



รายงานผลการวิจัย

เรื่อง

การสกัดและเก็บรวบรวมความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’
จากเอกสารบนเว็บ

Extraction and Collection of “HerbPlantPart-MedicinalProperty Relation”
from Web-Documents

โดย

รองศาสตราจารย์ ดร. จวีวรรณ เพ็ชรศิริ

สำราญ ไข่มวล

รายงานการวิจัยนี้ได้รับทุนอุดหนุนจากมหาวิทยาลัยธุรกิจบัณฑิตย์

พ.ศ. 2560

ชื่อเรื่อง: การสกัดและเก็บรวบรวมความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’

จากเอกสารบนเว็บ

ผู้วิจัย: รศ.ดร. ณวิวรรณ เพ็ชรศิริ

สถาบัน: มหาวิทยาลัยธุรกิจบัณฑิต

ปีที่พิมพ์: 2560

สถานที่พิมพ์ : มหาวิทยาลัยธุรกิจบัณฑิต

แหล่งที่เก็บรายงานวิจัยฉบับสมบูรณ์: มหาวิทยาลัยธุรกิจบัณฑิต

จำนวนหน้างานวิจัย: 27 หน้า

บทคัดย่อ

จุดประสงค์ของงานวิจัยนี้คือการสกัดและเก็บรวบรวมการสกัดและเก็บรวบรวมความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารบนเว็บสมุนไพรของหน่วยงานหรือองค์กรที่เกี่ยวข้องกับดูแลรักษาพันธ์พืชสมุนไพรรวมทั้งการพัฒนาผลิตภัณฑ์ ผลของงานวิจัยนี้คือองค์ความรู้คุณสมบัติหรือสรรพคุณทางยาของพืชสมุนไพรบนพื้นฐานประโยคความเดียวหรืออิตียู (EDU, Elementary Discourse Unit) จำนวนหลายๆประโยค ซึ่งเป็นเป็นประโยชน์ต่อชาวบ้านหรือบุคคลทั่วไปในการใช้รักษาสุขภาพผ่านระบบถามตอบอัตโนมัติ อย่างไรก็ตามงานวิจัยนี้มีสี่ปัญหาหลัก: ปัญหาแรกคือ ซึ่งมักจะมีลักษณะปัญหาเป็น Zero Anaphora (การละคำสรรพนามที่ใช้อ้างอิงเอนทิตี เช่น พืชสมุนไพร) และ Textual Ellipsis (การละคำนาม เช่น เอนทิตีพืชสมุนไพร) ปัญหาที่สองคือ การระบุ EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพรของเอนทิตีพืชสมุนไพรและปัญหาที่สามคือ การหาขอบเขตของ ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร ปัญหาที่สี่คือ การจัดเก็บความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่สกัดได้อย่างไรเพื่อให้สามารถเข้าถึงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ได้อย่างไม่ซับซ้อน ดังนั้นสำหรับปัญหาที่หนึ่งงานวิจัยนี้ได้ใช้กฎทางฮิวริสติก คือใช้คำนามที่เป็น Agent ของ EDU ก่อนหน้า มาเป็น Agent ของ EDU ที่ปรากฏ Zero Anaphora และการใช้ Topic name สำหรับปัญหาTextual Ellipsis ส่วนปัญหาที่สองงานวิจัยนี้เสนอการใช้ N-Word-Co ที่สกัดได้จากกริยาวลีของ EDU ต่างๆที่มีแนวความคิดสรรพคุณทางยา ของพืชสมุนไพร และเก็บรวบรวมไว้มาทำการระบุ EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร นอกจากนี้งานวิจัยนี้ใช้ N-Word-Co ที่เก็บรวบรวมนั้น กับ Similarity Score แก้ไขปัญหาที่สามและงานวิจัยนี้ได้ใช้เมตริกซ์ที่ประกอบด้วย รหัสพืชสมุนไพร , ชื่อพืชสมุนไพร, ชื่อส่วนของพืชสมุนไพร, แนวความคิดสรรพคุณทางยา, และรหัสความสัมพันธ์ มาทำการเก็บรวบรวมข้อมูลเป็นตารางเพื่อแสดงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ในรูปแบบเครือข่ายเหมือน CODASYL โดยการเรียงลำดับตาม ชื่อพืชสมุนไพรและ ชื่อส่วนของพืชสมุนไพร ฉะนั้นผลลัพธ์จากงานวิจัยที่ได้เสนอนี้ สามารถสกัดความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารภาษาไทยได้ค่าความแม่นยำ และค่าระลึก 90% และ 77.5% ตามลำดับ

Research Topic: Extraction and Collection of “HerbPlantPart-MedicinalProperty Relation” from Web-Documents

Researcher: Assoc. Prof. Chaveevan Pechsiri **Institution:** Dhurakij Pundit University

Year of Publication: 2017 **Place of Publication:** Dhurakij Pundit University

Place of Collection: Dhurakij Pundit University

Number of Pages: 27

Abstract

This research aims to collect extracted Herb-MedicinalProperty relations from downloaded herbal-plant documents from several Thai herabal medicinal product associations for creating the herbal-medicinal-property-network based representation. An Herb-MedicinalProperty relation is a semantic relation between one herbal-plant-component concept and several herbal-medicinal-property-concept expressions on texts and vice versa. An herbal-plant-component occurrence is a noun-phrase expression and each herbal-medicinal-property concept occurrence is an event expression by a verb-phrase of an EDU (where ‘EDU’ is an elementary discourse unit which is a simple sentence or a clause). The extracted Herb-MedicinalProperty relations as the property knowledge represented by a diagram of an herbal-medicinal-property network are the essential information for the social health-care problems. In particularly, the herbal-medicinal-property network based representation benefits a recommendation system of solving health-problems on web-boards. The research has four problems: 1) the problems of the zero anaphora and the textual ellipsis for the herbal named-entity expression 2)how to identify an EDU with the herbal medicinal property concept 3)how to determine the boundary of several EDUs with the herbal medicinal property concepts 4)how to collect the extracted Herb-MedicinalProperty relations for facilely accessing through the relations for the herbal-medicinal-property-network representation. Therefore, we apply the heuristic rules for solving the first problem; using the previous noun as an agent to be the agent of the EDU containing the zero-anaphora agent and using the topic name to solve the textual ellipsis. We propose applying a co-occurrence of N-Words (or N-Word-Co) including N-Word-Co size learning on the verb phrase to identify EDU with the herbal medicinal property concept from the documents. We also apply the similarity-score determination for solving the boundary of several EDUs with the herbal medicinal property concepts. We then apply the matrix consisting 5 features of Herb ID, HerbName, HerbPlantPart Concept, Herbal-Medicinal-Property Concept, and Relation ID to create the network representation of the herbal-medicinal-property network liked CODASYL. The research results provide the 90% precision with 77.5% recall of the Herb-MedicinalProperty relation extraction from the documents.

กิตติกรรมประกาศ

ขอขอบพระคุณ ผู้อำนวยการศูนย์บริการวิจัย มหาวิทยาลัยธุรกิจบัณฑิตที่ให้โอกาสข้าพเจ้าและอาจารย์ สรรพ ใฝ่เนวล ทำการศึกษาค้นคว้าวิจัยเรื่อง “การสกัดและเก็บรวบรวมความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารบนเว็บ” จนสำเร็จ

ขอขอบพระคุณ คณบดีวิทยาลัยนวัตกรรมการเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิต ที่สนับสนุนในเรื่องของเวลาที่ใช้ไปสำหรับงานวิจัยนี้

ขอขอบพระคุณ มหาวิทยาลัยธุรกิจบัณฑิต ที่ให้เงินทุนสำหรับสนับสนุนโครงการวิจัยนี้

ท้ายที่สุด ขอขอบพระคุณครอบครัว ญาติพี่น้องและเพื่อนๆ ที่ให้กำลังใจในการทำโครงการวิจัยที่มีค่านี้

(รองศาสตราจารย์ ดร. จวีวรรณ เพ็ชรศิริ)

หัวหน้าโครงการ

18 ตุลาคม 2560

สารบัญ

	หน้า
สารบัญ	i
สารบัญตาราง	iv
สารบัญรูป	v
บทนำ	1
1. ความเป็นมาของปัญหา	1
2. วัตถุประสงค์	6
3. นิยามคำศัพท์	6
4. ขอบเขตงานวิจัย	7
งานวิจัยที่เกี่ยวข้อง	8
ความรู้พื้นฐาน	8
1. Maximun Entropy	8
2. Similarity_Score	8
งานวิจัยก่อนหน้า	9
1. แนวทางแพตเทิร์นหรือกฎ (Pattern/Rule Based Approach)	9
2. แนวทางสถิติรวมถึงการเรียนรู้ของเครื่อง (Statistical Based Approach	
Including Machine Learning)	10

สารบัญ (ต่อ)

หน้า

ปัญหาการสกัดความสัมพันธ์ “ส่วนของพืชสมุนไพร-สรรพคุณทางยา” จาก เอกสารบนเว็บ	12
ปัญหาการสกัดความสัมพันธ์	12
1. ปัญหาการระบุเอนทิตีพืชสมุนไพร (Herb-Plant Entity) และส่วนของพืช สมุนไพร (Herb-Plant Part)	12
2. ปัญหาการระบุ EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร	12
3. ปัญหาการหาขอบเขตของ EDUs (EDU Boundary) ที่มีแนวความคิด สรรพคุณทางยาสมุนไพร	13
ปัญหาการจัดเก็บความสัมพันธ์	14
กรรมวิธีดำเนินงาน	15
1. การเตรียมคลังข้อมูล (Corpus Preparation)	15
2. การเรียนรู้ขอบเขต N-Word-Co (N-Word-Co Boundary Learning)	16
3. การสกัด N- Word-Co (N-Word-Co Extraction)	17
4. การสกัดความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทาง ยา’(HerbPlantPart-MedicinalProperty Relation Extraction)	20

สารบัญ (ต่อ)

5. การรวบรวมเก็บสะสมและแทนความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ (Collection and Representation of HerbPlantPart-MedicinalProperty Relation)	22
ผลการทดลองและการประเมินผล	24
1. การสกัด N-Word-Co ที่มีแนวความคิดสรรพคุณทางยา	24
2. การสกัด ความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากคลังข้อมูล	24
สรุป	26
เอกสารอ้างอิง	27

สารบัญตาราง

ตาราง		หน้า
1	ตัวอย่าง ลิสต์ N-Word-Co	19
2	แสดงเมตริกซ์ของ ความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่เรียงโดย HerbName and HerbPlantPart Concept	22
3	แสดงเมตริกซ์ของ ความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่เรียงโดย Herbal-Medical-Property	23

สารบัญรูป

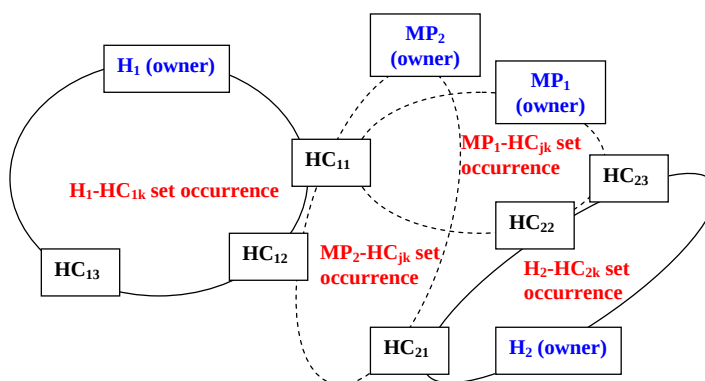
รูป		หน้า
1	แสดงความสัมพันธ์เชิงเน็ตเวิร์ค ระหว่างส่วนของพืชสมุนไพร กับรายการ สรรพคุณทางยา	2
2	กรอบงานวิจัย	15
3	การกำกับคำของ N-Word-Co ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร และNP1 ที่มีแนวความคิดส่วนของพืชสมุนไพร	16
4	อัลกอริทึมหาขอบเขต N-Word-Co	18
5	อัลกอริทึมหาขอบเขต การหาขอบเขต EDU ที่มีแนวความคิดสรรพคุณทางยา สมุนไพร	21

