Explanation Knowledge Graph Construction Through Causality Extraction from Texts

Chaveevan Pechsiri\textsuperscript{1} and Rapepun Piriyakul\textsuperscript{2}

\textsuperscript{1}Department of Information Technology, Dhurakij Pundit University, Bangkok, Thailand
\textsuperscript{2}Department of Computer Science, Ramkumheang University, Bangkok, Thailand

E-mail: itdpu@hotmail.com; rapepunnight@yahoo.com

Received December 10, 2008; revised February 2, 2010.

Abstract Explanation knowledge expressed by a graph, especially in the graphical model, is essential to comprehend clearly all paths of effect events in causality for basic diagnosis. This research focuses on determining the effect boundary using a statistical based approach and patterns of effect events in the graph whether they are consequence or concurrence without temporal markers. All necessary causality events from texts for the graph construction are extracted on multiple clauses/EDUs (Elementary Discourse Units) which assist in determining effect-event patterns from written event sequences in documents. To extract the causality events from documents, it has to face the effect-boundary determination problems after applying verb pair rules (a causative verb and an effect verb) to identify the causality. Therefore, we propose Bayesian Network and Maximum entropy to determine the boundary of the effect EDUs. We also propose learning the effect-verb order pairs from the adjacent effect EDUs to solve the effect-event patterns for representing the extracted causality by the graph construction. The accuracy result of the explanation knowledge graph construction is 90% based on expert judgments whereas the average accuracy results from the effect boundary determination by Bayesian Network and Maximum entropy are 90% and 93%, respectively.

Keywords elementary discourse unit, explanation knowledge graph, causality boundary, effect-event pattern

1 Introduction

The explanation knowledge graph constructed automatically through the extracted causality from texts or textual data is a challenge. According to Trnkova J. and Theilmann W. (2004)\cite{1}, explanation knowledge is knowing the reason why something is the way it is. This explanation knowledge involved with the causal relations is pivoted on the distinction between causality and causation\cite{2}, whereas causality is “the relation between causes and effects” (http://wordnet.princeton.edu/) and is “a law-like relation between types of events”\cite{2}, and causation is “the actual causal relation that holds between individual events”\cite{2}. An example for the difference between the types of events and the individual events are shown below.

Example. “The aphids suck sap from corn leaves making the leaves become yellow, and shrink.”

From the above sentence the types of events consist of:

\[ \alpha = \text{objects consume}, \]
\[ \beta = \text{other objects change in color}, \]
\[ \gamma = \text{other objects change in shape}; \]

whereas the individual events consist of:

\[ a = \text{the aphids suck sap}, \]
\[ b = \text{the corn leaves become yellow}, \]
\[ c = \text{[the corn leaves] shrink}. \]

The main concept of “causality” and “causation” is that one or more things/events can cause one or more things/events to happen as the effect. This research focuses on “causality” of one causative concept causing multiple events of effect concepts for gaining the explanation knowledge being useful for the expert system in diagnosis problems and also for the question answering system (QA) as the knowledge source. The explanation knowledge will be more comprehensible for users if the knowledge is represented by the graphical model\cite{3}. Therefore, this research concerns the automatic explanation knowledge graph construction within the graphical model\cite{3} through extracting “causality” from documents to learn each event concept node effecting to the other event concept nodes with the probability notion between nodes of the graph. Likewise, two contributions of this paper are statistical based approach. First is the causality extraction giving the better results than
using the linguistic rule based approach to several domains. The second is the extracted explanation knowledge (causality) which can be represented by a graph being constructed according to one event implying another event occurrence patterns on text. Moreover, Murphy (2001)\(^3\) stated the graphical model consisted of probability theory and graph theory with the fundamental of a complex system built by combining simpler parts. In addition, there are two kinds of the graphical models: undirected graphical model (e.g., Markov networks, log-linear models) and directed graphical model (e.g., Bayesian networks, belief networks)\(^3\). Our research is based on the directed graphical model (acyclic chain, e.g., Bayesian networks by which an arc from \(A\) to \(B\) can be informally interpreted as indicating that \(A\) “causes” \(B\)).

Moreover, the causality in our research has been expressed through documents in the form of EDU (Elementary Discourse Unit) as in \([4]\) where EDU is defined by \([5]\) as a clause which is equivalent to a Thai simple sentence. This research emphasizes the main verb of each EDU. This causality expression has been classified as the inter-causal EDU and the intra-causal EDU by \([4]\). \([4]\) defined the inter-causal EDU as a causality expression of multiple EDUs in both causative unit and effect unit, for example:

**Causative Unit: \((EDU_1 + EDU_2)\)**

\(EDU_1\) “If the aphids infest rice plants.”

\(EDU_2\) “in the way that they suck sap from leaves.”

**Effect Unit: \((EDU_3 + EDU_4 + EDU_5)\)**

\(EDU_3\) “[it] will make the leaves become yellow.”

\(EDU_4\) “Then [the leaves] shrink”

\(EDU_5\) “and the rice plants will stop growing.”

(Here a symbol [..] means ellipsis.)

The intra-causal EDU is defined as a causal expression within one EDU, e.g., “โรคเกิดจากไข้สิ้น / The stunt disease is caused by virus”. However, our current research of automatically constructing explanation knowledge graphs is based on the inter-causal EDU extraction, especially that one causative concept implies multiple effect EDUs to show the consequent and concurrent effect events clearly in diagnosis. Therefore, the explanation knowledge graph has to involve temporal reasoning\(^6\) as a key role, and provide the ability to answer time-related queries over sets of events mentioned in the text whether a particular event precedes another one or an event occurred concurrently with other events. According to Mani \(et\ al.\)\(^6\), temporal reasoning is represented by an event graph for supporting inference clearly, where the nodes are events, and the edges are temporal relations with ordering.

Previous causality extraction works were based on the rule/pattern matching approach, the statistical approach, or the combination between pattern and statistics (see Section 2). The explicit cue, cue-phrase, or discourse marker, e.g., ‘because’, ‘as the result of’, ‘and’ etc., are necessary for most of the previous research to identify the causal relation or the causality. And also, most of their researches do not have the causal-boundary determination. Meanwhile, our research concerns the effect boundary determination without discourse marker because about 30% of discourse markers are implicit for the causal relation in our corpora while the boundary determination is necessary for the enhancement of our extracted explanation knowledge. To construct the explanation knowledge graph, two major problems are confronted: the explanation knowledge boundary determination (especially the effect boundary determination) and the effect-event pattern (a consequence or a concurrence) determination (see Section 3). Therefore, we propose using two different machine learning techniques, Maximum Entropy (ME)\(^7\) and Bayesian Network (BN)\(^8\), for comparing the effect boundary determination by having effect verbs (from the effect clauses or EDUs) with concepts as features of ME and BN. We also propose learning the effect-verb order pairs from the adjacent effect EDUs to solve the effect-event patterns for the graph construction.

In Section 2, related works are summarized. Problems of the effect boundary determination and the effect-event-pattern determination for the graph construction from texts are described in Section 3. Our framework of the explanation knowledge graph construction through causality extraction from textual data is shown in Section 4. Section 5 evaluates and discusses our proposed methodology and Section 6 concludes the paper.

### 2 Related Work

Several strategies have been proposed to approach the explanation knowledge graph construction through semantically extracting causality from textual data. In 1995, Khoo\(^9\) used linguistic patterns from Wall Street Journal (e.g., “[Noun-phrase: cause]’ is due to [Noun-phrase: cause]’), [Clause: cause] because [Clause: cause]’) and cues (e.g., ‘because’, ‘since’, ‘due to’) to extract causal relations within one or two adjacent sentences without any cause/effect boundary determination from documents, hence achieving 64% precision and 68% recall.

Marcu and Echihabi\(^10\) presented the unsupervised...
learning of Naïve Bayes classifier (NB) to recognize the discourse relations by using word pair probabilities between two adjacent sentences or clauses for identifying the rhetorical relation, such as “Contrast”, “Cause-Explanation Evidence” (or causal relation), “Condition”, and “Elaboration”. The result of extracting the causal relation based on two adjacent sentences without any cause/effect boundary determination from the BLIPP corpus showed 75% precision.

