PARSING THAI TEXT WITH SYNTACTIC ANALYSIS USING DIGRAPH REPRESENTATION

ANUKUL CHIMPIPOP

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
1999

ISBN 974-662-522-5

3937015 SCCS/M : MAJOR : COMPUTER SCIENCE ; M.Sc.(COMPUTER SCIENCE)
KEY WORDS : PARSING / DICTIONARY / SYNTACTIC STRUCTURE / DIGRAPH / THAI

ANUKUL CHIMPIPOP : PARSING THAI TEXT WITH SYNTACTIC ANALYSIS USING DIGRAPH REPRESENTATION. THESIS ADVISOR : DAMRAS WONGSAWANG Ph.D. 73 p. ISBN 974-662-522-5

Parsing Thai text is the method to determine the word boundary in Thai language sentences. Many parsing methods have been proposed and implemented. However, most of the methods do not take the grammar of Thai language into consideration. Among them, the longest matching is one of the most effective methods implemented. This thesis proposed the algorithm, called Parsing Thai Text with Syntactic Analysis (PTTSA), applying the longest matching method enhanced by analyzing Thai grammar.

For analyzing Thai grammar we proposed a syntactic structure model that is structured from Thai language structure. Digraph is the way to represent our model. The probabilities of segmentation patterns are calculated when each segmentation pattern traverses in digraph. The highest probability is selected to be the best of the segmentations. We simulated the test environments and compared the parsing results between our approach with parsing without syntactic analysis. We found that our approach gives higher accuracy of parsing than parsing without syntactic analysis for most documents tested. We further found that the accuracy of parsing results depends on the syntactic structure model, probabilities of edges in digraph, and style of expression in documents. Adjusting the probabilities of edges in digraph, the accuracy of parsing may be better or worse than that previous one depending on type of documents. Thus, the accuracy of parsing results will be increased only when the probabilities of edges in digraph are appropriate to the style of expression in documents.

This thesis described PTTSA in detail including the formulation, analysis and implementation of the model. The prototype of PTTSA were developed and tested. The parsing results were presented and discussed. Finally, improvements of the model have been proposed.

3937015 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม. (วิทยาการคอมพิวเตอร์)
คำสำคัญ : การตัดคำ / พจนานุกรม / โครงสร้างทางไวยากรณ์ / ไดกราฟ / ไทย

อนุกูล ฉิมพิภพ : การตัดคำภาษาไทยโดยการวิเคราะห์โครงสร้างทางไวยากรณ์ของประโยคในรูปแบบของไดกราฟ (PARSING THAI TEXT WITH SYNTACTIC ANALYSIS USING DIGRAPH REPRESENTATIOIN). คณะกรรมการผู้ควบคุมวิทยานิพนธ์ : ดำรัส วงศ์สว่าง Ph.D. 73 หน้า ISBN 974-662-522-5

การตัดคำภาษาไทยคือวิธีการกำหนดขอบเขตของคำในประโยคของภาษาไทย วิธีการตัดคำภาษาไทยหลายวิธีได้ถูกเสนอและถูกนำมาใช้ อย่างไรก็ตามวิธีส่วนใหญ่ไม่ได้นำโครงสร้างทางไวยากรณ์ของภาษาไทยเข้ามาพิจารณา ในจำนวนนั้นวิธี longest matching เป็นวิธีที่มีประสิทธิภาพในการนำมาใช้ วิทยานิพนธ์นี้เสนอวิธีการที่เรียกว่า การตัดคำภาษาไทยโดยการวิเคราะห์โครงสร้างทางไวยากรณ์ของประโยค (Parsing Thai Text with Syntactic Analysis) หรือ PTTSA โดยนำเอาวิธี longest matching เข้ามาใช้และเพิ่มเติมด้วยการวิเคราะห์โครงสร้างทางไวยากรณ์ของภาษาไทย

สำหรับการวิเคราะห์โครงสร้างทางภาษาไทย เราเสนอรูปแบบโครงสร้างทางไวยากรณ์ซึ่งออกแบบจากโครงสร้างทางภาษาไทย ไดกราฟเป็นวิธีนำเสนอรูปแบบของเรา ความน่าจะเป็นของรูปแบบการตัดคำถูกคำนวณเมื่อแต่ละรูปแบบของการตัดคำเดินทางในไดกราฟ ความน่าจะเป็นที่สูงที่สุดจะถูกเลือกเป็นรูปแบบการตัดคำที่ดีที่สุด เราจำลองสภาพแวดล้อมการทดลองและเปรียบเทียบผลการตัดคำระหว่างวิธีการของเรากับการตัดคำโดยไม่ได้วิเคราะห์โครงสร้างทางไวยากรณ์ เราพบว่าวิธีการของเราให้ความถูกต้องสูงกว่าการตัดคำโดยไม่ได้วิเคราะห์โครงสร้างทางไวยากรณ์สำหรับเอกสารทดสอบส่วนใหญ่ เราพบมากไปกว่านั้นว่าความถูกต้องของผลการตัดคำขึ้นอยู่กับรูปแบบโครงสร้างทางไวยากรณ์, ความน่าจะเป็นของเอ็จในไดกราฟและลักษณะการใช้ภาษาในเอกสาร การปรับค่าความน่าจะเป็นของเอ็จในไดกราฟนั้นความถูกต้องของผลการตัดคำอาจดีขึ้นหรือแย่ลงกว่าเก่าขึ้นอยู่กับลักษณะของเอกสาร ดังนั้นความถูกต้องของผลการตัดคำจะเพิ่มขึ้นเมื่อความน่าจะเป็นของเอ็จในไดกราฟเหมาะสมกับรูปแบบการใช้ภาษาในเอกสารนั้น ๆ

วิทยานิพนธ์นี้อธิบายวิธี PTTSA ในรายละเอียดรวมถึงหลักเกณฑ์ การวิเคราะห์และการนำไปใช้ของวิธี แบบจำลองของวิธี PTTSA ถูกพัฒนาและทดสอบ ผลการตัดคำถูกนำมาแสดงและพิจารณา สุดท้ายเสนอการแก้ไขให้ดีขึ้น