การเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษาด้วยเทคนิควรอลเน็ตเวิร์ก

นางสาว ทัศนวรรณ ศูนย์กลาง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2543

ISBN 974-346-944-3

THAI/ENGLISH TRANSLITERATED WORD ENCODING FOR CROSS-LANGUAGE RETRIEVAL USING NEURAL NETWORKS

Miss Tasanawan Soonklang

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

Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University

Academic Year 2000
ISBN 974-346-944-3

หัวข้อวิทยานิพนธ์	การเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษาด้วยเทคนิควรอลเน็ตเวิร์ก
โดย	นางสาวทัศนวรรณ ศูนย์กลาง
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.สมชาย ประสิทธิ์จูตระกูล
อาจารย์ที่ปรึกษาร่วม	อาจารย์ ดร.บุญเสริม กิจศิริกุล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

……..คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สมศักดิ์ ปัญญาแก้ว)

คณะกรรมการสอบวิทยานิพนธ์

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

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

……..อาจารย์ที่ปรึกษาร่วม
(อาจารย์ ดร.บุญเสริม กิจศิริกุล)

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

ทัศนวรรณ ศูนย์กลาง : การเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษาด้วยเทคนิคนิวรอลเน็ตเวิร์ก. (THAI/ENGLISH TRANSLITERATED WORD ENCODING FOR CROSS-LANGUAGE RETRIEVAL USING NEURAL NETWORKS) อ.ที่ปรึกษา : ผศ.ดร.สมชาย ประสิทธิ์จูตระกูล, อ.ที่ปรึกษาร่วม : อ.ดร. บุญเสริม กิจศิริกุล, 50 หน้า. ISBN 974-346-944-3.

วิทยานิพนธ์ฉบับนี้นำเสนอขั้นตอนวิธีการเข้ารหัสคำทับศัพท์ภาษาไทย/ภาษาอังกฤษโดยใช้เทคนิคนิวรอลเน็ตเวิร์ก เพื่อการค้นคืนข้ามภาษา คำทับศัพท์ที่อนุญาตให้ใช้เป็นข้อคำถามจะเป็นคำทับศัพท์ระหว่างคำภาษาอังกฤษกับคำภาษาไทย ซึ่งจะสามารถทำการค้นคืนข้ามภาษาได้โดยไม่ต้องอาศัยพจนานุกรม

ขั้นตอนวิธีการเข้ารหัสคำทับศัพท์ ใช้นิวรอลเน็ตเวิร์กแบบแบ็กพรอพาเกชันเรียนรู้วิธีการเข้ารหัสคำ โดยรับข้อมูลขาเข้าเป็นตัวอักขระที่สนใจทีละตัวพร้อมทั้งตัวอักขระข้างเคียงหน้าหลังข้างละสี่ตัวของคำ และให้ข้อมูลขาออกเป็นรหัสเสียงของตัวอักขระขาเข้านั้น แล้วนำรหัสคำที่ได้ไปเปรียบเทียบรหัสคำแบบประมาณ ผลการทดลองแสดงให้เห็นว่า ขั้นตอนวิธีการเข้ารหัสคำที่นำเสนอให้ค่าเฉลี่ยของค่าแม่นยำและค่าเรียกคืนสูงถึง 81.91 เปอร์เซ็นต์ ในกรณีคำภาษาไทยทับศัพท์ภาษาอังกฤษ และ 84.41 เปอร์เซ็นต์ในกรณีคำภาษาอังกฤษทับศัพท์ภาษาไทย เมื่ออนุญาตให้มีความแตกต่างของรหัสที่นำมาเปรียบเทียบได้ไม่เกินหนึ่ง

ภาควิชา วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่ออาจารย์ที่ปรึกษา
ปีการศึกษา 2543	ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

4170322821 : MAJOR COMPUTER SCIENCE

KEY WORD: TRANSLITERATED WORD/ CROSS-LANGUAGE/ MACHINE LEARNING/ NEURAL NETWORK

TASANAWAN SOONKLANG : THAI/ENGLISH TRANSLITERATED WORD ENCODING FOR CROSS-LANGUAGE RETRIEVAL USING NEURAL NETWORKS. THESIS ADVISOR : ASSIST. PROF. SOMCHAI PRASITJUTRAKUL, Ph.D., THESIS CO-ADVISOR : BOONSERM KIJSIRIKUL, Ph.D., 50 pp. ISBN 974-346-944-3.

This thesis presents an algorithm for Thai-English transliterated word encoding using backpropagation neural networks for a cross-language information retrieval system. The query of Thai-English transliterated words can be cross-language retrieved without using the dictionary.

By successively feeding each character of the word along with its eight neighboring (preceeding and following) characters as the network inputs, we can obtain a sequence of phonetic codes of the word from the network output. The codes are then approximately matched with the codes of keywords in the index. Experimental results using K-fold cross validation technique showed that the average recall and precision of the Thai-to-English and English-to-Thai transliterated word cross-language retrieval are 81.91% and 84.41% , respectively with allowable edit distance of one.

Department Computer Engineering Student's signature

Fields of study Computer Science Advisor's signature

Academic year 2000 Co-Advisor's signature

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของผู้ช่วยศาสตราจารย์ ดร.สมชาย ประสิทธิ์จูตระกูล อาจารย์ที่ปรึกษาวิทยานิพนธ์ และดร.บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ซึ่งท่านทั้งสองได้ให้คำแนะนำและข้อคิดเห็นต่าง ๆ ในการวิจัยมาด้วยดีตลอด รวมทั้งตรวจแก้วิทยานิพนธ์ฉบับนี้อย่างละเอียด ผู้เขียนขอขอบพระคุณท่านอาจารย์ภาควิชาวิศวกรรมคอมพิวเตอร์ทุกท่านที่ประสิทธิประสาทวิชา

ขอขอบคุณพี่ประยุทธ สุวรรณวิสารท สำหรับแรงบันดาลใจ แนวคิดและข้อแนะนำ ขอขอบคุณสมาชิกห้องปฏิบัติการอัจฉริยภาพเครื่องกลและการค้นพบความรู้ (MIND LAB) ที่ให้ความรู้และข้อคิดเห็นที่ดี ๆ และเป็นประโยชน์ต่อการวิจัย รวมทั้งเหล่าเพื่อน พี่ น้อง ร่วมรุ่นที่ช่วยเหลือทั้งในด้านวิชาการ สันทนาการ กิจการพิเศษในระหว่างที่ศึกษาและทำงานวิจัยจนสำเร็จลุล่วงมาได้ด้วยดี

ท้ายนี้ ผู้วิจัยใคร่ขอขอบพระคุณบิดา มารดา ที่คอยสนับสนุน เฝ้ากระตุ้นถามความคืบหน้า และให้กำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันการค้นหาข้อมูลต่าง ๆ โดยใช้สื่ออิเล็กโทรนิกส์ได้รับความนิยมเป็นอย่างมาก เนื่องจากสื่อดังกล่าวนี้มีข้อมูลที่สามารถเข้าถึงได้เป็นจำนวนมากและยังมีเพิ่มมากขึ้นตลอดเวลา ทำให้สามารถค้นหาข้อมูลได้ในเกือบทุกสาขาที่สนใจ นอกจากนี้ยังมีความสะดวกอย่างมากในการค้นหาเนื่องจากสามารถเข้าถึงได้ง่ายและเข้าถึงได้จากระยะไกล โดยใช้ระบบสารสนเทศ

การค้นคืนสารสนเทศข้ามภาษา (cross-language information retrieval) หมายถึง การค้นคืนสารสนเทศซึ่งภาษาที่แสดงในเอกสารไม่ตรงกับภาษาที่แสดงในการสอบถาม[1] ปัจจุบันเอกสารทางวิชาการในประเทศไทยมักจะจัดทำทั้งในรูปภาษาไทยและภาษาอังกฤษเพื่อประโยชน์ในการเผยแพร่ทั้งภายในและภายนอกประเทศ ซึ่งเอกสารเหล่านี้โดยเฉพาะอย่างยิ่งเอกสารทางด้านวิทยาศาสตร์และวิศวกรรมศาสตร์โดยมากแล้วมักจะปรากฏคำนามเฉพาะ (proper noun) และคำศัพท์เทคนิคต่าง ๆ เป็นจำนวนมากซึ่งจะพบได้ทั้งในรูปของคำในภาษาอังกฤษ คำภาษาไทยทับศัพท์ภาษาอังกฤษ คำภาษาอังกฤษทับศัพท์ภาษาไทย หรือคำในภาษาไทยเอง ดังนั้น ถ้าระบบค้นคืนสารสนเทศไม่สนับสนุนการทำงานข้ามภาษาก็จะทำให้ประสิทธิภาพในการค้นคืนต่ำ และใช้ประโยชน์จากสารสนเทศที่มีอยู่ได้ไม่เต็มที่

ปัญหาในการค้นคืนสารสนเทศมีหลายประการด้วยกันโดยเฉพาะการค้นคืนข้ามภาษาซึ่งคำในภาษาหนึ่งอาจจะถูกเขียนในอีกภาษาหนึ่งได้หลายรูปแบบ เช่น “carbohydrate” ในภาษาไทยอาจพบได้ทั้ง “คาร์โบไฮเดรต” “คาร์โบไฮเดรท” หรือ “คาร์โบฮัยเดรต” หรือชื่อเฉพาะในภาษาไทย เช่น “ประภาส” อาจปรากฏในเอกสารที่ใช้เผยแพร่ในต่างประเทศในรูป “Prapass” หรือ “Prabhas” ซึ่งระบบค้นคืนควรจะคืนคำเหล่านี้มาทั้งหมดหรือให้ได้มากที่สุด แม้ว่าจะมีการนำพจนานุกรมสองภาษา (bilingual dictionary) มาใช้ในระบบค้นคืนสารสนเทศก็ไม่อาจแก้ปัญหานี้ได้มากนัก เนื่องจากมีคำศัพท์เทคนิคใหม่ ๆ มากมายในหลากหลายสาขาเกิดขึ้นแทบทุกวัน และ
คำทับศัพท์ส่วนมากมักไม่ปรากฏในพจนานุกรม โดยเฉพาะปัจจุบันสารสนเทศในสื่ออิเล็กโทรนิกส์มีจำนวนเพิ่มขึ้นอย่างรวดเร็วมาก ทำให้ปัญหาดังกล่าวยิ่งเพิ่มมากขึ้นจนระบบค้นคืนทั่วไปที่มีอยู่ไม่สามารถแก้ปัญหาได้

---

[1] D. 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.

การวิจัยนี้มุ่งเน้นการเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษาโดยใช้เทคนิคนิวรอลเน็ตเวิร์ก ซึ่งขั้นตอนการเข้ารหัสทั้งคำภาษาไทยทับศัพท์ภาษาอังกฤษ และคำภาษาอังกฤษทับศัพท์ภาษาไทยจะใช้แนวทางเดียวกันในการแก้ปัญหา แม้ว่าทั้ง 2 กรณีจะใช้ความรู้ของการทับศัพท์ที่แตกต่างกันในการแก้ปัญหาก็ตาม

การวิจัยนี้มีข้อสมมุติฐานว่าขั้นตอนที่นำเสนอจะสามารถทำการสืบค้นคำทับศัพท์ข้ามภาษาไทย-ภาษาอังกฤษได้โดยไม่ต้องอาศัยพจนานุกรม

1.2 วัตถุประสงค์ของการวิจัย

เพื่อออกแบบและพัฒนาวิธีเข้ารหัสคำเพื่อการค้นคืนคำทับศัพท์ข้ามภาษาไทย-อังกฤษโดยใช้การเรียนรู้ด้วยเครื่องแบบนิวรอลเน็ตเวิร์ก

1.3 ขอบเขตของการวิจัย

1. คำทับศัพท์ที่ใช้เป็นการทับศัพท์ระหว่างคำภาษาอังกฤษกับคำภาษาไทยเท่านั้น
2. คำศัพท์ในภาษาอังกฤษที่ใช้ไม่รวมถึงคำย่อ (abbreviation) และรัสพจน์ (acronym)

1.4 ขั้นตอนและวิธีการดำเนินการวิจัย

1. ศึกษาขั้นตอนวิธีการค้นคืนสารสนเทศข้ามภาษา
2. ศึกษาหลักภาษาในการถอดอักษร และหลักเกณฑ์การทับศัพท์
   2.1. จากภาษาอังกฤษเป็นภาษาไทย
   2.2. จากภาษาไทยเป็นภาษาอังกฤษ
3. ศึกษาขั้นตอนวิธีการเรียนรู้ด้วยเครื่อง
4. ออกแบบและพัฒนาวิธีการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์ก
5. ออกแบบวิธีการทดสอบขั้นตอนวิธี
6. ทดสอบและปรับปรุงคุณภาพของขั้นตอนวิธี
7. สรุปผลการวิจัย และจัดทำรายงานวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

งานวิจัยนี้สามารถนำไปใช้ในการค้นคืนคำทับศัพท์ข้ามภาษาไทย-อังกฤษทั้งคำทับศัพท์ที่ใช้กันทั่วไปในปัจจุบันและคำทับศัพท์ใหม่ ๆ ที่อาจเพิ่มขึ้นมาในอนาคตได้ อีกทั้งยังอาจใช้เป็นแนวทางในการพัฒนาขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาอื่น ๆ ต่อไป

1.6 ผลงานที่ตีพิมพ์จากงานวิจัย

วิทยานิพนธ์นี้ได้ตีพิมพ์และนำเสนอในงานประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ 2543 (The National Computer Science and Engineering Conference : NCSEC 2000) เมื่อวันที่ 16-17 พฤศจิกายน พ.ศ.2543 ในบทความเรื่อง “ขั้นตอนวิธีการเข้ารหัสคำทับศัพท์ด้วยเทคนิคนิวรอลเน็ตเวิร์ก เพื่อการค้นคืนข้ามภาษา” โดยผู้นำเสนอ คือ ทัศนวรรณ ศูนย์กลาง สมชาย ประสิทธิ์จูตระกูล และบุญเสริม กิจศิริกุล

1.7 โครงสร้างของวิทยานิพนธ์

เนื้อหาของวิทยานิพนธ์ฉบับนี้ถูกแบ่งออกเป็น 6 บทดังนี้ คือ บทที่ 1 เป็นบทนำ บทที่ 2 จะกล่าวถึงทฤษฎีและงานวิจัยต่าง ๆ ที่เกี่ยวข้อง เช่น หลักเกณฑ์การออกเสียง การเรียนรู้ด้วยเครื่องแบบนิวรอลเน็ตเวิร์ก เป็นต้น บทที่ 3 กล่าวถึงขั้นตอนวิธีการฝึกนิวรอลเน็ตเวิร์กเพื่อการเข้ารหัสคำ ส่วนบทที่ 4 จะกล่าวถึงขั้นตอนวิธีการเข้ารหัสคำทับศัพท์และขั้นตอนการค้นคืนข้ามภาษา บทที่ 5 กล่าวถึงการทดลองและผลการทดลองของขั้นตอนวิธีที่นำเสนอ และบทที่ 6 ซึ่งเป็นบทสุดท้ายจะเป็นบทสรุปของการวิจัย รวมทั้งข้อเสนอแนะต่าง ๆ ในการพัฒนาขั้นตอนวิธีการเข้ารหัสคำเพื่อการค้นคืนข้ามภาษาไทย-อังกฤษให้ดียิ่งขึ้น

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีต่าง ๆ ที่เกี่ยวกับงานวิจัยนี้ ได้แก่ หลักเกณฑ์การออกเสียง การเรียนรู้ด้วยเครื่องแบบนิวรอลเน็ตเวิร์ก

