2018 4th International Conference on Science in Information Technology (ICSITech)
October 30-31, 2018 // Melaka, Malaysia

SUBMISSION >>
REGISTRATION >>
PREVIOUS CONFERENCES >>
FUTURE CONFERENCES >>
CONTACT >>
Event-Concept Pair Series Extraction to Represent Medical Complications from Texts

Chaveevan Pechsiri*, Sumran Phainoun
College of Innovative Technology and Engineering, Dhusit University
110/1-4 Prachachuen, Bangkok, Thailand
*chaveevan.pec@dpu.ac.th

Abstract

This research aims to determine an event-concept pair series as consequent events, particularly a cause-effect-concept pair series on disease documents downloaded from hospital-web-boards. These series are used for representing medical complication complications which benefit for solving system. Each causal-effect event concept is expressed by a verb phrase of an elementary discourse unit which is a simple sentence. The research had three problems: how to determine each adjacent-simple-sentence pair having the cause-effect relation, how to determine each cause-effect-concept pair series mingled with simple sentences having non-cause-effect-relations, and how to identify the complication of several extracted cause-effect-concept pair series from the documents. Therefore, we extract NWordCo-concept set having the causative/effect concepts from the sentences’ verb phrases including a support vector machine to solve each NWordCo size. We apply the Naive Bayes classifier to extract an NWordCo-concept pair set as a knowledge template having the cause-effect relation from the documents. We then propose using the knowledge template to extract several cause-effect-concept pair series. We also apply the intersection of the NWordCo-concept sets to identify the common-cause/effect for representing the complication-development parts of these extracted series. The research results provide a high percent correctness of the cause-effect-concept-pair series determination from the documents.

Keywords: Event-Concept Pair Series, Elementary Discourse Unit, NWordCo, Complication.

Copyright © 2016 Institute of Advanced Engineering and Science. All rights reserved.

1. Introduction

The objective of this paper is to determine each event-concept pair series, particularly a cause-effect-concept pair (called ‘CEPair’) series of disease information downloaded from hospital-web-boards (i.e. http://www.sj.mahidol.ac.th/siddoctor/e-pj/), such as diabetes documents, kidney-disease documents, and artery-disease documents. The CEPair series is used for representing medical complications, particularly the disease complications, including complication development parts which benefit for the solving system. 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 then a group of CEPair elements which are cause-effect-event ordered pairs occurring as a sequence of the CEPair elements on a document. Each CEPair element 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. Moreover, the ‘Complication’ term in medicine is ‘is an event or occurrence that is associated with a disease or a healthcare intervention, is a departure from the desired course of events, and may cause, or be associated with suboptimal outcome’ [1], e.g. a diabetic patient may develop complication in the artery system. Thus, each causative/effect event concept on each CEPair element, CEPairi (i=1,2,...,last which is an integer), is expressed by an EDU pair (where an EDU is an elementary discourse unit which is a simple sentence, [2]) from two adjacent EDUs; one causative-event concept EDU and one effect-event concept EDU as shown in the following CEPairi sequence to represent Example 1.

Example 1: Topic Name: ติ่งความไม่ดี/Diabetic Problems

... EDU1: ติ่งความไม่ดี (A patient gets a diabetes disease.)

EDU2: ติ่งความไม่ดี (Since the pancreas produces less insulin.)

Received September 2, 2016; Revised December 25, 2016; Accepted January 11, 2017
EDU3: "Insulin has a function of signaling cells to take sugar for use."
(Insulin has a function of signaling cells to take sugar for use.)

EDU4: "When the body lacks insulin,"
(When the body lacks insulin.)

EDU5: "[Lack of insulin](EDU4) makes the body unable to take sugar for use."
([Lack of insulin](EDU4) makes the body unable to take sugar for use.)

EDU6: "[Being unable to use the sugar](EDU5) is a cause of blood-sugar level being higher than normal."
([Being unable to use the sugar](EDU5) is a cause of blood-sugar level being higher than normal.)

EDU7: "[Which is a catalyst for artery deterioration occurrence through the body.]"

EDU8: "[The artery deterioration occurrence](EDU7) causes the arteries to constrict."
([The artery deterioration occurrence](EDU7) causes the arteries to constrict.)

EDU9: "[The constriicted arteries](EDU8) is the cause of the ischemic heart disease."
([The constriicted arteries](EDU8) is the cause of the ischemic heart disease.)

EDU10: "Thus, the diabetes disease will be a significant risk factor to a brain disease, a heart disease, and a kidney disease..."
(Thus, the diabetes disease will be a significant risk factor to a brain disease, a heart disease, and a kidney disease...)

where the [...] symbol means ellipsis.

Example 1 is then represented by the CEPair series containing EDU3 as a non-causative and non-effect concept EDU and EDU6 as an intervening EDU of the stimulation relation as shown in the following cause-effect relation expressions.

