Extraction of Explanation Based Symptom-Treatment Relation from Texts

Chaveevan Pechsiri[1], Uraiwan Janviriyasopak[2]
[1]*Dept. of Information Technology, Dhurakij Pundit University, Bangkok, Thailand.*
[2]*Eastern Industry Co.Ltd.,Bangkok, Thailand.*
itdpu@hotmail.com

***Abstract*— This paper aims to extract the explanation-based Problem-Solving relation, especially the Symptom-Treatment relation, from hospital-web-board documents. The extracted relations benefit people who are learning how to solve their health problems. The research includes three main problems: 1) how to identify symptom-concept EDUs (where an EDU is an elementary discourse unit or a simple sentence/clause) and treatment concept EDUs, 2) how to identify the symptom-concept-EDU boundary and the treatment-concept-EDU boundary as an explanation, 3) how to determine Symptom-Treatment relations from documents. Therefore, we propose collecting each Multi-Word-Co occurrence with either a symptom concept or a treatment concept from a verb-phrase to identify each symptom-concept EDU and each treatment-concept EDU including their boundaries. Collecting Multi-Word-Co involves two more problems of the ambiguous Multi-Word-Co and the Multi-Word-Co size. Thus, we apply the Bayesian Network to solve both problems of Multi-Word-Co after applying word rules. The Symptom-Treatment relation can be solved by Naive Bayes learning vector pairs of symptom vectors and treatment vectors. The research results can provide high precision when extracting Symptom-Treatment relations through texts.**

***Index Terms*— Multi-word-co expression; Problem-solving relation; Symptom vector.**

I. INTRODUCTION

Identifying and extracting Problem-Solving relations, based on the explanation of both problems and solving methods from texts, are very useful for both information retrieval and the Question Answering (QA) system. To extract the Problem-Solving relation, especially a Symptom-Treatment relation between two explanation groups, a disease-symptom/problem group and a treatment/solving-procedure group from documents, is a challenge. Thus, the research focuses on extracting the Symptom-Treatment relation from Thai documents on medical-care consultation edited by patients and professional medical practitioners on the hospital's web-board on a Non-Government-Organization (NGO) website. Both the disease symptoms and the treatments in the medical-care-consulting documents are event explanations of several consequences of events expressed by several verb phrases in several EDUs (where an EDU is an Elementary Discourse Unit, which is a simple sentence/clause defined by [1]). Each EDU is expressed by the following linguistic pattern after stop word removal.

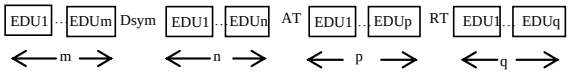
$EDU_{sym} \rightarrow NP1_{sym}\ V_{sym}\ NP2_{sym}$
$V_{sym} \rightarrow V_{weak} | V_{strong}$
$NP1_{sym} \rightarrow$ pronoun $| w_{sym1}$
$NP2_{sym} \rightarrow w_{1\text{-}sym1}\ w_{2\text{-}sym2}\ w_{3\text{-}sym2} \ldots\ w_{ns\text{-}sym2}$
$EDU_{treat} \rightarrow NP1_{treat}\ V_{treat}\ NP2_{treat}$
$NP1_{treat} \rightarrow$ pronoun $| w_{treat1}$
$NP2_{treat} \rightarrow w_{1\text{-}treat1}\ w_{2\text{-}treat2}\ w_{3\text{-}treat2} \ldots\ w_{nt\text{-}treat2}$
$V_{weak} \rightarrow$ {'เป็น/*be*', 'มี/*have*', '*ปรากฏ/occur*'}
$V_{strong} \rightarrow$ {'*คลื่นไส้/nauseate*','*อาเจียน/vomit*','*ปวด/pain*','*เจ็บ/pain*','*แน่น/constrict*','*คัน/itchy*',…}
$V_{treat} \rightarrow$ {'*ใช้/use*','*ทา/apply*','*กิน/consume*','*รักษา/treat*','*ฉีด/inject*',..}

where EDU_{sym} and EDU_{treat} are a symptom concept EDU and a treatment concept EDU, respectively. V_{strong} is a strong verb set with the symptom concept. V_{weak} is a weak verb set which needs more information to determine the symptom concept. V_{treat} is a treatment/procedural verb concept set. NP1 and NP2 are noun phrases

$w_{1\text{-}sym1} \in W_{sym1}$ (where W_{sym1} is a noun word set with symptom concepts); $w_{i\text{-}sym2} \in W_{sym2}$ (W_{sym2} is a word set with symptom concepts, i=2,3,..*ns*, and *ns* is the number of words in $NP2_{sym}$)
$w_{1\text{-}treat1} \in W_{treat1}$ (where W_{treat1} is a noun word set with treatment concepts) , $w_{i\text{-}treat2} \in W_{treat2}$($W_{treat2}$ is a word set with treatment concepts, i=2,3,..*nt*, and *nt* is the number of words in $NP2_{treat}$)

