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

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

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 Na?ve 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%

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 Na?ve 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 Na?ve Bayes and 86% by the centering theory. And, the result from the question answering system is 72% of answering correctly from 50 random questions

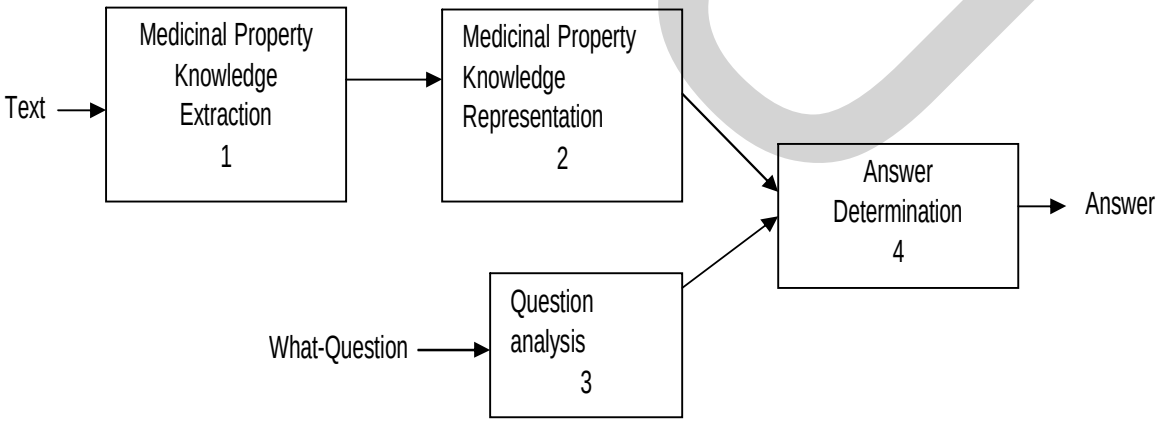
การสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทยเพื่อสนับสนุนระบบการตอบคำถามอัตโนมัติ

บทนำ

1. ความเป็นมาของปัญหา

ระบบศูนย์บริการความรู้อัตโนมัติ (Automatic Knowledge Service Center System) ในปัจจุบันมีปัญหาในด้านการให้บริการความรู้เชิงบรรยายแบบอัตโนมัติได้อย่างมีประสิทธิภาพ โดยเฉพาะอย่างยิ่งความรู้เกี่ยวกับคุณสมบัติหรือสรรพคุณต่างๆของเอนติตี้ (Entity)ที่เป็นวัตถุ (Object) มีอยู่จริง (เช่น โหระพา กระเพรา ขมิ้น ขิง ข่า พลู มะนาว เป็นต้น ซึ่งคือพืช สมุนไพรไทย) ความรู้เกี่ยวกับการเตรียมยาสมุนไพร และความรู้วิธีการใช้ยาสมุนไพรไทย เป็นต้น ความรู้เหล่านี้ถือว่าเป็นความรู้ที่มีประโยชน์อย่างมากสำหรับประชาชนทั่วไปเพื่อนำไปใช้บำรุงรักษาสุขภาพของตนเองด้วยการตอบคำถามความรู้เกี่ยวกับสมุนไพรไทยผ่านทางระบบอัตโนมัติของศูนย์บริ-การความรู้ด้วยคำถามปะเภทต่างๆ เช่น “อะไร(What-question)” “อย่างไร(How-question)” “ทำไม(Why-question)” “ที่ไหน(Where-question)” “เมื่อไร(“When-question)” เป็นต้น นอกจากนี้(Takahashi K., et. al., 2004; Metzler D. and Croft W.B., 2005) ได้แบ่งคำถาม “อะไร(What-question)” ออกเป็นประเภทต่างๆเช่น ประเภทลิสต์(List)/รายการ ประเภทเอนติตี้ ประเภทนิยาม ประเภทเวลา เป็นต้น ตัวอย่างเช่นเช่น “ใบโหระพามีสรรพคุณทางยาอะไรบ้าง” “พืชสมุนไพรอะไรมีสรรพคุณขับลม” “สมุนไพรคืออะไร” “ควรใช้ยาสมุนไพรเมื่อไร” เป็นต้น ทั้งนี้จะทำให้ประชาชนทั่วไปมีความรู้ได้โดยไม่ต้องทำการสืบค้นและอ่านจากเอกสารต่างๆ ซึ่งทำให้เสียเวลามาก ดังนั้นระบบศูนย์บริการความรู้อัตโนมัติ ดังกล่าวซึ่งแสดงในรูปที่1 ควรประกอบด้วยสี่ส่วนหลักคือ

1) การสกัดความรู้เรื่องสรรพคุณทางยาของสมุนไพรไทย (Medicinal Property Knowledge Extraction) เก็บลงในฐานความรู้
2) การแทนความรู้เรื่องสรรพคุณทางยาของสมุนไพรไทย (Medicinal Property Knowledge Representation)
3) การวิเคราะห์คำถาม (Question Analysis)
4) การหาคำตอบ (Answer Determination)



รูปที่1 แสดงกรอบงานของระบบศูนย์บริการความรู้อัตโนมัติ

สำหรับงานวิจัยครั้งนี้จะเป็นการศึกษาเฉพาะการสกัดความรู้เรื่องสรรพคุณทางยาของสมุนไพรไทย และระบบการตอบคำถามเกี่ยวกับสรรพคุณทางยาของสมุนไพรไทยซึ่งจะเป็นคำถาม “อะไร/what” ประเภทลิสต์ (เช่น ... สรรพคุณอะไรบ้าง) ประเภทเอนติตี้ (เช่น...สมุนไพรอะไร) นอกจากนี้ความรู้เรื่องสรรพคุณของสมุนไพรไทยปรากฏ

ในเอกสารพืชสมุนไพรไทยในรูปแบบของหลายๆ EDU (Elementary Discourse Unit, คือประโยคง่ายๆ ธรรมดาไม่ซับซ้อน, Carlson and et. al, 2003) ดังตัวอย่างต่อไปนี้

EDU1: กระเทียมเป็นยาขับลมในลำไส้
EDU2: [กระเทียม]แก้ไอ
EDU3: [กระเทียม]ขับเสมหะ
EDU4: [กระเทียม] ช่วยย่อยอาหาร
EDU5: [กระเทียม]รักษากลาก เกลื้อน
หมายเหตุ: สัญลักษณ์ '[..]' หมายถึงการละคำ

ดังนั้นจากตัวอย่างเอกสารพืชสมุนไพรไทยต่อไปนี้

พริกไทย (Piper nigrum Linn.)
สรรพคุณ เปลือกของพริกไทยมีน้ำย่อยสำหรับย่อยไขมัน ด้วยเหตุนี้ตำราโบราณจึงเชื่อกันว่า พริกไทยสามารถลดความอ้วนได้, พริกไทยช่วยกระตุ้นปุ่มรับรสที่ลิ้น เพื่อให้กระเพาะอาหารหลั่งน้ำย่อยได้มากขึ้น, พริกไทยดำมีรสเผ็ดอุ่น เมื่อรับประทานเข้าไปจะรู้สึกอุ่นวาบที่ท้อง ช่วยขับลม ขับเหงื่อ ขับปัสสาวะ แก้ท้องอืดท้องเฟ้อ แก้ไข้มาลาเรีย แก้อหิวาตกโรค, ใช้ก้านพริกไทย 10 ก้าน บดให้ละเอียดแล้วต้มกับน้ำ 8 แก้ว ใช้เป็นยาล้างแผลที่อักเสบ, สารพิเพอรีนในพริกไทยสามารถใช้เป็นยาฆ่าแมลง ซึ่งไม่เป็นอันตรายต่อมนุษย์โดยนำผลพริกไทยมาทุบให้แตกแล้วใช้โรยบริเวณตู้เสื้อผ้าหรือบริเวณที่ต้องการ

จากตัวอย่างเอกสารสมุนไพรไทยข้างต้นส่วนที่ขีดเส้นใต้คือสรรพคุณหรือคุณสมบัติของพืชสมุนไพร "พริกไทย" ดังนั้นปัญหาในส่วนของการสกัดความรู้เกี่ยวกับคุณสมบัติของเอนติตี้โดยเฉพาะเรื่องสรรพคุณทางยาของสมุนไพรไทยจากเอกสารภาษาไทย ประกอบด้วยสามปัญหาคือ ปัญหาการระบุเอนติตี้สมุนไพรไทย ปัญหาการระบุสรรพคุณทางยาของสมุนไพรไทย ปัญหาการหาขอบเขตของ EDUs สรรพคุณทางยาของสมุนไพรไทย

ส่วนปัญหาจากระบบสอบถามเกี่ยวกับคุณสมบัติของเอนติตี้สมุนไพรไทย ประกอบด้วยปัญหาการวิเคราะห์ "คำถามอะไร" ปัญหาการระบุโฟกัส (Focus)ของคำถาม และปัญหาการหาคำตอบจากความรู้เกี่ยวกับคุณสมบัติของเอนติตี้ที่ได้สกัดมา

งานวิจัยที่เกี่ยวข้องสามารถแบ่งออกเป็นสองส่วนหลักคือ ส่วนการสกัดและแทนความรู้เกี่ยวกับคุณสมบัติของเอนติตี้ และส่วนการตอบคำถามของคำถามประเภท"อะไรบ้าง" กล่าวคือมีการใช้เทคนิคต่างๆสกัดความรู้เกี่ยวกับคุณสมบัติของเอนติตี้จากงานวิจัยที่เกี่ยวข้องเช่น Weeber M. and Vos. R., 1998; Fang et al.,2008; และ Paşca M., 2008 โดย Weeber M. and Vos. R.(1998)ได้กล่าวถึงคุณสมบัติของฤทธิ์ยาจำเป็นต้องคำนึงถึง 3 เรื่องหลักคือ ยา(A) ผลที่แสดงออกทางกายภาพ (Physiological Effect)(B) และ โรค (C) และความสัมพันธ์ระหว่าง 3 เรื่องดังกล่าวเป็น A -> B , B -> C, ทำให้ได้ A -> C ดังนั้น Weeber M. and Vos. R.(1998) เสนอการสกัดความรู้ทางการแพทย์โดยการหาความสัมพันธ์ระหว่างคำ(ซึ่งอยู่ในรูปของนามวลี) A, B, และ C จากเอกสารทางการแพทย์โดยใช้แนวทางสถิติด้วยการความสัมพันธ์ชนิดAssociationระหว่างคำที่อยู่รอบๆ Side-Effect Words ภายใต้กรอบหน้าต่างขนาด 64 คำ Expert I ได้ recall = 0.19 precision = 0.14 Expert II ได้ recall = 0.24 precision = 0.07

Fang et al.,(2008) ได้ค้นพบความสัมพันธ์(Association Discovery) ระหว่างคำนามต่างๆที่เป็นชื่อยาสมุนไพรจีน โรค พันธุกรรม ผลกระทบ (Side Effect) ของยาสมุนไพรจีน และส่วนผสม โดยการวิเคราะห์การเกิดร่วมกัน (Collocation Analysis) จากเอกสารที่มีการกำกับ และมีการนำเอา IE (Information Extraction) และแบบจำลอง Swanson's ABC (A -> Bและ B -> C ทำให้ได้ ความสัมพันธ์แบบการส่งผ่าน (Transitive Association) คือ A -> C) มาประยุกต์ใช้ โดยกำหนดให้ A คือพันธุ์กรรม B คือ ส่วนผสมที่สามารถควบคุม A และ C คือ ยาสมุนไพรจีน เพื่อการบอกเป็นนัยของ A -> C เมื่อ A -> Bและ B -> C ปรากฏขึ้นในเอกสารอย่างมีนัยสำคัญ ผลการวิจัยของ Fang et al.,(2008) ได้ precision =0.91 อย่างไรก็ตามวิธีของ Fang et al.,(2008)อยู่บนพื้นฐานของการใช้แต่เพียงนามวลี

Paşca M. (2008) ระบุความรู้ที่เป็นจริงเกี่ยวกับคลาสวัตถุ (Object Class) ต่างๆได้โดยการใช้ อิสอะแพทเทิร์น (Is-A pattern) กับการสอบถามหรือคิวรี (Query) ที่มีคีย์เวิร์ด (Key Word) อยู่ด้วยทำการสกัดตุณสมบัติต่างๆที่อยู่ในรูปนามวลี จากเอกสารบนเว็บและจากส่วนบันทึกคิวรี (Query Log) ด้วยค่า precision of 0.8 สำหรับ 100 ดลาสที่สามารถระบุได้ จาก5 คุณสมบัติหรือ แอททริบิวท์ (Attribute) ที่ สกัดได้
อย่างไรก็ตาม วิธีดังกล่าว(Weeber M. and Vos. R., 1998; Fang et al.,2008; Paşca M., 2008) ไม่เหมาะสำหรับงานวิจัยนี้ในส่วนของการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของเอนติตี้พืชสมุนไพรไทยจากเอกสารภาษาไทย เพราะสรรพคุณเหล่านี้จะแสดงอยู่ในรูปของกริยาวลี และหลายๆกริยาวลีอยู่ต่อเนื่องกันต่อหนึ่งเอนติตี้สมุนไพรซึ่งจะอยู่ในรูปของนามวลีที่ส่วนใหญ่มักจะละ

