การเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษเพื่อการค้นคืนข้ามภาษาด้วยเทคนิคนิวรอลเน็ตเวิร์ก

Thai–English Transliterated Word Encoding for Cross–Language Retrieval using Neural Networks

ทัศนวรรณ ศูนย์กลาง สมชาย ประสิทธิ์จูตระกูล บุญเสริม กิจศิริกุล
Tasanawan Soonklang Somchai Prasitjutrakul Boonserm Kijsirikul
Department of Computer Engineering
Chulalongkorn University
Bangkok, Thailand
E-Mail: g41tsk@cp.eng.chula.ac.th, somchaip@chula.ac.th, boonserm@cp.eng.chula.ac.th

บทคัดย่อ: บทความนี้นำเสนอการเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษโดยใช้นิวรอลเน็ตเวิร์กเพื่อการค้นคืนข้ามภาษา นิวรอลเน็ตเวิร์กที่ใช้เป็นแบบแบ็กพรอพาเกชันซึ่งรับข้อมูลขาเข้าเป็นตัวอักขระที่สนใจให้รหัสในคำ พร้อมทั้งตัวอักขระข้างเคียงหน้าหลังข้างละสี่ตัวของคำ และให้ข้อมูลขาออกเป็นรหัสเสียงของตัวอักขระขาเข้านั้น ผลที่ได้คือ นิวรอลเน็ตเวิร์กที่ใช้เข้ารหัสที่ใช้ได้กับคำอังกฤษที่ทับศัพท์คำไทย และคำไทยที่ทับศัพท์คำอังกฤษ ขั้นตอนการค้นคืนคำข้ามภาษาอาศัยการเปรียบเทียบรหัสของคำแบบประมาณ โดยอนุญาตให้มีความแตกต่างของรหัสที่นำมาเปรียบเทียบได้ไม่เกินหนึ่ง ผลการทดลองด้วยวิธี K–fold cross validation พบว่า การค้นคืนได้ค่าตัววัด F1 (ซึ่งเป็นผลการเฉลี่ยของค่าความเที่ยง และค่าเรียกคืน) ของการค้นคืนทั้งกรณีคำอังกฤษทับศัพท์คำไทยและกรณีคำไทยทับศัพท์อังกฤษที่สูงเกิน 80%

Abstract: This paper presents Thai–English transliterated word encoding for cross–language retrieval system using backpropagation neural networks. By successively feeding each character of the word along with its eight neighboring (preceding 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 approximately matched (with allowable edit distance of one) during retrieval. The system can be used in both Thai–to–English and English–to–Thai transliterated words. Experimental results using K–fold cross validation technique showed that a high F1–measure (which is an average of precision and recall measurements) of more than 80% can be achieved.
Key words: Information retrieval, Cross-language, Backpropagation neural network, Transliterated word

1. บทนำ

การค้นคืนสารสนเทศข้ามภาษา (Cross–Language Information Retrieval) หมายถึง การค้นคืนสารสนเทศซึ่งภาษาที่แสดงในเอกสารไม่ตรงกับภาษาที่แสดงในการสอบถาม [1] ปัจจุบันเอกสารทางวิชาการในประเทศไทยมักจะจัดทำทั้งในรูปภาษาไทยและภาษาอังกฤษเพื่อประโยชน์ในการเผยแพร่ทั้งภายในและภายนอกประเทศ ซึ่งเอกสารเหล่านี้โดยเฉพาะอย่างยิ่งเอกสารทางด้านวิทยาศาสตร์และวิศวกรรมศาสตร์โดยมากแล้วมักจะปรากฏคำนามเฉพาะ (Proper Noun) และคำศัพท์เทคนิคต่าง ๆ เป็นจำนวนมาก ซึ่งคำทับศัพท์จะเป็นคำอังกฤษทับศัพท์คำไทย เช่น “Somchai” ทับศัพท์ “สมชาย” หรือ

คำไทยทับศัพท์คำอังกฤษ เช่น “คอมพิวเตอร์” ทับศัพท์ “Computer” ก็ได้ ดังนั้นระบบค้นคืนสารสนเทศข้ามภาษาจึงมีความสำคัญในการเพิ่มประสิทธิภาพของการใช้ประโยชน์จากสารสนเทศที่มีอยู่ให้ได้อย่างเต็มที่

