การพัฒนาระบบสกัดองค์ความรู้จากตำรายาอีสานโบราณโดยใช้เทคนิคการรู้จำชื่อเฉพาะแบบ Bi-LSTMs

Development of a knowledge extraction system from ancient Isan medical texts using Bi-LSTM-based named entity recognition

จินตนา พลศรี[1]*

Jintana Polsri[1]*

Received: 15 August 2025 ; **Revised**: 15 October 2025 ; **Accepted**: 27 October 2025

บทคัดย่อ

ตำรายาอีสานโบราณบันทึกองค์ความรู้ด้านการแพทย์พื้นบ้านที่สะท้อนถึงภูมิปัญญาท้องถิ่นในการดูแลสุขภาพของคนอีสานในอดีต ซึ่งยังคงได้รับการสืบทอดและนำมาประยุกต์ใช้ในปัจจุบัน โดยเฉพาะสมุนไพรไทยหลายชนิดถูกนำมาบูรณาการใช้ร่วมกันกับยาแผนปัจจุบันอย่างมีประสิทธิภาพและปลอดภัยยิ่งขึ้น รวมถึงศักยภาพในการพัฒนาเป็นผลิตภัณฑ์ที่มีมูลค่าเพิ่มในภาคอุตสาหกรรม อย่างไรก็ตาม ตำรายาอีสานโบราณมีลักษณะเฉพาะด้านภาษาและคำศัพท์เฉพาะทางที่ซับซ้อนจึงเป็นเรื่องยากต่อการอ่านโดยเฉพาะสำหรับผู้ที่ไม่มีพื้นฐานทางภาษาและเนื้อหา งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาระบบสกัดองค์ความรู้จากตำรายาอีสานโบราณแบบอัตโนมัติ ประกอบด้วย ชื่อยา ชื่อโรค และชื่อสมุนไพร วิธีดำเนินการวิจัยประกอบด้วย การรวบรวมข้อมูลและฝึกฝนโมเดลรู้จำชื่อเฉพาะด้วยเทคนิค Bi-LSTMs พัฒนาระบบสกัดองค์ความรู้ด้วยภาษา Python และจัดเก็บองค์ความรู้ที่สกัดได้ในฐานข้อมูล SQLiteStudio ประสิทธิภาพของระบบประเมินจากความแม่นยำของโมเดลด้วย F1-score แยกรายคลาส และแบบค่าเฉลี่ยถ่วงน้ำหนักทุกคลาส ประเมินความพึงพอใจของผู้ใช้งานที่มีต่อการออกแบบและพัฒนาระบบและการออกแบบและพัฒนาฐานข้อมูล โดยกลุ่มผู้ประเมินประกอบด้วยผู้เชี่ยวชาญด้านการออกแบบและพัฒนาโปรแกรมคอมพิวเตอร์ นักภาษาโบราณ และผู้ใช้งานทั่วไป รวมทั้งสิ้น 12 คน การประเมินประสิทธิภาพของโมเดล ค่าเฉลี่ย F1-score แบบถ่วงน้ำหนักอยู่ที่ร้อยละ 94 ระดับรายคลาส ได้แก่ ชื่อสมุนไพร ร้อยละ 96 ชื่อโรค ร้อยละ 82 และชื่อยา ร้อยละ 7 สำหรับการประเมินความพึงพอใจของผู้ใช้งานที่มีต่อระบบพบว่า การออกแบบและพัฒนาระบบอยู่ในระดับมากที่สุด ($\bar{x}$= 4.44, SD = 0.57) และการออกแบบฐานข้อมูลอยู่ในระดับมาก ($\bar{x}$= 4.05, SD = 0.50) ผลการวิจัยชี้ให้เห็นความเป็นไปได้ในการใช้เทคนิคการรู้จำชื่อเฉพาะเพื่อพัฒนาระบบสกัดองค์ความรู้จากตำรายาอีสานโบราณแบบอัตโนมัติได้อย่างมีประสิทธิภาพ

คำสำคัญ: ไบ-แอลเอสทีเอ็ม, การรู้จำชื่อเฉพาะ, การสกัดองค์ความรู้, ตำรายาอีสานโบราณ

Abstract

Ancient Isan medicinal manuscripts encapsulate traditional medical knowledge that reflects the indigenous wisdom of the Isan people in healthcare practices of the past. This body of knowledge continues to be preserved and applied in contemporary contexts, particularly through the effective and safe integration of various Thai herbs with modern medicine. Furthermore, these herbs hold significant potential for development into value-added products within the industrial sector. Nonetheless, the interpretation of ancient Isan medicinal manuscripts remains challenging due to their distinctive linguistic features and domain-specific terminologies, especially for individuals lacking foundational knowledge of the language and subject matter. This research aims to develop an automated knowledge extraction system from ancient Isan medicinal manuscripts, focusing on the identification of medicine names, disease names,

[1] สาขาวิชาเทคโนโลยีสารสนเทศและธุรกิจดิจิทัล, คณะบริหารธุรกิจ, มหาวิทยาลัยเทคโนโลยีราชมงคลกรุงเทพ

[1] Information Technology and Digital Business Department, Faculty of Business Administration, Rajamangala University of Technology Krungthep

* Corresponding author: Jintana Polsri, Email: jintana.p@mail.rmutk.ac.th

and herbal names. The research methodology includes data collection and the training of a named entity recognition model using the Bi-LSTMs technique. The knowledge extraction system is implemented using the Python programming language, and the extracted knowledge is stored in an SQLiteStudio database. The system's performance is evaluated based on the model's precision using F1-scores, both per class and as a weighted average across all classes. User satisfaction is also assessed regarding the system and database design and development. The evaluation group consists of 12 participants, including experts in computer program design and development, specialists in ancient languages, and general users. The model's performance was evaluated using the weighted average F1-score, which reached 94%. The class-wise F1-scores were 96% for herbal names, 82% for symptom names, and 7% for medicine names. User satisfaction with the system was also assessed. The results indicated that the system design and development were rated at the highest level ($\bar{x}$= 4.44, SD = 0.57), while the database design was rated at a high level ($\bar{x}$= 4.44, SD = 0.57). These findings demonstrate the potential of applying named entity recognition techniques to effectively develop an automated knowledge extraction system for ancient Isan medicinal texts.

Keywords: Bi-LSTM, named entity recognition, knowledge extraction, ancient Isan medical texts

บทนำ

ประเทศไทยให้ความสำคัญกับการใช้ฐานความรู้เป็นกลไกในการขับเคลื่อนพัฒนาเศรษฐกิจและสังคมอย่างต่อเนื่อง ปรากฏชัดในแผนพัฒนาเศรษฐกิจและสังคมแห่งชาติที่มุ่งเน้นการนำเอาภูมิปัญญาท้องถิ่นมาใช้ในการเพิ่มศักยภาพเศรษฐกิจและสุขภาพ โดยเฉพาะการต่อยอดภูมิปัญญาดังกล่าวเพื่อนำมาใช้เพิ่มมูลค่าให้กับสินค้าและบริการ ตลอดจนประยุกต์ใช้ในภาคอุตสาหกรรมหลายประเภท (สำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ, 2555) รวมถึงใช้ขับเคลื่อนให้ประเทศไทยเป็น Medical & Wellness Hub (สำนักเลขาธิการนายกรัฐมนตรี ทำเนียบรัฐบาล, 2567) ในขณะเดียวกันรัฐธรรมนูญแห่งราชอาณาจักรไทย พ.ศ. 2560 ได้กำหนดให้รัฐมีหน้าที่ส่งเสริมและสนับสนุนการพัฒนาภูมิปัญญาด้านการแพทย์แผนไทยและนวัตกรรมยาสมุนไพรให้เกิดประโยชน์สูงสุด โดยมีเป้าหมายยกระดับการสร้างนวัตกรรมให้กับประเทศ เช่น การนำสมุนไพรไทยมาพัฒนาและบูรณาการใช้ในอุตสาหกรรมยา เวชภัณฑ์ เครื่องสำอาง อาหารและเครื่องดื่มเพื่อสุขภาพ เป็นต้น นโยบายดังกล่าวสะท้อนให้เห็นถึงบทบาทของสมุนไพรไทยทั้งในมิติของการเป็นแพทย์ทางเลือก (Alternative Medicine) ที่สนับสนุนการใช้รักษาโรคควบคู่ไปกับการแพทย์แผนตะวันตก และในมิติของการขับเคลื่อนเศรษฐกิจเชิงสร้างสรรค์ทั้งในระดับชาติและระดับสากล ด้วยเหตุนี้ การศึกษาภูมิปัญญาท้องถิ่นในด้านการใช้สมุนไพรจึงมีความสำคัญเป็นอย่างยิ่ง ซึ่งไม่เพียงแต่ช่วยเสริมสร้างความมั่นคงด้านสุขภาพและสาธารณสุขของประเทศ แต่ยังสนับสนับสนุนการพึ่งพาตนเองที่มีความยั่งยืนในระยะยาว

ตำรายาอีสานเป็นแหล่งข้อมูลสำคัญที่บันทึกองค์ความรู้ด้านการแพทย์พื้นบ้านและการใช้สมุนไพรในการรักษาโรคของชาวอีสานในอดีตเอาไว้เป็นจำนวนมาก ไม่เพียงเฉพาะเนื้อหาที่เกี่ยวข้องกับตำรับยาเท่านั้น แต่ยังมีเนื้อหาอื่น ๆ ที่ไม่ใช่ตำรับยารวมอยู่ด้วย โดยตำรับยาอีสานสามารถจำแนกออกได้เป็น 3 ประเภทหลัก ได้แก่ ตำรับยาที่ใช้สมุนไพรเป็นองค์ประกอบหลักในการรักษา ตำรับยาที่ใช้คาถาอาคมในการรักษา และตำรับยาที่ผสมผสานทั้งสมุนไพรและคาถา นอกจากนี้ยังมีเนื้อหาประเภทอื่น ๆ ที่ไม่ใช่ตำรับยาโดยตรง เช่น วิธีการเก็บสมุนไพร ขั้นตอนหรือพิธีกรรมที่ต้องปฏิบัติก่อน ระหว่างและหลังการรักษา การกำหนดฤกษ์ยามที่เป็นมงคลสำหรับประกอบกิจกรรมที่เกี่ยวข้องกับการรักษาโรค หรือพิธีกรรมอื่น เช่น ฤกษ์ตัดผม (จินตนา พลศรี, 2564) เป็นต้น ข้อมูลเหล่านี้สะท้อนให้เห็นถึงความซับซ้อนขององค์ความรู้ทางการแพทย์พื้นบ้านอีสานที่เชื่อมโยงทั้งมิติทางกายภาพ ความเชื่อ และวัฒนธรรมท้องถิ่น

ลักษณะการบันทึกตำรายาอีสานโบราณมีความซับซ้อนทั้งในด้านภาษาและคำศัพท์เฉพาะทาง อาทิ ชื่อยา ชื่อโรค หรืออาการของโรค ชื่อสมุนไพร อุปกรณ์ปรุงยา ตลอดจนองค์ประกอบของตำรับยาที่มีความหลากหลาย ตำรับยาบางตำรับอาจมีหลายขนาน เพื่อใช้ตามลักษณะอาการของโรค เพศ อายุ ของผู้ป่วย หรือใช้ยาตามช่วงเวลาที่เกิดอาการเจ็บป่วย เช่น เช้า สาย บ่าย เย็น อีกทั้งตำรับยาบางตำรับยังมีข้อกำหนดเฉพาะ เช่น กำหนดวัน เวลา พิธีกรรม บทสวด หรือคาถา เมื่อต้องไปเก็บสมุนไพร ไปจนถึงขั้นตอนการปรุงยา ใช้ยา กินยา และเนื้อหาบันทึกเป็นภาษาถิ่นอีสานโบราณที่มีคาถาบาลีสอดแทรก (จินตนา พลศรี และคณะ, 2565) ทำให้ตำรายาอีสานโบราณมีลักษณะเฉพาะที่ซับซ้อนยากต่อการอ่านหรือทำความเข้าใจ ส่งผลให้การสกัดองค์ความรู้จากตำรายาอีสานโบราณต้องอาศัยผู้เชี่ยวชาญด้านสมุนไพรพื้นบ้าน

