Comparative Study of Using Word Co-occurrence to Extract Disease Symptoms from Web Documents

Chaveevan Pechsiri[1], Sumran Phainoun[1] ,Renu Sukharomana [2]

[1]College of Innovative of Technology and Engineering, Dhurakij Pundit University,Thailand
[2] Institute of Social and Economic Science, TRF, Thailand

{ chaveevan.pec@dpu.ac.th,
sumran.pha@dpu.ac.th,sukharenu@gmail.com }

Abstract. The research aim is a comparative study of using different word co-occurrence sizes as the two word co-occurrence and the N word co-occurrence on verb phrases to extract disease symptom explanations from downloaded hospital documents. The research results are applied to construct the semantic relations between disease-topic names and symptom explanations for enhancing the automatic problem-solving system. The machine learning technique, Support Vector Machine, and the similarity score determination are proposed to solve the boundary of simple sentences explaining the symptoms for the two word co-occurrence and the N word co-occurrence respectively. The symptom extraction result by the N word co-occurrence provides the higher precision than the two word co-occurrence from the documents.

Keywords: word co-occurrence, event boundary, symptom explanation

1 Introduction

The research objective is the comparative study of using different word co-occurrence sizes as the two word co-occurrence and the N word co-occurrence on verb phrases to extract the disease symptom explanations from the downloaded health-care documents on the hospital web-boards. The research results are beneficial to the automatic problem-solving system after each semantic relation is constructed between a disease name from a document topic name and the extracted disease-symptom explanation from the document. Moreover, the disease symptom explanation mostly consists of event expressions on several EDUs (where EDU is an Elementary Discourse Unit expression defined as a simple sentence or a clause, [1]) as the symptom-concept explanation on a document of a certain disease as follow.

Example1.

Topic name: *โรคหลอดลมอักเสบ* ***Bronchitis Disease***

EDU1 (symptom) : " *เมื่อผมไอออกมา*/***When I cough,*** "

(*เมื่อ*/*When* (*ผม*/*I*)/NP (*ไอออกมา*/*cough*)/VP)

EDU2 (symptom): " [*มัน*]*จะมีเสมหะเป็นเลือด*/[***It***]***will have phlegm containing blood.*** "

([*มัน/It*] (*จะมี/will have เสมหะ/phlegm* (*sputum*) *เป็นเลือด/as blood*)/VP)

EDU3 (symptom): " *แต่ว่า*[*ผม*]*ไม่เป็นไข้/**But** [**I**] **have no fever.***"

(*แต่ว่า/But* [*ผม/I*] (*ไม่เป็น/have no ไข้/fever*)/VP)

EDU4 (symptom): " [*ผม*]*เป็นมาประมาณ 2 วันครับ/*[*I*] ***have the symptoms about 2 days.***"

([*ผม/I*] (*เป็นมา/get* [*symptom*] *ประมาณ 2 วันครับ/about 2 days*)/VP)

EDU5: "[*ผม*]*สงสัยเป็นโรคหลอดลมอักเสบ/* [**I**] ***doubts to get bronchitis***? "

([*ผม/I*] (*สงสัย/doubt เป็น/get โรคหลอดลมอักเสบ/bronchitis*)/VP)

where the [..] symbol means ellipsis, NP is a noun phrase, and VP is a verb phrase.

A symptom-concept EDU boundary occurs on EDU1, EDU2, EDU3 and EDU4. According to Example1, the research emphasizes on the event expressions by verb phrases because of most symptom-concept expressions on the verb phrases of EDUs. Each EDU is based on the following Thai linguistic pattern after stemming words and eliminating stop words.

EDU → NP1 VP

VP → V1 | V1 NP2 | V1 Adverb | V2 NP3 | V2 NP3 VP | V2 Adj

V1 → V_{strong} | Preverb V_{strong}

V2 → V_{weak} | Preverb V_{weak}

NP1→ Noun1 | Noun2 | Noun3

NP2 → Noun2 | Noun2 NP2 | Noun2 AdjectivePhrase

NP3→ Noun3 | Noun3 Adj prep NP2

Noun1→{'*ผู้ป่วย/patient*' '*โรค/disease*'...} ;

Noun2→{'*อวัยวะ/organ*''*บริเวณ/area*' '*อุจจาระ/stool* ' ...}

Noun3→{'*อาการ/symptom*' '*แผล/scar*' '*รอย/mark*' '*ไข้/fever*' '*ผื่น/rash*' '*หนอง/pus*' ...}

