Chain of Causation Determination from Texts

Chaveevan Pechsiri and Renu Sukharomana

***Abstract*—This research aims to determine the chain of causation of problem events, especially drug-addiction, expressed by several simple sentences from web documents. The chain of causation determination benefits for the problem-solving system. The research has three problems; how to determine a sentence having causative/effect event concept, how to determine the causative/effect event-concept vector size, and how to determine several consecutive causality relations (each causality is a relation between a causative-event-concept vector and an effect-event-concept vector) occurring as the chain of causation. Therefore, we apply WordCo to solve the cause/effect event concepts. We also use Support Vector Machine and WordCo features to solve the causative-event/effect-event vector size/boundary. We then propose using Na ïve Bayes to determine the consecutive causality relations between causative event-concept vectors and effect event-concept vectors. The research results provide the high precision of the chain of causation determination from the documents.**

***Index Terms*—Chain of causation, effect boundary, elementary discourse unit, WordCo.**

I. Introduction

The objective of this paper is to determine the chain of causation with concepts of problem events, especially drug-addiction, from downloaded web documents from hospital web-boards (i.e. http://haamor.com). In Regard to (https://www.merriam-webster.com/), 'chain of causation' is the causal connection between an original cause and its subsequent effects especially as a basis for criminal or civil liability. The drug addiction problems are increasing concern to people because they worry about the crime and violence that is associated with drugs. They also worry that drugs are becoming more widespread and are becoming increasingly easy for children to use. Beyond the harmful consequences for the person with the addiction, drug abuse can cause serious health problems for others, i.e. negative effects of prenatal drug exposure on infants and children. Therefore, the research concerns to determine the chain of causation of teen drug addiction from the documents for enhancing the warning system on the social web. The chain of causation of addiction contains two kinds of causative events, an external causative event (as the root cause) caused by the outside environments of addicts (i.e. a broken family, friends, etc) and an internal causative event caused by addicts themselves. Each consecutive causality relation are expressed by several EDUs (EDU is an Elementary Discourse Unit expression defined as a simple sentence or a clause, [1]) as shown in Example 1.

Example 1:

EDU1: "*พ่อแม่มีหนี้สินค้างจ่ายมาก/Parents have a lot of accrued liability*"

EDU2: "*ทะเลาะกันทุกวัน/* [*they*] *fight each other in every day*"

EDU3: "*ทำให้เด็กรู้สึกไม่อยากอยู่บ้าน/It causes the teen don't want to stay home*."

EDU4: "[*เด็ก*] *เริ่มเที่ยวกลางคืน/*[*He*] *start to stay out overnight.*

EDU5: "*และ*[*เด็ก*] *รู้สึกเครียด/ and* [*he*] *feels stress* ."

EDU6: "*ทำให้*[*เด็ก*]*เริ่มใช้ยาเสพติดเพื่อแก้ปัญหา/It causes* [*he*] *starts to <u>use drug</u> for solving problems*."

EDU7: "*สารเสพติดจะกระตุ้นระบบประสาท/The drug will activate the nerve system.*"

EDU8:"*เด็กจะหัวเราะร่าเริงได้ตลอดเวลา/He will laugh cheerfully all day* ."

EDU9: "*และ*[*เด็ก*]*เสพ*[*ยา*]*เพิ่มขึ้นเรื่อยๆ/and* [*he*] *use* [*drug*] *increasingly*."

EDU10: "*ทำให้*[*เด็ก*]*มีอาการประสาทหลอน/ It causes* [*he*] *has hallucination*."

EDU11:"*ต่อมา*[*เด็ก*]*มีอาการหงุดหงิด/Then*[*he*] *has a fidgety symptom*."

EDU12:"*เพราะ*[*เด็ก*]*ต้องการใช้ยา/because*[*he*]*craves to use drug*." ....

(where [..] means ellipsis.)

Example1 can be expressed as the chain of causation as follow.

Step 1 (EDU1∧EDU2): Cause → (EDU3∧EDU4∧EDU5): Effect.
Step 2 (EDU3∧EDU4∧EDU5): Cause → (EDU6): Effect.
Step 3 (EDU6): Cause → (EDU7∧EDU8∧EDU9∧EDU10): Effect.
Step 4 (EDU12): Cause→(EDU11): Effect.

where Step1-Step4 are the consecutive causality relations having the causality relation of Step1 with EDU1 and EDU2 as the root cause, the causality relation of Step2 with EDU5 as the internal cause, the causality relation of Step3 with EDU6 as the internal cause and EDU7 through EDU10 as addiction effects, and the causality relation of Step4 with EDU12 as the internal cause and EDU11 as a carving effect.