Inui et al.[11] proposed extracting causal knowledge from two adjacent sentences or clauses (without any cause/effect boundary determination) by using the explicit connective markers, e.g., ‘because’, ‘if...then’, with the problem of the connective marker ambiguity for classifying the causal relation types. Support Vector Machine (SVM) was used to solve their problem. Their precision is as high as 90% but the recall is as low as 30% because of unsolved anaphora.

However, the techniques from [9-11] cannot be applied to our proposed multiple EDUs for explanation knowledge extraction with graphical representation. Pechsiri and Kawtrakul (2007)[4] proposed verb-pair rules learned by two different machine learning techniques (NB and SVM) to extract causality with multiple EDUs of a causative unit and multiple EDUs of an effect unit. The verb-pair rules have been represented by the following formula where \( V_c \) is the causative verb concept set, \( V_e \) is the effect verb concept set, \( C \) is the Boolean variables of causality and non-causality. Each causative verb concept \( (v_c \in V_c) \) and each effect verb concept \( (v_e \in V_e) \) are referred to WordNet[12] (http://wordnet.princeton.edu/) and the predefined plant disease information from the Department of Agriculture (http://www.doa.go.th/).

\[
\text{CausalityFunction}: V_c \land V_e \rightarrow C \quad (1)
\]

where the elements of \( V_c \) and \( V_e \) are Catesian products. [4] also proposed to use \( V_c \) and \( V_e \) to solve the boundary of the causative unit and using Centering Theory[13] (which is the center of attention from a discourse segment, and is expressed by a noun) along with \( V_e \) to solve the boundary of the effect unit. When to apply Centering Theory (CT)[4] is whenever the transition state of the center of attention is the smooth shift occurrence (the attention agent, mostly being a subject of a sentence, is changed), the boundary is ended. For example: “If the brown leaphopper aphids suck sap from rice plant, leaves will be yellow. [Leaves] shrink. These aphids destroy plant very fast.”. The effect boundary is ended at ‘[Leaves] shrink’ because the next center of attention is changed to ‘aphids’. However, there are some inter-causal EDUs containing effect units with the smooth shift occurrence although the boundary is not ended, e.g., “The earthquake occurred in China. It caused many buildings were collapsed. Public utilities were cut down. More than 100 people died.”, where ‘buildings’, ‘Public utilities’, and ‘people’ are in the effect boundary with different attentions. Therefore, we propose BN and ME for solving the effect boundary determination without any concern with the attention agent as in [4]. Finally, the major outcomes of their research are the verb-pair rules, with the correctness of the causality-boundary determination varied from 80% to 96% depending on the corpus behaviors, especially the global warming corpus (to which CT could not be applied efficiently).

Chang and Choi’s work[14] has been modified[10,15], that aims to extract causal relations based on one complex sentence to construct the causal network/graph for the term protein with the two relations of the causal relation and the hypernym relation. The edge between a cause node and an effect node of each causal relation represents the causal probability with the directed/indirected causal relation. However, their causal relations cannot show the effect events occurring either concurrence or consequence which is necessary to comprehend effect events for assisting in diagnosis.

Mani et al.[6] reviewed that the representation of event graphs as temporal constraint networks has proved very apropos for TimeML annotation tool (http://nrrc.mitre.org/NRRC/TangoFinalReport.pdf) which is the metadata standard for markup of events and their temporal anchoring of tensed verb, and grammatical aspect in the English documents. The TimeML tool cannot be applied to Thai language which does not have tensed verb in Thai grammar.

Li et al.[16] extracted the temporal relation from Chinese news by using temporal concept frames with constructed rule sets containing the explicit reference time as a temporal indicator which is the temporal marker (Grote[17] defined a temporal marker as “a word or phrase signals the temporal relation between events”). Their temporal concept frames are linked by several events from several sentences with an explicit time expression as the time indicator. [16]’s temporal concept frames with constructed rule sets can achieve a 93% accuracy of temporal relation extraction.

Han and Lavie[18] proposed a time resolution containing a temporal indicator within the framework of temporal constraint satisfaction problems (TCSP) from the Penn Treebank corpora for automatic extraction and reasoning over temporal properties in natural language discourse. In terms of semantics, real calendars (which are explicit time expression on texts) are
modeled as their constraint systems in the TCSP. Solving this TCSP using all-pair-shortest-path algorithm combined with a backtracking search method.

However, there are some implicit temporal markers or expressions in our corpora, to which the methods from [14, 16, 18] cannot be applied to extract automatically the effect-event patterns whether it is consequence or concurrence. Finally, we are aiming at constructing the graph for representing the extracted knowledge (which is the extracted inter-causal EDU) from Thai textual data (which has specific characteristics, e.g., the sentence-like name entity, zero anaphora, and the lack of sentence delimiter) in natural language description, by applying the statistical model and language processing to improve the effect-boundary determination and also to construct the explanation knowledge graph.

3 Problems of Explanation Knowledge Graph Construction from Textual Data

There are two sets of problems: the first problem set consists of the effect boundary determination problems from the inter-causal EDU extraction after applying verb-pair rules in (1) from [4] to identify the inter-causal EDU and to determine the causative boundary. The second problem set is the effect-event pattern determination problem for the explanation knowledge graph construction.

3.1 Effect Boundary Determination Problems

Like other languages, how to determine the effect boundary is by using a discourse marker set, {‘and’, ‘but’, ‘then finally’, ‘after’, ‘because’, ‘since’, ‘when’...} (http://www.usingenglish.com/glossary/discourse-marker.html). Although the discourse marker is used to identify whether the effect boundary ends, we still have some problems of the discourse marker’s connection, the multiple locations of discourse markers, and the implicit discourse marker cue elements in the inter-causal EDU. These problems will effect the graph construction quality.

3.1.1 Discourse Marker’s Connection

Some discourse markers, e.g., ‘and’, ‘but’, ‘then’, are used as either connecting the sequential effect EDUs to the ending effect or connecting between two EDUs other than the ending EDU of the effect boundary.

Example a)

\[
\begin{align*}
\text{EDU}_1: & \quad \text{‘It is late. ’} \\
\text{EDU}_2: & \quad \text{‘The leaves are wilting.’} \\
\text{EDU}_3: & \quad \text{‘And the aphids live inside the rolled leaves...’}
\end{align*}
\]

where \( \text{EDU}_1 \) is the cause, \( \text{EDU}_2, \text{EDU}_3, \) and \( \text{EDU}_4 \) are the effects.

Example b)

\[
\begin{align*}
\text{EDU}_1: & \quad \text{‘It is late.’} \\
\text{EDU}_2: & \quad \text{‘The leaves are wilting.’} \\
\text{EDU}_3: & \quad \text{‘And the aphids live inside the rolled leaves...’}
\end{align*}
\]

where \( \text{EDU}_1 \) is the cause, \( \text{EDU}_2 \) and \( \text{EDU}_3 \) are the effects.

‘And’ in a) marks the end of the effect boundary where in b) marks the connection between \( \text{EDU}_1 \) and \( \text{EDU}_4 \) with the ‘aphid’ attention. From these two examples, whether to determine the ending of the effect boundary is the challenge.

3.1.2 Multiple Locations of Discourse Markers

Discourse Marker can occur in several locations within the effect boundary, as in the following example.

\[
\begin{align*}
\text{EDU}_1: & \quad \text{‘It is late.’} \\
\text{EDU}_2: & \quad \text{‘The leaves are wilting.’} \\
\text{EDU}_3: & \quad \text{‘And the aphids live inside the rolled leaves...’}
\end{align*}
\]

where \( \text{EDU}_2 \) and \( \text{EDU}_3 \) are the effect EDUs resulting from the causative EDU \( \text{EDU}_1 \). From this example, the discourse marker ‘and’ in \( \text{EDU}_3 \) is a connection between two events of \( \text{EDU}_2 \) and \( \text{EDU}_3 \) whereas the discourse marker ‘And’ in \( \text{EDU}_5 \) connects \( \text{EDU}_5 \) to all effect EDUs. In addition, the discourse marker ‘when’ in \( \text{EDU}_6 \) is the connection of \( \text{EDU}_5 \).

