บทคัดย่อ

การประมวลผลภาษาธรรมชาติสำหรับภาษาไทยนั้นเป็นเรื่องที่ยากเนื่องจากภาษาไทยเป็นภาษาที่ไม่เข้มงวดในโครงสร้างของไวยกรณ์ แต่เน้นเอาความหมายเป็นหลัก คำที่ถูกสร้างขึ้นมาใหม่ส่วนใหญ่เกิดขึ้นจากเป็นการนำเอาคำโดดมาผสมกันเพื่อให้เกิดเป็นคำใหม่ที่อาจจะมีความหมายคล้ายกับคำเดิมหรือแตกต่างไป ซึ่งบางครั้งหากต้องการเน้นความหมายก็อาจใช้วิธีเขียนคำเดิมซ้ำหรือเพิ่มคำขยาย หรืออาจจะใช้วิธีการผันเสียงที่สูงหรือห้วนขึ้น สำหรับคำในภาษาไทยจะไม่มีการเปลี่ยนรูปคำ ตามกาลเวลา หรือความเป็นเอกพจน์และพหูพจน์ หรือการใช้อักษรตัวใหญ่ ตัวเล็ก เพื่อบ่งบอกคำเฉพาะ ซึ่งทำให้ไม่สามารถวิเคราะห์หน้าที่ของคำจากโครงสร้าง ไวยกรณ์ได้ อีกทั้งคำที่เขียนในภาษาไทยจะเขียนในลักษณะที่ติดต่อกันโดยไม่มีช่องว่างระหว่างคำ ซึ่งจำเป็นต้องมีขบวนการการตัด ในส่วนของการวิเคราะห์

โครงงานวิจัยนี้จึงมีแนวคิดในการจัดทำเหมืองข้อมูลสำหรับเอกสารงานวิจัยและผลงานทางวิชาการของสำนักงานกองทุนสนับสนุนการวิจัย ทั้งที่มีอยู่ในปัจจุบันและในอดีตที่มีมากกว่า 5000 เรื่องที่อาจอยู่ในไฟล์รูปแบบต่างๆหรืออาจจะยังอยู่ในรูปของกระดาษ โดยจะทำการวิเคราะห์เอกสารโดยการประยุกต์ใช้หลักการประมวลผลภาษาธรรมชาติสำหรับการวิเคราะห์โครงสร้างของข้อความภาษาไทยเพื่อคัดแยกโครงสร้าง คำสำคัญและจัดเอกสารออกเป็นหมวดหมู่ต่างๆ พร้อมทั้งจัดทำดัชนี สำหรับการสืบค้น พร้อมทั้งจัดทำเป็นโปรแกรมประยุกต์เพื่อให้เหมาะสมต่อการนำไปใช้งานและสำหรับผู้ใช้งานในระดับต่างๆ

Abstract

Natural Language Processing (NLP) for Thai language is difficult. This is because Thai language gives less important in the grammartical structure when forming the sentences and rather more concerned about the meaning. Words usually are formed from several root words which provides similar or total new meaning to the original words. Specific meanings are sometimes presented by repeating words or adding more words to extending the meanings, and sometimes by changing the tones. Thai grammar doesn't have word inflection. The word form remain when presenting plurals or tenses. Thai language also has no capital letters for specific words or as beginning of sentences. Words are written altogether without a word break, therefore, word segmentation is needed during the analysis processes.

The objective of this research is to apply natural language processings to analyse the research papers of Thailand Research Funds (TRF) which have been collected for more than 5000 topics. The papers are in different formats and mostly are still in paper formats. The analysis processes consists of the grammatical structure analysis, words and keywords extraction. The documents are then classified into groups with pre-defined indexes as purposes of documents retrieval. Finally, the software application can be developed for an ease of uses.