ๆ จะไม่สามารถนำมาใช้เป็นชื่อเฉพาะอ้างข้ามประเภทได้ เช่น แม่น้ำ ถนนอนุสาวรีย์ สะพาน เป็นต้น

สิ่งสำคัญที่ช่วยบ่งบอกว่าชื่อเฉพาะใดเป็นชื่ออ้างข้ามประเภทคือบริบทข้างเคียงของชื่อนั้น ๆ เช่น ชื่อองค์กรมักกลายเป็นชื่อสถานที่เมื่อตามหลังคำบุพบท ตัวอย่างเช่น

> “...ต่อมาในเวลา 14.00 <abb>น.</abb>*ที่*<orgName ref=“loc”>**กระทรวงสาธารณสุข**</orgName><persName>นางสุดารัตน์</persName>เปิดแถลงข่าวยืนยันการตรวจพบผู้ป่วย...”

> “…ได้รับตัวเด็กแฝดผู้พี่ของ<persName><abb>ด.ช.</abb>วิรัตน์</persName>ที่มีอาการปอดบวมและมีไข้ไว้รักษาและให้การดูแลใกล้ชิด*ใน*<orgName ref=“loc”><abb>**รพ.**</abb>**เจ้าพระยายมราช**</orgName>…”

อีกกรณีคือชื่อองค์กรตามหลังคำกริยาที่มีความหมายเกี่ยวข้องกับสถานที่ โดยที่คำนามที่ตามหลังคำกริยานั้นมักเป็นชื่อสถานที่ เช่น ไป มา เป็นต้น ตัวอย่างเช่น

> “...ไก่ของ<persName>นายบุญสม</persName>ตาย<persName>นายบุญสม</persName>ได้เชือดคอแล้วนำไปฝังกลางที่นาพอกลับบ้านไม่ถึงครึ่งชั่วโมงมีอาการหอบเหนื่อยไข้ขึ้นสูงตัวร้อนเหงื่อออกปวดท้องจึงพา*ไป*<orgName ref=“loc”><abb>**รพ.**</abb>**หนองจอก**</orgName>…”

> “...ดังนั้นเมื่อวันที่ 22 <abb>ม.ค.</abb>จึงพา*มา*<orgName ref=“loc”><abb>**รพ.**</abb>**แม่และเด็ก**</orgName>...”

สำหรับชื่อสถานที่ที่กลายเป็นชื่อองค์กรนั้นมักมีคำกริยาตามหลัง เนื่องจากมีเพียงบุคคลและองค์กรเท่านั้นที่สามารถกระทำการหรือดำเนินงานต่าง ๆ ได้ ดังนั้นหากสถานที่สามารถกระทำการได้เช่นเดียวกับบุคคลหรือองค์กร นั่นหมายความว่าชื่อสถานที่นั้นได้นำมาใช้อ้างข้ามประเภทเป็นชื่อองค์กรแล้ว เช่น

“<orgName>องค์การอนามัยโลก</orgName>ยังยืนยันเหมือนเดิมว่า
<placeName ref=“org”>**ไทย**</placeName>*ควรจะเร่งส่ง*เชื้อไข้หวัดนกที่ตรวจ
พบไปตรวจที่แล็บต่างประเทศ”

“...ขณะที่<placeName ref=“org”><abb>**จ.**</abb>**พิจิตร**</placeName>
*เดินหน้าตรวจสอบ*โกดังข้าวของ<orgName><abb>อคส.
</abb></orgName>...”

นอกจากนี้ชื่อสถานที่จะกลายเป็นชื่อองค์กรได้ หากชื่อสถานที่นั้นปรากฏร่วมกับคำหรือวลีที่มีความหมายเชื่อมโยงกับองค์กร เช่น

“<persName>นายนาม</persName>ยื่นคำให้การต่อ<orgName>ศาลแพ่ง
</orgName>เพื่อแก้ต่างในคดีที่ถูก<persName>นายสมัครสุนทรเวช
</persName>อดีต*ผู้ว่าราชการ*<placeName ref=“org”>**กรุงเทพมหานคร**
</placeName>ฟ้องเรียกค่าเสียหายในคดีการจัดซื้อรถ-เรือดับเพลิง*ของ*
<placeName ref=“org”><abb>**กทม.**</abb></placeName>...”

จากตัวอย่างชื่อสถานที่ “กรุงเทพมหานคร” อยู่ตามหลังตำแหน่งงาน “ผู้ว่าราชการ” ดังนั้นชื่อสถานที่นี้จึงเป็นชื่อองค์กรเพราะมีลักษณะเป็นสังกัดของตำแหน่งงาน และอีกตัวอย่างคือ “กทม.” ปรากฏร่วมกับคำบุพบท “ของ” ซึ่ง “ของ” มักปรากฏอยู่หน้าคำนามหรือสรรพนามซึ่งทำหน้าที่เป็นผู้เป็นเจ้าของ ดังนั้นจึงมีเพียงบุคคลและองค์กรเท่านั้นที่แสดงความเป็นเจ้าของได้ “กทม.” ในที่นี้จึงเป็นชื่อสถานที่อ้างข้ามประเภทเป็นชื่อองค์กร

จากข้อมูลชื่อองค์กรที่ใช้อ้างถึงสถานที่มีจำนวน 417 ชื่อ และชื่อสถานที่ที่ใช้อ้างถึงองค์กรมีจำนวน 1,405 ชื่อ เมื่อเปรียบเทียบจำนวนแล้วจะเห็นว่าต่างกันค่อนข้างมาก นั่นเพราะปกติชื่อสถานที่มักใช้เป็นส่วนเติมเต็มของประโยค ทำให้ทราบว่าเหตุการณ์ในข่าวเกิดขึ้นหรือเกี่ยวข้องกับสถานที่ใด ดังนั้นจึงไม่จำเป็นต้องมีการกล่าวถึงซ้ำอีกหากเหตุการณ์ต่าง ๆ ยังคงเกิดขึ้นในสถานที่เดิม ซึ่งจะต่างจากองค์กรที่สามารถเป็นผู้กระทำการต่าง ๆ ได้ ชื่อองค์กรจึงมักอยู่ในตำแหน่งประธานของประโยค ดังนั้นจึงเป็นไปได้ที่จะมีการกล่าวซ้ำชื่อองค์กรเดิมอีกหากต้องมีการกล่าวถึงการกระทำต่าง ๆ ขององค์กรนั้น ๆ จากเหตุผลดังกล่าวทำให้จำนวนชื่อองค์กรมีมากกว่าชื่อสถานที่และเป็นเหตุให้ชื่อสถานที่อ้างถึงองค์กรมีมากกว่าชื่อองค์กรอ้างถึงสถานที่

## 5.2 อภิปรายคุณสมบัติที่ใช้ที่มีความเกี่ยวข้องกับลักษณะทางภาษาศาสตร์

คุณสมบัติสำหรับระบบการรู้จำชื่อเฉพาะที่ผู้วิจัยใช้มีทั้งคุณสมบัติที่เป็นเชิงสถิติ และเชิงภาษาศาสตร์ โดยคุณสมบัติที่มีพื้นฐานมาจากความรู้เชิงภาษาศาสตร์ คือ คุณสมบัติคำบริบทและคุณสมบัติคำทั่วไป

คุณสมบัติคำบริบทมาจากพื้นฐานความคิดที่ว่าคำบริบทมีส่วนสำคัญในการตัดสินชนิดของชื่อเฉพาะโดยเฉพาะชื่อองค์กรและชื่อสถานที่ จากชื่อเฉพาะทั้ง 3 ชนิด ชื่อบุคคลมีปัญหาในการระบุประเภทน้อยที่สุดเพราะไม่ต้องนำไปใช้อ้างถึงชื่อเฉพาะชนิดอื่น ซึ่งจะต่างจากชื่อองค์กรและชื่อสถานที่ที่สามารถใช้อ้างข้ามประเภทกันได้ ในการอ้างข้ามประเภทนั้นเราไม่สามารถตัดสินชนิดของชื่อเฉพาะได้จากรูปภาษาโดยตรงแต่ต้องใช้บริบทในการกำหนดชนิดของชื่อเฉพาะ ดังนั้นในงานวิจัยนี้จึงได้นำเอาบริบทมาเป็นคุณสมบัติหนึ่งของระบบเพื่อใช้ช่วยในการตัดสิน

สำหรับคุณสมบัติคำทั่วไป ผู้วิจัยกำหนดขึ้นจากลักษณะการปรากฏของชื่อโดยเฉพาะชื่อบุคคลที่จะต่างจากคำทั่วไป เนื่องจากชื่อบุคคลมักเป็นภาษาบาลี สันสกฤต และเกิดจากการนำคำมาสมาสหรือสนธิกัน ทำให้บางครั้งได้คำใหม่ที่ไม่พบในพจนานุกรม เช่น หัสดาภรณ์ เป็นชื่อที่เกิดจากการสนธิ มาจาก หัสดี+อาภรณ์ เป็นต้น และชื่อบุคคลอาจมีการปรับเปลี่ยนพยัญชนะภายในชื่อเพื่อสร้างเอกลักษณ์เฉพาะตัวหรือตามความเชื่อได้ ดังนั้นชื่อบุคคลจึงมักเป็นคำที่ไม่ใช่คำที่ใช้กันทั่วไปในชีวิตประจำวัน นอกจากนี้ชื่อที่ถ่ายถอดเสียงมาจากภาษาต่างประเทศก็เป็นคำที่ไม่พบโดยทั่วไปเช่นกันเช่น บริษัท อีซูซุ, บริษัท เอ พี ไอ เนท จำกัด เป็นต้น อย่างไรก็ตามชื่อเหล่านี้จะต้องผ่านโปรแกรมตัดคำและพยางค์ก่อนนำไปใช้ประมวลผลจริง จึงเป็นไปได้ว่าอาจมีบางส่วนของชื่อที่ตรงกับคำทั่วไป ดังนั้นชื่อที่ผ่านการตัดคำและพยางค์แล้วจึงจะมีทั้งชื่อที่คำหรือพยางค์นั้นพบในคำทั่วไปทั้งหมด ไม่พบในคำทั่วไปเลย หรือมีทั้งส่วนที่พบและไม่พบในคำทั่วไปอยู่ร่วมกัน เช่น วัลยาภรณ์ ปกติเป็นคำที่ไม่พบโดยทั่วไป แต่เมื่อตัดคำแล้วได้เป็น 3 token คือ วัล|ยา|ภรณ์ “วัล” และ “ภรณ์” เป็น token ที่ไม่พบในคำทั่วไปในขณะที่ “ยา” ตรงกับคำทั่วไป เป็นต้น

จากผลการทดสอบระบบพบว่า คำบริบทไม่ได้ช่วยให้ประสิทธิภาพของแบบจำลองดีขึ้นมากเท่าใดนัก ปัญหาที่พบคือคำบริบทที่ดึงออกมาไม่ใช่ว่าคำเหล่านั้นทั้งหมดจะสามารถช่วยบ่งบอกชนิดของชื่อเฉพาะได้ จะมีแค่เพียงส่วนหนึ่งเท่านั้น เช่น ที่ ของ เป็นต้น แต่แม้คำที่ช่วยบ่งบอกเหล่านี้เมื่อนำไปเทียบกับจำนวนครั้งในการปรากฏร่วมกับคำอื่น ๆ ที่ไม่ใช่ชื่อเฉพาะจะพบว่าสัดส่วนที่เกิดร่วมกับชื่อเฉพาะจริง ๆ น้อยมาก และแม้ว่าจะปรับลดช่วงคำบริบทให้เหลือเพียง 2 หรือ 1 คำซึ่งเทียบเท่ากับ 3 และ 2 พยางค์ เพื่อแก้ปัญหาที่ว่าใช้ช่วงคำบริบทกว้างเกินไป คือ 3 คำก่อนหน้าและหลังชื่อเฉพาะหรือเทียบเท่ากับ 4 พยางค์ทำให้มีคำอื่นที่ไม่ใช่คำบ่งบอกมาปนอยู่ด้วยค่อนข้างมาก แต่ผลที่ออกมาทั้งของระบบที่ใช้ข้อมูลตัดคำและตัดพยางค์ก็ไม่ต่างกับ

การใช้ช่วง 3 คำ ดังจะเห็นได้จากประสิทธิภาพของระบบเมื่อใช้คุณสมบัติคำบริบทที่ยังไม่ได้ปรับลดกับปรับลดแล้วร่วมกับคุณสมบัติ unigram และ bigram ในตารางที่ 5.1-5.3 (รายละเอียดการประมวลผลทั้ง 10 ครั้งดูได้จากภาคผนวก ช)

ตารางที่ 5.1 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติคำบริบท 3 คำและ 4 พยางค์

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 89.29 | 91.67 | 82.76 | 84.79 | 85.86 | 88.04 |
| ชื่อองค์กร | 84.63 | 82.11 | 67.61 | 70.45 | 75.12 | 75.79 |
| ชื่อสถานที่ | 80.91 | 79.38 | 65.85 | 68.01 | 72.55 | 73.15 |
| ทั้งหมด | 85.62 | 85.20 | 72.54 | 74.98 | 78.52 | 79.74 |

ตารางที่ 5.2 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติคำบริบท 2 คำและ 3 พยางค์

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 89.57 | 91.73 | 82.92 | 84.66 | 86.06 | 87.99 |
| ชื่อองค์กร | 84.55 | 82.88 | 67.49 | 70.23 | 75.02 | 75.98 |
| ชื่อสถานที่ | 80.97 | 79.34 | 65.80 | 68.15 | 72.55 | 73.22 |
| ทั้งหมด | 85.71 | 85.45 | 72.52 | 74.86 | 78.55 | 79.78 |

ตารางที่ 5.3 ประสิทธิภาพของแบบจำลองเมื่อใช้คุณสมบัติคำบริบท 1 คำและ 2 พยางค์

| | P (%) | | R (%) | | F (%) | |
|---|---|---|---|---|---|---|
| | WSG | SSG | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 89.87 | 91.65 | 81.69 | 84.42 | 85.53 | 87.82 |
| ชื่อองค์กร | 85.23 | 83.43 | 66.98 | 69.48 | 74.95 | 75.76 |
| ชื่อสถานที่ | 81.53 | 79.30 | 65.58 | 68.13 | 72.65 | 73.19 |
| ทั้งหมด | 86.14 | 85.63 | 71.82 | 74.48 | 78.32 | 79.64 |

หากพิจารณาจากโครงสร้างของภาษาแล้วจะพบว่าคำบางคำมีหลายหน้าที่ แม้รูปจะเหมือนกันแต่มีหน้าที่ต่างกันเมื่อไปปรากฏในบริบทที่ต่างกัน ตัวอย่างเช่น คำว่า “ที่” ทุกคนมัก

เข้าใจว่า “ที่” เป็นคำช่วยบ่งบอกชื่อสถานที่หรือกำหนดประเภทของชื่อเฉพาะได้ เพราะชื่อสถานที่มักปรากฏตามหลัง “ที่” แต่เมื่อพิจารณาหน้าที่ของคำว่า “ที่” แล้วจะพบว่าสามารถเป็นได้หลายหน้าที่ โดย “ที่” ที่เกิดกับชื่อสถานที่นั้นทำหน้าที่เป็นคำบุพบท เช่น “ไปพบกัน**ที่**เชียงใหม่” “เชียงใหม่” เป็นชื่อสถานที่บอกให้รู้ว่าไปพบกันที่ไหน หรือ “...เนื่องจากได้แท้งลูก**ที่**โรงพยาบาลลาดกระบัง” ปกติ “โรงพยาบาลลาดกระบัง” เป็นชื่อองค์กรแต่เมื่ออยู่ในบริบทนี้จะกลายเป็นชื่อสถานที่ แต่นั่นไม่ได้หมายความว่าชื่อเฉพาะที่ตามหลัง “ที่” จำเป็นต้องเป็นชื่อสถานที่เสมอไป เช่น “...เป็นสิ่ง**ที่**ประเทศญี่ปุ่นอยากได้” จากตัวอย่างจะเห็นว่า “ประเทศญี่ปุ่น” อยู่ตามหลัง “ที่” แต่ “ที่” ในที่นี้ไม่ได้ทำหน้าที่เป็นคำบุพบทแต่เป็นคำเชื่อมส่วนขยาย และ “ประเทศญี่ปุ่น” มีคำกริยา “อยาก” ตามมาแสดงว่า “ประเทศญี่ปุ่น” ในบริบทนี้เป็นชื่อสถานที่อ้างข้ามประเภทเป็นชื่อองค์กร ดังนั้นการดึงเฉพาะคำบริบทที่ปรากฏร่วมกับชื่อเฉพาะออกมาโดยไม่พิจารณาหน้าที่ของคำนั้น ๆ จึงไม่ได้ช่วยให้ระบบมีประสิทธิภาพในการรู้จำชื่อเฉพาะดีขึ้น

สำหรับคุณสมบัติคำทั่วไปที่คาดว่าจะช่วยในส่วนของชื่อเฉพาะที่ประกอบด้วยคำที่ไม่พบโดยทั่วไป แต่จากผลการทดสอบแบบจำลองพบว่าคุณสมบัติคำทั่วไปช่วยให้แบบจำลองที่ใช้ข้อมูลตัดคำดีขึ้นเพียงเล็กน้อยเท่านั้นและแทบไม่ช่วยแบบจำลองที่ใช้ข้อมูลแบบตัดพยางค์เลย ดังแสดงไว้ในตารางต่อไปนี้

ตารางที่ 5.4 ผลการเปรียบเทียบค่า F-measure ระหว่างแบบจำลองที่ใช้คุณสมบัติ unigram และ bigram กับแบบจำลองที่ใช้คุณสมบัติ unigram และ bigram ร่วมกับคุณสมบัติคำทั่วไป

| F-measure (%) | Unigram และ bigram | | Unigram และ bigram + คำทั่วไป | |
|---|---|---|---|---|
| | WSG | SSG | WSG | SSG |
| ชื่อบุคคล | 85.25 | 88.22 | 85.36 | 88.18 |
| ชื่อองค์กร | 74.21 | 75.50 | 75.07 | 75.92 |
| ชื่อสถานที่ | 72.41 | 73.37 | 71.96 | 73.09 |
| ทั้งหมด | 77.93 | 79.72 | 78.15 | 79.80 |

จากตารางเมื่อดูจากค่า F-measure ของชื่อเฉพาะทั้งหมดจะเห็นว่าข้อมูลแบบตัดคำเพิ่มขึ้นจากการใช้เฉพาะคุณสมบัติ unigram และ bigram เพียง 0.22% และข้อมูลตัดพยางค์เพิ่มขึ้นเพียง 0.08% เท่านั้น สาเหตุที่เป็นเช่นนี้ ผู้วิจัยคาดว่ามาจากลักษณะของคำและพยางค์ที่เมื่อผ่านโปรแกรมแล้ว ในข้อมูลแบบตัดคำจะมีคำบางส่วนมีลักษณะเป็นวลี เช่น ก่อนหน้านี้

ห้องปฏิบัติการ ที่ผ่านมา เป็นต้น ซึ่งคำที่เป็นวลีนี้จะไม่มีอยู่ในรายการคำทั่วไปที่ผู้วิจัยเตรียมไว้ ทำให้การกำกับคุณสมบัติมีข้อผิดพลาด เพราะกำกับส่วนที่เป็นวลีให้มีค่าเดียวกับส่วนที่เป็นชื่อเฉพาะจริง ๆ สำหรับข้อมูลแบบตัดพยางค์เนื่องจากข้อมูลอยู่ในระดับที่แยกย่อย เมื่อชื่อถูกแยกออกเป็นพยางค์ โอกาสที่แต่ละพยางค์จะไปตรงกับส่วนของคำในรายการคำทั่วไปจึงมีสูง เช่น กรมประชาสัมพันธ์ เมื่อตัดพยางค์จะได้เป็น กรม-ประ-ชา-สัม-พันธ์ เห็นได้ว่าแต่ละ token ในที่นี้สามารถไปตรงกับส่วนของคำทั่วไปอื่น ๆ ได้ ดังนั้นในข้อมูลแบบตัดพยางค์จึงมีปัญหาว่าส่วนของชื่อเฉพาะจริง ๆ ตรงกับคำทั่วไปมีมาก จากสาเหตุดังกล่าวจึงส่งผลให้คุณสมบัติคำทั่วไปไม่ช่วยให้ระบบดีขึ้นเท่าใดนัก

### 5.3 ลักษณะทางภาษาศาสตร์ที่มีผลต่อประสิทธิภาพของแบบจำลอง

จากลักษณะของชื่อเฉพาะ เมื่อเราวิเคราะห์รูปแบบการปรากฏของชื่อจะเห็นว่าชื่อบุคคลที่มีลักษณะโครงสร้างที่ค่อนข้างแน่นอนมากกว่าชื่อองค์กรและชื่อสถานที่ เพราะชื่อบุคคลมักปรากฏโดยมีคำนำหน้าชื่อซึ่งสามารถใช้เป็นทั้งคำบ่งชี้ชนิดรวมถึงบอกจุดเริ่มต้นของชื่อได้ แม้ว่าชื่อองค์กรและชื่อสถานที่จะมีคำบ่งชี้ชื่อเช่นกันก็ตาม แต่เมื่อเทียบสัดส่วนแล้วจะพบว่าชื่อบุคคลจะปรากฏโดยมีคำนำหน้าชื่อร่วมด้วยมากที่สุด รองลงมาคือชื่อองค์กร และชื่อสถานที่ อีกทั้งชื่อบุคคลมีการใช้ช่องว่างในการแยกส่วนของชื่อและนามสกุลออกจากกันอย่างชัดเจน ซึ่งการใช้ช่องว่างภายในชื่อนี้ จะพบกับชื่อบุคคลมากกว่าชื่อชนิดอื่นด้วยลักษณะดังกล่าวทำให้ระบบสามารถรู้จำชื่อบุคคลได้มากกว่าชื่อองค์กรและชื่อสถานที่อย่างเห็นได้ชัด โดยหากดูผลจากการใช้จำนวนชื่อค่า F-measure ของระบบที่ใช้ข้อมูลตัดคำและตัดพยางค์เท่ากับ 89.16% และ 89.01% ตามลำดับ ในขณะที่ชื่อองค์กรเท่ากับ 77.99% และ 78.02% และชื่อสถานที่เท่ากับ 74.98% ในทั้งสองระบบ

ในงานวิจัยนี้ ไม่ได้ใช้คุณสมบัติที่เกี่ยวกับเชิงภาษาศาสตร์มากนัก เนื่องจากมีข้อจำกัดในเรื่องของเวลา เพราะการให้ข้อมูลเชิงภาษากับระบบจำเป็นต้องอาศัยการกำกับข้อมูลเหล่านั้นให้กับข้อมูลในคลังฝึกฝนทั้งหมดทำให้ใช้เวลาค่อนข้างมากแต่อย่างไรก็ตาม หลังจากที่ผู้วิจัยได้ทดสอบระบบแล้ว พบว่าระบบมีปัญหาในเรื่องของชื่อเฉพาะอ้างข้ามประเภท เนื่องจากมีชื่อจำนวนหนึ่งที่ระบบสามารถระบุขอบเขตของชื่อได้ถูกต้องแต่ระบุประเภทของชื่อเฉพาะผิดจึงส่งผลให้ค่าความถูกต้องของชื่อองค์กรและชื่อสถานที่ต่ำลงไปด้วย สำหรับชื่อเฉพาะชนิดนี้จำเป็นต้องอาศัยคำบริบทรอบข้างในการกำหนดชนิดของชื่อ แต่ทั้งนี้การนำคำบริบทมาใช้จำเป็นต้องดูหน้าที่ของคำด้วย เพราะการใช้เพียงแค่รูปของภาษาไม่ได้ช่วยให้ระบบสามารถแยกความแตกต่างระหว่างชื่อเฉพาะปกติกับชื่อเฉพาะอ้างข้ามประเภทได้ ดังนั้นสิ่งที่อาจช่วยได้คือเพิ่มข้อมูลเกี่ยวกับ

หน้าที่ของคำให้กับระบบ เพื่อให้ระบบสามารถแยกความแตกต่างของคำได้ และรู้ว่าคำใดมีหน้าที่อะไร เพราะชนิดของคำที่ปรากฏร่วมกับชื่อองค์กรและสถานที่ค่อนข้างต่างกัน โดยชื่อองค์กรมักปรากฏร่วมกับคำกริยา ในขณะที่ชื่อสถานที่จะปรากฏร่วมกับคำบุพบทหรือคำกริยาบางประเภทที่เกี่ยวข้องกับสถานที่ เช่น เดินทางไป จากข้อมูล หาก “เดินทางไป” ตามด้วยคำนาม คำนามนั้นจะเป็นชื่อสถานที่เสมอ หรือหากตามด้วยคำบุพบท ก็มักเป็นคำบุพบทที่ชี้บ่งสถานที่ เช่น ยัง ที่ ถึง เป็นต้น โดยคำบุพบทเหล่านี้มักตามด้วยคำนามที่เป็นชื่อสถานที่ แต่ถ้าหากว่า “เดินทางไป” ตามด้วยคำชนิดอื่น เช่น คำกริยา หรือคำวิเศษณ์ ก็จะไม่เกี่ยวข้องกับชื่อสถานที่ เป็นต้น จากที่กล่าวมาจะเห็นได้ว่าการกำกับหมวดคำให้กับข้อมูลมีแนวโน้มว่าน่าจะช่วยในการสกัดชื่อเฉพาะอ้างข้ามประเภทได้ กล่าวคือ หากระบบพบว่าชื่อสถานที่ไปปรากฏร่วมกับคำกริยา หรือชื่อองค์กรปรากฏร่วมกับคำบุพบท ก็มีแนวโน้มว่าชื่อสถานที่หรือชื่อองค์กรนั้นจะเป็นชื่ออ้างข้ามประเภท

จากโครงสร้างของชื่อพบว่าเป็นคำประสมหรือวลีก็ได้ แต่เนื่องจากการตั้งชื่อไม่จำเป็นต้องเป็นไปตามหลักภาษา ดังนั้นชื่อบางชื่อจึงสามารถเกิดจากการประสมคำไทยกับคำต่างประเทศได้ เช่น ไทยเบฟเวอเรจ, บริษัท เจริญพลาซ่า จำกัด เป็นต้น หรือมีลักษณะการประสมคำหรือวลีที่ต่างจากวลีทั่วไป เช่น มูลนิธิปวีณาหงสกุลเพื่อเด็กและสตรี เป็นการนำชื่อเฉพาะ “ปวีณา หงสกุล” มาประสมเข้ากับบุพบทวลี “เพื่อเด็กและสตรี” ซึ่งในกรณีเช่นนี้ หากข้อมูลมีการกำกับชนิดของคำรวมถึงให้ข้อมูลว่าคำใดเป็นภาษาต่างประเทศด้วย ก็อาจช่วยในการสกัดชื่อเฉพาะดังกล่าวได้

# บทที่ 6

# สรุปผลการวิจัย ปัญหาและข้อเสนอแนะ

ในส่วนนี้ผู้วิจัยจะสรุปผลการวิจัยทั้งหมด และจะกล่าวถึงปัญหาในการรู้จำชื่อเฉพาะที่พบในงานวิจัยนี้รวมถึงข้อเสนอแนะที่จะเป็นประโยชน์ต่องานวิจัยอื่นต่อไป

