Indonesian Journal of Electrical Engineering and Computer Science
Vol. 18, No. 2, May 2020, pp. 970~978
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v18.i2.pp970-978

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

Chaveevan Pechsiri[1], Titirut Mekbunditkul[2]
[1]College of Innovative of Technology and Engineering, Dhurakij Pundit University, Thailand
[2]Research Service Center, Dhurakij Pundit University, Thailand

Article Info

Article history:

Received Sep 9, 2019
Revised Nov 10, 2019
Accepted Nov 24, 2019

Keywords:

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

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.

Corresponding Author:

Chaveevan Pechsiri,
College of Innovative of Technology and Engineering,
Dhurakij Pundit University, Thailand.
Email: chaveevan.pec@dpu.ac.th

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 particulary diabetic, cardiovascular, kidney dieases [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, $CEpair_i$, is expressed by an elementary discourse unit (EDU which is a simple sentence, [4]) as follows:

$$CEpair_1, CEpair_2, .., CEpair_{last}$$

where $CEpair_i$ (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: "*ผู้ป่วยเป็นโรคเบาหวาน*" (***A patient gets a diabetes disease***.)
"*ผู้ป่วย*/*patient* *เป็น*/*is* *โรคเบาหวาน*/*diabetes disease*."
EDU2:"*เนื่องจาก ร่างกาย ไม่สามารถนำ น้ำตาล ในร่างกายไปใช้ได้อย่างเต็มที่*" (***since the body cannot fully use sugar in the body***.)
"*เนื่องจาก*/*since* *ร่างกาย*/*body* *ไม่สามารถนำ*/*cannot take* *น้ำตาล*/*sugar* *ใน*/*inside* *ร่างกาย*/ *body* *ไปใช้*/*to use* *ได้อย่างเต็มที่*/*fully*"
EDU3:"*เพราะ*[*ร่างกาย*] *ขาดฮอร์โมนอินซูลิน*" (***Because* [*the body*] *lacks of hormone insulin*.**)
"*เพราะ*/*because* [*ร่างกาย*/*body*] *ขาด*/*lack of* *ฮอร์โมนอินซูลิน*/*hormone insulin*"
EDU4:"*บางครั้งร่างกายไม่ขาดฮอร์โมน*" (***The body sometimes does not lack of the hormone*.**)
"*บางครั้ง*/*sometimes* *ร่างกาย*/*body* *ไม่*/*not* *ขาด*/*lack_of* *ฮอร์โมน*/*hormone* "
EDU5:"*แต่*[*ร่างกาย*] *ไม่ตอบสนองต่อฮอร์โมน*" (***But* [*the body*] *does not respond to the hormone*.**)
"*แต่*/*but*[*ร่างกาย*//*body*] *ไม่*/*not* *ตอบสนองต่อ*/*respond_to* *ฮอร์โมน*/*hormone* "
EDU6:"[*ไม่ตอบสนองต่อฮอร์โมน*/EDU5] *ทำให้ระดับน้ำตาลในเลือดสูงกว่าปกติ*"
(**[*No responding to hormone*/EDU5] *causes Blood-sugar level to be higher than normal*.**)
"[*ไม่ตอบสนองต่อฮอร์โมน*//*no responding to the hprmone*/EDU5] *ทำให้*/*cause* *ระดับน้ำตาล*/*sugar-level* *ใน*/*in* *เลือด*/*blood* *สูงกว่า*/*higher than* *ปกติ*/*normal* "
EDU7:"[*ระดับน้ำตาลในเลือดสูง*] *เป็นตัวเร่งให้เกิดการเสื่อมของหลอดเลือดแดงทั่วร่างกาย*"
(**[*The high blood-sugar*/EDU6] *is a catalyst for artery deterioration occurrence through the body*.**)
"[*ระดับน้ำตาลในเลือดสูง*/*high blood-sugar level*/EDU6] *เป็น*/*is* *ตัวเร่ง*/*catalyst* *ให้เกิด*/*to_occur* *การเสื่อม*/*deterioration* *ของ*/*of* *หลอดเลือดแดง*/*artery* *ทั่ว*/*through* *ร่างกาย*/*body* "
EDU8:"[*การเสื่อมของหลอดเลือดแดง*] *ทำให้หลอดเลือดแดงตีบ*" (**[*The artery deterioration occurrence*/EDU7] *causes the arteries to constrict*.**)
"[*การเสื่อมของหลอดเลือดแดง*/*arteries deterioration*/EDU7] *ทำให้*/*cause* *หลอดเลือดแดง*/*artery* *ตีบ*/*constrict*"
EDU9: "[*หลอดเลือดแดงตีบ*/EDU8] *ทำให้เกิดโรคหัวใจขาดเลือด*" (**[*The constricted arteries* /EDU8] *causes of the ischemic heart disease*.**)
"[*หลอดเลือดแดงตีบ*/*constricted arteries*/EDU8] *ทำให้เกิด*/*cause of* *โรคหัวใจ*/*heart disease* *ขาด*/*lack of* *เลือด*/*blood* "
EDU10: "*ดังนั้น โรคเบาหวาน จึงเป็นปัจจัยเสี่ยงทำสำคัญต่อโรคทางสมอง โรคหัวใจ และโรคไต เป็นต้น*"…
(***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 EDU7 as an intervening EDU of the stimulation relation as shown in the following.
EDU1-EDU2 Pair as $CEpair_1$: EDU2 (Cause) → EDU1 (Effect)
EDU2-EDU3 Pair as $CEpair_2$: EDU3 (Cause) → EDU2 (Effect)
EDU5-EDU6 Pair as $CEpair_3$: EDU5 (Cause) → EDU6 (Effect)
EDU7 as an intervening EDU having the stimulation relation:
<highBloodSugar>… *beStimulationRelation*...<artery Deterioration>
EDU7-EDU8 Pair as $CEpair_4$: EDU7 (Cause) → EDU8 (Effect)
EDU8-EDU9 Pair as $CEpair_5$: EDU8 (Cause) → EDU9 (Effect)
where EDU4 is a non-cause/non-effect concept EDU and the stimulation relation on EDU7 co-occurs with the cause-effect relation on $CEpair_4$ as the part of the CEpair series which consists of two sub-series, $CEpair_1$-$CEpair_2$ and $CEpair_3$-$CEpair_5$.