Moreover, the cause and effect events of the consecutive causality relation of the research mostly are expressed by several EDUs'verb phrases. 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 AdvPhrase_{dose}

Verb→Preverb Verb| V_{weak}-noun2| V_{weak}-noun2 Verb| V_{strong}| V_{strong} Verb

NP1 → pronoun | Noun1 | Noun1 modify| Noun2 | Noun2 modify

NP2→ Noun2 | Noun2 modify | modify

modify→Adj | Adj modify | V_s | V_s modify | Noun1 modify|Noun2 modify

V_{weak}→ {'*เป็น/be*', '*มี/have*', '*ใช้/use*', '*นำ/take*', '*เอา/get*', '*รู้สึก/feel*'}

V_{strong}→{'*ยากจน/be-poor*', '*ว่างงาน/be-jobless*', '*ทะเลาะ/quarrel, fight*', '*แยก/separate*','*ชักชวน/induce*',..,'*ดื่ม,กิน,เสพ/consume*','*ใช้/use*','*ฉีด/inject*','*สูดดม/sniff*',..,'*กระตุ้น/stimulate*','*ออกฤทธิ์/activate*', '*หวาดระแวง / be-mistrustful*' , '*ตื่นตัว/be-awakened-to*' , '*เสียสติ,บ้า /be-insane*',

Manuscript received July 15, 2017; revised October 9, 2017. This work was supported by Thailand Research Fund.

Chaveevan Pechsiri is with the College of Innovative of Technology and Engineering, Dhurakij Pundit University, Bangkok, Thailand (e-mail: chaveevan.pec@dpu.ac.th).

Renu Sukharomana was with Dhurakij Pundit University, Bangkok, Thailand. She is now with the Institute of Social and Economic Science, Thailand Research Fund, Bangkok, Thailand (e-mail: sukharenu@gmail.com).

doi: 10.18178/ijmlc.2017.7.5.627

‘คลุ้มคลั่ง / *be-manic-depression*’ , ‘ชัก/*convulse*’,‘หมดสติ / *lose-consciousness*’, ‘เสียชีวิต/*die*’, ‘เสื่อม/*deteriorate*’, ‘เคลิบเคลิ้ม/*be-absent-minded*’,‘กด/*depress*’,‘ลด/*reduce*’,..,‘ติด/*be-addicted-to*’,‘ขาด/*withdraw*’,‘อยาก,ต้องการ/*crave*’,..,‘เครียด/*be-stressed-out*’ ,‘หงุดหงิด/*fidget*’,‘วิตกกังวล/*be-anxious*’,‘กระวนกระวาย/*be-nervous*’,‘ก้าวร้าว/*be-aggressive*’, ‘ทำร้าย/*harm*’, ‘ซึมเศร้า/*sadden*’, ‘อ่อนเพลีย/*be-weak*’, …}

Noun1→{‘พ่อแม่/*parents*’ , ‘ครอบครัว/*family*’ ,‘เด็ก,วัยรุ่น / *youth, teenager*’ ,…}

Noun2→{‘ ’,‘ยา/*drug,addicted-substance*’,‘อาการ/*symptom*’, ‘หัวใจ/*heart*’,‘ประสาท/*nerve*’,‘สมอง/*brain*’,‘จิตใจ/*mental*’,‘ประสาทหลอน/*hallucination*’,..}

Adj→{‘สูง/*high*’,‘ต่ำ/*low*’..};

Adv→{‘อย่างแรง/*intensely*’,‘ซ้ำ/*repeatly*’..}; Preverb→{‘ไม่/*not*’..}

where NP1 and NP2,are noun phrases. VP is a verb phrase. V_{strong} is a strong verb concept set consisting of the causative verb concept set,V_{sc}, and the effect verb concept set, V_{se}, ($V_{strong}= V_{sc}\cup V_{se}$). V_{weak} is a weak verb concept set requiring more information, i.e. V_{weak}-Noun2, to have either the causative-event concept or the effect-event concept. Adv is an adverb concept set. Adj is the adjective concept set. In addition to Example1, there are several causality relation (CR*i*) occurrences in the consecutive order (or called the consecutive causality relations where i=1,2,..,*num* ; *num* is the number of causality relations). These consecutive causality relations consist of several events expressed by EDUs’ verb phrases with causative concepts and effect concepts as shown in the following.

<CR1><CR2> … <CRlast>

where

VP_{EDUc} = an EDU’s verb phrase with a causative concept.

VP_{EDUe} = an EDU’s verb phrase with an effect concept.

Causality Relation (CR*i*) consists of a causative vector of $VP_{EDUc\text{-}i}$ and an effect vector of $VP_{EDUe\text{-}i}$

CR*1*: $\langle VP_{EDUc\text{-}11}VP_{EDUc\text{-}12}..VP_{EDUc\text{-}1lastC1}\rangle \rightarrow$
$\langle VP_{EDUe\text{-}11}VP_{EDUe\text{-}12}..VP_{EDUe\text{-}1lastE1}\rangle$;

CR*2*: $\langle VP_{EDUc\text{-}11}VP_{EDUc\text{-}12}..VP_{EDUc\text{-}1lastC2}\rangle \rightarrow$
$\langle VP_{EDUe\text{-}11}VP_{EDUe\text{-}12}..VP_{EDUe\text{-}1\,lastE2}\rangle$; ....

