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

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในสถานการณ์ของโลกยุคปัจจุบัน คอมพิวเตอร์ได้เข้ามามีบทบาท และส่งผลกระทบต่อวิถีการดำรงชีวิตของมนุษย์มากขึ้น ดังจะเห็นได้จากจำนวนของผู้ใช้งานคอมพิวเตอร์ที่สูงมากขึ้นเรื่อยๆ ซึ่งการเพิ่มขึ้นของจำนวนผู้ใช้งานดังกล่าวนั้น ก็เป็นสาเหตุหนึ่งของการเพิ่มขึ้นของสารสนเทศในรูปแบบต่างๆ มากมาย ทำให้การแลกเปลี่ยนข้อมูลของมนุษย์ในยุคสมัยนี้เกิดขึ้นกับกองสารสนเทศจำนวนมาก

ด้วยสาเหตุดังกล่าวจึงจำเป็นอย่างยิ่งที่จะต้องมีการพัฒนาระบบค้นคืนสารสนเทศ (Information Retrieval Systems) ขึ้นใช้งาน เพื่อสร้างระบบที่สามารถจัดการกับกองสารสนเทศเหล่านั้นได้อย่างมีประสิทธิภาพ ซึ่งในปัจจุบันก็มีระบบค้นคืนสารสนเทศให้เลือกใช้งานเป็นจำนวนมาก แต่จะแตกต่างกันตรงที่ ความสามารถที่จำเพาะเจาะจงลงไปในแต่ละระบบ ที่จะสร้างขึ้นให้เป็นจุดเด่นของระบบของตนเอง แต่ทุกระบบก็จะมุ่งเน้นไปที่สารสนเทศในหลากหลายรูปแบบ ที่ตัวระบบนั้นจะต้องมีความสามารถในการค้นคืนออกมาให้ได้มากที่สุด ทำให้ผลลัพธ์ที่ได้จากการค้นคืนของระบบนั้น พบว่ามีทั้งสารสนเทศที่เกี่ยวข้อง และที่ไม่เกี่ยวข้องกับความต้องการของผู้ใช้งานสารสนเทศนั้น

ปัจจุบันเอกสารทางวิชาการในประเทศไทย มักจะจัดทำทั้งในรูปแบบของภาษาไทย และในรูปแบบของภาษาอังกฤษ เพื่อประโยชน์ในการเผยแพร่ทั้งภายในและภายนอกประเทศ ซึ่งเอกสารเหล่านี้โดยเฉพาะอย่างยิ่งเอกสารทางด้านวิทยาศาสตร์และวิศวกรรมศาสตร์โดยมากแล้วมักจะปรากฏคำนามเฉพาะ (Proper Noun) และคำศัพท์เทคนิคต่างๆ เป็นจำนวนมาก ซึ่งจะพบได้ทั้งในรูปแบบของคำในภาษาอังกฤษ คำภาษาไทยทับศัพท์คำภาษาอังกฤษ คำภาษาอังกฤษทับศัพท์คำภาษาไทย หรือคำในภาษาไทยเอง ดังนั้นถ้าระบบค้นคืนสารสนเทศไม่สนับสนุนการทำงานข้ามภาษา ก็จะทำให้ประสิทธิภาพในการค้นคืนนั้นลดต่ำลง และเป็นการใช้ประโยชน์จากสารสนเทศที่มีอยู่ในเครือข่ายได้อย่างไม่เต็มที่

การค้นคืนสารสนเทศข้ามภาษา (Cross-Language Information Retrieval) หมายถึงการค้นคืนสารสนเทศซึ่งภาษาที่แสดงในเอกสารไม่ตรงกับภาษาที่แสดงในการสอบถาม [1]

ปัญหาในระบบค้นคืนสารสนเทศมีอยู่หลายประการด้วยกัน โดยเฉพาะอย่างยิ่งปัญหาในเรื่องของการค้นคืนข้ามภาษา ซึ่งพบว่าคำในภาษาหนึ่งอาจจะถูกเขียนในอีกภาษาหนึ่งได้หลาย

รูปแบบ ตัวอย่างเช่น “Carbohydrate” ในภาษาไทยอาจพบได้หลายแบบทั้ง “คาร์โบไฮเดรต” “คาร์โบไฮเดรท” หรือ “คาร์โบฮัยเดรต” หรือชื่อเฉพาะในภาษาไทย เช่น “โอภาส” อาจปรากฏในเอกสารที่ใช้เผยแพร่ในต่างประเทศในรูปของ “Ophas” หรือ “Opas” เป็นต้น ซึ่งระบบค้นคืนสารสนเทศควรจะมีความสามารถในการค้นคืนเอกสารที่มีคำเหล่านี้ปรากฏอยู่ออกมาให้ได้ทั้งหมดหรือให้ได้มากที่สุด และถึงแม้ว่าจะมีการนำพจนานุกรมสองภาษา (Bilingual Dictionary) มาใช้ในระบบค้นคืนสารสนเทศก็ไม่อาจจะแก้ปัญหานี้ได้มากนัก เนื่องจากมีคำศัพท์เทคนิคและคำทับศัพท์ใหม่ ๆ เกิดขึ้นมากมายในหลากหลายสาขาแทบทุกวัน และคำศัพท์ใหม่ ๆ เหล่านี้ส่วนมากมักจะไม่ปรากฏพบในพจนานุกรม [2] ดังนั้นจึงทำให้การสอบถามด้วยคำหลักในภาษาหนึ่ง อาจจะทำให้พลาดเอกสารที่มีคำหลักที่ตรงกันในอีกภาษาหนึ่งได้

ด้วยเหตุนี้ นักวิจัยจำนวนมากจึงได้หันมาสนใจกับปัญหาของการค้นคืนข้ามภาษา โดยมุ่งเน้นแก้ปัญหาด้วยการใช้การเข้ารหัสคำเพื่อนำมาช่วยในการค้นคืนข้ามภาษา โดยรหัสคำดังกล่าวนั้นจะเป็นสัญลักษณ์ที่แทนเสียงอ่านของคำ เพราะเนื่องจากคำที่อ่านออกเสียงเหมือนกันจะมีรหัสคำที่ตรงกันหรือใกล้เคียงกัน งานวิจัยต่างๆ ที่เกี่ยวข้องกับการเข้ารหัสคำและการค้นคืนข้ามภาษาไทย-ภาษาอังกฤษ ได้แก่ [3] [4] และ [5]

เนื่องจากรหัสคำที่ได้จากขั้นตอนของการเข้ารหัสคำนั้น อาจมีรหัสคำไม่เหมือนกันทุกตัวอักษร ถึงแม้ว่ารหัสคำนั้นจะเป็นรหัสคำที่อ่านออกเสียงตรงกันจากทั้งสองภาษา ดังนั้นในงานวิจัยดังกล่าวที่ผ่านมาข้างต้น จึงได้นำเสนอวิธีในการเปรียบเทียบรหัสคำ ซึ่งต้องใช้วิธีการเปรียบเทียบแบบประมาณ (Approximate Matching) ด้วยเทคนิคระยะแก้ไขสั้นสุด (Minimum Edit Distance) [6] ซึ่งเป็นการคำนวณความคล้ายคลึงกันระหว่างสายอักขระ 2 สาย โดยคำนวณจากต้นทุน (Cost) ที่ใช้ในการเปลี่ยนสายอักขระหนึ่งไปเป็นอีกสายอักขระหนึ่ง ด้วยการเพิ่ม ลบ หรือแทนที่อักขระ ซึ่งในขั้นตอนของการเปรียบเทียบรหัสคำนั้น ในงานวิจัยก่อนหน้าจะทำการแยกส่วนของการเปรียบเทียบออกเป็น 2 ส่วนจากกัน คือ ส่วนของพยัญชนะ และส่วนของสระ ตัวอย่างเช่น “kanokpiravut_” ซึ่งเป็นรหัสคำที่ได้จากการเข้ารหัสจากคำว่า “กนกพีระวุฒิ” ซึ่งในงานวิจัยก่อนหน้าจะทำการสลับตำแหน่งของตัวอักษรที่เป็นสัญลักษณ์แทนเสียงพยัญชนะ และเสียงสระออกจากกัน ด้วยการเลื่อนสัญลักษณ์ที่แทนเสียงของสระไปไว้ด้านหลังกลายเป็น “knkprvt, aoiau” แล้วแยกการเปรียบเทียบในแต่ละส่วนออกจากกัน

แต่ในงานวิจัยนี้ได้มีการปรับเปลี่ยนวิธีการดังกล่าว โดยจะทำการเข้ารหัสใหม่เพื่อแยกรหัสเสียงของเสียงพยัญชนะต้น เสียงสระ และเสียงของพยัญชนะที่เป็นตัวสะกดออกจากกัน เพื่อทำให้การเข้ารหัสคำนั้นสามารถแบ่งแยกได้อย่างชัดเจนว่าเป็นรหัสคำในส่วนไหน เพื่อให้การค้นคืนนั้นสามารถทำการเปรียบเทียบถูกต้องมากยิ่งขึ้น และการวิจัยนี้มีข้อสมมุติฐานว่าขั้นตอนที่

นำเสนอนั้น จะสามารถทำการสืบค้นคำทับศัพท์ข้ามภาษาไทย-ภาษาอังกฤษ ได้โดยไม่จำเป็นต้องใช้พจนานุกรม

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อปรับปรุงการเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษาโดยการตัดพยางค์ของรหัสเสียง

1.3 ขอบเขตของการวิจัย

1. คำทับศัพท์ที่ใช้เป็นคำทับศัพท์ระหว่างภาษาไทยและภาษาอังกฤษเท่านั้น
2. คำศัพท์ในภาษาอังกฤษที่ใช้ไม่รวมถึงคำย่อ (Abbreviation) และคำรัสพจน์ (Acronym)ยึดหลักเกณฑ์การออกเสียงของคำตามหลักของราชบัณฑิตยสถาน

1.4 ประโยชน์ที่คาดว่าจะได้รับ

สามารถนำไปใช้ในระบบการสืบค้นสารสนเทศ ให้มีสามารถในค้นคืนข้ามภาษาไทย และภาษาอังกฤษได้โดยไม่ต้องอาศัยพจนานุกรม และสามารถรองรับคำที่ศัพท์ที่เกิดขึ้นใหม่ได้ รวมทั้งสามารถใช้เป็นแนวทางในการสร้างระบบการค้นคืนข้ามภาษาในภาษาอื่นหรือการค้นคืนด้วยวิธีการที่ดียิ่งขึ้นได้

1.5 วิธีดำเนินการวิจัย

1. ศึกษาทฤษฎีพื้นฐานทางด้านภาษาศาสตร์ โดยเฉพาะอย่างยิ่งเรื่องของการตัดพยางค์
2. ศึกษาวิธีการเข้ารหัสคำโดยใช้นิวรอลเน็ตเวิร์ก
3. รวบรวมและจัดเก็บชุดข้อมูลคำทับศัพท์เพื่อใช้ในการทดลอง และกำหนดรหัสคำของแต่ละคำศัพท์
4. แปลงข้อมูลคำศัพท์ให้อยู่ในรูปที่ใช้สำหรับฝึกสอนนิวรอลเน็ตเวิร์ก
5. ฝึกสอนและทดสอบนิวรอลเน็ตเวิร์ก เพื่อใช้เป็นตัวสร้างรหัสคำ
6. ออกแบบขั้นตอนวิธี และสร้างลำดับของกฎที่จะนำมาใช้ในการแยกกลุ่มของรหัสคำให้เป็นรหัสของพยางค์ โดยใช้ความรู้ทางด้านภาษาศาสตร์ และการตัดพยางค์
7. ทำการทดลองและปรับปรุงผลการทดลอง
8. สรุปผลการทดลอง และจัดทำวิทยานิพนธ์

1.6 ผลงานที่ตีพิมพ์จากงานวิจัย

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้ตีพิมพ์และนำเสนอในงานประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ 2549 (The National Computer Science and Engineering Conference: NCSEC'06) เมื่อวันที่ 25-27 ตุลาคม พ.ศ. 2549 ในบทความเรื่อง

“การเข้ารหัสและนับจำนวนพยางค์ของรหัสคำด้วยแบ็คพรอพาเกชั่นนิวรอลเน็ตเวิร์ก” (Word Encoding and Syllable Enumeration of Phonetic Codes Using by A Back-Propagation Neural Networks) [7] โดยผู้นำเสนอคือ โอภาส วงษ์ทวีทรัพย์ บุญเสริม กิจศิริกุล และ สมชาย ประสิทธิ์จูตระกูล

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้ได้นำความรู้ทางด้านภาษาศาสตร์ รวมไปถึงงานวิจัยที่เกี่ยวข้องกับวิธีการเข้ารหัสเสียง และโดยเฉพาะอย่างยิ่งวิธีการตัดพยางค์ เพื่อนำมาปรับใช้ในงานวิจัยนี้ โดยมีรายละเอียดดังต่อไปนี้

2.1 หลักเกณฑ์การออกเสียง

ระบบเสียงในทุกภาษาย่อมประกอบด้วยหน่วยเสียง (Phoneme) [8] จำนวนหนึ่งหน่วยเสียงนี้หมายถึง เสียงสำคัญในภาษาที่มีหน้าที่แยกความหมายของคำในภาษา หน่วยเสียงยังสามารถแบ่งประเภทได้เป็น หน่วยเสียงพยัญชนะ หน่วยเสียงสระ และสำหรับภาษาไทยนั้น จะมีหน่วยเสียงวรรณยุกต์ด้วย และเพื่อให้การพิจารณาศึกษาเรื่องเสียงที่ปรากฏในภาษาพูดได้สะดวก นักภาษาศาสตร์จึงได้กำหนดอักษรแทนเสียงขึ้นชุดหนึ่งเรียกว่า สัทอักษร (Phonetic Alphabet)

สัทอักษร คือ อักษรที่ใช้เฉพาะในทางสัทศาสตร์ กำหนดขึ้นเพื่อแสดงลักษณะในการออกเสียงของภาษาอย่างเป็นสากล จึงไม่มุ่งแสดงลักษณะเฉพาะของภาษาใดภาษาหนึ่ง โดยมีสมาคมสัทศาสตร์ระหว่างชาติ (the International Phonetic Association - IPA) เป็นผู้กำหนดสัทอักษรขึ้นใช้แทนเสียงพูดของมนุษย์ทั่วโลก สัทอักษรนี้ส่วนใหญ่เป็นอักษรโรมัน และมีเครื่องหมายประกอบบ้าง เพื่อให้มีอักษรเพียงพอที่จะบันทึกเสียงของภาษาต่างๆ ได้ทั่วโลก และแม้ว่าหน่วยเสียงแต่ละหน่วยในทุก ๆ ภาษา แต่ละคนอาจออกเสียงแตกต่างกันไปเล็กน้อย แต่นักภาษาศาสตร์ยังคงถือว่าเป็นหน่วยเสียงเดียวกัน เมื่อนักภาษาศาสตร์จะทำการวิเคราะห์ภาษาใดก็สามารถเลือกสัทอักษรที่มีลักษณะของเสียงตรงกันไปใช้ สัทอักษรที่ใช้สำหรับภาษาไทยก็นำมาจากสัทอักษรของ IPA นี้เช่นกัน

2.2 ระบบเสียงในภาษาไทยและระบบเสียงในภาษาอังกฤษ

ระบบเสียงในภาษาไทย และระบบเสียงในภาษาอังกฤษ [9] นั้นจะมีความแตกต่างกันบ้าง เนื่องจากจำนวนของหน่วยเสียงพื้นฐานที่ไม่เท่ากัน รวมไปถึงระบบเสียงในภาษาอังกฤษนั้นไม่มีหน่วยเสียงวรรณยุกต์เหมือนในระบบเสียงภาษาไทย แต่เนื่องจากในงานวิจัยนี้ ต้องการจะเข้ารหัสให้เป็นหน่วยเสียงที่เหมือนกัน หรือใกล้เคียงกันมากที่สุด จึงจำเป็นต้องมีกระบวนการในการประมวลผลตัวอักษรเบื้องต้น สำหรับกรณีของคำภาษาไทยด้วย

2.2.1 ระบบเสียงในภาษาไทย

หน่วยเสียงภาษาไทย มี 47 หน่วยเสียง ดังนี้

1.**หน่วยเสียงพยัญชนะ** มี 21 หน่วยเสียง ซึ่งหน่วยเสียงพยัญชนะทั้ง 21 หน่วยเสียงมีการปรากฏในพยางค์ดังนี้

- เป็นพยัญชนะต้นได้ 21 หน่วยเสียง
- เป็นพยัญชนะต้นควบได้ 9 หน่วยเสียง เรียงต่อกันได้ 12 แบบ
- เป็นพยัญชนะท้าย หรือตัวสะกดได้ 9 หน่วยเสียง

2.**หน่วยเสียงสระ** มี 21 หน่วยเสียง ซึ่งจำแนกเป็น 2 ชนิด ดังนี้

- สระเดี่ยว มี 18 หน่วยเสียง
- สระประสม มี 3 หน่วยเสียง

3.**หน่วยเสียงวรรณยุกต์** มี 5 หน่วยเสียง ดังนี้

- หน่วยเสียงวรรณยุกต์สามัญ
- หน่วยเสียงวรรณยุกต์เอก
- หน่วยเสียงวรรณยุกต์โท
- หน่วยเสียงวรรณยุกต์ตรี
- หน่วยเสียงวรรณยุกต์จัตวา

2.2.2 ระบบเสียงในภาษาอังกฤษ

หน่วยเสียงภาษาอังกฤษ มี 44 หน่วยเสียง ดังนี้

1.**หน่วยเสียงพยัญชนะ** มี 24 หน่วยเสียง

2.**หน่วยเสียงสระ** มี 20 หน่วยเสียง

2.3 ความหมายของพยางค์

พยางค์ในภาษาไทยนั้นเกิดจากการเปล่งเสียงพยัญชนะ เสียงสระ และเสียงวรรณยุกต์ตามกันออกมาอย่างกระชั้นชิด จนฟังดูเหมือนกับเปล่งเสียงออกมาในครั้งเดียวกัน ซึ่งเรียกว่า การประสมเสียงในภาษา เสียงที่เกิดจากการประสมเสียง จึงเรียกว่า พยางค์ แต่ในภาษาอังกฤษนั้น พยางค์ (Syllable) นั้นจะแตกต่างกับพยางค์ในภาษาไทย ตรงที่ในภาษาอังกฤษนั้นไม่มีหน่วยเสียงของวรรณยุกต์

พระยาอุปกิตศิลปสาร [10] กล่าวว่า “ถ้อยคำที่เราใช้พูดกันนั้น บางทีก็เปล่งเสียงออกครั้งเดียว บางทีก็หลายครั้ง เสียงที่เปล่งออกมาครั้งหนึ่ง ๆ นั้น ท่านเรียกว่า “พยางค์” คือ ส่วนของคำพูด”

กาญจนา นาคสกุล [11] กล่าวว่า “พยางค์ จึงหมายถึง จำนวนเสียงที่ดังเด่น ซึ่งปรากฏในกลุ่มเสียงที่เรียงเป็นคำพูด เสียงอื่น ๆ ที่อยู่ข้างเคียงก็จะประกอบเข้าเป็นส่วนของพยางค์ โดยปกติเสียงสระจะเป็นเสียงที่มีลักษณะประจำตัว เป็นเสียงก้องที่ดังกว่าเสียงอื่น ฉะนั้นเสียงสระจึงมักจะเป็นเสียงที่ทำให้เกิดพยางค์”

จากคำอธิบายความหมายของพยางค์ดังกล่าว สรุปได้ว่าพยางค์คือ เสียงที่เปล่งออกมาครั้งหนึ่ง ๆ ซึ่งมีเสียงสระเป็นเสียงที่ดังเด่นอยู่จำนวน 1 เสียง และเสียงที่อยู่ข้างเคียงอีกอย่างน้อย 2 เสียง ได้แก่ เสียงของพยัญชนะ และเสียงของวรรณยุกต์สำหรับระบบเสียงภาษาไทย (ในภาษาอังกฤษจะไม่มีหน่วยเสียงวรรณยุกต์) ซึ่งเสียงสระนี้จะเป็นเสียงที่ทำให้เกิดพยางค์ขึ้นในภาษา พยางค์อาจจะเป็นคำก็ได้ถ้าพยางค์นั้นมีความหมาย เช่น

นา	เป็น 1 พยางค์ 1 คำ
สาหร่าย	เป็น 2 พยางค์ 1 คำ
อภิเษก	เป็น 3 พยางค์ 1 คำ

ซึ่งในงานวิจัยนี้ได้ใช้นิยามความหมายของพยางค์ เพื่อจำแนกและตัดพยางค์ เนื่องจากถ้าเราทราบจำนวนของสระที่ปรากฏอยู่ในคำใดๆ แล้ว เราจะสามารถทราบได้ถึงจำนวนของพยางค์ ซึ่งจำนวนของพยางค์จะเป็นตัวบ่งบอกถึงการจะแบ่งคำว่าจะสามารถแบ่งคำได้ออกเป็นกี่พยางค์

