ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์	วิศวกรรมคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง การสกัดและเข้าถึงความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารภาษาไทย

Object-Property Knowledge Extraction and Accessing from Thai Texts

นามผู้วิจัย นายอรรถพล คงหวาน

ได้พิจารณาเห็นชอบโดย

ประธานกรรมการ .....
(รองศาสตราจารย์อัศนีย์ ก่อตระกูล, D.Eng.)

กรรมการ .....
(อาจารย์ยอดเยี่ยม ทิพย์สุวรรณ, Ph.D.)

กรรมการ .....
(อาจารย์จิตร์ทัศน์ ฝักเจริญผล, Ph.D.)

หัวหน้าภาควิชา .....
(อาจารย์พีรวัฒน์ วัฒนพงศ์, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....
(รองศาสตราจารย์วินัย อาจดงหาญ, M.A.)

คณบดีบัณฑิตวิทยาลัย

วันที่ 5 **เดือน** เมษายน **พ.ศ.** 2545

วิทยานิพนธ์

เรื่อง

การสกัดและเข้าถึงความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารภาษาไทย

Object-Property Knowledge Extraction and Accessing from Thai Texts

โดย

นายอรรถพล คงหวาน

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2549

ISBN 974-16-1420-9

อรรถพล คงหวาน 2549: การสกัดและเข้าถึงความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารภาษาไทย ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ ประธานกรรมการที่ปรึกษา: รองศาสตราจารย์อัศนีย์ ก่อตระกูล, D.Eng. 82 หน้า
ISBN 974-16-1420-9

ปัจจุบันเอกสารเป็นแหล่งเก็บความรู้ที่สำคัญของมนุษย์ ซึ่งความรู้ต่าง ๆ ได้กระจายอยู่ในเอกสารและมีการเติบโตอย่างรวดเร็ว แต่ผู้ใช้จะต้องเสียเวลาในการอ่านเอกสารจำนวนมากเพื่อค้นหาความรู้ และนำความรู้นั้นไปใช้ประโยชน์ ในงานวิจัยนี้จึงมีวัตถุประสงค์ในการพัฒนาเทคนิคสำหรับสกัดความรู้และเข้าถึงความรู้จากเอกสารภาษาไทย การสกัดความรู้หมายถึงการสกัดเฉพาะใจความสำคัญในขอบเขตที่ผู้ใช้สนใจ โดยในงานวิจัยนี้จะสนใจเฉพาะความรู้ที่เกี่ยวกับคุณสมบัติของอ็อบเจกต์ เนื่องจากคุณสมบัติของอ็อบเจกต์สามารถใช้ระบุถึงอ็อบเจกต์ที่สนใจได้ ตัวอย่างเช่น ในโดเมนการเกษตร การสืบค้นชื่อของแมลงศัตรูพืชสามารถสืบค้นได้จากคุณสมบัติของแมลง เป็นต้น ปัญหาการสกัดความรู้จากเอกสารภาษาไทยนั้นจะมีปัญหาที่เกี่ยวกับการประมวลผลภาษา ได้แก่ การใช้สิ่งอ้างอิงร่วมแบบสูญรูปและการละคำ ซึ่งทำให้องค์ประกอบของความรู้ที่ปรากฏในรูปประโยคมีความไม่สมบูรณ์ และในการเข้าถึงความรู้นั้นจะมีปัญหาเกี่ยวกับความไม่ชัดเจนของค่าคุณสมบัติในการสืบค้นอ็อบเจกต์ เนื่องจากค่าคุณสมบัติอาจจะอยู่ในรูปของค่าที่เป็นนามบัญญัติและส่วนขยาย เช่น “เขียวเข้ม”, “แดงอ่อน” เป็นต้น วิทยานิพนธ์นี้จึงเสนอเทคนิคสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์รวมทั้งการเข้าถึงความรู้ด้วยคอมพิวเตอร์ โดยในส่วนสกัดความรู้ การประมวลผลภาษาระดับบทความสำหรับการประมวลผลสิ่งอ้างอิงร่วมแบบสูญรูปและการประมวลผลการละคำ เป็นสิ่งจำเป็นเพื่อทำรูปประโยคให้สมบูรณ์และใช้เทคนิคการเปรียบเทียบแม่แบบเพื่อทำการสกัดความรู้ และได้มีการใช้ทฤษฎีฟัซซี่และการวัดความคล้ายคุณสมบัติแบบฟัซซี่ ในส่วนของการเข้าถึงความรู้

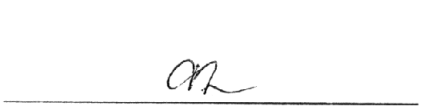
การวัดผลของงานวิจัยนี้ประกอบด้วยการวัดผลใน 4 ส่วน ได้แก่ อัลกอริทึมประมวลผลการใช้สิ่งอ้างอิงร่วมแบบสูญรูป อัลกอริทึมประมวลผลการละคำ การสกัดความรู้ และ การสืบค้นอ็อบเจกต์ โดยเน้นโดเมนการเกษตรซึ่งมีขนาดคลังเอกสารประมาณ 2,000 ประโยค โดยผลค่าความถูกต้องของการประมวลผลสิ่งอ้างอิงร่วมแบบสูญรูปอยู่ที่ 82.14% ค่าความระลึก 71.42% และค่า F-measure 76.40% ผลค่าความถูกต้องของการประมวลผลการละคำอยู่ที่ 96.96% ค่าความระลึก 96.96% และค่า F-measure 96.96% ผลค่าความถูกต้องของเทคนิคสกัดความรู้อยู่ที่ 88.88% ค่าความระลึก 47.50% และค่า F-measure 61.53% และผลค่าความถูกต้องของเทคนิคการสืบค้นอ็อบเจกต์อยู่ที่ 81.81% ค่าความระลึก 90.90% และค่า F-measure 86.11%

ลายมือชื่อนิสิต ลายมือชื่อประธานกรรมการ 29 / 03 / 49

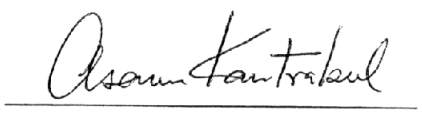
Authapon Kongwan 2006: Object-Property Knowledge Extraction and Accessing from Thai Texts. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Asanee Kawtrakul, D.Eng. 82 pages. ISBN 974-16-1420-9

Nowadays, documents are the important storage of human knowledge. The knowledge is spread through the document and also grows rapidly. However, the users have to take a lot of time to read the document to find the knowledge and utilize it. This thesis has the objective to develop the technique to extract the knowledge from Thai texts and also to access the knowledge. The knowledge extraction is the tool to assist the user extract the salient of the document in the interested domain. This research attended to the knowledge about the properties of the object. The properties of the object are able to identify the interested object such as to search the pest by given the pest properties. To extract the knowledge from Thai texts, there are problems involved with the linguistic phenomena such as the zero anaphora and textual ellipsis that make the components of knowledge incomplete in the sentence. To access the knowledge, there are the problems about the obscurity of the property value such as "dark green", "light red". Accordingly, this research proposed the technique for extracting the knowledge about the object properties from Thai texts and the technique for accessing the knowledge by computer. For extracting the knowledge, the discourse processing for solving the zero anaphora and textual ellipsis are needed in order to make the sentence complete and template matching is used for extracting the knowledge from the sentence. For accessing knowledge about the property of the object, the fuzzy theory and fuzzy property similarity measurement are utilized.

Four major components of the system, which are the zero anaphora resolution, the textual ellipsis resolution, the knowledge extraction module, and the query module, were evaluated with 2,000 sentences from corpus in agricultural domain. In the zero anaphora resolution, the precision, the recall and F-measurement are 82.14%, 71.42% and 76.40% respectively. In the textual ellipsis resolution, those are 96.96%, 96.96% and 96.96%. In the knowledge extraction module, those are 88.88%, 47.50% and 61.53%. And finally, in the query module, those are 81.81%, 90.90% and 86.11%.



Student's signature



Thesis Advisor's signature

29 / 03 / 06

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลงได้ด้วยความช่วยเหลือจากบุคคลหลายท่าน ข้าพเจ้าขอขอบพระคุณรองศาสตราจารย์อัศนีย์ ก่อตระกูล ประธานกรรมการที่ปรึกษา ผู้ที่ให้โอกาส และผลักดันข้าพเจ้าจนสำเร็จการศึกษา พร้อมทั้งให้แนวทางในการทำวิจัย และข้อเสนอแนะที่เป็นประโยชน์ต่อการทำวิทยานิพนธ์ฉบับนี้ และขอขอบคุณ นิสิตปริญญาเอก คุณธนา สุขวารี, คุณตระกูล เพิ่มพูน, คุณฉวีวรรณ เพ็ชรศิริ, คุณเฉลิมพล ศิริกายน และคุณอรวรรณ อิ่มสมบัติ ที่กรุณาให้คำปรึกษา และข้อเสนอแนะที่มีคุณค่า เพื่อให้วิทยานิพนธ์ฉบับนี้สมบูรณ์ยิ่งขึ้น

ข้าพเจ้าขอขอบคุณ คุณมุกข์ดา สุขธาราจาร และคุณพัชรี วราศรัย ที่ให้คำปรึกษาด้านภาษาศาสตร์ และเตรียมข้อมูลเพื่อใช้ในงานวิจัยนี้ ขอขอบคุณ คุณอารีรัตน์ ทองใบ ที่คอยช่วยเหลือติดต่อประสานงานธุระต่าง ๆ ขอขอบคุณ เพื่อน ๆ พี่ ๆ และน้อง ๆ ห้องปฏิบัติการวิจัยเชี่ยวชาญเฉพาะการประมวลผลภาษาธรรมชาติและเทคโนโลยีสารสนเทศอัจฉริยะ (NAiST Laboratory) ทุกท่านที่ให้ความช่วยเหลือในการทำงานวิจัย ขอขอบคุณห้องปฏิบัติการ NAiST ที่สนับสนุนทรัพยากร และสิ่งอำนวยความสะดวกต่าง ๆ ในการทำงานวิจัย และขอขอบคุณเจ้าหน้าที่ธุรการโครงการปริญญาโท และเจ้าหน้าที่ธุรการภาคภาควิชาวิศวกรรมคอมพิวเตอร์มหาวิทยาลัยเกษตรศาสตร์ทุกท่าน สำหรับความช่วยเหลือในการประสานงาน และงานด้านเอกสารต่าง ๆ

คุณงามความดี หรือประโยชน์อันใดที่เกิดจากวิทยานิพนธ์ฉบับนี้ ขออุทิศให้แก่ บิดา มารดา บุพการี และผู้มีพระคุณทุกท่าน

อรรถพล คงหวาน

มีนาคม 2549

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

สารบัญภาพ (ต่อ)

การสกัดและเข้าถึงความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารภาษาไทย

Object-Property Knowledge Extraction and Accessing from Thai Texts

คำนำ

ในการประยุกต์ใช้งานหรือการพัฒนาระบบที่สามารถประมวลผลและอนุมานความรู้เพื่อหาคำตอบต่าง ๆ ให้แก่ผู้ใช้ เช่น ระบบผู้เชี่ยวชาญ (Expert System), ระบบถาม-ตอบ (Question-Answering System) เป็นต้น เป็นระบบที่ต้องการฐานความรู้เป็นองค์ประกอบสำคัญ ซึ่งระบบเหล่านี้จะเรียกว่า ระบบอิงฐานความรู้ (Knowledge Based System) ซึ่งในอดีตฐานความรู้ของระบบอิงฐานความรู้นั้นจะถูกสร้างด้วยคน ซึ่งมีข้อจำกัดในการเพิ่มความรู้เข้าไปยังฐานความรู้ เนื่องจากความรู้ต่าง ๆ ที่จะนำไปจัดเก็บในฐานความรู้นั้นจะต้องถูกแปลงให้อยู่ในรูปแบบตัวแทนความรู้ (Knowledge Representation) ที่สามารถประมวลผลด้วยคอมพิวเตอร์เสียก่อน จึงทำให้การเพิ่มความรู้ให้กับระบบฐานความรู้เป็นไปได้ช้า แหล่งความรู้ที่สำคัญที่จะนำมาใช้สร้างฐานความรู้นั้นนอกจากการสร้างด้วยคนโดยตรงแล้ว เอกสารก็สามารถใช้เป็นแหล่งความรู้ให้กับระบบอิงฐานความรู้ได้เช่นกัน ซึ่งการใช้เอกสารเป็นแหล่งความรู้ในการสร้างฐานความรู้มีข้อดีคือ เอกสารเป็นแหล่งความรู้ของมนุษย์ที่มีการเติบโตอย่างรวดเร็ว และมีอยู่เป็นจำนวนมากทั้งในห้องสมุดและอินเตอร์เน็ต (Internet) ดังนั้นจึงมีงานวิจัยและพัฒนาระบบสกัดความรู้ที่สนใจจากเอกสารและแปลงความรู้เหล่านั้นให้อยู่ในรูปแบบตัวแทนความรู้เพื่อจัดเก็บในฐานความรู้เพื่อนำไปประยุกต์ใช้ต่อไป

ระบบถาม-ตอบคือ ระบบอิงฐานความรู้ที่ผู้ใช้สามารถส่งชุดคำถามต่าง ๆ ให้กับระบบแล้วระบบจะหาคำตอบที่เหมาะสมให้กับผู้ใช้ ซึ่งระบบถาม-ตอบนั้นถือว่าเป็นระบบสืบค้นข้อมูล (Information Retrieval) ชนิดหนึ่ง ระบบสืบค้นข้อมูลที่เป็นที่รู้จักกันมากที่สุดคือ ระบบสืบค้นอินเตอร์เน็ต (Internet Search Engine) เช่น Yahoo.com, Google.com เป็นต้น โดยระบบถาม-ตอบจะมีความแตกต่างกับระบบสืบค้นอินเตอร์เน็ตคือ ระบบสืบค้นอินเตอร์เน็ตจะเป็นระบบที่ค้นหาเอกสารหรือเว็บเพจ (Web Page) ที่ตรงกับชุดคำถาม (Query) หรือ คำสำคัญ (Key Word) มาให้ผู้ใช้ ซึ่งผู้ใช้จะต้องอ่านเอกสารเพื่อหาคำตอบหรือ ข้อมูลที่ต้องการด้วยตัวเอง แต่ระบบถาม-ตอบ จะอ่านเอกสารและสกัดเอาคำตอบมาให้ตรงกับความต้องการของผู้ใช้มากที่สุด ในระบบถาม-ตอบนั้นอาจจะมีการตอบสนองต่อชนิดของคำถามที่ไม่เหมือนกัน เนื่องจากชนิดของคำถามในแต่ละชนิด

นั้นจะมีวิธีการสกัดความรู้และการประมวลผลเพื่อหาคำตอบที่ไม่เหมือนกัน โดยชนิดของคำถามเมื่อแบ่งตามลักษณะของคำตอบสามารถแบ่งได้ดังนี้

(ก) รู้ว่าอะไร (Know-What)
(ข) รู้ว่าทำไม (Know-Why)
(ค) รู้ว่าอย่างไร (Know-How)

จากข้างต้นจะเห็นว่าลักษณะของคำตอบที่แตกต่างคือ “รู้ว่าอะไร” คำตอบจะเป็นอ็อบเจกต์ (Object) และสิ่งที่อธิบายอ็อบเจกต์ (Object Description) นั้น ๆ ส่วน “รู้ว่าทำไม” คำตอบจะเป็นเหตุผล (Reason) หรือเงื่อนไข (Condition) ของเหตุการณ์ (Event) หรือสภาพ (State) ต่าง ๆ และ “รู้ว่าอย่างไร” คำตอบจะเป็นขั้นตอน (Procedure) หรือวิธีการ (Method) ของสิ่งใดสิ่งหนึ่ง สำหรับงานวิจัยนี้สนใจความรู้เกี่ยวกับการสกัดคุณสมบัติ (Property) ของอ็อบเจกต์ ซึ่งอยู่ในกลุ่มของ “รู้ว่าอะไร” โดยมีสมมุติฐานที่ว่า “คุณสมบัติของอ็อบเจกต์” เป็นสิ่งที่สามารถใช้ระบุและอธิบายถึงตัวอ็อบเจกต์ได้ เป็นผลให้เราสามารถสืบค้นชื่อของอ็อบเจกต์ต่าง ๆ ด้วยการป้อนคุณสมบัติต่าง ๆ เช่น สี ขนาด เป็นต้น ในโดเมนการเกษตร ระบบนี้จะสามารถประยุกต์ใช้เพื่อสืบค้นศัตรูพืช โดยผู้ใช้อาจจะสังเกตตั้งคำถามด้วย สีและขนาดของแมลงศัตรูพืชที่พบเห็น และผลลัพธ์ที่ได้คือชื่อของแมลงศัตรูพืช ซึ่งนำไปสู่การค้นหาวิธีการป้องกันและกำจัดแมลงศัตรูพืชเหล่านั้นต่อไปได้

ในการสกัดความรู้จากเอกสารภาษาไทยนั้น มักจะมีปัญหาที่เกี่ยวกับการประมวลผลภาษา ได้แก่ การใช้สิ่งอ้างอิงร่วมแบบสูญรูปและการละคำ ซึ่งทำให้การสกัดความรู้จากประโยคไม่สมบูรณ์ และในการเข้าถึงความรู้นั้นจะมีปัญหาเกี่ยวกับความไม่ชัดเจนของค่าคุณสมบัติในการสืบค้นอ็อบเจกต์ เนื่องจากค่าคุณสมบัติอาจจะอยู่ในรูปของค่าแบบสัญลักษณ์ที่เป็นนามบัญญัติ เช่น ค่าสี, ค่ากลิ่น เป็นต้น โดยวิทยานิพนธ์นี้ได้เสนอเทคนิคสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารภาษาไทยรวมทั้งการเข้าถึงความรู้ด้วยคอมพิวเตอร์ การสกัดความรู้จากเอกสารนั้น งานวิจัยนี้ได้เสนออัลกอริทึมเพื่อใช้แก้ปัญหาที่เกิดจากพฤติกรรมทางภาษา และเสนอเทคนิคการเทียบแม่แบบในการวิเคราะห์ประโยคเพื่อทำการสกัดองค์ประกอบของความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์ เช่น สี, น้ำหนักและขนาด ต่าง ๆ ของอ็อบเจกต์ที่อธิบายไว้ในเอกสารมาจัดเก็บในฐานความรู้ สำหรับการเข้าถึงความรู้นั้น ได้มีการนำทฤษฎีฟัซซี่มาใช้เป็นตัวแทนความรู้ของค่าคุณสมบัติ รวมทั้งเสนอวิธีการคำนวณความคล้ายทางคุณสมบัติของอ็อบเจกต์ เพื่อใช้ในการสืบค้นอ็อบเจกต์

วัตถุประสงค์และขอบเขต

วัตถุประสงค์

1. ศึกษาทฤษฎีประยุกต์ที่มีอยู่ ในส่วนทางด้านการประมวลผลภาษาธรรมชาติ และการประมวลผลความรู้ เพื่อพัฒนากระบวนการสกัดความรู้ในเอกสารภาษาไทย ได้อย่างเหมาะสมและมีประสิทธิภาพ
2. พัฒนาเทคนิคสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารภาษาไทยที่มีความสามารถสกัดความรู้จากเอกสารที่มีการใช้สิ่งอ้างอิงร่วมแบบสูญรูปและการละคำได้
3. พัฒนาเทคนิคสืบค้นอ็อบเจกต์ด้วยคุณสมบัติของอ็อบเจกต์ โดยสามารถรองรับข้อมูลค่าคุณสมบัติทั้งชนิดแบบตัวเลขและแบบสัญลักษณ์

ขอบเขตงานวิจัย

งานวิทยานิพนธ์นี้มุ่งเน้นที่การพัฒนาเทคนิคสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสาร โดยจะเน้นในเอกสารในโดเมนการเกษตร เทคนิคสกัดความรู้ดังกล่าวนี้ จะสกัดความรู้จากเอกสารด้วยเทคนิคการเทียบแม่แบบ (Template matching) โดยแม่แบบที่ใช้ในการสกัดและฐานความรู้สนับสนุนทั้งหมดจะมีการกำหนดไว้ล่วงหน้า (Predefined) โดยผู้พัฒนา และอ็อบเจกต์ที่จะทำการสกัดนั้นประกอบด้วย แมลงและ พืชต่าง ๆ โดย เอกสารที่ใช้ในการทดลองจะต้องมีรูปแบบการเขียนที่ดี (well-style written) ซึ่งข้อมูลนำเข้าคือเอกสารที่ผ่านการตัดคำและมีการแจกแจงประโยคแล้ว และผลลัพธ์ที่ได้คือความรู้ที่อยู่ในรูปของตรรกะภาคแสดง รวมทั้งพัฒนาเทคนิคการสืบค้นอ็อบเจกต์ด้วยคุณสมบัติของอ็อบเจกต์จากฐานความรู้ที่สร้างจากระบบสกัดความรู้ดังกล่าว ซึ่งข้อมูลนำเข้าจากผู้ใช้คือชุดของคุณสมบัติของอ็อบเจกต์ และผลลัพธ์ที่ได้คือค่าความคล้ายของอ็อบเจกต์เทียบกับค่าคุณสมบัติที่นำเข้า

ขั้นตอนการวิจัย

1. ศึกษาทฤษฎีประยุกต์ต่าง ๆ ที่ใช้ในงานสกัดความรู้จากเอกสารและสืบค้น รวมทั้งการวิเคราะห์ข้อดีข้อด้อยในงานก่อนหน้า เพื่อนำข้อมูลมาใช้ในงานวิจัย
2. รวบรวมคลังเอกสารในโดเมนการเกษตร เพื่อนำมาใช้ในการศึกษาพฤติกรรมทางภาษาที่จะเป็นปัญหาของการสกัดความรู้รวมทั้งการสืบค้น
3. ศึกษาพฤติกรรมทางภาษาที่เกิดขึ้นในคลังเอกสาร เพื่อรวมรวมปัญหาต่าง ๆ เพื่อนำมาใช้ในการพัฒนาเทคนิคและอัลกอริทึมในการสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารภาษาไทยและการสืบค้น
4. พัฒนาเทคนิคการสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารภาษาไทยและการสืบค้น
5. ทดสอบและวัดผลของเทคนิคสกัดการความรู้และการสืบค้นที่พัฒนาขึ้น
6. สรุปผลการวิจัยและประโยชน์ที่ได้รับ

การตรวจเอกสาร

ความรู้พื้นฐาน

1. การประมวลผลภาษาธรรมชาติ (Natural Language Processing)

การประมวลผลภาษาธรรมชาติเป็นส่วนสำคัญของงานวิจัยนี้ ซึ่งประกอบด้วย การตัดคำ และกำกับหน้าที่ของคำ การแจกแจงประโยคและการใช้สิ่งอ้างอิงร่วม ซึ่งมีรายละเอียดดังนี้

1.1 การตัดคำและกำกับหน้าที่ของคำ(Word Segmentation and Part-Of Speech Tagging)

การตัดคำเป็นขั้นตอนแรกที่จำเป็นสำหรับการพัฒนาโปรแกรมประยุกต์ด้านการประมวลผลภาษาธรรมชาติ เช่น การแปลเอกสาร การย่อความ การสืบค้นเอกสาร ซึ่งรวมทั้งการสกัดความรู้ด้วย ปัญหาสำคัญในการตัดคำในภาษาไทยคือการหาขอบเขตของคำได้อย่างถูกต้อง เนื่องจากภาษาไทยเป็นภาษาที่ไม่มีช่องว่างในการแบ่งคำ เช่นเดียวกับภาษาอังกฤษ อีกทั้งภาษาไทยไม่มีการใช้สัญลักษณ์พิเศษหรือชุดอักษรเพื่อบอกถึงศัพท์ที่ เป็นคำยืมจากภาษาต่างประเทศหรือชื่อเฉพาะ เช่น ภาษาญี่ปุ่นที่ใช้อักษรฮิราคานะในการเขียนคำยืมจากภาษาต่างประเทศ และไม่มีการใช้ตัวพิมพ์ใหญ่สำหรับเขียนชื่อเฉพาะ เหมือนใน ภาษาอังกฤษ (สุธี, 2547)

งานวิจัยด้านการตัดคำภาษาไทยที่มีมาก่อน สามารถแบ่งออกได้เป็น 2 แนวทางใหญ่ (ศักดิ์ชาติ, 2548)

วิธีอิงพจนานุกรม(ยืน, 2529) วิธีการนี้ได้รับความนิยมอย่างมากเพราะเป็นวิธีที่ง่าย รวดเร็ว แต่ความถูกต้องขึ้นอยู่กับขนาดของพจนานุกรมที่ใช้ ซึ่งในความเป็นจริงเป็นเรื่องยากที่จะเก็บคำศัพท์ทุกคำได้ครบถ้วน โดยเฉพาะศัพท์ที่เป็น ชื่อเฉพาะ หรือ คำยืมจากภาษาต่างประเทศ ซึ่งต่อไปนี้จะเรียกว่า คำไม่รู้จัก นอกจากนี้การใช้พจนานุกรมยังทำให้เกิดปัญหาความคลุมเครือในการแบ่งขอบเขตคำ เช่น “ตากลม” สามารถตัดได้เป็น “ตา-กลม” และ “ตาก-ลม” ในกรณีที่พจนานุกรมมีคำว่า “ตา” “กลม” “ตาก” และ “ลม”

วิธีอิงข้อมูลสถิติจากคลังประโยค(Kawtrakul *et al.*, 1995) เนื่องจากการตัดคำโดยการอิงพจนานุกรมไม่สามารถแก้ปัญหาคำไม่รู้จักและความคลุมเครือในการแบ่งขอบเขตคำได้อย่างมีประสิทธิภาพ จึงได้มีการนำสถิติจากคลังประโยคมาใช้เพื่อแก้ปัญหา

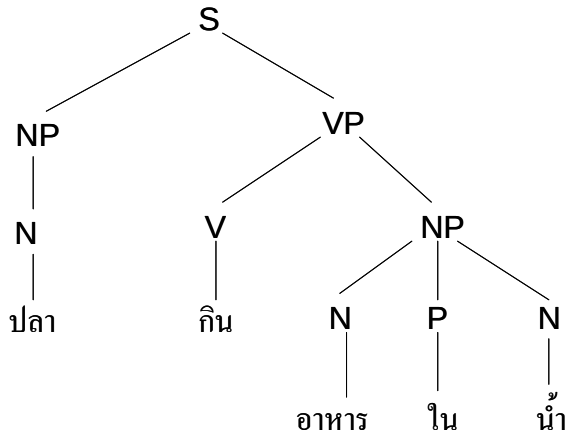
สำหรับการกำกับหน้าที่ของคำเป็นขั้นตอนหลังจากการตัดคำ ปัญหาในการกำกับหน้าที่ของคำได้แก่การที่คำมีหลายหน้าที่ตัวอย่างเช่นคำว่า “ขัน” สามารถทำหน้าที่เป็นคำนามที่หมายถึงขันน้ำ หรือทำหน้าที่เป็นคำกริยาที่หมายถึงอาการส่งเสียงร้องของไก่ โปรแกรมที่ทำหน้าที่ตัดคำจะต้องอาศัยคำแวดล้อมในการแก้ปัญหาความกำกวมนี้ ในงานวิจัยนี้ใช้โปรแกรมตัดคำของ (สุธี, 2547) ซึ่งมีการกำกับหน้าที่ของคำโดยอัตโนมัติ