Thus, the disease causation pathway represented by the extracted 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 concerns to extract the CEpair series with the event concepts from texts for providing the knowledge 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 → $Verb_{weak}$ Noun | $Verb_{strong}$
NP1 → pronoun | Noun | Noun Adj | Noun AdjPhrase
NP2→ Noun | Noun Adj | Noun AdjPhrase

$Verb_{weak}$→{'*เป็น*/*be*','*ไม่เป็น*/*not_be*','*มี*/*have*','*ไม่มี*/*not_have*','*ใช้*/*use*'}; $Verb_{strong}$→{'*ทำให้*/*cause*','*เกิด*/*occur*','*ตัน*/*block-up*','*ตีบ*/*constrict*',
'*เสื่อม*/*deteriorate*','*ไม่ตอบสนอง*/*not_respond*','*ขับ*/*excrete*','*เปลี่ยนแปลง*/*change*','*บวม*/*swell*','*อาเจียน*/*vomit*','*ชัก*/*convulse*','*หมดสติ*/*be_ unconscious*','*เพิ่มขึ้น*/*increase*','*สูง*/*high*','*ตาย*/*die*','*กระตุ้น*/*stimulus*', '*เร่ง*/ *catalyze*', ..}; Adv→{'*ยาก*/ *difficultly*', '*เหลว*/ *liquidly*', …};
Noun→{' ','*ผู้ป่วย*/*patient*','*อาการ*/*symptom*','*ตระคริว*/*contraction*','*อวัยวะ*/*human organ*','*แผล*/*scar*', '*เลือด*/*blood*', '*น้ำตาล*/*sugar*', '*ไขมัน*/*fat*','*โปรตีน*/*protein*' , '*ปัสสาวะ*/*urine*','*ความดัน*/*pressure*', '*ตัวเร่ง*/*catalyst*',…}; Adj → {'*สูง*/*high*', …}…

