Event-Concept Pair Series Extraction to Represent Medical Complications from Texts

Chaveevan Pechsiri, Sumran Phainoun
College of innovative Technology and Engineering, Dhurakij Pundit University
110/1-4 Pracha Chuen, Bangkok, Thailand

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/disease complications which benefit for solving system. Each causative/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.

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.si.mahidol.ac.th/sidoctor/e-pl/), 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 benefits 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, CEpair, \((i=1,2,\ldots, last)\) 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 CEpair, sequence to represent Example 1.

CEpair\(_1\), CEpairs, ..., CEpairs\(_{last}\)

Example 1: Topic Name: ปัญหาโรคเบาหวาน/Diabetic Problems

... EDU1: “ผู้ป่วยเป็นโรคเบาหวาน” *(A patient gets a diabetes disease.)*

“ผู้ป่วย patient เป็น disease โรคเบาหวาน diabetes disease.”

EDU2: “เนื่องจากตับอ่อนสร้างฮอร์โมนอินซูลินได้น้อย” *(Since the pancreas produces less insulin.)*

“เนื่องจาก cause pancreas สร้าง produce ฮอร์โมนอินซูลิน insulin”

EDU3: “อินซูลินมีหน้าที่ส่งสัญญาณให้เซลล์นำน้ำตาลไปใช้” *(insulin has a function of signaling cells to take sugar for use.)*

“อินซูลิน insulin มีหน้าที่ function ของ signaling cells ผ่านน้ำตาล ไปใช้ take sugar for use”

EDU4: “เนื่องจาก cause ขาดฮอร์โมนอินซูลิน” *(When the body lacks of insulin.)*

“เนื่องจาก cause ขาด lack of ฮอร์โมนอินซูลิน insulin”

EDU5: “[ขาดฮอร์โมนอินซูลินในช่วง EDU4] ทำให้ขาด lack เหล่า far หลอดเลือด artery ตีบ constrict” *(lacking of hormone insulin/EDU4 makes the body unable to take the sugar for use.)*

“ขาดฮอร์โมนอินซูลิน lack of hormone insulin/EDU4] ทำให้ make ร่างกาย body ไม่สามารถ not able to take นำน้ำตาล ไปใช้ take sugar for use”

EDU6: “[ไม่สามารถนำน้ำตาลไปใช้ EDU5]เป็นสาเหตุให้เกิดระดับน้ำตาลสูงกว่าปกติ” *(Being unable to use the sugar/EDU5 is a cause of blood-sugar level being higher than normal.)*

“ไม่สามารถนำน้ำตาลไปใช้/Being unable to use the sugar/EDU5 เป็นสาเหตุ cause a cause of ระดับน้ำตาล blood-sugar-level สูงกว่า normal”

EDU7: “ซึ่งเป็นเรื่องของคุณกำลังของหลอดเลือดartery ทำให้ made ร่างกาย body” *(which is a catalyst for artery deterioration occurrence through the body.)*

“ซึ่ง which เป็น catalyst ทำให้ make การเสื่อม deterioration ของ หลอดเลือด artery ทำให้ make ร่างกาย body”

EDU8: “[การเสื่อมของหลอดเลือด artery EDU7]ทำให้ make หลอดเลือด artery ตีบ constrict.” *(The artery deterioration occurrence/EDU7 causes the arteries to constrict.)*

“การเสื่อมของหลอดเลือด artery/EDU7 ทำให้ make หลอดเลือด artery ตีบ constrict.”

EDU9: “[หลอดเลือด artery/EDU8]เป็นเหตุทำให้เกิดโรคหัวใจขาดเลือด” *(The constricted arteries /EDU8 is the cause of the ischemic heart disease.)*

“หลอดเลือด artery/EDU8 ตีบ constrict”

EDU10: “ดังนั้น โรคเบาหวาน เป็นปัจจัยเสี่ยงสำหรับต่อมอร์ทัวส์และสมอง 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 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 CEpairs1: EDU2 (Cause)  EDU1 (Effect)
EDU4-EDU5 Pair as CEpairs2: EDU4 (Cause)  EDU5 (Effect)
EDU5-EDU6 Pair as CEpairs3: EDU5 (Cause)  EDU6 (Effect)
EDU6-EDU7 as an intervention relation as the stimulation relation:
<highBloodSugar>… StimulationRelation…<artery Deterioration>
EDU7-EDU8 Pair as CEpairs4: EDU7 (Cause)  EDU8 (Effect)
EDU8-EDU9 Pair as CEpairs5: EDU8 (Cause)  EDU9 (Effect)

