Causal Web Determination from Texts

Chaveevan Pechsiri[1], Narongdech Keeratipranon[1], and Intaka Piriyakul[2]
[1] College of Innovative Technology and Engineering, Dhurakij Pundit University, Bangkok, Thailand
[2] Faculty of Social Sciences, Srinakharinwirot University, Bangkok, Thailand
Email: {chaveevan.pec, narongdech.ken}@dpu.ac.th, intaka@hotmail.com

***Abstract*—The research aim is to determine a causal web from downloaded guru web-board documents. The causal web which benefits a diagnosis service assistant of a problem-solving system consists of several cause-effect pair sequences where each cause-effect pair has a cause-effect relation and the last cause-effect pair of each cause-effect pair sequence has the same effect concept. Each causative/effect concept is expressed by an elementary discourse unit or a simple sentence. The research has three problems; how to determine the cause-effect pair with an overlap problem between a causative-verb concept set and an effect-verb concept set, how to determine cause-effect pair sequences including causative/effect boundary determination, and how to determine the causal web on the extracted cause-effect pair sequences without redundant sequences. We use a word co-occurrence to represent a sentence's event/state with a causative/effect concept. We then propose using a self-Cartesian product on a collected word co-occurrence set and Naïve Bayes including categorized verb groups to extract each cause-effect pair sequence including the boundary determination without the verb-concept-overlap influence. And we use a dynamic template matching technique to determine the causal web without the redundancy. The research result has a high percentage correctness of the causal web determination.**

***Index Terms*—word co-occurrence, elementary discourse unit, template matching**

I. Introduction

Determination of a causal web from downloaded documents on the guru web-boards is a challenge where a causal web is a metaphor that emphasizes the interconnectedness of casual components in a population including direct causes and indirect causes comprising causal webs [1]. The causal web expression of a certain domain benefits a diagnosis service assistant of a problem -solving system in that domain. The aim of this research is then to determine the causal web from the downloaded guru documents, particularly on the car-problem guru web-boards (i.e. AskGuru [2], MThai [3], and etc.). Regard to the causal web explanation by [1], the causal web of our research consists of several cause-effect pair sequences expressed on the documents where each cause-effect pair has a cause-effect relation between one or more event/state expressions with causative concepts and one or more event/state expressions with effect concepts. The event/state expression with the causative/effect concept is based on an Elementary Discourse Unit (EDU) defined as a simple sentence / a clause by [4]. Moreover, the last cause-effect pairs from all cause-effect pair sequences have events/states with the same effect concept to construct the causal web as shown in Fig. 1 where a node represents an event/state expressed by EDU with a causative/effect concept and a link represents a cause-effect relation of a cause-effect pair from a cause node to an effect node pointed by an arrow. Fig. 1 also shows the alternative causes for assisting automatic human reasoning in diagnosis of the car problems/ symptoms. Each EDU (see Fig. 2) is expressed by the general Thai linguistic expression after stemming words and eliminating stop words where NP1 and NP2 are noun phrases, VP is a verb phrase, V is a non-terminal verb expression, $\text{Verb}_{\text{strong}}$ is a strong-verb concept set, $\text{Verb}_{\text{weak}}$ is a weak-verb concept set needed more information, Adv is an adverb concept set, Noun is a noun concept set, and Adj is an adjective concept set. All concepts of these concept sets are based on WordNet [5] and Thai Encyclopedia [6] after translating from Thai to English by Lexitron [7]. For example: Example1 (see Fig. 3) shows the cause-effect relation occurs between an EDU1 expression with an effect concept and both EDU2 and EDU3 expressions with causative concepts as shown in the following cause-effect pair expression.

EDU2 $\wedge$ EDU3: Cause → EDU1: Effect

Example2 (see Fig. 4 where a [..] symbol means ellipsis) contains the cause-effect pair sequence as shown in the following where EDU2, EDU3 and EDU4 express the causative concepts and also the effect concepts.

EDU1: Cause → EDU2: Effect
EDU2: Cause → EDU3: Effect
EDU3: Cause → EDU4: Effect
EDU4: Cause → EDU5: Effect

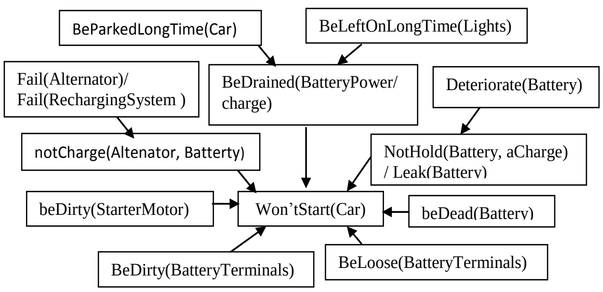

Figure 1. Causal web of car problems, i.e. "A car won't start."

Manuscript received September 26, 2019; revised April 7, 2020.

doi: 10.12720/jait.11.2.64-70

EDU → NP1 VP | VP
NP1 → pronoun | Noun | Noun modify
NP2 → Noun | Noun modify
VP → V NP2 | V | V adv | V AdvPhrase
V → Verb_{weak} Word | Verb_{strong}
modify → Adj| Adj modify| Noun modify
Word = Noun ∪ Verb_{strong} ∪ Adv
Verb_{weak}→ {‘เป็น/*be*’, ‘มี/*have*’, ‘ใช้/*use*’, ‘นำ/*take*’, ‘เอา/*get*’}

Verb_{strong}→{‘สตาร์ทไม่ติด/*fail-to-start*’ ‘เสื่อม,เสื่อมสภาพ,สึกหรอ/*deteriorate*’ ‘ปิดไม่สนิท/*be-not-tightly-closed*’ ‘ไม่สามารถเก็บ/*not-hold*’ ‘หลวม/*loose*’ ‘ไม่สว่าง/*not lighten*’ ‘ลืม/*forget*’ ‘เปิด/*turn-on*’ ‘ไหล/*leak-out*’ ‘จอด/*park*’ ‘ชำรุด/*be-worn-out*’ ‘หลุด/*drop-off*’ ‘หมดอายุ/*expire*’ ‘ปล่อยให้/*leave*’ ‘โก่งตัว/*bend*’ ‘ลื่น/*slip*’ ‘ติด/*light*’ ‘ไม่สามารถกวาด/*not-sweep*’ ‘ไม่สามารถกั้น/*not-block*’ ‘ไม่สามารถปั่น/*not-spin*’ ‘ไม่สามารถเลี้ยง/*not-support*’ ‘ดับ/*stop*’ ‘ขาด/*lack-of*’ ‘กัด/*bite*’...}

