Extraction of cause-effect-concept pair series from web documents

Chaveevan Pechsiri¹, Titirut Mekbunditkul²
¹College of Innovative of Technology and Engineering, Dhurakij Pundit University, Thailand
²Research Service Center, Dhurakij Pundit University, Thailand

ABSTRACT

This research aims to extract a cause-effect-concept pair series of consequent event occurrences in health information of hospital web-boards. The extracted cause-effect-concept pair series representing a disease causation pathway benefits for the automatic diagnosis and solving system. Where each causative/effect event concept is expressed by an elementary discourse unit (EDU which is a simple sentence). The research has three problems; how to determine causative/effect concept EDUs from the documents containing some EDU occurrences with both causative concepts and effect concepts, how to determine the cause-effect relation between two adjacent EDUs having the discourse cue ambiguity, and how to extract cause-effect-concept pair series mingled with either a stimulation relation EDU or other non-cause-effect relation EDUs from the documents. Therefore, we apply annotated NWordCo pairs with causative-effect concepts to represent EDU pairs with causative-effect concept where the NWordCo size solved by Naïve Bayes. We also apply Naïve Bayes to solve NWordCo-concept pairs having the cause-effect relation from the adjacent EDU pairs. We then propose using cue words and the collected NWordCo-concept pairs with the cause-effect relation to extract the cause-effect-concept pair series. The research results provide the high precision of the cause-effect-concept pair series determination from the documents.

Keywords:
Cause-effect-concept pair series
Elementary discourse unit
NWordCo
Ordered pair

1. INTRODUCTION

The objective of this paper is to extract a Cause-Effect-concept pair (called ‘CEpair’) series which is a series of cause-effect-event concept pairs of a disease causation pathway, from hospital web-board documents (i.e. http://haamor.com; http://www.si.mahidol.ac.th/sidoctor/e-pl/). Whilst ‘series’ means ‘a group or a number of related or similar things, events, etc., arranged or occurring in temporal, spatial, or other order or succession; sequence.’ (http://www.dictionary.com/). The CEpair series of the research is a group of cause-effect-event ordered pairs occurring in the CEpair sequence as in a document. Regard to the disease causation pathway for chronic disease particular diabetic, cardiovascular, kidney diseases [1-3], each CEpair is an ordered pair (c, e) with the cause-effect relation where c is a causative-event concept and e is an effect-event concept. Where each cause/effect event concept on each CEpair element, CEpair, is expressed by an elementary discourse unit (EDU which is a simple sentence, [4]) as follows:

CEpair1, CEpair2, ..., CEpairlast

where CEpairi (i=1,2, ..., last which is an integer) is an expression of the cause-effect relation between a
causative-event concept EDU and an effect-event concept EDU, from two adjacent EDUs as an EDU pair as shown in Example 1.

Example 1:

...EDU1: “A patient gets a diabetes disease.”
“เป็นโรคเบาหวาน”

EDU2: “since the body cannot fully use sugar in the body.”
“เพราะถังน้ำตาลไม่สามารถใช้ได้อย่างเต็มที่”

EDU3: “the diabetes disease will be a significant risk factor to a brain disease, a heart disease, and a kidney disease.”
“โรคเบาหวานจะเป็นปัจจัยเสี่ยงที่สำคัญต่อโรคสมอง, โรคหัวใจ และโรคไต”.

Thus, the diabetes disease will be a significant risk factor to a brain disease, a heart disease, and a kidney disease. Therefore, the research concerns to extract the CEpair series benefits for improvement of the public’s understanding of a complex problem of a certain chronic disease to follow up physician’s suggestion of solving steps. Therefore, the research emphasizes on the EDU’s verb phrase expressions because the CEpair series is based on several events that each event concept is mostly expressed by an EDU’s verb phrase. The EDU expression has the following Thai linguistic patterns after stemming words and the stop word removal.

Example 1 is then represented by the CEpair series containing EDU7 as an intervening EDU of the stimulation relation as shown in the following.

