นายธุวชิต แฉล้มเขตต์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมสำรวจ ภาควิชาวิศวกรรมสำรวจ
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

The Development of Geoparsing and Automated Classification from Thai Twitter Text Data

Mr. Tuvachit Chalamkate

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Geomatic Engineering
Department of Survey Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2022

หัวข้อวิทยานิพนธ์	การพัฒนาตัวแบบแปลความหมายทางภูมิศาสตร์และจำแนกประเภทอัตโนมัติจากข้อมูลภาษาไทยบนทวิตเตอร์
โดย	นายธุวชิต แฉล้มเขตต์
สาขาวิชา	วิศวกรรมสำรวจ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.ชนินทร์ ทินนโชติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	รองศาสตราจารย์ ดร.อรรถพล ธำรงรัตนฤทธิ์

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.เฉลิมชนม์ สถิระพจน์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รองศาสตราจารย์ ดร.ชนินทร์ ทินนโชติ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(รองศาสตราจารย์ ดร.อรรถพล ธำรงรัตนฤทธิ์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.กรวิก ตนักษรานนท์)

..... กรรมการ
(อาจารย์ ดร.ธงทิศ ฉายากุล)

..... กรรมการภายนอกมหาวิทยาลัย
(ดร.ปรัชญา บุญขวัญ)

ธุวชิต แฉล้มเขตต์ : การพัฒนาตัวแบบแปลความหมายทางภูมิศาสตร์และจำแนกประเภทอัตโนมัติจากข้อมูลภาษาไทยบนทวิตเตอร์. (The Development of Geoparsing and Automated Classification from Thai Twitter Text Data) อ.ที่ปรึกษาหลัก : รศ. ดร. ชนินทร์ ทินนโชติ, อ.ที่ปรึกษาร่วม : รศ. ดร.อรรถพล ธำรงรัตนฤทธิ์

ทวิตเตอร์เป็นแหล่งข้อมูลข่าวสารที่มีความรวดเร็วอย่างมาก ในข้อความปริมาณมหาศาลที่มีการสื่อสารกันนั้น มีข้อมูลเกี่ยวกับสถานที่ใหม่ ๆ ทั้งชื่อและข้อความที่อธิบายตำแหน่งที่ตั้ง จึงนับเป็นแหล่งข้อมูลที่สำคัญสำหรับช่วยในการปรับปรุงฐานข้อมูลภูมิสารสนเทศในระบบสารสนเทศต่าง ๆ เช่น ระบบแผนที่นำทาง ให้ทันสมัยอยู่อย่างต่อเนื่อง โดยขั้นตอนสำคัญ 2 ขั้นตอนคือ การสกัดภูมินาม เพื่อค้นหาและสกัดชื่อของสถานที่ในข้อความ และการเข้ารหัสภูมิศาสตร์ เพื่อวิเคราะห์ประมาณค่าตำแหน่งที่ตั้งทางภูมิศาสตร์ของสถานที่นั้น ในปัจจุบันการนำงานวิจัยและเครื่องมือการสกัดภูมินามที่ได้มีการพัฒนาไว้กับภาษาอื่นมาใช้กับข้อมูลภาษาไทยยังมีอยู่ค่อนข้างจำกัด และเทคนิคการเข้ารหัสภูมิศาสตร์ที่มีอยู่ก็ยังให้ค่าความถูกต้องทางตำแหน่งไม่ดีเท่าที่ควร งานวิจัยนี้พัฒนาตัวแบบเพื่อแปลความหมายทางภูมิศาสตร์ภาษาไทย โดยในการสกัดภูมินามนั้น ได้นำเทคนิคการเรียนรู้ของเครื่องได้แก่ แบบจำลอง CRF ซึ่งมีการสร้างฟังก์ชันคุณลักษณะเฉพาะทางด้านภูมิศาสตร์เพิ่มเติม โครงข่ายประสาทเทียมแบบวกกลับได้แก่ LSTM และ GRU และสุดท้ายคือแบบจำลองการถ่ายโอนความรู้ คือ BERT โดย BERT คือแบบจำลองที่ให้ค่าความถูกต้องโดยรวมในระดับคำที่สมบูรณ์ (F1-Phrase) อยู่ที่ 0.919 การเข้ารหัสภูมิศาสตร์เพื่อหาตำแหน่งของชื่อสถานที่ใหม่ที่สกัดได้นั้น ได้มีการพัฒนาอัลกอริทึมใหม่ขึ้นงานวิจัยนี้โดยการนำข้อมูลความสัมพันธ์เชิงพื้นที่ระหว่างชื่อสถานที่อื่น ๆ ที่ทราบตำแหน่งที่ตั้งในข้อความมา ใช้เป็นค่าถ่วงน้ำหนักในการประมาณตำแหน่งของสถานที่ใหม่ ให้ชื่อว่า Topology words ซึ่งจากผลการวิจัยพบว่า แบบจำลอง Topology words ให้ประสิทธิภาพดีที่สุดจากค่าเฉลี่ยกำลังสอง (Root mean square error) ต่ำที่สุดคือ 0.947 กิโลเมตร และเป็นค่าความถูกต้องที่ดีกว่าเทคนิคเดิม ๆ ที่มีอยู่ทั้ง DBSCAN, K-means, K-medoids และ Agglomerative clustering

สาขาวิชา วิศวกรรมสำรวจ ลายมือชื่อนิสิต .....

ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6071457521 : MAJOR GEOMATIC ENGINEERING

KEYWORD: Geoparsing, Toponym recognition, Geocoding

Tuvachit Chalamkate : The Development of Geoparsing and Automated Classification from Thai Twitter Text Data. Advisor: Assoc. Prof. CHANIN TINNACHOTE, D.Eng. Co-advisor: Assoc. Prof. Attapol Thamrongrattanarit, Ph.D.

Twitter is a rapid news source with a wealth of geo-referenced information. Geoparsing is the transformation of textual place names into geospatial data. For locating new locations, navigation systems and geospatial data retrieval systems are utilized. There is no such instrument for Thai language data currently. In this study, it is necessary to create a model for the geoparsing of Thai. It includes two crucial steps: Toponym recognition. geocoding In the first stage of topographic extraction, additional geographic feature functions are generated using a machine learning technique called the CRF model, the recurrent neural networks, LSTM, and GRU; and lastly, the knowledge transfer model, BERT, where BERT is the model with the highest absolute word-level accuracy (F1-Phrase). The final step is geocoding. This research extends to the estimation of a place if it cannot be determined using the existing database. An algorithm known as "topology words" incorporates the properties of referencing relationships between locations in the text. Also utilized are clustering machine learning models, including DBSCAN, K-means, K-medoids, and Agglomerative clustering. Used to designate a group of place names that will be used to estimate the location. According to the research findings, the topology word model provided the greatest performance, with the lowest root mean square error of 0.94 km.

Field of Study:	Geomatic Engineering	Student's Signature .....
Academic Year:	2022	Advisor's Signature .....
		Co-advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ไปได้ด้วยดีโดยได้รับความช่วยเหลือจากหลายฝ่าย ขอขอบพระคุณ รองศาสตราจารย์ ดร.ชนินทร์ ทินนโชติ และ รองศาสตราจารย์ ดร.อรรถพล ธำรงรัตนฤทธิ์ ที่ปรึกษาวิทยานิพนธ์ ที่ให้คำแนะนำ ความรู้และข้อคิดเห็นในการทำงานวิจัย และช่วยตรวจสอบและแก้ไขวิทยานิพนธ์นี้ให้มีความสมบูรณ์ และขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ทุกท่าน ประกอบด้วย ศ. ดร. เฉลิมชนม์ สถิระพจน์ ผศ. ดร. กรวิก ตนักษรานนท์ อ. ดร. ธงทิศ ฉายากุล และ ดร. ปรัชญา บุญขวัญ กรรมการภายนอกมหาวิทยาลัย ที่กรุณาให้คำแนะนำ และตรวจสอบวิทยานิพนธ์ฉบับนี้ให้มีความสมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณคณาจารย์ภาควิชาสำรวจ ทุกท่านที่ให้ความรู้อันมีคุณค่า

ขอขอบคุณพ่อแม่ที่ดูแลและคอยสนับสนุนลูกคนนี้มาตลอด

ขอขอบคุณเพื่อนๆ พี่ๆ นิสิตปริญญาดุษฎีบัณฑิต ภาควิชาวิศวกรรมสำรวจ (PhD Survival) ภาควิชาวิศวกรรมคอมพิวเตอร์ และเพื่อนๆ พี่ๆ น้องๆจากการประปาส่วนภูมิภาค (pwa gisdev) ที่ช่วยให้ความรู้ ประสบการณ์การพัฒนาโปรแกรม

แหล่งเรียนรู้ที่มีคุณค่า ได้แก่ stack overflow, geeks for geeks, medium, papers with code และ github

ขอขอบคุณ คุณศุภาวีร์ เปี่ยมด้วยธรรม ภรรยาที่คอยสนับสนุนและให้กำลังใจมาโดยตลอด

ธุวชิต แฉล้มเขตต์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ข้อมูลที่อ้างถึงตำแหน่งในรูปแบบของข้อความถูกพบในหลากหลายแหล่งข้อมูล เช่น สื่อสังคมออนไลน์ เว็บไซต์ข่าวสารออนไลน์ โฆษณาขายที่อยู่อาศัยบนเว็บไซต์ รีวิวร้านอาหาร ฯลฯ ซึ่งมีบทบาทสำคัญมากขึ้นในการนำไปประยุกต์ใช้ทางด้านภูมิสารสนเทศ มีการอธิบาย บอกเล่าเหตุการณ์ (เช่น ภัยพิบัติ) หรือ ข้อมูลสถานที่ที่อยู่ภายในข้อความ (Cadorel et al., 2021) สื่อสังคมออนไลน์ถือเป็นแหล่งข้อมูลที่มีความรวดเร็วในการสื่อสารและยังมีข้อมูลที่มีความถี่ รวมถึงปริมาณที่มาก โดยประเทศไทยมีผู้รับข่าวสารจากสื่อสังคมออนไลน์เป็นจำนวนถึงร้อยละ 78 จากประชากรทั้งหมด เป็นสัดส่วนที่มากที่สุดในโลก สำหรับทวิตเตอร์มีผู้ใช้งานในประเทศไทยอยู่อันดับที่ 10 ของโลกและเป็นแพลตฟอร์มที่ได้รับความนิยมในการรับรู้ข่าวสารรองจาก Facebook (Kemp, 2021) ดังนั้นจึงเป็นไปได้ว่ามีข้อมูลอีกมากมายที่เราสามารถนำมาใช้ประโยชน์ได้ในหลากหลายวัตถุประสงค์ ซึ่งพบว่าข้อมูลดิจิตอลเหล่านี้ทั้ง สื่อสังคมออนไลน์ เว็บบล็อก ฯลฯ กว่าร้อยละ 60 มีการอ้างอิงเชิงตำแหน่ง (Hahmann & Burghardt, 2013)

ในอดีตที่ผ่านมามีการศึกษาความสัมพันธ์ระหว่างตำแหน่งทางภูมิศาสตร์และภาษาที่ใช้โดยให้คำจำกัดความว่าศาสตร์ภาษาถิ่น (Dialectology) มีสมติฐานว่าการใช้ภาษาจะมีลักษณะเฉพาะแตกต่างกันไปตามแต่ละภูมิภาค (Chambers & Trudgill, 1998) นอกจากนี้พบว่ามีการศึกษาทางด้านนี้ในอีกหลายงานวิจัย เช่น ศึกษาการกระจายตัวของภาษาถิ่นที่ใช้ในสหรัฐอเมริกาจากข้อมูลทวิตเตอร์ (Huang et al., 2016) และงานวิจัยของ Gonçalves and Sánchez (2014) นำข้อมูลจีโอแท็ก (geotagged) จากไมโครบล็อค (micro blog) มาจัดกลุ่มการใช้ภาษาสเปนแต่ละพื้นที่ในทวีปอเมริกาทั้งเหนือและใต้ หรือการนำข้อมูลทวิตเตอร์ที่มีจีโอแท็กมาศึกษาเปรียบเทียบกับข้อมูลในคลังคำ (corpora) ซึ่งพบตัวแปรที่ซ่อนอยู่ในข้อมูลเหล่านั้นโดยเฉพาะทางด้านประชากรศาสตร์ที่เกี่ยวเนื่องกับการใช้ภาษาตามลักษณะทางภูมิศาสตร์ (Pavalanathan & Eisenstein, 2015)

และในปัจจุบันนักวิจัยในศาสตร์ของการรวบรวมข้อมูล (information retrieval) และการประมวลผลภาษาธรรมชาติ (natural language processing : NLP) มีความพยายามสร้างเครื่องมือเพื่อที่จะสกัดเอาข้อมูลเชิงพื้นที่ออกมาจากข้อความซึ่งเป็นข้อมูลที่ไม่มีโครงสร้าง (unstructured data) เพื่อที่จะนำข้อมูลเหล่านี้มาใช้ประโยชน์ เช่น การศึกษาด้านภูมิศาสตร์มนุษย์ การศึกษาทางสังคมศาสตร์ โดยมุ่งเน้นในการศึกษาแง่มุมทางวิทยาศาสตร์ที่เป็นเชิงปริมาณ (Melo & Martins, 2017)

กระบวนการเปลี่ยนข้อมูลที่อยู่ในรูปของตัวอักษร (Text File) เป็นข้อมูลที่มีค่าพิกัดทางภูมิศาสตร์หรือข้อมูลที่อ้างอิงได้ในเชิงตำแหน่ง มีชื่อเรียกต่างกันไปในหลากหลายงานวิจัย เช่น Geographic Information Retrieval (GIR), Geographic Information Extraction (GIE), Geographic Analysis (GA) (Steinberger et al., 2013) โดยทั่วไปมีขั้นตอนสำคัญ 2 ขั้นตอน คือ 1) การรู้จำภูมินาม (Toponym recognition, Toponym extraction, Geotagging) และ 2) การเข้ารหัสทางภูมิศาสตร์ (Geocoding, Toponym resolution, Toponym disambiguating) โดยขั้นตอนแรกใช้สำหรับบอกหน้าที่ของคำเพื่อสกัดข้อมูลของสถานที่นั้นออกมาจากข้อความอื่น ส่วนขั้นตอนที่ 2 ทำหน้าที่ลดความกำกวมของคำที่สกัดออกมาและเชื่อมต่อไปยังค่าพิกัดทางภูมิศาสตร์ (Gritta et al., 2018) การรู้จำภูมินามเป็นวิธีการที่นักภูมิสารสนเทศนำมาจากสาขาหนึ่งในการประมวลผลภาษาธรรมชาติ เรียกว่า “การรู้จำชื่อเฉพาะ (Named Entity Recognition : NER) ” อย่างไรก็ตาม การพัฒนาเครื่องมือเพื่อทำ NER ในอดีตใช้คลังข้อมูลที่มาจากการรวบรวมข้อมูลจากบทความ งานวิจัย ข่าว แต่ในปัจจุบันข้อมูลที่ได้มาจากหลายแหล่งโดยเฉพาะสื่อสังคมออนไลน์ดังที่กล่าวไว้ในข้างต้น ซึ่งข้อมูลเหล่านี้นอกจากจะไม่มีโครงสร้างแล้ว ยังใช้ภาษาที่ไม่เป็นทางการ มีการใช้ตัวย่อ มีการสะกดผิด ขาดเครื่องหมายวรรคตอนที่เหมาะสม รวมทั้งปัญหาอีกประการหนึ่งคือ ผู้ให้บริการบางราย (เช่น ทวิตเตอร์) จำกัดจำนวนตัวอักษรที่ใช้ต่อข้อความ ส่งผลให้เครื่องมือในการทำ NER ที่มีการสร้างไว้ไม่สามารถรู้จำชื่อเฉพาะจากข้อมูลเหล่านี้ได้ตามที่ควรจะเป็น (Lingad et al., 2013; Murnane, 2010)

นอกจากปัญหาที่ได้กล่าวถึงข้างต้นแล้ว ในแต่ละภาษายังมีโครงสร้างทางภาษาและการใช้งานที่แตกต่างกัน สำหรับภาษาไทย Chanlekha and Kawtrakul (2004) ได้ศึกษาลักษณะปัญหาของภาษาไทยที่ทำให้ยากต่อการรู้จำชื่อเฉพาะ สรุปได้ดังนี้

1) ภาษาไทยไม่มีข้อมูลบ่งบอกถึงชื่อเฉพาะ เช่น ภาษาอังกฤษใช้ตัวอักษรพิมพ์ใหญ่นำหน้าชื่อเฉพาะ

2) ภาษาไทยไม่มีการเว้นวรรคหรือใช้อักขระพิเศษในการแบ่งคำทำให้ตัดแบ่งคำมีความกำกวมและตัดคำได้หลายแบบ

3) ลักษณะการสร้างชื่อเฉพาะไม่มีหลักเกณฑ์ที่แน่นอน สามารถสร้างด้วยคำใดก็ได้ ทำให้ยากต่อการสร้างกฎเพื่อใช้จำแนก

4) ลักษณะงานเขียนของภาษาไทยนิยมเขียนชื่อเต็มในครั้งแรก หลังจากนั้นหากกล่าวถึงชื่อเฉพาะอีกครั้งจะใช้ตัวย่อ หรือลดทอนบางคำไป ทำให้เกิดความกำกวมระหว่างชื่อเฉพาะและคำนามทั่วไปได้ เช่นตัวอย่างประโยคต่อไปนี้ “ แหล่งข่าวจากบริษัท **ห้างสรรพสินค้าโรบินสัน จำกัด (มหาชน)** เปิดเผยกับฐานเศรษฐกิจว่า **ห้างสรรพสินค้าโรบินสัน** ได้ตัดสินใจ ... สำหรับ**โรบินสัน** สาขาสีลมเปิดให้บริการมามากกว่า 30 ปี ” จะสังเกตเห็นว่าทั้งสามคำที่มีการเน้นในประโยคตัวอย่างเป็นคำที่สื่อความหมายถึงสิ่งเดียวกันแต่เขียนไม่เหมือนกันในแต่ละครั้งที่กล่าวถึง

จากสภาพปัญหาที่กล่าวมา งานวิจัยนี้จึงมุ่งเน้นที่จะศึกษาและพัฒนากระบวนการ การแปลความหมายทางภูมิศาสตร์ (Geographic Information Retrieval : GIR, Geographic Information Extraction : GIE, Geoparsing) สำหรับข้อความภาษาไทยบนทวิตเตอร์ เพื่อนำไปใช้ประโยชน์ในการสำรวจหาสถานที่ใหม่ รวมถึงการประมาณตำแหน่งเบื้องต้นจากข้อความ ซึ่งกระบวนการแปลความหมายทางภูมิศาสตร์ (Geoparsing) มีส่วนสำคัญ 2 ส่วน ประกอบไปด้วย 1) การรู้จำภูมินาม และ 2) การเข้ารหัสทางภูมิศาสตร์ ในขั้นตอนการเข้ารหัสภูมิศาสตร์ ชื่อเฉพาะสถานที่ที่เป็นภาษาไทย ผู้ให้บริการออนไลน์ เช่น Google map API , Bing map ฯลฯ ยังมีข้อจำกัดในด้าน ความครบถ้วน ของชื่อภาษาไทย (Manoruang & Asavasuthirakul, 2019b) ในงานวิจัยนี้จึงเพิ่มเติมการประมาณตำแหน่งทางภูมิศาสตร์ (Location Estimation) สำหรับสถานที่ที่ไม่สามารถเข้ารหัสภูมิศาสตร์จากผู้ให้บริการออนไลน์ได้

โดยการรู้จำทางภูมินามจะศึกษาการใช้เทคนิคสำคัญคือ แบบจำลอง Conditional Random Fields (CRFs) การใช้โครงข่ายประสาทเทียมวกกลับ แบบประตูสัญญาณวกกลับ (Gated Recurrent Unit : GRU) แบบหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory : LSTM) การนำโครงข่ายประสาทเทียมแบบ GRU และ LSTM มาใช้ร่วมกับแบบจำลอง CRF (Iu & Zou, 2018; Sagcan & Karagoz, 2015) และสุดท้ายคือการนำแบบจำลอง Bidirectional Encoder Representations from Transformers : BERT ซึ่งเป็นแบบจำลองการเรียนรู้เชิงลึก (Deep Learning) (Ma et al., 2022)

ในส่วนที่สอง ผลจากการดึงภูมินามจากข้อความข้างต้นอาจจะมีบางส่วนที่เป็นชื่อสถานที่ที่ไม่ได้มีอยู่จริงซึ่งอาจเกิดจากข้อจำกัดบางประการของแบบจำลองหรือสิ่งรบกวนอื่น (noise) การเข้ารหัสทางภูมิศาสตร์จึงเป็นการกรองชื่อสถานที่ที่ไม่อยู่จริงออกไปหรือนัยหนึ่งชื่อนั้นอาจเป็นชื่อของสถานที่ใหม่ซึ่งยังไม่มีอยู่ในฐานข้อมูลใด ในส่วนที่สองนี้จะเป็นการประมาณขอบเขตของภูมินามที่อาจเป็นสถานที่ใหม่นี้ โดยการเข้ารหัสทางภูมิศาสตร์นี้จะเป็นการนำชื่อมาจับคู่กับข้อมูลในฐานข้อมูลที่จัดเตรียมไว้ โดยมีการทำความสะอาด จัดรูปแบบ และประมวลผลความคล้ายระหว่างภูมินามที่สกัดออกมากับชื่อที่อยู่ในฐานข้อมูล ในงานวิจัยนี้จะใช้ฐานข้อมูล 2 ส่วนคือ ฐานข้อมูลที่สร้างขึ้นเองจากแหล่งข้อมูลขอบเขตการปกครอง จังหวัด อำเภอ ตำบล จากบริษัท Nostra ซึ่งได้รับความอนุเคราะห์จากสำนักเทคโนโลยีภูมิสารสนเทศ การประปาส่วนภูมิภาค ข้อมูลภูมินามและสถานที่สำคัญจากกรมแผนที่ทหาร และสุดท้ายข้อมูลจาก Google API

โดยสำหรับข้อมูลสถานที่ซึ่งไม่สามารถจับคู่กับข้อมูลในอักขรานุกรมทางภูมิศาสตร์ข้างต้น นำการประมาณค่าเชิงตำแหน่งจากบริบทของคำโดยรอบมาประยุกต์ใช้เรียกวิธีนี้ว่า Location Indicative Words : LIW (Han et al., 2012) หลังจากนั้นจึงหาขอบเขตของตำแหน่งสถานที่นั้นด้วยอัลกอริทึมที่ทางผู้วิจัยออกแบบโดยนำแนวคิดมาจาก Location Indicative Words : LIW สร้าง

ความสัมพันธ์ระหว่างชื่อสถานที่ในประโยค และคำที่บ่งบอกตำแหน่ง เช่น “อยู่ใกล้” “ติดกับ” “ถัดจาก” ฯลฯ มาสร้างกฎ (Rule based) ให้คำจำกัดความวิธีนี้ว่า Topology words และมีการเปรียบเทียบความถูกต้องกับเทคนิคการเรียนรู้ของเครื่อง (Machine learning) แบบจัดกลุ่ม (Clustering algorithm) อีก 4 วิธีคือ 1) เทคนิค Density-Based Spatial Clustering of Applications with Noise : DBSCAN (Tang et al., 2015) 2) เทคนิคการจัดกลุ่มด้วย K-means (Cosentino et al., 2022) 3) การจัดกลุ่มด้วย K-medoids (Sureja et al., 2022) และสุดท้ายคือ 4) การจัดกลุ่มแบบลำดับขั้น Agglomerative clustering (Tie et al., 2019) เพื่อระบุขอบเขตที่มีความเป็นไปได้ให้จำกัดลงจากชุดข้อมูล

สำหรับซอร์สโค้ดและตัวอย่างชุดข้อมูลที่ใช้ในงานวิจัยนี้สามารถเข้าถึงได้จาก https://github.com/crescendonow/thai_geoparsing

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 ศึกษาและพัฒนากระบวนการแปลความหมายทางภูมิศาสตร์ (geoparsing) สำหรับข้อความภาษาไทยจากทวิตเตอร์ เพื่อให้การสกัดข้อมูลเชิงพื้นที่จากข้อความมีประสิทธิภาพเพียงพอในการใช้งานทั้งในทางภาษาศาสตร์และภูมิศาสตร์

1.2.2 พัฒนากระบวนการจำแนกประเภทของชื่อสถานที่อัตโนมัติจากข้อความ เพื่อลดขั้นตอนในการจัดระเบียบข้อมูลก่อนเข้าสู่ระบบภูมิสารสนเทศ

1.3 สมมติฐานของงานวิจัย

1.3.1 การพัฒนาแบบจำลองการรู้จำภูมินามโดยเฉพาะสำหรับข้อความภาษาไทยด้วยอัลกอริทึม CRF หรือโครงข่ายประสาทเทียม GRU, LSTM หรือ BERT สามารถให้ความถูกต้องที่เหมาะสมกับการสกัดข้อมูลสถานที่จากทวิตเตอร์

1.3.2 อัลกอริทึมที่พัฒนาขึ้นจากคำระบุตำแหน่ง Topology words รวมทั้งการนำอัลกอริทึมในการจัดกลุ่มที่กล่าวข้างต้น DBSCAN, K-means ฯลฯ มาใช้ในการประมาณตำแหน่งของสถานที่ จะสามารถช่วยในการจำกัดขอบเขตของสถานที่ให้มีความถูกต้องใกล้เคียงกับตำแหน่งจริง

1.4 ขอบเขตงานวิจัย

1.4.1 รวบรวมข้อมูลจากสื่อสังคมออนไลน์ โดยในงานวิจัยนี้ใช้ข้อมูลจากทวิตเตอร์เป็นหลัก มีพื้นที่ กรุงเทพมหานครและปริมณฑลเป็นพื้นที่ต้นแบบในการทดลอง

1.4.2 งานวิจัยนี้มุ่งเน้นศึกษาทดลองในข้อความที่เป็นภาษาไทยเท่านั้นสำหรับในบางประโยคที่มีการใช้คำภาษาไทยเขียนทับคำศัพท์ภาษาอังกฤษ ให้ถือว่าเป็นส่วนหนึ่งของข้อความภาษาไทย

1.4.3 ประสิทธิภาพของแบบจำลองหมายถึงประสิทธิภาพที่ได้จากชุดข้อมูลในงานวิจัยนี้เท่านั้น

1.4.4 สำหรับการเข้ารหัสภูมิศาสตร์ ในงานวิจัยนี้เข้ารหัสภูมิศาสตร์ในลักษณะที่เป็นจุด (Point) เท่านั้น จึงไม่รวมข้อมูลที่เป็น เส้นทางคมนาคมและขอบเขตการปกครองที่เป็นข้อมูลแบบเส้น (Line) และพื้นที่รูปปิด (Polygon) โดยข้อมูลทั้ง 2 ชนิดนี้จะทำฐานข้อมูลไว้อ้างอิงโดยเฉพาะ

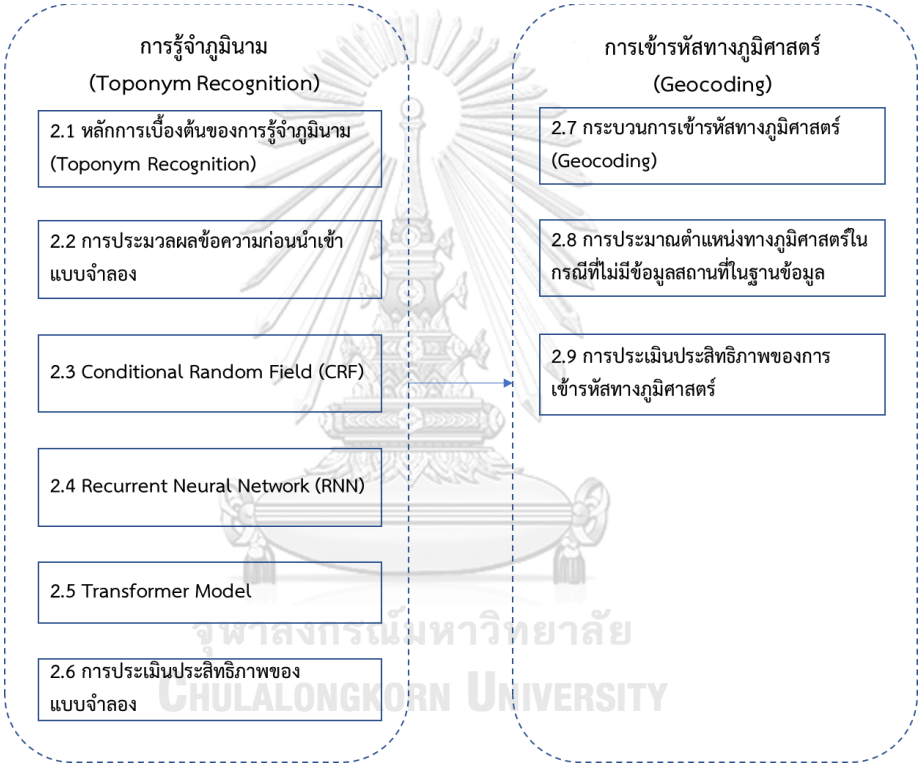
1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 แบบจำลองการรู้จำภูมินามที่เหมาะสมกับข้อความภาษาไทยบนทวิตเตอร์ รวมทั้งการนำไปประยุกต์ใช้กับข้อมูลสื่อสังคมออนไลน์อื่นในอนาคต

1.5.2 ลดขั้นตอนในการจำแนกประเภทของสถานที่ สามารถนำข้อมูลที่ได้จากแบบจำลองไปใช้ได้อย่างรวดเร็วและถูกต้อง

บทที่ 2
หลักการ แนวคิดและทฤษฎีที่เกี่ยวข้อง

จากที่ได้กล่าวถึงในบทนำ กระบวนการแปลความหมายทางภูมิศาสตร์ มีขั้นตอนสำคัญ 2 ขั้นตอนประกอบไปด้วย ขั้นตอนที่ 1 การรู้จำภูมินาม ทำให้ทราบว่าคำไหนคือชื่อเฉพาะของสถานที่ และขั้นตอนที่ 2 คือการเข้ารหัสทางภูมิศาสตร์เพื่อให้ได้ข้อมูลเชิงตำแหน่ง สำหรับขั้นตอนที่ 1 จะกล่าวถึงในหัวข้อที่ 2.1 - 2.5 ขั้นตอนที่ 2 จะกล่าวถึงในหัวข้อที่ 2.6 -2.9 โดยภาพรวมรายละเอียดสำคัญแสดงในรูปที่ 1



รูปที่ 1 แผนผังแสดงขั้นตอนในการสร้างกระบวนการแปลความหมายทางภูมิศาสตร์

2.1 หลักการเบื้องต้นของการรู้จำภูมินาม (Toponym Recognition)

การรู้จำภูมินามมีวัตถุประสงค์หลักคือ การเปลี่ยนชื่อเฉพาะของสถานที่ (Place names) ไปเป็นค่าพิกัดภูมิศาสตร์จากข้อความทั่วไปซึ่งไม่มีโครงสร้างข้อมูล (Free text) โดยมีขั้นตอนหลักๆ สำคัญ 2 ขั้นตอนคือ 1) การสกัดภูมินามออกมาจากข้อความ (Toponym Recognition หรือ Geotagging) และ 2) การลดความกำกวมและเชื่อมต่อไปยังการระบุค่าพิกัดภูมิศาสตร์ (Toponym Resolution หรือ Geocoding) (Gritta et al., 2019) สำหรับขั้นตอนการสกัดภูมินามออกจาก

ข้อความมีการนำวิธีรู้จำชื่อเฉพาะหรือ Name Entity Recognition : NER มาประยุกต์ใช้งาน การรู้จำชื่อเฉพาะ หรือ NER เป็นศาสตร์ในการประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) ที่มุ่งเน้นการสกัดข้อมูล (Information Extraction) เพื่อเป็นเครื่องมือในการรู้จำหน้าที่ของคำจากข้อความ โดยข้อมูลที่สกัดออกมาผ่าน NER หลัก ๆ จะเป็น บุคคล องค์และ สถานที่ ในงานด้านภูมิสารสนเทศจะใช้ NER เพื่อสกัดข้อมูลที่บ่งบอกถึงสถานที่ออกมา (Nur Yasir Utomo et al., 2018) โดยในงานวิจัยนี้นำอัลกอริทึมในการสร้าง NER คือ Conditional Random Fields : CRF แบบจำลอง Gated Recurrent Unit : GRU และแบบจำลองการถ่ายโอนความรู้ BERT มาสร้าง Toponym Recognition โดยจะกล่าวถึงรายละเอียดต่อในหัวข้อที่ 2.3 – 2.4

2.2 การประมวลผลข้อความก่อนนำเข้าแบบจำลอง

การประมวลผลข้อความจากแหล่งที่ได้มา ในขั้นตอนนี้มีลำดับการทำงานที่สำคัญ คือ การแบ่งคำในประโยคพร้อมทั้งเก็บข้อมูลไปใช้งานต่อ (Word tokenization) และการแทนข้อความ (Text representation) ด้วยการสร้างคำฝังตัว (Word embedding)

2.2.1 การตัดคำ (Word tokenization)

ในการทำงานเกี่ยวกับข้อมูลที่เป็นข้อความหรือตัวอักษร การเปลี่ยนข้อมูลที่เป็นลำดับเกี่ยวเนื่องกัน (Sequence data) ของตัวอักษรให้กลายเป็นข้อมูลลำดับของคำพร้อมนำไปใช้งานต่อ โดยสำหรับภาษาที่เป็นระบบภาษาเขียนแทนเสียง (alphabetic languages) เช่นภาษาอังกฤษ หรือภาษาไทย คำหนึ่งคำมักจะประกอบด้วยตัวอักษรมากกว่า 1 ตัวเสมอ ทำให้ต้องมีการตัดคำหรือตัดพยางค์ (Syllable) เพื่อความสะดวกในการนำข้อมูลไปใช้งาน (Eisenstein, 2019) สำหรับการตัดคำในภาษาไทยมีคลังคำสั่งให้สามารถเลือกใช้ได้ในภาษาไพทอน เช่น newmm ใน pythainlp ซึ่งใช้อัลกอริทึม maximal matching ในการตัดคำ หรือ แบบจำลอง attacut พัฒนาขึ้นโดยใช้โครงข่ายประสาทเทียมแบบคอนโวลูชัน (Chormai et al., 2019)

2.2.2 คำฝังตัว (Word embedding)

เป็นการแทนข้อความด้วยกลุ่มของเวกเตอร์ โดยการสร้างเวกเตอร์ของคำจะวิเคราะห์ข้อความจากชุดข้อมูลทั้งหมดก่อนแล้วจึงสร้างเวกเตอร์ของคำโดยให้คู่ของคำที่มีความหมายใกล้เคียงควรจะต้องมีระยะห่างของเวกเตอร์ใกล้เคียงกัน (Tretasayuth, 2017) สำหรับภาษาไทยมีคลังคำสั่งในภาษาไพทอนที่นิยมใช้งานคือ thai2fit โดย thai2fit ได้รับการพัฒนามาจากข้อมูล Thai-Wikipedia ปัจจุบัน thai2fit มีคลังคำศัพท์กว่า 55,677 คำ พัฒนาขึ้นด้วยวิธี ULMFit (Howard, 2018) แต่ละคำนั้นถูกแทนโดยเวกเตอร์ขนาด 300 มิติ (Polpanumas, 2019)

2.3 Conditional Random Field (CRF)

แบบจำลอง Conditional Random Fields (CRF) 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 (Tirasaroj & Aroonmanakun, 2009) โดยแบบจำลอง CRF ที่นำมาใช้งานเป็นแบบห่วงโซ่ตรงหรือ Linear-Chain CRF ซึ่งผู้เขียนจะขอใช้อักษรย่อว่า CRF ในการอ้างถึงต่อไป แสดงเป็นความสัมพันธ์ได้ตามสมการที่ (1) (Sutton & McCallum, 2012)

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

โดยที่

x คือ ลำดับของผลลัพธ์ค่าสังเกต

y คือ ลำดับของผลลัพธ์ค่าที่ทำนาย

θ_k คือ ค่าน้ำหนักของฟังก์ชันคุณสมบัติ

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

T คือ ลำดับของเหตุการณ์ที่ต่อเนื่องตั้งแต่ $t_1, \dots, t_T$

K คือ จำนวนของฟังก์ชันคุณสมบัติที่นำมาหาค่าน้ำหนักในตำแหน่งของเหตุการณ์นั้นๆตั้งแต่ $k_1, \dots, k_K$

$Z(x)$ คือ การปรับข้อมูลให้เข้ากับบรรทัดฐาน (normalization) ซึ่งคำนวณได้จากสมการที่ (2)

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

2.3.1 ฟังก์ชันคุณสมบัติ (Feature function)

ฟังก์ชันคุณสมบัติเป็นองค์ประกอบหนึ่งที่สำคัญในแบบจำลอง CRF แสดง ความสัมพันธ์ดังที่อ้างถึงไว้ในสมการที่ (1) ดังนี้ $f_k(y_t, y_{t-1,} x_t)$

โดยที่

k	คือ ดัชนีของฟังก์ชัน
y_{t-1}	คือ แท็กของผลลัพธ์ค่าที่ทำนายก่อนหน้า
y_t	คือ แท็กของผลลัพธ์ค่าที่ทำนายปัจจุบัน
x_t	คือ ลำดับของข้อมูลนำเข้า ณ ตำแหน่งใด ๆ

จากตัวอย่างของ Zhu (2009) การกำหนดค่าของฟังก์ชันคุณสมบัติที่ให้ค่าแบบมีสองค่า คือ 0 กับ 1 เท่านั้น เช่น กำหนดให้ค่าเป็น 1 เมื่อคำปัจจุบัน x_t เป็น John และแท็กปัจจุบัน y_t เป็น

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

จากตัวอย่างของฟังก์ชันคุณสมบัติลำดับที่ 1 หรือ f_1 หมายความว่า หากมีคำว่า John ปรากฏอยู่ ณ ตำแหน่งใด และแท็กของ y_t คือ NER ชนิด PERSON ค่าของฟังก์ชันคุณสมบัติจะมีค่าเป็น 1 และเมื่อพิจารณาถึงค่าน้ำหนักของฟังก์ชันคุณสมบัติ θ_1 หากค่าของ $\theta_1 > 0$ เมื่อพบคำว่า John และ PERSON ปรากฏอยู่ ณ ตำแหน่งเดียวกัน จะเป็นการเพิ่มค่าความน่าจะเป็นให้กับ y ตามสมการ (1) ส่งผลต่อความน่าจะเป็นในแบบจำลอง CRF หากพบคำว่า John ในข้อมูลอื่นที่เรานำไปใช้งาน CRF จะทำนายแท็ก NER เป็น PERSON ในทางกลับกัน หาก $\theta_1 < 0$ แบบจำลอง CRF การทำนายแท็ก NER เป็น PERSON หากพบคำว่า John และกรณีสุดท้ายคือ $\theta_1 = 0$ จะหมายถึงฟังก์ชันนี้ไม่มีผลต่อค่าความน่าจะเป็นของ y

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

ตัวอย่างของฟังก์ชันคุณสมบัติลำดับที่ 2 หรือ f_2 จะให้ค่าเป็น 1 ก็ต่อเมื่อ แท็กของคำในตำแหน่งปัจจุบันเป็น NER ชนิด PERSON และคำถัดไป หรือ x_{t+1} เป็นคำว่า “said” และหากไม่เป็นตามเงื่อนไขจะให้ค่าเป็น 0 จากฟังก์ชันคุณสมบัติทั้งสองตามตัวอย่างข้างต้น สามารถนำมาใช้

