

การปรับปรุงการเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ
เพื่อการค้นคืนข้ามภาษาโดยการตัดพยางค์ของรหัสเสียง



นาย โสภาส วงษ์ทวีทรัพย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2549
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

AN IMPROVEMENT OF THAI/ENGLISH TRANSLITERATED WORD ENCODING
FOR CROSS-LANGUAGE RETRIEVAL
BY SYLLABLE SEGMENTATION OF PHONETIC CODES

Mr. Opas Wongtaweessap

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Computer Science Program in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2006

Copyright of Chulalongkorn University

491744

นาย โอบาส วงษ์ทวีทรัพย์ : การปรับปรุงการเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษาโดยการตัดพยางค์ของรหัสเสียง. (AN IMPROVEMENT OF THAI/ENGLISH TRANSLITERATED WORD ENCODING FOR CROSS-LANGUAGE RETRIEVAL BY SYLLABLE SEGMENTATION OF PHONETIC CODES) อ. ที่ปรึกษา : รศ.ดร.บุญเสริม กิจศิริกุล, อ. ที่ปรึกษาร่วม : รศ.ดร.สมชาย ประสิทธิ์จตุระกุล, 64 หน้า.

วิทยานิพนธ์ฉบับนี้นำเสนอการค้นคืนข้ามภาษา สำหรับคำทับศัพท์ภาษาไทย/อังกฤษโดยได้ใช้วิธีการของนิรอลเน็ตเวิร์กในการเข้ารหัสคำ และใช้ขั้นตอนการตัดพยางค์ของรหัสเสียง วิธีการที่นำเสนอช่วยให้สามารถค้นคืนคำทับศัพท์ข้ามภาษาได้โดยไม่ต้องอาศัยพจนานุกรม

ในการค้นคืนข้ามภาษาโดยไม่อาศัยพจนานุกรมนั้นจำเป็นต้องใช้หลักการเข้ารหัสซึ่งเป็นสัญลักษณ์แทนเสียงอ่านของคำและประกอบด้วยรหัสเสียงของแต่ละตัวอักษรของคำมาเรียงต่อกัน ใน การที่จะทราบว่าตัวอักษรที่กำลังสนใจในคำนั้นให้รหัสเสียงใดจำเป็นต้องอาศัยการพิจารณาตัวอักษรข้างเคียงด้วย ดังนั้นการเข้ารหัสคำสามารถจัดได้ว่าเป็นปัญหาการจำแนกอย่างหนึ่ง ด้วยเหตุนี้จึงได้นำวิธีการนิรอลเน็ตเวิร์กมาใช้ในการเข้ารหัสคำ แต่เนื่องจากว่ารหัสคำของคำไทยและอังกฤษที่มีเสียงอ่านตรงกัน อาจมีความแตกต่างกันบ้าง จึงได้ใช้ขั้นตอนการเปรียบเทียบแบบประมาณสำหรับการค้นคืนคำที่มีเสียงอ่านคล้ายกันมากที่สุด จากผลการทดลองด้วยวิธี K-fold cross validation พบว่าเมื่อได้ปรับปรุงนิรอลเน็ตเวิร์ก สามารถให้ผลการค้นคืนในแบบที่ 1 ด้วยตัววัด F1 เป็น 83.28% สำหรับกรณี คำไทยทับศัพท์คำอังกฤษและให้ผลการค้นคืน F1 90.54% สำหรับคำอังกฤษทับศัพท์คำไทยที่ค่าความแตกต่างของรหัสเสียงเป็น 0

ภาควิชา วิศวกรรมคอมพิวเตอร์.....ลายมือชื่อนิสิต.....โอบาส
 สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์.....ลายมือชื่ออาจารย์ที่ปรึกษา.....
 ปีการศึกษา 2549.....ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....


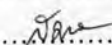
4670618021 : MAJOR COMPUTER SCIENCE

KEY WORD: TRANSLITERATED WORD / CROSS-LANGUAGE RETRIEVAL / NATURAL LANGUAGE PROCESSING / SYLLABLE SEGMENTATION OF PHONETIC CODES

OPAS WONGTAWESAP : AN IMPROVEMENT OF THAI/ENGLISH TRANSLITERATED WORD ENCODING FOR CROSS-LANGUAGE RETRIEVAL BY SYLLABLE SEGMENTATION OF PHONETIC CODES. THESIS ADVISOR : ASSOC. PROF. BOONSERM KIJSIRIKUL, Ph.D., THESIS COADVISOR : ASSOC. PROF.SOMCHAI PRASITJUTRAKUL, Ph.D., 64 pp.