ส่วนการตอบคำถามของคำถามประเภท"อะไรบ้าง" ได้มีการใช้เทคนิคต่างๆจากงานวิจัยที่เกี่ยวข้อง (Riloff E. and Thelen M.,2000; Quaresma P. and Rodrigues I., 2005; Fan S. et. al., 2008) โดย Riloff E. and Thelen M.(2000) ได้ใช้กฎเป็นพื้นฐาน (Rule Base)ต่างๆพร้อมกับการให้คะแนน สำหรับระบบตอบคำถามอย่างอัตโนมัติ ทั้งนี้เพื่อทดสอบความเข้าใจจากการอ่านบทความภาษาอังกฤษ โดยระบบอัตโนมัติหลังจากที่ได้ผ่านซอฟท์แวร์แจงประโยค "Sundance" วิธีการเลือกคำตอบโดยการให้คะแนนสำหรับประโยคทีมีคำตรงกับคำในประโยคคำถาม ถ้าประโยคใดมีคะแนนสูงประโยคนั้นคือประโยคคำตอบ สำหรับการตอบคำถาม "What"ได้ความถูกต้องเป็น 0.31 สำหรับระบบอัตโนมัติ 0.28 สำรับคนตอบ อย่างไรก็ตามกฎของ คำถาม "What"เหล่านั้นไม่สามารใช้กับงานวิจัยนี้

Quaresma P. and Rodrigues I.(2005) ได้ เสนอระบบการตอบคำถามที่มีการใช้ซอฟท์แวร์แจงประโยคภาษาโปรตุเกส (Protuguese Parser) กับเอกสารทางคดีความและประโยคคำถาม สำหรับสถาบันเกี่ยวกับความยุติธรรมของชาวโปรตุเกส เช่น ศาล สำนักงานทนายความ เป็นต้น โดยศึกษาคำถามเกี่ยวกับคดีความ โดยวิธี ยูนิไฟด์ (Unify) ด้วยโปรแกรมภาษาโปรลอค (Prolog) ระหว่างประโยคในเอกสารทางคดีความกับคำถามหลังจากผ่านซอฟท์แวร์แจงประโยคได้ความถูกต้อง 25% จาก 200 คำถาม

Fan S. et. al., (2008) ได้เสนอแบบจำลอง CRF (Conditional Random Field Model) สำหรับระบบการตอบคำถามทีคำถามมีลักษณะซับซ้อนมาก โดยเขาแก้ไขปัญหาสำหรับคำถามที่ซับซ้อนด้วยการให้มีการกำกับความหมายระดับก้อน (Chunk Semantic) ให้กับคำถาม ซึ่งคำถามนี้จะถูกนำไปหาค่าความคล้าย (Similarity) กับคำถามที่มีคู่คำตอบ (Question-Answer Pair) จากเว็บไซด์ภาษาจีน CRFคล้ายกับ Maximum Entropy(ME) เซทฟีเจอร์(Feature Set) ที่ใช้โดย CRFประกอบด้วยคำที่กำหนดความหมายไว้ , tag, Pattern, และKey word (ดูในหัวข้องานวิจัยก่อนหน้า) เพื่อใช้หาค่าความคล้าย ซึ่งได้ค่าความถูกต้องเฉลี่ย 93.07% precision 93.07%recall

อย่างไรก็ตาม วิธีดังกล่าว(Riloff E. and Thelen M.,2000; Quaresma P. and Rodrigues I., 2005; Fan S. et. al., 2008) ไม่เหมาะสำหรับงานวิจัยนี้ในส่วนของการตอบคำถาม เนื่องคำถาม "what" ของงานวิจัยนี้เป็นประเภทลิสต์ ส่วนของ Riloff E. and Thelen M.(2000)และ Quaresma P. and Rodrigues I.(2005) เป็นประเภทอื่นที่ไม่ใช่ลิสต์ ฉะนั้นจะมีแพทเทิร์นที่ต่างกันออกไป และลักษณะภาษาไทยต่างจากภาษาอังกฤษและภาษาโปรตุเกส

ซึ่งมีการใช้ "Question Mark, ?" เป็นตัวระบุประโยคคำถาม ในขณะที่ภาษาไทยไม่มี นอกจากนี้คู่คำถามคำตอบเกี่ยวกับสรรพคุณสมุนไพรไทยบนเว็บไซต์มีไม่มากเหมือนของจีนฉะนั้นวิธีของ Fan S. et. al.(2008) ไม่สามารถนำมาประยุกต์ใช้กับงานวิจัยนี้

ดังนั้นในส่วนของการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของเอนติตี้พืชสมุนไพรไทยจากเอกสารภาษาไทย งานวิจัยนี้จึงขอเสนอการใช้แพทเทิร์น ทางภาษาศาสตร์คือ NP1 V_{mp} NP2 (เมื่อ NP1 คือเซทความคิดนามวลีหรือ Noun Phrase Concept เกี่ยวกับเอนติตี้ เช่นพืชสมุนไพร(Herb) และสาร(Substance)ของสมุนไพร NP2 คือเซทความคิดนามวลีเกี่ยวกับอาการ (Symptom) โรค(Disease) เชื้อ(Pathogen) และ V_{mp} คือเซทความคิดกริยาที่แสดงให้เห็นสรรพคุณทางยา(Medicinal-Property Verb Concept) ดังนี้

$np1_i$= part + TopicName เมื่อ i=1,2,..m

part∈{null,"ใบ" ,"ดอก", "ราก", "ต้น", "เมล็ด", "ผล", ...}

TopicName ∈{"โหระพา/basil" "กระเพรา/sweet basil" "กระเทียม/garlic" "พริกไทย/pepper" "พริก/chili" "ขิง/ginger" "ข่า/galangal" "ตระไคร้/lemon glass" "ว่านหางจระเข้/aloe" "กานพลู/cinnamon" "พลู/betel" .....}

NP1 = {$np1_1$, $np1_2$,.., $np1_m$}

NP2 = {"[อาการ] ปวดท้อง/abdominal pain", "[อาการ] คลื่นไส้/nausea", "[อาการ] อาเจียน/vomit" ,"[อาการ] เวียนศีรษะ/dizziness" , "ไข้/fever", "[อาการ] ปวดศีรษะ /headache", "[อาการ] [แผล] อักเสบ/inflame", "[อาการ] ภูมิแพ้/allergy", "[อาการ] ท้องเสีย/diarrhea" ,"ปัสสาวะ/urine", "ลม/gas", "กรดในกระเพาะ/stomach acid", "โรคผิวหนัง/skin disease", "โรคหัวใจ/heart disease", "ความดัน/blood pressure"...}

V_{mp}= {"รักษา/cure" "บรรเทา/relieve" "แก้/stop,prevent" "ขับ/release" "ลด/reduce" "เพิ่ม/increase" ...}

มาทำการระบุความรู้เกี่ยวกับสรรพคุณทางยาของเอนติตี้พืชสมุนไพรไทย โดยที่ความคิดนามวลี และความคิดกริยาได้จากสารานุกรมไทย เวิอร์ดเนต (Wordnet) และ www.longdo.com นอกจากนี้ขอเสนอ คู่ความคิดกริยาที่เกี่ยวสรรพคุณทางยาและอยู่ต่อเนื่องกันภายในหนึ่งกรอบหน้าต่างที่เคลื่อนไปด้วยระยะทางหนึ่ง EDU เพื่อใช้หาขอบเขตของสรรพคุณทางยาของเอนติตี้พืชสมุนไพรนั้นด้วยการใช้Naive Bayesทดสอบ Hypothesis นอกจากนี้ความรู้ที่สกัดได้จะอยู่ในรูปของเมตริกซ์เวคเตอร์(V) ของความคิดกริยาแสดงสรรพคุณยาของพืชสมุนไพร(V_i) ดังนี้

V_i = {v_{i1}, v_{i2}....v_{ik} } สำหรับพืชสมุนไพรหนึ่งชนิด และ v_{ik}∈V_{mp} (v_{ik} คือ $v_{mp\_at\_ik}$ และ V_i คือ V_{mp})

V = {V_i} where i=1..n

ส่วนการตอบคำถามเกี่ยวกับสรรพคุณทางยาของสมุนไพรไทยงานวิจัยนี้จึงขอเสนอการใช้แพทเทิร์นของคำถาม (Question Word) "what" ประเภทลิสต์"อะไรบ้าง" และประเภทเอนติตี้ "x อะไร" (เมื่อ x คือเอนติตี้) ร่วมกับ NP1, NP2, และ V_{mp} มาทำการระบุประเภทคำถามและระบุโฟกัสของคำถาม เพื่อหาคำตอบจากความรู้ที่สกัดได้นั้น

2. วัตถุประสงค์

2.1 ศึกษาวิธีการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของเอนติตี้พืชสมุนไพรไทยจากเอกสารภาษาไทย

2.2 ศึกษาระบบสอบถามเกี่ยวกับสรรพคุณทางยาของสมุนไพรไทยด้วยคำถามประเภท "อะไรบ้าง(What-question)"

3. สมมติฐาน

3.1 คู่ความคิดระหว่างนามวลี NP1กับนามวลีNP2 และความคิดกริยาVmp จากแพทเทิร์นทางภาษาศาสตร์NP1 Vmp NP2 สามารถระบุว่าเป็นEDUอธิบายสรรพคุณทางยา

3.2 คู่ความคิดกริยาที่เกี่ยวกับสรรพคุณทางยาและอยู่ต่อเนื่องกันภายในหนึ่งกรอบหน้าต่าง ขนาด 2 EDUs แสดงว่าขอบเขตของ EDUs สรรพคุณทางยาของสมุนไพรไทยยังไม่สิ้นสุด

4. นิยามคำศัพท์

Medicinal Effect: ผลทางยาซึ่งเป็นคุณสมบัติของยา (Medicinal Property) หรือเรียกว่าสรรพคุณทางยา
EDU: Elimentary Discourse Unit คือประโยคง่าย ๆ ธรรมดาไม่ซับซ้อน
QA System: Question Answering Systemคือระบบการตอบคำถาม
What-question: คำถามประเภทอะไร
NP:Noun Phrase Conceptคือความคิดนามวลี
V:Verb Concept คือความคิดกริยา
Question Word: คำที่ใช้แสดงคำถาม
Key Word:คำสำคัญ

5. ขอบเขตของการวิจัย

5.1 สามารถสกัดความรู้เกี่ยวกับสรรพคุณทางยาของเอนติตี้พืชสมุนไพรไทยจากเอกสารภาษาไทย

5.2 สามารถสอบถามเกี่ยวกับสรรพคุณทางยาของสมุนไพรไทยด้วยคำถามประเภท “อะไรบ้าง(What-question)”ได้อย่างอัตโนมัติ

5.3 การวิจัยนี้จะเป็นการตอบคำถาม “อะไร(what-question)”ประเภทลิสต์ “อะไรบ้าง” และประเภทเอนติตี้ “X อะไร” เมื่อ X คือเอนติตี้ เท่านั้น

งานวิจัยที่เกี่ยวข้อง

การวิจัยการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทยเพื่อสนับสนุนระบบการตอบคำถามอัตโนมัติประกอบด้วยความรู้พื้นฐาน และงานวิจัยก่อนหน้าดังนี้

ความรู้พื้นฐาน

1) Na?ve Bayes Classifier(Mitchell 1997)

ตัวจัดประเภทเนอีฟเบย์ (Na?ve Bayes classifier, NB) หรือ ตัวเรียนรู้ NB เป็นวิธีการเรียนรู้ที่นิยมใช้กันมาก และเป็นการเรียนรู้ที่อยู่บนพื้นฐานของความน่าจะเป็น (Probability) กับข้อมูลที่สังเกต (Observed Data) ตามที่ Mitchell T.M., (1997) ได้กล่าวว่าตัวจัดประเภท NB สามารถประยุก๖ใช้กับงานเรียนรู้ที่ซึ่งแต่ละตัวอย่าง x (Instance x) ได้ถูกอธิบายโดยการเชื่อมโยงค่าแอททริบิวท์ (Attribute Values) ต่างๆ และที่ซึ่งฟังก์ชันเป้าหมาย (Target Function, f(x)) สามารถแสดงค่าคลาส (Class Value, v) จาก คลาสไฟไนท์เซท (Class Finite Set, V) ดังนั้นเซทของตัวอย่างการเรียนรู้ของฟังก์ชันเป้าหมายได้ถูกกำหนดไว้ให้ และเมื่อมีตัวอย่างใหม่เกิดขึ้นก็สามารถอธิบายได้ คือบอกค่าคลาสได้ด้วยทูปเพิล (Tuple) ของค่าแอททริบิวท์ $<a_1, a_2,..a_n>$ นั่นคือตัวเรียนรู้ทำนายค่าเป้าหมายหรือการจัดแบ่งประเภทสำหรับตัวอย่างใหม่ที่เข้ามา

แนวทางเบย์ที่จะจัดประเภทให้กับตัวอย่างใหม่ที่เข้ามานั้นเป็นการกำหนดค่าเป้าหมายที่มีโอกาสเป็นไปได้มากสุด หรือที่เรียกว่า $v_{\text{maximum a posterior}}$ (v_{MAP}) เมื่อกำหนดค่าแอททริบิวท์ต่างๆให้ $<a_1, a_2,..a_n>$ที่ใช้อธิบายตัวอย่าง ดังแสดงในสมการ(2) และ (3)

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2 ... a_n) \quad . \qquad (1)$$

$$. v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2 ... a_n \mid v_j) P(v_j) \qquad (2)$$

