การระบุอาการและอาการแสดงจากข้อความบอกเล่าอาการสำคัญภาษาไทยตามมาตรฐาน ICD-10 โดยขั้นตอนการประมวลผลภาษาธรรมชาติ

ภวินท์ แซ่คู[1], จารุณี ดวงสุวรรณ[2], และ ลัดดา ปรีชาวีรกุล[3]
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ 15 ถ.กาญจนวนิช อ.หาดใหญ่ จ.สงขลา 90110
E-mail: tamakosan14@gmail.com[1], jarunee.d@psu.ac.th[2], ladda.p@psu.ac.th [3]

บทคัดย่อ

บทความนี้นำเสนอกระบวนเพื่อระบุอาการและอาการแสดงจากการสกัดข้อความบอกเล่าอาการผู้ป่วยด้วยกระบวนการประมวลผลภาษาธรรมชาติ สำหรับข้อความบอกเล่าอาการผู้ป่วยหรือ Chief Complains (CCs) ถูกบันทึกอยู่ในรูปแบบภาษาธรรมชาติด้วยภาษาไทย ดังนั้นการที่จะนำสกัดข้อความดังกล่าว และแปลงให้อยู่ให้อยู่ในรูปแบบที่สามารถนำมาใช้ในประมวลผลทางคอมพิวเตอร์ได้มีความยุ่งยากและซับซ้อน เพราะนอกจากจะต้องเข้าใจถึงลักษณะ ข้อจำกัด วิธีการจัดการกับข้อจำกัดของภาษาไทย แล้วยังต้องกำหดนกระบวนการและขั้นตอนสำหรับสกัดอาการและอาการแสดงให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถนำไปประมวลผลได้เช่นกัน

คำสำคัญ: การประมวลผลภาษาธรรมชาติ, บัญชีจำแนกทางสถิติระหว่างประเทศของโรคและปัญหาสุขภาพที่เกี่ยวข้อง, ข้อความบอกเล่าอาการสำคัญ

Abstract

This article describes a signs and symptoms extraction from Chief Complains (CCs) in Thai language. CCs describes about what an illness leading the patient to the hospital. There are useful information contained in CCs, but due the barrier in Thai natural language, to collective information from Thai CCs and preparing to use for computation, there need to understand and overcome the circumstance in Thai language and method to extract signs and symptoms to the form ready to computation has been talk in this article.

Keywords: natural language processing, International Statistical Classification of Diseases and Related Health Problems(ICD), Chief complaints

1. คำนำ

ข้อความบอกเล่าอาการสำคัญ หรือ Chief Complaints (CCs) เป็นข้อความที่อธิบายถึงอาการสำคัญที่นำผู้ป่วยมายังโรงพยาบาลเพื่อทำการรักษา โดยแพทย์จะเป็นผู้บันทึกข้อมูล CCs ของผู้ป่วย โดยในข้อมูล CCs จะไม่มีการบันทึกอาการสำคัญในรูปแบบของชื่อโรค แต่จะบันทึกเพียงคำอธิบายที่บ่งบอกอาการที่สำคัญที่ได้จากการพูดคุย สอบถามระหว่างแพทย์และผู้ป่วย สำหรับ CCs นั้นเป็นส่วนประกอบหนึ่งของเวชระเบียน (medical record) ซึ่งโดยทั่วไปข้อมูล CCs ดังกล่าวจะถูกบันทึกอยู่ในรูปภาษาธรรมชาติทั้งในรูปแบบของภาษาพูด หรือภาษาเขียน ซึ่งส่งผลให้การที่จะนำข้อมูล CCsไปใช้นั้นเป็นเรื่องยากและจำเป็นต้องอาศัยผู้เชี่ยวชาญในการจำแนกข้อมูลCCs เพื่อทำให้ได้ข้อมูลที่ถูกจำแนกออกมามีคุณภาพ และสามารถนำไปใช้ในระบบทางการแพทย์ที่เกี่ยวข้องได้อย่างมีประสิทธิภาพ รวมถึงสามารถนำข้อมูลที่ได้ไปใช้ในการประมวลผลในระบบคอมพิวเตอร์ได้เช่น ระบบการเก็บข้อมูลทางการแพทย์อิเล็กทรอนิกส์ มีชื่อเรียกว่าเวชระเบียนอิเล็กทรอนิกส์(Electronic health record : EHR หรือ Electronic medical record : EMR) ซึ่งเป็นระบบของการสนับสนุนให้เกิดการใช้ข้อมูลร่วมกันระหว่างโรงพยาบาลต่างๆ เช่นข้อมูลของโรค อาการ และ อาการแสดงใน EHR โดยข้อมูลเหล่านี้จะถูกบันทึกในรูปแบบรหัสของ International Statistical Classification of Diseases and Related Health Problems(ICD) ซึ่งปัจจุบันเป็นฉบับที่10(ICD-10) ทำให้สามารถนำข้อมูลดังกล่าวไปใช้ประมวลผลทางคอมพิวเตอร์ได้โดยไม่เกิดปัญหาจากผลกระทบทางภาษา