## 6.1 สรุปผลการวิจัย

งานวิจัยนี้ได้นำเสนอระบบการรู้จำชื่อเฉพาะ ได้แก่ ชื่อบุคคล ชื่อองค์กร และชื่อสถานที่ โดยใช้แบบจำลองทางสถิติคอนดิชันนอลแรนดอมฟิลด์สพร้อมกับเปรียบเทียบประสิทธิภาพของระบบระหว่างรับข้อมูลเข้าเป็นคำกับพยางค์ โดยผู้วิจัยตั้งสมมติฐานว่าแบบจำลองที่รับข้อมูลเข้าเป็นคำสามารถรู้จำชื่อเฉพาะได้ดีกว่าแบบจำลองที่รับข้อมูลเข้าเป็นพยางค์เนื่องจากชื่อเฉพาะในภาษาไทยส่วนใหญ่เกิดจากการประสมคำเป็นหลัก

การเรียนรู้ของระบบเป็นแบบ Supervised learning คือ มีการให้คำตอบแก่ระบบในคลังข้อมูลสำหรับฝึกฝน แบบคำตอบที่ใช้ทั้งหมดมี 5 แบบด้วยกัน โดยแต่ละแบบให้ข้อมูลมากน้อยแตกต่างกันไป ดังนี้

แบบที่ 1 P, O, L, X

แบบที่ 2 B, I, X – PER, ORG, LOC

แบบที่ 3 B, I, X – P, O, L, OL, LO

แบบที่ 4 B, I, E, X – PER, ORG, LOC

แบบที่ 5 B, I, E, X – P, O, L, OL, LO

โดยที่ B: จุดเริ่มต้นของชื่อ, I: ภายในชื่อ, E: จุดสิ้นสุดของชื่อ, X: ส่วนที่ไม่ใช่ชื่อ คำตอบแบบที่ 1, 2 และ 4 แบ่งชนิดออกเป็นชื่อบุคคล ชื่อองค์กร และชื่อสถานที่ ส่วนคำตอบแบบที่ 3 และ 5 แบ่งชนิดออกเป็นชื่อบุคคล ชื่อองค์กร ชื่อสถานที่ ชื่อองค์กรอ้างถึงสถานที่ และชื่อสถานที่อ้างถึงองค์กร

คุณสมบัติที่ใช้ในการฝึกฝนระบบ ได้แก่ รายการชื่อเฉพาะ คำบริบท คำย่อ รายการคำทั่วไป สถิติ และ template ที่ใช้คือ unigram และ bigram เมื่อเปรียบเทียบประสิทธิภาพของระบบที่ได้รับคำตอบต่างกัน พบว่า ผลของระบบที่ใช้ข้อมูลตัดคำกับตัดพยางค์ไม่ต่างกัน คือ ระบบที่ได้รับคำตอบแบบที่ 1 มีประสิทธิภาพน้อยที่สุดเนื่องจากมีการระบุขอบเขตของชื่อผิดพลาดมากกว่าแบบอื่น และแบบที่ 4 มีประสิทธิภาพมากที่สุด ดังนั้นในการทดสอบระบบขั้นต่อไปผู้วิจัย

จึงเลือกใช้เฉพาะคำตอบแบบที่ 4 เท่านั้น สำหรับคำตอบแบบที่ 5 ซึ่งมีรายละเอียดคำตอบมากที่สุดได้ผลต่ำกว่าแบบที่ 4 เล็กน้อย สาเหตุมาจากคลังข้อมูลที่ใช้ในการฝึกฝนมีไม่มากพอ เพราะการได้รับรายละเอียดคำตอบที่มากขึ้น จำเป็นต้องใช้คลังข้อมูลสำหรับฝึกฝนมากขึ้นตามไปด้วย

เมื่อทดสอบคุณสมบัติแต่ละชนิดว่ามีผลต่อประสิทธิภาพของระบบมากน้อยเพียงใดโดยการใช้คุณสมบัติที่ต้องการทดสอบควบคู่กับคุณสมบัติ unigram และ bigram ซึ่งเป็น template และสำหรับคุณสมบัติ unigram และ bigram จะประมวลผลโดยไม่มีคุณสมบัติอื่นร่วมด้วย ผลที่ออกมาแสดงให้เห็นว่าคุณสมบัติ unigram และ bigram ช่วยให้ประสิทธิภาพของระบบที่ใช้ข้อมูลตัดพยางค์ดีกว่าข้อมูลตัดคำ เพราะระดับพยางค์แยกย่อยกว่าระดับคำ ทำให้สามารถหาความสัมพันธ์ของพยางค์ภายในชื่อเฉพาะซึ่งเป็นคำ ๆ เดียวและปรากฏโดยไม่มีคำบ่งชี้ได้ ส่วนคุณสมบัติรายการชื่อเฉพาะสนับสนุนข้อมูลตัดคำมากกว่าตัดพยางค์ เนื่องจากลักษณะการกำหนดคุณสมบัติ คือ นำแต่ละ token ไปเทียบกับรายการชื่อเฉพาะว่าตรงกับส่วนใดของชื่อหรือไม่ ลักษณะเช่นนี้ทำให้ข้อมูลแบบตัดคำได้เปรียบกว่าเพราะการตัดพยางค์ทำให้โอกาสที่พยางค์ของคำทั่วไปจะไปเป็นส่วนหนึ่งของชื่อมีสูง ส่งผลให้ระบบนำพยางค์ที่ไม่ใช่ชื่อเฉพาะไปประมวลผลร่วมกับส่วนที่เป็นชื่อเฉพาะจริง ๆ และคุณสมบัติที่ทำให้ประสิทธิภาพของทั้งสองระบบต่ำลง คือ คุณสมบัติค่าทางสถิติเนื่องจาก token ที่ตรงตามคุณสมบัติส่วนใหญ่ไม่ใช่ชื่อเฉพาะ สำหรับคุณสมบัติอื่น ๆ ช่วยให้ประสิทธิภาพของระบบดีขึ้นเพียงเล็กน้อย

เมื่อนำข้อมูลมาผ่านกระบวนการประมวลผลภายหลัง ทำให้ค่าความครบถ้วนของทั้งสองระบบมากขึ้นจาก 77.64% เป็น 80.15% และ 80.06% ในข้อมูลตัดคำและตัดพยางค์ตามลำดับ แต่ค่าความแม่นยำกลับลดลงโดยเฉพาะข้อมูลแบบตัดพยางค์ สาเหตุเนื่องจากขั้นตอนนี้เน้นไปที่การสกัดชื่อเฉพาะที่ระบบไม่ได้สกัดออกมาและมีการนำชื่อเฉพาะที่ระบบสกัดออกมาได้มาสร้างเป็นรายการชื่อเฉพาะใหม่เพื่อสกัดชื่อที่เหลือโดยไม่ได้ปรับแก้ให้ขอบเขตและชนิดของชื่อเฉพาะถูกต้องเสียก่อน เมื่อจำนวนชื่อเฉพาะที่สกัดผิดเพิ่มขึ้นจึงส่งผลให้ค่าความแม่นยำลดลง

แม้ว่าเมื่อผ่านขั้นตอนประมวลผลภายหลังแล้วจะทำให้ค่า F-measure ของระบบที่ใช้ข้อมูลตัดพยางค์ต่ำกว่าระบบที่ใช้ข้อมูลตัดคำ แต่เมื่อพิจารณาถึงประสิทธิภาพของระบบก่อนขั้นตอนนี้แล้วจะเห็นว่าไม่ต่างกันเพราะทั้งสองระบบได้ค่า F-measure เท่ากัน คือ 81.30% โดยสิ่งที่มีผลต่อระบบคือคุณสมบัติที่ใช้ว่าออกแบบได้เหมาะสมกับข้อมูลหรือไม่ แม้ว่าชื่อเฉพาะส่วนใหญ่จะเกิดจากการประสมคำแต่ภายในคำเมื่อตัดแบ่งเป็นพยางค์แล้วก็สามารถหาความเกาะเกี่ยวกันระหว่างพยางค์ภายในคำได้ดังจะเห็นได้จากการใช้คุณสมบัติ unigram และ bigram

อย่างไรก็ตาม เมื่อพิจารณาถึงแต่ละคุณสมบัติที่ใช้นอกจากคุณสมบัติทางสถิติที่ทำให้ระบบทั้งสองประสิทธิภาพลดลงแล้ว คุณสมบัติอื่น ๆ ต่างช่วยให้ประสิทธิภาพของระบบที่ใช้ข้อมูลตัดคำดีกว่าข้อมูลตัดพยางค์ยกเว้นเพียงคุณสมบัติ unigram และ bigram เท่านั้น ดังนั้นการเลือก

คุณสมบัติให้ข้อมูลตัดพยางค์จึงดูเหมือนมีข้อจำกัดมากกว่าด้วยลักษณะของข้อมูลที่แยกย่อยเกินไปอีกทั้งหากต้องให้ข้อมูลทางภาษาศาสตร์ที่มากขึ้นแก่ระบบ เช่น การกำกับหมวดคำ (part of speech) หรือการให้ข้อมูลทางความหมาย ก็น่าจะเหมาะสมกับระดับคำมากกว่า

สำหรับลักษณะทางภาษาศาสตร์ที่มีผลต่อประสิทธิภาพของระบบนั้น ในงานวิจัยนี้ สิ่งที่มีส่วนช่วยมากที่สุด คือ คำบ่งชี้ที่ปรากฏร่วมกับชื่อเฉพาะ เช่น คำนำหน้าชื่อ ตำแหน่งทางวิชาการ คำบ่งสถานที่และองค์กร เป็นต้น โดยชื่อบุคคลจะปรากฏร่วมกับคำบ่งชี้มากกว่าชื่อเฉพาะชนิดอื่น ๆ เพราะชื่อบุคคลมักมีคำนำหน้าชื่อปรากฏร่วมด้วย อีกทั้งชื่อบุคคลไม่ต้องมีการอ้างข้ามประเภทดังเช่นชื่อองค์กรและชื่อสถานที่ ดังนั้นระบบจึงสามารถรู้จำชื่อบุคคลได้ถูกต้องและมากกว่าชื่อเฉพาะชนิดอื่น โดยค่า F-measure ของชื่อบุคคลในข้อมูลตัดคำและตัดพยางค์เท่ากับ 89.16% และ 89.01% ตามลำดับ ในขณะที่ค่า F-measure ของชื่อองค์กรเท่ากับ 77.99% และ 78.02% และชื่อสถานที่เท่ากับ 74.98% ในทั้งสองระบบ โดยส่วนของชื่อองค์กรและชื่อสถานที่ระบบมักมีปัญหาในการรู้จำชื่อเฉพาะอ้างข้ามประเภท

แม้ว่าในงานวิจัยได้นำเอาคำบริบทที่คาดว่าจะช่วยในการระบุชนิดของชื่อเฉพาะอ้างข้ามประเภทมาเป็นส่วนหนึ่งของคุณสมบัติ แต่จากผลการทดสอบที่ออกมาพบว่าไม่ได้ช่วยทำให้ระบบมีประสิทธิภาพในการรู้จำมากขึ้น เนื่องจากผู้วิจัยนำคำบริบทมาใช้โดยไม่คำนึงถึงหน้าที่หรือหมวดคำของคำบริบทนั้น ๆ จึงทำให้ได้คำทั่วไปและคำที่ไม่ได้ช่วยบ่งชี้มารวมอยู่ในรายการคำบริบทเป็นจำนวนมาก อีกทั้งเมื่อนำรายการคำไปเทียบกับแต่ละ token ในข้อมูล คำที่มีรูปเดียวกันแต่หน้าที่ต่างกัน เช่น “ไป” ที่เป็นคำกริยาหลักเกิดหลังประธานซึ่งสามารถช่วยบ่งชี้ชื่อสถานที่อ้างถึงองค์กรได้ เช่น “ไทย***ไป***ดูงานที่ญี่ปุ่น” เป็นต้น กับ “ไป” ที่เป็นคำกริยาช่วยขยายความหมายของกริยาหลัก เช่น”ตึกใบหยกสูง***ไป***” เป็นต้น ซึ่ง “ไป” ประเภทหลังมักไม่ได้ช่วยระบุชนิดของชื่อเฉพาะ ก็จะถูกกำหนดว่าเป็นคำบริบทช่วยบ่งชี้เหมือนกัน จึงส่งผลให้คำบริบทไม่มีส่วนช่วยให้ระบบมีประสิทธิภาพในการรู้จำชื่อเฉพาะมากขึ้น

นอกจากคุณสมบัติคำบริบทแล้ว อีกคุณสมบัติหนึ่งที่เกี่ยวข้องกับลักษณะทางภาษาศาสตร์ คือ คุณสมบัติคำทั่วไปที่มีพื้นฐานมาจากลักษณะการตั้งชื่อ เพราะมีชื่อจำนวนหนึ่งที่เกิดจากการสมาสหรือสนธิคำ ยืมคำจากภาษาต่างประเทศ หรือปรับเปลี่ยนคำ ดังนั้นจึงอาจมีบางส่วนของชื่อที่แปลก ไม่พบเห็นได้ทั่วไป ผู้วิจัยจึงใช้คุณสมบัตินี้เพื่อช่วยสกัดชื่อเฉพาะในลักษณะดังกล่าว แต่จากผลการทดสอบแสดงให้เห็นว่าคุณสมบัตินี้ช่วยการรู้จำของระบบได้เพียงเล็กน้อยเท่านั้น เนื่องจากในข้อมูลแบบตัดคำมี token จำนวนหนึ่งที่เป็นวลี ซึ่งวลีเหล่านี้ไม่มีอยู่ในรายการคำทั่วไป ดังนั้นส่วนที่เป็นชื่อเฉพาะจริง ๆ กับวลีเหล่านี้จึงมีค่าคุณสมบัติเดียวกัน ในขณะที่ระบบที่ใช้ข้อมูลตัดพยางค์พบปัญหา token ที่เป็นส่วนของชื่อเฉพาะจริง ๆ ไปตรงกับส่วนของคำทั่วไปเป็นจำนวนมาก เนื่องจากการตัดพยางค์ที่แยกย่อยทำให้โอกาสที่แต่ละ token จะตรงกับ

ส่วนของคำมีสูง ด้วยเหตุผลดังกล่าวคุณสมบัติคำทั่วไปจึงไม่มีประสิทธิภาพในการช่วยให้ระบบรู้จำชื่อเฉพาะได้เท่าใดนัก

## 6.2 ปัญหาที่พบในการรู้จำชื่อเฉพาะ

ปัญหาหลักที่พบในระบบการรู้จำชื่อเฉพาะทั้งในข้อมูลแบบตัดคำและตัดพยางค์จะคล้ายกัน สามารถแบ่งได้เป็น 2 ปัญหา คือ ปัญหาที่เกิดจากคลังข้อมูลฝึกฝนและทดสอบ และปัญหาด้านระบบการรู้จำชื่อเฉพาะ

### 6.2.1 ปัญหาด้านคลังข้อมูล

1) ปัญหาเรื่องเครื่องหมายวรรคตอน ปัญหาที่พบบ่อยในคลังข้อมูล คือ ข้อมูลไม่มีการใช้เครื่องหมายวรรคตอน เช่น จุลภาค (,) หรือเว้นวรรค (<s>) เช่น “ท่องเที่ยวที่ฝรั่งเศสญี่ปุ่นจีนฮ่องกงและมัลดีฟส์” “โรงสีธัญญาเรืองอ.ท่าตะโกจ.นครสวรรค์” “ร.ท.อภินันท์ สุมนะเศรณีกรรมการผู้อำนวยการใหญ่บริษัท การบินไทยนางกัลยา ผกากรองรักษาการกรรมการ” เป็นต้น อีกปัญหาคือการเว้นวรรคตอนผิด เช่น “นายบวรศักดิ์ลา<s>ออกไปเตรียมเล่นการเมือง” เป็นต้น ลักษณะเช่นนี้แม้ระบบสามารถรู้จำว่าเป็นชื่อเฉพาะได้แต่มักระบุขอบเขตของชื่อผิด ซึ่งเครื่องหมายวรรคตอนถือเป็นสิ่งสำคัญสิ่งหนึ่งที่ช่วยในการบอกขอบเขตของชื่อเฉพาะได้

2) ปัญหาการตัดคำผิด เป็นอีกปัญหาหนึ่งที่มีผลต่อการรู้จำชื่อเฉพาะค่อนข้างมาก เพราะจะทำให้ระบบระบุขอบเขตของชื่อผิด ดังเช่น “นางคนึงนิจ วาทีสาธกกิจ” โปรแกรมตัดคำเป็น “นางค|นึง|นิจ|<s>|วาที|สาธก|กิจ” ทำให้ระบบรู้จำเพียง “นึงนิจ วาทีสาธกกิจ” เนื่องจากระบบไม่สามารถรู้จำส่วนที่เป็นคำนำหน้าชื่อได้

### 6.2.2 ปัญหาด้านระบบการรู้จำชื่อเฉพาะ

1) ปัญหาการรู้จำชื่อเฉพาะข้ามประเภท ปัญหาที่พบส่วนใหญ่ในงานวิจัยนี้ คือ ระบบสามารถระบุขอบเขตของชื่อเฉพาะได้ แต่จะระบุชนิดของชื่อเฉพาะผิดโดยเฉพาะชื่อเฉพาะอ้างข้ามประเภท แม้ว่าการอ้างข้ามประเภทส่วนใหญ่มักเกิดจากบริบทรอบข้างเป็นตัวกำหนด แต่ทั้งนี้ก็ต้องคำนึงถึงหน้าที่ของคำด้วยเช่นกันดังที่ได้กล่าวไปแล้วในบทสรุป

2) ปัญหาการระบุขอบเขตของชื่อเฉพาะผิด ปัญหานี้ส่วนหนึ่งมาจากข้อมูลไม่มีการใช้เครื่องหมายวรรคตอน แต่อีกส่วนก็มาจากลักษณะของชื่อเฉพาะที่ทำให้ระบบเกิดความสับสนในการแบ่งขอบเขตของชื่อ สามารถแบ่งได้เป็น 2 ลักษณะ ดังนี้

ประเภทแรก คือ ชื่อเฉพาะ 2 ชื่อปรากฏร่วมกัน แต่ชื่อเฉพาะที่ตามหลังใช้เพื่อขยายชื่อเฉพาะก่อนหน้า เช่น มหาวิทยาลัยเกษตรศาสตร์บางเขน กระทรวงเกษตรจีน เป็นต้น จากตัวอย่าง “บางเขน” เป็นชื่อเฉพาะสถานที่ใช้ขยายเพื่อให้รู้ว่าเป็น “มหาวิทยาลัยเกษตรศาสตร์” ที่วิทยาเขตนี้ไม่ใช่วิทยาเขตอื่น และ “จีน” เป็นชื่อเฉพาะองค์กรใช้ขยาย “กระทรวงเกษตร” เพื่อให้รู้ว่าสังกัดประเทศอะไร แต่ทั้งนี้เนื่องจากชื่อเฉพาะหลักและชื่อเฉพาะขยายปรากฏอยู่ติดกันทำให้ระบบรู้จำว่าเป็นชื่อเฉพาะเดียว

อีกประเภทหนึ่ง คือ ภายในชื่อเฉพาะมีคำบ่งซ้อนกันโดยคำซ้อนนั้นเป็นส่วนหนึ่งของชื่อเฉพาะ เช่น เจ้าหน้าที่ตำรวจจาก*ศูนย์ปฏิบัติการสำนักงานตำรวจแห่งชาติส่วนหน้า* เป็นต้น ส่วนที่ขีดเส้นใต้คือชื่อเฉพาะองค์กร คำบ่งชื่อคือ “ศูนย์” ส่วน “สำนักงาน” โดยปกติสามารถเป็นคำบ่งชื่อเฉพาะได้เช่นกัน แต่ในที่นี้ “สำนักงาน” เป็นส่วนหนึ่งของชื่อ แต่ระบบกลับรู้จำเพียง “สำนักงานตำรวจแห่งชาติส่วนหน้า” เป็นชื่อองค์กรเท่านั้น ทั้งนี้อาจเป็นไปได้ว่าระบบพบ “สำนักงานตำรวจแห่งชาติ” ในข้อมูลมากกว่า “ศูนย์ปฏิบัติการสำนักงานตำรวจแห่งชาติส่วนหน้า” จึงทำให้ระบบรู้จำเช่นนี้

3) ปัญหาระบบรู้จำชื่อเฉพาะผิดเนื่องจากการสับสนระหว่างชื่อเฉพาะและคำทั่วไป ปัญหานี้สามารถแบ่งได้เป็น 2 กรณี คือ ระบบรู้จำคำทั่วไปให้เป็นชื่อเฉพาะ และระบบไม่รู้จำชื่อเฉพาะที่ตรงกับคำทั่วไป

กรณีแรกมักเกิดกับชื่อองค์กรและสถานที่ เพราะส่วนใหญ่คำบ่งชื่อโดยเฉพาะชื่อองค์กรและสถานที่ เช่น บริษัท จังหวัด เป็นต้น สามารถทำหน้าที่เป็นคำนามทั่วไปได้ จึงไม่จำเป็นต้องมีชื่อเฉพาะตามมาเสมอไป ดังนั้นในบางกรณีระบบจึงรู้จำชื่อเฉพาะผิดเพราะเข้าใจว่าคำนามนั้นเป็นคำบ่งชื่อจึงกำกับให้คำที่ตามมาเป็นชื่อเฉพาะ เช่น “องค์การอนามัยโลกเป็นองค์การเกี่ยวกับสุขภาพอนามัย” จากตัวอย่างระบบรู้จำว่า “องค์การอนามัยโลก” และ “องค์การเกี่ยวกับสุขภาพอนามัย” เป็นชื่อองค์กร จากทั้งสองชื่อที่ระบบรู้จำจะเห็นว่าขึ้นต้นด้วย “องค์การ” เหมือนกันโดย “องค์การ” แรกเป็นคำบ่งชื่อเฉพาะจริง ในขณะที่ “องค์การ” ที่สองเป็นเพียงคำนามทั่วไปที่ใช้แทนองค์การอนามัยโลกเท่านั้น สำหรับชื่อบุคคลจะมีปัญหากรณีที่คำไปตรงคำบ่งชื่อบุคคล เช่น “เวสนาย ไวรัส” ระบบจะสกัด “นาย ไวรัส” ออกมา หรือ ชื่อสถานที่ “เดิมบางนางบวช” ระบบก็สกัดเฉพาะ “นางบวช” ให้เป็นชื่อบุคคลแทน เป็นต้น

อีกกรณีคือระบบเข้าใจว่าชื่อเฉพาะเป็นคำทั่วไป เพราะเป็นคำที่มีรูปเหมือนกัน เช่น “ผศ.น.สพ.ธีระพล ศิรินฤมิตร และผศ.น.สพ.ดร.ทวีศักดิ์ *ส่งเสริม* ร่วมกันแถลงถึง

ผลการทดสอบ” จากตัวอย่าง “ส่งเสริม” เป็นส่วนหนึ่งของชื่อเฉพาะเพราะเป็นนามสกุล แต่เนื่องจากคำนี้มีรูปตรงกับคำว่า “ส่งเสริม” ที่เป็นคำทั่วไป ทำให้ระบบเข้าใจว่า “ส่งเสริม” นี้ไม่ใช่ส่วนหนึ่งของชื่อ จึงรู้จำเพียง “ผศ.น.สพ.ดร.ทวีศักดิ์” อีกตัวอย่าง คือ “นายสมพร ประชุมอดีตกำนันตำบลตลิ่งชัน” จากตัวอย่าง “ประชุม” เป็นนามสกุลที่ตรงกับคำทั่วไปอีกทั้งปรากฏโดยไม่มีการเว้นวรรคระหว่างนามสกุลกับวลีขยาย ดังนั้นระบบจึงไม่รู้จำ “ประชุม” เป็นส่วนหนึ่งของชื่อ

4) ปัญหาการรู้จำชื่อเฉพาะที่มีลักษณะคล้ายวลีทั่วไป เช่น องค์การเฝ้าระวังโรคระบาดในสัตว์ ศูนย์ปฏิบัติการโรคไข้หวัดนก เป็นต้น โดยชื่อเหล่านี้ส่วนใหญ่มักเป็นชื่อองค์กรและระบบไม่สามารถรู้จำได้ว่าเป็นชื่อเฉพาะ

## 6.3 ข้อเสนอแนะ

เนื่องจากคุณสมบัติต่าง ๆ ที่ใช้ในงานวิจัยนี้ส่วนใหญ่ไม่ได้ลงลึกไปในรายละเอียดเชิงภาษาศาสตร์มากนัก อีกทั้งลักษณะการกำหนดคุณสมบัติที่มีการใช้รายการชื่อหรือรายการคำมาเทียบโดยการนำเอา token ไปเทียบว่าเป็นส่วนหนึ่งของคำหรือชื่อหรือไม่ ลักษณะเช่นนี้ทำให้เกิดการกระจายของข้อมูล เพราะมีคำหรือส่วนที่ไม่ใช่ชื่อเฉพาะมาปะปนด้วยค่อนข้างมาก ดังนั้นในการเทียบชื่อหรือคำจึงควรดูคำข้างเคียงร่วมด้วยหรือดูไปทั้งชื่อ

จากงานวิจัยจะเห็นว่าเพียงแค่การใช้คำบริบทไม่สามารถช่วยให้ระบบรู้จำได้ดีขึ้นหรือระบุชนิดของชื่อเฉพาะได้ถูกต้อง ดังนั้นสิ่งที่อาจจะช่วยได้คือการกำกับหมวดคำ เพราะจะช่วยแยกคำที่มีรูปเดียวกันแต่มีหน้าที่ต่างกันออกไปได้ รวมถึงช่วยในเรื่องชื่อเฉพาะอ้างข้ามประเภท เช่น คำกริยามักช่วยบ่งเรื่องชื่อเฉพาะสถานที่อ้างเป็นชื่อองค์กร หรือชื่อองค์กรที่ตามหลังคำบุพบทและปิดท้ายด้วยช่องว่างมักใช้อ้างข้ามเป็นชื่อสถานที่ เป็นต้น อย่างไรก็ตาม การกำกับหมวดคำในคลังข้อมูลมีข้อด้อยคือต้องใช้เวลาค่อนข้างมาก แต่หากสามารถทำได้ประสิทธิภาพในการรู้จำของระบบน่าจะดีขึ้น

สำหรับคุณสมบัติ unigram และ bigram อาจเพิ่มจำนวน gram ให้มากขึ้น เพื่อให้ระบบได้เห็นช่วงคำหรือพยางค์ที่ปรากฏร่วมกันในช่วงที่กว้างขึ้น ซึ่งจะมีส่วนช่วยในการรู้จำชื่อเฉพาะที่มีลักษณะโครงสร้างเป็นวลียาว แต่ทั้งนี้การเพิ่มจำนวน gram อาจมีผลทำให้ระบบต้องใช้เวลาในการประมวลผลด้วย

ในส่วนของวลีทั่วไปที่ระบบเข้าใจว่าเป็นชื่อเฉพาะเนื่องจากวลีนั้นขึ้นต้นด้วยคำบ่งชื่อเฉพาะ หรือมีคำบ่งชื่อเฉพาะอยู่ เช่น “ตำบล<s>52” “เดิมบางนางบวช” เป็นต้น หรือกรณีที่ชื่อสกุลของบุคคลประกอบด้วยคำทั่วไป เช่น “นางฉลวย<s>ป้องกัน<s>เจ้าหน้าที่สาธารณสุขอ.หล่มสัก” “พ.ต.อ.ชำนาญ<s>รวดเร็ว<s>รองผู้บังคับการตำรวจภูธร” เป็นต้น ในกรณีเหล่านี้อาจเขียน

กฎเพิ่มโดยใช้ตัวเลขหรือช่องว่างมาเป็นตัวชี้บ่งขอบเขตของชื่อเฉพาะหรือสกัดแยกส่วนที่ไม่ใช่ชื่อเฉพาะออกไปได้

นอกจากตัวเลขและช่องว่างที่กล่าวไปข้างต้นแล้ว ยังอาจใช้รายการคำที่มักไม่เป็นส่วนหนึ่งของชื่อมาสกัดเอาชื่อเฉพาะที่ระบบสกัดผิดออกไปได้ เช่น จากรายการชื่อองค์กรที่ใช้ ชื่อองค์กรมักมีคำว่า “แห่ง” “และ” “เพื่อ” เป็นส่วนหนึ่งของชื่อ แต่จะไม่พบคำว่า “หรือ” และ “เกี่ยวกับ” เลย เป็นต้น ดังนั้นจึงอาจใช้คำที่ไม่พบเหล่านี้มาช่วยคัดกรองชื่อเฉพาะที่ระบบสกัดออกมาผิดได้ เช่น ระบบสกัด “องค์การเกี่ยวกับสุขภาพอนามัย” ให้เป็นชื่อเฉพาะเนื่องจากวลีนี้มีคำบ่งชื่อเฉพาะ “องค์การ” อยู่ โดยในการคัดกรองนี้อาจเขียนเป็นกฎตรวจสอบหรือนำไปเป็นคุณสมบัติหนึ่งของระบบก็ได้

นอกจากนี้ในขั้นตอนประมวลผลภายหลังควรเพิ่มขั้นตอนการคัดคำทั่วไปออกไปก่อนหากต้องมีการนำชื่อที่ระบบสกัดออกมาไปใช้ รวมถึงมีขั้นตอนตรวจสอบความถูกต้องของชื่อเฉพาะที่ระบบสกัดออกมาได้ เพื่อลดปัญหาจำนวนชื่อเฉพาะผิดเพิ่มขึ้นหรือมีคำทั่วไปเข้ามาปนกับชื่อเฉพาะ ซึ่งจะส่งผลให้ค่าความแม่นยำของระบบลดลง

อย่างไรก็ตาม การเลือกใช้คุณสมบัติหรือการเขียนกฎขึ้นมาช่วยนั้น จำเป็นต้องดูลักษณะของคลังข้อมูลด้วย เพราะข้อมูลแต่ละประเภทจะมีลักษณะแตกต่างกันไป เช่น ถ้าข้อมูลที่ใช้มีลักษณะการเขียนแบบเป็นทางการ ชื่อเฉพาะมักปรากฏในรูปเต็มและมักปรากฏร่วมกับคำบ่งชี้ ในขณะที่ถ้าเป็นข่าวทั่วไป ลักษณะการเขียนมักเป็นทางการไม่มากนัก ลักษณะของชื่อเฉพาะจึงมีทั้งชื่อรูปแบบเต็มและแบบลดรูป อาจปรากฏร่วมกับคำบ่งชี้หรือไม่ก็ได้ แต่ถ้าหากเป็นข่าวกีฬาลักษณะของชื่อก็จะต่างออกไปคือมีการใช้นามแฝงหรือฉายาค่อนข้างมาก เป็นต้น ดังนั้นการกำหนดคุณสมบัติหรือกฎจึงต้องออกแบบให้เหมาะสมกับลักษณะข้อมูลที่จะนำมาสกัดชื่อเฉพาะนั้น ๆ ด้วย

# รายการอ้างอิง

## ภาษาไทย


กำชัย ทองหล่อ. 2547. หลักภาษาไทย. กรุงเทพมหานคร: รวมสาส์น (1977).

