ระบบสกัดนิพจน์ทางการแพทย์สำหรับระบบสนับสนุนการตัดสินใจของแพทย์

ศศลักษณ์ อุบลวิโรจน์

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีการศึกษา 2565

CLINICAL ENTITY RECOGNITION SYSTEM FOR CLINICAL DECISION SUPPORT SYSTEM

SASALUK UBOLVIROJ

A Thematic Paper Submitted in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering
Department of Big Data Engineering,
College of Innovation Technology and Engineering
Dhurakij Pundit University
Academic Year 2022

ใบรับรองสารนิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่

หัวข้อสารนิพนธ์ ระบบสกัดนิพจน์ทางการแพทย์สำหรับระบบสนับสนุนการตัดสินใจของแพทย์

เสนอโดย ศศลักษณ์ อุบลวิโรจน์

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาสารนิพนธ์ ดร.ธนภัทร ฆังคะจิตร

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบสารนิพนธ์แล้ว

(ดร.สรรพฤทธิ์ มฤคทัต) ประธานกรรมการ

(ดร.ธนภัทร ฆังคะจิตร) กรรมการที่ปรึกษาสารนิพนธ์

(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น) กรรมการ

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ รับรองแล้ว

(ดร.ชัยพร เขมะภาตะพันธ์) คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 31 เดือน กรกฎาคม พ.ศ. 2566

หัวข้อสารนิพนธ์ ระบบสกัดนิพจน์ทางการแพทย์สำหรับระบบสนับสนุนการตัดสินใจของแพทย์
ชื่อผู้เขียน ศศลักษณ์ อุบลวิโรจน์
อาจารย์ที่ปรึกษา ดร. ธนภัทร ฆังคะจิตร
หลักสูตร วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา 2565

บทคัดย่อ

ในปัจจุบันแพทย์ผู้เชี่ยวชาญต้องให้บริบาลผู้ป่วยจำนวนมากขึ้นเนื่องปัญหาความคลาดแคลนแพทย์ ประกอบกับนโยบายสาธารณสุขของประเทศไทยทำให้สามารถเข้าถึงการรักษาในโรงพยาบาลได้มากขึ้น ส่งผลทำให้แพทย์มีเวลาที่ใช้ในการรักษาน้อยลง เพื่อช่วยการทำงานของแพทย์และลดความคลาดเคลื่อนที่อาจเกิดขึ้น จึงต้องการนำระบบสนับสนุนการตัดสินใจของแพทย์มาช่วยเสริมการทำงานของแพทย์

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

คำสำคัญ: การระบุนิพจน์สำคัญ, ระบบสนับสนุนการตัดสินใจของแพทย์

(อาจารย์ที่ปรึกษา)

Thematic Paper Title	CLINICAL ENTITY RECOGNITION SYSTEM FOR CLINICAL DECISION SUPPORT SYSTEM
Author	SASALUK UBOLVIROJ
Thematic Paper Advisor	Dr. Thanapat Kangkachit
Program	Big Data Engineering
Academic Year	2022

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, the available time for each patient has been reduced. To support physicians and reduce medication errors, there is a need for decision support systems.

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 databases and symptom databases in the Thai language. The system achieves an F1-score of 79.62% for extracting medical terms related to urinary tract disorders.

Keywords: Named Entity Recognition, Clinical Decision Support System

(Advisor)

กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้โดยการให้ความช่วยเหลือของ ดร.ธนภัทร ฆังคะจิตร ซึ่งเป็นอาจารย์ที่ปรึกษาสารนิพนธ์ ที่ได้กรุณาให้คำแนะนำ ตรวจสอบ และแก้ไขข้อบกพร่องต่าง ๆ มาโดยตลอด เพื่อให้สารนิพนธ์ฉบับนี้สมบูรณ์ผู้เขียนจึงขอกราบขอบพระคุณไว้ ณ โอกาสนี้

ผู้เขียนขอกราบขอบพระคุณ ดร.สรรพฤทธิ์ มฤคทัต ที่กรุณาให้เกียรติเป็นประธานโดยมี ผศ.ดร. ดวงใจ จิตคงชื่น เป็นกรรมการในการสอบสารนิพนธ์ ซึ่งได้กรุณาตรวจ แก้ไขสารนิพนธ์ฉบับนี้ให้ถูกต้องสมบูรณ์ยิ่งขึ้น และ นางสาวกุลธิดา รอดบุญ รวมถึงเจ้าหน้าที่บัณฑิตมหาวิทยาลัยธุรกิจบัณฑิตย์ทุกท่านที่ให้ความสะดวก และประสานงาน ในการทำสารนิพนธ์ให้กับผู้เขียนในครั้งนี้ลุล่วงไปด้วยดี

ศศลักษณ์ อุบลวิโรจน์

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ประเทศไทยเป็นประเทศหนึ่งที่ให้ความสำคัญต่อระบบสาธารณสุข ระบบสาธารณสุขได้ถูกพัฒนาอย่างต่อเนื่องทำให้ประชาชนสามารถเข้าถึงการรักษาในโรงพยาบาลได้มากขึ้น และมีแนวโน้มสูงขึ้น คาดการณ์อัตราส่วนแพทย์ตาม “แผนกำลังคน ตามการจัดระบบบริการโดยเขตสุขภาพ กระทรวงสาธารณสุข ปี 2563 – 2567” พบว่าแพทย์ 1 คนต้องดูแลคนผู้ป่วยจำนวนกว่า 1,814 คน[1] นอกจากนี้ยังพบปัญหาการคลาดแคลนแพทย์บางพื้นที่ในอีกด้วย ทำให้ในแพทย์ต้องแลผู้ป่วยจำนวนมากกว่าจำนวนดังกล่าว ส่งผลให้แพทย์ไม่มีเวลาการดูแลผู้ป่วย พร้อมทั้งแพทย์มีภาระงานที่มากส่งผลทำให้เกิดความเหนื่อยล้าในขณะการอ่านประวัติเก่าของผู้ป่วย ทำให้อาจอ่านบางอาการผิดพลาดและส่งผลทำให้เกิดความคลาดเคลื่อนทางการรักษาได้

เพื่อช่วยลดปัญหาความคลาดเคลื่อนที่เกิดขึ้นจากการรักษาที่เกิดจากการผิดพลาดของการอ่านบันทึกการรักษา จึงมีแนวที่นำเทคโนโลยีด้านการประมวลผลทางภาษาศาสตร์ (Natural language processing) มาประยุกต์ใช้เพื่อให้แพทย์อ่านข้อมูลการรักษาหรือประวัติการรักษาได้รวดเร็ว พร้อมทั้งมีความถูกต้องในอาการสำคัญ ส่งผลทำให้แพทย์มีเวลาในการรักษามากขึ้นและยังเป็นส่วนที่ช่วยให้แพทย์สามารถวินิจฉัยโรคที่ผู้ป่วยอาจจะเป็นได้มั่นใจยิ่งขึ้น

การพัฒนาระบบดังกล่าวได้นำเทคนิคการสกัดนิพจน์เฉพาะหรือ Named Entity Recognition (NER) ซึ่งเป็นเทคนิคในการประมวลทางภาษาศาสตร์ (Natural language processing) มาประยุกต์ใช้ในการสกัดนิพจน์สำคัญโดยจะเป็นการสกัดนิพจน์ที่มีความจำเพาะเจาะจงในด้านการแพทย์ โดยสามารถนิพจน์ที่สำคัญในบันทึกทางการแพทย์ได้ เช่น ตำแหน่งที่ผิดปกติหรืออวัยะที่มีความผิดปกติ, อาการเจ็บป่วย เป็นต้น โดยงานวิจัยนี้จะนำเทคนิคดังกล่าวสามารถนำมาประยุกต์ใช้เพื่อให้เหมาะสมลักษณะของบันทึกทางการแพทย์ของประเทศไทย

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลในการสกัดนิพจน์สำคัญในบันทึกข้อมูลทางการแพทย์ของไทย เพื่อลดความผิดพลาดที่อาจจะเกิดจากความคลาดเคลื่อนในการอ่านบันทึกข้อมูลทางการแพทย์ พร้อมทั้งยังสามารถนำใช้ในสกัดอาการสำคัญเพื่อนำไปใช้ในการวิเคราะห์ข้อมูลได้อีกด้วย โดยประยุกต์ใช้เทคนิคการเรียนรู้เชิงลึก (Deep Learning) ที่พัฒนาพื้นฐานมาจาก Bidirectional Encoder

Representations from Transformers (BERT) โดยมีการปรับปรุงให้มีขนาดเล็กกกว่า BERT แต่ยังมีคงประสิทธิภาพเช่นเดิม ได้แก่ Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) ด้วยการใช้เทคนิคการกระจาย (Distillation) ทำให้สามารถทำงานได้รวดเร็วขึ้นเหมาะสมกับการประมวลผลที่ต้องการความเร็วและประหยัดพื้นที่ในการจัดเก็บข้อมูล และนำไปแสดงโดยผ่าน Application Programming Interface (API) พร้อมทั้งแสดงผลผ่านแอพพลิเคชัน Line โดยการทำงานจะมีการรับข้อความผ่านทาง Line และเรียกใช้ระบบการสกัดนิพจน์ทางการแพทย์ที่อยู่บนเซิฟเวอร์ พร้อมทั้งจำแนกระบุข้อความที่จำแนกนิพจน์กลับมาที่ผู้ใช้งาน ผลที่ได้จากงานวิจัยพบว่า โมเดลสามารถสกัดนิพจน์ที่สำคัญได้

1.2 วัตถุประสงค์งานวิจัย

1.2.1 เพื่อสร้างระบบสกัดนิพจน์ที่มีความจำเพาะเจาะจงทางกับศัพท์ทางด้านการแพทย์ และสามารถใช้งานกับบันทึกทางการแพทย์ของประเทศไทยได้

1.2.2 เพื่อสร้างคลังคำศัพท์ตัวอักษรย่อทางการแพทย์และคำศัพท์อาการป่วยในภาษาไทย

1.2.3 เพื่อ API สำหรับระบบสกัดนิพจน์ทางการแพทย์เพื่อนำไปเชื่อมต่อกับระบบต่างๆ

1.3 ขอบเขตงานวิจัย

1.3.1 บันทึกทางการแพทย์ทางแผนกศัลยกรรม ในกลุ่มโรคทางเดินปัสสาวะ จำนวน 200 ข้อความ

1.3.2 สามารถสกัดนิพจน์ที่สำคัญได้โดยแบ่งเป็น 3 หมวด ได้แก่ ข้อมูลทั่วไป, ตำแหน่งหรืออวัยวะที่พบ, อาการที่พบ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ได้คลังคำศัพท์ตัวอักษรย่อทางการแพทย์และอาการป่วยภาษาไทย

