รายงานวิจัยฉบับสมบูรณ์

โครงการ: การหาความรู้โดยรวมจากความรู้เหตุและผลที่สกัดได้จากเอกสารภาษาไทยสำหรับการตอบคำถามเหตุและผลโดยอาศัยคำถามจากข้อความและภาพ

โดย รองศาสตราจารย์ ฉวีวรรณ เพ็ชรศิริ

กรกฎาคม 2557

สัญญาเลขที่ MRG5580030

รายงานวิจัยฉบับสมบูรณ์

โครงการ: การหาความรู้โดยรวมจากความรู้เหตุและผลที่สกัดได้จากเอกสารภาษาไทยสำหรับการตอบคำถามเหตุและผลโดยอาศัยคำถามจากข้อความและภาพ

ผู้วิจัย สังกัด มหาวิทยาลัยธุรกิจบัณฑิตย์

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย

(ความเห็นในรายงานนี้เป็นของผู้วิจัย สกว.ไม่จำเป็นต้องเห็นด้วยเสมอไป)

Abstract (บทคัดย่อ)

Project Code : MRG5580030
(รหัสโครงการ)

Project Title : Knowledge Generalization from Causality Knowledge Extracted from Texts for Answering Why- Question Expressed by Text and Image

(ชื่อโครงการ) การหาความรู้โดยรวมจากความรู้เหตุและผลที่สกัดได้จากเอกสารภาษาไทยสำหรับการตอบคำถามเหตุและผลโดยอาศัยคำถามจากข้อความและภาพ

Investigator : Assoc. Prof. Chaveevan Pechsiri (Dhurakij Pundit University)
(ชื่อนักวิจัย) รองศาสตราจารย์ ดร. ฉวีวรรณ เพ็ชรศิริ (มหาวิทยาลัยธุรกิจบัณฑิตย์)

E-mail Address :itdpu@hotmail.com

Project Period : July 2012 – July 2014
(ระยะเวลาโครงการ) กรกฎาคม 2555 –กรกฎาคม 2557

Abstract

The research aims to extract and generalize the causality knowledge for supporting a Why Question Answering (QA) system integrated with image processing (called the Embedded-Image Why-QA system) for providing the knowledge used in the problem diagnosis, especially in plant diseases. The image expression is applied on the Why-question part for providing Why-question contents (i.e. plant symptoms) that are difficult to be explained by text. There are three main problems involved with this current research. The first problem of the causality knowledge extraction, especially the effect boundary determination problems, is confronted after applying the verb-pair (a causative verb and an effect verb) rules to identify the causality. Then, the research applies Maximum Entropy, Supported Vector Machine, and Naïve Bayes for the comparative study of the effect boundary determination, having the effect verb concepts from the effect EDUs as the features. The second problem is the knowledge generalization problems which come from the extracted causality knowledge containing the uncertainty nuance expression and the incompleteness knowledge. Thus, the research proposes applying the basic linguistic rules to solve the uncertainty problems and the Monte Carlo simulation technique to solve incompleteness problems by imputation of the effect unit. And then, we apply the fuzzy function right after the imputation to determine the generality value of each effect event expressed by the effect verb concept feature of the effect EDU from several documents having the same cause concept. The third problem is from the Embedded-Image Why-QA system which consists of how to determine the Why-question type from the text part of the question, how to determine the Why-question contents from the image part of the question, and how to determine the corresponding answer to the Why-question from the extracted causality. Therefore, the research applies a Why-question cue set to solve the Why-question type, a Bag-of-Visual-Words to solve the Why-question contents, and determining the corresponding answers by ranking the similarity scores between the question content and the extracted causality knowledge including the symptom generality value. Then, the results of this research have shown that the effect boundary determination based on ME has the highest correctness 92% on average and the extracted causality can support the embedded image Why-QA system by answering correctly at 78% correctness at the first rank.

Keywords: Generality value, effect boundary, Embedded-Image Why-QA system, visual word, Why-question cue

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์สกัดและหาความรู้เหตุและผลโดยรวม เพื่อสนับสนุนระบบการตอบคำถามเหตุและผล ที่ได้บูรณาการกับการประมวลผลภาพเข้าไว้ (เรียกระบบคำถามนี้ว่า ระบบเอมเบดเดด-อิมเมจวายคิวเอ Embedded-Image Why-QA system) ทั้งนี้เพื่อการจัดหาความรู้ให้ซึ่งใช้ในการวินิจฉัยปัญหา โดยเฉพาะปัญหาโรคพืช การนำภาพมาประยุกต์ใช้ในระบบการตอบคำถามเหตุและผลนี้ก็เพื่อง่ายต่อการอธิบายลักษณะอาการของโรค ซึ่งบางครั้งยากที่จะอธิบายด้วยข้อความ ปัญหาสำหรับงานวิจัยนี้ประกอบด้วยสามปัญหาหลักดังนี้ ปัญหาแรกเป็นปัญหาด้านการสกัดความรู้เหตุและผล โดยเฉพาะเรื่องการหาขอบเขตของผลหลังจากที่ได้ใช้กฎคู่กริยา (Verb-Pair Rule, กริยาเหตุและกริยาผล) ระบุความรู้เหตุและผล ฉะนั้นงานวิจัยนี้จึงทำการประยุกต์ใช้แม็กซิมัมเอนโตรพีหรือเอ็มอี (Maximum Entropy, ME) :ซับพอร์ตเวคเตอร์แมชชีนหรือเอสวีเอม (Support Vector Machine, SVM) และเนย์อีฟเบย์หรือเอนบี (Naïve Bayse, NB) เพื่อศึกษาการเปรียบเทียบการหาขอบเขตของผล โดยมีฟีเจอร์ (Feature) ที่ใช้คือกริยาผล ปัญหาที่สองเป็นปัญหาเกี่ยวกับการหาความรู้โดยรวม ซึ่งเกิดจากความรู้เหตุและผลที่สกัดได้นั้นพบปัญหาเกี่ยวกับ ความไม่แน่นอนของนูแอนซ (Nuance)ที่แสดงอาการโรค ปัญหาความไม่สมบูรณ์ของข้อมูลที่สกัดได้ ดังนั้นงานวิจัยนี้ขอเสนอกฎทางภาษาศาสตร์ แก้ปัญหาเกี่ยวกับความไม่แน่นอนของนูแอนซ และใช้เทคนิคการจำลอง มอนติคาร์โล(Monte Carlo Simulation Technique) ทำการเติมเต็มข้อมูลเพื่อแก้ปัญหาความไม่สมบูรณ์ของข้อมูล ก่อนที่จะทำการหาความรู้โดยรวมด้วยฟังก์ชั่นฟัซซี (Fuzzy Function) สำหรับหาค่าเจนเนอรัลลิตี (Generality Value) ของแต่ละกริยาผลที่มาจากสาเหตุเดียวกัน ปัญหาที่สามเป็นปัญหาเกี่ยวกับระบบเอมเบดเดด-อิมเมจวายคิวเอ ที่ประกอบด้วยปัญหาการระบุประเภทคำถามเหตุและผลจากส่วนที่เป็นข้อความของคำถาม ปัญหาการหาเนื้อหาคำถามจากส่วนที่เป็นภาพของคำถาม และปัญหาการหาคำตอบจากความรู้เหตุและผลที่สกัดได้พร้อมด้วยค่าเจนเนอรัลลิตี ดังนั้นงานวิจัยนี้จึงประยุกต์ใช้คูเซตประเภทคำถามเหตุและผล (Why-question cue set) ระบุประเภทคำถามเหตุและผล ใช้ถุงของวิสชวลเวอร์ด (Bag of Visual Word) หาเนื้อหาคำถาม และใช้การจัดลำดับคะแนนความคล้าย ระหว่างเนื้อหาคำถามกับความรู้เหตุและผล ที่สกัดได้มาทำการหาคำตอบจากการทดลองของงานวิจัยนี้ ได้แสดงให้เห็นว่า การหาขอบเขตของผลด้วยวิธีเอ็มอีให้ความถูกต้องเฉลี่ยสูงสุดคือ 92% และความรู้เหตุและผลที่สกัดได้สามารถใช้ตอบคำถามได้ถูกต้อง 78% ที่ลำดับที่1 (Rank1)

คำสำคัญ: Generality value, effect boundary, Embedded-Image Why-QA system, visual word, Why-question cue

Output จากโครงการวิจัยที่ได้รับทุนจาก สกว.

1. ผลงานตีพิมพ์ในวารสารวิชาการนานาชาติ (ระบุชื่อผู้แต่ง ชื่อเรื่อง ชื่อวารสาร ปี เล่มที่ เลขที่ และหน้า) หรือผลงานตามที่คาดไว้ในสัญญาโครงการ

   ชื่อผู้แต่ง: ผู้ช่วยศาสตราจารย์ ดร. ฉวีวรรณ เพ็ชรศิริ

   ชื่อเรื่อง: The Integration of Text-Based Why Question Answering System And Image Processing For Root-Cause Diagnosis

   ชื่อวารสาร ปี เล่มที่ เลขที่ และหน้า: International Journal on Artificial Intelligence Tools รอการตอบรับ

   ชื่อเรื่อง: Introducing Why - How Question Answering System and Integrated Causality Graph through Online Community

   ชื่อวารสาร ปี เล่มที่ เลขที่ และหน้า: Journal of Universal Computer Science รอการตอบรับ

2. การนำผลงานวิจัยไปใช้ประโยชน์

   - เชิงชุมชน โดยได้มีการนำผลการวิจัยไปพัฒนาต่อในขั้น Prototype ที่1 เพื่อทดลองใช้กับกลุ่มเกษตรกรที่ปลูกข้าวในภาคกลางว่าสามารถช่วยทำให้เกษตรกรและครอบครัวมีความเข้าใจในเรื่องสาเหตุและอาการของโรคได้ดีขึ้น

3. อื่นๆ (เช่น ผลงานตีพิมพ์ในวารสารวิชาการในประเทศ การเสนอผลงานในที่ประชุมวิชาการ หนังสือ การจดสิทธิบัตร)

   -

Summary Report

The situation today of using the internet is quite different from the previous time which emphasizes on sending emails, searching the required information, and e-business. Now using the internet emphasizes on the social network, e.g. Face Book, Lines, and etc. When people have some of problems, questions, the interesting information, and suggestions, they prefer to post them on the social media based on social network. In order to enhance Know-Why knowledge to people in the social network for solving their problem through problems' diagnosis, the research aims to develop a Why Question Answering system integrated with image processing to provide root-cause analysis or to support knowledge used in the problem diagnosis, especially in plant diseases through a mobile phone or a computer as a solution center. The image expression is applied on the Why-question part for providing Why-question contents (i.e. plant symptoms) that are difficult to be explained by text. There are several problems involved to this research on the Why-question part, which includes how to determine the Why-question type from the textual question probably containing the ambiguous question word, how to determine the Why-question contents from the image embedded within the textual question, and how to determine the Why-question focus. Therefore, we propose using a Why-question cue set to solve the Why-question type, a Bag-of-Visual-Words to solve the Why-question contents, and a causative verb concept /an effect verb concept gained from our previous research to solve the Why-question focus. Moreover, there are two problems on the Why-answering part; how to generalize the previous extracted causality knowledge as the answer source with the incomplete knowledge problem, and how to determine the corresponding answer. We apply the Monte-Carlo technique to solve the incomplete knowledge and the verb-pair rules along with the noun phrase similarity to solve the answers with reasoning. Finally, the research achieves 78% correctness of answering.

Executive Summary

Disease diagnostics and nosologic studies often require a combination of a broad knowledge of diseases and symptoms' prevalence, and probabilistic concepts in their reasoning (Miller,1994). The compilation of experiences and the capacity to perform the root-cause determination including the cause and effect reasoning allows diagnosticians to recognize common disease states and perform efficient and ethical diagnostic evaluations. However, some diagnosticians are often required to make decisions with the lack of information and knowledge. Thus, a Why-Question Answering system (a Why-QA system) with the generalized knowledge from the causality knowledge extracted from text approach would assist them to obtain the generalized causality knowledge through a Why question expressed in either the text form or the image and text form. The generalized causality knowledge is required to achieve an effective diagnosis at the fundamental level and to provide better services in the solution centers or the service centers.

In recent years, an automatic Why-QA system has been involved with several strategies: Information Retrieval, Information Extraction, Knowledge Extraction, Machine Learning, Image Processing, Natural Language Processing, and Reasoning for its answer determination. However, our research concerns of the knowledge generalization from the extracted causality knowledge from texts for the problem diagnosis through the Why-QA system. According to our current research, both the causality knowledge extraction from texts, especially the improvement of the effect-boundary determination and the causality knowledge generalization are necessary for automatically answering the Why question expressed in either the text-based question or the image embedded question (called "an Embedded-Image Why question" or "an EIWhy question") under the closed-domain QA system in each specific domain study. The reason of generalizing the knowledge is the extracted causality knowledge containing various causality expression contents varying on explanation with the same cause, and also varying on nuance expression on the documents. In addition to the Why question, it is very difficult to determine the root cause determination from the plant disease symptom (especially the lesion color and the lesion shape) explained on the text based Why question because several people have several ideas of color and shape explanations. According to the camera on the mobile telephone, it can assist the people with the plant-disease-symptom-explanation problem by taking a picture of the suspected symptom

of the plant disease and sending the EIWhy question to the server as shown in the following of the rice disease.

"ข้าว/rice เป็นโรค/get disease อะไร/what"
("What *disease does the rice plant get*?")

Then, the EIWhy-QA system (the Embedded-Image Why-QA system) will answer the basic cause of this rice disease for approaching how to control the disease. Thus, the Why-QA system is supported by both the causality knowledge extraction with the effect boundary consideration, and the causality knowledge generalization is very desirable for the enhancement of the preliminary diagnosis. Moreover, the causality in our research has been expressed through documents in the form of EDU (Elementary Discourse Unit) defined by Carlson et. al.(2003) as a clause which is equivalent to a Thai simple sentence. And, this research concerns only the inter-causal EDU (a causality expression of either one EDU or multiple EDUs on both the causative unit and the effect unit) defined by Pechsiri and Kawtrakul, 2007) for example:
Causative unit: EDU1 "ถ้าเพลี้ยทำลายต้นข้าว / If the aphids infest rice plants,"
Effect unit (EDU2+EDU3+EDU4):

EDU2 "จะทำให้ใบเหลือง/[it] will make the leaves become yellow."

EDU3 "ต่อมาหงิกงอ/Then [the leaves] shrink"

EDU4 "และต้นข้าวจะหยุดการเจริญเติบโต/and the rice plants will stop growing."

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

However, there are three main problems involved with this current research. The first problem of the causality extraction, especially the effect boundary determination problems, is confronted after applying the verb pair (a causative verb and an effect verb) rules from (Pechsiri and Kawtrakul, 2007) to identify the causality. The previous research (Pechsiri and Kawtrakul, 2007) applied the linguistic rules as Centering Theory(Walker et. al., 1998) to determine the effect EDU boundary performed inefficiently in some domains. Then, Maximum Entropy (ME, Csiszar, 1996), Supported Vector Machine (SVM, Cristianini and Shawe-Taylor, 2000), and Naïve Bayes (NB, Mitchell. 1997) are proposed by this research for the comparative study of the effect boundary determination, having the effect verb concepts from the effect EDUs as the features.

The second problem is the knowledge generalization problems where some extracted

inter-causal EDUs contain the uncertainty nuance expression and the incompleteness causality knowledge. We propose applying the basic linguistic rules to solve the uncertainty problems and the Monte Carlo simulation technique (Woller.,1996) to solve incompleteness problems by imputation of the effect unit. And then, we apply the fuzzy function right after the imputation to determine the generality value of each effect event expressed by the effect verb concept feature of the effect EDU from several documents having the same cause concept. The generality value from the fuzzy logic can represents subjective belief of the effect-verb concept feature are provided as the knowledge base for answering the Why question.

The third problem of our Why-QA research can be separated into two different parts according to the Why question expression: the text part and the image part. According to (Vazquez-Reyes and Black., 2008)I, the text part involves the Why question as the question word ambiguity. Previously, wh-questions have been approached by determining answers from noun phrases and question words (Verberne, 2006), which is suitable for the causal question or the Why question with the answer based on the lexico-syntactic pattern (Girju, 2003) as NP1 Verb NP2 (where NP1 and NP2 are the noun phrases), i.e. “What causes Tsunami? → Earthquake causes Tsunami”. However, it is not suitable for the Why questions with the answers based on the explanation as in our research, i.e. What are the effect symptoms after the aphid has destroyed the rice? This research proposes using a Why-question cue set to solve Why questions type determination. And, the image part of the Why question involves the recognition of plant image for finding lesion shape and infected area color. According to (Weizherg et al., 2008), the detection and identification of plant disease in practice is always performed by the naked eye observation of experts. This approach is expensive and time consuming because it requires an expertise from the experts. It could be improved by the assistance of advancement technology. In (Patil and Kumar, 2011), the technology that most research focused on is automatic detection of plant diseases by analyzing symptom and observing the lesion on the leaves or stems of the plant. This research also proposes using a bag of visual words from the image processing and a symptom-concept-frame structure to determine a question content from the image part of the question especially lesions occurring on rice leaves. The answer determination is based on the highest rank of the similarity scores between the extracted causality knowledge and the question content.

In addition to our methodology of generalizing the extracted causality knowledge from texts for the Embedded-Image Why-QA system" (EIWhy-QA system), it has been evaluated with two main parts based on three experts with max win voting; the first part is based on the %Correctness of effect boundary determination of causality knowledge extraction. The second part is based on the precision and the recall for determining the Why question type and the highest rank of the similarity score determination between the extracted causality knowledge and the question content for determining the answer including the generality value. The error of the knowledge extraction especially the effect boundary determination is resulted by the effect verb feature dependency. The errors from the EIWhy-QA system are resulted by the sarcastic questions from the textual questions of the EIWhy questions and the incorrect patch generation which results in incorrectly determining the visual words of the images of the EIWhy questions. The results of this research have shown that the effect boundary determination based on ME has the highest correctness 92% on average and the extracted causality knowledge can support the EIWhy-QA system by answering correctly at 78% correctness at the first rank.

In conclusion, our research includes 3 major phases: Causality Knowledge Extraction as the source of answers, Causality Knowledge Generalization, and EIWhy-QA System. The EIWhy-QA system includes the image processing technique to enhance the ability in diagnosing problems, especially the plant diseases. Therefore, the extracted causality knowledge including the generality values can successfully support the EIWhy QA system which benefits to inexperienced persons in preliminary diagnostics. Once integrated with mobile phones, such capacities allows the EIWhy-QA system to have a profound effect on several business areas as a tool to assist inexperience participants/people and amateur diagnosticians to diagnose problems.

TABLE OF CONTENT

TABLE OF CONTENT (Continued)

TABLE OF CONTENT (Continued)

LIST OF TABLES

LIST OF FIGURES

Knowledge Generalization from Causality Knowledge Extracted from Texts for Answering Why- Question Expressed by Text and Image

(การหาความรู้โดยรวมจากความรู้เหตุและผลที่สกัดได้จากเอกสารภาษาไทยสำหรับการตอบคำถามเหตุและผลโดยอาศัยคำถามจากข้อความและภาพ)

1. Introduction

Nosology studies and Disease diagnostics, especially the root-cause diagnosis, often require a combination of a broad knowledge of diseases and symptoms' prevalence, and probabilistic concepts in their reasoning (Miller,1994). The compilation of experiences and the capacity to perform the root-cause determination including the cause and effect reasoning allows diagnosticians to recognize common disease states and perform efficient and ethical diagnostic evaluations. However, some diagnosticians are often required to make decisions with the lack of information and knowledge. Thus, to provide the generalized causality knowledge extracted from technical documents for people in the preliminary diagnosis through a Why-Question Answering system (a Why-QA system) is challenge. Our research concerns of the knowledge generalization from the extracted causality knowledge from texts for the problem diagnosis through the Why-QA system (where 'Causality' is defined as 'a law-like relation between cause event types and effect event types (Lehmann et. al., 2004)). Both the causality knowledge extraction from texts, especially the improvement of the effect-boundary determination and the causality knowledge generalization are necessary for automatically answering the Why question expressed in either the text-based question or the image embedded question under the closed-domain QA system in each specific domain study. The reason of generalizing the knowledge is the extracted causality knowledge containing various causality expression contents varying on explanation with the same cause. Then, the generalized causality knowledge with the generality value determination is required to achieve an effective diagnosis at the fundamental level and to provide better services in the solution centers or the service centers. Furthermore, (Hovy et. al., 2002) there are about 5% of Why questions occurring in the Question Answering (QA) system. Although the frequency Why questions posed to QA systems is lower than that of other types of question such as who and what questions, it is necessary for diagnosis with reasoning. Therefore, our research concerns of generalizing the extracted causality knowledge from texts, especially from the plant disease domain or the hospital health-care domain, with the comparative study of the boundary determination to previous research (Pechsiri and Piriyakul, 2010) as the answer source of Why questions for supporting the problem diagnosis. And, the research also concerns of the text-based Why-QA system with an image embedded

(called “an Embedded-Image Why-QA system” or “an EIWhy-QA system”) for providing the clearer Why question as comparing to the text-based Why question without an image embedded (called “a regular Why question” or “a textual Why question”) of the regular Why-QA system / the Why-QA system. In addition to the regular Why-QA system, it is very difficult to determine the root-cause determination from the plant disease symptom explained on the text-based Why question without an image embedded; especially the lesion color and the lesion shape, because several people have several ideas of color and shape explanations. According to the camera on the mobile telephone, it can assist the people (who have the problem of the plant-disease-symptom explanation) by sending the embedded image question (called “an Embedded-Image Why question” or “an EIWhy question”) of the EIWhy-QA system under the closed-domain QA system in the specific domain study as shown in Fig. 1 of the rice leaf disease.

(a) “*ทำไมใบข้าวมีอาการผิดปกติ*”
(“*Why do rice leaves have unusual symptoms*?”)

(b) “*ข้าว/rice เป็นโรค/get disease อะไร*/what”
(“What *disease* does the rice plant *get*?”)

Fig.1 Examples of EIWhy questions

Therefore, our EIWhy-QA system is separated into two main parts, a Why-question part and a Why-answer part. The Why-question part from the user (notifying a problem) consists of two portions; a textual portion and an image portion which provides the Why question contents of the disease symptoms. The Why-answer part is a textual answer retrieved from the answer source which contains several extracted causality knowledge with the adjustment of boundary determination from (Pechsiri and Piriyakul, 2010) on agricultural-technical documents downloaded from the agricultural department web sites. Thus, our EIWhy-QA system can provide the causality knowledge for supporting the user’s diagnostic of a plant disease by answering the user’s question with all possible corresponding answers with the generality values of symptoms from the answer source.

Previous literature on knowledge generalization from the extracted causality knowledge from texts as the answer source of the automatic Why-QA systems have involved several strategies including, but not limited to, Information Retrieval, Information Extraction, Knowledge Extraction, Machine Learning, Image Processing, Natural Language Processing, and Reasoning (Burhans and Shapiro,2001;Girju, 2003; Verberne,2006; Yeh et al.,2008; Verberne et al.,2007;

Pechsiri and Kawtrakul,2007; Pechsiri and Piriyakul, 2010). The knowledge generalization of this research is based on the boundary determination after the adjustment of the verb features from the previous research (Pechsiri and Piriyakul, 2010). The knowledge generalization also aims to provide the core concept of causality knowledge as the answer source for the problem diagnosis through the Why-QA system (under the closed-domain QA system). Yeh et al.,2008 worked on the photo-based question answering system, especially the What and Where question types, where the information retrieval was applied for finding the possible answers from web sites. However, our EIWhy-QA system allows web base submissions of a textual Why question including an image (see Fig. 1) from the users. Meanwhile the corresponding answers of the EIWhy question for supporting the plant disease diagnosis must be obtained from the scientific research papers or the technical documents (in agriculture) that have been accepted by the specialists, such as our answer source.

The EIWhy question is emphasized on the corresponding answers based on the causality between a causative event and an effect event, which can be represented by a causative verb concept set (V_c) and an effect verb concept set (V_e) respectively, (see Table 1). This causality is also expressed in the form of the inter-causal EDUs (where EDU is the elementary discourse unit or a simple sentence/clause and the inter-causal EDUs is the causal relation between one/multiple causative EDU(s) and one/multiple effect EDU(s)) (Pechsiri and Kawtrakul, 2007). For example, the extracted causality in the answer source:

Causative unit: EDU1 "*ถ้าเพลี้ยทำลายต้นข้าว* / *If the aphids infest rice plants*,"

Effect unit (EDU2+EDU3+EDU4) :

EDU2 "*จะทำให้ใบเหลือง* / [*it*] *will make the leaves become yellow*."

EDU3 "*ต่อมาหงิกงอ* / *Then [the leaves] shrink*"

EDU4 "*และต้นข้าวจะหยุดการเจริญเติบโต* / *and the rice plants will stop growing*."

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

Table 1 Causative Verb Concept Set (V_c) and Effect Verb Concept Set (V_e) (Pechsiri and Kawtrakul, 2007)

Verb Type	***Surface form***	**V_c**
Causative Verb	*ดูด/suck, ดูดกิน/suck. กิน/eat, กัด/bite,*	*suck,eat,bite/ consume/ destroy*
	ทำลาย/destruct, กำจัด/eliminate, ฆ่า/kill, หัก/break, ระเบิด/explode, บุกรุก/infest	*destruct,eliminate, break,explode,infest/ destroy*
	เป็น+โรค/ be+ disease,	*be disease/ get dis ase*
	ได้รับ+เชื้อโรค/get+ pathogen,	*get pathogen*
	ติด+เชื้อ/contract	*contract/ infect*
	เกิด/occur, ปรากฏ/appear	*occur/ appear*
	.....	.....
Effect Verb	**Surface form**	**V_e**
	หงิก /shrink, งอ/bend, บิด/twist, โค้งงอ/curl	*shrink, bend, twist, curl/ beAbnormalShape / beSymptom*
	แห้ง/dry, ไหม้/blast	*dry, blast/ beSymptom,*
	เหี่ยว/wilt	*wilt / loseWater/ beSymptom*
	แคระแกรน/stunt	*stunt/ notGrow/ b Symptom*
	ร่วง/drop off	*comeOff/ beSymptom*
	เหลือง/be yellow	*beYellow/ beAbnormalColor/ beSymptom*
	เป็น+จุด/be+spot, มี+จุด/have+spot (be/have a spot)	*beMark , haveMark / beSymptom , haveSymptom*
	เป็น+แผล+รูปตา/ be+lesion+eye-shape , มี+แผล+รูปตา/ have+lesion+eye-shape (be/have an eye-shape lesion)	*beEyeShapeMark/ beSymptom , haveEyeShapeMark/ haveSymptom*
	เป็น+สี/be+Color, มี+สี/have+Color (be/have Color) *เป็น+แผล+สี/ be+lesion+Color* (be/have a Color lesion) where Color ={'สีเหลือง/yellow' 'สีน้ำตาล/brown' 'สีส้ม/orange' 'สีเทา/grey' 'สีดำ/black'..}	*beColor/ beAbnormalColor/ beSymptom , haveColor/ haveAbnormalColor/ haveSymptom , beColorMark/ beSymptom , haveColorMark/ haveSymptom*
	ขยาย/expand, รวม/combine	*ncrease*
	เกิด/occur, ปรากฏ/appear	*occur/ appear*
	.....	.....

Previous causality extraction works were based on the rule/pattern matching approach, the statistical approach, or the pattern and statistics combination (see Section 2). The explicit cue, cue-phrase, or discourse marker, e.g. 'because' 'as the result of' 'and' etc., is necessary for most of the previous research to identify the causal relation or the causality. However, most of their researches do not have the causal-boundary determination and causality generalization. Meanwhile, our research concerns the effect boundary determination without discourse marker because about 30% of discourse markers for the causality are implicit in our corpora. Moreover, the boundary determination followed by generalization is necessary for clearly supporting diagnosis of problems through the Why-QA system.

