Available at www.sciencedirect.com

INFORMATION PROCESSING IN AGRICULTURE 3 (2016) 36–53

journal homepage: www.elsevier.com/locate/inpa

Developing a Why–How Question Answering system on community web boards with a causality graph including procedural knowledge

C. Pechsiri [a,*], R. Piriyakul [b]

[a] Department of Information Technology, Dhurakij Pundit University, Bangkok, Thailand
[b] Department of Computer Science, Ramkhamhaeng University, Bangkok, Thailand

ARTICLE INFO

Article history:
Received 12 June 2015
Received in revised form
15 January 2016
Accepted 19 January 2016
Available online 22 January 2016

Keywords:
Why-Q
How-Q
Visualized answer
Integrated causality graph

ABSTRACT

The research aims to develop an automatic Question Answering system, in particular *Why* and *How* questions, on community web-boards to support ordinary people in preliminary diagnosis and problem solving, such as plant disease problems. The research includes two main problems: *Why* and *How* question identification and *Why* and *How* answer determination, where Why and How questions are based on explanations. Therefore, the research applies machine learning techniques for question type identification. We also propose an integrated causality graph with extracted procedural knowledge from text to determine the visualized answers based on the information retrieval technique. The experiment shows the Question Answering system can achieve answers at Rank 1 with 91.1% and 88.9% correctness for Why questions and How questions, respectively.

1. Introduction

In the online community, most people prefer to post their problems or queries on a certain thread on their community's web page and then wait for times ranging from a few minutes to several days to receive the answers and recommendations made by the problem-solving experts on the web page. However, it is time consuming for people to wait for the answers. In a rural community, there are inexperienced farmers and others who know how to use information technology but lack experience in other areas, e.g. agriculture, health-care, etc. For example, on community web-boards, people with an illness try to explain their disease symptoms by asking a *Why* question (Why-Q) type, asking for reasons, and/or a *How* question (How-Q) type, asking for a problem solving approach. However, the speed of response to questions depends on the question domain, the chat room type of a certain web-board, the web-board domain, etc. Most plant disease questions receive responses within a week through web-boards. While waiting, an automatic *Why–How* Question–Answering (QA) system could be developed to provide a preliminary diagnosis including possible solutions before or during an epidemic. Therefore, this research aims to develop a *Why* and *How* QA system based on questions that require explanation of problems, especially plant-disease symptoms, on a certain web board. The corresponding answers are the visualized as causality graphs [1] integrated with procedural knowledge extracted from texts for the preliminary diagnosis and problem solving of plant disease symptoms. There are several types of *How* question [2] e.g. Causality How-Q (which is used to determine the causes of a certain event: "*How did*

* Corresponding author. Tel.: +66 2 954 7300.
E-mail address: itdpu@hotmail.com (C. Pechsiri).
Peer review under the responsibility of China Agricultural University.
http://dx.doi.org/10.1016/j.inpa.2016.01.002

John die?"), Instrumental How-Q (which is used to learn about instruments as in "*How is couscous eaten in Morocco?*", answer: "*by hand*"), and Instructional How-Q (which corresponds to an organized set of instructions designed to reach a goal: "*How do you change a car wheel?*"), etc. However, How-Q in this research is Instructional How-Q, which emphasizes the organized instruction set for problem solving which depends on the cause of the problems/symptoms. The *Why* and *How* questions with explanations are expressed in the form of Elementary Discourse Units (where each EDU is defined as a simple sentence or a clause, [3]) with the following question patterns (called 'Qpattern') through the community web board.

Qpattern-1: $EDU_{ct\text{-}1}$ $EDU_{ct\text{-}2}$ ...$EDU_{ct\text{-}n}$ EDU_q
Qpattern-2: $EDU_{ct\text{-}1}$ $EDU_{ct\text{-}2}$ ...$EDU_{ct\text{-}n}$ EDU_q $EDU_{ct\text{-}(n+1)}$
Qpattern-3: EDU_q $EDU_{ct\text{-}1}$ $EDU_{ct\text{-}2}$ ...$EDU_{ct\text{-}n}$

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 NP2 Qword | NP1 V 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}
pre-verb → 'จะ/*will*' 'ต้อง/*must*'
$v_{q\text{-Strong}}$ → 'ทำ/*solve*' 'แก้/*solve*' 'แสดง/*express*' 'เกิดจาก/*be caused by*' 'แห้ง/*dry*' 'ร่วง/*come off*' 'แคระแกรน/*stunt*' 'หงิก/*change shape*' ...
$v_{q\text{-weak}}$ → '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*' 'แสดงวิธี/*Show method*'}

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

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

$EDU_{ct\text{-}a}$ → NP1 VP
VP → $v_{ct\text{-}a}$ NP2 | $v_{ct\text{-}a}$ | $v_{ct\text{-}a}$ AdjectivePhrase |
pre-verb $v_{ct\text{-}a}$ NP2 | pre-verb $v_{ct\text{-}a}$ | pre-verb $v_{ct\text{-}a}$ AdjectivePhrase

(where $v_{ct\text{-}a}$ is a causative verb concept (v_c) or an effect verb concept (v_e) as shown in Table 1 (v_c∈V_c; v_e∈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, without sentence delimiters (e.g. without a question mark), etc as shown in Fig. 1.

All of these characteristics are involved in determining the question type and its answer in the *Why–How* QA system of this research based on Qpattern, which contains several EDUs as explanations. It attempts to determine the answer with Qpattern, whilst previous QA researches, especially *on Why–How* QA systems, were based on one or two EDUs. It also attempts to answer a How-Q which expresses only the sequence of events of the effect/symptom EDUs without mention of their cause. In this research, the How-Q expression results in diagnosing the effect/symptom events before determining the solution whereas previous How-Q researches are based on direct instruction guidelines or an event description graph without including problem/symptom diagnosis.

Table 1 – 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, กัด/bite, ทำลาย/destruct, กำจัด/eliminate, ฆ่า/kill, หัก/break,	consume/destroy destroy
	Weak Verb + Noun or Information	เป็น + โรค/be + disease, ได้รับ + เชื้อโรค/get + pathogen, ...	getDisease getPathogen ...
V_e (Effect-Verb Concept set)	Strong Verb	หงิก/shrink, งอ/bend, บิด/twist, โค้งงอ/curl	be_abnormal_shape
		แห้ง/dry, ไหม้/blast, เหี่ยว/wilt	dry/be_symptom lose_water/be_symptom
		แคระแกรน/stunt	stunt/be_symptom
	Weak Verb + Noun or Information	เป็น + จุด/be + spot, เป็น + ขีด/be + scratch, เป็น + แผล/be + lesion	be_spot_mark/be_symptom, be_scratch_ mark/be_symptom be_ mark/be_symptom
		มี + จุด/have + spot, มี + ขีด/have + scratch, มี + แผล/have + lesion	have_spot_mark/have_symptom have_scratch_mark/have_symptom have_ mark/have_symptom
		มี + สี + น้ำตาลไหม้/have + color + dark brown	have_brown_color/have_symptom
		...	...

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).

Several techniques of the Why-QA system and the How-QA system [4–8] have been considered in this research (see Section 2). Several techniques [9–12] have also been previously applied to extract procedural knowledge (see Section 2), where procedural knowledge is the knowledge about how to perform a specific task, such as how to remove a smoke detector [5]. However, a working Why–How QA system must involve two main problems: (1) how to identify the Why-Q and How-Q question types, where some question words are ambiguous, (2) how to determine the corresponding answers of the explanations questions for *Why* and *How* questions including How-Q with the effect/symptom explanation but without notifying the effect/symptom cause (see Section 3.2.3). It is necessary to know the effect/symptom cause or the disease name to determine the method sets to solve the effect/symptom events from textual data. All of these problems result in the research applying machine learning techniques including linguistic phenomena to solve the research problems. Therefore, different machine learning techniques such as Naïve Bayes (NB), Maximum Entropy (ME) and Multilayer Perceptron (MLP) are proposed to identify the question types of Why-Q and How-Q from two adjacent EDUs of EDU_q and EDU_{ct-k} (where a is k and $k = 1$, n, or $n + 1$ as in Qpattern). We then apply the word co-occurrence (Word-Co) with the procedural concept including Support Vector Machine (SVM) and ME to extract the procedural knowledge vectors, especially plant disease prevention and treatment, from downloaded documents from several websites, e.g. the Department of Agriculture website (http://www.doa.go.th/), involved with plant-disease problems. The research then integrates the extracted plant-disease prevention and treatment into the previously constructed causality graph of plant diseases, e.g. the causality graphs of rice diseases [1] (see Fig. 2) where [1] has been applied on (http://www.web3point2.com/rice/indexApp.php) to provide the causality knowledge with four categories of causing agent (Fungi, Virus, Bacteria, and Aphid). This integrated causality graph is used as the knowledge source to answer *Why* and *How* questions with explanations.

In addition, each causality graph [1] of plant diseases represents the extracted causality knowledge from documents on the Department of Agriculture website. The extracted causality knowledge with stop word removal has been kept in a repository as a cause-effect-EDU vector $\langle EDU_{cause}, EDU_{effect-1}, EDU_{effect-2}, \ldots, EDU_{effect-m}\rangle$ as shown in Example 1 of each disease under a certain causing agent category (where EDU_{cause} is a causative concept EDU, $EDU_{effect-b}$ is an effect concept EDU with $b = 1,2,\ldots,m$; and m is an integer number).