1.4.2 ป้องกันการเกิดความคลาดเคลื่อนทางการรักษาที่เกิดจากความผิดพลาดในการอ่านข้อความ

1.5 นิยามศัพท์

1.5.1 Clinical Decision Support Systems (CDSS) หมายถึง ระบบสนับสนุนการตัดสินใจทางคลินิก ซึ่งจะมีระบบซอฟต์แวร์ที่ช่วยในการตัดสินใจเกี่ยวกับการจัดการ การรวบรวมข้อมูล การวิเคราะห์ข้อมูลทางคลินิกของผู้ป่วย ซึ่งระบบจะมีข้อมูลความรู้ทางคลินิกและข้อมูลที่เกี่ยวข้องกับผู้ป่วย มีระบบการแจ้งเตือน การวิจารณ์ การตีความ การวินิจฉัย ตลอดจนการให้คำแนะนำในการดูแลผู้ป่วย อย่างเหมาะสม

บทที่ 2
แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานศึกษาวิจัยนี้มีวัตถุประสงค์เพื่อสกัดคำสำคัญที่มีอยู่ในบันทึกทางการแพทย์ของประเทศไทย ด้วยการประยุกต์ใช้เรียนรู้เชิงลึก (Deep Learning) โดยมีการศึกษาข้อมูลเพื่อเติมและงานวิจัยที่เกี่ยวข้องดังรายการต่อไปนี้

2.1 Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT)

2.2 Natural Language Processing

2.3 Name Entity Recognition

2.4 การเรียนรู้เชิงลึก (Deep Learning)

2.5 Distilled Bidirectional Encoder Representations from Transformers (DistilBERT)

2.6 การวัดประสิทธิภาพของโมเดล

2.7 งานวิจัยที่เกี่ยวข้อง

2.1 Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT)

Systematized Nomenclature of Medicine Clinical Terms [2] หรือ SNOMED-CT เป็นระบบมาตรฐานศัพท์ทางการแพทย์สากลที่มีความสมบูรณ์มากที่สุดในปัจจุบันที่ใช้กับระบบคอมพิวเตอร์ นอกเหนือจากส่วนที่นำมาใช้เป็นฐานความรู้ในการพัฒนารหัสยา อาจนำไปใช้พัฒนาระบบข้อมูลเวชภัณฑ์ เครื่องและอุปกรณ์ทางการแพทย์ได้อีกด้วย ระบบ SNOMET-CT เป็นระบบศัพท์ทางการแพทย์ที่มีความครอบคลุมการแพทย์ในสาขาต่าง รวมทั้ง ทันตแพทย์ เภสัชศาสตร์ พยาบาลศาสตร์ เทคนิคการแพทย์ และสัตวแพทย์ และในประเทศไทยได้มีการเข้าร่วมเป็นสมาชิก SNOMED-CT ในเดือนมกราคม 2022 อีกด้วย

SNOMED-CT พัฒนาขึ้นมาเนื่องจาการพยายามแก้ไขข้อบกพร่องที่พบจากการใช้รหัส ICD โดยแตกต่างจาก ICD คือ ICD เป็นระบบการให้รหัสเพื่อใช้ในการวินิจฉัยโรค และสรุปผลเป็นข้อมูลทางสถิติซึ่งไม่สามารถนำข้อมูลในบันทึกทางการแพทย์มาใช้ในการวิเคราะห์ร่วมได้ ในทางกลับกัน SNOMED-CT คือคลังคำศัพท์ทางการแพทย์ที่มีขนาดใหญ่กว่า 300,000 คำ ทำให้คลอบคลุมคำทางคลินิกมากที่สุด

คลังคำศัพท์ SNOMED-CT เป็นคลังคำศัพท์ที่มีความสำคัญต่อการแพทย์ สามารถช่วยปรับปรุงความถูกต้องของบันทึกทางการแพทย์ ช่วยสนับสนุนการตัดสินใจของแพทย์ เพิ่มคุณภาพด้านการรักษาและช่วยลดความคลาดเคลื่อนของงารแพทย์

2.2 การประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) [3,4]

NLP หรือ Natural Language Processing เป็นเทคโนโลยีที่ใช้การประมวลผลข้อมูลภาษาธรรมชาติโดยใช้คอมพิวเตอร์เป็นเครื่องมือหลัก ซึ่งมีวัตถุประสงค์เพื่อช่วยให้คอมพิวเตอร์เข้าใจและวิเคราะห์ข้อมูลภาษาธรรมชาติอย่างมีประสิทธิภาพ โดยการใช้เทคโนโลยี NLP จะช่วยให้เราสามารถทำงานกับข้อมูลภาษาธรรมชาติได้อย่างมีประสิทธิภาพและรวดเร็วมากยิ่งขึ้น

การนำเทคโนโลยี NLP มาใช้งานมีประโยชน์อย่างมากมาย เช่น การพัฒนา chatbot ที่สามารถตอบคำถามและสื่อสารกับผู้ใช้งานได้เหมือนกับมนุษย์ การสร้างระบบแปลภาษาอัตโนมัติที่สามารถแปลภาษาได้แม้ว่าภาษาที่นำมาแปลจะไม่เหมือนกัน และการสร้างระบบคัดกรองเนื้อหาออกจากข้อความเพื่อค้นหาข้อมูลที่ต้องการได้อย่างง่ายดาย

การทำงานของเทคโนโลยี NLP นั้นสามารถแบ่งออกเป็นหลายขั้นตอน โดยมีหลักการทำงานหลักดังต่อไปนี้

2.2.1 การตัดคำ (Words tokenization)

การตัดคำเป็นการแบ่งส่วนคำ (Tokenization) คำถือเป็นงานพื้นฐานการประมวลผลภาษาธรรมชาติ สำหรับการตัดคำภาษาไทยนั้นจะมีความยุ่งยากมากกว่า เช่นเดียวกับภาษาจีน ญี่ปุ่นและเกาหลี เพราะเป็นภาษาที่ไม่มีการแบ่งกลุ่ม ลักษณะการเขียนมันเรียงเป็นคำต่อเนื่องและไม่มีรูปร่างของประโยคที่ชัดเจน

ในกรณีที่เป็นการตัดคำในภาษาอังกฤษ ฝรั่งเศส หรือสเปน ซึ่งเป็นภาษาที่มีรากฐานมาจากภาษาละตินมักจะไม่มีปัญหาและทำการตัดคำง่ายกว่าภาษาที่ไม่มีได้มีรากฐานมาจากภาษาละติน เพราะภาษาเหล่านี้มันมีเครื่องหมายคั่นคำ (Delimiting) เช่น ช่องว่าง (Spaces) เซมิโคลอน (Semi-Colon) จุลภาค (Comma) เครื่องหมายคำพูด (Quote) และจุด (Period)

สำหรับภาษาที่ไม่แบ่งส่วนมีการศึกษาเทคนิคเพื่อแก้ไขปัญหาการแบ่งคำ สามารถแบ่งออกเป็น 2 วิธี คือ ตามพจนานุกรม (Dictionary-Based: DCB) และการเรียนรู้ด้วยเครื่อง (Machine Learning Based: MLB)

ภาพที่ 2.1 ตัวอย่างการตัดคำในภาษาไทย (Word tokenization)

ที่มา: (https://www.nectec.or.th/innovation/innovation-software/lextoplus.html)

จากการศึกษาพบว่าการตัดคำภาษาไทยที่มีประสิทธิภาพและเป็นที่น่าพอใจ มักจะเป็นการตัดคำที่ได้จากวิธีการตัดคำตามพจนานุกรม โดยวิธีการตัดคำตามพจนานุกรม จะเป็นการใช้ชุดคำศัพท์จากพจนานุกรมในการแยกและแบ่งข้อความขั้นตอนการแยกจะค้นหาชุดของอักขระตามพจนานุกรมเพื่อค้นหาคำที่ตรงกัน ประสิทธิภาพการตัดคำตามพจนานุกรมจะขึ้นกับคุณภาพและขนาดของคำที่ต้องการตัด พจนานุกรมที่ใช้มีคำที่ค่อนข้างง่ายและตรงไปตรงมาซึ่งมักจะมีปัญหา เช่น ปัญหาคำที่ไม่รู้จัก เป็นคำที่ไม่พบในพจนานุกรม หรือปัญหาความกำกวมของคำที่ทำการตัด ปัญหาเหล่านี้สามารถแก้ไขด้วยเทคนิคต่างๆ เช่น

(1) เทคนิคการตัดคำภาษาไทยด้วยพจานุกรมแบบเปรียบเทียบคำที่ยาวที่สุด (Longest Matching)

ในภาษาไทยมีการเขียนตัวอักษรโดยไม่มีขอบเขตของคำที่ชัดเจน โดยคำที่เขียนมักอยู่กับบริบทซึ่งมีหลากหลายวิธีในการแบ่งเป็นคำ ยกตัวอย่าง เช่น

คำว่า “อาจอง” สามารถแบ่งออกเป็น “อา - จอง” หรือ “อาจ - อง”

หรือคำว่า “นั่งตากลม” สามารถแบ่งออกเป็น “นั่ง - ตา – กลม” หรือ “นั่ง – ตาก – ลม”

จากตัวอย่างดังกล่าวจะเห็นได้ว่ามีความซับซ้อนในการระบุคำจึงได้มีการนำเทคนิคการ

เปรียบเทียบคำที่ยาวที่สุดมากใช้ในการแบ่งคำ การทำงานจะอ่านข้อความจากซ้ายไปขวาแล้วนำคำที่ได้ไปเปรียบเทียบกับคำในพจนานุกรมและเลือกคำที่ยาวที่สุด อย่างไรก็ตามคำที่ยาวที่สุดที่ได้อาจจะไม่สอดคล้องกับความเป็นจริง

(2) เทคนิคการตัดคำภาษาไทยด้วยพจนานุกรมแบบที่สอดคลองมากที่สุด (Maximal Matching)

เทคนิคการตัดคำภาษาไทยด้วยพจนานุกรมแบบที่สอดคล้องมากที่สุด เป็นอีกเทคนิคที่ใช้ในการแก้ไขข้อบกพร่องของเทคนิคการตัดคำภาษาไทยด้วยพจนานุกรมแบบเปรียบเทียบคำที่ยาวที่สุด และเลือกตัดคำที่เป็นไปได้ทั้งหมดของประโยคนั้นๆ ก่อน จากนั้นจึงจะทำการเลือกรูปแบบที่เหมาะสมที่สุดโดยการพิจารณาจาก จำนวนคำที่ตัดได้ และรูปแบบของประโยคที่มีจำนวนคำน้อยที่สุดจะถูกคัดเลือกให้เป็นรูปแบบที่สอดคล้องที่สุด