ตัวจัดประเภท NBดำเนินงานบนพื้นฐานของข้อสมมติฐานแบบง่ายๆที่มีเงื่อนไขว่าค่าแอททริบิวท์แต่ละแอททริบิวท์จะต้องเป็นอิสระต่อกันเมื่อกำหนดค่าเป้าหมายไว้ให้ กล่าวคือข้อสมมติฐานเป็นการกำหนดค่าเป้าหมายของตัวอย่าง (คือคลาสของตัวอย่าง) ฉะนั้นความน่าจะเป็นของการสังเกตการเชื่อมโยงกันของ $a_1, a_2,..a_n$ คือผลคูณของค่าความน่าจะเป็นของแอททริบิวท์ต่างๆ ดังนั้นตัวจัดประเภท NB, v_{NB},สามารถแสดงได้ดังต่อไปนี้

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j) \qquad (3)$$

สำหรับงานวิจัยนี้ เราได้ประยุกต์ใช้ตัวจัดประเภท NB ที่เป็นสมการ (4) สำหรับการเรียนรู้แยกประเภทของขอบเขตของประโยคธรรมดาต่างๆ (Simple Sentences หรือ EDUs) ที่แสดงคุณสมบัติเป็นยาสมุนไพรได้สิ้นสุดหรือยังไม่สิ้นสุดดังต่อไปนี้

$$MedicinalPropertyBoundaryClass = \arg\max_{class \in Class} P(class | v_{ij}, v_{ij+1})$$
$$= \arg\max_{class \in Class} P(v_{ij} | class) P(v_{ij+1} | class) P(class) \quad (4)$$

$$where\ v_{ij} \in V_i \quad and\ v_{ij+1} \in V_i \quad (V_i\ is\ a\ medicinal\_property\_verb\_concept\ vector)$$
$$i = \{1,2,..n\} \qquad j = \{1,2,..k\}$$

เมื่อตัวแปร "*Class*" เป็นไฟไนท์เซท (Finite Set) ของประเภท ขอบเขตของสรรพคุณทางยาของสมุนไพร ได้สิ้นสุดหรือยังไม่สิ้นสุด {end, continue} และแอททริบิวท์ a_1, a_2,.. to a_n คือ ฟีเจอร์กริยาต่างๆ (Verb Features, v_{ij} และ v_{ij+1}) ที่เป็นสมาชิกของ V_i (V_i คือ เวคเตอร์ความคิดกริยาที่แสดงคุณสมบัติเป็นยาสมุนไพร) และ $V_i \subseteq V_{mp}$ (V_{mp}คือ เซทความคิดกริยาทีแสดงให้เห็นสรรพคุณทางยาของสมุนไพร, Herbal Medicinal-Property Verb Concept Set) (ดูข้อ กรรมวิธี, Method Section)

2) Centering Theory

(Walker, M., A. Joshi, and E. Prince,1998)ได้เสนอทฤษฎีเซนเทอร์ริง (Centering Theory) โดยกล่าวว่าเซนเทอร์ริงเป็นโมเดลหรือแบบจำลองของความซับซ้อนของการอนุมานทีต้องการผสมผสานระหว่างความหมายของถ้อยแถลงให้เป็นความหมายของบทความก่อนหน้า และกล่าวโดยรวมคือ นามวลี (Noun Phrase, NP) เป็นเอนติตีที่เป็นศูนย์กลาง หรือเซนเตอร์ (Center) Grosz and Sidner (Grosz, 1977; Sidner, 1979; Grosz and Sidner, 1986)ได้เสนอสถานะความสนใจ (Attentional State) ในบทความ จะต้องประกอบด้วยสองระดับของการโฟกัส (Focusing) คือ โกลบอล (Global) และ โลคัล(Local) ต่อมา Walker et al.(1998)ได้สรุปไว้ว่าเซนเทอร์ริงเป็นโมเดลของศูนย์กลางที่มีประสบการณ์ของความสนใจในบทความที่เน้นในเรื่องของความสัมพันธ์ของสถานะความสนใจ ความซับซ้อนของการอนุมาน และรูปแบบ (Form) ของการอ้างอิงการแสดงออก

จาก Walker et al.(1998) เซนเทอร์ริงโมเดล (Centering Model) อยู่บนพื้นฐานของข้อจำกัดต่อไปนี้: ส่วนของบทความ (Discourse Segment) ประกอบด้วยลำดับของถ้อยแถลง Ui, i=1,2,...n, ที่ซึ่งแต่ละถ้อยแถลง Un ถูกเชื่อมโยงกับลิสต์ (List) ของเซนเตอร์ที่มองไปข้างหน้า (แทนด้วย Cf(Ui)) ซึ่งประกอบด้วยเรื่องราวต่างๆที่มาก่อน (Antecedences) ที่เป็นไปได้ ซึ่งเป็นลำดับบางส่วนตามจำนวนของปัจจัย การจัดลำดับของเอนติตีในลิสต์ต้องสอดคล้องกันที่ว่าจะต้องเป็นโฟกัสที่สำคัญหรือเป็นพื้นฐานของบทความต่อๆมา ฉะนั้นองค์ประกอบแรกของลิสต์จะถูกระบุให้เป็นเซนเตอร์ที่ชอบมากกว่าหรือให้ความสำคัญเป็นอันดับแรก (Cp) ส่วนเซนเตอร์ที่มองไปข้างหลังของ Ui (แทนด้วย Cb(Ui)) แทนเอนติตีปัจจุบันซึ่งกำลังโฟกัสอยู่ในบทความหลังจากที่ Ui ถูกตีความหมาย เอนติตีในลิสต์ถูกจัดลำดับได้ดังต่อไปนี้

subject > direct object > indirect object> adjuncts

ตัวอย่างเช่น

Ui-1: "John helps Jim washing a car."

Ui: "He cleans the windshield very well."

Where Cp(Ui) is "He" and Cb(Ui) is "John" .

กฎของอัลกอริทึม ทฤษฎีเซนเทอร์ริง (Rules of the Centering Theory algorithm) :

Rule 1: ถ้าองค์ประกอบใดๆของ Cf(Ui-1) ถูกรู้จักโดยคำสรรพนามของถ้อยแถลง Ui, แล้ว Cb(Ui) จะต้องถูกรู้จักในรูปแบบของคำสรรพนามด้วย

Rule 2: สถานการณ์ส่งผ่าน (Transition States) ถูกจัดลำดับตามความชอบมากกว่าดังนี้ :

Continue > Retain > Smooth Shift > Rough Shift

เมื่อ “Continue” คือ Cb(Ui) ที่เท่ากับ Cb(Ui-1) (หรือ Cb(Ui-1) เป็นค่าว่าง) และ Cb(Ui) เท่ากับ Cp(Ui).

“Retain” เป็น Cb(Ui) ที่เท่ากับ Cb(Ui-1) และ Cb(Ui) ไม่เท่ากับ Cp(Ui).

“Smooth Shift” เป็น Cb(Ui) ที่ไม่เท่ากับ Cb(Ui-1) และ Cb(Ui) เท่ากับ Cp(Ui).

“Rough Shift” เป็นCb(Ui) ที่ไม่เท่ากับ Cb(Ui-1) และ Cb(Ui) ไม่เท่ากับ Cp(Ui).

Rule 2 กล่าวอ้างว่าบางครั้งการส่งผ่านระหว่างถ้อยแถลงเป็นโคฮีเรนท์ (Coherent) มากกว่าอันอื่นโดยการระบุเงื่อนไขที่ว่าการส่งผ่านเหล่านั้นจะต้องถูกเป็นที่ชอบมากกว่าอันอื่นๆ ตัวอย่างเช่นบทความที่เป็น Continue Centering แล้ว เอนติตีที่ไม่เปลี่ยนแปลงจะเป็นโคฮีเรนท์มากกว่าพวกที่เคลื่อนย้ายอย่างซ้ำๆจากเซนเตอร์หนึ่งไปยังเซนเตอร์อื่น

อัลกอริทึม:

1. Generate possible Cb and Cf combinations for each possible set of reference assignments.
2. Filter by constraints (Grosz, 1977), e.g., centering rules and constraints.
3. Rank by transition orderings.

ในอัลกอริทึมนี้เรื่องราวต่างๆที่มาก่อนและชอบมากกว่าจะถูกคำนวณได้จากความสัมพันธ์ระหว่างเซนเตอร์ที่มองไปข้างหน้า (Forward) และที่มองไปข้างหลัง (Backward) ในประโยคที่อยู่ติดกัน สี่ความสัมพันธ์ระดับระหว่างประโยคระหว่าง Ui และ Ui-1 ถูกระบุไว้ชัดเจน ซึ่งขึ้นอยู่กับความสัมพันธ์ระหว่าง Cb(Ui-1), Cb(Ui), and Cp(Ui) ถ้าเซนเตอร์ที่ชอบมากกว่า , Cp(Ui-1), ถูกรู้จักใน Ui, แล้วมันก็จะถูกทำนายเป็น Cb(Ui) ดังแสดงในรูปที่2 ในขณะที่โทโปโลจี ของการส่งผ่าน (Topology of Transition)จากถ้อยแถลงหนึ่ง, Ui-1, ไปยังถ้อยแถลงถัดไป, Ui, นั้นอยู่บนพื้นฐานของสองปัจจัยคือ

1 เซนเตอร์ที่มองไปข้างหลัง, Cb,เหมือนกันระหว่าง Ui-1 กับ Ui

2 เอนติตีของบทความเหมือนกับเซนเตอร์ที่ชอบมากกว่า, Cp, ของ Ui

	$Cb(U_i) = Cb(U_{i-1})$ OR $Cb(U_{i-1}) =$ null	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	Continue	Smooth-shift
$Cb(U_i) \neq Cp(U_i)$	Retain	Rough-shift

รูปที่2 แสดงกฎสถานการณ์ส่งผ่านของเซนเทอร์ริง (Centering transition state rule, Walker et al.,1998)

พิจารณาบทความต่อไปนี้:

U_1: "*Jane likes Mary*."

U_2: "*She often brings her food*."

U_3: "*She chats with the young woman for kids*."

Question: What do the pronouns and the description (underlined) refer to?

From sentence U_1

"*Jane likes Mary*."

1. **Generate**:

$Cf(U_1)$: <Jane, Mary>

$Cb(U_1)$: NIL

$Cp(U_1)$: Jane

2. **Filter**: non

3. **Rank by transition state ordering**: non

From sentence U_2

"*She often brings her food*."

1. **Generate**:

$Cf(U_2)$: <Jane, Mary, food> or <Mary, Jane, food>
or <Jane, Jane, food> or <Mary, Mary, food>

$Cb(U_2)$: Jane or Mary

$Cp(U_2)$: Jane or Mary

2. **Filter**:

2.1. "she" or "her" refers to $Cb(U_2)$ which can be Jane, Marry.

2.2. $Cb(U_2)$ is Jane

2.3. <Jane, Jane, food> & <Mary, Mary, food> are ruled out

3. **Rank by transition state ordering**:

(a) $Cf(U_2)$: <Jane, Mary, food>

$Cb(U_2)$: Jane

$Cp(U_2)$: Jane

So, $Cb(U_2) \neq Cb(U_1)$, $Cb(U_2) = Cp(U_2)$

i.e **smooth shift**

(b) $Cf(U_2)$: <Mary, Jane, food>

$Cb(U_2)$: Jane

$Cp(U_2)$: Mary

So, $Cb(U_2) \neq Cb(U_1)$, $Cb(U_2) \neq CP(U_2)$

i.e **rough shift**

Then, **select** *smooth shift* and the result is:
"She often brings her food. = Jane often brings Mary food."

From sentence U_3
"*She chats with the young woman for kids*."
1. **Generate**:
$Cf(U_3)$: <Jane, Mary> or <Mary, Jane> or <Jane, Jane> or <Mary, Mary>
$Cb(U_3)$: Jane or Mary or food or NIL
$Cp(U_3)$: Jane or Mary
2. **Filter**:
2.1. $Cb(U_3)$ is Jane, Mary
2.2. $Cb(U_3)$ is Jane
2.3. <Jane, Jane> & <Mary, Mary> are ruled out
3. **Rank by transition state ordering**:
(a) $Cf(U_3)$: <Jane, Mary>
$Cb(U_3)$: Jane
$Cp(U_3)$: Jane
So, $Cb(U_3) = Cb(U_2)$, $Cb(U_3) = Cp(U_3)$
i.e **continue**
(b) $Cf(U_3)$: <Mary, Jane>
$Cb(U_3)$: Jane
$Cp(U_3)$: Mary
So, $Cb(U_3) = Cb(U_2)$, $Cb(U_3) \neq Cp(U_3)$
i.e **retain**

Then, **select** *continue* and the result is:
"*She chats with the young woman for kids. = Jane chats with Mary for kids*."