Example 1: Rice Brown Spot disease:

EDU_{cause}: "(*Bipolaris Oryzae*|virus)/NP1 ทำลาย(destroy)/v_c ใบและกาบใบข้าว(leaf and leaf-sheath)/NP2"
("*The Bipolaris Oryzae virus damages rice leaves and rice leaf sheaths*")

$EDU_{effect-1}$: "ระยะแตกกอ(Tillering Stage):ใบ(leaf)/NP1 มีจุด(have_spot_mark)/v_e สีน้ำตาล (*brown*)/*adj*"
("*Tillering Stage: Leaves have brown spots.*")

$EDU_{effect-2}$: "กาบใบ(leaf_sheath)/NP1 มีแผล (have_ mark)/v_e สีน้ำตาลไหม้(brown)/adj "
("*Leaf sheaths have dark brown lesions.*")

$EDU_{effect-3}$: "และ/*and* กาบใบ(leaf_sheath)/NP1 มีแผล (have_ mark)/v_e สีดำ(*black*)/adj"
("*Leaf sheaths have black lesions.*")

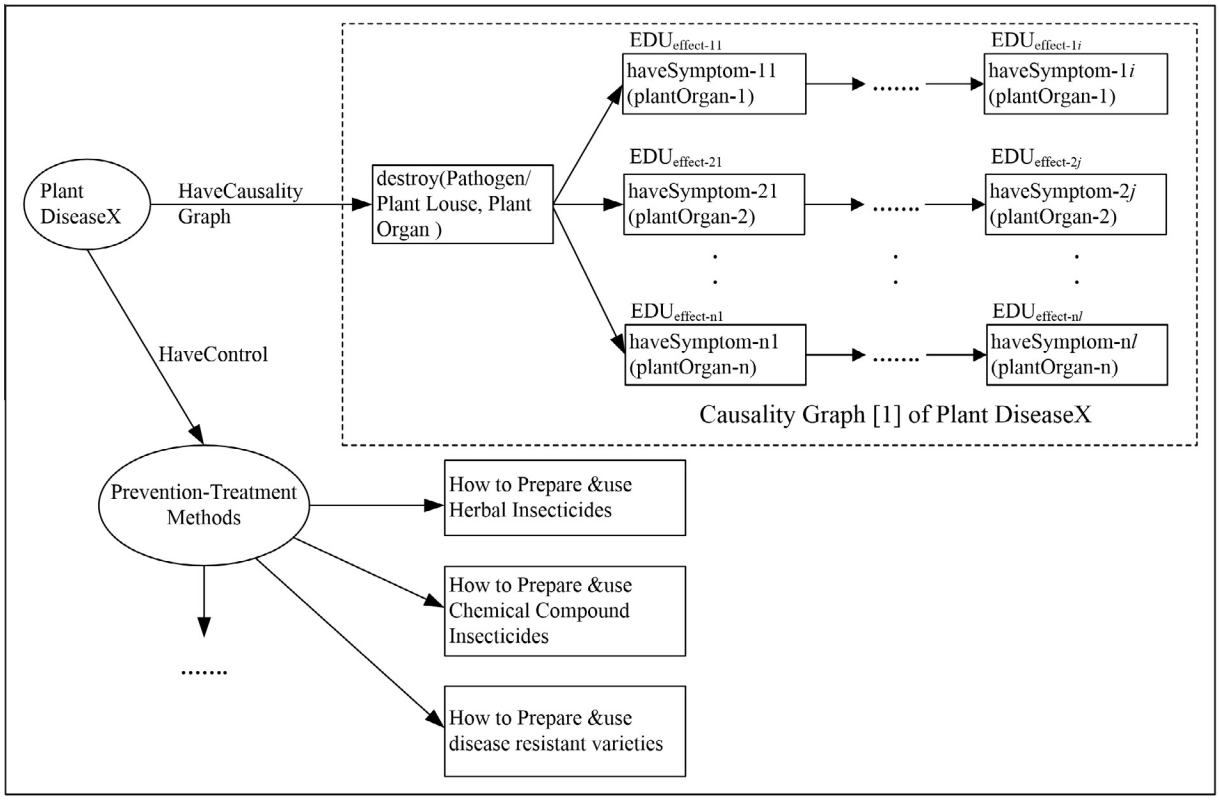


Fig. 2 – Visualization of the causality graph integrated with procedural knowledge.

$EDU_{effect-4}$: "แผล(mark)/NP1 เป็นรูป(be_shape)/v_e ไข่ (ellipse)/adj"
("*Lesions are an oval shape.*")

Finally, we determine the answers to Why-Q and How-Q by the number of matching EDUs based on the similarity-score determination between EDU_{ct-a} of the content EDU vector ($\langle EDU_{ct-1}\ EDU_{ct-2} \ldots EDU_{ct-k} \rangle$ where $k = n$ or $n + 1$, as shown in Example 2) and $EDU_{effect-b}$ of all cause-effect-EDU vectors of several diseases on the causality knowledge repository. Each cause-effect-EDU vector is equivalent to a certain causality graph [1] of a certain disease as shown in Fig. 2,

Example 2:

EDU_{ct-1}: "ใบ(leaf)/NP1 มีแผลจุด(have_spot_mark)/v_{ct} สีน้ำตาลไหม้(brown)/adj"
("*Leaves have dark brown spot lesions.*")

EDU_{ct-2}: "ต่อมา(*Then*) แผล(mark)/NP1 เป็นรูป(be_shape)/v_e (คล้าย(alike) ตา(eye))/adjPhrase"
("*Then, the lesions are eye shaped.*")

EDU_{ct-3}: "ตรงกลางแผล(middle_of_mark)/NP1 มีสี(have_color)/v_e เทา(grey)/adj"
(" *The middle of a Lesion has a grey color.*")

EDU_{ct-4}: "แผล(mark)/NP1 กระจายทั่ว(spread)/v_e ใบ(leaf)/NP2"
("*Lesions/spread over the leaf.*")

In Section 2, related works are summarized. Problems of the *Why–How* QA system and procedural knowledge extraction are described in Section 3. Our framework of the *Why–How* QA system including the integrated causality graph is shown in Section 4. We evaluate and discuss our proposed methodology in Section 5 and present conclusions in Section 6.

2. Related works

Other related works to address several techniques required for Why-Q and How-Q in our research and also for the procedural knowledge extraction, have involved Natural Language Processing, machine learning, and information retrieval approaches.

2.1. *Why–How QA system*

Most techniques from the previous approach to a QA system, especially a Why-QA system and a How-QA system, are Natural Language Processing (NLP), Machine Learning, Information Retrieval (IR), Knowledge Base, Rule Base, or mixed techniques. Girju [4] 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? → Earthquakes cause Tsunami*". However, it is not suitable for our research which is mostly based on several effect-event explanations which are expressed by verbs/verb phrases. Schwitter et al. [5] worked on the procedural questions/*How* questions with their answers being extracted from technical documents by the ExtrAns system. Their procedural answer is often expressed in a procedural

writing style with guidelines. High performance in their QA system is best achieved through logic-based and pattern-matching techniques. Verberne et al. [6] 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 answer correctness of 91.8% and recall of 53.3%. Baral et al. [7] 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 corresponding to *Why* and *How* questions on the biology domain. Their questions were 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?". Their answer expression is an event description graph based on frame base knowledge without having an implicit noun phrase or an NP ellipsis. Oh et al. [8] 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 by employing a supervised classifier (SVM). Their Why-QA system achieved an average correctness of 41.4%.

However, most of the previous researches on a *Why* QA system and a *How* QA system [4,5,7] are based on a single sentence/one EDU of a *Why* question and also a *How* question, except [6,8], which were based on two EDUs of a *Why* question, whereas our Why-Q and How-Q are based on several EDUs.

2.2. *Procedural knowledge extraction*

Several techniques have been applied to extract the procedural knowledge varying from one sentence/EDU to multiple sentences/EDUs with/without numbering in front of each step in the process. The extracted procedural knowledge from Web pages by [9] is based on HTML list tags, e.g. <OL>,<UL>, learned by SVM to determine the Procedural class. Delpech and Saint-Dizier [10] recognized the procedural knowledge by using HTML tags, e.g. <p>,<b>, and <h>, bold letters to identify the title/goal and by using a procedural writing style that contained the numbering form, hyphens or bullets in front of each process step to identify the procedure/instruction. There are several zero-anaphora occurrences in our corpora whilst our procedural knowledge is still based on verb or verb phrases whereas [11,12] involved noun phrases. In addition, the treatment and the prevention of our research are separated by their topic names. And, each document of our research describes several treatment-procedure sets for solving the same problem (the same target) and also several prevention-procedure sets. Each procedure set of either the treatment or the prevention contains several EDUs as process steps without the numbering form, hyphens or bullets in front of each process step. Most of the previous researches on procedural knowledge extraction from documents had different structure occurrences from our research. Therefore, we apply word co-occurrences and different machine learning techniques such as SVM and ME to extract procedural knowledge from texts to answer How-Q.

3. Research problems

This research work involves two major areas of problems: procedural knowledge extraction and the Why–How QA system.

3.1. *Problems of procedural knowledge extraction*

There are two main problems in procedural knowledge extraction: the first problem is how to identify the procedural knowledge from documents after identifying the target as the problem solution e.g. Prevention and Treatment of plant diseases. The target is identified by a target word pair, *tw1 tw2*, existing in either a topic name or an EDU in the plant disease documents (where *tw1* ∈ TW, and TW is a target word set collected from corpus study).

TW = {'ป้องกัน/prevent' 'รักษา/treat' 'ควบคุม/control' 'กำจัด/eliminate' 'การป้องกัน/prevention' 'การรักษา/treatment' 'การควบคุม/control''การกำจัด/elimination'...}
tw2 ∈Tname, and Tname is a target name set collect from corpus study
Tname = {'เพลี้ยกระโดดสีน้ำตาล/Brown Planthopper' 'เพลี้ยจักจั่นสีเขียว/Green Leafhopper' 'โรคใบไหม้/Blast Disease' 'โรคใบจุดสีน้ำตาล/Brown Spot Leaf disease '...}

The second problem is how to determine the procedural knowledge boundary.

3.1.1. *Procedural knowledge identification problem*

There are two problems: the implicit starting-procedural cue and the ambiguous starting-procedural cue.

3.1.1.1. *Implicit starting-procedural cue.* The starting procedural EDUs can be identified by using the starting-procedural cue set {'ดังต่อไปนี้/*the following*' 'ดังนี้/*as follows*' 'โดย/*By*'...} right after the target of the procedural knowledge has been identified by the target word pair, *tw1 tw2*, as shown in Fig. 3(a). Where the topic name containing *tw1* as 'การควบคุม/*control*' and *tw2* as 'โรคใบไหม้/*Blast disease*', is followed by EDU1 having the starting-procedural cue, 'โดย/*By*'or'ดังต่อไปนี้/*the following*', and then EDU2 as the starting-procedural EDU. According to Fig. 3 (b), EDU2 contains *tw1* as 'ควบคุม/*control*' and *tw2* as 'โรคใบไหม้/*Blast disease*'. And, there is an implicit starting-procedural cue, 'โดย/*By*', occurring in EDU3 as "[*By*] *using Bacillus*....", which results in the lack of ability to identify EDU3 as the starting-procedural EDU.

