การพัฒนากระบวนการสกัดชื่อภูมิศาสตร์จากข้อความทวิตเตอร์ด้วยแบบจำลองคอนดิชันนอลแรนดอมฟิลด์

ธุวชิต แฉล้มเขตต์[1*] ชนินทร์ ทินนโชติ[2] อรรถพล ธำรงรัตนฤทธิ์[3]

[1,2]ภาควิชาวิศวกรรมสำรวจ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพมหานคร, ประเทศไทย*
[3]ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพมหานคร, ประเทศไทย

*ผู้ประพันธ์บรรณกิจ อีเมล : tuvachitchalamkate@gmail.com

รับต้นฉบับ: 18 พฤษภาคม 2565; รับบทความฉบับแก้ไข: 27 กรกฎาคม 2565; ตอบรับบทความ: 19 กันยายน 2565;
เผยแพร่ออนไลน์: 22 ธันวาคม 2565

บทคัดย่อ

ระบบนำทางแผนที่ออนไลน์ทั้งบนเว็บบราวเซอร์ โทรศัพท์เคลื่อนที่ และแพลตฟอร์มอื่น ๆ ปัจจุบันมีความสำคัญขึ้นอย่างมาก เนื่องจากการเพิ่มขึ้นของผู้ใช้งาน ผู้ให้บริการ โดยชื่อสถานที่หรือชื่อภูมิศาสตร์คือแหล่งข้อมูลสำคัญที่ผู้ใช้งานมักจะใช้เป็นคำสำคัญในการค้นหา รวมทั้งการจัดเก็บข้อมูลเหล่านี้ให้อยู่เป็นหมวดหมู่ต่าง ๆ ด้วย ซึ่งงานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างแบบจำลองที่สามารถสกัดเอาชื่อภูมิศาสตร์พร้อมทั้งจัดหมวดหมู่แบบอัตโนมัติโดยมีแหล่งข้อมูลสำคัญคือสื่อสังคมออนไลน์จากทวิตเตอร์ซึ่งเป็นแพลตฟอร์มที่ได้รับความนิยมในประเทศไทยแพลตฟอร์มหนึ่ง เป็นแหล่งข้อมูลข่าวสารที่มีความรวดเร็วและมีการอัพเดตข้อมูลตลอดเวลาทำให้มีโอกาสพบกับชื่อภูมิศาสตร์ที่เป็นสถานที่ใหม่ ๆ ซึ่งเป็นประโยชน์กับการรวบรวมข้อมูลภูมิสารสนเทศโดยที่ไม่จำเป็นต้องออกไปสำรวจข้อมูลภาคสนามเพียงอย่างเดียว ซึ่งเครื่องมือทางด้านการรู้จำชื่อเฉพาะในภาษาไทยนั้นไม่สามารถนำมาใช้ได้โดยตรงเนื่องจากมีการแบ่งประเภทของชื่อเฉพาะที่ไม่ได้จำแนกหมวดหมู่ของชื่อภูมิศาสตร์ไว้ โดยในส่วนของงานวิจัยนี้นำเอาอัลกอริทึมคอนดิชันนอลแรนดอมฟิลด์มาประยุกต์ใช้ร่วมกับคุณลักษณะทางภาษา เช่น คำบุพบทที่แสดงถึงสถานที่ (ใกล้ ไกล ถัดจาก ถัดไป ฯลฯ) คำนำหน้า เช่น โรงเรียน ตลาด วัด หมู่บ้าน ฯลฯ ซึ่งในงานวิจัยนี้มีการสร้างคลังข้อมูลภาษา (corpus) ขึ้นมาจากข้อความทวิตเตอร์ จำนวน 28,082 ข้อความ แบ่งเป็นข้อมูลสำหรับฝึกสอน 22,445 ข้อความ คิดเป็นร้อยละ 80 และข้อมูลสำหรับทดสอบจำนวน 5,617 ข้อความ คิดเป็นร้อยละ 20 โดยออกแบบการทดลองออกเป็น 2 กลุ่มหลัก ตามอัลกอริทึมที่ใช้ในการตัดคำ ซึ่งผลการศึกษาแบบจำลองที่ให้ค่าความถูกต้องโดยรวม (F1) สูงที่สุดคือ 0.946 ซึ่งให้ความถูกต้องโดยรวมเพียงพอสำหรับการนำไปใช้งานในส่วนที่เกี่ยวข้องต่อไป

คำสำคัญ **:** ภูมิสารสนเทศ การเรียนรู้ของเครื่องจักร การประมวลผลภาษาธรรมชาติ

The Development of Geo-Names Extraction from Twitter Texts Data by Conditional Random Fields

Tuvachit Chalamkate[1*] Chanin Thinnachote[2] Attapol Thamrongrattanarit[2]

[1,2]*Department of Survey Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand*
[3]*Department of Linguistics, Faculty of Arts, Chulalongkorn University, Bangkok, Thailand*

*Corresponding Author. E-mail address: tuvachitchalamkate@gmail.com

Received: 18 May 2022; Revised: 27 July 2022; Accepted: 19 September 2022;
Published online: 22 December 2022

Abstract

Navigation systems and online maps, Mobile application, and other platforms, are becoming increasingly important due to increasing users and providers. Place names or geonames (geographic names) are essential sources of information that users tend to use as keywords in their searches. Including storing these data in different categories. This research aims to create a model capable of extracting geonames and automatically categorizing them from the social media source of Twitter, one of the popular platforms in Thailand. It is a fast and always up-to-date information source, providing the opportunity to discover new geographic locations and helpful in gathering geospatial information without needing a field survey. Named-entity recognition standard tool cannot be used directly because of the classification of name entities that are not categorized by geographic names. As for the model, the conditional random field algorithm is applied to linguistic features such as place prepositions (near, far, next, next to, etc.) and prefixes, for instance, school, market, temples, villages, etc. This study, the Corpus was created from 28,082 Twitter messages, representing 80 percent of the 22,445 training set and 20 percent of the test set of 5,617 messages. According to the algorithm used to word tokenize, the experiment was designed into two main groups. The study result of the model with the highest overall accuracy (F1) was 0.946, which provided sufficient overall accuracy for relevant applications both on the web browser.

***Keywords*:** Geoinformatics, Machine learning, Natural language processing

1) บทนำ

ข้อมูลจากสื่อสังคมออนไลน์มีการเพิ่มขึ้นอย่างต่อเนื่องในแต่ละปี โดยมีทั้งข้อมูลที่เป็นข่าวสาร กระดานข่าว ข้อมูลที่เป็นการโต้ตอบระหว่างผู้ใช้งาน ประกาศ ฯลฯ ซึ่งประเทศไทยมีผู้รับข่าวสารจากสื่อสังคมออนไลน์เป็นจำนวนถึงร้อยละ 78 จากประชากรทั้งหมด เป็นสัดส่วนที่มากที่สุดในโลก โดยทวิตเตอร์มีผู้ใช้งานในประเทศไทยอยู่อันดับที่ 10 ของโลกและเป็นแพลตฟอร์มที่ได้รับความนิยมในการรับรู้ข่าวสารรองจาก Facebook [1] ดังนั้นจึงเป็นไปได้ว่ามีข้อมูลอีกมากมายที่เราสามารถนำมาใช้ประโยชน์ได้ในหลากหลายวัตถุประสงค์ โดยพบว่าข้อมูลดิจิตอลเหล่านี้ทั้ง สื่อสังคมออนไลน์ เว็บบล็อก ฯลฯ กว่าร้อยละ 60 มีการอ้างอิงเชิงตำแหน่ง [2] การสกัดเอาข้อมูลที่มีการกล่าวถึงหรืออ้างถึงในเชิงตำแหน่ง เช่น ชื่อสถานที่ หรือบริบทที่ทำให้เข้าใจได้ว่าเป็นการสื่อถึงสถานที่ใดสถานที่หนึ่ง ในปัจจุบันมีการศึกษาวิจัยในภาษาต่าง ๆ โดยเฉพาะภาษาอังกฤษ ซึ่งเรียกกระบวนการนี้ว่า การสกัดข้อมูลทางภูมิศาสตร์ (Geographic Information Retrieval : GIR) ประกอบไปด้วย 2 ขั้นตอนที่สำคัญ คือ 1) การรู้จำภูมินาม (toponym recognition, toponym extraction, geotagging) และ 2) การเข้ารหัสทางภูมิศาสตร์ (Geocoding, Toponym resolution, Toponym ambiguation) สำหรับงานวิจัยนี้จะพัฒนาแบบจำลองในส่วนแรกซึ่งเป็นขั้นตอนที่ใช้ในการสกัดข้อมูลชื่อเฉพาะที่เป็นข้อมูลชื่อสถานที่ หรือชื่อทางภูมิศาสตร์ออกมาจากข้อความ [3]

จากเนื้อความข้างต้นเครื่องมือสำคัญในการสกัดชื่อเฉพาะหรือ Named Entities Recognition : NER เป็นหนึ่งในเครื่องมือที่นักภูมิสารสนเทศนำมาใช้เมื่อต้องประมวลผลข้อมูลจำนวนมากที่อยู่ในรูปของข้อความ เพื่อสกัดเอาชื่อเฉพาะ เช่น ชื่อบุคคล สถานที่ องค์กร วันและเวลา ชื่อยา หรือ คำเฉพาะที่เราสนใจ ฯลฯ โดยความท้าทายของการสกัดชื่อภูมิศาสตร์ออกจากข้อความคือ ความกำกวมที่อยู่ภายในประโยค เช่น ในตัวอย่างประโยค “Paris Hilton is in Paris” จากตัวอย่างข้างต้น Paris ที่เป็นชื่อคน และ Paris ที่เป็นชื่อเมืองเขียนนำด้วยอักษรตัวใหญ่เช่นเดียวกันทำให้ต้องพิจารณาถึงบริบทของคำในประโยค ซึ่งในที่นี้มีบุพบทที่ใช้คือ คำว่า “in” ทำให้สามารถแยกหน้าที่ของคำว่า “Paris” ที่เป็นชื่อบุคคลและชื่อสถานที่ได้ สำหรับภาษาไทยนั้นยังไม่พบว่ามีผู้พัฒนาเครื่องมือในการทำ NER เพื่อสกัดข้อมูลที่เป็นชื่อสถานที่โดยเฉพาะหรือ จำแนกสถานที่ออกมาจากข้อความ