Noun → { ‘ขั้วแบตเตอรี่/*terminals*’ ‘ไดชาร์จ/*alternator*’ ‘มอเตอร์สตาร์ท/*starter-motor*’ ‘แบตเตอรี่/*battery*’ ‘เกียร์/*gear*’ ‘ระบบไฟ/*electric-system*’ ‘ประจุไฟฟ้า/*electric-charge*’ ‘วาล์วน้ำ/*valve*’ ‘ท่อไอเสีย/*intake*’ ‘น้ำมันเครื่อง/*engine-oil*’ ‘เสื้อสูบ/*cylinder-block*’ ‘ลูกสูบ/*piston*’ ‘ตัวแหวนลูกสูบ/*piston-ring*’ ‘เครื่องยนต์/*engine*’ ‘ฝาสูบ/*cylinder-head*’ ‘ซีลฝาสูบ/*cylinder-head-seal*’ ‘ซีลก้านวาวล์ไอดี/*intake-valve-stem-seal*’ ‘คลัตช์/*clutch*’ ‘แผงหน้าปัด/*panel*’ ‘กระแสไฟ/*current*’ ‘สายไฟ/*wire*’ ‘เซ็นเซอร์เพลาลูกเบี้ยว/*sensor-camshaft*' 'ห้องเผาไหม้/*combustion chamber*’ ‘ห้องโดยสาร/*cabin*’ ‘น้ำหม้อน้ำ/*radiator*’ ‘ดอกยาง/*tread*’ ‘ควัน/*smoke*' 'หนู/*rat*' ‘อาการ/*symptom*’‘เสียง/*noise*’ ‘เวลา/*time*’,…}
Adv → {‘นาน/*long*’ ‘ยาก/*hard*’ ‘หมด/*clearly*’ …} ;
Adj→ {‘ดัง/*loud*’ ‘หลวม/*loose*’ ‘สกปรก/*dirty*’ ‘แข็ง/*hard*’,…}

Figure 2. Linguistic pattern after stemming words and stop word removal

Example1:
... **EDU1**: “*รถสตาร์ทไม่ติด*” (**“*A car won’t start.*”**)
(“(*รถ/car*)/NP1 ((*สตาร์ท/start ไม่ติด/not*))/VP”)
EDU2: “*เพราะแบตเตอรี่เสื่อม*” (**“*because a battery deteriorates.*”**)
(“*เพราะ/because* (*แบตเตอรี่/battery*)/NP1 ((*เสื่อม/deteriorate*)/V)/VP”)
EDU3: “*และ[แบต] หมดอายุ*” (**“And [*the battery*] *expires.*”**) ...
(“*และ/And* [(*แบต/battery*)/NP1] ((*หมดอายุ/expire*)/V)/VP”) ...
Where a [..] symbol means ellipsis.

Figure 3. Show cause-effect relation between 2 causative-concept EDUs and 1 effect-concept EDU

Example2:
EDU1: “*เมื่อไส้กรองอากาศหมดสภาพ*” (“***When the air filter is worn out.***”)
(“*เมื่อ/When* (*ไส้กรองอากาศ/air filter*)/NP1 ((*หมดสภาพ/be worn out*)/V)/ VP”)
EDU2: “*ทำให้ [ไส้กรองอากาศ] เกิดอาการอุดตัน*”
(“***Cause*** [***the air filter***] ***to have obstructed symptom***”)
(“((***ทำให้เกิด/cause***)/ConjVerb [(*ไส้กรองอากาศ/ air filter*)/NP1]
(*เกิด/occur with*)/V *อาการ/symptom อุดตัน/obstructed*)/VP”)
EDU3: “*ส่งผลให้อากาศเข้ากระบอกสูบน้อย*”
(“***Cause the air to enter the cylinder slightly***”)
(“((*ส่งผลให้/cause*)/ConjVerb (*อากาศ/air*)/NP1
((*เข้า/enter*)/V *กระบอกสูบ/cylinder น้อย/slightly*)/VP”)
EDU4: “*ทำให้การเผาไหม้ในห้องเครื่องยนต์ไม่สมบูรณ์*”
(“***Cause the combustion in the engine room to be incomplete***”)
(“((*ทำให้/cause*)/ConjVerb (*การเผาไหม้/combustion ห้อง/room เครื่องยนต์/engine*)/NP1 ((*ไม่สมบูรณ์/be not complete*)/V)/VP”)
EDU5: “*เครื่องยนต์กำลังตก*” (“***The engine power drops***.”)
(“(*เครื่องยนต์/engine กำลัง/power*)/NP1 (*ตก/drop*)/V)/VP”)

Figure 4. Show a sequence of cause-effect EDU pairs

There are several techniques, [8]-[14], having been applied for determining the cause-effect/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 a cause-effect pair from an EDU pair having a cause-effect relation with the overlap problem between a causative-verb concept set (V_{cause}) and an effect-verb concept set (V_{effect}). Where $V_{strong} = V_{cause} \cup V_{effect}$. The second problem is how to determine the cause-effect pair sequences including the EDU boundary determination with either the causative-event/state concepts or the effect-event/state concepts. And the third problem is how to determine the causal web from the extracted cause-effect pair sequences with the problem of redundant cause-effect pair sequences. Regarding all of these problems, we develop a framework which combines machine learning and the linguistic phenomena to learn the several EDUs having the cause-effect relation on the downloaded documents. Therefore, we collect a word co-occurrence (called wrdCo) pattern from each EDU expression after stemming words and eliminating stop words into a wrdCo set (called WC). Each wrdCo expression/element has the pattern as shown in the following to represent an EDU occurrence with a causative-event/state concept or an effect-event/state concept where each causative/effect concept is expressed by VP of an agent expressed by NP1.

$$\text{wrdCo Pattern: } w_1 + w_2 + w_3 + w_4$$

where w_1 is a head noun of NP1 and has a null value if NP1 is ellipsis; w_2 , w_3, and w_4 exist on VP having $w_2 \in (V_{strong} \cup V_{weak})$; w_3 and w_4 are a word sequence right after w_2 and have a null value if they don’t exist. And $w_3, w_4 \in$ Word.

