ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์	วิศวกรรมคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง การแบ่งขอบเขตอนุพากย์ปริจเฉทในภาษาไทยโดยใช้คำระบุนัยและข้อสนเทศเชิงวากยสัมพันธ์

Thai Elementary Discourse Unit Segmentation by Using Discourse Segmentation Cues and Syntactic Information

นามผู้วิจัย นางสาวจิรวรรณ เจริญสุข

ได้พิจารณาเห็นชอบโดย

ประธานกรรมการ .....

(รองศาสตราจารย์อัศนีย์ ก่อตระกูล, D.Eng.)

กรรมการ .....

(อาจารย์จิตร์ทัศน์ ฝักเจริญผล, Ph.D.)

กรรมการ .....

(อาจารย์ยอดเยี่ยม ทิพย์สุวรรณ์, Ph.D.)

หัวหน้าภาควิชา .....

(อาจารย์พีรวัฒน์ วัฒนพงศ์, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์วินัย อาจคงหาญ, M.A.)

คณบดีบัณฑิตวิทยาลัย

วันที่ 5 **เดือน** เมษายน **พ.ศ.** 2549

วิทยานิพนธ์

เรื่อง

การแบ่งขอบเขตอนุพากย์ปริจเฉทในภาษาไทยโดยใช้คำระบุนัยและข้อสนเทศเชิงวากยสัมพันธ์

Thai Elementary Discourse Unit Segmentation by Using Discourse Segmentation Cues and Syntactic Information

โดย

นางสาวจิรวรรณ เจริญสุข

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2549

ISBN 974-16-1465-9

จิรวรรณ เจริญสุข 2549: การแบ่งขอบเขตอนุพากย์ปริจเฉทในภาษาไทยโดยใช้คำระบุนัยและข้อสนเทศเชิงวากยสัมพันธ์ ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) ภาควิชาวิศวกรรมคอมพิวเตอร์ ประธานกรรมการที่ปรึกษา: รองศาสตราจารย์อัศนีย์ ก่อตระกูล, D.Eng. 80 หน้า
ISBN 974-16-1465-9

อนุพากย์ปริจเฉท หมายถึง หน่วยที่เล็กที่สุดของการแบ่งข้อความในระดับปริจเฉทซึ่งอนุพากย์ปริจเฉทเหล่านี้จะต้องเป็นอนุพากย์อิสระ ที่สามารถสื่อความหมายได้อย่างสมบูรณ์ และข้อความในแต่ละหน่วยจะต้องไม่ทับซ้อนกัน จากคุณสมบัติดังกล่าวอนุพากย์ปริจเฉทโดยทั่วไปจึงมีโครงสร้างทางไวยากรณ์เป็นอนุประโยค หรือประโยคความเดียว แต่ในบางกรณีอนุพากย์ปริจเฉทสามารถมีโครงสร้างเป็นวลีได้ ทั้งนี้วลีเหล่านั้นจะต้องขึ้นต้นตัวเชื่อมปริจเฉทแบบชัดแจ้งบางตัวเท่านั้น เช่น "เพราะ" "เช่น" "ดังนั้น" เนื่องจากภาษาไทยเป็นภาษาที่เขียนเรียงคำต่อกันไปเรื่อยๆ โดยไม่มีข้อสนเทศใดเป็นตัวบ่งชี้ขอบเขตของแต่ละประโยค ดังนั้นการประมวลผลภาษาเพื่อหาขอบเขตของอนุพากย์ปริจเฉทจึงเป็นสิ่งสำคัญมาก โดยเฉพาะอย่างยิ่ง การประมวลผลภาษาธรรมชาติในระดับปริจเฉท เช่น การย่อความอัตโนมัติ การสกัดความรู้ ต้องมีกระบวนการแบ่งขอบเขตข้อความให้มีหน่วยเป็นอนุพากย์ปริจเฉท เพื่อใช้อนุพากย์เหล่านี้เป็นข้อมูลนำเข้าของระบบดังกล่าว

อย่างไรก็ตามกระบวนการแบ่งขอบเขตข้อความภายในเอกสารให้เป็นอนุพากย์ปริจเฉท ไม่สามารถทำได้โดยง่ายสำหรับภาษาไทย ทั้งนี้เนื่องมาจากคุณลักษณะของภาษาไทย 3 ประการ ปัญหาประการแรก คือปัญหาของภาษาไทยที่ไม่ปรากฏข้อสนเทศ ในการระบุจุดสิ้นสุดของอนุประโยคหรือประโยคเหมือนภาษาอื่นๆ เช่น มหัพภาคหรือจุด จุลภาคหรือจุดลูกน้ำ ปัญหาที่สอง คือ ปัญหาที่เกิดจากนิพจน์ระบุนาม ซึ่งนิพจน์ระบุนามของภาษาไทยมีโครงสร้างทางไวยากรณ์เหมือนกับอนุพากย์ปริจเฉทและปัญหาสุดท้าย คือ ปัญหาที่เกิดจากโครงสร้างของประโยคภาษาไทยที่สามารถละประธาน กริยา และกรรมของประโยคได้ ซึ่งปัญหาเหล่านี้เป็นปัญหาที่สำคัญที่ก่อให้เกิดความคลุมเครือในการระบุขอบเขตของอนุพากย์ปริจเฉท งานวิจัยนี้จึงทำการวิจัยและพัฒนาเทคนิคการระบุขอบเขตอนุภาคปริจเฉทภาษาไทยเพื่อแก้ปัญหาดังกล่าว โดยใช้หลักการผสมผสานระหว่างแนวทางการฝึกฝนและการเรียนรู้โดยเครื่องจักรกล ซึ่งใช้เทคนิคการเรียนรู้แบบต้นไม้ตัดสินใจร่วมกับแนวทางการใช้กฎจากผู้เชี่ยวชาญ ผลการทดลองการแบ่งขอบเขตอนุพากย์ภาษาไทยในโดเมนการเกษตร พบว่าระบบสามารถแบ่งขอบเขตอนุภาคปริจเฉท โดยมีค่าความถูกต้องเท่ากับ 0.80 และค่าความระลึก เท่ากับ 0.69

จิรวรรณ เจริญสุข
ลายมือชื่อนิสิต

อัศนีย์ ก่อตระกูล
ลายมือชื่อประธานกรรมการ

29 / 03 / 49

Jirawan Charoensuk 2006: Thai Elementary Discourse Unit Segmentation by Using Discourse Segmentation Cues and Syntactic Information. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Asanee Kawtrakul, D.Eng. 80 pages.
ISBN 974-16-1465-9

Elementary discourse unit (EDU) is the minimal discourse unit that was a production from discourse segmentation process. EDU should have independent unit that represent meaning full unit, and have non-overlapping unit. From these properties, EDU boundary is a clause or a simple sentence. In some case, EDU boundary might be phrase but the position of starting phrasal EDU must proceed with strong discourse markers such as "because", "for example" and "therefore". Since Thai language does not have special signals to identify sentence boundary, therefore, EDU segmentation is a significant process for discourse processing especially Text Summarization and Knowledge Extraction. These applications used EDU segmentation process to separated full text into EDU units and used these EDUs as inputs.

In additional, there are three major problems in Thai EDUs segmentation that cause EDU boundary ambiguity. Firstly, Thai does not have punctuation marks or special symbols to signal EDU boundary. Secondly, Thai Name Entity and EDU have the similar patterns. Finally, subject, verb and object could be omitted in Thai sentence. To solve these problems, this research developed and proposed a hybrid approach for Thai EDU segmentation by using decision-tree learning system and heuristic rules. The experiment in Thai agriculture domain shows that the precision and recall of the system are 0.80 and 0.69 respectively.

Jirawan Charoensuk Asanee Kawtrakul 29 / 03 / 06

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลงได้ด้วยความช่วยเหลือจากบุคคลหลายท่าน ข้าพเจ้าขอขอบพระคุณ รองศาสตราจารย์ ดร. อัศนีย์ ก่อตระกูล ประธานกรรมการที่ปรึกษา ผู้ให้แนวทางการวิจัย ให้ข้อเสนอแนะที่เป็นประโยชน์ อาจารย์ ดร. จิตร์ทัศน์ ฝักเจริญผล กรรมการที่ปรึกษาวิชาเอก อาจารย์ ดร. ยอดเยี่ยม ทิพย์สุวรรณ์ กรรมการที่ปรึกษาวิชารอง ที่กรุณาให้คำปรึกษา และข้อเสนอแนะที่มีคุณค่า เพื่อให้วิทยานิพนธ์นี้สมบูรณ์ยิ่งขึ้น

นอกจากนี้ข้าพเจ้าขอขอบพระคุณอาจารย์ธนา สุขวากรี ที่แนะแนวทางในการทำย่อความเอกสารภาษาไทย ขอขอบคุณอาจารย์อรวรรณ อิ่มสมบัติ ที่ให้คำแนะนำการใช้ซอฟต์แวร์ที่ใช้ในการฝึกฝนระบบ อีกทั้งขอขอบคุณ คุณมุกข์ดา สุขธาราจาร, คุณพัชรี วราศรัย และคุณอัจฉรา นภาโชติ ที่จัดเตรียมคลังเอกสารเพื่อใช้พัฒนางานวิทยานิพนธ์นี้ เพื่อนๆ และพี่ๆ ห้องปฏิบัติการวิจัยเชี่ยวชาญเฉพาะการประมวลผลภาษาธรรมชาติและเทคโนโลยีสารสนเทศอัจฉริยะ (NAiST Lab.) ทุกท่านที่ให้คำปรึกษาเกี่ยวกับการทำงานวิจัย และเจ้าหน้าที่ธุรการภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ทุกท่านสำหรับความช่วยเหลือด้านการประสานงาน และดำเนินการเอกสารต่างๆ

จิรวรรณ เจริญสุข

มีนาคม 2549

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

สารบัญภาพ (ต่อ)

การแบ่งขอบเขตอนุพากย์ปริจเฉทในภาษาไทยโดยใช้คำระบุนัยและข้อสนเทศเชิงวากยสัมพันธ์

Thai Elementary Discourse Unit Segmentation by Using Discourse Segmentation Cues and Syntactic Information

คำนำ

ปริจเฉท (discourse) หมายถึง ข้อความต่อเนื่องที่มิใช่เพียงการนำประโยคมาเรียงต่อกัน แต่ประโยคต้องมีสัมพันธภาพกัน (วิไลศักดิ์ ;สมทรง,2537) ดังนั้นการวิเคราะห์ปริจเฉท (discourse analysis) จึงเป็นสิ่งที่นักภาษาศาสตร์ในปัจจุบันให้ความสนใจเป็นอย่างมาก เนื่องจากการวิเคราะห์ดังกล่าวต้องทำการวิเคราะห์ทั้งในระดับไวยากรณ์ (syntactic level) และระดับความหมาย (semantic level) ของปริจเฉท โดยกระบวนการวิเคราะห์ปริจเฉทประกอบด้วย 3 ขั้นตอน คือ ขั้นตอนการแบ่งขอบเขตข้อความภายในเอกสาร (discourse segmentation) ออกเป็นหน่วยปริจเฉทย่อยๆ (discourse units), ขั้นตอนการหาชนิดปริจเฉทสัมพันธ์ (discourse relations) ระหว่างหน่วยดังกล่าว และขั้นตอนการสร้างรูปแทนโครงสร้างปริจเฉท (discourse representation) ดังนั้นขั้นตอนที่มีความสำคัญมากในการวิเคราะห์ปริจเฉท คือ ขั้นตอนการแบ่งขอบเขตข้อความภายในเอกสารออกเป็นหน่วยย่อยๆ เนื่องจากขั้นตอนนี้เป็นขั้นตอนแรกของการวิเคราะห์ปริจเฉท อีกทั้งยังเป็นขั้นตอนการสร้างข้อมูลนำเข้า (input) ให้กับขั้นตอนต่อไป ซึ่งหากแบ่งขอบเขตของอนุพากย์ปริจเฉทผิดพลาด จะส่งผลกระทบทำให้การหาชนิดปริจเฉทสัมพันธ์ผิดพลาด อีกทั้งยังส่งผลกระทบต่อเนื่องทำให้มีการสร้างรูปแทนโครงสร้างปริจเฉทผิดพลาดอีกด้วย ขั้นตอนการแบ่งขอบเขตข้อความภายในเอกสารออกเป็นหน่วยปริจเฉทย่อยๆ ซึ่งหน่วยปริจเฉทที่เล็กที่สุดของขั้นตอนการแบ่งดังกล่าว เรียกว่า อนุพากย์ปริจเฉท (Elementary Discourse Unit หรือ EDU) โดยอนุพากย์ปริจเฉทเหล่านี้จะต้องเป็นอนุพากย์อิสระ (independent) ที่สามารถสื่อความหมายได้อย่างสมบูรณ์ และข้อความในแต่ละหน่วยนั้นจะต้องไม่ทับซ้อนกัน (non-overlapping) อีกทั้งอนุพากย์ปริจเฉทเหล่านี้ต้องสามารถกำหนดชนิดปริจเฉทสัมพันธ์ และสถานะความสำคัญ (nuclearity status) ให้กับอนุพากย์ปริจเฉทแต่ละหน่วยได้ด้วย จากคุณสมบัติดังกล่าวอนุพากย์ปริจเฉทโดยทั่วไปจึงมีโครงสร้างทางไวยากรณ์เป็นอนุประโยค (clause) (Marcu, 1997) ในบางกรณีอนุพากย์ปริจเฉทสามารถมีโครงสร้างทางไวยากรณ์เป็นวลี (phrase) (Carlson et al. , 2001)

แต่วลีเหล่านั้นจะต้องขึ้นต้นด้วยตัวเชื่อมปริจเฉทแบบชัดแจ้ง (strong discourse marker) บางตัวเท่านั้น เช่น “เพราะ”, “อย่างไรก็ตาม” และ “ดังนั้น” เป็นต้น

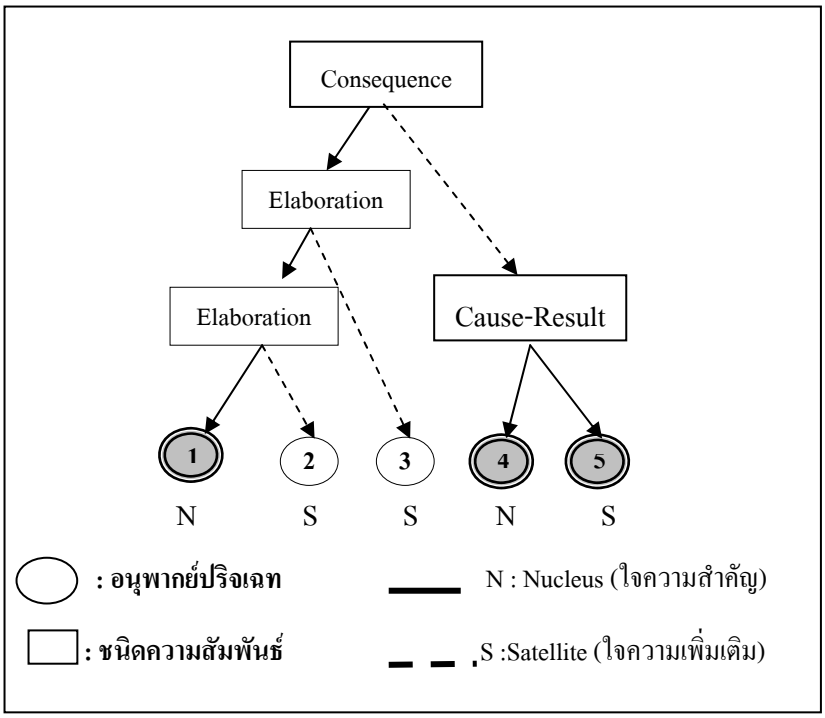
การแบ่งข้อความของเอกสารเป็นอนุพากย์ปริจเฉทมีประโยชน์ต่อการพัฒนาระบบประมวลผลเอกสารในระดับปริจเฉท เช่น การย่อความ (Text Summarization) และการสกัดความรู้ (Knowledge Discovery) เนื่องจากระบบเหล่านี้ต้องการใช้อนุพากย์ปริจเฉทเป็นข้อมูลนำเข้าของระบบ ตัวอย่างเช่น ระบบย่อความเอกสารที่ใช้ทฤษฎี Rhetorical structure theory : RST (Mann et al. , 2001) ซึ่งระบบการย่อความที่ใช้ทฤษฎีดังกล่าวจะประกอบด้วยขั้นตอนหลัก 3 ขั้นตอน โดยขั้นตอนแรก คือ ขั้นตอนการแบ่งขอบเขตข้อความเอกสารให้มีหน่วยเป็นอนุพากย์ปริจเฉท ซึ่งขั้นตอนนี้จะนำข้อความที่เป็นข้อมูลเข้าของระบบที่แสดงดังภาพที่ 1.1 มาแบ่งขอบเขตให้มีหน่วยเป็นอนุพากย์ปริจเฉทย่อยๆ จากข้อมูลเข้าดังกล่าวสามารถแบ่งขอบเขตของอนุพากย์ปริเฉทได้ทั้งหมด 5 อนุพากย์ แสดงดังภาพที่ 1.2 จากนั้นจึงนำอนุพากย์ปริจเฉทเหล่านี้ไปใช้เป็นข้อมูลเข้าสำหรับขั้นตอนที่ 2 ซึ่งเป็นขั้นตอนของการกำหนดชนิดปริจเฉทสัมพันธ์ (Rhetorical relation) และสถานะความสำคัญ (nucleus/satellite)ให้กับแต่ละอนุพากย์ปริจเฉท โดยระบบกำหนดชนิดของปริจเฉทสัมพันธ์และสถานะความสำคัญให้กับอนุพากย์ปริจเฉททั้งหมด และนำเสนอความสัมพันธ์ของอนุพากย์เหล่านี้ในรูปของโครงสร้างต้นไม้ (tree structure) ดังภาพที่ 1.3 ซึ่งการกำหนดชนิดปริจเฉทสัมพันธ์และสถานะความสำคัญให้กับแต่ละอนุพากย์นั้นเป็นวิธีการจัดลำดับความสำคัญให้กับแต่ละอนุพากย์ปริจเฉทให้กับระบบย่อความ และขั้นตอนสุดท้าย คือ ขั้นตอนการสกัดใจความสำคัญของเอกสาร โดยระบบจะสกัดอนุพากย์ปริจเฉทที่สำคัญออกมาเป็นผลลัพธ์ของการย่อความ โดยคำนวณระดับความสำคัญของแต่ละอนุพากย์จากสถานะความสำคัญและตำแหน่งของอนุพากย์บนโครงสร้างต้นไม้ จากข้อความที่เป็นข้อมูลเข้าของระบบ ส่วนที่เป็นใจสำคัญของข้อความนี้ คือ อนุพากย์ปริจเฉทที่ 1, 4 และ 5 แสดงดังภาพที่ 1.4 เพราะฉะนั้นขั้นตอนการแบ่งขอบเขตข้อความภายในเอกสารให้มีหน่วยเป็นอนุพากย์ปริจเฉทจึงเป็นขั้นตอนที่มีความสำคัญต่อระบบย่อความเป็นอย่างมาก

โรคนี้พบได้เกือบทุกระยะการเจริญเติบโต โดยพบมากในระยะที่กะหล่ำปลีห่อหัว ในระยะแรกพบเป็นจุด ต่อมาแผลจะขยายลุกลามออกไปทำให้เกิดการเน่าเละ

ภาพที่ 1. 1 ข้อมูลเข้า

[โรคนี้พบได้เกือบทุกระยะการเจริญเติบโต]$_{EDU1}$
[โดยพบมากในระยะที่กะหล่ำปลีห่อหัว]$_{EDU2}$
[ในระยะแรกพบเป็นจุด]$_{EDU3}$
[ต่อมาแผลจะขยายลุกลามออกไป]$_{EDU4}$
[ทำให้เกิดการเน่าเละ]$_{EDU5}$

ภาพที่ 1. 2 รายการอนุพากย์



ภาพที่ 1. 3 โครงสร้างต้นไม้

[โรคนี้พบได้เกือบทุกระยะการเจริญเติบโต]$_{EDU1}$
[ต่อมาแผลจะขยายลุกลามออกไป]$_{EDU4}$
[ทำให้เกิดการเน่าเละ]$_{EDU5}$

ภาพที่ 1.4 ผลลัพธ์การย่อความ

ภาพที่ 1 ตัวอย่างการย่อความ

อย่างไรก็ตามขั้นตอนการแบ่งขอบเขตข้อความภายในเอกสารให้เป็นอนุพากย์ปริจเฉท ไม่สามารถทำได้โดยง่ายสำหรับภาษาไทย ทั้งนี้เนื่องมาจากคุณลักษณะของภาษาไทย ซึ่งทำให้ต้องประสบปัญหาในการทำงานภายของระบบ โดยคุณลักษณะภาษาเหล่านี้ ได้แก่

1. ภาษาไทยไม่มีข้อสนเทศ ในการระบุจุดสิ้นสุดของอนุประโยคหรือประโยคเหมือนภาษาอื่นๆ เช่น มหัพภาคหรือจุด (“ . ”) , จุลภาคหรือจุดลูกน้ำ (“,”) , อัฒภาค (“;”)

2. นิพจน์ระบุนาม (Named Entity) และนามวลีของภาษาไทยมีโครงสร้างเหมือนกับอนุพากย์ปริจเฉท ทำให้เกิดความคลุมเครือในการระบุขอบเขตของอนุพากย์ปริจเฉท

3. โครงสร้างของประโยคภาษาไทยสามารถละประธาน, กริยา และกรรมของประโยคได้ ซึ่งทำให้เกิดปัญหาในการระบุขอบเขตของอนุพากย์ปริจเฉท

จากการที่ภาษาไทยมีลักษณะหลายประการที่แตกต่างจากภาษาอื่น จึงทำให้ไม่สามารถนำเทคนิคการระบุขอบเขตอนุภาคปริจเฉทของภาษาอื่นมาใช้ได้โดยตรง ดังนั้นวิทยานิพนธ์นี้ จึงมีขึ้นเพื่อศึกษาทฤษฎีต่างๆ และนำมาประยุกต์ เพื่อพัฒนาระบบแบ่งขอบเขตอนุภาคปริจเฉทสำหรับเอกสารภาษาไทย สำหรับวิทยานิพนธ์นี้ เป็นการใช้หลักการผสมผสานระหว่างแนวทางการฝึกฝนและการเรียนรู้โดยใช้เครื่องจักรกล (Machine learning) ร่วมกับการแนวทางการใช้กฎที่สร้างขึ้นโดยผู้เชี่ยวชาญ (Heuristic rules) เพื่อช่วยเพิ่มประสิทธิภาพของการระบุขอบเขตอนุภาคปริจเฉทภาษาไทยให้ถูกต้องมากยิ่งขึ้น

วัตถุประสงค์และขอบเขต

วัตถุประสงค์

1. เพื่อศึกษาปัญหาและวิเคราะห์การแบ่งขอบเขตของอนุพากย์ปริจเฉทสำหรับเอกสารภาษาไทย

2. เพื่อศึกษาและพัฒนาเทคนิคการแบ่งขอบเขตอนุพากย์ปริจเฉท เพื่อใช้เป็นข้อมูลเข้าของระบบประมวลผลภาษาธรรมชาติ เช่น การย่อความอัตโนมัติ และการสกัดความรู้

3. เพื่อพัฒนาระบบต้นแบบของการแบ่งขอบเขตอนุพากย์ปริจเฉทสำหรับภาษาไทย

ขอบเขตของงานวิจัย

1. วัดประสิทธิภาพและทดสอบระบบโดยใช้เอกสารภาษาไทยในโดเมนการเกษตร

2. ระบบสามารถทำงานได้กับเอกสารที่มีรูปแบบการเขียนที่ดี (well-style written)

การตรวจเอกสาร

ปริจเฉท

ปริจเฉท (Discourse) เป็นหน่วยภาษาในการสื่อสาร (functional unit) ที่เกิดจากการนำประโยคมาเรียงต่อกัน คุณสมบัติที่สำคัญของประโยคในปริจเฉท คือ ความต่อเนื่อง (coherence) ซึ่งเป็นความสัมพันธ์ที่เชื่อมโยงประโยคต่างๆ เข้าด้วยกัน เนื่องจากปริจเฉทไม่มีหน่วยทางวากยสัมพันธ์ที่สามารถให้นิยามทางด้านโครงสร้างได้อย่างชัดเจน จึงมีการให้นิยามปริจเฉทที่แตกต่างกันไปดังต่อไปนี้

Holiday *et al.* (1976) กล่าวว่าโครงสร้างของปริจเฉทมีระดับที่สูงกว่าประโยค เช่น ย่อหน้า บทในหนังสือ ซึ่ง Holiday ได้กล่าวเกี่ยวกับการเชื่อมโยงความว่าเป็นการเชื่อมโยงความของหน่วยต่างๆ เข้าไว้ด้วยกัน โดยการเชื่อมโยงดังกล่าวจะทำให้เกิดความสัมพันธ์กันภายในปริจเฉท

Hovy (1993) กล่าวว่า โครงสร้างปริจเฉท (Discourse Structure) คือ โครงสร้างเอกสารที่ประกอบด้วยอนุประโยคและประโยค โดยระหว่างอนุประโยคหรือประโยคด้วยกันจะประกอบด้วยความสัมพันธ์ระหว่างประโยคในเชิงความหมาย (Semantic Relation)

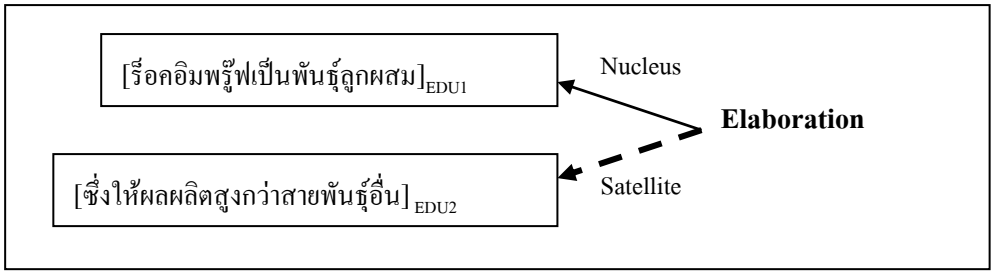
วิไลศักดิ์ (2549) นิยามว่า ปริจเฉท คือ ข้อความในระดับต่อเนื่อง ซึ่งไม่สามารถอธิบายได้ภายในกรอบของหน่วยภาษาระดับประโยค เพราะปริจเฉทมิใช่เพียงการนำประโยคมาเรียงต่อกัน แต่ประโยคเหล่านั้นต้องมีสัมพันธภาพกันด้วย และปริจเฉทถูกจำแนกออกเป็นชนิดต่างๆ โดยมีเกณฑ์ตัดสินที่แตกต่างกันออกไปตามที่นักภาษาศาสตร์แต่ละคนที่ได้ทำการศึกษา เช่น ปฏิบัติปริจเฉท, สนทนาปริจเฉท และปราศรัยปริจเฉท เป็นต้น