เครื่องมือในการทำ NER ภาษาไทยที่ใช้กันแพร่หลายปัจจุบันคือ PyThaiNLP [4] แต่จำแนกประเภทของชื่อเฉพาะในกลุ่มหลักที่สำคัญ คือ ชื่อบุคคล ชื่อองค์กร ชื่อสถานที่ วัน เวลา URLs ฯลฯ ซึ่งจะพบว่าชื่อสถานที่นั้นไม่ได้แยกประเภทออกมาเป็นหมวดหมู่ทำให้การสกัดข้อมูลเพื่อที่จะเก็บรวบรวมข้อมูลชื่อสถานที่สำคัญหรือนำไปวิเคราะห์เชิงพื้นที่ (spatial analysis) อาทิเช่น หากต้องการนำไปใช้ในการเก็บข้อมูลชื่อสถานที่เพื่อทำฐานข้อมูลสำหรับระบบนำทาง พบว่ามักจะมีการแบ่งประเภทหมวดหมู่ในการให้บริการ หรือหากนำไปวิเคราะห์เชิงพื้นที่ การจำแนกข้อมูลที่เป็นขอบเขตการปกครอง ซึ่งเป็นข้อมูลแบบพื้นที่รูปปิด (polygon) หากแสดงในระบบภูมิสารสนเทศ และข้อมูลตำแหน่ง โรงเรียน วัด อาคารสูง ร้านค้า ฯลฯ ซึ่งเป็นข้อมูลรูปแบบจุด (point) โดยในกรณีหากข้อมูลชื่อเฉพาะมีความน่าจะเป็นชื่อสถานที่เหล่านั้นไม่สามารถหาค่าพิกัดอ้างอิงได้ การคาดคะเนโดยใช้ขอบเขตการปกครองมาช่วยจึงสามารถกำหนดขอบเขตของสถานที่เหล่านั้นได้แคบลง [5], [6]

เนื่องจากข้อมูลทวิตเตอร์มีปริมาณและความถี่ของข้อมูลที่เกิดขึ้นสูงมากในแต่ละวันทำให้มีโอกาสที่จะได้รับข้อมูล ข่าวสารใหม่ ๆ อยู่ตลอด [7] ดังนั้นสำหรับงานวิจัยนี้จึงมีวัตถุประสงค์ที่จะพัฒนากระบวนการสร้างเครื่องมือเพื่อสกัดข้อมูลชื่อภูมิศาสตร์จากสื่อสังคมออนไลน์คือ ทวิตเตอร์รวมทั้งจำแนกชนิดของชื่อเฉพาะเหล่านี้เพื่อให้สามารถนำข้อมูลไปใช้เป็นฐานข้อมูลอ้างอิงในการสำรวจภาคสนามต่าง ๆ รวมทั้งยังเป็นขั้นตอนแรกของการพัฒนาระบบแปลความหมายทางภูมิศาสตร์ให้มีความถูกต้องและใช้งานได้อย่างมีประสิทธิภาพ

2) ทบทวนวรรณกรรม

สำหรับงานวิจัยทางด้านการสกัดชื่อเฉพาะทางภูมิศาสตร์นั้นปัจจุบันยังอยู่ในวงที่จำกัดเนื่องจากเป็นการนำข้อมูลที่ได้ไปใช้ในงานทางด้านภูมิสารสนเทศซึ่งเป็นงานเฉพาะทางเป็นส่วนใหญ่ โดยเฉพาะสำหรับภาษาไทย ซึ่งวิธีการที่ใช้จะเป็นการนำเอาแนวทางการแก้ปัญหาทางด้านการรู้จำชื่อเฉพาะหรือ Named-Entity Recognition : NER มาประยุกต์ใช้งาน ซึ่งเนื้อหาในส่วนนี้จะนำเสนอใน 2 หัวข้อสำคัญ คือ 2.1) งานวิจัยทางด้านการรู้จำภูมินาม และ 2.2) งานวิจัยทางด้านการรู้จำชื่อเฉพาะซึ่งเป็นภาษาไทย

2.1) งานวิจัยทางด้านการรู้จำภูมินาม

ปัจจุบันมีการศึกษาวิจัยในด้านนี้ค่อนข้างจำกัดโดยหัวข้อที่สำคัญในแนวทางวิจัยนี้คือ แก้ไขปัญหาในเรื่องการรู้จำชื่อที่มีความกำกวม และชื่อที่ไม่ได้ถูกบรรจุอยู่ในอักขรานุกรมทางภูมิศาสตร์ การนำคลังคำสั่ง Stanford-NER มาปรับใช้เพื่อสร้างคุณสมบัติในการหาคำที่เป็นคำนามเฉพาะหรือ proper-nouns เนื่องจากชื่อทางภูมิศาสตร์นั้นก็จัดเป็นชื่อเฉพาะประเภทหนึ่งพร้อมทั้งใช้พจนานุกรมเอนทิตี้ซึ่งเป็นพจนานุกรมที่เก็บข้อมูลคุณสมบัติของชื่อเฉพาะต่าง ๆ เช่น ตำแหน่งที่ตั้ง (หากเป็นสถานที่) สี ฤดูกาล [8], [9] การนำอัลกอริทึม Conditional Random Fields : CRFs มาใช้ร่วมกับการกำหนดคุณลักษณะจาก หน้าที่ของคำ (POS-tagss) บริบทของคำในประโยคได้แก่ คำอุปสรรค (prefix) และคำปัจจัย (suffix) เพื่อนำมาสร้างแบบจำลองสกัดชื่อเฉพาะที่เป็นโบราณสถาน โดยเปรียบเทียบกับอีก 2 วิธีคือ SVM และ HMM โดยเปรียบเทียบความถูกต้องจาก google geocoding API ซึ่งพบว่า CRF ให้ผลลัพธ์เป็นอัตราการเชื่อมโยงข้อมูลระหว่างชื่อภูมิศาสตร์ที่สกัดออกด้วยแบบจำลองกับฐานข้อมูลอย่าง google geocoding API ได้จำนวนมากที่สุด [10] นอกจากนี้ยังมีการสร้างคุณลักษณะจากการสร้างกฎ (rule-based) ขึ้นมา เช่น คำว่า “street” หากเจอวลี “treet” แล้วพบตัวอักษรข้างหน้าเป็น [s, S] ให้ติดแท็กคำนั้นเป็น location [11]

2.2) งานวิจัยทางด้านการรู้จำชื่อเฉพาะภาษาไทย (Thai NER)

งานวิจัยทางด้านการรู้จำชื่อเฉพาะภาษาไทยยังมีอยู่จำกัดหากเทียบกับภาษาที่มีความนิยม เช่น ภาษาอังกฤษ ฝรั่งเศส หรือจีน ในเริ่มแรกของงานด้านนี้มีการใช้วิธีการทางสถิติ ร่วมกับแบบจำลองฮิวริสติก โดยสร้างกฎขึ้นมาจากบริบทของชื่อเฉพาะ เช่น คำสำคัญหรือคำที่อยู่ใกล้กับชื่อเฉพาะ ซึ่งผลที่ได้พบว่าใช้ได้ดีกับงานเขียนประเภทนิตยสารแต่หากเป็นหนังสือพิมพ์จะให้ผลที่ไม่ดีเท่าที่ควรเนื่องจากหนังสือพิมพ์จะมีการใช้สำนวนที่เป็นทางการน้อยกว่านิตยสาร [12] หลังจากนั้นมีการนำแบบจำลอง maximum entropy มาใช้ร่วมกับแบบจำลองฮิวริสติกและพจนานุกรมศัพท์เฉพาะซึ่งผ่านการตัดคำมาแล้ว โดยผลที่ได้พบว่าให้ผลดีกับชื่อบุคคล รองลงมาคือชื่อองค์กร แต่สำหรับชื่อสถานที่นั้นมีความถูกต้องน้อยที่สุดเนื่องจากมีความคลุมเครือมากกว่าชื่อประเภทอื่น [13] ในปี 2009 มีการนำเอาแบบจำลอง CRF มาใช้เพื่อสร้าง NER ภาษาไทย โดยใช้คลังข้อมูล “BEST 2009” ของ NECTEC จำนวน 90,000 คำ ในงานวิจัยนี้แบ่งข้อมูลออกเป็น 2 ส่วนคือ 1) ตัดคำ และ 2) ตัดพยางค์ (syllable segmentation) ซึ่งพบว่าให้ค่าความถูกต้องโดยรวมใกล้เคียงกันคือ 0.8039 และ 0.808 ตามลำดับ หากพิจารณาเฉพาะส่วนของชื่อสถานที่พบว่า ค่าความถูกต้องโดยรวมเป็น 0.7372 และ 0.7692 ตามลำดับ [14] สำหรับในปัจจุบันมีการนำโครงข่ายประสาทเทียมแบบวกกลับ คือ Bi-Directional Long Short Term Memory : Bi-LSTM ร่วมกับ CRF จำแนกเป็น 13 ชั้นข้อมูล โดยชื่อสถานที่ก็เป็นหนึ่งในจำนวนนั้น ซึ่งสำหรับชื่อสถานที่มีค่าความถูกต้องโดยรวมอยู่ที่ 0.8773 ซึ่งเป็นงานวิจัยที่ให้ค่าความถูกต้องสูงที่สุดในปัจจุบัน [15]

