การเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ

นาย ประยุทธ สุวรรณวิสารท

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2541
ISBN 974-332-123-3

TRANSLITERATED WORD ENCODING FOR THAI-ENGLISH CROSS-LANGUAGE RETRIEVAL

Mr. Prayut Suwanvisat

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 1998

ISBN 974-332-123-3

หัวข้อวิทยานิพนธ์	การเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ
โดย	นาย ประยุทธ สุวรรณวิสารท
ภาควิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จูตระกูล

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ นายแพทย์ สุกวัฒน์ ชุติวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร. วันชัย ริ้วไพบูลย์)

..... อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จูตระกูล)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. ประภาส จงสถิตย์วัฒนา)

..... กรรมการ
(อาจารย์ ดร. บุญเสริม กิจศิริกุล)

ประยุทธ สุวรรณวิสารท : การเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ (TRANSLITERATED WORD ENCODING FOR THAI-ENGLISH CROSS-LANGUAGE RETRIEVAL) อ.ที่ปรึกษา : ผศ. ดร. สมชาย ประสิทธิ์จูตระกูล. 85 หน้า. ISBN 974-332-123-3.

วิทยานิพนธ์ฉบับนี้นำเสนอขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ ซึ่งอนุญาตให้ใช้ข้อคำถามที่เป็นคำทับศัพท์ภาษาอังกฤษหรือภาษาไทยในการค้นคืนเอกสารที่มีคำหลักตรงกันในอีกภาษา โดยมีข้อสมมุติฐานว่าสามารถทำการค้นคืนข้ามภาษาไทย-อังกฤษได้โดยไม่ต้องอาศัยพจนานุกรม ขั้นตอนวิธีที่นำเสนอแบ่งออกเป็นสองส่วนคือ (1) ขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามแบบภาษาไทยทับศัพท์ภาษาอังกฤษ และ (2) ขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามแบบภาษาอังกฤษทับศัพท์ภาษาไทย ขั้นตอนวิธีการค้นคืนข้ามภาษานี้จะทำงานโดยการเข้ารหัสคำในข้อคำถามแล้วนำรหัสคำที่ได้ไปเปรียบเทียบกับรหัสคำในดัชนีคำหลัก การเปรียบเทียบรหัสคำในการข้ามภาษาแบบภาษาไทยทับศัพท์ภาษาอังกฤษจะอาศัยวิธีการเปรียบเทียบแบบเหมือนกันทุกประการ ส่วนการเปรียบเทียบรหัสคำในการข้ามภาษาแบบอังกฤษทับศัพท์ภาษาไทยจะอาศัยวิธีการเปรียบเทียบเชิงประมาณและแยกเปรียบเทียบส่วนพยัญชนะและสระออกจากกัน โดยใช้เทคนิคกำหนดการพลวัต ผลการทดลองแสดงให้เห็นว่าขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษแบบภาษาไทยทับศัพท์ภาษาอังกฤษมีค่าเรียกคืนสูงถึง 90 เปอร์เซ็นต์ และค่าแม่นยำสูงถึง 78 เปอร์เซ็นต์ เมื่อคำทับศัพท์มีความยาวมากกว่า 7 ตัวอักษร และแบบภาษาอังกฤษทับศัพท์ภาษาไทยมีค่าเรียกคืนสูงถึง 73 เปอร์เซ็นต์ และค่าแม่นยำสูงถึง 69 เปอร์เซ็นต์

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต ประยุทธ สุวรรณวิสารท

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา สมชาย ประสิทธิ์จูตระกูล

ปีการศึกษา 2541 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม .....

4070330721 : MAJOR COMPUTER SCIENCE
KEY WORD: TRANSLITERATED WORD / ENCODING / CROSS-LANGUAGE / SOUNDEX
PRAYUT SUWANVISAT : TRANSLITERATED WORD ENCODING FOR THAI-ENGLISH CROSS-LANGUAGE RETRIEVAL. THESIS ADVISOR : ASSIST. PROF. SOMCHAI PRASITJUTRAKUL, Ph.D. 85 pp. ISBN 974-332-123-3.