2. งานวิจัยและทฤษฎีที่เกี่ยวข้อง

ที่ผ่านมางานวิจัย [1] ได้มีการใช้ข้อมูล ICD-9 จากข้อมูลอิเล็กทรอนิกของ 189 โรงพยาบาลในไต้หวันเพื่อการจำแนกประเภทของ CCs และในงานวิจัย [2] ได้เสนอการจำแนกการเฝ้าระวังการระบาดของโรคที่สนับสนุนหลายๆภาษาโดยใช้กรณีศึกษาเป็นภาษาจีน

เมื่อกล่าวถึงคุณลักษณะของภาษาไทยในงานวิจัย [3] ได้อธิบายถึงคุณลักษณะที่สำคัญของภาษาไทยไว้ว่า เป็นภาษาที่มีรูปแบบโครงสร้างในลักษณะของ ประธาน-กริยา-กรรม หรือ Subject-Verb-Object (SVO) ซึ่งมีลักษณะทั่วไปเช่นเดียวกับภาษาอังกฤษ จากคุณลักษณะในงานวิจัยดังกล่าว เมื่อพิจารณาถึงปัญหาในการสื่อถึงปัญหาของการบันทึกข้อมูล CCs นั้นคือการที่แพทย์สามาถบันทึก ข้อมูล อาการหรืออาการแสดง ในภาษาไทย โดยการที่คำหนึ่งคำสามารถแทนได้หลายความหมาย หรือสามารถใช้คำหลายคำแทนความหมายเดียวกันได้ เช่น ปวดหัว เวียนหัว ปวดศีรษะ เวียนศีรษะ ซึ่งคำทั้งสี่คำนี้เป็นการสื่อถึงอาการเดียวกัน หรืออีกตัวอย่างหนึ่งเช่น เจ็บไหล่ ปวดไหล่ โดยวลีเหล่านี้มีรูปแบบโครงสร้างที่เหมือนกันคือ Verb-Object(VO) หรือ Object-Verb(OV) แต่คำที่นำมาใช้ประกอบมีความแตกต่างกัน อย่างไรก็ตามข้อความที่เกิดขึ้นจริงในภาษาไทยอาจมีองค์ประกอบของวลีที่สื่อถึงความหมายเดียวกันและอาจมีรูปแบบอื่นๆซึ่งมีความหมายเดียวกันได้ การแก้ปัญหาถัดมาคือการแก้ปัญหาการตัดคำเนื่องจากภาษาไทยไม่มีสัญลักษณ์ที่ใช้กำหนดขอบเขตคำ ในการแก้ไขปัญหาดังกล่าวได้เสนอแนวทางในงานวิจัย [4] โดยเสนอวิธีการตัดพยางค์เพื่อหาขอบเขตหน้าและขอบเขตหลังของประโยคโดยอาศัยกฎที่สร้างขึ้นตามคุณลักษณะของอักขระภาษาไทย สำหรับงานวิจัยงาน [5] และ [6] ได้นำเสนอวิธีการตัดคำด้วยพจนานุกรม โดยใช้วิธีที่เรียกว่าLongest matching(LM) ซึ่งเป็นวิธีที่จะเลือกตัดคำจากคำที่ยาวที่สุดที่พบในพจนานุกรม ซึ่งหากคำที่ยาวที่สุดที่ถูกเลือก เป็นคำที่ไม่มีอยู่ในพจนานุกรม ระบบก็จะทำการเลือกคำใหม่ซึ่งยาวรองลงมา โดย Maximum Matching(MM) จะทำ LM ก่อนจากนั้นจะกลับไปทำ LM ซ้ำอีกครั้งในแต่ละคำที่ถูกตัด