W_{sym1}={' ', '*ผู้ป่วย/patient*-noun','*อาการ/symptom*-noun','*แผล/scar*-noun','*รอย/mark*-noun','*ไข้/fever*-noun','*ผื่น/rash*-noun','*หนอง/pus*-noun',…}
W_{sym2}={'*ยาก/difficultly*-adv','*สี..../…color*-adj','*เป็น/be*-adv','*เหลว/watery*-adj','*คลื่นไส้/nauseate*-verb','*อาเจียน/vomit*-verb','*ปวด/pain*-verb','*แน่น/constrict*-verb','*แผล/scar*-noun','*รอย/mark*-noun','*ไข้/fever*-noun','*ผื่น/rash*-noun','*หนอง/pus*-noun','*อวัยวะ/organ*-noun',..}
W_{treat1}={' ','*ผู้ป่วย/patient*-noun','*ยา/medicine*-noun','*อาหาร/food*-noun','*สมุนไพร/herb*-noun','*รังสี/radiation*-noun',..}
W_{treat2}={'*แก้/reduce*-verb','*ลด/reduce*-verb','*ฆ่า/kill*-verb', *เชื้อ/pathogen*-noun', '*อาการ/symptom*-noun','*ปวด/pain*-verb/ noun','*ไข้/fever*-noun','*ใน/reduce*-verb','*อวัยวะ/organ*-noun',..}

Moreover, there are two kinds of treatment on web-board documents; the actual treatment notified by the patient/user from his experience, and the recommended treatment determined by the professional medical practitioner. 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 follows:
  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

Figure1 shows the Symptom-Treatment relation examples: Dsym → AT and Dsym → RT where Dsym is EDU1-EDU3, AT is EDU5, and RT is EDU8-EDU10.

Topic name: หนูเป็นโรคกระเพาะหรือเปล่า/**Do I get a stomach disease?**
EDU1:"[หนู]*ปวดท้องอย่างหนอ*" ("[หนู/***patient***] *ปวด*/***pain*** *ท้อง*/***stomach*** *อย่างหนอ*/***heavily***")
([*A patient*] *has a stomachache heavily*.)
EDU2: "[หนู]*มีแก๊สมากในกระ เพาะ* "
("[หนู/***patient***] *มี*/***has*** *แก๊ส*/***gas*** *มาก*/***a lots*** *ในกระ เพาะ*/***inside stomach***")
([*The patient*] *has lots of gas in the stomach*.)
EDU3: "*อาการมันจะ เป็นหลังอาหารเย็นแล้ ตอนกลางคืน*" ("*อาการ*/***Symptom*** *มักเป็น*/***mostly occurs*** *หลังอาหารเย็น*/***after dinner***แล้ /***and***ตอนกลางคืน/***night***")
([*The symptom mostly occurs after dinner and at night*.)
EDU4 : "[หนู]*สงสัยเป็นโรคกระ เพาะ* " **([*The patient*] *doubts to get gastropathy*.)**
EDU5: "[หนู]*กินยาลดกรดเพื่อแก้ปวดท้อง*" ("[หนู/***patient***] *กิน*/***consume*** *ยา*/***medicine*** *ลด*/***reduce*** *กรด*/*acid* *เพื่อแก้*/***to solve*** *ปวด*/***pain*** *ท้อง*/***stomach***")
([*The patient*] *takes an antacid to solve the stomach ache*.)
EDU6 : "*แต่มันก็ไม่หายปวด*" ("*แต่*/***But*** *มัน*/***it*** *ก็ไม่หายปวด* / ***cannot work***")
(*But it cannot work*.)

Physician Suggestion

EDU7 : "*ไปหาหมอหรือยัง* /***Have you seen the doctor?***"
EDU8: *ถ้าหนูเป็นโรคกระ เพาะ* " ("*ถ้า* /***If*** [หนู/***patient***] *เป็น*/***get*** *โรคกระ เพาะ* / ***gastropathy***")
(*If* [*the patient*] *gets gastropathy* ,)
EDU9:"[หนู]*ก็อาจต้องกินยาลดการหลั่งกรดในกระ เพาะ อาหาร*" ("[หนู/***patient***] *ก็อาจต้อง*/***may*** *กิน*/***consume*** *ยา*/***medicine*** *ลด*/***reduce*** *การหลั่ง*/ ***secretion*** *กรดในกระ เพาะ อาหาร*/***gastric acid***")
([*the patient*] *may take a medicine to reduce the gastric acid secretion*.)
EDU10: "[หนู]*ควรหลีกเลี่ยงอาหารที่ทำให้เกิดแก๊สในกระ เพาะ* " ("[หนู/***patient***] *ควรหลีกเลี่ยง*/***should avoid*** *อาหาร*/***food*** *ที่ทำให้เกิด*/ ***causing*** *แก๊ส*/***gassy*** *ในกระ เพาะ* /***in the stomach***")
([The patient] *should void food causing gassy in the stomach*.)

