Transliterated Word Encoding and Retrieval Algorithms for Thai-English Cross-Language Retrieval

Prayut Suwanvisat
Department of Computer Engineering
Chulalongkorn University
Bangkok, Thailand
g40psw@cp.eng.chula.ac.th

Somchai Prasitjutrakul
Department of Computer Engineering
Chulalongkorn University
Bangkok, Thailand
somchaip@chula.ac.th

Abstract: This paper presents transliterated word encoding and retrieval algorithms for Thai-English cross-language retrieval. The encoding is used for Thai-to-English transliterated words and their corresponding Thai words. The encoding algorithm transforms Thai words into a canonical form using romanization rules and transforms English characters to Thai using transliteration rules. The retrieval algorithm searches codes of the query words approximately with codes in the index using a dynamic programming technique. Experimental results showed that the system can achieve both precision and recall as high as 70% simultaneously.
Key words: Transliterated word encoding, Cross-language retreival

ขั้นตอนวิธีการเข้ารหัสและการค้นคืนคำทับศัพท์ข้ามภาษาไทย-อังกฤษ

ประยุทธ สุวรรณวิสารท
ภาควิชาวิศวกรรมคอมพิวเตอร์
จุฬาลงกรณ์มหาวิทยาลัย
prayut@cp.eng.chula.ac.th

สมชาย ประสิทธิ์จูตระกูล
ภาควิชาวิศวกรรมคอมพิวเตอร์
จุฬาลงกรณ์มหาวิทยาลัย
somchaip@chula.ac.th

บทคัดย่อ: บทความนำเสนอขั้นตอนวิธีการเข้ารหัสคำทับศัพท์และการค้นคืนสำหรับระบบการค้นคืนข้ามภาษาไทย-อังกฤษ การเข้ารหัสคำนี้ใช้กับกรณีของคำอังกฤษที่ทับศัพท์คำไทย ซึ่งประกอบด้วยขั้นตอนการแปลงรูปคำไทยโดยอาศัยหลักการถ่ายเสียง และการแปลงตัวอักษรอังกฤษเป็นตัวอักษรไทยโดยอาศัยหลักการถอดอักษร ส่วนการค้นคืนนั้นอาศัยการเทียบรหัสคำของข้อคำถาม กับรหัสคำในดัชนีแบบประมาณ โดยใช้เทคนิคกำหนดการพลวัต จากผลการทดลองพบว่าได้ค่าค่าเรียกคืนสูง และค่าแม่นยำประมาณ 70 เปอร์เซ็นต์
คำสำคัญ: การเข้ารหัสคำทับศัพท์, การค้นคืนข้ามภาษา

1. บทนำ

ระบบค้นคืนสารสนเทศเป็นเครื่องมือที่สำคัญอย่างยิ่งในการบริหารสารสนเทศที่มีอยู่จำนวนมากในรูปของสื่ออิเล็กทรอนิกส์ โดยเฉพาะสารสนเทศที่จัดเก็บในเวิลด์ไวด์เว็บและซีดีรอม เนื่องจากสื่อดังกล่าวสามารถจัดเก็บสารสนเทศได้เป็นจำนวนมาก และสามารถเข้าถึงได้ง่ายทั้งจากระยะใกล้และไกล โดยปกติแล้วการวัดประสิทธิผลของระบบค้นคืนสารสนเทศใด ๆ มักจะวัดจากค่าแม่นยำ (Precision) และ ค่าเรียกคืน (Recall) [1] ซึ่ง ค่าแม่นยำ หมายถึงการวัดความสามารถของระบบในการที่จะขจัดเอกสารที่ไม่เกี่ยวข้องออกไป ค่าที่ได้จะเป็นอัตราส่วนระหว่างจำนวนของเอกสารที่เกี่ยวข้องที่คืนกลับมา กับจำนวนเอกสารทั้งหมดที่กลับคืนมา ส่วนค่าเรียกคืนหมายถึงการวัดความสามารถของระบบในการคืนเอกสารที่เกี่ยวข้องกลับมา ค่าที่ได้จะเป็นอัตราส่วนระหว่างจำนวนของเอกสารที่เกี่ยวข้องที่คืนกลับมา กับจำนวนทั้งหมดของเอกสารที่เกี่ยวข้อง

ในกรณีที่ผู้ใช้ป้อนคำหลักด้วยภาษาใดภาษาหนึ่ง ในขณะที่คำหลักในเอกสารจัดเก็บด้วยภาษาอื่น ตัวอย่างเช่น ผู้ใช้ต้องการสืบค้นคำว่า *สมชาย* แต่ระบบไม่ได้คืนเอกสารที่มีคำว่า *SOMCHAI* (คำทับศัพท์ที่ตรงกัน) ทำให้ค่าเรียกคืนของระบบค้นคืนสารสนเทศน้อยกว่าที่ควรจะเป็น ถ้าระบบดังกล่าวไม่สนับสนุนการทำงานแบบข้ามภาษา (cross-language retrieval) [2]

