# การค้นคืนข้ามภาษาสำหรับคำทับศัพท์ภาษาไทย/อังกฤษโดยใช้ทรานฟอร์เมอร์

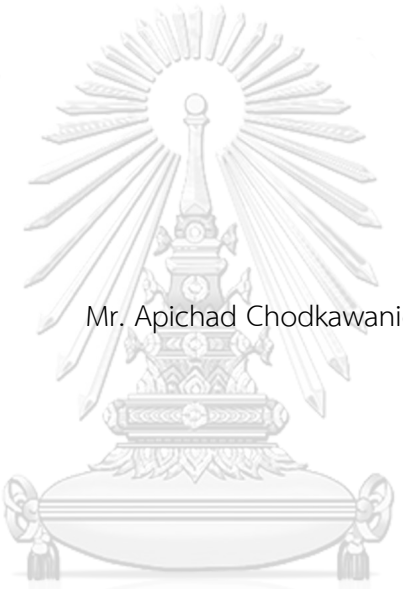
นายอภิชัจจ์ โชดกวณิชย์



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

# Thai/English cross-language transliterated word retrieval using Transformer

Mr. Apichad Chodkawanich





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 2020

| | |
|---|---|
| หัวข้อวิทยานิพนธ์ | การค้นคืนข้ามภาษาสำหรับคำทับศัพท์ภาษาไทย/อังกฤษ โดยใช้ทรานฟอร์เมอร์ |
| โดย | นายอภิชัจจ์ โชดกวณิชย์ |
| สาขาวิชา | วิทยาศาสตร์คอมพิวเตอร์ |
| อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก | ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล |

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.พีรพล เวทีกูล)

..... กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.ทัศนวรรณ ศูนย์กลาง)

อภิชัจจ์ โชดกวณิชย์ : การค้นคืนข้ามภาษาสำหรับคำทับศัพท์ภาษาไทย/อังกฤษโดยใช้ทรานฟอร์เมอร์. ( Thai/English cross-language transliterated word retrieval using Transformer) อ.ที่ปรึกษาหลัก : ศ. ดร.บุญเสริม กิจศิริกุล

การค้นคืนข้ามภาษานั้นเป็นงานที่ท้าทายในวิทยาการด้านการประมวลผลภาษาธรรมชาติของไทย ด้วยเหตุผลในด้านของความแตกต่างระหว่างภาษา เช่น การออกเสียง และ กฎการทับศัพท์ วิทยานิพนธ์เล่มนี้ได้นำเสนอ ขั้นตอนวิธีการค้นคืนข้ามภาษาสำหรับคำทับศัพท์ภาษาไทย/อังกฤษโดยใช้ทรานฟอร์เมอร์ วิธีการที่นำเสนอนี้ช่วยให้สามารถค้นคืนคำทับศัพท์ข้ามภาษาได้โดยไม่ต้องอาศัยพจนานุกรม ซึ่งการค้นคืนข้ามภาษาโดยไม่อาศัยพจนุกรมนั้นจำเป็นต้องใช้หลักการเข้ารหัสเสียงซึ่งเป็นสัญลักษณ์แทนเสียงอ่านของคำ จากผลการทดลองของโมเดลการเรียนรู้แบบกึ่งสอน (Semi-supervised) ด้วยวิธี K-Fold cross validation แสดงให้เห็นว่า ขั้นตอนวิธีการเข้ารหัสคำที่นำเสนอให้ค่าเฉลี่ยของค่าแม่นยำ ค่าเรียกคืน และค่า F1 อยู่ที่ 85.08%, 88.25% และ 86.63% ตามลำดับ สำหรับชุดข้อมูลภาษาไทย และค่าเฉลี่ยของค่าแม่นยำ ค่าเรียกคืน และค่า F1 ของชุดข้อมูลภาษาอังกฤษอยู่ที่ 80.44%, 87.15% และ 83.66% ตามลำดับ



สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต .....
ปีการศึกษา 2563 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

# # 6270310421 : MAJOR COMPUTER SCIENCE
KEYWORD: TRANSLITERATED WORD, INFORMATION RETRIEVAL, TRANSFORMER, SUPERVISED LEARNING, SEMI-SUPERVISED LEARNING, THAI LANGUAGE

Apichad Chodkawanich : Thai/English cross-language transliterated word retrieval using Transformer. Advisor: Prof. BOONSERM KIJSIRIKUL, Ph.D.

Cross-language transliterated word retrieval is a challenging task for Thai Natural Language Processing due to the difference between languages such as pronunciation and transliteration rules. This thesis presents Thai/English cross-language transliterated word retrieval using Transformer. The proposed method enables transliterated word retrieval without using a dictionary. The phonetic code is used for cross-language retrieval. The phonetic code of a word represents the sound of a word. The results from our semi-supervised model using K-Fold cross validation showed that the model yielded precision, recall and F1 at 85.08%, 88.25% and 86.63% respectively for Thai-based datasets, and 80.44%, 87.15% and 83.66% respectively for English-based datasets.



| | | |
|---|---|---|
| Field of Study: | Computer Science | Student's Signature ..... |
| Academic Year: | 2020 | Advisor's Signature ..... |

# กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดีด้วยความกรุณาอย่างยิ่งของ ศาสตราจารย์ ดร. บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาหลัก โดยท่านได้ให้โอกาส คำแนะนำในการทำวิจัย แนวทางการแก้ปัญหา และข้อคิดเห็นต่าง ๆ รวมทั้งการตรวจแก้วิทยานิพนธ์ฉบับนี้อย่างละเอียด ทำให้การวิจัยลุล่วงและประสบความสำเร็จมาโดยตลอดระยะเวลาการศึกษาและการทำวิจัย ขอกราบขอบพระคุณเป็นอย่างสูงมา ณ ที่นี้ด้วย

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.พีรพล เวทีกูล และ ผู้ช่วยศาสตราจารย์ ดร.ทัศนวรรณ ศูนย์กลาง กรรมการการสอบวิทยานิพนธ์ที่ได้ให้คำแนะนำและแนวทางที่เป็นประโยชน์ต่อการทำวิทยานิพนธ์ในครั้งนี้

ขอขอบพระคุณ พี่ศิริพจน์ สุรบถโสภณ สำหรับการแบ่งปันข้อมูล ที่ถูกนำมาในการวิจัยในครั้งนี้ รวมทั้งบรรดาเพื่อน ๆ ทั้งรุ่นพี่ และ รุ่นน้อง ภาควิชาวิศวกรรมคอมพิวเตอร์ภาคนอกเวลาทุกคนที่ร่วมแลกเปลี่ยนความรู้ แง่คิดต่าง ๆ

ท้ายที่สุดนี้ ผู้วิจัยขอกราบขอบพระคุณบิดา มารดา รวมถึงญาติ ทุก ๆ คนที่คอยให้ความสนับสนุน และความห่วงใย รวมถึงให้กำลังใจเสมอมา

อภิชัจจ์ โชดกวณิชย์

# สารบัญ

**หน้า**


..... ค

บทคัดย่อภาษาไทย ..... ค

..... ง

บทคัดย่อภาษาอังกฤษ ..... ง

กิตติกรรมประกาศ ..... จ

สารบัญ ..... ฉ

สารบัญตาราง ..... ฌ

สารบัญภาพ ..... ฎ

บทที่ 1 บทนำ ..... 1

1.1 ความเป็นมาและความสำคัญ ..... 1

1.2 วัตถุประสงค์ของการวิจัย ..... 2

1.3 ขอบเขตการวิจัย ..... 2

1.4 ขั้นตอนและวิธีการดำเนินการวิจัย ..... 2

1.5 ประโยชน์ที่คาดว่าจะได้รับ ..... 3

1.6 ผลงานตีพิมพ์จากงานวิจัย ..... 3

บทที่ 2 ทฤษฎีที่เกี่ยวข้อง ..... 4

2.1 การเรียนรู้ของเครื่อง (Machine Learning) ..... 4

2.2 ทรานฟอร์เมอร์ (Transformer) ..... 5

2.2.1 ตัวเข้ารหัส (Encoder) ..... 6

2.2.2 ตัวถอดรหัส (Decoder) ..... 6

2.2.3 ความสนใจ (Attention) ..... 7

2.2.4 ความสนใจแบบผลคูณจุดปรับขนาด (Scaled Dot-Product Attention) ..... 7
2.2.5 ความสนใจหลายหัว (Multi-Head Attention) ..... 8
2.3 การถอดอักษร (Transliteration) ..... 8
2.4 การถ่ายเสียงด้วยตัวอักษรโรมัน (Romanization) ..... 9
2.5 การวัดผลการค้นคืน ..... 10
2.6 ขั้นตอนวิธีระยะการแก้ไขสั้นที่สุด (Minimum Edit Distance) ..... 10
2.7 ขั้นตอนวิธีซาวด์เด็กซ์ภาษาอังกฤษ (Soundex) ..... 11
บทที่ 3 งานวิจัยที่เกี่ยวข้อง ..... 13
3.1 งานวิจัยของ วรรณี อุดมพาณิชย์ ..... 13
3.2 งานวิจัยของ ประยุทธ สุวรรณวิสารท และ สมชาย ประสิทธิ์จูตระกูล ..... 15
3.3 งานวิจัยของ ทัศนวรรณ ศูนย์กลาง สมชาย ประสิทธิ์จูตระกูล และบุญเสริม กิจศิริกุล ..... 19
3.4 งานวิจัยของ ศิริพจน์ สุรบถโสภณ และ บุญเสริม กิจศิริกุล ..... 20
3.5 งานวิจัยของ S. Tasanaprasert, W. Pokasame, S. Rattanaliam ..... 21
บทที่ 4 การเข้ารหัสคำ ..... 24
4.1 รหัสคำ ..... 24
4.2 การประมวลผลเบื้องต้น ..... 27
4.3 การเข้ารหัสคำ ..... 28
บทที่ 5 การค้นคืนข้ามภาษา ..... 30
5.1 การคำนวณความต่างของรหัสคำ ..... 30
5.2 เกณฑ์การประเมินการค้นคืน ..... 31
บทที่ 6 การทดลอง ..... 32
6.1 โมเดล (Model) ..... 32
6.2 ชุดข้อมูล (Dataset) ..... 32
6.1.1 ชุดข้อมูลที่มีฉลากกำกับ (Labeled data) ..... 32

6.1.2 ชุดข้อมูลที่ไม่มีฉลากกำกับ (Unlabeled data) ..... 33
6.3 การปรับจูนค่าพารามิเตอร์ (Hyperparameter Tuning) ..... 34
6.4 การเข้ารหัสคำโดยใช้ทรานฟอร์เมอร์ ..... 36
6.3.1 การผสมชุดข้อมูลที่มีฉลากกำกับและไม่มีฉลากกำกับ ..... 36
6.3.2 การปรับแต่งฟังก์ชันต้นทุน (Loss function) ..... 37
6.5 ผลการทดลอง ..... 38
6.6 วิเคราะห์ผลการทดลองการเข้ารหัสคำ ..... 45
6.7 สรุป ..... 46
บทที่ 7 สรุปผลการวิจัยและข้อเสนอแนะ ..... 47
7.1 สรุปผลการวิจัย ..... 47
7.2 ข้อเสนอแนะ ..... 47
บรรณานุกรม ..... 49
ภาคผนวก ..... 51
ภาคผนวก ก การใช้อักษรโรมันแทนอักขระไทย ..... 52
ภาคผนวก ข หน่วยเสียงในภาษาไทยและภาษาอังกฤษ ..... 54
หน่วยเสียงในภาษาไทย ..... 54
ระบบเสียงในภาษาอังกฤษ ..... 55
ภาคผนวก ค ตัวอย่างข้อมูลคำทับศัพท์ที่ใช้ในงานวิจัย ..... 56
ตัวอย่างคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ ..... 56
ตัวอย่างคำไทยและคำอังกฤษทับศัพท์คำไทย ..... 57
ภาคผนวก ง ค่าพารามิเตอร์ที่ดีที่สุดใช้ในการทดลอง ..... 58
ประวัติผู้เขียน ..... 59

# สารบัญตาราง

**หน้า**

*ตารางที่ 1 การกำหนดรหัสซาวด์เด็กซ์ภาษาอังกฤษของ Odell และ Russel* ..... 12

ตารางที่ 2 การกำหนดรหัสตัวอักษรของรหัสซาวด์เด็กซ์ภาษาไทย จากงานวิจัยของวรรณี อุดมพาณิชย์ ..... 14

ตารางที่ 3 การกำหนดรหัสตัวเลขของรหัสซาวด์เด็กซ์ภาษาไทย จากงานวิจัยของของวรรณี อุดมพาณิชย์ ..... 14

ตารางที่ 4 การกำหนดรหัสสำหรับคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ ..... 16

ตารางที่ 5 การกำหนดรหัสของพยัญชนะสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัยของประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล ..... 16

ตารางที่ 6 การกำหนดรหัสของสระสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัยของประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล ..... 17

ตารางที่ 7 ตัวอย่างการกำหนดต้นทุนการแทนที่อักขระสำหรับพยัญชนะ จากงานวิจัยของประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล ..... 18

ตารางที่ 8 ตัวอย่างการกำหนดต้นทุนการแทนที่อักขระสำหรับสระ จากงานวิจัยของประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล ..... 18

ตารางที่ 9 การกำหนดรหัสสำหรับอักขระตัวแรก ..... 21

ตารางที่ 10 การกำหนดรหัสสำหรับอักขระถัดจากตัวแรก ในกรณีที่เป็นพยัญชนะ ..... 22

ตารางที่ 11 การกำหนดรหัสสำหรับอักขระถัดจากตัวแรก ในกรณีที่เป็นสระ ..... 22

ตารางที่ 12 รหัสเสียงสำหรับคำไทยทับศัพท์คำอังกฤษ ..... 25

ตารางที่ 13 รหัสเสียงสำหรับคำอังกฤษทับศัพท์คำไทย ..... 26

ตารางที่ 14 จำนวนคู่คำที่มีรหัสคำไม่เหมือนกัน ..... 33

ตารางที่ 15 ตัวอย่างคำและรหัสคำ ..... 34

ตารางที่ 16 ผลการปรับค่าพารามิเตอร์ ..... 35

ตารางที่ 17 ค่าความถูกต้องของโมเดลการเรียนรู้แบบสอน ..... 38

ตารางที่ 18 ค่าความถูกต้องของโมเดลการเรียนรู้แบบกึ่งสอน ..... 39
ตารางที่ 19 การเปรียบเทียบผลของค่าความถูกต้องของงานวิจัยนี้ และงานวิจัยของศิริพจน์ สุรบถโสภณ และบุญเสริม กิจศิริกุล ..... 39
ตารางที่ 20 ค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต ของชุดข้อมูลคำไทย ..... 40
ตารางที่ 21 ค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต ของชุดข้อมูลคำอังกฤษทับศัพท์คำไทย ..... 40
ตารางที่ 22 ค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต ของชุดข้อมูลคำอังกฤษ ..... 41
ตารางที่ 23 ค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต ของชุดข้อมูลคำไทยทับศัพท์คำอังกฤษ ..... 41
ตารางที่ 24 สรุปผลค่าความถูกต้องเฉลี่ยแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต ..... 42
ตารางที่ 25 ค่าความแม่นยำของโมเดลการเรียนรู้แบบกึ่งสอน ..... 43
ตารางที่ 26 ค่าเรียกคืนของโมเดลการเรียนรู้แบบกึ่งสอน ..... 44
ตารางที่ 27 ตัววัด F1 ของโมเดลการเรียนรู้แบบกึ่งสอน ..... 44
ตารางที่ 28 สรุปผลค่าเฉลี่ยความแม่นยำ ค่าเรียกคืน และตัววัด F1 ของโมเดลการเรียนรู้แบบกึ่งสอน ..... 45
ตารางที่ 29 การใช้อักษรโรมันแทนพยัญชนะไทยของราชบัณฑิตยสถาน ..... 52
ตารางที่ 30 การใช้อักษรโรมันแทนสระไทยของราชบัณฑิตยสถาน ..... 53
ตารางที่ 31 หน่วยเสียงพยัญชนะในภาษาไทย ..... 54
ตารางที่ 32 หน่วยเสียงพยัญสระในภาษาไทย ..... 54
ตารางที่ 33 หน่วยเสียงพยัญชนะในภาษาอังกฤษ ..... 55
ตารางที่ 34 หน่วยเสียงสระในภาษาอังกฤษ ..... 55
ตารางที่ 35 ค่าพารามิเตอร์ที่ดีที่สุดใช้ในการทดลอง ..... 58

# สารบัญภาพ

**หน้า**

ภาพที่ 1 ประเภทการเรียนรู้ของเครื่อง ..... 4
ภาพที่ 2 โครงสร้างโมเดลของทรานฟอร์เมอร์ [3] ..... 5
ภาพที่ 3 ความสนใจแบบผลคูณจุดปรับขนาด (Scaled Dot-Product Attention) [3] ..... 7
ภาพที่ 4 ความสนใจหลายหัว (Multi-Head Attention) [3] ..... 8
ภาพที่ 5 ตัวอย่างการเข้ารหัสคำไทยจากงานวิจัยของทัศนวรรณ ศูนย์กลาง และคณะ ..... 19
ภาพที่ 6 ตัวอย่างการเข้ารหัสคำอังกฤษจากงานวิจัยของทัศนวรรณ ศูนย์กลาง และคณะ ..... 20
ภาพที่ 7 การเข้ารหัสคำโดยใช้ทรานฟอร์เมอร์ ..... 24
ภาพที่ 8 การฝึกสอนโมเดล ..... 36
ภาพที่ 9 การปรับแต่งฟังก์ชันต้นทุน (Loss function) ..... 37
ภาพที่ 10 ค่าความถูกต้องเฉลี่ยแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต ..... 42

# บทที่ 1
# บทนำ

## 1.1 ความเป็นมาและความสำคัญ

ในปัจจุบันอินเตอร์เน็ต คอมพิวเตอร์ และเครื่องมือสื่อสารอิเล็กทรอนิกส์นั้นเป็นส่วนหนึ่งในการดำรงชีวิตประจำวันของมนุษย์ เกือบจะเรียกได้ว่าสิ่งเหล่านี้เป็นสิ่งที่จำเป็นต่อการใช้ชีวิตไปเลยก็ว่าได้ เพราะสื่ออิเล็กทรอนิกส์นั้นมีข้อมูลที่สามารถเข้าถึงเป็นจำนวนมาก อีกทั้งยังสามารถค้นหาข้อมูลได้ในเกือบทุกสาขาที่สนใจ เนื่องจากสามาถเข้าถึงได้ง่ายทั้งจากระยะใกล้และระยะไกล ทำให้การแลกเปลี่ยนข้อมูลเกิดขึ้นกับกองข้อมูลสารสนเทศมีจำนวนมาก และทำให้ข้อมูลนั้นเพิ่มมากขึ้นตลอดเวลาอีกด้วย

การค้นคืนสารสนเทศข้ามภาษา หรือการสืบค้นข้ามภาษา (Cross-Language Information Retrieval หรือ CLIR) เป็นฟิลด์ย่อยของการค้นคืนข้อมูล (Information Retrieval) หมายถึง การค้นคืนสารสนเทศซึ่งภาษาที่แสดงในเอกสารไม่ตรงกับภาษาที่ใช้เป็นคำค้นหา [1]