ปัญหาในการค้นคืนสารสนเทศมีหลายประการ โดยเฉพาะการค้นคืนข้ามภาษาซึ่งคำในภาษาหนึ่งอาจจะถูกเขียนในอีกภาษาหนึ่งได้หลายรูปแบบ เช่น “Carbohydrate” ในภาษาไทยอาจพบได้ทั้ง “คาร์โบไฮเดรต” “คาร์โบไฮเดรท”หรือ “คาร์โบฮัยเดรต” หรือ “ประภาส” อาจเขียนอยู่ในรูป “Prapass” หรือ “Prabhas” ซึ่งระบบค้นคืนควรจะคืนคำเหล่านี้มาให้ได้ทั้งหมดหรือให้ได้มากที่สุด การนำพจนานุกรมสองภาษา (Bilingual Dictionary) มาใช้ในระบบค้นคืนสารสนเทศก็ไม่อาจแก้ปัญหานี้ได้มากนัก เนื่องจากมีคำศัพท์เทคนิคใหม่ ๆ มากมายในหลากหลายสาขาเกิดขึ้นแทบทุกวัน และคำทับศัพท์ส่วนมากมักไม่ปรากฏในพจนานุกรม โดยเฉพาะในปัจจุบันสารสนเทศในสื่ออิเล็กทรอนิกส์มีจำนวนเพิ่มขึ้นอย่างรวดเร็วมาก ทำให้ปัญหาดังกล่าวยิ่งเพิ่มมากขึ้นจนระบบค้นคืนทั่วไปที่มีอยู่ไม่สามารถแก้ปัญหาได้

งานวิจัยต่าง ๆ ที่เกี่ยวข้องกับปัญหาการค้นคืนข้ามภาษาไทย–อังกฤษ ได้แก่ [2] นำเสนอขั้นตอนวิธีสำหรับเข้ารหัสคำภาษาอังกฤษ ซึ่งรหัสคำที่ได้จะเป็นกลุ่มของเสียงอ่านที่เป็นไปได้ในภาษาไทย การเข้ารหัสคำจะอาศัยตารางการกำหนดรหัสและกฎ แต่งานวิจัยนั้นไม่ได้เสนอตารางการกำหนดรหัส และไม่ได้รายงานผลการทดลอง [3] นำเสนอขั้นตอนวิธีการเข้ารหัสคำสำหรับคำไทยทับศัพท์คำอังกฤษโดยใช้เทคนิคการเข้ารหัสคำแบบซาวเด็กซ์ และ [4] นำเสนอขั้นตอนวิธีการเข้ารหัสคำสำหรับคำอังกฤษทับศัพท์คำไทยโดยใช้อาศัยหลักการทางภาษาศาสตร์ในการเข้ารหัสคำ และอาศัยตารางการกำหนดต้นทุนในการแทนที่อักขระในการคำนวณหาค่าความแตกต่างของรหัสคำในขั้นตอนการค้นคืน แต่งานวิจัยนี้ไม่ได้นำเสนอตารางการกำหนดต้นทุนในการแทนที่อักขระ

การวิจัยครั้งนี้มุ่งเน้นการเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษโดยใช้เทคนิคนิวรอลเน็ตเวิร์ก เพื่อการค้นคืนข้ามภาษา การเข้ารหัสคำทับศัพท์ทั้งคำอังกฤษทับศัพท์คำไทย และคำไทยทับศัพท์คำอังกฤษ นำเสนอโดยใช้ขั้นตอนวิธีเดียวกัน และให้ประสิทธิภาพการค้นคืนที่ดีกว่า [3] และ [4] ในขณะที่ [3] และ [4] ใช้ขั้นตอนวิธีที่แตกต่างกัน สำหรับการเข้ารหัสคำอังกฤษทับศัพท์คำไทย และคำไทยทับศัพท์คำอังกฤษ โดยใช้กฎการเข้ารหัสที่ตายตัว และใช้ตารางการกำหนดต้นทุนในการแทนที่อักขระ เพื่อนำไปใช้ในการเปรียบเทียบรหัสที่ค่อนข้างซับซ้อนและไม่มีหลักเกณฑ์แน่นอนที่ใช้ในการสร้างตาราง

บทความนี้จะนำเสนอขั้นตอนวิธีการเข้ารหัสคำอย่างละเอียดในหัวข้อต่อไป และอธิบายขั้นตอนการค้นคืนในหัวข้อที่สาม หัวข้อที่สี่นำเสนอวิธีการทดลองประสิทธิผลและผลการทดลอง จากนั้นสรุปเนื้อหาของบทความในหัวข้อที่ห้า

