ระบบสกัดนิพจน์ทางการแพทย์สำหรับระบบสนับสนุนการตัดสินใจของแพทย์

CLINICAL ENTITY RECOGNITION SYSTEM FOR CLINICAL DECISION SUPPORT SYSTEM

ศศลักษณ์ อุบลวิโรจน์[1]
ดร.ธนภัทร ฆังคะจิตร[2]

บทคัดย่อ

ในปัจจุบันแพทย์ผู้เชี่ยวชาญต้องให้บริบาลผู้ป่วยจำนวนมากขึ้นเนื่องจากปัญหาความขาดแคลนแพทย์ ประกอบกับนโยบายสาธารณสุขของประเทศไทยทำให้สามารถเข้าถึงการรักษาในโรงพยาบาลได้มากขึ้น ส่งผลทำให้แพทย์มีเวลาที่ใช้ในการรักษาน้อยลง เพื่อช่วยการทำงานของแพทย์และลดความคลาดเคลื่อนที่อาจเกิดขึ้น จึงต้องการนำระบบสนับสนุนการตัดสินใจของแพทย์มาช่วยเสริมการทำงานของแพทย์

งานวิจัยนี้จึงนำเสนอระบบการสกัดนิพจน์ทางการแพทย์ที่สามารถใช้งานได้กับข้อมูลบันทึกการรักษาของประเทศไทยโดยการนำเทคนิคการระบุนิพจน์ (Named Entity Recognition) มาประยุกต์ร่วมกับใช้ฐานข้อมูลตัวอักษรย่อและฐานข้อมูลอาการป่วยภาษาไทย โดยสามารถให้มีค่า F1-score สำหรับการสกัดนิพจน์ทางการแพทย์ของกลุ่มโรคทางเดินปัสสาวะอยู่ที่ร้อยละ 79.62

คำสำคัญ : การระบุนิพจน์สำคัญ, ระบบสนับสนุนการตัดสินใจของแพทย์

ABSTRACT

In the current healthcare landscape, physicians are faced with the challenge of managing a larger number of patients due to physician shortages. Thailand's healthcare policies, which aim to enhance accessibility to hospital care, coupled with the increasing number of patients, have reduced the available time for each patient. To support physicians and reduce medication errors, there is a need for decision support systems.

---

[1] นักศึกษาหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมข้อมูลขนาดใหญ่

[2] ที่ปรึกษาสารนิพนธ์

This research presents a medical term extraction system that can be applied to the treatment records in Thailand. The system utilizes Named Entity Recognition (NER) techniques and leverages both abbreviation and symptom databases in Thai. The system achieves an F1-score of 79.62% for extracting medical terms related to urinary tract disorders.

Keyword: Named Entity Recognition, Clinical Decision Support System

บทนำ

ประเทศไทยเป็นประเทศหนึ่งที่ให้ความสำคัญต่อระบบสาธารณสุข ส่งผลทำให้ประชาชนสามารถเข้าถึงการรักษาในโรงพยาบาลได้มากขึ้น และมีแนวโน้มสูงขึ้น คาดการณ์อัตราส่วนแพทย์ตาม “แผนกำลังคนตามการจัดระบบบริการโดยเขตสุขภาพ กระทรวงสาธารณสุขปี 2563 – 2567” พบว่าแพทย์ 1 คนต้องดูแลคนผู้ป่วยจำนวนกว่า 1,814 คน นอกจากนี้ยังพบปัญหาการขาดแคลนแพทย์ในบางพื้นที่อีกด้วย ทำให้แพทย์ต้องดูแลผู้ป่วยจำนวนมากกว่าจำนวนดังกล่าว ส่งผลให้แพทย์ไม่มีเวลาการดูแลผู้ป่วย พร้อมทั้งมีภาระงานมาก ส่งผลทำให้เกิดความเหนื่อยล้าในขณะการอ่านประวัติเก่าของผู้ป่วย ทำให้เกิดการอ่านข้อมูลการรักษาผิดพลาดและส่งผลทำให้เกิดความคลาดเคลื่อนทางการรักษาได้

เพื่อช่วยลดปัญหาความคลาดเคลื่อนที่เกิดขึ้นจากการรักษาที่เกิดจากการผิดพลาดของการอ่านบันทึกการรักษา จึงมีแนวคิดที่นำเทคโนโลยีด้านการประมวลผลทางภาษาศาสตร์ (Natural language processing) มาประยุกต์ใช้เพื่อให้แพทย์อ่านข้อมูลการรักษาหรือประวัติการรักษาได้รวดเร็ว พร้อมทั้งมีความถูกต้องในการอ่านข้อมูลอาการป่วย ส่งผลทำให้แพทย์มีเวลาในการรักษามากขึ้นและยังเป็นส่วนที่ช่วยให้แพทย์สามารถวินิจฉัยโรคที่ผู้ป่วยอาจจะเป็นได้มั่นใจยิ่งขึ้น การพัฒนาระบบดังกล่าวได้นำเทคนิคการสกัดนิพจน์เฉพาะหรือ Named Entity Recognition (NER) ซึ่งเป็นเทคนิคในการประมวลทางภาษาศาสตร์ (Natural language processing) มาประยุกต์ใช้ในการสกัดนิพจน์สำคัญโดยจะเป็นการสกัดนิพจน์ที่มีความจำเพาะเจาะจงในด้านการแพทย์ประยุกต์ใช้ให้เหมาะสมลักษณะของบันทึกทางการแพทย์ของประเทศไทย

