# วิธีการกำจัดความกำกวมของคำหลายความหมายโดยใช้ความหมายเชิงการกระจายและความหมายแฝง

นางสาวสุนีย์ พงษ์พินิจภิญโญ



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2547
ISBN 974-53-1612-1

# A DISTRIBUTIONAL SEMANTICS AND LATENT SEMANTICS APPROACH FOR WORD SENSE DISAMBIGUATION

Miss Sunee Pongpinigpinyo



A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic year 2004

ISBN 974-53-1612-1

| | |
|---|---|
| Thesis Title | A DISTRIBUTIONAL SEMANTICS AND LATENT SEMANTICS APPROACH FOR WORD SENSE DISAMBIGUATION |
| By | Miss Sunee Pongpinigpinyo |
| Field of Study | Computer Engineering |
| Thesis Advisor | Associate Professor Dr.Wanchai Rivepiboon |

---

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctor's Degree

... ... ... ... ... ... ... ... ... ... ... ... ... ... Dean of the Faculty of Engineering
(Professor Direk Lavansiri, Ph.D.)

THESIS COMMITTEE

... ... ... ... ... ... ... ... ... ... ... ... ... ... Chairman
(Associate Professor Boonserm Kijsirikul, D.Eng.)

... ... ... ... ... ... ... ... ... ... ... ... ... ... Thesis Advisor
(Associate Professor Wanchai Rivepiboon, Ph.D.)

... ... ... ... ... ... ... ... ... ... ... ... ... ... Member
(Yunyong Teng-amnuay, Ph.D.)

... ... ... ... ... ... ... ... ... ... ... ... ... ... Member
(Assistant Professor Wirote Aroonmanakun, Ph.D.)

... ... ... ... ... ... ... ... ... ... ... ... ... ... Member
(Assistant Professor Arnon Rungsawang, Ph.D.)

สุนีย์ พงษ์พินิจภิญโญ : วิธีการกำจัดความกำกวมของคำหลายความหมายโดยใช้ความหมายเชิงการกระจายและความหมายแฝง. (A DISTRIBUTIONAL SEMANTICS AND LATENT SEMANTICS APPROACH FOR WORD SENSE DISAMBIGUATION) อ. ที่ปรึกษา : รศ. ดร.วันชัย ริ้วไพบูลย์ จำนวนหน้า 86 หน้า. ISBN 974-53-1612-1.

วิทยานิพนธ์นี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการแก้ปัญหาความกำกวมของคำหลายความหมายในภาษาไทยโดยใช้วิธีการเรียนรู้แบบไม่ต้องใช้ผู้สอน (unsupervised learning technique) ซึ่งไม่จำเป็นต้องใช้แหล่งความรู้ต่างๆมากมาย เช่น ข้อความที่กำกับความหมาย หรือพจนานุกรม วิธีการที่นำเสนอในวิทยานิพนธ์นี้เป็นวิธีการใหม่เพราะยังไม่เคยมีใครนำวิธีการนี้มาใช้กับภาษาไทยที่ยังไม่ได้รับความสนใจมากนักในงานวิจัยด้านการประมวลผลภาษาธรรมชาติ วิธีการของผู้วิจัยที่นำเสนอคือการใช้ความหมายเชิงกระจายที่มีพื้นฐานจากสมมติฐานของการกระจายที่ว่าคำที่มีความหมายคล้ายกันจะปรากฏอยู่ในบริบทที่คล้ายกัน การนำเสนอวิธีนี้เกี่ยวข้องกับการใช้ฐานข้อมูลของภาษา (corpora) เพื่อที่ตรวจสอบบริบทของแต่ละคำที่ปรากฏอยู่แล้วจึงคำนวณหาความคล้ายกันระหว่างบริบทที่กระจาย กลุ่มความหมายของคำเป็นกลุ่มคำที่เกิดขึ้นด้วยกันเสมอกับคำหลายความหมาย ซึ่งแต่ละกลุ่มความหมายประกอบไปด้วยกลุ่มคำที่เกิดขึ้นด้วยกันเสมอที่ให้ความหมายเหมือนกัน

