J Inf Process Syst, Vol.13, No.5, pp.1103~1125, October 2017
https://doi.org/10.3745/JIPS.04.0046

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Extraction of ObjectProperty-UsageMethod Relation from Web Documents

Chaveevan Pechsiri*, Sumran Phainoun*, and Rapeepun Piriyakul**

Abstract

This paper aims to extract an ObjectProperty-UsageMethod relation, in particular the HerbalMedicinalProperty-UsageMethod relation of the herb-plant object, as a semantic relation between two related sets, a herbal-medicinal-property concept set and a usage-method concept set from several web documents. This HerbalMedicinalProperty-UsageMethod relation benefits people by providing an alternative treatment/solution knowledge to health problems. The research includes three main problems: how to determine EDU (where EDU is an elementary discourse unit or a simple sentence/clause) with a medicinal-property/usage-method concept; how to determine the usage-method boundary; and how to determine the HerbalMedicinalProperty-UsageMethod relation between the two related sets. We propose using N-Word-Co on the verb phrase with the medicinal-property/usage-method concept to solve the first and second problems where the N-Word-Co size is determined by the learning of maximum entropy, support vector machine, and naïve Bayes. We also apply naïve Bayes to solve the third problem of determining the HerbalMedicinalProperty-UsageMethod relation with N-Word-Co elements as features. The research results can provide high precision in the HerbalMedicinalProperty-UsageMethod relation extraction.

Keywords

Medicinal Property, N-Word-Co, Semantic Relation, Usage-Method

1. Introduction

The objective of this research is to extract an ObjectProperty-UsageMethod relation, especially a HerbalMedicinalProperty-UsageMethod relation of an herb-plant object, from downloaded documents from several websites. The downloaded document contents comprise the object names (i.e., the herb plant names) as the topic names and the explanation of several kinds of property (i.e., physical properties, chemical properties, and medicinal properties) and the methods of usage of the objects. The explanation content for herb plants is indigenous knowledge about curing certain diseases effectively even though some disease treatments by medicinal plants are time consuming. However, the result of searching for the herb plant knowledge on both the medicinal properties and usage methods from the web-sites to solve health problems is a list of documents that the user has to read in order to extract the required knowledge. Therefore, it is necessary to automatically extract the HerbalMedicinalProperty-

Manuscript received March 8, 2017; first revision June 27, 2017; accepted June 28, 2017.
Corresponding Author: Chaveevan Pechsiri (chaveevan.pec@dpu.ac.th)
* College of Innovative Technology and Engineering, Dhurakijpundit University, Bangkok, Thailand (chaveevan.pec@dpu.ac.th, sumran.pha @dpu.ac.th)
**Dept. of Computer Science, Ramkhamhaeng University, Bangkok, Thailand (rapepunnight@yahoo.com)

UsageMethod relation from the documents on the web-pages. The ObjectProperty-UsageMethod /HerbalMedicinalProperty-UsageMethod relation is a semantic relation between two related sets; a property-concept set of an object, i.e. a medicinal-property concept set of an herb-plant object, and a usage-method concept set (where a usage-method concept is the procedure concept of using the object).

In addition to the research, a usage-method occurrence on a herb-plant document mostly consists of herbal-preparation procedure concept and a treatment procedure concept whilst both semantics/ concepts and relations are the foundation of the knowledge structures [1] which are necessary for the search engine and the reasoning and inference in the information retrieval, question answering, text summarization, and problem-solution applications. From another view point, the extracted semantic relations from the unstructured documents are collected to provide the core information of the web documents for supporting the information retrieval, question answering, text summarization, and problem-solution applications. Thus, the extracted semantic relation, in particular, the HerbalMedicinalProperty-UsageMethod relation, benefits people by providing alternative treatment/solution knowledge to health problems and also benefits the problem-solution system. The occurrences of both the herbal medicinal property concept and the procedure-step concept of the usage-method concept on the documents are mostly the event expressions on EDUs (where an EDU is an Elementary Discourse Unit which is a simple sentence or a clause [2]). According to [1], the extracted HerbalMedicinalProperty-UsageMethod relation from the documents can be represented as <*hmp*> ----(relation)---- <*um*> where the '<..>' and '(..)' symbols represent a concept and a relation type, respectively; $hmp \in$ HMP, $um \in$ UM, HMP is a herbal-medicinal-property concept set, and UM is a usage-method concept set. Each medicinal-property concept and each procedure-step/usage-method-step concept of the research are the event expressions by EDU verb phrases (VP_{EDUs}) as shown in the following.

$HMP=\{VP_{EDUmp\text{-}1}, VP_{EDUmp\text{-}2}, .., VP_{EDUmp\text{-}a}\}$;

$UM = \{M_1, M_2, .. , M_b\}$;

$M_l = VP_{EDUum\text{-}1} + VP_{EDUum\text{-}2} + .. + VP_{EDUum\text{-}c}$;

(relation) is (HerbalMedicinalProperty-UsageMethod) as the HerbalMedicinalProperty-UsageMethod relation type.

where $VP_{EDUmp\text{-}j}$ is *hmp* or a VP_{EDU} with the herbal-medicinal-property concept; j=1,2,..,a ; and a is the number of HMP's EDUs.

M_l is *um* or a usage-method concept; l=1,2,..,b; and b is the number of M_l of UM

$VP_{EDUum\text{-}k}$ is a VP_{EDU} with a procedure-step concept; k=1,2,..,c; and c is the number of procedure-step concepts of M_l

Each EDU is expressed by the following general linguistic expression after stemming words and eliminating stop words:

EDU → NP1 VP | VP

VP → verb NP2 | verb adv | verb $AdvPhrase_{dose}$

verb → $Verb_{weak}$-noun2| $Verb_{weak}$-noun2 verb| $Verb_{strong}$| $Verb_{strong}$ verb

NP1 → pronoun | Noun1 | Noun1 modify

NP2 → Noun2 | Noun2 modify

modify →Adj| Adj modify| Noun1 modify |Noun2 modify| Unit

$AdvPhrase_{dose}$ → {'*วันละ...../....per day*', …}

$Verb_{weak}$ → {'*เป็น/be*', '*มี/have*', '*ใช้/use*', '*นำ/take*', '*เอา/get*'}

$Verb_{strong}$→{'*แก้,รักษา/cure,treat*','*ห้าม/stop*','*ลด/reduce*','*บรรเทา/relieve,remody*','*ขับ/expel,release*', '*บำรุง/enrich*','*ปวด/pain*','*บวม/swell*','*อาเจียน/vomit*','*จุกเสียด/be-colic*','*ถ่าย/defecate*',…,'*บด,ขยี้/grind*', '*ทุบ/pound*', '*ผสม/mix*', '*เติม/add*', '*เคี่ยว/stew*', '*ชง/brew*', '*ต้ม/boil*', '*กรอง,แยก/separate*', '*ตาก/dry*','*คั้น/squeeze*',…, '*ดื่ม,กิน/consume*','*พอก/apply*','*ทา/apply*','*สูดดม/sniff*',…}

Noun1 → {'*ส่วนของพืช/plant part*','*สมุนไพร/herb*',…}

Noun2→ {' ', '*อาการ/symptom*','*ตระคริว/contraction*', '*แผล/ scar*', '*พุพอง/blister*', '*ผื่น/ rash*', '*สี../..color*', '*อวัยวะ/human organ*', '*ยา/medicine*', '*ส่วนผสม/ingredient*', '*ส่วนของพืช/plant part*', '*สมุนไพร/herb*',…}

Adv → {'*ยาก/ difficultly*', '*เหลว/ liquidly*', '*ละเอียด/ thoroughly*', '*แน่นท้อง/ uncomfortably*', …}

Adj → {'*ช้ำ/ bruised*', '*แน่นท้อง/ uncomfortable*',…}

Unit → ('*mกรัม/gm.*','*mกำมือ/handful*','*mใบ/leaves*','*mดอก/flowers*',…);

where NP1 and NP2 are noun phrases, VP is a verb phrase, and include the following sets: $Verb_{strong}$ is a strong-verb concept set, $Verb_{weak}$ is a weak-verb concept set, Adv is an adverb concept set, Noun1 and Noun2 are noun concept sets, Adj is an adjective concept set, and $AdvPhrase_{dose}$ is an adverb phrase with the dosage concept set. All concepts of these concept sets are based on WordNet [3] and Thai Encyclopedia (http://kanchanapisek.or.th/kp6/sub/book/book.php?book=14&chap=10&page=chap10.htm) after translating from Thai to English by Lexitron (https://dict.longdo.com).

For example 1: "*ขิง* / ***Ginger***"

EDU1:"(*ขิง/ginger*)/NP1 (*เป็น/is สมุนไพร/herb แก้/to-treat จุกเสียด/be-colic*)/VP"
(*Ginger is an herb for treating colic*.)

EDU2:"([*Ginger*])/NP1 (*บรรเทา/relives อาการ/symptom แน่นท้อง/uncomfortable*)/VP"
([*Ginger*] *relives the abdominal distension symptom*)

EDU3: "([*Ginger*])/NP1 (*ขับ/ release ลม/gas*)/VP"
([*Ginger*] *release gas*.)

EDU4: "(*นำ/take เหง้าขิง/ginger-rootstock แก่/old* 5*กรัม/5gm.*)VP"
(*Take about 5 gm old ginger rootstock*.)

EDU5:"(*ทุบ/pound* [*ginger rootstock*] *ให้แตก/to crash*)/VP"
(*Pound* [*ginger rootstock*] *to crash*)

EDU6: "(*ต้ม/boil* [*ginger rootstock*] *น้ำ/water*)/VP"
(*Boil* [*ginger rootstock*] *in water*.)

