บทคัดย่อ

ชื่อวิทยานิพนธ์	:	การเพิ่มประสิทธิผลของวิธีการ Longest Matching ในการตัดคำภาษาไทย โดยใช้ Prediction by Partial Matching และ Logistic Regression
ชื่อผู้เขียน	:	น.ส. ปวีณา ชัยวนารมย์
ชื่อปริญญา	:	วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
ปีการศึกษา	:	2546

T158859

การตัดคำเป็นขั้นตอนสำคัญในงานหลายประเภท เช่น การค้นคืนเอกสารและการสรุปใจความสำคัญของเอกสาร การตัดคำเป็นปัญหาที่สำคัญสำหรับภาษาไทย เนื่องจากเป็นภาษาที่มีการเขียนคำในลักษณะที่ต่อเนื่องกันและโครงสร้างของคำไม่แน่นอน

Longest Matching เป็นวิธีการตัดคำโดยใช้พจนานุกรมที่มีประสิทธิผลค่อนข้างสูง อย่างไรก็ตาม การเลือกคำที่ยาวที่สุดเสมอทำให้เกิดข้อผิดพลาดในการตัดคำ โดยสาเหตุของความผิดพลาดนี้สามารถแบ่งได้เป็น 2 ประเภท คือ ความกำกวมในระดับตัวอักษรและความกำกวมในระดับพยางค์

งานวิจัยนี้มุ่งประเด็นที่การเพิ่มประสิทธิผลของวิธีการ Longest Matching โดยแบ่งการตัดคำออกเป็น 2 ขั้นตอน คือ การตัดพยางค์และการรวมชุดของพยางค์ให้เป็นคำ การตัดพยางค์ให้ถูกต้องสามารถลดความกำกวมในระดับตัวอักษรลงได้ เนื่องจากพยางค์มีโครงสร้างที่แน่นอนกว่าคำ ส่วนการนำชุดของพยางค์มารวมเป็นคำนั้นใช้หลักการของ Longest Matching ในระดับพยางค์ ร่วมกับตัวแบบ Logistic Regression ที่นำบริบทของพยางค์ที่กำกวมมาพิจารณาประกอบ

จากการทดลอง พบว่าวิธีการที่นำเสนอมีความถูกต้องในการตัดพยางค์มากกว่า 96% และความถูกต้องในการตัดคำทั้งกระบวนการกว่า 97%

ABSTRACT

Title of Thesis	:	Improving the Performance of the Longest Matching Approach for Thai Word Segmentation Using Prediction by Partial Matching and Logistic Regression
Author	:	Ms. Paweena Chaiwanarom
Degree	:	Master of Science (Computer Science)
Year	:	2003

Word segmentation is an important process in many applications, including information retrieval and document summarization. In Thai language, the process is complicated since words are written continuously and their structures are not well-defined.

A recognized effective approach to word segmentation is the longest matching, method based on dictionary. Nevertheless, this method suffers from two sources of ambiguities: one in the character level and the other in the syllable level.

This research aims to improve the performance of the longest matching method by proposing a two-step approach to word segmentation. First, text is segmented, by Prediction by Partial Matching, into syllables whose structures are more well-defined, reducing the earlier type of ambiguity. Then, syllables are combined into words by applying a syllable-level longest matching method, together with a logistic regression model, take into account contextual data.

The experimental results show the syllable segmentation accuracy of more than 96% and the overall word segmentation accuracy of 97%.