3.1.1.2. *Ambiguous starting-procedural cue.* There are some EDUs expressing as the non procedure even though they contain a starting-procedural cue, as shown in Fig. 4.

3.1.2. *Procedural knowledge boundary determination problem*

The problem is how to identify the ending of each procedure, especially where there is no cue, e.g. 'และ/*and*', 'หรือ/*or*',

(a) **Explicit Starting-Procedural Cue**

Topic-Name: "*การควบคุมโรคใบไหม้ข้าว/* ***Rice's Blast Disease Control*** "
EDU1: "*โดยใช้วิธีดังต่อไปนี้/* ***By using the following method***."
EDU2: "*ใช้พันธ์ต้านทานโรค/* ***Use the resistant varieties***"
EDU3: ……
where, EDU2 is the starting EDU of the procedural knowledge.

(b) **Implicit Starting-Procedural Cue**

EDU1: "[*โรคใบไหม้ข้าว*]*ระบาดที่ภาคเหนือ/* **[*The Blast-Rice disease*] *spreads out in the north part***."
EDU2: "*เจ้าหน้าที่กำลังควบคุมโรคใบไหม้ข้าว/* ***The expert is controlling the Blast-Rice disease***."
EDU3: "*ใช้เชื้อบาซิลลัสควบคุมโรคใบไหม้ข้าว/* ***Use Bacillus Subtilis to control the Blast-Rice disease***"
EDU4: ……
where EDU2 is the target, EDU3 is the starting EDU of the procedural knowledge.

Fig. 3 – Examples of explicit starting-procedural cue and implicit starting-procedural cue.

EDU1: "*วิธีทำสารชีวภาพกำจัดศัตรูพืชแบบชาวบ้านเป็นที่นิยมมาก/* ***The method of making indigenous biopesticides is very well known***."
EDU2: "*โดยใช้ต้นทุนเพียง500บาท/* ***By having cost only 500 Bath.*** "
Where, the cue '*โดย /By*' in EDU2 is not the starting EDU of the procedural knowledge.

Fig. 4 – An example of ambiguous cue.

'*ในที่สุด/finally*' etc., to mark the ending boundary. And, there are 2–3 different-procedural-knowledge sets solving the same plant-disease problem occurring in one document as shown in Fig. 5.

Therefore, we apply learning the relatedness value (see Section 4.2.1.2) between two consecutive words as the word co-occurrence or Word-Co with the concept of procedural knowledge. Word-Co is then used to identify the starting EDU of the procedural knowledge where the first word of Word-Co is a verb, υ_{proc} ($\upsilon_{proc} \in V_{proc}$, V_{proc} is the procedural verb concept set), and the second word of Word-Co is a noun, n_{proc} ($n_{proc} \in N_{proc}$, N_{proc} is the noun concept set with the procedural concept approach).

V_{proc} = {'*ใช้/use*', '*นำ/take*', '*หว่าน/scatter*', '*ทำลาย/destroy*', '*ปลูก/grow*', '*ตาก/dry*', '*ตำ/hit*', …}

N_{Proc} = {' ', '*ส่วนประกอบพืช/Plant Organ*', '*พันธุ์ต้านทาน/resistant variety*', '*สารเคมี/chemical substance*', '*ยา/pesticide*', '*เชื้อ/micro-organism*', '*น้ำ/water*', …}

We apply SVM, and ME to learn the procedural knowledge boundary from υ_{proc} (the procedural verb concept), of two adjacent EDUs by the sliding window size of two consecutive EDUs with the sliding distance of one EDU.

3.2. *Problems of Why–How QA system*

There are three main problems: how to identify Why-Q and How-Q on Qpattern when their question words are ambiguous, how to determine the corresponding answer of Why-Q, and how to determine the corresponding answer of How-Q without problem-cause notification.

EDU1: "*น้อยหน่าสามารถใช้กำจัด เพลี้ยกระโดดสีน้ำตาล/****A sugar apple can be used to kill Brown PlantHopper***."
EDU2: "*ใช้เมล็ดน้อยหน่า 1 กก./* ***Use 1kgs.sugar apple seeds***."
EDU3: "*ตำละเอียด/* ***Grind finely***."
EDU4: "*แช่น้ำ 10 ลิตร นาน 12-24 ชั่วโมง/* ***Soak in 10 liters water for 12-24hrs***."
EDU5: "*กรองน้ำผสมน้ำสบู่ 1 ช้อนโต๊ะ/* ***Filtrate mixes with 1tb. soap solution***."
EDU6: "*ฉีดพ่นทุกๆวันนาน 6-10 วัน ช่วงเวลาเย็น/* ***Spray*[***the plant***] ***every day for 6-10 days***."
EDU7: "*ใช้ใบสด 2 กก./* ***Use 2kgs.fresh sugar apple leaves***."
EDU8: "*ตำละเอียด /* ***Grind finely***."
EDU9: "*แช่น้ำ 15 ลิตร นาน 24 ชั่วโมง /* ***Soak in 15 liters. water for 24hrs***."
EDU10: "*กรองน้ำผสมน้ำสบู่ 1 ช้อนโต๊ะ/* ***Filtrate mixes with 1tb.soap solution***."
EDU11: "*ฉีดพ่นทุกๆวันช่วงเวลาเย็น/* ***Spray*[***the plant***] ***every evening***."
EDU12: "[*เราสามารถ*]*ใช้น้อยโหน่งแทนน้อยหน่าได้เช่นกัน/* ***[We can also] use a custard apple to replace a sugar apple***"
where EDU2 through EDU5 are the procedural knowledge of the herbal-insecticide preparation.
And EDU7 through EDU10 are another herbal-insecticide preparation.

Fig. 5 – An example of boundary determination problem.

3.2.1. *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 Why-Q is a reasoning question, a 'อะไร/*What*' function of What-Q is asking for information about something (http://www.englishclub.com/vocabulary/wh-question-words.htm). However, there is question word ambiguity, e.g. 'อะไร/*What*' as in reasoning, as shown in Fig. 6. Therefore, we propose using different machine learning, NB, ME and MLP, to classify three question types as Why-Q (a reasoning question or a causality question), How-Q (the instructional How for solving problems), and Other-Q (Other-questions). 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); from two adjacent EDUs (EDU_q and $EDU_{ct\text{-}k}$ where $k = 1$, n or $n + 1$).

3.2.2. *Why-answer determination problem*

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 the USA?" Ans: "**Obama is the president of the 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 [13] whereas the *Why* question word in Qpattern cannot be applied to determine the answer. Moreover, the *Why* question word has previously been approached by determining the corresponding *Why* answer based on the question EDU/sentence having a causal verb [4] or noun phrases with a question word [6], which is not suitable for our *Why* question based on several effect-event explanations. Therefore, we solve the answers of causes for Why-Q with Qpattern by ranking the candidate answers of causes from the number of matching EDUs based on the similarity-score. 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 4.2.3) after the stop word removal. The similarity score determination in this research is based on WordNet and a Thai Encyclopedia after using a Thai-to-English dictionary.

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

Fig. 6 – An example of question word ambiguity.

EDU_{ct1}: "*ช่วงแตกกอ:ใบข้าวหงิกงอ/* ***In the tillering stage: rice leaves shrink***."
EDU_{ct2}: "*ต้นไม่เติบโต/* ***The rice plant stunts***."
EDU_q: "*[เรา]จะทำอย่างไร/* ***How should[we]solve***?"

Fig. 7 – How-Q contains the symptom-explanation EDUs without the symptom-cause notification.

3.2.3. *How-answer determination problem*

The questions based on problem solving are difficult to answer if How-Q of the research contains the explanation of symptoms/problems without notifying the cause of symptoms, as shown in Fig. 7. It is necessary to solve the disease names (PlantDiseaseX) or the causes of the symptoms to determine the methods to solve the symptoms through the causality graph [1] integrated with the current extracted procedural knowledge. Therefore, the disease name can be determined by ranking the candidate causes from the number of matching EDUs based on the similarity-score which is determined among $EDU_{ct\text{-}a}$ of the content EDU vector and $EDU_{effect\text{-}b}$ of all cause-effect-EDU vectors after the stop word removal.

4. Framework of Why and How QA system

The *Why–How* QA system of this research consists of two major parts, a question part and an answering part (including the procedural knowledge extraction). There are three steps in the question part, Question Corpus Preparation, Learning of Why-Q and How-Q on Qpattern, and Identification of Why-Q and How-Q. The answering part consists of three main steps, Procedural Knowledge Extraction from Textual Data, Integration of Causality Graph and Extracted Procedural Knowledge, and Answer Determination, as shown in the lower part of Fig. 8.