จากงานวิจัยที่ผ่านมาพบว่าการสร้างแบบจำลองเพื่อสกัดข้อมูลชื่อเฉพาะนั้นมีหลายงานวิจัยที่นำคุณลักษณะทางด้านภาษาและการใช้อักขรานุกรมหรือพจนานุกรมศัพท์เฉพาะมาใช้งาน แม้ในปัจจุบันจะมีการนำโครงข่ายประสาทเทียมมาประยุกต์ใช้งานแต่หากมีการสร้างคุณลักษณะเพิ่มไปด้วยก็มีแนวโน้มจะส่งผลให้แบบจำลองสามารถทำงานได้ดียิ่งขึ้นซึ่ง CRF มีข้อได้เปรียบแบบจำลองที่สร้างมาจากโครงข่ายประสาทเทียมในแง่ของเวลา (ทรัพยากร) ในการประมวลผล

3) แบบจำลอง Conditional Random Fields : CRFs

แบบจำลอง Conditional Random Fields (CRFs) เป็นแบบจำลองความน่าจะเป็นแบบหนึ่งที่นิยมใช้กับข้อมูลที่เป็นลำดับ (sequence) เช่น การประยุกต์ใช้กับการประมวลผลธรรมชาติโดยเฉพาะอย่างยิ่ง การทำ Named Entity Recognition (NER) ซึ่งแบบจำลองนี้พิสูจน์แล้วว่าให้ผลที่ดีกับการตัดคำ (segmentation) หรือติดป้าย (labelling) ให้กับข้อมูลที่เป็นลำดับ แบบจำลองนี้พัฒนามาจาก maximum entropy Markov models (MEMMs) โดย CRF สามารถลดข้อจำกัดของ MEMMs ลงได้ ซึ่งแบบจำลอง CRF เป็นแบบห่วงโซ่ตรงหรือ linear-chain CRF แสดงความสัมพันธ์ตามสมการที่ (1) และ (2) [16]

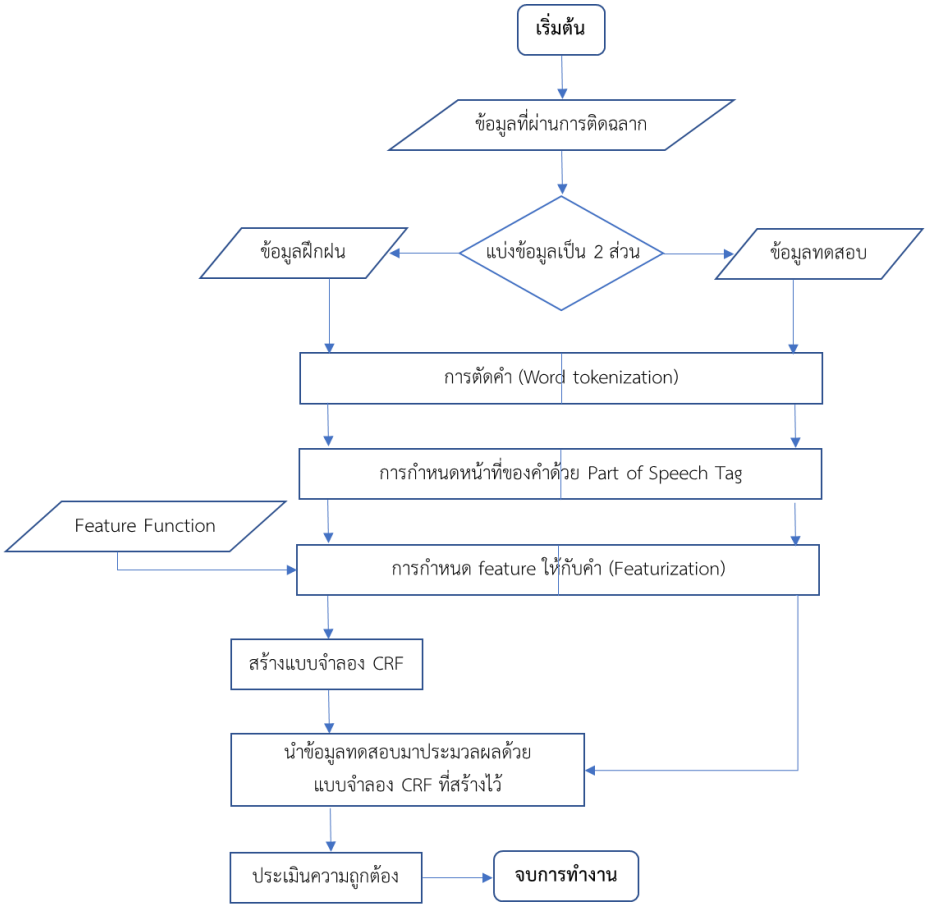
$$p(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} exp\left(\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right) \quad (1)$$

เมื่อ $f_k(y_t, y_{t-1}, x_t)$ คือฟังก์ชันคุณลักษณะ; x คือลำดับผลลัพธ์ค่าสังเกต; y คือลำดับผลลัพธ์ค่าที่ทำนาย; θ_k คือค่าน้ำหนักของฟังก์ชันคุณลักษณะ; *T, t* คือลำดับของเหตุการณ์ที่ต่อเนื่อง; *K, k* คือจำนวนของฟังก์ชันคุณลักษณะ; *Z(x)* คือการปรับข้อมูล (normalization) ซึ่งคำนวณได้จากสมการที่ (2)

$$Z(x) = \sum_{y} \prod_{t=1}^{T} exp\left(\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\right) \quad (2)$$

4) วิธีการดำเนินงานวิจัย

ในงานวิจัยชิ้นนี้มีการเก็บรวบรวมข้อมูลจากแหล่งข้อมูลทวิตเตอร์ จำนวน 28,082 ข้อความ ซึ่งใช้ twitter API เป็นเครื่องมือหลักในการดึงข้อมูลผ่านการโปรแกรมด้วยภาษา python จากสมมติฐานของงานวิจัยนี้ตัวแปรอิสระประกอบไปด้วยส่วนการดำเนินงาน ส่วนสำคัญในหัวข้อที่ 4.1) – 4.6) แสดงสรุปขั้นตอนตามรูปที่ 1



รูปที่ 1 : แสดงขั้นตอนการดำเนินงานสร้างแบบจำลองด้วย CRF

4.1) การออกแบบการทดลอง

งานวิจัยนี้มีสมมติฐานคือ CRF และฟังก์ชันคุณสมบัติที่นำมาใช้จะสามารถทำให้ได้แบบจำลองในการสกัดข้อมูลชื่อภูมิศาสตร์ที่มีประสิทธิภาพ (วัดประสิทธิภาพจาก F1-score) โดยตัวแปรอิสระของงานวิจัยนี้คือ ฟังก์ชันคุณสมบัติที่ส่งผลต่อประสิทธิภาพของแบบจำลอง โดยฟังก์ชันคุณสมบัติต่าง ๆ ที่ใช้แสดงในหัวข้อที่ 4.6 เช่น แบบจำลอง CRF ที่มีฟังก์ชันคุณสมบัติ POS-tagss + บุพบทที่บ่งบอกถึงสถานที่ ให้ค่า F1-score ที่น้อยกว่า POS-tagss + บุพบทที่บ่งบอกถึงสถานที่ + คำอุปสรรค (prefix) เป็นต้น ซึ่งทั้งนี้ไม่แน่ว่าการมีฟังก์ชันคุณสมบัติที่มากกว่าจะให้ค่า F1-score ได้มากกว่า จึงจำเป็นที่จะต้องมีการทดลองและสรุปผลว่าฟังก์ชันคุณสมบัติใดที่ส่งผลต่อประสิทธิภาพของแบบจำลอง

4.2) ส่วนการรวบรวมและประมวลผลข้อมูลเบื้องต้น

การรวบรวมข้อมูลจะใช้เครื่องมือ ค้นหาออกเป็น 2 ส่วนหลัก คือ 1. ค้นหาจากคำสำคัญ หรือ hashtag ที่มักจะมีคำสำคัญเกี่ยวกับสถานที่ เช่น อร่อยบอกต่อ js100radio รีวิวไทยแลนด์ กรุงเทพมหานคร นนทบุรี สมุทรปราการ ฯลฯ โดยจำกัดข้อมูลที่มาจากกรุงเทพและปริมณฑลเป็นหลัก และ 2. ค้นหาจาก bounding box กำหนดขอบเขตให้ครอบคลุมกรุงเทพ และปริมณฑล หลังจากนั้นจึงประมวลผลข้อมูลเบื้องต้นด้วยการตัดข้อความที่สั้นเกินไป หรือ ข้อความที่ซ้ำ รวมทั้งคำที่มีการซ้ำของตัวอักษรมาก ๆ เช่น “ยากกกกกกกกกก” “มากมายก่ายกองงงงงงงง” ได้ข้อความจำนวนทั้งหมด 28,082 ข้อความ แบ่งเป็น

ข้อมูลสำหรับฝึกสอน 22,445 ข้อความ คิดเป็นร้อยละ 80 และข้อมูลสำหรับทดสอบจำนวน 5,617 ข้อความ คิดเป็นร้อยละ 20

4.3) การทำฉลากให้กับข้อมูลเพื่อสร้างคลังข้อมูลทางภาษา (Corpus)