EDU7:"(*ดื่ม/drink*[*solution*] *ระหว่างอาหารแต่ละมื้อ/ during each meal*)/VP"
(*Drink* [*solution*] *during each meal*.)

EDU8:"*หรือ/or* (*ใช้/use ขิงผง/ginger-powder* 1*ช้อนโต๊ะ/1table spoon ginger powder*)/VP"
(*Or use1table spoon of ginger powder*.)

EDU9: "(*ต้ม/boil* [*ginger powder*] *น้ำ/water*"
(*Boil* [*ginger powder*] *in water*.)

EDU10: "(*ดื่ม/drink* [*solution*])/VP"
(*Drink* [*solution*].)

where the [..] symbol means the ellipses of words in this symbol. EDU1–EDU3 have the herbal-medicinal-property concepts. EDU4–EDU7 and EDU8–EDU10 have the usage-method-step concepts with EDU4–EDU6 and EDU8–EDU9 as preparation-procedure-step concepts and EDU7 and EDU10 as the treatment-procedure-step concepts.

There are several techniques [4-9] that have been used to extract the semantic relations involved with the object properties and the procedural knowledge based on explanations from documents (see Section 2) whilst the other previous researches [10] on the semantic relation determination from texts relies on the relations, i.e. *is-a*, *part-of*, and *cause-effect*, between two entities of noun phrases without boundary consideration. The ObjectProperty-UsageMethod relation in our research is extracted from the downloaded Thai documents of the RSPG website (which is the website of Plant Genetic Conservation Project Under the Royal Initiation of Her Royal Highness Princess Maha Chakri Sirindhorn, http://www.rspg.or.th/plants_data/herbs/herbs_200.htm) and the GoToKnow website (https://www.gotoknow.org/posts/339687). However, the Thai documents have some specific characteristics, such as zero anaphora or implicit noun phrases, without word and sentence delimiters, etc., as shown in Example 1. Where both herbal-medicinal-property concept EDUs and procedure-step concept EDUs are the event expressions by verb phrases with the $Verb_{strong}$ occurrences (as shown in EDU2–EDU3, EDU5–EDU7, and EDU9–EDU10) and the $Verb_{weak}$ occurrences (as shown in EDU1, EDU4, and EDU8). All of these characteristics are involved in three main problems in extracting the ObjectProperty-UsageMethod relation from the downloaded documents (see Section 3): the first problem is how to determine an EDU with a verb phrase and the medicinal-property concept (VP_{EDUmp}) or the usage-method-step/procedure-step concept (VP_{EDUum}). With regard to VP_{EDUs}, some verb phrases contain the $Verb_{weak}$ elements which require other words to provide the certain concepts, i.e., an object-property concept, and a usage-method-step concept. Thus we propose using a word co-occurrence of N-Words (or N-Word-Co; N is the number of co-occurred words) with the certain concepts, either medicinal-property concepts or usage-method-step concepts, to solve the first problem whilst the N-Word-Co size or the N value is determined by maximum entropy (ME), support vector machine (SVM), and naïve Bayes (NB) [11] learning from VP after stemming words and eliminating stop words. N-Word-Co is then extracted and collected to solve this first problem. The second problem is how to determine the M_l element of UM, with the boundary consideration. In particular, there are some non-usage-method-step concept EDUs mingled in the M_l boundary. The extracted N-Word-Co expressions are also applied to solve the M_l boundary. Moreover, the N-Word-Co expression is also used to represent the $VP_{EDUmp\text{-}j}$ or $VP_{EDUum\text{-}k}$ occurrence. The third problem is how to determine the ObjectProperty-UsageMethod/HerbalMedicinalProperty-UsageMethod relation between the HMP element and the UM element. We then apply naïve Bayes [11] with the Cartesian product between HMP and UM to determine the ObjectProperty-UsageMethod/HerbalMedicinalProperty-UsageMethod relation where a $VP_{EDUmp\text{-}j}$ element of HMP is represented by an N-Word-Co with a medicinal-property concept, and an M_l element of UM is represented by a vector of several N-Word-Co expressions with usage-method-step concepts from the documents whilst all word concepts of the research are referred to WordNet and Thai-Herb-Encyclopedia after translating from Thai to English by Lexitron.

Our research is organized into 5 sections. In Section 2, related work is summarized. Problems in this research are described in Section 3 and Section 4 is the research framework. In Section 5, we evaluate and conclude our proposed model.

2. Related Works

Several strategies [4-9] have been proposed to determine the semantic relation from the textual data. In 2008, [4] presented TCMGeneDIT, a database that provides associations on Traditional Chinese Medicine (TCM), genes, diseases, the effects of TCM, ingredients, and the TCM effect and effecter relationships, which are mined and extracted from literature. The association discovery on noun phrases (TCM, disease, gene, ingredient, and effect) was conducted by using hypothesis testing and collocation analysis on annotated documents where a rule-based information extraction was performed. The Swanson's model was also applied to derive the transitive association genes → TCMs from genes → ingredients and ingredients → TCMs. The precision result of the associations between effects and effecters is 0.91. In 2011, [5] extracted the semantic relations between medical entities (as the treatment relations between a medical treatment and a problem, i.e. a disease symptom) by using the linguistic patterns to extract the relation from the 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. [5]'s results showed 75.72% precision and 60.46% recall. [6] extracted the procedural knowledge from MEDLINE abstracts as shown in the following example by using SVM compared to Conditional Random Field (CRF). "….<*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. 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 the Target. The other features: word feature, context feature, predicate-argument structure, and ontological feature, were utilized to classify procedures from several sentences. The results were 0.7279 and 0.8369 precisions for CRF and SVM, respectively with 0.7326 and 0.7957 recalls for CRF and SVM, respectively. In 2014, [7] applied the semi-automatic pipeline detection and the extraction of drug-adverse event (drug-AE) pairs from unstructured data, i.e. user-comment blogs and MEDLINE abstracts, and the structure database (Food and Drug Administration Adverse Event Reporting System). The drugs, diseases and symptoms or adverse events were based on noun phrases, including name entity recognition by using the PubMed dictionary. The Information Component (IC) value by using the Bayesian Confidence Propagation Neural Network was a measure of the disproportionality between entities of the drug-adverse event pairs. The IC was thus a measure of the strength of the dependency between a drug and an AE (Adverse Event). In 2015, [8] focused on extracting and classifying relations with three classes, cure, prevent, and side effect, occurring between disease and treatment on the MEDLINE abstracts including their titles. SVM and NB algorithms were used for the classification of relations. The learning features of the classifiers were obtained by using UMLS (Unified Medical Language System) to rank the words of verb phrases and noun phrases from the abstracts. [8]'s results of classifying the abstract relations showed an average F-measure of 93.5%. In 2016, [9] used linguistic patterns with semantic constraints to extract the semantic relations between entities or noun phrases

from French documents. [9] also showed that adding constraints improved both recall and precision, without having to rely of a POS tagger or syntactic analyzer.

However, unlike our research, the semantic relations extracted by previous researches [4,5,7-9] mostly occurred within one sentence containing either the relation between two NPs (NP1 and NP2 of a sentence expression(S) as S→NP1 VP; VP→ verb NP2 |…) or the relation between NP1 and VP. However, there are a few researches on a certain semantic relation occurrence with only one noun phrase expression as one procedural-target related to one vector of events expressed by the sentences' verb phrase (VP) expressions connected to explicit NP1 occurrences as the predicate-argument structure features of sentences [6] whereas the extraction of the semantic relation: the HerbalMedicinalProperty-UsageMethod relation, in our research is based on several events expressed by VP expressions from two related event sets as follows. Each event element (as the verb phrase element) of one event set, HMP, is related to several event vectors (verb phrase vectors) as the elements of the other event set, UM, where all of these event expressions on EDUs contain some NP1-ellipsis occurrences on the documents. Therefore, this research applied the Natural Language Processing (NLP) technique along with the machine learning techniques, i.e., SVM ME, and NB to determine and extract the N-Word-Co occurrences on verb phrases. The extracted N-Word-Co expressions are collected to identify either a medicinal-property-concept EDU or a usage-method-step-concept EDU, solving the boundary of each usage method (M_l), and also determining the HerbalMedicinalProperty-UsageMethod relation by NB learning.

3. Research Problems

There are three problems that must be solved: how to determine EDU with a verb phrase with a medicinal-property concept or a usage-method-step concept, how to determine each usage method (M_l) boundary, and how to determine the HerbalMedicinalProperty-UsageMethod relation between the HMP element and the UM element.

3.1 How to Determine a Verb Phrase with a Medicinal-Property Concept / Usage-Method-Step Concept

In the herb documents, there are some verb phrases with/without the medicinal property concepts or the usage-method-step concepts as shown in the following examples:

Example 2: "*กระวาน*/***Cardamom***" (Medicinal-Property)

EDU1: "(*ใบกระวาน*/*Cardamom leaf*)/NP (*มี*/*have* *รส*/*taste* *เผ็ดร้อน*/ *spicy*)/VP"

(*A Cardamom leaf has a spicy taste*)

EDU2: "*และ*/*and* [*it*] (*มี*/*has* *กลิ่นหอม*/*pleasing scent*)/VP"

([*it*] *has pleasing scent*)

EDU3: "*มี*/*has* *ฤทธิ์*/*effect* *ขับ*/*release* *ลม*/ *gas*"

([*It*] *has a carminative effect*.)

Example 2 has only EDU3's verb phrase with the medicinal-property concept (VP_{EDUmp}).

Example 3: "*โหระพา*/***basil***" (Usage-Method)

a) EDU1:"(*นำ*/*take* *เมล็ด*/*seed* *โหระพา*/*basil* *มาต้ม*/*boil*)/VP"

(*Take basil seeds to boil*.)