2.4 การถอดอักษร

การถอดอักษร (Transliteration) หมายถึง การนำคำในภาษาหนึ่งมาเขียนด้วยตัวอักษรอีกภาษาหนึ่งแบบอักษรต่ออักษร โดยพยายามใช้หน่วยเสียงของอักษรทั้งสองภาษาใกล้เคียงกันมากที่สุด [12] ตัวอย่างเช่น คำว่า “OXFORD” ในภาษาอังกฤษถอดอักษรเป็น “ออกซ์ฟอร์ด” ในภาษาไทย เป็นต้น การถอดอักษรแบ่งเป็น 3 ขั้นตอนหลัก ๆ ดังนี้

1. ถอดหน่วยอักษรในภาษาต้นแบบ (Source Language) เป็นหน่วยเสียงในภาษาต้นแบบ เช่น ถอดหน่วยอักษร “B” ในภาษาอังกฤษเป็นหน่วยเสียง /b/ ในภาษาอังกฤษ เป็นต้น
2. แทนหน่วยเสียงในภาษาต้นแบบ ด้วยหน่วยเสียงในภาษาเป้าหมาย (Target Language) โดยพยายามใช้หน่วยเสียงที่ใกล้เคียงกันมากที่สุด เช่น แทนหน่วยเสียง /b/ ในภาษาอังกฤษเป็นหน่วยเสียง /b/ ในภาษาไทย เป็นต้น
3. ถอดหน่วยเสียงในภาษาเป้าหมาย เป็นหน่วยอักษรในภาษาเป้าหมาย เช่น ถอดหน่วยเสียง /b/ ในภาษาไทยเป็นหน่วยอักษร “บ” ในภาษาไทย เป็นต้น

ปัญหาต่าง ๆ ในการถอดอักษรได้แก่

1. ความสัมพันธ์ของหน่วยอักษรและหน่วยเสียง มีความสัมพันธ์แบบหนึ่งตัวอักษรแทนหลายหน่วยเสียง เช่น ในภาษาอังกฤษ “C” แทนด้วย /k/ หรือ /s/ เป็นต้น และมี

ความสัมพันธ์แบบหลายหน่วยอักษรแทนหนึ่งหน่วยเสียง เช่น ในภาษาอังกฤษ “N, TN, GN, PN” แทนด้วย /n/ ในภาษาไทย “ร, ฤ, หร” แทนด้วย /r/ และ “ฉ, ช, ฌ” แทนด้วย /ch/ เป็นต้น

2. การแบ่งพยางค์ในภาษาต้นแบบ เมื่อมีพยัญชนะตัวเดียวอยู่ระหว่างสระ เช่น คำว่า money ในภาษาอังกฤษ จะแบ่งพยางค์อย่างไร จะถอดพยัญชนะซ้ำสองตัวเพื่อให้อ่านได้สะดวกเป็น มัน-นีย์ หรือจะถอดอักษรเพียงตัวเดียวตามที่ปรากฏในภาษาอังกฤษเป็น มะ-นีย์ หรือ มัน-อีย์
3. ปัญหาอันเนื่องมาจากช่วงเวลาของการยืมคำทับศัพท์ คำทับศัพท์บางคำยืมมาเป็นเวลานาน ซึ่งในอดีตมีหลักเกณฑ์การทับศัพท์ไม่ตรงกับหลักเกณฑ์ในปัจจุบัน เช่น “C” ที่แทน /k/ ในอดีตนิยมถอดเป็นอักษร “ก” เช่น กุ๊ก (Cook) กัปตัน (Captain) กะรัต(Carat) แก๊ป(Cap) เป็นต้น แต่ปัจจุบัน “C” ที่แทน /k/ มักจะถอดเป็น “ค” ในตำแหน่งพยัญชนะต้น เช่น คอนโดมิเนียม (Condominium) แคปซูล (Capsule) แคร็อต (Carrot) เป็นต้น

2.5 การถ่ายเสียงด้วยตัวอักษรโรมัน

การถ่ายเสียงด้วยตัวอักษรโรมัน (Romanization) คือการถ่ายเสียงตัวอักษรของภาษาอื่นที่ไม่ใช่อักษรโรมัน เช่น ไทย จีน ญี่ปุ่น ฯลฯ ให้เป็นตัวอักษรโรมัน เพื่อให้ผู้ที่ไม่รู้จักภาษานั้น ๆ สามารถอ่านออกเสียงได้ ทางราชบัณฑิตยสถานจึงได้กำหนดระบบการใช้ตัวอักษรโรมันเพื่อการถ่ายเสียงสำหรับตัวอักษรไทยออกเป็น 2 ระบบ คือระบบทั่วไปและระบบพิสดาร โดยระบบทั่วไปจะใช้สำหรับกรณีที่การออกเสียงสำคัญกว่าการเขียนตัวสะกด ซึ่งจะอาศัยหลักการออกเสียงเป็นสำคัญ ต้องสอดคล้องกับไวยากรณ์ของไทย และสามารถขยายเป็นระบบเฉพาะได้ เช่น คำว่า “กษัตริย์” ถ่ายเสียงเป็น Kasat ส่วนระบบพิสดารจะใช้ในกรณีที่จะแสดงตัวอักษรให้ละเอียดแม่นยำ เพื่อให้คงความหมายของคำนั้นไว้ เช่น คำว่า “กษัตริย์” ถ่ายเสียงเป็น Kasatriy

2.6 การตัดพยางค์

การตัดพยางค์ [13] นั้นจะกระทำโดยการหาขอบเขตความยาวของแต่ละพยางค์ที่ปรากฏในคำนั้นๆ ซึ่งจะแบ่งการพิจารณาออกเป็น 2 ชนิด คือ

- **การหาขอบเขตหน้า** (Front boundary) จะใช้สำหรับการกำหนดการเริ่มต้นของพยางค์
- **การหาขอบเขตหลัง** (Tail boundary) จะใช้สำหรับการกำหนดตำแหน่งท้ายสุดของพยางค์

เนื่องจากการเขียนประโยคภาษาไทย เป็นการเขียนคำต่อเนื่องโดยไม่มีการเว้นวรรคช่องว่างระหว่างคำ ดังนั้นจำเป็นต้องมีการตัดพยางค์ก่อน ซึ่งวิธีการตัดพยางค์นั้นจะใช้การอ้างอิงกับกฎที่สร้างขึ้นจากโครงสร้างของแต่ละพยางค์ที่เป็นไปได้จากหลักภาษาไทยที่ถูกต้อง ทั้งนี้เนื่องจากแต่ละพยางค์ในภาษาไทยนั้น มีโครงสร้างตายตัว กฎที่ใช้ภายในบทความนี้ แบ่งออกเป็น 2 หมวด

1.หมวดกฎที่สร้างจากสระ ตัวอย่างกฎจากสระที่ใช้หาขอบเขตหน้า เช่น สระ อา อะ อิ อี อึ อื อุ อู อำ อั ต้องมีอักษรข้างหน้าอย่างน้อย 1 ตัวเสมอ ตัวอย่างกฎจากสระที่ใช้หาขอบเขตหลัง เช่น สระ อี ถ้ามีวรรณยุกต์ ไม่มีตัวสะกดเสมอ สำหรับสระที่มี หรือ ไม่มีตัวสะกดก็ได้ให้ตัดแบบไม่มีตัวสะกดก่อน ผลลัพธ์ที่ได้จากการผ่านกฎจากสระ แบ่งออกเป็น 2 แบบ

- กลุ่มอักษรที่มีสระหรือวรรณยุกต์ประกอบร่วมอยู่ เช่น พ่อ, ข้าว เป็นต้น
- กลุ่มอักษรที่มีแต่พยัญชนะประกอบอย่างเดียว เช่น ขวด, ครอบครอง เป็นต้น

2.หมวดกฎที่สร้างจากพยัญชนะ ใช้กับผลลัพธ์ที่ผ่านกฎจากสระมาแล้ว โดยพิจารณาเฉพาะกลุ่มอักษรที่มีแต่พยัญชนะประกอบอย่างเดียว เพื่อใช้หาขอบเขตหน้า และขอบเขตหลังที่แท้จริงของพยางค์ เนื่องจากขอบเขตของพยางค์อาจขยายออกไปอีกได้ เช่น

- พยัญชนะต้นของพยางค์มีได้มากกว่า 1 ตัว (คำควบกล้ำหรืออักษรนำ) เช่น “กลอง”
- ตัวสะกดของพยางค์สามารถมีได้มากกว่า 1 ตัว เช่น “จิตร”

2.7 นิวรอลเน็ตเวิร์ก (Neural Networks)

นิวรอลเน็ตเวิร์ก (Neural Networks) เป็นวิทยาการแขนงหนึ่งทางปัญญาประดิษฐ์ (Artificial Intelligence) ที่พยายามลอกเลียนแบบการทำงานของเซลล์ประสาทในสมองมนุษย์ เรียกว่า เซลล์ประสาทเทียม (Artificial Neural) หรือยูนิต (Unit) ซึ่งเซลล์ประสาทเทียมที่จำลองขึ้นนั้นจะมีความเร็วในการประมวลผลที่สูงกว่าเซลล์ประสาทจริงในสมองมนุษย์ แต่เซลล์ประสาทในสมองมนุษย์สามารถทำงานได้หลากหลายและมีความสลับซับซ้อนได้ดีและเร็วกว่าเซลล์ประสาทเทียมในคอมพิวเตอร์ เนื่องจากโครงสร้างของเซลล์ประสาทมนุษย์นั้นมีโครงสร้างที่ขนานกันอย่างมาก (Massively Parallel Structure) ซึ่งการจำลองการทำงานของเซลล์ประสาทเทียมนั้น ได้อาศัยโครงสร้างที่ประกอบไปด้วยหน่วยประมวลผล (Processing Element) ที่เรียกว่า นิวรอล (Neural) จำนวนมากเชื่อมต่อกัน (Connection) ทำให้การทำงานของนิวรอลเน็ตเวิร์กนั้นเป็นไปในรูปแบบของการประมวลผลพร้อม ๆ กันจำนวนมาก (Massively Parallel Processing) โดยการเชื่อมต่อถึงกันของหน่วยประมวลผลนั้นจะมีค่าน้ำหนัก (Weight) ของแต่ละการเชื่อมต่อและเมื่อมีการให้ตัวอย่างที่ใช้ในการเรียนรู้ นิวรอลเน็ตเวิร์กก็จะทำการปรับค่าน้ำหนักให้เหมาะสม จนได้

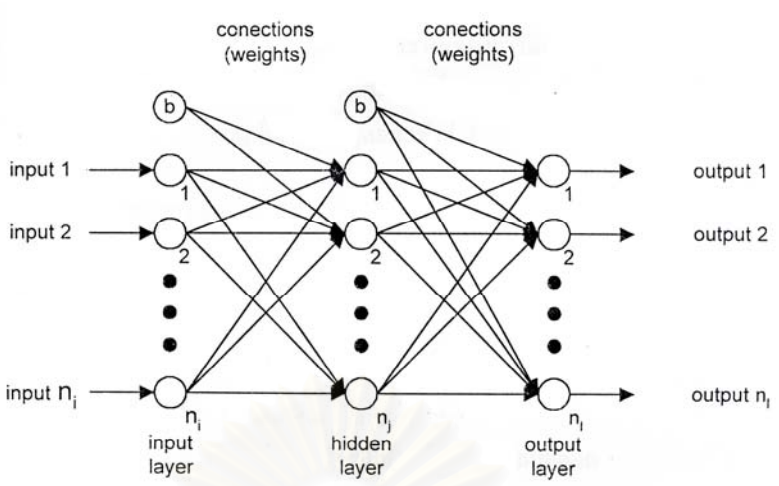
ผลลัพธ์ที่ถูกต้องหรือมีข้อผิดพลาดน้อยที่สุด และสามารถนำค่าน้ำหนักที่ถูกปรับนี้ไปใช้ในการเรียนรู้ครั้งต่อไปได้

นิวรอลเน็ตเวิร์กเป็นวิธีการเรียนรู้ของเครื่อง (Machine Learning) วิธีหนึ่ง ซึ่งมักใช้ในปัญหาการจำแนกประเภท (Classification) หรือปัญหาการจำแนกรูปแบบ (Pattern Classification) นิวรอลเน็ตเวิร์กประกอบด้วยนิวรอน (Neuron) จำนวนมากเชื่อมต่อกัน นิวรอนที่ใช้นั้นมีได้หลายรูปแบบ แบบที่นิยมใช้ได้แก่นิวรอนแบบซิกมอยด์ (Sigmoid unit) ซึ่งคำนวณผลลัพธ์จากอินพุตได้ตามสมการ (2.1) โดยกำหนดให้ W คือเวกเตอร์ของค่าน้ำหนัก (Weight) *X* คือเวกเตอร์ของค่าอินพุต และ σ คือฟังก์ชันกระตุ้น (Activation function) ประเภทฟังก์ชันซิกมอยด์ (Sigmoid function)

$$\mathbf{O} = \sigma(\mathbf{W} \cdot \mathbf{X}) \tag{2.1}$$

โดย $$\sigma(y) = \frac{1}{1+e^{-y}} \tag{2.2}$$

รูปที่ 2.1 แสดงนิวรอลเน็ตเวิร์กที่มีโครงสร้าง 3 ชั้นคือชั้นอินพุต (Input layer) ชั้นซ่อน (Hidden layer) และชั้นเอาท์พุต (Output layer) แต่ละการเชื่อมต่อของนิวรอนมีค่าน้ำหนักกำหนดไว้ ค่าน้ำหนักเหล่านี้จะถูกปรับค่าให้เหมาะสมในระหว่างการเรียนรู้ จนกระทั่งได้ผลการเรียนรู้ถูกต้องที่สุดหรือมีความผิดพลาดน้อยที่สุด วิธีการเรียนรู้ของนิวรอลเน็ตเวิร์กแบบหนึ่งคือการเรียนรู้แบบแพร่กระจายย้อนกลับ (Backpropagation Learning Algorithm) ซึ่งทำโดยการผ่านข้อมูลฝึกสอนเข้าไปยังนิวรอลเน็ตเวิร์ก แล้วนำผลลัพธ์ไปเปรียบเทียบกับค่าเป้าหมาย (Target) ผลต่างที่ได้คือค่าผิดพลาด ค่าผิดพลาดนี้จะนำไปใช้ในการคำนวณปริมาณการปรับค่าน้ำหนักเพื่อให้ค่าน้ำหนักของการเชื่อมต่อมีค่าที่เหมาะสม กระบวนการนี้จะทำซ้ำจนกระทั่งค่าผิดพลาดมีค่าน้อยในระดับที่ยอมรับได้หรือจนกระทั่งครบจำนวนรอบที่กำหนด การเรียนรู้แบบแพร่กระจายย้อนกลับสามารถอธิบายได้ดังรูปที่ 2.2



รูปที่ 2.1 นิวรอลเน็ตเวิร์กที่มี 3 ชั้น

X : vector of network input values

T : vector of target network output values

η : learning rate

x_{ji} : input from unit i to unit j

w_{ji} : weight from unit i to unit j

Create feed-forward network with n_{in} inputs, n_{hidden} hidden units, and n_{out} output units

Initialize all network weights to small random numbers

Until the termination condition is met, Do

For each training examples (X, T) Do

Input training instance X to the network and compute output O

For each network output unit k, calculate its error term δ_k

$$\delta_k \leftarrow O_k (1-O_k)(T_k-O_k)$$

For each hidden unit h, calculate its error term δ_h

$$\delta_h \leftarrow O_h (1-O_h)\sum_{k \in outputs} w_{kh}\delta_k$$

Update each network weight w_{ji}

$$w_{ji} \leftarrow w_{ji} + \Delta w_{ji} \quad , \quad \Delta w_{ji} = \eta \delta_k x_{ji}$$

รูปที่ 2.2 ขั้นตอนวิธีการเรียนรู้แบบแพร่กระจายย้อนกลับ

2.8 การวัดผลการค้นคืน

มาตรวัดที่ใช้วัดประสิทธิภาพของการค้นคืนได้แก่ ค่าความเที่ยง (Precision) ค่าเรียกคืน (Recall) [14] และตัววัด F1 (F1 Measurement) [15] ซึ่งมีวิธีคำนวณจากการนับจำนวนข้อมูลที่เกี่ยวข้อง (Relevant Data) และข้อมูลที่ระบบค้นคืนกลับมา (Retrieved Data) ได้ดังนี้

$$\text{ค่าความเที่ยง} = \frac{\text{จำนวนคำที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำที่คืนกลับมาทั้งหมด}}$$

$$\text{ค่าเรียกคืน} = \frac{\text{จำนวนคำที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำที่เกี่ยวข้องทั้งหมด}}$$

$$F1 = \frac{2 \times \text{ค่าความเที่ยง} \times \text{ค่าเรียกคืน}}{\text{ค่าความเที่ยง} + \text{ค่าเรียกคืน}}$$

คำที่เกี่ยวข้อง หมายถึง คำที่มีเสียงอ่านตรงกันจริง ส่วนคำที่คืนกลับมาหมายถึง คำที่ผ่านเกณฑ์การเปรียบเทียบรหัสคำ โดยที่ในความเป็นจริงแล้วอาจมีเสียงอ่านตรงกันหรือไม่ก็ได้

2.9 ขั้นตอนวิธีระยะแก้ไขสั้นที่สุด (Minimum Edit Distance)

ระยะแก้ไขสั้นที่สุด [6] เป็นเทคนิคหนึ่งที่ถูกนำมาใช้ในการวัดความคล้ายคลึงกันระหว่าง 2 สายอักขระใดๆ ซึ่งจะทำการคำนวณหาจำนวนคำสั่งที่น้อยที่สุดที่จะใช้ในการเพิ่ม การลบ และการแทนที่แต่ละตัวอักขระ เพื่อทำให้สายอักขระทั้งสองสายเหมือนกัน ตัวอย่างเช่น ระยะห่างของการแก้ไขให้ EXSAMBL เป็น EXAMPLE เท่ากับ 3 ซึ่งมีวิธีการคำนวณดังนี้

ตัวอย่างที่ 2.1 การคำนวณระยะห่างของการแก้ไขให้ EXSAMBL เป็น EXAMPLE

1. การลบตัวอักษร S EXSAMBL → EXAMBL
2. การแทนที่ตัวอักษร B ด้วย P EXAMBL → EXAMPL
3. การเพิ่มตัวอักษร E EXAMPL → EXAMPLE

ดังนั้นระยะห่างของการแก้ไขให้ EXSAMBL เป็น EXAMPLE มีค่าเท่ากับ 3

จากวิธีการคำนวณข้างต้นสามารถเขียนในอยู่ในรูปการคำนวณด้วยความสัมพันธ์เวียนเกิด Edit (P_j, W_k) ได้ดังนี้

$$\text{Edit}(P_0, W_0) = 0$$

$$\text{Edit}(P_j, W_0) = j$$

$$\text{Edit}(P_0, W_k) = k$$

$$\text{Edit}(P_j, W_k) = \min[\ \text{Edit}(P_{j-1}, W_k) + 1,$$

$$\text{Edit}(P_j, W_{k-1}) + 1,$$

$$Edit(P_{j-1}, W_{k-1}) + r(p_j, w_k)]$$

โดยที่ $P_j = p_1 p_2 p_3 \ldots p_j$ เป็นสายอักขระต้นแบบ มีความยาว j ตัวอักษร

$W_k = w_1 w_2 w_3 \ldots w_k$ เป็นสายอักขระเป้าหมาย มีความยาว k ตัวอักษร

$r(p_j, w_k) = 0$ ถ้า p_j เท่ากับ w_k

1 ถ้า p_j ไม่เท่ากับ w_k

2.10 งานวิจัยของ วรรณี อุดมพาณิชย์

งานวิจัย [16] นำแนวคิดของการเข้ารหัสซาวด์เด็กซ์ภาษาอังกฤษ (ดูตารางที่ 2.1) มาพัฒนา และเสนอกฎเกณฑ์การสร้างรหัสคำได้แก่ (1) ไม่ใช้สระ วรรณยุกต์ และไม้ไต่คู้ มาสร้างรหัส ยกเว้นสระที่ให้เสียงตัวสะกดเป็นพยัญชนะ ได้แก่ ไ- ใ- –ำ (2) เปลี่ยน ไ- ใ- ไ-ย -ย เป็น –ัย (3) เปลี่ยน รร เป็น –ัน (4) ตัดการันต์ พยัญชนะที่มีการันต์กำกับ รวมทั้งสระและอักษรที่ควบการันต์ทิ้ง (5) ใช้รหัสความยาว 7 หลัก รหัสตัวแรกเป็นตัวอักษรซึ่งใช้ตารางเทียบรหัสดังแสดงในตารางที่ 2.2 ส่วนรหัสตัวที่เหลือเป็นตัวเลขซึ่งใช้ตารางที่ 2.3 ในการเทียบรหัส ถ้ารหัสที่ได้มีความยาวน้อยกว่า 7 หลัก ให้เติม 0 จนครบ นอกจากนี้ยังได้เพิ่มกฎเกณฑ์เพื่อให้เหมาะสมกับภาษาไทย ได้แก่