ในงานวิจัยนี้แบ่งประเภทของชื่อภูมิศาสตร์ออกเป็น 18 ชนิด โดยใน 18 ชนิดนั้นจัดอยู่ใน 5 หมวดสำคัญ ได้แก่ 1. สถานที่ธรรมชาติ 2. สิ่งปลูกสร้าง 3. ขอบเขตการปกครอง 4. สถานที่ที่ไม่อยู่ในอาณาเขตของประเทศไทย 5. สถานที่อื่น ๆ ใช้วิธีติดฉลากของข้อมูลด้วยมนุษย์ (human annotation) ซึ่งมีการสร้างคู่มือเพื่อเป็นแนวทางในการตัดสินใจเลือกฉลากที่เหมาะสมกับชื่อภูมิศาสตร์ในแต่ละประเภท แสดงตัวอย่างรูปแบบของฉลากที่ใช้ดังตารางที่ 1

ตารางที่ 1 : ตัวอย่างข้อความและลักษณะการติดฉลาก

ลำดับ	ข้อความ
1.	@Ordinary_Champ วิว<NAT>แม่น้ำโขงนครพนม</NAT> สวยสุดละ นี่ชอบมากกกกกก
2.	141062 @7.57 ฝนตกหนักเลย <ROAD>แจ้งวัฒนะ</ROAD> น่าจะตกก่อนออกจากบ้าน อุตสาห์จะไปเดินออกกำลังกาย<RCT>สวนจตุจักร</RCT> ท่าจะไม่รอด😅
...	.....
3.	ตักบาตรเข้าสารแห้ง... เราก้อทำน่ะ 👍😊😊 — ที่ <MKT>ตลาดน้ำขวัญ-เรียม</MKT> / <RP>วัดบางเพ็งใต้+วัดบำเพ็ญเหนือ กรุงเทพมหานคร</RP>

4.4) การตัดคำจากประโยคภาษาไทยเพื่อนำไปใช้ในการประมวลผล (Thai Tokenization)

การตัดคำถือเป็นขั้นตอนแรกในการดำเนินการเพื่อสร้างแบบจำลองการสกัดชื่อภูมิศาสตร์ออกมาจากข้อความ เนื่องจากข้อความที่ไม่ว่าจะเป็นประโยค หรือวลีต่างก็ประกอบไปด้วยคำที่มากกว่า 1 คำเรียงร้อยกัน โดยการตัดคำภาษาไทยในปัจจุบันมีผู้ศึกษาและมีการพัฒนาคลังคำสั่งในภาษา python ออกมาให้ใช้งาน ในงานวิจัยนี้เลือกเครื่องมือในการตัดคำจาก PyThaiNLP 2 แบบคือ newmm ซึ่งเป็นการใช้อัลกอริทึม maximal matching กับพจนานุกรมคำศัพท์ภาษาไทยเป็นพื้นฐาน และอีกวิธีหนึ่งคือ Attacut ซึ่งเป็นการนำเทคนิคการเรียนรู้เชิงลึก (deep learning) มาใช้งาน โดยมีการนำบริบททางความหมายของคำและประโยคมาใช้ประกอบทำให้ได้ผลที่ดีกว่าการตัดคำโดยใช้อัลกอริทึม maximal matching มาก [17] สำหรับในงานวิจัยนี้จะนำทั้งสองวิธีนี้มาใช้ร่วมกับแบบจำลองทางคณิตศาสตร์อย่าง CRF เพื่อเป็นการทดสอบอีกทางหนึ่งว่าเครื่องมือที่ใช้ในการตัดคำมีผลต่อแบบจำลองในทางใดทางหนึ่ง

4.5) การเปลี่ยนรูปแบบฉลากข้อมูลให้เป็นรูปแบบลำดับ (Sequence Tagging) แบบ IOB tags

การสร้าง tag แบบ IOB tags ซึ่งเป็นรูปแบบการสร้างฉลากให้กับข้อมูลแบบเป็นลำดับ (sequence tagging) โดย B คือ หน่วยแรกของชื่อเฉพาะ (beginning nouns phrase) I คือ ส่วนประกอบของชื่อเฉพาะซึ่งต่อเนื่องมาจากหน่วยแรก (inside current noun phrase) และสุดท้าย O คือ คำอื่นหรือ token อื่นที่เราไม่ได้จะสกัดเอาชื่อเฉพาะออกมา [18] ซึ่งเราจะดำเนินการต่อเนื่องมาจากขั้นตอนการตัดคำเนื่องจากตัวตัดคำที่ไม่เหมือนกันจะทำให้หน่วยย่อยของคำแตกต่างกัน เช่น ตัวอย่างข้อความที่ 1 ในตารางที่ 1

“@Ordinary_Champ วิว<NAT>แม่น้ำโขงนครพนม</NAT> สวยสุดละ นี่ชอบมากกกกกก”

(@Ordinary_Champ, O), (วิว, O), (แม่น้ำ, B-NAT), (โขง, I-NAT), (นครพนม, I-NAT), (สวย, O), (สุด, O), (ละ, O), (นี่, O), (ชอบมากกกกกก, O) จากตัวอย่างข้างต้นนี้จะพบว่า คำว่า “แม่น้ำโขงนครพนม” คือคำที่เราต้องการสกัดออกมาจากประโยคที่ 1 ซึ่งประกอบไปด้วยหน่วยย่อย 3 ส่วน คือ แม่น้ำ, โขง, นครพนม โดย แม่น้ำ คือหน่วยแรกของชื่อ และ โขงกับนครพนม คือส่วนประกอบของชื่อตามลำดับ และคำที่เหลือมีฉลากเป็น O ซึ่งเราไม่ต้องการสกัดออกมา

4.6) การกำหนดคุณลักษณะเพื่อสร้างแบบจำลอง (Featurization)

ในการกำหนดคุณลักษณะเป็นการคัดเลือกเอาคุณลักษณะเด่นของชื่อแต่ละประเภทที่เราต้องการจำแนกออกจากกัน เช่น หากมีคำบุพบทนำหน้า “อยู่ที่” “ที่” “ถัดจาก” ฯลฯ มักจะเป็นคำที่บ่งบอกถึงสถานที่ หรือ การติด tag หน้าที่ของคำในประโยค หรือ Part of speech tagging (POS-tagss) โดยชื่อเฉพาะนั้นมักจะทำหน้าที่เป็นคำนามเป็นหลักในประโยค โดยแบ่งคุณลักษณะสำคัญดังนี้

4.6.1) หน้าที่ของคำในประโยค (Part of Speech Tagging: POS-tagss) โดยส่วนมากชื่อเฉพาะมักจะทำหน้าที่เป็นคำนามในประโยค การนำ POS-tagss มาใช้จะสามารถระบุชื่อเฉพาะได้ดีขึ้น

4.6.2) บุพบทที่แสดงถึงการระบุสถานที่ หากมีคำบุพบทนำหน้าชื่อเฉพาะเหล่านี้ก็อาจมีแนวโน้มที่จะเป็นชื่อภูมิศาสตร์ได้ เนื่องจากมีความเป็นไปได้ที่จะอ้างอิงเชิงตำแหน่ง เช่น “อยู่” “อยู่ที่” “ที่” “ถึง” “ไปถึง” ฯลฯ

4.6.3) คำอุปสรรค (Prefix) เป็นคำนำหน้าชื่อเฉพาะ เช่น มีคำว่า “โรงเรียน” “มหาวิทยาลัย” ก็มีแนวโน้มว่ามีโอกาสเป็นชื่อภูมิศาสตร์ที่เกี่ยวข้องกับสถานศึกษา

4.6.4) คำท้าย (Suffix) เป็นคำที่อยู่ท้ายชื่อซึ่งสื่อถึงประเภทของชื่อเฉพาะนั้น ๆ ได้ เช่น “วิทยาลัย” มีแนวโน้มที่จะเป็นชื่อภูมิศาสตร์ที่เกี่ยวกับสถานศึกษา หรือ “ดีพาร์ทเมนต์สโตร์” ที่มีแนวโน้มจะเป็นห้างสรรพสินค้า

4.6.5) การสร้างอักขรานุกรมเพื่อนำมาเป็นหนึ่งในคุณลักษณะสำคัญของแบบจำลอง (Gazetteer) การสร้างอักขรานุกรมนำข้อมูลชื่อภูมิศาสตร์ของกรมแผนที่ทหารมาเป็นส่วนหนึ่งในการสร้างอักขรานุกรม อีกทั้งมีการใช้เทคนิคการขูดเว็บ (web scraping) จากแหล่งข้อมูลที่มีชื่อสถานที่หลากหลาย เช่น wongnai retty รวมทั้งข้อมูล open data จาก https://data.go.th/ ซึ่งเป็นข้อมูลที่เข้าใช้งานได้แบบไม่มีค่าใช้จ่ายโดยผ่านการจัดทำจากหน่วยงานภาครัฐที่เป็นเจ้าของข้อมูล ตัวอย่างการใช้งานอักขรานุกรมชื่อภูมิศาสตร์ เช่น เรา สามารถรวบรวมรายชื่อเกี่ยวกับสถานศึกษามาได้ดังนี้ [‘โรงเรียนกรุงเทพคริสเตียนวิทยาลัย’, ‘โรงเรียนสวนกุหลาบวิทยาลัย’, ‘โรงเรียนสตรีวิทยา’] ซึ่งมีการเก็บข้อมูลเป็น 2 กลุ่ม คือ กลุ่มของชื่อที่ไม่มีการตัดคำ และ กลุ่มที่มีการตัดคำ ตัวอย่างจากลิสต์ของชื่อสถานศึกษาข้างต้น เมื่อมีการตัดคำเก็บข้อมูลได้เป็น [[‘โรงเรียน’, ‘กรุงเทพ’, ‘คริสเตียน’, ‘วิทยาลัย’], [‘โรงเรียน’, ‘สวนกุหลาบ’, ‘วิทยาลัย’], [‘โรงเรียน’, ‘สตรี’, ‘วิทยา’]] โดยจะกล่าวถึงส่วนของการนำคุณลักษณะมาใช้สำหรับการสร้างแบบจำลองในส่วนถัดไป

