

การตัดคำภาษาไทยโดยใช้คุณลักษณะ



นายไพศาล เจริญพรสวัสดิ์



สถาบันวิทยบริการ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

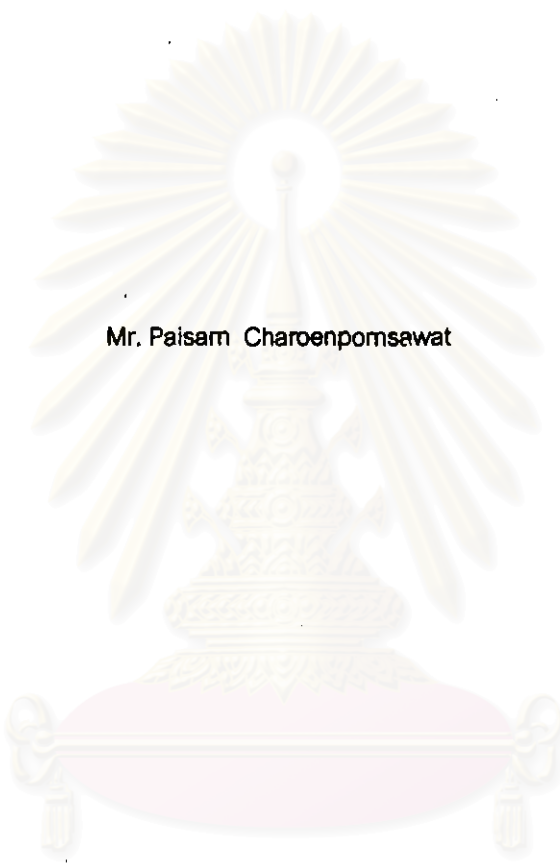
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2541

ISBN 974-332-382-1

ลิขสิทธิ์ของ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

FEATURE-BASED THAI WORD SEGMENTATION



Mr. Paisarn Charoenpomsawat

สถาบันวิทยบริการ  
จุฬาลงกรณ์มหาวิทยาลัย

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

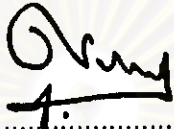
Chulalongkorn University

Academic Year 1998

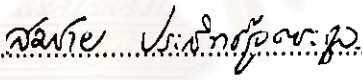
ISBN 974-332-382-1

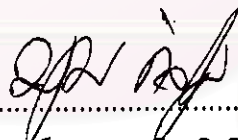
หัวข้อวิทยานิพนธ์      การตัดคำภาษาไทยโดยใช้คุณลักษณะ  
โดย                              นายไพศาล เจริญพรสวัสดิ์  
ภาควิชา                            วิศวกรรมคอมพิวเตอร์  
อาจารย์ที่ปรึกษา              อาจารย์ ดร. บุญเสริม กิจศิริกุล  
อาจารย์ที่ปรึกษาร่วม        อาจารย์ ดร. สุรพันธ์ เมฆนาวิน

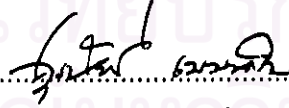
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้วิทยานิพนธ์ฉบับนี้เป็นส่วน  
หนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

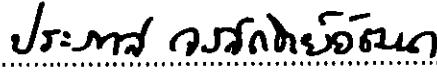
  
.....คณบดีบัณฑิตวิทยาลัย  
(ศาสตราจารย์ นายแพทย์ศุภวัฒน์ ชูติวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

  
.....ประธานกรรมการ  
(ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จตุระกุล)

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

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

  
.....กรรมการ  
(ผู้ช่วยศาสตราจารย์ ดร. ประภาส จงสิตติชัยวัฒนา)

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

เนื่องจากลักษณะการเขียนของภาษาไทยนั้นไม่มีการใช้ตัวอักษรหรือสัญลักษณ์ที่นำมาใช้คั่นระหว่างคำ และงานต่างๆ ในด้านการประมวลผลภาษารวมชาตินั้นจำเป็นต้องทราบขอบเขตของคำก่อนถึงจะสามารถนำไปประมวลผลต่อไปได้ ดังเช่นการแปลภาษาไทย-อังกฤษ การสังเคราะห์เสียงภาษาไทย หรือการแก้ไขคำที่สะกดผิด เป็นต้น ทำให้การตัดคำนั้นถือได้ว่าเป็นปัญหาที่สำคัญปัญหาหนึ่งสำหรับงานด้านการประมวลผลภาษารวมชาติ