นววรรณ พันธุเมธา. 2549. ไวยากรณ์ไทย. พิมพ์ครั้งที่ 3. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

พระยาอุปกิตศิลปสาร. 2546. หลักภาษาไทย. พิมพ์ครั้งที่ 12. กรุงเทพมหานคร: ไทยวัฒนาพานิช.

ราชบัณฑิตยสถาน. 2546. พจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ.2542. กรุงเทพมหานคร: นานมีบุ๊คส์พับลิเคชั่นส์.

สุฤดี ฉัตรไตรมงคล. 2548. การรู้จำและจำแนกของชื่อเฉพาะภาษาไทย. วิทยานิพนธ์ปริญญามหาบัณฑิต, ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

อัศนีย์ ก่อตระกูล. 2549. การประมวลผลภาษามนุษย์ด้วยคอมพิวเตอร์. พิมพ์ครั้งที่ 2. กรุงเทพมหานคร: จรัลสนิทวงศ์การพิมพ์.

อัศนีย์ ก่อตระกูล. 2550. เทคนิคสำคัญสำหรับการประมวลผลภาษา. ใน รายงานการพัฒนาระบบสกัดข้อสนเทศและความรู้จากเอกสารไร้โครงสร้างภาษาไทย, หน้า 5-8 - 5-14. (ม.ป.ท).


## ภาษาอังกฤษ


Aroonmanakun, W. 2002. Collocation and Thai Word Segmentation. In Proceedings of SNLP-Oriental COCOSDA 2002. Prachuapkhirikhan.

Black, W., Rinaldi, F., and Mowatt, D. 1998. FACILE: Description of the NE System used for MUC-7. In Proceedings of the 7th Message Understanding Conference.

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

Chanlekha, H., and Kawtrakul, A. 2004. Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. In International Joint Conference of Natural Language Processing (IJCNLP-2004). Hainan Island.

Charoenpornsawat, P., Kijsirikul, B., and Meknavin, S. 1998. Feature-based Proper Name Identification in Thai. In Proceedings of National Computer Science and Engineering Conference. Bangkok.

Chieu, H. L., and Ng, H. T. 2002. Named Entity Recognition: A Maximum Entropy Approach Using Global Information. In Proceedings of the 19th International Conference on Computational Linguistics. Taipei.

Chinchor, N. 1998. MUC-7 Named Entity Task Definition (version 3.5). In Proceedings of 7th Message Understanding Conference. Fairfax.

Chiong, R. 2008. A Hybrid Learning for Named Entity Recognition Systems. In INFOCOMP Journal of Computer Science, vol. 7(4), pp. 92-98. (n.p.)

Fang, X., and Sheng, H. 2002. A Hybrid Approach for Chinese Named Entity Recognition. In Proceedings of the Fifth International Conference on Discovery Science. Luebeck.

Feng, Y., Huang, R., and Sun, L. 2008. Two Step Chinese Named Entity Recognition Based on Conditional Random Fields Models. In Proceeding of the Sixth SIGHAN Workshop on Chinese Language Processing, pp. 120-123. Hyderabad.

Feng, Y., Sun, L., and Lv, Y. 2006. Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields Models. In Proceeding of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 181-184. Sydney.

Feng, Y., Sun, L., and Zhang, J. 2005. Early Results for Chinese Named Entity Recognition Using Conditional Random Fields Model, HMM and Maximum Entropy. In Proceedings of IEEE Natural Language Processing and Knowledge Engineering 2005 (IEEE NLP-KE'05), pp. 549-552. Wuhan.

Hanks, P. 2006. Proper Names: Linguistic Status. In Brown, E. K., and Anderson, A. (eds), Encyclopedia of Language & Linguistics (second edition), pp. 134-135. London: Elsevier.

Haruechaiyasak, C., Kongyoung, S., and Dailey, M. N. 2008. A Comparative Study on Thai Word Segmentation Approach. In Proceedings of ECTI-CON. Krabi.

He, J., and Wang, H. 2008. Chinese Named Entity Recognition and Word Segmentation Based on Character. In Proceeding of the Sixth SIGHAN Workshop on Chinese Language Processing, pp. 128-131. Hyderabad.

Jing, H., Florian, R., Luo, X., Zhang, T., and Ittycheriah, A. 2003. HowtogetaChinese Name(Entity): Segmentation and Combination Issues. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Cited in Yu, K., Kurohashi, S., Liu, H., and Nakazawa, T. 2006. Chinese Word Segmentation and Named Entity Recognition by Character Tagging. In Proceeding of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney.

Krishnarao, A. A., Gahlot, H., Srinet, A., and Kushwaha, D. S. 2009. A Comparative Study of Named Entity Recognition for Hindi Using Sequential Learning Algorithm. In Proceedings of 2009 IEEE International Advance Computing Conference (IACC 2009), pp. 1164-1169. Patiala.

Kruengkrai, C., Sornlertlamvanich, V., and Isahara, H. 2006. A Conditional Random Field Framework for Thai Morphological Analysis. In Proceeding of the Fifth International Conference on Language Resources and Evaluation. Genoa.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceeding of 18th ICML. San Francisco.

Lehrer, A. 2006. Proper Names: Semantic Aspects. In Brown, E. K., and Anderson, A. (eds), Encyclopedia of Language & Linguistics (second edition), pp. 141-143. London: Elsevier.

McCallum, A., and Li, W. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons. In Proceedings of the Conference on Natural Language Learning, pp. 188-191. Edmonton.

Mao, X., He, S., Bao, S., Dong, Y., and Wang, H. 2008. Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields. In Proceeding of the Sixth SIGHAN Workshop on Chinese Language Processing, pp. 90-93. Hyderabad.

Palmer, D. D. 1997. A Trainable Rule-Based Algorithm for Word Segmentation. In Proc of 35th of ACL & 8th conf. of EACL, pp. 321-328. Cited in Ye, S., Chua, T., Jimin, L. 2002. An Agent-based Approach to Chinese Named Entity Recognition. In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002). Taipei.

Phanarangsan, K., Arnold, S., Mandel, M., and Walker, C. 2006. Simple Named Entity Guidelines Version 6.4-Thai. (n.p).

Reimer, R. 2006. Proper Names: Philosophical Aspects. In Brown, E. K., and Anderson, A. (eds), Encyclopedia of Language & Linguistics (second edition), pp. 137-139. London: Elsevier.

Sekine, S., Grishman, R., 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. Montreal.

Sutton, C. and McCallum, A. 2007. An Introduction to Conditional Random Fields for Relational Learning. MIT Press.

Wu, C., Jan, S., Tsai, R., and Hsu, W. 2006. On Using Ensemble Methods for Chinese Named Entity Recognition. In Proceeding of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 142-145. Sydney.

Wu, X., et al. 2008. An Improved CRF based Chinese Language Processing System for SIGHAN Bakeoff 2007. In Proceeding of the Sixth SIGHAN Workshop on Chinese Language Processing, pp. 155-159. Hyderabad.

Wu, Y., Yang, J., and Lin, Q. 2006. Description of the NCU Chinese Word Segmentation and Named Entity Recognition System for SIGHAN Bakeoff 2006. In Proceeding of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 209-212. Sydney.

Yang, F., Zhao, J., and Zou, B. 2008. CRFs-Based Named Entity Recognition Incorporated with Heuristic Entity List Searching. In Proceeding of the Sixth SIGHAN Workshop on Chinese Language Processing, pp. 171-174. Hyderabad.

Yu, X., Lam, W., Chan, S., Wu, Y., and Chen, B. 2008. Chinese NER Using CRFs and Logic for the Fourth SIGHAN Bakeoff. In Proceeding of the Sixth SIGHAN Workshop on Chinese Language Processing, pp. 102-105. Hyderabad.

Zhang, S., Qin, Y., Wen, J., and Wang, X. 2006. Word Segmentation and Named Entity Recognition for SIGHAN Bakeoff3. In Proceeding of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 158-161. Sydney.

Zhao, H., and Kit, C. 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. In Proceeding of the Sixth SIGHAN Workshop on Chinese Language Processing, pp. 106-107. Hyderabad.

Zhou, J., He, L., Dai, X., and Chen, J. 2006. Chinese Named Entity Recognition with a Multi-Phase Model. In Proceeding of the Fifth SIGHAN Workshop on Chinese Language Processing, pp. 213-216. Sydney.

Zhu, X. 2008. Conditional Random Fields. Lecture on Advanced Natural Language Processing. University of Wisconsin–Madison.

# ภาคผนวก

# ภาคผนวก ก
# ตัวอย่างข้อมูลสำหรับฝึกฝน

ตัวอย่างข้อมูลฝึกฝนแบบตัดคำใช้คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC)*

| token | รายการชื่อเฉพาะ** | | | | | | | | | | | คำบริบท | คำย่อ | คำทั่วไป | คำทางสถิติ | คำตอบ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | | | | | |
| น.พ. | Y | N | N | N | N | N | N | N | N | N | N | N | Y | Y | N | B-PER |
| จรัล | N | Y | N | N | N | N | N | N | N | N | N | N | Y | N | N | I-PER |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | Y | N | N | I-PER |
| ตฤณ | N | Y | N | N | N | N | N | N | N | N | N | N | N | N | N | I-PER |
| วุฒิ | N | Y | Y | Y | N | N | N | N | N | N | Y | N | N | N | N | I-PER |
| พงษ์ | N | Y | Y | Y | N | N | N | Y | N | N | Y | N | N | Y | N | E-PER |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| อธิบดี | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| กรม | N | Y | N | Y | N | Y | Y | N | N | N | N | Y | N | N | N | B-ORG |
| ควบคุม | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N | I-ORG |
| โรค | N | Y | N | N | N | N | Y | N | N | N | N | Y | N | N | N | E-ORG |

* คำตอบแบบอื่นสามารถปรับเปลี่ยนได้ในคอลัมน์สุดท้าย

** รายละเอียดรายการชื่อเฉพาะ

1 - คำนำหน้าชื่อ
2 - ส่วนต้นชื่อบุคคล
3 - ส่วนกลางชื่อบุคคล
4 - ส่วนท้ายชื่อบุคคล
5 - ชื่อย่อองค์กร
6 - ส่วนต้นชื่อองค์กร
7 - ส่วนกลางชื่อองค์กร
8 - ส่วนท้ายชื่อองค์กร
9 - ส่วนต้นชื่อสถานที่
10 - ส่วนกลางชื่อสถานที่
11 - ส่วนท้ายชื่อสถานที่

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| กล่าว | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | Y | X |
| ว่า | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| วันนี้ | N | N | N | N | N | N | N | N | N | N | N | Y | N | Y | N | X |
| (27 | N | N | N | N | N | N | N | N | N | N | N | N | N | Y | N | X |
| ก.พ. | N | N | N | N | Y | N | N | N | N | N | N | Y | Y | Y | N | X |
| ) | N | N | N | N | N | N | N | N | N | N | N | Y | Y | Y | N | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | Y | N | N | X |
| คณะกรรมการ | N | N | N | N | N | Y | Y | N | N | N | N | Y | N | Y | N | X |
| ผู้เชี่ยวชาญ | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| โรค | N | Y | N | N | N | N | Y | N | N | N | N | Y | N | N | N | X |
| ไข้หวัด | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | Y | X |
| นก | N | Y | N | Y | N | N | Y | N | N | Y | Y | Y | N | N | Y | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ซึ่ง | N | Y | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ประกอบด้วย | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ตัวแทน | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| จาก | N | Y | N | N | N | N | Y | N | N | N | Y | Y | N | N | N | X |
| กระทรวง | N | N | N | N | N | Y | Y | Y | N | N | N | Y | N | N | N | B-ORG |
| สาธารณสุข | N | N | N | N | N | N | Y | Y | N | N | N | Y | N | N | N | E-ORG |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| องค์การ | N | N | N | N | N | Y | Y | N | N | N | N | Y | N | N | N | B-ORG |
| อนามัย | N | Y | N | N | N | N | Y | Y | N | N | N | N | N | N | N | I-ORG |
| โลก | N | Y | N | Y | N | N | Y | Y | N | N | Y | Y | N | N | N | E-ORG |
| และ | N | Y | Y | Y | N | N | Y | N | N | Y | N | Y | N | N | N | X |
| หน่วยงาน | N | N | N | N | N | Y | N | N | N | N | N | Y | N | Y | N | X |
| ที่ | N | N | N | N | N | N | Y | N | N | Y | N | Y | N | N | N | X |
| เกี่ยวข้อง | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ได้ | N | N | N | Y | N | N | N | N | N | N | N | Y | N | N | N | X |
| ชื้น | N | N | N | Y | N | N | N | N | N | N | N | Y | N | N | N | X |
| บัญชี | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N | X |
| ผู้ป่วย | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N | X |
| ไข้หวัดใหญ่ | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| จาก | N | Y | N | N | N | N | Y | N | N | N | Y | Y | N | N | N | X |
| สัตว์ | N | N | N | Y | N | N | Y | Y | N | N | N | Y | N | N | N | X |
| ปีก | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| หรือ | N | Y | N | Y | N | N | N | N | N | N | N | Y | N | N | N | X |
| ไข้หวัด | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| นก | N | Y | N | Y | N | N | Y | N | N | Y | Y | Y | N | N | Y | X |
| ... | | | | | | | | | | | | | | | | |

# ภาคผนวก ข
# ตัวอย่างข้อมูลสำหรับทดสอบ

ตัวอย่างข้อมูลทดสอบแบบตัดคำ

| token | รายการชื่อเฉพาะ* | | | | | | | | | | | ค่าบริบท | ค่าย่อ | ค่าทั่วไป | ค่าทางสถิติ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | | | | |
| นาง | Y | Y | Y | Y | N | N | Y | N | N | Y | Y | Y | N | N | N |
| สุดารัตน์ | N | Y | N | N | N | N | N | N | N | N | N | N | N | Y | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| เก | N | Y | Y | N | N | N | Y | N | N | Y | N | N | N | N | N |
| ยุ | N | Y | Y | Y | N | N | N | N | N | N | N | Y | N | N | N |
| รา | N | Y | Y | Y | N | N | Y | N | N | Y | Y | N | N | N | N |
| พันธุ์ | N | Y | N | Y | N | N | Y | Y | N | N | N | N | N | N | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| รมว. | N | N | N | N | Y | N | N | N | N | N | N | Y | Y | Y | N |
| สาธารณสุข | N | N | N | N | N | N | Y | Y | N | N | N | Y | Y | N | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | Y | N | N |

*รายละเอียดรายการชื่อเฉพาะ

1 - คำนำหน้าชื่อ
2 - ส่วนต้นชื่อบุคคล
3 - ส่วนกลางชื่อบุคคล
4 - ส่วนท้ายชื่อบุคคล
5 - ชื่อย่อองค์กร
6 - ส่วนต้นชื่อองค์กร
7 - ส่วนกลางชื่อองค์กร
8 - ส่วนท้ายชื่อองค์กร
9 - ส่วนต้นชื่อสถานที่
10 - ส่วนกลางชื่อสถานที่
11 - ส่วนท้ายชื่อสถานที่

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| กล่าว | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| ว่า | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| ขณะนี้ | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| พบ | N | Y | Y | Y | N | N | N | N | N | Y | N | Y | N | N | N |
| ผู้ป่วย | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N |
| ที่ | N | N | N | N | N | N | Y | N | N | Y | N | Y | N | N | N |
| ติด | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N |
| เชื้อ | N | Y | Y | Y | N | N | N | N | N | N | N | Y | N | N | N |
| ไข้หวัด | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| นก | N | Y | N | Y | N | N | Y | N | N | Y | Y | Y | N | N | N |
| แล้ว | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| 3 | N | N | N | N | N | N | N | Y | N | N | Y | Y | N | Y | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| ราย | N | N | N | N | N | N | Y | Y | N | Y | Y | Y | N | N | Y |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| ซึ่ง | N | Y | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| เสีย | N | N | N | N | N | N | N | Y | N | N | N | Y | N | N | N |
| ชีวิต | N | Y | N | Y | N | N | Y | Y | N | N | N | Y | N | N | N |
| ไป | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | N | N | N | N | N | N | N | Y | N | N | N | Y | N | Y | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| ราย | N | N | N | N | N | N | Y | Y | N | Y | Y | Y | N | N | Y |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| ที่ | N | N | N | N | N | N | Y | N | N | Y | N | Y | N | N | N |
| รพ. | N | N | N | N | Y | N | Y | N | N | N | N | Y | Y | Y | N |
| ดิ | N | Y | Y | N | N | Y | Y | N | N | Y | N | N | Y | N | N |
| ริ | N | Y | Y | Y | N | N | Y | Y | N | Y | Y | N | Y | N | N |
| ราช | N | Y | Y | Y | N | Y | Y | Y | N | Y | Y | Y | N | N | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| เป็น | N | Y | Y | N | N | N | Y | N | N | N | N | Y | N | N | N |
| เด็กชาย | Y | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| จาก | N | Y | N | N | N | N | Y | N | N | N | Y | Y | N | N | N |
| จังหวัด | N | N | N | N | N | N | Y | Y | N | N | Y | Y | N | N | N |
| กาญจนบุรี | N | N | N | N | N | N | N | Y | Y | Y | Y | N | N | Y | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| ส่วน | N | Y | N | Y | N | N | Y | N | N | N | N | Y | N | N | N |
| อีก | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| 2 | N | N | N | N | N | N | Y | Y | N | N | Y | Y | N | Y | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| ราย | N | N | N | N | N | N | Y | Y | N | Y | Y | Y | N | N | Y |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| ได้แก่ | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| เด็กชาย | Y | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| ที่ | N | N | N | N | N | N | Y | N | N | Y | N | Y | N | N | N |
| จ. | Y | N | N | N | Y | N | N | N | N | Y | N | Y | Y | Y | N |
| สุพรรณ | N | Y | N | Y | N | N | Y | N | Y | Y | Y | N | Y | N | N |
| บุรี | N | Y | N | Y | N | N | Y | Y | Y | Y | Y | N | Y | N | N |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| และ | N | Y | Y | Y | N | N | Y | N | N | Y | N | Y | N | N | N |
| จ. | Y | N | N | N | Y | N | N | N | N | Y | N | Y | Y | Y | N |
| สุ | N | Y | Y | Y | N | N | Y | N | Y | Y | N | N | Y | N | N |
| โข | N | Y | Y | Y | N | N | Y | N | N | Y | N | N | Y | N | N |
| ทัย | N | Y | N | Y | N | N | Y | Y | N | N | Y | N | N | Y | N |

# ภาคผนวก ค
# ตัวอย่างผลลัพธ์ที่ได้จากแบบจำลอง CRF

ตัวอย่างผลลัพธ์ที่ได้จากแบบจำลอง CRF ของข้อมูลทดสอบในภาคผนวก ข

| token | รายการชื่อเฉพาะ* | | | | | | | | | | | ค่าบริบท | ค่าย่อ | ค่าทั่วไป | ค่าทางสถิติ | คำตอบจาก CRF** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | | | | | |
| นาง | Y | Y | Y | Y | N | N | Y | N | N | Y | Y | Y | N | N | N | B-PER |
| สุดารัตน์ | N | Y | N | N | N | N | N | N | N | N | N | N | N | Y | N | I-PER |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | I-PER |
| เก | N | Y | Y | N | N | N | Y | N | N | Y | N | N | N | N | N | I-PER |
| ยุ | N | Y | Y | Y | N | N | N | N | N | N | N | Y | N | N | N | I-PER |
| รา | N | Y | Y | Y | N | N | Y | N | N | Y | Y | N | N | N | N | I-PER |
| พันธุ์ | N | Y | N | Y | N | N | Y | Y | N | N | N | N | N | N | N | E-PER |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| รมว. | N | N | N | N | Y | N | N | N | N | N | N | Y | Y | Y | N | X |

* รายละเอียดรายการชื่อเฉพาะ

1 - คำนำหน้าชื่อ
2 - ส่วนต้นชื่อบุคคล
3 - ส่วนกลางชื่อบุคคล
4 - ส่วนท้ายชื่อบุคคล
5 - ชื่อย่อองค์กร
6 - ส่วนต้นชื่อองค์กร
7 - ส่วนกลางชื่อองค์กร
8 - ส่วนท้ายชื่อองค์กร
9 - ส่วนต้นชื่อสถานที่
10 - ส่วนกลางชื่อสถานที่
11 - ส่วนท้ายชื่อสถานที่

** รูปแบบคำตอบของ CRF จะยึดตามรูปแบบคำตอบในคลังข้อมูลฝึกฝน

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| สาธารณสุข | N | N | N | N | N | N | Y | Y | N | N | N | Y | Y | N | N | B-ORG |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | Y | N | N | X |
| กล่าว | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ว่า | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ขณะนี้ | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| พบ | N | Y | Y | Y | N | N | N | N | N | Y | N | Y | N | N | N | X |
| ผู้ป่วย | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N | X |
| ที่ | N | N | N | N | N | N | Y | N | N | Y | N | Y | N | N | N | X |
| ติด | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N | X |
| เชื้อ | N | Y | Y | Y | N | N | N | N | N | N | N | Y | N | N | N | X |
| ไข้หวัด | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| นก | N | Y | N | Y | N | N | Y | N | N | Y | Y | Y | N | N | N | X |
| แล้ว | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| 3 | N | N | N | N | N | N | N | Y | N | N | Y | Y | N | Y | N | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ราย | N | N | N | N | N | N | Y | Y | N | Y | Y | Y | N | N | Y | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ซึ่ง | N | Y | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| เสีย | N | N | N | N | N | N | N | Y | N | N | N | Y | N | N | N | X |
| ชีวิต | N | Y | N | Y | N | N | Y | Y | N | N | N | Y | N | N | N | X |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ไป | N | N | N | N | N | N | Y | N | N | N | N | Y | N | N | N | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| 1 | N | N | N | N | N | N | N | Y | N | N | N | Y | N | Y | N | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ราย | N | N | N | N | N | N | Y | Y | N | Y | Y | Y | N | N | Y | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ที่ | N | N | N | N | N | N | Y | N | N | Y | N | Y | N | N | N | X |
| รพ. | N | N | N | N | Y | N | Y | N | N | N | N | Y | Y | Y | N | B-LOC |
| ศิ | N | Y | Y | N | N | Y | Y | N | N | Y | N | N | Y | N | N | I-LOC |
| ริ | N | Y | Y | Y | N | N | Y | Y | N | Y | Y | N | Y | N | N | I-LOC |
| ราช | N | Y | Y | Y | N | Y | Y | Y | N | Y | Y | Y | N | N | N | E-LOC |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| เป็น | N | Y | Y | N | N | N | Y | N | N | N | N | Y | N | N | N | X |
| เด็กชาย | Y | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| จาก | N | Y | N | N | N | N | Y | N | N | N | Y | Y | N | N | N | X |
| จังหวัด | N | N | N | N | N | N | Y | Y | N | N | Y | Y | N | N | N | B-LOC |
| กาญจนบุรี | N | N | N | N | N | N | N | Y | Y | Y | Y | N | N | Y | N | E-LOC |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ส่วน | N | Y | N | Y | N | N | Y | N | N | N | N | Y | N | N | N | X |
| อีก | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| 2 | N | N | N | N | N | N | Y | Y | N | N | Y | Y | N | Y | N | X |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ราย | N | N | N | N | N | N | Y | Y | N | Y | Y | Y | N | N | Y | X |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ได้แก่ | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| เด็กชาย | Y | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| ที่ | N | N | N | N | N | N | Y | N | N | Y | N | Y | N | N | N | X |
| จ. | Y | N | N | N | Y | N | N | N | N | Y | N | Y | Y | Y | N | B-LOC |
| สุพรรณ | N | Y | N | Y | N | N | Y | N | Y | Y | Y | N | Y | N | N | I-LOC |
| บุรี | N | Y | N | Y | N | N | Y | Y | Y | Y | Y | N | Y | N | N | E-LOC |
| <s> | N | N | N | N | N | N | N | N | N | N | N | Y | N | N | N | X |
| และ | N | Y | Y | Y | N | N | Y | N | N | Y | N | Y | N | N | N | X |
| จ. | Y | N | N | N | Y | N | N | N | N | Y | N | Y | Y | Y | N | B-LOC |
| สุ | N | Y | Y | Y | N | N | Y | N | Y | Y | N | N | Y | N | N | I-LOC |
| โข | N | Y | Y | Y | N | N | Y | N | N | Y | N | N | Y | N | N | I-LOC |
| ทัย | N | Y | N | Y | N | N | Y | Y | N | N | Y | N | N | Y | N | E-LOC |
| ... | | | | | | | | | | | | | | | | |