หรือหมอยาพื้นบ้านอีสานมาช่วยอ่าน วิเคราะห์ และระบุองค์ประกอบต่าง ๆ ในตำรับยา นอกจากนี้ รูปแบบการบันทึกข้อมูลในแบบหนังสือตำรายา หรือไฟล์ข้อความทั่วไป (Text file) ซึ่งเป็นข้อมูลแบบไม่เป็นโครงสร้าง (Unstructured text) ยังเป็นอุปสรรคสำคัญที่ทำให้การสืบค้นข้อมูล เช่น ชื่อโรค ชื่อสมุนไพรที่ใช้รักษา มีความยุ่งยากและใช้เวลานาน

การรู้จำชื่อเฉพาะ หรือ นิพจน์ระบุนาม (Named Entity Recognition: NER) เป็นเครื่องมือที่ถูกนำมาใช้อย่างแพร่หลายในงานประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) เพื่อสกัดสารสนเทศหรือสรุปข้อมูลจากข้อความที่ไม่มีโครงสร้าง การสกัดชื่อเฉพาะ (Proper Name) ของคน สัตว์ สิ่งของ สถานที่ ชื่อองค์กร รวมถึงข้อมูลเชิงเวลาและตัวเลข (Li *et al*., 2020) NER ถูกนำไปประยุกต์ใช้ในงานอย่างหลากหลาย เช่น การแปลภาษาแบบอัตโนมัติ การสืบค้นข้อมูล การตอบคำถาม ใคร ทำอะไร เมื่อไร ที่ไหน การสรุปข้อมูล การสกัดและดึงข้อมูลชื่อ สถานที่ วันที่ เวลา และตัวเลข เป็นต้น (Mansouri *et al*., 2008; Chalamkate *et al*., 2022) ซึ่งเทคนิควิธีการรู้จำชื่อเฉพาะแบ่งตามวิธีสร้างกฎและการเรียนรู้แบ่งเป็น 3 กลุ่ม (Yadav & Bethard, 2018) ได้แก่ วิธีที่ 1 การใช้กฎ (Rule-based) หรือฐานความรู้ (Knowledge-based) เป็นวิธีที่อาศัยผู้เชี่ยวชาญในการสร้างกฎจากพจนานุกรมชื่อเฉพาะและองค์ความรู้ทางภาษาศาสตร์ โดยความแม่นยำของวิธีนี้จะขึ้นอยู่กับความครอบคลุมของกฎที่สร้างขึ้น วิธีที่ 2 การเรียนรู้ด้วยเครื่อง (Machine Learning-based) ที่ใช้สถิติและการเรียนรู้ความสัมพันธ์จากคลังข้อความ (Corpus) เช่น เทคนิค Decision Tree, Support Vector Machine (SVM), Hidden Markov Model (HMM), Maximum Entropy, Maximum Entropy Markov Models (MEMMs), Condition Random Field (CRFs) และ Long Short-Term Memory (LSTM) เป็นต้น วิธีนี้ลดภาระในการสร้างกฎด้วยมือ แต่ประสิทธิภาพจะขึ้นอยู่กับขนาดของคลังข้อความที่ใช้ฝึกโมเดล และในบางโดเมนอาจขาดแคลนคลังข้อความสำหรับใช้ฝึกโมเดล และวิธีที่ 3 วิธีผสม (Hybrid-based) ได้แก่ การใช้กฎร่วมกับการเรียนรู้ด้วยเครื่อง เพื่อผสานข้อดีของการใช้กฎและการเรียนรู้ด้วยเครื่อง ลดข้อจำกัดและความผิดพลาดของแต่ละวิธี ตัวอย่างเช่น การใช้เทคนิค CRFs ร่วมกับกฎทางภาษาศาสตร์ (Zhang & Wang, 2023) การใช้ CRFs ร่วมกับ Bi-LSTMs (Kuran *et al*., 2024) เป็นต้น

การบันทึกตำรายาอีสานโบราณและไวยากรณ์ภาษาอีสานโดยทั่วไปแม้ว่าจะมีลักษณะคล้ายคลึงกับเอกสารข้อความภาษาไทยสมัยใหม่ คือ เขียนข้อความต่อเนื่องกันโดยไม่เว้นวรรคระหว่างคำ และรูปแบบการลำดับเนื้อหาจะดูเหมือนว่าค่อนข้างมีแบบแผน กล่าวคือ ตำรับยาแต่ละตำรับมักจะเขียนบรรยายด้วยการบอก ชื่อยา/ชื่อโรค/อาการ และตามมาด้วยรายการชื่อสมุนไพรที่ใช้รักษา วิธีเตรียมยา และวิธีใช้ยา เป็นต้น แต่ยังคงมีลักษณะเฉพาะที่ซับซ้อนและอ่านยากมากกว่า โดยเฉพาะในส่วนของชื่อเฉพาะต่าง ๆ เช่น ชื่อยา ชื่อสมุนไพร หรืออาการของโรค ซึ่งส่วนใหญ่มักประกอบด้วยคำประสมที่เกิดจากคำโดดอย่างน้อย 2 คำ มาประสมกัน ทั้งในรูปแบบที่มีความหมายแตกต่างกัน หรือในบางกรณีไม่มีความหมายแยกส่วนอย่างชัดเจน เช่น “ยาออกทกทื้น” “ยางัวควยตาส่อน” “ยาฟกยาไค่” “ผักหวานบ้าน” “ฮากหมากเขือ” “ฮากตำนิน” (รากตำลึง) “ต้นดอกเข็ม” “ดูกคันคาก” (กระดูกคางคก) เป็นต้น นอกจากนี้การสะกดคำตามเสียงภาษาอีสานที่ไม่มีแบบแผนข้อกำหนดในการสะกดคำที่ตายตัวทำให้คำเดียวกันสะกดได้หลายแบบ และมีความยาวหลายพยางค์ที่ยากต่อการระบุขอบเขตของคำที่ชัดเจน ตัวอย่างโครงสร้างและความสัมพันธ์ขององค์ประกอบต่าง ๆ ในตำรายาอีสานโบราณ ดัง Figure 1

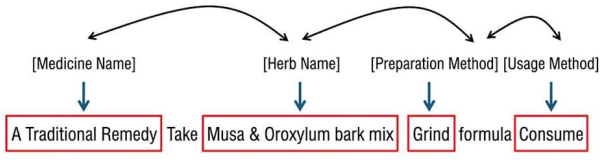


Figure 1 Structure of an ancient Isan medicinal recipe

นอกจากนี้การแบ่งแยกคำบางคำ (Word Segmentation) ให้ถูกต้องนั้นยังขึ้นอยู่กับบริบท เช่นคำว่า “น้ำอุ่น” บางบริบทแยกเป็น “น้ำ” (นาม) “อุ่น” (กริยา) หมายถึงเอาไปทำให้อุณหภูมิสูงขึ้น) และบางบริบทแบ่งแยกเป็นคำเดียว “น้ำอุ่น” (นาม) ด้วยเหตุนี้การสกัดองค์ความรู้จากตำรายา เช่น ชื่อยา ชื่อสมุนไพร หรือคำที่เกี่ยวข้องกับตำรับยาจึงมีความซับซ้อนและใช้เวลาในการประมวลผลสูง ไม่ว่าจะเป็นการดำเนินการโดยผู้เชี่ยวชาญซึ่งทำด้วยมือ หรือผ่านระบบอัตโนมัติที่ใช้กฎในการวิเคราะห์ ข้อจำกัดเหล่านี้สะท้อนให้เห็นถึงความจำเป็นในการพัฒนาเทคโนโลยีที่มีความสามารถในการทำความเข้าใจบริบทด้านภาษาและโครงสร้างเฉพาะของการเขียนตำรายาอีสานโบราณ เพื่อให้การสกัดองค์ความรู้มีประสิทธิภาพมากยิ่งขึ้น

ดังนั้นเพื่อให้ผู้ที่ไม่มีพื้นฐานความรู้ทั้งด้านภาษาอีสานและความรู้เฉพาะด้านตำรายาอีสานโบราณ สามารถเข้าถึงองค์ความรู้ในตำรายาอีสานโบราณได้ง่ายขึ้น อีกทั้งเพื่อสนับสนุนการพัฒนาเครื่องมือที่ช่วยสกัดองค์ความรู้จากตำรา

ยาอีสานโบราณ ซึ่งจัดเก็บอยู่ในรูปแบบข้อมูลแบบไม่มีโครงสร้าง มาจัดเก็บไว้ในรูปแบบที่มีโครงสร้างอย่างเป็นระบบ งานวิจัยนี้จึงมีเป้าหมายในการศึกษาและพัฒนาระบบสกัดองค์ความรู้จากตำรายาอีสานโบราณแบบอัตโนมัติ โดยองค์ความรู้ที่มุ่งเน้นในการสกัด ประกอบไปด้วย ชื่อยา ชื่อโรค และชื่อสมุนไพร พร้อมจัดเก็บไว้ในฐานข้อมูลเพื่อความสะดวกต่อการค้นคืน หรือการนำไปใช้ประโยชน์ในงานประมวลผลภาษาธรรมชาติ และงานด้านอื่น ๆ ต่อไป

กระบวนการวิจัยประกอบด้วยการฝึกโมเดลรู้จำชื่อเฉพาะด้วยเทคนิค Bi-Directional Long Short-Term Memory (Bi-LSTMs) ซึ่งเป็นเทคนิคที่มีข้อดีคือ สามารถเรียนรู้ลำดับของข้อมูลจากทั้งทิศทางซ้ายไปขวาและขวาไปซ้าย ซึ่งจะช่วยให้การวิเคราะห์บริบทของคำในตำรายาอีสานมีความแม่นยำมากขึ้น คำถามการวิจัยหลัก ได้แก่ (1) การรู้จำชื่อเฉพาะจากตำรายาอีสานโบราณด้วยเทคนิค Bi-LSTMs ให้ผลลัพธ์ในด้านความแม่นยำ (Precision) ความครอบคลุม (Recall) และค่าเฉลี่ยความแม่นยำ F1-score ของคลาสที่เป็นองค์ความรู้ที่ต้องการสกัดอยู่ในระดับใด? และ (2) ระบบสกัดองค์ความรู้จากตำรายาอีสานโบราณที่พัฒนาขึ้น สามารถสนับสนุนงานด้านการสกัดและจัดเก็บองค์ความรู้จากตำรายาอีสานโบราณแบบอัตโนมัติได้อย่างมีประสิทธิภาพเพียงใด?