การค้นคืนสารสนเทศข้ามภาษา หมายถึงการค้นคืนสารสนเทศ โดยภาษาที่ใช้ในข้อคำถามแตกต่างจากภาษาที่ใช้ในการจัดเก็บเอกสาร การใช้พจนานุกรมสองภาษา (Bilingual Dictionary) ในลักษณะของอรรถาภิธาน (Thesaurus) กับระบบค้นคืนสารสนเทศไม่สามารถแก้ไขปัญหาดังกล่าวได้มากนัก เนื่องจากคำทับศัพท์ส่วนมากมักเป็นคำเฉพาะที่ไม่ปรากฏในพจนานุกรม [3]

มีงานวิจัยต่าง ๆ ที่เกี่ยวข้องกับปัญหาการค้นคืนข้ามภาษาไทย-อังกฤษ คือ [4] ได้นำเสนอขั้นตอนวิธีสำหรับเข้ารหัสคำภาษาอังกฤษ ซึ่งรหัสคำที่ได้จะเป็นกลุ่มของเสียงอ่านที่เป็นไปได้ในภาษาไทย การเข้ารหัสจะอาศัยตารางการกำหนดรหัสและกฎ แต่ในงานวิจัยนั้นไม่ได้แสดงรายละเอียดของตารางการกำหนดรหัส และไม่ได้รายงานผลการทดลอง [5] ได้นำเสนอขั้นตอนวิธีการเข้ารหัสคำสำหรับคำไทยที่ทับศัพท์คำอังกฤษ โดยใช้เทคนิคการเข้ารหัสคำแบบซาวด์เด็กซ์

งานวิจัยนี้จะนำเสนอขั้นตอนวิธีการการค้นคืนคำทับศัพท์ข้ามภาษาไทย-อังกฤษ ซึ่งคำทับศัพท์จะเป็นคำภาษาอังกฤษทับศัพท์คำภาษาไทย เช่น *BOONSERMSAP* ทับศัพท์ *บุญเสริมทรัพย์* เป็นต้น การเข้ารหัสประกอบด้วยขั้นตอนการแปลงรูปคำไทยซึ่งอาศัยหลักการถ่ายเสียง และใช้หลักการถอดอักษรเพื่อแปลงตัวอักษรอังกฤษเป็นตัวอักษรไทย จากนั้นนำรหัสคำที่ได้ไปเทียบอักษรแบบประมาณโดยอาศัยเทคนิคกำหนดการพลวัตกับรหัสคำในดัชนี โดยมีการกำหนดเกณฑ์การยอมรับในการเทียบสายอักษร บทความนี้จะนำเสนอขั้นตอนวิธีการเข้ารหัสคำอย่างละเอียดในหัวข้อที่สอง นำเสนอขั้นตอนการค้นคืนในหัวข้อที่สาม หัวข้อที่สี่นำเสนอวิธีการทดลองประสิทธิผลและผลการทดลอง จากนั้นสรุปเนื้อหาของบทความในหัวข้อที่ห้า

2. ขั้นตอนวิธีการเข้ารหัสคำ

การเข้ารหัสคำมีจุดประสงค์เพื่อแปลงคำไทย และคำอังกฤษที่ทับศัพท์คำไทย ให้อยู่ในรูปแบบรหัสคำรูปแบบเดียวกัน รหัสคำของคำเฉพาะทั้งคำไทย และคำอังกฤษที่ทับศัพท์คำไทยต่างๆในเอกสารจะถูกสร้างขึ้นในขั้นตอนการสร้างดัชนีในระบบการจัดเก็บสารสนเทศ เมื่อผู้ใช้ป้อนข้อคำถามที่เป็นคำเฉพาะคำไทย หรือคำอังกฤษทับศัพท์คำไทย ระบบค้นคืนก็จะสร้างรหัสคำของคำต่างๆในข้อคำถาม เพื่อใช้ค้นหากับรหัสคำที่จัดเก็บไว้ในดัชนี

ขั้นตอนวิธีการเข้ารหัสคำแบ่งออกเป็นสองกระบวนการ คือการแปลงรูปคำไทยโดยอาศัยหลักการถ่ายเสียง และการแปลงตัวอักษรอังกฤษเป็นตัวอักษรไทยโดยอาศัยหลักการถอดอักษร

2.1 การแปลงรูปคำไทย

ขั้นตอนการแปลงรูปคำไทยมีจุดประสงค์หลักเพื่อแปลงคำไทยที่อ่านออกเสียงคล้ายกัน แต่เขียนได้หลายรูปแบบ ให้อยู่ในรูปแบบเดียวกัน เพื่อให้ขั้นตอนการเทียบรหัสกระทำได้ง่ายขึ้น การแปลงรูปประกอบด้วยประกอบด้วยการตัดวรรณยุกต์และไม้ไต่คู้ การเปลี่ยนรูปการสะกด และการแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล

2.1.1 การตัดวรรณยุกต์และไม้ไต่คู้

เนื่องจากในการถอดอักษรไทยเป็นอังกษรอังกฤษนั้น จะไม่พิจารณาวรรณยุกต์และไม้ไต่คู้ เช่น ช้าง ถอดอักษรเป็น