2.2.2 การตัดคำหยุด (Stop-Word Removal)

การตัดคำหยุด เป็นการตัดคำหรือสัญลักษณ์ที่พบบ่อยในเอกสาร แต่คำหรือสัญลักษณ์เหล่านั้นไม่ได้ส่งผลต่อใจความสำคัญและไม่มีนัยสำคัญต่อการวิเคราะห์ข้อมูลในเอกสาร ดังนั้นเมื่อทำการตัดคำเหล่านั้นออกไปแล้วก็ไม่ทำให้ใจความสำคัญในเอกสารนั้นๆ เปลี่ยนแปลงตัวอย่างคำหยุดที่มักปราฏในเอกสาร เช่น คำในกลุ่มบุพบท (Prepositions) เป็นคำที่นำหน้าคำนามเพื่อแสดงความสัมพันธ์ของคำนามอีกคำในประโยค เช่น in, on, with, so, ใต้, บน, ริม เป็นต้น

คำในกลุ่มสันธาน (Conjuction) เป็นคำที่เชื่อมต่อคำอื่นหรือกลุ่มคำ เช่น and, or, but, ทั้ง...และ , ทั้ง...หรือ, ...หรือ...และอื่นๆ

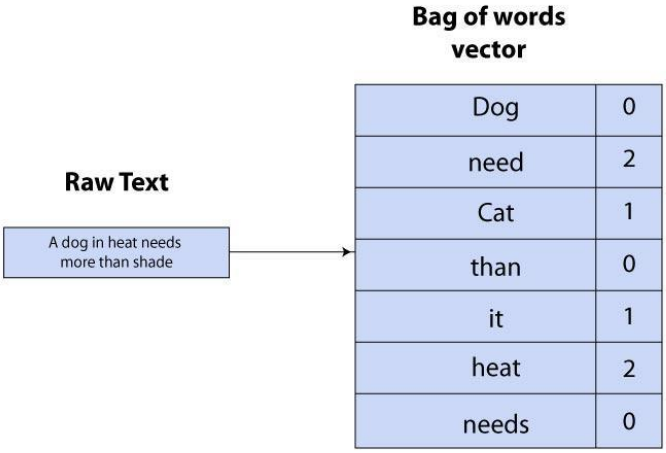
คำในกลุ่มคุณศัพท์ (Adjective) เป็นคำที่ใช้บอกลักษะและคุณสมบัติต่างๆ ของคำนามว่ามีลักษณะอย่างไร เช่น one, two, many, little, เล็ก, ใหญ่, น้อย เป็นต้น

คำในกลุ่มคำสรรพนาม (Pronoun) เป็นคำที่ใช้เรียกแทนคำนาม อันได้แก่ คน สัตว์ สิ่งของ สถานที่ เพื่อหลีกเลี่ยงการเรียกชื่อนั้นซ้ำๆ ตัวอบ่างคำในกลุ่มนี้ เช่น ผม, ฉัน, ข้าพเจ้า, I, me, it, mine เป็นต้น

2.2.3 การสร้างตัวแทนข้อความ (Text Representation)

เนื่องจากปัจจุบันคอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของข้อความที่เป็นภาษาธรรมชาติได้โดยตรง จึงต้องมีการจำลองข้อความให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถอ่านเข้าใจและสามารถเรียนรู้ได้ โดยการสร้างตัวแทนข้อความที่นิยมใช้ก็คือการจำลองเอสารให้อยู่ในแบบจำลองปริเวกเตอร์ (Vector Space Model: VSM)

การสร้างตัวแทนข้อความในรูปแบบของเวกเตอร์ เป็นหนึ่งในวิธีการแทนเอกสารที่ไม่มีโครงสร้าง (Unstructured Text Document) โดยการแบ่งข้อความให้อยู่ในรูปของถุงคำ (Bag-of-Words: BOW) ซึ่งจะเก็บอยู่ในรูปแบบของเวกเตอร์ โดยกำหนดให้เอกสารแต่ละฉบับเปรียบเสมือนเวกเตอร์ของคำ หรือเรียกว่าการหาค่าน้ำหนักของคำ (Tern Weighting) มักแทนค่าด้วยเลขฐานสอง คือจะมีค่าตั้งแต่ 0 ถึง 1 หากค่าเป็น 0 จะหมายความว่าไม่มีคำนั้นอยู่ในเอกสาร และถ้าหากค่าเป็น 1 ก็หมายความว่าพบคำนั้นในเอกสาร ซึ่งจะได้รูปแบบที่มีลักษณะของการแทนความสัมพันธ์ระหว่างคำ (Words: W) และเอกสาร (Documents: D) ด้วยเวกเตอร์ 2 มิติ



ภาพที่ 2.2 ตัวอย่างการทำ Bag of words

ที่มา: (https://www.analyticssteps.com/blogs/an-optimum-approach-towards-the-bag-of-words-with-code-illustration-in-python)

2.2.4 การให้น้ำหนักคำ (Tern Weighting)

ในการจัดหมวดหมู่เอกสารหรือข้อความมักจะถูกจำลองให้อยู่ในรูปแบบของเวกเตอร์ และเอกสารแต่ละฉบับจะแสดงเป็นเวกเตอร์ ซึ่งประกอบด้วยน้ำหนักของคำศัพท์หลายคำ การจัดหมวดหมู่มักจะเริ่มต้นด้วยการให้น้ำหนักคำ (Term Weighting) ซึ่งจะแสดงให้เห็นถึงความสัมพันธ์ของข้อความที่เกี่ยวข้องกันได้ชัดเจนยิ่งขึ้น

การให้น้ำหนักคำคือการกำหนดค่าน้ำหนักให้กับคำหรือเอกสาร เพื่อแสดงถึงความสำคัญของคำ ซึ่งจะจัดให้อยู่ในรูปแบบของ Vector Space Model (VSM) หรือ Bag-of-Words (BOW) ซึ่งหากคำใดที่พบเป็นจำนวนมากในเอกสารหรือคำที่พบบ่อย แสดงว่าคำเหล่านั้นไม่มีความสำคัญจึงไม่สามารถนำมาใช้เป็นตัวแทนของเอกสารได้

2.3 การสกัดนิพจน์สำคัญ (Named Entity Recognition: NER)

การสกัดนิพจน์สำคัญหรือ Named Entity Recognition ซึ่งเป็นเทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ที่ทำหน้าที่ระบุประเภทของ Noun phrase ที่ปรากฏในข้อความ 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) แล้วจึงแปลงหน่วยย่อยของคำดังกล่าวเป็น Features และนำเข้าสู่กระบวนการเรียนรู้ Deep Learning ในลำดับถัดไป ได้แก่ Bi-LSTM เป็นต้น

ในปัจจุบันมี Library program หรือ API ที่ทำ NER สำเร็จรูปพร้อมใช้งานทั้งภาษาต่างประเทศและภาษาไทย โดยในกรณีภาษาต่างประเทศได้แก่ Python library ของ NLTK (NLTK, n.d.) และ spaCy (spaCy, n.d) ในขณะที่ของไทยมี AI for Thai และ PythaiNLP เป็นต้น

ภาพที่ 2.3 ตัวอย่างการทำ Named Entity Recognition

ที่มา: (https://th.shaip.com/blog/named-entity-recognition-and-its-types)

2.4 การเรียนรู้เชิงลึก (Deep Learning)

การเรียนรู้เชิงลึกเป็นส่วนหนึ่งของการเรียนรู้ของเครื่องจักร (Machine Learning) มีพื้นฐานมาจากโครงข่ายประสาทเทียม (Artificial Neural Network: ANN) เป็นวิธีที่สร้างขึ้นเพื่อนทำให้เครื่องจักรสามารถเรียนรู้ได้โดยใช้ต้นแบบมาจากระบบประสาทของมนุษย์ โดยต้องใส่ข้อมูลเข้าไปในชั้นรับข้อมูล (Input Layer) จากนั้นเครื่องจักรจะนำข้อมูลไปประมวลผลในชั้นซ่อน (Hidden Layer) แล้วจะนำเสนอข้อมูลผลลัพธ์ในชั้นแสดงผล (Output Layer) ทั้งนี้ เนื่องจาก Deep Learning จำเป็นต้องใช้ข้อมูลจำนวนมากในการเรียนรู้ จำเป็นต้องมีการกำกับข้อมูล (Data Tagging) เช่น กำกับว่าข้อความใดสอดคล้องกับเจตนาไหน จงทำให้การกำกับข้อความเป็นงานหลักของการเรียนรู้ประเภทนี้ อย่างไรก็ตามการกำกับข้อความมักมีความซับซ้อนน้อยกว่าการออกแบบ Pattern หรือสกัดคุณลักษณะสำคัญของข้อความ (Feature Extraction) ที่มักต้องอาศัยผู้เชี่ยวชาญ ตัวอย่าง Deep Learning ได้แก่ Convolutional Neural Network (CNN), Long-Term Memory Networks (LSTM) และ Bidirectional Encoder Representations from Transformers (BERT) เป็นต้น

2.5 Distilled Bidirectional Encoder Representations from Transformers (DistilBERT)

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

โครงสร้างของ DistilBERT คล้ายกับ BERT แต่มีขนาดเล็กลง เพื่อลดการใช้ทรัพยากรในการประมวลผล โดย DistilBERT ประกอบด้วยส่วนต่าง ๆ ดังนี้:

1. Transformer Encoder: DistilBERT ใช้สถาปัตยกรรม Transformer Encoder เช่นเดียวกับ BERT เพื่อเข้ารหัสคำหรือประโยคในข้อความ ซึ่งประกอบด้วยหลายเลเยอร์ Transformer Encoder ที่ซ้ำกัน โดยแต่ละเลเยอร์ประกอบด้วยเลเยอร์ Self-Attention และเลเยอร์ Feed-Forward Neural Network เพื่อเรียนรู้ความสัมพันธ์ระหว่างคำหรือประโยคในข้อความ

2. Distillation Layer: เป็นส่วนที่แตกต่างจาก BERT ใน DistilBERT เพื่อลดขนาดของโมเดล โดยจะใช้โค้ดของ BERT ใหญ่ (pre-trained BERT) เข้าไปภายในโมเดลขนาดเล็ก (DistilBERT) และเทรน DistilBERT ให้เรียนรู้จากโค้ด BERT ใหญ่ ในกระบวนการนี้ เนื้อหาสำคัญจะถูกถ่ายทอดจาก BERT ใหญ่ไปยัง DistilBERT เพื่อให้โมเดลขนาดเล็กสามารถรับรู้และสร้างความหมายได้ด้วยความแม่นยำ

