2018 4th International C

Not secure | icsitech.org

ICSITech

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

HOME // KEYNOTE SPEAKERS // COMMITTEES // VENUE // REGISTRATION // SUBMISSION // PROGRAM

SUBMISSION >>

REGISTRATION >>

PREVIOUS CONFERENCES >>

FUTURE CONFERENCES >>

CONTACT>>

icsitech.org/#myCarousel

registration receipt.pdf | img-820123954-0....pdf | hotelpayment.pdf | Show all

ES18 conference | 2018 4th Internati... | ES18_เอกสารใช้งาน

EN 13:41 8/20/2018

Indonesian Journal of Electrical Engineering and Computer Science
Vol. 5, No. 3, March 2017, pp. 401 ~ 408
DOI: 10.11591/ijeecs.v5.i3.pp401-408

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
*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/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.

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

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, $CEpair_i$ (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 $CEpair_i$ sequence to represent Example 1.

$CEpair_1$, $CEpair_2$, $CEpair_3$,..., $CEpair_{last}$

Example 1: Topic Name: *ปัญหาโรคเบาหวาน/Diabetic Problems*
... **EDU1**: "*ผู้ป่วยเป็นโรคเบาหวาน*" (***A patient gets a diabetes disease.***)
"*ผู้ป่วย/patient เป็น/is โรคเบาหวาน/diabetes disease.*"
EDU2:"*เนื่องจากตับอ่อนสร้างฮอร์โมนอินซูลินได้น้อย*" (***Since the pancreas produces less insulin.***)
"*เนื่องจาก/since ตับอ่อน/pancreas สร้าง/produces ฮอร์โมนอินซูลิน/insulin*"

Received September 2, 2016; Revised December 25, 2016; Accepted January 11, 2017

EDU3:"อินซูลินมีหน้าที่ส่งสัญญาณให้เซลนำน้ำตาลไปใช้" (***insulin has a function of signaling cells to take sugar for use .***)
"อินซูลิน/*insulin* มีหน้าที่/*has a function of* ส่งสัญญาณให้เซล/*signaling cells* นำน้ำตาลไปใช้/*to take sugar for use*"

EDU4:"เมื่อร่างกาย ขาดฮอร์โมนอินซูลิน" (***When the body lacks of insulin.***)
"เมื่อ/*When* ร่างกาย/*body* ขาด/*lack of* ฮอร์โมนอินซูลิน/*insulin*"

EDU5:"[ขาดฮอร์โมนอินซูลิน/EDU4] ทำให้ร่างกายไม่สามารถนำน้ำตาลไปใช้ได้" (***[lacking of hormone insulin/EDU4] makes the body unable to take the sugar for use.***)
"[ขาดฮอร์โมนอินซูลิน/*lacking of hormone insulin*/EDU4] ทำให้/*make* ร่างกาย/*body* ไม่สามารถนำ/*unable to take* น้ำตาล ไปใช้ได้/ *sugar for use*"

EDU6:"[ไม่สามารถนำน้ำตาลไปใช้/EDU5]เป็นสาเหตุให้ระดับน้ำตาลในเลือดสูงกว่าปกติ" (***[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] เป็นสาเหตุให้/*is a cause of* ระดับน้ำตาล/*sugar-level* ใน/*in* เลือด/*blood* สูงกว่า/*higher than* ปกติ/*normal* "

EDU7:"ซึ่งเป็นตัวเร่งให้เกิดการเสื่อมของหลอดเลือดแดงทั่วร่างกาย" (***which is a catalyst for artery deterioration occurrence through the body.***)
"ซึ่ง/*which* เป็น/*is* ตัวเร่ง/*catalyst* ให้เกิด/*occur* การเสื่อม/*deterioration* ของ/*of* หลอดเลือดแดง/*artery* ทั่ว/*through* ร่างกาย/*body* "