ร่วมกันได้ ซึ่งแบบจำลอง CRF สามารถมีฟังก์ชันคุณสมบัติแล้วแต่ผู้ใช้งานจะกำหนด โดยค่าที่ได้จากฟังก์ชันคุณสมบัติสามารถกำหนดได้มากกว่า 2 ค่า ไม่ใช่เพียงแค่ 1 กับ 0 ตามตัวอย่างที่กล่าวถึงเท่านั้น โดยมีเงื่อนไขคือค่าที่กำหนดต้องเป็นสมาชิกของจำนวนจริง (Zhu, 2009)

2.3.2 การฝึกฝนแบบจำลองเพื่อนำไปใช้งาน (Training model)

วัตถุประสงค์ในการฝึกฝนแบบจำลองคือการหาค่าพารามิเตอร์ที่สำคัญในแบบจำลอง CRF นั่นคือ ชุดค่าน้ำหนักของฟังก์ชันคุณสมบัติ หรือ $\theta = \{\theta_1, \theta_2, \ldots \theta_k\}$ โดยใช้วิธีการประมาณค่าความน่าจะเป็นสูงที่สุด (Maximum Likelihood Estimation : MSE) เขียนแบบจำลองได้ตามสมการที่ (3)

$$\ell(\theta) = \sum_{i=1}^{N} log\, p\left(y^i \middle| x^i\right) \qquad (3)$$

โดยที่

i	คือ ดัชนีของ x^i, y^i
x^i	คือ ลำดับของข้อมูลเข้าหรือค่าสังเกต
y^i	คือ ลำดับของผลลัพธ์หรือข้อมูลที่ทำนายได้

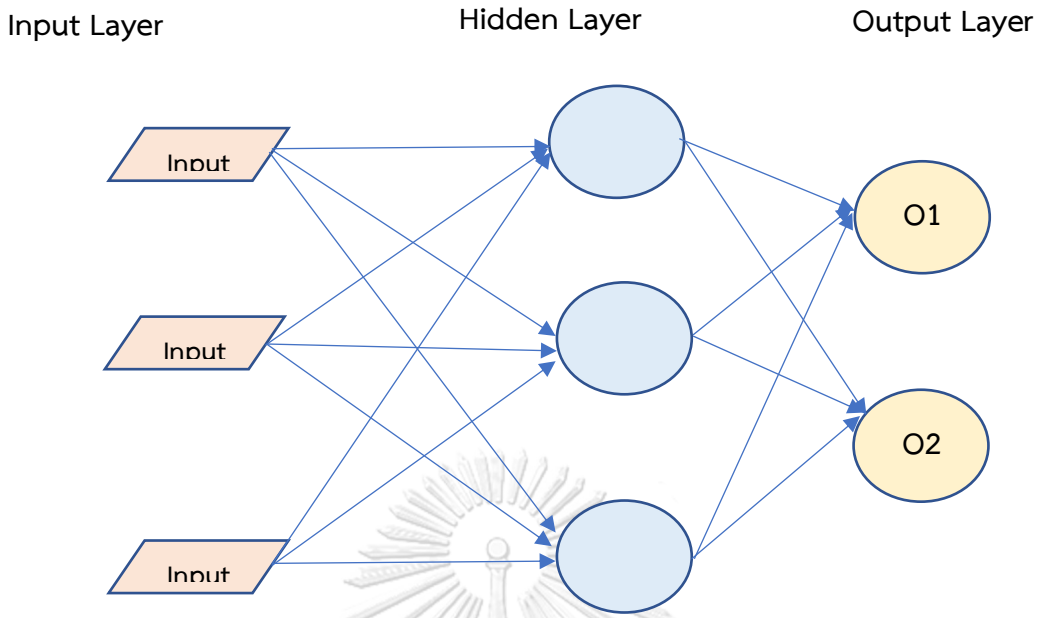
ในการประมาณค่าเพื่อหาคำตอบของฟังก์ชันสามารถทำให้ลู่เข้าสู่คำตอบ หรือ Global Optima ได้ช้า จำเป็นต้องใช้ฟังก์ชันเพื่อหาค่าเหมาะสม (Optimization Function) เช่น Limited memory quasi-newton method (L-BFGS) หรือ Stochastic Gradient Descent (SGD) ซึ่งสามารถเลือกใช้งานได้ตามความเหมาะสมของข้อมูล (Sutton & McCallum, 2012)

2.4 Recurrent Neural Network (RNN)

โครงข่ายประสาทเทียมแบบวกกลับ หรือ RNN ถูกออกแบบมาให้ตอบสนองต่อการประมวลผลข้อมูลที่มีลำดับ (Sequential) โดยในหัวข้อนี้จะอธิบายถึง สถาปัตยกรรมทั่วไปของโครงข่ายประสาทเทียม รวมทั้งโครงข่ายประสาทเทียมแบบวกกลับ (Recurrent Neural Network :RNN) แล้วจึงสรุปที่ GRU และ LSTM เนื่องจากเป็นโครงข่ายประสาทเทียมที่พัฒนาขึ้นมาจากโครงข่ายประสาทเทียมแบบวกกลับ

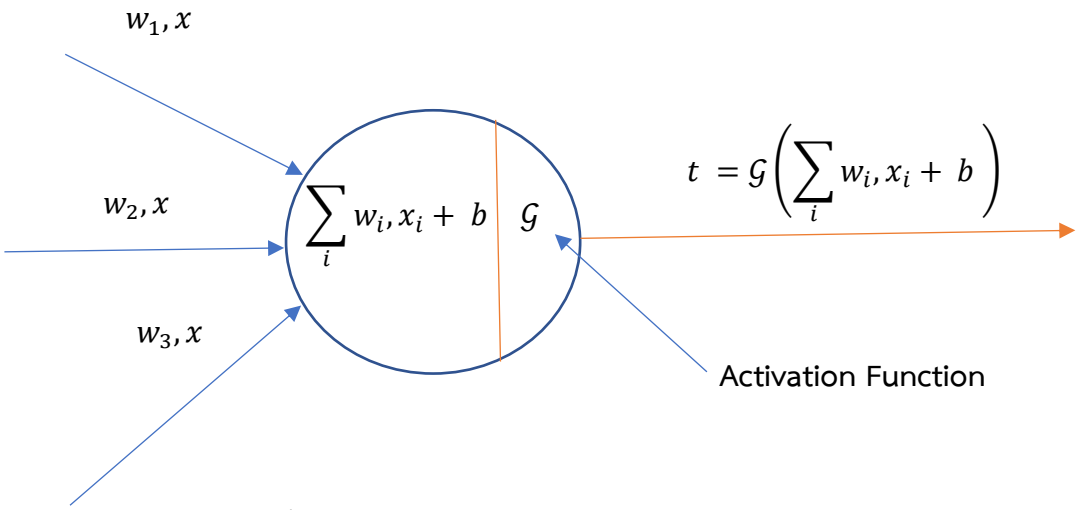
2.4.1 สถาปัตยกรรมทั่วไปของโครงข่ายประสาทเทียม

สถาปัตยกรรมโดยทั่วไปของโครงข่ายประสาทเทียมจะประกอบไปด้วย 3 ส่วนหลักๆที่สำคัญ คือ ชั้นข้อมูลนำเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และ ชั้นข้อมูลส่งออก (Output Layer) แสดงตามรูปที่ 2



รูปที่ 2 สถาปัตยกรรมทั่วไปของโครงข่ายประสาทเทียม

จากรูปที่ 2 โครงข่ายประสาทเทียมจะประกอบไปด้วย Perceptron ค่าน้ำหนักของเส้นที่เชื่อมระหว่าง Perceptron ค่าความลำเอียง (Bias) และค่าส่งออก (Output) โดยโครงข่ายประสาทเทียมจะใช้ Perceptron ในการรับค่านำเข้า (Input) จากนั้นจะรวมค่าที่ส่งเข้ามาจากชั้นก่อนหน้าทั้งหมดด้วยสมการเชิงเส้น และกระตุ้นด้วยฟังก์ชันกระตุ้นที่กำหนด (เป็นสมการที่ไม่เป็นเชิงเส้น Non-Linear equation) ก่อนส่งต่อค่าที่ได้จากการคำนวณไปยัง Perceptron ชั้นต่อไป หรือส่งเป็นค่าส่งออกไปยังข่ายประสาทในชั้นถัดไป



รูปที่ 3 แสดงลักษณะการรับส่งข้อมูลของ Perceptron

จากรูปที่ 3 กำหนดฟังก์ชันของ Perceptron แทนด้วย $f(x)$ เป็นฟังก์ชันที่ใช้กำหนดค่าส่งออก แสดงตามสมการที่ (4)

$$t = f(x) = \begin{cases} 1, \text{ if } \mathcal{G}(\sum_{i=1}^{m} w_i x_i + b) > 0 \\ 0, \text{ if Otherwise} \end{cases} \quad (4)$$

โดยที่

t	คือ ค่าส่งออก
x	คือ ข้อมูลรับเข้า
w_i	คือ เวกเตอร์ของค่าน้ำหนัก
b	คือ ค่าความลำเอียง
m	คือ ขนาดของข้อมูลรับเข้า
$\mathcal{G}$	คือ ฟังก์ชันกระตุ้น (Activation Function)

จากสมการที่ (4) ค่าส่งออกหลังจากที่ผ่านฟังก์ชันกระตุ้นแล้วจะมีค่าเป็น 1 หรือ 0 โดยฟังก์ชันที่รับค่าไปคำนวณได้ผลลัพธ์ออกมาและไปสอดคล้องกับเงื่อนไขใด ค่าน้ำหนักของ Perceptron จะถูกปรับอัตราการเรียนรู้ เพื่อให้ได้มาซึ่งผลรวมเชิงเส้นถ่วงน้ำหนัก อัตราการเรียนรู้ (Learning rate) มีลักษณะเป็นค่าคงที่บวกที่ส่งผลต่อการลู่เข้าสู่คำตอบของโครงข่ายประสาทเทียม หากอัตราการเรียนรู้มีค่ามากโครงข่ายสามารถเรียนรู้ได้แต่มีความเสี่ยงที่คำตอบที่ได้อาจจะไม่ใช่ค่าที่เป็นจุดต่ำสุดทั้งหมด (Global Minima) แต่หากปรับค่าอัตราการเรียนรู้น้อยจนเกินไปอาจได้คำตอบที่ละเอียดถูกต้องดีกว่าแต่ใช้เวลาในการคำนวณที่มากขึ้น (Daniel, 2013)

2.4.1.1 โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า

โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้าหรือ Feed Forward Neural Network มีลำดับการส่งผ่านข้อมูลในทิศทางเดียว ซึ่งโครงสร้างจะแบ่งเป็นลำดับชั้น ในแต่ละลำดับชั้นเดียวกันจะมี Perceptron ที่ไม่มีการเชื่อมต่อกัน แต่จะมีเส้นที่เชื่อมต่อกันในชั้นก่อนหน้าและชั้นถัดไป โดยข้อมูลที่ส่งออกจากชั้นก่อนหน้าจะกลายเป็นข้อมูลเข้าของชั้นปัจจุบัน โดยสามารถคำนวณหาค่าของผลลัพธ์ในชั้นถัดไปได้จากสมการที่ (5) และ (6) แสดงตัวอย่างในรูปที่ 4

$$z_j^l = \sum_{k=1}^{n} w_{jk}^l a_k^{l-1} + b_j^l \quad (5)$$

$$a_j^l = \mathcal{G}(z_j^l) \quad (6)$$

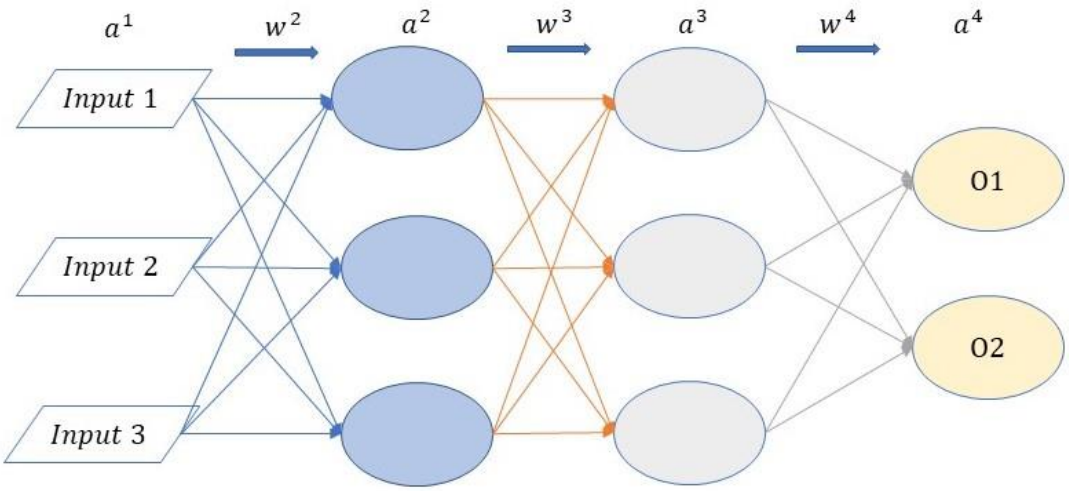
โดยที่

a_k^{l-1} คือ ผลลัพธ์ของ *Perceptron* ตัวที่ k ในลำดับชั้น $l-1$

w_{jk}^{l} คือ ค่าน้ำหนักของ Perceptron ตัวที่ j ในลำดับชั้น l ที่มีเส้นเชื่อมมาจาก Perceptron ตัวที่ k ในลำดับชั้นก่อนหน้า

b_j^{l} คือ ค่าความลำเอียง

$\mathcal{G}$ คือ ฟังก์ชันกระตุ้น



รูปที่ 4 แสดงลำดับของการส่งผ่านข้อมูลในโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า

2.4.1.2 ฟังก์ชันกระตุ้น (Activation function)

ข้อมูลก่อนที่จะส่งออกจาก Perceptron ไปสู่ Perceptron ในชั้นถัดไป จะต้องผ่านฟังก์ชันกระตุ้น ซึ่งในหัวข้อก่อนหน้านี้ แทนสัญลักษณ์ด้วย $\mathcal{G}$ ซึ่งมีลักษณะเป็นฟังชันแบบไม่เชิงเส้น (Non-Linear Function) ซึ่งรูปแบบของฟังก์ชันกระตุ้นที่งานวิจัยนี้คาดว่าจะนำมาใช้งานมีดังต่อไปนี้

2.4.1.2.1 ฟังก์ชันซิกมอยด์ (Sigmoid Function)

ซึ่งจะทำหน้าที่เปลี่ยนค่าที่นำเข้าฟังก์ชันให้อยู่ในช่วง 0-1 แสดงการคำนวณตาม สมการที่ (7)

$$\sigma(z) = \frac{1}{1+\exp(-z)} \qquad (7)$$

2.4.1.2.2 ฟังก์ชันแทนเจนต์ไฮเพอร์โบลิก (Hyperbolic Tangent Function)

ทำหน้าที่เปลี่ยนค่าที่เข้ามาในฟังก์ชันให้อยู่ในช่วง -1 ถึง 1 แสดงการคำนวณตาม

สมการที่ (8)

$$tanh(z) = \frac{exp(z)-exp(-z)}{exp(z)+exp(-z)} \quad (8)$$

2.4.1.2.3 ฟังก์ชันเรกติไฟด์เชิงเส้น (Rectified Linear Unit Function) ทำหน้าที่เปลี่ยนค่าที่เข้ามาในฟังก์ชันหากติดลบจะให้ค่าเป็น 0 และหากมากกว่า 0 ให้ค่าที่ได้ออกมาเป็นค่าเดิม โดยมีเงื่อนไขดังนี้

$$f(z) = \begin{cases} 0, if\ z < 0 \\ z, if\ z \geq 0 \end{cases}$$

2.4.1.2.4 ฟังก์ชันค่าสูงสุดแบบอ่อน (Softmax function)

ทำหน้าที่เปลี่ยนค่าที่เข้ามาให้ผลลัพธ์อยู่ในช่วง 0 ถึง 1 ซึ่งเป็นค่าความน่าจะเป็นของ ค่าที่นำเข้ามาแต่ละตัวโดยผลรวมของค่าความน่าจะเป็นที่ได้จะมีค่าเป็น 1 ซึ่งในงานทางด้านการจำแนก (Classification) นิยมใช้ฟังก์ชันนี้อย่างมาก (Hammerton, 2003)

2.4.1.3 ฟังก์ชันต้นทุน (Cost function)

ในการเรียนรู้ของโครงข่ายประสาทเทียมนั้น จำเป็นต้องมีฟังก์ชันที่สามารถใช้วัดผลการเรียนรู้หรือ ฟังก์ชันต้นทุน โดยเป้าหมายของการเรียนรู้จะกำหนดให้เป็นการพยายามลดค่าที่ได้จากฟังก์ชันต้นทุนให้เข้าใกล้ 0 มากที่สุด โดยตัวอย่างฟังก์ชันต้นทุนที่นิยมนำมาใช้งาน เช่น ค่าเฉลี่ยความผิดพลาดกำลังสอง (Mean Square Error : MSE) เอนโทปีไขว้ (Cross-Entropy Loss) หรือค่าติดลบลอการิทึมของความเป็นไปได้ (Negative Log Likelihood) (Tretasayuth, 2017)

2.4.1.4 ฟังก์ชันการหาค่าเหมาะสม (Optimization function)

การหาค่าเหมาะสมที่สุดเป็นวิธีการปรับปรุงอัตราการเรียนรู้ เพื่อให้สามารถลดค่าของฟังก์ชันต้นทุนได้มากที่สุดในแต่ละรอบ เพื่อเพิ่มโอกาสในการลดค่าให้ไปถึงจุดต่ำสุดทั้งหมด (Global Minima) โดยฟังก์ชันที่นิยมนำใช้งานหลายตัว เช่น สโตแคสติกเกรเดียนเดสเซนท์ (Stochastic Gradient Descent : SGD) วิธีโมเมนตัม เป็นต้น (Tretasayuth, 2017)

2.4.1.5 การแพร่กระจายย้อนกลับ (Back propagation)

หลังจากข้อมูลนำเข้าถูกป้อนเข้าสู่โครงข่ายประสาทเทียม และได้ผ่านฟังก์ชันกระตุ้น รวมทั้งคำนวณความผิดพลาดที่เกิดขึ้นจากการคำนวณด้วยฟังก์ชันต้นทุน ซึ่งหากต้องการหาค่าความผิดผลาดที่เกิดขึ้นของ Perceptron ในชั้นโครงข่ายก่อนหน้าไม่สามารถหาค่าโดยตรงได้

วิธีการแพร่กระจายย้อนกลับจึงเป็นวิธีที่มีวัตถุประสงค์เพื่อปรับค่าน้ำหนักก่อนหน้า โดยมีความคาดหวังว่าค่าน้ำหนักในรอบถัดไปเมื่อส่งกลับเข้ามาในโครงข่ายอีกครั้งจะส่งผลลัพธ์ที่ใกล้เคียงกับค่าที่คาดหวังไว้มากขึ้นจนฟังก์ชันต้นทุนลดลงอยู่ในเกณฑ์ที่ยอมรับได้ แสดงตามสมการที่ (9)

$$\frac{\partial Error_{total}}{\partial w_j^l} = \frac{\partial Error_{total}}{\partial out_j^{l+1}} \times \frac{\partial out_j^{l+1}}{\partial net^l} \times \frac{\partial net^l}{\partial w_j^l} \qquad (9)$$

โดยที่

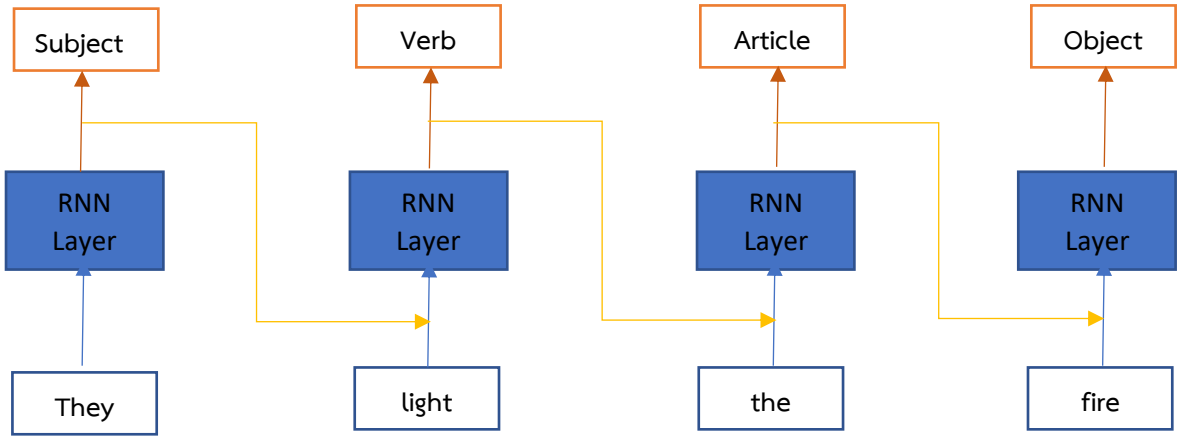
$\frac{\partial Error_{total}}{\partial w_j^l}$ คือ อัตราการเปลี่ยนแปลงของค่าความคลาดเคลื่อนเทียบกับ ค่าน้ำหนัก j ในชั้นที่ l

out_j^{l+1} คือ ค่าผลลัพธ์ในชั้นถัดไป

net^l คือ ผลรวมของโครงข่าย

2.4.2 โครงข่ายประสาทเทียมแบบวกกลับ (Recurrent Neural Network)

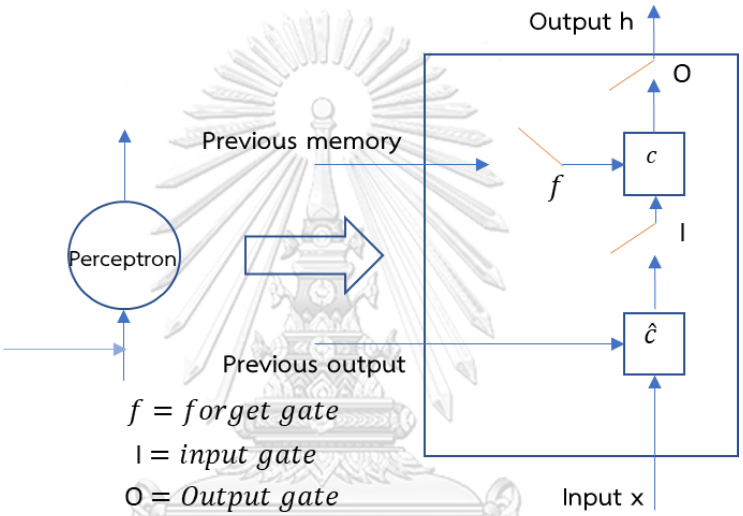
โครงข่ายประสาทเทียมแบบวกกลับ หรือ RNN ถูกออกแบบมาให้ตอบสนองต่อการประมวลผลข้อมูลที่มีลำดับ (Sequential) ซึ่งจะมีการส่งผ่านผลการประมวลผลจากข้อมูลในช่วงเวลาก่อนหน้าไปยังช่วงเวลาถัดไปดังตัวอย่างในรูปที่ 2.4 ทั้งนี้โครงสร้างของ RNN จะคล้ายคลึงกับโครงข่ายประสาทเทียมทั่วไป ส่วนที่แตกต่างคือ มีการส่งต่อชั้นซ่อน ไปเป็นข้อมูลนำเข้าของ Perceptron ในช่วงเวลาถัดไป โครงข่ายประสาทเทียมประเภทนี้มีการนำไปใช้ในงานที่หลากหลาย เช่น การแปลข้อความ การแบ่งประเภทของข้อความ รวมถึงการทำสรุปบทความ (Tretasayuth, 2017) แสดงตัวอย่างภาพรวมของ RNN ตามรูปที่ 5 โดยมีข้อมูลนำเข้าเป็น “คำ” จากประโยค “They light the fire” และผลลัพธ์เป็น “หน้าที่ของคำ”



รูปที่ 5 แสดงภาพรวมตัวอย่างการทำงานของ RNN

2.4.3 โครงข่ายประสาทเทียมแบบแบบหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory : LSTM)

โครงข่ายประสาทเทียมชนิดนี้พัฒนาขึ้นมาเพื่อแก้ปัญหาบางอย่างจาก RNN แบบเดิม โดยเพิ่มประตูสัญญาณว่าเมื่อไรข้อมูลที่ผ่านเข้ามาควรจะส่งผ่านข้อมูลที่ได้รับไปยังหน่วยความจำอื่น หรือ รับข้อมูลมาแต่เก็บไว้ใช้เฉพาะในหน่วยความจำของตัวเอง หรือสุดท้ายคือลืมหรือลบข้อมูลนั้นไป (forget, delete) ทั้งนี้เพื่อเป็นการทำให้แบบจำลองสามารถเข้าใจได้โดยอัตโนมัติเองว่าข้อมูลไหนเป็นข้อมูลที่สำคัญ (Hochreiter & Schmidhuber, 1997) แสดงตัวอย่างตามรูปที่ 6



รูปที่ 6 แสดงภาพรวมการทำงานภายใน Perceptron ของ LSTM

จากรูปที่ 6 แสดงการคำนวณในแต่ละ Perceptron ได้ตามสมการที่ (10) ดังนี้

$$h_t^j = o_t^j \tanh\left(c_t^j\right) \tag{10}$$

จากสมการที่ (10) จะพบว่ามีพารามิเตอร์ที่ไม่ทราบค่า คือ o_t^j และ c_t^j ซึ่งสามารถคำนวณหาค่าได้จากสมการที่ (11) และ (12) ตามลำดับ

$$o_t^j = F^j(W_o x_t + U_o h_{t-1} + V_o c_t) \tag{11}$$

$$c_t^j = f_t^j c_{t-1}^j + i_t^j \hat{c}_t^j \tag{12}$$

จากสมการที่ (12) มีพารามิเตอร์อีก 1 ตัวที่ต้องการจะหาค่าคือ f_t^j , i_t^j และ $\hat{c}_t^j$ ซึ่งคำนวณได้จากสมการที่ (13) (14) และ (15) ตามลำดับ

$$f_t^j = F^j\left(W_f x_t + U_f h_{t-1} + V_j c_{t-1}\right) \quad (13)$$

$$i_t^j = F^j(W_i x_t + U_i h_{t-1} + V_o c_{t-1}) \quad (14)$$

$$\hat{c}_t^j = tanh^j(W_c x_t + U_c h_{t-1}) \quad (15)$$

โครงสร้างของ LSTM ประกอบไปด้วยประตูสัญญาณที่สำคัญ 3 ตัวได้แก่

2.4.3.1 Forget gate

เป็นประตูสัญญาณที่ทำหน้าที่ตัดสินใจว่าจะลบหรือไม่ลบข้อมูลที่มาจาก cell state (c) ก่อนหน้า ในการสร้าง forget gate จะดูจาก input และ hidden state ก่อนหน้า โดยใช้ฟังกชัน sigmoid เป็นเครื่องมือตัดสินใจ

2.4.3.2 Input gate

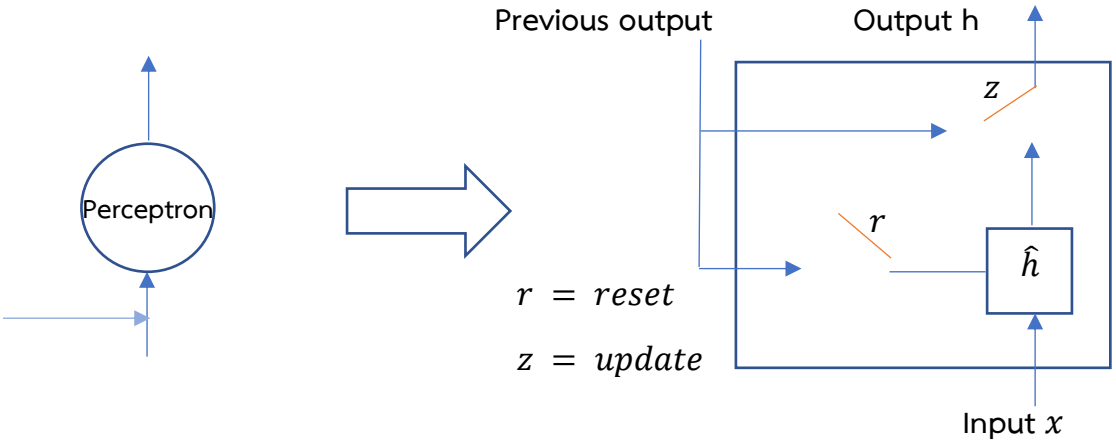
เป็นประตูสัญญาณที่ทำหน้าที่ในการเลือกปรับปรุงข้อมูลที่รับมาจาก cell state ก่อนหน้า โดยใช้ฟังก์ชัน tanh เป็นฟังก์ชันในการตัดสินใจเลือกปรับปรุงข้อมูล

2.4.3.3 Output gate

เป็นประตูสัญญาณที่เลือกว่าจะส่งข้อมูลจาก cell state ออกไปยังหน่วยความจำอื่นหรือไม่ เพื่อให้หน่วยความจำอื่นสามารถนำข้อมูลนี้ไปใช้ต่อ

2.4.4 โครงข่ายประสาทเทียมแบบประตูสัญญาณวกกลับ (Gated Recurrent Unit : GRU)

โครงข่ายประสาทเทียมแบบประตูสัญญาณวกกลับ เรียกต่อไปเป็นตัวย่อว่า GRU โครงข่ายประสาทเทียมนี้นำเอาความทรงจำระยะสั้นแบบยาว (Long Short Term Memory : LSTM) มาทำการปรับปรุงให้ลดจำนวนประตูสัญญาณลง โดยการใช้ประตูสัญญาณอัพเดท (Update gate) มาใช้แทนประตูสัญญาณการจำและประตูสัญญาณผลลัพธ์ อีกทั้งยังทำการรวมหน่วยความจำเข้ากับชั้นซ่อน ทำให้โครงข่ายมีความรวดเร็วมากขึ้นและใช้หน่วยความจำน้อยลง (Cho et al., 2014) แสดงการทำงานของ Perceptron ใน GRU ได้ตามรูปที่ 7



รูปที่ 7 แสดงภาพรวมการทำงานภายใน Perceptron ของ GRU

จากรูปที่ 7 แสดงการคำนวณในแต่ละ Perceptron ได้ตามสมการที่ (10) ดังนี้

$$h_t^j = \left(1 - z_t^j\right)h_t^j + z_{t-1}^j \hat{h}_t^j \qquad (16)$$

โดยที่

j คือ ดัชนีของ Perceptron

t คือ ดัชนีของเวลา

จากสมการที่ (16) จะพบว่ามีพารามิเตอร์ที่ไม่ทราบค่า คือ z_t^j และ $\hat{h}_t^j$ ซึ่งสามารถคำนวณหาค่า $\hat{h}_t^j$ ได้จากสมการที่ (17) และ (18) ตามลำดับ

$$\hat{h}_t^j = tanh^j\left(Wx_t + U(r_t \odot h_{t-1})\right) \qquad (17)$$

$$z_t^j = sigmoid^j(W_z x_t + U_z h_{t-1}) \qquad (18)$$

จากสมการที่ (17) มีพารามิเตอร์อีก 1 ตัวที่ต้องการจะหาค่าคือ reset gate หรือ r_t^j ซึ่งคำนวณได้จากสมการที่ (19)

$$r_t^j = sigmoid^j(W_r x_t + U_r h_{t-1}) \qquad (19)$$

โดยโครงข่ายประสาทเทียมแบบชนิด RNN จะแตกต่างจาก Feed Forward Neural Network ในส่วนของการคำนวณค่าภายใน Perceptron ส่วนกระบวนการอื่นนั้นทำงานเช่นเดียวกัน

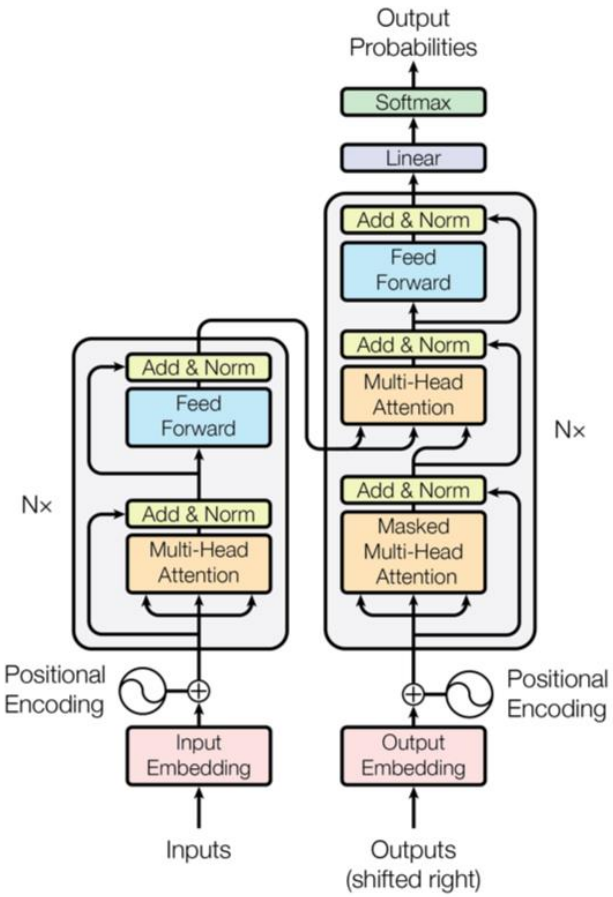
2.5 Transformer Model

เนื่องจากภาษาเป็นข้อมูลแบบลำดับที่มีความละเอียดอ่อนและซับซ้อน ทำให้โครงข่ายประสาทเทียมแบบวกกลับ (RNN) ที่แม้จะเป็นแบบจำลองที่ดีแต่ก็ยังมีข้อจำกัดบางประการอยู่ ได้แก่ เรื่องของลำดับและบริบททางภาษา เช่น “เมื่อปีก่อนขับรถยนต์ของพ่อไปชนต้นไม้ ปัจจุบันรถคันนี้ก็ยังใช้งานอยู่” จากตัวอย่างจะสังเกตเห็นว่า “รถยนต์” กับ “รถ” ในประโยคนี้หมายถึงรถคันเดียวกัน แต่ RNN เอง เมื่อเป็นประโยคยาวๆมีการส่งค่าผ่านมาหลายทอดอาจทำให้สื่อความหมายบางอย่างได้ผิดพลาดไป โดยอาจมองว่า “รถยนต์” และ “รถ” ในประโยคดังกล่าวไม่มีความเกี่ยวข้องกัน จากปัญหาดังกล่าวจึงมีกลุ่มนักวิจัยคิดค้นและนำเสนอสถาปัตยกรรม Transformer ขึ้นมา (Vaswani et al., 2017) ซึ่งกลไกของ Transformer จะมีกลไกหลักเพิ่มขึ้นมาคือ Attention และ Self-Attention โดยจะทำหน้าที่เก็บบันทึกว่าส่วนใดของประโยคที่เป็นใจความสำคัญ คล้ายกับที่คนเราอ่านหนังสือแล้วจดจำใจความสำคัญทำให้สามารถเข้าใจความหมายและบริบทของคำที่ใช้ได้อย่างดีและรวดเร็ว

Transformer เป็นแบบจำลองที่ถูกฝึกฝนมาเพื่อให้เป็นแบบจำลองบริบททางภาษา (language model) เช่น การทำนายคำถัดไป หรือการทำนายคำที่เว้นว่างไว้ในประโยค (masked language modeling) ผ่านคลังข้อมูลทางภาษาขนาดใหญ่จากการเรียนรู้ด้วยตนเอง (self-supervised) โดยที่ไม่ต้องอาศัยการทำฉลาก (label) จากมนุษย์ แบบจำลองประเภทนี้จึงมักถูกนำไปใช้ในกระบวนการถ่ายโอนความรู้ (transfer learning) ซึ่งต้องนำไปฝึกฝนเพิ่มเติม (fine-tune) ในงานอื่นที่มีการทำเฉลยจากมนุษย์ (Vaswani et al., 2017)

2.5.1 สถาปัตยกรรมของแบบจำลอง Transformer

แบบจำลอง Transformer มีองค์ประกอบหลักที่สำคัญ 2 ส่วน คือ ส่วนการเข้ารหัส (encoder) และส่วนการถอดรหัส (decoder) ซึ่งทั้งสองส่วนอาศัยการเรียงซ้อนกันของชั้น Self-attention และ ชั้น Fully-connected แสดงตัวอย่างตามรูปที่ 8



รูปที่ 8 แสดงตัวอย่างสถาปัตยกรรมของแบบจำลอง Transformer

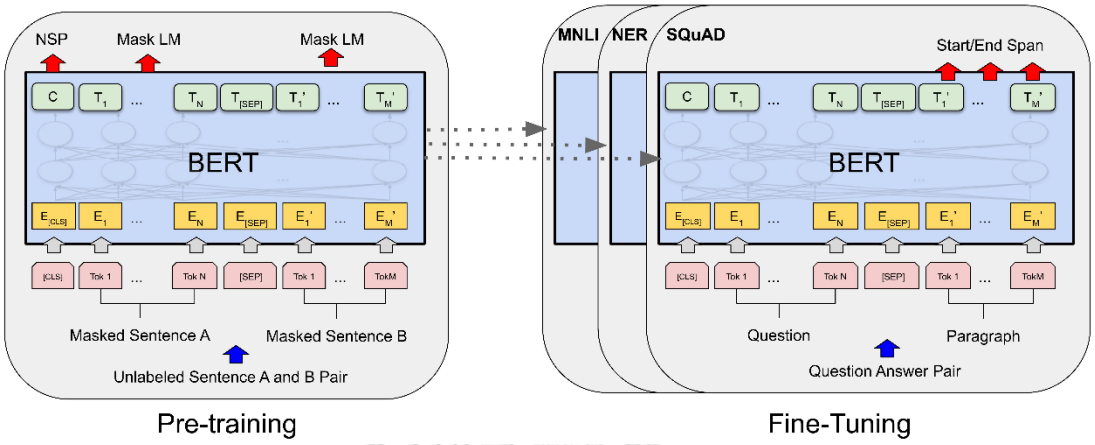
ที่มา: Attention is all you need (Vaswani et al., 2017)

จากรูปที่ 8 ด้านซ้ายมือคือส่วนของ encoder และทางด้านขวามือ คือ decoder ในส่วนประกอบของ encoder จะมี Self-attention และ fully-connected เรียงต่อกัน 6 ชั้น ซึ่งแต่ละชั้นประกอบไปด้วย 2 ชั้นย่อย ชั้นย่อยแรกคือ Multi-head self-attention และชั้นย่อยที่ 2 คือ Position-wise fully-connected feed forward และในส่วนของ decoder มีส่วนประกอบคล้าย decoder มีข้อแตกต่างที่มีการเพิ่มชั้นย่อยที่ 3 ซึ่งเป็นชั้น Multi-head attention ในตัว decoder เพื่อป้องกันไม่ให้เข้าถึงข้อมูลการทำนายผล เพื่อให้แน่ใจว่าแบบจำลองจะทำนายตำแหน่งที่ i โดยตำแหน่งที่ i-1 หรือน้อยกว่า i เท่านั้น (Sae-Lim, 2021)

2.5.2 แบบจำลอง Bidirectional Encoder Representation from Transformer (BERT)

แบบจำลอง Bidirectional Encoder Representation from Transformer (BERT) เป็นหนึ่งในแบบจำลอง Transformer ที่ใช้เพียงส่วน encoder ซึ่งออกแบบมาเพื่อเป็นแบบจำลองฝึกฝนก่อนหน้า (pretrained model) จากคลังข้อมูลขนาดใหญ่ เป็นการเรียนรู้แบบสองทิศทาง โดย