Event-Concept Pair Series Extraction to Represent Medical Complications… (Chaveevan Pechsiri)
where the stimulation relation on EDU6 co-occurs with the cause-effect relation on CEpair1 and CEpair3, as the part of the CEpair series. The CEpair Series representation of Example 1 is then shown in Figure 1.

![CEpair Series Representation](image)

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.

| EDU1 | NP1 VP | VP | Verb VP | Verb adv | Verb |
|------|--------|----|---------|----------|------|
| EDU2 | (BeDiabetic) | cause | (haveHigh BloodSugar) | (Deteriorate Artery) | effect |
| EDU3 | (unableToUse Sugar) | cause | (haveHigh BloodSugar) | (Contracted Artery) | effect |
| EDU4 | (unableToUse Sugar) | cause | (haveHigh BloodSugar) | (Constrict Artery) | effect |
| EDU5 | (Deteriorate Artery) | cause | (StimulationRelation) | (Constrict Artery) | effect |
| EDU6 | (Deteriorate Artery) | cause | (Constrict Artery) | (Constrict Artery) | effect |
| EDU7 | (Deteriorate Artery) | cause | (Constrict Artery) | (Constrict Artery) | effect |
| EDU8 | (Deteriorate Artery) | cause | (Constrict Artery) | (Constrict Artery) | effect |
| EDU9 | (Deteriorate Artery) | cause | (Constrict Artery) | (Constrict Artery) | effect |

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, i.e. ‘kapراด / catalyst’, ‘กระตุ้น / stimulus’, etc. Verbweak is a weak verb concept set requiring more information, i.e. ‘เป็น/be’, ‘ไม่เป็น/not_be’, ‘มี/have’, ‘ไม่มี/not_have’, ‘ใช้/use’.

Verbweak → { ทำให้/cause, เกิด/occur, ตีบ/constrict, ตบ/block up, ระหว่าง/terminate, ไม่ตอบสนอง/not_respond, เสื่อม/deteriorate, ขับ/excrete, เพิ่มขึ้น/increase, เปลี่ยนแปลง/change, อาเจียน/vomit, บวม/swell, ชัก/convulse, หมดสติ/be unconscious, สูง/high’, ‘ตาย/die’, ‘ เชื้อ/catalyze’, ‘ กระตุ้น/stimulus’, … }

Adj → { สูง/high’, ‘ … }.

Adv → { ‘ยาก/difficulty’, ‘น่า nota liquidity’, ‘ … }.

Noun → { ‘’, แผล/scar, ผู้ป่วย/patient, ไข้บาง/human organ, เลือด/blood, ปัสสาวะ/urine’, ความดัน/pressure, น้ำตาล:sugar’, ‘ไขมัน/fat’, ‘โปรตีน/protein’, อาการ/symptom’, ตรวจวิเคราะห์/contraction’, ‘สี/color’, ‘คัดรัง/catalyst’, ‘ … }.

There are several techniques [3-9] having been applied for determining the cause-effect causality/cause 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 CEpair series occurrence mingled with non-cause-effect relation EDUs including a stimulation relation EDU from the documents, and how to identify the common-cause/effect of the disease complications including their development parts from several CEpair 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 NWordCo) on the EDU’s verb phrase. The reason of using NWordCo 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 NWordCo expression on an EDU’s verb phrase of the research starts with a word, \( w_1 \) (where \( w_1 \in \text{Verb}_{\text{strong}} \cup \text{Verb}_{\text{weak}} \)), 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.

\[
\text{NWordCo expression} = w_1 + w_2 + \ldots + w_N
\]

(1)

where \( w_i \in \text{Verb}_{\text{strong}} \cup \text{Verb}_{\text{weak}} \); \( w_1, \ldots, w_N \in \text{Noun} \cup \text{Adj} \cup \text{Adv} \cup \text{Verb} \).