This thesis presents Thai/English cross-language transliterated word retrieval by using neural networks and syllable segmentation of phonetic codes. The proposed method enables the transliterated word retrieval without using the dictionary.

Without dictionary, the phonetic code is employed for cross-language retrieval. The phonetic code of a word represents the sound of the word and it consists of a sequence of phonetic codes of characters in the word. In order to determine the code of a particular character, it is necessary to consider its surrounding characters. Hence this problem can be identified as a classification problem. For this reason, neural networks are used in phonetic encoding. However, as the codes generated from a pair of corresponding Thai/English words are sometimes slightly different, the approximate string matching is applied to determine of character editing. The experimental results, using K-fold cross validation, show that the F1-measure values are 83.28% for Thai/English cross-language transliterated and 90.54% for English/Thai cross-language transliterated with zero distance between phonetic codes.

Department Computer Engineering.....	Student's.....	Opas
Field of study Computer Science.....	Advisor's.....	
Academic year 2006.....	Co-advisor's.....	

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สามารถสำเร็จลุล่วงไปได้ด้วยดี ก็ด้วยความช่วยเหลืออย่างดียิ่งของ รองศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ และรองศาสตราจารย์ ดร.สมชาย ประสิทธิ์จตุระกุล อาจารย์ที่ปรึกษาร่วมวิทยานิพนธ์ โดยอาจารย์ทั้งสองท่านได้คอยชี้แนะ ให้คำแนะนำ และข้อคิดเห็นต่างๆ ในการดำเนินการวิจัยตลอดมา รวมทั้งได้ช่วยตรวจแก้ไข วิทยานิพนธ์ฉบับนี้อย่างละเอียด รวมทั้งคณะกรรมการสอบวิทยานิพนธ์ทั้งสองท่าน คือ รองศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา และ ผู้ช่วยศาสตราจารย์ ดร.ญาใจ ลิ้มปิยะกรณีย์ ที่ได้เสียสละเวลามาเป็นคณะกรรมการ ให้ข้อเสนอแนะ และแนวทางอันเป็นประโยชน์ยิ่งในการทำวิจัย

กราบขอบพระคุณอย่างสูงสำหรับ ผู้ช่วยศาสตราจารย์ ดร.ปราณี นิลกรณีย์ และ ผู้ช่วยศาสตราจารย์ ดร.จรุงแสง ลักษณะบุญส่ง คณบดีคณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร และ ขอขอบคุณศูนย์ปฏิบัติการวิจัย และพัฒนาระบบสารสนเทศอันชาญฉลาด (Intelligent Information Systems Development and Research Laboratory Centre) คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร ที่ได้ให้ความเอื้อเฟื้อเครื่องคอมพิวเตอร์ และสถานที่เพื่อใช้ในการ ดำเนินงานวิจัยของข้าพเจ้า

กราบขอบพระคุณบุพการีของข้าพเจ้า ทั้งคุณพ่อและคุณแม่ ลูกดีใจมากที่ได้เกิดมาใน ครอบครัวที่มีแต่ความรักและความอบอุ่นนี้ ลูกได้ทำทุกสิ่งทุกอย่าง ที่ไม่ทำให้พ่อกับแม่ผิดหวังแล้ว กราบขอบพระคุณคณาจารย์ทุกท่านที่ได้ประสิทธิ์ประสาทวิชา อบรมสั่งสอนและให้ความรู้มา ตั้งแต่อนุบาลจนถึงระดับบัณฑิตศึกษา เพื่อเป็นรากฐานอันแข็งแกร่งให้กับข้าพเจ้าในวันนี้

ขอบคุณความคิด ความฝันของข้าพเจ้า ที่เป็นเครื่องชี้นำและเป็นกำลังใจให้กับข้าพเจ้า เสมอมาตั้งแต่ในวัยเด็ก ว่าอย่ามัวแต่คิด อย่ามัวแต่เพื่อฝัน แต่ต้องทำสิ่งที่ตนเองคิดและฝันนั้นให้ สำเร็จลุล่วง เป็นจริงไปให้ได้

และสุดท้ายนี้ต้องขอบคุณเพื่อน ๆ พี่ ๆ และน้อง ๆ ในภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่เป็นกำลังใจและคอยเป็นห่วงช่วยเหลือกันเสมอ มา ตลอดระยะเวลาที่เรียน และตลอดระยะเวลาทำงาน จนสำเร็จการศึกษา