EDU4-EDU7 Pair as CEpair: EDU4 (Cause) → EDU5 (Effect) → EDU6 (Cause) → EDU7 (Effect)

where the […] symbol means ellipsis.
There are several techniques [5-12] having been applied for determining the cause-effect/causality/causal relation but not including the stimulation relation from texts (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 three main problems (see Section 3). The first problem is how to determine causeative-concept/ effect-concept EDUs from the documents containing some EDU occurrences with both causeative-concepts and effect-concepts. The second problem is how to determine the cause-effect relation between two adjacent EDUs as an EDU pair with a discourse cue ambiguity. And the third problem is how to extract CEPair series mingled with either a stimulation relation EDU or other non-cause-effect relation EDUs from the documents. Regarding these problems, we need to develop a framework which combines machine learning and the linguistic phenomena to represent each EDU event concept by n-word co-occurrence (called NWordCo) on the EDU verb phrase as shown in (1) where NWordCo is expressed as compound terms with/without any pattern or restriction depending on each research perspective as [13-16]. The reason of using NWordCo to represent an EDU event is the Verbreak element which needs more information from some linguistic sets, i.e. Noun, Adj, Verb and Adv, to form the causative/effect concept or the stimulating concept. The NWordCo expression of the research starts with v1 (where v1 ∈ Verb ∪ Verbbreak) followed by the N-1 co-occurred words (N is an integer) from the EDU verb phrase as shown in the following (1) after stemming words and eliminating stop words.

\[ \text{NWordCo} \text{' expression } = v_1 + w_2 + ... + w_N \]

where \( v_1 \in \text{Verb} \cup \text{Verbbreak} \); \( w_2, w_3, w_4 \in \text{Noun} \cup \text{Adj} \cup \text{Adv} \cup \text{Verb} \)

Thus, we apply an annotated NWordCo-expression pairs with causative-effect-event concepts to represent a cause-effect relation including an annotated NWordCo with stimulating-event concept. We then apply Naïve Bayes (NB) [17] to learn the NWordCo size (which is an N value) to extract and collect NWordCo with concepts into an NWordCo-Concept (NWC) set from the testing corpus. We also use NB to learn probabilities of NWordCo-concept pairs with a CauseEffectRelation class and a non-CauseEffectRelation class from the learning corpus having the discourse cue ambiguity. We then identified and extract all NWordCo-concept pairs having the cause-effect relation by using the NB-learning probabilities of NWordCo-concept pairs with the CauseEffectRelation class from the learning corpus to the Cartesian product of the NWC sets from the testing corpus. Later, we collect the extracted NWordCo-concept pairs into an NWCPce set (which is an ordered pair set of NWordCo-concept pairs with the CauseEffectRelation class) as shown in the following.

\[
\text{NWCPce} = \{ \text{NWordCo}, \text{NWordCo}_{-}\text{-pair}, \text{NWordCo}_{-}\text{NWordCo}_{-}\text{-pair}, ... , \text{NWordCo}_{-}\text{NWordCo}_{-}\text{-pair}_\text{1}\} \\
\text{where } \text{NWordCo}_{-}\text{NWordCo}_{-}\text{-pair} \text{ is an NWordCo-concept pair having the cause-effect relation between NWordCo, and NWordCo (in which NWordCo is an NWordCo with a causative concept and NWordCo is an NWordCo with an effect concept);} i=1,2,...,\text{last which is an integer.}
\]

We then propose using NWCPce and the stimulating-cue-word set, \{‘เข้ารับกระตุ้น’-Verbreak catalysis- Noun’, ‘เข้ารับกระตุ้น’-Vstrong’, ‘เข้ารับกระตุ้น’-Vweak’ \} to extract the CEPair series including a stimulation relation EDU from another testing corpus (see section 3).

Our research is organized into 5 sections. In Section 2, related work is summarized. Problems in extracting the CEPair series from texts are described in Section 3 and Section 4 shows our framework of extracting the CEPair series. In Section 5, we evaluate and conclude our proposed model.