1.2 การแจกแจงประโยค (Syntactic Parser)

การแจกแจงประโยคคือ กระบวนการวิเคราะห์วากยสัมพันธ์ (Syntax) ของประโยค เพื่อระบุโครงสร้างของความสัมพันธ์ของคำที่เรียงกันในประโยค ที่กำหนดจากไวยากรณ์ (Grammar) ของภาษานั้น ๆ ซึ่งไวยากรณ์ที่เกี่ยวกับวากยสัมพันธ์นั้น สามารถแบ่งออกได้เป็น 2 ชนิด ไวยากรณ์ส่วนประชิด และไวยากรณ์ดีเพนเดนซี (วีร์, 2548)

1.2.1 ไวยกรณ์ส่วนประชิด (Constituency Grammar)

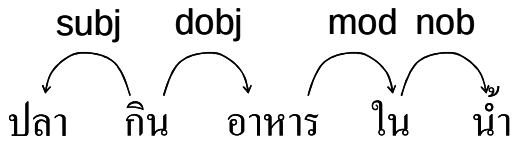
ไวยากรณ์ส่วนประชิดเป็นไวยากรณ์ที่อธิบายโครงสร้างของประโยคโดยการสร้างส่วนประชิด (Constituent) ขึ้นมานอกเหนือจากคำ โดยมีส่วนประชิดระดับบนสุดคือประโยค โดยประโยคจะประกอบไปด้วยวลีและอนุประโยค วลีและอนุประโยคจะประกอบไปด้วยวลีหรืออนุประโยคหรือคำ (Schneider, 1998) ตัวอย่างเช่น ประโยคประกอบด้วยนามวลีและกริยาวลี กริยาวลีประกอบคำกริยาและนามวลี โครงสร้างของส่วนประชิดสามารถแสดงได้ด้วยโครงสร้างต้นไม้ดังภาพที่ 1 ซึ่ง NP VP เป็นส่วนประชิดที่แทนนามวลีและกริยาวลี ตามลำดับ และ S เป็นส่วนประชิดที่แทนประโยค



<u>ภาพที่ 1</u> โครงสร้างต้นไม้ตามแบบไวยากรณ์ส่วนประชิด (วีร์, 2548)

1.2.2 ไวยากรณ์ดีเพนเดนซี (Dependency Grammar)

ไวยากรณ์ดีเพนเดนซีเป็นไวยากรณ์ที่แสดงความสัมพันธ์ของคำคำหนึ่งที่ขึ้นกับคำอื่นโดยไม่มีส่วนประกอบอื่น ๆ นอกเหนือจากคำ ดังในภาพที่ 2 คำว่า “ปลา” และ “อาหาร” ขึ้นอยู่กับคำว่า “กิน” คำว่า “ใน” ขึ้นอยู่กับคำว่า “อาหาร” และคำว่า “ น้ำ” ขึ้นอยู่กับคำว่า “ใน”



<u>ภาพที่ 2</u> โครงสร้างต้นไม้ตามแบบไวยากรณ์ดีเพนเดนซี (วีร์, 2548)

ไวยากรณ์ส่วนประชิดได้รับความนิยมและมีการใช้งานอย่างกว้างขวางทั้งยังได้ผลดีในการแจงประโยคภาษาอังกฤษซึ่งมีลำดับของคำค่อนข้างตายตัว เช่น ตัวแจงประโยคของคอลลินส์ (Collins, 1999) และตัวแจงประโยคของชาเนียคร์ (Charniak, 1997)

1.3 การใช้สิ่งอ้างอิงร่วม (Anaphora)

สิ่งอ้างอิงร่วม (Anaphora) คือ เครื่องมือทางภาษาชนิดหนึ่ง ซึ่งใช้เพื่อลดความความฟุ่มเฟือยในการอ้างอิงถึงนามวลีซึ่งเคยกล่าวมาก่อนแล้วในประโยคก่อนหน้า ซึ่งสิ่งอ้างอิงร่วมนั้นสามารถแบ่งได้เป็น 3 ชนิด คือ สิ่งอ้างอิงร่วมแบบสรรพนาม (Pronominal Anaphora) สิ่งอ้างอิงร่วมแบบคำนาม (Nominal Anaphora) และสิ่งอ้างอิงร่วมแบบสูญรูป (Zero Anaphora)

1.3.1 สิ่งอ้างอิงร่วมแบบสรรพนาม

การใช้สิ่งอ้างอิงร่วมแบบสรรพนาม คือ การใช้สรรพนามแทนนามวลีที่เคยกล่าวมาก่อนหน้าแล้วจากประโยคก่อนหน้า ดังตัวอย่าง

(1) สมชายไปดูหนังที่โรงภาพยนตร์
(2) จากนั้น **<u>เขา</u>**ก็ไปกินพิซซ่าที่ร้านพิซซ่า

ในประโยคที่ 2 จะเห็นว่าจะมีการใช้สรรพนาม “เขา” แทนคำว่า “สมชาย” ซึ่งเป็นนามวลีที่เคยกล่าวไว้ก่อนแล้วในประโยคที่ 1 นั่นเอง

1.3.2 สิ่งอ้างอิงร่วมแบบคำนาม

ในการใช้สิ่งอ้างอิงร่วมนั้น นอกจากมีการใช้สรรพนามแล้ว ยังสามารถใช้คำนามร่วมกับคำบ่งชี้ เพื่อใช้เป็นสิ่งอ้างอิงร่วมได้เช่นกัน ดังตัวอย่าง

(1) เพลี้ยเป็นแมลงศัตรูพืชที่สำคัญ
(2) **<u>แมลงชนิดนี้</u>**จะระบาดมากในช่วงหน้าร้อน

ในประโยคที่ 2 จะมีนามวลีคำว่า “แมลงชนิดนี้” ซึ่งเป็นสิ่งอ้างอิงร่วมที่อ้างถึง “เพลี้ย” ในประโยคแรกนั่นเอง

1.3.3 สิ่งอ้างอิงร่วมแบบสูญรูป

สิ่งอ้างอิงร่วมแบบสูญรูปคือ การละนามวลีออกจากประโยคในการอ้างอิงถึงนามวลีที่ได้เคยกล่าวมาแล้วในประโยคอื่นก่อนหน้า โดยที่ผู้อ่านสามารถเข้าใจได้อย่างถูกต้องอยู่ การใช้สิ่งอ้างอิงร่วมแบบสูญรูปมักจะใช้ในภาษาไทย

(1) เพลี้ยไฟเป็นแมลงขนาดเล็ก
(2) **<u>∅</u>** มีสีเหลืองหรือสีน้ำตาลอ่อน

ในประโยคที่ 2 นั้น ประธานจะถูกละไว้ที่ตำแหน่งเครื่องหมาย ∅ ซึ่งประธานที่ละไว้คือ นามวลี “เพลี้ยไฟ” นั่นเอง

2. การประมวลผลความรู้ (Knowledge Processing)

การประมวลผลความรู้เป็นกระบวนการสำคัญสำหรับระบบอิงความรู้ ปกติความรู้จะอยู่ที่ตัวมนุษย์ซึ่งเป็นความรู้ชนิดโดยนัย (Tacit Knowledge) การถ่ายทอดความรู้ที่อยู่ในตัวมนุษย์ไปยังระบบคอมพิวเตอร์นั้น ความรู้ชนิดโดยนัยจะต้องถูกแปลงให้เป็นความรู้ชนิดชัดเจน (Explicit Knowledge) ก่อน แล้วจึงทำการแปลงความรู้ชนิดชัดเจนให้อยู่ในรูปแบบตัวแทนความรู้ (Knowledge Representation) ที่คอมพิวเตอร์สามารถประมวลผลได้ ตัวแทนความรู้ที่สำคัญมี 3 ชนิด คือ ตรรกะภาคแสดง (Predicate Logic) กรอบ (Frame) และสคริปท์ (Script)

2.1 ตรรกะภาคแสดง (Predicate Logic)

ตรรกะภาคแสดงเป็นภาษาคณิตศาสตร์ชนิดหนึ่งที่ใช้แสดงตรรกะเพื่อการอนุมานความรู้ โดยมีส่วนประกอบคือ ภาคแสดง (Predicate) อาร์กิวเมนต์ (Argument) และตัวบ่งปริมาณ (Quantifier) ในตรรกะภาคแสดงจะมีตัวปฏิบัติการที่สำคัญ การเชื่อมแบบ “และ” (AND), การเชื่อมแบบ “หรือ” (OR) นอต (NOT) และ การส่อความ (IMPLICATION) ซึ่งความรู้ที่แสดงด้วยตรรกะภาคแสดงนั้นจะแบ่งได้เป็น 2 ชนิดคือ ข้อเท็จจริง (Fact) และกฎอนุมาน (Inference Rule) ในการแทนข้อเท็จจริงด้วยตรรกะภาคแสดง จะใช้ตรรกะที่ไม่มีตัวปฏิบัติ ดังตัวอย่างข้อเท็จจริง “Spot is a dog.”

dog(Spot)

จากตัวอย่างข้างบนภาคแสดงคือ “dog” ซึ่งแสดงถึงคอนเซ็ปต์สุนัข และมีอาร์กิวเมนต์เป็นค่าคงที่ (Constant) “Spot” ซึ่งคือชื่อของสุนัขนั่นเอง ซึ่งในกรณีที่แสดงข้อเท็จจริง อาร์กิวเมนต์จะเป็นค่าคงที่ (Constant) แต่ในกรณีที่ใช้แทนกฎอนุมานนั้น อาร์กิวเมนต์สามารถเป็นตัวแปรได้ด้วย ในการแทนกฎอนุมานด้วยตรรกะภาคแสดงจะใช้ตรรกะร่วมกับตัวปฏิบัติและตัวบ่งชี้ปริมาณ ดังตัวอย่างกฎอนุมานว่า “all dogs have tails”

$\forall x : dog(x) \rightarrow hastail(x)$

จากตัวอย่างข้างต้น ตรรกะ “dog” และ “hastail” มีอาร์กิวเมนต์เป็นตัวแปร x และมีการกำหนดตัวบ่งชี้ปริมาณ $\forall x$ ซึ่งหมายถึงทั้งหมด (For all) นั่นคือ สุนัขทุกตัวจะมีหาง นั่นเอง ซึ่งตัว

บ่งชี้ปริมาณในตรรกะภาคแสดงจะมี 2 ชนิด คือ ทั้งหมด ∀ (For all) และ บางส่วน ∃ (For some) ดังตัวอย่างกฎอนุมานว่า “ some boy can swim.”

$\exists x : boy(x) \rightarrow canswim(x)$

จากตัวอย่างข้างต้น ตัวแปร x จะมีปริมาณบ่งชี้เป็น ∃ ซึ่งหมายถึงเด็กผู้ชายบางคนไม่ใช่เด็กผู้ชายทุกคนที่ว่ายน้ำได้

2.2 กรอบ (Frame)

กรอบคือ กลุ่มของคุณสมบัติ (Attribute)หรือสล๊อต (Slot) และค่าที่เกี่ยวข้อง (Associated Value) ที่ใช้อธิบายสิ่งต่าง ๆ ในโลก (Minsky, 1975) ในระบบอิงความรู้จะมีการใช้กรอบในการเก็บความรู้กันอย่างแพร่หลาย ดังเช่นแทนความรู้เกี่ยวกับ “เมือง”

(City

<Province provincename>

<Country countryname>

<Population populationsize>

)

จากตัวอย่างข้างต้น เป็นกรอบที่อธิบายองค์ประกอบของกรอบ “เมือง” ซึ่งจะมีค่าคุณสมบัติ 3 ค่าหรือที่เรียกว่าสล๊อต (Slot) ได้แก่ จังหวัด (Province), ประเทศ (Country) และจำนวนประชากร (Population) ในกรอบ เพื่อใช้เก็บข้อมูลที่เกี่ยวข้องกับ “เมือง” จากนั้นเมื่อการเก็บข้อมูลจริงก็จะมีการสร้างตัวกรณีตัวอย่าง (Instance) ขึ้นมาสำหรับกรอบ ดังตัวอย่าง “เมืองหาดใหญ่”

(“หาดใหญ่”

<Province “สงขลา”>

<Country “ไทย”>

<Population 300K>

)

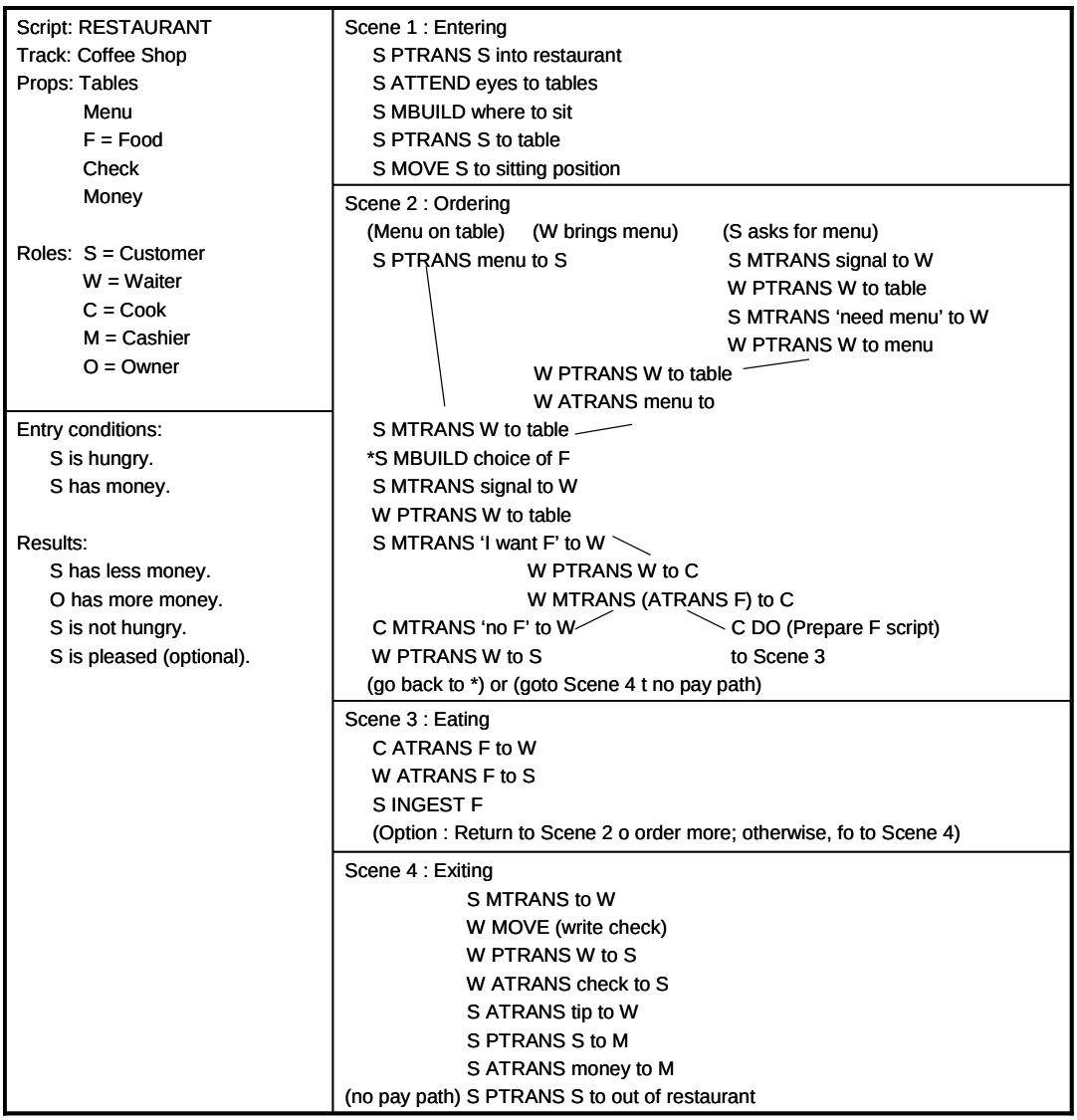
จากตัวอย่างข้างต้น เป็นตัวอย่างของกรอบกรณีตัวอย่างของ “เมืองหาดใหญ่” ซึ่งจะมีข้อมูลในสล๊อต Province เป็น “สงขลา” สล๊อต Country เป็น “ไทย” และสล๊อต Population เป็น 300K

2.3 สคริปท์ (Script)

สคริปท์คือ โครงสร้างที่ใช้อธิบายความต่อเนื่องกัน (Sequence) ของเหตุการณ์ (Event) ในบริบท (Context) สคริปท์จะประกอบด้วยเซตของสล๊อต (Set of slots) ที่เกี่ยวข้องกับสิ่งที่ใช้และสิ่งที่เกิดขึ้นในเหตุการณ์ โดยเซตของข้อมูลจะประกอบด้วย 6 ชุด (Schank and Abelson, 1977) ได้แก่

(ก) Entry conditions คือ เงื่อนไขที่จำเป็นต้องมี ก่อนจะเกิดเหตุการณ์ในสคริปท์
(ข) Result คือ เงื่อนไขที่จะเกิดขึ้นเมื่อเหตุการณ์ในสคริปท์ได้ผ่านไปแล้ว
(ค) Props คือ สล๊อตที่อธิบายถึงอ็อบเจกต์ที่เกี่ยวข้องกับเหตุการณ์
(ง) Roles คือ สล๊อตที่อธิบายถึงบทบาทของบุคคลที่เกี่ยวข้องกับเหตุการณ์
(จ) Track คือ ตำแหน่งหรือทิศทางที่เกิดขึ้นในเหตุการณ์
(ฉ) Scenes คือ ลำดับของเหตุการณ์ที่เกิดขึ้น

สคริปท์เป็นตัวแทนความรู้ที่สำคัญในการแสดงเหตุการณ์ที่เกิดขึ้นในโลกจริง (Real World) เพราะสคริปท์ได้แสดงถึงลำดับและความสัมพันธ์ต่าง ๆ ระหว่างบุคคลและเหตุการณ์ต่าง ๆ ในสคริปท์ โดยภาพที่ 3 ได้แสดงถึงตัวอย่างของสคริปท์ของร้านอาหาร ที่แสดงถึงลำดับเหตุการณ์การทานอาหารที่ร้านอาหาร โดยจะมีฉาก (Scene) ทั้งหมด 4 ฉาก คือ Entering, Ordering, Eating และ Exiting และมีบุคคลที่เกี่ยวข้องกับเหตุการณ์ 5 บุคคลคือ Customer, Waiter, Cook, Cashier และ Owner



ภาพที่ 3 สคริปท์ของร้านอาหาร (Schank and Abelson, 1977)

3. ทฤษฎีฟัซซี่ (Fuzzy Theory)

ทฤษฎีฟัซซี่คือทฤษฎีที่เกี่ยวกับการประมวลผลข้อมูลที่มีความไม่ชัดเจนอยู่ในค่าของข้อมูลเอง ในการพัฒนางานประยุกต์บางโดเมนนั้น ข้อมูลที่ทำการประมวลผลอาจจะอยู่ในรูปที่ไม่สามารถระบุได้อย่างชัดเจน หรือข้อมูลอาจจะอยู่ในรูปของคำหรือความหมายมากกว่าที่จะเป็นค่าตัวเลข ตัวอย่างเช่น การจำแนกบุคคลว่าเป็นคนสูงหรือคนเตี้ย ถ้าเราสร้างกฎว่าคนที่สูงมากกว่าหรือเท่ากับ 180 ซม. คือ คนสูงแล้ว แสดงว่าคนที่สูงน้อยกว่า 180 ซม. ต้องเป็นคนเตี้ย แต่สำหรับคนที่สูง 179.9 ซม. ถ้านับจากกฎแล้วอาจจะจำแนกว่าเป็นคนเตี้ย ทั้ง ๆ ที่คนที่สูง 179.9 ซม. กับ คนที่ 180 ซม. นั้นความสูงแทบจะไม่ต่างกันเลย ทั้งนี้เพราะว่า ค่าข้อมูลของคำว่า “คนสูง” กับคำว่า “คน

เตี้ย” นั้นไม่สามารถระบุได้อย่างชัดเจน ฉะนั้นทฤษฎีฟัซซี่จึงมีขึ้นเพื่ออธิบายวิธีการประมวลข้อมูลที่มีความไม่ชัดเจนเหล่านี้

3.1 ฟัซซี่เซต (Fuzzy Sets)

ในการอธิบายลักษณะของกลุ่มข้อมูลนั้น เราอาจจะใช้เซตในการอธิบายได้ เช่น เราจะอธิบายข้อมูลของคนที่มีอายุน้อยคือ คนที่อายุน้อยกว่า 20 ปี เราอาจจะอธิบายได้ดังนี้

$$young = \{x \in P \mid age(x) \le 20\}$$

นั่นคือ x ที่อยู่ในโดเมน P จะอยู่ในเซต young ก็ต่อเมื่อฟังก์ชั่น age(x) นั้นมีค่าน้อยกว่าหรือเท่ากับ 20 โดยฟังก์ชั่น age(x) คือ ค่าอายุของบุคคล x นั่นเอง ซึ่งเราสามารถอธิบายฟังก์ชั่นคุณสมบัติ (Characteristic Function) ได้ดังนี้

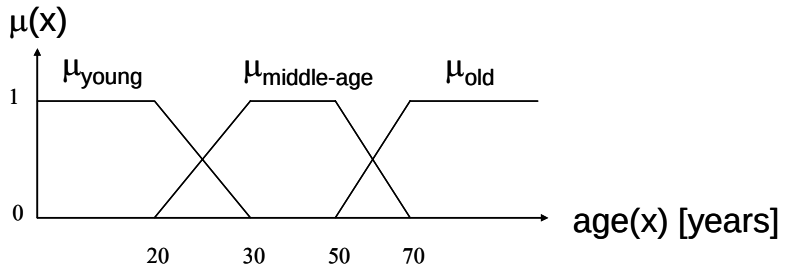
$$m_{young}(x) = \begin{cases} 1 : age(x) \le 20 \\ 0 : age(x) > 20 \end{cases}$$

ค่าของฟังก์ชั่นจะมีค่าเป็น 1 เมื่อ x นั้นอยู่ในเซตคนที่มีอายุน้อยและ ค่าของฟังก์ชั่นจะมีค่าเป็น 0 เมื่อ x นั้นไม่อยู่ในเซตคนที่มีอายุน้อย ซึ่งเราสามารถเรียกฟังก์ชั่นคุณสมบัตินี้ว่าฟังก์ชั่นความเป็นสมาชิก (Membership Function) ของเซตคนที่มีอายุน้อย

เมื่อเราพิจารณาที่ขอบเขตของเซตคนที่มีอายุน้อยจะเห็นว่าจะมีการแบ่งขอบเขตที่ชัดเจนเกินไปนั่นคือถ้าอายุมากกว่า 20 ปีแล้วจะถือว่าไม่เป็นสมาชิกของเซตซึ่งจะมีความไม่เหมาะสม ซึ่งในฟังก์ชั่นความเป็นสมาชิกของฟัซซี่เซตจะไม่จำเป็นต้องมีค่าสมาชิกเป็นเพียง 0 หรือ 1 เท่านั้น นั่นคือคนที่มีอายุ 21 ปี ก็ยังคงอยู่ในเซตของคนที่มีอายุน้อยเช่นกัน เพียงแต่ค่าความเป็นสมาชิกของเซตอาจจะอยู่ระหว่าง 0 ถึง 1 เช่น อาจจะเป็น 0.9 ก็ได้ เป็นต้น ฉะนั้นเราจึงสามารถสร้างฟังก์ชั่นความเป็นสมาชิกของฟัซซี่เซตคนที่มีอายุน้อยได้ดังนี้

$$\mu_{young}(x) = \begin{cases} 1 & : age(x) \le 20 \\ 1 - \dfrac{(age(x) - 20)}{10} & : 20 < age(x) \le 30 \\ 0 & : 30 < age(x) \end{cases}$$

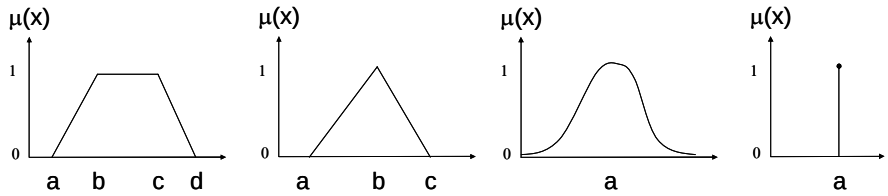
จากข้างต้นจะเห็นว่าเซตคนที่มีอายุน้อยจะมีขอบเขตถึง 30 ปี โดยที่ช่วงระหว่าง 20 – 30 ปี นั้นค่าความเป็นสมาชิกจะลดลงด้วยกราฟเส้นตรงและถ้าเราสร้างกราฟของฟังก์ชั่นความเป็นสมาชิกของเซตคนที่มีอายุน้อย คนที่มีอายุปานกลาง และคนที่มีอายุมาก ก็จะได้กราฟดังรูปที่ 4



ภาพที่ 4 ฟังก์ชั่นความเป็นสมาชิกเซตคนที่มีอายุน้อย คนที่มีอายุปานกลาง และคนที่มีอายุมาก

3.2 ฟังก์ชั่นความเป็นสมาชิก (Membership Function)

ในงานประยุกต์ทั่วไปที่มีการใช้ฟัซซี่เข้ามาเกี่ยวข้องนั้น จะมีการสร้างฟังก์ชั่นความเป็นสมาชิกขึ้นมาใช้งาน ซึ่งรูปร่างกราฟของฟังก์ชั่นความเป็นสมาชิกนั้นอาจจะมีรูปร่างแตกต่างกันไป แต่ที่นิยมใช้กันจะมีอยู่ 4 แบบ คือ แบบสี่เหลี่ยมคางหมู (Trapezoidal) แบบสามเหลี่ยม (Triangular) แบบเก้าส์เซี่ยน (Gaussian) และแบบเส้นเดี่ยว (Singleton) ดังตัวอย่างในภาพที่ 5



ภาพที่ 5 รูปร่างที่นิยมของฟังก์ชั่นความเป็นสมาชิก แบบสี่เหลี่ยมคางหมู แบบสามเหลี่ยม แบบเก้าส์เซี่ยน และแบบเส้นเดี่ยว ตามลำดับ

จากภาพที่ 5 กราฟฟังก์ชั่นความเป็นสมาชิกแบบสี่เหลี่ยมคางหมูจะนิยมใช้แทนค่าที่เป็นช่วงโดย ค่าความเป็นสมาชิกจะเป็น 1 ในช่วง b ถึง c และค่าความเป็นสมาชิกจะลดลงเป็นกราฟเส้นตรงที่ b ไป a และ c ไป d ในกราฟแบบสามเหลี่ยมและแบบเก้าส์เซี่ยนจะนิยมใช้แทนค่าที่เป็นค่าเดี่ยว โดยแบบสามเหลี่ยมจะมีการลดค่าความเป็นสมาชิกแบบเส้นตรง แต่แบบเก้าส์เซี่ยนจะลดแบบเก้าส์เซี่ยน ส่วนกราฟแบบเส้นเดี่ยวจะใช้แทนค่าเดี่ยวแบบชัดเจน