Working on Why-QA emphasizing on the explanation knowledge especially on events is different from working on other wh-QA (such as who, what, and where)which emphasizes on name entities or noun phrases (Verberne, 2006). Previous Why-QA works were based on reasoning (Burhans and Shapiro, 2001) (Verberne, 2006) and discourse structures (Verberne et. al., 2007) (see Section 2)

There are three main problems involved with this research. **The first problem of the the causality extraction**, especially the boundary determination with the problem of the V_c V_e intersection which affects to the boundary determination, is confronted (see Section 3) after applying the verb pair (a causative verb and an effect verb) rules from (Pechsiri and Kawtrakul, 2007) to identify the causality. The previous research (Pechsiri and Kawtrakul, 2007) applied the linguistic rules as Centering Theory(Walker et. al., 1998) to determine the boundary of the effect EDUs which performed inefficiently in some domains (see Section 2). Therefore, we apply the verb feature adjustment rules to identify a verb element in the V_c V_e intersection as the causative concept or the effect concept before learning the effect boundary from effect verb features by different machine learning techniques for comparative study; Maximum Entropy (ME) (Csiszar, 1996), Supported Vector Machine (SVM) (Cristianini and Shawe-Taylor, 2000), and Naïve Bayes (NB) (Mitchell. 1997). **The second problem** is the knowledge generalization problems where some extracted inter-causal EDUs contain the linguistic uncertainty of the nuance expression and the incompleteness causality knowledge (see Section 3). We propose applying the basic linguistic rules to solve the linguistic uncertainty problems and the Monte Carlo simulation technique (Woller.,1996) to solve incompleteness problems by the random number to fulfill the missing data of the effect unit. And then, we apply the fuzzy function right after the data fulfillment or imputation to determine the generality value of each effect event expressed by the effect verb concept of the effect EDU from several documents having the same cause. The causality with each generality value of each effect-verb concept are provided as the knowledge base for answering the Why question. Furthermore, the generality value from the fuzzy logic can represents subjective belief of the effect-verb concept feature being better than the crisp data set which is the binary logic.

The Why-question part from the user (notifying a problem) consists of two portions; a textual portion and an image portion which provides the Why question contents of the disease symptoms.

The third problem is the problems of EIWhy-QA **system** which consist of two main parts of problems; the EIWhy question part and the EIWhy answering part. The EIWhy question part consists of three problems; first is how to identify a Why-question type from the textual portion of the EIWhy question with the problem of the question word ambiguity (see Section 3). Identifying the textual-question expression without using the question symbol (i.e. '?'), commonly practiced in some languages as in our research is a challenge. Previously, wh-questions have been approached by determining answers from noun phrases and question words (Verberne, 2006), which is suitable for the causal question or the Why question with the

answer based on the lexico-syntactic pattern (Girju, 2003) as NP1 Verb NP2 (where NP1 and NP2 are the noun phrases), i.e. “What causes Tsunami? → Earthquake causes Tsunami”. However, it is not suitable for both the Why questions with the answers based on the explanation as in our research (see Section 2) and other non-factoid questions as portrayed (Verberne, 2006; Verberne et. al., 2007; Verberne et. al., 2008; Pechsiri et. al.,2008; Quarteroni and Saint-Dizier, 2009). Therefore, our research proposes using a Why-question cue set (YQC) to determine the Why-question type from the testing corpus instead of using only the question words (see Section 3). Second is how to determine Why-question contents from the image portion of the EIWhy question. According to (Weizherg et al., 2008), the detection and identification of plant disease in practice is always performed by the naked eye observation of experts. This approach is expensive and time consuming because it requires an expertise from the experts. It could be improved by the assistance of advancement technology. In (Patil and Kumar, 2011), the technology that most research focused on is automatic detection of plant diseases by analyzing symptom and observing the lesion on the leaves or stems of the plant. Therefore, We also propose the use of image processing techniques as a Bag-of-Visual-Words (BOW) to represent the region of interest (ROI) on the image as the Why-question contents where a visual word is a small patch on the array of pixels containing the interesting feature space of color, texture...etc.(Sivic et al., 2005) and (http://en.wikipedia.org/wiki/Visual Word#cite_note-val u99-3). Their BOW (Sivic et al., 2005) represents an image containing several patches or several visual words whereas the BOW in our research represents the ROI on an image where the ROI contains several visual words of lesion shape, lesion color, image background color, and image background texture (see Section 3). We also apply the symptom-concept-frame structure to interpret the ROI’s BOW to the conceptual predicate content of the EIWhy question. And, third is how to determine the EIWhy question’s focus, especially on the root cause determination. We apply v_e (where $v_e \in V_e$, see Table 1) to determine the Why-question focus based on an event mostly expressed by a verb or a verb phase.

The second part of problems is the EIWhy answering part as how to determine the corresponding Why answers from the knowledge source. We apply the similarity scores between the conceptual predicate content and the EDUs from the cause-effect vector of the knowledge source (Pechsiri and Piriyakul, 2010) after stop-word elimination, to solve the Why answer.

2. Literature Review

Other related works to address the Why-QA techniques and the knowledge extraction from text followed by the knowledge generalization as the answer source have been involved with NLP (Natural Language Processing), image processing techniques, and machine learning techniques. To have better understanding, we need to mention the related work aspects of the answer supporter provided by the causality knowledge extraction followed by generalization first and then the Why-QA system.

Causality Knowledge Extraction

In 1995, Khoo used linguistic patterns from Wall Street Journal, e.g. '[Noun-phrase: effect] is due to [Noun-phrase: cause]', '[Clause: effect] because [Clause: cause]', and etc., and cues, e.g. 'because', 'since', 'due to' and etc., to extract causal relations within one or two adjacent sentences without any cause/effect boundary determination from documents with 64% precision and 68% recall.

Marcu and Echihabi (2002) presented the unsupervised learning methodology of Naïve Bayes classifier (NB) to recognize the discourse relations by using word pair probabilities between two adjacent sentences or clauses for identifying the rhetorical relation, such as "Contrast", "Cause-explanation Evidence" (or causal relation), "Condition", and "Elaboration". The result of extracting the causal relation based on two adjacent sentences without any cause/effect boundary determination from the BLIPP corpus showed 75%precision.

Inui et al. (2004) proposed extracting causal knowledge from two adjacent sentences or clauses (without any cause/effect boundary determination) by using the explicit connective markers, e.g. 'because', 'if...then', etc., with the problem of the connective marker ambiguity for classifying the casual relation types. Support Vector Machine (SVM) was used for solving their problem and their precision is as high as 90% but the recall is as low as 30% because of unsolved anaphora.

However, the techniques from (Khoo, 1995), (Marcu and Echihabi,2002), (Inui et al., 2004) cannot be applied to our causality knowledge extraction with the inter-causal EDU for clear explanation of cause and effect. Then, Pechsiri and Kawtrakul (2007) proposed verb-pair rules learned by two different machine learning techniques (NB and SVM) to extract causality with multiple EDUs from a causative unit and multiple EDUs from an effect unit. The verb-pair rules have been represented by Eq. (1) (see Section 1). Each causative verb concept (v_c , where $v_c \in V_c$) and each effect verb concept (v_e , where $v_e \in V_e$) are referred to WordNet24

(http://wordnet.princeton.edu/) and the predefined plant disease information from the Department of Agriculture (http://www.doa.go.th/).

(Pechsiri and Kawtrakul, 2007) also proposed to use V_c and V_e (Table 1) to solve the boundary of the causative unit and using Centering Theory (Walker et. al., 1998) (which is the center of attention from a discourse segment, and is expressed by a noun) a long with V_e to solve the boundary of the effect unit. How to apply Centering Theory in (Pechsiri and Kawtrakul, 2007) is whenever the transition state of the center of attention is the smooth shift occurrence (the attention agent, mostly being a subject of a sentence, is changed), the boundary ends. For example: “If the brown Leaphopper aphids suck sap from rice plant, leaves will be yellow. [Leaves] shrink. These aphids destroy plant very fast.” The effect boundary ended at ‘[leaves] shrink’ because the next center of attention is changed to ‘aphids’. However, sometimes there are some inter-causal EDUs containing effect units with the smooth shift occurrence even when the boundary is not ended. For example, “The earthquake occurred in China. It caused many buildings were collapsed. Public utilities were cut down. More than 100 people died.”, where ‘buildings’, ‘Public utilities’, and ‘people’ are in the effect boundary with different attentions. Then, Pechsiri and Piriyakul (2010) solved the effect boundary determination problem by learning the v_e v_{e+1} pairs with ME(Maximum Entropy) comparing to BN (Bayesian Network) where ME has the better results. Finally, the major outcomes of their research are the verb-pair rules, with the correctness of the causality-boundary determination varied from 80% to 96% depending on the corpus behaviors, especially the global warming corpus (which Centering Theory could not be applied efficiently). However, Pechsiri and Piriyakul (2010) still have the problem of the V_c V_e intersection which affects to the causative boundary determination and the effect boundary determination. Therefore, we apply the verb feature adjustment rule for solving the verb feature intersection before learning $v_e v_{e+1}$ pairs by ME, SVM, and NB with a sliding window size of two EDUs and sliding with one EDU distance.

Generalization

Mitchell et al., (1986) Most researchers have proposed generalization methods that contrasted sharply with the data-intensive, similarity-based methods and relying on many training examples with an inductive bias to constrain the search for a correct generalization. Whereas Mitchell et.al., (1986) ‘s method relied on Knowledge of the specific domain with a single training example. Their methods were based on the knowledge of concepts (in the hierarchy structure) which were generalization, and were called “Explanation Based Methods” providing a more reliable means of generalization and being able to extract more information from individual training examples.. These methods analyze the training example by first

constructing an explanation of how the example satisfies the definition of the concept under study.

Angryk and Petry (2003) stated that Data generalization is a process of grouping of data, enabling transformation of similar data collections, expressed originally in a database at the low (primitive) level, into more abstract conceptual representations. They focused on attribute-oriented induction in data mining with fuzzy generalization model through concept hierarchies. They have defined a fuzzy concept hierarchy (FCH) as an order pair (*C,L*), where *C* is a set of concepts utilized to generalize a particular domain and *L* is a set of links between these concepts, reflecting ideas applied for the generalization process. Each concept c has its unique name (Label) and abstraction level, placing it on a specific height of the generalization hierarchy. This can simplify notation as v^j and to refer directly to the specific concept at the given level of the generalization hierarchy by using v_i^j , where i symbolizes the index of the concept v at the f^{th} abstract level. If j=0 then c is the surface form and l (where $l \in L$) is the link between concept nodes s and t at the concept level j and $j+1$ respectively with the membership function (μ_{st}) under hierarchical generalized concepts. These concepts are classified into 5 levels starting from 0 to 4. There is fuzzy sets, {white, lightGrey, gray, darkGrey, black} and { light, dark} at the 1st level and the 3rd level, respectively. Furthermore, Angryk and Petry (2003) reduced two hierarchy levels for effectively determining the generalized color with the sum of weights at all links leaving the color concept in the fuzzy hierarchy remained exactly 1 (means the completeness of the generalization model after its height reduction)

However the research works of (Mitchel et.al., 1986) and (Angryk and Petry 2003) can not applied to our research because their completeness data in the database format whereas there is incompleteness in our matrix data. Therefore, we propose using the Monte Carlo simulation for the data imputation followed by fuzzy generalization since there are the uncertainty nuance and the incompleteness in our matrix data.

Why Question Answering system

Verberne(2006), working on Why-QA emphasizing on events is different from working on other wh-QA (such as who, what, and where) which emphasizes on object or name entities. Previously, wh-questions have been approached by determining answers from noun phrases and question words (Verberne, 2006), which is suitable for the causal question with the lexico-syntactic pattern based answer as NP1 Verb NP2 (Girju, 2002; Girju, 2003; Vazquez-Reyes S. and Black W.J., 2008) (where NP1 and NP2 are the noun phrases with the cause concept and the effect concept, respectively), e.g. “What causes Tsunami? Earthquake causes Tsunami”. However, wh-QA is not suitable for other non-factoid questions as portrayed by previous

literatures (Verberne et al., 2007; Verberne et al.,2008) including the Why questions with the answers based on the explanation as in our research, for example:

Why-Question EDU: “ทำไม/Why ใบมะม่วง/mango leaves หงิก/shrink”

(Why do mango leaves shrink?)

Answer: EDU1 “เพลี้ย/Aphids ทำลาย/destroy ใบมะม่วง/mango leaves”

(Aphids destroy mango leaves.)

EDU2 “ทำให้/make ใบ/leaves หงิก/shrink ” ([it] makes leaves shrink.)

Verberne S. et al (2007) proposed using RST (Rhetorical Structure Theory) structures to approach Why questions by matching the question topic with a nucleus in the RST tree while yielding the answer from the satellite. The author compared manual RST analysis with a system constructed using Perl script where the likelihood of the nucleus and the discourse relations are calculated. The RST approach to the Why-QA system achieved the answer correctness of 91.8% and a recall of 53.3%. However, their technique would not perform effectively with frequent occurrences of zero anaphora in our corpus.

Cheng J. (1996) discussed the role of entailment in knowledge representation and reasoning. According to Cheng J. (1996), entailment is the abstracted notion in conditional sentences, which is the usual sentence form of causal relations and is often used in abduction. Entailment also plays an important role in determining the validity of the abduction, since the key of abduction is “how to get and use genuine logical entailments that are certainly relevant to the premise”. The validity of abduction is vital when dealing with diagnosis since diagnosis is usually framed as abduction where the cause is often inferred from the effect (Kate and Mooney, 2009). In addition to (Carlson et. al., 2003), when abduction is viewed as a type of question answering, abductive hypotheses can be seen as a subset of hypothetical (or conditional) answers. According to (Druzdzel, 1993), abduction is a type of uncertain reasoning that generates hypothesis, and hence the certainty factor can be applied. However, (Yamada, 1995) suggested using the possibility theory in Fuzzy for abduction in diagnosis due to the risk involved in the determination of the causes of symptoms in diagnosis. Thus, our research proposes to apply the verb-pair rules in term of abduction to answer the Why questions with the calculation of the possibility values which are the generality values from the fuzzy concepts with some incomplete knowledge expressed on text for bringing up answering with confidence

Image Processing Application

Sivic et al. (2005), their image layout was analogous to topic determination in text by using the bag of words or BOW. Thus, the visual words were applied to determine the image topic. Their visual words were formed by vector quantizing the local appearance descriptors of

images. The probabilistic Latent Semantic Analysis (pLSA) of Hofmann using the bag of 'visual words' representation was applied to determine the object categories as the topics. Their results of the topic determination approach were successfully to identify the object categories for each image with the high reliabilities. However, our research applies the BOW to determine the Why question contents.

Yeh et al.(2008), worked on the photo-based question answering system based on five categories of images: books, movies, groceries, modern landmarks and classical landmarks, where a question is expressed by both a photo/image as an object and a caption or text as an image-searching scope. The information retrieval was applied for searching the possible answers from web sites or the internal repository of resolved photo-based questions. However, unlike the more holistic information retrieval approach, our research derives answers from specific facts within a specific domain

Patil and Kumar (2011) discussed the role of image processing in agricultural. They concluded that it can be used for detecting diseased plant, quantifying affected area and finding shape and color of affected area. Woodford et al. (1999) proposed using wavelet transform technique and neural network to help identify pest damages in fruit. In additional, Ei-Helly et al. (2004) proposed a novel approach to integrate image analysis technique into diagnostic expert system for plant diseases. However, the objective of their system is for plant disease classification only. Another interesting approach purposed by Kaundal et al. (2006), they developed weather based prediction models of plant diseases using SVM. While Weizheng et al. (2008) focused on how to grading of grape leaf disease by calculating the quotient of disease spot and affected leaf area, Meunkaewjinda et al. (2008) tried to classify grape leaf disease using self organizing map and back propagation neural networks. Ying et al. (2008) proposed a method of image pre-processing for crop diseases and also suggested effective characteristic parameters for the disease diagnoses.

As mention above, most of research focused on detection of the plant diseases using both image processing techniques and machine learning techniques. None of them, however, exploited the usage of plant disease detection and classification to find a root cause of the disease, which help us find a solution on how to cure for the disease.

3. Crucial Problems

In order to achieve the knowledge generalization of the extracted causality from text as the answer source for the EIWhy-QA system in diagnosis, there are three main areas of problems; the causality-knowledge-extraction problem of boundary determination, the causality-knowledge-generalization problems of the uncertainty linguistic expressions and the incomplete knowledge, and the EIWhy-QA problems of the question-word ambiguity, how to determine the Why question contents from an image of an EIWhy question, how to determine the Why-question focus, and how to determine the corresponding EIWhy-answer.

3.1 Causality-Knowledge-Extraction Problem of Boundary Determination

According to the linguistic expression, EDU can be expressed as:

EDU → NP1 VP | NP1 V NP2

NP1, NP2 → N NP | N

NP → N NP I N

VP → V NP2 | V

V → verb

N → noun

where NP1 and NP2 are noun phrases, VP is a verb phrase.

(Pehsiri and Kawtrakul,2007) and (Pechsiri and Piriyakul,2010) applied the verb-pair rules using V_c and V_e (see Table 1) to identify the causality and the causative boundary where V_c and V_e are equivalent to VP in the above EDU expression. Then, (Pechsiri and Piriyakul,2010) learned v_e v_{e+1} pairs by ME and BN to determine the effect boundaries. Even though ME has the highest precision of 96% of the effect boundary determination, they still have the problem of the V_c V_e intersection which effects to determine both boundaries. Therefore, the following verb-noun co-occurrence patterns are applied before learning $v_e v_{e+1}$ pairs by ME, SVM, and NB with a sliding window size of two EDUs and sliding with one EDU distance through the document.

Verb Feature (V_c and V_e) Adjustment Rules:

If $v_{EDUx} \in V_s \wedge np1_{EDUx} \in$ {'disease' 'pathogen' 'plant louse', 'phenomenon',..} then $v_{EDUx} \rightarrow v_c$
If $v_{EDUx} \in V_s \wedge np1_{EDUx} \in$ {'symptom', 'casualty'. 'damage',..} then $v_{EDUx} \rightarrow v_e$

where: $V_s = V_c \cap V_e$ = {'occur' 'appear} (see Table1)

v_{EDUx} is a verb element of VP in EDUx and $np1_{EDUx}$ is a noun element of NP1 in EDUx where EDUx is the first detected EDU as the causality from a document by the causality extraction algorithm (Pechsiri and Kawtrakul,2007) and (Pechsiri and Piriyakul,2010).

3.2 Causality-Knowledge- Generalization Problems

In addition to the extracted causality knowledge from texts as the answer source of Why questions, there are two problems existing in the extracted causality knowledge, the uncertainty of the linguistic expression on nuance and the incomplete causality knowledge expression.

3.2.1 *Uncertainty of Linguistic Expression on Nuance*

One of the uncertainty linguistic expressions from the extracted causality knowledge is fuzzy data of nuance which mostly occurs in the extracted causality. For example:

Report1 EDU1: "*เพลี้ยทำลายใบข้าว /Aphids destroy rice leaves*."

EDU2:"*ทำให้ใบเกิดจุดสีเทาอ่อนปนเขียว*/[*It*] *makes leaves have greenish light grey spots.*"

Fuzzy data

Report2 EDU1: "*เพลี้ยทำลายใบข้าว /Aphids destroy rice leaves*."

EDU2: "*ทำให้ใบเกิดจุดสีเทาปนเขียว*/ [*It*] *makes leaves have greenish grey spots*."

Fuzzy data

The expression of nuance is varied in various reports/documents causing the problem of answering the Why questions whose effects are not contained in the documents used for knowledge extraction. For example:

"*ทำไมใบข้าวมีจุดสีเทา*/ *Why do rice leaves have grey spots*?"

(whilst "grey spots", "greenish light grey spots", and "greenish grey spots" are the same object but different expressions by several writers.)

This nuance expression problem can be solved by using the nuance concept from the linguistic head noun pattern to align the nuance expression in the question to the nuance expression in the extracted causality

3.2.2 Incomplete Causality Knowledge

The problem of incomplete causality knowledge commonly occurs when certain effect verbs or symptoms are not mentioned consistently in the corpus from the same causing agent as shown in the following Document1 and Document2.

Document1 EDU1: "*หากเพลี้ยทำลายใบพืช /If aphids destroy leaves*."

EDU2: "*ทำให้ใบแห้ง*/ [*it*] *makes leaves dry*."

EDU3: “*และร่วง/ and* [*leaves*] *come off*.”

Document2 EDU1: “*หากเพลี้ยทำลายใบพืช /If aphids destroy leaves*.”

EDU2: “*ทำให้ใบร่วง/* [*it*] *makes leaves come off*.”

Therefore, we propose using the Monte Carlo simulation technique for imputing the effect events to solve the incompleteness problem. After the imputation technique have been done, the fuzzy function needs to process for determining a generality value of an effect event expressed by an effect verb concept, v_e, from each cause agent type (e.g. aphid, virus, fungi, .. etc.). Then, the extracted causality knowledge generalization from texts with the generality value determination of effect-verb concepts are provided as the answer source of the Why-QA system for the problem diagnosis. Furthermore, the generality value from the fuzzy logic can represents subjective belief of the effect-verb concept feature being better than the crisp data set which is the binary logic.

3.3 EIWhy-QA Problems

The research contains two major parts of problems: the EIWhy question part and the EIWhy answering part

3.3.1 EIWhy question part

There are three problems as how to identify a Why-question type (from the textual portion of the EIWhy question with the question word ambiguity), how to determine Why-question contents (from the image portion of the EIWhy question), and how to determine the EIWhy question’s focus.

3.3.1.1 *How to identify a Why-question type*

The problem of identifying the question expression without using the question symbol (i.e. ‘?’) is solved by using the question words or the wh-question word set {‘*ทำไม*/*Why*’, ‘*อะไร*/*What*’, ..}; where a ‘*ทำไม*/*Why*’ function is a reasoning question, a ‘*อะไร*/*What*’ function is asking for information about something, (http://www.englishclub.com/vocabulary/wh-question- words.htm). However, there is a wh-question word’s function ambiguity, e.g. ‘*อะไร*/*What*’ as in reasoning: “*เกิดผลกระทบอะไรเมื่อเพลี้ยทำลายพืช*/*What are the effects when aphids destruct plant*?”. Therefore, our research proposes using a Why-question cue set (YQC) to determine the Why-question type from the testing question corpus.

YQC = YQWord ∪ YQCuephrase

where YQWord is a Why-question-word set, and YQCuephrase is a Why-question-cue-phase set.

YQWord = {'*ทำไม/Why*', '*เหตุใด/Why*' ,..}

YQCuephrase = {'*...ผลลัพธ์จาก/result(s)from+อะไร/what*', '*...ผลลัพธ์/result(s)+อะไรบ้าง/what...*', '*…ผลกระทบจาก/effect(s)from+อะไรบ้าง/what*', '*…ผลกระทบ/effect(s)+อะไรบ้าง/what…*', '*…สาเหตุ/cause(s)+อะไรบ้าง/*what…', '*..เพราะ/reason+อะไร/what…*', '*…เกิดจากสาเหตุ/be caused by+อะไร/what*', '*...ส่งผล/affect+อย่างไร/how...*', '*อะไร/**what**+เป็น/**be**+สาเหตุ/**cause***..', ..}

For example: Determine a Why-question type by using YQC:

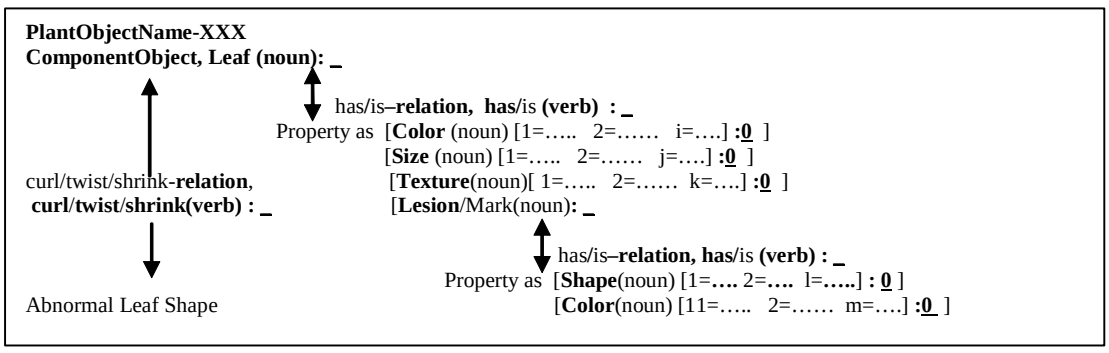
"*อะไรเป็นสาเหตุให้ใบข้าวมีอาการผิดปกติ*"

"*อะไร**/what** เป็นสาเหตุให้/**is the cause** that makes ใบข้าว/rice leaves มี/have อาการผิดปกติ/unusual symptom*"

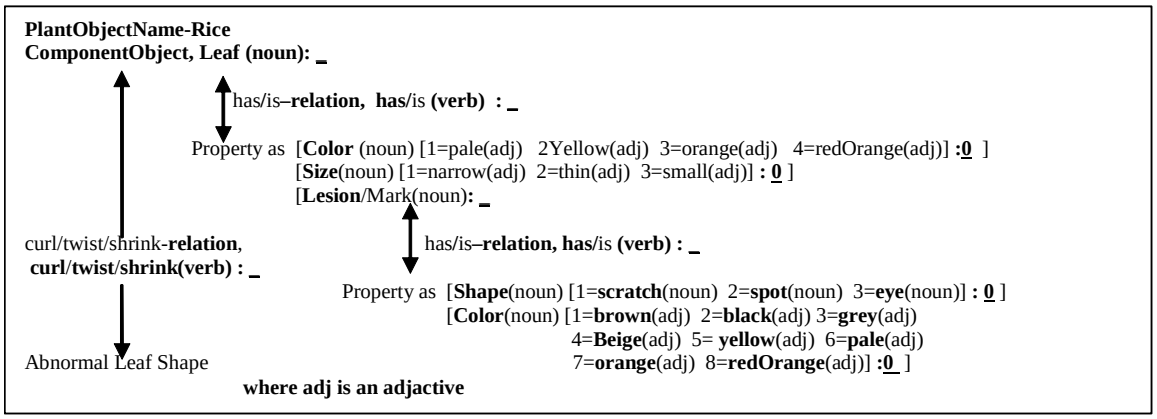
(***What is the cause*** *that makes the rice leaves have an unusual symptom*?)

3.3.1.2 *How to determine Why-question contents*

Since the textual portion of the EIWhy question is expressed in the general concept of the diagnostic problem, for example: the '*symptom*' word of Fig. 1(a) on the textual portion of the EIWhy question is the general concept of an *eyeShape* lesion with brown color expressed on the image portion of the EIWhy question (see Table 1). Therefore, we propose using the BOW to represent a ROI on the image portion of the EIWhy question followed by the symptom-concept-frame structure (see Fig. 2) to interpret the BOW to the conceptual predicate content of the EIWhy question. Fig. 2(a) shows a general symptom-concept-frame structure of leaf symptoms which consist of properties and relations (where a property is expressed by a noun phrase and a relation is expressed by a verb/a verb phrase). Fig. 2(b) is an example of the symptom-concept-frame structure of the rice leaf symptoms, contains three main symptom features (Leaf Color, Leaf Shape, and Leaf Lesion/Mark) with the default "zero" value or null.