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลในการสกัดนิพจน์สำคัญในบันทึกข้อมูลทางการแพทย์ของไทย เพื่อลดความผิดพลาดที่อาจจะเกิดจากความคลาดเคลื่อนในการอ่านบันทึกข้อมูลทางการแพทย์ พร้อมทั้งยังสามารถนำมาใช้ในการสกัดอาการสำคัญเพื่อนำไปใช้ในการวิเคราะห์ข้อมูลได้อีกด้วย โดยประยุกต์ใช้เทคนิคการเรียนรู้เชิงลึก (Deep Learning) ที่พัฒนาพื้นฐานมาจาก Bidirectional Encoder Representations from Transformers (BERT) โดยมีการปรับปรุงให้มีขนาดเล็กกว่า BERT แต่ยังคงมีประสิทธิภาพเช่นเดิม ได้แก่

Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) เพื่อให้ทำงานได้รวดเร็วขึ้น และนำไปแสดงโดยผ่าน Application Programming Interface (API) พร้อมทั้งแสดงผลผ่านแอพพลิเคชัน Line

วัตถุประสงค์ของงานวิจัย

1. เพื่อสร้างระบบสกัดนิพจน์ที่มีความจำเพาะเจาะจงทางกับศัพท์ทางด้านการแพทย์ และสามารถใช้งานกับบันทึกทางการแพทย์ของประเทศไทยได้
2. เพื่อสร้างคลังคำศัพท์ตัวอักษรย่อทางการแพทย์และคำศัพท์อาการป่วยในภาษาไทย
3. เพื่อสร้าง API สำหรับระบบสกัดนิพจน์ทางการแพทย์เพื่อนำไปเชื่อมต่อกับระบบต่างๆ

ขอบเขตงานวิจัย

1. บันทึกทางการแพทย์ทางแผนกศัลยกรรม ในกลุ่มโรคทางเดินปัสสาวะ จำนวน 200 ข้อความ
2. สามารถสกัดนิพจน์ที่สำคัญได้โดยแบ่งเป็น 3 หมวด ได้แก่ ข้อมูลทั่วไป, ตำแหน่งหรืออวัยวะที่พบ, อาการที่พบ

ทฤษฎี และผลงานที่เกี่ยวข้อง

สารนิพนธ์นี้มีวัตถุประสงค์เพื่อสร้างระบบสกัดนิพจน์สำคัญทางการแพทย์เพื่อสนับสนุนการตัดสินใจของแพทย์นั้นมีการศึกษารายละเอียดและงานวิจัยที่เกี่ยวข้องดังนี้

1. Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) เป็นระบบมาตรฐานศัพท์ทางการแพทย์สากลที่มีความสมบูรณ์มากที่สุดในปัจจุบันที่ใช้กับระบบคอมพิวเตอร์ นอกเหนือจากส่วนที่นำมาใช้เป็นฐานความรู้ในการพัฒนารหัสยา อาจนำไปใช้พัฒนาระบบข้อมูลเวชภัณฑ์ เครื่องและอุปกรณ์ทางการแพทย์ได้อีกด้วย ระบบ SNOMET-CT เป็นระบบศัพท์ทางการแพทย์ที่มีความครอบคลุมการแพทย์ในสาขาต่างๆ รวมทั้งทันตแพทย์ เภสัชศาสตร์ พยาบาลศาสตร์ เทคนิคการแพทย์ และสัตวแพทย์ และในประเทศไทยได้มีการเข้าร่วมเป็นสมาชิก SNOMED-CT ในเดือนมกราคม 2022 อีกด้วย

2. การประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) เป็นเทคโนโลยีที่ใช้การประมวลผลข้อมูลภาษาธรรมชาติโดยใช้คอมพิวเตอร์เป็นเครื่องมือหลัก ซึ่งมีวัตถุประสงค์เพื่อช่วยให้คอมพิวเตอร์เข้าใจและวิเคราะห์ข้อมูลภาษาธรรมชาติอย่างมีประสิทธิภาพ โดยการใช้เทคโนโลยี NLP จะช่วยให้เราสามารถทำงานกับข้อมูลภาษาธรรมชาติได้อย่างมีประสิทธิภาพและรวดเร็วมากยิ่งขึ้น

การนำเทคโนโลยี NLP มาใช้งานมีประโยชน์อย่างมากมาย เช่น การพัฒนา chatbot ที่สามารถตอบคำถามและสื่อสารกับผู้ใช้งานได้เหมือนกับมนุษย์ การสร้างระบบแปลภาษาอัตโนมัติที่สามารถแปลภาษาได้

แม้ว่าภาษาที่นำมาแปลจะไม่เหมือนกัน และการสร้างระบบคัดกรองเนื้อหาออกจากข้อความเพื่อค้นหาข้อมูลที่ต้องการได้อย่างง่ายดาย

3. การสกัดนิพจน์สำคัญ (Named Entity Recognition: NER) ซึ่งเป็นเทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ที่ทำหน้าที่ระบุประเภทของ Noun phrase ที่ปรากในข้อความ เช่น ประเภท บุคคล สถานที่ วันที่ เวลา เป็นต้น NER มักถูกนำไปใช้เพื่อร่วมวิเคราะห์งานทางด้าน NLP ต่าง เนื่องจาก NER สามารถช่วยระบุ Entity ที่ถูกกล่าวถึงในข้อความ เช่น ชื่อบุคคล สถานที่ เวลา เพื่อนำไปหาความสัมพันธ์ระหว่าง Entity เหล่านั้นในลำดับถัดไป