ปัจจุบันเอกสารทางวิชาการในประเทศไทยมักจะจัดทำทั้งในรูปภาษาไทยและภาษาอังกฤษเพื่อประโยชน์ในการเผยแพร่ทั้งภายในและภายนอกประเทศ ซึ่งเอกสารเหล่านี้โดยเฉพาะอย่างยิ่งเอกสารทางด้านวิทยาศาสตร์และวิศวกรรมศาสตร์โดยมากแล้วมักจะปรากฏคำนามเฉพาะ (Proper Noun) และคำศัพท์เทคนิคต่าง ๆ เป็นจำนวนมากซึ่งจะพบได้ทั้งในรูปแบบของคำภาษาอังกฤษ คำภาษาไทยทับศัพท์ภาษาอังกฤษ คำภาษาอังกฤษทับศัพท์ภาษาไทย หรือคำในภาษาไทยเอง ดังนั้น ถ้าระบบค้นคืนสารสนเทศไม่สนับสนุนการทำงานข้ามภาษาก็จะทำให้ประสิทธิภาพในการค้นคืนต่ำ และใช้ประโยชน์จากสารสนเทศที่มีอยู่ได้ไม่เต็มที่



ปัญหาในการค้นคืนคำทับศัพท์ข้ามภาษามีหลายประการโดยเฉพาะการที่คำในภาษาหนึ่งอาจจะถูกเขียนในอีกภาษาหนึ่งได้หลายรูปแบบ ตัวอย่างเช่น คำภาษาอังกฤษ “Clinic” เมื่อเขียนทับศัพท์ในภาษาไทยอาจพบได้ทั้ง “คลินิก” “คลีนิก” หรือ “คลีนิค” หรือ คำภาษาไทย “ชุมพร” อาจเขียนทับศัพท์ในภาษาอังกฤษเป็น “Chumphon” หรือ “Chumporn” เป็นต้น ซึ่งระบบค้นคืนควรจะคืนคำเหล่านั้นมาทั้งหมดหรือให้ได้มากที่สุด แม้ว่าจะมีการนำพจนานุกรมสองภาษา (Bilingual Dictionary) มาใช้ในระบบ ก็ไม่อาจจะแก้ปัญหานี้ได้ตลอด เนื่องจากมีคำศัพท์เทคนิคใหม่ ๆ มากมายหลายสาขาเกิดขึ้นอยู่เสมอ และคำเหล่านี้ส่วนมากมักไม่ปรากฏในพจนานุกรม [2] ดังนั้นจึงทำให้การค้นหาด้วยคำหลักในภาษาหนึ่ง อาจจะทำให้พลาดเอกสารที่มีคำหลักตรงกันในอีกภาษาหนึ่งได้

ด้วยเหตุนี้จึงได้มีการใช้รหัสคำมาช่วยในการค้นคำทับศัพท์ข้ามภาษา โดยรหัสคำนี้จะเป็นสัญลักษณ์แทนเสียงอ่านของคำ คำที่มีเสียงอ่านตรงกันจะมีรหัสคำที่ตรงกันหรือใกล้เคียงกัน โดยการวิจัยนี้มุ่งเน้นการเข้ารหัสคำทับศัพท์ภาษาไทย/ภาษาอังกฤษ เพื่อการค้นคืนข้ามภาษาโดยใช้เทคนิคทรานฟอร์เมอร์ ซึ่งขั้นตอนการเข้ารหัสคำทั้งภาษาไทยทับศัพท์ภาษาอังกฤษ และคำภาษาอังกฤษทับศัพท์คำภาษาไทยจะใช้แนวทางเดียวกันในการแก้ปัญหา แม้ว่าทั้ง 2 กรณีจะใช้ความรู้ของการทับศัพท์ที่แตกต่างกันในการแก้ปัญหาก็ตาม

การวิจัยนี้มีข้อสมมุติฐานว่าขั้นตอนที่นำเสนอจะสามารถทำการสืบค้นคำทับศัพท์ข้ามภาษาไทย-ภาษาอังกฤษได้โดยไม่ต้องอาศัยพจนานุกรม

**1.2 วัตถุประสงค์ของการวิจัย**

ออกแบบและพัฒนาวิธีการเข้ารหัสคำและการการค้นคืนข้ามภาษาสำหรับคำทับศัพท์ภาษาไทย-อังกฤษโดยใช้ทรานฟอร์เมอร์

**1.3 ขอบเขตการวิจัย**

1.3.1 คำทับศัพท์ที่ใช้เป็นคำทับศัพท์ระหว่างภาษาไทยและภาษาอังกฤษเท่านั้น

1.3.2 คำศัพท์ในภาษาอังกฤษที่ใช้ไม่รวมถึงคำย่อ (Abbreviation) และคำรัสพจน์ (Acronym)

1.3.3 ยึดหลักเกณฑ์การออกเสียงของคำตามหลักของราชบัณฑิตยสถาน

1.3.4 นำข้อมูลที่มีป้ายติดกำกับ (Labeled data) และข้อมูลที่ไม่มีป้ายติดกำกับ (Unlabeled data) มาผสมกันและนำชุดข้อมูลทั้งหมดมาเรียนรู้แบบกึ่งสอน (Semi-Supervised Learning)

1.3.5 เปรียบเทียบผลการทดลองกับงานวิจัย [13] ของ ศิริพจน์ สุรบถโสภณ โดยเปรียบเทียบกับวิธีการนิวรอลเน็ตเวิร์กเท่านั้น

**1.4 ขั้นตอนและวิธีการดำเนินการวิจัย**

1.4.1 ศึกษาขั้นตอนวิธีการเข้ารหัสคำโดยใช้ทรานฟอร์เมอร์

1.4.2 รวบรวมและจัดเก็บชุดข้อมูลคำทับศัพท์เพื่อใช้ในการทดลอง และกำหนดรหัสคำของแต่ละคำศัพท์

1.4.3 แปลงข้อมูลคำศัพท์ให้อยู่ในรูปที่ใช้สำหรับฝึกสอนทรานฟอร์เมอร์

1.4.4 ฝึกสอนและทดสอบทรานฟอร์เมอร์ เพื่อใช้เป็นตัวสร้างรหัสคำ

1.4.5 ทำการทดลองและปรับปรุงผลการทดลอง

1.4.6 สรุปผลการทดลอง

**1.5 ประโยชน์ที่คาดว่าจะได้รับ**

สามารถนำไปใช้ในระบบการสืบค้นข้อมูลให้สามารถค้นคืนข้ามภาษาไทยและภาษาอังกฤษได้โดยไม่ต้องอาศัยพจนานุกรมและสามารถรองรับคำที่ศัพท์ที่เกิดขึ้นใหม่ได้ รวมทั้งสามารถใช้เป็นแนวทางในการสร้างระบบการค้นคืนข้ามภาษาในภาษาอื่นหรือการค้นคืนด้วยวิธีการที่ดียิ่งขึ้นได้

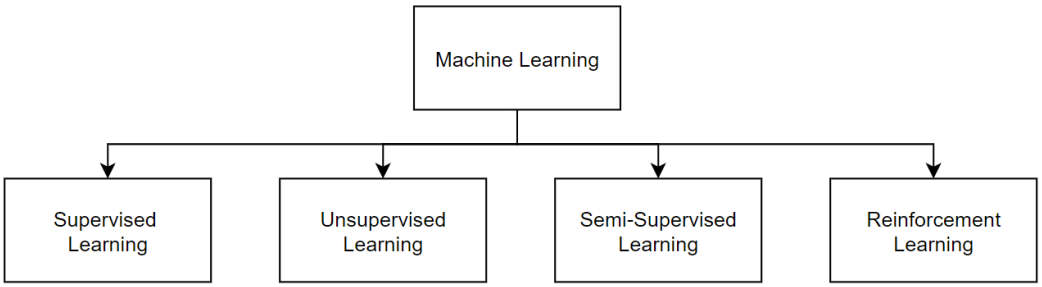
**1.6 ผลงานตีพิมพ์จากงานวิจัย**

วิทยานิพนธ์นี้ได้ตีพิมพ์และนำเสนอในงานประชุมทางวิชาการ International Conference on Natural Language Processing (ICNLP 2021) เมื่อวันที่ 26-28 มีนาคม พ.ศ. 2564 ในบทความเรื่อง “การค้นคืนข้ามภาษาสำหรับคำทับศัพท์ภาษาไทย/อังกฤษโดยใช้ทรานฟอร์เมอร์” (Thai/English cross-language transliterated word retrieval using Transformer) โดยผู้นำเสนอคือ อภิชัจจ์ โชดกวณิชย์ และ บุญเสริม กิจศิริกุล

# บทที่ 2
# ทฤษฎีที่เกี่ยวข้อง

## 2.1 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง คือศาสตร์และเทคนิคในการทำให้คอมพิวเตอร์สามารถเรียนรู้ และทำงานเหมือนมนุษย์ และปรับปรุงการเรียนรู้เมื่อเวลาผ่านไปในรูปแบบอิสระโดยการป้อนข้อมูล ข้อมูลที่อยู่ในรูปแบบของการสังเกต และข้อมูลที่มาจากโลกแห่งความจริงจากแหล่งต่าง ๆ การเรียนรู้ของเครื่องจะสามารถแบ่งออกได้เป็น 4 ประเภท ได้แก่

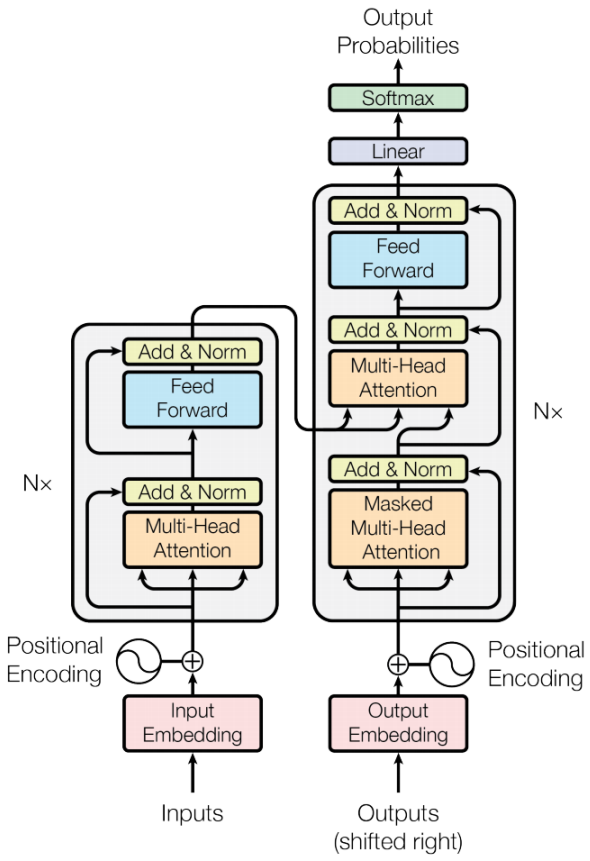



*ภาพที่ 1 ประเภทการเรียนรู้ของเครื่อง*

1. การเรียนรู้แบบสอน (Supervised Learning) คือ ชุดข้อมูลที่เป็นตัวอย่างและมีฉลากกำกับ นั้นจะถูกป้อนเข้าไปยังคอมพิวเตอร์ โดยมีเป้าหมายในการหากฎความเชื่อมโยงระหว่างชุดข้อมูลขาเข้าและขาออก
2. การเรียนรู้แบบไม่สอน (Unsupervised Learning) คือ ชุดข้อมูลตัวอย่างจะไม่มีการทำฉลากใด ๆ นั้นจะถูกป้อนเข้าไปยังคอมพิวเตอร์ และให้คอมพิวเตอร์หาโครงสร้างของข้อมูลขาเข้าและขาออก
3. การเรียนรู้แบบกึ่งสอน (Semi-Supervised Learning) คือ คือการเรียนรู้ที่อยู่ระหว่างการเรียนรู้แบบสอน และ การเรียนรู้แบบไม่สอน โดยชุดข้อมูลตัวอย่างที่มีฉลาก และชุดข้อมูลตัวอย่างที่ไม่มีข้อมูลฉลากนั้นจะถูกป้อนเข้าไปยังคอมพิวเตอร์ โดยที่ข้อมูลที่มีฉลากนั้นจะมีจำนวนที่น้อยกว่าข้อมูลที่ไม่มีฉลาก
4. การเรียนรู้แบบเสริมกำลัง (Reinforcement Learning) คือ คอมพิวเตอร์จะมีปฏิสัมพันธ์กับสิ่งแวดล้อมต่าง ๆ ที่มีเป้าหมายในการเรียนรู้บางอย่าง เช่นการขับรถ หรือเล่นเกม ในขณะที่ทำการทดลองนั้นโปรแกรมจะให้ข้อเสนอแนะกลับมาและพยายามทำให้ประสิทธิภาพดีขึ้น

**2.2 ทรานฟอร์เมอร์ (Transformer)**

ทรานฟอร์เมอร์ [3] คือ โมเดลการเรียนรู้เชิงลึก (Deep Learning) ที่ถูกสร้างขึ้นในปี ค.ศ. 2017 ซึ่งเป็นที่นิยมใช้งานในด้านการประมวลผลภาษาธรรมชาติหรือภาษามนุษย์ (Natural Language Processing) ทรานฟอร์เมอร์ ถูกออกแบบมาเพื่อจัดการกับข้อมูลที่เป็นลำดับ เช่น ภาษาธรรมชาติหรือภาษามนุษย์ เหมาะสำหรับงานในด้านการแปลภาษา และการสรุปข้อความต่าง ๆ




*ภาพที่ 2 โครงสร้างโมเดลของทรานฟอร์เมอร์ [3]*

สถาปัตยกรรมของทรานฟอร์เมอร์จะเป็นแบบ ตัวเข้ารหัส (Encoder) และตัวถอดรหัส (Decoder) จากภาพที่ 2 ด้านซ้ายจะเป็นตัวเข้ารหัส ส่วนด้านขวาจะเป็นตัวถอดรหัส ทั้งสองตัวประกอบด้วยโมดูลที่สามารถซ้อนทับได้หลาย ๆ ครั้ง จะเห็นว่าโมดูลต่าง ๆ จะประกอบไปด้วยเลเยอร์ความสนใจหลายหัว (Multi-Head Attenion) และแบบป้อนไปด้านหน้า (Feed Forward) เป็นหลัก

2.2.1 ตัวเข้ารหัส (Encoder)

ตัวเข้ารหัสแต่ละตัวนั้นประกอบไปด้วย 2 ส่วนหลักๆ ส่วนแรกคือ กลไกความสนใจตนเอง (Self-attention) และโครงข่ายประสาทเทียมแบบป้อนไปด้านหน้า (Feed-forward Neural Network) การทำงานของกลไกความสนใจตนเองนั้นจะรับชุดข้อมูลจากตัวเข้ารหัสก่อนหน้าและทำการปรับค่าน้ำหนัก และสร้างชุดข้อมูลขาออก จากนั้นโครงข่ายประสาทเทียมจะประมวลผลข้อมูลแต่ละตัวแยกกัน ซึ่งข้อมูลขาออกเหล่านี้จะถูกส่งไปยังตัวเข้ารหัสและตัวถอดรหัสในชั้นต่อไป ตัวเข้ารหัสตัวแรกจะทำการรับข้อมูลตำแหน่ง (Positional Information) และ ฝังคำ (Word Embedding) ของลำดับชุดข้อมูลขาเข้าแทนการเข้ารหัส ซึ่งการรับข้อมูลตำแหน่งนี้สำคัญในทรานฟอร์เมอร์ เพราะไม่มีส่วนอื่นของทรานฟอร์เมอร์ที่สามารถทำงานในส่วนตรงนี้ได้

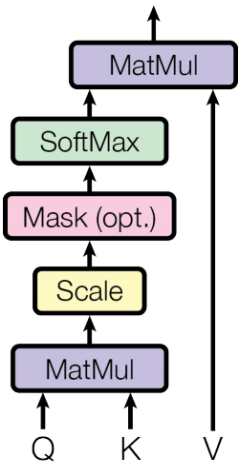
2.2.2 ตัวถอดรหัส (Decoder)

ตัวถอดรหัสแต่ละตัวนั้นประกอบไปด้วย 3 ส่วนหลักๆ ได้แก่ กลไกความสนใจตนเอง กลไกความสนใจหลายหัว บนข้อมูลขาออกจากตัวเข้ารหัส และโครงข่ายประสาทเทียมแบบป้อนไปด้านหน้า (Feed-forward Neural Network) การทำงานของตัวถอดรหัสนั้นจะมีลักษณะการทำงานคล้ายกับตัวเข้ารหัส แต่มีการเพิ่มเลเยอร์ความสนใจ (Attention) เพิ่มเติมเข้ามาซึ่งจะทำหน้าที่ดึงข้อมูลที่ได้มาจากตัวเข้ารหัส ตัวถอดรหัสตัวแรกนั้นจะทำงานเหมือนตัวเข้ารหัสโดยจะรับข้อมูลตำแหน่ง (Positional Information) และ ฝังคำ (Word Embedding) ของลำดับชุดข้อมูลขาออกแทนการเข้ารหัส เนื่องจากตัวถอดรหัสไม่ควรจะเห็นข้อมูลในปัจจุบันและข้อมูลในชุดถัดไป จึงมีการทำการปิดบังข้อมูล (Masking) เพื่อป้องกันการรั่วไหลของข้อมูล ในเลเยอร์สุดท้ายของตัวถอดรหัสจะเป็นการทำการแปลงเชิงเส้น (Linear Transformation) และซอฟต์แม็กซ์ (Softmax) เพื่อนำมาหาความน่าจะเป็นของข้อมูลขาออกแต่ละตัว

2.2.3 ความสนใจ (Attention)

ฟังก์ชันความสนใจ (Attention) คือการจับคู่ คิวรี่ (Query) ชุดข้อมูลของ คู่คีย์ - แวลยู (Key-Values) ไปยังเอาต์พุต (Output) โดยที่คิวรี่ (Query), คีย์ (Key), แวลยู (Value) และเอาต์พุต จะอยู่ในรูปแบบเวกเตอร์ (Vector) ข้อมูลเอาต์พุตคือการคำนวณโดยคิดจากผลรวมน้ำหนัก (Weight) ทั้งหมดของค่าแวลยู โดยค่าน้ำหนักแต่ละตัวที่กำหนดเข้ากับแวลยูจะถูกคำนวณโดยฟังก์ชันที่สอดคล้องกันกับคิวรี่ที่ความตรงกันกับคีย์

2.2.4 ความสนใจแบบผลคูณจุดปรับขนาด (Scaled Dot-Product Attention)



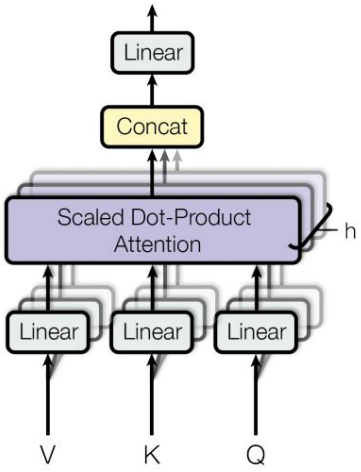

*ภาพที่ 3* ความสนใจแบบผลคูณจุดปรับขนาด (Scaled Dot-Product Attention) [3]



ส่วนประกอบพื้นฐานของทรานฟอร์เมอร์นั้นคือ ความสนใจแบบผลคูณจุดปรับขนาด (Scaled Dot-Product Attention) ที่มีลักษณะเฉพาะ โดยวิธีการคำนวณของฟังก์ชัน Attention นั้นจะคำนวณบนคิวรี่แบบพร้อม ๆ กัน และได้ออกมาเป็นเมทริกซ์ (Metric) Q ส่วนคีย์และแวลยูนั้นจะถูกเก็บไว้ในเมทริกซ์ K และ V โดยมีสูตรตามนี้

$$\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \qquad (1)$$

2.2.5 ความสนใจหลายหัว (Multi-Head Attention)