2. RELATED WORKS

Several strategies [5-12] have been proposed to determine the cause-effect relation from texts without the cause-effect series consideration except [12]. Reference [5] applied Text Mining to cluster the effects/symptoms of the causes/diseases from pathology reports having effect expressions as complicated technical terms based on NP. All clusters benefitted of the ability in grouping patients with the similar condition. Regarding [6], Girju proposed decision tree learning the causal relation from a sentence based on the lexico syntactic pattern (NP1 causal-verb NP2). Reference [7] determined event knowledge as a causal relation (based on the lexico-syntactic pattern, NP1 verb NP2) including the causal association/strength measurement from web-texts. Reference [8] extracted the causal knowledge from two adjacent sentences by using SVM to learn several features as a shared agent (NP1) from causative and effective clauses, causal volition, the verb class from the dictionary, verbal semantic attributes, the connective marker, and the modality for classifying the causal knowledge into four classes of causal relations: cause, precondition, mean, and effect relations. Reference [9] applied verb-pair rules and machine learning techniques to extract the causality occurrence within several effect EDUs. There are more research works based on the lexico syntactic

Indonesian J Elec Eng & Comp Sci, Vol. 18, No. 2, May 2020 : 970 - 978
pattern with the causal concept as in [10] proposed the Restricted Hidden Naïve Bayes model to learn and extract the causality from the English documents. The learning features as in [10] include contextual, syntactic, position, and connective features. Reference [11] applied the rule-based, Support Vector Machine and the temporal reasoning to extract the causal relation on a complex sentence or two simple sentences from English documents. Reference [12] made causal chains by adding the causal chains obtained from latent topics to the causal chains obtained from word matching. The model’s [12] is based on noun features including hidden causal chains solved by latent topics. However, most of the previous works on the cause-effect relation are based on noun/NP features existing on one/two sentences without the series consideration except [12] whereas our work has NP ellipsis occurrences on documents. There are few works on extracting the CEpair series as a disease causation pathway.

3. PROBLEMS OF EXTRACTING CEPAIR SERIES FROM TEXTS

3.1. How to Determine Causative-Concept/Effect-Concept EDUs from Documents

There are some EDU occurrences with both causative-concepts and effect-concepts, i.e. EDU2 and EDU8 of Example 1 on CEpair1 to CEpair3 and CEpair4, respectively. It is difficult to identify the certain EDU occurrence as the causative concept or the effect concept. Therefore, after stemming words and eliminating stop words, we apply the annotated NWordCo pairs with cause-effect relation on the learning corpus to represent the causative/effect concept EDUs and the annotated NWordCo with stimulating concept. If the first word of each EDU verb phrase is the element of Verbset1, the NWordCo size is then solved by NB learning on the consecutive words of each annotated verb phrase with a slide window size of two adjacent words with one word sliding distance on each EDU verb phrase. The NWordCo extraction is then occurred after the NWordCo sizes have been solved.

3.2. How to Determine CEpairs as Cause-Effect Relation with Discourse-Cue Ambiguity

The CEpairs, expression as the cause-effect relation between two adjacency EDUs as an EDU pair can be determined by using the discourse-cue set, {‘เพราะ’, ‘because’, ‘เมื่อจาก’, ‘since’, ‘ทำให้’, ‘cause’...}, see Example 1. However, some discourse-cue set elements are ambiguity. For example: CEpairs of Example 1 has a discourse cue, ‘เมื่อจาก’, on EDU2 whereas an EDU1-EDU2 pair of the following Example 2 having ‘because’ on EDU2 is not the CEpairs expression.

Example 2

EDU1: “ผู้ป่วยเบาหวานอาจเป็นโรคหัวใจ” (A diabetic patient might get the heart disease.)
EDU2: “เพราะการรักษาไม่ดี” (Since a blood sugar level is high.)
EDU3: “เวลานี้ความสูงน้ำตาลในเลือด” (In blood sugar level high.)

The high blood sugar level (EDU2) causes of having some increased chemical substance types in blood.) ...
4.1. Corpus Preparation

This step is to prepare an EDU corpus from the chronic disease documents, i.e. diabetes, heart disease, artery disease etc., downloaded from hospitals web-boards (http://www.bangkokhealth.com; http://haamor.com/). The step involves using Thai-word-segmentation tools [18] and Named-Entity recognition [19-20]. After the word segmentation is achieved, EDU Segmentation [21] based on [22-24] is then operated to provide a 2500 EDUs’ corpus. The corpus included stemming words and the stop word removal is separated into 3 parts; the first part of 1000 EDUs for learning the NWordCo sizes/boundaries having causative/effect/stimulating concepts and also learning the NWordCo-concept pairs having the cause-effect relation. The second part of 1000 EDUs is the testing corpus used for the NWordCo size determination to extract and collect NWordCo occurrences with causative/effect/stimulating concepts into the NWC set. The NWC set is used for collecting NWordCo-concept pairs with the cause-effect relation into the NWCp set. The third part of 500 EDUs is used for the CEpairs series extraction. This step also includes semi-automatic annotation of each NWordCo size along with the causative/effect/stimulating concept as shown in Figure 2 [25]. This step also annotates the EDU pairs as the NWordCo-concept pairs with the cause-effect relation. All word concepts of each NWordCo expression is referred to Wordnet (http://wordnet.princeton.edu/ obtain) [26] and MeSH after translating from Thai to English by Lexitron (http://lexitron.nectec.or.th/).