จากคำนิยามของปริจเฉทต่างๆข้างต้น สามารถสรุปนิยามของปริจเฉทดังนี้ ปริจเฉทหมายถึง ข้อความต่อเนื่องที่มิใช่เพียงการนำประโยคมาเรียงต่อกัน แต่ข้อความเหล่านั้นจะต้องมีสัมพันธภาพ

ในปัจจุบันนี้นักภาษาศาสตร์มีความสนใจการวิเคราะห์ปริจเฉท (discourse analysis) อย่างมาก เนื่องจากการวิเคราะห์ดังกล่าวต้องทำการวิเคราะห์ทั้งในระดับไวยากรณ์ (syntactic level) และระดับความหมาย (semantic level) ของปริจเฉท กระบวนการวิเคราะห์ปริจเฉทประกอบด้วย 3 ขั้นตอน คือ ขั้นตอนการแบ่งขอบเขตข้อความภายในเอกสาร (discourse segmentation) ออกเป็นหน่วยปริจเฉทย่อยๆ (discourse units), ขั้นตอนหาชนิดปริจเฉทสัมพันธ์ (discourse relations) ระหว่างหน่วยดังกล่าว และขั้นตอนสร้างรูปแทนโครงสร้างปริจเฉท (discourse representation) ดังนั้นขั้นตอนที่มีความสำคัญมากในการวิเคราะห์ปริจเฉท คือ ขั้นตอนการแบ่งขอบเขตข้อความภายในเอกสารออกเป็นหน่วยย่อยๆ เนื่องจากขั้นตอนนี้เป็นขั้นตอนแรกของการวิเคราะห์ปริจเฉท อีกทั้งยังเป็นขั้นตอนการสร้างข้อมูลนำเข้า (input) ให้กับขั้นตอนต่อไป ซึ่งหากมีการแบ่งขอบเขตของแต่ละหน่วยปริจเฉทผิดพลาด จะส่งผลกระทบทำให้การหาชนิดปริจเฉทสัมพันธ์ผิดพลาด และยังส่งผลกระทบต่อเนื่องทำให้มีการสร้างรูปแทนโครงสร้างปริจเฉทผิดพลาดอีกด้วย

อนุพากย์ปริจเฉท

อนุพากย์ปริจเฉท (Elementary Discourse Unit หรือ EDU) คือ หน่วยที่เล็กที่สุดของการแบ่งข้อความในระดับปริจเฉท (Marcu, 1997) ซึ่งอนุพากย์ปริจเฉทเหล่านี้จะเป็นอนุพากย์อิสระที่สามารถสื่อความหมายได้สมบูรณ์ และข้อความในแต่ละหน่วยจะต้องไม่ทับซ้อนกัน อีกทั้งอนุพากย์ปริจเฉทเหล่านี้ต้องสามารถกำหนดชนิดปริจเฉทสัมพันธ์ (rhetorical relation) (Mann, 1988) เช่น เหตุผล (cause-result) ขัดแย้ง (contrast) และเงื่อนไข (condition) และสถานะความสำคัญให้กับอนุพากย์ปริจเฉทแต่ละหน่วยได้ ซึ่งสถานะความสำคัญนี้สามารถแบ่งออกเป็น 2 ชนิด คือ สถานะนิวเคลียส (Nucleus) เป็นสถานะที่บ่งบอกว่าอนุพากย์ปริจเฉทนี้เป็นอนุพากย์ที่มีใจความสำคัญหลักปรากฏอยู่ และถ้าข้อความนี้ถูกลบออกไปจะมีผลทำให้อีกอนุพากย์หนึ่งที่สัมพันธ์กันอยู่นั้นไม่สามารถแสดงความหมายที่สมบูรณ์ตามชนิดปริจเฉทสัมพันธ์นั้นไว้ได้ ส่วนสถานะแซตเทินซ์ไรท์ (Satellite) เป็นสถานะของอนุพากย์ปริจเฉทที่เป็นส่วนขยายของอนุพากย์หลัก และอนุพากย์นี้ต้องพึ่งพารายละเอียดจากอนุพากย์หลัก ถึงแม้จะลบอนุพากย์นี้ออกไป ก็จะไม่กระทบต่อความหมายที่แท้จริงของอนุพากย์หลักตามชนิดปริจเฉทสัมพันธ์นั้นแต่อย่างใด จากตัวอย่างดังภาพที่ 2 อนุพากย์ปริจเฉททั้งสองอนุพากย์มีความสัมพันธ์กัน โดยมีชนิดปริจเฉทความสัมพันธ์แบบ Elaboration โดยที่อนุพากย์ปริจเฉทที่ 1 มีสถานะความสำคัญเป็นนิวเคลียสและอนุพากย์ปริจเฉทที่ 2 มีสถานะความสำคัญเป็นแซตเทินซ์ไรท์



<u>ภาพที่ 2</u> ตัวอย่างชนิดปริจเฉทสัมพันธ์และสถานะความสำคัญ

จากคุณสมบัติดังกล่าวอนุพากย์ปริจเฉทโดยทั่วไปจะมีโครงสร้างทางไวยากรณ์เป็นอนุประโยคหรือประโยคความเดียว เช่น “โรคจะระบาดมากในภาคกลาง” ในบางกรณีอนุพากย์ปริจเฉทสามารถมีโครงสร้างเป็นวลีได้ ทั้งนี้วลีเหล่านั้นจะต้องขึ้นต้นด้วยตัวเชื่อมปริจเฉทแบบชัดแจ้งบางตัวเท่านั้น เช่น “เนื่องจาก” “ดังนั้น” “เช่น” ตัวอย่างของอนุพากย์ปริเฉทที่มีโครงสร้างทางไวยากรณ์เป็นวลี เช่น “**<u>เช่น</u>** นครปฐม ปทุมธานี” โดยคำว่า “เช่น” เป็นตัวเชื่อมปริจเฉทแบบชัดแจ้งที่สื่อถึงชนิดปริจเฉทความสัมพันธ์แบบ Elaboration

ดังนั้นการแบ่งขอบเขตข้อความให้มีหน่วยเป็นอนุพากย์ปริจเฉทมักจะทำการแบ่งขอบเขตโดยใช้ข้อมูลนำเข้ามีหน่วยเป็นประโยค ซึ่งประโยคภาษาไทยสามารถแบ่งตามลักษณะและจำนวนของหน่วยประโยคได้ 3 ประเภท คือ ประโยคความเดียว หรือ เอกรรถประโยค (Simple sentence), ประโยคความรวม หรือ อเนกรรถประโยค (Compound sentence) และประโยคความซ้อน หรือ สังกรประโยค (Complex sentence)

1. ประโยคความเดียว

คือ ประโยคสามัญที่มีความหมายเพียงอย่างเดียว ซึ่งประกอบด้วย ภาคประธาน และภาคกริยา อย่างละ 1 หน่วย ส่วนกรรมตรง กรรมรอง ที่อยู่หลังกริยาจะมีหรือไม่มีก็ได้ ตัวอย่างประโยคความเดียวแสดงดังตารางที่ 1

ประธาน กริยา (กรรมตรง) (กรรมรอง)*

ตารางที่ 1 ตัวอย่างประโยคความเดียว

ประโยค	ภาคประธาน	ภาคกริยา	หมายเหตุ
ฝนตก	ฝน	ตก	ประโยคที่กริยาไม่ต้องมีกรรมมารับ
เพลี้ยกัดใบ	เพลี้ย	กัดใบ	กัด =กริยา ใบ = กรรม

2. ประโยคความรวม

คือ ประโยคใหญ่ที่มีความหมายมากกว่าหนึ่งความหมาย หรือ เป็นประโยคที่รวมเอาประโยคความเดียวตั้งแต่ 2 ประโยคขึ้นไปมารวมกัน โดยอาจมีสันธาน เช่น “และ” “ทั้ง” “หรือ” และ “จึง” เป็นบทเชื่อมระหว่างประโยคความเดียวเหล่านั้นอยู่ด้วย หรืออาจจะละสันธานไว้ในฐานเข้าใจ ซึ่งนววรรณ (2527) แบ่งชนิดของประโยคความรวมตามสันธานได้ 2 แบบ คือ

2.1 ประโยคความรวมที่มีสันธานเป็นบทเชื่อม ประกอบด้วยประโยคตั้งแต่ 2 ประโยคขึ้นไปและมีหน่วยเชื่อมอยู่ระหว่างประโยค ตัวอย่างประโยคความรวมที่มีสันธานแสดงดังตารางที่ 2

ตารางที่ 2 ตัวอย่างประโยคความรวมที่มีสันธานเป็นบทเชื่อม

ประโยคความรวม	ประโยคความเดียว	ประโยคความเดียว	สันธาน
เมื่อกะหล่ำปลีโตเต็มที่แล้ว ควรลดปริมาณน้ำลง	เมื่อกะหล่ำปลีโตเต็มที่แล้ว	ควรลดปริมาณน้ำลง	เมื่อ
หัวของแมลงมีสีแดงและด้านข้างมีลายสีเงิน	หัวของแมลงมีสีแดง	ด้านข้างมีลายสีเงิน	และ

2.2 ประโยคความรวมที่ไม่มีสันธานเป็นบทเชื่อม จะมีหน่วยกริยาเรียงต่อกันตั้งแต่ 2 หน่วยขึ้นไป โดยอาจมีการลดคำหรือละคำส่วนใดส่วนหนึ่งภายในประโยค และลักษณะพิเศษของภาษาไทยอย่างหนึ่งคือ คำกริยาที่เรียงต่อกันหลายคำ คำที่เรียงนั้นบางกรณีอาจเป็นหน่วยเดียวกันหรือคนละหน่วยก็ได้ ตัวอย่างของประโยคความรวมแสดงดังประโยคข้างล่าง

ประโยคที่ 1 “เกษตรกรนั่งเล่น” เป็นประโยคความเดียว เนื่องจากคำกริยาหลัก คือ นั่ง และคำกริยาที่ใช้ขยาย คือ เล่น

ประโยคที่ 2 “เกษตรกรนั่งยิ้ม” เป็นประโยคความรวม เนื่องจาก มีหน่วยกริยา 2 หน่วย คือ นั่งและยิ้ม

ประโยคความรวมที่มีหน่วยกริยา 2 หน่วยจะมีรูปแบบโครงสร้างประโยคดังนี้
ประธาน กริยา กริยา (กรรมตรง) (กรรมรอง)*
หรือ ประธาน กริยา กรรมตรง กริยา (กรรมตรง) (กรรมรอง)*

ตัวอย่างของประโยคความรวมที่ไม่มีสันธานแสดงดังตารางที่ 3

ตารางที่ 3 ตัวอย่างประโยคที่ไม่มีสันธานเป็นบทเชื่อม

ประโยคความรวม	ประโยคความเดียว	ประโยคความเดียว
ข้าวหอมมะลิหวานอร่อย	ข้าวหอมมะลิหวาน	ข้าวหอมมะลิอร่อย
เพลี้ยกัดกินใบ	เพลี้ยกัดใบ	เพลี้ยกินใบ

เนื่องจากประโยคความรวมต้องประกอบด้วยประโยคตั้งแต่ 2 ประโยคขึ้นไปมารวมกัน ประโยคที่นำมารวมกันอาจเกิดจากการรวมประโยคระหว่าง ประโยคความเดียว, ประโยคความรวม หรือประโยคความซ้อน (กำชัย, 2545) ซึ่งตัวอย่างของการรวมประโยคแสดงดังตารางที่ 4

3. ประโยคความซ้อน

คือ ประโยคที่ประกอบด้วยประโยคหลัก (มุขยประโยค) ที่เป็นเนื้อความสำคัญของประโยคและประโยคย่อย (อนุประโยค) ซึ่งส่วนขยายหรือประกอบประโยคหลัก มารวมเป็นประโยคเดียวกัน โดยมีประพันธสรรพนาม (“ผู้” “ที่” “ซึ่ง” และ “อัน”) ประพันธวิเศษณ์หรือบุพบทเป็นบทเชื่อม (นววรรณ, 2527) ตัวอย่างของประโยคความซ้อนแสดงดังตารางที่ 5

<u>ตารางที่ 4</u> ตัวอย่างการรวมประโยค

ประโยคความรวม	ประโยคที่ 1	ประโยคที่ 2	ประเภทของประโยค
หัวของแมลงมีสีแดงและด้านข้างมีลายสีเงิน	หัวของแมลงมีสีแดง	ด้านข้างมีลายสีเงิน	ประโยคความเดียว + ประโยคความเดียว
แมลงเป็นศัตรูข้าวและเพลี้ยที่ทำลายข้าวจะพบมากในฤดูร้อน	แมลงเป็นศัตรูข้าว	เพลี้ยที่ทำลายข้าวจะพบมากในฤดูร้อน	ประโยคความเดียว + ประโยคความซ้อน
เพลี้ยกระโดดสีน้ำตาลและเพลี้ยไฟพริกเป็นศัตรูของข้าวแต่เพลี้ยจักจั่นที่มีสีเขียวเป็นศัตรูของกะหล่ำปลี	เพลี้ยกระโดดสีน้ำตาลและเพลี้ยไฟพริกเป็นศัตรูของข้าว	เพลี้ยจักจั่นที่มีสีเขียวเป็นศัตรูของกะหล่ำปลี	ประโยคความรวม + ประโยคความซ้อน

<u>ตารางที่ 5</u> ตัวอย่างประโยคความซ้อน

ประโยคความซ้อน	ประโยคหลัก	ประโยคย่อย	ตัวเชื่อม
เพลี้ยที่ทำลายข้าวพบมากในฤดูร้อน	เพลี้ยพบมากในฤดูร้อน	เพลี้ยที่ทำลายข้าว	ที่ (แทนคำว่า "เพลี้ย")
กะหล่ำปลีเป็นพืชซึ่งปลูกมากในแถบเมดิเตอร์เรเนียน	กะหล่ำปลีเป็นพืช	ซึ่งปลูกมากในแถบเมดิเตอร์เรเนียน	ซึ่ง (แทนคำว่า "พืช")

ประโยคความซ้อนสามารถแบ่งตามตำแหน่งการขยายประโยคได้ 3 รูปแบบ คือ ประโยคความซ้อนที่มีประโยคขยายในหน่วยนาม ,หน่วยกริยา และหน่วยนามและหน่วยกริยา

<u>ประโยคความซ้อนที่มีประโยคขยายในหน่วยนาม</u> ประโยคขยายในส่วนนามจะมีคำเชื่อม คือ “ที่” “ซึ่ง” และ “อัน” นำหน้าเสมอ คำเชื่อม “ที่” “ซึ่ง” และ “อัน” นี้จะสามารถใช้แทนกันได้ในบางกรณีเท่านั้น แต่เราจะพบว่า ประโยคขยายที่มีคำว่า “ที่” นำหน้ามีจำนวนมากกว่าประโยคขยายที่มีคำว่า “ซึ่ง” และ “อัน” โดยปกติประโยคขยายที่มี “ที่” “ซึ่ง” และ “อัน” นำหน้า จะเป็นประโยคที่ใช้จำกัดความหมายหรือพูดอย่างใดอย่างหนึ่ง ซึ่งเป็นประโยคที่ไม่สมบูรณ์ เนื่องจากคำนามซึ่งซ้ำกับหน่วยหลักไม่ได้ปรากฏ (ถูกละคำไป) ถ้าคำนามนั้นปรากฏขึ้นจะถือว่าผิดไวยากรณ์ ดังตัวอย่างในตารางที่ 6

ตารางที่ 6 ตัวอย่างประโยคความซ้อนที่มีประโยคขยายในหน่วยนาม

ประโยคความซ้อน	ประโยคหลัก	ประโยคย่อย	คำนามที่ถูกขยาย	ตัวเชื่อม
กะหล่ำปลีที่นิยมปลูกทางภาคเหนือเป็นพันธุ์หนัก	กะหล่ำปลีเป็นพันธุ์หนัก	กะหล่ำปลีที่นิยมปลูกทางภาคเหนือ	กะหล่ำปลี	ที่
โรคเน่าดำเกิดจากเชื้อราซึ่งเข้าทำลายทางรูใบ	โรคเน่าดำเกิดจากเชื้อรา	เชื้อราซึ่งเข้าทำลายทางรูใบ	เชื้อรา	ซึ่ง

ประโยคความซ้อนที่มีประโยคขยายในหน่วยกริยา ประโยคขยายในหน่วยกริยาที่มีคำเชื่อมว่า “ที่” และ “ว่า” นำหน้า ประโยคขยายไม่จำเป็นต้องอยู่ติดกับส่วนหลักเสมอ อาจจะอยู่ท้ายประโยค หรือมีสิ่งอื่นคั่นระหว่างส่วนหลักหรือส่วนขยายก็ได้ โดยที่คำเชื่อม “ว่า” มักจะใช้เมื่อคำกริยาที่เป็นส่วนหลักเป็นคำกริยาที่ต้องการประโยคมาช่วยเสริมความหมายให้สมบูรณ์ ได้แก่ กริยาเกี่ยวกับอารมณ์ การสื่อสาร การคิด และประสาทสัมผัส ส่วนคำเชื่อม “ที่” มักจะใช้กับคำกริยาเกี่ยวกับอารมณ์และแสดงอาการ ตัวอย่างแสดงดังตารางที่ 7

ตารางที่ 7 ตัวอย่างประโยคความซ้อนที่มีประโยคขยายในหน่วยกริยา

ประโยคความซ้อน	ประโยคหลัก	ประโยคย่อย	คำกริยาที่ถูกขยาย	ตัวเชื่อม
เกษตรกรหวังว่า โรคเน่าดำจะไม่ระบาด	เกษตรกรหวัง	ว่าโรคเน่าดำจะไม่ระบาด	หวัง	ว่า
เกษตรกรดีใจที่ราคาข้าวไม่ตกต่ำ	เกษตรกรดีใจ	ที่ราคาข้าวไม่ตกต่ำ	ดีใจ	ที่

ประโยคความซ้อนที่มีประโยคขยายในหน่วยนามและหน่วยกริยา ทั้งหน่วยนามและกริยาในประโยคความซ้อนอาจมีประโยคขยาย และในหน่วยนามและกริยาในส่วนขยายเองอาจมีประโยคขยายอยู่ด้วย ทำให้ประโยคมีความซับซ้อนยิ่งขึ้น ตัวอย่างแสดงดังตารางที่ 8

<u>ตารางที่ 8</u> ตัวอย่างประโยคความซ้อนที่มีประโยคขยายในหน่วยนามและหน่วยกริยา

ประโยคความซ้อน	ประโยคหลัก	ประโยคย่อยที่ 1	ประโยคย่อยที่ 2	ตัวเชื่อม
กะหล่ำปลีที่ปลูกทางภาคเหนือเป็นพันธุ์หนักซึ่งนิยมบริโภคภายในประเทศ	กะหล่ำปลีเป็นพันธุ์หนัก	ที่ปลูกทางภาคเหนือ	ซึ่งนิยมบริโภคภายในประเทศ	ที่ ,ซึ่ง
เพลี้ยเป็นพาหนะนำเชื้อวิสาซึ่งทำให้เกิดโรคใบหงิกที่เกิดขึ้นเฉพาะกับข้าวพันธุ์ กข	เพลี้ยเป็นพาหนะนำเชื้อวิสา	ซึ่งทำให้เกิดโรคใบหงิก	ที่เกิดขึ้นเฉพาะกับข้าวพันธุ์ กข	ซึ่ง ,ที่

หลักเกณฑ์การแบ่งขอบเขตของอนุพากย์ปริจเฉท

หลักเกณฑ์เบื้องต้นในการแบ่งขอบเขตของอนุพากย์ปริจเฉทสำหรับภาษาอังกฤษ จะพิจารณาจากโครงสร้างไวยากรณ์ (Carlson et al., 2003) ของอนุพากย์เป็นหลัก ซึ่งอนุพากย์ปริจเฉทจะมีโครงสร้างไวยากรณ์เป็นอนุประโยคหรือประโยคความเดียว (Marcu, 1997) ต่อมาในปี ค.ศ. 2001 Carlson (Carlson et al., 2001) ได้กำหนดหลักเกณฑ์ในการแบ่งขอบเขตอนุพากย์ปริเฉทเพิ่มขึ้น โดยกำหนดให้อนุพากย์ปริจเฉทสามารถมีโครงสร้างไวยากรณ์เป็นวลี แต่วลีเหล่านั้นจะต้องนำหน้าวลีด้วยเซตของตัวเชื่อมปริจเฉทแบบชัดแจ้ง ซึ่งเซตของตัวเชื่อมปริจเฉทแบบชัดแจ้งแสดงดังภาพที่ 3

คำระบุนัยที่ไม่มีความคลุมเครือ ∈ {“เพราะ” (because), “ทั้งๆที่” (in spite of), “ถึงอย่างไรก็ตาม” (despite), “ไม่คำนึงถึง” (regardless), “ไม่คำนึงถึง” (irrespective), “โดยปราศจาก” (without), “ตามที่” (according to), “ด้วยเหตุที่” (as a result of,) “ไม่เพียงแต่...ยัง” (not only ... but also)}

<u>ภาพที่ 3</u> เซตของตัวเชื่อมปริจเฉทแบบชัดแจ้งสำหรับภาษาอังกฤษ

จากหลักเกณฑ์ในการแบ่งขอบเขตอนุพากย์ปริจเฉทของ Lynn Carlson สามารถแบ่งชนิดของอนุพากย์ปริจเฉทตามลักษณะการปรากฏของแต่ละอนุพากย์ได้ 2 ชนิด คือ EDU และ Embedded EDU โดยที่ EDU จะเป็นอนุพากย์ปริเฉทที่มีโครงสร้างทางไวยากรณ์เป็นอนุประโยค

หรือวลีดังที่กล่าวมาแล้ว ส่วน Embedded EDU จะเป็นอนุพากย์ปริจเฉทที่มีโครงสร้างเหมือนกับ EDU แต่ว่าตำแหน่งของ Embedded EDU จะแทรกกลางอยู่ภายใน Basic EDU ทำให้ EDU ถูกแบ่งออกเป็น 2 ส่วน ซึ่งตัวอย่างของ EDU และ Embedded EDU ของภาษาอังกฤษแสดงดังตารางที่ 9

ตารางที่ 9 ตัวอย่างของ EDU และ Embedded EDU ของภาษาอังกฤษ

ชนิดของ EDU	ตัวอย่างประโยค
EDU	[The company will shut down its plant]$_{EDU1}$ [although it will not dismiss any employees.]$_{EDU2}$
Embedded EDU	[The plant]$_{EDU1.1}${that the company will shutdown}$_{Embedded\ EDU}$[is in Ohio]$_{EDU1.2}$

งานวิจัยที่เกี่ยวข้อง

จากงานวิจัยที่ผ่านมาได้มีการนำข้อสนเทศทางภาษาต่างๆ มาใช้ในการแบ่งขอบเขตของอนุพากย์ปริจเฉทจากข้อความภายในเอกสารภาษาต่างประเทศ เช่น ภาษาอังกฤษ และภาษาสเปน โดยอาศัยข้อสนเทศทางภาษา ซึ่งประกอบด้วย คำระบุนัย (Discourse segmentation cues) (Marcu, 1997, 1998; Alonso et al., 2001; Carlson et al. 2003) เครื่องหมายวรรคตอน (punctuation marks) (Marcu, 1998; Carlson et al. 2003) และข้อมูลจากโครงสร้างไวยากรณ์ (syntactic information) ในระดับต่างๆ เช่น ระดับคำ (Marcu, 1999) ระดับประโยค (Soricut et al.,2003; Polanyi et al.,2004; Thanh et al.,2004) มาใช้เป็นเครื่องมือประกอบการตัดสินใจในการแบ่งขอบเขตและกำหนดชนิดให้กับแต่ละอนุพากย์ปริจเฉท จากงานวิจัยที่ผ่านมาสามารถแบ่งแนวทางในการแบ่งขอบเขตและการกำหนดชนิดของอนุพากย์ปริจเฉทจากการใช้ข้อสนเทศทางภาษาเหล่านี้ได้เป็นกลุ่มได้ 2 กลุ่ม ได้แก่ แนวทางการฝึกฝนและการเรียนรู้โดยใช้เครื่องจักรกล (Marcu, 1999; Soricut et al.,2003) และแนวทางการใช้กฎที่สร้างขึ้นโดยผู้เชี่ยวชาญ (Marcu, 1998; Alonso et al., 2001; Thanh et al.,2004) ซึ่งข้อสนเทศทางภาษาของแต่ละแนวทางแสดงดังตารางที่ 10

ตารางที่ 10 แนวทางการฝึกฝนและการเรียนรู้โดยใช้เครื่องจักรกลและแนวทางการใช้กฎที่สร้างขึ้นโดยผู้เชี่ยวชาญ

ปี	นักวิจัย	ภาษา	ประสิทธิภาพ		ข้อสนเทศที่ใช้
			ค่าความถูกต้อง (%)	ค่าความระลึก (%)	
แนวทางการฝึกฝนและการเรียนรู้โดยใช้เครื่องจักรกล					
1999	D. Marcu	อังกฤษ	60.45	-	คำระบุนัย,เครื่องหมายวรรคตอน, ชนิดของคำ
2003	R. Soricut et al.	อังกฤษ	83.5	82.7	ชนิดของคำ, รูปคำ, การแจกแจงประโยค
แนวทางการใช้กฎที่สร้างขึ้นโดยผู้เชี่ยวชาญ					
1998	D. Marcu	อังกฤษ	90.3	-	คำระบุนัย, เครื่องหมายวรรคตอน
2004	H. L. Thanh et al.	อังกฤษ	81.43	79.26	คำระบุนัย, เครื่องหมายวรรคตอน, การแจกแจงประโยค

แนวทางที่ 1: การแบ่งขอบเขตอนุพากย์ปริจเฉทด้วยแนวทางการฝึกฝนและการเรียนรู้โดยใช้เครื่องจักรกล

แนวทางการแบ่งขอบเขตของอนุพากย์ปริจเฉทในกลุ่มนี้ เป็นระบบที่สร้างขึ้นโดยใช้การฝึกฝนและการเรียนรู้จากเครื่องจักรการเรียนรู้ (Machine learning) ระบบในกลุ่มนี้จะใช้ข้อสนเทศทางภาษา ได้แก่ คำระบุนัย, เครื่องหมายวรรคตอน, ข้อมูลจากโครงสร้างไวยากรณ์ในระดับคำ (Marcu, 1999) และข้อมูลจากโครงสร้างไวยากรณ์ในระดับประโยค (Soricut et al. , 2003) เป็นคุณสมบัติของเครื่องจักรกลในการตัดสินใจในการแบ่งขอบเขตและกำหนดชนิดให้กับอนุพากย์ปริจเฉท