4.6.6) คำข้างเคียงที่อยู่ติดกัน (n–gram) สำหรับนิยามของ n-gram คือของที่อยู่ติดกันเป็น n ลำดับ เช่น bi-gram หมายถึงคำสองคำที่อยู่ติดกัน ดังตัวอย่าง [‘โรงเรียน’, ‘กรุงเทพ’, ‘คริสเตียน’, ‘วิทยาลัย’] หากกำหนดให้คำว่า ‘คริสเตียน’ เป็นตำแหน่งที่ n คำว่า ’กรุงเทพ’ ที่อยู่ก่อนหน้าจะเป็น n-1 และ ‘วิทยาลัย’ ซึ่งเป็นคำถัดไปจะอยู่ในตำแหน่งที่ n+1 ซึ่งสำหรับงานวิจัยนี้จะมีการใช้ quadigram หรือคำที่อยู่ก่อนหน้า ถัดไปมากที่สุด 4 คำ โดยจากมีงานวิจัยที่อธิบายเหตุผลหลักพบว่าในภาษาไทยหากคำอยู่ห่างกันมากกว่า 4 คำขึ้นไปจะมีความสัมพันธ์กันน้อยมาก ดังนั้นใช้เพียง quadigram ก็เพียงพอแล้วสำหรับภาษาไทย [19]

4.6.7) คุณสมบัติอื่น ๆ นอกเหนือจากคุณสมบัติที่กล่าวถึงมาในข้างต้นแล้วยังมีคุณสมบัติอื่น ๆ ที่นำมาใช้เพิ่มเติมได้แก่ ความยาวของของคำ (word length) รูปแบบของคำ (word shape) เช่น มีตัวเลขประกอบในคำ มีการเว้นช่องว่างระหว่างคำ มีสัญลักษณ์เชื่อมคำ มีภาษาไทยผสมกับภาษาอังกฤษ การมีอักขระพิเศษ รวมทั้ง อีโมติคอนต่าง ๆ ประกอบอยู่ในคำด้วย แสดงตัวอย่างฟังก์ชันคุณสมบัติตามตารางที่ 2

ตารางที่ 2 : สรุปตัวอย่างฟังก์ชันคุณสมบัติที่ใช้ในงานวิจัย

Feature Function	Return of function
POS-tags	<NCMN>,... , etc.
Preposition	word, False
Prefix	word, False
Suffix	word, False
gazetteer	word, False
N-gram	word, False
Space	True, False
Word length	28
Word shape	Eeedd_TTTTTT

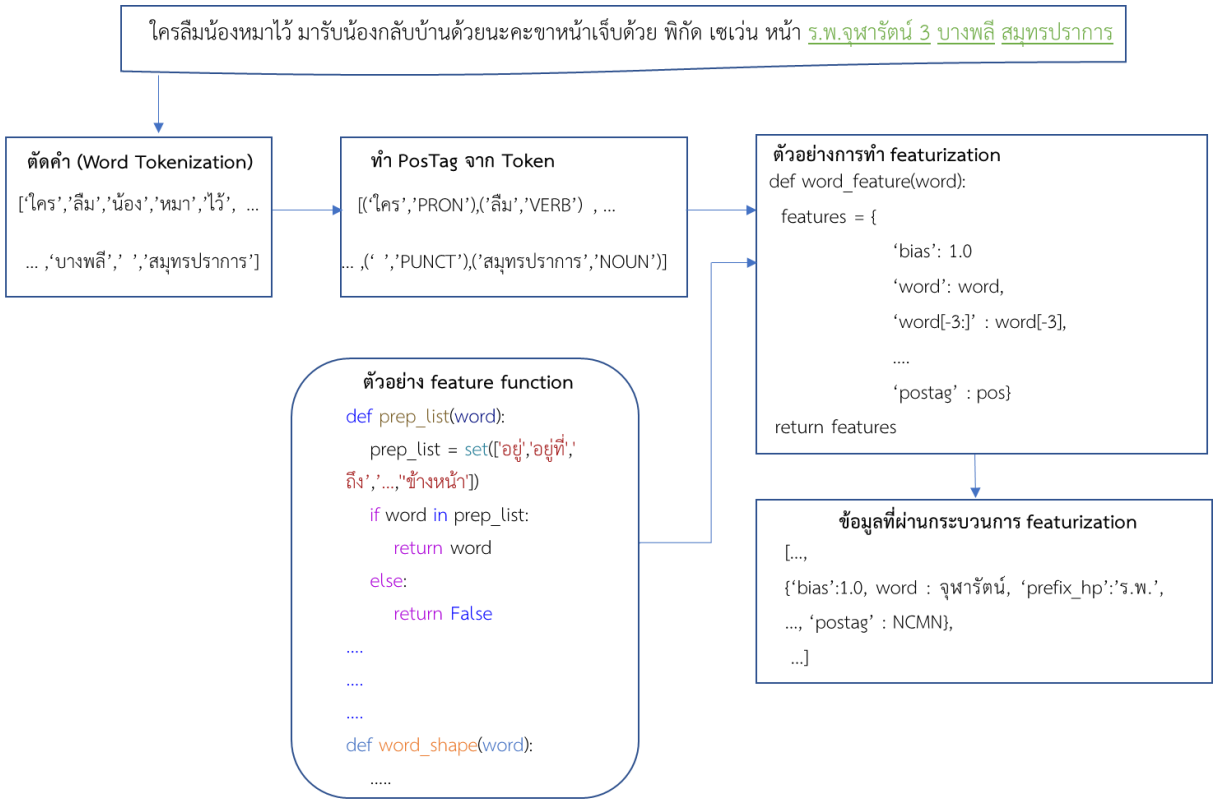
จากตารางที่ 2 แสดงตัวอย่างของฟังก์ชันคุณสมบัติและตัวอย่างค่าที่ฟังก์ชันคืนมา โดยสำหรับ POS-tagss จะใช้คลังคำสั่งจาก PyThaiNLP ซึ่งคืนค่า POS-tagss ของคำในรูปแบบ Orchid corpus, ฟังก์ชันบุพบทเชิงตำแหน่ง จนถึง N-gram ฟังก์ชันจะคืนค่ามาเป็นคำที่พบ แต่หากไม่เข้าเงื่อนไขกำหนดให้คืนค่าเป็น False เช่น หากมีคำที่อยู่ใน set ของคำดังต่อไปนี้ set(['อยู่','อยู่ที่','ถึง','ไปถึง','มาถึง','มุ่งหน้า','เข้าสู่','เส้น','เยื้อง','ถัดไป','ถัดจาก','ตรงข้าม','ตรงกับ','ตรงข้ามกับ',.....,'ติด','ติดกับ','ข้าง','ข้างหน้า']) ให้คืนค่ากลับมาเป็นคำที่พบใน set ข้างต้น และสุดท้ายคือฟังก์ชันคุณสมบัติอื่น ๆ ได้แก่ ภายในคำมีช่องว่างหรือไม่, ความยาวของคำ เช่น คำว่า “สยามวัน” จะคืนค่ามาเป็นตัวเลขแทนด้วยจำนวนตัวอักษรของคำคือ 7 และรูปร่างของคำ ซึ่งในที่นี้คือ ประกอบไปด้วยตัวอักษรภาษาอังกฤษตัวใหญ่หรือตัวเล็ก, ตัวอักษรภาษาไทย, ตัวเลข มีช่องว่างระหว่างตัวอักษรหรือไม่ เช่น คำว่า “ร้าน Coffeebeans สยามพารากอน” เมื่อผ่านฟังก์ชันรูปร่างของคำจะคืนค่าเป็น “TTTTEeeeeeeeeee_TTTTTTTTTT” เป็นต้น

4.7) เครื่องมือที่ใช้ในการประมวลผล

ในหัวข้อนี้จะกล่าวโดยสรุปในส่วนของภาษาโปรแกรม คลังคำสั่งที่ใช้ รวมถึงทรัพยากรในการประมวลผล

4.7.1) ภาษาโปรแกรมและคลังคำสั่งที่ใช้ในการทดลอง ในงานวิจัยนี้ใช้ภาษาไพธอนเป็นภาษาหลัก โดยมีการใช้คลังคำสั่งจาก scikit-learn และ sklearn-crfsuite เพื่อสร้างแบบจำลอง CRF รวมถึงคลังคำสั่งจาก pythanlp เพื่อประมวลผลข้อความเบื้องต้นและตัดคำ (tokenization)

4.7.2) เครื่องมือที่ใช้ในการประมวลผล เป็น gaming laptop dell inspiron 7559; ram 16 gb; gpu nvidia geforce gtx 960m; เวลาที่ใช้ในการฝึกฝนแบบจำลอง 3 ชม. 30 นาที



รูปที่ 2 : แสดงตัวอย่างการสร้างฟังก์ชันคุณลักษณะ (featurization)