Figure 1. System overview

```
Figure 2. Annotation of NWordCo and CEpairs series
```

---

 Indonesian J Elec Eng & Comp Sci, Vol. 18, No. 2, May 2020 : 970 - 978
4.2. NWordCo Size Learning

This step is an NWordCo size/boundary learning (N value) by the NB classifier [17] from the annotated verb phrases with the concepts from the corpus preparation step. The annotated NWordCo occurrences with causative/effect/stimulating concepts are separated into 2 word-concept vectors (Wj) in a matrix (W).

W_{\text{Co}} = \{w_{j1}, w_{j2}, ..., w_{jk}\} \text{ where CorEorS (non-CorEorS)} where CorEorS is an NWordCo/a word vector with a causative/effect/stimulating concept and non-CorEorS is an NWordCo/a word vector with a non-causative/effect/stimulating concept, existing in EDU1, EDU2, ..., EDUm.

W = \{Wj\} where j=1..m; after we have obtained the annotated word features including the stop word removal and stemming words, we then determine the probabilities of CorEorS concept and non-CorEorS concept from a slide window size of two consecutive words on the verb phrase with the one-sliding-word distance by using Weka (http://www.cs.wakato.ac.nz/ml/weka/).

4.3. Collection of NWordCo with Event Concepts

After stemming word and eliminating stop words of the testing corpus, if w_j is the first word of EDUj verb phrase, the NWordCo size is then determined by using NB in (2) and the learnt probability of CorEorS concept and non-CorEorS concept from the previous step of IV.B to determine the consecutive words on the verb phrase with a one-sliding window size of two adjacent words with the one-sliding-word distance. As soon as class=‘non-CorEorS’ is determined, the NWordCo boundary/size is solved as shown in the NWordCo extraction algorithm of Figure 3. In regard to Figure 3, the extracted NWordCo expressions in NWCSet (which is the NWordCo-concept set, NWC) from the testing corpus is collected with the concepts according to the sequence of word concepts as shown in Table 1 consisting of the causative-NWordCo, effect-NWordCo, and stimulating-NWordCo concepts.

Assume that each EDU is represented by (NP1 VP).
L is a list of EDUs after stemming words and the stop word removal.
Verbs=W_{\text{break}} \cup W_{\text{strong}} (where w_j is the first word of EDUj verb phrase), the NWordCo size is then determined by using NB in (2) and the learnt probability of CorEorS concept and non-CorEorS concept from the previous step of IV.B to determine the consecutive words on the verb phrase with a one-sliding window size of two words on the verb phrase with the one-sliding-word distance. As soon as class=‘non-CorEorS’ is determined, the NWordCo size is then determined by using NB in (2) and the learnt probability of CorEorS concept and non-CorEorS concept from the previous step of IV.B to determine the consecutive words on the verb phrase with a one-sliding window size of two words on the verb phrase with the one-sliding-word distance. As soon as class=‘non-CorEorS’ is determined, the NWordCo size is then determined by using NB in (2) and the learnt probability of CorEorS concept and non-CorEorS concept from the previous step of IV.B to determine the consecutive words on the verb phrase with a one-sliding window size of two words on the verb phrase with the one-sliding-word distance.