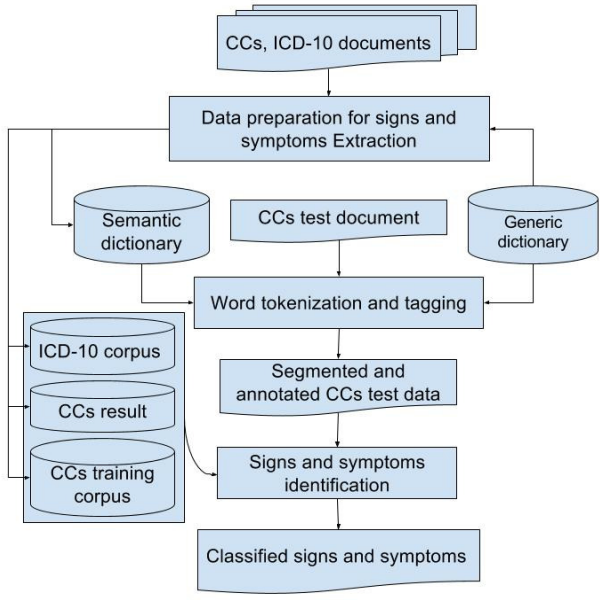
คลังข้อความคือเอกสารอิเล็กทรอนิกส์หรือไฟล์ที่บรรจุข้อความหรือคำในรูปแบบข้อมูลเชิงโครงสร้างโดยมีโครงสร้างแตกต่างกันไปตามบริบทของงานที่นำคลังข้อความดังกล่าวไปใช้ การติดป้าย(tagging) เป็นการกำกับคำศัพท์ลงไปในคลังข้อความซึ่งเป็นวิธีการหนึ่งที่ช่วยในการอธิบายนัยสำคัญของคำนั้นๆต่อบริบทของเรื่องที่สนใจได้อย่างมีประสิทธิภาพ เนื่องจากคำบางคำในบริบทใดบริบทหนึ่งอาจมีความหมายหรือหน้าที่ต่างไปจากการใช้คำดังกล่าวโดยบริบททั่วไป การติดป้ายเพื่ออธิบายคำศัพท์ตามบริบทของเรื่องที่สนใจจึงมีความจำเป็นในการอธิบายถึงความหมายของคำศัพท์ต่อบริบทที่สนใจ