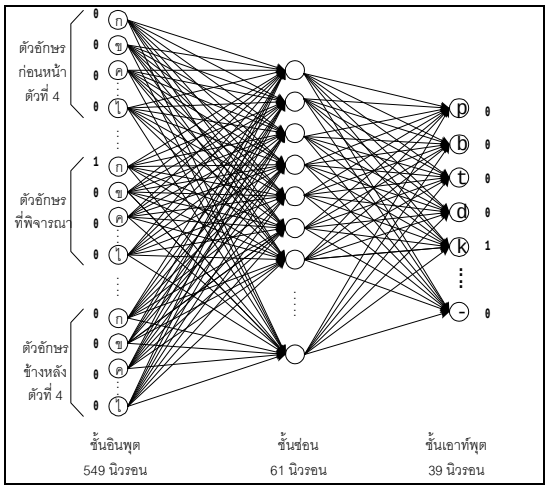
2. ขั้นตอนวิธีการเข้ารหัสคำ

ในงานวิจัยนี้ การเข้ารหัสคำมีจุดประสงค์เพื่อแปลงคำทับศัพท์ ทั้งที่อยู่ในรูปคำไทยและคำอังกฤษ ให้อยู่ในรูปรหัสคำอ่านรูปแบบเดียวกัน เพราะหากเป็นคำทับศัพท์ที่ตรงกันของทั้งสองภาษาจะอ่านออกเสียงได้เหมือนหรือคล้ายคลึงกัน รหัสคำของคำทับศัพท์ต่าง ๆ ในเอกสารจะถูกสร้างขึ้นในขั้นตอนการสร้างดัชนี ในระบบการจัดเก็บสารสนเทศ เมื่อผู้ใช้ป้อนข้อคำถามที่เป็นคำทับศัพท์ ระบบค้นคืนก็จะสร้างรหัสคำของคำต่าง ๆ ในข้อคำถามเพื่อใช้ค้นหากับรหัสคำที่จัดเก็บไว้ในดัชนี

ขั้นตอนวิธีการเข้ารหัสคำจะใช้แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก (Backpropagation Neural Networks) [5] ซึ่งเป็นวิธีการเรียนรู้ของเครื่องที่เหมาะสมสำหรับการนำไปใช้ในการแก้ปัญหาที่เกี่ยวข้องกับการจำแนกหรือแบ่งประเภทมาช่วยในการเรียนรู้รหัสคำอ่านของคำทับศัพท์ โดยจะใช้นิวรอลเน็ตเวิร์กให้เรียนรู้การสร้างรหัสคำอ่าน

ทั้งหมด 4 ชุด สำหรับ (1) คำไทย (2) คำอังกฤษทับศัพท์คำไทย และ (3) คำอังกฤษ (4) คำไทยทับศัพท์คำอังกฤษ โครงสร้างของนิวรอลเน็ตเวิร์กแสดงในรูปที่ 1

รูปที่ 1 ตัวอย่างโครงสร้างแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กที่ใช้ในการเรียนรู้คำไทย (ในกรณีคำอังกฤษทับศัพท์คำไทย) ซึ่งมีอินพุตเป็น (_ , _ , _ , _ , ก , น , ก , พ , ี) และมีเอาท์พุตเป็น 'k'



จากรูปที่ 1 โครงสร้างเน็ตเวิร์กประกอบด้วย 3 ชั้นดังนี้

<u>ชั้นอินพุต</u> (Input layer) ประกอบด้วยจำนวนนิวรอนเท่ากับ จำนวนอักขระทั้งหมดของภาษาที่พิจารณา คูณด้วย *จำนวนตัวอักษรทั้งหมดที่ใช้พิจารณา* (กำหนดให้เป็นค่าคงตัว m) ดังนั้น กรณีคำอังกฤษ ข้อมูลเข้าจะมี 26 x m นิวรอน กรณีคำไทย ข้อมูลเข้าจะมี 61 x m นิวรอน โดยค่า m ซึ่งคือจำนวนตัวอักษรทั้งหมดที่ใช้พิจารณานั้นมีค่าเป็น 9 เนื่องจากคำไทยมีการใช้สระที่เกิดจากการใช้ตัวอักษรตั้งแต่ 2 ตัวขึ้นไป เช่น สระเอีย ดังนั้นจึงควรพิจารณาตัวอักษรข้างเคียงด้วย เช่น คำว่า "เสถียร" เมื่อตัวอักษรที่กำลังพิจารณา คือ "เ" ซึ่งจะรู้ว่าเป็นส่วนประกอบของสระเอียก็โดยพิจารณาจากตัวอักษรข้างหลัง 4 ตัว หรือเมื่อตัวอักษรที่กำลังพิจารณา คือ "ย" จะรู้ได้ก็โดยพิจารณาจากตัวอักษรข้างหน้า 4 ตัว ในการทดลองนี้ได้ทำการทดสอบ เพื่อหาจำนวนตัวอักษรข้างเคียงที่ใช้พิจารณาร่วมแล้วได้ว่า จำนวนที่ให้ผลในการเรียนรู้ที่ดีคือ พิจารณาตัวอักษรข้างหน้า 4 ตัว ข้างหลังอีก 4 ตัว รวมกับตัวที่กำลังพิจารณาอีก 1 ตัว