Thus, we apply each annotated NWordCo-expression pair with one NWordCo with a causative-event concept and another NWordCo with an effect-event concept to represent a cause-effect relation including an annotated NWordCo with a stimulating-event concept. We then apply Support Vector Machine (SVM) [10] to learn the NWordCo size (which is an \( N \) value) for extracting and collecting NWordCo expressions with the causative/effect/stimulating event concepts into an NWordCo-concept set. NWC. However, some NWordCo occurrences lack of information because their verb phrases consist of only one word, i.e. “\( \text{ไขมัน} \text{เพิ่ม} \)” (Fat level in the blood increases). Thus, we collect the NWC element from the one-word VP by adding two more words of the head noun of NPI as the following NWordCo expression: \( \text{ไขมัน} \text{เพิ่ม} \text{กระตุ้น} \text{เลือด} \).

We also apply the intersection set with the causative/effect concepts of NWC. NWC and NWC and which are the NWordCo-concept sets extracted from diabetes documents, kidney-disease documents, and artery-disease documents respectively as shown in equation (2). We then determine NWCP (which is an order pair set of the NWordCo-concept pairs having the cause-effect relation) by the Cartesian product of NWC along with the NB learning of relation-class probabilities from the annotated NWordCo-concept pairs. Therefore, the extracted NWCP can be expressed as the knowledge template, particularly the cause-effect-relation template in equation (3) for extracting the CEpair series from the documents.

\[
\text{NWC=} \text{NWC}_{\text{dbd}} \cup \text{NWC}_{\text{kd}} \cup \text{NWC}_{\text{artd}}
\]

(2)

\[
\text{NWCP}_{ce}=\{(\text{nwc}_{ci},\text{nwc}_{cj}), (\text{nwc}_{ci},\text{nwc}_{cj})\ldots (\text{nwc}_{ci},\text{nwc}_{cj}) (\text{nwc}_{ci},\text{nwc}_{cj})\ldots\}
\]

(3)

where: \( \text{nwc}_{ci}, \text{nwc}_{cj}, \ldots \text{nwc}_{ci}, \text{nwc}_{cj}, \ldots \in \text{NWC}; \)

\( \text{nwc}_{ci}, \text{nwc}_{cj} \) is 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 \( \text{nwc}_{cj} \) as an NWordCo with an effect-event concept; \( i, j \) are an integer.

And, we assign \( \text{nwc}_{ce,k} \in \text{NWCP}_{ce} \); therefore \( \text{nwc}_{ce,k} = (\text{nwc}_{ci}, \text{nwc}_{cj}) ; k=1,2,\ldots,\text{theNumberOfElementOfNWCP}_{ce} \).

We then propose using the cause-effect-relation template, NWCP ce, and the stimulating-cue-word set, \{‘\text{ไขมัน}\text{เพิ่ม}’\text{Verb}_{\text{weak}}\text{catalyst-Noun’}, ‘\text{กระตุ้น}’\text{Verb}_{\text{weak}}\text{stimulus-Verb}_{\text{strong}}’\} 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/effect concepts of NWC dbd, NWC kd, and NWC artd 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). Chang [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 latent topics to 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 ellipsis 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, {‘เพราะ/because’, ‘เนื่องจาก/since’, ‘ทำให้/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: โรคหวัดจากโรคเบาหวาน/Heart Disease from Diabetes

... EDU1: “ผู้ป่วยเบาหวานสามารถเป็นโรคหวัดได้” (A diabetic patient might get the heart disease.)
   “ผู้ป่วยเบาหวาน/โรคได้ที่เริ่มต้น โรคหวัดได้”

EDU2: “เนื่องจากภาวะน้ำตาลในเลือดสูง” (Since a blood sugar level is high.)
   “เนื่องจาก/since ภาวะน้ำตาลในเลือดสูง/มีโรคหวัด”

EDU3: “[ภาวะน้ำตาลในเลือดสูงทำให้มีการเกิดโรคที่เสี่ยงต่อในเลือด” (The high blood sugar level/EDU2) causes of having some increased chemical substance types in blood.)
   “[ภาวะน้ำตาลในเลือดสูง/มีโรคหวัด/blood sugar level/EDU2] ทำให้cause มีการเกิด/increase สารเคมีบางชนิด/some chemical substance type เพิ่มสูงขึ้น”

Example 2 contains the following CEpair occurs.