EDU8:"[การเสื่อมของหลอดเลือดแดง/EDU7] ทำให้หลอดเลือดแดงตีบ" (***[The artery deterioration occurrence/EDU7] causes the arteries to constrict.***)
"[การเสื่อมของหลอดเลือดแดง/*arteries deterioration*/EDU7] ทำให้/*cause* หลอดเลือดแดง/*artery* ตีบ/*constrict*"

EDU9: "[หลอดเลือดแดงตีบ/EDU8] เป็นเหตุทำให้เกิดโรคหัวใจขาดเลือด" (***[The constricted arteries /EDU8] is the cause of the ischemic heart disease.***)
"[หลอดเลือดแดงตีบ/*constricted arteries*/EDU8] เป็นเหตุทำให้เกิด/*is the 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 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:
<highBloodSugar>... 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 $CEpair_3$ and $CEpair_4$ as the part of the CEpair series. The CEpair Series representation of Example 1 is then shown in Figure.1.

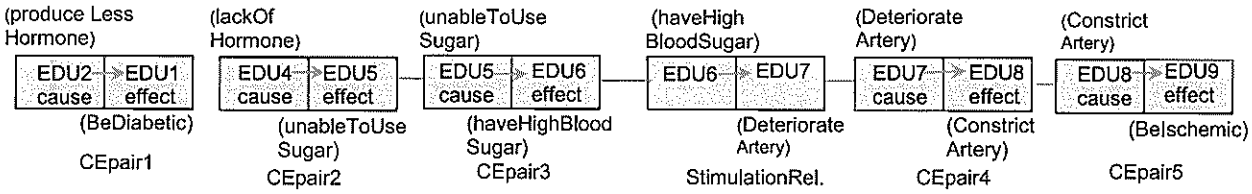

Figure 1. CEpair Series Representation of Example 1.

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 → $Verb_{weak}$ 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*', 'รีด/*constrict*', 'ตัน/*block up*', 'วาย/*terminate*', 'ไม่ตอบสนอง/*not_respond*', 'เสื่อม/*deteriorate*', 'ขับ/*excrete*', 'เพิ่มขึ้น/*increase*', 'เปลี่ยนแปลง/*change*','อาเจียน/*vomit*','บวม/*swell*','ชัก/*convulse*', 'หมดสติ/*beUnconscious*','สูง/*high*','ตาย/*die*','เร่ง/*catalyze*','กระตุ้น/ *stimulus*', ..}
Adj → {'สูง/*high*', ...} Adv→{'ยาก/ *difficultly*', 'เหลว/ *liquidly*', ...}
Noun→{' ', 'แผล /*scar*', 'ผู้ป่วย/*patient*', 'อวัยวะ/*human organ*', 'เลือด/*blood*','ปัสสาวะ/*urine*','ความดัน/*pressure*', 'น้ำตาล/*sugar*', 'ไขมัน/*fat*', 'โปรตีน/*protein*','อาการ/*symptom*','ตะคริว/*contraction*','สี../..*color*', 'ตัวเร่ง/*catalyst*',...}

where NP1 and NP2,are noun phrases. VP is a verb phrase. $Verb_{strong}$ 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, i.e. $Verb_{weak}$ Noun, to become either the cause-event/effect-event concept, i.e. 'เป็น/*be*+ลิ่มเลือด/*clot*', or the stimulation-event concept, i.e. 'เป็น/*be*+ตัวเร่ง/*catalyst*'. Noun is a noun concept set. Adv is an adverb concept set. Adj is the adjective concept set and Adjphrase is an adjective 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 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 $Verb_{weak}$ 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 $w_1 \in Verb_{strong} \cup Verb_{weak}$), 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.

