การตัดคำภาษาไทยโดยใช้คุณลักษณะ

นายไพศาล เจริญพรสวัสดิ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2541

ISBN 974-332-382-1

FEATURE-BASED THAI WORD SEGMENTATION

Mr. Paisarn Charoenpornsawat

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Computer Engineering
Department of Computer Engineering
Graduate School
Chulalongkorn University
Academic Year 1998

ISBN 974-332-382-1

หัวข้อวิทยานิพนธ์	การตัดคำภาษาไทยโดยใช้คุณลักษณะ
โดย	นายไพศาล เจริญพรสวัสดิ์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	อาจารย์ ดร. บุญเสริม กิจศิริกุล
อาจารย์ที่ปรึกษาร่วม	อาจารย์ ดร. สุรพันธ์ เมฆนาวิน

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้วิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ นายแพทย์ศุภวัฒน์ ชุติวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จูตระกูล)

.....อาจารย์ที่ปรึกษา
(อาจารย์ ดร. บุญเสริม กิจศิริกุล)

.....อาจารย์ที่ปรึกษาร่วม
(อาจารย์ ดร. สุรพันธ์ เมฆนาวิน)

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

ไพศาล เจริญพรสวัสดิ์ : การตัดคำภาษาไทยโดยใช้คุณลักษณะ (Feature-based Thai Word Segmentation) อาจารย์ที่ปรึกษา : อ. ดร. บุญเสริม กิจศิริกุล, อ. ที่ปรึกษาร่วม : อ. ดร. สุรพันธ์ เมฆนาวิน ; 70 หน้า. ISBN 974-332-382-1

เนื่องจากลักษณะการเขียนของภาษาไทยนั้นไม่มีการใช้ตัวอักษรหรือสัญลักษณ์ที่นำมาใช้คั่นระหว่างคำ และงานต่างๆ ในด้านการประมวลผลภาษาธรรมชาตินั้นจำเป็นต้องทราบขอบเขตของคำก่อนถึงจะสามารถนำไปประมวลผลต่อไปได้ ดังเช่นการแปลภาษาไทย-อังกฤษ การสังเคราะห์เสียงภาษาไทย หรือการแก้ไขคำที่สะกดผิด เป็นต้น ทำให้การตัดคำนั้นถือได้ว่าเป็นปัญหาที่สำคัญปัญหาหนึ่งสำหรับงานด้านการประมวลผลภาษาธรรมชาติ

ในการตัดคำนั้นประกอบไปด้วยปัญหาหลัก 2 ปัญหาคือ 1. ปัญหาความกำกวม 2. ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม สำหรับแนวคิดในการตัดคำนั้นมีอยู่หลายแนวคิด เช่นการตัดคำแบบเลือกคำยาวที่สุด การตัดคำโดยเลือกแบบเหมือนมากที่สุด และการตัดคำโดยโมเดลไตรแกรม อย่างไรก็ตามแนวคิดต่างๆ เหล่านั้นไม่สามารถให้ความถูกต้องที่สูงในการแก้ปัญหาการตัดคำ เพราะว่ามีการใช้เพียงวิทยาการศึกษาสำนึก สำหรับการตัดคำโดยแบบเลือกคำยาวที่สุดและการตัดคำโดยเลือกแบบที่เหมือนมากที่สุด และสำหรับการตัดคำโดยใช้โมเดลไตรแกรมนั้นมีการพิจารณาแค่คำบริบทก่อนหน้าแค่เพียง 2 คำเท่านั้น ส่วนความถูกต้องในการแก้ปัญหาความกำกวมนั้นมีความถูกต้องประมาณ 53% และ 73% สำหรับการตัดคำโดยเลือกแบบเหมือนมากที่สุดและการตัดคำโดยใช้โมเดลไตรแกรมตามลำดับ

ในวิทยานิพนธ์นี้เสนอแนวคิดการนำคุณลักษณะโดยใช้การเรียนรู้ของเครื่อง 2 แบบคือริปเปอร์และวินโนว์ในการแก้ปัญหาการตัดคำภาษาไทย โดยคุณลักษณะคือข้อมูลที่อยู่รอบๆ ซึ่งสามารถนำมาประยุกต์ใช้ในการแก้ปัญหาได้ สำหรับคุณลักษณะที่นำมาใช้ในการแก้ปัญหาการตัดคำทั้ง 2 ปัญหา คือคำบริบท และสิ่งที่เกิดร่วมกันโดยมีลำดับ ในการทดลองมีการนำคลังข้อความที่มีการกำหนดหน้าที่คำจำนวน 80% เข้ามาใช้ในการเรียนรู้และส่วนที่เหลือนำมาใช้ในการทดสอบ สำหรับการวัดประสิทธิภาพนั้นได้มีการแบ่งออกเป็น 2 ส่วนคือ 1. วัดค่าความถูกต้องของการแก้ปัญหาความกำกวม 2. วัดค่าความถูกต้องของการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม สำหรับความถูกต้องโดยการใช้ริปเปอร์และวินโนว์ในการแก้ปัญหาความกำกวมนั้นให้ความถูกต้องมากกว่า 85% และ 90% ตามลำดับ ส่วนความถูกต้องในการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้นให้ความถูกต้องมากกว่า 70% และ 80% สำหรับริปเปอร์และวินโนว์ตามลำดับ