- กรณีพบ ใ- ไ- ไ-ย และ –ัย จะเปลี่ยนให้อยู่ในรูปแบบเดียวกันคือ –ัย ก่อนทำการเข้ารหัส เนื่องจากสระดังกล่าวอ่านออกเสียงเหมือนกัน เช่น ใท ไท ไทย และ ทัย จะเปลี่ยนเป็น ทัย
- กรณีพบ รร จะเปลี่ยนเป็น –ัน ในกรณีที่ไม่มีตัวสะกดตามหลัง และเปลี่ยนเป็น –ั กรณีที่มีตัวสะกดตามหลัง เช่น เปลี่ยนคำว่า สรรเพชร รังสรรค์ พรรณนที และ ธรรมรัตน์ เป็น สันเพชร รังสัน พัณนที และ ธัมรัตน์ ตามลำดับ
- กรณีพบการันต์ จะตัดการันต์และพยัญชนะที่มีตัวการันต์กำกับรวมทั้งสระและอักษรควบการันต์ทิ้ง เช่น คำว่า จันทร์ ศักดิ์ และ พันธุ์ เปลี่ยนเป็น จัน ศัก และ พัน ตามลำดับ

ตัวอย่างการเข้ารหัสคำ เช่น คำ “อัมพร” หรือ “อำภรณ์” จะได้รหัสเป็น “อ059000” คำ “พรรณศักดิ์” หรือ “พันธุ์ศักดิ์” จะได้รหัสเป็น “พ341000” คำ “เนืองนิตย์” หรือ “เนืองนิจ” จะได้รหัสเป็น “น623400”

ตารางที่ 2.1 การกำหนดรหัสซาวด์เด็กซ์ภาษาอังกฤษของ Odell และ Russel

ตัวอักษร	รหัสตัวเลข
A E I O U H W Y	0
B F P V	1
C G J K Q S X Z	2
D T	3
L	4
M N	5
R	6

ตารางที่ 2.2 การกำหนดรหัสตัวอักษรของรหัสซาวด์เด็กซ์ภาษาไทย
จากงานวิจัยของวรรณี อุดมพาณิชย์

รหัสตัวอักษร	ตัวอักษร
ก	ก
ข	ข ฃ ค ฅ ฆ
ง	ง
จ	จ
ช	ช ฌ ฉ
ส	ซ ศ ษ ส
ด	ด ฎ
ต	ต ฏ
ท	ฐ ฑ ฒ ถ ท ธ
น	ณ น

รหัสตัวอักษร	ตัวอักษร
บ	บ
ป	ป
พ	พ ภ ผ
ฟ	ฝ ฟ
ม	ม
ย	ญ ย
ร	ร ล ฬ ฤ ฦ
ว	ว
อ	อ
ห	ห ฮ

ตารางที่ 2.3 การกำหนดรหัสตัวเลขของรหัสซาวด์เด็กซ์ภาษาไทย
จากงานวิจัยของของวรรณี อุดมพาณิชย์

รหัสตัวเลข	ตัวอักษร
0	ม ว –ำ
1	ก ข ฃ ค ฅ ฆ
2	ง ย
3	ญ ณ น
4	ฎ ฏ ด ต ศ ษ ส
5	บ ป พ ภ
6	ผ ฝ ฟ ห อ ฮ
7	จ ฉ ช ซ ฌ
8	ฐ ฑ ฒ ถ ท ธ
9	ร ฤ ล ฦ ฬ

2.11 งานวิจัยของ นิลเนตร อรุณวงศ์ ณ อยุธยา

งานวิจัย [17] มีการปรับวิธีการจาก [16] ได้แก่ (1) ใช้อักขระไทยทั้งหมดในการเข้ารหัส โดยใช้ตารางที่ 2.4 เป็นตารางกำหนดรหัสสำหรับอักขระตัวแรกของรหัส และตารางที่ 2.5 สำหรับอักขระตัวที่เหลือของรหัส (2) ตัดตัวควบกล้ำทิ้ง เช่น ปรับ เป็น ปับ หรือ คลอง เป็น คอง (3) อักษรนำเสียงสนิท (ได้แก่ อย, หง, หญ, หน, หม, หย, หร, หล, หว ให้ตัด “อ“ หรือ “ห“ ทิ้ง (4) เปลี่ยนตำแหน่งสระหน้า (เช่น เ- แ- โ- ไ- ใ-) ไปไว้หลังสุด (5) ตัวอักษรติดกันและเหมือนกันให้ยุบเหลือรหัสตัวเดียว (6) ไม่นำสระ -ะ และ ิ- มาเข้ารหัส ตัวอย่างการเข้ารหัสคำ เช่น “กิตติพรรณ” แทนด้วย “กตบน” ซึ่งได้จากการตัด ิ– , ตัว “ต” ติดกัน 2 ตัว ตัดเหลือตัวเดียว, ตัว “พ” แทนด้วยรหัส “บ”, แทน “รร” ด้วย –ัน แล้วตัดตัว –ั ออก

ตัวอย่างการเข้ารหัสคำ เช่น คำ “อัมพร” หรือ “อำภรณ์” จะได้รหัสเป็น “อมบน” คำ “พรรณศักดิ์” หรือ “พันธุ์ศักดิ์” จะได้รหัสเป็น “พนดก” คำ “เนืองนิตย์” หรือ “เนืองนิจ” จะได้รหัสเป็น “นืองนดเ”

ตารางที่ 2.4 การกำหนดรหัสสำหรับอักขระตัวแรก
จากงานวิจัยของนิลเนตร อรุณวงศ์ ณ อยุธยา

รหัสตัวอักษร	ตัวอักษร
ก	ก
ข	ข ฃ ค ฅ ฆ
ง	ง
จ	จ
ช	ช ฌ ฉ
ซ	ซ ศ ษ ส
ด	ด ฎ
ต	ต ฏ
ท	ฐ ฑ ฒ ถ ท ธ
น	ณ น

รหัสตัวอักษร	ตัวอักษร
บ	บ
ป	ป
พ	พ ภ ผ
ฟ	ฝ ฟ
ม	ม
ย	ญ ย
ร	ร ล ฬ ฤ ๅ
ว	ว
อ	อ
ฮ	ห ฮ

ตารางที่ 2.5 การกำหนดรหัสสำหรับอักขระถัดจากตัวแรก
จากงานวิจัยของนิลเนตร อรุณวงศ์ ณ อยุธยา

รหัสตัวอักษร	ตัวอักษร
ก	ก ข ค ฆ
ง	ง
ด	จ ฉ ช ซ ฌ ฎ ฏ ฐ ฑ ฒ ด ต ถ ท ธ ศ ส ษ
น	ญ ณ น ร ล ฬ
บ	บ ป พ ฟ ภ ผ ฝ
ม	ม ำ
ย	ย
ว	ว
ไ	ไ-ย -ย ใ-

2.12 งานวิจัยของ ประยุทธ สุวรรณวิสาท และ สมชาย ประสิทธิ์จูตระกูล

สำหรับงานวิจัย [3] และ [4] ได้แบ่งการทดลองออกเป็น 2 ส่วนหลักๆ คือ การค้นคืนข้ามภาษาไทยทับศัพท์ภาษาอังกฤษ และ การค้นคืนข้ามภาษาอังกฤษทับศัพท์ภาษาไทย โดยที่กรณี

การค้นคืนข้ามภาษาไทยทับศัพท์ภาษาอังกฤษ นั้นใช้รหัสคำเป็นตัวเลขทั้งหมดโดยไม่จำกัดความยาวของรหัสคำ ตารางที่ 2.6 แสดงการกำหนดรหัสสำหรับคำไทยทับศัพท์คำอังกฤษ สำหรับการค้นคืนนั้นใช้การเปรียบเทียบรหัสคำแบบเหมือนกันทุกประการ นั่นคือรหัสคำต้องตรงกันทุกตัวจึงจะถือว่าสองคำนั้นมีเสียงอ่านตรงกัน จากผลการทดลองพบว่าเมื่อใช้รหัสคำความยาวมากกว่า 4 หลักขึ้นไป ได้ค่าความเที่ยง 78% และค่าเรียกคืน 90% สำหรับกรณีการค้นคืนข้ามภาษาอังกฤษทับศัพท์ภาษาไทย ใช้รหัสคำเป็นตัวอักษรไทยผสมกับสัญลักษณ์เสียงสากล

ตารางที่ 2.7 และตารางที่ 2.8 แสดงรหัสสำหรับพยัญชนะและสระตามลำดับ สำหรับคำไทยมีการประมวลผลเบื้องต้นก่อนการเข้ารหัสคำ ซึ่งได้แก่ การลดรูป การตัดวรรณยุกต์และตัวการันต์ การแทนที่ ไ- ใ- ไ-ย -ย ด้วย -ัย แทนที่ รร เป็น -ัน เป็นต้น ในการเปรียบเทียบรหัสคำนั้นใช้วิธีการเปรียบเทียบเชิงประมาณด้วยเทคนิคระยะแก้ไขสั้นสุด และมีการกำหนดต้นทุนในการแทนที่อักขระสำหรับแต่ละคู่อักขระตามกฎเกณฑ์ที่ได้สร้างขึ้น ค่าของต้นทุนมี 4 ระดับคือ C_1 C_2 C_3 และ C_4 ซึ่งมีค่า 0 1 4 และ 7 ตามลำดับ ดังตัวอย่างในรูปที่ 2.3 และรูปที่ 2.4 จากผลการทดลองพบว่าได้ค่าความเที่ยง 69% และค่าเรียกคืน 73% สำหรับกรณีคำไทยทับศัพท์คำอังกฤษ และได้ค่าความเที่ยง 96% และค่าเรียกคืน 75% สำหรับกรณีคำอังกฤษทับศัพท์คำไทย

2.13 งานวิจัยของ ทัศนวรรณ ศูนย์กลาง สมชาย ประสิทธิ์จูตระกูล และบุญเสริม กิจศิริกุล

งานวิจัย [5] เสนอการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์กแบบแพร่กระจายย้อนกลับ โดยการพิจารณาตัวอักษรที่อยู่ข้างเคียงกับตัวที่กำลังพิจารณา ในงานวิจัยนี้ใช้ตัวอักษรที่อยู่ข้างเคียงข้างหน้าและข้างหลังอย่างละ 4 ตัว ดังนั้นเมื่อรวมตัวอักษรที่กำลังพิจารณาแล้ว ข้อมูลเข้าของนิวรอลเน็ตเวิร์กจึงมี 9 ตัวอักษร ส่วนข้อมูลออกคือรหัสเสียงจากข้อมูลขาเข้า ดังแสดงในรูปที่ 2.5 และรูปที่ 2.6 นอกจากนี้การเข้ารหัสคำของคำไทยยังมีการใช้การประมวลผลเบื้องต้นด้วยเช่นเดียวกับในงานวิจัย [12] ในการค้นคืนใช้วิธีการเปรียบเทียบรหัสคำแบบประมาณด้วยวิธีระยะแก้ไขสั้นที่สุด โดยยอมให้ความต่างของรหัสคำมีได้ไม่เกิน 1 และกำหนดค่าต้นทุนการแก้ไขอักขระทั้งการเพิ่ม ลบ และแทนที่ ให้มีค่าเท่ากับ 1 จากผลการทดลองพบว่าได้ค่าความเที่ยง 87.28% และค่าเรียกคืน 77.19% สำหรับกรณีคำไทยทับศัพท์คำอังกฤษ และได้ค่าความเที่ยง 96.34% และค่าเรียกคืน 75.15% สำหรับกรณีคำอังกฤษทับศัพท์คำไทย

ตารางที่ 2.6 การกำหนดรหัสสำหรับคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ จากงานวิจัยของ ประยุทธ สุวรรณวิสาทร

ภาษาอังกฤษ	ภาษาไทย	รหัส
A E I O U H W Y [2]	อ ห ฮ ย ญ	0
B F P V	บ ฝ ฟ ป ผ พ ภ ว	1
C G J K Q S X Z	ข ฃ ค ฅ ฆ ฉ ช ฌ ก จ ซ ศ ษ ส	2
D T	ฎ ด ฏ ต ฐ ฑ ฒ ถ ท ธ	3
L	ล ฬ	4
M N	ม ณ น	5
R	ร	6
A E I O U [1]	อ	7
H [1]	ห ฮ	8
W [1]	ว	1
Y [1]	ย ญ	9
	ง	52

1 : สำหรับตัวอักษรแรกของคำ

2 : สำหรับตัวอักษรตั้งแต่ตัวที่ 2 เป็นต้นไปของคำ

ตารางที่ 2.7 การกำหนดรหัสของพยัญชนะสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัยของ ประยุทธ สุวรรณวิสาทร

อักษรอังกฤษ	อักษรไทย	รหัส
b	บ	บ
bh	พ	พ
c	ซ	ซ
ch	ช ฉ ฌ	ช
ck	ก	ก
d	ด ฎ	ด
dh	ท	ท
f	ฟ ฝ	ฟ

อักษรอังกฤษ	อักษรไทย	รหัส
n	น ณ	น
ng	ง	ง
p	ป	ป
ph	พ ผ ภ	พ
q	ค	ค
r	ร ฤ	ร
s	ส ซ ศ ษ	ส
t	ต ฏ	ต

ตารางที่ 2.7 (ต่อ) การกำหนดรหัสของพยัญชนะสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัยของ ประยุทธ สุวรรณวิสาทร

อักษรอังกฤษ	อักษรไทย	รหัส
g	ก	ก
h	ห ฮ	ห
j	จ	จ
k	ก	ก
kh	ข ฃ ค ฅ ฆ	ข
l	ล ฦ ฬ	ล
m	ม	ม

อักษรอังกฤษ	อักษรไทย	รหัส
th	ท ฐ ฑ ฒ ถ ธ	ท
v	ว	ว
w	ว	ว
x	ก	ก
y	ย ญ	ย
z	ซ	ซ

ตารางที่ 2.8 การกำหนดรหัสของสระสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัยของ ประยุทธ สุวรรณวิสาทร

ตัวอักษรอังกฤษ	ตัวอักษรไทย	รหัส
-a	ะ	ะ
-aa	า	า
-ae	แ-ะ แ-	x
-ai	-ัย	-ัย
-ao	เ-า	@
-aiu	เ-ียว	I
-arn	าน	าน
-art	าท	าท
-e	เ-ะ เ-	เ
-ee	-ี	-ี
-eo	แ-ว	แ-ว
-er	เ-อ เ-ิ	q

ตัวอักษรอังกฤษ	ตัวอักษรไทย	รหัส
-eu	-ึ	-ึ
-i	-ิ	-ิ
-ia	เ-ียะ เ-ีย	I
-ie	เ-ียะ เ-ีย	I
-o	-อ	อ
-oe	เ-อ เ-	Q
-oi	อย	อย
-oo	-ู	-ู
-orn	-อน	ร
-u	-ึ -ื -ุ -ู	-ุ
-ua	เ-ือะ เ-ือ -ัวะ -ัว	U
-ue	-ื	-ื

	ก	ข	ค	…	ท	ธ	น	บ	…
ก	C_1	C_2	C_2	…	C_4	C_4	C_4	C_4	…
ข	C_2	C_1	C_1	…	C_4	C_4	C_4	C_4	…
ค	C_2	C_1	C_1	…	C_4	C_4	C_4	C_4	…
…	…	…	…	…	…	…	…	…	…
ท	C_4	C_4	C_4	…	C_1	C_1	C_4	C_4	…
ธ	C_4	C_4	C_4	…	C_1	C_1	C_4	C_4	…
น	C_4	C_4	C_4	…	C_4	C_4	C_1	C_4	…
บ	C_4	C_4	C_4	…	C_4	C_4	C_4	C_1	…
…	…	…	…	…	…	…	…	…	…

รูปที่ 2.3 ตัวอย่างการกำหนดต้นทุนการแทนที่อักขระสำหรับพยัญชนะ
จากงานวิจัยของประยุทธ สุวรรณวิสาทร

	ะ	ั	า	ิ	ี	ึ	ื	ุ	…
ะ	C_1	C_1	C_1	C_4	C_4	C_4	C_4	C_4	…
ั	C_1	C_1	C_2	C_4	C_4	C_4	C_4	C_2	…
า	C_1	C_2	C_1	C_4	C_4	C_4	C_4	C_4	…
ิ	C_4	C_4	C_4	C_1	C_1	C_3	C_4	C_4	…
ี	C_4	C_4	C_4	C_1	C_1	C_4	C_4	C_4	…
ึ	C_4	C_4	C_4	C_3	C_4	C_1	C_1	C_2	…
ื	C_4	C_4	C_4	C_4	C_4	C_1	C_1	C_2	…
ุ	C_4	C_2	C_4	C_4	C_4	C_2	C_2	C_1	…
…	…	…	…	…	…	…	…	…	…

รูปที่ 2.4 ตัวอย่างการกำหนดต้นทุนการแทนที่อักขระสำหรับสระ
จากงานวิจัยของประยุทธ สุวรรณวิสาทร

ลำดับตัวอักษร	รหัส
_ , _ , _ , _ , ก , น , ก , พ , ◌ี	k
_ , _ , _ , ก , น , ก , พ , ◌ี , ร	a, n
_ , _ , ก , น , ก , พ , ◌ี , ร , ะ	o, k
_ , ก , น , ก , พ , ◌ี , ร , ะ , ว	p
ก , น , ก , พ , ◌ี , ร , ะ , ว , ◌ุ	i
น , ก , พ , ◌ี , ร , ะ , ว , ◌ุ , ฒ	r
ก , พ , ◌ี , ร , ะ , ว , ◌ุ , ฒ , ◌ิ	a
พ , ◌ี , ร , ะ , ว , ◌ุ , ฒ , ◌ิ , _	v
◌ี , ร , ะ , ว , ◌ุ , ฒ , ◌ิ , _ , _	u
ร , ะ , ว , ◌ุ , ฒ , ◌ิ , _ , _ , _	t
ะ , ว , ◌ุ , ฒ , ◌ิ , _ , _ , _ , _	_

รูปที่ 2.5 ตัวอย่างการเข้ารหัสคำไทยของทัศนวรรณ ศูนย์กลาง และคณะ

ลำดับตัวอักษร	รหัส
_ , _ , _ , _ , R , H , O , D , I	r
_ , _ , _ , R , H , O , D , I , U	_
_ , _ , R , H , O , D , I , U , M	o
_ , R , H , O , D , I , U , M , _	d
R , H , O , D , I , U , M , _ , _	l
H , O , D , I , U , M , _ , _ , _	_
O , D , I , U , M , _ , _ , _ , _	m

รูปที่ 2.6 ตัวอย่างการเข้ารหัสคำอังกฤษของทัศนวรรณ ศูนย์กลาง และคณะ

2.14 สรุป

ในบทนี้ได้กล่าวถึงทฤษฎีที่เกี่ยวข้องกับงานวิจัยนี้ ได้แก่ การถอดอักษร การถ่ายเสียงนิวรอลเน็ตเวิร์ก การวัดผลการค้นคืน และขั้นตอนวิธีระยะแก้ไขสั้นสุด จากนั้นได้กล่าวถึงงานวิจัยในอดีตที่เกี่ยวข้องกับการเข้ารหัสคำ และการค้นคืนข้ามภาษาไทย-อังกฤษ สำหรับในบทต่อๆ ไปจะกล่าวถึงการนำทฤษฎี และงานวิจัยที่เกี่ยวข้องมาประยุกต์ใช้ในการทำงานวิจัยนี้

บทที่ 3

การเข้ารหัสคำ

โดยปกติกระบวนการที่ใช้ในการพัฒนาระบบค้นคืนข้ามภาษาโดยไม่อาศัยพจนานุกรมนั้น มีความจำเป็นอย่างยิ่งที่จะต้องอาศัยวิธีการเข้ารหัสคำ และนำรหัสคำที่ผ่านการเข้ารหัสแล้วไปสร้างเป็นดัชนีเก็บไว้ในฐานข้อมูลเพื่อใช้ในขั้นตอนของการสืบค้นต่อไป โดยเมื่อมีการสืบค้นข้อมูลก็จะนำข้อมูลที่ต้องการสืบค้นไปสร้างเป็นรหัสคำ แล้วนำรหัสคำที่ได้ไปค้นในดัชนีรหัสคำที่ระบบได้สร้างและเก็บเอาไว้ สำหรับเนื้อหาในบทนี้จะกล่าวถึงกระบวนการที่ใช้ในการเข้ารหัสคำ ซึ่งได้แก่ การกำหนดรหัสเสียงในภาษาไทย และภาษาอังกฤษ ทั้งแบบเก่า และแบบใหม่ การประมวลผลเบื้องต้นสำหรับคำในภาษาไทย และขั้นตอนวิธีที่ใช้ในการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์ก

3.1 รหัสคำ

รหัสคำ คือ สัญลักษณ์แทนเสียงอ่านของคำ ดังนั้นคำที่มีเสียงอ่านตรงกันจะมีรหัสคำที่เหมือนกัน หรืออย่างน้อยก็ต้องมีรหัสใกล้เคียงกัน รหัสคำจะประกอบไปด้วยรหัสของเสียง ซึ่งรหัสของเสียงในแต่ละเสียงนั้นจะแทนเสียงของตัวอักขระแต่ละตัวที่ปรากฏอยู่ในคำนั้นมาเรียงต่อกัน สำหรับในงานวิจัยชิ้นนี้ได้ใช้หลักเกณฑ์การอ่านออกเสียงของคำทั้งในภาษาไทยและภาษาอังกฤษของราชบัณฑิตยสถาน [18] [19] มาสร้างเป็นตารางรหัสเสียงแบบเดิมดังที่แสดงไว้ในตารางที่ 3.1 และตารางที่ 3.2 โดยในกรณีคำไทยทับศัพท์คำอังกฤษจะใช้หน่วยเสียงภาษาอังกฤษเป็นหลัก แล้วนำหน่วยเสียงภาษาไทยไปเทียบเพื่อหากลุ่มเสียงที่ตรงกันหรือใกล้เคียงกัน ในทางกลับกันในกรณีคำอังกฤษทับศัพท์คำไทยจะใช้หน่วยเสียงภาษาไทยเป็นหลัก แล้วนำหน่วยเสียงภาษาอังกฤษไปเทียบ