การสกัดและเก็บรวบรวมความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารบนเว็บ

บทนำ

1. ความเป็นมาของปัญหา

สมุนไพร ตามพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2525 หมายถึง พืชที่ใช้ ทำเป็นเครื่องยาสมุนไพรกำเนิดมาจากธรรมชาติและมีความหมายต่อชีวิตมนุษย์โดยเฉพาะ ในทางสุขภาพ อันหมายถึง ทั้งการส่งเสริมสุขภาพและการรักษาโรค สำหรับสมุนไพรไทย โดยเฉพาะที่เป็นพืชมีมากกว่า 1500 ชนิด ซึ่งมีคุณสมบัติและสรรพคุณแตกต่างกันไปและถูกบรรยายอยู่บนเว็บไซต์มากมาย ทำให้เสียเวลาไปกับการค้นหาพืชสมุนไพรต่างๆและส่วนของพืชสมุนไพร (เช่นราก ลำต้น ใบ ดอก เมล็ด เป็นต้น) มีความสัมพันธ์หรือเกี่ยวข้องกับสรรพคุณต่างๆทางยาอย่างไรบ้าง หรือค้นหาสรรพคุณทางยาที่ต้องการมีอยู่ในส่วนไหนของพืชสมุนไพร ตัวอย่างสรรพคุณทางยา เช่น ลดไข้ บรรเทาอาการปวด แก้ปวดท้อง เป็นต้น ดังนั้นงานวิจัยนี้จึงมีจุดประสงค์ในการสกัดและเก็บรวบรวม ความสัมพันธ์ ‘พืช/ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ (Herb-MedicinalProperty Relation) จากเอกสารบนเว็บ เช่น เว็บไซต์ของโครงการอนุรักษ์พันธุกรรมพืชอันเนื่องมาจากพระราชดำริฯ (http://www.rspg.or.th/plants_data/herbs/herbs_200.htm)และสารานุกรมไทย(<http://kanchanapisek.or.th/kp6/sub/book/book.php?book=14&chap=10&page=chap10.htm>) ฉะนั้นความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่สกัดได้และเก็บรวบรวมจากเว็บไซต์สู่ศูนย์ให้บริการความรู้ (Knowledge Service-Center Portal) โดยความสัมพันธ์ที่สกัดได้สามารถแทนด้วยเครือข่ายสรรพคุณทางยาของพืชสมุนไพร (Herbal-Medicinal-Property Network) ซึ่งแสดงให้เห็นความสัมพันธ์เชิงเน็ตเวิร์ค (Network) หรือ CODASYL (Rayward and Bowden, 2004) ระหว่างส่วนของพืชสมุนไพรกับสรรพคุณทางยา ดังแสดงในรูปที่ 1 โดยมีการปรากฏของเซต (Set Occurrence) แต่ละวงในเน็ตเวิร์คประกอบด้วย เจ้าของ (Owner) เซตที่เป็นวงบนเน็ตเวิร์ค (เช่น H_1 เป็นเจ้าของการปรากฏวงเน็ตเวิร์ค H_1-HC_{1k} (H_1-HC_{1k} set occurrence)) และสมาชิก(Member) ของเซตบนวงเน็ตเวิร์ค (เช่น HC_{11} HC_{12} HC_{13} เป็นสมาชิกของเซตบนวงเน็ตเวิร์ค H_1-HC_{1k})



รูปที่ 1. แสดงความสัมพันธ์เชิงเน็ตเวิร์ค ระหว่างส่วนของพืชสมุนไพร กับรายการสรรพคุณทางยา (where H_1 = ‘มะนาว/Lemon’, H_2 = ‘สะเดา/Neem’, HC_{11} = ‘เมล็ดมะนาว/Lemon.seed’, HC_{12} = ‘รากมะนาว/Lemon.root’, HC_{13} = ‘ผลมะนาว/Lemon.fruit’, HC_{21} = ‘เปลือกไม้สะเดา/Neem.bark’, HC_{22} = ‘รากสะเดา/Neem.root’, HC_{23} = ‘แก่นไม้สะเดา/Neem.heart-of-wood’, MP_1 = ‘ละลายเสมหะ/clear-phlegm’, MP_2 = ‘ลดไข้/Reduce-fever’)

ดังนั้นความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่แสดงด้วยเครือข่ายสรรพคุณทางยาของพืชสมุนไพรบนเว็บไซต์ศูนย์ให้บริการความรู้ที่ประชาชนทั่วไปสามารถเข้าถึงและสามารถสอบถามความสัมพันธ์ ‘พืช/ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่สนใจเพื่อนำไปใช้แก้ไขปัญหาสุขภาพ นอกจากนี้ที่เว็บไซต์ศูนย์ให้บริการความรู้ควรที่จะเป็นแหล่งให้บริการความรู้ด้านอื่นๆอีก เช่น ความรู้เรื่องการเตรียมยาสมุนไพรและความรู้วิธีการใช้ยาสมุนไพร เป็นต้น อย่างไรก็ตามความรู้เรื่องสรรพคุณของสมุนไพรปรากฏในเอกสารพืชสมุนไพรไทยในรูปแบบของหลายๆ EDU (Elementary Discourse Unit, คือประโยคง่ายๆ ธรรมดาไม่ซับซ้อน, Carlson and et. al, 2003) ดังตัวอย่างต่อไปนี้

EDU1: “กระเทียมเป็นยาขับลมในลำไส้”

EDU2: “[กระเทียม] แก้อาเจียน”

EDU3: “[กระเทียม] ขับเสมหะ”

EDU4: “[กระเทียม] ช่วยย่อยอาหาร”

EDU5: “[กระเทียม] รักษากลากเกลื้อน”

หมายเหตุ: สัญลักษณ์ ‘[...]’ หมายถึงการละคำ

ดังนั้นจากตัวอย่างเอกสารพืชสมุนไพรไทยต่อไปนี้

พริกไทย (Piper nigrum Linn.)

สรรพคุณ เปลือกของพริกไทยมีน้ำย่อยสำหรับย่อยไขมัน ด้วยเหตุนี้ตำราโบราณจึงเชื่อกันว่า พริกไทยสามารถลดความอ้วนได้, พริกไทยช่วยกระตุ้นปฏิกิริยาที่ลิ้น เพื่อให้กระเพาะอาหารหลั่งน้ำย่อยได้มากขึ้น, พริกไทยดำมีรสเผ็ดอุ่น เมื่อรับประทานเข้าไปจะรู้สึกอุ่นวาบที่ท้อง ช่วยขับลม ขับเหงื่อ ขับปัสสาวะ แก้อืดท้องเฟ้อ แก้ไข้มาลาเรีย แก้หิวาตกโรค, ใช้ก้านพริกไทย 10 ก้าน บดให้ละเอียดแล้วต้มกับน้ำ 8 แก้ว ใช้เป็นยาล้างแผลที่อันตราย, สารพิเพอรินในพริกไทยสามารถใช้เป็นยาฆ่าแมลง ซึ่งไม่เป็นอันตรายต่อมนุษย์โดยนำผลพริกไทยมาทบให้แตกแล้วใช้

จากตัวอย่างเอกสารสมุนไพรไทยข้างต้นพบว่า การปรากฏของประโยคต่างๆ (ที่ขีดเส้นใต้) แสดงสรรพคุณหรือคุณสมบัติของพืชสมุนไพร ‘พริกไทย’ นั้นมีลักษณะเป็นกลุ่มกระจายอยู่บนเอกสาร กล่าวคือ จะทำอย่างไรให้คอมพิวเตอร์สามารถระบุความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารดังกล่าว ดังนั้นปัญหาสำหรับงานวิจัยนี้ (ดูบทปัญหาการสกัด ‘ความสัมพันธ์ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารบนเว็บ) ประกอบด้วยสองปัญหาหลักคือ ปัญหาการสกัดความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารสมุนไพร และปัญหาการรวบรวมจัดเก็บความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่สกัดได้เพื่อทำให้สามารถแสดงหรือแทนความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ในรูปแบบเครือข่าย (รูปที่ 1) ได้อย่างมีประสิทธิภาพ นอกจากนี้ปัญหาการสกัด ‘ความสัมพันธ์ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารสมุนไพร ประกอบด้วยปัญหาย่อย 3 ปัญหาคือ 1) ปัญหาการระบุเอนทิตีพืชสมุนไพร (Herb-Plant Entity) และส่วนของพืชสมุนไพร (Herb-Plant Part) ซึ่งมักจะมีลักษณะปัญหาเป็น Zero Anaphora (การละคำสรรพนามที่ใช้อ้างอิงเอนทิตีพืชสมุนไพร) และ Textual Ellipsis (การละคำนามที่เป็นเอนทิตีพืชสมุนไพร) 2) ปัญหาการระบุ EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพรของเอนทิตีพืชสมุนไพร 3) ปัญหาการหาขอบเขตของ EDUs (EDU Boundary) ที่มีแนวความคิดสรรพคุณทางยาสมุนไพรของเอนทิตีพืชสมุนไพร

งานวิจัยก่อนหน้านี้เกี่ยวกับการสกัดความสัมพันธ์เชิงความหมาย (Semantic Relation) ที่เกี่ยวข้องกับการสกัดคุณสมบัติ/สรรพคุณของเอนทิตีที่มีดังต่อไปนี้ Weeber M. and Vos. R., 1998; Fang et al., 2008; และ Paşca M., 2008 ; Haruechaiyasak et al., 2011; Riaz M. and Girju R., 2014; Recupero, D.R et al., 2015 (ดูในบทงานวิจัยที่เกี่ยวข้อง) โดยวิธีดังกล่าวไม่เหมาะสำหรับงานวิจัยนี้ที่มุ่งสกัดความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ซึ่งอยู่บนพื้นฐานการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของเอนทิตีพืชสมุนไพรไทยจากเอกสารภาษาไทย ซึ่งบรรยายสรรพคุณ

เหล่านี้ในรูปของกริยาวลี ที่มีหลายๆกริยาวลีอยู่ต่อเนื่องกันภายใต้หัวข้อสมุนไพรหนึ่งชนิด (คือหนึ่งเอ็นทีดีสมุนไพร) และการบรรยายสรรพคุณทางยาของสมุนไพรส่วนใหญ่จะกล่าวถึงส่วนของสมุนไพรซึ่งจะอยู่ในรูปของนามวลีที่ส่วนใหญ่มักจะละ (Zero Anaphora) เช่น “เปลือกของพริกไทย.....[เปลือกของพริกไทย]ช่วยขับลม [เปลือกของพริกไทย]ขับเหงื่อ [เปลือกของพริกไทย]ขับปัสสาวะ [เปลือกของพริกไทย]แก้ท้องอืดท้องเฟ้อ” ([..] หมายถึงละคำที่อยู่ในวงเล็บก้ามปู)

ดังนั้นจากงานวิจัยก่อนหน้าและปัญหามานี้ที่ได้กล่าวข้างต้น กรณีปัญหาการสกัดความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารสมุนไพร ซึ่งประกอบด้วย 3 ปัญหาย่อย โดยมีปัญหาย่อยแรกคือ ปัญหา Zero Anaphora และ Textual Ellipsis ของการระบุเอ็นทีดีพืชสมุนไพรและส่วนของพืชสมุนไพร ปัญหานี้สามารถแก้ไขโดยใช้กฎทางอิวิเดนต์ คือใช้คำนามที่เป็นเอเจนต์ (Agent) ของ EDU ก่อนหน้ามาเป็นเอเจนต์ของ EDU ที่ปรากฏ Zero Anaphora สำหรับ Textual Ellipsis สามารถแก้ปัญหาโดยใช้คำนามชื่อสมุนไพรซึ่งเป็น Topic name สำหรับปัญหาย่อยที่สองคือปัญหาการระบุ EDU ที่แสดงแนวความคิดสรรพคุณทางยาของพืชสมุนไพร โดย EDU ดังกล่าวมีลักษณะรูปแบบทางภาษาศาสตร์ดังนี้