งานวิจัยที่เกี่ยวข้องและอิทธิพลต่องานวิจัยนี้ ได้แก่ ขั้นตอนวิธีการเข้ารหัสคำทับศัพท์ และขั้นตอนวิธีระยะแก้ไขสั้นที่สุด

2.1 หลักเกณฑ์การออกเสียง[1]

ระบบเสียงในทุกภาษาย่อมประกอบด้วยหน่วยเสียง (phoneme) จำนวนหนึ่งหน่วยเสียงนี้หมายถึง เสียงสำคัญในภาษาที่มีหน้าที่แยกความหมายของคำในภาษา หน่วยเสียงยังสามารถแบ่งประเภทได้เป็น หน่วยเสียงพยัญชนะ หน่วยเสียงสระ และสำหรับภาษาไทยจะมีหน่วยเสียงวรรณยุกต์ด้วย และเพื่อให้การพิจารณาศึกษาเรื่องเสียงที่ปรากฏในภาษาพูดได้สะดวก นักภาษาศาสตร์จึงได้กำหนดอักษรแทนเสียงขึ้นชุดหนึ่งเรียกว่า สัทอักษร (phonetic alphabet)

สัทอักษร คือ อักษรที่ใช้เฉพาะในทางสัทศาสตร์กำหนดขึ้นเพื่อแสดงลักษณะในการออกเสียงของภาษาอย่างเป็นสากล จึงไม่มุ่งแสดงลักษณะเฉพาะภาษาใดภาษาหนึ่ง โดยมีสมาคมสัทศาสตร์ระหว่างชาติ (the International Phonetic Association – IPA) เป็นผู้กำหนดสัทอักษรขึ้นใช้แทนเสียงพูดของมนุษย์ทั่วโลก สัทอักษรนี้ส่วนใหญ่เป็นอักษรโรมัน และมีเครื่องหมายประกอบบ้าง เพื่อให้มีตัวอักษรเพียงพอที่จะบันทึกเสียงของภาษาต่าง ๆ ได้ทั่วโลก และแม้ว่าหน่วยเสียงแต่ละหน่วยในทุก ๆ ภาษา แต่ละคนอาจออกเสียงแตกต่างกันไปเล็กน้อย แต่นักภาษาศาสตร์ยังคงถือว่าเป็นหน่วยเสียงเดียวกัน เมื่อนักภาษาศาสตร์จะทำการวิเคราะห์ภาษาใดก็สามารถเลือกสัทอักษรที่มีลักษณะของเสียงตรงกันไปใช้ สัทอักษรที่ใช้สำหรับภาษาไทยก็นำมาจากสัทอักษรของ IPA นี้เช่นกัน

ในงานวิจัยนี้ได้ทำการศึกษาหน่วยเสียงและสัทอักษรทั้งในภาษาไทยและภาษาอังกฤษเพื่อนำไปประยุกต์ใช้ในการสร้างตารางรหัสเสียง ดังต่อไปนี้

<u>ระบบเสียงในภาษาไทย</u>[2]

หน่วยเสียงภาษาไทย มี 47 หน่วยเสียง ดังนี้

---

[1] จินดา เฮงสมบูรณ์, <u>ภาษาศาสตร์เบื้องต้น</u> (กรุงเทพมหานคร : สุวีริยาสาส์น, 2542).

[2] เรื่องเดียวกัน, หน้า 116-124.

1. หน่วยเสียงพยัญชนะ มี 21 หน่วยเสียง ดังในตารางที่ 2.1 ซึ่งหน่วยเสียงพยัญชนะทั้ง 21 หน่วยเสียงมีการปรากฏในพยางค์ ดังนี้
   - เป็นพยัญชนะต้นได้ 21 หน่วยเสียง
   - เป็นพยัญชนะต้นควบได้ 9 หน่วยเสียง เรียงต่อกันได้ 12 แบบ หน่วยเสียงพยัญชนะต้นหน่วยเสียงที่ 1 ได้แก่ / p / , / ph / , / t / , / th / , / k / , / kh / หน่วยเสียงพยัญชนะต้นหน่วยที่ 2 ได้แก่ / r / , / l / และ / w /
   - เป็นพยัญชนะท้าย หรือตัวสะกดได้ 9 หน่วยเสียง ได้แก่ / p / , / t / , / k / , / m /, / n / , / ŋ / , / j / , / w / และ / ? /

ตารางที่ 2.1 หน่วยเสียงพยัญชนะในภาษาไทย

สัญลักษณ์แทนเสียง	รูปพยัญชนะ	สัญลักษณ์แทนเสียง	รูปพยัญชนะ
/ k /	ก	/ n /	ณ น หน
/ kh /	ข ฃ ค ฅ ฆ	/ b /	บ
/ ŋ /	ง หง	/ p /	ป
/ c /	จ จร	/ ph /	ผ พ ภ
/ ch /	ฉ ช ฌ	/ f /	ฝ ฟ
/ s /	ซ ศ ษ ส ทร	/ m /	ม หม
/ j /	ญ ย หย หญ	/ r /	ร
/ d /	ฎ ด ฑ	/ l /	ล ฬ หล
/ t /	ฏ ต	/ w /	ว หว
/ th /	ฐ ฑ ฒ ถ ท ธ ทร	/ h /	ห ฮ
		/ ? /	อ

2. หน่วยเสียงสระ มี 21 หน่วยเสียง ดังแสดงในตารางที่ 2.2 ซึ่งจำแนกเป็น 2 ชนิด ดังนี้
   - สระเดี่ยว มี 18 หน่วยเสียง ได้แก่ / i / , / ii / , / e / , / ee / , / ɛ / , / ɛɛ / , / ɯ /, / ɯɯ / , / a / , / aa / , / u / , / uu / , / o / , / oo / , / ɔ / , / ɔɔ / , / ɔ / และ / ɔɔ /
   - สระประสม มี 3 หน่วยเสียง ได้แก่
     หน่วยเสียง / ia / มีเสียงย่อยเป็น / ia / และ / iia /
     หน่วยเสียง / ɯa / มีเสียงย่อยเป็น / ɯa / และ / ɯɯa /
     หน่วยเสียง / ua / มีเสียงย่อยเป็น / ua / และ / uua /

ตารางที่ 2.2 หน่วยเสียงสระในภาษาไทย

สัญลักษณ์แทนเสียง	รูปสระ	สัญลักษณ์แทนเสียง	รูปสระ
/ i /	อิ	/ a /	อะ
/ ii /	อี	/ aa /	อา
/ e /	เอะ	/ u /	อุ
/ ee /	เอ	/ uu /	อู
/ ɛ /	แอะ	/ o /	โอะ
/ ɛɛ /	แอ	/ oo /	โอ
/ ɯ /	อึ	/ ɔ /	เอาะ
/ ɯɯ /	อื	/ ɔɔ /	ออ
/ ə /	เออะ	/ ia / , / iia /	เอียะ, เอีย
/ əə /	เออ	/ ɯa / , / ɯɯa /	เอือะ, เอือ
		/ ua / , / uua /	อัวะ, อัว

3. หน่วยเสียงวรรณยุกต์ มี 5 หน่วยเสียง ดังนี้
   - หน่วยเสียงวรรณยุกต์สามัญ
   - หน่วยเสียงวรรณยุกต์เอก
   - หน่วยเสียงวรรณยุกต์โท
   - หน่วยเสียงวรรณยุกต์ตรี
   - หน่วยเสียงวรรณยุกต์จัตวา

ระบบเสียงในภาษาอังกฤษ[3]

หน่วยเสียงภาษาอังกฤษ มี 44 หน่วยเสียง ดังนี้

1. หน่วยเสียงพยัญชนะ มี 24 หน่วยเสียง ดังแสดงในตารางที่ 2.3
2. หน่วยเสียงสระ มี 20 หน่วยเสียง ดังตารางที่ 2 4

[3] J. C. Wells, Longman Pronunciation Dictionary (Harlow,Essex : Longman, 1990).

ตารางที่ 2.3 หน่วยเสียงพยัญชนะในภาษาอังกฤษ

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

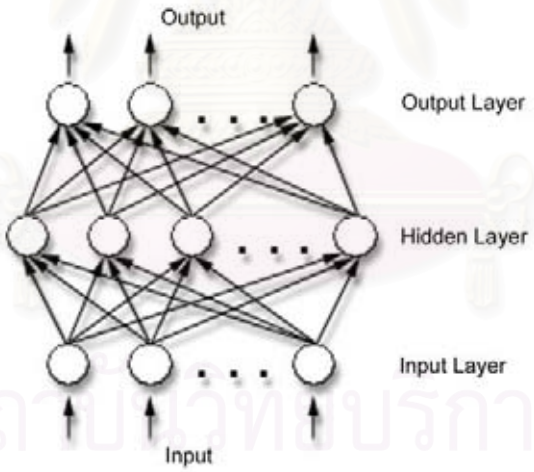
ตารางที่ 2.4 หน่วยเสียงสระในภาษาอังกฤษ

สัญลักษณ์แทนเสียง	คำตัวอย่าง	สัญลักษณ์แทนเสียง	คำตัวอย่าง
/ i: /	see	/ ɜ: /	fur
/ ɪ /	sit	/ ə /	ago
/ e /	ten	/ eɪ /	page
/ æ /	hat	/ əʊ /	home
/ ɑ: /	arm	/ aɪ /	tie
/ ɒ /	got	/ aʊ /	now
/ ɔ: /	saw	/ ɔɪ /	join
/ ʊ /	put	/ ɪə /	near
/ u: /	too	/ eə /	hair
/ ʌ /	cup	/ ʊə /	tour

2.2 การเรียนรู้ด้วยเครื่องแบบนิวรอลเน็ตเวิร์ก[4 5]

นิวรอลเน็ตเวิร์ก เป็นการเรียนรู้ด้วยเครื่อง (machine learning) รูปแบบหนึ่ง ซึ่งมีแนวคิดในการทำงานโดยการจำลองการทำงานบางส่วนของสมองมนุษย์ ที่ประกอบด้วยนิวรอลจำนวนมากเชื่อมต่อกัน โดยนิวรอลเน็ตเวิร์กจะจำลองให้มีนิวรอลจำนวนหนึ่งซึ่งเชื่อมต่อถึงกัน โดยมีค่าน้ำหนักของแต่ละการเชื่อมต่อ เมื่อมีการให้ตัวอย่างที่ใช้ในการเรียนรู้ นิวรอลเน็ตเวิร์กก็จะปรับค่าน้ำหนักให้เหมาะสม จนได้ผลลัพธ์ที่ถูกต้องหรือมีข้อผิดพลาดน้อยที่สุด และสามารถนำค่าน้ำหนักนี้ไปใช้ในงานที่ต้องการได้ นิวรอลเน็ตเวิร์กเหมาะสมสำหรับการนำไปใช้ในการแก้ปัญหาที่เกี่ยวข้องกับการจำแนกหรือแบ่งประเภท ในงานวิจัยนี้จึงนำนิวรอลเน็ตเวิร์กมาใช้ในการสร้างรหัสคำอ่านของคำทับศัพท์

งานวิจัยนี้จะใช้แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก (backpropagation neural network) ซึ่งเป็นนิวรอลเน็ตเวิร์กแบบหลายชั้นที่ใช้ขั้นตอนวิธีแบ็กพรอพาเกชัน (the backpropagation algorithm) ซึ่งในขั้นตอนการทำงานจะไม่มีการป้อนผลลัพธ์ที่ได้ในแต่ละโหนดย้อนกลับไปยังโหนดที่ส่งข้อมูลมาให้ (feed forward) โครงสร้างของแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กประกอบด้วยชั้นอินพุต (input layer) ชั้นซ่อน (hidden layer) และชั้นเอาต์พุต (output layer) แสดงดังรูปที่ 2.1 โดยจำนวนชั้นซ่อนสามารถมีได้มากกว่า 1 ชั้น



รูปที่ 2.1 โครงสร้างแบ็กพรอพาเกชันเน็ตเวิร์ก

ในแต่ละโหนดของนิวรอลเน็ตเวิร์กแบบหลายชั้นจะให้ค่าผลลัพธ์ ตามสมการที่ 2.1

$$o = \sigma(\vec{w} \cdot \vec{x}) \quad \text{..... (2.1)}$$

---

[4] E. Rich and K. Knight, Artificial Intelligence (Singapore : Prentice–Hill, 1991).

[5] T. M. Michell, Machine Learning (The McGraw-Hill Companies, Inc., 1997), pp. 95-112.

โดย σ เป็นฟังก์ชันกระตุ้น (activation function) ซึ่งนิยมใช้ฟังก์ชันซิกมอยด์ (sigmoid function) ตามสมการที่ 2.2

$$\sigma(y) = \frac{1}{1+e^{-y}} \quad \text{..... (2.2)}$$

เมื่อ o คือเอาต์พุต

$\vec{x}$ คืออินพุต

$\vec{w}$ คือค่าน้ำหนักของอินพุตนั้นๆ

ส่วนขั้นตอนวิธีแบ็กพรอพาเกชัน จะเป็นการเรียนรู้เพื่อปรับค่าน้ำหนักสำหรับนิวรอลเน็ตเวิร์กแบบหลายชั้น โดยที่ค่าน้ำหนักที่ได้จะเป็นค่าน้ำหนักที่ทำให้ค่าผลต่างกำลังสองที่น้อยที่สุด ระหว่างเอาต์พุตที่ได้จากเน็ตเวิร์กและค่าเป้าหมายของอินพุต โดยมีขั้นตอนสำหรับการปรับน้ำหนักดังนี้

กำหนดให้ตัวอย่างที่ใช้ในการเรียนรู้แต่ละตัวอย่างอยู่ในรูป $(\vec{x}, \vec{t})$

เมื่อ $\vec{x}$ เป็นเวกเตอร์ของอินพุตของเน็ตเวิร์ก

$\vec{t}$ เป็นเวกเตอร์ของเป้าหมายของเอาต์พุตของเน็ตเวิร์ก

η เป็นค่าอัตราการเรียนรู้ (learning rate)

x_{ji} เป็นอินพุตขององค์ประกอบ *j* ซึ่งมาจากองค์ประกอบ *i*

w_{ji} เป็นค่าน้ำหนักขององค์ประกอบ *j* ซึ่งมาจากองค์ประกอบ *i*

1. สร้างนิวรอลเน็ตเวิร์กตามโครงสร้างที่ต้องการ
2. กำหนดจำนวนนิวรอลของแต่ละชั้น
3. กำหนดค่าน้ำหนักเริ่มต้นแบบสุ่มให้มีค่าน้อยๆ (เช่น ระหว่าง -0.05 ถึง 0.05)
4. ทำการปรับค่าน้ำหนักด้วยขั้นตอนวิธีดังนี้

   สำหรับ $(\vec{x}, \vec{t})$ แต่ละตัว ให้ทำดังนี้

   - อินพุต $\vec{x}$ ในเน็ตเวิร์ก และคำนวณเอาต์พุต O_u ในโหนด *u* ทุกโหนด
   - คำนวณค่าความผิดพลาด δ_k ของแต่ละโหนดเอาต์พุต *k* โดยที่

$$\delta_k = o_k(1-o_k)(t_k - o_k) \quad \text{.....(2.3)}$$

- คำนวณค่าความผิดพลาด δ_h ของแต่ละโหนดที่ถูกซ่อน *h* โดยที่

$$\delta_h = o_h(1 - o_h) \sum_{k \in outputs} w_{kh}\delta_k \quad \text{..... (2.4)}$$

- ทำการปรับค่าน้ำหนัก w_{ji} โดย

$$w_{ji} = w_{ji} + \Delta w_{ji} \quad \text{..... (2.5)}$$

เมื่อ $w_{ji} = \eta\delta_j x_{ji}$

2.3 ขั้นตอนวิธีการเข้ารหัสคำทับศัพท์[6]

ประยุทธ สุวรรณวิสารท ได้ออกแบบขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ โดยขั้นตอนวิธีที่นำเสนอแบ่งเป็น 2 ส่วน คือ

1. ขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามแบบภาษาไทยทับศัพท์ภาษาอังกฤษ เป็นการนำขั้นตอนวิธีซาวด์เด็กซ์ภาษาอังกฤษของ Odell และ Russell[7] มาดัดแปลงเพียงเล็กน้อย โดยการเปลี่ยนแปลงตารางการกำหนดรหัสคำให้ใช้กับตัวอักษรไทยและไม่จำกัดความยาวของรหัสที่ได้ ดังแสดงในตารางที่ 2.5

หลักเกณฑ์การสร้างรหัสซาวด์เด็กซ์มีดังนี้

- ใช้รหัสตัวเลขทั้งหมดในการเข้ารหัสคำ และสำหรับรหัส 7 8 9 ใช้ในกรณีที่เป็นตัวอักษรตัวแรกของคำเท่านั้น
- ไม่จำกัดความยาวของรหัสคำ
- ไม่นำวรรณยุกต์ในภาษาไทย และสระทั้งในภาษาไทยและภาษาอังกฤษมาพิจารณาในการเข้ารหัสคำ

ตัวอย่างการเข้ารหัสคำทับศัพท์แบบภาษาไทยทับศัพท์ภาษาอังกฤษ

คลินตัน → 24535

CLINTON → 24535

---

[6] ประยุทธ สุวรรณวิสารท, "การเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ" (วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2541).