4.1. *Question part*

4.1.1. *Question corpus preparation*

The preparation of the question corpora was conducted 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 to learn the question types based on ten fold cross validation, and the second part of 150 questions for testing. For all of these questions, a Thai word segmentation tool is employed, which includes tagging the part of speech [14], and solving Named Entity [15]. EDU segmentation [16] is then carried out to generate EDUs for the semi-automatic annotation (based on experts) of question type concepts, a causative-verb concept (v_c) and an effect-verb concept (v_e) as shown in Fig. 9. 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 1 used to identify a causative EDU and an effect EDU, respectively. All concepts from Table1 are referred to Word Net [17]

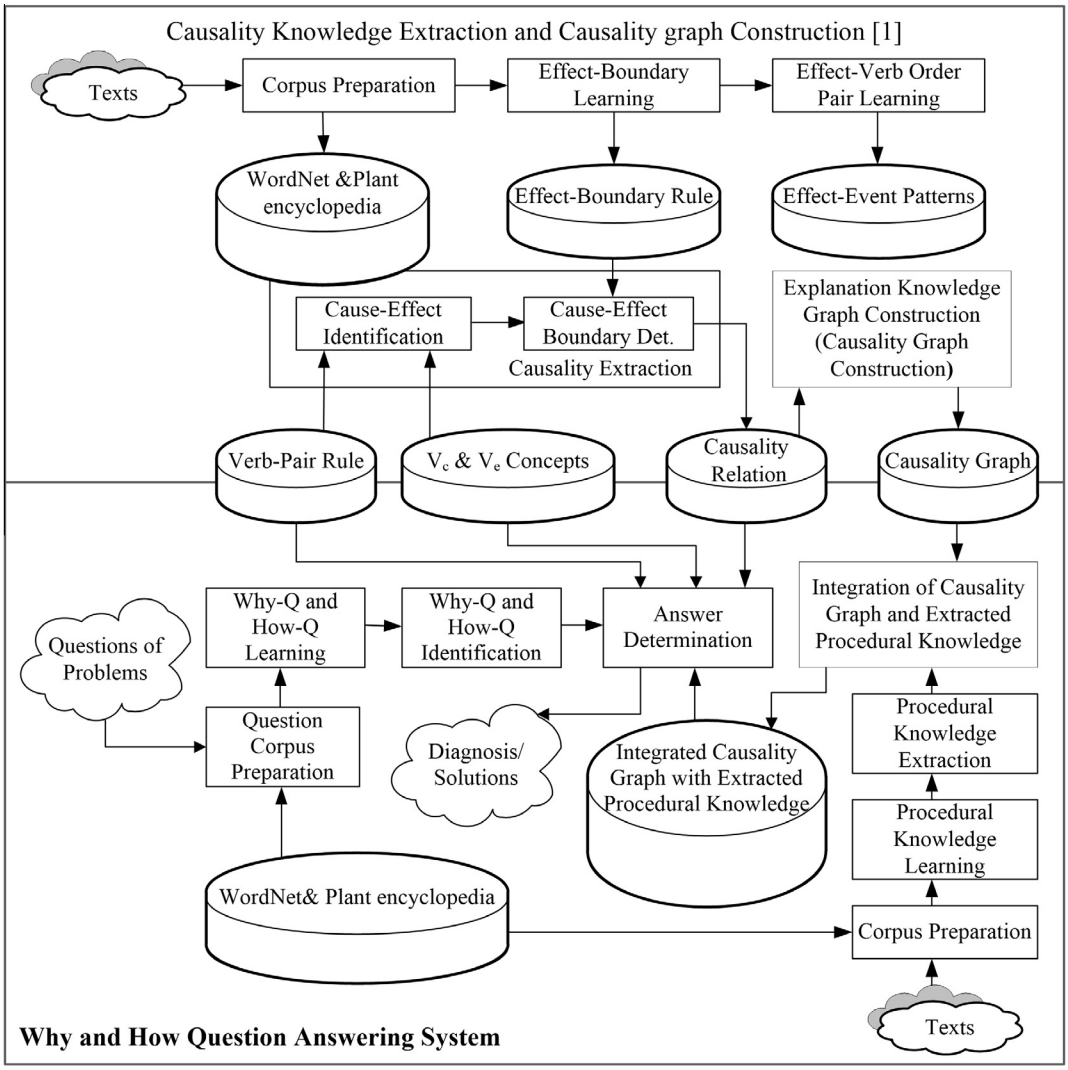


Fig. 8 – System overview.

EDUct-1: "ใบข้าว/*Rice leaves* หงิก/*shrink* งอ" ("***Rice leaves shrink***.")
EDUct-2: "ต้น/*Plants* แคะแกรน/*stunt* ช่วงแตกกอ/*at the tillering stage*" ("***Plants stunt at the tillering stage***.")
EDUq: "[เรา/*we*] จะแก้/*should solve* อย่างไรครับ/*how*" ("***How should* [*we*] *solve* [*it*]?")

<**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'>แคะแกรน/vi </**Vct**></**Qfocus**> [ช่วง/ncn แตก/vi กอ/nct]/NP]/VP
</**EDUct-2**>
<**EDUq**> ϕ =we]/NP
[จะ/prev <**Vq** : concept= 'solve'>แก้/vt </**Vq**>
<**Qword**=How: **concept=ComplicateHow-Q**> อย่างไร/pint </**Qword**> ครับ/aff]/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 set, 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.

Fig. 9 – An example of the question annotation.

(http://wordnet.princeton.edu/obtain) and Thai Encyclopedia of Plant Diseases (http://kanchanapisek.or.th/kp6/) after using the Thai-to-English dictionary (http://longdo.com).

4.1.2. *Learning of Why-Q and How-Q*

In this step, three different machine learning techniques are applied, NB, ME, and MLP to learn Why-Q, How-Q, and Other-Q 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 a is k and $k = 1$ or n or $n + 1$) from the annotated corpora are used in learning the question type classification by different machine learning techniques NB, ME, and MLP.

Naïve Bayes (NB): According to [18], NB learning is a generic classification to determine the feature probabilities of three classes of the question types based on Qpattern (class1 = 'Why-Q', class2 = 'How-Q', class3 = 'Other-Q'). The features of NB classifiers consist of Qword, V_{ct}, and V_q, from the annotated corpora of EDU_q and $EDU_{ct\text{-}k}$.

Maximum Entropy (ME): The ME model will be the one that is consistent with the set of constraints imposed by the evidence, but otherwise is as uniform as possible [19,20]. They modeled the probability of a semantic role r given a vector of features x according to the ME formulation below:

$$p(r|x) = 1/z_x \exp\left[\sum_{j=0}^{n} \lambda_j f_j(r, x)\right] \tag{1}$$

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 Eq. (1), ME can be used as the classifier of the r class when $p(r|x)$ 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 = 'How-Q' if $r = 2$, and class3 = 'Other-Q' if $r = 3$) and x is the binary vector consisting of all the consecutive elements of three feature sets: Qword, V_{ct}, and V_q, from EDU_q and $EDU_{ct\text{-}k}$ as shown in Eq. (2).

$$\begin{aligned} p(r|x) = \arg\max_r \frac{1}{z} \exp\Bigg(& \sum_{j=1}^{n} \lambda_j f_{class1,\ ct-k,j}(r, v_{ct-k}) \\ & + \sum_{j=1}^{n} \lambda_j f_{class2,\ ct-k,j}(r, v_{ct-k}) + \sum_{j=1}^{n} \lambda_j f_{class3,\ ct-k,j}(r, v_{ct-k}) \\ & + \sum_{j=1}^{n} \lambda_j f_{class1,\ q,j}(r, v_q) + \sum_{j=1}^{n} \lambda_j f_{class2,\ q,j}(r, v_q) \\ & + \sum_{j=1}^{n} \lambda_j f_{class3,\ q,j}(r, v_q) + \sum_{j=1}^{n} \lambda_j f_{class1,\ qw,j}(r, qw) \\ & + \sum_{j=1}^{n} \lambda_j f_{class2,\ qw,j}(r, qw) + \sum_{j=1}^{n} \lambda_j f_{class3,\ qw,j}(r, qw) \Bigg) \end{aligned} \tag{2}$$

Multi-Layer Perceptrons (MLPs): According to [21], Artificial Neural Networks (ANNs) are composed of neuron-like units connected together through input and output paths that have adjustable weights. Each node (neuron) produces an output signal, which is a function of the sum of its inputs. This function is formulated as in Eq. (3)

$$y_i = f\left(\sum x_i w_i\right) \tag{3}$$

where w_i represents the weight, x_i is the input feature of the input node. There are three input nodes of qw, $v_{ct\text{-}k}$, v_q (where $k = 1$ or n or $n + 1$) from two adjacent EDUs (EDU_q and $EDU_{ct\text{-}k}$). $f(\cdot)$ is the activation function such as a sigmoid function, and y_i is the output of the i^{th} node. MLP consists of an input layer, hidden layers, and an output layer which produce the output pattern/class. At the output layer, there are three nodes of three different classes: Why-Q (a reasoning class), How-Q (a procedural class), and Other-Q. Thus, MLP applied in determining the question type is based on the binary classes with multilevels since the activation function is generally a binary-value function (either 0 or 1). Each layer includes a different number of processing nodes. The net weighted input can then be solved by Eq. (4) where n is the number of neuron inputs, θ_j is the threshold value of the neuron at the j^{th} node in the hidden layer, and the number of hidden layers $p = 2$.

$$y_j(p) = \sum_{i=1}^{n} x_i(p) w_{ij}(p) - \theta_j \tag{4}$$

4.1.3. *Identification of Why-Q and How-Q*

All probabilities or weights from the previous learning step by NB, ME, and MLP are used to identify the question types.

Naïve Bayes: According to [18], Eq. (5) and the feature-probabilities determined by the previous step of NB are used to identify the question-type classes of Why-Q, How-Q, and Other-Q on Qpattern by the algorithm shown in Fig. 10.

$$\begin{aligned} QpatternClass &= \arg\max_{class \in Class} P(class|v_{ct-k}, v_q, qw) \\ &= \arg\max_{class \in Class} P(v_{ct-k}|class) P(v_q|class) P \\ &\quad (qw|class) P(class) \\ &\text{where } v_{ct-k} \in V_{ct} \text{ where } V_{ct} \text{ is a verb concept set expressed} \\ &\quad \text{on } EDU_{ct-k} \\ &\quad k = 1 \text{ or } n \text{ or } n+1 \\ &\quad v_q \in V_q \text{ where } V_q \text{ is a verb concept} \\ &\quad \text{set expressed on } EDU_q \\ &\quad qw \in Qword \text{ where } Qword \\ &\quad question - word \text{ concept set} \\ &\quad Class = \{class1, class2, class3\} \end{aligned} \tag{5}$$

Maximum Entropy: We use λ_j (the weight for a given feature function of the binary vector) which resulted from learning Why-Q, How-Q, and Other-Q to identify the question type classes by Eq. (2) as shown in the algorithm of Fig. 10 with the ME case.

Multi-Layer Perceptrons: The weight **w** from the results of learning Why-Q, How-Q, and Other-Q is used to determine the classes of the question types by Eq. (4) as shown in the algorithm of Fig. 10 with the MLP case.

4.2. *Answering part*

4.2.1. *Procedural knowledge extraction from texts*

There are three steps including Corpus Preparation, Procedural Knowledge Learning, and Procedural Knowledge Extraction as shown in Fig. 8.