สามารถนำแบบจำลอง BERT ไปฝึกฝนเพิ่มเติมเพื่อแก้ปัญหา เช่น การรู้จำชื่อเฉพาะ การจัดประเภทของข้อความ ฯลฯ ได้โดยที่ไม่จำเป็นต้องปรับเปลี่ยนสถาปัตยกรรมของแบบจำลองเดิม แสดงตามรูปที่ 9



รูปที่ 9 แสดงการนำ Pretrained จาก BERT มาฝึกฝนเพิ่มเติม

ที่มา: https://paperswithcode.com/method/bert

เพื่อให้แบบจำลอง BERT สามารถจัดการงานได้หลากหลาย ข้อมูลนำเข้าที่ใช้สำหรับแบบจำลอง BERT จึงถูกออกแบบให้สามารถแสดงรูปประโยคทั้งแบบเดี่ยวและแบบคู่ นอกจากนี้ยังมีการใช้ Word piece embedding โดยมีจำนวนชิ้นส่วนคำ (Token) จำนวน 30,000 คำ และประกอบไปด้วยชิ้นส่วนคำพิเศษ คือ [CLS] ทำหน้าที่เป็น Token แรกของประโยคแรก และ [SEP] ซึ่งเป็น Token ที่ทำหน้าที่แบ่งแยกระหว่างประโยคที่หนึ่งและประโยคที่สอง (Sae-Lim, 2021)

2.6 การประเมินประสิทธิภาพของแบบจำลอง

ในการประเมินความสามารถของแบบจำลองที่สร้างขึ้น ใช้การคำนวณแม่นยำ (Evaluation metrics) โดยวัดจากค่าเอฟวัน (F1 score) ซึ่งตัวชี้วัดที่ใช้ในการคำนวณพารามิเตอร์มีดังต่อไปนี้

TP คือ จำนวนของคำที่แบบจำลองให้คำตอบตรงเฉลย

FP คือ จำนวนของคำที่แบบจำลองให้คำตอบเกินกว่าเฉลย

FN คือ จำนวนของคำที่แบบจำลองไม่ให้คำตอบแต่พบในเฉลย

ตัววัดประสิทธิภาพ (F1 Measurement)

การคำนวณหาค่าความเที่ยงตรง (Precision) ค่าความระลึก (Recall) และค่าเอฟวัน (F1) นั้น สามารถคำนวณได้จาก สมการที่ (14) – (16)

$$Precision = \frac{TP}{TP+FP} \quad (14)$$

$$Recall = \frac{TP}{TP+FN} \quad (15)$$

$$F_1 = \frac{2*Pecision*Recall}{Precision+Recall} \quad (16)$$

ในงานวิจัยนี้จะใช้การประเมินความถูกต้องโดยรวมในระดับคำ (phrase-level) เป็นหลัก เนื่องจากเป็นการประเมินผลของคำที่ถูกต้องทั้งหมด ไม่ได้นับเอาเฉพาะส่วนใดส่วนหนึ่งของคำที่ถูกต้องเพียงอย่างเดียว สำหรับการประเมินประสิทธิภาพด้วย F1-Token เพียงอย่างเดียวยังไม่ถูกต้องครบถ้วนนักเนื่องจากการสร้างแบบจำลองเราต้องการที่จะสกัดชื่อภูมิศาสตร์ทั้งชื่อออกมาจากข้อความไม่ใช่เพียงแค่ส่วนใดส่วนหนึ่งของชื่อ ตัวอย่างเช่น การหาค่า F1-Token ประโยคว่า “วิวแม่น้ำโขงนครพนมสวยสุดละ นี่ชอบมาก” แสดงตัวอย่างตามรูปที่ 10

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 10 ตัวอย่างการคำนวณค่าความถูกต้องโดยรวม (F1)

จากรูปที่ 10 และสมการที่ (14) – (16) คำนวณ F1-Token ได้ดังนี้ TP = 2, FP = 1, FN = 1 Precision = 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 10 สำหรับคำว่าแม่น้ำโขง แบบจำลองให้คำตอบ token สุดท้ายผิดจึงถือว่าแบบจำลองให้

คำตอบที่ผิดในกรณีนี้ จากตัวอย่างค่าที่คำนวน F1-Phrase ได้จาก TP = 1, FP = 1, FN = 1, Precision = 1/(1+1), Recall = 1/(1+1), F1 = 2*[(0.5*0.5)/(0.5+0.5)] = 0.25 จากตัวอย่างข้างต้นจะพบว่าควรพิจารณาที่ระดับ F1-Phrase มากกว่า เนื่องจากผลลัพธ์สุดท้ายของแบบจำลองที่ต้องการคือ ชื่อภูมิศาสตร์ที่ครบสมบูรณ์

2.6 กระบวนการเข้ารหัสทางภูมิศาสตร์ (Geocoding)

สำหรับกระบวนการเข้ารหัสทางภูมิศาสตร์เป็นขั้นตอนในการเชื่อมโยงภูมินามไปสู่พิกัดตำแหน่งในทางภูมิศาสตร์ โดยหากภูมินามที่สกัดออกมาไม่สามารถเข้ารหัสทางภูมิศาสตร์ได้อาจเป็นไปได้ว่า ภูมินามที่สกัดออกมาไม่ถูกต้อง เช่น “เดอะมอลล์บางแคถึงเดอะมอลล์งามวงศ์วาน” หากเครื่องมือในการรู้จำภูมินามสกัดข้อมูลออกมาเป็น “เดอะมอลล์บางแค” และ “เดอะมอลล์งามวงศ์วาน” อาจจะสามารถเข้ารหัสภูมิศาสตร์ได้ แต่หากเครื่องมือสกัดภูมินามจากวลีข้างต้นออกมาเป็นคำเดียวกันจะทำให้การเข้ารหัสภูมิศาสตร์มีความผิดพลาด หรือในอีกกรณีหนึ่งคือภูมินามที่สกัดมาได้นั้นอาจเป็นสถานที่ใหม่หรือเป็นชื่อสถานที่ที่ใช้กันในท้องถิ่นทำให้ไม่สามารถจับคู่กับข้อมูลที่มีอยู่ในฐานข้อมูลได้

การเข้ารหัสภูมิศาสตร์ (Geocoding) คือกระบวนการที่มีวัตถุประสงค์เพื่อระบุพิกัดตำแหน่งทางภูมิศาสตร์ (เช่น ละติจูด ลองจิจูด) จากชื่อสถานที่หรือข้อมูลที่อยู่ (Address) โดยประกอบไปด้วยส่วนสำคัญ 2 ส่วนคือ 1) อัลกอริทึมในการค้นหาหรือจับคู่ข้อมูล และ2) ฐานข้อมูลที่ใช้อ้างอิง (Wilson & Knoblock, 2007; Zandbergen, 2008) กระบวนการเข้ารหัสทางภูมิศาสตร์อาศัยอัลกอริทึมในการจับคู่เป็นหลัก (Matching algorithm) ซึ่งการตรวจสอบตำแหน่งนั้นข้อมูลนำเข้าจะต้องอยู่ในขอบเขตของข้อมูลอ้างอิงจึงจะคืนค่ามาเป็นพิกัดภูมิศาสตร์ สำหรับกระบวนการที่เกี่ยวข้องกับการเข้ารหัสภูมิศาสตร์จากที่อยู่หรือชื่อสถานที่เรียกว่า Address geocoding (Cetl et al., 2018)

2.6.1 การเข้ารหัสทางภูมิศาสตร์โดยใช้ข้อมูลที่อยู่ (Address) ร่วมกับชื่อถนนซึ่งอ้างอิงกับฐานข้อมูลโครงข่ายถนน

เทคนิคที่ได้รับความนิยมและใช้กันอยู่อย่างแพร่หลายคือ ข้อมูล TIGER (Topological Integrated Geographic Encoding and Referencing) ซึ่งหน่วยงานที่เป็นเจ้าของข้อมูลคือสำนักงานสถิติแห่งชาติสหรัฐฯ (U.S. Census Bureau) โดยการทำงานของอัลกอริทึมคือ การประมาณค่าในช่วงแบบเชิงเส้น (Linear interpolation) เพื่อตรวจสอบข้อมูลที่อยู่นั้นอยู่ในช่วงของเลขถนนและจุดตัดของถนนเส้นใด โดยแบ่งลำดับขั้นในการประมวลผลดังนี้

2.6.1.1 จับคู่ชื่อถนนจากข้อมูลที่อยู่ เป็นการจับคู่ระหว่างชื่อถนนในข้อมูลที่อยู่กับชื่อถนนที่เตรียมไว้ในชุดข้อมูลอ้างอิง (Reference dataset)

2.6.1.2 ระบุฝั่งของถนน ระบุฝั่งถนนโดยนำข้อมูลจากเลขที่ซอย เลขที่บ้าน ว่าเป็นเลขคู่หรือเลขคี่

2.6.1.3 ค่าพิกัดสุดท้ายจะได้จากการคำนวณโดยวิธีการประมาณค่าในช่วงแบบเชิงเส้น ไปตัดกับค่า offset ของถนนในระบบภูมิสารสนเทศจากเส้นกึ่งกลางถนนทำให้ระบุได้ว่าตำแหน่งของที่อยู่ตกในช่วงถนนใด (street segment) แสดงตัวอย่างตามรูปที่ 11 (Owusu et al., 2017)

รูปที่ 11 แสดงอัลกอริทึมในการเข้ารหัสภูมิศาสตร์โดยอ้างอิงฐานข้อมูลโครงข่ายถนน

จากรูปที่ 11 สัญลักษณ์ V คือตำแหน่งที่ได้จากการประมาณค่าในช่วงแบบเชิงเส้นซึ่งตัดกับ D ซึ่งเป็นค่า offset จากเส้นกลางถนน

2.7 บริการการเข้ารหัสทางภูมิศาสตร์แบบออนไลน์ (Online Geocoding)

จากหัวข้อที่ 2.5 การเข้ารหัสทางภูมิศาสตร์ฐานข้อมูลอ้างอิงผู้ใช้งานจะต้องเตรียมข้อมูลอ้างอิงเอง ซึ่งเรียกอีกอย่างว่า Conventional Geocoding โดยในปัจจุบันมีผู้ให้บริการสำหรับขอข้อมูลค่าพิกัดจากผู้ให้บริการออนไลน์หลากหลายผ่านระบบ Application Programing Interface (API) เช่น Google Geocoding API , Nominatim จาก Open Street Map (OSM) จากงานวิจัยของ Manoruang and Asavasuthirakul (2019a) ทดสอบเปรียบเทียบประสิทธิภาพของผู้ให้บริการออนไลน์ในการเข้ารหัสภูมิศาสตร์ของข้อมูลที่เป็นตัวอักษรภาษาไทยโดยมีการวัดประสิทธิภาพใน 2 ส่วนหลัก คือ อัตราการจับคู่ระหว่างชื่อสถานที่ (Match rate) ซึ่งแสดงเป็นอัตราส่วนระหว่างชื่อสถานที่ที่ส่งไปกับค่าพิกัดที่ระบบส่งคืนกลับมาให้ และความถูกต้องเชิงตำแหน่ง (Positional accuracy) พบว่าจากการเปรียบเทียบประสิทธิภาพระหว่างผู้ให้บริการ 5 ราย ได้แก่ Google Geocoding , Mapquest, Bing, Yahoo และ Opencage ข้อมูลที่ใช้ในการทดลองคือ กลุ่มตัวอย่างข้อมูลที่อยู่ (Address) และข้อมูลชื่อสถานที่ (ในงานวิจัยนี้เรียกว่า POIs names) จำนวน 200 แห่ง โดยผลการทดลอง Google Geocoding ให้ประสิทธิภาพในแง่ของ Match rate สูงกว่าทั้ง 4 รายที่เหลือ โดยข้อมูลการใช้งานเบื้องต้นของทั้ง 5 API แสดงตามตารางที่ 1

ตารางที่ 1 แสดงรายละเอียดการเข้าใช้งานฟรีของผู้ให้บริการเข้ารหัสทางภูมิศาสตร์ออนไลน์

ชื่อของเซอร์วิส	บริษัทที่ให้บริการ	ข้อจำกัดในการเข้าใช้งานฟรี
Google Geocoding API	Google Inc.	2,500 requests/วัน
MapQuest Maps	MapQuest Inc.	15,000 transactions/เดือน
Bing Maps	Microsoft Corporation	50 jobs/วัน
Yahoo! Place Finder	Yahoo! Inc.	5,000 queries/IP/วัน
OpenCage Geocoder	OpenCage Data Ltd.	2,500 queries/วัน

2.8 การประมาณตำแหน่งทางภูมิศาสตร์ในกรณีที่ไม่มีข้อมูลสถานที่ในฐานข้อมูลอ้างอิง

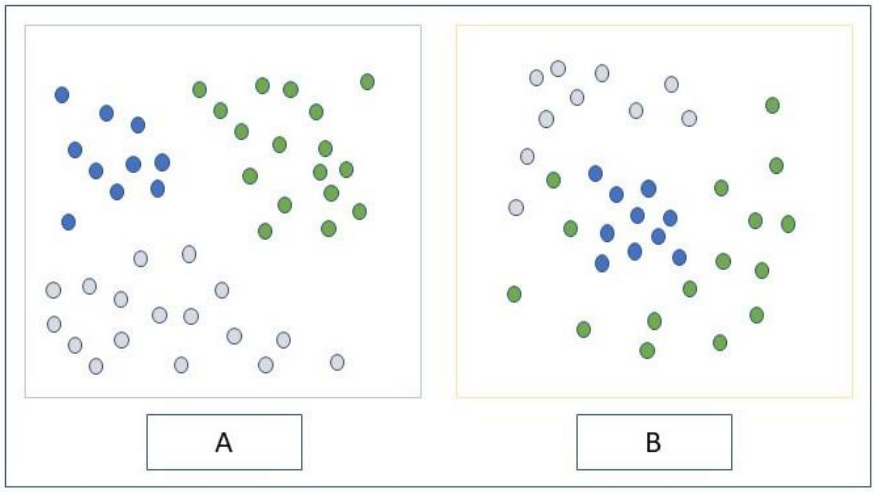
จากหัวข้อที่ 2.6 - 2.7 การเข้ารหัสทางภูมิศาสตร์จะใช้การจับคู่ระหว่างข้อมูลนำเข้ากับข้อมูลค่าพิกัดของสถานที่ที่เก็บไว้ในฐานข้อมูลหรือฐานข้อมูลของผู้ให้บริการออนไลน์ แต่หากข้อมูลนำเข้ายังไม่ถูกบันทึกไว้จะไม่สามารถให้ค่าพิกัดทางภูมิศาสตร์กลับมาได้ดังนั้นเพื่อแก้ปัญหาข้างต้น ในงานวิจัยนี้จึงนำอัลกอริทึมในการจัดกลุ่ม (Clustering) มาใช้เพื่อประมาณขอบเขตของสถานที่ที่ได้มาจากขั้นตอนของการรู้จำภูมินาม โดยอาศัยตำแหน่งของภูมินามอื่นๆที่เกี่ยวข้องจากข้อความของแหล่งข้อมูล เช่น ตำแหน่งของผู้ส่งข้อความและตำแหน่งของสถานที่ในข้อความเดียวกัน

2.8.1 การจัดกลุ่ม (Clustering) และอัลกอริทึมที่นิยมใช้งาน

การจัดกลุ่ม (Clustering) เป็นอัลกอริทึมประเภทหนึ่งในสาขาของการเรียนรู้ของเครื่อง (Machine learning) โดยจัดเป็นการเรียนรู้แบบไม่กำกับดูแล (Unsupervised Learning) ซึ่งจุดเด่นของวิธีนี้คือ ไม่ต้องมีการติดฉลาก (Label) ให้กับข้อมูลฝึกสอน (Training Data) ทำให้ผลที่ได้จากอัลกอริทึม คือ การจัดแบ่งข้อมูลออกเป็นกลุ่มตามลักษณะเฉพาะของข้อมูลนั้น เช่น การเกาะกลุ่มกันของข้อมูลเป็นต้น (Omran et al., 2007) โดยอัลกอริทึมที่นิยมใช้มีหลายอัลกอริทึม เช่น K-Means , Fuzzy C-Means, Gaussian Mixture Model : GMM ฯลฯ (Ullman et al., 2014) แต่จากงานวิจัยของ Mai et al. (2017) ให้ผลดีในการจัดกลุ่มค่าพิกัดของสถานที่สำคัญ รวมไปถึง Geotagged ที่มากับข้อความสื่อสังคมออนไลน์ เช่น Flickr, Twitter หรือแม้แต่ Facebook เป็นต้น นอกจากนี้ยังมีงานวิจัยที่ประมาณตำแหน่งของภูมินามในภาษาจีนโดยใช้ข้อมูลจาก Point of Interest (POI) เป็นข้อมูลหลักซึ่งประมวลผลในการประมาณตำแหน่งด้วย DBSCAN (Kuai et al., 2020)

2.8.2 การวิเคราะห์เพื่อประมาณตำแหน่งจากสถานที่ใหม่ด้วยอัลกอริทึม DBSCAN

อัลกอริทึม Density-based clustering algorithm and noise หรือ DBSCAN เป็นหนึ่งในอัลกอริทึมการแบ่งกลุ่มแบบไม่กำกับดูแล (Unsupervised Learning) โดยจุดเด่นของวิธีนี้เมื่อเปรียบเทียบกับอัลกอริทึมในประเภทเดียวกันอย่าง k-means หรือ การจัดกลุ่มแบบลำดับขั้น (Hierarchical Clustering) คือสามารถใช้กับกลุ่มข้อมูลที่มีลักษณะกลุ่มเป็นรูปทรงต่างๆไม่เป็นกลุ่มก้อนชัดเจน รวมทั้งสามารถตัดข้อมูล noise ที่ไม่ได้อยู่ในกลุ่มข้อมูลใดได้ ตัวอย่างของข้อมูลที่เป็นกลุ่มก้อนชัดเจนและไม่ชัดเจนแสดงตามรูปที่ 12

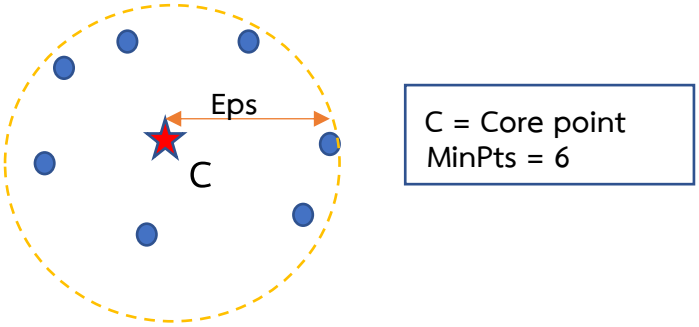


รูปที่ 12 แสดงข้อมูลที่เป็นกลุ่มก้อนชัดเจน (A) และไม่ชัดเจน (B)

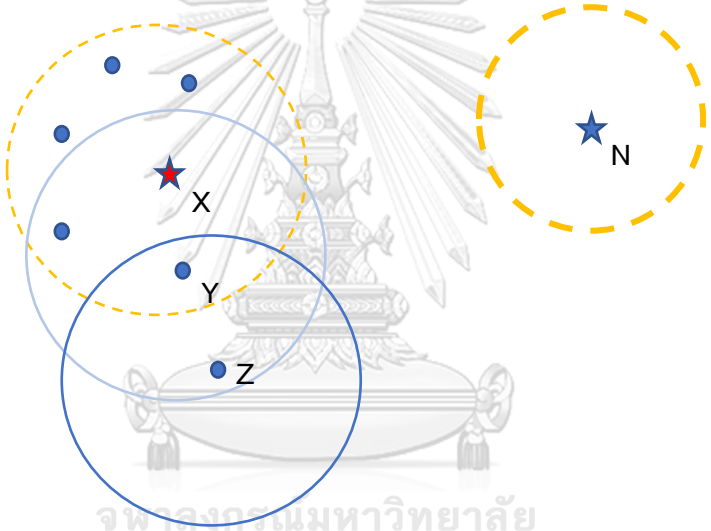
จากรูปที่ 12 หากต้องการแบ่งกลุ่มข้อมูลตามรูป B อัลกอริทึม DBSCAN สามารถแก้ปัญหานี้ได้ซึ่งมีลักษณะการทำงานคือ การหาบริเวณที่ข้อมูลมีการเกาะกลุ่มกันโดยใช้การคำนวณจากจุดข้อมูล (Data point) ที่อยู่โดยรอบ โดยอัลกอริทึม DBSCAN มีพารามิเตอร์ที่สำคัญ 2 ตัวคือ

1) Eps คือรัศมีของข้อมูลศูนย์กลาง (Core point) ไปยังจุดที่อยู่ไกลที่สุดที่เป็นเพื่อนบ้าน (Neighborhood) ของข้อมูลศูนย์กลาง (Neighborhood มีจำนวนกี่จุดกำหนดจาก Minpts) และ

2) Minpts คือจำนวนจุดข้อมูลขั้นต่ำที่จะกำหนดจุดศูนย์กลาง (Kolatch, 2001) โดยการกำหนดค่า Minpts มีอยู่ด้วยกันหลายวิธีแต่ในงานวิจัยนี้เลือกนำวิธีของ Devkota et al. (2019) ซึ่งใช้การคำนวณ Minpts = dimension of data + 1 แต่ต้องไม่น้อยกว่า 3 แสดงตัวอย่างพารามิเตอร์ทั้งสองตัวตามรูปที่ 13



รูปที่ 13 แสดงภาพรวมพารามิเตอร์ของ DBSCAN



รูปที่ 14 แสดงตัวอย่างการทำงานของ DBSCAN

จากรูปที่ 14 สรุปได้ดังนี้ (C. Lowphansirikul, 2018)

1) กำหนด Minpts = 6 จุด

2) ตำแหน่ง X เป็น Core point เนื่องจากมีจุดใกล้เคียงเท่ากันกับ Minpts คือ 6 จุด (รวมตัวเอง)

3) ตำแหน่ง Y เป็น Border เนื่องจาก Z เป็นตำแหน่งที่มีจุดใกล้คียงแต่มีเพียง 4 จุด (รวมตัวเอง) ซึ่งน้อยกว่า MinPts และอยู่ในรัศมีของ Core point คือ X

4) ตำแหน่ง Z เป็นBorder เนื่องจาก Z เป็นตำแหน่งที่มีจุดใกล้เคียงแต่มีเพียง 2 จุด (รวมตัวเอง) ซึ่งน้อยกว่า MinPts และอยู่ในรัศมีของ Y ซึ่ง Y อยู่ในรัศมี Core point คือ X จึงจัดให้ Z อยู่ในกลุ่มเดียวกับ X และ Y

5) ตำแหน่ง N คือ noise เนื่องจากตำแหน่งของ N นั้นไม่ได้อยู่ในรัศมีของ Core point จุดใดเลย ซึ่ง noise ถือเป็นข้อมูลที่ต้องตัดออกไปไม่รวมอยู่ในกลุ่มใด

6) สร้างขอบเขตประมาณตำแหน่งของสถานที่ด้วยวิธี Convex hull และสุดท้าย

7) สร้างพิกัดตัวแทนของตำแหน่งโดยที่ใช้ centroid ของรูปปิดซึ่งได้จาก convex hull

2.8.3 K-means

เป็นเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอน ใช้สำหรับแก้ปัญหาการจัดกลุ่มที่รู้จักกันทั่วไป โดยอัลกอริทึม K-Means จะตัดแบ่ง (Partition) วัตถุออกเป็น K กลุ่ม และแทนค่าแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม ซึ่งใช้เป็นจุดศูนย์กลาง (centroid) ของกลุ่มในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน (Nanda et al., 2023) มีวิธีการโดยสรุปคือ

1) กำหนดหรือสุ่มค่าเริ่มต้น จำนวน k ค่า(กลุ่ม) และกำหนดจุด ศูนย์กลางเริ่มต้น k จุด เรียกว่า cluster centers หรือ(centroid)

2) นำวัตถุทั้งหมดจัดเข้ากลุ่ม โดยทำการหาค่าระยะห่างระหว่างข้อมูล กับจุดศูนย์กลาง หากข้อมูลไหนใกล้ค่าจุดศูนย์กลางตัวไหนที่สุดอยู่กลุ่มนั้น

3) หาค่าเฉลี่ย แต่ละกลุ่ม ให้เป็นค่าจุดศูนย์กลางใหม่

4) ทำซ้ำข้อ 2 จนกระทั่งค่าเฉลี่ยหรือจุดศูนย์กลางในแต่ละกลุ่มจะไม่เปลี่ยนแปลงจึงหยุดทำ

2.8.4 K-medoids

เป็นเทคนิคที่คล้ายกับ K-means แต่กำหนดจุดที่เป็น medoids จากจุดใดจุดหนึ่งในชุดข้อมูลแทนที่จะเป็นการกำหนดจุดขึ้นมาจากศูนย์กลางของกลุ่มข้อมูล (Jin & Han, 2010)

1) กำหนดหรือสุ่มค่าเริ่มต้น จำนวน k ค่า(กลุ่ม) และกำหนดจุดใดจุดหนึ่งของข้อมูลเป็นศูนย์กลางเริ่มต้น k จุด เรียกว่า cluster centers หรือ(centroid)

2) นำวัตถุทั้งหมดจัดเข้ากลุ่ม โดยทำการหาค่าระยะห่างระหว่างข้อมูล กับจุดศูนย์กลาง หากข้อมูลไหนใกล้ค่าจุดศูนย์กลางตัวไหนที่สุดอยู่กลุ่มนั้น

3) หาค่าเฉลี่ย แต่ละกลุ่ม ให้เป็นค่าจุดศูนย์กลางใหม่

4) ทำซ้ำข้อ 2 จนกระทั่งค่าเฉลี่ยหรือจุดศูนย์กลางในแต่ละกลุ่มจะไม่เปลี่ยนแปลงจึงหยุดทำ

2.8.5 Agglomerative clustering

เป็นเทคนิคที่ไม่ต้องมีการกำหนดจำนวนกลุ่มที่ต้องการจัดกลุ่มข้อมูลก่อน เป็นการวิเคราะห์แบบเป็นขั้นตอน ประกอบด้วยขั้นตอนที่สำคัญคือ

1) กำหนดให้ข้อมูลแต่ละข้อมูลคือแต่ละกลุ่ม เช่น มีชื่อภูมิศาสตร์ที่สามารถใช้อ้างอิงตำแหน่งได้ 5 ชื่อ ก็คือสามารถแบ่งได้ 5 กลุ่มในเบื้องต้น

2) ทำการสร้างเมตริกขนาด N x N จากชุดข้อมูล คำนวณระยะห่างของแต่ละจุดข้อมูล

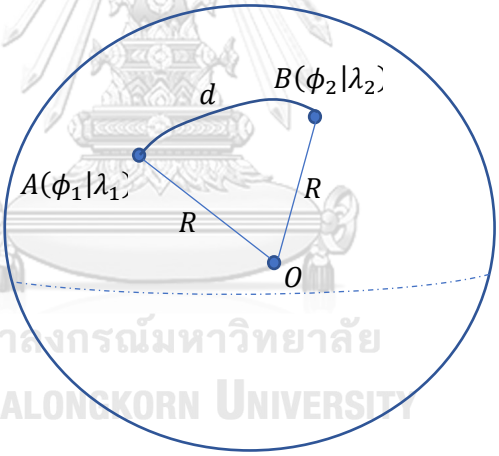
3) หาค่าระยะห่างระหว่างข้อมูลเหล่านั้นทีมีค่าน้อยที่สุดและทำการรวมข้อมูลเป็นกลุ่มเดียวกัน และปรับปรุงเมตริกโดยเลือกค่าที่มากที่สุดของทั้งสองกลุ่มเอาไว้

4) ปรับปรุงตารางข้อมูลเมื่อรวมกลุ่มกันแล้ว ให้เหลือค่าที่มากที่สุดของ 2 กลุ่มเอาไว้ เช่น จุดที่ 1 มีค่า 10 จุดที่ 2 มีค่า 7 ให้เลือกเก็บค่า 10 เอาไว้

5) ทำซ้ำขั้นตอนที่ 3-4 จนกว่าจะเหลือเพียงกลุ่มเดียว ซึ่งผลลัพธ์จะแสดงให้เห็นความสัมพันธ์ของแต่ละข้อมูล ซึ่งเป็นอีกอัลกอริทึมหนึ่งที่เหมาะกับการนำมาเป็นตัวช่วยในเรื่องการประมาณตำแหน่งของสถานที่ (Subasi, 2020)

2.8.6 การคำนวณระยะทางด้วยสมการ Haversine

การคำนวณระยะทางสำหรับข้อมูลที่เป็นค่าพิกัดภูมิศาสตร์ เนื่องจากพิกัดภูมิศาสตร์เป็นพิกัดเชิงขั้วดังนั้นในการคำนวณจึงจำเป็นต้องคำนึงถึงความโค้งของผิวโลกด้วย สมการที่เป็นเครื่องมือในการคำนวณหาระยะทางที่สั้นที่สุดสำหรับพื้นผิวบนรูปทรงกลม คือ สมการ Haversine (Jiangping Wang, 2009) โดยตัวอย่างการคำนวณระยะทางระหว่างจุด 2 จุดของสมการ Haversine แสดงตามรูปที่ 15



รูปที่ 15 แสดงการหาระยะทางจาก A ไป B บนพื้นผิวทรงกลม

จากรูปที่ 15 ค่าพิกัดระหว่างจุด A และจุด B คือ $A\ (\phi_1, \lambda_1)$ และ $B(\phi_2, \lambda_2)$ ตามลำดับ โดยที่

R	คือ รัศมีของโลก (6,371 กิโลเมตร)
d	คือ ระยะทางระหว่างจุด 2 จุด
ϕ_1, ϕ_2	คือ ละติจูดของจุดที่ 1 และละติจูดของจุดที่ 2
λ_1, λ_2	คือ ลองจิจูดของจุดที่ 1 และลองจิจูดของจุดที่ 2

$$haversine\left(\frac{d}{R}\right) = haversine(\triangle \phi) + cos\phi_1 cos\phi_2 haversine(\triangle \lambda) \quad (17)$$

โดยที่

$$haversine(\theta) = sin^2\left(\frac{\theta}{2}\right) \quad (18)$$

จากสมการที่ 17 จัดรูปหาระยะทาง (d) ได้ดังนี้

$$d = 2R \times \arcsin\left(\sqrt{haversine(\triangle \phi) + cos\phi_1 cos\phi_2 haversine(\triangle \lambda)}\right) \quad (19)$$

จากสมการที่ 18 และ 19 เขียนสมการที่ 20 ได้ดังนี้

$$d = 2R \times \arcsin\left(\sqrt{sin^2\left(\frac{\phi_2-\phi_1}{2}\right) + cos\phi_1 cos\phi_2 \sin\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right) \quad (20)$$

โดยสมการที่ 20 คือสมการที่นำไปใช้งานเพื่อหาระยะทางระหว่างจุด 2 จุดในงานวิจัยนี้

2.9 การประเมินประสิทธิภาพการเข้ารหัสภูมิศาสตร์

การประเมินประสิทธิภาพทางภูมิศาสตร์ส่วนใหญ่จะใช้การประเมินความถูกต้องเชิงตำแหน่ง เช่น Mean Square Error (MSE) , Root Mean Square Error (RMSE) แต่ในกรณีที่ไม่สามารถเข้ารหัสทางภูมิศาสตร์จากฐานข้อมูลที่มีอยู่ได้ต้องประมาณตำแหน่งของภูมินามขึ้นมา การหาค่าคลาดเคลื่อนเชิงตำแหน่งเพียงอย่างเดียวจึงอาจจะไม่เพียงพอเนื่องจากผลลัพธ์จากอัลกอริทึม DBSCAN จะให้ขอบเขตของตำแหน่งภูมินามที่ไม่ทราบค่าพิกัดโดยนำภูมินามที่อยู่ข้างเคียงซึ่งมีค่าพิกัดมาช่วยในการประมาณขอบเขต หากพบว่าค่าพิกัดของภูมินามอยู่ในขอบเขตที่ประมาณขึ้น (Point in Polygon) ให้ถือว่าเป็น True แต่หากอยู่นอกขอบเขตให้ถือว่าเป็น False คำนวณหา Precision Recall และ ค่า F1-score ซึ่งในหลายงานวิจัยทางด้านการประมาณค่าพิกัดจากสื่อสังคมออนไลน์คิดค่าความถูกต้องในขอบเขต X กิโลเมตร จากค่าพิกัดที่ทราบค่า โดยในงานวิจัยของ Grover et al. (2010) กำหนดขอบเขตไว้ที่ 5 กิโลเมตร ซึ่งพื้นที่ศึกษาของงานวิจัยคือพื้นที่เกาะอังกฤษและไอร์แลนด์บางส่วน J. Santos et al. (2015) แบ่งค่าความถูกต้องออกเป็นระดับ เช่น เมืองใหญ่ (Cities) ตัวเมือง (Town) ประเทศ (Country) ซึ่งมีค่ามัธยฐาน (Median) ของค่าคลาดเคลื่อนเชิงตำแหน่งอยู่ที่ 77 40 และ 635 กิโลเมตร ตามลำดับ หรืองานวิจัยของ Dredze et

al. (2013) แบ่งการประเมินความถูกต้องออกเป็น 4 ระดับ คือ 25 50 100 และ 250 ไมล์ นอกจากนี้ ในบางงานวิจัยยังมีการกำหนดขอบเขตไว้ที่ 161 กิโลเมตร หรือ 100 ไมล์ ซึ่งเป็นการประเมินความถูกต้องในระดับของเมือง (City level) (Han, 2014; Roller et al., 2012; Zheng et al., 2018)

บทที่ 3
งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องแบ่งออกเป็นหัวข้อสำคัญ 4 หัวข้อดังนี้ 3.1 งานวิจัยทางด้านการรู้จำภูมินาม (Toponym) ในภาษาต่างประเทศ 3.2 งานวิจัยทางด้านการรู้จำชื่อเฉพาะ (NER) ในภาษาไทย 3.3 งานวิจัยทางด้านการสกัดข้อมูลตำแหน่งจากสื่อสังคมออนไลน์ และหัวข้อสุดท้าย 3.4 คือสรุปการทบทวนงานวิจัยซึ่งมีสาระดังนี้

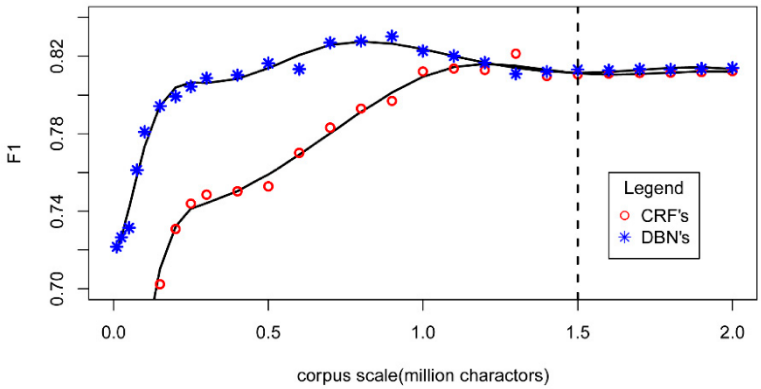
3.1 งานวิจัยทางด้านการรู้จำภูมินาม (Toponym Recognition)

ปัจจุบันมีการศึกษาวิจัยในด้านนี้ค่อนข้างจำกัดโดยหัวข้อที่สำคัญในแนวทางวิจัยนี้คือ แก้ไขปัญหาในเรื่องการรู้จำชื่อเฉพาะที่มีความกำกวม และชื่อที่ไม่ได้ถูกบรรจุอยู่ในอักขรานุกรมทางภูมิศาสตร์

ในงานของ Lieberman and Samet (2011) เสนอวิธีการรู้จำภูมินามด้วยการใช้เครื่องมือ Stanford-NER (Finkel et al., 2005) เพื่อสร้างคุณสมบัติในการหาคำที่ทำหน้าที่เป็น คำนามที่แสดงถึงชื่อเฉพาะ (proper nouns) เนื่องจากสถานที่นั้นมักจะเป็นคำนามที่แสดงถึงชื่อเฉพาะ นอกจากนี้ยังมีการใช้พจนานุกรมเอนทิตี้ (entity dictionary) ซึ่งมีข้อมูลที่อธิบายคุณสมบัติ เช่น ชื่อสถานที่ สี ฤดูกาล เพื่อให้สามารถตรวจสอบได้อีกชั้นหนึ่งว่าชื่อสถานที่ที่สกัดได้จาก Stanford-NER เป็นชื่อสถานที่จริง Sobhana et al. (2010) ใช้การกำหนดคุณสมบัติที่ต่างจาก Lieberman and Samet (2011) ในระบบนี้ใช้การหาหน้าที่ของคำ (POS-tags) ในงานนี้เพิ่มเติมบริบทของคำอุปสรรค (prefix) ซึ่งเป็นคำนำหน้า และ คำปัจจัย (suffix) ซึ่งเป็นคำที่อยู่ส่วนท้ายคำ รวมทั้งคุณสมบัติเฉพาะบางอย่างของคำ โดยใช้แบบจำลอง Conditional Random Fields (CRF) ในงานวิจัยนี้ใช้เทคนิคการรู้จำชื่อเฉพาะเพื่อหาคำทางด้านธรณีวิทยา Chasin et al. (2014) นำเทคนิคทางด้านการเรียนรู้ของเครื่อง 3 วิธีมาศึกษาเปรียบเทียบกันในขั้นตอนของการสกัดเอาชื่อสถานที่จากเอกสาร ได้แก่ 1) Support Vector Machine (SVM) 2) Hidden Markov Model (HMM) และ 3) เครื่องมือประมวลผลภาษาธรรมชาติของ Stanford Group 11 หลังจากนั้นจึงใช้วิธี CRF ร่วมกับ Google Geocoder เป็นตัวทดสอบว่าชื่อที่สกัดออกมาได้เป็นชื่อสถานที่จริงหรือไม่ งานของ Sagcan and Karagoz (2015) เสนอเทคนิคแบบผสม (hybrid) โดยใช้เทคนิคการจำแนกแบบใช้กฎ (Rule Based) ร่วมกับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) สำหรับการเรียนรู้ของเครื่องเลือกใช้แบบจำลอง CRF ในการใช้งาน โดยใช้ข้อมูลที่ไม่เป็นทางการอย่างทวิตเตอร์ และในส่วนของเทคนิคการจำแนกแบบใช้กฎถูกนำมาใช้โดยการกำหนด regular expression เช่น เจอคำว่า “treet” กำหนดให้หากมีตัวอักษรข้างหน้าเป็น [s,S] และ คำข้างหน้ามีหน้าที่ของคำเป็นชื่อเฉพาะ ให้ติดแท็กคำนี้เป็น

“LOCATION” ในการทดลองนี้แบ่งคุณสมบัติที่ใช้ในการศึกษาออกเป็น 9 กลุ่ม โดยกลุ่มที่มีค่าความถูกต้องโดยรวม (F1-score) สูงที่สุดคือ 59%

สังเกตได้ว่างานวิจัยทางด้านการรู้จำภูมินาม (Toponym Recognition) ในช่วงแรกใช้วิธีการสกัดข้อมูลจากวิธีการใช้กฎทางภาษาศาสตร์ (การใช้อักขรานุกรมทางภูมิศาสตร์รวมอยู่ในวิธีนี้) ร่วมกับการใช้วิธีการเรียนรู้ของเครื่อง (Machine Learning) โดยในช่วงปี ค.ศ. 2014 จนถึงปัจจุบันนำวิธีการเรียนรู้แบบลึก (Deep Learning) ซึ่งเป็นโครงข่ายประสาทเทียมที่มีความซับซ้อน โดยมากมักจะมีขนาดของชั้น (Layer) มากกว่า 2 ชั้นขึ้นไปมาใช้งาน เช่น ในงานของ S. Wang et al. (2018) นำโครงข่ายที่ชื่อว่า “Deep Belief Networks” มาใช้ในการสร้างการรู้จำภูมินามจากภาษาจีน ซึ่งใช้คลังข้อความภาษาจีนที่มีการติดแท็กบอกหน้าที่ของคำที่เป็น ภูมินาม นำมาเป็นข้อมูลฝึกฝนให้กับโครงข่าย (Zhang et al., 2012) โดยเปรียบเทียบความถูกต้องกับวิธี CRF แสดงผลลัพธ์ตามรูปที่ 16