(a) A general symptom-concept-frame structure of leaf symptoms



(b) Example of a symptom-concept-frame structure of rice leaf symptoms

Fig.2 Show a symptom-concept-frame structure

3.3.1.3 *How to determine the EIWhy question focus*

The determination of the EIWhy question focus is necessary for the answer determination. The EIWhy question focus is always expressed by the effect event, especially for the root cause determination, mostly represented by a verb/a verb phase based on V_e. Thus, the focus of EIWhy question can be determined from V_e.

3.3.2 EIWhy answering part

The problem of this part is how to determine the corresponding answer to the EIWhy-question content from the image portion. However, it is unlike wh-questions from text-based questions, the answer of the ImageWhy question can not be determined by the question word (*qw*). For example:

a) Q : Who is the president of USA? Ans: Obama is the president of USA.

b) Q: "*ทำไม /Why ใบมะม่วง/mango leaves หงิก /shrink*" (Why do mango leaves shrink?)

Ans: EDU1 "*เพลี้ย/Aphids ทำลาย/destroy ใบมะม่วง /mango leaves*" (Aphids destroy mango leaves.)

EDU2 "*ทำให้/make ใบ/leaves หงิก/shrink*" ([it] makes leaves shrink.)

The answer of the question in a) can be determined by a question word "Who" (Agichtein et al., 2005) whereas the question word "Why" cannot be applied to determine the answer in b). Moreover, wh-questions have previously been approached by determining answers from noun phrases and question words (Verberne, 2006), which is suitable for the Why question with the answer based on the lexico-syntactic pattern (Girju , 2003) 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 the ImageWhy-QA system mostly based on several effect-event explanations which

are always expressed by verbs/verb phrases. And, it is not suitable for other non-factoid questions either as portrayed by (Verberne, 2006 ; Verberne et al., 2007; Verberne et al., 2008; Pechsiri and Piriyakul, 2012; Quarteroni and Saint-Dizier, 2009). Therefore, we use the similarity scores between the Why-question content and EDU_{effect} from the cause-effect vector to determine the root-cause answer of the EIWhy question. Where all word concepts are referred to WordNet (http://wordnet.princeton.edu/) and the predefined plant disease information from the Department of Agriculture (http://www.doa.go.th/) including Encyclopedia (http://kanchanapisek.or.th/kp6/New/) after using the Thai-to-English dictionary (Longdo.com)

4. Research Method

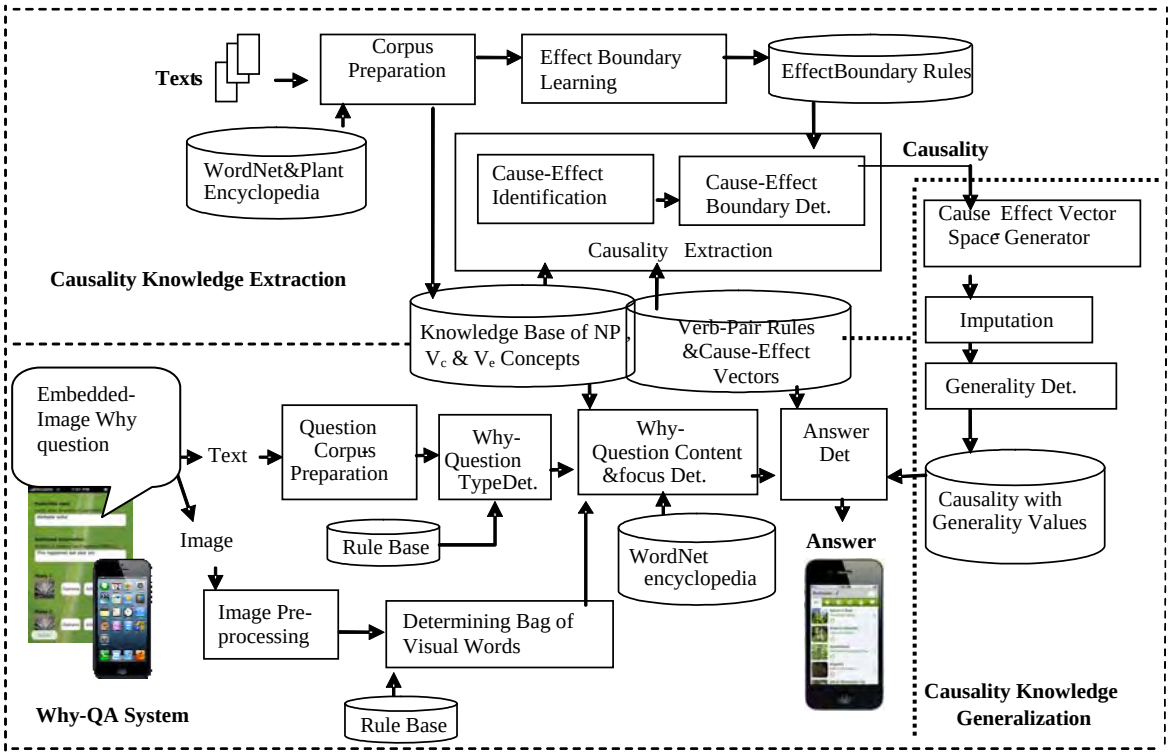


Fig. 3 System architecture

There are two layers in our System Architecture (see Fig. 3), the first layer is the Causality Knowledge Extraction system (with the verb feature adjustment) as the source of the EIWhy answers. The second layer is the EIWhy-QA system which can be separated into two parts: the EIWhy question part and the EIWhy answering part which includes the Causality Knowledge Generalization providing the effect/symptom generality value for supporting diagnosis

4.1 Causality Knowledge Extraction

There are three steps in the causality knowledge extraction part. First is a Corpus Preparation step followed by an Effect-Boundary Learning step. And, the next step is Causality Extraction.

4.1.1 Corpus Preparation

The corpus preparation are similar to [6] where 4000 EDUs of the agricultural domain of plant disease documents, and the news domain of global environment involves using a Thai word segmentation tool (Sudprasert and Kawtrakul,2003) including Name Entity (Chanlekha and Kawtrakul, 2004) followed by EDU segmentation(Chareonsuk et al., 2005). Then, all inter-causal EDUs semi-automatically annotate with the causative/effect verb concepts from Table1

(referred to WordNet (Miler et al., 1993) after using Thai to English dictionary (www.longdo.com) and the predefined plant disease information from Department of Agriculture (http://www.doa.go.th/)) shown in Table 1. Fig.4 shows the annotation example of the inter-causal EDU.

(<C id=1 type=causality>
<EDU>เมื่อ /When <npEntity concept = 'plant louse'> เพลี้ยกระโดดสีน้ำตาล /leaf hoppers </npEntity> <VC concept= 'consume'> ดูดกิน/suck </VC> < npEntity concept ='solution' > น้ำเลี้ยง/sap </npEntity > ของต้นข้าว /of rice plant</EDU></C>
<R id=1 >
<EDU>จะทำให้ต้นข้าว/will make the rice plant <VE concept='be symptom'> มีสีเหลือง/have yellow color</VE> </EDU>
<EDU> และ/and<VE concept='prevent growth'> แคระแกรน/stunt</VE> </EDU> </R>)
EDU = Elementary Discourse Unit tag C = causative tag, R=result tag, VC=causative verb tag, VE=effect verb tag, npEntity = noun phrase entity tag

Fig.4 Example of Causality Knowledge Annotation.

In addition to the causality extraction, 4000 annotated EDUs is divided into 2 groups: 3000 annotated EDUs for effect-boundary learning and 1000 annotated EDUs for the effect-boundary evaluation. Moreover, the annotated concepts of verbs and nouns with their surface forms are kept as the knowledge base for generating the cause-effect vector space in the causality-knowledge-generalization part.

4.1.2 Effect-Boundary Learning

The effect boundary is learned by using different machine learning techniques for the comparative study; ME, SVM and NB are shown in the following (based on ten-fold cross-validation):

Maximum Entropy (ME): The best model, ME, is consistent with the set of constrains imposed by the evidence, but otherwise is as uniform as possible (Csiszar, 1996; Berger et al., 1996). (Fleischman , 2002) modeled the probability of a semantic role r given a vector of features x according to the ME formulation below:

$$p(r \mid x) = \frac{1}{z_x} \exp[\sum_{j=0}^{n} \lambda_j f_j(r, x)] \qquad (2)$$

where Z_x is a normalization constant, $f_i(r,x)$ 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. The final classification is the role with highest probability given its feature vector and the model. According to Eq. (2), ME can

be used as the classifier of the r class when the probability $p(r|x)$ is argmax $p(r|x)$ to determine the effect boundary classes. Where r is the effect boundary classes (boundary is ending when r =0, otherwise r =1) and x is the binary vector of the effect-verb concept (v_e) features containing an effect-verb concept pair ($v_{ei}v_{ei+1}$), where $v_{ei} \in V_e$ and $v_{ei+1} \in V_e$), as shown in Eq. (3). All pairs of $v_{ei}v_{ei+1}$ are gained by sliding a window size of two adjacent effect EDUs with one EDU distance through the effect EDU unit (Eq. (3) where λ_j are shown in Table 2).

$$p(r \mid x) = \underset{r}{\operatorname{argmax}} \frac{1}{z} \exp\left(\sum_{j=1}^{n} \lambda_j f_{yes,ei,j}(r, v_{ei}) + \sum_{j=1}^{n} \lambda_j f_{no,ei,j}(r, v_{ei}) + \sum_{j=1}^{n} \lambda_j f_{yes,ei+1,j}(r, v_{ei+1}) + \sum_{j=1}^{n} \lambda_j f_{no,ei+1,j}(r, v_{ei+1})\right) \quad (3)$$

Table 2 Show λ_j of v_e from the plant aphid documents

v_e	λ
beAbnormalShape (leaf)	-4.4553
stunt(plant)	-3.0448
occur(symptom)	0.2294
dry(leaf)	-4.754
…..	…..

Support Vector Machine (SVM) (Cristianini and Shawe-Taylor, 2000) is a binary or linear classification applied in this research to classify the effect boundary ending of each effect verb pair, $v_{ei}v_{ei+1}$, gained by sliding a window size of two adjacent effect EDUs with one EDU distance through the effect EDU unit of the learning corpus. According to (Cristianini and Shawe-Taylor, 2000) this linear function, $f(x)$, of the input $x = (x_1 \ldots x_n)$ assigned to the positive class if $f(x) \geq 0$, and otherwise to the negative class if $f(x) < 0$, can be written as

$$\begin{aligned} f(x) &= \langle w \cdot x \rangle + b \\ &= \sum_{i=1}^{n} w_i x_i + b \end{aligned} \quad (4)$$

where x is a dichotomous vector number, w is a weight vector, b is a bias, and $(w,b) \in R^n \times R$ are the parameters that control the function. The SVM learning is to determine w_i and b for each verb feature (x_i) in effect verb pairs from the annotated corpus (Table 3).

Table 3. List v_e features and w_i by SVM learning

v_e	w of v_e
wilt(leaf)	1
dry(eaf)	0.4007
die(plant)	0.3996
reduce(yield)	0.3993
occur(symptom	-1.5992
change_color(leaf)	0.4004
beAbnormalShape (leaf)	0.4005
……..	……

Naïve Bayes (NB): According to (Mitchell, 1997), we applied NB for learning to classify the boundary of the effect EDUs as a verb concept vector (V_{ei}) in matrix vector V_e

V_{ei} = {v_{ei1}, v_{ei2}….$v_{ei\,k}$ end/not-end} where end is the end of the effect boundary., and not-end is the continue of the effect EDU.

V_e = {V_{ei}} where i=1..n

After we have obtained the effect verb features, we then determine the effect verb probability of end/not-end relation from a sliding window size of two effect verbs from consecutive EDUs with the one-EDU sliding distance, shown in Table4, by using Weka (http://www.cs. wakato.ac.nz/ml/ weka/).

Table 4. Show Probabilities of Effect Verbs in the Effect verb pairs

v_{eij}	end	not-end	v_{eij+1}	end	not-end
change_color(leaf)	0.053	0.035	change_color(leaf)	0.007	0.03
beAbnormalShape (leaf)	0.046	0.144	beAbnormalShape (leaf)	0.007	0.084
stunt(plant)	0.038	0.035	stunt(plant)	0.014	0.046
come-off(leaf)	0.061	0.07	come-off(leaf)	0.014	0.088
……	..	..	….	..	..

4.1.3 Causality Extraction

The causality extraction algorithm (Pechsiri and Piriyakul, 2010) has been modified with the verb feature adjustment rule to identify the causality knowledge from text followed by determining the causative boundary (where the causality form is Causative_Unit – Effect_Unit) or the starting effect EDU (where the causality form is Effect_Unit – One_Causative_EDU). Then, the results of the effect boundary learning step from different machine learning techniques used for extracting/recognizing the effect-boundary with the comparative study amoung three machine learning techniques, ME, SVM, and NB from the tested corpus.

Maximum Entropy: From the effect boundary learning by ME, we use λ_j (the weight for a given feature function of the effect boundary with a vector of effect-verb-concept features containing the $v_{ei}v_{ei+1}$ pair) to determine the effect boundary by Eq. (3) with the Effect_Boundary_Determination_by_ME algorithm including where the verb feature adjustment rule is applied (Fig.5).

Support Vector Machine: Then, the effect boundary is determined by using the weight vector and the bias learned by SVM from section 4.1.2. After the causality has been identified, the effect boundary is started to determine by using the $v_{ei}v_{ei+1}$ pair (from a sliding window size of two effect verbs from consecutive EDUs with the one-EDU sliding distance) along with the weight vector and the bias (from section 4.1.2) on Eq. (4) to identify the effect boundary (see Fig.5).

EFFECT_BOUNDARY_DETERMIATON /*by Maximum Entropy

1 **r ←1 /* r** is the effect boundary classes (boundary is ending when r=0, otherwise r=1)
2 **while r=1 do**
3 **If (($v_{ei} \in V_b$) ∧ ($np_1 \in$ {'symptom' 'spread over' 'destruction'})) ∨ ($v_{ei} \in V_e - V_b$)**
4 **If (($v_{ei+1} \in V_b$) ∧ ($np_1 \in$ {'symptom' 'spread over' 'destruction'})) ∨ ($v_{ei+1} \in V_e - V_b$)**
5 **{ case ME**

$$p(r \mid x) = \arg\max_r \frac{1}{z} \exp\Big(\sum_{j=1}^{n} \lambda_j f_{yes,ei,j}(r, v_{ei}) + \sum_{j=1}^{n} \lambda_j f_{no,ei,j}(r, v_{ei}) + \sum_{j=1}^{n} \lambda_j f_{yes,ei+1,j}(r, v_{ei+1}) + \sum_{j=1}^{n} \lambda_j f_{no,ei+1,j}(r, v_{ei+1})\Big)$$

case SVM

$$f(x) = \sum_{j=1}^{n} wt_j x_j + b \quad \text{where } x \text{ is } v_{ei} v_{ei+1}$$

If $f(x) > 0$ then bound ← *yes*

}

6 **If** (r =1) ∨ (bound= 'yes'
EC ← EC ∪ {i}, i ← i + 1,
7 **} return**

Fig.5 Effective boundary determination algorithm by using ME/SVM

Naïve Bayes: The NB Classifier shown in Eq. (5) , with Class={end, not-end}, is applied by sliding the window size of two adjacent effect EDUs with the one-EDU sliding distance along with using verb probabilities in Table 4. Whenever, the determined class is 'end', the effect boundary is end (Fig.6).

$$\begin{aligned} EDUclass &= \arg\max_{class \in Class} P(class \mid v_{eij}, v_{eij+1}) \\ &= \arg\max_{class \in Class} P(ve_{ij} \mid class) P(ve_{ij+1} \mid class) P(class) \end{aligned} \quad (5)$$

$v_{eij} \in Ve_i$ *where* Ve_i *is an effect verb concept vector*

$v_{eij+1} \in Ve_i$ *where* Ve_i *is an effect verb concept vector*

$i = \{1,2,..n\}$ $j = \{1,2,..k\}$

EFFECT_BOUNDARY_DETERMIATON /*byNB

1 **EC←∅** , EDUcl=not-end /*EDUcl is EDUClass for identifying the boundary end
2 **while EDUcl=not-end do** /*effect boundary determination
3 **begin If (($ve_{ij} \in V_b$) ∧ ($np_1 \in$ {'symptom' 'spread over' 'destruction'})) ∨ ($ve_{ij} \in V_{ei} - V_b$)**
4 **If (($ve_{ij+1} \in V_b$) ∧ ($np_1 \in$ {'symptom' 'spread over' 'destruction'})) ∨ ($ve_{ij+1} \in V_{ei} - V_b$)**
5 EDUcl ← $\arg\max_{class \in \{end, not-end\}} P(ve_{ij} \mid class) P(ve_{ij+i} \mid class) P(class)$
6 **if** EDUcl=not-end **then**
7 EC ← EC ∪ {i}; i ← i + 1
8 **end_while ; return R**

Fig.6. Effective boundary determination algorithm by Naïve Bayes

4.2 EIWhy-QA system

4.2.1 EIWhy-Question Part

There are two processing areas of the EIWhy-question part, the text processing for the textual portion of the EIWhy question and the image processing for the image portion of the EIWhy question. Thus, there are several steps involved to the EIWhy-question part: Textual-Question-Corpus Preparation, Why-Question Type Determination, Image Pre-processing, BOW determination, EIWhy-Question-Content Determination, and Question Focus Determination.

4.2.1.1 *Textual-Question-Corpus Preparation*

All 800 textual questions with/without images embedded are collected for this research corpus by interviewing farmers who have the plant disease problems and by downloading from several QA sites and web blogs, e.g. http://www.gotoknow.org/blogs/ books/view/agriculture , with all question types, i.e. 'Why', 'How', 'What', 'When', 'Where', and 'Who', of the rice-plant-disease domains. The collected textual questions consist of 400 textual questions with images embedded, called "the Embedded-Image questions", and the rest 400 textual questions without images embedded (which are the regular Why questions).

Question: "ทำไมยอดใบหงิกงอ /Why do top leaves shrink?"
<Why-Q-type><EDU>[<*Why-Q-Cue*>ทำไม(**Why**)/pint </*Why-Q-Cue*> ยอดใบ(**top leaves**)/ncn]/NP
<*Why-Q-Focus*>[หงิก(**shrink**)/vi]/VP</*Why-Q-Focus*> </EDU></Why-Q-type >
Question: "อะไรเป็นสาเหตุทำให้ต้นข้าวแคระแกรน/ What is the cause making rice plants stunt?"
<Why-Q-type><EDU>[<*Why-Q-Cue*>อะไร(**What**)/pint เป็น(**is**)/vcs สาเหตุ(**the cause**)/ncn</*Why-Q-Cue*>
]/VP</EDU>
<EDU>[ทำให้(make)/vcau [ต้นข้าว(rice plants)/ncn]/NP <*Why-Q-Focus*>[[แคระแกรน(stunt)/vi]/VP
</*Why-Q-Focus*>]/VP </EDU></Why-Q type >

Fig. 7 Examples of Why-Question Annotation for the corpus studies of the Why-question type

In addition to all collected textual questions with/without images embedded, their texts have to be prepared by using a Thai word segmentation tool (Sudprasert and Kawtrakul,2003) with the part of speech annotation including Name Entity (Chanlekha and Kawtrakul, 2004) followed by EDU segmentation (Chareonsuk et al., 2005). The 500 textual questions from the 800 collected textual questions contain 250 Embedded-Image questions and 250 regular Why question (or 250 textual questions without images embedded). These 500 textual questions with/without images embedded are used for a corpus study of the Why-question type by semi-automatically annotating a Why-question type tag (Why-Q-Type), a Why-question cue tag (Why-Q-Cue), and a Why-question focus tag (Why-Q-Focus), see Fig. 7. All concepts are referred to WordNet (http://wordnet.princeton.edu/) and Thai Encyclopedia (http://kanchanapisek.or.th/kp6/New/) after using the Thai-English dictionary (www.longdo.com) for translation. Finally, the

rest 300 textual questions consisted of 150 Embedded-Image questions and 150 regular Why questions (or 150 textual questions without images embedded) are used as a testing corpus of the Why-question type determination.

4.2.1.2 *Why-Question Type Determination*

The Why-question cue set (YQC) is obtained by the results of the corpus study of Why-question type. The Why-question focus is also gained by this corpus study where the Why-question focus is mostly expressed by a verb or a verb phrase containing v_c or v_e. Then, we use YQC to determine the Why question type along with v_c or v_e to determine the Why question focus on the testing corpus of the Why-question type determination.

4.2.1.3 *Image Pre-processing*

All 400 plant disease images, especially the rice leaf symptom, are collected from the image portions of the Embedded-Image questions (from section 4.2.1.1). Image enhancement is constructed from low pass and high pass filter for adjusting intensities of the image in order to highlight areas considered. After this pre-processing step, the image is ready for segmentation. The segmentation process is to differentiate between background and target object (which is the region showing the current symptoms of the disease) to eliminate the back ground away from the leaf area having the disease symptom. Then, a ROI is identified from this leaf area by using region growing algorithm (Stanciu, 2012). Therefore, the ROI of each image contains several major features of the target object as color, texture, lesion shape, etc.

4.2.1.4 *BOW Determination*

Our research applies BOW as shown in Fig. 8 to represent ROI being the image salience, especially the disease symptom, where each visual word represents each symptom feature. Thus, the BOW determination step is to detect the salient features which are the symptom features as lesion color (ROI object color), lesion shape (ROI object shape), leaf texture (ROI area color), and leaf color.

Fig. 8 Show all patches of visual words in the BOW from the image of Fig. 1 (b)

The ROI object shape is determined by using shape contexts, i.e. an eye shape, a scratch shape, and a spot shape, where the reference point captures the distribution of the remaining points relative to it (Stanciu,2012). Then, the corresponding points on two similar shapes have

similar shape contexts. After the missed shape contexts have been filtered out, the color detection is determined. There are two areas of color detection, a ROI area for texture detection and a ROI object. To detect the color and the texture, we apply the following two classification levels based on 400 sample images from the image portions of the Embedded-Image questions (from section 4.2.1.3) where these sample images are supervised data and consist of 250 sample images for learning and 150 sample images for testing with ten folders cross validation.

First Classification Level

The objective of this level is to filter out the normal properties of the color and the texture of the ROI by learning of the binary classifier as the logistic regression model (Ng and Jordan, 2002). The logistic regression model as shown in Eq.(6) is applied to classify both color and texture properties with two classes of Normal and Abnormal where ROI pixels are based on the HSI color model, for hue (H), saturation (S), and intensity (I). The twelve features (Feature Vector) as Min, Max, Mean, and Entropy of H, S, and I are used in the binary classification. The Entropy feature as shown in Eq. (7) (Shannon and Weaver, 1949) is applied in this research for determining local spatial variations of color intensity which express the textural property of an image as the roughness. And, the roughness texture property in our research expresses the leaf shape symptom of the curl/twist/shrink occurrence.

$$\text{LogisticRegression}: \quad \gamma = \frac{\exp^{(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n)}}{1+\exp^{(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n)}} \tag{6}$$

where γ = the probability of a case which is in a particular category: AbnormalClass

exp = the base of natural logarithms

FeatureVector $X = x_1, x_2, .., x_n$

α = the constant of the equation

β = the coefficient of the predictor variables

$$\text{Entropy of P}: \quad H(P) = \sum_{i=1}^{n} p_i * \log(1/p_i) \tag{7}$$

where P is a set of a probability distribution of information as all features of ROI $P = \{p_1, p_2, .., p_n\}$

Second Classification Level

The results of the Abnormal class samples from the first classification level are used in this second classification by learning of a multi-class-classifier as Multi-Layer Perceptrons (MLPs)(Haykin, 1999) for detecting the color symptom and the texture symptom of the ROI

object (which emphasizes on the disease lesion) and the ROI background (which is the leaf containing the disease lesion) respectively. There are twelve input features used in the MLPs classifier as Min, Max, Mean, and Entropy of H, S, and I. The MLPs classifier has eight classes (Brown, Black, Grey, Beige, Yellow, Pale, Orange, Red-Orange) of irregular color occurrences on the ROI object or the ROI background.

Multi-Layer Perceptrons (MLPs)

According to (Haykin,1999), 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. (8).

$$y_i = f(\sum x_i w_i) \tag{8}$$

where w_i represents the weight, x_i is the input feature of the ROI, $f(\cdot)$ is the activation function, and y_i is the output of the i^{th} node. A sigmoid function is often used as the activation function. MLP consists of successive layers, each of which includes a different number of processing nodes.

$$X = \sum_{i=1}^{n} x_i w_i - \theta \tag{9}$$

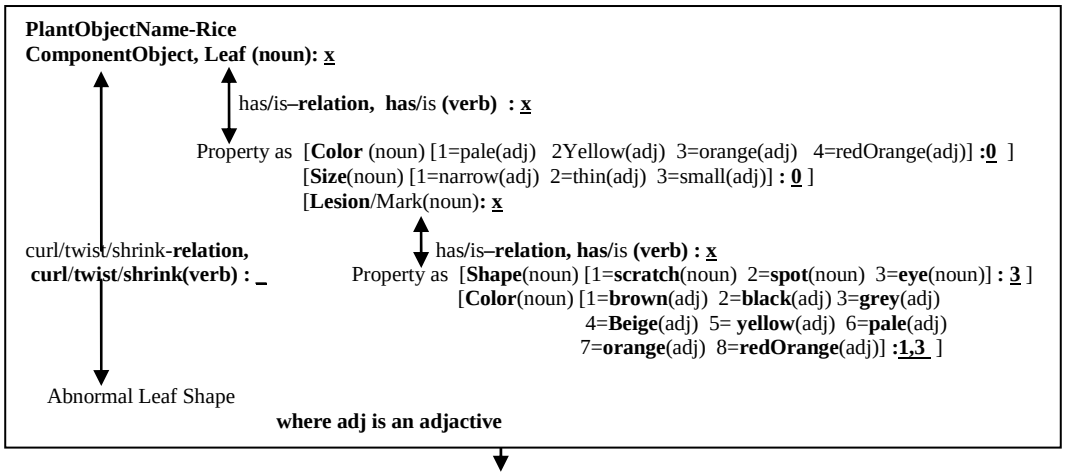
The nodes in the first layer receive inputs from the outside world and are fully connected to nodes in the hidden layer where each node in the hidden layer is connected to the output layer nodes to produce the output pattern or the output class of the MLP. Furthermore, the net weighted input can be solved by Eq.(10) which contains the activation function.

$$y_j(p) = \sum_{i=1}^{n} x_i(p) w_{ij}(p) - \theta_j \tag{10}$$

where n is the number of neuron inputs, and θ_j is the threshold value of neuron at the j^{th} node in the hidden layer.

4.2.1.5 *EIWhy-Question-Content Determination*

This step is to generate the conceptual predicate content of the image portion of the EIWhy question by using a symptom-concept-frame structure shown in Fig. 2 to interpret the BOW after the Why question type has been determined from the textual portion of the EIWhy question. Therefore, the BOW from Fig. 8 can be interpreted as the Why question content (the conceptual predicate content which contains a content word set generated by its symptom-concept frame) as shown in Fig 9.



- Why-question content: hasEyeShape_mark(leaf)∧ be_brown_and_grey_color(eyeShaped_mark)
- Set of content words : leaf(noun), has/is(verb), lesion/Mark(noun), has/is(verb), shape(noun), eye(noun), color(noun), brown(adj), grey(adj)

Fig. 9 Examples of EIWhy-Question-Content Determination in the form of the conceptual predication of symptoms.

4.2.1.6. *Question-Focus Determination*

The question-focus is necessary for pointing to what the answer is. The EIWhy question 's focus for the root cause determination is expressed on the question-image portion and expressed by the symptom conceptual predication which is the relation expression (V_e) in the symptom-concept-frame structure. For example: the EIWhy-Question content in Fig. 9 based on the symptom-concept-frame structure (see section 4.2.1.5), has the question focus on the symptom conceptual predication from the following contents:
has_Lesion relation, *has_eyeShape* relation, and *is_brown-greyColor* relation.

