รายงานการวิจัย

เรื่อง

การหาความสัมพันธ์ระหว่างปัญหาและวิธีการแก้ปัญหาจากเอกสารภาษาไทยบนเว็บบอร์ด

Determination of Problem-Solving Relation from Thai Documents on Web-Board

โดย

รองศาสตราจารย์ ดร. ฉวีวรรณ เพ็ชรศิริ
อรอุมา มูลวัตร

มหาวิทยาลัยธุรกิจบัณฑิตย์

รายงานผลการวิจัยนี้ได้รับทุนอุดหนุนจากมหาวิทยาลัยธุรกิจบัณฑิตย์

พ.ศ. 2559

ชื่อเรื่อง : การหาความสัมพันธ์ระหว่างปัญหาและวิธีการแก้ปัญหาจากเอกสารภาษาไทยบนเว็บบอร์ด
ผู้วิจัย : รองศาสตราจารย์ ดร. ฉวีวรรณ เพ็ชรศิริ
ผู้วิจัยร่วม : อรอุมา มูลวัตร
สถาบัน : มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีที่พิมพ์ : พุทธศักราช 2559 **สถานที่พิมพ์** : มหาวิทยาลัยธุรกิจบัณฑิตย์
แหล่งที่เก็บรายงานการวิจัยฉบับสมบูรณ์ : ศูนย์วิจัยมหาวิทยาลัยธุรกิจบัณฑิตย์
จำนวนหน้าวิจัย : 52 หน้า **ลิขสิทธิ์** : มหาวิทยาลัยธุรกิจบัณฑิตย์
คำสำคัญ : ความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา, ความสัมพันธ์อาการ-วิธีการรักษา, คู่คำ

บทคัดย่อ

จุดประสงค์ของงานวิจัยนี้คือการหาและสกัดความสัมพันธ์ระหว่างปัญหาและวิธีการแก้ปัญหา โดยเฉพาะความสัมพันธ์อาการ-วิธีการรักษา ซึ่งเป็นความสัมพันธ์ระหว่างปัญหาคืออาการกับวิธีการแก้ปัญหาคือวิธีการรักษาจากเอกสารบนเว็บบอร์ดของโรงพยาบาล ผลของงานวิจัยนี้คือ องค์ความรู้อย่างมีเหตุและผลบนพื้นฐานเอกัตถประโยคหรืออีดียู (EDU, Elementary Discourse Unit) หลาย ๆ ประโยค ซึ่งเป็นเป็นประโยชน์ต่อชาวบ้านหรือบุคคลทั่วไปในการแก้ไขปัญหาสุขภาพระดับเบื้องต้นผ่านระบบถามตอบอัตโนมัติ อย่างไรก็ตามงานวิจัยนี้มีสามปัญหาหลัก: ปัญหาแรกคือ การระบุอีดียูที่มีแนวความคิดอาการและอีดียูที่มีแนวความคิดวิธีการรักษา ปัญหาที่สองคือ การหาขอบเขตของอีดียูที่มีแนวความคิดอาการ และการหาขอบเขตของอีดียูที่มีแนวความคิดวิธีการรักษา และปัญหาที่สามคือ การหาความสัมพันธ์อาการ-วิธีการรักษาจากเอกสาร ดังนั้นงานวิจัยนี้ได้นำคู่คำ (Word-occurrence) ที่มีแนวความคิดอาการและแนวความคิดวิธีการรักษามาระบุอีดียูที่มีแนวความคิดอาการและอีดียูที่มีแนวความคิดวิธีการรักษาตามลำดับ งานวิจัยนี้ได้ใช้เทคนิคเอสวีเอม (SVM, Support Vector Machine) มาทำการหาขอบเขตของอีดียูที่มีแนวความคิดอาการและขอบเขตของอีดียูที่มีแนวความคิดวิธีการรักษา และสุดท้ายงานวิจัยนี้ขอเสนอการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning Techniques) ด้วยเนย์อีฟเบย์ (Naïve Bayes) มาทำการเรียนรู้ความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา ฉะนั้นจากงานวิจัยนี้ผลลัพธ์การสกัดความสัมพันธ์อาการ-วิธีการรักษามีความถูกต้องคิดเป็นค่าความแม่นยำ (Precision) และค่าระลึก (Recall) คือ 84% และ72% ตามลำดับ

Title : Determination of Problem-Solving Relation from Thai Documents on Web-Board

Researcher : Assoc. Prof. Chaveevan Pechsiri

Co-Researcher : Onuma Moolwat

Institution : Dhurakijpundit University.

Year of Publication : 2016 **Publisher** : Dhurakijpundit University.

Sources : Dhurakijpundit University Research Center.

Number of Pages : 52 Pages **Copyright** : Dhurakijpundit University.

Keyword : Problem-Solving Relation, Symptom-Treatment Relation, Word-Co

Abstract

This paper aims to determine and extract the Problem-Solving relation, especially the Symptom-Treatment relation, which is the relation between the problems as the disease symptoms and the solving steps as the treatment steps from hospital-web-board documents downloaded from the non-governmental organization websites of the certain hospitals. The research results are the reasoning knowledge (based on multiple simple sentence or EDUs, Elementary Discourse Units) which benefits for inexpert-people to solve their health problems in preliminary through the automatic question-answering system. The research contains three problems: first is how to identify a symptom-concept EDU and a treatment-concept EDU. Second is how to determine a symptom-concept-EDU boundary and a treatment-concept-EDU boundary. Third is how to determine the Symptom-Treatment relation from documents. Therefore, we apply a word co-occurrence having a symptom/treatment concept to identify a disease-symptom-concept/treatment-concept EDU, respectively, and also apply Support Vector Machine as the machine learning technique to solve their boundaries. We propose using Naïve Bayes to determine the Symptom-Treatment relation from documents with two feature groups, a symptom-concept-EDU group and a treatment-concept-EDU group. Finally, the result of extracting the Symptom-Treatment relation shows successfully the precision and recall of 84% and 72%, respectively.

กิตติกรรมประกาศ

ขอขอบพระคุณ รองอธิการบดีฝ่ายวิจัยและวิทยาบริการ มหาวิทยาลัยธุรกิจบัณฑิตย์ที่ให้โอกาสข้าพเจ้าและอาจารย์ อรอุมา มูลวัตร ในการศึกษาค้นคว้าวิจัยเรื่อง “การหาความสัมพันธ์ระหว่างปัญหาและวิธีการแก้ปัญหาจากเอกสารภาษาไทยบนเว็บบอร์ด” จนสำเร็จ

ขอขอบพระคุณ คณบดีคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์ ที่ช่วยตรวจสอบการใช้ภาษาไทยให้เหมาะสม

ขอขอบพระคุณ มหาวิทยาลัยธุรกิจบัณฑิตย์ ที่ให้เงินทุนสำหรับสนับสนุนโครงงานวิจัยนี้

ท้ายที่สุด ขอกราบขอบพระคุณ คุณพ่อ ครอบครัว ญาติพี่น้องและเพื่อนๆ ที่ให้กำลังใจในการทำโครงงานวิจัยที่มีค่านี้

รองศาสตราจารย์ ดร. ฉวีวรรณ เพ็ชรศิริ
หัวหน้าโครงการ
18 พฤษภาคม 2559

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตาราง **หน้า**

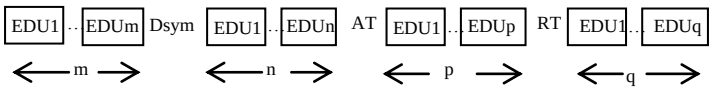
สารบัญรูป

การหาความสัมพันธ์ระหว่างปัญหาและวิธีการแก้ปัญหาจากเอกสารภาษาไทยบนเว็บบอร์ด

บทนำ

1. ความเป็นมาของปัญหา

ในสังคมปัจจุบันประชาชนส่วนใหญ่รู้จักใช้เทคโนโลยีสารสนเทศและการสื่อสารมาช่วยสนับสนุนการดำเนินธุรกิจรวมทั้งการขอคำปรึกษาแนวทางแก้ไขปัญหาต่าง ๆโดยผ่านทางเว็บบอร์ด (Web-Board) ซึ่งใช้เป็นศูนย์กลางของการชี้แนะแนวทางแก้ไขปัญหา (Solving Center) กล่าวคือ เป็นศูนย์กลางของการแสดงปัญหาต่าง ๆและวิธีแก้ปัญหาของชาวบ้านหรือประชาชนทั่วไป รวมทั้งการแสดงแนวทางแก้ปัญหาเหล่านั้นจากผู้เชี่ยวชาญ โดยเฉพาะอย่างยิ่งปัญหาสุขภาพที่แสดงบนเว็บบอร์ด (http://haamor.com/) ซึ่งเป็นเว็บที่ดำเนินการโดยเอนจีโอ (NGO, non-governmental organization) ทำให้ชาวบ้านของชุมชนต่าง ๆที่ไม่สะดวกไปโรงพยาบาลเพื่อรอรับการรักษา ได้รับความรู้เกี่ยวกับแนวทางการแก้ไขปัญหาสุขภาพ หรือแนวทางเบื้องต้นของการปฐมพยาบาลรักษาโรคตามอาการที่ปรากฏ อย่างไรก็ตามบางครั้งชาวบ้านเหล่านั้นได้แสดงปัญหาเกี่ยวกับอาการต่าง ๆของโรคบนเว็บบอร์ดดังกล่าวแต่ใช้เวลาสองสามวันในการได้รับคำตอบจากแพทย์ผู้เชี่ยวชาญ ทำให้อาการของโรคมีโอกาสค่อย ๆรุนแรงขึ้นในระหว่างที่รอคำปรึกษาจากแพทย์ ดังนั้นงานวิจัยนี้มุ่งเน้นหาความสัมพันธ์ระหว่างปัญหาและวิธีการแก้ปัญหาซึ่งก็คือการหาความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา (Problem-Solving Relation) โดยเฉพาะความสัมพันธ์อาการ-วิธีการรักษา (Symptom-Treatment Relation) จากเอกสารภาษาไทยบนเว็บบอร์ด (http://haamor.com/) ทั้งปัญหา/อาการ และวิธีการแก้ปัญหา/วิธีการรักษา แสดงออกมาในรูปแบบของหลาย ๆประโยค แต่ละประโยคเรียกว่า EDU (Elementary Discourse Unit ซึ่งก็คือประโยคบอกเล่าแบบง่าย ๆ (Simple Sentence หรือ Clause)) โดย EDU เหล่านี้จะอยู่เรียงกันตามรูปแบบต่อไปนี้พร้อมตัวอย่างในรูปที่ 1



เมื่อ

- Dsym, AT, และ RT เป็นกลุ่มของ EDU ที่มีแนวความคิดอาการ, กลุ่มของ EDU ที่มีแนวความคิดวิธีการักษาตามปฏิบัติจริง, กลุ่มของ EDU ที่มีแนวความคิดวิธีการักษาตามคำแนะนำของผู้เชี่ยวชาญตามลำดับ

Dsym = ($EDU_{sym-1}EDU_{sym-2}$ .. EDU_{sym-a}) เมื่อ a คือเลขจำนวนเต็มมีค่า>0,

AT= (EDU_{at-1} EDU_{at-2} ..EDU_{at-b}) เมื่อ b คือเลขจำนวนเต็มมีค่า≥0,

RT= (EDU_{rt-1} EDU_{rt-2} ..EDU_{rt-c}) เมื่อ c คือเลขจำนวนเต็มมีค่า≥0

- m, n, p, and q คือจำนวน EDU และมีค่า≥0

Problem-Topic: Stomachache

EDU1(symptom):[*ผู้ป่วย*]*ปวดท้องอย่างหนัก*; [*ผู้ป่วย*]/[*A patient*] *ปวดท้อง*/*has a stomachache* *อย่างหนัก*/*heavily*
([A patient] has a stomachache heavily.)

EDU2 (symptom): [*ผู้ป่วย*]*มีแก๊สในกระเพาะมาก*; [*ผู้ป่วย*]/[*The patient*] *มี*/*has* *แก๊ส*/*gas* *ในกระเพาะ*/ *inside stomach* *มาก* /a lots
([The patient] has lots of gas in the stomach)

EDU3 (symptom): *อาการมักจะเป็นหลังทานข้าวเย็นและตอนกลางคืน*; *อาการ*/*symptom* *มักจะเป็น*/*mostly occurs* *หลัง*/*after* *ทานข้าวเย็น*/*having dinner* *และ*/*and* *ตอนกลางคืน*/*at night*
(The symptom mostly occurs after having dinner and at night)

EDU4: [*ผู้ป่วย*]*สงสัยเป็นโรคกระเพาะ*; [*ผู้ป่วย*]/[*The patient*] *สงสัย*/*doubts* *เป็นโรคกระเพาะ*/*to get a gastropathy*
([The patient] doubts to get a gastropathy)

EDU5 (treatment): [*ผู้ป่วย*]*กินยาลดกรดเพื่อแก้ปวดท้อง*; [*ผู้ป่วย*]/[*The patient*] *กินยาลดกรด*/*takes an antacid* *เพื่อแก้ปวดท้อง* /*to solve the stomach ache*
([The patient] takes an antacid to solve the stomach ache)

EDU6: *แต่ก็ไม่หายปวด*; *แต่*/*But* *ก็ไม่หายปวด* /*it cannot work.* **(But it cannot work)**

Physician Suggestion

EDU7 *ไปหาหมอหรือยัง* /*Have you seen the doctor?*

EDU8 (recommendation): *ถ้า*[*ผู้ป่วย*]*เป็นโรคกระเพาะ*; *ถ้า* /*If [the patient*] *เป็นโรคกระเพาะ*/ *get a gastropathy*
(If [the user] gets a gastropathy)

EDU9 (recommendation): [*ผู้ป่วย*]*ก็อาจต้องกินยาลดการหลั่งกรดในกระเพาะอาหาร*; [*ผู้ป่วย*]/[*The patient*] *ก็อาจต้องกินยา*/*may take a medicine* *ลด*/*to reduce* *การหลั่งกรดในกระเพาะอาหาร*/ *gastric acid secretion*
([the user] may take a medicine to reduce the gastric acid secretion)

EDU10 (recommendation): *หลีกเลี่ยงอาหารที่ทำให้เกิดแก๊สในกระเพาะ*; *หลีกเลี่ยง*/*avoid* *อาหาร*/*food* *ที่ทำให้เกิด*/ *causing* *แก๊สใน กระเพาะ*/*stomach gassy* **(Avoid food causing stomach gassy.)**

รูปที่ 1 แสดงตัวอย่างความสัมพันธ์อาการ–วิธีการรักษาในเอกสารบนเว็บบอร์ด (สัญญลักษณ์ [..] หมายถึง การละคำที่อยู่ในสัญญลักษณ์)

จากรูปที่ 1 Dsym คือ EDU1-EDU3, AT คือ EDU5, และ RT คือ EDU8 - EDU10 ดังนั้นการหาหรือสกัดความสัมพันธ์ปัญหา-วิธีการแก้ไขปัญหา เช่น ความสัมพันธ์อาการ-วิธีการรักษา จากงานวิจัยนี้ สามารถนำไปใช้ในระบบการตอบคำถามอัตโนมัติสำหรับคำถามอย่างไร (How Question) ประเภท แสดงวิธีการแก้ปัญหา/วิธีการรักษาโรค บนเครือข่ายสังคมออนไลน์ (Social Network) อย่างไรก็ตามงานวิจัยการหาความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา (โดยเฉพาะความสัมพันธ์อาการ-วิธีการรักษา) ประกอบด้วยสามปัญหาหลักดังนี้

ก. จะสามารถระบุ EDU ที่มีแนวความคิดอาการ/ปัญหา (Symptom Concept EDU, EDU_{sym}) และระบุ EDU ที่มีแนวความคิดวิธีการรักษา/วิธีการแก้ไขปัญหา (Treatment Concept EDU, EDU_{treat}) จากเอกสารได้อย่างไร

ข. จะสามารถระบุขอบเขต EDU ที่มีแนวความคิดอาการ/ปัญหา (Dsym) และขอบเขต EDU ที่มีแนวความคิดวิธีการรักษา/วิธีการแก้ปัญหา (AT/RT) ได้อย่างไร

ค. การหาความสัมพันธ์ ปัญหา-วิธีการแก้ปัญหา (เช่น ความสัมพันธ์อาการ-วิธีการรักษา) จากคู่เวคเตอร์ระหว่างเวคเตอร์ EDU ทีแสดงแนวความคิดอาการ และเวคเตอร์ EDU ที่แสดงแนวความคิดวิธีการรักษา ($\langle EDU_{sym\text{-}1}EDU_{sym\text{-}2}\ldots EDU_{sym\text{-}m}\rangle\langle EDU_{treat\text{-}1}EDU_{treat\text{-}2}\ldots EDU_{treat\text{-}n}\rangle$) ได้อย่างไร

ดังนั้นจากงานวิจัยที่เกี่ยวข้องเช่น Rosario B.(2005) ได้สกัดความสัมพันธ์เกี่ยวกับความหมาย (Semantic Relation) ระหว่างสองเอนทิตี้ที่เป็นคำนามเช่น “โรค/Disease” “วิธีการรักษา/Treatment” ในหนึ่งประโยค จากเอกสารวิทยาศาสตร์ทางชีวะภาพ ฉะนั้นความสัมพันธ์ที่สกัดได้คือ ความสัมพันธ์ระหว่าง DIS กับ TREAT (Disease and Treatment Relation) เช่น ชื่อโรค และชื่อยาที่ใช้รักษา โดย Rosario B.(2005) ใช้ MeSH เป็นฐานความรู้สำหรับหาแนวความคิดของ DISและTREAT และใช้ตัวแบบกราฟ_ไวยกรณ์ที่รวมความหมายเข้าไว้ด้วยมาเป็นตัวระบุความสัมพันธ์ระหว่างคำนามสองเอนทิตี้ Abacha A. B. and Zweigenbaum P. (2011) ได้สกัดความสัมพันธ์ทางความหมายเกี่ยวกับวิธีการรักษา (Treatment Relation) ระหว่างวิธีการรักษาและโรค บนพื้นฐาน UMLS และรูปแบบทางภาษา (Linguistic Pattern) มาทำการสกัดความสัมพันธ์นี้ที่เกิดขึ้นในหนึ่งประโยค

Song S. et al, (2011) ได้สกัดความรู้ทางกรรมวิธี (Procedural Knowledge) จากบทคัดย่อของเมดลาย (MEDLINE abstract) โดยใช้ SVM (Support Vector Machine) และ CRF (Conditional Random Field) ร่วมกับการประมวลภาษาธรรมชาติ เพื่อให้ได้กรรมวิธีขั้นตอนการแก้ปัญหาที่สอดคล้องกันกับเป้าประสงค์ (Target)