การสกัดข้อมูลสารสนเทศคือการสกัดข้อมูลเชิงโครงสร้างจากข้อมูลที่ไม่เป็นโครงสร้างหรือข้อมูลกึ่งโครงสร้างที่เครื่องคอมพิวเตอร์สามารถเข้าใจได้ ในงานวิจัย [9] ได้แสดงให้เห็นว่างานวิจัยด้านการสกัดข้อมูล (Information Extraction : IE) ได้รับความนิยมในการนำไปประยุกต์ใช้กับงานทางด้านชีวการแพทย์ เนื่องจากงานวิจัยดังกล่าวสามารถนำไปสู่การ ค้นหาข้อมูล การได้ความองค์ความรู้ใหม่ และการสร้างสมมติฐานอย่างมีประสิทธิภาพ ตัวอย่างงานวิจัยด้านการสกัดข้อมูลอื่นๆเช่นงานวิจัย [7] และ [8] ได้นำเสนอการสกัด Name Entity ทางการแพทย์ โดยอาศัยหลักการของ Machine learning ในขณะที่งานวิจัย [9] เป็นงานวิจัยเกี่ยวกับการสกัดข้อมูลโดยอาศัยวิธี dictionary-based ในการสกัด Entity และ rule-based ในการสกัด Relation จากสื่อตีพิมพ์ biomedical ตัวอย่างงานวิจัยด้านการสกัดข้อมูลที่เป็นภาษาไทยคืองานวิจัย [10] ได้กล่าวถึงการสกัด Name Entity ในรูปแบบหลายคำในภาษาไทยบนคลังข้อความข่าวการเมืองโดยอาศัยวิธี Maximum Entropy Model โดยที่ผ่านมาพบว่างานวิจัยภาษาธรรมชาติด้านทางการแพทย์ในภาษาไทยยังไม่ได้รับความสนใจและมีน้อย

3. ขั้นตอนวิธีการที่นำเสนอ

ในบทความนี้ได้นำเสนอขั้นตอนวิธีการระบุอาการและอาการแสดงจากข้อความบอกเล่าอาการสำคัญภาษาไทยตามมาตรฐาน ICD-10 โดยขั้นตอนการประมวลผลภาษาธรรมชาติ ดังรูปที่ 1 โดยเริ่มจากการสกัดข้อความสำคัญ ได้แก่ อาการ อาการแสดง จากข้อความ CCs ซึ่งถูกบันทึกด้วยภาษาไทย โดยอาศัยขั้นตอนประมวลภาษาธรรมชาติ ได้แก่ การตัดคำ (tokenization) การทำรากศัพท์ (Word stemming) การตัดคำที่ไม่มีนัยสำคัญ (Stop word removal), การแทนข้อความให้อยู่ในรูปเวกเตอร์สเปซโมเดล (Vector Space Model) และการคำนวณค่าความเหมือน (similarity calculation) แล้วจึงส่งผ่านข้อความสำคัญซึ่งเป็นผลลัพธ์ที่ได้ไปทำการจับคู่กับรหัสอาการตามที่กำหนดโดยมาตรฐาน ICD-10เพื่อให้ได้รหัสของ ICD-10 ที่ตรงกับอาการที่ระบุในข้อมูล CCs ดังนั้นเพื่อเป็นการเตรียมข้อมูลให้พร้อมและเหมาะสมกับหัวข้อหรือบริบทของงานวิจัยซึ่งจะมีผลโดยตรงต่อผลลัพธ์และกระบวนการสกัดข้อมูล จึงจำเป็นที่จะต้องเข้าใจธรรมชาติของลักษณะข้อมูล เพื่อสามารถเตรียมข้อมูลให้ง่ายต่อการสกัดและได้ผลลัพธ์ที่มีความแม่นยำโดยขั้นตอนวิธีประกอบไปด้วย 2 ขั้นตอนหลัก คือ การเตรียมข้อมูลสำหรับการสกัดอาการและอาการแสดง และการสกัดอาการและอาการแสดง รูปที่ 1 แสดงภาพรวมของกระบวนการ

