โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

นางสาวสมปรารถนา รัทยานนท์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
ภาควิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย จุฬาล กรณ์มหาวิทยาลัย
พ.ศ. 2535

ISBN 974-581-176-9

DATA STRUCTURE FOR THAI ELECTRONIC DICTIONARY

Miss Somprathana Ratthayanond

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science

Department of Computer Engineering

Graduate Scool

1992

ISBN 974-581-176-9

หัวข้อวิทยานิพนธ์	โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย
โดย	นางสาวสมปรารถนา รัทยานนท์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	รองศาสตราจารย์ ดร. วิลาศ วูวงศ์
	ผู้ช่วยศาสตราจารย์ สุยุชน์ สัตยประกอบ

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้วิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ ดร. ถาวร วัชราภัย)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รศ. เดือน สินธุพันธ์ประทุม)

..... อาจารย์ที่ปรึกษา
(รศ.ดร. วิลาศ วูวงศ์)

..... อาจารย์ที่ปรึกษาร่วม
(ผศ. สุยุชน์ สัตยประกอบ)

..... กรรมการ
(อ.ดร. ยรรยง เต็งอำนวย)

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสีเขียวนี้เพียงแผ่นเดียว

สมปรารถนา รัทยานนท์ : โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย (DATA STRUCTURE FOR THAI ELECTRONIC DICTIONARY) อ.ที่ปรึกษา : รศ.ดร.วิลาศ วูวงศ์, ผศ.สุยุทน์ สัตยประกอบ, 90 หน้า. ISBN 974-581-176-9.

พจนานุกรมอิเล็กทรอนิกส์เป็นแหล่งเก็บข้อมูลสำหรับงานประมวลผลด้านภาษาศาสตร์ เช่น การตัดคำและการตรวจสอบตัวสะกดในโปรแกรมประมวลผลคำ การวิเคราะห์ไวยากรณ์ในงานประมวลผลภาษาธรรมชาติ จากความก้าวหน้าของงานการประมวลผล ภาษาไทยด้วยคอมพิวเตอร์ที่มีมากขึ้น ทำให้มีความสนใจในการพัฒนาพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยมากขึ้นตามไปด้วย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยโดยใช้โครงสร้างข้อมูลแบบดับเบิลอะเรย์ที่มีการสืบค้นแบบดิจิตอล 2 แบบ ดังนี้คือ แบบแรก เป็นพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่จัดเก็บคำศัพท์ต่าง ๆ โดยตรง ส่วนแบบที่ 2 เป็นพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่จัดเก็บคำโดดที่ได้มาจากการแยกคำศัพท์ ซึ่งผลของการจัดเก็บได้ว่า เนื้อที่ที่ใช้ในการเก็บพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยแบบที่ 2 น้อยกว่าแบบแรก นอกจากนี้ ยังพัฒนาอัลกอริทึมการตัดคำโดยใช้พจนานุกรมที่ให้ผลลัพธ์เป็นคำศัพท์ทุกคำที่ปรากฏในพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

สำหรับผลการทดสอบประสิทธิภาพการทำงานพบว่า อัลกอริทึมการสืบค้นคำศัพท์ของพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่พัฒนาขึ้นใช้เวลามากกว่าพจนานุกรมของมหาวิทยาลัยเกษตรศาสตร์ แต่สามารถนำอัลกอริทึมดังกล่าวมาเป็นแนวทางในการพัฒนาอัลกอริทึมการตัดคำโดยใช้พจนานุกรมได้โดยง่าย

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต สมปรารถนา รัทยานนท์

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา

ปีการศึกษา 2538 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม สุยุทน์ สัตยประกอบ

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสีเขียวนี้เพียงแผ่นเดียว

##C116994 : MAJOR COMPUTER SCIENCES
KEY WORD : DATA STRUCTURE/DICTIONARY/ELECTRONIC/THAI
SOMPRATHANA RATTHAYANOND : DATA STRUCTURE FOR THAI ELECTRONIC DICTIONARY. THESIS ADVISOR : ASSOCIATE PROF.DR. VILAS WUWONGSE ASSISTANT PROF. SUYUT SATAYAPRAKORB. 90 pp., ISBN 974-581-176-9.

Electronic dictionaries are storages of machine readable lexical items. They are used in advanced word processors for word segmentation and spelling checkers, and natural language processing for syntactic, semantic and discourse analysis. Due to the advancement in computer processing of Thai language, more attention has recently been paid to the development of Thai electronic dictionaries.

This study proposes and develops two frameworks for Thai electronic dictionaries employing double-array digital search tree. The first framework treats a Thai word as a lexical entity and directly applies the double-array digital search tree to store lexical entities. The second one recognizes the fact that many Thai words are composed of few isolated words, and therefore stores a word in terms of a few isolated words, resulting in less storage space. Based on the dictionaries, a Thai word segmentation algorithm has been developed which produces all possible segmentations.

Experiments have been conducted to evaluate performance of the proposed frameworks. It has been found out that it takes more time for the frameworks to retrieve words than the Kasetsart University's approach but they allow the development of simple word segmentation algorithm.

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต สมปราถนา วิทยานนท์

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา

ปีการศึกษา 2534 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ รศ.ดร.วิลาส วุวงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ และ ผศ.สุยุชน์ สัตยประกอบ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ซึ่งท่านได้สละเวลาให้คำแนะนำ และข้อคิดเห็นต่างๆ ของการวิจัยมาด้วยดีตลอด

ขอขอบพระคุณ รศ.ยืน ภู่วรวรรณ ที่ได้ อนุเคราะห์ข้อมูลสำหรับการทำวิทยานิพนธ์ ขอขอบพระคุณ คุณพรพจน์ สุริยะวงศ์ ที่เสียสละเวลาให้คำแนะนำ และข้อคิดเห็นต่างๆ ในการทำวิทยานิพนธ์ครั้งนี้ ขอขอบพระคุณเจ้าหน้าที่ประจำห้องปฏิบัติการคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ที่เอื้อเฟื้ออุปกรณ์บริภัณฑ์ และเนื่องจากทุนการวิจัยครั้งนี้บางส่วนได้รับมาจากทุนอุดหนุนการวิจัยของบัณฑิตวิทยาลัย จึงขอขอบคุณบัณฑิตวิทยาลัยมา ณ ที่นี้ด้วย นอกจากนี้ขอขอบพระคุณเพื่อนๆ พี่ๆ น้องๆ ชาวนิสิตปริญญาโททุกๆ ท่านที่คอยเป็นกำลังใจตลอดมา

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา-มารดา ซึ่งสนับสนุนในด้านการเงิน รวม ทั้ง พี่ และน้องที่คอยให้กำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า