Figure 3. NWordCo Extraction Algorithm

```
NWORDCO_EXTRACTION
1  NWCSet<->; NWco<> ; i=1 ; j=1 ; k=0 ; fl='no'
2  while j< Length[L] do
3     {  If i=1 then /* identify the 1\text{st} word of NWordCo
4        {  If (evp \in Verbs_{\text{break}} \cup Verbs_{\text{strong}}) then {  NWco< evp; w ; fl='yes' }
5        Else If ( evp \in Verbs_{\text{break}}) \lor ( evp \in Verbs_{\text{strong}}) then
6            {  NWco< ( evp; w; w_{i+1} ) ; i++; fl='yes' }
7            i++; /* determine N-Word-Co size
8        while (fl='yes') \lor ( evp; w; w_{i} ) \lor (\text{endOfVerbPhrase}) do
9            i=i-1;
10       Equation(2) ;
11        If class=‘nonCorEorS\_concept’ then fl<‘no’
12        Else fl<‘yes’;
13        If class=‘yes’ then NWco< NWco \cup w_{i}
14            i++ ;
15        if NWco<>\emptyset \land fl='no' then /* append new NWordCo
16        {  NWco<> NWco<i=1 ; j+= ; NWco<>\emptyset>}
17 } return NWSet
```

Table 1. NWC Set Collection

| NWC Set Collection | Concept |
|--------------------|---------|
| occur-sugar-blood-high | occur-sugar-blood-high |
| lackOf-hormone | lackOf-hormone |
| have-complication-kidney | have-complication-kidney |
| cause-protein-blood-low | cause-protein-blood-low |
| collect-fat-artery | collect-fat-artery |
| deteriorate-artery | deteriorate-artery |
| loss-protein-urine | loss-protein-urine |

Extraction of cause-effect-concept pair series from web documents (Chaveevan Pechsiri)
\[ NWordCoBoundaryClass = \arg \max_{class} P(\text{class} | w_i, w_{i+1}). \]
\[ = \arg \max_{class} P(w_i | \text{class})P(w_{i+1} | \text{class})P(\text{class}). \]

where \( w_i \in W_i \) and \( w_{i+1} \in W_j \) (\( W_i \) is a CorEorSet_word_concept vector)

\[ \text{Class} \; \in \{ \text{CorEorSet_concept}, \text{non_CorEorSet_concept} \} \]

(2)

4.4. NWordCo-Concept Pair Learning

This step is the NB learning [17] the feature set of NWordCo-concept pairs with the CauseEffectRelation class on several two adjacent EDUs with CEpairID annotation of the learning corpus from the corpus preparation step (section 4.1) after stemming words and eliminating stop words. The learning results of this step by using Weka (http://www.cs.wakato.ac.nz/ml/weka/) are the probabilities of the annotated NWordCo-concept pairs as shown in Table 2.

| NWordCo-Concept Pair: (CausativeNWordCoConcept)(EffectNWordCoConcept) | CauseEffect Rel. Probability | Non-Cause EffectRel. Probability |
|---|---|---|
| (lackOf-hormone)(occur-sugar-blood-high) | 0.0171 | 0.0116 |
| (occur-deteriorated-artery)(constrict-artery) | 0.0053 | 0.0029 |
| (collect-fat-artery)(cause-arteriosclerosis) | 0.0053 | 0.0029 |
| (lossOf-protein-urine)(cause-protein-blood-low) | 0.0132 | 0.0116 |
| (cause-protein-blood-low)(have-symptom-swell) | 0.0020 | 0.0025 |
| (cause-protein-blood-low)(occur-state-kidneyFailure) | 0.0038 | 0.0048 |
| (occur-sugar-blood-high)(deteriorate-artery) | 0.0038 | 0.0048 |

4.5. Extraction of NWCP<sub>e</sub>

The collected NWC set from the previous step of IV.C is used as the CauseConcept set and also the EffectConcept set for determining the Cartesian product of CauseConcept x EffectConcept as NWordCo-concept order pair set, NWCP<sub>e</sub>. We then extract and collect each NWordCo-concept pair with the cause-effect relation into NWCP<sub>e</sub> from NWCP<sub>e</sub> by using the NB classifier in (3) with the NWordCo-concept pair probabilities from Table 2 as the NB feature probabilities.

\[ nwcpClass = \arg \max_{\text{class}} P(\text{class} | nwcpOrdpair_{ik}). \]
\[ = \arg \max_{\text{class}} P(nwcpOrdpair_{ik} | \text{class})P(\text{class}). \]

where \( nwcpClass \) is an NWordCo_conceptpair_class; nwcpOrdpair<sub>ik</sub> \( \in NWCP<sub>e</sub> \);

\[ \text{Class} \; \in \{ \text{CauseEffectRelation, 'nonRelation'} \} \]

\[ k = 1,2,...\text{num}; \text{num} \; \text{is the number of NWCP<sub>e</sub> elements}; \]

