Explanation Based Why Question Answering System

Chaveevan Pechsiri. *Member, IAENG*

***Abstract*— The research aims to develop an automatic Why Question Answering system on community web-boards for supporting ordinary people in preliminary diagnosis problems such as plant disease problems. The research contains two main problems, how to identify the Why question and how to determine the answer of the Why question where all Why questions of the research are based on explanation. Therefore, the Maximum Entropy classifier was proposed to identify the Why question. The research applies the causality graph for determining the visualized answers based on the information retrieval technique. The experiment shows the Why Question Answering system of the research can achieve answers at Rank 1 with 90% correctness.**

***Index Terms*— Why-Q, visualized answer, causality graph**

I. Introduction

IN the online community, most people prefer to post their problems or queries on a certain thread on their community's web page. Then, they wait for several minutes to several days to receive the answers posted by the experts for consulting their problems on the web page. However, it is time consuming for people to receive the answers. For example: According to the countryside community, there are some beginning farmers or other people in this generation know well how to use the information technology but lack experience in a certain area, e.g. Agriculture, Health-Care, and etc. They confront their problems of disease symptom occurrences by explaining their problems including a *Why* question, asking for a reason, and/or a *How* question, asking for problem solving approach, on the community web-boards. However, this research concerns only the *Why* question with explanation for problem diagnosis first, then followed by the next research of providing the solution of solving their posted problems through the *How* question answering system. Therefore during the waiting time of receiving the expert answers, an automatic *Why* Question-Answering (Why-QA) system is introduced for providing a preliminary diagnosis before or during an epidemic. Thus, this research aims to develop the Why-QA system based on questions with explanation of problems, i.e. plant-disease symptoms, on a web board. And, the corresponding answers are determined by the visualized causality graphs [1] for the preliminary diagnosis problems of plant disease symptoms. According to (http://class.uark.edu/), there are two kinds of the *Why* questions (or Cause and effect questions): "first, the question that gives you a *cause* and asks you to trace the probable *effects* of that cause; and Second, the question that gives you an *effect* and asks you to discuss or analyze the probable *cause*(*s*) of that effect".

The *Why* questions with explanation on the community web board are expressed in the form of Elementary Discourse Units (where an EDU is defined as a simple sentence or a clause, [2]) with the following question patterns (called 'Qpattern').

Qpattern-1: EDU_{ct-1} EDU_{ct-2} …EDU_{ct-n} EDU_q

Qpattern-2: EDU_{ct-1} EDU_{ct-2} …EDU_{ct-n} EDU_q $EDU_{ct-(n+1)}$

Qpattern-3: EDU_q EDU_{ct-1} EDU_{ct-2} …EDU_{ct-n}

where: EDU_q is a question EDU containing a question word (*qw*) as shown in the following linguistic pattern of a Thai-question EDU.

EDU_q→Qword NP1 V NP2| Qword NP1 V| NP1 V Qword
| NP1 V NP2 Qword | V NP2 Qword| V Qword

V→ v_q | pre-verb v_q

v_q→ $v_{q\text{-Strong}}$ | $v_{q\text{-weak}}$ w_{info}

$v_{q\text{-Strong}}$→ 'แสดง/*express*' 'เกิดจาก/*be caused by*' 'แห้ง/*dry*' 'ร่วง/*come off*' 'แกระแกรน/*stunt*' 'หงิก/*change shape*' 'ทำ/*solve*' 'แก้/*solve*'…

$v_{q\text{-weak}}$→ 'เป็น/*be*' 'มี/*have*'

w_{info}→'อาการ/*symptom*' 'แผล/*mark*' 'สี/*color*' 'เพราะ/*reason*' 'สาเหตุ/*cause*' 'ผลลัพธ์/*result*'…

Qword → {'ทำไม/*Why*' 'อย่างไร/*How*' 'อะไร/*What*' }

pre-verb → 'จะ/*will*' 'ต้อง/*must*'…

(where Qword is a question-word set and $qw \in$ Qword; v_q is a verb concept expressed on EDU_q; NP1 and NP2 are noun phrases.)

EDU_{ct-a} is a content EDU expressing a content of EDU_q, where a=1,2,..,n or n+1. n is an integer number and is greater than 0. EDU_{ct-a} has the following Thai linguistic pattern.

EDU_{ct-a}→ NP1 VP

VP → v_{ct-a} NP2 | v_{ct-a} | v_{ct-a} AdjectivePhrase | pre-verb v_{ct-a} | pre-verb v_{ct-a} NP2 | pre-verb v_{ct-a} AdjectivePhrase

(where v_{ct-a} is a causative verb concept (v_c) or an effect verb concept (v_e) as shown in Table I ($v_c \in V_c$; $v_e \in V_e$; V_c and V_e are a causative verb concept set and an effect verb concept set respectively)) Moreover, the Thai documents have several specific characteristics, such as zero anaphora or implicit noun phrases, without word delimiters, and without sentence delimiters (e.g. without a question mark), as shown in Fig.1.

Manuscript received December 27, 2015; revised January 20, 2016. This work was supported in part by the Thai Research Fund Grant MRG5580030.