ส่วนที่แสดงในตารางที่ 3.4 และตารางที่ 3.5 นั้นใช้สำหรับการเข้ารหัสเสียงแบบใหม่ที่ลดความกำกวมของรหัสที่ซ้ำกันในหน่วยเสียงของพยัญชนะต้นกับหน่วยเสียงของตัวสะกด เพื่อให้ง่ายต่อการตัดแบ่งพยางค์ของรหัสเสียงในกระบวนการถัดไป

ตารางที่ 3.1 รหัสเสียงสำหรับคำไทยทับศัพท์คำอังกฤษแบบเดิม

เสียงพยัญชนะ		รหัสเสียง	เสียงสระ		รหัสเสียง
ไทย	อังกฤษ		ไทย	อังกฤษ	
พ	p	p	◌ิ ◌ี	ee, ei, ea, ey, i	i
บ	b	b	เ	e, ay	e
ท,ต	t, th	t	แ	a, air, are	w
ด	d, th	d	เ◌ียว	a, aw, au	@
ก,ค	c, k, g	k	◌ุ , ◌ู	u oo	u
ช	ch, sh	c	เ–อ, เ◌ิ	ur, er, ir, or	W
จ	j, ch, g	j	◌ะ, ◌ั	a	a
ฟ	f, ph	f	โ	ome, o	o
ว	w, v	v	ไ, ใ, ◌ัย, -าย	ie, ai	!
ส,ซ	s, z	s	-าว, เ-า	ow, ou, our, au	R
ฮ	h	h	-วย	oi	O
ม	m	m	เ◌ีย	ear, ia	I
น	n	n	◌ัว	our, ua	Y
ง	ng	g	◌ิว	ew, eua	X
ล	l	l	เ◌ิล	le	Q
ร	r	r			
ย	y	y			
ตัวอักษรที่ไม่ออกเสียง		_			

ตารางที่ 3.2 รหัสเสียงพยัญชนะสำหรับคำอังกฤษทับศัพท์คำไทยแบบเดิม

เสียงพยัญชนะ		รหัสเสียง		เสียงสระ		รหัสเสียง
ไทย	อังกฤษ	ตัวต้น	ตัวสะกด	ไทย	อังกฤษ	
ก ข ค ฆ	ck, g, k, x, c, kh, q	k	k	◌ิ ◌ี	i, ee	i
ง	ng	g	g	ะ า ◌ั	a, u, ar	a
จ ฉ ช ฌ	j, ch, x	c	t	เ–ะ เ–	e	e
ซ ส ศ ษ สร ทร	s, z	s	t	แ–ะ แ–	ae	w
ญ ย หย หญ	y	y	n	◌ึ ◌ื ◌ือ ◌ุ ◌ู	u, eu, ue, eo, oo	u
ด ฎ ฑ	d	d	t	โ–ะ โ– เ–าะ –อ	o	o
ต ฏ ถ ฐ ท ฑ ธ ฒ	t, th, dh	t	t	เ–อะ เ–อ เ◌็	er, oe	W
ณ น หน	n	n	n	เ◌ียะ เ◌ีย	ia, ie, aiu	I
บ	b	b	p	เ◌ือะ เ◌ือ –ัวะ –ัว	ua, ue, ea, ui	Y
ป ผ พ ภ	p, ph, bh	p	p	ไ– ใ– ไ–ย –ัย –าย	ai, ie, uy	!
ฝ ฟ	f	f	p	เ–า –าว	ao, ou, ow	R
ม	m	m	m	โ–ย –อย	oi, oy	x
ร ฤ	r	r	n	◌ิว	iu	X
ล ฬ ฦ	l	l	n	เ–ว	eo	q
ว	w, v	v	–	เ-ย	oei	Q
ห ฮ	h	h	–	เ◌ือย –วย	uai, uay, ou	O
ตัวอักษรที่ไม่ออกเสียง	–	–		แ-ว	aeo, eo, aew	$

ตารางที่ 3.2 (ต่อ) รหัสเสียงพยัญชนะสำหรับคำอังกฤษทับศัพท์คำไทยแบบเดิม

เสียงพยัญชนะ		รหัสเสียง		เสียงสระ		รหัสเสียง
ไทย	อังกฤษ	ตัวต้น	ตัวสะกด	ไทย	อังกฤษ	
				เ-ียว	ieo, eaw, eo, ew, iow, iau, iew, iaw	@

3.2 การประมวลผลเบื้องต้น

ในกรณีของคำที่เขียนในรูปแบบของภาษาไทยนั้น ก่อนการนำคำที่ต้องการมาเข้ารหัสจะต้องผ่านการประมวลผลตัวอักษรเบื้องต้นก่อน เพื่อช่วยลดความซับซ้อนในการเข้ารหัสคำ สำหรับรายละเอียดของการประมวลผลในส่วนนี้ ผู้วิจัยได้ยึดตามหลักการของ [12] ดังนี้

1. ทำการตัดรูปของวรรณยุกต์ และไม้ไต่คู้ทิ้ง
2. เปลี่ยนตัว รร ให้เป็น –ัน หรือ –ั ขึ้นกับชนิดของแต่ละคำ
3. เปลี่ยน ไ- ใ- ไ-ย ให้เป็น –ัย
4. เปลี่ยน –ำ ให้เป็น –ัม
5. ตัดตัวการันต์ และอักษรควบตัวการันต์
6. เปลี่ยน ฤ เป็น รึ ริ หรือ เรอ และเปลี่ยน ฤๅ เป็น รือ โดยตัว ฤ ต้องพิจารณาดังนี้
   6.1. ฤ ออกเสียงเป็น เรอ มีคำเดียว คือ ฤกษ์ โดยเปลี่ยนเป็น เริก
   6.2. ฤ ออกเสียงเป็น ริ ถ้าประสมกับ ก ต ท ป ศ ส เช่น กฤษณา ตฤณ ทฤษฎี ปฤงคพ ศฤงคาร สฤษฏ์
   6.3. ฤ ออกเสียงเป็น รึ ถ้าประสมกับตัวอื่น เช่น คฤหาสน์ พฤศจิกายน มฤตยู หฤทัย
7. อักษร “ห” นำ ให้ตัด ห ทิ้งถ้าอักษร “ห” นำตัวอักษร ร ล ว ง ญ น ม เพราะไม่ออกเสียงพยัญชนะ “ห” แต่ออกเสียงพยัญชนะต้นตามตัวอักษรที่ตามหลัง “ห” เช่น หรู ไหล หวี เหงา หญิง หนา หมู

3.3 วิธีการเข้ารหัสคำ และการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์ก

การเข้ารหัสคำ คือ การแปลงแต่ละตัวอักษรที่ปรากฏอยู่ในคำให้ไปเป็นรหัสเสียงโดยการเปรียบเทียบเสียงจากตารางรหัสเสียงที่ได้ทำการออกแบบไว้ แล้วนำรหัสเสียงมาเรียงต่อกันจนเป็นรหัสคำ ประเด็นหนึ่งที่ต้องพิจารณาในขั้นตอนการหารหัสเสียง นั่นก็คือ ความสัมพันธ์ระหว่าง

ตัวอักษรและรหัสเสียง สำหรับความสัมพันธ์แบบหนึ่งต่อหนึ่งซึ่งสามารถเทียบรหัสเสียงได้โดยตรงจากตาราง แต่ถ้าเป็นความสัมพันธ์แบบหนึ่งต่อหลายจะมีวิธีการพิจารณาดังนี้

1. <u>แบบหลายตัวอักษรต่อหนึ่งหน่วยเสียง</u> ในกรณีนี้จะมีตัวอักษรตัวเดียวที่ให้รหัสเสียง ส่วนตัวที่เหลือให้รหัสเป็น "_" (หมายถึงตัวอักษรที่ไม่ออกเสียง) ตัวอย่างเช่น คำภาษาอังกฤษ "king" จะมีรหัสเป็น kig_ โดยรหัสเสียง g เกิดจากกลุ่มตัวอักษร "ng" หรือเช่นคำไทย "เกรียง" จะมีรหัสเป็น lkr__g โดยรหัสเสียง l เกิดจากกลุ่มตัวอักษร "เ–ีย"

2. <u>แบบหลายหน่วยเสียงต่อหนึ่งตัวอักษร</u> ในบางกรณีสำหรับคำอังกฤษ ตัวอักษรตัวหนึ่งในคำอาจให้เสียงมากกว่า 1 เสียง เช่น ตัว "x" ใน "toxin" ให้ทั้งเสียง /k/ และ /s/ ในกรณีนี้จะเลือกเพียงเสียงเดียวโดยให้มีรหัสเป็น tosin สำหรับคำไทยนั้นมีการใช้สระลดรูป เช่น ลดา (สระ –ะ ลดรูป) นก (สระ โ-ะ ลดรูป) สุนทร (สระ –อ ลดรูป) ในรหัสคำจะแทรกรหัสเสียงตามเสียงสระที่ลดรูปไป ซึ่งทำให้บางกรณีให้รหัสเสียงมากกว่า 1 รหัส ดังนั้นจากตัวอย่าง "ลดา" จะได้รหัสเป็น l<u>a</u>da "นก" จะได้รหัสเป็น n<u>o</u>k และ "สุนทร" จะได้รหัสเป็น sunt<u>o</u>n

นอกจากนี้ในงานวิจัยยังได้แบ่งการเข้ารหัสคำตามภาษาและการทับศัพท์ ซึ่งได้แก่ คำไทย (เช่น คำว่า "เจริญ") คำอังกฤษทับศัพท์คำไทย (เช่น คำว่า "charoen") คำอังกฤษ (เช่น คำว่า "interface") และคำไทยทับศัพท์คำอังกฤษ (เช่น คำว่า "อินเตอร์เฟส") ดังนั้นในงานวิจัยนี้จึงต้องมีตัวเข้ารหัสคำทั้งหมด 4 ชุดด้วยกัน

ในบางครั้งการจะทราบว่าตัวอักษรที่กำลังพิจารณานั้นให้เสียงอย่างไร มีความจำเป็นอย่างยิ่งที่จะต้องพิจารณาตัวอักษรที่อยู่ข้างเคียงด้วย ซึ่งปัญหานี้เป็นปัญหาการจำแนกประเภท (Classification) รูปแบบหนึ่ง ในงานวิจัยนี้จึงได้นำเสนอการเข้ารหัสคำโดยใช้วิธีการของนิวรอลเน็ตเวิร์ก โดยจำนวนตัวอักษรที่จะใช้ในการพิจารณานั้นจะขึ้นอยู่กับว่าคำที่ต้องการจะเข้ารหัสนั้นเขียนอยู่ในรูปแบบใด ถ้าหากเขียนอยู่ในรูปแบบของตัวอักษรภาษาไทย (ซึ่งในการทดลองนี้ ได้แก่ คำไทย และคำไทยทับศัพท์คำอังกฤษ) จะใช้ตัวอักษรจำนวน 9 ตัว ในการพิจารณาแต่ละครั้ง แต่ถ้าเขียนอยู่ในรูปแบบตัวอักษรภาษาอังกฤษ (ได้แก่ คำอังกฤษ และคำอังกฤษทับศัพท์คำไทย) จะใช้จำนวนตัวอักษรจำนวน 7 ตัว ในการพิจารณาแต่ละครั้ง (จำนวนตัวอักษรที่เหมาะสมในการนำมาใช้พิจารณานั้น ได้มาจากการทดลอง) ซึ่งจำนวนของตัวอักษรที่ใช้ในการพิจารณาดังที่กล่าวมานั้น จะประกอบไปด้วย ตัวอักษรที่กำลังจะพิจารณาว่าจะให้รหัสเสียงเป็นเสียงใดจำนวน 1 ตัว และส่วนที่เหลือจะเป็นตัวอักษรที่อยู่ข้างหน้า และตัวอักษรที่อยู่ข้างหลังอย่างละเท่าๆ กัน ดังตัวอย่างที่ได้แสดงไว้ใน รูปที่ 3.1 และ รูปที่ 3.2 ซึ่งเป็นตัวอย่างแสดงการเข้ารหัสคำไทย คำว่า "สุรเกียรติ" และการเข้ารหัสคำอังกฤษทับศัพท์คำไทย คำว่า "surakiat" ในด้านซ้ายมือของภาพ คือ

ลำดับของตัวอักษรที่ประกอบด้วยตัวอักษรข้างเคียงและตัวอักษรที่กำลังพิจารณา โดยเครื่องหมาย '_' ในภาพจะหมายถึง ตัวอักษรว่าง (Blank) ส่วนทางด้านขวามือของภาพ คือรหัสเสียงของแต่ละลำดับโดยเครื่องหมาย '_' หมายความว่าไม่มีการออกเสียงสำหรับลำดับตัวอักษรนั้น เมื่อนำรหัสเสียงจากทุกลำดับมาต่อกันเรียงกัน จากรูปที่ 3.1 จะได้รหัสคำที่ผ่านการเข้ารหัสเป็น "suralk___t_" และจากรูปที่ 3.2 จะได้รหัสคำที่ผ่านการเข้ารหัสเป็น "surakl_t"

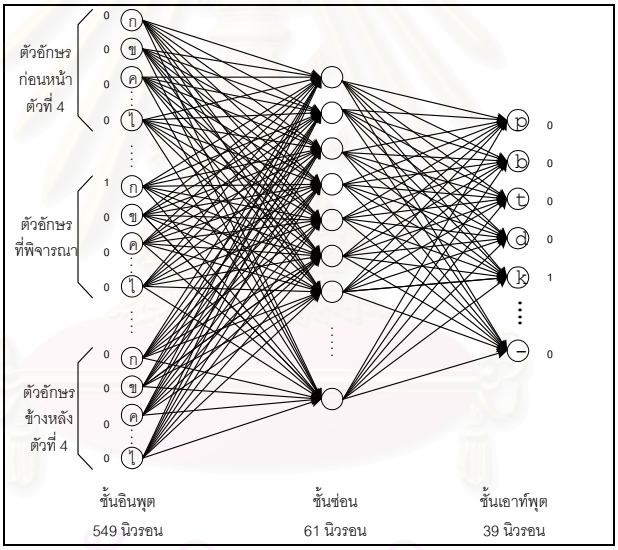
นิวรอลเน็ตเวิร์กที่ใช้เป็นแบบแบ็กพรอพาเกชัน (Backpropagation Neural Network) ซึ่งมีจำนวนชั้นของการทำงานจำนวน 3 ชั้น ดังแสดงในรูปที่ 3.3 โดยข้อมูลขาเข้าเป็นตัวอักษรภาษาไทย ประกอบด้วยตัวอักษรที่ต้องการทราบรหัสเสียง และตัวอักษรที่อยู่ข้างเคียงกับมัน ข้างละ 4 ตัว รวมเป็น 9 ตัว (เนื่องจากข้อมูลเข้าอยู่ในรูปอักษรไทย จึงใช้จำนวนตัวอักษร 9 ตัว) ส่วนข้อมูลขาออกจะเป็นรหัสเสียงของข้อมูลขาเข้า (ตามที่ผู้วิจัยได้ทำการออกแบบและกำหนดไว้ในตารางที่ 3.1 และตารางที่ 3.2)

ลำดับตัวอักษร	รหัส
_ , _ , _ , _ , ส , ุ , ร , เ , ก	s
_ , _ , _ , ส , ุ , ร , เ , ก , ี	u
_ , _ , ส , ุ , ร , เ , ก , ี , ย	r
_ , ส , ุ , ร , เ , ก , ี , ย , ร	a I
ส , ุ , ร , เ , ก , ี , ย , ร , ต	k
ุ , ร , เ , ก , ี , ย , ร , ต , ิ	_
ร , เ , ก , ี , ย , ร , ต , ิ , _	_
เ , ก , ี , ย , ร , ต , ิ , _ , _	_
ก , ี , ย , ร , ต , ิ , _ , _ , _	t
ี , ย , ร , ต , ิ , _ , _ , _ , _	_

รูปที่ 3.1 การเข้ารหัสคำโดยอาศัยการพิจารณาอักษรข้างเคียงสำหรับคำ "สุรเกียรติ"

ลำดับตัวอักษร	รหัส
_ , _ , _ , s , u , r , a	s
_ , _ , s , u , r , a , k	u
_ , s , u , r , a , k , i	r
s , u , r , a , k , i , a	a
u , r , a , k , i , a , t	k
r , a , k , i , a , t , _	I
a , k , i , a , t , _ , _	_
k , i , a , t , _ , _ , _	T

รูปที่ 3.2 การเข้ารหัสคำโดยอาศัยการพิจารณาอักษรข้างเคียงสำหรับคำ “surakiat”



รูปที่ 3.3 โครงสร้างของนิวรอลเน็ตเวิร์กและการเข้ารหัสคำด้วยแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก

รายละเอียดโครงสร้างของนิวรอลเน็ตเวิร์กในแต่ละระดับชั้นมีดังต่อไปนี้

ในระดับชั้นอินพุต (Input Layer) อินพุตของนิวรอลเน็ตเวิร์ก จะได้มาจากการแทนแต่ละตัวอักษรที่พิจารณาด้วยจำนวนนิวรอนเท่ากับจำนวนอักขระในแต่ละภาษา เพราะฉะนั้นชั้นอินพุตจึงมีจำนวนนิวรอนเท่ากับผลคูณระหว่างจำนวนอักขระของภาษาคูณกับจำนวนอักษรที่ใช้พิจารณา จำนวนอักขระสำหรับกรณีคำที่สะกดด้วยตัวอักษรอังกฤษมีจำนวน 26 ตัว (a-z) จึงใช้จำนวนนิวรอนเป็น 26×7=182 นิวรอน และในกรณีคำที่สะกดด้วยตัวอักษรไทยมีจำนวน 61 ตัว (พยัญชนะและสระเดี่ยว) จึงใช้จำนวนนิวรอนเป็น 61×9=549 นิวรอน อินพุตของแต่ละนิวรอนจะ

มีค่าเป็น 0 หรือ 1 ซึ่งแทนการปรากฏของตัวอักษรแต่ละตัว และฟังก์ชันการกระตุ้น (Activation Function) ในระดับชั้นนี้เป็นฟังก์ชันแบบเชิงเส้น (Linear Function)

ในระดับชั้นซ่อน (Hidden Layer) ผู้วิจัยได้เลือกใช้จำนวนชั้นซ่อนเพียง 1 ชั้น ส่วนจำนวนนิวรอนในระดับชั้นนี้ จะได้มาจากการทดลอง ซึ่งจะเลือกจำนวนของนิวรอลจากการให้ค่าความถูกต้องสูงสุด ซึ่งการทดลองนี้จะกล่าวถึงในภายหลัง และฟังก์ชันการกระตุ้นในระดับชั้นนี้เป็นแบบซิกมอยด์ (Sigmoid Function)

ในระดับชั้นเอาท์พุต (Output Layer) มีจำนวนนิวรอน เท่ากับจำนวนรหัสเสียงที่กำหนด โดยมี 33 รหัสเสียงสำหรับคำอังกฤษ และคำไทยทับศัพท์คำอังกฤษ (ดูตารางที่ 3.1) ส่วนคำไทยและคำอังกฤษทับศัพท์คำไทยจะมี 35 รหัสเสียง (ดูตารางที่ 3.2) ส่วนการเข้ารหัสคำอังกฤษ และคำไทยทับศัพท์คำอังกฤษแบบใหม่จะมี 46 เสียง (ดูตารางที่ 3.4) และการเข้ารหัสคำไทย และคำอังกฤษทับศัพท์คำไทยจะมี 41 เสียง (ดูตารางที่ 3.5) นิวรอนที่ให้ค่าเอาท์พุตสูงสุดจะเป็นรหัสเสียงที่เป็นคำตอบ ในกรณีคำไทยนั้นมีการใช้สระลดรูป ดังนั้นในขั้นตอนของการฝึกจะกำหนดให้มีค่ารหัสเป้าหมายจำนวน 2 ตัวที่มีค่าเป็น 1 ซึ่งแทนการมีคำตอบ 2 คำตอบ ส่วนในขั้นตอนของการเข้ารหัสคำ ถ้านิวรอนที่ให้ค่าเอาท์พุตสูงสุด 2 อันดับแรกมีค่าต่างกันไม่เกิน 0.3 (ซึ่งกำหนดตามที่ระบุในงานวิจัย [9]) จะถือว่าทั้งสองนิวรอนนั้นเป็นคำตอบ และฟังก์ชันการกระตุ้นในระดับชั้นนี้เป็นแบบซิกมอยด์