จากผลการทดลองแสดงให้เห็นว่าการตัดคำโดยใช้คุณลักษณะให้ประสิทธิภาพในการแก้ปัญหาได้ดีกว่าการตัดคำโดยใช้ไตรแกรมโมเดลและการตัดคำโดยเลือกแบบเหมือนมากที่สุด และยังแสดงให้เห็นว่าวินโนว์สามารถดึงคุณลักษณะต่างๆ จากคลังข้อความ เพื่อใช้ในการแก้ปัญหาการตัดคำได้ดีกว่าริปเปอร์

ภาควิชา วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต ไพศาล เจริญพรสวัสดิ์
สาขาวิชา วิศวกรรมคอมพิวเตอร์	ลายมือชื่ออาจารย์ที่ปรึกษา
ปีการศึกษา 2541	ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

4070366321 COMPUTER ENGINEERING
THAI / WORD / SEGMENTATION / FEATURE / CONTEXT / COLLOCATION / WINNOW / RIPPER
PAISARN CHAROENPORNSAWAT: FEATURE-BASED THAI WORD SEGMENTATION. THESIS ADVISOR: BOONSERM KIJSIRIKUL, Ph.D. THESIS COADVISOR: SURAPANT MEKNAVIN, Ph.D. 70 pp. ISBN 974-332-382-1.

In a Thai text, a delimiter for indicating the word boundary is not explicitly used. Many tasks of Natural Language Processing (NLP) such as Thai-English machine translation, Thai speech synthesis and spelling correction require boundaries of words. Therefore, word segmentation is one of the main problems in NLP.

There are two main problems in word segmentation. The first is the ambiguity problem and the second is the unknown word boundary problem. Many approaches such as longest matching, maximal matching and trigram model have been proposed. However, these approaches can not give high accuracy because longest matching and maximal matching use only heuristics and trigram model consider only two previous context words for solving the problems. The accuracy in solving ambiguity problem is about 53% and 73% for maximal matching and trigram model respectively.

This thesis proposes to use a feature-based approach with two learning algorithms namely RIPPER and Winnow in solving the problems in Thai word segmentation. A feature can be anything that tests for specific information in the context around the word in question, such as context words and collocations. In the experiment, we train the system by using RIPPER and Winnow algorithm separately, on an 80% of part-of-speech tagged corpus and the rest is used for testing. We divided the evaluation into two parts. One is the accuracy in solving the ambiguity problem and the other is the accuracy in solving the unknown word boundary problem. The accuracy using RIPPER and Winnow in solving the ambiguity problem is more than 85% and 90% respectively. On the other hand, the accuracy in solving the unknown word boundary problem is more than 70% and 80% for RIPPER and Winnow respectively.

The experiment results show the feature-based approach outperform trigram model and maximal matching, and Winnow is superior to RIPPER for extracting the features from the corpus.

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....
สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา .....
ปีการศึกษา 2541 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีด้วยคำแนะนำอย่างดียิ่งของ อ. ดร. บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ และ อ. ดร. สุรพันธ์ เมฆนาวิน อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ผู้วิจัยขอขอบคุณ คุณเทพพิทักษ์ การุณบุญญานันท์ ผู้ช่วยนักวิจัย ห้องปฏิบัติการวิจัยและพัฒนาวิศวกรรมภาษาและซอฟต์แวร์ ศูนย์เทคโนโลยีอิเล็คทรอนิคส์และคอมพิวเตอร์แห่งชาติ ที่ใช้รหัสต้นฉบับ (Source code) โครงสร้างการจัดเก็บข้อมูลแบบทรัย ขอขอบคุณ คุณธนพงษ์ โพธิปิติ สำหรับการเขียนโปรแกรมในการทดลอง ขอขอบคุณ คุณวิรงรอง เทศประสิทธิ์ ที่ช่วยตรวจทานตัวสะกดในวิทยานิพนธ์ฉบับนี้ และขอขอบคุณ อ. ดร. วิรัช ศรเลิศล้ำ-วานิช อ. วันทนีย์ พันธชาติ และสมาชิกห้องปฏิบัติการวิจัยและพัฒนาวิศวกรรมภาษาและซอฟต์แวร์ทุกท่านที่คอยให้คำปรึกษา คำแนะนำ ความอนุเคราะห์ในการใช้คลังข้อความออร์คิด รายการคำศัพท์ภาษาไทย และอุปกรณ์ต่างๆ

ท้ายนี้ผู้วิจัยใคร่ขอกราบขอบพระคุณบิดา-มารดา ซึ่งให้การสนับสนุนด้านการเงินและคอยให้กำลังใจแก่ผู้วิจัยเสมอมา

สารบัญ

หน้า