Chaveevan Pechsiri is with DhurakijPundit University, Bangkok, Thailand. She is now with Department of Information Technology (corresponding author phone: 662-954-7300; e-mail: itdpu@hotmail.com).

TABLE I
LIST OF V_c AND V_e PROVIDED BY [1]

Verb type		Surface form	Conceptual class
V_c (Causative-Verb Concept set)	Strong Verb	ดูด/*suck*, ดูดกิน/*suck. กิน/eat, กด/bite,*	consume/destroy
		ทำลาย/destruct, กำจัด/eliminate, ฆ่า/kill, หด/break,	destroy
	Weak Verb +Noun or Information	*เป็น+โรค/ be+ disease,*	getDisease
		ได้รับ+เชื้อโรค/get+ pathogen,	getPathogen
		.....	.....
V_e (Effect-Verb Concept set)	Strong Verb	*หงิก /shrink, งอ/bend, บิด/twist, ใส้งอ/curl*	beAbnormalShape
		แห้ง/dry, ไหม้/blast,	dry/beSymptom
		เหี่ยว/wilt	loseWater/beSymptom
		แคระแกรน/stunt	stunt/beSymptom
	Weak Verb +Noun or Information	*เป็น+จุด /be+spot, เป็น+ขีด /be+scratch, เป็น+แผล /be+ lesion*	beSpotMark / beSymptom, beScratchMark / beSymptom beMark / beSymptom
		มี+จุด /have+spot, มี+ขีด /have+scratch, มี+แผล /have+ lesion	haveSpotMark / haveSymptom haveScratchMark/haveSymptom haveMark / haveSymptom
		มี+สี+น้ำตาลไหม้/have+color+dark brown	haveBrownColor/ haveSymptom
		.....	.....

Qpattern-1

EDU_{ct-1}: "*ระยะแตกกอ*(*Tillering Stage*): *ใบข้าว*(*rice leaves*)/NP1 *หงิกงอ*(*shrink* |beAbnormalShape)/v_{ct} "
(*ระยะแตกกอ*: *ใบข้าวหงิกงอ*/***Tillering Stage: Rice leaves shrink.***)

EDU_{ct-2}: "*ต้น*(*plant*)/NP1 *แคระแกรน*(*stunt*|stunt)/v_{ct}"
(*ต้นแคระแกรน*/***Plant stunts.***)

EDU_q: " *เป็นเพราะ*(*be reason*)/v_q *อะไร*(what)/***qw***"
(*เป็นเพราะอะไร*/***What are the reasons?***)

Qpattern-2

EDU_{ct-1}:"*ระยะแตกกอ*(*Tillering Stage*): *ใบข้าว*(*rice leaves*)/NP1 *หงิกงอ*(*shrink*|beAbnormalShape)/v_{ct}"
(*ระยะแตกกอ*: *ใบข้าวหงิกงอ*/***Tillering Stage: Rice leaves shrink.***)

EDU_{ct-2}: " *ต้น*(*plant*)/NP1 *แคระแกรน*(*stunt*|stunt)/v_{ct}"
(*ต้นแคระแกรน*/***Plant stunts.***)

EDU_q: "[*เรา*(*we*)/NP1] *จะ*(*should*)/*pre-verb ทำ*(*solve*|solve)/v_q *อย่างไร*(*how*)/***qw***"
([*เรา*] *จะทำอย่างไร*/***How should [we] solve?***)

$EDU_{ct-(n+1)}$: "*ต้น*(*plant*)/NP1 *จึงจะ*(*will*)/ *pre-verb แข็งแรง* (*be strong*|beStrong)/v_{ct}"
(*ต้นจึงจะแข็งแรง*/***Plants will be strong.***)

Qpattern-3

EDU_q: "*ทำไม* (*Why*)/***qw*** *ใบพืช*(*plant leaves*)/NP1 *มีแผล* (*have scar*|haveMark)/v_q *สีน้ำตาล* (*brown*) "
(*ทำไมใบพืชมีแผลสีน้ำตาล*/***Why do plant leaves have brown lesions?***)

EDU_{ct-1}: "*แผล*(*lesions*)/NP1 *เป็น ขีด*(*are linear spots*|beScratchMark)/ v_{ct}"
(*แผลเป็นขีด*/***Lesions are linear spots***,)

EDU_{ct-2}: " [*แผล*(*Lesions*)/NP1] *กระจาย*(*spread on*|spreadOut) /v_{ct} *ทั่วใบ*(***whole leaves.***)/ NP2"
([*แผล*] *กระจายทั่วใบ*/[***Lesions*] *spread on whole leaves.***)

Where the '[..]' symbol means ellipsis of word(s) inside it.