EDU → NP1 VP

VP → V_{prop} VP | V_{prop} NP2 | V_{prop} NP2 VP | V_{prop} Adv | V_{prop} NP2 Adv

V_{prop} → V_{treat} | V_{treat} V_{effect} | V_{cause} | V_{cause} V_{effect} | V_{weak} W_{info}

NP1 → Noun1 | Noun1 PlantName

NP2 → Phrase | Phrase NP2

Phrase → Noun2 | Noun3 | Nounx | Adj | DiseaseName

Noun1 → { ‘ราก/root’, ‘ใบ/leaf’, ‘ต้น/plant’, ‘ดอก/flower’, ‘ผล/fruit’, ‘เมล็ด/seed’, ... }

Noun2 → { ‘อาการ/symptom’, ‘แผล/scar’, ‘รอย/mark’, ‘ไข้/fever’, ‘ผื่น/rash’, ‘หนอง/pus’, ‘อุจจาระ/stool’, ‘เหงื่อ/sweat’, ‘ลม/gas’, ‘กรด/acidity’, ... }

Noun3 → { ‘อวัยวะ/organ’, ‘บริเวณ/area’, ‘มนุษย์/human’, ‘ศีรษะ/head’, ‘หน้าอก/chest’, ‘กระเพาะ/stomach’, ... }

Nounx → { ‘xxx’ }

(‘xxx’ is any noun words other than noun words of Noun1-Noun3)

PlantName → { ‘โหระพา/basil’, ‘ขิง/ginger’, ‘ขมิ้น/tumeric’, ... }

DiseaseName → { ‘โรคผิวหนัง/skin-disease’, ‘โรคกระเพาะ/gastritis-disease’, ... }

V_{treat} → { ‘บรรเทา/relieve’, ‘ขับ/release’, ‘ลด/reduce’, ‘แก้/stop’, ‘แก้/cure, treat’, ‘รักษา/cure, treat’, ... }

V_{cause} → { ‘ทำให้/cause’, ‘ทำให้เกิด/be caused by’ }

$V_{effect} \rightarrow \{ \text{'คลื่นไส้/nauseate'}, \text{'อาเจียน/vomit'}, \text{'ปวด/pain'}, \text{'เจ็บ/pain'}, \text{'แน่น/constrict'}, \text{'คัน/itchy'}, \dots \}$

$V_{weak} \rightarrow \{ \text{'เป็น/be'}, \text{'มี/have'}, \text{'ใช้เป็น/be'} \}$

$W_{info} \rightarrow \{ \text{'ยา/medicine'}, \text{'สรรพคุณ/property'} \}$

$Adv \rightarrow \{ \text{'ยาก/difficultly'} \}; Adj \rightarrow \{ \text{'เหลว/watery'} \}$

(เมื่อ NP1 และ NP2 คือนามวลี; VP คือกริยาวลี; V_{prop} คือกริยาที่มีแนวความคิดคุณสมบัติ; V_{treat} คือเซตกริยาที่มีแนวความคิดรักษา; V_{cause} คือเซตกริยาที่มีแนวความคิดเหตุ; V_{effect} คือเซตกริยาที่มีแนวความคิดผลหรืออาการ; V_{weak} คือเซตกริยาอ่อนซึ่งต้องตามด้วยคำต่างๆที่ให้สารสนเทศ (W_{info}) ร่วมกับกริยานี้เป็นแนวความคิดสรรพคุณทางยา)

เนื่องจากภาษาไทยมีจำนวนคำศัพท์ไม่มากเท่ากับภาษาอังกฤษ ฉะนั้นความหมายหรือแนวความคิดเกี่ยวกับเหตุการณ์ต่างๆ ตัวอย่างเช่น เหตุการณ์เกี่ยวกับสรรพคุณทางยา ส่วนใหญ่เกิดจากคำหลายคำร่วมกันให้ความหมายของเหตุการณ์นั้น เช่น ‘ขับ+เหวี่ยง’ ‘คลาย+กล้ามเนื้อ’ ‘ระงับ+อาการ+ปวด’ ‘รักษา+อาการ+คอ+อักเสบ’ ‘บรรเทา+อาการ+แผล+พุพอง+อักเสบ’ เป็นต้น ฉะนั้นงานวิจัยนี้จึงขอเสนอการสกัด N-Word-Co (คือกลุ่มคำที่เกิดร่วมกันมีจำนวน N คำ) ที่มีแนวความคิดสรรพคุณทางยา (a medicinal-property concept) จากกริยาวลีในเอกสารพืชสมุนไพรต่างๆ โดย N-Word-Co แสดงหรือปรากฏเป็นกลุ่มคำ $w_{mpi}; i=1,2,\dots,N$; ที่อยู่ต่อเนื่องในเอกสารดังนี้.

$$\text{N-Word-Co expression} = w_{mp1} + w_{mp2} + w_{mp3} + \dots + w_{mpN}$$

และมีการใช้กฎ N-Word-Co rule ต่อไปนี้เพื่อก่อให้เกิด N-Word-Co

“If $w_{mp1} \in V_{mp}$ then $w_{mp2}, w_{mp3}, \dots, w_{mpN} \in W_{mp}$

Else If $w_{mp1} \in V_{weak} \wedge w_{mp2} \in W_{info}$ then $w_{mp3}, \dots, w_{mpN} \in W_{mp}$.”

(where $V_{mp} = V_{treat} \cup V_{cause}$; V_{mp} is the medicinal-property verb concept set;

$W_{mp} = \text{Noun2} \cup \text{Noun3} \cup \text{Nounx} \cup \text{PlantName} \cup \text{DiseaseName} \cup V_{effect} \cup \text{Adj} \cup \text{Adv}$;

W_{mp} is the medicinal-property word concept set)

ดังนั้น N-Word-Co ที่สกัดได้ จะถูกรวบรวมจัดเก็บเป็นลิสต์ (List) ของ N-Word-Co ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร ทั้งนี้เพื่อใช้เป็นแหล่งคำศัพท์ของกลุ่มคำที่เกิดร่วมกัน และมีแนวความคิดสรรพคุณทางยา ของพืชสมุนไพร

อย่างไรก็ตาม การสกัด N-Word-Co ทำให้เกิดปัญหาเพิ่มขึ้นอีกสองปัญหา คือ 1). ปัญหาการระบุกริยาวลีที่มี N-Word-Co ปัญหานี้สามารถแก้ไขได้โดยใช้ เซตของกริยาต่อไปนี้ V_{cause} , V_{effect} , V_{weak} , และ W_{info} มาระบุกริยาวลีที่มี N-Word-Co และ 2). ปัญหาการหาขอบเขตของ N-Word-

Co หรือ ขนาด N-Word-Co (คือค่า N) งานวิจัยนี้ได้ประยุกต์ใช้ Maximum Entropy ซึ่งเป็นเทคนิคการเรียนรู้ของเครื่องมาเรียนรู้หาขนาด N-Word-Co ในกริยาวลี

ดังนั้นผลลัพธ์ของ N-Word-Co จะถูกนำไปใช้ ระบุ EDU ที่มีแนวความคิดสรรพคุณทางยาของพืชสมุนไพร ด้วยการหาค่า Max-Similarity (candidate_N-Word-Co, List.N-Word-Co) ≥ 0.85 นอกจากนี้ผลลัพธ์ของ N-Word-Co สามารถใช้แก้ปัญหาย่อยที่สามคือปัญหาการหาขอบเขตของ EDUs ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร โดยการหาค่า Max-Similarity (candidate_N-Word-Co, List.N-Word-Co) หากมีค่า < 0.85 แสดงว่าสิ้นสุดขอบเขต EDU ที่มีแนวความคิดสรรพคุณทางยา

และปัญหาสุดท้าย คือ การรวบรวมจัดเก็บความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่สกัดได้อย่างไร เพื่อให้สามารถเข้าถึงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ได้อย่างไม่ซับซ้อน เพื่อแสดงหรือแทนในรูปแบบเครือข่าย (รูปที่ 1) ได้อย่างมีประสิทธิภาพ ดังนั้นงานวิจัยนี้ใช้เมตริกซ์ที่ประกอบด้วยฟีเจอร์ (Feature) ต่อไปนี้ ชื่อพืชสมุนไพร ชื่อส่วนของพืชสมุนไพร และสรรพคุณทางยามาทำการเก็บรวบรวมข้อมูลตามฟีเจอร์ ที่กำหนดเพื่อแสดงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ในรูปแบบเครือข่ายเหมือน (Like Network Format) แบบ CODASYL โดยการเรียงลำดับ (Sorting) ตามชื่อพืชสมุนไพรและ ชื่อส่วนของพืชสมุนไพร (คือ H_i-HC_{ik} set occurrence) ทำให้ได้รายชื่อสรรพคุณทางยาของแต่ละส่วนของพืชสมุนไพร ในทางกลับกัน ถ้าเรียงลำดับตามสรรพคุณทางยา (คือ MP_i-HC_{jk} set occurrence) จะได้รายชื่อพืชสมุนไพรและ ชื่อส่วนของพืชสมุนไพรของแต่ละสรรพคุณทางยาสมุนไพร

2. วัตถุประสงค์

เพื่อสกัดความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารภาษาไทย

3. นิยามคำศัพท์

Medicinal Effect: ผลทางยาซึ่งเป็นคุณสมบัติของยา (Medicinal Property) หรือเรียกว่าสรรพคุณทางยา

HerbPlantPart-MedicinalProperty Relation: ความสัมพันธ์ระหว่างสรรพคุณทางยากับส่วนของพืชสมุนไพร

N-Word-Co: การปรากฏร่วมกัน N คำ ด้วยแนวความคิดสรรพคุณทางยาสมุนไพร

EDU: Elementary Discourse Unit คือประโยคง่ายๆ ธรรมดาไม่ซับซ้อน

NP: Noun Phrase Concept คือความคิดนามวลี

VP: Verb Phrase Concept คือความคิดกริยาวลี

Corpus: คลังคำ

Concept: แนวความคิด

Machine Learning: การเรียนรู้ของเครื่อง

4. ขอบเขตของการวิจัย

4.1. ข้อมูลที่ใช้ทำวิจัย คือ เอกสารภาษาไทยภายใต้โดเมนพืชสมุนไพรไทยซึ่งดาวน์โหลดจากเว็บ

4.2. เอกสารที่ใช้มีลักษณะเป็นแบบกิ่งโครงสร้าง

DRAFT

งานวิจัยที่เกี่ยวข้อง

การวิจัยการสกัดและเก็บรวบรวมความสัมพันธ์ “ส่วนของพืชสมุนไพร-สรรพคุณทางยา” จากเอกสารบนเว็บประกอบด้วยความรู้พื้นฐาน และงานวิจัยก่อนหน้านี้

ความรู้พื้นฐาน

1) Maximum Entropy, ME

ตัวแบบ ME ทำให้ได้ความรู้ที่ว่าตัวแบบที่ดีที่สุดคือ ตัวแบบที่สามารถสอดคล้องกับเซตของข้อบังคับ (Constraints) ที่ถูกกำหนดโดยหลักฐานหรือสถานการณ์ มิฉะนั้นต้องเป็นรูปแบบเดียวกันเท่าที่จะเป็นไปได้ (Csiszar 1996; Berger et al. 1996) Fleischman et al. (2003) ได้จำลองแบบด้วยค่าความน่าจะเป็นของบทบาทหน้าที่ทางความหมาย, r (semantic role, r), เมื่อบอกหรือให้ค่าเวกเตอร์ (Vector) ของ Feature x มาให้ ดังสมการ (3) ต่อไปนี้

$$p(r|x) = \frac{1}{Z_x} \exp\left[\sum_{j=0}^n \lambda_j f_j(r, x)\right] \quad (1)$$