3. Pooling Layer: หลังจากเลเยอร์ Transformer Encoder ทุกเลเยอร์ จะมีเลเยอร์ Pooling ที่ใช้ในการรวมคุณลักษณะ (features) ที่เกี่ยวข้องกับคำหรือประโยคในข้อความ โดยปกติ DistilBERT จะใช้เลเยอร์ Pooling แบบเลือกสกัดคุณลักษณะสุดท้าย (CLS Pooling) ซึ่งจะเลือกคุณลักษณะที่สำคัญจากข้อมูลของประโยค และแปลงเป็น Feature Vector ที่เหมาะสมสำหรับการใช้งานต่อไปได้

2.6 การวัดประสิทธิภาพของโมเดล

การวัดประสิทธิภาพของโมเดล NER (Named Entity Recognition) สามารถทำได้โดยใช้ metrics ต่อไปนี้:

2.6.1 Precision

สัดส่วนของจำนวนข้อความที่ถูกทำนายว่าเป็นนิพจน์ที่ถูกต้องกับจำนวนนิพจน์ที่ถูกทำนายว่าเป็นนิพจน์

2.6.2 Recall

สัดส่วนของจำนวนข้อความที่ถูกทำนายว่าเป็นนิพจน์ที่ถูกต้องกับจำนวนนิพจน์ที่แท้จริงทั้งหมด

2.6.3 F1-score

คำนวณจากสูตร (2x(PrecisionxRecall))/(Precision + Recall)

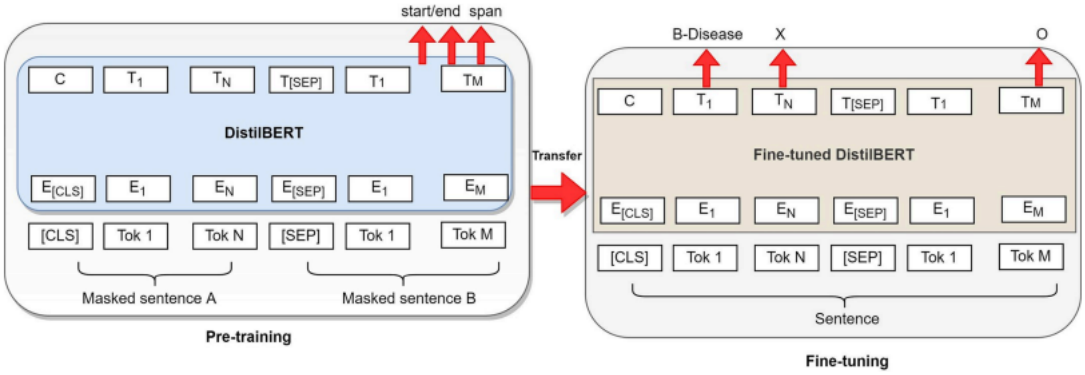
2.7 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่พัฒนาขึ้นเป็นงานวิจัยที่จะนำไปใช้ประโยชน์ทางด้านการแพทย์ โดยนำไปใช้งานกับข้อมูลที่มีทั้งภาษาไทยและภาษาอังกฤษอยู่ด้วยกัน เนื่องจากไม่มีงานวิจัยที่ทำการทดลองกับข้อมูลดังกล่าว ดังนั้นจึงทำการทำการศึกษาเกี่ยวกับรายละเอียดการสกัดนิพจน์ทางการแพทย์ที่ทำกับข้อมูลภาษาอังกฤษ ผู้วิจัยจึงรวบรวมงานวิจัยที่เกี่ยวของและสรุปรายละเอียดได้ดังนี้

2.7.1 Large-scale application of named entity recognition to biomedicine and epidermiology [5]

การวิจัยเกี่ยวกับการระบุนิพจน์ทางด้านงานชีวการแพทย์ค่อนข้างมีอุปสรรคในการพัฒนาอยู่หลายประการ เช่น ชุดข้อมูลที่มีลักษณะที่มีความจำเพาะเจาะจงในด้านต่างๆ และมีชุดข้อมูลค่อนข้างจำกัด รวมทั้งในข้อมูลยังมีข้อมูลอื่น ที่ได้เกี่ยวข้องกับด้านการแพทย์ เช่น ข้อมูลส่วนตัวด้านสังคม เป็นต้น ทำให้การพัฒนานั้นค่อนข้างเป็นไปได้ยาก

ในงานวิจัยดังกล่าวเลือกใช้โมเดล DistilBERT และทำการ fine tuning ข้อมูลในส่วน pre-trained โดยทำการเปลี่ยนแปลงเลเยอร์สุดท้ายของโมเดล DistilBERT ให้เหมาะสำหรับการระบุนิพจน์ทางด้านงานชีวิภาพเปลี่ยนจาก pre-trained DistillBERT เป็น fine-tuned Di



ภาพที่ 2.4 โครงสร้างการเปลี่ยนจาก Pre-training เป็น fine-tuning ของงานวิจัยที่เกี่ยวข้อง 1

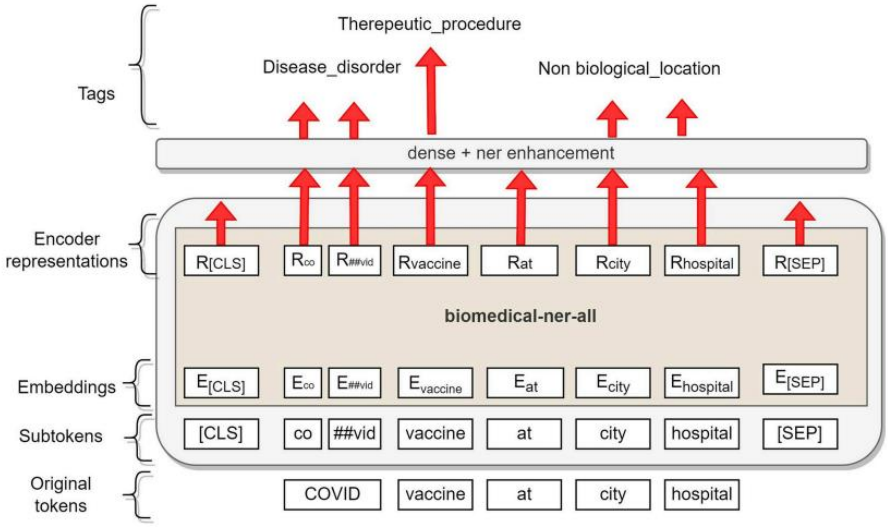
ในขั้นตอนการทำ fine-tuning นักวิจัยได้การแทนที่หัวข้อสำหรับการตรวจจับสิ่งที่เป็น entity ด้วยหัวข้อใหม่โดยหัวข้อทั้งหมดดนั้นจะครอบคลุมนิพจน์ที่เกี่ยวชีวการแพทย์ และมีการตั้งค่าต่าง ดังนี้

ตารางที่ 2.1 parameter ที่ใช้ในการเทรนโมเดลของงานวิจัยที่ 1

Training Data	CoNLL-2003
Optimizer	Adam
Batch size	16
Learning rate	2e-5
จำนวนรอบ	40 epoch
Weight decay	0.01
Drop out	0.1
การวัดผล	F1-score

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

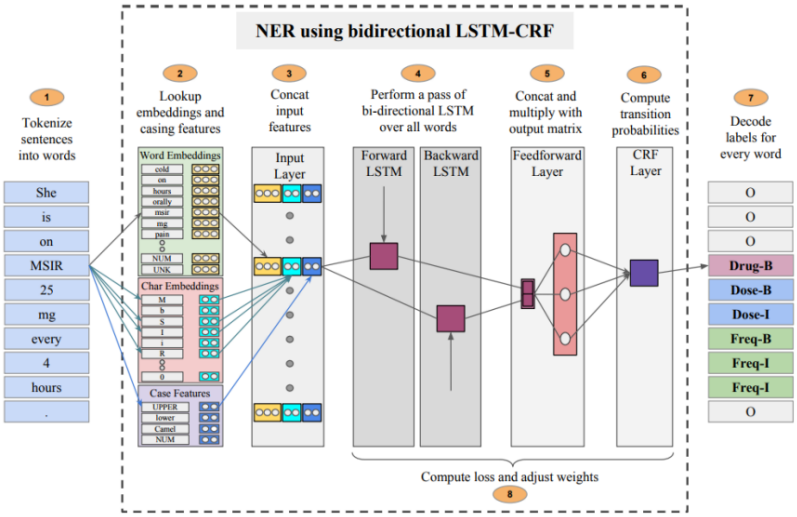
การแสดงผลของคำที่ได้จากนี้จะถูกส่งผ่านชั้น dense layer ที่รับผิดชอบในการเพิ่มความสามารถให้กับการแสดงผลข้อมูลนำเข้าและแปลงรูปแบบ IOB (Inside, Outside, Beginning) ของชุดข้อมูล CONLL-2003 และทำนายนิพจน์ในของแต่ละคำโดยจะแสดงลเฉพาะคำที่มีค่าความมั่นใจ (Confidence score) มากกว่า 0.4 เท่านั้น



ภาพที่ 2.5 ขั้นตอนการนำข้อมูลเข้าโมเดลของงานวิจัยที่เกี่ยวข้อง 1

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

2.7.2 Clinical NER and Relation Extraction using Bi-Char-LSTMs and Random Forest Classifiers [6]



ภาพที่ 2.5 โครงสร้างการสกัดนิพจน์ทางการแพทย์ของงานวิจัยที่เกี่ยวข้อง 2

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

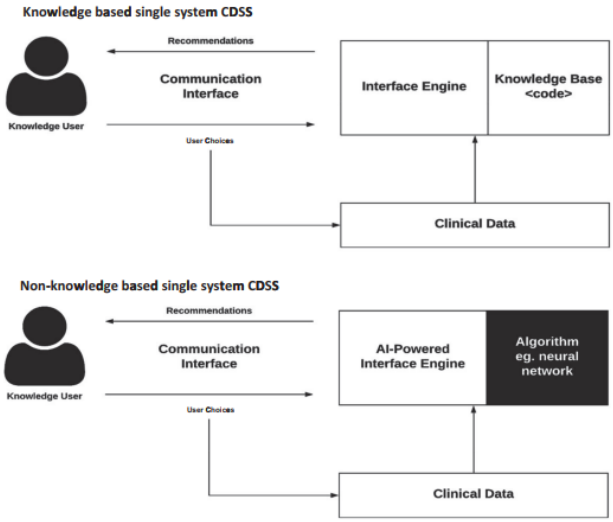
จากโมเดลดังกล่าวได้ประสิทธิภาพในการระบุนิพจน์ทางการแพทย์โดยมีค่า F1-score ของ micro-average อยู่ที่ 0.81