*ภาพที่ 4* ความสนใจหลายหัว (Multi-Head Attention) [3]

เมทริกซ์หนึ่งชุดของ $(W_Q, W_K, W_V)$ จะเรียกว่า Attention Head โดยแต่ละเลเยอร์ของทรานฟอร์เมอร์มี Attention Head หลายตัว โดยหนึ่ง Attention Head สนใจโทเค็น (Token) ที่มีความเกี่ยวข้องกันของแต่ละโทเค็น แต่ Multi-Head Attention นั้นสามารถเรียนรู้ถึงความแตกต่างกันของ “ความเกี่ยวข้อง”

**2.3 การถอดอักษร (Transliteration)**

การถอดอักษร (Transliteration) หมายถึง การนำคำในภาษาหนึ่งมาเขียนด้วยตัวอักษรอีกภาษาหนึ่งแบบอักษรต่ออักษร โดยพยายามใช้หน่วยเสียงของอักษรทั้งสองภาษาใกล้เคียงกันมากที่สุด [4] ตัวอย่างเช่น คำว่า “OXFORD” ในภาษาอังกฤษถอดอักษรเป็น “ออกซ์ฟอร์ด” ในภาษาไทย เป็นต้น การถอดอักษรแบ่งเป็น 3 ขั้นตอนหลัก ๆ ดังนี้

1. ถอดหน่วยอักษรในภาษาต้นแบบ (Source Language) เป็นหน่วยเสียงในภาษาต้นแบบ เช่น ถอดหน่วยอักษร “B” ในภาษาอังกฤษเป็นหน่วยเสียง /b/ ในภาษาอังกฤษ เป็นต้น
2. แทนหน่วยเสียงในภาษาต้นแบบ ด้วยหน่วยเสียงในภาษาเป้าหมาย (Target Language) โดยพยายามใช้หน่วยเสียงที่ใกล้เคียงกันมากที่สุด เช่น แทนหน่วยเสียง /b/ ในภาษาอังกฤษเป็นหน่วยเสียง /b/ ในภาษาไทย เป็นต้น
3. ถอดหน่วยเสียงในภาษาเป้าหมาย เป็นหน่วยอักษรในภาษาเป้าหมาย เช่น ถอดหน่วยเสียง /b/ ในภาษาไทยเป็นหน่วยอักษร “บ” ในภาษาไทย เป็นต้น

ปัญหาต่าง ๆ ในการถอดอักษรได้แก่

1. ความสัมพันธ์ของหน่วยอักษรและหน่วยเสียง มีความสัมพันธ์แบบหนึ่งตัวอักษรแทนหลายหน่วยเสียง เช่น ในภาษาอังกฤษ “C” แทนด้วย /k/ หรือ /s/ เป็นต้น และมีความสัมพันธ์แบบหลายหน่วยอักษรแทนหนึ่งหน่วยเสียง เช่น ในภาษาอังกฤษ “N, TN, GN, PN” แทนด้วย /n/ ในภาษาไทย “ร, ฤ, หร” แทนด้วย /r/ และ “ฉ, ช, ฌ” แทนด้วย /ch/ เป็นต้น
2. การแบ่งพยางค์ในภาษาต้นแบบ เมื่อมีพยัญชนะตัวเดียวอยู่ระหว่างสระ เช่น คำว่า money ในภาษาอังกฤษ จะแบ่งพยางค์อย่างไร จะถอดพยัญชนะซ้ำสองตัวเพื่อให้อ่านได้สะดวกเป็น มัน-นีย์ หรือจะถอดอักษรเพียงตัวเดียวตามที่ปรากฏในภาษาอังกฤษเป็น มะ-นีย์ หรือ มัน-อีย์

ปัญหาอันเนื่องมาจากช่วงเวลาของการยืมคำทับศัพท์ คำทับศัพท์บางคำยืมมาเป็นเวลานาน ซึ่งในอดีตมีหลักเกณฑ์การทับศัพท์ไม่ตรงกับหลักเกณฑ์ในปัจจุบัน เช่น “C” ที่แทน /k/ ในอดีตนิยมถอดเป็นอักษร “ก” เช่น กุ๊ก (Cook) กัปตัน (Captain) กะรัต(Carat) แก๊ป(Cap) เป็นต้น แต่ปัจจุบัน “C” ที่แทน /k/ มักจะถอดเป็น “ค” ในตำแหน่งพยัญชนะต้น เช่น คอนโดมิเนียม (Condominium) แคปซูล (Capsule) แคร็อต (Carrot) เป็นต้น

**2.4 การถ่ายเสียงด้วยตัวอักษรโรมัน (Romanization)**

การถ่ายเสียงด้วยตัวอักษรโรมัน (Romanization) คือการถ่ายเสียงตัวอักษรของภาษาอื่น ที่ไม่ใช่อักษรโรมัน เช่น ไทย จีน ญี่ปุ่น ฯลฯ ให้เป็นตัวอักษรโรมัน [4] เพื่อให้ผู้ที่ไม่รู้จักภาษานั้น ๆ สามารถอ่านออกเสียงได้ ทางราชบัณฑิตยสถานจึงได้กำหนดระบบการใช้ตัวอักษรโรมันเพื่อการถ่ายเสียงสำหรับตัวอักษรไทยออกเป็น 2 ระบบ คือระบบทั่วไปและระบบพิสดาร โดยระบบทั่วไปจะใช้สำหรับกรณีที่การออกเสียงสำคัญกว่าการเขียนตัวสะกด ซึ่งจะอาศัยหลักการออกเสียงเป็นสำคัญ ต้องสอดคล้องกับไวยากรณ์ของไทย และสามารถขยายเป็นระบบเฉพาะได้ เช่น คำว่า “กษัตริย์” ถ่ายเสียงเป็น Kasat ส่วนระบบพิสดารจะใช้ในกรณีที่จะแสดงตัวอักษรให้ละเอียดแม่นยำ เพื่อให้คงความหมายของคำนั้นไว้ เช่น คำว่า “กษัตริย์” ถ่ายเสียงเป็น Kasatriy

**2.5 การวัดผลการค้นคืน**

มาตรวัดที่ใช้วัดประสิทธิภาพของการค้นคืนได้แก่ ค่าแม่นยำ (Precision) ค่าเรียกคืน (Recall) [5] และตัววัด F1 (F1 Measurement) [6] ซึ่งมีวิธีคำนวณจากการนับจำนวนข้อมูลที่เกี่ยวข้อง (Relevant Data) และข้อมูลที่ระบบค้นคืนกลับมา (Retrieved Data) ได้ดังนี้

$$\text{ค่าแม่นยำ} = \frac{\text{จำนวนคำที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำที่คืนกลับมา}}$$

$$\text{ค่าเรียกคืน} = \frac{\text{จำนวนคำที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำที่เกี่ยวข้องทั้งหมด}}$$

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

**2.6 ขั้นตอนวิธีระยะการแก้ไขสั้นที่สุด (Minimum Edit Distance)**

ระยะแก้ไขสั้นที่สุด [7] เป็นวิธีการหนึ่งในการวัดความคล้ายคลึงกันระหว่าง 2 สายอักขระ ซึ่งจะทำการแทรก ลบ และการแทนที่อักขระ เพื่อเปลี่ยนอักขระหนึ่งไปเป็นอีกอักขระหนึ่งให้เหมือนกันด้วยชุดจำนวนคำสั่งที่น้อยที่สุด ตัวอย่างเช่น ระยะห่างของการแก้ไขให้ EXSAMBL เป็น EXAMPLE เท่ากับ 3 ซึ่งมีวิธีการคำนวณดังนี้



ตัวอย่างที่ 2.6.1 การคำนวณระยะห่างของการแก้ไขให้ EXSAMBL เป็น EXAMPLE

1. การลบตัวอักษร S EXSAMBL → EXAMBL
2. การแทนที่ตัวอักษร B ด้วย P EXAMBL → EXAMPL
3. การเพิ่มตัวอักษร E EXAMPL → EXAMPLE

ดังนั้นระยะห่างของการแก้ไขให้ EXSAMBL เป็น EXAMPLE มีค่าเท่ากับ 3

จากวิธีการคำนวณข้างต้นสามารถเขียนในอยู่ในรูปการคำนวณด้วยความสัมพันธ์เวียนเกิด 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 $w_k$

**2.7 ขั้นตอนวิธีซาวด์เด็กซ์ภาษาอังกฤษ (Soundex)**

ในปี ค.ศ. 1918 Robert C. Russell และ Margaret K. Odell ได้ออกแบบขั้นตอนวิธีการเข้ารหัสชื่อในภาษาอังกฤษ หรือที่เรียกว่า “ซาวด์เด็กซ์” (Soundex) [8] คือวิธีการเข้ารหัสชื่อในภาษาอังกฤษโดยยึดหลักการอ่านออกเสียง เพื่อให้ชื่อที่อ่านออกเสียงคล้ายกันได้รหัสเหมือนกัน โดยรหัสของซาวด์เด็กซ์นั้นจะมีความยาวคงที่ 4 ตัว ซึ่งรหัสเหล่านี้สามารถใช้ในการเปรียบเทียบคำสองคำที่ออกเสียงคล้าย ๆ กันได้ ขั้นตอนวิธีดังกล่าวได้ใช้แนวคิดทางภาษาศาสตร์และชวเลขที่ว่าชื่อในภาษาอังกฤษสามารถจำแนกความแตกต่างได้โดยพิจารณาเพียงพยัญชนะเท่านั้น

ขั้นตอนการเข้ารหัสซาวด์เด็กซ์ มีหลักเกณฑ์ที่สำคัญ ดังนี้

- ตัวอักษรตัวแรกของคำถูกนำไปเป็นรหัส
- ตัวอักษรที่เหลือจะถูกแปลงเป็นตัวเลข โดยใช้ตารางการกำหนดรหัสซาวด์เด็กซ์ ดังแสดงในตารางที่ 1 โดยพิจารณาเงื่อนไขเพิ่มเติมดังนี้

- ถ้าตัวอักษรใดที่มีรหัสเท่ากับศูนย์จะถูกตัดออกไป
- ถ้าตัวอักษรใดที่มีรหัสตัวเลขเหมือนกันอยู่ติดกันจะเก็บเพียงหนึ่งรหัสเท่านั้น
- ทำการแปลงไปเรื่อย ๆ จนได้รหัสที่ตรงตามรูปแบบที่ต้องการ ก็คือ 1 ตัวอักษร และ 3 ตัวเลข ถ้าเกิดกรณีที่เลขยังไม่ครบ 3 ตัว ให้เติมเลข 0 เข้าไปจนครบ

ตัวอย่างการเข้ารหัสซาวด์เด็กซ์ คำว่า “ROBERT” จะได้รหัสซาวด์เด็กซ์เท่ากับ R163 เป็นต้น

*ตารางที่ 1 การกำหนดรหัสซาวด์เด็กซ์ภาษาอังกฤษของ Odell และ Russel*

| ตัวอักษร | รหัสตัวเลข |
|---|---|
| A E I O U H W Y | 0 |
| B F P V | 1 |
| C G J K Q S X Z | 2 |
| D T | 3 |
| L | 4 |
| M N | 5 |
| R | 6 |

# บทที่ 3
# งานวิจัยที่เกี่ยวข้อง

## 3.1 งานวิจัยของ วรรณี อุดมพาณิชย์

งานวิจัย [9] ได้นำเสนอหลักเกณฑ์ที่สำคัญที่จะพิจารณาใช้ในการสร้างรหัสคำภาษาไทย ดังนี้

- ความยาวของรหัสจะมีจำนวน 7 หลัก
- รหัสตัวแรก เป็นพยัญชนะตัวแรกที่ปรากฏในชื่อ ซึ่งใช้ตารางเทียบรหัสในตารางที่ 2
- พยัญชนะอื่นที่ตามมาจะเปลี่ยนเป็นตัวเลขซึ่งใช้ตารางเทียบรหัสในตารางที่ 3
- ไม่ใช้สระ วรรณยุกต์ และไม้ไต่คู้ มาสร้างรหัส ยกเว้นสระที่ให้เสียงตัวสะกดเป็นพยัญชนะ ได้แก่ ไ- ใ- –ำ
- กรณีที่พบ ใ- ไ- ไ-ย และ –ัย จะเปลี่ยนให้อยู่ในรูปแบบเดียวกันคือ –ัย ก่อนทำการเข้ารหัส เนื่องจากสระดังกล่าวอ่านออกเสียงเหมือนกัน เช่น ใท ไท ไทย และทัย จะเปลี่ยนเป็น ทัย
- กรณีที่พบ รร จะเปลี่ยนเป็น –ัน ในกรณีที่ไม่มีตัวสะกดตามหลัง แต่ถ้าหลัง รร มีตัวสะกดก็จะใช้แทนด้วย – ั เปลี่ยนคำว่า สรรเพชร รังสรรค์ พรรณนที และธรรมรัตน์ เป็น สันเพชร รังสัน พัณนที และ ธัมรัตน์ ตามลำดับ
- กรณีที่มีตัวการันต์ ให้ตัดพยัญชนะที่มีการันต์กำกับ รวมทั้งสระและอักษรที่ควบการันต์ทิ้ง
- ถ้ารหัสที่ได้มีจำนวนน้อยกว่า 7 หลัก ให้เติม 0 จนครบ



ตัวอย่างการเข้ารหัสคำ เช่น คำ “อัมพร” หรือ “อำภรณ์” จะได้รหัสเป็น “อ059000” คำ “พรรณศักดิ์” หรือ “พันธุ์ศักดิ์” จะได้รหัสเป็น “พ341000” คำ “เนืองนิตย์” หรือ “เนืองนิจ” จะได้รหัสเป็น “น623400”

*ตารางที่ 2 การกำหนดรหัสตัวอักษรของรหัสซาวด์เด็กซ์ภาษาไทย*

*จากงานวิจัยของวรรณี อุดมพาณิชย์*

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

*ตารางที่ 3 การกำหนดรหัสตัวเลขของรหัสซาวด์เด็กซ์ภาษาไทย*

*จากงานวิจัยของของวรรณี อุดมพาณิชย์*

| รหัสตัวเลข | ตัวอักษร |
|---|---|
| 0 | ม ว ำ า |
| 1 | ก ข ฃ ค ฅ ฆ |
| 2 | ง ย |
| 3 | ญ ณ น |
| 4 | ฎ ฏ ด ต ศ ษ ส |
| 5 | บ ป พ ภ |
| 6 | ผ ฝ ฟ ห อ ฮ |
| 7 | จ ฉ ช ซ ฌ |
| 8 | ฐ ฑ ฒ ฌ ถ ท ธ |
| 9 | ร ฤ ล ฦ ฬ |

**3.2 งานวิจัยของ ประยุทธ สุวรรณวิสารท และ สมชาย ประสิทธิ์จูตระกูล**

งานวิจัยนี้ได้นำเสนอขั้นตอนการค้นคืนข้ามภาษาโดยสามารถแบ่งออกได้เป็น 2 ส่วนหลัก ๆ คือ การค้นคืนข้ามภาษาแบบภาษาไทยทับศัพท์ภาษาอังกฤษ [10] และ การค้นคืนข้ามภาษาแบบภาษาอังกฤษทับศัพท์ภาษาไทย [11]

โดยการค้นคืนข้ามภาษาแบบภาษาไทยทับศัพท์ภาษาอังกฤษ เสนอการเข้ารหัสโดยดัดแปลงต้นแบบการเข้ารหัสคำซาวด์เด็กซ์ของ Odell และ Russell ดังตารางที่ 4 แสดงการกำหนดรหัสคำไทยทับศัพท์คำอังกฤษ รหัสคำเป็นตัวเลขทั้งหมดโดยไม่จำกัดความยาวของรหัสคำ สำหรับการค้นคืนนั้นใช้การเปรียบเทียบรหัสคำแบบเหมือนกันทุกประการ นั่นคือรหัสคำของทั้งสองตัวต้องเหมือนกันทุกตัวจึงจะถือว่ามีเสียงอ่านที่ตรงกัน จากผลการทดลองพบว่า ได้ค่าแม่นยำ 78% และค่าเรียกคืน 90% เมื่อใช้รหัสคำความยาวมากกว่า 4 หลักขึ้นไป หรือใช้คำทับศัพท์ที่มีความยาวมากกว่า 7 ตัวขึ้นไป

สำหรับกรณีการค้นคืนข้ามภาษาแบบภาษาอังกฤษทับศัพท์ภาษาไทย จะมีขั้นตอนการทำงานแบ่งออกเป็น 2 ขั้นตอน คือ ขั้นตอนการเข้ารหัสคำและขั้นตอนวิธีการเปรียบเทียบรหัสคำสำหรับการค้นคืน โดยการเข้ารหัสคำสำหรับคำไทยจะมีการประมวลผลเบื้องต้นก่อนการเข้ารหัสคำ ซึ่งได้แก่ การลดรูป การตัดวรรณยุกต์ และไม้ไต่คู้ การตัดการันต์และอักษรควบการันต์ การเปลี่ยนรูป รร เป็น ัน การเปลี่ยนสระ ไ- ใ- ไ-ย -ย เป็น ัย และเปลี่ยนสระ ำ เป็น ัม เป็นต้น ขั้นตอนสุดท้ายหลังจากการประมวลผลเบื้องต้นแล้ว จะมีการย้ายตำแหน่งสระไปด้านหลังสุดของคำตามลำดับ รหัสคำนั้นจะเป็นตัวอักษรไทยผสมกับสัญลักษณ์เสียงสากล ดังแสดงในตารางที่ 5 และตารางที่ 6 ในส่วนของการเปรียบเทียบรหัสคำสำหรับการค้นคืนนั้นใช้การคำนวณหาค่าความแตกต่างของรหัสคำด้วยเทคนิคระยะแก้ไขสั้นที่สุด และมีการกำหนดต้นทุนในการแทนที่อักขระสำหรับแต่ละคู่อักขระตามกฎเกณฑ์ที่ได้สร้างขึ้น ค่าของต้นทุนมี 4 ระดับคือ C1 C2 C3 และ C4 ซึ่งมีค่า 0 1 4 และ 7 ตามลำดับ ดังตัวอย่างในตารางที่ 7 และตารางที่ 8 จากผลการทดลองพบว่าได้ค่าแม่นยำ 69% และค่าเรียกคืน 73%

*ตารางที่ 4 การกำหนดรหัสสำหรับคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ จากงานวิจัยของ ประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล*

| ภาษาอังกฤษ | ภาษาไทย | รหัส |
|---|---|---|
| A E I O U H W Y [2] | อ ห ฮ ย ญ | 0 |
| 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 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 [1] | อ | 7 |
| H [1] | ห ฮ | 8 |
| W [1] | ว | 1 |
| Y [1] | ย ญ | 9 |
| | ง | 52 |

1 : สำหรับตัวอักษรแรก (ซ้ายสุด) ของคำเท่านั้น

2 : สำหรับตัวอักษรตั้งแต่ตัวที่ 2 เป็นต้นไปเท่านั้น



*ตารางที่ 5 การกำหนดรหัสของพยัญชนะสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัยของประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล*

| อักษรอังกฤษ | อักษรไทย | รหัส |
|---|---|---|
| b | บ | บ |
| bh | พ | พ |
| c | ซ | ซ |
| ch | ช ฉ ฌ | ช |
| ck | ก | ก |
| d | ด ฎ | ด |

| อักษรอังกฤษ | อักษรไทย | รหัส |
|---|---|---|
| n | น ณ | น |
| ng | ง | ง |
| p | ป | ป |
| ph | พ ผ ภ | พ |
| q | ค | ค |
| r | ร ฤ | ร |