ในงานของ Marcu (1999) เป็นงานวิจัยที่เลือกใช้แบบจำลองต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคสำหรับฝึกฝนและเรียนรู้ ซึ่งคุณสมบัติ (feature) ที่ใช้พิจารณาในกระบวนการแบ่งขอบเขตของอนุพากย์ปริจเฉทของงานวิจัยนี้ คือ ข้อสนเทศทางภาษา ได้แก่ คำระบุนัย, เครื่องหมายวรรคตอน และชนิดของคำ (Part of Speech หรือ POS) ซึ่งคุณสมบัติเหล่านี้เป็นการนำข้อมูลจาก

โครงสร้างไวยากรณ์ในระดับคำมาใช้งาน โดยคุณสมบัติที่ใช้ในการแบ่งขอบเขตอนุพากย์ปริจเฉท มีทั้งหมดสองคุณสมบัติ คือ คุณสมบัติชนิดของคำ โดยพิจารณาจากชนิดของคำที่เป็นบริบทรอบข้างในขอบเขต ± 5 คำ และคุณสมบัติคำระบุนัยและเครื่องหมายวรรคตอน โดยจะพิจารณาว่าคำนั้นมีคุณสมบัติเป็นคำระบุนัยหรือเครื่องหมายวรรคตอนหรือไม่ ซึ่งงานวิจัยนี้ได้นำเอาคุณสมบัติทั้งสองคุณสมบัติมาร่วมในการพิจารณาแบ่งขอบเขตอนุพากย์ปริจเฉท โดยผลการทดลองของงานวิจัยนี้มีความถูกต้อง 60.45 %

ข้อเด่น

1. คุณสมบัติที่เลือกใช้ในแบบจำลองในงานวิจัยของ D. Marcu ใช้เพียงข้อมูลจากโครงสร้างไวยากรณ์ระดับคำเท่านั้น
2. ระบบสามารถแบ่งขอบเขตอนุพากย์ปริจเฉทที่ไม่ปรากฏคำระบุนัยหรือเครื่องหมายวรรคตอนได้

ข้อด้อย

1. ค่าความถูกต้องของระบบมีค่าไม่สูงมาก ปัญหานี้เกิดจากความผิดพลาดในขั้นตอนการเตรียมคลังเอกสารที่ใช้ในการฝึกฝนแบบจำลอง
2. คุณสมบัติที่ใช้ในแบบจำลองมีการกระจายตัวข้อมูลมากเกินไป ทำให้แบบจำลองไม่สามารถสร้างกฎที่ครอบคลุมการแบ่งขอบเขตอนุพากย์ปริจเฉทได้ทุกกรณีได้
3. งานวิจัยของ D. Marcu ไม่ได้ระบุวิธีการแบ่งขอบเขตอนุพากย์ปริจเฉทที่มีโครงสร้างทางไวยากรณ์เป็นวลี

Soricut et al. (2003) เป็นงานวิจัยที่เลือกใช้แบบจำลองการเรียนรู้เบย์อย่างง่าย (Naïve Bayes) เป็นเทคนิคสำหรับฝึกฝนและเรียนรู้ในการแบ่งขอบเขตอนุพากย์ปริเฉท ซึ่งงานวิจัยนี้เลือกใช้ข้อสนเทศทางภาษาทั้งในระดับคำ ได้แก่ รูปผิวของคำ (surface of word) ชนิดของคำ และข้อมูลจากผลลัพธ์ของการแจกแจงประโยค ซึ่งข้อสนเทศทางภาษาในระดับประโยค มาเป็นคุณสมบัติให้กับแบบ จำลองเพื่อใช้พิจารณาหาความน่าจะเป็นในการแบ่งขอบเขตให้กับแต่ละอนุพากย์ ซึ่งค่าความน่าจะเป็นในการแบ่งขอบเขตอนุพากย์ของงานวิจัยนี้มีค่าเท่ากับ 0.5 และงานวิจัยนี้ได้มีการทดลองโดยนำผลลัพธ์ของการแจกแจงประโยคมาจาก 2 แหล่ง คือ คลังเอกสาร Penn Treebank และผลลัพธ์การแจกแจงประโยคจากโปรแกรม Charniak parser มาใช้เป็นข้อมูลเข้าของระบบการแบ่งขอบเขตอนุพากย์ปริเฉท ซึ่งผลการทดลองของงานวิจัยนี้มีความถูกต้อง ดังตาราง

ที่ 11 จากผลการทดลองดังกล่าวจะเห็นว่าการนำผลลัพธ์ของการแจกแจงประโยคมาจากคลังเอกสาร Penn Treebank มาเป็นข้อมูลเข้าของระบบ จะให้ค่าความถูกต้องและค่าความระลึกมากกว่าการนำเอาผลลัพธ์ของการแจกแจงประโยคมาจากโปรแกรม Charniak parser ทั้งนี้เนื่องจากผลลัพธ์ของการแจกแจงประโยคมาจากคลังเอกสาร Penn Treebank ที่ทำด้วยคนมีความถูกต้องสูง จึงส่งผลทำให้แบ่งขอบเขตอนุพากย์ปริจเฉทถูกต้องสูงตามไปด้วย ส่วนผลลัพธ์จากโปรแกรม Charniak parser มีข้อผิดพลาดมาก จึงทำให้ผลการแบ่งขอบเขตอนุพากย์ปริจเฉทมีความผิดพลาดตามไปด้วย

ตารางที่ 11 สรุปค่าความถูกต้องและค่าความระลึกของระบบการแบ่งขอบเขตอนุพากย์ปริจเฉท

แหล่งที่มาของข้อมูลนำเข้า	ประสิทธิภาพ	
	ค่าความถูกต้อง (%)	ค่าความระลึก (%)
ผลลัพธ์การแจกแจงประโยคจากโปรแกรม Charniak's parser	83.5	82.7
การแจกแจงประโยคจากคลังเอกสาร Penn Treebank.	85.4	84.1

ข้อเด่น

1. ค่าความถูกต้องสูง เมื่อนำผลลัพธ์การแจกแจงประโยคจากคลังเอกสาร Penn Treebank มาเป็นข้อมูลนำเข้าของระบบ

ข้อด้อย

1. ความถูกต้องของระบบขึ้นอยู่กับค่าความถูกต้องของการแจกแจงประโยคเป็นหลัก
2. คุณสมบัติที่ใช้ในการระบุขอบเขตอนุพากย์ปริจเฉท จะเป็นข้อสนเทศทางภาษาในระดับประโยคที่มีความซับซ้อนมาก ซึ่งข้อมูลเหล่านี้เป็นข้อมูลที่ต้องใช้เวลาและแรงงานในการเตรียมเป็นอย่างมาก

โดยสรุปแนวทางการแบ่งขอบเขตของอนุพากย์ปริจเฉทซึ่งใช้การฝึกฝนและการเรียนรู้โดยใช้เครื่องจักรกลมีขึ้นเพื่อลดระยะเวลาและแรงงานของนักภาษาศาสตร์ในการสร้างกฎ รวมทั้งลดเวลาและแรงงานในการปรับระบบไปใช้ในโดเมนหรือภาษาใหม่ จากเทคนิคแนวทางนี้ถ้าต้องการเปลี่ยนระบบไปทำงานในโดเมนใหม่ ก็สามารถทำได้โดยนำข้อมูลจากคลังเอกสารของโดเมนใหม่

เข้าสู่แบบจำลองการเรียนรู้และฝึกฝน ซึ่งแบบจำลองจะฝึกฝนและเรียนรู้การแบ่งขอบเขตอนุพากย์ปริจเฉทได้ด้วยตัวเอง และสร้างกฎในการแบ่งขอบเขตอนุพากย์ปริจเฉทออกมา จากงานวิจัยที่ผ่านมาระบบการแบ่งขอบเขตอนุพากย์ปริจเฉทเป็นการใช้เทคนิคการฝึกฝนและการเรียนรู้แบบมีผลเฉลย (Supervised learning) ซึ่งมีอยู่ทั้งหมด 2 แบบจำลอง คือ ต้นไม้ตัดสินใจ (Decision tree) และการเรียนรู้เบย์อย่างง่าย (Naïve Bayes)

จากงานวิจัยที่ใช้แนวทางการฝึกฝนระบบให้สามารถเรียนรู้ สามารถสรุปข้อดีและข้อด้อยของระบบในกลุ่มนี้ได้ดังตารางที่ 12

ตารางที่ 12 ข้อดีและข้อด้อยของการใช้แนวทางการฝึกฝนและการเรียนรู้โดยใช้เครื่องจักรกล

ข้อดี	ข้อด้อย
1. ลดเวลาและแรงงานของนักภาษาศาสตร์ที่ใช้ในการสร้างกฎแบ่งขอบเขตอนุพากย์ปริจเฉท	1. ต้องการคลังเอกสารเพื่อใช้ในการฝึกฝนระบบเป็นจำนวนมาก ซึ่งทำให้ต้องเสียเวลาและแรงงานในการเตรียมคลังเอกสารขนาดใหญ่สำหรับการฝึกฝนระบบ
2. สามารถปรับระบบไปใช้ในโดเมนใหม่	2. ความถูกต้องของเอกสารที่ใช้ในการฝึกฝนระบบ มีผลต่อประสิทธิภาพของระบบ
3. เป็นแนวทางที่สร้างกฎการแบ่งขอบเขตอนุพากย์ปริจเฉท โดยไม่ต้องอาศัยความรู้ความชำนาญจากนักภาษาศาสตร์	3. การประมวลผลแต่ละครั้งต้องการใช้ทรัพยากรของคอมพิวเตอร์จำนวนมาก เช่น หน่วยความจำ และหน่วยประมวลผลกลาง

แนวทางที่ 2: การแบ่งขอบเขตอนุพากย์ปริจเฉทด้วยการสร้างกฎโดยผู้เชี่ยวชาญ

แนวทางการแบ่งขอบเขตของอนุพากย์ปริจเฉทในกลุ่มนี้ เป็นระบบที่สร้างขึ้นโดยอาศัยผู้เชี่ยวชาญเป็นผู้วิเคราะห์และสร้างกฎสำหรับระบุขอบเขตของอนุพากย์ปริจเฉท ระบบในกลุ่มนี้มักมีพื้นฐานการทำงานโดยใช้ Regular expression เป็นกฎฮิวริสติก (heuristic rules) โดยกฎที่ผู้เชี่ยวชาญสร้างขึ้นจะใช้ข้อมูลจากข้อสนเทศทางภาษาเป็นตัวแบ่งขอบเขตของอนุพากย์ปริจเฉท ซึ่งข้อสนเทศทางภาษาเหล่านี้ ได้แก่ คำระบุนัย (Marcu, 1998; Alonso et al., 2001) เช่น “เนื่องจาก”

“ดังนั้น” เป็นต้น เครื่องหมายวรรคตอน (Marcu, 1998) และข้อมูลจากโครงสร้างไวยากรณ์ (Thanh et al., 2004) เช่น ผลลัพธ์จากการแจกแจงประโยค

ในปี ค.ศ. 1998 Marcu ได้พัฒนาระบบในการแบ่งขอบเขตของอนุพากย์ปริจเฉทแบบหยาบ (shallow parser) โดยใช้ข้อสนเทศทางภาษา ได้แก่ คำระบุนัย เครื่องหมายวรรคตอน และข้อมูลจากบริบทรอบข้าง เช่น ช่องว่าง การเว้นบรรทัด และการย่อหน้า มาเป็นเครื่องมือในการตัดสินใจแบ่งอนุพากย์ปริจเฉท ซึ่งกฎที่นำมาใช้ภายในระบบนี้ได้มาจากการวิเคราะห์รูปแบบโครงสร้างทางไวยากรณ์ของอนุพากย์ในระดับคำ รวมทั้งการศึกษาลักษณะปรากฏการณ์ (phenomena) ของ คำระบุนัย เช่น ตำแหน่งของคำระบุนัยภายในอนุพากย์ปริจเฉท ลำดับของคำระบุนัย เป็นต้น เครื่องหมายวรรคตอน และข้อมูลจากบริบทรอบข้างที่เกิดขึ้นภายในคลังเอกสาร เพื่อใช้ข้อสนเทศทางภาษาดังกล่าวในการแบ่งขอบเขตของอนุพากย์ปริจเฉท กฎเหล่านี้ผู้เชี่ยวชาญจะเป็นผู้สร้างขึ้นมา ซึ่งผลการทดลองของงานวิจัยนี้มีความถูกต้องเท่ากับ 90.3%

ข้อเด่น

1. ระบบสามารถแบ่งขอบเขตอนุพากย์ปริจเฉทได้อย่างรวดเร็ว เนื่องจากระบบจะประมวลโดยพิจารณาเพียงรูปผิวของคำ (surface form) ของคำระบุนัย และเครื่องหมายวรรคตอนเท่านั้น
2. ระบบให้ความถูกต้องสูง เนื่องจากผู้เชี่ยวชาญสร้างกฎจากการวิเคราะห์คลังเอกสารที่มีขนาดใหญ่

ข้อด้อย

1. ค่าความถูกต้องของระบบจะลดลง เมื่อระบบทำการแบ่งขอบเขตอนุพากย์ปริจเฉทจากประโยคที่มีโครงสร้างทางไวยากรณ์ที่ซับซ้อน ได้แก่ ประโยคความรวม และประโยคความซ้อน เนื่องจากระบบจะพิจารณาจากรูปผิวของคำของคำระบุนัย และเครื่องหมายวรรคตอนเท่านั้น
2. กฎที่สร้างขึ้นไม่สามารถแบ่งขอบเขตอนุพากย์ปริจเฉทที่ไม่ปรากฏคำระบุนัย หรือเครื่องหมายวรรคตอน
3. งานวิจัยของ D. Marcu ไม่ได้ระบุวิธีการแบ่งขอบเขตอนุพากย์ปริจเฉทที่มาจากวลี และงานวิจัยสามารถหาขอบเขตของอนุพากย์ปริจเฉทชนิด Embedded EDU ได้ในกรณีที่ Embedded EDU อยู่ภายใต้เครื่องหมายวรรคตอนเท่านั้น

H. L. Thanh et al. (2004) ได้นำเอาข้อมูลจากผลลัพธ์ของการแจกแจงประโยคภายในคลังเอกสาร Penn Treebank ซึ่งเป็นข้อมูลทางไวยากรณ์ในระดับประโยค มาวิเคราะห์เพื่อสร้างกฎในการแบ่งขอบเขตอนุพากย์ปริจเฉทเป็นขั้นตอนแรก โดยผลลัพธ์จากขั้นตอนนี้จะทำการแบ่งขอบเขตอนุพากย์ปริเฉทที่มีโครงสร้างทางไวยากรณ์เป็นประโยคความเดียวหรืออนุประโยคออกมาเป็นอันดับแรก จากนั้นจึงนำตัวเชื่อมปริจเฉทแบบชัดแจ้ง เช่น "because", "without" ไปแบ่งขอบเขตของอนุพากย์ปริเฉทที่มีโครงสร้างทางไวยากรณ์เป็นนามวลี ซึ่งอนุพากย์ปริจเฉทที่มีโครงสร้างเป็นนามวลีจะต้องขึ้นต้นด้วยตัวเชื่อมปริจเฉทแบบชัดแจ้งดังที่กล่าวมา โดยผลการทดลองของงานวิจัยนี้มีความถูกต้อง 81.43 % และค่าความระลึก 79.26%

ข้อเด่น

1. การแบ่งขอบเขตอนุพากย์ปริจเฉทมีค่าความถูกต้องสูง เนื่องจากกฎสร้างขึ้นจากคลังเอกสาร Penn Treebank ที่มีความถูกต้องสูง จึงทำให้กฎที่สร้างขึ้นโดยผู้เชี่ยวชาญไม่มีความคลุมเครือหรือขัดแย้งกัน
2. สามารถแบ่งขอบเขตอนุพากย์ปริจเฉทที่ไม่ปรากฏคำระบุนัยหรือเครื่องหมายวรรค-ตอนได้ โดยใช้ข้อมูลจากผลลัพธ์ของการแจกแจงประโยค

ข้อด้อย

1. สามารถแบ่งขอบเขตอนุพากย์ปริจเฉทที่มีโครงสร้างที่เป็นวลีได้เฉพาะกรณีที่อนุพากย์ปริจเฉทมีโครงสร้างไวยากรณ์เป็นนามวลี
2. ต้องใช้ข้อมูลจากผลลัพธ์ของการแจกแจงประโยคเป็นส่วนหลักในการทำงาน ซึ่งข้อมูลเหล่านี้เป็นข้อมูลที่ต้องใช้เวลาและแรงงานในการเตรียมเป็นอย่างมาก

การแบ่งขอบเขตอนุพากย์ปริจเฉทด้วยการสร้างกฎโดยผู้เชี่ยวชาญจะให้ผลลัพธ์ที่มีความถูกต้องสูง เนื่องจากเป็นการนำความรู้ทางภาษาศาสตร์หรือความรู้พิเศษจากผู้เชี่ยวชาญใน แต่ละโดเมนมาใช้ในระบบโดยตรง อย่างไรก็ตาม แนวทางการทำงานดังกล่าวต่างก็มีข้อด้อยและข้อจำกัด เช่น กฎที่ผู้เชี่ยวชาญสร้างขึ้นจากเอกสารตัวอย่างนั้นอาจจะไม่ครอบคลุมการแบ่งขอบเขตอนุพากย์ปริจเฉททุกกรณี ปัญหานี้เกิดจากการสำรวจพฤติกรรมของการแบ่งขอบเขตอนุพากย์ปริจเฉทไม่ครบทุกกรณี หรือเกิดจากรูปแบบการเขียนบทความของผู้เขียนระหว่างเอกสารที่ใช้วิเคราะห์และเอกสารที่ใช้ประมวลผลมีความแตกต่างกันมาก ปัญหาอีกประการหนึ่งคือ กฎที่สร้างขึ้นจากผู้เชี่ยวชาญอาจมีการซ้ำซ้อนหรือขัดแย้งกัน ซึ่งปัญหานี้เกิดจากการสร้างกฎจำนวนมากจาก

ผู้เชี่ยวชาญหลายคนโดยไม่ได้มีการตรวจสอบความถูกต้องและความขัดแย้งของแต่ละกฎ ปัญหาประการสุดท้าย คือ เทคนิคที่พัฒนาขึ้นโดยวิธีนี้มักทำงานได้ดีเฉพาะในโดเมนที่ผู้พัฒนาสนใจเท่านั้น เมื่อต้องการเปลี่ยนไปประมวลผลในโดเมนอื่นก็จะต้องสร้างกฎขึ้นใหม่ จึงทำให้เสียเวลาและแรงงานเป็นอย่างมาก ซึ่งข้อดีและข้อด้อยของแนวทางนี้สรุปดังตารางที่ 13

ตารางที่ 13 ข้อดีและข้อด้อยของการใช้แนวทางการการสร้างกฎโดยผู้เชี่ยวชาญ

ข้อดี	ข้อด้อย
1. สามารถนำเอาความรู้จากผู้เชี่ยวชาญมาใช้ในระบบได้โดยตรง	1. ความถูกต้องจะขึ้นกับความเชี่ยวชาญของผู้ออกแบบและสร้างกฎการแบ่งขอบเขตอนุพากย์ปริจเฉท
2. ผลลัพธ์มีความถูกต้องสูงเมื่อวัดผลกับเอกสารในโดเมนเดียวกับกลุ่มเอกสารที่ใช้วิเคราะห์เพื่อสร้างกฎ	2. การสร้างกฎการแบ่งขอบเขตอนุพากย์ปริจเฉทจะต้องใช้เวลาและแรงงานจำนวนมาก
3. ไม่ต้องใช้การประมวลผลทางคอมพิวเตอร์ที่ซับซ้อนและในการประมวลผลแต่ละครั้งไม่ต้องใช้ทรัพยากรของคอมพิวเตอร์จำนวนมาก เช่น หน่วยความจำ และหน่วยประมวลผลกลาง	3. การสร้างกฎการแบ่งขอบเขตอนุพากย์ปริจเฉทให้ครอบคลุมทุกกรณี เป็นกระบวนการที่ค่อนข้างยาก
	4. กฎการแบ่งขอบเขตอนุพากย์ปริจเฉทที่สร้างขึ้นอาจมีการซ้ำซ้อนหรือขัดแย้งกัน

อุปกรณ์และวิธีการ

อุปกรณ์

1. ฮาร์ดแวร์ระบบ

เครื่องคอมพิวเตอร์พีซี 1 เครื่อง: ซีพียู Pentium IV 2.0 GHz. หน่วยความจำขนาด 256 MB, ฮาร์ดดิสก์ 20 GB.

2. ซอฟต์แวร์ระบบ

2.1 ระบบปฏิบัติการวินโดวส์ XP

2.2 โปรแกรมภาษาไพธอน (Python)

3. คลังเอกสารภาษาไทย

คลังเอกสารภาษาไทยในโดเมนการเกษตรสำหรับฝึกฝนระบบขนาด 615 EDUs (6,000 คำ) โดยผ่านการประมวลผลเบื้องต้น อันได้แก่ การตัดคำ, การกำกับชนิดของคำ, การกำกับขอบเขตของนิพจน์ระบุนาม, การกำกับขอบเขตของนามวลี และการกำกับขอบเขตและชนิดของแต่ละอนุพากย์ปริจเฉท

วิธีการ

การแบ่งขอบเขตของอนุพากย์ปริจเฉทในภาษาไทย

กระบวนการทำงานของการแบ่งขอบเขตอนุพากย์ปริจเฉท เป็นกระบวนการแบ่งข้อความภายในเอกสารให้เป็นข้อความอนุพากย์ปริจเฉท โดยแต่ละอนุพากย์ปริจเฉทนั้นต้องมีคุณสมบัติเป็นอนุพากย์อิสระที่สามารถสื่อความหมายได้อย่างสมบูรณ์ และข้อความในแต่ละหน่วยจะต้องไม่ทับซ้อนกัน จากการศึกษาโครงสร้างและชนิดของอนุพากย์ปริเฉทภาษาไทยที่เกิดขึ้นในโดเมนการเกษตรขนาด 615 EDUs (6,000 คำ) พบว่าโครงสร้างของอนุพากย์ปริจเฉทภาษาไทยมีความแตกต่างจากอนุพากย์ปริจเฉทภาษาอังกฤษ เนื่องจากทั้งสองภาษามีหลักโครงสร้างทางไวยากรณ์ของประโยคที่แตกต่างกัน ดังนั้นหลักเกณฑ์ในการแบ่งขอบเขตของอนุพากย์ปริจเฉทบางอย่างจึงมีความแตกต่างกัน ซึ่งจากการสำรวจคลังเอกสารภาษาไทยสามารถแบ่งชนิดของอนุพากย์ปริจเฉทได้

เป็น 2 ชนิด คือ Basic EDU และ Embedded EDU ซึ่ง Basic EDU เป็นอนุพากย์ปริจเฉทที่โครงสร้างทางไวยกรณ์เป็นอนุประโยคหรือประโยคความเดียว และวลีที่ขึ้นต้นด้วยตัวเชื่อมปริจเฉทแบบชัดแจ้ง เช่น “เนื่องจาก” ,“เพราะ” ,“เช่น” ส่วน Embedded EDU เป็นอนุพากย์ปริจเฉทที่มีโครงสร้างทางไวยากรณ์เหมือนกับ Basic EDU แต่ Embedded EDU เป็นอนุพากย์ปริจเฉทที่ปรากฏซ้อนอยู่ภายใน Basic EDU ทำให้ Basic EDU แบ่งออกเป็น 2 ส่วน ตัวอย่างของอนุพากย์ปริจเฉทสำหรับภาษาไทย แสดงดังตารางที่ 14 โดย Basic EDU จะอยู่ภายในเครื่องหมายวงเล็บ (“[…]”) และ Embedded EDU จะอยู่ภายในเครื่องหมายวงเล็บปีกกา (“{…}”)

ตารางที่ 14 ตัวอย่างของ EDU แต่ละชนิด

ชนิดของอนุพากย์		ตัวอย่าง
ชนิดหลัก	**ชนิดย่อย**	
Basic EDU	ประโยคความเดียว/ อนุประโยค	[กะหล่ำปลีมีสีเขียว] กะหล่ำปลีมีสีเขียว [ซึ่งปลูกได้ดีทางภาคเหนือ]
	นามวลี	โรคระบาดพบในภาคกลาง [เช่น ปทุมธานี ,นครปฐม]
Embedded EDU	อนุประโยค	กะหล่ำปลี {ที่ถูกทำลาย} จะมีสีเหลือง
	นามวลี	เกษตรกรควรใส่ปุ๋ยไนโตรเจน {เช่น ปุ๋ยแอมโมเนียซัลเฟต หรือยูเรีย} ลงในแปลงด้วย

ปัญหาในการแบ่งขอบเขตของอนุพากย์ปริจเฉทภาษาไทย

กระบวนการแบ่งขอบเขตข้อความภายในเอกสารให้เป็นอนุพากย์ปริจเฉทสำหรับภาษาไทยนั้น ไม่สามารถทำได้โดยง่ายเหมือนกับภาษาอื่นๆ ทั้งนี้มีสาเหตุเนื่องมาจากคุณลักษณะของภาษา ซึ่งทำให้ระบบเกิดความคลุมเครือในการแบ่งขอบเขตของอนุพากย์ปริเฉท โดยคุณลักษณะภาษาเหล่านี้ ได้แก่

1. ภาษาไทยไม่มีข้อสนเทศหรือสัญลักษณ์พิเศษ ในการระบุจุดสิ้นสุดของอนุประโยคหรือประโยคเหมือนภาษาอื่นๆ เช่น มหัพภาคหรือจุด (“ . ”), จุลภาคหรือจุดลูกน้ำ (,) และ อัฒภาค (“;”) ทำให้การแบ่งขอบเขตอนุพากย์ปริจเฉทของภาษาไทยมีความคลุมเครือในการแบ่งมากกว่าภาษาอื่น