Figure 1: Example of Symptom-Treatment Relation where [..] means ellipsis

There are several techniques ([2][3][4][5][6] and[7]) applied to extract either the Symptom-Treatment relation or the disease treatment relation from texts (see Section 2). However, the Thai documents have several specific characteristics, such as zero anaphora or an implicit noun phrase, without word and sentence delimiters, etc. All of these characteristics are involved in the three main problems of extracting the Symptom-Treatment relation from the NGO web-board documents (see Section 3): 1) identifying the symptom-concept EDU and the treatment concept EDU which are the event expressions by verb phrases moderately based on weak verbs, 2) identifying the symptom-concept-EDU boundary as Dsym and the treatment-concept-EDU boundary as AT/RT, and 3) determining the Symptom-Treatment relation from documents. For all of these problems, we need to develop a framework which combines a machine learning technique and the linguistic phenomena to learn the several EDU expressions of the Problem-Solving relation type, i.e. the Symptom-Treatment relations from documents. Therefore, we propose collecting multi-word co-occurrences with either the symptom concepts or the treatment concepts from verb phrases to identify the symptom-concept EDUs and the treatment concept EDUs. Where the multi-word co-occurrence (or 'Multi-Word-Co') is the co-occurrence of two or possibly more N-words; N=2,3, .., *num* and *num* is the number of words per EDU. Each EDU-verb-phrase expression in this research contains a Multi-Word-Co expression as the following expression form with either a symptom concept or a treatment concept after stemming words and stop word removal as shown in Table 1 based on WordNet [8] and Mesh (https://www.nlm.nih.gov/mesh/) after translation from Thai to English by http://www.longdo.com/.

Multi-Word-Co expression = $w_1 + w_2 + w_3 + .. + w_{num}$

The Multi-Word-Co expression is formed with the following word rules.

***w*Rule1**: If $w_1 \in V_{weak}$ then $(w_2 \in W_{sym1})$ and $(w_3, w_4, .., w_{num} \in W_{sym2})$ where $num \le ns$

***w*Rule2**:If $w_1 \in V_{strong}$ then $(w_2, w_3, .., w_{num} \in W_{sym2})$ where $num \le ns$

***w*Rule3**:If $w_1 \in V_{treat}$ then $(w_2 \in W_{treat1})$ and $(w_3, w_4, .., w_{num} \in W_{treat2})$ where $num \le nt$

Table 1
MultiWordCoMetrix with Symptom Concept and Treatment Concept

Multi-Word-Co expression	Symptom Location	Symptom concept	SymCon-ceptID
'เป็น/be ผื่น/rash แดง/red หน้า/face'	face	To occur red rash	S001
'มี/have อาการ/symptom คลื่นไส้/nauseate'	stomach(from 'nauseate'by WordNet)	To occur nauseated symptom	S002
...	...	...	...
Multi-Word-Co expression	-	Treatment concept	TreatCon-ceptID
'กิน/consume ยา/medicine ลด/reduce กรด/acid'	-	To consume an antacid	T001
'เลี่ยง/avoid อาหาร/food ทำให้เกิด/caus- ing แก๊ส/gassy กระเพาะ/stomach'	-	To avoid gassy food	T002
...	-	...	...

Moreover, to collect Multi-Word-Co with either the symptom concept or the treatment concept has two problems of Multi-Word-Co ambiguity and Multi-Word-Co size (see Section 3). Thus, we apply the Bayesian Network (BN)[9] to solve the ambiguity and the size of the Multi-Word-Co expression after applying the word rules to the verb phrase. The Multi-Word-Co expressions with the symptom/treatment concepts are then, determined and collected in terms of M_{sym} (MultiWordCoMatrix with symptom concepts) having four attributes and M_{treat} (MultiWordCoMatrix with treatment concepts) having three attributes (see Table 1 where the symptom/treatment concepts are based on WordNet and MeSH). Moreover, the Multi-Word-Co expressions are also applied to solve Dsym and AT/RT, and the Symptom-Treatment relation can be solved by using the Naive Bayes (NB) [9] with the symptom feature vector or Dsym and the treatment feature vector or AT/RT(see Section 4).

This paper consists of 5 sections. In Section 2, related work is summarized. The research problems are described in Section 3, and Section 4 presents the research framework to extract the Symptom-Treatment relation. In Section 5, we evaluate and conclude our proposed model.

II. RELATED WORKS

Several strategies ([2][3][4][5][6]and[7]) have been proposed to extract the Symptom-Treatment relation or the disease treatment relation from textual data.

Rosario B. [2] extracted the semantic relations from bioscience text. The goals of her work were to identify the semantic roles DIS (Disease) and TREAT (Treament), and to identify the semantic relations between DIS and TREAT in bioscience abstracts. She identified the DIS and TREAT entities 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 had an average accuracy 88.3% in the relation classification. Abacha A. B. and Zweigenbaum P. [3] extracted semantic relations between medical entities (as the treatment relations between a medical treatment and a problem, i.e. disease) by using the linguistic pattern-based method to extract the relation from selected MEDLINE articles.