Where NP1 and NP2,are noun phrases. VP is a verb phrase. Noun is a noun concept set. $Verb_{strong}$ is a strong verb concept set consisting of the causative/effect verb concept set and the stimulating verb concept set,{'*เร่ง*/*catalyze*','*กระตุ้น*/ *stimulus*', ..}. $Verb_{weak}$ is a weak verb concept set requiring more information as $Verb_{weak}$+Noun to become either the cause-event/ effect-event concept, i.e. '*เป็น*/*be+สิ่งเลือด*/*clot*', or the stimulating-event concept, i.e. '*เป็น*/*be+ตัวเร่ง*/*catalyst*'. Adj is an adjective concept set. AdjPhrase is an adjective phrase component. Adv is an adverb concept set.

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 causative-concept/effect-concept EDUs from the documents containing some EDU occurrences with both causative-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 $Verb_{weak}$ 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 v_1 (where $v_1 \in Verb_{strong} \cup Verb_{weak}$) 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.

$$\text{'NWordCo' expression} = v_1 + w_2 + \ldots + w_N \quad (1)$$

where $v_1 \in Verb_{strong} \cup Verb_{weak}$; $w_2,\ldots,w_N \in Noun \cup Adj \cup Adv \cup Verb$

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 $NWCP_{ce}$ set (which is an ordered pair set of NWordCo-concept pairs with the CauseEffectRelation class) as shown in the following.

$NWCP_{ce} = \{NWordCo_cNWordCo_e\text{-}pair_1, NWordCo_cNWordCo_e\text{-}pair_2, \ldots, NWordCo_cNWordCo_e\text{-}pair_{last}\}$

where $NWordCo_cNWordCo_e$-$pair_i$ is an NWordCo-concept pair having the cause-effect relation between $NWordCo_c$ and $NWordCo_e$ (in which $NWordCo_c$ is an NWordCo with a causative concept and $NWordCo_e$ is an NWordCo with an effect concept); i=1,2,…*last* which is an integer.

We then propose using $NWCP_{ce}$ and the stimulating-cue-word set, {'*เป็นตัวเร่ง*/*be*-$Verb_{weak}$ *catalys*-Noun', '*เร่ง*/*catalyze*-V_{strong}','*กระตุ้น*/*stimulu*-V_{strong}'…} 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 benifited 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

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 $CEpair_1$ to $CEpair_2$ and $CEpair_4$ to $CEpair_5$ 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 $Verb_{strong} \cup Verb_{weak}$, 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 a 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 $CEpair_i$ as Cause-Effect Relation with Discourse-Cue Ambiguity

The $CEpair_i$ 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*','***เนื่องจาก***/*since*','***ทำให้***/*cause*',..}, see Example 1. However, some discourse-cue set elements are ambiguity. For example: $CEpair_1$ 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 $CEpair_1$ expression.