Fig .1 Examples of question patterns (Qpattern)

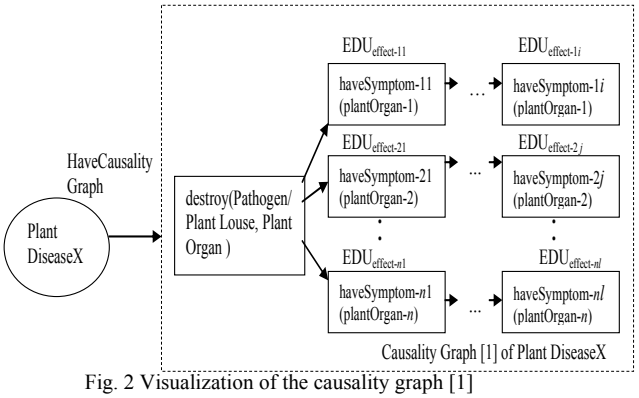

Fig. 2 Visualization of the causality graph [1]

All of these characteristics are involved in determining a *Why* question type and its answer of the QA system in this research based on several EDUs as the explanation question (Qpattern). It challenges in determining the answer from Qpattern whilst the previous QA researches, especially Why-QA system, are based on one to two EDUs. Several techniques of the Why-QA system [3]-[6] have been considered in this research (see section II). However, working Why-QA system must involve in two main problems:1) how to identify the Why-Q type with the question word ambiguity, 2) how to determine the corresponding answer of the explanation question as the Why-Q type (see section III). Therefore, the Maximum Entropy classifier including the linguistic phenomena was proposed to determine the *Why* question based on Qpattern. The research also applies the causality graphs of rice diseases (caused by Fungi, Virus, Bacteria, and Aphid) [1] (see Fig. 2) (http://www.web3point2.com/rice/indexApp .php) and the information retrieval (IR) technique for the *Why* answer determination.

In section II, related works are summarized. Problems of the Why-QA system are described in section III. Our framework of the Why-QA system is shown in section IV. We evaluate and discuss our proposed methodology and give a conclusion in section V.

II. Related Works

Other related works [3]-[6] to address several techniques, required for the Why-QA system, have been involved with Natural Language Processing, machine learning, and information retrieval approach. In 2003,Girju [3] worked on the *Why* question with the answer based on the lexico-syntactic pattern as 'NP1 Verb NP2' (where NP1 and NP2 are the noun-phrase expressions of a causative event and an effect event, respectively), i.e. "*What causes Tsunami*? → *Earthquake causes Tsunami*". However, it is not suitable for this research mostly based on several effect-event explanations which express by verbs/verb phrases. In 2007,Verberne et al.[4] proposed using RST (Rhetorical Structure Theory) structures to approach Why questions by matching a question topic with the nucleus in the RST tree while yielding an answer from the satellite. The RST approach to the Why-QA system achieved the answer correctness of 91.8% and a recall of 53.3%. In 2012,Baral et al. [5] developed a formal theory of answers to *Why* and *How* questions by developing the biological-graph model having event nodes and compositional edges as the knowledge-base with corresponding to *Why* and *How* questions on the biology domain. Their questions are based on the frame-base knowledge base in the forms: "How are X and Y related in the process Z?" and "Why is X important to Y. Where their answer expression is the event description graph based on frame base knowledge without having the implicit noun phrase or the NP ellipsis. In 2013,Oh et al.[6] used intra- and inter- sentential causal relations between terms or clauses as evidence for answering *Why* questions. Their answer candidates were obtained by answer candidate extraction with 83.2% precision of their causal relation recognition from Japanese web pages. They ranked their candidate answers with the ranking function including re-ranking the answer candidates done by a supervised classifier (SVM). Their Why-QA system achieves an average correctness of 41.4%.

However, most of previous researches on the Why-QA

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

system [3], [5] are based on a single sentence/one EDU of a *Why* question, and [4],[6] based on two EDUs of a *Why* question, whereas our research of Why-QA system are based on several EDUs including several zero-anaphora occurrences in the corpora.

III. Research Problems

There are two main problems: how to identify the *Why* question type based on Qpattern with the question word ambiguity and how to determine the corresponding answer of the Why question.

A. Question word Ambiguity

The problem of identifying the question expression without having the question mark symbol ('?') is solved by using a question word set {'*ทำไม/Why*', '*อย่างไร/How*' '*อะไร/What*', ..}. Where a '*ทำไม/Why*' function of a Why question is a reasoning question, a '*อะไร/What*' function of a What question is asking for information about something, and a '*อย่างไร/How*' function of a How question is asking about manner, condition, quality, or degree (http://www.englishclub.com/vocabulary/wh-question-words.htm). However, there is a question word's function ambiguity, e.g. '*อะไร/What*' as in reasoning and '*อย่างไร/How*' as asking for reason as shown in the following examples.

Example1: Causality What Question
EDU_{ct1} : "*ช่วงแตกกอใบข้าวหงิกงอ/* ***In the tillering stage, rice leaves shrink***."
EDU_{ct2}: " *ต้นไม่เติบโต/**Plants stunt***."
EDU_q : " *เป็นเพราะอะไร/**What are the reasons***?"

Example2: Causality How Question
EDU_{ct1}: "*อากาศไม่ได้ร้อนมาก/**The weather was not so hot***."
EDU_q: "*ต้นข้าวตายได้อย่างไร/ **How did the rice plant die***?"
where EDU_q is asking about the reason for sudden dying of the rice plant.