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

where E1, E2, or Ei is the medical entity identified by MetaMap. Their treatment relation extraction was based on a couple of medical entities or noun phrases occurring within a single sentence. Their results showed 75.72% precision and 60.46% recall. Song S. et al. [4], extracted procedural knowledge from MEDLINE abstracts as shown in the following by using a Supporting Vector Machine (SVM) compared to the 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 a Target and a corresponding solution. SVM and CRF were utilized with the following features: Content in a target sentence, Position, Neighbor, and Ontology as the concepts to classify the Target. In addition the other features to classify the procedures from several sentences were Word, Context, PredicateArgumentStructure, and Ontology. SVM yielded higher precision and higher recall of 0.8369 and 0.7957, respectively. In most of the previous works, i.e. [2] and [3], the treatment relation between the medical treatment and the problem (as the disease) occurs within one sentence whereas our Symptom-Treatment relation occurs within several sentences/EDUs in both the problems/ symptoms and the solving-procedure/treatment-steps. However, [4] had several sentences for the treatment method, but there was only one sentence for the problem as the Target disease or symptom. Therefore, we propose collecting Multi-word Co expressions with either symptom concepts or treatment concepts from verb phrases (after stemming words and eliminating stop words) to identify either the symptom-concept EDUs or the treatment concept EDUs. [5] introduced a syntactic constraint including an intuitive lexical constraint to identify *the relation phrases* expressed by verb phrases in the *verb-noun* combination for the Open Information Extraction system. The lexical constraint is used to extract relation phrases, i.e. "Faust made a deal with the devil." is extracted as "Faust, *made a deal with*, the devil" instead of "Faust, made, a deal". Their relation phrase identification was 80% precision. [6] proposed using the positive (Harmless) probability of each word co-occurrence in a certain sentence from a Social Network Service for filtering harmful sentences. The research achieved greater than 90% precision for three-Word-Co and lower than 50% precision for two-Word-Co. [7] learned the causal relation from verb-noun pairs of verb phrases by applying Integer Linear Programming with FrameNet, WordNet and linguistic features, i.e."*People died in hurricane*" had '*hurricane*-noun' and '*die*-verb' as the causal relation. They achieved a 14.74% F-score.

The previous researches [5] and [7] worked on verb phrases with two-Word-Co of the 'verb-noun' to determine relations and [6] worked on two/three-Word-Co to filter harmful sentences. However, our research focuses on determining and collecting Multi-Word-Co with two problems, the Multi-Word-Co ambiguity and the Multi-Word-Co size, solved by BN. The Multi-Word-Co collection is also applied to solve the symptom-concept-EDU boundary (the symptom concept vector, Dsym) and the treatment-concept-EDU boundary (the treatment concept vector, AT/RT) which are used to determine the Symptom-Treatment relation by NB.

III. RESEARCH PROBLEMS

There are three main problems in identifying a problem/ symptom-concept EDU and a solving/treatment-concept EDU, determining Dsym and AT/RT, and determining Symptom-Treatment relations from documents.

A. How To Identify a Symptom Concept EDU and a Treatment Concept EDU

According to the medical care domain, most of the symptom concept EDUs and the treatment concept EDUs are expressed as verb phrases. For example:

Symptom Concept

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

(b) EDU:"*ผู้ป่วย มีอาการปวดศีรษะ*"
("*ผู้ป่วย*/*A patient* *มี*/***have*** *อาการ*/***symptom*** *ปวด*/***pain*** *ศีรษะ*/***head***")
(*A patient has a headache symptom*.)

Treatment Concept

(c) EDU: "[*ผู้ป่วย*]*กินยาลดกรด*"
("[*ผู้ป่วย*/*A patient*]*กิน*/***consume*** *ยา*/***medicine*** *ลด*/***reduce*** *กรด*/***acid***")
([*A patient*] *takes an antacid*.)

where [..] means ellipsis. However, some verb phrases of the symptom concepts are ambiguous. For example:

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

(f) EDU1: "*ห้องน้ำสกปรกมาก*" ("*ห้องน้ำ*/***toilet*** *สกปรก*/***be dirty*** *มาก*/***very***.")
(the *toilet is very dirty*.)
EDU2: *ฉันจึงถ่ายยาก*" ("*ฉัน*/***I*** *จึง*/***then*** *ถ่าย*/***defecate*** *ยาก* /***difficultly***")
(*Then, I defecate with difficultly*.)

From (e) and (f), 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 determining and collecting the Multi-Word-Co with either the problem/symptom concept or the solving/treatment concept after stemming words and eliminating stop words from the health-care documents. However, to determine these Multi-Word-Co expressions for collection involves more problems of ambiguous Multi-Word-Co and the various sizes of the Multi-Word-Co expressions as follows.

Ambiguous Multi-Word-Co

(a) "((*ผื่น*/*rash*)/noun)/NP1
((*เป็น*/***be***)/verb (*เม็ด*/***bumps***)/noun (*สีน้ำตาล*/***brown***)/Adj)/VP"
Multi-Word-Co='*เป็น*/***be***-verb *เม็ด*/***bumps***-noun *สีน้ำตาล*/***brown***-Adj'

(b) "((*ไฝ*/*mole*)/noun)/NP1
((*เป็น*/*be*)/verb (*เม็ด*/*bumps*)/noun (*สีน้ำตาล* /*brown*)/Adj)/VP"
Multi-Word-Co='*เป็น*/***be***-verb *เม็ด*/***bumps***-noun *สีน้ำตาล*/***brown***-Adj'

Thus, the VP of (a) contains a Multi-Word-Co with symptom concepts whereas the VP of (b) contains Multi-Word-Co with the property concept of NP1 or 'mole'. This problem can be solved by applying the word rules with ***w*Rule1**/***w*Rule2** adjustment after the stop word removal as follows:

***w*Rule1**: If $w_1 \in V_{weak} \wedge AnyWordOfNP1 \in W_{sym1}$ then ($w_2, w_3, w_4, .., w_{num} \in W_{sym2}$) where $num \leq ns$
Else If $w_1 \in V_{weak} \wedge w_2 \in W_{sym1}$ then ($w_3, w_4, .., w_{num} \in W_{sym2}$) where $num \leq ns$

***w*Rule2**:If $w_1 \in V_{strong} \wedge AnyWordOfNP1 \in W_{sym1}$then ($w_2, w_3, .., w_{num} \in W_{sym2}$)

Various Sizes of Multi-Word-Co Expressions

(c)VP= '(*ปวด*/***pain***)/*verb* (*หัว*/***head***)/*noun*' ('***have a headache***')
Multi-Word-Co = '*ปวด*/***pain***-verb *หัว*/***head***-/*noun*'
SymptomConcept = 'To have a headache'

(d)VP='(*เป็น*/*be*)/*verb* (*เม็ด*/*bumps*)/*noun* (*พอง*/*blister*)/*noun* (*น้ำ*/*watery*)/*Adj*(*จำนวนมาก*/*a lot*)/*Adj*' ('***be lot of watery blister bumps***')
Multi-Word-Co='*เป็น*/***be***-verb *เม็ด*/***bump***-noun *พอง*/***blister***-verb *น้ำใส*/***watery***-Adv'
SymptomConcept = 'To occur watery blister bump'

The Multi-Word-Co expressions from (a) to (d) vary in terms of the number of words, which results in an algorithm to determine the symptom-EDU occurrences or the treatment-EDU occurrences. This problem can be solved by BN learning Multi-Word-Co with the symptom concept or the treatment concept by sliding the window size of two consecutive words with a sliding distance of one word in a verb phrase after stemming words and eliminating stop words.

B. How to Determine Dsym and AT/RT

According to Figure1, there is no clue (i.e. '*และ*/*and*', '*หรือ*/*or*', ..) in both EDU3 to identify the symptom boundary (EDU1-EDU3) and EDU10 to identify the treatment boundary (EDU8-EDU10). Therefore, we use the collected Multi-Word-Co expressions to solve both boundaries.

C. How to Determine the Symptom-Treatment Relation

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

(a) $EDU1_{sym}$: "*ผู้ป่วยปวดท้องอย่างหนัก*"
(***A patient has a bad stomachache***.)
$EDU2_{sym}$: "[*ผู้ป่วย*] *มีแก๊สในกระเพาะมาก*"
(**[*The patient*] *has lots of gas in the stomach*.**)
$EDU3_{treat}$: "[*ผู้ป่วย*]*กินยาลดกรด*" (**[*The patient*] *takes an antacid*.**)
EDU4: "*แต่ก็ไม่หายปวด*" (***But it does not work*.**)

(b) $EDU1_{sym}$: **"*ผู้ป่วยปวดท้อง*" (*A patient has a stomachache*.)**
$EDU2_{sym}$:[*ผู้ป่วย*] *มีแก๊สในกระเพาะ*
(**[*The patient*] *has gas in the stomach*.**)
$EDU3_{treat}$:[*ผู้ป่วย*] *กินยาลดกรด* (**[*The patient*] *takes an antacid* .**)
EDU4:[*ผู้ป่วย*] *รู้สึกดีขึ้น* (***[The patient] feels better***.)

Thus, the Symptom-Treatment relation occurs only in (b) because the EDU4 of (b) contains "*รู้สึกดีขึ้น*/*Feel better*" as the Class-cue-word of the Symptom-Treatment relation. Therefore, we apply NB learning the Symptom-Treatment relation with two feature vectors of a Symptom vector, $\langle s_1, s_2, .. s_a \rangle$(where s_j is a symptom concept id on Table1; j=1,2,..,a; a is the number of EDUs on Dsym), and a treatment vector, $\langle t_1, t_2, .. t_y \rangle$ (where t_l is a treatment concept id on Table1; l=1,2,..,y; y is b or c; b is the number of EDUs on AT; c is the number of EDUs on RT).

IV. Research Framework

There are several steps in our framework: Corpus Preparation, Multi-Word-Co Size/boundary Learning, Multi-Word-Co Expression Determination, Dsym and AT/RT Determination, Symptom-Treatment relation Learning, and Symptom-Treatment Relation Extraction as shown in Figure 2.