EDU2-EDU3 Pair as CEpair(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 CEpairs to CEpairs and CEpairs to CEpairs 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 NWCp,#, nwcp,#k, (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.

Indonesian J Elec Eng & Comp Sci, Vol. 12, No. 3, December 2018 : 1320 – 1333
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 concept 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.

\[ \text{IntNWC} = \text{NWC}_{\text{dia}} \cap \text{NWC}_{\text{kd}} \cap \text{NWC}_{\text{and}} \]

Example 3 Topic Name: โรคไตจากโรคเบาหวาน Diabetic Nephropathy

EDU1 : “ผู้ป่วยเป็นโรคเบาหวานประเภทที่2” (A patient gets type2 diabetes.)
“ผู้ป่วย/patient เป็น /get โรคเบาหวานประเภทที่2 /type2 diabetes”

EDU2 : “เพราะเจ้าหน้าที่ไม่ตอบสนองต่อฮอร์โมน” (because the body does not respond to the hormone.)
“เพราะ/because เจ้าหน้าที่/body ไม่/not ตอบสนอง/don’t respond_to ฮอร์โมน/hormone”

EDU3 : “เขาจะมีระดับน้ำตาลในเลือดสูงเกิน” (he will have too high blood sugar level.)
“เขา/he จะมี will have ระดับน้ำตาล/level น้ำตาล/sugar เลือด-blood สูง/high”

EDU4 : “รันได้รับจากโรคที่เกิดขึ้นในระยะที่สูง” (The high blood sugar level /EDU3 causes the kidneys to have extra-work in absorbing food nutrients and filtering waste.)
“[high blood sugar level/EDU] ทำให้/cause ไต/kidneys ทำงาน/ability to have extra work ดูดซึม/absorb สารอาหาร/food nutrients และ/and สร้างกรอง/and filter ของสิ่งส่ง/waste”

EDU5 : “ซึ่งเป็นสาเหตุที่ทำให้เกิดภาวะในระยะที่สูง” (which is the cause of deterioration in the kidney function afterwards.)
“ซึ่ง/which เป็น/cause สาเหตุ/event_of การเปลี่ยน/deterioration การทำงานของ/working kidney function ในระยะต่อมา/afterwards”

EDU6 : “และก็สุดท้าย [การเปลี่ยนในการทำงานของไต] ทำให้เกิดภาวะ” (And finally, [deterioration in the kidney function/EDU4] generates the kidney failure.)
“และก็สุดท้าย/and finally [deterioration kidney function/EDU4] ทำให้เกิด/cause เกิดภาวะ/kidney failure”

![Figure 2. CEPair Series Representation of Example 3](image-url)
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 with Event Concepts, NWordCo-Concept Pair Learning, Extraction of NWCPce, Extraction of CEpair Series, and Representation of Complication Development Parts as shown in Figure 4.

4.1. Corpus Preparation

This step is to prepare an EDU corpus from the chronic disease documents, i.e. diabetes, kidney disease, and artery disease, downloaded from hospitals web-boards (http://haamor.com/; http://www.bangkokhealth.com; http://www.si.mahidol.ac.th/sidoc tor/e-pl/). The step involves using Thai-word-segmentation tools [13] and Named-Entity recognition [14]. 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: the diabetes, kidney disease, and artery disease. The second part of 1200 EDUs having cause/effect relatio...
along with the causative/effect/stimulating concept as shown in Figure 5. 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://wordnet.princeton.edu/obtain) and MeSH after translating from Thai to English by Lexitron (http://lexitron.nectec.or.th/).

![Figure 5. Annotation of NWordCo and CEpair Series](image)

### 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\textsubscript{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

\[
\begin{align*}
(x) &= (w_\text{t} \cdot x) + b \\
&= \sum_{j=1}^{n} w_j x_j + b
\end{align*}
\]

where \( x \) is a dichotomous vector number, \( w_\text{t} \) is a weight vector, \( b \) is a bias, and \( (w_\text{t}, b) \in \mathbb{R}^n \times \mathbb{R} \) are the parameters that control the function. The SVM learning is to determine the weight, \( w_\text{t} \), 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 \( (w_j, w_j) \) with a CausativeOrEffectOrStimulating concept, after checking the first word occurrence on VP as follows.