การวิจัยนี้เป็นส่วนหนึ่งของโครงการวิจัยเรื่อง การออกแบบและพัฒนาระบบสกัดสารสนเทศด้านสมุนไพรจากตำรายาอีสานโบราณ และได้รับการรับรองจากคณะกรรมการจริยธรรมการวิจัยในมนุษย์ มหาวิทยาลัยนเรศวร เลขที่ 0033/2565 ลงวันที่ 16 สิงหาคม พ.ศ. 2566

วัตถุประสงค์ของงานวิจัย

เพื่อออกแบบและพัฒนาระบบสกัดองค์ความรู้จากตำรายาอีสานโบราณด้วยเทคนิคการรู้จำชื่อเฉพาะแบบ Bi-LSTMs

แนวคิดและหลักการของ Bi-LSTMs

Bi-LSTMs (Bidirectional Long Short-Term Memory) เป็นโมเดลเครือข่ายประสาทเทียมที่พัฒนาต่อยอดมาจาก LSTMs (Long Short-Term Memory) โดยการเพิ่มการเรียนรู้ข้อมูลที่มีลำดับแบบ 2 ทิศทาง ได้แก่ ทิศทางไปข้างหน้า (Forward) และย้อนกลับ (Backward) ทำให้โมเดลสามารถเรียนรู้บริบท (Context) ของข้อมูลที่มีลำดับ (Sequential Data) ได้ดียิ่งขึ้น (Schuster & Paliwal, 1997) สถาปัตยกรรม Bi-LSTM ดัง Figure 2 โมเดลประเภทนี้มักถูกใช้ในงานประมวลผลภาษาธรรมชาติที่ต้องเข้าใจบริบทของคำที่อยู่ก่อนหน้าและคำที่อยู่ถัดไปข้างหน้า เช่น การแปลภาษา การรู้จำชื่อเฉพาะ การรู้จำเสียง การจำแนกเอกสารข้อความ การวิเคราะห์ความรู้สึก (Sentiment Analysis) เป็นต้น

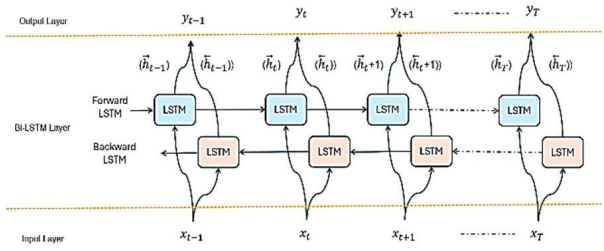


Figure 2 Bi-LSTM architecture

สถาปัตยกรรม Bi-LSTMs ประกอบด้วยชั้นหลายชั้น (Layers) ทำงานร่วมกันในการเรียนรู้ข้อมูลที่มีลำดับ กระบวนการทำงานของ Bi-LSTMs ข้อมูลที่ถูกนำเข้าสู่โมเดลในแต่ละช่วงเวลา (t) จะถูกแปลงให้อยู่ในรูปของเวกเตอร์ก่อนโดยใช้เทคนิค Word Embedding เช่น Word2Vec หรือ Glove เพื่อให้โมเดลสามารถเรียนรู้ความสัมพันธ์เชิงความหมายได้อย่างมีประสิทธิภาพ (Makarenkov *et al*., 2019) จากนั้นเวกเตอร์เหล่านั้นจะถูกนำเข้าสู่เครือข่าย LSTM 2 ชุด (Kannikaklang *et al*., 2024) ดังนี้

1) Forward LSTM ประมวลผลข้อมูลตามลำดับจากซ้ายไปขวา (จากเวลา t = 1 ถึง t = T และสร้างสถานะซ่อนก่อนหน้า (Forward Hidden States) ($\overrightarrow{h}_t$)

2) Backward LSTM ประมวลผลข้อมูลในทิศทางย้อนกลับจากขวาไปซ้าย (จากเวลา t = T ถึง t = 1) และสร้างสถานะซ่อนย้อนกลับ (Backward Hidden States) ($\overleftarrow{h}_t$)

เครือข่ายทั้งสองจะทำงานแบบคู่ขนาน เพื่อเรียนรู้บริบทจากข้อมูลทั้งที่อยู่ในลำดับข้างหน้าและหลังในลำดับของข้อความ โดยสถานะซ่อน (Hidden States) ของทั้งสองเครือข่ายจะถูกรวม (Concatenate) เข้าด้วยกันในแต่ละตำแหน่ง เพื่อสร้างผลลัพธ์สุดท้ายของทั้งสองเครือข่ายสำหรับใช้ในการประมวลผลขั้นตอนถัดไป

องค์ประกอบของเครือข่าย Bi-LSTMs ขยายแนวคิดมาจาก LSTM มาตรฐาน ที่มีองค์ประกอบภายในเหมือนกัน ดัง Figure 3 กล่าวคือ ภายในแต่ละหน่วย LSTM จะประกอบไปด้วยกลไกที่เรียกว่า ประตู (Gates) ซึ่งทำหน้าที่ควบคุมการไหลของข้อมูลในแต่ละช่วงเวลา (Time Step) ได้แก่ Forget Gate ทำหน้าที่ตัดสินใจว่าข้อมูลใดจากสถานะก่อนหน้าควรถูกลืมหรือเก็บไว้ Input Gate ทำหน้าที่ควบคุมว่าข้อมูลใหม่ใดควรถูกเพิ่มเข้ามาในสถานะซ่อนของลำดับเวลาปัจจุบัน และ Output Gate (o_t) ตัดสินใจว่าข้อมูลนั้นควรที่จะส่งออกไปเป็น

สถานะซ่อนของลำดับเวลาปัจจุบัน (h_t) หรือไม่ โดยใช้ค่าจากสถานะเซลล์ (C_t) ในการตัดสินใจ (Hochreiter & Schmidhuber, 1997)

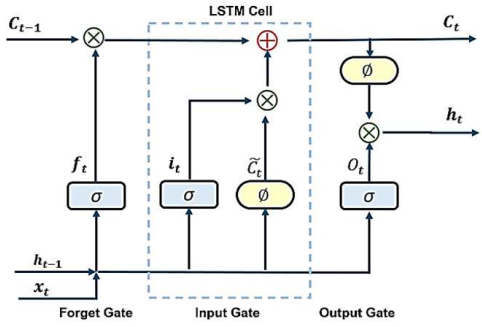


Figure 3 LSTM architecture

สถานะเซลล์ปัจจุบัน (C_t) จะเป็นกระบวนที่เกิดหลังกระบวนการคำนวณ 3 ส่วน ได้แก่ Forget Gate, Input Gate และ Output Gate โดย C_t จะถูกปรับปรุงสถานะจากการรวม Forget Gate และ Input Gate ดังสมการ (1)

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (1)$$

โดยที่

$f_t * C_{t-1}$ = ส่วนคำนวณเพื่อตัดสินใจที่จะลืม (Forget Gate) โดยค่าการตัดสินใจ f_t จะอยู่ระหว่าง 0 - 1

$i_t * \widetilde{C}_t$ = ส่วนควบคุมการนำข้อมูลเพิ่มเข้ามา โดย $\widetilde{C}_t$ คือค่าใหม่จะถูกนำเข้ามา และ i_t ค่าการตัดสินใจ

หลังจากการคำนวณและได้ C_t แล้ว จะเป็นขั้นตอนการคำนวณสถานะซ่อนของลำดับเวลาปัจจุบัน (h_t) โดยใช้ Output Gate (o_t) ดังสมการ (2)

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (2)$$

โดยที่

o_t = ค่าการตัดสินใจของ Output Gate

σ = ฟังก์ชัน Sigmoid

W_o = Weight Matrix ของ Output Gate

$[h_{t-1}, x_i]$ = Concatenation Input สถานะซ่อนก่อนหน้า และข้อมูลเข้าปัจจุบัน

b_o = Bias Vector ที่เรียนรู้ได้สำหรับ Output Gate

กลไกเหล่านี้ช่วยให้ LSTM จัดการปัญหา Vanishing Gradient ได้ดีขึ้น โดยโมเดลจะสามารถเรียนรู้และจดจำข้อมูลที่มีลำดับที่ยาว (Long-term Dependencies) ได้ดี

การคำนวณในเครือข่าย Forward LSTM มีองค์ประกอบหลักคือ Forget Gate, Input Gate และ Output Gate ที่ใช้ควบคุมการไหลของข้อมูลในแต่ละช่วงเวลา t มีดังนี้

Forget Gate ทำหน้าที่ตัดสินใจว่าค่าสถานะจากเซลล์ก่อนหน้า ($\vec{C}_{t-1}$) ควรถูกลืมหรือเก็บเอาไว้มากน้อยเพียงใด โดยพิจารณาจากข้อมูลนำเข้าในช่วงเวลาปัจจุบัน x_t และสถานะซ่อนก่อนหน้า (h_{t-1}) การคำนวณใช้สมการ (3)

$$\overrightarrow{f}_t = \sigma(W_f \cdot [\overrightarrow{h}_{t-1}, x_t] + b_f) \quad (3)$$

โดยที่

$\vec{f}_t$ = ค่าการตัดสินใจ (อยู่ระหว่าง 0-1)

W_f = น้ำหนักของ Forget Gate

$\vec{h}_{t-1}$= สถานะซ่อนของลำดับก่อนหน้า

x_t = ข้อมูลนำเข้าขณะนั้น

b_f = ค่าคงที่ (Bias)

σ = ฟังก์ชัน Sigmoid

โดยค่าการตัดสินใจ 0 = ลืม และ 1 = เก็บ

Input Gate ทำหน้าที่ตัดสินใจว่าจะให้มีการเพิ่มข้อมูลใหม่เข้าไปในสถานะเซลล์ปัจจุบันของหน่วยความจำหรือไม่ โดยมีการการคำนวณค่าการเปิดของประตูด้วยสมการ (4) ค่าที่ได้จากสมการนี้จะถูกใช้ร่วมกับข้อมูลใหม่ที่ถูกประมวลผลผ่านฟังก์ชัน tanh เพื่อนำเข้าสู่สถานะเซลล์ (Cell State) โดยมีการคำนวณค่าข้อมูลใหม่นี้ด้วยสมการ (5)

$$\overrightarrow{i}_t = \sigma(W_i \cdot [\overrightarrow{h}_{t-1}, x_t] + b_i) \quad (4)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (5)$$

โดยที่

$\overrightarrow{i_t}$ = ค่าการตัดสินใจ

$\widetilde{C}_t$ = ข้อมูลใหม่ที่อาจถูกเพิ่มเข้าไปในเซลล์ (Candidate Cell State)

W_i , W_C = น้ำหนักของประตูนำเข้าและเซลล์คู่แข่ง

tanh = ฟังก์ชัน hyperbolic tangent กำหนดข้อมูลใหม่อยู่ในช่วง -1 ถึง 1

Output Gate ทำหน้าที่ตัดสินใจว่าค่าจากสถานะเซลล์ที่ผ่านการปรับปรุงแล้ว ($\vec{C}_t$) นั้นควรจะแสดงข้อมูลอะไรออกมาเป็นสถานะซ่อนของเครือข่าย LSTM ณ เวลานั้น

ประกอบด้วยการคำนวณค่าสถานะเปิดของ Output Gate ด้วยสมการ (6) และ คำนวณค่าสถานะซ่อนที่เป็นผลลัพธ์ของเครือข่าย LSTM ณ เวลาปัจจุบัน ($\vec{h}_t$) ด้วยสมการ (7)

$$\vec{o}_t = \sigma(W_o \cdot [\vec{h}_{t-1}, x_t] + b_o) \quad (6)$$

$$\vec{h}_t = \vec{o}_t \odot \tanh(\vec{C}_t) \quad (7)$$

โดยที่

$\vec{O}_t$ = ค่าการตัดสินใจของประตูนำเข้า
$\vec{C}_t$ = สถานะของเซลล์หลังการปรับปรุง
$\vec{h}_t$ = สถานะซ่อนที่ส่งออกไป

สำหรับเครือข่าย Backward LSTM จะมีการทำงานคู่ขนานกับ Forward LSTM โดยมีการประมวลผลข้อมูลที่มีลำดับแบบย้อนกลับ คือตำแหน่งสุดท้าย ณ เวลา $t = T$ ไปยังตำแหน่งเริ่มต้น $t = 1$ เพื่อให้โมเดลเรียนรู้บริบทแบบย้อนกลับ โดยจะสร้างค่า Backward Hidden States ($\overleftarrow{h}_t$) และค่า Backward Cell State ($\overleftarrow{C}_t$) ณ ตำแหน่งเวลา t ซึ่งการทำงานประกอบด้วยโมเดลรับค่าข้อมูลนำเข้าปัจจุบัน x_t ค่า Backward Hidden State ตำแหน่งถัดไป ($\overleftarrow{h}_{t+1}$) และค่า Backward Cell State ($\overleftarrow{C}_{t+1}$) ถัดไป เพื่อใช้ในการคำนวณผ่านประตูต่าง ๆ เช่นเดียวกับเครือข่าย LSTM ปกติ แต่จะเป็นแบบทิศทางย้อนกลับ ผลลัพธ์จากการประมวลผลจากเครือข่ายทั้งสองจะถูกรวมเข้าด้วยกันเพื่อสร้างค่าผลลัพธ์จากข้อมูล (Combined Output) แล้วส่งไปยังชั้นที่อยู่ถัดไป เช่น Fully Connected Layer เพื่อประมวลผลในขั้นตอนต่อไป (Makarenkov *et al*., 2019) ตัวอย่างวิธีรวมผลลัพธ์ เช่น Concatenation ดังสมการ (8) Weighted Sum ดังสมการ (9) เป็นต้น

$$h_t = [\vec{h}_t; \overleftarrow{h}_t] \quad (8)$$

$$h_t = W_f \vec{h}_t + W_b \overleftarrow{h}_t + b \quad (9)$$

ขั้นตอนการวิจัย

การดำเนินการวิจัย แบ่งเป็น 2 ส่วน แต่ละส่วนมีขั้นตอนการดำเนินการ ดังนี้

1. การสร้างโมเดลรู้จำชื่อเฉพาะ

มีวัตถุประสงค์เพื่อฝึกและประเมินประสิทธิภาพของโมเดล สำหรับนำไปใช้ในขั้นตอนการพัฒนาระบบ แบ่งเป็น 5 ขั้นตอน ได้แก่ รวบรวมข้อมูล สร้างคลังข้อความ สร้างชุดข้อมูลฝึกฝน ฝึกฝนโมเดล และประเมินประสิทธิภาพโมเดล