การพัฒนาแบบจำลอง NER สามารถทำได้หลายวิธี ตั้งแต่วิธี Classical machine learning ที่อาศัยผู้เชี่ยวชาญช่วยระบุคุณสมบัติของข้อมูล (Feature engineering) ในเบื้องต้นให้ระบบ ได้แ Conditional Random Fields (CRF), Support Vector Machines (SVM) ไปจนถึงวิธี Deep Learning ที่ทำการแบ่งข้อความเป็นหน่วยย่อยของคำ (Word) หรือพยัญชนะ (Character) แล้วจึงแปลงหน่วยย่อยของคำ ดังกล่าวเป็น Features และนำเข้าสู่กระบวนการเรียนรู้ Deep Learning ในลำดับถัดไป ได้แก่ Bi-LSTM เป็นต้น

4. Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) เป็นตัวแบบ (model) ที่ถูกสร้างขึ้นจาก BERT (Bidirectional Encoder Representations from Transformers) โดยมีวัตถุประสงค์ในการลดขนาดของโมเดลเพื่อทำให้มีประสิทธิภาพในการทำงานได้รวดเร็วขึ้น โมเดล BERT เป็นโมเดลที่สร้างขึ้นโดยทีมวิจัยของ Google และได้รับความนิยมสูงเนื่องจากความสามารถในการเข้าใจและสร้างความหมายของประโยคหรือข้อความที่มีความซับซ้อนได้ดี

DistilBERT ใช้เทคนิคที่เรียกว่า "distillation" เพื่อลดขนาดของโมเดล วิธีนี้มีกระบวนการให้โมเดลใหญ่ (เช่น BERT) เป็นโมเดลเล็กขึ้น โดยการโค้ดโมเดลใหญ่เข้าไปในโมเดลเล็ก เพื่อให้โมเดลเล็กเรียนรู้จากข้อมูลการสอนที่อยู่ในโมเดลใหญ่ โดยเก็บสิ่งที่สำคัญสำหรับการเข้าใจประโยคหรือข้อความและลบสิ่งที่ไม่สำคัญ ซึ่งจะทำให้โมเดลมีขนาดเล็กลง แต่ยังคงความสามารถในการเข้าใจและสร้างความหมายของประโยคหรือข้อความได้ดีเช่นเดิม

5. การวัดประสิทธิภาพของโมเดล สามารถทำได้โดยใช้ metrics ต่อไปนี้:

Precision : สัดส่วนของจำนวนข้อความที่ถูกทำนายว่าเป็นนิพจน์ที่ถูกต้องกับจำนวนนิพจน์ที่ถูกทำนายว่าเป็นนิพจน์

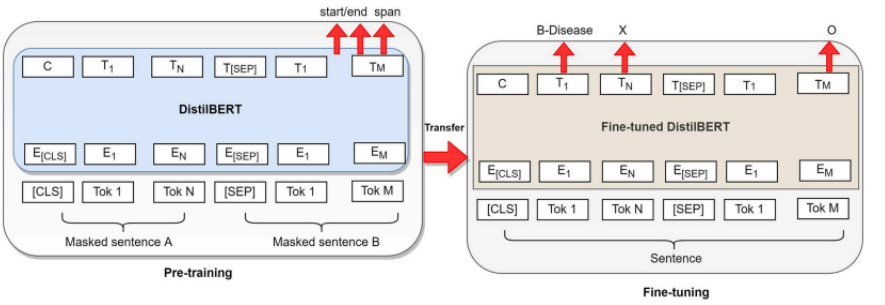
Recall : สัดส่วนของจำนวนข้อความที่ถูกทำนายว่าเป็นนิพจน์ที่ถูกต้องกับจำนวนนิพจน์ที่แท้จริงทั้งหมด

F1-score : คำนวณจากสูตร (2x(PrecisionxRecall))/(Precision + Recall)

งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่พัฒนาขึ้นเป็นงานวิจัยที่จะนำไปใช้ประโยชน์ทางด้านการแพทย์ โดยนำไปใช้งานกับข้อมูลที่มีทั้งภาษาไทยและภาษาอังกฤษอยู่ด้วยกัน เนื่องจากไม่มีงานวิจัยที่ทำการทดลองกับข้อมูลดังกล่าว ดังนั้นจึงทำการทำการศึกษาเกี่ยวกับรายละเอียดการสกัดนิพจน์ทางการแพทย์ที่ทำกับข้อมูลภาษาอังกฤษ ผู้วิจัยจึงรวบรวมงานวิจัยที่เกี่ยวข้องและสรุปรายละเอียดได้ดังนี้

DistilBERT (Raza S, 2022) : การใช้โมเดล DistilBERT และทำการ fine tuning ข้อมูลในส่วน pre-trained โดยทำการเปลี่ยนแปลงเลเยอร์สุดท้ายของโมเดล DistilBERT ให้เหมาะสำหรับการระบุนิพจน์ทางด้านงานชีวิภาพเปลี่ยนจาก pre-trained DistillBERT เป็น fine-tuned DistillBERT