CR*num*: $\langle VP_{EDUc\text{-}11}\ VP_{EDUc\text{-}12}\ ..\ VP_{EDUc\text{-}1lastCnum}\rangle \rightarrow \langle VP_{EDUe\text{-}11}$
$VP_{EDUe\text{-}12}\ ..\ VP_{EDUe\text{-}1\,lastEnum}\rangle$

There are several techniques [2]-[7] having been applied for determining the causality/causal relation from texts (see Section II). 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 III). The first problem is how to determine an EDU having the causative/effect event concepts. The second problem is how to determine the causative and effect event-concept vector size/boundary effected by the vector order. The third is how to determine each causality relation (CR*i*) between the causative event-concept vector and the effect event-concept vector. According to these problems, we need to develop a framework which combines machine learning and the linguistic phenomena to learn the several EDUs of the cause/effect expressions on the downloaded documents. Therefore, we collect a co-occurrence of two adjacent word components (called ‘WordCo’) with a causative event concept or the effect event concept from an EDU_j’s verb phrase (VP_{EDUj}; j is the EDU number) into the WordCo concept Matrix which is used for identifying cause/effect EDU on the testing corpus. A WordCo, $v_{co}\ w_{co}$, on VP_{EDUj} consists of the first component, v_{co} , as a group of 1-2words having the first word as a verb; and the second component, w_{co}, as a co-occurred word. Where $v_{co} \in V_{sc}\cup V_{se} \cup V_{wc}\cup V_{we}$; $w_{co} \in$ Noun2 $\cup\ V_{strong}\cup$ Adj $\cup$Adv; $V_{wc}=\{v_1+w_{c\text{-}1},\ v_2+w_{c\text{-}2},\ \ldots,v_\alpha+w_{c\text{-}\alpha}\}$; $V_{we}=\{v_1+w_{e\text{-}1},\ v_2+w_{e\text{-}2},\ \ldots,\ v_\beta+w_{e\text{-}\beta}\}$; and $v_j\in V_{weak}$; $w_{c\text{-}j},w_{e\text{-}j}\in$Noun2 with j=1, 2,.. α/β. Thus, all WordCo occurrences with causative/effect event concepts from the annotated corpus are collected into a WordCo set, WC. Where $WC= WC_c\cup WC_e$; WC_c is a WordCo set having causative event concepts and WC_e is a WordCo set having effect event concepts The WC elements are also used as features for the causative/effect event-concept vector determination through Support Vector Machine (SVM) [8]. We then propose using Na ïve Bayes (NB) [8] to determine CR*i* of the consecutive causality relations as the chain of causation.

Our research is separated into 5 sections. In Section II, related work is summarized. Problems in determining the chain of causation from texts are described in Section III and Section IV shows our framework of determining the chain of causation. In Section V, we evaluate and conclude our proposed model.

II. Related Works

Several strategies [2]-[7] have been proposed to determine the causal relation from texts without the chain of causation consideration except [7]. In 2003, [2] proposed decision tree learning the causal relation from a sentence based on the lexico syntactic pattern (NP1 causal-verb NP2). In 2004, [3] used cue-phrase and the statistical approach to NP-pair probabilities to solve the causal relation occurrence within two EDUs. In 2010, [4] applied verb-pair rules and machine learning techniques to extract the individual causality occurrence within several effect EDUs. There are more research works based on the lexico syntactic pattern with the causal concept as in [5] proposed the Restricted Hidden Na ïve Bayes model to learn and extract the causality from the English documents. The learning features [5] include contextual, syntactic, position, and connective features. In 2016, [6] 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. In 2012, [7] made causal chains by adding the causal chains obtained from latent topics to the causal chains obtained from word matching. The model’s [7] is based on noun features including hidden causal chains solved by latent topics.

However, most of the previous works on the individual causal/causality relation are based on NP1 and NP2 features of a sentence expression as NP1 verb NP2 existing on one/two sentences without the boundary consideration except [4] based on several EDUs’ verb phrases. However, [4]’s causality is mainly based on the effect boundary but without considering about the chain of causation. There are few works on determining the causal chain [7] based on NP1 occurrences whereas our work has NP1 ellipsis occurrences on the consecutive causality expressions as the chain of causation on the documents.

III. Problems of Determining Chain of Causation

A. How to Determine Causative/Effect Event Concept EDUs

Most of the causative/effect event occurrences on our documents are based on verb phrases with the causative/effect concepts provided by V_{strong} elements, i.e '*ซึมเศร้า/sadden*' as an effect concept, or V_{weak} elements along with Noun2 elements, i.e. '*ใช้/use*'+'*ยา/drug*' as a cause/effect concept. However, some V_{strong} elements or some V_{weak} elements along with Noun2 elements cannot provide the causative/effect event concepts as shown in the following.

Example 1

EDU1: "(*เขา/He*)/NP1 ((*ฉีด/inject*)/strong-verb (*เฮโรอิน/heroin*)/noun2 (*ด้วยตัวเอง/by himself*)/preproposition-phrase)/VP" ("*He inject heroin by himself*")

EDU2: "[(*เขา/He*)/NP1] ((*มี/has*)/weak-verb (*อาการ/symptom*)/noun2 (*เคลิบเคลิ้ม/be-absent-minded*)/strong-verb)/VP"

("[*He*] *has an absent-minded symptom*")

EDU1 contains the V_{strong} element as *ฉีด/inject* and EDU2 contains the V_{we} element as (*มี/has*)/weak-verb (*อาการ/symptom*)/noun2 where both elements cannot identify the causative/effect event concept. We then apply the WordCo concept to solve the above problems of identifying EDUs having the causative/effect concepts as follow: '*inject/consume-heroin/narcotic*' as a causative-event concept and '*have_symptom-be-absent-minded*' as an effect -event concept. However, there is another problem of $WC_c \cap WC_e \neq \varnothing$, i.e. '*ใช้/use*'+'*ยา/drug*' '*consume- narcotic*' , as shown in the following:

a) EDU1(cause):"*เขารู้สึกเครียดกับชีวิตของเขา/He feels stress with his life .*"