We then propose using a self-Cartesian product on the collected WC with concepts along with the Naïve Bayes (NB) learning technique [15] to determine WCP_{ce} which is a wrdCo ordered pair set having the cause-effect relation as follow: $WCP_{ce} = \{wcpair_1, wcpair_2, .., wcpair_{last}\}$; and each WCP_{ce} element ($wcpair_i$; where i=1,2,..,*last*) has the cause-effect relation between two wrdCo expressions, one wrdCo expression with a causative-event/state concept and the other one with an effect-event/state concept. We apply an experimental Event Rate (ER) [16] between two related event/state-concept occurrences for the verb categorization into verb groups/sets (see Section III.*A*). WCP_{ce} and categorized verb groups are used to identify each EDU pair with the cause-effect relation without the influence of the overlap between V_{cause} and V_{effect} ($V_{cause} \cap V_{effect} \neq \varnothing$). We then extract the wrdCo pair sequence as the cause-effect pair sequence from the documents by using WCP_{ce} including the categorized verb groups to solve the wrdCo/EDU boundary with the causative concept or the effect concept (see Section III.*B*). And we apply a dynamic template matching the last cause-effect pair of each extracted cause-effect pair sequence to discard the redundancy problem (see Section III.*C*).

Our research is organized into 5 sections. In Section II, Related Works are summarized. Problems in determining the causal web from texts are described in section III and Section IV shows a framework of Causal Web Determination from Texts. In Section V, we evaluate and conclude our model.

II. Related Works

Several strategies, [8]-[14], have been proposed to determine the cause-effect relation from texts without considering the cause-effect pair sequence except [13]. Reference [11] applied Integer Linear Programming to learn the causal relation on a sentence from annotated verb-noun pairs based on FrameNet, WordNet and linguistic features. Reference [8] identified the causal relation between two adjacent sentences by using Support Vector Machine to learn several features from the two sentences as causal volition, the verb class, verbal semantic attributes, the connective marker, and modality. Reference [10] determined a predicate pair, which is an event pair from two sentences, having a causality by measuring a cause-effect association based the point wise mutual information with minimally supervised approach. Reference [9] applied verb-pair rules resulted by machine learning techniques to extract the causality from several EDUs including the cause/effect EDU boundary determination without the cause-effect pair sequence consideration. Reference [12] proposed the Restricted Hidden Naïve Bayes model to the lexico syntactic pattern on a sentence to learn and extract the causality with the contextual, syntactic, positional, and connective features from the English documents. Reference [14] extracted the causal relation within one or two sentences by using the linguistic rules based along with Bayesian inference to reduce the number of pairs produced by ambiguous patterns. Reference [13] applied the Granger causality model with features, i.e. N-words, topics, sentiments and etc., to detect cause - effect relationships from text for a time series and also used a neural reasoning algorithm to construct chain of cause and effect pairs as an explanation result with 57% accuracy.

However, most of the previous works on the cause-effect relation are based on event/state expressed by either NP or VP features mostly existing in one/two sentences without considering the cause-effect pair sequences enclosed in the causal web. However, there are few works on determining cause-effect pair sequences as causal pathways.

III. Problems of Determining Causal Web

A. How to Determine a Cause-Effect Pair Having V_{cause} *and* V_{effect} *Overlap*

To determine the cause-effect pair having the cause-effect relation among several EDUs, we apply the wrdCo expression as mention in section I to represent an EDU's event/state concept along with NP1 as an EDU's agent. We use V_{cause} and V_{effect} collected from an annotated learning corpus to identify the EDU occurrence with the causative-event/state concept or the effect-event/state concept if $V_{cause} \cap V_{effect} = \varnothing$. However, there is the V_{cause} and V_{effect} overlap problem in our documents as shown in Example2 having $V_{cause} \cap V_{effect} \neq \varnothing$ on EDU2, EDU3 and EDU4. Therefore, the categorized verb groups and WCP_{ce} including the String Matching technique are integrated to identify a wrdCo pair of an EDU pair as the cause-effect pair having the cause-effect relation from the documents. Where WCP_{ce} is determined by the NB-learning probabilities of wrdCo concept pairs with the CauseEffectRelation class from the annotated learning corpus to the self-Cartesian product of the collected wrdCo set or WC including concept expressions after stemming words and eliminating stop words from the testing corpus. We apply ER to measure the frequencies of the v_s occurrences ($v_s \in V_{strong}$) and the ($v_w + wd$) or (v_w wd) occurrences ($v_w \in V_{weak}$ and $wd \in$Word on the linguistic pattern) as causative-event/state concepts and/or effect-event/state concepts on the annotated learning corpus for the verb categorization into three verb groups/sets, a cause group (VC), an cause/effect group (VCE), and an effect group (VE) as follow.

$$\text{ER-of-}v_{s\text{-}c} = \mathit{theNumberOf}\, v_{s\text{-}c} / (\mathit{theNumberOf}\, v_{s\text{-}c} + \mathit{theNumberOf}\, v_{s\text{-}e}) \quad (1)$$

$$\text{ER-of-}v_{s\text{-}e} = \mathit{theNumberOf}\, v_{s\text{-}e} / (\mathit{theNumberOf}\, v_{s\text{-}c} + \mathit{theNumberOf}\, v_{s\text{-}e}) \quad (2)$$

where $v_{s\text{-}c}$ is v_s with a causative-event/state concept; $v_{s\text{-}e}$ is v_s with an effect-event/state concept;

$$\text{ER-of-}v_{w\text{-}c}wd = \mathit{theNumberOf}\, v_{w\text{-}c}wd / (\mathit{theNumberOf}\, v_{w\text{-}c}wd + \mathit{theNumberOf}\, v_{w\text{-}e}wd) \quad (3)$$

$$\text{ER-of-}v_{w\text{-}e}wd = \mathit{theNumberOf}\, v_{w\text{-}e}wd / (\mathit{theNumberOf}\, v_{w\text{-}c}wd + \mathit{theNumberOf}\, v_{w\text{-}e}wd) \quad (4)$$

where $v_{w\text{-}c}wd$ is $v_w + wd$ with a causative-event/state concept; $v_{w\text{-}e}wd$ is $v_w + wd$ with an effect-event/state concept.

Equation (1)-(4), the verb expressions can be categorized by their ER values into VC (if ER-of-$v_{s\text{-}c}$ or ER-of-$v_{w\text{-}c}wd \geq 0.9$), VE (if ER-of-$v_{s\text{-}e}$ or ER-of-$v_{w\text{-}e}wd \geq 0.9$), or otherwise VCE.

