

การตัดคำก่ากวมในข้อความภาษาไทยด้วยการโปรแกรมตรรกะเชิงอุปนัย

นางสาว ชมภูษุช คุปติวุฒิ



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2542

ISBN 974-332-872-6

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

SEGMENTATION OF AMBIGUOUS THAI WORDS BY
INDUCTIVE LOGIC PROGRAMMING

Miss Chompunuch Kooptiwot

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science

Department of Computer Engineering

Graduate School

Chulalongkorn University

Academic Year 1999

ISBN 974-332-872-6

หัวข้อวิทยานิพนธ์ การตัดคำกำกับมโนในข้อความภาษาไทยด้วยการโปรแกรมตรรกะ
เชิงอุปนัย
โดย นางสาว ชมภุชช คุปติวุฒิ
ภาควิชา วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา อาจารย์ ดร. บุญเสริม กิจศิริกุล

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้บัณฑิตวิทยาลัยฉบับนี้เป็นส่วนหนึ่ง
ของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีบัณฑิตวิทยาลัย
(รองศาสตราจารย์ ดร. สุชาดา กิระนันท์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร. วันชัย รั้วไพบูลย์)

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

..... กรรมการ
(อาจารย์ ดร. ยรรยง เต็งอำนวย)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จตุระกุล)

ชมนุช คุปติวุฒิ : การตัดคำกำกวมในข้อความภาษาไทยด้วยการโปรแกรมตรรกะเชิงอุปนัย
(SEGMENTATION OF AMBIGUOUS THAI WORDS BY INDUCTIVE LOGIC
PROGRAMMING)

อ. ที่ปรึกษา : อ. ดร. บุญเสริม กิจศิริกุล; 71 หน้า. ISBN 974-332-872-6.

งานวิทยานิพนธ์ฉบับนี้มีวัตถุประสงค์เพื่อประยุกต์ใช้การโปรแกรมตรรกะเชิงอุปนัยหรือไอแอลพีในการตัดคำกำกวมในข้อความภาษาไทย ระบบไอแอลพีที่เลือกใช้คือระบบ FOIL โดยจะนำไปเปรียบเทียบกับระบบการเรียนรู้แบบประพจน์ โดยระบบการเรียนรู้แบบประพจน์ที่เลือกใช้คือระบบ RIPPER ขั้นตอนการวิจัยเริ่มจากการใช้ FOIL และ RIPPER เรียนรู้คุณลักษณะของคำกำกวม ผลที่ได้จากการเรียนรู้ของระบบ FOIL คือกลุ่มของอนุประโยคฮอร์นอันดับที่หนึ่ง ส่วนผลที่ได้จากการเรียนรู้ของระบบ RIPPER คือกลุ่มของกฎประพจน์ ซึ่งแต่ละอนุประโยคหรือกฎจะนิยามคุณลักษณะของคำกำกวมแต่ละคำ จากการทดลองพบว่ากลุ่มของอนุประโยคที่ได้จากการเรียนรู้ด้วยระบบ FOIL สามารถนิยามคุณลักษณะของคำกำกวมได้ถูกต้องมากกว่ากลุ่มของกฎที่ได้จากการเรียนรู้ด้วยระบบ RIPPER ขั้นตอนถัดมาคือการนำกลุ่มของอนุประโยคที่ได้จากการเรียนรู้ด้วยระบบ FOIL มาช่วยในการตัดคำกำกวมในข้อความภาษาไทย จากการทดลองได้ว่าการตัดคำกำกวมในข้อความภาษาไทยที่มีการใช้กลุ่มของอนุประโยคที่ได้จากการเรียนรู้ด้วยระบบ FOIL ให้ความถูกต้องมากกว่าวิธีการตัดคำแบบจำลองไตรแกรม

ภาควิชา.....วิศวกรรมคอมพิวเตอร์.....
สาขาวิชา.....วิทยาศาสตร์คอมพิวเตอร์.....
ปีการศึกษา2542.....

ลายมือชื่อนิติ.....ชมนุช คุปติวุฒิ.....
ลายมือชื่ออาจารย์ที่ปรึกษา.....
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

3970382921: MAJOR COMPUTER SCIENCE

KEY WORD:

INDUCTIVE LOGIC PROGRAMMING / PROPOSITIONAL LEARNING SYSTEM / FIRST-
ORDERHORN CLAUSE / TRIGRAM MODEL

CHOMPUNUCH KOOPTIWOOT : SEGMENTATION OF AMBIGUOUS THAI WORDS BY

INDUCTIVE LOGIC PROGRAMMING. THESIS ADVISOR : BOONSERM KIJSIRIKUL, Ph.D. 71 pp.

ISBN 974-332-872-6.

The purpose of this thesis is to apply Inductive Logic Programming (ILP) to the segmentation of ambiguous Thai words. The ILP system which has been chosen is FOIL. Another learning system, which is used to be compared with FOIL, is a propositional learning system, RIPPER. First, FOIL and RIPPER are used to learn features of ambiguous Thai words. The outputs of FOIL and RIPPER are a set of first-order Horn clauses and a set of propositional rules respectively, each of which defines the features of the ambiguous Thai words. Experimental results show that the clauses learned by FOIL are more accurate than rules of RIPPER. Then, these clauses from FOIL are used to help for segmentation of ambiguous Thai words. Experimental results show that the usage of the clauses improves the accuracy of word segmentation which uses trigram model.