ตารางที่ 2.2 ผลการวัดประสิทธิภาพของงานวิจัยที่ 2

Task	Label	Precision	Recall	F1-score
Named Entity Recognition	Drug	0.87	0.84	0.86
	Dose	0.88	0.83	0.86
	Route	0.89	0.91	0.90
	Frequency	0.82	0.81	0.82
	Duration	0.74	0.82	0.78
	Indication	0.47	0.65	0.55
	Severity	0.80	0.79	0.80
	SSLIF	0.83	0.82	0.82
	ADE	0.38	0.68	0.49
	Micro-Avg	**0.80**	**0.82**	**0.81**
Relationship Extraction*	Dosage	0.88	0.94	0.91
	Manner/Route	0.93	0.97	0.95
	Frequency	0.85	0.96	0.90
	Duration	0.88	0.96	0.92
	Severity Type	0.95	0.98	0.97
	Reason	0.60	0.82	0.70
	Adverse	0.66	0.88	0.76
	Micro-Avg	**0.82**	**0.94**	**0.88**

2.7.3 An overview of clinical decision support systems: benefits, risks, and strategies for success [7]

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



ภาพที่ 2.6 การพัฒนาระบบ CCDS จากงานวิจัยที่ 3

เปรียบเทียบข้อดีข้อเสียของระบบสนับสนุนการตัดสินใจของแพทย์

ตารางที่ 2.3 สรุปข้อดีข้อเสียของระบบ CCDS ในงานวิจัยที่ 3

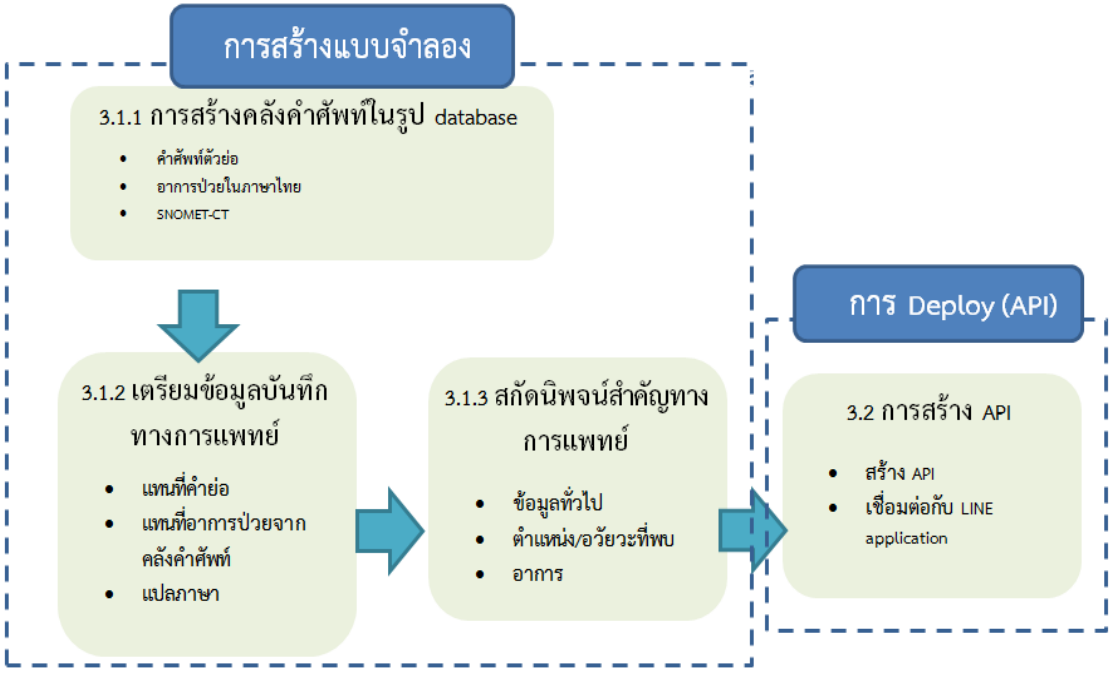
หัวข้อ	ข้อดี	ข้อเสีย
ความปลอดภัยของผู้ป่วย	ลดอันตรายที่เกิดจากความคลาดเคลื่อนที่อาจจะเกิดขึ้น เช่น การจ่ายยาผิด	ได้รับการแจ้งเตือนที่มากเกินไปทำให้ไม่สนใจและเมื่อมีข้อแจ้งที่อันตรายทำให้แพทย์ไม่สนใจ
การจัดการผู้ป่วย	ช่วยวางแผนการรักผู้ป่วยได้ดีขึ้น	ทำให้แพทย์ขาดความเชี่ยวชาญ และเชื่อในระบบมากเกินไป
ค่าใช้จ่าย	ช่วยลดค่าการรักษา เช่นการจ่ายยาหรือการรักษาที่ซ้ำซ้อน	ค่าติดตั้งและค่าบำรุงรักษาค่อนข้างมีราคาสูง
ฟังก์ชันการบริหารจัดการ/ระบบอัตโนมัติ	ช่วยให้เลือกหมายเลข ICD10 และทำเอกสารต่างๆ ได้อัตโนมัติ	ต้องมีการอับเดตระบบอยู่เสมอ
การตัดสินใจในการรักษา	ช่วยให้คำแนะนำเกี่ยวกับการรักษาโดยขึ้นกับข้อมูลของผู้ป่วยอัตโนมัติ	แพทย์อาจไม่เห็นด้วยกับสิ่งที่ระบบแนะนำ หรืออาจจะมี bias ทำให้ยึดตามระบบมากกว่า

ตารางที่ 2.3 (ต่อ)

หัวข้อ	ข้อดี	ข้อเสีย
ระบบเอกสาร	ได้ระบบเอกสารที่ดีขึ้น	ระบบอาจจะรวบรวมข้อมูลจากหลายแหล่งที่มาทำให้ข้อมูลนั้นไม่เป็นแพทเทิร์นเดียวกัน ผู้ใช้งานต้องมาปรับเอกสารใหม่อีกครั้ง

บทที่ 3
ระเบียบวิธีวิจัย

การศึกษาวิจัยครั้งนี้เป็นการนำเสนอระบบสกัดนิพจน์ทางการแพทย์สำหรับระบบสนับสนุนการตัดสินใจทางการแพทย์ โดยทำการสกัดนิพจน์จากบันทึกทางการแพทย์ของไทยซึ่งมีทั้งคำภาษาไทย, คำศัพท์ย่อโดยมีแนวทางการวิจัยดังนี้



ภาพที่ 3.1 แผนผังการทำงานของงานวิจัย

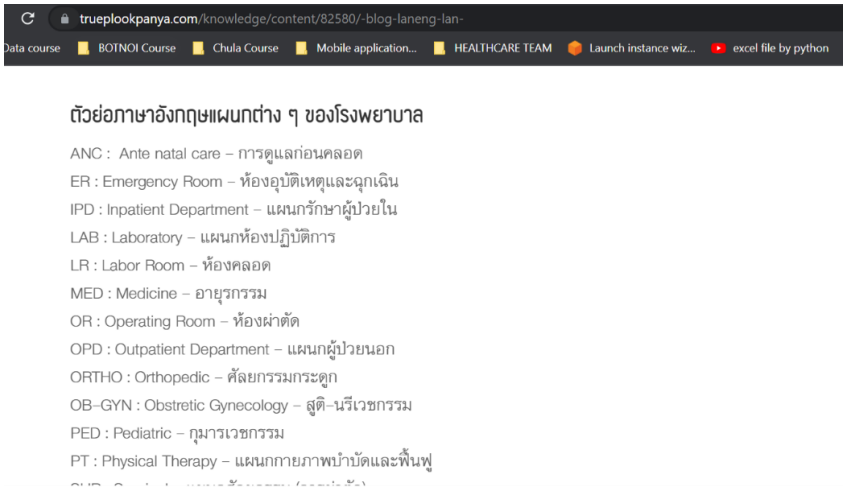
3.1 การสร้างแบบจำลอง

3.1.1 การสร้างคลังคำศัพท์ในรูปฐานข้อมูล

(1) ฐานข้อมูลตัวอักษรย่อทางการแพทย์

การสร้างฐานข้อมูลตัวอักษรย่อทางการแพทย์ถูกพัฒนาจากการนำข้อมูลจากเว็บไซต์ที่มีคำศัพท์ย่อและความหมายเต็มของคำศัพท์ย่อนั้นโดยดึงข้อมูลจากเว็บไซต์ (Web scraping) โดยแบ่งเป็นคำศัพท์ที่ใช้โดยทั่วไปในบันทึกทางการแพทย์, ตำแหน่งของอวัยวะ, คำที่ใช้ในการวินิจฉัยโรค เป็นต้น โดยดึงข้อมูลจากเว็บไซต์โดยใช้ชุดเครื่องมือ Selenium และนำข้อมูลมากเว็บไซต์ต่างๆ ดังนี้

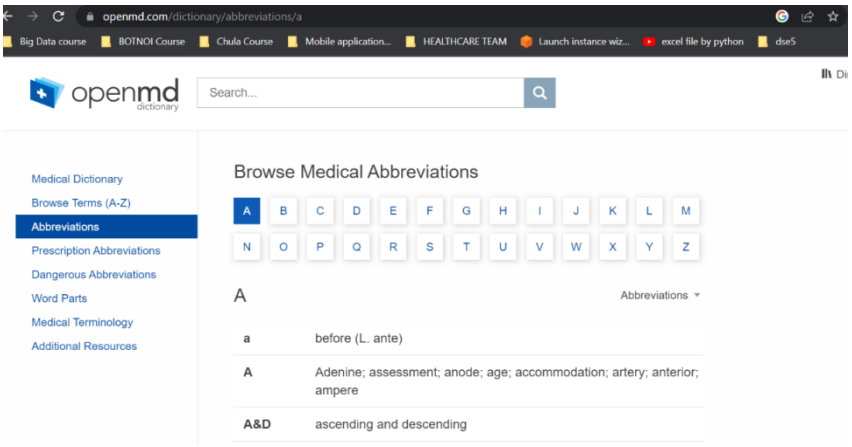
- เว็บไซต์ทรูปลูกปัญญา



ภาพที่ 3.2 ตัวอย่างตัวย่อภาษาภาษาอังกฤษจากเว็บไซต์ทรูปลูกปัญญา

ที่มา: (https://www.trueplookpanya.com/knowledge/content/82580/-blog-laneng-lan)