ฟังก์ชั่นความเป็นสมาชิกจะมีคุณสมบัติที่เกี่ยวข้องอยู่ 4 ชนิด ได้แก่ สนับสนุน (Support) แกน (Core) ตัด α (α-cut) และความสูง (Height) ดังนี้

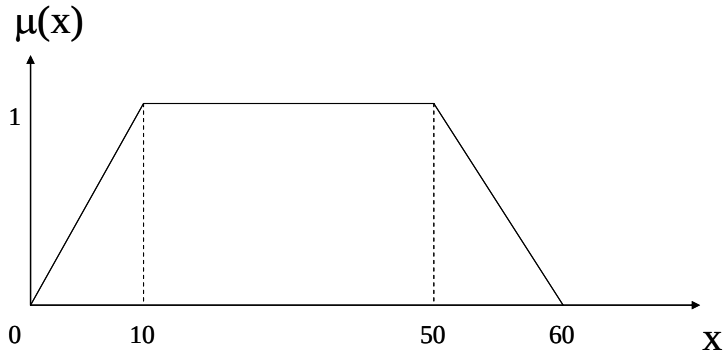
สนับสนุน (Support): $s_A := \{x : \mu_A(x) > 0\}$, หมายถึงสมาชิกทุกตัวในเซต A ที่มีค่าความเป็นสมาชิกที่มากกว่า 0

แกน (Core): $c_A := \{x : \mu_A(x) = 1\}$, หมายถึงสมาชิกทุกตัวในเซต A ที่มีค่าความเป็นสมาชิกเป็น 1

ตัด α (α-cut): $A_\alpha := \{x : \mu_A(x) \geq \alpha\}$, หมายถึงสมาชิกทุกตัวในเซต A ที่มีค่าความเป็นสมาชิกมากกว่าหรือเท่ากับค่า α

ความสูง (Height): $h_A := \max_x\{\mu_A(x)\}$, หมายถึงค่าความเป็นสมาชิกที่สูงที่สุดในเซต A

ตัวอย่างเช่น กำหนดให้มีฟัซซี่เซต A ดังแสดงในภาพที่ 6 เราสามารถหาค่าคุณสมบัติทั้ง 4 คุณสมบัติของฟัซซี่เซต A ได้ดังนี้ สนับสนุนคือ เซตที่มีสมาชิกตั้งแต่ 0 – 60 แกนคือ เซตที่มีสมาชิกตั้งแต่ 10 – 50 ตัด α ถ้ากำหนดให้ α มีค่าเท่ากับ 0.5 ฉะนั้น ตัด 0.5 คือ เซตที่มีสมาชิกตั้งแต่ 5 – 55 และสุดท้ายความสูงคือ 1 ในตัด α นั้น ถ้าค่า α มีค่าเท่ากับ 1 ค่าสมาชิกเซตจะเท่ากับแกน



ภาพที่ 6 ตัวอย่างฟัซซี่เซต A

3.3 ตรรกะฟัซซี่ (Fuuzy Logic)

ในทฤษฎีฟัซซี่นั้น ฟัซซี่เซตสามารถมีตัวปฏิบัติการ (operator) เช่น การเชื่อมแบบ “และ” (Conjunction) การเชื่อมแบบ “หรือ” (Disjunction) และส่วนเติมเต็ม (Complement) โดยจะมีการกำหนดการปฏิบัติการดังนี้

การเชื่อมแบบ “และ” (Conjunction): $\mu_{A\wedge B}(x) := \min\{\mu_A(x), \mu_B(x)\}$

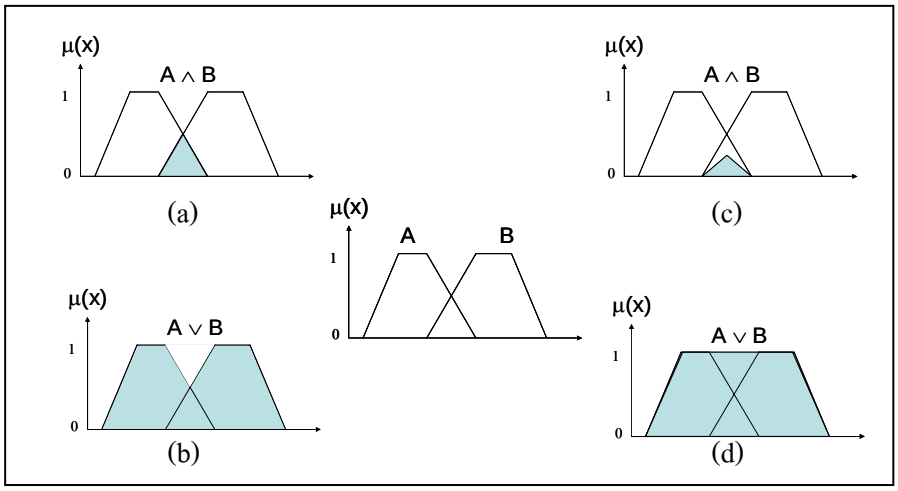
การเลือกแบบ “หรือ” (Disjunction): $\mu_{A\vee B}(x) := \max\{\mu_A(x), \mu_B(x)\}$

ส่วนเติมเต็ม (Complement): $\mu_{\neg A}(x) := 1 - \mu_A(x)$

สมมุติเรากำหนดให้ x มีค่าความเป็นสมาชิกในเซตคนสูงเป็น 0.8 และมีค่าความเป็นสมาชิกในเซตคนเตี้ย 0.4 เราจะกล่าวได้ว่า x มีค่าความเป็นสมาชิกในเซตคนสูงและ (Conjunction) คนเตี้ย คือ min(0.8,0.4) นั่นคือเป็น 0.4 นั่นเอง หรือ x มีค่าความเป็นสมาชิกในเซตคนสูงหรือ (Disjunction) คนเตี้ยคือ max(0.8,0.4) นั่นคือ 0.8 นั่นเอง การกำหนดการปฏิบัติการนอกจากการใช้ค่าสูงสุด (Max) และค่าต่ำสุด (Min) แล้วยังมีการกำหนดในแบบอื่น ๆ อีกด้วย เช่น การใช้การคูณ และ การบวก ภาพที่ 7 แสดงการปฏิบัติการการเชื่อมแบบ “และ” และการเชื่อมแบบ “หรือ” ในแบบต่าง ๆ ของฟัซซี่เซต A และ ฟัซซี่เซต B

การเชื่อมแบบ “และ” (Conjunction): $\mu_{A\wedge B}(x) := \mu_A(x) \bullet \mu_B(x)$

การเลือกแบบ “หรือ” (Disjunction): $\mu_{A\vee B}(x) := \min(\mu_A(x) + \mu_B(x), 1)$



ภาพที่ 7 ตัวอย่างการปฏิบัติการฟัซซี่ (a) การปฏิบัติการการเชื่อมแบบ “และ” (Conjunction) โดยใช้ค่าต่ำสุด (b) การเชื่อมแบบ “หรือ” (Disjunction) โดยใช้ค่าสูงสุด (c) การปฏิบัติการการเชื่อมแบบ “และ” (Conjunction) โดยใช้การคูณ (d) การเชื่อมแบบ “หรือ” (Disjunction) โดยใช้การบวก

3.4 กฎฟัซซี่ (Fuzzy Rules)

ในงานประยุกต์ที่ใช้ระบบกฎ (Rule-Based System) ในการทำงานนั้น จะประมวลผลข้อมูลโดยใช้กฎต่าง ๆ ในการอนุมาน (Inferences) ผลลัพธ์ออกมา ซึ่งกฎจะอยู่ในรูปแบบที่เป็นเงื่อนไขกับผลลัพธ์ที่ประกอบด้วยตัวแปร (Variable) และค่าเปรียบเทียบ ดังนี้

$$\text{IF age(x)} \leq 25 \text{ THEN risk(x)} > 60\%$$

กฎข้างต้นเป็นตัวอย่างของกฎอนุมานเกี่ยวกับความเสี่ยงในการทำประกันภัยรถยนต์ โดยจะมีความหมายว่า ถ้าผู้ทำประกันภัย x มีอายุน้อยกว่าหรือเท่ากับ 25 จะมีความเสี่ยงในการประกันภัยมากกว่า 60 % นั่นเอง ในระบบฟัซซี่กฎสามารถใช้ภาษา (Linguistic) แทนค่าตัวเลข ซึ่งจากกฎข้างต้น สามารถนำมาเขียนเป็นกฎใหม่ได้ดังนี้

IF age(x) IS young THEN risk(x) IS high

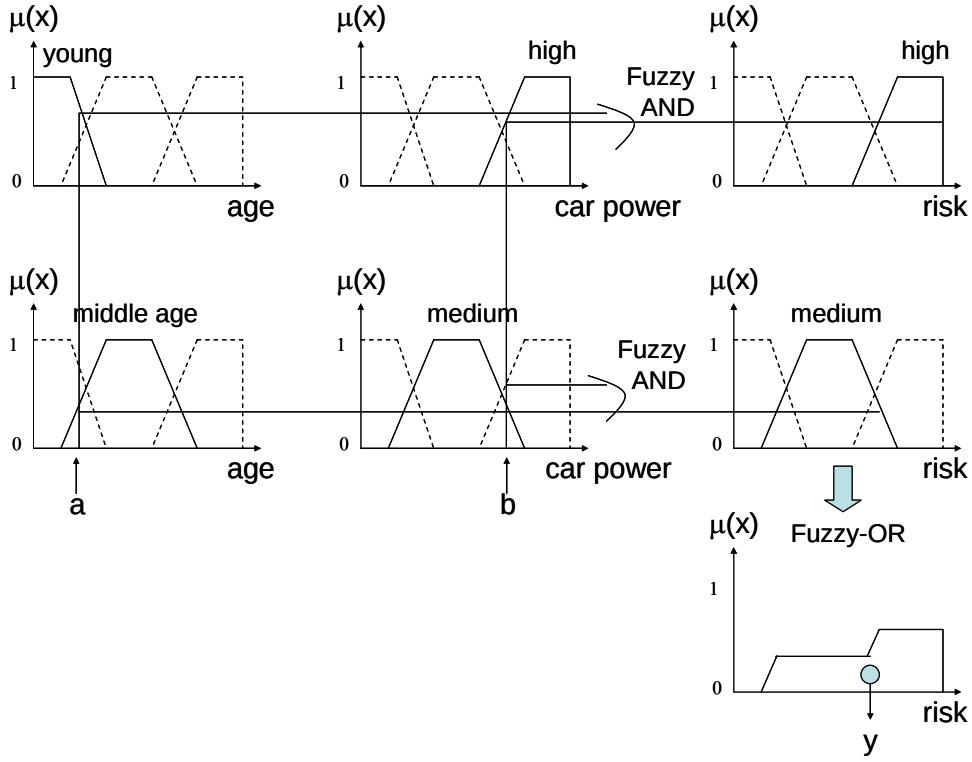
จากกฎ จะมีความหมายว่า ถ้าผู้ทำประกันภัยมีอายุน้อย (young) ก็จะมีความเสี่ยงในการทำประกันภัยสูง (high) นั่นเอง โดยในระบบเราต้องกำหนดฟังก์ชั่นความเป็นสมาชิกของค่าอายุน้อย (young age) และค่าความเสี่ยงสูง (high risk) ขึ้นในระบบก่อน ซึ่งกฎฟัซซี่จะอยู่ในรูปของการเชื่อมแบบ “และ” กันของเงื่อนไข ดังนี้

$$\text{IF } x_1 \text{ IS } A_1 \text{ AND } \ldots \text{ AND } x_n \text{ IS } A_n \text{ THEN y IS B}$$

ในงานประยุกต์หนึ่งอาจจะประกอบด้วยกฎมากกว่าหนึ่งกฎ ซึ่งทุกกฎก็จะทำการเชื่อมแบบ “หรือ” (Disjunction) กัน ในการประมวลผลหาผลลัพธ์ โดยเราสมมุติว่ากฎฟัซซี่ 2 กฎดังนี้

IF age IS young AND car power IS high THEN risk IS high
IF age IS middle age AND car power IS medium THEN risk IS medium

ในการอนุมานทั้งสองกฎนี้ เราจะต้องกำหนดฟังก์ชั่นความเป็นสมาชิกของค่าอายุ (age) ค่ากำลังรถ (car power) และค่าความเสี่ยง (risk) จากนั้นจึงนำเงื่อนไขคือค่าอายุและค่ากำลังรถมาทำการเชื่อมแบบ “และ” กัน และนำผลลัพธ์ของทั้งสองกฎคือค่าความเสี่ยงมาเชื่อมแบบ “หรือ” กัน ดังภาพที่ 7 โดยที่กำหนดให้ มีอายุ a และกำลังรถ b จะได้ความเสี่ยงออกมาเป็น y จากกฎฟัซซี่ โดยที่ค่า y จะได้จากการหาจุดศูนย์ถ่วงของพื้นที่ความเสี่ยงทั้งหมดออกมาเป็นค่า y นั่นเอง



<u>ภาพที่ 8</u> การอนุมานกฎฟัซซี่ (Berthold, M. and D. J. Hand, 2003)

4. การคำนวณความคล้าย (Similarity Measure)

การวัดความคล้ายคือ การคำนวณความใกล้เคียงกันของข้อมูลสองชุด โดยถ้าข้อมูลชุดนั้นมีความใกล้เคียงกันมากจะให้ค่าความคล้ายที่สูง แต่ถ้าข้อมูลชุดนั้นมีความใกล้เคียงกันน้อยจะให้ค่าความใกล้เคียงที่ต่ำ ซึ่งงานประยุกต์ (Application) ที่ต้องใช้การวัดความคล้ายจะมีตัวอย่างเช่น

(ก) การแยกความแตกต่างของอ็อบเจกต์ 2 อ็อบเจกต์

(ข) การจัดกลุ่ม (Clustering) ของอ็อบเจกต์ หรือเอกสาร หรือกลุ่มข้อมูล

(ค) การสืบค้นอ็อบเจกต์ หรือเอกสาร หรือรูปภาพ ต่าง ๆ ในระบบสืบค้น

(ง) การจำแนก (Classify) อ็อบเจกต์ หรือเอกสาร ไปยังกลุ่มที่กำหนด

ทั้งนี้วิธีการคำนวณความคล้ายในแต่ละงานประยุกต์ จะมีวิธีการคำนวณแตกต่างกันไป ทั้งนี้ขึ้นอยู่กับชนิดของข้อมูลในแต่ละรูปแบบงานนั้น ๆ โดยจะมีชนิดข้อมูลอันได้แก่

(ก) ข้อมูลเชิงทวิภาค (Binary Variable)

(ข) ข้อมูลเชิงนามบัญญัติ (Nominal Variable)

(ค) ข้อมูลเชิงอันดับ (Ordinal Variable)

(ง) ข้อมูลเชิงปริมาณ (Quantitative Variable)

ในการคำนวณความคล้ายสำหรับข้อมูลในแต่ละชนิดนั้นจะมีวิธีการคำนวณที่แตกต่างกันไปดังจะได้อธิบายในหัวข้อย่อยดังนี้

4.1 การคำนวณความคล้ายของข้อมูลเชิงทวิภาค

ข้อมูลเชิงทวิภาคคือ ข้อมูลที่มีค่าเพียง 2 ค่า เช่น ถูก (true), ผิด (false), ใช่, ไม่ใช่, เห็นด้วย, ไม่เห็นด้วย, จริง, ไม่จริง เป็นต้น ซึ่งการคำนวณความคล้ายสำหรับข้อมูลชนิดนี้จะเป็นการคำนวณความแตกต่างของข้อมูลในแต่ละคุณสมบัติ (Feature) ดังตัวอย่างในตารางที่ 1

ตารางที่ 1 ตัวอย่างข้อมูลเชิงทวิภาค

ผลไม้	รูปร่างกลม	รสหวาน	รสเปรี้ยว	มีความกรอบ
แอปเปิ้ล	ใช่	ใช่	ใช่	ใช่
กล้วย	ไม่ใช่	ใช่	ไม่ใช่	ไม่ใช่

จากตารางที่ 1 เราสามารถเขียนข้อมูลของแอปเปิ้ลโดยนำคุณสมบัติมาเขียนเป็นเวกเตอร์ได้คือ (1,1,1,1) และสามารถเขียนข้อมูลของกล้วยได้คือ (0,1,0,0) ซึ่งตัวเลขเหล่านี้จะเป็นตัวแทนของอ็อบเจกต์ทั้งสอง ซึ่งเราอาจจะกล่าวได้ว่าข้อมูลของอ็อบเจกต์ทั้งสองมีขนาด 4 มิติ โดยในการคำนวณความคล้ายนั้นเราจะกำหนดตัวแปรสำหรับอ็อบเจกต์สองอ็อบเจกต์ที่กำหนด คือ อ็อบเจกต์ i และ อ็อบเจกต์ j ได้ดังนี้

p คือ จำนวนคุณสมบัติที่อ็อบเจกต์ i และ อ็อบเจกต์ j มีค่าเป็น 1 ทั้งสองอ็อบเจกต์

q คือ จำนวนคุณสมบัติที่อ็อบเจกต์ i มีค่าเป็น 1 และอ็อบเจกต์ j มีค่าเป็น 0

r คือ จำนวนคุณสมบัติที่อ็อบเจกต์ i มีค่าเป็น 0 และอ็อบเจกต์ j มีค่าเป็น 1

s คือ จำนวนคุณสมบัติที่อ็อบเจกต์ i และอ็อบเจกต์ j มีค่าเป็น 0 ทั้งสองอ็อบเจกต์

t คือ ค่าตัวแปรทั้งหมดรวมกัน ซึ่งก็คือ p+q+r+s นั่นเอง

ดังนั้นจากตัวอย่างในตารางที่ 1 นั้น เราจะสามารถกำหนดตัวแปรในการคำนวณความคล้ายของแอปเปิ้ลกับกล้วย ถ้าเรากำหนดให้แอปเปิ้ลเป็นอ็อบเจกต์ i และกล้วยเป็นอ็อบเจกต์ j เราจะเห็นว่ากรณีที่คุณสมบัติของอ็อบเจกต์ทั้งสองจะมีค่าเป็น 1 ทั้งคู่เพียงคุณสมบัติเดียวคือรสหวาน ฉะนั้นได้ว่า p=1 และคุณสมบัติที่อ็อบเจกต์ i เป็น 1 และอ็อบเจกต์ j เป็น 0 มี 3 คุณสมบัติคือ รูปร่างกลม รสเปรี้ยวและมีความกรอบ ฉะนั้นจะได้ q=3 ส่วนค่าตัวแปรที่เหลือจะได้ r=0 และ s=0 รวมทั้งตัวแปร t = p+q+r+q = 4 และในการคำนวณความคล้ายของข้อมูลเชิงทวิภาคนั้นจะมีสูตรที่นิยมดังนี้

(ก) Simple Matching Distance: $d_{ij} = \frac{q+r}{t}$

(ข) Jaccard's Distance: $d_{ij} = \frac{q+r}{p+q+r}$

(ค) Hamming's Distance: $d_{ij} = q+r$

4.2 การคำนวณความคล้ายของข้อมูลเชิงนามบัญญัติ

ข้อมูลเชิงนามบัญญัติคือ ข้อมูลที่มีลักษณะข้อมูลเป็นคำนาม เช่น รสของผลไม้ รูปร่างของผลไม้ สัญชาติของบุคคล สัตว์เลี้ยงที่ชอบ เป็นต้น ดังตัวอย่างในตารางที่ 2

ตารางที่ 2 ตัวอย่างข้อมูลเชิงนามบัญญัติ

ผลไม้	รส	รูปร่าง
แอปเปิ้ล	หวาน	กลม
กล้วย	หวาน	ยาวรี
มะม่วง	เปรี้ยว	ยาวรี

ในการแสดงค่าข้อมูลเชิงนามบัญญัตินั้นจะต้องแปลงค่าเชิงนามบัญญัติให้อยู่ในรูปของข้อมูลเชิงทวิภาคก่อน ซึ่งจากตารางที่ 2 ผลไม้จะมีคุณสมบัติ 2 คุณสมบัติ คือรสและรูปร่าง โดยรสจะมีค่า 2 ค่า คือ หวานและเปรี้ยว จึงแปลงเป็นค่าเชิงทวิภาคได้ 2 คุณสมบัติคือ มีรสหวานและมีรสเปรี้ยว ส่วนรูปร่างก็จะมี 2 ค่าเช่นกัน คือ กลมและยาวรี จึงแปลงเป็นค่าเชิงทวิภาคได้ 2 คุณสมบัติ

เช่นกัน คือ มีรูปร่างกลมและมีรูปร่างยาวรี จึงทำให้เมื่อแปลงเป็นข้อมูลเชิงทวิภาคแล้วจะเป็นข้อมูลขนาด 4 หลัก ซึ่งข้อมูลในตารางที่ 2 จะสามารถแสดงความเป็นชนิดของผลไม้โดยค่าเวกเตอร์ของรสและรูปร่างดังต่อไปนี้

ชนิดผลไม้ = ((หวาน,เปรี้ยว),(กลม,ยาวรี))

จากตารางที่ 2 จึงได้ข้อมูลดังนี้

แอปเปิ้ล = ((1,0),(1,0))
กล้วย=((1,0),(0,1))
มะม่วง=((0,1),(0,1))

ในการคำนวณความคล้ายนั้นจะสามารถใช้วิธีแบบเดียวกับข้อมูลเชิงทวิภาคในการคำนวณความคล้ายนั่นเอง

4.3 การคำนวณความคล้ายของข้อมูลเชิงอันดับ

ข้อมูลเชิงอันดับคือ ข้อมูลที่เป็นอันดับของค่าข้อมูล ซึ่งจะเป็นข้อมูลที่จำเป็นในงานที่เกี่ยวกับการสำรวจข้อมูลต่าง ๆ ดังเช่น เห็นด้วยอย่างมาก, เห็นด้วย, เฉย ๆ, ไม่เห็นด้วย, ไม่เห็นด้วยอย่างยิ่ง เป็นต้น ซึ่งตัวอย่างของข้อมูลเชิงอันดับได้แสดงดังตารางที่ 3

ตารางที่ 3 ตัวอย่างข้อมูลเชิงอันดับ

ที่จอดรถ	ความปลอดภัย	ความกว้างขวาง	ความเหมาะสม	ความใกล้
A	-1	2	0	-2
B	1	0	1	1

จากตารางที่ 3 จะเป็นข้อมูลเกี่ยวกับความคิดเห็นเกี่ยวกับที่จอดรถสองแห่ง โดยมีคุณสมบัติ 4 อย่างคือ ความปลอดภัย ความกว้างขวาง ความเหมาะสมและความใกล้ ซึ่งค่าของคุณสมบัติจะมี 5 ค่าคือ -2 = ไม่เห็นด้วยอย่างยิ่ง, -1 = ไม่เห็นด้วย, 0 = เฉย ๆ , 1 = เห็นด้วย และ 2 = เห็นด้วยอย่างยิ่ง

ซึ่งการจะคำนวณค่าความคล้ายของข้อมูลเชิงอันดับจะต้องทำข้อมูลให้เป็นบรรทัดฐาน (Normalized) ก่อนคือ จะต้องจัดข้อมูลที่อยู่ในรูปอันดับให้กลายเป็นค่า 1 ถึง R ก่อน เช่น ไม่เห็นด้วยอย่างยิ่ง=1, เห็นด้วย=2, เฉย ๆ=3, เห็นด้วย=4, เห็นด้วยอย่างยิ่ง=5 เป็นต้น ซึ่งในที่นี่ R = 5 จากนั้นจึงแปลงค่าให้อยู่ในช่วงค่า 0 ถึง 1 โดยใช้สูตรดังนี้

$$x = \frac{r-1}{R-1}$$

โดยที่ x คือ ค่าที่แปลงแล้ว
r คือ ค่าที่ถูกจัดให้กับค่าอันดับนั้น ๆ
R คือ ค่าสูงสุดของค่า r

จากนั้นจึงสามารถคำนวณค่าความคล้ายได้ด้วยสูตรคำนวณเดียวกับข้อมูลเชิงปริมาณได้โดยตรง เช่น Euclidean distance, City block distance, Canberra distance, Angular separation เป็นต้น

4.4 การคำนวณความคล้ายของข้อมูลเชิงปริมาณ

ข้อมูลเชิงปริมาณคือ ข้อมูลที่มีค่าอยู่ในรูปของเลขจำนวนจริง ซึ่งข้อมูลชนิดนี้จะเป็นข้อมูลที่ใช้กันมากที่สุด ซึ่งตัวอย่างของข้อมูลเชิงปริมาณได้แสดงดังในตารางที่ 4

ตารางที่ 4 ตัวอย่างข้อมูลเชิงปริมาณ

คอมพิวเตอร์โน๊ตบุค	**ราคา (พันบาท)**	**น้ำหนัก (กิโลกรัม)**	**ขนาดจอภาพ (นิ้ว)**
A	48.5	3.5	12
B	60.8	2.6	14

จากตารางที่ 4 เราสามารถเขียนข้อมูลของอ็อบเจกต์ A ในรูปของเวกเตอร์ได้คือ (48.5, 3.5, 12) และอ็อบเจกต์ B คือ (60.8, 2.6, 14) ในการคำนวณความคล้ายของข้อมูลเชิงปริมาณนั้นจะมีสูตรที่นิยมดังนี้ โดยกำหนดให้ k คือ คุณสมบัติ, x คือค่าคุณสมบัติ, i และ j คือ อ็อบเจกต์ที่ i และ j ดังนี้

(ก) Euclidean distance: $d_{ij} = \sqrt{\sum_{k=1}^{n}\left(x_{ik} - x_{jk}\right)^2}$

(ข) City block distance: $d_{ij} = \sum_{k=1}^{n}\left|x_{ik} - x_{jk}\right|$

(ค) Canberra distance: $d_{ij} = \sum_{k=1}^{n}\frac{\left|x_{ik} - x_{jk}\right|}{\left|x_{ik}\right| + \left|x_{jk}\right|}$

(ง) Bray Curtis distance: $d_{ij} = \frac{\sum_{k=1}^{n}\left|x_{ik} - x_{jk}\right|}{\sum_{k=1}^{n}\left(x_{ik} + x_{jk}\right)}$

(จ) Angular separation: $d_{ij} = \frac{\sum_{k=1}^{n} x_{ik} \bullet x_{jk}}{\sqrt{\sum_{k=1}^{n} x_{ik}^2 \bullet \sum_{k=1}^{n} x_{jk}^2}}$