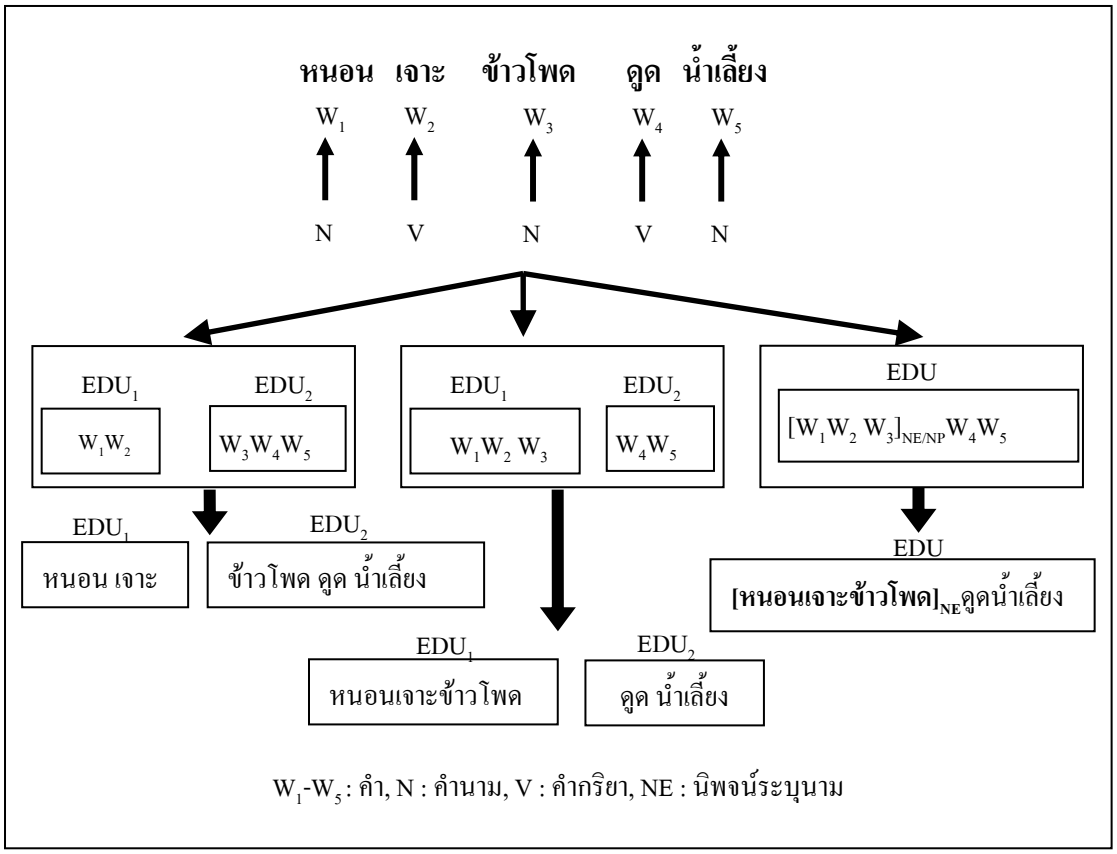
เช่น ภาษาอังกฤษ และภาษาเยอรมัน ซึ่งตัวอย่างการแบ่งขอบเขตของอนุพากย์ปริจเฉทระหว่างภาษาอังกฤษและภาษาไทยแสดงดังตารางที่ 15

<u>ตารางที่ 15</u> ตัวอย่างการแบ่งขอบเขตของอนุพากย์ปริจเฉทระหว่างภาษาอังกฤษและภาษาไทย

ประโยค	ข้อมูลนำเข้า	ผลลัพธ์การแบ่งขอบเขตของอนุพากย์ปริจเฉท
ภาษาอังกฤษ	If it is severe symptom, then cabbage will turn to rotten.	[If it is severe symptom,]$_{EDU1}$ [then cabbage will turn to rotten.]$_{EDU2}$
ภาษาไทย	เมื่ออาการรุนแรงจะทำให้กะหล่ำปลีเน่าและ	[เมื่ออาการรุนแรง]$_{EDU1}$ [จะทำให้กะหล่ำปลีเน่าและ]$_{EDU2}$

จากตารางที่ 15 จะเห็นว่าในประโยคภาษาอังกฤษสามารถใช้เครื่องหมายวรรคตอนเป็นข้อสนเทศสำหรับแบ่งขอบเขตของแต่ละอนุพากย์ปริจเฉทได้ โดยขอบเขตของอนุพากย์ปริจเฉทที่ 1 (EDU_1) จะใช้จุลภาคหรือจุดลูกน้ำ (“ , ”) เป็นตัวแบ่งขอบเขต และขอบเขตของอนุพากย์ปริจเฉทที่ 2 (EDU_2) ใช้มหัพภาคหรือจุด (“ . ”) เป็นตัวแบ่งขอบเขต ส่วนตัวอย่างประโยคภาษาไทยในตารางที่ 15 จะไม่มีปรากฏข้อสนเทศใดๆ ที่เป็นตัวช่วยในการแบ่งขอบเขตของแต่ละอนุพากย์ปริจเฉท ดังนั้นการแบ่งขอบเขตของแต่ละอนุพากย์ปริจเฉทในภาษาไทยจึงมีความยากลำบากกว่าการแบ่งขอบเขตของอนุพากย์ปริจเฉทภาษาอังกฤษ

2. นิพจน์ระบุนามและนามวลี เนื่องจากนิพจน์ระบุนามและนามวลีบางวลีเกิดจากการนำคำชนิดต่างๆ เช่น คำนาม มาประกอบกับกริยา ทำให้นิพจน์ระบุนามและนามวลีดังกล่าวมีโครงสร้างทางไวยากรณ์เหมือนกับโครงสร้างไวยากรณ์ของอนุพากย์ปริจเฉท กล่าวคือ มีโครงสร้างไวยากรณ์เป็นประโยคความเดียวหรืออนุประโยค จากการที่นิพจน์ระบุนาม นามวลีและอนุพากย์ปริจเฉทมีโครงสร้างทางไวยากรณ์เหมือนกันนั้น จะทำให้ระบบทำการระบุขอบเขตของอนุพากย์ปริจเฉทผิดพลาด โดยระบบอาจระบุขอบเขตของนิพจน์ระบุนามหรือนามวลีเป็นอนุพากย์ปริเฉทได้ ซึ่งตัวอย่างแสดงดังภาพที่ 3



ภาพที่ 4 ตัวอย่างปัญหาที่เกิดจากนิพจน์ระบุนาม

จากภาพที่ 4 เมื่อข้อมูลเข้า คือ ประโยค “หนอนเจาะข้าวโพดดูดน้ำเลี้ยง” เราสามารถแบ่งขอบเขตอนุพากย์ปริจเฉทโดยพิจารณาจากโครงสร้างทางไวยากรณ์ได้ 3 รูปแบบคือ

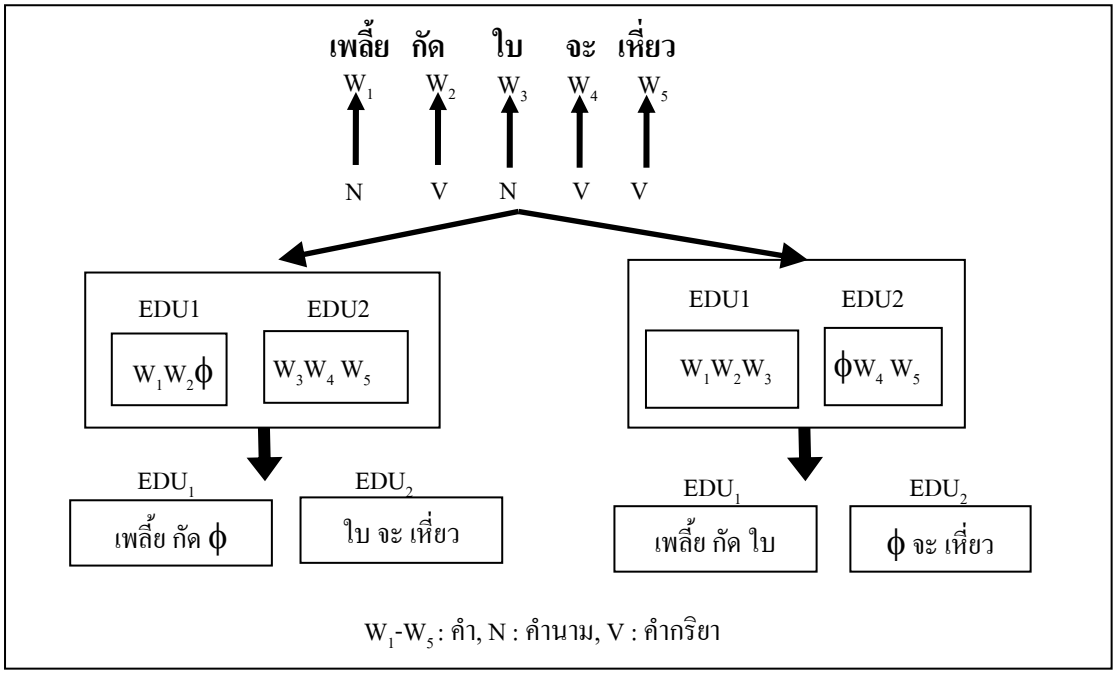
แบบที่ 1: [หนอนเจาะ]$_{EDU1}$ [ข้าวโพดดูดน้ำเลี้ยง]$_{EDU2}$
แบบที่ 2: [หนอนเจาะข้าวโพด]$_{EDU1}$ [ดูดน้ำเลี้ยง]$_{EDU2}$
แบบที่ 3: [หนอนเจาะข้าวโพดดูดน้ำเลี้ยง]$_{EDU}$

การแบ่งขอบเขตอนุพากย์ปริจเฉทในแบบที่ 1 สามารถแบ่งขอบเขตของอนุพากย์ปริจเฉทได้ 2 อนุพากย์ โดยที่อนุพากย์ปริจเฉทที่ 1 ประธาน คือ “หนอน” และ กริยา คือ “เจาะ” ส่วนอนุพากย์ปริจเฉทที่ 2 ประธาน คือ “ข้าวโพด” กริยา คือ “ดูด” และกรรม คือ “น้ำเลี้ยง” ในการแบ่งขอบเขตอนุพากย์ปริจเฉทในแบบที่ 2 สามารถขอบเขตของอนุพากย์ 2 อนุพากย์เช่นกัน โดยที่อนุพากย์ปริจเฉทที่ 1 ประธาน คือ “หนอน” กริยา คือ “เจาะ” และกรรม คือ “ข้าวโพด” ส่วน

อนุพากย์ปริจเฉทที่ 2 กริยา คือ “ดูด” กรรม คือ “น้ำเลี้ยง” และในแบบที่ 3 สามารถแบ่งขอบเขตของอนุพากย์ปริจเฉทได้เพียง 1 อนุพากย์เท่านั้น โดยที่ประธานของอนุพากย์ คือ “หนอนเจาะข้าวโพด” กริยา คือ “ดูด” และกรรม คือ “น้ำเลี้ยง”

จากการแบ่งขอบเขตของอนุพากย์ปริจเฉททั้ง 3 แบบ แบบที่ถูกต้อง คือ การแบ่งขอบเขตอนุพากย์ปริจเฉทในแบบที่ 3 เนื่องจาก คำว่า “หนอนเจาะข้าวโพด” เป็น นิพจน์ระบุนาม 1 นิพจน์ และทำหน้าที่เป็นประธานของประโยค จึงทำให้ไม่สามารถแยกคำภายในนิพจน์ออกเป็นส่วนๆ ได้ ดังรูปแบบที่ 1 และ 2 ดังนั้นการแบ่งขอบเขตของอนุพากย์ปริจเฉทในแบบที่ 1 และ 2 จึงไม่ถูกต้อง

3. การละรูปคำ (Zero Anaphora) และการละรูปกริยา (Verb Ellipsis) จากการศึกษาคลังเอกสารพบว่า ประโยคภาษาไทยสามารถเกิดการละรูปคำและการละรูปกริยาได้ ซึ่งปัญหาเหล่านี้จะส่งผลให้ระบบระบุขอบเขตของอนุพากย์ปริจเฉทผิดพลาด เนื่องจากการละรูปของคำจะส่งผลให้ระบบไม่สามารถระบุคำที่เป็นประธานหรือกรรมของประโยคได้ถูกต้อง เนื่องจากไม่ปรากฏรูปของคำดังกล่าว จึงส่งผลทำให้ระบุขอบเขตของอนุพากย์ปริจเฉทไม่ถูกต้อง โดยอาจมีการสลับตำแหน่งประธานหรือกรรมระหว่างอนุพากย์ปริจเฉทที่อยู่ติดกัน ทำให้เกิดความกำกวมในการระบุขอบเขตของอนุพากย์ปริจเฉทในภาษาไทย ซึ่งตัวอย่างแสดงดังภาพที่ 5



ภาพที่ 5 ตัวอย่างปัญหาที่เกิดจากการละรูปคำ

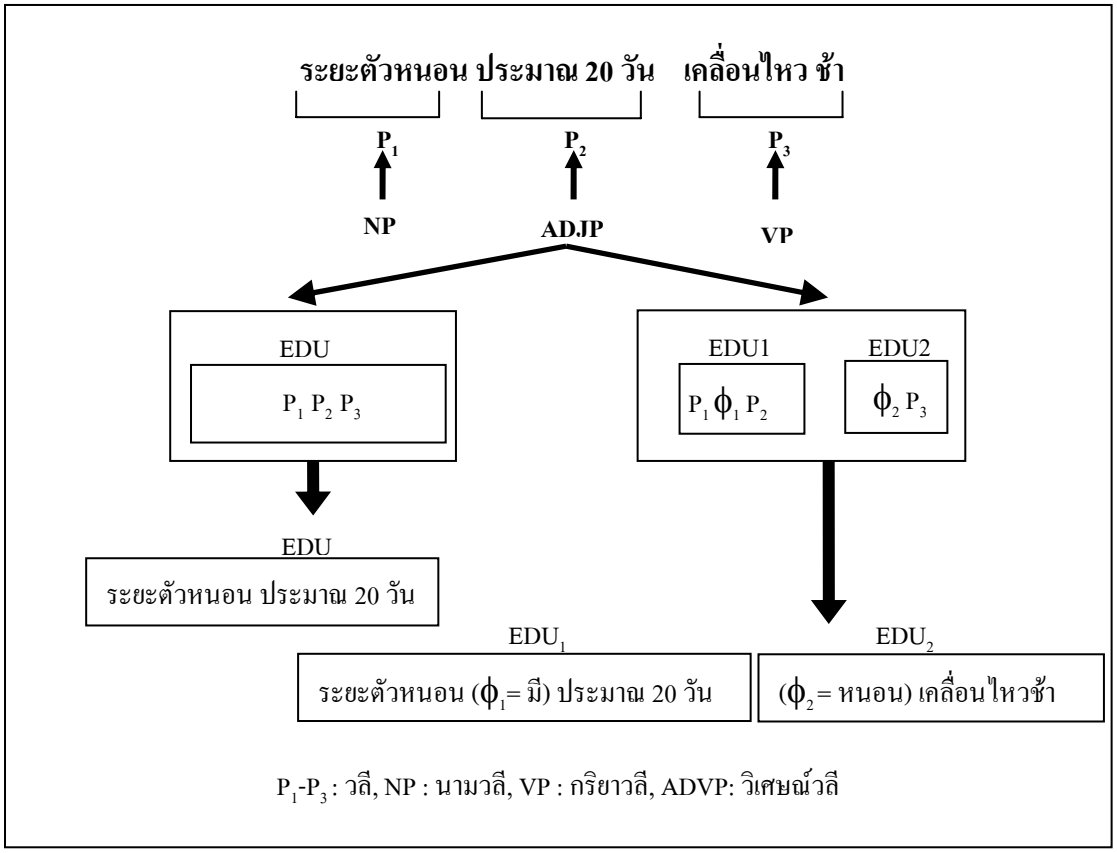
จากภาพที่ 4 เมื่อข้อมูลเข้าคือประโยค “เพลี้ยกัดใบจะเหี่ยว” สามารถแบ่งขอบเขตอนุพากย์ปริจเฉทโดยพิจารณาจากโครงสร้างทางไวยากรณ์ได้ 2 รูปแบบคือ

แบบที่ 1: [เพลี้ยกัด]$_{EDU1}$ [ใบจะเหี่ยว]$_{EDU2}$
แบบที่ 2: [เพลี้ยกัดใบ]$_{EDU1}$ [จะเหี่ยว]$_{EDU2}$

โดยแบบที่ 1 สามารถแบ่งขอบเขตของอนุพากย์ปริจเฉทได้ 2 อนุพากย์ โดยที่อนุพากย์ปริจเฉทที่ 1 ประธาน คือ “เพลี้ย” และ กริยา คือ “กัด” ซึ่งอนุพากย์ที่ 1 มีการละรูปกรรมของประโยคไว้ ส่วนอนุพากย์ปริจเฉทที่ 2 ประธาน คือ “ใบ” กริยา คือ “จะเหี่ยว” การแบ่งขอบเขตอนุพากย์ปริจเฉทในแบบที่ 2 สามารถขอบเขตของอนุพากย์ 2 อนุพากย์เช่นกัน โดยที่อนุพากย์ปริจเฉทที่ 1 ประธาน คือ “เพลี้ย” กริยา คือ “กัด” กรรม คือ “ใบ” ส่วนอนุพากย์ปริจเฉทที่ 2 ประกอบด้วย คำกริยา คือ “จะเหี่ยว” ซึ่งอนุพากย์ที่ 2 มีการละรูปประธานของไว้

จากตัวอย่างประโยคข้างต้นเราสามารถการแบ่งขอบเขตของอนุพากย์ปริจเฉททั้ง 2 แบบ และจะสังเกตเห็นว่าประโยคอาจมีการละรูปของคำได้ 2 ตำแหน่ง คือ ตำแหน่งประธานหรือกรรมของประโยค โดยการแบ่งขอบเขตอนุพากย์ปริจเฉทในแบบที่ 1 อนุพากย์ปริจเฉทที่ 1 นั้นมีการละรูปกรรม จึงทำให้คำที่อยู่ถัดไป คือ คำว่า “ใบ” ไปทำหน้าที่เป็นประธานให้กับอนุพากย์ปริจเฉทที่ 2 ส่วนการแบ่งขอบเขตอนุพากย์ปริจเฉทในแบบที่ 2 อนุพากย์ปริจเฉทที่ 2 นั้นมีการละรูปประธาน จึงทำให้คำว่า “ใบ” ที่อยู่ในตำแหน่งก่อนหน้าไปทำหน้าที่เป็นกรรมให้กับอนุพากย์ปริจเฉทที่ 1

ส่วนประโยคภาษาไทยที่ละรูปของกริยาไปนั้น จะทำให้โครงสร้างทางไวยากรณ์ของประโยคดังกล่าว มีโครงสร้างเทียบเท่าวลีเท่านั้น ดังนั้นประโยคดังกล่าวจะถูกระบุว่าไม่มีคุณสมบัติเป็นอนุพากย์ปริจเฉท ทำให้ข้อความภายในประโยคเหล่านั้นถูกนำไปรวมกับ EDU อื่นๆ ซึ่งจะส่งผลให้ระบบระบุขอบเขตและจำนวนของอนุพากย์ปริจเฉทผิดพลาด ตัวอย่างเช่น ประโยค: “ระยะตัวหนอนประมาณ 20 วัน เคลื่อนไหวช้า” มีรูปแบบการแบ่งขอบเขตอนุพากย์ปริจเฉทแสดงดังภาพที่ 6



ภาพที่ 6 ตัวอย่างปัญหาที่เกิดจากการละรูปกริยา

จากภาพที่ 5 เมื่อข้อมูลเข้าคือประโยค “ระยะตัวหนอนประมาณ 20 วัน เคลื่อนไหวช้า” สามารถแบ่งขอบเขตอนุพากย์ปริจเฉทโดยพิจารณาจากโครงสร้างทางไวยากรณ์ได้ 2 รูปแบบคือ

แบบที่ 1: [ระยะตัวหนอนประมาณ 20 วัน เคลื่อนไหวช้า]$_{EDU1}$
แบบที่ 2: [ระยะตัวหนอนประมาณ 20 วัน]$_{EDU1}$ [เคลื่อนไหวช้า]$_{EDU2}$

โดยการแบ่งขอบเขตอนุพากย์ปริจเฉทในแบบที่ 1 สามารถแบ่งขอบเขตของอนุพากย์ปริจเฉทได้เพียง 1 อนุพากย์เท่านั้น โดยที่อนุพากย์ที่ 1 ประธาน คือ “ระยะตัวหนอนประมาณ 20 วัน” และ กริยา คือ “เคลื่อนไหวช้า ส่วนการแบ่งขอบเขตอนุพากย์ปริจเฉทในแบบที่ 2 สามารถแบ่งขอบเขตของอนุพากย์ 2 อนุพากย์ โดยที่อนุพากย์ปริจเฉทที่ 1 ประธาน คือ “ระยะตัวหนอน” กริยาของประโยคที่ละรูป คือ “มี” และมีวิเศษณ์วลี คือ “ประมาณ 20 วัน” เป็นส่วนขยายกริยา ส่วนอนุพากย์ปริจเฉทที่ 2 ประกอบด้วย ประธานของประโยคที่ละไป คือ “หนอน” และกริยาวลี คือ “เคลื่อนไหวช้า”

จากการแบ่งขอบเขตของอนุพากย์ปริจเฉททั้ง 2 แบบ แต่แบบที่ถูกต้อง คือ การแบ่งขอบเขตในแบบที่ 2 เนื่องจาก อนุพากย์ปริจเฉทที่ 1 “ระยะตัวหนอนประมาณ 20 วัน” มีการละคำกริยาคำว่า “มี” และ อนุพากย์ปริจเฉทที่ 2 “เคลื่อนไหวช้า” มีการละประธานของประโยคคำว่า “หนอน” จึงทำให้ไม่สามารถรวมข้อความทั้งหมดเป็น 1 อนุพากย์ปริจเฉทได้ดังรูปแบบที่ 1 ดังนั้น การแบ่งขอบเขตอนุพากย์ปริจเฉทในแบบที่ 1จึงไม่ถูกต้อง

หลักการและเหตุผล

จากการที่ภาษาไทยมีลักษณะหลายประการที่แตกต่างจากภาษาอื่น อันได้แก่ ภาษาไทยไม่มีข้อสนเทศที่ใช้ระบุจุดสิ้นสุดของประโยค ภาษาไทยสามารถละรูปกริยาได้ จากลักษณะดังกล่าวทำให้ไม่สามารถนำแนวทางการระบุขอบเขตอนุภาคปริจเฉทของภาษาอื่นมาใช้ได้โดยตรง ดังนั้น วิทยานิพนธ์นี้ จึงมีขึ้นเพื่อศึกษาแนวทางในการแบ่งขอบเขตอนุพากย์ปริจเฉทจากงานวิจัยที่ผ่านมา และนำแนวทางเหล่านั้นมาประยุกต์เพื่อพัฒนาระบบระบุขอบเขตอนุภาคปริจเฉทสำหรับเอกสารภาษาไทย เพื่อให้สามารถระบุขอบเขตอนุภาคปริจเฉทแบบอัตโนมัติ ซึ่งการระบุขอบเขตอนุภาคปริจเฉทภาษาไทยสำหรับวิทยานิพนธ์นี้ เป็นการใช้หลักการผสมผสานระหว่างแนวทางการฝึกฝนและการเรียนรู้โดยใช้เครื่องจักรกล ร่วมกับแนวทางการใช้กฎที่สร้างขึ้นจากผู้เชี่ยวชาญ เพื่อช่วยเพิ่มประสิทธิภาพของการระบุขอบเขตอนุภาคปริจเฉทภาษาไทย อีกทั้งข้อสนเทศทางภาษาที่ใช้ในการระบุขอบเขตของอนุพากย์ปริเฉทของงานวิจัยนี้จะใช้ข้อสนเทศทางภาษาในระดับคำและวลีเท่านั้น

การแบ่งขอบเขตอนุพากย์ปริจเฉทสำหรับภาษาไทย

การแบ่งขอบเขตอนุพากย์ปริจเฉทสำหรับภาษาไทยเป็นการหาขอบเขตอนุพากย์ปริจเฉทสำหรับภาษาไทย โดยจะทำการแบ่งขอบเขตอนุพากย์ทั้ง Basic EDU และ Embedded EDU ซึ่งงานวิจัยนี้จะใช้หลักการผสมผสานระหว่างแนวทางการฝึกฝนและการเรียนรู้โดยใช้เครื่องจักรกลร่วมกับการแนวทางการใช้กฎที่สร้างขึ้นโดยผู้เชี่ยวชาญ โดยขั้นตอนแรกจะใช้แนวทางการฝึกฝนและการเรียนรู้โดยใช้เครื่องจักรกล ซึ่งจะเทคนิคการฝึกฝนและการเรียนรู้ภายในงานวิจัยนี้ได้