# ภาคผนวก ง
# ผลการทดสอบข้อมูลทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

## ผลการทดสอบโดยประเมินจากจำนวนชื่อ

ข้อมูลแบบตัดคำ

ตารางที่ ง-1 ค่าความแม่นยำของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| WSG | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 81.94 | 86.99 | 86.84 | 84.43 | 86.11 | 86.74 | 84.53 | 87.91 | 85.02 | 86.38 | 85.69 |
| ORG | 72.20 | 83.66 | 78.85 | 74.29 | 67.20 | 70.91 | 82.60 | 80.19 | 82.61 | 80.25 | 77.28 |
| LOC | 80.00 | 82.52 | 79.08 | 72.15 | 73.99 | 79.56 | 81.89 | 69.72 | 80.26 | 69.06 | 76.82 |
| ALL | 77.34 | 84.56 | 82.16 | 77.10 | 76.46 | 80.15 | 83.19 | 79.83 | 82.89 | 80.24 | 80.39 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 88.50 | 92.44 | 92.07 | 92.38 | 89.02 | 91.06 | 89.28 | 91.83 | 89.22 | 90.11 | 90.59 |
| ORG | 82.01 | 88.65 | 83.47 | 80.98 | 75.99 | 74.62 | 86.06 | 85.27 | 85.93 | 83.85 | 82.68 |
| LOC | 83.30 | 86.42 | 78.80 | 77.54 | 80.08 | 84.77 | 85.26 | 75.06 | 82.13 | 73.84 | 80.72 |
| ALL | 84.17 | 89.49 | 86.13 | 83.98 | 82.27 | 84.43 | 87.12 | 84.56 | 86.11 | 84.13 | 85.24 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 88.10 | 92.18 | 92.07 | 92.21 | 89.20 | 89.26 | 89.09 | 90.57 | 88.45 | 90.29 | 90.14 |
| O | 81.52 | 89.17 | 82.72 | 84.19 | 76.33 | 73.78 | 85.71 | 83.74 | 85.23 | 83.27 | 82.57 |
| L | 83.54 | 84.96 | 79.18 | 78.30 | 81.59 | 84.71 | 81.65 | 74.29 | 81.56 | 74.66 | 80.44 |
| LO | 80.12 | 87.01 | 83.55 | 76.14 | 80.39 | 77.32 | 87.15 | 82.42 | 83.56 | 76.67 | 81.43 |
| OL | 83.72 | 94.74 | 70.59 | 63.64 | 60.00 | 77.78 | 87.50 | 78.95 | 72.00 | 65.00 | 75.39 |
| ALL | 83.80 | 89.35 | 85.94 | 84.31 | 82.67 | 83.59 | 86.28 | 83.46 | 85.17 | 83.66 | 84.82 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 88.86 | 94.25 | 92.51 | 92.56 | 92.98 | 92.96 | 90.42 | 93.47 | 90.16 | 92.36 | 92.05 |
| ORG | 81.42 | 88.89 | 82.67 | 80.46 | 72.22 | 74.29 | 86.77 | 86.00 | 86.63 | 82.57 | 82.19 |
| LOC | 85.10 | 85.71 | 77.74 | 76.91 | 76.33 | 83.48 | 82.01 | 73.46 | 83.56 | 74.92 | 79.92 |
| ALL | 84.60 | 90.02 | 85.75 | 83.59 | 81.27 | 84.81 | 87.10 | 84.88 | 87.10 | 84.61 | 85.37 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 88.41 | 93.56 | 91.93 | 92.06 | 92.18 | 92.45 | 90.42 | 92.76 | 89.17 | 92.21 | 91.52 |
| O | 82.47 | 89.58 | 79.57 | 83.72 | 74.54 | 74.14 | 86.13 | 86.46 | 86.17 | 83.96 | 82.67 |
| L | 84.82 | 86.22 | 76.90 | 75.41 | 79.11 | 82.19 | 81.07 | 75.26 | 81.92 | 77.54 | 80.04 |
| LO | 81.55 | 87.01 | 84.46 | 77.91 | 79.09 | 74.00 | 87.78 | 82.11 | 82.67 | 73.77 | 81.04 |
| OL | 83.72 | 94.44 | 72.22 | 72.22 | 51.61 | 86.96 | 82.35 | 78.26 | 83.33 | 68.42 | 77.35 |
| ALL | 84.69 | 90.20 | 84.59 | 83.64 | 82.36 | 84.25 | 86.78 | 85.20 | 85.98 | 85.19 | 85.29 |

ตารางที่ ง-2 ค่าความครบถ้วนของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| WSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 79.24 | 87.62 | 80.93 | 85.46 | 79.91 | 84.96 | 78.96 | 86.74 | 88.16 | 87.32 | 83.93 |
| ORG | 68.38 | 78.06 | 72.64 | 67.20 | 59.26 | 65.31 | 76.26 | 70.44 | 76.84 | 71.09 | 70.55 |
| LOC | 67.02 | 78.38 | 75.41 | 64.90 | 70.51 | 70.76 | 73.12 | 61.29 | 75.50 | 68.83 | 70.57 |
| ALL | 70.58 | 81.23 | 76.59 | 72.20 | 70.45 | 75.04 | 76.55 | 72.93 | 80.45 | 76.60 | 75.26 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 83.80 | 91.99 | 83.79 | 89.14 | 81.41 | 87.01 | 81.77 | 87.31 | 88.35 | 89.13 | 86.37 |
| ORG | 74.64 | 82.41 | 73.78 | 70.97 | 64.20 | 69.62 | 79.33 | 77.01 | 81.99 | 72.19 | 74.61 |
| LOC | 67.02 | 81.08 | 68.03 | 60.91 | 70.86 | 72.48 | 73.80 | 64.31 | 76.24 | 72.40 | 70.71 |
| ALL | 74.17 | 85.25 | 76.70 | 73.72 | 72.66 | 77.66 | 78.96 | 76.48 | 82.64 | 78.47 | 77.67 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 84.30 | 91.50 | 83.79 | 89.14 | 80.51 | 86.67 | 81.35 | 87.31 | 87.77 | 89.31 | 86.17 |
| O | 76.67 | 85.26 | 72.40 | 71.25 | 67.21 | 65.53 | 78.67 | 77.69 | 82.06 | 74.73 | 75.15 |
| L | 68.32 | 85.71 | 67.64 | 63.43 | 74.57 | 75.78 | 76.56 | 63.80 | 77.53 | 79.27 | 73.26 |
| LO | 65.52 | 70.53 | 75.15 | 70.16 | 58.57 | 78.13 | 76.10 | 60.00 | 70.11 | 48.94 | 67.32 |
| OL | 52.94 | 51.43 | 52.17 | 43.75 | 42.00 | 37.50 | 36.84 | 34.88 | 46.15 | 39.39 | 43.71 |
| ALL | 73.48 | 85.33 | 76.14 | 74.10 | 72.88 | 76.60 | 78.10 | 74.46 | 81.27 | 78.53 | 77.09 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 82.78 | 91.50 | 82.43 | 89.32 | 83.36 | 88.03 | 82.05 | 86.74 | 88.93 | 89.86 | 86.50 |
| ORG | 73.84 | 83.08 | 74.50 | 70.83 | 61.90 | 68.42 | 80.59 | 76.68 | 82.17 | 70.31 | 74.23 |
| LOC | 68.76 | 81.08 | 65.85 | 63.19 | 70.33 | 71.99 | 70.62 | 62.50 | 76.73 | 76.62 | 70.77 |
| ALL | 74.23 | 85.41 | 75.97 | 74.38 | 72.50 | 77.59 | 78.80 | 75.63 | 83.05 | 78.80 | 77.64 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 83.04 | 91.75 | 82.29 | 89.32 | 83.06 | 87.86 | 82.05 | 87.31 | 87.96 | 90.04 | 86.47 |
| O | 76.19 | 85.66 | 70.70 | 71.61 | 66.51 | 66.77 | 80.23 | 79.13 | 83.15 | 73.81 | 75.38 |
| L | 69.70 | 86.61 | 65.01 | 65.05 | 74.57 | 74.93 | 75.81 | 64.46 | 76.99 | 80.36 | 73.35 |
| LO | 67.49 | 70.53 | 73.96 | 70.16 | 62.14 | 77.08 | 77.07 | 62.40 | 71.26 | 47.87 | 68.00 |
| OL | 52.94 | 48.57 | 56.52 | 40.63 | 32.00 | 35.71 | 36.84 | 41.86 | 51.28 | 39.39 | 43.57 |
| ALL | 73.73 | 85.65 | 74.47 | 74.65 | 73.66 | 77.02 | 78.75 | 75.44 | 81.75 | 78.60 | 77.37 |

ตารางที่ ง-3 ค่า F-measure ของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| WSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 80.57 | 87.30 | 83.78 | 84.94 | 82.89 | 85.84 | 81.65 | 87.32 | 86.56 | 86.85 | 84.77 |
| ORG | 70.24 | 80.76 | 75.62 | 70.57 | 62.98 | 68.00 | 79.30 | 75.00 | 79.62 | 75.39 | 73.75 |
| LOC | 72.93 | 80.40 | 77.20 | 68.33 | 72.21 | 74.90 | 77.26 | 65.24 | 77.81 | 68.94 | 73.52 |
| ALL | 73.81 | 82.86 | 79.27 | 74.57 | 73.33 | 77.51 | 79.73 | 76.22 | 81.65 | 78.38 | 77.73 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 86.09 | 92.21 | 87.73 | 90.73 | 85.04 | 88.99 | 85.36 | 89.51 | 88.78 | 89.62 | 88.41 |
| ORG | 78.15 | 85.42 | 78.33 | 75.64 | 69.60 | 72.03 | 82.56 | 80.93 | 83.91 | 77.58 | 78.42 |
| LOC | 74.27 | 83.67 | 73.02 | 68.23 | 75.19 | 78.15 | 79.12 | 69.27 | 79.08 | 73.11 | 75.31 |
| ALL | 78.85 | 87.32 | 81.14 | 78.52 | 77.17 | 80.90 | 82.84 | 80.32 | 84.34 | 81.20 | 81.26 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 86.16 | 91.84 | 87.73 | 90.65 | 84.63 | 87.94 | 85.04 | 88.91 | 88.11 | 89.80 | 88.08 |
| O | 79.02 | 87.17 | 77.22 | 77.18 | 71.48 | 69.41 | 82.04 | 80.60 | 83.61 | 78.76 | 78.65 |
| L | 75.16 | 85.33 | 72.96 | 70.09 | 77.92 | 80.00 | 79.02 | 68.65 | 79.49 | 76.90 | 76.55 |
| LO | 72.09 | 77.91 | 79.13 | 73.02 | 67.77 | 77.72 | 81.25 | 69.44 | 76.25 | 59.74 | 73.43 |
| OL | 64.86 | 66.67 | 60.00 | 51.85 | 49.41 | 50.60 | 51.85 | 48.39 | 56.25 | 49.06 | 54.89 |
| ALL | 78.30 | 87.29 | 80.74 | 78.88 | 77.47 | 79.94 | 81.99 | 78.71 | 83.18 | 81.02 | 80.75 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 85.71 | 92.86 | 87.18 | 90.91 | 87.91 | 90.43 | 86.03 | 89.98 | 89.54 | 91.09 | 89.16 |
| ORG | 77.44 | 85.89 | 78.37 | 75.34 | 66.67 | 71.23 | 83.56 | 81.08 | 84.34 | 75.95 | 77.99 |
| LOC | 76.06 | 83.33 | 71.30 | 69.38 | 73.21 | 77.31 | 75.89 | 67.54 | 80.00 | 75.76 | 74.98 |
| ALL | 79.08 | 87.66 | 80.57 | 78.71 | 76.63 | 81.04 | 82.74 | 79.99 | 85.02 | 81.60 | 81.30 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 85.64 | 92.65 | 86.84 | 90.67 | 87.38 | 90.10 | 86.03 | 89.95 | 88.56 | 91.11 | 88.89 |
| O | 79.21 | 87.58 | 74.87 | 77.19 | 70.30 | 70.26 | 83.08 | 82.63 | 84.63 | 78.56 | 78.83 |
| L | 76.52 | 86.41 | 70.46 | 69.85 | 76.77 | 78.39 | 78.35 | 69.44 | 79.38 | 78.93 | 76.45 |
| LO | 73.85 | 77.91 | 78.86 | 73.83 | 69.60 | 75.51 | 82.08 | 70.91 | 76.54 | 58.06 | 73.72 |
| OL | 64.86 | 64.15 | 63.41 | 52.00 | 39.51 | 50.63 | 50.91 | 54.55 | 63.49 | 50.00 | 55.35 |
| ALL | 78.83 | 87.86 | 79.21 | 78.89 | 77.77 | 80.47 | 82.57 | 80.03 | 83.81 | 81.76 | 81.12 |

ข้อมูลแบบตัดพยางค์

ตารางที่ ง-4 ค่าความแม่นยำของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| SSG | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 84.44 | 87.74 | 89.44 | 86.25 | 83.82 | 88.50 | 87.30 | 87.14 | 85.66 | 89.24 | 86.95 |
| ORG | 75.16 | 83.63 | 76.14 | 73.44 | 63.85 | 68.78 | 81.53 | 81.67 | 81.57 | 79.23 | 76.50 |
| LOC | 78.01 | 77.13 | 77.23 | 73.88 | 75.47 | 79.89 | 75.38 | 70.52 | 79.09 | 68.93 | 75.55 |
| ALL | 78.40 | 83.66 | 81.63 | 77.70 | 74.91 | 80.26 | 82.26 | 80.41 | 82.44 | 80.79 | 80.25 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 90.14 | 92.67 | 92.34 | 91.51 | 89.15 | 91.76 | 91.19 | 88.93 | 89.57 | 91.25 | 90.85 |
| ORG | 83.39 | 88.67 | 83.93 | 80.99 | 72.05 | 74.28 | 86.19 | 81.87 | 86.73 | 82.45 | 82.06 |
| LOC | 83.61 | 82.21 | 81.27 | 78.27 | 79.75 | 84.43 | 83.81 | 74.38 | 81.55 | 76.71 | 80.60 |
| ALL | 85.11 | 88.67 | 86.88 | 83.81 | 80.97 | 84.60 | 87.52 | 82.29 | 86.38 | 84.64 | 85.09 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 90.26 | 92.67 | 92.92 | 91.65 | 89.35 | 91.77 | 90.73 | 87.91 | 89.61 | 90.72 | 90.76 |
| O | 84.71 | 89.57 | 82.24 | 82.68 | 74.25 | 74.73 | 87.21 | 83.04 | 87.67 | 83.93 | 83.00 |
| L | 83.90 | 84.72 | 79.40 | 80.10 | 81.25 | 83.39 | 83.33 | 73.23 | 83.24 | 78.11 | 81.07 |
| LO | 79.88 | 77.53 | 83.69 | 75.53 | 64.81 | 76.92 | 90.29 | 82.65 | 80.77 | 74.32 | 78.64 |
| OL | 89.74 | 63.16 | 55.56 | 58.82 | 50.00 | 70.97 | 85.71 | 71.43 | 73.33 | 41.18 | 65.99 |
| ALL | 85.41 | 88.40 | 85.89 | 83.93 | 80.84 | 84.42 | 88.01 | 82.04 | 86.60 | 83.90 | 84.94 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 91.20 | 93.86 | 95.28 | 93.89 | 90.77 | 93.73 | 92.71 | 90.94 | 90.08 | 93.01 | 92.55 |
| ORG | 83.48 | 87.54 | 82.11 | 81.41 | 71.61 | 75.19 | 84.21 | 83.69 | 87.43 | 81.69 | 81.84 |
| LOC | 85.47 | 82.06 | 74.46 | 78.37 | 78.79 | 84.26 | 82.22 | 74.75 | 81.22 | 75.32 | 79.69 |
| ALL | 86.05 | 88.46 | 85.85 | 84.69 | 81.12 | 85.64 | 86.89 | 83.77 | 86.70 | 84.56 | 85.37 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 91.79 | 93.17 | 94.83 | 94.04 | 89.98 | 93.04 | 92.51 | 90.57 | 89.88 | 92.80 | 92.26 |
| O | 83.78 | 89.24 | 81.51 | 82.77 | 74.39 | 74.39 | 84.70 | 83.98 | 87.99 | 85.21 | 82.80 |
| L | 85.78 | 85.04 | 75.00 | 82.02 | 80.04 | 82.05 | 81.62 | 77.07 | 82.35 | 80.22 | 81.12 |
| LO | 80.12 | 78.82 | 82.73 | 78.19 | 65.14 | 77.17 | 89.66 | 80.81 | 77.92 | 70.67 | 78.12 |
| OL | 86.11 | 76.19 | 68.75 | 60.00 | 45.00 | 71.43 | 86.36 | 84.21 | 72.41 | 42.42 | 69.29 |
| ALL | 86.00 | 88.81 | 85.66 | 85.45 | 80.57 | 84.54 | 87.45 | 84.36 | 86.41 | 85.31 | 85.46 |

ตารางที่ ง-5 ค่าความครบถ้วนของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| SSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 76.96 | 88.59 | 78.47 | 84.59 | 76.91 | 85.47 | 78.12 | 85.98 | 88.16 | 87.14 | 83.04 |
| ORG | 73.84 | 78.73 | 69.48 | 69.49 | 58.55 | 64.83 | 75.84 | 73.89 | 78.13 | 70.31 | 71.31 |
| LOC | 65.62 | 76.83 | 73.22 | 65.46 | 70.33 | 68.30 | 68.34 | 60.28 | 73.02 | 69.16 | 69.06 |
| ALL | 71.65 | 81.55 | 73.92 | 73.02 | 69.06 | 74.40 | 74.95 | 73.67 | 80.25 | 76.27 | 74.87 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 78.73 | 91.99 | 80.52 | 86.87 | 80.06 | 87.52 | 79.80 | 86.74 | 88.35 | 88.77 | 84.94 |
| ORG | 78.17 | 82.58 | 74.07 | 74.46 | 61.38 | 67.70 | 79.33 | 76.35 | 82.90 | 72.66 | 74.96 |
| LOC | 71.20 | 80.31 | 66.39 | 60.15 | 68.06 | 69.29 | 73.12 | 60.28 | 75.50 | 72.73 | 69.70 |
| ALL | 75.80 | 85.17 | 75.14 | 74.21 | 70.39 | 76.38 | 78.05 | 74.83 | 82.78 | 78.60 | 77.14 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 79.75 | 91.99 | 80.52 | 86.51 | 80.51 | 87.69 | 79.66 | 86.74 | 88.74 | 88.59 | 85.07 |
| O | 83.10 | 83.86 | 70.89 | 75.95 | 63.47 | 64.29 | 80.04 | 78.93 | 84.03 | 72.71 | 75.73 |
| L | 73.27 | 86.61 | 69.68 | 64.24 | 72.08 | 74.36 | 74.81 | 61.59 | 77.53 | 75.27 | 72.94 |
| LO | 64.53 | 72.63 | 69.82 | 74.35 | 50.00 | 72.92 | 77.07 | 64.80 | 72.41 | 58.51 | 67.70 |
| OL | 51.47 | 34.29 | 43.48 | 31.25 | 38.00 | 39.29 | 47.37 | 34.88 | 56.41 | 42.42 | 41.89 |
| ALL | 75.42 | 84.78 | 74.14 | 75.14 | 70.50 | 76.10 | 77.78 | 74.40 | 82.64 | 77.47 | 76.84 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 78.73 | 92.72 | 82.56 | 88.79 | 81.11 | 89.40 | 80.22 | 87.50 | 88.16 | 89.13 | 85.83 |
| ORG | 77.05 | 82.41 | 72.35 | 74.73 | 60.49 | 69.62 | 78.21 | 76.68 | 81.80 | 72.50 | 74.58 |
| LOC | 70.86 | 83.01 | 66.12 | 63.95 | 70.68 | 71.01 | 72.67 | 60.28 | 75.99 | 75.32 | 70.99 |
| ALL | 75.24 | 85.88 | 75.25 | 76.00 | 71.33 | 78.23 | 77.68 | 75.20 | 82.43 | 79.20 | 77.64 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 79.24 | 92.72 | 82.43 | 88.44 | 80.81 | 89.06 | 79.66 | 87.31 | 87.96 | 88.77 | 85.64 |
| O | 81.19 | 84.26 | 71.64 | 74.68 | 64.64 | 66.77 | 79.06 | 80.17 | 83.37 | 74.91 | 76.07 |
| L | 74.06 | 88.84 | 66.47 | 67.27 | 72.08 | 72.93 | 75.31 | 61.59 | 76.71 | 79.64 | 73.49 |
| LO | 65.52 | 70.53 | 68.05 | 76.96 | 50.71 | 73.96 | 76.10 | 64.00 | 68.97 | 56.38 | 67.12 |
| OL | 45.59 | 45.71 | 47.83 | 37.50 | 36.00 | 35.71 | 50.00 | 37.21 | 53.85 | 42.42 | 43.18 |
| ALL | 74.92 | 85.73 | 74.42 | 76.55 | 70.89 | 76.81 | 77.57 | 74.95 | 81.68 | 79.00 | 77.25 |

ตารางที่ ง-6 ค่า F-measure ของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| SSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 80.53 | 88.16 | 83.60 | 85.41 | 80.22 | 86.96 | 82.46 | 86.56 | 86.89 | 88.18 | 84.90 |
| ORG | 74.49 | 81.10 | 72.66 | 71.41 | 61.09 | 66.75 | 78.58 | 77.59 | 79.81 | 74.50 | 73.80 |
| LOC | 71.28 | 76.98 | 75.18 | 69.42 | 72.81 | 73.64 | 71.68 | 65.00 | 75.93 | 69.04 | 72.10 |
| ALL | 74.88 | 82.59 | 77.58 | 75.29 | 71.87 | 77.22 | 78.43 | 76.89 | 81.33 | 78.46 | 77.45 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 84.05 | 92.33 | 86.03 | 89.13 | 84.36 | 89.59 | 85.12 | 87.82 | 88.95 | 89.99 | 87.74 |
| ORG | 80.70 | 85.52 | 78.69 | 77.59 | 66.29 | 70.84 | 82.62 | 79.01 | 84.77 | 77.24 | 78.33 |
| LOC | 76.91 | 81.25 | 73.08 | 68.03 | 73.45 | 76.11 | 78.10 | 66.59 | 78.41 | 74.67 | 74.66 |
| ALL | 80.19 | 86.89 | 80.58 | 78.72 | 75.31 | 80.28 | 82.51 | 78.38 | 84.54 | 81.51 | 80.89 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 84.68 | 92.33 | 86.28 | 89.01 | 84.70 | 89.69 | 84.84 | 87.32 | 89.17 | 89.64 | 87.77 |
| O | 83.89 | 86.63 | 76.14 | 79.17 | 68.43 | 69.12 | 83.47 | 80.93 | 85.81 | 77.92 | 79.15 |
| L | 78.22 | 85.65 | 74.22 | 71.30 | 76.39 | 78.61 | 78.84 | 66.91 | 80.28 | 76.67 | 76.71 |
| LO | 71.39 | 75.00 | 76.13 | 74.93 | 56.45 | 74.87 | 83.16 | 72.65 | 76.36 | 65.48 | 72.64 |
| OL | 65.42 | 44.44 | 48.78 | 40.82 | 43.18 | 50.57 | 61.02 | 46.88 | 63.77 | 41.79 | 50.67 |
| ALL | 80.11 | 86.55 | 79.58 | 79.29 | 75.32 | 80.04 | 82.58 | 78.03 | 84.58 | 80.55 | 80.66 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 84.51 | 93.28 | 88.47 | 91.27 | 85.67 | 91.51 | 86.02 | 89.19 | 89.11 | 91.03 | 89.01 |
| ORG | 80.13 | 84.90 | 76.92 | 77.93 | 65.58 | 72.30 | 81.10 | 80.03 | 84.52 | 76.82 | 78.02 |
| LOC | 77.48 | 82.53 | 70.04 | 70.43 | 74.52 | 77.07 | 77.15 | 66.74 | 78.52 | 75.32 | 74.98 |
| ALL | 80.28 | 87.15 | 80.20 | 80.11 | 75.91 | 81.76 | 82.02 | 79.25 | 84.51 | 81.79 | 81.30 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 85.05 | 92.94 | 88.19 | 91.16 | 85.15 | 91.00 | 85.61 | 88.91 | 88.91 | 90.74 | 88.77 |
| O | 82.47 | 86.68 | 76.26 | 78.52 | 69.17 | 70.38 | 81.78 | 82.03 | 85.62 | 79.73 | 79.26 |
| L | 79.49 | 86.90 | 70.48 | 73.92 | 75.86 | 77.22 | 78.34 | 68.47 | 79.43 | 79.93 | 77.00 |
| LO | 72.09 | 74.44 | 74.68 | 77.57 | 57.03 | 75.53 | 82.32 | 71.43 | 73.17 | 62.72 | 72.10 |
| OL | 59.62 | 57.14 | 56.41 | 46.15 | 40.00 | 47.62 | 63.33 | 51.61 | 61.76 | 42.42 | 52.61 |
| ALL | 80.08 | 87.24 | 79.64 | 80.76 | 75.42 | 80.49 | 82.21 | 79.38 | 83.98 | 82.04 | 81.12 |