EDU2(effect): "*เขาจึงเสพยาเสพติดเพื่อผ่อนคลาย/He then consumes drug for relax.*"

b) EDU1(cause): "*เมื่อวัยรุนเสพยาบ่อยครั้ง/When a teen consumes drug quite often.*"

EDU2(effect):"[*เขา*]*ก็เริ่มมีอาการหลอน/*[*He*] *starts to have hallucination symptom.*"

Therefore, it is necessary to separate WC into three sets, WC_c, WC_e =, and WC_{ce} =, which are used for identifying causative/effect event concept EDUs. Where WC_c is a WordCo set with the causative concept as the external cause which is necessary to be identified before the internal cause identification, WC_e is a WordCo set with the effect concept from the internal cause, and WC_{ce} is a WordCo set with the causative concept as the internal cause in one relation and with the effect concept in another relation having the external/internal cause. Each WordCo set, WC_c, WC_e, and WC_{ce}, contain the high probability of $v_{co}w_{co}$ occurrences from several EDUs' verb phrases on the annotated corpus (see part B of Section IV).

B. How to Determine Causative and Effect Event-Concept Vector Size

The problem of how to determine the causative/effect event-concept vector size/boundary with the vector order consideration is challenge , i.e. CR1 CR2 where CR1 has the causality expression as <a *cause* event vector><an *effect* event vector> and CR2 has the causality expression as <an *effect* event vector><a *cause* event vector>. For example:

Example 2

EDU1: "*เมื่อเด็กใช้ยาเสพติดเพื่อแก้ปัญหา/Cause* [*him*] *starting to use drug for solving problems.*"

EDU2: "*สารเสพติดมีผลต่อสมอง/ The drug has an affect to the brain.*"

EDU3:"*เด็กเริ่มมีปัญหากับการเรียนในชั้น/He starts to have the problem of studying in the class.*"

EDU4:"*ต่อมา*[*เด็ก*]*มีอาการกระวนกระวาย/Then*[*he*]*has the impatient symptom.*"

EDU5:"*เพราะ*[*เด็ก*]*ต้องการใช้ยาอีก/because*[*he*]*craves to use drug again.*"

where CR1 has EDU1 as a *cause* vector and EDU2-EDU3 as an *effect* vector, CR2 has EDU4 as an *effect* vector and EDU5 as a *cause* vector.

Moreover, there is another problem of the cause/effect EDU boundary mingled with non-cause/-effect concept EDUs as shown in EDU4 of the following Example 3.

Example 3

EDU1 (cause): "*เมื่อวัยรุ่นได้เสพกัญชา/ "When a teen consumes opium.*"

EDU2(effect): "[*มัน*]*จะกระตุ้นการกดประสาท/*[*It*] *will stimulate sedation.*"

EDU3(effect):"*ทำให้ผู้เสพมีอาการประสาทหลอน/ Cause addicts to have hallucination symptom.*"

EDU4:"*สารที่อยู่ในกัญชามีหลายชนิด/There are several kinds of opium substances.*"

EDU5(effect): "*สารออกฤทธิ์จะมีผลต่อสมอง/The activator substance have an effect to brain.*"

Therefore, after we apply SVM having WC_c,WC_e, and WC_{ce} as the feature sets to solve both the causative event-concept vector (which is the causative boundary determination) and the effect event-concept vector(which is the effect boundary determination).

C. How to Determine Causality Relation,CRi

There is an effect event concept existing between two causative event concepts as shown in the Example 4.

Example 4

EDU1:"*สารเสพติดออกฤทธิ์ต่อระบบประสาท/*The drug activates the nerve system."

EDU2: "*ผู้เสพเริ่มรู้สึกหงุดหงิด/the addict starts to feel fidgety.*"

EDU3:" *และ[ผู้เสพ] มีอาการทุรนทุราย/*and[*he]* has a restlessness symptom."

EDU4: "*เพราะ[ผู้เสพ]ต้องการใช้ยา/*because[he]craves to use drug."

……………

where EDU1 and EDU4 are the causative event concepts. EDU2 through EDU3 are the effect event concepts. CR*1* occurs between EDU4 as a cause and EDU2-EDU3 as an effect but does not occur between EDU1 as a cause and EDU2-EDU3 as effects. Therefore, we propose using NB to determine CR*i* from the pair (A,B) where A is a WordCo feature vector with the causative event concept and B is a WordCo feature vector with the effect event concept.

IV. Framework of Determining Cain of Causation

There are five steps in our framework, corpus preparation, determining WordCo sets, feature vector extraction, learning consecutive causality relations, and determining the consecutive causality relations as the chain of causation from texts as shown in Fig. 1.