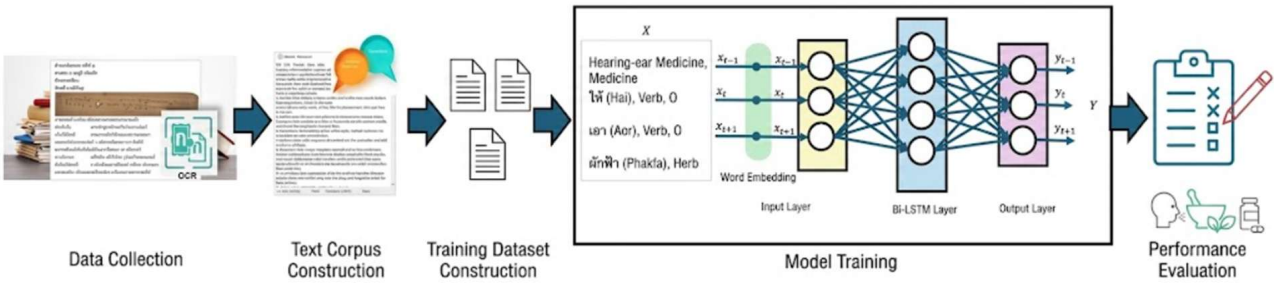


Figure 4 Steps in building the named entity recognition model for ancient Isan medicinal texts

รายละเอียดวิธีการสร้างโมเดลรู้จำชื่อเฉพาะตามขั้นตอนใน Figure 4 มีดังนี้

รวบรวมข้อมูล ได้แก่ การรวบรวมหนังสือตำรายาอีสานโบราณจำนวน 7 เล่ม ได้แก่ ตำรายาวัดมหาชัย จังหวัดมหาสารคาม เล่ม 1 - 5 ตำรายาวัดท่าม่วง อำเภอเสลภูมิ จังหวัดร้อยเอ็ด เล่ม 1 - 2 ปริวรรตจากต้นฉบับคัมภีร์ใบลานตำรายาอักษรไทยน้อยและอักษรธรรมอีสาน จากสถาบันวิจัยศิลปะและวัฒนธรรมอีสาน มหาวิทยาลัยมหาสารคาม ซึ่งได้มาในรูปแบบไฟล์ข้อความอิเล็กทรอนิกส์และหนังสือ ข้อมูลในรูปแบบหนังสือจะใช้โปรแกรมรู้จำอักขระด้วยแสง (Optical Character Recognition: OCR) จากเว็บไซต์ Ai for Thai ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ แปลงเป็นภาพข้อความเป็นข้อความอิเล็กทรอนิกส์ ตัวอย่างดัง Figure 5



Figure 5 Optical character recognition (OCR) program

หลังจากแปลงข้อมูลเสร็จจะนำข้อมูลทุกไฟล์มาทำความสะอาด (Data Cleaning) ได้แก่ ตรวจความถูกต้องของตัวสะกด และปรับการสะกดคำให้เป็นมาตรฐานเดียวกัน จากนั้นคัดเลือกเนื้อหาเอาเฉพาะตำรับยาเท่านั้น

การสร้างคลังข้อความ ได้แก่ การนำไฟล์ข้อความตำรายาอีสานจากขั้นตอนที่ 1 มาแบ่งแยกเป็นตำรับย่อย ๆ จากนั้นนำมาตัดคำและกำกับหมวดคำ (Part-of-Speech: POS) การแบ่งแยกข้อความตำรายาเป็นตำรับยาย่อย ๆ ใช้วิธีกึ่งอัตโนมัติคือ เขียนโปรแกรมแบ่งแยกประโยคตำรับยาด้วยกฎ เช่น จะตัดประโยค ณ ตำแหน่งก่อนหน้าคำดังนี้ เช่น “ยา” “ฯ” “พาก” “ผิว่า” เป็นต้น หากประโยคมีความยาวมากเกินไปจะทำการแบ่งแยกประโยคนั้นด้วยมืออีกครั้งหนึ่ง และการตัดคำจะใช้โปรแกรม pythainlp.tokenize โดยใช้คลังศัพท์ภาษาอีสาน (จินตนา พลศรี และคณะ, 2565) มาเป็นฐานความรู้ในการตัดคำ ดัง Figure 6 เนื่องจากโปรแกรมตัดคำ pythainlp เมื่อนำมาตัดคำภาษาอีสานโบราณพบว่ามีความผิดพลาดค่อนข้างสูง ซึ่งเกิดจากฐานข้อมูลคำศัพท์ของโปรแกรมไม่ครอบคลุมคำศัพท์ในตำรายาอีสานโบราณ และในคำหลายพยางค์บางคำอาจตัดคำได้มากกว่า 1 แบบ ขึ้นอยู่กับบริบท ดังนั้น ในกรณีตัดคำผิดจากคำศัพท์ในพจนานุกรมไม่ครอบคลุมหรือตัดคำผิดที่ทำให้ความหมายไม่สอดคล้องกับบริบท จะทำการตัดคำให้ถูกต้องด้วยมือ

	Vocab	Isan word	Thai word	POS	Definition
111	156	เคือ	เครือ	PDMN	เถาไม้, เรียกพรรณไม้ที่เป็นเถา
112	158	เคือเขามวก	เครือเขามวก	NPMED	รสฝาดเมา แก้สลบ แก้พิษทั้งปว
113	160	เคือเข้าหมื่นน้อย	เครือเข้าหมื่นน้อย	NPMED	ชื่อสมุนไพร
114	161	เคือเขาฮอ	บอระเพ็ด	NPMED	ดู "เขาฮอ"
115	163	เคือเขียว	เครือเขียว	NPMED	
116	164	เคือเหลือมเหลา	เครือเหลือมเหลา	NPMED	ดู "เหลือมเหลา"
117	170	เคือเห่า	เครืองูเห่า*	NPMED	ดู "เห่า
118	172	เคือเอ็นอ่อน	เครือเอ็นอ่อน	NPMED	ชื่อสมุนไพร
119	173	เคือแข้	เครือแข้	NPMED	ดู "แข้"
120	175	เคือไมไผ่บ้าน	เครือไมไผ่บ้าน	NPMED	ดู "ไผ่บ้าน"
121	177	เคือขามผี	เครือขามผี	NPMED	ดู "ขามผี"
122	181	เคือขี้กา	เครือขี้กา	NPMED	ดู "ขี้กา"

Figure 6 Example of the Isan-language vocabulary database

หลังจากกระบวนการตัดคำเสร็จสิ้น คำศัพท์แต่ละคำจะถูกกำกับด้วยหมวดคำที่ผู้วิจัยได้ออกแบบขึ้นใหม่ แบ่งออกเป็น 14 หมวด รองรับลักษณะเฉพาะของภาษาที่ปรากฏในตำรายาอีสานโบราณซึ่งประกอบด้วยคำในภาษาอีสาน คำยืมจากภาษาบาลี-สันสกฤต คำโบราณที่มักไม่ปรากฏในคลังคำมาตรฐานของภาษาไทยทั่วไป รวมถึงครอบคลุมบทบาททางไวยากรณ์ของคำตามบริบทเฉพาะของตำรับยา หมวดคำช่วยให้โมเดล Bi-LSTM เรียนรู้โครงสร้างไวยากรณ์ เช่น ตำแหน่งของคำนาม กริยา หรือคุณศัพท์ รวมถึงลดความคลุมเครือในคำบางคำในภาษาอีสานที่อาจมีความหมายหรือหน้าที่มากกว่าหนึ่งอย่าง เช่น “ฮาก” หมายถึง อาเจียน (กริยา) หรือ ราก (คำนาม) เป็นต้น รายละเอียดของแต่ละหมวดคำพร้อมตัวอย่างแสดงไว้ใน Table 1 และการกับหมวดคำจะเป็นแบบกึ่งอัตโนมัติ คือ เขียนโปรแกรมกำกับหมวดคำโดยใช้คลังศัพท์ภาษาอีสาน (จินตนา พลศรี และคณะ, 2565) และตรวจสอบความถูกต้องและแก้ไขความผิดพลาดด้วยมือโดยผู้เชี่ยวชาญด้านภาษาศาสตร์ที่รู้ภาษาอีสาน

เมื่อตัดแยกคำและกำกับหมวดคำเสร็จเรียบร้อยแล้ว ขั้นตอนต่อไปคือ การกำกับคลาสคำตอบ (Labeled Data) แบ่งเป็น 4 หมวด ได้แก่ ชื่อยา (Medicine) ชื่อสมุนไพร (Herb) อาการของโรค (Symptom) และ คำอื่น ๆ (Other) รายละเอียดคำอธิบายและตัวอย่างคำในแต่ละคลาส แสดงใน Table 2 โดยการกำกับคลาสคำตอบนั้น ช่วยให้โมเดลเข้าใจความหมายและความสัมพันธ์ของคำต่าง ๆ ในบริบทตำรายา และใช้ในการประเมินประสิทธิภาพของโมเดล ผลลัพธ์การดำเนินการที่ได้จากขั้นตอนนี้ คือไฟล์คลังข้อความตำรับยาอีสานโบราณที่พร้อมสำหรับนำไปใช้ฝึกฝนโมเดล โดยได้ตำรับยาจำนวนทั้งสิ้น 3,921 ตำรับ/ประโยค และมีคำศัพท์ทั้งหมด 102,868 คำ

Table 1 Part-of-speech (POS) categories.

POS category	Description	Example words (Isan)
NOUN	Common noun	yaa-mahok, phak-tamnin (names of a medicine and an herb)
VERB	Action verb	kin (eat), hom [rom] (to fumigate/steam-treat), fon (grind/pound)
PRON	Pronoun	khao (he/she), phu-khai (the patient), man (it)
PREP	Positional preposition	thoeng [bon] (above/on top), lum [lang] (below/under)

Table 1 Part-of-speech (POS) categories. (Cont.)

POS category	Description	Example words (Isan)
ADJ	Adjective (noun modifier)	daeng (red), khao (white), baan (domestic/village), pa (wild/forest)
ADV	Adverb (verb modifier)	dii (good/well), khuen (increase/up), long (decrease/down)
AUX	Auxiliary verb	si, ja ("will," future/intention markers)
CONJ	Conjunction	jueng (so/therefore), laeo-jing, laeo-jueng (and then)
DET	Determiner	dai (any/which), su [thuk] (every/all)
EAFF	Sentence-final particle	thoe (please/go ahead), dii-lae (that's fine/all right)
PUNC	Punctuation mark	ฯ (Thai section/ellipsis mark)
SPELL	Incantation/ chant	sitthikan, namo (sacred Pali invocation words)
NUM	Number	1, 2, 3
X	Unknown meaning	(none)

Table 2 Answer (Entity) Classes.