Yeleswarapu et al., (2014) ได้สกัดหาคู่ความสัมพันธ์ระหว่างเหตุการณ์ที่ไมพึงประสงค์จากการใช้ยา (drug-Adverse Event, drug-AE) จากเอกสารการวิจารณ์ของผู้ใช้บนบล็อค (Blogs) และบทคัดย่อของเมดลาย (MED-LINE Abstract) โดยคำต่าง ๆที่เกี่ยวกับยา โรค และ อาการ อยู่ในรูปนามวลี และมีความสัมพันธ์กันภายในหนึ่งประโยค พวกเขาใช้ Bayesian Confidence Propagation Neural Network มาทำการหาคู่ความสัมพันธ์ระหว่างเหตุการณ์ของ drug-AE

อย่างไรก็ตามงานวิจัยก่อนหน้าที่กล่าวมานี้ ทำการหาความสัมพันธ์ ปัญหา-วิธีการแก้ปัญหา ปรากฏบนหนึ่งประโยคดังงานวิจัยของ Rosario B.(2005) Abacha A. B. and Zweigenbaum P. (2011) และ Yeleswarapu et al., (2014) แต่งานวิจัยของ Song S. et al, (2011) จะประกอบด้วย หนึ่งประโยคของปัญหา และหลายประโยคของวิธีแก้ปัญหา ในขณะที่งานวิจัยที่เสนอในครั้งนี้ เป็นการหาความสัมพันธ์ ปัญหา-วิธีการแก้ปัญหา ซึ่งปรากฏบนหลายประโยค/EDU กล่าวคือ 'ปัญหา' ซึ่งก็คือ 'อาการโรค'มักจะปรากฏบนหลาย EDU อย่างต่อเนื่องกันเพื่อบรรยายลักษณะอาการโรค ในทำนองเดียวกัน 'วิธีการแก้ปัญหา' มักจะปรากฏบนหลาย EDU อย่างต่อเนื่องกันเพื่อบรรยายวิธีการแก้ปัญหา ซึ่งก็คือวิธีการรักษา/ป้องกันโรค นอกจากนี้อาการ/ปัญหา และ วิธีการรักษา/วิธีการแก้ปัญหา สำหรับงานวิจัยนี้อยู่ในรูปของเหตุการณ์ (Event) ซึ่งสามารถแสดงด้วยกริยาวลี (Verb Phrase) ดังนั้นงานวิจัยนี้ได้นำคู่คำ (Word-occurrence หรือเรียกว่า Word-Co) จากสองคำที่อยู่ติดกันหลังจากได้กำจัดคำหยุด (Stop Word) ออกไป ในแต่ละ EDU โดยมีคำแรกเป็นคำกริยา (v_{co}) คำที่สองเป็นคำถัดมาที่เข้าคู่ (w_{co}) เมื่อ $v_{co} \in V_{co}$; $V_{co} = V_{co1} \cup V_{co2}$; V_{co1} คือเซตของคำกริยาที่มีแนวโน้มเป็นแนวความคิดอาการ V_{co2} คือเซตของคำกริยาที่มีแนวโน้มเป็นแนวความคิดวิธีการรักษา และ $w_{co} \in W_{co}$; $W_{co} = W_{co1} \cup W_{co2}$; W_{co1}และW_{co2}คือเซตของคำที่เข้าคู่แล้วทำให้คู่คำมีแนวความคิดอาการ (Symptom Concept) และแนวความคิดวิธีการรักษา (Treatment Concept) ตามลำดับ โดยแนวความคิดทั้งหมดได้มาจากการกำกับคลังคำด้วย WordNet และ MeSH

$V_{co1} = V_{co1\text{-}strong} \cup V_{co1\text{-}weakPlusInfomation}$ (เมื่อ $V_{co1\text{-}strong}$ คือ เซตของคำกริยาที่มีแนวความคิดอาการโดยตรง $V_{co1\text{-}weakPlusInfomation}$ คือ เซตของคำกริยา ได้แก่ '*มี*-verb'/'*have*', '*เป็น*-verb' /'*be*', และ '*รู้สึก*-verb'/'*feel*' ซึ่งไม่มีแนวความคิดอาการโดยตรง แต่จะมีแนวความคิดอาการเมื่อมีคำตามเฉพาะ เช่น ('*มี*-verb'/ '*have*' + '*อาการ*-noun'/'*symptom*') / 'have_a_symptom'

$V_{co1\text{-}strong}$ = {'*ถ่าย*-verb'/'*defecate*','*เบ่ง*-verb'/'*push*','*ปวด*-verb'/' *pain*','*เจ็บ*-verb'/'*pain*', ..}

$V_{co1\text{-}weakPlusInfomation}$={('*มี*-verb'/'*have*'+ '*อาการ*-noun'/ '*symptom*') / 'have_a_symptom', ('*เป็น*-verb'/'*be*'+'*หวัด*-noun'/'*flu*') / 'get_flu', ('*รู้สึก*-verb'/'*feel*'+'*อึดอัด*-verb'/'*be uncomfortable*') / 'feel_ uncomfortable' ,..}

V_{co2}={'*กิน*-verb'/'*consume*', '*ทา*-verb'/'*apply*', '*ใช้*-verb' /*apply*', '*รักษา*-verb'/'*remedy*', '*บำรุง*-verb'/'*nourish*', '*ลด*-verb'/'*reduce*',.. }

W_{co1}={‘ ’, ‘*ยาก*-adverb’/‘*difficultly*’, ‘*ถ่าย*-noun’/‘*stools*’, ‘*เชื้อ*-noun’/‘*germ*’, ‘*เหลว*-adverb’/‘*dissolvingly*’, ‘*ประจำเดือน*-noun’/‘*period*’, ‘*ท้อง,ศรีษะ,..ส่วนของร่างกาย*-noun’/‘*stomac,head,..human_organ*’, ‘*แน่น*[*ท้อง*]-adjective’/‘*fullness*’, ‘*ท้องเฟ้อ*-noun’/‘*flatulence*’, ‘*ไข้*-noun’/‘*fever*’, …}

W_{co2} = {‘*ยา*-noun’/‘*medicine*’, ‘*อาหาร*-noun’/‘*food*’, ‘*อาหารเสริม*-noun’/‘*supplement*’, ‘*ท้อง,ศรีษะ,..ส่วนของร่างกาย*-noun’/ ‘*stomac,head,..human_organ*’, ‘*ความดัน* [*เลือด*]-noun’/ ‘[*blood*] *pressure*’, ‘*ไขมัน*[*คลอเลสเตอรอล*]-noun’/‘[*cholesterol*]*fat*’,… }

ตัวอย่าง Word-Co ($v_{co}w_{co}$) ที่เป็น Word-Co_{sym} ($v_{co1}w_{co1}$; $v_{co1}\in V_{co1}$; $w_{co1}\in W_{co1}$):

- ‘*ถ่าย*-verb’/‘*defecate*’ + ‘*ยาก*-adverb’/‘*difficultly*’ → ‘be_constipated’
- ‘*ถ่าย*-verb’/‘*defecate*’ + ‘*เหลว*-adverb’/‘*dissolvingly*’→ ‘have_diarrhea’
- ‘*ปวด*-verb’/‘*pain*’ + ***‘กล้ามเนื้อ***-noun’/‘*muscle*’→ ‘have_muscle_pain’
- ‘*เจ็บ*-verb’/‘*pain*’ + ***‘หน้าอก***-noun’/‘*chest*’→ ‘have_chest_pain’
- (‘*มี*-verb’/‘*have*’ + ‘*อาการ*-noun’/‘*symptom*’) + ‘*ท้องเฟ้อ*-noun’/‘*flatulence*’/ ‘have_a_ flatulence_symptom’

ตัวอย่าง Word-Co ($v_{co}w_{co}$) ที่เป็น Word-Co_{treat} ($v_{co2}w_{co2}$; $v_{co2}\in V_{co2}$; $w_{co1}\in W_{co1}$):

- ‘***กิน-verb’/‘consume’*** + ‘*ยา*-noun’/‘*medicine*’ → ‘take_medicine’
- ‘*บำรุง*-verb’/‘*nourish*’+ ‘*ร่างกาย*-noun’/‘*body*’ → ‘nourish_the_body’
- ‘*ลด*-verb’/‘*reduce*’ + ‘*ไขมัน*-noun’/‘*fat*’ → ‘reduce_ cholesterol

ฉะนั้น Word-Co ($v_{co}w_{co}$) ที่เป็น Word-Co_{sym} ($v_{co1}w_{co1}$; $v_{co1}\in V_{co1}$; $w_{co1}\in W_{co1}$) และ Word-Co_{treat} ($v_{co2}w_{co2}$; $v_{co2}\in V_{co2}$; $w_{co2}\in W_{co2}$) ใช้ระบุ EDU_{sym} และ EDU_{treat} ตามลำดับ และงานวิจัยนี้ใช้เทคนิคของ SVM มาทำการหาขอบเขตของ Dsym และขอบเขตของ AT/RT และสุดท้ายงานวิจัยนี้ขอเสนอเทคนิคการเรียนรู้ของเครื่อง (Machine Learning Techniques) ด้วย Naïve Bayes มาทำการเรียนรู้ ความสัมพันธ์อาการ-วิธีการรักษา (ความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา) จากคู่เวคเตอร์ที่ประกอบด้วยเวคเตอร์คู่คำที่มีแนวความคิดอาการ (Word-Co_{sym} Vector) และเวคเตอร์คู่คำที่มีแนวความคิดวิธีการรักษา (Word-Co_{treat} Vector) โดย Word-Co_{sym} Vector แทนเวคเตอร์ EDU_{sym} (EDU_{sym} Vector) และ Word-Co_{treat} Vector แทนเวคเตอร์ EDU_{treat} (EDUtreat Vector)

2. วัตถุประสงค์

2.1. หา Word-Co_{sym} และ Word-Co_{treat} จากคู่คำของสองคำที่อยู่ติดกันและมีคำแรกเป็นคำกริยา ($v_{co}w_{co}$) ในแต่ละ EDU

2.2. หา ความสัมพันธ์อาการ-วิธีการรักษา (ความสัมพันธ์ปัญหา–วิธีการแก้ปัญหา) จากเอกสารภาษาไทย

3. สมมติฐาน

3.1. คู่คำของสองคำที่อยู่ติดกันหลังจากได้กำจัดคำหยุดออกไป และมีคำแรกเป็นคำกริยา ($v_{co}w_{co}$) ในแต่ละ EDUของโดเมนการดูแลสุขภาพทำให้คู่คำดังกล่าวมีแนวความคิดอาการสำหรับ $v_{co1}w_{co1}$ และแนวความคิดวิธีการรักษาสำหรับ $v_{co2}w_{co2}$

3.2. คู่เวคเตอร์ของ Word-Co_{sym}Vector และ Word-Co_{treat} Vector ภายใต้โดเมนการดูแลสุขภาพแสดงความสัมพันธ์ระหว่างอาการและวิธีการรักษาซึ่งก็คือความสัมพันธ์ปัญหา-วิธีการแก้ไขปัญหา

4. นิยามคำศัพท์

Problem-Solving Relation: ความสัมพันธ์ปัญหา–วิธีการแก้ปัญหา (ความสัมพันธ์ระหว่างปัญหาและวิธีการแก้ปัญหา)

Symptom-Treatment Relation: ความสัมพันธ์อาการ-วิธีการรักษา (ความสัมพันธ์ระหว่างอาการและวิธีการรักษา)

Word-occurrence / Word-Co: คู่คำ

Corpus: คลังคำ

5. ขอบเขตของการวิจัย

5.1. สามารถหาความสัมพันธ์ปัญหา-วิธีการแก้ปัญหาโดยเฉพาะความสัมพันธ์อาการ-วิธีการรักษา ที่ทราบสาเหตุปัญหาโดยเฉพาะชื่อโรค

5.2. ความสัมพันธ์อาการ–วิธีการรักษา (ความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา) ที่มาจากสาเหตุที่ไม่ซับซ้อนคือเป็นโรคเดียวหรือสาเหตุเดียว

งานวิจัยที่เกี่ยวข้อง

การวิจัยการสกัดความรู้เกี่ยวกับการหาความสัมพันธ์ปัญหา-วิธีการแก้ปัญหาจากเอกสารภาษาไทยบนเว็บบอร์ดประกอบด้วยความรู้พื้นฐาน และงานวิจัยก่อนหน้าดังนี้

ความรู้พื้นฐาน

1) Naïve Bayes Classifier

ตัวจัดประเภทเนอีฟเบย์ (Naïve Bayes classifier, NB) (Mitchell 1997) หรือตัวจัดประเภทNB เป็นวิธีการเรียนรู้ที่นิยมใช้กันมาก และเป็นการเรียนรู้ที่อยู่บนพื้นฐานของความน่าจะเป็น (Probability) กับข้อมูลที่สังเกต (Observed Data) ได้ใช้ในการทดลอง ตามที่ Mitchell T.M., (1997) ได้กล่าวว่าตัวจัดประเภท NB สามารถประยุกต์ใช้กับงานเรียนรู้ที่ซึ่งแต่ละตัวอย่าง x (Instance x) ได้ถูกอธิบายโดยการเชื่อมโยงค่าคุณลักษณะ (Attribute Values) ต่างๆ และที่ซึ่งฟังก์ชันเป้าหมาย (Target Function, f(x)) สามารถแสดงค่าคลาส (Class Value, v) จากคลาสไฟไนท์เซต (Class Finite Set, V) ดังนั้นเซตของตัวอย่างการเรียนรู้ของฟังก์ชันเป้าหมายได้ถูกกำหนดไว้ให้ และเมื่อมีตัวอย่างใหม่เกิดขึ้นก็สามารถอธิบายได้ คือบอกค่าคลาสได้ด้วยทูปเพิล (Tuple) ของค่าคุณลักษณะ หรือฟีเจอร์ (Feature) $<a_1, a_2,..a_n>$ นั่นคือตัวเรียนรู้ทำนายค่าเป้าหมายหรือการจัดแบ่งประเภทสำหรับตัวอย่างใหม่ที่เข้ามา

แนวทางเบย์ที่จะจัดประเภทให้กับตัวอย่างใหม่ที่เข้ามานั้นเป็นการกำหนดค่าเป้าหมายทีมีโอกาสเป็นไปได้มากที่สุด หรือที่เรียกว่า $v_{maximum\_a\_posterior}$ (v_{MAP}) เมื่อกำหนดค่าคุณลักษณะต่างๆให้ $<a_1, a_2,..a_n>$ ที่ใช้อธิบายตัวอย่าง ดังแสดงในสมการ(2) และ (3)

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2 ... a_n) \quad (1)$$

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2 ... a_n \mid v_j) P(v_j) \quad (2)$$

ตัวจัดประเภทNB ดำเนินงานบนพื้นฐานของข้อสมมติฐานแบบง่ายๆที่มีเงื่อนไขว่า ค่าคุณลักษณะแต่ละคุณลักษณะจะต้องเป็นอิสระต่อกันเมื่อกำหนดค่าเป้าหมายไว้ให้ กล่าวคือข้อสมมติฐานเป็นการกำหนดค่าเป้าหมายของตัวอย่าง (คลาสของตัวอย่าง) ฉะนั้นความน่าจะเป็นของการสังเกตการเชื่อมโยงกันของ $a_1, a_2,..a_n$ คือผลคูณของค่าความน่าจะเป็นของคุณลักษณะต่างๆ ดังนั้นตัวจัดประเภท NB, v_{NB}, สามารถแสดงได้ดังต่อไปนี้

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j) \quad (3)$$

สำหรับงานวิจัยนี้ เราได้ใช้ตัวจัดประเภท NB สำหรับการเรียนรู้สองเรื่องคือ

1.1) การเรียนรู้หาขอบเขตของ Dsym และของ AT/RT โดยการเรียนรู้แยกประเภทการสิ้นสุดขอบเขต EDU_{sym} และ ขอบเขต EDU_{treat} ใน Dsym และ AT/RT ตามลำดับ ดังแสดงในสมการ (4)

$$
\begin{aligned}
BoundaryClass &= \arg\max_{class \in Class} P(class \mid v_{coj-i} w_{coj-i}, v_{coj-i+1} w_{coj-i+1}) \\
&= \arg\max_{class \in Class} P(v_{coj-i} w_{coj-i} \mid class) P(v_{coj-i+1} w_{coj-i+1} \mid class) P(class) \qquad (4) \\
where\ & v_{coj-i} w_{coj-i} \in VW_j \quad and \quad v_{coj-i+1} w_{coj-i+1} \in VW_j \\
& i = \{1,2,..endOfboundary\} \qquad j = \{1,2\} \\
& IF \ \ j=1 \quad VW_j \ is \ VM_{symptom} \ which \ is \ a \ \ Word-Co_{sym} \ set \\
& IF \ \ j=2 \quad VW_j \ is \ VM_{treatment} \ which \ is \ a \ \ Word-Co_{treat} \ set \\
& Class = \{end, \ continue\}
\end{aligned}
$$

เมื่อตัวแปร Class เป็นไฟไนท์เซท (Finite Set) ของประเภทขอบเขตของ Dsym และ ของ AT/RT ได้สิ้นสุดหรือยังไม่สิ้นสุด {end, continue} และคุณลักษณะ $a_1, a_2, .. a_n$ คือ ฟีเจอร์คู่คำต่าง ๆ (Word-Co Features, $v_{coj\text{-}i} w_{coj\text{-}i}$ และ $v_{coj\text{-}i+1} w_{coj\text{-}i+1}$) ที่เป็นสมาชิกของ VW_j (ถ้า j=1 → $VW_j = VW_{symptom}$ เมื่อ $VW_{symptom}$ คือเซตคู่คำที่มีแนวคิดอาการ และถ้า j=2→ $VW_j = VW_{treatment}$ เมื่อ$VW_{treatment}$ คือเซตคู่คำที่มีแนวคิดวิธีการรักษา) นอกจากนี้ แต่ละ EDU สามารถถูกแทนด้วยหนึ่งคู่คำ ($v_{coj} w_{coj}$) ดังนั้นฟีเจอร์คู่คำเหล่านั้นได้จากการเลื่อนคู่ EDU ที่อยู่ติดกัน (EDUi, EDUi+1; ซึ่งสามารถถูกแทนด้วย $v_{coj\text{-}i} w_{coj\text{-}i}$, $v_{coj\text{-}i+1} w_{coj\text{-}i+1}$) ไปด้วยระยะทางหนึ่ง EDU (ดูข้อ กรรมวิธี, Method Section)