เมื่อ Z_x เป็นค่า Normalization Constant

$f_j(r, x)$ เป็น Feature Function ที่แมป (Map) แต่ละค่า r และองค์ประกอบเวกเตอร์ (Vector Element) ให้กับค่าทวิภาค (Binary Value)

n เป็นจำนวน Feature Functions

λ_j เป็นค่าน้ำหนักของ Feature Functions ที่กำหนดให้

ดังนั้นการจัดแบ่งประเภท/คลาส (class) ที่ได้สุดท้ายคือ r (คลาสใช่ คลาสไม่ใช่ เป็นต้น) คำนวนได้จาก ที่มีค่าความน่าจะเป็นสูงสุดสำหรับเวกเตอร์ของฟีเจอร์ x ใน (1) ที่กำหนดให้ นั้น กับค่าความน่าจะเป็นสูงสุดเมื่อบอกค่า Feature Vector ของ r มาให้

2) Similarity_Score

การคำนวณค่าความเหมือน (Similarity_score) (Biggins et al., 2012) ในสมการ (2) ระหว่างสองประโยค จากคำต่างๆที่อยู่ในเซตของคำที่เป็นเนื้อความในประโยค (Content Word Set) ทั้งสองนั้น หลังจากการกำจัดคำหยุด (Stop Word) และ การหาแก่นคำ (Stemming Words)

$$Similarity_Score = \frac{|S1 \cap S2|}{\sqrt{|S1| \times |S2|}} \quad (2)$$

เมื่อ S1 คือประโยคที่ประกอบด้วย แนวความคิดของคำที่อยู่ในเซตของคำที่เป็นเนื้อความในประโยค S1
 S2 คือประโยคที่ประกอบด้วย แนวความคิดของคำที่อยู่ในเซตของคำที่เป็นเนื้อความในประโยค S2
 (โดย แนวความคิดของคำสำหรับงานวิจัยนี้อยู่ภายใต้ เวิร์ดเน็ต (WordNet) และสารานุกรมไทย
 หลังจากแปลภาษาไทยเป็นภาษาอังกฤษโดยใช้ Longdo.com)

งานวิจัยก่อนหน้า

ได้มีงานวิจัยมากมายที่ได้เสนอเทคนิคต่างๆ สำหรับสกัด “ส่วนของพืชสมุนไพรคุณทางยา” จากเอกสารบนเว็บ เพื่อใช้ในการรักษาสุขภาพ โดยแบ่งออกเป็น 2 แนวทางคือ แนวทางแพทเทิร์นหรือกฎ (Pattern/Rule Based Approach) และแนวทางสถิติรวมถึงการเรียนรู้ของเครื่อง (Statistical Based Approach Including Machine Learning)

1. แนวทางแพทเทิร์นหรือกฎ (Pattern/Rule Based Approach)

Paşca M. (2008) ระบุความรู้ที่เป็นจริงเกี่ยวกับคลาสวัตถุ (Object Class) ต่างๆได้โดยการใช้ อิสระแพทเทิร์น (Is-A pattern) กับการสอบถามหรือคิวรี (Query) ที่มีคีย์เวิร์ด (Key Word) อยู่ด้วย ทำการสกัดคุณสมบัติต่างๆที่อยู่ในรูปนามวลี จากเอกสารบนเว็บและจากส่วนบันทึกคิวรี (Query Log) ด้วยค่า precision 0.8 สำหรับ 100 คลาสที่สามารถระบุได้ จาก 5 คุณสมบัติหรือแอททริบิวต์ (Attribute) ที่สกัดได้

Haruechaiyasak et al., 2011 ได้ประยุกต์ใช้เครื่องมือการทำเหมืองข้อความมาทำการสกัดความสัมพันธ์ที่เด่น (Salient Relation) จากเอกสารพบบนเว็บ (PubMed Articles) ที่สัมพันธ์กับสมุนไพรไทย โดยมีการรู้จำความสัมพันธ์บนพื้นฐานทางโครงสร้างไวยากรณ์คือ “NP1 causal-verb NP2” เมื่อ NP1 คือ เหนเมอนิตีสมุนไพร (Herb Name Entity) และ NP2 คือ เอนิตีสรรพคุณทางยา ความสัมพันธ์ที่สกัดได้แสดงออกมาในรูปแบบเครือข่ายที่มีลักษณะเป็นโครงสร้างต้นไม้ที่แทนความสัมพันธ์ด้านสรรพคุณระหว่างชื่อสมุนไพร หรือ เหนเมอนิตีสมุนไพร (NP1) ที่เป็นโหนดราก (Root Node) และเอนิตีสรรพคุณทางยาสมุนไพร (NP2) เป็นโหนดใบ (Leaf Node) รวมทั้งในทางกลับกันคือ NP1 เป็นเอนิตีสรรพคุณทางยา NP2 เป็นสมุนไพร อย่างไรก็ตาม เครื่องมือนี้ไม่สามารถประยุกต์ใช้กับงานวิจัยนี้ซึ่งสรรพคุณทางยาสมุนไพรที่มีลักษณะเป็นเหตุการณ์ที่แสดงด้วยกริยาวลี

2. แนวทางสถิติรวมถึงการเรียนรู้ของเครื่อง (Statistical Based Approach Including Machine Learning)

Weeber M. and Vos. R.(1998) ได้กล่าวถึงคุณสมบัติของฤทธิ์ยาจำเป็นต้องคำนึงถึง 3 เรื่องหลักคือ ยา (A) ผลที่แสดงออกทางกายภาพ (Physiological Effect)(B) และ โรค (C) และความสัมพันธ์ระหว่าง 3 เรื่องดังกล่าวเป็น $A \rightarrow B$, $B \rightarrow C$, ทำให้ได้ $A \rightarrow C$ ดังนั้น Weeber M. and Vos. R.(1998) เสนอการสกัดความรู้ทางการแพทย์โดยการหาความสัมพันธ์ระหว่างคำ (ซึ่งอยู่ในรูปของนามวลี) A, B, และ C จากเอกสารทางการแพทย์โดยใช้แนวทางสถิติด้วยการแยกประเภทคำก่อนที่จะทำการหาความสัมพันธ์ระหว่างคำ ซึ่งได้ค่า precision =0.08 recall =0.57

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 ($A \rightarrow B$ และ $B \rightarrow C$ ทำให้ได้ความสัมพันธ์แบบการส่งผ่าน (Transitive Association) คือ $A \rightarrow C$) มาประยุกต์ใช้ โดยกำหนดให้ A คือพันธุกรรม B คือ ส่วนผสมที่สามารถควบคุม A และ C คือ ยาสมุนไพรจีน เพื่อการบอกเป็นนัยของ $A \rightarrow C$ เมื่อ $A \rightarrow B$ และ $B \rightarrow C$ ปรากฏขึ้นในเอกสารอย่างมีนัยสำคัญ ผลการวิจัยของ Fang et al.,(2008) ได้ precision =0.91 อย่างไรก็ตามวิธีของ Fang et al.,(2008)อยู่บนพื้นฐานของการใช้แต่เพียงนามวลี

Riaz M. and Girju R.(2014) ได้ใช้โมเดลโปรแกรมเชิงเส้นแบบที่คำตอบที่ดีที่สุดเป็นเลขจำนวนเต็ม (ซึ่งมีลักษณะคล้าย Naïve Bayes) มาทำการเรียนรู้ความสัมพันธ์เหตุและผล (Causal Relation)จากคู่ กริยา-นามที่ได้กำกับทั้งหมด (annotated verb-noun pairs) จากเฟรมเน็ต (FrameNet) และเวิร์ดเน็ต(WordNet) รวมทั้งใช้ความรู้เกี่ยวกับประเภทความหมาย (Semantic Class knowledge) และทั้งปัจจัยต่างๆทางภาษาศาสตร์ ตัวอย่างเช่น “People died in hurricane” มีคำนาม ‘hurricane’ และกริยา ‘die’ เป็นความสัมพันธ์เหตุและผล อย่างไรก็ตามความสัมพันธ์เหตุและผลของ Riaz M. and Girju R. มีลักษณะเป็นความสัมพันธ์ระหว่างสองเหตุการณ์โดยเหตุการณ์หนึ่งแสดงโดยคำนาม และอีกเหตุการณ์หนึ่งแสดงโดยคำกริยา

Recupero, D.R et al.,(2015) ได้ใช้ SHELDON ซึ่งเป็นเครื่องมือ NLP(Natural Language Processing Tool) ทำการสกัดความรู้ โดยเฉพาะความรู้เกี่ยวกับความสัมพันธ์ระหว่างสองเอนทิตีที่เป็นคำศัพท์คำนาม (two noun-term entities) นอกจากนี้มีบางความสัมพันธ์ที่ต้องคำนึงถึงการให้คะแนนสำหรับเอนทิตีคำศัพท์คำนามที่มีความคิดเห็นเชิงลบเพื่อใช้เป็นทริกเกอร์ (Trigger) ตอบสนองต่อความคิดเห็นเชิงลบแล้วทำให้เกิดความสัมพันธ์หรือความเชื่อมโยงกับคำกริยา ‘Tell’ กับประโยคตัวอย่าง

อย่างไรก็ตามจากงานวิจัยก่อนหน้านี้ทั้งสองแนวทางนั้น (Weeber M. and Vos. R., 1998; Fang et al.,2008; Paşca M., 2008; Haruechaiyasak et al.,2011; Riaz M. and Girju R.2014;

Recupero, D.R et al.,2015) ไม่เหมาะสำหรับงานวิจัยนี้ที่มุ่งสกัดความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ซึ่งอยู่บนพื้นฐานการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของเอนทิตีพืชสมุนไพรไทยจากเอกสารภาษาไทยโดยมีสรรพคุณเหล่านี้แสดงอยู่ในรูปของกริยาวลี และหลายๆกริยาวลีอยู่ต่อเนื่องกันและมีความสัมพันธ์กับหนึ่งเอนทิตีสมุนไพรหรือส่วนของสมุนไพรซึ่งจะอยู่ในรูปของนามวลีที่ส่วนใหญ่มักจะละ

ปัญหาการสกัดความสัมพันธ์ “ส่วนของพืชสมุนไพร-สรรพคุณทางยา” จากเอกสารบนเว็บ

ปัญหาการสกัดความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสาร ประกอบด้วย 3 ปัญหาหลักคือ

1) ปัญหาการระบุเอนทิตีพืชสมุนไพร (Herb-Plant Entity) และส่วนของพืชสมุนไพร (Herb-Plant Part) ซึ่งมักจะมีลักษณะปัญหาเป็น Zero Anaphora (การละคำสรรพนามที่ใช้อ้างอิงเอนทิตีพืชสมุนไพร) และ Textual Ellipsis (การละคำนามที่เป็นเอนทิตีพืชสมุนไพร)

ตัวอย่างที่ 1 ปัญหา Zero Anaphora

กระเทียม

EDU1: “(กลีบกระเทียม)/NP1 (ใช้เป็นยาขับลม)/VP”

EDU2 : “[(กลีบกระเทียม)/NP1] (แก้อาเจียน)/VP”

คำนามที่อยู่ภายในวงเล็บก้ามปู [...] คือ Zero Anaphora ซึ่งปรากฏใน EDU2 ปัญหานี้แก้โดยอ้างอิงคำนามที่เป็นประธานของ EDU ก่อนหน้าซึ่งก็คือกระเทียม

ตัวอย่างที่ 2 ปัญหา Textual Ellipsis

EDU3: “(จาก [กระเทียม])/NP1 (ใช้เป็นยาขับปัสสาวะ)/VP”

คำนามที่อยู่ภายในวงเล็บก้ามปู [...] คือ Textual Ellipsis ปัญหานี้แก้โดยอ้างอิงชื่อเรื่อง (Topic name) ซึ่งก็คือกระเทียม

2) ปัญหาการระบุ EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพรของเอนทิตีพืชสมุนไพร

ตัวอย่างที่ 3 ต่อไปนี้ จะทราบได้อย่างไรว่า EDU ต่างๆต่อไปนี้มี EDU อันไหนที่มีแนวความคิดสรรพคุณทางยาสมุนไพร

ตัวอย่างที่ 3

“พริกไทย (Piper nigrum Linn.)”