EDU2:"(*ดื่ม*/*drink*)/VP"

(*Drink [it]*.)

b) EDU1:"(*นำ*/*take* *เมล็ด*/*seed* *โหระพา*/*basil* *มาโรย*/*scatter* *บน*/*on* *ดิน*/*soil*)/VP"

(*Take basil seeds to scatter on the soil*.)

EDU2: "(*รดน้ำ*/*Water* *เมล็ด*/*seed*)/VP"

(*Water the seed*.)

The EDU's verb phrase with usage-method-step concept (VP_{EDUum}) occurs only on EDU1 and EDU2 of Example 3a). According to $VP_{EDUmp\text{-}j}$ and $VP_{EDUum\text{-}k}$ identification problems, the research applies the first word (w_1) of the following N-Word-Co expression to identify an EDU occurrence with either the medicinal-property concept or the usage-method-step concept after stemming words and eliminating stop words of the EDU occurrence.

$$\text{N-Word-Co} = w_1 + w_2 + .. + w_N$$

(where $w_1 \in Verb_{strong} \cup Verb_{weak}$ as a starting word of a N-Word-Co expression; i=2,3,..,N; $w_i \in$ Noun1 $\cup$ Noun2 $\cup$ $Verb_{strong}$ $\cup$ Adj $\cup$ Adv $\cup$ Unit $\cup$ $AdvPhrase_{dose}$)

Thus, it is necessary to extract the N-Word-Co expression with the medicinal-property concept (N-Word-Co_{mp}) and the N-Word-Co expression with the usage-method concept (N-Word-Co_{um}) from EDUs of the testing corpus after stemming words and eliminating stop words to solve the $VP_{EDUmp\text{-}j}$ and $VP_{EDUum\text{-}k}$ identification problems. However, there are various sizes of the extracted N-Word-Co_{mp}/ N-Word-Co_{um} expressions on verb-phrases as shown in the following examples.

Example 4: Herbal-Medicinal-Property Concept

EDU: "(*ต้นฟ้าทะลายโจร*/*Kariyat plant*)/NP (*ใช้*/*use* *รักษา*/*cure* *กระเพาะอาหาร*/ *stomach* *อักเสบ*/ *inflammation* *เรื้อรัง*/*chronic*)/VP"

(*Use Kariyat plant to cure chronic inflammation of the stomach*)

N-Word-Co= <to use> <to cure> <stomach><inflammation> (N=4)

EDU: "(*ขิง*/*Ginger*)/NP1 (*แก้*/*stop* *ท้องอืดเฟ้อ*/*flatulence*)/VP"

(*Ginger stops flatulence*.)

N-Word-Co= <to stop> <flatulence> (N=2)

Example 5: Usage-Method Concept

EDU1:"(*นำ*/*take* *ใบว่านหางจระเข้* /*aloe-vera leaf* *ล้างน้ำ*/*wash* *ให้สะอาด*/ *cleanly*)/VP"

(*Take an aloe-vera leaf to cleanly wash*.)

N-Word-Co= <to take> <plant-part> <to wash> (N=3)

EDU2:"(*ปอก*/*peel* *เปลือก*/skin *สีเขียว*/*green* *มีด*/*knife* *สะอาด*/*clean*)/VP"

(*Peel the green skin with clean knife*)

N-Word-Co= <to peel><plant-part> (N=2)

EDU3:"(*ล้าง/clean น้ำยาง/latex สีเหลือง/yellow ออกให้หมด/clearly*)/VP"

(*Clean out the yellow latex*)

N-Word-Co= <to clean > <exudate> (N=2)

Thus, we apply SVM, ME, and NB with w_1 and w_i features to learn the N-Word-Co size/boundary from verb phrases where $\text{Verb}_{\text{strong}}$, $\text{Verb}_{\text{weak}}$, Noun1, Noun2, Adj, Adv, Unit, and $\text{AdvPhrase}_{\text{dose}}$ are obtained from the corpus preparation.

The extracted $\text{N-Word-Co}_{\text{mp}}$ and $\text{N-Word-Co}_{\text{um}}$ expressions are collected into PNWC (which is the -$\text{N-Word-Co}_{\text{mp}}$ set) and UNWC (which is the $\text{N-Word-Co}_{\text{um}}$ set) (see Table 1) used to determine $\text{VP}_{\text{EDUmp-}j}$ and $\text{VP}_{\text{EDUum-}k}$, respectively.

Table 1. N-Word-Co Concept Sets

VP of Property	PNWC	PNWC with Summarized Concepts
'(*เป็น/is สมุนไพร/herb แก้/treat อุจเสียด/be-colic*)/VP'	<to be><herb><to treat>< be-colic>	<to treat>< be-colic>
'(*บรรเทา/relives อาการ/symptom แน่นท้อง/abdominal discomfort*)/VP'	<to relieve> <symptom> <abdominal-discomfort>	<to relieve> <abdominal-discomfort>
'(*ขับ/* ***release*** *ลม/**gas***)/VP'	<to release> <gas>	<to release> <gas>
….	….	….
VP of Usage Method	**UNWC**	**UNWC with Summarized Concepts**
'(*นำ/take เหง้าขิง/ginger-rootstock แก่/old ประมาณ 5กรัม/about 5gm.*)/VP'	<to take><plant-part><old><unit>	<to take> < plant-part> <old>
'(*ทุบ/pound ให้แตก/to crash*)/VP'	<to pound>	<to pound>
'(*ต้ม/boil กับ/in น้ำ/water*)/VP'	<to boil>	<to boil>
'(*ดื่ม/drink ระหว่างอาหารแต่ละมื้อ/ during each meal*)/VP'	<to drink> <$\text{AdvPhrase}_{\text{dose}}$>	<to drink>
….	…	.

Where the element concepts of the following linguistic sets: Noun1, Noun2, Adj, Adv, Unit, and $\text{AdvPhrase}_{\text{dose}}$ (but not including the common word occurrences on herb domain, i.e., '*อาการ/ symptom*' '*เป็น-สมุนไพร/be-herb*' '*มี-สรรพคุณ/have-medicinal-property*' etc.), are used to provide the summarized medicinal-property concepts of PNWC and the summarized usage-method-step concepts of UNWC. With regard to the linguistic phenomena, the main concept of each N-Word-Co expression is based on the word, w_1 or w_i , as the element of $\text{Verb}_{\text{strong}}$. For example:

EDU:"(*ใช้/use* [*น้ำมันสมุนไพร/herb_oil*] *ทา/apply ผิวหนัง/skin*)/VP"

N-Word-Co= <to use> <to apply> <skin>

where <to use> ∈ $Verb_{weak}$ and <to apply> ∈ $Verb_{strong}$. Therefore <to apply> is the main concept of N-Word-Co which is an element of UNWC.

3.2 How to Determine the Usage-Method Boundary

In regard to UM, the usage-method concept (M_l) extraction confronts two problems as to how to determine the first EDU of M_l and how to determine the M_l boundary mingled with non usage-method-step concept EDUs as shown in the following example.

Example 6: "Aloe Vera Preparation and Medicinal Property"

EDU1:"*นำใบว่านหางจระเข้ ล้างน้ำให้สะอาด*"

(*Take an aloe-vera leaf to cleanly wash.*)

EDU2:"*ปอกเปลือกสีเขียวออกด้วยมีดสะอาด*"

(*Peel the green skin with clean knife*)

EDU3:"*ล้างน้ำยางสีเหลืองออกให้หมด**clearly***" (*Clean out the yellow latex*)

EDU4:"*เพราะ*[*latex*]*อาจระคายเคือง ผิวหนัง*"

(*because* [*latex*] *may irritate skin.*)

EDU5:"*และ*[*latex*] *ทำให้ มี อาการแพ้ได้.*"

(*and* [*latex*]*causes to have allergy symptom.*)

EDU6:"*ฝาน* [*the peeled aloe-vera*] *เป็นแผ่นบาง*"

(*Slice* [*the peeled aloe-vera*] *into the thin flat form.*)

EDU7:"*ปิดบาดแผล*" (*Cover the wound.*)

EDU8:"*ช่วยรักษาบาดแผลให้ดีขึ้น*" (*Promote wound healing.*)

EDU9: "*รักษาผมร่วง*" (*and treat hair loss.*)

EDU10: "*และ ทำให้ ผิวหนังชุ่มชื้น*" (*and moisturize skin.*)

where EDU8-EDU10 are the medicinal property concept EDUs, and EDU1-EDU7 are the usage-method-step concept EDUs except EDU4 and EDU5. Moreover, there is another problem that *hmp* can occur before or right after *um* as shown in Example1 and Example6.

These problems can be solved by using the N-Word-Co expressions (see Table 1) to determine either the usage-method-step concept or the medicinal property concept of each EDU from the testing corpus.

3.3 How to Determine the ObjectProperty-UsageMethod Relation

Some documents contain two or more medicinal-property-concept-EDU occurrences around a certain usage-method-concept-EDU occurrence, which results in determining an incorrect relation of the ObjectProperty-UsageMethod/HerbalMedicinalProperty-UsageMethod relation between the semantic pair of the medicinal-property-concept-EDU occurrence and the usage-method-concept-EDU occurrence. For example:

Example 7: "ถั่วฝักยาว/***String-Bean***"

EDU1: "ถั่วฝักยาว/*string-bean* มี/*has* กากใยอาหาร/*food fiber* จำนวนมาก/*high amount*"

(*String beans have a high amount of fiber*.)

EDU2: "[มัน/*It*] ช่วย/*help* ลด/*reduce* คอเลสเตอรอล/*cholesterol*."

([*It*] *helps to lower cholesterol*.)

