

การรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทย



นางสาว สุฤดี นัครไตรมงคล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต

สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2548

ISBN 974-53-2979-7

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

NAMED ENTITY RECOGNITION AND CLASSIFICATION IN THAI

Miss Surudee Chattrimongkol

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Arts Program in Linguistics

Department of Linguistics

Faculty of Arts

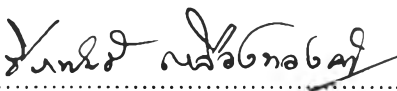
Chulalongkorn University

Academic Year 2005


ISBN 974-53-2979-7


หัวข้อวิทยานิพนธ์ การรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทย
โดย นางสาว สุฤดี ฉัตรไตรมงคล
สาขาวิชา ภาษาศาสตร์
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล


คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้หัวข้อวิทยานิพนธ์ฉบับนี้เป็นส่วน
หนึ่งของการศึกษาตามหลักสูตรปริญญาโทบัณฑิต


..... คณบดีคณะอักษรศาสตร์
(ศาสตราจารย์ ดร.ธีระพันธ์ เหลืองทองคำ)

คณะกรรมการสอบวิทยานิพนธ์


..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุดาพร ลักษณะียนาวิน)


..... อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ ดร.วิโรจน์ อรุณมานะกุล)


..... กรรมการ
(รองศาสตราจารย์ ดร.กิงกาญจน์ เทพกาญจนา)

สุฤดี ฉัตรไตรมงคล : การรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทย. (NAMED ENTITY RECOGNITION AND CLASSIFICATION IN THAI) อาจารย์ที่ปรึกษา : ผศ. ดร. วิโรจน์ อรุณมานะกุล, 75 หน้า. ISBN 974-53-2979-7.

งานวิจัยนี้มีจุดประสงค์เพื่อพัฒนาระบบการรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทยโดยใช้แนวทางแบบลูกผสม (hybrid approach) โดยแนวทางดังกล่าวจะแบ่งออกเป็นสองส่วนคือส่วนที่เป็นระบบทางสถิติและส่วนที่เป็นระบบกฎ

สำหรับส่วนของระบบทางสถิตินั้นจะใช้วิธีทางสถิติร่วมกับโลคอลแมกซ์อัลกอริทึมเพื่อคัดเลือกกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะออกมา ซึ่งวิธีการทางสถิติที่ใช้ในการวัดความสัมพันธ์ระหว่างพยางค์ในนี้มี 5 วิธี ได้แก่ การใช้ค่ามิวชวลอินฟอร์เมชัน ค่าโคก่าล่งสอง ค่าควิกแอสโซซิเอชันเรโซ ค่าล็อกไลค์ลิสต์ และค่ามิวชวลเอ็กซ์เป็กเตชันนั้น ผลพบว่าวิธีที่ใช้ค่ามิวชวลเอ็กซ์เป็กเตชันร่วมกับการใช้โลคอลแมกซ์ อัลกอริทึม ในการรู้จำชื่อเฉพาะนั้นให้อัตราการรู้จำได้ผลดีที่สุด แต่วิธีดังกล่าวก็มีข้อเสียตรงที่ใช้เวลาในการประมวลผลที่นานเกินไป ทำให้ในงานวิจัยนี้จะใช้วิธีทางสถิติที่ให้ผลอัตราการรู้จำที่รองลงมา นั่นคือ การใช้ค่ามิวชวลอินฟอร์เมชันร่วมกับการใช้โลคอลแมกซ์ อัลกอริทึม จากนั้นเมื่อได้ชื่อเฉพาะที่เลือกมาด้วยวิธีการทางสถิติแล้ว จะเข้าสู่ส่วนที่เป็นระบบกฎ ซึ่งระบบถูกเขียนขึ้นโดยอิงกับหลักฐานที่ได้จากบริบทภายใน เช่น คำนำหน้าชื่อและใช้บริบทข้างเคียง เช่น คำปรากฏรวม เพื่อช่วยในการรู้จำและจำแนกประเภทของชื่อเฉพาะและจากการทดสอบพบว่าระบบกฎที่สร้างขึ้นสามารถจำแนกประเภทของชื่อเฉพาะโดยให้อัตราการรู้จำ (ค่า F) สำหรับชื่อเฉพาะประเภทชื่อคน 69.15% ชื่อองค์กร 62.95% และชื่อสถานที่ 38.87% ตามลำดับ โดยมีค่าความแม่นยำและค่าความครบถ้วนสำหรับชื่อเฉพาะประเภทชื่อคน 54.00% และ 96.12% ชื่อองค์กร 47.60% และ 92.93% ชื่อสถานที่ 31.67% และ 50.32% ตามลำดับ

ภาควิชา ภาษาศาสตร์

สาขาวิชา ภาษาศาสตร์

ปีการศึกษา 2548

ลายมือชื่อนิสิต ...สุฤดี ฉัตรไตรมงคล.....

ลายมือชื่ออาจารย์ที่ปรึกษา

##4580247422 : MAJOR LINGUISTICS

KEY WORD: NAMED ENTITY / RECOGNITION / CLASSIFICATION / MUTUAL EXPECTATION
/ LOCALMAX ALGORITHM

SURUDEE CHATTRIMONGKOL : NAMED ENTITY RECOGNITION AND
CLASSIFICATION IN THAI. THESIS ADVISOR : ASST. PROF. WIROTE
ARONMANAKUN, Ph.D., 75 pp. ISBN 974-53-2979-7.

