การตรวจแก้การสะกดผิดแบบเป็นคำจริงในภาษาไทยโดยใช้แบบจำลองไตรแกรม

นายพลวัฒน์ ไหลมนู

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทคัดย่อและแฟ้มข้อมูลฉบับเต็มของวิทยานิพนธ์ตั้งแต่ปีการศึกษา 2554 ที่ให้บริการในคลังปัญญาจุฬาฯ (CUIR)
เป็นแฟ้มข้อมูลของนิสิตเจ้าของวิทยานิพนธ์ ที่ส่งผ่านทางบัณฑิตวิทยาลัย
The abstract and full text of theses from the academic year 2011 in Chulalongkorn University Intellectual Repository (CUIR)
are the thesis authors' files submitted through the University Graduate School.

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2559
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

THAI REAL-WORD SPELLING ERROR CORRECTION USING A TRIGRAM MODEL

Mr. Ponlawat Laimanoo

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Arts Program in Linguistics

Department of Linguistics

Faculty of Arts

Chulalongkorn University

Academic Year 2016

Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	การตรวจแก้การสะกดผิดแบบเป็นคำจริงในภาษาไทยโดยใช้แบบจำลองไตรแกรม
โดย	นายพลวัฒน์ ไหลมนู
สาขาวิชา	ภาษาศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะอักษรศาสตร์

(รองศาสตราจารย์ ดร.กิ่งกาญจน์ เทพกาญจนา)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(ดร. วรรณชัย คำภีระ)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล)

.....กรรมการภายนอกมหาวิทยาลัย

(ดร. เทพชัย ทรัพย์นิธิ)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

พลวัฒน์ ไหลมนู : การตรวจแก้การสะกดผิดแบบเป็นคำจริงในภาษาไทยโดยใช้แบบจำลองไตรแกรม (THAI REAL-WORD SPELLING ERROR CORRECTION USING A TRIGRAM MODEL) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร. วิโรจน์ อรุณมานะกุล, 80 หน้า.

งานวิจัยนี้มีวัตถุประสงค์เพื่อรวบรวมและวิเคราะห์การสะกดผิดแบบเป็นคำจริงในภาษาไทยที่พบบนอินเทอร์เน็ต พร้อมกับพัฒนาระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงในภาษาไทยด้วยแบบจำลองไตรแกรมและประเมินประสิทธิภาพของระบบที่พัฒนาขึ้น

งานวิจัยนี้แบ่งออกเป็นสองส่วน ส่วนแรกเป็นการวิเคราะห์การสะกดผิดแบบเป็นคำจริงในภาษาไทยจำนวน 1,674 คำ จากหนังสือคำไทยที่มักเขียนผิดรวบรวมโดยผู้เชี่ยวชาญภาษาไทย ซึ่งทุกคำล้วนผ่านการตัดคำสำเร็จและพบตัวอย่างการใช้จริงบนอินเทอร์เน็ต จากการวิเคราะห์พบว่าคำที่สะกดผิดเหล่านี้ส่วนใหญ่หรือร้อยละ 80 เป็นคำที่สะกดผิดหนึ่งตำแหน่งซึ่งมักจะสะกดผิดที่พยัญชนะต้นมากที่สุด และส่วนที่เหลืออีก 20% เป็นคำที่สะกดผิดหลายตำแหน่งและส่วนใหญ่จะยังออกเสียงเหมือนเดิม ในส่วนที่สองเป็นการพัฒนาระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงในภาษาไทยด้วยแบบจำลองไตรแกรมพร้อมกับประเมินประสิทธิภาพของระบบ ข้อมูลที่นำมาใช้ทดสอบเป็นข้อความที่มีคำสะกดผิดอยู่อย่างน้อยหนึ่งคำและคำนั้นจะต้องเป็นคำที่สะกดผิดแบบเป็นคำจริง จำนวน 1,000 ข้อความ ซึ่งระบบจะทำการตรวจจับคำที่สะกดผิดทั้งหมดในข้อความโดยนำสายคำเรียงสามแต่ละสายของข้อความเทียบกับคลังข้อมูลไตรแกรม หากไม่พบแสดงว่าสายคำเรียงสามนั้นต้องสงสัยว่าสะกดผิด โดยสายคำเรียงสามที่ต้องสงสัยทั้งหมดจะถูกนำไปปรับแก้ด้วยวิธีการปรับแก้น้อยสุด จากนั้นสายเรียงสามคำที่ถูกปรับแก้แล้วจะถูกนำไปแทนที่การสะกดผิดเดิมแล้วคำนวณหาค่าความน่าจะเป็นของข้อความ ซึ่งระบบจะเลือกสายคำเรียงสามที่ให้ค่าความน่าจะเป็นของข้อความสูงสุดมาใช้แก้ไขการสะกดผิด ผู้วิจัยได้ประเมินประสิทธิภาพของระบบในสามด้าน ได้แก่ ด้านระยะเวลาในการประมวลผล พบว่าระบบแบบจำลองไตรแกรมใช้เวลาในการประมวลผลทั้งหมด 128 วินาที ด้านประสิทธิภาพในการตรวจจับคำที่สะกดผิดแบบเป็นคำจริงในภาษาไทยพบว่ามีค่าความแม่นยำ (precision) และค่าความครบถ้วน (recall) เท่ากัน คือ 0.47 ส่วนด้านประสิทธิภาพในการแก้ไขคำที่สะกดผิดแบบเป็นคำจริงในภาษาไทยพบว่ามีค่าความครบถ้วนและค่าความแม่นยำอยู่ที่ 0.85

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ภาควิชา ภาษาศาสตร์ ลายมือชื่อนิสิต .....

สาขาวิชา ภาษาศาสตร์ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2559

5680132222 : MAJOR LINGUISTICS
KEYWORDS: REAL-WORD ERROR / SPELLING CORRECTION / TRIGRAM MODEL / THAI REAL WORD ERROR / THAI SPELLING CORRECTION

PONLAWAT LAIMANOO: THAI REAL-WORD SPELLING ERROR CORRECTION USING A TRIGRAM MODEL. ADVISOR: ASSOC. PROF. WIROTE AROONMANAKUN, Ph.D., 80 pp.

This research aims to collect and analyze Thai real-word spelling errors found on the internet, develop a Thai real-word error spelling correction program using a trigram model, and evaluate its performance.

This research consists of two parts; first is an analysis of 1,674 Thai real-word spelling errors found in ‘Thai often misspelled words’ books. It is found that 80 percent of these analyzed errors contain only one spelling error which mostly occurs at word initial position. The other 20 percent of the analyzed errors have more than one spelling errors individually, most of which are pronounced the same. The latter part of the research is about developing a Thai real-word spelling error correction program using a trigram model and evaluating its performance. The test data are 1,000 Thai strings of words. Each contains at least one real-word spelling error. To detect a real-word error, word trigrams of each string are checked with the trigram corpus and those which do not exist in the corpus are considered misspelling suspects. Then, all misspelling suspects are edited to generate possible candidates. Only one candidate, that gives the highest probability of the observed string when replacing the detected error, is chosen as the correct one. The program’s efficiency is measured in three aspects. First is the processing duration. The program’s execution takes 128 seconds to finish. Second is the error detection efficiency. It is found that the values of precision and recall are similar, which is 0.47. Last is the error correction efficiency. The program’s values of precision and recall are also equal, which is 0.85.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Department: Linguistics Student's Signature .....
Field of Study: Linguistics Advisor's Signature .....
Academic Year: 2016

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณ รองศาสตราจารย์ ดร. วิโรจน์ อรุณมานะกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์เป็นอย่างสูงที่ได้ให้ความรู้ ความช่วยเหลือ ความเมตตากรุณา คำแนะนำและข้อคิดเห็นต่างๆ อันเป็นประโยชน์อย่างยิ่งในการทำวิจัย อีกทั้งยังช่วยแก้ปัญหาต่างๆ ที่เกิดขึ้นระหว่างการทำวิจัย รวมถึงช่วยปรับแก้วิทยานิพนธ์ฉบับนี้ให้สำเร็จลุล่วงไปด้วยดี และผู้วิจัยขอขอบพระคุณอาจารย์ ดร.วรรณชัย คำภีระและดร. เทพชัย ทรัพย์นิธิ กรรมการสอบวิทยานิพนธ์เป็นอย่างสูงที่ได้ให้คำแนะนำต่างๆ ในการทำวิจัยและเสียสละเวลาเพื่อตรวจแก้วิทยานิพนธ์เล่มนี้ให้มีความสมบูรณ์มากยิ่งขึ้น

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. พิทยาวัฒน์ พิทยาภรณ์ ที่ให้โอกาสผู้วิจัยได้ช่วยทำงานวิจัยซึ่งเป็นการเพิ่มพูนความรู้และเสริมสร้างประสบการณ์การทำงานให้แก่ผู้วิจัย รวมถึงคณาจารย์ภาควิชาภาษาศาสตร์ทุกท่านที่ได้ประสิทธิ์ประสาทความรู้แก่ผู้วิจัย

สุดท้ายผู้วิจัยขอขอบพระคุณบิดาและมารดา คุณวิฑูรย์และคุณเสาวภาคย์ ไหลมนู ที่คอยให้ความรัก ความเป็นห่วงเป็นใยพร้อมกับคอยมอบกำลังใจและการสนับสนุนที่ดีมาโดยตลอด ขอขอบคุณพี่สาวและน้องชาย คุณพลอยไพฑูรย์และคุณภาคภูมิ ไหลมนู ที่ให้ความช่วยเหลือคำแนะนำและกำลังใจแก่ผู้วิจัย ขอขอบคุณ คุณวีรชัย อำพรไพบูลย์และคุณนัชชา ถิระสาโรช ที่ช่วยเหลือผู้วิจัยในด้านการเขียนโปรแกรมคอมพิวเตอร์ รวมถึงขอขอบคุณเพื่อนๆ พี่ๆ น้องๆ และเจ้าหน้าที่ภาควิชาภาษาศาสตร์สำหรับการสนับสนุน กำลังใจและความช่วยเหลือในด้านต่างๆ ที่มอบให้แก่ผู้วิจัยตลอดมา

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญตาราง 11

สารบัญรูปภาพ 12

บทที่ 1 บทนำ 13

1.1 ที่มาและความสำคัญของปัญหา 13

1.2 วัตถุประสงค์ของการวิจัย 15

1.3 สมมติฐานของการวิจัย 15

1.4 ขอบเขตของการวิจัย 15

1.5 ประโยชน์ที่คาดว่าจะได้รับ 15

1.6 วิธีดำเนินการวิจัย 16

1.7 เครื่องมือที่ใช้ในการวิจัย 16

บทที่ 2 ทบทวนวรรณกรรม 17

2.1 หลักการทำงานเบื้องต้นของการตรวจแก้การสะกดผิด 17

2.1.1 การตรวจจับการสะกดผิด (Error Detection) 17

2.1.2 การแก้ไขการสะกดผิด (Error Correction) 18

2.2 ประเภทของการสะกดผิด 19

2.2.1 การสะกดผิดแบบไม่เป็นคำ (non-word spelling errors) 19

2.2.2 การสะกดผิดแบบเป็นคำจริง (real-word spelling errors) 20

2.3 ประเด็นและข้อเท็จจริงต่างๆ ที่น่าสนใจเกี่ยวกับการสะกดผิด 20

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

หน้า

2.4 วิธีการต่างๆ ที่มีการนำมาใช้ในงานด้านการตรวจแก้การสะกดผิดด้วยระบบคอมพิวเตอร์ 23
2.4.1 การตรวจจับการสะกดผิด (spelling errors detection) 24
2.4.2 การแก้ไขการสะกดผิด (spelling errors correction) 26
บทที่ 3 การจัดเตรียมคลังข้อมูล 31
3.1 คลังข้อมูลไตรแกรมคำ (word-trigram corpus) 31
3.1.1 การเตรียมคลังข้อมูลไตรแกรมคำ 32
3.1.2 ลักษณะโครงสร้างของคลังข้อมูลไตรแกรมคำ 33
3.2 คลังข้อมูลยูนิแกรมคำ (word-unigram corpus) 34
3.2.1 การเตรียมคลังข้อมูลยูนิแกรมคำ 34
3.2.2 ลักษณะโครงสร้างของคลังข้อมูลยูนิแกรมคำ 34
3.3 คลังข้อมูลชุดคำสับสน (confusion set corpus) 35
3.3.1 การเตรียมคลังข้อมูลชุดคำสับสน 35
3.3.2 ลักษณะโครงสร้างของคลังข้อมูลชุดคำสับสน 36
3.3.3 ข้อมูลฝึกฝนและทดสอบ (training and test data) 37
3.3.3.1 การเตรียมข้อมูลฝึกฝน (training data) 37
3.3.3.2 การเตรียมข้อมูลทดสอบ (test data) 37
3.3.3.3 ลักษณะโครงสร้างของข้อมูลฝึกฝนและทดสอบ 38
บทที่ 4 การวิเคราะห์คำไทยที่มักเขียนผิด 39
4.1 ขั้นตอนการวิเคราะห์คำไทยที่มักเขียนผิด 39
4.2 รูปแบบการสะกดผิด 41
4.2.1 คำที่สะกดผิดหนึ่งตำแหน่ง 41
4.2.1.1 คำสะกดผิดที่พยัญชนะต้น 41
4.2.1.2 คำสะกดผิดที่ตัวสะกด 42

หน้า

4.2.1.3 คำสะกดผิดที่สระ 43
4.2.1.4 คำสะกดผิดที่ตัวการันต์ 44
4.2.1.5 คำสะกดผิดที่วรรณยุกต์ 45
4.2.2 คำที่สะกดผิดหลายตำแหน่ง 46
4.2.2.1 คำที่สะกดผิดหลายตำแหน่งแต่ออกเสียงเหมือนเดิม 46
4.2.2.2 คำที่สะกดผิดหลายตำแหน่งและออกเสียงเปลี่ยนไป 47
บทที่ 5 การตรวจแก้การสะกดผิดแบบเป็นคำจริง 51
5.1 ข้อมูลที่ใช้ทดสอบ (test data) 51
5.2 การตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรม 51
5.2.1 หลักการทำงานของระบบตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรม 52
5.2.1.1 ขั้นตอนที่หนึ่ง: ตรวจจับคำที่ต้องสงสัยในข้อความ 52
5.2.1.2 ขั้นตอนที่สอง: ปรับแก้คำที่ต้องสงสัย 53
5.2.1.3 ขั้นตอนที่สาม: เลือกคำที่เหมาะสมเพื่อใช้แก้ไขคำที่สะกดผิด 57
5.2.2 การประเมินประสิทธิภาพการตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรม 59
5.2.2.1 ประสิทธิภาพในการตรวจจับการสะกดผิดด้วยแบบจำลองไตรแกรม 60
5.3 การตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองยูนิแกรม 61
5.3.1 หลักการทำงานของระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรม 62
5.3.1.1 ขั้นตอนการเลือกคำที่ถูกต้องเพื่อใช้แก้ไขการสะกดผิด 62
5.3.2 ประสิทธิภาพการตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรม 63
5.4 การตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยชุดคำสับสน 64
5.4.1 หลักการทำงานของระบบตรวจแก้การสะกดผิดด้วยชุดคำสับสน 64
5.4.2 ประสิทธิภาพการตรวจแก้การสะกดผิดด้วยชุดคำสับสน 66
5.5 ผลการเปรียบเทียบประสิทธิภาพในการตรวจแก้การสะกดผิดของระบบทั้งสาม 67

หน้า

5.5.1 ด้านระยะเวลาที่ใช้ในการประมวลผล 67
5.5.2 ด้านการตรวจจับการสะกดผิดแบบเป็นคำจริง 68
5.5.3 ด้านการแก้ไขการสะกดผิดแบบเป็นคำจริง 69
บทที่ 6 สรุปผลการวิจัย ปัญหาและข้อเสนอแนะ 71
6.1 สรุปผลการวิจัย 71
6.1.1 สรุปผลการศึกษาวิเคราะห์คำไทยที่มักสะกดผิด 71
6.1.2 สรุปผลการทดสอบประสิทธิภาพในการทำงานของระบบตรวจแก้การสะกดผิดแบบเป็นคำจริง 72
6.2 ปัญหาที่พบ 74
6.2.1 ปัญหาด้านระบบ 75
6.2.2 ปัญหาด้านข้อมูล 75
6.3 ข้อเสนอแนะ 75
รายการอ้างอิง 77
ประวัติผู้เขียนวิทยานิพนธ์ 80

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญตาราง

หน้า

ตารางที่ 4.1แสดงตัวอย่างข้อมูลการสะกดผิดแบบเป็นคำจริงในภาษาไทยที่นำมาวิเคราะห์ 40

ตารางที่ 5.1 แสดงตัวอย่างรูปแบบที่นำไปใช้ปรับแก้คำที่ต้องสงสัยว่าสะกดผิด 53

ตารางที่ 5.2 แสดงตัวอย่างคำสะกดผิดที่มีจำนวนแกรมเปลี่ยนไปเมื่อได้รับการปรับแก้..... 55

ตารางที่ 5.3 แสดงค่าประสิทธิภาพการตรวจจับการสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรม..... 60

ตารางที่ 5.4 แสดงค่าประสิทธิภาพในการเลือกคำมาแก้ไขการสะกดผิดได้ถูกต้องด้วยแบบจำลองยูนิแกรม.... 64

ตารางที่ 5.5 แสดงประสิทธิภาพในการตรวจจับการสะกดผิดด้วยชุดคำสับสน 66

ตารางที่ 5.6 แสดงผลการเปรียบเทียบระยะเวลาที่ใช้ในการประมวลของแต่ละระบบ 68

ตารางที่ 5.7 แสดงค่าประสิทธิภาพในการตรวจจับการสะกดผิดด้วยระบบที่แตกต่างกัน 68

ตารางที่ 5.8 แสดงค่าประสิทธิภาพในการแก้ไขการสะกดผิดด้วยระบบที่แตกต่างกัน 69

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญรูปภาพ

หน้า

รูปภาพที่ 3.1 แสดงขั้นตอนในการเตรียมคลังข้อมูลไตรแกรมและคลังข้อมูลยูนิแกรม 35
รูปภาพที่ 3.2 แสดงขั้นตอนในการเตรียมคลังข้อมูลชุดคำสับสน 36
รูปภาพที่ 4.1 ประเภทของการสะกดผิดจำแนกตามจำนวนการสะกดผิดที่พบในหนึ่งคำ 47
รูปภาพที่ 4.2 แผนภูมิแท่งแสดงคำที่สะกดผิดหนึ่งตำแหน่งประเภทต่างๆ 48
รูปภาพที่ 4.3 แผนภูมิแท่งแสดงคำที่สะกดผิดหลายตำแหน่งสองประเภท 49
รูปภาพที่ 4.4 แผนภูมิแท่งแสดงคำที่สะกดผิดประเภทต่างๆ ที่พบในการวิเคราะห์ครั้งนี้ 49
รูปภาพที่ 5.1 แสดงขั้นตอนในการตรวจจับคำที่ต้องสงสัยในข้อความ 52
รูปภาพที่ 5.2 แสดงขั้นตอนการปรับแก้คำที่ต้องสงสัย 56
รูปภาพที่ 5.3 แสดงการเลือกคำที่เหมาะสมเพื่อแก้ไขการสะกดผิดในข้อความ 58
รูปภาพที่ 5.4 กระบวนการตรวจแก้การสะกดผิดด้วยชุดคำสับสน 65

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันผู้คนส่วนใหญ่ให้ความสำคัญกับโลกออนไลน์และสื่อโซเชียลเน็ตเวิร์คเป็นอย่างมาก ทำให้มีการสื่อสารผ่านตัวอักษรด้วยการพิมพ์มากขึ้น ซึ่งข้อความที่พิมพ์นั้นควรจะมีการสะกดคำที่ถูกต้องตามหลักภาษาเพราะจะส่งผลต่อความหมายของข้อความที่สื่อสาร ถ้าหากว่าข้อความนั้นสะกดได้ถูกต้องตามหลักภาษาก็จะทำให้ข้อความนั้นสื่อความหมายตรงตามที่ผู้พิมพ์ต้องการ แต่ในทางกลับกัน ถ้าข้อความนั้นมีคำที่สะกดผิดปะปนอยู่ก็อาจจะทำให้ข้อความนั้นไม่สื่อความหมายใดๆ หรือสื่อความหมายที่ผิดเพี้ยนไปจากที่ผู้พิมพ์ต้องการก็เป็นได้ ด้วยเหตุนี้ระบบตรวจแก้การสะกดคำจึงมีบทบาทสำคัญในการช่วยตรวจสอบความถูกต้องและแก้ไขข้อผิดพลาดด้านการเขียนสะกดคำที่อาจเกิดขึ้น เพื่อช่วยทำให้การสื่อสารผ่านทางตัวอักษรนั้นมีประสิทธิภาพสูงสุด

งานทางด้านการตรวจแก้การสะกดคำนั้นเป็นหนึ่งในงานด้านการประมวลผลภาษา (Language Processing) ที่สำคัญมากงานหนึ่ง ซึ่งนักวิศวกรคอมพิวเตอร์ นักวิทยาศาสตร์คอมพิวเตอร์ และนักภาษาศาสตร์คอมพิวเตอร์ ต่างคิดค้นและพัฒนาระบบช่วยตรวจแก้การสะกดคำมาอย่างต่อเนื่องเพื่อช่วยตรวจหาและแก้ไขข้อผิดพลาดในการสะกดคำของข้อความก่อนที่จะถูกส่งต่อหรือเผยแพร่ให้ผู้อื่นได้รับรู้

ซึ่งข้อผิดพลาดในการสะกดคำสามารถแบ่งออกได้เป็น 2 ประเภท คือ

1. ข้อผิดพลาดในรูปของคำที่สะกดผิดแบบไม่เป็นคำ (non-word error) คือ คำในข้อความที่ผู้พิมพ์สะกดคำไม่ถูกต้องทำให้ได้รูปคำที่ไม่ตรงกับคำใดในภาษา ตัวอย่างเช่น “วันนี้**อท้ง**ฟ้าแจ่มใส” มีสายอักขระ **อท้ง** ซึ่งไม่ตรงคำใดในภาษาและต้องแก้ไขให้เป็นประโยคที่ถูกต้องคือ “วันนี้**ท้อง**ฟ้าแจ่มใส”

2. ข้อผิดพลาดในรูปของคำที่สะกดผิดแบบเป็นคำ (real-word error) คือ คำในข้อความที่ผู้พิมพ์สะกดผิดแต่ได้บังเอิญตรงกับรูปคำที่มีอยู่ในภาษา (Islam & Inkpen, 2009) ตัวอย่างเช่น ในภาษาไทย “น้ำ**ไส**จนเห็นตัวปลา” ผู้พิมพ์พิมพ์คำว่า ใส ผิดเป็น ไส แม้คำว่า ไส จะเป็นคำที่มีจริงในภาษาไทย แต่เมื่อให้ความหมายที่ผิดเพี้ยนไป ไม่สอดคล้องกับคำอื่นๆ ในประโยค ซึ่งต้องได้รับการแก้ไขให้เป็นประโยคที่ถูกต้องคือ“น้ำ**ใส**จนเห็นตัวปลา”

หากพิจารณาในเรื่องความยากง่ายในการตรวจจับการสะกดผิดแต่ละประเภท การตรวจจับข้อผิดพลาดประเภทแรกทำได้ง่ายกว่าเนื่องจากในข้อความมีสายอักขระที่ไม่เป็นคำอยู่และสามารถตรวจจับคำที่สะกดผิดโดยเทียบกับรายการคำในพจนานุกรมได้ ซึ่งสายอักขระใดที่ไม่ปรากฏในพจนานุกรมแสดงว่าสายอักขระนั้นมีการสะกดผิดแบบไม่เป็นคำ ในขณะที่ข้อผิดพลาดในการสะกดผิดประเภทที่สองจะตรวจจับได้ยากกว่า เนื่องจากต้องพิจารณาหาว่าคำใดที่ไม่ควรปรากฏร่วมกับคำอื่นๆ ในบริบทหรือประโยคหนึ่งๆ งานตรวจแก้การสะกดคำผิดจึงมักแยกเป็นงานสองประเภทขาดจากกันได้ คือ งานตรวจแก้การสะกดผิดแบบไม่เป็นคำ (non-word spelling error correction) และงานตรวจแก้การสะกดผิดแบบเป็นคำจริง (real-word spelling error correction)

ในกรณีของภาษาไทย หากพบข้อผิดพลาดในการสะกดประเภทแรก ระบบตัดคำภาษาไทยจะสามารถบ่งชี้ข้อผิดให้ได้ เพราะถ้ามีคำใดที่ระบบตัดคำตัดไม่ได้แสดงว่าคำๆ นั้นสะกดผิดแบบไม่เป็นคำ แต่ในกรณีข้อผิดพลาดในการสะกดประเภทที่สอง ระบบตัดคำภาษาไทยจะไม่สามารถบ่งชี้ข้อผิดพลาดในลักษณะเช่นนี้ได้ เพราะต้องอาศัยบริบทช่วยในการระบุการสะกดผิด