CHANG ดังนั้นวรรณยุกต์ และไม้ไต่คู้ จะถูกตัดทิ้งจากคำไทย

2.1.2 การเปลี่ยนรูปการสะกด

การเปลี่ยนรูปการสะกดอาศัยหลักทางภาษาศาสตร์ [6] และข้อมูลทางสถิติ [7] เพื่อเปลี่ยนรูปแบบการสะกดของคำไทยที่เขียนได้หลายรูปแบบให้เหมือนกัน แบ่งเป็นกรณีต่าง ๆ ดังนี้

- รร ทำการเปลี่ยน รร เป็น –ัน ในกรณีที่ไม่มีตัวสะกดตามหลัง เช่น *จรรยา บรรจบ ครรภ์* เป็นต้น และเป็น –ั ในกรณีที่มีตัวสะกดตามหลัง เช่น *ธรรม พรรณ กรรม* เป็นต้น
- สระ ใ- ไ- และไ-ย ทำการเปลี่ยนสระดังกล่าวทั้งหมดเป็น -ัย
- สระ –ำ และ –ัม ทำการเปลี่ยนสระ –ำ เป็น –ัม
- การันต์และอักษรควบการันต์ ทำการตัดอักษรและอักษรควบที่มีตัวการันต์กำกับออก เนื่องจากการถอดอักษรโดยปกติจะไม่ถอดอักษรที่มีการันต์กำกับ เช่น *สิทธิ์* ถอดเป็น *SITH พันธุ์* ถอดเป็น *PHAN* เป็นต้น

2.1.3 การแทนที่สระประสมด้วยสัญลักษณ์เสียงสากล

การแทนที่สระประสมด้วยสัญลักษณ์เสียงสากลมีจุดประสงค์เพื่อแปลงทำให้คำไทยที่มีการใช้สระประสมและสระเดี่ยวที่ใช้อักขระตั้งแต่สองตัวขึ้นไป ให้อยู่รูปแบบที่ง่ายต่อการประมวลผล และลดจำนวนอักขระที่ต้องเปรียบเทียบ ซึ่งขั้นตอนนี้จะใช้สัญลักษณ์เสียงสากล [8] แทนเสียงสระดังกล่าวโดยจะวางสัญลักษณ์เสียงสากลไว้หลังพยัญชนะต้น เพื่อให้มีรูปแบบเหมือนกับผลลัพธ์ที่ได้จากการถอดอักษรคำอังกฤษทับศัพท์คำไทย การใช้สัญลักษณ์เสียงสากลแทนที่สระมีดังนี้

- ใช้สัญลักษณ์เสียง e แทนสระ เ-ะ เช่น *เละ* เปลี่ยนเป็น *ละe*
- ใช้สัญลักษณ์เสียง x แทนสระ แ-ะ แ--ะ เช่น *แกะ แหละ* เปลี่ยนเป็น *กx หลx* ตามลำดับ
- ใช้สัญลักษณ์เสียง q แทนสระ เ-ิ เ--ิ เช่น *เกิด เผชิญ* เปลี่ยนเป็น *กqด ผชqญ* ตามลำดับ
- ใช้สัญลักษณ์เสียง I แทนสระ เ-ีย เ--ียะ เ-ีย และ เ--ีย เช่น *เสียง เกเวียน* เปลี่ยนเป็น *สIง กวIน* ตามลำดับ
- ใช้สัญลักษณ์เสียง U แทนสระ เ-ือะ เ-ือ เ--ือ -ัวะ และ -ัว เช่น *เรือง เกลือ ผัวะ* และ *วัว* เปลี่ยนเป็น *รUง กลU ผU* และ *วU* ตามลำดับ
- ใช้สัญลักษณ์เสียง @ แทนสระ เ-า เ--า เ-าะ และ เ--าะ เช่น *เงา เพรา เงาะ* และ *เฉพาะ* เปลี่ยนเป็น *ง@ พร@ ง@* และ *ฉพ@* ตามลำดับ

2.2 การถอดอักษรคำอังกฤษทับศัพท์คำไทย

ในกรณีที่ข้อคำถามเป็นคำอังกฤษทับศัพท์คำไทย จะทำการถอดอักษรอังกฤษเป็นไทย การถอดอักษรที่นำเสนอในที่นี้เป็นการเปลี่ยนพยัญชนะอังกฤษเป็นพยัญชนะไทย ส่วนสระอังกฤษจะใช้หลักเกณฑ์การแทนที่สระประสมด้วยสัญลักษณ์เสียงสากลคือจะถอดสระอังกฤษเป็นสระไทย แต่ถ้าสระไทยนั้นเป็นสระที่ใช้อักขระตั้งแต่สองตัวขึ้นไปจะใช้สัญลักษณ์เสียงสากล แทนเสียงสระดังกล่าว

หลักเกณฑ์ในการถอดอักษรในส่วนพยัญชนะ (ดูดังตารางที่ 1) จะใช้หลักการเทียบตัวอักษรโรมัน-ไทยของ ISO [9] โดยได้ดัดแปลงบางส่วนเพื่อให้สมบูรณ์ยิ่งขึ้น เช่น ถอด DH เป็น ท และ ถอด BH เป็น พ เนื่องจากเป็นการเขียนแบบบาลีสันสกฤต ซึ่งยังมีใช้กันอยู่มากในปัจจุบัน [10]

