

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันสื่อข้อมูลข่าวสารทางอิเล็กทรอนิกส์เกิดขึ้นมากมาย และเป็นแหล่งรวบรวมข่าวสารจำนวนมาก ข่าวสารเหล่านี้มีเหตุการณ์สำคัญต่างๆในแต่ละช่วงเวลาเกิดขึ้นจะถูกบันทึกเก็บไว้ หากผู้ใช้ต้องการสืบค้นว่ามีใครทำกิจกรรมอะไรบ้างในช่วงเวลาที่ผ่านมาจะใช้เวลาค่อนข้างมากในอ่านและสรุป การสกัดนิพจน์ระบุนามสำหรับเหมืองข้อมูลข่าวภาษาไทยตามกิจกรรมของบุคคล เป็นอีกวิธีหนึ่งที่ถูกนำมาใช้ในการแก้ปัญหา การสกัดนิพจน์ระบุนามเป็นงานวิจัยสาขาหนึ่งทางด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing) เป็นวิธีการทำให้คอมพิวเตอร์สามารถเข้าใจภาษาธรรมชาติของมนุษย์ สำหรับการสกัดนิพจน์ระบุนามภาษาไทย ปัญหาหลักก็คือการกำหนดตำแหน่งและขอบเขตของนิพจน์ระบุนาม เนื่องจากภาษาไทยไม่มีอักขระหรือเครื่องหมายที่บ่งบอกตำแหน่งและขอบเขตที่แน่นอน อีกทั้งยังมีคำที่เกิดขึ้นใหม่จากการทับศัพท์คำภาษาต่างประเทศ และคำที่เกิดจากการผสมคำใหม่อยู่ตลอดเวลา

งานวิจัยทางด้านการสกัดนิพจน์ระบุนามจะเน้นที่การสกัดนิพจน์ระบุนามประเภทชื่อบุคคล, ชื่อสถานที่ และชื่อองค์กร โดยใช้เทคนิคต่างๆ เช่น ใช้กฎร่วมกับคลังคำศัพท์ และใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ปัจจุบันงานวิจัยทางด้านการสกัดนิพจน์ระบุนามไม่ได้นำสิ่งที่สกัดได้มาประยุกต์ใช้ หากดึงเอาความสัมพันธ์ที่ถูกซ่อนไว้ในบริบทที่อยู่รอบข้างของนิพจน์ระบุนาม จะทำให้ข้อมูลมีมูลค่าเพิ่มและใช้ประโยชน์ได้อย่างมีประสิทธิภาพ

สำหรับงานวิจัยนี้เป็นการนำเสนอวิธี การสกัดนิพจน์ระบุนามสำหรับทำเหมืองข้อมูลข่าวภาษาไทยตามกิจกรรมของบุคคล โดยมีกระบวนการและขั้นตอนของการสร้างเริ่มจากนำบทความข่าวภาษาไทย จากการวัดเปรียบเทียบสมรรถนะเพื่อพัฒนามาตรฐานการประมวลผลภาษาไทยปี 2009 (Benchmark for Enhancing the Standard of Thai language processing :BEST2009) มาทำกับนิพจน์ระบุนามประเภทบุคคลด้วยวิธีของ Text Encoding Initiative Consortium (TEI) ต่อมาสร้างไฟล์เพื่อฝึกฝนระบบด้วยเทคนิค Conditional Random Fields (CRFs) โดยใช้คุณสมบัติคำบ่งชี้ ผลลัพธ์ที่ได้จากการฝึกฝนระบบจะเป็นไฟล์ของโมเดลการเรียนรู้จำนิพจน์ระบุนาม ต่อมานำบทความข่าวจากการวัดเปรียบเทียบสมรรถนะเพื่อพัฒนามาตรฐานการประมวลผลภาษาไทย (BEST2009) มาตัดคำด้วย 2 เทคนิค เทคนิคแรก ใช้ทีเล็กซ์ (Tlexs) ในการ