<u>ชั้นซ่อน</u> (Hidden layer) ได้ทำการทดลองเพื่อหาจำนวนนิวรอนที่เหมาะสม (โดยได้จากค่าที่ใช้ในการฝึกแล้วให้ผลการเรียนรู้ดีที่สุด) สำหรับแต่ละเน็ตเวิร์ก ซึ่งสำหรับคำไทยจะมี 61 นิวรอน ส่วนคำอังกฤษจะมี 234 นิวรอน

ตารางที่ 1 รหัสเสียงสำหรับคำไทยทับศัพท์คำอังกฤษ

เสียงพยัญชนะ		รหัสเสียง	เสียงสระ		รหัสเสียง
ไทย	อังกฤษ		ไทย	อังกฤษ	
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	ู	–oo	U
ว	w, v	v	ั	U	V
ส,ซ	s, z	s	เ–อ	–ur,er,–ir	W
ฮ	h	h	ะ	a	a
ม	m	m	โ	–ome,o	o
น	n	n	ไ, ใ, –ั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
ตัวอักษรที่ไม่ออกเสียง		_	เ–า	ou,au	x
			เ–ิ	–or	q
			–ิว	–ew,eua	X
			เ–ิล	–le	Q

ชั้นเอาท์พุต (Output layer) จะมีจำนวนนิวรอนเท่ากับรหัสเสียงพยัญชนะและเสียงสระที่เป็นไปได้ทั้งหมด ซึ่งในกรณีคำไทย และคำอังกฤษทับศัพท์คำไทยจะมี 35 นิวรอน และในกรณีคำอังกฤษ และคำไทยทับศัพท์คำอังกฤษจะมี 39 นิวรอน (ดูตารางที่ 1 ตารางที่ 2 และตารางที่ 3 ประกอบ)

ตารางที่ 2 รหัสเสียงพยัญชนะสำหรับคำอังกฤษทับศัพท์คำไทย

เสียงพยัญชนะ		รหัสเสียง	
ไทย	อังกฤษ	ตัวต้น	ตัวสะกด
ก ข ค ฆ	ck, g, k, x, c, kh, q	k	K
ง	ng	g	G
จ ฉ ช ฌ	j, ch, x	c	T
ซ ส ศ ษ สร ทร	s, z	s	T
ญ ย หย หญ	y	y	N
ด ฎ ฑ	d	d	t
ต ฏ ถ ฐ ท ฑ ธ ฒ	t, th, dh	t	t
ณ น หน	n	n	n
บ	b	b	p
ป ผ พ ภ	p, ph, bh	p	p
ฝ ฟ	f	f	p
ม	m	m	m
ร ฤ	r	r	n
ล ฬ ฦ	l	l	n
ว	w, v	v	-
ห ฮ	h	h	-
ตัวอักษรที่ไม่ออกเสียง		_	_

ในการฝึก (Train) นิวรอลเน็ตเวิร์กให้เรียนรู้การสร้างรหัสคำอ่านนั้น ข้อมูลที่ใช้ฝึกได้มาจากการนำความรู้ทางภาษาศาสตร์ ได้แก่ หลักเกณฑ์ในการถอดอักษร [6] หลักการถ่ายเสียง [7] หลักการอ่านออกเสียงทั้งภาษาไทยและภาษาอังกฤษ [6] มาสร้างเป็นตารางรหัสเสียง

ตารางที่3 รหัสเสียงสระสำหรับคำอังกฤษทับศัพท์คำไทย

เสียงสระ		รหัสเสียง
ไทย	อังกฤษ	
— ี	i, ee	i
ะ า ั	a, u, ar	a
เ–ะ เ–	e	e
แ–ะ แ–	ae	w
– ึ ื –ือ–ุ –ู	u,eu,ue,eo,oo	u
โ–ะ โ– เ–าะ –อ	o	o
เ–อะ เ–อ เ็	er, oe	W
เ ีย–ะ เ ีย	ia, ie, aiu	I
เ–ือะ เ–ือ –ัวะ ัว –ว–	ua, ue, ea, ui	Y
ไ– ใ– ไ–ย ัย –าย	ai, ie, uy	!
เ–า –าว	ao, ou, ow	R
โ–ย –อย	oi, oy	x
ิว	iu	X
เ–ว	eo	q
เ–ย	oei	Q
เ–ือย –วย	uai, uay, ou	O
แ–ว	aeo, eo, aew	$
เ–ียว	ieo, eaw, eo, ew, iow, iau, iew, iaw	@