4.2.2 EIWhy answering part

The objective of this part is to determine the corresponding Why answers including the calculated generality values of effects from the answer source. There are two steps involved in this part, the causality knowledge generalization step and the answer determination step.

4.2.2.1 *Causality Knowledge Generalization*

The examples of the extracted causality knowledge (from Section 4.1) as the answer source are shown in Fig. 10 where the extracted causality knowledge contains the uncertainty of nuance expression for lesions on the plant leaves and the incomplete knowledge. This nuance expression problem can be solved by using the nuance concept from the linguistic head noun pattern to align the nuance expression. For example: (a) "Leaves have greenish light grey

spots." (b) "Leaves have dark grey spots." Then, both nuance expressions, (a) and (b), can be solved as "Leaves have grey spots." The incomplete knowledge can be solved by applying Monte Carlo the Monte Carlo simulation technique for imputing the effect events to solve the incompleteness problem as shown in the following Imputation step. All of the extracted causality knowledge with the verb feature adjustment is generalized by using the statistical based approach and the appropriate fuzzy concept of the triangular membership function (Jang et al., 1997). The plant disease knowledge from the electronic Thai encyclopedia has been used for determining the super concept set of the disease cause such as a pathogen type,{virus, bacteria, fungus}, or an insect type, {plant louse}. The generalization process consists of three steps of Cause-Effect Vector Space Generator, Imputation, and Generalization for Generality Value Determination

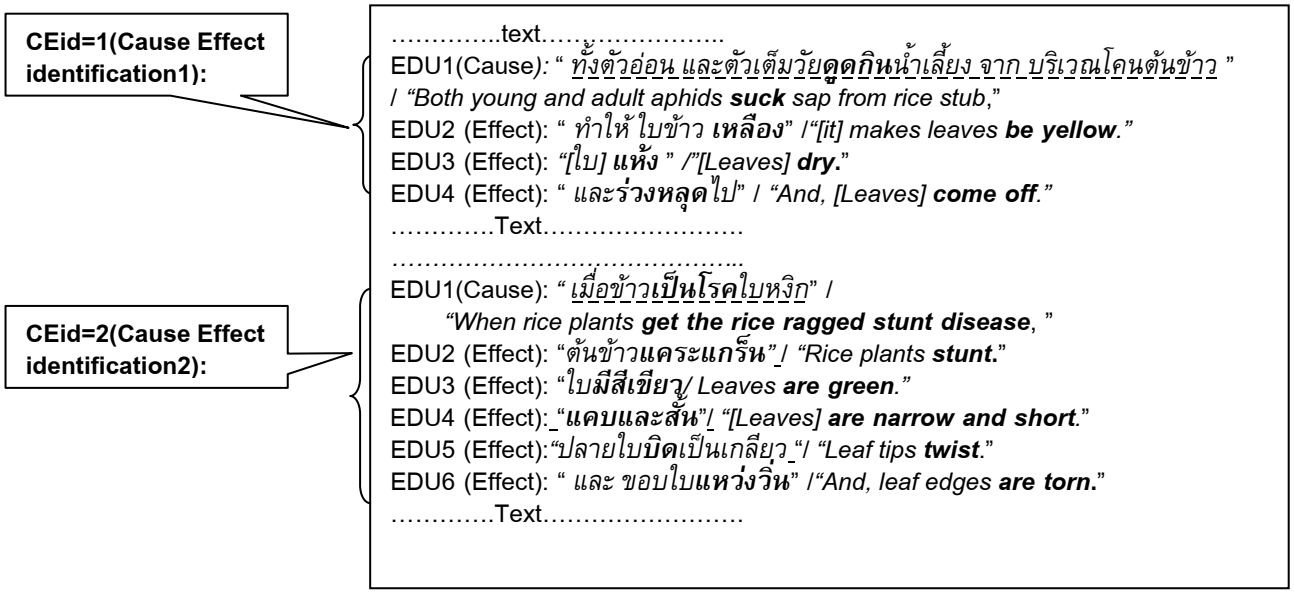


Fig. 10 Show examples of the extracted causality knowledge on the documents

4.2.2.1.1 *Cause-Effect Vector Space Generator*

According to the answer source, several extracted inter-causal EDUs (or causality knowledge) from several documents are collected to formulate as the cause-effect vector space. From Fig. 10, each CEid is the ''inter-causal EDUs' represented in the form of predication by using the cause and effect patterns as shown in the following:

Cause pattern: Vc (NP1,NP2) Effect pattern: Ve(NP3)

where NP1, and NP3 are noun phrase concepts with their function as 'Agent', NP2 is a noun phrase with the function of 'Patient', Vc is a causative verb concept, and Ve is an effect verb concept. All concepts of verbs and nouns are referred to WordNet and Thai Encyclopedia

(http://kanchanapisek. or.th/kp6/New/). For example: CEid=1 from Fig. 10 can be represented as follow:

CEid=1: Consume(plant-louse,solution) → Be_yellow(leaf)
^Dry(leaf) ^ Come_off(leaf)

After each causality predication has been constructed, all binary cause-effect vectors are generated to form the cause-effect vector space as shown in Table 5 by coding the symptom/effect predication to Si (where i=1,2,..,n).

4.2.2.1.2 *Imputation*

According to Table 5, there is only the symptom matrix (Sx) being concerned with the imputation of the incomplete knowledge occurring on Sx (where Sx is the m × n matrix with m row vectors of extracted causalities and n column vectors of symptoms).

Table 5. Cause-Effect Vector Space
(where S_1= Wilt(leaf), S_2=Change_shape(leaf), S_3= Have_eyeShape_Mark(leaf), S_4=Stunt(plant), S_5= Come_off(leaf), S_6= Dry(leaf), S_7= Be_yellow(leaf)),…

Cause	S_1	S_2	S_3	S_4	S_5	S_6	5_7	…	S_n
consume(plant-louse,solution)	1	1		1	1			…	…
consume(plant-louse,solution)				1	1			…	…
consume(plant-louse,solution)	1	1		1				…	…
consume(plant-louse,solution)		1		1	1		1	…	…
consume(plant-louse,solution)		1	1	1				…	…
destroy(plant-louse,plant)							1	…	…
destroy(plant-louse,plant)		1						…	…
destroy(plant-louse,plant)		1		1	1	1		…	…
……	…	…	…	…	…	…	…	…	…

Table 6. Results of imputation of undefined symptoms by using Monte Carlo simulation technique

Cause	S_1	S_2	S_3	S_4	S_5	S_6	S_7	…	S_n
consume(plant-louse,solution)	1	1	0	1	1	1	0	…	…
consume(plant-louse,solution)	0	1	0	1	1	0	1	…	…
consume(plant-louse,solution)	1	1	0	1	1	0	0	…	…
consume(plant-louse,solution)	1	1	0	1	1	0	1	…	…
consume(plant-louse,solution)	1	1	1	1	0	1	0	…	…
destroy(plant-louse,plant)	0	0	1	0	0	0	1	…	…
destroy(plant-louse,plant)	0	0	1	0	0	0	0	…	…
destroy(plant-louse,plant)	0	1	0	1	1	1	1	…	…
……	…	…	…	…	…	….	…	…	…

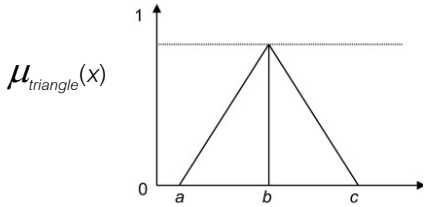
We apply the Monte Carlo technique to solve the incomplete knowledge occurrences on some consequences of causality. The Monte Carlo technique is a method that uses random numbers and probability statistics to perform simulations (Pengelly, 2007). Therefore, the random number (r) and the probability (ρ_i) of the Si occurrence (where i=1,2,..,n; and n is the number of symptoms) from 103 observed documents are applied to simulate the imputation of

undefined symptoms of Si, as shown in Table 6, by using the following imputation algorithm with m samples of the extracted causality knowledge.

```
Function   Imputation_Algorithm
{ For  i=1 to  n
   {For  j=1 to m
     { If  Sxji = “  ”
        { Generate Random Numner r
          If  r> ρi of Si, then Sxji = 1
          Else      Sxji =0
        } } }
}
```

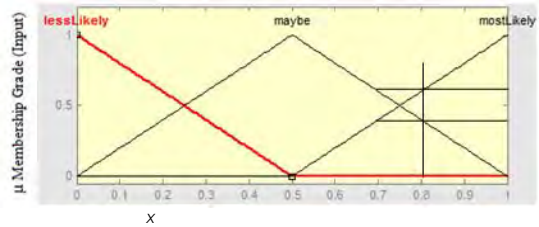
4.2.2.1.3 *Generalization for Generality Value Determination*

The fuzzy technique as explained by (Zadeh ,1965), uses the input membership values as weighting factors to determine their influence on the fuzzy output sets of the final output conclusion or defuzzification into a crisp output driving the system. The simplest membership functions as the triangular membership function, $\mu_{triangle}$, specified by three parameters {a,b,c}(as shown in Eq. (11)-(14)) is applied in our research. $\mu_{triangle}$ determines the degree of membership in the [0,1] interval from a single input (x) (Jang et al., 1997).



$$\mu_{triangle}(x;a,b,c)=\begin{cases}0, x\le a & (11)\\ \dfrac{x-a}{b-a}, a\le x\le b & (12)\\ \dfrac{c-x}{c-b}, b\le x\le c & (13)\\ 0, c\le x & (14)\end{cases}$$

According to the generalization of symptom occurrences from Table 6, the input (x) is an average weight of the Si occurrence based on the causing-agent type of the causing-agent type set {‘plant louse’, ‘fungus’, ‘virus’, ‘bacteria’}. There are three membership grades (lessLikely, maybe, and mostLikely) for $\mu_{triangle}$ to describe the specified input (x) (Fig.11).



μ maybe = (1-0.8)(1-0.5)=0.4

μ mostLikely = (0.8-0.5)(1-0.5)=0.6

Fig. 11 Triangular membership function of S_i with three membership grades where x=0.8

Then, the defuzzification can be solved by the Weight Average Formula (WAF) method (Jang et al., 1997) as shown in Eq. (12) which is the average weight of generality (called the generality value).

$$Weight\ Average\ Formula\ (WAF) = \frac{\sum w_i x_i}{\sum w_i} \quad (12)$$

$$where\ \ w_i = membership\ grade\ \ from\ \ input;$$

$$x_i = weight\ average\ \ or\ \ weight\ of\ \ generality\ \ \ from\ output$$

Table 7 Average Weight of Generality

Cause	Effect/Symptom	Average Weight of Generality
Destroy(plant-louse, plantORplant_component)	stunt(plant)	0.75
Destroy(plant-louse, plant ORplant_component)	beAbnormalShape (leaf)	0.8
Destroy(fungus, plantORplant_component)	Have_eyeShape_mark (leaf)	0.2
Destroy(virus, plantORplant_component)	stunt(plant)	0.78
…	……..	……

Table 7 shows the average weight of generality of each effect or plant symptom from 103 random plant disease documents caused by plant louses, fungus, bacteria, or virus

4.2.2.2 *Answer Determination*

After both the Why-question type determination and the Why-question content determination, the correct EIWhy questions are used for the answer determination from the knowledge source which contains cause-effect vectors of plant diseases. The answer is solved by determining the similarity score (Biggins et al., 2012) in Eq. (8) between a set of content words existing in the Why-question content and each EDU element of the cause-effect vectors after eliminating stop words.

$$Similarity\ \_\ Score = \frac{|S1 \cap S2|}{\sqrt{|S1| \times |S2|}} \quad (8)$$

where

S1 is all word concepts from the set of content words existing in the Why-question content.

S2 is all word concepts from a set of words from $EDU_{effect\text{-}i}$ after eliminating stop words (through stemming words for some languages) where $EDU_{effect\text{-}i}$ exists in the cause-effect vector $\langle EDU_{cause}, EDU_{effect\text{-}1}, EDU_{effect\text{-}2}, .., EDU_{effect\text{-}m} \rangle$.

The all word concepts of S1 and S2 are based on WordNet and Thai Encyclopedia after using the Thai-to-English dictionary. For example (Fig.1(b)):

S1 : Set of content words

{ leaf(*noun*), has/is(*verb*), lesion/Mark(*noun*), has/is(*verb*), shape(*noun*),

eye(*noun*), color(*noun*), brown(*adj*), grey(*adj*)}

= { leaf, have/be, lesion/mark, shape, eye, color, brown, grey}

Knowledge Source:

Cause-Effect Vector ID=1 DiseaseName: Rice Blast disease

EDU_{cause} : "*เชื้อราไพรีคิวลาเรีย/**Pyricularia fungus** ทำลาย/**destroy** ต้นข้าว/**rice plant** *"

(The Pyricularia fungus destroy the rice plant.)

$EDU_{effect1}$: "*ระยะ/**period** กล้า/**seedling** ใบ/**leaf** มี/**have** แผล/**lesion** รูป/**shape** ตา/**eye** สี/**color** น้ำตาล/**brown***"

(Seedling Period: Leaves have the brown eye shape lesions.)

$EDU_{effect2}$: "*แผล/**lesion** ขยายลุกลาม/**spread over** ทั่ว/**whole** ใบ/**leaf** *"

(The lesions spread over the whole leaf.)

$EDU_{effect3}$: ……..

S2: Cause-Effect Vector ID=1

$EDU_{effect1}$: {seedling, period, leaf, have, lesion, shape, eye, brown, color} ⟶

Similarity_Score =0.8 (where Have_eyeShape_mark (leaf) has the generality value=0.2)

$EDU_{effect2}$: {lesion, spread, whole, leaf} ⟶ Similarity_Score =0.4 (where Have_eyeShape_mark (leaf) has the generality value=0.2)

……..

The candidate answers can be selected from all Cause-Effect Vector IDs which have S2 of $EDU_{effect-i}$ being similar to S1 of the question-image portion with Similarity_Score >0.5. Then, the candidate answers can be ranking according to Similarity_Score of the selected Cause-Effect Vector IDs. We select only the top five ranks of Similarity_Score as the possible answers where the first rank is the highest correct answer.

5. Evaluation and Discussion

There are two main evaluation parts of our system, the first part is the boundary determination of causality knowledge extraction. The second part is the EIWhy-QA system.

5.1 Knowledge Extraction with Boundary Determination

The corpora used to evaluate the proposed model of the effect-boundary determination by using three different machine learning techniques, SVM, ME, and NB consist of 1,000 EDUs collected from on line of the plant disease technical papers and the news domain, especially global warming news. Each of these corpora has different characteristics of the effect verb frequency and the diversity of verb occurrence. The results of the effect-boundary determination by SVM, ME, and NB (Table 8) are based on two experts and one linguist with max-win voting. In addition to the causality extraction by using verb-pair rules (Pechsiri and Kawtrakul, 2007), the evaluation of the causality is expressed in terms of the precision (0.85 by average) and the recall (0.72 by average).

Table 8. Accuracies of boundary determination of the inter-causal EDU extraction from different methodologies.

Document type (≈250EDUs each)	No. of different effect verbs	%Correctness of effect boundary determination		
		SVM	ME	NB
Plant Disease by aphids	40	94	91	86
Plant Disease by fungi	63	87	89	82
Plant Disease by virus	35	93	93	85
Global Warming news	48	90	95	84

Table 8 shows that the wide variety of the effect verbs affects the % correctness of each methodology. ME has the highest correctness 92% by average whereas NB has the lowest average, 84.2%, because there are some dependencies among effect verbs or effect events.

5.2. EIWhy-QA system

The testing-question corpus for the evaluation of the EIWhy QA system is collected by interviewing farmers and by downloading from several QA sites and web blogs and contains the 300-textual questions consisted of 150 Embedded-Image questions and 150 regular Why questions. Moreover, the 300-textual questions are based on the plant disease domain, especially the rice leaf diseases, and contain several type of question as mention in section

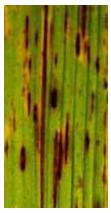
4.2.1.1. There are three necessary evaluations of the EIWhy-QA system comparing to the regular Why-QA system (where the image processing is not required for the regular Why-QA system): the Why-question type determination, the question content determination, and the corresponding Why answer determination from the answer source

5.2.1. Why-Question-Type Determination

Table 9 Evaluation of the Why-question-type determination from the textual questions without images embedded (the regular Why questions) and the textual portions of the Embedded-Image questions

Textual Questions	*Why-Question Type Determination*		
	#Correct Why Questions	Precision	Recall
150 regular Why questions contain 60 Why questions	54	0.95	0.90
150 Embedded-Image questions contain 60 Why questions	51	0.93	0.85

The testing-question corpus that is used for the Why-question type determination consists of two-150 textual questions, one with images embedded and another one without images embedded, having 60 Why questions for each 150 textual questions. According to determine the Why question type by using YQC, we can evaluate the Why-question type determination by calculating Precision and Recall based on experts with max win voting, as shown in Table 9.
The experimental results from Table 9 illustrate that the recall and the precision of the Why-question determination from the Embedded-Image questions are lower than the regular Why question. The reason of the lower recall is some vague questions occurring on the textual portions of the EIWhy questions, as shown in Fig. 12 (a) , where the complete question is "เกิดอะไรขึ้นกับใบและทำให้เกิดอาการเป็นแผล/What happens to the leaf and causes the lesion symptom?"). And, the reason of the lower precision is that some EIWhy questions are sarcastic questions since they are not required the Why answer, as shown in Fig. 12 (b).



(a) "*เกิดอะไรขึ้น*"
("*What happens?*")

(b) "*ทำไม[คุณ]ไม่แก้ไขปัญหา[โรคระบาดในข้าว] งานเยอะใช่ไหม*"
("*Why don't [you] solve this [rice epidemic] problem? Are there lots of works?*")
where [..] means ellipsis.

Fig.12 Show the examples of errors from the textual portions of the EIWhy Questions

5.2.2. Question Content Determination

According to the question contents determination by using the BOW of the ROI ‘s image portion of the Embedded-Image questions, the authors evaluate the question content determination by calculating the correctness of the visual word determination from the testing data of 150 sample images (collected from the image portions of the Embedded-Image questions based on supervised training data). The question content determination includes the lesion color (the ROI object color), lesion shape (the ROI object shape), leaf texture (the ROI area color), and leaf color. Then, the visual word determination consists of the ROI-object-shape determination, the ROI-object-color determination, and the ROI-area-color determination which is applied for the ROI texture determination. The correctness of the ROI object shape determination is 88 % depending on the perspective. The correctness of the ROI color determination is 94% of the binary classification whilst the precision and the recall of the multiclass classification is 0.766 and 0.768 respectively, with ten folders cross validation of 132 correct sample images after filtering out the incorrect ROI object shape from the ROI object shape determination (see Table 10).

Table 10 Evaluation of the ROI color determination of two classification levels (based on ten folders cross validation).

Binary Classification

Class	True Positive	False Positive	False Negative	True Negative	% correctness
Abnormal	112	7	-	-	94
Normal	-	-	2	11	85

Multiclass Classification

Classifier	True Positive Rate	False Positive Rate	Precision	Recall	F-Measure
MLP	0.768	0.087	0.766	0.768	0.765

According to the multi-class classification, there are 110 questions having the correctness of determining the visual words without the question type consideration. The errors of the multi-class classification are caused by the incorrect patch generation which results in incorrectly determining the visual words. However, the total correctness of the EIWhy question determination (after determining both the visual words (the question contents) and the Why-question type) are 47 questions from the testing corpus containing 60 EIWhy questions.

Table 11 Evaluation of the corresponding answer determination for the EIWhy-QA system and the regular Why-QA system

Why Questions	*# Correct Answers*	*% Correctness of Why Answer Determination At the first rank*
54 correct Why questions of the regular Why QA system	43	43/60=72
47 correct EIWhy questions of the EIWhy-QA system	47	47/60=78

5.2.3. Corresponding Why Answer Determination

According to the root cause determination, the evaluation of the Why answer determination for the EIWhy-QA system and the regular Why-QA system is based on three experts with max win voting. The 47 correct questions of the EIWhy questions (from section 5.2.2) are used to evaluate the Why answer determination of the EIWhy-QA system. The 54 correct questions of the Why question type determination from the regular Why questions (from section 5.2.1) are used to evaluate the Why answer determination of the regular Why-QA system. The corresponding Why answers of the EIWhy-QA system and the regular Why-QA system can be solved by determining the answer with the highest rank of the similarity score between a set of content words existing in the Why-question content and each EDU element of the cause-effect vectors after eliminating stop words (as shown in Table 11). According to Table 11, the regular Why-QA system has the lower % correctness of the Why answer determination because of the vague questions in the regular Why-QA system. For example: “ทำไมใบมีแผล /Why do leaves have lesions?” (The example is vague because it does not notify the lesion characteristics, i.e. shape, color, and etc. The lesion characteristics can determine the cause of disease.)

6. Conclusion

This research presents Causality Knowledge Extraction and Generalization for supporting the EIWhy-QA system. Thus, our research includes 3 major phases: Causality Knowledge Extraction as the source of answers, Causality Knowledge Generalization, and the EIWhy-QA system. The EIWhy-QA system includes the image processing technique to enhance the ability in diagnosing problems, especially the plant diseases. The EIWhy-QA approach diagnosis requires the knowledge extraction from technical documents based approach rather than the information retrieval based approach commonly practiced in previous studies. This is because the answer source for supporting diagnostics must be the knowledge proven by experiments or specialists unlike the information retrieval system. The evaluation of the EIWhy-QA system suggests that the image processing in Why questions improves the %correctness substantially from 72% to 78%.

The EIWhy-QA system requires evaluations for two major parts of problems: the EIWhy question part and the EIWhy answering part. Although, the result of the correct-Why-answer determination of the EIWhy-QA system is higher than the one of the regular Why-QA system, the result of the correct-Why-question-type determination of the EIWhy-QA system is lower than the one of the regular Why-QA system. The main reasons for the lower correctness of the Why-question type determination of the EIWhy-QA system are the vague question, the sarcastic question, the image perspective errors, and the image-noise-like symptoms. Moreover, issues regarding vague questions and sarcastic questions can affect the correctness to both the EIWhy-QA system and the regular Why-QA system. However, potentials to improve the EIWhy-QA system's performance exist where vague questions, sarcastic questions, and the BOW determination must be further studied, especially in pragmatics.

Therefore, the successful EIWhy QA system including the generality value determination of each symptom occurrence on the texts is valuable to people in supporting preliminary analysis and performing root-cause analysis for inexperienced persons in diagnostics. Once integrated with mobile phones, such capacities allows the EIWhy-QA system to have a profound effect on several business areas as a tool to assist inexperience participants/people and amateur diagnosticians to diagnose problems.

LITERATURE CITED

Agichtein E., Cucerzan S., & Brill E. Analysis of Factoid Questions for Effective Relation Extraction. In ACM SIGIR International Conference on Research and Development in Information Retrieval, SIGIR'05, Salvador, Brazil, 2005

Angryk R. & Petry F. Consistent Fuzzy Concept Hierarchies for Attribute Generalization. In Proc. IASTED Int.Conf on Information and Knowledge Sharing, pp. 158-163, Scottsdale AZ, 2003

Berger A. L., Della Pietra S. A., Della Pietra V. J. A Maximum Entropy approach to natural language processing. In Computer Linguist., 1996, 22(1):39–71.

Biggins S., Mohammed S., and Oakley S. Two Approaches to Semantic Text Similarity, In Proc. Of the First Joint Conference on Lexical and Computational Semantics, (Montre´al, Canada, 2012), pp 655–661.

Burhans D.T., & Shapiro S.C. Abduction and Question Answering. In Proc. of the IJCAI Workshop on Abductive Reasoning, IJCAI-01, Seattle, Washington, 2001

Carlson L., Marcu D., and Okurowski M. E. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Current Directions in Discourse and Dialogue, 2003, pp.85-112.

Chanlekha H., & Kawtrakul A. Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. In Proc for IJCNLP' 04, Hainan Island, China, 2004

Chareonsuk J., & Sukvakree T., & Kawtrakul A. Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information. In Proc for NCSEC, Thailand, 2005

Cheng J. The Fundamental Role of Entailment in Knowledge Representation and Reasoning. Journal of Computing and Information, 1996, 2(1).

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

Csiszar I. Maxent, mathematics, and information theory. In Proc. 15th Int. Workshop Maximum Entropy and Bayesian Methods, Santa Fe, USA, Jul. 31-Aug 4, 1996, pp. 35-50.

Druzdzel M.J. Probabilistic Reasoning in Decision Support Systems: From Computation to Common Sense [PhD Thesis], Department of Engineering and Public Policy, Carnegie Mellon University, USA, 1993

Ei-Helly M., Rafea A., Ei-Gamal S., & Ei-Whab R. Integrating Diagnostic Expert System with Image Processing via Loosely Coupled Technique. Central Laboratory for Agricultural Expert System (CLAES), 2004

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, pp.49-56.

Girju R. and Moldovan D. Mining answers for causation questions. In AAAI symposium on mining answers from texts and knowledge bases, 2002.

Girju R. Automatic detection of causal relations for question answering. In The 41st annual meeting of the assoc. for computational linguistics, workshop on multilingual summarization and question answering-Machine learning and beyond, Japan, 2003

Haykin, S. Neural networks: a comprehensive foundation. (2nded,). Upper Saddle Rever, New Jersey, USA: Prentice Hall, 1999

Hovy E., Hermjakob U., & Ravichandran D. A Question/Answer Typology with Surface Text Patterns. In Proc. of the Human Language Technology conference (HLT), San Diego, California, 2002

Inui T., Inui K., and Matsumoto Y. Acquiring causal knowledge from text using the connective markers. Journal of the information processing society of Japan, 2004, 45(3):919-933.

Jang J-S. R., Sun C.T., & Mizutani E. Neuro–Fuzzy AND Soft Computing: A Computational Approach to Learning and Machine Intelligence. USA: Prentice Hall, 1997

Kate R.J. & Mooney R.J. Probabilistic Abduction using Markov Logic Networks. In Proceedings of the IJCAI-09 Workshop on Plan, Activity, and Intent Recognition, PAIR-09, Pasadena, California, 2009

Kaundal R., Kapoor A., & Raghava G. Machine Learning Techniques in Disease Forecasting: A Case Study on Rice Blast Prediction. BMC Bioinformatics, 2006

Khoo C.S.G. Automatic Identification of Causal Relations in Text and Their Use for Improving Precision in Information Retrieval [Ph.D. Dissertation], School of Information Studies, Syracuse University, 1995.

Lehmann J., Maes S., Dirkx E. Causal Models for Parallel Performance Analysis. In Fourth PA3CT-Symposium, Edegem, Belgium, 2004.

Marcu D., Echihabi A. An Unsupervised Approach to Recognizing Discourse Relations. In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics Conference, Philadelphia, 2002, pp.368–375.

Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince. Centering in Naturally Occuring Discource: An Overview in Centering Theory of Discourse. Calendron Press, Oxford, 1998, pp. 1-28.

Meunkaewjinda A., Kumsawat P., Attakitmongcol K. & Srikaew A. Grape Leaf disease Detection from Color Imagery System Using Hybrid Intelligent System. In Proc. Of IEEE ECTI-CON, 2008, pp 513-516.

Miler G. A., Beckwith R., Fellbuan C., Gross D., and Miller K.. Introduction to Word Net. An Online Lexical Database, 1993.

Miller R.A. Medical Diagnostic Decision Support Systems – Past, Present, and Future: A threaded Bibliography and Brief Commentary. Journal of the American Medical Informatics Association, 1994, 1(1): 8 – 27

Mitchell T M. Machine Learning. The McGraw-Hill Companies Inc. and MIT Press, Singapore, 1997.