(3)

4.6. Extraction of CEpair Series

The objective of this step is to extract the CEpair series by matching \( cnwcp \) to \( nwcp_{e,k} \) as shown in Figure 4 where \( nwcp_{e,k} \in NWCP<sub>e</sub> \; k=1,2,...\text{numberOf_NWCP<sub>e</sub> element}; and \( tmwcp \) is a testing NWordCo-concept pair which is the CEpair expression consisting of two consecutive NWordCo-concept expressions as the testing NWordCo concepts (\( mw_{c1}, mw_{c2} \)) extracted from the testing corpus. If match(\( mw_{c1}, mw_{c2} \)) then Series \( \cup \) Series \( \cup \) tmwcp where Series is the research output. Moreover, the stimulation relation occurrence on one EDU as the part of CEpair series can be identified by using the stimulating-cue-word set.
Assume that each EDU is represented by (NP1 VP)
L is a list of EDU after stemming words and the stop word removal.
NWCPce is the NWordCo-concept pair set with the cause-effect relation.
tnwc is a testing NWordCo-concept pair from the series testing corpus.

5. EVALUATION AND CONCLUSION

There are three evaluations of the proposed research being evaluated by three expert judgments with max win voting: the first evaluation is the extraction of the NWC set with the NWordCo size/boundary consideration from 1000 EDUs of the testing corpus which is also used for the second evaluation. The extraction of NWCPce is evaluated as the second evaluation and the third evaluation is the CEpair series extraction from the other testing corpus of 500 EDUs. The first and the second evaluation are based on the precisions and the recalls within ten fold cross validation whilst the third evaluation is the percentage of correctness. The precisions of extracting the NWC set and the NWCPce set are 0.866 and 0.852 with recall of 0.798 and 0.715 respectively whilst the correctness of the CEpair series extraction is 87.5%. The reason of low recalls in extracting the NWC set and the NWCPce set is that some information of the certain event expressions by verb phrases exists on both NP1 and VP which results in lack of information/concept on the NWordCo expression, i.e., a)EDU:”The deterioration of artery occurs”/NP1 (deterioration of artery)/VP (“The deterioration of artery occurs”) and b) EDU:”The sugar in blood is low”/NP1 (sugar in blood)/VP (“The Sugar in blood is low”), Moreover, these a) and b) examples also effect to the % of correctness of the CEpair series extraction. Hence, the research contributes the methodology to determine the CEpair series for clearly communicating health information and improving health literacy, particularly the disease causation pathway, to people on the social network. And, this network should also provide how to solve problems/effects [27]. Finally, our research can also enhance the diagnosis and solving system of the other areas i.e. the financial services industry.

REFERENCES
[1] J. A. Quinlivan and D. Lam, “Cholesterol Abnormalities are Common in Women with Prior Gestational Diabetes,” J. of Diabetes & Metabolism, Vol.4, DOI:10.4172/2155-6156.1000255, No.4 :255, 2013.
[2] J. R. Petrie, T. J. Guzik, and R.M. Touyz. “Review Diabetes, Hypertension, and Cardiovascular Disease: Clinical Insights and Vascular Mechanisms,” Canadian J. of Cardiology, Vol.34, No.5, pp.575-584, 2018.
[3] H.Shahbazian and I.Rezaei, “Diabetic kidney disease;review of the current knowledge,” JIP, Vol.2, No.2, pp.73-80,2013.

Figure 4. CEpair series extraction
[4] L.Carlson, D. Marcu, M.E. Okurowski, “Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory,” Current and New Directions in Discourse and Dialogue, vol.22, 2003, pp. 85-112.

[5] U. Raja, T. Mitchell, T. Day, and J. M. Hardin, “Text mining in healthcare. Applications and opportunities,” J. of healthcare information management, Vol. 22, No.3, pp.52-56, 2008.

[6] R.Girju, “Automatic detection of causal relations for question answering,” In Proc. of MultiSumQA '03 Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering, Japan, 2003,pp.76-83.

[7] Y.Cao, P.Zhang, J.Guo, and L.Guo, “Mining Large-scale Event Knowledge from WebText,” Procedia Comput. Sci., Vol.29, pp.478–487, 2014.