B. How to Determine Cause-Effect Pair Sequences Including Boundary Determination

The cause-effect pair sequence sometimes contains an EDU boundary with the causative concept as shown in Fig. 5 as Example3 having EDU1 and EDU2 as the causes of EDU3, and also an EDU boundary with the effect concept as shown in Fig. 5 as Example4 having EDU4, EDU5, and EDU6 as the effects of EDU3.

Example3:

... **EDU1**: "[*เรา*] *บางครั้งจอดรถ*" **("*[We] sometimes park a car.*")**

("[(*เรา*/we)/NP1] ((*จอด*/*park*)/V *รถ*/*car*)/VP")

EDU2: "*และ* [*เรา*] *เผลอเปิดไฟค้างไว้ข้ามคืน*"

("*and [we] carelessly leave a light on overnight.*")

("*และ*/*and* [(*เรา*/*we*)/NP1] ((*เผลอ*/*be carless* *เปิด*/*leave*)/V *ไฟค้างไว้* /*light on* *ข้ามคืน*/*overnight*)/VP")

EDU3: "*จนกระทั่งแบตเตอรี่หมด*" **("*until the battery's power drains.*")**

("*จนกระทั่ง*/*until* (*แบตเตอรี่*/*battery power*)/NP1 ((*หมด*/*drain*))/V)/VP")

EDU4: "*ทำให้รถสตาร์ทไม่ติด*" **("*Cause the car to fail of starting.*")** ...

("(*ทำให้*/*cause*)/ConjVerb (*รถ*/*car*)/NP1 (*สตาร์ทไม่ติด*/*will not start*)/ V)/VP") …

Example4:

... **EDU1**: "[*เรา*] *เหยียบเบรกตลอดเวลาลงจากภูเขา*"

("*[We] pedal the brakes all the time down from the mountain.*")

("[(*เรา*/we)/NP1] ((*เหยียบ*/*pedal*)/V *เบรก*/*brake* *ตลอดเวลา*/*all time* *ลง*/*down* *ภูเขา*/*mountain*)/VP")

EDU2: "*ผ้าเบรกเสียดสีกับจานเบรกนานๆ*"

("*Brake pads rub with the brakes for a long time.*")

("*ผ้าเบรก*/*Brake pad*)/NP1 ((*เสียดสี*/*rub*)/V *จานเบรก*/*brake* *นานๆ*/*for a long time*)/VP")

EDU3:"[*จานเบรก*]*จึงเกิดความร้อนสูง*"

("*[The brakes] then have high heat.*")

("[(*จานเบรก*/*brake*)/NP1] ((*เกิด*/*occur*)/V *ความร้อน*/*heat* *สูง*/*high*)/VP")

EDU4: "*ทำให้ผ้าเบรกไหม้*" **("*Cause the brake pads to be burnt.*")**

("(*ทำให้*/*cause*)/ConjVerb (*ผ้าเบรก*/*brake pad*)/NP1 (*ไหม้*/*burn*)/V)/VP")

EDU5: "*และ*[*จานเบรก*]*มีกลิ่นไหม้*"

("*and [the brake pads]have a burning smell.*") ...

("*และ*/*and* [(*จานเบรก*/*brake pad*)/NP1] ((*มี*/*have* *กลิ่น*/*smell*)/V *ไหม้*/*burn*)/VP") …

EDU6: "*บางครั้งน้ำมันเบรกเดือดกลายเป็นไอ*"

("*The brake oil sometimes boils into vapor.*") ...

("(*น้ำมันเบรก*/*brake oil*)/NP1 ((*เดือด*/*boil*)/V *กลายเป็น*/*to become* *ไอ*/*vapor*)/VP") …

Where a [..] symbol means ellipsis.

Figure 5. Show a causative-EDU boundary on Example3 and an effect-EDU boundary on Example 4

Example3 on Fig. 5 shows a cause-effect pair sequence having EDU1 and EDU2 as a causative-EDU boundary as follow:

(EDU1$\wedge$EDU2): Cause $\rightarrow$ EDU3: Effect
EDU3: Cause $\rightarrow$ EDU4: Effect

Example4 on Fig. 5 also shows a cause-effect pair sequence having EDU4, EDU5, and EDU6 as an effect-EDU boundary as follow:

EDU1: Cause $\rightarrow$ EDU2: Effect
EDU2: Cause $\rightarrow$ EDU3: Effect
EDU3: Cause $\rightarrow$ (EDU4 $\wedge$ EDU5$\wedge$EDU6): Effect

Therefore, we solve the event/state boundary with either causative concept or effect concept by using the categorized verb groups by ER along with WCP_{ce} through String Matching.

C. *How to Determine Causal Web with Redundancy Problem*

There is variety of the extracted cause-effect pair sequences with some redundant sequences from the documents. Therefore, we apply a dynamic template based on concept expressions for String Matching in the last cause-effect pair of each extracted cause-effect pair sequence to eliminate the redundant cause-effect pair sequences which match patterns hold by the template. Whereas, a simple template (which is a concept pattern of the last cause-effect pairs of the extracted cause-effect pair sequences) consists of a wrdCo ordered pair with concept expression as follow:

SimpleTemplate

($wrdCo_{cause}$: *ConceptExpression*)
($wrdCo_{effect}$: *ConceptExpression*)

where $wrdCo_{cause}$ is a word co-occurrence with a causative- event/state concept; $wrdCo_{effect}$ is a word co-occurrence with an effect-event/state concept; $wrdCo_{effect}.w_2$ /$wrdCo_{effect}.w_2w_3 \in VE$; and $wrdCo_{cause}.w_2$ / $wrdCo_{cause}.w_2w_3 \in VC \cup VCE$.

The simple template will discard only the cause-effect pair sequence having the last cause-effect pair being the same as the template pattern. With regard to several extracted cause-effect pair sequences, the simple template pattern can be automatically adjusted to contain several wrdCo ordered pairs having only the same $wrdCo_{effect}$: *ConceptExpression* with different $wrdCo_{cause}$: *ConceptExpression* as follow:

DynamicTemplate

($wrdCo_{cause\text{-}i}$: *ConceptExpression*)
($wrdCo_{effect}$: *ConceptExpression*)

where i=1,2,..n; n is an integer; $wrdCo_{effect}.w_2$/$wrdCo_{effect}.w_2w_3 \in VE$; and $wrdCo_{cause\text{-}i}.w_2$/$wrdCo_{cause\text{-}i}.w_2w_3 \in VC \cup VCE$.