อย่างไรก็ตามสิ่งที่ยังคงเป็นปัญหาสำหรับการตรวจแก้การสะกดผิดในภาษาไทยนั้นก็คือการที่ยังไม่สามารถตรวจแก้การสะกดผิดแบบเป็นคำจริงได้เพราะเครื่องตรวจแก้การสะกดผิดที่ใช้กันอยู่ทั่วไปในขณะนี้จะยังไม่สามารถตรวจจับการสะกดผิดแบบเป็นคำจริงนี้ได้ เนื่องจากหลักการทำงานพื้นฐานของเครื่องมือเหล่านี้ก็คือการนำคำแต่ละคำในข้อความที่ผู้ใช้ป้อนให้ไปเปรียบเทียบกับคลังศัพท์พจนานุกรม ถ้าคำไหนตรงกับคำในคลังศัพท์พจนานุกรมก็จะถือว่าคำนั้นเป็นคำที่สะกดถูกต้อง ในทางกลับกันถ้าไม่พบคำที่ป้อนเข้าไปในคลังศัพท์พจนานุกรม คำนั้นก็จะถูกทำเครื่องหมายเพื่อบอกว่าเป็นคำที่สะกดไม่ถูกต้องแล้วให้ผู้ใช้ได้ดำเนินการปรับแก้ต่อไป ตัวอย่างเช่น คำว่า “**อรง**เท้า” โดยที่ “**อรง**” เป็นคำที่ไม่พบในคลังศัพท์พจนานุกรมของ Microsoft Word 2010 จึงได้รับการขีดเส้นหยักสีแดงใต้คำเพื่อบอกว่าเป็นคำที่สะกดผิดและให้ผู้ใช้กลับไปแก้ไขให้ถูกต้อง ซึ่งคำที่ถูกต้องก็คือ คำว่า “**รอง**เท้า” แต่ถ้าหากว่าผู้ใช้ต้องการจะพิมพ์คำว่า “รองเท้า” เช่นกันนี้แต่กลับพิมพ์ผิดเป็น “**นอง**เท้า”และ“**ยอง**เท้า” ดังในประโยค “รีบไปใส่**นอง**เท้าสิลูก” หรือ “**ยอง**เท้าคู่นี้สวยดี” เครื่องช่วยตรวจการสะกดคำใน Microsoft Word กลับไม่ทำเครื่องหมายใดๆ ไม่ว่าจะเป็นคำเดี่ยว (isolated word) หรือเมื่อคำนี้ปรากฏอยู่ในประโยค ด้วยสาเหตุที่ว่าคำทั้งสองเป็นคำที่สะกดผิดแบบเป็นคำจริง เพราะฉะนั้นการที่จะตรวจหาและแก้ไขคำที่สะกดผิดแบบเป็นคำจริงนี้จึงต้องอาศัยวิธีการที่ต่างไปจากการตรวจแก้การสะกดผิดแบบไม่เป็นคำเพื่อจัดการกับข้อผิดพลาดในลักษณะนี้ ด้วยเหตุนี้ผู้วิจัยจึงสนใจที่จะศึกษาและพัฒนาระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงเพื่อแก้ไขปัญหานี้

1.2 วัตถุประสงค์ของการวิจัย

1.2.1. เพื่อรวบรวมและวิเคราะห์การสะกดผิดแบบเป็นคำจริงที่พบบนอินเทอร์เน็ต

1.2.2. เพื่อพัฒนาระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงในภาษาไทยโดยใช้แบบจำลองไตรแกรม

1.2.3. เพื่อประเมินประสิทธิภาพของระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงที่พัฒนาขึ้น

1.3 สมมติฐานของการวิจัย

1.3.1 การสะกดผิดแบบเป็นคำจริงในภาษาไทยมักมีความผิดพลาดที่พยัญชนะสะกดและยังออกเสียงเหมือนเดิม

1.3.2 การตรวจแก้การสะกดผิดแบบเป็นคำจริงโดยใช้แบบจำลองไตรแกรมจะได้ผลถูกต้องมากกว่าการใช้แบบจำลองยูนิแกรม

1.3.3 การใช้เซตคำสับสนที่ถูกกำหนดไว้แล้ว (pre-defined confusion set) จะใช้เวลาในการประมวลผลน้อยกว่าการใช้เซตคำสับสนที่สร้างแบบอัตโนมัติด้วยวิธีการตรวจแก้น้อยสุด (minimum edit distance)

1.4 ขอบเขตของการวิจัย

การวิจัยครั้งนี้ศึกษาและพัฒนาระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงในภาษาไทย ซึ่งข้อมูลที่ใช้ในการฝึกฝนและทดสอบระบบนั้นเป็นข้อความภาษาไทยที่มีคำสะกดผิดแบบเป็นคำจริงปนอยู่จำนวน 4,787 ข้อความ และทุกข้อความต้องผ่านการตัดคำด้วยระบบตัดคำ ThaiSegmentation Version 2.2 (Aroonmanakul, 2002)

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ได้เครื่องมือพื้นฐานในการตรวจแก้การสะกดผิดที่จะนำไปประยุกต์ใช้กับงานประมวลผลภาษาไทยด้วยคอมพิวเตอร์ด้านอื่นๆ เช่น การแปลภาษาด้วยคอมพิวเตอร์ เป็นต้น

1.5.2 เป็นแนวทางการศึกษาระบบการตรวจแก้การสะกดผิดในภาษาอื่นๆ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1.6 วิธีดำเนินการวิจัย

1.6.1 ศึกษาทบทวนวรรณกรรมที่เกี่ยวข้องกับการตรวจแก้การสะกดผิด (Spelling Correction)

1.6.2 เก็บรวบรวมข้อมูลและสร้างคลังข้อมูลที่จะนำไปใช้ฝึกฝนและทดสอบระบบ

1.6.3 พัฒนาระบบช่วยตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรม

1.6.4 ทดสอบวิธีการตรวจแก้สะกดคำด้วยแบบจำลองไตรแกรม

1.6.5 ประเมินค่าประสิทธิภาพการทำงานของระบบที่เสนอเปรียบเทียบกับระบบขั้นพื้นฐาน

1.6.6 วิเคราะห์และสรุปผลการวิจัย

1.7 เครื่องมือที่ใช้ในการวิจัย

1.7.1. ระบบ Strawberry Perl version 5.24.2.1 เป็นระบบภาษา Perl ที่ผู้วิจัยใช้เขียนระบบตรวจแก้การสะกดผิดในงานวิจัยนี้

1.7.2. ระบบตัดคำ ThaiSegmentation version 2.2 ของภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

1.7.3. ข้อมูลจาก คลังข้อมูลภาษาไทยแห่งชาติ 2 (Thai National Corpus 2) ของภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 2

ทบทวนวรรณกรรม

เนื้อหาในส่วนนี้จะกล่าวถึงแนวคิด ทฤษฎี รวมถึงงานวิจัยต่างๆ ที่เกี่ยวข้องกับการตรวจแก้การสะกดผิดด้วยระบบคอมพิวเตอร์ ซึ่งได้แก่ 2.1 หลักการทำงานเบื้องต้นของการตรวจแก้การสะกดผิด เพื่อให้ผู้อ่านมองเห็นภาพรวมของงานด้านนี้ว่าประกอบด้วยกระบวนการทำงานเบื้องต้นอะไรบ้าง จากนั้นผู้วิจัยจะกล่าวถึง 2.2 ประเภทของการสะกดผิด สามารถแบ่งออกได้เป็น 2 ประเภท คือ การสะกดผิดแบบไม่เป็นคำและการสะกดผิดแบบเป็นคำจริง เพื่อให้เข้าใจว่าการสะกดผิดแต่ละประเภทนั้นหมายถึงอะไรและมีลักษณะอย่างไร ซึ่งการสะกดผิดแต่ละประเภทนั้นต้องการการแก้ไขจากวิธีการที่เหมาะสมจึงจะสามารถแก้ไขได้อย่างมีประสิทธิภาพมากที่สุด 2.3 รูปแบบหรือปัจจัยที่อาจส่งผลให้เกิดการสะกดผิด เพื่อเป็นแนวทางในการหาหรือสร้างตัวอย่างคำที่สะกดผิดสำหรับนำมาใช้เป็นข้อมูลในการฝึกฝนและทดสอบระบบได้ 2.4 วิธีการต่างๆ ที่มีการนำมาใช้ตรวจแก้การสะกดผิด ในส่วนนี้จะกล่าวถึงการตรวจแก้การสะกดผิดด้วยวิธีการแต่ละวิธีที่แตกต่างกัน เพื่อแสดงให้เห็นถึงหลักการทำงานในเบื้องต้นของวิธีการเหล่านั้นโดยอ้างอิงจากงานวิจัยต่างๆ ที่เกี่ยวข้อง ดังนั้นแนวคิดทั้งหมดที่ได้ในส่วนนี้จะสามารถช่วยให้ผู้วิจัยสามารถเลือกใช้วิธีการที่เหมาะสมสำหรับตรวจแก้การสะกดผิดแบบเป็นคำจริงได้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.1 หลักการทำงานเบื้องต้นของการตรวจแก้การสะกดผิด

โดยปกติแล้วในการจะตรวจแก้การสะกดผิดในข้อความหนึ่งๆ ผู้ที่จะสามารถทำการแก้ไขคำที่สะกดผิดให้ถูกต้องได้ในขั้นแรกจะต้องสามารถระบุถึงตำแหน่งที่มีสะกดผิดในข้อความนั้นก่อน หลังจากนั้นจึงนำคำที่สะกดผิดนั้นมาปรับแก้ให้ถูกต้อง ในทางคอมพิวเตอร์ก็เช่นกัน การตรวจแก้การสะกดผิด (Spelling Correction) ด้วยคอมพิวเตอร์คือ การตรวจจับและแก้ไขคำที่สะกดผิดในเอกสารข้อความ (Mishra & Kaur, 2013) ด้วยระบบคอมพิวเตอร์ ซึ่งหลักการทำงานเบื้องต้นของตัวระบบตรวจแก้การสะกดผิดประกอบไปด้วยกระบวนการหลัก 2 กระบวนการ คือ การตรวจจับการสะกดผิด (Error Detection) และการแก้ไขการสะกดผิด (Error Correction)

2.1.1 การตรวจจับการสะกดผิด (Error Detection)

การตรวจจับการสะกดผิดเป็นกระบวนการสกัดเอาคำจากข้อความเอกสารที่ป้อนเข้ามาไปเปรียบเทียบกับคลังศัพท์ที่บรรจุคำที่สะกดถูกต้อง เช่น คลังศัพท์จากพจนานุกรม ว่าตรงกันหรือไม่

จากนั้นหากพบว่ามีคำที่สะกดไม่ถูกต้องก็จะมีการทำเครื่องหมายกำกับคำที่สะกดผิดนั้นเพื่อเป็นการแสดงผลของการตรวจจับ เช่น ขีดเส้นใต้หรือเปลี่ยนสีของคำที่สะกดผิด โดยวิธีการดังกล่าวจะสามารถตรวจจับได้เฉพาะความผิดพลาดที่เป็นการสะกดผิดแบบไม่เป็นคำเท่านั้น เพราะคำที่สะกดผิดประเภทนี้เมื่อนำไปเปรียบเทียบกับคลังศัพท์จากพจนานุกรมแล้วจะไม่พบว่าตรงกับคำใดเลย ซึ่งกระบวนการตรวจจับการสะกดผิดข้างต้นนี้จะไม่สามารถตรวจจับข้อผิดพลาดที่เป็นการสะกดผิดแบบเป็นคำจริงได้ เพราะคำที่สะกดผิดแบบเป็นคำจริงจะมีรูปร่างหน้าตาเหมือนกับคำที่สะกดถูกต้องและปรากฏในคลังศัพท์พจนานุกรม ด้วยเหตุนี้กระบวนการตรวจจับการสะกดผิดแบบเป็นคำจริงจึงมีความท้าทายและยุ่งยากซับซ้อนมากกว่า นั่นคือในการที่จะตรวจจับการสะกดผิดแบบเป็นคำจริงได้นั้นจะต้องอาศัยบริบทข้างเคียงเข้ามาช่วยในการตัดสิน ซึ่งถ้าหากสงสัยว่าคำหนึ่งๆ นั้นเป็นคำที่สะกดผิดแบบเป็นคำจริงหรือไม่ ตัวบริบทที่อยู่ข้างเคียงคำนั้นๆ จะสามารถช่วยบอกได้ว่าคำที่น่าสงสัยนั้นเป็นคำที่ควรปรากฏอยู่ในบริบทนั้นหรือไม่

2.1.2 การแก้ไขการสะกดผิด (Error Correction)

หลังจากที่ตรวจพบคำที่สะกดผิด ไม่ว่าจะเป็นคำที่สะกดผิดแบบไม่เป็นคำหรือคำที่สะกดผิดแบบเป็นคำจริงแล้วต่างก็ต้องนำคำที่สะกดเหล่านั้นไปผ่านกระบวนการแก้ไขคำที่สะกดผิดให้ถูกต้อง ซึ่งกระบวนการการแก้ไขการสะกดผิดสามารถแบ่งย่อยออกเป็น ขั้นตอนการสรรหาหรือสร้างคำหรือรายการคำที่น่าจะเป็นไปได้และขั้นตอนการคัดเลือกคำที่ถูกต้องเหมาะสมมากที่สุด ในขั้นตอนของการสรรหาหรือสร้างคำที่น่าจะเป็นไปได้ โดยส่วนใหญ่แล้วคำที่น่าจะเป็นมักจะได้มาจากการปรับแก้คำที่สะกดผิดนั้นให้ถูกต้องโดยมีจำนวนของการแก้ไขน้อยครั้งที่สุด ซึ่งถ้าหากพบว่ามีคำที่น่าจะเป็นเพียงหนึ่งคำ ก็จะถือว่าคำๆ นั้นก็คือคำที่ถูกต้องเหมาะสมที่สุดและนำไปใช้แก้ไขคำที่สะกดผิด แต่ถ้าพบว่าคำที่น่าจะเป็นไปได้มีจำนวนมากกว่าหนึ่งคำ รายการคำเหล่านั้นจะต้องผ่านขั้นตอนสุดท้าย คือ การคัดเลือกคำที่ถูกต้องเหมาะสมมากที่สุดจากรายการคำที่น่าจะเป็น ซึ่งวิธีการคัดเลือกคำที่เหมาะสมที่สุดนั้นจะขึ้นอยู่กับผู้วิจัยแต่ละท่าน เช่น ในงานของ (Mays, Damerau, & Mercer, 1991) เลือกคำที่เหมาะสมที่สุดโดยใช้แบบจำลอง tri-gram ส่วนในงานของ (Islam & Inkpen, 2009) ใช้ noisy channel model เพื่อเลือกคำที่เหมาะสมที่สุด เมื่อได้คำที่เหมาะสมที่สุดแล้วจึงนำคำนั้นไปแก้ไขคำที่สะกดผิด นอกจากนี้ในส่วนของวิธีการดำเนินการแก้ไขคำที่สะกดผิดสามารถทำได้ 2 แบบ (Kukich, 1992) ได้แก่ แบบที่หนึ่งเป็นการแก้ไขแบบอัตโนมัติ (Automatic Correction) คือ การแก้ไขคำที่สะกดผิดให้เสร็จสรรพโดยอัตโนมัติไม่จำเป็นต้องถามความเห็นจากผู้ใช้ ซึ่งในกรณีนี้จำเป็นต้องใช้วิธีการที่มีความสามารถเพียงพอที่จะคัดเลือกเอาเฉพาะคำที่ถูกต้องและเหมาะสมที่สุดเพียงหนึ่งคำ ส่วนแบบที่สองเป็นการแก้ไขแบบมีปฏิสัมพันธ์ (Interactive Correction) คือ เมื่อได้

รายการคำที่น่าจะเป็นคำที่สะกดถูกต้องแล้วก็เสนอรายการคำเหล่านั้นให้ผู้ใช้ได้ทำการตัดสินเลือกคำที่เหมาะสมที่สุดด้วยตัวเอง

2.2 ประเภทของการสะกดผิด

การตรวจแก้การสะกดผิดโดยทั่วไปแล้วจะหมายถึงการปรับแก้คำที่มีการสะกดที่ไม่ตรงกับคำในพจนานุกรมหรือการสะกดที่ไม่เป็นคำให้ถูกต้อง แต่สำหรับการตรวจแก้การสะกดผิดของงานด้านการประมวลผลภาษาธรรมชาติจะหมายถึงระบบที่สามารถตรวจความถูกต้องของข้อความที่ป้อนให้ว่าข้อความนั้นมีข้อผิดพลาดในด้านการสะกดคำในข้อความนั้นหรือไม่ ถ้าหากตรวจพบว่ามีข้อผิดพลาดระบบก็จะทำการแก้ไขให้ถูกต้องหรือเสนอแนะรายการคำที่น่าจะเหมาะสมที่สุด (Dembitz, Gledec, & Randić, 2009; Kaur & Garg, 2014) ซึ่งข้อความที่ถูกต้องนั้นหมายถึงข้อความที่นอกจากจะต้องมีการสะกดคำที่ถูกต้องตามหลักภาษาและปรากฏในพจนานุกรมของภาษานั้นๆ แล้วยังต้องมีความถูกต้องและเหมาะสมกับบริบทและคำข้างเคียงเมื่อปรากฏร่วมกันอีกด้วย ตัวอย่างเช่น “โปรดพิจารณาควมเสี่ยวก่อนตัดสินใจลงทุน” ในข้อความนี้มีจุดผิดพลาดสองจุดที่แตกต่างกัน จุดแรกคือ “ควม” เป็นผลของการสะกดผิดแบบไม่เป็นคำคือ “ควม” นั้นไม่มีความหมายและไม่ปรากฏอยู่ในพจนานุกรม คำนี้จึงเป็นคำที่สะกดผิด ซึ่งคำที่ถูกต้องก็คือ “ความ” ส่วนข้อผิดพลาดในจุดที่สองนี้เป็นข้อผิดพลาดที่มีความซับซ้อนกว่าอันแรก คือคำว่า “เสี่ยว” ซึ่งถ้าดูแบบผิวเผินโดยไม่สนใจคำบริบทข้างเคียงจะไม่สามารถทราบได้ว่า “เสี่ยว” นั้นเป็นคำที่สะกดผิดเพราะว่า “เสี่ยว” เป็นคำที่มีความหมายและปรากฏในพจนานุกรม แต่ในประโยคนี้ “เสี่ยว” เป็นการสะกดผิดแบบเป็นคำจริง คือคำที่สะกดผิดนั้นเป็นคำที่ปรากฏในพจนานุกรม แต่เมื่อดูจากคำแวดล้อมอื่นๆ ในประโยค เช่น คำบริบท “ลงทุน” จะเห็นได้ว่าคำว่า “เสี่ยว” เมื่อปรากฏกับคำบริบทนี้แล้วทำให้ความหมายของประโยคฟังดูแปลกประหลาดจึงทำให้ทราบได้ว่า “เสี่ยว” เป็นคำที่สะกดผิด ซึ่งคำที่ถูกต้องเหมาะสมที่สุดควรจะเป็นคำว่า “เสี่ยง” แทน

จากตัวอย่างข้างต้นจะเห็นได้ว่า การสะกดผิดนั้นสามารถแบ่งออกได้เป็น 2 ประเภท (Mishra & Kaur, 2013; Mitton, 1987; S, Madi, D, & P, 2012; Verberne, 2002) ตามลักษณะของคำผิดที่ปรากฏ ซึ่งได้แก่ การสะกดผิดแบบไม่เป็นคำ (non-word spelling errors) และการสะกดผิดแบบเป็นคำจริง (real-word spelling errors)

2.2.1 การสะกดผิดแบบไม่เป็นคำ (non-word spelling errors)

คือ ความผิดพลาดในการประกอบตัวอักษรขึ้นเป็นคำทำให้คำนั้นมีการสะกดที่ผิดและคำที่ได้นั้นไม่ตรงกับคำใดเลยในพจนานุกรม ซึ่งอาจจะมีสาเหตุมาจากการเติม (insertion) การลบ

(deletion) หรือการแทนที่ (substitution) ตัวอักษรตั้งแต่หนึ่งตัวขึ้นไปผิดพลาดไปจากคำเดิมที่มีการสะกดถูกต้อง (Kaur & Garg, 2014; Stehouwer, 2011) ตัวอย่างเช่น

สะกด → สะกกด เป็นความผิดพลาดที่เกิดจาก การเติม คือพิมพ์อักษร ก เกินมา

สะกด → สะด เป็นความผิดพลาดที่เกิดจาก การลบ คือพิมพ์ตกอักษร ก

สะกด → สักด เป็นความผิดพลาดที่เกิดจาก การแทนที่ คือพิมพ์สระ ั แทนที่จะเป็นสระ -ะ

2.2.2 การสะกดผิดแบบเป็นคำจริง (real-word spelling errors)

คือ ความผิดพลาดในการประกอบตัวอักษรขึ้นเป็นคำโดยมีการพิมพ์ผิดพลาดเช่นเดียวแบบแรก แต่คำที่ได้จากการสะกดผิดบังเอิญยังเป็นคำที่มีในพจนานุกรมแต่สื่อความหมายที่แตกต่าง จึงทำให้โครงสร้างหรือความหมายของประโยคผิดเพี้ยนไป (Mitton, 1987) ซึ่งความผิดพลาดที่เกิดขึ้นนั้นอาจจะเกิดจากสาเหตุเดียวกับการสะกดผิดในประเภทแรกที่เกิดจากการเติม การลบ หรือการแทนที่ตัวอักษรภายในคำ

ตัวอย่างเช่น

หมอนกอด → หมอนอกอด เป็นความผิดพลาดที่เกิดจาก การเติม คือพิมพ์อักษร อ เกินมา

หมอนกอด→ หมอกอด เป็นความผิดพลาดที่เกิดจาก การลบ คือพิมพ์อักษร น ตกไป

หมอนกอด→ หมอรกอด เป็นความผิดพลาดที่เกิดจาก การแทนที่ คือพิมพ์ ร แทนที่จะเป็น น

2.3 ประเด็นและข้อเท็จจริงต่างๆ ที่น่าสนใจเกี่ยวกับการสะกดผิด

จากหลักการทำงานเบื้องต้นของการแก้ไขการสะกดผิดที่อธิบายในหัวข้อที่แล้วจะเห็นได้ว่าก่อนที่จะสามารถทำการแก้ไขคำที่สะกดผิดให้ถูกต้องได้นั้นจะต้องมีการตรวจจับและระบุคำที่สะกดผิดภายในข้อความให้ได้ก่อน เพื่อที่จะนำคำที่สะกดผิดนั้นไปแก้ไขให้ถูกต้องต่อไป ดังนั้นในการที่จะตรวจสอบว่าวิธีการหรือระบบที่จะใช้นั้นมีประสิทธิภาพที่จะใช้แก้ไขการสะกดผิดหรือไม่ ในเบื้องต้นอาจจะดูได้จากความสามารถในการตรวจจับคำที่สะกดผิดว่าสามารถตรวจจับความผิดพลาดนี้ได้ในระดับที่น่าพึงพอใจหรือไม่ ซึ่งในการที่จะพัฒนาระบบให้สามารถตรวจจับและระบุความผิดพลาดในข้อความได้อย่างมีประสิทธิภาพนั้นอาจจะต้องใช้ข้อมูลความรู้เกี่ยวกับการสะกดผิด กล่าวคือก่อนที่ระบบแก้ไขการสะกดผิดจะสามารถนำมาใช้ปฏิบัติงานได้นั้นจะต้องได้รับการฝึกฝนให้เรียนรู้ว่าข้อมูลตัวอย่างที่ถูกต้องและข้อมูลที่มีข้อผิดพลาดมีลักษณะอย่างไร ซึ่งโดยส่วนใหญ่แล้วประสิทธิภาพของระบบการตรวจสะกดคำนั้นจะขึ้นอยู่กับจำนวนข้อมูลตัวอย่างที่หลากหลายและครอบคลุมที่ได้รับการฝึกฝน ถ้ายิ่งมีข้อมูลตัวอย่างมาใช้สำหรับการฝึกฝนระบบมากเท่าไหร่ ประสิทธิภาพในการแก้ไขการ

สะกดผิดของระบบก็จะมากขึ้นตามไปด้วย จะเห็นได้ว่าข้อมูลตัวอย่างนั้นมีความสำคัญต่อการออกแบบและพัฒนาระบบ แต่อย่างไรก็ตามการที่จะหาข้อมูลที่มีหลากหลายเพื่อมาพัฒนาตัวระบบนั้นไม่ใช่งานที่ง่ายเลย ผู้วิจัยจึงได้ตระหนักถึงความสำคัญของการศึกษาข้อมูลต่างๆ ที่เกี่ยวข้องหรืออาจส่งผลให้เกิดการสะกดผิด และผู้วิจัยเห็นว่าเป็นเรื่องที่จำเป็นต้องศึกษาเพื่อที่จะสามารถนำความรู้ในส่วนนี้ไปประยุกต์ใช้ระบุข้อผิดพลาดเหล่านั้นและนำมาใช้สำหรับฝึกฝนตัวระบบให้สามารถทำงานได้อย่างมีประสิทธิภาพ

สำหรับการหารูปแบบหรือสาเหตุของการสะกดผิดนั้น Kukich ได้ทำงานวิจัยที่ศึกษาเกี่ยวกับสาเหตุเหล่านั้นขึ้นในปี 1992 ซึ่งได้มีการกล่าวถึงประเด็นที่น่าสนใจเกี่ยวกับการสะกดผิด โดยอ้างอิงข้อมูลจากงานวิจัยต่างๆ ที่เกี่ยวข้องที่ส่วนใหญ่เป็นงานที่ทดลองกับข้อมูลภาษาอังกฤษ เป้าหมายของการศึกษาของนักวิจัยเหล่านี้คือเพื่อนำข้อมูลที่ได้ไปออกแบบระบบหรือวิธีการจัดการกับข้อผิดพลาดเหล่านั้น ซึ่งในงานวิจัยนี้ได้นำเสนอประเด็นต่างๆ ที่น่าสนใจเกี่ยวกับการสะกดผิดไว้ดังต่อไปนี้

1 พบว่าคำที่สะกดผิดส่วนใหญ่มักจะมีจุดผิดพลาดเพียงที่เดียว

Kukich ได้อ้างข้อมูลจากงานวิจัยของ Mitton (1987) ซึ่งพบว่ามากกว่า 80% ของคำที่สะกดผิด 4,218 คำ เป็นคำที่สะกดพลาดเพียงหนึ่งจุด ซึ่งใช้หลักของ minimal edit distance ในการระบุจำนวนของข้อผิดพลาด โดยข้อมูลในส่วนนี้สามารถนำไปใช้ในการคาดเดาลักษณะของคำที่น่าจะสะกดผิดเพื่อให้สามารถตรวจจับคำที่สะกดผิดเหล่านั้นได้ดีขึ้น

2 พบว่าคำที่มีจำนวนตัวอักษรน้อยกว่าจะตรวจจับความผิดพลาดได้ยากกว่าคำที่มีจำนวนตัวอักษรมาก

จากงานวิจัยของ (POLLOCK, & ZAMORA,1983) พบว่าคำสั้นหรือคำที่ประกอบขึ้นจากจำนวนตัวอักษรเพียงแค่ 2-4 ตัวนั้นมักจะตรวจจับได้ยากกว่าคำที่ยาวหรือก็คือคำที่ประกอบขึ้นจากตัวอักษรจำนวนมากกว่า 4 ตัวขึ้นไป ซึ่งข้อมูลในส่วนนี้ช่วยให้ผู้วิจัยเพิ่มความระมัดระวังในการที่จะตรวจสอบความผิดพลาดในเหล่าคำสั้นเพื่อที่จะตรวจจับความผิดพลาดให้ได้ครบถ้วนที่สุด.