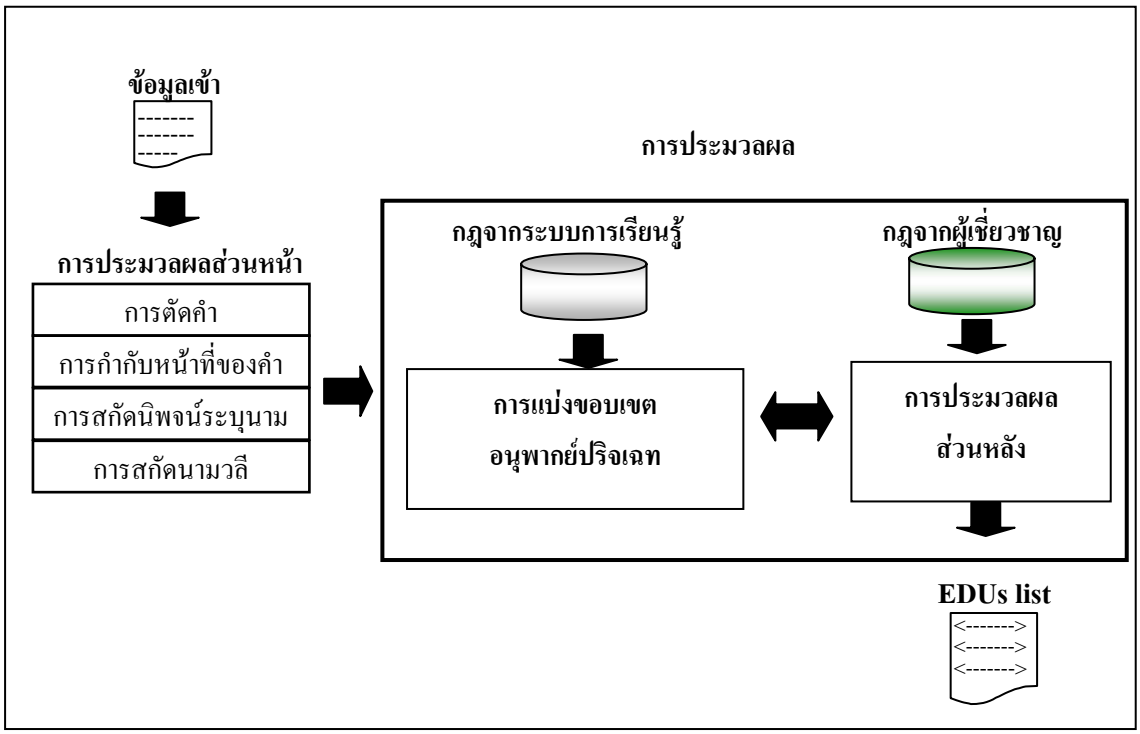
เลือกใช้แบบจำลองต้นไม้ตัดสินใจ (Decision-tree learning) เพื่อแบ่งขอบเขตของอนุพากย์ปริจเฉท เนื่องจากกฎที่สร้างจากแบบจำลองนี้จะมีโครงสร้างเป็นต้นไม้ ซึ่งจะทำให้แปลความหมายของแต่ละกฎได้ง่าย และคุณลักษณะที่เลือกใช้ในการแบ่งขอบเขตของอนุพากย์ปริจเฉท ได้แก่ คำระบุนัย เนื่องจากการสำรวจคลังเอกสารภาษาไทยในโดเมนการเกษตรขนาด 615 EDUs (6,000 คำ) พบว่า เอกสารภาษาไทยมักปรากฏคำระบุนัยประมาณ 18.37% และคำระบุนัยเหล่านี้สามารถใช้เป็นตัวบ่งชี้ตำแหน่งจุดเริ่มต้นและจุดสิ้นสุดของแต่ละอนุพากย์ปริจเฉทได้เป็นอย่างดี อย่างไรก็ตาม คำระบุนัยบางคำยังมีความคลุมเครืออยู่ ซึ่งจะส่งผลให้แบ่งขอบเขตของอนุพากย์ปริจเฉทผิดพลาด ดังนั้นในการหาขอบเขตของอนุพากย์ปริจเฉทจึงจำเป็นต้องใช้ข้อมูลจากคุณลักษณะอื่นๆ เพื่อช่วยลดความคลุมเครือของคำระบุนัย และช่วยแบ่งขอบเขตของอนุพากย์ที่ไม่ปรากฏคำระบุนัย ซึ่งคุณลักษณะอื่นๆ ดังกล่าว ได้แก่ ชนิดของคำ (Part of speech หรือ POS) ของคำระบุนัย, ชนิดของคำที่อยู่รอบข้างคำระบุนัย และช่องว่าง ส่วนขั้นตอนที่ 2 เป็นการใช้แนวทางการใช้กฎที่สร้างขึ้นโดยผู้เชี่ยวชาญ โดยผู้เชี่ยวชาญจะสร้างกฎเพื่อแก้ไขปัญหากรณีที่กฎที่ได้จากแนวทางการฝึกฝนและการเรียนรู้จากแบบจำลองต้นไม้ตัดสินใจยังมีความผิดพลาดอยู่ เช่น กรณีของคำระบุนัยที่ปรากฏมากกว่า 1 คำต่ออนุพากย์ อีกทั้งผู้เชี่ยวชาญจะสร้างกฎเพิ่มเติมโดยเน้นการสร้างกฎใน การแบ่งขอบเขตของอนุพากย์ปริจเฉทที่ไม่ปรากฏคำระบุ ทั้งนี้การสร้างกฎดังกล่าวจะช่วยทำให้การแบ่งขอบเขตอนุพากย์ปริจเฉทภาษาไทยมีประสิทธิภาพดีขึ้น

ภาพรวมของการแบ่งขอบเขตอนุพากย์ปริจเฉทสำหรับภาษาไทย

การแบ่งขอบเขตอนุพากย์ปริจเฉทในภาษาไทยจะรับข้อมูลเข้าเป็นลำดับของตัวอักษรภายในเอกสาร (Sequence of lexical token) โดยเรียงลำดับจากตัวอักษรจากซ้ายไปขวา เพื่อทำการหาขอบเขตของอนุพากย์ปริจเฉทแต่ละอนุพากย์ โดยภาพรวมของการแบ่งขอบเขตอนุพากย์ปริจเฉทออกเป็น 2 ขั้นตอน คือ การประมวลผลส่วนหน้า และการประมวลผล ซึ่งแสดงดังภาพที่ 7

1. การประมวลผลส่วนหน้า

เป็นกระบวนการแรกของระบบที่ประมวลผลภาษาในระดับคำและวลี เพื่อเตรียมเอกสารให้อยู่ในรูปที่สามารถประมวลผลในส่วนต่อไปได้ (Computable Format) อีกทั้งยังช่วยแก้ปัญหาความกำกวมของการแบ่งขอบเขตของอนุพากย์ในภาษาไทย การประมวลผลส่วนนี้ประกอบด้วย 4 ขั้นตอนย่อย คือ การตัดคำ การกำกับชนิดของคำ (Sudprasert et al. , 2003) การระบุนิพจน์ระบุนาม (Chanlekhaet al. , 2004) และการสกัดนามวลี (Pengphon et al. , 2002)



<u>ภาพที่ 7</u> ภาพรวมของการแบ่งขอบเขตอนุพากย์ปริจเฉทสำหรับภาษาไทย

1.1 การตัดคำ (Word Segmentation)

ขั้นตอนนี้เป็นการตัดคำในเอกสาร (Sudprasert et al. , 2003) จากคุณลักษณะของภาษาไทยที่เขียนคำเรียงติดกันไปจนจบประโยคหรือจบย่อหน้า ขั้นตอนนี้จะทำการแบ่งขอบเขตของคำภาษาไทยแต่ละคำออกเป็นส่วนๆ โดยผลลัพธ์ของการตัดคำจะเป็นดังภาพที่ 8

เมื่อ มี เพลี้ย กระโดด สี น้ำตาล จำนวน มาก ดูด กิน น้ำเลี้ยง ต้น ข้าว จะ ทำให้ ต้น ข้าว แสดง อาการ ใบ เหลือง แห้ง _ คล้าย ถูก น้ำ ร้อน ลวก _ ซึ่ง เรียกว่า อาการ ไหม้ เป็น หย่อม ถ้า รุนแรง มาก _ ต้น ข้าว จะ แห้ง ตาย เพลี้ย กระโดด สี น้ำตาล สามารถ ทำลาย ได้ ทุก ระยะ การ เจริญเติบโต ของ ข้าว นอกจากนี้ ยัง เป็น พาหะ นำ เชื้อ วิสา _ ซึ่ง ทำให้ เกิด โรค ใบหงิก มา สู่ ต้น ข้าว อีกด้วย

ภาพที่ 8 เอกสารที่ผ่านกระบวนการตัดคำแล้ว

1.2 การกำกับชนิดของคำ (POS Tagging)

ขั้นตอนนี้เป็นการกำกับชนิดของคำ (POS Tagging) ภาษาไทย (Sudprasert et al. , 2003) โดยอาศัยเทคนิคการประมวลผลทางสถิติทำงานร่วมกับพจนานุกรม โดยคำใดที่ไม่พบในพจนานุกรมหรือมีความคลุมเครือก็จะใช้ข้อมูลทางสถิติของการเกิดขึ้นของคำและชนิดของคำบริบทรอบข้างมาเป็นคุณลักษณะในการวิเคราะห์ชนิดของคำ ผลลัพธ์ของการกำกับชนิดของคำที่ได้แสดงดังภาพที่ 9

เมื่อ/conj มี/vt เพลี้ย/ncn กระโดด/npn สี/ncn น้ำตาล/adj จำนวน/ncn มาก/adj ดูด/vt กิน/vt น้ำเลี้ยง/ncn ต้น/ncn ข้าว/ncn จะ/prev ทำให้/vcau ต้น/ncn ข้าว/ncn แสดง/vt อาการ/ncn ใบ/ncn เหลือง/adj แห้ง/adj _/blk คล้าย/vcs ถูก/psm น้ำ/ncn ร้อน/adj ลวก/vt _/blk ซึ่ง/prel เรียกว่า/vcs อาการ/ncn ไหม้/vi เป็น/advm2 หย่อม/nct ถ้า/conj รุนแรง/vi มาก/adv _/blk ต้น/ncn ข้าว/ncn จะ/prev แห้ง/vi ตาย/advเพลี้ย/ncn กระโดด/npn สี/ncn น้ำตาล/adj สามารถ/prev ทำลาย/vt ได้/vpost ทุก/qubo ระยะ/ncn การ/pref1 เจริญเติบโต/vi ของ/prep ข้าว/ncn นอกจากนี้/conj ยัง/prev เป็น/vcs พาหะ/ncn นำ/vt เชื้อ/ncn วิสา/npn _/blk ซึ่ง/prel ทำให้/vcau เกิด/vt โรค/ncn ใบหงิก/npn มา/vt สู่/prep ต้น/ncn ข้าว/ncn อีกด้วย/adv

ภาพที่ 9 เอกสารที่ผ่านกระบวนการกำกับชนิดของคำ

1.3 การระบุนิพจน์ระบุนาม (Name Entity Extraction)

เป็นขั้นตอนระบุขอบเขตของนิพจน์ระบุนาม (Name Entity หรือ NE) ที่ปรากฏภายในเอกสาร (Chanlekhaet al., 2004) โดยใช้วิธีการผสมระหว่างการใช้การคำนวณเชิงสถิติจากการฝึกฝนระบบร่วมกับการใช้ฐานความรู้ โดยเทคนิคการเรียนรู้ใช้แบบจำลองแมกซิมั่มเอนโทรปีเรียนรู้จากข้อสนเทศระดับคำและฐานความรู้ ได้แก่ คลังคำศัพท์ พจนานุกรมชื่อพร้อมประเภทของชื่อนั้น และกฎฮิวริสติก เพื่อช่วยเพิ่มประสิทธิภาพของการสกัดนิพจน์ระบุนามโดยใช้สถิติ ผลลัพธ์ของการระบุนิพจน์ระบุนามแสดงดังภาพที่ 10

เมื่อ/conj มี/vt **[เพลี้ยกระโดดสีน้ำตาล]/NE** จำนวน/ncn มาก/adj ดูด/vt กิน/vt น้ำเลี้ยง/ncn ต้น/ncn ข้าว/ncn จะ/prev ทำให้/vcau ต้น/ncn ข้าว/ncn แสดง/vt อาการ/ncn ใบ/ncn เหลือง/adj แห้ง/adj _/blk คล้าย/vcs ถูก/psm น้ำ/ncn ร้อน/adj ลวก/vt _/blk ซึ่ง/prel เรียกว่า/vcs อาการ/ncn ไหม้/vi เป็น/advm2 หย่อม/nct ถ้า/conj รุนแรง/vi มาก/adv _/blk ต้น/ncn ข้าว/ncn จะ/prev แห้ง/vi ตาย/adv **[เพลี้ยกระโดดสีน้ำตาล]/NE** สามารถ/prev ทำลาย/vt ได้/vpost ทุก/qubo ระยะ/ncn การ/pref1 เจริญเติบโต/vi ของ/prep ข้าว/ncn นอกจากนี้/conj ยัง/prev เป็น/vcs พาหะ/ncn นำ/vt **[เชื้อวิสา]/NE** _/blk ซึ่ง/prel ทำให้/vcau เกิด/vt **[โรคใบหงิก]/NE** มา/vt สู่/prep ต้น/ncn ข้าว/ncn อีกด้วย/adv

ภาพที่ 10 เอกสารที่ผ่านการระบุนิพจน์ระบุนาม

1.4 การสกัดนามวลี

เป็นขั้นตอนระบุขอบเขตของนามวลีที่ปรากฏภายในเอกสาร (Pengphon et al. , 2002) เพื่อช่วยลดความคลุมเครือในการแบ่งขอบเขตอนุพากย์ปริเฉทที่มีโครงสร้างไวยากรณ์เหมือนกับนามวลี ผลลัพธ์ของการขอบเขตนามวลีแสดงดังภาพที่ 11

เมื่อ/conj มี/vt [เพลี้ยกระโดดสีน้ำตาล]/NE จำนวน/ncn มาก/adj ดูด/vt กิน/vt น้ำเลี้ยง/ncn **[ต้นข้าว]/Cpn** จะ/prev ทำให้/vcau **[ต้นข้าว]/cpn** แสดง/vt อาการ/ncn ใบ/ncn เหลือง/adj แห้ง/adj _/blk คล้าย/vcs ถูก/psm น้ำ/ncn ร้อน/adj ลวก/vt _/blk ซึ่ง/prel เรียกว่า/vcs อาการ/ncn ไหม้/vi เป็น/advm2 หย่อม/nct ถ้า/conj รุนแรง/vi มาก/adv _/blk **[ต้นข้าว]/cpn** จะ/prev แห้ง/vi ตาย/adv [เพลี้ยกระโดดสีน้ำตาล]/NE สามารถ/prev ทำลาย/vt ได้/vpost ทุก/qubo ระยะ/ncn การ/pref1 เจริญเติบโต/vi ของ/prep ข้าว/ncn นอกจากนี้/conj ยัง/prev เป็น/vcs พาหะ/ncn นำ/vt [เชื้อวิสา]/NE _/blk ซึ่ง/prel ทำให้/vcau เกิด/vt [โรคใบหงิก]/NE มา/vt สู่/prep **[ต้นข้าว]/cpn** อีกด้วย/adv

ภาพที่ 11 เอกสารที่ผ่านการสกัดนามวลี

2. การประมวลผล (Processing)

กระบวนการนี้เป็นกระบวนการระบุขอบเขตของจุดเริ่มต้นและจุดสิ้นสุดของขอบเขตอนุพากย์ปริจเฉท โดยใช้แนวทางการฝึกฝนและการเรียนรู้โดยเครื่องจักรกลและแนวทางการสร้างกฎด้วยผู้เชี่ยวชาญในการสร้างกฎสำหรับการแบ่งขอบเขตของอนุพากย์ปริจเฉท ซึ่งส่วนของการประมวลผลนี้ประกอบด้วยขั้นตอนย่อย 2 ขั้นตอน คือ ขั้นตอนการแบ่งขอบเขตอนุพากย์ปริจเฉทโดยใช้กฎการแบ่งขอบเขตอนุพากย์ปริจเฉทจากแบบจำลองการฝึกฝนการเรียนรู้ และขั้นตอนการประมวลผลส่วนหลังโดยใช้กฎจากผู้เชี่ยวชาญ

การแบ่งขอบเขตอนุพากย์ปริจเฉทโดยใช้กฎจากแบบจำลองการฝึกฝนและการเรียนรู้

คุณลักษณะที่ใช้ในการสร้างกฎสำหรับการแบ่งขอบเขตของอนุพากย์ในภาษาไทย สามารถแบ่งออกเป็น 4 คุณลักษณะ ได้แก่ ประเภทของคำระบุนัย (Discourse Segmentation Cues) คำระบุนัยแบบคู่ (Correlative Discourse Markers) ช่องว่าง และชนิดของคำ (POS)

1. ประเภทของคำระบุนัย (Discourse Segmentation Cues)

คำระบุนัย คือ คำที่ใช้บ่งชี้ขอบเขตของจุดเริ่มต้นและจุดสิ้นสุดในแต่ละอนุพากย์ปริจเฉท โดยจากการสำรวจจากคลังเอกสารปรากฏว่า คำระบุนัยสำหรับภาษาไทยประกอบด้วยข้อสนเทศทางภาษาทั้งหมด 3 ชนิด ได้แก่ ตัวเชื่อมปริจเฉท (discourse markers) คำระบุนัยสำคัญ (cue words) และ เครื่องหมายวรรคตอน (punctuation)

1.1 ตัวเชื่อมปริจเฉท

เป็นคำที่ใช้ระบุชนิดความสัมพันธ์ระหว่างอนุพากย์ปริจเฉท อีกทั้งยังเป็นคำที่ใช้ระบุจุดเริ่มต้นของอนุพากย์ปริจเฉทอีกด้วย เช่น คำว่า “เนื่องจาก” เป็นคำที่ใช้สื่อถึงชนิดความสัมพันธ์แบบ Cause-result และคำว่า “แต่” เป็นคำที่ใช้สื่อถึงชนิดความสัมพันธ์แบบ Contrast

1.2 คำระบุนัยสำคัญ

เป็นคำที่ใช้บ่งชี้ขอบเขตจุดเริ่มต้นและจุดสิ้นสุดของอนุพากย์ปริจเฉท ซึ่งสามารถแบ่งคำระบุนัยสำคัญเหล่านี้ตามตำแหน่งการปรากฏของคำระบุนัยดังกล่าวได้ทั้งหมด 2 ชนิด ได้แก่

1.2.1 คำระบุนัยเริ่มต้น (Starting EDUs cues) เป็นคำระบุนัยสำคัญที่ได้บ่งชี้จุดเริ่มต้นของอนุพากย์ปริจเฉท เช่น “โดย” และ “ที่” เป็นต้น

1.2.2 คำระบุนัยสิ้นสุด (Ending EDUs cues) เป็นคำระบุนัยสำคัญที่ได้บ่งชี้จุดเริ่มสิ้นสุดของอนุพากย์ปริจเฉท เช่น “แล้ว” และ “เสีย” เป็นต้น

1.3 เครื่องหมายวรรคตอน

เป็นเครื่องหมายวรรคตอนที่ใช้บ่งชี้ขอบเขตจุดเริ่มต้นและจุดสิ้นสุดของอนุพากย์ปริจเฉทเหมือนกับคำระบุนัยสำคัญ เช่น เครื่องหมายวงเล็บ “( )”, เครื่องหมายจุลภาค (,)

จากการสำรวจจำนวนการปรากฏของคำระบุนัยจากคลังเอกสารในโดเมนการเกษตร ที่มีจำนวนอนุพากย์ปริจเฉททั้งหมด 615 อนุพากย์ (6,000 คำ) ดังตารางที่ 16 พบว่าอนุพากย์ปริจเฉทที่ปรากฏคำระบุนัยมีจำนวน 18.37% และอนุพากย์ปริจเฉทที่ไม่ปรากฏคำระบุนัยมีจำนวน 81.63%

ตารางที่ 16 ผลการสำรวจจำนวนปรากฏของคำระบุนัย

ลักษณะอนุพากย์ปริเฉท	จำนวนการปรากฏ (%)
ปรากฏคำระบุนัย	18.37
ไม่ปรากฏคำระบุนัย	81.63

นอกจากนี้ยังพบว่าคำระบุนัยเหล่านี้เป็นตัวบ่งชี้ในการระบุขอบเขตของอนุพากย์ปริจเฉทที่มีประสิทธิภาพดี แต่อย่างไรก็ตามคำระบุนัยดังกล่าวอาจมีความคลุมเครืออยู่ ดังนั้นเราจึงจัดประเภทของคำระบุนัยตามประสิทธิภาพของคำระบุนัยได้ 2 ประเภท คือ คำระบุนัยที่ไม่มีความคลุมเครือ (strong cues) และ คำระบุนัยที่มีความคลุมเครือ (weak cues)

คำระบุนัยที่ไม่มีความคลุมเครือ เป็นคำระบุนัยที่ไม่มีความคลุมเครือในการระบุขอบเขตของจุดเริ่มต้นของอนุพากย์ปริจเฉท ดังนั้นการแบ่งขอบเขตอนุพากย์ปริจเฉทสามารถใช้เพียงรูปคำ (surface form) ในการแบ่งขอบเขตของอนุพากย์ เช่น “เพราะ” “เพื่อ” “หาก” และ “ดังนั้น” นอกจากนี้เรายังแบ่ง คำระบุนัยที่ไม่มีความคลุมเครือออกเป็น 2 ประเภท ตามตำแหน่งของคำระบุนัย คือ Strong-start-basic cues และ Strong-end-basic cues ตัวอย่างประเภทย่อยของคำระบุนัยที่ไม่มีความคลุมเครือ ดูตารางที่ 17 และรายชื่อของคำระบุนัยที่ไม่มีความคลุมเครือทั้งหมดดูตารางผนวกที่ ก1

ตารางที่ 17 ประเภทของคำระบุนัยที่ไม่มีความคลุมเครือ

ประเภทย่อยของคำระบุนัย	ตำแหน่งการปรากฏของคำระบุนัย		ชนิดของอนุพากย์ปริจเฉท		จำนวนของคำระบุนัย	ตัวอย่างของคำระบุนัย
	เริ่มต้น	สิ้นสุด	Basic EDU	Embedded EDU		
Strong-start-basic cues	✓	-	✓	-	32	ถ้า, เพราะ, นอกจากนี้, ต่อมา
Strong-end-basic cues	-	✓	✓	-	5	ก็ได้, ทีเดียว, เช่นกัน, ก็ตาม,นัก

- Strong-start-basic cues เป็นคำระบุนัยที่ไม่มีความคลุมเครือ ที่ใช้ในการระบุตำแหน่งเริ่มต้นของ อนุพากย์ปริจเฉทชนิด Basic EDU ซึ่งมีจำนวนทั้งหมด 32 คำ เช่น “เพราะ” “เนื่องจาก” และ “นอกจากนี้” เป็นต้น

- Strong-end-basic cues เป็นคำระบุนัยที่ไม่มีความคลุมเครือ ที่ใช้ในการระบุตำแหน่งสิ้นสุดของ อนุพากย์ปริจเฉทชนิด Basic EDU ซึ่งมีจำนวนทั้งหมด 5 คำ คือ “ก็ได้” “ทีเดียว” “เช่นกัน” “ก็ตาม” และ “นัก”

<u>คำระบุนัยที่มีความคลุมเครือ</u> เป็นคำระบุนัยที่มีความคลุมเครือในการระบุขอบเขตของจุดเริ่มต้นของอนุพากย์ปริจเฉทอยู่ จำเป็นต้องใช้ข้อมูลจากคุณลักษณะอื่นๆ ประกอบการพิจารณาเพื่อช่วยลดความคลุมเครือของคำระบุนัย เช่น ชนิดของคำของคำระบุนัย, ชนิดของคำของบริบทรอบข้าง ตัวอย่างของคำระบุนัยดูตารางที่ 18 และรายชื่อของคำระบุนัยที่มีความคลุมเครือทั้งหมดดูตารางผนวกที่ ก1

<u>ตารางที่ 18</u> ตัวอย่างการใช้ชนิดของคำเพื่อลดความคลุมเครือ

คำระบุนัยที่มีความคลุมเครือ	ชนิดของคำทั้งหมด (POS)	ชนิดของคำที่แบ่งขอบเขตอนุพากย์
โดย	สันธาน, บุพบท	สันธาน
ที่	ประพันธสรรพนาม, คำบุพบท	ประพันธสรรพนาม

จากตารางที่ 18 คำว่า “โดย” เป็นคำระบุชนิดมีความคลุมเครือ ที่ใช้บอกจุดเริ่มต้นของอนุพากย์ปริจเฉท ซึ่งเราไม่สามารถใช้รูปคำในการแบ่งขอบเขตอนุพากย์ได้ เนื่องจากยังมีความคลุมเครืออยู่ จึงจำเป็นต้องใช้ข้อมูลอื่นๆ ประกอบการพิจารณาแบ่งขอบเขตของอนุพากย์ ซึ่งคำว่า “โดย” จะพิจารณาจากชนิดของคำของคำว่า “โดย” ที่มีอยู่ 3 หน้าที่ คือ สันธาน (conjunction) และบุพบท (preposition) จากการศึกษาในคลังเอกสารพบว่า คำว่า “โดย” ที่สามารถบอกว่าเป็นจุดเริ่มต้นของอนุพากย์ปริจเฉทต้องมีชนิดของคำเป็นคำสันธาน เท่านั้น

นอกจากนี้เรายังแบ่งคำระบุนัยที่มีความคลุมเครือนี้ออกเป็น 4 ประเภท ตามตำแหน่งในการคำระบุนัย คือ Weak-start-basic cues, Weak-start-embedded cues, Weak-end-basic cues และ Weak-end-embedded cues ซึ่งประเภทย่อยของคำระบุนัยที่มีความคลุมเครือ ดูตารางที่ 19

- Weak-start-basic cues เป็นคำระบุนัยที่มีความคลุมเครือ ที่ใช้ในการระบุตำแหน่งเริ่มต้นของ อนุพากย์ปริจเฉทชนิด Basic EDU ซึ่งมีจำนวนทั้งหมด 24 คำ เช่น “โดย” “จะ” และ “ให้” เป็นต้น

- Weak-start-embedded cues เป็นคำระบุนัยที่มีความคลุมเครือ ที่ใช้ในการระบุตำแหน่งเริ่มต้นของ อนุพากย์ปริจเฉทชนิด Embedded EDU ซึ่งมีจำนวนทั้งหมด 4 คำ ได้แก่ “ที่” “(”, “ได้แก่” และ “เช่น”

- Weak-end-basic cues เป็นคำระบุนัยที่มีความคลุมเครือ ที่ใช้ในการระบุตำแหน่งสิ้นสุดของ อนุพากย์ปริจเฉทชนิด Basic EDU ซึ่งมีจำนวนทั้งหมด 7 คำ เช่น “แล้ว” “ลง” และ “ไป” เป็นต้น

- Weak-end-embedded cues เป็นคำระบุนัยที่มีความคลุมเครือ ที่ใช้ในการระบุตำแหน่งสิ้นสุดของอนุพากย์ปริจเฉทชนิด Embedded EDU ซึ่งมีจำนวนทั้งหมด 3 คำ ได้แก่ ")" "เป็นต้น" และ "ฯลฯ"

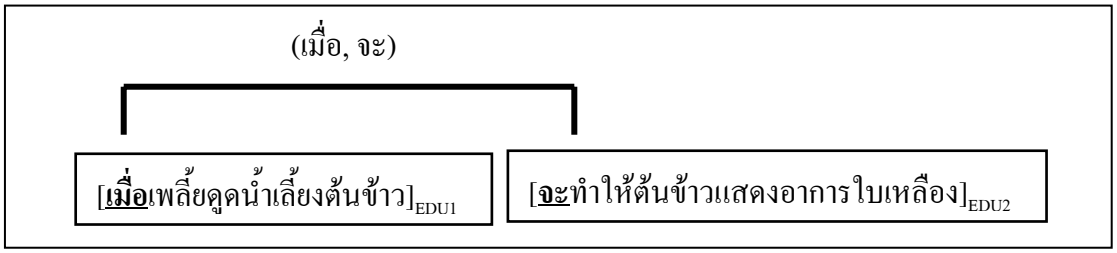
<u>ตารางที่ 19</u> ประเภทของคำระบุนัยที่มีความคลุมเครือ

ประเภทย่อยของคำระบุนัย	ตำแหน่งการปรากฏของคำระบุนัย		ชนิดของอนุพากย์ปริจเฉท		จำนวนของคำระบุนัย	ตัวอย่างของคำระบุนัย
	เริ่มต้น	สิ้นสุด	Basic EDU	Embedded EDU		
Weak-start-basic cues	✓	-	✓	-	25	มัก, ควร, จะ
Weak-start-embedded cues	✓	-	-	✓	4	ที่, ได้แก่, เช่น, (
Weak-end-basic cues	-	✓	✓	-	11	แล้ว, ลง,ไป
Weak-end-embedded cues	-	✓	-	✓	3	) , เป็นต้น ,ฯลฯ