Mitchell T.M., Keller R.M., & Kedar-Cabelli S.T. Explanation-Based Generalization: A Unifying View, Machine Learning, 1:47-80, Boston: Kluwer Academic Publishers. 1986

Ng A.Y., & Jordan M.I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Proc of Neural Information Processing Systems, 2002.

Patil J., & Kuma R. Advances in Image Processing for Detection of Plant Diseases. Journal of Advance Bioinformatics Application and Research, 2011, Vol.2, Issue 2, pp 135-141

Pechsiri C., & Kawtrakul A. Mining causality for explanation knowledge from text. Journal of Computer Science and Technology, 2007, 22(6): 877-889.

Pechsiri C., & Piriyakul R. Developing the UCKG-Why-QA System. In Proc for 7th International Conference on Computing and Convergence Technology, Korea, 2012

Pechsiri C., & Piriyakul R. Explanation Knowledge Graph Construction through Causality Extraction from Texts. Journal of Computer Science and Technology, 2010, 25(5): 1055-1070

Pechsiri C., Sroison P., and Janviriyasopa U. Know-Why Extraction from Textual Data. In Proc. KRAQ, 2008.

Pengelly J. Monte Carlo Methods, Student Tutorial for course COSC453 in 2007, Department of Computer Science, University of Otago 2007, http://www.cs.otago.ac.nz/cosc453/student_t utorials/monte_carlo.pdf

Quarteroni S., & Saint-Dizier P. Addressing How-to Questions using a Spoken Dialogue System: a Viable Approach? In Proc. of the 2009 Workshop on Knowledge and Reasoning for Answering Questions, ACL-IJCNLP 09, Suntec, Singapore, 2009

Shannon, C. E. & Weaver, W. The Mathematical Theory of Communication. Urbana, Illinois, USA: University of Illinois Press, 1949

Sivic J., Russell, B. C. Efros, A. A., Zisserman, A., & Freeman, W. T. Discovering Objects and Their Location in Images. Robotics Institute. Paper 286, 2005. Retrieved from http://repository.cmu.edu/robotics/286 on 2013-03-04.

Stanciu S.G. Digital Image Processing. Rijeka, Croatia: InTech, 2012

Sudprasert S., & Kawtrakul A. Thai Word Segmentation based on Global and Local Unsupervised Learning. In Proc for NCSEC'03, Chonburi, Thailand, 2003

Vazquez-Reyes S. and Black W.J. Evaluating Causal Questions for Question Answering. In 9th Mexican International Conference on Computer Science, ENC'2008, pp.132-142. 2008.

Verberne S. Developing an approach for why-question answering. In Proc. from Conference of the European Chapter of the Assoc. for Computational Linguistics, 2006

Verberne S., Boves L., Coppen P-A., & Oostdijk N. Discourse-based answering of why-questions. Traitement Automatique des Langues, 2007, 47(2).

Verberne S., Boves L., Oostdijk N., & Coppen P-A. Using Syntactic Information for Improving Why-Question Answering. In Proc. of the 22nd International Conference on Computational Linguistics, Coling '08, Manchester, UK, 2008

Weizheng S., Yachun W., Zhanliang C., & Hongda W. Grading Method of Leaf Spot Disease Based on Image Processing. In Proc. Of International Conference on Computer Science and Software Engineering, Cambridge, 2008

Woller J. The Basics of Monte Carlo Simulations, University of Nebraska-Lincoln, 1996. Retrieved from http://www.unl.edu/zeng /joy/mclab/mcintro.html on 2009-06-16

Woodford B., Kasabov N., & Wearing C. Fruit Image Analysis Using Wavelets. In Proc. Of ICONIP/ANZIIS/ANNES, 1999.

Yamada K. Fuzzy Abductive Reasoning for Diagnostic Problems. In IFSA World Congress 95, Sao Paulo,1995, pp.649-652.

Yeh T., Lee J.J. & Darrell T. Photo-based Question Answering. In MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada, 2008

Ying G., Miao L., Yuan Y., & Zelin H. A Study on the MEthoid of Image Pre-Processing for Recognition of Crop Diseases. In Proc of IEEE International Conference on Advances Computer Control, 2008

Zadeh L.A. Fuzzy Sets. Information and Control, 1965,

APPENDIX

Introducing Why - How Question Answering System and Integrated Causality Graph through Online Community

Chaveevan Pechsiri
(Department of Information Technology, Dhurakij Pundit University, Bangkok, Thailand
itdpu@hotmail.com)

Rapeepun Piriyakul
(Department of Computer Science, Ramkumheang University,Bangkok, Thailand
rapepunnight@yahoo.com)

Abstract: The research aims to develop a Question Answering system, especially 'Why' and 'How' questions, on online community web-boards for supporting the community in diagnosis and solving problems, e.g. plant disease, health-care problems, etc. Unlike factoid question, the Why and How questions of this research are based on preface questions which give some background and scene-setting of the questions by expressing in several Elementary Discourse Units (EDUs), where each EDU is a simple sentence or a clause. The research consists of several problems as 'How to identify the question types', 'How to solve the complicated question' and 'How to determine the answer'. Therefore, the research proposes using different machine learning techniques for the question type identification. The research also integrates the procedural knowledge, extracted from text by machine learning approach, into the previous causality graphs for solving the complicated question and answering How/Why questions based on similarity-score collection. The experiment shows that the system can achieve correct answers to 93.3% of the questions.

Keywords: Why-Q, How-Q, prefaced question, complicated question, Causality Graph
Categories: I.2.7, I.2.1, M.7, J.7, L2.1

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. Then, they wait for several minutes to several days to receive the recommended answers posted by the experts for solving their problems on the web page. However, it is time consuming for people to receive the answers. For example, 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 the problems with a why question (Why-Q) type, asking for a reason, and/or a how question (How-Q) type, asking for problem solving approach, on the community web-boards. However, there are some responses to some questions depending on a question domain, a chat room type of a certain web-board, a web-board domain, and etc. It is approximately 68% on average for plant disease questions to receive responses within a week on the Thai community web-boards (unpublished data). During the waiting time, an automatic Why-How Question-Answer (QA) system is introduced for providing a preliminary diagnosis includ-

ing solving methods before or during an epidemic. Therefore, this research aims to develop an automatic QA system of Why-Q and How-Q with the integrated causality graph [Pechsiri and Piriyakul, 10] for the preliminary diagnosis problems, e.g. plant diseases, including the recommendation of solving these problems on the online community. According to [Aouladomar , 06], there are several types of How questions e.g. the causality How, the instrumental How, the instructional How, and etc. whereas How-Q of our research is equivalent to the causality How question which is used to know the causes or the circumstances of a certain event. Most of the posted plant disease questions on the web-board resemble prefaced questions rather than factoid questions. A prefaced question is a question that provides background and scene-setting for the questions (http://www.dpc.nsw.gov.au/merit/module_7). The prefaced questions in our research are expressed in the form of Elementary Discourse Unit(s) (where each EDU is defined as a simple sentence or a clause, [Carlson et. al., 03]) with the following question patterns (called „Qpathern').

Qpattern-1: EDU_{ct1} EDU_{ct2} …EDU_{ctn} EDU_q

Qpattern-2: EDU_{ct1} EDU_{ct2} …EDU_{ctn} EDU_q EDU_{ctn+1}

QPattern-3: EDU_q EDU_{ct1} EDU_{ct2} …EDU_{ctn}

where:

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

$EDU_q \rightarrow$ Qword NP1 V_q NP2 | NP1 V_q NP2 Qword | V_q Qword

Qword → „*ทำไม*/*Why*' „*อย่างไร*/*How*' „*อะไร*/*What*' „*แสดงวิธี*/*Show method*'

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

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

$EDU_{cta} \rightarrow$ NP1 VP | conj NP1 VP

VP → V_{ct} NP2

(where $V_{ct} = V_c \cup V_e$; v_{cta} is a content verb element of EDU_{cta}, $v_{cta} \in V_{ct}$; V_c is a causative verb concept set; V_e is an effect verb concept set (see Table1[Pechsiri and Piriyakul, 10] in section 4.1.1); conj is Conjuction)

Moreover, there is no question mark, word delimeter, and sentence delimiter in the Thai language. For example:

Qpattern-1 EDU_{ct1}:"*ระยะแตกกอ*(*Tillering Stage*):*ใบข้าว*(*rice leaves*)/NP1 *หงิกงอ*(*shrink*)/v_{ct1}"
(*ระยะแตกกอ*: *ใบข้าวหงิกงอ*/***Tillering Stage: Rice leaves shrink.***)

EDU_{ct2}: "*ต้น*(*plant*)/NP1 *แคระแกรน*(*stunt*)/v_{ct2}" (*ต้นแคระแกรน*/***Plant stunts.***)

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

Qpattern-2 EDU_{ct1}:"*ระยะแตกกอ*(*Tillering Stage*):*ใบข้าว*(*rice leaves*)/NP1 *หงิกงอ*(*shrink*)/v_{ct1}"
(*ระยะแตกกอ*: *ใบข้าวหงิกงอ*/***Tillering Stage: Rice leaves shrink.***)

EDU_{ct2}: "*ต้น*(*plant*)/NP1 *แคระแกรน*(*stunt*)/v_{ct2}" (*ต้นแคระแกรน*/***Plant stunts.***)

EDU_q: "[*เรา*(*we*)/NP1] *จะทำ*(*should solve*)/v_q *อย่างไร*(*how*)/*qw*"
("[*เรา*] *จะทำอย่างไร*/***How should [we] solve?***)

EDU_{ctn+1}: "*ต้น*(*plant*)/NP1 *จึงจะแข็งแรง*(***will be strong***)/v_{ctn+1}"
(*ต้นจึงจะแข็งแรง*/***Plants will be strong.***)
(where [..] means ellipsis.)

Qpattern-3 EDU_q: "*จะเกิด*(*will happen*)/v_q*อะไรขึ้น*(*what*)/*qw*"
(*จะเกิดอะไรขึ้น*/***Whatwill happen?***)
EDU_{ct1}: "*ถ้า*(if)/conj *เพลี้ยจักจั่น*(leafhoppers)/NP1 *ระบาด*(spread out)/v_{ct1}"
(*ถ้าเพลี้ยจักจั่นระบาด*/***if leafhoppers spread out***,)
EDU_{ct2}:"*ขณะ*(*whilst*)/conj *ข้าว*(*rice*)/NP1 *กำลังออก*(*is giving*)/v_{ct2} *รวง*(*paddies*)/NP2"
(*ขณะข้าวกำลังออกรวง*/***whilst rice plants are giving paddies.***)

However, working on Qpattern of Why-Q and How-Q, must involve in three main problems:1) how to identify Why-Q and How-Q on Qpatterns with their question words being ambiguous, 2) how to solve the complicated question of How-Q where the complicated question is a question that cannot answer immediately, see section 3.2.2, and 3) how to determine the Why answer and the How answer of Why-Q and How-Q respectively. Therefore, different machine learning techniques such as Naïve Bayes (NB), Maximum Entropy (ME) and Multilayer Perceptron (MLP) are proposed to classify a question type, Why-Q, How-Q, and Other-Q, from two adjacent EDUs of EDU_q and EDU_{ctk} (where k = 1 or n or n+1), for diagnosis and solving the problems. We then apply the relatedness value determination (see section 4.2.1.2) and machine learning techniques to extract several kinds of the procedural knowledge, e.g. the disease prevention and treatment, from downloaded documents on several websites, e.g. the Department of Agriculture website (http://www.doa.go.th/). This research integrates the extracted procedural knowledge into our previous causality graph [Pechsiri and Piriyakul, 10] (Figure 1) (http://www.web3point2.com/rice/index App.php) of the rice plant diseases with four categories of a causing agent (Fungi, Virus, Bacteria, and Aphid).

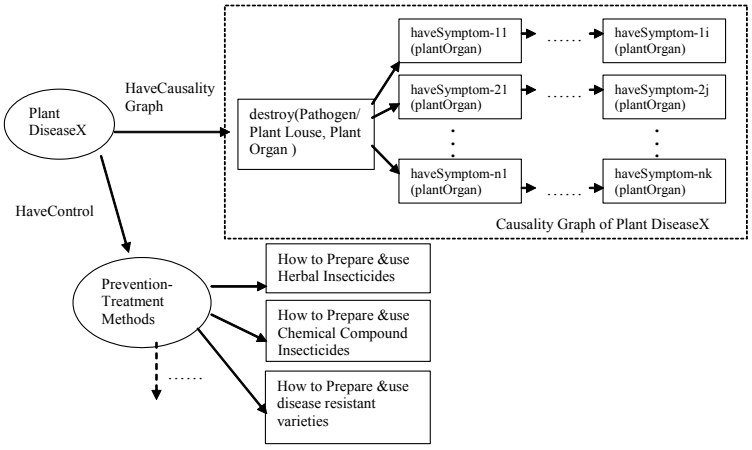


Figure1: Show the integrated causality-graph with the extracted procedural knowledge.

The integrated causality graph is used as the knowledge source for answering Why-Q and How-Q including the complicated question. In addition, each causality graph represents the causality knowledge previously extracted from documents downloaded from the Department of Agriculture website. The extracted causality knowledge has been kept in the repository as a cause-effect-EDU vector $\langle EDU_{cause}, EDU_{effect-1}, EDU_{effect-2},.., EDU_{effect-m}\rangle$ of each disease under a certain causing agent category (where EDU_{cause}, is a causative EDU, EDU_{effect} is an effect EDU).

Several techniques of the Why-QA system and the How-QA system, [Girju, 03], [Schwitter et al., 04] [Verberne et al., 07] [Baral et al., 12][Oh et al., 13], have been considered in this research (see section 2). Also several techniques, [Takechi et al., 03][Delpech and Saint-Dizier, 08][Soison and Pechsiri, 09][Song et al., 11] [Zhang et al., 12], have been previously applied for extracting the procedural knowledge (see section 2), where the procedural knowledge is the knowledge of how to perform a specific task or how to solve a problem [Schwitter et al.,04]. However, the procedural knowledge extraction in this research is developed from [Soison and Pechsiri, 09] by using different machine learning techniques with different domains. Finally, we determine the answers of Why-Q and How-Q by applying the similarity-score collection of two EDUs among each EDU from the content EDU vector ($\langle EDU_{ct1}$ $EDU_{ct2}\ldots EDU_{ctm}\rangle$ where m=n or n+1) and each EDU from the cause-effect-EDU vectors of several diseases from the causality knowledge repository. Each cause-effect-EDU vector is the component of a certain causality graph of a certain disease as shown in Figure 1.

In section 2, related works are summarized. Problems of Why/How-question determination and problems in extracting the procedural knowledge from documents are described in section 3. Our framework of a Why and How QA system including procedural knowledge extraction from textual data is shown in section 4. We evaluate and discuss our proposed methodology in section 5 and give a conclusion in section 6.

2 Related Works

Other related works to address several techniques, required for Why-Q and How-Q of our system and also for the procedural knowledge extraction, have been involved with Natural Language Processing and machine learning.

Why and 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, 03] 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 our research mostly based several effect-event explanations which express by verbs/verb phrases. [Schwitter et al., 04] 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.

The high performance in their QA system is best achieved through logic-based and pattern-matching techniques. [Verberne et al.,07] 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%. [Baral et al.,12] 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 forms: "How are X and Y related in the process Z?" and "Why is X important to Y. [Oh et al.,13] used intra- and inter- sentential causal relations between terms or clauses as evidence for answering Why-questions. They ranked their candidate answers (from documents retrieval Japanese web texts) with the ranking function including re-ranking the answer candidates done by a supervised classifier (SVM). Their why-QA system achieve 83.2% precision.

However, most of previous researches on a Why-QA system / a How-QA system [Girju, 03][Schwitter et al., 04][Baral et al., 12] are based on a single sentence/one EDU of a Why question and also a How question, except [Verberne et al., 07] and [Oh et al.,13] based on two EDUs of Why question, whereas our Why-Q and How-Q are based on several EDUs (see section 1).

Procedural Knowledge Extraction

Several techniques have been applied for extracting 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 [Takechi et al., 03] is based on HTML list tags, e.g. <OL>,<UL>, learned by SVM to determine the Procedural class. [Delpech and Saint-Dizier, 08] recognized the procedural knowledge by using XML tag, e.g. <p>, <b>, and <h>, bold letter to identify the title/goal and using a procedural writing style that contained the numbering form, hyphens or bullets in front of each process step to identify the procedure/instruction. Since there are several zero anaphora in our corpora, our procedural knowledge are still based on verb or verb phrase as in [Soison and Pechsiri, 09] whereas [Song et al., 11] and [Zhang et al., 12] involve with noun phrases. Moreover, the procedural knowledge of this research consists of several procedure sets in Natural Language description existing in one document for solving the same problem or having the same target such as Prevention & Treatment of a certain plant disease. Each procedure set contains several EDUs as process steps without the numbering form, hyphens or bullets in front of each process step. [Soison and Pechsiri, 09] also has several procedure sets in one document but each set has its own target of solving problem. The existences of the procedural knowledge on documents of the previous researches have different structure occurrences from our research. Therefore, we apply Word-Co and different machine learning techniques as NB, SVM, and ME for comparative study of extracting all procedural knowledge as the answer of How-Q.

3 Research Problems

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

3.1 Procedural Knowledge Extraction Problems

There are two main problems: the first problem is how to identify the procedural knowledge from documents after knowing the target as the problem solving e.g. Prevention & Treatment of plant diseases. The target is identified by using a target word, *tw*, existing in either a topic-name or an EDU of the plant disease domain (where *tw*∈TW, and TW is a target word set collected from corpus studying).
TW = {‚*ป้องกัน*/*prevent*' ‚*รักษา*/*treat*' ‚*ควบคุม*/*control*' ‚*กำจัด*/*eliminate*' ‚*การป้องกัน*/*prevention*' ‚*การรักษา*/*treatment*' ‚*การควบคุม*/*control*'‘*การกำจัด*/*elimination*'...}
The second problem is how to determine the procedural knowledge boundary.

3.1.1 Procedural Knowledge Identification Problem

There are two problems: the implicit cue and the ambiguous cue.

(a) Implicit Cue. The procedural knowledge can be identified by using the starting-procedural cue set { ‚*ดังต่อไปนี้*/ *the following*' ‚*ดังนี้*/*as follows*' „*โดย*/*By*'…}, as shown in the following examples of an explicit cue and an implicit cue.

Explicit 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.

Implicit Cue EDU1:"[*เรา*]*ต้องควบคุมโรคใบไหม้ข้าว*/[*We*]*must control the Blast-Rice disease*."
EDU2: "*มันเริ่มระบาดแล้ว*/ *It has started spreading out*."
EDU3: "*ใช้เชื้อบาซิลลัส*/ *Use Bacillus Subtilis*." EDU4……
where EDU1 is the target, EDU3 is the starting EDU of the procedural knowledge.

(b) Ambiguous Cue. There are some EDUs expressing as the non procedure even though they contain a cue, as in the following example:
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.

3.1.2 Procedural Knowledge Boundary Determination Problem

The problem is how to identify the ending of each procedure, especially there is no any cue, e.g. ‚*และ*/*and*', ‚*หรือ*/*or*',„*ในที่สุด*/*finally*' etc., telling the ending boundary. And there are 2-3 different procedural knowledge solving the same plant-disease problem occurred in one document. For example:
EDU1:"*น้อยหน่าสามารถกำจัด เพลี้ยกระโดดสีน้ำตาล* /*A sugar apple can kill Brown Plant Hopper*."
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: "*ฉีดพ่นทุกๆวันเนาน6-10 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: "*ผลลัพธ์จากการใช้น้อยหน่า…./ The results of using 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.

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

V_{proc} ={ „*ใช้/use*',„*นำ/take*',„*หว่าน/scatter*',„*ทำลาย/destroy*',„*ปลูก/grow*', „*ปล่อย/release*', …}
N_{Proc} ={ „ ', „*ส่วนประกอบพืช/Plant Organ*', „*พันธุ์ต้านทาน/resistant variety*', „*สารเคมี/chemical substance*', „*ยา/pesticide*', „*เชื้อ/micro-organism*', „*น้ำ/water*', …}

We also apply machine learning techniques (NB, SVM, and ME) to learn the procedural verb pairs from the consecutive EDUs by a sliding window size of two consecutive EDUs with one EDU sliding distance for the procedural-knowledge-boundary determination

3.2 Why and How QA System Problems

There are three main problems: how to identify Why-Q, How-Q, and Other-Q on Qpatterns with their question words are ambiguity, how to solve the Complicated-Q, and how to solve the Why and How answers..

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 is a reasoning question, a „*อะไร/What*' function is asking for information about something (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. For example:

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

3.2.2 Complicated-Question Problem

The questions based on problem solving are difficult to answer such as How-Q. For example:
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*?"
This type of question can be answered after knowing the disease name or the cause of the symptoms.

Therefore we propose using different machine learning as NB, ME and MLP to classify three question types as Why-Q (a reasoning question or a causality question), How-Q (the causality How question including the complicated question), and Other-Q (Other-question). The features used in this classification after stemming words consist of Qword, V_{ct} , and V_q from two adjacent EDUs (EDU_q and EDU_{ctk} where k=1or n or n+1).

3.2.3 Determination of Why and How answers

Unlike the question word sets from the factoid questions, the answers of the Why and How questions can not 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 a question word „*Who*' [Agichtein et.al., 05] whereas the question word „*Why*' in Qpattern cannot be applied to determine the answer. Moreover, the „*Why*' question word have previously been approached by determining answers from noun phrases and question words [Verberne, 08], which is not suitable for our „Why' question based on several effect-event explanations. Therefore, we solve the answers of Why-Q and How-Q on Qpatterns by applying the similarity-score collection of two EDUs among EDU_{cta} of the content EDU vector and each EDU from the cause-effect-EDU vectors of several diseases, see section 4.2.3, after the stop word elimination (where a cause-effect-EDU vector is the causality graph component of a certain disease). And, the similarity score determination is based on WordNet and Thai Encyclopedia after using Thai-to-English dictionary.

4. Framework of Why and How QA System

The Why and How QA system of this research consists of two major parts, a question part and an answering part included procedural knowledge extraction. There are three steps in the question part, the first is Question Corpus Preparation. The second is Learning of Why-Q, How-Q, and Other-Q on Qpatterns and the third is Identification of Why-Q, How-Q, and Other-Q. The answering part consists of three main steps, the first is Procedural knowledge Extraction from Texts. The second is Integration of Causality Graph and Extracted Procedural Knowledge and the third is Answer Determination, as shown in Figure 2.

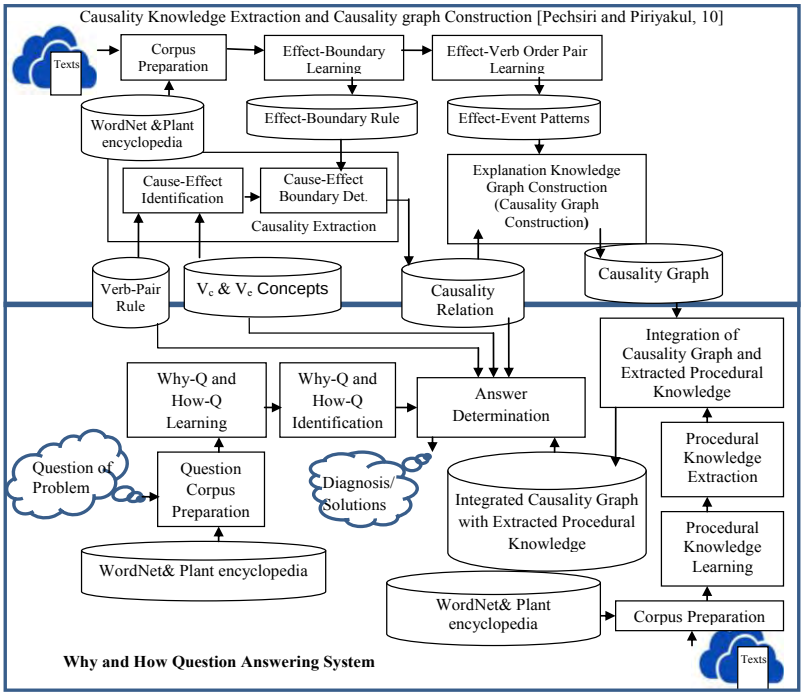


Figure 2: System Overview

.....

EDUct1: "ใบข้าว/*Rice leaves* หงิกงอ/*shrink*" ("*Rice leaves shrink.*")

EDUct2: "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*?")

<**EDUct1**>[ใบ/ncn ข้าว/ncn]/NP

[<**Qfocus**> <**Vct** : **Ve-concept**= 'shrink/change shape'>หงิก/vi งอ/adv **</Vct>**

</Qfocus>]/VP**</EDUct1**>

<**EDUct2**> [ต้น/ncn]/NP

[<**Qfocus**> <**Vct**: **Ve-concept**= 'stunt'>แคะแกรน/vi **</Vct>**</**Qfocus**> [ช่วง/ncn แตก/vi กอ/ nct]/NP]/VP**</EDUct2**>

<**EDUq**> [ϕ=we]/NP

[จะ/prev <**Vq** : concept= 'solve'>แก้/vt **</Vq>**

<**Qword**=How: **concept=Complicate-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, 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, ϕ stands for a zero anaphora or ellipsis.

Figure3: Examples of question annotation

4.1 Question Part

4.1.1 Question Corpus Preparation

The preparation of the question's corpora with 8000 EDUs downloaded from the online community websites with three different communities; a farmer community (650 questions in plant diseases from farmer-community web-boards, e.g. www.kasetporpeang clu.com), a health-care community (650 questions from health-care community web-boards, e.g. http://haamor.com), and a technology-and-indigenous-technology community (650 questions from echnology-and-indigenous-community web-boards, e.g. http://www. gotoknow.org/posts/325634). All of these questions involve using Thai word segmentation tool to solve a boundary of a Thai word and to tag its part of speech [Sudprasert and Kawtrakul ,03], including Name Entity [Chanlekha and Kawtrakul, 04]. EDU segmentation [Chareonsuk et al., 05] is then to be dealt with to generate EDUs for the semi-automatic annotation of question type concepts, a causative-verb concept (v_c) and an effect-verb concept (v_e) as shown in Figure 3 based on word stems. Where the causative-verb concept set (V_c, and $v_c \in V_c$) and the effect- verb concept set (V_e, and $v_e \in V_e$) are also provided by [Pechsiri and Piriyakul, 10] shown in Table 1 and the concepts are referred to Word Net [Miler et al., 93](http://wordnet.princeton.edu/obtain) and Thai Encyclopedia of plant disease(http://kanchanapisek.or.th/kp6/) after using the Thai-to-English dictionary (http://longdo.com). In addition, 1950 annotated questions from those three online communities based on web-boards are divided into 2 parts for the question classification, the 1500 questions'part for learning based on ten folds cross validation and the other part of 450 questions for testing.

Verb type		Surface form	Conceptual class
V_c (Causative verb)	strong verb	ดูด/*suck*, ดูดกิน/*suck. กิน/eat*, กัด/*bite*,	Consume/destroy
		ทำลาย/*destruct*, กำจัด/*eliminate*, ฆ่า/*kill*, หัก/*break*,	Destroy
	weak verb	เป็น+โรค/ *be+ disease*,	get disease
		ได้รับ+เชื้อโรค/*get+ pathogen*,	get pathogen
		……..	……
V_e (Effect verb)	strong verb	หงิก /*shrink*, งอ/*bend*, บิด/*twist*, โค้งงอ/*curl*	change shape
		แห้ง/*dry*, ไหม้/*blast*,	dry/be symptom
		เหี่ยว/*wilt*	lose water/be symptom
		แคระแกรน/*stunt*	stunt/be symptom
	weak verb	เป็น+จุด /*be+spot*, เป็น+แผล /*be+ scar*	be mark / be symptom
		มี+จุด /*have+spot*, มี+แผล /*have+ scar*	have mark / have symptom
		มี+สี/*have+color*	change in color/ have symptom
		.....	.....