3 พบว่าตัวอักษรตัวแรกของคำมักจะไม่ผิด (First-position errors)

จากงานวิจัยต่างๆ ที่ได้ศึกษาเกี่ยวคำลักษณะของคำที่สะกดผิดที่กล่าวไว้ในงานของ Kukich พบว่า ความผิดพลาดของการสะกดคำมักจะไม่เกิดขึ้นที่ตัวอักษรตัวแรก กล่าวคือตัวอักษรต้นของคำโดยส่วนใหญ่แล้วมักจะถูกต้องเสมอ ข้อมูลในส่วนนี้ก็สามารถนำไปประยุกต์ในการกระบวนการตรวจจับคำที่สะกดผิดได้เช่นเดียวกัน

4 พบว่าความใกล้ชิดในตำแหน่งของตัวอักษรบนแป้นพิมพ์ก็สามารถทำให้เกิดความผิดพลาดได้

เนื่องจากแป้นพิมพ์โดยทั่วไปจะประกอบด้วยปุ่มจำนวนมากเรียงต่อกันบนแป้นพิมพ์ ซึ่งปุ่มแต่ละปุ่มบนแป้นพิมพ์จะมีตัวอักษรแต่ละตัวกำกับอยู่ซึ่งในบางปุ่มอาจจะสามารถมีอักษรกำกับอยู่

มากกว่า 1 ตัว จึงมีความเป็นไปได้สูงมากที่ความผิดพลาดในการสะกดคำนั้นจะเกิดจากความใกล้ชิดของตำแหน่งตัวอักษรบนแป้นพิมพ์ ซึ่งในงานวิจัยของ Kukich ก็ได้ยืนยันถึงความผิดพลาดที่เกิดจากสาเหตุนี้ โดยได้รายงายไว้ว่าข้อผิดพลาดของผู้ชำนาญการพิมพ์ส่วนใหญ่มักจะเกิดจากการทำการเพิ่มตัวอักษรโดยไม่ได้ตั้งใจเนื่องจากนิ้วมือพลาดไปกดตัวอักษรบางตัวอย่างไม่ตั้งใจ ส่วนสำหรับผู้ไม่ชำนาญการพิมพ์มักจะเกิดจากการทำการแทนที่ตัวอักษรทำให้คำผิดเพี้ยนไป ซึ่งข้อมูลในส่วนนี้เป็นอีกส่วนหนึ่งที่มีประโยชน์มากในการช่วยคาดเดาคำที่ถูกต้องเมื่อตรวจพบคำที่สะกดผิด โดยสามารถนำไปตรวจสอบดูได้ว่าความผิดพลาดนั้นเกิดจากความใกล้ชิดของตัวอักษรบนแป้นพิมพ์หรือไม่

5 พบว่าความผิดพลาดของคำที่สะกดผิดอาจจะมีสาเหตุมาจากความรู้ด้านสัทศาสตร์ (Phonetic errors)

ความรู้ทางด้านสัทศาสตร์ของผู้ใช้ เช่น ความรู้ด้านการออกเสียงของคำ ก็เป็นอีกปัจจัยหนึ่งที่อาจจะทำให้เกิดความผิดพลาดในการสะกดคำได้ ตัวอย่างเช่น หากผู้ใช้ต้องพิมพ์ข้อความจากการฟัง และในข้อความนั้นมีคำพ้องเสียงที่มีรูปเขียนต่างกันอยู่ก็อาจจะทำให้ผู้ใช้พิมพ์คำที่ไม่ถูกต้องได้ เพื่อให้ชัดเจนยิ่งขึ้น ผู้วิจัยจึงขอยกตัวอย่างในภาษาไทย เช่น “เมื่อเขาเดินนับ**ก้าว**” อาจจะผิดเพี้ยนไปเป็น “เมื่อเขาเดินนับ**เก้า**” หรือพบคำที่มีคำที่มีการออกเสียงที่ใกล้เคียงกันภายในข้อความ ก็อาจจะทำให้เกิดความผิดพลาดได้เช่นกัน “ตัวอย่าง**เส้น**” อาจจะผิดเพี้ยนไปเป็น “ตัวอย่าง**เช่น**” ข้อมูลในส่วนนี้ก็สามารถนำไปใช้ในการช่วยหาคำที่ถูกต้องเมื่อทราบคำที่สะกดผิดได้เช่นเดียวกัน

6 พบว่าความผิดพลาดของคำที่สะกดผิดอาจเกิดจากสาเหตุอื่นๆ

นอกจากสาเหตุต่างๆ ที่ได้กล่าวมาแล้วข้างต้น ความผิดพลาดที่เกิดขึ้นอาจจะยังมีสาเหตุมาจากปัจจัยอื่นๆ อีก ดังที่กล่าวไว้ในงานของ Kukich เช่น รูปแบบข้อความที่ป้อน เช่น ข้อความที่ได้จากการเขียนซึ่งลายมือในการเขียนอาจจะทำให้เกิดความผิดพลาดในขั้นตอนถอดข้อความที่เขียน หรือความทันสมัยของเครื่องมือก็มีอาจจะส่วนในการระบุความผิดพลาดภายในข้อความคลาดเคลื่อนไปได้

นอกจากงานวิจัยของ Kukich แล้วยังมีงานวิจัยของ Baba and Suzuki (2012) ที่ศึกษาถึงปัจจัยต่างๆ ที่อาจทำให้เกิดการสะกดผิด ซึ่งงานของ Baba & Suzuki นั้นจะแตกต่างจากงานของ Kukich ตรงที่งานของ Baba & Suzuki นั้นจะนำเสนอปัจจัยต่างๆที่มีผลกระทบต่อหรือก่อให้เกิดการสะกดผิด โดย Baba & Suzuki ได้เสนอปัจจัยต่างๆ ที่มีผลต่อการสะกดผิดไว้ ซึ่งสามารถแบ่งออกเป็น 3 กลุ่ม ดังนี้

1 ความผิดพลาดที่เกิดจากปัจจัยทางด้านการสัมผัส (Physical factors) ได้แก่

ก. ความผิดพลาดที่เกิดจากความพลาดพลั้งของนิ้วมือ คือ การสะกดผิดที่เกิดจากการที่นิ้วมือบังเอิญไปกดโดนปุ่มบนแป้นพิมพ์โดยไม่ตั้งใจในระหว่างที่กำลังพิมพ์ข้อความอยู่ โดยที่ปุ่มที่เผลอไปกด

โดนนั้นไม่จำเป็นต้องเป็นปุ่มที่อยู่ติดกับปุ่มที่ต้องการจะกด เช่น ตกใจ พิมพ์ผิดเป็น ต กใจ (เผลอกดเว้นวรรค)

ข. ความผิดพลาดที่เกิดจากตำแหน่งที่อยู่ติดกันหรือใกล้กันของตัวอักษรบนแป้นพิมพ์ คือ การสะกดผิดที่เกิดจากตัวอักษรที่ต้องการจะพิมพ์นั้นอยู่ในตำแหน่งที่ใกล้ชิดกันบนแป้นพิมพ์ เช่น จริง เป็น ขริง

2 ความผิดพลาดที่เกิดจากปัจจัยทางด้านการมองเห็น (Visual factors) ได้แก่

ก. ความผิดพลาดที่เกิดจากรูปเขียนที่คล้ายคลึงกันของตัวอักษร คือ การสะกดผิดที่เกิดจากตัวอักษรนั้นมีรูปเขียนที่ใกล้เคียงกัน ตัวอย่างสมมติ เช่น ในภาษาไทย อาจมีความผิดพลาดจากการมองตัวอักษร ช หรือ ซ, ก หรือ ถ เป็นต้น

ข. ความผิดพลาดที่เกิดจากการใช้ตัวอักษรซ้ำภายในคำ คือ ความผิดพลาดที่มักจะเกิดกับคำที่มีการใช้ตัวอักษรตัวเดิมซ้ำกันทำให้เกิดการลบตัวอักษรซ้ำที่ปรากฏติดกันออก เช่น ‘บุคคล’ -> ‘บุคล'

3 ความผิดพลาดที่เกิดจากปัจจัยทางด้านสัทวิทยา (Phonological factors)

เป็นความผิดพลาดที่เกิดจากการแทนที่ตัวอักษรที่มีการออกเสียงใกล้เคียงกัน คือ การพิมพ์ตัวอักษรหนึ่งแทนที่อีกตัวอักษรหนึ่งทำให้เกิดการสะกดที่ผิด ซึ่งมีทั้งการแทนพยัญชนะผิดตัว ตัวอย่างเช่น “โอกาส” พิมพ์ผิดเป็น “โอกาศ” และการแทนสระผิดตัว เช่น “หลงใหล” พิมพ์ผิดเป็น “หลงไหล” เป็นต้น โดย Baba & Suzuki ยังพบว่าคำที่สะกดผิดจากการแทนพยัญชนะผิดตัวนั้น

2.4 วิธีการต่างๆ ที่มีการนำมาใช้ในงานด้านการตรวจแก้การสะกดผิดด้วยระบบคอมพิวเตอร์

ในงานด้านการตรวจแก้การสะกดผิดนั้นเป็นงานที่มีนักวิจัยหลายท่านเสนอวิธีการต่างๆ เพื่อจัดการกับปัญหาการสะกดผิด แต่ไม่มีการระบุที่แน่ชัดว่าควรจะใช้วิธีใดมาตรวจแก้การสะกดผิดเหล่านั้น โดยวิธีการที่นักวิจัยแต่ละท่านนำมาประยุกต์ใช้นั้นจะแตกต่างกันไปตามวัตถุประสงค์และเป้าหมายของนักวิจัยแต่ละคนว่าประสงค์ที่จะนำตัวระบบที่พัฒนาขึ้นมานั้นไปประยุกต์ใช้จัดการกับปัญหาการสะกดผิดประเภทใดและใช้อัลกอริทึมใดจึงจะสามารถจัดการกับปัญหานั้นได้ดีที่สุด โดยการที่จะประยุกต์ใช้เทคนิคหรือวิธีการหนึ่งๆ ให้สามารถทำงานได้อย่างมีประสิทธิภาพสูงสุดนั้นผู้ใช้จะต้องรู้ถึงความถนัดของวิธีการที่จะนำมาใช้ด้วยว่ามันสามารถจัดการกับข้อมูลประเภทไหนและในสิ่งแวดล้อมแบบใดได้ดีที่สุด ด้วยเหตุนี้ผู้วิจัยจึงเห็นว่าการศึกษาการทำงานเบื้องต้นของวิธีการต่างๆ ที่นำมาประยุกต์ใช้ตรวจแก้การสะกดผิดแต่ละวิธีนั้นเป็นสิ่งที่จำเป็นและควรที่จะมีการอธิบายรายละเอียดเพื่อช่วยให้ผู้อ่านเข้าใจในงานด้านการตรวจแก้การสะกดผิดมากขึ้น ซึ่งเนื้อหาในส่วนนี้จะกล่าวถึงวิธีการที่นำมาใช้จัดการตรวจแก้ปัญหาการสะกดผิดเหล่านั้น

จากการศึกษาทบทวนงานวิจัยที่เกี่ยวข้องพบว่าวิธีการที่นำมาใช้นั้นอาจมีความแตกต่างกันไปตามขั้นตอนการทำงานของการตรวจแก้การสะกดผิดซึ่งประกอบไปด้วย 2 ขั้นตอนหลัก ได้แก่ ขั้นตอนแรก คือ การตรวจจับคำที่สะกดผิด (Detection) เป็นกระบวนการตรวจหาคำที่สะกดผิดในข้อความ วิธีการที่นำมาใช้ในขั้นตอนนี้จึงมีเป้าหมายเพื่อตรวจหาข้อผิดพลาดในการสะกดคำให้ได้ครบถ้วนมากที่สุด เช่น งานของ Wilcox-O'Hearn (2014) ได้ใช้ lexicon-based model เป็นการนำคำในข้อความแต่ละคำไปค้นในคลังศัพท์ว่าพบหรือไม่ หากไม่พบแสดงว่าคำนั้นอาจเป็นคำที่มีการสะกดผิด เป็นต้น ส่วนขั้นตอนที่สอง คือ การแก้ไขคำที่สะกดผิด (Correction) เป็นขั้นตอนที่ประกอบด้วยขั้นตอนย่อยอีก 2 ขั้นตอน คือ ขั้นตอนการจัดหาหรือสร้างคำ (Generation) หรือรายการคำที่น่าจะเป็น อาจทำได้ด้วยวิธีการใช้ n-gram model เช่น ในงานของ Bassil (2012) ได้ใช้ letter-bi-gram model เพื่อหาคำที่มีการจัดเรียงอักษร 2 ตัวที่คล้ายคลึงกันแล้วเลือกเอาคำที่ปรากฏมากที่สุด เป็นต้น และขั้นตอนย่อยของการแก้ไขการสะกดผิดอีกขั้นตอนหนึ่งคือ การคัดเลือกคำที่เหมาะสมที่สุดจากรายการคำที่น่าจะเป็นนั้น ตัวอย่างเช่น งานวิจัยของ Islam and Inkpen (2009) ได้ใช้ noisy channel model เพื่อหาค่าความน่าจะเป็นของคำที่เป็นไปได้แล้วเลือกเอาคำที่ให้ค่าความน่าจะเป็นสูงที่สุดเพื่อนำมาแก้ไขคำที่สะกดผิด เป็นต้น ด้วยเหตุนี้ผู้วิจัยจึงได้อธิบายรายละเอียดของวิธีการต่างๆ ที่นำมาใช้แก้ไขปัญหาการสะกดผิดตามขั้นตอนต่างๆ ไว้ดังต่อไปนี้

2.4.1 การตรวจจับการสะกดผิด (spelling errors detection)

การตรวจจับการสะกดผิด เป็นขั้นตอนแรกของการตรวจแก้การสะกดคำที่มีหน้าที่ตรวจหาคำที่สะกดผิดในข้อความ อาจจะสามารถทำได้โดยการใช้แบบจำลองคลังศัพท์ (Lexicon-based model) ที่บรรจุคำที่สะกดถูกต้องเอาไว้ เช่น คลังศัพท์พจนานุกรม เป็นต้น วิธีการนี้เป็นวิธีการพื้นฐานที่ใช้ตรวจหาคำที่สะกดผิด (Hirst & Budanitsky, 2005) ซึ่งหลักการทำงานของวิธีการนี้ก็คือการนำคำแต่ละคำในข้อความไปค้นหาในคลังศัพท์ว่าพบหรือไม่ หากพบก็แสดงว่าคำนั้นสะกดถูกต้องแล้วผ่านไปค้นหาคำอื่นต่อไป แต่ถ้าหากไม่พบแสดงว่าคำนั้นอาจจะเป็นคำที่สะกดผิด โดยที่ประสิทธิภาพของวิธีการนี้อยู่ที่จำนวนคำภายในคลังศัพท์ คือ ต้องมีจำนวนคำในคลังศัพท์มากเพียงพอจึงจะสามารถทำการตรวจจับคำที่สะกดผิดได้อย่างครบถ้วน (Bassil, 2012) อย่างไรก็ตามวิธีการนี้สามารถที่จะใช้ตรวจหาคำที่สะกดผิดแบบไม่เป็นคำเท่านั้น (non-word spelling errors) แต่จะไม่สามารถตรวจหาคำที่มีการสะกดผิดแบบเป็นคำได้ (real-word spelling errors) ยกตัวอย่างเช่น ในภาษาไทย ถ้าใช้วิธีการตรวจหาคำที่สะกดผิดภายในข้อความด้วยแบบจำลองคลังศัพท์ “เขา**ตื้น**ไปโรง**เรียบน**ตั้งแต่เช้า” จะสามารถตรวจพบเฉพาะคำที่สะกดผิดอยู่เพียงหนึ่งคำ คือ “**เรียบน**” เพราะไม่

พบคำนี้ในคลังศัพท์ ทั้งๆ ที่ประโยคตัวอย่างข้างต้นมีคำที่สะกดผิดอยู่อีกหนึ่งคำ คือ “**ตื้น**” แต่เนื่องจากคำว่า “ตื้น” เป็นคำที่พบในคลังศัพท์ จึงตรวจจับไม่พบความผิดพลาดในการสะกดของคำๆ นี้ ด้วยสาเหตุนี้วิธีการนี้จึงไม่เหมาะที่จะนำมาใช้ตรวจจับคำที่สะกดผิดแบบเป็นคำจริง

นอกจากนี้ ปัญหาด้านการสะกดผิดแบบเป็นคำจริงนั้นมีความซับซ้อนมากกว่าและตรวจพบได้ยากกว่าการสะกดผิดแบบไม่เป็นคำ การหาวิธีที่จะจัดการกับปัญหานี้จึงถือเป็นเรื่องที่ท้าทายสำหรับงานทางด้านการตรวจแก้การสะกดผิด ซึ่งงานวิจัยต่างๆ ที่เกี่ยวข้องส่วนมากจึงมุ่งพัฒนาแบบจำลองที่สามารถตรวจหาคำที่สะกดผิดแบบเป็นคำจริง โดยมีการประยุกต์ใช้วิธีการต่างๆ เพื่อที่จะจัดการแก้ไขปัญหานี้ ตัวอย่างเช่น การใช้แบบจำลองเอ็นแกรม (N-gram model) ในงานของ Bassil (2012) ได้มีการประยุกต์ใช้คลังข้อมูล unigram ที่ได้จาก Yahoo! N-Grams Dataset ซึ่งคลังข้อมูล unigram นี้คือคลังข้อมูลที่บรรจุคำจาก Yahoo! N-Grams Dataset ไว้เป็นคำเดี่ยวๆ โดยนำคลังข้อมูล unigram นี้มาช่วยในการตรวจหาคำที่สะกดผิดในข้อความด้วยการนำคำแต่ละคำภายในข้อความมาค้นหาในคลังข้อมูล unigram หากไม่พบแสดงว่าคำๆ นั้นสะกดผิด ซึ่งในงานวิจัยของ Bassil ได้ใช้เป็นคลังข้อมูล unigram นี้ทำการตรวจหาคำที่สะกดผิดแบบไม่เป็นคำได้สำเร็จ 87% และการสะกดผิดแบบเป็นคำจริงได้สำเร็จเพียงแค่ 13% ซึ่งจะเห็นได้ว่าเปอร์เซ็นต์ในการใช้คลังข้อมูล unigram เพื่อตรวจหาคำที่สะกดผิดแบบเป็นคำนั้นยังไม่ใช่วิธีที่เหมาะสมนัก

วิธีที่เหมาะสมกว่า น่าจะเป็นการนำคำบริบทใกล้เคียงมาช่วยในการตรวจหาคำที่สะกดผิดแบบเป็นคำจริง เช่นในงานวิจัยของ Mays et al. (1991) ได้เสนอวิธีจัดการกับปัญหานี้ด้วย Word-trigrams statistical language model ซึ่งเป็นวิธีการเชิงสถิติที่ช่วยตัดสินความถูกต้องเหมาะสมของคำแต่ละคำด้วยการคำนวณหาค่าความน่าจะเป็นของสายคำหนึ่งๆ ซึ่งความน่าจะเป็นของประโยค $w_1, w_2, w_3, \ldots, w_n$ จะได้จากการคำนวณหา ค่าความน่าจะเป็นของ $P(w_1) \times P(w_2|w_1) \times P(w_3|w_1w_2) \times P(w_4|w_2w_3) \times \ldots P(w_n|w_n\text{-}2w_n\text{-}1)$ และสามารถประมาณด้วยความน่าจะเป็นของไตรแกรมของคำสามารถเขียนเป็นสูตรได้ดังนี้

$$P(w) = \prod_{i=1}^{n} P(w_n|w_{n-2}w_{n-1})$$

โดย P(w) คือ ความน่าจะเป็นของสายคำหนึ่งๆ

w คือ คำในประโยค

n คือ จำนวนของคำทั้งหมดภายในประโยค

ในงานวิจัยของ Mays และคณะได้มีการใช้ คำศัพท์ที่ถูกต้องจำนวน 20,000 คำและมีการสกัดหาค่าความน่าจะเป็นของ trigrams จากโครงการ IBM Speech Recognition (Bahl, Jelinek, & Mercer, 1983) แต่ไม่พบว่ามีการกล่าวถึงขนาดของคลังข้อมูลที่นำมาใช้เพื่อสกัดข้อมูลออกมา ซึ่งข้อมูลทีนำมาทดสอบนั้น Mays และคณะใช้ประโยค 100 ประโยคที่ล้วนประกอบด้วยคำศัพท์ที่ถูกต้องที่จัดเตรียมไว้ ต่อจากนั้นก็นำประโยคเหล่านั้นมาสร้างประโยคใหม่ที่คล้ายคลึงกันแต่ภายในประโยคใหม่แต่ละประโยคที่ถูกสร้างขึ้นนั้นจะประกอบด้วยคำที่สะกดผิดแบบเป็นคำจริงอยู่ โดยคำที่สะกดผิดแบบเป็นคำจริงนั้นได้มาจากการปรับแก้คำเดิมที่ถูกต้องด้วยการเพิ่ม การลด การแทนที่ หรือการสลับ อย่างใดอย่างหนึ่งเพียงครั้งเดียวเท่านั้น ซึ่งสามารถสร้างประโยคใหม่ที่มีคำที่สะกดผิดอยู่ได้ทั้งหมด 8,628 ประโยค แล้วจึงนำแต่ละประโยคที่มีอยู่ทั้งหมดไปคำนวณหาค่าความน่าจะเป็นของ trigrams ในการตรวจจับการสะกดผิดของ Mays และคณะใช้หลักการเทียบหาคำที่สะกดไม่เหมือนกับคำในประโยคดั้งเดิมเป็นคำที่สะกดผิด และหลักการแก้ไขคำที่สะกดผิดคือการนำคำที่น่าจะเป็นจากชุดคำสับสนที่เตรียมไว้มาแทนที่แล้วเลือกคำที่ให้ค่าความน่าจะเป็นของประโยคสูงสุดเป็นคำที่สะกดถูกต้อง งานวิจัยชิ้นนี้ได้สามารถทำการตรวจจับคำที่สะกดผิดได้สำเร็จ 76% และแก้ไขคำสะกดผิดได้สำเร็จ 74% แต่อย่างไรก็ตามในงานวิจัยชิ้นนี้ไม่ได้มีการนำไปใช้ในการตรวจจับและแก้ไขคำที่สะกดผิดแบบเป็นคำด้วยข้อมูลจริงที่ไม่ใช่ข้อมูลที่จัดทำขึ้นเอง ผลการวิจัยที่ได้จึงอาจจะไม่แสดงถึงความสามารถในการนำไปใช้งานได้จริงของวิธีการแก้ไขคำที่สะกดผิดด้วย Word-trigrams statistical language model

2.4.2 การแก้ไขการสะกดผิด (spelling errors correction)

หลังจากที่สามารถตรวจพบคำที่สะกดผิดแล้วก็จะต้องมีการจัดหาคำหรือรายการคำที่น่าจะเป็นไปได้เพื่อแก้ไขคำที่สะกดผิดนั้นซึ่ง minimum edit distance เป็นวิธีพื้นฐานที่สามารถจัดหาคำที่น่าจะเป็นได้ง่ายที่สุด โดยการปรับแก้คำที่สะกดผิดด้วยจำนวนการแก้ไขน้อยครั้งที่สุด ซึ่งถ้าหากว่าพบคำที่น่าจะเป็นมีจำนวนมากกว่าหนึ่งคำก็จะต้องเสนอเป็นรายการคำเพื่อให้ผู้ใช้ตัดสินใจเลือกคำที่ถูกต้องเอง แต่ถ้าหากพบว่าคำที่น่าจะเป็นมีจำนวนมากเกินไป เช่น มากกว่า 20 คำขึ้นไป การจัดหาคำที่เหมาะสมด้วย edit distance คงเป็นวิธีที่ไม่เหมาะสม จึงมีการเสนอวิธีอื่นๆ เพื่อนำมาใช้ในการจัดหาคำที่น่าจะเป็นเพื่อแก้ไขคำที่สะกดผิด เช่น การใช้วิธีการทางสถิติมาช่วยในการแก้ไข เป็นต้น

ในงานของ Bassil (2012) ได้ทำการคัดเลือกคำที่น่าจะเป็นด้วยวิธีการทางสถิติจากการใช้ letter-based bi-gram model เป็นการนำเอาคำที่สะกดผิดมาแบ่งออกเป็น อักษรเรียงคู่ (bi-gram letter sequence) ดังที่ Bassil ได้ยกตัวอย่างไว้ในงานวิจัยของเขา คือ ‘modil’ จะถูกทำให้เป็นข้อมูลอักษรเรียงคู่ ดังนี้ mo, od, di, il แล้วนำไปเทียบกับคำในคลังข้อมูล unigram ที่ได้จาก

Yahoo! N-Grams Dataset ว่า มีคำใดบ้างที่มีจำนวนตัวอักษรใกล้เคียงกับคำที่สะกดผิดและมีอักษรเรียงคู่: mo, od, di, il เหล่านี้อยู่ภายในคำมากที่สุดแล้วจึงนำเอา 5 คำแรกซึ่งได้แก่ modal, model, radian, mother, lading เป็นตัวเลือกของคำที่น่าจะเป็น โดยที่ modal และ model นั้นพบอักษรเรียงคู่ 2 ตัว (mo และ od) อยู่ภายในคำ radian และ lading พบอักษรเรียงคู่ 1 ตัวคือ di อยู่ภายในคำ ส่วน mother ก็พบอักษรเรียงคู่ 1 ตัวคือ mo อยู่ภายในคำ หลังจากที่ได้รายการคำที่น่าจะเป็นมาแล้วจะนำคำที่น่าจะเป็นเหล่านี้ไปแทนที่คำที่สะกดผิดในรูปแบบของประโยคคำเรียงห้า (5-gram word sentence) ดังสูตร

$$N_q = w_{q-4}w_{q-3}w_{q-2}w_{q-1}c_{qf}$$

โดย N คือ ประโยคคำเรียงห้า

w คือ คำในประโยคที่อยู่ก่อนหน้าคำที่สะกดผิด

c คือ คำน่าจะเป็น

q คือ ลำดับของคำที่สะกดผิดในประโยคเริ่มต้น

f คือ ลำดับของคำน่าจะเป็นที่สร้างขึ้น

จะได้

“also work with plastic (modal)”

“also work with plastic (model)”

“also work with plastic (radian)”

“also work with plastic (mother)”

“also work with plastic (lading)”