EDU3: "นำ/*take* ถั่วฝักยาว/*string-bean* ไปต้ม/*to boil* "

(*Take the string beans to boil*.)

EDU4: "เอา/*get* น้ำ/*liquid* ดื่ม/*to drink* " (*Get the liquid to drink*.)

EDU5: "[*It*] จะช่วย/*will help* รักษาบำรุง/*to treat* ไต/*kidney*" ([*It*] *helps to treat kidneys*.)

where EDU2 is the medicinal-property concept EDU whereas EDU3–EDU5 are the HerbalMedicinalProperty-UsageMethod relations with EDU5 with the medicinal property concept. This problem can be solved by using Naïve Bayes to learn the HerbalMedicinalProperty-UsageMethod relation between the semantic pair of the medicinal-property-concept-EDU occurrence and the vector of usage-method-concept-EDU occurrences.

4. A Framework of ObjectProperty-UsageMethod Relation Extraction from Texts

There are five steps in our framework. The first step is the corpus preparation step followed by the learning step of N-Word-Co size/boundary learning and ObjectProperty-UsageMethod/ HerbalMedicinalProperty-UsageMethod relation learning. The N-Word-Co extraction step is then operated and followed by the ObjectProperty-UsageMethod/HerbalMedicinalProperty-UsageMethod Relation extraction as shown in Fig. 1.

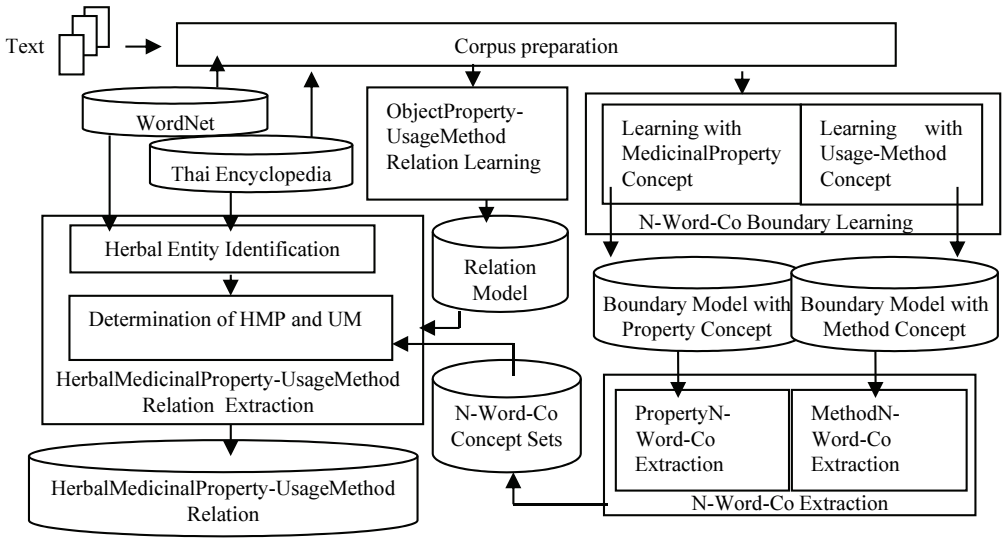

Fig. 1. System overview.

"ฟ้าทะลายโจร.....

ฟ้าทะลายโจรเป็นสมุนไพรรักษากระเพาะอักเสบเรื้อรัง[EDU1]ลดไข้[EDU2]ใช้ใบและก้าน[EDU3]หั่นเป็นชิ้นๆจำนวน9กรัม[EDU4]ต้มกับน้ำ[EDU5]ดื่ม[EDU6] ..."

"Kariyat ……

Kariyat is an herb used to heal a chronic inflammation of the stomach.[EDU1] Reduce fever[EDU2] Use leaves and stems.[EDU3] Cut [it] into pieces for 9gm.[EDU4] Boil in water.[EDU5] Drink [it].[EDU6]…"

<Topic_name Entity-concept= Kariyat /herb>ฟ้าทะลายโจร</Topic_name>.....

<PropertyMethodRelation ID=1> <EDU1 concept= 'medicinalProperty' >
<NP1 concept= 'Kariyat /herb-component'> ฟ้าทะลายโจร/herbNE </NP1>
<VP Type:'medicinalProperty'><N-Word-Co N=5words, concept= 'heal stomach inflammation'>
< w-mp1: POSset='verb-weak' ; concept= 'be' boundary ='yes'>เป็น </w-mp1>
< w-mp2: POSset='Noun2' ; concept= 'herb' boundary ='yes'>สมุนไพร </w-mp2>
< w-mp3: POSset ='verb-strong' ; concept= 'heal' boundary ='yes'>รักษา</w-mp3>
< w-mp4: POSset ='Noun2'; concept=' stomach'; boundary='yes'>กระเพาะ</w-mp4>
< w-mp5: POSset ='verb-strong' ; concept= 'inflammation' boundary ='yes'>อักเสบ</w-mp5>
< w-mp6: POSset ='Adj' ; concept=' chronic' boundary ='no'>เรื้อรัง</ w-mp6> </N-Word-Co></VP></EDU1>
<EDU2 concept= 'medicinalProperty' > <NP1 concept= 'Kariyat /herb-component'> ϕ </NP1>
<VP Type:'medicinalProperty'> <N-Word-Co N=2 words, concept= 'reduce fever'>
< w-mp1: POSset ='verb-strong' ; concept= 'reduce' boundary ='yes'>ลด </w-mp1>
< w-mp2: POSset ='Noun2' ; concept= 'fever' boundary ='yes'>ไข้ </w-mp2> </N-Word-Co></VP></EDU2>
<Method ID=1><EDU3 concept= 'usageMethod' > <NP1 concept= 'user'> ϕ </NP1>
<VP Type:'methodp'><N-Word-Co N=3words, concept='to prepare medicinal herb', Status=**FirstStep**>
<w-um1: POSset ='verb-weak' ; concept='use' boundary ='yes'>ใช้</w-um1>
<w-um2: POSset ='Noun2' ; concept='leaf' boundary ='yes'>ใบ</w-um2>
<w-um3: POSset =' Noun2';concept='stem'; boundary='yes'>ก้าน</w-um3></N-Word-Co ></VP></EDU3>
<EDU4 concept= 'usageMethod' > <NP1 concept= 'user'> ϕ </NP1>
<VP Type:'methodp'><N-Word-Co set='UNWC', N=3words, concept= 'to prepare medicinal herb', Status= **Step**>
<w-um1: POSset ='verb-strong' ; concept='cut' boundary ='yes'>หั่น</w-um1>
<w-um2: POSset ='Noun2' ; concept='piece' boundary ='yes'>ชิ้น</w-um2>
<w-um3: POSset ='Unit';concept='9 gm.'; boundary='yes'>9 กรัม</w-um3></N-Word-Co ></VP></EDU4>
<EDU5 concept= 'usageMethod' > <NP1 concept= 'reader'> ϕ </NP1>
<VP Type:'methodp'><N-Word-Co set='UNWC', N=2words, concept= 'to prepare medicinal herb' , Status= **Step**>
<w-um1: POSset ='verb-strong' ; concept='boil' boundary ='yes'>ต้ม</w-um1>
<w-um2: POSset ='Noun2' ; concept='water' boundary ='yes'>น้ำ</w-um2></N-Word-Co ></VP></EDU5>
<EDU6 concept= 'usageMethod' > <NP1 concept= 'reader'> ϕ </NP1>
<VP Type:'methodt'><N-Word-Co set='UNWC', N=1word, concept= 'to treat', Status= **LastStep** >
<w-um1: POSset ='verb-strong' ; concept= 'consume' boundary ='yes'>ดื่ม</w-um1></N-Word-Co ></VP>
</EDU6> </Method></PropertyMethodRelation>

where the PropertyMethodRelation tag is the ObjectProperty-UsageMethod relation tag, the Method tag is the usage-method boundary tag, the N-Word-Co tag is the word boundary tag of an N-Word-Co expression, the EDU tag is the simple sentence tag, the NP tag is the noun phrase tag, and the VP tag is the verb phrase tag. The w-mp*i* tag and the w-um*i* tag are the medicinal-property-word-*i* tag and the usage-method-word-*i* tag respectively where *i*=1,2,..,*num*. The [..] symbol or ϕ means ellipsis (Zero Anaphora).

Fig. 2. Annotation of ObjectProperty-UsageMethod relation and N-Word-Co expressions with medicinal-property concepts or usage-method-step concepts.