There are no capital letters or a sentence delimiter, e.g., ‘.’ or ‘,’ in Thai language.

3.1.3 Implicit Discourse Marker

Causality expressions do not always contain the boundary delimiter expressed by the discourse marker, especially on the effect boundary, causing a problem of identifying the effect boundary.

For example:

\[
\begin{align*}
\text{EDU}_1: & \quad \text{‘It is late. The leaves are wilting...’}
\end{align*}
\]

where \( \text{EDU}_1 \) is the cause, \( \text{EDU}_2, \text{EDU}_3, \) and \( \text{EDU}_4 \) are the effects.

Example b)

\[
\begin{align*}
\text{EDU}_1: & \quad \text{‘It is late.’} \\
\text{EDU}_2: & \quad \text{‘The leaves are wilting.’} \\
\text{EDU}_3: & \quad \text{‘And the aphids live inside the rolled leaves...’}
\end{align*}
\]

where \( \text{EDU}_1 \) is the cause, \( \text{EDU}_2 \) and \( \text{EDU}_3 \) are the effects.

‘And’ in a) marks the end of the effect boundary where in b) marks the connection between \( \text{EDU}_1 \) and \( \text{EDU}_4 \) with the ‘aphid’ attention. From these two examples, whether to determine the ending of the effect boundary is the challenge.

3.1.2 Multiple Locations of Discourse Markers

Discourse Marker can occur in several locations within the effect boundary, as in the following example.

\[
\begin{align*}
\text{EDU}_1: & \quad \text{‘It is late.’} \\
\text{EDU}_2: & \quad \text{‘The leaves are wilting.’} \\
\text{EDU}_3: & \quad \text{‘And the aphids live inside the rolled leaves...’}
\end{align*}
\]

where \( \text{EDU}_2 \) and \( \text{EDU}_3 \) are the effect EDUs resulting from the causative EDU \( \text{EDU}_1 \). From this example, the discourse marker ‘And’ in \( \text{EDU}_3 \) is a connection between two events of \( \text{EDU}_2 \) and \( \text{EDU}_3 \) whereas the discourse marker ‘And’ in \( \text{EDU}_5 \) connects \( \text{EDU}_5 \) to all effect EDUs. In addition, the discourse marker ‘when’ in \( \text{EDU}_6 \) is the connection of \( \text{EDU}_5 \).

There are no capital letters or a sentence delimiter, e.g., ‘.’ or ‘,’ in Thai language.

3.1.3 Implicit Discourse Marker

Causality expressions do not always contain the boundary delimiter expressed by the discourse marker, especially on the effect boundary, causing a problem of identifying the effect boundary.

For example:

\[
\begin{align*}
\text{EDU}_1: & \quad \text{‘It is late. The leaves are wilting...’}
\end{align*}
\]

where \( \text{EDU}_1 \) is the cause, \( \text{EDU}_2, \text{EDU}_3, \) and \( \text{EDU}_4 \) are the effects.
EDU2: “ทำให้น้ำผึ้ง /[it] makes leaves shrink.”
EDU3: “ใบผุ /The leaf edges will dry.”
EDU4: “ทำให้ข้าวเหนื่อย /Rice yield will be reduced.”
EDU5: “รายได้ของนายอย่างกลับ /Farmer’s income will be reduced.”
EDU6: “สัตว์สงครามกลายเป็นโรคในช่วงฤดูร้อน /Rice plant will be destroyed heavily during the summer.”

where EDU2, EDU3, EDU4, and EDU5 are the effects from the cause of EDU1. There is no discourse marker in EDU5 or EDU6. Moreover, EDU4 and EDU5 have the smooth shift occurrence because of changing the attention from ‘rice’ to ‘income’.

These effect boundary determination problems can be solved by learning the effect-verb pair (which is the conceptual-effect-verb order pairs, $v_{ei}v_{ei+1}$, where $v_{ei} \in V_e$, from $EDU_i$ and $EDU_{i+1}$, where $i > 1$) with two different machine learning techniques, BN and ME, to solve the boundary of effect unit without considering the attention agent in CT.

### 3.2 Problem of Effect-Event Pattern Determination for Explanation Knowledge Graph Construction

There is the problem of how to determine the pattern of effect events in the graph whether it is the consequence or the concurrence without any temporal marker (as shown in the following) from the written sequential effect EDUs.

**Temporal marker =** \{‘แต่’/’next’, ‘ในขณะที่’/’later’, ‘ในที่สุด’/’finally’, ‘แล้ว’/’then’, ‘ในขณะที่’/’while, whilst’, ‘ทำให้’/’cause, make’\}.

For example, Explicit Temporal Marker:

EDU1: “ทำให้น้ำผึ้ง /A whole leaf gets disease.”
EDU2: “ทำให้ใบเหี่ยงก่ออยู่ในใบเสื่อมข้างสั้น /[it] will make the leaf curl into the leaf axis.”
EDU3: “ทำให้ใบเหี่ยงก่ออยู่ในใบเสื่อมข้างสั้น /Finally, [it] makes the leaf dry.”

where EDU1 is a cause, EDU2 and EDU3 are effect EDUs with ‘Finally’ as a temporal marker for a consequence pattern as the following.

**Effect EDU**

$EDU_1 \rightarrow EDU_2 \rightarrow EDU_3 = \text{consequence}.$

Implicit Temporal Marker:

EDU4: “ทำให้ผักผุ /[it] makes leaves shrink.”
EDU6: “แล้วในพื้น กระดิ่ง /and leaves dry.”

where $EDU_4$ is a cause and $EDU_5$ and $EDU_6$ are effect EDUs which cannot determine the effect-event pattern about whether it is the consequence or the concurrence as in the following representation.

$EDU_4 \rightarrow EDU_5 \rightarrow EDU_6 = \text{consequence or concurrence}.$

Therefore, we propose to learn the effect-verb order pairs (which is the conceptual-effect-verb order pairs, $v_{ei}v_{ei+1}$ (where $v_{ei} \in V_e$, from $EDU_i$ and $EDU_{i+1}$ (where $i > 1$)) resulted from sliding a window size of two adjacent effect EDUs with the sliding distance of one EDU through the effect unit to resolve the effect-event patterns by determining the odds value\[^{[19]}\] from each effect-verb order pairs.

### 4 Framework of Explanation Knowledge Graph Construction Through Causality Extraction from Textual Data

To extract the inter-causal EDU, there are four steps in our framework. First is a corpus preparation step followed by a learning step: effect-boundary learning and effect-verb order pair learning. The next step is causality extraction followed by the last step of explanation knowledge graph construction, as shown in Fig.1.

#### 4.1 Corpus Preparation

The preparation of corpora was from 8000 EDUs of the agricultural domain of plant disease documents, health news domain and global environment news domain (e.g., global warming). This involved using Thai word segmentation tool to solve the boundary of a Thai word and to tag its part of speech\[^{[12]}\] and to determine its causative/effect concept referred to WordNet\[^{[14]}\].

After the word segmentation is achieved, EDU segmentation\[^{[23]}\] is then to be dealt with to generate EDUs for causality annotation\[^{[4]}\] with the causative/effect concept referred to WordNet and the predefined plant disease information from Department of Agriculture (http://www.doa.go.th/). The annotation example of a causative verb and an effect verb for the inter-causal EDU from [4] is shown in Fig.2. Some of the causative verb concepts and the effect verb concepts are semi-automatically predefined by [4] are shown in Table 1. In addition, 8000 annotated EDUs were divided into 2 parts: the first part is 6000 annotated EDUs to be used for learning. The second part, 2000 annotated EDUs to be used for the effect-boundary evaluation and be randomized for the graph pattern evaluation.
Fig. 1. System architecture.