## ผลการทดสอบโดยประเมินจากจำนวน token

ข้อมูลแบบตัดคำ

ตารางที่ ง-7 ค่าความแม่นยำของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| WSG (token) | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 93.01 | 94.16 | 94.62 | 93.79 | 94.80 | 94.55 | 93.26 | 92.36 | 93.11 | 90.43 | 93.41 |
| ORG | 78.15 | 87.63 | 85.08 | 79.31 | 72.71 | 76.77 | 85.01 | 80.88 | 86.32 | 82.05 | 81.39 |
| LOC | 88.06 | 85.86 | 86.74 | 83.76 | 81.17 | 86.04 | 90.99 | 86.03 | 88.46 | 75.21 | 85.23 |
| ALL | 86.76 | 90.56 | 90.33 | 86.86 | 85.32 | 88.47 | 90.61 | 87.35 | 90.34 | 84.99 | 88.16 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 88.17 | 93.23 | 95.64 | 96.61 | 95.36 | 96.57 | 95.29 | 94.21 | 94.21 | 90.79 | 94.01 |
| ORG | 80.42 | 89.87 | 84.31 | 78.38 | 74.78 | 76.91 | 83.95 | 78.31 | 85.93 | 81.78 | 81.46 |
| LOC | 89.03 | 86.74 | 86.12 | 82.00 | 84.80 | 87.84 | 91.88 | 86.83 | 87.62 | 75.63 | 85.85 |
| ALL | 85.87 | 91.06 | 90.54 | 87.28 | 86.90 | 89.92 | 91.45 | 87.29 | 90.53 | 85.23 | 88.61 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 88.56 | 94.55 | 95.34 | 96.29 | 95.32 | 96.07 | 95.16 | 92.49 | 93.64 | 91.02 | 93.84 |
| O | 78.60 | 87.44 | 82.08 | 82.66 | 73.15 | 75.13 | 80.81 | 75.59 | 85.59 | 81.21 | 80.22 |
| L | 88.78 | 85.28 | 84.32 | 82.45 | 87.12 | 89.50 | 90.53 | 88.18 | 87.04 | 75.46 | 85.87 |
| LO | 77.21 | 88.18 | 88.46 | 72.40 | 75.62 | 80.43 | 87.95 | 83.24 | 81.82 | 80.42 | 81.57 |
| OL | 88.89 | 95.89 | 65.22 | 65.22 | 47.90 | 76.29 | 100.00 | 78.38 | 65.69 | 54.41 | 73.79 |
| ALL | 85.28 | 90.82 | 89.62 | 87.75 | 86.06 | 89.60 | 90.81 | 86.21 | 89.44 | 84.70 | 88.03 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 86.86 | 94.96 | 96.29 | 95.94 | 96.42 | 96.67 | 95.65 | 95.45 | 94.65 | 92.73 | 94.56 |
| ORG | 82.20 | 90.38 | 83.03 | 80.06 | 74.83 | 76.06 | 85.73 | 77.98 | 87.96 | 83.75 | 82.20 |
| LOC | 89.64 | 86.18 | 88.72 | 83.64 | 82.97 | 87.93 | 92.59 | 86.51 | 88.76 | 75.09 | 86.20 |
| ALL | 86.08 | 91.95 | 90.82 | 87.95 | 86.85 | 89.76 | 92.29 | 87.59 | 91.57 | 86.55 | 89.14 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 86.56 | 93.78 | 95.69 | 95.33 | 96.04 | 96.54 | 96.17 | 94.74 | 94.04 | 92.81 | 94.17 |
| O | 82.22 | 89.71 | 77.36 | 83.08 | 73.75 | 75.36 | 84.14 | 78.43 | 86.89 | 84.61 | 81.55 |
| L | 89.47 | 86.48 | 86.38 | 83.36 | 85.78 | 87.80 | 91.80 | 87.18 | 88.02 | 80.03 | 86.63 |
| LO | 80.65 | 85.09 | 89.06 | 78.37 | 80.07 | 79.46 | 87.76 | 83.33 | 83.21 | 81.76 | 82.87 |
| OL | 86.57 | 94.29 | 79.41 | 74.29 | 59.06 | 85.37 | 83.87 | 74.71 | 76.09 | 55.07 | 76.87 |
| ALL | 85.73 | 91.24 | 89.01 | 88.27 | 86.98 | 89.63 | 92.12 | 87.64 | 90.45 | 87.44 | 88.85 |

ตารางที่ ง-8 ค่าความครบถ้วนของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| WSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (token) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 93.12 | 96.50 | 92.20 | 94.85 | 88.65 | 94.92 | 88.91 | 91.63 | 94.41 | 94.29 | 92.95 |
| ORG | 79.30 | 79.75 | 75.97 | 69.82 | 62.79 | 70.88 | 74.35 | 70.56 | 72.45 | 75.03 | 73.09 |
| LOC | 69.80 | 82.49 | 79.01 | 79.26 | 78.17 | 80.09 | 79.78 | 68.11 | 79.16 | 73.68 | 76.96 |
| ALL | 81.73 | 88.01 | 84.62 | 82.63 | 78.81 | 85.52 | 82.95 | 78.24 | 84.58 | 83.88 | 83.10 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 94.13 | 97.23 | 92.51 | 95.44 | 89.09 | 95.09 | 90.23 | 91.34 | 94.73 | 94.71 | 93.45 |
| ORG | 80.84 | 81.66 | 74.50 | 70.55 | 70.27 | 71.72 | 77.64 | 75.81 | 76.41 | 74.52 | 75.39 |
| LOC | 70.10 | 81.34 | 75.54 | 75.37 | 77.50 | 78.32 | 79.03 | 65.65 | 77.61 | 71.63 | 75.21 |
| ALL | 82.69 | 88.88 | 83.68 | 82.19 | 80.83 | 85.39 | 84.38 | 78.93 | 85.46 | 83.53 | 83.60 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 94.13 | 96.95 | 92.69 | 95.52 | 88.32 | 95.27 | 89.72 | 91.49 | 94.59 | 94.06 | 93.27 |
| O | 84.22 | 84.44 | 74.51 | 72.31 | 68.60 | 68.45 | 74.18 | 77.36 | 75.09 | 76.47 | 75.56 |
| L | 71.20 | 86.86 | 74.64 | 80.05 | 80.05 | 81.72 | 83.55 | 66.25 | 79.48 | 79.55 | 78.34 |
| LO | 58.25 | 61.01 | 64.72 | 65.59 | 66.58 | 77.89 | 79.56 | 58.33 | 60.34 | 48.12 | 64.04 |
| OL | 50.91 | 56.45 | 47.37 | 49.18 | 47.06 | 42.77 | 39.13 | 29.74 | 48.20 | 32.74 | 44.36 |
| ALL | 81.37 | 88.69 | 82.37 | 82.87 | 79.67 | 84.39 | 83.38 | 77.24 | 83.94 | 82.69 | 82.66 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 94.07 | 95.77 | 92.09 | 94.93 | 88.95 | 95.40 | 89.99 | 90.75 | 94.96 | 94.11 | 93.10 |
| ORG | 80.03 | 81.16 | 75.58 | 71.02 | 69.33 | 72.00 | 78.39 | 75.08 | 76.96 | 72.81 | 75.24 |
| LOC | 70.85 | 82.65 | 76.34 | 77.43 | 78.59 | 78.32 | 78.95 | 65.52 | 77.97 | 76.08 | 76.27 |
| ALL | 82.64 | 88.22 | 83.97 | 82.63 | 80.81 | 85.61 | 84.45 | 78.44 | 85.82 | 83.46 | 83.61 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 93.95 | 96.22 | 92.13 | 94.93 | 89.06 | 95.40 | 89.93 | 91.34 | 94.82 | 94.57 | 93.23 |
| O | 84.59 | 84.20 | 74.95 | 71.37 | 70.07 | 71.07 | 75.54 | 78.81 | 77.03 | 75.95 | 76.36 |
| L | 72.64 | 86.65 | 75.52 | 81.74 | 81.14 | 82.61 | 83.18 | 67.95 | 79.27 | 79.69 | 79.04 |
| LO | 61.40 | 61.01 | 63.42 | 65.00 | 64.40 | 77.37 | 80.11 | 60.61 | 63.69 | 50.63 | 64.76 |
| OL | 52.73 | 53.23 | 56.84 | 42.62 | 44.12 | 40.46 | 37.68 | 33.33 | 50.36 | 33.63 | 44.50 |
| ALL | 82.05 | 88.14 | 82.38 | 82.56 | 80.36 | 85.05 | 83.71 | 78.24 | 84.67 | 82.96 | 83.01 |

ตารางที่ ง-9 ค่า F-measure ของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| WSG (token) | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 93.06 | 95.32 | 93.40 | 94.32 | 91.62 | 94.73 | 91.03 | 92.00 | 93.76 | 92.32 | 93.16 |
| ORG | 78.72 | 83.50 | 80.26 | 74.27 | 67.39 | 73.71 | 79.32 | 75.37 | 78.78 | 78.38 | 76.97 |
| LOC | 77.87 | 84.14 | 82.69 | 81.45 | 79.64 | 82.96 | 85.02 | 76.03 | 83.55 | 74.44 | 80.78 |
| ALL | 84.17 | 89.27 | 87.38 | 84.69 | 81.93 | 86.97 | 86.61 | 82.54 | 87.36 | 84.43 | 85.54 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 91.05 | 95.19 | 94.05 | 96.02 | 92.12 | 95.83 | 92.69 | 92.75 | 94.47 | 92.71 | 93.69 |
| ORG | 80.63 | 85.57 | 79.10 | 74.26 | 72.46 | 74.22 | 80.67 | 77.04 | 80.89 | 77.98 | 78.28 |
| LOC | 78.44 | 83.95 | 80.49 | 78.54 | 80.99 | 82.81 | 84.97 | 74.77 | 82.31 | 73.58 | 80.08 |
| ALL | 84.25 | 89.96 | 86.98 | 84.66 | 83.76 | 87.60 | 87.78 | 82.90 | 87.93 | 84.37 | 86.02 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 91.26 | 95.73 | 93.99 | 95.90 | 91.68 | 95.67 | 92.36 | 91.98 | 94.11 | 92.51 | 93.52 |
| O | 81.31 | 85.91 | 78.11 | 77.14 | 70.80 | 71.63 | 77.35 | 76.46 | 80.00 | 78.77 | 77.75 |
| L | 79.02 | 86.06 | 79.19 | 81.23 | 83.44 | 85.43 | 86.90 | 75.66 | 83.09 | 77.45 | 81.75 |
| LO | 66.40 | 72.12 | 74.75 | 68.83 | 70.81 | 79.14 | 83.55 | 68.60 | 69.45 | 60.21 | 71.39 |
| OL | 64.74 | 71.07 | 54.88 | 56.07 | 47.48 | 54.81 | 56.25 | 43.12 | 55.60 | 40.88 | 54.49 |
| ALL | 83.28 | 89.75 | 85.84 | 85.24 | 82.74 | 86.92 | 86.94 | 81.48 | 86.60 | 83.68 | 85.25 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 90.32 | 95.36 | 94.14 | 95.43 | 92.53 | 96.03 | 92.74 | 93.04 | 94.81 | 93.41 | 93.78 |
| ORG | 81.10 | 85.52 | 79.13 | 75.27 | 71.97 | 73.98 | 81.89 | 76.51 | 82.09 | 77.90 | 78.54 |
| LOC | 79.14 | 84.38 | 82.06 | 80.41 | 80.72 | 82.85 | 85.23 | 74.57 | 83.02 | 75.58 | 80.80 |
| ALL | 84.33 | 90.05 | 87.26 | 85.21 | 83.72 | 87.64 | 88.20 | 82.76 | 88.60 | 84.98 | 86.27 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 90.10 | 94.98 | 93.88 | 95.13 | 92.42 | 95.97 | 92.94 | 93.01 | 94.43 | 93.68 | 93.65 |
| O | 83.39 | 86.87 | 76.13 | 76.78 | 71.87 | 73.15 | 79.61 | 78.62 | 81.66 | 80.05 | 78.81 |
| L | 80.18 | 86.56 | 80.58 | 82.54 | 83.39 | 85.12 | 87.28 | 76.37 | 83.42 | 79.86 | 82.53 |
| LO | 69.72 | 71.06 | 74.08 | 71.06 | 71.39 | 78.40 | 83.76 | 70.18 | 72.15 | 62.53 | 72.43 |
| OL | 65.54 | 68.04 | 66.26 | 54.17 | 50.51 | 54.90 | 52.00 | 46.10 | 60.61 | 41.76 | 55.99 |
| ALL | 83.85 | 89.66 | 85.57 | 85.32 | 83.54 | 87.28 | 87.71 | 82.67 | 87.46 | 85.14 | 85.82 |

ข้อมูลแบบตัดพยางค์

ตารางที่ ง-10 ค่าความแม่นยำของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| SSG (token) | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 94.82 | 95.13 | 96.99 | 95.38 | 94.85 | 96.93 | 97.01 | 92.37 | 93.53 | 92.95 | 94.99 |
| ORG | 80.76 | 89.71 | 82.64 | 80.55 | 73.35 | 78.81 | 85.87 | 84.44 | 84.69 | 81.76 | 82.26 |
| LOC | 87.02 | 85.76 | 82.01 | 83.53 | 78.90 | 87.05 | 87.82 | 81.76 | 87.14 | 77.28 | 83.83 |
| ALL | 87.30 | 91.35 | 89.26 | 86.80 | 83.95 | 89.75 | 91.37 | 87.11 | 89.21 | 85.89 | 88.20 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 94.97 | 94.14 | 96.96 | 96.60 | 95.31 | 97.70 | 97.14 | 91.41 | 95.03 | 92.88 | 95.21 |
| ORG | 84.85 | 90.93 | 83.75 | 83.24 | 72.16 | 79.20 | 85.90 | 80.01 | 86.43 | 81.70 | 82.82 |
| LOC | 88.75 | 85.92 | 88.58 | 84.84 | 81.09 | 91.09 | 90.35 | 85.63 | 88.72 | 78.04 | 86.30 |
| ALL | 89.46 | 91.49 | 90.84 | 88.66 | 84.25 | 91.01 | 91.94 | 85.98 | 90.74 | 86.10 | 89.05 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 95.18 | 93.69 | 96.84 | 96.93 | 95.09 | 97.74 | 96.65 | 90.77 | 94.37 | 92.10 | 94.94 |
| O | 84.16 | 90.39 | 81.30 | 85.34 | 75.25 | 79.29 | 84.51 | 77.78 | 87.19 | 82.43 | 82.76 |
| L | 88.93 | 87.37 | 84.83 | 86.22 | 86.17 | 89.98 | 91.41 | 84.95 | 89.10 | 80.25 | 86.92 |
| LO | 78.60 | 77.89 | 86.16 | 72.07 | 70.00 | 79.23 | 91.10 | 87.00 | 82.33 | 75.67 | 80.00 |
| OL | 91.16 | 78.81 | 46.32 | 57.69 | 31.89 | 69.08 | 94.59 | 72.97 | 70.43 | 43.40 | 65.64 |
| ALL | 89.12 | 90.70 | 88.95 | 88.53 | 84.07 | 90.24 | 92.00 | 84.98 | 90.04 | 85.18 | 88.38 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 94.34 | 95.15 | 98.83 | 96.27 | 95.94 | 97.82 | 98.13 | 93.80 | 94.15 | 93.35 | 95.78 |
| ORG | 86.33 | 91.74 | 82.10 | 86.85 | 76.52 | 80.40 | 86.33 | 83.36 | 88.48 | 82.29 | 84.44 |
| LOC | 89.56 | 85.27 | 86.13 | 87.18 | 83.93 | 89.63 | 91.10 | 85.87 | 87.50 | 76.06 | 86.22 |
| ALL | 90.05 | 92.11 | 90.52 | 90.57 | 86.78 | 91.23 | 92.65 | 88.26 | 90.81 | 86.15 | 89.91 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 95.28 | 94.71 | 98.31 | 96.53 | 94.97 | 97.10 | 98.00 | 92.90 | 94.11 | 93.39 | 95.53 |
| O | 85.15 | 91.11 | 78.00 | 85.88 | 77.66 | 80.50 | 85.12 | 78.33 | 88.84 | 83.86 | 83.44 |
| L | 89.73 | 88.58 | 87.29 | 89.71 | 87.78 | 89.31 | 90.54 | 87.76 | 87.73 | 81.53 | 88.00 |
| LO | 77.96 | 81.05 | 86.65 | 78.99 | 71.92 | 77.52 | 91.00 | 82.33 | 83.06 | 75.33 | 80.58 |
| OL | 85.88 | 83.06 | 76.25 | 54.84 | 33.76 | 67.05 | 95.20 | 81.73 | 72.16 | 43.13 | 69.30 |
| ALL | 89.49 | 91.76 | 89.49 | 89.77 | 85.17 | 90.08 | 92.61 | 86.39 | 90.28 | 86.33 | 89.14 |

ตารางที่ ง-11 ค่าความครบถ้วนของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| SSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (token) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 91.91 | 96.57 | 89.69 | 93.17 | 86.43 | 94.14 | 87.52 | 91.77 | 93.91 | 93.35 | 91.85 |
| ORG | 81.33 | 82.46 | 74.75 | 75.82 | 67.16 | 76.37 | 79.79 | 76.00 | 78.04 | 79.27 | 77.10 |
| LOC | 70.74 | 83.82 | 75.40 | 75.96 | 76.71 | 76.68 | 79.34 | 63.46 | 75.26 | 71.01 | 74.84 |
| ALL | 81.69 | 88.44 | 81.81 | 82.20 | 77.94 | 85.32 | 83.30 | 78.51 | 84.14 | 83.78 | 82.71 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 93.11 | 97.28 | 90.58 | 93.14 | 87.83 | 94.73 | 88.84 | 92.04 | 93.46 | 94.64 | 92.57 |
| ORG | 83.16 | 83.23 | 75.16 | 77.27 | 70.38 | 78.23 | 81.08 | 77.36 | 80.93 | 78.63 | 78.54 |
| LOC | 74.47 | 82.31 | 76.65 | 73.99 | 73.84 | 76.05 | 77.73 | 60.49 | 77.66 | 67.47 | 74.06 |
| ALL | 83.88 | 88.80 | 82.58 | 82.32 | 78.80 | 85.99 | 84.04 | 78.26 | 85.50 | 83.44 | 83.36 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 93.49 | 97.28 | 90.34 | 93.04 | 88.23 | 94.83 | 88.76 | 92.11 | 93.80 | 94.53 | 92.64 |
| O | 86.44 | 84.66 | 74.37 | 78.85 | 66.67 | 74.78 | 79.42 | 78.82 | 80.16 | 77.97 | 78.21 |
| L | 77.34 | 89.85 | 77.02 | 78.71 | 77.51 | 80.49 | 81.43 | 63.91 | 77.89 | 74.42 | 77.86 |
| LO | 57.88 | 62.18 | 60.96 | 74.91 | 65.37 | 78.40 | 81.66 | 65.25 | 73.21 | 64.31 | 68.41 |
| OL | 51.89 | 49.47 | 45.32 | 31.58 | 41.20 | 52.38 | 50.48 | 29.14 | 59.55 | 33.01 | 44.40 |
| ALL | 82.90 | 88.33 | 80.55 | 82.65 | 77.57 | 85.02 | 83.51 | 77.58 | 84.48 | 82.31 | 82.49 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 93.25 | 96.97 | 90.31 | 94.11 | 87.95 | 95.45 | 89.11 | 92.34 | 93.76 | 94.89 | 92.82 |
| ORG | 83.33 | 83.18 | 76.08 | 77.24 | 70.70 | 77.33 | 81.97 | 78.40 | 80.37 | 78.52 | 78.71 |
| LOC | 72.71 | 83.06 | 78.18 | 77.93 | 77.43 | 77.82 | 79.67 | 60.44 | 78.20 | 69.37 | 75.48 |
| ALL | 83.46 | 88.78 | 83.05 | 83.57 | 79.88 | 86.50 | 84.84 | 78.72 | 85.56 | 83.84 | 83.82 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 93.68 | 97.28 | 90.31 | 93.91 | 88.12 | 95.39 | 88.68 | 92.30 | 93.69 | 94.50 | 92.79 |
| O | 86.12 | 83.94 | 76.54 | 77.15 | 70.97 | 75.38 | 79.75 | 81.20 | 80.43 | 80.74 | 79.22 |
| L | 76.83 | 90.26 | 79.46 | 79.12 | 77.62 | 80.76 | 84.62 | 61.96 | 77.89 | 76.21 | 78.47 |
| LO | 58.37 | 64.71 | 61.70 | 76.03 | 64.79 | 80.49 | 80.65 | 65.25 | 71.79 | 64.02 | 68.78 |
| OL | 47.80 | 54.79 | 43.88 | 35.79 | 42.40 | 43.22 | 57.21 | 30.58 | 57.73 | 33.01 | 44.64 |
| ALL | 82.56 | 88.40 | 81.56 | 82.65 | 78.61 | 85.20 | 84.20 | 77.93 | 84.38 | 83.50 | 82.90 |

ตารางที่ ง-12 ค่า F-measure ของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยรูปแบบคำตอบทั้ง 5 แบบ

| SSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (token) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คำตอบแบบที่ 1 (P, O, L, X) | | | | | | | | | | | |
| PER | 93.34 | 95.84 | 93.20 | 94.26 | 90.44 | 95.52 | 92.02 | 92.07 | 93.72 | 93.15 | 93.36 |
| ORG | 81.05 | 85.93 | 78.50 | 78.11 | 70.12 | 77.57 | 82.72 | 80.00 | 81.23 | 80.50 | 79.57 |
| LOC | 78.04 | 84.78 | 78.57 | 79.56 | 77.79 | 81.54 | 83.37 | 71.45 | 80.76 | 74.01 | 78.99 |
| ALL | 84.40 | 89.87 | 85.37 | 84.44 | 80.84 | 87.48 | 87.15 | 82.58 | 86.60 | 84.82 | 85.35 |
| คำตอบแบบที่ 2 (B, I, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 94.03 | 95.68 | 93.66 | 94.84 | 91.42 | 96.20 | 92.80 | 91.72 | 94.24 | 93.75 | 93.84 |
| ORG | 84.00 | 86.91 | 79.22 | 80.14 | 71.26 | 78.71 | 83.42 | 78.66 | 83.59 | 80.14 | 80.61 |
| LOC | 80.99 | 84.08 | 82.18 | 79.04 | 77.30 | 82.89 | 83.57 | 70.90 | 82.82 | 72.37 | 79.61 |
| ALL | 86.58 | 90.12 | 86.52 | 85.37 | 81.43 | 88.43 | 87.81 | 81.94 | 88.04 | 84.75 | 86.10 |
| คำตอบแบบที่ 3 (B, I, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 94.33 | 95.45 | 93.48 | 94.95 | 91.54 | 96.26 | 92.54 | 91.44 | 94.08 | 93.30 | 93.74 |
| O | 85.29 | 87.43 | 77.68 | 81.96 | 70.70 | 76.97 | 81.88 | 78.30 | 83.53 | 80.13 | 80.39 |
| L | 82.73 | 88.59 | 80.73 | 82.29 | 81.61 | 84.97 | 86.13 | 72.94 | 83.12 | 77.23 | 82.04 |
| LO | 66.67 | 69.16 | 71.40 | 73.46 | 67.61 | 78.81 | 86.12 | 74.57 | 77.50 | 69.53 | 73.48 |
| OL | 66.13 | 60.78 | 45.82 | 40.82 | 35.95 | 59.58 | 65.83 | 41.65 | 64.53 | 37.50 | 51.86 |
| ALL | 85.90 | 89.50 | 84.54 | 85.49 | 80.69 | 87.55 | 87.55 | 81.11 | 87.17 | 83.72 | 85.32 |
| คำตอบแบบที่ 4 (B, I, E, X – PER, ORG, LOC) | | | | | | | | | | | |
| PER | 93.79 | 96.05 | 94.38 | 95.18 | 91.77 | 96.62 | 93.40 | 93.06 | 93.96 | 94.12 | 94.23 |
| ORG | 84.80 | 87.25 | 78.97 | 81.77 | 73.49 | 78.83 | 84.09 | 80.80 | 84.23 | 80.36 | 81.46 |
| LOC | 80.26 | 84.15 | 81.96 | 82.30 | 80.55 | 83.31 | 85.00 | 70.94 | 82.59 | 72.56 | 80.36 |
| ALL | 86.63 | 90.41 | 86.63 | 86.93 | 83.19 | 88.80 | 88.57 | 83.22 | 88.11 | 84.98 | 86.75 |
| คำตอบแบบที่ 5 (B, I, E, X – P, O, L, OL, LO) | | | | | | | | | | | |
| P | 94.48 | 95.98 | 94.14 | 95.20 | 91.42 | 96.24 | 93.11 | 92.60 | 93.90 | 93.94 | 94.10 |
| O | 85.63 | 87.38 | 77.26 | 81.28 | 74.16 | 77.86 | 82.34 | 79.74 | 84.42 | 82.27 | 81.24 |
| L | 82.78 | 89.41 | 83.19 | 84.08 | 82.39 | 84.82 | 87.48 | 72.64 | 82.51 | 78.78 | 82.81 |
| LO | 66.76 | 71.96 | 72.08 | 77.48 | 68.17 | 78.97 | 85.51 | 72.80 | 77.01 | 69.22 | 74.00 |
| OL | 61.41 | 66.03 | 55.71 | 43.31 | 37.59 | 52.56 | 71.47 | 44.50 | 64.14 | 37.40 | 53.41 |
| ALL | 85.88 | 90.05 | 85.34 | 86.06 | 81.76 | 87.57 | 88.21 | 81.94 | 87.23 | 84.90 | 85.89 |

# ภาคผนวก จ
# ผลการทดสอบข้อมูลทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

## ผลการทดสอบโดยประเมินจากจำนวนชื่อ

ข้อมูลแบบตัดคำ

ตารางที่ จ-1 ค่าความแม่นยำของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| WSG | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 90.25 | 91.56 | 91.74 | 91.90 | 88.99 | 89.78 | 91.05 | 93.72 | 88.96 | 94.28 | 91.22 |
| ORG | 87.42 | 90.35 | 86.96 | 87.13 | 80.79 | 85.05 | 89.52 | 90.37 | 91.62 | 85.91 | 87.51 |
| LOC | 87.83 | 87.56 | 75.35 | 82.34 | 82.21 | 86.73 | 84.97 | 79.11 | 82.61 | 79.15 | 82.79 |
| ALL | 88.30 | 90.25 | 86.68 | 87.59 | 84.57 | 87.67 | 89.07 | 88.28 | 87.99 | 87.99 | 87.84 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 88.05 | 93.18 | 92.74 | 92.48 | 91.62 | 91.73 | 91.63 | 93.00 | 92.34 | 96.45 | 92.32 |
| ORG | 86.62 | 88.80 | 86.70 | 85.05 | 76.09 | 83.09 | 88.40 | 88.52 | 88.79 | 86.25 | 85.83 |
| LOC | 84.89 | 88.56 | 74.52 | 77.67 | 79.01 | 86.81 | 85.75 | 75.31 | 83.83 | 75.74 | 81.21 |
| ALL | 86.40 | 90.26 | 86.74 | 85.60 | 83.04 | 87.96 | 89.07 | 86.20 | 88.73 | 87.77 | 87.18 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 90.46 | 91.84 | 91.33 | 92.44 | 88.93 | 91.24 | 90.71 | 92.52 | 88.82 | 94.64 | 91.29 |
| ORG | 86.40 | 90.27 | 84.84 | 86.63 | 79.66 | 82.46 | 88.87 | 90.35 | 91.22 | 84.75 | 86.55 |
| LOC | 86.89 | 85.65 | 75.43 | 82.92 | 80.89 | 87.34 | 84.10 | 78.63 | 82.75 | 80.16 | 82.48 |
| ALL | 87.63 | 89.96 | 85.69 | 87.78 | 83.74 | 87.74 | 88.50 | 87.77 | 87.94 | 87.83 | 87.46 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 85.96 | 91.07 | 92.87 | 90.64 | 89.77 | 88.21 | 89.03 | 88.84 | 87.70 | 88.81 | 89.29 |
| ORG | 85.69 | 88.25 | 83.16 | 83.02 | 79.25 | 79.35 | 86.71 | 88.63 | 88.71 | 83.54 | 84.63 |
| LOC | 86.68 | 85.71 | 77.35 | 75.87 | 79.07 | 85.76 | 83.05 | 77.17 | 81.77 | 76.62 | 80.91 |
| ALL | 86.09 | 88.76 | 86.21 | 83.81 | 83.43 | 85.06 | 86.79 | 85.32 | 86.41 | 84.27 | 85.62 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 89.75 | 91.56 | 91.40 | 91.57 | 89.44 | 90.38 | 90.86 | 93.13 | 88.06 | 94.09 | 91.02 |
| ORG | 86.38 | 90.35 | 85.13 | 86.32 | 81.10 | 82.15 | 89.29 | 90.42 | 91.71 | 84.25 | 86.71 |
| LOC | 87.59 | 87.14 | 76.16 | 81.03 | 80.22 | 86.64 | 83.94 | 77.40 | 82.14 | 80.31 | 82.26 |
| ALL | 87.67 | 90.17 | 86.01 | 86.82 | 84.19 | 87.07 | 88.66 | 87.60 | 87.62 | 87.53 | 87.33 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 89.44 | 91.07 | 91.54 | 90.12 | 89.07 | 90.16 | 91.39 | 92.92 | 88.82 | 93.25 | 90.78 |
| ORG | 86.53 | 90.76 | 83.80 | 86.15 | 81.32 | 83.89 | 89.53 | 89.16 | 92.01 | 87.27 | 87.04 |
| LOC | 87.83 | 86.26 | 75.62 | 81.82 | 81.60 | 86.22 | 84.30 | 78.65 | 84.57 | 81.01 | 82.79 |
| ALL | 87.74 | 90.00 | 85.45 | 86.46 | 84.55 | 87.39 | 89.07 | 87.55 | 88.66 | 88.44 | 87.53 |