รูปที่ 16 แสดงการเปรียบเทียบค่าความถูกต้องโดยรวม (F1-Score) ระหว่างวิธี Deep Belief Networks และ CRF (S. Wang et al., 2018)

จากรูปที่ 16 ค่าความถูกต้องโดยรวมของ Deep Belief Networks มีแนวโน้มเพิ่มขึ้นอย่างรวดเร็วเมื่อเทียบกับวิธี CRF สำหรับคลังข้อมูลที่มีขนาดน้อยกว่า 1 ล้านอักขระ แต่เมื่อข้อมูลคลังคำในช่วง 1.3 ล้านอักขระขึ้นไปพบว่าค่าความถูกต้องใกล้เคียงกันมากจนถึงช่วงที่คลังคำมีขนาดมากกว่า 1.5 ล้านคำทั้งสองเทคนิคพบว่าไม่มีความแตกต่างกัน ดังนั้นจากงานวิจัยนี้พบว่า หากใช้ข้อมูลฝึกฝนเป็นคลังคำที่มีขนาดเล็ก Deep Belief Networks ให้ความถูกต้องที่ดีกว่า CRF แต่หากชุดข้อมูลฝึกฝนคลังคำมีขนาดใหญ่ วิธี CRF ก็ให้ผลที่ดีไม่ต่างกัน โดยจากงานวิจัยนี้พบว่าเมื่อนำวิธี Deep Belief Networks + CRF ให้ผลดีที่สุด แสดงตามตารางที่ 2

ตารางที่ 2 ตารางเปรียบเทียบสรุปความถูกต้อง ระหว่างเทคนิคการรู้จำภูมินาม 3 แบบ (S. Wang et al., 2018)

Model	Precision (P)	Recall (R)	F1 Value
DBNs	0.8146	0.7749	0.7943
CRF	0.8198	0.7634	0.7906
DBNs + CRF	**0.7901**	**0.9375**	**0.8575**

งานวิจัยของ (R. Santos et al., 2018) นำเทคนิคโครงข่ายประสาทเทียมแบบวกกลับ (Recurrent Neural Network : RNN) แบบประตูสัญญาณวกกลับ (Gated Recurrent Unit : GRU) มาใช้ในการสกัดข้อมูลภูมินามและจับคู่ (Matching) กับข้อมูลจากอักขรานุกรมทางภูมิศาสตร์ของ Geonames พบว่าให้ค่าความถูกต้องโดยรวมถึง 0.8875

3.2 งานวิจัยทางด้านการรู้จำชื่อเฉพาะภาษาไทย (Thai NER)

งานวิจัยในการรู้จำชื่อเฉพาะสำหรับภาษาไทยนั้นในปัจจุบันถือว่ายังมีอยู่จำกัดเมื่อเทียบกับภาษาอื่นที่มีความนิยม เช่น ภาษาอังกฤษ ฝรั่งเศส จีน ในงานของ Chanlekha et al. (2002) ใช้วิธีการทางสถิติร่วมกับแบบจำลองที่ใช้กฎฮิวริสติก (Heuristic rule-based model) โดยสร้างกฎที่ยึดตามบริบทของชื่อเฉพาะ เช่น คำสำคัญหรือคำที่อยู่ใกล้กับชื่อเฉพาะ ซึ่งผลที่ได้พบว่าใช้ได้ดีกับงานเขียนประเภทนิตยสารแต่ในขณะที่หากเป็นข้อมูลจากหนังสือพิมพ์จะให้ผลที่ไม่ดีเท่าที่ควรเนื่องจากงานเขียนในนิตยสารมีลักษณะการใช้คำเป็นทางการ มีรูปแบบที่ชัดเจนกว่าหนังสือพิมพ์

หลังจากนั้น Chanlekha and Kawtrakul (2004) นำแบบจำลอง Maximum Entropy หรือ Logistic Regression มาใช้ร่วมกับกฎฮิวริสติก และพจนานุกรมศัพท์เฉพาะ โดยคลังข้อมูลที่นำมาใช้ผ่านการตัดคำ (tokenization) มาแล้วซึ่งผลที่ออกมาพบว่าให้ผลดีกับชื่อเฉพาะที่เป็นชื่อบุคคล ส่วนชื่อองค์กรนั้นพบว่าหากเป็นคำที่ยาวจะมีปัญหาในการสกัดข้อมูลเนื่องจากข้อมูลมีการกระจายตัวมากทำให้ประสิทธิภาพในการคำนวณค่าน้ำหนักฟังก์ชันคุณสมบัติลดลง (เช่น ถ้าเป็นคำที่ถัดไป 2 คำ หรือก่อนหน้า 2 คำ ความถูกต้องเหลือ 0.776 แต่หากเป็นคำที่อยู่ก่อนหน้าและถัดไป 1 คำ ความถูกต้องจะเพิ่มเป็น 0.8987 ส่วนชื่อเฉพาะของสถานที่มีความถูกต้องน้อยที่สุด เนื่องจากมีความกำกวมมากกว่าชื่อเฉพาะประเภทอื่น Tirasaroj and Aroonmanakun (2009) นำแบบจำลอง CRF มาใช้ในการสร้างระบบการรู้จำชื่อเฉพาะภาษาไทย โดยใช้ข้อมูลศึกษาจากคลังข้อมูล “BEST 2009” ของ NECTEC จำนวนโดยประมาณ 90,000 คำ ในการทดลองนี้เปรียบเทียบการฝึกฝนแบบจำลอง CRF ระหว่างข้อมูลที่ผ่านการตัดคำและข้อมูลที่ผ่านการตัดพยางค์ (Syllable segmentation) จากผลการทดลองพบว่าค่าความถูกต้องโดยรวมของแบบจำลองที่ผ่านการตัดคำและแบบจำลองที่ผ่าน

การตัดพยางค์นั้นให้ค่าความถูกต้องที่ใกล้เคียงกัน คือ 0.8039 และ 0.808 ตามลำดับ แต่หากพิจารณาเฉพาะ NER ของสถานที่ (LOCATION) จะพบว่าแบบจำลองให้ค่าความถูกต้องที่ 0.7372 และแบบจำลองการตัดพยางค์ให้ค่าความถูกต้องที่ดีกว่าคือ 0.7692 ในปี 2019 Thattinaphanich and Prom-On (2019) สร้าง NER ขึ้นมาโดยใช้โครงข่ายประสาทเทียมวกกลับแบบ Bi-Directional Long Short Term Memory : Bi-LSTM ร่วมกับ CRF จำแนก 13 ชั้นข้อมูลซึ่ง Location เป็นหนึ่งในชั้นข้อมูล แบบจำลองที่ให้ผลดีที่สุดในงานวิจัยนี้ให้ค่าความถูกต้องโดยรวมอยู่ที่ 0.8773 และปัจจุบันแบบจำลองการถ่ายโอนความรู้ BERT ที่ผ่านการฝึกฝนเพิ่มเติมกับคลังข้อมูลภาษาไทยขนาดใหญ่เรียกว่า WangchanBERTa เป็นแบบจำลองสำหรับภาษาไทยที่ทันสมัยที่สุด

3.3 งานวิจัยทางด้านการสกัดข้อมูลตำแหน่งจากสื่อสังคมออนไลน์

งานวิจัยทางด้านการประมาณตำแหน่งจากสื่อสังคมออนไลน์ส่วนใหญ่มีการกล่าวถึงข้อมูลที่เป็นตัวบ่งชี้เชิงพื้นที่ (Spatial indicators) โดยสรุปมี 7 อย่าง (Ajao et al., 2015) ได้แก่ 1) สถานที่ที่ถูกอ้างถึงในข้อความ 2) เครือข่ายทางสังคม 3) ที่อยู่ของผู้ใช้งาน (User profiles) 4) ข้อมูลแท็กภูมิศาสตร์ (Geotags) 5) การเชื่อมโยงข้อมูลจากแหล่งอื่นเพิ่มเติม (Third-party source) 6) โซนเวลา 7) ตำแหน่งของผู้ให้บริการเว็บไซต์ (IP Address)

Backstrom et al. (2010) ใช้วิธีการจัดอันดับความสัมพันธ์ของเพื่อนที่อยู่ใกล้กันโดยให้อันดับความสัมพันธ์แปรผกผันกับระยะทาง (Inverse Distance Weight) ร่วมกับการใช้วิธีหาความเป็นไปได้สูงสุด (Maximum likelihood) เพื่อทำนายตำแหน่ง ผลที่ได้สามารถคาดเดาได้ถูกต้อง 69.1% หากเครือข่ายสังคมของผู้ใช้งานห่างกันไม่เกิน 25 ไมล์ หรือประมาณ 40.23 กิโลเมตร เปรียบเทียบกับข้อมูลที่ได้จาก Ip address ซึ่งมีความถูกต้องน้อยกว่า คือ 57.2%

ในงานของ Xu et al. (2014) นำเสนออัลกอริทึมชื่อว่า “Location Propagation Probability” โดยมีวัตถุประสงค์เพื่อคาดคะเนถิ่นฐานของผู้ใช้งานจากข้อมูลเครือข่ายทางสังคมและที่อยู่ของผู้ใช้งานบนเว่ยป๋อ (Weibo) ซึ่งเป็นสื่อสังคมออนไลน์ของจีน ผลที่ได้พบว่าให้ความถูกต้องที่ 68.2% ในระดับเมือง และ 73.7% ในระดับจังหวัด

Williams et al. (2017) ใช้ข้อมูลตำแหน่งจากที่อยู่ของผู้ใช้งานและเครือข่ายทางสังคมจากทวิตเตอร์วิเคราะห์โดยนำอัลกอริทึม Density-based clustering algorithm and noise : DBSCAN ร่วมกับ K-means ใช้ในการจัดกลุ่มของความสัมพันธ์เชิงตำแหน่งและคาดคะเนพิกัดภูมิศาสตร์ ซึ่งมีความถูกต้อง 30% ในระดับสถานที่ รัศมี 5 กิโลเมตร

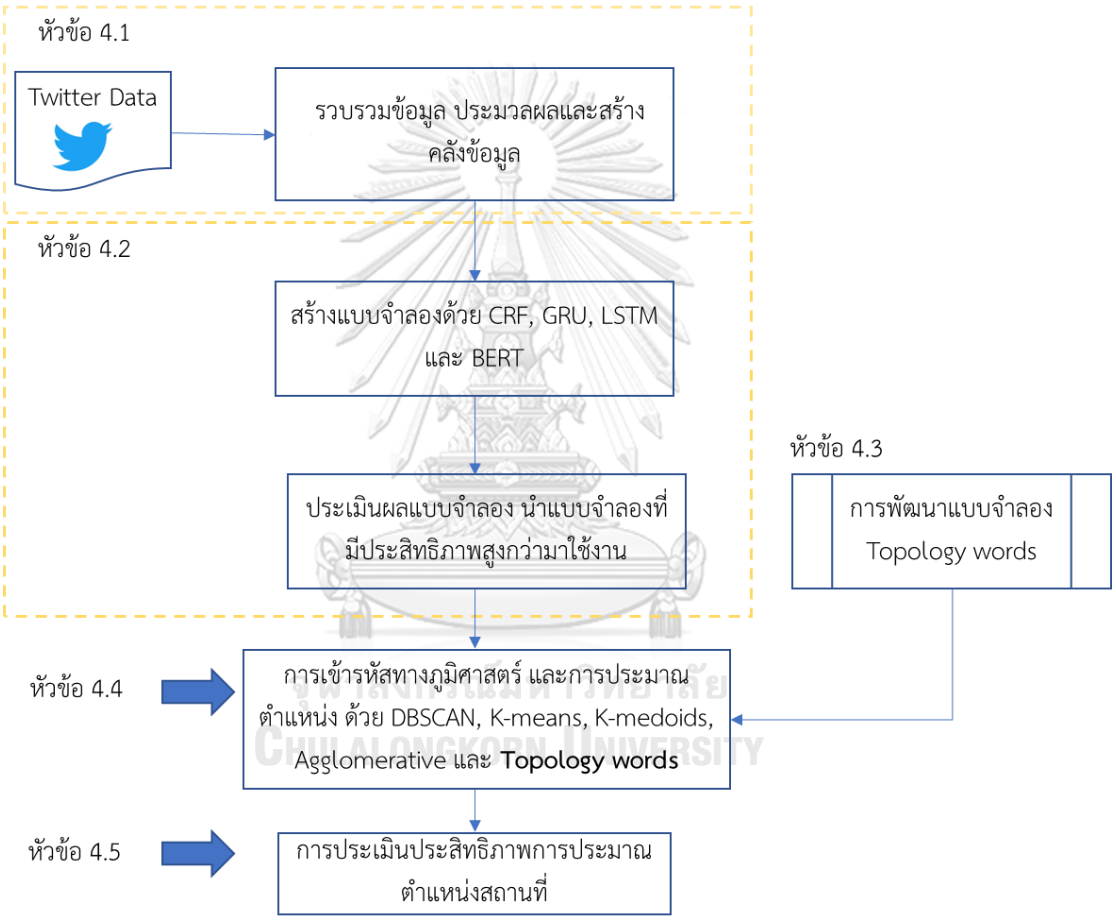
3.4 สรุปการทบทวนงานวิจัย

จากการทบทวนงานวิจัยทั้งสามหัวข้อที่สำคัญในส่วนของเทคนิคการรู้จำภูมินามนั้นจะพบว่าเทคนิคที่เป็นฐานของการรู้จำชื่อเฉพาะในภาษาไทยนั้นคือ Conditional random fields (CRF) จากงานวิจัยของ Tirasaroj and Aroonmanakun (2009) และการใช้โครงข่ายประสาทเทียมวกกลับชนิด Long short term memory (LSTM) แบบสองทิศทาง (Bi-directional) หรือ Bi-LSTM ร่วมกับแบบจำลอง CRF (Thattinaphanich & Prom-On, 2019) รวมถึงวิธีการของ R. Santos et al. (2018) ที่ใช้โครงข่ายประสาทเทียมแบบประตูสัญญาณวกกลับ (Gated recurrent unit : GRU) นั้นให้ผลดีที่สุดแต่ เทคนิคของ S. Wang et al. (2018) ซึ่งใช้โครงข่ายแบบ Deep Belief ร่วมกับ CRF นั้นให้ค่าความถูกต้องโดยรวมที่ใกล้เคียงกัน และในปัจจุบันแบบจำลองการถ่ายโอนความรู้ด้วยสถาปัตยกรรม Bi-directional encoder and decoder (BERT) ถูกนำมาใช้สำหรับงานรู้จำชื่อเฉพาะภาษาไทยและให้ผลที่ดี ในงานวิจัยนี้จึงทำการทดลองเปรียบเทียบประสิทธิภาพในแบบจำลอง 3 ประเภทใหญ่ คือ แบบจำลองการเรียนรู้ของเครื่องแบบมีลำดับ (CRF) แบบจำลองโครงข่ายประสาทเทียมแบบวกกลับ (LSTM และ GRU) รวมทั้งแบบจำลองการถ่ายโอนความรู้ (BERT) เพื่อนำแบบจำลองที่มีประสิทธิภาพดีที่สุดมาใช้เป็นแบบจำลองการรู้จำภูมินาม

สำหรับส่วนของการประมาณค่าพิกัดจากสถานที่ที่ไม่สามารถเข้ารหัสทางภูมิศาสตร์ได้นั้น จากหัวข้อที่ 3.3 ข้อมูลที่เป็นตัวบ่งชี้เชิงพื้นที่ (Spatial indicators) ในงานวิจัยของ Ajao et al. (2015) โดย สถานที่ที่ถูกอ้างถึงในข้อความเป็นหนึ่งในตัวบ่งชี้ที่สำคัญอีกทั้งยังรวมถึงคำที่เป็นตัวบ่งบอกความสัมพันธ์ระหว่างสถานที่แต่ละแห่งนั้นๆ ในงานวิจัยนี้จึงให้ความสำคัญของชื่อสถานที่ที่ถูกอ้างถึงในข้อความเป็นพารามิเตอร์หลักแต่เนื่องจากในแต่ละข้อความมีคำที่อาจจะกล่าวถึงความสัมพันธ์ระหว่างสถานที่นั้นอยู่หลายคำ หรืออาจจะเป็นอ้างถึงแบบไม่จำเพาะเจาะจง เช่น “แถวรามอินทรารถติดน่าเบื่อมาก อยากไปเที่ยวเกาะกูด เกาะกง เกาะกวม เกาะเสม็ด” จากประโยคข้างต้นจะเห็นว่ามีการอ้างถึงสถานที่ที่ตำแหน่งไม่ได้มีความสัมพันธ์กันจำนวนมาก การกรองชื่อสถานที่ที่ไม่เกี่ยวข้องกับ”รามอินทรา” ออกไปจะทำให้ทราบตำแหน่งของเจ้าของข้อความที่ทวีตได้ชัดเจนขึ้นอัลกอริทึมแบบจัดกลุ่มจึงเข้ามาแก้ไขปัญหาในส่วนนี้ของงานวิจัย โดยจากการทบทวนงานวิจัยที่ผ่านมา งานวิจัยนี้จึงเลือกใช้แบบจำลอง DBSCAN ที่ Williams et al. (2017) โดยนำพารามิเตอร์สถานที่ที่ถูกอ้างถึงในข้อความมาเป็นพารามิเตอร์หลักและมีการเปรียบเทียบกับแบบจำลอง K-means , K-medoids, Agglomerative clustering รวมทั้งมีการนำเสนอแบบจำลองใหม่ที่นำเอาชื่อสถานที่ที่ถูกอ้างอิงในข้อความและคำบ่งชี้ความสัมพันธ์ระหว่างสถานที่มาใช้งานเรียกอัลกอริทึมนี้ว่า Topology word

บทที่ 4
วิธีการดำเนินงานวิจัย

ในบทนี้นำเสนอในรายละเอียดของขั้นตอนการดำเนินงานวิจัย แบ่งเนื้อหาหลักออกเป็น 4 ส่วนดังนี้ 4.1 การเตรียมข้อมูล 4.2 การพัฒนาแบบจำลองสำหรับการรู้จำภูมินาม 4.3 การเข้ารหัสทางภูมิศาสตร์และการประมาณตำแหน่ง 4.4 การประมวลผลและการวิเคราะห์ข้อมูล โดยแสดงภาพรวมของเนื้อหาตามรูปที่ 17



รูปที่ 17 แสดงขั้นตอนของการดำเนินงานวิจัยโดยภาพรวม

4.1 การเตรียมข้อมูล

ในหัวข้อนี้จะกล่าวถึงองค์ประกอบและขั้นตอนต่างๆในการเตรียมข้อมูลจนเป็นคลังข้อมูล เพื่อนำไปพัฒนาแบบจำลองการรู้จำภูมินามและการเข้ารหัสทางภูมิศาสตร์ มีรายละเอียดดังนี้

4.1.1 พื้นที่ศึกษา

พื้นที่ศึกษาในงานวิจัยกำหนดให้ใช้ข้อมูลในเขตพื้นที่กรุงเทพฯและปริมณฑล (นครปฐม นนทบุรี ปทุมธานี สมุทรปราการ สมุทรสาคร) เนื่องจากมีจำนวนประชากรและความหนาแน่นของประชากรต่อพื้นที่มากที่สุดในประเทศไทย (กระทรวงมหาดไทย, 2562) โดยการกำหนดขอบเขตพื้นที่ศึกษาจากพารามิเตอร์ Location บน Endpoint Filter ซึ่งกำหนดเป็นกรอบสี่เหลี่ยม (Bounding Box : BBOX) ประกอบด้วยพิกัดมุมซ้ายล่าง มุมขวาบนตามลำดับ โดยค้นหาจากค่าพิกัดที่ตกอยู่ใน BBOX หากข้อความไม่มีพิกัดจะค้นหาจากชื่อสถานที่ซึ่งอยู่ใน Place Field แทนว่าอยู่ในภูมิภาคที่ระบุในเขต BBOX หรือไม่ โดยหากไม่มีทั้งพิกัดและชื่อสถานที่จะข้ามข้อความนั้นไป

4.1.2 การรวบรวมข้อมูลจากทวิตเตอร์

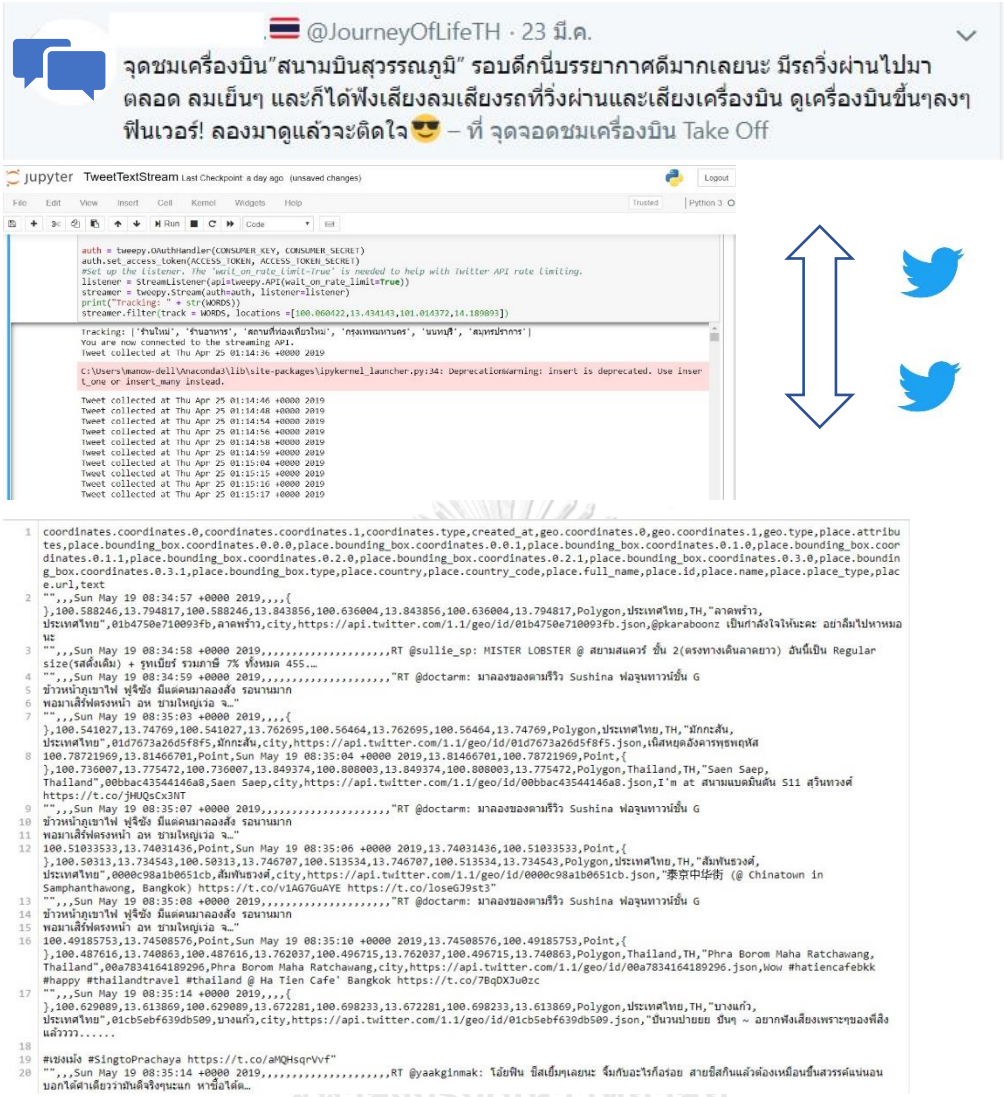
ข้อมูลที่ใช้ในงานวิจัยนี้มี 2 ส่วนหลักที่สำคัญ คือ ส่วนที่เป็นข้อมูลสำหรับฝึกฝนแบบจำลอง และสองคือส่วนของข้อมูลที่นำมาเป็นข้อมูลทดสอบในการประมาณตำแหน่งของสถานที่

4.1.2.1 ข้อมูลสำหรับฝึกฝนแบบจำลอง

ข้อมูลที่ใช้ในงานวิจัยนี้นำมาจากข้อมูลทวิตเตอร์จำนวน 28,082 ข้อความ 1,974,211 ตัวอักษร โดยเริ่มเก็บข้อมูลตั้งแต่เดือนกันยายนถึงพฤศจิกายน 2562 มาใช้พัฒนาแบบจำลอง เครื่องมือที่ใช้ในการดาวน์โหลดคือ คลังคำสั่งบนภาษา python ชื่อ tweepy สำหรับการดาวน์โหลดกำหนด BBOX ด้วยค่าละติจูดและลองจิจูดให้ครอบคลุมบริเวณพื้นที่กรุงเทพและปริมณฑล ไวยากรณ์ของ tweepy กำหนดให้ใส่เฉพาะพิกัดภูมิศาสตร์ของมุมซ้ายล่าง (Bottom Left: BL) และพิกัดภูมิศาสตร์ของมุมขวาบน (Top Right: TR) โดยลักษณะลิสต์ของค่าพิกัดคือ [LonBL, LatBL, LonTR, LatTR] ตามลำดับ ตัวอย่างเช่น BBOX = [100.060422, 13.434143, 101.014372, 14.189893] เมื่อข้อมูลผ่านตัวกรองแล้วจึงจัดเก็บข้อมูลเป็นไฟล์ JSON แสดงตามรูปที่ 18

4.1.2.2 ข้อมูลจากทวิตเตอร์ที่นำมาเป็นข้อมูลทดสอบในการประมาณตำแหน่งของสถานที่

มีการใช้ tweepy ร่วมกับ google custom search API และเทคนิค web scraping เนื่องจากข้อมูลที่ได้จาก twitter API ที่เป็นแบบฟรีนั้นจะได้ข้อมูลภายในช่วง 1-2 วันเท่านั้น ทำให้ต้องใช้เทคนิคอื่นร่วมเพื่อให้ได้ข้อมูลมากที่สุด โดยมีข้อมูลที่เกี่ยวกับสถานที่จำนวน 100 ชุด และแต่ละชุดประกอบไปด้วยข้อความตั้งแต่ 3 ข้อความ/ชุด จัดเก็บข้อมูลเป็นไฟล์ JSON



รูปที่ 18 แสดงตัวอย่างการจัดเตรียมข้อมูลทวิตเตอร์ด้วยคลังคำสั่ง Tweepy

4.1.3 การสร้างคลังข้อมูลเพื่อใช้ในงานวิจัย

ข้อมูลที่รวบรวมมาจากทวิตเตอร์จะนำมาประมวลผลเบื้องต้น เช่น การลบข้อความซ้ำ สัญลักษณ์ emoticon ในข้อความ ฯลฯ เพื่อให้ข้อมูลอยู่ในรูปแบบที่พร้อมใช้งาน โดยในงานวิจัยนี้ข้อมูลจะถูกเก็บใน format csv แล้วนำไปติดฉลากข้อมูล (Label) ด้วยวิธี Human annotation โดยมีการสร้างคู่มือและประเมินผลความรู้ความเข้าใจในรายละเอียดของการดำเนินงานด้วยแบบทดสอบเป็นตัวอย่างข้อมูลจำนวน 100 บรรทัด จากจำนวน Annotator ทั้งหมดพบว่าให้คำตอบตรงกันคิดเป็นร้อยละ 91 สำหรับประเภทของสถานที่ในงานวิจัยนี้ นำตัวอย่างการจัดประเภทและชนิดมาจากชนิดของสถานที่จากผู้ให้บริการออนไลน์ (Google, 2023; Here, 2023; Nostra, 2022) โดยมีการปรับให้เหมาะสมกับการใช้งานสำหรับงานวิจัยนี้แบ่งออกเป็น 18 ประเภทอยู่ในกลุ่มประเภทหลัก 5

กลุ่ม ได้แก่ 1) สถานที่ธรรมชาติ 2) สิ่งปลูกสร้าง สถานที่ที่มนุษย์สร้างขึ้น 3) ขอบเขตการปกครอง 4) สถานที่ที่ตั้งอยู่นอกเขตประเทศไทย 5) สถานที่อื่นที่ไม่สามารถจัดอยู่ใน 4 ประเภทข้างต้นได้ โดยแสดงรายละเอียดและสัญลักษณ์ของฉลากข้อมูล ตามตารางที่ 3 และ 4 ตามลำดับดังนี้

ตารางที่ 3 แสดงรายละเอียดการแบ่งประเภทของสถานที่ในงานวิจัย

กลุ่มของสถานที่	**ประเภทของสถานที่**	**นิยาม**
1) สถานที่ธรรมชาติ	1 สถานที่ธรรมชาติ	พื้นที่ป่า อุทยาน ภูเขา แหล่งน้ำ สถานที่ท่องเที่ยวตามธรรมชาติ
2) สิ่งปลูกสร้าง	2 ที่พักอาศัย	ชื่อหมู่บ้าน หมู่บ้านจัดสรร อาคารชุดพักอาศัย โรงแรม ที่พัก
	3 ศาสนสถาน	สถานที่สำคัญทางศาสนา เช่น วัด โบสถ์ มัสยิด สถานปฏิบัติธรรม
	4 สถานศึกษา	โรงเรียน วิทยาลัย มหาวิทยาลัย โรงเรียนกวดวิชา สถาบันอื่นๆที่จัดหลักสูตรการเรียนการสอนทุกประเภท ห้องสมุดทุกประเภท
	5 สถานพยาบาล	โรงพยาบาลทั้งของรัฐและเอกชน โรงพยาบาลชุมชน สถานีอนามัย ศูนย์สุขภาพ คลินิกทุกประเภท
	6 การคมนาคมขนส่ง	สถานที่เกี่ยวกับการคมนาคมอ้างอิงตำแหน่งแบบเป็นจุด (Point) เช่น สถานีรถไฟ สถานีรถไฟฟ้า ท่าเรือ สนามบิน (ในงานวิจัยนี้ไม่รวม ถนน ตรอก ซอย)
	7 ร้านค้าขนาดย่อม	ร้านสะดวกซื้อ ร้านค้าปลีก
	8 ตลาด	ตลาดทุกประเภท หมายถึงตลาดที่มีที่ตั้งแน่นอนและตลาดนัด เช่น ตลาดบางเขน ตลาดสามย่าน ตลาดนัดสวนจตุจักร
	9 ร้านอาหาร	ร้านอาหารทุกประเภท เช่น ภัตตาคาร ร้านอาหารในโรงแรม ร้านอาหารในห้างสรรพสินค้า ร้านอาหารริมทาง ร้านกาแฟ

ตารางที่ 3 (ต่อ) แสดงรายละเอียดการแบ่งประเภทของสถานที่ในงานวิจัย

กลุ่มของสถานที่	ประเภทของสถานที่	นิยาม
	10 ห้างสรรพสินค้า	ศูนย์การค้าขนาดใหญ่ เช่น มาบุญครอง เซ็นทรัลเวิลด์ สยามพารากอน ซุปเปอร์มาร์เก็ต เช่น โลตัส บิ๊กซี ฯลฯ
2) สิ่งปลูกสร้าง	11 อาคารสำนักงาน	บริษัท (Office) สถานที่ที่เป็นอาคารสำนักงานของเอกชน อาคารสูง ยกเว้นร้านอาหาร ร้านค้า ห้างสรรพสินค้า ในข้อ 7-10
	12 สถานที่ราชการ	สถานที่ อาคาร ของหน่วยงานราชการ รัฐวิสาหกิจ หรือหน่วยงานอื่นที่รัฐกำกับดูแล
3) ขอบเขตการปกครอง	13 สถานที่สำหรับสันทนาการ	สนามกีฬา สวนสาธารณะ สวนสัตว์ สวนสนุก โรงภาพยนตร์ โรงละคร สตูดิโอ สถานที่ท่องเที่ยวที่เกิดจากมนุษย์สร้างขึ้น เช่น เมืองโบราณ ฯลฯ พิพิธภัณฑ์
4) สถานที่ และ ขอบเขตการปกครองในพื้นที่อื่นนอกจากประเทศไทย	17 สถานที่ที่อยู่นอกประเทศ	สถานที่ที่อยู่นอกประเทศ (Foreign Place) เช่น ลอนดอน ตลาดฮงแด กัวลาลัมเปอร์ ฯลฯ
5) สถานที่อื่นๆ	18 สถานที่อื่นๆ	สถานที่ที่ไม่สามารถจัดให้เข้าอยู่ใน 4 ประเภท 17 ชนิด ข้างต้นได้

ตารางที่ 4 แสดงรายละเอียดสัญลักษณ์ของการติดฉลากข้อมูล

ชนิดข้อมูล	สัญลักษณ์ที่ใช้
1 สถานที่ธรรมชาติ (Natural Place)	<NAT> ... </NAT>
2 ที่พักอาศัย (Residential)	<RES> ... </RES>
3 ศาสนสถาน (Religious Place)	<RP> ... </RP>
4 สถานศึกษา (Academic Place)	<ACP> ... </ACP>
5 สถานพยาบาล (Healthcare Place)	<HP> ... </HP>
6 การคมนาคมขนส่ง (Transportation)	<TRAN> ... </TRAN>
7 ร้านค้าปลีก (Store)	<STORE> ... </STORE>
8 ตลาด (Market)	<MKT> ... </MKT>
9 ร้านอาหาร (Restaurant)	<RES> ... </RES>
10 ห้างสรรพสินค้า (Department store)	<DEP> ... </DEP>
11 อาคารสำนักงาน (Business office)	<BSN> ... </BSN>
12 สถานที่ของราชการ (Government)	<GOV> ... </GOV>
13 สถานที่สำหรับสันทนาการ (Recreation)	<RCT> ... </RCT>
14 อนุสาวรีย์ (Monument)	<MON> ... </MON>
15 ตรอก ซอย ถนน (Road)	<ROAD> ... </ROAD>
16 ขอบเขตการปกครอง (Admin)	<ADMIN> ... </ADMIN>
17 สถานที่ที่อยู่นอกประเทศไทย (Foreign Place)	<FPLACE> ... </FPLACE>
18 สถานที่อื่นๆ (Others)	<OTHER> ... </OTHER>

4.1.4 หลักเกณฑ์การติดฉลากข้อมูล

ขั้นตอนที่ 1 - สถานที่ในงานวิจัยนี้นิยามถึงชื่อเฉพาะของสถานที่เท่านั้น เช่น เช่น “ฉันมาถึงสนามบิน เวลา 18:00” ในรูปประโยคแบบนี้ไม่สามารถระบุได้ว่าเป็นสนามบินไหน ดังนั้นไม่ต้องติดฉลากให้ข้อมูล แต่หากเป็น “ฉันมาถึงสนามบินสุวรรณภูมิเวลา 18:00” สามารถระบุได้ว่าเป็นสนามบินไหน ตัวอย่างการติดฉลากข้อมูล คือ “ฉันมาถึง<TRAN>สนามบินสุวรรณภูมิ</TRAN>เวลา 18:00”

ขั้นตอนที่ 2 - ตัวย่อของสถานที่ เมื่อดูตามบริบทของประโยคแล้วสามารถระบุได้ทันที ให้ติดฉลากข้อมูลตามประเภท เช่น “เดี๋ยวพรุ่งนี้วันหยุดเราไปกินกุ้งเผาที่ อย กัน” คำว่า อย ในประโยคนี้น่าจะหมายถึงจังหวัดอยุธยา

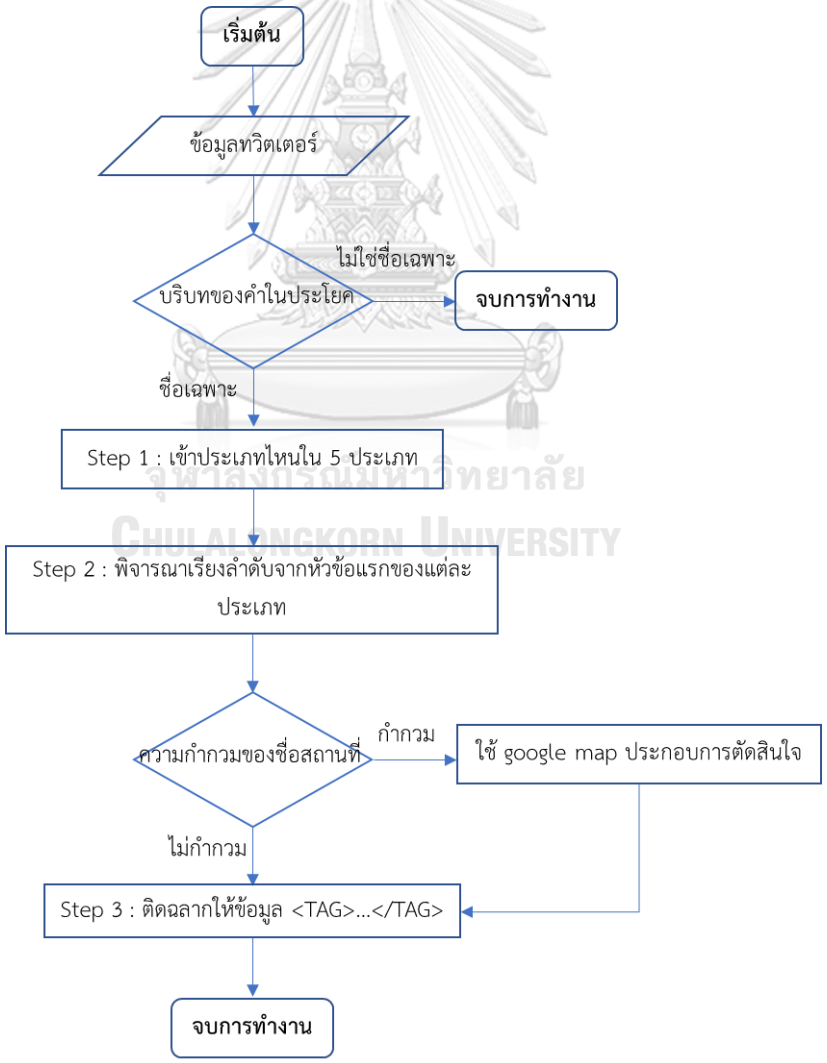
ขั้นตอนที่ 3 - ในการเลือกติดฉลากข้อมูลให้เลือกจากประเภทก่อน หลังจากนั้นในแต่ละประเภทให้เลือกพิจารณาจากหัวข้อแรกในแต่ละประเภทก่อน เช่น “เดี๋ยวเจอกันที่ BTS สยาม”

Step 1 : สถานที่จัดอยู่ในประเภทที่ 2

Step 2 : พิจารณาตั้งแต่หัวข้อที่ 4 มาก่อนจะพบว่ามาเข้าเกณฑ์ที่หัวข้อ 6 คือ Transportation จึงติดฉลากข้อมูลเป็น <TRAN>BTS สยาม</TRAN>

4) หากเป็นชื่อของย่าน เช่น ย่านวังหลัง ให้ติดฉลากข้อมูลสถานที่ที่คาดว่าเป็นแหล่งที่มาของชื่อย่านนั้น โดยย่านวังหลังอาจมาจาก ถนนวังหลัง (พิจารณาจากอันดับที่ google map คืนค่ามาให้) จึงติดฉลากข้อมูลเป็น <ROAD>ถนนวังหลัง</ROAD> ในกรณีนี้ใช้เครื่องมือคือ google map มาช่วยในการตัดสินใจ

จากหลักเกณฑ์ที่กล่าวมา 4 ข้อข้างต้นสรุปเป็นผังดำเนินงาน (Flow Chart) ได้ตามรูปที่ 19