Table 1: Show the causative-verb concept set (Vc) and the effect-verb concept set (Ve) [Pechsiri and Piriyakul, 10]

4.1.2 Learning of Why-Q, How-Q, and Other-Q

This step is using Weka (http://www.cs.waikato.ac.nz/ml/weka/) to learn Why-Q, How-Q, and Other-Q from the annotated question corpora based on Qpatterns by applying three different machine learning techniques, NB, ME, and MLP. The features used in these learning techniques are Qword (where $qw \in Qword$), the V_{ct} and V_q (where $v_{ctk} \in V_{ct}$, $V_{ct}=V_c \cup V_e$, $v_q \in V_q$) of two adjacent EDUs as EDU_q and EDU_{ctk} (where k=1 or n or n+1) from the annotated corpora.

Naïve Bayes (NB) According to [Mitchell, 97], the NB learning is a generic classification to determine the feature probabilities of three classes according to the question types with Qpatterns(class1='Why-Q',class2='How-Q',class3='Other-Q'). The features of NB classifiers consist of three feature sets: Qword, V_{ct}, and V_q, from the annotated corpora of two adjacent EDUs (EDU_q and EDU_{ctk}).

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 [Fleischman et al., 03]. They modeled the probability of a semantic role r given a vector of features *x* according to the ME formulation below:

$$p(r \mid x) = {}^{1}\!/_{z_x} \exp[\sum_{j=0}^{n} \lambda_j f_j(r,x)] \quad (1)$$

Where Z_x is a normalization constant, $f_i(r,x)$ 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 equation 1, ME can be used as the classifier of the r class when p(r|x) is the highest probability or argmax p(r|x) to determine four question-type classes. Where r is the question-type class value (a question-type class is „Why-Q' if r=1, „How-Q' if r=2, and „Other-Q' if r=3) and x is the binary vector consisted of all consecutive elements of three feature sets: Qword, V_{ct}, and V_q , from two adjacent EDUs of EDU_q and EDU_{ctk} as shown in equation 2.

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

Multi-Layer Perceptrons (MLPs) According to [Haykin, 99] 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 equation 3.

$$y_i = f(\sum x_i w_i) \quad (3)$$

where w_i represents the weight, x_i is the input feature from all consecutive elements of three feature sets: Qword, V_{ct}, and V_q , from two adjacent EDUs (EDU_q and EDU_{ctk}). $f(\cdot)$ 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. Each layer includes a different number of processing nodes. Then, the net weighted input can be solved by equation 4 where n is the number of neuron inputs, θj is the threshold value of neuron at the *jth* node in the hidden layer, and the number of hidden layers p=2.

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

4.1.3 Identification of Why-Q, How-Q, and Other-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 [Mitchell , 97], equation 5 and the feature-probabilities determined by the previous step of NB are used to identify the class of the question type with Qpattern by the algorithm shown in Figure 4.

$$\begin{aligned} QpatternClass &= \arg\max_{class \in Class} P(class \mid v_{ctk}, v_q, qw) \\ &= \arg\max_{class \in Class} P(v_{ctk} \mid class) P(v_q \mid class) P(qw \mid class) P(class) \end{aligned} \quad (5)$$

$$\begin{aligned} &where \quad v_{ctk} \in V_{ct} \ where \ V_{ct} \ is \ \text{a verb conceptset expressed on } EDU_{ctk} \\ &\quad v_q \in V_q \ where \ V_q \ is \ \text{a verb conceptset expressed on } EDU_q \\ &\quad qw \in Qword \quad where \ Qword \ is \ \text{a question-word conceptset} \\ &\quad \text{Class} = \{\text{class1}, \text{class2}, \text{class3}\} \end{aligned}$$

Maximum Entropy We use λ_j (the weight for a given feature function of the binary vector) resulted from learning Why-Q, How-Q, and Other-Q to determine the classes of the question types by equation 2 as shown in the algorithm of Figure 4 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 equation 4 as shown in the algorithm of Figure 4 with the MLP case.

```
Assume that each EDU is represented by (NP VP).  L is a list of EDUs with Qpattern.
EDUq →Qword NP1 Vq NP2 | NP1 Vq NP2 Qword | Vq Qword, vq∈Vq w∈Qword
EDUctk →NP1 Vct NP2   vct∈Vct which is the verb concept set of the EDUctk where
k=1or n or n+1
QUESTION_TYPE_DETERMINATION ( L )

 1  i ← 1, flagQ←0count←0
 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
 5     { If EDUi is EDUq ;  flagQ=1
 6       If i=1 then { EDUi+1 is EDUct1  };
 7       If i=count-1 then { EDUi-1 is EDUctn  and  EDUi+1 is EDUctn+1}
         If i=count  then { EDUi-1 is EDUctn }   }
12    i++ }
13    If  flagQ=1
14       Case: use NB
                Equation5
          Case: use ME
                Equation 2
           Case: use MLP
                Equation 4
            End_case
15  Return
```

Figure 4: Determination of Question Types with Qpatterns by NB, ME, or MLP

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 Figure2.

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 (downloaded from the online community web-board, http://www.kasetporpeang.com/forums), a plant disease domain (downloaded from the Department of Agriculture website, http://www.doa.go.th/), and a news domain (especially in indigenous technology, http://info.matichon.co.th/techno/). The step involves with using Thai word segmentation tools with tagging its part of speech [Sudprasert and Kawtrakul, 03], including Name entity [Chanlekha and Kawtrakul, 04], and EDU segmentation [Charoensuk et al., 05]. These 6000 EDUs corpora from three domains are separated into 2 parts, the 4500 EDUs' part for learning procedural knowledge based on 10 folds cross validation and the 1500EDUs' part for testing. In addition to the learning part, we semi-automatically annotate the procedural EDUs, as shown in Fig.5, with the target tag as the problem solving, a verb concept and a noun concept referred to WordNet and Thai Encyclopedia after using the Thai-to-English dictionary.

"เมล็ดน้อยหน่า สามารถกำจัดแมลง เช่นเพลี้ยกระโดดสีน้ำตาล ใช้เมล็ดน้อยหน่า 1 กก.ต้าละเอียด แช่น้ำ 10 ลิตร นาน 12-24 ชั่วโมง กรองน้ำผสมน้ำสบู่ 1 ช้อนโต๊ะ ฉีดพ่นทุกๆ 6-10 วัน ช่วงเวลาเย็น ใช้น้อยใหน่งแทนน้อยหน่าได้เช่นกัน"
("*Sugar Apple seeds. [It] can kill insects, i.e. Brown Plant Hopper . 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 a custard apple to replace a sugar apple.*")

<**Topic**><np concept=herb#1 type=**title**>**น้อยหน่า / *Sugar Apple*** </np></**Topic**>
<EDU type=**target id=1**> <**V** concept=use#1>**ใช้**</V><**Vt** concept=**kill**#1>**กำจัด /*kill***</**Vt**> <np>แมลง
เช่นเพลี้ยกระโดดสีน้ำตาล/ *insects i.e. Brown Plant Hopper*</np> </EDU>
<EDU type=**PrepProc** of id1 ><**Vproc** concept= **use**#1>**ใช้/*use***</Vproc>
<nproc concept= **plant organ** >**เมล็ดน้อยหน่า** 1 กก./ *1kgs.sugar apple* ***seeds*** </nproc></EDU>
<EDU type= **PrepProc** of id1>< **Vproc** concept=**hit**#1 >**ต้าละเอียด / *Grind finely*** </V></EDU>
……
<EDU type= **TreatProc** of id1>< **Vproc** concept=**spray**#2 >**ฉีดพ่น/ *Spray*** </V> ทุกๆ 6-10 วัน ช่วงเวลา
เย็น/every day for 6-10 days in the evening</EDU>
<EDU type=**non procedure** of id1 ><**V** concept= **use**#1> **ใช้/*use***</V>
<np concept= **plant** >**น้อยโหน่ง/ *a custard apple*** </np>
<EDU type=**non procedure** of id1 ><V concept=replace#1 >แทน</V><np concept= plant >น้อยหน่าได้
เช่นกัน / *a sugar apple.*</np> </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 Vt tag is a target verb tag, a Vproc tag is a procedural verb tag, a nproc tag is a procedural noun tag, a V tag is a verb tag of an EDU, and a np tag is a noun phrase tag.

Figure 5: Example of annotated corpus

4.2.1.2 Procedural Knowledge Learning

There are two necessary learning, learning RelatednessValue and learning Boundary.

(a) Learning Relatedness Value The objective of this learning is to learn the relatedness value (r) [Guthrie et al., 91] between two consecutive words, v_{proc} n_{proc} as Word-Co (see section 3.1.2) with the procedural knowledge concept as shown in equation (6). Thus, Word-Co is used to identify the starting procedural knowledge after a target topic or a target EDU has been identified by *tw* (where $tw \in$ TW (from section3.1).

$$r(v_{proc}, n_{proc}) = \frac{fv_{proc}n_{proc}}{fv_{proc} + fn_{proc} - fv_{proc}n_{proc}} . \quad (6)$$

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} *occurences*. fw_{proc} *is the numbers of* n_{proc} *occurences*.

$fv_{proc}n_{proc}$ *is the numbers of* v_{proc} *and* n_{proc} *occurences*.

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

(b) Learning Procedural Knowledge Boundary. We use Weka to learn the procedural knowledge boundary by three different machine learning techniques, NB, ME, and SVM. The features used in learning the procedural knowledge boundary are based on the events expressed by verbs. Thus, all annotated verbs from the corpus 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 from a procedural EDU, non-p is non procedural verb from a non procedural EDU.
V = {Vi} where i=1..n

Naïve Bayes We using Weka to determine the probability of procedural relation and non procedural relation from a verb concept pair (v_{ih} v_{ih+1}) by sliding a window size of two consecutive EDUs (EDU_{ih} EDU_{ih+1}) with one EDU sliding distance.

Maximum Entropy According to equation (1), 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 classes (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 one EDU sliding distance, as shown in equation 7.

$$p(r \mid x) = \arg\max_{r} \frac{1}{z} \exp\Big(\sum_{j=1}^{n} \lambda_j f_{yes,proc-ih,j}(r, v_{ih}) + \sum_{j=1}^{n} \lambda_j f_{no,proc-ih,j}(r, v_{ih}) + \sum_{j=1}^{n} \lambda_j f_{yes,proc-ih+1,j}(r, v_{ih+1}) + \sum_{j=1}^{n} \lambda_j f_{no,proc-ih+1,j}(r, v_{ih+1})\Big) \quad (7)$$

Support Vector Machine The linear binary classifier, SVM, applies in this research to classify the procedural knowledge boundary with ending or with continuing of each procedural verb pairs from the annotated corpus by using Weka. According to [Vapnik, 95] 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:

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

where x is a dichotomous vector number, w is weight vector, b is bias, and $(w,b) \in R^n \times R$ 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 one EDU sliding distance.

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 TW. Then, the Word-Co set from the learning step in section 4.2.1.2 is 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 follow by the algorithm shown in Figure6.

Naïve Bayes According to [Mitchell T M., 97], NB Classifier (equation 9) is applied to solve the boundary by sliding a window size of two consecutive EDUs with one EDU sliding distance along with verb concept probabilities from the learning step.

$$\begin{aligned} EDUclass &= \arg\max_{class \in Class} P(class \mid v_{ih} v_{ih+1}) \\ &= \arg\max_{class \in Class} P(v_{ih} \mid class) P(v_{ih+1} \mid class) P(class) \end{aligned} \quad (9)$$

$$where \;\; v_{ih} \in V_i \; is \; a \; verb \; concept \; vector, \quad v_{ih+1} \in V_i \; is \; a \; verb \; concept \; vector$$

$$i = \{1,2,..n\} \quad h = \{1,2,..m\} \quad Class = \{0,1\}$$

As soon as the class 0 or non procedural relation is determined, the procedural knowledge boundary is ended as shown in Fig. 6.

```
Assume that each EDU is represented by ( NP1 V NP2).
L is a list of EDUs.  Word-Co is a Word-Co set with procedural concepts.
V_ih, V_ih+1  are learning verb sets. TW is a target word set.
nt∈NT which is a target-noun concept set
PROCEDURAL_KNOWLEDGE_EXTRACTION ( L, V_ih, V_ih+1,TW)

1   i ← 1; R← ∅;   TG← ∅ ; PROC← ∅; flag=no
2    While flag=no
3   { If TargetEDUorTargetToppicFound  then {TG ← TG ∪ {i}; flag=yes }
4            Else i ++ }
    { If TG<> ∅ then { flag=no ; flagP=yes; count=1}
5   While i ≤ length[L] do
6   {  While flagP=yes ∧ i ≤ length[L] /* FindStartProcEDU
7     { If FindStartProcKnowledgeEDU then { flagP=no; flag=yes}
8       Else i ++; }
9     While (v_i ∈V_ih )∧(v_i+1 ∈ V_ih+1)∧ flag=yes ∧ i ≤ length[L] do /*BoundaryDet.
10      {     Case: useNB
                Equation 9   If class=0  then flag=no
              Case: useME
                Equation 7   If r=0 then flag=no
              Case: useSVM
                Equation 8   If f(x) ≥0 then flag=no
              EndCase
11            if flag= no ∧ TG <>∅ then PROC ← PROC ∪ {i};
12            i ++ ;
13          }; R = R ∪ (TG,PROC); flagP=yes;
14   };  return R
```

Figure 6: Procedural Knowledge Extraction algorithm

Maximum Entropy We use λ_j resulted from the ME learning to determine the procedural knowledge boundary by equation 7 as shown in Figure6. 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 one EDU sliding distance.

Support Vector Machine The results from SVM learning are weight, w_i, and bias, b, of each verb feature (x_i). According to equation (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 one EDU sliding distance) including their weights and bias are used to determine the boundary. If $f(x)\geq 0$,an ending class is occurs, otherwise a continuing class as shown in Figure 6.

4.2.2 Integration of Causality Graph and Extracted Procedural Knowledge

According to [Pechsiri and Piriyakul, 10], the causality graph has been constructed from the extracted causality knowledge from documents. The extracted causality knowledge including a disease name from a document topic are based on a causative event with several effect events. The causative event is expressed by a causative verb concept set (V_c) and the effect events are expressed by an effect verb concept set (V_e) (Table 1). Thus, a causality graph consists of a disease name, effect nodes which are all graph nodes except a root node, and a causative node which is a root node, as shown in Figure 7.

Therefore, we integrate our previous causality graph with the extracted procedural knowledge as shown in Figure 8 after the plant disease name of the causality graph is a substring of either the topic name or EDU_{target} of the extracted procedural knowledge.

Figure7: Causality Graph [Pechsiri and Piriyakul, 10] [Pechsiri and Piriyakul, 12] of Rice Blast Disease caused by Fungus (http://www.web3point2.com/rice/indexApp.php)

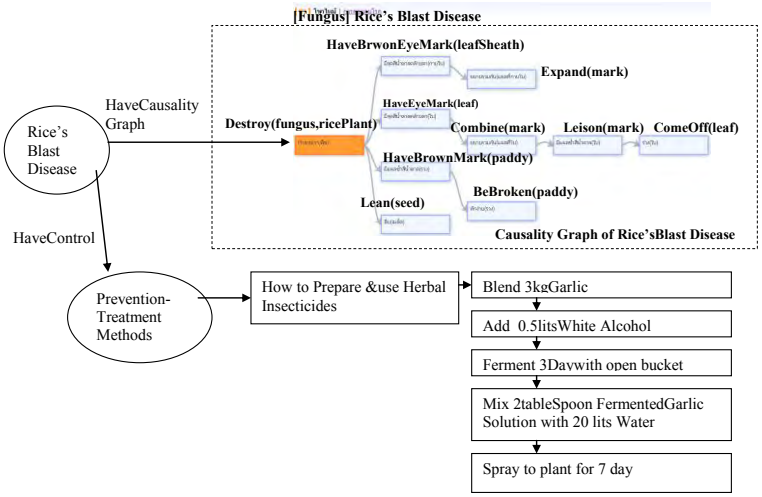


Figure 8: Show an example of the causality graph of Rice Blast Disease (caused by fungus) integrated with the extracted procedural knowledge of using herbal insecticide

4.2.3 Answer Determination

The 180 questions, randomly selected from the 418 correct-question-type Identification from section 4.1.3, consist of both 45 questions of Why-Q and 45questions of How-Q from the plant disease domains, especially on rice diseases, 45 questions of Why-Q from the health-care domain, and 45 questions of Why-Q from the indigenous-technology domain. The selected questions is used for determining the anwers based on the Information Retrieval (IR) approach by ranking the similarity-score collection of two EDUs among EDU_{cta} of the content EDU vector and each EDU from the knowledge base as the cause-effect-EDU vectors of several diseases after the stop word elimination. The possibilty answer can be solved from the selected cause-effect-EDU vectors that has Rank 1 which is the highest rank of the similarity-score collection. The answer representation of our research can be expressed by the visualization of the the integrated causality graph for the rice diseases where the previous research have extracted the knowledge of cause-effect-EDU vectors as much as to contruct the causality graph. Whereas the answers from the other domains can be present by the the cause-effect-EDU vectors for Why-Q instead of the visualized causality graph because of lacking the number of high occurences of Thai documents on a certain area to construct the graph. Since the focuses of Why-Q and How-Q from Qpatterns are based on the events expressed by V_{ct} where $V_{ct} = V_e \cup V_c$, both Why answers and How answers can be solved by determining the similarity score [Biggins, 12] in equation (10) including the similarity-score collection among EDU_{cta} and each EDU element of the cause-effect-EDU vectors after eliminating stop words.

$$Similarity\_Score = \frac{|S1 \cap S2|}{\sqrt{|S1| \times |S2|}} \quad (10)$$

where S1 is an EDU_{cta} of the content EDU vector (having a=1,2,..,n or n+1) after eliminating stop words.

S2 is an EDU_{cause} or an $EDU_{effect\text{-}i}$ of the cause-effect-EDU vector $\langle EDU_{cause}, EDU_{effect\text{-}1}, EDU_{effect\text{-}2},.., EDU_{effect\text{-}m}\rangle$ after eliminating stop words.

All word concepts of a S1 EDU and a S2 EDU are based on WordNet and Thai Encyclopedia after using the Thai-to-English dictionary. The number of words in the S1 EDU and the number of words in the S2 EDU are not significantly different. In addition, there are four categories of causality graphs based on causing agents as Fungus, Virus, Bacteria and Plant-Louse. Each causality graph represents each cause-effect-EDU vector of each disease is integrated with its procedural hnowledge of prevention and treatment. There are 69 different S2 EDUs by the union operation of thirteen cause-effect vectors (or 13 diseases) after eliminating stop words. The candidate disease ($Disease_i$) can be selected if its S2 EDUs are similar to any S1 EDUs with Similarity_Scores ≥ 0.5. Then, the answers can be ranked according to the number of the candidate disease selection. For example:

Qpattern-1: EDU_{ct1}-->$S1_1EDU_{ct2}$-->$S1_2$...EDU_{ctn}-->$S1_n$ EDU_q:What are the causes?

where $EDU_{ct1} \neq EDU_{ct2} \neq \ldots \neq EDU_{ctn}$

The candidate answers are ranked by sorting the number of selected $Disease_i$ after determining the collection of the Similarity_Scores (see Table 2)

Diseases	If Similarity_Score($S1_a$,$S2_j$)of $Disease_i$ > 0.5 then $Disease_i$ is selected with $S2_{ij}$=1 where a=1,2,..,n i=1,2,..,13 j=1,2,..,69						The number of Selected $Disease_j$ (NSD)	Rank by sorting NSD
	$S2_1$	$S2_2$	$S2_3$	$S2_4$	…	$S2_{69}$		
$Disease_1$							0	
$Disease_2$			1	1		1	3	1
$Disease_3$				1			1	3
…..							0	
$Disease_{13}$			1			1	2	2

Table 2: Show how to rank $Disease_i$ as the candidate answers for Qpattern

From Table2, the answer having the highest rank is $Disease_2$ (Rank1) and the answer having the lowest rank is $Disease_3$ (Rank3). Moreover, the answer of How-Q can be solved after ranking the number of selected $Disease_i$ where each disease is connected to the certain integrated causality graph.

5 Evaluation

5.1 Data

There are two categories of corpora for evaluation our propose model, the question corpora and the procedural text corpora. The question corpora for evaluating the proposed model of classifying the question types, Why-Q, How-Q, and Other-Q, based on prefaced questions contain 450 question equally collected by three different

domains from the online community websites; the rice-disease domain, the health-care domain, and the technology-and-indigenous-technology domain. The 180 selected questions from the correct-question-type identification are used for the answer evaluation based on IR approach. The corpora for the procedural knowledge extraction are collected from three domains, the herbal pest control documents, the plant disease domain, and a news domain (especially in indigenous technology). Both corpora categories are emphasized on events expressed by verbs whilst the procedural corpora have different characteristics of the frequency of verb features, and the diversity of verb feature occurrences including the feature dependencies. All of these characteristics make this research analyze how verb features effect to the results of using the different machine learning techniques for question identification and knowledge extraction.

5.2 Question Part

Domain (Each domain contains 150 questions)	#of Feature-Dependency Occurrences (v_{ctk}-v_q-qw)	#of verb Diversity Occurrences	MLP		ME		NB	
			Precision	Recall	Precision	Recall	Precision	Recall
PlantDisease	medium	89	**0.927**	0.836	0.910	0.827	0.859	0.777
HealthCare	medium	98	0.919	**0.840**	**0.930**	0.838	0.851	0.789
Indigenous Techno. &Auto. Techno.	low	115	**0.905**	0.823	0.886	0.805	**0.877**	**0.795**

Table 3: The Correctness of Why-Q and How-Q Classification

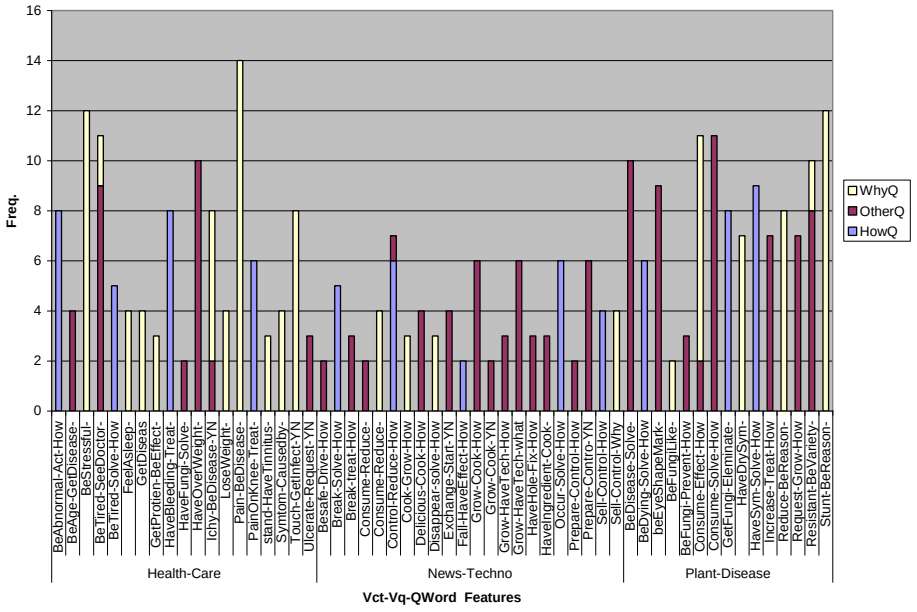


Figure 9: Show the feature dependency occurrences among different domains

The evaluation of the Why-Q, How-Q, and Other-Q classified in this research is expressed in terms of the precision and the recall based on human judgements (two experts and one linguistic) with max win voting. Table3 shows ME results in the highest precision of 0.93 for the health-care domain contains more feature dependency occurrences. The new domain of technology contains the highest diversity of verb feature occurrences (or low frequency of verb feature occurrences) and the lowest feature dependcy occurrences (see Figure 9) which result in a higher precision of 0.877 by NB comparing 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 0.777 for the plant disease corpus containing more question-word-ellipsis occurrences of the posted problems on the web-boards.

5.3 Answering Part

The procedural knowledge extraction as the knowledge source of How-Q is also evaluated in term of the precision and the recall based on three experts with max win voting as shown in Table 4. Word-Co, v_{proc} w_{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 the medium verb-pair-feature-dependency occurrences and the medium diversity of verb feature occurrences whereas NB gives the lowest %correctness of 75.3 for the indigenous technology corpus containing the very highest diversity of verb feature occurrences. And, ME results in the boundary determination of the plant disease corpus (containing the high verb-pair-feature-dependency occurrences and the lowest diversity of verb feature occurrences) to have the highest %correctness of 94.4 comparing to SVM and NB of the plant disease one.

* Same Herb domain as [Soison and Pechsiri, 09]based on NB

Each domain contains 500 EDUs	#of verb-pair feature Dependency Occurrences	#of verb Feature Diversity Occurrence	Procedural Knowledge Identification by Word-Co		Boundary Determination		
					SVM	ME	NB
			Precision	Recall	%correctness	%correctness	%correctness
Plant Disease	high	74	0.96	0.92	91.5	**94.4**	87.6
*Herbal Pest Control	medium	156	0.97	0.93	**95.8**	92.3	89.7
Indigenous Techno.	medium	228	0.94	0.97	85.2	87.8	**75.3**

Table 4: The evaluation of procedural knowledge extraction from texts

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

Answer Expression	Correct Answer (rank1)			
	HealthCare	Indigenous Techno	RiceDisease	
	Why-Q(45)	Why-Q(45)	Why-Q(45)	How-Q (45)
Integrated Causality Graph	-	-	42 (93.3%)	40(88.9%)
the cause-effect-EDU vector	41(91.1%)	38(84.5%)	-	-

Table 5: The evaluation of the answer determination

Table 5 shows that the integrated causality graph representation of the answers on the rice disease domain can provide the answers correctly at rank1 of Why-Q and How-Q as 93.3%, and 88.9% respectively. Whereas the Indigenous Technology domain has the lowest % correct answer of 84.5 by the cause-effect-EDU vector representation The reason of lower %correctness of either Why-Q or How-Q is that there are more zero anaphora occurrences (the ellipsis of noun phrases) on several EDU_{cta} occurrences resulted in the low similarity scores, especially on an EDU_{cta} containing three explicit words including one zero anaphora.

6 Conclusion

This paper introduces the automatic Why and How Question Answering system on the online community web-boards that provides preliminary diagnosis including the suggestion of how to solve problems to people/users while they are waiting for an expert response. The machine learning is proposed to solve the question type identification problems, especially Why-Q, How-Q, and Other-Q, and also the procedural knowledge extraction problems from text. The integration of the extracted procedural knowledge and the previous causality graph [Pechsiri and Piriyakul, 10] is provide as the knowledge source for answering the Why and How QA systemes. Thus, our Why and How QA system provide the answers with the visualization of the integrated causality graphs which make more understanding to the community than only textual answer. Howerve, the zero anaphora problem should be solved in the future work for increasing the correctness of answers. Moderately high performance has been achieved for the proposed system (tables 2 – 5 and Figure 9) showing the corpus behaviours, especially the feature dependency and the feature dyversity, effect to the application of machine learning approach. Finally, the model of our Why and How 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.

Acknowledgements

The research is supported by Thai Research Fund 2012.

References

[Agitchtein 05] Agichtein, E., Cucerzan, S., Brill, E.: “Analysis of Factoid Questions for Effective Relation Extraction“; Proc. SIGIR’05, ACM SIGIR International Conference on Research and Development in Information Retrieval, Salvador, Brazil (2005).