อังกฤษ	ไทย	
B	บ	
BH	พ	*
C	ซ	*
CH	ช (ฉ ฌ)	
CK	ก	*
D	ด (ฎ)	
DH	ท	*
F	ฟ (ฝ)	
G	ก	*
H	ห (ฮ)	
J	จ	*
K	ก	
KH	ข (ฃ ค ฅ ฆ)	
L	ล (ฦ ฬ)	
M	ม	

อังกฤษ	ไทย	
N	น (ณ)	
NG	ง	
P	ป	
PH	พ (ผ ภ)	
Q	ค	*
R	ร (ฤ)	
S	ส (ซ ศ ษ)	
T	ต (ฏ)	
TH	ท (ฐ ฑ ฒ ถ ธ)	
V	ว	
W	ว	
X	ก	*
Y	ย (ญ)	
Z	ซ	*

* ส่วนที่ปรับเปลี่ยนและเพิ่มเติมจากแบบของ ISO

ตารางที่ 1 การถอดอักษรพยัญชนะอังกฤษเป็นไทย

ส่วนหลักเกณฑ์ในการถอดอักษรในส่วนสระอังกฤษเป็นสระไทยนั้น ผู้วิจัยพบว่ามีปัญหาอย่างมากในการถอด

อักษร คือ หนึ่งหน่วยอักษรโรมันสามารถถอดได้เป็นหลายหน่วยอักษรไทยเช่น A ถอดอักษรเป็น เ-แ- -ะ –า และ O ถอดอักษรเป็น –อ โ- เป็นต้น และหลายหน่วยอักษรโรมันสามารถถอดเป็นหนึ่งหน่วยอักษรไทยได้ เช่น U หรือ OO ถอดเป็น –ุ เป็นต้น และจากการศึกษาของผู้วิจัยพบว่ามีความหลากหลายในการใช้อักษรโรมันแทนอักษรไทย เช่น คำว่า *พร* มีการเขียนเป็น *Phorn Phon Porn Pon* เป็นต้น (การใช้ *orn* เป็นความนิยมใช้ซึ่งไม่ถูกต้องตามหลักภาษาที่ราชบัณฑิตสถานกำหนด) ดังนั้นในงานวิจัยนี้จึงได้พยายามยึดหลักการถอดอักษรของราชบัณฑิตยสถาน [11] และ การใช้อักษรโรมันแทนอักษรไทยของจันทร์เพ็ญ โวหารสุนทร [12] (ได้จากการสำรวจความนิยมในการใช้อักษรโรมันแทนอักษรไทย) เป็นต้นแบบและเพิ่มเติมบางส่วนจากการศึกษาของผู้วิจัยที่มีผู้นิยมใช้เข้าไปดังตารางที่ 2

อังกฤษ	ถอดอักษรเป็น	หมายเหตุ
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	-ื	

* ส่วนที่เพิ่มเติมจากแบบของราชบัณฑิตยสถาน

ตารางที่ 2 การถอดอักษรสระอังกฤษเป็นไทย

<u>หมายเหตุ</u> ถ้าตัวอักษรแรกของคำเป็นสระได้แก่ A E I O และ U ให้ถอด A เป็น อ ถอด E เป็น เอ (สระ เ- กับ อ) ถอด I เป็น อิ (อ กับสระ -ิ) ถอด O เป็น โอ (สระ โ- กับ อ) และถอด U เป็น อุ (อ กับสระ -ุ) และถ้าตัวอักษรถัดไปเป็นสระอีกให้นำอักษรตัวแรกไปรวมด้วยในการถอดอักษรโดยเทียบตามตารางที่ 2

3. ขั้นตอนการค้นคืน

รหัสคำที่ได้ของคู่คำไทยและคำอังกฤษทับศัพท์คำไทยนั้นอาจไม่ตรงกันทุกตัวอักษร แต่จะมีลักษณะคล้ายกัน ทั้งนี้เนื่องจากหลักการทับศัพท์ที่ใช้กันในปัจจุบันมีหลายรูปแบบ การเปรียบเทียบรหัสคำแบบเปรียบเทียบตัวต่อตัวในรหัสให้ตรงกัน จะได้ความแม่นยำของการค้นคืนที่สูง แต่ค่าเรียกคืนจะต่ำมาก เพื่อสร้างความสมดุลย์ของระบบการค้นคืน เพื่อให้ได้ค่าแม่นยำและค่าเรียกคืนที่ดีทั้งคู่ การค้นคืนรหัสคำจะอาศัยสองเทคนิคคือ การแยกเทียบส่วนพยัญชนะและส่วนสระ และการเทียบรหัสคำแบบประมาณ