1.2) การเรียนรู้หาความสัมพันธ์อาการ-วิธีการรักษา ดังสมการ (5) โดยมีค่าความน่าจะเป็นของฟีเจอร์ต่าง ๆที่สกัดได้ เช่น Dsym และAT/RT นำมาใช้เป็นคุณลักษณะหรือฟีเจอร์ต่าง ๆ

$$
\begin{aligned}
SymptomTreatmentRelation &= \arg\max_{class \in Class} P(class \mid Dsym, Treatment) \\
&= \arg\max_{class \in Class} P(class \mid v_{co1-1} w_{co1-1}, v_{co1-2} w_{co1-2}, .., v_{co1-a} w_{co1-a}, v_{co2-1} w_{co2-1}, v_{co2-2} w_{co2-2}, .., v_{co2-y} w_{co2-y}) \\
&= \arg\max_{class \in Class} P(class) \prod_{num=1}^{a} P(v_{co1-num} w_{co1-num} \mid class) \prod_{num=1}^{y} P(v_{co2-num} w_{co2-num} \mid class) \qquad (5)
\end{aligned}
$$

ที่ซึ่ง Class = {“yes”, “no”}; a คือจำนวน EDU ใน Dsym และมีแนวความคิดอาการ y คือจำนวน EDU ใน AT/RT และมีแนวความคิดวิธีการรักษา (y =b ถ้าเป็น AT ; y =c ถ้าเป็น RT)

2) Support Vector Machine

ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine,SVM) (Cristianini N. and Shawe-Taylor J. ,2000) เป็นการเรียนรู้เชิงเส้นของเครื่องสำหรับการจำแนกประเภท (Classification) ข้อมูล ออกเป็นสองคลาส (Binary Classes) และสามารถประยุกต์กับกรณีหลายคลาส (Multiple Classes) หลักการของซัพพอร์ตเวคเตอร์แมชชีน คือ การสร้างไฮเปอร์เพลน (Hyper Plane) ที่เหมาะสมบนระนาบของข้อมูลตัวอย่างที่ใช้สอนการเรียนรู้ (Training Data) เพื่อแบ่งแยกกลุ่มข้อมูลที่แตกต่างกัน ในการสร้างไฮเปอร์เพลนที่เหมาะสม ระยะห่างระหว่างจุดของข้อมูลที่อยู่ใกล้กับไฮเปอร์เพลนมากที่สุดทั้งสองด้าน คือ d+ และ d- ระยะห่าง (Margin , γ) เกิดจากระยะ d+ + d- ไฮเปอร์เพลนที่เหมาะสม คือไฮเปอร์เพลนที่มีค่าระยะห่าง γ กว้างที่สุด โดยข้อมูลตัวอย่างที่อยู่บนขอบของมาร์จิ้น γ จะถูกเรียกว่า “เวคเตอร์สนับสนุน หรือ ซัพพอร์ตเวคเตอร์ (Support Vector)” ตัวอย่างข้อมูลสำหรับการเรียนรู้หรือข้อมูลอินพุต (Input Data) ปกติอยู่ในรูปของเวคเตอร์คุณลักษณะ (Attribute Vectors) ที่มี ปริภูมิอินพุท (Input Space)เป็นเซตย่อยของ R^n ฉะนั้นการจำแนกประเภทข้อมูลออกเป็นสองคลาส (Binary Classification) จะถูกแสดงออกมาด้วยฟังก์ชันของค่าจำนวนจริง(Real-Valued Function $f : X \subseteq R^n \rightarrow R$) ดังนั้น $x = (x_1,\ldots,x_n)$ ถูกกำหนดเป็นคลาสบวกถ้า $f(x) \geq 0$ มิฉะนั้นเป็นคลาสลบ ทั้งนี้โดยพิจารณาจากกรณี f(x) ที่เป็นฟังก์ชันเชิงเส้นของ $x \in X$ ดังสมการต่อไปนี้

$$\begin{aligned} f(\mathbf{x}) &= \langle \mathbf{w} \cdot \mathbf{x} \rangle + b \\ &= \sum_{i=1}^{n} w_i x_i + b \end{aligned} \tag{6}$$

เมื่อ $(w,b) \in R^n \times R$ เป็นพารามิเตอร์ที่ใช้ควบคุมฟังก์ชัน

สำหรับงานวิจัยนี้ได้ประยุกต์วิธีการจำแนกประเภทข้อมูลแบบSVM มาใช้ในการหาขอบเขตของอาการ และขอบเขตของวิธีการรักษาว่าอาการเหล่านั้นหรือวิธีการรักษาได้ถูกกล่าวสิ้นสุดหรือยัง โดยฟีเจอร์ที่ใช้ (ซึ่งก็คือเวคเตอร์คุณลักษณะ, x) คืออาการต่าง ๆที่แทนด้วย $v_{co1}w_{co1}$ หรือวิธีการรักษาต่าง ๆที่แทนด้วย $v_{co2}w_{co2}$ (ดังแสดงในหัวข้อ “กรรมวิธีดำเนินงาน”)

งานวิจัยก่อนหน้า

ได้มีงานวิจัยมากมายที่ได้เสนอเทคนิคต่าง ๆเพื่อหาความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา โดยเฉพาะความสัมพันธ์อาการ–วิธีการรักษาจากเอกสารโดเมนการดูแลสุขภาพ (Health Care Domain) โดยแบ่งออกเป็น 2 แนวทางคือ แนวทางรูปแบบหรือกฎ (Pattern/Rule Based Approach) และแนวทางสถิติรวมถึงการเรียนรู้ของเครื่อง (Statistical Based Approach Including Machine Learning)

1. แนวทางรูปแบบหรือกฎ (Pattern/Rule Based Approach)

Rosario B.(2005) ได้สกัดความสัมพันธ์เกี่ยวกับความหมาย (Semantic Relation) จากเอกสารวิทยาศาสตร์ทางชีวะภาพ โดยความสัมพันธ์นี้เป็นความสัมพันธ์ตามไวยกรณ์ภาษาอังกฤษซึ่งเป็นความสัมพันธ์ระหว่างสองเอนทิตี้ที่เป็นคำนามในหนึ่งประโยค ฉะนั้นความสัมพันธ์ที่สกัดได้คือ DIS and TREAT (Disease and Treatment) ดังประโยคตัวอย่างต่อไปนี้ “Therefore administration of TJ-135 may be useful in patients with severe acute hepatitis ac-companying cholestasis or in those with autoimmune hepatitis.” มีคำนาม “TJ-135” และคำนาม “hepatitis” เป็นความสัมพันธ์แบบ DIS and TREAT โดย “hepatitis” คือ DIS และ“TJ-135” คือ TREAT Rosario B.(2005) ใช้ MeSH เป็นฐานความรู้สำหรับหาแนวความคิดของ DISและTREAT และใช้ตัวแบบกราฟ ไวยกรณ์ที่รวมความหมายเข้าไว้ด้วย มาเป็นตัวระบุความสัมพันธ์ระหว่างคำนามสองเอนทิตี้ ผลงานวิจัยของ Rosario ได้ความถูกต้อง 96.9% เมื่อมีเอนทิตี้ทั้งสองปรากฏ และ79.6% เมื่อบางเอนทิตี้ไม่ปรากฏ

Abacha A. B. and Zweigenbaum P. (2011) ได้สกัดความสัมพันธ์ ทางการรักษา (Treatment Relation) ซึ่งเป็นความสันธ์ทางความหมาย (บนพื้นฐานUMLS) ระหว่างวิธีการรักษาและโรค โดยใช้รูปแบบทางภาษา (Linguistic Pattern) มาทำการสกัดความสัมพันธ์นี้ที่เกิดขึ้นในหนึ่งประโยค

Linguistic Pattern: … E1 … be effective for L1… | … E1 was found to reduce E2 …

where E1, E2, or Ei is the medical entity and L1 isชื่อโรคหรืออาการ

ตัวอย่างเช่น “Fosfomycin (E1) and amoxicillin-clavulanate (E2) appear to be effective for cystitis (L1) caused by susceptible isolates”ผลการทดลองได้ 75.72% precisionและ 60.46% recall.

2. แนวทางสถิติรวมถึงการเรียนรู้ของเครื่อง (Statistical Based Approach Including Machine Learning)

Song S. et al, (2011)ได้สกัดความรู้ทางกรรมวิธี(Procedural Knowledge)จากบทคัดย่อของเมดลาย(MEDLINE abstract)โดยใช้SVM(Support Vector Machine) และ CRF (Conditional Random Field) ร่วมกับการประมวลภาษาธรรมชาติ“….[In a total gastrostomy](Target), [clamps are placed on the end of the esophagus and the end of the small intestine](P1). [The stomach is removed] (P2) and [the esophagus is joined to the intestine] (P3).…”เมื่อP1, P2, และP3 เป็นกรรมวิธีการแก้ปัญหา พวกเขาได้นิยามความรู้ทางกรรมวิธี ว่า เป็นการรวมกันของ เป้าประสงค์(Target) และวิธีแก้ปัญหาที่สอดคล้องกันกับเป้าประสงค์ ความรู้ทางกรรมวิธีประกอบด้วยหลายขั้นตอน SVM และCRF ถูกใช้ร่วมกับ 4 ฟีเจอร์: content feature (หลังจากหาแก่นคำ (word stemming) และกำจัดคำหยุด ทั้งunigram และbigrams ในประโยคที่เป็นประโยคเป้าประสงค์), position feature, neighbor feature, และ ontological feature เพื่อที่จะจำแนกประเภทเป้าประสงค์ นอกจากนี้มีฟีเจอร์อื่น ๆเช่น word feature, context feature, Predicate-argument structure, และontological feature สำหรับจำแนกประเภทกรรมวิธีที่ประกอบด้วยหลาย ๆประโยค ผลงานวิจัยของพวกเขามีความถูกต้องเป็นค่าแม่นยำ (Precision) 0.7279 และ 0.8369 ของ CRFและSVM ตามลำดับ และค่าระลึก (Recall) 0.7326, 0.7957 ของ CRF และ SVM ตามลำดับ

Yeleswarapu et al., (2014) ได้ประยุกต์การตรวจหาและการสกัดแบบไปป์ไลน์กึ่งอัตโนมัติ (Semi-Automatic Pipeline) กับคู่ระหว่างเหตุการณ์ไม่พึงประสงค์จากการใช้ยา (drug-Adverse Event, drug-AE) จากข้อมูลที่ไม่เป็นโครงสร้าง เช่น เอกสารการวิจารณ์ของผู้ใช้บนบล็อค (Blogs) และบทคัดย่อของเมดลาย (MED-LINE Abstract) โดยคำต่าง ๆที่เกี่ยวกับยา โรค และ อาการ อยู่ในรูปนามวลี ซึ่งรวมถึงเนมเอนติตี้ (Name Entity) ด้วย และมีความสัมพันธ์กันภายในหนึ่งประโยค พวกเขาหาค่าส่วนประกอบสารสนเทศ (Information Component, IC) ด้วย Bayesian Confidence Propagation Neural Network โดยค่า IC ใช้วัดความไม่ได้สัดส่วนของคู่เหตุการณ์ไม่พึงประสงค์จากการใช้ยา ส่วนค่าความเบี่ยงเบนมาตรฐานของแต่ละ IC ใช้กำหนดค่าความแข็งแกร่ง (Robustness) ของค่า IC ฉะนั้นค่า IC สามารถแสดงถึงความแข็งแรงของการขึ้นต่อกันระหว่างการใช้ยาและเหตุการณ์ไม่พึงประสงค์จากการใช้ยา ถ้าค่า IC เป็น 0 แสดงว่าไม่มีการขึ้นต่อกันระหว่างการใช้ยาและเหตุการณ์ไม่พึงประสงค์จากการใช้ยา ถ้าค่า IC เพิ่มขึ้นเป็นบวก แสดงว่าการเชื่อมโยงทางบวกระหว่างเการใช้ยา

และเหตุการณ์ไม่พึงประสงค์นั้นเพิ่มขึ้น นั่นก็คือแต่ละคู่ drug-AE ที่สกัดได้สามารถบอกเป็นนัยของความสัมพันธ์ระหว่างการใช้ยาและเหตุการณ์ไม่พึงประสงค์จากการใช้ยา

ปัญหาการหาความสัมพันธ์ระหว่างปัญหาและวิธีการแก้ปัญหาจากเอกสารภาษาไทยบนเว็บบอร์ด

ปัญหาสำหรับงานวิจัยนี้ (การหาความสัมพันธ์ปัญหา-วิธีการแก้ปัญหาโดยเฉพาะความสัมพันธ์อาการ-วิธีการรักษา จากเอกสารภาษาไทยบนเว็บบอร์ด) ประกอบด้วยสามปัญหาหลักคือ

ก. ปัญหาเกี่ยวกับการระบุ EDU ที่มีแนวความคิดอาการ (EDU_{sym}) และระบุ EDU ที่มีแนวความคิดวิธีการรักษา (EDU_{treat}) จากเอกสารภาษาไทยนั้นในโดเมนการดูแลสุขภาพ

จากการศึกษาพฤติกรรมคลังข้อมูล (corpus behavior study) พบว่าแนวความคิดอาการและแนวความคิดวิธีการรักษา ของ EDU_{sym} และ EDU_{treat} ตามลำดับ แสดงออกมาในรูปของกริยาวลี ตัวอย่างเช่น

แนวความคิดอาการ (Symptom Concept)

a) EDU: "ผู้ป่วยรู้สึกเวียนศีรษะ" ; "ผู้ป่วย/A patient รู้สึก/feels เวียนศีรษะ/dizzy"
(A patient feels dizzy.)

b) EDU: "ฉัน[มีอาการ]ปวดศีรษะ"; "ฉัน/I [มีอาการ/have symptom] ปวดศีรษะ/headache"
(I have a headache symptom.)
เมื่อสัญลักษณ์[..]แทนการละคำหรือวลี

แนวความคิดวิธีการรักษา (Treatment Concept)

c) EDU: "กินยาลดกรด" ; "กิน/consume ยาลดกรด/antacid"
(Take an antacid.)

อย่างไรก็ตามบางครั้งเกิดความกำกวมของความหมายของกริยาวลีที่แสดงแนวความคิดอาการ ตัวอย่างเช่น

e) EDU: "[คนไข้]ถ่ายยาก" ; "[คนไข้/patient] ถ่าย/defecate ยาก /difficultly"
([A patient] defecates difficultly.)

f) EDU1: "ห้องน้ำสกปรกมาก" ; "ห้องน้ำ/toilet สกปรกมาก/is very dirty."
(A toilet is very dirty.)
EDU2: ฉันจึงถ่ายยาก" ; "ฉัน/I จึง/then ถ่าย/defecate ยาก /difficultly"
(Then, I defecate difficultly.)

จากตัวอย่าง e) และ f) กริยาวลีที่แสดงแนวความคิดอาการคือ ตัวอย่าง e) ด้วยแนวความคิด ‘ท้องผูก/be constipated’

เนื่องจากอาการ/ปัญหาและวิธีการรักษา/วิธีการแก้ปัญหาในเอกสารโดเมนการดูแลสุขภาพสำหรับงานวิจัยนี้ แสดงอยู่ในรูปของเหตุการณ์ (Event) ที่แสดงด้วยกริยาวลี (Verb Phrase) ดังนั้นงานวิจัยนี้จึงได้ดำเนินการทดสอบหาคู่คำ ด้วย ค่าความสัมพันธ์ r (relatedness) (Guthrie et al., 1991) จากสองคำที่อยู่ติดกันหลังจากได้กำจัดคำหยุดออกไป และมีคำแรกเป็นคำกริยา ในแต่ละ EDU ของโดเมนการดูแลสุขภาพ ว่ามีแนวความคิดอาการสำหรับ Word-Co_{sym} ($v_{co1}w_{co1}$) / แนวความคิดวิธีการรักษาสำหรับ Word-Co_{treat} ($v_{co2}w_{co2}$) ทั้งนี้เพื่อการประยุกต์ใช้ Word-Co_{sym} ($v_{co1}\in V_{co1}$; $w_{co1}\in W_{co1}$) ซึ่งใช้แทน EDU_{sym} และ Word-Co_{treat} ($v_{co2}\in V_{co2}$; $w_{co2}\in W_{co2}$) ซึ่งใช้แทน EDU_{treat} (ตามที่ได้กล่าวในบทนำ) มาระบุ EDU_{sym} และ EDU_{treat} ตามลำดับ

ข. ปัญหาเกี่ยวกับการหาขอบเขต EDU ที่มีแนวความคิดอาการ (Dsym) และขอบเขต EDU ที่มีแนวความคิดวิธีการรักษา (AT/RT)

จากรูปที่ 1 ไม่มีคำคลู (Clue) เช่น (i.e. ‘และ/and’, ‘หรือ/or’,..) ใน EDU3 เพื่อบอกขอบเขตการสิ้นสุดของแนวความคิดอาการ และ EDU10 เพื่อบอกขอบเขตการสิ้นสุดของแนวความคิดวิธีการรักษา ฉะนั้นหลังจากได้ทำการระบุ EDU_{sym} หรือ EDU_{treat} แล้วเราก็จะทำการประยุกต์ใช้ NB และSVM (ซึ่งเป็นเทคนิคการเรียนรู้เชิงเส้นของเครื่องสำหรับการจำแนกประเภท) เพื่อหาขอบเขต EDUs ของ Dsym และขอบเขต EDUs ของ AT/RT ด้วยคู่ Word-Co ($v_{coj\text{-}i}w_{coj\text{-}i}$ $v_{coj\text{-}i+1}w_{coj\text{-}i+1}$ สำหรับ j= {1,2} i=1,2,..,endOfboundary) ที่ได้จากการเลื่อนคู่ EDU ที่อยู่ติดกัน (EDUi EDUi+1) ไปด้วยระยะทางหนึ่ง EDU

ค. ปัญหาเกี่ยวกับการระบุความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา

โดยเฉพาะความสัมพันธ์อาการ-วิธีการรักษา จากแต่ละคู่ของอีดียูเวคเตอร์แพร์ (EDU Vector Pair) ที่ประกอบด้วยเวคเตอร์ของ EDUs ทีมีแนวความคิดอาการ และเวคเตอร์ของ EDUs ที่มีแนวความคิดวิธีการรักษา ($\langle EDU_{sym\text{-}1}\ EDU_{sym\text{-}2}\ldots EDU_{sym\text{-}m}\rangle\langle EDU_{treat\text{-}1}EDU_{treat\text{-}2}\ldots EDU_{treat\text{-}n}\rangle$)

เนื่องจากความสัมพันธ์อาการ-วิธีการรักษา แปรเปลี่ยนตามผู้ป่วย สภาพแวดล้อม เวลา และ อื่นๆ แม้ว่าจะเป็นโรคเดียวกันก็ตาม ตัวอย่างเช่น

(a) EDU1symtom: “[ผู้ป่วย]ปวดท้อง/have a stomachache อย่างหนัก/heavily”
([A patient] has a stomachache heavily.)