แล้วจึงเลือกเอาคำที่แทนแล้วทำให้ประโยคนั้นมีค่าความถี่สูงสุดใน Yahoo! N-Grams Dataset เป็นคำที่เหมาะสมที่สุด ซึ่งคำที่ถูกเลือกคือ model เพราะ “also work with plastic **model**” เป็นประโยคคำเรียงห้าที่พบความถี่ใน Yahoo! N-Grams Dataset มากที่สุด โดยผลงานวิจัยนี้พบว่าวิธีการแก้ไขคำที่สะกดผิดด้วย N-gram model นั้นสามารถแก้ไขคำที่สะกดผิดแบบไม่เป็นคำได้ถูกต้อง 99% และแก้ไขคำที่สะกดผิดแบบเป็นคำได้ถูกต้อง 65% จะเห็นได้ว่าการเลือกใช้ 5-grams model มาช่วยแก้ไขคำที่สะกดผิดแบบเป็นคำนั้นอาจจะไม่ได้ผลดีมากนัก เพราะสำหรับคำที่ถูกต้องบางคำอาจจะพบความถี่ในการปรากฏแบบคำเรียงห้าน้อยมากจนทำให้เลือกคำที่ยังไม่ใช่คำที่เหมาะสมที่สุดไปใช้แก้ไข

นอกจากการนำวิธีการทางสถิติมาใช้ในกระบวนการแก้ไขคำที่สะกดผิดแบบเป็นคำแล้ว การใช้ความน่าจะเป็นก็เป็นอีกวิธีหนึ่งที่สามารถนำมาช่วยจัดการกับปัญหานี้ เช่นงานวิจัยของ Golding

(1995) ที่ใช้บริบทในการแก้ไขการสะกดผิดแบบเป็นคำจริง โดยมีการใช้ confusion sets ทั้งหมด 18 ชุด แต่ละชุดจะประกอบด้วยคำที่มีรูปเขียน การออกเสียง หรือประเภทคำ ที่เหมือนกันหรือใกล้เคียงกัน โดยส่วนใหญ่แล้วconfusion sets นี้นำมาจาก “Words Commonly Confused” ที่อยู่ด้านหลังของ พจนานุกรม Random House ของ Flexner ปีค.ศ. 1983 ซึ่งในงานวิจัยของ Golding (1995) นี้ได้ใช้ข้อมูลในการฝึกจำนวนหนึ่งล้านคำจาก Brown Corpus ของ Kucera and Francis และข้อมูลในการทดสอบจำนวนเจ็ดแสนห้าหมื่นคำจาก Wall Street Journal Text ของ Marcusและคณะ มาใช้ในการเปรียบเทียบประสิทธิภาพของวิธีการที่ต่างกัน 5 วิธี ได้แก่

วิธีที่หนึ่ง baseline method เป็นวิธีการพื้นฐานที่นำมาใช้เพื่อใช้เป็นตัวเปรียบเทียบกับวิธีการอื่นๆ โดยเมื่อระบบตรวจพบว่าในข้อความมีคำใน confusion sets อยู่ก็จะนำคำในเซตนั้นมาเปรียบเทียบความถี่ในการปรากฏโดยไม่สนใจบริบทแล้วเลือกเอาคำที่พบความถี่สูงสุด

วิธีที่สอง Context words เป็นวิธีการที่ใช้คำบริบทข้างเคียงที่อยู่ในระยะ $\pm$k คำของคำเป้าหมาย (k เป็นค่าที่ Golding กำหนดขึ้นเองตามที่เห็นสมควร ซึ่งเขาได้ลองกำหนดให้ k = 3, 6, 12, และ 24 ซึ่งพบว่า k=3 ให้ผลดีที่สุด) มาช่วยในการตัดสินเลือกคำใน confusion sets ที่ให้ค่าความน่าจะเป็นสูงสุดในบริบทเหล่านั้น ซึ่งวิธีนี้พบปัญหาในเรื่องของคำบริบทที่นำมาช่วยในการเลือกคำที่เหมาะสมที่สุดเพราะไม่ใช่คำในบริบททุกคำจะสามารถระบุความเหมาะสมของคำเป้าหมายได้ เช่น คำไวยากรณ์ต่างๆ เป็นต้น

วิธีที่สาม Collocations เป็นวิธีการที่ใช้คำบริบทที่ปรากฏร่วมและอยู่ติดกับคำเป้าหมาย (collocation) มาช่วยในการเลือกคำเหมาะสมที่สุดจากใน confusion sets ซึ่งข้อบกพร่องของวิธีการนี้อยู่ที่คำบริบทที่อยู่ติดกับคำเป้าหมายจำนวนมากไม่มีคุณลักษณะที่จะสามารถนำมาใช้ในการคัดเลือกคำที่เหมาะสมได้ เพราะคำที่อยู่ติดกับคำเป้าหมายส่วนใหญ่อาจจะเป็นคำไวยากรณ์ที่สามารถปรากฏได้กับคำจำนวนมหาศาล จึงไม่สามารถช่วยในการคัดเลือกคำที่เหมาะสมได้

วิธีที่สี่ Decision lists เป็นวิธีการที่ประยุกต์ใช้ทั้ง context words และ collocations โดยนำคำบริบทและคำข้างเคียงที่ได้จากวิธีการทั้งสองมารวมกันแล้วหาค่าความน่าเชื่อถือของคำบริบททั้งหมดจากสูตร

$$reliability'(f) = \max p(w_i|f)$$

โดย w_i แทน คำใน confusion set

f แทน คำ context หรือ collocation

แล้วนำคำบริบทเหล่านั้นมาเรียงลำดับตามค่าความน่าเชื่อถือจากมากสุดไปหาน้อยสุด จากนั้นก็เลือกใช้คำบริบทคำแรกที่มีค่าความน่าเชื่อถือมากสุดและสามารถช่วยคัดเลือกคำใน

confusion set ที่เหมาะสมได้ จุดอ่อนของวิธีการนี้อยู่ที่การใช้คำบริบทที่มีค่าความน่าเชื่อถือมากที่สุดเพียงคำเดียวในการช่วยเลือกคำที่เหมาะสมเพราะคำบริบทที่มีค่าความน่าเชื่อถือมากที่สุดนั้นอาจจะไม่ได้เป็นคำบริบทที่จะช่วยเลือกคำที่เหมาะสมที่สุดได้

วิธีที่ห้าคือ Bayesian Classifiers เป็นวิธีการที่ประยุกต์ใช้ทั้ง context words และ collocations เช่นเดียวกันกับ Decision lists แต่วิธีการนี้จะต่างไปตรงที่การเลือกใช้คำบริบทมาช่วยในการคัดเลือกคำที่เหมาะสมจากหนึ่งคำเป็นทุกคำบริบทที่จะสามารถช่วยเลือกคำที่เหมาะสมที่สุดได้ ซึ่งเมื่อนำผลการปฏิบัติการของวิธีการทั้งหมดรวมถึง trigrams มาเปรียบเทียบกันได้ผลว่าวิธีการนี้สามารถจัดการกับปัญหาการสะกดผิดแบบเป็นคำได้ใกล้เคียงกับวิธีการอื่นๆ โดยจะสามารถจัดการคำที่สะกดผิดที่เป็นคำประเภทเดียวกันกับคำใน confusion set ได้ดีกว่า trigrams แต่จะแย่กว่าหากประเภทของคำต่างกัน

ซึ่งในเวลาต่อมา Golding and Schabes (1996) ได้เสนอวิธีการแบบผสม (hybrid method) ที่รวมเอาวิธีการเชิงสถิติอย่าง trigrams และ วิธีการหาความน่าจะเป็นอย่าง Bayes เข้าไว้ด้วยกันแล้วเสนอเป็นวิธีใหม่คือ Tribayes โดยใช้ confusion sets เดียวกันกับที่ใช้ในงานของ Golding เมื่อปีค.ศ. 1995 ซึ่ง เมื่อตรวจพบคำใน confusion set ภายในข้อความ Tribayes จะนำคำแต่ละคำใน confusion set มาแทนในประโยคและหาค่าความน่าจะเป็นของประโยคนั้นจาก POS trigrams แล้วเลือกเสนอคำที่ให้ค่าความน่าจะเป็นสูงสุด ซึ่งผลปรากฏว่า Tribayes ทำงานได้อย่างน่าพอใจ แต่อย่างไรก็ตาม งานวิจัยของ Golding ทั้งในปี 1995 และ 1996 วัดผลการทำงานของวิธีการต่างๆ จากการตรวจจับและแก้ไขคำใน confusion sets เพียง 18 เซตเท่านั้น หากต้องการที่จะวัดความสามารถของการแก้ไขการสะกดผิดแบบเป็นคำจริงในความเป็นจริง คงจะต้องใช้ confusion sets จำนวนมากกว่านี้ถึงจะสามารถตรวจแก้คำที่สะกดผิดได้อย่างมีประสิทธิภาพ อีกทั้งในปัจจุบันภาษาไทยยังไม่มีคลังข้อมูลที่มีการติดประเภทของคำ (Part-Of-Speech Tagging) เพราะฉะนั้นการใช้ Part-Of-Speech N-gram เพื่อช่วยในการแก้ไขการสะกดผิดจึงยังไม่ใช่วิธีการที่สมควรนักสำหรับสถานการณ์ปัจจุบัน

เนื่องจากข้อบกพร่องต่างๆ ของวิธีการทางสถิติและความน่าจะเป็น จึงได้มีนำวิธีการด้านการเรียนรู้ของเครื่อง (Machine learning) มาประยุกต์ใช้ในการแก้ไขการสะกดผิดแบบเป็นคำจริง เช่น งานของ Golding and Roth (1999) ซึ่งได้มีการใช้วินโนว์ (Winnow) เพื่อช่วยในการแก้ไขคำที่สะกดผิดแบบเป็นคำจริง โดยวินโนว์เป็นวิธีการที่เรียนรู้และพัฒนาจากการอัพเดทน้ำหนักความน่าเชื่อถือด้วยการคูณ ซึ่งวิธีการนี้จำเป็นต้องใช้ confusion set เพื่อสอนให้วินโนว์เรียนรู้การปรากฏของคำในบริบทที่เหมาะสม โดยที่คำต่างๆ ใน confusion set เปรียบเสมือน กลุ่มก้อนเมฆ (clouds) ที่เชื่อมโยงกับโหนดคำบริบท (context words node) และ collocations ต่างๆ ทำให้มีโครงสร้างคล้ายกับโครงข่ายใยประสาท (neuron-like network) โดยคำบริบทและ collocations เหล่านั้นจะ

เป็นคุณลักษณะที่นำมาช่วยในการตัดสินว่าคำแต่ละคำใน confusion set ควรที่จะปรากฏในบริบทแวดล้อมใด เพื่อเป็นการแก้ไขความกำกวมในระดับคำซึ่งเป็นปัญหาที่เกิดจากการสะกดผิดแบบเป็นคำจริง วิธีการจัดการกับปัญหานี้ด้วยวินโนว์ดูเหมือนจะเป็นวิธีการที่น่าจะสามารถแก้ไขปัญหานี้ได้อย่างมีประสิทธิภาพ แต่มีความยากลำบากในการหาข้อมูลที่ครอบคลุมคำใน confusion set ที่มากพอจะใช้ในการฝึกฝนระบบเพื่อให้สามารถหา feature ที่เกี่ยวข้องในการเลือกคำตอบได้ถูกต้อง

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 3

การจัดเตรียมคลังข้อมูล

เนื้อหาในบทนี้จะกล่าวถึงข้อมูลและคลังข้อมูลต่างๆ ที่นำมาใช้ในการตรวจแก้การสะกดผิดแบบเป็นคำจริงในงานวิจัยนี้ เนื่องจากแบบจำลองที่เสนอในการตรวจแก้คำผิดในที่นี้ใช้ข้อมูลไตรแกรมของคำเป็นสำคัญ หัวข้อ 3.1 จึงจะกล่าวถึงการเตรียมและสร้างคลังข้อมูลไตรแกรมคำ (word-trigram corpus) สำหรับหัวข้อ 3.2 คลังข้อมูลยูนิแกรมคำ (word-unigram corpus) เป็นรายการคำพร้อมความถี่ที่สร้างขึ้นมาเพื่อใช้กับระบบที่เป็นตัวเทียบพื้นฐาน หัวข้อถัดมา กล่าวถึงการเตรียมคลังข้อมูลชุดคำสับสน (confusion set corpus) ที่จะเป็นข้อมูลสำหรับหาตัวอย่างการสะกดผิดแบบเป็นคำจริงมาใช้และสร้างเป็นข้อมูลฝึกฝนและทดสอบ (training and testing data) โดยข้อมูลเหล่านี้เป็นส่วนที่มีความสำคัญมากที่สุดส่วนหนึ่งของงานวิจัยชิ้นนี้ เนื่องจากความสามารถและประสิทธิภาพในการทำงานของระบบช่วยตรวจแก้การสะกดผิดที่ผู้วิจัยได้พัฒนานั้นขึ้นอยู่กับข้อมูลทั้งหลายเหล่านี้

3.1 คลังข้อมูลไตรแกรมคำ (word-trigram corpus)

คลังข้อมูลไตรแกรมคำที่ใช้ในงานวิจัยชิ้นนี้ เป็นคลังข้อมูลที่เก็บข้อมูลเชิงสถิติของชุดคำเรียงต่อกันสามคำคู่กับความถี่ในการปรากฏของชุดคำเรียงนั้นเอาไว้ ซึ่งงานวิจัยนี้ผู้วิจัยต้องการทดลองพัฒนาระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรม ซึ่งแบบจำลองไตรแกรมที่ว่านั้นก็คือการนำข้อมูลไตรแกรมคำมาใช้ในการตรวจจับและแก้ไขการสะกดผิดแบบเป็นคำจริงในข้อความ เพราะฉะนั้นคลังข้อมูลไตรแกรมคำจึงเป็นเสมือนหัวใจสำคัญของระบบ ซึ่งวิธีการในการนำข้อมูลไตรแกรมในคลังข้อมูลไตรแกรมคำไปใช้ตรวจแก้การสะกดผิดอย่างไรนั้นผู้วิจัยได้อธิบายไว้ในบทที่ 5 และในการจัดเตรียมคลังข้อมูลไตรแกรมคำเพื่อที่จะสามารถนำมาใช้ในการตรวจแก้การสะกดผิดแบบเป็นคำจริงในภาษาไทยนั้นจำเป็นต้องอาศัยคลังข้อมูลภาษาไทยขนาดใหญ่ที่มีความหลากหลายและเป็นข้อมูลที่ผู้ใช้ภาษาไทยใช้จริงซึ่งจะช่วยให้ได้คลังข้อมูลไตรแกรมคำที่จะสามารถช่วยให้การตรวจแก้การสะกดผิด ผู้วิจัยจึงได้เลือกใช้คลังข้อมูลภาษาไทยแห่งชาติ (Thai National Corpus II) (Aroonmanakul, 2007) ในการจัดเตรียมคลังข้อมูลไตรแกรมคำ ซึ่งคลังข้อมูลภาษาไทยแห่งชาติเป็นคลังข้อมูลภาษาไทยที่รวบรวมการใช้ภาษาไทยทั้งภาษาเขียนและภาษาพูดที่ใช้จริงในปัจจุบันกว่า 32 ล้านคำ และมีความหลากหลายทั้งทางด้านเนื้อหา (เชิงวิชาการ กึ่งวิชาการ เป็นต้น) และสื่อที่ใช้เป็นแหล่งข้อมูล (หนังสือ วารสาร หนังสือพิมพ์ นิยาย เป็นต้น) เพราะฉะนั้นผู้วิจัยจึงเชื่อ

ว่าคลังข้อมูลไตรแกรมคำที่ได้จากข้อมูลในคลังข้อมูลภาษาไทยแห่งชาติจะสามารถนำไปใช้ตรวจหาและแก้ไขการสะกดผิดแบบเป็นคำจริงได้

3.1.1 การเตรียมคลังข้อมูลไตรแกรมคำ

เนื่องจากคลังข้อมูลไตรแกรมคำในงานวิจัยชิ้นนี้ใช้ข้อมูลจากคลังข้อมูลภาษาไทยแห่งชาติ ซึ่งมีการกำกับข้อมูลคำพร้อมคำอ่าน (ดังตัวอย่าง 3.1.1) แต่ข้อมูลที่ผู้วิจัยต้องการมีเพียงแค่รูปคำไทยเท่านั้น เพราะฉะนั้นขั้นตอนแรกในการเตรียมคลังข้อมูลไตรแกรมคำก็คือการสกัดเอาเฉพาะคำไทยออกมาจากคลังข้อมูลภาษาไทยแห่งชาติแล้วใช้เครื่องหมาย “|” คั่นระหว่างคำแต่ละคำ จากข้อมูลในตัวอย่าง 3.1 เมื่อสกัดเอาคำไทยออกมาจากข้อมูลในส่วนนี้แล้วใช้เครื่องหมาย “|” เชื่อมคำแต่ละคำเข้าด้วยกันจะได้ประโยคข้อความดังนี้ “โครงการ|วิจัย|เพื่อ|ปรับปรุง|วิธีการ|กำหนด|ช่อง|สัญญาณ|แบบ|พลวัต|ใน|ระบบ|โทรศัพท์|เคลื่อนที่|แบบ|เซล|ลูลาร์” และหากพบตัวเลขไม่ว่าจะเป็นตัวเลขไทยหรืออารบิก ตัวอักษรภาษาอังกฤษ หรือสัญลักษณ์เครื่องหมายต่างๆ ที่ไม่ใช่เครื่องหมายแบ่งคำ (|) ไม้ยมก (ๆ) ไม้ไต่คู้ (็) และไปยาลน้อย (ฯ) ปรากฏในข้อมูลที่สกัดออกมาจากคลังข้อมูลภาษาไทยแห่งชาติ ดังตัวอย่าง 3.2 ตัวเลขและตัวอักษรภาษาอังกฤษเหล่านั้นจะถูกเปลี่ยนให้อยู่ในรูปของรหัสข้อมูลคือ num และ en โดย num หมายถึง ตัวเลข และ en หมายถึง ตัวอักษรหรือคำภาษาอังกฤษ ส่วนสัญลักษณ์เครื่องหมายต่างๆ นอกเหนือจากที่ยกเว้นจะถูกตัดทิ้ง ดังตัวอย่างที่ 3.3 หลังจากที่ได้สกัดเอาประโยคข้อความออกมาจากคลังข้อมูลภาษาไทยแห่งชาติได้ทั้งหมดแล้ว ขั้นตอนที่สองคือแบ่งคำในแต่ละประโยคออกเป็นสายคำเรียงสามคำ เช่น “โครงการ|วิจัย|เพื่อ” “วิจัย|เพื่อ|ปรับปรุง” “เพื่อ|ปรับปรุง|วิธีการ” เป็นต้น แล้วนับความถี่ในการปรากฏของสายคำนั้นทั้งหมดที่พบในคลังข้อมูลภาษาไทยแห่งชาติ จากนั้นขั้นตอนสุดท้ายก็คือ จัดเก็บข้อมูลสายคำเรียงสามคำคู่กับความถี่ในการปรากฏของสายคำนั้นโดยใช้เครื่องหมาย “=” เชื่อมระหว่างข้อมูลทั้งสองเอาไว้ในคลังข้อมูลไตรแกรมคำ

ตัวอย่าง 3.1 แสดงตัวอย่างข้อมูลภายในคลังข้อมูลภาษาไทยแห่งชาติ

```
</tncHeader>
<text><body>
        <p n="1"><w tran="khrooN0 kaan0">โ ค ร ง ก า ร </w><w tran="wi3 caj0">วิ จั ย
</w><w tran="phUUa2 ">เ พื่ อ </w><w tran="prap1 pruN0 ">ป รั บ ป รุ ง </w><w
tran="wi3thii0kaan0">วิธีการ</w><w tran="kam0not1">กำหนด</w><w tran="chOON2">
ช่ อ ง </w><w tran="san4 jaan0 ">สั ญ ญ า ณ </w><w tran="bxxp1 ">แ บ บ </w><w
```

tran="phon0 la3 wat3">พล วัต</w><w tran="naj0">ใน</w><w tran="ra3 bop1">ระ บบ
</w><w tran="thoo0ra3sap1">โทรศัพท์</w><w tran="khlUUan2thii2">เคลื่อนที่</w><w
tran="bxxp1 " >แ บ บ </w><w tran="seel0 " >เ ซ ล </w><w tran="luu0 " >ลู </w><w
tran="laa0">ลาร์</w> </p>ccc

ตัวอย่างที่ 3.2 แสดงตัวอย่างที่สกัดได้จากคลังข้อมูลภาษาไทยแห่งชาติ

|วิธีการ|ที่|จะ|เพิ่ม|ความ|จุ|ของ|ระบบ|เพื่อ|รองรับ|การ|ใช้งาน|ที่|เพิ่ม|ขึ้น| |อาจ|ทำ|ได้|โดย|การ|ลด|ขนาด|ของ|พื้นที่|ครอบคลุม|ของ|สถานี|ฐาน|ให้|เล็ก|ลง| |ซึ่ง|เรียกว่า|ไมโคร|เซลล์| |Microcell| |Greenstein| |et| |al|. 1992| |และ| |Sarnecki| |et| |al|. 1993| |ได้|กล่าวถึง|ข้อดี|ของ|ไมโคร|เซลล์|

ตัวอย่างที่ 3.3 แสดงตัวอย่างที่จัดการกับตัวเลข คำภาษาอังกฤษ และเครื่องหมายต่างๆ เรียบร้อยแล้ว

|วิธีการ|ที่|จะ|เพิ่ม|ความ|จุ|ของ|ระบบ|เพื่อ|รองรับ|การ|ใช้งาน|ที่|เพิ่ม|ขึ้น| |อาจ|ทำ|ได้|โดย|การ|ลด|ขนาด|ของ|พื้นที่|ครอบคลุม|ของ|สถานี|ฐาน|ให้|เล็ก|ลง| |ซึ่ง|เรียกว่า|ไมโคร|เซลล์| |en| |num| |และ| |en| |num| |ได้|กล่าวถึง|ข้อดี|ของ|ไมโคร|เซลล์|

3.1.2 ลักษณะโครงสร้างของคลังข้อมูลไตรแกรมคำ

คลังข้อมูลไตรแกรมคำที่ใช้ในการวิจัยนี้จะประกอบด้วยข้อมูลสองส่วน ส่วนแรกคือข้อมูลไตรแกรมคำ (word trigram) หรือสายคำสามคำเรียงต่อเนื่องกันโดยมีเครื่องหมาย “|” ขั้นระหว่างคำ ตัวอย่างเช่น “ฉัน|กำลัง|รับประทาน” “กำลัง|รับประทาน|อาหาร” และ “รับประทาน|อาหาร|กลางวัน” ซึ่งเป็นไตรแกรมคำของประโยค “ฉัน|กำลัง|รับประทาน|อาหาร|กลางวัน” และส่วนที่สองคือข้อมูลความถี่ในการปรากฏของไตรแกรมคำแต่ละชุด ข้อมูลในส่วนแรกและส่วนที่สองจะถูกเก็บไว้ร่วมกันเป็นคู่ โดยมีเครื่องหมาย “=” คั่นกลาง ดังนั้นข้อมูลในไตรแกรมคำจึงมีลักษณะโครงสร้างดังตัวอย่างด้านล่าง

ตัวอย่าง 3.4 แสดงตัวอย่างข้อมูลในคลังข้อมูลไตรแกรมคำ (ความถี่ในตัวอย่างเป็นความถี่สมมติ)

ไตรแกรมคำ=ความถี่

ฉัน|กำลัง|รับประทาน=57

กำลัง|รับประทาน|อาหาร=63

รับประทาน|อาหาร|กลางวัน=45

3.2 คลังข้อมูลยูนิแกรมคำ (word-unigram corpus)

คลังข้อมูลยูนิแกรมคำในงานวิจัยนี้ หมายความถึงคลังข้อมูลที่เก็บข้อมูลเชิงสถิติของคำแต่ละคำคู่กับความถี่ในการปรากฏของคำๆ นั้นเอาไว้ ซึ่งผู้วิจัยเตรียมคลังข้อมูลนี้ขึ้นเพื่อนำไปใช้ในระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองยูนิแกรม เพื่อนำผลการทำงานที่ได้มาเปรียบเทียบกับผลการทำงานของระบบที่ใช้แบบจำลองไตรแกรม ตามสมมติฐานข้อที่ 2 ของงานวิจัยนี้ที่ผู้วิจัยระบุไว้ว่าผลการทำงานของระบบตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรมจะได้ผลถูกต้องมากกว่าด้วยแบบจำลองยูนิแกรม

3.2.1 การเตรียมคลังข้อมูลยูนิแกรมคำ

การเตรียมคลังข้อมูลยูนิแกรมคำนั้นคล้ายคลึงกับการเตรียมคลังข้อมูลไตรแกรมคำที่อธิบายไว้ในหัวข้อ 3.1.1 ในส่วนของการสกัดเอาประโยคข้อความภาษาไทยทั้งหมดออกมาจากคลังข้อมูลภาษาไทยแห่งชาจิ แต่จะต่างกันในขั้นตอนการนำคำแต่ละคำไปหาความถี่ในการปรากฏ คลังข้อมูลยูนิแกรมคำจะเก็บรวบรวมความถี่ในการปรากฏทั้งหมดของคำแต่ละคำในคลังข้อมูลภาษาไทยแห่งชาติไม่ใช่ชุดคำเรียงต่อกันสามคำ

3.2.2 ลักษณะโครงสร้างของคลังข้อมูลยูนิแกรมคำ

ข้อมูลในคลังข้อมูลยูนิแกรมคำประกอบด้วยข้อมูลสองส่วน ส่วนแรกคือข้อมูลยูนิแกรมคำ (word unigram) ตัวอย่างเช่น “ฉัน” “กำลัง” “รับประทาน” “อาหาร”เป็นยูนิแกรมคำของประโยค “ฉัน|กำลัง|รับประทาน|อาหาร” และส่วนที่สองคือข้อมูลความถี่ในการปรากฏของยูนิแกรมคำ ข้อมูลในส่วนแรกและส่วนที่สองจะถูกเก็บไว้ร่วมกันเป็นคู่ โดยมีเครื่องหมาย “=” คั่นกลาง ซึ่งมีลักษณะโครงสร้างดังตัวอย่างต่อไปนี้

ตัวอย่าง 3.5 แสดงตัวอย่างข้อมูลในคลังข้อมูลยูนิแกรมคำ (ความถี่ในตัวอย่างเป็นความถี่สมมติ)