Class	Description and example Isan words
Medicine	Medicine name, e.g., yaa-hu-nak ("heavy-ear remedy," for deafness), yaa-puat-thong (stomachache remedy)
Herb	Herb name, e.g., khao-min (turmeric), nam-phueng (honey)
Symptom	Name of a symptom, e.g., khai (fever), fok (bruised), buam (swollen)
O (Other)	Words outside the three main classes, e.g., preparation methods (grind, boil), usage instructions (eat, apply), incantations (sit-thikariya, namo), preparation utensils (cup, pot)

โครงสร้างไฟล์ข้อมูล ประกอบด้วยข้อมูล 4 คอลัมน์ ได้แก่ รหัส (ID) คำศัพท์ (WORD) หมวดคำ (POS) และคลาสของคำ (CLASS) ซึ่งจะถูกจัดเก็บในรูปแบบไฟล์ CSV ตัวอย่างโครงสร้างข้อมูล ดังปรากฏใน Figure 7

```
ID,WORD,POS,CLASS
1,ยาปวดห้อง,NOUN,Medicine
2,เอา,VERB,O
3,กล้วยหานีสุก,NOUN,Herb
4,ตำ,VERB,O
5,ใส่,VERB,O
6,พริกน้อย,NOUN,Herb
7,7,NUM,O
8,เทียม,NOUN,Herb
9,7,NUM,O
```

Figure 7 Data file structure

การสร้างคลังข้อความจากตำรายาอีสานโบราณเป็นขั้นตอนเตรียมข้อมูลให้มีความเหมาะสมสำหรับการฝึกฝนโมเดล (Pre-Processing) มีความสำคัญเป็นอย่างมาก ขณะเดียวกันก็ใช้เวลาดำเนินการค่อนข้างมากจากข้อจำกัดสำคัญ คือ การขาดแคลนเครื่องมือประมวลผลข้อมูลแบบอัตโนมัติที่มีประสิทธิภาพ นำไปสู่ความจำเป็นในการพัฒนาเครื่องมือเฉพาะทางเพื่อสนับสนุนการดำเนินงานขึ้นมาใหม่ เช่น การเขียนโปรแกรมเพื่อแยกตำรับยาออกจากกันด้วยวิธีใช้กฎ ตลอดจนการปรับปรุงฐานความรู้ที่ใช้สำหรับการตัดคำ อย่างไรก็ตาม ด้วยข้อจำกัดของการกำหนดกฎซึ่งไม่สามารถครอบคลุมรูปแบบการเขียนตำรับยาที่มีลักษณะหลากหลายได้อย่างครบถ้วน ส่งผลให้เกิดข้อผิดพลาดในการแยกประโยค เช่น การแบ่งแยกตำรับยาไม่ถูกต้อง หรือมีความยาวมากเกินไป อีกทั้งการตัดคำยังประสบปัญหาจากพจนานุกรมที่มีคำศัพท์ไม่เพียงพอ ซึ่งนอกจากจะนำไปสู่การตัดคำที่คลาดเคลื่อนแล้ว ยังส่งผลให้เกิดคำไม่รู้จัก (Unknown Words) และการระบุหมวดคำที่ผิดพลาดด้วย ทำให้จำเป็นต้องมีการตรวจสอบและแก้ไขด้วยมนุษย์ในทุกขั้นตอน เพื่อป้องกันไม่ให้ข้อผิดพลาดดังกล่าวส่งผลกระทบต่อกระบวนการเรียนรู้ของโมเดลในลำดับถัดไป

เตรียมข้อมูลสำหรับฝึกโมเดล ได้แก่ แบ่งชุดข้อมูลออกเป็น 3 ไฟล์ ได้แก่ ข้อมูลชุดฝึก (Training Dataset) ชุดทดสอบ (Test Dataset) และชุดประเมินประสิทธิภาพ (Validation Dataset) ในสัดส่วนร้อยละ 80:10:10 ตามลำดับ การระบุสัดส่วนดังกล่าวจะถูกกำหนดไว้ในโปรแกรมฝึกฝนและทดสอบโมเดล

ฝึกโมเดล Bi-LSTMs ได้แก่ การนำชุดข้อมูลที่สร้างไว้ไปฝึกโมเดลรู้จำชื่อเฉพาะด้วยเทคนิค Bi-LTSM โดยใช้เทคนิคฝังเวกเตอร์ คือ Word2Vec แปลงคำให้อยู่ในรูปเวกเตอร์ ชุดคำสั่งฝึกโมเดลเขียนด้วยภาษา Python โดยใช้ไลบรารี gensim สำหรับสร้างเวกเตอร์คำ และ Keras สำหรับ

ฝึกโมเดล Bi-LSTM ซึ่งการฝึกโมเดล โมเดลจะเรียนรู้จากข้อมูลที่มีป้ายกำกับ (Labeled Data) เพื่อปรับปรุงพารามิเตอร์ภายใน เช่น น้ำหนักของการเชื่อมต่อและค่าความลำเอียง (Bias) ให้สามารถคาดการณ์ลำดับข้อมูลได้อย่างถูกต้องและแม่นยำยิ่งขึ้น กระบวนการนี้ใช้เทคนิคต่าง ๆ เช่น Backpropagation Through Time (BPTT) ซึ่งขยายแนวคิดของ Backpropagation สำหรับข้อมูลที่มีลำดับ (Werbos, 1990)

ประเมินประสิทธิภาพโมเดล ใช้ Metrics สำหรับการจำแนกประเภท (Classification) ได้แก่ Precision, Recall, F1-score ทั้งแบบค่าเฉลี่ยในแต่ละคลาส (Macro Average) และเฉลี่ยแบบถ่วงน้ำหนัก (Weighted Average) เพื่อดูประสิทธิภาพของโมเดลในกรณีที่ข้อมูลแต่ละคลาสมีจำนวนไม่เท่ากัน

2. การพัฒนาระบบรู้จำชื่อเฉพาะจากตำรายาอีสานโบราณ

กำหนดขั้นตอนการดำเนินการ ดังนี้

สำรวจปัญหาและความต้องการของผู้ใช้ระบบ โดยการศึกษารูปแบบการบันทึกตำรับยาและลักษณะชื่อเฉพาะ สอบถามความต้องการของผู้ใช้ โดยโครงการวิจัยนี้จะมุ่งเน้นผู้ใช้งานเป็นนักภาษาโบราณเป็นหลัก

วิเคราะห์และออกแบบระบบ ได้แก่ การออกแบบส่วนติดต่อกับผู้ใช้ (User Experience/User Interface: UX/UI) การออกแบบขั้นตอนการประมวลผลข้อมูล และออกแบบฐานข้อมูลสำหรับจัดเก็บองค์ความรู้ที่สกัดได้ ประกอบด้วย 2 ตาราง ได้แก่ ตารางจัดเก็บองค์ความรู้ที่สกัดได้ (ExtractVocab) ตารางคำศัพท์ (Vocabulary) ใช้สำหรับจัดเก็บรายละเอียดหมวดคำ และความหมายของคำศัพท์แต่ละคำ ฐานข้อมูลนี้รองรับการนำไปสร้างพจนานุกรม ความสัมพันธ์ระหว่างตารางดังแสดงใน Figure 8

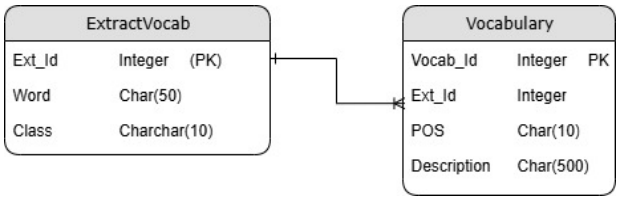


Figure 8 Database structure

รายละเอียดคำอธิบายโครงสร้างตาราง ดังแสดงใน table 3 และ 4

Table 3 "ExtractVocab" Table Structure.

Field name	Data type	Key	Description
Ext_Id	Integer	PK	Extracted-knowledge record ID
Word	Char		Word/term
Class	Char		Entity class

Table 4 "Vocabulary" table structure.

Field name	Data type	Key	Description
Vocab_Id	Integer	PK	Vocabulary ID
Ext_Id	Integer		Reference to the extracted-knowledge record ID
POS	Varchar		Part-of-speech tag
Description	Varchar		Meaning/definition of the word

พัฒนาระบบ ได้แก่ ดำเนินการพัฒนาระบบตามที่ออกแบบเอาไว้ด้วยภาษา Python โดยนำโมเดลรู้จำชื่อเฉพาะที่พัฒนาขึ้น มาใช้ในกระบวนการรู้จำชื่อเฉพาะเพื่อจำแนกชื่อเฉพาะที่ต้องการสกัดออกมา และสร้างฐานข้อมูลจัดเก็บชื่อเฉพาะที่สกัดได้ด้วยโปรแกรม SQLiteStudio

ทดสอบระบบและปรับปรุงระบบ ระบบที่พัฒนาเสร็จแล้วจะถูกนำมาทดสอบโดยใช้ข้อมูลจากตำรายาอีสานโบราณมาเป็นข้อมูลนำเข้า เพื่อตรวจสอบความสมบูรณ์ถูกต้องของกระบวนการทำงานในทุกขั้นตอน และดำเนินการปรับปรุงแก้ไขหากพบข้อผิดพลาด

ติดตั้งแอปพลิเคชันที่เสร็จสมบูรณ์แล้วไว้บนเครื่องแม่ข่าย (Web Server) เพื่อใช้งานระบบผ่านเครือข่ายอินเทอร์เน็ต

การประเมินประสิทธิภาพของระบบ โครงการวิจัยนี้กำหนดวิธีวัดประสิทธิภาพของระบบที่พัฒนาจากความพึงพอใจต่อการพัฒนาระบบ ผู้ประเมินประกอบด้วย นักพัฒนาโปรแกรมคอมพิวเตอร์จำนวน 5 คน นักภาษาโบราณจำนวน 5 คน ผู้ใช้ประโยชน์จากตำรายาอีสานโบราณจำนวน 2 คน รวมทั้งสิ้น 12 คน เครื่องมือเก็บรวบรวมข้อมูล ได้แก่ แบบสอบถามความพึงพอใจต่อการพัฒนาและการใช้งานระบบ การให้คะแนนความพึงพอใจ แบ่งเป็น 5 ระดับ (Rating Scale) ตามแบบของลิเคอร์ท (Likert, 1961) โดยมีมาตราส่วนการให้คะแนนน้อยที่สุด 1 คะแนน และมากที่สุด 5 คะแนน ความพึงพอใจน้อยที่สุดถึงมากที่สุด การประเมินผลจะดำเนินการหลังจากผู้ประเมินได้ทดลองใช้ระบบงานเรียบร้อยแล้ว สถิติ

สำหรับวิเคราะห์ข้อมูล ได้แก่ ค่าเฉลี่ย ($\bar{x}$) และส่วนเบี่ยงเบนมาตรฐาน (SD)

เกณฑ์การแปลผลค่าเฉลี่ยความพึงพอใจ กำหนดไว้ดังนี้

ค่าเฉลี่ย	ระดับประสิทธิภาพ
4.21 - 5.00	ดีมาก
3.41 - 4.20	มาก
2.61 - 3.40	ปานกลาง
1.81 - 2.60	น้อย
1.00 - 1.80	น้อยมาก

ผลการทดลองและอภิปรายผล

ประสิทธิภาพของโมเดลสกัดองค์ความรู้จากตำรายาอีสานโบราณด้วยเทคนิค Bi-LSTM

การวิจัยนี้ประเมินประสิทธิภาพของโมเดลที่สร้างขึ้นโดยใช้ข้อมูลชุดประเมินประสิทธิภาพ (Validation Dataset) ที่ได้แบ่งแยกมาจากคลังข้อความตำรายาอีสานโบราณที่สร้างเตรียมไว้ในขั้นตอนรวบรวมข้อมูลเพื่อการสร้างโมเดล โดยมีจำนวนคำศัพท์ทั้งหมด 8,691 คำ สัดส่วนจำนวนคำศัพท์ (Support) ของทั้ง 4 คลาส มีรายละเอียดดังแสดงใน Table 5

Table 5 Data count by class.