ตารางที่ 3.3 สรุปจำนวนนิวรอลที่ใช้ในการทดลองในนิวรอลเน็ตเวิร์กแต่ละโครงสร้าง

ชนิดของข้อมูล ที่ใช้ในการทดลอง	จำนวน เสียง ในข้อมูลเข้า	จำนวนตัวอักษร ที่ใช้ในการพิจารณา ในแต่ละรอบ	จำนวน นิวรอล ในชั้นเข้า	จำนวนเสียง ในข้อมูลออก
คำไทยแบบเดิม	61	9	549	35
คำอังกฤษทับศัพท์คำไทยแบบเดิม	26	7	182	35
คำอังกฤษแบบเดิม	26	7	182	33
คำไทยทับศัพท์อังกฤษแบบเดิม	61	9	549	33
คำไทยแบบใหม่	61	9	549	41
คำอังกฤษทับศัพท์คำไทยแบบใหม่	26	7	182	41
คำอังกฤษแบบใหม่	26	7	182	46
คำไทยทับศัพท์อังกฤษแบบใหม่	61	9	549	46

ตารางที่ 3.4 รหัสเสียงสำหรับคำไทยทับศัพท์คำอังกฤษแบบใหม่

เสียงพยัญชนะ		รหัสเสียง		เสียงสระ		รหัสเสียง
ไทย	อังกฤษ	ตัวต้น	ตัวสะกด	ไทย	อังกฤษ	
พ	p	P	p	◌ิ ◌ี	ee, ei, ea, ey, i	i
บ	b	B	b	เ	e, ay	e
ท,ต	t, th	T	t	แ	a, air, are	w
ด	d, th	D	d	เ◌ียว	a, aw, au	@
ก,ค	c, k, g	K	k	◌ุ , ◌ู	u oo	u
ช	ch, sh	C	c	เ◌อ, เ◌ิ	ur, er, ir, or	W
จ	j, ch, g	J	j	◌ะ, ◌ั	a	a
ฟ	f, ph	F	f	โ	ome, o	o
ว	w, v	v	-	ไ, ใ, -◌ัย, -าย	ie, ai	!
ส,ซ	s, z	S	s	-าว, เ-า	ow, ou, our, au	R
ฮ	h	h	-	-วย	oi	O
ม	m	M	m	เ◌ีย	ear, ia	I
น	n	N	n	-◌ัว	our, ua	Y
ง	ng	G	g	◌ิว	ew, eua	X
ล	l	L	l	เ◌ิล	le	Q
ร	r	r	-			
ย	y	y	-			
ตัวอักษรที่ไม่ออกเสียง		_	_			

ตารางที่ 3.5 รหัสเสียงพยัญชนะสำหรับคำอังกฤษทับศัพท์คำไทยแบบใหม่

เสียงพยัญชนะ		รหัสเสียง		เสียงสระ		รหัสเสียง
ไทย	อังกฤษ	ตัวต้น	ตัวสะกด	ไทย	อังกฤษ	
ก ข ค ฆ	ck, g, k, x, c, kh, q	K	k	◌ิ ◌ี	i, ee	i
ง	ng	G	g	ะ า ◌ั	a, u, ar	a
จ ฉ ช ฌ	j, ch, x	c	t	เ–ะ เ–	e	e
ซ ส ศ ษ สร ทร	s, z	s	t	แ–ะ แ–	ae	w
ญ ย หย หญ	y	y	n	◌ึ ◌ื ◌ือ ◌ุ ◌ู	u, eu, ue, eo, oo	u
ด ฎ ฑ	d	d	t	โ–ะ โ– เ–าะ –อ	o	o
ต ฏ ถ ฐ ท ฑ ธ ฒ	t, th, dh	T	t	เ–อะ เ–อ เ◌็	er, oe	W
ณ น หน	n	N	n	เ◌ียะ เ◌ีย	ia, ie, aiu	I
บ	b	b	p	เ◌ือะ เ◌ือ –ัวะ –ัว	ua, ue, ea, ui	Y
ป ผ พ ภ	p, ph, bh	P	p	ไ– ใ– ไ–ย –ัย –าย	ai, ie, uy	!
ฝ ฟ	f	f	p	เ–า –าว	ao, ou, ow	R
ม	m	M	m	โ–ย –อย	oi, oy	x
ร ฤ	r	r	n	◌ิว	iu	X
ล ฬ ฦ	l	l	n	เ–ว	eo	q
ว	w, v	v	–	เ-ย	oei	Q
ห ฮ	h	h	–	เ◌ือย –วย	uai, uay, ou	O
ตัวอักษรที่ไม่ออกเสียง	–	–		แ-ว	aeo, eo, aew	$

ตารางที่ 3.5 (ต่อ) รหัสเสียงพยัญชนะสำหรับคำอังกฤษทับศัพท์คำไทยแบบใหม่

เสียงพยัญชนะ		รหัสเสียง		เสียงสระ		รหัสเสียง
ไทย	อังกฤษ	ตัวต้น	ตัวสะกด	ไทย	อังกฤษ	
				เ-ีย	ieo, eaw, eo, ew, iow, iau, iew, iaw	@

สำหรับตารางที่ 3.4 และตารางที่ 3.5 นั้นผู้วิจัยได้ทำการปรับปรุงมาจากตารางที่ 3.1 และตารางที่ 3.2 ตามลำดับ เนื่องจากในงานวิจัยนี้ต้องการทดลองแบ่งพยางค์ของรหัสเสียงของคำให้เป็นรหัสเสียงในระดับพยางค์ แต่ตารางที่ 3.1 และตารางที่ 3.2 มีการกำหนดรหัสเสียงที่กำกวมและซ้ำซ้อนกันระหว่างรหัสเสียงที่เป็นรหัสเสียงของพยัญชนะต้น กับรหัสเสียงของพยัญชนะที่เป็นตัวสะกด ตัวอย่างเช่นในตารางที่ 3.2 พบว่ามีรหัสเสียงที่กำกวมอยู่จำนวน 6 หน่วยเสียง คือ เสียง k g t n p และ m ผู้วิจัยจึงได้ปรับหน่วยเสียงในส่วนนี้ โดยถ้าเป็นรหัสเสียงของพยัญชนะต้นให้ใช้ตัวอักษรตัวใหญ่จำนวน 6 ตัวแทน คือ K G T N P และ M ส่วนถ้าเป็นรหัสเสียงตัวเล็กจะเป็นรหัสเสียงของตัวสะกด ดังแสดงในตารางที่ 3.5 ส่วนกรณีของหน่วยเสียงอื่นๆ ไม่มีการเปลี่ยนแปลง

3.4 วิธีการนับจำนวนพยางค์ของรหัสคำ และการแบ่งพยางค์ของรหัสคำ

สำหรับในส่วนนี้จะเป็นการอธิบายถึงวิธีการนับจำนวนพยางค์ของรหัสเสียง เนื่องจากว่าการที่เราจะตัดแบ่งพยางค์ของรหัสเสียงได้นั้น เราจำเป็นต้องทราบถึงจำนวนของพยางค์ที่จะใช้ในการแบ่งรหัสเสียงก่อน ซึ่งในขั้นตอนนี้ ผู้วิจัยพบว่าโครงสร้างของนิวรอลเน็ตเวิร์กที่ใช้ในการเข้ารหัสนั้น มีความสามารถในการจะบอกถึงจำนวนของพยางค์ของรหัสเสียงได้ โดยดูจากผลลัพธ์ที่ได้หลังจากการเข้ารหัสคำของนิวรอลเน็ตเวิร์ก ซึ่งพบว่า จำนวนของพยางค์จะมีค่าเท่ากับจำนวนของรหัสเสียงในส่วนที่เป็นรหัสเสียงสระของรหัสเสียงที่ได้ทั้งหมด ดังแสดงในรูปที่ 3.4 ที่เป็นการเข้ารหัสคำไทยคำว่า “สุรเกียรติ” โดยรหัสคำนี้พบว่ามีรหัสของเสียงสระปรากฏอยู่จำนวน 3 เสียงซึ่งตรงกับจำนวนพยางค์ที่มี 3 พยางค์ และรูปที่ 3.5 ที่เป็นการเข้ารหัสคำอังกฤษทับศัพท์คำไทยคำว่า “surakiat” โดยรหัสคำนี้พบว่ามีรหัสของเสียงสระปรากฏอยู่จำนวน 3 เสียงซึ่งตรงกับจำนวนพยางค์ที่มี 3 พยางค์เช่นกัน

ส่วนขั้นตอนวิธีที่ใช้ในการแบ่งพยางค์นั้น เนื่องจากผู้วิจัยได้ทดลองปรับปรุงตารางสำหรับการเข้ารหัสคำ จากตารางที่ 3.2 มาเป็นตารางที่ 3.5 เพื่อไม่ให้รหัสเสียงของพยัญชนะต้น รหัสเสียงของสระ และรหัสเสียงของตัวสะกดซ้ำกัน ดังนั้นถ้ารหัสเสียงทั้ง 3 แยกจากกันอย่างชัดเจน

รวมถึงทราบจำนวนของพยางค์ที่จะต้องใช้ในการแบ่งพยางค์ด้วยแล้ว เราก็สามารถที่จะแบ่งพยางค์ของรหัสคำ ให้เป็นหน่วยย่อยของรหัสเสียงของพยางค์ได้

ลำดับตัวอักษร	รหัส	จำนวนเสียงของรหัสที่เป็นสระ
_ , _ , _ , _ , ส , ุ , ร , เ , ก	s	0
_ , _ , _ , ส , ุ , ร , เ , ก , ี	u	1
_ , _ , ส , ุ , ร , เ , ก , ี , ย	r	0
_ , ส , ุ , ร , เ , ก , ี , ย , ร	a I	2
ส , ุ , ร , เ , ก , ี , ย , ร , ต	K	0
ุ , ร , เ , ก , ี , ย , ร , ต , ิ	_	0
ร , เ , ก , ี , ย , ร , ต , ิ , _	_	0
เ , ก , ี , ย , ร , ต , ิ , _ , _	_	0
ก , ี , ย , ร , ต , ิ , _ , _ , _	t	0
ี , ย , ร , ต , ิ , _ , _ , _ , _	_	0
จำนวนเสียงสระรวม		3

รูปที่ 3.4 การนับจำนวนพยางค์จากรหัสเสียงสระของรหัสคำคำว่า “สุรเกียรติ”

ลำดับตัวอักษร	รหัส	จำนวนเสียงของรหัสที่เป็นสระ
_ , _ , _ , s , u , r , a	s	0
_ , _ , s , u , r , a , k	u	1
_ , s , u , r , a , k , i	r	0
s , u , r , a , k , i , a	a	1
u , r , a , k , i , a , t	K	0
r , a , k , i , a , t , _	I	1
a , k , i , a , t , _ , _	_	0
k , i , a , t , _ , _ , _	t	0
จำนวนเสียงสระรวม		3

รูปที่ 3.5 การนับจำนวนพยางค์จากรหัสเสียงสระของรหัสคำคำว่า “surakiat”

ตารางที่ 3.6 ความแม่นยำของการนับจำนวนพยางค์จากจำนวนรหัสเสียงสระ

กรณีการทดลอง	จำนวนข้อมูลทั้งหมด (คำ)	จำนวนคำที่นับพยางค์ผิด (คำ)		ความแม่นยำในการนับพยางค์ (%)	
		ด้วยมือ	นิวรอล	ด้วยมือ	นิวรอล
คำอังกฤษ	1876	0	139	100.00	92.59
คำไทยทับศัพท์คำอังกฤษ	1876	0	39	100.00	97.92
คำไทย	2000	0	151	100.00	92.45
คำอังกฤษทับศัพท์คำไทย	2000	0	41	100.00	97.95

จากตารางที่ 3.6 ผู้วิจัยได้ลองทดสอบคำศัทพ์ที่จะนำมาใช้ในการทดลอง มาทดสอบนับจำนวนพยางค์แล้วเก็บค่าไว้ พร้อมทั้งได้ทดลองนับจำนวนพยางค์จากจำนวนของรหัสเสียงสระที่ได้หลังจากการเข้ารหัสคำ โดยการนับจำนวนพยางค์ดังกล่าว ผู้วิจัยได้ทดลองนับใน 2 กรณี โดยในกรณีที่ 1 ได้ลองทดสอบกับข้อมูลสำหรับเตรียมสอน ซึ่งเป็นข้อมูลของการเข้ารหัสคำด้วยมือ ที่เป็นรหัสคำที่ใช้สำหรับสอนให้กับนิวรอลเน็ตเวิร์ก โดยในส่วนนี้พบว่า ขั้นตอนการเข้ารหัสคำด้วยมือนั้น จะต้องมีการใช้รหัสของเสียงสระในการสอนปรากฏอยู่ในทุกๆ พยางค์ของเสียง ดังนั้นรหัสเสียงของคำที่ได้ จึงทำให้เราสามารถนับจำนวนพยางค์จากรหัสของเสียงสระที่ปรากฏอยู่ในรหัสเสียงของคำได้แม่นยำ (Accuracy) สูงถึง 100% ในข้อมูลทุกชุด และในกรณีที่ 2 ได้ลองทดสอบกับข้อมูลจริงซึ่งเป็นรหัสคำที่ได้จริงๆ จากการเข้ารหัสของนิวรอลเน็ตเวิร์กตัวที่เก่งที่สุด เราพบว่ารหัสคำที่ได้นั้น ยังคงสามารถนับจำนวนพยางค์จากจำนวนรหัสของเสียงสระได้แม่นยำสูงถึงประมาณ 92% สำหรับคำไทย และคำอังกฤษ และประมาณ 97% สำหรับคำทับศัพท์ในทั้งสองภาษา (คำอังกฤษทับศัพท์คำไทย และคำไทยทับศัพท์คำอังกฤษ)

โดยสาเหตุใหญ่ของการนับจำนวนพยางค์ที่ผิดนั้น มาจากขั้นตอนของการฝึกสอนนิวรอลเน็ตเวิร์ก ที่รหัสคำที่ได้จากการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์กนั้น มีความผิดพลาดในการเข้ารหัสเสียงที่ผิดไป จึงทำให้ได้เสียงของรหัสสระที่เกินมา หรือขาดหายไปบ้าง รวมถึงจากการทดลองผู้วิจัยยังได้สังเกตเห็นอีกสิ่งหนึ่งคือ จำนวนพยางค์ของคำที่อยู่คู่กันในทั้งสองภาษานั้นมีจำนวนของพยางค์ที่ไม่เท่ากัน ดังแสดงในตารางที่ 3.7 อันเนื่องมาจากรูปที่ใช้ในการเขียนของคำในแต่ละภาษาที่ไม่เหมือนกัน ซึ่งในส่วนนี้จึงเป็นอีกปัญหาหนึ่งที่เราจำเป็นต้องพิจารณาในช่วงของการค้นคืนข้ามภาษาที่เราจะทำการเปรียบเทียบคำข้ามภาษาเฉพาะคำที่มีจำนวนพยางค์ หรือจำนวนของรหัสเสียงสระเท่ากันอย่างเดียวไม่ได้ แต่ยังคงต้องเปรียบเทียบกับคำที่มีจำนวนพยางค์หรือจำนวนของรหัสเสียงสระที่มากกว่า หรือน้อยกว่าด้วย ซึ่งจะได้กล่าวต่อไปในบทที่ 4 ในส่วนของการค้นคืนข้ามภาษา

ตารางที่ 3.7 ความแม่นยำของจำนวนพยางค์ที่ตรงกันในคำคู่กันของทั้งสองภาษา

กรณีการทดลอง	จำนวนข้อมูลทั้งหมด (คู่)	จำนวนพยางค์ไม่ตรงกันในทั้งสองภาษา (คู่)	ความแม่นยำ (Accuracy)
คำอังกฤษและคำไทยทับศัพท์คำอังกฤษ	1876	13	99.31%
คำไทยและคำอังกฤษทับศัพท์คำไทย	2000	68	96.60%

3.5 สรุป

ในบทนี้ได้กล่าวถึงวิธีการเข้ารหัสคำ ซึ่งประกอบไปด้วยการกำหนดรหัสคำซึ่งแทนเสียงอ่านของแต่ละตัวอักษรในภาษา ดังตารางที่ 3.1 ตารางที่ 3.2 และตารางที่ 3.4 การประมวลผลเบื้องต้นสำหรับคำที่เขียนอยู่ในรูปแบบภาษาไทย เพื่อช่วยลดความซับซ้อนในการเข้ารหัสคำ รวมถึงแนวคิดในการเข้ารหัสคำโดยพิจารณาว่าปัญหาการเข้ารหัสคำนั้นเป็นปัญหาการจำแนกประเภท และอธิบายถึงวิธีการการเข้ารหัสคำโดยใช้เทคนิคของนิวรอลเน็ตเวิร์ก ขั้นตอนวิธีที่ใช้ในการนับจำนวนพยางค์ของรหัสคำจากการนับจำนวนรหัสของเสียงสระของรหัสคำ

บทที่ 4

การค้นคืนข้ามภาษา

ขั้นตอนการค้นคืนข้ามภาษาประกอบด้วยการนำรหัสคำของคำที่เป็นข้อคำถาม ไปเปรียบเทียบกับรหัสคำในดัชนีคำหลักที่ถูกเก็บไว้ในฐานข้อมูลทุกคำ โดยถ้าผลการเปรียบเทียบออกมาอยู่ในเกณฑ์ที่กำหนด (ที่ตั้งไว้และยอมรับได้) ก็จะถือว่ารหัสคำทั้งสองนั้นมีเสียงอ่านที่ตรงกัน และระบบจะค้นคืนคำนั้นกลับออกไปให้กับผู้ใช้งาน สำหรับเนื้อหาในบทนี้จะอธิบายถึงวิธีการที่ใช้ในการเปรียบเทียบรหัสคำ เกณฑ์ที่ใช้ในการเปรียบเทียบรหัสคำ

4.1 การคำนวณความต่างของรหัสคำ

เนื่องจากรหัสคำที่ได้จากขั้นตอนการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์กนั้น แม้ว่าคู่คำจากทั้งสองภาษาจะออกเสียงเหมือนกัน ก็อาจมีรหัสไม่เหมือนกันทุกตัวอักษร ดังนั้นในการเปรียบเทียบรหัสคำ จึงต้องใช้วิธีการเปรียบเทียบแบบประมาณ (Approximate Matching) ซึ่งอาศัยการคำนวณความแตกต่าง (Distance) ของรหัสคำด้วยเทคนิคระยะแก้ไขสั้นสุด (Minimum Edit Distance) ซึ่งเป็นการคำนวณความคล้ายคลึงกันระหว่างสายอักขระ 2 สาย โดยคำนวณหาจากจำนวนครั้งที่น้อยที่สุดที่ใช้ในการเพิ่ม การลบ และการแทนที่แต่ละตัวอักขระ เพื่อทำให้รหัสคำทั้งสองเหมือนกัน การคำนวณค่าความแตกต่างนี้อาศัยเทคนิคกำหนดการพลวัต (Dynamic Programming) ซึ่งวิธีการคำนวณสามารถเขียนให้อยู่ในรูปการคำนวณด้วยความสัมพันธ์เวียนเกิด Edit (P_j, W_k) ดังนี้

$$\text{Edit}\ (P_0, W_0) = 0$$

$$\text{Edit}\ (P_j, W_0) = j$$

$$\text{Edit}\ (P_0, W_k) = k$$

$$\text{Edit}\ (P_j, W_k) = \min \{ \text{Edit}\ (P_{j-1}, W_k) + 1,\ \text{Edit}\ (P_j, W_{k-1}) + 1,\ \text{Edit}\ (P_{j-1}, W_{k-1}) + R(p_j, w_k) \}$$

โดยที่

$P_j = p_1\, p_2\, p_3 \ldots p_j$ เป็นสายอักขระต้นแบบ มีความยาว j ตัวอักษร

$W_k = w_1\, w_2\, w_3 \ldots w_k$ เป็นสายอักขระเป้าหมาย มีความยาว k ตัวอักษร

$R(p_j, w_k) = 0$ ถ้า p_j เท่ากับ w_k

$= 1$ ถ้า p_j ไม่เท่ากับ w_k

4.2 เกณฑ์การเปรียบเทียบรหัสคำ

สำหรับขั้นตอนของการค้นคืนข้ามภาษา เพื่อให้ได้ค่าความเที่ยง (Precision) และค่าเรียกคืน (Recall) ที่ดี จึงได้ใช้การเปรียบเทียบรหัสคำแบบประมาณโดยได้มีการกำหนดเกณฑ์ที่จะใช้ในการเปรียบเทียบขึ้นมา สำหรับเกณฑ์ที่จะใช้ในการเปรียบเทียบรหัสคำนั้น จะหาได้จากค่าความแตกต่างระหว่างรหัสคำ โดยถ้าค่าความแตกต่างที่ได้มีค่าไม่เกินเกณฑ์ (ค่าคงที่ d) ก็จะสรุปได้ว่ารหัสคำทั้งสองเป็นรหัสคำที่มาจากคำหลักที่ตรงกันในอีกภาษา สำหรับการค้นคืนข้ามภาษาในงานวิจัยนี้ได้ลองทำการทดลองทั้งหมด 4 กรณีด้วยกัน คือ

กรณีที่ 1 ใช้การเปรียบเทียบแบบประมาณกับรหัสคำแบบดั้งเดิม (ใช้ตารางที่ 3.1 และตารางที่ 3.2 ในการเข้ารหัส) โดยในขั้นตอนนี้จะใช้นิวรอลตัวที่ให้ผลการเรียนรู้ดีที่สุดมาใช้ในการทดสอบการเข้ารหัสคำ เมื่อได้รหัสเสียงของคำของทุกตัวอักษรแล้ว จะนำรหัสเสียงทั้งหมดมาเรียงต่อกันตามลำดับเพื่อสร้างเป็นรหัสคำอ่าน โดยสุดท้ายเมื่อได้รหัสคำอ่านแล้ว จะทำการตัดรหัสที่ไม่ออกเสียง (_) ออก และทำการย้ายรหัสเสียงของสระไปต่อท้ายเสียงของพยัญชนะ ดังตัวอย่างที่ 1 ด้านล่างนี้