*ตารางที่ 5 การกำหนดรหัสของพยัญชนะสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัยของประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล (ต่อ)*

| อักษรอังกฤษ | อักษรไทย | รหัส |
|---|---|---|
| dh | ท | ท |
| f | ฟ ฝ | ฟ |
| g | ก | ก |
| h | ห ฮ | ห |
| j | จ | จ |
| k | ก | ก |
| kh | ข ฃ ค ฅ ฆ | ข |
| l | ล ฦ ฬ | ล |
| m | ม | ม |

| อักษรอังกฤษ | อักษรไทย | รหัส |
|---|---|---|
| s | ส ซ ศ ษ | ส |
| t | ต ฏ | ต |
| th | ท ฐ ฑ ฒ ถ ธ | ท |
| v | ว | ว |
| w | ว | ว |
| x | ก | ก |
| y | ย ญ | ย |
| z | ซ | ซ |
| | | |

*ตารางที่ 6 การกำหนดรหัสของสระสำหรับคำไทยและคำอังกฤษทับศัพท์คำไทย จากงานวิจัยของ ประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล*

| ตัวอักษรอังกฤษ | ตัวอักษรไทย | รหัส |
|---|---|---|
| -a | ะ | ะ |
| -aa | า | า |
| -ae | แ-ะ แ- | x |
| -ai | -ัย | -ัย |
| -ao | เ-า | @ |
| -aiu | เ-ีย | I |
| -arn | าน | าน |
| -art | าท | าท |
| -e | เ-ะ เ- | เ |
| -ee | -ี | -ี |
| -eo | แ-ว | แ-ว |
| -er | เ-อ เ-ิ | q |

| ตัวอักษรอังกฤษ | ตัวอักษรไทย | รหัส |
|---|---|---|
| -eu | -ึ | -ึ |
| -i | -ิ | -ิ |
| -ia | เ-ียะ เ-ีย | I |
| -ie | เ-ียะ เ-ีย | I |
| -o | -อ | อ |
| -oe | เ-อ เ- | q |
| -oi | อย | อย |
| -oo | -ู | -ู |
| -orn | -อน | ร |
| -u | -ึ -ี -ุ -ู | -ุ |
| -ua | เ-ือะ เ-ือ -ัวะ -ัว | U |
| -ue | -ึ | -ึ |

*ตารางที่ 7 ตัวอย่างการกำหนดต้นทุนการแทนที่อักขระสำหรับพยัญชนะ จากงานวิจัยของประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล*

| | ก | ข | ค | ... | ท | ธ | น | บ | ... |
|---|---|---|---|---|---|---|---|---|---|
| ก | $C_1$ | $C_2$ | $C_2$ | ... | $C_4$ | $C_4$ | $C_4$ | $C_4$ | ... |
| ข | $C_2$ | $C_1$ | $C_1$ | ... | $C_4$ | $C_4$ | $C_4$ | $C_4$ | ... |
| ค | $C_2$ | $C_1$ | $C_1$ | ... | $C_4$ | $C_4$ | $C_4$ | $C_4$ | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ท | $C_4$ | $C_4$ | $C_4$ | ... | $C_1$ | $C_1$ | $C_4$ | $C_4$ | ... |
| ธ | $C_4$ | $C_4$ | $C_4$ | ... | $C_1$ | $C_1$ | $C_4$ | $C_4$ | ... |
| น | $C_4$ | $C_4$ | $C_4$ | ... | $C_4$ | $C_4$ | $C_1$ | $C_4$ | ... |
| บ | $C_4$ | $C_4$ | $C_4$ | ... | $C_4$ | $C_4$ | $C_4$ | $C_1$ | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

*ตารางที่ 8 ตัวอย่างการกำหนดต้นทุนการแทนที่อักขระสำหรับสระ จากงานวิจัยของประยุทธ สุวรรณวิสารท และสมชาย ประสิทธิ์จูตระกูล*

| | ะ | ั | า | ิ | ี | ึ | ื | ุ | ... |
|---|---|---|---|---|---|---|---|---|---|
| ะ | $C_1$ | $C_1$ | $C_1$ | $C_4$ | $C_4$ | $C_4$ | $C_4$ | $C_4$ | ... |
| ั | $C_1$ | $C_1$ | $C_2$ | $C_4$ | $C_4$ | $C_4$ | $C_4$ | $C_2$ | ... |
| า | $C_1$ | $C_2$ | $C_1$ | $C_4$ | $C_4$ | $C_4$ | $C_4$ | $C_4$ | ... |
| ิ | $C_4$ | $C_4$ | $C_4$ | $C_1$ | $C_1$ | $C_3$ | $C_4$ | $C_4$ | ... |
| ี | $C_4$ | $C_4$ | $C_4$ | $C_1$ | $C_1$ | $C_4$ | $C_4$ | $C_4$ | ... |
| ึ | $C_4$ | $C_4$ | $C_4$ | $C_3$ | $C_4$ | $C_1$ | $C_1$ | $C_2$ | ... |
| ื | $C_4$ | $C_4$ | $C_4$ | $C_4$ | $C_4$ | $C_1$ | $C_1$ | $C_2$ | ... |
| ุ | $C_4$ | $C_2$ | $C_4$ | $C_4$ | $C_4$ | $C_2$ | $C_2$ | $C_1$ | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

**3.3 งานวิจัยของ ทัศนวรรณ ศูนย์กลาง สมชาย ประสิทธิ์จูตระกูล และบุญเสริม กิจศิริกุล**

งานวิจัย [12] ได้นำเสนอขั้นตอนวิธีการเข้ารหัสคำทับศัพท์ภาษาไทย/ภาษาอังกฤษโดยใช้เทคนิคนิวรอลเน็ตเวิร์กแบบแพร่กระจายย้อนกลับ (Backpropagation Neural Network) ซึ่งจะมีนิวรอลเน็ตเวิร์กในการสร้างรหัสคำอ่านทั้งหมด 4 ชุด สำหรับ (1) คำไทย (2) คำอังกฤษทับศัพท์คำไทย และ (3) คำอังกฤษ (4) คำไทยทับศัพท์คำอังกฤษ งานวิจัยนี้ยังคงมีการใช้การประมวลผลเบื้องต้นด้วยเช่นเดียวกับในงานวิจัย [11] ในชั้นอินพุต จะพิจารณาตัวอักษรครั้งละ 9 ตัวอักษร โดยตำแหน่งตรงกลางจะเป็นตัวที่ถูกพิจารณา ส่วนตัวอักษร 4 ตัวหน้าและ 4 ตัวหลังจะใช้ประกอบเพื่อช่วยในการพิจารณา ในส่วนของชั้นเอาท์พุตจะได้รหัสเสียงจากข้อมูลขาเข้า ดังแสดงในภาพที่ 5 และภาพที่ 6 ในส่วนของการเปรียบเทียบรหัสคำสำหรับการค้นคืนนั้นใช้การคำนวณหาค่าความแตกต่างของรหัสคำด้วยเทคนิคระยะแก้ไขสั้นที่สุด จากผลการทดลองพบว่าเมื่อกำหนดเกณฑ์ในการยอมรับค่าความแตกต่างของรหัสคำมีค่าเท่ากับ 1 จะได้ค่าแม่นยำ 87.28% และค่าเรียกคืน 77.19% สำหรับกรณีคำไทยทับศัพท์คำอังกฤษ และได้ค่าแม่นยำ 96.34% และค่าเรียกคืน 75.15% สำหรับกรณีคำอังกฤษทับศัพท์คำไทย

| ลำดับตัวอักษร | รหัส |
|---|---|
| _ , _ , _ , _ , 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 , ◌ิ , _ , _ , _ , _ | _ |

*ภาพที่ 5 ตัวอย่างการเข้ารหัสคำไทยจากงานวิจัยของทัศนวรรณ ศูนย์กลาง และคณะ*

| ลำดับตัวอักษร | รหัส |
|---|---|
| _ , _ , _ , _ , 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 , _ , _ | l |
| H , O , D , I , U , M , _ , _ , _ | _ |
| O , D , I , U , M , _ , _ , _ , _ | m |

*ภาพที่ 6 ตัวอย่างการเข้ารหัสคำอังกฤษจากงานวิจัยของทัศนวรรณ ศูนย์กลาง และคณะ*

**3.4 งานวิจัยของ ศิริพจน์ สุรบถโสภณ และ บุญเสริม กิจศิริกุล**

งานวิจัย [13] ได้นำเสนอการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์กแบบแพร่กระจายย้อนกลับแบบซ้ายไปขวาในการจำลองรหัสคำ และใช้ขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithm) เพื่อเพิ่มประสิทธิภาพของการค้นคืนให้มากขึ้น และการเข้ารหัสคำของคำไทยยังใช้การประมวลผลเบื้องต้นด้วยเช่นเดียวกับในงานวิจัย [12] แต่มีการแก้ไขเพิ่มเติม และในการค้นคืนใช้การคำนวณหาค่าความแตกต่างของรหัสคำด้วยเทคนิคระยะแก้ไขสั้นที่สุด จากผลการทดลองของงานวิจัยนี้พบว่าการใช้นิวรอลเน็ตเวิร์กร่วมกับขั้นตอนวิธีเชิงพันธุกรรมในการหาตารางต้นทุนการแก้ไขอักขระให้ผลการค้นคืนได้ดีที่สุด เมื่อเทียบกับวิธีการของประยุทธ โดยผลที่ดีที่สุดของการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์กจะมีค่าแม่นยำ 96.35% และค่าเรียกคืน 87.30% สำหรับกรณีคำไทยทับศัพท์คำอังกฤษ และได้ค่าแม่นยำ 97.22% และค่าเรียกคืน 94.35% สำหรับกรณีคำอังกฤษทับศัพท์คำไทย

**3.5 งานวิจัยของ S. Tasanaprasert, W. Pokasame, S. Rattanaliam**

งานวิจัย [14] ได้นำเสนอการเข้ารหัสโดยดัดแปลงต้นแบบการเข้ารหัสคำซาวด์เด็กซ์ของ Odell และ Russell โดยมุ่งเน้นในการแปลงรหัสคำไทยทับศัพท์ภาษาอังกฤษ ซึ่งรหัสคำจะมีความยาว 9 ตัวอักษร โดยรหัสตัวแรกหรือตัวที่สอง เป็นพยัญชนะตัวแรกที่ปรากฏในชื่อ ซึ่งใช้ตารางเทียบรหัสในตารางที่ 9 ในส่วนของพยัญชนะอื่นที่ตามมาจะเปลี่ยนเป็นตัวเลขซึ่งใช้ตารางเทียบรหัสในตารางที่ 10 และในกรณีที่ตัวอักษรนั้นเป็นสระจะใช้ตารางเทียบรหัสในตารางที่ 11 ถ้ารหัสที่ได้มีจำนวนน้อยกว่า 9 หลัก ให้เติม 0 จนครบ จากผลการทดลองพบว่ามีค่าความแม่นยำประมาณ 90% จากคำศัพท์ทั้งหมด 7,240 คำ

ตัวอย่างการเข้ารหัสคำ เช่น คำ “thai” หรือ “ไทย” จะได้รหัสเป็น “T97105000” คำ “sab” หรือ “ทราบ” จะได้รหัสเป็น “S97100000”

*ตารางที่ 9 การกำหนดรหัสสำหรับอักขระตัวแรก*

| ตัวอักษรอังกฤษ | ตัวอักษรไทย | รหัส |
|---|---|---|
| A E I O U | อ | A E I O U |
| B | บ | B |
| F V | ฝ ฟ | F |
| PH P | ป ผ พ ภ | P |
| K KH C G | ก ข ฃ ค ฅ ฆ | K |
| C CH | จ ฉ ช ฌ | C |
| S | ซ ศ ษ ส ทร | S |
| D | ฎ ด | D |
| T | ฏ ต ฐ ฑ ฒ ถ ท ธ | T |
| L | ล ฬ | L |
| M | ม | M |
| N | ณ น | N |
| R | ร | R |
| NG | ง | NG |
| H | ห ฮ | H |
| Y | ญ ย | Y |
| W | ว | W |

*ตารางที่ 10 การกำหนดรหัสสำหรับอักขระถัดจากตัวแรก ในกรณีที่เป็นพยัญชนะ*

| ตัวอักษรอังกฤษ | ตัวอักษรไทย | รหัส |
|---|---|---|
| B F P V PH | บ ฝ ฟ ป ผ พ ภ | 1 |
| K C G KH CH S | ก ข ฃ ค ฅ ฆ จ ฉ ช ฌ ซ ศ ษ ส | 2 |
| D T | จ ฉ ช ฌ ซ ศ ษ ส ฎ ด ฏ ต ฐ ฑ ฒ ถ ท ธ | 3 |
| L | ล ฬ | 4 |
| M | ม | @ |
| N | ณ น ร ญ ล ฬ | 5 |
| R | ร | 6 |
| H | ห ฮ | 7 |
| W | ว | 8 |
| Y | ญ ย | 9 |
| NG | ง | 52 |

*ตารางที่ 11 การกำหนดรหัสสำหรับอักขระถัดจากตัวแรก ในกรณีที่เป็นสระ*

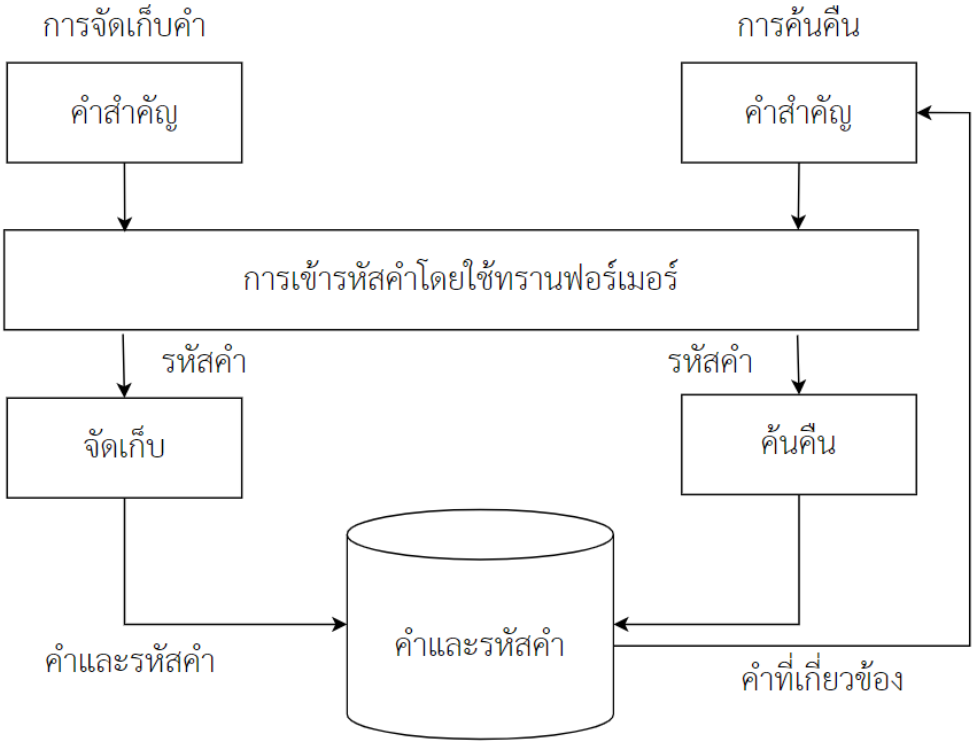
| ตัวอักษรอังกฤษ | ตัวอักษรไทย | รหัส |
|---|---|---|
| อะ, อา, อั | A | 97 |
| อัวะ, อัว | UA | 11797 |
| อำ | MA | 9764 |
| อิ, อี | I | 105 |
| อุ, อู | U | 117 |
| อึ, อื | UE | 11769 |
| เอย | OEI | 111101105 |
| เอะ, เอ็, เอ | E | 101 |
| เออะ, เออ, เอิ | OE | 111101 |
| เอา | AO | 97111 |
| เอียะ, เอีย | IA | 10597 |

*ตารางที่ 11 การกำหนดรหัสสำหรับอักขระถัดจากตัวแรก ในกรณีที่เป็นสระ (ต่อ)*

| ตัวอักษรอังกฤษ | ตัวอักษรไทย | รหัส |
|---|---|---|
| เอือะ, เอือ | UEA | 11710197 |
| แอะ, แอ็, แอ | AE | 97101 |
| โอะ, โอ | O | 111 |
| เอาะ, อ็, ออ | O | 111 |
| ใอ, ไอ, ไอย | AI | 97105 |
| เอือย | UEAI | 11710197105 |
| -วย | UAI | 11797105 |
| ิว | IO | 105111 |
| เ็ว, เ-ว | EO | 101111 |
| เอียว | IAO | 10597111 |

# บทที่ 4
# การเข้ารหัสคำ

การเข้ารหัสคำนั้นได้ถูกนำมาใช้ในการแก้ปัญหาการค้นคืนข้ามภาษาโดยไม่อาศัยพจนานุกรม โดยการนำรหัสคำที่ได้ทำการเข้ารหัสคำไว้แล้ว มาทำการสร้างดัชนีและจัดเก็บไว้ในฐานข้อมูล จากนั้นเมื่อต้องการที่จะทำการสืบค้นข้อมูล ก็จะนำข้อมูลที่ต้องการสืบค้นมาทำการสร้างรหัสคำ และนำรหัสคำที่ได้ไปสืบค้นในฐานข้อมูลที่ได้ทำดัชนีรหัสคำไว้แล้ว ในบทนี้จะกล่าวถึงวิธีการเข้ารหัสคำ การประมวลผลเบื้องต้น และการเข้ารหัสคำ




*ภาพที่ 7 การเข้ารหัสคำโดยใช้ทรานฟอร์เมอร์*

## 4.1 รหัสคำ

รหัสคำคือสัญลักษณ์แทนเสียงอ่านของคำ ดังนั้นคำที่มีเสียงอ่านตรงกัน จะมีรหัสคำที่เหมือนกัน หรือใกล้เคียงกัน รหัสคำนั้นจะประกอบไปด้วยรหัสเสียงของแต่ละตัวอักขระที่เรียงต่อกัน ยกตัวอย่างเช่นคำว่า “ชุมพร” และ “Chumporn” จะได้รหัสคำที่เหมือนกัน เพราะเสียงอ่านของทั้งสองคำนั้นตรงกัน ในงานวิจัยนี้ใช้หลักเกณฑ์การออกเสียงทั้งภาษาไทยและภาษาอังกฤษของราชบัณฑิตยสถาน [15, 16] มาสร้างเป็นตารางรหัสเสียงดังแสดงในตารางที่ 12 ซึ่งเป็นกรณีคำไทย

ทับศัพท์คำอังกฤษจะใช้หน่วยเสียงภาษาอังกฤษเป็นหลัก แล้วนำหน่วยเสียงภาษาไทยไปเทียบเพื่อหากลุ่มเสียงที่ตรงกันหรือใกล้เคียงกัน ในทางกลับกันดังแสดงในตารางที่ 13 จะเป็นกรณีคำอังกฤษทับศัพท์คำไทยจะใช้หน่วยเสียงภาษาไทยเป็นหลัก แล้วนำหน่วยเสียงภาษาอังกฤษไปเทียบ ทั้งสองตารางนี้ได้มาจากงานวิจัย [13] โดยมีการแก้ไขเพิ่มเติม

*ตารางที่ 12 รหัสเสียงสำหรับคำไทยทับศัพท์คำอังกฤษ*

| เสียงพยัญชนะ | | รหัสเสียง | เสียงสระ | | รหัสเสียง |
|---|---|---|---|---|---|
| ไทย | อังกฤษ | | ไทย | อังกฤษ | |
| พ, ป | p | p | ิ ี | ee, ei, ea, ey, i | i |
| บ | b | b | เ | e, ay | e |
| ท,ต,ธ | 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, ั -า | a | a |
| ฟ | f, ph | f | โ , -อ | 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 | | | |