- เว็บไซต์ openmd



ภาพที่ 3.3 ตัวอย่างตัวย่อภาษาภาษาอังกฤษจากเว็บไซต์ Openmd

ที่มา: (https://openmd.com/dictionary/abbreviations)

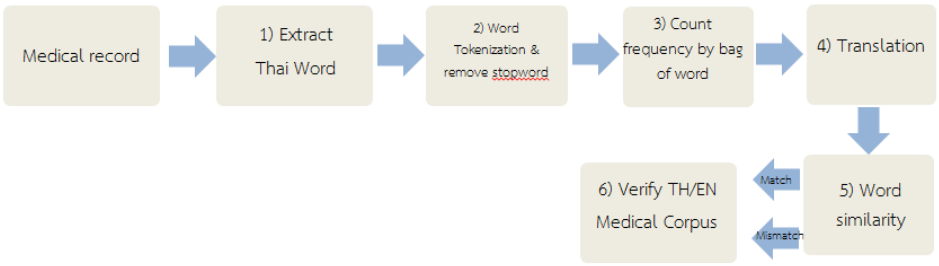
เมื่อได้ฐานข้อมูลตัวอักษรย่อทางการแพทย์ทั้งหมด ทำการตรวจสอบความถูกต้องของข้อมูล หากพบตัวอักษรย่อตัวไหนซ้ำให้ทำการจัดกลุ่มของคำดังกล่าวตามแผนกของการรักษาของคำนั้น

(2) ฐานข้อมูล SNOMET-CT

เพื่อให้การสกัดนิพจน์สำคัญนั้นได้คำศัพท์ที่เป็นมาตรฐานจึงมีการนำฐานข้อมูล SNOMED-CT มาใช้ โดยข้อมูลดังกล่าวอยู่ในรูปแบบฐานข้อมูล SQL โดยทำการติดตั้งและเรียกดูข้อมูลโดยผ่านโปรแกรม MySQL และนำคำศัพท์มาใช้ในเฉพาะหัวข้อ finding และ body of structure

(3) ฐานข้อมูลอาการป่วยภาษาไทย

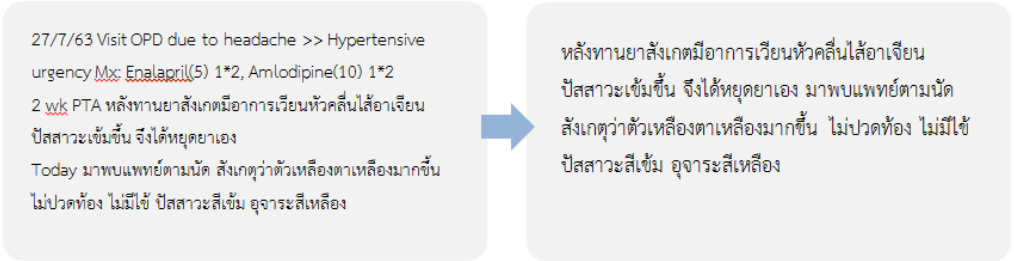
เนื่องจากในปัจจุบันยังมีข้อมูลฐานข้อมูลอาการป่วยภาษาไทยพร้อมคำแปลที่เป็นศัพท์เฉพาะทางการแพทย์ที่ไม่เพียงพอ จึงมีการพัฒนาฐานข้อมูลคลังคำศัพท์ข้อมูลอาการป่วยภาษาไทยขึ้นโดยสร้างขึ้นจากการค้นหาคำที่พบบ่อยจากข้อมูลบันทึกทางการแพทย์ของแผนกศัลยกรรมจำนวน 200,000 ข้อมูล มีขั้นตอนการทำดังต่อไปนี้



ภาพที่ 3.4 ขั้นตอนการสกัดอาการสำคัญ

(4) การสกัดข้อความภาษาไทย (Regex)

ข้อความบันทึกทางการแพทย์เป็นบันทึกที่มีการใช้ทั้งภาษาไทยและภาษาอังกฤษ เพื่อสร้างการสกัดข้อความภาษาไทยนำมาสร้างเป็นฐานข้อมูลคลังคำศัพท์โดยใช้ Regex



ภาพที่ 3.5 ขั้นตอนการสกัดอาการภาษาไทย

(5) การตัดคำ (Word tokenization)

นำข้อมูลทางการแพทย์ทั้งหมดนำมาตัดคำเพื่อนำไปใช้สร้างเป็นคำศัพท์โดยใช้เครื่องมือ pythainlp และตัดคำตามการเว้นวรรค โดยมีการรายละเอียดการเลือกใช้เครื่องมือดังนี้

ตารางที่ 3.1 parameter ที่ใช้ในการตัดคำ

เครื่องมือที่ใช้ตัดคำ	pythainlp, วรรค
การลบ stop word	thai_stopwords
อัลกอริทึมการตัดคำ	multi_cut, dictionary

(6) การนับความถี่ที่พบโดยใช้ Bag of word

นำข้อมูลที่ได้จากข้อ 2 นำมาสร้างเป็นชุดข้อมูล bag of word เพื่อนับความถี่ที่พบทั้งหมดในเอกสารทั้งหมด 200,000 ข้อมูล โดยใช้ Countvectorizer

feat	กกหู	กด	กรด	กรน	กระ	กระดก	กระดิก	กระดูก	กระดูกคอ	กระดูกสันหลัง	กระตุก	กระบอกตา	กระยะ	กระพริบตา	กระวนกระวาย	กระวนกระวายใจ	กระสับกระส่าย	กระเดิน	กระเบนเหน็บ
เกร็ง	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
เกล็ดเลือดต่ำ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
เกาเป็นแผล	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
เกาจนเป็นแผล	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
เกิดเหตุที่โรงงาน	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
อุจจาระสีเหลือง	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
อุจจาระสีเหลืองดี	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

ภาพที่ 3.6 ตัวอย่างทำ Bag of word

	feat	freq
19553	ปวดหลังส่วนล่าง	727
30015	ร้าวลงขาซ้าย	399
17704	ปวดเข่าสองข้าง	248
41628	ปวดเข่าทั้งสองข้าง	245
30033	ร้าวลงขาทั้งสองข้าง	206
30009	ร้าวลงขาข้างซ้าย	199
30331	ร้าวลงแขนซ้าย	199
16998	ปวดหลังลดลง	180
47740	ไม่มีความเสี่ยงต่อการหกล้ม	177
19525	ปวดหลัง	170
16883	ปวดหลังร้าวลงขาขวา	163
5442	ชาลงขาสองข้าง	154
16525	ปวดหลังช่วงบั้นเอวร้าวลงขาทั้ง	154
30316	ร้าวลงแขนขวา	149
5358	ชาร้าวลงขา	148
4742	ชาขาซ้าย	146
47794	ไม่มีชาร้าวลงขา	141

ภาพที่ 3.7 ตัวอย่างการนับความถี่ของคำที่พบ

(7) การแปลภาษา

นำข้อความจาก bag of word เข้าโมเดลแปลภาษาไทยเป็นภาษาอังกฤษ โดยใช้เครื่องมือ Google translate API เพื่อนำไปเทียบกับฐานข้อมูล Snomed-CT ในข้อ 5

(8) การหาความคล้ายคลึงของคำ (Word similarity)

เมื่อได้คู่ศัพท์ภาษาไทยและภาษาอังกฤษจากข้อ 4 นำข้อความทั้งหมดที่ได้จากการทำ Bag of word เปรียบเทียบความคล้ายคลึงกันกับศัพท์ในฐานข้อมูล Snomed-CT โดยการหา word similarity โดยใช้เครื่องมือ spaCy โดยเลือกโมเดลที่สร้างจากชุดข้อมูล ‘en_core_web_md’ ซึ่งพัฒนาจากโมเดลจากข้อมูลทั่วไปจากอินเตอร์เน็ทที่มาจากหลากหลายแหล่งที่มา เช่น เว็บไซต์ บทความข่าว เป็นต้น และเลือกข้อความที่มี word similarity สูงที่สุดในแต่ละคำ

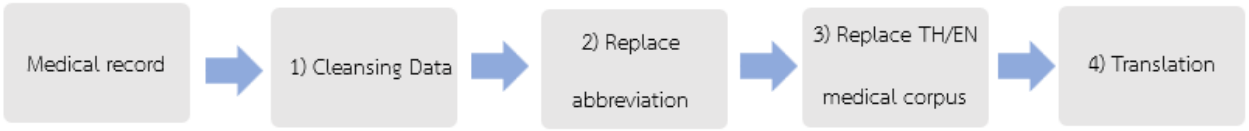
(9) ตรวจสอบความถูกต้องของข้อมูล (TH/EN Medical corpus verification)

เรียงลำดับข้อมูลโดยเรียงตามความถี่ที่พบและค่าความคล้ายคลึง (word similarity score) เพื่อเลือกคำศัพท์นำมาตรวจสอบความถูกต้องของข้อมูลและนำไปใช้ในการสร้างฐานข้อมูลอาการป่วยสำหรับการเตรียมข้อมูลบันทึกทางการแพทย์ในข้อม 3.1.2

feat	Google Translate	SNOMED-CT	Similarity score	freq
ปวดหลังส่วนล่าง	lower back pain	low back pain	1.0	727
ร้าวลงขาซ้าย	crack down left leg	pain radiating to left leg	0.6268353184	399
ปวดเข่า	knee pain	knee pain	1.0	248
ไม่มีความเสี่ยงต่อการหกล้ม	No risk of fall	at very low risk fall	0.5314480862	177
ปวดหลัง	backache	backache	1,0	170
ชาลงขาสองข้าง	numb of both legs	numbness of lower limb	0.8425106402	154
ปวดไหล่ซ้าย	left shoulder pain	pain of left shoulder joint	0.7198681139	60
ปวดสะบักขวา	right shoulder blade pain	pain of right shoulder blade	0.7717179248	134
ปวดเข่าซ้าย	left knee pain	pain in left knee	0.793310498	140

ภาพที่ 3.8 ตัวอย่างการเปลี่ยนเทียบคู่ศัพท์ที่แปลโดย google translate API และ SNOMED-CT