Table 1. Causative and Effect Verb Sets with Their Concepts[^4] Where $V_c$ = Causative Verb Concept Set, $V_e$ = Effect Verb Concept Set

<table>
<thead>
<tr>
<th>Verb Type</th>
<th>Regular Causative Verb Group</th>
<th>Compound Causative Verb Group</th>
<th>Regular Effect Verb Group</th>
<th>Compound Effect Verb Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Causative Verb Set</strong></td>
<td>Surface Form: $\psi$/suck, $\psi$/suck, $\psi$/eat, $\psi$/bite, $\theta$/drink, $\theta$/eat $\psi$/destruct, $\psi$/eliminate, $\psi$/kill, $\psi$/break, $\psi$/explode, $\psi$/infest $\psi$/spread out, $\psi$/diffuse</td>
<td>Conceptual Causative Verb, $v_c$ (where $v_c \in V_c$)</td>
<td>Surface Form: $\theta$/shrink, $\theta$/bend, $\theta$/twist, $\theta$/curl $\theta$/dry, $\theta$/blast, $\theta$/wilt $\xi$/drop off, $\psi$/come off $\theta$/rot, $\xi$/spoil $\theta$/die</td>
<td>Conceptual Effect Verb, $v_e$ (where $v_e \in V_e$)</td>
</tr>
<tr>
<td><strong>Effect Verb Set</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.2 Effect-Boundary Learning and Effect-Verb Order Pair Learning

There are two objectives in this learning step. The first objective is to determine the effect-boundary rules from each corpus by comparing two machine learning techniques, BN and ME. BN involves the conditional probability determination of each pair of \(v_{ei}, v_{ei+1}\), from the longest effect path (assuming to be the complete path) to solve the effect boundary determination in the next causality extraction step; whereas ME learns from the conditional probability of the effect boundary given a vector of effect verb concept features from sliding the window size of two adjacent effect EDUs with the sliding distance of one EDU through the effect unit from each learning corpus. The second objective is to determine the effect-event patterns (a consequent pattern or a concurrent pattern) by learning the effect-verb order pair resulted from sliding the window size of two adjacent effect EDUs with the sliding distance of one EDU through the effect unit appearing in the randomized corpus. These effect-event patterns will be used for constructing a graph in the explanation knowledge graph construction step.

4.2.1 Effect-Boundary Learning

4.2.1.1 Bayesian Network (BN) Learning

BN\(^8\) represents the joint probability distribution by specifying a set of conditional independent assumptions (represented by a directed acyclic graph), together with sets of local conditional probabilities. Each variable in the joint space is represented by a node in BN. For each variable, two types of information are specified. First, the network arcs represent the assertion that the variable is conditionally independent of its non-descendants in the network given its immediate predecessors in the network. We say \(X\) is a descendant of \(Y\) if there is a directed path from \(Y\) to \(X\). Second, a conditional probability table is given for each variable, describing the probability distribution for that variable given the values of its immediate predecessors. The joint probability for any desired assignment of values \((y_1, \ldots, y_n)\) to the tuple network variables \((Y_1, \ldots, Y_n)\) can be computed by the formula

\[
p(y_1, \ldots, y_n) = \prod_{i=1}^{n} P(y_i|\text{Parents}(Y_i)) \tag{2}
\]

where \(Y_0\) is the parents of \(Y_1\), and \(\text{Parents}(Y_i)\) denotes the set of immediate predecessors of \(Y_i\) in the network. The values of \(P(y_i|\text{Parents}(Y_i))\) are precisely the values stored in the conditional probability table associated with node \(Y_i\). \(^8\) also mentioned that the Bayesian structure could be constructed from the independence and dependence relationships from the data.

However, (2) is applied to our effect-boundary determination with \((Y_1, \ldots, Y_n)\) as the effect event set, \(\{E_1, \ldots, E_n\}\) where \(Y_0 = \text{cause}\). This effect event set is \(V\) because each event can be expressed by verb, especially EDU’s main verb. From the effect event order occurring on the text without any interrupt-EDU, each effect event, \(E_i\) (where \(i = 1, 2, 3, \ldots, n\)), is \(v_{ei}\) (where \(v_{ei} \in V_e\), in Table 1) from \(EDU_j\) (where \(j = i + 1\)), as shown in the following example with the predicate representation.

\[EDU_1: \text{""} \text{insect} \text{""} \text{effect} \text{""} \text{consume} \text{""} \text{sap from leaves}. \text{""} \]

\(\text{(Consumes(Aphid/insect, sap/solution) } \wedge \text{ Exist.in (sap, leaf/plant organ))}\)
EDU2: “ทำให้ไม่ปกติ ใบเดิมๆ/[it] makes leaves shrink.”
(\text{Be\_abnormal\_shape(leaf/plant organ)}) — E_1 \text{ is ‘Be\_abnormal\_shape’}
EDU3: “ใบเหลือง/leaves dry.”

(\text{Dry(leaf/plant organ)}) — E_2 \text{ is ‘Dry’}
EDU4: “ผลไม้/and leaves] come off.”
(\text{Be\_fallen\_off(leaf/plant organ)}) — E_3 \text{ is ‘Be\_fallen\_off’}

Table 2. Inter-Causal EDU Features from an Annotated Corpus Example of a Plant Disease

<table>
<thead>
<tr>
<th>Causality</th>
<th>NP1</th>
<th>VC (v_c)</th>
<th>NP2</th>
<th>NP3</th>
<th>NP4</th>
<th>VE (v_e)</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_1</td>
<td>aphid/plant</td>
<td>#2</td>
<td>sax/solution #1</td>
<td>leaves/plant organ #1</td>
<td>leaves/plant organ #1</td>
<td>shrink/be...yes</td>
<td></td>
</tr>
<tr>
<td>1_2</td>
<td>aphid/plant</td>
<td>#2</td>
<td>sax/solution #1</td>
<td>leaves/plant organ #1</td>
<td>leaves/plant organ #1</td>
<td>dry/be...yes</td>
<td></td>
</tr>
<tr>
<td>1_3</td>
<td>aphid/plant</td>
<td>#2</td>
<td>sax/solution #1</td>
<td>leaves/plant organ #1</td>
<td>leaves/plant organ #1</td>
<td>come/off...yes</td>
<td></td>
</tr>
<tr>
<td>1_4</td>
<td>plant/plant</td>
<td>#1</td>
<td>Bloom/</td>
<td>—</td>
<td>—</td>
<td>aphid/plant increase #1</td>
<td>no</td>
</tr>
<tr>
<td>1_5</td>
<td>aphid/plant</td>
<td>#2</td>
<td>Suck/consume #1</td>
<td>plant/plant organ #1</td>
<td>plant/plant organ #1</td>
<td>decrease/decrease</td>
<td>yes</td>
</tr>
<tr>
<td>1_6</td>
<td>plant/plant</td>
<td>#1</td>
<td>sprout/sprout #1</td>
<td>spread/out/</td>
<td>spread #1</td>
<td>Organ/plant be incomplete/</td>
<td>yes</td>
</tr>
<tr>
<td>1_7</td>
<td>aphid/plant</td>
<td>#2</td>
<td>spread out/</td>
<td>spread #1</td>
<td>—</td>
<td>plant #1</td>
<td>yield #1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 3. Show the Sequence of E_i or v_e, Appearing in the Example Documents of Plant Disease from Aphids, with the Conditional Probabilities from BN Learning