ระบุโรคจากการสกัดข้อความบอกเล่าอาการผู้ป่วยโดยอาศัยเทคนิคการประมวลผลภาษาธรรมชาติ โดยเริ่มจากเอกสาร CCs และ ICD-10 จะถูกส่งผ่านกระบวนการเตรียมข้อมูลเพื่อนาไปใช้ในการระบุอาการและอาการแสดง โดยก่อนที่จะนำเอกสารข้อความ CCs ทดสอบเข้าสู่กระบวนการการระบุอาการและอาการแสดง ระบบจะทำการตัดคำและติดป้ายเชิงความหมายโดยใช้พจนานุกรมทั่วไปและพจนานุกรมเชิงความหมายที่ได้จากการเตรียมข้อมูลเมื่อเตรียมข้อความ CCs ทดสอบก่อนเพื่อให้อยู่ในรูปแบบที่พร้อมสำหรับการระบุอาการและอาการแสดง จากนั้นจะนำเข้าสู่ขั้นตอนการสกัดอาการและอาการแสดงโดยอาศัยข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูล เพื่อให้สามารถจำแนกอาการและอาการแสดงจากข้อความ CCs ให้อยู่ในรหัส ICD-10 ได้ในที่สุด



รูปที่ 1 กระบวนการระบุโรคจากการสกัดข้อความบอกเล่าอาการผู้ป่วยโดยอาศัยเทคนิคการประมวลผลภาษาธรรมชาติ

3.1 การเตรียมข้อมูลสำหรับสกัดอาการและอาการแสดง

การเตรียมข้อมูลประกอบไปด้วยขั้นตอนย่อยดังแสดงในรูปที่ 2 โดยผลลัพธ์ของขั้นตอนดังกล่าวประกอบไปด้วยคลังข้อความ CCs (CCs corpus : A), คลังข้อความ ICD-10(ICD-10 corpus : B), ผลลัพธ์ของแต่ละกรณี CCs(CCs result : C) และ พจนานุกรมเชิงความหมาย (semantic dictionary : D)

3.1.1การเตรียมคลังข้อความอาการที่มาจาก ICD-10

คลังข้อความอาการและอาการแสดงจาก ICD-10 มีความสำคัญในการระบุหรือจำแนกอาการและอาการแสดงที่ถูกระบุหรืออธิบายอยู่ในข้อความ CCs โดยทำการตัดคำและติดป้ายในการอธิบายแต่ละคำจากเอกสาร ICD-10 ผลลัพธ์ของขั้นตอนนี้คือคลังข้อความ ICD-10 ตัวอย่างรหัส ICD-10 รหัส R51 "ปวดศีรษะ"[headache] ในการตัดคำด้วยโปรแกรม LongLexTo (http://www.sansarn.com/lexto/) ซึ่งเป็นโปรแกรมประยุกต์ที่ถูกพัฒนาบนภาษา Java โดยอาศัยวิธี LM โดยผลลัพธ์ที่จากการโปรแกรม LongLexTo คือ "ปวดศีรษะ" อย่างไรก็ตามเนื่องจากรูปแบบของภาษาไทยทั่วไปคำว่า "ปวดศีรษะ" สามารถแสดงในรูปอื่นๆที่มีความหมายเดียวกันได้ เช่น "ปวดหัว" หรือ "ปวดบริเวณศีรษะ" ซึ่งจะเห็นได้ว่าหากใช้คำว่า "ปวดศีรษะ" จะไม่สามารถใช้ในการระบุ"ปวดบริเวณศีรษะ" ได้ ดังนั้นหลังจากการตัดคำด้วยโปรแกรม LongLexTo แล้วจะต้องมีกระบวนการเพิ่มเติมเพื่อพิจารณาว่าคำดังกล่าวสามารถตัดคำเป็นส่วนย่อยโดยสามารถคงความหมายเดิมไว้ได้หรือไม่เช่น ปวด[ache]|ศีรษะ[head] เมื่อทำการพิจารณาเช่นนี้แล้วจะสามารถแก้ไข้ปัญหาในการระบุ" ปวดบริเวณศีรษะ" ได้