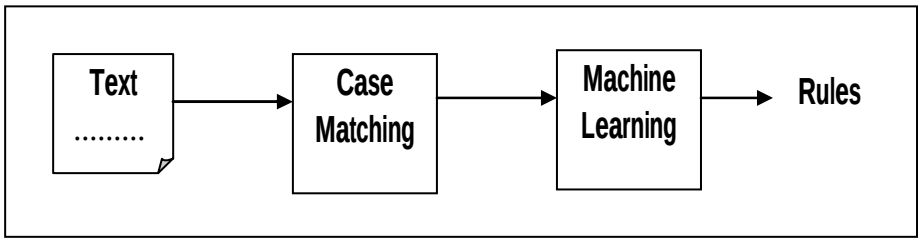
งานวิจัยที่เกี่ยวข้อง

การค้นพบความรู้จากเอกสาร (Knowledge Discovery from Text) เป็นงานวิจัยที่ท้าทาย ทั้งในงานวิจัยทางด้านปัญญาประดิษฐ์ (Artificial Intelligent) และฐานข้อมูล (Database) การค้นพบความรู้จากเอกสารนั้น เราสามารถแบ่งวิธีการค้นพบได้ 3 วิธี คือ เทคนิคการเรียนรู้ (Machine Learning), การเทียบแม่แบบ (Template Matching) และการทำเหมืองเอกสาร (Text Data Mining)

1. เทคนิคการเรียนรู้ (Machine Learning)

เทคนิคการเรียนรู้อย่างเช่นระบบ Inductive Logic Programming (ILP) (Muggleton, 1992) และ Explanation-based Generalization หรือ EBL (Mitchell *et al.*, 1986) ได้ถูกนำมาใช้ในการค้นพบความรู้จากเอกสาร โดยในปี 1993 นั้น Delannoy *et al.* (1993) ได้เสนอระบบ MaLTe (Machine Learning from Text) ซึ่งจะทำการสกัดประโยคทั้งหมดมาเป็นตัวแทนความรู้ ซึ่งในงานชิ้นนั้นจะใช้ตรรกะภาคแสดง (Predicate Logic) มาเป็นตัวแทนความรู้ทั้งประโยคที่ปรากฏ แล้วจึงใช้เทคนิคการเรียนรู้ ILP มาทำการหาความสัมพันธ์ของตรรกะภาคแสดง แล้วนำมาสร้างเป็นกฎ ซึ่ง

การแปลงประโยคให้เป็นตรรกะภาคแสดง (Predicate Logic) นั้น ได้ใช้เทคนิค Case-Matching (Copeck, 1992) ซึ่งตัวอย่างกระบวนการได้แสดงในภาพที่ 9



ภาพที่ 9 กระบวนการโดยรวมของระบบ MaLTe

เมื่อมีเอกสารถูกป้อนเข้าไปในระบบ MaLTe จะผ่านกระบวนการ Case-Matching เพื่อทำการแปลงประโยคภาษาอังกฤษให้กลายเป็นตรรกะภาคแสดง (Predicate Logic) ซึ่งตัวอย่างประโยคที่แปลงเป็นตรรกะภาคแสดงได้แสดงในภาพที่ 10

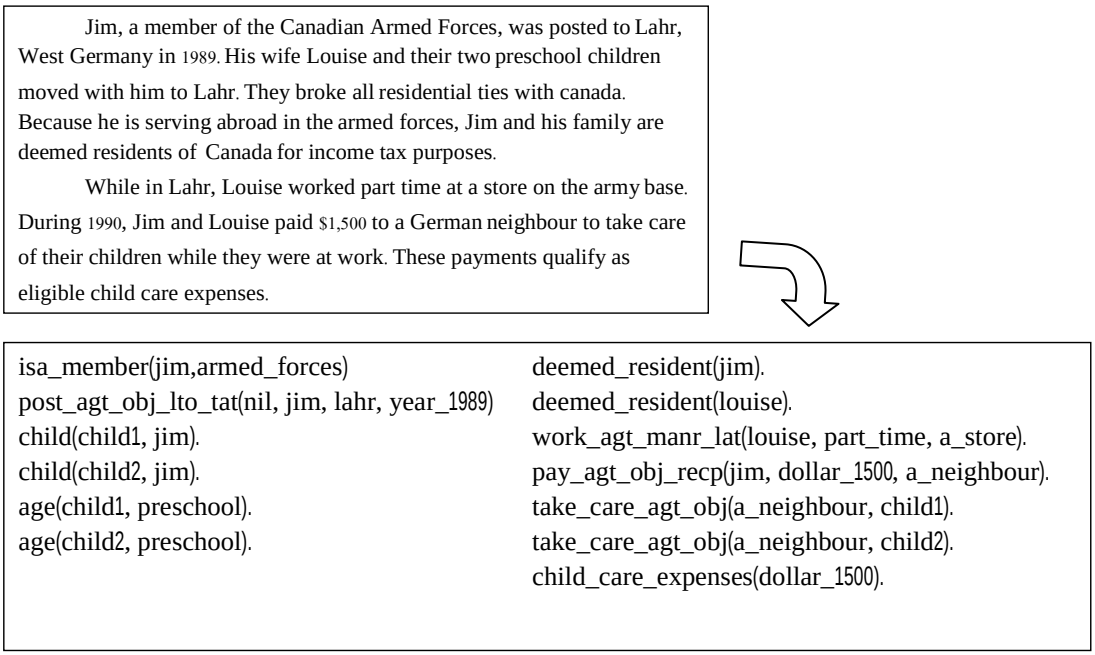
Jim, a member of the Canadian Armed Forces, was posted to Lahr, West Germany in 1989

isa_member(jim,armed_forces)

post_agt_obj_lto_tat(nil, jim, lahr, year_1989)

ภาพที่ 10 ตัวอย่างประโยคที่ถูกแปลงเป็นตรรกะภาคแสดง

ตัวอย่างเอกสารในระบบ MaLTe ได้แสดงในภาพที่ 11 ระบบจะทำการแปลงประโยคทั้งหมดที่มีอยู่ในเอกสารฉบับนั้นให้เป็นตรรกะภาคแสดงทั้งหมด



ภาพที่ 11 การแปลงเอกสารข้อความเป็นตรรกะภาคแสดง

หลังจากได้แปลงประโยคในเอกสารทั้งหมดเป็นตรรกะภาคแสดงหมดทั้งเอกสารดังแสดงในภาพที่ 11 แล้ว จากนั้นจึงทำการเรียนรู้ความสัมพันธ์ของอาร์กิวเมนต์ในตรรกะภาคแสดงทั้งหมดด้วย ILP และ EBL ออกมาเป็นกฎดังแสดงไว้ในภาพที่ 12

```
claim_agt_obj(Person, Exp) :-
        isa_member(Person, armed_forces),
        serve_agt_obj_lat(Person, armed_forces, abroad),
        child(FirstChild, Person),
        child(SecondChild, Person),
        age(FirstChild, preschool),
        age(SecondChild, preschool),
        pay_agt_obj_recp(Person, Exp, SomeOne),
        take_care_agt_obj(SomeOne, FirstChild),
        take_care_agt_obj(SomeOne, SecondChild).
```

ภาพที่ 12 ตัวอย่างกฎที่เรียนรู้แล้วในระบบ MaLTe

การค้นพบความรู้ในระบบ MaLTe นี้สามารถค้นพบความรู้ใหม่ ๆ ได้จากการเรียนรู้ความสัมพันธ์ของอาร์กิวเมนต์ของตรรกะภาคแสดงซึ่งเหมาะในโดเมนเปิด (Open Domain) แต่ปัญหาสำคัญของการเรียนรู้แบบนี้ คือ ความไม่สมบูรณ์ของความรู้ (Incomplete Knowledge) นั่นคือ ถ้าเอกสารมีองค์ประกอบของความรู้ไม่ครบถ้วนจะทำให้การเรียนรู้เพื่อสร้างกฎจะได้กฎที่ไม่สมบูรณ์ หรือผิดพลาดได้

2. การเทียบแม่แบบ (Template Matching)

การใช้แม่แบบมาใช้ในการสกัดความรู้นั้น เป็นวิธีการสกัดความรู้ที่ใช้กันมากในกระบวนการเสาะหาความรู้ ซึ่งจะเหมาะกับงานที่เป็นโดเมนเฉพาะ (Specific Domain) และสกัดความรู้ที่มีลักษณะปรากฏชัดเจน (Explicit Knowledge) ในเอกสาร ในปี 1994 นั้น Gomez *et al.* (1994) ได้นำเสนอระบบสกัดความรู้เกี่ยวกับสัตว์ อาหารของสัตว์ และที่อยู่ของสัตว์ จากเอกสารสารานุกรม (Encyclopedia) โดยได้มีการพัฒนาระบบที่ชื่อว่า SNOWY ขึ้นมา โดยประโยคในเอกสารสารานุกรมจะถูกวิเคราะห์วากยสัมพันธ์ (Syntax) ด้วยโปรแกรมแจงประโยค (Parser) จนกระทั่งระบุความหมายของกิริยาในประโยคได้ ซึ่งการระบุความหมายของกิริยาในประโยคนั้นจะถูกระบุด้วยกฎที่ชื่อว่า VM rules โดยภาพที่ 13 แสดงถึงตัวอย่างของ VM rules และภาพที่ 14 แสดงตัวอย่างของประโยค “The crowned eagle of Africa lives in the rain forests and eats monkeys.” ที่ผ่านการวิเคราะห์วากยสัมพันธ์และระบุความหมายของประโยคแล้ว

```
(ACTION
 (is-a (relation))
 (subj (thing (actor))) (obj (thing (theme)))
 (adverb
  (time (at-time))
  (negation (negation)) (frequency (frequency))
  .....
 (prep
  (in (ltm-ctgy (physical-thing (at-loc (weak)))
     (time-unit (at-time (strong)))))
  (at (ltm-ctgy (physical-thing (at-loc (strong)))
     (time-unit (at-time (strong)))))
  (during (ltm-ctgy (physical-thing (at-time (strong)))))
  (with (ltm-ctgy (animate-body-part (instrument (strong))))
     (state (state (strong)))))
```

```
(INGEST
 (is-a (action))
 (subj (animate (actor)))
 (prep
  (with (ltm-ctgy
    (utensil (instrument (strong)))
    (human (accompany (strong)))
    (physical-thing
      (co-theme (weak))))))
 (obj (physical-thing (theme)))
```

```
(DRINK
 (is-a (ingest))
 (obj (liquid (theme))))
```

<u>ภาพที่ 13</u> ตัวอย่างของกฎที่ใช้ระบุความหมายของกิริยาในประโยค (Gomez et al, 1994)

จากภาพที่ 13 จะเป็นตัวอย่างของกฎที่ใช้ระบุความหมายของกิริยา ซึ่งจะแสดงตัวอย่างกฎของกิริยา 3 ตัว คือ ACTION, INGEST และ DRINK โดยภายในกฎจะมีอาร์กิวเมนต์ (Argument) ที่ใช้ระบุกฎการระบุความสัมพันธ์ทางความหมายของกิริยาเช่นในกิริยา INGEST จะมีอาร์กิวเมนต์ (subj (animate (actor))) จะหมายความว่า กิริยา INGEST ถ้าประธานเป็น animate แล้ว animate จะมีความสัมพันธ์ทางความหมายกับกิริยา INGEST เป็น actor นั่นเอง ในส่วนของบุพบท (prep (with (ltm-ctgy (utensil (instrument (strong))) หมายถึงบุพบท with ถ้าตามด้วย utensil แล้ว utensil จะมีความสัมพันธ์ทางความหมายกับกิริยา INGEST เป็น instrument นั่นเอง นอกจากนี้กฎที่ใช้ระบุความหมายนี้สามารถมีการถ่ายทอด (Inherit) ไปยังกฎอื่นได้ด้วย ดังเช่นอาร์กิวเมนต์ (is-a (action)) ของกิริยา INGEST จะหมายถึง กิริยา INGEST นั้นถ่ายทอดมาจากกิริยา ACTION ฉะนั้นอาร์กิวเมนต์ต่าง ๆ ของกิริยา ACTION ก็จะใช้ได้กับกิริยา INGEST ด้วยเช่นกัน

```
(REPS ((VRS (G233)))
  SUBJ ((PARSER ((DFART THE) (VERB CROWNED) (NOUN EAGLE)))
    (REF (DEFINITE)) (PLURAL NIL)
    (INTERP (CROWNED-EAGLE (Q (ALL)))) (SEMANTIC-ROLE (ACTOR)))
  PREP ((PARSE (OF ((PN AFRICA)))) (INTERP (AFRICA (Q (CONSTANT))))
    (ATTACH-TO (CROWNED-EAGLE (LOCATION-R (AFRICA)))))
  VERB ((MAIN-VERB EAT EATS) (TENSE PRES)(NUM SING) (PRIM (INGEST)))
  OBJ ((PARSE ((NOUN MONKEYS))) (PLURAL T)) (INTERP (MONKEY (Q (?))))
    (SEMANTIC-ROLE (THEME))))
Output for Structure G233:
(SUBJ ((PARSE ((DFART THE) (VERB CROWNED) (NOUN EAGLE)))
  (REF (DEFINITE)) (PLURAL NIL)
  (INTERP (CROWNED-EAGLE (Q (ALL)))) (SEMANTIC-ROLE (ACTOR)))
VERB ((MAIN-VERB LIVE LIVES) (TENSE PRES) (NUM SING) (PRIM (INHABIT)))
PREP ((PARSE (IN ((DFART THE) (NOUN RAIN) (NOUN FOREST))))
 (REF (DEFINITE)) (PLURAL NIL)
 (INTERP (RAIN-FOREST (Q (?)))) (SEMANTIC-ROLE (AT-LOC))
 (ATTACH-TO (VERB (STRONGLY))))
```

<u>ภาพที่ 14</u> ประโยคที่ผ่านการวิเคราะห์วากยสัมพันธ์และความหมายของประโยคในระบบ SNOWY (Gomez et al, 1994)

จากภาพที่ 14 จะเป็นตัวอย่างของการแสดงความหมายของประโยค เช่น SUBJ ((PARSER ((DFART THE)(VERB CROWNED)(NOUN EAGLE)))(REF (DEFINITE))(PLURAL NIL) … จะหมายถึงที่ตำแหน่งประธานจะมีคำ THE CROWNED EAGLE ซึ่ง DFART, VERB และ NOUN เป็นชนิดของคำของ THE, CROWNED และ EAGLE ตามลำดับ โดยเป็นคำนามแบบ DEFINITE และ ไม่เป็น PLURAL ต่อมา (INTERP (CROWNED-EAGLE (Q (ALL))))(SEMANTIC-ROLE

(ACTOR))) หมายถึงแปลว่า CROWNED-EAGLE มีปริมาณคือทั้งหมด และมี SEMANTIC-ROLE เป็น ACTOR นั่นเอง หลังจากแปลความหมายของประโยคออกมาแล้ว ระบบ SNOWY จะทำการสกัดความรู้ออกมาให้อยู่ในรูปของกรอบความรู้ โดยภาพที่ 15 แสดงตัวอย่างของกรอบความรู้ (Knowledge Frame) ที่ระบบ SNOWY สร้างขึ้นมา

```
CROWNED-EAGLE
(is-a (eagle))

RAIN-FOREST
(is-a (forest))
(related-to (@a239))

@X235
(cf (is-a (crowned-eagle)) (@a235))
(location-r (africa ($more (@a237))))
(ingest (monkey ($more (@a237))))
(inhabit ($null ($more (@a239))))

@A235
(instance-of (description))
(args (@x235) (africa))
(pr (location-r))
(descr-subj (@x235 (q (all))))
(descr-obj (africa (q (constant))))

AFRICA
(location-of (@x235 ($more (@a236))))

MONKEY
(ingest%by (@x235 ($more (@a237))))

@A237
(args (@x235) (monkey))
(pr (ingest))
(actor (@x235 (q (all))))
(theme (monkey (q (?))))
(instance-of (action))

@A239
(args (@x235) (rain-forest))
(pr (inhabit))
(actor (@x235 (q (all))))
(at-loc (rain-forest (q (?))))
(instance-of (action))
```

ภาพที่ 15 ตัวอย่างกรอบความรู้ที่สร้างจากระบบ SNOWY (Gomez et al, 1994)

จากภาพที่ 15 จะแสดงถึงกรอบความรู้ที่สร้างขึ้นจากตัวแทนทางความหมายเช่น กรอบ CROWNED-EAGLE จะมีสล๊อต (is-a (eagle)) หมายถึง crowned-eagle มีความสัมพันธ์แบบ is-a กับ eagle และกรอบ AFRICA จะมีสล๊อต (location-of (@x235 ($more (@a236)))) จะหมายถึง Africa มีความสัมพันธ์กับกรอบ @a235 แบบ location-of ซึ่งก็หมายถึง crowned-eagle อาศัยอยู่ที่ Africa นั่นเอง การสกัดความรู้ด้วยวิธีเทียบแม่แบบ จะเหมาะกับการเก็บความรู้จากเอกสารที่เป็นโดเมนเฉพาะ ซึ่งโครงสร้างของความรู้นั้น จะต้องมีการออกแบบโครงสร้างซึ่งในกรณีนี้ก็คือโครงสร้างของกรอบความรู้ไว้ล่วงหน้าก่อนแล้ว ระบบสกัดความรู้แบบเทียบแม่แบบนั้น จะทำการค้นหาองค์ประกอบต่าง ๆ ที่ระบุไว้ในโครงสร้างของความรู้นั้นจากเอกสารแล้วนำมาเก็บไว้ในฐานความรู้ ซึ่งฐานความรู้นี้จะสามารถนำไปใช้ในระบบสืบค้นหรือใช้แก้ปัญหาแบบเฉพาะทางได้ต่อไป

3. การทำเหมืองเอกสาร (Text Data Mining)

การทำเหมืองเอกสารเป็นการหากฎความสัมพันธ์ (Association Rules) ด้วยเทคนิคการทำเหมืองข้อมูล หรือData Mining (Agrawal *et al.*, 1995) จากสารสนเทศในเอกสารจำนวนมากเพื่อค้นหาแนวโน้ม หรือรูปแบบ (Pattern) ที่น่าสนใจออกมา โดยความรู้ที่ค้นพบออกมาจะเป็นความรู้ที่ไม่เน้นการอนุมานมาก โดยกระบวนการทำเหมืองเอกสารจะมี 2 ขั้นตอน คือ ขั้นแรกสกัดสารสนเทศ หรือกรอบความคิด (Concept) ออกมาจากเอกสารมาเก็บไว้ในรูปแบบของฐานข้อมูลก่อน ขั้นที่สองก็จะนำเอาเทคนิคการทำเหมืองข้อมูลมาทำการค้นหากฎความสัมพันธ์ (Association rules) หรือรูปแบบ (Pattern) ที่น่าสนใจออกมา ในปี 2000 นั้น Loh *et al.* (2000)ได้นำเอกสารเวบเพจ (Web pages) มาทำการสกัดกรอบความคิดที่ปรากฏออกมา ในแต่ละเอกสาร แล้วจึงใช้เทคนิคของการทำเหมืองข้อมูล เพื่อค้นหากฎความสัมพันธ์ ของคอนเซ็ปต์ที่เกิดขึ้นในแต่ละเอกสาร ซึ่งการสกัดคอนเซ็ปต์ในเอกสาร จะมีการกำหนดกลุ่มของคำที่เกี่ยวพันธ์กับคอนเซ็ปต์นั้น ๆ เพื่อใช้ในการกำหนดคอนเซ็ปต์ที่ปรากฏในเอกสารออกมาดังภาพที่ 16

"crimes" = { crime, crimes, fraud, fraudulent, illegal, ...}
"elections" = { election, elections, term, voter, reelection, elected, electorate, ...}

ภาพที่ 16 การกำหนดกลุ่มคำที่เกี่ยวข้องกับคอนเซ็ปต์ที่กำหนด

เมื่อสกัดคอนเซ็ปต์ที่ปรากฏในเอกสารออกมาแล้ว จึงใช้เทคนิคของการทำเหมืองข้อมูลในการค้นหากฎความสัมพันธ์ (Association Rules) ของคอนเซ็ปต์ออกมาดังตัวอย่างในภาพที่ 17

loans => politicians (Confidence = 82.1%, support = 23 documents)
loans AND politicians => education (Confidence = 17.2%)

ภาพที่ 17 ตัวอย่างกฎความสัมพันธ์ของคอนเซ็ปต์ที่ได้จากเหมืองเอกสาร

Pierre (2002) ได้เสนอการค้นพบความรู้ในเอกสารด้วยการสร้าง Meta Data ซึ่งจะทำการกำหนดกลุ่มของ Meta Data ขึ้นมาเพื่อเป็นการกำหนดกลุ่มของคอนเซ็ปต์ ที่ปรากฏในเอกสารต่าง ๆ แล้วจึงทำการค้นหาความรู้ด้วยเทคนิคการทำเหมืองข้อมูล ในการค้นหากฎความสัมพันธ์ที่

ปรากฏระหว่าง Meta Data ต่าง ๆ กลุ่มของ Meta Data ในที่ Pierre กำหนดไว้ จะใช้ในโดเมนของผลิตภัณฑ์สินค้าต่าง ๆ เป็น 4 กลุ่มดังตัวอย่างในภาพที่ 18

Metadata Facet	Number of Concepts
Category	11
Subcategory	49
Products	1610
Rating	2

ภาพที่ 18 กลุ่มของ Meta Data

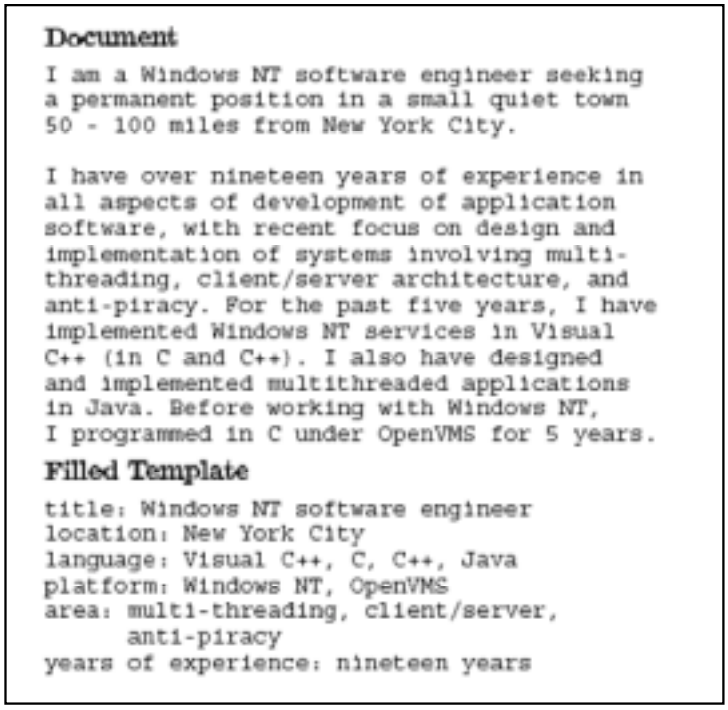
จากภาพที่ 18 จะแบ่งกลุ่มข้อมูลเป็น 4 กลุ่ม คือ Category, Subcategory, Products และ Rating ซึ่งในกลุ่ม Category จะมีการกำหนดไว้ทั้งหมด 11 Category ในกลุ่ม Subcategory มี 46 Subcategory ในกลุ่ม Products มี 1610 Products และในกลุ่ม Raing มี 2 Rating โดยการกำหนดสมาชิกในกลุ่มทั้งหมดจะเป็นการกำหนดไว้ล่วงหน้า (Predefined) เพื่อใช้ในการกำหนดกลุ่มของคอนเซ็ปต์ที่จะสกัดในเอกสาร หลังจากสกัดกลุ่มของคอนเซ็ปต์ ที่ปรากฏในเอกสารแล้ว จึงทำการค้นหากฎความสัมพันธ์ออกมา ด้วยเทคนิคของการทำเหมืองข้อมูล ออกมาเป็นกฎความสัมพันธ์ดังตัวอย่างที่แสดงในภาพที่ 19

Rule Type	Example
Subcategory(A) → Category(B):	
	A/V Receivers → Amplification
	DVD Players → Home Video
	Main Speaker → Speakers
Product(A) → Category(B):	
	Yamaha RX/V795 → Amplification
	Samsung DVD/611 → Home Video
	Paradigm Atom → Speakers
Product(A) → Subcategory(B):	
	Yamaha RX/V795 → A/V Receivers
	Samsung DVD/611 → DVD Players
	Paradigm Atom → Main Speaker
Product(A) → Rating(B):	
	Yamaha RX/V795 → GOOD
	Samsung DVD/611 → BAD
	Paradigm Atom → GOOD

ภาพที่ 19 ตัวอย่างกฎความสัมพันธ์ที่ได้

จากภาพที่ 19 จะเป็นตัวอย่างของกฎความสัมพันธ์ระหว่างคอนเซ็ปต์ เช่น A/V Receivers → Amplification จะหมายถึง Subcategory A/V Receivers จะอยู่ใน Category Amplification เป็นต้น

Nahm and Mooney (2002) ได้เสนอการค้นหาความรู้ จากเอกสารด้วยการใช้การสกัดสารสนเทศในเอกสาร(Information Extraction) ออกมาเก็บในรูปแบบของกรอบข้อมูล (Frame) แล้วจากนั้นจึงใช้เทคนิคการทำเหมืองข้อมูล เพื่อหากฎความสัมพันธ์ที่เกิดขึ้นในสล๊อต (Slot) ของกรอบสารสนเทศ (Information Frame) ซึ่งตัวอย่างของการสกัดสารสนเทศในเอกสาร แสดงไว้ในภาพที่ 20



Document

I am a Windows NT software engineer seeking
a permanent position in a small quiet town
50 - 100 miles from New York City.

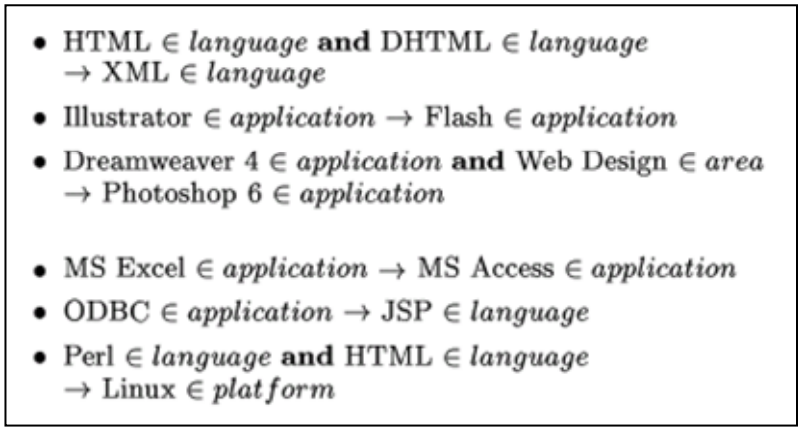
I have over nineteen years of experience in
all aspects of development of application
software, with recent focus on design and
implementation of systems involving multi-
threading, client/server architecture, and
anti-piracy. For the past five years, I have
implemented Windows NT services in Visual
C++ (in C and C++). I also have designed
and implemented multithreaded applications
in Java. Before working with Windows NT,
I programmed in C under OpenVMS for 5 years.