การเทียบรหัสคำแบบประมาณนั้นอาศัยการคำนวณค่าความแตกต่าง (distance) ของรหัสคำด้วยเทคนิคระยะแก้ไขเสียงอ่านสั้นที่สุด (Minimum Phonetic Edit Distance) [13] โดยแยกหาค่าความแตกต่างของส่วนพยัญชนะ และส่วนสระของรหัสคำ จากนั้นนำค่าความแตกต่างของคู่รหัสคำที่ได้มาทดสอบกับเงื่อนไขในการเปรียบเทียบ ถ้าผ่านการทดสอบจะสรุปได้ว่ารหัสคำทั้งสองรหัสนี้เป็นรหัสที่มาจากคำหลักที่ตรงกันในอีกภาษา

3.1 การคำนวณหาค่าความแตกต่างของรหัสคำ

การคำนวณหาค่าความแตกต่างของรหัสคำจะได้มาจากการคำนวณหาต้นทุนน้อยที่สุดในการแก้ไขอักขระ (ประกอบด้วยการลบการเพิ่มและการแทนที่อักขระ) ให้รหัสคำทั้งสองเหมือนกัน ซึ่งจะพิจารณาความแตกต่างกันทางเสียงของอักขระแทนรูปของอักขระ โดยใช้กลุ่มอักขระของซาวด์เด็กซ์ช่วยในการกำหนดกลุ่มเสียงที่คล้ายกัน โดยผู้วิจัยได้ทำการปรับเปลี่ยนการกำหนดต้นทุนในการแก้ไขอักขระของขั้นตอนวิธีระยะแก้ไขเสียงอ่านสั้นที่สุด [13] เพื่อความเหมาะสมเมื่อใช้กับคำในภาษาไทยโดยมีแนวคิดดังนี้

- กำหนดต้นทุนต่อการแก้ไขอักขระให้มี 4 ระดับ คือ C_1 C_2 C_3 และ C_4 โดยกำหนดให้ C_1=0 C_2=1 C_3=4 และ C_4=7
- กำหนดต้นทุนในการแก้ไขแต่ละอักขระให้มีค่าไม่เท่ากันได้
- อนุญาตให้มีการแทนที่อักขระแบบหนึ่งอักขระต่อสองอักขระ เช่น การแทนที่ ทร ด้วย ซ เป็นต้น

การคำนวณค่าความแตกต่างอาศัยเทคนิคกำหนดการพลวัต (dynamic programming) จากสมการเวียนเกิด *Edit*(P_j, W_k) ดังนี้

$Edit(P_0, W_0) = 0$

$Edit(P_j, W_0) = Edit(P_{j-1}, W_0) + D(p_{j-1}, p_j)$

$Edit(P_0, W_k) = Edit(P_0, W_{k-1}) + D(w_{k-1}, w_k)$

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

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

$p_{j-1} p_j$ เป็นสองอักขระใด ๆ ที่มีตำแหน่งติดกัน

$D(p_j, w_k)$ มีค่าเท่ากับ min(*DELADD*[p_j] , *REPLACE* [p_{j-1}pj , w_k])

$R(p_{j-1}p_j, w_k)$ มีค่าเท่ากับ min(*REPLACE*[p_j, w_k], *REPLACE* [$p_{j-1} p_j$, w_k])

REPLACE [$p_{j-1} p_j$, w_k] หมายถึงตารางค้นหาต้นทุนในการแทนที่อักขระ $p_{j-1} p_j$ ด้วยอักขระ w_k

DELADD [pj] หมายถึงตารางค้นหาต้นทุนในการลบเพิ่มอักขระ p_j

การกำหนดต้นทุนในการแทนที่อักขระ (ตาราง *REPLACE* [p_j, w_k] สำหรับการแทนที่อักขระ p_j ด้วยอักขระ w_k) มีหลักการดังนี้

- พยัญชนะไทยมี 44 รูป แยกความแตกต่างทางเสียงได้ 21 เสียง [8] กำหนดต้นทุนเท่ากับ C_1 เช่น การแทนที่ ญ ด้วย ย มีต้นทุนเท่ากับ C_1 เป็นต้น
- อักขระที่ไม่สามารถแยกความแตกต่างในการถอดอักษร กำหนดต้นทุนเท่ากับ C_2 ได้แก่ อักษร K ถอดเป็น ก หรือ ค อักษร T ถอดเป็น ต หรือ ท อักษร P ถอดเป็น ป หรือ พ อักษร CH ถอดเป็น จ หรือ ช อักษร E ถอดเป็น เ- หรือ ◌ี และ อักษร U ถอดเป็น ◌ุ หรือ ◌ู
- การถอดอักษรสำหรับสระเสียงสั้นกับสระเสียงยาว ที่ไม่สามารถแยกความแตกต่างทางเสียงในภาษาอังกฤษ กำหนดต้นทุนในการแทนที่อักขระเท่ากับ C_1 เช่น อักษร A ถอดอักษรเป็น –ะ หรือ –า อักษร I ถอดอักษรเป็น ◌ิ หรือ ◌ี เป็นต้น
- มาตราของอักขระไทย จะเลือกบางอักขระมาเป็นตัวแทนมาตรา ดังตารางที่ 3 โดยจะกำหนดต้นทุนในการแทนที่อักขระตัวแทนมาตรากับอักขระที่อยู่ในมาตราเดียวกันเท่ากับ C_3 ซึ่งวิธีนี้จะทำให้ความแตกต่างของเสียงมีได้มากกว่าวิธีจัดกลุ่มแบบซาวด์เด็กซ์ภาษาไทย ที่รวมอักขระต่าง ๆ ที่อยู่มาตราเดียวกันเข้าเป็นกลุ่มเสียงเดียวกัน [14] เช่น คำว่า *สาร* กับ *ชาญ* วิธีที่นำเสนอสามารถบอกได้ว่า ส-ช และ ร-ญ ต่างกันสิ้นเชิง เนื่องจากทั้ง ส และ ช มีความสัมพันธ์กับ ด หรือ ต (มาตรา กด) แต่ ส และ ช ไม่มีความสัมพันธ์กันเอง แต่ถ้าเป็นวิธีซาวด์เด็กซ์จะบอกว่าอักษร ส และ ช มีความสัมพันธ์กันเนื่องจากอยู่ในมาตรากด และอักษร ร และ ญ มีความสัมพันธ์กันเนื่องจากอยู่ในอยู่ในมาตรากน