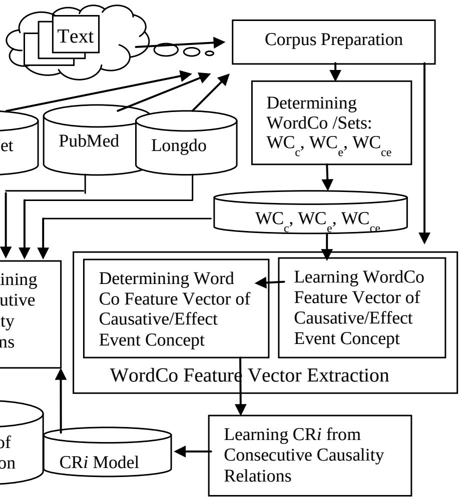
Text
Corpus Preparation
WordNet
PubMed
Longdo
Determining WordCo /Sets: WC_c, WC_e, WC_ce
WC_c, WC_e, WC_ce
Determining Consecutive Causality Relations
Determining Word Co Feature Vector of Causative/Effect Event Concept
Learning WordCo Feature Vector of Causative/Effect Event Concept
WordCo Feature Vector Extraction
Chain of Causation
CRi Model
Learning CRi from Consecutive Causality Relations

Fig.1. System overview.

A. Corpus Preparation

“กัญชา……………

วัยรุ่นใช้สารเสพติดระยะแรก $^{\text{EDU1}}$เพราะ[เขา] อยากลองสิ่งใหม่ๆ$^{\text{EDU2}}$ เมื่อ[เขา]ได้เสพกัญชาในระยะต่อมา$^{\text{EDU3}}$

กัญชาจะออกฤทธิ์กล่อมประสาท$^{\text{EDU4}}$..”

“Cannabis………………

A teenager uses an addictive substance at first .$^{\text{EDU1}}$

because [he] wants to try the new thing.$^{\text{EDU2}}$ When [he] consumes narcotic drug later.$^{\text{EDU3}}$

the drug activates sedation $^{\text{EDU4}}$…**”**

<Topic_name Entity-concept= Ectasy/drug>ยาอี</Topic_name>…………

<EDU1 ><NP1 concept= teenager/person>วัยรุ่น/ncn </NP1>

<VP **Type=effect**><WordCo **Set3=Wce** concept= ‘**use addictive substance**’>

<*v*co: Type=‘**weak-verb**’><*w*1 type= ‘verb’ concept= ‘**use**’>ใช้ </*w*1>

<*w2*: Type=‘**noun2**’ ; concept= ‘substance’ >สาร</*w2*></*v*co >

< *w*co: Type=‘adj’ ; concept= ‘**addictive**’ >เสพติด </*w*co ></WordCo>ระยะแรก

</VP> </EDU1>

<EDU2><Conj concept=because>เพราะ</Conj><NP1 concept= teen/person>ϕ</NP1>

<VP **Type=cause** ><WordCo **Set1=Wc** concept= **‘want to try’**>

<*v*co:Type=‘**strong-verb**’><*w*1 type=‘**cause-verb**’concept=‘**want**’>อยาก</*w*1></*v*co >

< *w*co: Type=‘**verb**’ ; concept= ‘**try**’ >ลอง</*w*co ></WordCo>สิ่งใหม่</VP></EDU2>

<EDU3 ><Conj concept=when>เมื่อ</Conj><NP1 concept= teen/person>ϕ</NP1>

<VP **Type=cause** ><WordCo **Set3= Wce** concept= ‘**consume drug’**>

<*v*co:Type=‘**strong-verb**’><*w*1 type=‘**cause-verb**’ concept=‘**consume**’>เสพ</*v*co >

< *w*co:Type=**‘noun2**’; concept=‘**cannabis/drug**’>กัญชา</*w*co ></WordCo>ในระยะต่อมา

</VP> </EDU3>

<EDU4 ><NP1 concept= ‘cannabis /drug’ >กัญชา</NP1>

<VP **Type=effect**><WordCo **Set2=We** concept= ‘**activate the neuron’**>

<*v*co:Type=‘**strong-verb**’><*w*1 type=‘**cause-verb**’ concept=‘**activate**>ออกฤทธิ์</*v*co >

<*w*co: Type=‘ **strong-verb**’ ; concept= ‘**sedate**’ >กล่อมประสาท</*w*co ></WordCo>

</VP> </EDU4>

……………………

The WordCo tag consists of a *v*co tag and *w*co tag. The *v*co tag is a verb tag as the first word of the WordCo tag. The *w*co tag is a co-occurred word tag to the *v*co tag. The [..] symbol or ϕ means ellipsis (Zero Anaphora)

Fig. 2. Annotation corpus.

This step is to prepare an EDU corpus from the addiction-problem documents downloaded from hospital web-boards. The step involves using Thai-word-segmentation tools [9] and Named-Entity recognition [10]. After the word segmentation is achieved, EDU Segmentation [11] is then operated to provide a 2900 EDUs’ corpus. The corpus included stemming words and the stop word removal is separated into 3 parts; an 800-EDUs’ part as the studying corpus for determining WordCo sets with causative/effect event concepts. The next 1600-EDUs’part as the learning and extracting/testing corpus is used for 1) learning the WordCo feature vector size/boundary within 800 EDUs and 2) extracting WordCo feature vectors with the causative/effect event concepts and also learning the causality relation from the consecutive pair(A,B) expressions within the other 800 EDUs. The last500-EDUs’part as the testing corpus is for determining the consecutive CR*i* occurrences as the chain of causation. With Regard to Fig. 2 on the studying corpus and the learning corpus, we semi-automatically annotate the causative/effect event concepts of all WordCo occurrences along with three property sets of the WordCo-set tags; a ‘Wc’ is a property set of the WordCo tag with the causative concept which is the external cause as the root cause, a ‘We’ is a property set of the WordCo tag with the effect concept from the internal cause, and a ‘Wc$_e$’ is a property set of the WordCo tag with the causative concept as the internal cause in one relation and with the effect concept in another relation having the external/internal cause. These property sets are then collected into WordCo sets as WC$_c$, WC$_e$, and WC$_{ce}$ respectively in the next step. All concepts of WordCo are referred to Wordnet (http://word-net.princeton.edu/obtain) and MeSH after translating from Thai to English, by Lexitron (the Thai-English dictionary) (http://lexitron.nectec.or.th/).