$$\text{NWordCo expression} = w_1 + w_2 + \ldots + w_N \quad (1)$$

where $w_1 \in Verb_{strong} \cup Verb_{weak}$; $w_2,\ldots,w_N \in$ Noun ∪Adj ∪Adv ∪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. "(ระดับ/*level* ไขมัน/*fat:lipid* เลือด/*blood:liquid_body_substance*)/NP1 <u>(เพิ่ม/*increase*)VP</u>" (***Fat level in the blood <u>increases</u>***). 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 NWordCo expression: ระดับ/*level*+ ไขมัน/*fat*+ <u>เพิ่ม/***increase***</u> . We then apply Naive Bayes (NB) [11] to learn probabilities of NWordCo-concept pairs with a relation-class set, {CauseEffectRelation, nonCauseEffectRelation}, from the annotated corpus having the discourse cue ambiguity and some NWordCo 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 NWC_{dbd} , NWC_{kd}, and NWC_{artd} 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 $NWCP_{ce}$ (which is as an order pair set of the NWordCo-concept pairs having the cause-effect relation) by the Cartesian product of NWC×NWC along with the NB learning of the relation-class probabilities from the annotated NWordCo-concept pairs.

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.

$$NWC = NWC_{dbd} \cup NWC_{kd} \cup NWC_{artd} \quad (2)$$

$$NWCP_{ce}=\{(nwc_{c1},nwc_{e2}), (nwc_{c1},nwc_{e3})\ldots (nwc_{c2},nwc_{e1}) (nwc_{c2},nwc_{e3})\ldots.\} \quad (3)$$

where: $nwc_{c1}, nwc_{c2},\ldots, nwc_{e1}, nwc_{e2},\ldots \in NWC$;

(nwc_{ci}, nwc_{ej}) is an ordered pair of an NWordCo-concept pair having the cause-effect relation between nwc_{ci} as an NWordCo with a causative-event concept and nwc_{ej} as an NWordCo with an effect-event concept; *i and j* are an integer.

And, we assign $nwcp_{ce\text{-}k} \in NWCP_{ce}$; therefore $nwcp_{ce\text{-}k} = (nwc_{ci}, nwc_{ej})$; k=1,2,..,*theNumberOfElementOfNWCP*$_{ce}$

We then propose using the cause-effect-relation template, $NWCP_{ce}$, and the stimulating-cue-word set, {'*เป็นตัวเร่ง/be*-Verb$_{weak}$+*catalyst*-Noun', '*เร่ง/catalyze*-Verb$_{strong}$','*กระตุ้น/stimulus*-Verb$_{strong}$'...} 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 $CEpair_i$ 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',..}. 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: *โรคหัวใจจากโรคเบาหวาน/Heart Disease from Diabetes*

... **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)

Moreover, there are some EDU occurrences with both causative-concepts and effect-concepts, i.e. EDU5 and EDU8 of Example 1 on CEpair$_2$ to CEpair$_3$ and CEpair$_4$ to CEpair$_5$ 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$_{ce}$, $nwcp_{ce\text{-}k}$,(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.

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$_{ce}$ 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$_{ce}$ 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$_i$ 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.

$$\text{IntNWC} = \text{NWC}_{dbd} \cap \text{NWC}_{kd} \cap \text{NWC}_{artd} \quad (4)$$

Example 3. Topic Name: *โรคไตจากโรคเบาหวาน/Diabetic Nephropathy*

EDU1: "*ผู้ป่วยเป็นโรค เบาหวานประเภทที่2*" (***A patient gets type2 diabetes,***)
"*ผู้ป่วย/patient เป็น/get โรค เบาหวานประเภทที่/2type2 diabetes*"

EDU2:"*เพราะร่างกาย ไม่ตอบสนองต่อฮอร์โมน*" (***because the body does not respond to the hormone.***)
"*เพราะ/because ร่างกาย/body ไม่/not ตอบสนองต่อ/respond_to ฮอร์โมน/hormone*"

EDU3: "*เขาจะมีระดับน้ำตาลในเลือดสูงเกิน*" (***he will have too high blood sugar level.***)
"*เขา/he จะมี/will have ระดับ/level น้ำตาล/sugar เลือด/blood สูง/high*"

EDU4:"[ระดับน้ำตาลในเลือดสูง/EDU3]ทำให้ไตทำงานหนักในการดูดซึมสารอาหารและการกรองของเสีย([*The high blood sugar level* /EDU3] ***causes the kidneys to have extra-work in absorbing food nutrients and filtering waste.***)