จากรูปที่ 2 แสดงตัวอย่างขั้นตอนการเตรียมข้อมูลและการทำ featurization ไล่เรียงจากการตัดคำในแต่ละประโยคด้วย newmm หรือ arttacut จาก PyThaiNLP หลังจากนั้นจึงติด POS-tagss ด้วยรูปแบบของ Orchid corpus และติดฉลากตามรูปแบบของ IOB tags พร้อมทั้งจัดรูปแบบของข้อมูลที่ใช้ในการฝึกฝนและทดสอบแบบจำลองในรูปแบบ Conll2003 หลังจากนั้นจึงจะนำข้อมูลผ่านฟังก์ชันคุณลักษณะที่สร้างเตรียมไว้ โดยผลลัพธ์จากขั้นตอนนี้เมื่อทำ featurization จะออกมาเป็นข้อมูลแบบ dictionary ซึ่งเก็บ key เป็นชื่อฟังก์ชันที่ใช้ และ value เป็นส่วนที่ฟังก์ชันได้คืนค่ากลับมาดังแสดงตัวอย่างไว้ในหัวข้อย่อยที่ 4.6.7

5) ผลการวิจัยและอภิปรายผล

การจำแนกข้อมูลชื่อภูมิศาสตร์ออกจากข้อความโดยใช้แบบจำลองคอนดิชันนอลแรนดอมฟิลด์แบ่งผลลัพธ์ในการจำแนกออกเป็นกลุ่มตามคุณลักษณะที่มีแนวโน้มจะส่งผลต่อประสิทธิภาพของแบบจำลอง

5.1) การประเมินประสิทธิภาพของแบบจำลอง

ในงานวิจัยนี้เลือกใช้เมทริกซ์ 3 ตัว คือ ความแม่นยำ (precision), ความครบถ้วน (recall) และ ความถูกต้องโดยรวม (F1) ซึ่งในการประเมินจะประเมินจากความถูกต้องในระดับคำ (phrase level) พร้อมทั้งมีความถูกต้องในระดับ token แสดงไว้ควบคู่กัน แต่ประสิทธิภาพจะยึดตามความถูกต้องในระดับคำเป็นหลัก

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F1-Score} = 2^* \frac{\text{Precision}^*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (3)$$

โดย TP คือ จำนวนข้อมูลที่แบบจำลองสกัดออกมาได้ตรงกับข้อมูลกำกับ

FP คือ จำนวนข้อมูลที่ไม่ได้มีการกำกับไว้แต่แบบจำลองสกัดออกมา

FN คือ จำนวนข้อมูลที่การกำกับไว้แต่แบบจำลองไม่สามารถสกัดออกมาได้

สำหรับการประเมินประสิทธิภาพด้วย F1-Token เพียงอย่างเดียวยังไม่ถูกต้องครบถ้วนนักเนื่องจากการสร้างแบบจำลองเราต้องการที่จะสกัดชื่อภูมิศาสตร์ทั้งชื่อออกมาจากข้อความไม่ใช่เพียงแค่ส่วนใดส่วนหนึ่งของชื่อ ตัวอย่างเช่น การหาค่า F1-Token ประโยคว่า “วิวแม่น้ำโขงนครพนมสวยสุดละ นี่ชอบมาก” แสดงตัวอย่างตามรูปที่ 3

หน่วยของคำ	ค่าที่กำกับ	ค่าที่ทำนาย	
วิว	O	B-NAT	FP
แม่น้ำ	B-NAT	B-NAT	TP
โขง	I-NAT	O	FN
นครพนม	B-ADMIN	B-ADMIN	TP
สวย	O	O	
สุดละ	O	O	
นี่	O	O	
ชอบมาก	O	O	

รูปที่ 3 : แสดงตัวอย่างการหาค่า F1-Token

จากรูปที่ 3 คำนวณ F1-Token ได้ดังนี้ TP = 2, FP = 1, FN = 1 Precission = 2/(2+1), Recall = 2/(2+1), F1 = 2* [(0.67* 0.67)/(0.67+0.67)] = 0.34 และสำหรับ F1-Phrase ข้อมูลกำกับที่มี 2 token นั้นจะถูกรวมเป็น tag เดียว ซึ่งถ้ามีส่วนหนึ่งส่วนใดของ token ผิดไปให้ถือว่าคำตอบที่ได้จากแบบจำลองผิด เช่น (แม่น้ำ, B-NAT), (โขง, I-NAT), (นครพนม, B-ADMIN) จะรวม token และ tag เป็น แม่น้ำโขง, NAT และ นครพนม, ADMIN ซึ่งจากรูปที่ 3 สำหรับคำว่าแม่น้ำโขง แบบจำลองให้คำตอบ token สุดท้ายผิดจึงถือว่าแบบจำลองให้คำตอบที่ผิดในกรณีนี้ จากตัวอย่างค่าที่คำนวณ F1-Phrase ได้จาก TP = 1, FP = 1, FN = 1, Precission = 1/(1+1), Recall = 1/(1+1), F1 = 2*[(0.5*0.5)/(0.5+0.5)] = 0.25 จากตัวอย่างข้างต้นจะพบว่าควรพิจารณาที่ระดับ F1-Phrase มากกว่า เนื่องจากผลลัพธ์สุดท้ายของแบบจำลองชื่อภูมิศาสตร์ที่ครบสมบูรณ์

5.2) ประสิทธิภาพของแบบจำลองสกัดชื่อภูมิศาสตร์จากข้อความ

โดยแสดงผลลัพธ์ตามตารางที่ 2 ซึ่งทุกกลุ่มนั้นจะมีพื้นฐานอยู่บนแบบจำลอง CRF มีฟังก์ชัน optimizer คือ LBFGS มีการใช้ฟังก์ชันคุณลักษณะคำนำหน้า คำตามหลัง รูปร่างของคำ และหน้าที่ของคำ (POS-tags) โดยผลจากการจำแนกในงานวิจัยแบ่งออกเป็นกลุ่มตามฟังก์ชันคุณลักษณะดังนี้

ตารางที่ 3 : แสดงการเปรียบเทียบค่าความถูกต้องโดยรวมระหว่างฟังก์ชันคุณลักษณะในแบบจำลอง CRF

Tokenize Library	Gram	Gazetteer	F1-Token	F1-Phrase
newmm	3	-	0.909	0.881
newmm	3	✓	0.91	0.882
newmm	4	-	0.908	0.885
newmm	4	✓	0.908	0.885
attacut	3	-	0.95	0.938
attacut	3	✓	**0.956**	**0.946**
attacut	4	-	0.951	0.941
attacut	4	✓	**0.956**	**0.946**

จากตารางที่ 3 กลุ่มของแบบจำลองจะแบ่งออกเป็น 2 กลุ่ม จากอัลกอริทึมที่ใช้ตัดคำ คือ newmm และ attacut โดยพบว่าแบบจำลองที่ใช้ attacut เป็นตัวตัดคำให้ค่าความถูกต้องโดยรวมในระดับคำสูงกว่า newmm คือ 0.946 ส่วน newmm คือ 0.885 โดยแบบจำลองที่ใช้ตัวตัดคำ newmm นั้นพบว่า แบบจำลองที่ใช้ 4 gram จะใช้ gazetteer เป็นคุณลักษณะหรือไม่ ก็ให้ค่า F1-Phrase เท่ากันทั้ง 2 แบบ คือ 0.885 และสำหรับกลุ่มที่ใช้ตัวตัดคำเป็น attacut พบว่า 3 gram หรือ 4 gram และจากแบบจำลองที่ให้ค่าความถูกต้องโดยรวมสูงที่สุด แสดงผลความถูกต้องของการจำแนกแยกตามประเภทของชื่อภูมิศาสตร์ ตามตารางที่ 4

ตารางที่ 4 : แสดงรายละเอียดการจำแนกชื่อภูมิศาสตร์ตามชนิดของสถานที่

TAG	ความหมาย	Precision	Recall	F1
ACP	สถานศึกษา	0.946	0.946	0.946
ADMIN	ขอบเขตการปกครอง	0.964	0.951	0.958
BSN	อาคารสำนักงาน	0.913	0.808	0.857
DEP	ห้างสรรพสินค้า-ขนาดใหญ่	0.949	0.903	0.925
FPLACE	สถานที่ที่อยู่นอกประเทศไทย	0.956	0.869	0.91
GOV	สำนักงาน สถานที่ราชการ	1.0	0.84	0.913
HP	สถานพยาบาล	1.0	0.958	0.979
MKT	ตลาด	1.0	1.0	1.0
MON	อนุสาวรีย์ วงเวียน	1.0	1.0	1.0
NAT	สถานที่ตามธรรมชาติ	0.909	0.909	0.909
RCT	สวนสาธารณะ สวนสนุก สนามกีฬา สถานที่สันทนาการต่าง ๆ	0.87	0.909	0.889
RES	สถานที่พักอาศัย	0.964	0.9	0.931
ROAD	ถนน ซอย	0.979	0.903	0.94
RP	สถานที่สำคัญทางศาสนา	0.962	0.98	0.971
RT	ร้านอาหาร	0.946	0.953	0.95
STORE	ร้านค้าขนาดย่อม	1.0	0.917	0.957
TRAN	สถานีขนส่งมวลชน ท่ารถ สถานีรถไฟ รถไฟฟ้า ท่าเรือ	0.925	0.949	0.937
OTHER	สถานที่อื่น ๆ	1.0	1.0	1.0

จากตารางที่ 4 ความถูกต้องโดยรวมซึ่งสำหรับตารางนี้จะกล่าวถึง F1-Phrase เพียงอย่างเดียว เนื่องจากเป็นการสกัดคำที่ครบถ้วนแล้วออกมาจากประโยค ของชื่อภูมิศาสตร์ที่เป็น ตลาด อนุสาวรีย์ หรือสถานที่อื่น ๆ มีความถูกต้องโดยรวมที่สูงมากกว่าชั้นข้อมูลอื่น F1 คือ 1.0 ทั้งนี้อาจเนื่องจาก ตลาดมักมีคำหน้าขึ้นต้นด้วย “ตลาด” เป็นคุณลักษณะที่เด่นชัดจึงแยกออกจากข้อมูลอื่นได้ง่าย ส่วนชนิดของสถานที่ที่ค่า F1 น้อยที่สุด คืออาคารสำนักงาน (BSN) คือ 0.857 ทั้งนี้อาจจะเนื่องจากเป็นชื่อที่ค่อนข้างมีความหลากหลายและกำกวมระหว่างชื่อบุคคล รวมถึงชื่อหน่วยงาน องค์กรต่าง ๆ อย่างทวิตเตอร์

