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 ส 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 เ็	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-ีย	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"

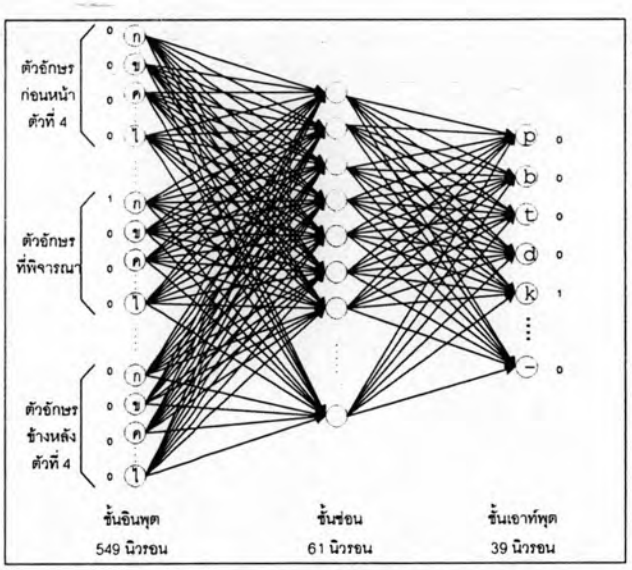
นิวรอลเน็ตเวิร์กที่ใช้เป็นแบบแบ็กพรอพาเกชัน (Backpropagation Neural Network) ซึ่งมีจำนวนชั้นของการทำงานจำนวน 3 ชั้น ดังแสดงในรูปที่ 3.3 โดยข้อมูลขาเข้าเป็นตัวอักษรภาษาไทย ประกอบด้วยตัวอักษรที่ต้องการทราบรหัสเสียง และตัวอักษรที่อยู่ข้างเคียงกับมัน ข้างละ 4 ตัว รวมเป็น 9 ตัว (เนื่องจากข้อมูลเข้าอยู่ในรูปอักษรไทย 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

ชนิดของข้อมูลที่ใช้ในการทดลอง	จำนวนเสียงในข้อมูลเข้า	จำนวนตัวอักษรที่ใช้ในการพิจารณาในแต่ละรอบ	จำนวนนิวรอลในชั้นเข้า	จำนวนเสียงในข้อมูลออก
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 ส ร 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-ียว	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	aI	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

I23504560