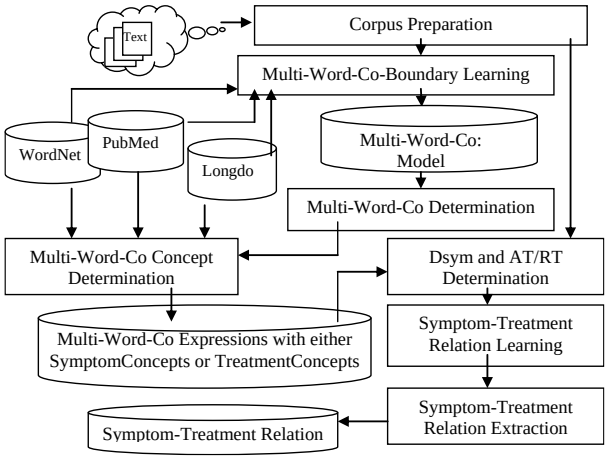


Figure 2: System Overview

A. 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 using Thai word segmentation tools [10], including Name entity [11] followed by EDU segmentation [12] as an EDU corpus. The corpus contains 3000 EDUs of gastrointestinal tract diseases and childhood diseases. The corpus is separated into 2 parts: 2000 EDUs for learning the size/boundary of the Multi-Word-Co expression with either the symptom concept or the treatment concept and the symptom-Treatment relation; and 1000 EDUs for Multi-Word-Co determination and the symptom-Treatment relation extraction. This step also includes semi-automatic annotation of the Multi-Word-Co concepts of the symptoms or treatments as shown in Figure3. All word concepts of Multi-Word-Co are referred to Wordnet(http://word-net.prince ton.edu/obtain) and MeSH after translating from Thai to English, by using Lexitron (the Thai-English dictionary) (http://lexitron.nectec. or.th/).

Disease Topic : โรคเกี่ยวกับทางเดินอาหาร **/ Gastrointestinal tract disease**
EDU1: *ผู้ป่วยมีอาการจุกเสียดอย่างมาก*
ผู้ป่วย/A patient มี/has อาการ/symptom จุกเสียด/be colic อย่างมาก/badly
<EDU1>
(*ผู้ป่วย/A patient*-ncn)/NP1
(<**MultiWordCo** Concept=symptom location= intestinal from WordNet of 'colic'>
< **w1:** setType='weak-verb' ; concept='has/occur' boundary ='y'>*มี*</ **w1**>
< **w2:** setType='Noun3' ; concept= 'symptom' boundary ='y'>*อาการ*</**w2**>
< **w3:** setType='strong-verb' ; concept=' be colic' boundary ='y'>*จุกเสียด*</**w3**>
< **w4:** setType='Adv' ; concept= 'badly' boundary= 'n'>*อย่างมาก*</**w4**>
</**MultiWordCo**>)/VP </EDU1>

Figure 3: Multi-Word-Co annotation

B. Multi-Word-Co Size Learning

BN represents the joint probability distribution by specifying a set of conditional independence assumptions (represented by a directed-acyclic graph), together with sets of local conditional probabilities. For each node variable, the arcs represent the variable which is conditionally independent of its non-descendants in the network given its immediate predecessors. The conditional probability table is given for each variable, describing the probability distribution for that variable given the values of its immediate predecessors. The joint probability for the values $\langle y_1,...,y_n\rangle$to the tuple-network variables $\langle Y_1,..,Y_n\rangle$ can be computed by Equation (1).

$$P(y_1...,y_n)=\prod_{i=1}^{n}P(y_i \mid Parents(Y_i)) \quad (1)$$

where Y_0 *is the parents of* Y_1 *, and Parents* (Y_i) denotes the set of immediate predecessors of Y_i in the network. The values of $P(y_i|Parents(Y_i))$ are the values stored in the conditional probability table associated with node Y_i.

However, Equation (1) is applied to the Multi-Word-Co size/boundary determination with $\langle Y_1,..,Y_n\rangle$ as the consequence of word set after stemming words and eliminating stop words, $\{w_1 \ldots w_n\}$,where Y_0=Disease Topic from the document name. Each word, w_i (where i=1..n), is a consequence word concept where $w_1 \in V_{weak} \cup V_{strong} \cup V_{treat}$; $n = num$; i=2,3,..,num; $w_i \in W_{sym1} \cup W_{sym2} \cup W_{treat1} \cup W_{treat2}$. All annotated concepts of w_i (which is w_i) are features used in determining the conditional probabilities of consequence words as shown in Table 2. According to Table2, it can be concluded that the least probability of $P(w_i|w_1,..w_{i-1})$ is 0.00714 as the Multi-Word-Co Boundary threshold with an actual Multi-Word-Co Boundary threshold of 0.005 to determine the size or boundary of the Multi-Word-Co expression with the symptom concepts or the treatment concepts on the health care corpus as shown in the following rule:

MultiWordCoBoundary Rule

IF $P(w_i \mid w_{i-1},..w_2 w_1) < MWC\_Threshold$ THEN
$Boundary\ is\ Found with\ MultiWordCoBoundary = \{w_1..w_i\}$

where the *MWC_Threshold* is the actual Multi-Word-Co Boundary threshold, 0.005, and w_i= a consequence word concept after stemming words and the stop word removal.

C. Multi-word-Co Determination

After stemming words and stop word removal, the first word of the Multi-Word-Co expression is identified by the word rule with the ***w*Rule1/*w*Rule2** adjustment. The Multi-Word-Co boundary is determined by using the MultiWordCoBoundary rule.