สำหรับคำไทยทับศัพท์คำอังกฤษ (ดังในตารางที่ 1) และตารางรหัสเสียงสำหรับคำอังกฤษทับศัพท์คำไทย (ดังในตารางที่ 2 และ 3) ซึ่งรหัสคำอ่านจะประกอบด้วยสองส่วน คือ ส่วนที่เป็นเสียงพยัญชนะและส่วนที่เป็นเสียงสระ และสำหรับการฝึกคำไทยจะต้องนำคำเหล่านี้มาประมวลผลตัวอักษรเบื้องต้นก่อน ได้แก่ การตัดไม้ไต่คู้ วรรณยุกต์ การันต์และอักษรควบที่มีตัวการันต์ออก เพราะตามหลักการถอดอักษรไทยเป็นอังกฤษจะไม่พิจารณาตัวอักษรเหล่านี้ [4] จากนั้นจึงนำคำไทยที่ได้หลังผ่านกระบวนการนี้ส่งให้นิวรอลเน็ตเวิร์กเรียนรู้ ส่วนคำอังกฤษสามารถส่งไปเรียนรู้ได้เลย หลังจากทำการฝึกเสร็จก็จะนำน้ำหนัก (weight) ของแต่ละเน็ตเวิร์ก

ที่ให้ผลการเรียนรู้ที่ดีที่สุดมาใช้ในการเข้ารหัสคำ เมื่อได้รหัสคำอ่านจากนิวรอลเน็ตเวิร์กแล้ว จะทำการตัดรหัสที่ไม่ออกเสียง (_) ออกและทำการย้ายรหัสเสียงสระไปต่อท้ายรหัสเสียงพยัญชนะ

ในการฝึก เช่นคำว่า “กนกพีระวุฒิ” เป็นรหัสคำ “kanokpiravut_” จะฝึกโดยทำการเลื่อนคำไปทีละหนึ่งตัวอักษร โดยในครั้งแรกจะฝึกด้วย (_ , _ , _ , _ , ก , น , ก , พ , ี) → k ซึ่งในชั้นอินพุตที่โหนด “ก” ของอักษรที่พิจารณา จะถูกกำหนดให้มีค่าเป็น 1 ในชั้นเอาท์พุตที่โหนด “k” ของรหัสเสียงจะถูกกำหนดให้มีค่าเป็น 1 ส่วนโหนดที่เหลือจะมีค่าเป็น 0 ในการทดสอบเอาท์พุตที่มีค่ามากที่สุดจะถูกเลือก ส่วนในกรณีของคำไทยที่มีการให้เสียงสระลดรูป ทำให้บางครั้งเอาท์พุตอาจมี 2 โหนด เราจะเลือกจำนวนเอาท์พุตว่าเป็น 1 หรือ 2 โดยนำเอาท์พุต 2 โหนดที่มีค่ามากที่สุดมาหาผลต่าง และเมื่อผลต่างมีค่าไม่เกินค่า threshold จะได้ว่ามีจำนวนเอาท์พุตเป็น 2 โหนด จากการทดลองได้ทำการคำนวณหาค่า threshold จนได้ค่าที่เหมาะสม คือ 0.3

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

1. ประมวลผลตัวอักษรเบื้องต้น จะได้ กนกพีระวุฒิ → กนกพีระวุฒิ เช่นเดิม
2. สร้างเป็นอินพุตเพื่อส่งให้นิวรอลเน็ตเวิร์ก โดยพิจารณาตัวอักษรทั้งหมดครั้งละ 9 ตัว โดยตัวที่สนใจคือตัวที่ 5 และพิจารณาตัวอักษรข้างเคียงข้างหน้า 4 ตัวและข้างหลังอีก 4 ตัว
   (_ , _ , _ , _ , 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 , ิ , _ , _ , _ , _) → _
3. ตัดรหัสที่ไม่ออกเสียง (_) ออกและทำการย้ายรหัสเสียงสระไปต่อท้ายรหัสเสียงพยัญชนะ
   kanokpiravut_ → kanokpiravut → knkprvtaoiau

3. ขั้นตอนการค้นคืน

รหัสคำที่ได้ของคำคู่ที่ตรงกันทั้งสองภาษาอาจจะไม่ตรงกันทุกตัวอักษร แต่จะมีลักษณะคล้ายกัน เนื่องจากหลักเกณฑ์การทับศัพท์ที่ใช้ในปัจจุบันมีหลายรูปแบบ เพื่อให้ได้ค่าความเที่ยง (Precision) และค่าเรียกคืน (Recall) ที่ดีจะใช้การเปรียบเทียบรหัสคำแบบประมาณ (Approximate matching) ซึ่งอาศัยการคำนวณความแตกต่างของรหัสคำด้วยเทคนิคระยะแก้ไขสั้นที่สุด (Minimal Edit distance) [8] โดยคำนวณหาจากจำนวนครั้งที่น้อยที่สุดที่ใช้ในการเพิ่ม การลบ และการแทนที่แต่ละตัวอักขระ เพื่อทำให้รหัสคำทั้งสองเหมือนกัน

