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

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

177369

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.