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