ตารางที่ จ-2 ค่าความครบถ้วนของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| WSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 72.66 | 86.89 | 75.61 | 81.44 | 75.11 | 78.12 | 75.60 | 84.85 | 84.47 | 86.59 | 80.13 |
| ORG | 68.06 | 73.70 | 66.91 | 63.71 | 54.14 | 61.24 | 70.39 | 64.70 | 64.34 | 58.13 | 64.53 |
| LOC | 63.00 | 70.66 | 58.47 | 57.50 | 63.70 | 65.85 | 66.97 | 61.09 | 70.72 | 66.56 | 64.45 |
| ALL | 67.38 | 77.37 | 68.74 | 67.43 | 64.91 | 69.57 | 71.57 | 70.12 | 73.19 | 70.33 | 70.06 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 76.46 | 89.56 | 80.11 | 86.16 | 78.71 | 85.30 | 81.35 | 85.61 | 88.93 | 88.59 | 84.08 |
| ORG | 72.71 | 77.05 | 70.06 | 66.53 | 55.56 | 68.18 | 75.56 | 69.62 | 74.26 | 64.69 | 69.42 |
| LOC | 66.67 | 80.69 | 63.11 | 62.05 | 69.63 | 69.53 | 71.30 | 61.49 | 77.17 | 75.00 | 69.66 |
| ALL | 71.46 | 81.86 | 72.75 | 71.34 | 68.57 | 75.67 | 76.77 | 72.32 | 80.23 | 75.60 | 74.66 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 74.43 | 87.38 | 76.02 | 83.54 | 75.86 | 81.88 | 75.32 | 84.28 | 84.85 | 86.41 | 81.00 |
| ORG | 69.34 | 77.72 | 67.34 | 65.32 | 57.32 | 64.11 | 71.37 | 67.65 | 68.75 | 62.50 | 67.14 |
| LOC | 62.48 | 69.11 | 59.56 | 57.12 | 63.53 | 66.09 | 66.29 | 60.08 | 70.22 | 66.88 | 64.14 |
| ALL | 68.13 | 79.10 | 69.30 | 68.62 | 66.13 | 72.06 | 71.68 | 70.73 | 74.83 | 72.20 | 71.28 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 75.95 | 89.08 | 81.61 | 84.76 | 81.56 | 81.88 | 76.30 | 82.95 | 85.83 | 87.68 | 82.76 |
| ORG | 70.14 | 76.72 | 68.62 | 65.05 | 59.26 | 64.35 | 72.91 | 67.82 | 69.30 | 61.88 | 67.61 |
| LOC | 64.75 | 71.81 | 60.66 | 57.87 | 65.27 | 68.06 | 66.97 | 62.70 | 71.22 | 69.16 | 65.85 |
| ALL | 69.64 | 79.73 | 72.30 | 69.11 | 69.40 | 72.70 | 72.81 | 71.16 | 75.65 | 72.87 | 72.54 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 73.16 | 86.89 | 75.34 | 81.79 | 76.16 | 80.34 | 75.32 | 84.66 | 84.47 | 86.59 | 80.47 |
| ORG | 68.22 | 76.88 | 68.05 | 65.32 | 54.50 | 63.88 | 72.21 | 66.67 | 67.10 | 60.16 | 66.30 |
| LOC | 62.83 | 70.66 | 58.47 | 56.74 | 63.70 | 65.36 | 67.88 | 60.08 | 68.49 | 66.23 | 64.04 |
| ALL | 67.50 | 78.86 | 69.08 | 67.97 | 65.41 | 71.13 | 72.38 | 70.48 | 73.60 | 71.13 | 70.75 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 72.91 | 86.65 | 75.20 | 79.86 | 75.71 | 78.29 | 75.88 | 84.47 | 84.85 | 85.14 | 79.90 |
| ORG | 68.06 | 74.04 | 64.47 | 64.38 | 54.50 | 59.81 | 71.65 | 64.86 | 65.63 | 60.00 | 64.74 |
| LOC | 63.00 | 70.27 | 58.47 | 56.36 | 64.22 | 66.09 | 66.06 | 58.67 | 70.72 | 67.86 | 64.17 |
| ALL | 67.44 | 77.37 | 67.63 | 66.88 | 65.41 | 69.29 | 71.95 | 69.32 | 73.80 | 70.87 | 70.00 |

ตารางที่ จ-3 ค่า F-measure ของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| WSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 80.50 | 89.17 | 82.90 | 86.35 | 81.46 | 83.55 | 82.61 | 89.07 | 86.65 | 90.27 | 85.25 |
| ORG | 76.53 | 81.18 | 75.63 | 73.60 | 64.84 | 71.21 | 78.81 | 75.41 | 75.59 | 69.34 | 74.21 |
| LOC | 73.37 | 78.21 | 65.85 | 67.71 | 71.78 | 74.86 | 74.90 | 68.94 | 76.20 | 72.31 | 72.41 |
| ALL | 76.43 | 83.31 | 76.67 | 76.20 | 73.45 | 77.58 | 79.37 | 78.16 | 79.91 | 78.18 | 77.93 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 81.84 | 91.34 | 85.96 | 89.21 | 84.68 | 88.40 | 86.18 | 89.15 | 90.60 | 92.35 | 87.97 |
| ORG | 79.06 | 82.51 | 77.50 | 74.66 | 64.22 | 74.90 | 81.48 | 77.94 | 80.88 | 73.93 | 76.71 |
| LOC | 74.68 | 84.44 | 68.34 | 68.99 | 74.03 | 77.22 | 77.86 | 67.70 | 80.36 | 75.37 | 74.90 |
| ALL | 78.22 | 85.86 | 79.13 | 77.82 | 75.11 | 81.36 | 82.46 | 78.65 | 84.27 | 81.23 | 80.41 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 81.67 | 89.55 | 82.97 | 87.76 | 81.88 | 86.31 | 82.30 | 88.21 | 86.79 | 90.34 | 85.78 |
| ORG | 76.94 | 83.53 | 75.08 | 74.48 | 66.67 | 72.14 | 79.16 | 77.37 | 78.41 | 71.94 | 75.57 |
| LOC | 72.69 | 76.50 | 66.56 | 67.64 | 71.16 | 75.24 | 74.14 | 68.11 | 75.97 | 72.92 | 72.09 |
| ALL | 76.66 | 84.18 | 76.63 | 77.03 | 73.90 | 79.13 | 79.21 | 78.33 | 80.86 | 79.25 | 78.52 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 80.65 | 90.06 | 86.87 | 87.60 | 85.47 | 84.93 | 82.18 | 85.80 | 86.75 | 88.24 | 85.86 |
| ORG | 77.14 | 82.08 | 75.20 | 72.95 | 67.81 | 71.07 | 79.21 | 76.84 | 77.81 | 71.10 | 75.12 |
| LOC | 74.13 | 78.15 | 67.99 | 65.66 | 71.51 | 75.89 | 74.15 | 69.19 | 76.13 | 72.70 | 72.55 |
| ALL | 77.00 | 84.00 | 78.64 | 75.75 | 75.77 | 78.39 | 79.18 | 77.60 | 80.67 | 78.16 | 78.52 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 80.61 | 89.17 | 82.60 | 86.40 | 82.27 | 85.07 | 82.36 | 88.69 | 86.22 | 90.19 | 85.36 |
| ORG | 76.23 | 83.08 | 75.64 | 74.37 | 65.19 | 71.87 | 79.85 | 76.75 | 77.49 | 70.19 | 75.07 |
| LOC | 73.17 | 78.04 | 66.15 | 66.74 | 71.01 | 74.51 | 75.06 | 67.65 | 74.70 | 72.60 | 71.96 |
| ALL | 76.28 | 84.14 | 76.62 | 76.25 | 73.62 | 78.30 | 79.69 | 78.11 | 80.00 | 78.48 | 78.15 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 80.33 | 88.81 | 82.57 | 84.68 | 81.85 | 83.81 | 82.91 | 88.49 | 86.79 | 89.02 | 84.93 |
| ORG | 76.19 | 81.55 | 72.87 | 73.69 | 65.26 | 69.83 | 79.60 | 75.10 | 76.61 | 71.11 | 74.18 |
| LOC | 73.37 | 77.45 | 65.95 | 66.74 | 71.88 | 74.83 | 74.07 | 67.21 | 77.03 | 73.85 | 72.24 |
| ALL | 76.26 | 83.21 | 75.50 | 75.42 | 73.76 | 77.29 | 79.60 | 77.38 | 80.55 | 78.68 | 77.77 |

ข้อมูลแบบตัดพยางค์

ตารางที่ จ-4 ค่าความแม่นยำของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| SSG | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 92.31 | 93.55 | 92.68 | 93.54 | 91.03 | 94.03 | 92.83 | 93.32 | 91.58 | 95.25 | 93.01 |
| ORG | 84.56 | 88.89 | 86.63 | 83.04 | 78.62 | 75.29 | 86.32 | 84.76 | 87.64 | 84.51 | 84.03 |
| LOC | 86.71 | 83.67 | 73.29 | 77.86 | 79.49 | 85.63 | 85.01 | 74.38 | 81.39 | 74.09 | 80.15 |
| ALL | 87.25 | 89.37 | 86.43 | 85.34 | 83.84 | 86.52 | 88.49 | 84.77 | 87.38 | 86.26 | 86.57 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 91.44 | 94.07 | 93.33 | 92.97 | 91.27 | 93.76 | 91.46 | 93.89 | 91.90 | 96.18 | 93.03 |
| ORG | 83.24 | 88.62 | 85.69 | 82.12 | 75.28 | 75.41 | 85.90 | 85.29 | 86.93 | 82.88 | 83.14 |
| LOC | 85.39 | 83.33 | 74.45 | 77.75 | 79.31 | 84.76 | 82.22 | 74.82 | 83.06 | 73.75 | 79.88 |
| ALL | 85.99 | 89.35 | 86.51 | 84.64 | 82.64 | 85.92 | 87.11 | 85.25 | 87.72 | 85.82 | 86.10 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 92.59 | 92.79 | 93.17 | 93.74 | 91.18 | 95.29 | 92.68 | 93.69 | 92.21 | 95.42 | 93.28 |
| ORG | 82.75 | 88.80 | 84.20 | 81.29 | 78.40 | 77.72 | 85.99 | 85.04 | 89.43 | 83.33 | 83.70 |
| LOC | 87.39 | 82.73 | 73.83 | 79.11 | 79.11 | 86.86 | 84.76 | 74.94 | 81.44 | 75.60 | 80.58 |
| ALL | 86.75 | 88.88 | 85.88 | 84.96 | 83.63 | 87.98 | 88.24 | 85.16 | 88.26 | 86.25 | 86.60 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 89.77 | 93.63 | 93.87 | 91.85 | 89.50 | 92.47 | 91.43 | 91.20 | 90.22 | 92.79 | 91.67 |
| ORG | 83.67 | 87.55 | 83.98 | 81.61 | 74.71 | 74.79 | 82.64 | 84.29 | 84.83 | 83.05 | 82.11 |
| LOC | 85.74 | 81.85 | 73.60 | 78.45 | 78.12 | 84.23 | 83.60 | 75.38 | 81.77 | 71.05 | 79.38 |
| ALL | 85.92 | 88.44 | 86.03 | 84.30 | 81.66 | 85.15 | 86.13 | 84.23 | 86.03 | 84.14 | 85.20 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 92.00 | 92.59 | 92.50 | 93.41 | 90.92 | 94.19 | 93.16 | 93.51 | 91.39 | 95.44 | 92.91 |
| ORG | 84.66 | 89.06 | 86.01 | 81.45 | 76.65 | 75.35 | 85.35 | 85.19 | 87.11 | 84.89 | 83.57 |
| LOC | 86.61 | 83.06 | 73.29 | 80.32 | 79.83 | 85.80 | 84.49 | 74.50 | 81.01 | 76.68 | 80.56 |
| ALL | 87.18 | 89.04 | 86.12 | 85.26 | 83.31 | 86.53 | 88.11 | 85.04 | 87.04 | 87.18 | 86.48 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 92.07 | 93.56 | 92.04 | 93.70 | 90.28 | 94.36 | 92.54 | 93.48 | 91.15 | 95.45 | 92.86 |
| ORG | 84.19 | 88.37 | 84.53 | 82.75 | 78.26 | 76.22 | 85.26 | 83.83 | 88.21 | 85.41 | 83.70 |
| LOC | 86.40 | 82.68 | 73.18 | 78.57 | 80.99 | 85.40 | 84.81 | 76.07 | 81.84 | 76.92 | 80.69 |
| ALL | 86.94 | 88.93 | 85.39 | 85.41 | 83.90 | 86.82 | 87.92 | 85.03 | 87.55 | 87.43 | 86.53 |

ตารางที่ จ-5 ค่าความครบถ้วนของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| SSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 75.95 | 91.50 | 79.29 | 86.16 | 79.16 | 88.89 | 78.12 | 87.31 | 86.60 | 87.14 | 84.01 |
| ORG | 70.30 | 76.38 | 71.49 | 69.76 | 56.44 | 61.96 | 74.02 | 68.47 | 71.69 | 65.63 | 68.61 |
| LOC | 69.46 | 81.08 | 61.48 | 56.74 | 65.62 | 67.32 | 71.07 | 60.89 | 72.70 | 72.40 | 67.88 |
| ALL | 71.40 | 82.26 | 72.64 | 71.12 | 67.74 | 74.68 | 74.89 | 72.26 | 77.22 | 74.93 | 73.91 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 75.70 | 92.48 | 80.11 | 85.64 | 78.41 | 87.35 | 78.12 | 87.31 | 88.16 | 86.59 | 83.99 |
| ORG | 72.55 | 79.56 | 71.20 | 70.97 | 59.61 | 66.75 | 74.02 | 71.43 | 73.35 | 66.56 | 70.60 |
| LOC | 71.38 | 81.08 | 64.48 | 59.01 | 68.24 | 68.30 | 72.67 | 61.69 | 75.43 | 72.08 | 69.44 |
| ALL | 72.91 | 84.07 | 73.47 | 72.10 | 69.29 | 75.74 | 75.27 | 73.61 | 79.14 | 75.07 | 75.07 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 75.95 | 90.53 | 79.97 | 86.51 | 79.01 | 89.91 | 78.12 | 87.12 | 87.38 | 86.78 | 84.13 |
| ORG | 72.39 | 78.39 | 71.78 | 71.24 | 58.91 | 66.75 | 75.42 | 70.94 | 74.63 | 67.19 | 70.76 |
| LOC | 68.94 | 79.54 | 60.11 | 57.50 | 65.45 | 66.58 | 69.70 | 60.89 | 72.95 | 71.43 | 67.31 |
| ALL | 72.03 | 82.57 | 72.75 | 72.04 | 68.40 | 76.31 | 75.11 | 73.12 | 78.66 | 75.27 | 74.63 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 77.72 | 92.72 | 81.34 | 86.87 | 80.51 | 88.21 | 77.84 | 86.36 | 87.77 | 88.59 | 84.79 |
| ORG | 73.19 | 78.89 | 71.35 | 70.97 | 56.26 | 65.31 | 75.14 | 72.25 | 72.98 | 68.13 | 70.45 |
| LOC | 70.33 | 78.38 | 60.93 | 59.39 | 66.67 | 69.53 | 70.84 | 60.48 | 73.45 | 70.13 | 68.01 |
| ALL | 73.29 | 83.28 | 73.30 | 72.58 | 68.51 | 76.03 | 75.16 | 73.24 | 78.32 | 76.07 | 74.98 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 75.70 | 91.02 | 79.02 | 86.87 | 79.61 | 88.72 | 78.26 | 87.31 | 86.60 | 87.14 | 84.03 |
| ORG | 69.98 | 77.72 | 71.35 | 70.83 | 57.32 | 64.35 | 74.02 | 68.97 | 72.06 | 69.38 | 69.60 |
| LOC | 69.98 | 77.61 | 61.48 | 57.31 | 64.92 | 66.83 | 69.48 | 60.69 | 71.96 | 70.45 | 67.07 |
| ALL | 71.40 | 82.02 | 72.47 | 71.93 | 67.96 | 75.18 | 74.57 | 72.38 | 77.15 | 76.13 | 74.12 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 76.46 | 91.75 | 78.75 | 85.99 | 77.96 | 88.72 | 78.26 | 86.93 | 86.02 | 87.32 | 83.82 |
| ORG | 70.95 | 76.38 | 68.91 | 69.62 | 57.14 | 63.64 | 74.30 | 68.97 | 71.51 | 68.59 | 69.00 |
| LOC | 68.76 | 81.08 | 60.38 | 58.44 | 65.45 | 66.09 | 69.93 | 60.89 | 72.70 | 71.43 | 67.52 |
| ALL | 71.53 | 82.33 | 71.19 | 71.50 | 67.46 | 74.75 | 74.79 | 72.32 | 76.95 | 76.07 | 73.89 |

ตารางที่ จ-6 ค่า F-measure ของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| SSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 83.33 | 92.52 | 85.46 | 89.70 | 84.68 | 91.39 | 84.84 | 90.22 | 89.02 | 91.01 | 88.22 |
| ORG | 76.77 | 82.16 | 78.34 | 75.82 | 65.71 | 67.98 | 79.70 | 75.75 | 78.87 | 73.88 | 75.50 |
| LOC | 77.13 | 82.35 | 66.86 | 65.64 | 71.89 | 75.38 | 77.42 | 66.96 | 76.80 | 73.23 | 73.37 |
| ALL | 78.53 | 85.67 | 78.94 | 77.58 | 74.93 | 80.17 | 81.12 | 78.02 | 81.99 | 80.20 | 79.72 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 82.83 | 93.27 | 86.22 | 89.15 | 84.35 | 90.44 | 84.27 | 90.48 | 89.99 | 91.13 | 88.21 |
| ORG | 77.53 | 83.85 | 77.78 | 76.14 | 66.54 | 70.81 | 79.52 | 77.75 | 79.56 | 73.83 | 76.33 |
| LOC | 77.76 | 82.19 | 69.11 | 67.10 | 73.36 | 75.65 | 77.15 | 67.62 | 79.06 | 72.91 | 74.19 |
| ALL | 78.91 | 86.63 | 79.46 | 77.87 | 75.38 | 80.51 | 80.76 | 79.00 | 83.21 | 80.09 | 80.18 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 83.45 | 91.65 | 86.07 | 89.98 | 84.66 | 92.52 | 84.78 | 90.28 | 89.73 | 90.89 | 88.40 |
| ORG | 77.23 | 83.27 | 77.49 | 75.93 | 67.27 | 71.81 | 80.36 | 77.35 | 81.36 | 74.39 | 76.65 |
| LOC | 77.07 | 81.10 | 66.27 | 66.59 | 71.63 | 75.38 | 76.50 | 67.19 | 76.96 | 73.46 | 73.22 |
| ALL | 78.71 | 85.61 | 78.77 | 77.97 | 75.25 | 81.73 | 81.15 | 78.68 | 83.18 | 80.38 | 80.14 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 83.31 | 93.17 | 87.15 | 89.29 | 84.77 | 90.29 | 84.09 | 88.72 | 88.98 | 90.64 | 88.04 |
| ORG | 78.08 | 83.00 | 77.15 | 75.92 | 64.19 | 69.73 | 78.71 | 77.81 | 78.46 | 74.85 | 75.79 |
| LOC | 77.28 | 80.08 | 66.67 | 67.60 | 71.94 | 76.18 | 76.70 | 67.11 | 77.39 | 70.59 | 73.15 |
| ALL | 79.10 | 85.78 | 79.16 | 78.00 | 74.51 | 80.33 | 80.27 | 78.35 | 81.99 | 79.90 | 79.74 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 83.06 | 91.80 | 85.23 | 90.02 | 84.89 | 91.37 | 85.06 | 90.30 | 88.93 | 91.10 | 88.18 |
| ORG | 76.63 | 83.01 | 78.00 | 75.77 | 65.59 | 69.42 | 79.28 | 76.23 | 78.87 | 76.35 | 75.92 |
| LOC | 77.41 | 80.24 | 66.86 | 66.89 | 71.61 | 75.14 | 76.25 | 66.89 | 76.22 | 73.43 | 73.09 |
| ALL | 78.51 | 85.39 | 78.71 | 78.03 | 74.86 | 80.46 | 80.78 | 78.20 | 81.80 | 81.28 | 79.80 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 83.54 | 92.65 | 84.88 | 89.68 | 83.67 | 91.45 | 84.80 | 90.09 | 88.51 | 91.20 | 88.05 |
| ORG | 77.00 | 81.94 | 75.93 | 75.62 | 66.06 | 69.36 | 79.40 | 75.68 | 78.98 | 76.08 | 75.61 |
| LOC | 76.58 | 81.87 | 66.17 | 67.03 | 72.39 | 74.52 | 76.65 | 67.64 | 77.00 | 74.07 | 73.39 |
| ALL | 78.48 | 85.50 | 77.65 | 77.84 | 74.79 | 80.34 | 80.82 | 78.16 | 81.91 | 81.35 | 79.68 |

## ผลการทดสอบโดยประเมินจากจำนวน token

ข้อมูลแบบตัดคำ

ตารางที่ จ-7 ค่าความแม่นยำของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| WSG | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (token) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 88.74 | 94.32 | 97.73 | 96.89 | 97.89 | 97.29 | 96.75 | 97.07 | 94.74 | 93.50 | 95.49 |
| ORG | 87.28 | 93.42 | 88.51 | 89.06 | 83.93 | 85.04 | 92.64 | 89.62 | 93.83 | 86.68 | 89.00 |
| LOC | 92.35 | 88.35 | 89.76 | 88.25 | 89.66 | 91.36 | 93.84 | 89.32 | 87.99 | 80.03 | 89.09 |
| ALL | 89.20 | 93.12 | 93.67 | 92.56 | 92.17 | 93.32 | 95.14 | 92.98 | 92.93 | 89.19 | 92.43 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 91.15 | 94.74 | 96.98 | 98.00 | 98.20 | 98.46 | 98.24 | 96.56 | 96.87 | 96.60 | 96.58 |
| ORG | 87.26 | 91.93 | 91.71 | 87.20 | 82.81 | 84.51 | 90.37 | 86.18 | 89.89 | 87.52 | 87.94 |
| LOC | 89.95 | 88.91 | 87.10 | 84.68 | 86.39 | 91.46 | 93.61 | 87.82 | 89.13 | 76.17 | 87.52 |
| ALL | 89.65 | 92.87 | 93.71 | 91.58 | 90.95 | 93.77 | 95.26 | 91.27 | 93.29 | 89.96 | 92.23 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 89.05 | 94.85 | 97.61 | 96.25 | 97.77 | 97.31 | 96.36 | 96.71 | 93.87 | 94.16 | 95.40 |
| ORG | 85.89 | 92.88 | 86.79 | 88.13 | 83.90 | 83.09 | 91.72 | 90.42 | 93.66 | 86.01 | 88.25 |
| LOC | 91.24 | 87.94 | 88.86 | 88.78 | 88.66 | 91.52 | 93.58 | 88.97 | 88.47 | 80.12 | 88.81 |
| ALL | 88.62 | 93.16 | 92.89 | 92.15 | 91.81 | 92.95 | 94.64 | 92.96 | 92.58 | 89.32 | 92.11 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 82.19 | 91.04 | 96.99 | 93.98 | 94.73 | 93.77 | 93.41 | 91.66 | 92.18 | 86.74 | 91.67 |
| ORG | 85.19 | 91.00 | 82.96 | 83.39 | 80.83 | 79.82 | 85.99 | 85.10 | 89.18 | 83.59 | 84.71 |
| LOC | 90.45 | 85.66 | 89.55 | 81.93 | 86.15 | 87.83 | 91.47 | 88.03 | 87.53 | 76.48 | 86.51 |
| ALL | 85.15 | 90.22 | 91.46 | 87.95 | 88.91 | 89.34 | 91.13 | 88.87 | 90.41 | 84.04 | 88.75 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 88.12 | 94.68 | 96.91 | 97.60 | 97.44 | 97.16 | 97.68 | 96.56 | 93.39 | 92.12 | 95.17 |
| ORG | 85.81 | 93.21 | 89.40 | 87.29 | 83.84 | 81.20 | 92.21 | 90.45 | 94.05 | 86.03 | 88.35 |
| LOC | 91.76 | 88.16 | 90.27 | 87.26 | 87.97 | 91.02 | 93.23 | 89.04 | 88.06 | 80.15 | 88.69 |
| ALL | 88.33 | 93.21 | 93.58 | 92.06 | 91.53 | 92.24 | 95.36 | 92.89 | 92.32 | 88.33 | 91.98 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 85.50 | 92.95 | 97.52 | 96.39 | 96.24 | 97.05 | 96.92 | 95.44 | 93.89 | 92.50 | 94.44 |
| ORG | 86.39 | 93.42 | 83.71 | 87.65 | 84.49 | 84.32 | 92.73 | 84.80 | 94.56 | 87.43 | 87.95 |
| LOC | 92.12 | 87.14 | 89.88 | 87.51 | 88.60 | 90.79 | 94.25 | 89.02 | 89.87 | 81.80 | 89.10 |
| ALL | 87.38 | 92.22 | 91.91 | 91.70 | 91.24 | 92.92 | 95.32 | 90.74 | 93.11 | 89.21 | 91.58 |