สารบัญ

	หน้า
บทคัดย่อวิทยานิพนธ์ภาษาไทย	ง
บทคัดย่อวิทยานิพนธ์ภาษาอังกฤษ	จ
กิตติกรรมประกาศ	ฉ
สารบัญ	ช
สารบัญตาราง	ญ
สารบัญภาพ	ฎ
บทที่ 1 บทนำ	1
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์ของการวิจัย	3
1.3 ขอบเขตของการวิจัย	3
1.4 ประโยชน์ที่คาดว่าจะได้รับ	3
1.5 วิธีดำเนินการวิจัย	3
1.6 ผลงานที่ตีพิมพ์จากงานวิจัย	3
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	5
2.1 หลักเกณฑ์การออกเสียง	5
2.2 ระบบเสียงในภาษาไทยและระบบเสียงในภาษาอังกฤษ	5
2.2.1 ระบบเสียงในภาษาไทย	6
2.2.2 ระบบเสียงในภาษาอังกฤษ	6
2.3 ความหมายของพยางค์	6
2.4 การถอดอักษร	7
2.5 การถ่ายเสียงด้วยตัวอักษรโรมัน	8
2.6 การตัดพยางค์	8
2.7 นิวรอลเน็ตเวิร์ก (Neural Networks)	9
2.8 การวัดผลการค้นคืน	12
2.9 ขั้นตอนวิธีระยะแก้ไขสั้นที่สุด (Minimum Edit Distance)	12
2.10 งานวิจัยของ วรณี อุดมพานิชย์	13
2.11 งานวิจัยของ นิลเนตร อรุณวงศ์ ณ อยุธยา	15
2.12 งานวิจัยของ ประยุทธ์ สุวรรณวิสาท และ สมชาย ประสิทธิ์จิตรระกุล	16

2.13 งานวิจัยของ ทศนวรรณ ศูนย์กลาง สมชาย ประสิทธิ์จตุระกุล และบุญเสริม กิจศิริกุล	17
2.14 สรุป	21
บทที่ 3 การเข้ารหัสคำ	22
3.1 รหัสคำ	22
3.2 การประมวลผลเบื้องต้น	25
3.3 วิธีการเข้ารหัสคำ และการเข้ารหัสคำด้วยนิรอลเน็ตเวิร์ก	25
3.4 วิธีการนับจำนวนพยางค์ของรหัสคำ และการแบ่งพยางค์ของรหัสคำ	32
3.5 สรุป	35
บทที่ 4 การค้นคืนข้ามภาษา	36
4.1 การคำนวณความต่างของรหัสคำ	36
4.2 เกณฑ์การเปรียบเทียบรหัสคำ	37
4.3 สรุป	39
บทที่ 5 การทดลอง	40
5.1 วิธีการทดลอง	40
5.2 การเข้ารหัสคำด้วยนิรอลเน็ตเวิร์ก	41
5.3 วิเคราะห์ผลการทดลองการเข้ารหัสคำ	44
5.4 ผลการทดลองจากการค้นคืน	45
5.5 วิเคราะห์ผลการทดลองการค้นคืน	50
5.6 สรุป	52
บทที่ 6 สรุปผลการวิจัยและข้อเสนอแนะ	53
6.1 สรุปผลการวิจัย	53
6.2 ข้อเสนอแนะ	53
รายการอ้างอิง	55
ภาคผนวก	57
ภาคผนวก ก การใช้อักษรโรมันแทนอักขระไทย	58
ภาคผนวก ข หน่วยเสียงในภาษาไทยและภาษาอังกฤษ	60
หน่วยเสียงในภาษาไทย	60
ระบบเสียงในภาษาอังกฤษ	60
ภาคผนวก ค ตัวอย่างข้อมูลคำทับศัพท์ที่ใช้ในงานวิจัย	62
ตัวอย่างคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ	62
ตัวอย่างคำไทยและคำอังกฤษทับศัพท์คำไทย	63

