การสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทยเพื่อสนับสนุนระบบการตอบคำถามอัตโนมัติ

บทนำ

1. ความเป็นมาของปัญหา

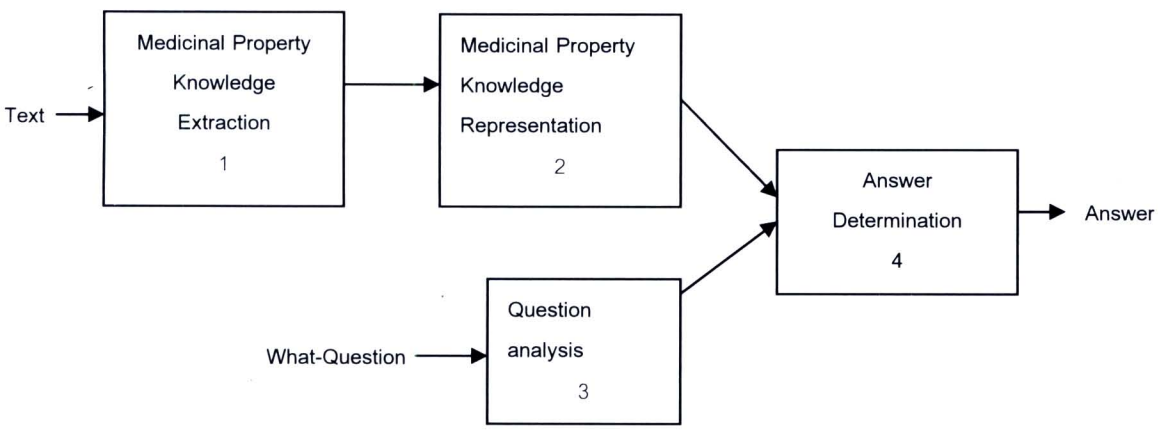
ระบบศูนย์บริการความรู้อัตโนมัติ (Automatic Knowledge Service Center System) ในปัจจุบันมีปัญหาในด้านการให้บริการความรู้เชิงบรรยายแบบอัตโนมัติได้อย่างมีประสิทธิภาพ โดยเฉพาะอย่างยิ่งความรู้เกี่ยวกับคุณสมบัติหรือสรรพคุณต่างๆของเอนติตี้ (Entity)ที่เป็นวัตถุ (Object) มีอยู่จริง (เช่น โหระพา กระเพรา ขมิ้น ขิง 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 ใช้เป็นยาล้างแผลที่อัณฑะ, สารพิเพอรีนในพริกไทยสามารถใช้เป็นยาฆ่าแมลง ซึ่งไม่เป็นอันตรายต่อมนุษย์โดยนำผลพริกไทยมาทุบให้แตกแล้วใช้โรยบริเวณตู้เสื้อผ้าหรือบริเวณที่ต้องการ

จากตัวอย่างเอกสารสมุนไพรไทยข้างต้นส่วนที่ขีดเส้นใต้คือสรรพคุณหรือคุณสมบัติของพืชสมุนไพร "พริกไทย" ดังนั้นปัญหาในส่วนของการสกัดความรู้เกี่ยวกับคุณสมบัติของเอนติตี้โดยเฉพาะเรื่องสรรพคุณทางยาของสมุนไพรไทยจากเอกสารภาษาไทย ประกอบด้วยสามปัญหาคือ ปัญหาการระบุเอนติตี้สมุนไพรไทย ปัญหาการระบุสรรพคุณทางยาของสมุนไพรไทย ปัญหาการหาขอบเขตของ EDUs สรรพคุณทางยาของสมุนไพรไทย

ส่วนปัญหาจากระบบสอบถามเกี่ยวกับคุณสมบัติของเอนติตี้สมุนไพรไทย ประกอบด้วยปัญหาการวิเคราะห์ "คำถามอะไร" ปัญหาการระบุโฟกัส (Focus)ของคำถาม และปัญหาการหาคำตอบจากความรู้เกี่ยวกับคุณสมบัติของเอนติตี้ที่ได้สกัดมา

งานวิจัยที่เกี่ยวข้องสามารถแบ่งออกเป็นสองส่วนหลักคือ ส่วนการสกัดและแทนความรู้เกี่ยวกับคุณสมบัติของเอนติตี้และส่วนการตอบคำถามของคำถามประเภท"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 คลาสที่สามารถระบุได้ 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) จากเว็บไซต์ภาษาจีน 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) ไม่สามรถนำมาประยุกต์ใช้กับงานวิจัยนี้

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} คือเซทความคิดกริยาทีแสดงให้เห็นสรรพคุณทางยา(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 $v_{ik} \in V_{mp}$ (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 คือประโยคง่ายๆ 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