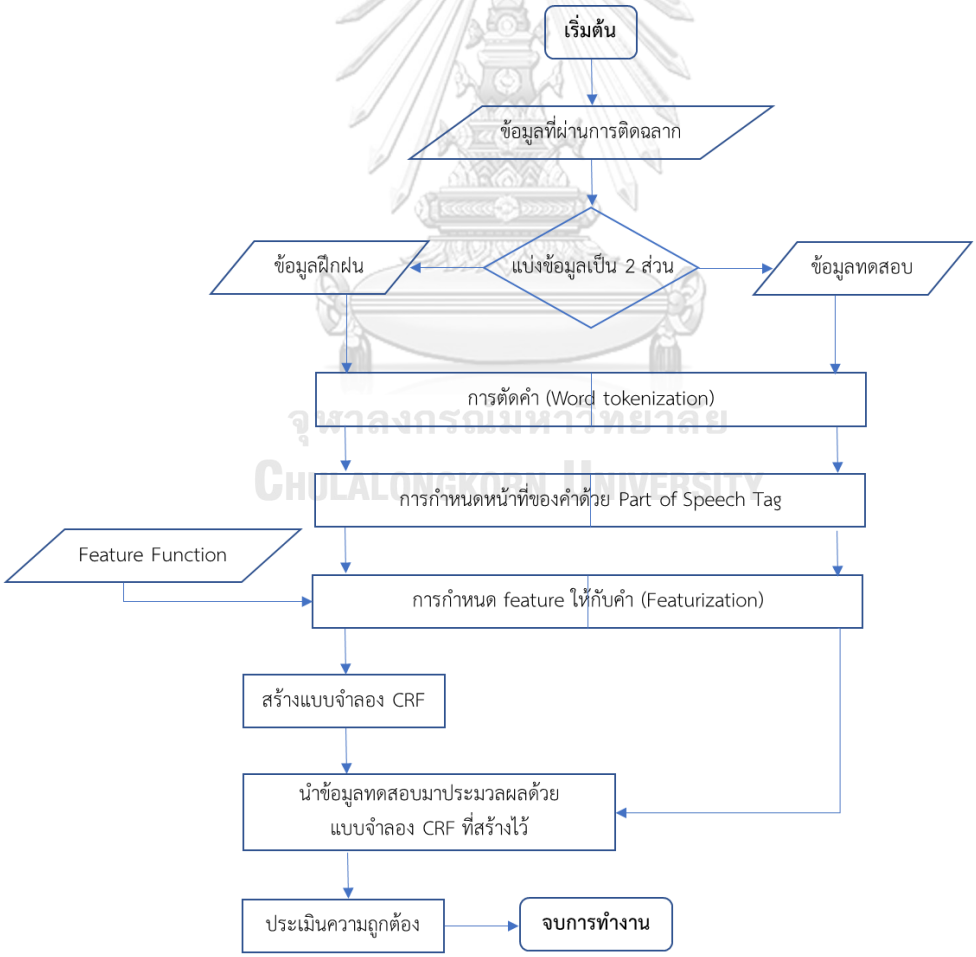
รูปที่ 19 ผังดำเนินการขั้นตอนการติดฉลากข้อมูล

4.2 การพัฒนาแบบจำลองเพื่อรู้จำภูมินาม

การพัฒนาเครื่องมือส่วนนี้เพื่อให้การสกัดข้อมูลเชิงพื้นที่จากข้อความมีประสิทธิภาพเพียงพอสำหรับข้อมูลภาษาไทยบนทวิตเตอร์โดยมีการพัฒนาแบบจำลองเป็น 2 แบบหลักๆคือ 1) การสร้างแบบจำลองเพื่อสกัดภูมินามด้วย CRF 2) แบบจำลองเพื่อสกัดภูมินามด้วยโครงข่ายประสาทเทียมทั้ง GRU และ LSTM และ 3) แบบจำลองเพื่อสกัดภูมินามจาก BERT ซึ่งจะแสดงรายละเอียดในการสร้างเครื่องมือและผลการทดลองเบื้องต้นดังนี้

4.2.1 การสร้างแบบจำลองเพื่อรู้จำภูมินามด้วย CRF

สำหรับขั้นตอนแรกก่อนการสร้างแบบจำลองจะเริ่มต้นด้วยแบ่งข้อมูลออกเป็น 2 ส่วนหลักคือ ชุดข้อมูลฝึกฝน (Training data) จำนวน 80% (22,445 ข้อความ) ข้อมูลปรับแบบจำลอง (Validate data) ประมาณ 10% (4,510 ข้อความ) และข้อมูลทดสอบ (Testing data) ประมาณ20% (5,617 ข้อความ) โดยสรุปขั้นตอนเป็นผังดำเนินการดังรูปที่ 20



รูปที่ 20 ผังดำเนินการแสดงขั้นตอนการสร้างแบบจำลองด้วย CRF

ขั้นตอนที่ 1 การตัดคำ - ในส่วนนี้ใช้เครื่องมือจากคลังคำสั่งชื่อว่า Attacut (Chormai et al., 2019) โดยผลที่ได้จะถูกเก็บเป็นลิสต์ของคำ

ขั้นตอนที่ 2 การทำ Part of Speech Tagging (PosTag) - ในส่วนนี้เป็นการบอกหน้าที่ของคำในแต่ละประโยค เช่น ‘ฉันไปโรงเรียน’ เมื่อผ่านการทำ PosTag จะกลายเป็น [(ฉัน,PPRS) , (ไป,VERB) , (โรงเรียน,NCMN)] โดยสัญลักษณ์ของ PosTag ในประโยคข้างต้นมีดังนี้

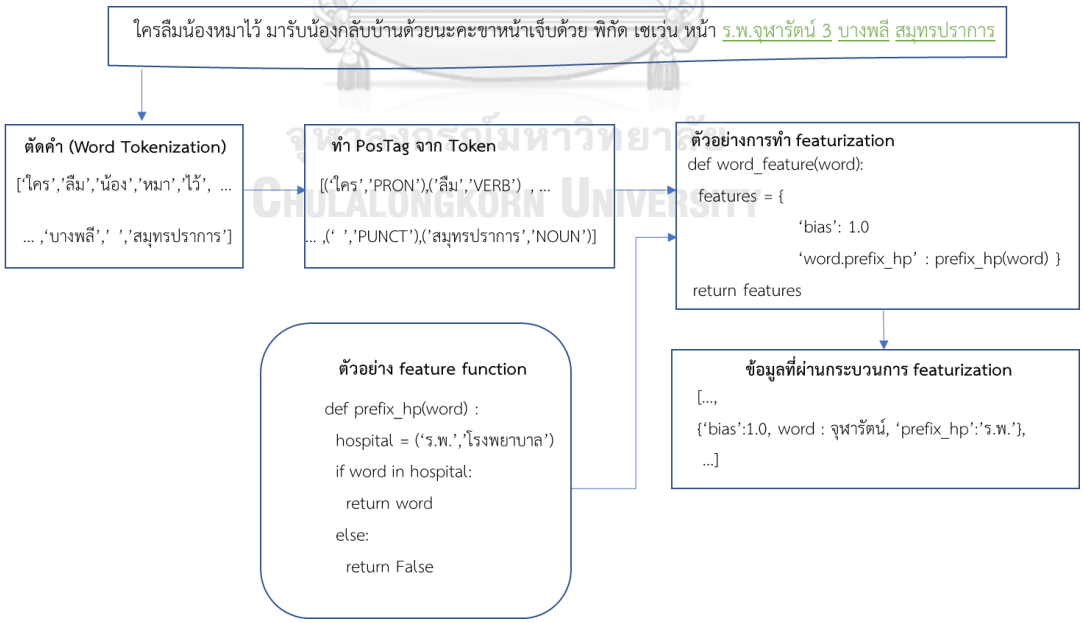
PPRS คือ Personal Pronoun

VERB คือ คำกริยา

NCMN คือ Common Noun

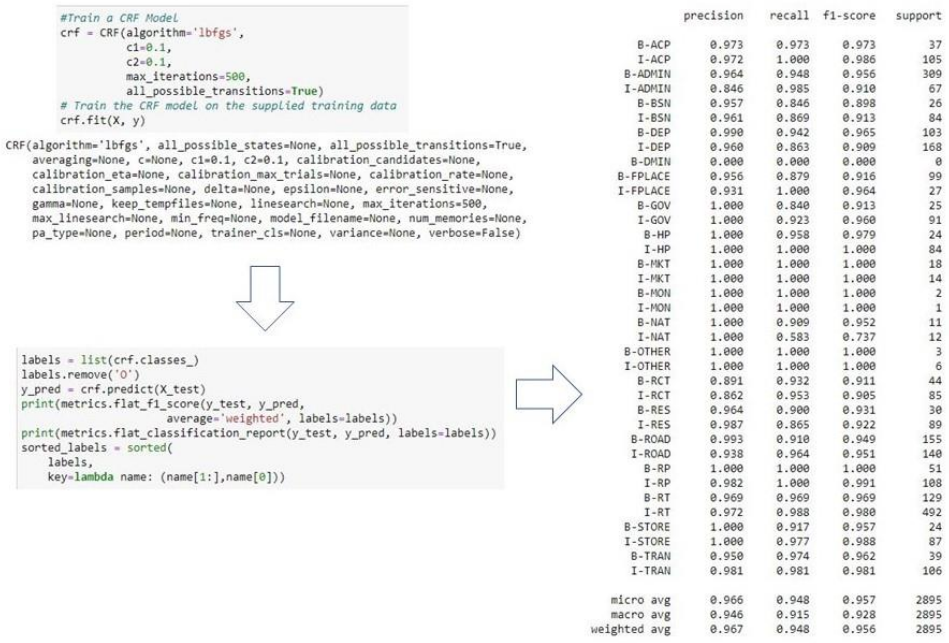
โดยนำเครื่องมือจากคลังคำสั่งในภาษาไพธอน ชื่อว่า pythainlp มาใช้งาน

ขั้นตอนที่ 3 - การสร้าง feature สำหรับแบบจำลอง (featurization) - ในการสร้างแบบจำลองด้วยวิธี CRF ขั้นตอนที่สำคัญคือการสร้าง feature เพื่อให้แบบจำลองเรียนรู้ feature ยิ่ง feature มีจำนวนมากเท่าไหร่ค่าที่ทำนายได้จากแบบจำลองมักจะให้ผลที่ดีแปรผันตามกัน ตัวอย่างในงานวิจัยนี้หนึ่งใน featureที่สำคัญคือ PosTag คำนำหน้า (Prefix) และ คำตามหลัง (Suffix) ฯลฯ เช่น ในการสร้าง feature ของชั้นข้อมูลที่อยู่อาศัย จะมีตัวอย่างคำนำหน้า เช่น ‘หมู่บ้าน’ , ’โรงแรม’ ฯลฯ หรือคำตามหลัง (Suffix) เช่น ‘วิลล่า’ , ’วิลเลจ’ , ’ธานี’ ฯลฯ ซึ่งจะสร้างเป็นฟังก์ชันขึ้นมาแล้วจึงเรียกใช้งานพร้อมกันทุกฟังก์ชัน ในขั้นตอน featurization แสดงตัวอย่างตามรูปที่ 21



รูปที่ 21 แสดงตัวอย่างการประมวลผลข้อมูลก่อนฝึกฝนแบบจำลอง

ขั้นตอนที่ 4 - การฝึกฝนแบบจำลอง - หลังจากขั้นตอนการทำ featurization เสร็จเรียบร้อย นำข้อมูลที่ได้ มาฝึกฝนแบบจำลองด้วยอัลกอริทึม CRF หลังจากนั้นนำไปทดสอบด้วยข้อมูลทดสอบที่เตรียมไว้ ประเมินประสิทธิภาพด้วยค่า Precision Recall และ F1 แสดงตัวอย่างตามรูปที่ 22



```
#Train a CRF Model
crf = CRF(algorithm='lbfgs',
          c1=0.1,
          c2=0.1,
          max_iterations=500,
          all_possible_transitions=True)
# Train the CRF model on the supplied training data
crf.fit(X, y)
```

```
CRF(algorithm='lbfgs', all_possible_states=None, all_possible_transitions=True,
    averaging=None, c=None, c1=0.1, c2=0.1, calibration_candidates=None,
    calibration_eta=None, calibration_max_trials=None, calibration_rate=None,
    calibration_samples=None, delta=None, epsilon=None, error_sensitive=None,
    gamma=None, keep_tempfiles=None, linesearch=None, max_iterations=500,
    max_linesearch=None, min_freq=None, model_filename=None, num_memories=None,
    pa_type=None, period=None, trainer_cls=None, variance=None, verbose=False)
```

```
labels = list(crf.classes_)
labels.remove('O')
y_pred = crf.predict(X_test)
print(metrics.flat_f1_score(y_test, y_pred,
                            average='weighted', labels=labels))
print(metrics.flat_classification_report(y_test, y_pred, labels=labels))
sorted_labels = sorted(
    labels,
    key=lambda name: (name[1:],name[0]))
```

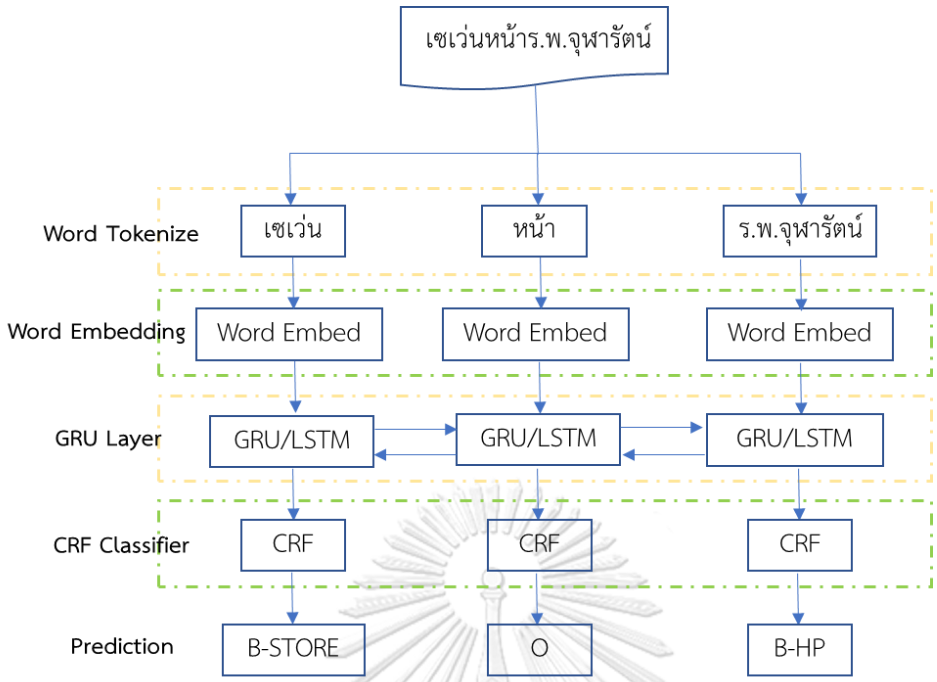
	precision	recall	f1-score	support
B-ACP	0.973	0.973	0.973	37
I-ACP	0.972	1.000	0.986	105
B-ADMIN	0.964	0.948	0.956	309
I-ADMIN	0.846	0.985	0.910	67
B-BSN	0.957	0.846	0.898	26
I-BSN	0.961	0.869	0.913	84
B-DEP	0.990	0.942	0.965	103
I-DEP	0.960	0.863	0.909	168
B-DMIN	0.000	0.000	0.000	0
B-FPLACE	0.956	0.879	0.916	99
I-FPLACE	0.931	1.000	0.964	27
B-GOV	1.000	0.840	0.913	25
I-GOV	1.000	0.923	0.960	91
B-HP	1.000	0.958	0.979	24
I-HP	1.000	1.000	1.000	84
B-MKT	1.000	1.000	1.000	18
I-MKT	1.000	1.000	1.000	14
B-MON	1.000	1.000	1.000	2
I-MON	1.000	1.000	1.000	1
B-NAT	1.000	0.909	0.952	11
I-NAT	1.000	0.583	0.737	12
B-OTHER	1.000	1.000	1.000	3
I-OTHER	1.000	1.000	1.000	6
B-RCT	0.891	0.932	0.911	44
I-RCT	0.862	0.953	0.905	85
B-RES	0.964	0.900	0.931	30
I-RES	0.987	0.865	0.922	89
B-ROAD	0.993	0.910	0.949	155
I-ROAD	0.938	0.964	0.951	140
B-RP	1.000	1.000	1.000	51
I-RP	0.982	1.000	0.991	108
B-RT	0.969	0.969	0.969	129
I-RT	0.972	0.988	0.980	492
B-STORE	1.000	0.917	0.957	24
I-STORE	1.000	0.977	0.988	87
B-TRAN	0.950	0.974	0.962	39
I-TRAN	0.981	0.981	0.981	106
micro avg	0.966	0.948	0.957	2895
macro avg	0.946	0.915	0.928	2895
weighted avg	0.967	0.948	0.956	2895

รูปที่ 22 แสดงตัวอย่างคำสั่งและผลลัพธ์ที่ได้จากการฝึกฝนแบบจำลอง CRF

ขั้นตอนที่ 5 - การประเมินประสิทธิภาพแบบจำลอง - หากได้ประสิทธิภาพของแบบจำลองที่เพียงพอแล้ว นำแบบจำลองที่ได้ไปสร้างเป็นฟังก์ชันแล้วนำไปทดสอบกับข้อมูลที่ไม่อยู่ในชุดฝึกฝนและชุดทดสอบที่สร้างไว้ข้างต้น ประเมินประสิทธิภาพ Precision Recall และ F1 เก็บค่าที่ได้ไว้รอเปรียบเทียบกับแบบจำลองที่สร้างด้วยอัลกอริทึมอื่นที่เหลือ

4.2.2 การสร้างแบบจำลองเพื่อรู้จำภูมินามด้วยโครงข่ายประสาทเทียมร่วมกับ CRF

ในแบบจำลองนี้จะมีขั้นตอนที่แตกต่างจากแบบจำลอง CRF คือ ในส่วนของการเตรียมข้อมูลจะต้องมีการสร้างคำฝังตัว (Word Embedding) เพิ่มขึ้นมาถัดจากการตัดคำ แต่ก็มีขั้นตอนที่ลดลงไปคือการสร้าง feature function แบบใน CRF เนื่องจาก แบบจำลองที่เป็นโครงข่ายประสาทเทียมจะเรียนรู้ feature ด้วยตนเองจากข้อมูลที่เตรียมไว้ เมื่อได้ feature จากโครงข่ายประสาทเทียมแล้วจึงส่งไปที่ CRF เพื่อใช้ในการจำแนก (Classification) โดยใช้สถาปัตยกรรมพื้นฐานของ Thattinaphanich and Prom-On (2019) สำหรับ LSTM และ Gridach and Haddad (2017) สำหรับ GRU ตัวอย่างสรุปขั้นตอนการสร้างแบบจำลองแสดงตามรูปที่ 23



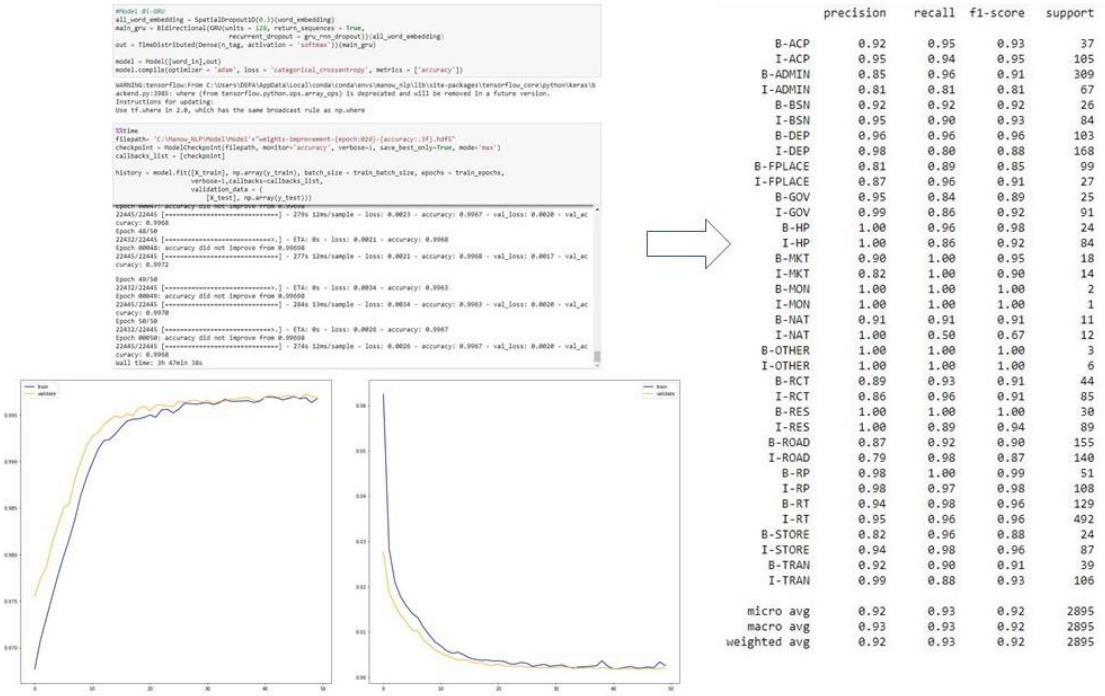
รูปที่ 23 แสดงตัวอย่างโดยสรุปของวิธี GRU/LSTM + CRF

ขั้นตอนที่ 1 - การตัดคำในส่วนนี้ใช้เครื่องมือเดียวกับวิธี CRF คือ Attacut ซึ่งเป็นคลังคำสั่งในภาษาไพธอน

ขั้นตอนที่ 2 - คำฝังตัวเป็นการเปลี่ยนข้อมูลจากตัวอักษรไปเป็นเวกเตอร์ที่เก็บความหมายของคำและคุณสมบัติของคำไว้ โดยใช้เครื่องมือ Thai2fit เมื่อได้เวกเตอร์ของคำแล้วจะส่งเวกเตอร์นี้เข้าไปสู่โครงข่ายประสาทเทียมแบบ GRU หรือ LSTM รวมทั้งมีการทำเชื่อมต่อเวกเตอร์ด้วย Character embedding

ขั้นตอนที่ 3 - ในชั้นของโครงข่ายประสาทเทียมแบบ GRU หรือ LSTM โครงข่ายที่นำมาใช้จะเป็นโครงข่ายแบบสองทิศทาง Bi-Directional ซึ่งข้อมูลที่ป้อนเข้าในแต่ละ perceptron ของโครงข่าย จะมาจากสองทิศทางคือ มาจากคำก่อนหน้าและคำถัดไปตามที่แสดงในรูปที่ 24 และจากนั้นจึงส่งไปที่ชั้น CRF เพื่อจำแนกชนิดของภูมินาม

ขั้นตอนที่ 4 - การประเมินประสิทธิภาพของแบบจำลองจะใช้ค่า Precision recall และ F1 หลังจากนั้นค่อยนำแบบจำลองไปทดสอบกับข้อมูลทดสอบ (Test set) แล้วเปรียบเทียบกับแบบจำลองอื่น ก่อนเลือกแบบจำลองที่ให้ประสิทธิภาพดีที่สุดไปใช้งาน

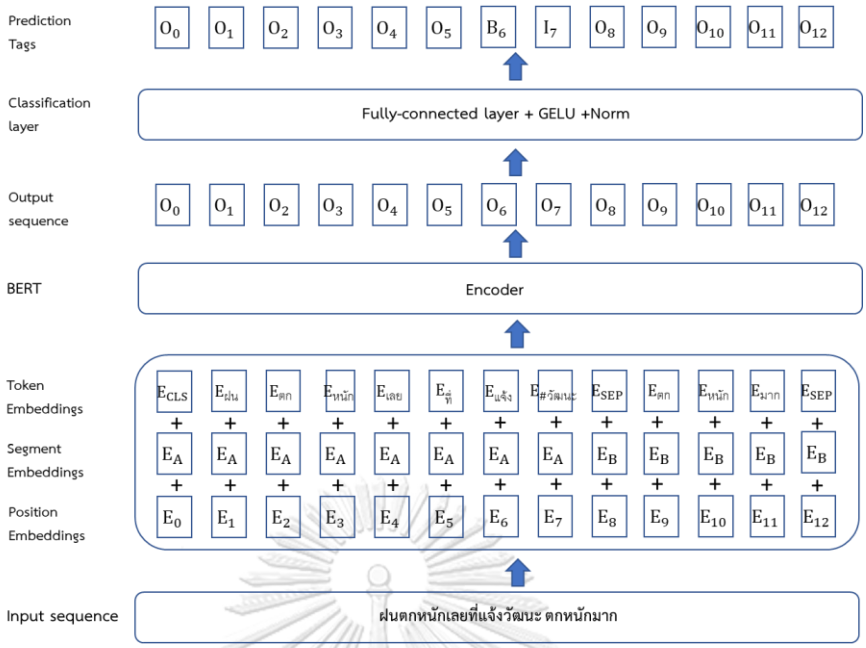


	precision	recall	f1-score	support
B-ACP	0.92	0.95	0.93	37
I-ACP	0.95	0.94	0.95	105
B-ADMIN	0.85	0.96	0.91	309
I-ADMIN	0.81	0.81	0.81	67
B-BSN	0.92	0.92	0.92	26
I-BSN	0.95	0.90	0.93	84
B-DEP	0.96	0.96	0.96	103
I-DEP	0.98	0.80	0.88	168
B-FPLACE	0.81	0.89	0.85	99
I-FPLACE	0.87	0.96	0.91	27
B-GOV	0.95	0.84	0.89	25
I-GOV	0.99	0.86	0.92	91
B-HP	1.00	0.96	0.98	24
I-HP	1.00	0.86	0.92	84
B-MKT	0.90	1.00	0.95	18
I-MKT	0.82	1.00	0.90	14
B-MON	1.00	1.00	1.00	2
I-MON	1.00	1.00	1.00	1
B-NAT	0.91	0.91	0.91	11
I-NAT	1.00	0.50	0.67	12
B-OTHER	1.00	1.00	1.00	3
I-OTHER	1.00	1.00	1.00	6
B-RCT	0.89	0.93	0.91	44
I-RCT	0.86	0.96	0.91	85
B-RES	1.00	1.00	1.00	30
I-RES	1.00	0.89	0.94	89
B-ROAD	0.87	0.92	0.90	155
I-ROAD	0.79	0.98	0.87	140
B-RP	0.98	1.00	0.99	51
I-RP	0.98	0.97	0.98	108
B-RT	0.94	0.98	0.96	129
I-RT	0.95	0.96	0.96	492
B-STORE	0.82	0.96	0.88	24
I-STORE	0.94	0.98	0.96	87
B-TRAN	0.92	0.90	0.91	39
I-TRAN	0.99	0.88	0.93	106
micro avg	0.92	0.93	0.92	2895
macro avg	0.93	0.93	0.92	2895
weighted avg	0.92	0.93	0.92	2895

รูปที่ 24 แสดงตัวอย่างคำสั่งและผลลัพธ์ที่ได้จากการฝึกฝนแบบจำลอง GRU

4.2.3 การสร้างแบบจำลองเพื่อรู้จำภูมินามด้วย BERT

ในปี 2021 มีการฝึกฝนแบบจำลอง BERT บนชุดข้อมูลภาษาไทยขนาดกว่า 78.5 GB จากแหล่งข้อมูลต่างๆโดยใช้แบบจำลองชื่อ RoBERTa ในการฝึกฝนแบบจำลอง (L. Lowphansirikul et al., 2021) BERT เป็นการใช้การถ่ายโอนความรู้ และ กลไกของ Attention เพื่อเรียนรู้ความสัมพันธ์ระหว่างคำ หรือ ระหว่างวลี (sub-words) ในข้อความ โดยทั่วไประบบการถ่ายโอนความรู้แบ่งออกเป็น 2 ส่วนที่สำคัญ คือ ส่วนเข้ารหัส (encoder) ซึ่งเป็นส่วนของการรับข้อความที่ป้อนเข้ามา และ ส่วนของการถอดรหัส (decoder) ซึ่งเป็นส่วนที่ทำนายผลลัพธ์ แต่สำหรับ BERT ถูกออกแบบให้ใช้เฉพาะส่วนที่เป็น encoder แล้วเพิ่มส่วนที่เป็นการจำแนก (classifier) เข้ามาแทน สำหรับการนำ BERT มาใช้ในงานวิจัยนี้สรุปได้ตามรูปที่ 25



รูปที่ 25 สรุปการทำงานของสถาปัตยกรรม BERT ที่ใช้งานวิจัย

จากรูปที่ 25 พบว่าในส่วนแรกที่ถัดจากส่วนการนำเข้าข้อมูล จะเป็นส่วน embeddings ซึ่งรวมกันสามส่วนคือ position segment และ token โดย position หมายถึงการทำ embeddings ในตำแหน่งของคำ segment หมายถึงการทำ embeddings ของการแบ่งส่วนของประโยคตามตัวอย่างในรูปที่ 25 เนื่องจากมีการแบ่งประโยคเป็น 2 ส่วนคือ A และ B และสุดท้าย token คือ การทำ embeddings ในส่วนของคำที่ถูกตัดแบ่งซึ่งในงานวิจัยนี้ใช้ตัวตัดคำคือ Sentence Piece และเมื่อเข้าสู่ชั้นที่เป็น BERT คือเป็นส่วนของ encoder จากนั้นจึงเข้าสู่ส่วนของการ classifier และให้ผลลัพธ์ออกมาเป็น tag แบบ IOB-tags โดยแบบจำลองถูกฝึกฝนด้วยพารามิเตอร์พื้นฐานจากแบบจำลอง RoBERTa และ WangchanBERTa (L. Lowphansirikul et al., 2021; Y. Liu, 2019) แสดงรายละเอียดตามตารางที่ 5

ตารางที่ 5 แสดงรายการพารามิเตอร์ที่ใช้เป็นพื้นฐานสำหรับแบบจำลอง BERT

รายการพารามิเตอร์	**ค่าพารามิเตอร์**
Train Batch size:	16
Test Batch size:	16
Learning rate:	2e-5
Maximum length of input:	416
Number of hidden layers in Transformer encoder:	6
Number of attention heads for each attention layer:	12
Dimension of encoder layers and pooler layers:	768
Hidden layer dimension: 3072	3072
Drop out for attention probabilities: 0.1	0.1
Activation: GeLU	GeLU
Pretrained Model:	“wangchanberta-base-att-spm-uncased”
Optimizer:	Adam

4.3 การออกแบบอัลกอริทึม Topology word

คุณสมบัติอย่างหนึ่งในทางภูมิศาสตร์ คือความสัมพันธ์สภาพแวดล้อมระหว่างสถานที่ หรือ Topology เช่น ประโยคที่ว่า “ร้าน Japang เป็นร้านขนมเปิดใหม่ที่ศูนย์ประชุมสิริกิตต์” คำว่า “ที่” ในประโยคนี้เป็นการแสดงคุณสมบัติทาง Topology ระหว่าง “ร้าน Japang” และ “ศูนย์ประชุมสิริกิตต์” ในงานวิจัยนี้จะนำคุณสมบัติที่มีการอ้างอิงกันระหว่างคำเหล่านี้มาใช้เพื่อคาดการณ์ลำดับของความสัมพันธ์โดยจัดเก็บเป็น รูปแบบ dictionary บนภาษาไพธอน แสดงตัวอย่างเงื่อนไขดังรูปที่ 26

```
weight_topology_dict = {
'อยู่':3, 'อยู่ที่':3,'อยู่ใน':3, ใน':3,'ที่':3,'บน':3, 'อยู่บน':3, 'ตั้งอยู่':3, 'ที่อยู่':3,
'พิกัด':3, 'ติดกับ':2, ใกล้กับ':2, 'ถัดไป':2, 'ต่อกับ':2, 'ถัดจาก':2,ใกล้ๆ':2,'ตรงข้าม':2,
'ตรงข้ามกับ':2,'เยื้องๆ':2, 'เยื้องกับ':2,ใจกลาง':2, 'อยู่ห่างจาก':1, 'อยู่ห่าง':1,
'ห่างจาก':1, 'ห่างไป':1, 'ห่างออกไป':1,'อีกไม่ไกล':1,ในเขต':1,'พื้นที่':1,
'บริเวณ':1,'รอบๆ':1
}
```

รูปที่ 26 แสดงค่าน้ำหนักของ Topology words

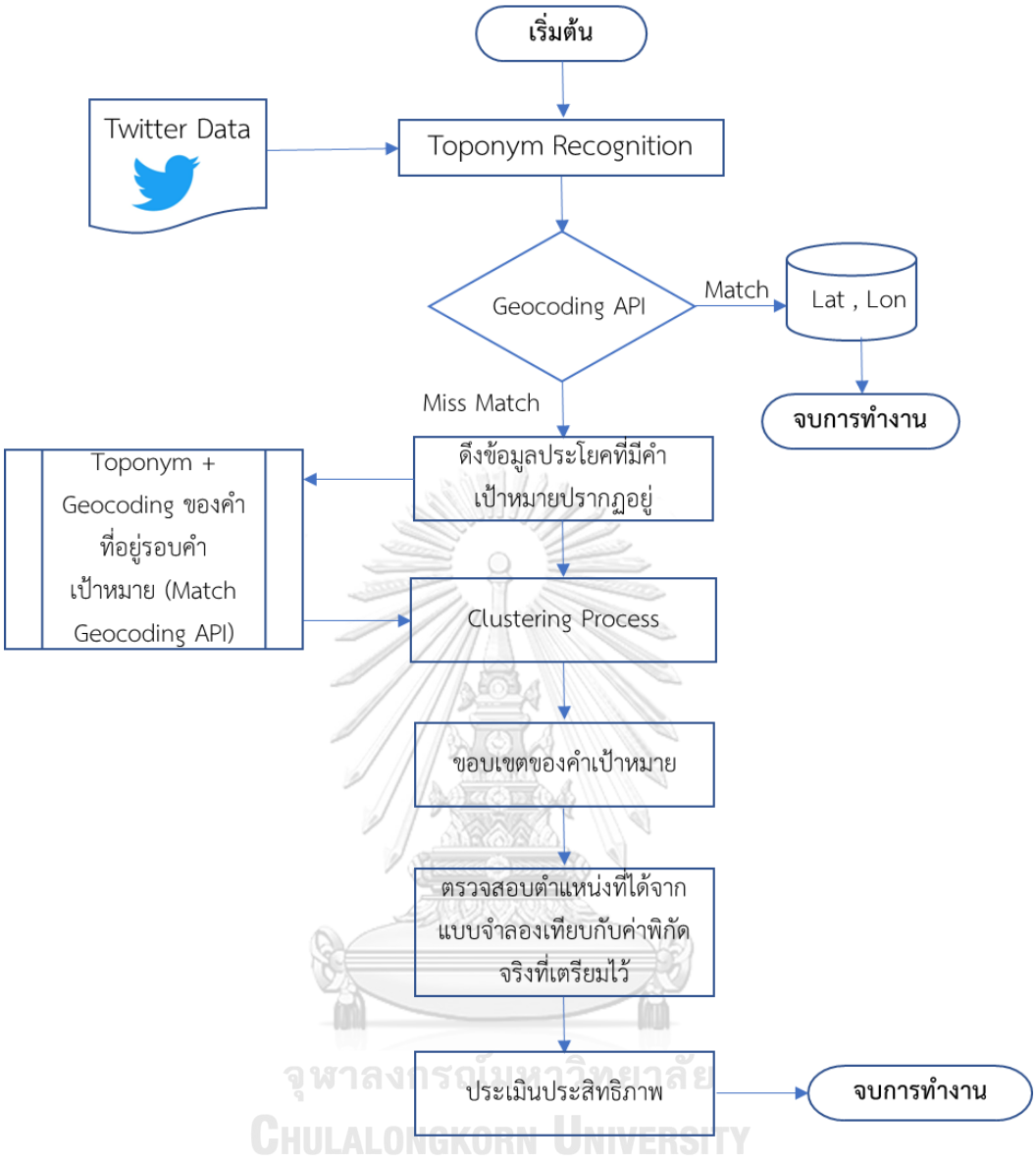
จากรูปที่ 26 คำที่มีการนำมาใช้เป็น Topology words คือคำที่ปรากฏบ่อยครั้งในชุดข้อมูลที่ศึกษาวิจัย โดยมีการให้ค่าน้ำหนักในเบื้องต้นเพื่อทดสอบอัลกอริทึมคือ 1-3 แต่ทั้งนี้ ชุดของคำและค่าน้ำหนักที่ใช้นั้นเป็นเพียงค่าเริ่มต้นซึ่งหากมีการนำไปใช้งานก็สามารถที่จะปรับเปลี่ยนให้เหมาะสมกับวัตถุประสงค์ของงานนั้นๆได้ จากตัวอย่างประโยค “ร้าน japing เปิดใหม่ที่ศูนย์ประชุมสิริกิตต์ ห่างจากสถานี MRT ไป 100 เมตร” ในประโยคข้างต้นจะพบว่ามี Topology words 2 คำ คือ “ที่” และ “ห่างจาก” เมื่อประโยคข้างต้นผ่านแบบจำลองการรู้จำภูมินาม จะได้ตัวอย่างผลลัพธ์ดังนี้ (ร้าน Japang, RES),(เปิดใหม่, O),(ที่,O),(ศูนย์ประชุมสิริกิตต์,GOV),(ห่างจาก,O),(สถานี MRT,TRAN) สังเกตว่า คำว่า “ที่” และ “ห่างจาก” เป็น tag O ซึ่งอัลกอริทึมเมื่อเจอ tag O ที่สามารถจับคู่กับสมาชิกใน dictionary ได้ ให้เปลี่ยน tag O เป็น GEO แล้วนำค่าน้ำหนักที่อยู่กับคำนั้นๆมาใส่ไว้กับคำถัดไปที่ติดกับ topology words เช่น “ศูนย์ประชุมสิริกิตต์” มีค่าน้ำหนักเป็น 3 ส่วน “สถานี MRT” มีค่าน้ำหนักเป็น 1 จากประโยคนี้จะได้ผลลัพธ์จากแบบจำลองเป็นตำแหน่งของ ศูนย์ประชุมสิริกิตต์ เนื่องจากมีค่าน้ำหนักมากที่สุด แต่หากในกรณีที่มีชื่อสถานที่ที่ค่าน้ำหนักที่มากที่สุดมีจำนวนมากกว่าหนึ่งชื่อก็ให้ใช้ทั้งหมดที่มีแล้วจึงทำการประมาณตำแหน่งจากค่าเฉลี่ยของทุกจุด แสดงรายละเอียดของอัลกอริทึมเป็นรหัสเทียมตามรูปที่ 27 และกำหนดให้ *T* คือ topology word dictionary G คือ list ของชื่อภูมิศาสตร์และ tag เช่น [[‘โรงเรียนกรุงเทพคริสเตียน’,‘ACP’], [‘อยู่ติด’,’GEO’], [‘ถนนประมวญ’,‘BSN’]] etc. และสุดท้าย R คือ list ซึ่งเป็นผลลัพธ์จากการประมวลผล

```
Requires: T, G, R
Ensure: G <> None
for i ∈ Index of G, g ∈ G :
   bound = length of G
   j = i+1
   k = i+2
   if i <= bound :
      if j < bound & g[j][index of tag] == 'GEO' :
         if g[j][index of tag] <> 'GEO' or 'O' :
            get score from T
         else :
            score = 0
         r = [g[k], score]
         append r To R
         Rank R by weight and get only max score
end for
Return R
```