IV. A Framework of Determining Causal Web

The causal web determination consists of six steps; Corpus Preparation, Categorized Verb Group Determination, Learning wrdCo Pair with CauseEffectRelation, Extraction of wrdCo ordered pairs having CauseEffectRelation, Extraction of Cause-Effect Pair Sequences, and Causal Web Determination (see Fig. 6.)

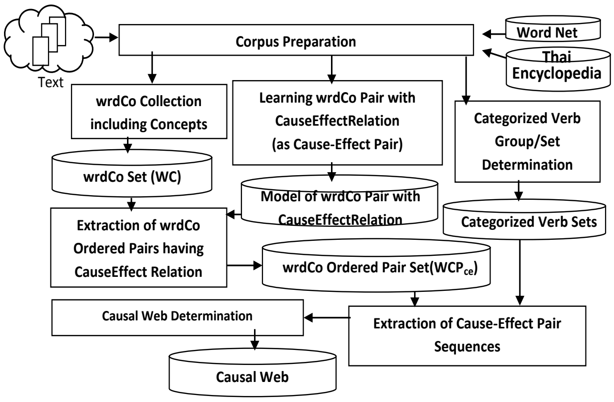

Figure 6. System overview

A. *Corpus Preparation*

“*ทำไมรถสตาร์ทไม่ติด*:
[เรา] บางครั้งจอดรถ[EDU1] *และ [เรา] เผลอเปิดไฟค้างไว้ข้ามคืน*[EDU2] *จนกระทั่งแบตเตอรี่หมด*[EDU3]
รถสตาร์ทไม่ติด[EDU4] ..”
“*Why a car won’t start*:
[*We*] *sometimes park a car*.[EDU1] *and* [*we*] *carelessly leave a light on overnight*.[EDU2]
until the battery’s power drains[EDU3] ***The car won’t start.***[EDU4] **………..”**
<Topic_name concept= Why a car won’t start >*ทำไมรถสตาร์ทไม่ติด*:</Topic_name>…
<**EDU1** CErelId=1><**wrdCo concept**=‘**park (ϕ, car)**’>
<NP1 concept= we/human><w*1*: headNoun concept=null>ϕ</w*1*></NP1>
<**VP Type=cause :**CErelId=1 **; Group=VC**>
<w*2*: setType=‘verb-strong as Causative-verb’ ; concept=‘park/place’>จอด</w*1*>
<w*3*: setType=‘Noun’ ; concept=‘motor vehicle’>รถ </w*2*>
<w*4*: concept=null>null</w*4*></VP></wrdCo></EDU1>
<**EDU2** CErelId=1><**wrdCo concept=‘Leave (ϕ,Electrical Device on)’**>
<Conj concept=and>และ</Conj>
<NP1 concept= we/human><w*1*: headNoun concept=null>ϕ</w*1*></NP1>
<**VP Type=cause :**CErelId=1 **; Group=VC**>
<w*2*: setType=‘verb-strong as Causative-verb’ ; concept=‘be careless’>เผลอ</w*2*>
<w*3*: setType=‘verb-strong as Causative-verb’ ; concept=‘leave open’>*เปิด*</w*3*>
<w*4*: setType=‘Noun’; concept=‘electrical device’>ไฟ</w*4*></VP ></wrdCo>
</EDU2>
<**EDU3** CErelId=1, 2><**wrdCo concept=‘drain(Battery Power)’**>
<Conj concept=until>จนกระทั่ง</Conj>
<NP1 concept=‘battery power’><w*1*: headNoun concept=‘battery/battery power’>
แบตเตอรี่</w*1*></NP1>
<**VP Type=effect :**CErelId=1 , **cause** :CErelId=2 **; Group=VCE**>
<w*2*: setType=‘verb-strong as Effect-verb’ ; concept=‘drain’>หมด</ w*2*>
<w*3*: concept=null>null</w*3*><w*4*: concept=null>null</w*4*></VP></wrdCo>
</EDU3>
<**EDU4** CErelId=2><**wrdCo concept=‘failToStart(Motor Vehicle)’**>
< NP1 concept=‘motor vehicle’><w*1*: headNoun concept=‘motor vehicle’>รถ</w*1*>
</NP1>
<**VP Type=effect :**CErelId=2 **; Group=VE**>
<w*2*: setType=‘verb-strong as Effect-verb’ ; concept=‘fail to start’>สตาร์ทไม่ติด
</w*2*>
<w*3*: concept=null>null</w*3*><w*4*: concept=null>null</w*4*></VP></wrdCo>
</EDU4>**……**

The wrdCo tag is the word co-occurrence pattern tag of each wrdCo expression within each EDU tag which consists of a NP1/ Noun Phrase tag and a VP/Verb Phrase tag . The w*i* tag is the word-*i* tag where *i*=1,2,3,4. The **[..]** or ϕ **symbol** means **ellipsis** , i.e. a noun ellipsis as Zero Anaphora.

Figure 7. Corpus preparation

This step is to prepare an EDU corpus of car-problem documents downloaded from the car guru web-boards. The step involves using Thai word segmentation tools [17], Named Entity recognition [18], and EDU Segmentation [19] to provide the 2500 EDUs’ corpus. The corpus included stemming words and the stop word removal is separated into 3 parts; the first 1000-EDUs’ part as the annotated learning corpus for categorizing verb groups by ER and learning the wrdCo pairs with CauseEffectRelation or NonCausaeEffectRelation. The second 1000-EDUs’part is for testing as extracting and collecting of the WCP_{ce} set. The third 500-EDUs’part is for testing as extracting the cause-effect pair sequences used for determining the causal web. We then semi-automatically annotate the wrdCo pattern with concepts along with the categorized verb groups as VC, VCE, and

VE on each EDU of the learning corpus (see Fig. 7). All wrdCo concepts are referred to WordNet [5] after the Thai-English translation by Lexitron [7].

B. Categorized Verb Group Determination

Regard to the annotated learning corpus, we collect each VP tag containing a group property as Group=VC, Group=VCE, or Group=VE for the ER determination. Equation (1)-(4) on Section III.*A*, the verb expression can be categorized by ER value into three verb groups/sets, VC, VCE, and VE (see Fig. 8), with regard to the following ER-rule where *wd* is a word occurrence annotated by the w3 tag (see Fig. 7).