EDU2symtom: “[ผู้ป่วย] มี/has แก๊ส/gas ในกระเพาะ/inside stomach มาก/a lots”

([The patient] has lots of gas in the stomach)

EDU3 treatment: “ [ผู้ป่วย] กิน/takes ยาลดกรด/an antacid”

([The patient] takes an antacid)

EDU4: “แต่/But ก็ไม่หายปวด/it cannot work”

(But it cannot work)

(b) EDU1symtom: “[ผู้ป่วย] ปวดท้อง/have a stomachache”

([A patient] has a stomachache.)

EDU2symtom: “[ผู้ป่วย] มี/has แก๊ส/gas ในกระเพาะ/inside stomach”

([The patient] has gassy in the stomach)

EDU3 treatment “[ผู้ป่วย] กินยาลดกรด//take an antacid”

([The patient] takes an antacid)

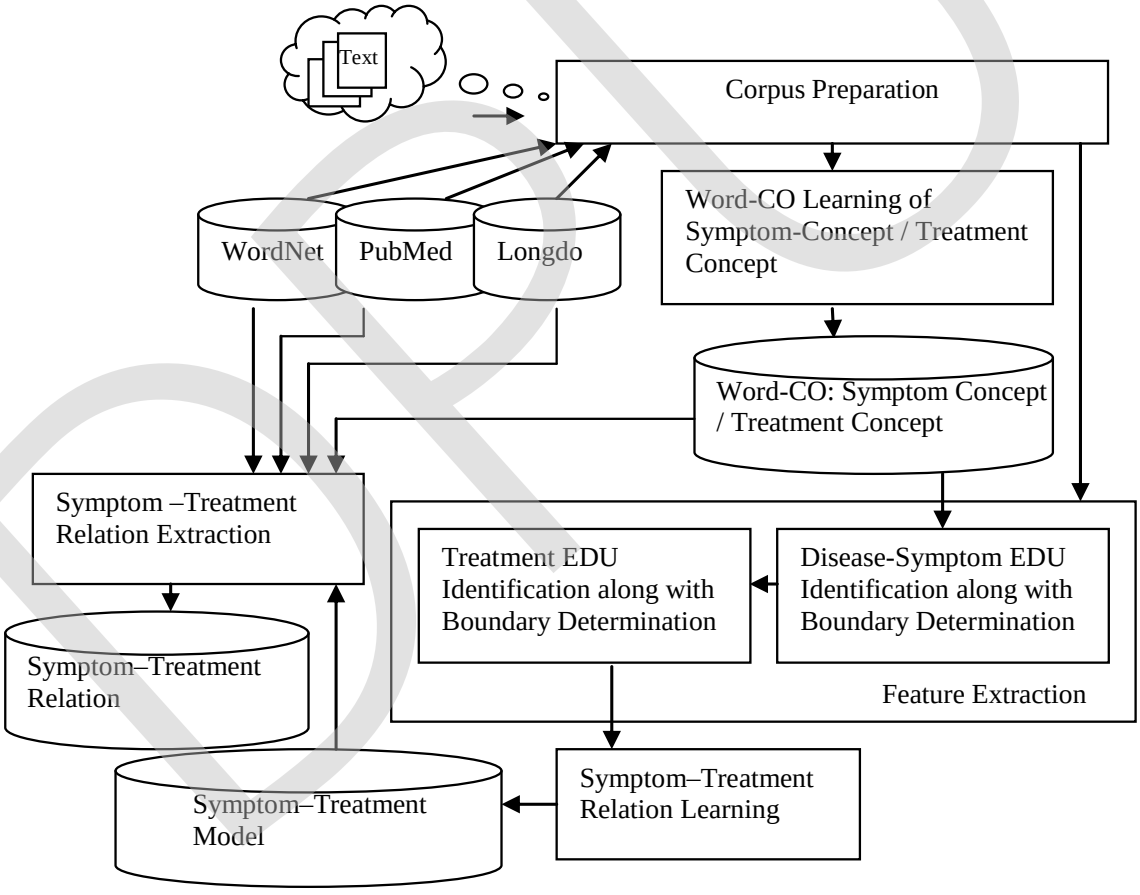
EDU4: “[ผู้ป่วย] รู้สึกดีขึ้น/ Feel better”

(The patient] feels better)

จากตัวอย่าง (a) และ (b) ความสัมพันธ์อาการ-วิธีการรักษา หรือความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา ปรากฏกับตัวอย่าง (b) เท่านั้น เพราะ EDU4 ของตัวอย่าง (b) มีกริยาวลี “รู้สึกดีขึ้น/ feel better” ซึ่งเป็น คลาสคิวเวิรด์ (Class-cue-word) ของความสัมพันธ์อาการ-วิธีการรักษา ดังนั้นงานวิจัยนี้ขอเสนอการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องด้วย Naïve Bayes มาทำการเรียนรู้ความสัมพันธ์อาการ-วิธีการรักษา จากคู่เวคเตอร์ EDU (EDU Vector Pair) ทั้งหมด ซึ่งคู่เวคเตอร์ EDU สามารถถูกแทนด้วย คู่เวคเตอร์คู่คำ (Word-Co Vector Pair) แต่ละคู่เวคเตอร์คู่คำประกอบด้วย เวคเตอร์คู่คำที่มีแนวความคิดอาการ (Word-Co_{sym} Vector) และเวคเตอร์คู่คำที่มีแนวความคิดวิธีการรักษา (Word-Co_{treat} Vector) จากปัญหาทั้ง 3 ที่ได้กล่าวข้างต้น งานวิจัยนี้ได้แสดงรายละเอียดกรรมวิธีการแก้ไขปัญหาดังกล่าวในหัวข้อถัดไปคือ “กรรมวิธีดำเนินงาน”

กรรมวิธีดำเนินงาน

ระบบงานโดยสรุปสำหรับการหาความสัมพันธ์อาการ-วิธีการรักษา (ความสัมพันธ์ระหว่างปัญหาและวิธีการแก้ปัญหา) จากเอกสารภาษาไทยบนเว็บบอร์ดประกอบด้วยขั้นตอนต่าง ๆ ดังต่อไปนี้ (ดังแสดงในรูปที่ 2) ขั้นตอนการเตรียมคลังข้อมูล (Corpus Preparation) ขั้นตอนการเรียนรู้แนวความคิดคู่คำ (Word-Co Concept Learning) ขั้นตอนการสกัดฟีเจอร์ (Feature Extraction) ขั้นตอนการเรียนรู้ความสัมพันธ์อาการ-วิธีการรักษา (Symptom-Treatment Relation Learning) และขั้นตอนการสกัดความสัมพันธ์อาการ-วิธีการรักษา (Symptom-Treatment Relation Extraction)



รูปที่ 2 ระบบงานโดยสรุป

1. ขั้นตอนการเตรียมคลังข้อมูล (Corpus Preparation)

ขั้นตอนนี้เป็นการเตรียมคลังข้อมูลภาษาไทยในโดเมนการดูแลสุขภาพและการดูแลรักษา (โดยเฉพาะโรคทางเดินอาหาร โรคในเด็ก และโรคทางสมองและหัวใจ) ภายใต้เว็บบอร์ดโรงพยาบาลที่เป็นเอนจีโอ (NGO, Non-Government-Organization) สามารถดาวโหลดได้จากเวปไซต์ (http://haamor.com/) เป็นจำนวนทั้งหมด 6000 EDUs ซึ่งถูกแบ่งออกเป็นสองส่วน คือ ส่วนที่ 1 จำนวน 4500 EDUs สำหรับการเรียนรู้แนวความคิดคู่คำ การเรียนรู้ขอบเขตของ Dsym และขอบเขตของ AT/RT และการเรียนรู้ ความสัมพันธ์อาการ-วิธีการรักษา ภายใต้ 10 Folds Cross Validation ส่วนที่ 2 จำนวน 1500 EDUs สำหรับการทดสอบหาขอบเขตของ Dsym และขอบเขตของ AT/RT และการทดสอบการสกัดความสัมพันธ์อาการ-วิธีการรักษา นอกจากนี้ขั้นตอนนี้รวมถึงขั้นตอนการตัดคำโดยการใช้ซอฟท์แวร์ตัดคำภาษาไทยที่สามารถแก้ปัญหาขอบเขตคำและขณะเดียวกันสามารถกำกับหน้าที่ของคำ (Part of Speech) ได้ (Sudprasert and Kawtrakul, 2003) ซึ่งรวมถึงการหา Name Entity (Chanlekha and Kawtrakul, 2004), และการรับรู้คำ (Word-Formation Recognition) (Pengphon et al., 2002) เพื่อที่จะแก้ปัญหาขอบเขตของ Thai Name Entity และนามวลี หลังจากนั้นทำการตัดประโยคในระดับ EDU ด้วยวิธีการของ (Chareonsuk et al., 2005) และสุดท้ายทำการกำกับคู่คำที่มี แนวความคิดอาการ และคู่คำที่มีแนวความคิดวิธีการรักษา พร้อมกับการกำกับ Class-cue-word Tag เพื่อระบุคิวเวิร์ด (Cue Word) ด้วยเซตประเภทคลาส (Class-type set {“yes”, “no”}) สำหรับความสัมพันธ์อาการ-วิธีการรักษา ดังแสดงในรูปที่ 3 นอกจากนี้แนวความคิดคู่คำทั้งหมดอ้างอิงจาก Wordnet (http:// word-net.princeton. edu/obtain) และ MeSH หลังจากแปลไทยเป็นอังกฤษโดยใช้ Lexitron (the Thai-English dictionary) (http://lexitron.nectec.or.th/).

ปวดท้อง/ Stomachache

<**Symptom Boundary**>

<EDU>*น้องชาย*<**verb-co:** class=symptom ; concept='has a symptom'>*มีอาการ*</verb-co><**word-co**: class=symptom ; concept='stomachache'>*ปวดท้อง*</word-co>*อย่างหนัก* </EDU>

(A user brother has a stomachache heavily.)

<EDU>[*น้องชาย*]<**verb-co**: class=symptom ; concept='has'>*มี* </verb-co><**word-co**: class=symptom ; concept='gassy'>*แก๊ส*</word-co>*ในกระเพาะมาก* </EDU>

([The user brother] has lots of gas in the stomach.)

<EDU>[*น้องชาย*]*มักจะ*<**verb-co**: class=symptom ; concept='has a symptom'>*มีอาการ*</verb-co><**word-co**: class=symptom ; concept= 'appearance'> *เป็น*</word-co>*ตอนหลังทานข้าวเย็นและตอนกลางคืน* </EDU> **([The user brother] mostlyhas symptoms occurring after having dinner and at night.)**

</**Symptom Boundary** >

<EDU>[*ผู้ใช้*]*สงสัยเป็นโรคกระเพาะ* </EDU> **([The user] doubts to get a gastropathy.)**

<**Treatment Boundary** >

<EDU>*เลย*<**verb-co**: class=treatment; concept='consume'>*กิน*</verb-co><**word-co**: class=treatment ;concept='antacid'>*ยาลดกรด*</word-co>*เพื่อแก้ปวดท้อง* </EDU> (**Then [The user brother] takes an antacid to solve the stomach ache.)**

</**Treatment Boundary**>

<EDU>*แต่ก็*<**Class-cue-word: class**=no>*ไม่หาย*</Class-cue-word>*ปวด* </EDU> **(But it cannot work.)**
<EDU>*และปวดเพิ่มขึ้น*</EDU> **(And, [The user brother] has more pain)**

รูปที่ 3. การกำกับความสัมพันธ์อาการ-วิธีการรักษา (ความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา)

2. ขั้นตอนการเรียนรู้แนวความคิดคู่คำ (Word-Co Concept Learning)

จากงานวิจัยของ Guthrie et al. (1991) ค่าความสัมพันธ์ r (relatedness) ได้ถูกนำมาประยุกต์ใช้ในงานวิจัยนี้สำหรับหาค่าความสัมพันธ์ของคู่คำด้วยแนวความคิดอาการ ($v_{co1}w_{co1}$) หรือแนวความคิดวิธีการรักษา ($v_{co2}w_{co2}$) จากคำสองคำที่อยู่ติดกันหลังจากได้กำจัดคำหยุดออกไป และมีคำแรกเป็นคำกริยา ดังสมการที่ (7) ซึ่งในแต่ละคู่คำ หรือ Word-Co ($v_{coj}\, w_{coj}$) ที่อยู่ในแต่ละ EDU ประกอบด้วยค่า r สองค่าคือค่าใช่แนวความคิดอาการ/แนวความคิดวิธีการรักษา และ ค่าไม่ใช่แนวความคิดอาการ/แนวความคิดวิธีการรักษา เช่น ถ้า $v_{coj} = v_{co1}$ จะมี ค่า r สองค่าคือค่าใช่แนวความคิดอาการ และค่าไม่ใช่แนวความคิดอาการ ถ้า $v_{coj} = v_{co2}$ จะมีค่า r สองค่าคือค่าใช่แนวความคิดวิธีการรักษา และค่าไม่ใช่แนวความคิดวิธีการรักษา ฉะนั้นค่า r($v_{coj}\ w_{coj}$) ของแต่ละ $v_{coj}\ w_{coj}$ ที่มีค่าใช่แนวความคิดอาการ/แนวความคิดวิธีการรักษา มากกว่าค่าไม่ใช่แนวความคิดอาการ/แนวความคิดวิธีการรักษา จะถูกเลือกเก็บสะสมเป็นสมาชิกของเซต $VW_{symptom}$ หรือ $VW_{treatment}$ ตามลำดับ ($v_{co1}w_{co1} \in VW_{symptom}$ เมื่อ $VW_{symptom}$ เป็นเซตของ Word-Co ที่มีแนวความคิดอาการ และ $v_{co2}w_{co2} \in VW_{treatment}$ เมื่อ $VW_{treatment}$ เป็นเซตของ

Word-Co ที่มีแนวความคิดวิธีการรักษา) ดังนั้น $VW_{symptom}$ และ $VW_{treatment}$ ใช้สำหรับการระบุ EDU_{sym} และ EDU_{treat} ตามลำดับ

$$r(v_{coj}, w_{coj}) = \frac{fv_{coj} w_{coj}}{fv_{coj} + fw_{coj} - fv_{coj} w_{coj}} . \quad (7)$$

where $r(v_{coj}, w_{coj})$ is the relatedness of Word – CO with

a symptom concept if $coj = co1$ or a treatment concept if $coj = co2$.

$v_{coj} \in V_{coj}$, $w_{coj} \in W_{coj}$ V_{co1} is a set of verbs with the symptom concepts.

V_{co2} is a set of verbs with the treatment concepts.

W_{co1} is the co – occurred word set having the symptom concept in the $v_{co1} w_{co1}$ co – occurrence .

W_{co2} is the co – occurred word set having the treatment concept in the $v_{co2} w_{co2}$ co – occurrence .

fv_{coj} is the numbers of v_{coj} occurences. fw_{coj} is the numbers of w_{coj} occurences.

$fv_{coj} w_{coj}$ is the numbers of v_{coj} and w_{coj} occurences.

3. ขั้นตอนการสกัดฟีเจอร์ (Feature Extraction)

ขั้นตอนนี้เป็นการสกัดกลุ่มฟีเจอร์สองกลุ่มหลักในรูปของเวคเตอร์ กลุ่มแรกคือ เวคเตอร์คู่คำที่มีแนวความคิดอาการ (Word-Co_{sym} Vector) กลุ่มที่สองคือ เวคเตอร์คู่คำที่มีแนวความคิดวิธีการรักษา (Word-Co_{treat} Vector) โดยเวคเตอร์คู่คำที่มีแนวความคิดอาการแทนเวคเตอร์ EDU_{sym} (Dsym) และ เวคเตอร์คู่คำที่มีแนวความคิดวิธีการรักษาแทนเวคเตอร์ EDU_{treat} (AT/RT) ทั้งนี้เพื่อใช้ในการหาความพันธ์อาการ-วิธีการรักษา ฉะนั้นขั้นตอนนี้จึงเป็นการหาขอบเขตของ Dsym และ ขอบเขตของAT/RT หลังจากการระบุ EDU_{sym} และ EDU_{treat} ตามลำดับ ดังนั้นงานวิจัยนี้ได้ทำการเรียนรู้ขอบเขตของ Dsym และ ขอบเขตของ AT/RT ด้วยเทคนิคที่แตกต่างกันสองแบบเพื่อเปรียบเทียบกัน คือ SVM และ NB กับคลังข้อมูลส่วนที่ใช้สำหรับขั้นตอนการเรียนรู้ความสัมพันธ์อาการ-วิธีการรักษา

SVM : เป็นเทคนิคการเรียนรู้เชิงเส้นของเครื่องสำหรับการจำแนกประเภท งานวิจัยนี้ได้ทำการเรียนรู้ ขอบเขตของ Dsym และ AT/RT ด้วย SVM ดังแสดงในสมการ(6) ซึ่งแทน x_i ด้วยคู่Word-Co (คือ $v_{coj-i}w_{coj-i}$ $v_{coj-i+1}w_{coj-i+1}$ สำหรับ j={1,2} i=1,2,..,endOfboundary) และฟีเจอร์ x ในสมการ (6) เป็นเวคเตอร์ของคู่ Word-Co โดย ถ้า j=1 $v_{co1}w_{co1}$แทน EDU_{sym} และ ถ้า j=2 $v_{co2}w_{co2}$แทน EDU_{treat} ดังนั้นการเรียนรู้ขอบเขตของ Dsym และ AT/RT โดยจากการเลื่อนคู่ EDU ที่อยู่ติดกัน (EDUiEDUi+1) ไปด้วยระยะทางหนึ่ง EDU จนกระทั่งค่า f(x)เปลี่ยนจากคลาสบวกเป็นคลาสลบจึงหยุดการการเลื่อนคู่ EDU

หลังจากการเรียนรู้ขอบเขตของ Dsym และ ขอบเขตของ AT/RTด้วย SVMทำให้ได้ ค่า w_i ของ w จาก $v_{co1\text{-}i}w_{co1\text{-}i}$ และค่า w_i ของ w จาก $v_{co2\text{-}i}w_{co2\text{-}i}$ ดังแสดงในตารางที่1 ทั้งนี้เพื่อนำไปใช้หาขอบเขตของ Dsym และ ขอบเขตของ AT/RT ดังแสดงใน อัลกอริทึมการหาขอบเขตรูปที่ 4.

ตารางที่1. แสดงฟีเจอร์ $v_{co1\text{-}i}w_{co1\text{-}i}$, $v_{co2\text{-}i}$ $w_{co2\text{-}i}$ และ ค่า w_i จากการเรียนรู้ด้วยSVM

$v_{co1\text{-}i}w_{co1\text{-}i}$		w ของ $v_{co1}w_{co1}$ทั้งหมด
Surface form	Concept	
Pain-stomach	haveStomachache	1.2053
HaveSymptom-distension	haveFlatulence	0.2528
Defecate-difficultly	haveConstipation	-0.1922
HaveGas-stomach	haveColic	0.3993
Occur-symptom	haveSymptom	-1.1924
	……..	…….
$v_{co2\text{-}i}w_{co2\text{-}i}$		w ของ $v_{co2}w_{co2}$ทั้งหมด
Surface form	Concept	
Consume-yogurt	haveFoodControl	-0.1281
Consume-antacid	consumeAntacidMed	-0.4005
See-physician	visitPhysician	1.5496
Drink-water	haveWaterControl	0.2517
Exercise-null	Relax	0.2051
	……..	…….

