

การตัดคำภาษาไทยคือการแยกแต่ละคำในประโยคในเอกสารภาษาไทยซึ่งมีการเขียนคำติดกัน การตัดคำนั้นจำเป็นจะต้องมีเพื่อนำไปใช้ประโยชน์ทางด้านประมวลผลภาษาธรรมชาติ เช่น การสังเคราะห์เสียงพูด และการแปลภาษา เป็นต้น วิธีการตัดคำที่นิยมใช้กันมานานได้แก่ วิธีการตัดคำโดยใช้กฎ วิธีการตัดคำโดยใช้พจนานุกรม และวิธีการตัดคำโดยใช้คลังข้อความ แต่เนื่องจากเอกสารในภาษาไทยปัจจุบันมักจะมีคำจากภาษาต่างประเทศซึ่งมักจะมีคำอ่านออกเสียงอยู่ในรูปแบบของภาษาไทย จึงทำให้การตัดคำด้วยวิธีการปัจจุบันไม่สามารถตัดคำได้อย่างถูกต้องกับเอกสารในปัจจุบัน งานวิจัยนี้จึงทำการเสนอวิธีการตัดคำโดยการใช้กฎร่วมกับการใช้พจนานุกรม เพื่อเพิ่มความถูกต้องของการตัดคำ จากผลการทดลองพบว่าเมื่อเปรียบเทียบกับวิธีการตัดคำวิธีอื่น วิธีการตัดคำที่นำเสนอได้ช่วยเพิ่มความถูกต้องของการตัดคำถูกต้องในเอกสารภาษาไทยหลายประเภท

Thai word segmentation is the process of separating words in a sentence in Thai documents which words are adjacent to one another. Thai word segmentation is necessary for natural language processing such as speech synthesis and language translation. Traditional techniques in Thai word segmentation include rule-based segmentation, dictionary-based segmentation, and segmentation using corpus. However, currently Thai documents often consist of words from foreign languages which are spelt in the form of Thai language. Therefore, current Thai word segmentation technique cannot separate words effectively in Thai documents nowadays. This research work proposes Thai word segmentation by using rules with dictionary to enhance the correctness of Thai word segmentation. Based on experimental results, compared with other techniques, the proposed technique can increase the correctness of word segmentation in various types of Thai documents.