4.1 Corpus Preparation

This step is the preparation of the corpus in the form of EDUs from herbal medicine documents downloaded from several Thai herbal medicine websites, i.e., http://www.rspg.or.th/plants_data/herbs/herbs_200.htm. The step involves using Thai word segmentation tools [12], including Name entity [13]. After the word segmentation is achieved, EDU segmentation is then carried out [14]. These annotated EDUs are kept as an EDU corpus. This downloaded corpus contains 3000 EDUs (as unstructured and semi-structured documents with 100 different herbal medicine types with w_1 and w_i feature occurrences) randomly separated into 3 parts; the first part is 1500 EDUs to learn the N-Word-Co boundary/size with the medicinal-property/procedure-step concept from EDU verb phrases and also learning ObjectProperty-UsageMethod/HerbalMedicinalProperty-UsageMethod relation from the semantic pairs of the medicinal-property-concept-EDU occurrence and the vector of procedure-step-concept-EDU occurrences on the annotated corpus. The second part is 1000 EDUs for the N-Word-Co extraction and collection, and the third part is 500 EDUs for the ObjectProperty-UsageMethod/HerbalMedicinalProperty-UsageMethod relation extraction. We semi-annotate the learning corpus with the ObjectProperty-UsageMethod Relation, the medicinalProperty/herbal-medicinal-property concept EDUs and the usageMethod/procedure-step concept EDUs along with N-Word-Co expressions, as shown in Fig. 2, with word concepts referred to WordNet and Thai Herb Encyclopedia after translation from Thai to English, by Lexitron. The VP tag (<VP...>…</VP>) contains a POSset (Part of Speech set) features with several values (i.e., a Verb set, a Noun set, an Adj set, an Adv set, a Unit set, and etc.) used in the learning step. Where the Verb set consists of a 'verb-strong' set/$\text{Verb}_{\text{strong}}$ and a 'verb-weak' set/$\text{Verb}_{\text{weak}}$ (Verb = $\text{Verb}_{\text{strong}} \cup \text{Verb}_{\text{weak}}$). And, $\text{Verb}_{\text{strong}} = \text{Verb}_{\text{strongMP}} \cup \text{Verb}_{\text{strongPP}} \cup \text{Verb}_{\text{strongTP}}$; $\text{Verb}_{\text{strongMP}}$ is a strong-verb concept set with the medicinal-property concept , $\text{Verb}_{\text{strongPP}}$ is a strong-verb concept set with the preparation-procedure-step concept, and $\text{Verb}_{\text{strongTP}}$ is a strong-verb concept set with the treatment-procedure-step concept)

$\text{Verb}_{\text{strongMP}}$={'*แก้,รักษา/cure,treat*','*ห้าม/stop*','*ลด/reduce*','*บรรเทา/relieve,remedy*','*ขับ/expel,release*','*บำรุง/enrich*','*บวม/swell*', '*ปวด/pain*', '*อาเจียน/vomit*', '*จุกเสียด/be-colic*','*ถ่าย/defecate*',.., }

$\text{Verb}_{\text{strongPP}}$={'*บด,ขยี้/grind*','*ทุบ/pound*','*ผสม/mix*','*เติม/add*','*เคี่ยว/stew*','*ชง/brew*','*ต้ม/boil*','*กรอง,แยก/separate*', '*ตาก/dry*','*คั้น/squeeze*',…}

$\text{Verb}_{\text{strongTP}}$={, *'ดื่ม,กิน/consume'*, *'พอก/apply'*, *'ทา/apply'*, *'สูดดม/sniff'*, ...}.

4.2 Learning Step

The annotated corpus including stemming words and the stop word removal is used as the learning corpus to learn the ObjectProperty-UsageMethod/HerbalMedicinalProperty-UsageMethod relation and the N-Word-Co size/boundary.

4.2.1 ObjectProperty-UsageMethod relation learning

In regard to [11], Naïve Bayes learning is a generic classification to determine the feature probabilities of two classes ('yes', 'no') of the ObjectProperty-UsageMethod/HerbalMedicinalProperty-UsageMethod relation identification. The Naïve Bayes classifier of the research relies on a feature vector, ⟨N-Word-

$Co_{mp\text{-}j}$, N-Word-$Co_{um\text{-}1}$, N-Word-$Co_{um\text{-}2}$,.., N-Word-$Co_{um\text{-}c}$⟩ of two different feature sets, PNWC and UNWC, from the annotated corpus. The feature vector is used to determine the probabilities of ObjectProperty-UsageMethod/HerbalMedicinalProperty-UsageMethod relation (class='yes'; class∈Class; Class={'yes', 'no'}) and non-ObjectProperty-UsageMethod/HerbalMedicinalProperty-UsageMethod relation (class='no') from P(N-Word-$Co_{mp\text{-}j}$ |class), P(N-Word-$Co_{um\text{-}1}$|class), P(N-Word-$Co_{um\text{-}2}$|class),.., P(N-Word-$Co_{um\text{-}c}$ |class) by using Weka (http://www.cs.wakato.ac. nz/ml/weka/).

4.2.2 N-Word-Co size learning

In addition to the following N-Word-Co expression (after stemming words and the stop word removal) and the N-Word-Co size problem with Examples 4–5 described in Section 3.1, the features used to learn the N-Word-Co size from the learning corpus by ME, SVM, and NB are obtained from the annotated corpus containing the following concept sets: $Verb_{strong}$, Noun1, Noun2, Adj, Adv, Unit, and AdvPhrasedose; where each element of these concept sets should occur in more than 50% of the number of documents.

$$\text{N-Word-Co} = w_1 + w_2 + .. + w_N$$

(where $w_1 \in Verb_{strong} \cup Verb_{weak}$ as a starting word of a N-Word-Co expression; i=2,3,..,N; $w_i \in Noun1 \cup Noun2 \cup Verb_{strong} \cup Adj \cup Adv \cup Unit \cup AdvPhrase_{dose}$)

ME [11,15] modeled the probability of a semantic role r given a vector of features x according to the ME formulation below:

$$p(r \mid x) = \frac{1}{z_x} \exp[\sum_{j=0}^{n} \lambda_j f_j(r,x)] \quad (1)$$

where Z_x is a normalization constant, $f_i(r,x)$ is a feature function which maps each role and vector element to a binary value, n is the total number of feature functions, and λ_j is the weight for a given feature function. The final classification is just the role with the highest probability given its feature vector and mode.

According to Eq. (1), ME can be used as the classifier of the r class when $p(r|x)$ is the highest probability or argmax $p(r|x)$ to determine two N-Word-Co size/boundary classes, ending and continuing, ofall verb phrases from the corpus preparation. Moreover, r is the N-Word-Co size/boundary class (boundary is ending when r =0, otherwise r =1). And, x is the binary vector of word-concept features containing a word-concept pair (w_{ki} w_{ki+1}) with either a medicinal-property concept where k=1 or a usage-method-step concept where k=2.

If k=1 ∧ i=1 ∧ ($w_{k1} \in Verb_{strongMP} \cup V_{weak}$) then w_{ki} is the first word of VP having the medicinal-property concept.

If k=2 ∧ i=1 ∧ ($w_{k1} \in Verb_{strongPP} \cup Verb_{strongTP} \cup V_{weak}$) then w_{ki} is the first word of VP having the usage-method-step concept.

If k=1 ∧ ($w_{k1} \in Verb_{strongMP} \cup V_{weak}$) ∧ i=2,3,..,N then $w_{ki} \in W_1$ and $w_{ki+1} \in W_1$.

If k=2 ∧($w_{k1} \in Verb_{strongPP} \cup Verb_{strongTP} \cup V_{weak}$)∧ i=2,3,..,N then $w_{ki} \in W_2$, and $w_{ki+1} \in W_2$

W_1=$Verb_{strongMP} \cup Noun1 \cup Noun2 \cup Adv \cup Adj$

W_2=$Verb_{strongPP} \cup Verb_{strongTP} \cup Noun1 \cup Noun2 \cup Adv \cup Adj \cup Unit \cup AdvPhrase_{dose}$

All pairs of w_{ki} w_{ki+1} are obtained by sliding the window size of two adjacent words with one word sliding distance through an EDU's verb phrase after stemming words and the stop word removal (as shown in Eq. (2)) with the following binary feature vector format of word occurrences (w_{ki}) as word elements: $w_{e1}w_{e2}..w_{et} \in W_1$ and t is W_1 Cardinality where k=1; and $w_{e1}w_{e2}..w_{et} \in W_2$ and t is W_2 Cardinality where k=2.

w_{e1} w_{e2}.. w_{et}	w_{e1} w_{e2}.. w_{et}	w_{e1} w_{e2}.. w_{et}	.....	w_{e1} w_{e2}.. w_{et}	w_{e1} w_{e2}.. w_{et}
0 1 … 0	0 0 ..1.. 0	1 0 … 0		0 1 … 0	0 0 … 1
w_{k2}	w_{k3}	w_{k4}		w_{kN}	

$$p(r \mid x) = \arg\max_{r} \frac{1}{z}\exp\Big(\sum_{j=1}^{n} \lambda_j f_{yes,k\,i,j}(r, w_k) + \sum_{j=1}^{n} \lambda_j f_{no,k\,i,j}(r, w_k) + \sum_{j=1}^{n} \lambda_j f_{yes,k\,i+1,j}(r, w_{k+1}) + \sum_{j=1}^{n} \lambda_j f_{no,k\,i+1,j}(r, w_{k+1})\Big) \quad (2)$$

SVM [11,16] with the linear kernel: The linear function, $f(x)$, of the input $x = (x_1 \ldots x_n)$ assigned to the positive class if $f(x) \geq 0$, and otherwise to the negative class if $f(x)<0$, can be written as

$$\begin{aligned} f(\mathrm{x}) &= \langle \mathrm{w} \cdot \mathrm{x} \rangle + b \\ &= \sum_{j=1}^{n} w_j x_j + b \end{aligned} \quad (3)$$

where x is a dichotomous vector number, w is a weight vector, b is a bias, and $(w,b) \in R^n \times R$ are the parameters that control the function. The SVM learning is to determine the weight, w_j , and the bias, b, of each word feature, w_{kj} (or x_j) in the above binary feature vector format containing each word-concept pair (w_{kj} w_{kj+1},) with either a medicinal-property concept where k=1 or a usage-method-step concept where k=2 after checking the first word occurrence on VP as follows.

If k=1 $\wedge$ i=1 $\wedge$ ($w_{ki} \in Verb_{strongMP} \cup V_{weak}$) then w_{ki} is the first word of VP with the medicinal-property concept.

If k=2 $\wedge$ i=1 $\wedge$ ($w_{ki} \in Verb_{strongPP} \cup Verb_{strongTP} \cup V_{weak}$) then w_{ki} is the first word of VP with the usage-method-step concept.