EDU1-EDU2 Pair as CEPair1: EDU2 (Cause) → EDU1 (Effect)
EDU4-EDU5 Pair as CEPair2: EDU4 (Cause) → EDU5 (Effect)
EDU5-EDU6 Pair as CEPair3: EDU5 (Cause) → EDU6 (Effect)
EDU6-EDU7 as an intervention relation as the stimulation relation:
<highBloodSugars>... StimulationRelation...<artery Deterioration>
EDU7-EDU8 Pair as CEPair4: EDU7 (Cause) → EDU8 (Effect)
EDU8-EDU9 Pair as CEPair5: EDU8 (Cause) → EDU9 (Effect)

where the stimulation relation on EDU6 co-occurs with the cause-effect relation on CEPair3 and CEPair4 as the part of the CEPair series. The CEPair Series representation of Example 1 is then shown in Figure 1.

Figure 1. CEPair Series Representation of Example 1.

Thus, the disease causation direction represented by the CEPair series determined by this research benefits for improvement of people's understanding and compliance to the physician suggestion of the appropriate treatment. Therefore, the research concerns to determine the CEPair series with the event concepts from texts for providing the knowledge representation to people and enhancing the solving system. In addition, this 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.

```
EDU → NP1 | VP | VP
VP → Verb NP2 | Verb adv | Verb
Verb → Verbal Noun | Verbword
NP1 → pronoun | Noun | Noun Adj | Noun Adjphrase
```

NP2 → Noun | Noun Adj | Noun Adphrase
VerbWeak → "be", "not be", "like", "unlike", "have", "not have", "use"
VerbStrong → "be after had", "in focus", "in constant", "in block up", "not determinate", "in ever not respond", "end deterorate", "be excrete", "in increase", "in washout", "in change", "in wash", "in well", "in convince", "be mental be unconscious", "be high", "be wide", "be catalyze", "be of stimulus", ...
Adj → "very high", "very wide", "very catalyze", "very of stimulus", ...

where NP1 and NP2 are noun phrases. VP is a verb phrase. VerbStrong is a strong verb concept set consisting of the causative/effect verb concept set and the stimulating verb concept set, ("be catalyze", "be of stimulus", ...). VerbWeak is a weak verb concept set requiring more information, i.e. VerbWeak Noun, to become either the cause-event/effect-event concept, i.e. "be be + mental clot", or the stimulation-event concept, i.e. "be be + mental catalyze". Noun is a noun concept set. Adv is an adverb concept set. Adj is the adjectival concept set and Adphrase is an adjectival phrase.

There are several techniques [3-9] 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), how to determine each adjacent-EDU pair having the cause-effect relation from the documents containing word ambiguities i.e. a discourse-cue ambiguity and some EDU occurrences of both causative and effect concepts, how to determine the CE pair occurrence mingled with non-cause-effect-relation EDUs including a stimulation relation EDU from the documents, and how to identify the cause-event/effect of the disease complications including their development parts from several CE pair series of different diseases. 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 NWCo) on the EDU’s verb phrase. The reason of using NWCo to represent an EDU event is the VerbWeak element which needs more information from some linguistic sets, i.e. Noun, Adj, Verb and Adv, to form the causative/effect/stimulating concept where the stimulating concept is the concept of the stimulation relation occurring as the enhancement of the certain cause-effect relation. The NWCo expression on an EDU’s verb phrase of the research starts with a word, $w_1$ (where $w_1 \in \text{VerbStrong} \cup \text{VerbWeak}$), followed by the N-1 co-occurred words (N is an integer) as shown in the following equation (1) after stemming words and eliminating stop words.

$$\text{NWCo expression} = w_1 + w_2 + \ldots + w_N$$

where $w_1 \in \text{VerbStrong} \cup \text{VerbWeak}$ ; $w_2, \ldots , w_N \in \text{Noun} \cup \text{Adj} \cup \text{Adv} \cup \text{Verb}$

Thus, we apply each annotated NWCo-expression pair with one NWCo with a causative-event concept and another NWCo with an effect-event concept to represent a cause-effect relation including an annotated NWCo with a stimulating-event concept. We then apply Support Vector Machine (SVM) [10] to learn the NWCo size (which is an N value) for extracting and collecting NWCo expressions with the causative/effect/stimulating event concepts into a NWCo-concept set, NWC. However, some NWCo occurrences lack of information because their verb phrases consist of only one word, i.e. ("mental level" "fat" "fat lip phid blood"

"blood liquid-body substance")/NP1 ("be increase") VP (Fat level in the blood increases)

Thus, we collect the NWC element from the one-word VP by adding two more words from the head noun of NP1 as the following NWCo expression: "mental level" "fat" "fat increase ."

We then apply Naive Bayes (NB) [11] to learn probabilities of NWCo-concept pairs with a relation-class set, (CauseEffectRelation, nonCauseEffectRelation), from the annotated corpus having the discourse cue ambiguity and some NWCo occurrences of both causative and effect concepts depending on their context. The extracted NWC with the causative-event concepts, the effect-event concepts, and the stimulating-event concepts consists of NWCBe, NWCaes, and NWCars which are the NWCo-concept sets extracted from diabetes documents, kidney-disease documents, and artery-disease documents respectively as shown in equation (2). We then determine NWCo (which is as an order pair set of the NWCo-concept pairs having the cause-effect relation) by the Cartesian product of NWC\times NWC along with the NB learning of the relation-class probabilities from the annotated NWCo-concept pairs.