D. Multi-word-Co Concept Determination

The concept of multi-Word-Co can be determined by w_1 as the main verb concept ($V_{main} = V_{strong} \cup V_{weak} \cup V_{treat}$). If $w_1 \in V_{weak}$ then the symptom concept is defined by ($w_2 \in W_{sym1}$) $\wedge$ (w_3, w_4,.., $w_{num} \in W_{sym2}$). If $w \in V_{strong}$ then the symptom concept is defined by V_{strong}. If $w_1 \in V_{treat}$ then the treatment concept is defined by ($w_2 \in W_{treat1}$) $\wedge$ ($w_3,w_4,..,w_{num} \in W_{treat2}$). The location of the symptom can be solved by either the w_i concept of '*อวัยวะ/organ*' or the V_{strong} concept from WordNet, i.e.'*nauseate*-verb' having the 'stomach' location by WordNet. All Multi-Word-Co expressions are collected and sorted into M_{sym} and M_{treat} after determining their concepts and locations (Table 1).

E. Dsym and AT/RT Determination

Assume that each EDU is represented by (NP VP). **L** is a list of EDUs in corpus.
Dsym_AT/RT_DETERMINATION(L)

```
1  i=1; SymptomVector ← ∅; TreatmentVector ← ∅
2    match=true;
3    While  match=true ∧ i≤ length[L]
4    { Determine MWCo1i of EDUi after StopWordRemoval
5      Determine matching_score  between MWCo1i
                    and MWCo2k of Msym or Mtreat
6        Equation (2)
7      If matching_score >= .9  then
8       {match=True;
9        If ConceptIDk is SymConcept then
10          SymptomVector←SymptomVector∪ConceptIDk ;
11       ElseIf ConceptIDk is TreatConcept then
12       TreatmentVector←TreatmentVector∪ConceptIDk;
13      }Else  match=false;
14       i++ ;
15     }  Return SymptomVector or TreatmentVector
```

Figure 4: Symptom Vector and Treatment Vector Determination Algorithm

Table 2
The sequence of wi concepts appearing in documents

w_1	$P(w_1)$	w_2	$P(w_2\|w_1)$	w_3	$P(w_3\|w_2, w_1)$	w_4	$P(w_4\|w_3,w_2,w_1)$
มี/*have*		อาการ/*symptom*		อักเสบ/*inflame*	0.02857		
มี/*have*		อาการ/*symptom*		……...	……..		
มี/*have*		อาการ/*symptom*	0.07143	ไอ/*cough*	0.01429		
มี/*have*		ไข้/*fever*		สูง/*high*	0.02857		
มี/*have*		ไข้/*fever*	0.04286				
มี/*have*		ผื่น/*rash*		แดง/*red*	0.01429	หน้า/*face*	0.01429
มี/*have*		ผื่น/*rash*	0.02857				
มี/*have*	0.32857	…..	…..		…..		
เป็น/*be*		ตุ่ม/*bump*		พอง/*blister*	0.01429	น้ำ /*water*	0.01429
เป็น/*be*	0.27143	ผื่น/*rash*		แดง/*red*		อก/*chest*	0.00714
เป็น/*be*	0.27143	ผื่น/*rash*	0.01429	แดง/*red*	0.01429		
….	….	….	….	….	….		

M_{sym} and M_{treat} from the previous step are used to determine SymptomVector or Dsym and TreatmentVector or AT/RT from the EDUs of the tested corpus.

SymptomVector ,$\langle s_1,s_2,..s_a\rangle$, has $s_j \in$ SymConceptID which is the symptom-conceptID set on Table 1 and j=1,2,..a. TreatmentVector ,$\langle t_1,t_2,..t_y\rangle$, has $t_l \in$ TreatConceptID which is the treatment-conceptID set on Table 1 and l=1,2,..y. SymptomVector and TreatmentVector can be determined by the algorithm in Figure 4 based on the highest similarity score on Equation (2) between $MWCo1_i$ (is Multi-Word-Co of EDU_i after stemming words and stop word removal) and $MWCo2_k$ (is Multi-Word-Co_k of M_{sym} or M_{treat}). And $MWCo1_i$ matches $MWCo2_k$ if this highest similarity score is greater than 0.9.

$$matching\_score = ArgMaxSimilarity_{k=1}^{numTuple}\left(\frac{|MWCo1 \cap MWCo2_k|}{\sqrt{|MWCo1| \times |MWCo2_k|}}\right) \quad (2)$$

where numTuple is the number of tuples of M_{sym} *or* M_{treat}

F. Symptom-Treatment Relation Learning

Two feature vectors of SymptomVector and Treatment Vector from Section 4E are used to learn the Symptom-Treatment relation along with the Class-cue-word pattern occurrence in the learning corpus. The Class-cue-word pattern shown in the following contains the Class-type set {"yes", "no"} of the symptom-Treatment relation.

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',..}

Dsym and AT/RT are represented by the $\langle s_1,s_2,..s_a\rangle$ vector and the $\langle t_1,t_2,..t_y\rangle$ vector respectively which are used in determining the probabilities of Symptom-Treatment relation(*class*='yes'; *class*∈Class) and non-Symptom-Treatment relation (*class*='no') from $P(s_1|class)$,$P(s_2|class)$,..,$P(s_a|class)$, $P(t_1|class)$, $P(t_2|class)$,..,$P(t_y|class)$ by using Weka (http://www.cs.wakato.ac.nz/ml/weka/).