Therefore, we proposed using the Maximum Entropy classifier to identify the *Why* question type from two adjacent EDUs of EDU_q and $EDU_{ct\text{-}k}$ (where $k = 1$, n, or $n+1$ as in Qpattern). All features used in this classification consist of three feature sets: 1) Qword, 2) V_{ct} (where $V_{ct} = V_c \cup V_e$ and V_{ct} is a set of all verb concepts expressed on $EDU_{ct\text{-}a}$), 3) V_q (which is a set of all verb concepts expressed on EDU_q);

B. How to Determine Corresponding Why Answer

Unlike the question word sets from the factoid questions, the answer of the Why-Q cannot be determined by the question word. For example:
Factoid-Q:"*Who is the president of USA*?"
Ans:"***Obama is the president of USA***."
NonFactoid-Q: EDU_{ct1} "*ช่วงแตกกอใบข้าวหงิกงอ/**In the tillering stage, rice leaves shrink***."
EDU_{ct2} "*ต้นไม่เติบโต/**The rice plant stunts***."
EDU_q "*เป็นเพราะอะไร/**What are the reasons***?"
Ans: "*เพลี้ยกระโดดทำลายต้นข้าว/ **The Plant Hopper aphids destroy the rice plant***."
The answer of the Factoid question is solved by the *Who* question word [7] whereas the *Why* question word in Qpattern cannot be applied to determine the answer. Moreover, the *Why* question word have previously been approached by determining the corresponding *Why* answer based on the question EDU having a causal verb [3] or noun phrases with a question word [4], which is not suitable for the research's *Why* question based on explanations by having several effect-events (EDUs). Therefore, we solve the answers of causes or the answers of effects for the *Why* question with Qpattern by ranking the candidate answers of causes or the candidate answers of effects from the number of matching EDUs based on the similarity-score through the matrix of cause-effect-EDU vectors represented by the causality graphs (such as the rice disease causality graph, see Fig.2). The similarity-score is determined among $EDU_{ct\text{-}a}$ of the content EDU vector and $EDU_{effect\text{-}b}$ of all cause-effect-EDU vectors (see section IV.*D*) after the stop word removal. All word concepts in the similarity score determination are based on WordNet and Thai Encyclopedia after using Thai-to-English dictionary (http://longdo.com).

IV. Framework of Why-QA System

The Why-QA system of this research consists of four major steps, Question Corpus Preparation, Learning of *Why* question type on Qpattern , *Why* question type Identification, and Answer Determination, as shown in the lower part of Fig. 3.

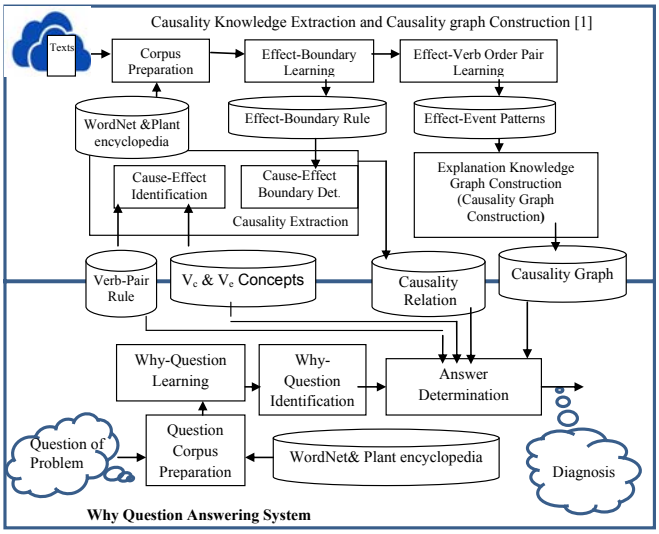

Fig.3 System Overview

A. Question Corpus Preparation

The preparation of the question's corpora with 8000 EDUs downloaded from the web-boards of three online community websites; a farmer community website based on plant diseases (www.kasetporpeangclu.com), a health-care community website (http://haamor.com), and a technology-and-indigenous-technology community website (http://www.gotoknow.org/posts/325634). Each community has 650 downloaded questions which are separated into two parts, the first part of 500 questions for learning the question types based on ten fold cross validation and the second part of 150 questions for testing. All of these questions include using Thai word segmentation tool which includes tagging the part of speech [8], and solving Named Entity [9]. EDU segmentation [10] is then to be dealt with to generate EDUs for the semi-automatic annotation (based

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

on experts) of question type concepts, a causative-verb concept (v_c) and an effect-verb concept (v_e) as shown in Fig. 4. Where the causative-verb concept set (V_c having $v_c \in V_c$) and the effect- verb concept set (V_e having $v_e \in V_e$) are provided by [1] shown in Table I used for identifying a causative EDU and an effect EDU respectively. All concepts from Table I are referred to Word Net [11] (http://wordnet.princeton.edu/obtain) and Thai Encyclopedia of plant disease (http://kanchanapisek. or.th/kp6/) after using the Thai-to-English dictionary .

EDUct-1: "ใบข้าว/*Rice leaves* หงิก/*shrink* งอ"
("***Rice leaves shrink***.")
EDUct-2: "ต้น/*Plants* แคร็ แกรน /*stunt* ช่วงแตกกอ/*at the tillering stage*"
("***Plants stunt at the tillering stage***.")
EDUq: "[อาการเหล่านี้/*these symptom*] เกิดขาก /*are caused by* อะไร/*what*"
("***What causes* [*these symptoms*]?")