โดยค่าของคุณลักษณะชนิดของคำระบุนัยจะมีค่าเป็น 0 และ 1 ซึ่งพิจารณาจากชนิดของคำระบุนัยแต่ละชนิด ซึ่งถ้าคำที่พิจารณาเป็นสมาชิกของคำระบุนัยชนิดไม่มีความคลุมเครือ และเป็นสมาชิกของคำระบุนัยที่มีชนิดย่อยเป็น Strong-start-basic cues หรือ Strong-end-basic cues แล้วค่าของคุณลักษณะของดังกล่าวมีค่าเท่ากับ 1 กรณีอื่นนอกจากนี้ค่าของคุณลักษณะดังกล่าวมีค่าเท่ากับ 0 แต่ถ้าคำที่พิจารณาเป็นสมาชิกของคำระบุนัยชนิดมีความคลุมเครือและเป็นสมาชิกของคำระบุนัยที่มีชนิดย่อยเป็น Weak-start-basic cues หรือ Weak-start-basic cues หรือ Weak-end-basic cues หรือ Weak-end-embed cues ค่าของคุณลักษณะดังกล่าวมีค่าเท่ากับ 1 เมื่อรูปคำและชนิดของคำของคำระบุนัยดังกล่าวตรงตามข้อกำหนด แต่ถ้ารูปคำหรือชนิดของคำของคำระบุนัยไม่ตรงตามข้อกำหนดแล้ว ค่าของคุณลักษณะดังกล่าวมีค่าเท่ากับ 0

2. คำระบุนัยแบบคู่ (Correlative Discourse Markers)

คำระบุนัยแบบคู่ (Correlative discourse markers หรือ Co-dm) เป็นคำระบุนัยแบบคู่ที่ใช้ระบุตำแหน่งเริ่มต้นของอนุพากย์ที่อยู่ติดกัน 2 อนุพากย์ จากการศึกษาคลังเอกสารพบว่ามีจำนวนของ Co-dm ทั้งหมด 10 คู่ ได้แก่ (“ถ้า”, “จะ”), (“ถ้า”, “ก็”), (“ถ้า”, “ให้”), (“หาก”, “จะ”), (“หาก”, “ก็”) , (“หาก”, “มัก”), (“เมื่อ”, “ก็”), (“เมื่อ”, “ให้”), (“เมื่อ”,”จะ”) และ (“เมื่อ”, “มัก”) ค่าของคุณลักษณะนี้ มีค่าเท่ากับ 1 เมื่อคำที่พิจารณาเป็นสมาชิกของคู่ลำดับแรกของ Co-dm และ คำต่อจากคำที่พิจารณานี้ที่อยู่ในช่วงระยะ 1- 10 คำ คำใดคำหนึ่งในช่วงระยะนี้ต้องเป็นสมาชิกตัวที่ 2 ของคู่ลำดับแรก ถ้าคำที่พิจารณาไม่ตรงกับเงื่อนไขดังกล่าว ค่าของคุณลักษณะนี้ มีค่าเท่ากับ 0 ตัวอย่างของ Co-dm แสดงในภาพที่ 12



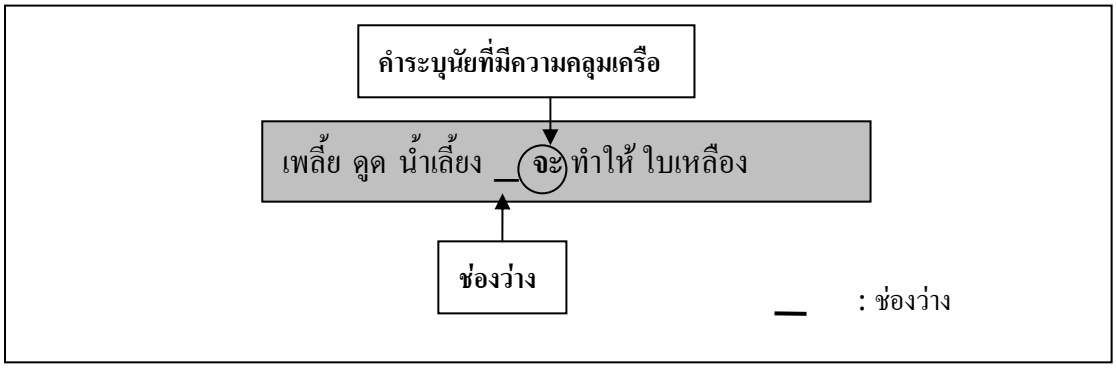
ภาพที่ 12 ตัวอย่างอนุพากย์ปริจเฉทที่ปรากฏคำระบุนัยแบบคู่

จากภาพที่ 16 อนุพากย์ปริจเฉทที่ 1 ปรากฏคำว่า “เมื่อ” และอนุพากย์ปริจเฉทที่ 2 ปรากฏคำว่า “จะ” คำว่า “เมื่อ” และ “จะ” เป็นสมาชิกของ Co-dm จะมีระยะห่างระหว่างคำอยู่ในช่วง 1-10 คำ ดังนั้นอนุพากย์ปริจเฉททั้งสองจึงมีค่าคุณลักษณะ Co-dm เท่ากับ 1

3. ช่องว่าง

เนื่องภาษาไทยไม่มีข้อสนเทศในการระบุจุดสิ้นสุดของอนุประโยคหรือประโยคเหมือนภาษาอื่นๆ เช่น มหัพภาคหรือจุด (“ . ”) , จุลภาคหรือจุดลูกน้ำ (“,”) และอัฒภาค (“;”) อีกทั้งจากผลการสำรวจดังตารางที่ 16 พบว่าอนุพากย์ปริจเฉทที่ไม่ปรากฏคำระบุนัยมีจำนวน 81.63% และคำระบุนัยบางตัวยังมีความคลุมเครืออยู่ ดังนั้นระบบจึงนำเอาช่องว่างมาเป็นคุณลักษณะหนึ่งที่ช่วยลดความคลุมเครือของคำระบุนัยในการแบ่งขอบเขตอนุพากย์และนำเอาการปรากฏช่องว่างมาเป็น

คุณลักษณะช่วยประกอบการตัดสินใจในการแบ่งขอบเขตอนุพากย์ปริจเฉทที่ไม่ปรากฏคำระบุนัย โดยค่าของคุณลักษณะมีค่าเท่ากับ 1 ถ้าคำที่พิจารณาเป็นช่องว่าง นอกจากนี้ค่าของคุณลักษณะนี้มีค่าเท่ากับ 0

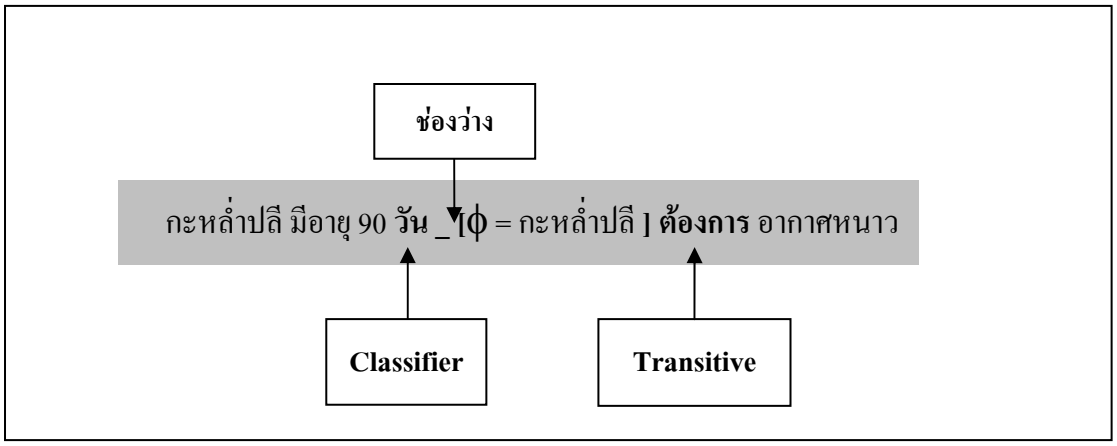


<u>ภาพที่ 13</u> อนุพากย์ปริจเฉทที่ใช้คุณลักษณะช่องว่างร่วมกับคำระบุนัยที่มีความคลุมเครือ

จากภาพที่ 13 ตัวอย่างข้อความคือ “เพลี้ยดูดน้ำเลี้ยง_จะทำให้ใบเหี่ยว” และถ้าคำที่พิจารณาว่าจะแบ่งขอบเขตอนุพากย์ปริจเฉทหรือไม่ คือ คำว่า “จะ” โดยจะเริ่มพิจารณาจากคุณลักษณะของ “จะ” ก่อน จากการพิจารณาพบว่า “จะ” เป็นสมาชิกคำระบุนัยประเภทย่อย คือ Weak-start-basic cues ที่ใช้ระบุจุดเริ่มต้นของอนุพากย์ปริจเฉท แสดงว่าคำระบุนัยนี้ยังมีความคลุมเครืออยู่ จึงจำเป็นต้องใช้ข้อสนเทศทางภาษาอื่นๆ เช่น ช่องว่าง มาประกอบการตัดสินใจในการแบ่งขอบเขตอนุพากย์ จากข้อความที่เป็นข้อมูลนำเข้า พบว่าหน้าคำว่า “จะ” ปรากฏช่องว่างอยู่ จากข้อมูลคุณลักษณะช่องว่างนี้ ระบบจึงตัดสินใจว่า คำว่า “จะ” คือ จุดเริ่มต้นของอนุพากย์ปริจเฉท ดังนั้นช่องว่างจึงเป็นคุณลักษณะที่สำคัญคุณลักษณะหนึ่งที่ช่วยในการแบ่งขอบเขตอนุพากย์ปริจเฉทภาษาไทย

4. ชนิดของคำ (POS)

ในกรณีที่อนุพากย์ปริจเฉทไม่ปรากฏข้อสนเทศทางภาษาที่ใช้บ่งชี้ขอบเขตของอนุพากย์ปริจเฉท เช่น คำระบุนัย และช่องว่าง หรือกรณีที่อนุพากย์ปริจเฉทปรากฏคำระบุนัยที่มีความคลุมเครือ อาจทำให้การแบ่งขอบเขตอนุพากย์ปริจเฉทผิดพลาด ดังนั้นจึงนำเอาคุณลักษณะของชนิดของคำเป็นข้อสนเทศที่ช่วยแบ่งขอบเขตอนุพากย์ปริจเฉทจากกรณีดังกล่าว โดยค่าของคุณลักษณะมีค่าเป็นชนิดของคำแต่ละคำ เช่น “vi” “ncn” และ “prep” เป็นต้น



<u>ภาพที่ 14</u> ตัวอย่างอนุพากย์ปริจเฉทที่ใช้คุณลักษณะชนิดของคำ

จากภาพที่ 14 ตัวอย่างข้อความคือ “กะหล่ำปลีมีอายุ _90_วัน _ [ϕ = กะหล่ำปลี] ต้องการอากาศหนาว” จะเห็นว่าอนุพากย์ปริจเฉทไม่ปรากฏคำระบุนัยที่ใช้บ่งชี้ขอบเขตของอนุพากย์ปริจเฉท ดังนั้นจึงต้องพิจารณาการแบ่งขอบเขตอนุพากย์โดยใช้คุณลักษณะชนิดของคำเป็นหลัก จากตัวอย่างข้อความข้างต้นถ้าคำที่ระบบพิจารณาว่าจะแบ่งขอบเขตอนุพากย์ปริจเฉทหรือไม่ คือ คำว่า “ต้องการ” ระบบจึงพิจารณาคุณลักษณะจากชนิดของคำของคำว่า “ต้องการ” จากนั้นจึงนำคุณลักษณะจากบริบทรอบข้างคำว่า “ต้องการ” มาพิจารณา ซึ่งคุณลักษณะดังกล่าวมีดังต่อไปนี้ คำว่า “ต้องการ” มีชนิดของคำเป็น transitive verb หน้าคำว่า “ต้องการ” ปรากฏช่องว่าง และคำถัดจากนั้นปรากฏคำที่มีชนิดของคำเป็น Classifier จากคุณลักษณะดังกล่าวจึงตัดสินใจว่า “ต้องการ” คือจุดเริ่มต้นของอนุพากย์ปริจเฉท ดังนั้นชนิดของคำจึงเป็นคุณลักษณะที่สำคัญที่ช่วยในการแบ่งขอบเขตอนุพากย์ปริจเฉทที่ไม่ปรากฏคำระบุนัย

การศึกษาเทคนิคการฝึกฝนและการเรียนรู้โดยเครื่องจักรกลที่ใช้ในการแบ่งขอบเขตอนุพากย์ปริจเฉท ดังปรากฏในภาคผนวก ง งานวิจัยนี้เลือกใช้เทคนิคการฝึกฝนและการเรียนรู้แบบจำลองต้นไม้ตัดสินใจ (C 4.5) เนื่องจากคุณลักษณะที่ใช้ในการแบ่งขอบเขตอนุพากย์เป็นคุณลักษณะที่มีความเป็นอิสระ ไม่มีความต่อเนื่อง อีกทั้งกฎสร้างจากแบบจำลองต้นไม้ตัดสินใจเป็นกฎที่อ่านเข้าใจได้ง่าย ซึ่งเหมาะสำหรับการนำเอากฎเหล่านี้ไปพัฒนาต่อโดยผู้เชี่ยวชาญ ซึ่งเป็นขั้นตอนการทำงานต่อไป หากกฎที่สร้างขึ้นจากเครื่องจักรกลเป็นกฎที่ผู้เชี่ยวชาญเข้าใจได้ง่ายจะช่วยทำให้ผู้เชี่ยวชาญทำความเข้าใจความหมายของกฎต่างๆได้ง่าย และรวดเร็วขึ้น อีกทั้งยังทำให้

ผู้เชี่ยวชาญสามารสร้างกฎและเพิ่มกฎในการแบ่งขอบเขตอนุพากย์ปริจเฉท โดยไม่มีความซ้ำซ้อนหรือมีความขัดแย้งกัน ดังนั้นงานวิจัยนี้จึงเลือกใช้แบบจำลองต้นไม้ตัดสินใจ (C 4.5) เป็นเทคนิคการฝึกฝนและการเรียนรู้ของเครื่องจักรกลที่ใช้ในการแบ่งขอบเขตอนุพากย์ปริจเฉทภาษาไทย

การประมวลผลส่วนหลังโดยใช้กฎจากผู้เชี่ยวชาญ

เนื่องจากกฎที่ได้จากการแบบจำลองการฝึกฝนและเรียนรู้ยังไม่สามารถสร้างกฎการแบ่งขอบเขตอนุพากย์ปริจเฉทได้คลอบครุมทุกกรณี เช่น กรณีที่อนุพากย์ปริจเฉทไม่ปรากฏคำระบุนัยที่ใช้ในการแบ่งขอบเขตอนุพากย์ปริจเฉท เป็นต้น ดังนั้นงานวิจัยนี้จึงต้องอาศัยผู้เชี่ยวชาญมาเพิ่มกฎที่แบบจำลองไม่สามารถสร้างได้ครอบคลุม ซึ่งคุณลักษณะที่ผู้เชี่ยวชาญนำมาใช้สร้างกฎเพื่อแบ่งขอบเขตอนุพากย์ปริจเฉทจะเหมือนกับคุณลักษณะที่ใช้ในระบบการฝึกฝนและเรียนรู้โดยเครื่องจักรกล ซึ่งคุณลักษณะดังกล่าว ได้แก่ ประเภทของคำระบุนัย คำระบุนัยแบบคู่ ช่องว่าง และชนิดของคำ

ตัวอย่างการสกัดคุณลักษณะ

คุณลักษณะที่ใช้ในการแบ่งขอบเขตอนุพากย์ปริจเฉทมีทั้งหมด 4 คุณลักษณะหลัก คือ ประเภทของคำระบุนัย คำระบุนัยแบบคู่ ช่องว่าง และชนิดของคำ ซึ่งคุณลักษณะประเภทของคำระบุนัยจะประกอบด้วยคุณลักษณะย่อยจำนวน 6 คุณลักษณะ คือ Strong-start-basic cues, Strong-end-basic cues, Weak-start-basic cues, Weak-start-embedded cues, Weak-end-basic cues และ Weak-end-embedded cues ดังนั้นคุณลักษณะที่ใช้ในการแบ่งขอบเขตอนุพากย์ปริจเฉทมีจำนวนทั้งหมด 9 คุณลักษณะ

โดยกำหนดให้

f_1 (w) = คุณลักษณะของ Strong-start-basic cues

f_2 (w) = คุณลักษณะของ Strong-end-basic cues

f_3 (w) = คุณลักษณะของ Weak-start-basic cues

f_4 (w) == คุณลักษณะของ Weak-start-embedded cues

f_5 (w) = คุณลักษณะของ Weak-end-basic cues

f_6 (w) = คุณลักษณะของ Weak-end-embedded cues

f_7 (w) = คุณลักษณะของคำระบุนัยแบบคู่

f_8 (w) = คุณลักษณะของช่องว่าง

f_9 (w) = คุณลักษณะของชนิดของคำ

ตัวอย่างการสกัดคุณลักษณะของข้อมูลเข้า เช่น "ถ้า/conj ระบาด/vi รุนแรง/adv จะ/prev ทำให้/vcau ข้าว/ncn แห้ง/vi ตาย/vi " ซึ่งผลลัพธ์ของการสกัดคุณลักษณะแสดงในตารางที่ 20

ตารางที่ 20 ตัวอย่างการสกัดคุณลักษณะของการแบ่งขอบเขตอนุพากย์ปริจเฉท

คำ	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
ถ้า/conj	1	0	0	0	0	0	0	0	conj
ระบาด/vi	0	0	0	0	0	0	0	0	vi
รุนแรง/adv	0	0	0	0	0	0	0	0	adv
จะ/prev	0	0	1	0	0	0	1	0	prev
ทำให้/vcau	0	0	1	0	0	0	0	0	vcau
ข้าว/ncn	0	0	0	0	0	0	0	0	ncn
แห้ง/vi	0	0	0	0	0	0	0	0	vi
ตาย/vi	0	0	0	0	0	0	0	0	vi

การวัดประสิทธิภาพของระบบ

การวัดประสิทธิภาพของการแบ่งขอบเขตอนุพากย์ปริจเฉทจะอ้างอิงความถูกต้องจากเอกสารที่ผ่านการกำกับจากผู้เชี่ยวชาญ ซึ่งค่าที่ใช้ในการวัดประสิทธิภาพของระบบ ได้แก่ ค่าความถูกต้อง (precision) ค่าความระลึก (recall) และค่า F-measure โดยทั้ง 3 ค่า สามารถคำนวณได้ดังนี้

ค่าความระลึก มีการคำนวณดังนี้

$$R = \frac{\text{จำนวนขอบเขตของอนุพากย์ปริจเฉทที่กำกับได้ถูกต้องจากระบบ}}{\text{จำนวนขอบเขตของอนุพากย์ปริจเฉทที่ได้รับการกำกับจากเอกสารที่ใช้อ้างอิง}}$$

ค่าความถูกต้อง มีการคำนวณดังนี้

$$P = \frac{\text{จำนวนขอบเขตของอนุพากย์ปริจเฉทที่กำกับได้ถูกต้อง}}{\text{จำนวนขอบเขตของอนุพากย์ปริจเฉททั้งหมดที่ได้รับการกำกับจากระบบ}}$$

F-measure เป็นการเฉลี่ยค่าความถูกต้องในการตรวจพบและค่าความระลึกในการตรวจพบเข้าด้วยกัน จึงเปรียบเหมือนค่าวัดความแม่นยำโดยรวม มีการคำนวณค่าดังนี้

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

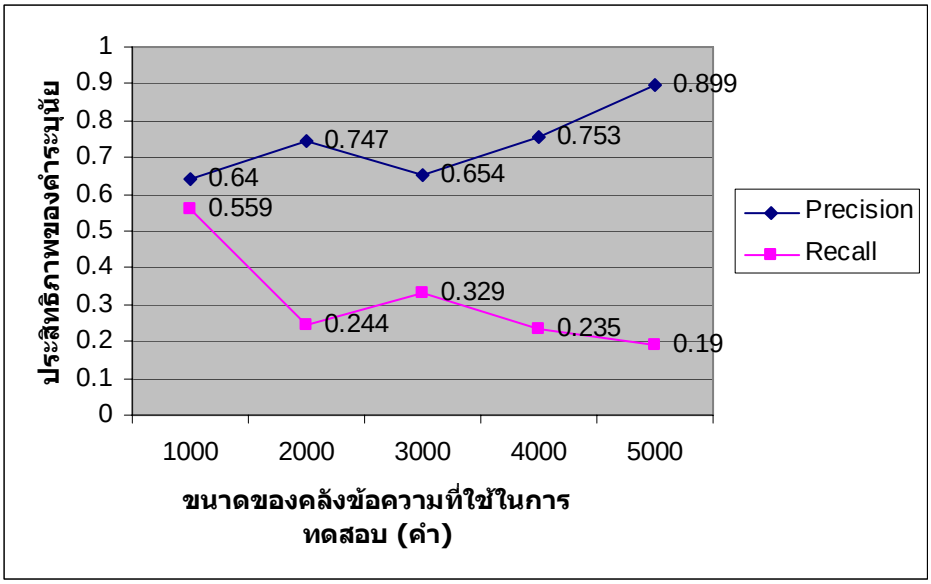
โดย β คือค่าพารามิเตอร์ที่แสดงสัดส่วนความสำคัญระหว่างค่าความถูกต้องและค่าความระลึก โดยทั่วไป จะใช้ค่า β เท่ากับ 1

ผลการทดลองและวิจารณ์

ผลการทดลอง

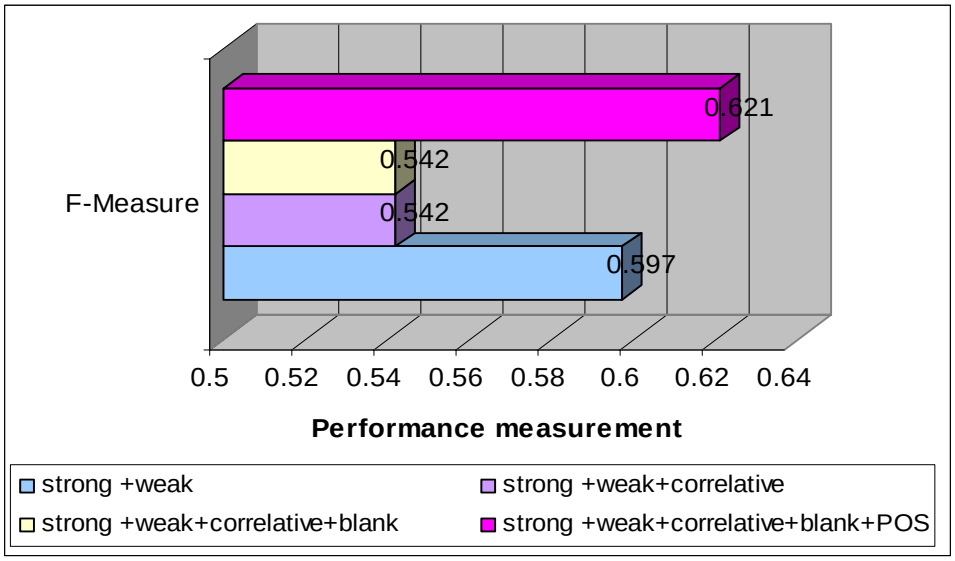
การทดลองวัดประสิทธิภาพของระบบ จะใช้เอกสารในโดเมนการเกษตร โดยแบ่งเป็นคลังเอกสารเพื่อฝึกฝนระบบขนาด 615 EDUs (6,000 คำ) และคลังเอกสารเพื่อทดสอบขนาด 220 EDUs (2,000 คำ) โดยการทดลองนี้มุ่งเน้นในการแบ่งขอบเขตของอนุพากย์ภาษาไทยให้มีความถูกต้องแม่นยำ ในกระบวนการ EDUs segmentation เราใช้เทคนิคการฝึกฝนและการเรียนรู้แบบต้นไม้ตัดสินใจโดยใช้ C4.5 จากซอฟต์แวร์ของ WEKA (http:// www.cs.waikato.ac.nz/ml/weka) คุณลักษณะที่ใช้ในการสร้างกฎสำหรับการแบ่งขอบเขตของอนุพากย์ในภาษาไทยแบ่งออกเป็น 4 คุณลักษณะ ได้แก่ ประเภทของคำระบุนัย (Discourse Segmentation Cues) คำระบุนัยแบบคู่ (Correlative Discourse Markers) ช่องว่าง และชนิดของคำ (POS)

การทดลองวัดประสิทธิภาพของคุณลักษณะประเภทของคำระบุนัยที่ใช้เป็นตัวแบ่งขอบเขตของอนุพากย์ปริจเฉท ซึ่งทำการทดลองเปรียบเทียบ ประสิทธิภาพของใช้ขนาดของคลังเอกสารจำนวน 615 EDUs (6,000 คำ) ปรากฏว่าคุณลักษณะประเภทของคำระบุนัยมีประสิทธิภาพในการแบ่งขอบเขตอนุพากย์ดี โดยคุณลักษณะนี้จะให้ค่าความถูกต้องสูงโดยเฉลี่ย 0.74 และมีค่าความระลึกโดยเฉลี่ย 0.31 จากผลการทดลองการวัดประสิทธิภาพของคุณลักษณะประเภทของคำระบุนัยแสดงในภาพที่ 15 ค่าความถูกต้องและค่าความระลึกดังกล่าว แสดงให้เห็นว่าถึงแม้ว่าคุณลักษณะประเภทของคำระบุนัยมีประสิทธิภาพในการแบ่งขอบเขตอนุพากย์ที่ดี แต่ในบางกรณีคุณลักษณะประเภทของคำระบุนัยนี้ยังมีความคลุมเครือ อีกทั้งยังแสดงให้เห็นว่าเอกสารที่ใช้ในการทดลองนี้ปรากฏคำระบุนัยจำนวนน้อย จึงทำให้ค่าความระลึกมีค่าต่ำ ดังนั้นเราจึงจำเป็นต้องใช้คุณลักษณะอื่นๆ มาเป็นคุณลักษณะร่วมในการแบ่งขอบเขตอนุพากย์ปริจเฉทในกรณีที่อนุพากย์ปริจเฉทปรากฏคำระบุนัยที่มีความคลุมเครือและกรณีที่อนุพากย์ปริจเฉทไม่ปรากฏคำระบุนัย