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 ; $qw \in$ word
$EDU_{ct\text{-}k} \rightarrow$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
2 count= length[L] /* the number of EDUs in Qpattern
3 while i ≤ length[L] and flagQ=0do
4 { If qw_in_EDU_i /*find the Question EDU or EDU_q
5 { flagQ=1
6 If i=1 then { EDU_{i+1} is $EDU_{ct\text{-}1}$ };
7 If i=count-1 then { EDU_{i-1} is $EDU_{ct\text{-}n}$ and EDU_{i+1} is $EDU_{ct\text{-}(n+1)}$}
8 If i=count then { EDU_{i-1} is $EDU_{ct\text{-}n}$ } }
9 i++ }
10 If flagQ=1
11 /* The features of NB, ME, MLP are based on verb concepts.
/*If the serial verbs occur in EDU_q or EDU_{ct}, the concept of the first verb in the serial verbs is considered as one of those features.
Case: use NB
Equation5
Case: use ME
Equation 2
Case: use MLP
Equation 4
End_case
12 **Return**

Fig. 10 – Algorithm of Identifying Why-Q and How-Q on Qpattern by NB, ME, and MLP.

4.2.1.1. *Corpus preparation.* This step is the preparation of corpora in the form of EDU from three domains: the natural-organic-pest-control domain (http://www.kasetporpeang.com/forums), a plant disease domain (http://www.doa.go.th/), and a news domain (particularly in indigenous technology, http://info.matichon.co.th/techno/). The step involves using Thai word segmentation tools which tag its part of speech [14], include Named entity [15], and EDU segmentation [16]. These EDU corpora from three domains consist of 2 parts: the first part of 4500 EDUs is for learning procedural knowledge based on 10-fold cross validation, and the second part of 1500 EDUs is for testing. In addition to the learning part, we semi-automatically annotate the procedural EDUs, as shown in Fig. 11, where a verb concept and a noun concept are referred to WordNet and Thai Encyclopedia after using the Thai-to-English dictionary.

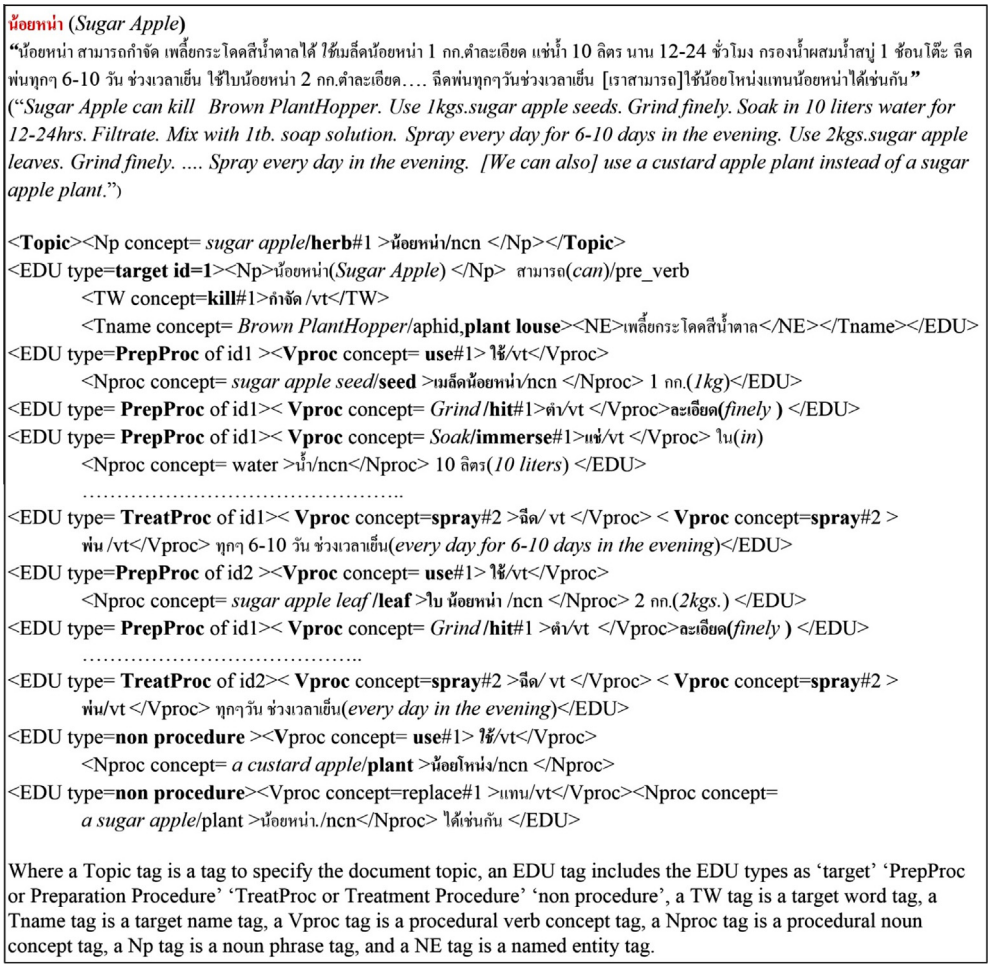
4.2.1.2. *Procedural knowledge learning.* This learning step includes two learning techniques: Learning RelatednessValue and Learning Boundary.

(a) Learning Relatedness Value. The objective of this learning step is to learn the relatedness value (r) [22] between two consecutive words, v_{proc} n_{proc}, as a word co-occurrence (Word-Co) with the procedural knowledge concept as a starting procedure as shown in Eq. (6). Thus, Word-Co is used to identify the starting procedural knowledge after a target topic or a target EDU has been identified by the target word pair, *tw1tw2* (see Section 3.1).

$$r(v_{proc}, n_{proc}) = \frac{fv_{proc}n_{proc}}{fv_{proc}+fn_{proc}-fv_{proc}n_{proc}}.$$
where $r(v_{proc}, n_{proc})$ *is the relatedness of Word–Co with a procedural concept.*
$v_{proc} \in V_{proc}$, V_{proc} *is a procedural verb concept set*
$n_{proc} \in N_{proc}$ N_{proc} *is the procedural noun concept set*
fv_{proc} *is the numbers of* v_{proc} *occurrences.*
fn_{proc} *is the numbers of* n_{proc} *occurrences.*
$fv_{proc}n_{proc}$ *is the numbers of* v_{proc} *and* n_{proc} *occurrences.* (6)

where each v_{proc} n_{proc} co-occurrence existing on documents contains two relatedness $r(v_{proc}, n_{proc})$ values, a procedural concept and a non-procedural concept. The only v_{proc} n_{proc} co-occurrence with the higher $r(v_{proc}, n_{proc})$ value of the procedural concept than the one of the non-procedural concept is collected as a Word-Co element of the Word-CO set with the procedural concepts. The Word-CO set is then used to identify the starting procedural EDU.

(b) Learning Procedural Knowledge Boundary. We use machine learning ME, and SVM by Weka to learn the procedural knowledge boundary. The features used in learning the procedural knowledge boundary are based on the events expressed by verbs or verb phrases. Moreover, some documents in our corpora contain a sequence of several procedural sets per document and each procedural set contains several procedural EDUs. A procedural EDU (EDU_{proc}) is expressed by a procedural verb concept, v (which is v_{proc}). Thus, all annotated verbs with the procedural concepts from the corpus


น้อยหน่า (*Sugar Apple*)
"น้อยหน่า สามารถกำจัด เพลี้ยกระโดดสีน้ำตาลได้ ใช้เมล็ดน้อยหน่า 1 กก.ตำละเอียด แช่น้ำ 10 ลิตร นาน 12-24 ชั่วโมง กรองน้ำผสมน้ำสบู่ 1 ช้อนโต๊ะ ฉีดพ่นทุกๆ 6-10 วัน ช่วงเวลาเย็น ใช้ใบน้อยหน่า 2 กก.ตำละเอียด…. ฉีดพ่นทุกๆวันช่วงเวลาเย็น [เราสามารถ]ใช้น้อยใหน่งแทนน้อยหน่าได้เช่นกัน"
("*Sugar Apple can kill Brown PlantHopper. Use 1kgs.sugar apple seeds. Grind finely. Soak in 10 liters water for 12-24hrs. Filtrate. Mix with 1tb. soap solution. Spray every day for 6-10 days in the evening. Use 2kgs.sugar apple leaves. Grind finely. .... Spray every day in the evening. [We can also] use a custard apple plant instead of a sugar apple plant.*")

<Topic><Np concept= *sugar apple*/herb#1 >น้อยหน่า/ncn </Np></Topic>
<EDU type=target id=1><Np>น้อยหน่า(*Sugar Apple*) </Np> สามารถ(*can*)/pre_verb
<TW concept=kill#1>กำจัด /vt</TW>
<Tname concept= *Brown PlantHopper*/aphid,plant louse><NE>เพลี้ยกระโดดสีน้ำตาล</NE></Tname></EDU>
<EDU type=PrepProc of id1 ><Vproc concept= use#1> ใช้/vt</Vproc>
<Nproc concept= *sugar apple seed*/seed >เมล็ดน้อยหน่า/ncn </Nproc> 1 กก.(*1kg*)</EDU>
<EDU type= PrepProc of id1>< Vproc concept= *Grind*/hit#1>ตำ/vt </Vproc>ละเอียด(*finely*) </EDU>
<EDU type= PrepProc of id1>< Vproc concept= *Soak*/immerse#1>แช่/vt </Vproc> ใน(*in*)
<Nproc concept= water >น้ำ/ncn</Nproc> 10 ลิตร(*10 liters*) </EDU>
……..
<EDU type= TreatProc of id1>< Vproc concept=spray#2 >ฉีด/ vt </Vproc> < Vproc concept=spray#2 >
พ่น /vt</Vproc> ทุกๆ 6-10 วัน ช่วงเวลาเย็น(*every day for 6-10 days in the evening*)</EDU>
<EDU type=PrepProc of id2 ><Vproc concept= use#1> ใช้/vt</Vproc>
<Nproc concept= *sugar apple leaf*/leaf >ใบ น้อยหน่า /ncn </Nproc> 2 กก.(*2kgs.*) </EDU>
<EDU type= PrepProc of id1>< Vproc concept= *Grind*/hit#1 >ตำ/vt </Vproc>ละเอียด(*finely*) </EDU>
……..
<EDU type= TreatProc of id2>< Vproc concept=spray#2 >ฉีด/ vt </Vproc> < Vproc concept=spray#2 >
พ่น/vt </Vproc> ทุกๆวัน ช่วงเวลาเย็น(*every day in the evening*)</EDU>
<EDU type=non procedure ><Vproc concept= use#1> ใช้/vt</Vproc>
<Nproc concept= *a custard apple*/plant >น้อยโหน่ง/ncn </Nproc>
<EDU type=non procedure><Vproc concept=replace#1 >แทน/vt</Vproc><Nproc concept=
a sugar apple/plant >น้อยหน่า./ncn</Nproc> ได้เช่นกัน </EDU>

Where a Topic tag is a tag to specify the document topic, an EDU tag includes the EDU types as 'target' 'PrepProc or Preparation Procedure' 'TreatProc or Treatment Procedure' 'non procedure', a TW tag is a target word tag, a Tname tag is a target name tag, a Vproc tag is a procedural verb concept tag, a Nproc tag is a procedural noun concept tag, a Np tag is a noun phrase tag, and a NE tag is a named entity tag.