*ตารางที่ 13 รหัสเสียงสำหรับคำอังกฤษทับศัพท์คำไทย*

| 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 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 เ็ | er, oe | W |
| ณ น หน | n | n | n | เ็ ียะ เ็ ีย | ia, ie, aiu | I |
| บ | b | b | p | เ็ อะ เ็ 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 | – | เ็ อย –วย | uai, uay, ou | O |

*ตารางที่ 13 รหัสเสียงสำหรับคำอังกฤษทับศัพท์คำไทย (ต่อ)*

| เสียงพยัญชนะ | | รหัสเสียง | | เสียงสระ | | รหัสเสียง |
|---|---|---|---|---|---|---|
| ไทย | อังกฤษ | ตัวต้น | ตัวสะกด | ไทย | อังกฤษ | ตัวต้น |
| | | | | แ-ว | aeo, eo, aew | $ |
| | | | | เ ี- ยว | ieo, eaw, eo, ew, iow, iau, iew, iaw | @ |

**4.2 การประมวลผลเบื้องต้น**

คำภาษาไทยนั้นมีความซับซ้อน จึงมีความจำเป็นที่จะต้องนำคำที่ต้องการเข้ารหัสมาผ่านการประมวลผลตัวอักษรเบื้องต้นก่อน เพื่อช่วยในการลดความซับซ้อนในการเข้ารหัสคำ โดยยึดหลักการตาม [9] ดังนี้

1. ตัดวรรณยุกต์และไม้ไต่คู้
2. เปลี่ยน รร เป็น –ัน หรือ –
3. เปลี่ยน ไ- ใ- ไ-ย เป็น -ัย
4. เปลี่ยน –ำ เป็น –ัม
5. ตัดตัวการันต์ และอักษรควบตัวการันต์
6. เปลี่ยน ฤ เป็น รึ ริ หรือ เรอ และเปลี่ยน ฤๅ เป็น รือ โดยตัว ฤ ต้องพิจารณาดังนี้ [17]

    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 สฤษฎ์

    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” [17] เช่น หรู ไหล หวี เหงา หญิง หนา หมู

**4.3 การเข้ารหัสคำ**

การเข้ารหัสคำ หมายถึง การแปลงตัวอักษรแต่ละตัวในคำไปเป็นรหัสเสียงโดยการเทียบเสียงจากตารางรหัสเสียงที่ได้ออกแบบไว้ จากนั้นนำรหัสเสียงที่ได้มาเรียงต่อกันจนเป็นรหัสคำ แต่จะมีอยู่ประเด็นหนึ่งในการหารหัสเสียงนั่นก็คือความสัมพันธ์ระหว่างตัวอักษรและรหัสเสียง สำหรับความสัมพันธ์แบบหนึ่งต่อหนึ่งนั่นจะสามารถเทียบรหัสเสียงได้โดยตรงจากตาราง แต่ในกรณีที่เป็นความสัมพันธ์แบบหนึ่งต่อหลายรหัสเสียง จะใช้วิธีการพิจารณา ดังนี้

หลักเกณฑ์สำหรับคำภาษาอังกฤษ

1. แบบหลายตัวอักษรต่อหนึ่งหน่วยเสียง ในกรณีนี้จะมีตัวอักษรตัวเดียวที่ให้รหัสเสียง ตัวอย่างเช่น คำภาษาอังกฤษ “young” จะมีรหัสเป็น yag โดยรหัสเสียง g เกิดจากกลุ่มตัวอักษร “ng”
2. แบบหลายหน่วยเสียงต่อหนึ่งตัวอักษร ในบางกรณีสำหรับคำอังกฤษ ตัวอักษรตัวหนึ่งในคำอาจให้เสียงมากกว่า 1 เสียง เช่น ตัว “x” ใน “boxer” ให้ทั้งเสียง /k/ และ /s/ ในกรณีนี้จะเลือกเพียงเสียงเดียวโดยให้มีรหัสเป็น bosW

หลักเกณฑ์สำหรับคำภาษาไทย

1. แบบหลายตัวอักษรต่อหนึ่งหน่วยเสียง ในกรณีนี้จะมีตัวอักษรตัวเดียวที่ให้รหัสเสียง ตัวอย่างเช่น คำไทย “เตรียม” จะมีรหัสเป็น trlm โดยรหัสเสียง l เกิดจากกลุ่มตัวอักษร “เ–ีย”
2. สระอะลดรูป ในกรณีนี้จะเพิ่มรหัสเสียง /a/ สำหรับสระอะลดรูป เช่น คำว่า “ขจิต” ตัวอักษร “ข” ให้เสียง /k/ และเสียง /a/ สำหรับสระอะลดรูป จะมีรหัสเป็น kacit
3. สระโอะลดรูป ในกรณีนี้จะเพิ่มเสียง /o/ สำหรับสระโอะลดรูป เช่นคำว่า “มง” มีเสียงสระโอะลดรูป จะมีรหัสเป็น mog
4. สระออลดรูป ในกรณีนี้จะเพิ่มเสียง /o/ สำหรับสระออลดรูป เช่นคำว่า “ถาวร” ตัวอักษร “ว” จะมีสระออ จะมีรหัสเป็น tavon

**ตัวอย่างการเข้ารหัสคำ 1** คำว่า “ลลิดา” ซึ่งเป็นคำไทย อ่านว่า ละ-ลิ-ดา จะได้รหัส lalida

**ตัวอย่างการเข้ารหัสคำ 2** คำว่า “kemthong” ซึ่งเป็นคำอังกฤษทับศัพท์คำไทย อ่านว่า เข็ม-ทอง จะได้รหัส kemtog

**ตัวอย่างการเข้ารหัสคำ 3** คำว่า “vector” ซึ่งเป็นคำอังกฤษ อ่านว่า เว็ก-เตอ จะได้รหัส vektW

**ตัวอย่างการเข้ารหัสคำ 4** คำว่า “ยูเรเนียม” ซึ่งเป็นคำไทยทับศัพท์คำอังกฤษ อ่านว่า ยู-เร-เนียม จะได้รหัส urenlm

# บทที่ 5
# การค้นคืนข้ามภาษา

วิธีการค้นคืนข้ามภาษานั้นจะเป็นการนำรหัสคำของคำที่เป็นคำถาม ไปเปรียบเทียบกับรหัสคำในดัชนีคำ ถ้าผลการเปรียบเทียบระหว่างรหัสคำสองคำอยู่ในเกณฑ์ที่กำหนด จะถือว่ารหัสคำทั้งสองมีเสียงอ่านที่ตรงกัน และระบบจะคืนคำนั้นออกมา ในบทนี้จะกล่าวถึงวิธีการคำนวณความต่างของรหัสคำ และ เกณฑ์การประเมินการค้นคืน

## 5.1 การคำนวณความต่างของรหัสคำ

เนื่องจากคู่คำของทั้งสองภาษาที่อ่านออกเสียงเหมือนกัน แต่ก็อาจจะมีรหัสคำที่ไม่เหมือนกันทุกตัวอักษร แต่จะมีลักษณะที่คล้ายกัน ดังนั้นในการเปรียบเทียบรหัสคำ จึงต้องใช้วิธีการเปรียบเทียบแบบประมาณ (Approximate Matching) โดยอาศัยการคำนวณความแตกต่าง (Distance) ของรหัสคำด้วยเทคนิคระยะการแก้ไขสั้นที่สุด (Minimum Edit Distance) ซึ่งเป็นวิธีการหนึ่งในการวัดความคล้ายคลึงกันระหว่าง 2 สายอักขระ ที่จะทำการแทรก ลบ และการแทนที่อักขระ เพื่อเปลี่ยนอักขระหนึ่งไปเป็นอีกอักขระหนึ่งให้เหมือนกัน ด้วยชุดจำนวนคำสั่งที่น้อยที่สุด

สามารถเขียนในอยู่ในรูปการคำนวณด้วยความสัมพันธ์เวียนเกิด 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 $w_k$

**ตัวอย่างที่ 1** การหาค่าความแตกต่างของคำว่า “อัตถศาสตร์” และ “athasart” เป็นคำทับศัพท์ที่ตรงกันหรือไม่ โดยทำการเข้ารหัสคำ แล้วคำนวณหาค่าความแตกต่าง

**การเข้ารหัสคำ**

อัตถศาสตร์ -> attasat

athasart -> atasat

**การค้นคืน**

minimum_edit_distance(“attasat”, “atasat”) = 1

จากตัวอย่างพบว่า รหัสคำทั้งสองมีค่าความแตกต่างเป็น 1 ถ้าหากกำหนดเกณฑ์ค่าความแตกต่างมีค่าเท่ากับ 1 หรือ $d$=1 ก็จะสามารถค้นคืนคู่คำนี้ได้

**5.2 เกณฑ์การประเมินการค้นคืน**

มาตรวัดที่ใช้วัดประสิทธิภาพของการค้นคืนได้แก่ ค่าความถูกต้อง (Accuracy) ซึ่งจะเป็นค่าความถูกต้องระดับตัวอักษรของคำ ค่าแม่นยำ (Precision) ค่าเรียกคืน (Recall) และตัววัด F1 (F1 Measurement) ซึ่งมีวิธีคำนวณจากการนับจำนวนข้อมูลที่เกี่ยวข้อง (Relevant Data) และข้อมูลที่ระบบค้นคืนกลับมา (Retrieved Data) ได้ดังนี้

$$\text{ค่าความถูกต้อง} = \frac{\text{ค่าสูงสุด}(\text{จำนวนตัวอักษรของอินพุต}, \text{จำนวนตัวอักษรเอาต์พุต}) - d}{\text{ค่าสูงสุด}(\text{จำนวนตัวอักษรของอินพุต}, \text{จำนวนตัวอักษรเอาต์พุต})}$$

เมื่อ $d$ แทน ค่าระยะห่างของการแก้ไขสั้นที่สุดของอินพุตและเอาต์พุต



$$\text{ค่าแม่นยำ} = \frac{\text{จำนวนคำที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำที่คืนกลับมา}}$$

$$\text{ค่าเรียกคืน} = \frac{\text{จำนวนคำที่เกี่ยวข้องที่คืนกลับมา}}{\text{จำนวนคำที่เกี่ยวข้องทั้งหมด}}$$

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

# บทที่ 6
# การทดลอง

ในบทนี้จะกล่าวถึงโมเดล ชุดข้อมูลและการแบ่งชุดข้อมูล การปรับจูนค่าพารามิเตอร์ของโมเดลทรานฟอร์เมอร์เพื่อให้ได้ค่าพารามิเตอร์ที่ดีที่สุดสำหรับโมเดลนี้ การเข้ารหัสคำโดยใช้ทรานฟอร์เมอร์ทั้งแบบการเรียนรู้แบบสอน และการเรียนรู้แบบกึ่งสอน ในหัวข้อต่อไปนี้จะกล่าวถึงแต่ละการทดลองและผลการทดลองที่ได้

## 6.1 โมเดล (Model)

ในการฝึกสอนโมเดลในปัจจุบันนั้นมีเทคนิคที่เรียกว่า การถ่ายทอดการเรียนรู้ (Transfer Learning) โดยใช้ค่าน้ำหนักที่ฝึกสอนไว้แล้ว (Pre-trained weight) มาใช้ เพื่อช่วยในการประหยัดเวลาในการฝึกสอนโมเดล ในงานวิจัยนี้จึงได้ทำการทดลองนำค่าน้ำหนักของตัวอักษร (Character embedding weight) ของโมเดลจากงานวิจัย [18] มาใช้ในการทดลอง พบว่าโมเดลนั้นไม่ได้รับการปรับปรุงไปในทางที่ดีขึ้น กล่าวคือโมเดลไม่ได้คืนค่าความถูกต้องที่มากกว่าเดิม เมื่อเทียบกับการฝึกสอนโมเดลโดยไม่ใช่ค่าน้ำหนักของตัวอักษร เพราะฉะนั้นในงานวิจัยนี้จะเป็นการฝึกสอนแบบตั้งแต่แรก (Scratch) โดยจะไม่มีการใช้ค่าน้ำหนักก่อนฝึกใด ๆ

## 6.2 ชุดข้อมูล (Dataset)

ชุดข้อมูลของการวิจัยนี้จะสามารถแบ่งชุดข้อมูลออกเป็น 2 กลุ่มหลัก ๆ ได้ดังนี้

6.1.1 ชุดข้อมูลที่มีฉลากกำกับ (Labeled data)

ชุดข้อมูลที่มีฉลากกำกับคือ ชุดข้อมูลที่มีการกำหนดรหัสคำของแต่ละคำไว้เพื่อใช้ในการฝึกสอนและเปรียบเทียบความถูกต้องเมื่อทำการทดลอง ประกอบไปด้วยข้อมูล 3 ชุดข้อมูล ซึ่งชุดข้อมูลที่ 1 และชุดข้อมูลที่ 2 นั้นได้มาจากงานวิจัย [13]

1. ชุดคำไทยและคำทับศัพท์ภาษาอังกฤษที่ตรงกัน ประกอบไปด้วย ชื่อ และชื่อสกุลของนิสิตจำนวน 2,000 คู่คำ
2. ชุดคำอังกฤษและคำทับศัพท์ภาษาไทยที่ตรงกัน ประกอบไปด้วย คำนามเฉพาะ คำศัพท์วิทยาศาสตร์ คำศัพท์คณิตศาสตร์ และคำศัพท์เคมี จำนวน 1,876 คู่คำ
3. ชุดคำไทยและคำทับศัพท์ภาษาอังกฤษที่ตรงกัน ประกอบไปด้วย ชื่อ และชื่อสกุล จำนวน 400 คู่คำ (ชุดข้อมูลนี้ถูกนำมาใช้เพื่อทำการทดสอบวัดค่าการทำงานของโมเดลว่าทำงานได้ดีแค่ไหน ในขั้นตอนการปรับจูนค่าพารามิเตอร์ต่าง ๆ)

เพื่อให้การทดลองไม่โน้มเอียงในการแบ่งชุดข้อมูลฝึก (Train set) และชุดข้อมูลทดสอบ (Test set) จึงใช้วิธี K-fold cross validation เข้ามาช่วยในการแบ่งข้อมูลออกเป็น K ส่วน เท่า ๆ กัน จากนั้นข้อมูลหนึ่งส่วนจะถูกนำมาเป็นชุดข้อมูลทดสอบ ส่วนที่เหลือจะถูกนำมาใช้เป็นข้อมูลฝึก ทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ จากนั้นนำค่าที่ได้จากการทดลองทั้งหมด K ครั้ง มาหาค่าเฉลี่ย ในการทดลองนี้ได้แบ่งข้อมูลออกเป็นส่วน ๆ ในกรณีชุดข้อมูลคำไทยและคำทับศัพท์ที่ภาษาอังกฤษที่ตรงกันจะแบ่งออกเป็น 5 ส่วน หรือ 400 คู่คำ และในกรณีชุดข้อมูลคำอังกฤษและคำทับศัพท์ภาษาไทยที่ตรงกันจะแบ่งออกเป็น 4 ส่วน หรือ 469 คู่คำ

ตารางที่ 14 แสดงถึงข้อมูลจำนวนคู่คำที่มีรหัสคำไม่เหมือนกัน สำหรับชุดข้อมูลคำไทยและคำทับศัพท์ภาษาอังกฤษที่ตรงกัน มีจำนวนคู่คำที่รหัสคำไม่เหมือนกันอยู่ที่ 33 คู่คำ หรือ 1.38% ต่อจำนวนคู่คำทั้งหมด ในส่วนของชุดข้อมูลอังกฤษและคำทับศัพท์ภาษาไทยที่ตรงกันมีจำนวนคู่คำที่รหัสคำไม่เหมือนกันอยู่ที่ 133 คู่คำ หรือ 7.09% ต่อจำนวนคู่คำทั้งหมด

*ตารางที่ 14 จำนวนคู่คำที่มีรหัสคำไม่เหมือนกัน*

| ชุดข้อมูล | จำนวนคู่คำทั้งหมด | จำนวนคู่คำที่ไม่เหมือนกัน | เปอร์เซ็นต์ |
|---|---|---|---|
| ชุดคำไทยและคำทับศัพท์ภาษาอังกฤษที่ตรงกัน | 2,400 | 33 | 1.38 |
| ชุดคำอังกฤษและคำทับศัพท์ภาษาไทยที่ตรงกัน | 1,876 | 133 | 7.09 |



6.1.2 ชุดข้อมูลที่ไม่มีฉลากกำกับ (Unlabeled data)

ชุดข้อมูลที่ไม่มีฉลากกำกับคือ ชุดข้อมูลที่ไม่มีการกำหนดรหัสคำของคำไว้ เพื่อใช้ในการฝึกสอนแบบกึ่งสอน (Semi-Supervised Learning) โดยลำดับของคำและคำทับศัพท์ที่ตรงกันนั้นจะอยู่ในลำดับที่ติดกัน ยกตัวอย่างเช่น คำไทย “เชี่ยวชาญ” อยู่ลำดับที่ 1 และคำทับศัพท์ภาษาอังกฤษ “chiaochan” อยู่ลำดับที่ 2 เป็นต้น ประกอบไปด้วยข้อมูล 2 ชุดข้อมูล

1. ชุดคำไทยและคำทับศัพท์ภาษาอังกฤษที่ตรงกัน ประกอบไปด้วย ชื่อ และชื่อสกุล จำนวน 5,000 คู่คำ
2. ชุดคำอังกฤษและคำทับศัพท์ภาษาไทยที่ตรงกัน ประกอบไปด้วย ชื่อ และคำศัพท์ในชีวิตประจำวัน จำนวน 4,496 คู่คำ

*ตารางที่ 15 ตัวอย่างคำและรหัสคำ*

| คำ | รหัสคำ | คำทับศัพท์ที่ตรงกัน | รหัสคำ |
|---|---|---|---|
| ปิยะวัฒน์ | piyavat | piyawat | piyavat |
| ศริพันธุ์ | saripan | saripant | saripan |
| กิจมีดี | kitmidi | kijmedee | kitmidi |
| soup | sup | ซุป | sup |
| aluminium | aluminIm | อะลูมิเนียม | aluminIm |
| polonium | polonIm | พอโลเนียม | polonIm |

**6.3 การปรับจูนค่าพารามิเตอร์ (Hyperparameter Tuning)**

การปรับจูนค่าพารามิเตอร์คือ การปรับค่าพารามิเตอร์ต่าง ๆ ให้เหมาะสมกับข้อมูลที่ป้อนเข้าไปเพื่อสอนโมเดลของเรา และเพื่อให้โมเดลเรามีประสิทธิภาพมากที่สุด โดยที่โมเดลสามารถทำการค้นคืนโดยมีค่าความถูกต้องมากที่สุด

วิธีการทดลองคือการนำชุดข้อมูลภาษาไทยที่มีฉลากกำกับทั้งหมด (ชุดข้อมูลที่ 1 จากหัวข้อ 6.1.1) มาทำการสอนแบบการเรียนรู้แบบสอน และใช้ชุดข้อมูลภาษาไทยชุดข้อมูลที่ 3 (จากหัวข้อ 6.1.1) ในการวัดผลค่าความถูกต้องของการค้นคืน เครื่องมือเพิ่มประสิทธิภาพ (Optimizer) ของการวิจัยนี้ใช้เป็น Adam จำนวนแบทช์ (Batch size) คือ 128 และจำนวนรอบการสอน (Epoch) คือ 10 รอบในแต่ละการทดลอง

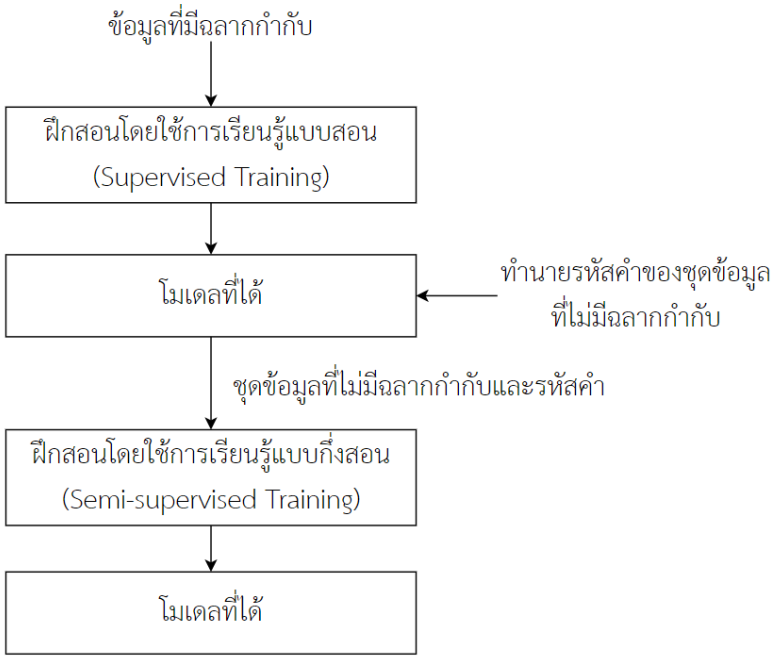


เพื่อหาค่าพารามิเตอร์ต่าง ๆ ที่มีค่าความถูกต้องมากที่สุดนั้น จะเริ่มจากการปรับค่าการเรียนรู้ (Learning rate) เพื่อหาว่าค่าไหนที่สามารถคืนค่าความถูกต้องมากที่สุด แล้วเราจะยึดค่าการเรียนรู้นั้นไว้ จากนั้นก็เริ่มปรับค่าพารามิเตอร์ตัวอื่นต่อไปในลักษณะเดียวกันกับการปรับค่าการเรียนรู้ จากตารางที่ 16 จะเห็นว่าค่าการเรียนรู้มีการปรับทั้งหมด 9 ค่า แต่การทดลองที่มีค่าความถูกต้องมากที่สุดนั้นก็คือค่าการเรียนรู้ 0.005 ที่มีค่าความถูกต้องอยู่ที่ 87.54% เมื่อทำการปรับพารามิเตอร์ครบทุกตัวแล้วพบว่าค่าความถูกต้องที่มากที่สุด มีค่าเท่ากับ 89.62% เมื่อใช้ค่า Learning rate ที่ 0.005 ค่า Embedding dimension ที่ 32 ค่า Encoder layer number ที่ 2 ค่า Decoder layer number ที่ 2 ค่า Head number ที่ 4 ค่า Hidden dimension ที่ 128 และค่า Dropout rate ที่ 0.3 ซึ่งค่าพารามิเตอร์ที่ได้จากผลการทดลองนี้จะถูกนำมาใช้กับการฝึกสอนโมเดลในการทดลองในลำดับต่อไป

*ตารางที่ 16 ผลการปรับค่าพารามิเตอร์*

| # | Learning Rate | Embedding dimension | Encoder layer number | Decoder layer number | Head number | Hidden dimension | Dropout rate | Accuracy |
|---|---|---|---|---|---|---|---|---|
| 1 | **0.03** | 32 | 2 | 2 | 4 | 128 | 0.05 | 14.76% |
| 2 | **0.003** | 32 | 2 | 2 | 4 | 128 | 0.05 | 86.99% |
| 3 | **0.0003** | 32 | 2 | 2 | 4 | 128 | 0.05 | 86.68% |
| 4 | **0.05** | 32 | 2 | 2 | 4 | 128 | 0.05 | 14.76% |
| 5 | **0.005** | 32 | 2 | 2 | 4 | 128 | 0.05 | 87.54% |
| 6 | **0.0005** | 32 | 2 | 2 | 4 | 128 | 0.05 | 86.50% |
| 7 | **0.07** | 32 | 2 | 2 | 4 | 128 | 0.05 | 14.76% |
| 8 | **0.007** | 32 | 2 | 2 | 4 | 128 | 0.05 | 51.71% |
| 9 | **0.0007** | 32 | 2 | 2 | 4 | 128 | 0.05 | 86.00% |
| 10 | 0.005 | **64** | 2 | 2 | 4 | 128 | 0.05 | 42.84% |
| 11 | 0.005 | **128** | 2 | 2 | 4 | 128 | 0.05 | 22.82% |
| 12 | 0.005 | 32 | **1** | 2 | 4 | 128 | 0.05 | 75.20% |
| 13 | 0.005 | 32 | **3** | 2 | 4 | 128 | 0.05 | 51.86% |
| 14 | 0.005 | 32 | **4** | 2 | 4 | 128 | 0.05 | 19.18% |
| 15 | 0.005 | 32 | 2 | **1** | 4 | 128 | 0.05 | 86.84% |
| 16 | 0.005 | 32 | 2 | **3** | 4 | 128 | 0.05 | 78.31% |
| 17 | 0.005 | 32 | 2 | **4** | 4 | 128 | 0.05 | 49.48% |
| 18 | 0.005 | 32 | **1** | **1** | 4 | 128 | 0.05 | 85.97% |
| 19 | 0.005 | 32 | **3** | **3** | 4 | 128 | 0.05 | 26.06% |
| 20 | 0.005 | 32 | 2 | 2 | **8** | 128 | 0.05 | 85.67% |
| 21 | 0.005 | 32 | 2 | 2 | 4 | **64** | 0.05 | 86.28% |
| 22 | 0.005 | 32 | 2 | 2 | 4 | **256** | 0.05 | 69.17% |
| 23 | 0.005 | 32 | 2 | 2 | 4 | 128 | **0.01** | 82.57% |
| 24 | 0.005 | 32 | 2 | 2 | 4 | 128 | **0.1** | 86.49% |
| 25 | 0.005 | 32 | 2 | 2 | 4 | 128 | **0.3** | **89.62%** |
| 26 | 0.005 | 32 | 2 | 2 | 4 | 128 | **0.5** | 88.04% |

### 6.4 การเข้ารหัสคำโดยใช้ทรานฟอร์เมอร์




*ภาพที่ 8 การฝึกสอนโมเดล*

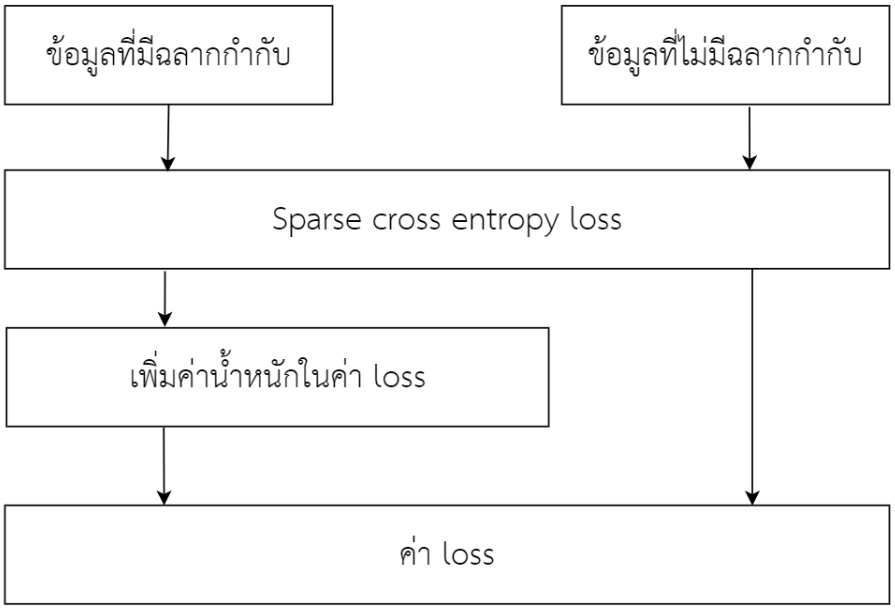
การเข้ารหัสคำโดยใช้ทรานฟอร์เมอร์จะทำการฝึกสอนโมเดลทั้งหมด 2 รอบ รอบแรกจะเป็นการฝึกสอนแบบการเรียนรู้แบบสอน (Supervised training) จากนั้นเมื่อได้โมเดลออกมาแล้ว จะทำการทำนายรหัสคำของชุดข้อมูลที่ไม่มีฉลากกำกับ เพื่อนำคำและรหัสคำที่ได้จากการทำนาย มาใช้ในการฝึกสอนแบบการเรียนรู้แบบกึ่งสอน (Semi-supervised training) และเพื่อให้โมเดลที่ได้นั้นไม่เกิดการแกว่งในชุดข้อมูลที่ฝึกสอน และทำงานอย่างเต็มประสิทธิภาพแล้วนั้น จึงจำเป็นต้องใช้วิธีการเหล่านี้เข้ามาช่วยในการฝึกสอนเพิ่มเติม ดังนี้

6.3.1 การผสมชุดข้อมูลที่มีฉลากกำกับและไม่มีฉลากกำกับ

การผสมชุดข้อมูลที่มีฉลากกำกับและไม่มีฉลากกำกับเข้าด้วยกัน และทำการฝึกสอนจะช่วยลดการแกว่งของชุดข้อมูลที่ฝึกสอนได้ ด้วยวิธีการนำรหัสคำของคู่คำที่ตรงกันสำหรับชุดข้อมูลที่ไม่มีฉลากกำกับมาเทียบกัน ถ้าหากมีรหัสคำที่ได้เหมือนกันจะถูกนำมาไว้ในชุดข้อมูลเพื่อทำการฝึกสอน ยกตัวอย่างเช่น คำว่า “บางทอง” ได้รหัสเสียงออกมาเป็น “bagtog” และคำทับศัพท์ที่ตรงกันคำว่า “bangthong” ได้รหัสเสียงออกมาเป็น “bagtog” ซึ่งรหัสเสียงของทั้งสองคำนี้เป็นรหัสเสียงที่ตรงกัน คู่คำนี้ก็จะถูกนำมาใช้ในการฝึกสอน

จำนวนคำที่ถูกนำมาใช้ในการเรียนรู้นั้น ยกตัวอย่างในกรณีของ ชุดคำภาษาไทยจะมีจำนวนทั้งหมด 1,600 คู่คำ (อีก 400 คู่คำที่เหลือจะเป็นชุดข้อมูลทดสอบ) เพื่อให้การผสมข้อมูลอยู่ในจำนวนที่เท่ากัน คู่คำที่รหัสเสียงเหมือนกันของชุดข้อมูลที่ไม่มีฉลากกำกับจำนวน 1,600 คู่แรกจะถูกนำมาใช้ในการเรียนรู้ในครั้งนี้ หลังจากนั้น เราจะทำการเรียงลำดับข้อมูลใหม่ โดยอิงจากขนาดของแบทช์เป็นหลัก ซึ่งในการวิจัยนี้จำนวนแบทช์ที่ใช้คือ 128 เราจะทำการเรียงลำดับโดยใช้ชุดข้อมูลที่มีฉลากกำกับจำนวน 64 คำแรก (32 คู่คำ) มาใส่ไว้ในชุดข้อมูล และใช้คำจากชุดข้อมูลที่ไม่มีฉลากกำกับจำนวน 64 คำ (32 คู่คำ) มาต่อท้าย ทำแบบนี้วนไปเรื่อย ๆ จนขนาดของข้อมูลครบ 3,200 คู่คำ

6.3.2 การปรับแต่งฟังก์ชันต้นทุน (Loss function)




*ภาพที่ 9 การปรับแต่งฟังก์ชันต้นทุน (Loss function)*

การปรับแต่งฟังก์ชันต้นทุนโดยการเพิ่มค่าน้ำหนักบนชุดข้อมูลที่มีฉลากกำกับนั้นจะช่วยทำให้โมเดลให้ความสำคัญกับชุดข้อมูลที่มีฉลากกำกับมากกว่าชุดข้อมูลที่ไม่มีฉลากกำกับ วิธีการปรับแต่งนั้นจะยึดจากจำนวนแบทช์ที่ใช้ในงานวิจัยนี้ นั่นก็คือจำนวน 128 จากหัวข้อการผสมข้อมูลที่ผ่านมา จะเห็นว่าชุดข้อมูลที่ถูกนำมาเรียงใหม่นั้น ข้อมูลจำนวน 64 แถวแรกจะเป็นชุดข้อมูลที่มีฉลากกำกับ และ 64 แถวต่อมาจะเป็นชุดข้อมูลที่ไม่มีฉลากกำกับ ดังนั้นเราจะสามารถแบ่งการคำนวณค่าต้นทุนออกเป็น 2 ส่วนได้ ส่วนแรกจะเป็นค่าต้นทุนของ

ชุดข้อมูลที่มีฉลากกำกับ และส่วนที่สองจะเป็นค่าต้นทุนของชุดข้อมูลที่ไม่มีฉลากกำกับ ในกรณีของการคำนวณค่าต้นทุนของชุดข้อมูลที่มีฉลากกำกับ เราจะทำการเพิ่มค่าน้ำหนัก (weight) เข้าไปเพิ่มเติม ซึ่งในการวิจัยครั้งนี้ค่าน้ำหนักที่เพิ่มเข้าไป มีค่าเท่ากับ 25% ของค่าน้ำหนักเดิม

**6.5 ผลการทดลอง**

ผลการทดลองในงานวิจัยนี้จะประกอบไปด้วย การวัดค่าความถูกต้อง ค่าความแม่นยำ ค่าเรียกคืน และตัววัด F1

โดยตารางที่ 17 จะแสดงถึงค่าความถูกต้องที่ได้จากโมเดลการเรียนรู้แบบสอน ตารางที่ 18 จะแสดงถึงค่าความถูกต้องที่ได้จากโมเดลการเรียนรู้แบบกึ่งสอน และตารางที่ 19 จะแสดงการเปรียบเทียบผลค่าความถูกต้องของงานวิจัยนี้และงานวิจัยเดิมของศิริพจน์ สุรบถโสภณ [13] โดยนำผลค่าความถูกต้องจากการเข้ารหัสคำด้วยนิวรอลเน็ตเวิร์กมาเทียบเท่านั้น จากผลการทดลองนี้ พบว่าโมเดลของงานวิจัยนี้นั้นดีขึ้นใน 2 ส่วน ส่วนแรกเมื่อเทียบกับงานวิจัยเดิมของศิริพจน์ สุรบถโสภณ ค่าความถูกต้องเฉลี่ยของโมเดลการเรียนรู้แบบสอนนั้นมีค่าความถูกต้องเฉลี่ยมากกว่างานวิจัยเดิมของศิริพจน์ สุรบถโสภณ ถึง 5.91% และค่าความถูกต้องเฉลี่ยของโมเดลการเรียนรู้แบบกึ่งสอนนั้นมีค่าความถูกต้องเฉลี่ยมากกว่างานวิจัยเดิมของศิริพจน์ สุรบถโสภณ ถึง 7.38% ส่วนที่สองคือค่าความถูกต้องเฉลี่ยของโมเดลการเรียนรู้แบบกึ่งสอนนั้นมีค่าความถูกต้องเฉลี่ยมากกว่าโมเดลการเรียนรู้แบบสอนถึง 1.47%



*ตารางที่ 17 ค่าความถูกต้องของโมเดลการเรียนรู้แบบสอน*

| ชุดข้อมูล | ค่าความถูกต้อง (เปอร์เซ็นต์) | | | | | |
|---|---|---|---|---|---|---|
| | ชุดที่ 1 | ชุดที่ 2 | ชุดที่ 3 | ชุดที่ 4 | ชุดที่ 5 | ค่าเฉลี่ย |
| คำไทย | 92.57 | 92.56 | 91.14 | 90.95 | 91.56 | 91.76 |
| คำอังกฤษทับศัพท์คำไทย | 95.66 | 95.58 | 94.69 | 94.76 | 94.97 | 95.13 |
| คำอังกฤษ | 84.96 | 83.69 | 84.46 | 83.34 | - | 84.11 |
| คำไทยทับศัพท์คำอังกฤษ | 94.83 | 95.87 | 96.24 | 94.55 | - | 95.37 |

*ตารางที่ 18 ค่าความถูกต้องของโมเดลการเรียนรู้แบบกึ่งสอน*

| ชุดข้อมูล | ค่าความถูกต้อง (เปอร์เซ็นต์) | | | | | |
|---|---|---|---|---|---|---|
| | ชุดที่ 1 | ชุดที่ 2 | ชุดที่ 3 | ชุดที่ 4 | ชุดที่ 5 | ค่าเฉลี่ย |
| คำไทย | 93.98 | 94.35 | 92.02 | 94.01 | 93.43 | 93.56 |
| คำอังกฤษทับศัพท์คำไทย | 96.42 | 96.09 | 95.01 | 94.73 | 95.94 | 95.64 |
| คำอังกฤษ | 86.22 | 87.23 | 88.19 | 86.89 | - | 87.13 |
| คำไทยทับศัพท์คำอังกฤษ | 95.54 | 96.49 | 96.35 | 95.28 | - | 95.92 |

*ตารางที่ 19 การเปรียบเทียบผลของค่าความถูกต้องของงานวิจัยนี้*
*และงานวิจัยของศิริพจน์ สุรบถโสภณ และบุญเสริม กิจศิริกุล*

| ชุดข้อมูล | ค่าความถูกต้อง (เปอร์เซ็นต์) | | |
|---|---|---|---|
| | งานวิจัยของศิริพจน์ | งานวิจัยนี้ | |
| | | โมเดลการเรียนรู้แบบสอน | โมเดลการเรียนรู้แบบกึ่งสอน |
| คำไทย | 85.00 | 91.76 | 93.56 |
| คำอังกฤษทับศัพท์คำไทย | 91.41 | 95.13 | 95.64 |
| คำอังกฤษ | 75.00 | 84.11 | 87.13 |
| คำไทยทับศัพท์คำอังกฤษ | 91.31 | 95.37 | 95.92 |
| ค่าเฉลี่ย | 85.68 | 91.59 | **93.06** |

เนื่องจากวิธีการที่ใช้วัดค่าความถูกต้องของงานวิจัยนี้เป็นการวัดค่าความถูกต้องระดับตัวอักษรของคำ จึงได้มีการทดลองการวัดค่าความถูกต้องของคำโดยจัดแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต โดยตารางที่ 20 แสดงถึงค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุตของชุดข้อมูลคำไทย ตารางที่ 21 แสดงถึงค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุตของชุดข้อมูลคำอังกฤษทับศัพท์คำไทย ตารางที่ 22 ค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุตของชุดข้อมูลคำอังกฤษ ตารางที่ 23 ค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุตของชุดข้อมูลคำไทยทับศัพท์คำอังกฤษ ตารางที่ 24 แสดงการสรุปผลค่าความถูกต้องเฉลี่ยแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต จากภาพที่ 10 เป็นการนำผลที่ได้จากตารางที่ 25 มาวาดกราฟเพื่อเปรียบเทียบข้อมูลแต่ละชุด พบว่าค่าความถูกต้องเฉลี่ยของชุดข้อมูลคำไทย และชุดข้อมูลคำอังกฤษทับศัพท์คำไทย มีการลดลงเพียงเล็กน้อยเท่านั้นเมื่อจำนวนตัวอักษรของอินพุตมีจำนวนมากขึ้น ในส่วนของชุดข้อมูลคำอังกฤษ และชุดข้อมูลคำไทยทับศัพท์คำอังกฤษนั้นพบว่า ค่าความถูกต้องเฉลี่ยใน

กลุ่มจำนวนตัวอักษรอินพุตที่ 16 – 20 ตัวอักษร มีค่าความถูกต้องที่น้อยกว่ากลุ่มอื่นอย่างเห็นได้ชัด ซึ่งมีค่าความถูกต้องน้อยกว่ากลุ่มอื่น ๆ อยู่ที่ประมาณ 20 – 30%

*ตารางที่ 20 ค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต ของชุดข้อมูลคำไทย*

| จำนวนตัวอักษรอินพุต | | ชุดที่ 1 | ชุดที่ 2 | ชุดที่ 3 | ชุดที่ 4 | ชุดที่ 5 |
|---|---|---|---|---|---|---|
| 1 - 5 | จำนวนคำ | 46 | 47 | 56 | 46 | 58 |
| | ค่าความถูกต้อง (%) | 93.99 | 96.15 | 93.77 | 94.45 | 91.51 |
| 6 - 10 | จำนวนคำ | 249 | 249 | 238 | 248 | 259 |
| | ค่าความถูกต้อง (%) | 95.37 | 95.14 | 92.77 | 94.93 | 95.00 |
| 11 - 15 | จำนวนคำ | 98 | 96 | 102 | 98 | 75 |
| | ค่าความถูกต้อง (%) | 91.09 | 92.02 | 89.17 | 92.25 | 89.88 |
| 16 – 20 | จำนวนคำ | 7 | 8 | 4 | 8 | 8 |
| | ค่าความถูกต้อง (%) | 84.89 | 87.11 | 95.10 | 84.33 | 90.07 |

*ตารางที่ 21 ค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต ของชุดข้อมูลคำอังกฤษทับศัพท์คำไทย*

| จำนวนตัวอักษรอินพุต | | ชุดที่ 1 | ชุดที่ 2 | ชุดที่ 3 | ชุดที่ 4 | ชุดที่ 5 |
|---|---|---|---|---|---|---|
| 1 - 5 | จำนวนคำ | 25 | 10 | 19 | 23 | 14 |
| | ค่าความถูกต้อง (%) | 100 | 100 | 94.04 | 95.65 | 97.14 |
| 6 - 10 | จำนวนคำ | 248 | 264 | 245 | 232 | 277 |
| | ค่าความถูกต้อง (%) | 96.40 | 97.16 | 95.65 | 95.68 | 96.94 |
| 11 - 15 | จำนวนคำ | 109 | 112 | 116 | 127 | 91 |
| | ค่าความถูกต้อง (%) | 96.14 | 94.22 | 94.75 | 93.23 | 92.96 |
| 16 – 20 | จำนวนคำ | 18 | 14 | 20 | 18 | 18 |
| | ค่าความถูกต้อง (%) | 93.33 | 88.17 | 89.55 | 91.98 | 94.74 |

*ตารางที่ 22 ค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต ของชุดข้อมูลคำอังกฤษ*

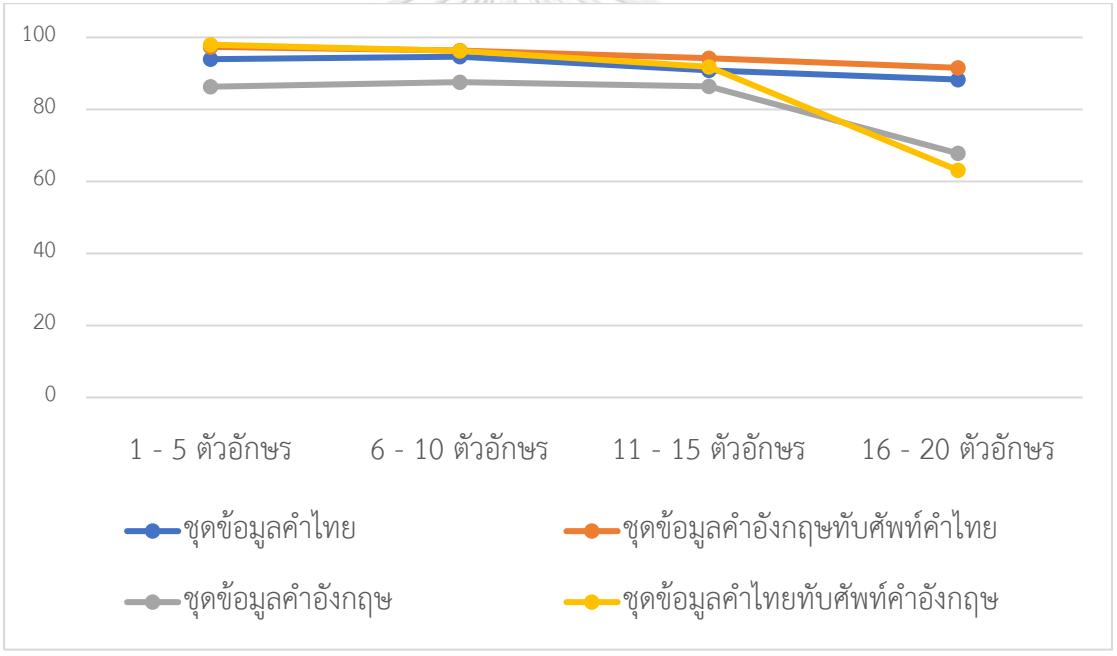
| จำนวนตัวอักษรอินพุต | | ชุดที่ 1 | ชุดที่ 2 | ชุดที่ 3 | ชุดที่ 4 |
|---|---|---|---|---|---|
| 1 - 5 | จำนวนคำ | 125 | 116 | 128 | 128 |
| | ค่าความถูกต้อง (%) | 84.73 | 84.91 | 88.36 | 87.29 |
| 6 - 10 | จำนวนคำ | 309 | 324 | 310 | 304 |
| | ค่าความถูกต้อง (%) | 86.86 | 88.05 | 88.37 | 87.02 |
| 11 - 15 | จำนวนคำ | 35 | 28 | 29 | 36 |
| | ค่าความถูกต้อง (%) | 85.83 | 88.16 | 86.87 | 84.77 |
| 16 – 20 | จำนวนคำ | - | 1 | 2 | 1 |
| | ค่าความถูกต้อง (%) | - | 62.5 | 67.5 | 73.33 |

*ตารางที่ 23 ค่าความถูกต้องแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต ของชุดข้อมูลคำไทยทับศัพท์คำอังกฤษ*

| จำนวนตัวอักษรอินพุต | | ชุดที่ 1 | ชุดที่ 2 | ชุดที่ 3 | ชุดที่ 4 |
|---|---|---|---|---|---|
| 1 - 5 | จำนวนคำ | 110 | 103 | 101 | 111 |
| | ค่าความถูกต้อง (%) | 95.91 | 99.48 | 98.75 | 97.78 |
| 6 - 10 | จำนวนคำ | 287 | 309 | 304 | 296 |
| | ค่าความถูกต้อง (%) | 96.27 | 96.87 | 96.47 | 95.45 |
| 11 - 15 | จำนวนคำ | 72 | 54 | 62 | 58 |
| | ค่าความถูกต้อง (%) | 92.06 | 90.09 | 93.18 | 91.92 |
| 16 – 20 | จำนวนคำ | - | 3 | 2 | 4 |
| | ค่าความถูกต้อง (%) | - | 70.83 | 56.25 | 62.18 |

*ตารางที่ 24 สรุปผลค่าความถูกต้องเฉลี่ยแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต*

| | ค่าความถูกต้อง (เปอร์เซ็นต์) | | | |
|---|---|---|---|---|
| จำนวนตัวอักษรอินพุต | ชุดข้อมูลคำไทย | ชุดข้อมูลคำอังกฤษทับศัพท์คำไทย | ชุดข้อมูลคำอังกฤษ | ชุดข้อมูลคำไทยทับศัพท์คำอังกฤษ |
| 1 – 5 | 93.97 | 97.37 | 86.32 | 97.98 |
| 6 – 10 | 94.64 | 96.37 | 87.58 | 96.27 |
| 11 – 15 | 90.88 | 94.26 | 86.41 | 91.81 |
| 16 - 20 | 88.30 | 91.55 | 67.78 | 63.09 |




*ภาพที่ 10 ค่าความถูกต้องเฉลี่ยแบ่งกลุ่มตามจำนวนตัวอักษรของอินพุต*

ผลการทดลองถัดมาคือ การคำนวณหาค่าความแม่นยำ ค่าเรียกคืน และตัววัด F1 ที่ระยะห่างของค่าความต่างของรหัสคำ ซึ่งผู้วิจัยได้ทำการทดลองโดยการเปลี่ยนแปลงค่าพารามิเตอร์ *d* หรือค่าระยะห่างของการแก้ไขสั้นที่สุด เพื่อหาค่าความแตกต่างที่น้อยที่สุดที่ให้ค่าเฉลี่ยของค่าความแม่นยำ และค่าเรียกคืนสูงที่สุด ในกรณีที่ค่า *d*=0 หมายความว่าการเปรียบเทียบรหัสคำที่ได้แบบเหมือนกันทุกประการ (Exactly Matching) และยังทดลองกับค่า *d* ในค่าอื่น ๆ อีกด้วยโดยมีค่าตั้งแต่ 0 ถึง 3 และ 10% ของความยาวอินพุต โดยตารางที่ 25 แสดงค่าความแม่นยำของโมเดลการเรียนรู้แบบกึ่งสอน ตารางที่ 26 แสดงเรียกคืนของโมเดลการเรียนรู้แบบกึ่งสอน ตารางที่ 27 แสดงค่าตัววัด F1 ของโมเดลการเรียนรู้แบบกึ่งสอน และตารางที่ 28 สรุปผลค่าเฉลี่ยความแม่นยำ ค่าเรียกคืน และตัววัด F1 ของโมเดลการเรียนรู้แบบกึ่งสอน จากตารางนี้พบว่า ค่าเฉลี่ยของตัววัด F1 จะมีค่าสูงที่สุดเมื่อ *d* มีค่าเท่ากับ 10% ของความยาวอินพุต ซึ่งตัววัด F1 มีค่าเฉลี่ยสูงถึง 86.63% ในชุดข้อมูลคำไทยและคำอังกฤษทับศัพท์คำไทย และค่าเฉลี่ยตัววัด F1 มีค่าสูงถึง 82.51% ในชุดข้อมูลคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ

*ตารางที่ 25 ค่าความแม่นยำของโมเดลการเรียนรู้แบบกึ่งสอน*

| ชุดข้อมูล | *d* | ค่าความแม่นยำ (เปอร์เซ็นต์) | | | | | |
|---|---|---|---|---|---|---|---|
| | | ชุดที่ 1 | ชุดที่ 2 | ชุดที่ 3 | ชุดที่ 4 | ชุดที่ 5 | ค่าเฉลี่ย |
| ชุดคำไทยและคำอังกฤษทับศัพท์คำไทย | 0 | 75.75 | 73.62 | 68.25 | 70.50 | 72.00 | 72.02 |
| | 1 | 83.14 | 82.12 | 79.98 | 82.48 | 83.29 | 82.20 |
| | 2 | 73.33 | 75.38 | 72.63 | 71.71 | 75.33 | 73.68 |
| | 3 | 48.63 | 51.29 | 53.07 | 49.79 | 48.35 | 50.23 |
| | 10% ของความยาวอินพุต | 86.01 | 85.38 | 83.98 | 84.48 | 85.54 | 85.08 |
| ชุดคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ | 0 | 67.16 | 69.19 | 68.66 | 65.35 | - | 67.59 |
| | 1 | 78.99 | 79.42 | 79.24 | 75.11 | - | 78.19 |
| | 2 | 59.91 | 61.80 | 62.45 | 57.75 | - | 60.48 |
| | 3 | 32.13 | 34.03 | 34.68 | 34.48 | - | 33.83 |
| | 10% ของความยาวอินพุต | 79.26 | 80.27 | 79.77 | 75.54 | - | 78.71 |

*ตารางที่ 26 ค่าเรียกคืนของโมเดลการเรียนรู้แบบกึ่งสอน*

| ชุดข้อมูล | *d* | ค่าเรียกคืน (เปอร์เซ็นต์) | | | | | |
|---|---|---|---|---|---|---|---|
| | | ชุดที่ 1 | ชุดที่ 2 | ชุดที่ 3 | ชุดที่ 4 | ชุดที่ 5 | ค่าเฉลี่ย |
| ชุดคำไทยและคำอังกฤษทับศัพท์คำไทย | 0 | 75.00 | 73.19 | 68.19 | 70.31 | 72.00 | 71.74 |
| | 1 | 86.88 | 86.12 | 81.56 | 85.56 | 86.62 | 85.35 |
| | 2 | 94.38 | 95.50 | 92.69 | 93.06 | 95.25 | 94.18 |
| | 3 | 98.50 | 98.25 | 97.50 | 97.75 | 98.62 | 98.12 |
| | 10% ของความยาวอินพุต | 89.88 | 89.38 | 85.56 | 87.56 | 88.88 | 88.25 |
| ชุดคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ | 0 | 65.94 | 67.32 | 65.72 | 63.01 | - | 65.50 |
| | 1 | 85.50 | 86.73 | 87.37 | 84.91 | - | 86.13 |
| | 2 | 94.35 | 95.95 | 96.91 | 94.35 | - | 95.39 |
| | 3 | 98.19 | 98.77 | 98.72 | 98.24 | - | 98.48 |
| | 10% ของความยาวอินพุต | 85.82 | 87.58 | 87.90 | 85.50 | - | 86.70 |

*ตารางที่ 27 ตัววัด F1 ของโมเดลการเรียนรู้แบบกึ่งสอน*

| ชุดข้อมูล | *d* | ตัววัด F1 (เปอร์เซ็นต์) | | | | | |
|---|---|---|---|---|---|---|---|
| | | ชุดที่ 1 | ชุดที่ 2 | ชุดที่ 3 | ชุดที่ 4 | ชุดที่ 5 | ค่าเฉลี่ย |
| ชุดคำไทยและคำอังกฤษทับศัพท์คำไทย | 0 | 75.37 | 73.41 | 68.22 | 70.41 | 72.00 | 71.88 |
| | 1 | 84.96 | 84.08 | 80.76 | 83.99 | 84.93 | 83.74 |
| | 2 | 82.53 | 84.26 | 81.44 | 81.00 | 84.13 | 82.67 |
| | 3 | 65.11 | 67.40 | 68.73 | 65.97 | 64.89 | 66.42 |
| | 10% ของความยาวอินพุต | 87.90 | 87.33 | 84.76 | 85.99 | 87.18 | 86.63 |
| ชุดคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ | 0 | 66.55 | 68.24 | 67.16 | 64.16 | - | 66.53 |
| | 1 | 82.12 | 82.91 | 83.10 | 79.71 | - | 81.96 |
| | 2 | 73.28 | 75.18 | 75.96 | 71.65 | - | 74.02 |
| | 3 | 48.42 | 50.62 | 51.32 | 51.05 | - | 50.35 |
| | 10% ของความยาวอินพุต | 82.41 | 83.77 | 83.64 | 80.21 | - | 82.51 |

*ตารางที่ 28 สรุปผลค่าเฉลี่ยความแม่นยำ ค่าเรียกคืน และตัววัด F1 ของโมเดลการเรียนรู้แบบกึ่งสอน*

| | | เปอร์เซ็นต์ | | |
|---|---|---|---|---|
| ชุดข้อมูล | *d* | ค่าแม่นยำ | ค่าเรียกคืน | ตัววัด F1 |
| ชุดคำไทยและคำอังกฤษทับศัพท์คำไทย | 0 | 72.02 | 71.74 | 71.88 |
| | 1 | 82.20 | 85.35 | 83.74 |
| | 2 | 73.68 | 94.18 | 82.67 |
| | 3 | 50.23 | 98.12 | 66.42 |
| | 10% ของความยาวอินพุต | 85.08 | 88.25 | **86.63** |
| ชุดคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ | 0 | 67.59 | 65.50 | 66.53 |
| | 1 | 78.19 | 86.13 | 81.96 |
| | 2 | 60.48 | 95.39 | 74.02 |
| | 3 | 33.83 | 98.48 | 50.35 |
| | 10% ของความยาวอินพุต | 78.71 | 86.70 | **82.51** |

**6.6 วิเคราะห์ผลการทดลองการเข้ารหัสคำ**

เมื่อนำผลการทดลองสำหรับคำไทยมาวิเคราะห์ พบว่ารหัสเสียงที่มีความผิดพลาดในจำแนกรหัสเสียงที่พบบ่อยนั้นจะเป็นกลุ่มคำที่มีตัวอักษรของสระอะลดรูป ยกตัวอย่างเช่นคำว่า “อุปลา” รหัสเสียงที่ถูกต้องคือ “upala” แต่เอาต์พุตที่ได้คือ “upla” หรือคำว่า “ศุภชัย” รหัสเสียงที่ถูกต้องคือ “suppac!” แต่เอาต์พุตที่ได้คือ “supc!” เป็นต้น ส่วนในกรณีของคำอังกฤษทับศัพท์คำไทย เมื่อนำผลการทดลองมาวิเคราะห์แล้วพบว่า รหัสเสียงที่มีความผิดพลาดในการจำแนกรหัสเสียง ส่วนใหญ่แล้วจะเป็นกลุ่มของรหัสเสียง ของตัวอักษร “a” และ “u” เช่นคำว่า “ponsup” รหัสเสียงที่ถูกต้องคือ “ponsap” แต่เอาต์พุตที่ได้คือ “ponsup” หรือคำว่า “rungroj” รหัสเสียงที่ถูกต้องคือ “rugrot” แต่เอาต์พุตที่ได้คือ “ragrot” เป็นต้น

สำหรับผลการทดลองของคำอังกฤษ และคำไทยทับศัพท์คำอังกฤษ เมื่อวิเคราะห์แล้วพบว่าไม่มีจุดสังเกตที่ชัดเจน หรือจุดที่รหัสเสียงที่มีการจำแนกผิดอยู่บ่อยครั้ง เท่ากับการทดลองของชุดข้อมูลที่ได้กล่าวมาทั้งสองชุด

## 6.7 สรุป

ในบทนี้ได้กล่าวถึงโมเดลทรานฟอร์เมอร์ ผลการทดลองการเข้ารหัสคำโดยใช้ทรานฟอร์เมอร์ทั้งแบบการเรียนรู้แบบสอน และการเรียนรู้แบบกึ่งสอน โดยมีการปรับจูนค่าพารามิเตอร์ต่าง ๆ และเทคนิคที่ใช้ในการฝึกสอนโมเดลในการเรียนรู้แบบกึ่งสอน การทดลองการวัดค่าความถูกต้องพบว่าโมเดลทรานฟอร์เมอร์ทั้งสองแบบให้ค่าความถูกต้องที่สูงกว่างานวิจัยเดิมของศิริพจน์ สุรบถโสภณ หลังจากนั้นได้แสดงผลการทดลองการวัดค่าความแม่นยำ ค่าเรียกคืน และตัววัด F1 ซึ่งค่าตัววัด F1 มีค่าสูงถึง 86.63% ในกรณีคำไทยและคำอังกฤษทับศัพท์คำไทย และ 82.51% ในกรณีคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ

# บทที่ 7
# สรุปผลการวิจัยและข้อเสนอแนะ

## 7.1 สรุปผลการวิจัย

งานวิจัยนี้เสนอการค้นคืนข้ามภาษาสำหรับคำทับศัพท์ภาษาไทย-อังกฤษ โดยใช้ทรานฟอร์เมอร์ ในการค้นคืนนั้นอาศัยรหัสคำซึ่งแทนเสียงอ่านของคำในการเปรียบเทียบคำคู่คำว่ามีเสียงอ่านที่ตรงกันหรือไม่ โดยใช้เทคนิคระยะการแก้ไขสั้นที่สุดเข้ามาช่วยในการเปรียบเทียบคำ

ชุดข้อมูลคำอังกฤษและชุดข้อมูลคำไทยทับศัพท์คำอังกฤษในกลุ่มตัวอักษรที่มีขนาด 16 – 20 ตัวอักษรนั้นมีจำนวนที่น้อย อาจจะเป็นสาเหตุหนึ่งที่ทำให้โมเดลคืนค่าความถูกต้องออกมาได้น้อยกว่ากลุ่มตัวอักษรกลุ่มอื่น ๆ หากมีการเพิ่มชุดข้อมูลในกลุ่มนี้เพิ่มเข้าไปในข้อมูลชุดฝึก ก็จะทำให้โมเดลให้ผลการทดลองดียิ่งขึ้น

สำหรับการฝึกสอนโมเดล ในงานวิจัยนี้ได้นำเสนอการฝึกสอนโมเดลแบบสองขั้นตอน โดยใช้การเรียนรู้แบบสอนและการเรียนรู้แบบกึ่งสอน สำหรับการเรียนรู้แบบกึ่งสอนงานวิจัยนี้ได้เสนอเทคนิควิธีการฝึกสอนโดยใช้การผสมชุดข้อมูลระหว่างชุดข้อมูลที่มีฉลากกำกับ และชุดข้อมูลที่ไม่มีฉลากกำกับเข้าด้วยกัน อีกทั้งยังเสนอวิธีการปรับแต่งฟังก์ชันต้นทุน เพื่อทำให้การทำงานของโมเดลนั้นมีประสิทธิภาพมากยิ่งขึ้น จากผลการทดลองพบว่าโมเดลการเรียนรู้แบบสอนนั้นคืนค่าความถูกต้องเฉลี่ยที่ 91.59% และ 93.06% สำหรับโมเดลการเรียนรู้แบบกึ่งสอน และเมื่อเปรียบเทียบค่าความถูกต้องเฉลี่ยกับงานวิจัยเดิมของศิริพจน์ สุรบถโสภณ พบว่าโมเดลการเรียนรู้แบบสอนมีค่าความถูกต้องเฉลี่ยมากกว่า 5.91% และ 7.38% สำหรับโมเดลการเรียนรู้แบบกึ่งสอน ในส่วนของตัววัด F1 ของงานวิจัยนี้อยู่ที่ 86.63% สำหรับชุดข้อมูลคำไทยและคำอังกฤษทับศัพท์คำไทย และ 82.51% ในกรณีคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ



## 7.2 ข้อเสนอแนะ

7.2.1 คำไทยที่มีการใช้สระลดรูป ได้แก่ สระอะลดรูป สระโอลดรูป และสระออลดรูป ถ้าหากสามารถสร้างขั้นตอนวิธีขึ้นมาเพื่อจัดการกับปัญหานี้ได้ จะทำให้ค่าความถูกต้องนั่นมีค่าเพิ่มมากยิ่งขึ้น

7.2.2 การเพิ่มจำนวนข้อมูล และข้อมูลที่มีความหลาหลายมากขึ้น สำหรับใช้ในการฝึกสอนจะทำให้ผลการทดลองนี้ดียิ่งขึ้น

7.2.3 ถ้าหากมีการปรับปรุงฟังก์ชันต้นทุน โดยการนำเทคนิคระยะการแก้ไขสั้นที่สุดมาใช้ในการคำนวณค่าต้นทุนของคู่คำในชุดข้อมูลที่ไม่มีฉลากกำกับ อาจจะทำให้โมเดลมีประสิทธิภาพมากยิ่งขึ้น

# บรรณานุกรม


1. Oard, D.W. and B.J. Dorr, *A survey of multilingual text retrieval*. Computer science technical report series UMIACS-TR-96-19 CS-TR-3615. 1996, College Park: University of Maryland.
2. Knight, K. and J. Graehl. *Machine Transliteration*. 1997. Madrid, Spain: Annual Meeting of the Association for Computational Linguistics (ACL-97/EACL-97).
3. Vaswani, A., et al., *Attention is all you need.* In Proc. of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA., 2017.
4. อุไรรัตน์ บุญภานนท์, การถอดอักษรภาษาอังกฤษเป็นไทยโดยใช้หลักภาษาศาสตร์ วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาภาษาศาสตร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย. 2526, จุฬาลงกรณ์มหาวิทยาลัย.
5. Frakes, W.B., and Baeza Yates, R., *Information Retrieval : Data Structures & Algorithms*. 1992, Englewood Cliffs, N.J.: Prentice Hall.
6. Blair, D.C., *Information Retrieval, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; 1979: 208 pp.* Journal of the American Society for Information Science, 1979. **30**(6): p. 374-375.
7. Zobel, J. and P. Dart, *Phonetic String Matching: Lessons from Information Retrieval.* In Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996.
8. Binstock, A. and J. Rex. *Practical algorithms for programmers*. 1995. New York: Addison Wesley.
9. วรรณี อุดมพาณิชย์, การใช้หลักคำพ้องเสียง เพื่อค้นหาชุดอักขระภาษาไทยที่ออกเสียงเหมือนกัน. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาภาษาศาสตร์ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิยาลัย. 2526, จุฬาลงกรณ์มหาวิยาลัย.
10. Suwanvisat, P. and S. Prasitjutrakul. *Thai-English Cross-Language Transliterated Word Retrieval using Soundex Technique*. in *Proc. of the National Computer Science and Engineering Conference 1998*. 1998. Bangkok, Thailand.
11. Suwanvisat, P. and S. Prasitjutrakul. *Transliterated Word Encoding and Retrieval Algorithms for Thai–English Cross–Language Retrieval*. in *In Proc. of the National*

*Computer Science and Engineering Conference 1999*. 1999. Bangkok Thailand.

12. ทัศนวรรณ ศูนย์กลาง, สมชาย ประสิทธิ์จูตระกูล และบุญเสริม กิจศิริกุล. เข้ารหัสคำทับศัพท์ด้วยเทคนิคนิวรอลเน็ตเวิร์ก เพื่อการค้นคืนข้ามภาษา. ในรายงานการประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ (*NCSEC 2000)*. กรุงเทพฯ, ประเทศไทย.
13. ศิริพจน์ สุรบถโสภณ และ บุญเสริม กิจศิริกุล. การค้นคืนข้ามภาษาสำหรับคำทับศัพท์ภาษาไทย/อังกฤษด้วยวิธีการนิวรอลเน็ตเวิร์ก แบบจำลองฮิดเด็นมาร์คอฟ และขั้นตอนวิธีเชิงพันธุกรรม. ในรายงานประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ *2546 (NCSEC 2003)*. 2546. ชลบุรี, ประเทศไทย.
14. Supawan, T., P. Wimonsiri, and R. Sittan. *Romanized Thai Input Method Editor*. in *Proc. of 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. 2017.
15. หลักเกณฑ์การถอดอักษรไทยเป็นอักษรโรมันแบบถ่ายเสียง. 2542: กรุงเทพมหานคร: ราชบัณฑิตยสถาน.
16. หลักเกณฑ์การทับศัพท์ภาษาอังกฤษ ฉบับราชบัณฑิตยสถาน. 2532: กรุงเทพมหานคร: ราชบัณฑิตยสถาน.
17. พระยา อุปกิตศิลปสาร, หลักภาษาไทย. พิมพ์ครั้งที่ *11*. 2545, กรุงเทพมหานคร: สำนักพิมพ์ไทยวัฒนาพานิช.
18. Thattinaphanich, S. and S. Prom-on. *Thai Named Entity Recognition Using Bi-LSTM-CRF with Word and Character Representation*. in *2019 4th International Conference on Information Technology (InCIT)*. 2019.

# ภาคผนวก

# ภาคผนวก ก
# การใช้อักษรโรมันแทนอักขระไทย

*ตารางที่ 29 การใช้อักษรโรมันแทนพยัญชนะไทยของราชบัณฑิตยสถาน*

| พยัญชนะไทย | อักษรโรมัน | |
|---|---|---|
| | พยัญชนะต้น | ตัวสะกด |
| ก | K | K |
| ข | KH | K |
| ฃ | KH | K |
| ค | KH | K |
| ฅ | KH | K |
| ฆ | KH | K |
| ง | NG | NG |
| จ | CH | T |
| ฉ | CH | T |
| ช | CH | T |
| ซ | S | T |
| ฌ | CH | T |
| ญ | Y | N |
| ฎ | D | T |
| ฏ | T | T |
| ฐ | TH | T |
| ฑ | D | T |
| ฒ | TH | T |
| ณ | TH | T |
| ณ | N | N |
| ด | D | T |
| ต | T | T |
| ถ | TH | T |

| พยัญชนะไทย | อักษรโรมัน | |
|---|---|---|
| | พยัญชนะต้น | ตัวสะกด |
| ธ | TH | T |
| น | N | N |
| บ | B | P |
| ป | P | P |
| ผ | PH | P |
| ฝ | F | P |
| พ | PH | P |
| ฟ | F | P |
| ภ | PH | P |
| ม | M | M |
| ย | Y | - |
| ร | R | N |
| ฤ | R | R |
| ล | L | N |
| ฦ | L | L |
| ว | W | - |
| ศ | S | T |
| ษ | S | T |
| ส | S | T |
| ห | H | - |
| ฬ | L | N |
| อ | H | - |
| ฮ | H | - |

*ตารางที่ 29 การใช้อักษรโรมันแทนพยัญชนะไทยของราชบัณฑิตยสถาน (ต่อ)*

| พยัญชนะไทย | อักษรโรมัน | |
|---|---|---|
| | พยัญชนะต้น | ตัวสะกด |
| ท | TH | T |

| พยัญชนะไทย | อักษรโรมัน | |
|---|---|---|
| | พยัญชนะต้น | ตัวสะกด |
| ทร | S | T |

*ตารางที่ 30 การใช้อักษรโรมันแทนสระไทยของราชบัณฑิตยสถาน*

| สระไทย | อักษรโรมัน |
|---|---|
| ะ ั า | A |
| ำ | AM |
| ิ ี ิย | I |
| ึ ื ือ ุ ู | U |
| เ-ะ เ็ เ- | E |
| แ-ะ แ็ แ- | AE |
| โ-ะ โ- เ-าะ ็อ- -อ อ- | O |
| เ-อะ เ-อ เ ิ | OE |
| เ ียะ เ ีย | IA |
| เ ือะ เ ือ ัวะ ัว -ว- | UA |
| ใ- ไ- ไ-ย ัย -าย | AI |
| เ-า -าว | AO |
| ุย ียี | UI |
| โ-ย -อย | OI |
| ิว | IU |
| เ ็ว เ-ว | EO |
| เ-ย | OEI |
| เ ือย -วย | UAI |
| แ-ว | AEU |
| เ ียว | IEU |

# ภาคผนวก ข
# หน่วยเสียงในภาษาไทยและภาษาอังกฤษ

## หน่วยเสียงในภาษาไทย

ภาษาไทยมีหน่วยเสียงพยัญชนะ 21 หน่วยเสียง ดังตารางที่ 31 หน่วยเสียงสระ 21 หน่วยเสียง ดังตารางที่ 32 และหน่วยเสียงวรรณยุกต์ 5 หน่วยเสียงได้แก่ สามัญ เอก โท ตรี จัตวา

*ตารางที่ 31 หน่วยเสียงพยัญชนะในภาษาไทย*

| | | | |
|---|---|---|---|
| ก | ช ศ ษ ส ทร | ณ น หน | ม หม |
| ข ฃ ค ฅ ฆ | ญ ย หย หญ | บ | ร |
| ง หง | ฎ ด ฑ | ป | ล ฬ หล |
| จ จร | ฏ ต | ผ พ ภ | ว หว |
| ฉ ช ฌ | ฐ ฑ ถ ท ธ | ฝ ฟ | ห ฮ |
| | | | อ |

*ตารางที่ 32 หน่วยเสียงพยัญสระในภาษาไทย*

| | | | | | |
|---|---|---|---|---|---|
| อิ | แอะ | เออะ | อุ | เอาะ | อัวะ อัว |
| อี | แอ | เออ | อู | ออ | |
| เอะ | อึ | อะ | โอะ | เอียะ เอีย | |
| เอ | อื | อา | โอ | เอือะ เอือ | |

**ระบบเสียงในภาษาอังกฤษ**

ภาษาอังกฤษมีหน่วยเสียงพยัญชนะ 24 หน่วยเสียง ดังตารางที่ 33 และหน่วยเสียงสระ 20 หน่วยเสียง ดังตารางที่ 34

*ตารางที่ 33 หน่วยเสียงพยัญชนะในภาษาอังกฤษ*

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

*ตารางที่ 34 หน่วยเสียงสระในภาษาอังกฤษ*

| เสียง | คำตัวอย่าง | เสียง | คำตัวอย่าง | เสียง | คำตัวอย่าง |
|---|---|---|---|---|---|
| / i: / | see | / ʊ / | put | / aI / | tie |
| / I / | sit | / u: / | too | / aʊ / | now |
| / e / | ten | / ʌ / | cup | / ɔI / | join |
| / æ / | hat | / ɜ: / | fur | / Iə / | near |
| / a: / | arm | / ə / | ago | / eə / | hair |
| / o / | got | / eI / | page | / ʊə / | tour |
| / ɔ: / | saw | / əʊ / | home | | |

# ภาคผนวก ค
# ตัวอย่างข้อมูลคำทับศัพท์ที่ใช้ในงานวิจัย

## ตัวอย่างคำอังกฤษและคำไทยทับศัพท์คำอังกฤษ

| | | | | | |
|---|---|---|---|---|---|
| liberty | ลิเบอร์ตี | gyro | ไจโร | gowland | เกาว์แลนด์ |
| europe | ยุโรป | berkelium | เบอร์คีเลียม | anderson | แอนเดอร์สัน |
| aurora | ออโรรา | juice | จูซ | frisch | ฟรีซ |
| cytosol | ไซโทซอล | lothar | โลทาร์ | bowman | โบว์แมน |
| playfair | เพลย์แฟร์ | collenchyma | คอลเลงคิมา | helium | ฮีเลียม |
| zeta | ซีตา | boson | โบซอน | stokes | สโตกส์ |
| iodide | ไอโอไดด์ | einstein | ไอน์สไตน์ | marconi | มาร์โคนี |
| okhotsk | โอค็อตสก์ | chrysler | ไครส์เลอร์ | warren | วาร์เรน |
| maclagan | แมกลาเกน | gaussian | เกาส์เซียน | delta | เดลต้า |
| rye | ไรย์ | mil | มิล | allantois | แอลแลนทอยส์ |
| arc | อาร์ก | broadway | บรอดเวย์ | sikh | ซิก |
| mozambique | โมซัมบิก | thomson | ทอมสัน | hankel | ฮันเกล |
| suzuki | ซูซูกิ | romer | เรอเมอร์ | micelle | ไมเซลล์ |
| lantis | แลนทิส | peta | เพตะ | factor | แฟกเตอร์ |
| dewey | ดิวอีย์ | klein | ไคลน์ | zygomata | ไซโกมาตา |
| cotangent | โคแทนเจนต์ | manganese | แมงกานีส | karl | คาร์ล |
| cosecant | โคเซแคนต์ | golf | กอล์ฟ | bract | แบร็กต์ |
| chromosphere | โครโมสเฟียร์ | harsh | ฮาร์ช | joule | จูล |

**ตัวอย่างคำไทยและคำอังกฤษทับศัพท์คำไทย**

| | | | | | |
|---|---|---|---|---|---|
| weerachai | วีรชัย | kijluakiat | กิจลือเกียรติ | worawas | วรวัสส์ |
| plianrungsi | เปลี่ยนรังษี | puttakrong | พุทธกรอง | tongchai | ทองชัย |
| vilaiphan | วิไลพันธุ์ | rossukon | รสสุคนธ์ | wasana | วาสนา |
| sukhaboon | สุขาบูรณ์ | mitisubin | มิติสุบิน | achaporn | อัชพร |
| prangsiri | ปรางศิริ | veerasak | วีรศักดิ์ | poonpissamai | พูนพิศมัย |
| puengrusme | พึงรัศมี | kanokrat | กนกรัตน์ | wisanupong | วิษณุพงษ์ |
| intapuntee | อินทปันติ | sarakit | ศารกิจ | sapatporn | สภัทร์พร |
| onumpai | อรอำไพ | ekarat | เอกรัฐ | silarujisun | ศิลารุจิสรรค์ |
| muendech | หมื่นเดช | siriphap | ศิริภาพ | wilailak | วิไลลักษณ์ |
| chatkaew | ฉัตรแก้ว | khrongwong | ครองวงศ์ | wichian | วิเชียร |
| ruethai | ฤทัย | rapeephan | รพีพรรณ | marianukroh | มารีอนุเคราะห์ |
| ratree | ราตรี | oranuch | อรนุช | warit | วริษฐ์ |
| jittima | จิตติมา | sawart | สวาท | pisithsak | พิสิษฐ์ศักดิ์ |
| phamornthep | ภมรเทพ | tiawsirisup | เตียวศิริทรัพย์ | laosunthara | เหล่าสุนทร |
| saowarat | เสาวรัตน์ | charanvas | จรัญวาศน์ | suproongruing | ทรัพย์รุ่งเรือง |
| manoon | มนูญ | kanitta | ขนิษฐา | nattaphan | ณัฐพันธุ์ |
| keng | เก่ง | phongphew | ผ่องแผ้ว | somjate | สมเจตน์ |
| limlawan | ลิมป์ลาวัณย์ | weerasak | วีระศักดิ์ | leelawat | ลีละวัฒน์ |
| sudarat | สุดารัตน์ | chumpot | จุมภฏ | kuankid | ควรคิด |
| suthip | สุทิพย์ | charinee | ฌาริณีย์ | pongput | ผองพุทธ |

# ภาคผนวก ง
# ค่าพารามิเตอร์ที่ดีที่สุดใช้ในการทดลอง

*ตารางที่ 35 ค่าพารามิเตอร์ที่ดีที่สุดใช้ในการทดลอง*

| พารามิเตอร์ | ค่า |
|---|---|
| Optimizer | Adam |
| Batch size | 128 |
| Epoch | 10 |
| Learning rate | 0.005 |
| Embedding dimension | 32 |
| Encoder layer number | 2 |
| Decoder layer number | 2 |
| Head number | 4 |
| Hidden dimension | 128 |
| Dropout rate | 0.3 |

# ประวัติผู้เขียน

| | |
|---|---|
| **ชื่อ-สกุล** | นายอภิชัจจ์ โชดกวณิชย์ |
| **วัน เดือน ปี เกิด** | 6 กุมภาพันธ์ 2535 |
| **สถานที่เกิด** | ชุมพร |
| **วุฒิการศึกษา** | ปริญญาตรีวิศวกรรมศาสตร์ สาขาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยธุรกิจบัณฑิตย์ |
| **ที่อยู่ปัจจุบัน** | 98/5 หมู่ 4 หมู่บ้านชัยพฤกษ์ เวสต์เกต ถนนสายบึงบัว - คลองประปา ตำบลบางแม่นาง อำเภอบางใหญ่ จังหวัดนนทบุรี 11140 |
| **ผลงานตีพิมพ์** | บทความเรื่อง การค้นคืนข้ามภาษาสำหรับคำทับศัพท์ภาษาไทย/อังกฤษโดยใช้ทรานฟอร์เมอร์ (Thai/English cross-language transliterated word retrieval using Transformer) ได้รับการตีพิมพ์ในงาน 3rd International Conference on Natural Language Processing (ICNLP 2021) ที่กรุงปักกิ่ง ประเทศจีน เมื่อวันที่ 26-28 มีนาคม 2564 |