งานวิทยานิพนธ์นี้ให้ประโยชน์ไม่เพียงสำหรับการพัฒนาขั้นต่อไปของวิธีการแก้ปัญหาความกำกวมของคำหลายความหมายในภาษาไทยแต่ยังสามารถนำไปปรับใช้กับลักษณะของปัญหา (domain) ที่แตกต่างกัน และขอบเขตตัวอย่างข้อมูลขนาดใหญ่ขึ้นได้



| | | |
|---|---|---|
| ภาควิชา | วิศวกรรมคอมพิวเตอร์ | ลายมือชื่อนิสิต..... |
| สาขาวิชา | วิศวกรรมคอมพิวเตอร์ | ลายมือชื่ออาจารย์ที่ปรึกษา..... |
| ปีการศึกษา | 2547 | |

## 4371818121 : MAJOR COMPUTER ENGINEERING
KEY WORD: WORD SENSE DISAMBIGUATION / NATURAL LANGUAGE PROCESSING / DISTRIBUTIONAL SEMANTICS / LATENT SEMANTICS

SUNEE PONGPINIGPINYO : (A DISTRIBUTIONAL SEMANTICS AND LATENT SEMANTICS APPROACH FOR WORD SENSE DISAMBIGUATION. THESIS ADVISOR: ASSOC. PROF. WANCHAI RIVEPIBOON, Ph.D., 86 pp. ISBN 974-53-1612-1.

This thesis aims at developing a methodology of word sense disambiguation in Thai using unsupervised learning techniques that do not rely on any knowledge intensive resources like sense–tagged text or dictionaries. The thesis approach is novel since the work focuses on Thai language which has not received much attention in the Natural Language Processing literature. Our proposed method is the distributional semantics which is based on the *distributional hypothesis* that *similar words appear in similar contexts*. This approach involves using corpora to examine the contexts which each word appears in and then calculating the similarity between context distributions. The sense clusters are occurrences of words that have been grouped into clusters based from raw text where each cluster consists of occurrences having the same meaning.

This thesis work also gives benefit not only to the further development of word sense disambiguation for Thai language but also to be portable to different domains, and can be scaled up easily for larger samples of text.



Department Computer Engineering
Field of study Computer Engineering
Academic year 2004

Student's signature ...Sunee Pongpinigpinyo...
Advisor's signature .....

# ACKNOWLEDGEMENTS

I sincerely feel that this thesis would not have been possible, without the tremendous help, support, guidance and lastly, but not least, the true friendship and love, given by various people who I have both the honour and the luck to get acquainted with.

Firstly, I would like to thank my supervisor, Associate Professor Dr. Wanchai Rivepiboon, for the guidance and advice. Next, I would like to thank my thesis committee: Associate Professor Dr. Boonserm Kijsirikul, Dr. Yunyong Teng-amnuay, Assistant Professor Dr. Arnon Rungsawang, and Assistant Professor Dr. Wirote Aroonmanakun who provided valuable thesis comments. Especially, a very grateful thank you is given to Assistant Professor Dr. Arnon Rungsawang, without him I would probably never completed this thesis. He gives his time for thesis guidance and discussion on various aspects of the research.

I am also specially thankful to Associate Professor Dr. Prabhas Chongstitvatana, Yodthong Rodkaew and Chaiwat Jessadapakorn who lift me up when I felt down with stress. They give me the strength and encouragement thru all my Ph.D student life. They cheer me up and get me ready for another day of research.

Also I would like to express my gratitude towards my good friends Dr. Judith Young for taking the trouble to help by proof reading drafts of this thesis.

My sincere thanks are given to all fellows of the Software Engineering Laboratory for their helps with programming, technical computer problems and being my recreation. My special thanks are given to Miss Jutapuck Pugsee for her invaluable and kindness helps whenever I ask for.

I am indebted to the Commission on Higher Education Ministry of Education for funding source for my study.

Last but not least, I would like to express my greatest gratitude to the persons to whom I owe the most, my beloved parents who have always given me pure love, devotion and moral support throughout the various stages of my life, understanding and great patience waiting for me during these years to finish my study. Thank you for being who you are; the best parents for me.

Finally, I want to offer my wholehearted praises and thanksgiving to our Lord Buddha for giving enlightenment Dharma to make my body and soul be in peace when I feel suffer from attachment and stress.

# CONTENTS

Page

# CONTENTS (continued)



Page

# CONTENTS (continued)



Page

# LIST OF TABLES



Table Page

# LIST OF FIGURES



Figure Page