V_{strong}→{'*คลื่นไส้/nauseate*' '*อาเจียน/vomit*' '*ปวด/pain*' '*เจ็บ/pain*' '*แน่น/constrict*' '*คัน/itchy*'...}

V_{weak} → {'*เป็น/be*' '*มี/have*''*รู้สึก/feel*'}

Adv→ {'*ยาก/difficultly*' .... } ; Adj →{'*สี.../...color*' '*เหลว/watery*' ....} ;

Preverb → {'*ไม่/not*' ... }

where NP1, NP2, and NP3 are noun phrases. V_{strong} is a strong verb set with the symptom concept. V_{weak} is a weak verb set which need more information to have the symptom concept. Noun3 is a noun set with a symptom concept. Adv is an adverb set with the symptom concept. Adj is the adjective set with the symptom concept. prep is a preposition.

There are several techniques [2],[3],[4],[5] having been used for event extraction from text (see section 2). However, the Thai documents have several specific characteristics, such as zero anaphora or the implicit noun phrase, without word and sentence delimiters, and etc. All of these characteristics are involved in two main problems of extracting the explanation of the symptom-concept EDUs. The first problem is how to identify an EDU verb phrase having symptom concept whilst some verb phrases contain V_{weak} which needs some following words to provide the symptom concepts. Thus, we apply the different word co-occurrence sizes; the co-occurrence between two words (called Word-Co) and the co-occurrence between N words (called N-Word-Co), on the EDU verb phrase with the V_{strong}/V_{weak} element as the first word of the co-occurrence for the comparative study of using Word-Co and N-Word-Co to

determine the symptom-concept EDU. Where N-Word-Co size (or the N value) is solved by Support Vector Machine (SVM) learning [6]. The second problem is how to determine the symptom explanation as the symptom-concept EDU boundary, i.e. EDU1-EDU4 of Example1 (see section 3.2). With regard to the second problem, we need to develop a framework which combines the machine learning technique and the linguistic phenomena to learn the several EDU expressions of the disease-symptom explanation on the health-care hospital web boards. Therefore, we propose the SVM learning to solve the boundary having Word-Co as input features and the similarity score [7] to solve the boundary having N-Word-Co as input features.

Our research is organized into 5 sections. In section 2, related work is summarized. Problems in extracting the disease symptom explanation from the documents are described in section 3 and section 4 shows our framework for extracting the disease symptom explanation from the documents. In section 5, we evaluate and conclude our proposed model.

2 Related Work

Several strategies [2],[3],[4],[5] have been proposed to solve the event extraction from text.

In 2011, [2] applied syntactic and lexical constraints on binary relations expressed by verb phrases (called relation phrases) for the Open Information Extraction system, REVERB. They implemented a verb-noun combination on the relation phrase to match the POS tag pattern. The research results with more than 30% of REVERB's extractions are at precision 0.8.

S.Ando et al.[3] proposed methods for filtering harmful sentences based on multiple word co-occurrences. They compare harmless rate between two-word co-occurrence and three-word co-occurrence. The precision of identify and filtering the harmful sentences through three-word co-occurrence method exceeds 90% whereas the precision of the two-word co-occurrence is under 50%.

In 2014, [4] worked on a model for identifying causality in verb-noun pairs to encode cause or non-cause relation. The result of this research achieves 14.74% and 41.9% F-scores for the basic supervised classifier and the knowledge of semantic classes of verbs respectively.

In 2016, [5] studied the temporal variation in word co-occurrence (i.e. Noun-Noun, Verb-Noun) statistics, with application to event detection. [5] developed an efficient unsupervised spectral clustering algorithm that uncovers clusters of co occurring words which can be related to events in the dataset. The performance of [5] methods for event detection on F-score, obtaining higher recall at the expense of precision informative terms occurring in discrete time frames.

However, most of previous researches identify an event by two-word/three-word co-occurrence without the EDU/simple-sentence boundary consideration as our research. Whilst the symptom-concept expression on each EDU of our research mostly consists of several words, i.e. EDU2 of Example 1.

3 Problems of Extracting Disease-Symptom Concepts

Our research contains two problems of determining the symptom-concept explanation; how to identify an EDU verb phrase having symptom concept and how to determine the symptom-concept EDU boundary.