Filled Template

title: Windows NT software engineer
location: New York City
language: Visual C++, C, C++, Java
platform: Windows NT, OpenVMS
area: multi-threading, client/server,
anti-piracy
years of experience: nineteen years

ภาพที่ 20 ตัวอย่างกรอบสารสนเทศที่สกัดจากเอกสาร

จากภาพที่ 20 สารสนเทศในเอกสารจะถูกสกัดสารสนเทศออกมาเก็บไว้ในกรอบสารสนเทศ โดยเก็บในสล๊อต (Slot) ต่าง ๆของกรอบสารสนเทศ เช่น สล๊อต title มีสารสนเทศคือ Windows NT software engineer, สล๊อต location มีสารสนเทศคือ New York City เป็นต้น หลังจากนั้นจึงทำการ ค้นหากฎความสัมพันธ์ที่ปรากฏขึ้นในแต่ละสล๊อตของกรอบข้อมูลออกมา ดังแสดงในภาพที่ 21



ภาพที่ 21 ตัวอย่างกฎความสัมพันธ์

อย่างไรก็ตามความรู้ที่ได้มาจากเทคนิคการทำเหมืองเอกสารนั้น จะเป็นความรู้ที่ใช้ในการค้นหาแนวโน้ม หรือความสัมพันธ์ระหว่างสารสนเทศในเอกสาร ซึ่งจะเป็นความรู้ที่ไม่เน้นการใช้เหตุผล (Reasoning) หรือการอนุมาน (Inference) เพราะฉะนั้นความรู้ที่ค้นพบด้วยวิธีนี้ จึงเหมาะกับงานประเภททำนายแนวโน้มต่าง ๆ มากกว่าจะนำมาสร้างเป็นระบบผู้เชี่ยวชาญ (Expert System) หรือระบบถาม-ตอบ (Question-Answering System)

จากการตรวจเอกสารที่เกี่ยวข้อง เราสามารถสรุปข้อดีข้อเสียของวิธีการค้นหาความรู้จากเอกสารในแต่ละวิธี ดังตารางที่ 5

ตารางที่ 5 สรุปการตรวจเอกสาร

เทคนิค	ผู้วิจัย	ข้อมูลนำเข้า	ข้อมูลผลลัพธ์	ข้อดี	ข้อเสีย
เทคนิคการเรียนรู้	Dalannoy *et al*., 1993	English Technical Texts	Horn Clause Rules	- สามารถใช้ในโดเมนเปิด (Open Domain)	- อาจจะเกิดกฎที่ผิดพลาดได้ถ้าข้อมูลไม่สมบูรณ์
	Faure and Nedellec, 1999	Technical Texts	Ontological Kowledge		
การเทียบแม่แบบ	Moulin and Rousseau, 1992	Technical Texts	Rules	- มีความเที่ยงตรงในการดึงความรู้สูง - เหมาะกับงานเฉพาะทาง	- ต้องออกแบบโครงสร้างความรู้ไว้ล่วงหน้า - ต้องการความรู้สนับสนุนจำนวนมาก
	Gomez *et al*., 1994	Encyclopedia	Frame-like		
การทำเหมืองเอกสาร	Loh *et al.*, 2000	Web Pages	Association Rules	- เหมาะกับงานที่ใช้ค้นหาแนวโน้มหรือ รูปแบบที่ไม่เน้นด้านความหมาย	- ต้องการเอกสารในการทำเหมืองจำนวนมาก
	Pierre, 2002	Web Pages			
	Nahm *et al*., 2002	USENET newsgroup			

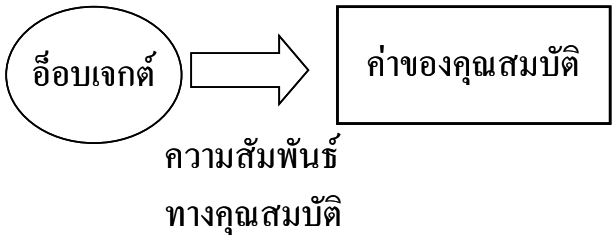
อุปกรณ์และวิธีการ

อุปกรณ์

1. เครื่องคอมพิวเตอร์พีซี 1 เครื่อง : ซีพียู Pentium IV 1.6 GHz, หน่วยความจำขนาด 256 MB, ฮาร์ดดิกส์ขนาด 40 GB
2. ระบบปฏิบัติการ Windows XP Professional
3. ชุดพัฒนาโปรแกรมภาษา JAVA 1 ชุด
4. ฐานข้อมูลคำ WordNet
5. คลังเอกสารในโดเมนการเกษตรขนาด 2,000 ประโยค สำหรับการทดลองและประเมินผล โดยผ่านการประมวลผลเบื้องต้น อันได้แก่ ตัดคำ, กำกับ POS และ วลี เป็นต้น

ปัญหาที่เกี่ยวข้องในการสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์

โครงสร้างของความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์นั้นมีองค์ประกอบหลัก อยู่ 3 องค์ประกอบดังนี้ คือ อ็อบเจกต์ (Object), ความสัมพันธ์ทางคุณสมบัติ (Property Relation) และ ค่าของคุณสมบัติ (Property Value) ดังนั้นโครงสร้างพื้นฐานของความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์สามารถแสดงได้ดังภาพที่ 22



ภาพที่ 22 โครงสร้างของความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์

ในส่วนของค่าของคุณสมบัติสามารถแบ่งได้เป็น 2 ชนิด คือ ค่าแบบตัวเลข (Numerical Value) และค่าแบบสัญลักษณ์ (Symbolic Value) ซึ่งโครงสร้างของค่าของคุณสมบัตินั้น ก็มี

องค์ประกอบที่แตกต่างกันดังนี้คือ ค่าแบบตัวเลขจะมีองค์ประกอบ 3 องค์ประกอบคือ ส่วนขยาย (Quantifier), ตัวเลข (Number) และหน่วยวัด (Measurement) ส่วนค่าแบบสัญลักษณ์นั้น ก็มีองค์ประกอบ 3 องค์ประกอบดังนี้คือ ส่วนขยาย (Quantifier), สัญลักษณ์ค่า (Symbol) และส่วนควบคุม (Operator) โดยโครงสร้างของความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์นั้นสามารถเขียนอธิบายได้ดังนี้

OPKB = (Ob, P, V)

P = Pnum $\cup$ Psym

Pnum = {height, length, width, diameter, perimeter, weight}

Psym = {color, taste, odor}

V = Vnum $\cup$ Vsym

Vnum = ([Qnum], M, N)

Vsym = ([Qsym], [Op], S)

*เครื่องหมาย “[]” หมายถึงสมาชิกในวงเล็บนั้นอาจจะมีหรือไม่มีก็ได้

โดยที่ :

OPKB คือ ความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์

Ob คือ อ็อบเจกต์

P คือ ความสัมพันธ์ทางคุณสมบัติ

Pnum คือ ความสัมพันธ์ทางคุณสมบัติแบบตัวเลข

Psym คือ ความสัมพันธ์ทางคุณสมบัติแบบสัญลักษณ์

V คือ ค่าของคุณสมบัติ

Vnum คือ ค่าของคุณสมบัติแบบตัวเลข

Vsym คือ ค่าของคุณสมบัติแบบสัญลักษณ์

Qnum คือ ส่วนขยายค่าของค่าแบบตัวเลข เช่น ประมาณ, มากกว่า, น้อยกว่า เป็นต้น

Qsym คือ ส่วนขยายค่าของค่าแบบสัญลักษณ์ เช่น เข้ม, อ่อน เป็นต้น

M คือ หน่วยวัด เช่น นิ้ว, เมตร, เซนติเมตร เป็นต้น

Op คือ ส่วนควบคุมค่าของค่าแบบสัญลักษณ์ เช่น ปน, อม, แกม เป็นต้น

N คือ ตัวเลข

S คือ สัญลักษณ์ค่า

ตัวอย่างของค่าคุณสมบัติเราสามารถเขียนอธิบายได้ ดังเช่น “ประมาณ 10 นิ้ว” จะอธิบายได้เป็น (“ประมาณ”, 10, “นิ้ว”) ,“แดงเข้ม” จะอธิบายได้เป็น (“เข้ม”, ϕ, “แดง”) เป็นต้น

ในระบบสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์นั้น มีปัญหาที่เกี่ยวข้องในการระบุถึงองค์ประกอบของความรู้ที่ปรากฏในแต่ละประโยคของเอกสารทั้งหมด 3 ปัญหาคือ ปัญหาในการระบุถึงอ็อบเจกต์, ปัญหาในการระบุถึงความสัมพันธ์ทางคุณสมบัติและปัญหาในการระบุถึงค่าของคุณสมบัติดังนี้

1. ปัญหาในการระบุถึงอ็อบเจกต์

ในภาษาไทยนั้นมีพฤติกรรมทางภาษาที่เกี่ยวข้องกับปัญหาในการระบุถึงอ็อบเจกต์ในประโยคนั้นมี 2 ชนิดคือ สิ่งอ้างอิงร่วม (Anaphora) และการละข้อความ (Textual Ellipsis) โดยการใช้สิ่งอ้างอิงร่วมนั้น การใช้สิ่งอ้างอิงร่วมแบบสูญรูป (Zero Anaphora) มีผลต่อการระบุอ็อบเจกต์ในประโยคค่อนข้างมาก ดังนี้

1.1 สิ่งอ้างอิงร่วมแบบสูญรูป (Zero Anaphora)

ในประโยคของภาษาไทย นามวลีที่ได้เคยกล่าวถึงมาก่อนแล้วในประโยคก่อนหน้านั้น ถ้าหากได้มีการกล่าวถึงนามวลีนั้นอีก เรามักจะเปลี่ยนนามวลีนั้นให้อยู่ในรูปของสิ่งอ้างอิงร่วม (Anaphora) โดยตารางที่ 6 แสดงถึงจำนวนประโยคที่มีการใช้สิ่งอ้างอิงร่วมชนิดต่าง ๆ ในเอกสารโดเมนการเกษตรขนาด 435 ประโยค

<u>ตารางที่ 6</u> จำนวนประโยคที่มีการใช้สิ่งอ้างอิงร่วมชนิดต่าง ๆ ในเอกสารโดเมนการเกษตรขนาด 435 ประโยค

ชนิดของสิ่งอ้างอิงร่วม	ปริมาณการปรากฏ
Pronominal Anaphora	0.45 %
Nominal Anaphora	9.20 %
Zero Anaphora	21.38 %

จากตารางที่ 6 แสดงให้เห็นว่าในเอกสารโดเมนการเกษตรนั้น มีการใช้สิ่งอ้างอิงร่วมแบบสูญรูปมากถึง 21.38 % ฉะนั้นการใช้สิ่งอ้างอิงร่วมแบบสูญรูปนั้นจึงมีผลต่อการสกัดความรู้เกี่ยวกับอ็อบเจกต์มาก ซึ่งสิ่งอ้างอิงรูปแบบสูญรูปคือ การละนามวลีออกจากประโยคในการอ้างอิงถึงนามวลีที่ได้เคยกล่าวมาแล้วในประโยคอื่นก่อนหน้าดังตัวอย่างข้างล่างนี้

(1) เพลี้ยไฟเป็นแมลงขนาดเล็ก
(2) ∅ มีสีเหลืองหรือสีน้ำตาลอ่อน

จากตัวอย่างในประโยคที่ 2 นั้น ประธานจะถูกละไว้ที่ตำแหน่งเครื่องหมาย ∅ ซึ่งประธานที่ละไว้คือ นามวลี “เพลี้ยไฟ” ผลกระทบของการใช้สิ่งอ้างอิงร่วมแบบสูญรูปที่มีต่อระบบสกัดความรู้คือ ทำให้อ็อบเจกต์ที่อยู่ในโครงสร้างของความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์นั้นสูญหายไป ดังประโยคที่ 2 จะเห็นว่าอ็อบเจกต์ที่จะมีสีเหลืองหรือสีน้ำตาลอ่อนนั้นก็คือ “เพลี้ยไฟ” ฉะนั้นถ้าไม่มีการแก้ปัญหาการใช้สิ่งอ้างอิงร่วมแบบสูญรูป ก็จะมีผลทำให้ไม่สามารถสกัดความรู้ออกมาจากประโยคนี้ได้

1.2 การละข้อความ (Textual Ellipsis)

ในภาษาธรรมชาตินั้น ประโยคหนึ่ง ๆ มักจะมีข้อความบางส่วนของประโยคที่ละไว้เนื่องจากได้กล่าวไว้แล้วในประโยคก่อนหน้าหรือ เป็นความรู้ที่เข้าใจกันเองระหว่างผู้สื่อสาร ซึ่งข้อความบางส่วนที่ละไว้นี่เองทำให้ ประโยคในภาษาธรรมชาติจะมีสารสนเทศทางความหมายที่ไม่สมบูรณ์ (Incomplete Semantic)ได้ ซึ่งความไม่สมบูรณ์ทางความหมายอาจจะเกิดจากการละข้อความในหลาย ๆ แบบ ซึ่งในระบบสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์นั้น จะเกิดกับการละบุพบทที่แสดงความเป็นเจ้าของในนามวลีดังตัวอย่างต่อไปนี้

(1) เพลี้ยไฟเป็นแมลงขนาดเล็ก
(2) ไข่ (ของเพลี้ยไฟ) มีสีขาว

จากตัวอย่าง ในประโยคที่ 2 นั้นจะเห็นว่านามวลี “ไข่” นั้น เป็นนามวลีที่มีความไม่สมบูรณ์ทางความหมาย เพราะว่านามวลี “ไข่” นั้น เป็นนามวลีที่มีความหมายเป็นส่วนประกอบหรือนามวลีที่จะต้องมีบุพบทที่แสดงความเป็นเจ้าของ ซึ่งบุพบทแสดงความเป็นเจ้าของที่ได้ถูกละไว้คือ “ของ

เพลี้ยไฟ” ไม่ใช่ “ไข่” ของสิ่งมีชีวิตชนิดอื่นที่มีสีขาว ซึ่งจากการละไว้ดังกล่าวนั้น ผู้อ่านจะเข้าใจได้เองว่าบุพบทที่ถูกละไว้คือ “ของเพลี้ยไฟ” เนื่องจากได้มีการกล่าวถึง “เพลี้ยไฟ” มาก่อนในประโยคที่ 1 ผู้อ่านจึงเข้าในว่าสิ่งจะเป็นบุพบทแสดงความเป็นเจ้าของของนามวลี “ไข่” ในประโยคที่ 2 นั้นเป็น “ของเพลี้ยไฟ” ฉะนั้นถ้าหากไม่มีการแก้ปัญหาการละข้อความแล้วจะทำให้ความรู้ที่สกัดมาจากเอกสารนั้น มีความไม่ถูกต้อง หรือไม่สมบูรณ์ได้

2. ปัญหาในการระบุความสัมพันธ์ทางคุณสมบัติ

การระบุความสัมพันธ์ทางคุณสมบัติที่ปรากฏในประโยคนั้น จะมีปัญหาที่เกี่ยวข้อง 2 ปัญหา คือ ปัญหาการแสดงความสัมพันธ์ทางคุณสมบัติแบบโดยตรงและแบบโดยนัย และปัญหาความกำกวมในการแสดงความสัมพันธ์ทางคุณสมบัติแบบโดยตรง

2.1 การแสดงความสัมพันธ์ทางคุณสมบัติแบบโดยตรงและโดยนัย

ในประโยคภาษาไทยที่สามารถสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์ได้นั้น จะมีคำนามหรือคำกิริยาที่สามารถระบุถึงชนิดของความสัมพันธ์ทางคุณสมบัติได้โดยตรง ดังตัวอย่างข้างล่างนี้

ดอกกุหลาบมี<u>สี</u>แดง

จากประโยคข้างบนจะเห็นว่ามีคำนาม “สี” ซึ่งเป็นคำนามที่จะระบุถึงชนิดของความสัมพันธ์ทางคุณสมบัติแบบสี (Color) อย่างชัดเจน แต่ก็มีกรณีที่มีการระบุชนิดของความสัมพันธ์ทางคุณสมบัติแบบโดยนัย ดังตัวอย่างข้างล่างนี้

ลูกมะม่วง**<u>เปรี้ยว</u>**มาก

จากประโยคข้างบน เราสามารถระบุชนิดความสัมพันธ์ทางคุณสมบัติแบบรส (Taste) ได้จากคำ “เปรี้ยว” ซึ่งเป็นค่าของคุณสมบัติแบบสัญลักษณ์ของความสัมพันธ์ทางคุณสมบัติ นอกจากนี้โครงสร้างของค่าของคุณสมบัติก็สามารถใช้ในการระบุชนิดของคุณสมบัติทางคุณสมบัติได้ด้วยเช่นกัน ดังตัวอย่างข้างล่างนี้

โต๊ะขนาด **120x60x75** เซนติเมตร

จากประโยคตัวอย่างข้างต้น ค่าของคุณสมบัติ “120x60x75” นั้น สามารถระบุถึงชนิดของความสัมพันธ์ทางคุณสมบัติโดยนัยถึง 3 คุณสมบัติ คือ ความกว้าง (Width), ความยาว (Length) และความสูง (Height) ซึ่งจากโครงสร้างของค่าคุณสมบัตินี้ เราได้ว่าค่าแรก “120” จะเป็นค่าของความสัมพันธ์ทางคุณสมบัติแบบความกว้าง ค่าที่สอง “60” จะเป็นค่าของความสัมพันธ์ทางคุณสมบัติแบบความยาว และค่าที่สาม “75” จะเป็นค่าของความสัมพันธ์ทางคุณสมบัติแบบความสูงนั่นเอง

2.2 ความไม่ชัดเจนในการแสดงความสัมพันธ์ทางคุณสมบัติแบบโดยตรง

คำนามที่สามารถใช้ระบุชนิดความสัมพันธ์ของคุณสมบัตินั้น มีบางคำที่มีความหมายไม่ชัดเจนในการระบุชนิดของความสัมพันธ์ทางคุณสมบัติ ดังตัวอย่างข้างล่างนี้

(1) ปลาคาร์ฟ**ขนาด** 1 กิโลกรัม
(2) ต้นปาล์ม**ขนาด** 3 เมตร

จากตัวอย่างประโยคข้างต้นจะเห็นว่าคำนาม “ขนาด” จะมีความหมายไม่ชัดเจนในการระบุชนิดความสัมพันธ์ทางคุณสมบัติ ประโยคที่ 1 จะระบุถึงความสัมพันธ์ทางคุณสมบัติแบบน้ำหนัก (Weight) แต่ในประโยคที่ 2 นั้นจะระบุถึงความสัมพันธ์ทางคุณสมบัติแบบความยาว (Length) ซึ่งในกรณีแบบนี้ หน่วยวัดจะสามารถเป็นตัวช่วยในการระบุถึงชนิดความสัมพันธ์ทางคุณสมบัติได้

3. ปัญหาในการอธิบายค่าของคุณสมบัติ

ค่าของคุณสมบัติที่ปรากฏในแต่ละประโยคของเอกสารนั้น จะมีรูปแบบที่ใช้ในการแสดงลักษณะของค่าของคุณสมบัติหลายรูปแบบ เนื่องจากค่าของคุณสมบัตินั้น มีหลายองค์ประกอบ อย่างเช่น ส่วนขยายค่า “ประมาณ”, “ไม่น้อยกว่า” และ ยังมีหน่วยวัด “นิ้ว”, “เมตร” เหล่านี้ประกอบอยู่ในค่าของคุณสมบัติแบบตัวเลข และในค่าคุณสมบัติแบบสัญลักษณ์ก็มีส่วนขยายค่า “มาก”, “เข้ม” ประกอบอยู่ด้วย นอกจากนี้แล้วยังมีส่วนควบคุมค่าที่จะทำให้รวมค่าของคุณสมบัติแบบสัญลักษณ์สองค่าเข้าด้วยกันอย่าง “ปน”, “อม”, “แกม” เช่น เขียวอมเหลือง เป็นต้น

ปัญหาที่เกี่ยวข้องในการสืบค้นอ็อบเจกต์

ปัญหาในระบบการสืบค้นอ็อบเจกต์ด้วยคุณสมบัตินั้นคือ จะทำอย่างไรให้สามารถสืบค้นอ็อบเจกต์ด้วยคุณสมบัติได้ แม้จะมีความคล้ายคลึงกันของค่าของคุณสมบัติที่ใช้ในการสืบค้นกับค่าของคุณสมบัติของอ็อบเจกต์ในฐานความรู้นั้น ผู้ใช้ที่ทำการสืบค้นอ็อบเจกต์นั้นจะต้องอธิบายค่าของคุณสมบัติที่ใช้ในการสืบค้นอ็อบเจกต์แก่ระบบ ซึ่งมีค่าของคุณสมบัตินั้นอาจจะมีความคล้ายคลึงกันอย่างเช่น สีแดง, สีแดงเข้ม และ สีส้ม เป็นต้น จึงมีความเป็นไปได้ผู้ใช้อาจจะใช้ค่าคุณสมบัติที่ไม่ตรงกับค่าคุณสมบัติของอ็อบเจกต์โดยตรง (Exactly Match) ในการสืบค้นอ็อบเจกต์แต่เป็นค่าที่ใกล้เคียงกันแทน

ดังนั้นในการสืบค้นอ็อบเจกต์ด้วยคุณสมบัตินั้น จะต้องมีการแทนค่าของคุณสมบัติที่ดีที่สามารถอธิบายความใกล้เคียงกันของค่าของคุณสมบัติได้ และสูตรคำนวณที่ใช้ในการคำนวณความคล้ายกันของค่าของคุณสมบัตินั้นจะต้องสามารถใช้การแทนค่านั้นในการคำนวณความคล้ายของค่าคุณสมบัติได้

หลักการและเหตุผล

ในการค้นพบความรู้จากเอกสาร (Knowledge Discovery from Texts) นั้นมีเทคนิคในการค้นพบอยู่ 3 วิธีคือ การใช้เทคนิคการเรียนรู้ (Machine Learning), การใช้การเทียบแม่แบบ (Template Matching) และ การทำเหมืองเอกสาร (Texts Data Mining) วิธีการค้นพบความรู้จากเอกสารในแต่ละวิธีนั้นก็จะมีความเหมาะสมกับงานที่จะต้องใช้แตกต่างกัน โดยการค้นพบความรู้ด้วยเทคนิคการเรียนรู้นั้นจะเหมาะกับเอกสารที่เป็นโดเมนเปิด (Open Domain) ซึ่งความรู้ที่ได้จากระบบนี้มักจะออกมาในรูปของกฎต่าง ๆ ที่เกี่ยวข้องกับความสัมพันธ์ของอาร์กิวเมนต์ (Argument) ของประโยคที่ถูกแปลให้อยู่ในรูปแบบของตรรกะ (logic) แล้ว ซึ่งกฎเหล่านี้ระบบสามารถจะนำมาอนุมาน (Inference) เพื่อสร้างคำตอบให้กับคำถามบางอย่างได้ แต่ในงานเฉพาะทาง (Specific Domain) บางอย่าง เช่น การสืบค้นอ็อบเจกต์ด้วยคุณสมบัติ ตัวโครงสร้างของความรู้เกี่ยวกับคุณสมบัติของ อ็อบเจกต์จะมีองค์ประกอบที่ซับซ้อนและเป็นความรู้ที่ชัดเจน (Explicit Knowledge) เทคนิคการเรียนรู้จึงไม่ใช่วิธีการที่จะนำมาใช้วิเคราะห์รูปประโยคเพื่อสกัดองค์ความรู้เกี่ยวกับคุณสมบัติ การค้นพบความรู้โดยการทำเหมืองเอกสารนั้น จะเหมาะกับการค้นพบความสัมพันธ์ของ

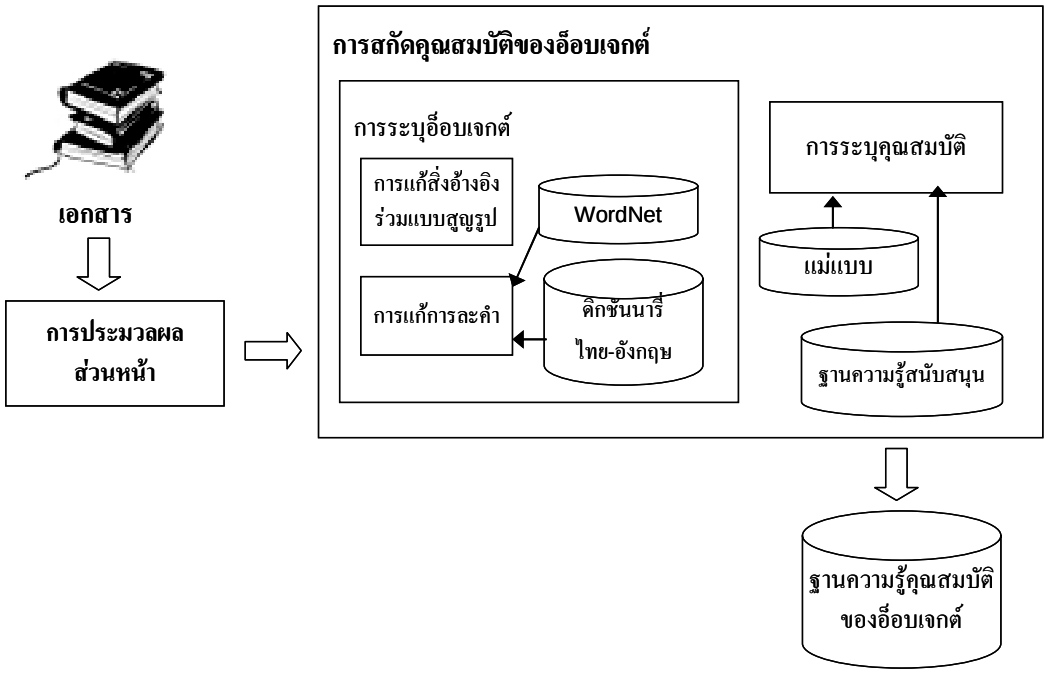
สารสนเทศที่มีความสัมพันธ์กันในเอกสาร ซึ่งความรู้ที่ได้มักจะอยู่ในรูปแบบของแนวโน้ม (Trend), กฎความสัมพันธ์ (Association Rules) หรือ รูปแบบที่มีนัยยะสำคัญ (Pattern) ซึ่งความรู้ประเภทนี้จะไม่สามารถนำมาสร้างเป็นฐานความรู้สำหรับระบบอิงฐานความรู้ (Knowledge Based System) เช่น ระบบผู้เชี่ยวชาญ (Expert System) และระบบช่วยตัดสินใจ (Decision Support System) เป็นต้น ฉะนั้นการทำเหมืองเอกสารจึงไม่เหมาะกับระบบสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์ การค้นพบความรู้ด้วยการเทียบแม่แบบ จะเป็นการสร้างแม่แบบ (Template) ขึ้นมาเพื่อใช้ในการสกัดสารสนเทศหรือองค์ประกอบของต่าง ๆ ในโครงสร้างความรู้ที่ได้ออกแบบไว้ล่วงหน้าแล้ว แล้วนำมาสร้างองค์ความรู้ขึ้นมาแล้วจึงนำมาจัดเก็บไว้ในฐานความรู้ การค้นพบความรู้ด้วยการเทียบแม่แบบซึ่งเหมาะกับงานที่มีการออกแบบโครงสร้างของความรู้ไว้ล่วงหน้าแล้วจึงเหมาะกับการสร้างฐานความรู้สำหรับงานเฉพาะทาง ซึ่งวิทยานิพนธ์นี้ก็ได้เลือกใช้เทคนิคการสกัดความรู้ด้วยการเทียบแม่แบบมาใช้ในระบบสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์