The N-Word-Co size/boundary learning from $w_{kj}w_{kj+1}$ of VP is then the supervised learning of SVM by sliding the window size of two consecutive words with one sliding word distance after stemming words and the stop word removal. Where j=1,2,..,n and n is End-of-Boundary and is equivalent to the N value of N-Word-Co size.

NB [11] An annotated verb phrase with either a medicinal-property concept or a usage-method-step concept in the learning corpus is obtained as a N-Word-Co with the medicinal-property concept vector (WV_{ki}, where k=1) or the usage-method-step concept vector (WV_{ki} , where k=2) into matrix vector, MW_k , of the herbal-medicinal properties (k=1) or the usage-method steps (k=2) respectively.

WV_{kj} = {$w_{k\text{-}j1}$, $w_{k\text{-}j2}$….$w_{k\text{-}jN}$ mp/non-mp} where k=1 ; mp is a N-Word-Co/a word vector with a medicinal-property concept and non-mp is a N-Word-Co/a word vector with a non-medicinal-property concept existing in an EDU verb phrase (*vp*) as $vp = w_{k\text{-}j1}w_{k\text{-}j2}...w_{k\text{-}jN}...w_{k\text{-}j\text{lastVPword}}$ respectively with the medicinal-property concept of a certain herbal plant where the first word as $w_{k\text{-}j1} \in \text{Verb}_{\text{strongMP}} \cup \text{V}_{\text{weak}}$.

WV_{kj} = {$w_{k\text{-}j1}$, $w_{k\text{-}j2}$….$w_{k\text{-}jN}$ um/non-um} where k=2; um is a N-Word-Co/a word vector with a usage-mothod-step concept and non-um is a N-Word-Co/a word vector with non-usage-method-step concept existing on an EDU verb phrase (*vp*) as $vp = w_{k\text{-}j1}w_{k\text{-}j2}...w_{k\text{-}jN}...w_{k\text{-}j\text{lastVPword}}$ respectively with the usage-method-step concept of the certain herbal plants where the first word as $w_{k\text{-}j1} \in \text{Verb}_{\text{strongPP}} \cup \text{Verb}_{\text{strongTP}} \cup \text{V}_{\text{weak}}$.

MW_k = {WV_{kj}} where j=1,2,..,*theNumberOfVPs*

After we have obtained the word feature vectors on verb phrases with the first word as $w_{k\text{-}j1} \in \text{Verb}_{\text{strongMP}} \cup \text{V}_{\text{weak}}$ (where k=1) and $w_{k\text{-}j1} \in \text{Verb}_{\text{strongPP}} \cup \text{Verb}_{\text{strongTP}} \cup \text{V}_{\text{weak}}$ where k=2 from the learning corpus, we then determine the probabilities of the medicinal-property concept/non the medicinal-property concept and the usage-method-step concept/non the usage-method-step concept respectively from a slide window size of two consecutive words from the verb phrase with the one-sliding-word distance, shown in Table 2 by using Weka(http://www.cs.wakato.ac.nz/ml/weak/).

Table 2. Show probability of $w_{k\text{-}ji}$ concept and $w_{k\text{-}ji+1}$ concept of words in the N-Word-Co

$w_{k\text{-}ji}$	Medicinal Property Word	NonMedicinal Property Word	$w_{k\text{-}ji}$	UsageMethodStep Word	NonUsageMethodStep Word
‘แก้/*to stop*’	0.00986842	0.03169014	‘บด/*to grind*’	0.00328947	0.01760563
‘คลื่นไส้/*to vomit*’	0.00328947	0.01056338	‘ตาก/*to dry*’	0.00328947	0.02464789
‘เป็น/*be,as*’	0.00986842	0.00352113	‘ต้ม/*to boil*’	0.00328947	0.01056338
‘เลือด/*blood*’	0.00328947	0.01760563	‘ใบ/*plantPart*’	0.00328947	0.01760563
‘อาการ/*symptom*’	0.05592105	0.05985915	‘ดื่ม/*to consume*’	0.03618421	0.01760563
…	…	…			
$w_{k\text{-}ji+1}$	**Medicinal Property Word**	**NonMedicinal Property Word**	$W_{k\text{-}ji+1}$	**UsageMethodStep Word**	**NonUsageMethodStep Word**
‘แก้/*to stop*’	0.02160494	0.00328947	‘บด/*to grind*’	0.0154321	0.03618421
‘คลื่นไส้/*to vomit*’	0.00925926	0.01644737	‘ตาก/*to dry*’	0.02160494	0.01644737
‘เป็น/*be,as*’	0.00925926	0.00328947	‘ต้ม/*to boil*’	0.00925926	0.00328947
‘เลือด/*blood*’	0.0154321	0.01644737	‘ใบ/*plantPart*’	0.0154321	0.00328947
‘อาการ/*symptom*’	0.10185185	0.02960526	‘ดื่ม/*to consume*’	0.0462963	0.00328947
	….	…..			

4.3 N-Word-Co Extraction

The extracted N-Word-Co occurrences on the documents by the following ME, SVM, and NB from the testing corpus are collated into two different N-Word-Co concept sets, PNWC and UNWC, in Table 1 of Section 3.

ME: The N-Word-Co size/boundary is then determined by using λ (the weight for the given feature function of the N-Word-Co size/boundary determination based on a vector of word features with medicinal-property-concepts/usage-method-step concepts) by Eq. (2) as shown in Fig. 3 after stemming words and the stop word removal of the EDU.

SVM: The N-Word-Co size/boundary is also solved by the weight vector from of all w_{kj} (where k=1) and the weight vector of all w_{kj} (where k=2). These weight vectors are obtained from the SVM learning and are also used to extract and collect the N-Word-Co occurrences with either the medicinal-property concepts or the usage-method-step concepts into PNWC or UNWC by Eq. (3) as shown in Fig. 3. Hence, Fig. 3 returns $NWCset_k$ which is PNWC if k=1 or UNWC if k=2.

```
Assume that each EDU is represented by (NP1  VP) after stemming words and the stop word removal.
L is a list of EDUs.   Verb=Verb_strongMP∪Verb_strongPP ∪Verb_strongTP ∪ Verb_weak
V_strong= Verb_strongMP∪Verb_strongPP ∪Verb_strongTP ; W_k=Verb_strongMP∪Noun1∪Noun2∪Adv∪Adj  where k=1
W_k = Noun1∪Noun2∪Verb_strong∪Adj∪Adv ∪ Unit∪AdvPhrase_dose where k=2 ;
 k=1: Medicinal-Property Concept;  k=2: Usage-Method Concept

N_WORD_CO_EXTRACTION
 1  NWCSet_k ←∅; NWC_1←∅ ;  NWC_2←∅ ;  temp←∅  ;  i=1 ; j=1;  k =0 ;  flag=yes
    /* NWC_k is the N-Word-Co expression,  k =1,2
    /* w_ki is a word at the i^th element of VP  having a medicinal property concept (k=1)  or
       an usage-method concept (k=2) , i=1,2,..endOfVerbPhrase
 2  while j≤ Length[L] do
 3   {1  If  i=1    then                              /* identify the 1st word of N-Word-Co and k
 4     {2   If  (w_ki∈ Verb_strongMP ) then  {k=1;  NWC_k←w_ki }
 5       Else If  w_ki∈(Verb_strongPP ∪ Verb_strongTP )   then   {k=2; NWC_k←w_ki}
 6       Else If  (w_ki∈ Verb_weak)      then   temp← w_ki ;
 7        i++ ;
 8        while (temp<>∅)  do
 9          {3  If  w_ki∈ Verb_strongMP  then  {k=1;  NWC_k←temp + w_ki ;  temp←∅ }
10            Else If w_ki∈(Verb_strongPP ∪Verb_strongTP ) then {k=2; NWC_k←temp + w_ki ;  temp←∅ }
11             Else  {  temp← temp+w_ki      } ;
12              i++   }3
13     }2                          /* determine N-Word-Co size/boundary according to the selected Case
14     r=1    /* r is the N-Word-CO boundary classes (r=0:boundary is ending,otherwise r=1)
15     while (flag=yes) ∧ (w_ki∈W_k) ∧ (i≤endOfVerbPhrase) do
16      {4  i=i-1;
17         Case: use_ME
18               Equation(2) ; If r=0 then flag←no ,otherwise flag←yes
19         Case: use_SVM
20               Equation(3) ;  If f(x)>0 then flag←no , otherwise flag←yes
           Case: use_NB
                 Equation(4) ;  If class= 'no'  then flag←no ,otherwise flag←yes
21         EndCase
22         If  (r=1) ∨ ( f(x)≤0) ∨ (class= 'yes') then  NWC_k← NWC_k  ∪  w_ki ;
23        i++   }4
24     NWCSet_k ←  NWCSet_k ∪ NWC_k;    i=1 ;  j++ ;   }1
25  } return   NWCSet_k                          /*  NWCSet_1 is PNWC ;  NWCSet_2 is UNWC
```

Fig. 3. N-Word-Co extraction algorithm.