[7] A. Binstock and J. Rex, Practical Algorithms for Programmers (New York : Addison Wesley, 1995), pp.157-172.

การค้นคืนทำได้โดยทำการเข้ารหัสคำในข้อคำถาม แล้วนำรหัสคำที่ได้ไปค้นหาจากดัชนีคำหลักของเอกสารที่ได้เข้ารหัสไว้แล้วในขั้นตอนการทำดัชนี โดยใช้การเปรียบเทียบแบบเหมือนกันทุกประการ (exact matching) ถ้ามีรหัสคำที่เหมือนกันจะถือว่าคำหลักนั้นเป็นคำหลักที่ตรงกันในอีกภาษาหนึ่ง

ผลการทดลองจากคำศัพท์ทั้งหมด 1,689 คู่พบว่า จะได้ประสิทธิผลสูงสุด คือ ค่าแม่นยำ 78 เปอร์เซ็นต์ และค่าเรียกคืน 90 เปอร์เซ็นต์ เมื่อใช้คำทับศัพท์ที่มีความยาวมากกว่า 7 ตัวอักษรขึ้นไปในการค้นคืนข้ามภาษาไทย-อังกฤษ

ตารางที่ 2.5 การกำหนดรหัสซาวด์เด็กซ์สำหรับอักษรไทยและอักษรอังกฤษ

ภาษาอังกฤษ	ภาษาไทย	รหัส	หมายเหตุ
A E I O U H W Y	อ ห ฮ ย ญ	0	#2
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 ช ฌ ฎ 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	อ	7	#1
H	ห ฮ	8	#1
W	ว	1	#1
Y	ย ญ	9	#1
	ง	52	

#1 : สำหรับตัวอักษรแรก (ซ้ายสุด) ของคำเท่านั้น

#2 : สำหรับตัวอักษรตั้งแต่ตัวที่ 2 เป็นต้นไปเท่านั้น

2. ขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามแบบภาษาอังกฤษทับศัพท์ภาษาไทย เป็นการนำหลักภาษาไทยมาช่วยในการออกแบบการประมวลผลตัวอักษรเบื้องต้นเพื่อให้คำศัพท์ทั้งสองภาษามีรูปแบบที่เหมือนกันก่อนที่จะนำไปเข้ารหัสคำ และใช้เทคนิคระยะแก้ไขการอ่านออกเสียงสั้นที่สุด (minimal phonetic edit distance) มาดัดแปลงให้เหมาะสมกับการอ่านออกเสียงของภาษาไทยเพื่อใช้ในการเปรียบเทียบกับคู่รหัสคำที่ได้ พร้อมทั้งออกแบบ

เงื่อนไขที่ใช้ทดสอบรหัสคำทั้งสองว่าเป็นรหัสคำที่ได้มาจากคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษ ซึ่งจะอธิบายแต่ละขั้นตอนให้ละเอียดขึ้นดังต่อไปนี้

ขั้นตอนวิธีการเข้ารหัสคำ จะประกอบด้วย 2 ขั้นตอนย่อย คือ (1) ขั้นตอนการประมวลผลตัวอักษรเบื้องต้น และ (2) การย้ายตำแหน่งสระ การประมวลผลตัวอักษรเบื้องต้น ในกรณีที่ข้อคำถามเป็นภาษาไทย จะประกอบด้วยการลดรูป ตัดวรรณยุกต์ และการแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล ส่วนในกรณีที่ข้อคำถามเป็นภาษาอังกฤษทับศัพท์ภาษาไทยจะต้องทำการประมวลผลตัวอักษรเบื้องต้นโดยทำการถอดอักษรเพื่อเปลี่ยนพยัญชนะอังกฤษเป็นพยัญชนะไทย ส่วนสระอังกฤษจะถอดเป็นสระไทย แต่ถ้าสระไทยนั้นเป็นสระประสมหรือสระเดี่ยวที่ใช้อักขระตั้งแต่ 2 ตัวขึ้นไปจะใช้สัญลักษณ์เสียงสากลหนึ่งตัวแทนเสียงสระดังกล่าว โดยนำเสนอตารางที่ใช้ในการถอดอักษรดังแสดงในตารางที่ 2.6 และ 2.7

ตารางที่ 2.6 การถอดอักษรอังกฤษเป็นอักษรไทยในส่วนของพยัญชนะ

อักษรอังกฤษ	อักษรไทย	อักษรอังกฤษ	อักษรไทย
B	บ	N	น (ณ)
BH	พ	NG	ง
C	ซ	P	ป
CH	ช (ฉ ฌ)	PH	พ (ผ ภ)
CK	ก	Q	ค
D	ด (ฎ)	R	ร (ฤ)
DH	ท	S	ส (ซ ศ ษ)
F	ฟ (ฝ)	T	ต (ฏ)
G	ก	TH	ท (ฐ ฑ ฒ ถ ธ)
H	ห (ฮ)	V	ว
J	จ	W	ว
K	ก	X	ก
KH	ข (ฃ ค ฅ ฆ)	Y	ย (ญ)
L	ล (ฦ ฬ)	Z	ซ
M	ม		

หลังจากคำภาษาอังกฤษทับศัพท์ภาษาไทยและคำภาษาไทยผ่านขั้นตอนการประมวลผลตัวอักษรเบื้องต้นแล้วจะได้คำทับศัพท์ที่อยู่ในรูปแบบเดียวกัน คือ อักษรไทยผสมกับสัญลักษณ์เสียงสากล จากนั้นจะทำการย้ายตำแหน่งสระ โดยจะย้ายสระไปข้างหลังสายอักขระตามลำดับ

ตารางที่ 2.7 การถอดอักษรอังกฤษเป็นอักษรไทยในส่วนของสระ

ตัวอักษร	ถอดอักษรเป็น
-A	= ะ
-AA	= า
-AE	= x (แ-ะ แ-)
-AI	= ั-ย
-AO	= @ (เ-า)
-AIU	= I (เ-ีย)
-ARN	= าน
-ART	= าท
-E	= เ (เ-ะ เ-)
-EE	= -ี
-EO	= แ-ว
-ER	= q (เ-อ เ-ิ) หลัง R ต้องไม่เป็นสระ
-EU	= -ึ
-I	= -ิ
-IA	= I (เ-ียะ เ-ีย)
-IE	= I (เ-ียะ เ-ีย)
-O	= อ (-อ)
-OE	= q (เ-อ เ-ิ)
-OI	= อย
-OO	= -ู
-ORN	= ร (-อน)
-U	= -ุ (-ึ-ึ - -ู)
-UA	= U (เ-ือะ เ-ือ ั-วะ ั-ว)
-UE	= -ื

ในขั้นตอนการเปรียบเทียบรหัสคำ จะนำรหัสคำทั้งสองภาษามาทำการคำนวณหาค่าความแตกต่างของรหัสคำด้วยเทคนิคระยะแก้ไขเสียงอ่านสั้นที่สุด ซึ่งการคำนวณหาค่าความแตกต่างของคำจะได้จากการคำนวณต้นทุนน้อยที่สุดในการแก้ไขอักขระให้คำทั้งสองเหมือนกัน โดยพิจารณาความแตกต่างกันทางเสียงของอักขระแทนรูปของอักขระ และสร้างเป็นตารางการกำหนดต้นทุนในการแทนที่อักขระในส่วนของพยัญชนะและสระ จากนั้นนำค่าความแตกต่างที่ได้มาเข้าสมการเงื่อนไขตามสมการที่ 2.6 และ 2.7

$$Edit(P_m^c, W_n^c) \leq \alpha \times Max(Len(P_m^c), Len(W_m^c)) \times C_4 \quad \text{..... (2.6)}$$

$$Edit(P_m^v, W_n^v) \leq \alpha \times Max(Len(P_m^v), Len(W_m^v)) \times C_4 \quad \text{..... (2.7)}$$

โดยที่

P_m^c	คือ รหัสคำ P เฉพาะในส่วนของพยัญชนะ มีความยาว m ตัวอักษร
P_m^v	คือ รหัสคำ P เฉพาะในส่วนของสระ มีความยาว m ตัวอักษร
$Edit(P,W)$	ค่าความแตกต่างของรหัสคำ P กับรหัสคำ W
α	คือ พารามิเตอร์ที่ปรับประสิทธิผล มีค่าระหว่าง 0-1
$Max(j,k)$	คือ ฟังก์ชันที่จะเลือกค่าที่มากที่สุดระหว่าง j กับ k
$Len(P)$	ความยาวของรหัสคำ P
C_4	คือ ต้นทุนในการแก้ไขอักขระที่มากที่สุด

ถ้ารหัสคำที่ได้จากคำหลักภาษาไทยกับภาษาอังกฤษทับศัพท์ภาษาไทยนั้นตรงกัน สมการเงื่อนไขต้องเป็นจริงทั้งส่วนของพยัญชนะและสระ

ตัวอย่างการทดสอบคำว่า “CHULERTTIYAWONG” และ “ชูเลิศติยวงศ์” เป็นคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษหรือไม่

<u>การเข้ารหัสคำ</u>

CHULERTTIYAWONG → ชุลqตตียะวอง → ชลตตยวอง ุq ิ ะ

ชูเลิศติยวงศ์ → ชูเลิศติยวง → ชูลqศติยวง → ชลศตยวง ูq ิ

<u>การคำนวณหาค่าความแตกต่าง</u>

ส่วนพยัญชนะของรหัสคำ $Edit($“ชลตตยวอง”, “ชลศตยวง”$) = 4$

ส่วนสระของรหัสคำ $Edit($“ ุq ิ ะ”, “ ูq ิ”$) = 1$

<u>การทดสอบเงื่อนไขในการเปรียบเทียบรหัสคำ</u>

กำหนดให้ α = 0.15

$C_4 = 7$

$$Edit(\text{“ชลตตยวอง”}, \text{“ชลศตยวง”}) \le 0.15 \times Max\,(Len(\text{“ชลตตยวอง”}),$$
$$Len(\text{“ชลศตยวง”})) \times 7$$

$$Edit(\text{“ุqะ”}, \text{“ูqะ”}) \le 0.15 \times Max(Len(\text{“ุqะ”}), Len(\text{“ูqะ”})) \times 7$$

$4 \le 8.4$ และ

$1 \le 4.2$

จากตัวอย่างพบว่า $4 \le 8.4$ และ $1 \le 4.2$ เป็นจริง เพราะฉะนั้นขั้นตอนวิธีสรุปคำศัพท์ทั้งสองคำเป็นคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษ

การค้นคืนทำได้โดยทำการเข้ารหัสคำในข้อคำถาม แล้วนำรหัสคำที่ได้ไปค้นหาจากดัชนีคำหลักของเอกสารที่ได้เข้ารหัสไว้แล้วในขั้นตอนการทำดัชนี โดยใช้การเปรียบเทียบดังเงื่อนไขที่ได้กล่าวไปแล้ว ถ้าเงื่อนไขเป็นจริงจะถือว่าคำหลักนั้นเป็นคำหลักที่ตรงกันในอีกภาษาหนึ่ง

ผลการทดลองจากคำศัพท์ทั้งหมด 5,000 คู่พบว่า จะได้ประสิทธิผลสูงสุด คือค่าแม่นยำ 69 เปอร์เซ็นต์ และค่าเรียกคืน 73 เปอร์เซ็นต์ เมื่อกำหนดค่าแอลฟาเท่ากับ 0.15

2.4 **ขั้นตอนวิธีแก้ไขระยะสั้นที่สุด** (minimal edit distance)

วิธีการวัดความคล้ายคลึงกันระหว่าง 2 สายอักขระมีอยู่ด้วยกันหลายวิธี[8] แต่ละวิธีก็จะมีการกำหนดฟังก์ชันในการคำนวณหาระยะห่างระหว่าง 2 สายอักขระ (distance function, *d*) ที่แตกต่างกันไป ซึ่งต้องมีคุณสมบัติดังต่อไปนี้

$d(s1,s1) = 0$,

$d(s1,s2) \ge 0$,

$d(s1,s3) \le d(s1,s2) + d(s2,s3)$

สำหรับฟังก์ชันที่นิยมใช้ มี 2 ฟังก์ชันหลักดังนี้

1) ระยะแฮมมิ่ง (hamming distance) กำหนดมาเพื่อใช้สำหรับคู่สายอักขระที่มีความยาวเท่ากัน โดยฟังชัน *d* ได้จากการนับจำนวนของสัญลักษณ์ที่แตกต่างกันในตำแหน่งที่ตรงกัน ตัวอย่างเช่น *d*(“text”, “that”) = 2

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

2) ระยะแก้ไขสั้น (minimal edit distance) คำนวณหาจากจำนวนครั้งที่น้อยที่สุดที่ใช้ในการเพิ่ม การลบ และการแทนที่แต่ละตัวอักขระ เพื่อทำให้สายอักขระทั้งสองเหมือนกัน โดยที่

$d(s1,s2) \geq |\, length(s1) - length(s2) \,|$ ตัวอย่างเช่น d(“text”, “tax”) = 2

ในงานวิจัยนี้ จะต้องทำการเปรียบเทียบรหัสคำของข้อคำถามกับรหัสคำของดัชนีคำหลัก โดยใช้ขั้นตอนวิธีการคำนวณหาค่าระยะแก้ไขสั้น ซึ่งใช้เวลาทำงานเป็น O(mn) โดย m และ n คือความยาวของสายอักขระที่ 1 และ 2 ตามลำดับ เมื่ออาศัยเทคนิคกำหนดการพลวัต (dynamic programming)[9] ซึ่งวิธีการคำนวณสามารถเขียนให้อยู่ในรูปการคำนวณด้วยความสัมพันธ์เวียนเกิด Edit (P_j, W_k) ได้ดังนี้

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

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

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

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

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

2.5 สรุป

ในบทนี้ได้กล่าวถึงทฤษฎีและงานวิจัยต่าง ๆ ที่เกี่ยวข้องในการพัฒนาขั้นตอนวิธีการเข้ารหัสคำทับศัพท์ภาษาไทย-อังกฤษเพื่อการค้นคืนข้ามภาษา โดยใช้แนวคิดในการแปลงคำทับศัพท์ทั้งในรูปภาษาไทย และภาษาอังกฤษให้เป็นรหัสคำอ่าน โดยใช้นิวรอลเน็ตเวิร์กในการเรียนรู้การสร้างรหัสคำ และใช้เทคนิคการเปรียบเทียบเชิงประมาณแบบขั้นตอนวิธีระยะแก้ไขสั้นที่สุดในการเปรียบเทียบรหัสคำ โดยรายละเอียดต่าง ๆ จะกล่าวไว้ในบทถัดไป

---

[9] J. Zobel and P. Dart, Phonetic String Matching: Lessons from Information Retrieval, Proceedings of the 19th Annual International ACM SIGR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp. 166-172, 1996.