รูปที่ 27 แสดงรหัสเทียมของอัลกอริทึม Topology word

4.4 การเข้ารหัสภูมิศาสตร์และการประมาณตำแหน่ง

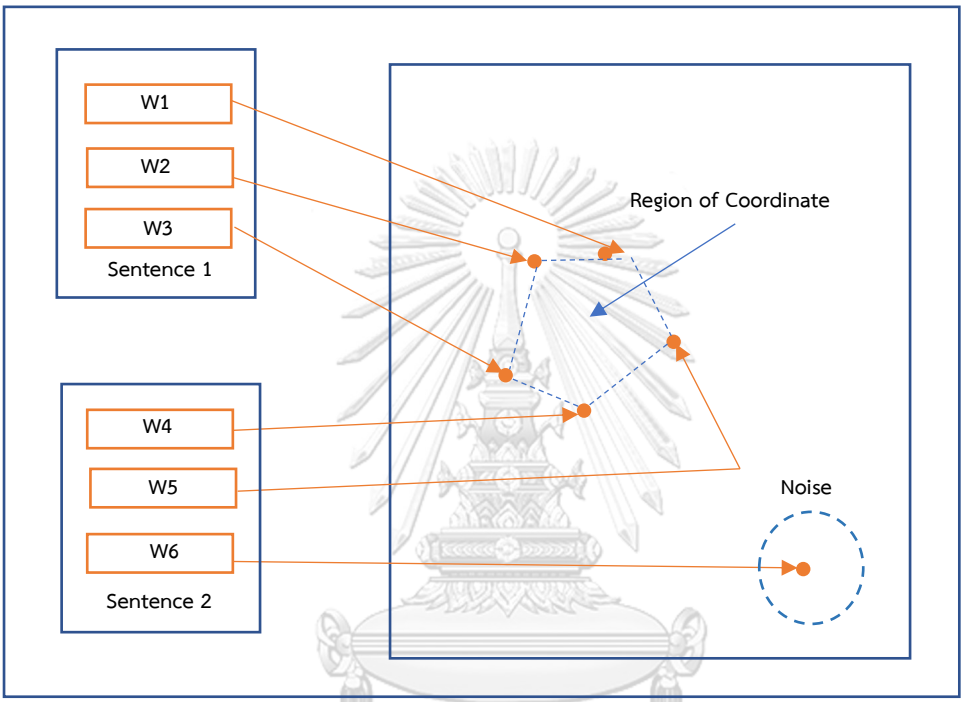
ในการเข้ารหัสภูมิศาสตร์หากชื่อสถานที่ที่ได้เป็นสถานที่ที่ไม่สามารถเข้ารหัสภูมิศาสตร์ได้จากผู้ให้บริการออนไลน์ เช่น Google API , Open Street Map (Nominatim) ซึ่งกระบวนการในการสร้างขอบเขตประมาณตำแหน่งของสถานที่สรุปเป็นผังดำเนินงานตามรูปที่ 28 ได้ ดังนี้



รูปที่ 28 แสดงผังดำเนินงานของการประมาณตำแหน่งด้วยแบบจำลองสำหรับจัดกลุ่ม

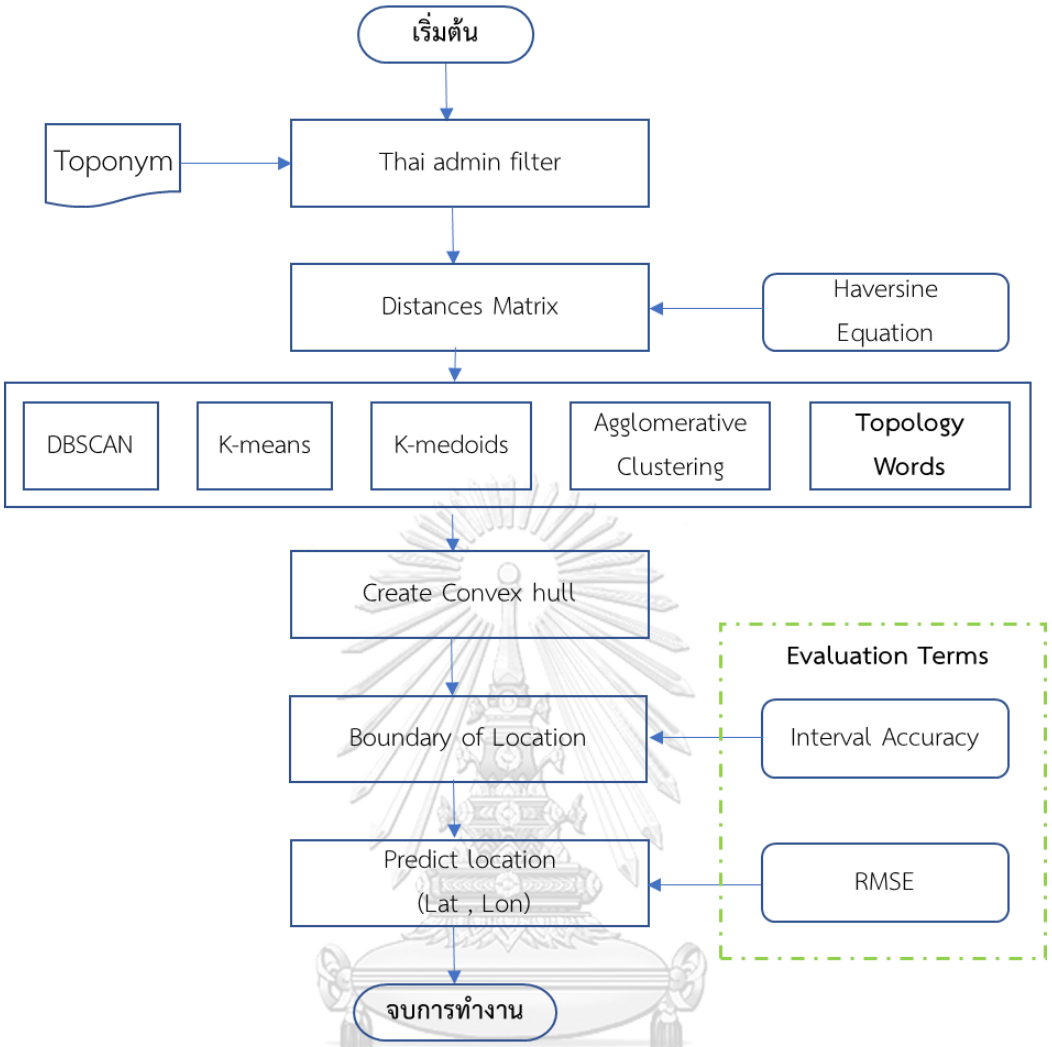
จากรูปที่ 28 เมื่อมีข้อมูลจากทวิตเตอร์เข้ามาผ่านเครื่องมือรู้จำภูมินามที่พัฒนาเตรียมไว้จากหัวข้อที่ 4.2 ในขั้นตอนต่อไปคือ การเข้ารหัสภูมิศาสตร์ ซึ่งหากสามารถจับคู่ภูมินาม กับผู้ให้บริการ เช่น Geocoding API ของ Google ให้คืนข้อมูลออกมาเป็นค่าพิกัดภูมิศาสตร์แต่หากไม่สามารถเข้ารหัสภูมิศาสตร์จากผู้ให้บริการออนไลน์ได้ให้ประมาณค่าพิกัดด้วยอัลกอริทึมสำหรับ จัดกลุ่ม (Clustering Algorithm) ได้แก่ แบบจำลอง Topology words แบบจำลอง DBSCAN แบบจำลอง K-means แบบจำลอง K-medoids และสุดท้ายแบบจำลอง Agglomerative clustering

โดยข้อมูลที่นำมาเข้าอัลกอริทึมให้ดูว่าคำเป้าหมาย (คำที่ไม่สามารถเข้ารหัสภูมิศาสตร์จาก Geocoding API) ไปปรากฏอยู่ในประโยคอื่นใดบ้าง ซึ่งในงานวิจัยนี้เตรียมข้อมูลสำหรับทดสอบแบบจำลองเป็นชุดตามภูมินาม 100 ชุด ละประมาณ 3-20 ข้อความ ทั้งหมด 430 ข้อความ หลังจากนั้นจึงนำตำแหน่งของคำที่อยู่โดยรอบมาประมวลผลเพื่อหาขอบเขตของคำเป้าหมาย แสดงตัวอย่างตามรูปที่ 29



รูปที่ 29 แสดงตัวอย่างการประมาณขอบเขตของคำเป้าหมาย

จากรูปที่ 29 คำที่อยู่รอบคำเป้าหมายในทุกประโยคหากทราบค่าพิกัดจะนำมาประมวลผลเพื่อประมาณขอบเขตของคำเป้าหมาย สรุปเป็นค่าร้อยละในแต่ละขอบเขตรัศมีระยะบัฟเฟอร์ของจุดที่ทราบค่ารวมทั้งค่า RMSE จากการเปรียบเทียบระหว่างค่าพิกัดที่ประมาณได้จากแบบจำลองและค่าพิกัดภูมิศาสตร์ของภูมินามที่ทราบค่า รายละเอียดการทำงานของอัลกอริทึม Clustering ในงานวิจัยสรุปเป็นผังดำเนินงานในรูปที่ 30 ได้ ดังนี้



รูปที่ 30 แสดงขั้นตอนในการประมาณขอบเขตและค่าพิกัดพร้อมทั้งการประเมินผล

จากรูปที่ 30 การทำ admin filter คือการเตรียมขอบเขตการปกครอง (admin) โดยในงานวิจัยนี้จะมีการเตรียมขอบเขตของปริมณฑลเป็นอันดับแรกหากมีจุดอ้างอิง toponym ซึ่งเป็นจุดอ้างอิงที่ไม่อยู่ในขอบเขตการปกครองให้เอาจุดเหล่านั้นออกไปก่อน หลังจากนั้นจึงพิจารณาขอบเขตการปกครองในข้อความต่อ นำมาเรียงลำดับหากขอบเขตการปกครองไหนที่มีการพูดถึงมากที่สุดให้นำ admin ตัวนั้นมาใช้งานเป็นตัวกรองต่อแต่หากมี rank ที่เท่ากันหลาย admin ให้นำ admin เหล่านั้นออกไปจากการพิจารณา โดยสมการที่นำมาใช้สร้างเมทริกซ์ของระยะทาง (Distance matrix) คือ สมการ Haversine โดยเมื่อได้เมทริกซ์ของระยะทางแล้วจึงนำไปใช้เป็นหนึ่งในพารามิเตอร์เพื่อเข้าแบบจำลองการจัดกลุ่ม ซึ่งในขั้นตอนนี้จะมีเพื่อกรองเอาข้อมูลตำแหน่งที่ไม่ควรนำมาคิดคำนวณในการประมาณสถานที่ออกไป หลังจากนั้นจึงสร้างขอบเขตของตำแหน่งที่จะ

ประมาณตำแหน่งออกมาแล้วจึงใช้ค่าเฉลี่ยหรือจุดกึ่งกลางของพื้นที่รูปปิดออกมาเป็นตัวแทนโดยแสดงตำแหน่งออกมาเป็นพิกัดภูมิศาสตร์ (latitude, longitude) และประเมินประสิทธิภาพของแบบจำลองด้วยการเทียบระยะที่ต่างกันระหว่างตำแหน่งที่ได้จากการประมาณของแบบจำลองและตำแหน่งจริงของสถานที่นั้น รวมทั้งจัดกลุ่มเป็นช่วงชั้นตามช่วงชั้นที่เป็นค่าควอไทล์ของระยะทางที่แบบจำลองให้ไม่ตรงกับตำแหน่งจริง ออกเป็น 4 ช่วงชั้น

4.4.1 การประมาณตำแหน่งจากอัลกอริทึมการจัดกลุ่ม (Clustering algorithm)

ในหัวข้อนี้อัลกอริทึมการจัดกลุ่มนำมาใช้งานเพื่อเป็นการกรองเอาตำแหน่งที่ไม่ควรจะเป็นขอบเขตของสถานที่เป้าหมายออกหลังจากนั้นจึงสร้างพื้นที่รูปปิดจาก convex hull แล้วนำค่าเฉลี่ยหรือจุดศูนย์กลางของพื้นที่รูปปิดมาเป็นตำแหน่งที่ประมาณ ซึ่งมีการตั้งค่าพารามิเตอร์ในแต่ละแบบจำลองตามตารางที่ 6

ตารางที่ 6 แสดงรายการพารามิเตอร์ที่ใช้เป็นพื้นฐานสำหรับแบบจำลอง BERT

แบบจำลอง	การตั้งค่าพารามิเตอร์
DBSCAN	Minpts: $\ln(n)$ Eps: 0.001 Algorithm: ball_tree Metric: haversine
K-means	N_clusters: 2 Random_state: 0 Metric: haversine
K-medoids	N_clusters: 2 Random_state: 0 Metric: haversine
Agglomerative Clustering	N_clusters: 2 Affinity: hevrsine

4.5 การประเมินประสิทธิภาพการประมาณตำแหน่งสถานที่

เมื่อได้ตำแหน่งของสถานที่ที่ประมาณขึ้นจากอัลกอริทึม 2 กลุ่มใหญ่ คือ Topology word ตามหัวข้อที่ 4.3 ซึ่งเป็นอัลกอริทึมที่งานวิจัยนี้ออกแบบและพัฒนาขึ้นมา และอัลกอริทึมแบบจัดกลุ่ม ตามหัวข้อที่ 4.4.1 ได้แก่ DBSCAN K-means K-medoids และ Agglomerative clustering หลังจากนั้นจึงคำนวณค่าต่างระยะทาง (Error) ระหว่างตำแหน่งของสถานที่นั้นๆที่เตรียมไว้เป็นตำแหน่งอ้างอิงซึ่งได้ค่าพิกัดอ้างอิงมาจาก Google geocoding API กับตำแหน่งที่ประมาณได้จากอัลกอริทึม เมื่อได้ค่า Error จากทุกจุดที่ทำการทดลองแล้วจึงนำข้อมูลทั้งหมดที่ได้มาคำนวณหาค่ารากที่สองของค่าเฉลี่ยผิดพลาดกำลังสอง (Root mean square error : RMSE) เนื่องจากในงานวิจัยนี้ชุดข้อมูลทดสอบอัลกอริทึมมีจำนวนที่ไม่ได้ใหญ่มาก คือ 100 ชุด ผลที่ได้จาก RMSE จะมีอิทธิพลมาจากค่าผิดปกติ (Outlier) ที่ค่อนข้างมาก ทำให้ในการทดสอบอัลกอริทึมนั้นสามารถสังเกตเห็นรูปแบบหรือขั้นตอนที่อาจจะส่งผลให้อัลกอริทึมที่เลือกมาหรือสร้างขึ้นมานั้นทำงานได้ไม่ดีเนื่องจากสาเหตุอะไร โดยมีการนำค่าเบี่ยงเบนมาตรฐาน (Standard deviation : S.D.) และ ค่าเฉลี่ยผลต่างสัมบูรณ์ (Mean absolute error : MAE) รวมถึงมีการสรุปผลออกมาเป็นรูปแบบกราฟแบบกล่อง (Box plot) และช่วงชั้นของ RMSE ออกมาเป็น 4 ช่วงควอไทล์

บทที่ 5
ผลการศึกษา

ผลการศึกษาวิจัยแบ่งแบ่งออกเป็นสามส่วนที่สำคัญคือ 5.1 การประเมินประสิทธิภาพของแบบจำลองที่นำมาใช้ในการสร้างเครื่องมือรู้จำภูมินาม ซึ่งเป็นการเปรียบเทียบค่าความถูกต้องโดยรวม (F1-Phrase) ของแบบจำลองทั้ง CRF, LSTM, GRU และ BERT หัวข้อ 5.2 แสดงผลการประมาณตำแหน่งของข้อมูลสถานที่ ว่าให้ความถูกต้องในแบบช่วงชั้น และ ค่าความคลาดเคลื่อนเมื่อเทียบกับค่าพิกัดจริงของภูมินามเป็น RMSE และสุดท้ายคือหัวข้อ 5.3 ซึ่งเป็นการแสดงกรณีตัวอย่างในการประมาณตำแหน่งของข้อมูลภูมินาม

5.1 การแสดงผลประสิทธิภาพของแบบจำลองต่างๆที่นำมาสร้างแบบจำลองรู้จำภูมินาม

หัวข้อนี้จัดเป็นส่วนแรกในการแสดงผลของแบบจำลองนี้เลือกมาใช้ ซึ่งถือว่าเป็นแบบจำลองที่ได้รับความนิยมและเป็นฐานสำคัญในการศึกษาเปรียบเทียบในการสกัดข้อมูลที่เป็นชื่อเฉพาะโดยในงานวิจัยนี้มุ่งเน้นที่จะสกัดข้อมูลภูมินามออกมาจากข้อความบนสื่อสังคมออนไลน์อย่างทวิตเตอร์ แบบจำลองที่นำมาใช้ในงานวิจัยนี้แบ่งออกเป็น 3 กลุ่มหลัก ดังนี้ 1) แบบจำลองที่เป็นการเรียนรู้ของเครื่องจักรแบบมีลำดับต่อเนื่อง (sequence machine learning model) ซึ่งในงานวิจัยนี้นำ CRF มาใช้ 2) แบบจำลองที่เป็นโครงข่ายประสาทเทียมแบบวกกลับ (Recurrent Neural network) ในงานวิจัยนี้คือ LSTM และ GRU 3) แบบจำลองที่เป็นการถ่ายโอนความรู้ (Transformer model) ในงานวิจัยนี้นำ WangchanBERTa ที่ผ่านการฝึกฝนด้วยคลังข้อมูลภาษาไทยขนาดใหญ่ มาเป็นค่าเริ่มต้นในการฝึกฝนแบบจำลอง และสุดท้ายคือ 4) แสดงผลเปรียบเทียบระหว่างแบบจำลองที่ให้ค่าความถูกต้องโดยรวมที่ดีที่สุดแต่ละประเภท

5.1.1 ผลจากการฝึกฝนแบบจำลอง CRF

ในการฝึกฝนแบบจำลอง CRF มีการกำหนดคุณลักษณะที่สำคัญทั้งหมด 4 ชุดที่ สำคัญคือ 1) คำนำหน้าและคำต่อท้าย (prefix and suffix) , 2) รูปร่างของคำและหน้าที่ของคำ (word shape and pos-tag) , 3) ส่วนใดส่วนหนึ่งของคำในอักขรานุกรมภูมิศาสตร์ และ 4) การจับคู่กับอักขรานุกรมภูมิศาสตร์แบบครบถ้วนทั้งคำ สำหรับการพิจารณาคุณลักษณะในด้านลำดับของคำแบบ n ลำดับ หรือ n-gram ซึ่งในงานวิจัยนี้ใช้แบบ 3 แกรม และ 4 แกรม ตามลำดับ ในส่วนของฟังก์ชันการหาค่าเหมาะสม (optimizer) ในงานวิจัยนี้ใช้แบบ LBFGS และมีการทดลองเปรียบเทียบเพิ่มเติมกับ L2SGD (L2 regularization stochastics gradient descent) สำหรับเครื่องมือในการตัดคำใช้เป็น คลังคำสั่ง Attacut แสดงผลลัพธ์ตามตารางที่ 7

ตารางที่ 7 แสดงผลประสิทธิภาพของแบบจำลอง CRF จากคุณลักษณะที่ต่างกัน

แบบจำลอง	แกรม	optimizer	1)	2)	3)	4)	F1-phrase
1	3	LBFGS					0.851
2	3	LBFGS	/				0.849
3	3	LBFGS	/	/			0.854
4	**3**	**LBFGS**	**/**	**/**	**/**		**0.863**
5	3	LBFGS	/	/	/	/	0.862
6	4	LBFGS					0.862
7	4	LBFGS	/				0.86
8	4	LBFGS	/	/			0.857
9	4	LBFGS	/	/	/		0.857
10	4	LBFGS	/	/	/	/	0.861
11	4	L2GSD	/	/			0.552
12	4	L2GSD	/	/	/		0.688
13	4	L2GSD	/	/	/	/	0.697

จากตารางที่ 7 ผลที่ได้จากแบบจำลองที่ 4 ซึ่งมีการใช้คุณลักษณะคือ 3 แกรม และคุณลักษณะในหัวข้อที่ 1) – 3) คือคำนำหน้า คำต่อท้าย รูปร่างของคำและหน้าที่ของคำ และสุดท้ายคือ ส่วนใดส่วนหนึ่งของคำในอักขรานุกรมภูมิศาสตร์ ซึ่งให้ค่า F1-phrase สูงที่สุดคือ 0.863 รองลงมาคือ แบบจำลองที่ 5 ซึ่งมีคุณลักษณะในหัวข้อที่ 5) คือ การจับคู่กับอักขรานุกรมภูมิศาสตร์แบบครบถ้วนทั้งคำ มีค่า F1-phrase ที่ 0.862 ซึ่งใกล้เคียงกันมาก จากผลอาจกล่าวได้ว่าสำหรับข้อมูลในงานวิจัยนี้การใช้ 3 แกรม พร้อมด้วยการใช้คุณลักษณะในการเทียบเคียงส่วนใดส่วนหนึ่งของชื่อเฉพาะในอักขรานุกรมภูมิศาสตร์ก็ได้ผลไม่ต่างจาก 4 แกรมมากนัก และเห็นได้ชัดว่าการใช้ฟังก์ชันหาค่าเหมาะสมแบบ L2SGD ให้ผลที่ต่างกับ LBFGS มากโดย LBFGS คำนวณจากขนาดของข้อมูลทั้งหมดในการปรับค่าแต่ละรอบ ซึ่ง L2GSD ใช้การคำนวณจากชุดข้อมูลย่อย (mini-batch) ดังนั้นหากต้องการนำแบบจำลอง CRF ไปใช้เป็นแบบจำลองการรู้จำภูมินาม จึงควรเลือกแบบจำลองที่ 4 ไปใช้

5.1.2 ผลจากการฝึกฝนแบบจำลองที่เป็นโครงข่ายประสาทเทียมแบบวกกลับ

ในงานวิจัยนี้นำเอาแบบจำลองที่เป็นโครงข่ายประสาทเทียมวกกลับ 2 แบบมาใช้ คือ LSTM และ GRU ซึ่งในการปรับค่าพารามิเตอร์สำหรับ LSTM มีการอ้างอิงมาจากสถาปัตยกรรมของ Thattinaphanich and Prom-On (2019) และสำหรับ GRU อ้างอิงจากสถาปัตยกรรมของ Gridach and Haddad (2017) ประกอบไปด้วยส่วนสำคัญตามตารางที่ 8 ดังนี้

ตารางที่ 8 แสดงรายละเอียดของพารามิเตอร์ที่ใช้ในการฝึกฝนแบบจำลองแต่ละชุด

ชุดที่	**Word embedding**	**Learning rate**	**optimizer**	**Dropout embedding**	**Dropout layer**	**dense**	**epochs**
1	Thai2vec	0.001	Adam	0.3	0.5	256	50
2	Thai2vec	0.001	Adam	0.3	0.3	128	50

จากตารางที่ 8 ชุดพารามิเตอร์ที่ 1 คือสภาพแวดล้อมของการฝึกฝนแบบจำลอง LSTM ส่วนชุดพารามิเตอร์ที่ 2 คือสภาพแวดล้อมของการฝึกฝนแบบจำลอง GRU โดยผลจากการฝึกฝนแบบจำลองแสดงตามตารางที่ 9

ตารางที่ 9 แสดงผลประสิทธิภาพของแบบจำลอง LSTM และ GRU

ลำดับที่	**แบบจำลอง**	**ชุดพารามิเตอร์**	**F1-phrase**
1	LSTM	ชุดที่ 1	0.60
2	Bi-LSTM	ชุดที่ 1	0.809
3	**Bi-LSTM-CRF**	**ชุดที่ 1**	**0.859**
4	GRU	ชุดที่ 2	0.66
5	Bi-GRU	ชุดที่ 2	0.812
6	Bi-GRU-CRF	ชุดที่ 2	0.831

จากตารางที่ 9 แบบจำลองที่ให้ค่าความถูกต้องโดยรวม (F1-Phrase) สูงที่สุดคือ Bi-LSTM-CRF ใช้ชุดพารามิเตอร์ที่ 1 ในการฝึกฝนแบบจำลอง ซึ่งเมื่อพิจารณาจากค่า F1-Phrase จะพบว่าหากเปรียบเทียบระหว่าง LSTM และ GRU พบว่า GRU ให้ค่า F1-Phrase ที่มากกว่าโดยตลอดจนกระทั่งมีการเพิ่มชั้นของ CRF เข้าไป LSTM จึงจะให้ค่า F1-Phrase ที่สูงกว่า

5.1.3 ผลจากการฝึกฝนแบบจำลองที่เป็นการถ่ายโอนความรู้ (Transformer Model)

ในงานวิจัยนี้นำแบบจำลองการถ่ายโอนความรู้มาใช้งาน โดยเลือกใช้แบบจำลองของ WangchanBERTa มาเป็นตัวเริ่มต้นในการฝึกฝนแบบจำลอง ซึ่งสภาพแวดล้อมของแบบจำลองประกอบไปด้วย seed=9, learning rate=0.00002, weight decay=0.01, epoch=10 แสดงผลจากการฝึกฝนแบบจำลองตามตารางที่ 10 ดังนี้

ตารางที่ 10 แสดงผลการฝึกฝนแบบจำลองจากสถาปัตยกรรม BERT

Model	Lr_schedular_type	warmup_ratio	F1_Phrase
1	-	-	0.918
2	linear	0.1	0.917
3	polynomial	0.05	0.917
4	**cosine_with_restarts**	**0.05**	**0.919**
5	constant_with_warmup	0.05	0.908

จากตารางที่ 10 พบว่าค่า F1-Phrase ที่ดีที่สุดคือ 0.919 ซึ่งใช้ learning rate เป็น cosine_with_restarts สำหรับตัว lr_schedular_type หรือ learning rate schedular type มีการทดลองใช้ทั้งหมด 4 แบบ ร่วมกับการใช้ warm up ratio (Mishra & Sarawadekar, 2019) และเมื่อพิจารณาจากค่า F1-Phrase ในตารางที่ 10 จะพบว่ามีค่าอยู่ในระดับที่ใกล้เคียงกันมาก

5.1.4 แสดงผลการศึกษาเปรียบเทียบระหว่างแบบจำลองแต่ละประเภท

โดยจากผลการทดลองใน 3 หัวข้อที่ผ่านมาเมื่อนำผลที่ดีที่สุดจากแบบจำลองในแต่ละประเภทมารวมทั้งแบบจำลอง CRF ที่มีการใช้ฟังก์ชันคุณลักษณะตามแบบของ PyThaiNLP และ BERT ที่ใช้แบบจำลองของ monsoon และ bert-base-multilingual นำมาแสดงผลเพื่อเปรียบเทียบตามตารางที่ 11

ตารางที่ 11 แสดงผลการเปรียบเทียบระหว่างแบบจำลองแต่ละประเภท

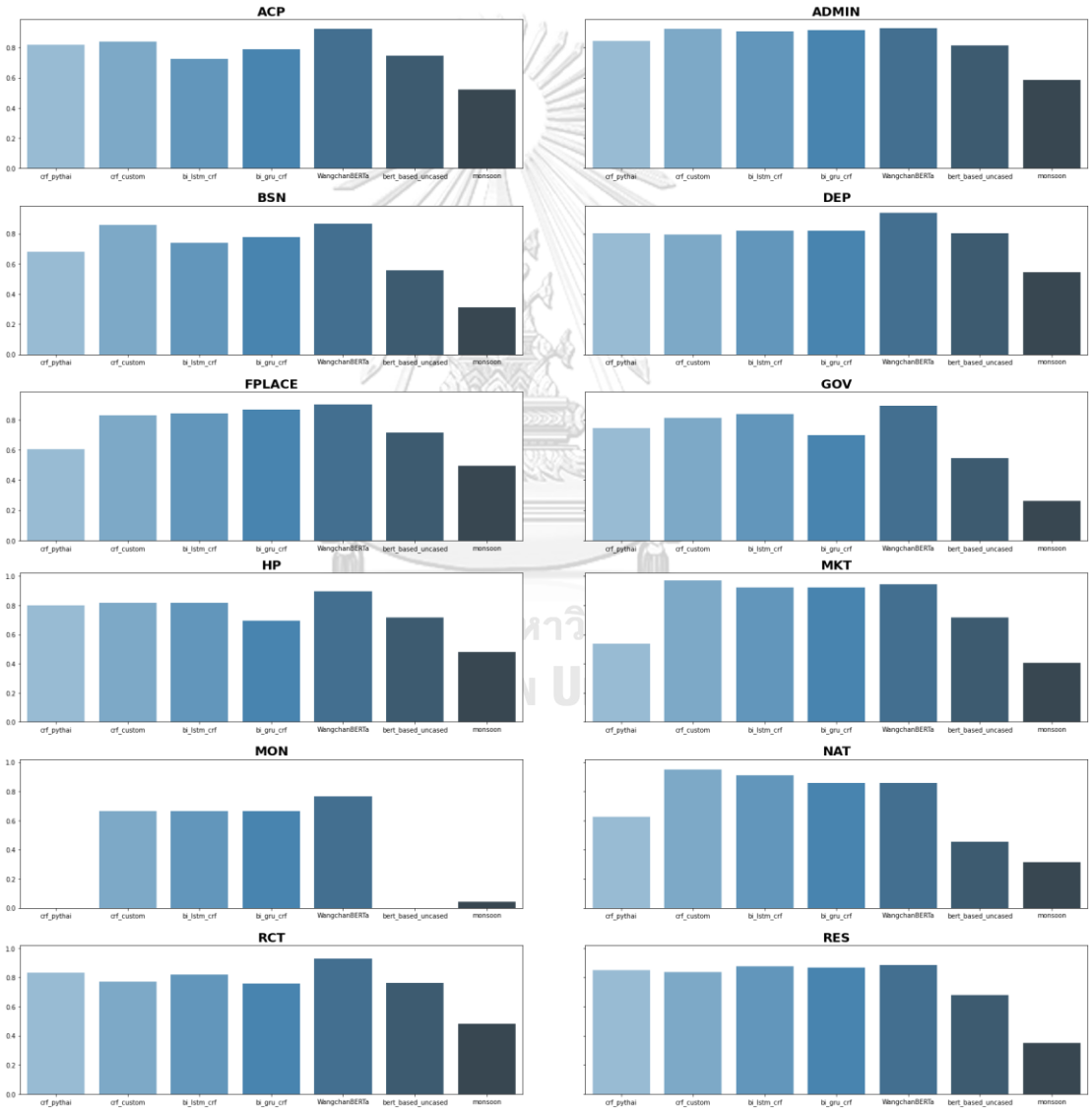
ลำดับที่	แบบจำลอง	F1-phrase
1	CRF PyThaiNLP	0.797
2	CRF Custom feature	0.862
3	Bi-LSTM-CRF	0.859
4	Bi-GRU-CRF	0.842
5	**BERT (WangchanBERTa)**	**0.919**
6	BERT (bert-base-multilingual)	0.747
7	BERT (Monsoon)	0.483

จากตารางที่ 11 พบว่าแบบจำลองที่ให้ค่าความถูกต้องโดยรวม F1-Phrase ดีที่สุดคือ BERT ซึ่งให้ค่า F1-Phrase ที่ 0.919 จากแบบจำลองอื่นๆที่แสดงผลเปรียบเทียบ และพบว่าแบบจำลอง CRF ที่มีการปรับใช้คุณสมบัติเฉพาะสำหรับการจำแนกประเภทสถานที่ให้ผลที่ดีกว่าแบบจำลองที่เป็นโครงข่ายประสาทเทียมทั้งแบบจำลอง Bi-LSTM-CRF และ Bi-GRU-CRF ดังนั้นจากผลการทดลองในส่วนของการสร้างแบบจำลองรู้จำภูมินาม การนำแบบจำลองที่ฝึกฝนเพิ่มเติมจาก WnagchanBERTa มาเป็นแบบจำลองที่ใช้ในงานวิจัยนี้จึงมีความเหมาะสมและแสดงผลของค่า F1 ในแต่ละประเภท ตามตารางที่ 12

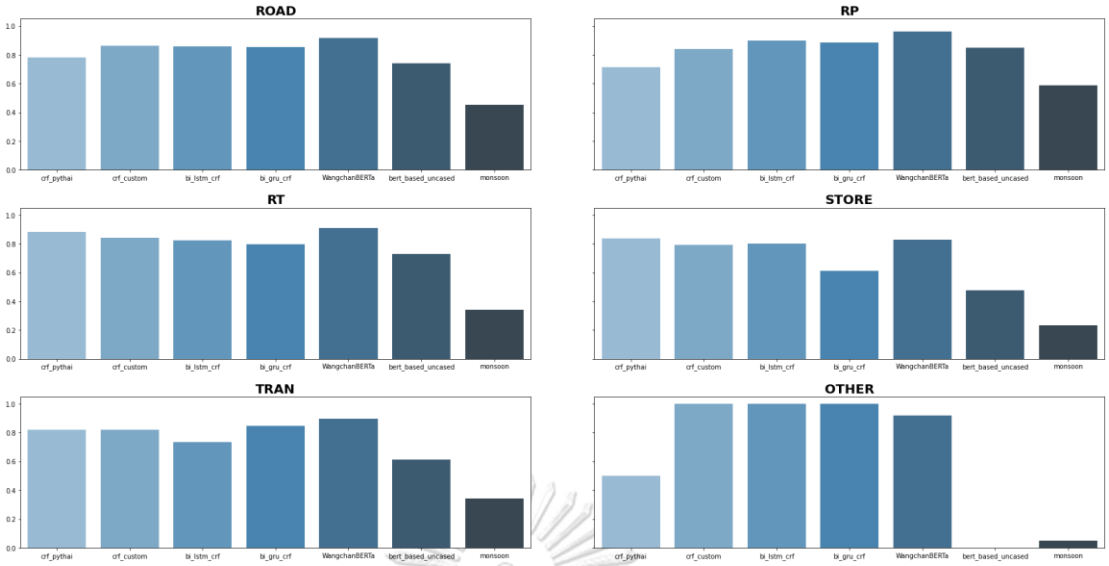
ตารางที่ 12 ตารางแสดงการเปรียบเทียบค่าความถูกต้องตามชนิดของภูมินามจาก WangchanBERTa

TAG	ความหมาย	Precision	Recall	F1-Phrase
ACP	Academic place	0.925	0.917	0.921
ADMIN	Admin boundary	0.922	0.934	0.928
BSN	Office building	0.829	0.899	0.863
DEP	Department store	0.930	0.950	0.940
FPLACE	Location outside thailand	0.883	0.925	0.904
GOV	Government office	0.903	0.873	0.888
HP	Healthcare place	0.892	0.933	0.912
MKT	market	0.941	0.938	0.939
MON	Monument or roundabout	0.842	0.744	0.790
NAT	Natural place	0.817	0.902	0.857
RCT	Recreations, parks, amusement, stadiums	0.909	0.964	0.936
RES	residential	0.842	0.909	0.874
ROAD	Highway, road, alley	0.907	0.927	0.917
RP	Regional place	0.957	0.967	0.962
RT	restaurant	0.889	0.931	0.909
STORE	Store, shops, local shops,	0.819	0.824	0.821
TRAN	Mass transit, train station, bus station, piers, port etc.	0.871	0.914	0.892
OTHER	Other places	0.933	0.875	0.903

จากตารางที่ 12 พบว่าผลลัพธ์ที่ได้จากการฝึกฝนแบบจำลอง ชนิดของภูมินามที่ให้ค่า F1 สูงที่สุด คือ RP หรือสถานที่สำคัญทางศาสนา ให้ค่า F1 ที่ 0.962 และชนิดของภูมินามที่มีค่า F1 น้อยที่สุด คือ MON ซึ่งเป็นชนิดของอนุสาวรีย์ วงเวียน หอนาฬิกา ฯลฯ ให้ค่า F1 ที่ 0.790 จากลักษณะของภูมินามทั้ง 2 ประเภท พบว่าชนิดของภูมินามที่ให้ค่า F1 ใกล้เคียงกับ DEP ได้แก่ ACP, ADMIN, MKT และ RCT โดยผลการเปรียบเทียบค่าความถูกต้องโดยรวมแต่ละประเภทของภูมินามทั้ง 7 แบบจำลองในตารางที่ 12 แสดงตามรูปที่ 31 และ 32



รูปที่ 31 แสดงกราฟแท่งเปรียบเทียบค่าความถูกต้องโดยรวมแต่ละแบบจำลองแยกตามชนิดของภูมินาม ACP - RES



รูปที่ 32 แสดงกราฟแท่งเปรียบเทียบค่าความถูกต้องโดยรวมแต่ละแบบจำลองแยกตามชนิดของภูมินาม ROAD – OTHER

จากรูปที่ 31 และ 32 แสดงผลเปรียบเทียบค่าความถูกต้องโดยรวม (F1-Phrase) ระหว่างแบบจำลองรู้จำภูมินามโดยแยกตามชนิดของภูมินามเรียงตามลำดับจากตารางที่ 5.6 และเรียงลำดับแบบจำลองจากซ้ายไปขวา ได้แก่ แบบจำลอง CRF PythaiNLP แบบจำลอง CRF custom feature แบบจำลอง Bi-LSTM-CRF แบบจำลอง Bi-GRU_CRF แบบจำลอง WangchanBERTa แบบจำลอง bert_based_uncased และ แบบจำลอง monsoon

ซึ่งพบว่าแบบจำลองที่สร้างจาก WangchanBERTa ให้ค่าความถูกต้องโดยรวมสูงที่สุดในแต่ละชนิดของภูมินาม และเป็นที่น่าสังเกตว่าบางชนิดของภูมินาม ได้แก่ MON ซึ่งหมายถึง อนุสาวรีย์ วงเวียน หอนาฬิกา และ OTHER ซึ่งเป็นชนิดของภูมินามที่ไม่สามารถจัดไว้ในหมวดหมู่ใดได้ แบบจำลองที่สร้างจาก CRF โดยอ้างอิงฟังก์ชันคุณลักษณะจาก PythaiNLP แบบจำลอง BERT ที่ใช้ค่าจากการฝึกฝนก่อนหน้า (pretrained) จาก bert_base_uncased และ monsoon นั้นให้ค่าความถูกต้องโดยรวมที่น้อยมาก หรือ ตอบไม่ถูกเลย ซึ่งแบบจำลอง CRF ที่มีการสร้างฟังก์ชันคุณลักษณะสำหรับสกัดภูมินาม และโครงข่ายประสาทเทียมทั้ง Bi-LSTM-CRF และ Bi-GRU-CRF ต่างรับมือได้ดีกว่า 2 แบบจำลองข้างต้นที่กล่าวมา และมีบางชนิดของภูมินามที่หลายแบบจำลองให้ค่าความถูกต้องโดยรวมที่ค่อนข้างสูง ได้แก่ ADMIN ซึ่งหมายถึง ขอบเขตการปกครอง MKT ซึ่งหมายถึงตลาด และ RES ซึ่งหมายถึงหมู่บ้าน ที่อยู่อาศัย

5.2 ผลการประมาณตำแหน่งของภูมินามจากคุณสมบัติการอ้างอิงสภาพแวดล้อม (Topology) และการเรียนรู้ของเครื่องแบบจัดกลุ่ม (Clustering)

ผลจากการประมาณตำแหน่งของภูมินามจะเทียบเคียงกับตำแหน่งที่อ้างอิงได้จาก google geocoding โดยใช้ค่า root mean square error เป็นตัวเปรียบเทียบมีหน่วยเป็นกิโลเมตร และการเปรียบเทียบผลแบบช่วงชั้นทั้งหมด 4 ช่วงชั้น โดยประมวลผลจากชุดข้อมูล 100 ชุด ซึ่งแต่ละชุดประกอบด้วย 5 - 20 ข้อความ แบ่งเป็น 2 หัวข้อย่อยคือ 5.2.1 ผลการศึกษาในการประมาณตำแหน่งของสถานที่ และ 5.2.2 เป็นการแสดงตัวอย่างและอภิปรายผลจากการทดลอง