<**EDUct-1**>[ใบ/ncn ข้าว/ncn]/NP
[<**Qfocus**><**Vct**: **Ve-concept**='shrink/be_abnormal_shape'>หงิก/vi</**Vct**> งอ/adv
</**Qfocus**>]/VP
</**EDUct-1**>
<**EDUct-2**> [ต้น/ncn]/NP
[<**Qfocus**> <**Vct**: **Ve-concept**= 'stunt'>แคระแกรน/vi </**Vct**></**Qfocus**> [ช่วง/ncn แตก/
vi กอ/ nct]/NP]/VP
</**EDUct-2**>
<**EDUq**> [ϕ=thses symptoms]/NP
[<**Vq** : concept= 'be caused by'>เกิดจาก/vt </**Vq**>
<**Qword**=What: **concept=why-Q**> อะไร/pint </**Qword**>]/VP
</**EDUq**>

Where: a 'Qfocus' tag is a question focus tag. A 'Vct' tag is a verb tag of a content EDU and has three verb concept sets for selection, a causative verb concept set, V_c , an effect verb concept, V_e , and the other verb concept set, V_{other}. A 'Vq' tag is a verb tag of an EDU containing the question word. A 'Qword' tag is a question word tag. An EDUct tag is an EDU content tag. An EDUq tag is a tag of an EDU having the question word. And, the symbol 'ϕ' represents a zero anaphora or ellipsis.
The [..] symbol means ellipsis.

Fig. 4 An example of the question annotation

B. *Learning of Why question type on Qpattern*

The Maximum Entropy classifier is applied to learn the question type with two classes: the *Why* question class and the Other question class from the annotated question corpora based on Qpattern by using Weka (http://www.cs.waikato .ac.nz/ml/weka/). The feature sets used in these learning techniques are Qword, V_{ct} and V_q (Qword is a question-word set and $qw \in$ Qword; V_{ct} is the verb concept set existing on $EDU_{ct\text{-}a}$, $v_{ct\text{-}a} \in V_{ct}$, $V_{ct}=V_c \cup V_e$; V_q is the verb concept set existing on EDU_q and $v_q \in V_q$). These three feature sets from two adjacent EDUs, EDU_q and $EDU_{ct\text{-}k}$ (where k=1 or n or n+1) from the annotated corpora are used in learning the question type classification by Maximum Entropy.

Maximum Entropy(ME) ME model will be the one that is consistent with the set of constrains imposed by the evidence, but otherwise is as uniform as possible [12], [13]. They modeled the probability of a semantic role r given a vector of features x according to the ME formulation below:

$$p(r \mid x) = \frac{1}{Z_x} \exp[\sum_{j=0}^{n} \lambda_j f_j(r,x)] \qquad (1)$$

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 λ_j is the weight for a given feature function. According to (1), ME can be used as the classifier of the r class when $p(r \mid x)$ is the highest probability or argmax $p(r \mid x)$ to determine three question-type classes. Where r is the question-type class value (class1='Why-Q' if r=1, class2='Other-Q' if r=2) and x is the binary vector consisted of all consecutive elements of three feature sets: Qword, V_{ct}, and V_q , from EDU_q and $EDU_{ct\text{-}k}$ as shown in (2).

$$p(r \mid x) = \operatorname*{argmax}_r \frac{1}{z} \exp\Big(\sum_{j=1}^{n} \lambda_j f_{class1,\ ct-k,\ j}(r, v_{ct-k}) + \sum_{j=1}^{n} \lambda_j f_{class2,\ ct-k,\ j}(r, v_{ct-k}) + \sum_{j=1}^{n} \lambda_j f_{class1,\ q,\ j}(r, v_q) + \sum_{j=1}^{n} \lambda_j f_{class2,\ q,\ j}(r, v_q) + \sum_{j=1}^{n} \lambda_j f_{class1,\ qw,\ j}(r, qw) + \sum_{j=1}^{n} \lambda_j f_{class2,\ qw,\ j}(r, qw) \qquad (2)$$

C. *Why question type Identification*

All weights from the previous learning step by ME are used to identify the question types. We use λ_j (the weight for a given feature function of the binary vector) resulted from learning Why-Q and Other-Q to identify the question type classes by (2) as shown in the algorithm of Fig. 5.

Assume that each EDU is represented by (NP VP). L is a list of EDUs with Qpattern.
$EDU_q \rightarrow$ Qword NP1 v_q NP2 | Qword NP1 v_q | NP1 v_q NP2 Qword | NP1 v_q Qword | v_q NP2 Qword | v_q Qword
v_q is a verb concept expressed on EDUq ; $qw \in$ Qword
$EDU_{ct\text{-}k} \rightarrow$ NP1 v_{ct} NP2 | NP1 v_{ct} | v_{ct} NP2
v_{ct} is a causative verb concept or an effect verb concept where k=1or n or n+1

