การปรับปรุงการเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษาโดยการตัดพยางค์ของรหัสเสียง

นาย โอภาส วงษ์ทวีทรัพย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2549

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

Mr. Opas Wongtaweesap

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

หัวข้อวิทยานิพนธ์	การปรับปรุงการเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษาโดยการตัดพยางค์ของรหัสเสียง
โดย	นาย โอภาส วงษ์ทวีทรัพย์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษา	รองศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล
อาจารย์ที่ปรึกษาร่วม	รองศาสตราจารย์ ดร.สมชาย ประสิทธิ์จูตระกูล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์

(ศาสตราจารย์ ดร.ดิเรก ลาวัณย์ศิริ)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

..... อาจารย์ที่ปรึกษา

(รองศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษาร่วม

(รองศาสตราจารย์ ดร.สมชาย ประสิทธิ์จูตระกูล)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ญาใจ ลิ่มปิยะกรณ์)

นาย โอภาส 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 ประสิทธิ์จูตระกูล, 64 หน้า.

วิทยานิพนธ์ฉบับนี้นำเสนอการค้นคืนข้ามภาษา สำหรับคำทับศัพท์ภาษาไทย/อังกฤษโดยได้ใช้วิธีการของนิวรอลเน็ตเวิร์กในการเข้ารหัสคำ และใช้ขั้นตอนการตัดพยางค์ของรหัสเสียง วิธีการที่นำเสนอช่วยให้สามารถค้นคืนคำทับศัพท์ข้ามภาษาได้โดยไม่ต้องอาศัยพจนานุกรม

ในการค้นคืนข้ามภาษาโดยไม่อาศัยพจนานุกรมนั้นจำเป็นต้องใช้หลักการเข้ารหัสซึ่งเป็นสัญลักษณ์แทนเสียงอ่านของคำและประกอบด้วยรหัสเสียงของแต่ละตัวอักษรของคำมาเรียงต่อกัน ในการที่จะทราบว่าตัวอักษรที่กำลังสนใจในคำนั้นให้รหัสเสียงใดจำเป็นต้องอาศัยการพิจารณาตัวอักษรข้างเคียงด้วย ดังนั้นการเข้ารหัสคำสามารถจัดได้ว่าเป็นปัญหาการจำแนกอย่างหนึ่ง ด้วยเหตุนี้จึงได้นำวิธีการนิวรอลเน็ตเวิร์กมาใช้ในการเข้ารหัสคำ แต่เนื่องจากว่ารหัสคำของคำไทยและอังกฤษที่มีเสียงอ่านตรงกัน อาจมีความแตกต่างกันบ้าง จึงได้ใช้ขั้นตอนการเปรียบเทียบแบบประมาณสำหรับการค้นคืนคำที่มีเสียงอ่านคล้ายกันมากที่สุด จากผลการทดลองด้วยวิธี K-fold cross validation พบว่าเมื่อได้ปรับปรุงนิวรอลเน็ตเวิร์ก สามารถให้ผลการค้นคืนในแบบที่ 1 ด้วยตัววัด F1 เป็น 83.28% สำหรับกรณีคำไทยทับศัพท์คำอังกฤษและให้ผลการค้นคืน F1 90.54% สำหรับคำอังกฤษทับศัพท์คำไทยที่ค่าความแตกต่างของรหัสเสียงเป็น 0

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต.....

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 WONGTAWEESAP : 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 translitarated 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 ก็ด้วยความช่วยเหลืออย่างดียิ่งของรองศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ และรองศาสตราจารย์ ดร.สมชาย ประสิทธิ์จูตระกูล อาจารย์ที่ปรึกษาร่วมวิทยานิพนธ์ โดยอาจารย์ทั้งสองท่านได้คอยชี้แนะให้คำแนะนำ และข้อคิดเห็นต่างๆ ในการดำเนินการวิจัยตลอดมา รวมทั้งได้ช่วยตรวจแก้ไขวิทยานิพนธ์ฉบับนี้อย่างละเอียด รวมทั้งคณะกรรมการสอบวิทยานิพนธ์ทั้งสองท่าน คือ รองศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา และ ผู้ช่วยศาสตราจารย์ ดร.ญาใจ ลิ่มปิยะกรณ์ ที่ได้เสียสละเวลามาเป็นคณะกรรมการ ให้ข้อเสนอแนะ และแนวทางอันเป็นประโยชน์ยิ่งในการทำวิจัย

กราบขอบพระคุณอย่างสูงสำหรับ ผู้ช่วยศาสตราจารย์ ดร.ปราณี นิลกรณ์ และ ผู้ช่วยศาสตราจารย์ ดร.จรุงแสง ลักษณบุญส่ง คณบดีคณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร และขอขอบคุณศูนย์ปฏิบัติการวิจัย และพัฒนาระบบสารสนเทศอันชาญฉลาด (Intelligent Information Systems Development and Research Laboratory Centre) คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร ที่ได้ให้ความเอื้อเฟื้อเครื่องคอมพิวเตอร์ และสถานที่เพื่อใช้ในการดำเนินงานวิจัยของข้าพเจ้า

กราบขอบพระคุณบุพการีของข้าพเจ้า ทั้งคุณพ่อและคุณแม่ ลูกดีใจมากที่ได้เกิดมาในครอบครัวที่มีแต่ความรักและความอบอุ่นนี้ ลูกได้ทำทุกสิ่งทุกอย่าง ที่ไม่ทำให้พ่อกับแม่ผิดหวังแล้ว กราบขอบพระคุณคณาจารย์ทุกท่านที่ได้ประสิทธิ์ประสาทวิชา อบรมสั่งสอนและให้ความรู้มาตั้งแต่อนุบาลจนถึงระดับบัณฑิตศึกษา เพื่อเป็นรากฐานอันแข็งแกร่งให้กับข้าพเจ้าในวันนี้

ขอบคุณความคิด ความฝันของข้าพเจ้า ที่เป็นเครื่องชี้นำและเป็นกำลังใจให้กับข้าพเจ้าเสมอมาตั้งแต่ในวัยเด็ก ว่าอย่ามัวแต่คิด อย่ามัวแต่เฟ้อฝัน แต่ต้องทำสิ่งที่ตนเองคิดและฝันนั้นให้สำเร็จลุล่วง เป็นจริงไปให้ได้

และสุดท้ายนี้ต้องขอบคุณเพื่อน ๆ พี่ ๆ และน้อง ๆ ในภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่เป็นกำลังใจและคอยเป็นห่วงช่วยเหลือกันเสมอมา ตลอดระยะเวลาที่เรียน และตลอดระยะเวลาทำงาน จนสำเร็จการศึกษา

สารบัญ

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า