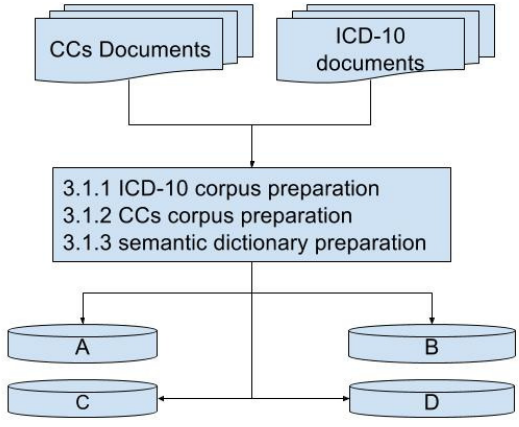
ตารางที่ 1 ตัวอย่างข้อมูลใน ICD-10 corpus

รหัส ICD-10	อาการ (ภาษาอังกฤษ)	อาการ (ภาษาไทย)	ผลลัพธ์จาก LongLexto	ตัวอย่างข้อมูล ICD-10 corpus
R00.2	Palpitations	ใจสั่น	ใจสั่น|	ใจ[heart] สั่น[beat]
R04.0	epistaxis	เลือดกำเดาไหล	เลือด กำเดา| ไหล|	เลือดกำเดา [epistaxis]ไหล[discharge]
R05	cough	ไอ	ไอ	ไอ[cough]
R07.0	Pain in throat	เจ็บในคอ	เจ็บ|ใน| คอ|	เจ็บ[pain]ใน [in]คอ [throat]
R07.1	Chest pain on breathing	เจ็บหน้าอกเวลาหายใจ	เจ็บ| หน้าอก | เวลา| หายใจ|	เจ็บ [pain]หน้าอก [chest]เวลา [when]หายใจ [breathe]
R07.4	Chest pain	เจ้าหน้าอก	เจ็บ| หน้าอก|	เจ็บ [pain]หน้าอก [chest]

รหัส ICD-10	อาการ (ภาษาอังกฤษ)	อาการ (ภาษาไทย)	ผลลัพธ์จาก LongLexto	ตัวอย่างข้อมูล ICD-10 corpus
R42	Dizziness and giddiness	เวียนศีรษะ	เวียน|ศีรษะ|	เวียน [dizziness]ศีร ษะ [head]
R51	headache	ปวดศีรษะ	ปวด ศีรษะ|	ปวด [ache]ศีรษะ [head]

3.1.2การเตรียมคลังข้อความบอกเล่าอาการ

ในขั้นตอนนี้จะประกอบไปด้วยส่วนของการตัดคำและการติดป้ายอธิบาย พร้อมทั้งระบุผลลัพธ์ของแต่ละกรณี โครงสร้างของคลัง CCs ประกอบไปด้วย ลำดับกรณี ประโยค CCs และ ประโยค CCs ที่ถูกตัดและติดป้ายแล้วผลลัพธ์ของขั้นตอนนี้คือ คลังข้อความ CCs และ ผลลัพธ์ของแต่ละกรณีของ CCs ตัวอย่างข้อความ CCs ที่ถูกตัดด้วย LongLexTo มีผลลัพธ์ ดังนี้ “|มีอา|การ|เกร็ง|และ|อ่อนแรง|ของ|แขน| |ขา| |ทั้ง| |2| |ข้าง| |มา|ประมาณ| |1| |ปี| |6| |เดือน|” จะเห็นได้ว่าคำว่า "มีอาการ" ถูกตัดออกเป็น "|มีอา|การ|" เป็นการตัดคำที่ไม่ตรงความหมาย หากเป็นคำที่ไม่มีนัยสำคัญแล้วจะไม่สนใจว่าตัดคำได้ถูกต้องหรือไม่ ทั้งนี้เนื่องจากคำดังกล่าวจะไม่ถูกนำมาพิจารณาในการระบุอาการหรืออาการแสดง แต่ในความเป็นจริงแล้วคำว่า "อาการ" หรือ "มีอาการ" เป็นคำที่มักขึ้นต้นในการกล่าวถึงอาการเสมอจึงถูกพิจารณาให้เป็นคำที่มีนัยสำคัญหรือมีส่วนในการระบุอาการหรืออาการแสดงนั่นเอง