"Fat level in the blood increases"
Therefore, the extracted NWCP_{ce} can be expressed as the knowledge template, particularly the cause-effect-relation template in equation (3) for extracting the CEPair series from the documents.

\[
\text{NWC} = \text{NWC}_{\text{ext}} \cup \text{NWC}_{\text{ed}} \cup \text{NWC}_{\text{end}} \\
\text{NWC}_{\text{ce}} = \left\{ \left( \text{nwc}_{c1}, \text{nwc}_{e1} \right), \left( \text{nwc}_{c1}, \text{nwc}_{e2} \right), \ldots \right\}
\]

where: \( \text{nwc}_{ci}, \text{nwc}_{ej} \) are an ordered pair of an NWordCo-concept pair having the cause-effect relation between \( \text{nwc}_{ci} \) as an NWordCo with a causative-event concept and \( \text{nwc}_{ej} \) as an NWordCo with an effect-event concept; \( i \) and \( j \) are an integer.

And, we assign \( \text{nwc}_{cei} \in \text{NWC}_{\text{ce}} \); therefore \( \text{nwc}_{cei} = (\text{nwc}_{ci}, \text{nwc}_{ej}) ; i=1,2,\ldots, \text{theNumberOfElementOfNWC}_{\text{ce}} \).

We then propose using the cause-effect-relation template, NWCP_{ce}, and the stimulating-cue-word set, \{'\text{be}, \text{be-Verb}\text{+catalyst-Noun}'; '\text{catalyze-Verb}\text{+stimulus-Verb}'\} to determine the CEPair series including a stimulation relation EDU from the testing corpus (see section 3). We also apply the intersection set with the causative/effects concepts of NWCP_{ed}, NWCP_{ce}, and NWCP_{end} to identify the common-cause/effect of disease complications for representing CEPair series containing the complication-development parts from several extracted-CEPair series.

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

2. Related Works

Several strategies [3-9] have been proposed to determine the cause-effect relation from texts without the cause-effect series consideration except [8]. Girju [3] proposed decision tree learning the causal relation from a sentence based on the lexico-syntactic pattern (NP1 causal-verb NP2). Cheng [4] used cue-phrase and the statistical approach to NP-pair probabilities to solve the causal relation occurrence within two EDUs. Verb-pair rules were applied along with machine learning techniques to extract the causality occurrence within several effect EDUs [5]. There are more research works based on the lexico-syntactic pattern with the causal concept as in [6] proposed the Restricted Hidden Naive Bayes model to learn and extract the causality from the English documents. Where the learning features in [6] include contextual, syntactic, position, and connective features. Mirza [7] 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. Whilst causal chains were generated by adding the causal chains obtained from word matching [8]. The model's [8] is based on noun features including hidden causal chains solved by latent topics. Events of automatic pathway curation using the popular mTOR pathway (mTOR is a kinase that in humans is encoded by the MTOR gene) [9] were extracted by using different training datasets and learning algorithms. Their event extraction based on the noun derivative extracts the entities (genes, proteins etc), reactions (e.g. phosphorylation) and their arguments (theme, cause, and product). Whereas event pairs of our research are based on verb phrases. Nevertheless, most of the previous works on the cause-effect relation are based on noun/NP features (except [5]) existing on one/two sentences without the series consideration (except [8]) whereas our work has NP1 ellipses occurrences on documents. Even though [5]'s work is based on verb phrases, their work emphasizes on a cause/effect boundary without the event-pair-series consideration. Whilst [8]'s work as the causal chain emphasizes on NP1 and the latent topics. However, there are few works on extracting the CEPair series as a disease causation direction including the complication development.

3. Problems of Extracting CEPair series from Texts

3.1 How to Determine EDU pair Having Cause-Effect Relation Including Word Ambiguities

The CEPair, expression as the cause-effect relation between two adjacency EDUs as an EDU pair can be determined by using the discourse-cue set, \{'\text{arrived because}', \text{\&\&}, \text{\&\&}, \text{\&\&}'}
cause',...). However, some discourse-cue set elements are ambiguity. For example: CEPair1 of Example 1 has a discourse cue, 'เนื่องจาก/since', on EDU2 whereas an EDU1-EDU2 pair of the following Example 2 having 'เนื่องจาก/since' on EDU2 is not the CEPair1 expression.

Example 2 Topic Name: โรคหัวใจบดทุพologically related to Heart Disease from Diabetes

... EDU1: "ผู้ป่วยที่มีหลอดเลือดผิดބจะมีโรคหัวใจ" (A diabetic patient might get the heart disease.)

"ผู้ป่วย / patient ที่มี/have หลอดเลือดผิด / disease จะ / will ไปยัง/ have โรคหัวใจ / heart disease "

EDU2: "เนื่องจาก การที่มี/since blood level มาก / high" (Since a blood sugar level is high.)