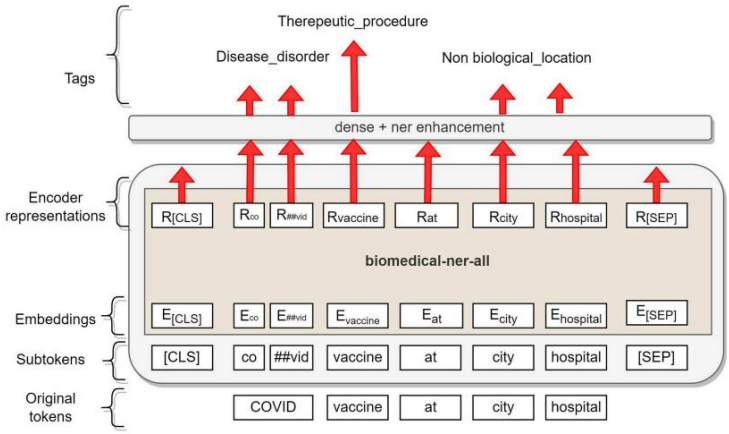


ภาพที่ 1 โครงสร้างการเปลี่ยนจาก Pre-training เป็น fine-tuning

ก่อนนำข้อมูลเข้าโมเดล จะนำข้อมูลมาตัดให้เป็นคำ (tokenized) แปลงข้อความแต่ละคำให้อยู่ในรูปแบบ embedding โดยจัดทำเป็น 3 แบบ ได้แก่ token embedding, segment embedding และ position embedding และรวมกันเพื่อนำเป็นข้อมูลไปใช้ในขั้นตอนถัดไป ซึ่ง token embedding จะเป็นข้อมูลที่แสดงความหมายของแต่ละคำ, segment embedding เป็นตัวช่วยให้โมเดลแยกส่วนต่างๆ ในประโยค และ position embedding จะเป็นการให้ข้อมูลเกี่ยวกับตำแหน่งของคำในประโยค

การแสดงผลของคำที่ได้จากนี้จะถูกส่งผ่านชั้น dense layer ที่รับผิดชอบในการเพิ่มความสามารถให้กับการแสดงผลข้อมูลนำเข้าและแปลงรูปแบบ IOB (Inside, Outside, Beginning) ของชุดข้อมูล CONLL-2003 และ

ทำนายนิพจน์ในของแต่ละคำโดยจะแสดงลเฉพาะคำที่มีค่าความมั่นใจ (Confidence score) มากกว่า 0.4 เท่านั้น



ภาพที่ 2 ขั้นตอนการนำข้อมูลเข้าโมเดล

ผลการทดลองของงานวิจัยดังกล่าวพบว่าโมเดลดังกล่าวมีประสิทธิภาพในการระบุนิพจน์สำคัญทางชีวการแพทย์โดยมีค่า F1-score สำหรับชุดข้อมูล MACCROBAT, NCBI-Disease และ I2b2-2012 เท่ากับ 91.89%, 90.28 และ 89.54 ตามลำดับ ซึ่งมีประสิทธิภาพสูงกว่าโมเดลในการระบุนิพจน์ทางการแพทย์หลายๆ โมเดล เช่น BioBERT v1.2, ClinicalBERT เป็นต้น

Bi-Char-LSTMs (Magge A, 2018) :งานวิจัยดังกล่าวได้นำบันทึกทางการแพทย์ (Clinical notes) โดยนำมาตัดคำและทำ word embedding และนำไปสร้างโมเดลการระบุนิพจน์โดยใช้ bidirectional LSTM โดยกำหนดชั้น hidden unit ไว้ที่ถึง 75 และใช้ Adam เป็น optimizer ด้วย learning rate ที่ 0.005 และมีการกำหนด dropout ที่ 0.5 เพื่อป้องกันการ overfit และใช้โมเดล Random Forest เป็น classifier โดยมีประสิทธิภาพในการระบุนิพจน์ทางการแพทย์โดยมีค่า F1-score ของ micro-average อยู่ที่ 0.81

ระบบสนับสนุนการตัดสินใจของแพทย์ (Clinical Decision Support System) (Sutton RT, 2020) : สามารถแบ่งตามประเภทตามรูปแบบการพัฒนาได้แก่ 1) ใช้องค์ความรู้เก่า โดยการทำมักจะพัฒนาโดยผู้เชี่ยวชาญด้านนั้นๆ เช่น แพทย์ เภสัชกร เป็นต้น มักจะมีการทำงานแบบ rule-base 2) ไม่ได้ใช้องค์ความรู้เก่า มักจะพัฒนาโดยนักพัฒนา โดยจะมีการนำอัลกอริทึมทางคณิตศาสตร์มาใช้ เช่น neural network

ตารางที่ 1 สรุปข้อดีข้อเสียของระบบ CCDS

หัวข้อ	ข้อดี	ข้อเสีย
ความปลอดภัยของผู้ป่วย	ลดอันตรายที่เกิดจากความคลาดเคลื่อนที่อาจจะเกิดขึ้น เช่น การจ่ายยาผิด	ได้รับการแจ้งเตือนที่มากเกินไปทำให้ไม่สนใจ และเมื่อมีข้อแจ้งที่อันตรายทำให้แพทย์ไม่สนใจ
การจัดการผู้ป่วย	ช่วยวางแผนการรักผู้ป่วยได้ดีขึ้น	ทำให้แพทย์ขาดความเชี่ยวชาญ และเชื่อในระบบมากเกินไป
ค่าใช้จ่าย	ช่วยลดค่าการรักษา เช่นการจ่ายยาหรือการรักษาที่ซ้ำซ้อน	ค่าติดตั้งและค่าบำรุงรักษาค่อนข้างมีราคาสูง
ฟังก์ชันการบริหารจัดการ/ระบบอัตโนมัติ	ช่วยให้เลือกหมายเลข ICD10 และทำเอกสารต่างๆ ได้อัตโนมัติ	ต้องมีการอับเดตระบบอยู่เสมอ
การตัดสินใจในการรักษา	ช่วยให้คำแนะนำเกี่ยวกับการรักษาโดยขึ้นกับข้อมูลของผู้ป่วยอัตโนมัติ	แพทย์อาจไม่เห็นด้วยกับสิ่งที่ระบบแนะนำหรืออาจจะมี bias ทำให้ยึดตามระบบมากกว่า
ระบบเอกสาร	ได้ระบบเอกสารที่ดีขึ้น	ระบบอาจจะรวบรวมข้อมูลจากหลายแหล่งที่มาทำให้ข้อมูลนั้นไม่เป็นแพทเทิร์นเดียวกัน ผู้ใช้งานต้องมาปรับเอกสารใหม่อีกครั้ง