QUESTION_TYPE_DETERMINATION (L)

```
1  i ← 1, flagQ←0, count←0 , class←2
2    count= length[L] / the number of EDUs in Qpattern
3  while i ≤ length[L] and flagQ=0do
4  { If qw_in_EDUi /*find the Question EDU or EDUq
5    { flagQ=1
6      If i=1 then { EDUi+1 is EDUct-1 };
7      If i=count-1 then { EDUi-1 is EDUct-n and EDUi+1 is EDUct-(n+1)}
8      If i=count then { EDUi-1 is EDUct-n } }
9    i++ }
10   If flagQ=1
11       /* The features of ME are based on verb concepts.
         /*If the serial verbs of the Thai Language occur in EDUq or
          EDUct, the concept of the first verb in the serial verbs is
          considered as one of those feature
       Equation (2) /*ME
12   If r=1      /* result from Equation (2)
13     class←1 /* The question type is 'Why Question'
14   Return
```

Fig. 5 Algorithm of Identifying the Why question based on Qpattern by ME

D. *Why Answer Determination*

The visualized answers of the *Why* questions are randomly applied on the plant disease domain, especially the rice diseases, through the integrated causality graph. According to our research, the focuses of the *Why* questions with Qpattern are based on the events expressed by v_{ct} which is v_c or v_e. The 50 correct *Why* questions for rice diseases are randomly selected from the 418 correct-question-type identification from section IV.*C*. Each selected question is used for determining its answers based on the Information Retrieval technique approach by ranking its candidate answers from TotalSimilarity_Score values. Each TotalSimilarity_Score value (5) [14] is determined by EDUmatching amoung the content EDU vector of the *Why* question and the cause-effect-EDU vectors in the repository. E_i is an effect-concept EDU set (a symptom-concept EDU set) of $Disease_i$ $\{EDU_{effect\text{-}1}, EDU_{effect\text{-}2},..,EDU_{effect\text{-}m}\}$

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$\eta = \bigcup_{i=1}^{\alpha} E_i \quad (3)$$

$$Similarity\_Score = \frac{|S1_a \cap S2_{ij}|}{\sqrt{|S1_a| \times |S2_{ij}|}} \quad (4)$$

$$TotalSimilarity\_Score = \sum_{1}^{\eta} Similarity\_Score \quad (5)$$

η is the number of different effect-concept EDUs (the number of different symptom-concept EDUs).

α is the number of different diseases.

where: $S1_a$ is an $EDU_{ct\text{-}a}$ of the content EDU vector (having a=1,2,..,n or n+1) after eliminating stop words.

$S2_{ij}$ is an $EDU_{effect\text{-}b}$ (having b=1,2,..,m; $m \leq \eta$; j=b) of the cause-effect-EDU vector $\langle EDU_{cause}, EDU_{effect\text{-}1}, EDU_{effect\text{-}2},.., EDU_{effect\text{-}m}\rangle$ of $Disease_i$ after the stop word removal.

In addition, our research emphasizes on two kinds of the *Why* question, Cause-Why-question and Effect-Why-question. Where Cause-Why-question is a *why* question of determining the root cause of effects/problems (e.g. "*ใบข้าวมีจุดสีน้ำตาล*/*The rice leaves have brown spots*. *และจุดอยู่กระจายทั่วทั้งใบ*/*And*, *the spots spread over the leaf*. *เป็นเพราะอะไร*/ *What is the cause?*"), and Effect-Why-question is a *why* question of determining the results/effects of the root cause (e.g. "*เพลี้ยกระโดดสีน้ำตาลปรากฏที่นาจำนวนมาก*/*Brown Planthopper fully occurs over a rice field* . *ต้นข้าวจะแสดงอาการอะไรบ้าง*/*What symptoms will rice plants show up?*"). According to the Cause-Why-question type, each TotalSimilarity_Score value is determined between $EDU_{ct\text{-}a}$ of the content EDU vector and $EDU_{effect\text{-}b}$ of each cause-effect-EDU vector. If the *Why* question is the Effect-Why-Q type, each TotalSimilarity_Score value is determined between $EDU_{ct\text{-}a}$ of the content EDU vector and EDU_{cause} of each cause-effect-EDU vector.

All word concepts of $S1_a$ and $S2_{ij}$ are based on WordNet and Thai Encyclopedia after using the Thai-to-English dictionary. The number of words in $S1_a$ and the number of words in $S2_{ij}$ are not significantly different. If Similarity_Scores($S1_a$, $S2_{ij}$) in equation (4) is calculated by having $|S1_a \cap S2_{ij}|$ =1 with one matched word concept of plant organ, e.g. 'leaf', 'seed', 'flower',..etc., it will result in this Similarity_Scores($S1_a$, $S2_{ij}$) value becoming to zero because there is no matching concept of an effect/symptom event of an $EDU_{effect\text{-}b}$. The Similarity_Score ($S1_a$, $S2_{ij}$) values of $Disease_i$ are collected for ranking the candidate answers of the cause-effect-EDU vectors for the answer selection. For example: the following Qpattern-1 of the Cause-Why-question type is expressed with all word concepts after stop word removal as follow.