"เนื่องจาก / since การที่มี/blood level มาก / high "

EDU3: "การที่มี/since การที่มี/blood level มาก / high เกิดขึ้น / happened เนื่อง / because กล่าวไปมาจาก / from some increased chemical substance types in blood" (The high blood sugar level / EDU2] causes of having some increased chemical substance types in blood.)...

"การที่มี/since การที่มี/blood level / EDU2 ไปยัง/ have กล่าวไปมาจาก / from some increased chemical substance type เกิดขึ้น / happened ใน / in เบื้องต้น / highblood "...

Example 2 contains the following CEPair occurrence.

EDU2-EDU3 Pair as CEPair1: EDU2 (cause) → EDU3 (effect)

Moreover, there are some EDU occurrences with both causative-concepts and effect-concepts, i.e. EDU5 and EDU8 of Example 1 on CEPair1 to CEPair2 and CEPair2 to CEPair3 respectively. It is difficult to identify the certain EDU occurrence as the causative concept or the effect concept. With regard to the above word ambiguity problem, we solved these examples of the word ambiguity problem by applying the NB machine learning technique to learn the annotated NWordCo-concept pairs with the cause-effect/non-cause-effect relation from each EDU pair on the learning corpus after stemming words and eliminating stop words. And also, the NWordCo size has to be solved by SVM learning on the consecutive words on equation (1) of each verb phrase with a slide window size of two adjacent words with a one word sliding distance on each EDU's verb phrase. The NWordCo extraction is then occurred after the NWordCo sizes have been solved. The extracted NWordCo expressions along with concepts according to the word sequence from the testing corpus are collected into NWC. NWC is then applied by the Cartesian product of NWC×NWC. The result of the Cartesian product is an NWordCo-concept ordered pair set containing some ordered pairs with the cause-effect relation. Therefore, we collect each element of NWC, nwcpc (see section 1) by using the relation-class learning results by NB from the annotated NWordCo-concept pairs to the result of the Cartesian product.

3.2 How to Determine CEPair Series Mingled with Non-Cause-Effect-Relation EDUs

Regarding Example 1, the CEPair series extraction including the cause-effect relation occurrences and the stimulation relation occurrences on the series mingled with non-relation EDU as EDU3 of this example is challenge. Therefore we propose using NWCP as the cause-effect-relation template to determine each CEPair series through the string matching by using the max_similarity scores (MaxSimilarityScore) [12] between each ordered pair of NWCP and each NWordCo-concept pair from the testing corpus and also using the stimulating-cue-word set to determine the stimulation relation occurrence on the determined CEPair series.

3.3 How to Identify Complication Development Parts for Representation

The disease complications do not occur on all extracted CEPair series from the disease documents. If two or more disease types have the related complication development, each disease will have at least one CEPair, having the same common-cause/effect. Therefore, we apply the intersection set, IntNWC, as in equation (4) with the causative/effect concepts including the element ranking of IntNWC to identify the common cause/effect of complications for representing the complication-development parts of the extracted CEPair series as shown in Figure 2 of Example 3 and Figure 3.

\[
\text{IntNWC} = \text{NWC}_{\text{ed}} \cap \text{NWC}_{\text{cd}} \cap \text{NWC}_{\text{and}}
\]  

Example 3. Topic Name: โรคไตยที่ขึ้นกับโรคตุ่นโรค Diabetic Nephropathy

EDU1: "ผู้ป่วยที่มี/have หลอดเลือดผิด / disease เกิดขึ้น / happened โรคตุ่น / Diabetic" (A patient gets type2 diabetes.)

"ผู้ป่วย / patient ที่มี / have หลอดเลือดผิด / disease เกิดขึ้น / happened โรคตุ่น / Diabetic"

EDU2: "เนื่องจาก การที่มี/because the body does not respond to the hormone" (because the body does not respond to the hormone)

"เนื่องจาก / because การที่มี/because the body ไม่ก่อ / does not respond to the hormone"

EDU3: "จะมีการส่งผลของโรคตุ่น / will have complications of the disease เกิดขึ้น / happened การที่มี/because the body ที่มี high blood sugar will have high blood เกิดขึ้น / happened high" (he will have too high blood sugar level.)

"จะมีการส่งผลของโรคตุ่น / will have complications of the disease เกิดขึ้น / happened การที่มี/because the body ที่มี high blood เกิดขึ้น / happened high"
EDU4: "(The high blood sugar level) causes the kidneys to have extra work in absorbing food nutrients and filtering waste."

"(which is the cause of deterioration in the kidney function afterwards.)"

"(which is the cause of deterioration in the kidney function afterwards)"

EDU5: "(which is the cause of deterioration in the kidney function afterwards)"

"(which is the cause of deterioration in the kidney function afterwards)"

"(which is the cause of deterioration in the kidney function afterwards)"

"(which is the cause of deterioration in the kidney function afterwards)"

"(which is the cause of deterioration in the kidney function afterwards)"

Figure 2. CEpair Series Representation of Example 3

Figure 3. Show two complication development parts inside the rectangular dash line having "***" as the common-effect of both complications of Figure 1 and Figure 2.

