งานวิจัยที่เกี่ยวข้อง

การวิจัยการสกัดความรู้เกี่ยวกับสรรพคุณทางยาของพืชสมุนไพรไทยจากเอกสารภาษาไทยเพื่อสนับสนุนระบบการตอบคำถามอัตโนมัติประกอบด้วยความรู้พื้นฐาน และงานวิจัยก่อนหน้าดังนี้

ความรู้พื้นฐาน

1) Naïve Bayes Classifier (Mitchell 1997)

ตัวจัดประเภทเนอีฟเบย์ (Naïve 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

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

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

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

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

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

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

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) โดยกล่าวว่าเซนเทอร์ริงเป็นโมเดลหรือแบบจำลองของความซับซ้อนของการอนุมานที่ต้องการผสมผสานระหว่างความหมายของถ้อยแถลงให้เป็นความหมายของบทความก่อนหน้า 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 U_n 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

ได้มีงานวิจัยมากมายที่ได้เสนอเทคนิคต่างๆเพื่อที่จะให้ได้มาซึ่งการสกัดความรู้สรรพคุณทางยาของสมุนไพร (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) 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) 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)

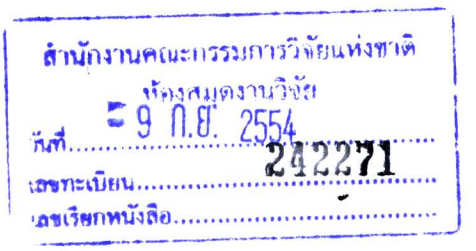
Riloff E. and Thelen M.(2000) ได้ใช้กฎเป็นพื้นฐาน (Rule Base)ต่างๆพร้อมกับการให้คะแนน สำหรับระบบตอบคำถามอย่างอัตโนมัติ กับคำถาม (Question Word) คือ "Who" "What" "When" "Where" และ "Why" หลังจากผ่านซอฟท์แวร์แจงประโยค ทั้งนี้เพื่อทดสอบความเข้าใจจากการอ่านบทความภาษาอังกฤษ โดยระบบอัตโนมัติให้คะแนนสำหรับประโยคที่มีคำตรงกับคำในประโยคคำถาม 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 ต่างๆบนเว็บไซด์ภาษาจีนจำนวน 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

ปัญหาจากระบบการตอบคำถามเกี่ยวกับคุณสมบัติของเอนติตี้สมุนไพรไทย

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: "ใบโหระพามีสรรพคุณทางยาอะไรบ้าง"

คำถาม2: "พืชสมุนไพรอะไรมีสรรพคุณขับลม"

คำถาม3: "สมุนไพรคืออะไร"

คำถาม4: "อะไรคือสาเหตุของโรค"

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"