[Aouladomar , 06] Aouladomar, F.: “Towards Answering Procedural Questions”; Proc. Knowledge and Reasoning for Answering Questions Workshop (KRAQ-05), International Joint Conference on Artificial Intelligence, Edinburgh, United Kingdom (2005)

[Baral 12] Baral, C., Vo, N.H., Liang, S.: “Answering Why and How questions with respect to a frame-based knowledge base: a preliminary report“; Proc. ICLP 2012, Hungary (2012).

[Biggins 12] Biggins, S., Mohammed, S., Oakley, S.: “University Of Sheffield: Two Approaches to Semantic Text Similarity“; Proc. First Joint Conference on Lexical and Computational Semantics, Montre´al, Canada (2012).

[Carlson (03)] Carlson, L., Marcu, D., Okurowski, M.E.: “Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory”; Current and New Directions in Discourse and Dialogue,22 (2003), 85-112.

[Chanlekha 04] Chanlekha, H., Kawtrakul, A.: “Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information“;Proc. IJCNLP’04, China (2004).

[Chareonsuk 05] Chareonsuk, J., Sukvakree, T., Kawtrakul, A.: “Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information“; NCSEC,,05, Thailand (2005).

[Delpech 08] Delpech, E., Saint-Dizier, P.: “Investigating the Structure of Procedural Texts for Answering How-to Question“; Proc. JADT 2008, 9es Journées internationales d’Analyse statistique des Données Textuelles, (2008).

[Fleischman 03] Fleischman, M., Kwon, N., Hovy, E.: “Maximum entropy models for Frame Net classification“; Proc. EMNLP,,03, Sapporo, Japan (2003).

[Girju 03] Girju, R.: “Automatic detection of causal relations for question answering“; Proc. 41st annual meeting of the assoc. for computational linguistics, workshop on multilingual summarization and question answering-Machine learning and beyond, Japan (2003).

[Guthrie 91]Guthrie, J.A., Guthrie, L., Wilks, Y., Aidinejad, H.: “Subject-dependent co-occurrence and word sense disambiguation“; Proc. 29th annual meeting on Association for Computational Linguistics (1991).

[Haykin (99)] Haykin, S.: “Neural networks: a comprehensive foundation“; Prentice Hall, USA (1999).

[Miller 93] Miler, G.A., Beckwith, R., Fellbuan, C., Gross, D., Miller, K.: “Introduction to Word Net“; An Online Lexical Database, (1993).

[Mitchell (97)] Mitchell, T.M.: “Machine Learning“; The McGraw-Hill Companies Inc. and MIT Press, Singapore (1997).

[Oh 13] Oh, J-H., Torisawa, K., Hashimoto, C., Sano, M., Saeger, S.D., Ohtake, K.: “Why-Question Answering using Intra- and Inter-Sentential Causal Relations“; Proc.

of the 51st Annual Meeting of the Association for Computational Linguistics, Bulgaria (2013).

[Pechsiri 10] Pechsiri, C. Piriyakul, R.: “Explanation Knowledge Graph Construction through Causality Extraction from Texts”; J.Comput.Sci.&Technol. (Journal of Computer Science and Technology), 25,5 (2010), 1055-1070.

[Pechsiri 12] Pechsiri, C., Piriyakul, R.: “Developing the UCKG-Why-QA System“; Proc. ICCCT„12, South Korea (2012).

[Schwitter 2004] Schwitter, R., Rinaldi, F., Clematide, S.: “The Importance of How-Questions in Technical Domains“; Proc. TALN-04, Workshop Question – Réponse, Fez, Morocco (2004).

[Soison 09] Soison, P., Pechsiri, C.: “Know-How Extraction from Textual Data for Problem Solving“; Proc. Artificial Intelligence and Applications, Innsbruck, Austria (2009).

[Song 11] Song, S-K., Oh, H-S., Myaeng, S.H., Choi, S-P., Chun, H-W., Choi, Y-S., Jeong, C-H.: “Procedural Knowledge Extraction on MEDLINE Abstracts“; Proc. AMT 2011, LNCS 6890 (2011), 345–354.

[Sudprasert 03] Sudprasert, S., Kawtrakul, A.: “Thai Word Segmentation based on Global and Local Unsupervised Learning“; Proc. NCSEC’2003, Chonburi, Thailand (2003)

[Takechi 03] Takechi, M., Tokunaga, T., Matsumoto, Y., Tanaka, H.: “Feature Selection in Categorizing Procedural Expressions”; Proc. of the Sixth International Workshop on Information Retrieval with Asian Languages, (2003).

[Vapnik (95)] Vapnik, V.N.: “The nature of statistical learning theory“; Springer, USA (1995).

[Verberne 07] Verberne, S., Boves, L., Coppen, P-A., Oostdijk, N.: “Discourse-based answering of why-questions”; Traitement Automatique des Langues, 47, 2 (2007).

[Verberne 08] Verberne, S., Boves, L., Oostdijk, N., Coppen, P-A.: “Using Syntactic Information for Improving Why-Question Answering“; Proc. COLING’08, the 22nd International Conference on Computational Linguistics, Manchester, UK (2008).

[Zhang 12] Zhang, Z., Webster, P., Uren, V., Varga, A., Ciravegna, F.: “Automatically Extracting Procedural Knowledge from Instructional Texts using Natural Language Processing“; Proc. LREC 12, Istanbul, Turkey (2012).

International Journal on Artificial Intelligence Tools
Vol. XX, No. X (2014) 1–20

World Scientific
www.worldscientific.com

THE INTEGRATION OF TEXT-BASED WHY QUESTION ANSWERING SYSTEM AND IMAGE PROCESSING FOR ROOT-CAUSE DIAGNOSIS

CHAVEEVAN PECHSIRI

Department of Information Technology, Dhurakij Pundit University,
Bangkok, Thailand
itdpu@hotmail.com

RAPEPUN PIRIYAKUL

Department of Computer Science, Ramkhamhaeng University ,
Bangkok, Thailand
rapepunnight @ yahoo.com

WORASIT CHOOCHAIWATTANA

Department of Information Technology, Dhurakij Pundit University,,
Bangkok, Thailand
worasit.cha@dpu.ac.th

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

This research aims to develop a Why Question Answering system integrated with image processing for providing the root-cause determination, especially in plant diseases. The image expression is applied on the Why question for providing the Why-question content (i.e. plant symptoms) that are difficult to be explained by text. There are several problems involved to this research including how to determine the Why-question type from the question-text portion, how to determine the Why-question content from the question-image portion, how to determine the Why-question focus, and how to determine the corresponding answer from the previous extracted causality knowledge as the knowledge source from technical documents. Therefore, we propose using two different techniques of Support Vector Machine and Maximum Entropy to identify the Why question, a Bag-of-Visual-Words to solve the Why-question content, and a causative verb concept /an effect verb concept to solve the Why-question focus. Then, we apply the "similarity" between the Why-question content of the conceptual predicate query as the question representation and the knowledge source. Finally, the research achieves the high correctness of answering at the first rank to 86.7%.

Keywords: ImageWhy-QA system, visual word, root-cause.

1. Introduction

Nosology studies and Disease diagnostics including the root-cause diagnosis often require a combination of a broad knowledge of diseases and symptoms' prevalence, and probabilistic concepts in their reasoning[1]. The compilation of experiences and the capacity to perform the root-cause determination including the cause and effect reasoning allows diagnosticians to recognize common disease states and perform efficient and

ethical diagnostic evaluations. However, some diagnosticians are often required to make decisions with the lack of information or knowledge. Therefore, it is necessary to have an automatic system that provides the reasoning knowledge to support their root-cause diagnosis through a Why Question Answering system (a Why-QA system). In respect to "A picture is worth a thousand words"[2], our research aims to develop an ImageWhy-QA system which is a Why-QA system integrated the Why-question text with an image for clearing an expression of a problem for the root-cause diagnosis to the amateur diagnostician or other people. The clear explanation of the questions of problems also results in acquiring the better answers. The ImageWhy-QA system is separated into two main parts, a ImageWhy-question part (which is a textual Why-question including an image) and a ImageWhy-answer part (which is a textual answer). Thus, the ImageWhy-question part consists of a question-text portion (containing a question word, e.g. 'Why', 'What', 'How', and etc.) and a question-image portion (providing its Why-question content of a problem, e.g. a disease symptom). In addition, the ImageWhy-QA system allows a user to post a problem in term of an ImageWhy question (see Fig. 1) on the online community web-board. Thus, it is a challenge to diagnose the root cause of the problems, e.g. plant disease symptoms, through the automatic root-cause identification by the proposed ImageWhy-QA system before approaching to solve these problems. The proposed ImageWhy-QA system are based on leaf-symptom images emphasized on the following symptom properties, lesion color, lesion shape, and leaf texture, which are typical symptoms of certain diseases.

(a) "*ทำไมใบข้าวมีอาการผิดปกติ*"

("*Why does a rice leaf have an unusual symptom*?")

(b) "*ข้าวเป็นโรคอะไร*"

("*What disease does the rice plant get*?")

Fig.1. Examples of ImageWhy questions

The ImageWhy-answer part is acquired from the extracted causality knowledge[3] from agricultural-technical documents downloaded from the agricultural department website. The causality is the relation between a causative event and an effect event, which can be represented by a causative verb concept set (V_c) and an effect verb concept set (V_e) respectively, (see Table 1). The extracted causality[3] is expressed in the form of an inter-causal EDU vector (a cause-effect vector) where EDU is defined by[4], as an elementary discourse unit or a simple sentence/clause. The cause-effect vector is a vector of the causal relation between one causative EDU and one/multiple effect EDU(s) as $\langle EDU_{cause}, EDU_{effect\text{-}1}, EDU_{effect\text{-}2},.., EDU_{effect\text{-}m}\rangle$. This cause-effect vector is kept in the repository as the knowledge source for answering to the ImageWhy question.

Example of a cause-effect-EDU vector:

Causative unit: EDU1 "*ถ้าราทำลายใบข้าว / If the fungus infest rice leaves,*"

Effect unit (EDU2+EDU3+EDU4) :
EDU2 "*จะทำให้ใบมีแผลจุดตาสีน้ำตาล / [it] will make the leaves have brown-spot lesions.*"
EDU3 "*ต่อมาแผลขยาย / Then the lesions expand.*"
EDU4 "*และ[แผล]รวมตัวกัน / and [the lesions] are combined.*"
(where a symbol [..] means ellipsis).

Table 1. Causative Verb Concept Set (V_c) and Effect Verb Concept Set (V_e) [3]

Verb Type	Surface form	V_c (Causative Verb Concept Set)
Causative Verb	*ดูด/suck, ดูดกิน/suck. กิน/eat, กัด/bite,*	consume/ destroy
	ทำลาย/destruct, กำจัด/eliminate, ฆ่า/kill, หัก/break, ระเบิด/explode, บุกรุก/infest	destroy
	เป็น+โรค/ be+ disease,	be disease/ get disease
	ได้รับ+เชื้อโรค/get+ pathogen,	get pathogen
	ติด+เชื้อ/contract	infect
	เกิด/occur, ปรากฏ/appear	appear
	……..	……
Effect Verb	Surface form	V_e (Effect Verb Concept Set)
	หงิก /shrink, งอ/bend, บิด/twist, โค้งงอ/curl	beAbnormalShape / beSymptom
	แห้ง/dry, ไหม้/blast	dry/ beSymptom,
	เหี่ยว/wilt	loseWater/ beSymptom
	แคระแกรน/stunt	stunt/ notGrow/ beSymptom
	ร่วง/drop off	comeOff/ beSymptom
	เหลือง/be yellow	beYellow/ beAbnormalColor/ beSymptom
	เป็น+จุด/be+spot, มี+จุด/have+spot (be/have a spot)	beMark , haveMark / beSymptom , haveSymptom
	เป็น+แผล+รูปตา/ be+lesion+eye-shape , มี+แผล+รูปตา/ have+lesion+eye-shape (be/have an eye-shape lesion)	beEyeShapeMark/ beSymptom , haveEyeShapeMark/ haveSymptom
	เป็น+สี/be+Color, มี+สี/have+Color (be/have Color) เป็น+แผล+สี/ be+lesion+Color (be/have a Color lesion) where Color ={'สีเหลือง/yellow' 'สีน้ำตาล/brown' 'สีส้ม/orange' 'สีเทา/grey' 'สีดำ/black'..}	beColor/ beAbnormalColor/ beSymptom , haveColor/ haveAbnormalColor/ haveSymptom , beColorMark/ beSymptom , haveColorMark/ haveSymptom
	ขยาย/expand, รวม/combine	increase
	……..	……

Moreover, the ImageWhy-QA system can be applied to a solution center or a service center for supporting the root–cause identification and also providing the causality knowledge to users. Although previous studies[5] indicated that there were about 5% of Why questions occurring in the Question Answering (QA) system, it is necessary for reasoning in diagnosis.

Previous literature on automatic Why-QA systems have involved several strategies including Information Retrieval, Information Extraction, Knowledge Extraction, Machine Learning, Image Processing, Natural Language Processing ,and Reasoning[6-10]. Yeh et al.[9] worked on the photo-based question answering system, especially the What and Where question types, where the information retrieval was applied for finding the possible answers from websites. Moreover, working on Why-QA that emphasizes on events is different from working on other wh-QA (such as who, what, and where) which

emphasizes on name entities or noun phrases[8]. Previous Why-QA works were based on reasoning [6, 8] and discourse structures[10] (see section 2).

There are two main parts of problems involved with this research; the ImageWhy question part and the ImageWhy answering part. The ImageWhy question part consists of three problems; first is how to identify a Why-question type from the question-text portion with the problem of the question word ambiguity (see section 3). Identifying the question-text expression without using the question symbol (i.e. '?'), commonly practiced in some languages as in our research is a challenge. Therefore, our research proposes using two different machine techniques, Support Vector Machine (SVM) and Maximum Entropy(ME), to identify the Why-question type with two feature sets, a question word set and a question verb set from the question-text portion (see section 3.1.1). Second is how to determine the Why-question content from the question-image portion. We also propose using a Bag-of-Visual-Words (BOW) to identify and represent the region of interest (ROI) on the image as the Why-question content where a visual word is a small patch on the array of pixels containing the interesting feature space of color, texture...etc.[11] and (http://en.wikipedia.org/wiki/VisualWord#cite _note-valu99-3). Their BOW[11] represents an image containing several patches or several visual words whereas the BOW in our research represents the ROI on an image where the ROI contains several visual words of lesion shape, lesion color, image background color, and image background texture (see section 3.1.2). We also apply the symptom-concept-frame structure to interpret the ROI's BOW to the conceptual predicate query. And, third is how to determine the ImageWhy question's focus. We apply v_c or v_e to determine the Why-question focus based on an event mostly expressed by a verb or a verb phrase (where $v_c \in V_c$, $v_e \in V_e$, V_c is a causative verb concept set, V_e is an effect verb concept set, and see Table1).

The second part of problems is the ImageWhy answering part as how to determine the corresponding Why answers from the knowledge source. We apply the similarity scores between the conceptual predicate query (contains the Why-question content) and the EDUs from the cause-effect vector of the knowledge source[3] after stop-word elimination, to solve the Why answer.

In section 2, related works are summarized. Problems of the ImageWhy-QA system is described in section 3. The system's architecture is described in section 4. We evaluate and discus our system in section 5 and give a conclusion in section 6.

2. Related Works

Other related works to address several techniques required for the ImageWhy-QA system have been involved with some of the following areas; Image Processing, Natural Language Processing, and knowledge reasoning.

In 2008, Yeh et al.[9] worked on the photo-based QA system based on five categories of images: books, movies, groceries, modern landmarks and classical landmarks, where a question is expressed by both a photo/image as an object and a caption or text. The question text is used to determine the scope of relevant images which are used for image

matching by indexing the images extracted from online multimedia resources. Then, the question text and the matched image are used for building the question template use for solving the answers based on the similarity scores. Their results have high recalls varying from 68% to 100% within the top five ranks and have at least one correct question. However, their QA system cannot be applied to our research because their images do not involve with a stage whereas each kind of a plant disease in our research has several symptom stages changing over time and variety. It is time consuming to find the root-cause by matching an image problem to all symptom stages for one disease.

Girju [7] 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 our research mostly based several effect-event explanations which express by verbs/verb phrases. Verberne et al.[10] proposed using RST (Rhetorical Structure Theory) structures to approach texted-based Why questions by matching the question topic with a nucleus in the RST tree while yielding the answer from the satellite. The author compared manual RST analysis with a system constructed using Perl script where the likelihood of the nucleus and the discourse relations are calculated. The RST approach to the Why-QA system achieved the answer correctness of 91.8% and a recall of 53.3%. Oh et al., 2013,[12] used intra- and inter-sentential causal relations between terms or clauses as evidence for answering Why-questions. They ranked their candidate answers (from documents retrieval Japanese web texts) with their ranking function including re-ranking the answer candidates done by a supervised classifier (SVM). Their Why-QA system achieves 83.2% precision. However, the only text-based question as in [7,10,12] could not explain the symptom problem as clear as an image.

Sivic et al.[11], their image layout was analogous to topic determination in text by using the bag of words or BOW. Thus, the visual words were applied to determine the image topic. Their visual words were formed by vector quantizing the local appearance descriptors of images. The probabilistic Latent Semantic Analysis (pLSA) of Hofmann using the bag of 'visual words' representation was applied to determine the object categories as the topics. Their results of the topic determination approach were successfully to identify the object categories for each image with the high reliabilities. However, our research applies the BOW to determine the ImageWhy-question content.

In 2011, Patil and Kuma[13] discussed the role of image processing in agricultural. They concluded that it can be used for detecting diseased plant, quantifying affected area and finding shape and color of affected area. Woodford et al.[14] proposed using wavelet transform technique and neural network to help identify pest damages in fruit. In additional, Ei-Helly et al.[15] proposed a novel approach to integrate image analysis technique into diagnostic expert system for plant diseases. However, the objective of their system is for plant disease classification only. Another interesting approach purposed by [16], they developed weather based prediction models of plant diseases using SVM. While [17] focused on how to grading of grape leaf disease by calculating the quotient of disease

spot and affected leaf area, Meunkaewjinda et al.[18] tried to classify grape leaf disease using self-organizing map and back propagation neural networks. Ying et al.[19] proposed a method of image pre-processing for crop diseases and also suggested effective characteristic parameters for the disease diagnoses. As mention above of image processing on the agriculture, most of research focused on detection of the plant diseases using both image processing techniques and machine learning techniques. None of them, however, exploited the usage of plant disease detection and classification to find a root cause of the disease through the ImageWhy-QA system.

3. Research problems

The research contains two main parts of problems: the ImageWhy question part and the ImageWhy answering part .

3.1. *ImageWhy question part*

There are three major problems as how to identify a Why-question type (from the question-text portion with the question word ambiguity), how to determine the Why-question content (from the question-image portion), and how to determine the ImageWhy question's focus.

3.1.1. *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 is a reasoning question, a '*อะไร/What*' function is asking for information about something (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. For example:

EDU$_1$ "*ช่วงแตกกอใบข้าวหงิกงอ/ In the tillering stage, rice leaves shrink.*"

EDU$_2$ " *เป็นเพราะอะไร/**What** are the reasons*?"

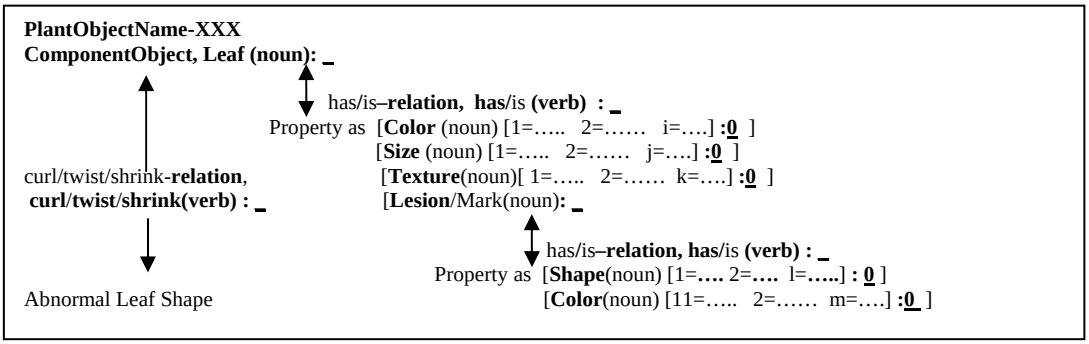
Therefore we propose using different machine learning as ME and SVM to classify a Why-question type. The features used in this classification consist of a question word set (QW) and a question verb set (QV) where $qv \in$QV and qv exists in the EDU having a question word (qw, $qw \in$QW).

QW={'*ทำไม/Why*' '*อย่างไร/How*' '*อะไร/What*' '*ใคร/Who*' '*ที่ไหน/Where*' '*เมื่อไร/When*' …}.

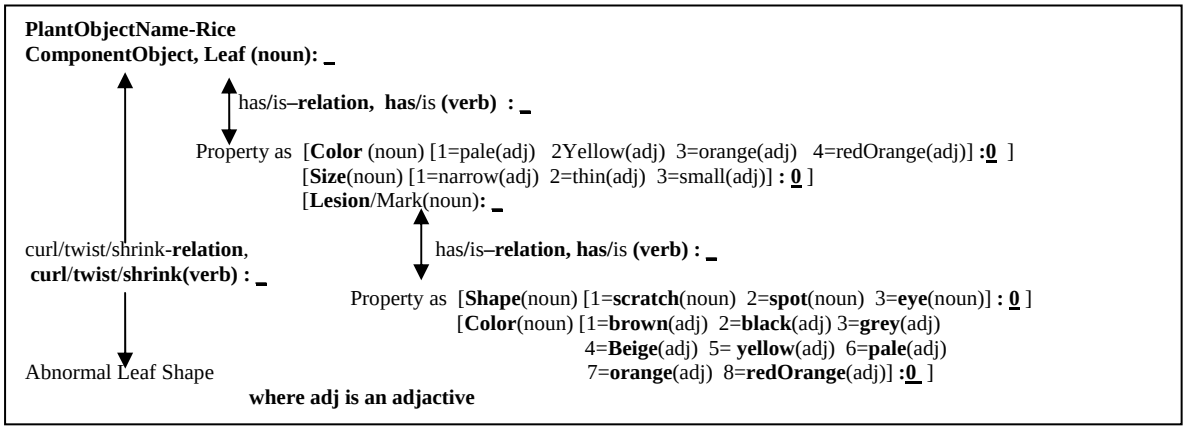
3.1.2. *How to determine Why-question content*

Most of the question-text portions of the ImageWhy questions always express in general concepts of problems as in Fig. 1(a) lacking of the symptom-problem content. Therefore, we propose using the BOW to identify and represent a ROI on the question-image portion followed by the symptom-concept-frame structure (see Fig. 2) to interpret the BOW to the conceptual predicate query having the Why-question content. Fig. 2(a) shows a general symptom-concept-frame structure of leaf symptoms which consist of properties

and relations (where a property is expressed by a noun phrase and a relation is expressed by a verb/a verb phrase). Fig. 2(b) is an example of the symptom-concept-frame structure of the rice leaf symptoms, contains three main symptom features (Leaf Color, Leaf Shape, and Leaf Lesion/Mark) with the default "zero" value or null.



(a) A general symptom-concept-frame structure of leaf symptoms



(b) Example of a symptom-concept-frame structure of rice leaf symptoms

Fig.2. Show a symptom-concept-frame structure

3.1.3. *How to determine the ImageWhy question's focus*

The determination of the ImageWhy question's focus is necessary for the answer determination. The focus of the ImageWhy question for the root cause determination is always expressed on the question-image portion by the relation expression of the symptom concept frame structure. Where the relation is expressed by the effect event represented by V_e from Table1. Moreover the effect event can also express on the question-text portion if the question verb, *qv*, on the question-text EDU is v_e .

.

3.2. *ImageWhy answering part*

The problem of this part is how to determine the corresponding answer to the Why-question content of the ImageWhy question. However, it is unlike wh-questions from text-based questions, the answer of the ImageWhy question can not be determined by the question word (*qw*). For example:

a) Q : Who is the president of USA? Ans: Obama is the president of USA.

b) Q: "*ทำไม /Why ใบมะม่วง/mango leaves หงิก /shrink*" (Why do mango leaves shrink?)

Ans: EDU1 "*เพลี้ย/Aphids ทำลาย/destroy ใบมะม่วง /mango leaves*" (Aphids destroy mango leaves.)

EDU2 "*ทำให้/make ใบ/leaves หงิก/shrink*" ([it] makes leaves shrink.)

The answer of the question in a) can be determined by a question word "Who"[20] whereas the question word "Why" cannot be applied to determine the answer in b). Moreover, wh-questions have previously been approached by determining answers from noun phrases and question words [8], which is suitable for the Why question with the answer based on the lexico-syntactic pattern [7] 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 the ImageWhy-QA system mostly based on several effect-event explanations which are always expressed by verbs/verb phrases. And, it is not suitable for other non-factoid questions either as portrayed by[8, 21, 22, 23]. Therefore, we use the similarity scores between the Why-question content and EDU_{effect} from the cause-effect vector to determine the root-cause answer of the ImageWhy question. Where all word concepts are referred to WordNet (http://wordnet.princeton.edu/) and the predefined plant disease information from the Department of Agriculture (http://www.doa.go.th/) including Encyclopedia (http://kanchanapisek.or.th/kp6/New/) after using the Thai-to-English dictionary (Longdo.com)

4. System architecture

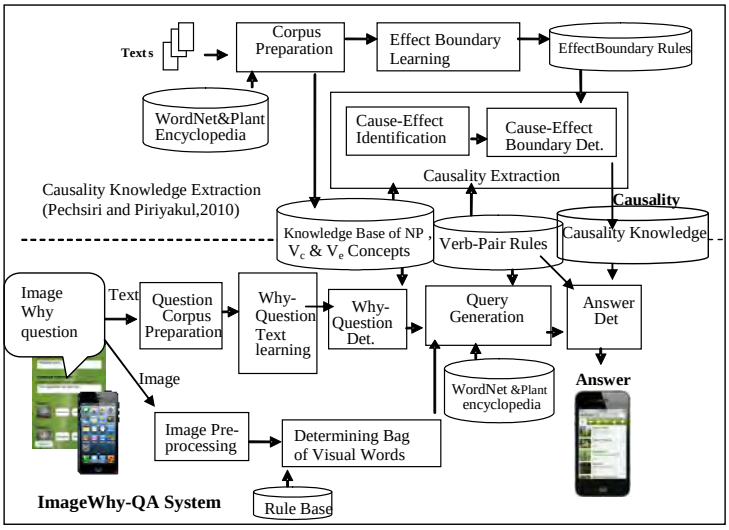


Fig.3. System architecture

There are two layers in our System Architecture (see Fig. 3), the first layer is the Causality Knowledge Extraction system developed by the previous research[3] as the knowledge source for the root-cause determination through the ImageWhy question. The second layer is the ImageWhy-QA system which can be separated into two parts: the ImageWhy question part and the ImageWhy answering part.

4.1. *ImageWhy-Question Part*

There are two areas of processing of the ImageWhy-question part, the text processing for the question-text portion and the image processing for the question-image portion. Thus, there are several steps involved to the ImageWhy-question part: Textual-Question-Corpus Preparation, Why-Question Type Determination, Image Pre-processing, BOW determination, and Query Generation.