การคำนวณความแตกต่างนี้อาศัยเทคนิคกำหนดการพลวัต (Dynamic programming) ซึ่งวิธีการคำนวณสามารถเขียนให้อยู่ในรูปการคำนวณด้วยความสัมพันธ์เวียนเกิด Edit (P_j, W_k) ดังนี้ [4]

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

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

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

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

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

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

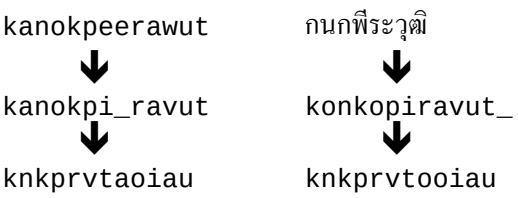
$r(p_j, w_k) = 0$ ถ้า p_j เท่ากับ w_k

$= 1$ ถ้า p_j ไม่เท่ากับ w_k

ถ้าความแตกต่างที่ได้มีค่าไม่เกินกว่าค่าที่ยอมรับได้ (ค่าคงที่ d) จะสรุปได้ว่า รหัสคำทั้งสองเป็นรหัสที่มาจากคำหลักที่ตรงกันในอีกภาษา

ตัวอย่าง ต้องการทดสอบคำว่า “kanokpeerawut” และ “กนกพีระวุฒิ” เป็นคำทับศัพท์ที่ตรงกันในภาษาไทย–อังกฤษหรือไม่ โดยทำการเข้ารหัสคำแล้วคำนวณหาค่าความแตกต่าง

การเข้ารหัสคำ



การค้นคืน

Edit (knkprvtaoiau, knkprvtooiau) = 1

จากตัวอย่าง รหัสคำทั้งสองมีค่าความแตกต่างเป็น 1 ถ้าเรากำหนดให้ d=1 ก็จะสามารถค้นคืน “kanokpeerawut” จาก “กนกพีระวุฒิ” ได้

4. ผลการทดลอง

ผู้วิจัยได้ทำการทดลองขั้นตอนวิธีที่ได้นำเสนอ ในกรณีคำไทยทับศัพท์คำอังกฤษ ใช้ชุดของคำอังกฤษและคำทับศัพท์ที่ตรงกัน ซึ่งเป็นคำนามเฉพาะ [9] คำศัพท์วิทยาศาสตร์ [10] คำศัพท์คณิตศาสตร์ [11] และคำศัพท์เคมี [12] จำนวน 1,876 คู่ และกรณีคำอังกฤษทับศัพท์คำไทย ใช้ชื่อและชื่อสกุลทั้งภาษาไทยและภาษาอังกฤษที่ตรงกันของนิสิต จำนวน 2,000 คู่ การทดลองกระทำโดยการแบ่งข้อมูลเป็นข้อมูลฝึก (train set) และข้อมูลทดสอบ (test set) นำข้อมูลทดสอบไปรับการเข้ารหัสด้วยนิวรอลเน็ตเวิร์กที่ได้มาจากชุดข้อมูลฝึก จากนั้นจัดเก็บคำพร้อมด้วยรหัสของคำเหล่านั้นที่ได้ในฐานข้อมูล แล้วนำคำต่างๆทั้งหมดไปค้นคืนทีละคำเพื่อวัดค่าความเที่ยง ค่าเรียกคืน [13] และตัววัด F1 (F1–Measure) [14] (ซึ่งเป็นการวัดค่าเฉลี่ยของค่าความเที่ยงและค่าเรียกคืน) มีสูตรต่างๆ ดังนี้

ค่าความเที่ยง

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

ค่าเรียกคืน

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

ค่าตัววัด F1

$$\frac{2 \times \text{ค่าแม่นยำ} \times \text{ค่าเรียกคืน}}{\text{ค่าแม่นยำ} + \text{ค่าเรียกคืน}}$$

เพื่อให้การทดลองไม่โน้มเอียงกับการแบ่งชุดฝึกกับชุดทดสอบ เราใช้วิธี K–fold Cross Validation [15] ซึ่งแบ่งข้อมูลทั้งหมดออกเป็น K ส่วนเท่าๆกัน แล้วใช้แต่ละส่วนเป็นชุดทดสอบ ส่วนละ 1 ครั้ง ทำการทดสอบทั้งหมด K ครั้ง ในแต่ละครั้งที่เลือกส่วนหนึ่งใดๆ เป็นชุดทดสอบ ส่วนที่เหลือ K–1 ส่วนจะถูกใช้เป็นชุดฝึก จากนั้นนำค่าที่ได้จากการทดลองทั้งหมด K ครั้งมาหาค่าเฉลี่ยเป็นผลการทดลอง ในการทดลองนี้ได้แบ่งข้อมูลออกเป็นส่วน ๆ ให้แต่ละส่วนมีคำประมาณ 400 คู่