NB: After the first word, $w_{k\text{-}i1}$,of a word vector on an EDU's verb phrase from the testing corpus has been identified by $w_{k\text{-}i1} \in \text{Verb}_{\text{strongMP}} \cup \text{V}_{\text{weak}}$ (where k=1) and $w_{k\text{-}i1} \in \text{Verb}_{\text{strongPP}} \cup \text{Verb}_{\text{strongTP}} \cup \text{V}_{\text{weak}}$ (where k=2), the N-Word-Co boundary is determined by using the Naive Bayes classifier in Eq. (4) and Table 2 to determine the boundary by sliding a window size of two words with the one-sliding-word distance on the consecutive words of the verb phrase (after stemming words and eliminating stop words). As soon as the class '*no*' (non medicinal-property concept) is determined, the N-Word-Co boundary with the medicinal-property concept (N-Word-Co_{mp}) is ended, and as soon as class '*no*' (non usage-method-step concept) is determined, the N-Word-Co boundary with the usage-method-step concept (N-Word-Co_{um}) is ended as shown in Fig. 3 of the N-Word-Co boundary determination.

$$\begin{aligned} MultiWordCoBoundaryClass &= \arg\max_{class \in Class} P(class \mid w_{k-ji}, w_{k-ji+1}). \\ &= \arg\max_{class \in Class} P(w_{k-ji} \mid class) P(w_{k-ji+1} \mid class) P(class). \\ where\ w_{k-ji} &\in W_{kj} \quad and \quad w_{k-ji+1} \in W_{kj} \\ (\text{if } k = 1, \text{then } &W_{kj} \text{ is a } Medicinal\,\text{Pr}operty\_word\_concept\ vector; \\ \text{if } k = 2, \text{then } &W_{kj} \text{ is a } UsageMethodStep\_word\_concept\ vector\ ) \\ i = \{1,2,..lastVPword\} \quad j &= \{1,2,..theNumberOfVPs\} \quad Class = \{"yes","no"\} \end{aligned} \tag{4}$$

Moreover, the N-Word-Co concepts are determined by $\text{Verb}_{\text{strongMP}}$, $\text{Verb}_{\text{strongPP}}$, $\text{Verb}_{\text{strongTP}}$, Noun1, Noun2, Adv, Adj, Unit, and AdvPhrasedose on the herbal-plant domain (see Table 1). All of these concepts are referred to WordNet and Thai-Herb-Encyclopedia after translation from Thai to English by Lexitron.

4.4 ObjectProperty-UsageMethod Relation Extraction

The objective of this step is to recognize and extract the ObjectProperty-UsageMethod/ HerbalMedicinalProperty-UsageMethod relation from the testing EDU corpus after the herb plant name/object has been identified from the document topic name by using the Thai Herb Encyclopedia. The determination of an N-Word-Co of the testing corpus's EDUs (after stemming words and stop word removal) with the medicinal-property concept or the usage-method-step concept relies on the similarity score determination as the Max Similarity Score (MaxSimScore) in Eq. (5). MaxSimScore is determined between the N-Word-Co of the testing corpus's EDUs and the candidate N-Word-Co expressions from either PNWC or UNWC.

$$MaxSimScore = ArgMaxSimilarity_{t=1}^{Cardinality}\left(\frac{|\,NWCcorpus \cap NWCcandidate_t\,|}{\sqrt{|NWCcorpus\,| \times |\,NWCcandidate_t\,|}}\right)$$

$$\begin{aligned} &where\ Cardinality \text{ is the number of N-Word-Co elements of the N-Word-Co concept set, PNWC or UNWC.} \\ &NWCcandidate \text{ is a candidate N-Word-Co element of the N-Word-Co concept set, PNWC or UNWC.} \\ &NWCcorpus \text{ is an N-Word-Co of EDU from the testing corpus} \end{aligned} \tag{5}$$

Assume that each EDU is represented by (NP1 VP) after stemming words and the stop word removal. L is a list of EDU

V_{strong}= $Verb_{strongMP} \cup Verb_{strongPP} \cup Verb_{strongTP}$; W_k=$Verb_{strongMP} \cup Noun1 \cup Noun2 \cup Adv \cup Adj$ where k=1

W_k = $Noun1 \cup Noun2 \cup Verb_{strong} \cup Adj \cup Adv \cup Unit \cup AdvPhrase_{dose}$ where k=2

k=1: Medicinal-Property Concept k=2: Usage-Method Concept

N-Word-Co_k = $w_{k1} + w_{k2} + .. + w_{kN}$ (where $w_1 \in Verb_{strong} \cup Verb_{weak}$ as a starting word of a N-Word-Co expression; i=2,3,..,N; $w_{ki} \in W_k$ if k=1,2)

wdji is an input word$j$ of a verb phrase of EDUi ($VP_{EDU\text{-}i}$) of the testing corpus

(where $VP_{EDU\text{-}i}$ contain $NWC_{EDU\text{-}i}$ (N-Word-Co of EDUi))

canPNWC is a candidate set of PNWC.

canUNWC is a candidate set of UNWC.

OBJECT_PROPERTY-USAGE_METHOD_RELATION_DETERMINATION

```
1  {i=1 ; j=1; HMP←∅; UM← ∅; Method←∅; HMPelement=0; UMelement=0
2   while i≤ Length[L] do                /*identify a starting word of N-Word-Co
3   {1 while (wdji∈Vstrong ∪ Verbweak) ∧ i ≤ (Length[L]) do
4     {2          /*identify N-Word-Co with either canPNWC or canUNWC
5       If (MaxMaxSimScore (NWCEDU-i, canPNWC, canUNWC) >0.9) ∧ (class ='MedicinalProperty')
6          {3  HMP←HMP ∪ NWCEDU-i;          /* HMP_Det.(Determination of Herbal- Medicinal-
                                                          Property Concept Set)
7              HMPelement++ ; i++;
8          }3
9        ElseIf  (MaxMaxSimScore (NWCEDU-i, canPNWC, canUNWC) >0.9) ∧ (class ='UsageMethod')
10         {4                                  /*UM_Det.(Determination of Usage-Method Concept Set
11           While ( wdji ∉ VstrongTP) ∧( wdji ∉ VstrongMP) ∧ (i≤ Length[L]) do
12             {5  /*to determine each usage-method (Ml) and the Ml boundary mingled with non usage-method
                   concept EDUs where VerbstrongTP is a strong verb concept set with the treatment-procedure
                   concept
13               If (MaxMaxSimScore (NWCEDU-i, canPNWC, canUNWC) >0.9) ∧ (class ='UsageMethod')
14                   Method←Method + NWCEDU-i;
15             i++ }5  ;
16           While (MaxMaxSimScore (NWCEDU-i, canPNWC, canUNWC) >0.9) ∧(class ='UsageMethod')
   ∧ (wdji ∈VstrongTP) ∧( wdji ∉ VstrongMP) ∧(i≤ Length[L]) do
17                 {6  Method←Method + NWCEDU-i;  i++ }6
18           UM← UM ∪ Method;  UMelement++;
19         }4  }2   i++ }1
20   For l =1 to UMelement                  /* Determination of ObjectProperty-UsageMethod Relation
21   For  j =1 to HMPelement
22     {  NWordComp-j  ← query_N-Word-Comp-j_from_HMP
23        Ml ← query_Ml_from_UM
24        Equation (7)
25        If Class= 'yes'  then  R← R∪ (NWordComp-j+Ml )
26      }
27 }Return  R
```

Fig. 4. ObjectProperty-UsageMethod Relation Determination Algorithm.

$$MaxMaxSimScore = \underset{class \in Class}{ArgMax}\ (MaxSimScore1, MaxSimScore2) \quad (6)$$

where $MaxSimScore1$ is MaxSimScore between the N - Word - Co of the testing corpus EDU and the candidate N - Word - Co from PNWC

$MaxSimScore2$ is MaxSimScore between the N - Word - Co of the testing corpus EDU and the candidate N - Word - Co from UNWC

$Class = \{'property', 'usage-method'\}$

There are two main steps in extracting the HerbalMedicinalProperty-UsageMethod relation from the documents as shown in Fig. 4. The first step is to identify the N-Word-Co concept of VP_{EDU} from two N-Word-Co concept sets, PNWC or UNWC, by calculating MaxMaxSimScore in Eq. (6) which is the highest similarity-score value between the MaxSimScore1 (the candidate N-Word-Co based on PNWC) and the MaxSimScore2 (the candidate N-Word-Co based on UNWC) where PNWC$\cap$ UNWC =$\varnothing$. The concept class (*class*) of MaxMaxSimScore of N-Word-Co of VP_{EDU} is also determined by Eq. (6) (where *class*= 'property' if MaxSimScore1>MaxSimScore2 and MaxMaxSimScore≥0.9; *class*= 'usage-method' if MaxSimScore2>MaxSimScore1 and MaxMaxSimScore≥0.9). The herbal-medicinal-property concept set (HMP) is formed if *class*= 'property'. The usage-method concept set (UM) is formed if *class*= 'usage-method' including the M_l boundary determination as shown in Fig. 4.

The second step is determining the HerbalMedicinalProperty-UsageMethod relation by the Cartesian product of matching each HMP element to each UM element through Naïve Bayes as in Eq. (7) where the HMP element is represented by N-Word-Co_{mp-j} of $VP_{EDUmp-j}$ and the UM element is represented by M_l. M_l and consists of several procedure steps of $VP_{EDUum-k}$ (M_l=$VP_{EDUum-1}$+$VP_{EDUum-2}$ +..+$VP_{EDUum-c}$ and k =1,2,..,c) where $VP_{EDUum-k}$ is represented by N-Word-Co_{um-k}.

$$\begin{aligned} Object\,Property-UsageMethod\_RelClass &= \underset{class \in Class}{\arg\max}\ P(class \mid NWordCo_{mp-j}, M_l) \\ &= \underset{class \in Class}{\arg\max}\ P(NWordCo_{mp-j} \mid class) P(M_l \mid class) P(class) \end{aligned} \quad (7)$$

where $NWordCo_{mp-j}$ is an N - Word - Co expression with the herbal medicinal property concept of VP_{EDU-j}

$j = 1,2,..,Cardinality\_Of\_HMP$; HMP is a herbal - medicinal - property concept set

M_l is *a* usage method concept; $l = 1,2,..,Cardinality\_Of\_UM$; UM is a usage - method concept set

$Class = \{"yes","no"\}$

5. Evaluation and Conclusion

The testing corpus of 1,500 EDUs randomly collected from several Thai herbal web sites in the corpus preparation step is used to evaluate the proposed ObjectProperty-UsageMethod relation extraction from texts. The testing corpus is separated into 2 parts; the first part of the testing corpus contains 1,000 EDUs to test N-Word-Co determination and extraction based on precision and recall which are evaluated by three expert judgments with max win voting as shown in Table 3. The correct extracted N-Word-Co expressions are then used in the second part of the testing corpus with 500 EDUs to evaluate the ObjectProperty-UsageMethod/HerbalMedicinalProperty-UsageMethod relation extraction based on precision and recall which are judged by three expert judgments with max win voting.