Assume that each EDU is represented by Word-Co ($v_{coj}\, w_{coj}$) . L is a list of EDU. VW_j is the Word-Co concept set.
If j=1 then VW_j is $VW_{symptom}$ which is the Word-Co_{sym} set.
If j=2 then VW_j is $VW_{treatment}$ which is the Word-Co_{treat} set.
BOUNDARY_DETERMINATION (L,j)

```
1  i ← 1, BOUNDARYj ← ∅, bd= 'end'
2  while i ≤ length[L]∧ bd=end do
3  begin_while1
4   If v_coj-i w_coj-i ∈VW_j /*if j=1 start Dsym , if j=2 start AT/RT
5     bd= 'continue' ; BOUNDARYj ← BOUNDARYj ∪ {i}
     Else  i=i+1
     end_while1
7     While ( i ≤ length[L])∧(v_coj-i w_coj-i ∈VW_j )∧( v_coj-i+1 w_coj-i+1 ∈VW_j)∧
           bd='continue' do

8     begin_while2        /* Boundary determination
        Case SVM:
           Equation(6) /* bd= 'end' or Positive Class If f(x) ≥ 0 otherwise bd=
                                                    'continue'
        Case NB:
           Equation(4) ; bd=BoundaryClass
9        End Case
10          If  bd = 'continue'
11           BOUNDARYj ← BOUNDARYj ∪ {i+1}; i=i+1
13    end_while2
17  Return
```

รูปที่ 4. อัลกอริทึมการหาขอบเขต Boundary Determination Algorithm

NB: การหาขอบเขตของ Dsym และ ขอบเขตของ AT/RT สามารถคำนวณได้โดยการหาค่า BoundaryClass จากสมการ NB (สมการ(4)) กับค่าความน่าจะเป็น ที่ได้จากการเรียนรู้ของเครื่องโดยการเลื่อนคู่ EDU ที่อยู่ติดกัน (EDUi EDUi+1) ไปด้วยระยะทางหนึ่ง EDU ดังแสดงในตารางที่ 2ของ $v_{co1\text{-}i}w_{co1\text{-}i}$, $v_{co1\text{-}i+1}w_{co1\text{-}i+1}$ สำหรับ Dsym และของ $v_{co2\text{-}i}w_{co2\text{-}i}$ และ $v_{co2\text{-}i+1}w_{co2\text{-}i+1}$ สำหรับ AT/RT ดังนั้นการหาขอบเขตหลังจากเรียนรู้ได้โดยการเลื่อนคู่ EDU ที่อยู่ติดกัน (EDUi EDUi+1) ไปด้วยระยะทางหนึ่ง EDU พร้อมกับค่าความน่าจะเป็นที่ได้จากการเรียนรู้ของเครื่องนั้น จนกระทั่งค่า BoundaryClass เป็น 'end' ดังแสดงในอัลกอริทึมการหาขอบเขตรูปที่ 4

ตารางที่ 2. แสดงค่าความน่าจะเป็นของ Word-Co$_{sym}$ จาก Word-Co$_{sym}$ Pair ($v_{co1\text{-}i}w_{co1\text{-}i}$, $v_{co1\text{-}i+1}w_{co1\text{-}i+1}$) และ Word-Co$_{treat}$ จากWord-Co$_{treat}$ Pair ($v_{co2\text{-}i}w_{co2\text{-}i}$, $v_{co2\text{-}i+1}w_{co2\text{-}i+1}$) ที่มีความคิดของแต่ละคู่คำเป็นสิ้นสุดขอบเขตและไม่สิ้นสุดขอบเขต

$v_{co1\text{-}i}w_{co1\text{-}i}$ (concept)	End-of-Dsym	Continue-of-Dsym	$v_{co1\text{-}i+1}w_{co1\text{-}i+1}$ (concept)	End-of-Dsym	Continue-of-Dsym
haveConstipation	0.4110	0.1731	haveConstipation	0.375	0.1091
haveFlatulence	0.1507	0.1346	haveFlatulence	0.1447	0.0273
haveStomachache	0.0069	0.0385	haveStomachache	0.0197	0.0091
haveColic	0.1367	0.0096	haveColic	0.0132	0.0182
be-drug	0.0137	0.0385	be-drug	0.0263	0.0091
…	…	…	…	…	…
$v_{co2\text{-}i}w_{co2\text{-}i}$ (concept)	End-of-AT/RT	Continue-of- AT/RT	$v_{co2\text{-}i+1}w_{co2\text{-}i+1}$ (concept)	End-of-AT/RT	Continue-of-AT/RT
haveFoodControl	0.385	0.1101	haveFoodControl	0.321	0.1041
consumeAntacidMed	0.1547	0.0283	consumeAntacidMed	0.1607	0.1356
visitPhysician	0.0297	0.0101	visitPhysician	0.0169	0.0395
haveWaterControl	0.0232	0.0192	haveWaterControl	0.1467	0.0106
Relax	0.0363	0.0101	Relax	0.0237	0.0395
…	…	…	…	…	…

4.ขั้นตอนการเรียนรู้ความสัมพันธ์อาการ-วิธีการรักษา (Symptom-Treatment Relation Learning)

ขั้นตอนนี้เป็นการเรียนรู้ความสัมพันธ์อาการ-วิธีการรักษา (ความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา) จากคู่เวคเตอร์ที่สกัดได้ในข้อก่อนหน้า คือ Dsym และตามด้วย AT/RT ดังนี้

Dsym–AT = $\langle v_{co1\text{-}1}\, w_{co1\text{-}1}$, $v_{co1\text{-}2}\, w_{co1\text{-}2}$,.., $v_{co1\text{-}a}\, w_{co1\text{-}a}\, v_{co2\text{-}1}\, w_{co2\text{-}1}$, $v_{co2\text{-}2}\, w_{co2\text{-}2}$,.., $v_{co2\text{-}b}w_{co2\text{-}b}\rangle$

Dsym–RT = $\langle\, v_{co1\text{-}1}\, w_{co1\text{-}1}$, $v_{co1\text{-}2}\, w_{co1\text{-}2}$,.., $v_{co1\text{-}a}\, w_{co1\text{-}a}\, v_{co2\text{-}1}\, w_{co2\text{-}1}$, $v_{co2\text{-}2}\, w_{co2\text{-}2}$,.., $v_{co2\text{-}c}w_{co2\text{-}c}\rangle$

พร้อมระบุประเภทหรือคลาสของความสัมพันธ์ของคู่เวคเตอร์ดังกล่าวภายใน 5 EDUs หลัง AT/RT ว่าใช่ความสัมพันธ์ดังกล่าวหรือไม่ ภายใต้เซตประเภทคลาส (Class-type set,{‘yes’ ‘no’}) โดยใช้เซตของแพตเทิร์นคำบอกเป็นนัยของคลาส (Set of Class-cue-word Pattern) ดังนี้

Class-cue-word pattern={‘cue:หาย/disappear=class:yes’, ‘cue:รู้สึกดีขึ้น/feel better=class: yes’, ‘cue:ไม่ปวด/do not pain=class:yes’, ‘cue:“ ”=class:yes’, ‘cue:ไม่หาย/appear=class: no’,‘cue:ยังปวดอยู่/still pain=class:no’, ‘cue:ปวดมากขึ้น/have more pain=class: no’,..}

ทั้งนี้เพื่อหาค่าความน่าจะเป็นของ Word-Co_{sym} ($v_{co1}w_{co1}$) และWord-Co_{treat}($v_{co2}w_{co2}$) ด้วยคลาสความสัมพันธ์ที่ใช่ (Symptom-Treatment Relation) และไม่ใช่ (Non Symptom-Treatment Relation) ดังแสดงในตารางที่ 3 ด้วยการใช้ Weka (http://www.cs.wakato.ac.nz/ml/weka/)

ตาราที่ 3. แสดงค่าความน่าจะเป็นของ Word-Co_{sym} และ Word-Co_{treat} ที่ทำให้เกิดคลาสแบบความสัมพันธ์อาการ-วิธีการรักษา (Symptom-Treatment Relation) และไม่ใช่ความสัมพันธ์อาการ-วิธีการรักษา (Non Symptom-Treatment Relation)

v_{co1} w_{co1}	Symptom-Treatment	Non Symptom-Treatment
haveStomachache	0.39130435	0.33928571
haveConstipation	0.06086957	0.125
haveColic	0.03636364	0.01960784
haveGastricDistress	0.03448276	0.07017544
haveAnusPain	0.03418803	0.01724138
vomit	0.14529915	0.17241379
haveDiarrhea	0.11304348	0.125
…	…	…
$v_{co2}w_{co2}$	Symptom-Treatment	Non Symptom-Treatment
consumeSuppositoryMed	0.03125	0.01449275
haveEnemaTreat	0.0078125	0.10144928
haveFoodControl	0.0546875	0.01449275
consumeAntacidMed	0.06140351	0.01818182
haveWaterControl	0.03508772	0.01818182
consumeAntiFlatulenceMed	0.03703704	0.02040816
….	….	….

5.ขั้นตอนการสกัดความสัมพันธ์อาการ-วิธีการรักษา (Symptom-Treatment Relation Extraction)

ขั้นตอนนี้เป็นการรู้จำและสกัดความสัมพันธ์อาการ-วิธีการรักษา(ความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา) จากคลังข้อมูลที่ใช้ทดสอบหลังจากที่ทราบชื่อโรคจากชื่อหัวข้อเอกสารแล้ว ความสัมพันธ์อาการ-วิธีการรักษาจะถูกดำเนินการหาออกมาด้วยสามขั้นตอนคือ ขั้นตอนการหา

Dsym หลังจากระบุ EDU_{sym} ขั้นตอนการหา AT/RT หลังจากระบุ EDU_{treat} และขั้นตอนการหาความสัมพันธ์อาการ-วิธีการรักษา

5.1 ขั้นตอนการหา Dsym

หลังจากการระบุ EDU_{sym} ด้วย v_{co1} w_{co1} (v_{co1} $w_{co1} \in VW_{symptom}$) ก็เริ่มทำการหาขอบเขตของ Dsym ด้วยสองเทคนิคที่แตกต่างคือ SVM และ NB ต่อไปนี้

SVM: เป็นการหาขอบเขตของ Dsym ด้วยค่า w_i ของ $v_{co1\text{-}i}w_{co1\text{-}i}$ (x_i) ในตารางที่ 1 ซึ่งเป็นค่าที่ได้จากการเรียนรู้ SVM (สมการ (6)) โดยการเลื่อนคู่ EDU ที่อยู่ติดกัน (EDUi EDUi+1) ไปด้วยระยะทางหนึ่ง EDU จนกระทั่งค่า f(x)เปลี่ยนจากคลาสบวกเป็นคลาสลบจึงหยุดการการเลื่อนคู่ EDU ซึ่งก็คือขอบเขตของ Dsym

NB: เป็นการหาขอบเขตของ Dsym ด้วยค่าความน่าจะเป็นของ $v_{co1\text{-}i}w_{co1\text{-}i}$ และ $v_{co1\text{-}i+1}w_{co1\text{-}i+1}$ ในตารางที่ 2 กับสมการ(4) โดยการเลื่อนคู่ EDU ที่อยู่ติดกัน (EDUi EDUi+1) ไปด้วยระยะทางหนึ่ง EDU จนกระทั่งค่า BoundaryClass เป็น‘end’

5.2 ขั้นตอนการหา AT/RT

หลังจากการระบุ EDU_{sym} ด้วย v_{co2} w_{co2} (v_{co2} $w_{co2} \in VW_{symptom}$) ก็เริ่มทำการหาขอบเขตของ AT/RT ด้วยสองเทคนิคที่แตกต่างคือ SVM และ NB ต่อไปนี้

SVM: เป็นการหาขอบเขตของ AT/RT ด้วยค่า w_i ของ $v_{co2\text{-}i}w_{co2\text{-}i}$ (x_i) ในตารางที่ 1 ซึ่งเป็นค่าที่ได้จากการเรียนรู้ SVM (สมการ (6)) โดยการเลื่อนคู่ EDU ที่อยู่ติดกัน (EDUi EDUi+1) ไปด้วยระยะทางหนึ่ง EDU จนกระทั่งค่า f(x)เปลี่ยนจากคลาสบวกเป็นคลาสลบจึงหยุดการการเลื่อนคู่ EDU ซึ่งก็คือขอบเขตของ AT/RT

NB: เป็นการหาขอบเขตของ AT/RT ด้วยค่าความน่าจะเป็นของ $v_{co2\text{-}i}w_{co2\text{-}i}$ และ $v_{co2\text{-}i+1}w_{co2\text{-}i+1}$ ในตารางที่ 2 กับสมการ(4) โดยการเลื่อนคู่ EDU ที่อยู่ติดกัน (EDUi EDUi+1) ไปด้วยระยะทางหนึ่ง EDU จนกระทั่งค่า BoundaryClass เป็น‘end’

5.3 ขั้นตอนการหาความสัมพันธ์อาการ-วิธีการรักษา

หลังจากการสกัดกลุ่มฟีเจอร์ในรูปของเวคเตอร์ ออกมาเป็นสองกลุ่มหรือสองเวคเตอร์ของคู่คำคือ Word-Co_{sym} Vector (〈 $v_{co1\text{-}1}$ $w_{co1\text{-}1}$, $v_{co1\text{-}2}$ $w_{co1\text{-}2}$,..,$v_{co1\text{-}a}$ $w_{co1\text{-}a}$ 〉 ซึ่งแทน Dsym) และWord-Co_{treat} Vector (〈 $v_{co2\text{-}1}$ $w_{co2\text{-}1}$, $v_{co2\text{-}2}$ $w_{co2\text{-}2}$,.., $v_{co2\text{-}y}$ $w_{co2\text{-}y}$ 〉 ซึ่งแทน AT/RT,) แล้ว ขั้นตอนนี้เป็นการนำเวคเตอร์คู่คำทั้งสองเวคเตอร์ที่สกัดได้มาทดสอบว่ามีความสัมพันธ์อาการ-วิธีการรักษาหรือไม่ด้วยการคำนวณหาค่าความสัมพันธ์อาการ-วิธีการรักษาจากตัวจัดประเภท NB (สมการ (5)) กับค่าความน่าจะเป็นของ Word-Co_{sym} และ Word-

Co_{treat} (ตารางที่ 3) ที่ทำให้เกิดคลาสแบบความสัมพันธ์อาการ-วิธีการรักษาและแบบไม่ใช่ความสัมพันธ์อาการ-วิธีการรักษา ดังแสดงในอัลกอริทึมการสกัดความสัมพันธ์อาการ-วิธีการรักษา (รูปที่ 5)

```
Assume that each EDU is represented by (NP VP). L is a list of EDU.
VW_symptom is a set of word-order-pairs having the symptom concepts and
VW_treatment is a set of word-order-pairs having the treatment concepts.
v_co1∈V_co1, v_co2∈V_co2, w_co1∈W_co1, w_co2∈W_co2
SYMPTOM_TREATMENT_REL_EXTRACTION( L, V_co1, V_co2, W_co1, W_co2 )

 1  i← 1;j←1; R←∅; flag←0; SymptomVector ← ∅;
 2  while i ≤ length[L] do
 3    { while flag = 0  /*findSymptomConceptEDU
 4        if v_s-i w_s-i∈ VW_symptom  then  flag=1
 5        else  i++ ;
 6      While notEndofBoundary ∧v_co1-i w_co1-i∈VW_symptom∧v_co1-i+1 w_co1-i+1∈VW_symptom
                        /*findSymptomFeatureVector
        { SVMequition(6)or NBequation(4) /*SVM  or NB classifier with a
          slide window size of two consecutive EDUs with one sliding EDU
          distance.
 7        SymptomVector← SymptomVector ∪v_co1-i w_co1-I;
 8          i ++ };
 9      Flag← 0 ; j←1; treatmentVector ←∅;
10      while flag = 0  /*findTreatmentConceptEDU
11        if  v_co2-j w_co2-j∈ VW_treatment  then  flag=1
12        else  {i++ ;j++};
13      While notEndofBoundary ∧v_co2-j w_co2-j∈VW_treatment∧v_co2-i+1 w_co2-i+1∈VW_treatment
                        /*findTreatmentFeatureVector
        { SVMequition(6)or NBequation(4) /*SVM  or NB classifier  with a
          slide window size of two consecutive EDUs with one sliding EDU
          distance.
14        treatmentVector←treatmentVector ∪v_co2-j w_co2-j;
15              j++; i++};
        SymptomTreatmentRelationExtraction by NBequation(5)
20      if SymptomTreatmentRelation = yes   then
21       {R ← R ∪{⟨SymptomVector⟩+⟨TreatmentVector⟩};
22        i++ };
23    }Return R
```

รูปที่5 แสดงอัลกอริทึมการสกัดความสัมพันธ์อาการ-วิธีการรักษาจากเอกสารที่ปรึกษาสุขภาพ

ผลการทดลองและการประเมินผล

คลังข้อมูลภาษาไทยระดับประโยค หรือ EDU จากเว็บบอร์ดดังที่ได้กล่าวข้างต้นสำหรับใช้ทดสอบเพื่อประเมินวิธีการหาความสัมพันธ์ปัญหา-วิธีการแก้ปัญหานั้น (โดยเฉพาะความสัมพันธ์อาการ-วิธีการรักษา) ประกอบด้วยโรคทางเดินอาหารจำนวน 500 EDUs โรคเด็กจำนวน 500 EDUs และโรคทางสมองและหัวใจจำนวน 500 EDUs ซึ่งใช้ประเมินความถูกต้องของวิธีการสกัดฟีเจอร์ และวิธีการหาความสัมพันธ์อาการ-วิธีการรักษาตามที่ได้เสนอ โดยใช้ผู้เชี่ยวชาญ 3 ท่านตัดสินโดยวิธีการลงคะแนนเสียงข้างมาก (Max Win Voting) ดังนี้

1. การประเมินความถูกต้องของวิธีการสกัดฟีเจอร์ Dsym และ AT/RT ประกอบด้วยการประเมินการระบุ EDU_{sym} (ที่เป็น EDU เริ่มต้นของ Dsym โดย$VW_{symptom}$) และการระบุ EDU_{treat} (ที่เป็นEDUเริ่มต้นของ AT/RT โดย $VW_{treatment}$) ด้วยค่าความแม่นยำ (Precision) และค่าระลึก (Recall) และการประเมินขอบเขตของ Dsym และขอบเขต AT/RT (โดยเทคนิค NB และ SVM) ด้วยค่าเปอร์เซ็นต์ความถูกต้อง (% Correctness) ดังแสดงในตารางที่ 4