ตารางที่ จ-8 ค่าความครบถ้วนของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| WSG (token) | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 88.37 | 94.64 | 86.26 | 88.61 | 82.53 | 89.53 | 84.62 | 89.52 | 91.59 | 93.46 | 88.91 |
| ORG | 75.04 | 73.37 | 65.80 | 62.70 | 62.04 | 63.86 | 67.20 | 65.38 | 65.00 | 64.77 | 66.52 |
| LOC | 66.19 | 74.47 | 71.19 | 70.66 | 71.50 | 73.08 | 74.79 | 62.74 | 74.31 | 66.47 | 70.54 |
| ALL | 77.48 | 83.47 | 76.98 | 75.52 | 73.97 | 79.35 | 77.71 | 74.23 | 79.98 | 78.62 | 77.73 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 90.98 | 95.54 | 89.28 | 92.17 | 84.94 | 92.33 | 88.67 | 89.86 | 93.02 | 93.60 | 91.04 |
| ORG | 75.92 | 76.63 | 70.36 | 63.42 | 65.12 | 68.45 | 72.24 | 68.77 | 71.81 | 67.55 | 70.03 |
| LOC | 68.60 | 81.34 | 74.85 | 74.78 | 77.38 | 75.05 | 77.95 | 64.36 | 78.70 | 72.24 | 74.52 |
| ALL | 79.49 | 86.22 | 80.61 | 78.25 | 77.53 | 82.38 | 81.85 | 75.86 | 83.63 | 80.70 | 80.65 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 89.74 | 94.69 | 86.26 | 90.37 | 83.12 | 92.03 | 84.96 | 89.67 | 92.05 | 92.81 | 89.57 |
| ORG | 75.55 | 74.79 | 66.45 | 63.37 | 63.23 | 64.89 | 67.39 | 66.45 | 67.85 | 65.84 | 67.58 |
| LOC | 65.74 | 73.98 | 71.88 | 70.37 | 71.56 | 72.62 | 75.12 | 61.58 | 73.67 | 65.87 | 70.24 |
| ALL | 78.03 | 83.95 | 77.31 | 76.42 | 74.58 | 80.77 | 78.01 | 74.25 | 80.84 | 78.58 | 78.27 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 89.80 | 95.20 | 90.69 | 92.17 | 88.61 | 92.95 | 86.27 | 89.81 | 93.76 | 94.11 | 91.34 |
| ORG | 76.87 | 78.05 | 69.60 | 64.26 | 67.32 | 67.79 | 68.26 | 69.44 | 69.20 | 67.81 | 69.86 |
| LOC | 69.05 | 75.29 | 72.97 | 73.68 | 74.71 | 75.51 | 74.96 | 64.23 | 75.05 | 69.95 | 72.54 |
| ALL | 79.47 | 85.61 | 80.73 | 78.27 | 79.04 | 82.65 | 78.89 | 76.00 | 82.37 | 80.61 | 80.36 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 89.32 | 94.41 | 86.05 | 88.69 | 83.23 | 90.06 | 84.08 | 89.71 | 91.45 | 93.27 | 89.03 |
| ORG | 72.83 | 74.86 | 67.43 | 63.94 | 60.34 | 64.70 | 69.81 | 66.71 | 66.35 | 65.97 | 67.29 |
| LOC | 65.29 | 74.30 | 70.69 | 71.03 | 72.29 | 71.96 | 75.62 | 63.07 | 72.12 | 64.54 | 70.09 |
| ALL | 76.89 | 83.90 | 77.31 | 76.07 | 74.06 | 79.55 | 78.34 | 74.80 | 79.76 | 78.60 | 77.93 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 88.49 | 94.58 | 85.83 | 88.44 | 82.57 | 89.36 | 84.99 | 89.57 | 91.59 | 93.27 | 88.87 |
| ORG | 74.60 | 73.44 | 69.00 | 64.62 | 62.98 | 62.92 | 68.88 | 67.84 | 66.03 | 66.10 | 67.64 |
| LOC | 64.99 | 74.30 | 71.19 | 70.07 | 72.10 | 72.80 | 74.96 | 59.77 | 73.77 | 66.47 | 70.04 |
| ALL | 77.03 | 83.45 | 77.80 | 75.96 | 74.41 | 78.96 | 78.41 | 74.07 | 80.13 | 78.99 | 77.92 |

ตารางที่ จ-9 ค่า F-measure ของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| WSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (token) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 88.56 | 94.48 | 91.64 | 92.56 | 89.56 | 93.25 | 90.28 | 93.14 | 93.14 | 93.48 | 92.01 |
| ORG | 80.69 | 82.19 | 75.48 | 73.59 | 71.34 | 72.94 | 77.90 | 75.61 | 76.80 | 74.14 | 76.07 |
| LOC | 77.11 | 80.82 | 79.40 | 78.48 | 79.55 | 81.20 | 83.24 | 73.71 | 80.57 | 72.62 | 78.67 |
| ALL | 82.93 | 88.03 | 84.51 | 83.18 | 82.07 | 85.77 | 85.55 | 82.55 | 85.97 | 83.57 | 84.41 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 91.07 | 95.14 | 92.97 | 94.99 | 91.09 | 95.30 | 93.21 | 93.09 | 94.91 | 95.07 | 93.68 |
| ORG | 81.19 | 83.58 | 79.63 | 73.43 | 72.91 | 75.63 | 80.29 | 76.50 | 79.84 | 76.25 | 77.93 |
| LOC | 77.83 | 84.96 | 80.51 | 79.42 | 81.64 | 82.44 | 85.07 | 74.28 | 83.59 | 74.15 | 80.39 |
| ALL | 84.27 | 89.42 | 86.67 | 84.39 | 83.70 | 87.71 | 88.05 | 82.85 | 88.19 | 85.08 | 86.03 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 89.39 | 94.77 | 91.58 | 93.22 | 89.86 | 94.60 | 90.30 | 93.05 | 92.95 | 93.48 | 92.32 |
| ORG | 80.39 | 82.86 | 75.27 | 73.73 | 72.11 | 72.87 | 77.69 | 76.60 | 78.70 | 74.59 | 76.48 |
| LOC | 76.42 | 80.36 | 79.47 | 78.51 | 79.19 | 80.98 | 83.34 | 72.78 | 80.40 | 72.30 | 78.37 |
| ALL | 82.99 | 88.31 | 84.39 | 83.55 | 82.30 | 86.43 | 85.52 | 82.55 | 86.31 | 83.60 | 84.60 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 85.83 | 93.07 | 93.73 | 93.07 | 91.57 | 93.36 | 89.70 | 90.73 | 92.96 | 90.27 | 91.43 |
| ORG | 80.82 | 84.03 | 75.69 | 72.58 | 73.46 | 73.32 | 76.11 | 76.47 | 77.93 | 74.88 | 76.53 |
| LOC | 78.31 | 80.14 | 80.41 | 77.58 | 80.03 | 81.21 | 82.40 | 74.27 | 80.81 | 73.07 | 78.82 |
| ALL | 82.21 | 87.86 | 85.76 | 82.83 | 83.69 | 85.87 | 84.57 | 81.93 | 86.20 | 82.29 | 84.32 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 88.72 | 94.54 | 91.16 | 92.94 | 89.78 | 93.48 | 90.37 | 93.01 | 92.41 | 92.69 | 91.91 |
| ORG | 78.79 | 83.03 | 76.88 | 73.81 | 70.18 | 72.02 | 79.46 | 76.79 | 77.81 | 74.68 | 76.34 |
| LOC | 76.29 | 80.64 | 79.29 | 78.31 | 79.36 | 80.38 | 83.51 | 73.84 | 79.30 | 71.50 | 78.24 |
| ALL | 82.21 | 88.31 | 84.67 | 83.30 | 81.87 | 85.43 | 86.01 | 82.87 | 85.58 | 83.18 | 84.34 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 86.97 | 93.76 | 91.31 | 92.25 | 88.88 | 93.04 | 90.56 | 92.41 | 92.72 | 92.88 | 91.48 |
| ORG | 80.06 | 82.24 | 75.65 | 74.39 | 72.16 | 72.06 | 79.04 | 75.38 | 77.76 | 75.28 | 76.40 |
| LOC | 76.21 | 80.21 | 79.45 | 77.83 | 79.51 | 80.81 | 83.50 | 71.52 | 81.02 | 73.34 | 78.34 |
| ALL | 81.88 | 87.62 | 84.27 | 83.09 | 81.97 | 85.38 | 86.04 | 81.56 | 86.14 | 83.79 | 84.17 |

ข้อมูลแบบตัดพยางค์

ตารางที่ จ-10 ค่าความแม่นยำของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| SSG (token) | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 92.43 | 96.71 | 98.52 | 95.98 | 97.58 | 98.41 | 97.02 | 94.44 | 95.21 | 95.21 | 96.15 |
| ORG | 87.72 | 92.26 | 88.73 | 87.66 | 82.69 | 83.91 | 89.61 | 86.39 | 91.13 | 85.31 | 87.54 |
| LOC | 90.95 | 88.70 | 86.11 | 86.95 | 86.09 | 91.57 | 92.17 | 84.84 | 88.38 | 73.56 | 86.93 |
| ALL | 90.31 | 93.60 | 93.01 | 90.86 | 90.17 | 93.13 | 93.65 | 89.47 | 92.41 | 87.79 | 91.44 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 95.10 | 95.81 | 97.75 | 96.92 | 97.57 | 98.63 | 98.22 | 97.41 | 95.60 | 95.85 | 96.88 |
| ORG | 86.76 | 91.07 | 90.21 | 87.07 | 81.88 | 82.52 | 90.19 | 86.88 | 90.72 | 84.24 | 87.15 |
| LOC | 90.14 | 87.31 | 85.96 | 87.14 | 83.75 | 90.39 | 90.80 | 85.03 | 89.21 | 73.21 | 86.29 |
| ALL | 90.67 | 92.52 | 93.15 | 90.97 | 89.18 | 92.53 | 94.12 | 90.82 | 92.63 | 87.50 | 91.41 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 92.98 | 96.32 | 98.09 | 95.72 | 98.09 | 98.64 | 98.14 | 95.03 | 95.40 | 95.45 | 96.39 |
| ORG | 85.73 | 92.07 | 87.54 | 86.28 | 82.68 | 83.64 | 89.10 | 86.52 | 92.00 | 83.56 | 86.91 |
| LOC | 91.55 | 87.67 | 86.87 | 87.44 | 85.63 | 92.63 | 92.66 | 84.95 | 88.24 | 74.50 | 87.21 |
| ALL | 89.85 | 93.19 | 92.51 | 90.29 | 90.23 | 93.37 | 94.04 | 89.78 | 92.73 | 87.39 | 91.34 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 89.75 | 94.52 | 98.22 | 94.16 | 96.10 | 97.28 | 96.51 | 93.03 | 92.89 | 91.13 | 94.36 |
| ORG | 86.54 | 91.45 | 85.40 | 86.31 | 79.43 | 82.53 | 84.80 | 81.61 | 87.82 | 83.42 | 84.93 |
| LOC | 90.20 | 87.16 | 87.33 | 88.19 | 83.88 | 90.37 | 91.40 | 84.91 | 87.81 | 71.32 | 86.26 |
| ALL | 88.70 | 92.09 | 91.86 | 89.86 | 87.98 | 91.85 | 91.51 | 87.12 | 90.15 | 84.89 | 89.60 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 89.88 | 94.84 | 98.37 | 95.83 | 97.40 | 98.38 | 98.11 | 94.45 | 94.60 | 93.85 | 95.57 |
| ORG | 87.67 | 91.96 | 89.03 | 86.76 | 81.40 | 81.73 | 88.27 | 86.82 | 90.33 | 85.61 | 86.96 |
| LOC | 91.18 | 88.37 | 85.80 | 89.52 | 85.33 | 92.03 | 91.92 | 83.77 | 88.26 | 75.58 | 87.18 |
| ALL | 89.42 | 92.65 | 93.00 | 90.97 | 89.50 | 92.54 | 93.61 | 89.38 | 91.86 | 87.75 | 91.07 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 87.19 | 95.77 | 98.02 | 96.30 | 97.35 | 98.38 | 96.96 | 94.54 | 94.70 | 95.05 | 95.43 |
| ORG | 86.80 | 91.77 | 84.59 | 88.90 | 82.44 | 84.02 | 88.48 | 83.08 | 91.34 | 84.49 | 86.59 |
| LOC | 91.00 | 87.00 | 85.59 | 86.60 | 86.59 | 91.85 | 92.70 | 86.25 | 88.95 | 76.49 | 87.30 |
| ALL | 88.02 | 92.72 | 91.19 | 91.36 | 90.08 | 93.19 | 93.31 | 88.63 | 92.37 | 87.99 | 90.89 |

ตารางที่ จ-11 ค่าความครบถ้วนของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| SSG (token) | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 91.67 | 96.79 | 88.92 | 92.77 | 86.57 | 95.27 | 87.09 | 91.51 | 92.60 | 93.75 | 91.69 |
| ORG | 79.69 | 79.59 | 74.55 | 72.80 | 69.23 | 74.02 | 76.03 | 73.44 | 76.78 | 73.30 | 74.94 |
| LOC | 70.68 | 82.96 | 73.06 | 71.00 | 73.12 | 74.70 | 75.74 | 60.91 | 74.57 | 66.01 | 72.27 |
| ALL | 81.00 | 87.16 | 80.98 | 79.69 | 77.72 | 84.79 | 81.11 | 76.82 | 82.99 | 80.66 | 81.29 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 91.87 | 96.75 | 89.45 | 92.51 | 86.34 | 94.80 | 87.84 | 91.28 | 93.84 | 92.46 | 91.71 |
| ORG | 80.40 | 79.85 | 73.49 | 73.65 | 70.86 | 74.97 | 76.99 | 75.22 | 77.62 | 73.68 | 75.67 |
| LOC | 72.12 | 80.91 | 76.21 | 72.97 | 74.18 | 74.63 | 78.76 | 61.22 | 76.49 | 66.01 | 73.35 |
| ALL | 81.76 | 86.91 | 81.37 | 80.39 | 78.40 | 84.80 | 82.38 | 77.43 | 84.25 | 80.27 | 81.80 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 91.91 | 96.75 | 89.37 | 92.74 | 86.57 | 95.77 | 87.17 | 91.39 | 92.71 | 92.96 | 91.73 |
| ORG | 82.22 | 79.46 | 75.67 | 73.71 | 69.94 | 74.86 | 78.36 | 73.94 | 77.90 | 73.30 | 75.94 |
| LOC | 69.88 | 82.09 | 72.18 | 71.27 | 72.93 | 73.00 | 74.77 | 60.23 | 75.05 | 64.80 | 71.62 |
| ALL | 81.76 | 86.95 | 81.45 | 80.11 | 77.89 | 84.90 | 81.75 | 76.76 | 83.53 | 80.12 | 81.52 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 92.63 | 96.79 | 90.20 | 93.31 | 88.95 | 95.42 | 87.33 | 91.58 | 93.69 | 93.96 | 92.39 |
| ORG | 80.84 | 81.43 | 74.51 | 74.31 | 68.87 | 75.53 | 77.36 | 75.51 | 77.52 | 75.59 | 76.15 |
| LOC | 72.12 | 80.58 | 74.67 | 72.81 | 74.89 | 77.18 | 77.28 | 59.76 | 75.53 | 65.66 | 73.05 |
| ALL | 82.18 | 87.55 | 81.83 | 80.92 | 79.09 | 85.84 | 81.98 | 77.25 | 83.93 | 81.62 | 82.22 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 91.39 | 96.70 | 88.95 | 93.11 | 86.80 | 95.11 | 87.03 | 91.58 | 92.86 | 93.82 | 91.74 |
| ORG | 79.02 | 79.25 | 74.11 | 74.64 | 68.83 | 74.07 | 77.14 | 72.95 | 76.87 | 76.64 | 75.35 |
| LOC | 70.58 | 80.37 | 73.87 | 70.15 | 72.02 | 73.64 | 73.94 | 60.44 | 74.23 | 64.11 | 71.33 |
| ALL | 80.63 | 86.55 | 80.96 | 80.36 | 77.41 | 84.49 | 81.11 | 76.54 | 83.05 | 81.71 | 81.28 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 91.87 | 96.79 | 88.81 | 92.37 | 86.14 | 95.24 | 87.06 | 91.28 | 92.64 | 94.07 | 91.63 |
| ORG | 79.51 | 78.78 | 73.15 | 73.35 | 70.06 | 74.64 | 77.03 | 73.40 | 77.38 | 75.67 | 75.30 |
| LOC | 70.10 | 82.31 | 71.30 | 72.33 | 72.55 | 73.49 | 75.23 | 60.07 | 74.50 | 64.28 | 71.62 |
| ALL | 80.82 | 86.71 | 80.12 | 80.07 | 77.64 | 84.68 | 81.34 | 76.49 | 83.20 | 81.45 | 81.25 |

ตารางที่ จ-12 ค่า F-measure ของการทดสอบแบบจำลองทั้ง 10 ครั้งด้วยคุณสมบัติแต่ละชนิด

| SSG (token) | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram | | | | | | | | | | | |
| PER | 92.05 | 96.75 | 93.47 | 94.35 | 91.75 | 96.81 | 91.79 | 92.95 | 93.89 | 94.47 | 93.83 |
| ORG | 83.51 | 85.46 | 81.02 | 79.54 | 75.36 | 78.65 | 82.27 | 79.39 | 83.34 | 78.85 | 80.74 |
| LOC | 79.54 | 85.73 | 79.05 | 78.17 | 79.08 | 82.28 | 83.15 | 70.91 | 80.89 | 69.58 | 78.84 |
| ALL | 85.40 | 90.27 | 86.58 | 84.91 | 83.48 | 88.76 | 86.93 | 82.66 | 87.45 | 84.07 | 86.05 |
| คุณสมบัติ unigram และ bigram + รายการชื่อเฉพาะ | | | | | | | | | | | |
| PER | 93.45 | 96.27 | 93.42 | 94.66 | 91.62 | 96.68 | 92.74 | 94.24 | 94.71 | 94.12 | 94.19 |
| ORG | 83.46 | 85.09 | 81.00 | 79.80 | 75.97 | 78.57 | 83.07 | 80.63 | 83.66 | 78.61 | 80.98 |
| LOC | 80.13 | 83.99 | 80.79 | 79.43 | 78.67 | 81.75 | 84.36 | 71.19 | 82.36 | 69.42 | 79.21 |
| ALL | 85.98 | 89.63 | 86.86 | 85.35 | 83.44 | 88.50 | 87.86 | 83.60 | 88.24 | 83.73 | 86.32 |
| คุณสมบัติ unigram และ bigram + คำย่อ | | | | | | | | | | | |
| PER | 92.44 | 96.53 | 93.53 | 94.21 | 91.97 | 97.18 | 92.33 | 93.18 | 94.04 | 94.19 | 93.96 |
| ORG | 83.94 | 85.30 | 81.18 | 79.50 | 75.78 | 79.01 | 83.38 | 79.73 | 84.36 | 78.10 | 81.03 |
| LOC | 79.26 | 84.79 | 78.85 | 78.53 | 78.77 | 81.65 | 82.76 | 70.48 | 81.11 | 69.31 | 78.55 |
| ALL | 85.61 | 89.96 | 86.63 | 84.90 | 83.61 | 88.93 | 87.47 | 82.76 | 87.89 | 83.60 | 86.13 |
| คุณสมบัติ unigram และ bigram + คำบริบท | | | | | | | | | | | |
| PER | 91.17 | 95.64 | 94.04 | 93.73 | 92.39 | 96.34 | 91.69 | 92.30 | 93.29 | 92.52 | 93.31 |
| ORG | 83.59 | 86.15 | 79.59 | 79.86 | 73.77 | 78.87 | 80.91 | 78.44 | 82.35 | 79.31 | 80.29 |
| LOC | 80.15 | 83.74 | 80.51 | 79.77 | 79.13 | 83.26 | 83.75 | 70.15 | 81.21 | 68.37 | 79.00 |
| ALL | 85.31 | 89.76 | 86.56 | 85.16 | 83.30 | 88.74 | 86.48 | 81.89 | 86.93 | 83.22 | 85.73 |
| คุณสมบัติ unigram และ bigram + คำทั่วไป | | | | | | | | | | | |
| PER | 90.63 | 95.76 | 93.42 | 94.45 | 91.80 | 96.72 | 92.24 | 92.99 | 93.73 | 93.83 | 93.56 |
| ORG | 83.12 | 85.13 | 80.89 | 80.25 | 74.59 | 77.72 | 82.33 | 79.28 | 83.06 | 80.88 | 80.72 |
| LOC | 79.57 | 84.18 | 79.39 | 78.66 | 78.11 | 81.81 | 81.95 | 70.21 | 80.64 | 69.37 | 78.39 |
| ALL | 84.80 | 89.50 | 86.56 | 85.34 | 83.02 | 88.33 | 86.92 | 82.46 | 87.23 | 84.62 | 85.88 |
| คุณสมบัติ unigram และ bigram + ค่าทางสถิติ | | | | | | | | | | | |
| PER | 89.47 | 96.28 | 93.19 | 94.30 | 91.40 | 96.78 | 91.74 | 92.88 | 93.66 | 94.56 | 93.43 |
| ORG | 83.00 | 84.78 | 78.46 | 80.38 | 75.75 | 79.05 | 82.36 | 77.94 | 83.78 | 79.83 | 80.53 |
| LOC | 79.19 | 84.59 | 77.80 | 78.83 | 78.95 | 81.65 | 83.06 | 70.82 | 81.09 | 69.85 | 78.58 |
| ALL | 84.27 | 89.61 | 85.30 | 85.34 | 83.40 | 88.73 | 86.91 | 82.11 | 87.55 | 84.59 | 85.78 |

# ภาคผนวก ฉ
# ผลการทดสอบข้อมูลก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

## ผลการทดสอบโดยประเมินจากจำนวนชื่อ

ข้อมูลแบบตัดคำ

ตารางที่ ฉ-1 ค่าความแม่นยำของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| WSG | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 88.86 | 94.25 | 92.51 | 92.56 | 92.98 | 92.96 | 90.42 | 93.47 | 90.16 | 92.36 | 92.05 |
| ORG | 81.42 | 88.89 | 82.67 | 80.46 | 72.22 | 74.29 | 86.77 | 86.00 | 86.63 | 82.57 | 82.19 |
| LOC | 85.10 | 85.71 | 77.74 | 76.91 | 76.33 | 83.48 | 82.01 | 73.46 | 83.56 | 74.92 | 79.92 |
| ALL | 84.60 | 90.02 | 85.75 | 83.59 | 81.27 | 84.81 | 87.10 | 84.88 | 87.10 | 84.61 | 85.37 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 90.40 | 92.20 | 91.75 | 90.19 | 92.11 | 93.44 | 81.18 | 93.23 | 85.12 | 90.37 | 90.00 |
| ORG | 80.10 | 84.72 | 78.84 | 74.93 | 70.65 | 71.26 | 80.95 | 84.41 | 86.78 | 78.24 | 79.09 |
| LOC | 84.15 | 81.51 | 77.04 | 74.18 | 74.54 | 79.95 | 82.34 | 70.20 | 80.25 | 73.54 | 77.77 |
| ALL | 84.07 | 86.45 | 83.81 | 79.63 | 79.77 | 83.02 | 81.33 | 83.11 | 84.42 | 81.79 | 82.74 |

ตารางที่ ฉ-2 ค่าความครบถ้วนของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| WSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 82.78 | 91.50 | 82.43 | 89.32 | 83.36 | 88.03 | 82.05 | 86.74 | 88.93 | 89.86 | 86.50 |
| ORG | 73.84 | 83.08 | 74.50 | 70.83 | 61.90 | 68.42 | 80.59 | 76.68 | 82.17 | 70.31 | 74.23 |
| LOC | 68.76 | 81.08 | 65.85 | 63.19 | 70.33 | 71.99 | 70.62 | 62.50 | 76.73 | 76.62 | 70.77 |
| ALL | 74.23 | 85.41 | 75.97 | 74.38 | 72.50 | 77.59 | 78.80 | 75.63 | 83.05 | 78.80 | 77.64 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 85.82 | 91.75 | 86.38 | 90.19 | 83.96 | 92.48 | 82.89 | 88.64 | 91.07 | 91.85 | 88.50 |
| ORG | 78.17 | 85.43 | 75.79 | 75.94 | 63.67 | 70.57 | 83.66 | 81.77 | 85.66 | 73.59 | 77.43 |
| LOC | 75.04 | 83.40 | 66.94 | 64.33 | 70.51 | 74.45 | 72.21 | 64.11 | 78.66 | 77.60 | 72.72 |
| ALL | 78.94 | 87.07 | 78.31 | 77.04 | 73.33 | 80.78 | 80.67 | 78.63 | 85.64 | 81.13 | 80.15 |

ตารางที่ ฉ-3 ค่า F-measure ของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| WSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 85.71 | 92.86 | 87.18 | 90.91 | 87.91 | 90.43 | 86.03 | 89.98 | 89.54 | 91.09 | 89.16 |
| ORG | 77.44 | 85.89 | 78.37 | 75.34 | 66.67 | 71.23 | 83.56 | 81.08 | 84.34 | 75.95 | 77.99 |
| LOC | 76.06 | 83.33 | 71.30 | 69.38 | 73.21 | 77.31 | 75.89 | 67.54 | 80.00 | 75.76 | 74.98 |
| ALL | 79.08 | 87.66 | 80.57 | 78.71 | 76.63 | 81.04 | 82.74 | 79.99 | 85.02 | 81.60 | 81.30 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 88.05 | 91.97 | 88.98 | 90.19 | 87.84 | 92.96 | 82.03 | 90.87 | 87.99 | 91.11 | 89.20 |
| ORG | 79.12 | 85.07 | 77.28 | 75.43 | 66.98 | 70.91 | 82.28 | 83.07 | 86.22 | 75.85 | 78.22 |
| LOC | 79.34 | 82.44 | 71.64 | 68.90 | 72.47 | 77.10 | 76.94 | 67.02 | 79.45 | 75.51 | 75.08 |
| ALL | 81.43 | 86.76 | 80.97 | 78.31 | 76.41 | 81.88 | 81.00 | 80.81 | 85.03 | 81.46 | 81.40 |

<u>ข้อมูลแบบตัดพยางค์</u>

ตารางที่ ฉ-4 ค่าความแม่นยำของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| SSG | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 91.20 | 93.86 | 95.28 | 93.89 | 90.77 | 93.73 | 92.71 | 90.94 | 90.08 | 93.01 | 92.55 |
| ORG | 83.48 | 87.54 | 82.11 | 81.41 | 71.61 | 75.19 | 84.21 | 83.69 | 87.43 | 81.69 | 81.84 |
| LOC | 85.47 | 82.06 | 74.46 | 78.37 | 78.79 | 84.26 | 82.22 | 74.75 | 81.22 | 75.32 | 79.69 |
| ALL | 86.05 | 88.46 | 85.85 | 84.69 | 81.12 | 85.64 | 86.89 | 83.77 | 86.70 | 84.56 | 85.37 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 89.30 | 82.72 | 77.93 | 80.16 | 70.61 | 84.20 | 68.67 | 74.45 | 83.33 | 89.63 | 80.10 |
| ORG | 82.15 | 81.66 | 77.96 | 73.76 | 67.86 | 68.13 | 78.91 | 82.97 | 84.28 | 77.74 | 77.54 |
| LOC | 81.94 | 81.04 | 72.92 | 71.13 | 75.97 | 81.25 | 81.55 | 68.97 | 77.42 | 73.07 | 76.53 |
| ALL | 83.78 | 81.90 | 77.02 | 75.25 | 71.39 | 78.66 | 75.06 | 76.02 | 82.10 | 81.22 | 78.24 |

ตารางที่ ฉ-5 ค่าความครบถ้วนของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| SSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 78.73 | 92.72 | 82.56 | 88.79 | 81.11 | 89.40 | 80.22 | 87.50 | 88.16 | 89.13 | 85.83 |
| ORG | 77.05 | 82.41 | 72.35 | 74.73 | 60.49 | 69.62 | 78.21 | 76.68 | 81.80 | 72.50 | 74.58 |
| LOC | 70.86 | 83.01 | 66.12 | 63.95 | 70.68 | 71.01 | 72.67 | 60.28 | 75.99 | 75.32 | 70.99 |
| ALL | 75.24 | 85.88 | 75.25 | 76.00 | 71.33 | 78.23 | 77.68 | 75.20 | 82.43 | 79.20 | 77.64 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 80.25 | 92.96 | 87.06 | 90.54 | 81.41 | 93.85 | 81.77 | 89.96 | 91.26 | 92.39 | 88.15 |
| ORG | 79.78 | 84.25 | 74.50 | 77.82 | 63.67 | 70.57 | 81.01 | 81.61 | 84.74 | 75.31 | 77.33 |
| LOC | 75.22 | 84.17 | 66.94 | 65.46 | 71.73 | 73.46 | 74.49 | 62.30 | 77.42 | 76.62 | 72.78 |
| ALL | 78.25 | 87.07 | 78.09 | 78.23 | 72.77 | 81.06 | 79.76 | 78.44 | 85.02 | 81.87 | 80.06 |