| E_i                  | P(E_i) | E_2 | P(E_2|E_i) | E_4 | P(E_4|E_i, E_2) |
|----------------------|--------|-----|---------|-----|--------------|
| \text{Be\_abnormal\_shape (leaf)} |        |      |         |     |              |
| \text{Be\_abnormal\_shape (leaf)} | Change-in-color (leaf) | 0.01667 |       |       |              |
| \text{Be\_abnormal\_shape (leaf)} | \text{Be\_fallen (leaf)} | 0.00832 | Stunt (plant) | 0.00832 |              |
| \text{Be\_abnormal\_shape (leaf)} | \text{Be\_low (tillering)} | 0.00832 |       |       |              |
| \text{Be\_abnormal\_shape (leaf)} | \text{Be\_mark (leaf)} | 0.00832 | \text{Be\_fallen (leaf)} | 0.00832 |              |
| \text{Be\_abnormal\_shape (leaf)} | \text{Be\_thin (leaf)} | 0.00832 | \text{Be\_rough (leaf)} | 0.00834 |              |
| \text{Be\_abnormal\_shape (leaf)} | \text{Be\_thin (leaf)} | 0.01667 | Stunt (plant) | 0.00834 |              |
| \text{Be\_abnormal\_shape (leaf)} | \text{Dry (leaf)} | 0.00833 | \text{Be\_fallen (leaf)} | 0.00832 |              |
| \text{Be\_abnormal\_shape (leaf)} | \text{Dry (leaf)} | 0.00833 | \text{Be\_flowerless} | 0.00833 |              |
| \text{Be\_abnormal\_shape (leaf)} | \text{Dry (leaf)} | 0.05000 |       |       |              |
| \text{Be\_abnormal\_shape (leaf)} | \text{Dry (leaf)} | 0.01667 | \text{Stop (growth)} | 0.00834 |              |
| \text{Be\_abnormal\_shape (leaf)} | \text{Stop (growth)} | 0.16667 | Stunt (plant) | 0.01667 |              |
| \text{Change-in-color (leaf)} | Reduce (leafsize) | 0.02600 |       |       |              |
| \text{Change-in-color (leaf)} | Sprout_slowly (leaf) | 0.00832 |       |       |              |
| \text{Change-in-color (leaf)} | \text{Be\_abnormal\_shape (leaf)} | 0.00832 |       |       |              |
| \text{Change-in-color (leaf)} | \text{Be\_fallen (leaf)} | 0.00832 |       |       |              |
| \text{Change-in-color (leaf)} | \text{Be\_fallen (leaf)} | 0.01667 |       |       |              |
| \text{Change-in-color (leaf)} | \text{Dry (leaf)} | 0.00832 |       |       |              |
| \text{Change-in-color (leaf)} | \text{Dry (leaf)} | 0.01667 | \text{Be\_fallen (leaf)} | 0.00834 |              |
| \text{Change-in-color (leaf)} | \text{Dry (leaf)} | 0.01667 | \text{Reduce (yield)} | 0.00834 |              |
| \text{Dry (leaf)} | \text{Be\_abnormal\_shape (leaf)} | 0.00832 |       |       |              |
| \text{Dry (leaf)} | \text{Be\_fallen (leaf)} | 0.00832 |       |       |              |
| \text{Dry (leaf)} | \text{Be\_fallen (leaf)} | 0.01667 |       |       |              |
| \text{Dry (leaf)} | \text{Dry (leaf)} | 0.01667 |       |       |              |
| \text{Dry (leaf)} | \text{Dry (leaf)} | 0.05833 | Change-in-color (leaf) | 0.00832 | \text{Be\_abnormal\_shape (leaf)} | 0.00832 |              |
| ... | ...              | ...       | ...          | ...          | ...          | ...      | ...   |
where \( EDU_1 \) is a causative EDU and \( EDU_2, EDU_3, \) and \( EDU_4 \) are effect EDUs.

All annotated concepts of the causative verbs, the effect verbs, and noun phrases from each annotated corpus from the previous step are transformed to a data file of inter-causal EDU features, in Table 2, for determining the conditional probabilities of the conceptual effect verbs as shown in Table 3.

From Table 3, we can conclude that the least probability of \( P(E_i|E_1, \ldots, E_{-1}) \) is 0.00832 is the effect-boundary threshold with the actual effect-boundary threshold of 0.005 for determining the effect boundary, as shown in the following rule (named the effect-boundary rule).

\[
\text{IF } P(v_{e1}, v_{e2}, v_{e3}|v_{e1} \leq i \leq v_{e2} > v_{e3}) > \text{EBThreshold} \\
\text{THEN EffectBoundary } = \{E_1, \ldots, E_i\}
\]

where \( \text{EBThreshold} \) is the actual effect-boundary threshold.

All the conditional probabilities of effect verb concepts from the plant disease corpus (shown in Table 3) are determined according to the sequence of effect verbs that appeared in the documents. From the experiment, each \( v_{ei} \) occurred in several long or short paths including paths containing some implicit effect verbs, causing the determination of the conditional probabilities of \( v_{ei} \), from their complete paths by sorting all paths (in Table 3) according to the longest path. Then, the conditional probability of each \( v_{ei} \) is determined from the first appearance in its longest paths for the effect boundary determination in the causality extraction step.

### 4.2.1.2 Maximum Entropy (ME) Learning

ME models implement the intuition that the best model will be the one that is consistent with the set of constraints imposed by the evidence, but otherwise it is as uniform as possible\(^{[7,24]}\). Fleischman M \textit{et al.}\(^{[25]}\) modeled the probability of a semantic role \( r \) given a vector of features \( x \) according to the ME formulation below:

\[
p(r|x) = \frac{1}{Z_x} \exp \left( \sum_{j=0}^{n} \lambda_j f_j(r, x) \right) \quad (3)
\]

where \( Z_x \) is a normalization constant, \( f_j(r, x) \) is a feature function which maps each role and vector element (or combination of elements) to a binary value, \( n \) is the total number of feature functions, and \( \lambda_j \) is the weight for a given feature function. The final classification is just the role with the highest probability given its feature vector and the model.

According to (3), ME can be used as the classifier of the \( r \) class when \( p(r|x) \) is the highest probability or \( \text{arg max } p(r|x) \) to determine two effect boundary classes, ending and continuing, where \( r \) is the effect boundary class (boundary is ending when \( r = 0 \), otherwise \( r = 1 \)) and \( x \) is the binary vector of the effect-verb concept \( (v_e) \) features containing an effect-verb concept pair \( (v_{ei}, v_{ei+1}) \), where \( v_{ei} \in V_e \) and \( v_{ei+1} \in V_e \), as shown in (4). All pairs of \( v_{ei} v_{ei+1} \) are gained by sliding a window size of two adjacent effect EDUs with one EDU distance through the effect EDU unit.

\[
\begin{align*}
p(r|x) &= \arg \max_r \frac{1}{2} \exp \left( \sum_{j=1}^{n} \lambda_j f_{yes,ei,j}(r, v_{ei}) + \sum_{j=1}^{n} \lambda_j f_{no,ei,j}(r, v_{ei}) + \sum_{j=1}^{n} \lambda_j f_{yes,ei+1,j}(r, v_{ei+1}) + \sum_{j=1}^{n} \lambda_j f_{no,ei+1,j}(r, v_{ei+1}) \right) \\
\lambda_j \text{ are shown in the following}
\end{align*}
\]

\[
\begin{array}{cccccc}
v_e & \lambda_j & v_{ei+1} & \lambda_j \\
\text{Abnormal} & \text{Leaf} & \text{Shape} & \text{Leaf} & \text{Resin} & 26.3399 \\
\text{Shape} & \text{Leaf} & \text{Shape} & \text{Leaf} & \text{Resin} & 25.8771 \\
\text{Stunt} & \text{Leaf} & \text{Leaf} & \text{Leaf} & \text{Resin} & 8.1437 \\
\end{array}
\]

### 4.2.2 Effect-Verb Order Pair Learning

This learning step is to determine the effect-event pattern of a consequence or a concurrence by determining the odds value (http://www.stat.ubc.ca/~rollin/teach/243w04/lec/node50.html) of the effect-verb order pair resulted by sliding a window size of two adjacent effect EDUs with one sliding effect-EDU distance through the effect boundary. The odds value is a numerical value given by (5) of two effect-verb order pairs with probability values, \( p \) and \( 1-p \).

\[
\text{Odds}[\text{EffectVerbOrderPair}] = \frac{p}{1-p} \quad (5)
\]

where \( p \) is the probability of the effect-verb order pair, \( v_{ex} v_{ey} \), within a slide window; \( 1-p \) is the probability of the effect-verb order pair, \( v_{ey} v_{ex} \), within a slide window.

