รายงานผลการวิจัย

เรื่อง

การสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทยเพื่อสนับสนุนการตอบคำถามอัตโนมัติ

KNOWLEDGE EXTRACTION OF MEDICINAL PROPERTIES OF THAI HERBS FROM THAI TEXTS FOR SUPPORTING AUTOMATIC QUESTION-ANSWERING SYSTEM

โดย

ผู้ช่วยศาสตราจารย์ ดร.ณัฐวรรณ เพ็ชรศิริ

รายงานการวิจัยนี้ได้รับทุนอุดหนุนจากมหาวิทยาลัยธุรกิจบัณฑิตย์

พ.ศ. 2553

รายงานผลการวิจัย

เรื่อง

การสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทยเพื่อสนับสนุนการตอบคำถามอัตโนมัติ

Knowledge Extraction of Medicinal Properties of Thai Herbs from Thai Texts for Supporting Automatic Question-Answering System

โดย

ผู้ช่วยศาสตรจารย์ ดร. ฉวีวรรณ เพ็ชรศิริ

รายงานการวิจัยนี้ได้รับทุนอุดหนุนจากมหาวิทยาลัยธุรกิจบัณฑิตย์

พ.ศ. 2553

กิตติกรรมประกาศ

ขอขอบพระคุณ รองศาสตราจารย์ ดร. ระพีพรรณ พิริยะกุล ที่กรุณาสละเวลา ให้ความรู้และคำแนะนำเกี่ยวกับการทำระบบถามตอบอัตโนมัติ

ขอขอบพระคุณ รองอธิการบดีฝ่ายวิจัยและวิทยาบริการ มหาวิทยาลัยธุรกิจบัณฑิตย์ที่ให้โอกาสข้าพเจ้าและอาจารย์ สำราญ ในการศึกษาค้นคว้าวิจัยเรื่อง “การสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทยเพื่อสนับสนุนระบบการตอบคำถามอัตโนมัติ” จนสำเร็จ

ขอขอบพระคุณ คณบดีคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์ ที่ช่วยตรวจสอบการใช้ภาษาไทยให้เหมาะสม

ขอขอบพระคุณ มหาวิทยาลัยธุรกิจบัณฑิตย์ ที่ให้เงินทุนสำหรับสนับสนุนโครงงานวิจัยนี้

ท้ายที่สุด ขอกราบขอบพระคุณ คุณพ่อ ครอบครัว ญาติพี่น้องและเพื่อนๆ ที่ให้กำลังใจในการทำโครงงานวิจัยที่มีค่านี้

(ผู้ช่วยศาสตราจารย์ ดร. ฉวีวรรณ เพ็ชรศิริ)
หัวหน้าโครงการ
8 เม.ย. 2554

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

Abstract

The aim of this research is to automatically extract the medicinal properties of an object, especially an herb, from technical documents as knowledge sources for health-care problem solving through the question-answering system, especially What-Question, for disease treatment. The extracted medicinal property knowledge is based on multiple simple sentence or EDUs (Elementary Discourse Units). There are three problems of extracting the medicinal property knowledge: the herbal object identification problem, the medicinal property identification problem for each object and the medicinal property boundary determination problem. According to the question-answering system, there are two main problems as how to determine the focus of What-Question and how to solve the question and answer alignment from the extracted medicinal-property knowledge base. This research applies NLP (Natural Language Processing) technique with statistical based approach to solve the research problems. We propose using the lexico syntactic pattern to identify the medicinal property along with machine learning technique as Naïve Bayes (with verb features) and the centering theory for comparative studying of solving the boundary problem. And, we also propose using the question patterns and the predicate representation for the alignment of the question and the extracted medicinal-property knowledge as the answer. The result shows successfully the medicinal property extraction of the precision and recall of 87% and 74%, respectively, along with the correctness of the boundary determination as 91.1% by Naïve Bayes and 86% by the centering theory. And, the result from the question answering system is 72% of answering correctly from 50 random questions

บทคัดย่อ

จุดประสงค์ของงานวิจัยนี้คือการสกัดความรู้ด้านสรรพคุณทางยาของพืชสมุนไพรโดยเฉพาะอย่างยิ่งพืชสมุนไพรไทยจากแหล่งความรู้ที่เป็นเอกสารทางวิชาการภาษาไทยเพื่อใช้แก้ไขปัญหาทางด้านสุขภาพโดยผ่านระบบการตอบคำถามอัตโนมัติกับคำถามประเภท "อะไร/What-Question" ซึ่งถามเกี่ยวกับคุณสมบัติของวัตถุ หรือสรรพคุณทางยาใช้รักษาโรคของพืชสมุนไพร โดยความรู้ที่สกัดได้นี้ต้องอยู่รูปของประโยคบอกเล่าแบบง่ายๆที่เรียกว่า "EDU (Elementary Discourse Unit)" ปัญหาจากการสกัดความรู้นี้ประกอบด้วย 3ปัญหาหลักคือ ปัญหาในการระบุพืชสมุนไพร ปัญหาในการระบุสรรพคุณทางยาของพืชสมุนไพรแต่ละชนิด และปัญหาในการหาขอบเขตของสรรพคุณดังกล่าว นอกจากนี้ยังมีปัญหาจากระบบการตอบคำถามอัตโนมัติ คือ ปัญหาการวิเคราะห์คำถามประเภท "อะไรบ้าง" ปัญหาการระบุโฟกัส (Focus)ของคำถาม และปัญหาการสกัดคำตอบ ดังนั้นงานวิจัยนี้ จึงขอเสนอการใช้กรรมวิธีการประมวลผลภาษาธรรมชาติร่วมกับแนวทางสถิติ เพื่อใช้แก้ปัญหา 2ส่วนคือ ส่วนของการสกัดความรู้สรรพคุณทางยาของพืชสมุนไพรไทยโดยใช้ในการระบุสรรพคุณทางยาของพืชสมุนไพร และใช้เทคนิคการเรียนรู้ของเครื่องด้วย Naïve Bayes (NB) เพื่อหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรไทยโดยเปรียบเทียบกับการใช้ทฤษฎีทางภาษาศาสตร์คือทฤษฎีเซนเทอร์ริง (Centering Theory, CT) และส่วนการตอบคำถามใช้การเรียนรู้แพทเทิร์นของคำถาม "อะไร/อะไรบ้าง" เพื่อทำอะไลเมนท์ (Alignment) กับคำตอบที่สกัดได้ซึ่งอยู่ในรูปแทนของเพรดดิเคต (Predicate Representation) ผลจากการวิจัยพบว่าส่วนการสกัดความรู้สรรพคุณทางยาของพืชสมุนไพรไทยมีการสกัดถูกต้องโดยเฉลี่ยของ พรีซิชัน (Precision) เป็น 87 % และของรีคอล (Recall) เป็น 74% และการหาขอบเขตสรรพคุณดังกล่าวได้ถูกต้องโดยเฉลี่ยของ NBเป็น 91.5 % และของ CT เป็น 86 % ส่วนการตอบคำถามระบบสามารถตอบได้ถูกต้อง72%