# A SEMI-AUTOMATIC CONSTRUCTION OF THAI WORDNET LEXICAL DATABASE FROM MACHINE READABLE DICTIONARIES

**PATANAKUL SATHAPORNRUNGKIJ**

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2004**

**ISBN 974-04-5256-6**

การสร้างฐานข้อมูลคำศัพท์ไทยเวิร์ดเน็ตจากพจนานุกรมอิเล็กทรอนิกส์แบบกึ่งอัตโนมัติ(A SEMI-AUTOMATIC CONSTRUCTION OF THAI WORDNET LEXICAL DATABASE FROM MACHINE READABLE DICTIONARIES)

พัฒณกุล สถาพรรุ่งกิจ 4237668 SCCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการควบคุมวิทยานิพนธ์ : ชาญยศ ปลื้มปีติวิริยะเวช, Ph.D., ธันวดี สุเนตนันท์, Ph.D.

บทคัดย่อ

จากความสำคัญของฐานข้อมูลคำศัพท์ที่เพิ่มขึ้นนั้นทำให้เวิร์ดเน็ต (wordnet) กลายเป็นฐานข้อมูลคำศัพท์ที่ใช้กันอย่างแพร่หลายสำหรับงานด้านการประมวลผลภาษาธรรมชาติเนื่องจาก WordNet มีข้อมูลเกี่ยวกับไวยากรณ์ซึ่งมีความสำคัญในงานด้านนี้มาก จากความแพร่หลายของเวิร์ดเน็ตทำให้เวิร์ดเน็ตได้ถูกนำมาสร้างในภาษาต่างๆ เว้นแต่ภาษาไทย ดังนั้นงานวิจัยนี้ได้นำเสนอการสร้างฐานข้อมูลคำศัพท์ไทยเวิร์ดเน็ตจากพจนานุกรมอิเล็กทรอนิกส์แบบกึ่งอัตโนมัติ ซึ่งมุ่งเน้นการสร้างฐานข้อมูลคำศัพท์เวิร์ดเน็ตภาษาไทยขึ้น

งานวิจัยนี้ได้นำเสนอแนวทางการสร้างเวิร์ดเน็ตซึ่งจะดึงข้อมูลเกี่ยวกับคำศัพท์ต่างๆ จากพจนานุกรมอิเล็กทรอนิกส์ที่มีอยู่ โดยจะเริ่มกระบวนการด้วยการเค้นข้อมูลจากพจนานุกรมอิเล็กทรอนิกส์คือ Princton WordNet ซึ่งเป็นฐานข้อมูลคำศัพท์เวิร์ดเน็ตภาษาอังกฤษที่มีโครงข่ายความหมายของคำ และ LEXiTRON ซึ่งเป็นพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย ⇔ ภาษาอังกฤษ ฐานข้อมูลคำศัพท์ทั้งสองนี้มีคำศัพท์จำนวนมาก กระบวนการถัดมาจะเป็นการสร้างชุดความสัมพันธ์ระหว่างคำภาษาไทยและความหมายจากเวิร์ดเน็ตภาษาอังกฤษโดยใช้ข้อมูลที่ได้จากการเค้นข้อมูล ซึ่งความสัมพันธ์ระหว่างคำภาษาไทยและความหมายจากเวิร์ดเน็ตภาษาอังกฤษนี้จะถูกนำเสนอในรูปของความสัมพันธ์ระหว่างคำต่างๆ และความสัมพันธ์เหล่านี้จะถูกตรวจสอบและนำมาสร้างแบบจำลองทางสถิติที่มาช่วยในการตรวจสอบ candidate link ให้สร้างไทยเวิร์ดเน็ตได้รวดเร็วยิ่งขึ้น

จากการนำ Logistic Regression ช่วยในการตรวจสอบ ชุดความสัมพันธ์ที่ได้สร้างขึ้นนั้นจะได้ความแม่นยำของ translation link เท่ากับ 76% ด้วยความครอบคลุมที่ 44,844 ความสัมพันธ์, 19,582 ความหมาย, and 13,730 คำ ซึ่งจะใช้เวลาในการสร้างพจนานุกรมน้อยกว่าการสร้างพจนานุกรม จากผลที่ได้รับนี้ถือได้ว่ามีประสิทธิผลดี

112 หน้า ISBN 974-04-5256-6

A SEMI-AUTOMATIC CONSTRUCTION OF THAI WORDNET LEXICAL DATABASE FROM MACHINE READABLE DICTIONARIES

PATANAKUL SATHAPORNRUNGKIJ 4237668 SCCS/M

M.Sc. (COMPUTER SCIENCE)

THESIS ADVISORS: CHARNYOTE PLUEMPITIWIRIYAWEJ, Ph.D., THANWADEE SUNETNANTA, Ph.D.

## ABSTRACT

A significant increase in the use of lexical databases has led WordNet to become one of the most widely used lexical information source for Natural Language Processing applications. WordNet is a lexical database that, apart from lexical information, also provides semantic information and is available in many languages. Despite its popularity, the database still lacks support for Thai language.

This research presents a semi-automatic method for constructing Thai WordNet lexical databases from machine readable dictionaries (MRDs). The method starts with extracting lexical information from available MRDs: the Princeton WordNet, also known as a semantic network, and the LEXiTRON, a Thai⇔English dictionary. Both MRDs contain hundreds of thousands of English and Thai words. Then sets of candidate translation links are generated with respect to the extracted lexical information. These translation links are the representations of the semantic relationships among the words. A subset of these candidate links is later on verified and used to generate a model, which is finally used to automatically construct the Thai WordNet.

Our Thai WordNet has a coverage of 44,844 links, 19,582 synsets, and 13,730 words. The logistic regression is used to choose translation links among candidate links. It shows an accuracy of 76%. This semi-automatic method takes less time than the manual. The result shows that our method is effective.

KEY WORDS: THAI WORDNET/ LEXICAL DATABASE/ MACHINE READABLE DICTIONARIES

112 pp. ISBN 974-04-5256-6