ER-rule:

If ER-of-$v_{s\text{-}c} \geq 0.9$ or ER-of-$(v_{w\text{-}c}\ wd) \geq 0.9$ then
$v_{s\text{-}c} \in VC$ or $(v_{w\text{-}c}\ wd) \in VC$ respectively
ElseIf ER-of-$v_{s\text{-}e} \geq 0.9$ or ER-of-$(v_{w\text{-}e} wd) \geq 0.9$ then
$v_{s\text{-}e} \in VE$ or $(v_{w\text{-}e} wd) \in VE$ respectively
Else $v_s \in VCE$ or $(v_w\ wd) \in VCE$ respectively.

VC={‘เสื่อม,เสอมสภาพ,ส อ/*deteriorate*’‘ชารด,เสย/*be-worn-out*’ ‘ลม/ *forget*’ ‘เปด/*turn-on*’‘ปลอยให/*leave*’ ‘จอด/*park*’ ‘หมดอาย/ *expire*’ ‘รว/*leak-out*’ ‘กัด/*bite*’ ‘ไมม/*notHave*+นามัน/*gas*’ …}
VCE={‘เกด/*occur*’‘หมด/*exhaust*’ ‘ไ บ/*not-hold*’ ‘ลน/*slip*’ ‘ไมสามารถจด ประกายไฟ/*not-ignite-spark*’ ‘ไ น/*not-spin*’ ‘ไมสามารถเลยง/*not-support*’ ‘เสยดส/*friction*’ ‘เส ดส/*rub*’ ‘โ งตัว/*bend*’ ‘ปดไม ท/*untightly-close*’ ‘มี/*have*+ อาการ/*symptom*’… }
VE={‘ส ด/*fail-to-start*’‘ดับ/*stop*’‘มี/*have*+ ควัน/*smoke*’ ‘มี/*have*+ กลน/*smell*’…. }

Figure 8. Categorized verb group: VC, VCE, and VE

C. Learning wrdCo Pairs with CauseEffectRelation

This step is the NB learning [15] the feature set of wrdCo-concept pairs with the CauseEffectRelation/ nonCauseEffectRelation class on several two adjacent EDUs with the CErelId/nonCErelId annotation from the corpus preparation step (Section IV.*A*) after stemming words and eliminating stop words. The learning results of this step by using Weka [20] are the probabilities of the wrdCo pairs with concept expressions from the annotated learning corpus as shown in Table I.

TABLE I. PROBABILITIES OF WRDCO PAIRS WITH CONCEPTS

wrdCo Pair with Concept Expression: Causative*wrdCo*Concept – Effect*wrdCo*Concept	CauseEffect Relation	Non CauseEffect Relation
Park(ϕ,carLeaveBehind) - FailToStart(car)	0.02367942	0.05732484
Leave(ϕ,lightOn) - Drain(battPower)	0.02367942	0.00636943
Loose(battTerminal) - NotStartCar(electricSystem)	0.00884956	0.00429185
BeWrongPosition(gear) - FailToStart(car)	0.00910747	0.00636943
………………….	………	…………

D. Extraction of wrdCo Ordered Pairs Having CauseEffectRelation

We collect all wrdCo expressions from the 1000EDU testing corpus including concepts from the corpus preparation into $WC_{concept}$ (which is WC including concept expressions) used for the self-Cartesian product as $WC_1 \times WC_2$ where WC_1 is $WC_{concept}$ used as the causative-event/state concept set; WC_2 is $WC_{concept}$ used as the effect-event/state concept set. The result of the self-Cartesian product on $WC_{concept}$ is WCP_{ce} which is the wrdCo order pair set (including the concept expression of each wrdCo order pair) having the cause-effect relation through NB as shown in (5) along with the probabilities of the wrdCo pair features with CauseEffectRelation and nonCauseEffectRelation from the learning corpus of section IV.*C*.

$$\begin{aligned} wcOrdpair\_Class &= \underset{class \in Class}{\arg\max}\ P(class \mid wcpair_k). \\ &= \underset{class \in Class}{\arg\max}\ P(wcpair_k \mid class)P(class). \end{aligned} \tag{5}$$

where $wcOrdpair\_Class$ is a wrdCo ordered pair class of $wcpair_k$;
$wcpair_k \in WC_1 \times WC_2$; $k = 1,2,..num$;
num is the number of $WC_1 \times WC_2$ elements;
$Class = \{'CauseEffect\ Relation', 'nonCauseEffect\ Relationt'\}$
If $wcOrdpair\_Class$ of $wcpair_k$ is $'CauseEffect\ Relation'$ then $wcpair_k$ is the element of WCP_{ce}.

E. Extraction of Cause-Effect Pair Sequences

This step is to extract the cause-effect pair sequences by matching *twcp* (a wrdCo concept pair from the testing corpus) to $wcp_{ce\text{-}k}$ as shown in Fig. 9 including VC, VCE, and VE to form a cause-effect pair sequence where $wcp_{ce\text{-}k} \in WCP_{ce}$; k=1,2,..*numberOfWCP*$_{ce}$*SetElements*. If match $(twcp, wcp_{ce\text{-}k})$ then $Sequence_a = Sequence_a \cup wcp_{ce\text{-}k}$ (see Fig. 9 as “Sequence$_a$.AddNewCauseEffectPair($wcp_{ce\text{-}k}$, pair1, or pair2)”).