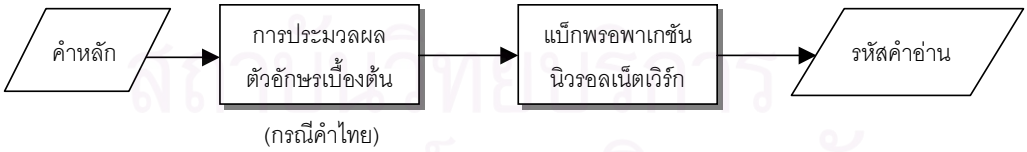
บทที่ 3
ขั้นตอนวิธีการฝึกนิวรอลเน็ตเวิร์ก

ในบทนี้จะกล่าวถึงขั้นตอนวิธีการฝึกนิวรอลเน็ตเวิร์กแบบแบ็กพรอพาเกชัน เพื่อใช้ในการเข้ารหัสคำทับศัพท์ภาษาไทย-อังกฤษ โดยจะกล่าวถึงขั้นตอนวิธีการฝึก การประมวลผลคำเบื้องต้น โครงสร้างของแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กที่ใช้ ลักษณะของข้อมูลเข้าและข้อมูลออก และหลักเกณฑ์ในการฝึก

3.1 ขั้นตอนวิธีการฝึก

ในการฝึก (train) นิวรอลเน็ตเวิร์กให้เรียนรู้การสร้างรหัสคำอ่านนั้น ข้อมูลที่ใช้ฝึกได้มาจากการนำความรู้ทางภาษาศาสตร์ ได้แก่ หลักเกณฑ์ในการถอดอักษร หลักการถ่ายเสียง หลักการอ่านออกเสียงทั้งภาษาไทยและภาษาอังกฤษ มาใช้ในการพิจารณาข้อมูลเข้า และข้อมูลออก

ในการฝึกนิวรอลเน็ตเวิร์กให้เรียนรู้การสร้างรหัส มีขั้นตอนดังรูปที่ 3.1 โดยในขั้นแรกจะนำคำหลักไปประมวลผลตัวอักษรเบื้องต้นก่อนในกรณีที่เป็นคำไทย ส่วนในกรณีคำอังกฤษสามารถส่งให้นิวรอลเน็ตเวิร์กเรียนรู้ได้เลย โดยจะใช้นิวรอลเน็ตเวิร์กให้เรียนรู้การสร้างรหัสคำอ่านทั้งหมด 4 ชุด สำหรับ (1) คำไทย (2) คำอังกฤษทับศัพท์คำไทย และ (3) คำอังกฤษ (4) คำไทยทับศัพท์คำอังกฤษ ข้อมูลเข้าจะเป็นตัวอักษรที่สนใจฝึกรหัสในคำ พร้อมทั้งตัวอักษรข้างเคียงหน้าหลังข้างละ 4 ตัวของคำ และให้ข้อมูลออกเป็นรหัสเสียงของตัวอักษรขาเข้านั้น ดังจะได้กล่าวรายละเอียดในหัวข้อถัดไป



รูปที่ 3.1 ขั้นตอนการฝึกแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กให้เรียนรู้การสร้างรหัสคำ

3.2 การประมวลผลคำเบื้องต้น

การเข้ารหัสคำนี้มีจุดประสงค์เพื่อแปลงคำทับศัพท์ทั้งที่อยู่ในรูปคำไทยและคำอังกฤษให้อยู่ในรูปแบบเดียวกัน คือ ในรูปรหัสคำอ่าน เพราะหากเป็นคำทับศัพท์ที่ตรงกันของทั้งสองภาษาจะอ่านออกเสียงได้เหมือนหรือคล้ายคลึงกัน ดังนั้น ตัวอักษรในภาษาไทยบางตัวที่ไม่อ่านออกเสียง และไม่มีเสียงที่ตรงกันในภาษาอังกฤษ ได้แก่ ไม้ไต่คู้ วรรณยุกต์ การันต์และทัณฑฆาต จึง

เป็นตัวอักษรที่ต้องทำการตัดออกก่อนเพื่อให้สะดวกแก่การประมวลผลต่อไป เพราะตามหลักการถอดอักษรไทยเป็นอังกฤษจะไม่พิจารณาตัวอักษรเหล่านี้

การประมวลผลตัวอักษรเบื้องต้นจะกระทำตามขั้นตอนโดยแบ่งเป็นกรณีต่าง ๆ ดังนี้[1] ถ้าเป็นคำที่พิจารณาเป็นคำในภาษาไทย

1. ตัดวรรณยุกต์และไม้ไต่คู้ออก
2. เปลี่ยน รร เป็น ั-น ในกรณีที่ไม่มีตัวสะกดตามหลัง และเปลี่ยนเป็น ั- ในกรณีที่มีตัวสะกด
3. เปลี่ยนสระ ใ- ไ- และ ไ-ย ให้อยู่ในรูปเดียวกันคือ –ัย
4. เปลี่ยนสระ -ำ เป็น ั-ม
5. ตัดตัวการันต์และทัณฑฆาตออก
6. เปลี่ยนสระ ฤ และ ฤา เป็น ริ และ รือ ตามลำดับ

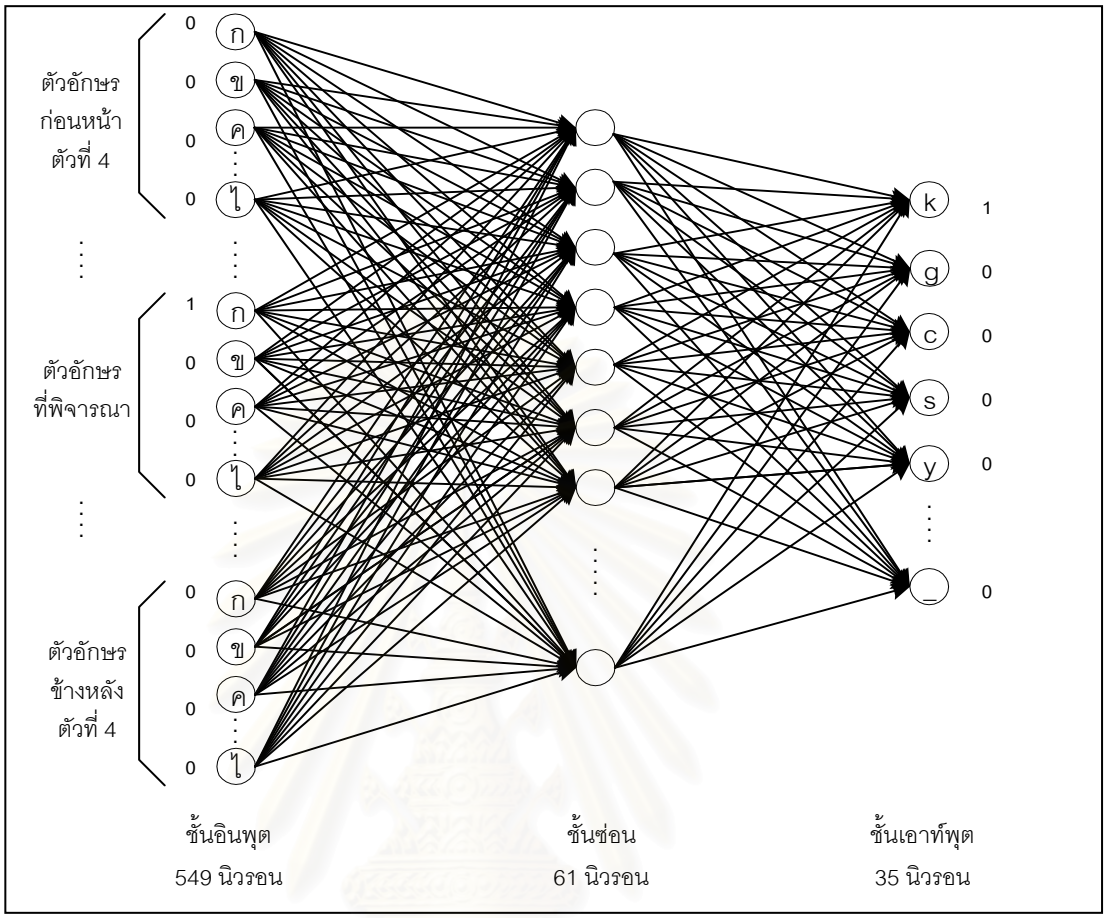
จากนั้นจึงนำคำที่ได้หลังผ่านกระบวนการนี้ส่งให้นิวรอลเน็ตเวิร์กเรียนรู้ ส่วนในกรณีคำอังกฤษสามารถส่งไปเรียนรู้ได้เลย

3.3 โครงสร้างนิวรอลเน็ตเวิร์ก

ขั้นตอนวิธีการเข้ารหัสคำจะใช้แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กมาช่วยในการเรียนรู้และสร้างรหัสคำอ่านของคำทับศัพท์ โดยจะใช้นิวรอลเน็ตเวิร์กให้เรียนรู้การสร้างรหัสคำอ่านของคำทับศัพท์ทั้งหมด 4 ชุด โดยแบ่งเป็น 2 กรณี คือ

- กรณีคำไทยทับศัพท์คำอังกฤษจะใช้นิวรอลเน็ตเวิร์ก 2 ชุด คือ สำหรับคำอังกฤษ เช่น CLINTON และคำไทยทับศัพท์คำอังกฤษ เช่น คลินตัน
- กรณีคำอังกฤษทับศัพท์คำไทยก็ใช้นิวรอลเน็ตเวิร์ก 2 ชุด คือ สำหรับคำไทย เช่น จุฬาลงกรณ์ และคำอังกฤษทับศัพท์คำไทย เช่น CHULALONGKORN

[1] ประยุทธ สุวรรณวิสารท, “การเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ” (วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2541), หน้า 34-38.



รูปที่ 3.2 ตัวอย่างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กที่ใช้ในการเรียนรู้คำไทย (ในกรณีคำอังกฤษทับศัพท์คำไทย) ซึ่งมีอินพุตเป็น (_ , _ , _ , _ , ก , น , ก , พ , ี) และมีเอาท์พุตเป็น 'k'

โครงสร้างของนิวรอลเน็ตเวิร์กที่ใช้ประกอบด้วยชั้นต่าง ๆ 3 ชั้น (รูปที่ 3.2) ดังนี้

- ชั้นอินพุต (input layer) ประกอบด้วยจำนวนนิวรอนเท่ากับจำนวนอักขระทั้งหมดของภาษาที่พิจารณา คูณด้วย จำนวนตัวอักษรทั้งหมดที่ใช้พิจารณา ซึ่งจะพิจารณาตัวอักษรครั้งละ 9 ตัวอักษรของคำที่ต้องการฝึก โดยตัวอักษรที่ตำแหน่งตรงกลางจะเป็นตัวที่ถูกพิจารณาเพื่อเรียนรู้และสร้างรหัสคำอ่าน ส่วนตัวอักษร 4 ตัวหน้าและอีก 4 ตัวหลังจะใช้ประกอบเพื่อช่วยในการพิจารณา เนื่องจากในการที่เราจะทราบได้ว่าตัวอักษรเรากำลังพิจารณาจะออกเสียงอย่างไร จะต้องพิจารณาจากตัวข้างเคียงที่อยู่ในคำนั้นด้วย ทั้งในคำอังกฤษและคำไทย ตัวอย่างเช่น ในกรณีภาษาไทยที่มีการใช้สระประสมและสระเดี่ยวที่เกิดจากการใช้ตัวอักษรตั้งแต่ 2 ตัวขึ้นไป เช่น สระเอีย ใน

คำว่า “เสถียร” เมื่อตัวอักษรที่กำลังพิจารณา คือ “เ” เราจะรู้ได้ว่าเป็นส่วนประกอบของสระเอียก็โดยที่พิจารณาจากตัวอักษรข้างหลัง 4 ตัว หรือเมื่อตัวอักษรที่กำลังพิจารณา คือ “ย” จะรู้ได้ก็โดยที่พิจารณาจากตัวอักษรข้างหน้า 4 ตัว ส่วนกรณีคำอังกฤษนั้นอาจใช้จำนวนตัวอักษรที่ใช้พิจารณาร่วมน้อยกว่าก็สามารถทราบลักษณะการออกเสียงของตัวอักษรที่สนใจได้ ดังนั้น ในการทดลองนี้จึงได้ทำการทดสอบ เพื่อหาจำนวนตัวอักษรข้างเคียงที่ใช้พิจารณาร่วมที่เหมาะสม ซึ่งพบว่า จำนวนที่ให้ผลในการเรียนรู้ที่ดี คือ พิจารณาตัวอักษรข้างหน้า 4 ตัว ข้างหลังอีก 4 ตัว รวมกับตัวที่กำลังพิจารณาอีก 1 ตัว

ดังนั้น กรณีคำอังกฤษ มีจำนวนอักขระทั้งหมด 26 ตัว ข้อมูลเข้าจึงมีจำนวน 26 x 9 นิวรอน กรณีคำไทย มีจำนวนอักขระซึ่งเป็นพยัญชนะและสระเดี่ยวทั้งหมด 61 ตัว จึงมีข้อมูลเข้าจำนวน 61 x 9 นิวรอน โดยตำแหน่งที่ตรงกับตัวอักษรที่พิจารณาจะถูกกำหนดให้มีค่าเป็น 1

ในแต่ละคำที่ใช้ฝึก จะประกอบด้วยชุดของข้อมูลเข้าจำนวนเท่ากับจำนวนตัวอักษรของคำนั้น เนื่องจากจะเลื่อนคำไปครั้งละหนึ่งตัวอักษรซึ่งจะทำให้ตัวอักษรทุกตัวได้รับการพิจารณาเพื่อสร้างรหัส

- ชั้นซ่อน (hidden layer) ได้ทำการทดลองเพื่อหาจำนวนนิวรอนที่เหมาะสม (โดยหาได้จากค่าที่ใช้ในการฝึกแล้วให้ผลการเรียนรู้ดีที่สุด) สำหรับแต่ละเน็ตเวิร์ก ซึ่งสำหรับคำไทย จะมี 61 นิวรอน ส่วนคำอังกฤษจะมี 234 นิวรอน
- ชั้นเอาท์พุต (output layer) จะมีจำนวนนิวรอนเท่ากับรหัสเสียงพยัญชนะและเสียงสระที่เป็นไปได้ทั้งหมด ซึ่งในกรณีคำอังกฤษ และคำไทยทับศัพท์คำอังกฤษจะมี 39 นิวรอน ดังแสดงในตารางที่ 3.1 และในกรณีคำไทย และคำอังกฤษทับศัพท์คำไทยจะมี 35 นิวรอน ดังแสดงในตารางที่ 3.2 โดยนิวรอนในตำแหน่งที่ตรงกับรหัสเสียงของตัวอักษรที่กำลังพิจารณา (ตัวที่ 5) จะถูกกำหนดให้มีค่าเป็น 1