ระเบียบวิธีวิจัย

1. การสร้างแบบจำลอง

1.1 การสร้างคลังคำศัพท์ในรูปฐานข้อมูล

(1) ฐานข้อมูลตัวอักษรย่อทางการแพทย์ การสร้างฐานข้อมูลตัวอักษรย่อทางการแพทย์ถูกพัฒนาจากการนำข้อมูลจากเว็บไซต์ที่มีคำศัพท์ย่อและความหมายเต็มของคำศัพท์ย่อนั้น โดยดึงข้อมูลจากเว็บไซต์ (Web scraping) โดยแบ่งเป็นคำศัพท์ที่ใช้โดยทั่วไปในบันทึกทางการแพทย์, ตำแหน่งของอวัยวะ, คำที่ใช้ในการวินิจฉัยโรค เป็นต้น โดยดึงข้อมูลจากเว็บไซต์โดยใช้ชุดเครื่องมือ Selenium และนำข้อมูลมากเว็บไซต์ทรูปลูกปัญญา (https://www.trueplookpanya.com/knowledge/content/82580/-blog-laneng-lan) และ openmd (https://openmd.com/dictionary/abbreviations)

(2) ฐานข้อมูล SNOMET-CT ข้อมูลดังกล่าวอยู่ในรูปแบบฐานข้อมูล SQL โดยทำการติดตั้งและเรียกดูข้อมูลโดยผ่านโปรแกรม MySQL และนำคำศัพท์มาใช้ในเฉพาะหัวข้อ finding และ body of structure

(3) ฐานข้อมูลอาการป่วยภาษาไทย เตรียมฐานข้อมูลอาการป่วยภาษาไทยคำที่พบบ่อยจากข้อมูลบันทึกทางการแพทย์ของจำนวนขั้นตอนการทำดังต่อไปนี้

3.1 การสกัดข้อความภาษาไทย (Regex) : ข้อความบันทึกทางการแพทย์เป็นบันทึกที่มีการใช้ทั้งภาษาไทยและภาษาอังกฤษ เพื่อสร้างการสกัดข้อความภาษาไทยนำมาสร้างเป็นฐานข้อมูลคลังคำศัพท์โดยใช้ Regex

3.2 การตัดคำ (Word tokenization) : นำข้อมูลทางการแพทย์ทั้งหมดนำมาตัดคำเพื่อนำไปใช้สร้างเป็นคำศัพท์โดยใช้เครื่องมือ pythainlp และตัดคำตามการเว้นวรรค โดยมีการรายละเอียดการเลือกใช้ pythaiNLP และเลือกอัลกอริทึมการตัดคำเป็น multi_cut และ dictionary การนับความถี่ที่พบโดยใช้ Bag of word : นำข้อมูลที่ได้จากข้อ 2 นำมาสร้างเป็นชุดข้อมูล bag of word เพื่อนับความถี่ที่พบทั้งหมดในเอกสารทั้งหมด 200,000 ข้อมูล โดยใช้ Countvectorizer

3.3 การแปลภาษา : นำข้อความจาก bag of word เข้าโมเดลแปลภาษาไทยเป็นภาษาอังกฤษ โดยใช้เครื่องมือ Google translate API เพื่อนำไปเทียบกับฐานข้อมูล Snomed-CT

3.4 การหาความคล้ายคลึงของคำ (Word similarity) : เมื่อได้คู่ศัพท์ภาษาไทยและภาษาอังกฤษจากข้อ 4 นำข้อความทั้งหมดที่ได้จากการทำ Bag of word เปรียบเทียบความคล้ายคลึงกันกับศัพท์ในฐานข้อมูล Snomed-CT โดยการหา word similarity โดยใช้เครื่องมือ spaCy โดยเลือกโมเดลที่สร้างจากชุดข้อมูล ‘en_core_web_md’ ซึ่งพัฒนาและเลือกข้อความที่มี word similarity สูงที่สุดในแต่ละคำ

3.5 ตรวจสอบความถูกต้องของข้อมูล (TH/EN Medical corpus verification) : เรียงลำดับข้อมูลโดยเรียงตามความถี่ที่พบและค่าความคล้ายคลึง (word similarity score) เพื่อเลือกคำศัพท์นำมาตรวจสอบความถูกต้องของข้อมูลและนำไปใช้ในการสร้างฐานข้อมูลอาการป่วยสำหรับการเตรียมข้อมูลบันทึกทางการแพทย์

1.2 การเตรียมข้อมูลบันทึกทางการแพทย์ (Data Preprocessing)