Qpattern-1: $EDU_{ct\text{-}1} \rightarrow S1_1$, $EDU_{ct\text{-}2} \rightarrow S1_2$, ..., $EDU_{ct\text{-}n} \rightarrow S1_n$, EDU_q

where $EDU_{ct\text{-}1} \neq EDU_{ct\text{-}2} \neq \ldots \neq EDU_{ct\text{-}n}$

$EDU_{ct\text{-}1}$: "*ใบ*(leaf)/NP1 *มีแผลจุด*(have_spot_mark)/ v_{ct} *สีน้ำตาลไหม้*(brown)/adj"
(haveBrownSpotMark(leaf))

$EDU_{ct\text{-}2}$:"*แผล*(mark)/NP1 *เป็นรูป*(be_shape)/v_e *คล้าย*(*alike*)*ตา*(eye))/adjphrase"
(beAlikeEyeShape(mark))

$EDU_{ct\text{-}3}$: "*แผล*(mark)/NP1 *กระจายทั่ว*(spread)/v_e *ใบ*leaf)/NP2"
(spread(mark,leaf))

$EDU_{ct\text{-}4}$: "*ทั้งต้น*(plant)/NP1 *แห้ง*(dry)/v_e"
(dry(plant))

EDU_q:"(*เป็นเพราะ*/be_reason)/v_q (*อะไร*/what)/pint" (What is the reason?)

The candidate answers are ranked by sorting the TotalSimilarity_Score values (see Table II) Then, the possibility answer can be solved by the selection of the cause-effect-EDU vector that has Rank 1 (which is the highest rank) of the TotalSimilarity_score value from the EDU matching . From Table II, the answer having the highest rank is the cause-effect-EDU vector with Disease2 (Rank1).

V. Evaluation and Conclusion

The question corpora for evaluating the proposed model of identifying the *Why* question type with explanation contain 450 questions equally collected from the three community web-boards with different domains, the plant-disease domain, the health-care domain, and the technology-and-indigenous-technology domain. The 50 questions of rice diseases from the correct-question-type identification are randomly selected for the answer evaluation based on IR approach. All corpora categories are emphasized on events expressed by verbs/verb phrases having different characteristics, e.g. the number of different verb features and feature dependencies. All of these characteristics make this research analyze how verb features effect to the results of using ME for question identification.

TABLE II
Ranking candidate answers form cause-effect-EDU vectors for Qpattern with α =13 and η=69

$Disease_i$	EDU_{cause}	$EDU_{effect\text{-}b}$ **:Similarity_Scores Determination** where a=1,2,..,n or n+1; i=1,2,..,α j=1,2,..,η						**Total Similarity_ Score (TSC)**	**Rank by sorting TSC**
		$S2_{i1}$ haveBrownSpotMark (leaf)	$S2_{i2}$ beAlikeEye Shape(mark)	$S2_{i3}$ Dry (plant)	$S2_{i4}$ beYellow Color(leaf)	…	$S2_{i\eta}$		
$Disease_1$: Stunt disease	Destroy(RaggedStuntVirus, plant)	0	0	0	0	…	0	0	0
$Disease_2$: RiceBlastDisease	destroy(RiceBlastFungus ,plant)	1	1	1	0	…	0	3	1
$Disease_3$: RiceBrownSpot disease	destroy(Brown_sput_fungus_of_ Rice, plant)	1	.0	0	0	…	0	1	2
…..	…	…	…	…	…	…	…	....	...
$Disease_\alpha$	…	0	0	0	0	…	0	0	0

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The evaluation of the *Why* question identification by this research is expressed in terms of the precision and the recall based on three experts with max wins voting. The *Why* questions with Qpattern are based on several events expressed by verbs or verb phrases which are used as the main features for the *Why* question identification by ME. Table III shows ME results in the highest precision of 0.930 for the health-care domain containing more feature dependency occurrences.

TABLE III
CORRECTNESS OF WHY-Q AND HOW-Q IDENTIFICATION

Domain (Each domain contains 150 questions)	#of Feature-Dependency Occurrences ($v_{ct\text{-}k}$-v_q-qw)	#of Different Verb Features (Diversity)	ME	
			Pre-cision	Re-call
PlantDisease	medium	89	0.910	0.827
HealthCare	medium	98	0.930	0.838
IndigenousTechno.	low	115	0.891	0.805

The news domain of technology contains the highest diversity of verb feature occurrences (result in low frequency of verb feature occurrences) and the lowest feature dependency occurrences, which result in the lowest precision of 0.891. However, the average precision of the *Why* question identification by ME is 0.910 with the average recall of 0.828. Whereas [15] applied five different machine learning algorithms, the Support Vector Machines (SVM), Nearest Neighbors (NN), Naïve Bayes (NB), Decision Tree (DT), and Sparse Network of Winnows (SNoW) , with the noun-based features to classify several question types, and each question type was occurred within one sentence. The average %correctness of their question classification [15] is 75% whilst SVM outperforms with 87.4% correctness. Moreover, the SVM algorithm is not concerned in the Why question identification of our research because the plant-disease domain and the health-care domain contain the verb-feature-dependency occurrences, e.g. “Plant stunts. What is the cause?” --->stunt/symptom - be_cause/be_a_reason, Then, we currently apply the SVM classifier to identify the *Why* question, with the following precisions as 0.879, 0.881, and 0.889 and the following recalls as 0.779, 0.796, and 0.801 for the following domains as PlantDisease, HealthCare, and IndigenousTechnology respectively.