ในการตัดคำนั้นประกอบไปด้วยปัญหาหลัก 2 ปัญหาคือ 1. ปัญหาความกำกวม 2. ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม สำหรับแนวคิดในการตัดคำนั้นมีอยู่หลายแนวคิด เช่นการตัดคำแบบเลือกคำยาวที่สุด การตัดคำโดยเลือกแบบเหมือนมากที่สุด และการตัดคำโดยโมเดลโครงข่าย อย่างไรก็ตามแนวคิดต่างๆ เหล่านี้ไม่สามารถให้ความถูกต้องที่สูงในการแก้ปัญหาคัดคำ เพราะว่ามีการใช้เพียงวิทยาการศึกษาคำศัพท์ สำหรับการตัดคำโดยแบบเลือกคำยาวที่สุด และการตัดคำโดยเลือกแบบที่เหมือนมากที่สุด และสำหรับการตัดคำโดยใช้โมเดลโครงข่ายนั้นมีการพิจารณาแค่คำบริบทก่อนหน้าแค่เพียง 2 คำเท่านั้น ส่วนความถูกต้องในการแก้ปัญหาคำกำกวมนั้นมีความถูกต้องประมาณ 53% และ 73% สำหรับการตัดคำโดยเลือกแบบเหมือนมากที่สุดและการตัดคำโดยใช้โมเดลโครงข่ายตามลำดับ

ในวิทยานิพนธ์นี้เสนอแนวคิดการนำคุณลักษณะโดยใช้การเรียนรู้ของเครื่อง 2 แบบคือริปเปอร์และวินโนวีในการแก้ปัญหาคัดคำภาษาไทย โดยคุณลักษณะคือข้อมูลที่ถูกรอบๆ ซึ่งสามารถนำมาประยุกต์ใช้ในการแก้ปัญหาคัดคำสำหรับคุณลักษณะที่นำมาใช้ในการแก้ปัญหาคัดคำทั้ง 2 ปัญหา คือคำบริบท และสิ่งที่เกิดร่วมกันโดยมีลำดับ ในการทดลองมีการนำคลังข้อความที่มีการกำหนดหน้าที่คำจำนวน 80% เข้ามาใช้ในการเรียนรู้และส่วนที่เหลือนำมาใช้ในการทดสอบ สำหรับการวัดประสิทธิภาพนั้นได้มีการแบ่งออกเป็น 2 ส่วนคือ 1. วัดค่าความถูกต้องของการแก้ปัญหาคำกำกวม 2. วัดค่าความถูกต้องของการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม สำหรับความถูกต้องโดยการใช้ริปเปอร์และวินโนวีในการแก้ปัญหาคำกำกวมนั้นให้ความถูกต้องมากกว่า 85% และ 90% ตามลำดับ ส่วนความถูกต้องในการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรมนั้นให้ความถูกต้องมากกว่า 70% และ 80% สำหรับริปเปอร์และวินโนวีตามลำดับ

จากผลการทดลองแสดงให้เห็นว่าการตัดคำโดยใช้คุณลักษณะให้ประสิทธิภาพในการแก้ปัญหาคัดคำดีกว่าการตัดคำโดยใช้โครงข่ายโมเดลและการตัดคำโดยเลือกแบบเหมือนมากที่สุด และยังแสดงให้เห็นว่าวินโนวีสามารถดึงคุณลักษณะต่างๆ จากคลังข้อความ เพื่อใช้ในการแก้ปัญหาคัดคำได้ดีกว่าริปเปอร์

ภาควิชา .....วิศวกรรมคอมพิวเตอร์.....  
สาขาวิชา .....วิศวกรรมคอมพิวเตอร์.....  
ปีการศึกษา ..... 2541.....

ลายมือชื่อนิติกร ..... *Travis* .....  
ลายมือชื่ออาจารย์ที่ปรึกษา ..... *Dr. Kiat* .....  
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม ..... *Dr. Suran* .....

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.

สถาบันวิทยบริการ  
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา ..... วิศวกรรมคอมพิวเตอร์  
สาขาวิชา ..... วิศวกรรมคอมพิวเตอร์  
ปีการศึกษา ..... 2541

ลายมือชื่อนิติกร ..... Boonserm Kijirikul  
ลายมือชื่ออาจารย์ที่ปรึกษา ..... Paisarn Charoenpornsawat  
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม ..... Surapant Meknavin

## กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีด้วยคำแนะนำอย่างดียิ่งของ อ. ดร. บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ และ อ. ดร. สุรพันธ์ เมฆนาวัน อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ผู้วิจัย ขอขอบคุณ คุณเทพพิทักษ์ การุณบุญญานันท์ ผู้ช่วยนักวิจัย ห้องปฏิบัติการวิจัยและพัฒนาวิศวกรรมภาษาและซอฟต์แวร์ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ที่ใช้รหัสต้นฉบับ (Source code) โครงสร้าง การจัดเก็บข้อมูลแบบทรีย์ ขอขอบคุณ คุณธนพงษ์ โพธิ์ปิติ สำหรับการเขียนโปรแกรมในการทดลอง ขอขอบคุณ คุณวิรงรอง เทศประสิทธิ์ ที่ช่วยตรวจทานตัวสะกดในวิทยานิพนธ์ฉบับนี้ และขอขอบคุณ อ. ดร. วิรัช ศรีเลิศล้ำวานิช อ. วันทนีย์ พันธชาติ และสมาชิกห้องปฏิบัติการวิจัยและพัฒนาวิศวกรรมภาษาและซอฟต์แวร์ทุกท่านที่ คอยให้คำปรึกษา คำแนะนำ ความอนุเคราะห์ในการใช้คลังข้อความออร์คิด รายการคำศัพท์ภาษาไทย และ อุปกรณ์ต่างๆ

ท้ายนี้ผู้วิจัยใคร่ขอกราบขอบพระคุณบิดา-มารดา ซึ่งให้การสนับสนุนด้านการเงินและคอยให้กำลังใจแก่ผู้วิจัยเสมอมา



สถาบันวิทยบริการ  
จุฬาลงกรณ์มหาวิทยาลัย

## สารบัญ

หน้า

บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ.....	ช

### บทที่

1. บทนำ .....	1
1.1 ความเป็นมา.....	1
1.2 ปัญหาการตัดคำ.....	2
1.3 วัตถุประสงค์ของวิทยานิพนธ์ .....	2
1.4 ขอบเขตของวิทยานิพนธ์.....	3
1.5 ขั้นตอนการวิจัย .....	3
1.6 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย .....	4
1.7 สิ่งตีพิมพ์ที่ได้จากงานวิทยานิพนธ์.....	4
2. งานวิจัยและทฤษฎีที่เกี่ยวข้อง.....	5
2.1 ยุคการใช้กฎ .....	5
2.2 ยุคการใช้พจนานุกรม .....	8
2.3 ยุคการใช้คลังข้อความ.....	11
3. การกำกับหน้าที่คำ.....	17
3.1 ลักษณะปัญหาของการกำกับหน้าที่คำ .....	18
3.2 วิธีการแก้ปัญหา.....	18
3.3 การเพิ่มประสิทธิภาพ .....	20
4. โครงสร้างของพจนานุกรม.....	22
4.1 โครงสร้างข้อมูลแบบทรี.....	22
4.2 ประสิทธิภาพด้านความเร็ว .....	24
4.3 ประสิทธิภาพในการใช้หน่วยความจำ .....	25
5. ปัญหาความกำกวมและคำศัพท์ที่ไม่ปรากฏในพจนานุกรม .....	28
5.1 ความกำกวม.....	26
5.2 คำศัพท์ที่ไม่ปรากฏในพจนานุกรม.....	27
6. การเรียนรู้ของเครื่อง .....	30



6.1 ริปเปอร์ (RIPPER: REPEATED INCREMENTAL PRUNING TO PRODUCE ERROR REDUCTION) .....	30
6.2 วินโรว์ (WINNOW).....	32
<b>7. การตัดคำภาษาไทยโดยใช้คุณลักษณะ .....</b>	<b>34</b>
7.1 คุณลักษณะ.....	34
7.2 การแก้ไขปัญหาคำกำกวม.....	35
7.3 การแก้ไขปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม.....	39
<b>8. ประสิทธิภาพการตัดคำโดยใช้คุณลักษณะ.....</b>	<b>45</b>
8.1 ขั้นตอนการนำคุณลักษณะเข้ามาใช้ในการแก้ปัญหาคำกำกวม.....	47
8.2 ผลการทดลองแก้ปัญหาคำกำกวม.....	47
8.3 สรุปผลการทดลองการแก้ปัญหาคำกำกวม.....	48
8.4 ขั้นตอนการนำคุณลักษณะเข้ามาใช้ในการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม.....	53
8.5 ผลการทดลองการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม .....	54
8.6 สรุปผลการทดลองการแก้ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม.....	54
<b>9. บทสรุปและแนวทางการพัฒนาต่อ.....</b>	<b>56</b>
9.1 ประสิทธิภาพการนำคุณลักษณะมาใช้ในการแก้ปัญหาคำตัดคำ.....	56
9.2 ข้อเสนอแนะ .....	57
<b>รายการอ้างอิง .....</b>	<b>58</b>
ภาษาไทย .....	58
ภาษาอังกฤษ.....	59
<b>ภาคผนวก.....</b>	<b>82</b>
ภาคผนวก ก .....	63
ภาคผนวก ข .....	65
ภาคผนวก ค.....	68
<b>ประวัติผู้เขียน.....</b>	<b>71</b>