(1) การทำความสะอาดข้อมูล (Cleansing Data) : นำสัญลักษณ์พิเศษต่างๆ ออกจากข้อมูลบันทึกทางการแพทย์ เช่น ‘<’, ‘>’, ‘-’, ‘\’, ‘\n’, ‘?’, ‘:’ เป็นต้น เพื่อเตรียมข้อมูลสำหรับขั้นตอนถัดไป

(2) การแทนที่ตัวอักษรย่อทางการแพทย์ (Replace abbreviation) : นำข้อมูลจากฐานข้อมูลทางการแพทย์มาตรวจสอบกับข้อมูลบันทึกทางการแพทย์โดยตรงพบตามการเว้นวรรค (space) หากตรงตามคลังคำศัพท์ทางการแพทย์ที่สร้างไว้ให้แทนที่ด้วยความหมายที่กำหนด

(3) แทนที่อาการป่วยจากคลังคำศัพท์ทางการแพทย์ (Replace TH/EN medical corpus)
นำข้อความที่ได้จากข้อที่ 2 แทนที่ด้วยฐานข้อมูลคลังคำศัพท์ทางการแพทย์ด้วยการเรียงคำศัพท์ในที่มีความยาวของตัวอักษรสูงสุดและเลื่อนตรวจเช็คข้อความ window slicing) ครั้งละ 1 ตัวอักษร หากตรวจพบข้อความที่ตรงตามข้อมูลจากคลังคำศัพท์ทางการแพทย์จะแทนที่ข้อความดังกล่าวด้วยคู่คำศัพท์นั้น

1.3 สกัดนิพจน์สำคัญทางการแพทย์

นำข้อมูลทางการแพทย์ที่ได้จากข้อ 3.1.2 นำมาสกัดนิพจน์สำคัญทางการแพทย์โดยเลือกที่จะแสดงข้อใน 3 หมวด ได้แก่ ข้อมูลทั่วไป, ตำแหน่ง/อวัยวะที่พบ และอาการสำคัญ โดยเลือกใช้โมเดล Bio-Epidemiology-NER (https://pypi.org/project/Bio-Epidemiology-NER/)

2. การนำไปใช้งาน

เพื่อนำไปสร้างเป็นเป็นผลิตภัณฑ์ต้นแบบ พร้อมทั้งสามารถเก็บข้อมูลต่างๆ จากแพทย์ผู้ทดลองใช้ระบบได้ง่าย จึงนำระบบสกัดนิพจน์สำคัญทางการแพทย์มีการนำไปแสดงผลบน Line application ดังนี้

ส่วนที่ 1 การพัฒนา API โดยรับ input เป็นข้อความทางการแพทย์และส่ง output ของผลการสกัดนิพจน์สำคัญให้อยู่ในรูปแบบ JSON เพื่อนำไปแสดงผลในขั้นตอนถัดไป

ส่วนที่ 2 สร้าง Line official account และเชื่อมต่อกับระบบแชทบอทของบอทน้อย เพื่อเตรียมระบบสำหรับการแสดงผล

ส่วนที่ 3 เชื่อมต่อ API กับระบบแชทบอทของบอทน้อยเพื่อแสดงผลข้อมูลการสกัดนิพจน์ตามรูปแบบข้อความที่ได้ออกแบบไว้

ผลการศึกษา

1. ผลการเตรียมฐานข้อมูลทางการแพทย์

1.1 ฐานข้อมูลคลังคำศัพท์ย่อ : ฐานข้อมูลตัวย่อที่ถูกดึงด้วยการดึงข้อมูลผ่านเว็บไซต์และทำการตรวจสอบความถูกต้องของข้อมูลทั้งหมดได้ชุดฐานข้อมูลจำนวน 1,210 รายการ

Abb	Meaning
PTA	Prior To Admission
bph	Benign Prostatic Hyperplasia
UTI	urinary tract infection
vag	vaginal
V/S	vital signs
Hx	history
ID	Infectious Diseases
IMI	Inferior Myocardial Infarction
K	Potassium
LDL	low-density lipoprotein
LLE	Left Lower Extremity
LLL	left lower lobe
LLQ	left lower quadrant
LN	Lymph Node
LUTS	Lower Urinary Tract Symptoms

ภาพที่ 3 ฐานข้อมูลคลังคำศัพท์ย่อทางการแพทย์

1.2 ฐานข้อมูล SNOMED-CT : ฐานข้อมูล SNOMED-CT ถูกสร้างให้อยู่ในฐานข้อมูล SQL เนื่องจากเป็นฐานข้อมูลที่มีขนาดใหญ่ ประกอบด้วยคลังคำศัพท์มากกว่า 1 ล้านรายการ แต่ในงานวิจัยนี้เลือกใช้เฉพาะข้อมูลที่อยู่ในหมวด finding และ body structure โดยนำข้อมูลดังล่าวออกจากฐานข้อมูล SQL ให้อยู่ในไฟล์ CSV เพื่อให้ง่ายต่อการนำไปใช้ต่อ