4. Framework of Event-Concept Pair Series Extraction to Present Disease Complications

There are seven steps in our framework, Corpus Preparation, NWordCo Size Learning, Collection of NWordCo, Extraction of NWordCo with Event Concepts, NWordCo-Concept Pair Learning, Extraction of NWordCo, Extraction of CEpair Series, and Representation of Complication Development Parts as shown in Figure 4.

IJEESCS Vol. 5, No. 3, March 2017: xxx – xxx
4.1 Corpus Preparation


After the word segmentation is achieved, EDU Segmentation [15] is then operated to provide a 3000 EDUs’ corpus (consists of 1000 EDUs from each disease: the diabetes, kidney disease, and artery disease). The corpus included stemming words and the stop word removal is separated into 3 parts: the first part of 1200 EDUs consists of 400EDUs from each disease 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 1200 EDUs having 400EDUs from each disease is the testing corpus used for the NWordCo size determination to extract and collect NWordCo occurrences with causative/effect/stimulating concepts into the NWordCo-concept set, NWC. NWC consists of three disease-NWordCo-concept sets as NWC_dir, NWC_ad, and NWC_ad. NWC are also used for collecting NWCPa. The third part of 600 EDUs consisting of 200 EDUs from each disease is used for CEpairs series extraction. This step also includes semi-automatic annotation of each NWordCo size along with the causative/effect/stimulating concept as shown in Figure5. The step annotates the EDU pairs through a CEpairID property of each EDU tag as the CEpair elements of their CEpair series annotated by a CEpairSeries tag. All word concepts of each NWordCo expression is referred to Wordnet (http://word-net.princeton.edu/obtain) and MeSH after translating from Thai to English by Lextron (http://lextron.nedec.or.th).

4.2 NWordCo Size Learning

With regard to NWordCo expression on equation (1) after stemming words and the stop word removal, the features used to learn the NWordCo size from the learning corpus by SVM are obtained from the annotated corpus containing the following concept sets: Verb_strong, Noun, Adj, Adv; where each element of these concept sets should occur in more than 50% of the number of documents.
SVM [10,11] 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

$$f(x) = \langle w \cdot x \rangle + b$$

$$= \sum_{j=1}^{n} w_t x_j + b$$

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

If $i = 1 \land (w_i \in \text{VerbMng} \lor \text{V_w Cape})$ then $w_i$ is the first word of VP with the CausativeOrEffectOrStimulating concept.

The N-Word-Co size/boundary learning from $w_jw_{j+1}$ of VP based on using Weka (http://www.cs.waikato.ac.nz/ml/weka/) is then the SVM supervised learning by sliding the window size of two consecutive words with one sliding distance after stemming words and the stop word removal. Where $j=1,2,\ldots,n$ and $n$ is End-of-Boundary and is equivalent to the N value of NWordCo size.

### 4.3 Collection of NWordCo with Event Concepts

Assume that each EDU is represented by (NFP, VP).

L is a list of EDUs after stemming words and the stop word removal.

Verb=Verbnum \lor VerbAdj, Nfp= NfpNum, Adv=AdvADJ, AdvP=AdvP

NFP is a noun phrase; VP is a verb phrase; exp1 is an EDU’s NFP; exp2 is an EDU’s VP; NWC is an NWordCo-concept set

**NWordCo-EXTRACTION**

1. **NWC<0; NWC<0; i=1; j=1; k=0; b=0**;
2. While k < Length(L) do
3. If i = 1 then
4. $\{
5. \text{Identify the first word of NWordCo} \}
6. $\{
7. \text{Determine VP consisting of only one word} \}
8. $\{
9. \text{If exp1, exp2, VerbMng \land numberOfWords exp1 != 1 then} \}
10. $\{
11. \text{Next NWC} \}
12. $\{
13. \text{If class -> 'nonCoreOnCore' continue then if f != 'no' \}
14. $\{
15. \text{Else f = 'yes';} \}
16. $\{
17. \text{If class -> 'year' then NWC = NWC \lor w_j;} \}
18. $\{
19. \text{If j = i + 1 then NWC = NWC \lor w_j;} \}
20. $\{
21. \text{Return NWC} \}

**Figure 6. NWordCo Extraction Algorithm**

<table>
<thead>
<tr>
<th>NWordCo Expression</th>
<th>WordSequenceConcept</th>
<th>Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>occur-sugar-blood-blood-high</td>
<td>&lt;occur-sugar-blood-blood-high&gt;</td>
<td>(haveHighBloodSugar)</td>
</tr>
<tr>
<td>occur-sugar-blood-bleed-high</td>
<td>&lt;occur-sugar-blood-bleed-high&gt;</td>
<td>(haveHighBloodSugar)</td>
</tr>
<tr>
<td>lackOf-hormone</td>
<td>&lt;lackOf-hormone&gt;</td>
<td>(lackOfHormone)</td>
</tr>
<tr>
<td>kidney-complication</td>
<td>&lt;kidney-complication&gt;</td>
<td>(kidneyComplication)</td>
</tr>
<tr>
<td>cause-to-blood-Protein</td>
<td>&lt;cause-protein-blood-low&gt;</td>
<td>(causeProteinBloodLow)</td>
</tr>
<tr>
<td>collect-fat-artery</td>
<td>&lt;collect-fat-artery&gt;</td>
<td>(collectFatArtery)</td>
</tr>
<tr>
<td>detect-protein</td>
<td>&lt;detect-protein&gt;</td>
<td>(detectProtein)</td>
</tr>
<tr>
<td>detect-urine</td>
<td>&lt;detect-urine&gt;</td>
<td>(detectUrine)</td>
</tr>
<tr>
<td>loss-protein-urine</td>
<td>&lt;loss-protein-urine&gt;</td>
<td>(lossProteinUrine)</td>
</tr>
</tbody>
</table>