"[*high blood sugar level*/EDU3[ทำให้/*cause* ไต/*kidneys* ทำงานหนัก/*have extra work* ดูดซึม/*absorb* สารอาหาร/*food nutrients* และการกรอง/*and filter* ของเสีย/*waste*"

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

"ซึ่ง/*which* เป็นสาเหตุให้เกิด/*is_cause_of* การเสื่อม/*deterioration* การทำงานของไต/*kidney function* ในเวลาต่อมา/*afterwards*"

EDU6: "และในที่สุด] การเสื่อมในการทำงานของไต [ทำให้เกิดไตวาย" (***And finally, [deterioration in the kidney function*/EDU4] *generates the kidney failure.***)

"และในที่สุด/*and finally*]*deterioration kidney function*/EDU4[ทำให้เกิด/*generate* ไตวาย/*kidney failure* "

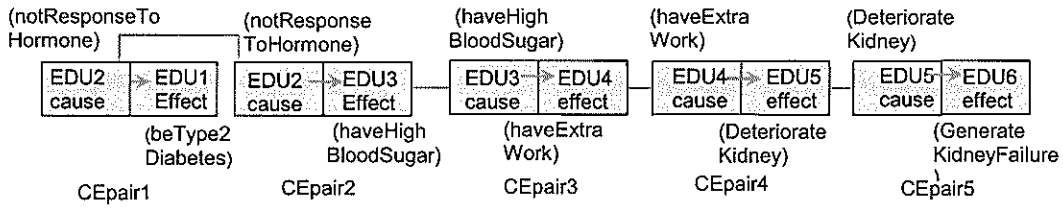

Figure 2. CEpair Series Representation of Example 3

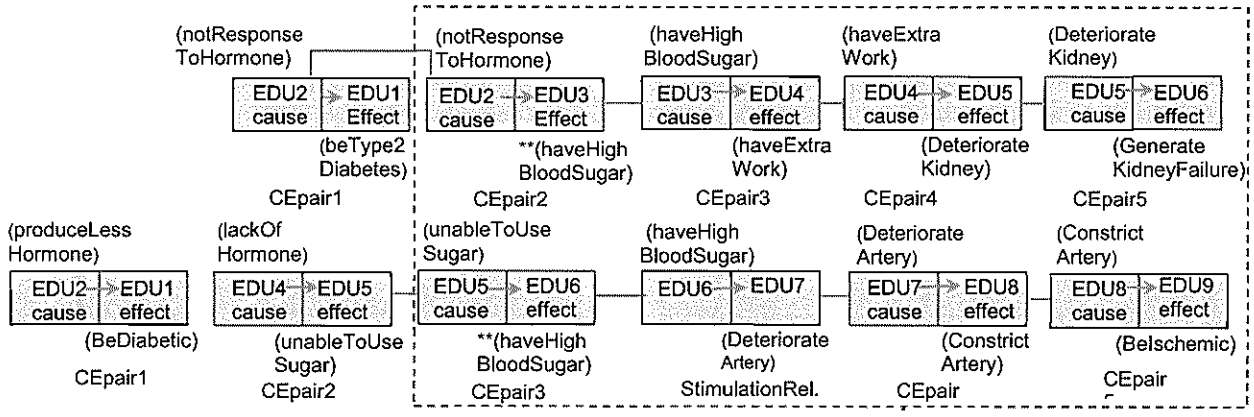

Figure 3. Show two complication development parts inside the rectangular dash line having "**" as the common-effect of both complications of Figure 1 and Figure 2.