5.3) Transition Score and Feature Learns

ซึ่งในหัวข้อนี้จะเป็นการสรุปภาพรวมของความน่าจะเป็นของแบบจำลองในการทำนาย tag ของชื่อภูมิศาสตร์ ว่าหากมี tag ไหนที่มีค่า transition สูงจะแสดงว่าแบบจำลองมีโอกาสที่จะคล้อยตามในแต่ละ tag นั้นและส่วนของ feature learns จะเป็นการแสดงผลของคุณลักษณะที่ส่งผลต่อประสิทธิภาพของแบบจำลอง โดยในหัวข้อนี้จะแสดงใน 30 ลำดับแรกของคุณลักษณะที่แบบจำลองเรียนรู้แสดงตามตารางที่ 5 6 และ 7

โดยจากตารางที่ 5 tag ที่ให้ค่า likely transition สูงที่สุด I-MKT -> I-MKT ซึ่งเป็น tag ที่เป็นส่วนย่อยหมายความถึงชื่อตลาด พบว่าสอดคล้องกับ MKT ที่ให้ค่า F1 = 1.0 หรือ 100% และลำดับที่ 2 กับ 3 คือ B-TRAN -> I-TRAN และ B-RES -> I-RES โดย TRAN และ RES ให้ค่า F1 เป็น 0.937 และ 0.931 ตามลำดับ

จากตารางที่ 7 ซึ่งเป็นส่วนของคุณลักษณะที่แบบจำลองเรียนรู้ได้ พบว่าใน 30 อันดับแรกซึ่งเป็นผลบวกกับแบบจำลองมักจะเป็นคุณลักษณะที่เป็น n-gram และคำนั้น ๆ โดยพบว่าเป็นคำก่อนหน้า 3 คำ (word[-3]) มากที่สุด รองลงมาคือคำก่อนหน้า 2 คำ (word[-2]) เมื่อสรุปจาก ตาราง feature learns จึงพบว่า n-gram แบบ quadigram ให้ผลบวกกับค่าความน่าจะเป็นของแบบจำลองสูงที่สุด และอีกคุณลักษณะหนึ่งที่น่าสนใจคือ รูปร่างของคำ หรือ ฟังก์ชัน shape_word พบว่าให้ผลที่ดีกับ tag ที่เป็น B-ADMIN และ B-ROAD ซึ่งสันนิษฐานได้ว่า ชื่อขอบเขตการปกครอง เช่น จังหวัด อำเภอ หรือตำบล มักจะมีอักษรย่อนำหน้าแล้วตามด้วยจุด เช่น ‘อ.พบพระ’ เมื่อผ่านฟังก์ชันนี้จะให้ค่าเป็น T.TTTTTT เมื่อแบบจำลองเจอคุณลักษณะนี้จึงเข้าใจได้ง่ายขึ้นว่าเป็น ADMIN รวมทั้ง tag ที่เป็นถนนซึ่ง มักจะมีตัวเลข

ต่อท้ายหรือเป็นส่วนประกอบอยู่ภายในคำ เช่น ‘ซอยลาดพร้าว 60’ เมื่อผ่านฟังก์ชันรูปร่างของคำจะคืนค่าเป็น TTT.TTTTTTT_dd เป็นต้น

ตารางที่ 5 : แสดงรายละเอียดของค่า transition ที่คล้อยตามกัน

Tag	Target tag	ค่า transition
I-MKT	I-MKT	7.166
B-TRAN	I-TRAN	6.979
B-RES	I-RES	6.927
I-NAT	I-NAT	6.713
B-GOV	I-GOV	6.633
I-ACP	I-ACP	6.603
B-MKT	I-MKT	6.599
I-ADMIN	I-ADMIN	6.516
B-DEP	I-DEP	6.491
I-OTHER	I-OTHER	6.478
I-GOV	I-GOV	6.472
B-HP	I-HP	6.391
I-TRAN	I-TRAN	6.379
B-ACP	I-ACP	6.321
B-RT	I-RT	6.316
I-RCT	I-RCT	6.304
I-RP	I-RP	6.291
I-RES	I-RES	6.278
I-STORE	I-STORE	6.239

ตารางที่ 6 : แสดงรายละเอียดของค่า transition ที่ไม่คล้อยตามกัน

Tag	Target tag	ค่า transition
I-BSN	I-RT	-1.809
I-ROAD	B-ROAD	-2.241
O	I-MON	-3.078
O	I-OTHER	-3.975
O	I-MKT	-4.825
O	I-NAT	-4.919
O	I-ADMIN	-5.694
O	I-TRAN	-5.789
O	I-GOV	-5.989
O	I-ACP	-6.024
O	I-FPLACE	-6.125
O	I-RES	-6.165
O	I-RCT	-6.272
O	I-HP	-6.385

ตารางที่ 6 : แสดงรายละเอียดของค่า transition ที่ไม่คล้อยตามกัน (ต่อ)

Tag	Target tag	ค่า transition
O	I-STORE	-6.451
O	I-DEP	-6.699
O	I-RP	-6.713
O	I-BSN	-6.822
O	I-ROAD	-6.985
O	I-RT	-8.325

ตารางที่ 7 : แสดงรายละเอียดคุณลักษณะที่แบบจำลองเรียนรู้

Tag	Feature	Feature score
B-RCT	word:ราชมัง	4.449
B-ROAD	-1:word:โครตไฮลาดพร้าว	3.342
O	word[-3:]:	3.329
O	BOS	3.284
B-RP	word[-3:]:หาร	3.254
B-RCT	word[-3:]:มัง	3.202
B-TRAN	word:สนามบินดอนเมือง	3.058
O	word:	3.015
O	word.shape_word():	3.015
B-DEP	word[-3:]:gna	2.911
B-ADMIN	word:กทม	2.789
B-ADMIN	word.shape_word():T.TTTTTTTTT	2.750
B-RT	word:บ้านหิรัณยกุล	2.733
B-ACP	word[-3:]:ตร์	2.642
B-DEP	word[-3:]:กอน	2.634
B-DEP	word[-2:]:bk	2.632
B-DEP	word:เอ็มควา5555	2.619
B-DEP	-2:word:พิกัด	2.600
B-ROAD	word.shape_word():TTTTTTd	2.589
O	word[-2:]:55	2.578
B-ADMIN	word[-3:]:uri	2.562
B-ADMIN	word.shape_word():T.TTTTTTT	2.542
O	word:มุ่งหน้า	2.536
O	word.prep_list():มุ่งหน้า	2.536
B-ACP	word[-3:]:ษตร	2.529
B-ROAD	word[-3:]:ีลม	2.518
B-DEP	-1:word:สยามสแควร์	2.504
O	word.shape_word():TTTT.....T	2.492
B-FPLACE	word[-3:]:ซาน	2.479
B-RCT	word:สวนจตุจักร	2.473

5.4) การเปรียบเทียบกับแบบจำลองอื่น

เนื่องจากแบบจำลองอื่นหรือเครื่องมือที่เป็นมาตรฐานทางด้านการรู้จำชื่อเฉพาะ (standard ner tools) ไม่ได้มีการแบ่งประเภทของชื่อภูมิศาสตร์ไว้ดังเช่นงานวิจัยนี้จึงไม่สามารถนำมาใช้ได้ทันที ดังนั้นผู้วิจัยจึงนำเอาคุณลักษณะที่แบบจำลองนั้นใช้หากเป็นแบบจำลอง CRF เหมือนกัน และนำสถาปัตยกรรมพร้อมทั้งพารามิเตอร์มาใช้หากเป็นโครงข่ายประสาทเทียม โดยนำมาจาก thai-ner CRF ของ PyThaiNLP , Bi-LSTM และ Bi-LSTM-CRF [15] แสดงตัวอย่างตามตารางที่ 8

ตารางที่ 8 : แสดงรายการเปรียบเทียบค่าความถูกต้องโดยรวมของแบบจำลอง CRF+attacut+4 gram กับแบบจำลองอื่น

Model	Detail	F1-Token	F1-Phrase
1	CRF+attacut+4 gram	**0.956**	**0.946**
2	CRF (PyThaiNLP)	0.894	0.80
3	Bi-LSTM	0.91	0.827
4	Bi-LSTM-CRF [15]	0.92	0.87