The results of learning the NWordCo size by SVM from the previous step is the weight vector of all $w_j$ and $w_{j+1}$. This weight vector is used to solve each NWordCo size with a CausativeOrEffectOrStimulating concept for extracting the solved-size NWordCo from the testing corpus into the NWordCo-concept set, NWC, by equation (5) as shown in Figure 6.

Moreover, some EDUs' verb phrases consist of only one word of a verb, i.e. 'rise', 'decrease' 'increase' 'be/behind' 'must/reduce', which results in the NWordCo size or N=1 with lacking of some
information to represent those EDUs. Thus, we add two more words from the head noun of NP1 to the NWordCo expression determined from the EDU’s verb phrase consisting of only one word of a verb. In regard to Figure 6, the extracted NWordCo expressions existing in 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 NWordCo expressions with the causative, effect, and/or stimulating concepts. Table 1 also includes the annotated concepts from the corpus preparation.

<table>
<thead>
<tr>
<th>NWordCo-Concept Pair</th>
<th>CauseEffect</th>
<th>Non-CauseEffect</th>
</tr>
</thead>
<tbody>
<tr>
<td>(CausativeNWordCoConcept)(EffectNWordCoConcept)</td>
<td>CauseEffect Rel. Probability</td>
<td>Non-CauseEffect Rel. Probability</td>
</tr>
<tr>
<td>(lackOfnormcore)(haveHighBloodSugar)</td>
<td>0.0171</td>
<td>0.0116</td>
</tr>
<tr>
<td>(deteriorateArtery)(constrictArtery)</td>
<td>0.0053</td>
<td>0.0029</td>
</tr>
<tr>
<td>(collectFatInArtery)(causeAtherosclerosis)</td>
<td>0.0053</td>
<td>0.0029</td>
</tr>
<tr>
<td>(lossProteinToUrine)(haveLowBloodProtein)</td>
<td>0.0132</td>
<td>0.0116</td>
</tr>
<tr>
<td>(haveLowBloodProtein)(getSwellSymptom)</td>
<td>0.0020</td>
<td>0.0025</td>
</tr>
<tr>
<td>(haveHighBloodSugar)(getKidneyFailure)</td>
<td>0.0038</td>
<td>0.0048</td>
</tr>
<tr>
<td>(haveHighBloodSugar)(deteriorateArtery)</td>
<td>0.0038</td>
<td>0.0048</td>
</tr>
</tbody>
</table>

### 4.4 NWordCo-Concept Pair Learning

This step is the NB learning [11] of the NWordCo-concept pair occurrence feature with the CauseEffectRelation class on several two adjacent EDUs as EDU pairs with CEpairID annotations of the annotated corpus from the corpus preparation step in section 4.1 as the learning corpus after stemming words and eliminating stop. The learning results of this step by using Weka(www.cs.waikato.ac.nz/ml/weka/) are the probabilities of the annotated NWordCo-concept pairs with the CauseEffectRelation and Non-CauseEffectRelation classes as shown in Table2.

### 4.5 Extraction of NWCPce

The collected NWC set from the previous step of in section 4.3 is used by the Cartesian product of NWC×NWC to become an NWordCo order pair set, NW CordP. Where NW CordP = NW CordP ; h=1,2, ... num ; num is the number of elements of NW CordP. We then extract and collect only NW OrdP with the cause-effect relation into the NWCPce set by equation (6) with the probabilities of NWordCo-concept pair occurrences from Table 2 resulted by the previous NB learning in section 4.4.

\[
\text{nwcOrdPRel} = \arg \max \limits_{\text{class}} P(\text{class}|\text{nwcOrdP}) \cdot P(\text{class}),
\]

\[
= \arg \max \limits_{\text{class}} P(\text{nwcOrdP} | \text{class}) \cdot P(\text{class}),
\]

where NWOrdPRel is the relation of nwcOrdP ;

\[
\text{nwcOrdP} \in \text{NW CordP} \text{ which is an NWordCo order pair set;}
\]

Class = "CauseEffect Relation", "NonCauseEffect Relation"

h = 1,2, ... num ; num is the number of elements of the NW CordP set.