term_x	concep	typeId_x	term_y	typeId	eptabili	type
Previous pregnancies (finding)	1.27E+08	Fully specified name	Previous pregnancies	Synonym	Preferred	finding
Basic activity of daily living (finding)	1.29E+08	Fully specified name	BADL	Synonym	Acceptabl	finding
Basic activity of daily living (finding)	1.29E+08	Fully specified name	Basic activity of daily living	Synonym	Preferred	finding
Inspiratory crepitation (finding)	61343002	Fully specified name	Inspiratory crepitation	Synonym	Preferred	finding
Radiating chest pain (finding)	10000006	Fully specified name	Radiating chest pain	Synonym	Preferred	finding
White blood cell abnormality (finding)	1.34E+08	Fully specified name	White blood cell abnormality	Synonym	Preferred	finding
Chlamydia trachomatis nucleic acid detection (finding)	1.34E+08	Fully specified name	Chlamydia trachomatis nucleic acid de	Synonym	Preferred	finding
Cytomegalovirus nucleic acid detection (finding)	1.34E+08	Fully specified name	Cytomegalovirus nucleic acid detectio	Synonym	Preferred	finding
Dengue nucleic acid detection (finding)	1.34E+08	Fully specified name	Dengue nucleic acid detection	Synonym	Preferred	finding
Hantavirus nucleic acid detection (finding)	1.34E+08	Fully specified name	Hantavirus nucleic acid detection	Synonym	Preferred	finding
Hepatitis C nucleic acid detection (finding)	1.34E+08	Fully specified name	Hepatitis C nucleic acid detection	Synonym	Preferred	finding
Herpes simplex nucleic acid detection (finding)	1.34E+08	Fully specified name	Herpes simplex nucleic acid detection	Synonym	Preferred	finding
HIV 1 nucleic acid detection (finding)	1.34E+08	Fully specified name	HIV 1 nucleic acid detection	Synonym	Preferred	finding
HTLV 1 nucleic acid detection (finding)	1.34E+08	Fully specified name	HTLV 1 nucleic acid detection	Synonym	Preferred	finding
Meningococcal nucleic acid detection (finding)	1.34E+08	Fully specified name	Meningococcal nucleic acid detection	Synonym	Preferred	finding
Neisseria gonorrhoeae nucleic acid detection (finding)	1.34E+08	Fully specified name	Neisseria gonorrhoeae nucleic acid de	Synonym	Preferred	finding
Parvovirus B19 nucleic acid detection (finding)	1.34E+08	Fully specified name	Parvovirus B19 nucleic acid detection	Synonym	Preferred	finding
Toxoplasma nucleic acid detection (finding)	1.34E+08	Fully specified name	Toxoplasma nucleic acid detection	Synonym	Preferred	finding
Influenza A antigen level (finding)	1.34E+08	Fully specified name	Influenza A antigen level	Synonym	Preferred	finding
Influenza B antigen level (finding)	1.34E+08	Fully specified name	Influenza B antigen level	Synonym	Preferred	finding

ภาพที่ 4 ฐานข้อมูล SNOMED-CT ในกลุ่มของ finding

1.3 ฐานข้อมูลคลังคำศัพท์อาการป่วยทางการแพทย์ภาษาไทย : ผลลัพธ์การเตรียมฐานข้อมูลคลังคำศัพท์อาการป่วยทางการแพทย์ภาษาไทยได้ถูกสร้างตามกระบวนการตัดคำ เทียบความคล้ายคลึงกับฐานข้อมูลใน SNOMED-CT และตรวจสอบความถูกต้องของข้อมูลได้ชุดข้อมูลทั้งหมด 2,553 รายการ

ThaiWord	Translation
กันกระแทกพื้น	injury of buttock
กรดไหลย้อน	gastroesophageal reflux disease
กระดูกสันหลังคด	scoliosis deformity of spine
กระตุกทั้งตัว	myoclonus
กระวนกระวาย	anxiety
กระวนกระวายใจ	anxiety
กระสับกระส่าย	anxiety
กรีดข้อมือ	cutting own wrists
กลั้นขับถ่ายไม่ได้	incontinence of feces
กลั้นปัสสาวะไม่ค่อยได้	urinary incontinence
กลั้นปัสสาวะไม่ได้	urinary incontinence
กลั้นปัสสาวะไม่อยู่	urinary incontinence
กลัว	fear
กลัวตาย	fear of death

ภาพที่ 5 ฐานข้อมูลคลังคำศัพท์อาการป่วยทางการแพทย์ภาษาไทย

2. ผลการวัดประสิทธิภาพความถูกต้องของโมเดล

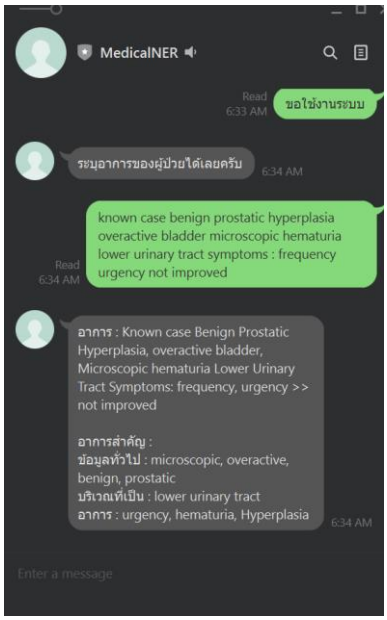
สุ่มข้อมูลบันทึกทางการแพทย์กลุ่มโรคทางเดินปัสสาวะโดยนำมาใช้ในการทดสอบทั้งหมด 200 ข้อมูล โดยผลของค่า Precision, Recall และ F1-score เมื่อเปรียบเทียบกับการสกัดชุดข้อมูลมาตรฐานโดยคำนวณจากนิพจน์ที่สำคัญมีรายละเอียดดังต่อไปนี้