ตารางที่ ฉ-6 ค่า F-measure ของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| SSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 84.51 | 93.28 | 88.47 | 91.27 | 85.67 | 91.51 | 86.02 | 89.19 | 89.11 | 91.03 | 89.01 |
| ORG | 80.13 | 84.90 | 76.92 | 77.93 | 65.58 | 72.30 | 81.10 | 80.03 | 84.52 | 76.82 | 78.02 |
| LOC | 77.48 | 82.53 | 70.04 | 70.43 | 74.52 | 77.07 | 77.15 | 66.74 | 78.52 | 75.32 | 74.98 |
| ALL | 80.28 | 87.15 | 80.20 | 80.11 | 75.91 | 81.76 | 82.02 | 79.25 | 84.51 | 81.79 | 81.30 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 84.53 | 87.54 | 82.24 | 85.03 | 75.63 | 88.76 | 74.65 | 81.48 | 87.12 | 90.99 | 83.80 |
| ORG | 80.94 | 82.93 | 76.19 | 75.74 | 65.70 | 69.33 | 79.94 | 82.28 | 84.51 | 76.51 | 77.41 |
| LOC | 78.43 | 82.58 | 69.80 | 68.18 | 73.79 | 77.16 | 77.86 | 65.47 | 77.42 | 74.80 | 74.55 |
| ALL | 80.92 | 84.40 | 77.55 | 76.71 | 72.07 | 79.85 | 77.34 | 77.22 | 83.53 | 81.54 | 79.11 |

## ผลการทดสอบโดยประเมินจากจำนวน token

ข้อมูลแบบตัดคำ

ตารางที่ ฉ-7 ค่าความแม่นยำของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| WSG | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (token) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 86.86 | 94.96 | 96.29 | 95.94 | 96.42 | 96.67 | 95.65 | 95.45 | 94.65 | 92.73 | 94.56 |
| ORG | 82.20 | 90.38 | 83.03 | 80.06 | 74.83 | 76.06 | 85.73 | 77.98 | 87.96 | 83.75 | 82.20 |
| LOC | 89.64 | 86.18 | 88.72 | 83.64 | 82.97 | 87.93 | 92.59 | 86.51 | 88.76 | 75.09 | 86.20 |
| ALL | 86.08 | 91.95 | 90.82 | 87.95 | 86.85 | 89.76 | 92.29 | 87.59 | 91.57 | 86.55 | 89.14 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 93.78 | 95.72 | 96.12 | 95.95 | 96.27 | 97.81 | 94.19 | 96.80 | 93.44 | 91.70 | 95.18 |
| ORG | 81.55 | 87.95 | 81.13 | 77.81 | 73.58 | 75.24 | 82.54 | 77.95 | 87.80 | 82.29 | 80.78 |
| LOC | 89.04 | 84.43 | 87.87 | 82.23 | 81.88 | 86.47 | 92.63 | 84.77 | 87.17 | 74.88 | 85.14 |
| ALL | 88.46 | 91.08 | 89.96 | 86.69 | 86.04 | 89.55 | 90.56 | 87.54 | 90.53 | 85.51 | 88.59 |

ตารางที่ ฉ-8 ค่าความครบถ้วนของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| WSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (token) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 94.07 | 95.77 | 92.09 | 94.93 | 88.95 | 95.40 | 89.99 | 90.75 | 94.96 | 94.11 | 93.10 |
| ORG | 80.03 | 81.16 | 75.58 | 71.02 | 69.33 | 72.00 | 78.39 | 75.08 | 76.96 | 72.81 | 75.24 |
| LOC | 70.85 | 82.65 | 76.34 | 77.43 | 78.59 | 78.32 | 78.95 | 65.52 | 77.97 | 76.08 | 76.27 |
| ALL | 82.64 | 88.22 | 83.97 | 82.63 | 80.81 | 85.61 | 84.45 | 78.44 | 85.82 | 83.46 | 83.61 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 93.95 | 95.88 | 93.99 | 95.35 | 89.13 | 93.69 | 91.01 | 92.22 | 94.87 | 94.80 | 93.49 |
| ORG | 83.41 | 83.22 | 76.45 | 74.09 | 70.21 | 73.69 | 80.19 | 78.67 | 78.62 | 76.87 | 77.54 |
| LOC | 73.85 | 84.29 | 76.73 | 77.94 | 78.90 | 81.21 | 79.45 | 66.24 | 78.88 | 76.68 | 77.42 |
| ALL | 84.56 | 89.30 | 85.27 | 83.97 | 81.22 | 85.84 | 85.58 | 80.31 | 86.46 | 85.30 | 84.78 |

ตารางที่ ฉ-9 ค่า F-measure ของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| WSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (token) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 90.32 | 95.36 | 94.14 | 95.43 | 92.53 | 96.03 | 92.74 | 93.04 | 94.81 | 93.41 | 93.78 |
| ORG | 81.10 | 85.52 | 79.13 | 75.27 | 71.97 | 73.98 | 81.89 | 76.51 | 82.09 | 77.90 | 78.54 |
| LOC | 79.14 | 84.38 | 82.06 | 80.41 | 80.72 | 82.85 | 85.23 | 74.57 | 83.02 | 75.58 | 80.80 |
| ALL | 84.33 | 90.05 | 87.26 | 85.21 | 83.72 | 87.64 | 88.20 | 82.76 | 88.60 | 84.98 | 86.27 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 93.87 | 95.80 | 95.04 | 95.65 | 92.57 | 95.70 | 92.57 | 94.46 | 94.15 | 93.22 | 94.30 |
| ORG | 82.47 | 85.52 | 78.72 | 75.91 | 71.86 | 74.46 | 81.35 | 78.31 | 82.96 | 79.49 | 79.10 |
| LOC | 80.74 | 84.36 | 81.92 | 80.03 | 80.36 | 83.76 | 85.54 | 74.36 | 82.82 | 75.77 | 80.97 |
| ALL | 86.47 | 90.18 | 87.55 | 85.31 | 83.56 | 87.65 | 88.00 | 83.77 | 88.45 | 85.40 | 86.63 |

<u>ข้อมูลแบบตัดพยางค์</u>

ตารางที่ ฉ-10 ค่าความแม่นยำของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| SSG | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (token) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 94.34 | 95.15 | 98.83 | 96.27 | 95.94 | 97.82 | 98.13 | 93.80 | 94.15 | 93.35 | 95.78 |
| ORG | 86.33 | 91.74 | 82.10 | 86.85 | 76.52 | 80.40 | 86.33 | 83.36 | 88.48 | 82.29 | 84.44 |
| LOC | 89.56 | 85.27 | 86.13 | 87.18 | 83.93 | 89.63 | 91.10 | 85.87 | 87.50 | 76.06 | 86.22 |
| ALL | 90.05 | 92.11 | 90.52 | 90.57 | 86.78 | 91.23 | 92.65 | 88.26 | 90.81 | 86.15 | 89.91 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 94.11 | 92.33 | 91.04 | 91.47 | 86.60 | 94.74 | 90.13 | 87.09 | 91.43 | 92.51 | 91.14 |
| ORG | 84.74 | 86.79 | 79.82 | 83.72 | 75.17 | 76.22 | 82.50 | 83.05 | 86.55 | 80.93 | 81.95 |
| LOC | 88.27 | 84.23 | 84.69 | 81.98 | 81.38 | 87.77 | 90.22 | 82.46 | 85.13 | 74.66 | 84.08 |
| ALL | 88.96 | 88.72 | 85.96 | 86.32 | 81.83 | 88.01 | 87.49 | 84.64 | 88.43 | 84.92 | 86.53 |

ตารางที่ ฉ-11 ค่าความครบถ้วนของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| SSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (token) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 93.25 | 96.97 | 90.31 | 94.11 | 87.95 | 95.45 | 89.11 | 92.34 | 93.76 | 94.89 | 92.82 |
| ORG | 83.33 | 83.18 | 76.08 | 77.24 | 70.70 | 77.33 | 81.97 | 78.40 | 80.37 | 78.52 | 78.71 |
| LOC | 72.71 | 83.06 | 78.18 | 77.93 | 77.43 | 77.82 | 79.67 | 60.44 | 78.20 | 69.37 | 75.48 |
| ALL | 83.46 | 88.78 | 83.05 | 83.57 | 79.88 | 86.50 | 84.84 | 78.72 | 85.56 | 83.84 | 83.82 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 93.35 | 97.15 | 92.45 | 95.38 | 88.09 | 96.49 | 91.46 | 93.36 | 94.63 | 95.32 | 93.77 |
| ORG | 85.64 | 84.08 | 78.02 | 79.42 | 72.05 | 78.06 | 83.59 | 81.33 | 82.10 | 80.96 | 80.53 |
| LOC | 75.80 | 84.14 | 78.55 | 79.05 | 77.95 | 80.86 | 80.76 | 61.43 | 79.23 | 70.66 | 76.84 |
| ALL | 85.26 | 89.42 | 84.82 | 85.17 | 80.50 | 87.89 | 86.70 | 80.40 | 86.76 | 85.23 | 85.21 |

ตารางที่ ฉ-12 ค่า F-measure ของการทดสอบแบบจำลองก่อนและหลังการประมวลผลภายหลังทั้ง 10 ครั้ง

| SSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (token) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| ก่อนการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 93.79 | 96.05 | 94.38 | 95.18 | 91.77 | 96.62 | 93.40 | 93.06 | 93.96 | 94.12 | 94.23 |
| ORG | 84.80 | 87.25 | 78.97 | 81.77 | 73.49 | 78.83 | 84.09 | 80.80 | 84.23 | 80.36 | 81.46 |
| LOC | 80.26 | 84.15 | 81.96 | 82.30 | 80.55 | 83.31 | 85.00 | 70.94 | 82.59 | 72.56 | 80.36 |
| ALL | 86.63 | 90.41 | 86.63 | 86.93 | 83.19 | 88.80 | 88.57 | 83.22 | 88.11 | 84.98 | 86.75 |
| หลังการประมวลผลภายหลัง | | | | | | | | | | | |
| PER | 93.73 | 94.68 | 91.74 | 93.38 | 87.34 | 95.60 | 90.79 | 90.12 | 93.00 | 93.89 | 92.43 |
| ORG | 85.19 | 85.42 | 78.91 | 81.51 | 73.58 | 77.13 | 83.04 | 82.18 | 84.27 | 80.95 | 81.22 |
| LOC | 81.56 | 84.19 | 81.50 | 80.49 | 79.63 | 84.18 | 85.23 | 70.41 | 82.07 | 72.61 | 80.19 |
| ALL | 87.07 | 89.07 | 85.39 | 85.74 | 81.16 | 87.95 | 87.09 | 82.47 | 87.58 | 85.07 | 85.86 |

# ภาคผนวก ช
## ผลการทดสอบข้อมูลทั้ง 10 ครั้งเมื่อใช้คุณสมบัติคำบริบทช่วงต่าง ๆ

ข้อมูลแบบตัดคำ

ตารางที่ ช-1 ค่าความแม่นยำของการทดสอบแบบจำลองทั้ง 10 ครั้งเมื่อใช้คุณสมบัติคำบริบทช่วงต่าง ๆ

| WSG | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram + ช่วงคำบริบท 3 คำ | | | | | | | | | | | |
| PER | 85.96 | 91.07 | 92.87 | 90.64 | 89.77 | 88.21 | 89.03 | 88.84 | 87.70 | 88.81 | 89.29 |
| ORG | 85.69 | 88.25 | 83.16 | 83.02 | 79.25 | 79.35 | 86.71 | 88.63 | 88.71 | 83.54 | 84.63 |
| LOC | 86.68 | 85.71 | 77.35 | 75.87 | 79.07 | 85.76 | 83.05 | 77.17 | 81.77 | 76.62 | 80.91 |
| ALL | 86.09 | 88.76 | 86.21 | 83.81 | 83.43 | 85.06 | 86.79 | 85.32 | 86.41 | 84.27 | 85.62 |
| คุณสมบัติ unigram และ bigram + ช่วงคำบริบท 2 คำ | | | | | | | | | | | |
| PER | 87.13 | 90.15 | 91.94 | 90.32 | 89.72 | 89.35 | 88.96 | 91.13 | 86.33 | 90.67 | 89.57 |
| ORG | 84.74 | 88.43 | 83.42 | 84.38 | 78.97 | 78.55 | 86.00 | 88.84 | 88.29 | 83.93 | 84.55 |
| LOC | 86.85 | 84.38 | 76.74 | 75.62 | 78.96 | 86.58 | 83.48 | 77.94 | 82.16 | 77.01 | 80.97 |
| ALL | 86.08 | 88.25 | 85.78 | 84.13 | 83.26 | 85.50 | 86.59 | 86.41 | 85.87 | 85.27 | 85.71 |
| คุณสมบัติ unigram และ bigram + ช่วงคำบริบท 1 คำ | | | | | | | | | | | |
| PER | 89.64 | 88.37 | 90.87 | 90.82 | 92.06 | 87.90 | 87.21 | 92.39 | 86.88 | 92.56 | 89.87 |
| ORG | 85.94 | 88.29 | 82.59 | 84.06 | 81.23 | 78.87 | 88.12 | 88.34 | 89.42 | 85.43 | 85.23 |
| LOC | 86.84 | 84.79 | 77.97 | 79.13 | 79.21 | 85.85 | 83.33 | 76.19 | 83.28 | 78.65 | 81.53 |
| ALL | 87.22 | 87.64 | 85.21 | 85.13 | 84.85 | 84.78 | 86.67 | 86.20 | 86.75 | 86.97 | 86.14 |

ตารางที่ ช-2 ค่าความครบถ้วนของการทดสอบแบบจำลองทั้ง 10 ครั้งเมื่อใช้คุณสมบัติคำบริบทช่วงต่าง ๆ

| WSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram + ช่วงคำบริบท 3 คำ | | | | | | | | | | | |
| PER | 75.95 | 89.08 | 81.61 | 84.76 | 81.56 | 81.88 | 76.30 | 82.95 | 85.83 | 87.68 | 82.76 |
| ORG | 70.14 | 76.72 | 68.62 | 65.05 | 59.26 | 64.35 | 72.91 | 67.82 | 69.30 | 61.88 | 67.61 |
| LOC | 64.75 | 71.81 | 60.66 | 57.87 | 65.27 | 68.06 | 66.97 | 62.70 | 71.22 | 69.16 | 65.85 |
| ALL | 69.64 | 79.73 | 72.30 | 69.11 | 69.40 | 72.70 | 72.81 | 71.16 | 75.65 | 72.87 | 72.54 |
| คุณสมบัติ unigram และ bigram + ช่วงคำบริบท 2 คำ | | | | | | | | | | | |
| PER | 75.44 | 88.83 | 80.79 | 84.94 | 81.11 | 81.71 | 76.86 | 85.61 | 85.83 | 88.04 | 82.92 |
| ORG | 69.50 | 75.54 | 69.20 | 65.32 | 59.61 | 64.83 | 72.91 | 66.67 | 69.30 | 62.03 | 67.49 |
| LOC | 64.57 | 72.97 | 60.38 | 58.25 | 66.14 | 66.58 | 66.74 | 64.11 | 69.73 | 68.51 | 65.80 |
| ALL | 69.20 | 79.34 | 72.14 | 69.38 | 69.62 | 72.34 | 72.97 | 72.01 | 75.24 | 72.93 | 72.52 |

| คุณสมบัติ unigram และ bigram + ช่วงคำบริบท 1 คำ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PER | 76.71 | 86.65 | 78.61 | 83.19 | 79.91 | 79.49 | 74.61 | 85.04 | 84.85 | 87.86 | 81.69 |
| ORG | 70.63 | 74.54 | 68.62 | 65.19 | 58.02 | 63.40 | 72.49 | 67.16 | 68.38 | 61.41 | 66.98 |
| LOC | 65.62 | 71.04 | 60.93 | 59.01 | 66.49 | 65.60 | 67.20 | 61.29 | 70.47 | 68.18 | 65.58 |
| ALL | 70.33 | 77.76 | 71.13 | 69.00 | 68.79 | 70.71 | 72.06 | 71.16 | 74.76 | 72.53 | 71.82 |

ตารางที่ ช-3 ค่า F-measure ของการทดสอบแบบจำลองทั้ง 10 ครั้งเมื่อใช้คุณสมบัติคำบริบทช่วงต่าง ๆ

| WSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram + ช่วงคำบริบท 3 คำ | | | | | | | | | | | |
| PER | 80.65 | 90.06 | 86.87 | 87.60 | 85.47 | 84.93 | 82.18 | 85.80 | 86.75 | 88.24 | 85.86 |
| ORG | 77.14 | 82.08 | 75.20 | 72.95 | 67.81 | 71.07 | 79.21 | 76.84 | 77.81 | 71.10 | 75.12 |
| LOC | 74.13 | 78.15 | 67.99 | 65.66 | 71.51 | 75.89 | 74.15 | 69.19 | 76.13 | 72.70 | 72.55 |
| ALL | 77.00 | 84.00 | 78.64 | 75.75 | 75.77 | 78.39 | 79.18 | 77.60 | 80.67 | 78.16 | 78.52 |
| คุณสมบัติ unigram และ bigram + ช่วงคำบริบท 2 คำ | | | | | | | | | | | |
| PER | 80.87 | 89.49 | 86.00 | 87.55 | 85.20 | 85.36 | 82.47 | 88.28 | 86.08 | 89.34 | 86.06 |
| ORG | 76.37 | 81.48 | 75.65 | 73.64 | 67.94 | 71.04 | 78.91 | 76.17 | 77.65 | 71.34 | 75.02 |
| LOC | 74.07 | 78.26 | 67.58 | 65.81 | 71.98 | 75.28 | 74.18 | 70.35 | 75.44 | 72.51 | 72.55 |
| ALL | 76.72 | 83.55 | 78.37 | 76.05 | 75.83 | 78.37 | 79.20 | 78.56 | 80.20 | 78.62 | 78.55 |
| คุณสมบัติ unigram และ bigram + ช่วงคำบริบท 1 คำ | | | | | | | | | | | |
| PER | 82.67 | 87.50 | 84.30 | 86.84 | 85.55 | 83.48 | 80.42 | 88.56 | 85.85 | 90.15 | 85.53 |
| ORG | 77.53 | 80.84 | 74.96 | 73.43 | 67.70 | 70.29 | 79.54 | 76.31 | 77.50 | 71.45 | 74.95 |
| LOC | 74.75 | 77.31 | 68.40 | 67.61 | 72.30 | 74.37 | 74.40 | 67.93 | 76.34 | 73.04 | 72.65 |
| ALL | 77.87 | 82.41 | 77.54 | 76.22 | 75.98 | 77.11 | 78.69 | 77.96 | 80.31 | 79.10 | 78.32 |

ข้อมูลแบบตัดพยางค์

ตารางที่ ช-4 ค่าความแม่นยำของการทดสอบแบบจำลองทั้ง 10 ครั้งเมื่อใช้คุณสมบัติคำบริบทช่วงต่าง ๆ

| SSG | Precision (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram + ช่วงบริบท 4 พยางค์ | | | | | | | | | | | |
| PER | 89.77 | 93.63 | 93.87 | 91.85 | 89.50 | 92.47 | 91.43 | 91.20 | 90.22 | 92.79 | 91.67 |
| ORG | 83.67 | 87.55 | 83.98 | 81.61 | 74.71 | 74.79 | 82.64 | 84.29 | 84.83 | 83.05 | 82.11 |
| LOC | 85.74 | 81.85 | 73.60 | 78.45 | 78.12 | 84.23 | 83.60 | 75.38 | 81.77 | 71.05 | 79.38 |
| ALL | 85.92 | 88.44 | 86.03 | 84.30 | 81.66 | 85.15 | 86.13 | 84.23 | 86.03 | 84.14 | 85.20 |
| คุณสมบัติ unigram และ bigram + ช่วงบริบท 3 พยางค์ | | | | | | | | | | | |
| PER | 90.64 | 93.63 | 94.14 | 92.68 | 88.31 | 91.52 | 91.12 | 92.29 | 90.46 | 92.56 | 91.73 |
| ORG | 83.96 | 87.97 | 84.41 | 81.36 | 77.62 | 73.84 | 84.60 | 85.74 | 85.59 | 83.72 | 82.88 |
| LOC | 86.55 | 81.38 | 73.31 | 77.78 | 77.35 | 84.89 | 82.71 | 74.08 | 81.79 | 73.56 | 79.34 |
| ALL | 86.56 | 88.54 | 86.16 | 84.28 | 81.75 | 84.65 | 86.62 | 84.65 | 86.38 | 84.94 | 85.45 |
| คุณสมบัติ unigram และ bigram + ช่วงบริบท 2 พยางค์ | | | | | | | | | | | |
| PER | 90.50 | 92.74 | 94.28 | 92.80 | 88.44 | 91.77 | 90.98 | 92.17 | 90.76 | 92.05 | 91.65 |
| ORG | 84.50 | 87.86 | 85.47 | 81.82 | 78.35 | 74.17 | 84.89 | 85.86 | 86.87 | 84.51 | 83.43 |
| LOC | 86.96 | 81.45 | 72.55 | 78.36 | 77.87 | 84.59 | 82.80 | 74.15 | 81.44 | 72.88 | 79.30 |
| ALL | 86.88 | 88.22 | 86.52 | 84.63 | 82.22 | 84.80 | 86.71 | 84.68 | 86.85 | 84.82 | 85.63 |

ตารางที่ ช-5 ค่าความครบถ้วนของการทดสอบแบบจำลองทั้ง 10 ครั้งเมื่อใช้คุณสมบัติคำบริบทช่วงต่าง ๆ

| SSG | Recall (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram + ช่วงบริบท 4 พยางค์ | | | | | | | | | | | |
| PER | 77.72 | 92.72 | 81.34 | 86.87 | 80.51 | 88.21 | 77.84 | 86.36 | 87.77 | 88.59 | 84.79 |
| ORG | 73.19 | 78.89 | 71.35 | 70.97 | 56.26 | 65.31 | 75.14 | 72.25 | 72.98 | 68.13 | 70.45 |
| LOC | 70.33 | 78.38 | 60.93 | 59.39 | 66.67 | 69.53 | 70.84 | 60.48 | 73.45 | 70.13 | 68.01 |
| ALL | 73.29 | 83.28 | 73.30 | 72.58 | 68.51 | 76.03 | 75.16 | 73.24 | 78.32 | 76.07 | 74.98 |
| คุณสมบัติ unigram และ bigram + ช่วงบริบท 3 พยางค์ | | | | | | | | | | | |
| PER | 78.48 | 92.72 | 80.93 | 86.51 | 79.31 | 88.55 | 77.70 | 86.17 | 88.35 | 87.86 | 84.66 |
| ORG | 72.23 | 78.39 | 71.35 | 70.97 | 57.50 | 64.83 | 74.44 | 71.10 | 73.16 | 68.28 | 70.23 |
| LOC | 69.63 | 77.61 | 62.30 | 58.44 | 67.36 | 69.04 | 70.84 | 61.09 | 74.69 | 70.45 | 68.15 |
| ALL | 72.85 | 82.89 | 73.41 | 72.20 | 68.68 | 75.89 | 74.84 | 72.93 | 78.93 | 75.93 | 74.86 |

| คุณสมบัติ unigram และ bigram + ช่วงบริบท 2 พยางค์ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PER | 77.22 | 92.96 | 80.79 | 85.81 | 79.16 | 87.69 | 77.84 | 86.93 | 87.77 | 88.04 | 84.42 |
| ORG | 71.75 | 77.55 | 70.77 | 70.16 | 56.79 | 63.88 | 73.74 | 69.79 | 72.98 | 67.34 | 69.48 |
| LOC | 69.81 | 77.99 | 60.66 | 59.77 | 66.32 | 68.80 | 71.30 | 61.29 | 72.95 | 72.40 | 68.13 |
| ALL | 72.41 | 82.65 | 72.80 | 72.04 | 68.07 | 75.18 | 74.73 | 72.75 | 78.18 | 76.00 | 74.48 |

ตารางที่ ช-6 ค่า F-measure ของการทดสอบแบบจำลองทั้ง 10 ครั้งเมื่อใช้คุณสมบัติคำบริบทช่วงต่าง ๆ

| SSG | F-measure (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| คุณสมบัติ unigram และ bigram + ช่วงบริบท 4 พยางค์ | | | | | | | | | | | |
| PER | 83.31 | 93.17 | 87.15 | 89.29 | 84.77 | 90.29 | 84.09 | 88.72 | 88.98 | 90.64 | 88.04 |
| ORG | 78.08 | 83.00 | 77.15 | 75.92 | 64.19 | 69.73 | 78.71 | 77.81 | 78.46 | 74.85 | 75.79 |
| LOC | 77.28 | 80.08 | 66.67 | 67.60 | 71.94 | 76.18 | 76.70 | 67.11 | 77.39 | 70.59 | 73.15 |
| ALL | 79.10 | 85.78 | 79.16 | 78.00 | 74.51 | 80.33 | 80.27 | 78.35 | 81.99 | 79.90 | 79.74 |
| คุณสมบัติ unigram และ bigram + ช่วงบริบท 3 พยางค์ | | | | | | | | | | | |
| PER | 84.12 | 93.17 | 87.03 | 89.49 | 83.57 | 90.01 | 83.88 | 89.13 | 89.39 | 90.15 | 87.99 |
| ORG | 77.65 | 82.91 | 77.33 | 75.81 | 66.06 | 69.04 | 79.20 | 77.74 | 78.89 | 75.22 | 75.98 |
| LOC | 77.18 | 79.45 | 67.36 | 66.74 | 72.01 | 76.15 | 76.32 | 66.96 | 78.08 | 71.97 | 73.22 |
| ALL | 79.11 | 85.62 | 79.28 | 77.78 | 74.65 | 80.03 | 80.30 | 78.36 | 82.49 | 80.18 | 79.78 |
| คุณสมบัติ unigram และ bigram + ช่วงบริบท 2 พยางค์ | | | | | | | | | | | |
| PER | 83.33 | 92.85 | 87.01 | 89.17 | 83.54 | 89.69 | 83.90 | 89.47 | 89.24 | 90.00 | 87.82 |
| ORG | 77.60 | 82.38 | 77.43 | 75.54 | 65.85 | 68.64 | 78.92 | 76.99 | 79.32 | 74.96 | 75.76 |
| LOC | 77.44 | 79.68 | 66.07 | 67.81 | 71.63 | 75.88 | 76.62 | 67.11 | 76.96 | 72.64 | 73.19 |
| ALL | 78.99 | 85.34 | 79.07 | 77.83 | 74.48 | 79.70 | 80.28 | 78.26 | 82.29 | 80.17 | 79.64 |

# ประวัติผู้เขียนวิทยานิพนธ์

นางสาว นัชชา ถิระสาโรช เกิดที่จังหวัดเชียงใหม่ สำเร็จการศึกษาระดับปริญญาตรี ศิลปศาสตรบัณฑิต สาขาภาษาอังกฤษ มหาวิทยาลัยเชียงใหม่ ในปีการศึกษา 2544 และเข้าศึกษาต่อในหลักสูตร อักษรศาสตรมหาบัณฑิต ภาควิชาภาษาศาสตร์ จุฬาลงกรณ์ลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2550