### 4.6 Extraction of CEpair Series

The objective of this step is to extract the CEpair series by using the similarity scores/MaxSimilarityScore [12] on the following equation (7) to determine the string matching between nwcep and nwcece. Where nwcep is an NWordCo-concept pair gained by sliding a window size of two consecutive EDUs/NWordCos as an NWordCo pair (nwce1 and nwce2) with one EDU/NWordCo distance from the 600EDUs testing corpus. And, nwcece \in NWCPce; nwce has a causative/effect concept whilst nwce has an effect/causative concept respectively.

\[
\text{MaxSimilarityScore} = \arg \text{MaxSimilarityScore}(\text{nwcep}) \cdot \frac{|\text{nwce1} \cap \text{nwce2}|}{\sqrt{|\text{nwce1}| \times |\text{nwce2}|}}
\]

where numCEpair is the number of NWCPce elements ;

nwce1 and nwce2 are the NWordCo concepts as a causative/effect concept and an effect/causative concept respectively from the testing corpus

\[
\text{nwcOrdP} \text{ is an NWordCo concept pair, nwce1 and nwce2}; \beta = 1,2;
\]

nwcep = nwce1 + nwce2 if nwce1 is a cause ;

nwcep = nwce1 + nwce2 if nwce2 is a cause;

NWCPce is an ordered pair set of NWordCo concept pairs having the cause/effect relation.

Title of manuscript is short and clear, implies research results (First Author)
If MaxSimilarityScore between either \( \text{mnwcp}.\text{nwcp.c}+\text{tnwcp}.\text{tnwcp.c} \) or \( \text{mnwcp}.\text{nwcp.c}+\text{tnwcp}.\text{tnwcp.e} \) and \( \text{mnwcp}_\text{ed.}+\text{tnwcp}_\text{ed.} \) as shown in Figure 7 is greater than or equal to 90%, then both \( \text{tnwcp} \) and \( \text{mnwcp}_\text{ed.} \) are equivalent which results in \( \text{mnwcp}_\text{ed.} \) appended to a series as Series, \( \text{Series} \cup \text{mnwcp}_\text{ed.} \), 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 stimulation-cue-word set.

![Figure 7. CEpair Series Extraction Algorithm](image)

4.7 Determination of Complication Development Parts for Representation

With regard to section 4.3, we collect three different NWordCo-concept sets: \( \text{NWC}_{\text{abd.}} \), \( \text{NWC}_{\text{kid.}} \), and \( \text{NWC}_{\text{en.d.}} \) from diabetes, kidney-disease, and artery-disease documents respectively. The interaction set, IntNWC, with the causative/effect concepts of \( \text{NWC}_{\text{abd.}} \), \( \text{NWC}_{\text{kid.}} \), and \( \text{NWC}_{\text{en.d.}} \) consists of the following NWordCo elements: \( \text{beHighbloodSugar} \), \( \text{beHighbloodFat} \), \( \text{inflameOrgan} \), \( \text{deteriorateArtery} \), \( \text{constrictArtery} \), \( \text{highHighBloodPressure} \), \( \text{getDisease} \), \( \text{beComplication} \), \( \text{beNonfunctional} \), and \( \text{malfunction} \). However, some elements of IntNWC occasionally occur on the documents. Therefore, it is necessary to count and rank the top 5 \( \text{intnwc} \) (where \( \text{intnwc} \in \text{IntNWC} \)) by the number of \( \text{intnwc} \) occurrences as shown in Table 3 to determine the most common-cue/effect (whose rank is equal to 1) of disease complications. The top 5 \( \text{intnwc} \) from Table3 are used for determining the complication development parts of several extracted CEpair series as shown in Figure 8 which shows only two extracted CEpair series in an ArrayList[2] object by the CEpair Series Extraction algorithm in Figure 7. The result of determining CEpair Series with complication development parts by the algorithm in Figure 8 is kept in ListSeries [], which is the Array of ArrayList data structure. Therefore, ListSeries [] is used to represent the CEpair series with complication development parts as in Figure 3.
<table>
<thead>
<tr>
<th>Disease has -150EDUs</th>
<th>Number Of NWordCo-Concept Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>betHighblood-</td>
</tr>
<tr>
<td></td>
<td>Sugar level</td>
</tr>
<tr>
<td>Kidney Disease</td>
<td>4</td>
</tr>
<tr>
<td>Diabetes</td>
<td>30</td>
</tr>
<tr>
<td>Artery Disease</td>
<td>6</td>
</tr>
<tr>
<td>total</td>
<td>40</td>
</tr>
<tr>
<td>Rank</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3. Show top 5 in NWC by number of occurrences

Assume that List adaptation is an array of ArrayList for representation of some CEPair series with the complication development parts for some disease types.

List adaptation has two diseases of ArrayList where each ArrayList element contains a CEPair series of one disease type document.