This thesis presents two algorithms for transliterated word encoding for Thai-English cross-language retrieval. The algorithms enable retrieval of documents containing either the English-to-Thai or Thai-to-English transliterated keywords. We have a hypothesis that cross-language retrieval does not use a dictionary. The proposed algorithms are (1) English-to-Thai transliterated word encoding for cross-language retrieval algorithm and (2) Thai-to-English transliterated word encoding for cross-language retrieval algorithm. This cross-language retrieval is done by encoding each word in the query terms and then matching the query code with codes of keywords in the index. The English-to-Thai cross-language retrieval uses exact code matching. On the other hand, the Thai-to-English uses approximate code matching (separatedly done for consonant and vowel parts) by using dynamic programming technique. Experimental results showed that for keywords of length longer than seven characters the recall and precision of the English-to-Thai transliterated word cross-language retrieval are 90% and 78%, respectively. The recall and precision of the Thai-to-English transliterated word are around 73% and 69%, respectively.

ภาควิชา วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต ประยุทธ สุวรรณวิสาท
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่ออาจารย์ที่ปรึกษา สมชาย ประสิทธิ์จุตระกูล
ปีการศึกษา 2541	ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จูตระกูล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่าง ๆ ในการวิจัยมาด้วยดีตลอด รวมทั้งตรวจแก้วิทยานิพนธ์ฉบับนี้อย่างละเอียด ผู้เขียนขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร. ประกาส จงสถิตย์วัฒนา ซึ่งท่านได้ให้แนวคิดในการทำงานวิจัยในภาควิชาวิศวกรรมคอมพิวเตอร์

ขอขอบพระคุณมูลนิธิเพื่อการศึกษาคอมพิวเตอร์และการสื่อสาร (Computer and Communication Education Foundation) และมูลนิธิเพื่อการศึกษาและวิจัยวิทยาศาสตร์คอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย ที่สนับสนุนทุนการศึกษาและทุนในการทำงานวิจัยชิ้นนี้

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณมารดา ซึ่งสนับสนุนและให้กำลังใจแก่ผู้วิจัยเสมอจนสำเร็จการศึกษา