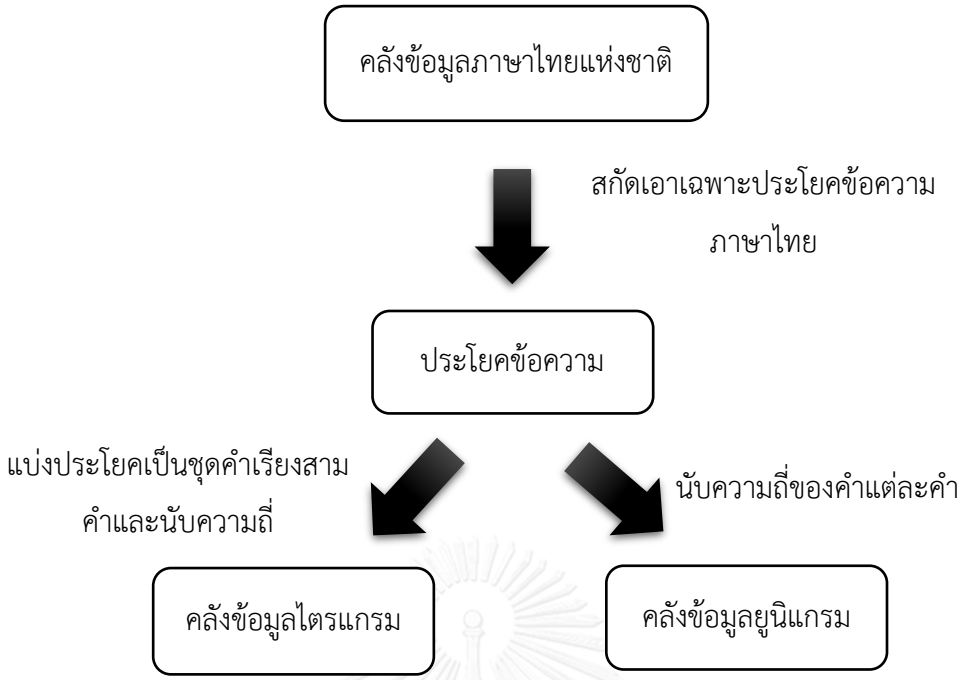
ยูนิแกรมคำ=ความถี่

ฉัน=2,278

กำลัง=4,152

รับประทาน=394

อาหาร=735



รูปภาพที่ 3.1 แสดงขั้นตอนในการเตรียมคลังข้อมูลไตรแกรมและคลังข้อมูลยูนิแกรม

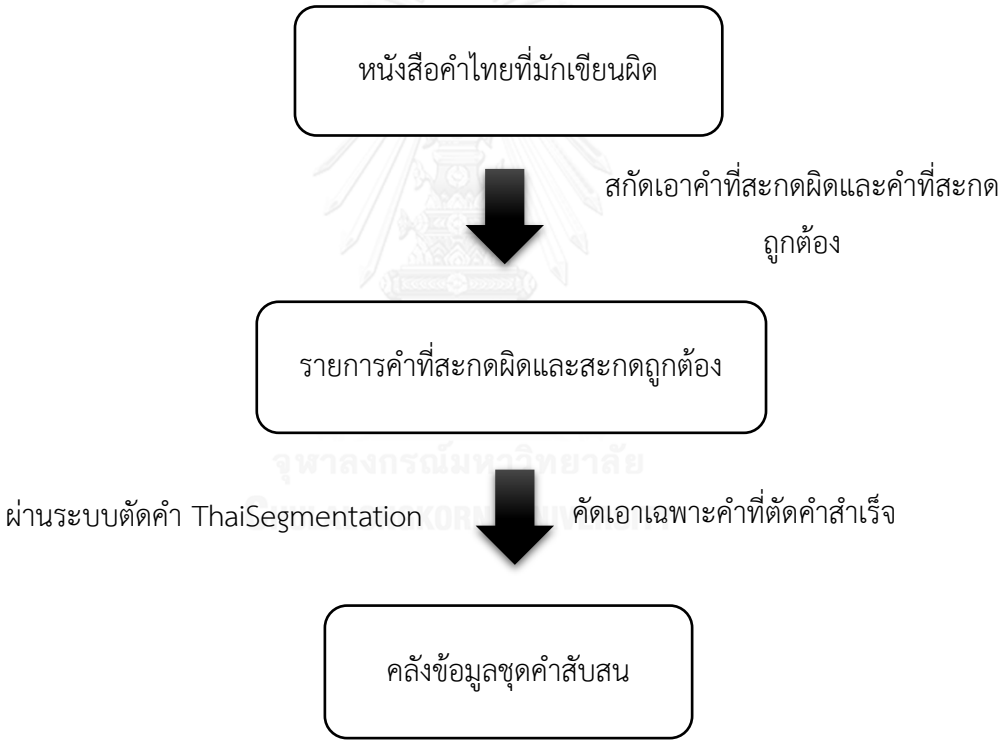
3.3 คลังข้อมูลชุดคำสับสน (confusion set corpus)

คลังข้อมูลชุดคำสับสนในงานวิจัยนี้ หมายถึง คลังข้อมูลที่ประกอบด้วยข้อมูล 2 ประเภท ประเภทแรกคือ คำไทยที่สะกดผิด และประเภทที่สองคือ คำที่สะกดถูกต้อง ซึ่งคำแต่ละคำในรายการคำที่สะกดผิดจะถูกเก็บไว้คู่กับคำที่ถูกต้องของคำผิดนั้น คลังข้อมูลชุดคำสับสนนี้เป็นคลังข้อมูลที่นำไปใช้ในการตรวจแก้สะกดผิดแบบเป็นคำจริงเช่นกัน ซึ่งผลการทำงานของระบบตรวจแก้การสะกดผิดที่ใช้คลังข้อมูลชุดคำสับสนในการตรวจแก้การสะกดผิดนี้จะถูกนำไปเปรียบเทียบกับผลการทำงานของระบบตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรมด้วย ตามที่ระบุในสมมติฐานข้อที่ 3 เอาไว้ว่าระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงที่ใช้คลังข้อมูลชุดคำสับสนในการแก้ไขการสะกดผิดจะใช้เวลาในการประมวลผลน้อยกว่าการใช้วิธีการตรวจน้อยที่สุด (minimum edit distance)

3.3.1 การเตรียมคลังข้อมูลชุดคำสับสน

ในการเตรียมคลังข้อมูลชุดคำสับสนนั้น ผู้วิจัยได้รวบรวมคำที่สะกดผิดจากหนังสือคำไทยที่มักเขียนผิดทั้งหมดห้าเล่ม ได้แก่ “อ่านอย่างไรและเขียนอย่างไร ฉบับราชบัณฑิตยสถาน (แก้ไขเพิ่มเติม) พิมพ์ครั้งที่ ๒๒” (ราชบัณฑิตยสถาน, 2557), “๒๘๐ คำไทยที่มักเขียนและใช้กันผิด” (ดนัย เมธิตานนท์, 2549) , “ร้อยแปด (๑๐๘) คำที่มักเขียนผิด” (ธนู ทดแทนคุณ, 2550), “คำไทยที่มักใช้

ผิด” (ตระการ เอี่ยมตระกูล, 2554) และ “ภาษาไทย คำที่มักเขียนผิด” (ฝ่ายวิชาการ พีบีซี, 2553) ที่ภายในบันทึกคำภาษาไทยที่มักเขียนผิดรวมถึงคำที่เขียนถูกต้องเอาไว้ เนื่องจากงานวิจัยนี้สนใจที่จะตรวจแก้เฉพาะการสะกดผิดแบบเป็นคำจริง ผู้วิจัยจึงได้นำข้อมูลคำเหล่านี้ไปผ่านกระบวนการตัดคำด้วยระบบตัดคำ“Thai Segmentation” (Aroonmanakun, 2002) แล้วคัดเอาเฉพาะคู่คำสะกดผิดและถูกต้องที่ผ่านกระบวนการตัดคำ เพราะคำที่มีความหมายและปรากฏในพจนานุกรมเท่านั้นที่ระบบจะสามารถตัดคำได้สำเร็จ ดังนั้นคำสะกดผิดที่ผ่านการตัดคำสำเร็จก็คือคำที่สะกดผิดแบบเป็นคำจริง ซึ่งมีจำนวนทั้งหมด 1,674 คู่คำ จากนั้นก็จัดเก็บคู่คำสะกดผิดและถูกต้องทั้งหมดนี้ไว้ในคลังข้อมูลชุดคำสับสน



รูปภาพที่ 3.2 แสดงขั้นตอนในการเตรียมคลังข้อมูลชุดคำสับสน

3.3.2 ลักษณะโครงสร้างของคลังข้อมูลชุดคำสับสน

ข้อมูลในคลังข้อมูลชุดคำสับสนประกอบด้วยข้อมูลสองส่วน ส่วนแรกคือข้อมูลคำสะกดผิด และส่วนที่สองคือข้อมูลคำที่สะกดถูกต้อง ข้อมูลในส่วนแรกและส่วนที่สองจะถูกเก็บไว้ร่วมกันเป็นคู่ โดยคั่นกลางด้วยแท็บ ซึ่งมีลักษณะโครงสร้างดังตัวอย่างต่อไปนี้

ตัวอย่าง 3.6 แสดงตัวอย่างข้อมูลในคลังข้อมูลชุดคำสับสน

คำที่สะกดผิด	**คำที่สะกดถูกต้อง**
|เกร็ด|ปลา|	|เกล็ด|ปลา|
|แกง|บวช|	|แกงบวด|
|คั่น|บันได|	|ขั้น|บันได|

3.3.3 ข้อมูลฝึกฝนและทดสอบ (training and test data)

เนื่องจากงานวิจัยนี้สนใจเฉพาะการตรวจแก้การสะกดผิดแบบเป็นคำจริงเท่านั้น เพราะฉะนั้นข้อมูลที่จะนำมาใช้ฝึกฝนและทดสอบตัวระบบก็ควรเป็นข้อมูลที่จะสะท้อนความสามารถในการตรวจแก้การสะกดผิดแบบเป็นคำจริงได้ ผู้วิจัยจึงได้นำเอาคำที่สะกดผิดในคลังข้อมูลชุดคำสับสนมาใช้เป็นคำสำคัญในการค้นหาประโยคตัวอย่างบนอินเตอร์เน็ต ด้วยเหตุผลที่ว่าคำที่สะกดผิดในคลังข้อมูลชุดคำสับสนนั้นล้วนผ่านการตัดคำสำเร็จ ในทางคอมพิวเตอร์ถือว่าเป็นคำที่สะกดผิดแบบเป็นคำจริงและการนำคำสะกดผิดไปค้นหาประโยคตัวอย่างบนอินเตอร์เน็ตนั้นก็เพื่อให้ได้ตัวอย่างประโยคที่ปรากฏการใช้จริง

3.3.3.1 การเตรียมข้อมูลฝึกฝน (training data)

ในขั้นแรกผู้วิจัยได้สุ่มเอาคำที่สะกดผิดในคลังข้อมูลชุดคำสับสนจำนวนหนึ่งไปค้นหาข้อความตัวอย่างที่มีคำสะกดผิดนั้นอยู่ในข้อความด้วยจาก google ซึ่งข้อความนั้นต้องมีความยาวระหว่าง 5-50 คำต่อหนึ่งข้อความ ตัวอย่างเช่น คำว่า “ขีดขั้น” เป็นคำที่สะกดผิดในคลังข้อมูลชุดคำสับสน ที่นำไปค้นหาข้อความตัวอย่างจาก google ได้ข้อความ “ตัวอย่างของเครื่องหมายขีดขั้นแบบไม่แบ่งส่วน” จากนั้นผู้วิจัยได้เก็บรวบรวมข้อความตัวอย่างในลักษณะเช่นนี้ไว้คู่กับคำสะกดผิดที่นำไปค้นและคำที่สะกดถูกต้องซึ่งในตัวอย่างนี้คือคำว่า “ขีดคั่น” เอาไว้ด้วยกันเป็นชุด ซึ่งรวบรวมได้จำนวนทั้งหมด 3,787 ชุด แล้วในขั้นตอนสุดท้ายผู้วิจัยก็นำข้อมูลทั้ง 3,787 ข้อมูล นี้ไปผ่านระบบตัดคำ “ThaiSegmentation 2.2”(Aroonmanakul, 2002) เพื่อคัดเลือกเอาเฉพาะตัวอย่างข้อความที่ตัดคำสำเร็จทั้งข้อความและเพื่อเป็นการช่วยระบุขอบเขตของคำแต่ละคำด้วยเครื่องหมาย |

3.3.3.2 การเตรียมข้อมูลทดสอบ (test data)

ขั้นตอนในการเตรียมข้อมูลทดสอบนั้นเหมือนกับการเตรียมข้อมูลฝึกฝน แต่จะแตกต่างกันตรงที่จำนวนข้อความภาษาไทยภายในข้อมูล ซึ่งจำนวนข้อความของข้อมูลทดสอบนั้นเท่ากับ 1,000 ข้อความ ที่ได้จากการสุ่มเลือกคำที่สะกดผิดแบบเป็นคำจริงจากคลังชุดคำสับสนมา 375 คำ แล้วนำไปสืบค้นบนอินเตอร์เน็ตเพื่อให้ได้ข้อความตัวอย่างของคำสะกดผิดที่เกิดขึ้นจริงและมีคำสะกดผิดที่สุ่มเลือกมาจากชุดคำสับสนนี้ปนอยู่อย่างน้อยหนึ่งคำต่อหนึ่งข้อความจำนวน 1,000 ข้อความ หรือ

ประมาณ 33,000 คำ แล้วนำข้อความที่จะใช้ทดสอบทั้งหมดนี้ไปผ่านกระบวนการตัดคำ แล้วคัดเลือกเอาเฉพาะข้อความที่ตัดคำได้สำเร็จทั้งข้อความ หากพบว่ามีข้อความใดตัดคำไม่สำเร็จ ผู้วิจัยก็จะหาข้อความตัวอย่างใหม่มาทดแทนให้มีจำนวนข้อมูลทดสอบครบ 1,000 ข้อความ ซึ่งภายในยังคงมีคำที่สะกดผิดแบบเป็นคำจริงอยู่เช่นเดิม

3.3.3.3 ลักษณะโครงสร้างของข้อมูลฝึกฝนและทดสอบ

ข้อมูลที่จะใช้ฝึกฝนและทดสอบแต่ละชุด จะประกอบด้วยข้อมูลทั้งหมด 3 ส่วน ส่วนแรกคือประโยคตัวอย่างซึ่งมีคำที่สะกดผิดแบบเป็นคำจริงอยู่ ส่วนที่สองคือคำที่สะกดผิดแบบเป็นคำจริงที่ปรากฏอยู่ในประโยค และส่วนสุดท้ายคือคำที่สะกดถูกต้อง โดยจะมีเครื่องหมาย “>>” คั่นกลางระหว่างข้อมูลแต่ละส่วน และมีเครื่องหมาย “|” เป็นตัวระบุขอบเขตของคำ ดังตัวอย่างด้านล่าง

ตัวอย่าง 3.7 แสดงตัวอย่างข้อมูลฝึกฝนและทดสอบ

ประโยคตัวอย่าง>>**คำที่สะกดผิด**>>**คำที่สะกดถูกต้อง**

|นอกจาก|ข้าวต้ม|มัด|แล้ว|ยัง|มี|แกง|บวช|ฟักทอง|>>แกง|บวช>>แกงบวด

คลังข้อมูลทั้งหมดที่จัดเตรียมไว้ในส่วนนี้จะถูกนำไปใช้เป็นองค์ประกอบสำคัญสำหรับตรวจจับและแก้ไขการสะกดผิดแบบเป็นคำจริงด้วยวิธีการที่แตกต่างกันไปตามประเภทของข้อมูลในคลังข้อมูล ซึ่งผู้วิจัยได้อธิบายไว้ในบทที่ 5 ส่วนในบทต่อไปผู้วิจัยจะกล่าวถึงการวิเคราะห์คำไทยที่มักเขียนผิด

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 4

การวิเคราะห์คำไทยที่มักเขียนผิด

เนื้อหาในบทนี้จะกล่าวถึงการนำคำไทยที่มักเขียนผิดมาศึกษาวิเคราะห์หาลักษณะและรูปแบบการสะกดผิดแบบเป็นคำจริง ซึ่งเนื้อหาส่วนแรกจะบอกถึงขั้นตอนวิธีการวิเคราะห์คำที่มักเขียนผิดและข้อมูลที่นำมาใช้ จากนั้นในส่วนที่สองจะกล่าวถึงผลการวิเคราะห์รูปแบบการสะกดผิดที่พบในการศึกษาครั้งนี้

4.1 ขั้นตอนการวิเคราะห์คำไทยที่มักเขียนผิด

ขั้นตอนที่ 1 เตรียมข้อมูล

สำหรับคำไทยที่มักเขียนผิดที่จะนำมาใช้ในการศึกษาวิเคราะห์ครั้งนี้ต้องเป็นคำที่สะกดผิดแบบเป็นคำจริงและต้องปรากฏการใช้จริงบนอินเตอร์เน็ตด้วยเพื่อให้ตรงตามวัตถุประสงค์ข้อที่ 1 ของงานวิจัยนี้ ผู้วิจัยจึงได้นำคำที่สะกดผิดจำนวน 1,674 คำจากคลังข้อมูลชุดคำสับสนในหัวข้อ 3.3 มาใช้ในการศึกษาวิเคราะห์ ซึ่งล้วนเป็นคำที่สะกดผิดแบบเป็นคำจริง จากนั้นจึงนำคำที่สะกดผิดเหล่านั้นไปค้นบนอินเตอร์เน็ตเพื่อหาตัวอย่างการใช้จริงของคำเหล่านั้นมาเป็นข้อมูล

ขั้นตอนที่ 2 กำกับข้อมูล

ในขั้นตอนนี้ ผู้วิจัยได้นำแนวคิดวิธีการปรับแก้น้อยครั้งสุด (minimum edit distance) มาใช้ช่วยในการหารูปแบบของการสะกดผิด โดยมองว่าความแตกต่างของคำที่สะกดถูกและคำที่สะกดผิดนั้นเกิดจากการปรับแก้อย่างน้อย 1 จาก 3 แบบต่อไปนี้ ได้แก่ การเติม (insertion) การลบ (deletion) และการแทนที่ (substitution) กล่าวคือ ถ้าหากว่าจำนวนตัวอักษรของคำที่สะกดผิดนั้นมากกว่าคำที่สะกดถูกแสดงว่าการสะกดผิดนี้เกิดจากการเติมตัวอักษร x เข้าไปในคำที่สะกดผิด ซึ่งคำที่สะกดผิดนี้จะถูกกำกับรูปแบบการสะกดผิดเป็น “0_x” โดย 0 หมายถึง ความว่างเปล่า เครื่องหมาย _ หมายถึง เปลี่ยนไปเป็น และ x หมายถึงตัวอักษรที่เติมเข้าไปในคำที่สะกดถูกต้อง ตัวอย่างเช่น คำว่า “มาตรการ” เขียนผิดเป็น “มาตราการ” จะเห็นว่าเมื่อเติม “สระ -า” เข้าไป คำที่สะกดถูกก็จะกลายเป็นคำที่สะกดผิด ดังนั้นข้อมูลนี้จะถูกกำกับรูปแบบของการสะกดผิดเป็น “0_า” แต่ถ้าหากว่าจำนวนตัวอักษรของคำที่สะกดผิดน้อยกว่าคำที่สะกดถูก แสดงว่าการสะกดผิดนี้เกิดจากการลบตัวอักษร x ออกจากคำที่ถูกต้อง โดยคำสะกดผิดในลักษณะนี้จะถูกกำกับข้อมูลด้วย “x_0” โดย x หมายถึงตัวอักษรที่ถูกลบออกจากคำสะกดถูก ตัวอย่างเช่น คำว่า “อุกกาบาต” เขียนผิดเป็น “อุกาบาต” จะเห็นได้ว่าเมื่อลบ “ก” ในคำที่สะกดถูกออกหนึ่งตัว ก็จะกลายเป็นคำที่สะกดผิด ดังนั้นข้อมูล

นี้จะถูกกำกับข้อมูลเป็น “ก_0” และถ้าหากว่าจำนวนตัวอักษรของคำที่สะกดถูกและสะกดผิดมีจำนวนเท่ากัน แสดงว่าการสะกดผิดนี้เกิดจากการแทนที่ตัวอักษร x ด้วยตัวอักษร y ซึ่งการสะกดผิดลักษณะนี้จะถูกกำกับข้อมูลเป็น “x_y” โดย x หมายถึงตัวอักษรในคำที่สะกดถูกซึ่งถูกแทนที่ด้วยตัวอักษร y และ y หมายถึงตัวอักษรในคำที่สะกดผิดซึ่งปรากฏแทนที่ตัวอักษร x ตัวอย่างเช่น คำว่า “ลมหวน” เขียนผิดเป็น “ลมหวล” จะเห็นได้ว่าเมื่อ น ในคำที่สะกดถูกนั้นถูกแทนที่ด้วย ล ก็จะกลายเป็นคำที่สะกดผิด ซึ่งข้อมูลที่มีการสะกดผิดในลักษณะนี้จะถูกกำกับรูปแบบการสะกดผิดเป็น “น_ล” ด้วยวิธีการนี้จะทำให้สามารถระบุและกำกับข้อมูลรูปแบบของการสะกดผิดหรือรูปตัวอักษรอะไรเปลี่ยนไปเป็นอะไรเทียบจากคำสะกดถูกไปเป็นคำสะกดผิด จำนวนการสะกดผิดที่ปรับแก้ และการสะกดผิดปรากฏที่ตำแหน่งใดบ้างตามโครงสร้างคำภาษาไทย ได้แก่ พยัญชนะต้น สระ วรรณยุกต์ ตัวสะกด และตัวการันต์ (สุนันท์ อัญชลีนุกูล, 2552) ผู้วิจัยได้กำกับข้อมูลต่างๆ เหล่านี้เพื่อนำไปวิเคราะห์ในขั้นตอนต่อไป ซึ่งข้อมูลที่ได้รับการกำกับข้อมูลเรียบร้อยแล้วจะมีลักษณะดังที่ปรากฏในตารางที่ 1

ตารางที่ 4.1แสดงตัวอย่างข้อมูลการสะกดผิดแบบเป็นคำจริงในภาษาไทยที่นำมาวิเคราะห์

คำที่สะกดถูกต้อง	**มักสะกดผิดเป็น**	**จำนวนการสะกดผิดที่พบในหนึ่งคำ**	**ตำแหน่งของการสะกดผิด**	**รูปแบบการสะกดผิด**
ราดหน้า	ลาดหน้า	1	พยัญชนะต้น	ร_ล (ร ถูกแทนที่ด้วย ล)
วิกฤติกาล	วิกฤตการ	2	สระ ตัวสะกด	ิ_0 (ลบสระ อิ) ล_ร (ล ถูกแทนที่ด้วย ร)
นานัปการ	นานับประการ	3	ตัวสะกด พยัญชนะต้น สระ	0_บ (เติม บ) 0_ร (เติม ร) 0_ะ (เติมสระ อะ)

ขั้นตอนที่ 3 วิเคราะห์ข้อมูล

หลังจากที่กำกับข้อมูลเสร็จเรียบร้อยแล้ว ผู้วิจัยได้นำข้อมูลเหล่านั้นมาศึกษาวิเคราะห์โดยอาศัยข้อมูลต่างๆ ที่ได้กำกับคำที่สะกดผิดแต่ละคำเอาไว้ไม่ว่าจะเป็นจำนวนการสะกดผิด ตำแหน่งที่สะกดผิด และรูปแบบของการสะกดผิดที่พบในแต่ละคำว่ามีความเหมือนหรือความต่างกันอย่างไรบ้างเพื่อที่จะสามารถนำมาใช้จำแนกหรือจัดประเภทของคำที่สะกดผิดตามลักษณะหรือรูปแบบที่ปรากฏได้

4.2 รูปแบบการสะกดผิด

จากการศึกษาวิเคราะห์คำที่สะกดผิดแบบเป็นคำจริงของคำที่มักเขียนผิดในครั้งนี้พบว่า เมื่อพิจารณาจำนวนการสะกดผิดที่ปรับแก้ในคำแต่ละคำดังตัวอย่างในตารางที่ 1 สามารถจำแนกข้อมูลคำที่สะกดผิดเหล่านี้ออกเป็น 2 กลุ่มใหญ่ คือ กลุ่มของคำที่มีการสะกดผิดหนึ่งตำแหน่ง และ กลุ่มของคำที่สะกดผิดหลายตำแหน่ง โดยพบว่ามีจำนวนคำที่สะกดผิดในกลุ่มแรก 1,339 คำ และกลุ่มที่สอง 335 คำ จากทั้งหมด 1,674 คำ หรือร้อยละ 80 และ 20 ตามลำดับ

4.2.1 คำที่สะกดผิดหนึ่งตำแหน่ง

เมื่อศึกษาวิเคราะห์คำที่สะกดผิดกลุ่มนี้ตามตำแหน่งที่ปรากฏการสะกดผิดพบว่าสามารถแบ่งคำในกลุ่มนี้ออกเป็น 5 กลุ่มตามโครงสร้างคำไทยที่ประกอบด้วย พยัญชนะต้น สระ วรรณยุกต์ ตัวสะกด และตัวการันต์ แล้ววิเคราะห์รูปแบบของการสะกดผิดในแต่ละกลุ่มตามวิธีการปรับแก้จากคำที่สะกดถูกไปเป็นคำที่สะกดผิดใน 3 ลักษณะคือ การเติม การลบ และการแทนที่ ซึ่งผลจากการศึกษาวิเคราะห์มีดังต่อไปนี้

4.2.1.1 คำสะกดผิดที่พยัญชนะต้น

จากการศึกษาพบว่ามีคำที่สะกดผิดในตำแหน่งพยัญชนะต้นมีจำนวนมากที่สุดคือมีจำนวน 382 คำ จากทั้งหมด 1,674 คำหรือร้อยละ 22.82 หรือเมื่อเทียบกับคำที่สะกดผิดหนึ่งตำแหน่งทั้งหมด 1,339 คำ จะคิดเป็นร้อยละ 28.53 ซึ่งตัวอย่างของการสะกดผิดที่พบมากที่สุดในกลุ่มนี้ คือ การแทนที่ ร ด้วย ล เช่น คำว่า “ร่องน้ำ” “ราดหน้า” “หารือ” สะกดผิดเป็น “ล่องน้ำ” “ลาดหน้า” “หาลือ” ตามลำดับ เป็นต้น และผลการวิเคราะห์รูปแบบการสะกดผิดตามวิธีการปรับแก้ 3 แบบมีดังนี้