The evaluation of the answer determination by the proposed methodology of applying the IR approach to the causality graph repository is expressed in term of the percentage of correctness based on the answer set proved by experts with max win voting. Therefore, the % correctness of the *Why* answer determination by this research is 90% based on the rice disease question and the rice-disease causality graph. Moreover, the zero anaphora occurrence on an $EDU_{ct\text{-}a}$ effects to the % correctness of the visualized answers of the *Why* question type. Most of the previous works of the *Why*-QA system have their questions based on one sentence except [6], [16]. Their answer method of Why-QA [6] was based on finding text fragments the web documents that include intra-and inter-sentential causal relations with an effect part that resembled a given why question (by using SVM with a linear kernel) and provided them as answers. [6] achieved the answer correctness of 41.8% (precision of the top answer) from the why questions extracted from the Japanese version of Yahoo! Answers and also created by annotators without several event explanation as in Qpattern. Whilst [16] works on interpreting of consumer health questions (which are the explanation question) without solving the answers. Whereas the *Why* questions of our research based on Qpattern with explanation of problems, e.g. plant-disease symptoms, are collected from the community web-boards after the misspelling-word correction, and the *Why* answers are determined by ranking the TotalSimilarity_Score values of the candidate answers from the knowledge repository of cause-effect-EDU vectors of the plant disease symptoms represented by the causality graph. Thus, the problem of zero anaphora occurrences should be solved in the future work for increasing the correctness of answers. Finally, the model of our Why-QA system can be applied not only by the people on the online community but also by the other on the business and financial industries.

REFERENCES

[1] 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.

[2] L.Carlson, D.Marcu, and ME.Okurowski,“Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory,” *In Current and New Directions in Discourse and Dialogue*,vol.*22*,2003, pp.85-112.

[3] R.Girju,“Automatic detection of causal relations for question answering,”*in Proc. 41st annual meeting of the assoc. for computational linguistics, workshop on multilingual summarization and question answering-Machine learning and beyond*, Japan, 2003,pp.76-83.

[4] S.Verberne,L.Boves,P-A.Coppen,and N.Oostduk,”Discourse-based answering of why-questions. *Traitement Automatique des Langues*," vol.47,no.2, 2007,pp.1-41.

[5] C. Baral, NH. Vo, and S.Liang, “Answering Why and How questions with respect to a frame-based knowledge base: a preliminary report,” in *Proc. 28th International Conference on Logic Programming ICLP 2012*, Hungary, 2012, pp.26-36.

[6] J-H. Oh, K.Torisawa, C.Hashimoto, M.Sano , SD.Saeger,andK.Ohtake, “Why-Question Answering using Intra-and Inter-Sentential Causal Relations,” in *Proc. 51st* Annual Meeting of the Association for Computational Linguistics, Bulgaria,2013,pp.1733-1743.

[7] E.Agichtein, S.Cucerzan,and E. Brill, “Analysis of Factoid Questions for Effective Relation Extraction,” in *Proc. 28th annual international ACM SIGIR conference on Research and development in information retrieval*, Salvador, Brazil, 2005, pp.567-568.

[8] S.Sudprasert, and A.Kawtrakul, “Thai Word Segmentation based on Global and Local Unsupervised Learning,” in *Proc.7thNational Computer Science and Engineering Conference*, Thailand;2003, pp.1-8.

[9] H.Chanlekha,and A.Kawtrakul,“Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information,” in *Proc.*1st *International Joint Conference IJCNLP*’2004, Hainan Island, China, 2004, pp.1-7.

[10] J. Chareonsuk , T. Sukvakree, and A. Kawtrakul, “Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information,” in *Proc.* 9th *National Computer Science and Engineering Conference*, Bangkok, Thailand,2005, pp.85-90.

[11] G A. Miller, “WordNet: A lexical database,” *Communication of the ACM*, vol.38,no.11, 1995, pp.39-41.

[12] AL.Berger, SA.Della Pietra, and VJ.Della Pietra, “A Maximum Entropy approach to natural language processing,” in *Computer Linguistics*, vol.22,no.1, p39–71, 1996.

[13] M.Fleischman, N.Kwon, and E.Hovy, “Maximum Entropy models for Frame Net classification,” in *Proc.20th international conference on conference on Empirical*, Sapporo, Japan ,2003,pp.49-56.

[14] S.Biggins,Mohammed,andS.Oakley,“UniversityOfSheffield:Two ApproachestoSemanticText.”Similarity,”First*JointConferenceOnLexicalAnd ComputationalSemantics,Montre´al,*Canada,2012,pp.655-661.

[15] D.Zhang, andW.S. Lee,“Question Classification using Support Vector Machines,” in *Proc.*26th*Annual International ACM SIGIR* on *Research and Development in Information Retrieva*l, Canada,2003,pp26-32.

[16] H.Kilicoglu, M.Fiszman, and D.Demner-Fushman, “Interpreting Consumer Health Questions: The Role of Anaphora and Ellipsis,”in *Proc. the 2013 Workshop on Biomedical Natural Language Processing* (BioNLP 2013), Sofia, Bulgaria, 2013,pp.54–62.