



242271



รายงานผลการวิจัย

เรื่อง

การสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสาร
ภาษาไทยเพื่อสนับสนุนการตอบคำถามอัตโนมัติ

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

โดย

ผู้ช่วยศาสตราจารย์ ดร.ฉวีวรรณ เพ็ชรศิริ

รายงานการวิจัยนี้ได้รับทุนอุดหนุนจากมหาวิทยาลัยสุรนารีเป็นทุน

พ.ศ. 2553



รายงานผลการวิจัย

เรื่อง

การสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสาร

ภาษาไทยเพื่อสนับสนุนการตอบคำถามอัตโนมัติ

Knowledge Extraction of Medicinal Properties of Thai Herbs from Thai Texts

for Supporting Automatic Question-Answering System

โดย

ผู้ช่วยศาสตราจารย์ ดร. จวีวรรณ เพ็ชรศิริ

รายงานการวิจัยนี้ได้รับทุนอุดหนุนจากมหาวิทยาลัยธุรกิจบัณฑิตย์

พ.ศ. 2553



กิตติกรรมประกาศ

ขอขอบพระคุณ รองศาสตราจารย์ ดร. ระพีพรรณ ปริยะกุล ที่กรุณาสละเวลา ให้ความรู้ และคำแนะนำเกี่ยวกับการทำระบบถามตอบอัตโนมัติ

ขอขอบพระคุณ รองอธิการบดีฝ่ายวิจัยและวิทยาบริการ มหาวิทยาลัยธุรกิจบัณฑิตที่ให้โอกาสข้าพเจ้าและอาจารย์ สำราญ ในการศึกษาค้นคว้าวิจัยเรื่อง “การสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทยเพื่อสนับสนุนระบบการตอบคำถามอัตโนมัติ” จนสำเร็จ

ขอขอบพระคุณ คณะบดีคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิต ที่ช่วยตรวจสอบการใช้ภาษาไทยให้เหมาะสม

ขอขอบพระคุณ มหาวิทยาลัยธุรกิจบัณฑิต ที่ให้เงินทุนสำหรับสนับสนุนโครงการวิจัยนี้

ท้ายที่สุด ขอกราบขอบพระคุณ คุณพ่อ ครอบครัว ญาติพี่น้องและเพื่อนๆ ที่ให้กำลังใจในการทำโครงการวิจัยที่มีค่านี้

(ผู้ช่วยศาสตราจารย์ ดร. จวีวรรณ เพ็ชรศิริ)

หัวหน้าโครงการ

8 เม.ย. 2554

สารบัญ

	หน้า
สารบัญ	i
สารบัญตาราง	iii
สารบัญรูป	iv
บทนำ	1
1. ความเป็นมาของปัญหา	1
2. วัตถุประสงค์	4
3. สมมติฐาน	4
4. นิยามคำศัพท์	5
5. ขอบเขตงานวิจัย	5
งานวิจัยที่เกี่ยวข้อง	6
ความรู้พื้นฐาน	6
1. Naïve Bayes Classifier	6
2. Centering Theory	7
งานวิจัยก่อนหน้า	11
การสกัดความรู้สรรพคุณทางยาของสมุนไพร	11
1. แนวทางสถิติ	11
2. แนวทางแพทเทิร์นหรือกฎร่วมกับแนวทางสถิติ	12
การตอบคำถามความรู้สรรพคุณทางยาของสมุนไพร	12
1. แนวทางแพทเทิร์นหรือกฎ	12

สารบัญ (ต่อ)

	หน้า
2. แนวทางสถิติ	12
ปัญหาการสกัดความรู้สรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทย	14
เพื่อสนับสนุนระบบการตอบคำถามอัตโนมัติ	
ปัญหาการสกัดความรู้สรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทย	14
1. ปัญหาการระบุเอนตีตี้สมุนไพรไทย	14
2. ปัญหาการระบุสรรพคุณทางยาของสมุนไพรไทย	14
3. ปัญหาการหาขอบเขตของ EDUs สรรพคุณทางยาของสมุนไพรไทย	14
ปัญหาจากระบบการตอบคำถามเกี่ยวกับคุณสมบัติของเอนตีตี้สมุนไพรไทย	15
1. ปัญหาการระบุคำถาม	15
2. ปัญหาความกำกวมของ Question Word	15
3. ปัญหาการระบุโฟกัสของคำถาม	15
กรรมวิธีดำเนินงาน	17
ส่วนสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรจากเอกสารภาษาไทย	17
1. การเตรียมคลังข้อมูล	17
2. การเรียนรู้ขอบเขตของสรรพคุณของพืชสมุนไพร	20
3. การสกัดความรู้สรรพคุณทางยาของพืชสมุนไพร	21
4. การแทนความรู้สรรพคุณทางยาของพืชสมุนไพร	22
ส่วนการตอบคำถามชนิด “ถามอะไร” ประเภทลิสต์ และประเภทเอนตีตี้เกี่ยวกับ	23
สรรพคุณทางยาของพืชสมุนไพร	

สารบัญ (ต่อ)