ตารางที่ 2 สรุปผลประสิทธิภาพของโมเดลกับชุดข้อมูลบันทึกทางการแพทย์ภาษาไทย

Dataset	Precision	Recall	F1-score
MACROBBAT 2020	92.10	91.68	91.89
NCBI-Disease	91.68	88.92	90.28
I2b2-2012	90.10	88.98	89.54
บันทึกทางการแพทย์ภาษาไทยกลุ่มโรคทางเดินปัสสาวะในแผนกศัลยกรรม	80.52	78.80	79.62

3. ผลการใช้งาน

นำ API เชื่อมต่อกับระบบแชทบอทและแสดงผลผ่าน Line Official Account ซึ่งได้ตัวอย่างผลการทำงานดังนี้



ภาพที่ 6 การใช้งานผ่าน Line Application

สรุปและอภิปรายผลการวิจัย

ได้พัฒนาที่มีประสิทธิภาพการสกัดนิพจน์สำคัญทางการแพทย์ ประกอบด้วยขั้นตอนดังนี้

(1) ค้นหาเว็บไซต์ที่มีข้อมูลตัวอักษรย่อทางการแพทย์ ใช้การดึงข้อมูลจากเว็บไซต์ (web scaping) และทำการตรวจสอบความถูกต้องของข้อมูล ได้ฐานข้อมูลตัวอักษรย่อทางการแพทย์จำนวน 1,210 รายการ

(2) สร้างฐานข้อมูล SNOMED-CT โดยการดึงข้อมูลที่อยู่ใน MySQL ให้อยู่ในรูปไฟล์ csv ในหมวด finding และ body structure

(3) สร้างฐานข้อมูลอาการป่วยภาษาไทยจากบันทึกข้อมูลทางการแพทย์โดยการตัดคำตามวรรคและใช้เครื่องมือ pythaiNLP นับจำนวนความถี่ที่พบเลือกข้อความที่พบบ่อยนำมาแปลภาษาเบื้องต้นโดยใช้ google translate API นำคำแปลที่ได้หาค่าความคล้ายคลึงกันกับข้อมูลในฐานข้อมูล SNOMED-CT ได้ฐานข้อมูลอาการป่วยภาษาไทยจำนวน 2,553 รายการ

(4) นำข้อมูลบันทึกทางการแพทย์ของแผนกศัลยกรรมในกลุ่มโรคทางเดินปัสสาวะจำนวน 200 รายการ นำมาแทนที่ด้วยฐานข้อมูลคลังคำศัพท์ตัวอักษรย่อทางการแพทย์และฐานข้อมูลอาการป่วยภาษาไทย และสกัดอาการสำคัญด้วยโมเดล Bio-Epidemiology-NER

โดยโมเดลในงานวิจัยนี้ให้ความแม่นยำในการสกัดอาการสำคัญทางการแพทย์ซึ่งได้ค่า Precision, Recall และ F1-score เท่ากัน 80.52%, 78.80% และ 79.62% ตามลำดับ

ข้อสังเกต

1. ผลการสกัดนิพจน์สำคัญทางการแพทย์ยังไม่คำศัพท์ไม่ครอบคลุมข้อมูลทั้งหมดในบันทึกทางการแพทย์และข้อมูลที่เป็นส่วนการปฏิเสธ เช่น ไม่มีอาเจียน ไม่มีปัสสาวะแสบขัด เป็นต้น ยังไม่สามารถสกัดข้อความออกมาในรูปปฏิเสธได้

2. บันทึกทางการแพทย์ที่เป็นข้อมูลรายละเอียดมีค่อนข้างยาว ไม่มีรูปประโยค และมีภาษาไทยปนภาษาอังกฤษ ระบบสกัดนิพจน์สำคัญทางการแพทย์ยังไม่สามารถสกัดข้อความได้แม่นยำ

3. ตัวอักษรย่อทางการแพทย์ที่ใช้ตัวย่อเดียวกัน ระบบยังไม่ระบุได้ว่าตัวอักษรย่อดังกล่าวหมายถึงคำเต็มรายการไหน

ข้อเสนอแนะ

1. วิธีการแปลงข้อมูลโดยใช้ฐานข้อมูลคลังคำศัพท์จำเป็นต้องมีฐานข้อมูลจำนวนมาก หากต้องการเพิ่มความแม่นยำหรือนำไปใช้กับข้อมูลอาการป่วยในโรคอื่นจำเป็นต้องมีการเพิ่มฐานข้อมูลจึงยังไม่เหมาะสมกับการนำขยายส่วน (scale-up)

2. สร้างเว็บแอปพลิเคชันเพื่อให้กำหนดสีของนิพจน์แต่ละประเภทได้ ทำให้มีประสบการณ์ใช้งานที่ดีกว่าการเชื่อมต่อ API กับระบบสารสนเทศของโรคพยาบาล

3. นำระบบสกัดนิพจน์ดังกล่าวนำไปใช้ต่อยอดในขั้นตอนการสกัด feature เพื่อนำข้อมูลไปใช้ในการทำนายโรคตามรหัสโรค ICD-10 หรือใช้ทำวิเคราะห์ข้อมูลสถิติเกี่ยวกับบันทึกการรักษาผู้ป่วย

บรรณานุกรม

Raza S, Reji DJ, Shajan F, Bashir SR. Large-scale application of named entity recognition to biomedicine and epidemiology. PLOS Digital Health. 2022 Dec 7;1(12):e0000152.

Magge A, Scotch M, Gonzalez-Hernandez G. Clinical NER and relation extraction using bi-char-LSTMs and random forest classifiers. InInternational workshop on medication and adverse drug event detection 2018 May 16 (pp. 25-30). PMLR.

Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine. 2020 Feb 6;3(1):17.