ตัวอย่างที่ 1 ต้องการทดสอบคำว่า “เปล่งสมบัติ” และ “plengsombut” ว่าเป็นคำทับศัพท์ที่ตรงกันหรือไม่ โดยจะทำการเข้ารหัส แล้วคำนวณหาค่าความแตกต่าง

การเข้ารหัส

เปล่งสมบัติ --> เปลงสมบัติ --> eplgsombat_ --> plgsmbteoa

plengsombut --> pleg_sombat --> plgsmbteoa

การค้นคืน

Edit(“plgsmbteoa” , ”plgsmbteoa”) = 0

จากตัวอย่างพบว่า รหัสคำทั้งสองมีค่าความแตกต่างเป็น 0 ถ้าเรากำหนดเกณฑ์มีค่าเท่ากับ 0 (d=0) ก็จะสามารถค้นคืนคำว่า “plengsombut” จากคำว่า “เปล่งสมบัติ” ได้

กรณีที่ 2 ใช้การเปรียบเทียบแบบประมาณกับการเข้ารหัสคำแบบใหม่ (ใช้ตารางที่ 3.4 ในการเข้ารหัส) ซึ่งรหัสคำอ่านในกรณีนี้จะมี 3 ส่วนคือ รหัสคำของเสียงพยัญชนะที่เป็นพยัญชนะต้น รหัสคำของเสียงสระ และรหัสคำของเสียงพยัญชนะที่เป็นตัวสะกด โดยในขั้นตอนนี้จะใช้นิวรอลตัวที่ให้ผลการเรียนรู้ดีที่สุดมาใช้ในการทดสอบการเข้ารหัสคำ เมื่อได้รหัสเสียงของคำของทุกตัวอักษรแล้ว จะนำรหัสเสียงทั้งหมดมาเรียงต่อกันตามลำดับเพื่อสร้างเป็นรหัสคำอ่าน โดยสุดท้ายเมื่อได้รหัสคำอ่านแล้ว จะทำการตัดรหัสที่ไม่ออกเสียง (_) ออก และทำการย้ายรหัสเสียงให้เรียงเป็นรหัสเสียงของพยัญชนะต้น รหัสของเสียงสระ และรหัสของเสียงตัวสะกด ดังตัวอย่างที่ 2 ด้านล่างนี้

ตัวอย่างที่ 2 ต้องการทดสอบคำว่า “รุ่งแสงมนูญ” และ “roongsangmanoon” ว่าเป็นคำทับศัพท์ที่ตรงกันหรือไม่ โดยจะทำการเข้ารหัส แล้วคำนวณหาค่าความแตกต่าง

การเข้ารหัส

รุ่งแสงมนูญ --> รุงแสงมนูญ --> rugwsgMaNun --> rsMNuwauggn

roongsangmanoon --> ru_g_swg_MaNu_n --> rsMNuwauggn

การค้นคืน

Edit(“rsMNuwauggn” , ”rsMNuwauggn”) = 0

จากตัวอย่างพบว่า รหัสคำทั้งสองมีค่าความแตกต่างเป็น 0 ถ้าเรากำหนดเกณฑ์มีค่าเท่ากับ 0 (d=0) ก็จะสามารถค้นคืนคำว่า “roongsangmanoon” จากคำว่า “รุ่งแสงมนูญ” ได้

กรณีที่ 3 ใช้การเปรียบเทียบแบบประมาณกับการเข้ารหัสคำแบบใหม่ คล้ายๆ กับกรณีที่ 2 แต่จะแตกต่างกันตรงที่จะไม่ย้ายตำแหน่งของรหัสคำอ่าน แต่จะใช้วิธีการประมวลผลรหัสคำอ่าน โดยทำการจัดวางตำแหน่งใหม่ ให้ตำแหน่งของรหัสคำอ่านเขียนเรียงเป็นพยางค์ๆ ต่อกันโดยให้ในแต่ละพยางค์มีรหัสคำ 3 ส่วนเขียนเรียงกัน คือ รหัสของเสียงพยัญชนะต้น รหัสเสียงสระ และรหัสเสียงตัวสะกด ดังตัวอย่างที่ 3 ด้านล่างนี้

ตัวอย่างที่ 3 ต้องการทดสอบคำว่า “รุ่งแสงมนูญ” และ “roongsangmanoon” ว่าเป็นคำทับศัพท์ที่ตรงกันหรือไม่ โดยจะทำการเข้ารหัส แล้วคำนวณหาค่าความแตกต่าง

การเข้ารหัส

รุ่งแสงมนูญ --> รุงแสงมนูญ --> rugwsgMaNun --> rugswgMaNun

roongsangmanoon --> ru_g_swg_MaNu_n --> rugswgMaNun

การค้นคืน

Edit(“rugswgMaNun” , ”rugswgMaNun”) = 0

จากตัวอย่างพบว่า รหัสคำทั้งสองมีค่าความแตกต่างเป็น 0 ถ้าเรากำหนดเกณฑ์มีค่าเท่ากับ 0 (d=0) ก็จะสามารถค้นคืนคำว่า “roongsangmanoon” จากคำว่า “รุ่งแสงมนูญ” ได้

กรณีที่ 4 คล้ายๆ กับกรณีที่ 3 แต่จะใช้วิธีการตัดแบ่งพยางค์ของรหัสคำอ่านให้เป็นหน่วยย่อยๆ โดยในหน่วยย่อยนั้น จะเป็นหน่วยย่อยของรหัสเสียงพยางค์ ดังตัวอย่างที่ 4 ด้านล่างนี้

ตัวอย่างที่ 4 ต้องการทดสอบคำว่า “รุ่งแสงมนูญ” และ “roongsangmanoon” ว่าเป็นคำทับศัพท์ที่ตรงกันหรือไม่ โดยจะทำการเข้ารหัส แล้วคำนวณหาค่าความแตกต่าง

การเข้ารหัส

รุ่งแสงมนูญ --> รุงแสงมนูญ --> rugwsgMaNun --> rugswgMaNun

--> rug-swg-Ma-Nun

roongsangmanoon --> ru_g_swg_MaNu_n --> rugswgMaNun

--> rug-swg-Ma-Nun

<u>การค้นคืน</u>

Edit(“rug-swg-Ma-Nun” , ”rug-swg-Ma-Nun”)

= Edit(“rug”,”rug”) + Edit(“swg”,”swg”) + Edit(“Ma”,”Ma”) + Edit(“Nun”,”Nun”)

= 0 + 0 + 0 + 0 = 0

จากตัวอย่างพบว่า รหัสคำจะถูกแยกออกมาเป็นส่วนๆ ในหน่วยของรหัสพยางค์ แล้วทำการเปรียบเทียบแบบประมาณทีละพยางค์ แล้วหาค่าผลรวมของค่าความแตกต่างๆ ในแต่ละพยางค์ ซึ่งจากตัวอย่างด้านบนพบว่า ค่าความแตกต่างของเสียงอ่านของทั้งสองคำมีค่าผลรวมของความแตกต่างเป็น 0 ถ้าเรากำหนดเกณฑ์มีค่าเท่ากับ 0 (d=0) ก็จะสามารถค้นคืนคำว่า “roongsangmanoon” จากคำว่า “รุ่งแสงมนูญ” ได้

4.3 **สรุป**

ในบทนี้ได้กล่าวถึงการค้นคืนข้ามภาษา โดยได้อธิบายถึงวิธีการคำนวณหาค่าความต่างของรหัสคำ ด้วยการเปรียบเทียบแบบประมาณด้วยเทคนิคระยะแก้ไขสั้นสุด จากนั้นได้อธิบายถึงเกณฑ์การเปรียบเทียบรหัสคำ เพื่อใช้ในการตัดสินว่ารหัสคำทั้งสองรหัสที่เปรียบเทียบกันนั้นมีเสียงอ่านที่ตรงกันหรือไม่ และได้กล่าวถึงวิธีการที่จะใช้ในการทดลองทั้ง 4 กรณีของงานวิจัยนี้

บทที่ 5

การทดลอง

5.1 วิธีการทดลอง

สำหรับการทดลองในงานวิจัยนี้ ได้ทำการเก็บรวบรวมคำศัพท์ที่จะใช้ในการทดลอง โดยแบ่งเป็นคำในกรณีคำไทยทับศัพท์คำอังกฤษ ใช้ชุดข้อมูลของคำในภาษาอังกฤษ และคำทับศัพท์ที่ตรงกัน ซึ่งเป็นคำนามเฉพาะ คำศัพท์วิทยาศาสตร์ คำศัพท์คณิตศาสตร์ และคำศัพท์เคมี จำนวน 1,876 คู่ และกรณีคำอังกฤษทับศัพท์คำไทย ได้ใช้ชื่อและชื่อสกุลทั้งภาษาไทยและภาษาอังกฤษที่ตรงกันของนิสิต จำนวน 2,000 คู่ โดยข้อมูลที่ใช้ในการทดสอบนั้นเป็นข้อมูลชุดเดียวกันกับการทดลองของทัศนวรรณ [13] (ดูตัวอย่างข้อมูลได้ที่ภาคผนวก ค) จากนั้นผู้วิจัยได้ทำการกำหนดรหัสคำอ่านของแต่ละคำศัพท์ไว้เพื่อใช้ในการฝึกสอน และเพื่อใช้เปรียบเทียบความถูกต้องเมื่อทำการทดลอง โดยการเข้ารหัสนั้นยังคงยึดหลักเกณฑ์ทางภาษาศาสตร์ใน [18] [19]

สำหรับข้อมูลที่ใช้ในการทดลองนั้นจะถูกแบ่งออกเป็นชุดข้อมูลฝึก (Training Set) และชุดข้อมูลทดสอบ (Test Set) ชุดข้อมูลฝึกจะถูกนำไปใช้ในการสอนนิวรอลเน็ตเวิร์ก เพื่อสร้างตัวเข้ารหัสคำ แล้วใช้ตัวเข้ารหัสคำที่ได้มาเข้ารหัสคำให้กับข้อมูลทั้งหมด และสุดท้ายก็จะนำรหัสคำของชุดทดสอบไปทดสอบการค้นคืน และบันทึกผลที่ได้ต่อไป

เพื่อให้การทดลองไม่โน้มเอียงกับปัญหาในการแบ่งชุดข้อมูลระหว่างชุดข้อมูลฝึกกับชุดข้อมูลทดสอบ งานวิจัยนี้จึงได้เลือกใช้วิธี K–fold Cross Validation [20] ซึ่งเป็นวิธีที่จะทำงานโดยการแบ่งชุดของข้อมูลที่ใช้ในการทดลองทั้งหมดออกเป็น K ส่วนย่อยๆ เท่าๆ กัน แล้วทำการทดลองทั้งหมดจำนวน K ครั้ง โดยในแต่ละครั้งของการทดลองนั้น จะทำการเลือกหนึ่งส่วนของข้อมูลขึ้นมาเป็นชุดข้อมูลทดสอบ ส่วนที่เหลืออีกจำนวน K–1 ส่วนจะถูกนำไปใช้เป็นข้อมูลในชุดฝึก จากนั้นนำผลการทดลองทั้งหมดที่ได้จากการทดลองทั้งหมดจำนวน K ครั้งมาหาค่าเฉลี่ย โดยในการทดลองนี้ได้แบ่งข้อมูลออกเป็นส่วนๆ โดยให้แต่ละส่วนมีคำอยู่ประมาณ 400 คู่

ตารางที่ 5.1 จำนวนข้อมูลทั้งหมด และจำนวนข้อมูลในแต่ละชุดตามค่า K

กรณีการทดลอง	จำนวนข้อมูลทั้งหมด (คู่)	ค่า K	จำนวนข้อมูลในแต่ละชุด (คู่)
คำไทยทับศัพท์คำอังกฤษ	1876	4	469
คำอังกฤษทับศัพท์คำไทย	2000	5	400

สำหรับขบวนการต่างๆ ที่ใช้ในการทดลองในงานวิจัยนี้จะประกอบไปด้วย การทดลองเพื่อหาจำนวนของนิวรอนที่อยู่ในระดับชั้นซ่อนที่เหมาะสมที่สุดสำหรับการรู้จำในแต่ละโครงสร้างของภาษา ซึ่งจะมีอยู่ทั้งหมดด้วยกัน 4 โครงสร้าง คือ โครงสร้างสำหรับการเข้ารหัสคำไทย โครงสร้างสำหรับการเข้ารหัสคำอังกฤษทับศัพท์คำไทย โครงสร้างสำหรับการเข้ารหัสคำอังกฤษ และโครงสร้างสำหรับการเข้ารหัสคำไทยทับศัพท์คำอังกฤษ ประสิทธิภาพที่ได้จากการทดลองการค้นคืนข้ามภาษาใน 4 กรณีที่ได้กล่าวไว้ในบทที่ 4 โดยการทดลองในส่วนของการค้นคืนนั้น ผู้วิจัยได้ทำการทดลองกับข้อมูลที่เป็นการเข้ารหัสด้วยมือ เพื่อให้ทราบถึงประสิทธิภาพสูงสุดที่เป็นไปได้ในการค้นคืนภาษา และได้ทดลองกับชุดข้อมูลที่เป็นการเข้ารหัสด้วยนิวรอลตัวที่เรียนรู้ดีที่สุดเปรียบเทียบกัน โดยหัวข้อต่อไปนี้จะกล่าวถึงแต่ละการทดลอง และผลการทดลองที่ได้

5.2 การเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์ก

ในการใช้งานนิวรอลเน็ตเวิร์กนั้น จำนวนนิวรอนที่อยู่ในชั้นซ่อนเป็นปัจจัยหนึ่งที่จะส่งผลต่อความแม่นยำในการรู้จำ ดังนั้นผู้วิจัยจึงได้ทำการทดลองเพื่อหาจำนวนนิวรอนที่เหมาะสมโดยกำหนดจำนวนนิวรอนต่างๆ กัน ซึ่งได้แก่ 10 50 100 150 และ 200 นิวรอลตามลำดับ และในการทดลองนี้ได้กำหนดอัตราการเรียนรู้ (Learning Rate) ของการทดลองเป็น 0.005 ค่าโมเมนตัม (Momentum) เป็น 0.95 จำนวนรอบในการฝึก (Ephoes) ของนิวรอลเน็ตเวิร์กเท่ากับ 300 รอบในแต่ละการทดลอง และกำหนดส่วนที่เพิ่มเติมขึ้นมา คือ มีการกำหนดค่าอคติ (Bias Value) มีค่าเท่ากับ 1 และกำหนดให้ลำดับในการปรับสอน (Presentation Order) เป็นการสอนแบบเรียงลำดับข้อมูลสอน ผลการทดลองดังแสดงในตารางที่ 5.2 และ 5.3 ที่จะแสดงให้เห็นว่าเมื่อใช้จำนวนนิวรอนในชั้นซ่อนเป็น 100 สำหรับคำอังกฤษ และ 200 สำหรับคำไทยทับศัพท์คำอังกฤษจะให้ค่าแม่นยำเป็น 90.65% และ 98.28% ตามลำดับ และจากผลการทดลองในตารางที่ 5.4 และตารางที่ 5.5 จำนวนนิวรอนที่ให้ความแม่นยำสูงสุดสำหรับทั้งคำไทย และคำอังกฤษทับศัพท์คำไทย คือ 150 โดยให้ความแม่นยำเป็น 90.50% และ 97.30% ตามลำดับ ทั้งนี้ค่าความแม่นยำในตารางเป็นค่าที่ได้จากการทดลองจากข้อมูลในแต่ละชุด

ตารางที่ 5.2 ความแม่นยำเมื่อใช้นิวรอนในชั้นซ่อนต่างๆ กันสำหรับข้อมูลคำอังกฤษ

จำนวนนิวรอลในระดับชั้นซ่อน	ค่าความแม่นยำ (เปอร์เซ็นต์)				
	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	เฉลี่ย
10	84.93	84.96	84.29	84.75	84.73
50	89.69	89.91	89.52	89.72	89.71
100	90.57	90.29	90.51	**90.65**	90.51
150	90.12	90.05	90.24	89.84	90.06
200	89.97	90.28	90.02	89.62	89.97

ตารางที่ 5.3 ความแม่นยำเมื่อใช้นิวรอนในชั้นซ่อนต่างๆ กัน สำหรับคำไทยทับศัพท์คำอังกฤษ

จำนวนนิวรอลในระดับชั้นซ่อน	ค่าความแม่นยำ (เปอร์เซ็นต์)				
	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	เฉลี่ย
10	92.25	91.47	91.49	92.18	91.85
50	97.28	97.68	97.33	97.57	97.47
100	97.59	97.76	97.70	97.73	97.70
150	98.01	98.10	98.12	98.13	98.09
200	98.13	**98.28**	98.12	**98.28**	98.20

ตารางที่ 5.4 ความแม่นยำเมื่อใช้นิวรอนในชั้นซ่อนต่างๆ กัน สำหรับข้อมูลคำไทย

จำนวนนิวรอลในระดับชั้นซ่อน	ค่าความแม่นยำ (เปอร์เซ็นต์)					
	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5	เฉลี่ย
10	86.79	87.32	87.12	86.81	86.88	86.98
50	89.51	89.12	89.72	88.55	89.88	89.36
100	89.48	89.91	89.76	89.68	89.90	89.75
150	89.87	**90.50**	90.04	90.11	89.94	90.09
200	88.17	87.76	87.54	88.90	87.95	88.06

ตารางที่ 5.5 ความแม่นยำเมื่อใช้นิวรอนในชั้นซ่อนต่างๆ กันสำหรับคำอังกฤษทับศัพท์คำไทย

จำนวนนิวรอลในระดับชั้นซ่อน	ค่าความแม่นยำ (เปอร์เซ็นต์)					
	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5	เฉลี่ย
10	92.61	92.89	92.90	92.86	92.47	92.75
50	96.97	96.50	96.87	96.52	96.85	96.74
100	96.85	96.71	96.85	96.44	96.79	96.73
150	97.23	97.10	**97.30**	97.06	97.21	97.18
200	96.71	96.50	96.13	96.51	96.70	96.51

ตารางที่ 5.6 สรุปค่าความแม่นยำสูงสุดเมื่อใช้จำนวนนิวรอนในชั้นซ่อนต่างๆ กัน สำหรับข้อมูลคำอังกฤษ และคำไทยทับศัพท์คำอังกฤษเทียบกัน

จำนวนนิวรอลในระดับชั้นซ่อน	ค่าความแม่นยำสูงสุด (เปอร์เซ็นต์)	
	คำอังกฤษ	คำไทยทับศัพท์คำอังกฤษ
10	84.96	92.25
50	89.91	97.68
100	**90.65**	97.76
150	90.24	98.13
200	90.28	**98.28**

ตารางที่ 5.7 สรุปค่าความแม่นยำสูงสุดเมื่อใช้จำนวนนิวรอนในชั้นซ่อนต่างๆ กัน สำหรับข้อมูลคำไทย และคำอังกฤษทับศัพท์คำไทยเทียบกัน

จำนวนนิวรอลในระดับชั้นซ่อน	ค่าความแม่นยำสูงสุด (เปอร์เซ็นต์)	
	คำไทย	คำอังกฤษทับศัพท์คำไทย
10	87.32	92.90
50	89.88	96.97
100	89.91	96.85
150	**90.50**	**97.30**
200	88.90	96.71

ตารางที่ 5.8 สรุปผลการเข้ารหัสคำจากนิวรอลตัวที่ดีที่สุดเปรียบเทียบกันกับงานวิจัยก่อนหน้า

คำศัพท์ที่ใช้ในการทดลอง	ค่าความแม่นยำสูงสุด (เปอร์เซ็นต์)	
	ทัศนวรรณ	งานวิจัยนี้
คำอังกฤษ	74.42	90.65
คำไทยทับศัพท์คำอังกฤษ	94.88	98.28
คำไทย	89.16	90.50
คำอังกฤษทับศัพท์คำไทย	93.68	97.30

5.3 วิเคราะห์ผลการทดลองการเข้ารหัสคำ

เมื่อนำผลการทดลองเปรียบเทียบในตารางที่ 5.8 มาวิเคราะห์พบว่านิวรอลเน็ตเวิร์กที่ได้จากงานวิจัยนี้มีการรู้จำสูงขึ้นในข้อมูลทุกชุด ซึ่งน่าจะเป็นเพราะว่าผู้ทดลองได้ทำการปรับเปลี่ยนเครื่องมือที่ใช้ในการทดลอง โดยเครื่องมือที่ใช้มีพารามิเตอร์ที่แตกต่างไปจากเดิม คือ ค่าอคติ ซึ่งในการทดลองนี้ได้กำหนดค่าดังกล่าวเป็น 1 สำหรับข้อมูลทั้ง 4 ชุดที่ใช้ในการทดลอง และลำดับข้อมูลที่ใช้สอนใช้แบบเรียงลำดับ