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

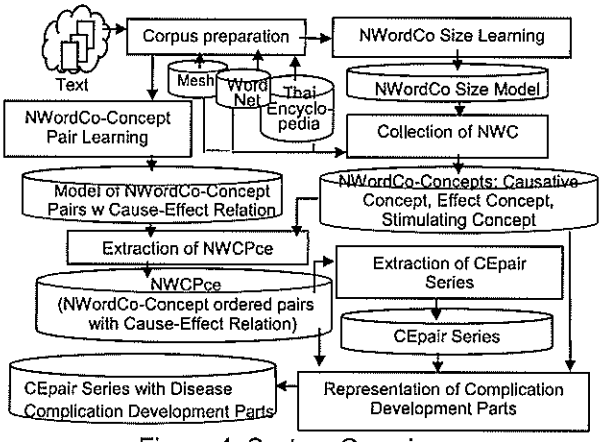

Figure 4. System Overview

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

4.1 Corpus Preparation

```
"..ผู้ป่วยเป็นโรคเบาหวาน" EDU1 "เนื่องจากร่างกายไม่สามารถนำน้ำตาลในร่างกายไปใช้ได้อย่างเต็มที่" EDU2 เพราะ [ร่างกาย] ขาดฮอร์โมนอินซูลิน EDU3 ทำให้ระดับน้ำตาลในเลือดสูงกว่าปกติ 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>โรคเบาหวาน</Topic_name>.....
<CEpairSeries ID= 1>
<EDU1  CEpairID = 1 type=effect><NP1 concept= patient/human>ผู้ป่วย/ncn </NP1>
<VP   marker =no><N-Word-CoExpression N=2 words  concept='getDiabetes' >
   <w1: setType='Verb-weak' ; concept= 'get' boundary ='yes'>เป็น </w1>
   <w2: setType='Noun' ; concept= 'diabetes' boundary ='yes'>โรคเบาหวาน</w2>
</N-Word-CoExpression ></VP></EDU1>
<EDU2 CEpairID = 1 type=cause | CEpairID=2 type=effect><NP1 concept= body/organ>ร่างกาย</NP1>
<VP  marker=yes><N-Word-CoExpression  N=4words  concept= 'notTakeSugarForUse' >
   <w1: setType='Verb-strong' ; concept='not take' boundary ='yes'>ไม่นำ</w1>
   <w2: setType='Noun' ; concept= 'sugar' boundary ='yes'>น้ำตาล</w2>
   <w3: setType='Noun' ; concept= 'body/organ' boundary ='yes'>ร่างกาย</w3>
   <w4: setType='Verb-weak' ; concept= 'use' boundary ='yes'>ใช้</w4>
   <w5: setType='Adv' ; concept= 'fully' boundary ='no'>อย่างเต็มที่</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= 'lackOf Hormone'>
   < w1: setType='Verb-strong' ; concept='lack of' boundary='yes'>ขาด</w1>
   < w2: setType='Noun' ; concept='insulin' boundary ='yes'>ฮอร์โมนอินซูลิน</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 wi tag is the word-i tag  where i=1,2,..,num.  .
The [..] symbol or ϕ means ellipsis (Zero Anaphora)
```

Figure 5. Annotation of NWordCo and CEpair Series

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 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_{dbd} , NWC_{kd}, and NWC_{artd}. NWC are also used for collecting $NWCP_{ce}$. The third part of 600 EDUs consisting of 200 EDUs from each disease is used for CEpair 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 Lexitron (http://lexitron.nectec.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

$$f(x) = \langle wt \cdot x \rangle + b = \sum_{j=1}^{n} wt_j x_j + b \quad (5)$$

where x is a dichotomous vector number, wt is a weight vector, b is a bias, and $(wt,b) \in R^n \times R$ are the parameters that control the function. The SVM learning is to determine the weight, wt_j, 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 $i=1 \wedge (w_i \in Verb_{strong} \cup V_{weak})$ then