Fig. 11 – An example of the procedural knowledge annotation.

preparation are extracted as a verb concept vector (Vi) in matrix vector V.

Vi = {v_{i1}, v_{i2}. . .v_{im} p/non-p} where p is a procedural-verb-concept vector class from the procedural EDUs, and non-p is non procedural-verb-concept vector class from the non procedural EDUs.

V = {Vi} where i = 1. . .*n*

Maximum Entropy: According to Eq. (1) [20], ME can be used as the classifier of the *r* class when *p*(*r*| x) is the highest probability to determine two procedural knowledge boundary classes, ending and continuing. Where *r* is the procedural knowledge boundary class (boundary is ending when *r* = 0, otherwise *r* = 1) and x is the binary vector of the verb concept pair (v_{ih} v_{ih+1}) features from a sliding window size of two consecutive EDUs with the sliding distance of one EDU (where *i* = 1,2, . . . ,*n*; *h* = 1,2, . . . ,*m*), as shown in Eq. (7).

$$p(r|x) = \arg\max_r \frac{1}{z} \times \exp\left(\sum_{j=1}^{n}\lambda_j f_{yes,proc-ih,j}(r,v_{ih}) + \sum_{j=1}^{n}\lambda_j f_{no,proc-ih,j}(r,v_{ih}) + \sum_{j=1}^{n}\lambda_j f_{yes,proc-ih+1,j}(r,v_{ih+1}) + \sum_{j=1}^{n}\lambda_j f_{no,proc-ih+1,j}(r,v_{ih+1})\right) \quad (7)$$

Support vector machine: The linear binary classifier, SVM, is applied in this research to classify the procedural knowledge boundary with ending or with continuing each procedural verb pair from the annotated corpus by using Weka. According to [23], this linear function, *f*(x), of the input x = (x_1x_2. . .x_n) assigned to the positive class if *f*(x) ⩾ 0, and otherwise to the negative class if *f*(**x**) < 0, can be written as:

$$f(\mathbf{x}) = \langle \mathbf{w}\cdot\mathbf{x}\rangle + b = \sum_{i=1}^{n} w_i x_i + b \quad (8)$$

where *x* is a dichotomous vector number, *w* is the weight vector, *b* is bias, and (*w*, *b*) ∈ $\mathbf{R}^n \times \mathbf{R}$ are the parameters that control the function. The SVM learning results are w_i and *b* for each verb concept feature (x_i) in a verb concept pair (v_{ih} v_{ih+1}) from a sliding window size of two consecutive EDUs ($EDU_{ih}EDU_{ih+1}$) with the sliding distance of one EDU (where i = 1,2, . . . ,*n*; *h* = 1,2, . . . ,*m*).

4.2.1.3. Procedural knowledge extraction. The objective of this step is to recognize and extract the procedural knowledge from the testing EDU corpora after the target or the problem solution is identified by the *tw1 tw2* pair. The Word-CO set from the learning step in Section 4.2.1.2 is then used to identify the starting procedural EDU of the procedural knowledge,

followed by solving the procedural knowledge boundary. The procedural knowledge boundary determination is performed as follows by the algorithm shown in Fig. 12.

Maximum Entropy: We use λ_j which resulted from the ME learning, to determine the procedural knowledge boundary by Eq. (7) as shown in Fig. 12. Where λ_j is the weight for a given feature function of the boundary determination with a vector of verb-concept features containing the verb concept pair, v_{ih} v_{ih+1}, by sliding a window size of two consecutive EDUs with the sliding distance of one EDU.

Support vector machine: The results from SVM learning are the weight, w_i, and bias, b, of each verb feature (x_i). According to Eq. (8), the input vector of verb features (**x**) in the verb-concept pair, v_{ih} v_{ih+1} (by sliding a window size of two consecutive EDUs with the sliding distance of one EDU) including their weights and bias are used to determine the boundary. If $f(\mathbf{x}) \geqslant 0$, an ending class occurs, otherwise a continuing class occurs as shown in Fig. 12.

4.2.2. *Integration of causality graph and extracted procedural knowledge*

According to [1,24] and (http://www.web3point2.com/rice/indexApp.php), the previous causality graph was constructed from the extracted causality knowledge from documents. Thus, the previously constructed causality graph including a disease name consists of a causative node as a root node representing a causative event expressed by v_c, $v_c \in V_c$, and effect-nodes representing effect events (where each effect node is expressed by v_e, $v_e \in V_e$) as shown in Fig. 13a. Therefore, we integrate this previous causality graph with the extracted procedural knowledge if the plant disease name of the previous causality graph is a substring of either the topic name or the EDU_{target} of the extracted procedural knowledge. The integrated causality graph consists of a root node representing the disease name where the root node connects two sub-trees including a causality sub-tree (the previous causality graph) and a procedural sub-tree as shown in Fig. 13b.

Assume that each EDU is represented by (NP1 VP | NP1 V NP2).
L is a list of EDUs. Word-CO is a word co-occurrence set of procedural concepts as a starting procedure.
V_{ih}, V_{ih+1} are learning verb sets of procedural concepts from V_{proc}.
V_{proc} is a procedural concept set. TW is a target word set, Tname is a target name set.
PROCEDURAL_KNOWLEDGE_EXTRACTION (L, V_{ih}, V_{ih+1},TW,Tname)

```
1   i ← 1; j ← 1; R ← ∅ ; TG ← ∅ ; PROC ← ∅ ;
2   If (word_j ∈ TW ∧ word_{j+1} ∈ Tname ) in TopicName then TG ← TopicName
3   Else while (word_ji ∉ TW ∧ word_{ji+1} ∉ Tname ) in EDU_i do /* determine the target EDU
4             { TG ← TG ∪ {i}; i ++ } /*TG is a target EDU
5   If TG <> ∅
6   { while i ≤ length[L] do
7       begin_while1
8       flag ← no; count ← 1;
9       while flag=no ∧ count < 5 do /*find the starting Know-how within 5 EDUs from
                                       corpus studying
10      begin_while2
11        If v_i n_i ∈ Word-CO then /* v_i n_i is a verb followed by a noun within EDU_i.
12          flag=yes;
13        i++;
14      end_while2
15      while (v_i ∈ V_ih ) ∧ (v_{i+1} ∈ V_{ih+1}) ∧ flag=yes ∧ i ≤ length[L] do
16      begin_while3      /*Procedural Knowledge boundary determination
17            Case: useME
                Equation 7
                    If r=0 then flag ← no otherwise flag ← yes
              Case: useSVM
                Equation 8
                    If f(x)>0 then flag ← no otherwise flag ← yes
              EndCase
18            PROC ← PROC {i};
19            i ++
20        end_while3
21        R = R ∪ { (TG,PROC) }
22  end_while1  }
23  return R
```

Fig. 12 – Procedural knowledge extraction algorithm.

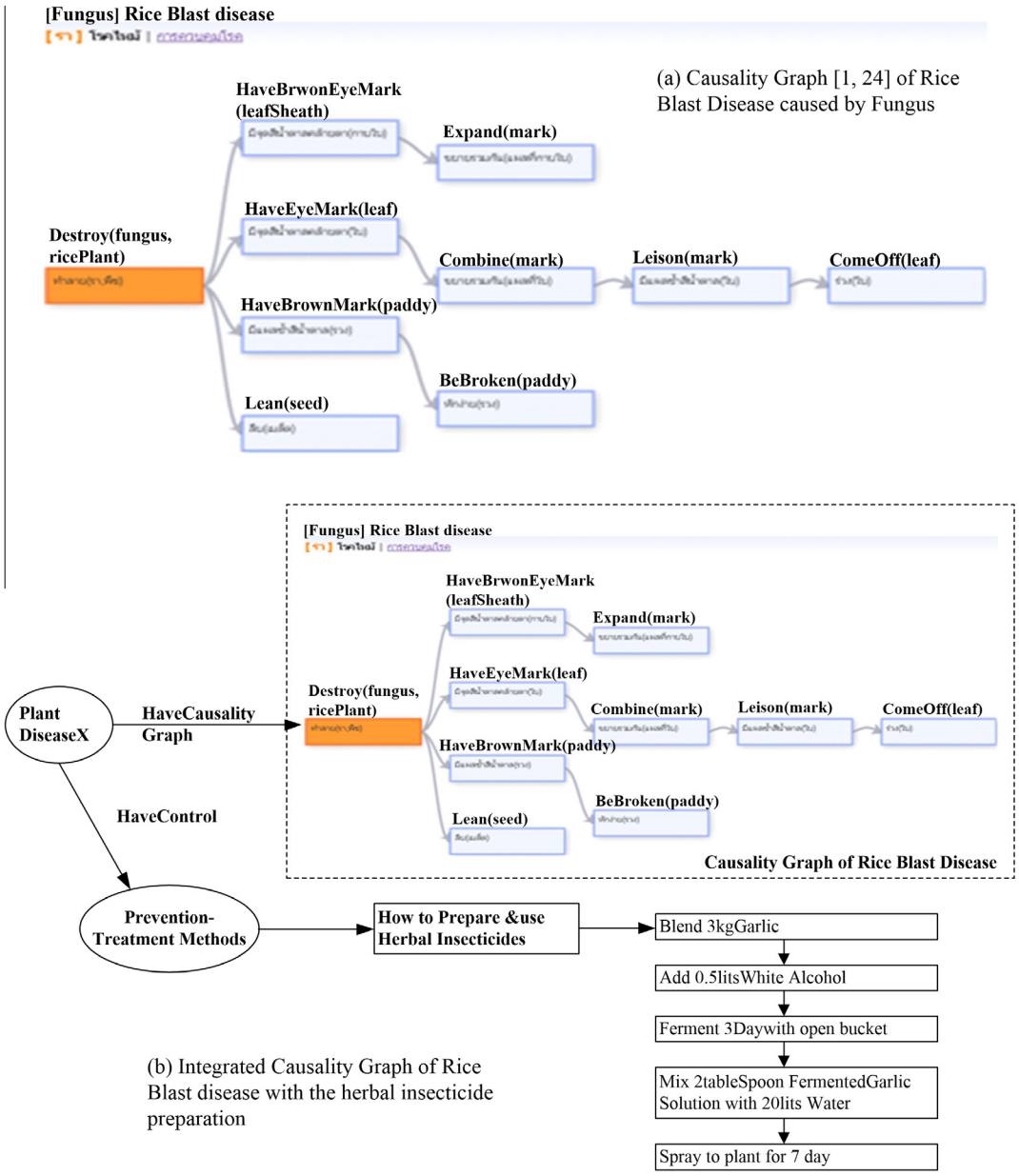


Fig. 13 – Example of the integration of the causality graph and procedural knowledge extracted from textual data.