B. Determining WordCo Set

According to the annotated corpus as the studying corpus, the WordCo-element tags , <v_{co}><w_{co}>, as $v_{co}w_{co}$ occurs on several EDUs’ verb phrases. We then determine the probabilities of v_{co} w_{co} according to the property sets as Set*k* (*k*=1, 2, 3) of the annotated corpus to separate the WC set into three subsets of WC$_c$, WC$_e$, and WC$_{ce}$ respectively as follow.

If Probability (v_{co} w_{co} ∈ Set1/Wc) ≥0.9

{ k=1; $v_{cok}w_{cok} = v_{co}w_{co}$; $v_{cok}w_{cok} \in$ WC$_c$ }

Else-If *Probability* (v_{co} w_{co} ∈ Set2/We) ≥0.9

{ k=2; $v_{cok}w_{cok} = v_{co}w_{co}$; $v_{cok}w_{cok} \in$ WC$_e$ }

Else-If *Probability* ($v_{co}w_{co}$∈ Set*k*/Wce) ≥0.9

{ k=3 ; $v_{cok}w_{cok} = v_{co}w_{co}$; $v_{cok}w_{cok} \in$ WC$_{ce}$ }

WC$_c$,WC$_e$, and WC$_{ce}$ then consist of the following elements.

WC$_c$={‘*ว่าง-งาน/be-jobless*’, ‘*ยากจน-‘ ’/be-poor*’, ‘*แยก-ตัว/separate-himself*’, ‘*หย่า-‘ ’/divorce*’, ‘*รู้สึก-ทรมาน/feel suffering*’, ‘*อยาก-ลอง/want-to-try*’,… }

WC$_{ce}$={‘*กิน,เสพ-ยาเสพติด/consume-narcotic*’,‘*ใช้-ยาเสพติด/use,consume-narcotic*’,‘*ฉีด-ยาเสพติด/inject,consume-narcotic*’,‘*อยาก,ต้องการ/-ยาเสพติดcrave*-narcotic’,‘*ขาด-ยาเสพติด/withdraw*’, ‘*ออกฤทธิ-‘ ’/activate*’, ‘*กระตุ้น-ประสาท/stimulate*’, ‘*กด-ประสาท/be-sedative*’,…}

WC$_e$={‘*รู้สึก-หวาดระแวง/be-mistrustful*’,‘*ทำร้าย-ร่างกาย/commit-bodily-harm*’,‘*รู้สึก-เคลิบเคลิ้ม/be-absent-minded*’,‘*รู้สึก-กระวนกระวาย/be-nervous*’,‘*รู้สึก-เครียด/be-stressed-out*’, ‘*มีอาการ-ประสาทหลอน/ have-hallucination-symptom*’, ‘*มีอาการ-ง่วงซึม/ have-drownsiness-symptom*’ ,..}

Thus, WC$_c$, WC$_e$, and WC$_{ce}$ are used for determining the causative/effect event-concept EDUs and also the WordCo feature vector with the causative/effect event concept.

C. Feature Vector Extraction

There are two steps for extracting the WordCo feature vector with the causative/effect event concept, the first step is a WordCo Feature Vector Size Learning step by SVM [8], [12] and the second step is a WordCo Feature Vector Determining step.

1) WordCo fearture vector size learning

This step applies SVM to learn the WordCo feature vector size/ boundary with either the causative event concept or the effect event concept of each the causative event concept vector / each effect event concept vector respectively. According to [12], the linear function in (1), $f(x)$ or $f(\mathrm{v}_{\mathrm{co}k}\mathrm{w}_{\mathrm{co}k})$, of the input $\mathrm{v}_{\mathrm{co}k}\mathrm{w}_{\mathrm{co}k}= \langle v_{\mathrm{co}k\text{-}1}\ w_{\mathrm{co}k\text{-}1}\dots\ v_{\mathrm{co}k\text{-n}}\ w_{\mathrm{co}k\text{-n}}\rangle$ assigned to the positive-class/BoundaryContinuing if $f(\mathrm{v}_{\mathrm{co}k}\mathrm{w}_{\mathrm{co}k}) \geq 0$; and otherwise to the negative-class/ EndOfBoundary as a vector size if $f(\mathrm{v}_{\mathrm{co}k}\mathrm{w}_{\mathrm{co}k}) < 0$. In addition, $v_{cok}w_{cok} \in \mathrm{WC_c}$ where k=1; $v_{cok}w_{cok} \in \mathrm{WC_e}$ where k=2; and $v_{cok}w_{cok} \in \mathrm{WC_{ce}}$ where k=3.