ผู้วิจัยได้นำหลักเกณฑ์การออกเสียงทั้งภาษาไทยและภาษาอังกฤษ ดังที่ได้กล่าวมาแล้วในบทที่ 2 มาใช้ในการสร้างตารางรหัสเสียงสำหรับคำทับศัพท์ทั้งสองประเภท (ตารางที่ 3.1 และ 3.2) โดยการสร้างตารางรหัสเสียงสำหรับคำไทยทับศัพท์คำอังกฤษ จะใช้หน่วยเสียงของภาษาอังกฤษเป็นหลัก แล้วนำหน่วยเสียงภาษาไทยไปเปรียบเทียบ เพื่อหากลุ่มเสียงที่ตรงกันหรือใกล้เคียงกันของทั้งสองภาษาแล้วกำหนดเป็นรหัสเสียง หากเป็นการสร้างตารางรหัสเสียงสำหรับคำอังกฤษทับศัพท์คำไทย ก็จะใช้หน่วยเสียงของภาษาไทยเป็นหลัก แล้วนำหน่วยเสียงภาษาอังกฤษไปเปรียบเทียบ เพื่อหากลุ่มเสียงที่ตรงกันหรือใกล้เคียงกันแล้วกำหนดเป็นรหัสเสียง

ตารางที่ 3.1 รหัสเสียงสำหรับคำไทยทับศัพท์คำอังกฤษ

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

ตารางที่ 3.2 รหัสเสียงสำหรับคำอังกฤษทับศัพท์คำไทย

พยัญชนะไทย	พยัญชนะอังกฤษ	รหัสตัวต้น	รหัสตัวสะกด	สระไทย	ตัวอย่างสระอังกฤษ	รหัส
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 ็	er, oe	W
ณ น ห	n	n	n	เ–ียะ เ–ีย	ia, ie, aiu	I
บ	b	b	p	เ–ือะ เ–ือ -ัวะ -ัว –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	r	r	n	-ิว	iu	X
ล ฬ	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	$
				เ–ียว	ieo, eaw, eo, ew, iow, iau, iew, iaw	@

3.4 หลักเกณฑ์การฝึก

ในการฝึก[2] สำหรับคำไทยทับศัพท์คำอังกฤษ จะเป็นการแปลงจากตัวอักษรเป็นรหัสเสียงแบบหนึ่งต่อหนึ่ง แต่สำหรับคำอังกฤษทับศัพท์คำไทย ตัวอักษรหนึ่งตัวในภาษาไทยอาจทำให้เกิดเสียงได้มากกว่าหนึ่งเสียง เนื่องจากภาษาไทยมีสระลดรูป ดังนั้นในการฝึกคำไทยในกรณีนี้ จึงไม่ใช่เป็นการแปลงจากตัวอักษรเป็นรหัสเสียงแบบหนึ่งต่อหนึ่งเสมอ แต่บางครั้งอาจเป็นแบบหนึ่งต่อสอง โดยจะให้รหัสสำหรับเสียงตัวอักษรและรหัสสำหรับเสียงสระลดรูป ผู้วิจัยได้เสนอหลักพิจารณาโดยแยกเป็นกรณีต่าง ๆ ดังนี้

- ตัวอักษรหลายตัวในคำอังกฤษทำให้เกิดเสียงหนึ่งเสียง ในกรณีนี้จะมีเพียงตัวอักษรเพียงตัวเดียวที่มีรหัส ส่วนตัวอักษรที่เหลือจะให้รหัสเสียงเป็น _ คือ เป็นตัวที่ไม่มีเสียง เช่น เสียง /n/ เกิดจากกลุ่มตัวอักษร “nd” ในคำว่า “ligand” จะให้รหัสเสียงสำหรับ “nd” เป็น /n_/
- ตัวอักษรหนึ่งตัวในคำอังกฤษทำให้เกิดเสียงมากกว่าหนึ่งเสียง ในกรณีนี้จะเลือกให้รหัสเสียงเพียงเสียงเดียว เช่น ตัว “x” ในคำว่า oxygen ทำให้เกิดทั้งเสียง /k/ และเสียง /s/ จะเลือกที่ให้รหัสเสียงเพียงเสียงเดียวสำหรับตัวอักษร “x” คือ /s/
- ตัวอักษรหลายตัวในคำไทยทำให้เกิดเสียงหนึ่งเสียง ในกรณีนี้จะเลือกให้รหัสเสียงสำหรับตัวอักษรตัวแรกของกลุ่ม ส่วนที่เหลือให้รหัสเสียงเป็น _ เช่น เสียง /I/ ที่เกิดจากกลุ่มตัวอักษร “เ–ีย” ในคำว่า “ไลบีเรีย” จะให้รหัสเสียงสำหรับ “เรีย” เป็น /Ir__/
- ตัวอักษร อ ในกรณีนี้จะเลือกให้รหัสเสียงเป็นรหัสเสียงของสระที่อยู่หลังตัวอักษร อ ส่วนตัวสระเองให้รหัสเสียงเป็น _ เช่น คำว่า “อา” ให้รหัสเสียงเป็น /a_/ หรือ คำว่า “อิน” ให้รหัสเสียงเป็น /i_n/
- สระอะลดรูป ในกรณีนี้จะเพิ่มรหัสเสียง /a/ สำหรับสระอะลดรูป เช่น ตัวอักษร “ห” ในคำว่า “หริ” ให้เสียง /h/ สำหรับตัว อักษร “ห” และเสียง /a/ สำหรับสระอะลดรูป ดังนั้นจะให้รหัสเสียงสำหรับ “หริ” เป็น /hari/
- สระโอะและสระออลดรูป ในกรณีนี้จะเพิ่มรหัสเสียง /o/ สำหรับสระโอะหรือสระออลดรูป เช่น คำว่า “สม” มีเสียงสระโอะลดรูป จึงให้รหัสเสียงเป็น /som/ หรือคำว่า ตัวอักษร “ว” ในคำว่า “วรชัย” มีเสียงสระออลดรูป จึงให้รหัสเสียงเป็น /vo/

---

[2] R. R. Leighton, The Aspirin/MIGRAINES Neural Network Software 6.0 (AM6) The MITRE Corporation, 1992.

ตัวอย่าง 1 ต้องการฝึกคำว่า “กนกพีระวุฒิ” ซึ่งเป็นคำไทย มีขั้นตอนการทำงานดังนี้

1. ประมวลผลตัวอักษรเบื้องต้น จะได้ กนกพีระวุฒิ → กนกพีระวุฒิ เช่นเดิม
2. ใช้ตาราง 3.2 สร้างเป็นข้อมูลเข้าเพื่อส่งให้นิวรอลเน็ตเวิร์ก โดยพิจารณาตัวอักษรครั้งละ 9 ตัว โดยตัวที่สนใจคือตัวที่ 5 และพิจารณาตัวอักษรข้างเคียงข้างหน้า 4 ตัวและข้างหลัง 4 ตัว จำนวนชุดข้อมูลเข้าจะเท่ากับจำนวนตัวอักษรของคำ คือ 11

ข้อมูลเข้า		ข้อมูลออก
(_,_,_,_,**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,**ิ**,_,_,_,_)	→	_

ตัวอย่าง 2 ต้องการฝึกคำว่า “อินเตอร์เฟส” ซึ่งเป็นคำไทยทับศัพท์คำอังกฤษ

1. ประมวลผลตัวอักษรเบื้องต้น จะได้ อินเตอร์เฟส → อินเตอเฟส
2. ใช้ตาราง 3.1 สร้างเป็นข้อมูลเข้าเพื่อส่งให้นิวรอลเน็ตเวิร์ก

ข้อมูลเข้า		ข้อมูลออก
(_,_,_,_,**อ**,ิ,น,เ,ต)	→	i
(_,_,_,อ,ิ,น,เ,ต,อ)	→	_
(_,_,อ,ิ,**น**,เ,ต,อ,เ)	→	n
(_,อ,ิ,น,**เ**,ต,อ,เ,ฟ)	→	w
(อ,ิ,น,เ,**ต**,อ,เ,ฟ,ส)	→	t
(ิ,น,เ,ต,**อ**,เ,ฟ,ส,_)	→	_
(น,เ,ต,อ,**เ**,ฟ,ส,_,_)	→	e
(เ,ต,อ,เ,**ฟ**,ส,_,_,_)	→	f
(ต,อ,เ,ฟ,**ส**,_,_,_,_)	→	s

ตัวอย่าง 3 ต้องการฝึกคำว่า “รักษ์ไชยวรรณ” ซึ่งเป็นคำไทย

1. ประมวลผลตัวอักษรเบื้องต้น จะได้ รักษ์ไชยวรรณ → รักชัยวัณ
2. ใช้ตาราง 3.2 สร้างเป็นข้อมูลเข้าเพื่อส่งให้นิวรอลเน็ตเวิร์ก

ข้อมูลเข้า		ข้อมูลออก
ั(_,_,_,_,**ร**,ั,ก,ช,ั)	→	r
(_,_,_,ร,ั,**ก**,ช,ั,ย)	→	a
(_,_,ร,ั,ก,**ช**,ั,ย,ว)	→	k
(_,ร,ั,ก,ช,**ั**,ย,ว,ั)	→	c
(ร,ั,ก,ช,ั,**ย**,ว,ั,ณ)	→	!
(ั,ก,ช,ั,ย,**ว**,ั,ณ,_)	→	_
(ก,ช,ั,ย,ว,**ั**,ณ,_,_)	→	v
(ช,ั,ย,ว,ั,**ณ**,_,_,_)	→	a
(ั,ย,ว,ั,ณ,_,_,_,_)	→	n

ตัวอย่าง 4 ต้องการฝึกคำว่า “mexico” ซึ่งเป็นคำอังกฤษ

1. ใช้ตาราง 3.1 สร้างเป็นข้อมูลเข้าเพื่อส่งให้นิวรอลเน็ตเวิร์ก

ข้อมูลเข้า		ข้อมูลออก
ั(_,_,_,_,**m**,e,x,i,c)	→	m
(_,_,_,m,**e**,x,i,c,o)	→	e
(_,_,m,e,**x**,i,c,o,_)	→	s
(_,m,e,x,**i**,c,o,_,_)	→	i
(m,e,x,i,**c**,o,_,_,_)	→	k
(e,x,i,c,**o**,_,_,_,_)	→	o

ตัวอย่าง 5 ต้องการฝึกคำว่า “rhodium” ซึ่งเป็นคำอังกฤษ

1. ใช้ตาราง 3.1 สร้างเป็นข้อมูลเข้าเพื่อส่งให้นิวรอลเน็ตเวิร์ก

ข้อมูลเข้า		ข้อมูลออก
ั(_,_,_,_,**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, _, _) → I

(h, o, d, i, **u**, m, _, _, _) → _

(o, d, i, u, **m**, _, _, _, _) → m

ตัวอย่าง 6 ต้องการฝึกคำว่า “visuth” ซึ่งเป็นคำอังกฤษทับศัพท์คำไทย

1. ใช้ตาราง 3.2 สร้างเป็นข้อมูลเข้าเพื่อส่งให้นิวรอลเน็ตเวิร์ก

ข้อมูลเข้า		ข้อมูลออก
(_, _, _, _, **v**, i, s, u, t)	→	v
(_, _, _, v, **i**, s, u, t, h)	→	i
(_, _, v, i, **s**, u, t, h, _)	→	s
(_, v, i, s, **u**, t, h, _, _)	→	u
(v, i, s, u, **t**, h, _, _, _)	→	t
(i, s, u, t, **h**, _, _, _, _)	→	_

3.5 สรุป

ในบทนี้ได้กล่าวถึงขั้นตอนในการฝึกแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กให้ทำการเรียนรู้วิธีการสร้างรหัสเสียงสำหรับคำที่ต้องการ หลังจากทำการฝึกเสร็จก็จะนำน้ำหนัก (weight) ของแต่ละเน็ตเวิร์กที่ให้ผลการเรียนรู้ที่ดีที่สุดไปใช้ในการเข้ารหัสคำในบทถัดไป

บทที่ 4
ขั้นตอนวิธีการเข้ารหัสคำและการค้นคืนข้ามภาษา

ในบทนี้จะกล่าวถึงขั้นตอนวิธีการเข้ารหัสคำทับศัพท์ภาษาไทย-อังกฤษ เพื่อการค้นคืนข้ามภาษา โดยอาศัยนิวรอลเน็ตเวิร์กที่ได้ผ่านกระบวนการฝึกแล้ว ซึ่งจะกล่าวถึงขั้นตอนวิธีการเข้ารหัสคำ ขั้นตอนการค้นคืนข้ามภาษา

4.1 ขั้นตอนวิธีการเข้ารหัสคำ

ขั้นตอนวิธีการเข้ารหัสคำจะใช้แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กมาใช้สร้างรหัสคำอ่านของคำทับศัพท์ โดยจะนำน้ำหนัก (weight) ของแต่ละเน็ตเวิร์กที่ผ่านการฝึกตามกระบวนการดังบทที่ 3 แล้วให้ผลการเรียนรู้ที่ดีที่สุดมาใช้ในการทดสอบการเข้ารหัสคำ โดยจะพิจารณาคำที่ต้องการทดสอบว่าเป็นคำทับศัพท์ประเภทใด เพื่อที่จะได้นำไปทดสอบกับน้ำหนักและโครงสร้างนิวรอลเน็ตเวิร์กที่ถูกต้อง โดยขั้นตอนการเข้ารหัสจะคล้ายกับขั้นตอนวิธีการฝึก แต่ไม่ต้องมีการกำหนดค่าของข้อมูลออก เนื่องจากนิวรอลเน็ตเวิร์กจะทำการกำหนดค่าให้กับแต่ละโหนดในชั้นเอาท์พุต ซึ่งในกรณีของคำอังกฤษที่มีเอาท์พุตเพียงโหนดเดียว โหนดเอาท์พุตที่มีค่ามากที่สุดจะถูกเลือกเป็นรหัสเสียง ส่วนในกรณีของคำไทยที่มีการให้เสียงของสระลดรูป ทำให้บางครั้งเอาท์พุตอาจมี 2 โหนด เราจะเลือกจำนวนเอาท์พุต โดยนำเอาท์พุตที่มีค่ามากที่สุด 2 โหนดมาหาผลต่าง และเมื่อผลต่างมีค่าไม่เกินเกณฑ์ (threshold) จะได้ว่ามีเอาท์พุตจำนวน 2 โหนด จากการทดลองได้ทำการคำนวณหาเกณฑ์จนได้ค่าที่เหมาะสม คือ 0.3

เมื่อได้รหัสเสียงของทุกตัวอักษรในคำจากนิวรอลเน็ตเวิร์กแล้ว จะนำรหัสเสียงทั้งหมดมาต่อกันตามลำดับเพื่อสร้างเป็นรหัสคำอ่าน โดยจะทำการตัดรหัสที่ไม่ออกเสียง (_) ออกและทำการย้ายรหัสเสียงสระไปต่อท้ายรหัสเสียงพยัญชนะ ดังแสดงในตัวอย่าง