Example 2

… EDU1: "***ผู้ป่วยเบาหวานอาจเป็นโรคหัวใจ***" (***A diabetic patient might get the heart disease***.)
"***ผู้ป่วย***/*patient* ***เบาหวาน***/*diabetes* ***อาจเป็น***/*might get* ***โรคหัวใจ***/*heart disease* "
EDU2:"***เนื่องจาก ภาวะน้ำตาลในเลือดสูง***" (***Since a blood sugar level is high***.)
"***เนื่องจาก***/*since* ***ภาวะน้ำตาล***/*sugar level* ***ใน***/*in* ***เลือด***/*blood* ***สูง***/*high* "
EDU3:" [***ภาวะน้ำตาลในเลือดสูง***]***ทำให้มีสารเคมีบางชนิดเพิ่มสูงขึ้นในเลือด***" …
(**[*The high blood sugar level* /EDU2]*causes of having some increased chemical substance types in blood***.) …
"[***ภาวะน้ำตาลในเลือดสูง***/*high blood sugar level*/EDU2] ***ทำให้***/*cause* ***มี***/*have* ***สารเคมีบางชนิด***/ *some chemical substance type* ***เพิ่มสูงขึ้น***/*increase* ***ใน***/*in* ***เลือด***/*blood* " …

Example 2 contains the following $CEpair_i$ occurrence.

EDU2-EDU3 Pair as $CEpair_1$:EDU2 (cause)→ EDU3(effect)

With regard to this problem, we can solved this this problem by apply the NB machine learning technique to learn the annotated NWordCo-concept pair (the annotated $NWordCo_c NWordCo_e$-$pair_i$, i=1,2,..,lastLearntPair) feature with the CauseEffectRelation class from each EDU pair on the learning corpus after stemming words and eliminating stop words. The extracted NWordCo expressions are collected into an NWordCo-concept set (NWC) used as the CauseConcept set (CauseConcept set =NWC) and also the EffectConcept set (EffectConcept set =NWC) for the Cartesian product of CauseConcept ×EffectConcept as an *NWordCo-concept ordered pair* set. We then collect the $NWCP_{ce}$ set (see section 1) by using NB [17] with the feature probabilities of the annotated NWordCo-concept pairs having the cause-effect relation to the *NWordCo-concept order pair* set.

3.3. How to Extract CEpair Series Mingled with Non-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 EDUs is challenge. Therefore we propose using the stimulating-cue-word set and $NWCP_{ce}$ collection as the knowledge base to extract CEpair series including the stimulation relation occurrence from the documents.

4. A FRAMEWORK OF CEPAIR SERIES EXTRACTION

There are six steps in our framework, Corpus Preparation, NWordCo Size Learning, Collection of NWordCo with Event Concepts, NWordCo-Concept Pair Learning, Extraction of $NWCP_{ce}$, and Extraction of CEpair Series as shown in Figure 1.

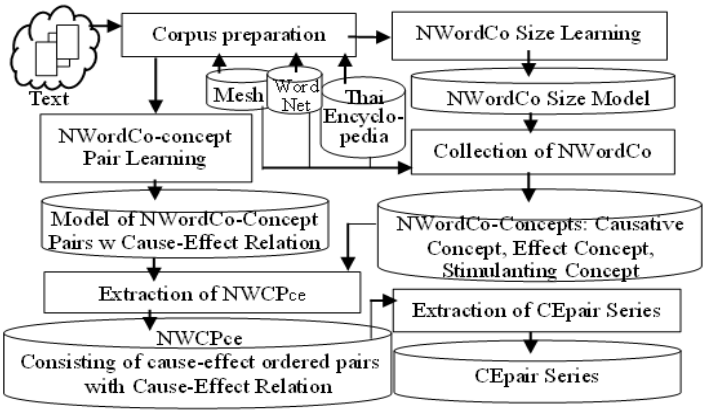

Figure 1. System overview

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 $NWCP_{ce}$ set. The third part of 500 EDUs is used for the CEpair 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://word-net.princeton.edu/ obtain) [26] and MeSH after translating from Thai to English by Lexitron (http://lexitron.nectec.or.th/).