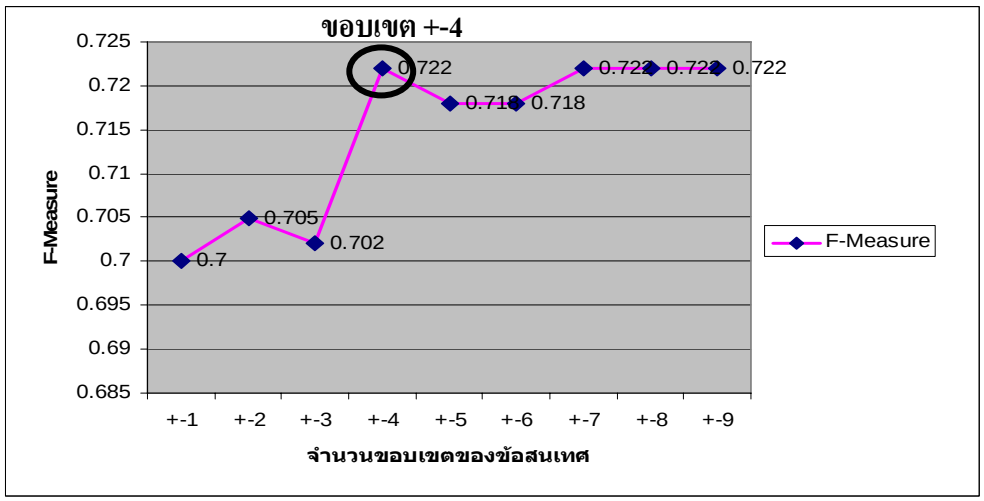
<u>ภาพที่ 15</u> ผลการทดลองการวัดประสิทธิภาพจากคุณลักษณะของคำระบุนัย

จากคุณลักษณะที่ใช้ในการฝึกฝนและเรียนรู้ระบบทั้ง 4 คุณลักษณะ ได้แก่ ประเภทของคำระบุนัย คำระบุนัยแบบคู่ ช่องว่าง และชนิดของคำ จึงทำการทดลองต่อเพื่อเปรียบเทียบประสิทธิภาพของคุณลักษณะทั้ง 4 คุณลักษณะที่ใช้ในการฝึกฝนและเรียนรู้ว่า คุณลักษณะใดบ้างที่นำใช้งานร่วมกันแล้ว จึงทำให้การแบ่งขอบเขตอนุพากย์ปริจเฉทมีประสิทธิภาพดีที่สุด จากผลการทดลองดังภาพที่ 16 แสดงให้เห็นว่าการนำคุณลักษณะประเภทของคำระบุนัย คำระบุนัยแบบคู่ ช่องว่าง และชนิดของคำมาใช้งานร่วมกันทั้ง 4 คุณลักษณะจะทำให้การแบ่งขอบเขตอนุพากย์ปริจเฉทมีประสิทธิภาพดีที่สุด จากผลการทดลองการนำคุณลักษณะทั้ง 4 คุณลักษณะมาใช้งานร่วมกัน จะให้ค่า F-measure เท่ากับ 0.621



ภาพที่ 16 ผลการทดลองการเปรียบเทียบประสิทธิภาพจากคุณลักษณะต่างๆที่ใช้ในการเรียนรู้

จากผลการทดลองการเปรียบเทียบประสิทธิภาพจากคุณลักษณะต่างๆที่ใช้ในการเรียนรู้ ดังภาพที่ 16 แสดงให้เห็นว่าการนำคุณลักษณะประเภทของคำระบุนัย คำระบุนัยแบบคู่ ช่องว่าง และชนิดของคำมาใช้งานร่วมกัน จะทำให้การแบ่งขอบเขตอนุพากย์ปริจเฉทมีประสิทธิภาพดี แต่ยังมีบางกรณีที่เกิดความคลุมเครือในการแบ่งขอบเขตอนุพากย์ปริจเฉท จึงทำให้แบ่งขอบเขตอนุพากย์ปริจเฉทได้ไม่ถูกต้อง ดังนั้นแบบจำลองการเรียนรู้และฝึกฝนจึงต้องการคุณลักษณะอื่นๆ



ภาพที่ 17 ผลการทดลองการเปรียบเทียบประสิทธิภาพจำนวนคำจากบริบทในขอบเขต ±1 คำ ถึง ±9 คำ

เพิ่มเติม เพื่อใช้คุณลักษณะเหล่านั้นมาช่วยลดความคลุมเครือในการแบ่งขอบเขตอนุพากย์ปริจเฉท โดยคุณลักษณะที่เพิ่มขึ้น คือ คุณลักษณะของคำบริบทรอบข้าง ดังนั้นงานวิจัยจึงทำการทดลองเพื่อหาว่าจำนวนคุณลักษณะจากคำจากบริบทในขอบเขตเท่าใด จึงทำให้การแบ่งขอบเขตอนุพากย์ปริจเฉทมีประสิทธิภาพดีที่สุด ซึ่งผลการทดลองดังกล่าวแสดงดังภาพที่ 17

การทดลองการเปรียบเทียบประสิทธิภาพของจำนวนคำจากบริบทในขอบเขต $\pm$1 คำ ถึง $\pm$9 คำ ในภาพที่ 21 ปรากฏว่าการใช้คุณลักษณะประเภทของคำระบุนัย คำระบุนัยแบบคู่ ช่องว่าง และชนิดของคำจากบริบทในขอบเขต $\pm$4 คำ จะทำให้ประสิทธิภาพการแบ่งขอบเขตอนุพากย์ปริจเฉทดีที่สุด จากผลการทดลองค่า F-measure ที่ได้มีค่าเท่ากับ 0.722

การทดลองวัดประสิทธิภาพของระบบ จะใช้เอกสารในโดเมนการเกษตรโดยใช้คลังเอกสารเพื่อทดสอบขนาด 2,000 คำ โดยใช้คุณลักษณะหลัก 4 คุณลักษณะ ได้แก่ ประเภทของคำระบุนัย คำระบุนัยแบบคู่ ช่องว่าง และชนิดของคำ โดยใช้คุณลักษณะดังกล่าวในบริบท $\pm$4 คำ ซึ่งกฎที่สร้างจากแบบจำลองต้นไม้ตัดสินใจมีทั้งหมด 32 กฎ และกฎจากผู้เชี่ยวชาญ 5 กฎ การวัดประสิทธิภาพของงานวิจัย จะวัดโดยใช้ค่าความถูกต้อง และค่าความระลึก โดยแสดงผลลัพธ์ไว้ในตารางที่ 21

กำหนดให้ $W = w_1, w_2 \ldots w_n$ เป็นข้อความที่ประกอบด้วยคำทั้งหมด n คำ และ w_i เป็นคำที่ระบบพิจารณาเพื่อกำหนดสถานะของ w_i ว่าเป็นจุดเริ่มต้นหรือจุดสิ้นสุดของอนุพากย์ปริจเฉท

<u>ตัวอย่างกฎที่ 1</u>

(1) if $w_i \in$ Strong-start-basic cue

then status(w_i): = starting EDUs

ตัวอย่างกฎที่ 1 เป็นกฎที่ใช้กำหนดสถานะของขอบเขตอนุพากย์ปริจเฉทชนิด Basic EDU โดยถ้า w_i เป็นสมาชิกของ Strong-start-basic cue ดังนั้นจึงกำหนดสถานะว่า w_i เป็นจุดเริ่มต้นของอนุพากย์ปริจเฉทชนิด Basic EDU

ตัวอย่างเช่น ข้อมูลเข้าเป็น “กะหล่ำปลีต้องการธาตุไนโตรเจน_/blank***เพื่อ***สร้างความเจริญเติบโต” ดังนั้นจึงกำหนดสถานะของคำว่า “เพื่อ” ว่าเป็นจุดเริ่มต้นของอนุพากย์ปริจเฉทชนิด Basic EDU เนื่องจาก คำว่า “เพื่อ” เป็นสมาชิกของ Strong-start-basic cue

ตัวอย่างกฎที่ 2

(2) EndPOS = {ncn, vt, vi, vcs, vcau, prep, conj, conjncl}

if ($w_i \notin$ Strong-start-basic cue $\wedge$ $w_i \notin$ Weak-start-basic cue $\wedge$ w_{i1+} = blank $\wedge$ $POS(w_{i+2}) \in$ EndPOS)

then status(w_i): = ending EDUs

ตัวอย่างกฎที่ 2 เป็นกฎที่ใช้กำหนดสถานะของ w_i ว่าเป็นจุดสิ้นสุดของอนุพากย์ปริจเฉทชนิด Basic EDU โดยที่ข้อความเข้านั้นไม่ปรากฏคำระบุนัยภายในข้อความ ดังนั้นจึงต้องใช้ข้อสนเทศอื่น เพื่อประกอบการตัดสินใจ ในการแบ่งขอบเขตอนุพากย์ปริจเฉท ซึ่งกฎที่ 2 นี้จะใช้ข้อสนเทศจากช่องว่าง และชนิดของคำรอบๆ w_i ซึ่งจะกำหนดสถานะให้ w_i เป็นจุดสิ้นสุดของอนุพากย์ปริจเฉท

ตัวอย่างเช่น ข้อความเข้าเป็น “กะหล่ำปลีมีอายุการเก็บเกี่ยว 90 ***วัน*** _/blank $\phi_{\text{กะหล่ำปลี}}$ ต้องการอากาศหนาว” โดยคำที่ระบบพิจารณาเป็น w_i คือ “วัน” ซึ่งระบบจะกำหนดให้ “วัน” มีสถานะเป็นจุดสิ้นสุดของอนุพากย์ปริจเฉท เนื่องจากคำว่า “วัน” มีเงื่อนไขตรงตามกฎที่ 2 คือ คำว่า “วัน” ไม่เป็นสมาชิกของ Strong-start-basic cue และ Weak-start-basic cue และ คำที่อยู่ต่อจาก w_i ไป 1 คำเป็นช่องว่าง อีกทั้งชนิดของคำที่อยู่ต่อจาก w_i ไป 2 คำ เป็นสมาชิกของ EndPOS ดังนั้น “วัน” จึงมีสถานะเป็นจุดสิ้นสุดของอนุพากย์ปริจเฉทตามกฎข้อที่ 2

ตารางที่ 21 ผลการทดลองของระบบ

ตำแหน่งการแบ่งขอบเขตอนุพากย์	ประสิทธิภาพของระบบ	
	ค่าความถูกต้อง (%)	ค่าความระลึก (%)
Start-Basic EDU	0.80	0.69
End-Basic EDU	0.77	0.69
Start-Embedded EDU	0.53	0.45
End-embed EDU	0.727	0.348

วิจารณ์

จากผลการทดลองในการระบุขอบเขตอนุพากย์ปริจเฉทแสดงให้เห็นว่า กฎที่ใช้ในการแบ่งขอบเขตอนุพากย์ทั้งกฎที่ได้จากการฝึกฝนและการเรียนรู้โดยเครื่องจักรกลและกฎที่ได้จากผู้เชี่ยวชาญ ยังไม่สามารถแบ่งขอบเขตอนุพากย์ได้ถูกต้องทุกกรณี

ตัวอย่างเช่น กรณีอนุพากย์ปริจเฉทปรากฏคำระบุนัยที่มีความคลุมเครือที่มีความกำกวมมาก ได้แก่ “และ”, “หรือ” ซึ่งคำระบุนัยที่มีความคลุมเครือเหล่านี้จะมีชนิดของคำเป็นสันธานที่สามารถทำหน้าที่เป็นคำเชื่อมได้ทั้งนามวลีหรือประโยคได้ ตัวอย่างแสดงดังประโยคข้างล่าง

(1) กะหล่ำปลีต้องการธาตุไนโตรเจน**และ**โปตัสเซียมสูง

(2) ตัวอ่อนมีสีน้ำตาล**และ**ด้านข้างมีลายสีเงินแวววาว

จากตัวอย่างข้างต้น ประโยคที่ 1 คำว่า “และ” เป็นคำระบุนัยที่มีความคลุมเครือ และมีชนิดของคำเป็นสันธาน ทำหน้าที่เชื่อมนามวลีระหว่างคำว่า “ไนโตรเจน” และ “โปตัสเซียม” ซึ่งประโยคที่ 2 คำว่า “และ” เป็นคำระบุนัยที่มีความคลุมเครือ และมีชนิดของคำเป็นสันธาน ทำหน้าที่เชื่อมประโยคระหว่างประโยค “ตัวอ่อนมีสีน้ำตาล” และ “ด้านข้างมีลายสีเงินแวววาว” จากการที่คำว่า “และ” เป็นคำระบุนัยที่มีความคลุมเครือ สามารถทำหน้าที่เป็นคำเชื่อมได้ทั้งนามวลีหรือประโยคได้นั้น จากความกำกวมดังกล่าวจึงส่งผลทำให้มีการแบ่งขอบเขตอนุพากย์ปริจเฉทผิดพลาด โดยระบบส่วนใหญ่จะระบุคำว่า “และ ดังกล่าวจะเป็นสันธาน หน้าที่เชื่อมประโยคระหว่างประโยค ดังนั้นระบบจึงระบุว่า คำว่า “และ” เป็นจุดเริ่มต้นของอนุพากย์ปริจเฉท

ผลลัพธ์จากการวัดประสิทธิภาพในการระบุขอบเขตอนุพากย์ปริจเฉทของแสดงให้เห็นว่า นอกจากพบปัญหาที่แสดงดังตัวอย่างข้างต้นแล้ว ปัญหาที่พบในการวิจัยนี้อีก คือ ปัญหาจากคลังเอกสารที่ใช้ในการฝึกฝนระบบ ยังมีข้อผิดพลาด ทั้งในส่วนของการตัดคำ การกำกับชนิดของคำ การกำกับนิพจน์ระบุนาม การกำกับนามวลี และการกำกับชนิดและขอบเขตของอนุพากย์ปริจเฉท ซึ่งต้องใช้เวลาและแรงงานจากนักภาษาศาสตร์ในการตรวจสอบและวิเคราะห์ โดยข้อผิดพลาดได้แก่ การแบ่งคำที่ไม่ถูกต้อง การกำกับขอบเขตของนิพจน์ระบุนาม การกำกับนามวลีที่ผิดพลาด รวมไปถึงผู้กำกับชนิดและขอบเขตแต่ละคน ยังมีความเห็นในการพิจารณาว่าขอบเขตของอนุพากย์ปริจเฉทอยู่ในตำแหน่งที่ต่างกัน จึงอาจทำให้เกิดความขัดแย้งในคลังเอกสารเพื่อฝึกฝนระบบได้ ข้อผิดพลาดในคลังเอกสารเหล่านี้ ส่งผลให้ขั้นตอนการเรียนรู้อาจมีความคลาดเคลื่อน จึงทำให้การแบ่งขอบเขตอนุพากย์ปริจเฉทอาจมีความคลาดเคลื่อนได้

สรุปและข้อเสนอแนะ

สรุป

จากผลการทดลอง การพัฒนาทฤษฎีโดยใช้แนวทางแบบผสมผสาน ระหว่างแนวทางเทคนิคการฝึกฝนและการเรียนรู้ และแนวทางการใช้กฎที่สร้างขึ้นโดยผู้เชี่ยวชาญ สามารถนำมาประยุกต์ใช้กับระบบการแบ่งขอบเขตอนุพากย์ปริจเฉทสำหรับเอกสารภาษาไทยได้เป็นอย่างดี โดยวิทยานิพนธ์นี้เป็นการพัฒนาทฤษฎีปริจเฉทและพัฒนาระบบต้นแบบเพื่อใช้ในการแบ่งขอบเขตอนุพากย์ปริจเฉทภาษาไทย โดยใช้ข้อมูลเพียงข้อสนเทศในระดับคำและวลีภายในเอกสาร คุณลักษณะที่ใช้ในการฝึกฝนและเรียนรู้จะใช้เทคนิคของต้นไม้ตัดสินใจ ซึ่งคุณลักษณะที่ใช้ในการฝึกฝนและเรียนรู้มีทั้งหมด 4 คุณลักษณะได้แก่ ประเภทของคำระบุนัย คำระบุนัยแบบคู่ ช่องว่าง และหน้าที่ของคำ และจะใช้คุณลักษณะดังกล่าวขอบเขตบริบทเท่ากับ w_{-4}-w_{+4} ; เมื่อ w_0 คือ คำที่กำลังพิจารณา ในขั้นตอนของการแบ่งขอบเขตอนุพากย์ปริจเฉทจากกฎที่สร้างขึ้นโดยผู้เชี่ยวชาญจะเป็นขั้นตอนการตรวจสอบการแบ่งขอบเขตอนุพากย์จากกฎที่ได้การฝึกฝนและเรียนรู้ของเครื่องจักรกล ซึ่งกฎเหล่านี้จะช่วยลดความผิดพลาดของการแบ่งขอบเขตอนุพากย์ปริจเฉท อีกทั้งขั้นตอนนี้ผู้เชี่ยวชาญจะสร้างกฎที่ระบบการฝึกฝนและเรียนรู้ของเครื่องจักรกลไม่สามารถเรียนรู้และสร้างเป็นกฎออกมาได้ โดยเฉพาะกรณีที่อนุพากย์ปริจเฉทไม่ปรากฏคำระบุนัย ซึ่งกฎเหล่านี้จะช่วยเพิ่มค่าความระลึกของการแบ่งขอบเขตอนุพากย์ปริจเฉท ความผิดพลาดของการแบ่งขอบเขตอนุพากย์ปริจเฉทเกิดจากการไม่ปรากฏคำระบุนัยภายในอนุพากย์ปริจเฉท และลักษณะโครงสร้างของภาษาไทยที่มีความซับซ้อน เนื่องจากประโยคส่วนใหญ่เป็นประโยคความซ้อน จึงทำให้แบ่งขอบเขตของอนุพากย์ปริจเฉทภาษาไทยผิดพลาด

ข้อเสนอแนะ

1. การนำเอาข้อมูลสารสนเทศอื่น เช่น ผลลัพธ์จากการใช้การแจกแจงประโยค มาพิจารณาร่วมเพื่อการตัดสินใจการแบ่งขอบเขตอนุพากย์ปริเฉทให้มีความถูกต้องมากยิ่งขึ้น

2. เพิ่มขนาดของคลังเอกสารที่ใช้ในการฝึกฝนและเรียนรู้ เพื่อทำให้กฎที่ได้มีความครอบคลุมการแบ่งขอบเขตอนุพากย์ปริจเฉทได้ทุกกรณี

3. เพิ่มจำนวนกฎที่สร้างจากผู้เชี่ยวชาญในการแบ่งขอบเขตอนุพากย์ปริจเฉทในภาษาไทย

4. แนวทางการเรียนรู้ที่ใช้ในงานวิจัยนี้ เป็นวิธีการฝึกฝนและการเรียนรู้แบบมีผลเฉลย (Supervised learning) ซึ่งได้ทำการเรียนรู้จากคลังเอกสารในโดเมนเดียว ถ้าหากมีการนำระบบไปใช้กับเอกสารในโดเมนอื่น ความถูกต้องของระบบอาจลดลง ดังนั้น หากต้องการนำระบบไปใช้กับเอกสารในโดเมนอื่นๆ ควรนำชุดเอกสารในโดเมนนั้นมาทำการเรียนรู้ใหม่ เพื่อทำให้ระบบสามารถแบ่งขอบเขตอนุพากย์ได้อย่างถูกต้องและมีประสิทธิภาพ

เอกสารและสิ่งอ้างอิง

กำชัย ทองหล่อ. 2545. **หลักภาษาไทย.** บริษัทรวมสาส์น (1997) จำกัด, กรุงเทพฯ.

นววรรณ พันธุเมธา. 2527. **ไวยากรณ์ไทย.** ม.ป.ท

วิจินตน์ ภาณุพงศ์. 2520. **โครงสร้างภาษาไทย: ระบบไวยากรณ์.** มหาวิทยาลัยรามคำแหง, กรุงเทพฯ.

วิไลศักดิ์ กิ่งคำ. ม.ป.ป. **อรรถศาสตร์ภาษาไทย**. แหล่งที่มา: http://cyberlab.lh1.ku.ac.th/elearn/faculty/human/ hm20/ lesson%207.htm. [01-09-2548]

อุปกิตศิลปสาร, พระยา. 2511. **หลักภาษาไทย**. พระนคร: ไทยวัฒนาพานิช.

Alonso, L. and I. Castellon. 2001. Towards a delimitation of discursive segment for Natural Language Processing applications. **In Proc. of International Workshop on Semantics, Pragmatics and Rhetorics**, San Sebastián.

Carlson, L. and D. Marcu. 2003. **Discourse Tagging Reference Manual**. Available Source: http://www.isi.edu/~marcu. [20-12-2004].

Carlson, L., D. Marcu, and M.E. Okurowski. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. **In Proc. of 2nd SIGDIAL Workshop on Discourse and Dialogue**, Denmark.

Chanlekha, H. and A. Kawtrakul. 2004. Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. **In Proc. of** IJCNLP , Hainan Island , China, 2004.

Haliday, M.A.K. and R. Ruqaiya. 1976. Cohesion in English. Hong Kong : Longman.

Hovy, E. (1993). Automated discourse generation using discourse structure relations. Artificia Intelligence 63:341-385.

Mann, W. and S. Thompson. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization, pp.243-281.

Marcu, D. 1997. **The Rhetorical Parsing, Summarisation and Generation of Natural Language Texts**. Ph.D. Thesis, University of Toronto.

Marcu, D. 1998. A Surface-based Approach to Identifying Discourse Markers and Elementary Textual Units in Unrestricted Texts, pp. 1-7. **In Proc. of COLING/ACL Workshop on Discourse Relations and Discourse Markers**, Montreal, Canada.

Marcu, D. 1999. A Decision-based Approach to Rhetorical Parsing, pp. 365-372. **In Proc. of 37th Annual Meeting of the Association for Computational Linguistics (ACL'1999)**, Maryland.

Marcu, D. 1999. **Instructions for Manually Annotating the Discourse Structure of Texts**. http://www.isi.edu/~marcu. [20-12-2004].

Marcu, D. 2000. **The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach,** pp. 395-448. Computational Linguistics

Marcu, D. 2000. **The Theory and Practice of Discourse Parsing and Summarization**. The MIT Press, A Bradford Book, MIT Press,Cambridge, Massachusetts, London, England.

Soricut, R. and D. Marcu. 2003. Sentence Level Discourse Parsing using Syntactic and Lexical Information. **In Proc. of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)**, Edmonton, Canada.

Thanh, H. L., G. Abeysinghe and C. Huyck. 2004. Automated Discourse Segmentation by Syntactic Information and Cue Phrases. **In Proc. of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2004),** Innsbruck, Austria.

Polanyi, L., C. Culy, M.V.D. Berg, G. L. Thione, and D. Ahn. 2004a. A Rule Based Approach to Discourse Parsing, pp. 108-117. **In Proc. of 5th SIGdial Workshop in Discourse and Dialogue**, Cambridge, USA.

Polanyi, L., C. Culy, M.V.D. Berg, G. L. Thione, and D. Ahn. 2004b. Sentential Structure and Discourse Parsing. **In Proc. of ACL2004 Workshop on Discourse Annotation**, Barcelona, Spain.

Pengphon, N., A. Kawtrakul, M. Suktarachan. 2002. Word Formation Approach to Noun Phrase Analysis for Thai. **In Proc. of Symposium of Natural Language Processing (SNLP'02)**, Thailand, 2002.