ก คำสะกดผิดที่มีการแทนที่พยัญชนะต้น

จากการวิเคราะห์คำที่สะกดผิดในตำแหน่งพยัญชนะต้น 382 คำ พบว่าร้อยละ 63.61 หรือ 243 คำ เป็นการสะกดผิดที่มีการแทนที่พยัญชนะต้นด้วยพยัญชนะตัวอื่นที่ออกเสียงคล้ายกันหรือเหมือนกันแต่มีรูปเขียนต่างกัน เช่น การใช้พยัญชนะ ร /r/ และ ล /l/ ซึ่งเป็นพยัญชนะในภาษาไทยที่ในความเป็นจริงแล้วมักจะถูกใช้สลับกันอยู่เป็นประจำ และถึงแม้ว่าจะออกเสียงพยัญชนะสองตัวนี้สลับกันผู้ฟังก็ยังคงเข้าใจในสิ่งที่ผู้พูดต้องการสื่อ ด้วยบริบทในการสนทนาจึงไม่ส่งผลต่อความหมายเช่น “นั่งเลือไปซื้อพลิก” ด้วยเหตุนี้ความถูกต้องในการออกเสียงของคำจึงอาจเป็นปัจจัยสำคัญที่ส่งผลให้พยัญชนะสองตัวนี้ถูกใช้สลับกันในรูปเขียนและทำให้เกิดการสะกดผิดในลักษณะนี้ จากผลการวิเคราะห์ยังพบว่ามีคำที่สะกดผิดโดยเขียนพยัญชนะสองตัวนี้สลับกัน คือใช้ ร

แทน ล หรือ ร แทน ล ในตำแหน่งพยัญชนะต้นมากถึง 135 คำ จาก 382 คำหรือร้อยละ 35.34 ตัวอย่างเช่น คำว่า “พรางตา” สะกดผิดเป็น “พลางตา” และ “เกล็ดปลา” สะกดผิดเป็น “เกร็ดปลา” เป็นต้น นอกจากนี้ยังพบว่ามีการสะกดผิดในลักษณะนี้เกิดขึ้นกับพยัญชนะตัวอื่นๆ ด้วย ตัวอย่างเช่น การใช้พยัญชนะ น แทน ณ, ใช้ ส แทน ศ, และใช้ ทร แทน ซ ในคำว่า “ปราณี” สะกดผิดเป็น “ปรานี”, “ประกาศิต” สะกดผิดเป็น “ประกาสิต” และ “ซาบซึ้ง” สะกดผิดเป็น “ทราบซึ้ง” ตามลำดับ ซึ่งจะเห็นได้ว่าพยัญชนะต้นแต่ละคู่ที่สะกดผิดในลักษณะนี้มักเป็นรูปเขียนที่ใช้แทนหน่วยเสียงเดียวกัน คือ น และ ณ แทนหน่วยเสียง /n/ และ ส ศ ซ และทร(ในคำว่า “ทราบ”) แทนหน่วยเสียง /s/

ข คำสะกดผิดที่มีการเติมพยัญชนะต้น

จากการวิเคราะห์พบว่ามีคำที่สะกดผิดในลักษณะนี้ร้อยละ 20.94 หรือจำนวน 80 คำจากคำที่สะกดผิดในตำแหน่งพยัญชนะต้นทั้งหมด 382 คำ โดยพบว่าคำที่สะกดผิดในลักษณะนี้ส่วนใหญ่เป็นคำควบกล้ำ โดยเฉพาะคำควบกล้ำที่มี ร และ ล ประสมในอักษรควบ และตัวอย่างของคำสะกดผิดในลักษณะนี้ที่พบมากที่สุดคือ การเติม ร หลังพยัญชนะต้น ก ได้เป็นอักษรควบกล้ำ กร เช่น “กะทันหัน” “กะพริบ” สะกดผิดเป็น “กระทันหัน” “กระพริบ” เป็นต้น และตัวอย่างที่พบรองลงมา คือ การเติม ล หลังพยัญชนะต้น ผ หรือ ก ได้เป็นอักษรควบกล้ำ ผล หรือ กล เช่น “ผัดผ่อน” สะกดผิดเป็น “ผลัดผ่อน” และ “วางก้าม” สะกดผิดเป็น “วางกล้าม” เป็นต้น

ค คำสะกดผิดที่มีการลบพยัญชนะต้น

คำที่สะกดผิดในลักษณะนี้มีจำนวน 59 คำจาก 382 คำหรือร้อยละ 15.45 ซึ่งคำสะกดผิดที่มีการลบพยัญชนะต้นส่วนใหญ่ก็เป็นคำควบกล้ำที่มีพยัญชนะต้นตัวที่สองเป็น ร หรือ ล ซึ่งตัวอย่างการสะกดผิดลักษณะนี้ที่พบบ่อยคือ การพิมพ์ตกพยัญชนะควบกล้ำ ร ที่ตามหลังพยัญชนะต้น ก ตัวอย่างเช่น “กระเพาะ” สะกดผิดเป็น “กะเพาะ” และไม่ใส่ ล หรือ ร หลังพยัญชนะต้น ผ และ ป ตามลำดับ เป็นตัวอย่างการสะกดผิดลักษณะนี้ที่พบรองลงมา ตัวอย่างเช่น “ผลัดเวร” สะกดผิดเป็น “ผัดเวร” และ “ประปา” สะกดผิดเป็น “ปะปา” เป็นต้น

4.2.1.2 คำสะกดผิดที่ตัวสะกด

จากการวิเคราะห์พบว่าจากคำที่สะกดผิดหนึ่งตำแหน่ง 1,339 คำ มีคำที่สะกดผิดที่ตัวสะกดจำนวน 338 คำ คิดเป็นร้อยละ 25.24 หรือเท่ากับร้อยละ 20.20 เมื่อคำนวณจากคำสะกดผิดที่ศึกษาทั้งหมด 1,674 คำ และเมื่อวิเคราะห์คำในกลุ่มนี้ตามวิธีการปรับแก้ 3 แบบ ได้ผลดังนี้

ก คำสะกดผิดที่มีการแทนที่ตัวสะกด

คำสะกดผิดในรูปแบบนี้มักเป็นคำที่แทนที่ตัวสะกดที่ถูกต้องด้วยตัวสะกดอื่นในมาตราตัวสะกดเดียวกัน จึงทำให้เกิดการสะกดผิดที่ยังคงออกเสียงเช่นเดิมอยู่ คำสะกดผิดกลุ่มนี้พบ

มากที่สุดคือ 315 คำ หรือ 93.2% จากคำสะกดผิดที่ตัวสะกดทั้งหมด 338 คำ ตัวอย่างการสะกดผิดที่พบมากเป็นอันดับต้นๆ ได้แก่ การใช้ตัวสะกด ล แทน น เช่น “ทูนหัว” สะกดผิดเป็น “ทูลหัว” การใช้ น แทน ล เช่น “กังวล” สะกดผิดเป็น “กังวน” และการใช้ น แทน ณ เช่น “เกษียณอายุ” สะกดผิดเป็น “เกษียนอายุ" เป็นต้น

ข คำสะกดผิดที่มีการเติมตัวสะกด

คำสะกดผิดในรูปแบบนี้เกิดจากการเติมตัวสะกดเพิ่มเข้าไปในคำที่ถูกต้องทำให้กลายเป็นคำที่สะกดผิด พบทั้งหมด 12 คำ หรือร้อยละ 3.55 ตัวอย่างเช่น การเติมตัวสะกด น ใน คำว่า “ทรมาทรกรรม” สะกดผิดเป็น “ทรมานทรกรรม” หรือเติมตัวสะกด ม ในคำว่า “กรรมาชีพ” ซึ่งสะกดผิดเป็น “กรรมมาชีพ” เป็นต้น

ค คำสะกดผิดที่มีการลบตัวสะกด

คำที่สะกดผิดเนื่องจากตัวสะกดของคำที่สะกดถูกต้องบางตัวหายไปทำให้คำนั้นสะกดผิด มีทั้งหมด 11 คำหรือ 3.25% เช่น การลบตัวสะกด ษ ออก จากคำว่า “อธิษฐาน” ทำให้สะกดผิดเป็น “อธิฐาน” เป็นต้น

4.2.1.3 คำสะกดผิดที่สระ

การสะกดผิดในลักษณะนี้เกิดจากการเลือกใช้สระไม่ถูกต้อง ซึ่งพบจำนวนคำที่สะกดผิดในลักษณะนี้จำนวน 306 คำ เท่ากับร้อยละ 18.28 จากคำสะกดผิดทั้งหมดที่ศึกษาและเท่ากับร้อยละ 22.85 จากคำที่สะกดผิดหนึ่งตำแหน่งทั้งหมด และเมื่อจำแนกคำที่สะกดผิดในลักษณะนี้ออกเป็น 3 กลุ่มย่อย ตามวิธีการปรับแก้ 3 แบบ ได้ผลดังนี้

ก คำสะกดผิดที่มีการเติมสระ

การสะกดผิดในลักษณะนี้เกิดจากการใส่รูปสระเพิ่มเข้าไปในคำ ซึ่งจากการวิเคราะห์พบว่าประมาณครึ่งหนึ่งของคำสะกดผิดที่สระจากทั้งหมด 306 คำ มีคำที่สะกดผิดในลักษณะนี้อยู่ 151 คำ หรือร้อยละ 49.35 นอกจากนี้ยังพบว่าคำที่สะกดผิดในลักษณะนี้มักเป็นคำที่ไม่ประวิสรรชนีย์ หรือก็คือคำที่มีเสียงสระ อะ ประสมอยู่แต่ไม่ปรากฏในรูปเขียน โดยธรรมชาติของคำที่ไม่ประวิสรรชนีย์นั้นจะออกเสียงสระ อะ เพียงกึ่งเสียง ตัวอย่างเช่น “ธุรกิจ” สะกดผิดเป็น “ธุระกิจ” “สไบ” สะกดผิดเป็น “สะไบ” “ชบา” สะกดผิดเป็น “ชะบา” “ลออ”สะกดผิดเป็น “ละออ” เป็นต้น อย่างไรก็ตามจากตัวอย่างข้างต้น หากลองออกเสียงสระ อะ ในคำเหล่านี้เต็มเสียงจะพบว่าไม่มีผลต่อการรับรู้หรือเข้าใจความหมายของคำเหล่านี้ ดังนั้นความยืดหยุ่นในวิธีการออกเสียงคำเหล่านี้อาจเป็นปัจจัยสำคัญที่ทำให้เกิดการสะกดผิดในลักษณะนี้ นอกจากการเติมสระ อะ ในคำที่ไม่ประวิสรรชนีย์แล้ว ยังพบการสะกดผิดจากการเติมรูปสระอื่นๆ อีก เช่น สระอา สระโอ สระอิ เป็นต้น ตัวอย่างเช่น

คำว่า “มาตรฐาน” สะกดผิดเป็น “มาตราฐาน” “หยักศก” สะกดผิดเป็น “หยักโศก” “กริยา” สะกดผิดเป็น “กิริยา” เป็นต้น

ข คำสะกดผิดที่มีการลบสระ

ลักษณะของการสะกดผิดนี้คือการที่มีรูปสระบางตัวตกหล่นไป ซึ่งมีจำนวน 99 คำ หรือร้อยละ 32.35 ตัวอย่างของการสะกดผิดในลักษณะนี้ที่พบมากที่สุดคือ การไม่ประวิสรรชนีย์ให้กับคำที่ประวิสรรชนีย์ ซึ่งคำที่ประวิสรรชนีย์นั้นจะออกเสียงสระเต็มเสียงเฉพาะพยางค์ที่อยู่ท้ายคำ ถ้าอยู่ตำแหน่งอื่นออกเสียงกึ่งเสียง ตัวอย่างเช่น คำว่า “อิสระ” สะกดผิดเป็น “อิสร” “ชะงัด” สะกดผิดเป็น “ชงัด” “สะดุ้ง สะกดผิดเป็น “สดุ้ง” เป็นต้น การสะกดคำเหล่านี้ผิดอาจจะมีสาเหตุหลักมาจากวิธีการออกเสียงดังที่กล่าวไว้ในหัวข้อที่แล้ว รูปสระอื่นในการสะกดผิดลักษณะนี้ที่พบมากรองลงมา ได้แก่ การลบไม้ไต่คู้ออก (ั) ซึ่งเป็นเครื่องหมายที่เป็นส่วนประกอบของสระแอะ (แ็ะ) และสระเอาะ(็อ_) ตัวอย่างเช่น “นอนแบ็บ” สะกดผิดเป็น “นอนแบบ” “ผล็อย” สะกดผิดเป็น “ผลอย” เป็นต้น

ค คำสะกดผิดที่มีการแทนที่สระ

การสะกดผิดนี้เกิดจากการเลือกใช้สระไม่ถูกต้อง ซึ่งพบว่ามีคำที่สะกดผิดในลักษณะนี้ 56 คำ จาก 306 คำ เท่ากับ 18.30% ซึ่งการสะกดผิดในลักษณะนี้ตัวอย่างที่พบมากที่สุดจะเป็นการเลือกใช้รูปสระผิดระหว่าง สระ ไ- และ ใ- ตัวอย่างเช่น “ไต้ฝุ่น” เขียนผิดเป็น “ใต้ฝุ่น” และ “เยื่อใย” เขียนผิดเป็น “เยื่อไย” เป็นต้น รองลงมาคือการเลือกใช้รูปสระอะ แทน สระอา ตัวอย่างเช่น “ประณีต” สะกดผิดเป็น “ปราณีต” “จะละเม็ด” สะกดผิดเป็น “จาละเม็ด” เป็นต้น

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

4.2.1.4 คำสะกดผิดที่ตัวการันต์

จากการวิเคราะห์พบว่ามีคำสะกดผิดในตำแหน่งตัวการันต์จำนวน 270 คำ จากคำสะกดผิดที่นำมาศึกษาทั้งหมด 1,674 คำ ซึ่งคิดเป็นเปอร์เซ็นต์ได้ 16.13% และ 20.16% ของกลุ่มคำที่สะกดผิดหนึ่งตำแหน่ง 1,339 คำ หลังจากที่ได้ศึกษาวิเคราะห์คำที่มักเขียนผิดในกลุ่มนี้ตามวิธีการปรับแก้ 3 แบบ ได้ผลดังนี้

ก คำสะกดผิดที่มีการเติมตัวการันต์

การสะกดผิดลักษณะนี้เป็นการสะกดผิดที่มีตัวอักษรหรือเครื่องหมายการันต์ (์) เพิ่มเข้าไปในส่วนของตัวการันต์ทำให้สะกดผิด คำสะกดผิดในกลุ่มนี้เป็นการสะกดผิดที่ตัวการันต์ที่พบว่ามีมากที่สุดคือ 130 คำ จากทั้งหมด 270 คำ หรือคิดเป็น 48.15% ตัวอย่างเช่น การเติม ย์ ในคำว่า “อินทรี” (นก) สะกดผิดเป็น “อินทรีย์” (สาร) หรือ “บัณฑิต” สะกดผิดเป็น “บัณฑิตย์” ซึ่งตัวอย่างที่พบรองลงมาได้แก่ การเติม ค์ หรือ ณ์ ตัวอย่างเช่น “คัดสรร” สะกดผิดเป็น “คัดสรรค์” “รักษาการ” สะกดผิดเป็น “รักษาการณ์” ตามลับดับ

ข คำสะกดผิดที่มีการลบตัวการันต์

คำที่สะกดผิดในกลุ่มนี้เป็นคำที่ขาดตัวการันต์ไปแล้วทำให้สะกดผิด ซึ่งพบว่ามีคำที่สะกดผิดในกลุ่มนี้ทั้งหมด 98 คำ หรือ 36.30% ตัวอย่างการสะกดผิดลักษณะที่พบมากที่สุด คือ คำที่ตก ณ์ เช่นคำว่า “ประสบการณ์” ได้เป็นคำที่สะกดผิด “ประสบการ” และตัวอย่างที่พบมากรองลงมา คือ คำที่ตก ย์ ค์ ห์ ฯ ตามลำดับ เช่น “ครองราชย์” สะกดผิดเป็น “ครองราช” “สังสรรค์” สะกดผิดเป็น “สังสรร” “มโนราห์” สะกดผิดเป็น “มโนรา” เป็นต้น

ค คำสะกดผิดที่มีการแทนที่ตัวการันต์

ส่วนคำในกลุ่มนี้เป็นคำที่สะกดผิดเนื่องจากใช้ตัวการันต์ไม่ถูกต้อง ซึ่งมีคำที่สะกดผิดในกลุ่มนี้ทั้งหมด 42 คำหรือร้อยละ 15.56 ซึ่งตัวอย่างการสะกดผิดในลักษณะนี้ที่พบมากที่สุด คือ การใช้ ร์ แทน น์ เช่นในคำว่า “จันทน์เทศ” “ดอกไม้จันทน์” “จันทน์ผา” ซึ่งสะกดผิดเป็น “จันทร์เทศ” ดอกไม้จันทร์” “จันทร์ผา” ตามลำดับ ตัวอย่างการสะกดผิดที่พบรองลงมา คือ การใช้ ส์ แทน ฆ์ และ ศ์ แทน ค์ เช่น “ผลานิสงส์” สะกดผิดเป็น “ผลานิสงฆ์” “ดุริยางค์” เขียนผิดเป็น “ดุริยางศ์” เป็นต้น

4.2.1.5 คำสะกดผิดที่วรรณยุกต์

จากการศึกษาพบว่าการสะกดผิดหนึ่งตำแหน่งที่เกิดจากความผิดพลาดในการใช้วรรณยุกต์นั้นเป็นการสะกดผิดที่พบน้อยที่สุดคือ 43 คำ จากทั้งหมด 1,674 คำ หรือเท่ากับร้อยละ 2.57 หรือเท่ากับร้อยละ 3.21 จากคำที่สะกดผิดหนึ่งตำแหน่งทั้งหมด 1,339 คำ หลังจากที่ได้ศึกษาวิเคราะห์คำที่มักเขียนผิดในกลุ่มนี้ตามวิธีการปรับแก้ 3 แบบ ได้ผลดังนี้

ก คำสะกดผิดที่มีการเติมวรรณยุกต์

คำที่สะกดผิดในกลุ่มนี้เป็นคำที่มีรูปวรรณยุกต์เกินมาซึ่งเป็นลักษณะการสะกดผิดที่วรรณยุกต์ที่พบมากที่สุด คือ 22 คำจากทั้งหมด 43 คำ หรือเท่ากับ 51.16% โดยรูปวรรณยุกต์ที่มักถูกเติมเข้าไปในคำแล้วทำให้เกิดการสะกดผิดในลักษณะนี้คือ วรรณยุกต์เอก ตรี โท และจัตวา ไล่เรียงตามจำนวนตัวอย่างคำสะกดผิดที่พบ ตัวอย่างเช่น “จักจั่น” สะกดผิดเป็น “จั๊กจั่น”, “ตังโอ๋” สะกดผิดเป็น “ตั้งโอ๋” และ “กวยจี๊” สะกดผิดเป็น “ก๋วยจี๊” เป็นต้น

ข คำสะกดผิดที่มีการแทนที่วรรณยุกต์

คำที่สะกดผิดในกลุ่มนี้เป็นคำที่มีการสะกดผิดเพราะใช้รูปวรรณยุกต์ผิดเพี้ยนไปจากเดิม ซึ่งมีทั้งหมด 15 คำ คิดเป็น 34.88% ซึ่งการสะกดผิดในลักษณะนี้มักจะมีรูปแบบดังนี้คือ ใช้รูปวรรณยุกต์ตรีแทนที่รูปวรรณยุกต์โท หรือใช้รูปวรรณยุกต์เอกแทนที่รูปวรรณยุกต์โท ตัวอย่างเช่น “ปลั๊กไฟ” สะกดผิดเป็น “ปลั้กไฟ” หรือ “เลื้อยเจื้อย สะกดผิดเป็น เลื่อยเจื้อย” เป็นต้น

ค คำสะกดผิดที่มีการลบวรรณยุกต์

ส่วนคำในกลุ่มนี้เป็นคำที่สะกดผิดเนื่องจากการขาดรูปวรรณยุกต์ที่ถูกต้องซึ่งจากการวิเคราะห์พบว่ามีคำที่สะกดผิดในลักษณะนี้ 6 คำ หรือเท่ากับ 13.96% และรูปแบบการสะกดผิดที่มักพบในลักษณะนี้ ได้แก่ การไม่ใส่รูปวรรณยุกต์ตรี ตัวอย่างเช่น “ก๊าซ” เขียนผิดเป็น “กาซ” เป็นต้น

4.2.2 คำที่สะกดผิดหลายตำแหน่ง

คำสะกดผิดที่ถูกจำแนกไว้ในกลุ่มนี้เป็นคำที่พบการสะกดผิดมากกว่าหนึ่งตำแหน่ง เช่นคำว่า “จันทน์กพ้อ” สะกดผิดเป็น “จันทร์กะพ้อ” จะเห็นได้ว่ามีจุดที่สะกดผิดอยู่สองตำแหน่ง ได้แก่ การสะกดผิดในการเลือกใช้ตัวการันต์คือใช้ ร แทน น และการสะกดผิดในการประวิสรรชนีย์ คือ ไม่ควรใส่รูปสระอะในคำๆ นี้ คำว่า “ขโมย” สะกดผิดเป็น “โขมย" ซึ่งพบการสะกดผิดสองตำแหน่งเช่นกัน คือ การลบพยัญชนะต้น ข หรือ สระ โ- แล้วจึงเติมสระ โ- หน้าพยัญชนะ ข หรือเติมพยัญชนะ ข หลังสระ โ- หรือคำว่า “ขันลงหิน” สะกดผิดเป็น “ขันรองหิน” ที่มีการใช้ ร แทน ล และเติม อ หรือ สระ ออ เพิ่มเข้าไปอีก ซึ่งจากการศึกษาพบว่ามีคำที่สะกดผิดในลักษณะเช่นนี้อยู่ทั้งหมด 335 คำ หรือร้อยละ 20 ของคำทั้งหมดในคลังคำที่ใช้ศึกษาวิเคราะห์ โดยสามารถจำแนกออกได้เป็น 2 กลุ่ม ดังนี้

4.2.2.1 คำที่สะกดผิดหลายตำแหน่งแต่ออกเสียงเหมือนเดิม

คำที่สะกดผิดหลายตำแหน่งในกลุ่มแรกเป็นกลุ่มของคำที่มีการสะกดผิดโดยเขียนพยางค์ในรูปแบบอื่นที่มีเสียงคงเดิมหรือเป็นคำพ้องเสียงที่มีรูปเขียนต่างกันและมีการสะกดผิดที่ต้องปรับแก้มากกว่าหนึ่งตำแหน่ง จากการวิเคราะห์พบว่ามีคำที่สะกดผิดในลักษณะนี้อยู่จำนวน 239 คำ หรือเท่ากับร้อยละ 71.34 ของคำที่สะกดผิดหลายตำแหน่ง และร้อยละ 14.28 ของคำสะกดผิดในการศึกษาครั้งนี้ทั้งหมด ตัวอย่างการสะกดผิดในลักษณะนี้ที่พบมากคือ การใช้รูป ั-น แทน รูป ร หัน - รร ซึ่งความผิดพลาดในลักษณะนี้มักจะเกิดขึ้นกับคำที่ขึ้นต้นด้วยคำว่า “กรร-” หรือ “บรร-” เช่นคำว่า “กรรแสง” สะกดผิดเป็น “กันแสง” และคำว่า “บรรทัด” สะกดผิดเป็น “บันทัด" เป็นต้น คำเหล่านี้เป็นคำที่มีการสะกดผิดหลายตำแหน่งเพราะต้องปรับแก้มากกว่าหนึ่งครั้ง คือต้อง แทนที่ ร ตัวแรกด้วย ั- และแทนที่ ร ตัวที่สองด้วย น นอกจากนี้ยังมี คำว่า “สูญ” กับ “ศูนย์” และ “ปัด” กับ “ปัทม์” ซึ่งทั้งหมดเป็นคำพ้องเสียงที่มีรูปเขียนไม่เหมือนกัน หากใช้ไม่ถูกบริบทก็อาจจะทำให้ความเหมายผิดเพี้ยนไปได้ ตัวอย่างเช่นคำว่า “ศูนย์หน้า” สะกดผิดเป็น “สูญหน้า" และ “หน้าปัทม์” สะกดผิดเป็น “หน้าปัด” ตามลำดับ นอกจากนี้ยังมีคำที่สะกดผิดในลักษณะนี้อยู่อีกกลุ่มหนึ่งที่พบตัวอย่างการสะกดผิดจำนวนไม่น้อย นั่นก็คือ กลุ่มคำอักษรนำ เช่นคำว่า “ขเยก” “สแตมป์” “สแลง” “ถไล” เป็นต้น เหล่านี้ล้วนเป็นคำที่สะกดผิด รูปคำที่สะกดถูกของคำเหล่านี้คือ “เขยก” “แสตมป์”

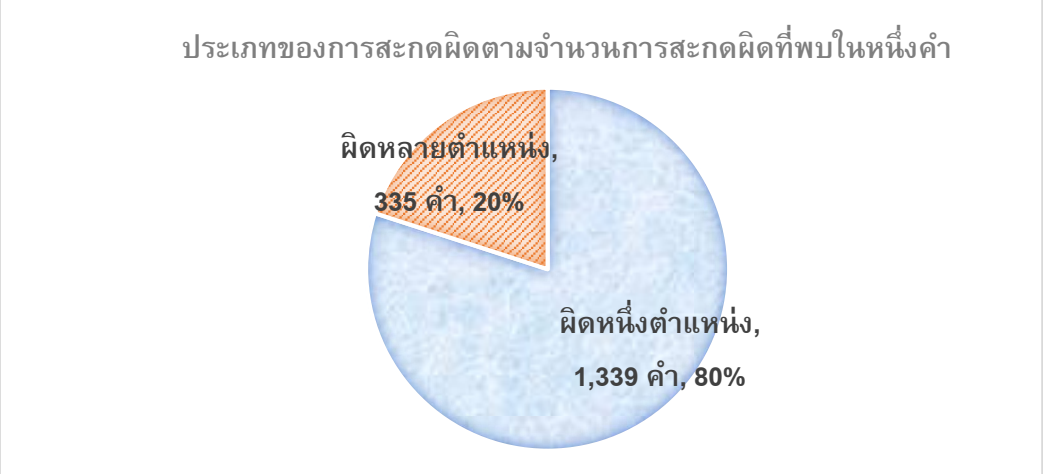
“แสลง” “ไถล” ตามลำดับ ซึ่งผู้วิจัยคิดว่าปัจจัยที่ทำให้เกิดการสะกดคำเหล่านี้ผิดนั้นอาจมีสาเหตุมาจากรูปแบบการสะกดคำของอักษรนำมีทั้งแบบที่ให้พยัญชนะต้นตัวแรกนำหน้าสระหน้า เช่น “ทแยง” “อเนก” “พยักพเยิด” เป็นต้น และแบบที่ให้พยัญชนะต้นตัวแรกตามหลังสระหน้า เช่น “โพยม” “ไศล” “แสยง” เป็นต้น โดยไม่ว่าสะกดแบบใดคำเหล่านี้ก็ยังอ่านออกเสียงเหมือนเดิม