$$\begin{aligned} f(x) &= \langle wt.x \rangle + b \\ f(v_{cok} w_{cok}) &= \left\langle wt.v_{cok} w_{cok} \right\rangle + b \qquad (1) \\ &= \sum_{j=1}^{n} wt_j (v_{\mathrm{co}k\text{-}j} w_{\mathrm{co}k\text{-}j}) + b \end{aligned}$$

where x is $\mathrm{v}_{cok}\mathrm{w}_{cok}$; $v_{\mathrm{cok}} w_{\mathrm{cok}} \in WC_c$ if $k=1$; $v_{\mathrm{cok}} w_{\mathrm{cok}} \in WC_e$ if $k=2$; $v_{\mathrm{cok}} w_{\mathrm{cok}} \in WC_{ce}$ if $k=3$

where $\mathrm{v}_{\mathrm{co}k}\mathrm{w}_{\mathrm{co}k}$ is a dichotomous vector number, wt is the weight vector, b is bias, and (wt, b)$\in R^n \times R$ are the parameters that control the function. The SVM learning is applied to the research to determine wt_j and b for each WordCo concept feature for (x_j) or ($v_{\mathrm{co}k\text{-}j}w_{\mathrm{co}k\text{-}j}$) in WordCo-concept pair ($v_{\mathrm{co}k\text{-}j}w_{\mathrm{co}k\text{-}j}\ v_{\mathrm{co}k\text{-}j+1}w_{\mathrm{co}k\text{-}j+1}$) from a sliding window size of two consecutive EDUs (EDUj EDUj+1) with the sliding distance of one EDU by using Weka(http://www.cs.wakato.ac.nz/ml/weka/) in each causality relation (CRi) from the annotated corpus as the learning corpus. (where n = *EndOfBoundary*)

2) WordCo fearture vector determination

The results from SVM learning are weight, *wt*, and bias, *b*, of each feature ($v_{\mathrm{co}k\text{-}j}w_{\mathrm{co}k\text{-}j}$). According to equation 1, the input vector of WordCo features ($\mathrm{v}_{\mathrm{co}k}\mathrm{w}_{\mathrm{co}k}$) having the WordCo-concept pair, $v_{\mathrm{co}k\text{-}j}w_{\mathrm{co}k\text{-}j}v_{\mathrm{co}k\text{-}j+1}w_{\mathrm{co}k\text{-}j+1}$, including their weights and bias are used to determine the boundary of the causative/effect event-concept vector. If $f(x)$<0, an ending class (EndOfBoundary) occurs, otherwise a continuing class (BoundaryContinuing) by sliding a window size of two consecutive EDUs with one EDU sliding distance to form the WordCo-concept pair as the input vector of (1) on the testing corpus.

D. Causality Relation Learning

Each pair (A,B) extracted by the previous step consists of several WordCo occurrences (with causative/effect event concepts) used as the learning features of this step. These learning features are used for learning the causality relation by using Weka (http://www.cs.wakato .ac.nz/ml/weka/) to determine probabilities of $a_1,..,a_g,b_1,..,b_h$ where $a_1,..,a_g \in \mathrm{WC_c}\cup\mathrm{WC_{ce}}$; $b_1,..,b_h \in \mathrm{WC_e}\cup\mathrm{WC_{ce}}$. A is a causative vector which consists of all elements of a WordCo feature vector with the causative event concept; B is an effect vector which consists of all elements of a WordCo feature vector with the effect event concept. Then $a_1,..,a_g$ can be represented by $v_{\mathrm{co}k\text{-}1}w_{\mathrm{co}k\text{-}1},..,v_{\mathrm{co}k\text{-g}}w_{\mathrm{co}k\text{-g}}$ where k=1 or 2; and $b_1,..,b_h$ can be represented by $v_{\mathrm{co}k\text{-}1}\ w_{\mathrm{co}k\text{-}1},..,\ v_{\mathrm{co}k\text{-}h}w_{\mathrm{co}k\text{-}h}$ where k=2 or 3 and k in A $\neq$ k in B with the Class-type set of the causality relation,{'yes' 'no'}. The Class-type set is specified by the experts.

E. Determining Consecutive Causality Relations

The objective of this step is to recognize and extract each CRi expression as the consecutive causality relations from the testing corpus by using Na ïve Bayes [8] in (2) with the probabilities of $a_1,..,a_g$, $b_1,..,b_h$ (or $v_{\mathrm{co}k\text{-}1}w_{\mathrm{co}k\text{-}1},..,v_{\mathrm{co}k\text{-g}}w_{\mathrm{co}k\text{-g}}$ (where k=1 or 2) , $v_{\mathrm{co}k\text{-}1}\ w_{\mathrm{co}k\text{-}1},..,\ v_{\mathrm{co}k\text{-}h}w_{\mathrm{co}k\text{-}h}$ (where k=2 or 3 and k in A $\neq$ k in B)) from the previous step by using the algorithm as shown in Fig. 3.

$$\begin{aligned} &Causality\,\mathrm{Re}\,lationClass = \underset{class\in\mathrm{Class}}{\arg\max}\, P(class \mid a_1, a_2, .., a_g, b_1, b_2, .., b_h) \\ &= \underset{class\in\mathrm{Class}}{\arg\max}\, P(a_1 \mid class) P(a_2 \mid class) P...(a_g \mid class) P(b_1 \mid class) P(b_2 \mid class) \\ &\qquad ..P(b_h \mid class) P(class) \end{aligned} \qquad (2)$$

where $a_1, a_2, .., a_g \in \mathrm{WC}_c \cup \mathrm{WC_{ce}}$.