ทฤษฎีเซนเทอร์ริงสามารถประยุกต์ใช้กับบทความภาษาไทยสำหรับคำนวณหาขอบเขตของ EDUs สรรพคุณทางยาของพืชสมุนไพร ถ้าสถานการณ์ส่งผ่านของประโยคที่แสดงสรรพคุณทางยาของสมุนไพรไทย EDUi (เทียบเท่ากับ Ui) และ EDUi+1 (เทียบเท่ากับ Ui+1) เป็น continue และ smooth shift, ตามลำดับ แล้วขอบเขตของสรรพคุณทางยาของสมุนไพรไทยจะจบที่ EDUi ดังตัวอย่างต่อไปนี้

(where the symbol "[..]" stands for elicit word(s):
EDU1: "พริกไทยช่วยขับลม" (Transition State = 'non')
EDU2: "[พริกไทย] ขับเหงื่อ" (Transition State = 'continue')
EDU3: "[พริกไทย] ขับปัสสาวะ" (Transition State = 'continue')
EDU4: "[พริกไทย] แก้ท้องอืดท้องเฟ้อ" (Transition State = 'continue')
EDU5: "[พริกไทย] แก้ไข้มาลาเรีย" (Transition State = 'continue')

EDU6: "[พริกไทย] แก้อหิวาตกโรค" (Transition State = 'continue')
EDU7: "[ผู้อ่าน]ใช้ก้านพริกไทย 10 ก้าน" (Transition State = 'smooth shift')

งานวิจัยก่อนหน้า

ได้มีงานวิจัยมากมายที่ได้เสนอเทคนิคต่าง ๆเพื่อที่จะให้ได้มาซึ่งการสกัดความรู้สรรพคุณทางยาของสมุนไพร (Herbal Medicinal-Property Knowledge Extraction) โดยแบ่งออกเป็น 2 แนวทางคือ แนวทางสถิติ (Statistical Based Approach) และแนวทางผสมระหว่างแพทเทิร์นและสถิติ (Hybrid Approach: Pattern and Statistical Based Approach) ส่วนแนวทางการวิจัยเกี่ยวกับการการตอบคำถามอัตโนมัติสามารถแบ่งออกเป็น 2 แนวทางคือ แนวทางแพทเทิร์นหรือกฎ (Pattern/Rule Based Approach) และแนวทางสถิติ

การสกัดความรู้สรรพคุณทางยาของสมุนไพร

1. แนวทางสถิติ (Statistical Based Approach)

Weeber M. and Vos. R.(1998)ได้กล่าวถึงคุณสมบัติของฤทธิ์ยาจำเป็นต้องคำนึงถึง 3 เรื่องหลักคือ ยา(A) ผลที่แสดงออกทางกายภาพ (Physiological Effect)(B) และ โรค (C) และความสัมพันธ์ระหว่าง 3 เรื่องดังกล่าวเป็น A -> B , B -> C, ทำให้ได้ A -> C ดังนั้น Weeber M. and Vos. R.(1998) เสนอการสกัดความรู้ทางการแพทย์โดยการหาความสัมพันธ์ที่เป็นแอสโซวิเอชัน (Association) ระหว่างคำ(ซึ่งอยู่ในรูปของนามวลี, Noun Phrase (NP)) A, B, และ C จากเอกสารบทคัดย่อทางการแพทย์จำนวน 7,000 บทคัดย่อเกี่ยวกับยา captopril และ enalapril จาก MEDLINE บทคัดย่อที่มีเนื้อหาหรือคำเกี่ยวกับผลข้างเคียง (side effect) ถูกเลือกออกมาจากคลังข้อมูล 7,000 บทคัดย่อ โดยใช้แนวทางสถิติด้วยหาคำที่อยู่รอบๆ คำที่เป็น Side Effect ซึ่งเป็น seed ของแต่ละกรอบหน้าต่างขนาด 2,4,8, 16, 32,และ64 แล้วนำคำเหล่านี้มาหาความสัมพันธ์กัน หรือ Association ด้วยวิธีสถิติแบบดั่งเดิม เช่น log-likelihood ratio, G^2 , เพื่อหาว่าคำที่อยู่รอบseed มีความสัมพันธ์กับ seed อย่างมีนัยสำคัญ ดังนั้นจากการหาความสัมพันธ์ระหว่างคำโดย Expert I ได้ประเมินคำที่มีจำนวนคำทีปรากฏรอบside effect words เป็น151 คำเมื่อใช้กรอบหนาต่างขนาด 16 ได้ 1785คำแตกต่างกัน แต่มีเพียง442คำที่มีนัยสำคัญที่0.05 มีเพียง 46 คำที่แสดงความสัมพันธ์ที่เป็นแอสโซวิเอชันได้อย่างมีนัยสำคัญ ฉะนั้น recall เป็น 46/151= 0.31 และ precision = 46/442 = 0.01 ส่วน Expert II ได้ recall = 0.31 precision=0.03 ในขณะที่กรอบหน้าต่างขนาด 64 คำ Expert I ได้ recall = 0.19 precision = 0.14 Expert II ได้ recall = 0.24 precision = 0.07

Fang et al.,(2008) ได้ค้นพบความสัมพันธ์(Association Discovery) ระหว่างคำนามต่าง ๆที่เป็นชื่อยาสมุนไพรจีน โรค พันธุกรรม ผลกระทบ (Side Effect) ของยาสมุนไพรจีน และส่วนผสม โดยการวิเคราะห์การเกิดร่วมกัน (Collocation Analysis) จากเอกสารที่มีการกำกับ และมีการนำเอา IE (Information Extraction) และแบบจำลอง Swanson's ABC (A -> Bและ B -> C ทำให้ได้ ความสัมพันธ์แบบการส่งผ่าน (Transitive Association) คือ A -> C) มาประยุกต์ใช้ โดยกำหนดให้ A คือพันธุ์กรรม B คือ ส่วนผสมที่สามารถควบคุม A และ C คือ ยาสมุนไพรจีน เพื่อการบอกเป็นนัยของ A -> C เมื่อ A -> Bและ B -> C ปรากฏขึ้นในเอกสารอย่างมีนัยสำคัญ ผลการวิจัยของ Fang et al.,(2008) จาก38,072 MEDLINE abstracts ได้ 570 (TCM, effect) ความสัมพันธ์ (Associations) ที่ 97.5% confidence level ด้วยค่า precision เป็น 96.5%. อย่างไรก็ตามวิธีของ Fang et al.,(2008) อยู่บนพื้นฐานของการใช้แต่เพียงนามวลี

2. แนวทางแพทเทิร์นหรือกฎ (Pattern/Rule Based Approach) ร่วมกับแนวทางสถิติ (Statistical Based Approach)

Paşca M. (2008) ระบุความรู้ที่เป็นจริงเกี่ยวกับคลาสวัตถุ (Object Class) ต่างๆที่เป็นนามวลีได้โดยการใช้ อิสอะแพทเทิร์น (Is-A pattern) กับ100ล้านเอกสารและการสอบถามหรือคิวรี (Query) จำนวน 50คิวรี เอกสารทั้งหมดเป็นภาษาอังกฤษและ ดาวน์โหลดจากเว็บ ผ่านการสกัดวลีออกมาขณะเดียวกันหาคลาสวัตถุโดยการมายนิ่ง (Mining) หากลุ่มคำที่เป็นอิสอะแพทเทิร์นและมีความถี่มากมาเป็นคลาส เช่น "... are commonwealth countries", "...are asia pacific countries" จะได้country เป็นคลาส นำคลาสที่สกัดได้มาทำการหาแอททริบิวท์ (Attribute หรือคุณสมบัติของคลาส) เช่นคลาส "Movie" มีอินสแตนซ์ (instance) คือ "jay and silent bob strike back" "kill bill" เป็นต้น จากคลาสที่ได้มาทำการคิวรีเพื่อหาแคนดิเดทแอททริบิวท์ (Candidate Attribute) เช่น อินสแตนซ์ "jay and silent bob strike back" มีคิวรี่ "cast jay and silent bob strike back" ทำให้ได้ "cast" เป็น แคนดิเดทแอททริบิวท์ ต่อมาสร้างอินเตอร์เสิร์ชซิกเนเจอร์เวคร์เตอร์ (Internal Search-Signature Vector) ด้วยเทมเพลทคิวรี (Template Query : X for Y, Xเป็นเป็นแคนดิเดทแอททริบิวท์ Yเป็นอินสแตนซ์) ตัวอย่างเช่น "cast for kill bill" ซึ่ง "cast" เป็นแคนดิเดทแอททริบิวท์ "kill bill" เป็นอินสแตนซ์ สำหรับแทนแต่ละแคนดิเดทแอททริบิวท์ ฉะนั้นสามารถหาความถี่ของแคนดิเดทแอทริบิวท์จากเวคร์เตอร์ที่ได้เหล่านี้ทั้งหมด ต่อมาหาแร็ง (Rank) ของแคนดิเดทแอททริบิวท์ทั้งหมดของแต่ละคลาสโดยการหาค่า ซิมมิลาริตีสคอร์ (Similarity Scores) ระหว่างเวคร์เตอร์ที่แทนแคนดิเดทแอททริบิวท์ (Individual Vector Representations) และเวคร์เตอร์อ้างอิงของซีดแอททริบิวท์ (Reference Vector of the seed attributes) ผลลัพธ์ที่ได้คือลิสต์ของแอททริบิวท์ที่ได้ถูกแร็งสำหรับคลาสหนึ่ง เช่น คลาส "cast" มี [opening song, cast,...]เป็นลิสต์ของแอททริบิวท์นั้น เป็นต้น ทำให้สามารถสกัดแอททริบิวท์ได้ถูกต้องด้วยค่า precisionคือ 0.8 สำหรับ 100 คลาสกับ5 ซีดแอททริบิวท์

การตอบคำถามความรู้สรรพคุณทางยาของสมุนไพร

1. แนวทางแพทเทิร์นหรือกฎ (Pattern/Rule Based Approach)

Riloff E. and Thelen M.(2000) ได้ใช้กฎเป็นพื้นฐาน (Rule Base)ต่างๆพร้อมกับการให้คะแนน สำหรับระบบตอบคำถามอย่างอัตโนมัติ กับคำถาม (Question Word) คือ "Who" "What" "When" "Where" และ "Why" หลังจากผ่านซอฟท์แวร์แจงประโยค ทั้งนี้เพื่อทดสอบความเข้าใจจากการอ่านบทความภาษาอังกฤษ โดยระบบอัตโนมัติให้คะแนนสำหรับประโยคที่มีคำตรงกับคำในประโยคคำถาม ถ้าประโยคใดมีคะแนนสูงประโยคนั้นคือประโยคคำตอบ แต่สำหรับคำถาม "What" กฎที่ใช้มีลักษณะดังนี้ "What occur...on the date.." "What kind..." "What is the name of ..<proper noun>" "What is it called.." และ"What is it made from.." ได้ความถูกต้องสำรับคนตอบเป็น 0.31 สำหรับระบบอัตโนมัติเป็น 0.28 สำหรับการตอบคำถาม "What" อย่างไรก็ตามกฎของ คำถาม "What"เหล่านั้นไม่สามารถใช้กับงานวิจัยนี้ เพราะรูปแบบของคำถาม "What.." นี้ไม่สามารถครอบคลุมรูปแบบของคำถาม "What.."ทั้งหมดที่ปรากฏในงานวิจัยนี้เช่น โหระพามีสรรพคุณอะไรบ้าง โหระพามีสรรพคุณอย่างไร จงแสดงสรรพคุณต่างๆของพืชสมุนไพร: โหระพา เป็นต้น ทั้งนี้เพราะรูปแบบลักษณะภาษาของประโยคคำถามในภาษาอังกฤษต่างจากในภาษาไทย ตัวอย่างเช่น ส่วนที่เป็นคำถาม "What/อะไร" ของภาษาอังกฤษจะอยู่ที่ส่วนหัวของประโยคคำถาม สำหรับภาษาไทยจะอยู่ที่ส่วนท้ายของประโยคคำถาม นอกจากนี้คำถาม "How/อย่างไร" ในภาษาไทยมีความหมายเป็นเป็นคำถาม "What/อะไร" ตัวอย่างเช่น โหระพามีสรรพคุณอย่างไร

2. แนวทางสถิติ (Statistical Based Approach)

Quaresma P. and Rodrigues I.(2005) ได้ เสนอระบบการตอบคำถามที่มีการใช้ซอฟท์แวร์แจงประโยคภาษาโปรตุเกส (Protuguese Parser) กับเอกสารทางคดีความและประโยคคำถามที่ต้องการคำตอบที่เป็นความรู้

เกี่ยวกับคดีที่มีลักษณะเหมือนคดีในอดีตที่มีการตัดสินผิดพลาด ฉะนั้นคำถามที่ศึกษาจะเป็นคำถามเกี่ยวกับสถานที่ ("Where") วัน ("When") นิยาม ("What is..") และเฉพาะเรื่อง ("How many time.." โดยนำคำถามเหล่านั้นมาผ่านตัวแจงประโยค (Parser) แล้วแทนคำถามนั้นด้วย Predicate เพื่อสามารถทำ ยูนิไฟด์ (Unify) ระหว่างคำถามที่ได้อยู่ในรูปของPredication กับประโยคต่างๆในเอกสารต่างๆที่ได้จากเว็บไซด์ด้วยเทคนิค IR (Information Retrieval) แล้วผ่านตัวแจงประโยค พร้อมกับการกำกับความหมายจาก Ontology และแทนประโยคเหล่านั้นด้วยภาษา Predicate ได้ความถูกต้อง 25% จาก 200 คำถาม

Fan S. et. al., (2008) ได้เสนอแบบจำลอง CRF (Conditional Random Field Model) สำหรับระบบการตอบคำถามที่มีการกำกับความหมายระดับก้อน (Chunk Semantic) ออกเป็น 4chunks เช่น "Topic" (the question subject), "Focus" (the additional information of topic), Restrict (เช่น time restriction, location restriction), Rubbish information (words no meaning for the question), และอื่นๆ ให้กับคำถาม ซึ่งคำถามนี้จะถูกนำไปหาค่าความคล้าย (Similarity) จากค่า Information Gain กับคำถามที่มีคู่คำตอบ (Question-Answer Pair) ซึ่งได้จากBlog ต่างๆบนเว็บไซต์ภาษาจีนจำนวน 14000ประโยคคำถาม CRFคล้ายกับ Maximum Entropy(ME) ต่างกันที่ MEใช้ค่าคงที่ที่ทำให้เป็นมาตรฐาน (Normalization Constant) เพียงตัวเดียว ในขณะที่ CRF ใช้หลายตัว CRF ใช้สำหรับเลือกเซทฟีเจอร์(Feature Set) จากฟีเจอร์ต่างๆดังนี้ คำที่อยู่ในเซทที่กำหนดความหมายไว้ , POS tag, Question Pattern, Question type, Pattern Key word, และ Pattern tag เพื่อใช้หาค่าความคล้าย ซึ่งได้ค่าความถูกต้องเฉลี่ย 93.07% precision 93.07%recall

ปัญหาการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทยเพื่อสนับสนุนระบบการตอบคำถามอัตโนมัติ

เนื่องจากงานวิจัยนี้มีเป้าหมาย 2 ประการหลัก คือ ศึกษาวิธีการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของเอนติตี้พืชสมุนไพรไทยจากเอกสารภาษาไทย และศึกษาระบบสอบถามเกี่ยวกับสรรพคุณทางยาของสมุนไพรไทยด้วยคำถามประเภท "อะไรบ้าง(What-question)" ทำให้เกิดแนวทางปัญหาหลัก 2 ทางที่ต้องศึกษาคือ ปัญหาการสกัดความรู้สรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทย และปัญหาจากระบบการตอบคำถามเกี่ยวกับคุณสมบัติของเอนติตี้สมุนไพรไทย

<u>ปัญหาการสกัดความรู้สรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทย</u>

ปัญหาในส่วนของการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของสมุนไพรไทยจากเอกสารภาษาไทย ประกอบด้วยสามปัญหาคือ ปัญหาการระบุเอนติตี้สมุนไพรไทย ปัญหาการระบุสรรพคุณทางยาของสมุนไพรไทย ปัญหาการหาขอบเขตของ EDUs สรรพคุณทางยาของสมุนไพรไทย

1. ปัญหาการระบุเอนติตี้สมุนไพรไทย สามารถแบ่งออกเป็นสองปัญหาย่อยคือ

1.1 ปัญหาการละนามที่อ้างอิง (Zero Anaphora Problem) ดังตัวอย่างต่อไปนี้

EDU1 "กระเทียมใช้เป็นยาขับลม"

EDU2 " ϕ แก้ไอ"

ซึ่ง ϕ แทน Zero Anaphora ที่อ้างอิงถึง กระเทียม

1.2 ปัญหาการละข้อความ (Textual Ellipsis Problem) ดังตัวอย่างต่อไปนี้

"ทับทิม/ Pomegranate

.....

EDU1: "ราก [ทับทิม] ใช้เป็นยาขับปัสสาวะ"

....."

ซึ่ง [..] หมายถึงการละ อักขระหรือข้อความใดๆที่อยู่ภายในเครื่องหมายวงเล็บก้ามปู

2. ปัญหาการระบุสรรพคุณทางยาของสมุนไพรไทย ดังตัวอย่างต่อไปนี้

EDU1: "ต้มน้ำใบบัวบก"

EDU2: "แช่เย็น"

EDU3: "แล้วดื่ม"

EDU4: "ช่วยลดไข้"

EDU5: "แก้เจ็บคอ/"

ฉะนั้นเครื่องจะทราบได้อย่างไรว่า EDU4 และ EDU5 คือสรรพคุณทางยา

3. ปัญหาการหาขอบเขตของ EDUs สรรพคุณทางยาของสมุนไพรไทย สามารถแบ่งออกเป็นสองปัญหาย่อยคือ

3.1 ปัญหาการละกริยา (Verb Ellipsis Problem) ดังตัวอย่างต่อไปนี้

EDU1:"กระเพราแก้ปวดท้อง"

EDU2: "[กระเพรา] [แก้] ท้องเสีย"

EDU3: "และ [กระเพรา] [แก้] คลื่นไส้"

3.2 ปัญหาการละคูบอกขอบเขตสิ้นสุด (Ending-Boundary-Cue Ellipsis) ดังตัวอย่างต่อไปนี้

(เมื่อ Ending-Boundary-Cue = {"และ" "ในที่สุด"...})

EDU1: "ขมิ้นใช้เป็นยาลดกรด"

EDU2: "[ขมิ้น] ขับ/releases ลม/gas"

EDU3: "[ขมิ้น] แก้ปวดท้อง"

EDU4: "[ขมิ้น] คลายอาการปวดเกร็งช่องท้อง"

EDU5: "การใช้ขมิ้นเป็นที่นิยมมาก...."

นั่นคือเครื่องจะทราบได้อย่างไรว่า EDU4 คือ ขอบเขตการสิ้นสุด (Ending Boundary) ของกลุ่ม EDU สรรพคุณทางยา

ฉะนั้นงานงานวิจัยนี้ขอเสนอการใช้แพทเทิร์น ทางภาษาศาสตร์หรือที่เรียกว่า "Lexico Syntactic Pattern" คือ NP1 V_{mp} NP2 (Girju R. and Moldovan D., 2002), มาเป็นตัวระบุ EDU ที่มีความหมายเป็นสรรพคุณทางยาของสมุนไพรไทย หลังจากที่ได้แก้ปัญหาการละนามที่อ้างอิงโดยใช้นามหรือนามวลี (NP1) ก่อนหน้าที่ไม่ได้ละมาแทนที่ และปัญหาการละข้อความโดยใช้ชื่อหัวเรื่อง (Topic Name) ดังนั้นหลังจากที่ EDUแรกของลำดับ EDU สรรพคุณทางยาของสมุนไพรไทยได้ถูกรู้จัก ปัญหาต่อมาคือการหาขอบเขตของ EDUs สรรพคุณทางยาของสมุนไพรไทย โดยงานวิจัยนี้ขอเสนอวิธีการแก้ปัญหาการหาขอบเขตนี้ด้วยวิธีที่แตกต่างกันสองวิธี คือ วิธีการใช้Naive Bayesทดสอบคู่ กริยา v_{mp} หรือ v_{mp} pair เมื่อ $v_{mp} \in V_{mp}$ ของคู่ EDU ที่อยู่ติดกันในหนึ่งกรอบหน้าต่างพร้อมทั้งเลื่อนกรอบหน้าต่างไปด้วยระยะทางหนึ่ง EDU ว่ามีความหมายเป็นสรรพคุณทางยาของสมุนไพรไทย ถ้าหากไม่มีความหมายเป็นสรรพ คุณทางยาสมุนไพรไทยก็ถือว่าขอบเขตได้สิ้นสุด (หลังจากที่ได้แก้ปัญหาการละกริยาโดยใช้กริยาก่อนหน้า) และวิธีการใช้ ทฤษฎีเซนเทอร์ริง (ซึ่งเป็นวิธีทางภาษาสาสตร์) กล่าวคือเมื่อไรก็ตามเกิดสถานะ Smooth Shift หมายถึงขอบเขตของ EDU ที่มีความหมายเป็นสรรพคุณทางยาสมุนไพรไทยได้สิ้นสุด

<u>ปัญหาจากระบบการตอบคำถามเกี่ยวกับคุณสมบัติของเอนติตี้สมุนไพรไทย</u>

ปัญหาระบบสอบถามเกี่ยวกับคุณสมบัติของเอนติตี้สมุนไพรไทย ประกอบด้วยสามปัญหาหลักคือ

1. ปัญหาการระบุคำถาม เนื่องจากไม่มี "เครื่องหมายคำถาม" ในภาษาไทย ทำให้ยากต่อการระบุว่าประโยคต่อไปนี้เป็นคำถาม ฉะนั้นแก้ไขโดยการใช้ "Question Word: "อะไร/What" "ที่ไหน/Where" "เมื่อไร/When" "ทำไม/Why" "อย่างไร/How" "ลิสต์/List" "แสดง/Show" เป็นต้น

2. ปัญหาความกำกวมของ Question Word เช่น "อย่างไร/How" มีความหมายเป็นการถาม "อะไร/What" ตัวอย่างเช่น

คำถาม1: "ใบโหระพามีสรรพคุณทางยาอย่างไร"

นอกจากนี้ Question Word "อะไร/What" ต้องการคำตอบที่แตกต่างกัน 3 แบบดังนี้

คำถาม2: "พืชสมุนไพรอะไรมีสรรพคุณขับลม" (ต้องการคำตอบเกี่ยวกับคุณสมบัติหรือสรรพคุณ)

คำถาม3: "สมุนไพรคืออะไร" (ต้องการคำตอบที่เป็นนิยาม)

คำถาม4: "อะไรคือสาเหตุของโรค" (ต้องการคำตอบที่เป็นเหตุ)

3. ปัญหาการระบุโฟกัสของคำถาม

คำถาม1: "ใบโหระพา<u>มีสรรพคุณทางยาอะไรบ้าง</u>"

คำถาม2: "<u>พืชสมุนไพรอะไร</u>มีสรรพคุณขับลม"

คำถาม3: "สมุนไพร<u>คืออะไร</u>"

คำถาม4: "<u>อะไรคือ</u>สาเหตุของโรค"

ซึ่งบริเวณที่ขีดเส้นใต้คือโฟกัสของคำถาม นั่นคือเครื่องคอมพิวเตอร์จะทราบได้อย่างไร ซึ่งโฟกัสที่ได้จะถูกนำไปหาคู่คำตอบจากฐานความรู้สมุนไพรที่สกัดได้

ฉะนั้นงานวิจัยนี้จึงขอเสนอการใช้แพทเทิร์นของคำถามชนิด “คำถามอะไร/what” ประเภทลิสต์“อะไรบ้าง” และประเภทเอนติตี้ “x อะไร” (เมื่อ x คือเอนติตี้) ร่วมกับ NP1, NP2, และ V_{mp} มาทำการระบุประเภทคำถามและระบุโฟกัสของคำถาม เพื่อหาคำตอบจากความรู้ที่สกัดได้นั้น โดยมีแพทเทิร์นดังนี้
(จากบทนำ เมื่อกำหนดให้ $np1_i \in NP1$ (i=1,2,..,m), $np2_l \in NP2$ (l=1,2,..,h), และ $v_{mp} \in V_{mp}$)
ประเภทลิสต์ มีทั้งหมด 5 แพทเทิร์นดังนี้

- “ลิสต์” + “สรรพคุณ” + [“ของ”] + $np1_i$
- [“แสดง | บอก”] + “สรรพคุณ” + [“ของ”] + $np1_i$ + “มี” + “อะไรบ้าง”
- $np1_i$ + “มี” + “สรรพคุณ” + “อะไรบ้าง/อย่างไรบ้าง”
- “ลิสต์ชื่อสมุนไพร” + [“มีสรรพคุณ”] + v_{mp} +$np2_l$
- [“แสดง | บอกชื่อ”] + “สมุนไพร” + “อะไรบ้าง” + [“มีสรรพคุณ”] + v_{mp} +$np2_l$

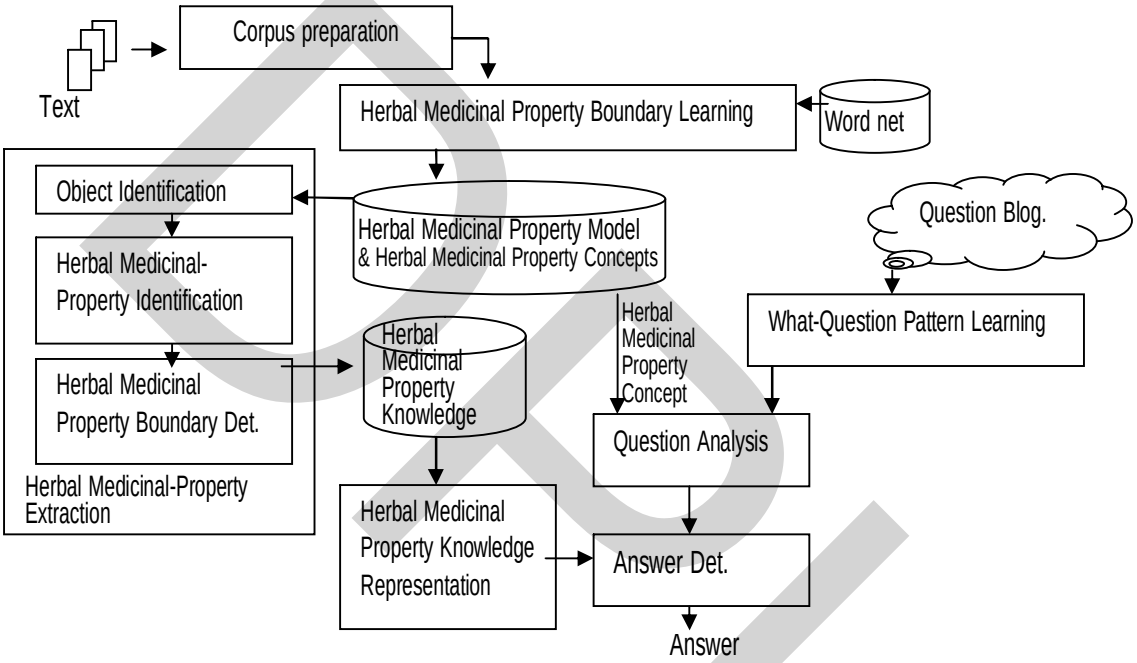
ประเภทเอนติตี้ “x อะไร” มีทั้งหมด 2 แพทเทิร์นดังนี้

- [“แสดง | บอกชื่อ”] + “สมุนไพร” + “อะไร + [“มีสรรพคุณ”] + v_{mp} + $np2_l$
- $np1_i$ + “มี” + “สรรพคุณ” + “อย่างไร”