Class	Count	Proportion (%)
Herb	1,838	21.14
Medicine	262	3.01
Symptom	489	5.63
Others	6,102	70.22
Total	8,691	100

ผลการวิจัยพบว่า การกำหนดค่าพารามิเตอร์ ได้แก่ Epochs และ Vector size ที่แตกต่างกัน ส่งผลต่อค่าความแม่นยำในการรู้จำชื่อเฉพาะในตำรายาอีสานโบราณที่แตกต่างกัน ดังจะเห็นได้จากผลการทดสอบประสิทธิภาพของโมเดล Bi-LSTM ภายใต้การกำหนดค่าพารามิเตอร์ที่แตกต่างกัน 4 รูปแบบ ผลการทดสอบใน Table 6 และแผนภูมิแสดงประสิทธิภาพในการเรียนรู้ของโมเดล ใน Figure 9 เส้น dev ซึ่งใช้ข้อมูลชุดประเมินประสิทธิภาพ (Validation Dataset) เปรียบเทียบกับเส้น train แสดงให้เห็นว่า โมเดลแบบที่ 4 (Epoch = 5, Vector Size = 300) ให้ผลลัพธ์ที่ดีที่สุดทั้งแบบ Micro F1-Score, Macro F1-Score และ Weighted F1-score เมื่อเพิ่มจำนวน Epoch จาก 3 เป็น 5 รอบ และปรับค่า Vector size ให้มีขนาดเพิ่มขึ้น จาก 150 เป็น 300 พบว่า ทั้ง Precision, Recall และ F1-score โดยเฉพาะในกลุ่มชื่อโรคและชื่อยา ค่าความแม่นยำมีแนวโน้มดีขึ้น โดยเฉพาะคลาสชื่อยา แม้ว่าจะมีจำนวนตัวอย่างน้อย แต่เมื่อเพิ่มรอบการฝึกซ้ำหลายรอบ พบว่า Recall เพิ่มขึ้นจาก 0.80 เป็น 0.89 และ F1-score เพิ่มจาก 0.06 เป็น 0.07 โดยแผนภูมิ Figure 9 แสดงความความถูกต้อง (Accuracy) ในแต่ละรอบการฝึกที่แสดงแนวโน้มความแม่นยำของโมเดลมีการเพิ่มขึ้นอย่างต่อเนื่อง และค่าความสูญเสีย (Loss) มีค่าลดลงต่อเนื่องเช่นกัน แสดงถึงความสามารถในการลดความคลาดเคลื่อนระหว่างค่าที่คาดการณ์กับค่าจริงได้ดีของโมเดล

สำหรับประสิทธิภาพความแม่นยำของโมเดล Bi-LSTM ในการรู้จำชื่อเฉพาะในตำรายาอีสานโบราณ เมื่อพิจารณาเฉพาะ 3 คลาสที่เป็นเป้าหมายหลักของงานวิจัย ได้แก่ ชื่อสมุนไพร ชื่อยา และชื่อโรค อ้างอิงตามผลการทดสอบประสิทธิภาพ Table 6 โมเดลที่ 4 พบว่า โมเดลรู้จำชื่อเฉพาะในคลาส ชื่อสมุนไพร ได้ดีที่สุด โดยสรุปผลในแต่ละคลาส ดังนี้ ชื่อสมุนไพร เป็นคลาสที่โมเดลรู้จำชื่อเฉพาะได้ดีที่สุด โดยมีค่า Precision เท่ากับ 0.96, Recall เท่ากับ 0.96 และ F1-score เท่ากับ 0.96 แสดงถึงโมเดลจำแนกคลาสชื่อสมุนไพรออกจากคลาสอื่นได้ดี อธิบายได้ว่า การปรากฏของชื่อสมุนไพรในตำรายามีความชัดเจน คือมักเขียนติดกันในบริบทเดียวกันกับคลาสที่มีความสัมพันธ์กันในทางความหมาย เช่น “ยาหูหนักเอาน้ำมันปลาฝาหนวดปลาเค้งใส่ถ้วยทองแหย่หูดีแล” จะเห็นได้ว่า ชื่อสมุนไพรมักจะอยู่หลังคลาส ชื่อยา หรือชื่อโรค หรืออาการ หรือ คลาสที่อยู่ก่อนหน้าหรือหลังคลาสชื่อสมุนไพรมักเป็นคลาสชื่อสมุนไพร เป็นต้น

ชื่อยา มีความแม่นยำต่ำที่สุด โดยค่า Precision เท่ากับ 0.04, Recall เท่ากับ 0.89, F1-scoreเท่ากับ 0.07 แต่ด้วยค่า Recall ที่ค่อนข้างดี แสดงถึงโมเดลจำแนกชื่อยาออกมาได้ครบถ้วนแต่มีความผิดพลาดสูง สาเหตุสำคัญเนื่องจากจำนวนคำประเภทชื่อยามีน้อยในชุดฝึก (ร้อยละ 3.01) มีความกำกวมและคล้ายคลึงกับชื่อโรค เช่น “ยาคนเป็นเปี้ยไปมาบ่ได้” “ยาเจ็บเสียบ” เป็นต้น จึงส่งผลต่อการเรียนรู้ของโมเดลโดยตรง ทั้งนี้ความไม่สมดุลของจำนวนคำศัพท์ในแต่ละคลาสโดยเฉพาะคลาสชื่อยา เป็นลักษณะปกติทั่วไปของชุดข้อมูลประเภทนี้ เนื่องจากตำรายา 1 เล่ม จะมีตำรับยาประมาณ 50 ตำรับ และแต่ละตำรับจะมีความสัมพันธ์กับสมุนไพรที่ใช้รักษาหลายรายการ การเพิ่มปริมาณข้อมูลชื่อยาให้มีสัดส่วนสูงขึ้นส่งผลต่อการเพิ่มขึ้นของจำนวนคำศัพท์ในคลาสชื่อสมุนไพรและคลาสอื่น ๆ ด้วยเช่นกัน อย่างไรก็ตามการเพิ่มจำนวนของชื่อยาให้มีจำนวนที่เพียงพอกับการเรียนรู้ของโมเดลเป็นสิ่งที่ต้องดำเนินการในงานวิจัยครั้งต่อไป

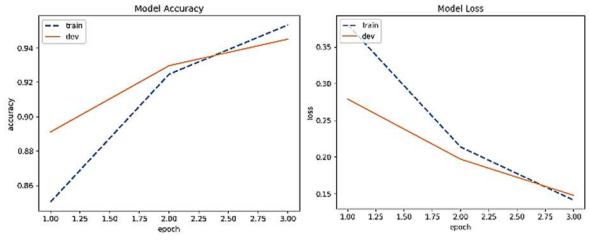
ชื่อโรค หรือชื่ออาการ ความแม่นยำอยู่ในระดับดี ค่า Precision เท่ากับ 0.83, Recall เท่ากับ 0.82 และ F1-score เท่ากับ 0.82 เมื่อเทียบกับโมเดล Bi-LSTM แบบที่ 1 - 3 สะท้อนว่าโมเดลมีแนวโน้มการเรียนรู้และความแม่นยำมากขึ้นเมื่อเพิ่มรอบการฝึกจาก 3 เป็น 5 Epoch แต่ยังตรวจจับชื่อโรคได้ไม่ครบถ้วน ที่เกิดจากชื่อโรคในตำรายาอีสานมักเป็นคำหลาพยางค์ที่ประสมจากคำโดดหลายคำ การระบุขอบเขตชื่อโรคทำได้ยาก ชื่อโรคกับชื่อยามีความคล้ายคลึงกัน หรือใช้ชื่อโรคแทนชื่อยา เช่น “ออกเห่าห้อมทำทาน” “ตาเปียกตาจิ้น” เป็นต้น ประกอบกับจำนวนข้อมูลมีน้อยทำให้ไม่เพียงพอสำหรับให้โมเดลเรียนรู้

สำหรับคลาสอื่น ๆ (Others) นั้น มีความแม่นยำใกล้เคียงกันทั้ง Precision, Recall และ F1-score อยู่ในระดับสูงถึง 0.98 แสดงให้เห็นว่า โมเดลสามารถระบุคำทั่วไปที่ไม่เกี่ยวกับชื่อสมุนไพร หรืออาการ ได้อย่างแม่นยำ ซึ่งส่วนหนึ่งเกิดจากจำนวนตัวอย่างข้อมูลมีเป็นจำนวนค่อนข้างมากกว่าทุกคลาส (6,102 คำ) คิดเป็นสัดส่วนร้อยละ 70.22 และลักษณะเฉพาะของคำกลุ่มนี้มักมีรูปแบบที่ง่ายต่อการจำแนก เช่น มีหมวดคำเป็นคำกริยา คำเชื่อม คำกริยาวิเศษณ์ เป็นต้น

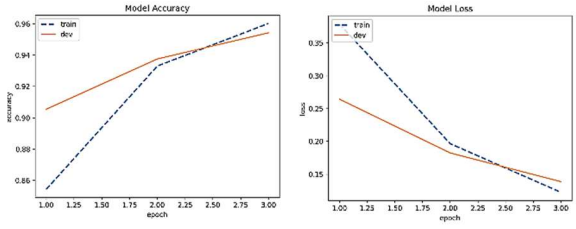
Table 6 Bi-LSTM Performance Under Four Parameter Settings, and CRF Comparison.

1) Epoch = 3, Vector size = 150, Batch size =16			
Bi-LSTM Model	Precision	Recall	F1
Herb	0.94	0.95	0.94
Medicine	0.03	0.80	0.06
Symptom	0.75	0.79	0.77
Others	0.98	0.97	0.98
Macro avg	0.68	0.88	0.69
Weighted avg	0.93	0.95	0.93
2) Epoch = 3, Vector size = 300, Batch size =16			
Bi-LSTM Setting	Precision	Recall	F1
Herb	0.93	0.96	0.95
Medicine	0.03	0.80	0.06
Symptom	0.78	0.80	0.79
Others	0.98	0.97	0.98
Macro avg	0.68	0.88	0.69
Weighted avg	0.93	0.96	0.93
3) Epoch = 5,Vector size = 150, Batch size = 16			
Bi-LSTM Setting	Precision	Recall	F1
Herb	0.95	0.96	0.96
Medicine	0.04	0.85	0.07
Symptom	0.75	0.85	0.79
Others	0.98	0.99	0.98
Macro avg	0.68	0.91	0.70
Weighted avg	0.94	0.96	0.94
4) Epoch = 5, Vector size = 300, Batch size =16			
Bi-LSTM Setting	Precision	Recall	F1
Herb	0.96	0.96	0.96
Medicine	0.04	0.89	0.07
Symptom	0.83	0.82	0.82
Others	0.99	0.98	0.98
Macro avg	0.70	0.91	0.71
Weighted avg	0.94	0.97	0.94
CRF Model	Precision	Recall	F1
Herb	0.91	0.97	0.94
Medicine	0.89	0.84	0.86
Symptom	0.88	0.73	0.80
Other	0.98	0.73	0.98
Macro avg	0.92	0.88	0.90
Weighted avg	0.96	0.96	0.96

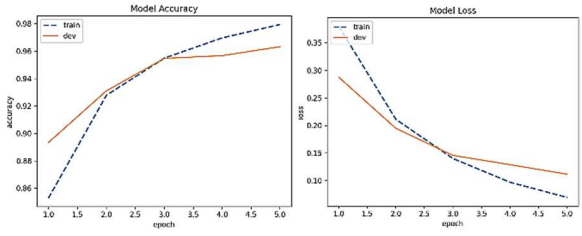
Model 1 Epoch = 3, Vector size = 150, Batch size = 16