"..[illegible]" [EDU1] "[illegible] [EDU2] [illegible] [EDU3] [illegible] [EDU4] …"
"..A patient gets a diabetes disease.[EDU1] since the body cannot fully use sugar inside the body. [EDU2] Because [the body] lacks of hormone insulin.[EDU3][lack of hormone insulin] causes Blood-sugar level to be high [EDU4]…"
<Topic_name Entity-concept=Diabetes/disease>[illegible]</Topic_name>…………
<CEpairSeries ID= 1>
<EDU1 CEpairID = 1 type=**effect**><NP1 concept= patient/human>[illegible]/ncn </NP1>
<VP marker =no><N-Word-CoExpression N=2 words **concept='get diabetes'** >
<w1: setType='Verb-weak' ; concept= 'get' boundary ='yes'>[illegible] </w1>
<w2: setType='Noun' ; concept= 'diabetes' boundary ='yes'>[illegible]</w2>
</N-Word-CoExpression ></VP></EDU1>
<EDU2 CEpairID = 1 type=**cause** | CEpairID=2 type=**effect**><NP1 concept= body/organ>[illegible]</NP1>
<VP marker=yes><N-Word-CoExpression N=4words **concept= 'not take body sugar to use'** >
<w1: setType='Verb-strong' ; concept='not take' boundary ='yes'>[illegible]</w1>
<w2: setType='Noun' ; concept= 'sugar' boundary ='yes'> [illegible]</w2>
<w3: setType='Noun' ; concept= 'body/organ' boundary ='yes'>[illegible]</w3>
<w4: setType='Verb-weak' ; concept= 'use' boundary ='yes'>[illegible]/w4>
<w5: setType='Adv' ; concept= 'fully' boundary ='no'> [illegible]/w5>
</N-Word-CoExpression></VP></EDU2>
<EDU3 CEpairID=2 type=**cause** | CEpairID=3 type=**cause** ><NP1 concept= body/organ> ϕ</NP1>
<VP marker=yes><N-Word-CoExpression N=2words concept= **'lack of insulin'**>
< w1: setType='Verb-strong' ; concept='lack of' boundary='yes'>[illegible]</w1>
< w2: setType='Noun' ; concept='insulin' boundary ='yes'>[illegible]</w2>
</N-Word-CoExpression></VP></EDU3>
……………………
The CEpairSeries tag is the CEpair series tag. The N-Word-CoExpression tag is the word boundary tag of each N-Word-Co expression. The w*i* tag is the word-*i* tag where *i*=1,2,..,*num*. .
The [..] symbol or ϕ means ellipsis (Zero Anaphora)

Figure 2. Annotation of NWordCo and CEpair series

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 (W_j) in a matrix vector (W).

$W_j=\{w_{j1},w_{j2},..,w_{jk}$ with 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,..,EDU*m*.

$W=\{W_j\}$ 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 $w_{j1} \in Verb_{weak} \cup Verb_{strong}$ (where w_{j1} 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 slide window size of two adjacent words with the one-sliding-word distance. As soon as *class*= 'non-CorEorS-concept' 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.
Verb=Verb_strong ∪ Verb_weak; W= Noun∪ Verb_strong ∪Adv∪Adj
NWCSet is an NWordCo-concept set, evp is an EDU's verb phrase