4.1.1. *Preparation of Question-Text Corpus*

All 560 questions having question-text portions and question-image portions are collected from several QA sites and web-boards, e.g. http://www.gotoknow.org /blogs/books/view/ agriculture , with several question types, i.e. 'Why', 'How', 'What', 'Where', and etc., of the plant-disease domains. The research emphasizes only the leaf symptoms of plant diseases on the following plants, rice, mango, and orange. The collected ImageWhy questions are separated into two parts, one part of 330 questions for learning with ten folds cross validation, and the other part of 230 questions (contain 90 Why questions which consists of 30 Why questions of each plant, rice, mango, and orange) for testing. The learning part of all question-text portions have to be prepared by using a Thai word segmentation tool[24] with the part of speech annotation including Name Entity[25] followed by EDU segmentation[26]. Then, we semi-automatically annotate textual data for learning the Why question type with the following tags: a *Why-question type* tag (Why-Q-Type), a *question word* tag (Qw), a *question-verb tag* (Qv, which consists of three types of verb concept set, a Why-question-cue-verb set (V_{cue}), an effect verb conceptset(V_e), and a causative verb concept set(V_c), and a *Why-question focus* tag (Why-Q-Focus), where all concepts are referred to WordNet (http://wordnet.princeton.edu/) and Thai Encyclopedia (http://kanchanapisek.or.th/kp6/New/) after using the Thai-English dictionary (www.longdo.com) (see Fig. 4). The learning part of the question-image portion is prepared in the image pre-processing step (in section 4.1.3) for the color, texture, and shape classification of symptoms.

Question: "ทำไมยอดใบหงิกงอ /Why do top leaves shrink?"
<Why-Q-type class=yes>
<EDU>[<*Qw type=why*>ทำไม(**Why**)/pint <*/Qw*> ยอดใบ(**top leaves**)/ncn]/NP
<*Why-Q-Focus*><*Qv type=Ve*>[หงิก(**shrink**)/vi]<*/Qv*>/VP<*/Why-Q-Focus*> </EDU></Why-Q-type >
Question: "อะไรเป็นสาเหตุทำให้ต้นข้าวแคระแกรน/ What is the cause making rice plants stunt?"
<Why-Q-type class=yes>
<EDU>[<*Qw type=what*>อะไร(**What**)/pint<*/Qw*><*Qv type=Vcue*> เป็น(**is**)/vcs สาเหตุ(**the cause**)/ncn<*/Qv*>]/VP</EDU>
<EDU>[ทำให้(**make**)/vcau [ต้นข้าว(**rice plants**)/ncn]/NP <*Why-Q-Focus*>[[แคระแกรน(**stunt**)/vi]/VP<*/Why-Q-Focus*>]/VP
</EDU></Why-Q type >

Fig.4. Examples of Why-Question Annotation from all question-text portions

4.1.2. *Why-Question Type Learning and Why-Question Type Determination*

The research applies two different techniques of machine learning, ME and SVM, to learn the Why-question type from the learning corpus with two feature sets, a question word set (QW) (see section 3.1.1) and a question verb set (QV) gained by the Qw tag and the Qv tag respectively from the annotated corpus (Fig. 4)

$QV = V_{cue} \cup V_e \cup V_c$ (see Table1 for V_c and V_e)

V_{cue} = {'*เป็นโรค../get a disease..*' '*เป็นเพราะ.../be the reason...*' '*เป็นสาเหตุ.../be the cause..*'
'*ทำให้../be the cause..*' '*เป็นผลจาก.../be the result from..*'….}

Maximum Entropy (ME) According to Eq. (1)[27,28], ME can be used as the classifier of the r class when p(r|x) is the highest probability to determine two question-type classes, Why-Q and non-Why-Q. Where r is the question-type classes (the question type is Why-Q when r=0, otherwise r=1) and x is the binary vector of the question-word concept features (*qw* where $qw \in QW$) and the question verb concept features (*qv* where $qv \in QV$) from the annotated corpus, as shown in Eq. (1).

$$p(r \mid x) = \arg\max_{r} \frac{1}{z} \exp\Big(\sum_{j=1}^{n} \lambda_j f_{yes,qw,j}(r, qw) + \sum_{j=1}^{n} \lambda_j f_{no,qw,j}(r, qw) + \sum_{j=1}^{n} \lambda_j f_{yes,qv,j}(r, qv) + \sum_{j=1}^{n} \lambda_j f_{no,qv,j}(r, qv) \Big) \quad (1)$$

Where z is a normalization constant. Then, we use λ_j (the weight for a given feature function of the binary vector) resulted from learning the Why-question type to determine the question-type classes by Eq. (1).

Support Vector Machine The linear binary classifier, SVM, applies in this research to classify the question types with Why-Q or non-Why-Q from the annotated corpus by using Weka (http://www.cs.waikato.ac.nz/ml/weka/). According to[29] this linear function, f(x), of the input $x = (x_1 x_2 \ldots x_n)$ assigned to the positive class (Why-Q) if $f(x) \geq 0$, and otherwise to the negative class (non-Why-Q) if $f(x) < 0$, can be written as:

$$f(x) = \langle w \cdot x \rangle = b = \sum_{i=1}^{n} w_i x_i + b \quad (2)$$

where x is a dichotomous vector number, w is weight vector, b is bias, and $(w,b) \in R^n \times R$ are the parameters that control the function. The SVM learning results are w_i and b for the input x which consists of the question word vector from QW and the question verb vector from QV. Therefore, the question-type classes can be identified from Eq. (2) with the input x and the learning results (w_i and b).

4.1.3. *Image Pre-processing*

The 560 plant disease images, especially the leaf symptoms on rice, mango, and orange, are collected from the question-image portions (from section 4.1.1). Image enhancement is constructed from low pass and high pass filter for adjusting intensities of the images in order to highlight areas considered. After this pre-processing step, each image is ready for segmentation. The segmentation process is to differentiate between background and target object (which is the region showing the current symptoms of the disease) to eliminate the back ground away from the leaf area having the disease symptom. Then, the target object is used for the next step of BOW Determination.

4.1.4. *BOW Determination*

The BOW determination step is to generate BOW by collecting the relevant visual words from the target object. Each visual word is generated by using region growing algorithm[30]. According to BOW in our research as in Fig.5, the only relevance visual words to plant symptoms are selected to represent our research's ROI (Region of Interest which is the image salience) by detecting the symptom features e.g. lesion color (as ROI object color), lesion shape (as ROI object shape), leaf texture (as ROI area color), and leaf color. Moreover, ROI equivalents to a noun phrase/a verb phrase after interpretation of an image to text. Hence, ROI is the content/the salient content of a Why-question image that has to be filled in the symptom-concept-frame structure for the image interpretation.

(a) Rice Blast Disease from Fig. 1(b) (b) Mango Anthracnose (b) Orange Star Melanose

Fig.5. Show all patches of visual words in the BOW from the images of leaf symptoms

The ROI object shape is determined by using shape contexts, i.e. an eye shape, a scratch shape, a spot shape, and a star shape, where the reference point captures the distribution of the remaining points relative to it [30]. Then, the corresponding points on two similar shapes have similar shape contexts. After the missed shape contexts have been filtered out, the color detection is determined. There are two areas of color detection, a ROI area for texture detection and a ROI object. To detect the color and the texture, we apply the following two classification levels based on 560 sample images of the question-image portions (from section 4.1.3) where these sample images are supervised data and consist of 330 sample images for learning based on ten folds cross validation and 230 sample images for testing.

First Classification Level

The objective of this level is to filter out the normal properties of the color and the texture of the ROI by learning of the binary classifier as the logistic regression model [31]. The logistic regression model as shown in Eq. (3) is applied to classify both color and texture properties with two classes of Normal and Abnormal where ROI pixels are based on the HSI color model, for hue (H), saturation (S), and intensity (I). The twelve features (Feature Vector) as Min, Max, Mean, and Entropy of H, S, and I are used in the binary classification. The Entropy feature as shown in Eq. (4)[32] is applied in this research for determining local spatial variations of color intensity which express the textural property of an image as the roughness. And, the roughness texture property in our research expresses the leaf shape symptom of the curl/twist/shrink occurrence.

$$\text{LogisticRegression}: \quad \gamma = \frac{\exp^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}{1+\exp^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}} \tag{3}$$

where γ = the probability of a case which is in a particular category: AbnormalClass

exp = the base of natural logarithms

FeatureVector $X = x_1, x_2, .., x_n$

α = the constant of the equation

β = the coefficient of the predictor variables

$$\text{Entropy of P}: \quad H(P) = \sum_{i=1}^{n} p_i * \log(1/p_i) \tag{4}$$

where P is a set of a probability distribution of information as all features of ROI $P = \{p_1, p_2, .., p_n\}$

Second Classification Level

The results of the Abnormal class samples from the first classification level are used in this second classification by learning of a multi-class-classifier as Multi-Layer Perceptrons (MLPs)[33] for detecting the color symptom and the texture symptom of the ROI object (which emphasizes on the disease lesion) and the ROI background (which is the leaf containing the disease lesion) respectively. There are twelve input features used in the MLPs classifier as Min, Max, Mean, and Entropy of H, S, and I. The MLPs classifier has eight classes (Brown, Black, Grey, Beige, Yellow, Pale, Orange, Red-Orange) of irregular color occurrences on the ROI object or the ROI background.

Multi-Layer Perceptrons (MLPs) According to[33] , 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. (5).

$$y_i = f(\sum x_i w_i) \tag{5}$$

where w_i represents the weight, x_i is the input feature of the ROI, $f(\cdot)$ is the activation function, and y_i is the output of the i^{th} node. A sigmoid function is often used as the activation function. MLP consists of successive layers, each of which includes a different number of processing nodes.

$$X = \sum_{i=1}^{n} x_i w_i - \theta \tag{6}$$

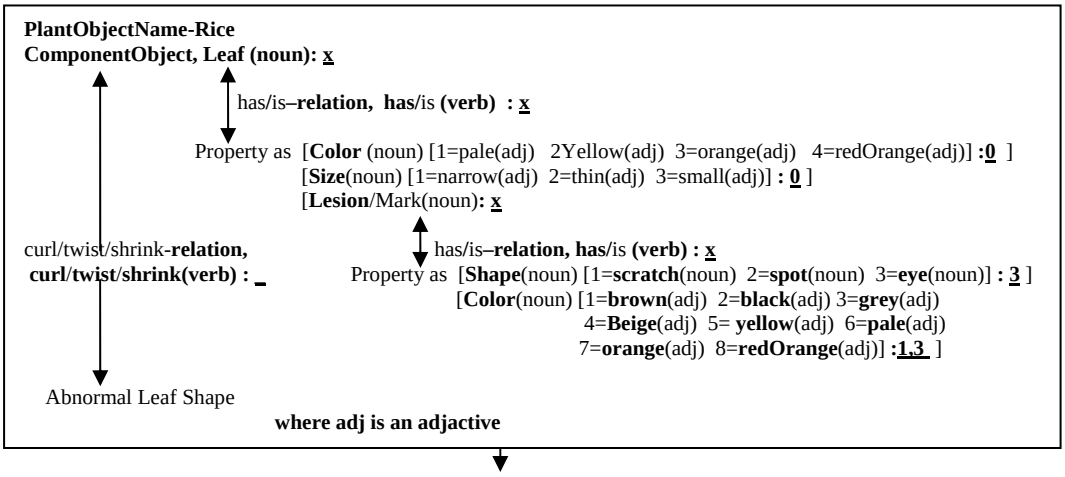
The nodes in the first layer receive inputs from the outside world and are fully connected to nodes in the hidden layer where each node in the hidden layer is connected to the output layer nodes to produce the output pattern or the output class of the MLP. Furthermore, the net weighted input can be solved by Eq. (7) which contains the activation function.

$$y_j(p) = \sum_{i=1}^{n} x_i(p) w_{ij}(p) - \theta_j \tag{7}$$

where n is the number of neuron inputs, and θ_j is the threshold value of neuron at the j^{th} node in the hidden layer.

4.1.5. *Query Generation*

This step is to generate the conceptual predicate query of the question-image portion by using a symptom-concept-frame structure shown in Fig. 2 to interpret the BOW. Therefore, the BOW from Fig. 5(a) can be interpreted as the Why-question content of the conceptual predicate query (which contains a content word set generated by its symptom-concept frame) as follow.



•Why-question content: hasEyeShape_mark(leaf) ∧ be_brown_and_grey_color(eyeShaped_mark)

• Set of content words : leaf(noun), has/is(verb), lesion/Mark(noun), has/is(verb),

shape(noun), eye(noun), color(noun), brown(adj), grey(adj)

4.1.6. *Question-Focus Determination*

The question-focus is necessary for pointing to what the answer is. The question-focus of the ImageWhy question, especially for the root-cause diagnosis, is based on the effect event expressed by V_e. The ImageWhy question 's focus is expressed on the question-image portion and expressed by the relation expression of V_e in the symptom-concept-frame structure. And the ImageWhy question's focus can also be expressed on the question-text portion if *qv* is v_e and $v_e \in V_e$.

For examples:

1) The ImageWhy question of a rice leaf (Fig.1(b))

"ข้าว/*rice* เป็นโรค/*get disease* อะไร/what"

("What *disease* does the rice plant *get*?")

- According to the symptom-concept-frame structure (see section 4.1.5), the question–image portion has the following relations: *has_Lesion* relation, *has_eyeShape* relation, and *is_brown-greyColor* relation.

-The question–text portion contains: *qw* is "What" and *vq* is "*get disease* "

The ImageWhy question's focus:

- Leaf has Lesion (a) Lesion has eyeShape (b) Lesion is brown-greyColor. (c)

a),b), and c) are one symptom relation expression as a Why-question Focus whilst all word in a),b), and c) are concept words of the content word set.

2) The ImageWhy question of orange leaves

"ทำไม/why ใบส้ม/*orange leaves* หงิก/*shrink*"

("Why *do orange leaves shrink*?")

- According to the symptom-concept-frame structure, the question–image portion has the following relations: *shrink relation*

-The question–text portion contains: *qw* is "Why" and *vq* is "*shrink* " (v_e)

The ImageWhy question's focus:

- Leaf shrink : The shrink relation is a symptom relation expression as the question-image focus and "shrink" is also a concept words of the content word set.

4.2. ***ImageWhy answering part***

After both the Why-question type determination and the Why-question content determination, the correct ImageWhy questions are used for the answer determination from the knowledge source which contains cause-effect vectors of plant diseases. The answer is solved by determining the similarity score [34] in Eq. (8) between a set of content words existing in the Why-question content and each EDU element of the cause-effect vectors after eliminating stop words.

$$Similarity\_Score = \frac{|S1 \cap S2|}{\sqrt{|S1| \times |S2|}} \quad (8)$$

where

S1 is all word concepts from the set of content words existing in the Why-question content.

S2 is all word concepts from set of words from $EDU_{effect\text{-}i}$ after eliminating stop words (through stemming words for some languages) where $EDU_{effect\text{-}i}$ exists in the cause-effect vector $\langle EDU_{cause}, EDU_{effect\text{-}1}, EDU_{effect\text{-}2},.., EDU_{effect\text{-}m}\rangle$.

The all word concepts of S1 and S2 are based on WordNet and Thai Encyclopedia after using the Thai-to-English dictionary. For example (Fig.1(b)):

S1 : Set of content words

{ leaf(*noun*), has/is(*verb*), lesion/Mark(*noun*), has/is(*verb*), shape(*noun*), eye(*noun*), color(*noun*), brown(*adj*), grey(*adj*)}

= **{** leaf, have/be, lesion/mark, shape, eye, color, brown, grey}

Knowledge Source:

Cause-Effect Vector ID=1 DiseaseName: Rice Blast disease

EDU_{cause} : "*เชื้อราไพรีคิวลาเรีย*/***Pyricularia fungus*** *ทำลาย*/***destroy*** *ต้นข้าว*/***rice plant*** "
(The Pyricularia fungus destroy the rice plant.)

$EDU_{effect1}$: "*ระยะ*/***period*** *กล้า*/***seedling*** *ใบ*/***leaf*** *มี*/***have*** *แผล*/***lesion*** *รูป*/***shape*** *ตา*/***eye*** *สี*/***color*** *น้ำตาล*/***brown***"
(Seedling Period: Leaves have the brown eye shape lesions.)

$EDU_{effect2}$: "*แผล*/***lesion*** *ขยายลุกลาม*/***spread over*** *ทั่ว*/***whole*** *ใบ*/***leaf*** "
(The lesions spread over the whole leaf.)

$EDU_{effect3}$: ……..

S2: Cause-Effect Vector ID=1

$EDU_{effect1}$: {seedling, period, leaf, have, lesion, shape, eye, brown, color} →
Similarity_Score =0.8

$EDU_{effect2}$: {lesion, spread, whole, leaf} → Similarity_Score =0.4

……..

The candidate answers can be selected from all Cause-Effect Vector IDs which have S2 of $EDU_{effect\text{-}i}$ being similar to S1 of the question-image portion with Similarity_Score >0.5. Then, the candidate answers can be ranking according to Similarity_Score of the selected Cause-Effect Vector IDs. We select only the top five ranks of Similarity_Score as the possible answers where the first rank is the highest correct answer.

5. Evaluation and Discussion

Evaluation is achieved by using 230 questions (containing 30 ImageWhy questions of each plant, rice , mango, and orange, based on leaf symptoms) downloaded from several QA sites and the community web-boards, and is conducted on the Why-Question-Type Determination, Why-Question Content Determination, and ImageWhy Answer Determination.

5.1. *Why-Question-Type Determination*

The evaluation of the Why-question classification in this research is expressed in terms of the precision and recall based on human judgments (two experts and one linguistic) with max win voting. Results (Table2) demonstrate that feature dependency occurrences between the question word features and the question verb features allows ME to attain the highest precision of 97.2%. Moreover, 79 correct Why questions are achieved as the Recall result of Why-question type determination by ME from the question-text portions.

Table 2. Why Question Classification from the question-text portions

230 Questions of Plant Diseases based on Leaf Symptoms	Why-Question Type Determination			
	SVM		ME	
	Precision	Recall	Precision	Recall
Question-Text Portions	96.3%	85.8%	97.2%	87.5%

5.2. *Why-Question Content Determination*

With respect to the question content determination, by using the BOW of the ROI 's question-image portions of the 230 ImageWhy questions, the authors evaluate the question content determination by calculating the correctness of the visual word determination from the ten folds cross validation based supervised training data. The question content determination includes the lesion color (the ROI object color), lesion shape (the ROI object shape), leaf texture (the ROI area color), and leaf color. Then, the visual word determination consists of the ROI-object-shape determination, the ROI-object-color determination, and the ROI-area-color determination which is applied for the ROI texture determination. The correctness of the ROI object shape determination is approximately 88 % and is perspective dependent. After filtering out the incorrect ROI object shape, the correctness of the ROI color determination is 94% of the binary classification whilst the precision and the recall of the multi-class classification is 0.766 and 0.768 respectively (see Table 3). There are 190 correct images after the binary classification.

With respect to the multi-class classification, there are 145 questions having the correctness of determining the visual words without the question type consideration. The errors of the multi-class classification are caused by the incorrect patch generation which results in incorrectly determining the visual words. However, the total correctness of the ImageWhy question determination (after determining the visual words (the question content) and the Why-question type) is 76 questions or 84.5% correctness from 90 ImageWhy questions of the testing question corpus.

Table 3. Evaluation of the ROI color determination of two classification levels from the question-image portions

Binary Classification					
Class	True Positive	False Positive	False Negative	True Negative	% correctness
Abnormal	**178**	11	-	-	**94**
Normal	-	-	2	12	85

Multiclass Classification					
Classifier	True Positive Rate	False Positive Rate	Precision	Recall	F-Measure
MLP	0.768	0.087	**0.766**	**0.768**	**0.765**

5.3. *Corresponding ImageWhy Answer Determination*

The evaluation of the ImageWhy answer determination is based on three experts with max win voting. The number of correct ImageWhy questions from ImageWhy-Question Part (section 4.1) is 76 questions (consisting of 28 rice leaf symptoms questions, 25 mango leaf symptoms questions, and 23 orange leaf symptoms questions, see section 5.2). Then, the correct ImageWhy questions are used to evaluate the Why answer determination of the root-cause identification through the ImageWhy-QA system as shown in Table 4

Table 4. Evaluation of the ImageWhy answer determination for the root-cause identification

30 ImageWhy questions of each plants based on leaf symptoms	Number of correct answers at the first rank	%correctness of answers at the first rank
Rice	26	86.7%
Mango	23	76.6%
Orange	22	73.3%

Table 4 shows that the ImageWhy-QA system can highly achieve the root-cause identification of the rice leaf symptoms with 86.7% correct answers at the first rank whereas the orange leaf symptoms have the low percent correctness of the answers. However, the image perspective that mainly causes two different lesion shapes having the same shape (as in an orange Star melanose, and an orange Melanose) has led to 73.3% correct answers for the orange leaf symptom.

6. Conclusion

This research approaches to integrate a text-based QA system with an image to enhance the ability in identifying the root-cause problems, especially the plant diseases and hence, we have proposed an ImageWhy-QA system. The previous research [9] is based on IR-based QA which retrieves the answers from webs by matching images within the scope of the question text. Whilst our proposed ImageWhy QA system is based on two processing techniques including machine learning and the knowledge source, Text Processing to determine the Why-question type and Image Processing to determine the Why-question

content by BOW and the symptom-concept-frame structure. Then, the answer is solved by determining Similarity_Score between set of content words from the Why-question content and set of words from $EDU_{effect\text{-}i}$ of the knowledge source. The possible correct answers of [9] vary from 68% to 100% at the top five ranks. However, the proposed ImageWhy QA system can achieve 86.7% as the highest correct answers at the first rank. The results of our research can be improved if the image perspective has been solved in the next research and the adaptive concept-frame structure should also be considered to enhance the answer correctness. Finally, the ImageWhy QA system can be applied not only to identify the root-cause of plant symptoms but also to perform the preliminary diagnosis in other problem cases discussed in online community websites involving health care, vehicle usage and maintenance, utilities, and among others.

7. Acknowledgements

This work has been supported by the Thai Research Fund grant.

References:

1. R.A. Miller, Medical Diagnostic Decision Support Systems – Past, Present, and Future: A threaded Bibliography and Brief Commentary, *Journal of the American Medical Informatics Association*, Volume 1 Number 1 (1994), pp. 8 – 27.
2. J. H Larkin.and H.A. Simon, Why a Diagram is (Sometimes) Worth Ten Thousand Words, *Cognitive Science*, 11 (1987), pp 65-99
3. C., Pechsiri and R. Piriyakul, Explanation Knowledge Graph Construction through Causality Extraction from Texts. Journal of Computer Science and Technology, 25(5, 2010),pp 1055-1070.
4. L.Carlson,, D. Marcu, , M. E Okurowski, Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue* (2003). pp.85-112
5. E. Hovy, U. Hermjakob, and D.Ravichandran, A Question/Answer Typology with Surface Text Patterns, In *Proc. of the Human Language Technology conference* (San Diego, California, 2002).
6. D.T. Burhans and S.C. Shapiro, Abduction and Question Answering, In Proc. *of the IJCAI Workshop on Abductive Reasoning* (Seattle, Washington, 2001)
7. R.Girju, Automatic detection of causal relations for question answering, In *The 41st annual meeting of the assoc. for computational linguistics, workshop on multilingual summarization and question answering-Machine learning and beyond* (Japan, 2003)
8. S.Verberne , Developing an approach for why-question answering, In Proc. Of *Conference of the European Chapter of the Assoc. for Computational Linguistics* (2006)
9. T.Yeh, J.J. Lee, and T.Darrell, Photo-based Question Answering, In *MM'08* (Vancouver, British Columbia, Canada, 2008)
10. S. Verberne, L. Boves, P-A. Coppen, and N. Oostdijk, Discourse-based answering of why-questions, Traitement Automatique des Langues 47(2, 2007)
11. J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, Discovering Objects and Their Location in Images, *Robotics Institute*. Paper 286 (2005). Retrieved from http://repository.cmu.edu/robotics/286 on 2013-03-04.
12. O. Jong-Hoon, K. Torisawa, C. Hashimoto, M. Sano, S. D. Saeger, and K. Ohtake, Why-Question Answering using Intra- and Inter-Sentential Causal Relations, In Proc. of *the 51st Annual Meeting of the Association for Computational Linguistics* (Sofia, Bulgaria, 2013), pp 1733–1743.

13. J. Patil, and R. Kuma, Advances in Image Processing for Detection of Plant Diseases, *Journal of Advance Bioinformatics Application and Research*, Vol.2 (2, 2011), pp 135-141
14. B. Woodford, N. Kasabov, and C. Wearing, Fruit Image Analysis Using Wavelets, In Proc. Of *ICONIP/ANZIIS/ANNES* (1999).
15. M. Ei-Helly, A. Rafea, S. Ei-Gamal, and R. Ei-Whab, Integrating Diagnostic Expert System with Image Processing via Loosely Coupled Technique. *Central Laboratory for Agricultural Expert System*, (2004)
16. R. Kaundal, A. Kapoor, and G. Raghava, Machine Learning Techniques in Disease Forecasting: A Case Study on Rice Blast Prediction, *BMC Bioinformatics* (2006)
17. S. Weizheng, W. Yachun, C. Zhanliang, and W. Hongda, Grading Method of Leaf Spot Disease Based on Image Processing, In Proc. Of *International Conference on Computer Science and Software Engineering* (Cambridge, 2008)
18. A. Meunkaewjinda, P. Kumsawat, K. Attakitmongcol and A. Srikaew, Grape Leaf disease Detection from Color Imagery System Using Hybrid Intelligent System, In Proc. *Of IEEE ECTI-CON* (2008), pp 513-516.
19. G. Ying, L. Miao, Y. Yuan, and H. Zelin, A Study on the Method of Image Pre-Processing for Recognition of Crop Diseases, In Proc of *IEEE International Conference on Advances Computer Control* (2008)
20. E. Agichtein, S. Cucerzan, and E. Brill, Analysis of Factoid Questions for Effective Relation Extraction, In *ACM SIGIR International Conference on Research and Development in Information Retrieval* (Salvador, Brazil, 2005)
21. S. Verberne, L. Boves, N. Oostdijk, and P-A. Coppen, Using Syntactic Information for Improving Why-Question Answering, In Proc. of *the 22nd International Conference on Computational Linguistics* (Manchester, UK, 2008).
22. C. Pechsiri and R. Piriyakul, Developing the UCKG-Why-QA System, In Proc of *7th International Conference on Computing and Convergence Technology* (Korea, 2012)
23. S. Quarteroni and P. Saint-Dizier, Addressing How-to Questions using a Spoken Dialogue System: a Viable Approach?, In Proc. of *the 2009 Workshop on Knowledge and Reasoning for Answering Questions* (Suntec, Singapore, 2009).
24. S. Sudprasert and A. Kawtrakul, Thai Word Segmentation based on Global and Local Unsupervised Learning, In Proc for *NCSEC'03* (Chonburi, Thailand,2003).
25. H. Chanlekha and A. Kawtrakul, Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information, In Proc of *IJCNLP' 04* (Hainan Island, China, 2004).
26. J. Chareonsuk, T. Sukvakree, and A. Kawtrakul, Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information, In Proc of *NCSEC* (Thailand, 2005)
27. A. L. Berger, S. A. Della Pietra, and V J. Della Pietra, A maximum entropy approach to natural language processing, *Computer Linguist*. 22 (1, 1996), pp.39–71.
28. M. Fleischman, N. Kwon, and E. Hovy, Maximum entropy models for Frame Net classification, In Proc. of *the 2003 conference on Empirical methods in natural language processing* (Sapporo, Japan, 2003), pp.49-56.
29. V. N. Vapnik, *The nature of statistical learning theory*, (Springer-Verlag New York, Inc., 1995).
30. S.G. Stanciu, Digital Image Processing, (InTech, Croatia, 2012)
31. A.Y. Ng and M.I. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, In Proc of *Neural Information Processing Systems* (2002).
32. C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, (University of Illinois Press, USA, 1949)
33. S. Haykin, *Neural networks: a comprehensive foundation*. 2nd ed. (Prentice Hall, USA, 1999)

34. S. Biggins, S. Mohammed, and S. Oakley, Two Approaches to Semantic Text Similarity, In Proc. Of the *First Joint Conference on Lexical and Computational Semantics*, (Montre´al, Canada, 2012**),** pp 655–661.