Model 2 Epoch = 3, Vector size = 300, Batch size = 16



Model 3 Epoch = 5, Vector size = 150, Batch size = 16



Model 4 Epoch = 5, Vector size = 300, Batch size = 16

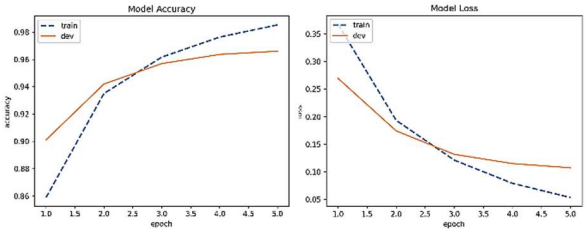


Figure 9 Learning performance of the Bi-LSTM model under four parameter configurations

แม้ว่าผลการทดลองจะแสดงให้เห็นว่าโมเดลมีค่าเฉลี่ยถ่วงน้ำหนัก (Weighted Average F1-score) อยู่ในระดับที่น่าพอใจที่ 0.94 แต่อย่างไรก็ตาม ค่าเฉลี่ยดังกล่าว อาจไม่สะท้อนประสิทธิภาพที่แท้จริงของการจำแนกคลาสเป้าหมายในงานวิจัยนี้ ทั้งนี้เพราะค่า F1-score ที่สูงส่วนหนึ่งมาจากคลาสที่มีข้อมูลจำนวนมากและมีรูปแบบบริบทที่ค่อนข้างชัดเจน เช่น คลาสอื่น ๆ (Other) และคลาสชื่อสมุนไพร ซึ่งโมเดลสามารถเรียนรู้ได้ดีจากข้อมูลที่มีความสมดุลและมีจำนวนเพียงพอ ในทางกลับกัน คลาสที่มีความสำคัญต่อการสกัดองค์ความรู้ เช่น ชื่อยา หรือชื่อโรค กลับมีข้อมูลจำนวนน้อยและมีความไม่สมดุลอย่างชัดเจน ส่งผลให้ Precision และ Recall ของคลาสเหล่านี้ต่ำกว่าค่าเฉลี่ย แม้ว่าภาพรวมอาจดูดี แต่ไม่สามารถใช้แทนการประเมินในกรณีที่ข้อมูลมีความไม่สมดุลสูงได้ ดังนั้น การประเมินประสิทธิภาพระหว่างโมเดลในบริบทนี้ ควรพิจารณาจากค่า Precision และ Recall รายคลาสเป็นหลัก เพื่อสะท้อนศักยภาพของโมเดลในการรู้จำคำสำคัญในแต่ละประเภทอย่างแท้จริง และเพื่อเพิ่มประสิทธิภาพของการสกัดองค์ความรู้ในโดเมนตำรายาอีสานโบราณ การแก้ไขปัญหาความไม่สมดุลของข้อมูล โดยเฉพาะการเพิ่มจำนวนข้อมูลในคลาสที่มีน้อย เช่น ชื่อยาและชื่อโรค ถือเป็นปัจจัยสำคัญลำดับแรกที่ต้องดำเนินการ

ผลการเปรียบเทียบประสิทธิภาพของโมเดล Conditional Random Fields (CRF) (Polsri, 2024) และ Bi-LSTM แบบที่ 4 เฉพาะสามคลาสหลักที่เป็นองค์ความรู้สำคัญในตำรายาอีสานโบราณที่ต้องการสกัด ได้แก่ ชื่อสมุนไพร (Herb), ชื่อยา (Medicine) และ ชื่อโรค (Symptom) โดยประเมินจากตัวชี้วัดประสิทธิภาพ ได้แก่ Precision, Recall และ F1-score พบว่าโมเดล Bi-LSTM แบบที่ 4 มีประสิทธิภาพสูงที่สุดในการจำแนกคลาสชื่อสมุนไพร โดยให้ค่า F1-score เท่ากับ 0.96 ซึ่งสอดคล้องกับค่า Precision และ Recall ที่มีค่าเท่ากันที่ 0.96 แสดงให้เห็นถึงความสามารถของโมเดลในการจำแนกคำประเภทสมุนไพรได้อย่างแม่นยำและครอบคลุม อย่างไรก็ตาม ประสิทธิภาพของโมเดลดังกล่าวลดลงอย่างมีนัยสำคัญในคลาสชื่อยา โดยมีค่า Precision ต่ำมากเพียง 0.04 แม้ว่า Recall จะอยู่ในระดับสูงถึง 0.89 ส่งผลให้ค่า F1-score ลดลงเหลือเพียง 0.07 สะท้อนถึงปัญหาการเกิดค่าบวกเทียม (False Positive) ในระดับสูง กล่าวคือ โมเดลมีแนวโน้มจำแนกคำจำนวนมากว่าเป็นชื่อยา แม้จะไม่ถูกต้องก็ตาม

ในส่วนของคลาสชื่อโรค โมเดล Bi-LSTM ให้ผลลัพธ์ในระดับปานกลาง โดยมีค่า Precision เท่ากับ 0.83 และ Recall เท่ากับ 0.82 ซึ่งถือว่าดีกว่าโมเดล CRF ที่ให้ค่า Recall ต่ำกว่าอย่างชัดเจน (0.73) ขณะที่โมเดล CRF แสดงความสามารถในการจำแนกทั้งสามคลาสอย่างสม่ำเสมอมากกว่า โดยเฉพาะในคลาสชื่อยาและชื่อโรคที่ให้ค่า Precision และ Recall อยู่ในระดับที่ใกล้เคียงกัน สะท้อนถึงเสถียรภาพของโมเดล CRF แม้ว่าค่า F1-score จะไม่สูงเท่ากับ Bi-LSTM ในบางคลาสก็ตาม

ผลการทดลองข้างต้นชี้ให้เห็นถึงข้อจำกัดของโมเดล Bi-LSTM ที่ได้รับผลกระทบจากปัญหาความไม่สมดุลของข้อมูล (Class Imbalance) อย่างชัดเจน โดยเฉพาะในคลาสชื่อยาซึ่งมีจำนวนตัวอย่างในชุดข้อมูลค่อนข้างน้อย ส่งผลให้โมเดลไม่สามารถเรียนรู้ลักษณะเฉพาะของคำในคลาสนี้ได้อย่างเพียงพอ แม้จะสามารถเรียกคืนตัวอย่างที่แท้จริงได้จำนวนหนึ่ง แต่กลับไม่สามารถจำแนกได้อย่างแม่นยำ ทำให้ Precision ต่ำมาก

ดังนั้น เพื่อยกระดับประสิทธิภาพของโมเดล Bi-LSTM ในการสกัดองค์ความรู้จากตำรายาอีสานโบราณ ควรมีการจัดการกับปัญหาความไม่สมดุลของข้อมูลอย่างเป็นระบบ เช่น การเพิ่มจำนวนข้อมูลในคลาสที่มีจำนวนน้อย การใช้เทคนิค Resampling หรือการปรับแต่งฟังก์ชันความสูญเสีย (Loss Function) ด้วยน้ำหนักถ่วง (Weighted Loss) เพื่อช่วยให้โมเดลสามารถเรียนรู้และจำแนกคำประเภทที่มีความสำคัญแต่มีจำนวนจำกัดในข้อมูลได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

การพัฒนาระบบสกัดองค์ความรู้จากตำรายาอีสานโบราณและผลการประเมินคุณภาพของระบบ

ระบบที่พัฒนาขึ้นนี้มีเป้าหมายเพื่อสร้างเครื่องมืออำนวยความสะดวกในการสกัดองค์ความรู้ที่บันทึกในตำรายาอีสานโบราณให้มีความสะดวก รวดเร็ว ใช้งานง่าย และสามารถใช้งานระบบได้จากทุกที่โดยไม่ต้องลงทะเบียน เข้าถึงได้ผ่านเครือข่ายอินเตอร์เน็ต ไม่มีข้อจำกัดการใช้งานจากทั้งเวลาและสถานที่ให้บริการ การพัฒนาระบบจึงนำเสนอในรูปแบบเว็บแอปพลิเคชัน ฟังก์ชันหลักประกอบด้วย ข้อมูลทั่วไปเกี่ยวกับโครงการและผู้พัฒนา ฐานข้อมูลคำศัพท์ที่ใช้ในการวิจัย และกระบวนการสกัดและจัดเก็บองค์ความรู้ในฐานข้อมูล งานวิจัยนี้ช่วยให้ข้อมูลองค์ความรู้จากเดิมที่มีรูปแบบข้อความที่ไม่เป็นโครงสร้างถูกนำมาจำแนกแยกออกเป็นคลาสจัดเก็บในแบบฐานข้อมูลแบบมีโครงสร้าง ซึ่งจะง่ายต่อการนำไปใช้งาน และสะดวกยิ่งขึ้น

กระบวนการสกัดและจัดเก็บองค์ความรู้ออกแบบให้สามารถรองรับการนำเข้าข้อมูลได้ทั้งในรูปแบบการป้อนข้อมูลด้วยตนเอง (Manual Input) และการอัปโหลดไฟล์ข้อความเข้าสู่ระบบ โดยข้อมูลที่นำเข้าเหล่านี้จะผ่านการประมวลผลด้วยโมดูลการรู้จำชื่อเฉพาะแบบ Bi-LSTM เพื่อระบุคลาสของคำศัพท์แต่ละรายการ (คำศัพท์:คลาส) ผลลัพธ์การจำแนกคลาสของคำที่ผู้ใช้งานสามารถปรับปรุงแก้ไขให้ถูกต้องก่อนการบันทึกเก็บในฐานข้อมูลช่วยให้ข้อมูลมีความถูกต้องยิ่งขึ้น ผลลัพธ์ที่ได้จากกระบวนการดังกล่าวจะแสดงตัวอย่างใน Figure 10 และ Figure 11

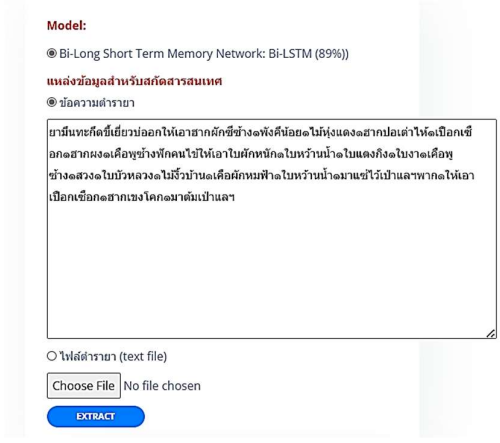


Figure 10 Example of knowledge extraction from a medicinal recipe

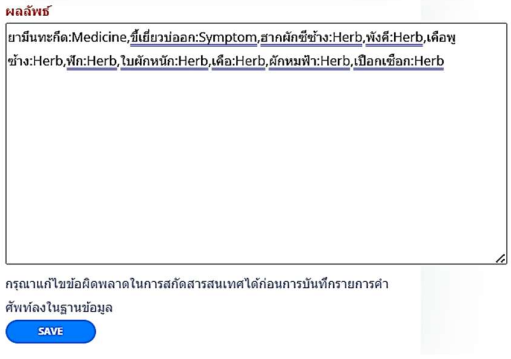


Figure 11 Knowledge extraction results

องค์ความรู้ที่ผ่านการจำแนกประเภทแล้วจะถูกจัดเก็บในฐานข้อมูลตามโครงสร้างที่แสดงใน Table 6 ซึ่งสามารถนำไปใช้งานในหลายลักษณะ เช่น นำไประบุชนิดของคำ การจัดทำคำอธิบายความหมายของคำศัพท์ และการสร้างชุดคำศัพท์เพื่อใช้ในการตัดคำสำหรับระบบประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) เช่น พจนานุกรมอิเล็กทรอนิกส์ หรือรายการคำศัพท์เฉพาะทาง เป็นต้น

เพื่อประเมินประสิทธิภาพของระบบที่พัฒนาขึ้น งานวิจัยนี้ได้ดำเนินการสำรวจความพึงพอใจจากผู้ใช้งานที่มีความเกี่ยวข้องกับระบบโดยตรง ประกอบด้วย นักภาษาโบราณ 5 คน ผู้เชี่ยวชาญด้านการพัฒนาโปรแกรมคอมพิวเตอร์ 5 คน และผู้ใช้ประโยชน์จากตำรายาโบราณอีสาน 2 คน รวมทั้งสิ้น 12 คน แบบสำรวจที่ใช้ในรวบรวมข้อมูลครอบคลุม 2 ด้านหลัก ได้แก่ (1) ความเหมาะสมของการออกแบบและพัฒนาระบบ และ (2) ความเหมาะสมของการออกแบบฐานข้อมูล โดยใช้มาตรวัดระดับความพึงพอใจ 5 ระดับ จากมากที่สุด ถึง น้อยที่สุด ผลการประเมิน ดังแสดงใน Table 7 พบว่า ผู้ใช้งานมี

ระดับความพึงพอใจทุกด้านอยู่ในระดับมากที่สุด โดยเฉพาะด้านการออกแบบและพัฒนาระบบ สะท้อนให้เห็นว่า ระบบที่พัฒนาขึ้นตอบสนองความต้องการของผู้ใช้งานได้อย่างมีประสิทธิภาพ ทั้งยังช่วยให้การสกัดองค์ความรู้จากตำราโบราณอีสานมีความสะดวกรวดเร็วขึ้น เมื่อเปรียบเทียบกับวิธีการแบบดั้งเดิมที่ต้องอาศัยการดำเนินการด้วยมือเท่านั้น แม้ว่าในด้านความแม่นยำของการรู้จำชื่อเฉพาะบางคลาสจะยังอยู่ในระดับต่ำที่จำเป็นต้องมีการปรับปรุง แต่การออกแบบระบบให้ผู้ใช้สามารถตรวจสอบและแก้ไขข้อมูลได้ก่อนการบันทึกข้อมูล ช่วยลดข้อผิดพลาดและเพิ่มความน่าเชื่อถือของระบบโดยรวมได้เป็นอย่างดี

Table 7 User satisfaction evaluation.