<u>ตัวอย่าง 1</u> ต้องการเข้ารหัสคำว่า “กนกพีระวุฒิ” ซึ่งเป็นคำไทย

<u>การเข้ารหัสคำ</u>

1. ทำการประมวลผลตัวอักษรเบื้องต้น จะได้ กนกพีระวุฒิ → กนกพีระวุฒิ เช่นเดิม
2. สร้างเป็นข้อมูลเข้าเพื่อส่งให้นิวรอลเน็ตเวิร์กผลิตข้อมูลออก โดยพิจารณาตัวอักษรทั้งหมดครั้งละ 9 ตัว โดยตัวที่สนใจ คือตัวที่ 5 และพิจารณาตัวอักษรข้างเคียงข้างหน้า 4 ตัวและข้างหลังอีก 4 ตัว

ข้อมูลเข้า		ผลลัพธ์ที่ได้
(_,_,_,_,**ก**,น,ก,พ,ี)	→	k
(_,_,_,ก,**น**,ก,พ,ี,ร)	→	a,n[1]
(_,_,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,ิ,_,_,_,_)	→	_

3. นำรหัสที่ได้มาเรียงต่อกัน แล้วตัดรหัสที่ไม่ออกเสียง (_) ออกและทำการย้ายรหัสเสียงสระไปต่อท้ายรหัสเสียงพยัญชนะ

```
kanokpiravut_
     ↓
kanokpiravut
     ↓
knkprvt,aoiau
     ↓
knkprvtaoiau
```

<u>ตัวอย่าง 2</u> ต้องการเข้ารหัสคำว่า “ไอโซเทอร์มัล” ซึ่งเป็นคำไทยทับศัพท์คำอังกฤษ

<u>การเข้ารหัสคำ</u>

1. ประมวลผลตัวอักษรเบื้องต้น จะได้ ไอโซเทอร์มัล→ ไอโซเทอมัล
2. สร้างเป็นข้อมูลเข้าเพื่อส่งให้นิวรอลเน็ตเวิร์กผลิตข้อมูลออก

ข้อมูลเข้า		ผลลัพธ์ที่ได้
(_,_,_,_,**ไ**,อ,โ,ซ,เ)	→	!
(_,_,_,ไ,**อ**,โ,ซ,เ,ท)	→	_
(_,_,ไ,อ,**โ**,ซ,เ,ท,อ)	→	o

[1] จากตัวอย่างที่นำเสนอนี้ กรณีที่มีเอาท์พุตเป็น 2 รหัสเสียง ลำดับของรหัสเสียงไม่มีความสำคัญ

(_,ไ,อ,โ,**ซ**,เ,ท,อ,ม) → s

(ไ,อ,โ,ซ,**เ**,ท,อ,ม,ั) → W

(อ,โ,ซ,เ,**ท**,อ,ม,ั,ล) → t

(โ,ซ,เ,ท,**อ**,ม,ั,ล,_) → _

(ซ,เ,ท,อ,**ม**,ั,ล,_,_) → m

(เ,ท,อ,ม,ั,ล,_,_,_) → V

(ท,อ,ม,ั,**ล**,_,_,_,_) → l

3. นำรหัสมาเรียงลำดับใหม่

```
!_soWt_mVl
     ↓
!soWtmVl
     ↓
stml,!oWV
     ↓
stml!oWV
```

<u>ตัวอย่าง 3</u> ต้องการเข้ารหัสคำว่า “dioxide” ซึ่งเป็นคำอังกฤษ

<u>การเข้ารหัสคำ</u>

1. สร้างเป็นข้อมูลเข้าเพื่อส่งให้นิวรอลเน็ตเวิร์กผลิตข้อมูลออก

ข้อมูลเข้า		ผลลัพธ์ที่ได้
(_,_,_,_,**d**,i,o,x,i)	→	d
(_,_,_,d,**i**,o,x,i,d)	→	!
(_,_,d,i,**o**,x,i,d,e)	→	o
(_,d,i,o,**x**,i,d,e,_)	→	s
(d,i,o,x,**i**,d,e,_,_)	→	!
(i,o,x,i,**d**,e,_,_,_)	→	_
(o,x,i,d,**e**,_,_,_,_)	→	_

2. นำรหัสมาเรียงลำดับใหม่

```
d!os!_ _
   ↓
d!os!
   ↓
ds,!o!
   ↓
ds!o!
```

ตัวอย่าง 4 ต้องการเข้ารหัสคำว่า “chitman” ซึ่งเป็นคำอังกฤษทับศัพท์คำไทย

การเข้ารหัสคำ

1. สร้างเป็นข้อมูลเข้าเพื่อส่งให้นิวรอลเน็ตเวิร์กผลิตข้อมูลออก

ข้อมูลเข้า		ผลลัพธ์ที่ได้
(_ , _ , _ , _ , **c** , h , i , t , m)	→	c
(_ , _ , _ , c , **h** , i , t , m , a)	→	_
(_ , _ , c , h , **i** , t , m , a , n)	→	i
(_ , c , h , i , **t** , m , a , n , _)	→	t
(c , h , i , t , **m** , a , n , _ , _)	→	m
(h , i , t , m , **a** , n , _ , _ , _)	→	a
(i , t , m , a , **n** , _ , _ , _ , _)	→	n

2. นำรหัสมาเรียงลำดับใหม่

c_itman
↓
citman
↓
ctmn,ia
↓
ctmnia

4.2 ขั้นตอนการค้นคืนข้ามภาษา

รหัสคำที่ได้ของคำคู่ที่ตรงกันทั้งสองภาษาอาจจะไม่ตรงกันทุกตัวอักษร แต่จะมีลักษณะคล้ายกัน เนื่องจากหลักเกณฑ์การทับศัพท์ที่ใช้ในปัจจุบันมีหลายรูปแบบ เพื่อให้ได้ค่าแม่นยำ (precision) และค่าเรียกคืน (recall) ที่ดี จะใช้การเปรียบเทียบรหัสคำแบบประมาณ (approximate matching) ซึ่งอาศัยการคำนวณความแตกต่าง (distance) ของรหัสคำด้วยเทคนิคระยะแก้ไขสั้นที่สุด (minimal edit distance) โดยคำนวณหาจากจำนวนครั้งที่น้อยที่สุดที่ใช้ในการเพิ่ม การลบ และการแทนที่แต่ละตัวอักขระ เพื่อทำให้รหัสคำทั้งสองเหมือนกัน

การคำนวณค่าความแตกต่างนี้อาศัยเทคนิคกำหนดการพลวัต (dynamic programming) ซึ่งวิธีการคำนวณสามารถเขียนให้อยู่ในรูปการคำนวณด้วยความสัมพันธ์เวียนเกิด Edit (P_j, W_k) ดังนี้

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

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

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

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

โดยที่ $P_j = p_1\ p_2\ p_{3...}\ p_j$ เป็นสายอักขระต้นแบบ มีความยาว j ตัวอักษร

$W_k = w_1\ w_2\ w_{3...}\ w_k$ เป็นสายอักขระเป้าหมาย มีความยาว k ตัวอักษร

$r(p_j, w_k) = 0$ ถ้า p_j เท่ากับ w_k

$= 1$ ถ้า p_j ไม่เท่ากับ w_k

ถ้าค่าความแตกต่างที่ได้มีค่าไม่เกินเกณฑ์ (ค่าคงที่ d) จะสรุปได้ว่ารหัสคำทั้งสองเป็นรหัสที่มาจากคำหลักที่ตรงกันในอีกภาษา

<u>ตัวอย่าง 1</u> ต้องการทดสอบคำว่า “kanokpeerawut” และ “กนกพีระวุฒิ” เป็นคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษหรือไม่ โดยทำการเข้ารหัสคำ แล้วคำนวณหาค่าความแตกต่าง

<u>การเข้ารหัสคำ</u>

kanokpeerawut → `kanokpi_ravut` → `knkprvtaoiau`

กนกพีระวุฒิ → `konkopiravut_` → `knkprvtooiau`

<u>การค้นคืน</u>

$Edit($ “`knkprvtaoiau`”, “`knkprvtooiau`” $) = 1$

จากตัวอย่างพบว่า รหัสคำทั้งสองมีค่าความแตกต่างเป็น 1 ถ้าเรากำหนดให้เกณฑ์มีค่าเท่ากับ 1 (d=1) ก็จะสามารถค้นคืน “kanokpeerawut” จาก “กนกพีระวุฒิ” ได้

<u>ตัวอย่าง 2</u> ต้องการทดสอบคำว่า “isothermal” และ “ไอโซเทอร์มัล” เป็นคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษหรือไม่ โดยทำการเข้ารหัสคำ แล้วคำนวณหาค่าความแตกต่าง

<u>การเข้ารหัสคำ</u>

isothermal → `!sot_W_mwl` → `stml!oWw`

ไอโซเทอมัล → `!sot_W_mVl` → `stml!oWV`

<u>การค้นคืน</u>

$Edit($ “`stml!oWw`”, “`stml!oWV`” $) = 1$

จากตัวอย่างพบว่า รหัสคำทั้งสองมีค่าความแตกต่างเป็น 1 ถ้าเรากำหนดให้เกณฑ์มีค่าเท่ากับ 1 (d=1) ก็จะสามารถค้นคืน “isothermal” จาก “ไอโซเทอร์มัล” ได้

4.3 สรุป

ในบทนี้ได้กล่าวถึงขั้นตอนวิธีการเข้ารหัสคำทับศัพท์ภาษาไทย-อังกฤษเพื่อการค้นคืนข้ามภาษา โดยนำนิวรอลเน็ตเวิร์กที่ผ่านการฝึกมาใช้ในการเข้ารหัสคำทับศัพท์ให้อยู่รูปรหัสคำอ่านในการค้นคืนข้ามภาษาทำได้โดยระบุคำหลักในการค้นหาพร้อมทั้งแจ้งให้ขั้นตอนวิธีทราบว่าจะข้ามภาษาแบบคำไทยทับศัพท์คำอังกฤษหรือคำอังกฤษทับศัพท์คำไทย จากนั้นระบบทำการเข้ารหัสคำหลักที่ต้องการค้นหาแล้วนำรหัสคำที่ได้ไปเปรียบเทียบกับรหัสคำในดัชนีคำหลักของเอกสารที่ได้เข้ารหัสไว้แล้วในขั้นตอนการทำดัชนี เมื่อเปรียบเทียบหาค่าความแตกต่างของรหัสคำใดแล้วได้ค่าน้อยกว่าหรือเท่ากับเกณฑ์ที่ใช้ในการยอมรับค่าความแตกต่างแล้ว จะถือว่าคำหลักนั้นเป็นคำหลักที่ตรงกันในอีกภาษาหนึ่ง

บทที่ 5
การทดลองและผลการทดลอง

ในบทนี้จะกล่าวถึงวิธีการทดลองและผลการทดลอง เมื่อใช้ขั้นตอนวิธีการเข้ารหัสคำทับศัพท์ภาษาไทย-อังกฤษดังที่ได้นำเสนอมาแล้ว

5.1 วิธีการทดลอง

ผู้วิจัยได้ทำการทดลองขั้นตอนวิธีที่ได้นำเสนอ ในกรณีคำไทยทับศัพท์คำอังกฤษ ใช้ชุดของคำอังกฤษและคำทับศัพท์ที่ตรงกัน ซึ่งส่วนใหญ่เป็นคำนามเฉพาะ คำศัพท์วิทยาศาสตร์ คำศัพท์คณิตศาสตร์ และคำศัพท์เคมี โดยนำมาจากหนังสือรวมคำศัพท์ของราชบัณฑิตยสถาน และหนังสือเรียนวิชาเคมี จำนวน 1,876 คู่ และกรณีคำอังกฤษทับศัพท์คำไทยใช้ชื่อและชื่อสกุลทั้งภาษาไทยและภาษาอังกฤษที่ตรงกันของนิสิตจุฬาลงกรณ์มหาวิทยาลัย จำนวน 2,000 คู่ ไปทำการเข้ารหัสด้วยขั้นตอนวิธีที่เสนอและจัดเก็บคำศัพท์และรหัสคำที่ได้เก็บในฐานข้อมูล หลังจากนั้นนำคำศัพท์ทั้งหมดไปค้นคืนทีละคำศัพท์กับฐานข้อมูลเพื่อทำการวัดค่าแม่นยำ ค่าเรียกคืน[1] และตัววัด F1 (F1-measure)[2] ซึ่งเป็นการวัดค่าเฉลี่ยของค่าแม่นยำและค่าเรียกคืน ซึ่งใช้สูตรดังนี้

$$\text{ค่าแม่นยำ} = \frac{\text{จำนวนคำศัพท์ที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำศัพท์ที่คืนกลับมา}} \times 100$$

$$\text{ค่าเรียกคืน} = \frac{\text{จำนวนคำศัพท์ที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำศัพท์ที่เกี่ยวข้องทั้งหมด}} \times 100$$

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

---

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

[2] C. J. Van Rijsbergen, Information Retrieval (Butterworths, London, 1979).

ในการทดลองจะทำการแบ่งข้อมูลเป็นข้อมูลฝึก (train set) และข้อมูลทดสอบ (test set) เพื่อให้การวัดผลที่ได้ไม่โน้มเอียงกับการแบ่งชุดฝึกกับชุดทดสอบ เราใช้วิธีการ K-fold cross validation[3] วิธีการนี้จะแบ่งข้อมูลทั้งหมดออกเป็น K ส่วนเท่าๆกัน แล้วใช้แต่ละส่วนเป็นชุดทดสอบ ส่วนละ 1 ครั้ง ทำการทดสอบทั้งหมด K ครั้ง ในแต่ละครั้งที่เลือกส่วนหนึ่งใดๆเป็นชุดทดสอบ ส่วนที่เหลือ K-1 ส่วนจะถูกใช้เป็นชุดฝึก จากนั้นนำค่าที่ได้จากการทดลองทั้งหมด K ครั้งมาหาค่าเฉลี่ยเป็นผลการทดลอง

5.2 ผลการทดลอง

ในการทดลองนี้ได้แบ่งข้อมูลออกเป็นส่วน ๆ ให้แต่ละส่วนมีคำประมาณ 400 คู่ คือ กรณีคำไทยทับศัพท์คำอังกฤษจะแบ่งข้อมูลออกเป็น 4 ส่วน (469 คู่) ส่วนในกรณีคำอังกฤษทับศัพท์คำไทยจะแบ่งข้อมูลออกเป็น 5 ส่วน (400 คู่) ส่วนค่าเฉลี่ยของการทดสอบหารหัสคำที่ได้จากนิวรอลเน็ตเวิร์กทั้ง 4 ชุดแสดงดังตารางที่ 5.1

ตารางที่ 5.1 ค่าเฉลี่ยของการทดสอบหารหัสคำที่ได้จากนิวรอลเน็ตเวิร์ก

คำทับศัพท์	นิวรอลเน็ตเวิร์ก สำหรับคำไทย	นิวรอลเน็ตเวิร์ก สำหรับคำอังกฤษ
คำอังกฤษทับศัพท์คำไทย	89.16 %	93.68 %
คำไทยทับศัพท์คำอังกฤษ	94.88 %	74.42 %

ผู้วิจัยได้ทำการทดลองโดยการเปลี่ยนแปลงค่าพารามิเตอร์ d (เป็นเกณฑ์ในการยอมรับค่าความแตกต่าง) เพื่อหาค่าความแตกต่างที่น้อยที่สุดที่ให้ค่าเฉลี่ยของค่าแม่นยำและค่าเรียกคืนสูงที่สุด ได้ผลดังแสดงในตารางที่ 5.2 และตารางที่ 5.3 กรณีที่ค่า d =0 คือการเปรียบเทียบแบบเหมือนกันทุกประการ (exact matching)

[3] Michell, T. M., Machine Learning (New York : McGraw-Hill Companies, 1997).