อีกสิ่งหนึ่งที่ผู้ทดลองสังเกตเห็นในขณะทำการทดลอง คือ คำที่ถูกเขียนในรูปแบบของคำภาษาไทย (คำไทยทับศัพท์คำอังกฤษ และคำไทยเอง) จะมีการกำหนดจำนวนของนิวรอลในระดับชั้นข้อมูลเข้าจำนวนเท่ากับ 61x9 = 549 นิวรอล ซึ่งในจำนวน 61 เสียงนี้ผู้วิจัยพบว่ามีบางหน่วยเสียงไม่ได้ใช้งาน เนื่องจากคำในภาษาไทยจะต้องผ่านการประมวลผลเบื้องต้น จึงทำให้มี 7 หน่วยเสียงที่ไม่ได้ถูกใช้ คือ หน่วยเสียง ฃ ฅ ฤ ฦ –ำ ไ และ ใ ดังนั้นผู้วิจัยจึงได้ทำการทดลองตัดหน่วยเสียงทั้ง 7 เสียงนี้ออก และลองทดสอบดูว่า นิวรอลยังคงรู้จำและให้ความแม่นยำสูงหรือไม่

ดังนั้นจึงทำให้จำนวนนิวรอลในระดับชั้นข้อมูลเข้าของคำที่เขียนในรูปแบบของคำภาษาไทย จากเดิมที่มีอยู่จำนวน 549 นิวรอล จะเหลือเพียง 486 นิวรอล (54 เสียง x 9 ตัวอักษรที่ใช้พิจารณา) ได้ผลการทดลองสำหรับคำไทยทับศัพท์คำอังกฤษดังตารางที่ 5.9 และตารางที่ 5.10 ซึ่งจะเห็นได้ว่าการตัดหน่วยเสียงที่ไม่ได้ใช้ไปสำหรับโครงสร้างของนิวรอลเน็ตเวิร์กที่ข้อมูลเข้าเป็นภาษาไทยนั้น ไม่ได้ทำให้ความสามารถในการรู้จำลดต่ำลงแต่อย่างใด แต่ถ้ามองกลับกันในส่วนดีจะพบว่าช่วยให้ลดจำนวนการนำเข้าข้อมูลสำหรับการสอน และช่วยให้การคำนวณนั้นเร็วขึ้น ซึ่งตรงจุดนี้ น่าจะเป็นแนวทางของการทำวิจัยต่อไปได้ในอนาคต สำหรับการลดจำนวนของรหัสเสียงในคำภาษาไทย ที่มีเสียงที่คล้ายคลึงกัน

ตารางที่ 5.9 ความแม่นยำเมื่อใช้นิวรอลในชั้นซ่อนต่างๆ กันสำหรับข้อมูลคำไทยทับศัพท์คำอังกฤษ ที่มีการตัดหน่วยเสียงของข้อมูลขาเข้าที่ไม่ได้ใช้ออก

จำนวนนิวรอลในระดับชั้นซ่อน	ค่าความแม่นยำ (เปอร์เซ็นต์)				
	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	เฉลี่ย
10	91.98	92.42	92.71	91.71	92.21
50	96.82	96.73	96.91	96.93	96.85
100	97.71	97.79	97.72	97.66	97.72
150	98.07	98.10	**98.13**	97.91	98.05
200	97.58	97.72	97.70	97.70	97.68

ตารางที่ 5.10 ความแม่นยำสูงสุดเมื่อใช้นิวรอลในชั้นซ่อนต่างๆ กันสำหรับข้อมูลคำไทยทับศัพท์คำอังกฤษ ก่อนและหลังการตัดหน่วยเสียงที่ไม่ได้ใช้ออก

จำนวนนิวรอลในระดับชั้นซ่อน	ค่าความแม่นยำสูงสุด (เปอร์เซ็นต์)	
	ข้อมูลเข้า 61 เสียง	ข้อมูลเข้า 54 เสียง
10	92.25	92.71
50	97.68	96.93
100	97.76	97.79
150	98.13	98.13
200	98.28	97.72
ค่าความแม่นยำเฉลี่ย	96.82	96.66

5.4 ผลการทดลองจากการค้นคืน

การทดลองกรณีคำไทยทับศัพท์คำอังกฤษ และคำไทยทับศัพท์คำอังกฤษ ผู้วิจัยได้นำโครงสร้างของนิวรอลเน็ตเวิร์กตัวที่เรียนรู้ดีที่สุดมาใช้ในการเข้ารหัสคำ และหลังจากผ่านกระบวนการเข้ารหัสคำ ลบรหัสที่ไม่มีเสียงออก และจัดเรียงลำดับรหัสคำใหม่ ที่ได้อธิบายขั้นตอนไว้ในบทที่ 3 เรื่องการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์ก และบทที่ 4 เรื่องการค้นคืนข้ามภาษาโดยกระบวนการในส่วนนี้จะแบ่งการทดลองออกเป็น 4 แบบ โดยการค้นคืนภาษานั้น จะนำคำแต่ละคำไปค้นคืนจากคำทั้งหมด แล้วตรวจสอบคำที่ถูกค้นคืนกลับออกมา ว่าเป็นคำเดียวกันเอง รวมถึงคำที่อยู่คู่กันข้ามภาษาหรือไม่ ทำเช่นนี้ไปเรื่อยๆ จนครบทุกคำ แล้วคำนวณหาค่าความความเที่ยง ค่าเรียกคืน และตัววัด F1 ที่ระยะห่างของค่าความแตกต่างของรหัสคำ หรือค่าของ d ที่ใช้ในการ

ทดลองไม่เท่ากัน โดยการทดลองผู้วิจัยได้ทดลองกับค่า d ที่เท่ากับ 0 กล่าวคือ เป็นการทดลองที่ระยะห่างระหว่างรหัสคำเป็น 0 หรือคือ รหัสคำทั้งสองต้องเหมือนกันทุกประการ (Exactly Matching) และยังทดลองกับค่า d ในค่าอื่นๆ อีกโดยมีค่าตั้งแต่ 1 ถึง 3

ในการทดลองผู้ทดลองได้ทำการทดลองกับรหัสคำที่ได้ทำการเข้ารหัสด้วยมือ ซึ่งรหัสคำนี้เป็นรหัสคำที่ถูกต้อง และใช้ในการสอนให้กับนิวรอลเน็ตเวิร์ก เปรียบเทียบกันกับรหัสคำที่ได้หลังจากผ่านกระบวนการการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์ก จากนั้นได้ทำการสรุปผลการทดลองเปรียบเทียบกันกับงานวิจัยก่อนหน้า ดังที่จะแสดงผลในตารางสรุปผลการทดลองต่อไป

ตารางที่ 5.11 ผลการทดลองจากการค้นคืนในแบบที่ 1
กรณีคำไทยทับศัพท์คำอังกฤษ และกรณีคำอังกฤษทับศัพท์คำไทยจากการเข้ารหัสด้วยมือ

d	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ความเที่ยง	เรียกคืน	F1	ความเที่ยง	เรียกคืน	F1
0	97.40	93.12	95.22	98.54	90.95	94.59
1	75.44	99.44	85.79	83.09	97.70	89.80
2	54.55	100.00	55.45	52.08	99.60	68.40
3	14.74	100.00	25.70	29.93	99.88	46.04

ตารางที่ 5.12 ผลการทดลองจากการค้นคืนในแบบที่ 2
กรณีคำไทยทับศัพท์คำอังกฤษ และกรณีคำอังกฤษทับศัพท์คำไทยจากการเข้ารหัสด้วยมือ

d	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ความเที่ยง	เรียกคืน	F1	ความเที่ยง	เรียกคืน	F1
0	97.70	93.04	95.31	98.73	91.10	94.76
1	78.07	99.33	87.43	83.97	97.00	90.01
2	42.40	99.97	59.55	53.64	99.20	69.63
3	16.66	100.00	28.57	30.89	99.78	47.17

ตารางที่ 5.13 ผลการทดลองจากการค้นคืนในแบบที่ 3

กรณีคำไทยทับศัพท์คำอังกฤษ และกรณีคำอังกฤษทับศัพท์คำไทยจากการเข้ารหัสด้วยมือ

d	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ความเที่ยง	เรียกคืน	F1	ความเที่ยง	เรียกคืน	F1
0	97.79	88.62	92.98	98.78	90.70	94.57
1	80.28	94.35	86.75	85.14	96.85	90.62
2	47.03	98.21	63.60	57.69	98.85	72.86
3	22.01	99.65	36.06	34.96	99.23	51.70

ตารางที่ 5.14 ผลการทดลองจากการค้นคืนในแบบที่ 4

กรณีคำไทยทับศัพท์คำอังกฤษ และกรณีคำอังกฤษทับศัพท์คำไทยจากการเข้ารหัสด้วยมือ

d	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ความเที่ยง	เรียกคืน	F1	ความเที่ยง	เรียกคืน	F1
0	94.80	88.66	91.63	97.11	90.70	93.79
1	74.48	94.19	83.18	80.62	96.11	87.69
2	44.89	97.37	61.45	54.19	97.59	69.69
3	23.75	98.56	38.28	34.84	98.03	51.41

ตารางที่ 5.15 เปรียบเทียบผลการค้นคืนทั้ง 4 แบบ (d=0) ด้วยข้อมูลจากการเข้ารหัสด้วยมือ

วิธีการ	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ความเที่ยง	เรียกคืน	F1	ความเที่ยง	เรียกคืน	F1
งานวิจัยนี้ 1	97.40	93.12	95.22	98.54	90.95	94.59
งานวิจัยนี้ 2	97.70	93.04	95.31	98.73	91.10	94.76
งานวิจัยนี้ 3	97.79	88.62	92.98	98.78	90.70	94.57
งานวิจัยนี้ 4	94.80	88.66	91.63	97.11	90.70	93.79

ตารางที่ 5.16 ผลการทดลองจากการค้นคืนในแบบที่ 1 ในกรณีคำไทยทับศัพท์คำอังกฤษ และกรณีคำอังกฤษทับศัพท์คำไทยจากการเข้ารหัสด้วยนิวรอลเน็ตเวิร์ก

d	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ความเที่ยง	เรียกคืน	F1	ความเที่ยง	เรียกคืน	F1
0	96.65	73.16	83.28	98.49	83.78	90.54
1	72.54	90.67	80.60	83.00	95.60	88.85
2	37.23	97.68	53.92	52.14	98.83	68.27
3	14.73	99.57	25.67	29.83	99.73	45.92

ตารางที่ 5.17 การเปรียบเทียบผลของตัววัด F1 จากการค้นคืนในแบบที่ 1 กรณีคำไทยทับศัพท์คำอังกฤษของงานวิจัยนี้กับงานวิจัยของทัศนวรรณ ศูนย์กลาง

d	ทัศนวรรณ	งานวิจัยนี้	
		เข้ารหัสด้วยมือ	เข้ารหัสด้วยนิวรอล
0	58.72	95.22	83.28
1	81.91	85.79	80.60
2	70.90	55.45	53.92
3	43.94	25.70	25.67

ตารางที่ 5.18 การเปรียบเทียบผลของตัววัด F1 จากการค้นคืนในแบบที่ 1 กรณีคำอังกฤษทับศัพท์คำไทยของงานวิจัยนี้กับงานวิจัยของทัศนวรรณ ศูนย์กลาง

d	ทัศนวรรณ	งานวิจัยนี้	
		เข้ารหัสด้วยมือ	เข้ารหัสด้วยนิวรอล
0	61.60	94.59	90.54
1	84.41	89.80	88.85
2	83.39	68.40	68.27
3	64.19	46.04	45.92

ตารางที่ 5.19 ผลการทดลองจากการค้นคืนในแบบที่ 2

กรณีคำไทยทับศัพท์คำอังกฤษ และกรณีคำอังกฤษทับศัพท์คำไทยจากการเข้ารหัสแบบใหม่

d	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ความเที่ยง	เรียกคืน	F1	ความเที่ยง	เรียกคืน	F1
0	97.04	68.18	80.09	98.86	67.55	80.26
1	75.56	84.99	80.00	85.83	80.20	82.92
2	41.34	95.07	57.63	56.17	91.38	69.57
3	16.69	98.67	28.55	31.79	96.85	47.87

ตารางที่ 5.20 ผลการทดลองจากการค้นคืนในแบบที่ 3

กรณีคำไทยทับศัพท์คำอังกฤษ และกรณีคำอังกฤษทับศัพท์คำไทยจากการเข้ารหัสแบบใหม่

d	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ความเที่ยง	เรียกคืน	F1	ความเที่ยง	เรียกคืน	F1
0	96.41	67.11	79.14	98.87	68.60	81.00
1	76.65	82.76	79.59	86.32	80.95	83.55
2	45.09	92.96	60.73	59.42	88.18	71.00
3	21.49	97.60	35.22	36.39	92.63	55.25

ตารางที่ 5.21 ผลการทดลองจากการค้นคืนในแบบที่ 4

กรณีคำไทยทับศัพท์คำอังกฤษ และกรณีคำอังกฤษทับศัพท์คำไทยจากการเข้ารหัสแบบใหม่

d	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ความเที่ยง	เรียกคืน	F1	ความเที่ยง	เรียกคืน	F1
0	96.46	67.24	79.25	96.94	69.10	80.69
1	77.89	81.05	79.44	81.25	80.49	80.87
2	49.88	87.78	63.61	55.52	84.28	66.94
3	28.99	90.55	43.92	35.78	86.39	50.60

ตารางที่ 5.22 ผลการค้นคืนด้วยตัววัด F1 สูงสุดของงานวิจัยนี้กับผลที่ได้จากวิธีการก่อนหน้า

วิธีการ	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ความเที่ยง	เรียกคืน	F1	ความเที่ยง	เรียกคืน	F1
ประยุทธ	72.43	90.25	80.33	92.27	76.00	83.33
ทัศนวรรณ	87.28	77.19	81.91	96.34	75.15	84.41
งานวิจัยนี้ 1	96.65	73.16	83.28	98.49	83.78	90.54
งานวิจัยนี้ 2	97.04	68.18	80.09	85.83	80.20	82.92
งานวิจัยนี้ 3	76.65	82.76	79.59	86.32	80.95	83.55
งานวิจัยนี้ 4	77.89	81.05	79.44	96.94	69.10	80.69

ตารางที่ 5.23 ผลการค้นคืนด้วยค่าความเที่ยงสูงสุดของงานวิจัยนี้กับผลที่ได้จากวิธีการก่อนหน้า

วิธีการ	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ความเที่ยง	เรียกคืน	F1	ความเที่ยง	เรียกคืน	F1
ประยุทธ	**72.43**	90.25	80.33	**92.27**	76.00	83.33
ทัศนวรรณ	**87.28**	77.19	81.91	**96.34**	75.15	84.41
งานวิจัยนี้ 1	**96.65**	73.16	83.28	**98.49**	83.78	90.54
งานวิจัยนี้ 2	**97.04**	68.18	80.09	**98.86**	67.55	80.26
งานวิจัยนี้ 3	96.41	67.11	79.14	**98.87**	68.60	81.00
งานวิจัยนี้ 4	96.46	67.24	79.25	**96.94**	69.10	80.69

5.5 วิเคราะห์ผลการทดลองการค้นคืน

จากผลการทดลองของการค้นคืนในกรณีของคำไทยทับศัพท์คำอังกฤษ และคำอังกฤษทับศัพท์คำไทยนั้น พบว่าเมื่อได้ทำการปรับปรุงการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์กใหม่ ผนวกกับการปรับสอนนิวรอลเน็ตเวิร์กด้วยพารามิเตอร์ต่างๆ ที่เปลี่ยนไปตามที่ได้บอกไว้ในบทที่ 3 จะเห็นได้ว่าผลลัพท์ที่ได้จากการค้นคืนนั้นดีขึ้นเมื่อเปรียบเทียบกับงานวิจัยก่อนหน้า ที่เกณฑ์การเปรียบเทียบความแตกต่างของรหัสคำ หรือค่า d มีค่าเท่ากับ 0 (เปรียบเทียบแบบเหมือนกันทุกประการ)

ดังนั้นจึงเป็นที่น่าสังเกตว่า การสอนนิวรอลเน็ตเวิร์กให้สามารถรู้จำได้ค่าความแม่นยำสูงๆ จะส่งผลทำให้ไม่จำเป็นต้องใช้การเปรียบเทียบแบบประมาณเข้ามาช่วย (ที่ค่าความแตกต่างมีค่าตั้งแต่ 1 ขึ้นไป) เพราะเนื่องจากโครงสร้างของนิวรอลเน็ตเวิร์กดังกล่าวสามารถทำให้รหัสคำที่ผ่านการเข้ารหัสแล้วนั้น มีรหัสคำที่ตรงกันออกมา จึงทำให้เราสามารถเปรียบเทียบรหัสคำที่ได้แบบ

เปรียบเทียบเหมือนกันทุกประการได้เลย ช่วยลดระยะเวลาที่จะต้องเสียไปสำหรับการค้นคืนอีกด้วย ส่วนการค้นคืนในแบบที่ 2 แบบที่ 3 และแบบที่ 4 นั้นผู้วิจัยได้ทดลองแล้วพบว่า ผลที่ได้จากการค้นคืนด้วยตัววัด F1 (ดูผลในตารางที่ 5.22) มีความใกล้เคียงกันกับงานวิจัยก่อนหน้า แต่ถ้าหากดูผลจากการค้นคืนด้วยค่าความเที่ยงตามตารางที่ 5.23 จะพบว่าผลการทดลองจากการค้นคืนทั้ง 4 แบบที่ผู้วิจัยได้ทำการออกแบบไว้ให้ผลการค้นคืนด้วยค่าความเที่ยงสูงถึงประมาณ 97% ในทุกการทดลอง (สูงกว่างานวิจัยก่อนหน้ามากถึง 10%) ซึ่งผลที่เกิดขึ้นนี้ เกิดขึ้นจากการเข้ารหัสแบบใหม่ ที่ลดความกำกวม และพยายามเข้ารหัสให้เสียงของทั้งสองภาษามีความใกล้เคียงกันมากที่สุด หรือมีความแตกต่างของรหัสเสียงให้ได้น้อยที่สุด รวมถึงตัวนิวรอลเน็ตเวิร์กที่เข้ารหัสแบบใหม่นั้นมีผลการเรียนรู้ที่สูง จึงส่งผลทำให้การค้นคืนข้ามภาษามีประสิทธิภาพที่สูงขึ้นด้วย

ตารางที่ 5.24 ความแม่นยำสำหรับคำอังกฤษทับศัพท์คำไทยด้วยการเข้ารหัสแบบใหม่

จำนวนนิวรอล ในระดับชั้นซ่อน	ความแม่นยำ (เปอร์เซ็นต์)					
	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5	เฉลี่ย
10	84.20	83.86	84.16	84.23	84.22	84.13
50	86.49	86.45	86.53	86.63	86.61	86.54
100	86.48	86.59	86.55	86.78	86.61	86.60
150	86.64	86.64	86.89	**86.91**	86.92	86.80
200	86.60	86.50	86.70	86.73	86.82	86.67

ตารางที่ 5.25 ความแม่นยำสำหรับคำอังกฤษด้วยการเข้ารหัสแบบใหม่

จำนวนนิวรอล ในระดับชั้นซ่อน	ค่าความแม่นยำ (เปอร์เซ็นต์)				
	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	เฉลี่ย
10	82.31	82.23	81.69	82.68	82.23
50	86.72	87.36	87.29	88.17	87.39
100	87.53	87.50	87.04	88.54	87.65
150	87.78	88.02	87.81	**88.63**	88.06
200	87.49	87.78	87.32	88.18	87.69

ตารางที่ 5.26 ความแม่นยำสำหรับคำไทยทับศัพท์คำอังกฤษด้วยการเข้ารหัสแบบใหม่

จำนวนนิวรอล ในระดับชั้นซ่อน	ค่าความแม่นยำ (เปอร์เซ็นต์)				
	ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	เฉลี่ย
10	89.81	89.42	88.45	90.04	89.43
50	96.01	96.08	96.00	95.82	95.98
100	97.30	**97.40**	97.11	97.06	97.22
150	95.94	96.31	95.83	96.22	96.08
200	97.26	96.31	97.12	96.22	96.73

5.6 สรุป

ในบทนี้ได้กล่าวถึงผลการทดลองในการใช้นิวรอลเน็ตเวิร์กในการเข้ารหัสคำ สำหรับนิวรอลเน็ตเวิร์กนั้นได้แสดงผลความแม่นยำเมื่อใช้จำนวนนิวรอนในชั้นซ่อนต่างๆ กัน เพื่อสร้างตัวเข้ารหัสคำที่ให้ค่าความแม่นยำสูงสุด หลังจากนั้นได้แสดงผลการทดลองการค้นคืนข้อมูลคำไทยทับศัพท์คำอังกฤษ และการค้นคืนข้อมูลคำอังกฤษทับศัพท์คำไทย พบว่าตัวเข้ารหัสสามารถให้การทดลองค้นคืนข้ามภาษาทั้ง 4 แบบ ที่ได้ออกแบบไว้ให้ผลการค้นคืนด้วยค่าความเที่ยงสูงถึงประมาณ 97% ในทุกการทดลอง ซึ่งเป็นค่าความเที่ยงที่สูงกว่างานวิจัยก่อนหน้าเกือบ 10% เมื่อยอมให้การค้นคืนมีค่าความแตกต่างของรหัสคำเป็น 0 หรือเป็นการค้นคืนแบบเหมือนกันทุกประการ ซึ่งดีกว่าการเข้ารหัสคำแบบเดิมที่ต้องอาศัยการเปรียบเทียบแบบประมาณเข้ามาช่วยในช่วงของการค้นคืน