ในระบบสืบค้นอ็อบเจกต์ด้วยคุณสมบัตินั้น จะต้องมีการเลือกตัวแทน (Representation) ของค่าของคุณสมบัติที่เหมาะสม เนื่องจากค่าของคุณสมบัตินั้นมี 2 แบบ คือ แบบตัวเลข และแบบสัญลักษณ์ และนอกจากนี้ในส่วนขยายค่าของค่าของคุณสมบัติทั้งสองแบบมีความเป็นนามธรรม (Subjective) มาก ตัวอย่างเช่น “ประมาณ 10 เซนติเมตร” , “สีเขียวเข้ม” เป็นต้น ตัวแทนที่เหมาะสำหรับค่าที่มีลักษณะเป็นนามธรรมจึงมีความจำเป็น ในทฤษฎีฟัซซี่ (Fuzzy Theory) ได้กล่าวถึงการใช้ฟังก์ชั่นของความเป็นสมาชิก (Membership Function) ในการแสดงถึงความเป็นสมาชิกของค่าใด ๆ กับเซ็ทของค่าที่อยู่ในรูปของนามธรรมมากกว่าที่จะอยู่ในรูปแบบที่เป็นตัวเลข ฉะนั้นในวิทยานิพนธ์นี้จึงได้เลือกการใช้ฟังก์ชั่นของความเป็นสมาชิกในทฤษฎีฟัซซี่ มาใช้เป็นตัวแทนของค่าของคุณสมบัติในระบบการสืบค้นอ็อบเจกต์ด้วยคุณสมบัติ

ภาพรวมของระบบสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์

ในระบบสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารนั้น มีกระบวนการที่สำคัญ 3 ขั้นตอนคือ การประมวลผลส่วนหน้า (Preprocessing), การระบุอ็อบเจกต์ (Object Identifier) และการระบุคุณสมบัติ (Property Identifier) ซึ่งขั้นตอนการสกัดจะเริ่มจากเอกสารจะผ่านการประมวลผลส่วนหน้าเพื่อวิเคราะห์คำและวิเคราะห์วลีต่าง ๆ ของประโยคในเอกสาร จากนั้นโปรแกรมในการระบุอ็อบเจกต์จะเป็นโปรแกรมที่ทำการแก้ปัญหาที่เกี่ยวกับพฤติกรรมทางภาษาเช่น การใช้สิ่งอ้างอิงร่วมแบบสูญรูป และการละข้อความ จากนั้นเอกสารก็จะประมวลผล

ต่อไปยังโปรแกรมการระบุคุณสมบัติ ซึ่งในส่วนนี้จะมีการใช้เทคนิคการเทียบแม่แบบ เพื่อทำการระบุถึงองค์ประกอบของความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์ในประโยค แล้วจึงทำการสร้างองค์ความรู้ขึ้นมาตามโครงสร้างของความรู้ที่ได้ออกแบบไว้ แล้วจึงจัดเก็บในฐานความรู้ โดยจะจัดเก็บให้อยู่ในรูปแบบของตรรกะศาสตร์ภาคแสดง (Predicate Logic) โดยภาพที่ 23 แสดงถึงภาพรวมของระบบสกัดความรู้เกี่ยวกับคุณสมบัติของวัตถุ



<u>ภาพที่ 23</u> ภาพรวมของระบบสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์

1. การประมวลผลส่วนหน้า (Preprocessing)

การประมวลผลส่วนหน้าจะเป็นการประมวลเอกสารในระดับคำและวากยสัมพันธ์เพื่อใช้ในการวิเคราะห์พฤติกรรมทางภาษาต่อไป โดยในการประมวลผลส่วนนั้นมีส่วนประกอบดังนี้

(ก) การตัดคำ (Word Segmentation) ลักษณะของคำในประโยคของภาษาไทยนั้นจะไม่มีเครื่องหมายหรือช่องว่างใด ๆ เป็นตัวแบ่งคั่นระหว่างคำอย่างเช่นภาษาอังกฤษ ก่อนที่จะมีการประมวลผลใด ๆ ในภาษาไทยนั้น กระบวนการแรกต้องมีการวิเคราะห์คำด้วยโปรแกรมตัดคำก่อน โดยในวิทยานิพนธ์นี้จะใช้โปรแกรม KU Word Cut (สุธี, 2547)

(ข) การกำกับชนิดของคำ (Part of Speech Tagging) เมื่อเอกสารภาษาไทยได้ผ่านการตัดคำแล้ว กระบวนการต่อไปคือการกำกับชนิดของคำเพื่อนำไปใช้ในการวิเคราะห์วากยสัมพันธ์ของประโยคต่อไป ในการกำกับชนิดของคำได้ใช้โปรแกรม KU Word Cut (สุธี, 2547) ซึ่งเป็นโปรแกรมในชุดเดียวกันกับที่ใช้ในการตัดคำ

(ค) การแจกแจงประโยค (Syntactic Parsing) ในกระบวนการสุดท้ายของการประมวลผลส่วนหน้าหลังจากที่เอกสารได้ผ่านการตัดคำและกำกับชนิดของคำแล้ว คือการแจกแจงประโยค ซึ่งในวิทยานิพนธ์นี้ได้เลือกใช้โปรแกรมแจกแจงประโยคที่พัฒนามาจากเทคนิกการเรียนรู้ทางสถิติของไวยากรณ์ในประโยค (Charniak, 1997; Johnson, 1998) ซึ่งประโยคในเอกสารจะมีการกำกับวลีเพิ่มเข้ามาในกระบวนการนี้ ภาพที่ 24 แสดงตัวอย่างของเอกสารที่ผ่านการประมวลผลส่วนหน้าแล้ว

```
[ การ/pref1 ปลูก/vt  [ หน่อไม้ฝรั่ง/ncn ]/NP  ]/NP  ]/S
[ [ ลักษณะ/ncn  ทั่วไป/quaf ]/NP ]/S
[ [ หน่อไม้ฝรั่ง/ncn ]/NP [ เป็น/vcs  [ พืช/ncn  ผัก/ncn  [ ที่/prel  [ ได้รับ/vt  ความ/
pref1 สนใจ/vt มาก/adv  [ ใน/prep [  ปัจจุบัน/ncn  ]/NP ]/PP ]/VP
]/RELC  ]/NP ]/VP  _/blk  ]/S
```

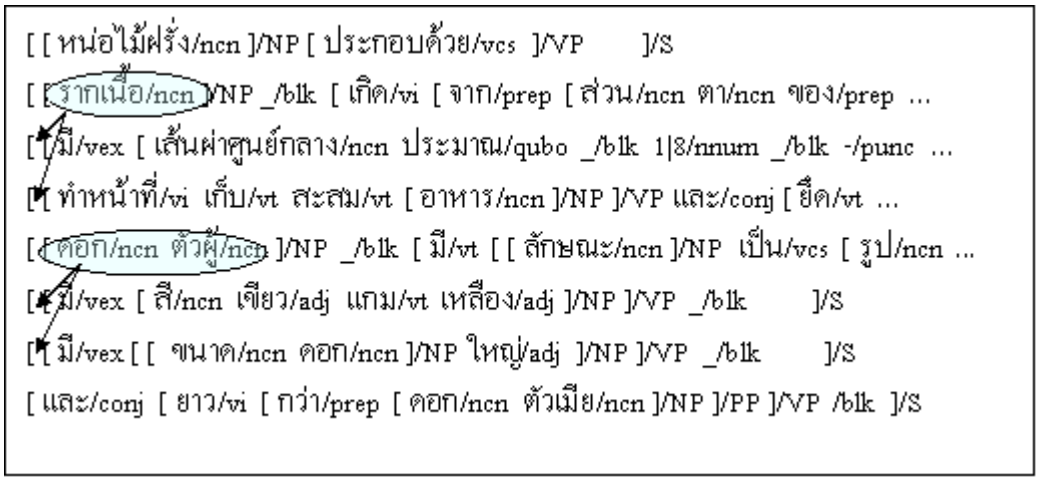
ภาพที่ 24 เอกสารที่ผ่านการประมวลผลส่วนหน้าแล้ว

2. การระบุอ็อบเจกต์ (Object Identifier)

หลังจากที่เอกสารได้ผ่านการวิเคราะห์คำและวากยสัมพันธ์แล้ว กระบวนการต่อไปจะเป็นการแก้ไขปัญหาที่เกี่ยวกับพฤติกรรมทางภาษาซึ่งมีผลต่อกระบวนการระบุอ็อบเจกต์ โดยพฤติกรรมทางภาษาที่จำเป็นต้องแก้ไขในระบบสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์มี 2 ชนิดคือ การใช้สิ่งอ้างอิงร่วมแบบสูญรูป และการละคำ ดังนี้

2.1 วิธีแก้การใช้สิ่งอ้างอิงร่วมแบบสูญรูป (Zero Anaphora Resolution)

จากการสำรวจคลังเอกสารในโดเมนการเกษตรนั้น การใช้สิ่งอ้างอิงร่วมแบบสูญรูปนั้นมักจะเกิดที่ตำแหน่งประธานของประโยค ซึ่งพิจารณาในประโยคที่มีการใช้สิ่งอ้างอิงร่วมแบบสูญรูปจะพบว่าส่วนใหญ่สิ่งอ้างอิงร่วมแบบสูญรูปนั้น จะอ้างอิงถึงประธานจากประโยคก่อนหน้า ฉะนั้นการแก้ปัญหาการใช้สิ่งอ้างอิงร่วมแบบสูญรูปจึงเลือกใช้กฎฮิวริสติก (Heuristic Rules) ด้วยการดึงประธานจากประโยคก่อนหน้ามาเติมในตำแหน่งที่มีการใช้สิ่งอ้างอิงร่วมแบบสูญรูป ภาพที่ 25 แสดงการดึงประธานจากประโยคก่อนหน้ามาแก้ปัญหาการใช้สิ่งอ้างอิงร่วมแบบสูญรูป



<u>ภาพที่ 25</u> การดึงประธานจากประโยคก่อนหน้ามาแก้ปัญหาการใช้สิ่งอ้างอิงร่วมแบบสูญรูป

2.2 วิธีแก้การละข้อความ (Textual Ellipsis Resolution)

นอกจากการใช้สิ่งอ้างอิงร่วมแบบสูญรูปแล้ว ยังมีพฤติกรรมทางภาษาอีกอย่างที่เป็นปัญหาในการระบุอ็อบเจกต์นั่นคือ การละข้อความ ในการสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์นั้นในตำแหน่งของอ็อบเจกต์บางอ็อบเจกต์ที่มีความหมายเป็นส่วนประกอบของสิ่งต่าง ๆ เช่น ปีก, ขา, ไข่ เป็นต้น ซึ่งอ็อบเจกต์เหล่านี้จะต้องมีบุพบทที่แสดงตัวความเป็นเจ้าของ ซึ่งในภาษาไทยมักจะมีการละบุพบทที่แสดงความเป็นเจ้าของ ฉะนั้นในการสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์นั้นจึงต้องมีอัลกอริทึมในการเติมบุพบทที่แสดงความเป็นเจ้าของให้กับอ็อบเจกต์ ในการค้นหานามวลีที่เป็นเจ้าของอ็อบเจกต์นั้น ระบบต้องใช้ฐานข้อมูลคำใน WordNet (Miller, 1990) มาใช้ในการวิเคราะห์นามวลีว่านามวลีก่อนหน้านี้ นามวลีใดที่มีความสัมพันธ์แบบส่วนประกอบ

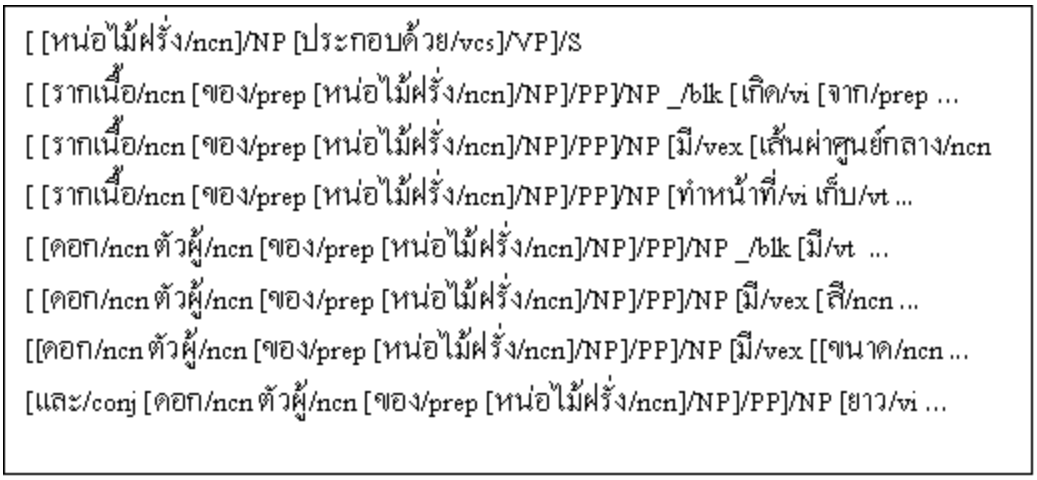
(Meronym) กับอ็อบเจกต์ได้บ้าง แต่ฐานข้อมูลคำใน WordNet นั้นเป็นฐานข้อมูลคำของภาษาอังกฤษ จึงต้องมีพจนานุกรม 2 ภาษาคือ ภาษาไทยและภาษาอังกฤษ มาใช้ในการแปลคำภาษาไทยให้เป็นภาษาอังกฤษก่อน แล้วจึงใช้คำภาษาอังกฤษนั้นมาใช้หาความสัมพันธ์แบบส่วนประกอบในฐานข้อมูลคำ WordNet อีกทอดหนึ่ง เช่น คำว่า “ดอก” และคำว่า “หน่อไม้ฝรั่ง” เมื่อใช้พจนานุกรมแปลเป็นภาษาอังกฤษจะได้คำว่า “flower” และ “Asparagus” จากนั้นจึงใช้คำภาษาอังกฤษนี้ค้นหาความสัมพันธ์แบบส่วนประกอบใน WordNet จะพบว่า “flower” เป็นส่วนประกอบของ “Asparagus” ภาพที่ 26 แสดงอัลกอริทึมในการแก้ปัญหาการละคำ

```
Input : S is the current sentence
Output : S is the solved sentence.

EllipsisRes (S):
1 N = getSubject(S)
2 if N is no possession preposition:
3    Sp = getPreviousSentence(S)
4    if N is "part, piece" or "plant part" sense:
5       while Sp != Null:
6          I = getSubject(Sp)
7          if N is same I:
8             if I is no prep(I):
9                Sp = getPreviousSentence(Sp)
10               continue
11            addPrep(N,prep(I))
12            return S
13         else if N is meronym of I:
14            addPrep(N,I)
15            return S
16         else if N is meronym of prep(I)
17            addPrep(N,prep(I))
18            return S
19         else:
20            Sp = getPreviousSentence(Sp)
21   else: return S
```

ภาพที่ 26 อัลกอริทึมการแก้ปัญหาการละคำ

เมื่อเอกสารผ่านการแก้ปัญหาการละคำแล้ว อ็อบเจกต์ของประโยคจะมีการเติมบุพบทที่แสดงความเป็นเจ้าของเพิ่มเข้ามา ภาพที่ 27 แสดงตัวอย่างเอกสารที่ผ่านการแก้ปัญหาการละคำมาแล้ว



ภาพที่ 27 เอกสารที่ผ่านการแก้ปัญหาการใช้สิ่งอ้างอิงร่วมและปัญหาการละคำแล้ว

3. การระบุคุณสมบัติ (Property Identifier)

หลังจากที่เอกสารที่ผ่านการประมวลผลส่วนหน้าและผ่านการระบุอ็อบเจกต์ ซึ่งเอกสารจะได้รับการแก้ปัญหาการใช้สิ่งอ้างอิงร่วมแบบสูญรูปและการละคำแล้ว กระบวนการต่อมาจะเป็นการระบุคุณสมบัติในที่ปรากฏในประโยค โดยสิ่งที่จะถูกระบุในกระบวนการนี้มี 2 อย่าง คือ ชนิดของความสัมพันธ์ทางคุณสมบัติ และค่าของคุณสมบัติ โดยจะใช้เทคนิคการเทียบแม่แบบในการสกัดความรู้ ซึ่งแม่แบบที่ใช้ในกระบวนการระบุคุณสมบัติจะมีแม่แบบ 2 ชนิด คือ แม่แบบประโยค (Sentence Template) และแม่แบบค่าคุณสมบัติ (Value Template)

ในกระบวนการระบุคุณสมบัตินั้น จะนำแม่แบบทั้งสองชนิดมาทำงานร่วมกันกับความรู้สนับสนุนเพื่อใช้ในการนำมาเปรียบเทียบกับประโยคต่าง ๆ ในเอกสาร เพื่อทำการระบุและดึงสารสนเทศในประโยคมาสร้างเป็นฐานความรู้ โดยตัวอย่างของแม่แบบประโยคนั้นได้แสดงไว้ในภาพที่ 28 และตารางที่ 7 ได้แสดงตัวอย่างของฐานความรู้คำเกี่ยวกับความสัมพันธ์ทางคุณสมบัติเพื่อสนับสนุนการระบุความสัมพันธ์ทางคุณสมบัติ ซึ่งแม่แบบประโยคประโยคทั้งหมดได้แสดงไว้ในภาคผนวก ก และ ฐานความรู้คำเกี่ยวกับความสัมพันธ์ทางคุณสมบัติทั้งหมดได้แสดงไว้ในภาคผนวก ข

```
[ object [ prop [ value ]/ADVP  ]/VP ]/S
[ object [มี/vt [ prop [ value ]/ADJP ]/NP  ]/VP ]/S
[ object [มี/vt [ prop value ]/NP  ]/VP ]/S
[ object [ value ]/VP ]/S
…
```

ภาพที่ 28 ตัวอย่างแม่แบบประโยค

ตารางที่ 7 ตัวอย่างฐานความรู้คำเกี่ยวกับความสัมพันธ์ทางคุณสมบัติ

ความสัมพันธ์ทางคุณสมบัติ	ฐานคำ
length	ยาว/vi
	ความ/pref1 ยาว/vi
height	สูง/vi
	ความ/pref1 สูง/vi
diameter	เส้นผ่าศูนย์กลาง/ncn
	เส้นผ่านศูนย์กลาง/ncn
weight	หนัก/vi
	น้ำหนัก/ncn
taste	รส/ncn
	รสชาติ/ncn

จากภาพที่ 28 นั้นได้แสดงตัวอย่างของแม่แบบประโยค ซึ่งแม่แบบประโยคจะมีโครงสร้างคล้ายกับประโยคที่ผ่านการประมวลผลส่วนหน้ามาแล้ว แต่ในแม่แบบประโยคจะมีคำสำคัญ (key word) 3 คำคือ object, prop และ valueซึ่งจะทำหน้าที่เป็นตัวระบุองค์ประกอบความรู้ที่ปรากฏในประโยค โดยคำสำคัญ object จะหมายถึงนามวลีที่เป็นอ็อบเจกต์ของความรู้ คำสำคัญ prop จะหมายถึงคำที่ใช้ระบุชนิดของความสัมพันธ์ทางคุณสมบัติซึ่งจะต้องทำงานร่วมกับฐานความรู้คำเกี่ยวกับความสัมพันธ์ทางคุณสมบัติด้วย และสุดท้ายคำสำคัญ value จะหมายถึงค่าของคุณสมบัติในประโยค ซึ่งตรงตำแหน่งนี้เองจะต้องทำงานร่วมกับแม่แบบค่าคุณสมบัติอีกทอดหนึ่ง เพื่อระบุ

องค์ประกอบต่าง ๆ ของค่าของคุณสมบัติออกมา โดยภาพที่ 29 แสดงตัวอย่างของแม่แบบค่าคุณสมบัติ ซึ่งแม่แบบค่าคุณสมบัติทั้งหมดได้แสดงไว้ในภาคผนวก ก

```
qnum [ num measure ]/NP
num measure
qnum [ num1 -/punc num2 measure ]/NP
sym
sym qsym
sym1 oper sym2
…
```

ภาพที่ 29 ตัวอย่างแม่แบบค่าคุณสมบัติ

จากตัวอย่างแม่แบบค่าคุณสมบัติในภาพที่ 29 จะเป็นรูปแบบของค่าคุณสมบัติทั้งแบบตัวเลขและแบบสัญลักษณ์ เช่น “ประมาณ 2-3 มิลลิเมตร” จะสามารถสกัดได้ด้วยแม่แบบ “qnum [num1 -/punc num2 measure]/NP” และ “เขียวอมเหลือง” จะสามารถสกัดได้ด้วยแม่แบบ “sym1 oper sym2” เป็นต้น โดยในแม่แบบค่าคุณสมบัติจะมีคำสำคัญ 10 คำด้วยกันในการระบุองค์ประกอบต่าง ๆ ของค่าของคุณสมบัติ โดยคำสำคัญ num จะหมายถึงค่าตัวเลขของค่าแบบตัวเลขในกรณีที่เป็นค่าเดี่ยว (Single Value) เช่น “5” เป็นต้น คำสำคัญ num1 และ num2 จะหมายถึงค่าตัวเลขของค่าแบบตัวเลขในกรณีที่เป็นค่าช่วง (Range Value) โดย num1 จะเป็นค่าแรกของช่วง และ num2 จะเป็นค่าสุดท้ายของช่วง เช่น “2 – 3” เป็นต้น คำสำคัญ measure จะหมายถึงหน่วยวัดของค่าแบบตัวเลข เช่น “มิลลิเมตร” คำสำคัญ qnum จะหมายถึงส่วนขยายค่าของค่าแบบตัวเลข เช่น “ประมาณ” คำสำคัญ sym จะหมายถึงค่าสัญลักษณ์ในกรณีของค่าแบบสัญลักษณ์ เช่น “เขียว” เป็นต้น คำสำคัญ sym1 และ sym2 จะเป็นค่าสัญลักษณ์ของค่าแบบสัญลักษณ์ในกรณีที่มีส่วนควบคุมร่วมอยู่ด้วยในค่าแบบสัญลักษณ์ โดย sym1 จะเป็นค่าแรกของค่าแบบสัญลักษณ์ และ sym2 จะเป็นค่าที่ 2 ของค่าแบบสัญลักษณ์ คำสำคัญ qsym จะหมายถึงส่วนขยายค่าของค่าแบบสัญลักษณ์ เช่น “เข้ม” เป็นต้น และสุดท้าย oper จะหมายถึงส่วนควบคุมค่าของค่าแบบสัญลักษณ์ เช่น “ปน” เป็นต้น

ในการระบุองค์ประกอบต่าง ๆ ภายในค่าของคุณสมบัติในประโยคนั้น แม่แบบค่าคุณสมบัติจะต้องทำงานร่วมกับความรู้สนับสนุน โดยความรู้สนับสนุนจะบรรจุฐานข้อมูลคำที่ใช้ร่วมกับคำสำคัญของแม่แบบค่าคุณสมบัติ โดยตารางที่ 8, 9, 10, 11 และ 12 แสดงตัวอย่างของ

ฐานความรู้สนับสนุนที่ต้องใช้ร่วมกับแม่แบบค่าคุณสมบัติ ซึ่งฐานความรู้สนับสนุนทั้งหมดได้แสดงไว้ในภาคผนวก ข

ตารางที่ 8 ตัวอย่างฐานความรู้สนับสนุนส่วนขยายค่าของค่าแบบตัวเลข

ฉลากค่า	ฐานคำ
approx	ประมาณ/qubo
notless	ไม่/neg น้อย/vi กว่า/qubo
morethan	มาก/vi กว่า/qubo

ตารางที่ 9 ตัวอย่างฐานความรู้สนับสนุนส่วนขยายค่าของค่าแบบสัญลักษณ์

ความสัมพันธ์ทางคุณสมบัติ	ฐานคำ	ค่าส่วนขยาย
color	เข้ม/adj	-3
color	อ่อน/adj	3
taste	เจี๊ยบ/adj	-3
common	มาก/adj	-3
common	น้อย/adj	3

ตารางที่ 10 ตัวอย่างฐานความรู้สนับสนุนหน่วยวัด

หน่วยวัด	ฐานคำ	ความสัมพันธ์ทางคุณสมบัติ	อัตราส่วน
millimeter	มิลลิเมตร/cl	length	1
inch	นิ้ว/cl	length	25.4
gram	กรัม/cl	weight	1000

ตารางที่ 11 ตัวอย่างฐานความรู้สนับสนุนค่าสัญลักษณ์

ความสัมพันธ์ทางคุณสมบัติ	ฐานคำ	ฉลากค่า	การแปลงค่า
color	เขียว/adj	green	hue=40
color	ส้ม/adj	orange	hue=10
taste	หวาน/adj	sweet	hue=60
odour	หอม/adj	aromatic	hue=20

ตารางที่ 12 ตัวอย่างฐานความรู้สนับสนุนส่วนควบคุมค่าของค่าแบบสัญลักษณ์

ความสัมพันธ์ทางคุณสมบัติ	ฐานคำ	ฟังก์ชั่น
color	อม/adj แกม/adj	means
taste	อม/adj	means

หลังจากผ่านกระบวนการระบุคุณสมบัติแล้ว องค์ความรู้ในประโยคจะถูกนำมาจัดเก็บให้อยู่ในรูปแบบของตรรกะศาสตร์ภาคแสดง (Predicate Logic) เพื่อเก็บไว้ในฐานความรู้ต่อไป โดยภาพที่ 30 แสดงตัวอย่างประโยคที่ถูกสกัดและแปลงให้อยู่ในรูปแบบตรรกะศาสตร์ภาคแสดง (Predicate Logic)

เมล็ดของหน่อไม้ฝรั่งมีเส้นผ่าศูนย์กลางประมาณ 0.2 นิ้ว

↓

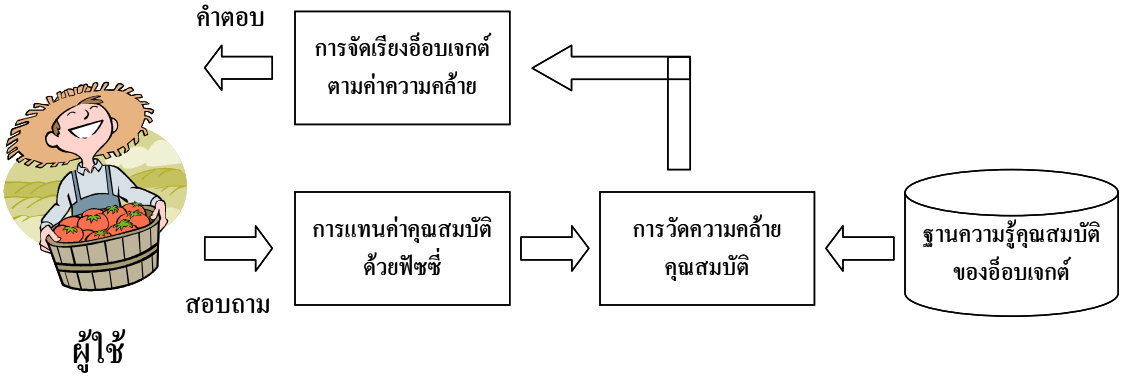
pname(x1, หน่อไม้ฝรั่ง),
part(x1, เมล็ด),
prop(diameter, x1, v1),
value(v1, 0.2, inch, approx)