รูปที่ 2 การเตรียมข้อมูลสำหรับการสกัดอาการและอาการแสดง

3.1.3 การเตรียมพจนานุกรมเชิงความหมาย

พจนานุกรมเชิงความหมายจะเป็นคำที่ถูกนำมาจากคลังข้อความทั้งสองในขั้นต้น ซึ่งจะถูกนำมาใช้ในกระบวนการต่อไปนี้ การตัดคำ การติดป้ายอธิบายคำศัพท์ และ การแก้ไขปัญหาความพ้องความหมาย

การตัดคำ เป็นการตัดคำโดยอาศัยพจนานุกรมมีความจำเป็นในการตัดคำที่จำเป็นที่จะต้องตัดคำสำคัญได้อย่างถูกต้อง เพื่อนำคำสำคัญดังกล่าวมาใช้ในการประมวลผลถัดไป

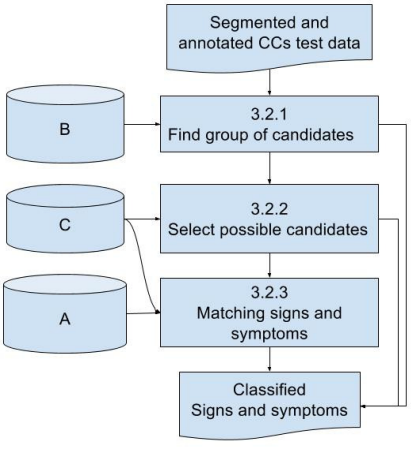
การติดป้ายอธิบาย โดยป้ายอธิบายถูกแบ่งออกเป็นสามประเภท ป้ายเชิงความหมาย ป้ายเชิงรูปแบบ และ ป้ายใดๆ

ในการแก้ไขปัญหาความพ้องความหมาย เช่น หัว/region,head ศีรษะ/region,head จะพบว่าทั้งสองมีรูปประกอบและคำอ่านออกเสียงที่ต่างกันแต่มีความหมายเหมือนกัน หรือ ที่/หน้าอก หรือ บริเวณ/หน้าอก บริเวณ/อก พบว่าการประกอบของวลีมีความแตกต่างกันแต่สื่อความหมายเดียวกัน

- ป้ายเชิงความหมาย คือป้ายที่ระบุถึงอาการสำคัญที่ประกอบอยู่ในคลังข้อความอาการ
- ป้ายเชิงรูปแบบ คือป้ายเชิงความหายที่ไม่ได้สื่อหรือระบุถึงอาการสำคัญแต่มีส่วนในการใช้พิจารณารูปแบบเพื่อหาอาการสำคัญ
- ป้ายใดๆ คือป้ายที่เป็นได้ทั้งป้ายเชิงความหมายและป้ายเชิงรูปแบบหรือไม่เป็นทั้งสองอย่างก็ได้

3.2 การสกัดอาการและอาการแสดง

ในการสกัดอาการและอาการแสดงจะใช้คลังข้อความ CCs ที่ตัดโดยพจนานุกรม โดยสามารถแบ่งขั้นตอนการสกัดอาการออกได้เป็น 3 ขั้นตอนดังนี้



รูปที่ 3 กระบวนการการสกัดอาการและอาการแสดง

3.2.1 การหาอาการและอาการแสดงที่เป็นไปได้

ในการหาอาการที่เป็นไปได้จากข้อความ CCs ที่ใช้ทดสอบจะพิจารณาจากความสัมพันธ์ระหว่างป้ายเชิงความหมายสำคัญที่ถูกติดบนข้อความ CCs และจากป้ายเชิงความหมายสำคัญที่ถูกติดบนคลังอาการ