ตารางที่ 5.2 ผลการทดลองกรณีคำไทยทับศัพท์คำอังกฤษ

d	ค่าแม่นยำ	ค่าเรียกคืน	ตัววัด F1
0	99.06	41.74	58.72
1	87.28	77.19	81.91
2	56.88	94.09	70.90
3	28.32	98.08	43.94

ตารางที่ 5.3 ผลการทดลองกรณีคำอังกฤษทับศัพท์คำไทย

d	ค่าแม่นยำ	ค่าเรียกคืน	ตัววัด F1
0	99.71	44.60	61.60
1	96.34	75.15	84.41
2	76.37	91.90	83.39
3	47.75	98.10	64.19

จากผลการทดลองพบว่า ทั้งกรณีคำไทยทับศัพท์อังกฤษและคำอังกฤษทับศัพท์ ตัววัด F1 จะมีค่าสูงที่สุด เมื่อ d มีค่าเป็น 1 และเมื่อทำการเปรียบเทียบผลการทดลองที่ได้กับผลการทดลองที่ได้จากการใช้วิธีเข้ารหัสของประยุทธ สุวรรณวิสารท[4] ด้วยข้อมูลชุดเดียวกัน โดยเลือกค่าความแม่นยำและค่าเรียกคืนที่ให้ค่าตัววัด F1 สูงสุด จะได้ดังแสดงในตารางที่ 5.4

ตารางที่ 5.4 การเปรียบเทียบผลการค้นคืนของงานวิจัยนี้กับผลที่ได้จากวิธีประยุทธ สุวรรณวิสารท

วิธีการเข้ารหัส	กรณีคำไทยทับศัพท์คำอังกฤษ			กรณีคำอังกฤษทับศัพท์คำไทย		
	ค่าแม่นยำ	ค่าเรียกคืน	ตัววัด F1	ค่าแม่นยำ	ค่าเรียกคืน	ตัววัด F1
ประยุทธ	72.43	90.25	80.33	92.27	76.00	83.33
งานวิจัยนี้	87.28	77.19	81.91	96.34	75.15	84.41

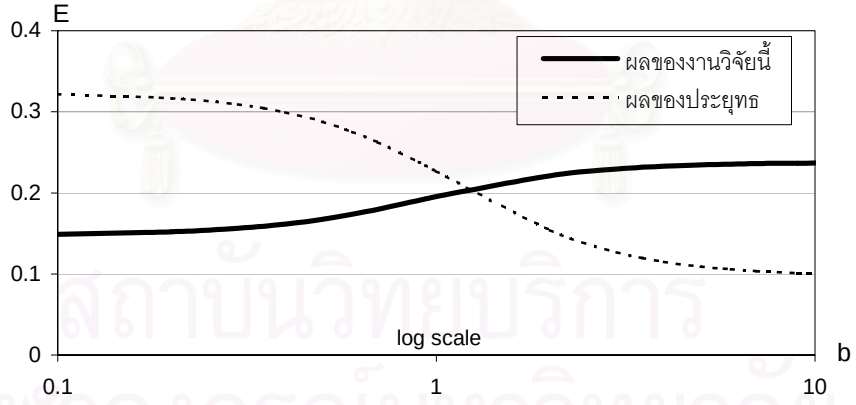
[4] ประยุทธ สุวรรณวิสารท, "การเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ" (วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2541).

จากตารางที่ 5.4 จะเห็นว่า ผลของการค้นคืนกรณีคำอังกฤษทับศัพท์คำไทยนั้นได้ผลใกล้เคียงกัน แต่สำหรับการค้นคืนกรณีคำไทยทับศัพท์คำอังกฤษนั้น ถึงแม้ว่าจะมีค่าตัววัด F1 ใกล้เคียงกัน (80.33% กับ 81.91%) แต่ผลของการค้นคืนมีพฤติกรรมของค่าความแม่นยำและค่าเรียกคืนที่ต่างกัน สามารถแสดงให้เห็นได้โดยใช้ตัววัด E ซึ่งเป็นอีกค่าหนึ่งที่นิยมใช้หาค่าโดยรวมของค่าแม่นยำและค่าเรียกคืน[5] ซึ่งมีสูตรดังนี้

$$E = 1 - \frac{(1 + b^2) * \text{ค่าแม่นยำ} * \text{ค่าเรียกคืน}}{(b^2 * \text{ค่าแม่นยำ}) + \text{ค่าเรียกคืน}}$$

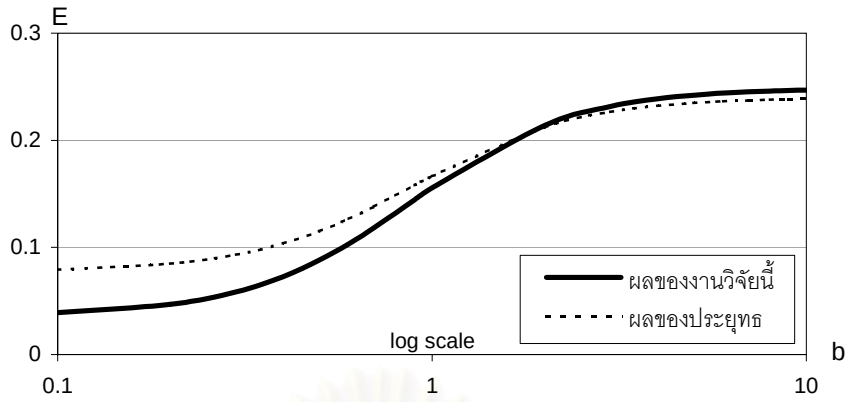
โดยที่ b เป็นค่ากำหนดน้ำหนักความสำคัญระหว่างค่าแม่นยำและค่าเรียกคืนที่สนใจในการค้นคืน เช่น ถ้า b = 1 หมายถึง กรณีให้ความสำคัญของค่าแม่นยำกับค่าเรียกคืนเท่ากัน (คล้ายตัววัด F1) ถ้า b = 0.1 แสดงว่าให้ความสำคัญของค่าแม่นยำมากกว่าค่าเรียกคืน 10 เท่า และในทางกลับกันถ้า b = 10 แสดงว่าให้ความสำคัญกับค่าเรียกคืนมากกว่าค่าแม่นยำ 10 เท่า การค้นคืนที่ให้ค่าของ E น้อยแสดงว่ามีความคุณภาพสูง

เมื่อนำผลที่ได้จากตารางที่ 5.4 มาวาดกราฟเปรียบเทียบพฤติกรรมการค้นคืนเมื่อแปรค่า b จาก 0.1 ถึง 10 จะได้ดังแสดงในรูปที่ 5.1 และ 5.2



รูปที่ 5.1 กราฟแสดงการเปรียบเทียบค่า E ของการค้นคืนคำไทยทับศัพท์คำอังกฤษที่ได้จากงานวิจัยกับผลที่ได้จากวิธีของประยุทธ สุวรรณวิสารท

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



รูปที่ 5.2 กราฟแสดงการเปรียบเทียบค่า E ของการค้นคืนคำอังกฤษทับศัพท์คำไทย ที่ได้จากงานวิจัยกับผลที่ได้จากวิธีของประยุทธ สุวรรณวิสารท

ในรูปที่ 5.2 ซึ่งเป็นกราฟของกรณีการค้นคืนคำอังกฤษทับศัพท์คำไทยนั้น แสดงให้เห็นว่าให้ผลของการค้นคืนในลักษณะคล้ายกัน แต่ในรูปที่ 5.1 ซึ่งเป็นกราฟของกรณีการค้นคืนคำไทยทับศัพท์คำอังกฤษนั้นมีพฤติกรรมต่างกัน กล่าวคือ การเปลี่ยนค่า b ในระบบการค้นคืนที่ได้จากงานวิจัยนี้จะมีผลต่อคุณภาพการค้นคืนที่น้อยกว่า (เส้นทึบมีความลาดชันน้อยกว่า) อีกทั้งได้ผลการค้นคืนที่ดีกว่า (E มีค่าต่ำกว่า) เมื่อให้ความสนใจกับค่าความแม่นยำมากกว่าค่าเรียกคืน แต่ผลการค้นคืนนี้ก็ไม่ลดคุณภาพไปมากนักเมื่อให้ความสำคัญที่มากขึ้นกับค่าเรียกคืน

นอกจากนี้ ผู้วิจัยยังได้ทำการทดลองโดยการเปลี่ยนแปลงค่าพารามิเตอร์ K (จำนวนส่วนของข้อมูลทั้งหมดที่ถูกแบ่งให้เท่า ๆ กัน) เพื่อหาว่าขนาดของข้อมูลจะมีผลอย่างไรต่อประสิทธิภาพในการค้นคืนของคำทับศัพท์ทั้งสองประเภทด้วยขั้นตอนวิธีที่นำเสนอเมื่อเทียบกับผลการทดลองที่ได้จากการเข้ารหัสคำโดยใช้ขั้นตอนวิธีของประยุทธ สุวรรณวิสารท ได้ผลดังแสดงในตารางที่ 5.5 และ 5.6

ตารางที่ 5.5 ผลการทดลองกรณีคำไทยทับศัพท์คำอังกฤษเมื่อแปรค่า K

K	จำนวนคำ(คู่)	ตัววัด F1	
		งานวิจัยนี้	ประยุทธ
4	469	81.91	80.36
3	625	80.46	77.34
2	938	76.56	73.16

ตารางที่ 5.6 ผลการทดลองกรณีคำอังกฤษทับศัพท์คำไทยเมื่อแปรค่า K

K	จำนวนคำ(คู่)	ตัววัด F1	
		งานวิจัยนี้	ประยุทธ
5	400	84.41	83.33
3	667	82.83	81.88
2	1,000	80.72	80.43

จากตารางที่ 5.5 และ 5.6 จะเห็นว่าเมื่อขนาดของข้อมูลมีค่าเพิ่มขึ้น ค่าตัววัด F1 ของงานวิจัยนี้ยังคงมีค่ามากกว่าผลการทดลองที่ได้จากการเข้ารหัสคำโดยใช้ขั้นตอนวิธีของประยุทธ แสดงให้เห็นว่า แนวโน้มพฤติกรรมของค่าตัววัด F1 ของงานวิจัยนี้น่าจะยังคงมีค่ามากกว่าเสมอ แม้ว่าข้อมูลจะมีขนาดใหญ่ขึ้น

5.3 **สรุป**

ผลการทดลองขั้นตอนวิธีที่นำเสนอ พบว่าเมื่อเกณฑ์ในการยอมรับค่าความแตกต่างของรหัสคำมีค่าเท่ากับ 1 ทั้งกรณีของคำไทยทับศัพท์คำอังกฤษและคำอังกฤษทับศัพท์คำไทยจะให้ค่าตัววัด F1 เป็นค่าที่สูงเกิน 80% ซึ่งเป็นค่าที่สูงกว่าผลการทดลองที่ได้จากการเข้ารหัสคำโดยใช้ขั้นตอนวิธีของประยุทธ สุวรรณวิสารทเล็กน้อย และมีแนวโน้มที่จะมากกว่าเมื่อข้อมูลมีขนาดใหญ่ขึ้น

บทที่ 6
สรุปผลการวิจัยและข้อเสนอแนะ

ระบบค้นคืนสารสนเทศมีการพัฒนามาเป็นเวลานาน แต่ยังคงมีปัญหาอีกมากมาย และปัญหาที่ผู้วิจัยค้นพบว่าเป็นหนึ่งในปัญหาที่น่าสนใจ คือ การค้นคืนสารสนเทศข้ามภาษา โดยเฉพาะการค้นคืนข้ามภาษาไทย-อังกฤษ ซึ่งจะทำให้ประสิทธิภาพการค้นคืนได้ดีขึ้น จากการศึกษาที่ผ่านมาระบบค้นคืนสารสนเทศข้ามภาษามีหลายแนวทางที่ใช้ในการแก้ปัญหาซึ่งมีข้อดีและข้อเสียแตกต่างกันไป สำหรับแนวทางที่ใช้ในการวิจัยนี้ จะเป็นการเข้ารหัสคำทับศัพท์โดยนำแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กมาใช้

6.1 สรุปผลการวิจัย

ผู้วิจัยได้เสนอขั้นตอนวิธีการเข้ารหัสคำ โดยนำแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กมาใช้ในการเรียนรู้และสร้างรหัสคำ ในขั้นตอนวิธีการฝึกให้นิวรอลเน็ตเวิร์กเรียนรู้การสร้างรหัสคำ จะใช้ความรู้ทางภาษาศาสตร์ในการพิจารณาข้อมูลเข้าและข้อมูลออก โดยข้อมูลที่ใช้เป็นคำหลักในชุดข้อมูลฝึก ในกรณีที่คำหลักนั้นเป็นคำอังกฤษสามารถนำไปสร้างเป็นข้อมูลเข้าได้เลย แต่หากเป็นคำไทยต้องผ่านการประมวลผลตัวอักษรเบื้องต้นก่อน เพื่อตัดตัวอักษรที่ไม่มีผลต่อการพิจารณารหัสออก และเปลี่ยนรูปสระบางตัวเพื่อให้สะดวกต่อการประมวลผล หลังจากนั้นนำคำที่ได้มาสร้างเป็นข้อมูลเข้าเพื่อส่งให้นิวรอลเน็ตเวิร์กเรียนรู้ ข้อมูลเข้ามีลักษณะเป็นชุดของตัวอักษรในคำ แต่ละชุดจะประกอบด้วยตัวอักษร 9 ตัว เป็นตัวที่สนใจให้รหัส 1 ตัวพร้อมทั้งตัวอักษรข้างเคียงหน้าและหลังข้างละ 4 ตัว ข้อมูลออกเป็นรหัสเสียงของตัวอักษรขาเข้าที่สนใจให้รหัส

ในขั้นตอนวิธีการเข้ารหัสคำ จะนำน้ำหนักของนิวรอลเน็ตเวิร์กที่ผ่านการฝึกแล้วให้ผลการเรียนรู้ที่ดีที่สุดมาใช้ในการสร้างรหัสเสียง โดยจะนำคำหลักมาสร้างเป็นข้อมูลเข้า แล้วส่งให้นิวรอลเน็ตเวิร์กสร้างข้อมูลออกซึ่งก็คือรหัสเสียง จากนั้นนำรหัสเสียงที่ได้มาสร้างเป็นรหัสคำ

ในขั้นตอนการค้นคืน จะทำการเข้ารหัสคำหลักที่ต้องการค้นหา แล้วนำรหัสคำที่ได้ไปเปรียบเทียบกับรหัสคำในดัชนีคำหลักของเอกสารที่ได้เข้ารหัสไว้แล้วในขั้นตอนการทำดัชนี ซึ่งใช้การเปรียบเทียบเชิงประมาณ โดยมีการกำหนดเกณฑ์ที่ใช้ในการยอมรับค่าความแตกต่าง ถ้ารหัสคำทั้งสองมีค่าความแตกต่างไม่เกินเกณฑ์ที่กำหนด ให้ถือว่าเป็นรหัสคำที่ได้มาจากคำทับศัพท์ที่ตรงกันระหว่างภาษาไทยและภาษาอังกฤษ

จากการทดลองพบว่า ประสิทธิภาพในการค้นคืนของคำทับศัพท์ทั้งสองประเภทซึ่งวัดโดยใช้ตัววัด F1 มีค่าสูงเกิน 80% เมื่อเกณฑ์ในการยอมรับค่าความแตกต่างของรหัสคำเป็น 1