มาตรา	ตัวอักษร	ตัวแทน
กก	ก ข ค ฆ	ก
กด	ด ต ถ ท ธ ฎ ฏ ฑ ฐ ฒ จ ช ศ ษ ส ซ	ด ต
กง	ง	ง
กน	น ณ ญ ล ฬ ร	น
กบ	บ ป พ ภ ฟ	บ
กม	ม	ม

ตารางที่ 3 อักขระแทนมาตราต่าง ๆ

- กำหนดต้นทุนในการแทนที่อักขระแบบหนึ่งอักขระต่อสองอักขระเท่ากัน เนื่องจากมีการอ่านออกเสียงเหมือนกัน โดยแบ่งเป็นกรณีต่าง ๆ ดังนี้
  1. อักษรควบไม่แท้ มี 5 ตัว [6] ดังนี้ ทร-ซ, จร-จ, สร-ส, ศร-ศ และ ซร-ซ
  2. อักษรนำเสียงสนิท มี 9 ตัว [6] ดังนี้ อย, หย, หง, หร, หญ, หล, หน, หว และ หม
  3. อักษรควบที่เป็นตัวสะกด [6] เช่น กร-*จักร*, คร-*สมัคร*, ตร-*จิตร*, ทร-*สมุทร*, ปร-*กอปร*, รค-*ชลมารค*, รถ-*สามารถ* และ หม-*พรหม*
- นอกจากนี้จะกำหนดต้นทุนในการแทนที่เท่ากับ C_4 (ซึ่งมีค่ามากที่สุด)

การกำหนดต้นทุนในการลบเพิ่มอักขระ (ตาราง *DELADD* [*a*]) สามารถแบ่งเป็นกรณีต่าง ๆ ดังนี้

- สระที่มีการแปลงรูป ได้แก่ -ะ โ- และ –อ ซึ่งจะไม่ปรากฏในหน่วยอักษรของคำไทย แต่จะปรากฏในคำอังกฤษทับศัพท์ภาษาไทย ทำให้หน่วยอักษรที่ได้จากถอดอักษรไม่ตรงกันกับคำไทย เช่น *Thavee-ทวี* *พร-Phon* และ *นก-Nok* เป็นต้น ดังนั้นจะกำหนดต้นทุนในการเพิ่มหรือลบอักขระกลุ่มนี้เท่ากับ C_2

- พยัญชนะที่เป็นส่วนประกอบของสระประสม ได้แก่ อ ย และ ว เช่น เ-ือ เ-ีย และ -ว- เป็นต้น ซึ่งมักจะเกิดความผิดพลาดได้ง่ายในการถ่ายอักษร เช่น คำว่า *Chuan (ชวน)* การถอดอักษรจะได้ *เชือน* ซึ่งไม่ได้อักษร ว ดังนั้นจะกำหนดต้นทุนในการเพิ่มหรือลบอักขระกลุ่มนี้เท่ากับ C_2
- การทับศัพท์ที่ไม่ตรงตามประกาศของราชบัณฑิตยสถาน เช่น คำว่า *ธรรม-Thama จิตร-Chitara* และ *รัตน-Ratana* เป็นต้น จากการทับศัพท์แบบนี้ทำให้เวลาถอดอักษรแล้ว จะได้สระ –ะ หรือ –า เกิน ดังนั้นจะกำหนดต้นทุนในการเพิ่มหรือลบอักขระกลุ่มนี้เท่ากับ C_2
- นอกจากนี้จะกำหนดต้นทุนในการลบเพิ่มเท่ากับ C_4 (ซึ่งมีมากที่สุด)