-สรรพคุณ

เปลือกของพริกไทยมีน้ำย่อยสำหรับย่อยไขมัน^{EDU} ด้วยเหตุนี้ตำราโบราณจึงเชื่อกันว่าพริกไทยสามารถลดความอ้วนได้^{EDU} พริกไทยช่วยกระตุ้นปฏิกิริยาที่^{EDU} เพื่อให้กระเพาะอาหารหลั่งน้ำย่อยได้มากขึ้น^{EDU} พริกไทยดำมีรสเผ็ด^{EDU} เมื่อรับประทานเข้าไป^{EDU} จะรู้สึกอุ่นวาบที่ท้อง^{EDU} ช่วยขับลม^{EDU} ขับเหงื่อ^{EDU} ขับปัสสาวะ^{EDU} แก้ท้องอืด^{EDU} ท้องเฟ้อ^{EDU} แก้ไข้มาลาเรีย^{EDU} แก้หิวาตกโรค^{EDU} ใช้ก้านพริกไทย 10 ก้าน^{EDU} บดให้ละเอียด^{EDU} แล้วต้มกับน้ำ 8 แก้ว^{EDU} ใช้เป็นยาล้างแผลที่^{EDU} ฝี, สารพิเพอรินในพริกไทยสามารถใช้เป็นยาฆ่าแมลง^{EDU} ซึ่งไม่เป็นอันตรายต่อมนุษย์^{EDU} โดยนำผลพริกไทยมาทุบให้แตก^{EDU} แล้วใช้^{EDU}”

สรรพคุณทางยาสมุนไพรที่ปรากฏบนเอกสารส่วนใหญ่แสดงออกมาในรูปของเหตุการณ์ (Event) ที่อธิบายด้วยกริยาวลี ตัวอย่างเช่น ‘ขับ+เหงื่อ’ ‘คลาย+กล้ามเนื้อ’ ‘ระงับ+อาการ+ปวด’ ‘รักษา+อาการ+คอ+อักเสบ’ ‘บรรเทา+อาการ+แผล+พุพอง+อักเสบ’ เป็นต้น ฉะนั้นงานวิจัยนี้จึงขอเสนอการสกัด N-Word-Co (คือกลุ่มคำที่เกิดร่วมกันมีจำนวน N คำ มีแนวความคิดสรรพคุณทางยา) โดย N-Word-Co ที่สกัดได้ จะถูกรวบรวมจัดเก็บเป็นลิสต์ (List) ทั้งนี้เพื่อใช้เป็นแหล่งคำศัพท์เกี่ยวกับแนวความคิดสรรพคุณทางยาของพืชสมุนไพรในรูปของ N-Word-Co อย่างไรก็ตามการสกัด N-Word-Co ประกอบด้วยสองปัญหา คือ 1) ปัญหาการระบุกริยาวลีที่มี N-Word-Co 2) ปัญหาการหาขอบเขตของ N-Word-Co ฉะนั้นงานวิจัยนี้แก้ปัญหาทั้งสองนี้โดยใช้ เซตของกริยาต่อไปนี้ V_{cause} , V_{effect} , V_{weak} , และ W_{info} มาระบุกริยาวลีที่มี N-Word-Co และได้ประยุกต์ใช้ Maximum Entropy ซึ่งเป็นเทคนิคการเรียนรู้ของเครื่องมาเรียนรู้หาขนาด N-Word-Co ในกริยาวลี

ดังนั้นลิสต์ของ N-Word-Co จะถูกนำไปใช้ ระบุ EDU ที่มีแนวความคิดสรรพคุณทางยาของพืชสมุนไพร ด้วยการหาค่า Max-Similarity (candidate_N-Word-Co, List.N-Word-Co) ≥ 0.85

3) ปัญหาการหาขอบเขตของ EDUs (EDU Boundary) ที่มีแนวความคิดสรรพคุณทางยาสมุนไพรของเอนทิตีพืชสมุนไพร

จากตัวอย่างที่ 3 จะทราบได้อย่างไรว่าสรรพคุณทางยาสมุนไพรต่อไปนี้

.....เมื่อรับประทานเข้าไป^{EDU} จะรู้สึกอุ่นวาบที่ท้อง^{EDU} ช่วยขับลม^{EDU} ขับเหงื่อ^{EDU} ขับปัสสาวะ^{EDU} แก้ท้องอืด^{EDU} ท้องเฟ้อ^{EDU} แก้ไข้มาลาเรีย^{EDU} แก้หิวาตกโรค^{EDU} ใช้ก้านพริกไทย 10 ก้าน^{EDU}

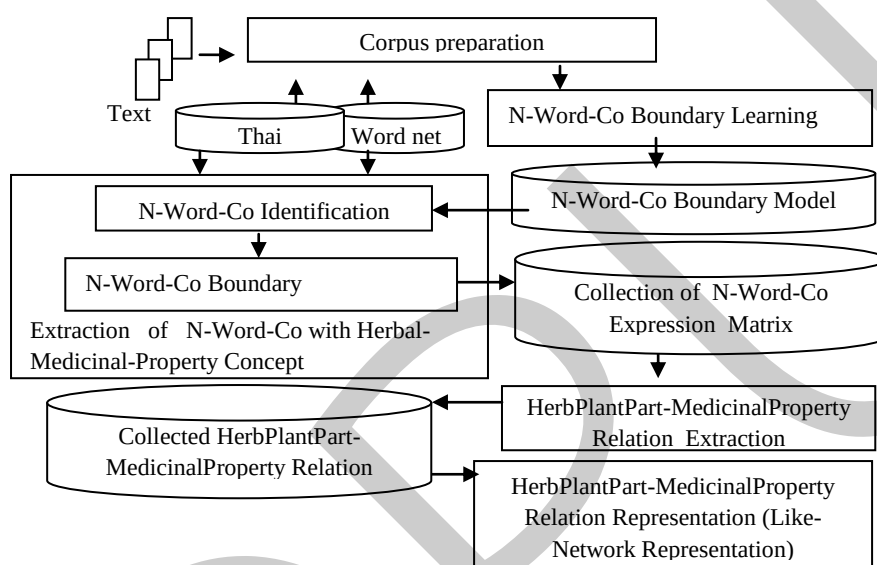
มี “แก้หิวาตกโรค^{EDU}” เป็นขอบเขตของ EDUs ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร

ปัญหานี้แก้ไขโดยใช้ N-Word-Co List โดยการหาค่า Max-Similarity (candidate_N-Word-Co, List.N-Word-Co) หากมีค่า < 0.85 แสดงว่าสิ้นสุดขอบเขต EDU ที่มีแนวความคิดสรรพคุณทางยา

ปัญหาการจัดเก็บความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่สกัดได้อย่างไร เพื่อให้สามารถเข้าถึงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ได้อย่างไม่ซับซ้อน เพื่อแสดงหรือแทนในรูปแบบเครือข่าย (รูปที่ 1) ได้อย่างมีประสิทธิภาพ ดังนั้นงานวิจัยนี้ใช้เมตริกซ์ที่ประกอบด้วยฟีเจอร์ (Feature) ต่อไปนี้ Herb ID , HerbName (ชื่อพืชสมุนไพร), HerbPlantPartConcept (ชื่อส่วนของพืชสมุนไพร), Herbal-Medicinal-Property Concept (ความคิดสรรพคุณทางยา) , และ Relation ID มาทำการเก็บรวบรวมข้อมูลเป็นตาราง เพื่อแสดงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ในรูปแบบเครือข่ายเหมือน (Like Network Format)แบบ CODASYL โดยการเรียงลำดับ (Sorting) ตาม ชื่อพืชสมุนไพรและ ชื่อส่วนของพืชสมุนไพร (คือ Hi-HCik set occurrence) ทำให้ได้รายชื่อสรรพคุณทางยาของแต่ละส่วนของพืชสมุนไพร ในทางกลับกัน ถ้าเรียงลำดับตามสรรพคุณทางยา (คือ MPi-HCjk set occurrence) จะได้รายชื่อ พืชสมุนไพรและ ชื่อส่วนของพืชสมุนไพรของแต่ละสรรพคุณทางยาสมุนไพร

กรรมวิธีดำเนินงาน

ระบบงานโดยสรุปสำหรับการการสกัดและเก็บรวบรวมความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารบนเว็บประกอบด้วยขั้นตอนต่าง ๆ ดังต่อไปนี้ (ดังแสดงในรูปที่ 2) ขั้นตอนการเตรียมคลังข้อมูล (Corpus Preparation) ขั้นตอนการเรียนรู้ขอบเขต N-Word-Co (N-Word-Co Boundary Learning) ขั้นตอนการสกัด N- Word-Co ขั้นตอน การสกัดความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ (HerbPlantPart-MedicinalProperty Relation Extraction) และขั้นตอนการรวบรวมเก็บสะสมและแทนความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ (Collection and Representation of HerbPlantPart-MedicinalProperty Relation)



รูปที่2. กรอบงานวิจัย

การทดลอง (ตามกรอบแนวคิดการวิจัย)

1) การเตรียมคลังคำ (Corpus Preparation)

ข้อมูลภาษาไทยที่ใช้ในการวิจัยนี้เป็นข้อมูลทางด้านพืชสมุนไพรไทย สามารถดาวน์โหลดได้จากเว็บไซต์
เว็บไซต์ของโครงการอนุรักษ์พันธุกรรมพืชอันเนื่องมาจากพระราชดำริฯ

(http://www.rspg.or.th/plants_data/herbs/herbs_200.htm) และสารานุกรมไทย (<http://kanchanapisek.or.th/kp6/sub/book/book.php?book=14&chap=10&page=chap10.htm>) แล้วผ่าน

ขั้นตอนการเตรียมคลังคำ ที่มีการกำกับ tag ตั้งแต่ระดับคำโดยใช้โปรแกรมตัดคำจาก Nias Lab (ของ
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์) จนถึงระดับประโยค
จำนวน 3,600 EDUs ซึ่งแบ่งออกเป็น 3 ส่วน คือ ส่วนที่ 1 จำนวน 2000 EDUs สำหรับการเรียนรู้หา
เซตกริยาสรพคุณทางยาสมุนไพรรซึ่งประกอบด้วย Vtreat และ Vcause และการเรียนรู้หาขนาด N-
Word-Co (ที่มีแนวความคิดสรพคุณทางยาสมุนไพรร) เพื่อใช้ระบุ EDU ที่มีแนวความคิดสรพคุณ

ทางยาสมุนไพร ส่วนที่ 2 จำนวน 1000 EDUs สำหรับสกัด N-Word-Co และรวบรวม เป็นลิสต์ N-Word-Co เพื่อใช้สำหรับสกัดความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ และส่วนที่ 3 จำนวน 600 EDUs สำหรับสกัดความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’

2) การเรียนรู้ขอบเขต N-Word-Co (N-Word-Co Boundary Learning)

นำคลังคำส่วนที่ 1 มาทำการกำกับคำที่เกี่ยวข้องกับ N-Word-Co พร้อมแนวความคิดสรรพคุณทางยาสมุนไพร รวมทั้ง NP1 ที่แสดงแนวความคิดส่วนของพืชสมุนไพรหลังจากที่ผ่าน หาแก่นคำ และกำจัดคำหยุด ดังแสดงในรูปที่ 3 โดยใช้ WordNet(<http://word-net.princeton.edu/obtain>) และ Thai-Herb-Encyclopedia (<http://kanchanapisek.or.th/kp6/BOOK14/chapter10/chap10.htm>) หลังจากแปลภาษาไทยเป็นภาษาอังกฤษโดย Lexitron (Thai-English Dictionary) (<http://lexitron.nectec.or.th/>)