5.2.1 ผลการศึกษาเปรียบเทียบแบบจำลองที่ใช้ในการประมาณตำแหน่งของภูมินาม

ในหัวข้อนี้จะแสดงการเปรียบเทียบผลการทดลองในแต่ละแบบจำลองซึ่งประกอบไปด้วย การเปรียบเทียบแบบช่วงชั้นทั้งหมด 4 ช่วงชั้นตามค่าควอไทล์ 0.25, 0.5, 0.75, > 0.75 ได้แก่ภายใน 0.145 0.515 1.71 และมากกว่า 1.71 กิโลเมตร ตามลำดับ ซึ่งได้ให้เหตุผลไว้ในบทที่ 3 ของงานวิจัยฉบับนี้ รวมถึงแสดงค่า RMSE ที่ได้จากค่าคลาดเคลื่อนของตำแหน่งที่ประมาณขึ้นจากแบบจำลองและตำแหน่งที่ได้จากฐานข้อมูลออนไลน์ คือ google geocoding API แสดงผลตามตารางที่ 13

ตารางที่ 13 แสดงตารางการเปรียบเทียบความแม่นยำในแต่ละช่วงชั้นและค่า RMSE ของแต่ละแบบจำลอง

แบบจำลอง	**รัศมี 0.145 กม.**	**รัศมี 0.515 กม.**	**รัศมี 1.71 กม.**	**> รัศมี 1.71 กม.**
Topology words	**38**	**29**	**22**	**11**
DBSCAN	16	29	26	29
K-means	25	27	22	26
K-medoids	22	16	26	36
Agglomerative clustering	26	25	26	23

จากตารางที่ 13 พบว่าแบบจำลอง topology words ให้ความแม่นยำในระดับรัศมี 0.145 กิโลเมตรมากที่สุด คืออยู่ภายในรัศมี 38 จุด เช่นเดียวกับอัลกอริทึม Agglomerative clustering ที่ให้ความแม่นยำในรัศมี 0.154 กิโลเมตรรองลงมาคือ 26 จุด สำหรับแบบจำลองที่มีค่า error และ accuracy น้อยที่สุดคือ K-medoids ซึ่งมีค่า RMSE 3.883 และ accuracy ในระดับ 18.73 กม. คือ 36 จุด

สำหรับข้อมูลตัวอย่างที่ใช้ในการทดลองจำนวน 100 ชุด แสดงรายละเอียดในการทดลองดังนี้ โดย คอลัมน์จำนวนภูมินาม หมายถึงภูมินามที่สกัดออกมาได้จากชุดข้อความทดสอบ จำนวนคำบ่งชี้สถานที่ หมายถึง คำบ่งชี้สถานที่ซึ่งสกัดออกมาได้จากข้อความ จำนวนตัวอักษร หมายถึงจำนวนตัวอักษรทั้งหมดในชุดข้อมูลแต่ละชุด และ 5 คอลัมน์สุดท้ายคือ ค่าคลาดเคลื่อนระหว่างพิกัดภูมิศาสตร์ที่ทำนายได้จากแบบจำลอง และพิกัดภูมิศาสตร์จริงของภูมินาม แสดงตามตารางที่ 14

ตารางที่ 14 แสดงรายละเอียดของผลการทดลองที่ได้ในแต่ละชุดข้อมูลภูมินามทดสอบ

ภูมินาม	จำนวนภูมินาม	จำนวนคำบ่งชี้สถานที่	จำนวนตัวอักษร	Topolgy Words	DBSCAN	K-means	K-medoids	Agglomerative clustering
ซัม ยอก ซัล	7	10	521	0.175	1.652	3.84	1.652	3.84
กอชุนฮวด	12	8	839	0.175	0.094	0.094	0.094	0.094
ทองหล่อ โยโคโจว	9	4	343	1.149	0.433	0.433	1.149	0.431
หน่องริมคลอง	17	5	1,414	0.077	0.145	0.044	0.145	0.145
Brunch Paradiso	9	13	1,393	**0.000**	2.255	1.499	2.255	1.499
ชฎาซาลอน	8	8	698	0.916	0.202	0.202	2.897	0.202
พิพิธภัณฑสถานแห่งชาติ ศิลป์ พีระศรี อนุสรณ์	35	39	1,468	0.113	0.651	0.651	16.11	0.651
พระราชวังพญาไท	12	20	972	0.669	0.207	1.443	0.207	0.785
กินใจ คอนเทมโพลารี	7	3	276	1.728	1.959	1.728	1.728	0.344
The Alphabet Book Café	14	3	548	0.220	1.612	1.612	2.458	0.207
หอสมุดเมืองกรุงเทพมหานคร	10	14	309	0.094	0.402	0.158	0.402	2.203
ศูนย์สร้างสรรค์งานออกแบบ	11	2	281	0.034	6.074	0.691	6.074	1.612
เยโลเฮาส์	11	5	307	1.837	1.932	0.699	2.83	0.158
แบมบินี วิลล่า	6	0	307	0.488	1.932	0.463	0.135	3.756
ศูนย์การเรียนรู้ธนาคารแห่งประเทศไทย	2	6	545	0.181	0.681	0.181	0.181	2.83
ไท้ซุ่นบาร์	5	4	351	0.005	0.101	0.252	0.101	1.023

ตารางที่ 14 แสดงรายละเอียดของผลการทดลองที่ได้ในแต่ละชุดข้อมูลภูมินามทดสอบ (ต่อ)

ภูมินาม	จำนวนภูมินาม	จำนวนคำบ่งชี้สถานที่	จำนวนตัวอักษร	Topolgy Words	DBSCAN	K-means	K-medoids	Agglomerative clustering
มีวนาคอฟฟี่บ้านสาทร	5	5	351	0.179	0.179	0.13	0.13	1.256
ร้านซานี่ส์	6	4	265	1.555	2.026	0.277	1.555	0.028
ชานพระนคร	6	2	266	2.876	2.543	6.269	0.559	0.13
Broccoli Revolution	3	2	320	2.497	2.497	1.149	3.471	1.739
The Commons Thonglor	8	10	594	0.407	0.412	0.859	0.412	6.243
สวนเบญจสิริ	7	14	490	0.567	0.394	0.394	0.585	1.149
ตลาดน้ำคลองลัดมะยม	11	9	631	2.718	3.266	4.952	4.952	0.036
พิพิธภัณฑ์ช้างเอราวัณ	2	4	370	0.833	0.833	0.833	0.833	0.394
ตลาดน้ำบางน้ำผึ้ง	10	7	714	3.385	2.456	5.451	1.056	3.686
Great harbour	13	14	1,460	0.000	1.136	0.02	1.136	0.833
พาร์เลอร์เรสเตอร์รองท์แอนด์ บาร์	14	10	949	0.091	3.275	3.275	0.432	5.451
ร้านจีฉ่อย	12	13	1,247	0.169	0.532	0.268	0.532	0.02
ร้านข้าวแกงนายอู๋	13	12	773	0.768	1.111	0.188	1.111	3.275
มูลนิธิกระจกเงา	12	4	695	0.033	1.142	1.142	1.433	0.268
กินลมชมสะพาน	7	3	457	0.026	0.052	0.107	0.052	0.188
โรงพยาบาลจุฬาภรณ์	14	7	1,003	0.078	0.314	1.209	0.599	1.142
สว่างบะหมี่ก้ามปู	6	3	883	0.348	1.289	0.348	2.123	0.107
ตั้งชุ่ยเฮง	11	6	838	0.383	0.942	2.69	0.942	1.209
หมูปลาร้าราชดำเนิน	9	8	549	0.272	2.456	5.451	1.056	0.348
วัดคลองครุ	10	10	1,280	0.554	0.489	0.032	0.489	2.69
โรงแรมโอไทเฮง	13	20	654	0.196	0.253	0.208	0.323	5.451
ร้านฟรุ๊ตฟิตส์	10	7	742	0.836	0.728	1.677	1.677	0.999
BBC:Big Black Cat	7	18	725	0.185	0.409	0.185	0.185	0.208

ตารางที่ 14 แสดงรายละเอียดของผลการทดลองที่ได้ในแต่ละชุดข้อมูลภูมินามทดสอบ (ต่อ)

ภูมินาม	จำนวนภูมินาม	จำนวนคำบ่งชี้สถานที่	จำนวนตัวอักษร	Topolgy Words	DBSCAN	K-means	K-medoids	Agglomerative clustering
อารีสัตวแพทย์	13	13	435	0.164	2.244	2.244	6.737	1.677
ชินจ่าว	16	12	569	1.078	1.078	1.078	8.361	0.645
ท่ามหาราช คอมมูนิตี้มอลล์	16	11	752	0.049	0.749	0.269	1.813	2.244
ช่างชุ่ย	10	8	433	1.255	2.791	2.791	2.791	1.078
จอมมารบู สุกี้ ชาบู	12	8	292	0.018	3.016	3.016	7.871	0.269
ก๋วยเตี๋ยวปาก หม้อนีมะห์ สาขา Jodd Fairs	12	11	476	0.736	6.989	6.989	1.888	2.791
คั่วไก่แฮปปี้แลนด์ 1999	30	8	820	0.639	1.159	1.159	13.122	3.016
สิรินารถข้าวมันไก่	13	5	527	0.224	3.019	18.73	0.569	0.918
ข้าวหมูแดงสีมรกต	11	17	758	0.188	0.667	2.397	0.667	1.159
แซมาอีลชิกดัง	11	11	698	0.095	0.483	0.002	0.048	3.019
seoga and cook	7	6	1,013	0.019	0.019	0.01	0.019	0.159
Kam's Roast	5	2	383	0.012	0.012	0.012	0.012	0.002
เฟิงฟู่	15	9	1,988	2.142	0.308	0.308	2.142	0.039
Japang	2	3	240	0.014	0.014	0.014	0.014	0.012
ช่างชุ่ย	10	8	433	1.255	2.791	2.791	2.791	1.078
จอมมารบู สุกี้ ชาบู	12	8	292	0.018	3.016	3.016	7.871	0.269
ก๋วยเตี๋ยวปาก หม้อนีมะห์ สาขา Jodd Fairs	12	11	476	0.736	6.989	6.989	1.888	2.791
คั่วไก่แฮปปี้แลนด์ 1999	30	8	820	0.639	1.159	1.159	13.122	3.016
สิรินารถข้าวมันไก่	13	5	527	0.224	3.019	18.73	0.569	0.918
ข้าวหมูแดงสีมรกต	11	17	758	0.188	0.667	2.397	0.667	1.159
แซมาอีลชิกดัง	11	11	698	0.095	0.483	0.002	0.048	3.019

ตารางที่ 14 แสดงรายละเอียดของผลการทดลองที่ได้ในแต่ละชุดข้อมูลภูมินามทดสอบ (ต่อ)

ภูมินาม	จำนวนภูมินาม	จำนวนคำบ่งชี้สถานที่	จำนวนตัวอักษร	Topolgy Words	DBSCAN	K-means	K-medoids	Agglomerative clustering
seoga and cook	7	6	1,013	0.019	0.019	0.01	0.019	0.159
Kam's Roast	5	2	383	0.012	0.012	0.012	0.012	0.002
เฟิงฟู่	15	9	1,988	2.142	0.308	0.308	2.142	0.039
Japang	2	3	240	0.014	0.014	0.014	0.014	0.012
Pasta AMA	4	1	397	0.542	0.542	0.004	0.542	0.308
Gimbocha BTSสนามกีฬาฯ	6	2	694	0.169	0.169	0.169	10.672	0.014
Yoko Donut and John's Lemon สาขาปิ่นเกล้า	2	1	135	0.021	3.189	6.366	0.021	0.004
FAMTIME พระราม 3	7	7	433	1.274	1.274	1.274	1.274	0.169
Mozza by Cocotte	2	3	216	0.028	0.015	0.015	0.028	0.021
Tangible	11	14	1,192	0.849	13.177	13.177	0.016	0.849
Azabusabo Thailand	1	1	249	0.027	0.027	0.027	0.027	0.027
เจ๊โอว	16	14	3,581	0.169	0.947	0.169	0.947	0.169
Mother Roaster	5	7	1,123	0.624	0.439	2.842	0.935	0.624
สุกี้จินดา	11	11	443	2.684	2.684	0.314	2.684	5.296
Abandoned Mansion	4	6	443	0.005	0.811	1.627	0.811	1.627
Bankara ramen siam paragon	4	0	206	2.087	2.087	4.234	4.234	0.059
Fran's	21	10	2,326	0.215	0.215	0.215	0.215	0.215
Bull & Bear	4	8	1,051	0.009	2.212	4.429	4.429	0.009
Jolly Porkii	3	4	167	0.002	0.002	0.002	0.002	0.002
THE HIDDEN MILKBAR	2	3	266	0.092	0.463	0.003	0.003	0.092
Saranghae Bingsu	3	3	127	0.033	0.033	0.033	0.033	0.033
HUUS of bread	14	10	1,162	0.004	1.572	0.004	1.572	3.142

ตารางที่ 14 แสดงรายละเอียดของผลการทดลองที่ได้ในแต่ละชุดข้อมูลภูมินามทดสอบ (ต่อ)

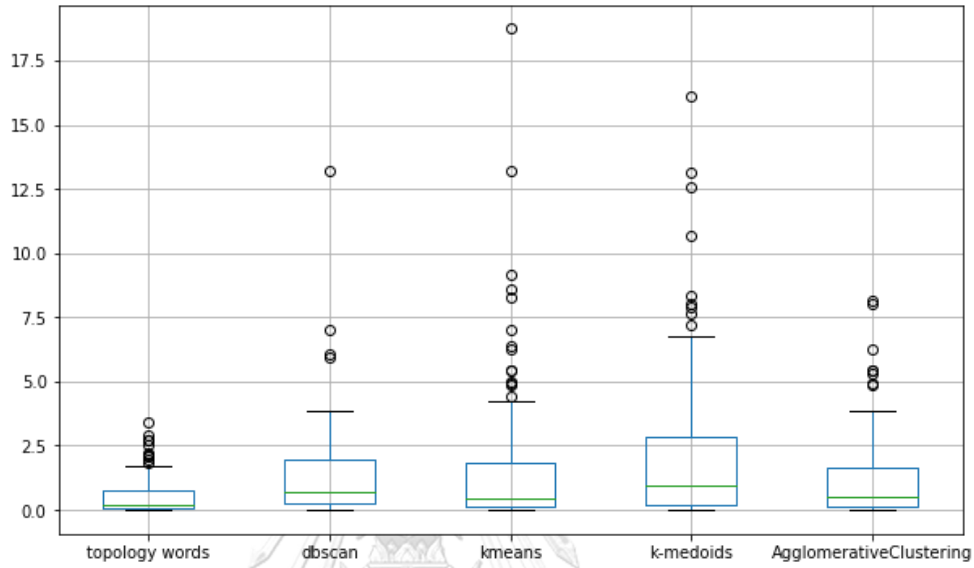
ภูมินาม	จำนวนภูมินาม	จำนวนคำบ่งชี้สถานที่	จำนวนตัวอักษร	Topolgy Words	DBSCAN	K-means	K-medoids	Agglomerative clustering
Tsuru Udon	9	3	372	0.108	2.411	4.982	4.982	1.649
Shoshana	2	1	137	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
Camin Cuisine & Cafe	2	1	93	0.002	2.454	0.002	0.002	4.907
Katsukura	3	3	213	0.123	0.123	0.123	0.123	0.123
Custard Nakamura	2	3	566	0.094	0.094	0.004	0.004	0.189
Semolina	5	2	566	0.314	0.314	0.314	0.314	0.314
Micho	5	3	534	0.105	0.105	0.105	0.105	0.105
Mingle	3	5	247	1.974	3.864	3.864	7.612	3.864
museumof contemporary art (moca bangkok)	10	3	386	0.887	0.732	0.732	12.556	0.732
Street Art สวนเฉลิมหล้าราชเทวี	10	4	403	0.362	0.473	1.223	1.223	1.223
ล้ง 1919	9	15	445	0.142	0.744	2.522	0.744	0.548
วัดโฝวกวงซัน	12	13	430	0.621	1.732	1.732	6.483	1.732
เฮียให้	7	2	474	2.095	2.223	0.302	3.315	0.302
Max Beef สาขาบางเขน	16	4	258	0.234	0.158	0.158	0.158	0.158
obanzai kitaro	14	3	486	0.002	0.273	0.273	2.942	0.273
THE HOUSE ON SATHORN	8	6	558	0.044	0.321	0.128	0.509	0.128
Mama Dolores	6	6	309	0.264	0.471	0.133	0.471	0.133
กะทิ บ้านอาหารไทยและขนม	10	12	366	0.253	1.741	4.885	1.741	4.885
Christoph Chocolate	4	3	280	0.04	0.04	0.04	0.04	0.045
MunJom.A	9	13	379	1.349	1.439	1.167	1.679	1.167
Street Art สวนเฉลิมหล้าราชเทวี	10	4	403	0.362	0.473	1.223	1.223	1.223

ตารางที่ 14 แสดงรายละเอียดของผลการทดลองที่ได้ในแต่ละชุดข้อมูลภูมินามทดสอบ (ต่อ)

ภูมินาม	จำนวนภูมินาม	จำนวนคำบ่งชี้สถานที่	จำนวนตัวอักษร	Topolgy Words	DBSCAN	K-means	K-medoids	Agglomerative clustering
ล้ง 1919	9	15	445	0.142	0.744	2.522	0.744	0.548
วัดโฝวกวงซัน	12	13	430	0.621	1.732	1.732	6.483	1.732
เฮียให้	7	2	474	2.095	2.223	0.302	3.315	0.302
Max Beef สาขาบางเขน	16	4	258	0.234	0.158	0.158	0.158	0.158
obanzai kitaro	14	3	486	0.002	0.273	0.273	2.942	0.273
THE HOUSE ON SATHORN	8	6	558	0.044	0.321	0.128	0.509	0.128
Mama Dolores	6	6	309	0.264	0.471	0.133	0.471	0.133
กะทิ บ้านอาหารไทยและขนม	10	12	366	0.253	1.741	4.885	1.741	4.885
Christoph Chocolate	4	3	280	0.04	0.04	0.04	0.04	0.045
MunJom.A	9	13	379	1.349	1.439	1.167	1.679	1.167
ร้าน 55 โภชนา	11	14	523	2.186	0.833	9.183	3.293	0.833
ร้านอาหารกลางซอย	12	20	585	0.245	0.515	0.515	5.928	0.515
ต้นเครื่อง	15	3	310	1.021	1.136	1.136	7.174	8.159
Noname noodle	10	7	318	0.281	0.154	8.246	2.577	0.154
Toronto.bkk	16	6	368	0.034	0.294	0.294	0.294	0.294
วิเสทสรร	12	3	509	0.251	0.483	8.614	3.052	0.483
13 Coins สุขุมวิท	13	15	405	0.083	5.922	0.684	8.048	8.048
RMSE				**0.947**	**2.211**	**3.430**	**3.833**	**2.183**
S.D.				**0.766**	**1.790**	**2.953**	**3.127**	**1.773**
Min	**1**	**0**	**93**	**0**	**0.001**	**0.001**	**0.001**	**0.001**
Max	**35**	**39**	**3581**	**3.385**	**13.177**	**18.73**	**16.11**	**8.159**

เมื่อพิจารณาจากตารางที่ 14 พบว่าแบบจำลอง topology words ให้ค่าคลาดเคลื่อนน้อยที่สุดคือ 0 หรือทำนายได้ตรงกับตำแหน่งของภูมินามนั้นจริงๆ และสำหรับแบบจำลองอื่น ค่าคลาดเคลื่อนที่น้อยที่สุด คือ 0.001 ซึ่งในความเป็นจริงคือทำนายได้ถูกต้องตามพิกัดของสถานที่จริงแล้ว แต่พบว่าค่าคลาดเคลื่อนที่มากที่สุดของ k-means และ k-medoids มีขนาดถึง 18.73 และ

16.11 กม. ตามลำดับ โดยส่วนเบี่ยงเบนมาตรฐานของค่าคลาดเคลื่อนนั้น ไม่ต่างจากค่า RMSE นัก แสดงให้เห็นว่าค่าคลาดเคลื่อนส่วนใหญ่มีการเกาะกลุ่มกัน แสดงแผนภูมิค่าคลาดเคลื่อนระหว่างแบบจำลองตามรูปที่ 33



รูปที่ 33 แผนภูมิแบบกล่องแสดงการกระจายตัวของค่าคลาดเคลื่อนระหว่างแบบจำลองการประมาณตำแหน่งภูมินาม

จากรูปที่ 33 แกนตั้งแทนค่าคลาดเคลื่อน มีหน่วยเป็น กิโลเมตร (กม.) แกนนอนแทนประเภทของแบบจำลอง พบว่าแบบจำลอง Topology words มีค่าคลาดเคลื่อนที่น้อยและเกาะกลุ่มกันมากที่สุด รองลงมาคือ Agglomerative clustering สำหรับแบบจำลอง DBSCAN มีค่าคลาดเคลื่อนที่น้อยแต่พบว่ามีบางจุดที่มีค่าคลาดเคลื่อนกระโดดออกจากกลุ่มค่อนข้างเยอะกว่า 2 แบบจำลองข้างต้น และแบบจำลองที่มีค่าคลาดเคลื่อนรวมทั้งการกระจายตัวของค่าคลาดเคลื่อนมากที่สุดคือ K-means และเมื่อพิจารณาความสัมพันธ์ระหว่างจำนวนภูมินาม จำนวนคำบ่งชี้สถานที่ และจำนวนตัวอักษร มีผลต่อแบบจำลองมากน้อยเพียงใด จึงสร้างตารางเมตริกค่าสหสัมพันธ์ แสดงตามรูปที่ 34

	จำนวนภูมินาม	จำนวนคำบ่งชี้สถานที่	จำนวนตัวอักษร	topology words	dbscan	kmeans	k-medoids	AgglomerativeClustering
จำนวนภูมินาม	1.00	0.60	0.49	-0.01	0.07	0.07	0.46	0.12
จำนวนคำบ่งชี้สถานที่	0.60	1.00	0.43	-0.06	0.08	0.03	0.25	0.00
จำนวนตัวอักษร	0.49	0.43	1.00	-0.08	-0.02	-0.05	-0.01	-0.16
topology words	-0.01	-0.06	-0.08	1.00	0.25	0.22	0.16	0.36
dbscan	0.07	0.08	-0.02	0.25	1.00	0.54	0.20	0.44
kmeans	0.07	0.03	-0.05	0.22	0.54	1.00	0.07	0.26
k-medoids	0.46	0.25	-0.01	0.16	0.20	0.07	1.00	0.24
AgglomerativeClustering	0.12	0.00	-0.16	0.36	0.44	0.26	0.24	1.00

รูปที่ 34 แสดงเมตริกค่าสหสัมพันธ์ระหว่างข้อมูลที่ได้จากแบบจำลองและชุดข้อมูล

จากรูปที่ 34 พบว่าจำนวนของภูมินามที่สกัดได้จากข้อความมีความสัมพันธ์กับแบบจำลอง k-medoids แต่จำนวนคำบ่งชี้สถานที่และจำนวนตัวอักษรไม่ได้มีความสัมพันธ์กับแบบจำลองการประมาณตำแหน่งภูมินามแบบใดเลย

5.3 กรณีตัวอย่างการประมาณตำแหน่งของภูมินามจากชุดข้อความ

การประมาณตำแหน่งของภูมินามจากข้อความเป็นส่วนที่สองของงานวิจัยนี้โดยเป็นการขยายผลจากการสกัดภูมินามออกจากข้อความแล้ว สามารถที่จะระบุตำแหน่งของสถานที่เหล่านั้นได้

จากหัวข้อที่ 5.1 ประสิทธิภาพของแบบจำลองที่ให้ประสิทธิภาพสูงที่สุดคือ Topology words แต่ในระหว่างการทดลองยังพบเห็นกรณีศึกษาในแต่ละรูปแบบที่มีความน่าสนใจจึงยกผลการทดลองในส่วนของการประมาณตำแหน่งมาไว้ในหัวข้อนี้

5.3.1 กรณีที่แบบจำลอง Topology words ให้ค่า RMSE ใกล้เคียงกับ 0 หรือใกล้เคียงสถานที่จริงมาก

แสดงตัวอย่างกรณีศึกษาในข้อมูลภูมินาม ชื่อ Brunch paradiso (บรันช์พาราดิโซ) จากชุดข้อความที่กล่าวถึง Brunch paradiso จำนวน 4 ข้อความ

Brunch paradiso

1. ไม่บ่อยนักที่กรุงเทพฯ จะมีร้านอาหารที่เน้นเสิร์ฟเมนูบรันช์เป็นหลักให้ได้ไปนั่งเอ็นจอยกัน ครั้งนี้ BKK. ขอแนะนำร้าน Brunch Paradiso คาเฟ่สายบรันช์ในย่านเย็นอากาศ ที่ตอบโจทย์สำหรับใครที่มองหามื้อบรันช์หรือคาเฟ่ที่นั่งสบาย ๆ ในช่วงเช้าถึงบ่าย มีที่จอดรถกว้างขวาง เพราะคาเฟ่แห่งนี้เป็นส่วนหนึ่งของโรงแรม Shama Yen-Akat Bangkok แบรนด์โรงแรมดังจากเกาะฮ่องกง และยังเวลคัมเหล่าบรรดาสัตว์เลี้ยงให้มาใช้ช่วงเวลาดี ๆ ด้วยกันอีกด้วย ทางร้านมาพร้อมคอนเซ็ปต์สนุก ๆ ที่ต้องการเน้นว่าทุกอย่างเป็นไปได้ เพราะอาหารเช้า ไม่จำเป็นต้องทานตอนเช้าเท่านั้น กิมมิกเหล่านี้ถูกถ่ายทอดผ่านภาพศิลปะที่ประดับตามมุมต่าง ๆ ภายในร้าน เช่น รูปขนมปังเดินได้ เป็นต้น ด้วยความตั้งใจที่อยากให้ที่นี่เป็นสวรรค์ของคนรักอาหารเช้า ให้มาใช้เวลาทานมื้ออร่อยกัน

2. ร้าน Brunch Paradiso ใช้พื้นที่ด้านหน้าของโรงแรม Shama บนถนนเย็นอากาศ ต้อนรับบรันช์เลิฟเวอร์ ที่นี่ยังเป็น Dog Friendly ที่มาสุนัขเข้ามาได้แต่ต้องมีสายจูงหรือรถเข็นเพื่อไม่ให้น้องไปกวนลูกค้าคนอื่น ๆ เราชอบงานอาร์ตที่ใช้ตกแต่ง ส่วนใหญ่เป็นสตอรีของมื้ออาหาร โดยเฉพาะ Art Piece ชิ้นเด่นที่เราสารภาพว่าไม่รู้ว่าเป็นของศิลปินท่านไหน เราขอเรียกว่า “ขนมปังเดินได้” ที่เป็นทั้งภาพวาดและไฟนีออน ตัวเป็นขนมปังขาเป็นขาไก่ขาเป็ด มันสื่อความเป็นบรันช์ได้ดีมาก

3. สายบรันช์ต้องแวะมาที่ Brunch Paradiso คาเฟ่ในย่านเย็นอากาศ ที่นอกจากจะมีหลากหลายเมนูบรันช์ เบเกอรี และเครื่องดื่มแล้ว ยังมาพร้อมสเปซสวย ๆ นั่งสบายและเวลคัมเหล่าบรรดาสัตว์เลี้ยงสุดน่ารักอีกด้วย

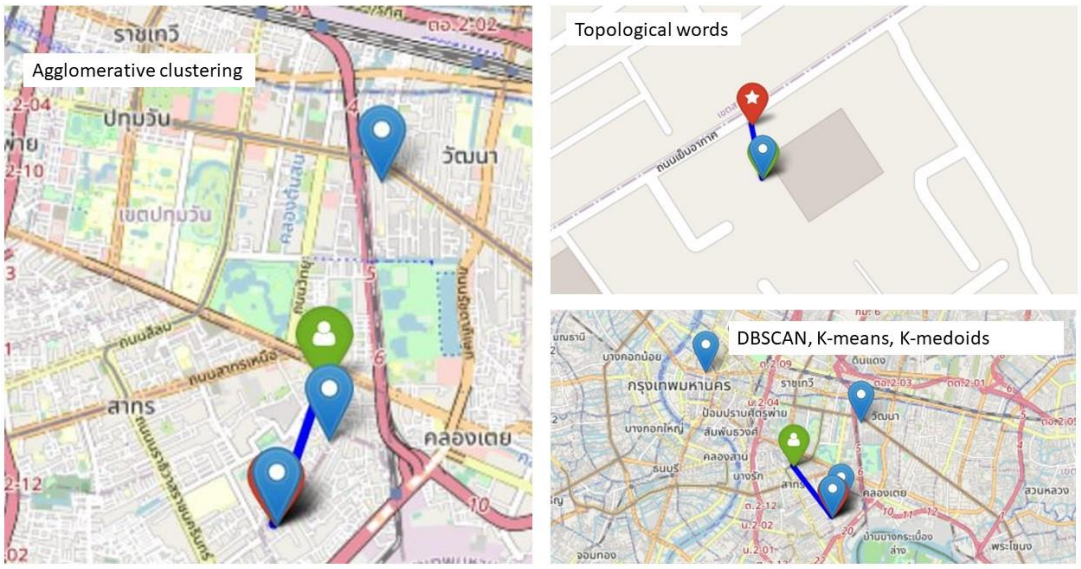
4. Brunch paradiso ที่อยู่ 69 ถ. เย็นอากาศ กรุงเทพมหานคร

จากข้อความข้างต้นเมื่อเข้าสู่การประมวลผล คำเป้าหมายและคำที่มีการอ้างอิงถึงตำแหน่งของสถานที่จะถูกสกัดออกมาจากข้อความ แสดงตัวอย่างตามรูปที่ 35 ซึ่งชื่อ Brunch Paradiso จะถูกดึงออกไปเนื่องจากเป็นคำเป้าหมายเหลือไว้เฉพาะคำที่อ้างอิงถึง Brunch Paradiso เพียงอย่างเดียว นอกจากนี้ tag FPLACE ซึ่งเป็น tag ที่บ่งบอกว่าสถานที่นั้นไม่ได้อยู่ในประเทศไทย ก็จะไม่ถูกนำมาประมวลผลด้วยเช่นกัน โดยอัลกอริทึมในการประมาณตำแหน่งจะมีพิกัดของ 4 จุดนี้เป็นข้อมูลสำหรับประมวลผลแสดงตามรูปที่ 35

```
[[['ที่', 'GEO'],
  ['กรุงเทพมหานคร', 'ADMIN'],
  ['ที่', 'GEO'],
  ['brunchparadiso', 'RT'],
  ['ใน', 'GEO'],
  ['ที่', 'GEO'],
  ['โรงแรมshama yen-akatbangkok', 'RES'],
  ['เกาะฮ่องกง', 'FPLACE'],
  ['ใน', 'GEO'],
  ['ที่', 'GEO']],
 [['ร้าน', 'RT'],
  ['brunchparadiso', 'RT'],
  ['ที่', 'GEO'],
  ['พื้นที่', 'GEO'],
  ['โรงแรมshama', 'RES'],
  ['ที่', 'GEO'],
  ['บน', 'GEO']],
 [['ที่', 'GEO'], ['brunchparadiso', 'RT'], ['ใน', 'GEO'], ['ที่', 'GEO']],
 [['ถ.เย็นอากาศ', 'ROAD']]]
```

รูปที่ 35 แสดงตัวอย่างข้อมูลภูมินามและคำที่อ้างอิงเชิงตำแหน่งซึ่งสกัดได้จากข้อความ

โดยผลลัพธ์ที่ได้ในแต่ละแบบจำลองมีดังนี้ topology words มีจุดที่ใช้อ้างอิงเพียงจุดเดียว คือโรงแรม Shama-yenakat Bangkok ส่วนแบบจำลอง DBSCAN, K-means, K-medoids ใช้ข้อมูลทั้งหมด 4 จุด และสุดท้ายคือ Agglomerative clustering 3 จุด โดยตัดเอา จุดที่เป็นพิกัดของกรุงเทพมหานครออกไป ซึ่งผลของค่า RMSE ต่ำที่สุดคือ 0.029 กม. หรือ 29 เมตร เนื่องจากแบบจำลองสามารถระบุได้ว่าร้าน Brunch Paradiso มีที่ตั้งอยู่ในโรงแรม Shama-yenakat แต่อัลกอริทึมอื่นไม่สามารถกรองจนเหลือเพียงข้อมูลที่เจาะจงเช่นนี้ แต่สำหรับอัลกอริทึม DBSCAN, K-means หรือ K-medoids ไม่สามารถเอาจุดที่เป็นตำแหน่งของกรุงเทพมหานครซึ่งห่างจากกลุ่มของจุดที่อยู่รอบๆ โรงแรม Shama-yenakat ออกไปได้ ทำให้ค่า RMSE เป็น 2.32 กม. และสุดท้าย คือ Agglomerative clustering ซึ่งให้ค่า RMSE ที่ 1.41 กม. เนื่องจากมีการกรองเอาตำแหน่งของกรุงเทพมหานครออกไป ทำให้ขอบเขตที่ประมาณตำแหน่งแคบลงกว่าแบบจำลองก่อนหน้า โดยแสดงผลบนแผนที่ตามรูปที่ 36 โดย marker สีน้ำเงินคือ ตำแหน่งที่ได้จากข้อความโดยตรงผ่านเครื่องมือ toponym recognition marker สีเขียว คือตำแหน่งที่ประมาณขึ้นจากแบบจำลองแต่ละแบบและสุดท้าย marker สีแดงคือ ตำแหน่งที่ตั้งจริงๆ ของภูมินามนั้น



รูปที่ 36 แผนที่แสดงผลลัพธ์จากแบบจำลองแต่ละแบบในกรณีของ Brunch paradiso

5.3.2 กรณีที่ภูมินามไม่ได้อยู่บนห้างร้าน ตึก หรือ อาคารสำนักงานใด

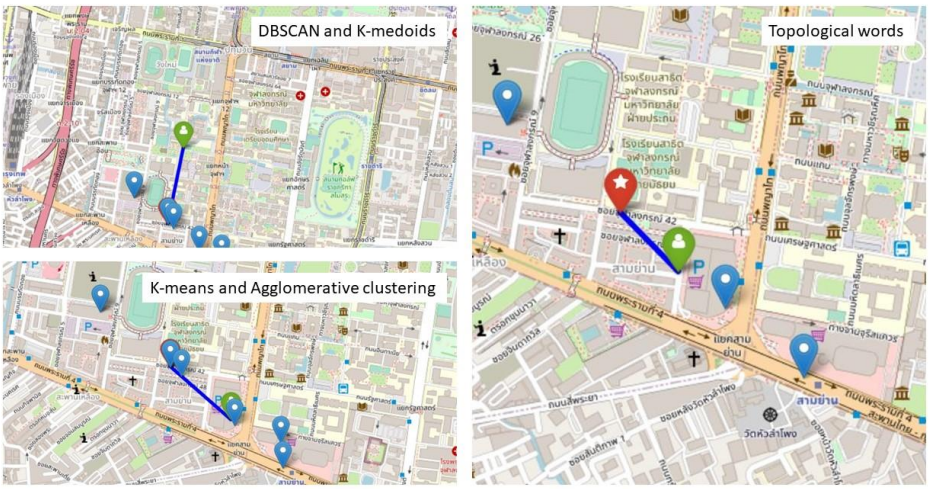
แสดงตัวอย่างกรณีศึกษาในข้อมูลภูมินาม ชื่อ ร้านจีฉ่อย จากชุดข้อความที่กล่าวถึง ร้านจีฉ่อย จำนวน 3 ข้อความ ซึ่งจากชุดข้อความนี้สามารถสกัดเอาภูมินามออกมาได้ ดังนี้ คือ จีฉ่อย ร้านจีฉ่อย ซึ่งเป็นชื่อเป้าหมาย ตลาดสามย่าน ถนนพญาไท อาคารยูเซ็นเตอร์ 1 สามย่านมิตรทาวน์ ร้านแว่นตาบุญชัยการแว่น บนถนนสามย่าน ซึ่งชื่อสุดท้ายที่สกัดออกมาได้จะพบว่า บนถนนสามย่านตัวแบบจำลองการรู้จำภูมินามสกัดคำว่า “บน” ซึ่งเกินจากชื่อออกมา เมื่อข้อความเข้าสู่การประมวลผล ผลลัพธ์ที่ได้ในแต่ละแบบจำลองมีดังนี้ topology words มีจุดที่ใช้อ้างอิงจากข้อความจำนวน 4 จุด ซึ่งมีค่า RMSE อยู่ที่ 0.169 กม. ส่วนแบบจำลอง K-means, Agglomerative clustering ใช้ข้อมูลทั้งหมด 6 จุด มีค่า RMSE อยู่ที่ 0.268 กม. และสุดท้ายคือ DBSCAN และ K-medoids ใช้ข้อมูลทั้งหมด 8 จุด ซึ่งผลของค่า RMSE คือ 0.532 กม. ซึ่งในกรณีนี้จะพบว่า K-means กลับมีค่า RMSE ต่ำกว่า DBSCAN อาจเป็นเพราะว่ามีข้อมูลจำนวนที่มากขึ้นทำให้มีผลต่อการกรองข้อมูลไปด้วย และสำหรับ Topology words อัลกอริทึม จะมีค่า RMSE ที่มากขึ้นเนื่องจากมีจุดที่มาเฉลี่ยขยายขอบเขตออกไป แต่ก็ยังให้ค่าคลาดเคลื่อนยังอยู่ในระยะ 0.2 กม. หรือ 200 เมตร แสดงผลลัพธ์ตามรูปที่ 36 โดย marker สีน้ำเงินคือ ตำแหน่งที่ได้จากข้อความโดยตรงผ่านเครื่องมือ toponym recognition marker สีเขียว คือตำแหน่งที่ประมาณขึ้นจากอัลกอริทึมแต่ละแบบและสุดท้าย marker สีแดงคือตำแหน่งที่ตั้งจริงๆ ของภูมินามนั้น

จีฉ่อย

1. เมื่อวันที่ 12 ธ.ค. เพจ “บูม บาม” ได้โพสต์เรื่องราวของร้านขายของชำที่ชื่อว่าร้าน “จีฉ่อย” สำหรับร้านจีฉ่อย เป็นร้านขายของชำขนาดหนึ่งคูหา ตั้งอยู่หน้าตลาดสามย่าน ถนนพญาไท ตรงข้ามจุฬาลงกรณ์มหาวิทยาลัย ขึ้นชื่อในบรรดานิสิตจุฬาลงกรณ์มหาวิทยาลัย ว่ามีของขายทุกอย่าง และถ้าของไหนไม่มีขายในร้าน จะสามารถมาเอาได้ภายใน 2 วัน ปัจจุบันย้ายร้านไปที่อาคารยูเซ็นเตอร์ 1 ซอยจุฬา 4 ถนนพระราม 4

2. เมื่อปี 47-48 ผมได้มีโอกาสผ่านอยู่บ่อย ๆ เพราะไปเรียนแถวนั้นแต่ก็ไม่รู้ว่านี่คือร้านดังประจำมหาวิทยาลัย บางวันก็เห็นแกเปิดประตูแค่ครึ่งเดียว ช่วงนั้นผมไม่คิดว่าร้านละแวกนี้จะโดนย้าย ถ้าเดินมาจากหัวมุมสามย่าน ที่ปัจจุบันเป็นสามย่านมิตรทาวน์ จะมีร้านแว่นตาบุญชัยการแว่น ร้านทำผม ร้านข้าวหน้าเป็ด ร้านโจ๊ก จีฉ่อย ถ้าจำไม่ผิดจะมีร้านทำกุญแจอยู่ด้วย