Assume that each EDU is represented by (NP1 VP).
L is a list of EDU after stemming words and the stop word removal.
WCP_{ce} is the wrdCo ordered pair set with concepts and the cause-effect relation.
twcp is a wrdCo concept pair on a cause-effect pair sequence from the testing corpus .
twc is a wrdCo concept on a cause-effect pair sequence from the testing corpus .
wc_j is an wrdCo concept of EDU*j* ‘s verb phrase.
CAUSE-EFFECT_PAIR_SEQUENCE_EXTRACTION

```
{ j=1;g=1;flendSeq ='no'; a=1 ;ctc =0;cte=0;wc_j←∅;
  pair1←∅; pair2 ←∅;
 ArrayList<string> Sequence_a,boundaryc,boundarye
                         = new ArrayList();
while j≤ Length[L] do
{1while g≤2 ∧ j≤ Length[L] do
 {2 If  wc_j=∅ ∧ j≤ Length[L] do  /* Get wrdCo from EDUj
       {j++ ; i=1; wc_j = Get_wrdCo }
    while  wc_j.w2w3 ∈ VC ∨ wc_j.w2 ∈ VC ∧j≤ Length[L] do
       {boundaryc.add(wc_j);ctc++; j++;  wc_j=Get_wrdCo}
    while  wc_j.w2w3 ∈ VE ∨ wc_j.w2 ∈ VE ∧j≤ Length[L] do
       {boundarye.add(wc_j);cte++; j++; wc_j=Get_wrdCo}
    If  wc_j.w2w3 ∉ VCE∪VE∪VC ∨ wc_j.w2 ∉ VCE∪VE∪VC then
         {  wc_j= twc_1; g=3;flendSeq ='yes'}
                                          /*EndSequence
    while  ctc>0 ∧ (wc_j.w2w3 ∈ VCE ∨ wc_j.w2 ∈ VCE) do
                                          /*havingCauseBoundary
       { twcp = boundaryc.get(ctc)+wc_j;
         For k=1 to numberOfWCP_ceSetElements
            {If twcp match wcp_ce-k then pair1=pair1∪wcp_ce-k;
             k++}
         ctc --}               /*CauseEffectPairWithCauseBound
    If pair1<>∅ then {Sequence_a.AddNewCauseEffectPair(pair1);
                        twc_1←wc_j; g=2;wc_j←∅ ; pair1←∅ };
    while cte>0∧(wc_j.w2w3 ∈ VCE ∨ wc_j.w2 ∈ VCE) do
                                         /*havingEffectBoundary
       { twcp = wc_j +boundarye.get(cte);
         For k=1 to numberOfWCP_ceSetElements
            {If twcp match wcp_ce-k then pair2=pair2∪wcp_ce-k;
             k++}
         cte --}               /*CauseEffectPairWithEffectBound
    If pair2<>∅ then {Sequence_a.AddNewCauseEffectPair(pair2);
                         twc_1←wc_j;g=2;wc_j←∅; pair2←∅ };
    If  wc_j <> ∅ then {twc_g ←wc_j ; wc_j←∅ g++;} }2

    If flendSeq ='no'∧ twc_1 <> ∅ ∧ twc_2 <> ∅ then
       {4 For k=1 to numberOfWCP_ceSetElements
           { If (twc_1 + twc_2) ∨ (twc_2 + twc_1) match wcp_ce-k then
             Sequence_a.AddNewCauseEffectPair(wcp_ce-k);
             k++ }
         twc_1← twc_2 ; g=2 ; wc_j←∅ }4
    If flendSeq ='yes' then
                        /*End of a cause-effect pair sequence
     {5 flendSeq='no';pair1←∅;pair2←∅;g=1;wc_j←∅;a++}5
                                                /*nextSequence
 }1
}Return  Sequence_a
```

Figure 9. Cause-Effect pair sequence extraction algorithm

F. Causal Web Determination

In regard to the downloaded documents, the research concerns only two different template patterns as shown in the following.

DynamicTemplate1:
(wrdCo$_{\text{cause-}i}$:*ConceptExpression*)
(wrdCo$_{\text{effect}}$: FailToStart(car))
DynamicTemplate2:
(wrdCo$_{\text{cause-}i}$:*ConceptExpression*)
(wrdCo$_{\text{effect}}$: drop(enginePower))

We search for matching between the last wrdCo ordered pair as the last cause-effect pair of each extracted cause-effect pair sequence (Sequence$_a$; a=1,2,..,*lastSequence*) and the specified templates, DynamicTemplate1 and DynamicTemplate2, where each template consists of one wrdCo$_{\text{cause-}i}$ and one wrdCo$_{\text{effect}}$. If the exact match of either DynamicTemplete1or DynamicTemplate2 is found, the extracted cause-effect pair sequence containing the match of the last cause-effect pair is discarded. If the match occurs only the concept expression of wrdCo$_{\text{effect}}$ on either DynamicTemplate1 or DnamicTemplate2, this will result in the adjustment of either DynamicTemplate1 or DynamicTemplate2 respectively with wrdCo$_{\text{cause-}(i+1)}$ having *ConceptExpression* from the Causative*wrdCo*Concept expression of the matched Effect*wrdCo*Concept expression on the last cause-effect pair of the extracted cause-effect pair sequence.

V. EVALUATTION AND CONCLUSION

There are three evaluations of the proposed research being evaluated by three expert judgments with max win voting: the first evaluation is the extraction of the WCPce set from 1000 EDUs testing corpus. The second evaluation is the cause-effect pair sequence extraction from the 500 EDUs testing corpus and the third evaluation is the causal web determination from the extracted cause-effect pair sequences. The first and second evaluations are based on the precision and the recall with tenfold cross validation whilst the third evaluation is the percentage of correctness (see Table II).

TABLE II. RESEARCH EVALUATION RESULTS

Evaluation	Precision	Recall	%Correctness
Extraction of WCP$_{ce}$, (from 1000EDU-Testing Corpus)	0.88	0.62	-
Extraction of Cause-Effect Pair Sequences (from500EDU-TestingCorpus)	0.90	0.65	-
Causal Web Determination	-	-	85%

The reasons of having low recalls in both extracting the WCPce set and the cause-effect pair sequences from the documents are 1) the noun clause occurrence as the direct object of the complex sentence expresses the effect event concept as "that the electric system is not fully functional" in EDU2 of Example5 on Fig. 10. 2) two cause-effect pair sequences express on the documents with sharing the last cause-effect pair as shown in Example6 of Fig 10. Example6 on Fig. 10 consists of two consecutive cause-effect pair sequences as shown in the following with sharing the last cause-effect pair (EDU5) where the first cause-effect pair sequence cannot be extracted.