If \( i = 1 \land (w_i \in \text{Verb\textsubscript{strong}} \lor \text{Noun}) \) then

\( \text{w}_i \) is the first word of VP with the CausativeOrEffectOrStimulating concept.

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

The results of learning the NWordCo size by SVM from the previous step is the weight vector of all \( w_i \) and \( w_{ij} \). 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.

Figure 6. NWordCo Extraction Algorithm

Moreover, some EDUs’ verb phrases consist of only one word of a verb, i.e. ‘เพิ่มขึ้น increase’ ‘สูงขึ้น high’ ‘ลดลง 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.

| NWordCo Expression | WordSequenceConcept | Concept |
|--------------------|---------------------|---------|
| เพิ่มขึ้นน้ำตาล-เลือดblood-sugar-blood-high | <occur-sugar-blood-high> | (haveHighBloodSugar) |
| น้ำตาล-เลือดblood-sugar-blood-high | <occur-sugar-blood-high> | (haveHighBloodSugar) |
| ขาดน้ำตาล-หอร์โมนlackOf-hormone | <lackOf-hormone> | (lackOfHormone) |
| มีภาวะแทรกซ้อน-สุญเสีย complication-kidney | <have-complication-kidney> | (haveKidneyComplication) |
| ทำหิวcauseTo- โปรตีน-เลือดprotein-blood-flow | <cause-protein-blood-low> | (haveLowBloodProtein) |
| กล้ามเนื้อ-ทำให้fat-collect-fat | <collect-fat> | (collectFatInArtery) |
| หลอดเลือดartery-เสื่อมdeteriorate | <artery-deteriorate> | (haveDeteriorationOfArtery) |
| สูญเสีย loseOSS- โปรตีน-protein-น้ำตาลurine | <loss-protein-urine> | (lossProteinToUrine) |
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(http://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 Table 2.

| NWordCo-Concept Pair: (CausativeNWordCoConcept)(EffectNWordCoConcept) | CauseEffect Rel. Probability | Non-CauseEffect Rel. Probability |
|---------------|-----------------------------|-------------------------------|
| (lackOfhormone)(haveHighBloodSugar) | 0.0171 | 0.0116 |
| (deteriorateArtery)(constrictArtery) | 0.0053 | 0.0029 |
| (collectFatInArtery)(causeArteriosclerosis) | 0.0053 | 0.0029 |
| (haveLowBloodProtein)(getSwellSymptom) | 0.0132 | 0.0116 |
| (haveLowBloodProtein)(getKidneyFailure) | 0.0020 | 0.0025 |
| (haveHighBloodSugar)(deteriorateArtery) | 0.0038 | 0.0048 |

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, NWCordP. Where nwcOrdpairh ∈ NWCordP ; h=1,2,...,num; num is the number of elements of NWCordP. We then extract and collect only nwcOrdpairh with the cause-effect relation into the NWCPce set by equation (6) with the probabilities of NWordCo-concept pairs occurrences from Table 2 resulted by the previous NB learning in section 4.4.

\[
\text{nwcOrdpRe l = arg \max_{\text{classClass}}} P(\text{class}|\text{nwcOrdpair}_h), \]

\[
= \arg \max_{\text{classClass}} P(\text{nwcOrdpair}_h | \text{class})P(\text{class}). \tag{6}
\]

where nwcOrdpRe l is the relation of nwcOrdpairh ;

\[
\text{nwcOrdpair}_h \in \text{NWCordP} \text{ which is an NWordCo order pair set};
\]

\[
\text{Class} = \{\text{CauseEffect Relation}, \text{nonCauseEffect Relation}\}
\]

\[
h = 1,2,...,\text{num}; \text{num is the number of elements of the NWCordP set};
\]