4.2.3. *Answer determination*

The visualized answers of Why-Q and How-Q are randomly applied on the plant disease domain, particularly rice diseases, through the integrated causality graph. According to our research, the focuses of Why-Q and How-Q with Qpattern are based on the events expressed by v_{ct} which is v_c or v_e. The 90 questions, randomly selected from the 418 correct-question-type identification from Section 4.1.3, consist of 45 questions of Why-Q and 45 questions of How-Q about rice diseases. Each selected question of the 90 questions is used to determine its answer based on the Information Retrieval (IR) approach by ranking its candidate answers from their TotalSimilarity Score values. Each TotalSimilarity_Score value (Eq. (10) [25] and Eq. (11)) is determined by EDU matching between the content EDU vector of Why-Q or How-Q and the cause-effect-EDU vectors in the repository.

E_i is a symptom/effect-concept EDU set of Disease$_i$ {$EDU_{effect-1}$, $EDU_{effect-2}$,…, $EDU_{effect-m}$}

η is the number of different symptom/effect-concept EDUs. α is the number of different diseases.

$$\eta = \bigcup_{i=1}^{\alpha} E_i \tag{9}$$

$$Similarity\_Score = \frac{|S1_a \cap S2_{ij}|}{\sqrt{|S1_a| \times |S2_{ij}|}} \tag{10}$$

$$TotalSimilarity\_Score = \sum_{1}^{\eta} Similarity\_Score \tag{11}$$

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

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

In addition, our research focuses on only two kinds of Why-Q: Cause-Why-Q and Effect-Why-Q where Cause-Why-Q is a *why* question to determine 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-Q is a *why* question to determine 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-Q type, each TotalSimilarity_Score value is determined between the $EDU_{ct\text{-}a}$ of the content EDU vector and the $EDU_{effect\text{-}b}$ of each cause-effect-EDU vector. If Why-Q is the Effect-Why-Q type, each TotalSimilarity_Score value is determined between the $EDU_{ct\text{-}a}$ of the content EDU vector and the EDU_{cause} of each cause-effect-EDU vector. In addition, if the question is How-Q asking the solving method/procedural knowledge, each TotalSimilarity_Score value is determined between the $EDU_{ct\text{-}a}$ of the content EDU vector and either the EDU_{cause} or the $EDU_{effect\text{-}b}$ of each cause-effect-EDU vector.

All the 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 Eq. (10) are 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 the Similarity_Scores ($S1_a$, $S2_{ij}$) value of 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 to rank the candidate answers of the cause-effect-EDU vectors for the answer selection. For example: the following Qpattern-1 of the Cause-Why-Q type is expressed with all word concepts after stop word removal as follows.

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

EDUq: (เป็นเพราะ/be_reason)/v_q (อะไร/what)/pint

The candidate answers are ranked by sorting the TotalSimilarity_Score values (see Table 2) The possibility answer can then 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 2, the answer having the highest rank is the cause-effect-EDU vector with $Disease_2$ (Rank1). Moreover, the answer of How-Q without notifying the symptom cause can be solved by the integrated causality graph having the highest rank of the TotalSimilarity_score value from the matched symptom EDUs of the cause-effect-EDU vector.

5. Evaluation

5.1. *Data*

There are two categories of corpora for the evaluation of our proposed model: the question corpora and the procedural text corpora. The question corpora for evaluating the proposed model of identifying the question types, Why-Q, How-Q, and Other-Q, based on the questions with explanation contain 450 questions collected equally from the three community web-boards with different domains: the plant-disease domain, the health-care domain, and the technology-and-in digenous-technology domain. The 90 questions on rice diseases from the correct-question-type identification are randomly selected for the answer evaluation based on an IR approach. The corpora for the procedural knowledge extraction are collected from three domains: the herbal pest control domain, the plant disease domain, and a news domain (particularly for indigenous technology). All corpora categories focus on events expressed by verbs having different characteristics, e.g. the number of different verb features and feature dependencies. All of these characteristics allow this research to analyze how verb features affect the results of using different machine learning techniques for question identification and knowledge extraction.

5.2. *Question part*

The evaluation of the Why-Q and How-Q identification in this research is expressed in terms of precision and recall based on three experts with max win voting. The *Why* and *How* questions with explanation or Qpattern are based on several events expressed by verbs or verb phrases which are used as the main features for the Why-Q, How-Q, and Other-Q identification by three different machine learning techniques (MLP, ME, and NB). Table 3 shows the ME results with the highest precision of 0.930 for the health-care domain, which contains more feature dependency occurrences. The news domain of technology contains the highest diversity of verb feature occurrences (which result in the low frequency of verb feature occurrences) and the lowest feature dependency occurrences, which result in the lowest precision of 0.851 by NB compared to the other domains. Moreover, MLP results in the best recall of 0.84 for the health-care corpus whereas NB gives the lowest recall of 0.776 for the plant disease corpus containing more question-word-ellipsis occurrences of the posted problems on the web-boards. However, the average precision of the question type identification by MLP, ME, and NB with three classes, Why-Q How-Q and other, is 0.897 with an average recall of 0.814. In contrast, [26] applied five

Table 2 – Ranking the candidate answers of the cause-effect-EDU vectors for Qpattern with $\alpha = 13$ and $g = 69$.

$Disease_i$	EDU_{cause}	$EDU_{effect\text{-}b}$: Similarity_Scores Determination where $a = 1,2,\ldots,n$ or $n+1$; $i = 1,2,\ldots,\alpha$; $j = 1,2,\ldots,\eta$						Total Similarity_Score (TSC)	Rank by sorting TSC
		$S2_{i1}$ have BrownSpotMark (leaf)	$S2_{i2}$ be AlikeEye Shape(mark)	$S2_{i3}$ dry(plant)	$S2_{i4}$ be Yellow Color(leaf)	…	$S2_{i\eta}$		
$Disease_1$ RiceRaggedStunt disease	destroy (Rice_ragged_stunt_virus, plant)	0	0	0	0	…	0	0	0
$Disease_2$ RiceBlast disease	destroy (Rice_blast_fungus,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

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 noun-based features to classify several question types, and each question type occurred within one sentence. The average% correctness of their question classification [26] is 75% whilst SVM outperforms with 87.4% correctness. Moreover, the SVM algorithm is not concerned in the Why-Q, How-Q, and Other-Q identification of our research because the plant-disease domain and the health-care domain contain verb-feature-dependency occurrences, e.g. "*Plant stunts. What is the cause*?" ---> stunt/symptom - be_cause/be_a_reason, "*I have a rash on my neck. What should I do*?" ---> have_rash/symptom - do/solve, which frequently occur in both Why-Q and How-Q. Moreover, we apply the one-against-all multi-class SVM classifier to identify three question types, Why-Q, How-Q, and Other-Q, with the precisions at 0.879, 0.881, and 0.889 and the recalls at 0.779, 0.796, and 0.801 for the PlantDisease, HealthCare, and IndigenousTechnology domains, respectively. However, MLP and ME yield better performances than SVM and NB.

5.3. *Answering part*

The procedural knowledge extraction as the knowledge source of How-Q is also evaluated in terms of precision and recall based on three experts with max win voting as shown in Table 4. Word-Co, v_{proc} n_{proc}, with the concept of procedural knowledge can successfully identify the starting sequence of EDUs with the procedural knowledge concept on an average precision and an average recall of 0.96 and 0.94, respectively. The boundary determination results show that SVM gives the highest %correctness of 95.8 for the herbal pest control corpus containing moderate verb-pair-feature-dependency occurrences and a moderate diversity of verb feature occurrences whilst ME achieved a high %correctness of the boundary determination (9 4.4%) in the plant disease corpus which contains lower verb diversity (resulting in higher verb frequency) along with high verb feature dependency.

According to [27], Procedural knowledge is the knowledge about the relationships between function and mechanisms to perform side-effects and to sequence events or procedures. Procedural knowledge is mostly expressed in documents in terms of one or several event expressions based on either noun phrases or verb phrases. If the procedural knowledge event is based on a noun phrase, the technical-term-concept library and Named Entity Recognition are required for procedural knowledge extraction. If the procedural knowledge event is based on a verb phrase, the procedural knowledge extraction from text can be solved by a parsing tree, a term based approach, or a frame based approach. The procedural knowledge of our research is based on the event expressed by the verb phrase. In addition, our corpora with about 40% of zero anaphora (an ellipsis noun phrase) from the corpus study result in using a term based approach to extract the procedural knowledge without solving the zero anaphora. However, [12] extracted the procedural knowledge of instructions from the instruction text by using finite-state grammars with a Stanford Parser with an average correctness

Table 3 – The 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)	MLP		ME		NB	
			Pre-cision	Re-call	Pre-cision	Re-call	Pre-cision	Re-call
Plant disease	Medium	89	0.927	0.836	0.910	0.827	0.877	0.776
Health care	Medium	98	0.919	0.840	0.930	0.838	0.859	0.789
Indigenous techno.	Low	115	0.905	0.823	0.891	0.805	0.851	0.795

Table 4 – The evaluation of procedural knowledge extraction from texts.