ตารางที่ 4 แสดงความถูกต้องของวิธีการสกัดฟีเจอร์ Dsym และ AT/RT

ข้อมูลโรค (500 EDUs)	จำนวน Word-Co_{sym} ที่แตกต่าง	จำนวน Word-Co_{treat} ที่แตกต่าง	ความถูกต้องของการระบุ EDU เริ่มต้นของ Dsym		ความถูกต้องของการระบุ EDU เริ่มต้นของ AT/RT		%ความถูกต้องของการหาขอบเขต EDU_{sym} (Dsym) และ EDU_{treat} (AT/RT)			
							หาDsym โดยเทคนิค		หา AT/RT โดยเทคนิค	
			Precision	Recall	Precision	Recall	NB	SVM	NB	SVM
โรคทางเดินอาหาร	69	38	0.889	0.762	0.882	0.857	79.8	81.9	87.2	89.7
โรคเด็ก	74	41	0.875	0.700	0.933	0.848	80.1	84.5	85.8	87.1
โรคทางสมองและหัวใจ	56	39	0.917	0.846	0.894	0.850	82.5	85.3	88.5	90.6
เฉลี่ย	66	39	0.894	0.769	0.903	0.852	80.8	83.9	87.2	89.1

จากตารางที่ 4 ค่าเฉลี่ยของความแม่นยำจากการใช้ $VW_{symptom}$ และ $VW_{treatment}$ ระบุ EDU เริ่มต้นของ Dsym และ AT/RT คือ 0.894 และ 0.903 พร้อมด้วยค่าเฉลี่ยของค่าระลึกคือ 0.769 และ 0.852 ตามลำดับ สาเหตุที่มีค่าระลึกของการระบุ EDUเริ่มต้นของ Dsym และ AT/RT ต่ำเพราะจำนวนคำใน Word-Co ที่ใช้เกิดจากสองคำที่อยู่ติดกันหลังกำจัดคำหยุดนั้น ไม่

เพียงพอต่อการระบุ EDU_{sym} และ EDU_{treat} เช่น ‘รู้สึก+มี+อะไร+ กดทับ+ หน้าอก’ (แน่นหน้าอก / feel tight chest)

ค่าความแม่ยำของการระบุ EDUเริ่มต้นของ Dsym ต่ำกว่าของ AT/RT เพราะ เกิดมีสัมพันธ์เหตุและผลเกิดขึ้นทำให้การระบุนั้นผิดไปเช่น

EDU_{sym-1}: “[คนไข้] (มีอาการ+ปวดท้อง) มาหลายวัน”

EDU_{sym-2}: “[คนไข้] (รู้สึก+แน่นท้อง)”

EDU_{at-1}as Effect: “แล้ว [คนไข้] (มีอาการ+ท้องเสีย)”

EDU_{at-2}as Cause: “เนื่องจาก [คนไข้] (กิน+ยาแก้ท้องอืด)”

EDU_{at-1} เป็น EDUเริ่มต้นของ Dsym ชุดใหม่หลังจากกินยาแก้ท้องอืด

นอกจากนี้ตารางที่ 4 แสดงให้เห็นว่าจำนวนความแตกต่างของ $Word\text{-}Co_{sym}$ ที่ปรากฏในเอกสารโรคทางสมองและหัวใจ มีจำนวนความแตกต่างน้อยกว่าที่ปรากฏในเอกสารโรคทางเดินอาหารและโรคเด็ก ซึ่งส่งผลให้ $Word\text{-}Co_{sym}$ ปรากฏในเอกสารของโรคทางสมองและหัวใจอย่างมีความถี่สูงซึ่งมีผลต่อการเรียนรู้ของเครื่องโดยเฉพาะ NB ทำให้โรคทางสมองและหัวใจมีค่า %ความถูกต้องของการหาขอบเขต Dsym สูงกว่าของโรคทางเดินอาหารและโรคเด็ก นอกจากนี้จำนวนความแตกต่างของ $Word\text{-}Co_{treat}$ ที่ปรากฏ ในเอกสารโรคทางเดินอาหาร โรคเด็ก และโรคทางสมองและหัวใจนั้น มีค่าใกล้เคียงกัน ทำให้ %ความถูกต้องของการหาขอบเขต AT/RT ด้วย NB ในแต่ละโรคนั้นมีค่าใกล้เคียงกัน อย่างไรก็ตามจำนวนความแตกต่างของ $Word\text{-}Co_{sym}$ และของ $Word\text{-}Co_{treat}$ (หรือความถี่การปรากฏ ของ $Word\text{-}Co_{sym}$ และของ $Word\text{-}Co_{treat}$) ในเอกสารโรคทางเดินอาหาร โรคเด็ก และโรคทางสมองและหัวใจนั้น ไม่มีผลต่อ%ความถูกต้องของการหาขอบเขตทั้ง Dsym และAT/RT ด้วย SVM

นอกจากนี้ความผิดพลาดของค่า%ความถูกต้องของการหาขอบเขตทั้ง Dsym และ AT/RT โดย SVM หรือ NB นั้นบางครั้งเกิดจาก มีการขัด (Interrupt) เกิดขึ้นระหว่างหาขอบเขต Dsym หรือ AT/RT ตัวอย่างเช่น

EDU_{sym-1}: หนู (มีอาการ+ท้องผูก) ค่ะ

EDU_{sym-1}: [หนู] (รู้สึก แน่น+ท้อง)

EDU_{at-1}: [หนู]พยายาม (ฝึก+ถ่าย) ทุกวัน

EDU_{at-2}: [หนู] (ออกกำลังกาย+null) ด้วย

$EDU_{interrupt}$: จึง ได้ผล (Class-cue-word Pattern)

EDU_{at-3}: แต่ หนู ต้อง (กิน+โยเกิร์ต) ด้วยคะ

2. การประเมินความถูกต้องของวิธีการหาความสัมพันธ์อาการ-วิธีการรักษาโดย NB นั้น เป็นการประเมินด้วยค่าความแม่นยำคือ 0.84 และค่าระลึกคือ 0.72 สาเหตุที่มีค่าระลึกต่ำเพราะมีบางอาการเช่น ‘เป็นหนอง+พุพอง’ ‘เป็นฝ้า+ขาว’ ‘มีผื่น+แดง’ ‘มีรอย+คล้ำ’ เป็นต้น ปรากฏน้อยมากในคลังข้อมูล โรคทางเดินอาหาร โรคในเด็ก และโรคทางสมองและหัวใจ ฉะนั้นหากคลังข้อมูลมีขนาดเพิ่มขึ้นค่าระลึกก็ควรจะเพิ่มขึ้นตาม

สรุป

งานวิจัยนี้มุ่งเน้นมุ่งเน้นหาความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา (ซึ่งก็คือความสัมพันธ์ระหว่างปัญหาและวิธีการแก้ปัญหา) โดยเฉพาะความสัมพันธ์อาการ-วิธีการรักษา จากเอกสารภาษาไทยบนเว็บบอร์ด (http://haamor.com/) ทั้งปัญหา/อาการ และวิธีการแก้ปัญหา/วิธีการรักษา แสดงออกมาในรูปแบบของหลาย ๆประโยค หรือ EDUs ที่ต่อเนื่องกัน ในขณะที่งานวิจัยก่อนหน้าที่กล่าวมานี้ ทำการหาความสัมพันธ์ ปัญหา-วิธีการแก้ปัญหา ในรูปของนามวลีที่สัมพันธ์กันภายในหนึ่งประโยคดังในงานวิจัยของ Rosario B.(2005) Abacha A. B. and Zweigenbaum P. (2011) และ Yeleswarapu et al., (2014) แต่งานวิจัยของ Song S. et al, (2011) จะประกอบด้วย ปัญหาซึ่งเป็นนามวลีในหนึ่งประโยคและหลายประโยคของวิธีแก้ปัญหา ในขณะที่งานวิจัยนี้เน้นถึงการบรรยายเหตุการณ์ของปัญหาและวิธีการแก้ปัญหา ออกมาในรูปของกริยาวลีจากหลาย EDUs ทั้งของปัญหาและของวิธีการแก้ปัญหา ทั้งนี้เพื่อให้เกิดความเข้าใจอย่างชัดเจนขึ้น

งานวิจัยนี้สามารถหาความสัมพันธ์อาการ-วิธีการรักษา (ความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา) ในรูปแบบหลาย ๆ EDUs ได้อย่างมีประสิทธิภาพ โดยเฉพาะอย่างยิ่งในขั้นตอนการสกัดฟีเจอร์ซึ่งประกอบด้วยฟีเจอร์เวคเตอร์ของปัญหาเช่น Dsym และฟีเจอร์เวคเตอร์ของวิธีการแก้ปัญหาเช่น AT/RT ด้วยการใช้วิธีการประมวลผลภาษาธรรมชาติเพื่อหาคู่คำจากกริยาวลีร่วมกับการเรียนรู้ของเครื่องด้วยเทคนิค ที่แตกต่างกันสองวิธี ME และ SVM หลังจากใช้ $VW_{symptom}$ และ $VW_{treatment}$ ทำการระบุจุดเริ่มต้นของ Dsym และ AT/RT ดังนั้นค่าเฉลี่ยความถูกต้องของการหาขอบเขตของ Dsym ด้วยวิธี NB และ SVM คือ 80.8 และ 83.9 ตามลำดับ และการหาขอบเขตของ AT/RT ด้วยวิธี NB และ SVM คือ 87.2 และ 89.1 ตามลำดับ ฉะนั้นการหาขอบเขตด้วยวิธี SVM ให้ผลลัพธ์ที่มีประสิทธิภาพกว่า NB เพราะบางฟีเจอร์คู่คำมีการขึ้นต่อกัน เช่น ‘มีอาการ+เวียนศรีษะ’→ ‘รู้สึก+คลื่นไส้’, ‘รู้สึก+คลื่นไส้’→ ‘อาเจียน+null’ เป็นต้น ดังนั้นหากมีการพัฒนาคู่คำ หรือ Word-Co จากที่ประกอบด้วยสองคำหรือคำเดียวไปเป็นสามคำหลังจากกำจัดคำหยุด ก็จะทำให้ได้ $VW_{symptom}$ และ $VW_{treatment}$ ที่สามารถระบุอาการ และวิธีการรักษาได้ถูกต้องมากขึ้น

อย่างไรก็ตามงานวิจัยที่ได้เสนอนี้สามารถหาความสัมพันธ์อาการ-วิธีการรักษาได้อย่างมีประสิทธิภาพด้วยค่าความแม่นยำ 0.84 และค่าระลึก 0.72 โดยการเรียนรู้ของเครื่องด้วย NB จาก Dsym และ AT/RT ที่สกัดถูกต้อง อย่างไรก็ตามค่าความแม่นยำและค่าระลึกจะสามารถ

เพิ่มขึ้นได้หากมีการจัดกลุ่มอาการ (Symptom Clustering) และ การจัดกลุ่มวิธีการรักษา (Treatment Clustering) ก่อนเรียนรู้ของเครื่องด้วย NB

นอกจากนี้ผลลัพธ์ที่ได้จากงานวิจัยนี้สามารถนำไปสร้างกราฟแสดงความสัมพันธ์อาการ-วิธีการรักษาทั้งหมด หรือแผนที่แสดงความสัมพันธ์ปัญหา-วิธีการแก้ปัญหา (Problem-Solving Map) ที่อยู่ในโดเมนอื่น ๆ เช่น การซ่อมรถยนต์ การแก้ปัญหาทางธุรกิจ เป็นต้น ซึ่งสามารถนำไปประยุกต์กับระบบถาม–ตอบอัตโนมัติ (Automatic Question Answering System) บนเว็บบอร์ด เพื่อทำให้ผู้ใช้สามารถเข้าใจได้ง่ายขึ้น และนำไปใช้แก้ไขปัญหา

บรรณานุกรม

Abacha, A.B. and Zweigenbaum, P. : Automatic extraction of semantic relations between medical entities: a rule based approach. Journal of Biomedical Semantics, 2 (Suppl 5):S4 (2011) (http://www.jbiomedsem .com/content/2/S5/S4)

Carlson, L., Marcu, D., Okurowski, M. E.: Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Current Directions in Discourse and Dialogue. pp.85-112 (2003)

Chanlekha, H., Kawtrakul, A.: Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. IJCNLP' 2004 proceedings (2004)

Chareonsuk, J ., Sukvakree, T., Kawtrakul, A.: Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information. NCSEC 2005 proceedings (2005)

Cristianini, N. and Shawe-Taylor, J.: An Introduction to Support Vector Machines, CambridgeUniversity Press, Cambridge, UK (2000).

Guthrie, J. A., Guthrie, L., Wilks, Y., Aidinejad, H.: Subject-dependent co-occurrence and word sense disambiguation. Proceedings of the 29th annual meeting on Association for Computational Linguistics (1991)

Mitchell, T.M.: Machine Learning. The McGraw-Hill Companies Inc. and MIT Press, Singapore (1997)

Rosario, B.: Extraction of semantic relations from bioscience text. A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Information Management and Systems. University of California, Berkeley. (2005)

Song, S-K., Oh, H-S. , Myaeng, S.H., Choi, S-P., Chun, H-W., Choi, Y-S., and Jeong, C-H.: Procedural Knowledge Extraction on MEDLINE. AMT 2011, LNCS 6890, pp. 345–354 (2011).

Sudprasert, S., Kawtrakul, A. Thai Word Segmentation based on Global and Local Unsupervised Learning. NCSEC'2003 Proceedings (2003)

Yeleswarapu S., Aditya Rao., Joseph T., Saipradeep V. G. and Srinivasan R., A pipeline to extract drug-adverse event pairs from multiple data sources. "Medical informatics and Decision Making" BMC Medical Informatics and Decision Making 2014, 14:13 (24 February 2014)

ภาคผนวก

Advances in Intelligent Systems and Computing

Volume 364

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

More information about this series at http://www.springer.com/series/11156

Andrzej M.J. Skulimowski · Janusz Kacprzyk
Editors

Knowledge, Information and Creativity Support Systems: Recent Trends, Advances and Solutions

Selected Papers from KICSS'2013—8th International Conference on Knowledge, Information, and Creativity Support Systems, November 7–9, 2013, Kraków, Poland

Springer

Editors
Andrzej M.J. Skulimowski
Decision Sciences Laboratory, Department of Automatic Control and Biomedical Engineering, Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering
AGH University of Science and Technology
Kraków
Poland

and

International Centre for Decision Sciences and Forecasting
Kraków
Poland

Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
Warsaw
Poland

and

Faculty of Physics and Applied Computer Science
AGH University of Science and Technology
Kraków
Poland

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-19089-1 ISBN 978-3-319-19090-7 (eBook)
DOI 10.1007/978-3-319-19090-7

Library of Congress Control Number: 2015958542

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer International Publishing AG Switzerland

Contents

Using Extracted Symptom -Treatment Relation from Texts to Construct Problem-Solving Map

Chaveevan Pechsiri[1], Onuma Moolwat[1], and Rapepun Piriyakul[2]

[1] Dept. of Information Technology, DhurakijPundit University, Bangkok, Thailand
[2] Dept. of Computer Science, Ramkhamhaeng university , Bangkok, Thailand
{ itdpu@hotmail.com , moolwat@hotmail.com, rapepunnight@yahoo.com }

Abstract. This paper aims to extract the relation between the disease symptoms and the treatments (called the Symptom-Treatment relation), from hospital-web-board documents to construct the Problem-Solving map which benefits for inexpert-people to solve their health problems primarily. Both symptoms and treatments expressed on documents are based on several EDUs (Elementary Discourse Units). Our research contains three problems: first is how to identify a symptom-concept EDU and a treatment-concept EDU. Second is how to determine a symptom-concept-EDU boundary and a treatment-concept-EDU boundary. Third is how to determine the Symptom-Treatment relation from documents. Therefore, we apply a word co-occurrence to identify a disease-symptom-concept/treatment-concept EDU and Naïve Bayes to determine a disease-symptom-concept boundary and a treatment concept boundary. We propose using k-mean and Naïve Bayes to determine the Symptom-Treatment relation from documents with two feature sets, a symptom-concept-EDU group and a treatment-concept-EDU group. Finally, the research achieves the 87.5% precision and 75.4% recall of the Symptom-Treatment relation extraction along with the Problem-Solving map construction.

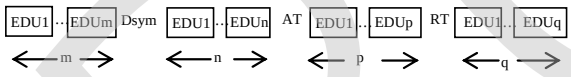
Keywords: word order pair, Elementary Discourse Unit, Symptom-Treatment relation

1 Introduction

The objective of this research is to develop a system of automatically extracting relation between the group of disease's symptoms and the treatment/treatment procedure (called the Symptom-Treatment relation) from the medical-care-consulting documents on the hospital's web-board of a non-government-organization (NGO) website edited by patients and professional medical practitioners. The extracted Symptom-Treatment relation is also conducted the construction of Problem-Solving map (PSM) which is a map representing how to solve problems, especially disease treatments. PSM benefits general people to understand how to solve their health problems in the preliminary stage. Each medical-care-consulting document contains the disease symptoms and the treatments are expressed in the form of EDUs (EDU is Elementary Discourse Unit, which is a simple sentence/clause defined by [3]. Each EDU is expressed by the following linguistic expression:

EDU → NP1 VP | VP
VP → verb NP2 | verb adv
verb → $verb_{weak}$-noun1 | $verb_{weak}$-noun2 |$verb_{strong}$
NP1 → pronoun | noun1 | noun1 AdjPhrase1
NP2→ noun2 | noun2 AdjPhrase2
noun1 → ‘ ’ , ‘ผู้ป่วย/patient’, ‘อวัยวะ/human organ’, ‘แผล/ scar’,…
noun2→ ‘ ’ , ‘อวัยวะ/human organ’, ‘อาการ/symptom’, ‘สี../..color’, ‘ยา/medicine’,…
$verb_{weak}$→ ‘เป็นแผล/have scar’, ‘มีอาการ/have symptom’,
$verb_{strong}$→’รู้สึกปวด/pain’, ‘ใช้/apply’ , ‘ทา/apply’, …
where NP1 and NP2 are noun phrases, VP is a verb phrase, adv is an adverb, and AdjPhrase1 and AdjPhrase2 are adjective phrases.

Moreover, there are two kinds of the treatments existing on the web-board documents; the actual treatment notified by patients/users from their experiences and the recommended treatment edited by professional medical practitioners. Thus, each medical-care-consulting document contains several EDUs of the disease-symptom-concepts along with the actual-treatment-concept EDUs and the recommended-treatment-concept EDUs as shown in the following form.