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 tmwcp and nwcpeck. Where tmwcp is an NWordCo-concept pair gained by sliding a window size of two consecutive EDUs/NWordCos as an NWordCo pair (tmwcp1 and tmwcp2) with one EDU/NWordCo distance from the 600EDUs testing corpus. And, nwcpeck ∈ NWCPce; tmwcp1 has a causative/effect concept whilst tmwcp2 has an effect/causative concept respectively.

\[
\text{MaxSimilarityScore} = \text{ArgMaxSimilarity}_{k=1}^{\text{numCEpair}} \frac{|tmwcp_p \cap nwcpeck|}{\sqrt{|tmwcp_p| \times |nwcpeck|}} \tag{7}
\]

where numCEpair is the number of NWCPce elements ;

\[
\text{tmwcp}_1 \text{ and tmwcp}_2 \text{ are the NWordCo concepts as a causative/effect concept and an effect/causative concept respectively from the testing corpus}
\]

\[
\text{tmwcp}_p \text{ is an NWordCo_concept pair, tmwcp}_1 \text{ and tmwcp}_2; \beta = 1,2;
\]

\[
\text{tmwcp}_1 = \text{tmwcp}_2 + \text{tmwcp}_2 \text{ if tmwcp}_1 \text{ is a cause; tmwcp}_2 = \text{tmwcp}_2 + \text{tmwcp}_1 \text{ if tmwcp}_2 \text{ is a cause;}
\]

\[
\text{nwcpeck} \in \text{NWCPce}; \text{ nwcpeck} = \text{nwcpeck}_1 \text{nwcpeck}_2 \text{nwcpeck}_3 \text{NWCPce}_j ;
\]

NWCPce is an ordered pair set of NWordCo_concept pairs having the cause_effect relation.
If MaxSimilarityScore between either tncw,tnwc1+tncw,tnwc2 or tncw,tnwc1+ tncw,tnwc3 and nwcpek,nwcpek+1,nwcpek+2,nwcpek+3,nwcpek+4,nwcpek+5 as shown in Figure 7 is greater than or equal to 90%, then both tncw and nwcpek are equivalent which results in nwcpek appended to a series as Seriesk leftarrow Seriesk+1 υ nwcpek+1 Where Seriesk 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.
NWCp,k is an ordered pair set of the NWordCo-concept pairs with the cause-effect relation.
NWCp,k ∈ NWCp,i; k is an index of an ordered pair element
nwcw is an NWordCo-concept pair from the testing corpus. tncw is an NWordCo concept from the testing corpus.
scue is an NWordCo concept of EDU’s’ verb phrase. Scue is the stimulating-cue-word set.

CEPAIR_SERIES_EXTRACTIONS

```java
public String getCause() {
  Public Relation(String Cause, String Effect{
    Class Relation{
      j=1;fl='no'; k=0; a=1; nwcj¬ ;
    2
    Class Relation{
      private String Cstring; private String Estring;
      private String Rstring;
      public Relation(String Cause, String Effect, String Relname){
        Cstring = Cause; Estring = Effect; Rstring = Relname
        public String getCAUSE() {
          return Cstring;
          etc …….
        }
      }
      ArrayList<Relation> Series = new ArrayList();
    3
    nwcj = NWordCo_Determination
      /*By Using NWORDCO_EXTRACTION algorithm
      Of Figure 6 from line no.3 through line no.16
    4
    While j< Length[L] do
    5     { While g< 2 ∧ j< Length[L] do
    6       { If nwcj ¬  then
    7         { tncwj ¬ nwcj; g++;
    8          j++;
    9          If g< 2 ∧ j< Length[L] then
    10            nwcj ¬ ; i=1;fl='no'; nwcj = NWordCo_Determination;
    11           If tncwj ≠  ∧ tncw ≠  ∧ tncw ≠  then
    12             /*determine the stimulation relation
    13              /*w1 and w2 is word1 and word2 of tncwj
    14            If (tncwj.w1)eScue)(tncw.j.w1+w2)eScue then
    15              /*the stimulation relation on the NWordCo occurrence.
    16              Series.add(new Relation(tncw, tncw, "stimulation Rel"));
    17          Else
    18            { tncw = tncw + tncw;
    19              /*if the result of Equation(7)≥90% then
    20              /*tncw=nWCpe, which is the CEpair element.
    21            If MaxSimilarityScore(tncw, NWCpe)≥90% then
    22              Series.add(new Relation(nwcpek,nwcpek,nwcpek+1,nwcpek+2,nwcpek+3,nwcpek+4,nwcpek+5, "CEpair Rel"); /*nwcpe and nwcpe is nwcj and nwcj;
    23              };
    24            };
    25            If g< Length[L] then
    26            { nwcj¬ ; i=1;fl='no'; nwcj = NWordCo_Determination;
    27          Return Series;
```