NWORDCO_EXTRACTION
  NWCSet←∅; NWco ←∅  ;  i=1 ; j=1;  k =0 ;  fl='no';
  while j≤ Length[L] do
   {1  If  i=1    then            /* identify the 1st word of NWordCo
      {2  If  (evp_j.w_i∈ Verb_strong ) then  {   NWco ←  evp_j.w_i ;  fl='yes' }
          Else If  (  evp_j.w_i ∈ Verb_weak)∧(  evp_j.w_i+1 ∈ W)      then
                {   NWco ← ( evp_j.w_i +  evp_j.w_i+1)  ;  i++ ;   fl='yes' }
         i++  }2                  /* determine N-Word-Co size
     while (fl='yes') ∧ ( evp_j.w_i ∈W) ∧ (i≤endOfVerbPhrase) do
       {3   i=i-1;
            Equation(2) ;
            If class= 'nonCorEorS_concept' then  fl← 'no'
            Else  fl←'yes';
            If  class= 'yes') then  NWco← NWco  ∪  w_i  ;
            i++ }3
     If   NWco<>∅  ∧  fl='no' then    /*append new NWordCo
     {  NWCSet← NWCSet ∪NWco;  i=1 ;  j++ ; NWco←∅}1
  } return    NWCSet
```

Figure 3. NWordCo extraction algorithm

Table 1. NWC Set Collection

NWordCo Expression	Concept
þÕ/occur- ¹éµÒÅ/sugar-ÅÍ´/blood-ÊÙ§/high	occur--sugar-blood-high
¢Ò/lackOff-ÎÍÃâÁ¹/hormone	lackOf-hormone
ÁÕ/have-ÀÒÇÐá·Ã¡«éÍ¹/complication-äµ/kidney	have-complication-kidney
·ÓãËé/causeTo-â»ÃµÕ/Protein-ÅÍ´/blood-µèÓ/low	cause-protein-blood-low
ÊÐÊÁ/collect-ä¢ÁÑ¹/fat-ËÅÍ´ÅÍ´à´§/artery	collect-fat-artery
àÊ×èÍÁ/deteriorate-ËÅÍ´ÅÍ´à´§/artery	occur-deteriorated-artery
ÊÙàÊÕÂ/lossOf-â»ÃµÕ/protein-»ÑÊÊÒÇÐ/urine	loss-protein-urine

$$NWordCoBoundaryClass = \arg\max_{class \in Class} P(class \mid w_{ij}, w_{ij+1}).$$
$$= \arg\max_{class \in Class} P(w_{ij} \mid class) P(w_{ij+1} \mid class) P(class).$$
$$where\ w_{ij} \in W_i\ ;\ and\ w_{ij+1} \in W_i\ (W_i\ is\ a\ CorEorS\_word\_concept\ vector)$$
$$i = \{1,2,..n\};\ j = \{1,2,..k\};\ CorEorS = a\ causative/effect/stimulating\ word\ concept;$$
$$Class = \{'CorEorS\_concept', 'non\_CorEorS\_concept'\} \quad (2)$$

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/weak/) are the probabilities of the annotated NWordCo-concept pairs as shown in Table 2.

Table 2. Show Probabilities of NWordCo-Concept Pair

NWordCo-Concept Pair: (CausativeNWordCoConcept)(EffectNWordCoConcept)	CauseEffect Rel. Probabilty	Non-Cause EffectRel. Probabilty
(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 $NWCP_{ce}$

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 ×EffectConcept as *NWordCo-concept order pair* set, NWCordP. We then extract and collect each NWordCo-concept pair class with the cause-effect relation into $NWCP_{ce}$ from NWCordP elements by using the NB classifier in (3) with the NWordCo-concept pair probabilities from Table 2 as the NB feature probabilities.

$$nwcpClass = \arg\max_{class \in Class} P(class \mid nwcOrdpair_k).$$
$$= \arg\max_{class \in Class} P(nwcOrdpair_k \mid class) P(class).$$
$$where\ nwcpClass\ \text{is an NWordCo - concept pair class};\ nwcOrdpair_k \in NWCordP;$$
$$Class = \{'CauseEffect\ \text{Re}lation', 'non\ \text{Re}lationt'\}$$
$$k = 1,2,..num;\ \ num\ \text{is the number of}\ NWCordP\ \text{elements}; \quad (3)$$

4.6. Extraction of CEpair Series

The objective of this step is to extract the CEpair series by matching *cnwcp* to $nwcp_{ce\text{-}k}$ as shown in Figure.4 where $nwcp_{ce\text{-}k} \in NWCP_{ce}$; k=1,2,..numberOf_$NWCP_{ce}$_element; and *tnwcp* is a testing NWordCo-concept pair which is the CEpair expression consisting of two consecutive NWordCo-concept expressions as the testing NWordCo concepts ($tnwc_1$,$tnwc_2$) extracted from the testing corpus. If match(*tnwcp*,$nwcp_{ce\text{-}k}$) then Series←Series ∪ *tnwcp* 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.
tnwcp is a testing NWordCo-concept pair from the series testing corpus.
tnwc is a testing NWordCo concept from the series testing corpus
nwc_j is an NWordCo concept of EDU*j* 's verb phrase. Scue is the stimulating-cue-word set
CEPAIR_SERIES_EXTRACTIONS