“ต้นสะเดา.....
ผลสะเดาทำให้ อหกรับประทานอาหารตลอดเวลา EDU1 ทำให้สดชื่น EDU2 รากแก้เสมหะติดคอ EDU3 เปลือกใช้เป็นยาแก้ไข้อย่างดี EDU4 ...”
“Neem Tree.....
A neem fruit makes [you] always feel like to eat food. EDU1 [The neem fruit] makes you being fresh EDU2 Root clears phlegm sticking in throat. EDU3 Bark is used as an antipyretic drug EDU4 ...”
<Topic_name Entity-concept=neem tree/herb>ต้นสะเดา</Topic_name>.....
<EDU1 ><NP1 concept= neem fruit/herb-component>ผล/ncn สะเดา/herbNE </NP1>
<VP medicinalPropertyType=Y> <N-Word-CoExpression N=4 words concept= ‘have an appetite’>
 < w-mp1: setType=‘cause-verb’; concept=‘make’ boundary =‘yes’>ทำให้</ w-mp1>
 < w-mp2: setType=‘effect-verb’; concept= ‘feel like’ boundary =‘yes’>อยาก</ w-mp2>
 < w-mp3: setType=‘effect-verb’; concept=‘eat’ boundary =‘yes’>รับประทาน</ w-mp3>
 < w-mp4: setType=‘noun2’; concept= ‘food’ boundary= ‘yes’>อาหาร</ w-mp4>
 < w-mp5: setType=‘noun2’; concept= ‘always’ boundary=‘no’>ตลอดเวลา</ w-mp5>
</N-Word-CoExpression ></VP> </EDU1>
<EDU2><NP1 concept= neem fruit/herb-component>φ</NP1>
<VP medicinalPropertyType=Y><N-Word-CoExpression N=2words concept= ‘feel refreshed’>
 < w-mp1: setType=‘cause-verb’; concept=‘make’ boundary =‘yes’>ทำให้</ w-mp1>
 < w-mp2: setType=‘effect-verb’; concept= ‘be fresh’ boundary =‘yes’>สดชื่น</ w-mp2>
</N-Word-CoExpression >)/VP </EDU2>
<EDU3 ><NP1 concept= root/herb-component>ราก/ncn </NP1>
<VP medicinalPropertyType=Y> <N-Word-CoExpression N=2words concept= ‘clear phlegm’>
 < w-mp1: setType=‘treat-verb’; concept=‘clear’ boundary=‘yes’>แก้</ w-mp1>
 < w-mp2: setType=‘noun2’; concept= ‘phlegm’ boundary =‘yes’>เสมหะ</ w-mp2>
 < w-mp3: setType=‘effect-verb’; concept=‘stick’ boundary =‘no’>ติด</ w-mp3>
 < w-mp4: setType=‘noun3’; concept= ‘throat’ boundary=‘no’>คอ</ w-mp4>
</N-Word-CoExpression ></VP> </EDU3>
.....
The N-Word-CoExpression tag is the word boundary tag of each N-Word-Co expression. The w-mpi tag is the medicinal-property-word-i tag where i=1,2,...,num. .
The [...] symbol or φ means ellipsis (Zero Anaphora)

รูปที่3. การกำกับคำของ N-Word-Co ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร และNP1 ที่มีแนวความคิดส่วนของพืชสมุนไพร

การเรียนรู้หาขนาดของ N-Word-Co ของ VP ของ EDU โดย เครื่องจักรเรียนรู้ Maximum Entropy (ME) (Berger et al.,1996) ซึ่งเป็นการหาค่าความน่าจะเป็นของส่วนความหมาย r (semantic role, r) เมื่อกำหนดเวกเตอร์ของฟีเจอร์ x ดังสมการ (1) ของหัวข้องานวิจัยที่เกี่ยวข้อง

จาก (1) สามารถนำมาประยุกต์ใช้กับงานวิจัยนี้เพื่อหาคลาส r ว่าขอบเขตสิ้นสุดหรือไม่ เมื่อ $p(r|x)$ คือ ค่าความน่าจะเป็นสูงสุด ($\arg\max p(r|x)$) เพื่อที่หาขนาดของ N-Word-Co หรือ คลาสขอบเขตของ N-Word-Co ว่าเป็นคลาส ending หรือ continuing จากกริยาวลีทั้งหมดในคลังคำที่เตรียมไว้ ก่อนที่จะมีการเรียนรู้ขนาดของ N-Word-Co จะต้องมีการหาจุดเริ่มต้นของ N-Word-Co โดยใช้ กฎ N-Word-Co rule กับ VP (ที่ได้กล่าวในหัวข้อ 5 ความเป็นมาของปัญหา) หลังจากที่ผ่านมา หาแก่นคำ และ กำจัดคำหยุด ดังนั้น r คือ คลาสขอบเขตของ N-Word-Co เมื่อ $r = 0$ แสดงว่าขอบเขตสิ้นสุด (ending) มิฉะนั้น $r = 1$ (continuing) และ x เป็นไบนารีเวกเตอร์ของพีเจอร์แนวความคิดคำ (A binary vector of word-concept features) และ x เป็นเวกเตอร์ที่มีคู่คำ (คู่แนวความคิดคำ, w_{mpi} w_{mpi+1}) โดยเฉพาะแนวความคิดสรรพคุณทางยาสมุนไพร รวมอยู่ด้วย เมื่อ $w_{mpi} \in W_{mp}$ และ $w_{mpi+1} \in W_{mp}$;

if $w_{mp1} \in V_{mp}$ then $i=2,3,...,N$ or if $w_{mp1} \in V_{weak} \wedge w_{mp2} \in W_{info}$ then $i=3,4,...,N$ ดังแสดงใน (2) โดยทุกคู่คำ ได้จากการเลื่อนกรอบหน้าต่างที่มีขนาดสองคำอยู่ติดกันไปด้วยระยะทางหนึ่งคำบน VP หลังจากที่ผ่านมา หาแก่นคำ และกำจัดคำหยุด

$$p(r | x) = \arg \max_r \frac{1}{Z} \exp \left(\sum_{j=1}^n \lambda_j f_{yes,mpi,j}(r, w_{mp}) + \sum_{j=1}^n \lambda_j f_{no,mpi,j}(r, w_{mp}) + \sum_{j=1}^n \lambda_j f_{yes,mpi+1,j}(r, w_{mp+1}) + \sum_{j=1}^n \lambda_j f_{yes,mpi+1,j}(r, w_{mp+1}) \right) \quad (2)$$

3) การสกัด N- Word-Co (Extraction of N-Word-Co with Herbal-Medicinal-Property Concept) เพื่อสะสมเป็นลิสต์ N-Word-Co

เนื่องจาก N-Word-Co แสดงถึงเหตุการณ์ด้วยกริยาวลี การหาคำเริ่มต้นของ N-Word-Co ซึ่งเป็น คำกริยาสามารถหาได้ด้วยการใช้ กฎ N-Word-Co rule (หลังจากที่ผ่านมา หาแก่นคำ และกำจัดคำหยุด) มาระบุหา EDU ที่มีกริยาตัวแรกของ VP เป็นสมาชิกใน V_{mp} ดังนี้ : If ($w_{mp1} \in V_{mp}$) $\vee$ ($w_{mp1} \in V_{weak} \wedge w_{mp2} \in W_{info}$) then w_{mp1} is the first word of N-Word-Co

เมื่อพบ VP มีคำแรกเป็นไปตามกฎ N-Word-Co rule การหาขอบเขตของ N-Word-Co ก็ จะ เริ่มขึ้นโดยใช้ค่า λ ที่ได้จากการเรียนรู้กับฟังก์ชันพีเจอร์สำหรับการหาขนาด N-Word-Co ดังแสดงใน สมการ (2) คำนวณหาขอบเขต N-Word-Co หรือ ขนาด N-Word-Co ตามอัลกอริทึมที่แสดงในรูปที่ 3

```

Assume that each EDU is represented by (NP1 VP).
NP1 → Noun1 | Noun1 Herb-Name ;
where Noun1 is the plant component set.
 $V_{mp} = V_{treat} \cup V_{cause}$ ;  $W_{mp} = Noun2 \cup Noun3 \cup Nounx \cup Name-Entity \cup V_{effect} \cup Adj \cup Adv$ 
where  $V_{mp}$  is the medicinal-property verb concept set
and  $W_{mp}$  is the medicinal-property word concept set.
 $W_{info} = \{ 'ยา/medicine', 'สรรพคุณ/property' \}$ 

N_WORD_CO_BOUNDARY_DETERMIATON /*by ME
1  $NWC \leftarrow \emptyset$ ;  $i=1$  /* NWC is the N-Word-Co expression.
   /* NP1word is a word occurrence on NP1.
   /* wmpi is a word at the  $i^{th}$  element of VP &
    $i=1,2,..endOfVerbPhrase$ 
2 If (NP1word ∈ Noun1) ∧ (wmpi ∈  $V_{mp}$ )
3 {  $NWC \leftarrow wmpi$ 
4 Else If (NP1word ∈ Noun1) ∧ (wmpi ∈  $V_{weak}$ ) ∧
   (wmpi+1 ∈  $W_{info}$ )
5  $NWC \leftarrow wmpi + wmpi+1$ ;
6  $i = i+1$ ;
7  $r \leftarrow 1$  /* r is the N-Word-CO boundary classes
   (r=0:boundary is ending, otherwise r=1)

8 while (r=1) ∧ (wmpi ∈  $W_{mp}$ ) ∧ ( $i \leq endOfVerbPhrase$ ) do
9 {  $NWC \leftarrow NWC \cup wmpi$ ;  $i \leftarrow i+1$ ;
   Equation (2)
0
1 } return NWC
1

```

รูปที่ 4. อัลกอริทึมหาขอบเขต N-Word-Co

ส่วนการหาแนวคิดของ N-Word-Co เกี่ยวกับ สรรพคุณทางยาสมุนไพร สามารถหาได้จาก V_{treat} , V_{effect} , V_{cause} , V_{weak} , $Noun2$, $Noun3$, และ Adj แต่ไม่รวมคำที่ปรากฏอยู่เสมอเหมือนบอกเป็นนัย (Cue word) เช่น 'ยา/medicine' และ 'สรรพคุณ/property' โดยแนวความคิดต่างๆที่เกี่ยวกับสรรพคุณทางยาสมุนไพรได้จากการอ้างอิง WordNet และ Thai-Herb-Encyclopedia หลังผ่านการแปลภาษาไทยเป็นภาษาอังกฤษด้วย Lexitron ดังนั้น N-Word-Co ที่สกัดได้พร้อมกับแนวความคิดที่ประมวลได้ ดังแสดงในตารางที่ 1 ซึ่งใช้สำหรับการสกัดความรู้เชิงความสัมพันธ์ 'ส่วนของพืชสมุนไพร-สรรพคุณทางยา' จากเอกสาร

ตารางที่ 1 ตัวอย่าง ลิสต์ N-Word-Co

Verb phrase example	N-Word-Co List	
	N-Word-Co Expression	N-Word-Co Concept
“ดอกกระเจียวแดงลดความดันเลือดของผู้ป่วย/The rosella flower reduces blood pressure of patients.” VP=“ลด/reduce-V_{treat} ความดัน/pressure-Noun2 เลือด/blood-Noun3 ของ/of-prep ผู้ป่วย/patients-Noun3”	‘ลด/reduce-V_{treat} ความดัน/pressure-Noun2 เลือด/blood-Noun3’	To reduce blood pressure
“ใบตำลึงทำให้สายตาดีขึ้น/The Ivy-Gourd leaves improve better eyesight.” VP=“ทำให้/improve-V_{cause} สายตา/eyesight-Noun3 ดีขึ้น/better-Adj”	‘ทำให้/improve-V_{cause} สายตา/eyesight-Noun3’	To improve eyesight
“ใบสะระแหน่บรรเทาเจ็บคอรุนแรง/The mint leaves relieve sore throat.” VP=“บรรเทา/relieve-V_{treat} เจ็บ/pain-V_{effect} คอ/throat-Noun3 รุนแรง/badly-Adv”	‘บรรเทา/relieve-V_{treat} เจ็บ/pain-V_{effect} คอ/throat-Noun3’	To relieve sore-throat
“ใบโหระพาขับของเสียในรูปแบบเหงื่อ/The sweet basil leaves excrete waste in the form of sweat.” VP=“ขับ/excrete-V_{treat} ของเสีย/waste-Nounx ใน/in-prep รูปแบบ/form-Nounx เหงื่อ/sweat-Noun2”	‘ขับ/excrete-V_{treat} ของเสีย/xxx-Nounx รูปแบบ/xxx-Nounx เหงื่อ/sweat-Noun2’	To excrete sweat
“ขิงใช้เป็นสมุนไพรขับเหงื่อ/Ginger is an herb for sweating” VP=“ใช้เป็น/be-V_{weak} สมุนไพร/herb-W_{info} ขับ/excrete-V_{treat} เหงื่อ/sweat-Noun2”	‘ใช้เป็น/be-V_{weak} สมุนไพร/herb-W_{info} ขับ/excrete-V_{treat} เหงื่อ/sweat-Noun2’	To excrete sweat
“รางจืดเป็นสมุนไพรถอนพิษทุกชนิด/Thunbergia is an herb for counteracting of all poisons.” VP=“เป็น/be-V_{weak} สมุนไพร/herb-W_{info} ถอน/counteract-V_{treat} พิษ/poison-Noun2”	‘ใช้เป็น/be-V_{weak} สมุนไพร/herb-W_{info} ถอน/counteract-V_{treat} พิษ/poison-Noun2’	To counteract poison

4) การสกัดความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ (HerbPlantPart-MedicinalProperty Relation Extraction)

การสกัดความรู้เชิงความสัมพันธ์ นี้ประกอบ 3 ขั้นตอน บนพื้นฐานการอ้างอิง WordNet และ Thai-Herb-Encyclopedia หลังผ่านการแปลภาษาไทยเป็นภาษาอังกฤษด้วย Lexitron : (1) การระบุชื่อพืชสมุนไพรและส่วนของพืชสมุนไพร (2) การระบุ EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร (3) การหาขอบเขต EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร

4.1. การระบุชื่อพืชสมุนไพรและส่วนของพืชสมุนไพร

ชื่อพืชสมุนไพรได้ชื่อเรื่อง (Topic name) ของเอกสาร เช่น ‘โหระพา/Basil’ ‘ขิง/Ginger’ ‘สะเดา/Neem’ เป็นต้น สำหรับส่วนของพืชสมุนไพร ได้จาก NP1 ของ EDU เช่น ‘ใบโหระพา/Basil leaves’ ‘เหง้าขิง/A Ginger root stock’ เป็นต้น แต่ถ้ามีปัญหา Zero-Anaphora และ Textual Ellipsis เกิดขึ้น ก็สามารถแก้ไขโดยใช้ Agent หรือประธานของ EDU ก่อนหน้า สำหรับการเกิด Zero-Anaphora ส่วนกรณี Textual Ellipsis สามารถแก้ไขโดยการอ้างอิงถึงชื่อของชื่อเรื่อง นอกจากนี้ Product Derivative (ผลผลิตที่พัฒนามา) ก็สามารถอ้างอิงถึง Agent ของ EDU ก่อนหน้า และชื่อเรื่องได้ด้วยตัวอย่างเช่น

Topic name: มะขามป้อม/ Indian Goose-Berry

EDU1: “เมล็ดมะขามป้อมมาเผา/ Indian goose-berry seeds are burnt.”

EDU2: “เมล็ดนั้นผสมกับน้ำมันพืช/The seeds are mixed with vegetable oil.”

EDU3: “น้ำมันผสมนี้ใช้ทาแก้คัน/The mixed oil is applied to stop itching.”

จากตัวอย่างน้ำมันผสมใน EDU3 คือ NP1 และเป็น Product Derivative อ้างถึง ‘เมล็ดมะขามป้อม’

4.2. การระบุ EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร

ถ้า NP1 ของ EDU มีแนวความคิด ส่วนของพืชสมุนไพร อยู่ใน Noun1 ทำให้ VP ของ EDU นี้ ใช้หา แนวความคิดสรรพคุณทางยาสมุนไพร จากลิสต์ N-Word-Co ในตารางที่ 1 ได้โดยใช้ กฎ N-Word-Co rule แล้วตามด้วย การหาร ขนาด N-Word-Co ได้ Candidate N-Word-Co ซึ่งถูกนำไปหาแนวความคิดสรรพคุณทางยาสมุนไพรโดยการหาค่า Max Similarity Score (MaxSimScore) ในสมการ (3) (Biggins et al., 2012) กับ N-Word-Co ที่สะสม (N-Word-Co collected) ในลิสต์ N-Word-Co ถ้าค่า MaxSimScore ≥ 0.85 แสดงว่า เกิดคู่ที่เหมือนกันมากที่สุด เพื่อที่จะได้แนวความคิดสรรพคุณทางยาสมุนไพร

$$MaxSimScore = \max_{MatrixSize} \frac{|NWordCo_{collected} \cap NWordCo_{candidate}|}{\sqrt{|NWordCo_{collected}| \times |NWordCo_{candidate}|}} \quad (3)$$

4.3.การหาขอบเขต EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร

ขั้นตอนนี้เป็น การหาขอบเขต EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร หลังจากที่สามารถระบุ EDU หรือ EDU_i ที่มีแนวความคิดสรรพคุณทางยาสมุนไพรได้ (ตามหัวข้อ 4.2) โดยการหาแนวความคิด ส่วนของพืชสมุนไพรใน NP1 ของ EDU_{i+1} (กล่าวคือมีแนวคิดเดียวกันกับ NP1 ของ EDU_i) และ VP ของ EDU_{i+1} มี Candidate N-Word-Co ที่ตรงกับ N-Word-Co ที่สะสมในลิสต์ N-Word-Co ด้วยค่า Max Similarity Score จากนั้นก็ดำเนินการ หัวข้อ 4.3 นี้ไปเรื่อยๆกับ EDU ถัดไป จนกระทั่ง แนวความคิด ส่วนของพืชสมุนไพรใน NP1 ของ EDU ถัดไป (EDU_{i+1}) ไม่เท่ากับของ EDU ก่อนหน้า หรือ VP ของ EDU ถัดไป (EDU_{i+1}) ไม่มี/มี Candidate N-Word-Co ที่ไม่อยู่ในลิสต์ N-Word-Co นั่นก็คือสิ้นสุดของขอบเขต EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร ดังแสดงในอัลกอริทึมการหาขอบเขต EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร (รูปที่ 4)

```

Assume that each EDU is represented by (NP1 VP) after stemming
words and the stop word removal. L is a list of EDUs.
WM is an N-Word-Co Matrix. Rel is a list of
HerbalMedicinalProperty Rel.
NP1 consists of HerbalPlantComponent +HerbalPlantName
TopicName is a herbal-plant name as the document topic name.
HMPB is an herbal-medicinal-property concept boundary

HERBAL_MEDICINAL_PROPERTY_RELATION_EXTRACTIONS
1  i=1 ; j=1; k=1; h=1; HMPB=∅
2  while i≤ Length[l] do
3  {If NP1i =∅ */solvingZeroAnaphora
4    NP1i.HerbalPlantComponent = NP1i-1.HerbalPlantComponent;
5    If NP1i.HerbalPlantName=∅ */solvingTextualEllipsis
6      NP1i.HerbalPlantName= TopicName;
7    };
8  i=1 ; HCj=NP1i.HerbalPlantComponent; temp=HCj;
9  while i≤ Length[l] do
10 {while (temp=HCj)^(i≤ Length[l])do
11 {CNWCj = Determine_Candidate_NWordCo(VPi);
12 For k=1 to nwcMatrixSize
13   if MaxSimScore(CNWCj, WM.NWordCoExpk)≥0.8
14     Conceptj= WM.NWordCoConceptk ;
15   If Conceptj ≠ ∅
16     {HMPB ← HMPB ∪ Conceptj ; Conceptj = ∅ };
17   j=j+1; i=i+1; HCj=NP1i.HerbalPlantComponent;
18 }
19   Relh ← NP1i-1 ∪ HMPB; h=h+1;
20   If temp ≠ HCj
21     temp=HCj;
22 } Return Relh

```

รูปที่ 5. อัลกอริทึมหาขอบเขต การหาขอบเขต EDU ที่มีแนวความคิดสรรพคุณทางยาสมุนไพร

5) การรวบรวมเก็บสะสมและแทนความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ (Collection and Representation of HerbPlantPart-MedicinalProperty Relation) หลังจากนั้นกลับไปดำเนินการหัวข้อที่ 1 – หัวข้อที่ 4 จนจบเอกสาร ก็ทำให้ได้ ความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ตามสำเนาต่างของพืช ฉะนั้นความรู้เชิงความสัมพันธ์ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่สกัดได้จากเอกสารพืชสมุนไพรต่างๆ ก็ถูกรวบรวมสะสมในรูปแบบของเมตริกซ์ ที่ประกอบด้วย 4 ฟีเจอร์คือ HerbName, HerbPlantPart Concept, HerbalMedical-Property Concept, และ Relation ID ดังแสดงใน ตารางที่ 2 ซึ่งสามารถแสดงหรือแทนความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ในรูปแบบเครือข่ายเหมือน (รูปที่ 1) โดยเรียงตาม HerbName และ HerbPlantPart Concept ทำให้ได้รายชื่อสรรพคุณทางยาของแต่ละส่วนของพืชสมุนไพร (คือ Hi-HCik set occurrence, รูปที่ 1) และ ถ้าเรียงลำดับตาม Herbal-Medical-Property จะได้รายชื่อ พืชสมุนไพรและ ชื่อส่วนของพืชสมุนไพรของแต่ละสรรพคุณทางยาสมุนไพร (ตารางที่3) ซึ่งก็คือ (MPi-HCjk set occurrence, รูปที่ 1)

ตารางที่ 2 แสดงเมตริกซ์ของ ความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่เรียงโดย HerbName and HerbPlantPart Concept

Herb ID	HerbName.	HerbPlantPart Concept	Herbal-Medicinal-Property Concept	Relation ID SortedBy PlantName& PlantPart
37	โหระพา/Basil	whole	ReleaseGas	37whRG
37	โหระพา/Basil	whole	RelieveNauseate	37whRN
37	โหระพา/Basil	whole	ExcreateSweat	37whES
37	โหระพา/Basil	whole	ReduceFever	37whRF
37	โหระพา/Basil	whole	ReduceInflame	37whRI
37	โหระพา/Basil	whole	RelieveCough	37whRC
38	สะเดา/Neem	leaf	HaveAppetite	38leHA
38	สะเดา/Neem	leaf	ReleaseGas	38leRG
38	สะเดา/Neem	leaf	ClearPhlegm	38leCP
38	สะเดา/Neem	bark	ReduceFever	38baRF
39	มะนาว/lime	seed	ClearPhlegm	39seCP
39	มะนาว/lime	seed	ReduceFever	39seRF
39	มะนาว/lime	root	ReduceFever	39rtRF
39	มะนาว/lime	juice	ClearPhlegm	39juCP
39	มะนาว/lime	juice	RelieveCough	39juRC
40	ขิง/Ginger	rootStock	ExcreateSweat	40rtsES
40	ขิง/Ginger	rootStock	ReduceFever	40rtsRF
40	ขิง/Ginger	rootStock	ReleaseGas	40rtsRG
40	ขิง/Ginger	rootSt ck	StopVomit	40rtsSV
...	...	...	...	...

ตารางที่ 3 แสดงเมตริกซ์ของ ความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่เรียงโดย Herbal-Medical-Property

Herb ID	HerbName.	HerbPlantPart Concept	Herbal-Medicinal- Property Concept	Relation ID SortedBy Herbal- Medical-Property
39	มะนาว/lime	juice	ClearPhlegm	CP39ju
39	มะนาว/lime	seed	ClearPhlegm	CP39se
38	สะเดา/Neem	leaf	ClearPhlegm	CP38le
37	โหระพา/Basil	whole	ExcreateSweat	ES37wh
40	ขิง/Ginger	rootStock	ExcreateSweat	ES40rts
38	สะเดา/Neem	leaf	HaveAppetite	HA38le
37	โหระพา/Basil	whole	ReduceFever	RF37wh
40	ขิง/Ginger	rootStock	ReduceFever	RF40rts
39	มะนาว/lime	root	ReduceFever	RF39rt
39	มะนาว/lime	seed	ReduceFever	RF39se
38	สะเดา/Neem	bark	ReduceFever	RF38ba
37	โหระพา/Basil	whole	ReduceInflame	RI37wh
37	โหระพา/Basil	whole	ReleaseGas	RG37wh
40	ขิง/Ginger	rootStock	ReleaseGas	RG40rts
38	สะเดา/Neem	leaf	ReleaseGas	RG38le
37	โหระพา/Basil	whole	RelieveCough	RC37wh
39	มะนาว/lime	juice	RelieveCough	RC39ju
37	โหระพา/Basil	whole	RelieveNauseate	RN37wh
40	ขิง/Ginger	rootStock	StopVomit	SV40rts
...	...	...	...	...