w_i is the first word of VP with the CausativeOrEffectOr Stimulating 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 $j=1,2,..,n$ 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 (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
NP1 is a noun phrase; VP is a verb phrase ; enp1 is an EDU's NP1 ; evp is an EDU's VP;
NWC is an NWordCo-concept set
NWORDCO_EXTRACTION
  NWC←∅; 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
                           /*determine VP consisting of only one word
    {2 If (evp_i.w_i∈ Verb_strong ) ∧ numberOfWords( evp_i ) =1 then
        { NWco ← enp1_i.w_1 + enp1_i.w_2 + evp_i.w_i ; fl='no' }
       ElseIf (evp_i.w_i∈ Verb_strong ) then { NWco ← evp_i.w_i ; fl='yes' }
        ElseIf ( evp_i.w_i ∈ Verb_weak)∧( evp_i.w_i+1 ∈ W)  then
            { NWco ← ( evp_i.w_i + evp_i.w_i+1) ; i++ ; fl='yes' }
       i++ }2          /* determine N-Word-Co size
    while (fl='yes') ∧ ( evp_i.w_i ∈W) ∧ (i≤endOfVerbPhrase) do
     {3  i=i-1;
        Equation(5) ;
        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
    { NWC← NWC ∪NWco; i=1 ; j++ ; NWco←∅} }1
} return  NWC
```

Figure 6. NWordCo Extraction Algorithm

Table 1. NWordCo-Concept Set (NWC) Collection

NWordCo Expression	WordSequenceConcept	Concept
เกิด/occur-น้ำตาล/sugar-เลือด/blood-สูง/high	<occur-sugar-blood-high>	(haveHighBloodSugar)
น้ำตาล/sugar-เลือด/blood-สูง/beHigh	<sugar-blood-beHigh>	(haveHighBloodSugar)
ขาด/lackOff-ฮอร์โมน/hormone	<lackOf-hormone>	(lackOfHormone)
มี/have-ภาวะแทรกซ้อน/complication-ไต/kidney	<have-complication-kidney>	(haveKidneyComplication)
ทำให้/causeTo-โปรตีน/Protein-เลือด/blood-ต่ำ/low	<cause-protein-blood-low>	(haveLowBloodProtein)
สะสม/collect-ไขมัน/fat-หลอดเลือดแดง/artery	<collect-fat-artery>	(collectFatInArtery)
หลอดเลือดแดง/artery-เสื่อม/deteriorate	<artery-deteriorate>	(haveDeteriorationOfArtery)
สูญเสีย/lossOf-โปรตีน/protein-ปัสสาวะ/urine	<loss-protein-urine>	(lossProteinToUrine)
.....	.....	.....

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. 'เพิ่มขึ้น/*increse*' 'สูง/*behigh*' '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.

Table 2. Show Probability of NWordCo-Concept Pair

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

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.wakato.ac.nz/ml/weak/) 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, NWCordP. Where $nwcOrdpair_h \in$ NWCordP ; h=1,2,..,num; num is the number of elements of NWCordP. We then extract and collect only $nwcOrdpair_h$ with the cause-effect relation into the $NWCP_{ce}$ 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.

$$\begin{aligned} nwcOrdp\mathrm{Rel} &= \arg\max_{class \in Class} P(class \mid nwcOrdpair_h). \\ &= \arg\max_{class \in Class} P(nwcOrdpair_h \mid class)P(class). \end{aligned} \tag{6}$$

where $nwcOrdp\mathrm{Rel}$ is the relation of $nwcOrdpair_h$;
$nwcOrdpair_h \in NWCordP$ which is an NWordCo order pair set;
$Class = \{'CauseEffect\,\mathrm{Relation}','nonCauseEffect\,\mathrm{Relation}'\}$