Figure 7. CEpair Series Extraction Algorithm

4.7. Determination of Complication Development Parts for Representation

With regard to section 4.3, we collect three different NWordCo-concept sets: NWCdb, NWCd, and NWCand from diabetes, kidney-disease, and artery-disease documents respectively. The intersection set, IntNWC, with the causative/effect concepts of NWCdb, NWCd, and NWCand consists of the following NWordCo elements: beHighbloodSugar, beHighbloodFat, inflameOrgan, deteriorateArtery, constrictArtery, highHighBloodPressure, getDisease, beComplication, beNonfunctional, and malfunction. However, some elements of IntNWC occasionally occur on the documents. Therefore, it is necessary to count and rank the top 5 intncw (where intncw ∈ IntNWC) by the number of intncw occurrences as shown in Table 3 to determine the most common-cause/effect (whose rank is equal to 1) of disease complications. The top 5 intncw from Table 3 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 8.
Swelling often begins at the feet.

Assume that ListSeries is an array of ArrayList for representation of some CEpair Series with the complication development for some disease types. ListSeries has two elements of ArrayList where each ArrayList element contains a CEpair series of one disease type.

### Table 3. Show top 5 intwv by number of occurrences

| Each randomed Disease has ≤150EDUs | beHighBlood-SugarLevel | beHighBlood-FatLevel | highHighBlood-Pressure | Inflame-Organ | Deteriorate-Artery |
|------------------------------------|-------------------------|----------------------|-------------------------|---------------|-------------------|
| KidneyDisease                      | 4                       | 2                    | 4                       | 3             | 5                 |
| Diabetes                           | 30                      | 6                    | 5                       | 5             | 3                 |
| ArteryDisease                      | 6                       | 14                   | 11                      | 8             | 3                 |
| total                              | 40                      | 22                   | 20                      | 16            | 11                |
| Rank                               | 1                       | 2                    | 3                       | 4             | 5                 |

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 NWCP_e 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 NWCP_e 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 NWCP_e are: 1) some causative event occurrences are based on an event expression by a preposition phrase whilst their effect events are expressed by their verbs, i.e. "(มักจะได้รับผลกระทบจากเส้นเลือดแดง)/NP1 ((สูงขึ้น)/degenerate) Verb ((จาก)/from) prep ((มี)/have) PP)/VP" (The arteries degenerate from having high blood lipids). 2) some effect event expressions occur on NP1, i.e. "(เสื่อม)/deteriorate (จาก)/from (การ)/of (บวม)/swelling) NP1 (ที่)/at (เท้า)/feet)/VP" (Swelling often begins at the feet.). 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.

**EDU1**-as Effect: “ผู้ป่วยเป็นโรคเบาหวาน” (A patient gets a diabetes disease.)

**EDU2**-as Cause: “เนื่องจากร่างกายขาดฮอร์โมนอินซูลิน” (Since the body lacks of insulin.)

**EDU3**-as nonCauseAndnonEffect: “อินซูลินมีหน้าที่ส่งสัญญาณให้เซลล์น้ำตาลไปใช้” (Insulin has a function of signaling cells to take sugar for use.)

**EDU4**-Effect: “[ขาดฮอร์โมนอินซูลิน/EDU2] ทำให้ร่างกายไม่สามารถนำน้ำตาลไปใช้ได้” ([lacking of insulin/EDU2] makes the body unable to take the sugar for use.)

where the following CEpairs can be determined except CEpairs2, CEpairs1: EDU2 (cause) $\rightarrow$ EDU1 (effect)

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

**Topic:** “ทำไมจึงเกิดภาวะแทรกซ้อนทางไต Why are there the kidney complication?”

**EDU1**-VPasCause: “ภาวะแทรกซ้อนทางไตในโรคเบาหวานเป็นผลจากการที่น้ำตาลในเลือดสูงกว่าระดับปกติ” (the kidney disease complication of diabetic disease is the result of the blood sugar level being higher than normal.)