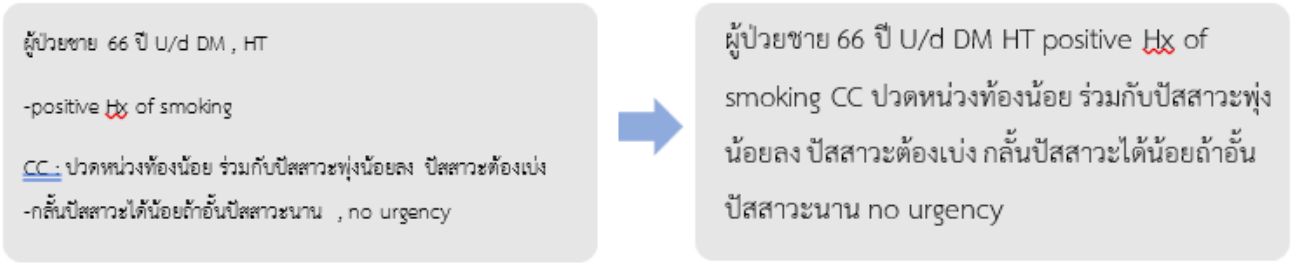
3.1.2 การเตรียมข้อมูลบันทึกทางการแพทย์ (Data Preprocessing)



ภาพที่ 3.9 ขั้นตอนการเตรียมบันทึกทางการแพทย์

(1) การทำความสะอาดข้อมูล (Cleansing Data)

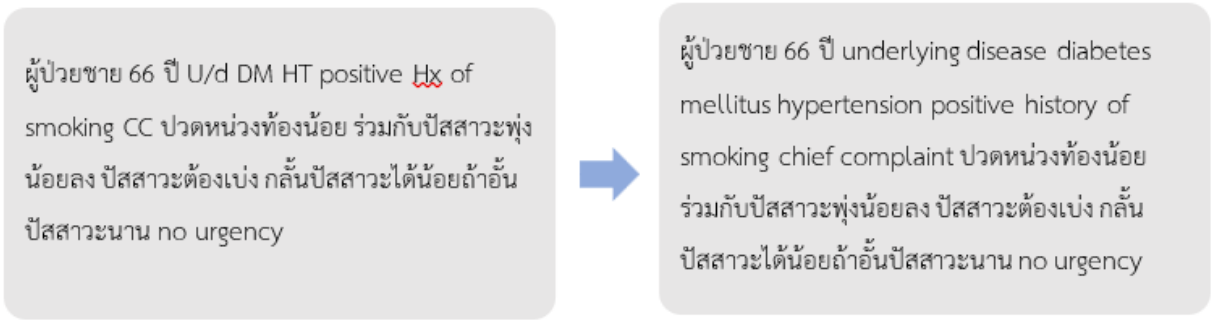
นำสัญลักษณ์พิเศษต่างๆ ออกจากข้อมูลบันทึกทางการแพทย์ เช่น ‘<’, ‘>’, ‘-’, ‘\’, ‘\n’, ‘?’, ‘:’ เป็นต้น เพื่อเตรียมข้อมูลสำหรับขั้นตอนถัดไป



ภาพที่ 3.10 ขั้นตอนการทำความสะอาดข้อมูล

(2) การแทนที่ตัวอักษรย่อทางการแพทย์ (Replace abbreviation)

นำข้อมูลจากฐานข้อมูลทางการแพทย์มาตรวจสอบกับข้อมูลบันทึกทางการแพทย์โดยตรงพบตามการเว้นวรรค (space) หากตรงตามคลังคำศัพท์ทางการแพทย์ที่สร้างไว้ให้แทนทีด้วยความหมายที่กำหนด



ภาพที่ 3.11 การแทนที่ตัวอักษรย่อทางการแพทย์

(3) แทนที่อาการป่วยจากคลังคำศัพท์ทางการแพทย์ (Replace TH/EN medical corpus)

นำข้อความที่ได้จากข้อที่ 2 แทนที่ด้วยฐานข้อมูลคลังคำศัพท์ทางการแพทย์ด้วยการเรียงคำศัพท์ในที่มีความยาวของตัวอักษรสูงสุดและเลื่อนตรวจเช็คข้อความ window slicing) ครั้งละ 1 ตัวอักษร หากตรวจพบข้อความที่ตรงตามข้อมูลจากคลังคำศัพท์ทางการแพทย์จะแทนที่ข้อความดังกล่าวด้วยคู่คำศัพท์นั้น

ภาพที่ 3.12 การแทนที่อาการป่วยจากคลังคำศัพท์ทางการแพทย์

3.1.3 สกัดนิพจน์สำคัญทางการแพทย์

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

ภาพที่ 3.13 การสกัดนิพจน์สำคัญทางการแพทย์

3.2 การนำไปใช้งาน

แนวคิดของงานวิจัยมีจุดประสงค์ที่จะนำระบบสกัดนิพจน์สำคัญทางการแพทย์ไปเชื่อมต่อกับระบบสารสนเทศของโรงพยาบาล (Hospital information system, HIS) โดยผ่านการเชื่อมต่อ API แต่เนื่องจากมีข้อจำกัดในการเชื่อมต่อในหลายประการ ดังนั้นจึงเลือกการแสดงผลผ่านแชทบาทบนช่องทาง Line application ในการแสดงผลของระบบนิพจน์สำคัญทางการแพทย์ เนื่องจากสามารถนำไปสร้างเป็นเป็นผลิตภัณฑ์ต้นแบบ พร้อมทั้งสามารถเก็บข้อมูลต่างๆ จากแพทย์ผู้ทดลองใช้ระบบได้ง่าย การนำระบบสกัดนิพจน์สำคัญทางการแพทย์มีการนำไปแสดงผลบน 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

3.3 เครื่องมือที่ใช้ในงานวิจัย

3.3.1 ภาษาไพธอน (Python)

ภาษาไพธอน (Python) เป็นภาษาโปรแกรมที่ได้รับความนิยมอย่างแพร่หลายในวงกว้าง เนื่องจากมีความอ่านง่ายและเขียนโค้ดได้ง่าย ภาษาไพธอนถูกสร้างขึ้นโดย Guido van Rossum ครั้งแรกในปี 1991 และต่อมาได้รับการพัฒนาและเป็นที่ยอมรับในชุดของนักพัฒนาซอฟต์แวร์ทั่วโลก ภาษาไพธอนเป็นภาษาโปรแกรมระดับสูง (high-level programming language) ที่สนับสนุนการโปรแกรมแบบวัตถุ (object-oriented programming) และสามารถใช้งานในหลายรูปแบบได้ เช่น การพัฒนาเว็บไซต์ (web development) การวิเคราะห์ข้อมูล (data analysis) การพัฒนาแอปพลิเคชันเดสก์ท็อป (desktop application development) และอื่นๆ

ภาษาไพธอนมีลักษณะที่อ่านง่าย สามารถเขียนโค้ดได้อย่างสั้นและกระชับ มีสัญลักษณ์เครื่องหมายที่ช่วยให้โค้ดอ่านง่าย เช่นการเว้นวรรค (indentation) แทนการใช้เครื่องหมายปีกกา ภาษาไพธอนยังมีไลบรารี (library) มากมายที่สามารถนำมาใช้เพื่อให้งานเป็นไปได้ง่ายและรวดเร็วมากขึ้น

3.3.2 Google Colab

Google Colab (ชื่อเต็มคือ Google Colaboratory) เป็นแพลตฟอร์มสำหรับการเขียนและรันโค้ด Python ออนไลน์แบบโฮสต์โดย Google ซึ่งให้บริการในรูปแบบของสมุดบันทึก (notebook) ที่เรียกว่า Colab Notebook หรือ Colab ให้คุณสามารถเขียนและรันโค้ด Python ได้โดยตรงในเบราว์เซอร์โดยไม่ต้องติดตั้งโปรแกรมใดๆ บนเครื่องของคุณ

3.3.3 MySQL

MySQL เป็นระบบฐานข้อมูลที่เป็นที่นิยมอย่างแพร่หลายใช้ในการจัดเก็บและจัดการข้อมูลในรูปแบบตาราง (table) และเขียนด้วยภาษาสอบถามฐานข้อมูล SQL (Structured Query Language) ซึ่งเป็นภาษามาตรฐานสำหรับการจัดการฐานข้อมูล โดย MySQL มีความเสถียรและมีประสิทธิภาพสูง ทำให้เป็นที่นิยมในการพัฒนาและดำเนินการกับฐานข้อมูลในการพัฒนาเว็บไซต์และแอปพลิเคชันต่างๆ

บทที่ 4
ผลการวิจัย

จากการพัฒนาโมเดลการสกัดนิพจน์สำคัญทางการแพทย์สำหรับข้อมูลภาษาไทยเพื่อช่วยสนับสนุนการตัดสินใจทางการแพทย์ โดยการสร้างฐานข้อมูลคลังคำศัพท์ตัวอักษรย่อและคำศัพท์ภาษาไทย/อังกฤษทางการแพทย์มาประยุกต์ในการสกัดนิพจน์สำคัญทางการแพทย์ผ่านโมเดลจาก Bio-Epidemiology-NER ซึ่งมีรายละเอียดดังนี้

4.1 ผลการเตรียมฐานข้อมูลทางการแพทย์

4.1.1 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 4.1 ฐานข้อมูลคลังคำศัพท์ย่อทางการแพทย์

4.1.2 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 4.2 ฐานข้อมูล SNOMED-CT ในกลุ่มของ finding

term_x	concep	typeId_x	term_y	typeId	eptabili	type
Undescended testis (body structure)	1.28E+08	Fully specified name	Undescended testis	Synonym	Preferred	body structure
Entire stylomastoid foramen (body structure)	1.34E+08	Fully specified name	Entire stylomastoid foramen	Synonym	Preferred	body structure
Entire occipitomastoid suture of skull (body structure)	1.34E+08	Fully specified name	Entire occipitomastoid suture of skull	Synonym	Preferred	body structure
Muscle belly (body structure)	1.34E+08	Fully specified name	Muscle belly	Synonym	Preferred	body structure
Trigger point (body structure)	1.34E+08	Fully specified name	Trigger point	Synonym	Preferred	body structure
Trigger point (body structure)	1.34E+08	Fully specified name	TP - trigger point	Synonym	Acceptabl	body structure
Respiratory structure (body structure)	1.34E+08	Fully specified name	Respiratory structure	Synonym	Preferred	body structure
Nodes of Kent (body structure)	1.34E+08	Fully specified name	Nodes of Kent	Synonym	Preferred	body structure
T6 spinous process (body structure)	1.34E+08	Fully specified name	T6 spinous process	Synonym	Preferred	body structure
Bilateral adrenal glands (body structure)	1.34E+08	Fully specified name	Bilateral adrenal glands	Synonym	Preferred	body structure
Bilateral adrenal glands (body structure)	1.34E+08	Fully specified name	Both adrenal glands	Synonym	Acceptabl	body structure
Histological tissue (body structure)	1.34E+08	Fully specified name	Histological tissue	Synonym	Preferred	body structure
Gland (body structure)	1.34E+08	Fully specified name	Gland	Synonym	Preferred	body structure
Bone tissue (body structure)	1.34E+08	Fully specified name	Bone tissue	Synonym	Preferred	body structure
Entire squamomastoid suture of skull (body structure)	1.35E+08	Fully specified name	Entire squamomastoid suture of skull	Synonym	Preferred	body structure
Entire squamous suture of skull (body structure)	1.36E+08	Fully specified name	Entire squamous suture of skull	Synonym	Preferred	body structure
Entire sphenofrontal suture of skull (body structure)	1.36E+08	Fully specified name	Entire sphenofrontal suture of skull	Synonym	Preferred	body structure
Entire sphenosquamous suture of skull (body structure)	1.36E+08	Fully specified name	Entire sphenosquamous suture of sku	Synonym	Preferred	body structure
Entire sphenoparietal suture of skull (body structure)	1.37E+08	Fully specified name	Entire sphenoparietal suture of skull	Synonym	Preferred	body structure
Entire sphenoid suture of skull (body structure)	1.37E+08	Fully specified name	Entire sphenoid suture of skull	Synonym	Preferred	body structure