ตารางที่ 4 ผลการทดลองการค้นคืนที่ได้เมื่อแปรค่า d กรณีคำไทยทับศัพท์คำอังกฤษ

d	ค่าความเที่ยง	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

ผู้วิจัยได้ทำการทดลองโดยการแปรค่าพารามิเตอร์ d เพื่อหาค่าความแตกต่างที่น้อยที่สุดที่ให้ค่าเฉลี่ยของค่าความเที่ยงและค่าเรียกคืนสูงที่สุด ได้ผลดังแสดงในตารางที่ 4 และ 5 (กรณีที่ค่า d=0 คือการเปรียบเทียบแบบเหมือนกันทุกประการ – exact matching) จากผลการทดลองพบว่าทั้งกรณีการค้นคืนคำไทยทับศัพท์อังกฤษ และคำอังกฤษทับศัพท์จะได้ค่า F1 สูงที่สุด (คือ 81.91% และ 84.40%

ตามลำดับ) เมื่อ d มีค่าเป็น 1 นั่นคือเมื่ออนุญาตให้มีความแตกต่างของรหัสที่นำมาเปรียบเทียบได้ไม่เกิน 1

ตารางที่ 5 ผลการทดลองการค้นคืนที่ได้เมื่อแปรค่า d กรณีคำอังกฤษทับศัพท์คำไทย

d	ค่าความเที่ยง	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

เมื่อนำผลที่ได้มาเปรียบเทียบกับผลการทดลองที่ได้จากการใช้วิธีเข้ารหัสของ [3] และ [4] ด้วยข้อมูลชุดเดียวกัน โดยเลือกค่าความเที่ยงและค่าเรียกคืนที่ให้ค่าตัววัด F1 สูงสุด จะได้ผลดังแสดงในตารางที่ 6 และ 7 จะเห็นได้ว่าผลของการค้นคืนกรณีคำอังกฤษทับศัพท์คำไทยนั้นได้ผลใกล้เคียงกัน (ตารางที่ 7) แต่สำหรับกรณีการค้นคืนกรณีคำไทยทับศัพท์คำอังกฤษ (ตารางที่ 6) นั้นถึงแม้ว่าจะมีค่าตัววัด F1 ใกล้เคียงกัน (80.33% กับ 81.91%) แต่ผลของการค้นคืนมีพฤติกรรมของค่าความเที่ยงและค่าเรียกคืนที่ต่างกัน ซึ่งสามารถแสดงให้เห็นได้โดยใช้ตัววัด E ซึ่งมีสูตรดังนี้ [13]

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

โดยที่ b เป็นค่ากำหนดน้ำหนักความสำคัญระหว่างค่าความเที่ยง และค่าเรียกคืนที่สนใจในการค้นคืน เช่น ถ้า b = 1 หมายถึงกรณีให้ความสำคัญของค่าความเที่ยงกับค่าเรียกคืนที่เท่ากัน (คล้ายตัววัด F1) ถ้า b = 0.1 แสดงว่าให้ความสำคัญกับค่าความเที่ยงมากกว่าค่าเรียกคืน 10 เท่า และในทางกลับกันถ้า b = 10 แสดงว่าให้ความสำคัญกับค่าเรียกคืนมากกว่าค่าความเที่ยง 10 เท่า การค้นคืนที่ให้ค่าของ E น้อยแสดงว่ามีคุณภาพสูง

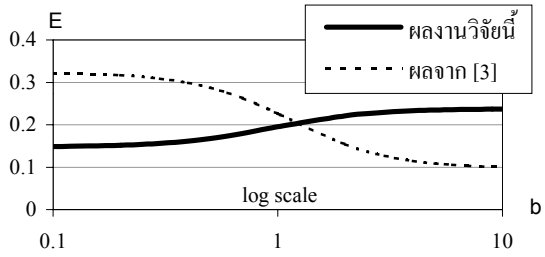
ตารางที่ 6 การเปรียบเทียบผลการค้นคืนของงานวิจัยนี้กับผลที่ได้จาก [3]

วิธีการเข้ารหัส	คำไทยทับศัพท์คำอังกฤษ		
	ค่าความเที่ยง	ค่าเรียกคืน	ตัววัด F1
[3]	72.43	90.25	80.33
งานวิจัยนี้	87.28	77.19	81.91

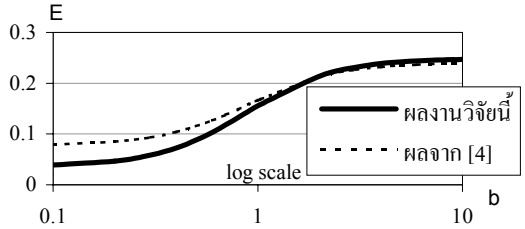
ตารางที่ 7 การเปรียบเทียบผลการค้นคืนของงานวิจัยนี้กับผลที่ได้จาก [4]