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญตาราง ..... ฌ

สารบัญภาพ ..... ฎ

1. บทนำ 1

1.1 ความเป็นมาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของการวิจัย 3

1.3 ขอบเขตของการวิจัย 3

1.4 ขั้นตอนและวิธีดำเนินการวิจัย 3

1.5 ประโยชน์ที่คาดว่าจะได้รับ 3

1.6 ผลงานที่ตีพิมพ์จากงานวิจัย 4

1.7 โครงสร้างของวิทยานิพนธ์ 4

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 5

2.1 การถอดอักษร 5

2.2 การใช้ตัวอักษรโรมันเพื่อการถ่ายเสียง 7

2.3 หลักเกณฑ์การทับศัพท์ 8

2.4 การค้นคืนข้ามภาษา 8

2.5 ขั้นตอนวิธีชาวด์เด็กซ์ภาษาอังกฤษ 9

2.6 ขั้นตอนวิธีชาวด์เด็กซ์ภาษาไทย 11

2.7 ขั้นตอนวิธีระยะแก้ไขสั้นที่สุด (Minimum Edit Distance) 22

2.8 สรุป 24

3. ขั้นตอนวิธีการค้นคืนข้ามภาษาแบบภาษาไทยทับศัพท์ภาษาอังกฤษ 25

3.1 โครงสร้างของระบบค้นคืนข้ามภาษาไทยทับศัพท์ภาษาอังกฤษ 25

3.2 ขั้นตอนวิธีการเข้ารหัสคำ 25

3.3 วิธีการทดลอง 28

3.4 ผลการทดลอง 29

3.5 สรุป 32

4. ขั้นตอนวิธีการค้นคืนข้ามภาษาแบบภาษาอังกฤษทับศัพท์ภาษาไทย 33
4.1 โครงสร้างของระบบค้นคืนข้ามภาษาอังกฤษทับศัพท์ภาษาไทย 33
4.2 ขั้นตอนวิธีการเข้ารหัสคำ 34
4.2.1 การประมวลผลตัวอักษรเบื้องต้น 34
4.2.1.1 การลดรูปและตัดวรรณยุกต์ 35
4.2.1.2 การแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล 39
4.2.1.3 การถอดอักษร 40
4.2.2 การย้ายตำแหน่งสระ 44
4.3 ขั้นตอนวิธีเปรียบเทียบรหัสคำ 44
4.3.1 การคำนวณหาค่าความแตกต่างของรหัสคำ 45
4.3.1.1 การกำหนดต้นทุนในการแก้ไขอักขระ 47
4.3.2 เงื่อนไขในการเปรียบเทียบรหัสคำ 55
4.4 วิธีการทดลอง 57
4.5 ผลการทดลอง 58
4.6 สรุป 60
5. สรุปผลการวิจัยและข้อเสนอแนะ 61
5.1 สรุปผลการวิจัย 61
5.2 ข้อดีและข้อเสียของขั้นตอนวิธี 62
5.3 ข้อเสนอแนะ 63
รายการอ้างอิง 64
ภาคผนวก 66
ภาคผนวก ก 67
ภาคผนวก ข 73
ภาคผนวก ค 81
ประวัติผู้เขียน 85

สารบัญตาราง

หน้า

ตารางที่ 2.1 ตัวอย่างการถอดอักษรจากภาษาต้นแบบไปยังภาษาเป้าหมาย 6
ตารางที่ 2.2 การกำหนดรหัสชาวด์เด็กซ์ภาษาอังกฤษ 11
ตารางที่ 2.3 การกำหนดรหัสตัวอักษรของรหัสชาวด์เด็กซ์ภาษาไทย 12
ตารางที่ 2.4 การกำหนดรหัสตัวเลขของรหัสชาวด์เด็กซ์ภาษาไทย 13
ตารางที่ 2.5 ตัวอย่างการเข้ารหัสชาวด์เด็กซ์ภาษาไทย 14
ตารางที่ 2.6 การกำหนดรหัสอักษรสำหรับอักขระตัวแรกของรหัสชาวด์เด็กซ์ภาษาไทย 15
ตารางที่ 2.7 การกำหนดรหัสอักษรสำหรับอักขระตัวถัดไปของรหัสชาวด์เด็กซ์ภาษาไทย 15
ตารางที่ 2.8 ตัวอย่างการเข้ารหัสชาวด์เด็กซ์ภาษาไทย 16
ตารางที่ 2.9 การเข้ารหัสสำหรับพยัญชนะต้น 17
ตารางที่ 2.10 การเข้ารหัสสำหรับสระ 18
ตารางที่ 2.11 การเข้ารหัสสำหรับตัวสะกด 18
ตารางที่ 2.12 ตัวอย่างการเข้ารหัสชาวด์เด็กซ์ภาษาไทย 21
ตารางที่ 3.1 กลุ่มเสียงของพยัญชนะไทย 21 กลุ่ม 26
ตารางที่ 3.2 กลุ่มอักษรไทยและกลุ่มอักษรอังกฤษที่ออกเสียงคล้ายกันในรหัสชาวด์เด็กซ์ 26
ตารางที่ 3.3 การกำหนดรหัสชาวด์เด็กซ์สำหรับอักษรไทยและอักษรอังกฤษที่นำเสนอ 28
ตารางที่ 3.4 รายละเอียดจำนวนคำศัพท์ที่ใช้ในการทดลอง 29
ตารางที่ 3.5 ความสัมพันธ์ระหว่างความยาวน้อยสุดของรหัสคำกับความยาวคำเฉลี่ย 30
ตารางที่ 3.6 ความสัมพันธ์ระหว่างความยาวน้อยสุดของรหัสคำกับประสิทธิผล 30
ตารางที่ 4.1 การใช้อักษรโรมันแทนพยัญชนะไทยของ ISO 41
ตารางที่ 4.2 การถอดอักษรอังกฤษเป็นอักษรไทยในส่วนของพยัญชนะที่นำเสนอ 42
ตารางที่ 4.3 การถอดอักษรอังกฤษเป็นอักษรไทยในส่วนของสระที่นำเสนอ 43
ตารางที่ 4.4 กลุ่มเสียงของพยัญชนะไทย 48
ตารางที่ 4.5 อักขระที่ไม่สามารถแยกความแตกต่างในการถอดอักษร 48
ตารางที่ 4.6 อักขระแทนมาตราต่าง ๆ 49
ตารางที่ 4.7 อักษรควบไม่แท้ 50
ตารางที่ 4.8 อักษรนำเสียงสนิท 50
ตารางที่ 4.9 อักษรควบที่เป็นตัวสะกด 50
ตารางที่ 4.10 ตัวอย่างตารางการกำหนดต้นทุนในการแทนที่อักขระที่เป็นพยัญชนะ 51
ตารางที่ 4.11 ตารางการกำหนดต้นทุนในการแทนที่อักขระที่เป็นสระ 52