CEPAIR SERIES WITH COMPLICATION DEVELOPMENT

```java
Arraylist<String> listSeries[] = new ArrayList[2];
listSeries[0] = Series1, /* m=1 from CEPair Series Extraction Algorithm for the 1st disease type.
listSeries[1] = Series2, /* n=2 from CEPair Series Extraction Algorithm for the 2nd disease type.

i := 0; j := 0; k := 0; match := 0;
while (i < 5 && k != 0)
do
 {While j < size0 && match := 0
  if listSeries[0].get(j).getRelname() == "CEPair" &&
    listSeries[1].get(k).getRelname() == "CEPair" then
    if (tempj = listSeries[0].get(j).getConcept();
        tempk = listSeries[1].get(k).getConcept();
        tempj = listSeries[0].get(j).getEffect();
        tempk = listSeries[1].get(k).getEffect();
        IF listSeries[0].get(j).getConcept() == ConceptRank[1] then
          listSeries[0].set(j, "**" + tempj, tempj);
        IF listSeries[0].get(j).getEffect() == ConceptRank[1] then
          listSeries[0].set(j, "**" + tempk, tempk);
        IF listSeries[1].get(k).getConcept() == ConceptRank[1] then
          listSeries[1].set(k, "**" + tempj, tempj);
        IF listSeries[1].get(k).getEffect() == ConceptRank[1] then
          listSeries[1].set(k, "**" + tempk, tempk);
        match := 1;
        k := k + 1;
      k := k + 1;
      j := j + 1;}

Return listSeries[]
```

Figure 8. Algorithm of Determining CEPair Series with Complication Development

5. Evaluation and Conclusion

There are four evaluations of the proposed research being evaluated by three expert judgments with max win voting: the first evaluation is the extraction of NWC with the NWordCo size/boundary consideration from 1200 EDU documents consisting of the diabetes, kidney, and artery diseases as a testing corpus which is also used for the second evaluation. The extraction of NWC家人 is evaluated as the second evaluation. The third and the fourth evaluations are the CEPair series extraction and the common-cause/effect identification from the other testing corpus of 600 EDUs consisting of the diabetes, kidney, and artery diseases. The first and the second evaluations are based on the precisions and the recalls within ten fold cross validation whilst the third and the fourth evaluations are the percentages of correctness. The precision of the NWC extraction based on the size/boundary determination is 0.876 with the recall of 0.801 whilst the precision of the NWC家人 extraction is 0.882 with the 0.757 recall. And the correctness of the CEPair series extraction and the common-cause/effect identification are 89.5% and 90% respectively. The reasons of low recalls in extracting NWC, and in determining NWC家人 are: 1) some causative event occurrences are based on an event expression by a prepositional phrase while their effect events are expressed by their verbs, i.e. "(swelling and arteries) NP1 (from degenerate) NP2 (having high blood lipids) Verb (NP3) (having high blood lipids) NP4 (NP5) (from feel) Verb" (Swelling often begins at the foot). Moreover, some problems that affect to the % correctness of the CEPair series extraction and also the common-cause/effect identification are:

1) The EDU sequence among a causative-event concept EDU, an effect-event concept EDU, and a non-causative-effect-event concept EDU as follow:

Title of manuscript is short and clear, implies research results (First Author)
EDU1-as Effect: ""ได้เป็นโรคเบาหวาน"" (A patient gets a diabetes disease.)
EDU2-as Cause: ""ศีรษะปวดมาก นอนไม่ที่นั่ง"" (Since the body lacks of insulin.)
EDU3-as CauseAndEffect: ""รู้สึกหิวถึงกับจะหาย"" (Insulin has a function of signaling cells to take sugar for use.)
EDU4-Effect: ""ปอดขาดสัมผัสน้ำมันที่ผิวหนัง"" (lacking of insulin/EDU2) makes the body unable to take the sugar for use.

where the following CEpair1 can be determined except CEpair2
CEpair1; EDU1 (cause) → EDU1 (effect)
CEpair2; EDU2 (cause) → EDU4 (effect)

2) the boundary of causative/effect event concept EDUs, for example:

Topic: ""Why are there the kidney complication?"
EDU1-VPasCause: ""การเพิ่ม/originalไม่ได้หมายให้ฉันมีอาการที่แสดงให้เห็นในสิ่งที่เกิดขึ้น"" (the kidney disease complication of diabetic disease is the result of the blood sugar level being higher than normal.)
EDU2-VPasEffect: ""การเพิ่ม/originalไม่ได้หมายให้ฉันมีอาการที่แสดงให้เห็นในสิ่งที่เกิดขึ้น"" (the blood sugar level/EDU1 causes to have changing of blood circulation in the kidneys.)
EDU3-VPasEffect: ""ทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เราทำให้เร

where EDU1 is a causative-event concept EDU having EDU2 and EDU3 as the effect-event concept EDU boundary.
CEpair1; EDU1 (cause) → EDU2 (effect) ∧ EDU3 (effect)

3) the complex sentence, e.g.

Complex Sentence: ""นักเรียนนั้นไม่ได้มีค่าที่สูง"" (This sugar level which is high causes problems as follows.) where 'This sugar level which is high' is equivalent to 'This high sugar level'.

Hence, the research contributes the methodology to determine the CEpair series with the complication development parts for clearly communicating health information and improving health literacy, particularly the disease causation pathway, to people on the social network. Finally, our research can also enhance the diagnosis and solving system of the other areas i.e. the business services industry analysis.

Reference