Therefore, the odds value is used to determine whether it is the consequent event or the concurrent event from two different sequences of the effect-verb order pairs, \( v_{ex} v_{ey} \) and \( v_{ey} v_{ex} \), from the randomized corpus which is the plant disease corpus (see Table 4).
Table 4. Effect-Event Pattern of Effect-Verb Order Pair, \( v_{ex}v_{ey} \), where \( p \geq 1 - p \), CS is a Consequent Event and CC is a Concurrent Event

<table>
<thead>
<tr>
<th>( v_{ex} )</th>
<th>( v_{ey} )</th>
<th>( v_{ey}v_{ex} )</th>
<th>( v_{ex}v_{ey} )</th>
<th>Odds</th>
<th>Effect-Event Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change-in-color (leaf)</td>
<td>Be_abnormal_shape (leaf)</td>
<td>0.800</td>
<td>0.200</td>
<td>4.00</td>
<td>CS</td>
</tr>
<tr>
<td>Be_abnormal_shape (leaf)</td>
<td>Dry (leaf)</td>
<td>0.860</td>
<td>0.140</td>
<td>6.00</td>
<td>CS</td>
</tr>
<tr>
<td>Be_abnormal_shape (leaf)</td>
<td>Stunt (plant)</td>
<td>0.511</td>
<td>0.489</td>
<td>1.05</td>
<td>CC</td>
</tr>
<tr>
<td>Change-in-color (leaf)</td>
<td>Stunt (plant)/reduce (leaf, size)</td>
<td>0.670</td>
<td>0.330</td>
<td>2.00</td>
<td>CS</td>
</tr>
<tr>
<td>Be_fallen (leaf)</td>
<td>Stunt (plant)</td>
<td>1.000</td>
<td>0.000</td>
<td>( \infty )</td>
<td>CS</td>
</tr>
<tr>
<td>Dry (leaf)</td>
<td>Be_fallen (leaf)</td>
<td>1.000</td>
<td>0.000</td>
<td>( \infty )</td>
<td>CS</td>
</tr>
<tr>
<td>Dry (leaf)</td>
<td>Die (leaf)</td>
<td>1.000</td>
<td>0.000</td>
<td>( \infty )</td>
<td>CS</td>
</tr>
<tr>
<td>Dry (plant)</td>
<td>Die (plant)</td>
<td>1.000</td>
<td>0.000</td>
<td>( \infty )</td>
<td>CS</td>
</tr>
<tr>
<td>Dry (leaf)</td>
<td>Reduce (yield)</td>
<td>1.000</td>
<td>0.000</td>
<td>( \infty )</td>
<td>CS</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

From Table 4, all effect-verb order pairs can be used to determine the effect-event patterns as the consequence of \( v_{ex}v_{ey} \) except the pattern of “Be\_abnormal\_shape(leaf)” to “Stunt(plant)” which trends to be the concurrence pattern with odd value = 1.05 at 95% confidence interval.

4.3 Causality Extraction

After the effect-boundary learning step, the next is the effect-boundary recognition or the causality extraction using a statistical baseline. This step can be separated into 2 parts: cause-effect identification and cause-effect boundary determination.

4.3.1 Cause-Effect Identification

The \( Vc \) set, the \( Ve \) set, and the verb-pair rule from [4] are used to identify the interesting locations of the inter-causal EDU, especially a cause consequence (a causative-EDUs unit immediately followed by an effect-EDU unit) and a nonadjacent cause consequence (a causative-EDUs unit followed by some non-causative/non-effect EDUs followed by an effect-EDU unit). From corpus behavior study (see Appendix), there are four EDUs as the maximum number of EDUs existing between a causative unit and an effect unit. And, there are two EDUs as the most likely number of EDUs existing between the causative unit and the effect unit. After the causative unit is identified by the \( Vc \) set, the ending boundary of a causative unit and the starting boundary of an effect unit are determined by the \( Ve \) set within five EDUs right after the first causative EDU.

4.3.2 Effect Boundary Determination

The ending boundary of an effect-EDU unit can be solved by two different methods of machine learning techniques, BN and ME based on 10-fold cross validation.

4.3.2.1 Bayesian Network (BN) Learning

The effect boundary is determined by using the effect-boundary rule with the conditional probability of each effect verb concept (\( v_{ey} \)) from its longest path of the BN learning (where the longest path is assumed to be the completed path), as shown in Fig.3 of the

![Fig.3. Multiple EDUs of causality extraction algorithm.](image)

Assume that each EDU is represented by a 3-tupple \((NP, VP, CONJ)\)

\( L \) is a list of EDU.

\( Vc \) is a set of causative verb. \( Vc \) or \( Ve \) is a set of effect verb.

\( DM \) is a discourse marker set.

MUltiple_EDUs_OF_CAUSALITY_EXTRACTION \((L, Vc, Ve)\)

```java
1  i ← 0, R ← ∅
2  while i ≤ length \([L]\) do
3      Begin
4          CA ← ∅, EC ← ∅, j ← 0 /*CA is a cause EDU, EC is an effect EDU.
5          if \((VP_i \in Vc)\) then /*determine a cause consequence and a nonadjacent cause consequence.
6              Concept ← VP_i
7              while \(((VP_i \in Vc) \land (VP_i = Concept)) \lor ((VP_i \notin Vc) \land (j < 6))\) do
8                  \{CA ← CA ∪ \{i\}, i ← i + 1, j ← j + 1\}
9                  EFFECT_BOUNDARY_DETERMINATION
10          else if \((VP_i \in Vp)\) then /*determine an adjacent consequence-antecedent.
11                 while \((VP_i \notin Vc)\) do
12                     EC ← EC ∪ \{i\}, i ← i + 1,
13                     CA ← CA ∪ \{i\}, i ← i + 1
14                 endif
15             R ← R ∪ \{(CA, EC)\}
16         End
17     return R
```

Fig.3. Multiple EDUs of causality extraction algorithm.
of each effect-verb order pair, \(v_{ex}v_{ey}\), from the learning step is used for constructing the graphical model of the explanation knowledge by comparing each \(v_{ex}v_{ey+1}\) pair (from the extracted inter-causal EDU) to \(v_{ex}v_{ey}\) of the learned effect-event pattern of the consequence or the concurrence (in Table 4). Meanwhile, all the effect verb concepts are provided by [4]. Fig.6 shows the EKGC (Explanation Knowledge Graph Construction) algorithm, where \(v_{EDU}\) is the effect verb concept, \(v_{ex}\) in \(EDU_i\) and \(v_{EDU_i+1}\) is the effect verb concept, \(v_{ey}\) in \(EDU_i+1\). The result from Fig.6 is the constructed graph of plant disease symptoms caused by aphids shown in Fig.7.

**EXPLANATION_KNOWLEDGE_GRAPH_CONSTRUCTION**

1. \(v_{EDU_i}, v_{EDU_i+1}, v_{EDU_i+1}\) are the effect verb concept vertices from \(EDU_i\), \(EDU_i+1\) and \(EDU_i+1\), respectively.
2. \(Graph = (VX, ED)\) where each vertex/node is an event represented by a causative verb concept (\(v_e\)) or an effect verb concept (\(v_{ex}\) or \(v_{ey}\)) and vertex \(\in VX\), each edge connects two event nodes and edge \(\in ED\)/*
3. If (Extracted_EffectVerb_Concept_Pair \((v_{ex}, v_{EDU_i+1})\) = consequence)
   4. ConsequentSubgraph (Vertex \(v_{EDU_i+1}\),
      \(Edge_{v_{ex}, v_{EDU_i+1}}: P(v_{ex}, v_{EDU_i+1})\))
   5. /*draw a consequent subgraph connecting from the \(v_{ex}\) to the vertex of \(v_{ey}\) vertex with assigning probability \(P(v_{ex}, v_{EDU_i+1}|v_{EDU_i})\) to the edge connecting these two vertices.
   6. */
7. Else if (Extracted_EffectVerb_Concept_Pair \((v_{ey}, v_{EDU_i+1})\) = concurrence)
   8. ConcurrentSubgraph (Vertex \(v_{EDU_i+1}\),
      \(Edge_{v_{ey}, v_{EDU_i+1}}: P(v_{ey}, v_{EDU_i+1}|v_{EDU_i})\))
   9. /*draw a concurrent subgraph connecting from the vertex of \(v_{ey}\) to the vertex of \(v_{ex}\) with assigning probability \(P(v_{ey}, v_{EDU_i+1}|v_{EDU_i})\) to the edge connecting these two vertices, where \(v_{EDU_i}\) is the parent node of \(v_{EDU_i}\).
   10. */
11. Endif
12. return

Fig.6. Explanation knowledge graph construction algorithm.