$h = 1,2,..num$; 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 $tnwcp$ and $nwcp_{ce\text{-}k}$. Where $tnwcp$ is an NWordCo-concept pair gained by sliding a window size of two consecutive EDUs/NWordCos as an NWordCo pair ($tnwc_1$ and $tnwc_2$) with one EDU/NWordCo distance from the 600EDUs testing corpus. And, $nwcp_{ce\text{-}k} \in NWCP_{ce}$; $tnwc_1$ has a causative/effect concept whilst $tnwc_2$ has an effect/causative concept respectively.

$$MaxSimilarityScore = ArgMaxSimilarity_{k=1}^{numCEpair}\left(\frac{|tnwcp_\beta \cap nwcp_{ce-k}|}{\sqrt{|tnwcp_\beta| \times |nwcp_{ce-k}|}}\right) \tag{7}$$

where $numCEpair$ is the number of $NWCP_{ce}$ elements;
$tnwc_1$ and $tnwc_2$ are the NWordCo concepts as a causative/effect concept and an effect/causative concept respectively from the testing corpus
$tnwcp_\beta$ is a an NWordCo_concept pair, $tnwc_1$ and $tnwc_2$; $\beta = 1,2$;
$tnwcp_1 = tnwc_1 + tnwc_2$ if $tnwc_1$ is a cause; $tnwcp_2 = tnwc_2 + tnwc_1$ if $tnwc_2$ is a cause;
$nwcp_{ce-k} \in NWCP_{ce}$; $nwcp_{ce-k} = nwcp_{ce-k}.nwc_{ci} + nwcp_{ce-k}.nwc_{cj}$;
$NWCP_{ce}$ is an ordered pair set of NWordCo_concept pairs having the cause_effect relation.

If MaxSimilarityScore between either $tnwcp.tnwc_1$+$tnwcp.tnwc_2$ or $tnwcp.tnwc_2$+ $tnwcp.tnwc_1$ and $nwcp_{ce\text{-}k}.nwc_{ci}$+$nwcp_{ce\text{-}k}.nwc_{ej}$ as shown in Figure 7 is greater than or equal to 90%, then both *tnwcp* and $nwcp_{ce\text{-}k}$ are equivalent which results in $nwcp_{ce\text{-}k}$ appended to a series as $Series_a \leftarrow Series_a \cup nwcp_{ce\text{-}k}$ where $Series_a$ 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_ce is an ordered pair set of the NWordCo-concept pairs with the cause-effect relation.
nwcp_ce-k ∈ NWCP_ce; k is an index of an ordered pair element
tnwcp is an NWordCo-concept pair from the testing corpus.  tnwc is an NWordCo concept from the testing corpus.
nwcj is an NWordCo concept of EDUj 's verb phrase.  Scue is the stimulating-cue-word set.
CEPAIR_SERIES_EXTRACTIONS
 1  j=1; g=1; i=1;fl='no'; k=0; a=1 ; nwcj←∅;
 2  Class Relation{₀ 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 ..... }₀
    ArrayList<Relation>  Series_a = 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  {₁ While g ≤ 2 ∧ j ≤ Length[L] do
 6      {₂ If nwcj <> ∅   then
 7           { tnwc_g ←nwcj ;   g++ };
 8         j++;
 9         If  g ≤ 2 ∧ j ≤ Length[L] then
10           { nwcj ←∅;i=1;fl='no'; nwcj = NWordCo_Determination};
        }₂ ;
11    If   tnwc_1 ≠ ∅ ∧ tnwc_2 ≠ ∅ ∧ tnwc_1 ≠tnwc_2 then
             /*determine the stimulation relation
             /* w1 and w2 is word1 and word2 of tnwc_g
12      If  (tnwc_2.w1∈Scue)∨(tnwc_2.(w1+w2)∈Scue)then
             /* the stimulation relation on the NWordCo occurrence.
13          Series_a.add(new Relation(tnwc_1, tnwc_2,"stimulation Rel")
14      Else
15      {₃  tnwcp = tnwc_1 + tnwc_2 ;
          /* if the result of Equation(7)≥90% then
          /* tnwcp=nwcp_ce-k which is the CEpair element.
16       If MaxSimilarityScore(tnwcp, NWCPce )≥90% then
17         Series_a.add(new Relation(nwcp_ce-k.nwc_cα,nwcp_ce-k.nwc_cγ,
                  "CEpair Rel"); /*nwc_cα and nwc_cγ is nwc_c1 and nwc_c1
        }₃ ;                          /* respectively of Equation(7)
18     tnwc_1← tnwc_2 ; g=2;
19     If  j ≤ Length[L] then
20        {  nwcj←∅; i=1;fl='no'; nwcj = NWordCo_Determination}  }₁
21 }Return  Series_a
```