[8] T. Inui, K. Inui, and Y. Matsumoto, “Acquiring causal knowledge from text using the connective markers,” J. of the information processing society of Japan, Vol.45, No.3, pp.919-933, 2004.

[9] C. Pechsiri and R. Piriyakul, “Explanation knowledge graph construction through causality extraction from texts,” J. of Computer Science and Technology, Vol.25, No.5, pp.1055-1070, 2010.

[10] S.Zhao, T.Liu, S.Zhao, Y. Chen, and J-Y. Nie, “Event causality extraction based on connectives analysis,” Neurocomputing, Vol. 173, pp.1943–1950, 2016.

[11] P.Mirza and S.Tonelli, “CATENA: Causal and Temporal relation extraction from NATural language texts,” In Proc. of COLING, Japan,2016,pp.64-75.

[12] H.Sawamaru and I. Kobayashi, “An approach to extraction of causal chain among events in multiple documents,” SCIS/ISIS, Japan, 2012, pp.1104-1108.

[13] F.Figueiredo, L.Rocha, T.Couto, T.Salles, MA.Gonçalves, and W.MeiraJr, “Word co-occurrence features for text classification,” Information Systems, Vol.36, No.5, pp.843-858.2011.

[14] G-B.Chen and H-Y.Kao, “Word co-occurrence augmented topic model in short text,”IJCLLP, Vol.20, No.2, pp.45-64, 2015.

[15] M. Sedighi, “Application of word co-occurrence analysis method in mapping of the scientific fields (case study: the field of Informetrics),” Library Review, Vol.65, No.3, pp.52-64, 2016.

[16] X. Chen, J. Chen, D. Wu, Y. Xie, and J. Li, “Mapping the research trends by co-word analysis based on keywords from funded project,” Procedia Comput. Sci., Vol.91. DOI: 10.1016/j.procs.2016.07.140. pp.547 – 555,2016.

[17] T.M. Mitchell, Machine Learning, The McGraw-Hill Co. Inc., and MIT Press, Singapore, 1997.

[18] S.Sudprasert and A.Kawtrakul,”Thai Word Segmentation based on Global and Local Unsupervised Learning,” NCSEC2003 Proceedings, Thailand, 2003, pp.1-8.

[19] H.Chanlekha and A. Kawtrakul, “Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information,” In Proc. IJCNLP, Hainan Island, China, 2004, pp.1-7.

[20] N. Tongtep and T. Theeramunkong, “Pattern-based Extraction of Named Entities in Thai News Documents,” Thammasat International Journal of Science and Technology, Vol.15, No.1, pp.70-81, 2010.

[21] J.Chareonsuk, T.Sukvaree, and A.Kawtrakul, “Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information,” NCSEC2005 proceedings, Thailand, 2005, pp.85-90.

[22] S. Sudprasert, A. Kawtrakul, Christian Bottet, and V. Berment, “Dependency Parsing with Lattice Structures for Resource-Poor Languages,” IEICE Transactions on Information and Systems, Vol.E92-D, No.10,pp.2122-2136,2009.

[23] S.Sinthupoun and O.Sornil, “ThaiRhetorical Structure Analysis,” IJCSIS, Vol.7, No.1, pp.95-105, 2010.

[24] N. Ketui, T. Theeramunkong, and C. Onsuwan, “Thai elementary discourse unit analysis and syntactic-based segmentation,” INFORMATION, Vol. 16, No.10, pp.7423-7436.2013.

[25] D.Albright, A.Lanfranchi, A.Fredriksen, W.Figaredo, C.Warner, J.D.Hwang, J.D.Chti, D.Dilgach, RD.Nielsen, J.Martin, W.Ward, M.Pal, -mer, G.K. Savoga, Towards comprehensive syntactic and semantic annotations of the clinical narrative,” Journal of the American Med. Informatics Association, Vol.20, No.5, pp.922-30, 2013.

[26] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller,”Introduction to WordNet: An On-line Lexical Database,” International Journal of Lexicography, Vol. 3. No.4, pp.1-86, 1991.

[27] C. Pechsiri and R. Piriyakul, “Extraction of a group-pair relation: problem-solving relation from web-board documents,” SpringerPlus, Vol. 5: 1265. DOI:https://doi.org/10.1186/s40064-016-2864-3, 2016.