3.2 เกณฑ์การเทียบรหัสคำแบบประมาณ

รหัสคำสอง P_m และ W_n ใด ๆ ที่มีค่าความแตกต่างผ่านเกณฑ์ที่แสดงข้างล่างนี้ ทั้งส่วนพยัญชนะ และส่วนสระของรหัส จะถือว่าเป็นรหัสคำของคำที่ตรงกันระหว่างภาษาไทยกับภาษาอังกฤษทับศัพท์ภาษาไทย

$$Edit(P_m, W_n) \leq \alpha \times \mathrm{Max}(m, n) \times C_4$$

$\mathrm{Max}(m, n) \times C_4$ หมายถึงต้นทุนมากสุดที่ใช้ในการแก้ไขอักขระ คือแก้ไขทุกอักขระของ W_n ให้เป็น P_m ในขณะที่ α นี้เป็นพารามิเตอร์ของระบบที่ผู้ใช้สามารถกำหนดได้ ซึ่งจะส่งผลต่อค่าแม่นยำและค่าเรียกคืนของระบบ α มีค่าระหว่าง 0 ถึง 1 ซึ่งจะเป็นตัวกำหนดเกณฑ์การยอมรับความเหมือนกันของรหัสคำแบบประมาณ โดยที่ 0 หมายถึงรหัสคำทั้งสองต้องเหมือนกันทุกประการจึงจะยอมรับ ในขณะที่ 1 หมายถึงการยอมรับทุก ๆ คู่รหัสคำไม่ว่าจะแตกต่างเท่าไรก็ตาม

ตัวอย่าง ต้องการทดสอบว่า "*ชูเลิศติยวงศ์*" และ "*CHULERTTIYAWONG*" เป็นคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษหรือไม่ โดยจะเริ่มต้นจากการเข้ารหัสคำ การคำนวณหาค่าความแตกต่าง และนำค่าความแตกต่างที่ได้ไปทดสอบกับเงื่อนไขเพื่อสรุปว่าเป็นคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษหรือไม่ กำหนดให้ $\alpha = 0.15$

การเข้ารหัสคำ

- *CHULERTTIYAWONG* → "ชุลqตติยะวอง"
- *ชูเลิศติยวงศ์* → ชูเลิศติยวง→ "ชูลqศติยวง"

การค้นคืน

ส่วนพยัญชนะ

Edit("ชลตตยวอง", "ชลศตยวง") = 5

$0.15 \times \mathrm{Max}(8, 7) \times 7 = 8.4$

ผ่านเกณฑ์ : $5 \leq 8.4$

ส่วนสระ

Edit("ุqิะ" , "ูqิ") = 1

$0.15 \times \mathrm{Max}(4, 3) \times 7 = 4.2$

ผ่านเกณฑ์ : $1 \leq 4.2$

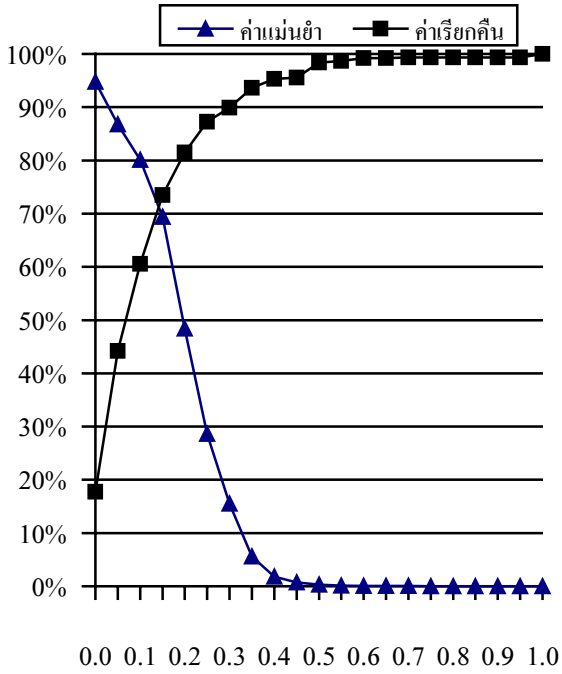
จากตัวอย่างพบว่าผ่านเกณฑ์ทั้งส่วนพยัญชนะ และส่วนสระ สรุปได้ว่าทั้งสองคำเป็นคำทับศัพท์ที่ตรงกันในภาษาไทย-อังกฤษ

4. ผลการทดลอง

ผู้วิจัยได้ทำการทดลองขั้นตอนวิธีที่ได้นำเสนอ โดยใช้ชื่อและชื่อสกุลทั้งภาษาไทยและภาษาอังกฤษที่ตรงกันของนิสิตปีการศึกษา 2541 ระดับปริญญามหาบัณฑิต จุฬาลงกรณ์มหาวิทยาลัย จำนวน 5,000 คู่ เป็นตัวอย่างข้อมูลในการทดลอง โดยนำคำศัพท์ทั้งหมดไปทำการเข้ารหัสด้วยขั้นตอนวิธีที่นำเสนอ และจัดเก็บคำศัพท์และรหัสคำในฐานข้อมูล หลังจากนั้นนำคำทั้งหมด 10,000 คำ ไปค้นคืนทีละคำจากฐานข้อมูล ด้วยขั้นตอนวิธีการค้นคืนที่ได้นำเสนอ เพื่อทำการคำนวณค่าแม่นยำ และค่าเรียกคืน โดยตั้ง α ให้มีค่าในระดับต่าง ๆ เพื่อหาความสัมพันธ์ระหว่าง α กับประสิทธิผลของการค้นคืน ได้ผลดังแสดงในรูปที่ 1