จากปัญหาที่กล่าวมาข้างต้นทั้งหมด คือ ปัญหาการสกัดความรู้สรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทย และปัญหาจากระบบการตอบคำถามเกี่ยวกับคุณสมบัติของเอนติตี้สมุนไพรไทย ทำให้เกิดแนวทางแก้ไขปัญหาดังกล่าวด้วยวิธีผสมผสานระหว่างการเรียนรู้ของเครื่อง (Machine Learning) กับการประมวลผลภาษาธรรมชาติ (Natural Language Processing, NLP) พร้อมกับการศึกษาพฤติกรรมทางภาษาของโดเมนพืชสมุนไพร ดังแสดงรายละเอียดกรรมวิธีการแก้ไขปัญหาดังกล่าวในหัวข้อถัดไปคือ “กรรมวิธีดำเนินงาน”

กรรมวิธีดำเนินงาน
Method

ระบบงานโดยสรุปสำหรับการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทยเพื่อสนับสนุนระบบการตอบคำถามอัตโนมัติประกอบด้วยขั้นตอนต่างๆดังต่อไปนี้ (ดังแสดงในรูปที่3) แบ่งออกเป็นสองส่วนคือ ส่วนสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรจากเอกสารภาษาไทยและส่วนการตอบคำถามชนิด "ถามอะไร"ประเภทลิสต์ และประเภทเอนติตี้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพร



รูปที่3 ระบบงานโดยสรุป

ส่วนสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรจากเอกสารภาษาไทย

การสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรจากเอกสารภาษาไทย ประกอบด้วยขั้นตอนหลักๆ ดังนี้

1. การเตรียมคลังข้อมูล (Corpus Preparation)

ก่อนที่จะมีการเตรียมคลังข้อมูลจะต้องมีการศึกษาพฤติกรรมทางภาษาของชุดข้อมูลที่ได้จากเอกสารสมุนไพรไทยดาวโหลดจากเว็บไซต์กรมส่งเสริมการเกษตร www.doae.go.th และโครงการอนุลักษณ์พันธ์พืชฯ http://www.rspg.or.th/index.html ว่ามีความเหมาะสมที่จะใช้เป็นคลังข้อมูลสำหรับงานวิจัยนี้ ประโยคต่างๆที่ปรากฏอยู่ในชุดข้อมูลดังกล่าวเราจะสนใจเฉพาะประโยคธรรมดา (Simple Sentence หรือ EDU) ซึ่งสามารถแสดงได้ด้วย Regular Express ดังนี้ NP1 V_{mp} NP2 เมื่อ NP1 คือ เซทของนามวลีที่เป็นเอนติตี้สมุนไพรหรือส่วนประกอบของเอนติตี้สมุนไพร V_{mp} คือเซทความคิดกริยาทีแสดงให้เห็นสรรพคุณทางยา(Medicinal-Property Verb Concept)

และ NP2 คือเซทความคิดนามวลีเกี่ยวกับอาการ, โรค, และ เชื้อโรค (ตามที่ได้กล่าวไว้ในบทนำ) อย่างไรก็ตาม งานวิจัยนี้สนใจสรรพคุณทางยาที่แสดงให้เห็นหรือเข้าใจได้ด้วยเหตุการณ์ (Event) ซึ่งสามารถแทนได้ด้วยกริยา v_{mp} เมื่อ $v_{mp} \in V_{mp}$

V_{mp}= {"รักษา/cure" "บรรเทา/relieve" "แก้/stop,prevent" "ขับ/release" "ลด/reduce" "เพิ่ม/increase" ...}

ดังแสดงสรุปได้ในตารางที่1 (จากข้อมูลตัวอย่างสุ่มจำนวน500 EDUsจากคลังข้อมูลที่ดาวโหลดมา)

ตารางที่1 แสดงพฤติกรรมทางภาษาของคำกริยาที่มีความคิดเป็นสรรพคุณทางยา(Medicinal-Property Verb Concept) และไม่มีความคิดเป็นสรรพคุณทางยา(Non Medicinal-Property Verb Concept) จาก Surface Form เดียวกัน จาก คลังข้อมูล500 EDUs

Medicinal-Property Verb Concept v_{mp}	Number of Occurrences (Surface Form)	Number of Medicinal-Property-Concept Occurrences	Number of Non-Medicinal-Property-Concept Occurrences
แก้(Stop, Relief)	68	60	8
ขับ (Discharge, Release)	25	23	2
บรรเทา (Relief)	5	5	-
รักษา (Treat, Cure)	11	8	3
ลด (Decrease)	15	9	6
ถ่าย (Excrete)	5	3	2
บำรุง (Norish)	10	6	4

จากตารางที่1จะเห็นว่า v_{mp} ที่ปรากฏส่วนใหญ่มีความหมายหรือแนวความคิดเป็นความคิดกริยาทีแสดงให้เห็นสรรพคุณทางยา (Medicinal-Property Concept) ดังนั้นคลังข้อมูลที่ดาวโหลดมานี้สามารถใช้ในงานวิจัยนี้ โดยเริ่มจากการเตรียมคลังข้อมูลประมาณ 3000 EDUs (2000 EDUs สำหรับการเรียนรู้ EDUs สรรพคุณทางยาของพืชสมุนไพรไทย 1000สำหรับการทดสอบและประเมินผล) ข้อมูลที่เป็นเอกสารเหล่านี้ก่อนที่จะนำมาประมวลผลทางภาษาธรรมชาติจะต้องผ่านขั้นตอนการตัดคำโดยการใช้ซอฟท์แวร์ตัดคำภาษาไทยที่สามารถแก้ปัญหาขอบเขตคำ และขณะเดียวกันสามารถกำกับหน้าที่ของคำ (Part of Speech) ได้ (Sudprasert and Kawtrakul, 2003) ซึ่งรวมถึงการหา Name Entity (Chanlekha and Kawtrakul, 2004), และการรับรู้คำ (Word-Formation Recognition) (Pengphon et al., 2002) เพื่อที่จะแก้ปัญหาขอบเขตของ Thai Name Entity และนามวลี หลังจากนั้นต้องทำการตัดประโยคในระดับ EDU ด้วยวิธีการของ (Chareonsuk et al., 2005) และสุดท้ายทำการกำกับ EDUs ที่เป็นความรู้สรรพคุณทางยาของพืชสมุนไพร (Medicinal Property Tag) ที่ได้ออกแบบดังแสดงในรูปที่4

“พริกไทยดำ……
พริกไทยดำมีรสเผ็ดอุ่น EDU1เมื่อรับประทานเข้าไป EDU2จะรู้สึกอุ่นวาบที่ท้อง EDU3 ช่วยขับลม EDU4ขับปัสสาวะ EDU5แก้ท้องอืดท้องเฟ้อ EDU6แก้ไข้มาลาเรีย EDU7แก้อหิวาตกโรค, EDU8ใช้ก้านพริกไทย 10 ก้าน”

<Topic_name Entity-concept=black pepper/herb>พริกไทยดำ</Topic_name>……

<EDU1> เมื่อ <NP1 concept=person>ϕ</NP1>
<VP><Vmp concept=consume>รับประทาน </Vmp>เข้าไป</EDU>

<id =1 class=Medicinal Property>
<EDU2><NP1 concept= black pepper/herb>ϕ</NP1>
<VP>จะ<Vmp concept=be warm>รู้สึกอุ่นวาบ </Vmp>ที่ท้อง</EDU>
<EDU3><NP1 concept= black pepper/herb>ϕ</NP1>
<VP><Vmp concept=release>ช่วยขับ </Vmp>
<NP2 concept=gas>ลม</NP2></VP></EDU>
<EDU4><NP1 concept= black pepper/herb>ϕ</NP>
<VP>< Vmp concept=discharge/release >ขับ </ Vmp >
<NP2 concept=urine>ปัสสาวะ</NP2></VP></EDU>
<EDU5><NP1 concept= black pepper/herb>ϕ</NP1>
<VP>< Vmp concept=stop>แก้ </ Vmp >
<NP2 concept=flatulence/symptom >ท้องอืดท้องเฟ้อ</NP2></VP></EDU>
<EDU6><NP1 concept= black pepper/herb>ϕ</NP1>
<VP>< Vmp concept=cure>แก้ </ Vmp >
<NP2 concept= malaria >ไข้มาลาเรีย</NP2></VP></EDU>
<EDU7><NP1 concept= black pepper/herb>ϕ</NP1>
<VP>< Vmp concept=cure>แก้ </ Vmp >
<NP2 concept= cholera >อหิวาตกโรค</NP2></VP></EDU>
</id>
Vmp is the medicinal property verb tag

รูปที่4 ตัวอย่างการกำกับ EDUที่เป็นความรู้สรรพคุณทางยาของพืชสมุนไพร

2. การเรียนรู้ขอบเขตของสรรพคุณของพืชสมุนไพร (Herbal Medicinal Property Boundary Learning)

จากคลังข้อมูลที่ได้กำกับความหมายสรรพคุณทางยาของพืชสมุนไพรในระดับ EDU ของขั้นตอนก่อนหน้านี้ ทำการสกัดลักษณะเฉพาะหรือฟีเจอร์ต่างๆ (Features) ที่เป็นทั้งแนวความคิด (Concept) และเซอร์เฟสฟอร์ม (Surface Form) ของกริยาที่มีความคิดเป็นสรรพคุณทางยาของพืชสมุนไพร รวมไปถึงนามวลีที่อยู่ก่อนหน้าและหลังกริยา เก็บเป็นฐานข้อมูลความคิดทางยาของพืชสมุนไพรของ v_{mp} (เมื่อ $v_{mp} \in V_{mp}$) , NP1, และ NP2 ดังแสดงในตารางที่2

ตารางที่2 แสดงฟีเจอร์ที่เป็นกริยาแสดงสรรพคุณทางยาของพืชสมุนไพร V_{mp} รวมทั้งสารสนเทศ(นามวลี)ที่อยู่รอบๆกริยา

no	ID	NP1/ concept	V_{mp}/ concept	NP2/ concept	class
1	001	พริกไทยดำ/black pepper	รู้สึกอุ่น/be warm	-	yes
2	001	พริกไทยดำ/black pepper	ช่วยขับ/ release	ลม/gas	yes
3	001	พริกไทยดำ/black pepper	ขับ/ release	ปัสสาวะ /urine	yes
4	002	พริกไทยดำ/black pepper	แก้ /stop	ท้องอืดท้องเฟ้อ /flatulence,symptom	yes
5	002	พริกไทยดำ/black pepper	แก้ /cure	ไข้มาลาเรีย /malaria disease	yes
6	003	พริกไทยดำ/black pepper	แก้ /cure	อหิวาตกโรค /cholera disease	No
7	004	ฟ้าทะลายโจร/ Andrographis paniculata	บรรเทา/ relief	อาการเจ็บคอ/sore throat Symptom	yes
8	004	ฟ้าทะลายโจร/ Andrographis paniculata	บรรเทา/ relief	อาการหวัด /flu symptom	yes
9	004	ใบฟ้าทะลายโจร/ Andrographis paniculata Leaf	แก้ /stop	ท้องเสีย /diarrhea	No
..	....	....	...	...	...

ซึ่งถูกนำไปใช้สำหรับการเรียนรู้ขอบเขตของ EDUs สรรพคุณทางยาของพืชสมุนไพรไทย ด้วยการหาความถี่ของ v_{mp} pair ของคู่ EDU ที่อยู่ติดกันในหนึ่งกรอบหน้าต่างพร้อมทั้งเลื่อนกรอบหน้าต่างไปด้วยระยะทางหนึ่ง EDU ว่ามีความหมายเป็นสรรพคุณทางยาของสมุนไพรไทยหรือไม่จากคลังข้อมูลที่ได้กำกับความหมายสรรพคุณทางยาของพืชสมุนไพรในระดับ EDU ของขั้นตอนก่อนหน้านี้ดังที่ได้แสดงในตารางที่3

ตารางที่3 แสดงค่าความน่าจะเป็นของ v_{mp} จาก v_{mp} pair คือ $v_{mp\_at\_ij}$ และ $v_{mp\_at\_ij+1}$ ที่มีความคิดเป็นสรรพคุณทางยาของพืชสมุนไพร และไม่มีความคิดเป็นสรรพคุณทางยาของพืชสมุนไพร

$v_{mp\_at\_ij}$	Medicinal Property Verb Concept	Non Medicinal Property Verb Concept
stop	0.4110	0.1731
release	0.1507	0.1346
relief	0.0069	0.0385
treat	0.1367	0.0096
discharge	0.0137	0.0385
be-drug	0.0205	0.0096
...	...	...
$v_{mp\_at\_ij+1}$	Medicinal Property Verb	Non Medicinal Property Verb.
stop	0.375	0.1091
release	0.1447	0.0273
reduce	0.0197	0.0091
treat	0.0132	0.0182
discharge	0.0263	0.0091
be-drug	0.0132	0.0182
	....	.....