This study aims to develop a Thai named entity recognition and classification system using a hybrid approach. The system is composed of two parts, which are statistical part and rule part.

Statistical part is used for extracting named entity candidates. Localmaxs algorithm and the statistical method are used for measuring associations between syllables. Five statistical methods namely Mutual Expectation, Mutual Information, Chi-square, Cubic Association ratio and Loglikelihood are tested in this part. Mutual Expectation combined with Localmaxs algorithm yields the best result, but this method uses much more times than other methods. Therefore, Mutual Information, which is the second best statistical method, combined with Localmaxs algorithm is used for extracting a chunk of syllables as a candidate of named entity. On the second part, named entity candidates will be recognized and classified by linguistic rules which are manually crafted. Internal evidence, i.e. title names, and external evidence, i.e. collocate words, are used in these rules. The system can recognize and classify named entities with the recognition rate (F-measure), precision and recall rates at 69.15% , 54.00% and 96.12% for person names, 62.95% , 47.60% and 92.93% for organization names, 38.87% , 31.67% and 50.32% for location names.

Department LINGUISTICS

Student's signature..

Surudee Chattrimongkol

Field of study LINGUISTICS

Advisor's signature..

Wrote Aronmanakun

Academic year 2005

กิตติกรรมประกาศ

ผู้วิจัยต้องขอขอบพระคุณ ผศ.ดร.วิโรจน์ อรุณมานะกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์เป็นอย่างสูง ที่ได้ให้ความช่วยเหลือและคำปรึกษาเกี่ยวกับวิทยานิพนธ์ฉบับนี้มาโดยตลอด หากไม่ได้คำแนะนำจากอาจารย์แล้ว วิทยานิพนธ์ฉบับนี้คงจะสำเร็จสมบูรณ์ไปไม่ได้เลย และผู้วิจัยขอขอบพระคุณ ผศ.ดร.สุดาพร ลักษณะียนาวิน และรศ.ดร. กิ่งกาญจน์ เทพกาญจนานุกรมการสอบวิทยานิพนธ์ที่ได้ให้คำปรึกษาและเสียสละเวลาเพื่อตรวจสอบวิทยานิพนธ์ฉบับนี้

สุดท้ายนี้ ผู้วิจัยขอขอบคุณคณาจารย์ภาควิชาภาษาศาสตร์ทุกท่านที่ได้ประสิทธิ์ประสาทความรู้ด้านภาษาศาสตร์ให้แก่ผู้วิจัย รวมทั้งขอขอบคุณเจ้าหน้าที่และเพื่อนพี่น้องภาควิชาภาษาศาสตร์ทุกคนที่ช่วยเหลือและให้กำลังใจผู้วิจัย และขอขอบพระคุณ คุณพ่อสงวน คุณแม่สุภาภรณ์ ฉัตรไตรมงคลสำหรับโอกาสและกำลังใจอันมากเหลือที่ได้รับ รวมทั้งขอขอบคุณพี่ชายคุณมงคล ฉัตรไตรมงคลที่คอยสนับสนุนและให้กำลังใจมาโดยตลอด

บทที่	หน้า
2.2.2.1.3	การหาค่า Pearson's Chi-square 13
2.2.2.1.4	การหาค่า Dunning's log likelihood 13
2.2.2.2	วิธีการทางสถิติแบบที่มองการสูญเสียหน่วยย่อยใดๆ ภายในว่ายอมรับได้มากหรือไม่ 15
2.2.2.2.1	การหาค่า Mutual Expectation 15
2.2.3	ระบบแบบลูกผสม (hybrid) 17
2.3	คลังข้อมูลที่ใช้ 18
2.4	ระบบการรู้จำชื่อเฉพาะภาษาไทย 21
3	การหาขอบเขตของชื่อเฉพาะ 23
3.1	วิธีทางสถิติที่ประยุกต์ใช้ 23
3.2	การกำหนดขอบเขตของชื่อเฉพาะ 26
3.3	เปรียบเทียบวิธีทางสถิติที่ใช้ในการหาขอบเขตชื่อเฉพาะ 28
3.4	อภิปรายผล 32
4	การรู้จำชื่อเฉพาะ 33
4.1	การจำแนกประเภทชื่อเฉพาะ 33
4.1.1	หลักฐานภายใน (internal evidence) 33
4.1.2	หลักฐานภายนอก (external evidence) 39
4.2	การสร้างกฎการจำแนกประเภทชื่อเฉพาะ 39
4.3	ผลการตัดสินใจและจำแนกประเภทชื่อเฉพาะ 45
4.4	อภิปรายผล 47
5	สรุปผลการวิจัยและข้อเสนอแนะ 50
5.1	สรุปผลการวิจัย 50
5.2	ข้อเสนอแนะ 51
	รายการอ้างอิง 54
	ภาคผนวก 58
	ประวัติผู้เขียนวิทยานิพนธ์ 75

สารบัญตาราง

ตารางที่		หน้า
1	(n-1) grams and missing words	16
2	ผลการทดสอบค่าความแม่นยำ ค่าความครบถ้วนและค่า F ที่ได้จากวิธีทางสถิติ 5 วิธี	30
3	ผลการทดสอบเวลาที่ใช้ในการรู้จำที่ได้จากวิธีทางสถิติ 5 วิธี	31
4	การแบ่งประเภทกลุ่มคำนำหน้าชื่อ	35
5	ผลการทดสอบค่าความแม่นยำ ค่าความครบถ้วน และค่า F เมื่อใช้โปรแกรมกฎ	45
6	ผลการจำแนกประเภทของโปรแกรมกฎที่เขียนขึ้น	46