จากรูปที่ 1 แสดงให้เห็นว่าค่าแม่นยำของระบบค้นคืนสูงประมาณ 94 เปอร์เซ็นต์ และจะลดค่าลงอย่างต่อเนื่องเมื่อค่าของ α เพิ่มขึ้น ส่วนพฤติกรรมของค่าเรียกคืนจะเริ่มต้นประมาณ 17 เปอร์เซ็นต์และจะเพิ่มค่าขึ้นเมื่อค่า α เพิ่มขึ้น ประสิทธิผลของการค้นคืนที่ได้จากการทดลองพบว่าสอดคล้องกับพฤติกรรมโดยปกติของค่าแม่นยำและค่าเรียกคืนในระบบค้นคืนใด ๆ คือแนวโน้มของค่าแม่นยำและค่าเรียกคืนจะตรงข้ามกัน

จากผลการทดลองพบว่าขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษจะเกิดประสิทธิผลสูงสุดคือค่าแม่นยำเท่ากับ 69 เปอร์เซ็นต์และค่าเรียกคืนเท่ากับ 73 เปอร์เซ็นต์ เมื่อกำหนดค่าแอลฟาเท่ากับ 0.15



รูปที่ 1 ความสัมพันธ์ระหว่างค่า α กับประสิทธิผลการค้นคืน

5. สรุป

บทความนี้ได้นำเสนอขั้นตอนวิธีการเข้ารหัสคำทับศัพท์และการค้นคืน เพื่อการค้นค้นข้ามภาษาไทย-อังกฤษ การเข้ารหัสคำนี้ใช้กับกรณีของคำอังกฤษที่ทับศัพท์คำไทย ซึ่งประกอบด้วยขั้นตอนการแปลงรูปคำไทยซึ่งอาศัยหลักการถ่ายเสียง และใช้หลักการถอดอักษรเพื่อแปลงตัวอักษรอังกฤษเป็นตัวอักษรไทย ขั้นตอนการค้นคืนนั้นจะเข้าหัสคำของคำในข้อถาม เพื่อนำได้ไปเทียบรหัสแบบประมาณกับรหัสคำต่างๆในดัชนี โดยอาศัยเทคนิคกำหนดการพลวัตรหัสคำใดผ่านเกณฑ์การยอมรับในการเทียบสายอักษรที่กำหนดไว้ ก็ถือว่าเป็นคำทับศัพท์ที่ตรงกันของอีกภาษาหนึ่ง ผลการทดลองแสดงให้เห็นว่าขั้นตอนวิธีที่ได้นำเสนอนี้ให้ประสิทธิผลของการค้นคืนที่มีค่าเรียกคืนและค่าแม่นยำประมาณ 70 เปอร์เซ็นต์

เอกสารอ้างอิง

[1] W. Frakes, “Introduction to Information Storage and Retrieval System”, *Information Retrieval : Data Structure & Algorithm* , W.B. Frakes and R. Baeza-Yates ed., Prentice Hall, 1992.

[2] 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.

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

[4] S. Ongroongruang, R. Prongsirivattana, and V. Jantarasukree, “English to Thai Word Retrieval Using Sound Index”, *Proc 2nd SNLP’95*, Bangkok Thailand, Aug. 2-4, 1995, pp. 4 7-413.

[5] 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*, Bangkok Thailand, Aug. 19-21, 1998.

[6] กำชัย ทองหล่อ, “หลักภาษาไทย”, พิมพ์ครั้งที่ 10, กรุงเทพมหานคร : อมรการพิมพ์, 2540.

[7] S. Sethaputra, “Thaisoft So Sethaputra Dictionary Version 1.5” [Computer Software], Bangkok : Thaisoft Co.,Ltd, 1996.

[8] ประจักษ์ ประภาพิทยากร และคณะ, “รู้จักภาษาไทย”, พิมพ์ครั้งที่ 1, กรุงเทพมหานคร : โอเดียนบุ๊คสโตร์, 2519.

[9] International Organization for Standardization, “Information and Documentation – Transliteration of Thai”, *Draft International Standard ISO/DIS 11940*, 1996.

[10] พยนต์ ทิมเจริญ, “การเขียนชื่อภาษาไทยด้วยอักษรโรมัน”, วารสารแผนที่ ปีที่ 27 ฉบับที่ 2 (ตุลาคม-ธันวาคม 2527) : 61-74.

[11] ราชบัณฑิตยสถาน, “ประกาศราชบัณฑิตยสถาน เรื่อง การถอดอักษรไทยเป็นโรมัน”, 2482.

[12] จันทร์เพ็ญ โวหารสุนทร, “การศึกษาการใช้อักษรโรมันแทนอักษรไทย”, ปริญญานิพนธ์, คณะอักษรศาสตร์ มหาวิทยาลัยศรีนครินทร์วิโรฒ ประสานมิตร, 2530.

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

[14] นิลเนตร อรุณวงศ์ ณ อยุธยา, “การเปลี่ยนอักขระของคำในภาษาไทย โดยใช้หลักการของซาวด์เด็กซ์”, ปริญญานิพนธ์, คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2534.