ประวัติผู้เขียนวิทยานิพนธ์..... 64

สารบัญตาราง

	หน้า
ตารางที่ 2.1 การกำหนดรหัสชาวเด็กซ์ภาษาอังกฤษของ Odell และ Russel	14
ตารางที่ 2.2 การกำหนดรหัสตัวอักษรของรหัสชาวเด็กซ์ภาษาไทย จากงานวิจัยของวรรณี อุดม พาณิชย์.....	14
ตารางที่ 2.3 การกำหนดรหัสตัวเลขของรหัสชาวเด็กซ์ภาษาไทย จากงานวิจัยของวรรณี อุดม พาณิชย์.....	15
ตารางที่ 2.4 การกำหนดรหัสสำหรับอักขระตัวแรก จากงานวิจัยของนิลเนตร อรุณวงศ์ ณ ออยุธยา	16
ตารางที่ 2.5 การกำหนดรหัสสำหรับอักขระถัดจากตัวแรก จากงานวิจัยของนิลเนตร อรุณวงศ์ ณ อยุธยา	16
ตารางที่ 2.6 การกำหนดรหัสสำหรับคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ จากงานวิจัยของ ประยูทธ สุวรรณวิสาทร.....	18
ตารางที่ 2.7 การกำหนดรหัสของพยัญชนะสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จาก งานวิจัยของ ประยูทธ สุวรรณวิสาทร.....	18
ตารางที่ 2.8 การกำหนดรหัสของสระสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัย ของ ประยูทธ สุวรรณวิสาทร.....	19
ตารางที่ 3.1 รหัสเสียงสำหรับคำไทยทับศัพท์คำอังกฤษแบบเดิม	23
ตารางที่ 3.2 รหัสเสียงพยัญชนะสำหรับคำอังกฤษทับศัพท์คำไทยแบบเดิม	24
ตารางที่ 3.3 สรุปจำนวนนิรอรลที่ใช้ในการทดลองในนิรอรลเน็ตเวิร์กแต่ละโครงสร้าง	29
ตารางที่ 3.4 รหัสเสียงสำหรับคำไทยทับศัพท์คำอังกฤษแบบใหม่	30
ตารางที่ 3.5 รหัสเสียงพยัญชนะสำหรับคำอังกฤษทับศัพท์คำไทยแบบใหม่.....	31
ตารางที่ 3.6 ความแม่นยำของการนับจำนวนพยางค์จากจำนวนรหัสเสียงสระ	34
ตารางที่ 3.7 ความแม่นยำของจำนวนพยางค์ที่ตรงกันในคำคู่กันของทั้งสองภาษา.....	35
ตารางที่ 5.1 จำนวนข้อมูลทั้งหมด และจำนวนข้อมูลในแต่ละชุดตามค่า K.....	40
ตารางที่ 5.2 ความแม่นยำเมื่อใช้นิรอรลในชั้นซ่อนต่างๆ กันสำหรับข้อมูลคำอังกฤษ.....	42
ตารางที่ 5.3 ความแม่นยำเมื่อใช้นิรอรลในชั้นซ่อนต่างๆ กัน สำหรับคำไทยทับศัพท์คำอังกฤษ ..	42
ตารางที่ 5.4 ความแม่นยำเมื่อใช้นิรอรลในชั้นซ่อนต่างๆ กัน สำหรับข้อมูลคำไทย.....	42
ตารางที่ 5.5 ความแม่นยำเมื่อใช้นิรอรลในชั้นซ่อนต่างๆ กันสำหรับคำอังกฤษทับศัพท์คำไทย ...	43