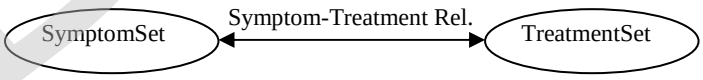


where
- Dsym, AT, and RT are a group of disease-symptom-concept EDUs, a group of actual-treatment-concept EDUs, and a group of recommended-treatment-concept EDUs respectively, as follow:
  Dsym = ($EDU_{sym-1}EDU_{sym-2}$ .. EDU_{sym-a}) where *a* is an integer number and is >0,
  AT = (EDU_{at-1} EDU_{at-2} .. EDU_{at-b}) where *b* is the number of EDU_{at} and is ≥0,
  RT = (EDU_{rt-1} EDU_{rt-2} .. EDU_{rt-c}) where *c* is the number of EDU_{rt} and is ≥0
- m, n, p, and q are the number of EDUs and are ≥0

Therefore, the Symptom-Treatment relation can be expressed as follow:

Dsym → AT
Dsym → RT

From Fig.1, Dsym is EDU1-EDU3, AT is EDU5, and RT is EDU8-EDU10. Thus, the extracted Symptom-Treatment relations (as shown in the following) from medical-care-consulting documents with several Problem-Topic names, e.g. diseases, are collected for constructing the general PSM representation (see Fig.5).



where SymptomSet={$Dsym_1$, $Dsym_2$,…..$Dsym_n$}, TreatmentSet= At∪RT, and $Dsym_1$≠ $Dsym_2$≠…..≠$Dsym_n$

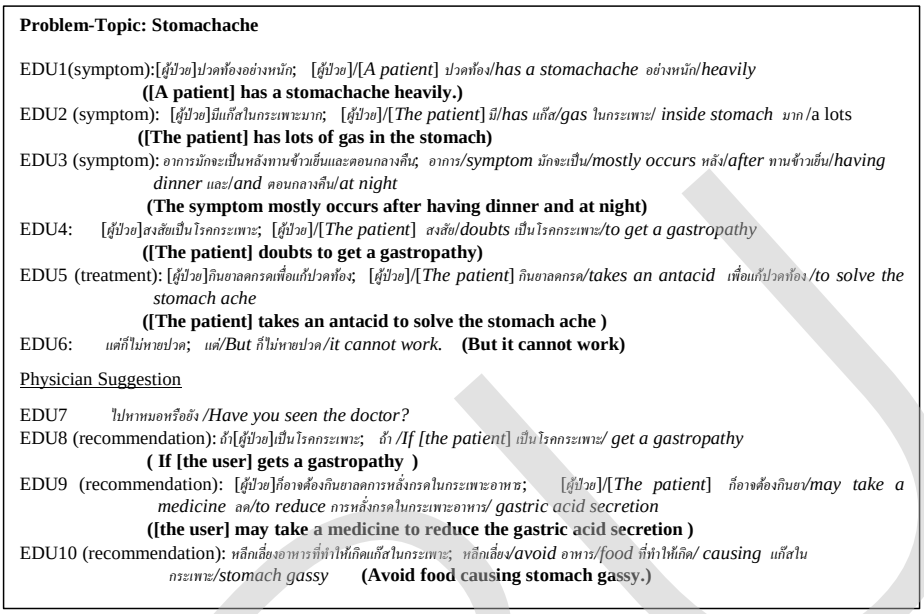


Fig 1. An example of a web-board document contains the Symptom-Treatment relations (where the [..] symbol means ellipsis.)

There are several techniques [8],[1],[9] having been used to extract the Symptom-Treatment relation or the disease treatment relation from texts (see section2). However, the Thai documents have several specific characteristics, such as zero anaphora or the implicit noun phrase, without word and sentence delimiters, and etc. All of these characteristics are involved in three main problems of extracting the Symptom-Treatment relation from the NGO web-board documents (see section 3): the first problem is how to identify the disease-symptom-concept EDU and the treatment-concept EDU. The second problem is how to identify the symptom-concept-EDU boundary as Dsym and the treatment-concept-EDU boundary as AT/RT. And, the third problem is how to determine the Symptom-Treatment relation from documents. From all of these problems, we need to develop a framework which combines the machine learning technique and the linguistic phenomena to learn the several EDU expressions of the Symptom-Treatment relations. Therefore, we apply learning relatedness value [6] of a word co-occurrence (called "Word-CO") to determine a symptom-concept EDU or a treatment-concept EDU. Word-CO in our research is the expression of two adjacent words as a word order pair (where the first word is a verb expression) existing in one EDU and having either a disease symptom concept or a treatment concept. The Naïve Bayes classifier [7] is also applied to solve the disease-symptom-concept-EDU boundary and also the treatment-concept-EDU boundary from the consecutive EDUs. We also propose using The Naïve Bayes classifier to determine the Symptom-Treatment relation from documents after clustering objects of posted problems (Dsym) on the web-board and clustering treatment features.

Our research is organized into five sections. In section 2, related work is summarized. Problems in extracting Symptom-Treatment relations from Thai

documents are described in section 3 and section 4 shows our framework for extracting the Symptom-Treatment relation. In section 5, we evaluate and conclude our proposed model.

2 Related Work

Several strategies [8], [1], [9] have been proposed to extract the disease treatment relation (or the Symptom-Treatment relation as in our research) from the textual data.

In 2005, Rosario [8] extracted the semantic relations from bioscience text. In general, the entities are often realized as noun phrases, the relationships often correspond to grammatical functional relations, as shown in the following example.
"*Therefore administration of TJ-135 may be useful in patients with severe acute hepatitis accompanying cholestasis or in those with autoimmune hepatitis.*"
Where the disease *hepatitis* and the treatment *TJ-135* are entities and the semantic relation is: *hepatitis* is treated or cured by *TJ-135*. The goals of her work is to identify the semantic roles DIS (Disease) and TREAT (Treament), and to identify the semantic relation between DIS and TREAT from bioscience abstracts. She identified the entities (DIS and TREAT) by using MeSH and the relationships between the entities by using a neural network based on five graphical models with lexical, syntactic, and semantic features. Her results were 79.6% accuracy in the relation classification when the entities were hidden and 96.9% when the entities were given.

In 2011, Abacha and Zweigenbaum [1] extracted semantic relations between medical entities (as the treatment relations between a medical treatment and a problem, e.g. disease) by using the linguistic pattern-based to extract the relation from the selective MEDLINE articles.

Linguistic Pattern: … E1 … be effective for E2… |… E1 was found to reduce E2 …,

where E1, E2, or Ei is the medical entity (as well as UMLS concepts and semantic types) identified by MetaMap.
Their treatment relation extraction was based on a couple of medical entities or noun phrases occurring within a single sentence, as shown in the following example:
"*Fosfomycin* (E1) *and amoxicillin-clavulanate* (E2) *appear to be effective for cystitis* (E3) *caused by susceptible isolates.*"
Finally, their results showed 75.72% precision and 60.46% recall.

In 2011, Song et al. [9] extracted the procedural knowledge from MEDLINE abstracts as shown in the following by using Supporting Vector Machine (SVM) comparing to Conditional Random Field (CRF), along with Natural language Processing.
"….*[In a total gastrectomy]* (Target), *[clamps are placed on the end of the esophagus and the end of the small intestine]* (P1). *[The stomach is removed]* (P2) *and [the esophagus is joined to the intestine]* (P3). …", where P1, P2, and P3 are the solution procedures. They defined procedural knowledge as a combination of Target and a corresponding solution consisting of one or more related procedures/methods. SVM and CRF were utilized with four feature types: content feature (after word stemming and stop-word elimination) with a unigram and bi-grams in a target sentence, position feature, neighbor feature, and ontological feature to classify Target. And, the other features:

word feature, context feature, predicate-argument structure, and ontological feature, were utilized to classify procedures from several sentences. Their results are 0.7279 and 0.8369 precision of CRF and SVM respectively with 0.7326 and 0.7957 recall of CRF and SVM respectively.

Most of the previous works, i.e. [8] and [1], the treatment relation between the medical treatment and the problem (as a disease) occurs within one sentence whereas our Symptom-Treatment relation occurs with in several sentences / EDUs on both the treatments and the problem. However, the [9] work has several sentences of the treatment method but there is one sentence of problem as the Target disease or symptom. Therefore, we propose using the Naïve Bayes classifier to learn the Symptom-Treatment relation with features from clustering objects of posted problems (Dsym) on the web-board and clustering treatment concepts from AT/RT.

3 Problems of Symptom –Treatment Relation Extraction

To extract the Symptom-Treatment relation, there are three problems that must be solved: how to identify a symptom-concept EDU and a treatment-concept EDU, how to determine a symptom-concept-EDU boundary and a treatment-concept-EDU boundary, and how to determine Symptom-Treatment relations from documents

3.1 How to Identify Symptom Concept EDU and Treatment Concept EDU

According to the corpus behavior study of the medical care domain, most of the symptom concept EDUs and the treatment concept EDUs are expressed by verb phrase. For example:

Symptom Concept

a) EDU: "*ผู้ป่วยรู้สึกเวียนศีรษะ*" / "***A patient feels dizzy*.**"
"(*ผู้ป่วย*/*A patient*)/NP1 (*รู้สึก*/*feels* *เวียนศีรษะ*/*dizzy*)/VP"

b) EDU: "*ฉัน*[*มีอาการ*]*ปวดศีรษะ*" / "***I have a headache symptom***."
"(*ฉัน*/*I*)/NP1 ((*มีอาการ*/*have symptom*])/verb (*ปวดศีรษะ*/*headache*)/NP2)/VP"

where [..] means ellipsis.

Treatment Concept

c) EDU: "*กินยาลดกรด*" / "***Take an antacid***."
"((*กิน*/*consume*)/verb (*ยาลดกรด*/*antacid*)/NP2)/VP"

However, some verb phrase expressions of the symptom concepts are ambiguities. For examples:

e) EDU: "[*คนไข้*]*ถ่ายยาก*" / "[***A patient***] ***has hard time bowel movement***."
"([*คนไข้*/*patient*])/NP1 ((*ถ่าย*/*defecate*)/verb (*ยาก* /*difficultly*)/adv)/VP"

f) EDU1: "*ห้องน้ำสกปรกมาก*" / "***A toilet is very dirty***."
"(*ห้องน้ำ*/*toilet*)/NP1 ((*สกปรกมาก* /*is very dirty*)/verb)/VP"
EDU2: "*ฉันจึงถ่ายยาก*" / "***Then, I has hard time bowel movement***."
"(*ฉัน*/*I*)/NP1 (*จึง*/*then*)/adv ((*ถ่าย*/*defecate*)/verb (*ยาก* /*difficultly*)/adv)/VP"

From e) and f) examples, the verb phrase expression of the symptom concept occurs only in e) with the concept of '*ท้องผูก*/be constipated'.

This problem can be solved by learning the relatedness from two consecutive words of Word-CO with the symptom concept or treatment concept. Where the first word of Word-CO is a verb expression, v_{co}, approaching to the symptom concept or the treatment concept (where $v_{co} \in V_{co}$, $V_{co} = V_{co1} \cup V_{co2}$, V_{co1} is a set of verbs with approaching to the symptom concepts , and V_{co2} is the treatment-verb concept set). And, the second word of Word-CO is a co-occurred word, w_{co} ($w_{co} \in W_{co}$; $W_{co} = W_{co1} \cup W_{co2}$). W_{co1} and W_{co2} are co-occurred word sets inducing the v_{co1} w_{co1} co-occurrence and the v_{co2} w_{co2} co-occurrence to have the symptom concept and treatment concept respectively, where $v_{co1} \in V_{co1}$, $w_{co1} \in W_{co1}$, $v_{co2} \in V_{co2}$ and $w_{co2} \in W_{co2}$. All concepts of V_{co1}, V_{co2}, W_{co1} , and W_{co2} from the annotated corpus are obtained from Wordnet and MeSH.

V_{co1} = {'*ถ่าย*'/'*defecate*','*เบ่ง*'/'*push*', '*ปวดท้อง*'/'*have an abdomen pain*','*ปวด*'/'*pain*', '*อึดอัด*'/'*be uncomfortable*','*รู้สึกไม่สบาย*'/'*be uncomfortable*','*มี*[*อาการ*]'/'*have*[*symptom*]',..}

V_{co2} ={'*กิน*'/'*consume*', '*ทา*'/'*apply*', '*ใช้/apply*' '*รักษา*'/'*remedy*', '*บำรุง*'/'*nourish*', '*ลด*'/'*reduce*',.. }

W_{co1}={' ', '*ยาก*'/'*difficultly*', '*ถ่าย*'/'*stools*', '*เชื้อ*'/'*germ*', '*เหลว*'/'*liquid*', '*ประจำเดือน*'/'*period*', '*แน่นท้อง*'/'*fullness*', '*ท้องเฟ้อ*'/'*flatulence*', '*ไข้*'/'*fever*', …}

W_{co2} = { '*ยา*'/'*medicine*', '*อาหาร*'/'*food*', '*อาหารเสริม*'/'*supplement*',…}

3.2 How to Solve Symptom-Concept-EDU Boundary (Dsym) and Treatment-Concept-EDU Boundary (AT, RT)

According to Fig. 1, there is no clue (i.e. '*และ*/and', '*หรือ*/or',..) on both EDU3 to identify the disease-symptom boundary (EDU1-EDU3) and EDU10 to identify the treatment boundary (EDU8-EDU10). After the symptom-concept EDU and the treatment-concept EDU has been identified by using Word-CO from the previous step, we then solve the symptom-concept-EDU boundary and the treatment-concept-EDU boundary by applying Naïve Bayes to learn a Word-CO pair from a window size of two consecutive EDUs with one sliding EDU distance.

3.3 How to Determine Symptom-Treatment Relation

The relations between symptoms and treatments are varied among patients, environments, times, etc. even though they have the same disease. For example:

(a) EDU_{sym-1}: "[*ผู้ป่วย*]/NP1 ((*ปวดท้อง* /*have a stomach ache*)/verb (*อย่างหนัก*/*badly*)/adv)/VP"
"[*A patient*] *has a stomachache badly*."

EDU_{sym-2}: "[*ผู้ป่วย*]/NP1 ((*มีแก๊ส* /*has* gas)/verb (*มาก*/*a lots*)/adv *ในกระเพาะ*/*inside stomach*)/VP"
"[*The patient*] *has lots of gas in the stomach.*"

EDU_{at-1}: "[*ผู้ป่วย*]/NP1 ((*กิน*/*takes*)/verb (*ยาลดกรด*/*an antacid*)/NP2)/VP"
"[*The patient*] *takes an antacid*."

EDU1: "(*แต่*/*But*)/conj "[*มัน*/*it*]/NP1 (*ก็ไม่หายปวด* / *cannot work*)/VP" **"*But*[*it*]*cannot work*."**

(b) EDU_{sym-1}:"[*ผู้ป่วย*]/NP1(*ปวดท้อง* //*have a stomachache*)/VP" **"[*A patient*] *has a stomachache*."**

EDU_{sym-2}: "[*ผู้ป่วย*]/NP1 ((*มี*/*has แก๊ส*/gas)/verb (*ในกระเพาะ*/*inside stomach*)/PrepPhrase)/VP"
"[*The patient*] *has gassy in the stomach*"

$EDU_{treat-3}$: "[*ผู้ป่วย*]/NP1 ((*กิน*/*takes*)/verb (*ยาลดกรด*/*an antacid*)/NP2)/VP"
"[*The patient*] *takes an antacid*."

EDU1: "[*ผู้ป่วย*]/NP1 ((*รู้สึกดีขึ้น*/*Feel better*)/verb)/VP" "[***The patient*] *feels better***."

According to (a) and (b) examples, the Symptom-Treatment relation occurs only on b) because EDU4 of (b) contains "*รู้สึกดีขึ้น/ feel better*" as Class-cue-word of the Symptom-Treatment relation. Therefore, we propose automatically learning the Symptom-Treatment relation on documents by using the Naïve Bayes classifier, with the features from clustering objects with Dsym as $\langle v_{co1\text{-}1}\ w_{co1\text{-}1}\ ,\ v_{co1\text{-}2}\ w_{co1\text{-}2}\ ,\ldots,\ v_{co1\text{-}a}\ w_{co1\text{-}a}\ \rangle$ (where each symptom-concept EDU is determined by Word-CO, $v_{co1}\ w_{co1}$), and clustering the treatment features from AT and RT as $\langle v_{co2\text{-}1}\ w_{co2\text{-}1}\ ,\ v_{co2\text{-}2}\ w_{co2\text{-}2}\ ,\ldots,\ v_{co2\text{-}b/c}\ w_{co2\text{-}b/c}\ \rangle$ (where (each treatment-concept-EDU is determined by Word-CO,$v_{co\text{-}2}\ w_{co\text{-}2}$)).

4 A Framework for Symptom-Treatment Relation Extraction

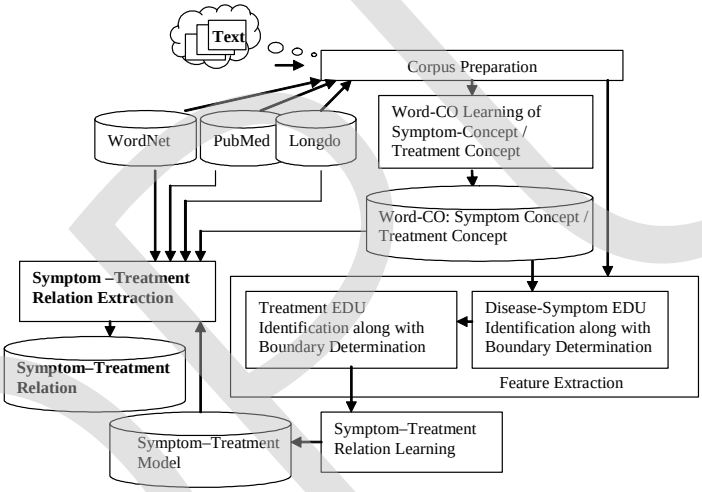


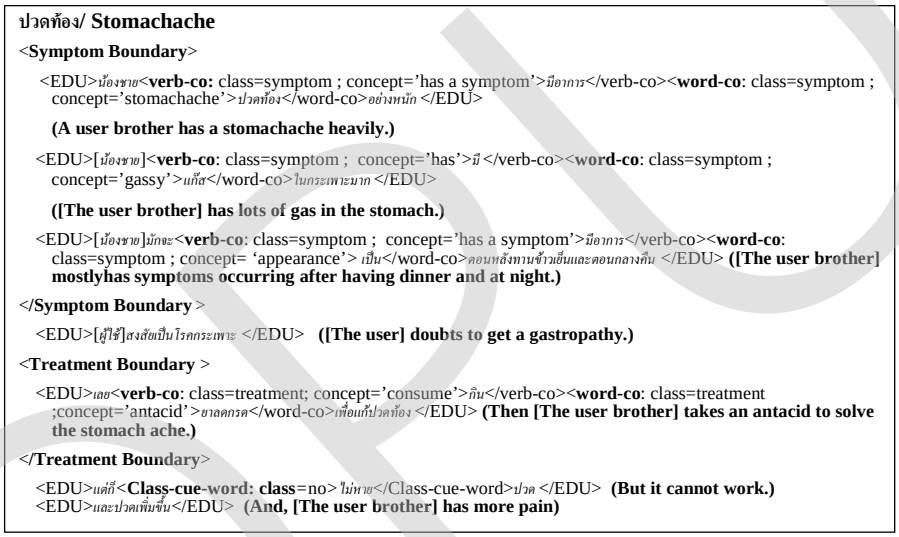
Fig.2 System Overview where the input is Text or downloaded documents and the output is Symptom-Treatment Relation

There are five steps in our framework. The first step is corpus preparation step followed by the step of Word-CO concept learning especially symptom concepts and treatment concepts. Then, the feature extraction step for Symptom-Treatment relation learning step which is followed by the Symptom-Treatment relation extraction step are operated as shown in Fig. 2.