The first cause-effect pair sequence:	The second cause-effect pair sequence:
EDU1:Cause→EDU2:Effect	EDU3:Cause→EDU4:Effect
EDU2:Cause→EDU5:Effect	EDU4:Cause→EDU5:Effect

Example5:
EDU1: "*ถ้าหากขั้วแบตหลวม*" (**"*If the spark plugs are loose.*"**)
EDU2: "*แสดงว่าระบบไฟทำงานได้ไม่เต็มที่*"
(**"[*It*] *shows that the electric system is not fully functional*"**)

Example6:
EDU1: "*เนื่องจากหัวเทียนอาจเสีย*"
(**"*Since a spark plug may be worn out.*"**)
EDU2: "*ทำให้ [หัวเทียน] ไม่สามารถจุดประกายไฟได้*"
(**"*Cause* [*the spark plug*] *to be unable to ignite the spark.*"**)
EDU3: "*หรือปั้มติ๊กเสีย*" (**"*or a fuel pump is worn out.*"**)
EDU4: "[*ปั้มติ๊ก*] *ไม่สามารถดูดน้ำมันเข้าสู่ระบบ*"
(**"[*The fuel pump*]*cannot suck the fuel oil into the system.*"**)
EDU5: "*ทำให้เครื่องยนต์สตาร์ทไม่ติด*"
(**"*Cause the engine to fail of starting.*"**) ...
Where a [..] symbol means ellipsis.

Figure 10. Show examples of having of low recalls

Moreover, An ellipse of EDU having wrdCo$_{\text{effect}}$.w_2 or wrdCo$_{\text{effect}}$.w_2 $w_3 \in$VCE as in EDU2 of Example7 on Fig. 11 results in the redundant cause-effect pair sequences which effect to the % of correctness of the causal web determination.

Example7:
EDU1: "*แบตเตอรี่เสื่อม*"
(**"*The battery deteriorates.*"**)
[EDU2: "*ฉะนั้นตัวแบตไม่สามารถเก็บประจุไฟฟ้าไว้ได้นาน*"
(**"*Then the battery won't hold a charge for a long time.*")]**
EDU3: "*ทำให้เครื่องยนต์สตาร์ทไม่ติด*"
(**"*Cause the engine to fail of starting.*"**) ...
Where a [..] symbol means ellipsis.

Figure 11. Show a redundant cause-effect pair sequence occurrence if EDU2 is explicit

Hence, the research contributes the methodology to determine the causal web for diagnosis service assistant, particularly diagnosis common car symptoms, for car troubleshooting and maintenance to people on social networks. Finally, our research methodology can also be applied to another area, e.g. environmental business, for representing a business activity and/or a human behavior to an environmental change whereas environmental business focuses on how to operate in an eco-friendly way and still remain profitable.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Chaveevan Pechsiri is the main author of this work. Chaveevan Pechsiri is responsible for the compilation and drafting of literature, methods, results, evaluation and discussion within this work. Chaveevan Pechsiri also designed the systems overview. Narongdech Keeratipranon is responsible for conducting the machine learning technique experiments. Intaka Piriyakul is responsible for corpus preparation for this study.

REFERENCES

[1] B. B. Gerstman, *Epidemiology Kept Simple: An Introduction to Traditional and Modern Epidemiology*, 2nd ed., John Wiley&Sons, Inc., 2003, ch. 2.

[2] E-ToyotaClub. (May 2017). Ask Guru. Toyota Motor Thailand Co, ltd. [Online]. Available: https://www.e-toyotaclub.net/site/Ask-Guru/ aft/83023

[3] Charlotte. (October 2017). MThai. [Online]. Available: https://auto.mthai.com/news/tips/46509.html

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

[5] Princeton University. (2010). About Wordnet. Princeton University. [Online]. Available: https://wordnet.princeton.edu/

[6] Kanchanapisek. (2019). Thai junior encyclopedia project by Royal Command of His Majesty the King. [Online]. Available: http://kanchanapisek.or.th/kp6/sub/other_sub.php?file=encyclopedia/saranugrom.htm

[7] Metamedia Technologies. (2005). Longdo. Metamedia Technologies. [Online]. Available: https://longdo.com/

[8] T. Inui, K. Inui, and Y. Matsumoto, "Acquiring causal knowledge from text using the connective markers," *Journal of the Information Processing Society of Japan*, vol. 45, no. 3, pp. 919-933, 2004.

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

[10] Q. X. Do, Y. S. Chan, and D. Roth, "Minimally supervised event causality identification," in *Proc. Conference on Empirical Methods in Natural Language Processing*, Edinburgh, United Kingdom, 2011, pp. 294-303.

[11] M. Riaz and R. Girju, "Recognizing causality in verb-noun pairs via noun and verb semantics," in *Proc. the EACL Workshop on Computational Approaches to Causality in Language*, Gothenburg, Sweden,2014, pp.48-57.

[12] 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.

[13] D. Kang, V.Gangal, A. Lu, Z. Z. Chen, and E. Hovy, "Detecting and explaining causes from text for a time series event," *in Proc. Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 2758-2767.

[14] A. Sorgente, G. Vettigli, and F. Mele, "A hybrid approach for the automatic extraction of causal relations from text," *Studies in Computational Intelligence*, vol. 746, 2018, pp. 15-29.

[15] T. M. Mitchell, *Machine Learning*, Singapore: The McGraw-Hill Co. Inc. and MIT Press, 1997, ch. 3.

[16] M. Elbarbary, "Understanding and expressing 'Risk'," *Journal of the Saudi Heart Association*, vol. 22, no. 3, pp. 159-164, 2010.

[17] S. Sudprasert and A. Kawtrakul, "Thai word segmentation based on global and local unsupervised learning," in *Proc. NCSEC*, Thailand, 2003, pp. 1-8.

[18] H. Chanlekha and A. Kawtrakul, "Thai named entity extraction by incorporateing maximum entropy model with simple heuristic information," in *Proc. IJCNLP*, Hainan Island, China, 2004, pp. 1-7.

[19] J. Chareonsuk, T. Sukvakree, and A. Kawtrakul, "Elementary discourse unit segmentation for Thai using discourse cue and syntactic information," in *Proc. of NCSEC*, Thailand, 2005, pp. 85-90.

[20] Waikato University. (2019). Weka 3: Machine learning software in Java. Waikato University. [Online]. Available: https://www.cs.waikato. ac.nz/ml/weka/

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 Assoc. Professor at Dhurakijpundit University, Thailand. Her general research interest is in natural language processing.

Narongdech Keeratipranon holds MEng.Sc. in com- puter and communication engineering and Ph.D. degrees from Queensland University of Technology, Australia. He is currently an Assist. Prof. at Dhurakijpundit University, Thailand. His research interests are robotic and computational intelligence.

Intaka Piriyakul holds B.E. (Computer Engineering), M. Econ, and DBA degrees from Ladkrabang University, National Institute of Administration and Ramkhamhaeng University respectively Thailand. He is a lecturer at Srinakrinwiroj University Thailand. His research interest is business intelligence.