3.1 How to Identify Verb Phrase having Symptom Concept

According to the hospital's health-care web-boards, there are several verb phrases with/without the symptom concepts as shown in the following Example2:

Example2

EDU1: "หลังจาก/*After* คนไข้/ ***patient*** ((ทาน/***consume***)/*verb* (อาหาร/ ***meal***)/*noun*)/VP
(*After a patient has consumed a meal,*)

EDU2: " [คนไข้/***he***] ((มี/***have***)/*weak-verb* (ไข้/***fever***)/*noun*)/VP"
([*he*] *has a fever.*)

EDU3: "และ/***and*** [คนไข้/***he***] ((มี/***have***)/*weak-verb* (อาการ/***symptom***)/*noun* (อุจจาระ/ ***stools***)/*noun* (เหลว/***watery***)/ *Adj* (หลาย/***several***)/*Adj* (ครั้ง/ ***times***)/*noun*)/VP"
(*and* [*he*] *has a symptom of watery stools within several times.*)

According to Example2, the verb phrases (VP) of EDU2 and EDU3 have the weak verbs with the symptom concepts whereas EDU1 having VP without the symptom concept. Thus, the research applies Word-Co and N-Word-Co on the verb phrases (which contain w_1 as either $v_{strong} \in V_{strong}$ or $v_{weak} \in V_{weak}$ and the co-occurred word as $w_2 \in$ Noun3 ; Noun3 exists in either NP3 or NP1) to identify the verb phrase having the symptom concepts. Using the N-Word-Co to identify the verb phrase with the symptom concept has another problem of how to determine the size of N-Word-Co or the N value, i.e. in Example2 having the EDU2 verb phrase with N-Word-Co as '***มี/have*** ***ไข้/fever***' (N=2) and the EDU3 verb phrase with N-Word-Co as '***มี/have*** ***อาการ/symptom*** ***อุจจาระ/stools*** ***เหลว/watery***' (N=4). Thus, we apply the SVM learning to solve the N value (by sliding the window size of two consecutive words with one sliding word distance after stemming words and the stop word removal).

3.2 How to Determine Symptom-Concept EDU Boundary

In regard to Example1 and the following Example3, how to determine the symptom-concept EDU boundary is challenge.

Example3

EDU1: "[ผม] ไม่มีน้ำมูก/ [*I*] *do not have mucus.*"

EDU2: "[ผม] ไม่ไอ /[*I*] *do not cough.*"

EDU3: "แต่[ผม] คัดจมูกข้าง /*But* [*I*] *have a congested nose.*"

EDU4: "[ผม] เจ็บคอ/[*I*] *have sore throat .*"

EDU5: "[ผม] ทานยาแก้อักเสบ/[*I*] *take the antibiotic medicine.*"

EDU6: "*แต่มันก็ไม่หาย/ but it does not work.*"

where Example3 has the symptom-concept EDU boundary occurrence on EDU1 through EDU4. Therefore, we propose the SVM learning to solve the boundary having Word-Co as input features (by sliding the window size of two consecutive EDUs with one sliding EDU distance) and the similarity score<0.9 to solve the boundary having N-Word-Co as input features.

4 A Framework for Extracting Disease-Symptoms

There are three main steps in determining the disease-symptom explanation for each document topic name by using the Word-Co or N-Word-Co technique, Corpus Preparation step, Learning Step and Symptom-Concept EDU Boundary Determination Step, as shown in Fig.1.