4.1 Corpus Preparation

This step is the preparation of a corpus in the form of EDU from the medical-care-consulting documents on the hospital's web-board of the Non-Government-Organization (NGO) website. The step involves with using Thai word segmentation tools [10], including named entities [4]. After the word segmentation is achieved, EDU segmentation is then to be dealt with [5]. These annotated EDUs will be kept as an EDU corpus. This corpus contains 6000 EDUs of several diseases and is separated

into 2 parts; one part is 4000 EDUs for learning the Word-CO concepts, boundaries of the symptom feature group and the treatment feature group, and the symptom-Treatment relations, based on ten folds cross validation. And, the other part of 2000 EDUs is for determining the boundaries and the symptom-Treatment relation extraction. In addition to this step of corpus preparation, we semi-automatically annotate the Word-CO concepts of symptoms and treatments along with the Class-cue-word annotation to specify the cue word with the Class-type set {"yes", "no"} of the symptom-Treatment relation as shown in Fig.3. All concepts of Word-CO are referred to Wordnet (http://word-net.princeton.edu/obtain) and MeSH after translating from Thai to English, by using Lexitron (the Thai-English dictionary) (http://lexitron.nectec.or.th/).


ปวดท้อง/ Stomachache

<**Symptom Boundary**>

<EDU>*น้องชาย*<**verb-co:** class=symptom ; concept='has a symptom'>*มีอาการ*</verb-co><**word-co**: class=symptom ; concept='stomachache'>*ปวดท้อง*</word-co>*อย่างหนัก* </EDU>

(A user brother has a stomachache heavily.)

<EDU>[*น้องชาย*]<**verb-co**: class=symptom ; concept='has'>*มี* </verb-co><**word-co**: class=symptom ; concept='gassy'>*แก๊ส*</word-co>*ในกระเพาะมาก* </EDU>

([The user brother] has lots of gas in the stomach.)

<EDU>[*น้องชาย*]*มักจะ*<**verb-co**: class=symptom ; concept='has a symptom'>*มีอาการ*</verb-co><**word-co**: class=symptom ; concept= 'appearance'> *เป็น*</word-co>*ตอนหลังทานข้าวเย็นและตอนกลางคืน* </EDU> **([The user brother] mostlyhas symptoms occurring after having dinner and at night.)**

</**Symptom Boundary** >

<EDU>[*ผู้ใช้*]*สงสัยเป็นโรคกระเพาะ* </EDU> **([The user] doubts to get a gastropathy.)**

<**Treatment Boundary** >

<EDU>*เลย*<**verb-co**: class=treatment; concept='consume'>*กิน*</verb-co><**word-co**: class=treatment ;concept='antacid'>*ยาลดกรด*</word-co>*เพื่อแก้ปวดท้อง* </EDU> **(Then [The user brother] takes an antacid to solve the stomach ache.)**

</**Treatment Boundary**>

<EDU>*แต่ก็*<**Class-cue-word: class**=no>*ไม่หาย*</Class-cue-word>*ปวด* </EDU> **(But it cannot work.)**
<EDU>*และปวดเพิ่มขึ้น*</EDU> **(And, [The user brother] has more pain)**

Fig. 3. Symptom-Treatment Relation Annotation

4.2 Word-CO Concept Learning

According to [6], the relatedness value, *r*, has been applied in this research for the relatedness between two consecutive word in Word-CO with either the symptom concept, v_{co1} w_{co1}, or the treatment concept, v_{co2} w_{co2} (as shown in equation (1)). Where each Word-CO, v_{coi} w_{coi} , existing on an EDU contains two relatedness r(v_{coi} , w_{coi}) values. If v_{coi} is v_{co1}, one relatedness value is the symptom concept and the other one is the non-symptom concept. If v_{coi} is v_{co2} , one relatedness value is the treatment concept and the other one is the non-treatment concept. The only v_{coi} w_{coi} co-occurrence with a higher r(v_{coi} , w_{coi}) value of the symptom concept or the treatment concept than the one of the non-symptom concept or the non-treatment concept respectively is collected as an element of $VW_{symptom}$ or $VW_{treatment}$ (v_{co1} $w_{co1} \in$ $VW_{symptom}$ where $VW_{symptom}$ is a set of Word-CO with the symptom concepts, and v_{co2} $w_{co2} \in$ $VW_{treatment}$ where $VW_{treatment}$ is a set of Word-CO with the treatment concepts).

$VW_{symptom}$ and $VM_{treatment}$ are used for identifying the disease-symptom concept EDU and the treatment concept EDU respectively.

$$r(v_{coi}, w_{coi}) = \frac{fv_{coi} w_{coi}}{fv_{coi} + fw_{coi} - fv_{coi} w_{coi}} . \quad (1)$$

where $r(v_{coi}, w_{coi})$ is the relatedness of $Word-CO$ with a symptom concept if $coi = co1$ or a treatment concept if $coi = co2$.

$v_{coi} \in V_{coi}$, $w_{coi} \in W_{coi}$ V_{co1} is a set of verbs with the symptom concepts.

V_{co2} is a set of verbs with the treatment concepts.

W_{co1} is the co-occurred word set having the symptom concept in the $v_{co1}\ w_{co1}$ co-occurrence.

W_{co2} is the co-occurred word set having the treatment concept in the $v_{co2}\ w_{co2}$ co-occurrence.

fv_{coi} is the numbers of v_{coi} occurences. fw_{coi} is the numbers of w_{coi} occurences.

$fv_{coi} w_{coi}$ is the numbers of v_{coi} and w_{coi} occurences.

4.3 Feature Extraction

This step is to extract two feature groups used for classifying the Symptom-Treatment relation in the next step, the symptom feature group (which is Dsym) and the treatment feature group (which is AT / RT). Therefore, the symptom feature group and the treatment feature group can be extracted from the consecutive EDUs by using $v_{co1}\ w_{co1}$ and $v_{co2}\ w_{co2}$ to identify the starting EDU of Dsym and the starting EDU of AT/RT respectively. Then, we learn the probability of a Word-CO pair, $v_{coi\text{-}j}\ w_{coi\text{-}j}\ v_{coi\text{-}j+1}\ w_{coi\text{-}j+1}$, with the symptom concept class (where *coi=co1*) and the treatment concept class (where *coi=co2*) from the learning corpus with a window size of two consecutive EDUs with one sliding EDU distance(where i={1,2}, j={1,2,..,endOfboundary}). The testing corpus of 2000EDUs is used to determine the boundary of the symptom feature group and the boundary of the treatment feature group by Naïve Bayes as shown in equation (2)

$$\begin{aligned} EDUBoundaryClass &= \arg\max_{class \in \text{Class}} P(class \mid v_{coi\_j} w_{coi\_j} \quad v_{coi\_j+1} w_{coi\_j+1}) \\ &= \arg\max_{class \in \text{Class}} P(v_{coi\_j} w_{coi\_j} \mid class) P(v_{coi\_j+1} w_{coi\_j+1} \mid class) P(class) \end{aligned} \quad (2)$$

where $v_{coi\_j} w_{coi\_j} \in \text{VW}_{\text{symptom}}$, $v_{coi\_j+1} w_{coi\_j+1} \in \text{VW}_{\text{symptom}}$ when $coi = co1$, $\text{VM}_{\text{symptom}}$ is a set of Word-CO with the symptom concepts.

$v_{coi\_j} w_{coi\_j} \in \text{VW}_{\text{treatment}}$, $v_{coi\_j+1} w_{coi\_j+1} \in \text{VW}_{\text{treatment}}$ when $coi = co2$, $\text{VM}_{\text{treatment}}$ is a set of Word-CO with the treatment concepts

$j = 1,2,..,endOfboundary$ $\text{Class} = \{"yes","no"\}$

4.4 Symptom-Treatment Relation Learning

It is necessary to cluster objects (or patients posting problems on the web-board) for enhancing the efficiency of learning the Symptom-Treatment relation because there is the high symptom diversity depending on patients, diseases, environment, and etc. We cluster the n samples of posted problems on the web-board by using k-mean as shown in equation (3)[2].

$$Cluster(\mathrm{x}_j) = \arg\min_{1 \le k \le K} \left\| \mathrm{x}_j - \mu_k \right\|^2 \quad (3)$$

where x_j is a disease-symptom vector, Dsym, of an object $\langle v_{co1\text{-}1} w_{co1\text{-}1}, v_{co1\text{-}2} w_{co1\text{-}2}, .., v_{co1\text{-}a} w_{co1\text{-}a} \rangle$ and j=1,2,..,n posted problems. μ_k is the mean vector of the k^{th} cluster. The highest number of $v_{co1\text{-}i} w_{co1\text{-}i}$ occurrences in each cluster is selected to its cluster representative. Thus, we have a symptom cluster set (Y) {rhinorrhoea-based-cluster, abdominalPain-based-cluster, brainSymptom-based-cluster, ….., nSymptom-based-cluster}.

From equation (3), we replace x_j with x_j to cluster the treatment features where x_j is a Word-CO element, $v_{co2\text{-}i} w_{co2\text{-}i}$, of AT ∪ RT and j=1,2,..,m Word-COs, $v_{co2} w_{co2}$. After clustering the treatment features, the highest number of the general concept (based on WordNet and MesH) of $v_{co2\text{-}i} w_{co2\text{-}i}$ occurrences in each cluster is selected to its cluster representatve. Then we have a treatment cluster set (Z) {relax-based-cluster, foodControl-based-cluster, injectionControl-based-cluster,…mTreatment-based-cluster}.

According to clustering the extracted feature vectors from section 4.3, we learn the Symptom-Treatment relation by using Weka (http://www.cs. wakato.ac.nz/ml/weka/) to determine probabilities of y , $z_1,..,z_h$ with the Class-type set of the Symptom-Treatment relation,{'yes' 'no'} where $y \in Y$, $z_1,..,z_h \in Z$, and h is max(b,c) from AT and RT. The Class-type set is specified on any five EDUs right after AT or RT. An element of the Class-type set is determined from the following set of Class-cue-word pattern.

Class-cue-word pattern={'**cue**:หาย/disappear=**class**:yes', '**cue**:รู้สึกดีขึ้น/feel better=**class**: yes', 'cue:ไม่ปวด/do not pain=**class**:yes', '**cue:**" "=**class**:yes', '**cue**:ไม่หาย/appear=**class**: no','**cue**:ยังปวดอยู่/still pain=**class**:no', '**cue**:ปวดมากขึ้น/have more pain=**class**: no',..}

4.5 Symptom-Treatment Relation Extraction

The objective of this step is to recognize and extract the Symptom-Treatment relation from the testing EDU corpus by using Naïve Bayes in equation (4) with probabilities of y , $z_1,..,z_h$ from the previous step with the algorithm shown in Fig.4.

$$SymTreat\_RelClass = \arg\max_{class \in Class} P(class \mid y, z_1, z_2, .., z_h)$$
$$= \arg\max_{class \in Class} P(y \mid class) P(z_1 \mid class) P(z_2 \mid class) .. P(z_h \mid class) P(class) \quad (4)$$

where $y \in Y$, Y is a symptom cluster set.

$z_1, z_2, .., z_h \in Z$, Z is a treatment cluster set.

Class = {"*yes*","*no*")

The extracted Symptom-Treatment relation of this step can be used for constructing PSM as shown in Fig.5.

```
Assume that each EDU is represented by (NP VP).        L is a list of
EDU.  VW_symptom is a set of word-order-pairs having the symptom concepts
and VW_treatment is a set of word-order-pairs having the treatment
concepts (see section4.2).  v_co1∈V_co1,  v_co2∈V_co2,  w_co1∈W_co1, w_co2∈W_co2 (see
section 3.1 )
MEDICINAL_PROPERTY_EXTRACTION( L, V_co1, V_co2, W_co1, W_co2 )

 1  i← 1;j←1; R←Ø; flag←0; SymptomVector ← Ø;
 2  while i ≤ length[L] do
 3    { while flag = 0  /*findSymptomConceptEDU
 4        if v_s-i w_s-i∈ VW_symptom  then  flag=1
 5        else  i++ ;
 6      While notEndofBoundary and v_co1-i w_co1-i∈VW_symptom
                        /*findSymptomFeatureVector
 7      { equation2, SymptomVector← SymptomVector ∪ v_co1-i w_co1-I;
 8          i ++ };
 9      cluster SymptomFeatureVector /*equation 3
10      Flag← 0 ; j←1; treatmentVector ←Ø;
11      while flag = 0  /*findTreatmentConceptEDU
12         if  v_co2-j w_co2-j∈ VW_treatment   then  flag=1
13         else  {i++ ;j++};
14      While notEndofBoundary and v_co2-j w_co2-j∈VW_treatment
                        /*findTreatmentFeatureVector
15      {equation2, treatmentVector←treatmentVector ∪ v_co2-j w_co2-j;
16              j++; i++};
17    cluster TreatmentFeatureVector /*equation 3
18    SymptomTreatmentRelationExtraction by equation 4
19     if SymptomTreatmentRelation = yes   then
20       {R ← R ∪{⟨SymptomVector⟩+⟨TreatmentVector⟩};
21        i++ };
```

Fig. 4 Symptom-Treatment Relation Extraction Algorithm

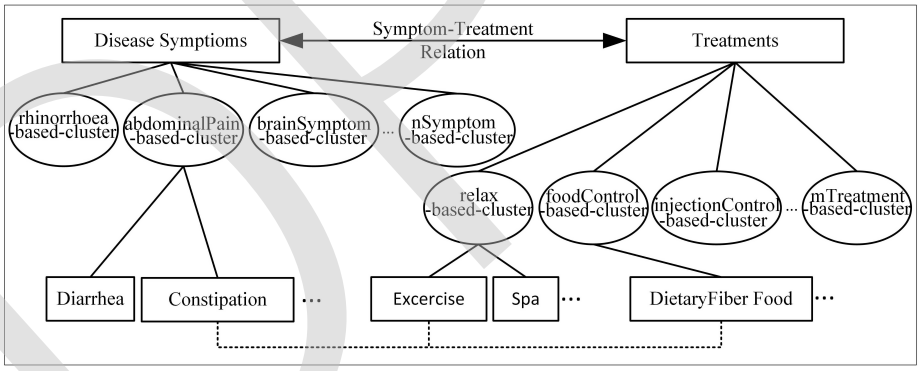


Fig. 5 Show the PSM representation of the Symptom-Treatment relation

5 Evaluation and Conclusion

The Thai corpora used to evaluate the proposed Symptom-Treatment relation extraction algorithm consist of about 2,000 EDUs collected from the hospital's web-board documents of medical-care-consulting. The evaluation of the Symptom-Treatment relation extraction performance of this research methodology is expressed in terms of the precision and the recall. The results of precision and recall are evaluated by three expert judgments with max win voting. The precision of the extracted Symptom-Treatment relation after clustering is 87.5% and 75.4% recall. These research results, especially the low recall, can be increased if the interrupt

occurrences on either a symptom boundary or a treatment boundary, as shown in the following, are solved.

EDU1: *หนูมีอาการท้องผูกค่ะ* (**I have a constipation symptom.**)

EDU2: [*หนู*]*พยายามฝึกถ่ายทุกวัน* (**[I] try to train excretion every day.**)

EDU3: *ได้ผล* (**It can work**)

EDU4: *แต่หนูต้องกินโยเกิร์ตด้วย*: (**But I must have yogurt too**)

where EDU3 is an interrupt to the treatment-concept-EDU boundary (EDU2 and EDU4). Moreover, our extracted Symptom-Treatment relation can be represented by PSM (Fig. 5) which is very beneficial for patients to understand the disease symptoms and their treatment. However, the extracted symptoms and the extracted treatments are various to the patient characteristics, environment, time, and etc. Therefore, the generalized symptoms and the generalized treatments have to be solved before constructing PSM.

6 Acknowledgement

This work has been supported by the Thai Research Fund grant MRG5580030

References

1. Abacha, A.B. and Zweigenbaum, P. : Automatic extraction of semantic relations between medical entities: a rule based approach. Journal of Biomedical Semantics, 2 (Suppl 5):S4 (2011) (http://www.jbiomedsem .com/content/2/S5/S4)
2. Aloise, D.; Deshpande, A.; Hansen, P.; Popat, P. (2009). "NP-hardness of Euclidean sum-of-squares clustering". Machine Learning 75: 245–249.
3. Carlson, L., Marcu, D., Okurowski, M. E.: Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Current Directions in Discourse and Dialogue. pp.85-112 (2003)
4. Chanlekha, H., Kawtrakul, A.: Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. IJCNLP'2004 proceedings,pp1-7(2004)
5. Chareonsuk, J ., Sukvakree, T., Kawtrakul, A.: Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information. NCSEC 2005 proceedings, Thailand, pp.85-90 (2005)
6. Guthrie, J. A., Guthrie, L., Wilks, Y., Aidinejad, H.: Subject-dependent co-occurrence and word sense disambiguation. Proceedings of the 29th annual meeting on Association for Computational Linguistics, University of California, Berkeley ,pp. 146-152 (1991)
7. Mitchell, T.M.: Machine Learning. The McGraw-Hill Companies Inc. and MIT Press, Singapore (1997)
8. Rosario, B.: Extraction of semantic relations from bioscience text. A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Information Management and Systems. University of California, Berkeley. (2005)
9. Song, S-K., Oh, H-S. , Myaeng, S.H., Choi, S-P., Chun, H-W., Choi, Y-S., and Jeong, C-H.: Procedural Knowledge Extraction on MEDLINE. AMT 2011, LNCS 6890, pp. 345–354 (2011).
10. Sudprasert, S., Kawtrakul, A. Thai Word Segmentation based on Global and Local Unsupervised Learning. NCSEC'2003 Proceedings, Chonburi, Thailand , pp.1-8 (2003)