3.2.2 การเลือกอาการที่เป็นไปได้จากคลังอาการ

ในขั้นตอนที่ผ่านมาทำให้สามารถระบุอาการที่อาจถูกระบุอยู่ในข้อความ CCs โดยอาการที่ยังไม่ได้ถูกยอมรับในขั้นตอนที่ 3.2.1 จะทำการเปรียบเทียบกับรูปแบบที่เกิดขึ้นในกับผลลัพธ์ของคลังข้อความ CCs ที่มีผลลัพธ์ตรงกับอาการที่มีตรงกับอาการที่น่าจะเป็นไปได้ในขั้นตอนที่ 3.2.1 หากเคยเกิดรูปแบบดังกล่าวขึ้นจริงและภายในรูปแบบของป้ายเชิงความสำคัญไม่ได้มีป้ายเชิงความหมายอื่นๆปะปนอยู่ก็สามารถระบุได้ว่าเป็นอาการดังกล่าวจริง

3.2.3 การจับคู่อาการและอาการแสดง

จากขั้นตอน 3.2.2 หากพบว่ามีรูปแบบถูกต้องแต่มีป้ายเชิงความหมายอื่นปนอยู่ก็จะทำการเปรียบเทียบรูปแบบจากข้อมูลทั้งผลลัพธ์ของข้อความทดสอบ และ ข้อความทดสอบจากคลังข้อความ CCs

4.บทสรุปและงานในอนาคต

บทความนี้เสนอกระบวนการเพื่อระบุชื่อโรคหรืออาการที่เป็นไปได้ในรูปแบบของรหัส ICD-10 โดยอาศัยกระบวนการตามทฤษฎีของการประมวลผลภาษาธรรมชาติเพื่อสกัดข้อความ CCs ภาษาไทย สำหรับงานในอนาคตผู้วิจัยจะทำการสร้างระบบเพื่อสนับสนุนการทำงานตามกระบวนการที่ได้นำเสนอไปแล้วนั้น และนำระบบที่ได้ไปทดสอบกับข้อความ CCs ที่เป็นข้อมูลจริง และจะนำผลลัพธ์ที่ได้มาตรวจสอบความถูกต้องและความแม่นยำของกระบวนการดังกล่าวโดยการเปรียบเทียบกับข้อความ CCs ที่ได้รับการวินิจฉัยและระบุรหัส ICD-10 โดยแพทย์จากโรงพยาบาล เพื่อประเมินประสิทธิภาพของกระบวนการที่นำเสนอในงานวิจัยนี้

เอกสารอ้างอิง

[1] Wu, Tsung-Shu Joseph, et al. "Establishing a nationwide emergency department-based syndromic surveillance system for better public health responses in Taiwan." BMC public health 8.1 (2008): 1.

[2] Lu, Hsin-Min, et al. "Multilingual chief complaint classification for syndromic surveillance: An experiment with Chinese chief complaints." international journal of medical informatics 78.5 (2009): 308-320.

[3] Ketui, Nongnuch, Thanaruk Theeramunkong, and Chutamanee Onsuwan. "Thai elementary discourse unit analysis and syntactic-based segmentation." International Information Institute (Tokyo). Information 16.10 (2013): 7423.

[4] Chamyapornpong. S. 1983. A Thai Syllable Separation Algorithm. Master thesis, Asian

[5] Poonwarawan. Y. 1986. Dictionary-based Thai Syllable Separation. In proceeding of the 9th Electrical Engineering Conference

[6] Sornlertlamvanich. V. 1993. Word Segmentation for Thai in a Machine Translation System. NECTEC, Bangkok.

[7] Jiang, Min, et al. "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries." Journal of the American Medical Informatics Association 18.5 (2011): 601-606.

[8] Chen, Yukun, et al. "A study of active learning methods for named entity recognition in clinical text." Journal of biomedical informatics 58 (2015): 11-18.

[9] Song, Min, et al. "PKDE4J: Entity and relation extraction for public knowledge discovery." Journal of biomedical informatics 57 (2015): 320-332.

[10] Chanlekha, Hutchatai, and Asanee Kawtrakul. "Thai named entity extraction by incorporating maximum entropy model with simple heuristic information." Proceedings of the IJCNLP. 2004.