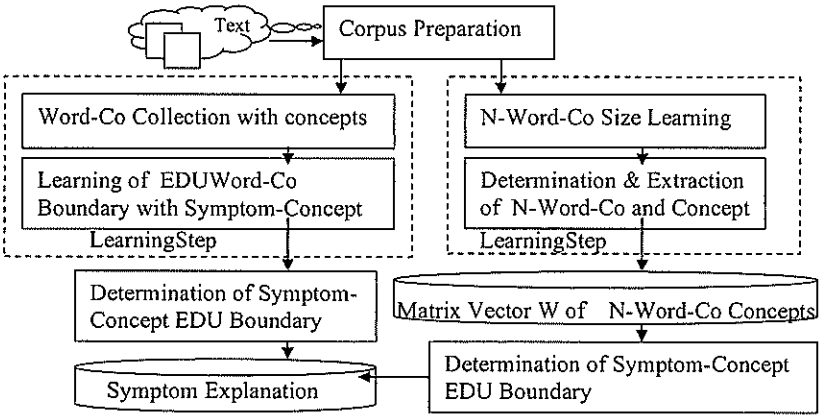

Fig.1. System Overview

4.1 Corpus Preparation

This step is the corpus preparation in the form of EDUs from the medical-care documents on the hospital's web-board (http://haamor.com/). The step involves using Thai word segmentation tools [8] including Name entity [9] followed by EDU segmentation [10]. These annotated EDUs are used as an EDU corpus which contains 3000 EDUs of gastrointestinal tract diseases and childhood diseases and is separated into 2 parts; the first part of 2000 EDUs for the learning step of both Word-Co and N-Word-Co; and the second part of 1000 EDUs for determining the symptom-concept EDU boundary. We then semi-automatically annotate the Word-Co expressions with symptom concepts for the w_1 and w_2 tags as Word-Co and for the w_1 through w_i as N-Word-Co after stemming words and the stop word removal as shown in Fig.2. All symptom concepts are referred to WordNet (http://word-net.princeton.edu/obtain) and

MeSH (https://www. nlm.nih.gov/mesh/) after translating from Thai to English, by Lexitron (http://lexitron.nectec.or.th/).

Disease Topic : โรคเกี่ยวกับทางเดินอาหาร / Gastrointestinal tract disease

EDU: [*ผู้ป่วย*]*รู้สึกแน่นที่หน้าอกด้านขวาเป็นบางครั้ง*

[*ผู้ป่วย/A patient*] *รู้สึก/feel แน่น/press against ที่/at หน้าอก/chest ด้านขวา/right side เป็นบางครั้ง*/sometime

<EDU1> ([(*ผู้ป่วย/A patient*)/ncn])/NP1

(<**Word-CoExpression** location = chest from Noun2>

< w_1: setType='weak-verb' ; concept= 'feel' boundary = 'y'>*รู้สึก*</ w_1>

< w_2: setType='strong-verb' ; concept='oppress/press against' boundary ='y'>*แน่น*</ w_2>

< w_3: setType='Noun2' ; concept= 'chest/organ' boundary= 'y'>*หน้าอก*</w_3>

< w_4: setType='Adj' ; concept= 'right side' boundary= 'y'>*ด้านขวา*</w_4>

< w_5: setType='Adv' ; concept= 'sometime' boundary= 'n'>*เป็นบางครั้ง*</w_5>

</ **Word-CoExpression**>)/VP </EDU1>.....

The Word-CoExpression tag is the word boundary tag including 2-Word-Co expression (w1 and w2) and N-Word-Co expression (w1 through w4 with boundary property= 'y'). The w1 tag is the verb tag and the w_i tag is the co-occurred word$_i$ tag where i=2,3,..,*num*.

The [..] symbol means ellipsis (Zero Anaphora)

Fig.2. Word Co-Occurrence Annotation

4.2 Learning

4.2.*1 Word-Co Learning*. We collect each Word-Co feature, w_1w_2 or $v_{co}\ w_{co}$, with the symptom concept into VW from annotated corpus where VW is a Word-Co set with the symptom concepts; w_1 is a verb represented by v_{co} ; and w_2 is a co-occurred word represented by w_{co}. VW is used for identifying and extracting the consecutive symptom-concept Word-Co occurrences for learning the EDU's Word-Co boundary with the symptom concept by SVM (using Weka, ,http://www.cs. wakato.ac.nz/ml/weka/). SVM is the linear kernel: the linear function, $f(x)$, of the input x =(x_1..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

$$f(\mathrm{x}) = \langle \mathrm{wt} \cdot \mathrm{x} \rangle + b = \sum_{j=1}^{n} wt_j x_j + b \qquad (1)$$

where x is a dichotomous vector number, wt 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 wt_j and b for each Word-Co feature, $v_{co\text{-}i}\, w_{co\text{-}i}(x_j)$ in each Word-Co pair, $v_{co\text{-}i}\, w_{co\text{-}i}\, v_{co\text{-}i-1}\, w_{co\text{-}i-1}$, from the supervised learning of SVM by sliding the window size of two consecutive EDUs with one sliding EDU distance where j = 1, 2, ..., n and n is End-of-Boundary.

4.2.2 *N-Word-Co Learning*. In regard to equation 1, the features used for learning N-Word-Co size by SVM are obtained by the following concept sets: $Verb_{strong}$,$Verb_{weak}$, Noun2, Noun3,Adj, and Adv . The SVM learning is to determine wt_j and b for each

word feature, w_j (or x_j) in each word-concept pair (w_j w_{j+1},) with a symptom concept. The N-Word-Co size/boundary learning from $w_j w_{j+1}$ (where $w_1 \in V_{strong} \cup V_{vweak}$; $w_j w_{j+1} \in$ Noun2∪Noun3∪$Verb_{strong}$∪$Verb_{weak}$∪Adj∪Adv; j=2,3,..n) of VP is 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.

4.3 Symptom-Concept EDU Boundary Determination

4.3.1 Symptom-Concept EDU Boundary Determination by Using Word-Co. After using VW to identify a symptom concept EDU from the testing corpus, the wt vector of all $w_{co\text{-}j}$ $w_{co\text{-}j}$ from the SVM learning are used to determine the boundary of the symptom-concept EDUs with equation 1 by sliding the window size of two consecutive EDUs with one sliding EDU distance. If $f(x)<0$ then the boundary is ended as the symptom-concept EDU boundary; otherwise continuing.

4.3.2 Symptom-Concept EDU Boundary Determination by Using N-Word-Co. The symptom- concept EDU boundary is determined after the N-Word-Co size determination and extraction. After $w_1 \in V_{strong} \cup V_{weak}$ and w_1 is the first word of VP on the testing corpus, the wt vector of all w_j from the SVM learning in section 2.2.2 which are used to determine and extract the N-WordCo size/boundary with symptom-concept collected into the matrix vector (W) of symptom concepts with equation 1 by sliding the window size of two consecutive words with one sliding word distance.

Table1. The N-Word-Co expression on the health care documents

N-Word-Co Occurrence on VP	Symptom concept
'เป็น/*be* ผื่น/*rash* แดง/*red*'	To occur red rash
'เป็น/*be* แผล/*scar* พุพอง/*blister*'	To occur blister mark
'เป็น/*be* แผล/*scar* อักเสบ/*inflame*'	To occur inflamed mark
……	……
'แน่น/*constrict* หน้าอก/*chest*'	To constrict chest pain
'แน่น/*constrict* ท้อง/*abdominal*'	To constrict abdominal pain
……	……
'รู้สึก/*feel* แน่น/*constrict* หน้าอก/*chest*'	To constrict chest pain
'รู้สึก/*feel* คลื่นไส้/*be nauseate*'	To be nauseate
'รู้สึก/*feel* เวียนศีรษะ/*dizzy*'	To be dizzy
'รู้สึก/*feel* ปวด/*pain* ศีรษะ/*head*'	To have an headache
'รู้สึก/*feel* ปวด/*pain* ท้อง/*abdominal*'	To have an abdominal pain
……	……
'มี/*have* ไข้/*fever*'	To have a fever
'มี/*have* รอย/*lesion* ช้ำ/*blue*'	To occur blue lesion
……	……
'มี/*have* อาการ/*symptom* คลื่นไส้/*nauseate*'	To occur nauseated symptom
'มี/*have* อาการ/*symptom* ปวด/*pain*'	To occur pain
'มี/*have* อาการ/*symptom* ปวด/*pain* ศีรษะ/*head*'	To have an headache
'มี/*have* อาการ/*symptom* ปวด/*pain* ท้อง/*abdominal*'	To have an abdominal pain
……	……

If $f(x) < 0$ then the boundary is ended as a word vector of N-Word-Co; otherwise continuing. All extracted N-WordCo expressions are collected into W of symptom concepts as shown in Table1. The symptom-concept EDU boundary is then deter-

mined by the similarity score determination as Max Similarity Score (MaxSimScore) [6] in equation 2. MaxSimScore is determined between the N-Word-Co of the testing corpus's EDU and the candidate N-Word-Co expressions from W. The N-Word-Co concept of each consecutive EDU verb phrase is the symptom concept if MaxSimScore≥0.9 to W; otherwise the symptom vector is ended.

$$MaxSimScore = ArgMaxSimlarity_{t=1}^{Cardinality}\left(\frac{|NWCcorpus \cap NWCcandidate_t|}{\sqrt{|NWCcorpus|\times|NWCcandidate_t|}}\right) \quad (2)$$

where Cardinality is thenumber of N - Word- Co elementsof W
W is theMatrix vectorof N - Word- Co (theN - Word- Co set) with thesymptomconcept
NWCcandidate is a candidate N - Word- Co element of theN - Word- Co set with thesymptomconcept
NWCcorpus is an N - Word- Co of EDU fromthetestingcorpus.