บทที่ 6

สรุปผลการวิจัยและข้อเสนอแนะ

6.1 สรุปผลการวิจัย

งานวิจัยนี้เสนอวิธีการปรับปรุงการทำงานของนิวรอลเน็ตเวิร์ก เพื่อการค้นคืนข้ามภาษาสำหรับคำทับศัพท์ภาษาไทย-อังกฤษ ในการค้นคืนนั้นอาศัยรหัสคำซึ่งแทนเสียงอ่านของคำในการเปรียบเทียบคำแต่ละคู่ว่ามีเสียงอ่านตรงกันหรือไม่ ในงานวิจัยนี้ใช้วิธีการนิวรอลเน็ตเวิร์ก แต่ได้ปรับปรุงคุณภาพของการเข้ารหัสให้ดีขึ้นด้วยการปรับค่าพารามิเตอร์ต่างๆ ที่ใช้ในการทดลอง รวมถึงหลังการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์กแล้ว ได้อาศัยความรู้ทางด้านภาษาศาสตร์มาช่วยในการแบ่งรหัสเสียงที่ได้ เพื่อให้ขั้นตอนในการเปรียบเทียบนั้นทำได้ดียิ่งขึ้น สำหรับการเปรียบเทียบรหัสคำในการค้นคืนนั้น ใช้วิธีเปรียบเทียบเชิงประมาณด้วยขั้นตอนวิธีระยะแก้ไขสั้นสุด โดยในขั้นตอนวิธีนี้เราได้กำหนดต้นทุนการแก้ไขอักขระไว้ระดับเดียว ในการทดลองได้แบ่งออกเป็นการทดลองสำหรับกรณีคำไทยทับศัพท์คำอังกฤษ และกรณีคำอังกฤษทับศัพท์คำไทย จากผลการทดลองพบว่าได้ผลการค้นคืนด้วยค่าความเที่ยงสูงถึงประมาณ 97% ในการทดลองทั้ง 4 แบบ ซึ่งเป็นผลการทดลองที่ให้ค่าความเที่ยงสูงกว่างานวิจัยก่อนหน้าเกือบ 10%

6.2 ข้อเสนอแนะ

1. หากข้อมูลคำทับศัพท์ที่ใช้ในการทดลองมีจำนวน และความหลากหลายมากกว่านี้ จะให้ผลการทดลองดียิ่งขึ้น
2. สำหรับคำไทยนั้นมีการใช้สระลดรูป ได้แก่ –ะ โ-ะ และ –อ ดังนั้นเราสามารถสร้างขั้นตอนวิธีขึ้นมาสำหรับจัดการกับปัญหานี้โดยเฉพาะได้ เพื่อให้สามารถจำแนกได้ชัดเจนว่าตำแหน่งใดของคำที่มีการใช้สระลดรูป
3. เมื่อจัดการกับปัญหาของสระลดรูปได้แล้ว ดังที่ได้นำเสนอไว้ในเรื่องของรหัสเสียงของข้อมูลเข้าของนิวรอล ที่มีมากเกินจำเป็น อาจจะลดจำนวนของนิวรอลในชั้นนี้ลง เพื่อเพิ่มความรวดเร็วในการรู้จำในช่วงปรับสอน และลดเวลาการทำงานในช่วงของการใช้งานจริง
4. ลองปรับสอนนิวรอลเน็ตเวิร์กด้วยค่าพารามิเตอร์อื่นๆ ที่แตกต่างกัน เพื่อให้ผลของการรู้จำนั้นสูงขึ้น ซึ่งผู้วิจัยให้เลือกปรับจำนวนรอบของการเรียนรู้เพิ่มขึ้นจากเดิมที่ 300 รอบ เป็น 500 รอบถึง 1000 รอบ รวมถึงอาจจะหาวิธีการเข้ารหัสในรูปแบบอื่นมาปรับใช้
5. เนื่องจากการเข้ารหัสนั้นยังคงแบ่งนิวรอลสำหรับเข้ารหัสเป็น 4 โครงสร้าง ผู้ที่สนใจอาจจะทำการปรับปรุง หรือหาวิธีการที่สามารถใช้งานได้ โดยทั่วๆ ไป เพื่อให้เหลือเพียง 2 โครงสร้าง

เท่านั้น คือ โครงสร้างสำหรับข้อมูลเข้าที่เป็นภาษาไทย และโครงสร้างสำหรับข้อมูลเข้าที่เป็นภาษาอังกฤษ เพื่อความง่ายต่อการใช้งานจริง

รายการอ้างอิง

[1] O. Oard and B. Dorr, A Survey of Multilingual Text Retrieval, Technical Report UMIACS-TR-96-19 CD-TR-3615, University of Maryland, College Park, April 1996.

[2] K. Knight and J. Graehl, Machine Transliteration, Annual Meeting of the Association for Computational Linguistics (ACL-97/EACL-97).

[3] Suwanvisat P., and Prasitjutrakul, S. Thai-English Cross-Language Transliterated Word Retrieval using Soundex Technique. In Proc. of the National Computer Science and Engineering Conference 1998, Bangkok Thailand, Aug. 19-21, 1998.

[4] Suwanvisat, P., and Prasitjutrakul, S. Transliterated Word Encoding and Retrieval Algorithms for Thai–English Cross–Language Retrieval. In Proc. of the National Computer Science and Engineering Conference 1999, Bangkok Thailand, Dec. 16–17, 1999.

[5] ทัศนวรรณ ศูนย์กลาง, สมชาย ประสิทธิ์จูตระกูล, และบุญเสริม กิจศิริกุล. การเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษเพื่อการค้นคืนข้ามภาษาด้วยเทคนิคนิวรอลเน็ตเวิร์ก. ในรายงานการประชุมวิชาการทางวิทยาศาสตร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ ครั้งที่ 7 (NCSEC 2003), กรุงเทพฯ, ประเทศไทย, 16-17 ธันวาคม, 2543.

[6] Zobel, J., and Dart, P. Phonetic String Matching: Lessons from Information Retrieval. In Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996.

[7] โอภาส วงษ์ทวีทรัพย์, บุญเสริม กิจศิริกุล, และสมชาย ประสิทธิ์จูตระกูล. การเข้ารหัสและนับจำนวนพยางค์ของรหัสคำด้วยแบ็คพรอพาเกชั่นนิวรอลเน็ตเวิร์ก. ในรายงานการประชุมวิชาการทางวิทยาศาสตร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ ครั้งที่ 10 (NCSEC 2006), ขอนแก่น, ประเทศไทย, 25-27 ตุลาคม, 2549.

[8] จินดา เฮงสมบูรณ์, ภาษาศาสตร์เบื้องต้น. กรุงเทพมหานคร : สุวีริยาสาส์น, 2542.

[9] J. C. Wells, Longman Pronunciation Dictionary. Harlow, Essex : Longman, 1990.

[10] อุปกิตศิลปสาร, พระยา. หลักภาษาไทย. พิมพ์ครั้งที่ 11. กรุงเทพมหานคร: สำนักพิมพ์ไทยวัฒนาพานิช, 2545.

[11] กาญจนา นาคสกุล, ระบบเสียงภาษาไทย. กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย, 2520.

[12] อุไรรัตน์ บุญภานนท์. การถอดอักษรภาษาอังกฤษเป็นไทยโดยใช้หลักภาษาศาสตร์. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาภาษาศาสตร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิยาลัย, 2526.

[13] พัชรวรรณ คล่องอาวุธ, โปรแกรมสร้างคำสัญลักษณ์แทนเสียงสำหรับประโยคภาษาไทย. ห้องปฏิบัติการวิจัยเชี่ยวชาญเฉพาะการประมวลผลภาษาธรรมชาติ และเทคโนโลยีสารสนเทศอัจฉริยะ ภาควิชาคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์.

[14] Frakes, W.B., and Baeza Yates, R. Information Retrieval : Data Structures & Algorithms. Englewood Cliffs, N.J.: Prentice Hall, 1992.

[15] van Rijsbergen, C.J. Information Retrieval. Butterworths, London, 1979.

[16] วรรณี อุดมพาณิชย์. การใช้หลักคำพ้องเสียง เพื่อค้นหาชุดอักขระภาษาไทยที่ออกเสียงเหมือนกัน. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาภาษาศาสตร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิยาลัย, 2526.

[17] นิลเนตร อรุณวงศ์ ณ อยุธยา. การเปลี่ยนอักขระของคำในภาษาไทย โดยใช้หลักการของซาวด์เด็กซ์. ปริญญานิพนธ์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิยาลัย, 2534.

[18] หลักเกณฑ์การถอดอักษรไทยเป็นอักษรโรมันแบบถ่ายเสียง. กรุงเทพมหานคร: ราชบัณฑิตยสถาน, 2542.

[19] หลักเกณฑ์การทับศัพท์ภาษาอังกฤษ ฉบับราชบัณฑิตยสถาน. กรุงเทพมหานคร: ราชบัณฑิตยสถาน, 2532.

[20] Michell, T. M. Machine Learning. The McGraw–Hill Companies, 1997.

ภาคผนวก

ภาคผนวก ก
การใช้อักษรโรมันแทนอักขระไทย

ตารางที่ ก.1 การใช้อักษรโรมันแทนพยัญชนะไทยของราชบัณฑิตยสถาน

พยัญชนะไทย	อักษรโรมัน	
	พยัญชนะต้น	ตัวสะกด
ก	K	K
ข	KH	K
ฃ	KH	K
ค	KH	K
ฅ	KH	K
ฆ	KH	K
ง	NG	NG
จ	CH	T
ฉ	CH	T
ช	CH	T
ซ	S	T
ฌ	CH	T
ญ	Y	N
ฎ	D	T
ฏ	T	T
ฐ	TH	T
ฑ	D	T
ฒ	TH	T
ณ	TH	T
ณ	N	N
ด	D	T
ต	T	T
ถ	TH	T

พยัญชนะไทย	อักษรโรมัน	
	พยัญชนะต้น	ตัวสะกด
ธ	TH	T
น	N	N
บ	B	P
ป	P	P
ผ	PH	P
ฝ	F	P
พ	PH	P
ฟ	F	P
ภ	PH	P
ม	M	M
ย	Y	-
ร	R	N
ฤ	R	R
ล	L	N
ฦ	L	L
ว	W	-
ศ	S	T
ษ	S	T
ส	S	T
ห	H	-
ฬ	L	N
อ	H	-
ฮ	H	-

ตารางที่ ก.1 (ต่อ) การใช้อักษรโรมันแทนพยัญชนะไทยของราชบัณฑิตยสถาน

พยัญชนะไทย	อักษรโรมัน	
	พยัญชนะต้น	ตัวสะกด
ท	TH	T

พยัญชนะไทย	อักษรโรมัน	
	พยัญชนะต้น	ตัวสะกด
ทร	S	T

ตารางที่ ก.2 การใช้อักษรโรมันแทนสระไทยของราชบัณฑิตยสถาน

สระไทย	อักษรโรมัน
ะ ั า	A
ำ	AM
ิ ี ึ	I
ึ ื ุ ู	U
เ-ะ เ็ เ-	E
แ-ะ แ็ แ-	AE
โ-ะ โ- เ-าะ ็อ- -อ อ-	O
เ-อะ เ-อ เ ิ	OE
เ ียะ เ ีย	IA
เ ือะ เ ือ ัวะ ัว -ว-	UA
ใ- ไ- ไ-ย ัย -าย	AI
เ-า -าว	AO
ุย ึย	UI
โ-ย -อย	OI
ิว	IU
เ ็ว เ-ว	EO
เ-ย	OEI
เ ือย -วย	UAI
แ-ว	AEU
เ ียว	IEU

ภาคผนวก ข
หน่วยเสียงในภาษาไทยและภาษาอังกฤษ

หน่วยเสียงในภาษาไทย

ภาษาไทยมีหน่วยเสียงพยัญชนะ 21 หน่วยเสียง ดังตารางที่ ข.1 หน่วยเสียงสระ 21 หน่วยเสียง ดังตารางที่ ข.2 และหน่วยเสียงวรรณยุกต์ 5 หน่วยเสียงได้แก่ สามัญ เอก โท ตรี จัตวา

ตารางที่ ข.1 หน่วยเสียงพยัญชนะในภาษาไทย

ก	ซ ศ ษ ส ทร	ณ น หน	ม หม
ข ฃ ค ฅ ฆ	ญ ย หย หญ	บ	ร
ง หง	ฎ ด ฑ	ป	ล ฬ หล
จ จร	ฏ ต	ผ พ ภ	ว หว
ฉ ช ฌ	ฐ ฑ ถ ท ธ	ฝ ฟ	ห ฮ
			อ

ตารางที่ ข.2 หน่วยเสียงพยัญสระในภาษาไทย

อิ	แอะ	เออะ	อุ	เอาะ	อัวะ อัว
อี	แอ	เออ	อู	ออ	
เอะ	อึ	อะ	โอะ	เอียะ เอีย	
เอ	อื	อา	โอ	เอือะ เอือ	

ระบบเสียงในภาษาอังกฤษ

ภาษาอังกฤษมีหน่วยเสียงพยัญชนะ 24 หน่วยเสียง ดังตารางที่ ข.3 และหน่วยเสียงสระ 20 หน่วยเสียง ดังตารางที่ ข.4

ตารางที่ ข.3 หน่วยเสียงพยัญชนะในภาษาอังกฤษ

เสียง	คำตัวอย่าง	เสียง	คำตัวอย่าง	เสียง	คำตัวอย่าง
/ p /	pen	/ f /	fall	/ h /	how
/ b /	bad	/ v /	voice	/ m /	man
/ t /	tea	/ θ /	thin	/ n /	no
/ d /	did	/ ð /	then	/ ŋ /	sing
/ k /	cat	/ s /	so	/ l /	leg
/ g /	got	/ z /	zoo	/ r /	red
/ tʃ /	chin	/ ʃ /	she	/ j /	yes
/ dʒ /	jam	/ ʒ /	vision	/ w /	wet

ตารางที่ ข.4 หน่วยเสียงสระในภาษาอังกฤษ

เสียง	คำตัวอย่าง	เสียง	คำตัวอย่าง	เสียง	คำตัวอย่าง
/ i: /	see	/ ʊ /	put	/ aI /	tie
/ I /	sit	/ u: /	too	/ aʊ /	now
/ e /	ten	/ ʌ /	cup	/ ɔI /	join
/ æ /	hat	/ ɜ: /	fur	/ Iə /	near
/ a: /	arm	/ ə /	ago	/ eə /	hair
/ o /	got	/ eI /	page	/ ʊə /	tour
/ ɔ: /	saw	/ əʊ /	home		

ภาคผนวก ค
ตัวอย่างข้อมูลคำทับศัพท์ที่ใช้ในงานวิจัย

ตัวอย่างคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ

Liberty	ลิเบอร์ตี	gyro	ไจโร	gowland	เกาว์แลนด์
europe	ยุโรป	berkelium	เบอร์คีเลียม	anderson	แอนเดอร์สัน
aurora	ออโรรา	juice	จูซ	frisch	ฟรีซ
cytosol	ไซโทซอล	lothar	โลทาร์	bowman	โบว์แมน
playfair	เพลย์แฟร์	collenchyma	คอลเลงคิมา	helium	ฮีเลียม
zeta	ซีตา	boson	โบซอน	stokes	สโตกส์
iodide	ไอโอไดด์	einstein	ไอน์สไตน์	marconi	มาร์โคนี
okhotsk	โอค็อตสก์	chrysler	ไครส์เลอร์	warren	วาร์เรน
maclagan	แมกลาเกน	gaussian	เกาส์เซียน	delta	เดลต้า
rye	ไรย์	mil	มิล	allantois	แอลแลนทอยส์
arc	อาร์ก	broadway	บรอดเวย์	sikh	ซิก
mozambique	โมซัมบิก	thomson	ทอมสัน	hankel	ฮันเกล
suzuki	ซูซูกิ	romer	เรอเมอร์	micelle	ไมเซลล์
lantis	แลนทิส	peta	เพตะ	factor	แฟกเตอร์
dewey	ดิวอีย์	klein	ไคลน์	zygomata	ไซโกมาตา
cotangent	โคแทนเจนต์	manganese	แมงกานีส	karl	คาร์ล
cosecant	โคเซแคนต์	golf	กอล์ฟ	bract	แบร็กต์
chromosphere	โครโมสเฟียร์	harsh	ฮาร์ช	joule	จูล

ตัวอย่างคำไทยและคำอังกฤษทับศัพท์คำไทย

weerachai	วีรชัย	kijluakiat	กิจลือเกียรติ	worawas	วรวัสส์
plianrungsi	เปลี่ยนรังษี	puttakrong	พุทธกรอง	tongchai	ทองชัย
vilaiphan	วิไลพันธุ์	rossukon	รสสุคนธ์	wasana	วาสนา
sukhaboon	สุขาบูรณ์	mitisubin	มิติสุบิน	achaporn	อัชพร
prangsiri	ปรางศิริ	veerasak	วีรศักดิ์	poonpissamai	พูนพิศมัย
puengrusme	พึงรัศมี	kanokrat	กนกรัตน์	wisanupong	วิษณุพงษ์
intapuntee	อินทปันติ	sarakit	ศารกิจ	sapatporn	สภัทร์พร
onumpai	อรอำไพ	ekarat	เอกรัฐ	silarujisun	ศิลารุจิสรรค์
muendech	หมื่นเดช	siriphap	ศิริภาพ	wilailak	วิไลลักษณ์
chatkaew	ฉัตรแก้ว	khrongwong	ครองวงศ์	wichian	วิเชียร
ruethai	ฤทัย	rapeephan	รพีพรรณ	marianukroh	มารีอนุเคราะห์
ratree	ราตรี	oranuch	อรนุช	warit	วริษฐ์
jittima	จิตติมา	sawart	สวาท	pisithsak	พิสิษฐ์ศักดิ์
phamornthep	ภมรเทพ	tiawsirisup	เตียวศิริทรัพย์	laosunthara	เหล่าสุนทร
saowarat	เสาวรัตน์	charanvas	จรัญวาศน์	suproongruing	ทรัพย์รุ่งเรือง
manoon	มนูญ	kanitta	ขนิษฐา	nattaphan	ณัฐพันธุ์
keng	เก่ง	phongphew	ผ่องแผ้ว	somjate	สมเจตน์
limlawan	ลิมป์ลาวัณย์	weerasak	วีระศักดิ์	leelawat	ลีละวัฒน์
sudarat	สุดารัตน์	chumpot	จุมภฏ	kuankid	ควรคิด
suthip	สุทิพย์	charinee	ฉาริณีย์	pongput	ผองพุทธ
natakit	ณัฐกิตติ์	puttipong	พุทธิพงศ์	thansathit	ตัณสถิตย์

ประวัติผู้เขียนวิทยานิพนธ์

นายโอภาส วงษ์ทวีทรัพย์ เกิดเมื่อวันที่ 21 มิถุนายน พ.ศ. 2524 ที่จังหวัดราชบุรี สำเร็จการศึกษาระดับปริญญาวิทยาศาสตรบัณฑิต เกียรตินิยมอันดับ 2 สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร เมื่อปี พ.ศ. 2545 หลังจากนั้นได้รับทุนพัฒนาอาจารย์สาขาขาดแคลนจากภาควิชาคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร ให้ศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อ พ.ศ. 2546 มีผลงานทางวิชาการจำนวน 3 บทความ คือ บทความเรื่อง “การเข้ารหัสและนับจำนวนพยางค์ของรหัสคำ ด้วยเทคนิคนิวรอลเน็ตเวิร์ก เพื่อเพิ่มประสิทธิภาพของการค้นคืนข้ามภาษา” (Word Encoding and Syllable Enumeration of Phonetic Codes by Neural Networks for Cross-Language Retrieval Improvement) ซึ่งได้รับการตีพิมพ์ และนำเสนอในงานประชุมวิชาการสถิติประยุกต์ระดับชาติ 2549 (The National Applied Statistics Conference: ASCONF’06) เมื่อวันที่ 10-11 สิงหาคม พ.ศ. 2549

บทความเรื่อง “การเข้ารหัสและนับจำนวนพยางค์ของรหัสคำด้วยแบ็คพรอพาเกชั่นนิวรอลเน็ตเวิร์ก” (Word Encoding and Syllable Enumeration of Phonetic Codes Using by A Back-Propagation Neural Networks) ซึ่งได้รับการตีพิมพ์ และนำเสนอในงานประชุมวิชาการวิทยาการคอมพิวเตอร์ และวิศวกรรมคอมพิวเตอร์แห่งชาติ 2549 (The National Computer Science and Engineering Conference: NCSEC’06) เมื่อวันที่ 25-27 ตุลาคม พ.ศ. 2549

บทความเรื่อง “ตัวแบบสำหรับเข้ารหัสและนับจำนวนพยางค์ของรหัสคำด้วยเทคนิคโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ” (Word Encoding and Phonetic Codes Syllable Enumeration Model Using by Back-Propagation Neural Networks) ซึ่งได้รับการตีพิมพ์ และนำเสนอในงานประชุมวิชาการเทคโนโลยีสารสนเทศแห่งชาติ 2549 (The National Conference on Information Technology: NCIT’06) เมื่อวันที่ 2-3 พฤศจิกายน พ.ศ. 2549