ภาพที่ 4.3 ฐานข้อมูล SNOMED-CT ในกลุ่มของ body structure

4.1.3 ฐานข้อมูลคลังคำศัพท์อาการป่วยทางการแพทย์ภาษาไทย

ผลลัพธ์การเตรียมฐานข้อมูลคลังคำศัพท์อาการป่วยทางการแพทย์ภาษาไทยได้ถูกสร้างตามกระบวนการตัดคำ เทียบความคล้ายคลึงกับฐานข้อมูลใน SNOMED-CT และตรวจสอบความถูกต้องของข้อมูล ได้ชุดข้อมูลทั้งหมด 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
กล้ามเนื้อแขนและขาซีกซ้ายอ่อนแรง	muscle weakness of limb
กล้ามเนื้อใบหน้าขวาอ่อนแรง	weakness of right facial muscle
กล้ามเนื้อใบหน้าข้างขวาอ่อนแรง	weakness of right facial muscle
กล้ามเนื้อใบหน้าข้างซ้ายอ่อนแรง	weakness of left facial muscle
กล้ามเนื้อหน้าข้างขวาอ่อนแรง	hemiparesis
กล้ามเนื้อหน้าข้างซ้ายอ่อนแรง	weakness of left facial muscle
กล้ามเนื้อหน้าด้านซ้ายอ่อนแรง	unilateral facial paresis
กล้ามเนื้อหลังตึง	stiff back
กล้ามเนื้ออ่อนแรงครึ่งซีกซ้าย	hemiplegia
กล้ามเนื้ออ่อนแรงซีกขวา	right hemiparesis
กลืนได้	able to swallow
กลืนติด	difficulty swallowing
กลืนไม่ได้	unable to swallow
กลืนไม่ลง	dysphagia

ภาพที่ 4.4 ฐานข้อมูลคลังคำศัพท์อาการป่วยทางการแพทย์ภาษาไทย

4.2 ผลการวัดประสิทธิภาพความถูกต้องของโมเดล

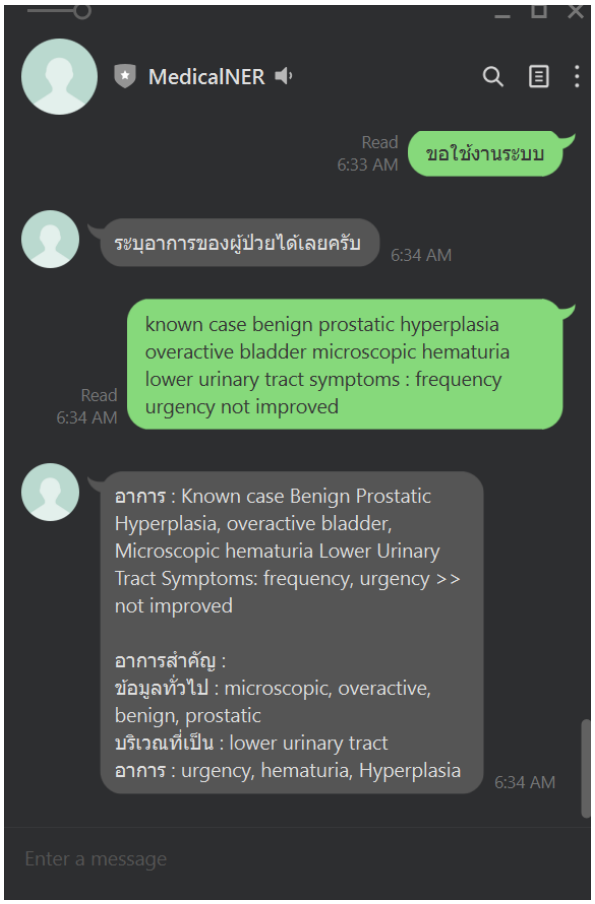
สุ่มข้อมูลบันทึกทางการแพทย์กลุ่มโรคทางเดินปัสสาวะในแผนกศัลยกรรมโดยนำมาใช้ในการทดสอบทั้งหมด 200 ข้อมูล โดยผลของค่า Precision, Recall และ F1-score เมื่อเปรียบเทียบกับการสกัดชุดข้อมูลมาตรฐานโดยคำนวณจากนิพจน์ที่สำคัญมีรายละเอียดดังต่อไปนี้

ตารางที่ 4.1 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

4.3 ผลการใช้งาน

นำ API เชื่อมต่อกับระบบแชทบอทและแสดงผลผ่าน Line Official Account ซึ่งได้ตัวอย่างผลการทำงานดังนี้



ภาพที่ 4.5 การใช้งานผ่าน Line Application

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอเกี่ยวกับกาพัฒนาระบบสกัดนิพจน์สำคัญสำหรับบันทึกทางการแพทย์ของประเทศไทย โดยการนำฐานข้อมูลตัวอักษรย่อ คลังคำศัพท์อาการป่วยภาษาไทย และการสกัดนิพจน์ทางการแพทย์โดยใช้ Bio-Epidemiology-NER ซึ่งระบบนี้จะช่วยแยกนิพจน์ที่สำคัญซึ่งช่วยให้แพทย์สามารถทวนสอบบันทึกทางการแพทย์ได้รวดเร็วขึ้นสามารถสรุปลงานวิจัยได้ดังนี้

5.1 สรุปผลการทดลอง

5.1.1 ได้พัฒนาที่มีประสิทธิภาพนารสกัดนิพจน์สำคัญทางการแพทย์ ประกอบด้วยขั้นตอนดังนี้

(1) ค้นหาเว็บไซต์ที่มีข้อมูลตัวอักษรย่อทางการแพทย์ ใช้การดึงข้อมูลจากเว็บไซด์ (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

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

5.2 ข้อสังเกต

5.2.1 ผลการสกัดนิพจน์สำคัญทางการแพทย์ยังไม่คำศัพท์ไม่ครอบคลุมข้อมูลทั้งหมดในบันทึกทางการแพทย์และข้อมูลที่เป็นส่วนการปฏิเสธ เช่น ไม่มีอาเจียน ไม่มีปัสสาวะแสบขัด เป็นต้น ยังไม่สามารถสกัดข้อความออกมาในรูปปฏิเสธได้

5.2.2 บันทึกทางการแพทย์ที่เป็นข้อมูลรายละเอียดมีค่อนข้างยาว ไม่มีรูปประโยค และมีภาษาไทยปนภาษาอังกฤษ ระบบสกัดนิพจน์สำคัญทางการแพทย์ยังไม่สามารถสกัดข้อความได้แม่นยำ

5.2.3 ตัวอักษรย่อทางการแพทย์ที่ใช้ตัวย่อเดียวกัน ระบบยังไม่ระบุได้ว่าตัวอักษรย่อดังกล่าวหมายถึงคำเต็มรายการไหน

5.3 ข้อเสนอแนะ

5.3.1 วิธีการแปลงข้อมูลโดยใช้ฐานข้อมูลคลังคำศัพท์จำเป็นต้องมีฐานข้อมูลจำนวนมาก หากต้องการเพิ่มความแม่นยำหรือนำไปใช้กับข้อมูลอาการป่วยในโรคอื่นจำเป็นต้องมีการเพิ่มฐานข้อมูลจึงยังไม่เหมาะสมกับการนำขยายส่วน (scale-up)

5.3.2 สร้างเว็บแอพพลิเคชันเพื่อให้กำหนดสีของนิพจน์แต่ละประเภทได้ ทำให้มีประสบการณ์ใช้งานที่ดีกว่าการเชื่อมต่อ API กับระบบสารสนเทศของโรคพยาบาล

5.3.3 นำระบบสกัดนิพจน์ดังกล่าวนำไปใช้ต่อยอดในขั้นตอนการสกัด feature เพื่อนำข้อมูลไปใช้ในการทำนายโรคตามรหัสโรค ICD-10 หรือใช้ทำวิเคราะห์ข้อมูลสถิติเกี่ยวกับบันทึกการรักษาผู้ป่วย

บรรณานุกรม

บรรณานุกรม

[1] อัตราส่วน ‘บุคลากรทางการแพทย์’ ไทย มีพอไหม ไหวหรือเปล่า https://www.thecoverage.info/news/content/2807

[2] เกี่ยวกับ SNOMED-CT https://www.this.or.th/about-snomed-ct/

[3] Aroonmanakun W, Nupairoj N, Muangsin V, Choemprayong S. Thai monitor corpus: Challenges and contribution to thai nlp. Vacana. 2018 Jul 17;6(2):1-4.

[4] Tapsai C, Meesad P, Unger H. An Overview on the development of Thai natural language processing. Information Technology Journal. 2019 Dec 28;15(2):45-52.

[5] 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.

[6] 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.

[7] 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.

ประวัติผู้เขียน

ชื่อ – นามสกุล ศศลักษณ์ อุบลวิโรจน์

ประวัติการศึกษา

พ.ศ. 2557 - ปริญญาตรี เภสัชศาสตรบัณฑิต สาขาเทคโนโลยีเภสัชภัณฑ์ มหาวิทยาลัยศรีนครินทรวิโรฒ

ประสบการณ์ทำงาน

พ.ศ. 2564 - Project manager, บริษัทไอบอทน้อย จำกัด

พ.ศ. 2563 - QA Pharmacist, บริษัทเจริญเภสัชแลป

พ.ศ. 2558 - QA Technical Pharmacist, บริษัทโอลิคจำกัด