Each domain contains 500 EDUs	Verb feature dependency occurrence	#of different verb features (Diversity)	Word-Co identification with procedural concept		Boundary determination	
			Precision	Recall	SVM %correct-ness	ME %correct-ness
Plant disease	High	74	0.96	0.92	91.5	94.4
Herbal pest control	Medium	156	0.97	0.93	95.8	92.3
Indigenous techno.	Medium	228	0.94	0.97	87.8	85.2

result of 81.4% F-measure without boundary consideration. [10] applied the pattern based approach to recognize each instruction as procedural knowledge (without boundary consideration) from 78 web pages over five domains with an average precision of 0.96 and an average recall of 0.59 where their highest precision and recall were 1.0 and 0.81 respectively from the cooking recipe domain. [11] extracted a unit-process vector from MEDLINE abstracts. Each unit process consisted of three elements: (1) Target (based on a noun-phrase expression e.g. a symptom/disease-name expression), (2) Action (based on a verb expression, e.g. '*treat*'), and Method (based on a noun/noun-phrase expression e.g. a treatment/prevention technical terms). [11] applied SVM and Conditional Random Fields (CRFs) to extract several unit processes without boundary consideration from abstracts but did not include partial matching in multi-word entities of Target and Method, which resulted in an average precision of 0.64 and an average recall of 0.61. However, the models or the methods from previous works on procedural knowledge extraction without boundary consideration can not be applied in our procedural knowledge extraction problems. Furthermore, most of the previous works can not be applied in our research without solving the zero anaphora occurrences.

The evaluation of the answer determination by the proposed model using the integration of the causality graph and the extracted procedural knowledge from text is expressed in terms of the percentage of correctness based on the answer set checked by experts with max win voting, as shown in Table 5.

Table 5 shows that the integrated causality graph representation of the answers for the rice disease domain can provide an average correctness of 90% for *Why* and *How* answers. Moreover, the zero anaphora occurrence on an $EDU_{ct\text{-}a}$ affects the %correctness of the visualized answers of both Why-Q and How-Q. Most of the previous works on the *Why* and *How* QA system based their questions on one sentence, except [8,28]. The answer method of Why-QA [8] was based on finding text fragments in 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. Oh et al. [8] achieved answer correctness of 41.8% (precision of the top answer) for why questions extracted from the Japanese version of Yahoo! Answers and also created by annotators without several event explanations as in Qpattern. In contrast, [28] worked on interpreting consumer health questions (which are explanation questions) without solving the answers. Our *Why* and *How* questions based on Qpattern with explanation of problems, e.g. plant-disease symptoms, are collected from community web-boards after misspelling-word correction, and our *Why* and *How* answers are determined by ranking the TotalSimilarity_Score values of the candidate answers

Table 5 – The evaluation of the answer determination.

Answer Expression	Correct Answer (Rank1) of Rice Disease	
	Why-Q (45)	How-Q (45)
By visualization of integrated causality graph	41 (91.1%)	40 (88.9%)

from the cause-effect-EDU vectors of the plant disease-symptom occurrences represented by the integrated causality graph from the knowledge repository.

6. Conclusion

This paper introduces a *Why* and *How* Question Answering system based on the questions that require explanation on community web-boards and provides preliminary diagnosis including methodologies for solving problems. The research benefits ordinary people of particular communities by providing the primary answers to their Why-Q and How/Q instead of waiting for expert responses. Machine learning is applied in question type identification, in particular Why-Q and How-Q, and also in the boundary determination of the procedural knowledge extraction from text. Thus, our proposed *Why* and *How* QA system can provide visualized answers by the integration of causality graphs [1,24] and the procedural knowledge extracted from text. The visualized *Why* and *How* answers with explanations in this research provide better results than previous researches. Moreover, our *Why* and *How* QA system can be applied to other languages because our methods of identifying question types (Why-Q, How-Q, and Other-Q) with explanations based on Qpattern, and extracting procedural knowledge with boundary considerations to answer How-Q with explanations based on the sequence of events mainly expressed by verb concepts from the verb phrases. The previous causality extraction [1] as the answers to our Why-Q with explanations is also based on consequence-event extraction by a verb concept pair, v_c and v_e. However, the problem of zero anaphora occurrences should be solved in future work to increase the correctness of answers. Finally, the model of our Why and How QA system can be applied not only by people in online communities but also by those in business and financial industries.

Acknowledgements

The research is supported by Thai Research Fund 2012 (MRG5580030).

References

[1] Pechsiri C, Piriyakul R. Explanation knowledge graph construction through causality extraction from texts. J Comput Sci Technol 2010;25(5):1055–70.
[2] Aouladomar F. Towards Answering Procedural Questions. In: Proc. of Knowledge and Reasoning for Answering Questions Workshop, International Joint Conference on Artificial Intelligence, Edinburgh, United Kingdom; 2005, p. 21–32.
[3] Carlson L, Marcu D, Okurowski ME. Building a discourse-tagged corpus in the framework of rhetorical structure theory. Curr New Direct Discour Dialog 2003;22:85–112.
[4] Girju R. Automatic detection of causal relations for question answering. In: Proc. of 41st annual meeting of the assoc. for computational linguistics, workshop on multilingual summarization and question answering-Machine learning and beyond, Japan; 2003, p. 76–83.
[5] Schwitter R, Rinaldi F, Clematide S. The Importance of How-Questions in Technical Domains. Proc. TALN-04, Workshop Question – Réponse, Fez, Morocco; 2004, p. 451–460.
[6] Verberne S, Boves L, Coppen P-A, Oostduk N. Discourse-based answering of why-questions. Traitement Automatique des Langues 2007;47(2):21–41.
[7] Baral C, Vo NH, Liang S. Answering Why and How questions with respect to a frame-based knowledge base: a preliminary report. In: Proc. of the 28th International Conference on Logic Programming ICLP 2012, Hungary; 2012, p. 26–36.
[8] Oh J-H, Torisawa K, Hashimoto C, Sano M, Saeger SD, Ohtake K. Why-Question Answering using Intra-and Inter-Sentential Causal Relations. In: Proc. of the 51st Annual Meeting of the Association for Computational Linguistics, Bulgaria; 2013, p. 1733–1743.
[9] Takechi M, Tokunaga T, Matsumoto Y, Tanaka H. Feature Selection in Categorizing Procedural Expressions. In: Proc. of the Sixth International Workshop on Information Retrieval with Asian Languages 2003; 11: 49–56.
[10] Delpech E, Saint-Dizier P. Investigating the Structure of Procedural Texts for Answering How-to Question. In: Proc. of JADT 2008, 9es Journées internationales d'Analyse statistique des Données Textuelles, France; 2008, p. 46–51.
[11] Song S-K, Oh H-S, Myaeng SH, Choi S-P, Chun H-W, Choi Y-S, Jeong C-H. Procedural Knowledge Extraction on MEDLINE Abstracts. In: Proc. of AMT 2011, LNCS 6890; 2011, p. 345–354.
[12] Zhang Z, Webster P, Uren V, Varga A, Fabio C. Automatically Extracting Procedural Knowledge from Instructional Texts using Natural Language Processing. International Conference on Language Resources and Evaluation; 2012, p. 520–527.
[13] Agichtein E, Cucerzan S, Brill E. Analysis of Factoid Questions for Effective Relation Extraction. In: Proc. of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil; 2005, p. 567–568.
[14] Sudprasert S, Kawtrakul, A. Thai Word Segmentation based on Global and Local Unsupervised Learning. In: Proc. of the 7th National Computer Science and Engineering Conference, Chonburi, Thailand; 2003, p. 1–8.
[15] Chanlekha H, Kawtrakul, A. Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. In: Proc. of the first International Joint Conference IJCNLP'2004, Hainan Island, China; 2004, p. 1–7.
[16] Chareonsuk J, Sukvakree T, Kawtrakul, A. Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information. In: Proc. of the 9th National Computer Science and Engineering Conference, Bangkok, Thailand; 2005, p. 85–90.
[17] Miller GA. WordNet: A lexical database. Commun ACM 1995;38(11):39–41.
[18] Mitchell TM. Machine Learning. Singapore: The McGraw-Hill Companies Inc. and MIT Press; 1997.
[19] Berger AL, Della Pietra SA, Della Pietra VJ. A Maximum Entropy approach to natural language processing. Comput Linguist 1996;22(1):39–71.
[20] Fleischman M, Kwon N, Hovy E. Maximum Entropy models for Frame Net classification. In: Proc. of the 2003 conference on Empirical methods in natural language processing, EMNLP, Sapporo, Japan; 2003, p. 49–56.
[21] Haykin S. Neural networks: a comprehensive foundation. USA: Prentice Hall; 1999.
[22] Guthrie JA, Guthrie L, Wilks Y, Aidinejad H. Subject-dependent co-occurrence and word sense disambiguation. In: Proc. of the 29th annual meeting on Association for Computational Linguistics, Berkeley, CA; 1991, p. 146–152.
[23] Vapnik VN. The nature of statistical learning theory. USA: Springer; 1995.

[24] Pechsiri C, Piriyakul R. Developing the UCKG-Why-QA System. In: Proc. of the 7th IEEE International Conference on Computing and Convergence Technology (ICCCT'12), South Korea; 2012, p. 679–683.
[25] Biggins S, Mohammed S, Oakley S. University Of Sheffield: Two Approaches to Semantic Text Similarity. In: Proc. of First Joint Conference on Lexical and Computational Semantics, Montre´al, Canada; 2012, p. 655–661.
[26] Zhang D., Lee W.S. Question Classification using Support Vector Machines. In: Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 2003, p. 26–32.
[27] Saxena Manoj K., Biswas K. K., Bhatt P. C. P. Knowledge representation in distributed blackboard architecture — Some issues. In: Proc, of International Conference KBCS '89 Bombay, India, 1989, p. 230–239.
[28] Kilicoglu H, Fiszman M, Demner-Fushman D. Interpreting Consumer Health Questions: The Role of Anaphora and Ellipsis. In: Proc. of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), Sofia, Bulgaria, 2013, p. 54–62.