Table 3. The evaluation of the N-Word-Co extraction from herbal-web documents

Herb corpus	Correctness of N-Word-Co determination (%)					
	SVM		ME		NB	
	Precision	Recall	Precision	Recall	Precision	Recall
MedicinalPropertyPart 500 EDUs	91.4	81.1	88.9	79.4	84.6	78.9
UsagePart 500 EDUs	90.2	80.8	89.1	81.3	86.3	79.5

The average precision of extracting N-Word-Co with the medicinal-property/usage-method-step concepts is 90.8%, 89%, and 85.5% with an average recall of 81.0%, 80.4%, and 79.2% by SVM, ME, and NB respectively, as shown in Table 3. The reason for low recall is the anaphora problem as shown in the following example (a).

(a) "*โหระพา/**Basil***"
EDU1:"*ใบ/leaves โหระพา/Basil บรรเทา/relieve อาการ/symptom ดังกล่าวข้างต้น/as mentioned above*"
(*Basil leaves relieve the symptom as mentioned above*.)

Moreover, the research results of N-Word-Co extraction show that NB yields the lowest precision because NB is based on feature probabilities, for examples (b) and (c):

(b) "*มะกรูด/ **Kaffir lime***"
EDU1: "*ใบ/ leaf มะกรูด/Kaffir lime ใช้/use แก้/stop อาเจียน/vomit เป็น/be เลือด/blood* "
(*The kaffir lime leaf is used to stop vomiting blood.*)
N-Word-Co_{mp} = <to use><to stop><to vomit><as><blood>
(c) **"*ขิง/Ginger* "**
EDU2: "*ขิง/Ginger เป็น/is สมุนไพร/herb แก้/stop อาเจียน/vomit ได้เป็นอย่างดี/very well*"
(*Ginger is an herb that stops vomiting very well*)
N-Word-Co_{mp} = <to use><to stop><to vomit>

NB determines N-Word-Co_{mp} boundaries of (b) and (c) as <to use><to stop><to vomit> because the probability of 'to vomit' as a N-Word-Co boundary is higher than the probability of 'to vomit' as a non N-Word-Co boundary.

The correct N-Word-Co concept sets, PNWC and UNWC, extracted from the documents are used to extract the HerbalMedicinalProperty-UsageMethod relation from the web documents with a precision of 94.5% and a recall of 85.3%. The reasons for lower recall are 1) some herbal-medicinal-property occurrences are based on an event expression by a preposition phrase as shown in the following example (d).

(d) **"*ชุมเห็ดเทศ/**Candle Bush***"
EDU1: "*ใช้/use ช่อดอก/inflorescence 2-3 ช่อ/2-3bunches*"
(*Use 2-3 inflorescences*.)
EDU2: "*ต้ม/boil* [*มัน*]" (*Boil* [*them*].)
EDU3: "*กิน/eat กับ/with น้ำพริก/chili sauce เพื่อ/for ระบายท้อง/excreting*".
(*Eat* [*them*] *with chili sauce for excretion*.)

where EDU1-EDU3 are the usage methods of using the herb plant of the candle bush. EDU3 also expresses the herbal medicinal property on the preposition phrase as '*for excreting*'. And 2) lacking either the explicit $VP_{EDUmp\text{-}j}$ or $VP_{EDUum\text{-}k}$ occurs in the corpus as shown in the following examples (e) and (f).

(e)"***พริกขี้หนู/guinea-pepper***"
EDU1 "*พริกขี้หนูแห้ง/dry guinea-pepper สามารถ/can ใช้เป็น/use as เครื่องปรุง/seasoning*"
(*Dry guinea-pepper can be used as seasoning*.)
EDU2 "*และ/and ใช้/use ผสม/mix วาสลิน/vaseline*"
(*and can be mixed with vaseline*.)
EDU3 "*ทา/apply แผลฟกช้ำ/bruise* "
(*Apply to a bruise*.)
(f) "***กะหล่ำปลี/Cabbage***"
EDU1"*กะหล่ำปลี/Cabbage ใช้/use เป็น/as อาหาร/food ใน/in การรักษา/curing โรคกระเพาะ/gastritis และ/and ป้องกัน/preventing มะเร็งลำไส้/Colonic Carcinoma ได้ดี/very well*"
(*Cabbage is used as food for curing gastritis and preventing Colonic Carcinoma* very *well*.)

where the example (e) lacks the explicit $VP_{EDUmp\text{-}}$ occurrence such as "*แก้ฟกช้ำ/ To heal a bruise* " and the example (f) lacks the explicit $VP_{EDUum\text{-}k}$ occurrence such as "*ต้มกะหล่ำปลี/ To cook cabbage.*".

However, the N-Word-Co concept sets, PNWC and UNWC, are useful not only to extract the HerbalMedicinalProperty-UsageMethod relation, but also for concision of the procedure step concepts through UNWC with the summarized concepts in Table 1, which people will understand effortlessly and rapidly. Hence, the research contributes the extraction of the semantic relation which is the HerbalMedicinalProperty-UsageMethod relation between two related sets, PNWC and UNWC, with the boundary consideration and also with the unordered pair consideration of the herbal-medicinal-property occurrence and the usage-method occurrence in texts. In regard to this research, the extracted HerbalMedicinalProperty-UsageMethod relation can enhance the Problem-Solution of the healthcare system by providing indigenous knowledge about using medicinal plants as the objects for healthcare through a question answering system. Finally, our research can also enhance Problem-Solution in other areas, e.g., solving industrial problems.

Acknowledgement

This work was supported by Dhurakij Pundit University.

References

[1] C. S. G. Khoo and J. C. Na, "Semantic relations in information science," *Annual Review of Information Science and Technology*, vol. 40, no. 1, pp. 157-228, 2006.

[2] 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*, Dordrecht, Netherlands: Springer, 2003, pp. 85-112.

[3] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.

[4] Y. C. Fang, H. C. Huang, H. H. Chen, and H. F. Juan, "TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining," *BMC Complementary and Alternative Medicine*, vol. 8, no. 58, pp. 1-11, 2008.

[5] A. B. Abacha and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach," *Journal of Biomedical Semantics*, vol. 2 (Suppl 5), pp. S4, 2011.

[6] 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," *Active Media Technology, Lecture Notes in Computer Science*, vol. 6890, pp. 345-354, 2011.

[7] S. Yeleswarapu, A. Rao, T. Joseph, V. G. Saipradeep, and R. Srinivasan, "A pipeline to extract drug-adverse event pairs from multiple data sources," *BMC Medical Informatics and Decision Making*, vol. 14, no. 13, pp. 1-16, 2014.

[8] A. W. Muzaffar, F. Azam, and U. Qamar, "A relation extraction framework for biomedical text using hybrid feature set," *Computational and Mathematical Methods in Medicine*, vol. 2015, article ID. 910423, 2015.

[9] M. Lafourcade and L. Ramadier, "Semantic relation extraction with semantic patterns: experiment on radiology report," in *Proceeding of the 10th LREC 2016 Conference on Language Resources and Evaluation*, Portoroz, Slovenia, 2016.

[10] S. J. Kim, Y. H. Lee, and J. H. Lee, "Method of extracting is-a and part-of relations using pattern pairs in mass corpus," in *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, China, 2009, pp. 260-268.

[11] T. M. Mitchell, *Machine Learning*. Singapore: McGraw-Hill Science, 1997.

[12] S. Sudprasert and A. Kawtrakul, "Thai Word segmentation based on global and local unsupervised learning," in *Proceedings of the 7th National Computer Science and Engineering Conference*, Chonburi, Thailand, 2003, pp. 1-8.

[13] H. Chanlekha and A. Kawtrakul, "Thai named entity extraction by incorporating maximum entropy model with simple heuristic information," in *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, Hainan Island, China, 2004, pp. 49-55.

[14] J. Chareonsuk, T. Sukvakree, and A. Kawtrakul, "Elementary discourse unit segmentation for Thai using discourse cue and syntactic information," in *Proceedings of the National Computer Science and Engineering Conference*, 2005, pp. 85-90.

[15] A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra, "A maximum Entropy approach to natural language processing," *Computer Linguistics*, vol. 22, no. 1, pp. 39-71, 1996.

[16] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, UK: Cambridge University Press, 2000.

Chaveevan Pechsiri http://orcid.org/0000-0001-6133-2285

She holds a Master's degree in computer science from Mississippi State University, USA, and a Doctoral degree in Computer Engineering from Kasetsart University, Thailand. She is currently an Associate Professor at Dhurakijpundit University, Thailand. Her general research interest is in natural language processing.

Sumran Phainoun http://orcid.org/0000-0002-5980-1596

She holds an MBA from and is currently a lecturer at Dhurakijpundit University, Thailand. Her general research interest is in databases in business and enterprises, statistics, information technology, and computer science.

Rapeepun Piriyakul http://orcid.org/0000-0002-7337-7728

She holds a Master's degree in Applied Statistics at the National Institute of Development Administration, Thailand and a Doctoral degree in Computer Engineering from Kasetsart University, Thailand. Her general research interests is in applied analytical statistics in computer engineering.