4.2.2.2 คำที่สะกดผิดหลายตำแหน่งและออกเสียงเปลี่ยนไป

จากการศึกษาพบว่ามีคำที่สะกดในลักษณะนี้จำนวน 96 คำจาก 335 คำ หรือร้อยละ 28.66 ของคำที่สะกดผิดหลายลักษณะ และ 5.72% ของคำผิดที่ศึกษาทั้งหมด 1,674 คำ และผู้วิจัยคาดว่าคำที่สะกดผิดในกลุ่มนี้เป็นการสะกดผิดที่มีสาเหตุมาจากการออกเสียงคำหรือการรับรู้คำที่ผิดเพี้ยนไป ทำให้รูปเขียนของคำเหล่านี้ผิดเพี้ยนตามไปด้วย ตัวอย่างเช่น คำว่า “กระบวนการ” สะกดผิดเป็น “ขบวนการ” “ฟั่น” สะกดผิดเป็น “ควั่น” “สะเพร่า” สะกดผิดเป็น “สับเพร่า” หรือ “ออกหาก” สะกดผิดเป็น “ออกห่าง” เป็นต้น

จากผลการศึกษาวิเคราะห์คำที่มักเขียนผิดที่เป็นการสะกดผิดแบบเป็นคำจริงจำนวน 1,674 คำ โดยเปรียบเทียบลักษณะของคำที่สะกดผิดแต่ละคำว่าเปลี่ยนแปลงไปไปจากรูปการสะกดที่ถูกต้องอย่างไรบ้าง สรุปได้ว่าคำสะกดผิดแบบเป็นคำจริงที่นำมาศึกษาวิเคราะห์ในครั้งนี้สามารถจำแนกตามจำนวนของการสะกดผิดที่พบในคำที่สะกดผิดแต่ละคำได้เป็น 2 กลุ่มใหญ่ คือ กลุ่มของคำที่สะกดผิดหนึ่งลักษณะ 1,339 คำ และกลุ่มของคำที่สะกดผิดหลายลักษณะ 335 คำ หรือร้อยละ 80 และร้อยละ 20 ตามลำดับ ดังที่แสดงในรูปภาพที่ 4.1

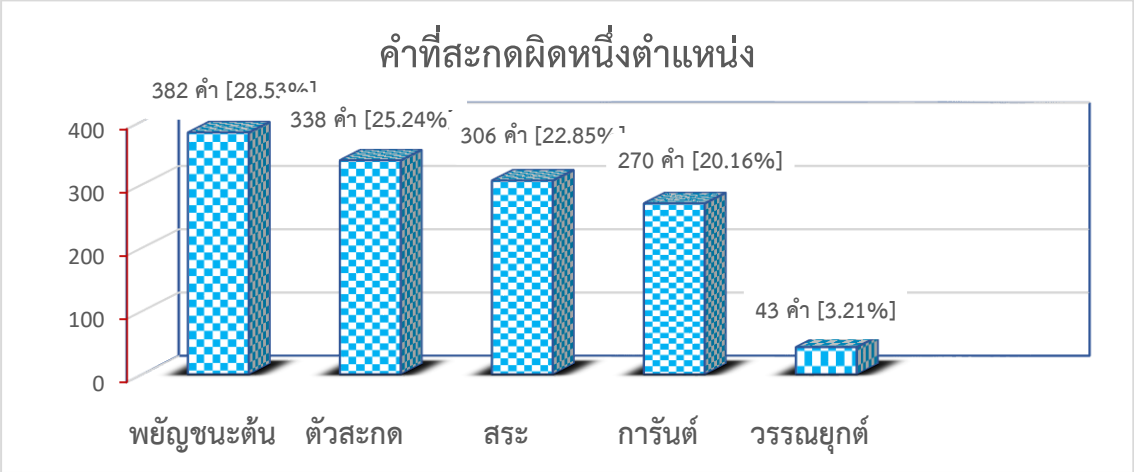
จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



รูปภาพที่ 4.1 ประเภทของการสะกดผิดจำแนกตามจำนวนการสะกดผิดที่พบในหนึ่งคำ

เมื่อพิจารณากลุ่มคำที่สะกดผิดหนึ่งตำแหน่งจำนวน 1,339 คำ พบว่าสามารถจำแนกออกเป็นกลุ่มย่อยได้ทั้งหมด 5 กลุ่มตามองค์ประกอบของโครงสร้างคำภาษาไทย (ดังรูปภาพที่ 4.2) ได้แก่

1. คำที่สะกดผิดที่พยัญชนะต้น จำนวน 382 คำ คิดเป็นร้อยละ 28.53
2. คำที่สะกดผิดที่ตัวสะกด จำนวน 338 คำ คิดเป็นร้อยละ 25.24
3. คำสะกดผิดที่สระ จำนวน 306 คำ คิดเป็นร้อยละ 22.85
4. คำที่สะกดผิดที่ตัวการันต์ จำนวน 270 คำ คิดเป็นร้อยละ 20.16%
5. คำที่สะกดผิดที่วรรณยุกต์ จำนวน 43 คำ คิดเป็นร้อยละ 3.21

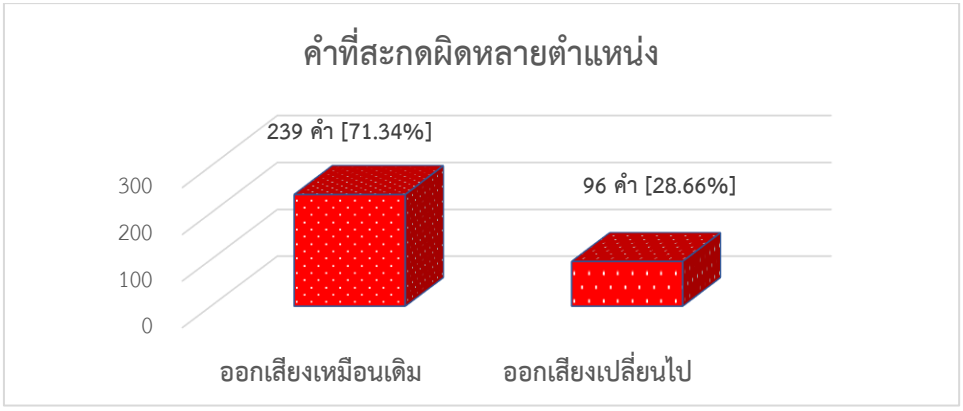


รูปภาพที่ 4.2 แผนภูมิแท่งแสดงคำที่สะกดผิดหนึ่งตำแหน่งประเภทต่างๆ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

นอกจากนี้ยังพบว่าสามารถจำแนกคำที่สะกดผิดหลายตำแหน่งจำนวน 335 คำ ออกเป็น 2 กลุ่ม (ดังรูปภาพที่ 4.3) ได้แก่

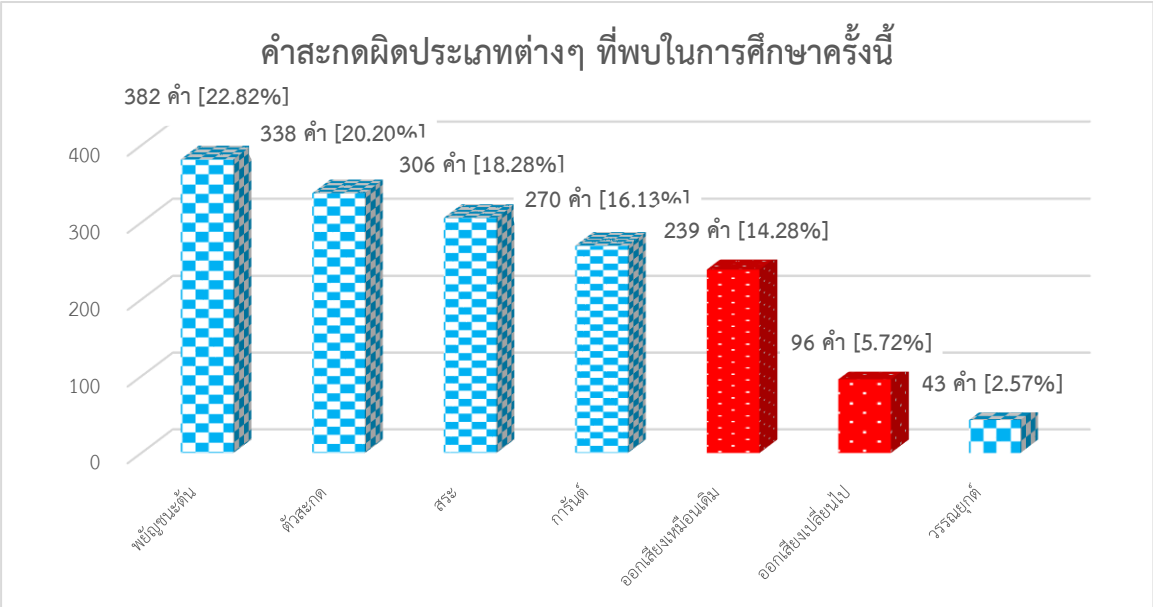
1. คำที่สะกดผิดหลายตำแหน่งแต่ออกเสียงเหมือนเดิม จำนวน 239 คำ คิดเป็นร้อยละ 71.34
2. คำที่สะกดผิดหลายตำแหน่งและออกเสียงเปลี่ยนไป จำนวน 96 คำ คิดเป็นร้อยละ 28.66



รูปภาพที่ 4.3 แผนภูมิแท่งแสดงคำที่สะกดผิดหลายตำแหน่งสองประเภท

เมื่อนำคำที่สะกดผิดแบบเป็นคำจริงในรูปแบบต่างๆ ทั้งประเภทสะกดผิดหนึ่งตำแหน่งและหลายตำแหน่งมาจัดเรียงรวมกันตามจำนวนตัวอย่างที่ปรากฏ พบว่าการสะกดผิดแบบเป็นคำจริงในภาษาไทยนั้นมักมีความผิดพลาดในการสะกดเกิดขึ้นที่พยัญชนะต้น รองลงมาเป็นที่ตัวสะกด และตำแหน่งที่พบตัวอย่างการสะกดผิดน้อยที่สุดคือที่วรรณยุกต์ ดังที่แสดงในรูปภาพที่ 4.4 แสดงให้เห็นว่าผลการวิเคราะห์ที่ได้นั้นไม่เป็นไปตามสมมติฐานข้อที่หนึ่งของงานวิจัยนี้เสียทั้งหมด ซึ่งคาดว่าตำแหน่งที่จะพบการสะกดผิดแบบเป็นคำจริงในภาษาไทยมากที่สุดนั้นเป็นที่ตัวสะกดและยังออกเสียงเหมือนเดิม คือส่วนที่ระบุว่าการสะผิดที่ตัวสะกดส่วนใหญ่จะยังออกเสียงเหมือนเดิมนั้นเป็นจริง เพราะตัวอย่างการสะกดผิดที่ตัวสะกดกว่าร้อยละ 90 นั้นล้วนเป็นการใช้ตัวสะกดในมาตราเดียวผิดตัวทำให้สะกดผิด

จุฬาลงกรณ์มหาวิทยาลัย



รูปภาพที่ 4.4 แผนภูมิแท่งแสดงคำที่สะกดผิดประเภทต่างๆ ที่พบในการวิเคราะห์ครั้งนี้

ผลการวิเคราะห์ที่ได้จากการศึกษาในครั้งนี้ ผู้วิจัยพบประเด็นที่น่าสนใจ คือ ข้อผิดส่วนใหญ่เป็นการสะกดผิดหนึ่งตำแหน่ง คือ ใช้อักษรผิดหนึ่งตัว หรือ ตกอักษรหนึ่งตัว หรือ ใส่อักษรเกินมาหนึ่งตัว เช่น “ครอก” สะกดผิดเป็น “คลอก” หรือ “หวน” สะกดผิดเป็น “หวล” ซึ่งคำแรกใช้ ล แทน ร และคำที่สอง ใช้ ล แทน น เป็นต้น ซึ่งการศึกษาครั้งนี้พบว่ามีคำที่สะกดผิดหนึ่งตำแหน่งถึง 80% และมีเพียง 20% ที่เป็นคำที่สะกดผิดหลายตำแหน่ง เช่น “จันทน์กพ้อ” สะกดผิดเป็น “จันทร์กะพ้อ” ซึ่งมีการสะกดผิดสองแห่ง คือ ใช้ ร แทน น และเติมสระอะเกินมา หรือ สูญ เขียนผิดเป็นศูนย์ มีการเขียนผิดรวมสี่แห่ง คือ จุดแรกจาก ส เป็น ศ จุดที่สองจาก ญ เป็น น จุดที่สามคือเติม ย และ จุดที่สี่คือเติมตัวการันต์ สัดส่วนของการสะกดผิดทั้งสองประเภทนี้แสดงให้เห็นว่าคำที่สะกดผิดส่วนใหญ่นั้นเกิดจากการเปลี่ยนแปลงเล็กน้อยเพียงครั้งเดียว ซึ่งตรงกับแนวคิดในการใช้โมดูล edit distance (Neil Bowers, 2015) เทียบหาคำที่มีความต่างของตัวอักษรเพียง 1 ตัวอักษรเพื่อแก้ไขคำที่สะกดผิดหรือสร้างคำที่น่าจะเป็นไปได้ในกรณีที่มีคำมากกว่าหนึ่งคำที่สามารถนำมาแก้ไขการสะกดผิด นอกจากนี้การกำกับข้อมูลรูปแบบของการสะกดผิดในการศึกษาครั้งนี้ทำให้ผู้วิจัยสามารถนำข้อมูลมาสร้างเป็นคลังข้อมูลรูปแบบการปรับแก้ตัวอักษรว่าตัวอักษรแต่ละตัวมักจะถูกลบออกไป ถูกเติมเข้าไป หรือถูกแทนที่ด้วยตัวอักษรใดในคำสะกดผิดเพื่อสร้างคำที่น่าจะเป็นสำหรับใช้แก้ไขการสะกดผิดที่ตรวจพบได้ซึ่งเป็นการเพิ่มประสิทธิภาพและความรวดเร็วในการตรวจจับและแก้ไขคำที่สะกดผิดให้ปฏิบัติการได้อย่างถูกต้องแม่นยำและรวดเร็วมากยิ่งขึ้น ส่วนเนื้อหาในบทต่อไปจะอธิบายถึงโครงสร้าง หลักการทำงาน และผลการทำงานของระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมซึ่งเป็นวิธีที่นำเสนอในงานวิจัยนี้ รวมถึงเปรียบเทียบผลการทำงานของวิธีที่นำเสนอกับวิธีพื้นฐานสองวิธีคือตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมและด้วยชุดคำสับสน

Chulalongkorn University

บทที่ 5

การตรวจแก้การสะกดผิดแบบเป็นคำจริง

ในบทนี้จะกล่าวถึงข้อมูลที่ใช้ทดสอบและระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมซึ่งเป็นวิธีการที่นำเสนอในงานวิจัยครั้งนี้ และอีกสองระบบพื้นฐานที่พัฒนาขึ้นเพื่อนำมาใช้เปรียบเทียบกับประสิทธิภาพของระบบที่นำเสนอ ซึ่งได้แก่ การตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองยูนิแกรมและด้วยการใช้ชุดคำสับสนในการตรวจแก้การสะกดผิด ในแต่ละวิธีการผู้วิจัยจะอธิบายถึงหลักการทำงานซึ่งประกอบด้วยสองกระบวนการหลักคือกระบวนการตรวจจับและกระบวนการแก้ไขคำที่สะกดผิด รวมถึงผลการทดสอบระบบ และในส่วนสุดท้ายของบทนี้จะกล่าวถึงประสิทธิภาพในการทำงานของระบบที่พัฒนาขึ้นเปรียบเทียบกับระบบพื้นฐานทั้งสอง

5.1 ข้อมูลที่ใช้ทดสอบ (test data)

ข้อมูลที่นำมาใช้ทดสอบเป็นข้อความภาษาไทยจำนวน 1,000 ข้อความ ที่ได้จากการสุ่มเลือกคำที่สะกดผิดแบบเป็นคำจริงจากคลังชุดคำสับสนมา 375 คำ แล้วนำไปสืบค้นบนอินเตอร์เน็ตเพื่อให้ได้ข้อความตัวอย่างของการสะกดผิดที่เกิดขึ้นจริงจำนวน 1,000 ข้อความที่มีจำนวนคำทั้งหมดประมาณ 33,000 คำ แล้วนำข้อความที่จะใช้ทดสอบทั้งหมดนี้ไปผ่านกระบวนการตัดคำ โดยผู้วิจัยอนุมานว่าทุกคำที่ผ่านการตัดคำได้สำเร็จเป็นคำจริง และภายในข้อความแต่ละข้อความจะมีคำที่สะกดผิดแบบเป็นคำจริงปนอยู่อย่างน้อยหนึ่งคำ ซึ่งเป็นคำสะกดผิดที่ต้องการให้ระบบตรวจจับและแก้ไขให้ถูกต้อง

5.2 การตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรม

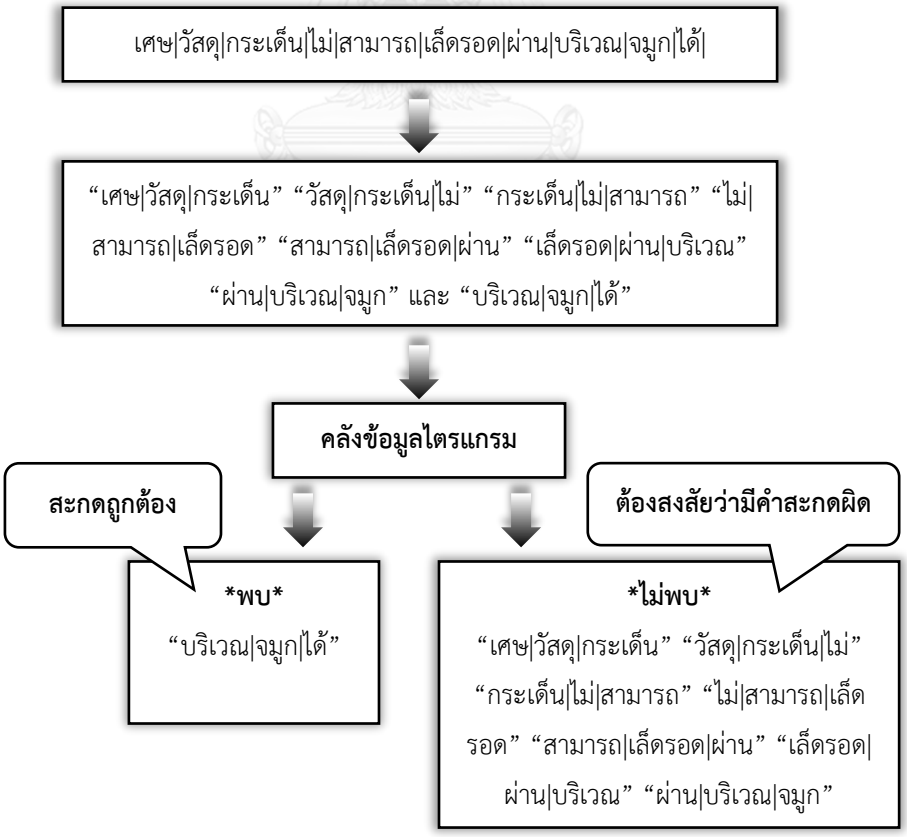
ผู้วิจัยพัฒนาระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมนี้ขึ้นด้วยเป้าหมายที่จะสร้างเครื่องมือที่สามารถแก้ไขปัญหาการสะกดผิดแบบเป็นคำจริง เหตุผลที่ผู้วิจัยเลือกใช้วิธีการนี้ก็คือวิธีการนี้เป็นวิธีการเชิงสถิติที่พิจารณาความถูกต้องของการสะกดจากความถี่และความน่าจะเป็นในการปรากฏของคำเรียงต่อกันสามคำหรือไตรแกรมคำในคลังข้อมูลภาษาไทยแห่งชาติ (Aroonmanakul, 2007) ซึ่งเป็นวิธีการที่ผู้วิจัยคาดว่าจะสามารถจัดการกับปัญหาการสะกดผิดประเภทนี้ได้

5.2.1 หลักการทำงานของระบบตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรม

ระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมที่ผู้วิจัยพัฒนาขึ้นนั้น มีขั้นตอนในการทำงานดังนี้

5.2.1.1 ขั้นตอนที่หนึ่ง: ตรวจจับคำที่ต้องสงสัยในข้อความ

ในขั้นแรกเมื่อป้อนข้อความทดสอบให้ระบบทำการตรวจแก้คำที่สะกดผิด ระบบจะอ่านข้อความในข้อมูลทดสอบที่มีคำสะกดผิดปนอยู่ทีละหนึ่งข้อความ จากนั้นระบบจะแบ่งข้อความที่กำลังอ่านออกเป็นสายคำเรียงต่อกันสามคำ หลังจากที่แบ่งข้อความออกเป็นสายคำเรียงสามเรียบร้อยแล้ว สายคำเรียงสามคำเหล่านี้จะถูกนำไปตรวจสอบว่าปรากฏอยู่ในคลังข้อมูลไตรแกรมหรือไม่ ทีละสายจนครบทุกสายของข้อความหนึ่งข้อความ ซึ่งถ้าหากว่าสายคำเรียงสามคำทุกสายปรากฏในคลังข้อมูลไตรแกรมคำแสดงว่าระบบตรวจไม่พบคำที่สะกดผิดในข้อความนั้น แต่ในทางกลับกัน ถ้าหากว่าสายคำเรียงสามคำสายใดไม่ปรากฏในคลังข้อมูลไตรแกรมแสดงว่าระบบตรวจพบว่าสายคำเรียงนั้นมีคำที่ต้องสงสัยว่าเป็นคำสะกดผิดปนอยู่ ดังตัวอย่างที่แสดงในรูปภาพที่ 5.1 จากนั้นสายคำเรียงสามที่มีคำที่สะกดผิดปนอยู่จะถูกส่งไปปรับแก้ในขั้นตอนต่อไป



รูปภาพที่ 5.1 แสดงขั้นตอนในการตรวจจับคำที่ต้องสงสัยในข้อความ

5.2.1.2 ขั้นตอนที่สอง: ปรับแก้คำที่ต้องสงสัย

หลังจากที่ตรวจพบว่ามีสายคำเรียงสามคำที่ไม่ปรากฏในคลังข้อมูลไตรแกรม สายคำเหล่านั้นจะถูกส่งมาทำการปรับแก้ ด้วยวิธีที่ใช้หลักการเดียวกันกับการปรับแก้น้อยที่สุดชื่อ “Levenshtein Distance” หรือ “minimum edit distance” (Bowers, 2015) ซึ่งเป็นการปรับแก้สายอักขระสายหนึ่งให้เหมือนกับอีกสายหนึ่งด้วยการปรับแก้น้อยที่สุด ซึ่งจากผลการวิเคราะห์ในบทที่แล้วพบว่าคำที่สะกดผิดส่วนใหญ่มักจะสะกดผิดพลาดไปด้วยการเติม ลบ หรือแทนที่ตัวอักษรตัวหนึ่งในคำเพียงแค่ตัวอักษรเดียว ดังนั้นวิธีการแก้ไขคำที่สะกดผิดส่วนใหญ่ก็จะสามารถทำได้โดยการปรับแก้เพียงครั้งเดียวเช่นกัน ซึ่งรูปแบบการสะกดผิดที่ได้จากการวิเคราะห์คำไทยที่มักเขียนผิดนั้นทำให้ทราบว่าตัวอักษรใดมักจะถูกเติมเกินเข้าไป ตกหล่น หรือถูกแทนที่ด้วยตัวอักษรใดจึงเกิดเป็นคำที่สะกดผิด ในทางกลับกันเมื่อมองย้อนกลับว่าจะปรับแก้คำที่สะกดผิดแต่ละคำให้เป็นคำที่สะกดถูกต้องอย่างไรก็จะสามารถหารูปแบบเพื่อแก้ไขการสะกดผิดของคำๆ นั้นได้ ดังตัวอย่างในตารางที่ 5.1

ตารางที่ 5.1 แสดงตัวอย่างรูปแบบที่นำไปใช้ปรับแก้คำที่ต้องสงสัยว่าสะกดผิด

คำที่สะกดถูกต้อง	คำที่สะกดผิด	รูปแบบการสะกดผิด	รูปแบบที่นำไปใช้ปรับแก้การสะกดผิด
|ร่ำ|เลือ|	|ล่ำ|เลือ|	ร_ล (ใช้ ล แทน ร)	ล_ร (แทนที่ ล ด้วย ร)
|กะพริบ|	|กระ|พริบ|	0_ร (เติม ร)	ร_0 (ลบ ร ออก)
|หวน|	|หวล|	ล_น (ใช้ ล แทน น)	น_ล (แทนที่ ล ด้วย น)
|ลำไย|	|ลำ|ใย|	ไ_ใ (ใช้ ใ แทน ไ)	ใ_ไ (แทนที่ ไ ด้วย ใ)
|สะกิด|	|สกิด|	ะ_0 (ตกรูปสระ -ะ)	0_ะ (เติมรูปสระ -ะ)
|มนุษยสัมพันธ์|	|มนุษย์|สัมพันธ์|	0_์ (เติม การันต์)	์_0 (ลบการันต์ออก)
|โน้ต|	|โน๊ต|	้_๊ (ใช้ ๊ แทน ้)	๊_้ (แทนที่ ้ ด้วย ๊)