ภาพที่ 30 ตัวอย่างประโยคที่ถูกแปลงให้อยู่ในรูปของตรรกะศาสตร์ภาคแสดง

จากภาพที่ 30 จะเห็นว่ามีภาคแสดง (Predicate) อยู่ 4 ชนิด คือ pname, part, prop และ value โดย pname จะหมายถึงชื่อเฉพาะหรือชื่อของอ็อบเจกต์นั้น คำว่า part จะแสดงถึงชิ้นส่วนของอ็อบเจกต์ คำว่า prop จะหมายถึงความสัมพันธ์ทางคุณสมบัติ และ value จะแสดงถึงค่าของคุณสมบัติแบบตัวเลขในกรณีค่าเดี่ยว ภาคแสดงในส่วนของค่าของคุณสมบัตินั้นนอกจาก value แล้วยังมีอีก 3 ภาคแสดงคือ range หมายถึงค่าของคุณสมบัติแบบตัวเลขในกรณีค่าช่วง เช่น “ประมาณ 10-20 นิ้ว” จะแสดงเป็น “range(v1, 10, 20, inch, approx)” คำว่า sym หมายถึงค่าของคุณสมบัติแบบสัญลักษณ์ เช่น “สีแดงเข้ม” จะแสดงเป็น “sym(v1, red, -3)” และ symoper หมายถึงค่าของคุณสมบัติในกรณีที่มีส่วนควบคุมค่าด้วย เช่น “สีเขียวปนเหลือง” จะแสดงเป็น “symoper(v1, green, yellow, means)”

ภาพรวมของระบบสืบค้นอ็อบเจกต์

หลังจากที่ฐานความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์ได้ถูกสร้างขึ้นมาจากการสกัดจากเอกสารแล้ว ฐานความรู้นี้จะสามารถนำมาใช้ในระบบสืบค้นอ็อบเจกต์ด้วยคุณสมบัติได้ โดยในระบบสืบค้นอ็อบเจกต์ด้วยคุณสมบัติจะมีขั้นตอนทั้งหมด 3 ขั้นตอนคือ การแทนค่าของคุณสมบัติด้วยทฤษฎีฟัซซี่, การคำนวณความคล้ายของค่าคุณสมบัติ และการเรียงลำดับความคล้าย โดยภาพรวมของระบบสืบค้นอ็อบเจกต์ด้วยคุณสมบัติได้แสดงในภาพที่ 31



ภาพที่ 31 ภาพรวมของระบบสืบค้นอ็อบเจกต์ด้วยคุณสมบัติ

ในการสืบค้นอ็อบเจกต์ด้วยคุณสมบัติจะเริ่มจากการที่ผู้ใช้ได้ป้อนค่าของคุณสมบัติต่าง ๆ เข้าไปยังระบบสืบค้น จากนั้นค่าของคุณสมบัติที่ผู้ใช้ป้อนเข้ามาและค่าคุณสมบัติของอ็อบเจกต์ในฐานความรู้จะถูกแทนค่าด้วยกราฟค่าความเป็นสมาชิกในทฤษฎีฟัซซี่ จากนั้นค่าของคุณสมบัติที่ผู้ใช้ป้อนเข้ามาและค่าคุณสมบัติของอ็อบเจกต์ที่ถูกแทนค่าด้วยฟัซซี่แล้วจะถูกนำมาคำนวณความคล้ายด้วยสูตรคำนวณความคล้าย จากนั้นระบบจะทำการเรียงลำดับความคล้ายระหว่างอ็อบเจกต์กับ

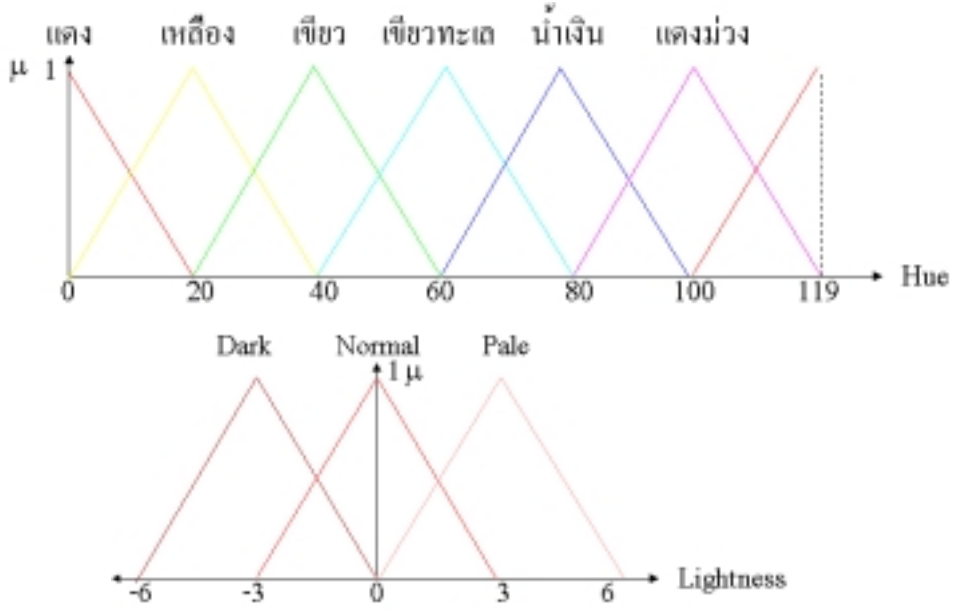
ค่าของคุณสมบัติที่ผู้ใช้ป้อนเข้ามา โดยเรียงลำดับจากมากไปหาน้อย แล้วจึงนำมาแสดงต่อผู้ใช้ โดยรายละเอียดในแต่ละขั้นตอนมีดังนี้

1. การแทนค่าของคุณสมบัติด้วยทฤษฎีฟัซซี่ (Fuzzy Property Representation)

ค่าของคุณสมบัติมี 2 ชนิด คือ ค่าแบบตัวเลข และค่าแบบสัญลักษณ์ โดยในการสืบค้นอ็อบเจกต์ด้วยคุณสมบัตินั้น ค่าของคุณสมบัติจะต้องถูกแทนค่าด้วยกราฟค่าความเป็นสมาชิก ซึ่งการแทนค่าของคุณสมบัติด้วยฟัซซี่นั้นจะมีการแทนค่า 2 แบบตามชนิดของค่าของคุณสมบัติเช่นกัน ดังนี้

1.1 การแทนค่าความเป็นสมาชิกของค่าแบบสัญลักษณ์(Symbolic Fuzzy Representation)

การแทนค่าความเป็นสมาชิกของค่าแบบสัญลักษณ์เช่น ค่าสี จะเป็นการแทนค่าด้วยกราฟค่าความเป็นสมาชิกของค่าเฉดสี (hue) และค่าความเข้มของแสง (lightness) มาใช้ในการแทนค่าของสัญลักษณ์ โดยข้อมูลการสร้างกราฟค่าความเป็นสมาชิกนั้น จะใช้ข้อมูลจากฐานความรู้สนับสนุนของค่าสัญลักษณ์ในส่วนของการแปลงค่า โดยภาพที่ 32 จะแสดงตัวอย่างของกราฟค่าความเป็นสมาชิกของค่าแบบสัญลักษณ์

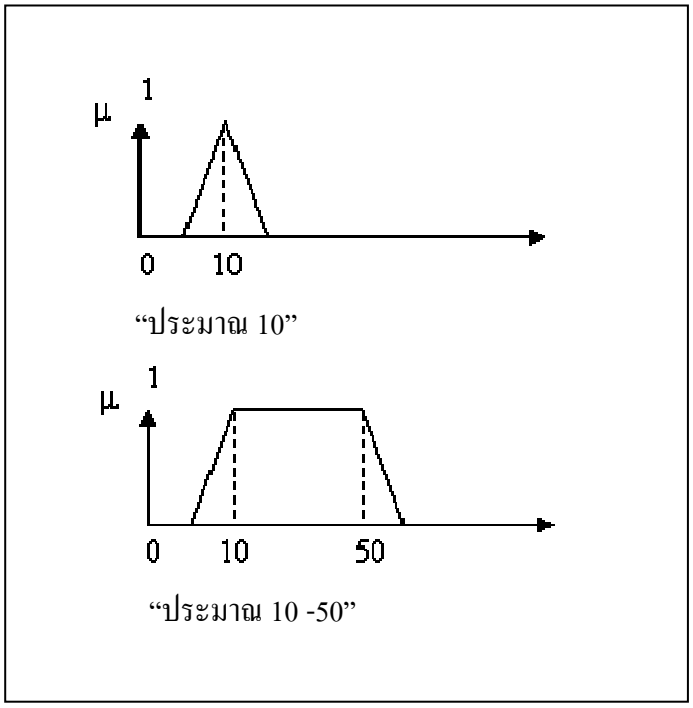


ภาพที่ 32 กราฟค่าความเป็นสมาชิกของค่าแบบสัญลักษณ์

จากภาพที่ 32 จะเห็นว่าค่าสีในแต่ละสีจะถูแทนด้วยกราฟค่าความเป็นสมาชิกของค่าเฉดสี เช่น สีเหลือง จะมีค่าเฉดสี ระหว่าง 0 ถึง 40 โดยค่าความเป็นสมาชิกจะเป็น 1 ที่ตำแหน่งค่าเฉดสี 20 จากนั้นค่าความเป็นสมาชิกจะลดลงทั้งสองข้างจนถึง 0 เป็นต้น ในส่วนของส่วนขยายค่านั้นจะแทนด้วยค่าความเป็นสมาชิกของค่าความเข้มแสง โดยถ้าส่วนขยายเป็น “เข้ม” จะแทนด้วยค่าความเข้มแสงที่ -3 ถ้าไม่มีส่วนขยาย จะแทนด้วยค่าความเข้มแสงที่ 0 และถ้าส่วนขยายเป็น “อ่อน” จะแทนด้วยค่าความเข้มแสงที่ 3 เป็นต้น

1.2 การแทนค่าความเป็นสมาชิกของค่าแบบตัวเลข (Numerical Fuzzy Representation)

การแทนค่าความเป็นสมาชิกของค่าแบบตัวเลขจะเป็นการแทนค่าด้วยกราฟของค่าความเป็นสมาชิกของช่วงค่าของตัวเลข ภาพที่ 33 แสดงตัวอย่างของกราฟค่าความเป็นสมาชิกของค่าแบบตัวเลข



<u>ภาพที่ 33</u> กราฟค่าความเป็นสมาชิกของค่า “ประมาณ 10- 50” และ “ประมาณ 10”

จากรูปที่ 33 ในกรณีค่าแบบตัวเลขเป็นช่วง กราฟค่าความเป็นสมาชิกจะเป็น 1 ตามช่วงของค่าที่ระบุไว้ และค่าความเป็นสมาชิกจะลดค่ากราฟเป็นกราฟลาดลง ลักษณะคล้ายสี่เหลี่ยมคางหมู เนื่องมาจากส่วนขยายค่าแบบตัวเลข “ประมาณ” ซึ่งถ้าค่าแบบตัวเลขนี้ไม่มีส่วนขยายค่าก็จะไม่มี

ส่วนของกราฟที่ลาดลง แต่กราฟค่าความเป็นสมาชิกจะมีรูปเป็นรูปสี่เหลี่ยมผืนผ้า ส่วนในกรณีค่าแบบตัวเลขเป็นค่าเดี่ยว ก็จะมีกราฟค่าความเป็นสมาชิกเป็น 1 ที่จุดเดียว ถ้าหากมีส่วนขยายค่า “ประมาณ” ก็จะมีกราฟลาดลง เป็นรูปสามเหลี่ยม แต่ถ้าไม่มีส่วนขยาย “ประมาณ” ก็จะเป็นเส้นตรงที่จุดของค่าเดี่ยวนั้น

2. การคำนวณความคล้ายของค่าคุณสมบัติ (Similarity Measure of Property)

หลังจากค่าของคุณสมบัติได้ถูกแทนค่าด้วยกราฟค่าความเป็นสมาชิกแล้ว กระบวนการต่อไปจะเป็นการคำนวณค่าความคล้าย (Similarity Measure) ของคุณสมบัติ สูตรคำนวณความคล้ายในวิทยานิพนธ์นี้ได้เลือกสูตร Angular Separation มาใช้ เนื่องจากค่าคุณสมบัติมีสองชนิด คือแบบตัวเลขและแบบสัญลักษณ์ ซึ่งมีหน่วยไม่ตรงกันจะนำมาคำนวณเทียบกันไม่ได้ ถ้าหากใช้วิธีการวัดความคล้ายแบบระยะ จึงเปลี่ยนมาใช้วัดค่ามุมทางเวคเตอร์ (Vector) แทนโดยสูตรคำนวณความคล้ายจะมี 2 สูตร ซึ่งแบ่งตามชนิดของค่าคุณสมบัติคือ สูตรความคล้ายของค่าแบบสัญลักษณ์ (Symbolic Similarity Measurement) และสูตรความคล้ายของค่าแบบตัวเลข (Numerical Similarity Measurement) โดยในการนำกราฟค่าความเป็นสมาชิกมาใช้ในสูตรคำนวณความคล้ายนั้น กราฟจะถูกแซมปลิ้ง (Sampling) ค่าออกมาเป็นจุด ๆ โดยจำนวนแซมปลิ้งจะขึ้นอยู่กับความเหมาะสมของค่าของคุณสมบัติ โดยรายละเอียดของสูตรคำนวณความคล้ายจะอธิบายในหัวข้อย่อยดังนี้

2.1 สูตรความคล้ายของค่าแบบสัญลักษณ์ (Symbolic Similarity Measurement)

ค่าของคุณสมบัติแบบสัญลักษณ์ที่ถูกแทนค่าด้วยกราฟความเป็นสมาชิกแล้วจะถูกแซมปลิ้ง ค่าออกมาเป็นจุด ๆ เพื่อใช้ในการคำนวณระยะห่างเชิงมุม โดยสูตรความคล้ายของค่าแบบสัญลักษณ์มีดังนี้

$$sim(i,\ j) = \frac{\sum_k \left(\mu(i)_k \times \mu(j)_k\right)}{\sqrt{\left(\sum_k \mu(i)_k^2\right) \times \left(\sum_k \mu(j)_k^2\right)}}$$

โดยที่ $\mu(i)_k$ คือ ค่าความเป็นสมาชิกของค่าคุณสมบัติสืบค้นที่ตำแหน่งแซมปลิ้ง k

$\mu(j)_k$ คือ ค่าความเป็นสมาชิกของค่าคุณสมบัติของอ็อบเจกต์เป้าหมายที่ตำแหน่งแซมปลิ้ง k

2.2 สูตรความคล้ายของค่าแบบตัวเลข (Numerical Similarity Measurement)

ค่าของคุณสมบัติแบบตัวเลขที่ถูกแทนที่ด้วยกราฟความเป็นเป็นสมาชิกแล้วจะถูกแซมปลิ้งค่าออกมาเป็นจุด ๆ เช่นเดียวกับกรณีค่าแบบสัญลักษณ์ แต่สูตรความคล้ายแบบตัวเลขจะมีการดัดแปลงสูตรเล็กน้อยเพื่อให้เหมาะสม เพราะค่าแบบตัวเลขนั้นจะต่างจากค่าแบบสัญลักษณ์ตรงที่ค่าแบบสัญลักษณ์นั้นค่าความเป็นสมาชิกจะอยู่ในช่วงเดียวกันหมด แต่ค่าแบบตัวเลขค่าความเป็นสมาชิกอาจจะอยู่คนละช่วงกันได้ โดยสูตรความคล้ายของค่าแบบตัวเลขมีดังนี้

$$sim(i,\ j) = \frac{\sum_k \left(\mu(i)_k \times \min\left(\mu(i)_k, \mu(j)_k\right)\right)}{\sqrt{\left(\sum_k \mu(i)_k^2\right) \times \left(\sum_k \min\left(\mu(i)_k, \mu(j)_k\right)^2\right)}}$$

โดยที่ $\mu(i)_k$ คือ ค่าความเป็นสมาชิกของค่าคุณสมบัติสืบค้นที่ตำแหน่งแซมปลิ้ง k

$\mu(j)_k$ คือ ค่าความเป็นสมาชิกของค่าคุณสมบัติของอ็อบเจกต์เป้าหมายที่ตำแหน่งแซมปลิ้ง k

min คือ ฟังก์ชั่นในการค่าที่ต่ำที่สุดของค่าสองค่า

หลังจากที่ระบบสืบค้นได้ทำการคำนวณความคล้ายระหว่างค่าคุณสมบัติสืบค้นกับค่าคุณสมบัติของอ็อบเจกต์ในฐานความรู้แล้วระบบสืบค้นก็จะทำการเรียงลำดับค่าความคล้ายที่คำนวณไว้แล้วในแต่ละอ็อบเจกต์ในฐานความรู้ โดยจะทำการเรียงจากค่าความคล้ายมากที่สุดไปยังค่าความคล้ายน้อยที่สุดแล้วจึงนำผลการเรียงค่าความคล้ายนั้นมาแสดงต่อผู้ใช้ต่อไป

ผลการทดลองและวิจารณ์

วิธีวัดผลการทดลอง

วิธีผลการทดลองจะมีการวัดผลการสกัดความรู้เกี่ยวกับคุณสมบัติจากเอกสารด้วยสูตรการวัดผลสามชนิดคือ ค่าความถูกต้อง (Precision), ค่าความระลึก (Recall) และค่าความถูกและระลึกได้โดยรวมแบบ F-measure โดยสูตรคำนวณค่าความถูกต้อง วิทยานิพนธ์นี้ได้ใช้สูตรดังข้างล่างนี้

$$P = \frac{c}{c + w}$$

โดยที่ P คือ ค่าความถูกต้อง
c คือ ผลลัพธ์ที่ถูกต้อง
w คือ ผลลัพธ์ที่ไม่ถูกต้อง

สำหรับสูตรคำนวณค่าความระลึกได้ วิทยานิพนธ์นี้ได้ใช้สูตรดังข้างล่างนี้

$$R = \frac{c}{t}$$

โดยที่ R คือ ค่าความระลึก
c คือ ผลลัพธ์ที่ถูกต้อง
t คือ ผลลัพธ์ที่ปรากฏอยู่ในเอกสารทั้งหมด

สูตรสำหรับคำนวณค่าค่าความถูกและระลึกได้โดยรวมแบบ F-measure

$$F = \frac{\left(\beta^2 + 1\right) PR}{\beta^2 R + P}$$

โดยที่ F คือ ค่าความถูกต้องและระลึกได้โดยรวมแบบ F-measure

P คือ ค่าความถูกต้อง

R คือ ค่าความระลึก

β คือ ค่าพารามิเตอร์ที่ใช้แสดงสัดส่วนความสำคัญระหว่างค่าความถูกต้องและค่าความระลึก โดยทั่วไปจะใช้ค่า เท่ากับ β 1

ในการทดลองวัดผลการสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์นั้น ได้มีการเตรียมเอกสารในโดเมนการเกษตรที่เกี่ยวกับศัตรูพืชขนาด 2,000 ประโยค โดยเอกสารนั้นได้ผ่านการประมวลผลส่วนหน้ามาเรียบร้อยแล้ว เพื่อใช้ในการทดลองสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารโดยอัตโนมัติ

ผลการทดลอง

จากการทดลองสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์ โดยในระบบสกัดความรู้ได้มีการวัดผลใน 3 ส่วนด้วยกัน คือ โปรแกรมแก้ปัญหาการใช้สิ่งอ้างอิงร่วมแบบสูญรูป (Zero Anaphora Resolution), โปรแกรมการแก้ปัญหาการละคำ (Textual Ellipsis Resolution) และโปรแกรมการระบุคุณสมบัติ (Property Identification) โดยทำการทดลองกับคลังเอกสารในโดเมนการเกษตรที่เกี่ยวกับศัตรูพืชขนาด 2,000 ประโยค โดยได้มีการแยกกันทดลองกับคลังเอกสารที่สะอาดแล้ว ไม่ได้นำผลลัพธ์จากโปรแกรมก่อนหน้ามาป้อนให้กับอีกโปรแกรม เพื่อการวัดผลที่ถูกต้องสำหรับแต่ละโปรแกรม

ในส่วนของระบบสืบค้นได้จัดทำชุดสืบค้นขึ้นมาพร้อมด้วยชุดคำตอบ เพื่อทำการวัดผลของระบบสืบค้น โดยในส่วนค่าความถูกต้องนั้น อ็อบเจกต์ที่ต้องการสืบค้นนั้นจะถือว่าถูกต้อง ถ้าหากค่าความคล้ายของอ็อบเจกต์นั้นมีค่าสูงสุดในการสืบค้นนั้น ๆ ในส่วนของค่าความระลึก ถ้าอ็อบเจกต์ที่ต้องการสืบค้นนั้นจะถือว่าระลึกได้ ถ้าหากค่าความคล้ายของอ็อบเจกต์นั้น ๆ อยู่ในลำดับไม่เกิน 3 ลำดับแรก โดยได้ผลการทดลองทั้งหมดดังตารางที่ 13

ตารางที่ 13 ผลการทดลองวัดค่าความถูกต้องและค่าความระลึกและค่า F-measure

	ค่าความถูกต้อง (%)	ค่าความระลึก (%)	ค่า F-measure (%)
การแก้ปัญหาการใช้สิ่งอ้างอิงร่วมแบบสูญรูป (Zero Anaphora Resolution)	82.14	71.42	76.40
การแก้ปัญหาการละคำ (Textual Ellipsis Resolution)	96.96	96.96	96.96
การระบุคุณสมบัติ (Property Identification)	88.88	47.05	61.53
การสืบค้นอ็อบเจกต์ (Object Query System)	81.81	90.90	86.11

วิจารณ์

จากผลการทดลองในส่วนของการแก้ปัญหาการใช้สิ่งอ้างอิงร่วมแบบสูญรูปนั้น เนื่องจากได้แก้ปัญหาด้วยวิธีฮิวริสติกโดยการดึงประธานจากประโยคก่อนหน้ามาเติมลงในสิ่งอ้างอิงร่วมแบบสูญรูปโดยตรง นั้นจะสามารถแก้ปัญหาได้ดีในเอกสารโดเมนการเกษตร แต่ก็มีผิดพลาดอยู่บ้าง เพราะบางกรณีต้องมีการวิเคราะห์ความหมายที่จำเป็นต้องใช้ความรู้สนับสนุนมาก ตัวอย่างเช่นประโยคข้างล่างนี้

(1) ทองพันชั่ง
(2) ลักษณะของพืชเป็นไม้พุ่มขนาดเล็ก
(3) $\underline{\varnothing}$ สูงประมาณ 0.5-2 เมตร

จากตัวอย่างข้างต้น ประโยคที่หนึ่งจะเป็นหัวข้อของเอกสาร ประโยคที่สองจะมีประธานคือ “ลักษณะของพืช” และประโยคที่สาม ประธานจะเป็นสิ่งอ้างอิงร่วมแบบสูญรูป ซึ่งถ้าใช้ฮิวริสติก จะทำให้ประโยคที่สามเป็น “ลักษณะของพืชสูงประมาณ 0.5-2 เมตร” ซึ่งจะถือว่าไม่ถูกต้อง เพราะถึงแม้ว่าคนอ่านจะอ่านแล้วเข้าใจเพราะคนจะมีความรู้สนับสนุนมากจึงทำให้เข้าใจ

ได้ว่า ทองพันชั่งนั้นสูง 0.5-2 เมตร แต่ระบบสกัดความรู้จะไม่สามารถสกัดความรู้ได้ นอกจากนี้ การใช้สิ่งอ้างอิงร่วมนั้น อาจจะเกิดขึ้นที่ตำแหน่งที่ไม่ใช่ตำแหน่งประธานก็ได้ เช่น กรรมตรง, กรรมรอง เป็นต้น

ในส่วนของการแก้ปัญหาการละคำนั้น ผลการทดลองได้ผลออกมาค่อนข้างสูง แต่เนื่องจากในโปรแกรมการแก้ปัญหาการละคำต้องใช้ฐานข้อมูลคำ WordNet มาใช้ในการค้นหาความสัมพันธ์แบบส่วนประกอบ (Meronym) ของอ็อบเจกต์ จึงทำให้เกิดปัญหาคำไม่พอ เช่น ทองพันชั่ง เป็นต้น ในฐานข้อมูลคำ WordNet อาจจะไม่มีคอนเซ็ปต์ของคำนามคำนี้ จึงทำให้ระบบไม่สามารถทำการแก้ปัญหาการละคำได้อย่างถูกต้องได้

ในส่วนของการระบุคุณสมบัตินั้นจะมีค่าความระลึกได้ที่ค่อนข้างต่ำ เนื่องจากการอธิบายคุณสมบัติของอ็อบเจกต์ในเอกสารนั้น จะอยู่ในรูปแบบของการเปรียบเทียบอยู่จำนวนมากซึ่งยากต่อการสกัดความรู้ ดังตัวอย่างข้างล่างนี้

ผลมังคุดจะมีขนาดเล็กกว่ากำปั้นเล็กน้อย

จากตัวอย่างข้างต้น ขนาดของผลมังคุดได้ถูกอธิบายในลักษณะของการเปรียบเทียบ ซึ่งยากในการสกัดความรู้ เนื่องจากระบบจะต้องมีฐานความรู้สนับสนุนในการอนุมานค่าที่แท้จริงจากสิ่งที่ถูกนำมาเปรียบเทียบด้วย

สรุปและข้อเสนอแนะ

สรุป

ในการพัฒนาระบบสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารนั้น ต้องมีกระบวนการประมวลทางภาษาที่ค่อนข้างซับซ้อน เช่น การประมวลการใช้สิ่งอ้างอิงร่วม (Anaphora) ชนิดต่าง ๆ การประมวลเกี่ยวกับการละคำ (Ellipsis) เนื่องจากพฤติกรรมทางภาษาต่าง ๆ เป็นอุปสรรคในการสกัดความรู้จากเอกสารที่เป็นภาษาธรรมชาติโดยตรง ดังนั้นก่อนที่จะมีกระบวนการสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์นั้นจึงจำเป็นต้องมีกระบวนการประมวลผลทางภาษาที่จำเป็นเหล่านี้ก่อน

