วิทยานิพนธ์ฉบับนี้ได้นำเสนอวิธีการตัดภาษาไทยซึ่งเป็นขั้นตอนพื้นฐานที่สำคัญสำหรับระบบงานทางด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing) โดยการตัดคำภาษาไทยประกอบด้วยปัญหาหลัก 2 ประการคือ ปัญหาความกำกวม และปัญหาคำที่ไม่ปรากฏในพจนานุกรม เนื่องมาจากลักษณะรูปแบบประโยคในภาษาไทยจะเขียนคำเรียงติดต่อกันไปโดยไม่มีการเว้นวรรคระหว่างคำ ทำให้ระบบไม่ทราบขอบเขตที่ชัดเจนของคำ จึงส่งผลกระทบโดยตรงต่อระบบงานที่เกี่ยวข้องกับการประมวลผลภาษาธรรมชาติ เช่น ระบบการสืบค้นข้อมูล ระบบการแปลภาษา และระบบการสังเคราะห์เสียง เป็นต้น ทำให้ประสิทธิภาพของระบบงานดังกล่าวไม่ดีเท่าที่ควร ทั้งนี้เพื่อให้การตัดคำภาษาไทยมีประสิทธิภาพสูงขึ้นงานวิจัยนี้จึงมุ่งเน้นศึกษาและพัฒนาวิธีการค้นหาคำที่ไม่ปรากฏในพจนานุกรม โดยได้นำเสนอการตัดคำภาษาไทยโดยใช้การเทียบสายอักษรโดยค้นหาคำจากพจนานุกรม ซึ่งได้ประยุกต์ใช้การสร้างกราฟโดยวิธีการหาเส้นทางที่สั้นที่สุดเพื่อหาเซตของกลุ่มคำที่ตัดได้จากข้อความนำเข้าและประยุกต์ใช้การระบุชนิดของคำตามหลักภาษาไทย เพื่อสร้างกฎสำหรับการหาขอบเขตของคำที่ทับกันและขอบเขตของคำที่ไม่ปรากฏในพจนานุกรม ในกรณีที่ไม่เป็นไปตามกฎจะใช้ตัวแบบฮิดเดนมาร์คอฟเข้ามาช่วยในการตัดสินใจในการแบ่งกลุ่มของคำที่ไม่ปรากฏในพจนานุกรม ซึ่งการค้นหาคำที่ไม่ปรากฏในพจนานุกรมเน้นเฉพาะที่เป็นแบบซ่อนเร้นบางส่วนและแบบชัดเจนเท่านั้น ในการประเมินประสิทธิภาพการตัดคำจากเอกสาร 90 ฉบับพบว่า ผลการวัดค่าประสิทธิภาพโดยเฉลี่ยของการตัดคำที่ไม่ปรากฏในพจนานุกรมของทุกกลุ่มเอกสารโดยใช้กฎเท่ากับ 75.01% และใช้กฎกับตัวแบบฮิดเดนมาร์คอฟเท่ากับ 81.74% โดยผลการตัดคำทั้งหมดในระดับพยางค์และระดับคำ 90.03% และ86.39% ตามลำดับ

216965

This thesis presents a method for Thai word segmentation which is the basic and importance step for Natural Language Processing (NLP). There are two main problems in word segmentation. The first is the ambiguity problem and the second is the unknown word boundary. In a Thai text, a delimiter for indicating the word boundary is not explicitly used. Therefore, this affects many tasks of Natural Language Processing such as Information retrieval, Thai-English machine translation, and Thai speech synthesis. In this research, we develop search for unknown word and present a method of Thai word segmentation using string matching and word identification in dictionary. A set of segmented word from input text and a graph was obtained to find the shortest path. In order to construct rules for matching unknown words, we classified words according to the Thai grammar and Hidden Markov Models (HMM). In this work, we search for partially hidden words and explicit unknown words in 90 documents. Finding unknown words in a dictionary emphasizes on partially hidden word and explicit unknown word. The results showed that the average segmentation efficiency for unknown word using rules was 75.01% and unknown word using rules and HMM were 81.74%, respectively. The results showed that average efficiency for syllable was 90.03% and word segmentation was 86.39%, respectively.