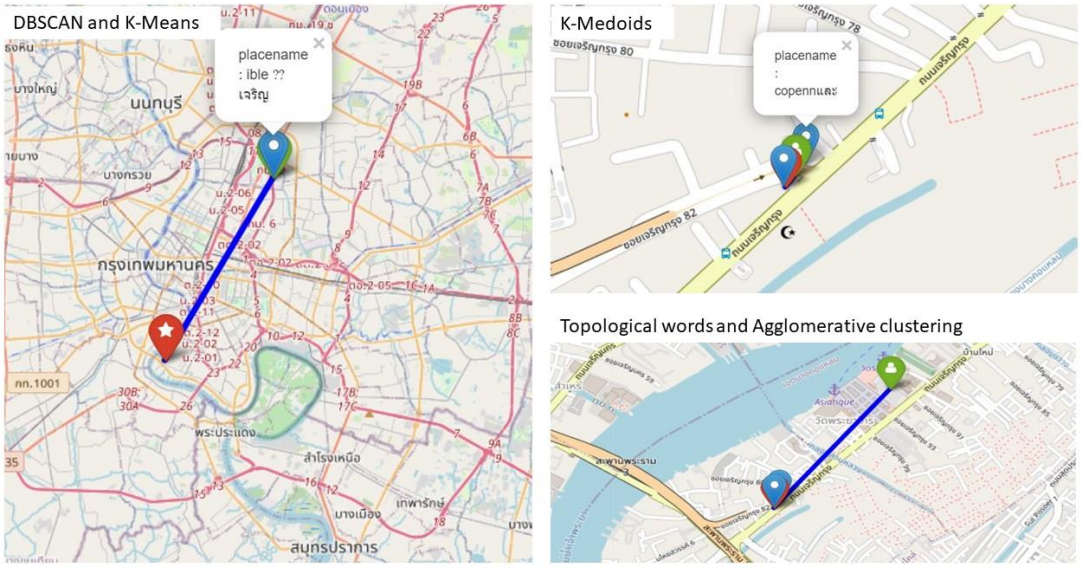
3. ร้านจีฉ่อย ร้านขายของชำในตำนานบนถนนสามย่าน ตึกแถวขนาด 1 คูหาอยู่หน้าตลาดสามย่านเปิด 24 ชม. (เมื่อก่อนจะย้ายร้าน) จีฉ่อยเป็นร้านขายของที่ขึ้นชื่อว่า มีของขายทุกอย่าง และถ้าของไหนไม่มีขายในร้าน จะสามารถมาเอาได้ภายใน 2 วันให้หลัง แอดมินเคยบังเอิญไปซื้อของที่นี่สิบกว่าปีได้แล้วจำได้ว่า งง ทำไมร้านนี้มันมีทุกอย่าง สีพลาสติก อุปกรณ์จัดพร็อบมีทุกอย่างเลยเหรอ แล้วก็มีของเล่นสังกะสีสภาพสวย แท้และเก่าแน่นอน แต่จีฉ่อยไม่ยอมขายบอกว่าอยากได้เหรอแล้วก็เงียบไป ไม่ยอมขายแต่ก็รู้ทีหลังว่าร้านนี้คือตำนานของเด็กจุฬาก็ที่งไปเลย



รูปที่ 37 แผนที่แสดงผลลัพธ์จากแบบจำลองแต่ละแบบในกรณีของร้านจีฉ่อย

5.3.3 กรณีที่แบบจำลองมีการประมาณตำแหน่งผิดพลาดมากที่สุด

แสดงตัวอย่างกรณีศึกษาในข้อมูลภูมินาม ชื่อ Tangible จากชุดข้อความที่กล่าวถึง ร้าน Tangible จำนวน 6 ข้อความ ซึ่งจากชุดข้อความนี้สามารถสกัดเอาภูมินามออกมาได้ ดังนี้ คือ tangible café ซึ่งเป็นชื่อเป้าหมาย และภูมินามอื่นที่ถูกสกัดออกมาได้แก่ ‘t’ ‘ang’ ‘ible’ ‘copenn และ’ ‘ible ?? เจริญ’ ‘กรุง’ ‘happieland’ ‘เจริญกรุง’ สังเกตเห็นว่าภูมินามที่สกัดออกมามีความผิดพลาดค่อนข้างมาก ซึ่งในส่วนของ t จะถูกตัดออกไปเนื่องจากมีอักษรตัวเดียว และในส่วนของชื่ออื่นที่ไม่สมบูรณ์หรือเกินจะถูกกรองออกไปจากการเข้ารหัสภูมิศาสตร์ ทำให้ในกรณีนี้ถึงแม้จะสกัดภูมินามออกมาได้จำนวนมากแต่กลับใช้ประโยชน์จริงได้น้อย โดยแบบจำลอง topology words และ Agglomerative Clustering ให้ค่า RMSE ที่ 0.849 และ สำหรับแบบจำลอง DBSCAN และ K-means คือ 13.177 โดย k-medoids ให้ค่า RMSE ที่ 0.16 ซึ่งเหตุที่ทำให้มีความคลาดเคลื่อนจากตำแหน่งจริงมาก เนื่องจากจุดที่นำมาใช้อ้างอิงมีจำนวนน้อยมาก และเป็นจุดที่ไม่ได้อยู่ใกล้กับชื่อเป้าหมายจึงให้ค่า RMSE ที่สูงกว่า 10 กม. ดังที่กล่าว แสดงผลตามแผนที่ ในรูปที่ 38 โดย marker สีน้ำเงินคือ ตำแหน่งที่ได้จากข้อความโดยตรงผ่านเครื่องมือ toponym recognition marker สีเขียว คือตำแหน่งที่ประมาณขึ้นจากอัลกอริทึมแต่ละแบบและสุดท้าย marker สีแดงคือตำแหน่งที่ตั้งจริงๆ ของภูมินามนั้น



รูปที่ 38 แผนที่แสดงผลลัพธ์จากแบบจำลองแต่ละแบบในกรณีของ tangible café

Tangible cafe'

1. ล่าสุดไปคาเฟ่ event replica maison margiela x tangible เป็นคาเฟ่ที่ไปซ้ำมา2รอบแล้ว ไปกี่ครั้งก็ชอบ เค้าจัดดีมาก event ไรก็ดูเข้ากับร้านไปหมด ร้านไม่ใหญ่มากแต่ถ่ายรูปสวยทุกมุม

2. แงเพิ่งรู้ว่า maison margiela จัดที่ tangible เมื่อวานวันสุดท้าย เศร้าอ่าอยากไป ᵀᵀ ᵀᵀ

3. งาน collab มีถึงเมื่อวันที่ 24 ต.ค. วันนี้เลย เราได้ไปแบบเฉียดฉิวก่อนหนึ่งวัน เลยขอมาลงย้อนหลังเป็นความทรงจำไว้ฮะพิกัด

4. Tangible 📍เจริญกรุง อีกหนึ่งคาเฟ่เท่ๆ เรามีโอกาสได้ไปช่วงที่เค้า collab กับแบรนด์น้ำหอมอย่าง Replica ภายในร้านก็จะมี tester น้ำหอมด้วย การตกแต่งร้านก็คือเหมือน art installation เลย ชอบด้านบนที่มีต้นไม้กับบึง แปลกดี@aroii

5. 'HAPPIELANDร้านขายเสื้อผ้าและ Smoking experience เปิดใหม่ที่เจริญกรุง ตัวร้านโดดเด่นที่การตกแต่งที่ใช้สีขาวล้วน รวมถึงการจัดวาง Product ในร้าน ที่ถือว่ามีสไตล์ที่โดดเด่นมากๆ แนะนำว่าถ้ามา tangible cafe แล้วให้มาเดินชิวต่อที่ copenn และ HAPPIELAND เลยครับ อยู่โซนเดียวกันหมดเลย

6. คือ เราจะให้ NFT เป็นการลงทุนในความตั้งใจของเราก็ได้ (เพราะคำจำกัดความคำว่าลงทุนมันเป็นเรื่องของ estimated value ที่เราคาดจะได้รับในอนาคต ไม่จำเป็นต้องเป็น tangible asset ก็ได้ จะเป็นในของความรู้ ความสัมพันธ์ สุขภาพ ฯลฯ ก็ได้ครับ อยู่กับที่เรามองและให้ค่า

บทที่ 6
อภิปราย สรุปผลและข้อเสนอแนะ

งานวิจัยนี้แบ่งการอภิปราย สรุปผล และข้อเสนอแนะของงานวิจัยออกเป็นสองส่วน คือ หัวข้อ 6.1 สำหรับการสร้างแบบจำลองรู้จำภูมินาม และหัวข้อที่ 6.2 สำหรับการประมาณตำแหน่งของภูมินามจากชุดข้อความ

6.1 อภิปราย สรุปผลและข้อเสนอแนะงานวิจัยสำหรับผลการสร้างแบบจำลองรู้จำภูมินาม

แบบจำลองการรู้จำภูมินามถือเป็นส่วนแรกของงานวิจัยนี้ในการแปลความหมายทางภูมิศาสตร์เพื่อสกัดเอาข้อมูลภูมินามในข้อความออกมา แต่เนื่องจากในปัจจุบันสำหรับภาษาไทยนั้นยังไม่มีงานวิจัยที่มุ่งเน้นศึกษาเฉพาะด้านในการสกัดภูมินามและจัดหมวดหมู่ โดยทั่วไปจะเป็นการรู้จำชื่อเฉพาะซึ่งไม่สามารถนำมาใช้งานในงานวิจัยนี้ได้โดยตรง จึงต้องสร้างแบบจำลองรู้จำภูมินามสำหรับภาษาไทยขึ้นมาโดยเฉพาะ ซึ่งมีการอ้างอิงสถาปัตยกรรมที่ใช้มาจากงานวิจัยอื่นที่เกี่ยวข้อง

ในงานวิจัยนี้มีการเลือกใช้แบบจำลองในการเรียนรู้ของเครื่องทั้งหมด 3 กลุ่ม โดยกลุ่มแรก คือ 1) แบบจำลองการเรียนรู้ของเครื่องแบบลำดับ (CRF) และ 2) คือโครงข่ายประสาทเทียมวกกลับทั้ง LSTM และ GRU รวมทั้งแบบที่ 3) คือการถ่ายโอนความรู้ (BERT)

ผลจากการสร้างแบบจำลองพบว่าเฉพาะแบบจำลองในกลุ่มที่ 1) และ แบบจำลองในกลุ่มที่ 2) นั้นให้ค่าความถูกต้องโดยรวม F1-Phrase ที่ใกล้เคียงกัน โดยแบบจำลอง CRF ที่สร้างฟังก์ชันคุณลักษณะเฉพาะสำหรับภูมินามให้ค่า F1-Phrase ที่สูงที่สุด (ตารางที่ 9) เนื่องจากการรู้จำภูมินามมีการแบ่งกลุ่มของภูมินามออกเป็น 18 ประเภท และในงานวิจัยนี้มีคลังข้อมูลทางภาษาที่ขนาดไม่ใหญ่ทำให้การสร้างฟังก์ชันคุณลักษณะเฉพาะด้านทางภูมิศาสตร์ เช่น คำบ่งชี้สถานที่ ใกล้ ไกล ถัดจาก ตรงไป ฯลฯ รวมถึงการนำอักขรานุกรมภูมิศาสตร์มาเป็นส่วนประกอบของฟังก์ชันคุณลักษณะ ทำให้ได้ผลลัพธ์ที่ใกล้เคียงหรือดีกว่าการใช้โครงข่ายประสาทเทียม หากมีคลังข้อมูลทางภาษาขนาดจำกัด เช่น งานวิจัยนี้ และเมื่อพิจารณาจากค่าความถูกต้องโดยรวมในทุกแบบจำลองแล้ว แบบจำลองในกลุ่มที่ 3) ซึ่งเป็นการถ่ายโอนความรู้นั้นให้ค่าความถูกต้องโดยรวมสูงที่สุดเมื่อเทียบกับทุกแบบจำลอง (ตารางที่ 9) โดยการใช้การฝึกสอนเพิ่มเติมกับแบบจำลอง WangchanBERTa เนื่องจากเป็นแบบจำลองที่ผ่านการฝึกสอนมาคลังข้อมูลทางภาษาขนาดใหญ่ ซึ่งเป็นผลลัพธ์ที่สอดคล้องกับงานวิจัยของ L. Lowphansirikul et al. (2021) และแม้ว่าแบบจำลองการรู้จำภูมินามที่สร้างจาก WangchanBERTa ให้ประสิทธิภาพที่ดีที่สุดในงานวิจัยนี้ แต่หากเปรียบเทียบกับแบบจำลองอื่นในบางประเภทของภูมินาม เช่น <NAT> หรือ สถานที่ธรรมชาติพบว่าแบบจำลองที่เป็น CRF ให้ค่าความถูกต้องโดยรวมที่สูงที่สุดอาจเนื่องจากเป็นชื่อที่มีความเฉพาะเจาะจง มีการใช้ภาษาถิ่น การมี

อักขรานุกรมภูมิศาสตร์เป็นส่วนหนึ่งของฟังก์ชันคุณลักษณะใน CRF จึงทำให้ค่าความถูกต้องโดยรวมของภูมินามประเภทนี้มีค่าสูงกว่า WangchanBERTa

ข้อจำกัดบางประการของแบบจำลองภูมินาม เช่น การสกัดเอาชื่อภูมิศาสตร์ที่อาจจะเกินขอบเขตชื่อจริงไป เช่น “ภายในซอยสุขุมวิท 24” ซึ่งความจริงควรจะเป็น “ซอยสุขุมวิท 24” หรือการสกัดชื่อภูมิศาสตร์ออกมาไม่ครบถ้วนโดยมีการแบ่งเป็นวลีและประภทของภูมินามที่แตกต่างกัน เช่น tangible café แยกเป็น [‘t’, RT], [‘ang’, RES], [‘ible’, BSN] และ [‘café’, RT] ซึ่งอาจเนื่องมาจากตัวตัดคำที่ใช้ในงานวิจัย โดยในงานวิจัยนี้ใช้ตัวตัดคำที่เป็น sentence piece หรือ การจัดประเภทของภูมินามผิด เช่น ‘กรุงเทพกรีฑา’ ให้ประเภทเป็น ‘ADMIN’ แทนที่จะเป็น ‘ROAD’ โดยงานวิจัยนี้แก้ปัญหากรณีนี้ด้วยการสร้างกฎเพื่อใช้ในการกรองข้อมูลและปรับแก้ข้อมูลเพื่อให้มีความถูกต้องมากขึ้น

จากการอภิปรายและสรุปผลที่ผ่านมาข้างต้น มีข้อเสนอแนะสำหรับงานวิจัยในอนาคต 3 หัวข้อดังต่อไปนี้

1) การนำเทคนิค few shot learning มาปรับใช้ เนื่องจากแหล่งข้อมูลภาษาไทยยังมีจำกัดเมื่อเทียบกับภาษาอื่นที่เป็นภาษาหลักของโลกที่นิยม เช่น อังกฤษ ฝรั่งเศส สเปน จีน ฯลฯ ซึ่งการทำ few shot learning สามารถที่จะฝึกฝนกับข้อมูลที่มีจำนวนน้อยได้จึงมีความน่าสนใจที่จะนำมาทดลองใช้งานกับข้อมูลภาษาไทย

2) การทดลองปรับเพิ่มส่วนของคำฝังตัวในการนำคุณลักษณะเกี่ยวกับ Topology word มาเป็นคุณลักษณะฝังตัว (feature embedding) ซึ่งแบบจำลอง BERT มีการทำ embeddings ทั้งในส่วนของคำ วลี และประโยค การเพิ่มส่วนของ feature embedding อาจเป็นการช่วยเพิ่มจำนวนตัวแปรในการเรียนรู้ของแบบจำลองได้ดียิ่งขึ้นอีกทั้งในโจทย์ของการรู้จำชื่อเฉพาะ การมีความรู้เดิมเป็นพื้นฐานมีแนวโน้มที่จะส่งผลต่อประสิทธิภาพของแบบจำลอง เช่นในการสร้างแบบจำลอง CRF พบว่าแบบจอลง CRF ที่มีการนำอักขรานุกรมภูมิศาสตร์ (gazetteer) มาเป็นหนึ่งในคุณลักษณะให้ประสิทธิภาพสูงกว่าแบบจำลอง CRF ที่ไม่ใช้ gazetteer

3) การจัดกลุ่มของภูมินามใหม่ โดยในหลายงานวิจัยมีการจัดหมวดหมู่ของสถานที่แตกต่างกันออกไป โดยส่วนใหญ่จะเป็นการแบ่งประเภทที่ไม่เยอะมาก เช่น 10 ชนิด หรือหากมีการแบ่งประเภทที่มีจำนวนประเภทมากขึ้นอาจทำให้ค่าความถูกต้องโดยรวม (F1-Phrase) มากขึ้นกว่าเดิม (Tao et al., 2022; Jimin Wang et al., 2020)

6.2 อภิปราย สรุปผลและข้อเสนอแนะงานวิจัยสำหรับการประมาณตำแหน่งของภูมินามจากชุดข้อความ

ผลที่ผ่านแบบจำลองรู้จำภูมินามยังมีความไม่สมบูรณ์หรือผิดพลาดอยู่บางส่วน การเข้ารหัสภูมิศาสตร์จึงเป็นการลดความกำกวมและกรองเอาข้อมูลภูมินามที่ไม่มีอยู่จริงออกไป แต่ฐานข้อมูลออนไลน์ หรือฐานข้อมูลที่เตรียมไว้อาจจะไม่ทันกับการเปลี่ยนแปลงของข้อมูลที่มาจากสื่อสังคมออนไลน์ เช่น ทวิตเตอร์ ทำให้อาจมีชื่อภูมินามบางส่วนไม่สามารถระบุค่าพิกัดจากฐานข้อมูลเหล่านั้นได้แต่เป็นสถานที่ที่เพิ่งเกิดขึ้นมาไม่นาน งานวิจัยนี้จึงสร้างแบบจำลองขึ้นเพื่อแก้ปัญหาในส่วนนี้ ในหลายงานวิจัยมักจะเป็นการประมาณตำแหน่งของทวิตเตอร์ และโดยส่วนใหญ่มีความคลาดเคลื่อนทางตำแหน่งในระดับหลาย 10 หรือ 100 กิโลเมตร (กม.) (Williams et al., 2017; Xu et al., 2014) เนื่องจากไม่มีการนำคำที่เป็นคำบ่งชี้เชิงตำแหน่ง เช่น ใกล้กับ ติดกับ บน เยื้อง ตรงข้าม ฯลฯ มาใช้ในการประมวลผลร่วมด้วย

ใช้เทคนิค geocoding ด้วย topology words ดีกว่า การทำ geocoding ด้วย clustering algorithm

ซึ่งเมื่อนำกลุ่มคำข้างต้นมาสร้างเป็นแบบจำลอง topology words พบว่ามีค่า Mean Error (ME) ที่ 0.561 กม. และ RMSE ที่ 0.947 กม. ตามตารางที่ 13 และ ตารางที่ 14 โดยมีประสิทธิภาพมากที่สุดเมื่อเทียบกับแบบจำลองการเรียนรู้ของเครื่องอีก 4 แบบ และให้ค่าความคลาดเคลื่อนในระดับที่ใกล้ 0 มากเมื่อเป็นภูมินามที่ตั้งอยู่ในห้าง ร้าน หรืออาคารสำนักงานต่างๆ แต่ทั้งนี้ยังมีข้อจำกัดในบางประการ เช่น หากภูมินามที่สกัดออกมาได้มีจำนวนที่ผิดมากอาจทำให้การประมาณตำแหน่งคลาดเคลื่อนไปมาก หรือหากในประโยคนั้นสั้นมากและไม่มีคำบ่งชี้สถานที่อยู่จะส่งผลให้ตำแหน่งที่ประมาณได้จากแบบจำลองมีแนวโน้มที่จะมีค่าคลาดเคลื่อนมากเช่นเดียวกัน อีกทั้งสำหรับการประมาณตำแหน่งของภูมินามที่มีจุดไม่ทราบค่าพิกัดมากกว่า 1 จุดในประโยคเดียวกัน ในงานวิจัยนี้จะเก็บชื่อทั้งหมดที่ไม่สามารถดึงค่าพิกัดจากฐานข้อมูลที่เตรียมไว้ได้มาเก็บไว้ก่อนแล้วจึงวนซ้ำนำชื่อแต่ละชื่อไปค้นหาประโยคที่กล่าวถึงชื่อเหล่านั้น หลังจากนั้นนำมาผ่านเครื่องมือการรู้จำภูมินามแล้วจึงนำมาประมาณตำแหน่งจากอัลกอริทึมข้างต้น

จากการอภิปรายและสรุปผลที่ผ่านมาข้างต้น มีข้อเสนอแนะสำหรับงานวิจัยในอนาคตหัวข้อดังต่อไปนี้

1) การทดลองปรับแบบจำลองในการให้ค่าน้ำหนักที่ต่างกันไป อาจให้ความถูกต้องของตำแหน่งดีขึ้นกว่างานวิจัยนี้ โดยมีการแบ่งช่วงชั้นของค่าน้ำหนักที่แตกต่างกันมากขึ้นซึ่งแต่เดิมในงานวิจัยนี้แบ่งไว้ที่ 1-3 แต่หากมีการกำหนดค่าน้ำหนักให้ต่างออกไปโดยที่มีช่วงชั้นของค่าน้ำหนักที่ถี่ขึ้น เช่น 1-10

2) การนำขอบเขตการให้บริการของไปรษณีย์มาใช้งานในกรณีที่ไม่มีคำบ่งชี้ขอบเขตการปกครอง เนื่องจากรหัสไปรษณีย์ อาจทำให้สามารถระบุขอบเขตของสถานที่ได้แคบลงกว่าขอบเขตการปกครองอย่างจังหวัด โดยในหนึ่งอำเภอจะใช้รหัสไปรษณีย์เดียวกัน เช่น อ.กุดชุม จ.ยโสธรจะมีรหัสไปรษณีย์เป็น 35140 แต่หากมีข้อมูลขอบเขตการปกครองในระดับ อำเภอ ตำบลอยู่แล้วไม่จำเป็นต้องนำรหัสไปรษณีย์มาใช้งาน

3) การปรับแบบจำลองโดยมีการตรวจสอบคำบ่งชี้สถานที่หากในประโยคนั้นไม่มีคำบ่งชี้สถานที่ หรือ topology words ให้นำอัลกอริทึมซึ่งอาจเป็น DBSCAN หรือ Agglomerative clustering มาเป็นส่วนประมวลผลในการจัดกลุ่มภูมินามที่คาดว่าอ้างอิงถึงสถานที่เป้าหมาย

4) การนำตัวแปรที่เป็นคุณลักษณะทางตำแหน่งที่สัมพันธ์กับเวลามาเป็นหนึ่งในการประมวลผล เช่น “เมื่อก่อนตลาดสามย่านตั้งอยู่บริเวณห้างสามย่านมิตรทาวน์ในปัจจุบัน ซึ่งในปี 2566 ไปใช้บริการได้ที่ U-center” จากข้อความข้างต้นจะพบว่ามีคำบ่งบอกสถานที่คือ ตั้งอยู่ซึ่งแบบจำลองอาจจะไปดึงตำแหน่งของสามย่านมิตรทาวน์ และ ที่ U-center โดยจากประโยคข้างต้นจะได้ภูมินามออกมาซึ่งเป็นคำที่อยู่หลังคำว่า ‘ตั้งอยู่’ และ ‘ที่’ แต่หากมีคำว่า ‘เมื่อก่อน’ ซึ่งเป็นคำบอกช่วงเวลา จะทำให้ชื่อภูมินามที่สกัดออกมาคือ U-center เพียงคำเดียว

5) การนำเทคนิคการจำแนกข้อความ (text classification) มาใช้เพื่อจำแนกข้อความที่เป็นที่อยู่ (address) โดยเฉพาะ

6) การนำคำที่บ่งบอกระยะทางมาใช้งาน เช่น วัดจันทรสโมสรอยู่ในเขตรับผิดชอบของเขตดุสิต ห่างจากตลาดสามเสนไปประมาณ 500 เมตร หรือการนำชั้นข้อมูลถนนเข้ามาเป็นตัวช่วยในการระบุตำแหน่ง โดยเมื่อมีการสกัดเอาข้อมูลประเภท ROAD ออกมาจากข้อความ เช่น “ตั้งต้นจากบริเวณปากซอยลาดพร้าว 31 เข้าไปประมาณ 500 เมตรจะเจอร้านห่านเอ๊ยท่าดินแดงอยู่ทางซ้ายมือ” จากประโยคข้างต้นจะพบว่ามีการนำถนนมาอ้างอิงตำแหน่งพร้อมระยะทางในกำกับไว้ด้วยโดยประมาณ โดยแสดงตัวอย่างตามรูปที่ 41

รูปที่ 39 แสดงตัวอย่างการนำข้อมูลชั้นถนนมาใช้ร่วมกับการประมาณตำแหน่งภูมินาม

บรรณานุกรม

Ajao, O., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science,* 41(6), 855-864. doi:10.1177/0165551515602847

Backstrom, L., Sun, E., & Marlow, C. (2010). *Find me if you can: improving geographical prediction with social and spatial proximity.* Paper presented at the Proceedings of the 19th international conference on World wide web.

Cadorel, L., Blanchi, A., & Tettamanzi, A. G. B. (2021). *Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text*. Paper presented at the Proceedings of the 11th on Knowledge Capture Conference, Virtual Event, USA. https://doi.org/10.1145/3460210.3493547

Cetl, V., Kliment, T., & Jogun, T. (2018). A comparison of address geocoding techniques – case study of the city of Zagreb, Croatia. *Survey Review,* 50(359), 97-106. doi:10.1080/00396265.2016.1252517

Chambers, J. K., & Trudgill, P. (1998). *Dialectology*: Cambridge University Press.

Chanlekha, H., & Kawtrakul, A. (2004). *Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information*.

Chanlekha, H., Kawtrakul, A., Varasrai, P., & Mulasas, I. (2002). *Statistical and Heuristic Rule Based Model for Thai Named Entity Recognition*.

Chasin, R., Woodward, D., Witmer, J., & Kalita, J. J. T. C. J. (2014). Extracting and displaying temporal and geospatial entities from articles on historical events. 57(3), 403-426.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. J. a. p. a. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation.

Chormai, P., Prasertsom, P., & Rutherford, A. (2019). *AttaCut: A Fast and Accurate Neural Thai Word Segmenter*.

Cosentino, R., Balestriero, R., Bahroun, Y., Sengupta, A., Baraniuk, R., & Aazhang, B. (2022). Spatial Transformer K-Means. *arXiv preprint arXiv:*2202.07829.

Daniel, G. (2013). *Principles of artificial neural networks* (Vol. 7): World Scientific.

Devkota, B., Miyazaki, H., Witayangkurn, A., & Kim, S. (2019). Using Volunteered Geographic Information and Nighttime Light Remote Sensing Data to Identify Tourism Areas of Interest. *Sustainability,* 11, 4718. doi:10.3390/su11174718

Dredze, M., Paul, M. J., Bergsma, S., & Tran, H. C. A. T. G. S. w. A. t. P. H. (2013). Carmen: A Twitter Geolocation System with Applications to Public Health. *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI),* 23, 5.

Eisenstein, J. (2019). *Introduction to Natural Language Processing*: MIT Press.

Finkel, J. R., Grenager, T., & Manning, C. (2005). *Incorporating non-local information into information extraction systems by gibbs sampling.* Paper presented at the Proceedings of the 43rd annual meeting on association for computational linguistics.

Gonçalves, B., & Sánchez, D. (2014). Crowdsourcing Dialect Characterization through Twitter. *PLOS ONE,* 9(11), e112074. doi:10.1371/journal.pone.0112074

Google. (2023). Place Types. Retrieved from https://developers.google.com/maps/documentation/places/web-service/supported_types

Gridach, M., & Haddad, H. (2017). *Arabic Named Entity Recognition : A Bidirectional GRU-CRF Approach*.

Gritta, M., Pilehvar, M. T., & Collier, N. (2019). A pragmatic guide to geoparsing evaluation. *Language Resources and Evaluation*. doi:10.1007/s10579-019-09475-3

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation,* 52(2), 603-623. doi:10.1007/s10579-017-9385-8

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., . . . Sciences, E. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. 368(1925), 3875-3889.

Hahmann, S., & Burghardt, D. J. I. J. o. G. I. S. (2013). How much information is geospatially referenced? Networks and cognition. 27(6), 1171-1189.

Hammerton, J. (2003). *Named entity recognition with long short-term memory.* Paper presented at the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4.

Han, B. (2014). *Improving the utility of social media with natural language processing.* (Doctor of Philosophy). University of Melbourne, Melmourne. (134 - 135)

Han, B., Cook, P., & Baldwin, T. (2012). *Geolocation Prediction in Social Media Data by Finding Location Indicative Words*.

Here. (2023). POI categories. Retrieved from https://developer.here.com/documentation/map-tile/dev_guide/topics/resource-meta-pois.html

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation,* 9, 1735-1780. doi:10.1162/neco.1997.9.8.1735

Howard, J., & Ruder, S. . (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv* : 1801.06146.

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems,* 59, 244-255. doi:https://doi.org/10.1016/j.compenvurbsys.2015.12.003

Iu, D., & Zou, X. (2018, 14-16 Dec. 2018). *Sequence Labeling of Chinese Text Based on Bidirectional Gru-Cnn-Crf Model.* Paper presented at the 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP).

Jin, X., & Han, J. (2010). K-Medoids Clustering. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 564-565). Boston, MA: Springer US.

Kemp, S. (2021). datareportal Retrieved from https://datareportal.com/reports/digital-2021-thailand

Kolatch, E. J. P. i. a. o. t. W. (2001). Clustering algorithms for spatial databases: A survey. 1-22.

Kuai, X., Guo, R., Zhang, Z., He, B., Zhao, Z., & Guo, H. (2020). Spatial Context-Based Local Toponym Extraction and Chinese Textual Address Segmentation from Urban POI Data. *ISPRS International Journal of Geo-Information,* 9(3). doi:10.3390/ijgi9030147

Lieberman, M. D., & Samet, H. (2011). *Multifaceted toponym recognition for streaming news.* Paper presented at the Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.

Lingad, J., Karimi, S., & Yin, J. (2013). *Location extraction from disaster-related microblogs.* Paper presented at the Proceedings of the 22nd international conference on world wide web.

Lowphansirikul, C. (2018). Clustering — DBSCAN. Retrieved from https://medium.com/@artificialcc/clustering-dbscan-%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3-116b5d5c9873

Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., & Nutanong, S. (2021). Wangchanberta: Pretraining transformer-based thai language models. *arXiv preprint arXiv:*2101.09635.

Ma, K., Tan, Y., Xie, Z., Qiu, Q., & Chen, S. (2022). Chinese toponym recognition with variant neural structures from social media messages based on BERT methods. *Journal of Geographical Systems,* 24(2), 143-169. doi:10.1007/s10109-022-00375-9

Mai, G., Janowicz, K., Gao, S., & Hu, Y. (2017). ADCN: An Anisotropic Density-Based Clustering Algorithm for Discovering Spatial Point Patterns with Noise. *Transactions in GIS,* 22. doi:10.1111/tgis.12313

Manoruang, D., & Asavasuthirakul, D. (2019a). Quality analysis of online geocoding services for Thai text addresses. *Engineering and Applied Science Research,* 46(2), 11. Retrieved from https://ph01.tci-thaijo.org/index.php/easr/article/view/140887

Manoruang, D., & Asavasuthirakul, D. (2019b). *A Tax-Map-Based Address Point Data Model for Geocoding Thai addresses*. Paper presented at the The 4th International Conference on Digital Arts, Media and Technology and 2nd ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering.

Melo, F., & Martins, B. (2017). Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS,* 21(1), 3-38. doi:10.1111/tgis.12212

Mishra, P., & Sarawadekar, K. (2019, 17-20 Oct. 2019). *Polynomial Learning Rate Policy with Warm Restart for Deep Neural Network.* Paper presented at the TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON).

Murnane, W. (2010). Improving accuracy of named entity recognition on social media data.

Nanda, A., Barik, R. C., & Bakshi, S. (2023). SSO-RBNN driven brain tumor classification with Saliency-K-means segmentation technique. *Biomedical Signal Processing and Control,* 81. doi:10.1016/j.bspc.2022.104356

Nostra. (2022). Points of interest (Landmark). Retrieved from https://www.nostramap.com/geodata/

Nur Yasir Utomo, M., Adji, T., & Ardiyanto, I. (2018). *Geolocation prediction in social media data using text analysis: A review*.

Omran, M., Engelbrecht, A., & Salman, A. (2007). An overview of clustering methods. *Intell. Data Anal.,* 11, 583-605. doi:10.3233/IDA-2007-11602

Owusu, C., Lan, Y., Zheng, M., Tang, W., & Delmelle, E. (2017). Geocoding Fundamentals and Associated Challenges. In (pp. 41-62).

Pavalanathan, U., & Eisenstein, J. (2015). Confounds and Consequences in Geotagged Twitter Data. *arXiv,* 1506.02275.

Polpanumas, C. (2019). ULMFit Language Modeling, Text Feature Extraction and Text Classification in Thai Language. Created as part of pyThaiNLP. Retrieved from https://github.com/cstorm125/thai2fit

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012). Supervised Text-based Geolocation Using Language Models on an Adaptive Grid.

Sae-Lim, T. (2021). *Virtual Adviserial Training With Weighted Token Reputation in Text Classification.* (Master). Chulalongkorn, Chulalongkorn. Retrieved from http://cuir.car.chula.ac.th/handle/123456789/80380

Sagcan, M., & Karagoz, P. (2015, 14-17 Nov. 2015). *Toponym Recognition in Social Media for Estimating the Location of Events.* Paper presented at the 2015 IEEE International Conference on Data Mining Workshop (ICDMW).

Santos, J., Anastácio, I., & Martins, B. (2015). Using machine learning methods for disambiguating place references in textual documents. *GeoJournal,* 80(3), 375-392. doi:10.1007/s10708-014-9553-y

Santos, R., Murrieta-Flores, P., Calado, P., & Martins, B. (2018). Toponym matching through deep neural networks. *International Journal of Geographical*

Information Science, 32(2), 324-348. doi:10.1080/13658816.2017.1390119

Sobhana, N., Mitra, P., & Ghosh, S. J. I. J. o. C. A. (2010). Conditional random field based named entity recognition in geological text. 1(3), 143-147.

Steinberger, R., Pouliquen, B., & Van Der Goot, E. (2013). An introduction to the Europe Media Monitor family of applications. *Information Access in a Multilingual World - Proceedings of the SIGIR* 2009 *Workshop (SIGIR-CLIR'*2009), 1-8.

Subasi, A. (2020). Chapter 7 - Clustering examples. In A. Subasi (Ed.), *Practical Machine Learning for Data Analysis Using Python* (pp. 465-511): Academic Press.

Sureja, N., Chawda, B., & Vasant, A. (2022). An improved K-medoids clustering approach based on the crow search algorithm. *Journal of Computational Mathematics and Data Science,* 3, 100034. doi:https://doi.org/10.1016/j.jcmds.2022.100034

Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. *Foundations*

Trends® in Machine Learning, 4(4), 267-373.

Tang, J., Liu, F., Wang, Y., & Wang, H. (2015). Uncovering urban human mobility from large scale taxi GPS data. *Physica A: Statistical Mechanics and its Applications,* 438, 140-153. doi:https://doi.org/10.1016/j.physa.2015.06.032

Tao, L., Xie, Z., Xu, D., Ma, K., Qiu, Q., Pan, S., & Huang, B. (2022). Geographic Named Entity Recognition by Employing Natural Language Processing and an Improved BERT Model. *ISPRS International Journal of Geo-Information,* 11(12). doi:10.3390/ijgi11120598

Thattinaphanich, S., & Prom-On, S. (2019). *Thai Named Entity Recognition Using Bi-LSTM-CRF with Word and Character Representation*.

Tie, J., Chen, W., Sun, C., Mao, T., & Xing, G. (2019). The application of agglomerative hierarchical spatial clustering algorithm in tea blending. *Cluster Computing,* 22(3), 6059-6068. doi:10.1007/s10586-018-1813-z

Tirasaroj, N., & Aroonmanakun, W. (2009, 20-22 Oct. 2009). *Thai named entity recognition based on conditional random fields.* Paper presented at the 2009 Eighth International Symposium on Natural Language Processing.

Tretasayuth, N. (2017). *Machine Reading Comprehension for Questions with Multiple*

Relationships Using Deep Learning. (Master of Engineering). Chulalonkorn University, Retrieved from http://cuir.car.chula.ac.th/handle/123456789/60126

Ullman, S., Poggio, T., Harari, D., Zysman, D., & Seibert, D. (2014). Unsupervised learning Custering. 54. Retrieved from http://www.mit.edu/~9.54/fall14/slides/Class13.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). *Attention Is All You Need*. Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://arxiv.org/abs/1706.03762

Wang, J. (2009). Geocoding Data Analysis and Processing in Relational Databases. *Communications of the IIMA,* 9(3), 11. Retrieved from https://scholarworks.lib.csusb.edu/ciima/vol9/iss3/8

Wang, J., Hu, Y., & Joseph, K. (2020). NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS,* 24(3), 719-735. doi:https://doi.org/10.1111/tgis.12627

Wang, S., Zhang, X., Ye, P., & Du, M. (2018). Deep Belief Networks Based Toponym Recognition for Chinese Text. *ISPRS International Journal of Geo-Information,* 7(6). doi:10.3390/ijgi7060217

Williams, E., Gray, J., & Dixon, B. (2017). Improving geolocation of social media posts. *Pervasive and Mobile Computing,* 36, 68-79. doi:https://doi.org/10.1016/j.pmcj.2016.09.015

Wilson, J., & Knoblock, C. (2007). From text to geographic coordinates: The current state of geocoding. *Urisa Journal,* 19, 33-46.

Xu, D., Cui, P., Zhu, W., & Yang, S. (2014). *Find you from your friends: Graph-based residence location prediction for users in social media.* Paper presented at the 2014 IEEE international conference on multimedia and expo (ICME).

Y. Liu, M. O., N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov. (2019). A Robustly Optimized BERT Pretraining Approach. *arXiv,* 1907.11692.

Zandbergen, P. (2008). Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National Standard for Spatial Data Accuracy. *T. GIS,* 12, 103-130. doi:10.1111/j.1467-9671.2008.01088.x

Zhang, X., Zhu, S., & Zhang, C. (2012). Annotation of Geographical Named Entitiesin Chinese Text. 41, 115-120.

Zheng, X., Han, J., & Sun, A. (2018). A Survey of Location Prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering,* 30(9), 19.

Zhu, X. (2009). "*Conditional Random Fields." In CS769 Spring 2009 Advanced Natural Language Processing*.

กระทรวงมหาดไทย, ก. (2562). จำนวนราษฎรทั่วราชอาณาจักร ตามหลักฐานการทะเบียนราษฎร ณ วันที่ 31 ธันวาคม ๒๕๖๒.

ประวัติผู้เขียน

ชื่อ-สกุล	ธุวชิต แฉล้มเขตต์
วัน เดือน ปี เกิด	11 มิถุนายน 2528
สถานที่เกิด	กาฬสินธุ์
วุฒิการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาระบบสารสนเทศปริภูมิทางวิศวกรรม จุฬาลงกรณ์มหาวิทยาลัย พ.ศ. 2555 วิทยาศาสตร์บัณฑิต สาขาเทคโนโลยีภูมิสารสนเทศ มหาวิทยาลัยบูรพา พ.ศ. 2551
ที่อยู่ปัจจุบัน	40/661 ม.กรุงทอง 5 ซ.นวมินทร์ 111 ถ.นวมินทร์ แขวงนวมินทร์ เขตบึงกุ่ม กรุงเทพฯ 10240
ผลงานตีพิมพ์	Chalamkate, T., Tinnachote, C., Rutherford, A.T. (2022). The Development of Geo-Names Extraction from Twitter Texts Data by Conditional Random Fields. Journal of Engineering and Digital Technology (JEDT)., Thai-Nichi Institute of Technology., 10(2), 97-109.