Evaluation item		SD	Satisfaction Level
System design and development	4.44	0.57	Highest
Database design and development	4.05	0.50	High
Overall average	4.25	0.54	Highest

นอกจากนี้ หากมีการขยายขอบเขตของระบบในอนาคต เช่น การเพิ่มปริมาณข้อมูลคลังข้อความสำหรับการฝึกโมเดล การพัฒนาฟังก์ชันสำหรับการเตรียมข้อมูล (Pre-Processing) หรือการเชื่อมโยงองค์ความรู้ด้านโรค สมุนไพร และยาพื้นบ้านไทย จากแต่ละภูมิภาคเข้าด้วยกัน ระบบนี้จะสามารถยกระดับจากคลังข้อมูลตำรายาอีสานโบราณดิจิทัลไปสู่การเป็น ระบบสนับสนุนการวิจัยและการประยุกต์ใช้เชิงวิชาการได้ ตัวอย่างเช่น การพัฒนาให้เป็นระบบให้บริการฐานข้อมูลสารสนเทศตำรายาอีสานโบราณออนไลน์ หรือเป็นแหล่งข้อมูลสำหรับงานด้านการประมวลผลภาษาธรรมชาติ (NLP) ของภาษาอีสานโบราณ เป็นต้น ซึ่งจะเพิ่มความสำคัญของระบบทั้งในมิติของการอนุรักษ์มรดกภูมิปัญญาท้องถิ่น สนับสนุนการเรียนการสอนด้านภาษาและวัฒนธรรม ตลอดจนต่อยอดองค์ความรู้ท้องถิ่นสู่การใช้ประโยชน์เชิงนวัตกรรมในระยะยาวได้อย่างเป็นรูปธรรม

สรุปผลการทดลองและข้อเสนอแนะ

การพัฒนาระบบสกัดองค์ความรู้จากตำรายาอีสานโบราณในครั้งนี้ เป็นความพยายามในการนำเอาเทคโนโลยีด้านการประมวลผลภาษาธรรมชาติมาพัฒนาเครื่องมือสำหรับช่วยให้ผู้ที่ทำงานเกี่ยวกับการรวบรวมภูมิปัญญาท้องถิ่นอีสานด้านการรักษาโรคไปจัดเก็บเป็นคลังข้อมูลตำรายาอีสานโบราณในแบบมีโครงสร้าง ที่ง่ายต่อการเข้าถึงและสืบค้น หลังการวิจัยผู้วิจัยมีข้อเสนอแนะ ดังนี้

1. เนื่องจากการรู้จำชื่อเฉพาะยังมีความผิดพลาดและระดับความถูกต้องยังไม่สูงมากนัก โดยเฉพาะการจำแนกชื่อยา ชื่อโรคหรืออาการของโรค จึงควรเพิ่มประสิทธิภาพการรู้จำชื่อเฉพาะให้สูงขึ้น ด้วยการฝึกโมเดลด้วยเทคนิควิธีใหม่ ๆ และการสร้างคลังข้อความที่มีขนาดใหญ่เพียงพอสำหรับฝึกฝนโมเดล และมีความสมดุลกันในทุกคลาส สำหรับด้านการพัฒนาระบบสกัดองค์ความรู้จากตำรายาอีสานโบราณ ควรมีฟังก์ชันสำหรับแก้ไขข้อมูลชื่อเฉพาะที่ถูกจัดเก็บในฐานข้อมูลเอาไว้แล้วได้ เพื่อให้สามารถปรับปรุงคลาสที่โมเดลทำนายผิดให้ถูกต้องได้ในภายหลัง ซึ่งจะช่วยให้ข้อมูลมีความถูกต้อง น่าเชื่อถือเพิ่มสูงขึ้น

2. นักภาษาโบราณมักต้องการทราบความถี่ของการเกิดคำศัพท์แต่ละคำ โดยเฉพาะชื่อยา ชื่อโรค อาการของโรค และชื่อสมุนไพร เพื่อใช้ในการวิเคราะห์ ยา โรค และสมุนไพร ที่มีการนำมาใช้รักษาโรคต่าง ๆ บ่อย ๆ จึงควรมีการเก็บข้อมูลความถี่ในการพบเห็นคำศัพท์ดังกล่าวด้วย

3. ควรมีการนำเอา ชื่อยา ชื่อโรค ชื่อสมุนไพร ซึ่งเป็นคำศัพท์ภาษาอีสานไปเชื่อมโยงกับ ชื่อยา ชื่อโรค ชื่อสมุนไพร ในถิ่นอื่น เพื่อให้เกิดความเชื่อมโยงกับตำรับยาในภูมิภาคอื่นที่มีภาษาถิ่นแตกต่างกัน เช่น ภาษาไทยถิ่นเหนือ ใต้ กลาง ที่ช่วยให้แพทย์แผนไทยสามารถนำไปศึกษาค้นคว้าเพิ่มเติมได้

4. ชุดข้อมูลสำหรับการนำมาฝึกฝนโมเดลอาจมี 2 รูปแบบ ทั้งแบบปรับปรุงการสะกดคำตามแบบปัจจุบัน และสะกดคำตามแบบโบราณ เพื่อให้รองรับการปริวรรตตำรายาอีสานโบราณแบบรักษาอักขรวิธีการสะกดคำตามแบบโบราณ (คำปริวรรต) ซึ่งคำอ่านจะเป็นแบบภาษาอีสานโบราณ และคำอ่านแบบแปลเป็นภาษาไทยปัจจุบัน (แปลคำศัพท์จากภาษาอีสานเป็นภาษาไทยมาตรฐาน) ด้วย

5. การสกัดองค์ความรู้ควรรองรับข้อมูลไฟล์ข้อมูลประเภทอื่น ๆ เช่น pdf หรือ รูปภาพ นอกเหนือจากนำเข้าข้อมูลในรูปแบบไฟล์ข้อความ (Text File) หรือการพิมพ์ข้อความเข้าไปเท่านั้น

6. ควรพัฒนาระบบแปลตำรายาอีสานโบราณต้นฉบับซึ่งอยู่ในรูปแบบเอกสารภาพใบลานตำรายาอักษรไทยน้อยหรืออักษรธรรมอีสาน เป็นเอกสารข้อความคำอ่านอักษรไทยปัจจุบัน (Optical Character Recognition: OCR อักษรไทยน้อย/อักษรธรรมอีสาน) เพื่อทำให้รวบรวมข้อมูลนำมาสร้างคลังข้อความขนาดใหญ่ทำได้ง่ายและรวดเร็วมากขึ้น

กิตติกรรมประกาศ

ขอขอบคุณมหาวิทยาลัยเทคโนโลยีราชมงคลกรุงเทพ ผู้สนับสนุนทุนวิจัย ประจำปี พ.ศ. 2565 และสถาบันวิจัยศิลปะและวัฒนธรรมอีสาน มหาวิทยาลัยมหาสารคาม ที่ให้ความอนุเคราะห์และอนุญาตการนำข้อมูลตำรายาของสถาบันมาใช้ในงานวิจัยนี้

เอกสารอ้างอิง

จินตนา พลศรี, นิศาชล จำนงศรี, และจีรยุทธ ไชยจารุวณิช. (2565). การพัฒนาคลังศัพท์สำหรับการแปลงคำปริวรรตของอักษรไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสานด้วยคอมพิวเตอร์ [Development of a lexicon for computerized transliteration of Thai Noi script into Isan Thai dialect readings]. *วารสารวิทยาศาสตร์ประยุกต์, 22*(1), 1–14.

จินตนา พลศรี. (2564). *กรอบการทำงานในการแปลงคำปริวรรตของอักษรไทยน้อยเป็นคำอ่านภาษาไทยถิ่นอีสาน* [A framework for transliterating Thai Noi script into Isan Thai dialect readings] [Doctoral dissertation, Suranaree University of Technology].

สำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ. (2555). *แผนพัฒนาเศรษฐกิจและสังคมแห่งชาติ ฉบับที่ 11 พ.ศ. 2555–2559* [The 11th National Economic and Social Development Plan, 2012–2016]. https://www.nesdb.go.th/download/article/article_20160323112431.pdf

สำนักเลขาธิการนายกรัฐมนตรี ทำเนียบรัฐบาล. (2567). *“เดชอิศม์” หนุนยกระดับสมุนไพรไทย นวดไทย ขับเคลื่อนเศรษฐกิจประเทศ* [“Dechis” supports elevating Thai herbal medicine and Thai massage to drive the national economy]. https://www.thaigov.go.th/news/contents/details/89674

Chalamkate, T., Thinnachote, C., & Thamrongrattanarit, A. (2022). The development of geo-names extraction from Twitter texts data by conditional random fields. *Journal of Engineering and Digital Technology (JEDT), 10*(2), 1–10.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kannikaklang, N., Thamviset, W., & Wongthanavasu, S. (2024). BiLSTCAN: A novel SRS-based bidirectional long short-term capsule attention network for dynamic user preference and next-item recommendation. *IEEE Access, 12*, 6879–6899. https://doi.org/10.1109/ACCESS.2024.3351283

Kuran, E. C., Elen, A., & Kuran, U. (2024). A hybrid BiLSTM-CRF model with BERT embeddings for named entity recognition on confidential synthetic data. *Proceedings of the V. International Science and Innovation Congress (INSI)*. https://www.researchgate.net/publication/385101301

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering, 34*(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

Likert, R. (1961). *New patterns of management.* McGraw-Hill.

Makarenkov, V., Rokach, L., & Shapira, B. (2019). Choosing the right word: Using bidirectional LSTM tagger for writing support systems. *Engineering Applications of Artificial Intelligence, 84*, 1–10.

Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition approaches. *International Journal of Computer Science and Network Security, 8*(2), 339–344.

Polsri, J. (2024). CRF-named entity recognition model for ancient Isan medicine texts. *In Proceedings of the 2024 12th International Electrical Engineering Congress* (iEECON) (pp. 1–5). IEEE.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, 45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE, 78*(10), 1550–1560. https://doi.org/10.1109/5.58337

Yadav, V., & Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. *In Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2145–2158). Association for Computational Linguistics.

Zhang, Y., & Wang, L. (2023). *Named entity recognition in electronic health records: A hybrid approach.* National Center for Biotechnology Information. https://pmc.ncbi.nlm.nih.gov/articles/PMC10651400/