### 5 Evaluation and Discussion

### 5.1 Effect-Boundary Extraction

The corpora used to evaluate the proposed model of the effect-boundary determination using two different machine learning techniques, BN and ME, consist of 2000 EDUs collected from online plant disease technical papers, bird flu news, health news, and environmental

**Fig.4. Effect boundary determination algorithm by BN learning.**

MECE (Multiple EDUs of Causality Extraction) algorithm connected to the effect boundary determination shown in Fig.4. Moreover, the effect boundary from the adjacent consequence-antecedent case (an effect unit followed by a causative unit) can be solved by the verb-pair rule [4] with its causative unit mostly containing one causative EDU from the corpus behavior studying by [4].

#### 4.3.2.2 Maximum Entropy (ME) Learning

From the learning step of the effect boundary learning by ME, we use \(\lambda_j\) (the weight for a given feature function of the effect boundary with a vector of \(v_e\) features containing the \(v_{ex}v_{ey+1}\) pair) to determine the effect boundary by (4) as shown in the effect boundary determination algorithm by ME (in Fig.5) called by the MECE algorithm (Fig.4).

**Fig.5. Effect boundary determination algorithm by ME learning.**

### 4.4 Explanation Knowledge Graph Construction

The extracted multiple EDUs of causality from the causality extraction step and each effect-event pattern
news on global warming. Each of these corpora has different characteristics of the effect verb frequency, the diversity of verb occurrence, and the center of attention for the effect unit, which mostly contains the same agent except the global warming corpus. This kind of characteristic causes this research to improve the effect-boundary determination from [4]. The results of using two different models of learning the effect boundary, are evaluated by comparing their correctness of the effect-boundary determination to the ones from [4] (see Table 5) based on two experts and one linguist with max-win voting. In addition to the causality extraction by using verb-pair rules [4], the evaluation of the causality is expressed in terms of the precision (0.85 by average) and the recall (0.72 by average).

The experimental results from Table 5 illustrate that the correctness of effect boundary determination achieved by BN, ME, and Applied CT are influenced by different factors. The BN learning technique is influenced by the verb frequency, as shown in Table 5 where BN results with the maximum correctness of the effect boundary determination to the plant disease technical documents which have the high frequency of verb occurrence and with the lowest correctness of the effect boundary determination to the corpora containing medium frequency of verb occurrence such as global warming news and the bird flu news. On the contrary, the correctness of the effect boundary determination is at the lowest when ME is applied to the plant disease technical document and at the highest when ME is applied to global warming news. This is due to the effect verb ambiguity increased by the number of two adjacent causality forms (a cause-effect form followed by an effect-cause form) in the corpus containing the high verb frequency occurrence and the low verb diversity occurrence, which effects $\lambda$ in (4) of the ME classification and leads to lower correctness of effect boundary determination. The following example is the effect verb ambiguity.

$EDU_1$ “ใบที่ถูกกัดกินแมลงแมว/Leaves that are sucked sap by Mango leafhopper.”

$EDU_2$ “จะพักไป/ will bend down”

$EDU_3$ “ใบเล้าใบกลมมีสีดำ/The leaflet tip has black color.”

$EDU_4$ “จากที่รวงตั้งฤดูกับ/since black mold cover it.”

where the effect verb concept in $EDU_3$ is “Change_in_color(leaf)” which can be the effect from either $EDU_1$ or $EDU_4$.

Table 5 also shows that the correctness of effect boundary determination is at the lowest when Applied CT is applied to the global warming news corpus. Applied CT will result in low percentage of the correctness of the effect boundary determination when the center of attention is mostly with a different agent as in the

<table>
<thead>
<tr>
<th>Document Type</th>
<th>No. Different $v_e$</th>
<th>Frequency of $v_e$ Occurrence</th>
<th>Center of Attention (in effect unit)</th>
<th>Correctness of Effect Boundary Determination (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health news</td>
<td>69</td>
<td>medium</td>
<td>Mostly same agent</td>
<td>90 95 94</td>
</tr>
<tr>
<td>Bird flu news</td>
<td>49</td>
<td>medium-high</td>
<td>Mostly same agent</td>
<td>89 92 94</td>
</tr>
<tr>
<td>Plant disease technical</td>
<td>43</td>
<td>high</td>
<td>Mostly same agent</td>
<td>92 91 91</td>
</tr>
<tr>
<td>document</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global warming news</td>
<td>65</td>
<td>medium</td>
<td>Mostly different agent</td>
<td>89 97 79</td>
</tr>
</tbody>
</table>

Fig.7. Explanation knowledge graph of plant disease symptoms caused by aphids.
global warming news corpus. However, in order to see whether there are any significant differences when these three techniques are applied, t-test will be used. The t-test measure defined in (6) is used to compare the ability of different techniques or methodologies to determine the effect boundary correctly with or without a significant difference between them in Table 6, Table 7, and Table 8.

\[ t = \frac{p_1 - p_2}{\sqrt{p_0 q_0 \left(\frac{2}{n}\right)}} \]  

(6)

where \( p_0, q_0 \) are proportion weights, \( p_0 = \frac{x_1 + x_2}{2n} \), \( q_0 = 1 - p_0 \); \( x_1 \) is the number of samples correctly classified by methodology 1; \( x_2 \) is the number of samples correctly classified by methodology 2; \( p_1 \) is the proportion of accuracy in classifying by methodology 1; \( p_2 \) is the proportion of accuracy in classifying by methodology 2; \( n \) is the number of experiment samples.

Although the correctness of the effect boundary determination shown in Table 5 varies between the methodologies due to specific corpora characteristics, the results shown in Table 6, Table 7, and Table 8 describe each methodology/technique the effect boundary determination remaining insignificant at 95% confidence interval with the exception between using ME and applied CT with the global warming news corpus (in Table 8). The results in Table 5 and Table 8 show that ME can achieve significant improvement in the correctness of the effect boundary determination when approaching corpora with a high verb diversity, medium or low verb frequency, and the center of attention in the effect unit to be mostly with different agents.

### 5.2 Explanation Knowledge Graph Construction

The evaluation of the causality graph construction is evaluated from the correctness of the odds value of the effect-verb order pair whether the effect-event pattern is a consequence or a concurrence. The evaluation of the effect-event pattern is evaluated by three expert judgments with max-win voting, as shown in Table 9 with the correctness of 90%.