```
1  j=1; k=1; g=1; Series=∅ ; flg=0 ;i=1;fl='no';
2  nwc_j = NWordCo_Determination
       /*By using NWORDCO_EXTRACTION alg. Of Figure.3 from
          line no.3    through line no.14
3  while j≤ Length[L] do
4  {1 while g≤2 ∧ j≤ Length[L] do
5   {2 while  nwc_j=∅ ∧ j≤ Length[L] do
6       {j++ ; i=1;fl='no'; nwc_j= NWordCo_Determination}
7     If  nwc_j <> ∅ ∧ j≤ Length[L] then
           /*determine the stimulation relation EDU
           /*tnwc_g is a testing NWordCo element concept Of tnwcp
           /* w1 and w2 is word1 and word2 of tnwc_g
8       {3  tnwc_g ←nwc_j ;
9         If(tnwc_g.w1∈Scue)∨(tnwc_g.(w1+w2)∈Scue)then
10          {Series← Series ∪ tnwc_g }
11        Else  g++;
12        If  g≤2 then
13          {j++;i=1;fl='no';nwc_j= NWordCo_Determination}
14      }3}2
15   If   tnwc_1 ≠ ∅ ∧  tnwc_2 ≠ ∅ then
16    {4 while k≤NumberOf_NWCPceElements ∧ flg=0 do
              /* tnwcp←tnwc_1cause + tnwc_2effect;
17      {5 If  tnwc_1 + tnwc_2 match nwcp_ce-k then
18         {Series←Series∪'CEpair'+tnwc_1+tnwc_2 ; flg=1 }
               /*  tnwcp←tnwc_2cause+tnwc_1effect
19        ElseIf   tnwc_2+tnwc_1 match nwcp_ce-k then
20          {Series←Series∪'CEpair'+tnwc_2+tnwc_1;flg=1};
21        k++ }5 }4  ;
22    tnwc_1← tnwc_2 ; g=2 ; flg=0; nwc_j←∅; }1
23 }Return Series
```

Figure 4. CEpair series extraction

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 $NWCP_{ce}$ 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 $NWCP_{ce}$ 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 $NWCP_{ce}$ 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:"(*ความเสื่อม*/deterioration *ของ*/of *หลอดเลือดแดง*/artery)/NP1 (*เกิดขึ้น*/occur)/VP" ("*The deterioration of artery occurs*") and b) EDU:"(*น้ำตาล*/sugar *ใน*/in *เลือด*/blood)/NP1 (*ต่ำ*/be low)/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.Rezaii, "Diabetic kidney disease;review of the current knowledge,"*JRIP*, Vol.2, No.2, pp.73-80,2013.

[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.Figueiredoa, 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,"*IJCLCLP*, 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.½, 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*," ProcediaComput.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.Sukvakree, and A.Kawtrakul, "*Elementary Dis- course unit Segmentation for Thai using Discourse Cue and Syntactic Information*," *NCSEC2005* proceedings, Thailand, 2005, pp. 85-90.
[22] S. Sudprasert, A. Kawtrakul, Christian Boitet, 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, WF.Styler, C.Warner, JD.Hwang, JD.Choi, D.Dligach, RD.Nielsen, J.Martin, W.Ward, M.Pal -mer, GK. Savova, "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.