ผลการทดลองและการประเมินผล

คลังข้อมูลภาษาไทยภายใต้โดเมนพืชสมุนไพรไทยจำนวน 1600EDUs (ซึ่งดาวน์โหลดจากเว็บพืชสมุนไพรไทย สามารถแบ่งออกเป็น 2 ส่วนคือ 1000EDUs สำหรับประเมินการสกัด N-Word-Co ที่มีแนวความคิดสรรพคุณทางยาจากเอกสารภาษาไทย ด้วยค่าความแม่นยำ (Precision) และค่าระลึก (Recall) และ 600EDUs สำหรับประเมินการสกัดความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารภาษาไทยนั้นด้วยค่าความแม่นยำ และค่าระลึก โดยการประเมินทั้งสองดังกล่าวใช้ผู้เชี่ยวชาญ 3 ท่านตัดสินด้วยวิธีการลงคะแนนเสียงข้างมาก (Max Win Voting) ดังนี้

1) การสกัด N-Word-Co ที่มีแนวความคิดสรรพคุณทางยา

ค่าความแม่นยำ และค่าระลึก ของการสกัด N-Word-Co ที่มีแนวความคิดสรรพคุณทางยาจากเอกสารภาษาไทยคือ 88.9% และ 79.4% ตามลำดับ สาเหตุที่มีค่าระลึกต่ำ เพราะปัญหา Anaphora ดังตัวอย่างต่อไปนี้

EDU1: “ใบโหระพา/*Basil leaves* บรรเทา/*relieve* อาการ/*symptom* ดังกล่าวข้างต้น/*as mentioned above*”
 (“*Basil leaves relieve the symptom as mentioned above.*”)

ซึ่งมีกลุ่มคำต่อไปนี้ ‘ดังกล่าวก่อน/as mentioned above’ เป็น Anaphora

2) การสกัด ความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากคลังข้อมูล โดยการใช้ N-Word-Co List ที่ถูกต้อง มาทำการสกัด ความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากคลังข้อมูลสำหรับทดสอบนั้น ได้ ด้วยค่าความแม่นยำ และค่าระลึก ของการสกัด ความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารภาษาไทยคือ 90% และ 77.5% ตามลำดับ สาเหตุที่มี ค่าระลึก ต่ำเพราะปัญหา ต่อไปนี้

a) ปัญหา NP1 ซึ่งประกอบด้วยประโยคย่อย.

“ใบชุมเห็ดเทศ/*Alata leaf* ซึ่งมีฤทธิ์เป็นยาระบายอ่อนๆ/*which has an effect as a soft laxative* สามารถนำไปต้ม/*can be boiled* น้ำ/*water* กินได้/*eatable*”
 (“*An Alata leaf which has an effect as a soft laxative can be boiled with water for drinking.*”)

ซึ่งมี NP1 คือ ‘ใบชุมเห็ดเทศ/*Alata leaf* ซึ่งมีฤทธิ์เป็นยาระบายอ่อนๆ/*which has an effect as a soft laxative*’ มีประโยคย่อยคือ ‘ซึ่งมีฤทธิ์เป็นยาระบายอ่อนๆ/*which has an effect as a soft laxative*’

b) ปัญหา NP2 ซึ่งมีหลาย Noun2 ปรากฏด้วยแนวความคิดหลักเดียวกันแต่แตกต่างกันรายละเอียด ดังนี้

EDU1: “(ใบเสลดพังพอน/Hophead leaf)/NP1 (ถอน/counteract (พิษไข้/fever-poison พิษแมลง/insect-poison และพิษงู/cobra-poison)/NP2)/VP”

(“A Hophead leaf counteracts in a fever-poison, an insect poison, and a cobra-poison.”)

ซึ่ง NP2 มีหลาย Noun2 ที่มีแนวความคิดหลักคือ ‘พิษ/poison’ แต่ไม่เหมือนกันในรายละเอียด เช่น พิษไข้/fever-poison พิษแมลง/insect-poison พิษงู/cobra-poison

ดังนั้นการสกัดความรู้เชิงความสัมพันธ์จากเอกสารสำหรับงานวิจัยนี้มีวิธีการหลักบนพื้นฐานของเหตุการณ์ที่แสดงออกด้วยกิริยาวิเศษณ์ที่สามารถประยุกต์ใช้ได้กับภาษาอื่น นอกจากนี้ความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ สามารถถูกจัดเก็บรวบรวมในรูปแบบ Matrix เพื่อง่ายต่อการเข้าถึงข้อมูลสำหรับการสร้างตัวแทนองค์ความรู้เชิงความสัมพันธ์ด้วยเครือข่ายวงแหวน เหมือนซึ่งง่ายสำหรับการทำความเข้าใจในการแก้ไขปัญหาสุขภาพ

สรุป

งานวิจัยนี้มุ่งเน้นการสกัดและเก็บรวบรวม ความสัมพันธ์ ‘พืช/ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ (Herb-MedicinalProperty Relation) จากเอกสารบนเว็บต่างๆที่เผยแพร่องค์ความรู้เกี่ยวกับพืชสมุนไพรไทย และควรเป็นเว็บของหน่วยงานหรือองค์กรที่เป็นที่รู้จักของคนทั่วไป เช่น เว็บไซต์ของโครงการอนุรักษ์พันธุกรรมพืชอันเนื่องมาจากพระราชดำริฯ (http://www.rspg.or.th/plants_data/herbs/herbs_200.htm)และสารานุกรมไทย(<http://kanchanapisek.or.th/kp6/sub/book/book.php?book=14&chap=10&page=chap10.htm>) เนื่องจากพืชสมุนไพรไทยมีอยู่มากมายหลายชนิดซึ่งมีคุณสมบัติและสรรพคุณแตกต่างกันไปตามสายพันธุ์ และแม้แต่สายพันธุ์เดียวกันก็ยังคงแตกต่างกันตามส่วนประกอบของพืช เช่น ราก ใบ ดอก ผล เป็นต้น ดังเช่นตามผลงานวิจัยนี้ที่ได้สกัดและเก็บรวบรวมความสัมพันธ์ ‘พืช/ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ไว้ในลักษณะตารางที่สามารถแสดงหรือแทนความสัมพันธ์ดังกล่าวออกมาในรูปแบบความสัมพันธ์เชิงเน็ตเวิร์ค ซึ่งทำให้สามารถเข้าใจได้ง่ายขึ้นและง่ายต่อการเข้าถึงระบบฐานข้อมูลสำหรับระบบตอบถามอัตโนมัติ ตัวอย่างเช่น คำถาม:ขอรายชื่อสมุนไพรและส่วนของสมุนไพร แก้วบรรเทาอาการไอ คำถาม:ใบโหระพามีสรรพคุณทางยาอะไรบ้าง เป็นต้น ฉะนั้นความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ ที่สกัดได้และเก็บรวบรวมจากงานวิจัยนี้สามารถนำไปพัฒนาต่อให้เข้ากับระบบตอบถามอัตโนมัติ สำหรับศูนย์ให้บริการความรู้ (Knowledge Service-Center Portal) แก่ชุมชน อย่างไรก็ตามงานวิจัยที่เกี่ยวกับภาษาไทยยังมีข้อจำกัดมากมาย เช่น การไม่มีตัวคัน การละคำ การละวลี เป็นต้น ทำให้เป็นอุปสรรคต่อพัฒนาแอปพลิเคชันต่างๆบนเว็บเช่น ระบบสอบถามองค์ความรู้อัตโนมัติ หรือ ระบบตอบถามอัตโนมัติ เป็นต้น ดังนั้นจึงเป็นเหตุผลหนึ่งที่งานวิจัยนี้ใช้กลุ่มของแก่นคำที่รวมเป็น N-Word-Co และมีแนวความคิดสรรพคุณทางยาทำหน้าที่แทนประโยคที่มีแนวความคิดสรรพคุณทางยา ฉะนั้นเมื่อประโยคคำถามมี N-Word-Co ที่กล่าวมาข้างต้นอยู่ด้วย ก็จะทำให้สามารถวิเคราะห์คำถามได้เร็วขึ้น อย่างไรก็ตามงานวิจัยนี้สามารถสกัด ได้ด้วยค่าความแม่นยำ และค่าระลึก 88.9% และ79.4% ตามลำดับจากเอกสารภาษาไทย และใช้ N-Word-Co ที่สกัดได้ถูกต้องสกัดความรู้เชิงความสัมพันธ์ ‘ส่วนของพืชสมุนไพร-สรรพคุณทางยา’ จากเอกสารภาษาไทยได้ค่าความแม่นยำ และค่าระลึก 90% และ77.5% ตามลำดับ ทั้งนี้ที่มีค่าระลึกต่ำเพราะปัญหาประโยคความซ้อน และปัญหานามวลี และบางครั้งรวมไปถึงการละนามวลี เป็นต้น งานวิจัยนี้นอกจากสามารถนำไปประยุกต์กับระบบถาม-ตอบอัตโนมัติเกี่ยวกับความรู้สรรพคุณทางยาของพืชสมุนไพรไทย แล้วยังสามารถเชื่อมต่อกับระบบการวินิจฉัยและการแก้ปัญหาอัตโนมัติ (Automatic Diagnosis and Problem-Solving System) เพื่อให้ความรู้และสนับสนุนการตัดสินใจในการแก้ไขปัญหาเบื้องต้นให้กับผู้ใช้

บรรณานุกรม

- Berger A.L., Della Pietra S.A., and Della Pietra V.J.(1996). A Maximum Entropy approach to natural language processing. *In Computer Linguistics*, vol.22,no.1, p.39–71.
- Biggins S., Mohammed S., and Oakley S. (2012). University Of Sheffield:Two Approaches to Semantic Text Similarity. In: Proc. of First Joint Conference on Lexical and Computational Semantics, Montréal, Canada, p. 655–661.
- Carlson L., Marcu D., Okurowski M. E.(2003). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. *In Current Directions in Discourse and Dialogue*, p.85-112.
- Chanlekha H. and Kawtrakul A.(2004). Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. *IJCNLP' 2004 proceedings*, p.1-7.
- Chareonsuk J., Sukvakree T., and Kawtrakul A.(2005). Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information. *NCSEC 2005 proceedings* , p.85-90.
- 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. *BioMed Central Complementary and Alternative Medicine*.,8:58.
- Haruechaiyasak C., Pailai J., Viratyosin W., and Kongkachandra R.(2011). ThaiHerbMiner: A Thai Herbal Medicine Mining and Visualizing Tool. In Proc. Of the 2011Workshop on Biomedical Natural Language Processing, Portland,Oregon,USA,June23-24,,ACL-HLT2011,p.186–187.
- Paşca M. (2008).Turning Web Text and Search Queries into Factual Knowledge: Hierarchical Class Attribute Extraction. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, p.1225-1230.
- Recupero D.R., Nuzzolese A.G., Consoli S., Presutti V., Peroni S., and Mongiovì M. (2015). Extracting knowledge from text using SHELDON, a Semantic Holistic framEwork for Linked ONtology data. The International World Wide Web Conference Committee (IW3C2), Florence, Italy, p.235–238.
- Riaz M. and Girju R.(2014). Recognizing Causality in Verb- Noun Pairs via Noun and Verb Semantics. In Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language BioNLP 2014, USA , p.48-57.
- Sudprasert S. and Kawtrakul A. (2003).. Thai Word segmentation based on Global and Local Unsupervised Learning. *NCSEC 2003 proceeding*, p.1-8,