ภาควิชา.....วิศวกรรมคอมพิวเตอร์
สาขาวิชา.....วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา.....2542

ลายมือชื่อนิสิต.....
ลายมือชื่ออาจารย์ที่ปรึกษา.....
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของอาจารย์ดร.บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำ ข้อคิดเห็นต่าง ๆ ในการวิจัยด้วยดีตลอดมา และตรวจแก้วิทยานิพนธ์ฉบับนี้อย่างละเอียด

ขอขอบคุณ คุณไพศาล เจริญพรสวัสดิ์ ที่ได้ให้ข้อมูลและโปรแกรมการตัดคำแบบจำลอง ไตรแกรมที่นำมาใช้ในงานวิจัย และขอขอบคุณพี่ๆ และเพื่อนๆ ที่ได้ให้คำปรึกษา กำลังใจและความช่วยเหลือในด้านต่างๆ ซึ่งทำให้การทำงานวิจัยเป็นไปอย่างราบรื่น

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา มารดา ซึ่งสนับสนุนในด้านการเงินและให้กำลังใจแก่ผู้วิจัยเสมอจนสำเร็จการศึกษา



สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ	ฉ
สารบัญ	ช
สารบัญตาราง	ฅ
สารบัญภาพ	ญ
บทที่	
1 บทนำ	1
ความเป็นมาและความสำคัญของปัญหา	1
วัตถุประสงค์	2
ขอบเขตของการวิจัย	2
ขั้นตอนการวิจัย	2
ประโยชน์ที่ได้จากการวิจัย	2
2 งานวิจัยที่เกี่ยวข้อง	3
วิธีการใช้กฎ	3
วิธีการใช้พจนานุกรม	5
วิธีการใช้คลังข้อความ	10
3 ทฤษฎีที่ใช้ในการวิจัย.....	16
การเรียนรู้ของเครื่อง	16
การโปรแกรมตรรกะเชิงอุปนัย	16
งานวิจัยด้านไอแอลพี	17
ระบบ FOIL	18
1. ความหมายของระบบ FOIL	18
2. การใช้งานระบบ FOIL	18
3. ขั้นตอนการทำงานของระบบ FOIL	20
ระบบ RIPPER	24

บทที่

1. ความหมายของระบบ RIPPER	24
2. การใช้งานระบบ RIPPER	25
3. ขั้นตอนการทำงานของระบบ RIPPER	26
4 กรรรมวิธีการเรียนรู้คำกำกวม	28
ประเภทของข้อความกำกวม	28
การเรียนรู้คำกำกวมโดยใช้ระบบ FOIL และ RIPPER	28
1. ข้อมูลที่ใช้ในการเรียนรู้	28
2. ข้อมูลตัวอย่าง	31
3. ความรู้ภูมิหลัง	34
ผลการเรียนรู้	37
การพัฒนาโมดูลย่อยจากการเรียนรู้โดยใช้ระบบ FOIL	38
5 สรุปการวิจัยและข้อเสนอแนะ	44
สรุปการวิจัย	44
ข้อเสนอแนะ	45
รายการอ้างอิง	46
ภาคผนวก	48
ภาคผนวก ก	49
ภาคผนวก ข	53
ภาคผนวก ค	55
ภาคผนวก ง	65
ประวัติผู้วิจัย	71

สารบัญตาราง

	หน้า
ตารางที่ 2.1 ตารางแสดงการตัดค่าแบบเลือกค่าที่ยาวที่สุด	7
ตารางที่ 2.2 ตารางแสดงการตัดค่าที่มีการใช้พจนานุกรมที่มีโครงสร้างข้อมูลแบบสองแถวลำดับ ..	8
ตารางที่ 4.1 ประเภทของคำ	29
ตารางที่ 4.2 ผลที่ได้จากการเรียนรู้	38
ตารางที่ 4.3 ผลการทดลอง	40

สารบัญภาพ

	หน้า
รูปที่3.1 รูปแบบข่า่งานที่แสดงทิศทาง	22
รูปที่4.1 ตัวอย่างข้อมูลในการเรียนรู้คำว่า มา กว่า	32
รูปที่4.2 ข้อมูลในแฟ้มข้อมูลที่จะนำไปเรียนรู้ด้วยระบบ FOIL	32
รูปที่4.3 ข้อมูลในแฟ้มข้อมูลที่จะนำไปเรียนรู้ด้วยระบบ RIPPER	33
รูปที่4.4 ข้อมูลในแฟ้มชื่อในการเรียนรู้คำว่า มา กว่า โดยใช้ระบบ RIPPER	33
รูปที่4.5 ตัวอย่างความรู้ภูมิหลังในแฟ้มข้อมูลในการเรียนรู้คำว่า มา กว่า ด้วยระบบ FOIL	34
รูปที่4.6 ขั้นตอนวิธีวิเทอร์บี (Viterbi Algorithm)	39
รูปที่4.7 ขั้นตอนการตัดคำโดยใช้อนุประโยคผลลัพธ์จาก FOIL	40
รูปที่4.8 ข้อมูลในแฟ้มข้อมูลที่จะนำไปผ่านขั้นตอนการตัดคำ	41
รูปที่4.9 ผลลัพธ์ที่ได้จากแฟ้มข้อมูลด้วยวิธีการตัดคำแบบจำลองไตรแกรม	42
รูปที่4.10 ผลลัพธ์ที่ได้จากแฟ้มข้อมูลด้วยวิธีการตัดคำที่ใช้อนุประโยคที่ได้จาก FOIL	42