3. การสกัดความรู้สรรพคุณทางยาของพืชสมุนไพร (Herbal Medicinal Property Extraction)

ขั้นตอนนี้แบ่งออกเป็นสองส่วนคือ ส่วนการระบุวัตถุหรือเอนติตี้สมุนไพรและการระบุสรรพคุณทางยาของพืชสมุนไพร และส่วนขั้นตอนการหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพร

3.1 การระบุวัตถุและการระบุสรรพคุณทางยาของพืชสมุนไพร (Object Identification and Herbal Medicinal-Property Identification)

ขั้นตอนนี้เป็นการระบุวัตถุหรือเอนติตี้สมุนไพร โดย ใช้ชื่อเอกสารเป็นTopicName (จากบทนำ) $np1_i$=part+TopicName

$Np1_i \in NP1 \quad (i-1,2,..,m)$

ต่อมาระบุ EDU ที่แสดงสรรพคุณทางยา โดยสแกนไปตามลำดับของ EDU ที่ปรากฏบนเอกสารพืชสมุนไพรไทยใช้เพื่อค้นหา Lexico Syntactic Pattern (NP1 V_{mp} NP2) หากพบให้ดำเนินการในขั้นตอนถัดไป หากไม่พบให้ระบุวัตถุและระบุ EDU ที่แสดงสรรพคุณทางยากับเอกสารใหม่เช่นนี้ไปเรื่อยๆ

3.2 การหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพร

ขั้นตอนนี้สามารถทำได้สองวิธีที่แตกต่างกัน

ก. การหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรสามารถหาได้โดยการหาค่าargmax จากสมการ Na?ve Bayesต่อไปนี้ กับค่าความน่าจะเป็น ของ $v_{mp\_at\_ij}$ และ $v_{mp\_at\_ij+1}$ ในตารางที่ 2 ดังแสดงในอัลกอริทึมของรูปที่ 5

$$HerbalMedicinalPropertyBoundaryClass = \arg\max_{class \in Class} P(class \mid v_{mp\_at\_ij}, v_{mp\_at\_ij+1})$$
$$= \arg\max_{class \in Class} P(v_{mp\_at\_ij} \mid class) P(v_{mp\_at\_ij+1} \mid class) P(class) \quad (5)$$

where $v_{mp\_at\_ij} \in V_{mp}$ *and* $v_{mp\_at\_ij+1} \in V_{mp}$ (V_{mp} *is a medicinal_ property_ verb_ concept set*)

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

```
Assume that each EDU is represented by (NP1 Vmp NP2).   L is a list
of EDU. Vmp is the medicinal-property verb concept set. NP1 is the
herbal noun phrase concept set. NP2 is the symptom/disease noun
phrase concept set.
vij or vmp_at_ij, vij+1 or vmp_at_ij+1 are learned verbs as elements of the Vmp set
HERBAL_MEDICINAL_PROPERTY_EXTRACTION1 ( L, Vmp,NP1,NP2 )
 1  i ← 1, j←1 R←∅     MEDPROPi← ∅
 2  while i ≤ length[L] do
 3  begin_while1
 4   If np1i∈NP1 ^ v ij∈Vpmp^np2i∈NP2  /*find the medicinal property EDU
 5      bd=yes ;  MEDPROPi ←  MEDPROPi ∪ {j}
 7      while( v ij ∈Vmp )∧( v ij+1∈Vmp)∧ bd=yes do
 8      begin_while2          /* Boundary determination
 9       bd= arg max_{c∈(yes,no)} P(v_ij | c) P(v_ij+1 | c) P(c)
10         if bd = yes   then
11             MEDPROPi ← MEDPROPi ∪ {j+1}; j=j+1
13      end_while2
14      R = R ∪ {MEDPROPi }
15      j=1;  i=i+1
16     end_while1
17 · Return
```

รูปที่ 5. อัลกอริทึม Medicinal Property Boundary Extractionโดย Naive Bayes

ข. การหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรโดยการใช้ทฤษฎีเซนเทอร์ริง กล่าวคือให้สแกน $np1_i$ ของคู่ EDU ที่อ

```
Assume that each EDU is represented by (NP1 Vmp NP2).   L is a list of EDU.
Vmp is the medicinal-property verb concept set. NP1 is the herbal noun phrase
concept set. NP2 is the symptom/disease noun phrase concept set.
vij or vmp_at_ij, vij+1 or vmp_at_ij+1 are learned verbs as elements of the Vmp set
HERBAL_MEDICINAL_PROPERTY_EXTRACTION2 ( L, Vmp,NP1,NP2 )
  1  i ← 1, j←1 R←∅     MEDPROPi← ∅
  2  while i ≤ length[L] do
  3  begin_while1
  4   If np1i∈NP1 ^ v ij∈Vpmp^np2i∈NP2  /*find the medicinal property EDU
  5      bd=yes ;  MEDPROPi ←  MEDPROPi ∪ {j}
  7      while( v ij ∈Vmp )∧( v ij+1∈Vmp)∧ (np1ij= np1ij+1) do
  8      begin_while2          /* Boundary determination
  9
 10            MEDPROPi ← MEDPROPi ∪ {j+1}; j=j+1
 11      end_while2
 12      R = R ∪ {MEDPROPi }
 13      j=1;  i=i+1
 14     end_while1
 15 · Return
```

รูปที่ 6. อัลกอริทึม Medicinal Property Boundary Extractionโดย Centering Theory

4. การแทนความรู้สรรพคุณทางยาของพืชสมุนไพร (Herbal Medicinal Property Knowledge Representation)

นำความรู้สรรพคุณทางยาของพืชสมุนไพรที่สกัดได้จากขั้นตอนก่อนหน้านี้มาเก็บในฐานข้อมูลในรูปแบบของฐานข้อมูลเชิงวัตถุที่มี แทททริบิวดังนี้

ID= ระหัสสมุนไพร
X=ชื่อสมุนไพร หรือ TopicName
Y=ส่วนของพืช หรือ part
$np1_i$= part+TopicName เมื่อ i=1,2,..,m
mp_j=สรรพคุณ หรือ V_{mp} NP2 เมื่อ j=1,2,..,n
P=วิธีการเตรียม (แอททริบิวท์"วิธีการเตรียม" สำหรับงานวิจัยครั้งต่อไป)
ฉะนั้นสามารถแทนความรู้สรรพคุณทางยาของพืชสมุนไพรที่เก็บอยู่ในฐานข้อมูลดังกล่าวออกมาในรูปแบบของ Predicate Representation ดังนี้

$$\text{Herb\_ID(ID)} \wedge \text{Herb(X)} \wedge \text{Part\_of\_Herb(Y)} \wedge \text{Medcinal\_Property\_List}(mp_1, mp_2, .., mp_n)$$

รหัสสมุนไพร(ID) ^ ชื่อสมุนไพร(X) ^ ส่วนประกอบพืชสมุนไพร (Y) ^ รายชื่อสรรพคุณ($mp_1, mp_2, .., mp_n$)

ส่วนการตอบคำถามชนิด "ถามอะไร" ประเภทลิสต์ และประเภทเอนติตี้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพร

การตอบคำถามชนิด "ถามอะไร" ประเภทลิสต์ และประเภทเอนติตี้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพร ประกอบด้วยขั้นตอนต่าง ๆดังต่อไปนี้

1. การเรียนรู้แพทเทิร์นของคำถามอะไร (what-Question Pattern Learning)

ขั้นตอนนี้จะต้องมีการรวบรวมคำถามจากบลอค (Blog) ต่าง ๆบนเว็บไซต์ต่าง ๆที่ถามเกี่ยวกับสรรพคุณของพืชสมุนไพรจำนวน 150 คำถามเพื่อใช้เรียนรู้หาแพทเทิร์นของคำถามประเภท "อะไร" และ "อะไรบ้าง" และรวบรวมคำถามอีก 50 คำถามเพื่อใช้ในการทดสอบระบบการตอบคำถาม โดยคำถามเหล่านี้ทั้งหมดจะต้องมีขนาดความยาวไม่เกิน 10 คำ และผ่านกระบวนการตัดคำโดยการใช้ซอฟท์แวร์ตัดคำภาษาไทยที่สามารถแก้ปัญหาขอบเขตคำและขณะเดียวกันสามารถกำกับหน้าที่ของคำ (Part of Speech) ได้ (Sudprasert and Kawtrakul, 2003) ซึ่งสามารถเข้าใช้ระบบได้ถูกต้องพอสมควรโดยผ่านระบบอินเตอร์เนต หลังจากนั้นนำมาศึกษาหาแพทเทิร์นของคำถามชนิด "ถามอะไร" ประเภทลิสต์ และประเภทเอนติตี้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรได้แพทเทิร์นทั้งหมดได้ แพทเทิร์นประเภทลิสต์ 6 แพทเทิร์น ประเภทเอนติตี้ 3 แพทเทิร์นดังนี้ (จากบทนำ กำหนดให้ $np1_i$= part + TopicName,

$$\text{TopicName} \in \text{Herb}$$
$$\text{part} \in \{\text{null}, \text{"ใบ"}, \text{"ดอก"}, \text{"ราก"}, \text{"ต้น"}, \text{"เมล็ด"}, \text{"ผล"}, ...\}$$
$$np1_i \in \text{NP1} \ (i=1,2,,,m), \ np2_l \in \text{NP2} \ (l=1,2,..,h), \text{ และ } v_{mp} \in V_{mp})$$

ประเภทลิสต์

- $np1_i$ + "มี" + Focus("สรรพคุณ" + "อะไรบ้าง")
- Focus("ลิสต์" + "สรรพคุณ") + ["ของ"] +$np1_i$
- Focus("ลิสต์" + TopicName) + ["มีสรรพคุณ"] + v_{mp} +$np2_l$
- Focus(["แสดง | บอก"] + "สรรพคุณ") + ["ของ"] + $np1_i$ + "มี" + Focus("อะไรบ้าง")
- Focus(["แสดง | บอกชื่อ"] + "สมุนไพร" + "อะไรบ้าง") + ["มีสรรพคุณ"] + v_{mp} +$np2_l$

ประเภทเอนติตี้ "X อะไร"

- $np1_i$ + "มี" + Focus("สรรพคุณ" + "อย่างไร")
- Focus(["แสดง | บอกชื่อ"] + "สมุนไพร" + "อะไร) + ["มีสรรพคุณ"] + v_{mp} + $np2_l$

2. การวิเคราะห์คำถาม(Question Analysis)

นำแพทเทิร์นคำถามชนิด “ถามอะไร” ประเภทลิสต์ และประเภทเอนติตี้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรพร้อม Focusของคำถาม จากขั้นตอนก่อนหน้ามาทำการวิเคราะห์คำถามที่ต้องการทดสอบว่าตรงกับแพทเทิร์นไหน

3. การหาคำตอบ(Answer Determination)

นำแพทเทิร์นที่ตรงกับคำถามที่ต้องการทดสอบจากขั้นตอนก่อนหน้ามาทำการหาคำตอบโดยแมชชิง (Matching) ระหว่าง TopicName จากแพทเทิร์น กับ X ใน Predicate Representation

$np1_i$จากแพทเทิร์น กับ Y ใน Predicate Representation

v_{mp} + $np2_l$ จากแพทเทิร์น กับ mp_j ใน Predicate Representation

และส่วนของ Predicate Representation (ที่ตรงกับส่วน Focus ของแพทเทิร์น) คือส่วนที่แสดงคำตอบ ตัวอย่างเช่น

คำถาม: ใบโหระพามีสรรพคุณอะไรบ้าง

ใบ/ncn โหระพา/ncp มี/vt สรรพคุณ/ncn อะไรบ้าง/qw

part +TopicName +“มี” + Focus(“สรรพคุณ” + “อะไรบ้าง”)

ตรงแพทเทิร์นประเภทลิสต์ที่3: $np1_i$ + “มี” + Focus(“สรรพคุณ” + “อะไรบ้าง”)

Predicate Representation:

Herb_ID(ID) ^Herb(X) ^ Part_of_Herb(Y) ^ Medcinal_Property_List(mp_1,mp_2,..,mp_n)

รหัสสมุนไพร(ID) ^ ชื่อสมุนไพร(X) ^ ส่วนประกอบพืชสมุนไพร (Y) ^ รายชื่อสรรพคุณ(mp_1,mp_2,..,mp_n)

Answer: รายชื่อสรรพคุณ(mp_1,mp_2,..,mp_n)

รายชื่อสรรพคุณ(ขับลม, แก้ปวดฟัน, แก้ไอ, หลอดลมอักเสบ**)**

ผลการทดลองและการประเมินผล

การสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพร

คลังข้อมูลที่ใช้ทดสอบแบบจำลองที่ได้เสนอเกี่ยวกับการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรด้วยการใช้เทคนิคที่แตกต่างกันสองวิธี คือวิธีการประยุกต์ใช้ความรู้ทางภาษาศาสตร์ร่วมกับการเรียนรู้ของเครื่องด้วยการใช้เทคนิค NB (Na?ve Bates Classfier) และวิธีการประยุกต์ใช้ความรู้ต่าง ๆทางภาษาศาสตร์จนถึงระดับบทความที่ใช้ทฤษฎีเซนเทอร์ริง (Centering Theory, CT) กับคลังข้อมูล(Corpus)เอกสารทางวิชาการเกี่ยวกับพืชสมุนไพร ที่ได้จากการดาวโหลดจากเว็บไซด์สองแหล่งที่แตกต่างกันคือ จากสำนักงานเกษตรอำเภอพระพุทธบาท (http://phraphutthabat.saraburi.doae.go.th/herbs.htm) และจากโครงการอนุลักษณ์พันธ์พืชอันเนื่องจากพระราชดำริสมเด็จพระเทพรัตนราชสุดาฯสยามบรมราชกุมารี (http://www.rspg.or.th/index.htm) โดยมีลักษณะของพฤติกรรมคลังข้อมูลของโครงการอนุลักษณ์พันธ์พืชฯเป็น semi-structure ที่มีความเป็นระเบียบมากกว่าของสำนักงานเกษตรอำเภอพระพุทธบาท ดังนั้นงานวิจัยนี้ที่ได้พัฒนาวิธีการสกัดความรู้พร้อมวิธีการหาขอบเขต EDU ที่เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรด้วยเทคนิคที่แตกต่างกันสองวิธีเพื่อศึกษาพฤติกรรมข้อมูลทางภาษามีผลต่อเทคนิคทั้งสองหรือไม่ ตารางที่ 4 แสดงค่า Precision (สมการที่6), Recall (สมการที่7), และ ค่าความถูกต้องของการหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรโดยเทคนิค NB และ CT โดยวัดความถูกต้องบนพื้นฐานของ Max Win Voting จากผู้เชี่ยวชาญ 3 ท่าน

$$Precision = \frac{\#\,of\ \ samples\,correctly\,extracted\,as\,R}{\#\,of\ \ all\,samples\,output\,as\,being\,R} \qquad (6)$$

$$Recall = \frac{\#\,of\ \ samples\,correctly\,extracted\,as\,R}{\#\,of\ \ all\,samples\,holding\,the\,target\,relation\,R} \qquad (7)$$