วิธีการเข้ารหัส	คำอังกฤษทับศัพท์คำอังไทย		
	ค่าความเที่ยง	ค่าเรียกคืน	ตัววัด F1
[4]	92.27	76.00	83.33
งานวิจัยนี้	96.34	75.15	84.41

รูปที่ 2 กราฟแสดงการเปรียบเทียบค่า E ของการค้นคืนคำไทยทับศัพท์คำอังกฤษจากงานวิจัยนี้กับผลจาก [3]



รูปที่ 3 กราฟแสดงการเปรียบเทียบค่า E ของการค้นคืนคำอังกฤษทับศัพท์คำไทยจากงานวิจัยนี้กับผลจาก [4]



เมื่อนำผลที่ได้จากตารางที่ 6 และ 7 มาวาดกราฟเปรียบเทียบพฤติกรรมการค้นคืนเมื่อแปรค่า b จาก 0.1 ถึง 10 จะได้ผลดังแสดงในรูปที่ 2 และ 3 ในรูปที่ 3 ซึ่งเป็นของกรณีการค้นคืนคำอังกฤษทับศัพท์คำไทยนั้น แสดงให้เห็นว่า ให้ผลของการค้นคืนในลักษณะคล้ายกัน แต่รูป

ที่ 2 ซึ่งเป็นของกรณีการค้นคืนคำไทยทับศัพท์อังกฤษนั้นมีพฤติกรรมต่างกัน กล่าวคือการเปลี่ยนค่า b ในระบบการค้นคืนที่ได้จากงานวิจัยนี้ จะมีผลต่อคุณภาพการค้นคืนที่น้อยกว่า (เส้นทึบมีความลาดชันน้อยกว่า) อีกทั้งได้ผลการค้นคืนที่ดีกว่า (E มีค่าต่ำกว่า) เมื่อให้ความสนใจกับค่าความเที่ยงมากกว่าค่าเรียกคืน (กรณี $b < 1$) แต่ผลการค้นคืนนี้ก็ไม่ลดคุณภาพไปมากนักเมื่อให้ความสำคัญที่มากขึ้นกับค่าเรียกคืน

5. สรุป

บทความนี้นำเสนอการใช้นิวรอลเน็ตเวิร์กในการเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษา นิวรอลเน็ตเวิร์กที่ใช้เป็นแบบแบ็กพรอพาเกชัน ซึ่งรับข้อมูลขาเข้าเป็นตัวอักขระที่สนใจให้รหัสในคำพร้อมทั้งตัวอักขระข้างเคียงหน้าหลังข้างละสี่ตัวของคำ และให้ข้อมูลขาออกเป็นรหัสเสียงของตัวอักขระขาเข้าสนใจ ผลที่ได้คือนิวรอลเน็ตเวิร์กที่ใช้เข้ารหัสที่ได้กับคำอังกฤษที่ทับศัพท์คำไทย และคำไทยที่ทับศัพท์คำอังกฤษ ขั้นตอนการค้นคืนคำข้ามภาษาอาศัยการเปรียบเทียบรหัสของคำแบบประมาณ โดยอนุญาตให้มีความแตกต่างของรหัสที่นำมาเปรียบเทียบได้ไม่เกิน 1 จากผลการทดลองกับชุดข้อมูลเกือบสี่พันคู่ของคำด้วยวิธี K–fold cross validation พบว่าการค้นคืนค่าตัววัด F1 (ซึ่งเป็นผลการเฉลี่ยของค่าความเที่ยง และค่าเรียกคืน) ของการค้นคืนกรณีคำอังกฤษทับศัพท์คำไทยและกรณีคำไทยทับศัพท์อังกฤษที่สูงเกิน 80%

6. เอกสารอ้างอิง

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

[2] Ongroongruang, S., Prongsirivattana, R. and Jantarasukree, V. English to Thai Word Retrieval Using Sound Index, Proc 2nd SNLP 95, Bangkok Thailand, pp.47–413, 1995.

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

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

[5] Rich, E. and Knight, K. Artificial Intelligence, Singapore, Prentice–Hill, 1991.

[6] บุญภานนท์, อุไรรัตน์. การถอดอักษรภาษาอังกฤษเป็นไทยโดยใช้หลักวิชาภาษาศาสตร์, วิทยานิพนธ์, คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

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

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

[9] ราชบัณฑิตยสถาน. หลักเกณฑ์การทับศัพท์, 2535.

[10] ราชบัณฑิตยสถาน. ศัพท์วิทยาศาสตร์, 2536.

[11] ราชบัณฑิตยสถาน. ศัพท์คณิตศาสตร์, 2540.

[12] ราชบัณฑิตยสถาน. หนังสือเรียนวิชาเคมี เล่ม 1 หลักสูตรมัธยมปลาย 2542, 2530.

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

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

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