4.4 Evaluation and Conclusions

Table 2. Evaluation of Symptom Vector Determination from Web Documents

Health-Care-Symptom Corpus	Correctness of Symptom Vector Determination			
	Using N-Word-Co		Using 2-Word-Co	
	Precision	Recall	Precision	Recall
Gastrointestinal tract diseases 500EDUs	92.4%	63.05%	84.2%	70%
Childhood diseases 500EDUs	90.2%	75.4%	85.4%	76.2%

The testing corpus of 500 EDUs of gastrointestinal tract diseases and 500 EDUs of childhood diseases collected from the hospital web sites is used for evaluating the symptom-concept EDU explanation/boundary determination from texts. Both evaluations of the symptom-concept EDU explanation determinations by using Word-Co and by using N-Word-Co from the testing corpus are based on the precision and the recall which are evaluated by three expert judgments with max win voting. The average of precisions of determining the symptom-concept EDU explanation are 91.3% and 84.8% with average recalls of 69.2% and 73.1% by using N-Word-Co and Word-Co respectively, as shown in Table2. The reason of low recall is the anaphora problem, especially with Noun3. For example: there are some pronoun words, i.e. '(*บางสิ่ง/something*)/*pronoun*' '(*อะไร/something*)/*pronoun*', appearing among the consequence words of some verb phrases with the symptom concept, which result in the low recall as shown in the following

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

However, the research results provide the higher precision by using N-Word-Co to determine the symptom-concept EDU explanation from the documents because N-Word-Co contains more information. However, the results also provide the higher recall by using Word-Co to determine the symptom-concept EDU explanation from the documents because Word-Co is more general than N-Word-Co. Thus, the symptom-concept EDU explanation are determined and extracted to construct the semantic relation as the diseaseName-symptomExplanation relation where the disease-names are obtained by the document topics. The diseaseName-symptomExplanation relation is beneficial to the automatic diagnosis of the problem solution. Moreover, the proposed method of using either N-Word-Co or Word-Co to determine the information or knowledge can also be applied to the other areas such as the industrial finance problems.

Acknowledgements

This work has been supported by The Thailand Research Fund. The medical-care knowledge and the pharmacology knowledge applied in this research are provided by Puangthong Kraipiboon, a clinician of Division of Medical Oncology, Department of Medicine, Ramathibodi Hospital, and Uraiwan Janviriyasopak, a pharmacist of Rex-Pharmcy, respectively.

References

1. Carlson, L., Marcu, D., and Okurowski , M. E.: Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Current Directions in Discourse and Dialogue, pp.85-112 (2003).
2. Fader,A., Soderland,S. and Etzioni ,O.: Identifying Relations for Open Information Extraction. In Proc. of the Conference on Empirical Methods in Natural Language Processing, pp.1535-1425(2011).
3. Ando,S., Fujii,Y. and Ito, T.: Filtering Harmful Sentences based on Multiple Word Co-occurrence. IEEE/ACIS 9th International Conference on Computer and Information Science(ICIS), (2010).
4. [4] Riaz , M. and Girju, R.: Recognizing Causality in Verb- Noun Pairs via Noun and Verb Semantics. In Proc. of the EACL 2014 Workshop on Computational Approaches to Causality in Language, pp. 48-57(2014).
5. Preoţiuc-Pietro, D. , Srijith, P.K., Mark ,H., and Trevor, C. Studying the temporal dynamics of word co-occurrences: An application to event detection. In LREC.(2016).
6. Mitchell, T.M.: Machine Learning. The McGraw-Hill Companies Inc. and MITPress, Singapore (1997).
7. Biggins, S., Mohammed, S., Oakley, S.: University Of Sheffield: Two Approaches to Semantic Text Similarity. In Proceedings of First Joint Conference on Lexical and Computational Semantics, Montre´al, Canada, pp.655-661 (2012).
8. Sudprasert , S. and Kawtrakul , A.: Thai Word segmentation based on Global and Local Unsupervised Learning. In Proceedings of the 7th National Computer Science and Engineering Conference, (2003).

9. Chanlekha, H. and Kawtrakul, A.: Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. First International Joint Conference,Hainan Island,China, (2004).
10. Chareonsuk,J. Sukvakree,T., and Kawtrakul, A.: Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information. In Proceedings of the 9th National Computer Science and Engineering Conference, (2005).