$b_1, b_2, .., b_h \in \mathrm{WC}_e \cup \mathrm{WC_{ce}}$. .

$\mathrm{Class} = \{"yes","no")$

V. Evaluation and Conclusion

There are two evaluations of the proposed research, the extraction of WordCo feature vectors with the causative/effect event concepts from 800 EDUs of the extracting corpus and the determination of consecutive causality relations as the chain of causation from 500 EDUs of testing corpus. Both evaluations are based on the precisions and the recalls which are evaluated by three expert judgments with max win voting.

TABLE I: WordCo Feature Vector Size/Boundary Determination by SVM

500EDUs Downloaded Corpus from Hospital web-boards	Precision	Recall%
Causative Event Concept Vector	0.917	0.852
Effect Event Concept Vector	0.891	0.833

According to Table I, the average precision of extracting WordCo feature vectors is 0.904 with the average recall of 0.843. The reason of low recall is the causative/effect event-concept occurrences on the NP1 expressions, i.e. EDUi ('*การเต้นของหัวใจ/Heart beat*ing')/NP ('*เร็ว/rapid*')/VP ("*The heart beats rapidly*"). Moreover, the precision of determining the chain of causation is 0.9 with the recall of 0.83. The recall result of determining the chain of causation is low because there are some CR*i* expressions having the effect event-concept EDUs around a cause event vector as < *effect* event-concept EDUs><a *cause* event vector><an effect event-concept EDU> as shown in the following example.

EDU1 "*วัยรุ่นรู้สึกก้าวร้าว /The teen feels aggressive*."

EDU2 ""'*และ*[*เด็ก*]*รู้สึกกระวนกระวาย* / [*he*] *feels nervous*

EDU3 "*เมื่อมีความต้องการใช้ยา* / *when* [*he*] *need to use the narcotic*."

EDU4 "*ถ้ามีใครเข้ามาขัดขวาง/if there are someone trying to stop him*."

EDU5 "*เด็กก็จะแสดงอาการหงุดหงิด/ he will show up the fidget symptom*.''
where EDU1, EDU2, and EDU5 are an effect event vector of the cause event vector on EDU3 and EDU4.

Hence, the research contributes the methodology to determine the chain of causation for finding the root cause which is very beneficial to people on the social network to clearly understand the sequence of causes and consequences for awareness. Finally, our research can also enhance the problem-solving system of the other areas i.e. the business financial system.

REFERENCES

[1] L. Carlson, D. Marcu, and M. E. Okurowski, "Building a discourse-tagged corpus in the framework of rhetorical structure theory," *Current and New Directions in Discourse and Dialogue*, vol. 22, pp. 85-112, 2003.

[2] R. Girju, "Automatic detection of causal relations for question answering," in *Proc. the 41st Annual Meeting of the Assoc. for Computational Linguistics, Workshop on Multilingual Summarization and Question Answering-Machine Learning and Beyond*, Japan, 2003.

[3] D. S. Chang and K. S. Choi, "Causal relation extraction using cue phrase and lexical pair probabilities," in *Proc. First International Joint Conference*, Hainan Island, China, March 22-24, 2004, pp. 61-70.

[4] C. Pechsiri and R. Piriyakul, "Explanation knowledge graph construction through causality extraction from texts," *Journal of Computer Science and Technology*, vol. 25, no. 5, pp. 1055-1070, 2010.

[5] S. Zhao, T. Liu, S. Zhao, Y. Chen, and J.-Y. Nie, "Event causality extraction based on connectives analysis," *Neurocomputing*, vol. 173, pp. 1943-1950, 2016.

[6] P. Mirza and S. Tonelli, "CATENA: Causal and temporal relation extraction from NAtural language texts," in *Proc. COLING 2016*, Osaka, Japan, December 11-17, 2016, pp. 64-75.

[7] H. Sawamaru and I. Kobayashi, "An approach to extraction of causal chain among events in multiple documents," in *Proc. The 6th International Conference on Soft Computing and Intelligent Systems and the 13th International Symposium on Advanced Intelligent Systems*, Japan, Nov. 20-24, 2012, pp. 1104-1108.

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

[9] S. Sudprasert and A. Kawtrakul, "Thai word segmentation based on global and local unsupervised learning," in *Proc. National Computer Science and Engineering Conference*, 2003, pp. 1-8.

[10] H. Chanlekha and A. Kawtrakul, "Thai named entity extraction by incorporating maximum entropy model with simple heuristic information," in *Proc. First International Joint Conference*, Hainan Island, China, March 22-24, 2004, pp. 1-7.

[11] J. Chareonsuk, T. Sukvakree, and A. Kawtrakul, "Elementary discourse unit segmentation for Thai using discourse cue and syntactic information," in *Proc. National Computer Science and Engineering Conference*, 2005, pp. 85-90.

[12] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK., 2000.

Chaveevan Pechsiri holds a master's degree in computer science from Mississippi State University, USA, and a doctoral degree in computer engineering from Kasetsart University, Thailand. She is currently an associate professor at Dhurakijpundit University, Thailand. Her general research interest is in natural language processing.

Renu Sukharomana is a professor and director of the Center for Children's Social Protection, Thailand research Fund, in Bangkok, Thailand. She holds a PhD in agricultural economics from the University of Nebraska and is currently engaged in multidisciplinary research from a wide range of topics in science and economics.