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: NWC_{dbd}, NWC_{kd}, and NWC_{artd} from diabetes, kidney-disease, and artery-disease documents respectively. The intersection set, IntNWC, with the causative/effect concepts of NWC_{dbd} , NWC_{kd}, and NWC_{artd} 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 *intnwc* (where *intnwc* ∈ IntNWC) by the number of *intnwc* 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 *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 3. Show top 5 *intnwc* by number of occurrences

Each randomed Disease has ≈150EDUs	Number Of NWordCo-Concept Occurrances				
	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

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

CEPAIR_SERIES_WITH_COMPLICATION_DEVELOPMENT

```
 1  ArrayList<Relation> listSeries[] = new ArrayList[2];
 2     listSeries[0] = Series1 ; /* a=1 from CEpair Series Extraction
                                  Algorithm for the 1st disease type.
 3     listSeries[1] = Series2 ; /* a=2 from CEpair Series Extraction
                                  Algorithm for the 2nd disease type.
 4  i=0;j=0;k=0; match=0;
 5  Size0=listSeries[0].size();   size1=listSeries[1].size();
 6  String[] ConceptRank = {"haveHighBloodSugar","haveHighBloodFat",
    "haveHighBloodPressure","inflame","detericrate"}
 7  while i < 5 ∧ match=0   do
 8   {3 While j <size0 ∧ match=0 do
 9     {4 While k<size1 ∧ match=0 do /*mark the common cause with "**"
10        {5 If listSeries[0].get(j).getRelname()= "CEpair" ∧
              listSeries[1].get(k).getRelname()= "CEpair" ∧ then
            { tempcj = listSeries[0].get(j).getCause();
              tempej = listSeries[0].get(j).getEffect();
              tempck = listSeries[1].get(k).getCause();
              tempek = listSeries[1].get(k).getEffect();
              If (listSeries[0].get(j).getCause()=(ConceptRank[i])then
                 listSeries[0].set(j,"** " + tempcj , tempej );
11            If(listSeries[0].get(j).getEffect()=(ConceptRank[i])then
12                listSeries[0].set(j, tempcj , "** " +tempej );
13            If (listSeries[1].get(k).getCause()=(ConceptRank[i])then
14                listSeries[1].set(k,"** " + tempck , tempek );
15            If(listSeries[1].get(k).getEffect()=(ConceptRank[i])then
16                listSeries[1].set(k, tempck ,"** " + tempek );
17             match=1;
18            }  k++; }5  k=0; j++  }4  k=0; j=0; i++; }3
20 }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 $NWCP_{ce}$ 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 $NWCP_{ce}$ 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 $NWCP_{ce}$ 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. "(*หลอดเลือดแดง/arteries*)/NP1 ((*เสื่อม/degenerate*)Verb (((*จาก/from*)/prep (*การมีไขมันในเลือดสูง/having high blood lipids*)/NP2)/PP)/VP" (***The arteries degenerate from having high blood lipids***). 2) some effect event expressions occur on NP1, i.e. "(*การบวม/swelling*)/NP1 (*มักจะ/often เริ่ม/begin ที่เท้า/on 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:"*เนื่องจากร่างกาย ขาดฮอร์โมนอินซูลิน*" (***Since the body lacks of insulin.***)
EDU3-as nonCauseAndnonEffect:"*อินซูลินมีหน้าที่ส่งสัญญาณให้เซลนำน้ำตาลไปใช้*" (***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 $CEpair_1$ can be determined except $CEpair_2$
$CEpair_1$:EDU2 (cause)→ EDU1(effect)
$CEpair_2$:EDU2 (cause)→ EDU4(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*/*and* [*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.
$CEpair_1$:EDU1 (cause)→ EDU2(effect) ∧ EDU3(effect)

3) the complex sentence, e.g.

Complex Stentence: "*ระดับน้ำตาลที่สูงนี้ทำให้เกิดปัญหาต่างๆ ตามมา*" (***This sugar level <u>which is high</u> causes problems as follows.***) where '***This sugar level <u>which is high</u>***' 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

[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*. InProc.of MultiSumQA'03Proceedings 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*. InProc.of BioNLPworkshop.Cannada.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.