R คือ ความสัมพันธ์ที่เกี่ยวกับการระบุ EDU เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพร

ตารางที่4 แสดงค่าPrecision, Recall, และ ค่าความถูกต้องของการหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรโดยเทคนิค NB และ CT

คลังข้อมูล	Lexico Syntactic Pattern for Herbal Medicinal Property Identification		Correctness of Boundary Determination	
	precision	recall	NB	CT
สำนักงาเกษตรอำเภอพระพุทธบาท (500 EDUs)	82	65	90	79
โครงการอนุลักษณ์พันธ์พืชฯ (500 EDUs)	92	84	93	93

จากตารางที่ 5 แสดงให้เห็นว่า ณ ที่ 95%Confident การหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรโดยเทคนิค NB ให้ผลดีกว่าการใช้ CT สำหรับคลังข้อมูลที่มีความเป็น Semi-structure น้อย ส่วนคลังข้อมูลที่มีความเป็น Semi-structure มากจะไม่มีความแตกต่างอย่างเป็นนัยสำคัญในระหว่างเทคนิคทั้งสองนี้

ตารางที่5 แสดงค่า t-test ของค่าความถูกต้องของการหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพร ระหว่างเทคนิคที่แตกต่างกันคือ NB กับ CT

คลังข้อมูล	Correctness of Boundary Determination		t-test
	NB	CT	
สำนักงาเกษตรอำเภอพระพุทธบาท	90	79	**2.15**
โครงการอนุลักษณ์พันธ์พืชฯ	93	93	0.27

จากตารางที่6 แสดงให้เห็นว่า ณ ที่ 95%Confident การหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรสำหรับคลังข้อมูลที่มีความแตกต่างกันในด้านความเป็น Semi-Structure พบว่าการใช้เทคนิค NB ให้ผลไม่มีความแตกต่างกันอย่างมีนัยสำคัญ แต่จะให้ผลแตกต่างอย่างมีนัยสำคัญสำหรับการใช้เทคนิค CT นั่นหมายความว่าหากใช้เทคนิค CT จะต้องคำนึงถึงการเลือกใช้คลังข้อมูลที่เหมาะสม

ตารางที่6 แสดงค่า t-test ของค่าความถูกต้องของการหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรระหว่างคลังข้อมูลที่แตกต่างกัน

คลังข้อมูล	Correctness of Boundary Determination	
	NB	CT
สำนักงาเกษตรอำเภอพระพุทธบาท	90	79
โครงการอนุลักษณ์พันธ์พืชฯ	93	93
t-test	0.76	**2.85**

จากตารางที่7 แสดงให้เห็นว่า ณ ที่ 95%Confident การระบุ EDU ที่เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรด้วยวิธีการใช้แพทเทิร์นทางภาษาศาสตร์คือ Lexico Syntactic Pattern สำหรับคลังข้อมูลที่มีความแตกต่างกันในด้านความเป็น Semi-Structure พบว่าค่า Precision และ ค่าRecall ที่ได้มีความแตกต่างกันอย่างมีนัยสำคัญในระหว่างคลังข้อมูลทั้งสองที่แตกต่างกันนั้น นั่นหมายความว่าพฤติกรรมของคลังข้อมูลมีผลต่อการใช้ Lexico Syntactic Pattern สำหรับการระบุ EDU ที่ต้องการ

ตารางที่7 แสดงค่า t-test ของค่า Precision และ Recall ระหว่างคลังข้อมูล(Corpus)ที่แตกต่างกัน

คลังข้อมูล	Lexico Syntactic Pattern for Herbal Medicinal Property Identification	
	precision	precision
สำนักงาเกษตรอำเภอพระพุทธบาท	82	82
โครงการอนุลักษณ์พันธ์พืชฯ	92	92
t-test	**2.1**	**3.08**

การตอบคำถามชนิด “ถามอะไร” ประเภทลิสต์ และประเภทเอนติตี้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพร

พืชสมุนไพรที่สกัดได้มีประมาณ 22 ชนิด คือกระชาย กระเทียม กระเพรา ขี้เหล็ก ชุมเห็ดเทศ โหระพา ขิง ข่า ตะไคร้ มะขาม พริก มะกูด มะนาว ทับทิม มังคุด คำฝอย ว่านหางจระเข้ ฟ้าทะลายโจร พลู กระวาน กระดังงา ขมิ้น เป็นต้น คำถามที่ใช้ทดสอบมีทั้งหมด 50 คำถาม ระบบสามารถตอบได้ถูกต้อง36 คำถาม (72%) ที่ไม่ถูกต้องส่วนใหญ่เนื่องมาจากคำถามที่ซับซ้อน

สรุป

งานวิจัยนี้มุ่งเน้นการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรโดยเฉพาะอย่างยิ่งสมุนไพรไทยจากเอกสารภาษาไทย โดยมีการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรนี้ออกมาในรูปแบบของEDU ซึ่งเป็นประโยคแบบง่ายๆธรรมดา แต่สกัดออกมาได้ครั้งละหลายๆ EDUs ที่ต่อเนื่องกัน คือได้หลายๆสรรพคุณติดต่อกันต่อพืชสมุนไพรหนึ่งชนิด ในขณะที่งานวิจัยก่อนหน้านี้ จะเป็นการสกัดออกมาได้ครั้งละหนึ่งหรือ หนึ่งถึงสามคุณสมบัติ/สรรพคุณ เท่านั้น ทั้งนี้เพราะงานวิจัยก่อนหน้าจะเป็นการสกัดที่อยู่บนพื้นฐานของนามวลีเท่านั้น ในขณะที่งานวิจัยนี้มุ่งเน้นการสัดความรู้บนพื้นฐานของเหตุการณ์ที่สามารถแทนได้ดีด้วยกริยา คือ ทำให้สามารถเข้าใจถึงเหตุการณ์ได้อย่างต่อเนื่อง เช่น “กระเพราใช้เป็นยาขับลม แก้คลื่นไส้ [แก้]อาเจียน บรรเทาอาการปวดท้อง แก้ท้องเสีย” ฉะนั้นความรู้เกี่ยวกับสรรพคุณทายาของพืชสมุนไพรที่สกัดได้นี้ช่วยสนับสนุนการตอบคำถามชนิด “คำถามอะไร/what-Question” ประเภทลิสต์ ได้อย่างชัดเจน

งานวิจัยนี้สามารถสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรในรูปแบบหลายๆ EDUs ได้อย่างมีประสิทธิภาพโดยเฉพาะอย่างยิ่งการสกัดขอบเขตของกลุ่มสรรพคุณทางยาของพืชสมุนไพรด้วยการใช้เทคนิคที่แตกต่างกันสองวิธี คือ วิธีการประมวลผลภาษาธรรมชาติด้วยการประยุกต์ใช้ความรู้ทางภาษาศาสตร์ระดับวลีของประโยคร่วมกับเทคนิคการเรียนรู้ของเครื่องด้วยเทคนิคของ NB และวิธีการประมวลผลภาษาธรรมชาติที่ประยุกต์ใช้ความรู้ทางภาษาศาสตร์ในระดับบทความโดยใช้ CT ซึ่งก็คือศูนย์กลางความสนใจ (Center of Attention) ของแต่ละ EDUภายใต้ขอบเขตของสรรพคุณทางยาของพืชสมุนไพร ที่ส่วนใหญ่จะมีผู้กระทำ (Agent) ที่แตกต่างกันน้อย ดังนั้นค่าเฉลี่ยความถูกต้องของการหาขอบเขตของสรรพคุณทางยาของพืชสมุนไพรสำหรับ NB เป็น 91.5% และ CT เป็น 86% กล่าวคือ NB ให้ผลลัพธ์ที่มีประสิทธิภาพกว่า CT ถึงแม้ว่าพฤติกรรมของข้อมูลจะแตกต่างกันก็ตาม ในขณะที่การระบุ EDU ที่เป็นสรรพคุณทางยาของพืชสมุนไพรด้วยการใช้ Lexico Syntactic Pattern (NP1 V_{mp} NP2) ไปพร้อมกับความหมายเชิงความคิด (Concept) ของกริยาและนามวลี ทำให้ได้ค่า Precision และ Recall เฉลี่ยคือ 89% และ 74% ตามลำดับ ทั้งนี้ขึ้นอยู่กับพฤติกรรมของคลังข้อมูล

ระบบการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรจากเอกสารทางวิชาการภาษาไทยนี้สามารถนำไปใช้สนับสนุน ระบบการตอบคำถามชนิด “ถามอะไร/what-Question” ประเภทลิสต์ และประเภทเอนติตี้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพร ด้วยวิธีการใช้แพทเทิร์นที่ได้จากการสังเกตจากการเรียนรู้ ซึ่งแพทเทิร์นเหล่านี้สามารถใช้สำหรับวิเคราะห์คำถามที่ป้อนเข้ามาโดยไม่มีการใช้ซอฟท์แวร์ Parser ฉะนั้นคำถามที่เข้ามาในระบบจะต้องเป็นคำถามที่ไม่ยาว คือ ไม่ควรยาวกว่า 10 คำ ผลลัพธ์ของระบบการตอบคำถามของงานวิจัยนี้สามารถตอบคำถามได้ถูกต้อง 72% ทั้งนี้เนื่องมาจากบางคำถามมีความซับซ้อน อย่างไรก็ตามหากระบบมีการพัฒนาต่อไปให้สามารถตอบคำถามได้ถูกต้องเพิ่มขึ้น ก็จะทำให้ระบบการตอบคำถามนี้มีประโยชน์ต่อสุขภาพของสังคมอย่างมาก โดยเฉพาะชาวบ้านที่อยู่ห่างไกลจากตัวเมือง เมื่อมีความจำเป็นต้องใช้ยาและไม่สามารถเดินทางไปหาซื้อได้ ฉะนั้นยาสมุนไพรก็เป็นอีกช่องทางที่สามารถหาใช้ได้ทันที่หากมีการปลูกพืชสมุนไพรที่ได้กล่าวมาข้างต้นไว้ในบ้านบ้าง

เอกสารอ้างอิง

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

Chang D S, Choi K S., 2004. Causal Relation Extraction Using Cue Phrase and Lexical Pair Probabilities, IJCNLP, pp61-70, Hainan Island, China.

Chanlekha, H. and A. Kawtrakul. 2004. Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. IJCNLP' 2004, HAINAN Island , China.

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

Fan S., Zhang Y., Ng W.W.Y., Wang X., and Wang X. 2008. Semantic Chunk Annotation for complex questions using Conditional Random Field Coling 2008: Proceedings of the workshop on Knowledge and Reasoning for Answering Questions, Manchester, pages 1–8.

Fang Y-C, Huang H-C, Chen H-H, and Juan H-F. 2008. TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. BMC Complementary and Alternative Medicine 2008, 8:58, Biomed Central.

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

Grosz, B. J. 1977. The Representation and Use of Focus in Dialogue Understanding. Ph.D. thesis, Stanford University.

Grosz, B. J. and C. L. Sidner. 1986. Attention, intention, and the structure of discourse. Computational Linguistics, 12(3):175-204.

Metzler D. and Croft W. B. 2005. Analysis of Statistical Question Classification for Fact-based Questions. in Information Retrieval, 8(3), pp. 481-504, Netherlands.

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

Paşca M. 2008. Turning Web Text and Search Queries into Factual Knowledge: Hierarchical Class Attribute Extraction. Proceedings of Association for the Advancement of Artificial Intelligence, AAAI 08.

Pengphon, N., A. Kawtrakul, and M. Suktarachan. 2002. Word Formation Approach to Noun Phrase Analysis for Thai. SNLP2002, Thailand.

Quaresma P. and Rodrigues I. 2005. A question answering system for Portuguese juridical documents. Proceedings of the 10th international conference on Artificial intelligence and law International, Conference on Artificial Intelligence and Law, pp. 256 - 257.

Riloff E. and Thelen M. 2000. A Rule-based Question Answering System for Reading Comprehension Tests. NAACL-ANLP 2000 Workshop: Syntactic and Semantic Complexity in Natural Language Processing Systems.

Takahashi K., Koike A., and Takagi T. 2004. Question Answering System in Biomedical Domain. Proceedings of the Genome Informatics 2004 (GIW 2004), 161-162.

Smith, J.G., and Duncan, A.J. "Elementary Statistics and Applications: Fundamentals of the Theory of Statistics", Mc Graw-Hill Book Company Inc., New York, London, pp. 323, 1944.

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

Walker M., Joshi A., and Prince E. 1998. Centering in Naturally Occuring Discource: An Overview in Centering Theory of Discourse. Calendron Press, Oxford.

Weeber M. and Vos. R. 1998. Extracting Expert Knowledge from Medical Texts. Intelligent Data Analysis in Medicine and Pharmacology, IDAMAP 98, A Workshop at the 13th European Conference on Artificial Intelligence ECAI-98, Brighton, UK.