<table>
<thead>
<tr>
<th>Pattern of Effect-Verb Pair ( v_{ex} ) to ( v_{ey} )</th>
<th>Odds = ( p/(1-p) )</th>
<th>Effect Event Pattern</th>
<th>Correctness by Experts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change-in-color(leaf) to Be_abnormal_shape(leaf)</td>
<td>4.00</td>
<td>consequence</td>
<td>true</td>
</tr>
<tr>
<td>Be_abnormal_shape(leaf) to Dry(leaf)</td>
<td>6.00</td>
<td>consequence</td>
<td>true</td>
</tr>
<tr>
<td>Be_abnormal_shape(leaf) to Stunt(plant)</td>
<td>1.05</td>
<td>consequence</td>
<td>false</td>
</tr>
<tr>
<td>Change-in-color(leaf) to Stunt(plant)/reduce(leaf_size)</td>
<td>2.00</td>
<td>consequence</td>
<td>true</td>
</tr>
<tr>
<td>Be_fallen(leaf) to Stunt(plant)</td>
<td>∞</td>
<td>consequence</td>
<td>true</td>
</tr>
<tr>
<td>Dry(leaf) to Be_fallen(leaf)</td>
<td>∞</td>
<td>consequence</td>
<td>true</td>
</tr>
<tr>
<td>Stunt(plant) to Be_less(flower)</td>
<td>∞</td>
<td>consequence</td>
<td>true</td>
</tr>
<tr>
<td>Stop(growth) to Stunt(plant)</td>
<td>∞</td>
<td>consequence</td>
<td>true</td>
</tr>
<tr>
<td>Dry(leaf) to Come_off(flower)</td>
<td>∞</td>
<td>consequence</td>
<td>true</td>
</tr>
<tr>
<td>Come_off(flower) to Reduce(yield)</td>
<td>∞</td>
<td>consequence</td>
<td>true</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

Table 6. t-Test of the Correctness of the Effect Boundary Determination by Different Techniques, BN and ME

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Boundary Determination (%)</th>
<th>t-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health news</td>
<td>90</td>
<td>95</td>
</tr>
<tr>
<td>Bird flu news</td>
<td>89</td>
<td>92</td>
</tr>
<tr>
<td>Plant disease and</td>
<td>92</td>
<td>91</td>
</tr>
<tr>
<td>technical document</td>
<td>89</td>
<td>97</td>
</tr>
</tbody>
</table>

Table 7. t-Test of the Correctness of the Effect Boundary Determination by Different Techniques, BN and Applied CT

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Boundary Determination (%)</th>
<th>t-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health news</td>
<td>90</td>
<td>94</td>
</tr>
<tr>
<td>Bird flu news</td>
<td>89</td>
<td>94</td>
</tr>
<tr>
<td>Plant disease and</td>
<td>92</td>
<td>91</td>
</tr>
<tr>
<td>technical document</td>
<td>89</td>
<td>79</td>
</tr>
</tbody>
</table>

Table 8. t-Test of the Correctness of the Effect Boundary Determination by Different Techniques, ME and Applied Centering Theory

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Boundary Determination (%)</th>
<th>t-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health news</td>
<td>95</td>
<td>94</td>
</tr>
<tr>
<td>Bird flu news</td>
<td>92</td>
<td>94</td>
</tr>
<tr>
<td>Plant disease and</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>technical document</td>
<td>97</td>
<td>79</td>
</tr>
</tbody>
</table>
The error of the effect-event pattern comes from two effect events being concurrence with one effect event occurring first which are a reflex on how to be expressed on texts. For example, the stunt(plant) event and the Be_abnormal_shape(leaf) event, e.g., shrink(leaf), are lately concurrent occurrences with the Be_abnormal_shape(leaf) event occurring first. Therefore, two appearance sequences between two effect events on text; ‘Stunt(plant) to Be_abnormal_shape(leaf)’ and ‘Be_abnormal_shape(leaf) to Stunt(plant)’, are not quite different which result in the effect event pattern. Then, the corrected causality graph is shown in Fig.8.

6 Conclusion

This research approaches constructing the graphical model for representation of the explanation knowledge through extracting the inter-causal EDU (multiple EDUs of causality) from textual data with the improvement of the effect-boundary determination. Previous researches concerned the discourse marker\cite{11} and NP pairs with the cue phrase\cite{14}. However, the problems of implicit and ambiguity discourse markers in combination with zero anaphora lead us to focus on verbs because verbs can express events with a consequence or a concurrence. This paper constructs the explanation knowledge graph from extracting more complete explanation knowledge from texts for supporting the expert system in diagnosis and for answering with reasoning in the QA system.

Our methodology of automatically constructing the explanation knowledge graph consists of the inter-causal EDU extraction and the construction of the graphical model. Our current methodology of the inter-causal EDU extraction can efficiently extract the multiple EDUs of causality, especially the boundary of the effect unit by using different machine learning techniques, BN and ME. The correctness average of an effect boundary determination for BN is 90%, for ME is 93.75%, and for Applied CT is 89.5%. Statistical analysis has shown that the differences between the results from ME, BN, and Applied CT were mostly insignificant. However, ME has shown a significant improvement when approaching the corpora with a high verb diversity, medium/low verb frequency, and the center of attention in the effect unit to be mostly with different agents. According to (4), ME gives a better result if there is a high correlation among \( v_e \) features without two adjacent causality forms (a cause-effect form followed by an effect-cause form) where there are relationships or high correlation among \( v_e \) features in our corpora along with the amount of two adjacent causality forms varied by different domains.

Our explanation knowledge graph construction can be computed by determining the effect-event patterns of the consequence/concerunce from the odds value of the effect-verb order pair by sliding a window-size of two adjacent effect EDUs with the sliding distance of one EDU through the effect unit. Therefore, our graph construction methodology can successfully construct the graph with 90% correctness from randomizing the corpus. However, the quality of the constructed explanation knowledge graph is based on the quality of the extracted multiple EDUs of causality which requires more improvement in the next research. Because there are some problems that our methodology does not consider, e.g., the interruption within the effect consequent unit and two adjacent causality forms (a cause-effect form or a cause-consequence causality form immediately followed by an effect-cause form or a consequence-antecedent causality form), as shown in the following example. These two problems will challenge the capability in boundary determination.

\begin{align*}
\text{EDU}_1 & \quad \text{“เพิ่ม ข้าวสำนัก เลี้ยงผึ้งใบก้าว ไปนอนกินน้ำ / Green leaf hopper infests sap from rice leaves.”} 
\end{align*}
EDU_1 "The root develops.”
EDU_2 “The plant grows.”
EDU_3 “The leaf edges will dry.”
EDU_4 “The rice plant will have severe symptoms.”
EDU_5 "มีแมลงสาปที่จะระบาด./when the aphids spread out.”

where EDU_1, EDU_3, and EDU_4 are the cause-effect form (with EDU_1 as a cause, while EDU_2 and EDU_3 are the effect from EDU_1). EDU_4 and EDU_5 is the effect-cause form (with EDU_5 as the cause while EDU_4 is an effect from EDU_5).

Furthermore, the constructed explanation knowledge graph through causality extraction from texts by this research is very useful for assisting the expert system to reasonably analyze and diagnose problems existing in which state of events for prediction of the next events, as shown in Fig.8. And also, our explanation knowledge graph will be useful for answering clearly the why-question in the automatic QA system. Finally, our methodology of constructing the graphical model of the explanation knowledge from the inter-causal EDU extraction can be applied to other languages other than the Thai language.

Acknowledgments
Patrick Saint Dizier, C. Yingseree, and Naist Lab have contributed greatly in this research, as they have shared their expertise in the field, which assisted in finalizing this research. We would also like to thank J. Pechsiri, N. Savavibool, and T. Amusas-Amornkul for their contribution in this work.

References
Chaveevan Pechsiri holds the Bachelor's degree of science in food science and technology from Kasetsart University, Thailand, the Master's degree in food science and the Master's degree in computer science both from Mississippi State University, USA, and the D.Eng. degree in computer engineering from Kasetsart University, Thailand. She is an associate professor at Dhurakij Pundit University, Thailand and her research interest is natural language processing.

Rapepun Piriyakul is currently an assistant professor at Ramkhamhaeng University, Thailand. She received the Bachelor's degree in mathematics from Chulalongkorn University, the Master's degree in applied statistics from National Institute of Development Administration, and the D.Eng. degree in computer engineering from Kasetsart University, Thailand. Her research interest is applied analytical statistics in computer engineering.

Appendix

Corpus study of the number of EDUs existing between a causative unit and an effect unit

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. number of EDUs</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>existing between</td>
<td>causative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>causative unit and</td>
<td>unit and</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>effect unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Most likely number</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>of EDUs existing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>between causative unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and effect unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>