จากตารางที่ 8 มีการเปรียบเทียบค่าความถูกต้องโดยรวมหรือ F1 กับอีก 3 แบบจำลอง โดยมีแบบจำลอง CRF ของ PyThaiNLP, แบบจำลอง Bi-LSTM และ Bi-LSTM-CRF [15] ซึ่งสำหรับแบบจำลองที่ 2 CRF ของ PyThaiNLP นั้นใช้ฟังก์ชันคุณลักษณะดังนี้คือ คุณลักษณะของ n-gram ใช้เป็น bi-gram ตรวจสอบว่าเป็นคำภาษาไทยทั้งหมดหรือไม่, คำฟุ่มเฟือย, ช่องว่างภายในคำ, เป็นตัวเลขหรือไม่ และสุดท้ายคือ POS-tags ให้ค่า F1-phrase อยู่ที่ 0.8 สำหรับแบบจำลองที่ 3 ซึ่งเป็น Bi-LSTM นั้นใช้ word embeddings จาก Thai2vec [4] มีพารามิเตอร์ดังนี้ จำนวน neuron 256 neuron, drop out 0.3, adam optimize, จำนวนในการฝึกฝนแบบจำลอง 50 epochs ให้ค่า F1-phrase อยู่ที่ 0.827 และสุดท้ายแบบจำลองที่ 4 Bi-LSTM-CRF ซึ่งอ้างอิงสถาปัตยกรรมจาก [15] โดยเป็นการใช้ word embeddings จาก Thai2vec ร่วมกับการทำ character embeddings มีพารามิเตอร์ดังนี้ จำนวน neuron 256 neuron, drop out 0.5, adam optimize, จำนวนในการฝึกฝนแบบจำลอง 50 epochs ให้ค่า F1-phrase อยู่ที่ 0.87

6) สรุปผล

ในงานวิจัยนี้มุ่งเน้นในการสร้างแบบจำลองเพื่อสกัดข้อมูลชื่อภูมิศาสตร์จากข้อความภาษาไทยที่อยู่บนสื่อสังคมออนไลน์อย่างทวิตเตอร์ โดยจากการนำอัลกอริทึม CRF มาใช้งานพบว่าให้ความถูกต้องในการสกัดเอาชื่อทางภูมิศาสตร์ออกมาได้มีความถูกต้องซึ่งสามารถนำไปใช้งานได้ และการจำแนกชนิดของชื่อสถานที่ก็ทำได้มีความถูกต้องสูงในหลายชั้นข้อมูล (>0.9) ซึ่งชั้นข้อมูลที่มีปัญหามากกว่าชั้นข้อมูลอื่นคือ BSN หรืออาคารสำนักงาน เนื่องจากลักษณะการตั้งชื่อของภาษาไทยไม่ได้มีกฎตายตัว รวมทั้งอาจมีความกำกวมกับชื่อเฉพาะแบบอื่น ทั้งชื่อองค์กร หรือบุคคล เช่น “เล่า เป้ง ง้วน” หากไม่มี gazetteer เป็นตัวช่วยในการกำหนดคุณลักษณะอาจจะไม่สามารถสกัดชื่ออาคารออกมาจากข้อความได้ แต่หากเขียนเป็น “อาคารเล่าเป้งง้วน” จะสามารถสกัดข้อมูลออกมาได้เนื่องจากมีคำอุปสรรคว่า “อาคาร” เป็นคุณลักษณะที่กำกับอยู่ ซึ่งงานวิจัยนี้มีการสร้างคุณลักษณะที่มีสมมติฐานว่ามีความเหมาะสมกับข้อมูลชื่อภูมิศาสตร์ โดยจากการวิเคราะห์สิ่งที่แบบจำลองเรียนรู้พบว่าฟังก์ชันที่เป็นบุพบทที่ใช้ระบุตำแหน่งให้ผลเชิงบวกกับค่าความน่าจะเป็นของแบบจำลอง CRF ที่ให้ค่าความถูกต้องโดยรวมดีที่สุดในงานวิจัยนี้ รวมทั้งฟังก์ชันรูปร่างของคำ

จากการเปรียบเทียบประสิทธิภาพของแบบจำลอง CRF ที่ให้ค่า F1 ดีที่สุดในงานวิจัยนี้กับแบบจำลองอื่นทั้ง CRF ของ PyThaiNLP กับ CRF ในงานวิจัยนี้ที่มีการสร้างคุณลักษณะต่างกัน โดยในงานวิจัยนี้มีการสร้างคุณลักษณะโดยเน้นที่ข้อมูลเชิงภูมิศาสตร์แบบจำเพาะจึงเป็นเหตุผลที่ทำให้สามารถปรับใช้กับข้อมูลชนิดนี้ได้ดีกว่าขณะที่แบบจำลองอื่นสร้างขึ้นเพื่อการใช้งาน NER โดยทั่วไป และอีกประการหนึ่งคลังข้อมูลในงานวิจัยนี้มีขนาดไม่ใหญ่มากทำให้การใช้แบบจำลอง CRF ให้ผลที่ดีกว่าหรือใกล้เคียงกันกับโครงข่ายประสาทเทียมซึ่งในอนาคตหากมีการเก็บรวบรวมข้อมูลที่มากขึ้นอาจจะมีผลทำให้ผลการวิจัยต่างไปจากนี้

จากงานวิจัยนี้พบว่าการสกัดข้อมูลชื่อภูมิศาสตร์เฉพาะในภาษาไทยนั้นจำเป็นที่จะต้องมีการสร้างฟังก์ชันคุณลักษณะที่จำเพาะเจาะจงกับงานทางด้านนี้โดยเฉพาะ เช่น บุพบทระบุตำแหน่ง รูปร่างของคำ หรืออักขรานุกรม (gazetteer) มาใช้เป็นหนึ่งในคุณลักษณะที่สำคัญเนื่องจากเป็นโจทย์ที่เฉพาะทางซึ่งในอนาคตหากมีการสร้างคลังข้อมูลที่มีขนาดใหญ่ขึ้น หรือมีการสกัดคุณลักษณะสำคัญร่วมกับโครงข่ายประสาทเทียม หรือการนำ transfer learning เช่น BERT มาใช้งานอาจทำให้ได้ผลลัพธ์ที่ดีมากขึ้นในอนาคต

REFERENCES

[1] S. Kemp. "DATAREPORTAL." DATAREPORTAL.com https://datareportal.com/reports/digital-2021-thailand (accessed Jan. 2, 2022).

[2] S. Hahmann and D. Burghardt, "How much information is geospatially referenced? Networks and cognition," *Int. J. Geogr. Inf. Sci.*, vol. 27, no. 6, pp. 1171–1189, 2013, doi: 10.1080/13658816.2012.743664.

[3] M. Gritta, M. T. Pilehvar, N. Limsopatham, and N. Collier, "What's missing in geographical parsing?," *Lang. Resour. Eval.*, vol. 52, no. 2, pp. 603–623, 2018, doi: 10.1007/s10579-017-9385-8.

[4] W. Phatthiyaphaibun, K. Chaovavanich, C. Polpanumas, A. Suriyawongkul, L. Lowphansirikul, and P. Chormai, "PyThaiNLP: Thai Natural language processing in Python." 2016. Distributed by Zenodo. doi: 10.5281/zenodo.3519354.

[5] J. Lingad, S. Karimi, and J. Yin, "Location extraction from disaster-related microblogs," in *Proc. 22nd Int. Conf. World Wide Web*, Rio de Janeiro, Brazil, May 2013, pp. 1017–1020.

[6] J. Wang, Y. Hu, and K. Joseph, "NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages," *Trans. GIS*, vol. 24, no. 3, pp. 719–735, 2020.

[7] J. A. de Bruijn, H. de Moel, B. Jongman, J. Wagemaker, and J. C. J. H. Aerts, "TAGGS: Grouping tweets to improve global geoparsing for disaster response," *J. Geovis. Spat. Anal.*, vol. 2, 2017, doi: 10.1007/s41651-017-0010-6.

[8] M. D. Lieberman and H. Samet, "Multifaceted toponym recognition for streaming news," in *Proc. 34th Int. ACM SIGIR Conf. Res. and Develop. Inf. Retrieval*, Beijing, China, Jul. 2011, pp. 843–852.

[9] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proc. 43rd Annu. Meeting Assoc. for Comput. Linguistics*, Ann Arbor, MI, USA, Jun. 2005, pp. 363–370.

[10] R. Chasin, D. Woodward, J. Witmer, and J. Kalita, "Extracting and displaying temporal and geospatial entities from articles on historical events," *Comput. J.*, vol. 57, no. 3, pp. 403–426, Mar. 2014.

[11] M. Sagcan and P. Karagoz, "Toponym recognition in social media for estimating the location of events," in *Proc. IEEE Int. Conf. Data Mining Workshop* (*ICDMW*), Atlantic City, NJ, USA, Nov. 2015, pp. 33–39.

[12] H. Chanlekha, A. Kawtrakul, P. Varasrai, and I. Mulasas, "Statistical and heuristic rule based model for Thai named entity recognition," 2002. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.295.7088

[13] H. Chanlekha and A. Kawtrakul, "Thai named entity extraction by incoperating maximum entropy model with simple heuristic information," 2004. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.1449

[14] N. Tirasaroj and W. Aroonmanakun, "Thai named entity recognition based on conditional random fields," in *Proc. 8th Int. Symp. Natural Lang. Process.*, Oct. 2009, pp. 216–220, doi: 10.1109/SNLP.2009.5340913.

[15] S. Thattinaphanich and S. Prom-on, "Thai named entity recognition using Bi-LSTM-CRF with word and character representation," in *4th Int. Conf. Inf. Technol.* (*InCIT*), Bangkok, Thailand, Oct. 2019, pp. 149–154.

[16] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Found. Trends Mach. Learn.*, vol. 4, no. 4, pp. 267–373, Apr. 2012.

[17] P. Chormai, P. Prasertsom, J. Cheevaprawatdomrong, and A. Rutherford, "Syllable-based neural Thai word segmentation," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, Dec. 2020, pp. 4619–4637.

[18] L. Ramshaw and M. Marcus, "Text chunking using transformation-based learning," in *Proc. 3rd Workshop on Very Large Corpora*, Cambridge, MA, USA, Jun. 1995, pp. 82–94.

[19] A. Ekwonganan, "Identification of Thai and transliterated words by N-gram models," M.S. thesis, Linguistics Dept., Chulalongkorn Univ., Bangkok, Thailand, 2005.