6.2 ข้อดีและข้อเสียของขั้นตอนวิธี

ขั้นตอนวิธีที่นำเสนอนี้ มีทั้งข้อดีและข้อเสียซึ่งสามารถแจกแจงได้ดังนี้

6.2.1 ข้อดี

- ขั้นตอนวิธีสำหรับคำทับศัพท์ทั้งสองประเภทเป็นขั้นตอนวิธีเดียวกัน ทำให้สะดวกในการใช้งาน
- ทำงานได้โดยไม่ต้องใช้พจนานุกรม ทำให้สามารถใช้กับคำทับศัพท์ใหม่ ๆ ที่เกิดขึ้นได้
- สามารถเพิ่มค่าเรียกคืนให้กับคำพ้องเสียงได้ เพราะใช้หลักการอ่านออกเสียงของคำในการสร้างรหัสคำ
- ใช้งานได้กับทั้งคำทับศัพท์ที่ถูกต้องตามหลักเกณฑ์ของราชบัณฑิตยสถานหรือเขียนตามความนิยม

6.2.2 ข้อเสีย

- ใช้หน่วยความจำในการเก็บค่าน้ำหนักของนิวรอลเน็ตเวิร์กมากกว่าขั้นตอนวิธีของงานวิจัยที่ผ่านมา

6.3 ข้อเสนอแนะ

ผู้วิจัยพบว่า มีข้อเสนอแนะบางประการที่น่าจะเป็นประโยชน์ และสามารถนำไปใช้ในการพัฒนาขั้นตอนวิธีการเข้ารหัสคำทับศัพท์ภาษาไทย-อังกฤษ เพื่อให้มีประสิทธิภาพในการค้นคืนข้ามภาษาที่ดีขึ้น ดังนี้

- นอกจากวิธีการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กซึ่งใช้ในการเข้ารหัสในการทดลองนี้แล้ว ยังมีการเรียนรู้ด้วยเครื่องวิธีอื่น ๆ อีกมากที่น่าสนใจซึ่งน่าจะนำมาประยุกต์ใช้ในการเข้ารหัสได้
- สำหรับคำไทย อาจไม่จำเป็นต้องทำการประมวลผลตัวอักษรเบื้องต้น เช่น ตัดตัวการันต์และวรรณยุกต์หรือเปลี่ยนรูปสระบางตัว แต่ส่งไปให้นิวรอลเน็ตเวิร์กทำการเรียนรู้ในส่วนนี้แทน
- ในขั้นตอนการฝึกคำไทย นอกจากวิธีที่นำเสนอแล้วยังมีวิธีฝึกที่น่าสนใจอีก ดังนี้
  - กรณีตัวอักษร อ ฝึกโดยให้รหัสเสียงสำหรับตัวอักษร อ เป็น _ ส่วนสระที่อยู่หลังตัวอักษร อ ให้รหัสเป็นรหัสเสียงของสระนั้น เช่น คำว่า “อา” ให้รหัสเสียงเป็น /_a/ หรือคำว่า “อิน” ให้รหัสเสียงเป็น /_in/

- กรณีสระประสม ฝึกโดยเลือกให้รหัสเสียงสำหรับตัวอักษรตัวสุดท้ายของกลุ่ม ส่วนตัวที่เหลือให้รหัสเป็น _ เช่น คำว่า “เตอ” ให้รหัสเสียงเป็น /_tW/ หรือคำว่า “เรีย” ให้รหัสเสียงเป็น /_r_I/

- การกำหนดชุดฝึกที่เหมาะสม โดยให้มีคำทับศัพท์ที่หลากหลายประเภท และให้ครอบคลุมในหลาย ๆ กรณี จะทำให้ประสิทธิภาพในการเข้ารหัสดีขึ้น

รายการอ้างอิง

ภาษาไทย

กระทรวงศึกษาธิการ. หนังสือเรียนวิชาเคมี เล่ม 1 หลักสูตรมัธยมปลาย 2524, 2530.

จินดา เฮงสมบูรณ์. ภาษาศาสตร์เบื้องต้น. พิมพ์ครั้งที่ 1. กรุงเทพมหานคร : สุวีริยาสาส์น, 2542.

ประยุทธ สุวรรณวิสารท. การเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2541.

พยนต์ ทิมเจริญ. การเขียนชื่อภาษาไทยด้วยอักษรโรมัน. วารสารแผนที่ ปีที่ 27 ฉบับที่ 2 (ตุลาคม–ธันวาคม 2527) : 61–74.

ราชบัณฑิตยสถาน. หลักเกณฑ์การทับศัพท์. (ม.ป.ท.), 2535.

ราชบัณฑิตยสถาน. ศัพท์วิทยาศาสตร์. กรุงเทพมหานคร : สหธรรมิก, 2536.

ราชบัณฑิตยสถาน. ศัพท์คณิตศาสตร์. กรุงเทพมหานคร : โรงพิมพ์ มหาจุฬาลงกรณราชวิทยาลัย, 2540.

อุไรรัตน์ บุญภานนท์. การถอดอักษรภาษาอังกฤษเป็นไทยโดยใช้หลักวิชาภาษาศาสตร์. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2529.

ภาษาอังกฤษ

A. Binstock and J. Rex. Practical Algorithms for Programmers. New York : Addison Wesley, 1995.

C. J. Van Rijsbergen. Information Retrieval. Butterworths, London, 1979.

D. 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.

E. Rich and K. Knight. Artificial Intelligence. Singapore : Prentice–Hill, 1991.

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

J. D. O'Connor. Better English Pronunciation. 2nd Edition. Cambridge : Cambridge University Press, 1980.

J. Zobel and P. Dart. Phonetic String Matching: Lessons from Information Retrieval. Proceedings of the 19th Annual International ACM SIGR Conference on Research and Development in Information Retrieval. 1996 : 166-172.

T. M. Michell. Machine Learning. New York : McGraw-Hill, 1997.

P. Suwanvisat and S. Prasitjutrakul. Thai–English Cross–Language Transliterated Word Retrieval using Soundex Technique. Proc. of the National Computer Science and Engineering Conference. 1998.

P. Suwanvisat and S. Prasitjutrakul. Transliterated Word Encoding and Retrieval Algorithms for Thai–English Cross–Language Retrieval. Proc. of the National Computer Science and Engineering Conference. 1999.

R. R. Leighton. The Aspirin/MIGRAINES Neural Network Software 6.0. The MITRE Corporation, 1992.

S. Ongroongruang, R. Prongsirivattana, and V. Jantarasukree. English to Thai Word Retrieval Using Sound Index. Proc. 2nd SNLP 95. 1995.

W. B. Frakes and R. B. Yates. Information Retrieval : Data Structures & Algorithms. Englewood Cliffs, NJ : Prentice Hall, 1992.

ภาคผนวก

ภาคผนวก ก
การใช้งานโปรแกรมในขั้นตอนการเรียนรู้

โปรแกรมที่ใช้

โปรแกรมที่ใช้ในการเรียนรู้และหาค่าน้ำหนัก สร้างขึ้นมาจากเครื่องมือที่ใช้สำหรับพัฒนาการจำลองนิวรอลเน็ตเวิร์คที่ชื่อ Aspirin/MIGRAINES 6.0[1]

<u>โครงสร้างนิวรอลเน็ตเวิร์ก</u>

กรณีคำไทยทับศัพท์คำอังกฤษ

1. สำหรับคำอังกฤษ เช่น CLINTON

```
DefineBlackBox a
{
  OutputLayer->  Output_Layer
  InputSize->    [234 x 1]
  Components->
  {
     PdpNode Output_Layer [39]
       {
              InputsFrom-> Hidden_Layer
       }
     PdpNode Hidden_Layer [234]
       {
              InputsFrom-> $INPUTS
       }
  }
}
```

2. สำหรับคำไทย เช่น คลินตัน

```
DefineBlackBox a
{
  OutputLayer->  Output_Layer
  InputSize->    [549 x 1]
  Components->
  {
```

---

[1] R. R. Leighton, <u>The Aspirin/MIGRAINES Neural Network Software 6.0 (AM6)</u> The MITRE Corporation, 1992.

```
        PdpNode Output_Layer [39]
          {
                InputsFrom-> Hidden_Layer
          }
        PdpNode Hidden_Layer [61]
          {
                InputsFrom-> $INPUTS
          }
    }
}
```

กรณีคำอังกฤษทับศัพท์คำไทย

1. สำหรับคำอังกฤษ เช่น CHULALONGKORN

```
DefineBlackBox a
{
  OutputLayer->  Output_Layer
  InputSize->    [234 x 1]
  Components->
  {
    PdpNode Output_Layer [35]
      {
            InputsFrom-> Hidden_Layer
      }
    PdpNode Hidden_Layer [234]
      {
            InputsFrom-> $INPUTS
      }
  }
}
```

2. สำหรับคำไทย เช่น จุฬาลงกรณ์

```
DefineBlackBox a
{
  OutputLayer->  Output_Layer
  InputSize->    [549 x 1]
  Components->
  {
    PdpNode Output_Layer [35]
      {
            InputsFrom-> Hidden_Layer
```

```
        }
      PdpNode Hidden_Layer [61]
        {
            InputsFrom-> $INPUTS
        }
    }
}
```

<u>พารามิเตอร์ที่ใช้</u>

รูปแบบ <executable file> -l -d <data file> -a <learning rate>

กำหนดค่าดังนี้ encode -l –d data.df –a 0.05

พารามิเตอร์ที่ใช้มีรูปแบบดังนี้

-l ให้โปรแกรมทำงานในขั้นตอนการเรียนรู้

-d <data file> ใช้กำหนดชุดข้อมูลที่ใช้ในโปรแกรม

-a <learning rate> ใช้กำหนดอัตราการเรียนรู้

ภาคผนวก ข
ตัวอย่างข้อมูลคำทับศัพท์ที่ใช้ในการทดลอง

คำภาษาไทยทับศัพท์ภาษาอังกฤษ

simon	ซิโมน	phylum	ไฟลัม	tartrazine	ตาร์ตราซีน
interphase	อินเตอร์เฟส	double	ดับเบิล	calvin	แคลวิน
celica	เซลิกา	hund	ฮุนด์	silicon	ซิลิคอน
metaphase	เมทาเฟส	leucoplast	ลิวโคพลาสต์	bridge	บริดจ์
kinetochore	ไคนีโทคอร์	glycerin	กลีเซอริน	kocher	โคเชอร์
brook	บรุก	hyde	ไฮด์	sure	ชัวร์
ribosome	ไรโบโซม	fermi	เฟอร์มี	burette	บิวเรตต์
aluminium	อะลูมิเนียม	volterra	โวลแตร์รา	oersted	เออร์สเตด
sony	โซนี่	barbary	บาร์บะริ	hopkins	ฮอปกินส์
polonium	พอโลเนียม	norm	นอร์ม	love	เลิฟ
celsus	เคลซุส	rectum	เรกตัม	halogen	แฮโลเจน
switzerland	สวิสเซอร์แลนด์	bonus	โบนัส	cambium	แคมเบียม
alessandro	อาเลสซันโดร	starling	สตาร์ลิง	chromosome	โครโมโซม
busy	บิซี	millikan	มิลลิแกน	correns	คอร์เรนส์
subnormal	ซับนอร์มอล	gustav	กุสตาฟ	americium	อะเมริเซียม
phosphorous	ฟอสฟอรัส	benthon	เบนทอน	johnstone	จอห์นสโตน
barents	แบเร็นตส์	hausdorff	เฮาส์ดอฟฟ์	cleanex	คลีเน็กซ์
placenta	พลาเซนตา	monoid	โมนอยด์	anode	แอโนด
polymer	พอลิเมอร์	ton	ตัน	molality	โมแลลิตี
crookes	ครูกส์	cosmic	คอสมิก	seymour	เซย์เมอร์
rocky	รอกกี	anastasio	อานัสตาซีโอ	agassiz	แอกาซี
antipodal	แอนติโพแดล	spain	สเปน	mulliken	มัลลิเกน
locket	ล็อกเกต	minor	ไมเนอร์	ricci	ริกชี
soup	ซุป	wilkins	วิลคินส์	tros	ทรอส

คำภาษาอังกฤษทับศัพท์ภาษาไทย

athasart	อัตถศาสตร์	munsaithong	มั่นไทรทอง	suntatisap	สันตติทรัพย์
plengsombut	เปล่งสมบัติ	varin	วรินทร์	sukhumavasi	สุขุมาวาสี
narinluk	นรินลักษณ์	chunvimonsiri	ชุนวิมลศิริ	mujananon	มุจนานนท์
piyawat	ปิยะวัฒน์	sommatat	โสมทัต	phiphob	พิภพ
saripant	ศริพันธุ์	supornkhunt	สุพลขันธ์	supornpan	สุพรพันธ์
kijmedee	กิจมีดี	wirat	วิรัช	lertvanasbodi	เลิศวนัสบดี
tanu	ฐานุ	dumrongsak	ดำรงค์ศักดิ์	wattanachai	วัฒนชัย
visuth	วิสูตร	chalomkwan	ชโลมขวัญ	sitthiwitch	สิทธิวิชญ์
rosaya	รสญา	chaiyos	ชัยยศ	smavatkul	ศมาวรรตกุล
laksanaprom	ลักษณะพรหม	kongsakul	คงสกุล	horanont	โหรานนท์
nitas	นิทรรศ	iramaneerat	ไอรมณีรัตน์	nirachorn	นิรชร
jithavech	จิตรถเวช	pairor	ไพเราะ	kamolporn	กมลพร
toasak	ต่อศักดิ์	jitpraneechai	จิตต์ปราณีชัย	mali	มะลิ
chitman	จิตต์มั่น	thunchai	ทุนชัย	kemthong	เข็มทอง
thunyawee	ธันยวีร์	patanapiradej	พัฒนพีระเดช	prompatima	พรหมปฏิมา
nuntiya	นันทิยา	santiwikranon	สันติวิกรานนท์	chittra	จิตรา
passanee	ภาษณี	piyavechvirat	ปิยะเวชวิรัตน์	wanwaew	วรรณแวว
numphung	น้ำผึ้ง	surakit	สุรกิจ	sumalee	สุมาลี
sunpoj	สรรพจน์	apakorn	อาภากร	umaporn	อุมาพร
ankanasopit	อังคณาโสภิต	kongmebhol	คงมีผล	waranya	วรัญญา
ponsup	ผลทรัพย์	sirirat	ศิริรัตน์	phisit	พิสิต
sivachat	ศิวชาติ	pornsukjantra	พรสุขจันทรา	oupala	อุปลา
pitakannop	พิทักษ์อรรณพ	nurack	นุรักษ์	yuenyongolan	ยืนยงโอฬาร
rakchaiwan	รักษ์ไชยวรรณ	piyada	ปิยดา	navee	นาวี
dusit	ดุสิต	rudeechanok	ฤดีชนก	roongruedee	รุ่งฤดี
srithongkul	ศรีทองกุล	patcharaporn	พัชราภรณ์	satamool	สาตมูล

ประวัติผู้เขียน

นางสาวทัศนวรรณ ศูนย์กลาง เกิดวันที่ 25 กุมภาพันธ์ พ.ศ.2518 ที่กรุงเทพมหานคร สำเร็จการศึกษาระดับปริญญาตรีวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร ในปีการศึกษา 2538 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ ที่จุฬาลงกรณ์มหาวิทยาลัย เมื่อ พ.ศ.2541 ปัจจุบันรับราชการที่คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร วิทยาเขตพระราชวังสนามจันทร์