ตารางที่ 5.6 สรุปค่าความแม่นยำสูงสุดเมื่อใช้จำนวนนิรอรอนในชั้นซ้อนต่างๆ กัน สำหรับข้อมูลคำ อังกฤษ และคำไทยทับศัพท์คำอังกฤษเทียบกัน.....	43
ตารางที่ 5.7 สรุปค่าความแม่นยำสูงสุดเมื่อใช้จำนวนนิรอรอนในชั้นซ้อนต่างๆ กัน สำหรับข้อมูลคำ ไทย และคำอังกฤษทับศัพท์คำไทยเทียบกัน.....	43
ตารางที่ 5.8 สรุปผลการเข้ารหัสคำจากนิรอรลตัวที่ดีที่สุดเปรียบเทียบกันกับงานวิจัยก่อนหน้า..	44
ตารางที่ 5.9 ความแม่นยำเมื่อใช้นิรอรลในชั้นซ้อนต่างๆ กันสำหรับข้อมูลคำไทยทับศัพท์คำ อังกฤษ ที่มีการตัดหน่วยเสียงของข้อมูลขาเข้าที่ไม่ได้ใช้ออก.....	45
ตารางที่ 5.10 ความแม่นยำสูงสุดเมื่อใช้นิรอรลในชั้นซ้อนต่างๆ กันสำหรับข้อมูลคำไทยทับศัพท์ คำอังกฤษ ก่อนและหลังการตัดหน่วยเสียงที่ไม่ได้ใช้ออก.....	45
ตารางที่ 5.11 ผลการทดลองจากการค้นคืนในแบบที่ 1.....	46
ตารางที่ 5.12 ผลการทดลองจากการค้นคืนในแบบที่ 2.....	46
ตารางที่ 5.13 ผลการทดลองจากการค้นคืนในแบบที่ 3.....	47
ตารางที่ 5.14 ผลการทดลองจากการค้นคืนในแบบที่ 4.....	47
ตารางที่ 5.15 เปรียบเทียบผลการค้นคืนทั้ง 4 แบบ (d=0) ด้วยข้อมูลจากการเข้ารหัสด้วยมือ....	47
ตารางที่ 5.16 ผลการทดลองจากการค้นคืนในแบบที่ 1 ในกรณีคำไทยทับศัพท์คำอังกฤษ และ กรณีคำอังกฤษทับศัพท์คำไทยจากการเข้ารหัสด้วยนิรอรลเน็ตเวิร์ก.....	48
ตารางที่ 5.17 การเปรียบเทียบผลของตัววัด F1 จากการค้นคืนในแบบที่ 1.....	48
ตารางที่ 5.18 การเปรียบเทียบผลของตัววัด F1 จากการค้นคืนในแบบที่ 1.....	48
ตารางที่ 5.19 ผลการทดลองจากการค้นคืนในแบบที่ 2.....	49
ตารางที่ 5.20 ผลการทดลองจากการค้นคืนในแบบที่ 3.....	49
ตารางที่ 5.21 ผลการทดลองจากการค้นคืนในแบบที่ 4.....	49
ตารางที่ 5.22 ผลการค้นคืนด้วยตัววัด F1 สูงสุดของงานวิจัยนี้กับผลที่ได้จากวิธีการก่อนหน้า...	50
ตารางที่ 5.23 ผลการค้นคืนด้วยค่าความเที่ยงสูงสุดของงานวิจัยนี้กับผลที่ได้จากวิธีการก่อนหน้า	50
ตารางที่ 5.24 ความแม่นยำสำหรับคำอังกฤษทับศัพท์คำไทยด้วยการเข้ารหัสแบบใหม่.....	51
ตารางที่ 5.25 ความแม่นยำสำหรับคำอังกฤษด้วยการเข้ารหัสแบบใหม่.....	51
ตารางที่ 5.26 ความแม่นยำสำหรับคำไทยทับศัพท์คำอังกฤษด้วยการเข้ารหัสแบบใหม่.....	52

สารบัญภาพ

	หน้า
รูปที่ 2.1 นิเวศน์เน็ตเวิร์กที่มี 3 ชั้น.....	11
รูปที่ 2.2 ขั้นตอนวิธีการเรียนรู้แบบแพร่กระจายย้อนกลับ.....	11
รูปที่ 2.3 ตัวอย่างการกำหนดต้นทุนการแทนที่อักขระสำหรับพยัญชนะ จากงานวิจัยของประยุทธ สุวรรณวิสาทร.....	20
รูปที่ 2.4 ตัวอย่างการกำหนดต้นทุนการแทนที่อักขระสำหรับสระ จากงานวิจัยของประยุทธ สุวรรณวิสาทร.....	20
รูปที่ 2.5 ตัวอย่างการเข้ารหัสคำไทยของทัศนวรรณ ศูนย์กลาง และคณะ.....	21
รูปที่ 2.6 ตัวอย่างการเข้ารหัสคำอังกฤษของทัศนวรรณ ศูนย์กลาง และคณะ.....	21
รูปที่ 3.1 การเข้ารหัสคำโดยอาศัยการพิจารณาอักษรข้างเคียงสำหรับคำ "สุรเกียรติ".....	27
รูปที่ 3.2 การเข้ารหัสคำโดยอาศัยการพิจารณาอักษรข้างเคียงสำหรับคำ "surakiat".....	28
รูปที่ 3.3 โครงสร้างของนิเวศน์เน็ตเวิร์กและการเข้ารหัสคำด้วยแบ็กพรอพาเกชันนิเวศน์เน็ตเวิร์ก	28
รูปที่ 3.4 การนับจำนวนพยางค์จากการหาค่าเสียงสระของรหัสคำคำว่า "สุรเกียรติ".....	33
รูปที่ 3.5 การนับจำนวนพยางค์จากการหาค่าเสียงสระของรหัสคำคำว่า "surakiat".....	33