Sudprasert, S., A. Kawtrakul. 2003. Thai Word Segmentation based on Global and Local Unsupervised Learning. **In Proc. of National Computer Science and Engineering Conference'03 (NCSEC'03)**, Chonburi, Thailand, 2003

ภาคผนวก

ภาคผนวก ก

คำระบุนัย

คำระบุนัย

<u>ตารางผนวกที่ ก1</u> คำระบุนัยในการแบ่งขอบเขตอนุพากย์ปริจเฉทภาษาไทย

ชนิดของคำระบุนัย	จำนวนคำ	รายการของคำระบุนัย
Strong-start-basic cues	32	ขณะ,ขณะที่,ขณะเดียวกัน,จนกระทั่ง,จากนั้น, ดังนั้น,โดยเฉพาะ, ต่อมา, แต่, ถ้า, ถ้าหาก, ทั้ง, ทั้งนี้,นอกจากนี้, นอกจากนั้น,เนื่องจาก, เนื่องมาจาก, ในขณะเดียวกัน,เพราะ, เพื่อ, เพื่อให้, เมื่อ, แม้ว่า, รวมทั้ง, แล้ว, ส่วน, ส่วนใหญ่, สำหรับ, หลังจาก, หาก, อย่างไรก็ตาม, ว่า, ซึ่ง
Strong-end-basic cues	5	ก็ได้, ทีเดียว, เช่นกัน, ก็ตาม,นัก
Weak-start-basic cues	25	ทั้ง, มัก, ควร, จะ, เรียกว่า, คือ, เป็น, ให้, ทำให้, ที่, ให้, ตั้งแต่, โดยทั่วไป, โดย, จึง, ก็, โดยเฉพาะ, เนื่องจาก, (, ได้แก่, เช่น, จน, และ, หรือ , ตัวเลข
Weak-start-embedded cues	4	ที่, (, ได้แก่, เช่น
Weak-end-basic cues	11	อยู่, ขึ้น, ลง, ได้, ไป, มา, ออก, แล้ว,), ฯลฯ, เป็นต้น
Weak-end- embedded cues	3	) , เป็นต้น,ฯลฯ
Correlative discourse segmentation cues	10	(ถ้า, จะ), (ถ้า, ก็), (ถ้า, ให้), (หาก, จะ), (หาก, ก็) , (หาก, มัก), (เมื่อ, ก็), (เมื่อ, ให้), (เมื่อ,จะ), (เมื่อ, มัก)

ภาคผนวก ข

ตัวย่อชนิดของคำ

ตัวย่อชนิดของคำ

เนื่องจากภาษาไทยมีชนิดของคำอยู่หลายชนิด เช่น คำนาม, คำกริยา ดังนั้นภายในวิทยานิพนธ์ฉบับนี้ จึงได้มีการอ้างอิงการแบ่งชนิดของคำภาษาไทยและการใช้คำย่อชนิดของคำ โดยใช้หลักเกณฑ์การแบ่งชนิดของคำ จากห้องปฏิบัติการวิจัยเชี่ยวชาญเฉพาะการประมวลผลภาษาธรรมชาติและเทคโนโลยีสารสนเทศอัจฉริยะ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ซึ่งทำการแบ่งชนิดของคำออกเป็น 15 ชนิด โดยคำนาม มีจำนวนทั้งหมด 16 ชนิด แสดงดังตารางผนวกที่ ข1, คำกิริยามีจำนวน 8 ชนิด แสดงดังตารางผนวกที่ ข2, คำบ่งชี้มีจำนวน 5 ชนิด แสดงดังตารางผนวกที่ ข3, คำคุณศัพท์มีจำนวน 7 ชนิด แสดงดังตารางผนวกที่ ข4, คำลักษณนามมีจำนวน 1 ชนิด แสดงดังตารางผนวกที่ ข5, คำสันธานมีจำนวน 3 ชนิด แสดงดังตารางผนวกที่ ข6, คำบุพบทมีจำนวน 2 ชนิด แสดงดังตารางผนวกที่ ข7, คำอุทานมีจำนวน 1 ชนิด แสดงดังตารางผนวกที่ ข8, คำอุปสรรคมีจำนวน 3 ชนิด แสดงดังตารางผนวกที่ ข9, คำลงท้ายมีจำนวน 2 ชนิด แสดงดังตารางผนวกที่ ข10, คำปฏิเสธมีจำนวน 1 ชนิด แสดงดังตารางผนวกที่ ข11, เครื่องหมายวรรคตอนมีจำนวน 1 ชนิด แสดงดังตารางผนวกที่ ข12, สำนวนมีจำนวน 1 ชนิด แสดงดังตารางผนวกที่ ข13,คำบ่งชี้กรรมวาจกมีจำนวน 1 ชนิด แสดงดังตารางผนวกที่ ข14, สัญลักษณ์มีจำนวน 1 ชนิด แสดงดังตารางผนวกที่ ข15

ตารางผนวกที่ ข1 คำย่อและตัวอย่างของคำนาม

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 ข1 คำย่อและตัวอย่างของคำนาม (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 ข2 คำย่อและตัวอย่างของคำกิริยา

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 ข3 คำย่อและตัวอย่างของคำบ่งชี้

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 ข4 คำย่อและตัวอย่างของคำคุณศัพท์

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 ข5 คำย่อและตัวอย่างของคำลักษณนาม

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

ตารางผนวกที่ ข6 คำย่อและตัวอย่างของคำสันธาน

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

ตารางผนวกที่ ข7 คำย่อและตัวอย่างของคำบุพบท

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

ตารางผนวกที่ ข8 คำย่อและตัวอย่างของคำอุทาน

ชนิดของคำอุทาน	ตัวย่อ	ตัวอย่าง
Interjection	int	เอ๊ะ อ๋อ อุ๊ย ว้าย

ตารางผนวกที่ ข9 คำย่อและตัวอย่างของคำอุปสรรค

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

ตารางผนวกที่ ข10 คำย่อและตัวอย่างของคำลงท้าย

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

ตารางผนวกที่ ข11 คำย่อและตัวอย่างของคำปฏิเสธ

ชนิดของคำปฏิเสธ	ตัวย่อ	ตัวอย่าง
Negative	neg	ไม่, มิ, ไร้

ตารางผนวกที่ ข12 คำย่อและตัวอย่างของเครื่องหมายวรรคตอน

ชนิดของเครื่องหมายวรรคตอน	ตัวย่อ	ตัวอย่าง
Punctuation	punc	. - , : ;

ตารางผนวกที่ ข13 คำย่อและตัวอย่างของสำนวน

ชนิดของสำนวน	ตัวย่อ	ตัวอย่าง
idiom	idm	รักวัวให้ผูก รักลูกให้ตี

ตารางผนวกที่ ข14 คำย่อและตัวอย่างของคำบ่งชี้กรรมวาจก

ชนิดของคำบ่งชี้กรรมวาจก	ตัวย่อ	ตัวอย่าง
Passive Voice Marker	psm	ถูก, โดน

<u>ตารางผนวกที่ ข15</u> คำย่อและตัวอย่างของสัญลักษณ์

ชนิดของสัญลักษณ์	ตัวย่อ	ตัวอย่าง
Symbol	sym	ฯลฯ, ฯ

ภาคผนวก ค

กฎที่ได้จากการฝึกฝนและการเรียนรู้

กฎที่ได้จากการฝึกฝนและการเรียนรู้

จากการใช้แนวทางการฝึกฝนและเรียนรู้จากแบบจำลองต้นไม้ตัดสินใจ โดยใช้เทคนิค C4.5 จากซอฟต์แวร์ของ WEKA มาสร้างกฎเพื่อแบ่งขอบเขตอนุพากย์ปริจเฉท ซึ่งใช้คลังเอกสารสำหรับฝึกฝนจำนวน 615 EDUs (6,000 คำ) และคุณลักษณะที่ใช้ในการฝึกฝนระบบ ได้แก่ ประเภทของคำระบุนัย คำระบุนัยแบบคู่ ช่องว่าง และชนิดของคำ จากการฝึกฝนและการเรียนรู้ดังกล่าวแบบจำลองต้นไม้ตัดสินใจได้ทำการสร้างกฎในการแบ่งขอบเขตของอนุพากย์ปริจเฉทได้ทั้งหมด 32 กฎ ซึ่งรายละเอียดของแต่ละกฎแสดงดังภาพภาคนวกที่ ค1

```
        if weak_start_embed == 0:
            if strong_start == 0:
                if weak_start_basic == 0:
                    if strong_start1 == 0:
                        if blank1 == 0:
                            if blank_1 == 0:
                                if weak_start_basic1 == 0:
                                    output = nullEDU(surface,output) #กฎที่ 1
                                else:
                                    if blank == 0:
                                        text_startEDU =
"none|null|/adj|/punc|/npn|/det|/sym"
                                        reg_startEDU = re.compile(text_startEDU)
                                        result_startEDU  = reg_startEDU .findall(pos_1)

                                        text_nullEDU = "/conj|/vcau|/conjd"
                                        reg_nullEDU = re.compile(text_nullEDU)
                                        result_nullEDU  = reg_nullEDU .findall(pos_1)
                                        text_endEDU =
"/cpn|/blk:|/vi|/vt|/prec|/neg|/nnum|/prev|/vpost|/advm3|"
                                        text_endEDU = text_endEDU +
"/advm2|/qubo|/advm1|/nct|/indet|/conjncl|/quaf|/nlab|/pref3|/adjbe|"
                                        text_endEDU = text_endEDU +
"/norm|/part|//punc|/ncn_/blk|/|/prepc|/vex|/adèj"

                                        reg_endEDU = re.compile(text_endEDU)
                                        result_endEDU  = reg_endEDU .findall(pos_1)
```

ภาพผนวกที่ ค1 กฎที่ได้จากการฝึกฝนและการเรียนรู้

ภาพผนวกที่ ค1 (ต่อ)

```
                        text_endEmbed =
"/cpn|/blk:|/vi|/vt|/prec|/neg|/nnum|/prev|/vpost|/advm3|"
                        reg_endEmbed = re.compile(text_endEmbed)
                        result_endEmbed  = reg_endEmbed.findall(pos_1)

                        if result_startEDU :
                           output =  start_EDU(surface,output) #กฎที่ 2
                        elif result_nullEDU:
                           output = nullEDU(surface,output) #กฎที่ 3
                        elif result_endEDU:
                           output = end_EDU(surface,output) #กฎที่ 4
                        elif result_endEmbed:
                           output = end_Embedded(sur,outdata) #กฎที่ 5
                        elif pos_1 == "/vcs":
                           text_nullEDU_vcs = "/vcs|/vcau"
                           reg_nullEDU_vcs =
re.compile(text_nullEDU_vcs)
                           result_nullEDU_vcs  = reg_nullEDU_vcs
.findall(pos1)

                           text_endEDU_vcs =
"/blk|/ncn|/npn|/nct|/cpn|/vi|/vt|/vpost|/vex|/adv|"
                           text_endEDU_vcs = text_endEDU_vcs +
"/adj|/advm3|/advm2|/advm1|/adjbe|/adj|/conj|/conjncl|/conjd|/prec|"
                           text_endEDU_vcs = text_endEDU_vcs +
"/prep|/prev|/pref3|/punc|/prel|/pref1|/psm|/part|//punc|/prepc|"
                           text_endEDU_vcs = text_endEDU_vcs +
"/neg|/nnum|/nlab|/norm|/cl|/qubo|/qube|/quaf|/indet|/det|"
                           text_endEDU_vcs = text_endEDU_vcs +
"none|null|/ncn_/blk|/|/sym"
                           reg_endEDU_vcs =
re.compile(text_endEDU_vcs)
                           result_endEDU_vcs  = reg_endEDU_vcs
.findall(pos1)

                           if result_nullEDU_vcs:
                              output = nullEDU(surface,output)#กฎที่ 6
                           elif result_endEDU_vcs:
                              output = end_EDU(surface,output)#กฎที่ 7
                           else:
                              output = nullEDU(surface,output)#กฎที่ 8
```

ภาพผนวกที่ ค1 (ต่อ)

```
                        elif pos_1 == "/ncn":
                            text_nullEDU_ncn_1 =
"|/blk|/vcs|/vcau|/vex|/conj|/conjncl|/conjd|/adj|/advm3|/advm2|"
                            text_nullEDU_ncn_1 = text_nullEDU_ncn_1 +
"/advm1|/adjbe|/adèj|/prec|/prep|/prev|/punc|/prel|/pref1|/qubo|"
                            text_nullEDU_ncn_1 = text_nullEDU_ncn_1 +
"/qube|/det|/quaf|/psm|/pref3|/part|//punc|/prepc|/npn|/nct|"
                            text_nullEDU_ncn_1 = text_nullEDU_ncn_1 +
"/neg|/nnum|/cl|/nlab|/norm|/ncn_/blk|/|/sym"
                            reg_nullEDU_ncn_1 =
re.compile(text_nullEDU_ncn_1)
                            result_nullEDU_ncn_1  = reg_nullEDU_ncn_1
.findall(pos)

                            text_endEDU_ncn_1 = "/vi|/vpost|/adv|/indet"
                            reg_endEDU_ncn_1 =
re.compile(text_endEDU_ncn_1)
                            result_endEDU_ncn_1  = reg_endEDU_ncn_1
.findall(pos)

                            if pos == "/cpn":
                               output = start_EDU(surface,output)#กฎที่ 9
                            elif result_nullEDU_ncn_1:
                               output = nullEDU(surface,output)#กฎที่ 10
                            elif result_endEDU_ncn_1:
                               output = end_EDU(surface,output)#กฎที่ 11
                            elif pos == "/ncn":
                               text_nullEDU_ncn =
"|none|null|/blk|/cpn|/npn|/vi|/vpost|/vcau|/vex|/conjncl|"
                               text_nullEDU_ncn = text_nullEDU_ncn +
"/conjd|/adv|/adj|/advm3|/advm2|/advm1|/adjbe|/adèj|/prev|/prec|"
                               text_nullEDU_ncn = text_nullEDU_ncn +
"/punc|/prel|/pref1|/qubo|/qube|/quaf|/psm|/pref3|/part|//punc|"
                               text_nullEDU_ncn = text_nullEDU_ncn +
"/prepc|/cl|/indet|/det|/|/sym|/neg|/nnum|/nct|/nlab|"
                               text_nullEDU_ncn = text_nullEDU_ncn +
"/norm|/ncn_/blk|conj"
                               reg_nullEDU_ncn =
re.compile(text_nullEDU_ncn)
                               result_nullEDU_ncn  = reg_nullEDU_ncn
.findall(pos_2)
```

ภาพผนวกที่ ค1 (ต่อ)

```
                                text_endEDU_ncn = "/vcs|/ncn|/vt"
                                reg_endEDU_ncn =
re.compile(text_endEDU_ncn)
                                result_endEDU_ncn  = reg_endEDU_ncn
.findall(pos_2)

                                if result_nullEDU_ncn:
                                   output = nullEDU(surface,output)#กฎที่ 12
                                elif result_endEDU_ncn:
                                   output = end_EDU(surface,output)#กฎที่ 13
                                elif pos_2 == "/prep":
                                   if pos1 == "/vcs":
                                      output = start_EDU(surface,output) #กฎที่
                                   elif pos1== "/vcau":
                                      output = end_EDU(surface,output)
                                        # กฎที่ 14
                                   else:
                                      output = nullEDU(surface,output)#กฎที่ 15

                             elif pos == "/vt":
                                if ((pos_2 == "/conj") or (pos_2 == "/vt") ):
                                   output = end_EDU(surface,output)#กฎที่ 16
                                else:
                                   output = nullEDU(surface,output)#กฎที่ 17
                             else:
                                 output = nullEDU(surface,output)#กฎที่ 18
                          elif pos_1 == "/adv":
                             if ((pos == "/cpn") or (pos == "/ncn")):
                                output = start_EDU(surface,output)#กฎที่ 19
                             elif (pos == "/punc"):
                                output = nullEDU(surface,output)#กฎที่ 20
                             else:
                                output = end_EDU(surface,output)#กฎที่ 21
                          elif pos_1 == "/prep":
                             if (strong_start_4 == 0):
                                output = nullEDU(surface,output)#กฎที่ 22
                             else:
                                output = end_EDU(surface,output)#กฎที่ 23
```

<u>ภาพผนวกที่ ค1</u> (ต่อ)

```
                        elif pos_1 == "/cl":
                            if pos_2 == "/blk":
                                output = start_EDU(surface,output)#กฎที่ 24
                            else:
                                output = nullEDU(surface,output)#กฎที่ 25
                        elif pos_1 == "/pref1":
                            if ((pos_2 == "/vcs") or (pos_2 == "/vt")):
                                output = end_EDU(surface,output)#กฎที่ 26
                            else:
                                output = nullEDU(surface,output)#กฎที่ 27
                        elif pos_1 == "/qube":
                            if blank_3 == 0:
                                output = nullEDU(surface,output)#กฎที่ 28
                            else:
                                output = end_EDU(surface,output)#กฎที่ 29
                        else:
                            output = nullEDU(surface,output)#กฎที่ 30
                    else:
                        output = nullEDU(surface,output) #กฎที่ 31
    else:
        output = nullEDU(surface,output) #กฎที่ 32
```

ภาคผนวก ง

การฝึกฝนและการเรียนรู้โดยใช้เครื่องจักรกล

การฝึกฝนและเรียนรู้โดยเครื่องจักรกล

การฝึกฝนและการเรียนรู้โดยเครื่องจักรกล (Machine learning) เป็นสาขาหนึ่งของปัญญาประดิษฐ์ โดยงานวิจัยนี้ได้นำแนวทางการฝึกฝนและการเรียนรู้ของด้วยเครื่องจักรกลมาใช้ในการแบ่งขอบเขตอนุพากย์ปริจเฉทสำหรับภาษาไทย ซึ่งการฝึกฝนและการเรียนรู้ของด้วยเครื่องจักรกลจะเน้นวิธีการสร้างกฎจากการวิเคราะห์ข้อมูลจากคลังเอกสาร ตัวอย่างเทคนิคการฝึกฝนและการเรียนรู้โดยเครื่องจักรกลที่ใช้กับการแบ่งขอบเขตอนุพากย์ปริจเฉท ได้แก่ ต้นไม้ตัดสินใจ (Decision tree) และการเรียนรู้เบย์อย่างง่าย (Naïve Bayes)

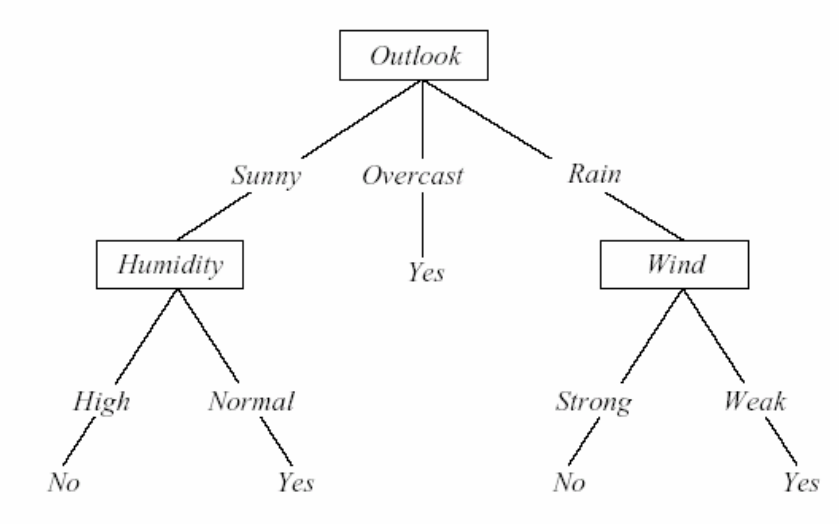
ต้นไม้ตัดสินใจ (Decision Tree)

การเรียนรู้แบบต้นไม้ตัดสินใจ (decision tree learning) เป็นการเรียนรู้ที่ใช้การแทนความรู้อยู่ในรูปของต้นไม้ตัดสินใจที่ประกอบด้วยบัพ (node) และ กิ่ง (link) ที่ต่อกับบัพ ดังแสดงในภาพภาคผนวกที่ ง1 บัพที่ปลายสุดเรียกว่าบัพใบ โดยบัพแสดงคุณสมบัติและกิ่งแสดงค่าของคุณสมบัตินั้น ส่วนบัพใบจะแสดงประเภท การสร้างต้นไม้ตัดสินใจทำโดยสร้างทีละบัพเพื่อตรวจสอบคุณสมบัติของตัวอย่าง แล้วแยกตัวอย่างลงตามค่าของกิ่ง ทำจนกระทั่งตัวอย่างในใบแต่ละใบอยู่ในประเภทเดียวกันทั้งหมด สำหรับการเลือกคุณสมบัติเพื่อมาสร้างเป็นบัพนั้น นิยมใช้ค่า Entropy และ Gain ของตัวอย่าง โดยสมการสำหรับคำนวณค่าทั้งสองนั้นเป็นดังนี้

$$Entropy(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_- \tag{1}$$

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

ตัวอย่างอัลกอริทึมที่เป็นแบบต้นไม้ตัดสินใจ เช่น CART ID3 C4.5 และ CHAID ข้อเสียของวิธีการแบบนี้คือ ไม่เหมาะกับระบบที่มีจำนวนคุณลักษณะมาก ๆ



<u>ภาพผนวกที่ ง1</u> ต้นไม้ตัดสินใจการเล่นเทนนิสโดยดูจากสภาพอากาศ

ข้อดีของแบบจำลองต้นไม้ตัดสินใจ คือ กฎที่สร้างจากแบบจำลองนี้จะง่ายต่อการเข้าใจและตีความผลลัพธ์ที่ได้จากการเรียนรู้ อีกทั้งยังสามารถรองรับข้อมูลได้หลายประเภททั้งข้อความเชิงกลุ่ม และข้อมูลเชิงปริมาณ และคำนวณของแบบจำลองนี้ยังไม่ซับซ้อนมาก ทำให้ใช้เวลาในการประมวลผลไม่มากนัก

การเรียนรู้เบย์อย่างง่าย (Naïve Bayes)

การเรียนรู้เบย์อย่างง่ายเป็นการเรียนรู้ที่อาศัยหลักการทางสถิติและความน่าจะเป็น โดยผลลัพธ์ที่ให้ค่าความน่าจะเป็นสูงที่สุดเป็นคำตอบของการรู้จำ ถ้าให้คำตอบที่เป็นไปได้ทั้งหมดคือ $v_1, v_2, ..., v_K$ และข้อมูลที่ใช้ในการตัดสินใจคือ $D = \langle a_1, a_2, ..., a_T \rangle$ หลักในการตัดสินใจคือเลือกตำตอบที่มีค่าความน่าจะเป็นสูงสุด ซึ่งเขียนเป็นสมการได้ดังนี้

$$v = \arg\max_{v \in \{v_1, v_2, ..., v_K\}} P(v \mid D)$$

$$v = \arg\max_{v \in \{v_1, v_2, ..., v_K\}} \frac{P(D \mid v)P(v)}{P(D)}$$

$$v = \arg\max_{v \in \{v_1, v_2, ..., v_K\}} P(D \mid v)P(v)$$

อย่างไรก็ตามการโมเดลความค่าจะเป็น $P(D \mid v)$ โดยตรงนั้นทำได้ยากเนื่องจากเป็นตัวแปรหลายมิติการเรียนรู้แบบ Naïve Bayes จึงตั้งสมมุติฐานว่าข้อมูลเข้าแต่ละตัวแปรไม่ขึ้นต่อกัน ซึ่งสมมุติฐานนี้ทำให้สามารถเขียนสมการใหม่ ได้เป็น

$$v = \underset{v \in \{v_1, v_2, ..., v_K\}}{\arg\max} \; P(v) \cdot \prod_{i=1}^{T} P(a_i \mid v)$$

จากข้อมูลการเล่นเทนนิสดังข้อมูลเราสามารถคำนวณการเล่นเทนนิสได้ดังนี้ สมมติต้องการทราบว่า ณ สภาพอากาศ Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong แล้วจะเล่นเทนนิสหรือไม่?

จาก

$$v_{NB} = \underset{v_k \in [yes, no]}{\arg\max} P(v_k) \prod_{t} P(a_t \mid v_k)$$

$$v_{NB} = \underset{v_k \in [yes, no]}{\arg\max} P(v_k) P(Outlook = sunny \mid v_k) P(Temp = cool \mid v_k)$$
$$P(Humidity = high \mid v_k) P(Wind = strong \mid v_k)$$

วิธีทำ

$$P(PlayTennis = yes) = 9/14 = 0.64$$
$$P(PlayTennis = no) = 5/14 = 0.36$$
$$P(Wind = strong \mid PlayTennis = yes) = 3/9 = 0.33$$
$$P(Wind = strong \mid PlayTennis = no) = 3/5 = 0.60$$
$$...$$
$$P(yes) P(sunny \mid yes) P(cool \mid yes) P(high \mid yes) P(strong \mid yes) = 0.0053$$
$$P(no) P(sunny \mid no) P(cool \mid no) P(high \mid no) P(strong \mid no) = \mathbf{0.0206}$$

คำตอบก็คือ “no” เพราะว่าค่าความน่าจะเป็นในการเกิด no มากกว่า yes

$$\Rightarrow answer : PlayTennis(x) = no$$

<u>ตารางผนวกที่ ง1</u> ตารางเปรียบเทียบข้อดีและข้อเสียของแต่ละเครื่องจักรเรียนรู้

เครื่องจักรเรียนรู้	ข้อเด่น	ข้อด้อย
Decision tree	1. แทนความรู้ที่ได้เป็นต้นไม้ทำให้แปลความหมายได้ง่าย (เข้าใจได้ง่าย) 2. สามารถนำกฎที่สร้างขึ้นไปพัฒนาต่อได้	1. ไม่เหมาะกับข้อมูลที่มีลักษณะที่เป็นข้อมูลต่อเนื่อง (continuous data) 2. เมื่ออัลกอริทึมที่ได้ทำการวิเคราะห์หาตัวแบ่ง (gain value)แล้วจะไม่สนค่าที่ตัดทิ้งไปเลย ซึ่งอาจมีความสำคัญต่อการตัดสินใจ 3. ไม่เหมาะกับต้นไม้ที่มีความลึก(tree level)มากๆเพราะโหนดจะถูกแตกเป็นชิ้นเล็กและไม่ค่อยมีประโยชน์ในการวิเคราะห์ต่อ
Naïve Bayes	1. เป็นอัลกอลิธึมที่ง่ายและเร็วในการคำนวณ 2. เป็นเครื่องจักรเรียนรู้ที่เหมาะกับการแบ่งแยกประเภทข้อมูล(classification problem) 3. หารูปแบบความสัมพันธ์ไม่ซับซ้อน	1. ความถูกต้องของข้อมูลจะมากถ้าข้อมูลไม่ขึ้นต่อกัน(independence data) 2. ไม่เหมาะกับข้อมูลแบบต่อเนื่อง

ภาคผนวก จ

ตัวอย่างการแบ่งขอบเขตอนุพากย์ปริจเฉทของภาษาอังกฤษ

ตัวอย่างการแบ่งขอบเขตอนุพากย์ปริจเฉทของภาษาอังกฤษ

<u>ตารางผนวกที่ จ1</u> ตัวอย่างการแบ่งขอบเขตอนุพากย์ปริจเฉทของภาษาอังกฤษ

โครงสร้างทางไวยากรณ์	ตัวอย่าง	หมายเหตุ
Main Clause	[The company will shut down its plant.]$_{\text{Main clause}}$	-
Subordinate Clause	[The company will shut down its plant] [although it will not dismiss any employees.]$_{\text{Subordinate Clause}}$	-
Coordinated Sentence	[The company will shut down its plant,] [and it will dismiss several hundred employees.] $_{\text{Coordinated Sentence}}$	-
Complements of Attribute Verbs	[The company says] [it will shut down its plants]$_{\text{Complements}}$	กริยาต้องเป็นกริยาที่แสดงอาการ อารมณ์ หรือความรู้สึกเท่านั้น เช่น say, feel, think, believe
Correlative Subordinators	[No sooner had they announced the closing of plants] [than massive protests erupted on the premises.]$_{\text{Correlative Subordinators}}$	-
Embedded Discourse units	<u>relative clause :</u> [The plant] {that the company will shutdown}$_{\text{Embedded}}$ [is in Ohio] <u>parentheticals:</u> [The plant]{(which is in Ohio)}$_{\text{Embedded}}$ [will b shut down in October]	มีโครงสร้างไวยากรณ์เป็น relative clause, nominal post modifiers, appositives, parenthetical
Discourse-Salient Phrase	[Today, no one gets in or out of the restricted area] [without De Beers's stingy approval]$_{\text{Discourse-Salient Phrase}}$	แต่ละอนุพากย์ปริจเฉทต้องเริ่มต้นวลีด้วยเซตของคำระบุนัยที่ไม่มีความคลุมเครือ โดยคำระบุนัยดังกล่าวจะแสดงถึงชนิดปริจเฉทสัมพันธ์

หมายเหตุ

การแบ่งขอบเขตของแต่ละอนุพากย์ปริจเฉท Basic EDU จะถูกแบ่งโดยใช้เครื่องหมายวงเล็บ (“[ ]”) ส่วนอนุพากย์ปริจเฉทชนิด Embedded EDU จะถูกแบ่งโดยใช้เครื่องหมายวงเล็บปีกกา (“{ }”)

ประวัติการศึกษาและการทำงาน

ชื่อ –นามสกุล	นางสาวจิรวรรณ เจริญสุข
วัน เดือน ปี ที่เกิด	20 ธันวาคม พ.ศ. 2522
สถานที่เกิด	ชลบุรี
ประวัติการศึกษา	วท.บ. (วิทยาการคอมพิวเตอร์) มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา (พ.ศ. 2544)
ตำแหน่งหน้าที่การงานปัจจุบัน	อาจารย์
สถานที่ทำงานปัจจุบัน	คณะทรัพยากรและสิ่งแวดล้อม มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา