3. ขั้นตอนวิธีการค้นคืนข้ามภาษาแบบภาษาไทยทับศัพท์ภาษาอังกฤษ

ในบทนี้จะกล่าวถึงขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษในส่วนของภาษาไทยทับศัพท์ภาษาอังกฤษ เช่น ค้นคืนเอกสารที่มีคำว่า "CLINTON" โดยใช้ข้อคำถาม "คลินตัน" เป็นต้น

3.1 โครงสร้างของระบบค้นคืนข้ามภาษาไทยทับศัพท์ภาษาอังกฤษ

ขั้นตอนวิธีของการค้นคืนข้ามภาษาแบบภาษาไทยทับศัพท์ภาษาอังกฤษมีขั้นตอนการทำงานหลัก ๆ คือ การเข้ารหัสคำ และการเปรียบเทียบรหัสคำ คือเมื่อผู้ใช้งานได้ระบุข้อคำถามให้กับระบบค้นคืนข้ามภาษาแล้ว ระบบจะทำการเข้ารหัสคำในข้อคำถาม เมื่อได้รหัสคำแล้วจะนำไปเปรียบเทียบกับรหัสคำในดัชนีคำหลักของเอกสารที่ได้เข้ารหัสไว้แล้วในขั้นตอนการทำดัชนี ถ้ามีรหัสคำเหมือนกันจะถือว่าคำหลักนั้นเป็นคำหลักที่ตรงกันในอีกภาษาหนึ่ง

3.2 ขั้นตอนวิธีการเข้ารหัสคำ

งานวิจัยนี้ใช้เทคนิคการเข้ารหัสคำในลักษณะคล้ายกับการเข้ารหัสซาวด์เด็กซ์ของ Odell และ Russell ซึ่งมีการใช้รหัสแทนกลุ่มตัวอักษรที่มีหน่วยเสียงใกล้เคียงกันดังที่ได้กล่าวไว้แล้วในบทที่ 2 โดยดัดแปลงให้ตารางการกำหนดรหัสซาวด์เด็กซ์สามารถใช้งานร่วมกับภาษาไทยได้ด้วยการจัดกลุ่มพยัญชนะไทย 44 ตัวอักษร เป็น 21 กลุ่มหน่วยเสียงตามหลักภาษาไทย [1] ดังแสดงในตารางที่ 3.1 และทำการรวม 21 กลุ่มหน่วยเสียงพยัญชนะไทยกับ 7 กลุ่มอักษรของซาวด์เด็กซ์เดิม โดยอาศัยหลักเกณฑ์ในการถอดอักษร ดังแสดงในตารางที่ 3.2

---

[1] ประจักษ์ ประภาพิทยากร และคณะ, รู้จักภาษาไทย. พิมพ์ครั้งที่ 1 (กรุงเทพมหานคร : ไอเดียนบุ๊คสโตร์, 2519), หน้า 12-68.

ก	ฎ ด	ฝ ฟ
ข ฃ ค ฅ ฆ	ฏ ต	ม
ง	ฐ ฑ ฒ ณ ท ธ	ร
จ	ญ น	ถ ผ
ฉ ช ฌ	บ	ว
ซ ศ ษ ส	ป	ห ฮ
ญ ย	ฝ พ ภ	อ

ตารางที่ 3.1 กลุ่มเสียงของพยัญชนะไทย 21 กลุ่ม

กาษาอังกฤษ	กาษาไทย
A E I O U H W Y	อ ห ฮ ว ญ ย
B F P V	บ ฝ ฟ ป ผ พ ภ ว
C G J K Q S X Z	ข ฃ ค ฅ ฆ ฉ ช ฌ ก จ ซ ศ ษ ส
D T	ฎ ด ฏ ต ฐ ฑ ฒ ณ ท ธ
L	ถ ผ
M N	ม ฌ น
R	ร

ตารางที่ 3.2 กลุ่มอักษรไทยและกลุ่มอักษรอังกฤษที่ออกเสียงคล้ายกันในรหัสซาวด์เด็กซ์

ผู้วิจัยเสนอการเข้ารหัสคำดังแสดงในตารางที่ 3.3 โดยดัดแปลงต้นแบบการเข้ารหัสคำซาวด์เด็กซ์ของ Odell และ Russell การกำหนดรหัสซาวด์เด็กซ์สำหรับอักษรไทยและอักษรอังกฤษที่นำเสนอมีรายละเอียดดังนี้

- ใช้รหัสตัวเลขแทนตัวอักษรตัวแรกแทนการใช้ตัวอักษรของคำ เนื่องจากพบว่ามีตัวอักษรอังกฤษหลายตัวถอดเป็นตัวอักษรไทยเหมือนกัน เช่น V และ W ถอดอักษรเป็นตัว ว และผู้วิจัยได้เสนอรหัสเพิ่มอีก 3 ตัว คือ 7 8 และ 9 ซึ่งจะใช้ในกรณีที่เป็นตัวอักษรแรกของคำเท่านั้น โดยมีรายละเอียดดังนี้

- รหัส 7 สำหรับคำที่ขึ้นต้น A E I O U หรือ อ เนื่องจากพบว่าคำภาษาอังกฤษที่ขึ้นต้นด้วยสระมักจะใช้ อ เป็นพยัญชนะต้นในการอ่านออกเสียง เช่น ABRAHAM (อับราฮัม) EDWARD (เอดเวิร์ด) IRIDIUM (อิริเดียม) OHM(โอห์ม) และ ULTRAVIOLET (อัลตราไวโอเลต) เป็นต้น
- รหัส 8 สำหรับคำที่ขึ้นต้น H เนื่องจาก H ที่เป็นตัวอักษรแรกของคำมักจะเป็นพยัญชนะและอ่านออกเสียง เช่น HOPKINS (ฮอปกินส์) ส่วน H ที่อยู่กลางคำมักจะเป็นอักษรควบและไม่อ่านออกเสียง เช่น WHITE (ไวต์) JOHN (จอห์น) SHOW (โชว์) เป็นต้น
- รหัส 9 สำหรับตัวอักษร Y เนื่องจากพบว่า Y ที่เป็นตัวอักษรแรกของคำมักจะเป็นพยัญชนะและอ่านออกเสียง เช่น YAHOO (ยาฮู) ส่วนY ที่อยู่กลางคำมักจะเป็นสระ เช่น ONYX (โอนิกซ์) PHYSICS (ฟิสิกส์) เป็นต้น

- เพิ่มรหัส 52 สำหรับตัวอักษร ง เนื่องจากอักษร ง ถอดมาจากตัวอักษรอังกฤษ NG หรือ NK เช่น KING สามารถถอดอักษรเป็น คิง (รหัส 5 สำหรับอักษร N และ รหัส 2 สำหรับอักษร G) และ LINK สามารถถอดอักษรเป็น ลิงค (รหัส 5 สำหรับอักษร N และ รหัส 2 สำหรับอักษร K) เป็นต้น
- ขยายความยาวของรหัสคำที่ได้จากเดิม 4 หลักเป็นไม่จำกัดความยาวของรหัสคำที่ได้ เนื่องจากพบว่าคำที่มีความยาวมากอาจได้รหัสตรงกันทั้ง ๆ ที่ออกเสียงไม่คล้ายกัน เช่น คริสต์เตียน ได้รหัส 262395 และ คริสต์โตฟเฟล ได้รหัส 262314 ซึ่งถ้าพิจารณารหัสเพียง 4 หลักคือ 2623 จะถือว่าทั้งสองคำออกเสียงคล้ายกัน
- เพิ่มพารามิเตอร์ k ซึ่งเป็นความยาวน้อยสุดของรหัสคำ โดยจะพิจารณาเฉพาะคำที่มีความยาวน้อยสุดของรหัสคำที่มากกว่า k หลัก ค่าของ k จะได้กล่าวในรายละเอียดต่อไป

ส่วนสระและวรรณยุกต์ในภาษาไทยจะไม่นำมาพิจารณาในการเข้ารหัสคำ เนื่องจากต้นแบบของการเข้ารหัสซาวด์เด็กซ์ภาษาอังกฤษของ Odell และ Russell ไม่นำสระมาพิจารณาในการเข้ารหัสคำ

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 ฉ ช ฌ 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

ตารางที่ 3.3 การกำหนดรหัสซาวด์เด็กซ์สำหรับอักษรไทยและอักษรอังกฤษที่นำเสนอ

3.3 วิธีการทดลอง

ผู้วิจัยได้ทำการทดลองขั้นตอนวิธีในการเข้ารหัสคำที่ได้นำเสนอในหัวข้อ 3.2 โดยใช้ชุดของคำอังกฤษและคำทับศัพท์ที่ตรงกันจำนวน 1,689 คู่ ดังแสดงในตารางที่ 3.4 ซึ่งส่วนใหญ่เป็นคำนามเฉพาะ คำศัพท์วิทยาศาสตร์ คำศัพท์คณิตศาสตร์ และคำศัพท์เคมี โดยนำมาจากหนังสือรวมคำศัพท์ของราชบัณฑิตยสถาน [2 3 4] และหนังสือเรียนวิชาเคมี [5]

---

[2] ราชบัณฑิตยสถาน, หลักเกณฑ์การทับศัพท์ ฉบับราชบัณฑิตยสถาน (2535).

[3] ราชบัณฑิตยสถาน, ศัพท์วิทยาศาสตร์ ฉบับราชบัณฑิตยสถาน (2536).

[4] ราชบัณฑิตยสถาน, ศัพท์คณิตศาสตร์ ฉบับราชบัณฑิตยสถาน (2540).

[5] กระทรวงศึกษาธิการ, หนังสือเรียนวิชาเคมี เล่ม 1 (ว 031) หลักสูตรมัธยมปลาย 2524 (2530).

ประเภทคำศัพท์	จำนวน (คู่)
คำนามเฉพาะทั่วไป	321
คำศัพท์คณิตศาสตร์	593
คำศัพท์วิทยาศาสตร์	604
คำศัพท์เคมี	171
รวม	1,689

ตารางที่ 3.4 รายละเอียดจำนวนคำศัพท์ที่ใช้ในการทดลอง

ผู้วิจัยได้ทำการทดลองโดยนำคำศัพท์ทั้งหมด 1,689 คู่ ไปทำการเข้ารหัสด้วยขั้นตอนวิธีที่นำเสนอในหัวข้อ 3.2 และจัดเก็บคำศัพท์และรหัสคำที่ได้เก็บในฐานข้อมูล หลังจากนั้นนำคำศัพท์ทั้งหมดไปค้นคืนทีละคำศัพท์กับฐานข้อมูลเพื่อทำการวัดค่าแม่นยำ และค่าเรียกคืนสำหรับการค้นคืน

ผู้วิจัยได้ทำการทดลองโดยการเปลี่ยนแปลงค่าพารามิเตอร์ k (ความยาวน้อยสุดของรหัสคำ) เพื่อหาความสัมพันธ์ระหว่างความยาวน้อยสุดของรหัสคำกับประสิทธิผลของระบบค้นคืน

การคำนวณหาค่าแม่นยำและค่าเรียกคืน [6] ใช้สูตรดังนี้

$$\text{ค่าแม่นยำ} = \frac{\text{จำนวนคำศัพท์ที่เกี่ยวข้องที่ค้นกลับมา}}{\text{จำนวนคำศัพท์ที่ค้นกลับมา}} \times 100$$

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

3.4 ผลการทดลอง

จากการทดลองในหัวข้อ 3.3 เพื่อหาความสัมพันธ์ระหว่างความยาวน้อยสุดของรหัสคำกับความยาวคำเฉลี่ย และความสัมพันธ์ระหว่างความยาวน้อยสุดของรหัสคำกับประสิทธิผลของ

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

ระบบค้นคืนโดยการเปลี่ยนแปลงค่าพารามิเตอร์ k ได้ผลแสดงในตารางที่ 3.5 ตารางที่ 3.6 และในรูปที่ 3.1 จากชุดคำทับศัพท์ที่ใช้ในการทดลองพบว่าคำศัพท์ที่มีความยาวน้อยสุดของรหัสคำที่มีค่ามากที่สุดคือ 10 หลัก

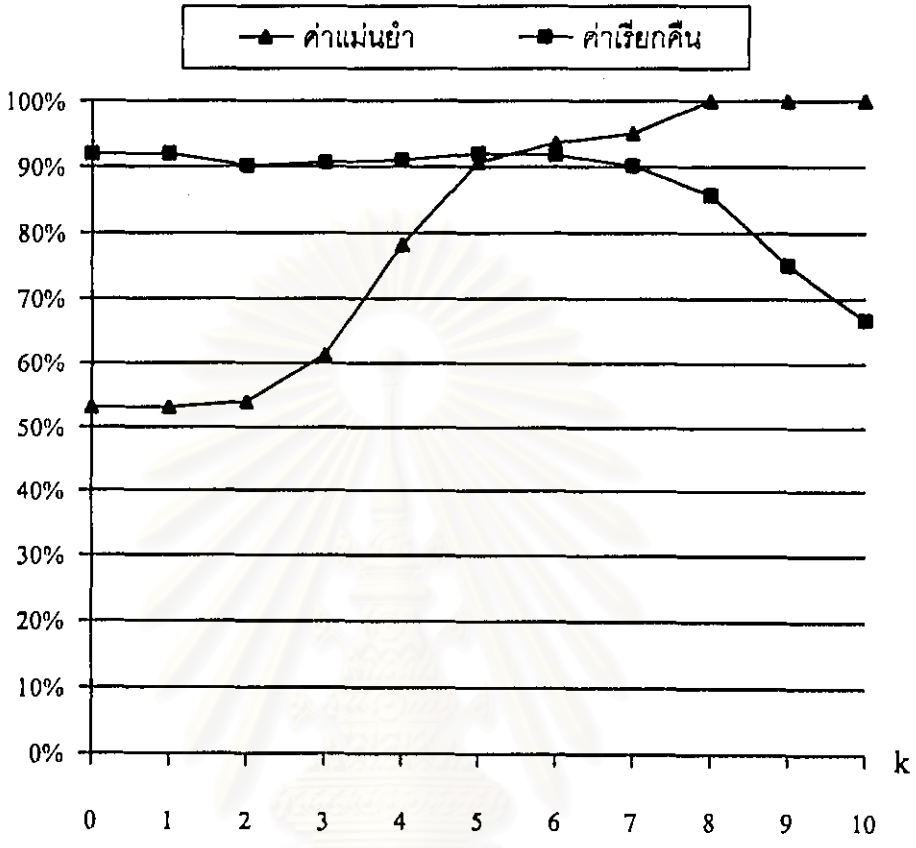
ความยาวน้อยสุดของรหัสคำ	ความยาวเฉลี่ยของคำ
1	4.13
2	4.99
3	6.23
4	7.65
5	9.00
6	10.49
7	12.05
8	13.85
9	17.67
10	17.83

ตารางที่ 3.5 ความสัมพันธ์ระหว่างความยาวน้อยสุดของรหัสคำกับความยาวคำเฉลี่ย

k	จำนวนคำ (คู่)	ค่าแม่นยำ	ค่าเรียกคืน
0	1689	0.529857	0.919941
1	1689	0.529857	0.919941
2	1649	0.538405	0.901152
3	1379	0.612388	0.905729
4	888	0.781167	0.909910
5	462	0.906050	0.919913
6	195	0.935531	0.917949
7	61	0.950820	0.901639
8	14	1.000000	0.857143
9	4	1.000000	0.750000
10	3	1.000000	0.666667

ตารางที่ 3.6 ความสัมพันธ์ระหว่างความยาวน้อยสุดของรหัสคำกับประสิทธิผล

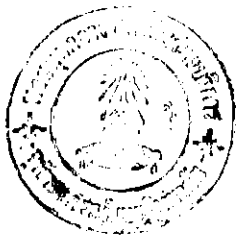
จากชุดคำทับศัพท์ที่ใช้ในการทดลองพบว่าคำศัพท์ที่มีความยาวน้อยสุดของรหัสคำที่มีค่ามากกว่า 7 หลัก มีจำนวนน้อยกว่า 1 เปอร์เซ็นต์ของคำศัพท์ทั้งหมดที่ใช้ในการทดลอง



รูปที่ 3.1 ความสัมพันธ์ระหว่างความยาวน้อยสุดของรหัสคำกับประสิทธิผล

จากรูปที่ 3.1 แสดงให้เห็นว่าค่าเรียกคืนของระบบค้นคืนสูงประมาณ 90 เปอร์เซ็นต์ และจะลดค่าลงอย่างช้า ๆ เมื่อความยาวน้อยสุดของรหัสคำเพิ่มขึ้น ส่วนพฤติกรรมของค่าแม่นยำจะเริ่มต้นประมาณ 52 เปอร์เซ็นต์และจะเพิ่มค่าขึ้นเมื่อความยาวน้อยสุดของรหัสคำเพิ่มขึ้น ประสิทธิผลของระบบที่ได้จากการทดลองพบว่า สอดคล้องกับพฤติกรรมโดยปกติของค่าแม่นยำและค่าเรียกคืนในระบบค้นคืนใด ๆ คือแนวโน้มของค่าแม่นยำและค่าเรียกคืนจะตรงข้ามกัน

จากผลการทดลองพบว่าขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษจะมีค่าแม่นยำสูงถึง 78 เปอร์เซ็นต์ และค่าเรียกคืนสูงถึง 90 เปอร์เซ็นต์ เมื่อรหัสคำที่ได้มีความยาวมากกว่า 4 หลักขึ้นไป หรือใช้คำทับศัพท์ที่มีความยาวมากกว่า 7 ตัวอักษรขึ้นไปในการค้นคืนข้ามภาษาไทย-อังกฤษ



3.5 สรุป

ในบทนี้ได้กล่าวถึงขั้นตอนวิธีการเข้ารหัสคำทับศัพท์เพื่อการค้นคืนข้ามภาษาไทย-อังกฤษ สำหรับภาษาไทยทับศัพท์ภาษาอังกฤษ การค้นคืนข้ามภาษานั้นสามารถทำได้โดยใช้รหัสที่ได้ไปค้นหาจากดัชนีคำหลักของเอกสารที่ได้เข้ารหัสไว้แล้วในขั้นตอนการทำดัชนี โดยขั้นตอนวิธีที่นำเสนอนี้จะอนุญาตให้ผู้ใช้สามารถระบุคำหลักในการค้นหาเป็นภาษาอังกฤษหรือเป็นภาษาไทยที่ทับศัพท์คำหลักนั้น

ขั้นตอนวิธีที่ได้นำเสนอนี้ ผู้วิจัยได้นำขั้นตอนวิธีซาวด์เด็กซ์ภาษาอังกฤษของ Odell และ Russell [7] มาดัดแปลงเพียงเล็กน้อยโดยการเปลี่ยนแปลงตารางการกำหนดรหัสคำให้ใช้ได้กับตัวอักษรไทยและไม่จำกัดความยาวของรหัสที่ได้

ผลการทดลองพบว่าค่าแม่นยำสูงถึง 78 เปอร์เซ็นต์ และค่าเรียกคืนสูงถึง 90 เปอร์เซ็นต์ เมื่อใช้คำทับศัพท์ที่มีความยาวมากกว่า 7 ตัวอักษรขึ้นไปในการค้นคืนข้ามภาษาไทย-อังกฤษ

---

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