ตัดคำเพื่อสร้างไฟล์สำหรับทดสอบระบบ จากนั้นทำการสกัดนิพจน์ระบุนามประเภทบุคคลและสร้างคลังข้อมูลนิพจน์ระบุนามที่สกัดได้ เทคนิคที่สองทำการตัดและกำกับชนิดของคำ ด้วยเทคนิคสวอธ (Swath) เลือกคำที่กำกับชนิดเป็นคำกริยา (Verb) ทำการคัดกรองเลือกคำกริยาที่เกี่ยวข้องกับการเมืองนำมาสร้างเป็นคลังคำศัพท์ ต่อมานำบทความข่าวการเมืองภาษาไทยมาทำการตัดคำด้วยเทคนิค เล็กซ์โต (Lexto) และทำการสกัดนิพจน์ระบุนามตามกิจกรรมของบุคคลโดยใช้กฎร่วมสกัด ผลลัพธ์ที่ได้สร้างเป็นฐานข้อมูลบุคคลตามกิจกรรม ขั้นตอนสุดท้ายนำเสนอความสัมพันธ์ระหว่างนิพจน์ระบุนามประเภทบุคคลตามกิจกรรม สำหรับการประเมินประสิทธิภาพของระบบจะทำการเปรียบเทียบกับผู้เข้ารับการทดลองจำนวน 5 ท่านโดยจะยึดการสกัดนิพจน์ระบุนามประเภทบุคคลตามกิจกรรมที่สกัดโดยมนุษย์เป็นเกณฑ์มาตรฐาน และประเมินผลจากค่าความระลึก (Recall) ค่าความแม่นยำ (Precision) และ ค่าประสิทธิภาพโดยรวมของระบบ (F-measure)

1.2 วัตถุประสงค์ของการวิจัย

1. ศึกษาแนวทางการสกัดนิพจน์ระบุนาม(Named Entity Extraction)ของข่าวภาษาไทยด้วยเทคนิค Conditional Random Fields (CRFs)
2. ศึกษาแนวทางการประยุกต์ใช้นิพจน์ระบุนามกับเหมืองข้อความ (Text Mining) ข่าวภาษาไทย

1.3 ขอบเขตการวิจัย

1. บทความข่าวภาษาไทยที่จะนำมาใช้ในการสกัดนิพจน์ระบุนาม เป็นบทความข่าวจากการวัดเปรียบเทียบสมรรถนะเพื่อพัฒนามาตรฐานการประมวลผลภาษาไทย ปี 2009 (Benchmark for Enhancing the Standard of Thai language processing : BEST2009)
2. กำกับนิพจน์ระบุนามประเภทบุคคลโดยจะใช้วิธีการของ Text Encoding Initiative Consortium (TEI) มาประยุกต์ใช้กับบทความข่าวภาษาไทย
3. การสกัดนิพจน์ระบุนามตามกิจกรรมของบุคคลด้วยมนุษย์ จะให้ผู้เข้าร่วมการทดลองจำนวน 5 ท่าน โดยกำหนดให้ผู้เข้าร่วมการทดลองขีดเส้นใต้คำที่เป็นชื่อบุคคลตามกิจกรรม ที่เกิดขึ้นภายในบทความข่าวการเมืองภาษาไทย

จำนวน 30 บทความ

4. ใช้การเรียนรู้ของเครื่อง (Machine Learning) เทคนิค Conditional Random Fields (CRFs) ด้วยคุณสมบัติตัวบ่งชี้เพื่อสร้างโมเดลไฟล์สำหรับการสกัดนิพจน์ระบุนาม
5. ใช้เทคนิคสวอธ (Swath) , เทคนิคทีเล็กซ์ (Tlexs) และเทคนิคเล็กซ์โต (Lexto) เป็นเครื่องมือในการตัดและกำกับชนิดของคำ
6. สกัดนิพจน์ระบุนามตามกิจกรรมของบุคคลโดยใช้กฎร่วมในการสกัด
7. การวัดผลจะประเมินจากการสกัดนิพจน์ระบุนามตามกิจกรรมของบุคคล จากค่าความถูกต้อง (Precision) ค่าความระลึก (Recall) และค่าประสิทธิภาพโดยรวมของระบบ (F-measure) เปรียบเทียบกับผู้เข้ารับการทดลอง 5 ท่าน

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถเข้าใจหลักการทำงานของเทคนิค Conditional Random Fields (CRFs) ในการสกัดนิพจน์ระบุนามสำหรับข่าวภาษาไทย
2. สามารถเข้าใจหลักการทำงานของเหมืองข้อความ (Text Mining) ในการประยุกต์ใช้กับข้อมูลนิพจน์ระบุนามสำหรับข่าวภาษาไทย
3. สามารถนำไปเป็นกรณีศึกษา สำหรับใช้ในงานวิจัยที่เกี่ยวข้องกับการสกัดนิพจน์ระบุนาม และเหมืองข้อความ(Text Mining)

1.5 นิยามคำศัพท์เฉพาะ

1. Named Entity (NE, นิพจน์ระบุนาม) หมายถึง คำนามที่เป็นชื่อเฉพาะ ในการเรียกชื่อบุคคล ชื่อกลุ่มบุคคล ชื่อองค์กร ชื่อสถานที่ กลุ่มคำบอกเวลา และกลุ่มคำที่บอกปริมาณ
2. Text Mining (เหมืองข้อความ) หมายถึงกระบวนการค้นหารูปแบบของความสัมพันธ์ที่ซ่อนอยู่ภายในบทความนั้น