**EDU2**-VPasEffect: “[การที่น้ำตาลในเลือดสูง/EDU1] ทำให้มีการเปลี่ยนแปลงของการไหลเวียนเลือดที่ไต” ([the blood sugar level/EDU1] causes to have changing of blood circulation in the kidneys.)

**EDU3**-VPasEffect: “และ [การที่น้ำตาลในเลือดสูง/EDU1] ยังทำให้มีการเปลี่ยนแปลงที่เนื้อไตโดยตรงด้วย” (and [the blood sugar level/EDU1] also makes changing the kidney cells.)

where EDU1 is a causative-event concept EDU having EDU2 and EDU3 as the effect-event concept EDU boundary.

CEpairs1: EDU1 (cause) $\rightarrow$ EDU2 (effect) $\land$ EDU3 (effect)

3) the complex sentence, e.g.

**Complex Stentence:** “ระดับน้ำตาลที่สูงนี้ทำให้เกิดปัญหาต่างๆ ตามมา” (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 CEpairs 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**

[1] Bird GL, Jeffries HE, Licht DJ, Wernovsky G, Weinberg PM, Pizarro C, Stellin G. Neurological complications associated with the treatment of patients with congenital cardiac disease: consensus definitions from the Multi-Societal Database Committee for Pediatric and Congenital Heart Disease. Cardiol Young. 2008; 18(S2): 234-239.

[2] Carlson L, Marcu D, Okurowski ME. Building a Discourse-TaggedCorpus in the Framework of Rhetorical Structure Theory. Current and New Directions in Discourse and Dialogue. 2003; 22: 85-112.

[3] Girju R. Automatic detection of causal relations for question answering. In Proc.of MultiSumQA’03-Proceedings of ACLworkshop on Multilingual Summarization and QuestionAnswering. Japan. 2003:76-83.

[4] Chang DS, Choi KS. Causal relation extraction using cue phrase and lexical pair probabilities. In Proc. IJCNLP. Hainan Island, China. 2004: 61-70.

[5] Pechsiri C, Piriyakul R. Explanation knowledge graph construction through causality extraction from texts. Journal of Computer Science and Technology. 2010; 25(5):1055-1070.

[6] Zhao S, Liu T, Zhao S, Chen Y, Nie J-Y. Event Causality Extraction Based on Connectives Analysis. Neurocomputing. 2016; 173:1943-1950.
[7] Mirza P, Tonelli S. CATENA: CAusal and TEmporal relation extraction from NATural language texts. In Proc. of COLING. Japan, 2016: 64-75.
[8] Sawamaru H, Kobayashi I. An Approach to Extraction of Causal Chain among Events in Multiple Documents. SCIS-ISIS 2012. Japan.2012: 1104-1108.
[9] Kusa W, Spranger M. External Evaluation of Event Extraction Classifiers for Automatic Pathway Curation: An extended study of the mTOR pathway. In Proc. of BioNLP workshop. Canada. 2017: 247–256.
[10] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines. Cambridge, UK: Cambridge University Press. 2000.
[11] Mitchell TM. Machine Learning. Singapore: The McGraw-Hill Co. Inc., and MIT Press. 1997.
[12] Mohammed S, Oakley S. University Of Sheffield: Two Approaches to Semantic Text Similarity. In: Proc. of First Joint Conference on Lexical and Computational Semantics. Canada. 2012: 655-661.
[13] Sudprasert S, Kawtrakul A. Thai Word Segmentation based on Global and Local Unsupervised Learning. NCSEC2003 proceedings. Thailand. 2003: 1-8.
[14] Chanlekha H, Kawtrakul A. Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. In Proc. IJCNLP. Hainan Island, China. 2004: 1-7.
[15] Chareonsuk J, Sukvakree T, Kawtrakul A. Elementary Dis-course unit Segmentation for Thai using Discourse Cue and Syntactic Information. NCSEC2005 proceedings. Thailand. 2005: 85-90.