ตารางที่ 4.12 ตารางการกำหนดต้นทุนในการเพิ่มหรือการลบอักขระ.....53
ตารางที่ 4.13 ความสัมพันธ์ระหว่างค่าแอลฟ่ากับประสิทธิผลของระบบค้นคืน.....58
ตารางที่ ก.1 ตารางกำหนดต้นทุนในการแก้ไขอักขระพยัญชนะ.....68
ตารางที่ ข.1 การใช้อักษรโรมันแทนพยัญชนะไทยของ ISO 73
ตารางที่ ข.2 การใช้อักษรโรมันแทนอักขระสระไทยของ ISO 74
ตารางที่ ข.3 การใช้อักษรโรมันแทนพยัญชนะไทยของราชบัณฑิตยสถาน 75
ตารางที่ ข.4 การใช้อักษรโรมันแทนสระไทยของราชบัณฑิตยสถาน 76
ตารางที่ ข.5 การใช้อักษรโรมันแทนพยัญชนะไทยของจันทร์เพ็ญ โวหารสุนทร 77
ตารางที่ ข.6 การใช้อักษรโรมันแทนสระไทยของจันทร์เพ็ญ โวหารสุนทร 78
ตารางที่ ข.7 กลุ่มเสียงพยัญชนะไทยและสัญลักษณ์เสียง 79
ตารางที่ ข.8 สัญลักษณ์เสียงสระไทย 80

สารบัญภาพ

หน้า

รูปที่ 2.1 โปรแกรมการเข้ารหัสชาวด์เด็กซ์กาษาอังกฤษ 10
รูปที่ 2.2 โครงร่างเครื่องเข้ารหัสชาวด์เด็กซ์ 19
รูปที่ 2.3 เครื่องเข้ารหัสชาวด์เด็กซ์ 20
รูปที่ 2.4 โปรแกรมระยะแก้ไขสั้นที่สุด โดยใช้เทคนิคกำหนดการพลวัต 23
รูปที่ 3.1 ความสัมพันธ์ระหว่างความยาวน้อยสุดของรหัสค่ากับประสิทธิผล 31
รูปที่ 4.1 ลำดับการทำงานของระบบค้นคืนข้ามภาษา 33
รูปที่ 4.2 การประมวลผลตัวอักษรเบื้องต้น 35
รูปที่ 4.3 การเปรียบเทียบรหัสค่าที่นำเสนอ 45
รูปที่ 4.4 ความสัมพันธ์ระหว่างค่าแอลฟากับประสิทธิผลของระบบค้นคืน 59