ในกระบวนการแก้ปัญหาการใช้สิ่งอ้างอิงร่วมแบบสูญรูปนี้ ได้ใช้ฮิวริสติกในการแก้ปัญหาการใช้สิ่งอ้างอิงร่วมแบบสูญรูป โดยได้มีการวัดผลค่าความถูกต้อง 82.14 %, ค่าความระลึก 71.42 % และค่า F-measure 76.40 % ซึ่งผลลัพธ์ที่ได้จากผลการทดลองถือได้ว่าได้ผลดีพอสมควร เนื่องจากในโดเมนการเกษตรนั้น การใช้สิ่งอ้างอิงร่วมแบบสูญรูปนั้น มักจะมีการใช้ที่ตำแหน่งประธานในประโยคเป็นส่วนมาก โครงสร้างประโยคไม่มีความซับซ้อนเกินไป ซึ่งถ้าหากนำไปใช้ในโดเมนอื่น อาจจะได้ผลที่ไม่ดีเท่าที่ควร

ในกระบวนการแก้ปัญหาการละคำ ได้มีการใช้ฐานข้อมูลคำ WordNet มาเป็นฐานความรู้สนับสนุนในการประมวลผลการละบุพบทแสดงความเป็นเจ้า โดยได้เสนออัลกอริทึมในการค้นหาบุพบทแสดงความเป็นเจ้าของ โดยได้มีการวัดผลค่าความถูกต้อง 96.96 %, ค่าความระลึก 96.96 % และค่า F-measure 96.96 % ซึ่งผลลัพธ์ที่ได้ออกมาค่อนข้างดี แต่ปัญหาหลักของอัลกอริทึมนี้ก็คือคำนามบางคำอาจจะไม่มีในฐานข้อมูลคำ WordNet เช่น ทองพันชั่ง เป็นต้น ซึ่งเป็นพืชชนิดหนึ่ง แต่ไม่ปรากฏในฐานข้อมูลคำ WordNet ซึ่งจะทำให้การค้นหาบุพบทแสดงความเป็นเจ้าของนั้น ได้ผลลัพธ์ที่ผิดพลาดได้

ในกระบวนการระบุความรู้ในเอกสารนั้น เทคนิคการใช้การเทียบแม่แบบนั้น เหมาะกับการสกัดความรู้ในเอกสารที่ได้มีการออกแบบโครงสร้างของความรู้ไว้ล่วงหน้าแล้ว โดยได้มีการวัดผลค่าความถูกต้อง 88.88 %, ค่าความระลึก 47.50 % และค่า F-measure 61.53 % ซึ่งกระบวนการสกัดความรู้ด้วยวิธีเทียบแม่แบบนั้นจะให้ค่าความถูกต้องที่ค่อนข้างสูง อย่างไรก็ดี การใช้เทคนิค

เทียบแม่แบบในการสกัดความรู้นั้น จะเหมาะกับการสกัดความรู้ชนิดชัดเจน (Explicit Knowledge) แต่จะมีปัญหากับการสกัดความรู้ชนิดโดยนัย (Implicit Knowledge) อย่างเช่นความรู้ที่อยู่ในรูปแบบการเปรียบเทียบ เป็นต้น

ในระบบการสืบค้นอ็อบเจกต์ด้วยคุณสมบัติ ได้มีการนำทฤษฎีฟัซซี่มาใช้เป็นตัวแทนค่าของคุณสมบัติ เพื่อให้สามารถคำนวณความคล้ายของค่าของคุณสมบัติที่มีความคล้ายคลึงกันได้ โดยในระบบสืบค้นได้ใช้สูตรคำนวณความคล้าย Angular Separation มาใช้ในระบบสืบค้นอ็อบเจกต์ ซึ่งสามารถทำการสืบค้นอ็อบเจกต์ได้ดี โดยได้มีการวัดผลค่าความถูกต้อง 81.81 %, ค่าความระลึก 90.90 % และค่า F-measure 86.11 %

ข้อเสนอแนะ

เพื่อให้ได้ผลลัพธ์ที่ดีขึ้นในการสกัดความรู้ในเอกสาร กระบวนการประมวลที่เกี่ยวกับพฤติกรรมทางภาษาต่าง ๆ จำเป็นต้องมีวิธีการที่ซับซ้อนมากขึ้น เช่นการแก้ปัญหาการใช้สิ่งอ้างอิงร่วมนั้น เพราะสิ่งอ้างอิงร่วมในภาษาไทยนั้นมีหลายชนิด ดังนั้นระบบควรจะต้องมีอัลกอริทึมที่แก้ปัญหาการใช้สิ่งอ้างอิงร่วมชนิดอื่น ๆ ด้วย จึงทำให้ผลลัพธ์ในการสกัดนั้นได้ผลลัพธ์ที่สูงขึ้น

นอกจากนี้ ในระบบสกัดความรู้ประเภทอื่น ๆ ปัญหาเรื่องการละคำบางประเภท เช่น การละคำกิริยา การละบุพบทชนิดอื่น ๆ เป็นต้น อาจจะจำเป็นนอกเหนือจากการละบุพบทแสดงความเป็นเจ้าของได้ สำหรับการสกัดความรู้จากเอกสารในโดเมนอื่น ๆ

เพื่อให้ระบบสกัดความรู้มีความสามารถสูงขึ้น ความสามารถในการอนุมานความรู้ใหม่ ๆ ขึ้นมาเองในฐานความรู้จากฐานความรู้เก่า และนำความรู้ใหม่ ๆ เหล่านั้นมาช่วยในการสกัดความรู้ต่อไปนั้น เป็นแนวทางที่ช่วยทำให้ระบบสกัดความรู้สามารถสกัดความรู้ชนิดโดยนัย (Implicit Knowledge) ได้ดียิ่งขึ้นได้ เช่น “ผลมังคุดมีขนาดเล็กกว่ากำมือเล็กน้อย” ในประโยคนี้ถ้าเราสามารถสกัดความรู้เกี่ยวกับขนาดมือคนโดยเฉลี่ยได้ เราก็สามารถอนุมานขนาดของผลมังคุดได้ เป็นต้น

เอกสารและสิ่งอ้างอิง

ยืน ภู่วรวรรณ. 2529. การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี, *ใน* **รายงานการประชุมวิชาการวิศวกรรมไฟฟ้า ครั้งที่ 9** . ขอนแก่น.

วีร์ สัตยมาศ. 2548. **การพัฒนากรอบระบบรวบรวมคลังต้นไม้ไวยากรณ์ภาษาไทย.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

ศักดิ์ชาติ งามล้วน. 2548. **การเพิ่มประสิทธิภาพระบบค้นคืนรูปภาพโดยใช้เทคนิคประมวลผลภาษาธรรมชาติ.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

สุธี สุดประเสริฐ. 2547. **การตัดคำภาษาไทยด้วยเทคนิคการเรียนรู้แบบไม่ใช้ตัวอย่าง.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

Agrawal, R., H. Mannila, R. Srikant, and A. I. Verkamo. 1995. Fast Discovery of Association Rules, pp. 307-328. *In* **Advances in Knowledge Discovery and Data Mining** .

Berthold, M. and D. J. Hand. 2003. **Intelligent Data Analysis.** 2 ed. Springer, Germany.

Delannoy, J.F., C. Feng, S.Matwin, and S. Szpakowicz. 1993. Knowledge extraction from text: Machine learning for text-to-rule translation, *In* **Proceedings of the Workshop on Machine Learning Techniques and Text Analysis, European Conference on Machine Learning (ECML-93)** . Vienna, Austria.

Delisle, S., K. Barker, J. F. Delannoy, S. Matwin, and S. Szpakowicz. 1994. From text to horn clauses: Combining linguistic analysis and machine learning, *In* **Proceedings of the 10th Canadian Artificial Intelligence Conference, CAI-94** . Canada.

Charniak, E. 1997. Statistical parsing with a context-free grammar and word statistics, p. 598–603. *In* **Proceedings of AAAI/IAAI** .

Collins, M. 1999. **Head-Driven Statistical Models for Natural Language Parsing.** Ph.D. Thesis thesis, University of Pennsylvania.

Copeck, T., S. Delisle and S. Szpakowicz. 1992. Parsing and case analysis in TANKA, *In* **Proceeding of the 15th Intl. Conf. of Computational Linguistics COLING-92** . Nantes.

Grosz, B., A. Joshi, and S. Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse, pp. 203-225. *In* **Computational Linguistics** .

Gomez, F., R. Hull, and C. Segami. 1994. Acquiring knowledge from encyclopedia texts, *In* **Proceedings of the 4th ACL conference on Applied Natural Language Processing** . Struttgart, Germany.

Johnson, M. 1998. PCFG models of linguistic tree representations, p. 613–632. *In* **Computational Linguistics** .

Kawtrakul, A., C. Thumkanon, and S. Seriburi. 1995. A statistical approach to Thai word filtering in WPA project, pp. 68-75. *In* **Proceedings of the Second Symposium on Natural Language Processing** . 1995. A statistical approach to Thai word filtering in . Bangkok, Thailand.

Loh, S., L. K. Wives, and J. P. M. de Oliveira. 2000. Concept-Based Knowledge Discovery in Texts Extracted from the Web, *In* **SIGKDD Explorations, ACM SIGKDD** .

Miller, G. 1990. Wordnet: an on-line lexical database, *In* **International Journal of Lexicography** .

Minsky, M. 1975. A framework for representing knowledge, *In* **The Psychology of Computer Vision** . McGraw-Hill, New York.

Mitchell, T. M., R. M. Keller and S. T. Kedar-Cabelli. 1986. Explanation-Based Generalization: A Unifying View, *In* **Machine Learning** .

Moulin, B. and D. Rousseau. 1992. Automate KA from regulatory texts, *In* **IEEE Expert** .

Muggleton, S. H. 1992. **Inductive Logic Programming.** Academic Press.

Nahm, U. Y. and R. J. Mooney. 2002. Text Mining with Information Extraction, pp. 60-67. *In* **Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases** . Stanford, CA.

Pierre, J. M. 2002. Mining Knowledge from Text Collections Using Automatically Generated Metadata, *In* **Proceeding of Fourth International Conference on Practical Aspects of Knowledge Management** .

Schank, R. C. and R. P. Abelson. 1977. **Scripts, Plans, Goals, and Understanding.** Erlbaum, Hillsdale, NJ.

Schneider, G. 1998. **A Linguistic Comparison Constituency, Dependency, and Link Grammar.** Master's Thesis thesis, University of Zurich.

Zadeh, L. A. 1965. Fuzzy sets, *In* **Information and Control** .

Zadeh, L. A. 1983. A computational approach to fuzzy quantifiers in natural languages, *In* **Computers and Mathematics with Applications** .

ภาคผนวก

ภาคผนวก ก

แม่แบบประโยคและแม่แบบค่าคุณสมบัติ

แม่แบบประโยคและแม่แบบค่าคุณสมบัติ

ในระบบสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์นั้น ได้ใช้เทคนิคเทียบแม่แบบในการระบุองค์ประกอบต่าง ๆ ของความรู้ในเอกสาร ซึ่งแม่แบบที่ใช้ในระบบสกัดความรู้นี้มี 2 ชนิดคือ แม่แบบประโยค และ แม่แบบค่าคุณสมบัติ ซึ่งได้แสดงในตารางภาคผนวกที่ ก1 และ ก2 ดังนี้

ตารางผนวกที่ ก1 แม่แบบประโยค

[object [prop [value]/ADVP]/VP]/S
[object [มี [prop]/NP value]/VP]/S
[object [จะ มี [prop value]/NP]/VP]/S
[object [มี [prop value]/NP]/VP]/S
[และ object [มี [prop]/NP value]/VP]/S
[object [value]/VP]/S

ตารางผนวกที่ ก2 แม่แบบค่าคุณสมบัติ

qnum [num measure]/NP
qnum num measure
num measure
qnum num1 -/punc num2 measure
num1 -/punc num2 measure
sym
sym qsym
sym1 oper sym2

ภาคผนวก ข

ฐานความรู้สนับสนุน

ฐานความรู้สนับสนุน

ในการสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์จากเอกสารนั้น ต้องใช้ฐานความรู้สนับสนุน เพื่อใช้ในทำงานร่วมกับแม่แบบในการสกัดความรู้จากเอกสาร ซึ่งฐานความรู้เบื้อหลังนี้มี 6 ชนิด คือ ฐานความรู้คำเกี่ยวกับความสัมพันธ์ทางคุณสมบัติ, ส่วนขยายค่าของค่าแบบตัวเลข, ส่วนขยายค่าของค่าแบบสัญลักษณ์, หน่วยวัด, ค่าสัญลักษณ์ และส่วนควบคุมค่าของค่าแบบสัญลักษณ์ ดังนี้

ตารางผนวกที่ ข1 ฐานความรู้คำเกี่ยวกับความสัมพันธ์ทางคุณสมบัติ

ความสัมพันธ์ทางคุณสมบัติ	ฐานคำ
length	ยาว/vi ความ/pref1 ยาว/vi
height	สูง/vi ความ/pref1 สูง/vi
width	กว้าง/vi ความ/pref1 กว้าง/vi
diameter	เส้นผ่าศูนย์กลาง/ncn เส้นผ่านศูนย์กลาง/ncn
perimeter	เส้นรอบวง/ncn
weight	หนัก/vi น้ำหนัก/ncn
taste	รส/ncn รสชาด/ncn
color	สี/ncn
odour	กลิ่น/ncn

ตารางผนวกที่ ข2 ส่วนขยายค่าของค่าแบบตัวเลข

ฉลากค่า	ฐานคำ
approx	ประมาณ/qubo
notless	ไม่/neg น้อย/vi กว่า/qubo
notless	ไม่/neg ต่ำ/vi กว่า/qubo
lessthan	น้อย/vi กว่า/qubo
lessthan	ต่ำ/vi กว่า/qubo
morethan	มาก/vi กว่า/qubo
morethan	สูง/vi กว่า/qubo
notmore	ไม่/neg มาก/vi กว่า/qubo
notmore	ไม่/neg สูง/vi กว่า/qubo

ตารางผนวกที่ ข3 ส่วนขยายค่าของค่าแบบสัญลักษณ์

ความสัมพันธ์ทางคุณสมบัติ	ฐานคำ	กำลัง
color	เข้ม/adj	-3
color	อ่อน/adj	3
taste	เจี๊ยบ/adj	-3
taste	ปี๋/adj	-3
common	มาก/adj	-3
common	น้อย/adj	3

ตารางผนวกที่ ข4 หน่วยวัด

หน่วยวัด	ฐานคำ	ความสัมพันธ์ทางคุณสมบัติ	อัตราส่วน
millimeter	มิลลิเมตร/cl	length	1
inch	นิ้ว/cl	length	25.4
mile	ไมล์/cl	length	1609344
meter	เมตร/cl	length	1000
centimeter	เซนติเมตร/cl	length	10
centimeter	ซม./cl	length	10
foot	ฟุต/cl	length	304.8
yard	หลา/cl	length	914.4
kilometer	กิโลเมตร/cl	length	1000000
milligram	มิลลิกรัม/cl	weight	1
pound	ปอนด์/cl	weight	453592.37
kilogram	กิโลกรัม/cl	weight	1000000
centigram	เซนติกรัม/cl	weight	10
gram	กรัม/cl	weight	1000
ton	ตัน/cl	weight	907184740
ounce	ออนซ์/cl	weight	28349.5231

ตารางผนวกที่ ข5 ค่าสัญลักษณ์

ความสัมพันธ์ทางคุณสมบัติ	ฐานคำ	ฉลากค่า	การแปลงค่า
taste	เปรี้ยว/adj	sour	hue=20
taste	หวาน/adj	sweet	hue=60
taste	เค็ม/adj	salt	hue=100
taste	ขม/adj	bitter	hue=140

ตารางผนวกที่ ข5 ค่าสัญลักษณ์ (ต่อ)

ความสัมพันธ์ทางคุณสมบัติ	ฐานคำ	ฉลากค่า	การแปลงค่า
taste	เผ็ด/adj	spicy	hue=180
taste	เผ็ดร้อน/adj	spicy	hue=180
taste	จืด/adj	insipid	hue=220
color	เขียว/adj	green	hue=40
color	เหลือง/adj	yellow	hue=20
color	แดง/adj	red	hue=0
color	ส้ม/adj	orange	hue=10
color	แสด/adj	orange	hue=9
color	น้ำเงิน/adj	blue	hue=80
color	คราม/adj	indigoblue	hue=90
color	ฟ้า/adj	lightblue	hue=80:;lightness=3
color	น้ำตาล/adj	brown	hue=10:;lightness=-3
color	ม่วง/adj	purple	hue=90:;lightness=3
color	ขาว/adj	white	hue=200
color	ดำ/adj	black	hue=240
color	เทา/adj	gray	hue=220
color	ชมพู/adj	pink	hue=110:;lightness=3
color	บานเย็น/adj	pink	hue=100
color	ทอง/adj	gold	hue=15
color	เงิน/adj	silver	hue=200
color	ทองแดง/adj	bronze	hue=10:;lightness=-3
odour	หอม/adj	aromatic	hue=20
odour	เหม็น/adj	stink	hue=60
odour	ฉุน/adj	stink	hue=60
odour	หืน/adj	rancid	hue=100

ตารางผนวกที่ ข6 ส่วนควบคุมค่าของค่าแบบสัญลักษณ์

ความสัมพันธ์ทางคุณสมบัติ	ฐานคำ	ฟังชั่น
color	อม/adj	means
	แกม/adj	
taste	อม/adj	means

ภาคผนวก ค

ตัวย่อชนิดของคำ

ตัวย่อชนิดของคำ

เนื่องจากภาษาไทยมีชนิดของคำอยู่หลายชนิด เช่น คำนาม (Noun), คำกริยา (Verb) ดังนั้น ภายในวิทยานิพนธ์ฉบับนี้ จึงได้มีการอ้างอิงการแบ่งชนิดของคำภาษาไทยและการใช้คำย่อชนิดของคำ โดยใช้หลักเกณฑ์การแบ่งชนิดของคำ จากห้องปฏิบัติการวิจัยเชี่ยวชาญเฉพาะการประมวลผลภาษาธรรมชาติและเทคโนโลยีสารสนเทศอัจฉริยะ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ซึ่งทำการแบ่งชนิดของคำออกเป็น 15 ชนิด โดยคำนาม มีจำนวนทั้งหมด 16 ชนิด แสดงดังตารางภาคผนวกที่ ค1, คำกิริยามีจำนวน 8 ชนิด แสดงดังตารางภาคผนวกที่ ค2, คำบ่งชี้มีจำนวน 5 ชนิด แสดงดังตารางภาคผนวกที่ ค3, คำคุณศัพท์มีจำนวน 7 ชนิด แสดงดังตารางภาคผนวกที่ ค4, คำลักษณนามมีจำนวน 1 ชนิด แสดงดังตารางภาคผนวกที่ ค 5, คำสันธานมีจำนวน 3 ชนิด แสดงดังตารางภาคผนวกที่ ค6, คำบุพบทมีจำนวน 2 ชนิด แสดงดังตารางภาคผนวกที่ ค7, คำอุทานมีจำนวน 1 ชนิด แสดงดังตารางภาคผนวกที่ ค8, คำอุปสรรคมีจำนวน 3 ชนิด แสดงดังตารางภาคผนวกที่ ค9, คำลงท้ายมีจำนวน 2 ชนิด แสดงดังตารางภาคผนวกที่ ค10, คำปฏิเสธมีจำนวน 1 ชนิด แสดงดังตารางภาคผนวกที่ ค11, เครื่องหมายวรรคตอนมีจำนวน 1 ชนิด แสดงดังตารางภาคผนวกที่ ค12, สำนวนมีจำนวน 1 ชนิด แสดงดังตารางภาคผนวกที่ ค13, คำบ่งชี้กรรมวาจกมีจำนวน 1 ชนิด แสดงดังตารางภาคผนวกที่ ค14, สัญลักษณ์มีจำนวน 1 ชนิด แสดงดังตารางภาคผนวกที่ ค15

ตารางผนวกที่ ค1 ชนิดคำประเภทคำนามและตัวอย่าง

ชนิดของคำนาม	ตัวย่อ	ตัวอย่าง
Proper noun	npn	น้ำดอกไม้, ตาขาวปะเหลียน
Cardinal number	nnum	พัน, หมื่น, แสน, ล้าน
Ordinal Number Marker	norm	ที่
Label noun	nlab	1, 2, ก, ข
Common Noun	ncn	ช้าง, ม้า
Collective Noun	nct	ฝูง, พวก
Title Noun	ntit	นาย, นาง, นางสาว

ตารางผนวกที่ ค1 ชนิดคำประเภทคำนามและตัวอย่าง (ต่อ)

ชนิดของคำนาม	ตัวย่อ	ตัวอย่าง
Personal Pronoun	pper	เขา, คุณ, ท่าน
Demonstrative Pronoun	pdem	นี้, นั้น, นั่น
Indefinite Pronoun	pind	ใครๆ, ผู้ใด, ต่าง, บ้าง
Possessive Pronoun	ppos	ของคุณ, ของเรา
Reflexive Pronoun	pref	เอง
Reciprocal Pronoun	prec	กัน
Relative pronoun	prel	ที่, ซึ่ง, อัน
Interrogative Pronoun	pint	ทำไม, อะไร

ตารางผนวกที่ ค2 ชนิดคำประเภทคำกิริยาและตัวอย่าง

ชนิดของคำกิริยา	ตัวย่อ	ตัวอย่าง
Intransitive Verb	vi	เดิน, นั่ง, ซึม, กดดัน
Transitive Verb	vt	กรุณา, กลัว, กวนใจ
Causative Verb	vcau	ให้, ทำให้
Complementary State Verb	vcs	เป็น, อยู่ , คือ, กล่าวคือ
Existential Verb	vex	มี
Pre-Verb	prev	จะ, ยัง, คง
Post-verb	vpost	ไป, มา, ขึ้น, ลง
Honorific marker	honm	พระ, ทรง, พระราช

ตารางผนวกที่ ค3 ชนิดคำประเภทคำบ่งชี้และตัวอย่าง

ชนิดของคำบ่งชี้	ตัวย่อ	ตัวอย่าง
Determiner	det	นี้,นั้น
Quantity which always occur before noun	qube	ทั่ว,ทั่วทุก, นานา
Quantity which occur after noun	quaf	ทั่วไป, ต่าง ๆ,
Quantity which can occur both before and after noun	qubo	บาง, หลาย, อีก, ประมาณ,
Indefinite determiner	indet	ใด, อื่น

ตารางผนวกที่ ค4 ชนิดคำประเภทคำคุณศัพท์และตัวอย่าง

ชนิดของคำคุณศัพท์	ตัวย่อ	ตัวอย่าง
Adjective	adj	ขยัน, กำยำ, กิตติมศักดิ์
Adverb	adv	กลางคัน,กว่า, แรก, สุดท้าย
Adverb Marker1	advm1	อย่าง
Adverb Marker2	advm2	เป็น
Adverb Marker3	advm3	โดย
Adverb Marker4	advm4	สัก
Adjective Before Noun	Adjbe	ต่าง, ต่างพ่อต่างแม่,

ตารางผนวกที่ ค5 ชนิดคำประเภทคำลักษณนามและตัวอย่าง

ชนิดของคำลักษณนาม	ตัวย่อ	ตัวอย่าง
Classifier	cl	เชือก,เซนติเมตร, ทาง,ประเทศ ,ชิ้น etc.

ตารางผนวกที่ ค6 ชนิดคำประเภทคำสันธานและตัวอย่าง

ชนิดของคำสันธาน	ตัวย่อ	ตัวอย่าง
Conjunction	conj	และ, ในที่นี้
Double Conjunction	conjd	ทั้ง…และ, ไม่… ก็
Noun Clause Conjunction	conjncl	ว่า, ให้, ได้แก่ , เช่น

ตารางผนวกที่ ค7 ชนิดคำประเภทคำบุพบทและตัวอย่าง

ชนิดของคำบุพบท	ตัวย่อ	ตัวอย่าง
Preposition	prep	กับ โดย เมื่อ
co-Preposition	prepc	ระหว่าง....กับ

ตารางผนวกที่ ค8 ชนิดคำประเภทคำอุทานและตัวอย่าง

ชนิดของคำอุทาน	ตัวย่อ	ตัวอย่าง
Interjection	int	เอ๊ะ อ๋อ อุ๊ย ว้าย

ตารางผนวกที่ ค9 ชนิดคำประเภทคำอุปสรรคและตัวอย่าง

ชนิดของคำอุปสรรค	ตัวย่อ	ตัวอย่าง
Prefix1	pref1	การ,ความ
Prefix2	pref2	ผู้,นัก
Prefix3	pref3	ชาว

ตารางผนวกที่ ค10 ชนิดคำประเภทคำลงท้ายและตัวอย่าง

ชนิดของคำลงท้าย	ตัวย่อ	ตัวอย่าง
Affirmative	aff	ค่ะ, ครับ, จ้า
Particle	part	นัก, นั่นเอง

ตารางผนวกที่ ค11 ชนิดคำประเภทคำปฏิเสธและตัวอย่าง

ชนิดของคำปฏิเสธ	ตัวย่อ	ตัวอย่าง
Negative	neg	ไม่, มิ, ไร้

ตารางผนวกที่ ค12 ชนิดคำประเภทเครื่องหมายวรรคตอนและตัวอย่าง

ชนิดของเครื่องหมายวรรคตอน	ตัวย่อ	ตัวอย่าง
Punctuation	punc	. - , : ;

ตารางผนวกที่ ค13 ชนิดคำประเภทสำนวนและตัวอย่าง

ชนิดของสำนวน	ตัวย่อ	ตัวอย่าง
idiom	idm	รักวัวให้ผูก รักลูกให้ตี

ตารางผนวกที่ ค14 ชนิดคำประเภทคำบ่งชี้กรรมวาจกและตัวอย่าง

ชนิดของคำบ่งชี้กรรมวาจก	ตัวย่อ	ตัวอย่าง
Passive Voice Marker	psm	ถูก, โดน

ตารางผนวกที่ ค15 ชนิดคำประเภทสัญลักษณ์และตัวอย่าง

ชนิดของสัญลักษณ์	ตัวย่อ	ตัวอย่าง
Symbol	sym	ฯลฯ, ฯ

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นายอรรถพล คงหวาน
วัน เดือน ปี ที่เกิด	วันที่ 15 พฤษภาคม 2520
สถานที่เกิด	ยะลา
ประวัติการศึกษา	ประถมศึกษา โรงเรียนเทศบาล 5 วัดหาดใหญ่ใน (พ.ศ.2532) มัธยมศึกษาตอนต้น โรงเรียนหาดใหญ่วิทยาลัย (พ.ศ.2535) ปวช. (อิเล็กทรอนิกส์) วิทยาลัยเทคนิคหาดใหญ่ (พ.ศ.2538) ปวส. (เทคนิคคอมพิวเตอร์) สถาบันเทคโนโลยีราชมงคลวิทยาเขตภาคใต้ (พ.ศ.2540) วศ.บ. (วิศวกรรมคอมพิวเตอร์) คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีราชมงคล (พ.ศ.2543)
ตำแหน่งหน้าที่การงานปัจจุบัน	อาจารย์ 1 ระดับ 4
สถานที่ทำงานปัจจุบัน	มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย วิทยาเขตสงขลา
ผลงานดีเด่นและรางวัลทางวิชาการ	
ทุนการศึกษาที่ได้รับ	