	หน้า
1. การเรียนรู้แพทเทิร์นของคำถามอะไร	23
2. การวิเคราะห์คำถาม	24
3. การหาคำตอบ	24
ผลการทดลองและการประเมินผล	25
การสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพร	25
การตอบคำถามชนิด “ถามอะไร” ประเภทลิสต์ และประเภทเอนดีตีเกี่ยวกับ	27
สรรพคุณทางยาของพืชสมุนไพร	
สรุป	28
เอกสารอ้างอิง	29

สารบัญตาราง

ตาราง		หน้า
1	แสดงพฤติกรรมทางภาษาของคำกริยาที่มีความคิดเป็นสรรพคุณทางยา (Medicinal-Property Verb Concept) และไม่มีความคิดเป็นสรรพคุณทางยา (Non-Medicinal-Property Verb Concept) จาก Surface Form เดียวกัน จากคลังข้อมูล 500 EDUs	18
2	แสดงพีเจอร์ที่เป็นกริยาแสดงสรรพคุณทางยาของพืชสมุนไพร V_{mp} รวมทั้ง สารสนเทศ (นามวลี) ที่อยู่รอบๆกริยา	20
3	แสดงค่าความน่าจะเป็นของ V_{mp} จาก V_{mp} pair คือ $V_{mp_at_ij}$ และ $V_{mp_at_ij+1}$ ที่มีความคิดเป็นสรรพคุณทางยาของพืชสมุนไพร และไม่มีความคิดเป็นสรรพคุณทางยาของพืชสมุนไพร	21
4	แสดงค่า Precision, Recall, และ ค่าความถูกต้องของการหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรโดยเทคนิค NB และ CT	25
5	แสดงค่า t-test ของค่าความถูกต้องของการหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรระหว่างเทคนิคที่แตกต่างกันคือ NB กับ CT	26
6	แสดงค่า t-test ของค่าความถูกต้องของการหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรระหว่างคลังข้อมูลที่แตกต่างกัน	26
7	แสดงค่า t-test ของค่า Precision และ Recall ระหว่างคลังข้อมูล (Corpus) ที่แตกต่างกัน	26

สารบัญรูป

รูป		หน้า
1	แสดงกรอบงานของระบบศูนย์บริการความรู้อัตโนมัติ	1
2	แสดงกฎสถานการณ์ส่งผ่านของเซนเทอร์ริง	8
3	ระบบงานโดยสรุป	17
4	ตัวอย่างการกำกับ EDUที่เป็นความรู้สรรพคุณทางยาของพืชสมุนไพร	19
5	อัลกอริทึม Medicinal Property Boundary Extractionโดย Naive Bayes	22
6	อัลกอริทึม Medicinal Property Boundary Extractionโดย Centering Theory	22

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 โดยความรู้ที่สกัดได้นี้ต้องอยู่รูปของประโยคบอกเล่าแบบง่าย ๆ ที่เรียกว่า “EDU (Elementary Discourse Unit)” ปัญหาจากการสกัดความรู้นี้ประกอบด้วย 3 ปัญหาหลักคือ ปัญหาในการระบุพืชสมุนไพร ปัญหาในการระบุสรรพคุณทางยาของพืชสมุนไพรแต่ละชนิด และปัญหาในการหาขอบเขตของสรรพคุณดังกล่าว นอกจากนี้ยังมีปัญหาจากระบบการตอบคำถามอัตโนมัติ คือ ปัญหาการวิเคราะห์คำถามประเภท “อะไรบ้าง” ปัญหาการระบุโฟกัส (Focus) ของคำถาม และปัญหาการสกัดคำตอบ ดังนั้นงานวิจัยนี้ จึงขอเสนอการใช้กรรมวิธีการประมวลผลภาษาธรรมชาติร่วมกับแนวทางสถิติ เพื่อใช้แก้ปัญหา 2 ส่วนคือ ส่วนของการสกัดความรู้สรรพคุณทางยาของพืชสมุนไพรไทยโดยใช้ในการระบุสรรพคุณทางยาของพืชสมุนไพร และใช้เทคนิคการเรียนรู้ของเครื่องด้วย Naïve Bayes (NB) เพื่อหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรไทยโดยเปรียบเทียบกับการใช้ทฤษฎีทางภาษาศาสตร์คือทฤษฎีเซนเทอร์ริง (Centering Theory, CT) และส่วนการตอบคำถามใช้การเรียนรู้แพทเทิร์นของคำถาม “อะไร/อะไรบ้าง” เพื่อทำอะไลเมนต์ (Alignment) กับคำตอบที่สกัดได้ซึ่งอยู่ในรูปแบบของเพรดิเคต (Predicate Representation) ผลจากการวิจัยพบว่าส่วนการสกัดความรู้สรรพคุณทางยาของพืชสมุนไพรไทยมีการสกัดถูกต้องโดยเฉลี่ยของ พรีซิชั่น (Precision) เป็น 87 % และของรีคอล (Recall) เป็น 74% และการหาขอบเขตสรรพคุณดังกล่าวได้ถูกต้องโดยเฉลี่ยของ NB เป็น 91.5 % และของ CT เป็น 86 % ส่วนการตอบคำถามระบบสามารถตอบได้ถูกต้อง 72%