G. Symptom-Treatment Relation Determination

According to the conditional probabilities, s_i of Dsym and t_j of AT/RT from the learning step (section 4.*F*), the Symptom-Treatment relation can be determined by Equation (3).

$$\begin{aligned} SymTreat\_RelClass &= \arg\max_{class \in Class} P(class|s_1,s_2,..s_a,t_1,t_2,..,t_y) \\ &= \arg\max_{class \in Class} P(s_1|class)P(s_2|class)...P(s_a|class) \\ &\quad P(t_1|class)P(t_2|class)...P(t_y|class)P(class) \end{aligned} \quad (3)$$

where s_i is a symptomconceptid; t_j is a treatment conceptid
a is the number of EDUs on Dsym; y is b or c;
b is the number of EDUs on AT; c is the number of EDUs on RT;
Class = {"*yes*","*no*")

V. Evaluation And Conclusion

The Thai corpus used to evaluate the extraction of explanation based relations, especially the Problem-Solving relation as the Symptom-Treatment relation, consist of 500 EDUs of gastrointestinal tract diseases and 500 EDUs of childhood diseases, collected from the hospital's web-boards for medical-care-consulting. The research performance is based on two evaluations: determining the Multi-Word-Co expression with the symptom/treatment concepts from documents and extracting the Symptom-Treatment relation from documents. Both evaluations are expressed in terms of precision and recall based on three experts with max-win voting.

Table 3
Evaluation of Multi-Word-Co Determination

Disease Type	Correctness of multi-Word-Co Determination		Symptom-Treatment Relation Extraction	
	Precision	*Recall*	*Precision*	*Recall*
GastrointestinalTract	91.4%	60.5%	90.1%	76.4%
Childhood diseases	89.2%	65.1%	87.5%	73.2%

The average precision in determining the Multi-Word-Co expressions with the symptom/treatment concepts is 90.3% with an average recall of 62.8% as shown in Table3. The reason for the low recall is the anaphora problem, i.e. '*สิ่ง*/***something***)/*pronoun*' as shown in the following:

VP='(*รู้สึก*/***feel***)/*pre-verb* (*มี*/***have***)/*weak-verb* (*บางสิ่ง*/***something***) /*pronoun* (*ข้างใน*/***inside***)/*prep* (*จมูก*/***nose***)/*noun* (*ระหว่าง*/***during***)/*prep* (*เวลาเช้า*/***morning***)/*noun*'

('***have something inside a nose during the morning***')

Multi-Word-Co=('***have ? nose morning***')

The average precision of the extracted Symptom-Treatment relation based on Multi-Word-Co expression is 88.8% and the average recall is 74.8%. The interrupt occurrences on the corpus cause the Symptom-Treatment relation extraction to have a low recall as shown in the following.

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

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

EDU3: *ได้ผล* (***It works***)

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

where EDU3 is an interrupt to the treatment-concept-EDU boundary (EDU2 and EDU4). Hence, the extraction of the explanation based Symptom-Treatment relation in this research is very beneficial not only application by ordinary people as to know how to solve their health problems through the QA system, but also for application by professional people in other areas, i.e. solving industrial finance problems.

REFERENCES

[1] L. Carlson, D. Marcu, and M. E. Okurowski, *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory*. In Current Directions in Discourse and Dialogue, 22: 85-112 2003.

[2] B. Rosario, *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, pp. 71-131, 2005.

[3] A. B. Abacha, and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach", J. of Biomedical Semantics, retrieved December, 20, from http://www.jbiomedsem .com/content/2/S5/S4, 2011.

[4] S-K.Song, H-S. Oh, S.H.Myaeng, S-P.Choi, H-W.Chun, Y-S.Choi, and C-H.Jeong, Procedural Knowledge Extraction on MEDLINE. AMT2011,LNCS6890, pp. 345–354, 2011.

[5] A. Fader, S. Soderland, and O. Etzioni, Identifying Relations for Open Information Extraction. In Proceedings of Empirical Methods in Natural Language Processing, 2011, pp. 1535-1425.

[6] S. Ando, Y. Fujii, and T. Ito., "Filtering Harmful Sentences based on Multiple Word Co-occurrence", in *Proc. 9th International Conference on Computer and Information Science*, 2010, pp. 581 – 586.

[7] M. Riaz and R. Girju, "Recognizing Causality in Verb- Noun Pairs via Noun and Verb Semantics", In *Proc. of the EACL2014 Workshop on Computational Approaches to Causality in Language*, 2014, pp. 48-57.

[8] G. Miller, "WordNet:Alexical database", *Communications of ACM*, vol. 38, no. 11, pp. 39 – 41, 1995.

[9] T. M. Mitchell, *Machine Learning*. The McGraw-Hill Companies Inc. and MIT Press, Singapore, pp. 154-199, 1997.

[10] S. Sudprasert and A. Kawtrakul, Thai Word Segmentation based on GlobalandLocalUnsupervisedLearning.NCSEC'2003Proceedings,pp1-8.

[11] H. Chanlekha and A. Kawtrakul, "Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information", in *Proc. IJCNLP'* 2004, pp. 1-7.

[12] J. Chareonsuk, T.Sukvakree, and A. Kawtrakul, "Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information", in *Proc. NCSEC* 2005, pp. 85-90.