ยกอย่างเช่นคำว่า “กะพริบ” สะกดผิดเป็น “ก**ร**ะพริบ” ตามวิธีการกำกับข้อมูลที่ผู้วิจัยได้กล่าวไว้ในหัวข้อ 4.1 รูปแบบที่กำกับการสะกดผิดในลักษณะนี้คือ “0_ร” ซึ่งหมายถึง ตัวอักษร ร ถูกเติมเข้าไป ดังนั้นการแก้ไขคำว่า “กระพริบ” ให้ถูกต้องก็คือลบตัวอักษร ร ออก และรูปแบบในการแก้ไขการสะกดผิดนี้คือ “ร_0” ผู้วิจัยได้เก็บรวมรวบรูปแบบการในแก้ไขการสะกดผิดที่ได้จากการวิเคราะห์การปรับแก้การสะกดผิดทั้งหมดนี้เอาไว้เป็นคลังข้อมูลอีกคลังหนึ่งเพื่อให้ง่ายต่อการนำไปปรับใช้ อีกทั้งการนำรูปแบบการแก้ไขการสะกดผิดเหล่านี้มาใช้ในขั้นตอนการปรับแก้คำที่ต้องสงสัยให้ถูกต้องนั้นจะช่วยย่นระยะเวลาในการหาคำที่เป็นไปได้ให้ดำเนินการได้รวดเร็วและแม่นยำยิ่งขึ้น ไม่สิ้นเปลืองเวลาในการนำทุกตัวอักษรที่มีในภาษาไทยมาทดลองแทนที่ตัวอักษรแต่ละตัวในคำที่ต้องสงสัย เพราะโดยส่วนใหญ่แล้วตัวอักษรแต่ละตัวนั้นมีความเป็นไปได้ที่จะถูกใช้สลับกับตัวอักษรเพียงบางตัวเท่านั้น เช่น

ตัวอักษร ด มักจะถูกใช้สลับกับตัวอักษรในมาตราตัวสะกดแม่กด เช่น ช (บวด กับ บวช) ต (ชาด กับ ชาต) ท (บาด กับ บาท) เป็นต้น แต่จะไม่ถูกใช้สลับกับ ม ย ว หรือ ง เป็นต้น ด้วยเหตุนี้ผู้วิจัยจึงได้นำเอารูปแบบการสะกดผิดที่รวบรวมได้จากการวิเคราะห์คำไทยที่มักสะกดผิดมาปรับใช้ในขั้นตอนปรับแก้คำที่ต้องสงสัย ซึ่งมีประกอบด้วยขั้นตอนย่อยอีก 3 ขั้นตอนดังต่อไปนี้

ขั้นตอนแรก สายคำเรียงสามต้องสงสัยที่ไม่ปรากฏในคลังข้อมูลไตรแกรมแต่ละสายจะถูกนำไปดัดแปลงด้วยการแทรกตัวเลข “0” ระหว่างทุกตัวอักษรในสายคำเรียงนั้น รวมถึงด้านหน้าและด้านหลังสายคำนั้นด้วย ตัวอย่างเช่น สมมติว่า “จะ|แก่|ไข” เป็นสายคำเรียงสามที่ไม่ปรากฏในคลังข้อมูลไตรแกรม เมื่อแทรก “0” ที่ด้านหน้า ระหว่างตัวอักษรแต่ละตัว และด้านหลังสายคำเรียงสามที่ต้องสงสัยข้างต้น จะได้เป็น 0 จ 0 ะ 0 แ 0 ก 0 ่ 0 ไ 0 ข 0 โดยเครื่องหมายแบ่งคำจะถูกลบออกก่อนที่จะแทรกตัวเลข 0 เข้าไป ซึ่งเหตุผลของการแทรกตัวเลขศูนย์เข้าไปนั้นก็เพื่อให้สอดคล้องกับรูปแบบการปรับแก้การสะกดผิดที่นำมาใช้ คือ 0_x หมายถึงการเติมตัวอักษร x กล่าวคือ ตัวเลข 0 ที่แทรกเข้าไปเป็นเสมือนการแทรกพื้นที่เอาไว้ให้สำหรับเติมตัวอักษรเพิ่มเข้าไปในคำแล้วสามารถแก้ไขให้คำๆ นั้นถูกต้องได้

ขั้นตอนที่สอง เมื่อแทรกตัวเลข 0 เข้าไปในสายคำเรียงสามที่มีคำต้องสงสัยว่าสะกดผิดเรียบร้อยแล้ว สายคำเรียงสามนั้นจะได้รับการปรับแก้เพื่อหาคำถูกต้องที่เป็นไปได้ ซึ่งจากการวิเคราะห์คำไทยที่มักเขียนผิดในบทที่ 4 พบว่าคำที่สะกดผิดส่วนใหญ่นั้นเกิดจากความผิดพลาดในพิมพ์ตัวอักษรเกินมาหนึ่งตัว พิมพ์ตัวอักษรตกไปหนึ่งตัว หรือพิมพ์ตัวอักษรผิดไปหนึ่งตัว จากผลการวิเคราะห์ที่พบนี้ ผู้วิจัยจึงอนุมานว่าภายในสายคำเรียงสามที่ต้องสงสัยแต่ละสายนั้นมีตัวอักษรหนึ่งตัวที่เกินมา ขาดหายไป หรือเปลี่ยนไป เป็นสาเหตุให้เกิดการสะกดผิด ซึ่งถ้าหากสามารถปรับแก้ตัวอักษรนั้นให้ถูกต้องได้ก็เท่ากับสามารถแก้ไขการสะกดผิดให้ถูกต้องได้สำเร็จ ซึ่งในขั้นตอนย่อยนี้ตัวอักษรแต่ละตัวในสายคำเรียงที่ต้องสงสัยจะถูกปรับแก้ทีละตัวโดยเริ่มจากอักษรตัวแรกจากทางด้านซ้าย ผู้วิจัยได้นำรูปแบบการปรับแก้การสะกดผิดที่กล่าวไปในตอนต้นมาใช้ในขั้นตอนนี้ คือ นำตัวอักษรแต่ละตัวในสายคำเรียงสามไปค้นหากลุ่มตัวอักษรที่น่าจะเป็นและนำตัวอักษรในกลุ่มที่น่าจะเป็น มาแทนที่ตัวอักษรเดิมในสายคำเรียงสามเพื่อเพิ่มความรวดเร็วและแม่นยำในการปรับแก้การสะกดผิด ซึ่งรูปแบบการปรับแก้การสะกดผิดที่รวบรวมไว้จะช่วยระบุว่าควรจะเติมตัวอักษรตัวใดเพิ่มเข้าไปในคำที่ต้องสงสัย ตัวอักษรตัวใดควรจะถูกลบออกไปหรือถูกแทนที่ด้วยตัวอักษรตัวใดเพื่อให้ได้เป็นคำที่สะกดถูกต้อง

ขั้นตอนที่สาม หลังจากที่ตัวอักษรหนึ่งตัวในสายคำเรียงสามที่ต้องสงสัยได้รับการปรับแก้ในแต่ละครั้ง ตัวเลขศูนย์ในชุดคำเรียงนั้นจะถูกลบออกทั้งหมด เมื่อเหลือแต่ตัวอักษรแล้วสายคำเรียงสามที่ได้รับการปรับแก้จะถูกส่งไปตรวจสอบว่ามีปรากฏอยู่ในคลังข้อมูลไตรแกรมหรือไม่ ถ้าหากว่าไม่พบก็ให้นำไปตรวจสอบในคลังข้อมูลไบแกรมและยูนิแกรมตามลำดับ เนื่องจากในช่วงฝึกฝน

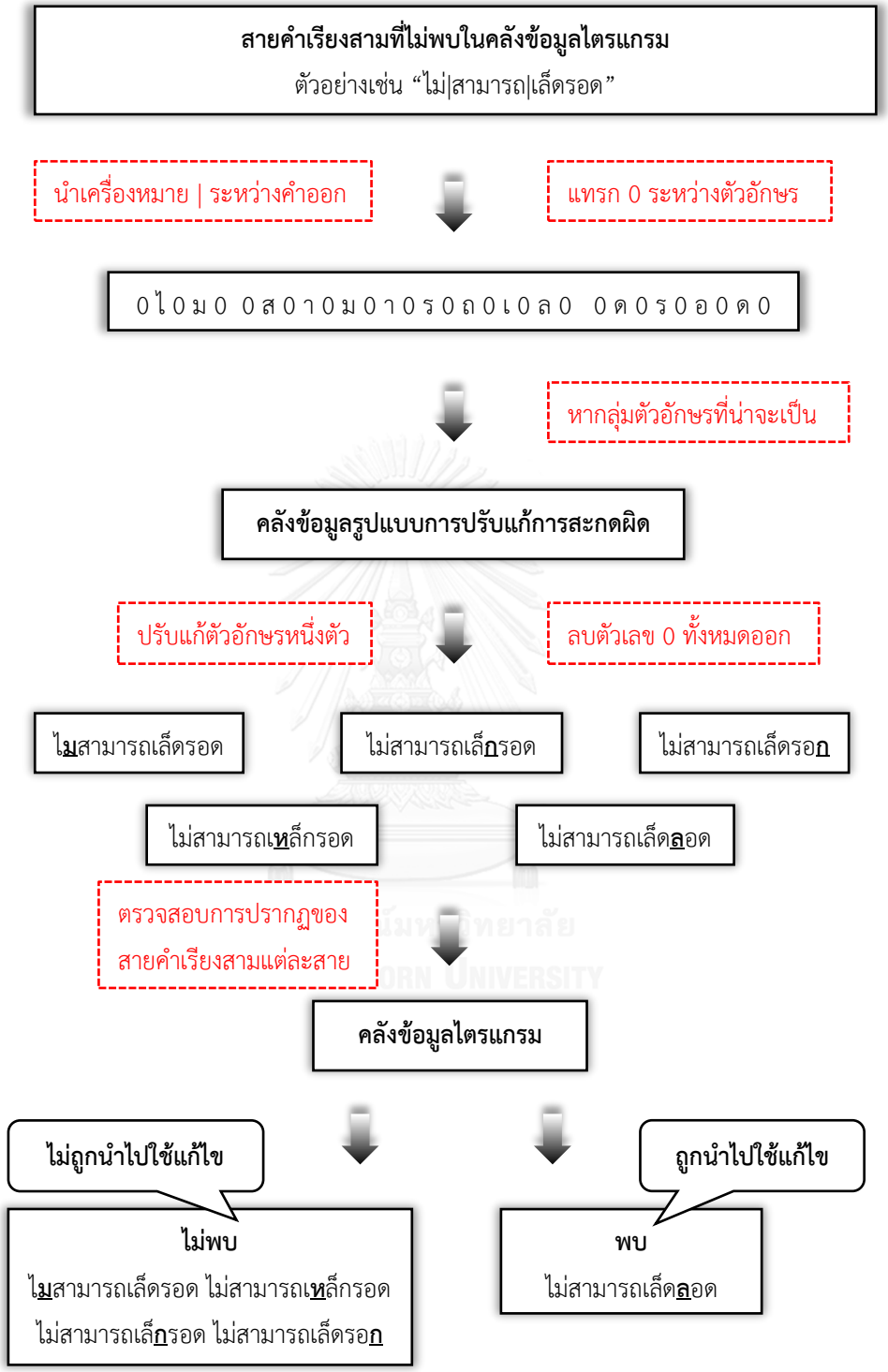
ระบบ ผู้วิจัยพบว่าสายคำเรียงสามที่ต้องสงสัยบางสายเมื่อได้รับการปรับแก้ให้สะกดถูกต้องแล้วจะมีจำนวนแกรมที่ลดลง ทำให้ไม่พบการปรากฏของสายคำเรียงที่ถูกต้องในคลังข้อมูลไตรแกรม แต่พบในคลังข้อมูลในคลังข้อมูลไบแกรมหรือคลังข้อมูลยูนิแกรม ดังตัวอย่างที่แสดงในตารางที่ 5.2

ตารางที่ 5.2 แสดงตัวอย่างคำสะกดผิดที่มีจำนวนแกรมเปลี่ยนไปเมื่อได้รับการปรับแก้

คำที่สะกดผิด	**จำนวนแกรม**	**ปรับแก้ด้วย**	**คำที่สะกดถูกต้อง**	**จำนวนแกรม**
|ธง|ไตร|รงศ์|	3	ศ_ค (แทน ศ ด้วย ค)	|ธง|ไตรรงค์|	2
|คลุม|เค|ลือ|	3	ล_ร (แทน ล ด้วย ร)	|คลุมเครือ|	1

แต่ถ้าหากว่าข้อความใดที่ระบบตรวจจับได้ว่ามีคำที่ต้องสงสัยว่าสะกดผิดปนอยู่ และถูกนำไปปรับแก้แล้ว แต่ไม่พบว่ามีสายคำเรียงสามสายใดในข้อความที่ได้รับการปรับแก้แล้วปรากฏอยู่ในคลังข้อมูลไตรแกรม ไบแกรม หรือยูนิแกรมเลย แสดงว่าสายคำเรียงสามที่ต้องสงสัยเหล่านั้นอาจจะไม่ได้สะกดผิดตามที่สงสัยจึงไม่ถูกนำไปใช้แก้ไขการสะกดผิด แต่ทันทีที่ระบบตรวจพบสายคำเรียงสามที่ได้รับการปรับแก้แล้วในคลังข้อมูล ระบบจะเก็บสายคำเรียงสามที่ตรวจพบการปรากฏในคลังข้อมูลเป็นสายแรกเอาไว้เพื่อที่จะนำไปใช้แก้ไขคำที่สะกดผิดจริงในขั้นตอนสุดท้าย ดังที่แสดงในรูปภาพที่ 5.2 เมื่อระบบตรวจสอบสายคำเรียงสามที่ปรับแก้แล้วสายหนึ่งเสร็จสิ้นแล้ว ระบบจะกลับไปอ่านสายคำเรียงสามสายต่อไปในข้อความเดียวกันและปรับแก้สายคำต้องสงสัยใหม่อีกซ้ำแบบนี้ไปจนครบทุกสายคำเรียงสามในข้อความที่ถูกป้อนเข้ามาหนึ่งข้อความ จากนั้นสายคำเรียงสามที่จะนำไปใช้แก้ไขการสะกดผิดทั้งหมดที่พบในหนึ่งข้อความจะถูกส่งต่อไปยังขั้นตอนสุดท้าย คือขั้นตอนการเลือกแก้ไขคำสะกดผิดจริงเพียงคำเดียวในหนึ่งข้อความ

Chulalongkorn University

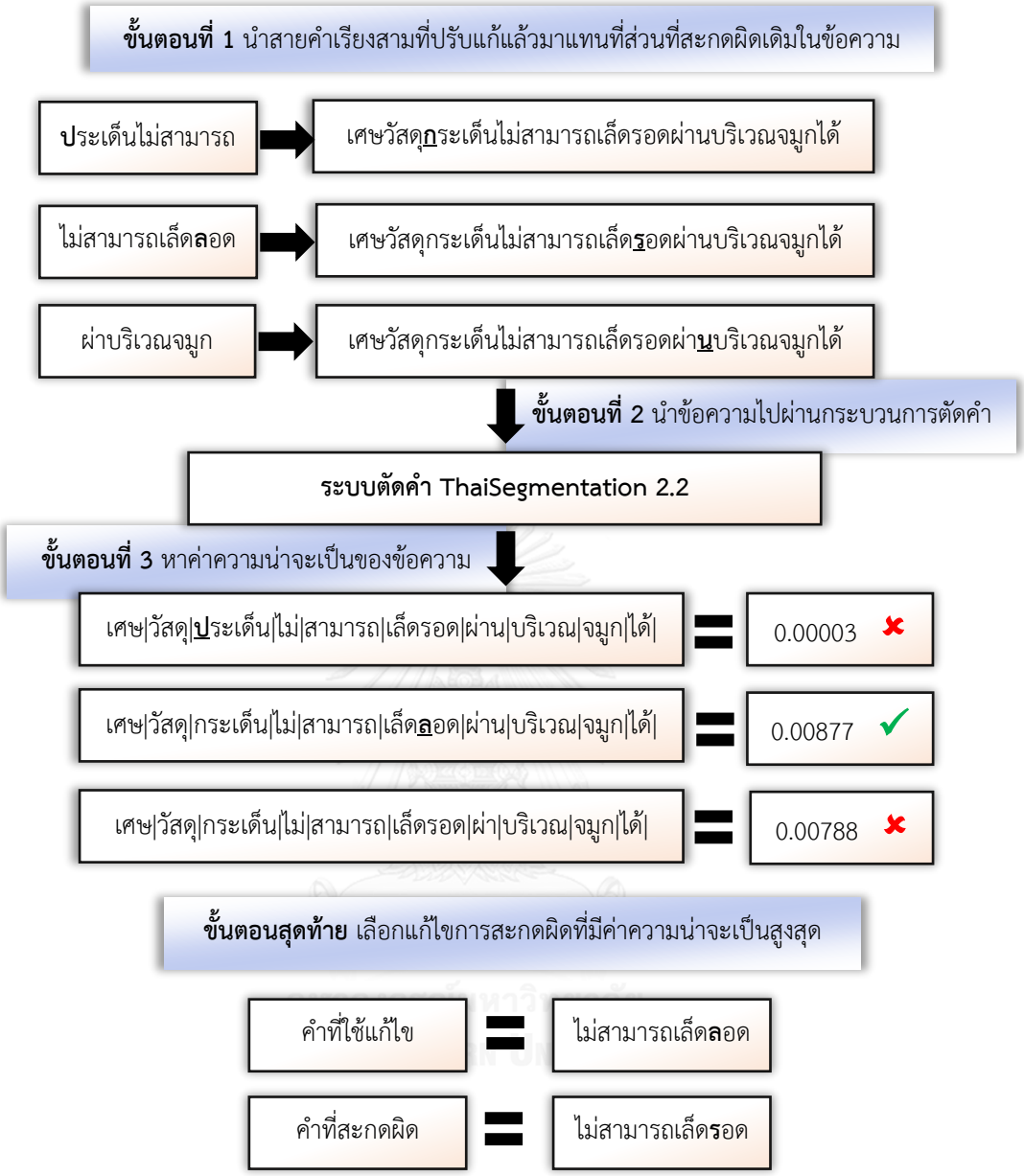


รูปภาพที่ 5.2 แสดงขั้นตอนการปรับแก้คำที่ต้องสงสัย

5.2.1.3 ขั้นตอนที่สาม: เลือกคำที่เหมาะสมเพื่อใช้แก้ไขคำที่สะกดผิด

ในขั้นตอนสุดท้ายนี้ระบบจะทำการเลือกแก้ไขคำที่สะกดผิดจริงเพียงคำเดียว ซึ่งมีขั้นตอนย่อยทั้งหมด 4 ขั้นตอนดังต่อไปนี้ **ขั้นตอนแรก** คือ นำสายคำเรียงสามที่ปรับแก้แล้วแต่ละสายมาแทนที่ส่วนที่ต้องสงสัยว่าสะกดผิดเดิมในข้อความ ซึ่งสายคำเรียงสามที่ปรับแก้แล้วแต่ละสายนั้นจะแตกต่างกันไป ไม่ซ้ำกัน และได้รับการแก้ไขเพียงครั้งเดียวทั้งหมด หลังจากนั้นใน **ขั้นตอนที่สอง** ข้อความที่ได้รับการแก้ไขแล้วทั้งหมดจะถูกนำไปผ่านกระบวนการตัดคำเพื่อระบุขอบเขตของคำแต่ละคำในข้อความ ต่อจากนั้นใน**ขั้นตอนที่สาม** ระบบจะคำนวณหาค่าความน่าจะเป็นของข้อความที่แก้ไขแล้วแต่ละข้อความ และใน**ขั้นตอนสุดท้าย** ระบบจะเลือกสายคำเรียงสามที่ให้ค่าความน่าจะเป็นของข้อความสูงสุดเพียงสายเดียวเป็นสายคำเรียงสามที่นำไปใช้แก้ไขการสะกดผิดในข้อความให้ถูกต้อง และสายคำเรียงสามต้องสงสัยที่นำมาปรับแก้เป็นสายคำเรียงที่ระบบเลือกนี้คือสายคำเรียงสามที่สะกดผิดจริง กล่าวคือ ระบบจะระบุว่าสายคำเรียงสามต้องสงสัยสายใดที่เป็นสายที่สะกดผิดหรือมีคำที่สะกดผิดอยู่จริงหลังจากที่ระบบสามารถทำการแก้ไขสายคำเรียงสามต้องสงสัยนั้นแล้วพบว่าให้ค่าความน่าจะเป็นของข้อความสูงสุด ดังที่ปรากฏในรูปภาพที่ 5.3

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



รูปภาพที่ 5.3 แสดงการเลือกคำที่เหมาะสมเพื่อแก้ไขการสะกดผิดในข้อความ

5.2.2 การประเมินประสิทธิภาพการตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรม

หลังจากที่ได้ทดสอบให้ระบบได้ทำการตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมโดยใช้ข้อมูลทดสอบที่จัดเตรียมไว้ซึ่งเป็นข้อความภาษาไทยจำนวน 1,000 ข้อความ ภายในข้อความแต่ละข้อความมีคำที่สะกดผิดแบบเป็นคำจริงปนอยู่ พบว่าระบบใช้เวลาในการประมวลผลข้อมูลทดสอบทั้งหมดประมาณ 128 วินาที ซึ่งระบบใช้เวลาในการอ่านคลังข้อมูลต่างๆ เพื่อเตรียมระบบให้พร้อมสำหรับการตรวจแก้การสะกดผิดประมาณ 55 วินาที และใช้เวลาประมาณ 73 วินาทีในการตรวจแก้ข้อความภาษาไทย 1,000 ข้อความ หรือประมาณ 33,000 คำ

ในส่วนของการประเมินประสิทธิภาพการตรวจแก้การสะกดผิดแบบเป็นคำจริงของระบบนั้น ผู้วิจัยประเมินจากค่าประสิทธิภาพ (F-measure) หรือ (F_1) ค่าความครบถ้วน (Recall) หรือ (R) และค่าความแม่นยำ (Precision) หรือ (P) จากสูตรด้านล่าง

สูตรสำหรับคำนวณค่าความครบถ้วนและค่าความแม่นยำในการตรวจจับคำที่สะกดผิด

$$Recall\ (\mathrm{R}) = \frac{\text{จำนวนคำสะกดผิดที่ตรวจจับได้ถูกต้อง}}{\text{จำนวนคำสะกดผิดทั้งหมดในชุดข้อมูลทดสอบหนึ่งชุด}}$$

$$Precision\ (P) = \frac{\text{จำนวนคำสะกดผิดที่ตรวจจับได้ถูกต้อง}}{\text{จำนวนคำสะกดผิดทั้งหมดที่ตรวจจับได้}}$$

สูตรสำหรับคำนวณค่าความครบถ้วนและค่าความแม่นยำในการแก้ไขคำที่สะกดผิด

$$Recall\ (\mathrm{R}) = \frac{\text{จำนวนคำสะกดผิดที่แก้ไขได้ถูกต้อง}}{\text{จำนวนคำสะกดผิดที่ตรวจจับได้ถูกต้องทั้งหมดในข้อมูลทดสอบหนึ่งชุด}}$$

$$Precision\ (P) = \frac{\text{จำนวนคำสะกดผิดที่แก้ไขได้ถูกต้อง}}{\text{จำนวนคำสะกดผิดทั้งหมดที่ได้แก้ไข}}$$

สูตรสำหรับคำนวณค่าประสิทธิภาพในการตรวจจับและแก้ไขคำที่สะกดผิด

$$F_1 = \frac{2PR}{P+R}$$

ซึ่งค่าประสิทธิภาพ ค่าความครบถ้วน และค่าความแม่นยำ นี้จะอยู่ระหว่าง 0 ถึง 1 ($0 < F_1/R/P < 1$) หากยิ่งมีค่าเข้าใกล้ 1 นั่นหมายถึงประสิทธิภาพของระบบยิ่งสูงตามไปด้วย

5.2.2.1 ประสิทธิภาพในการตรวจจับการสะกดผิดด้วยแบบจำลองไตรแกรม

จากข้อความที่ใช้ทดสอบทั้งหมดจำนวน 1,000 ข้อความ ซึ่งในแต่ละข้อความมีคำที่สะกดผิดแบบเป็นคำจริงที่ผู้วิจัยต้องการให้ระบบตรวจแก้ได้อยู่หนึ่งคำและคำสะกดผิดเหล่านั้นก็คือคำที่ผู้วิจัยสุ่มเลือกมาจากชุดคำสับสนแล้วนำไปใช้สืบค้นหาข้อความตัวอย่างมาใช้ในการทดสอบนั่นเอง ระบบสามารถตรวจจับคำสะกดผิดแบบเป็นคำจริงที่ผู้วิจัยต้องการให้ระบบตรวจจับได้ในข้อความทดสอบได้อย่างถูกต้อง 465 ข้อความ และมีข้อความจำนวน 3 ข้อความที่ระบบตรวจไม่พบการสะกดผิดใดๆ ภายในข้อความ เท่ากับว่าระบบตรวจจับคำสะกดผิดผิดพลาดไปจำนวน 532 ข้อความ เมื่อนำข้อมูลตัวเลขที่ได้จากการทดสอบไปคำนวณหาค่าความครบถ้วน (R) ในการตรวจจับการสะกดผิดได้เท่ากับ 0.465 (465/1,000) หมายถึงระบบตรวจจับการสะกดผิดด้วยแบบจำลองไตรแกรมนั้นสามารถตรวจจับการสะกดผิดแบบเป็นคำจริงได้ประมาณเกือบครึ่งหนึ่งของคำสะกดผิดที่ต้องการให้ระบบตรวจพบทั้งหมด และได้ค่าความแม่นยำ (P) เท่ากับ 0.466 (465/997) ซึ่งบอกให้ทราบว่าระบบสามารถตรวจจับคำสะกดผิดที่ผู้วิจัยต้องการให้ระบบตรวจจับได้ถูกต้องประมาณครึ่งหนึ่งของคำสะกดผิดที่ระบบตรวจจับได้ทั้งหมด และมีค่า F-1 เท่ากับ 0.466 ดังที่แสดงในตารางที่ 5.3 ผู้วิจัยคาดว่าวิธีการตรวจจับคำต้องสงสัยว่าสะกดผิดที่อาศัยคลังข้อมูลไตรแกรมในการระบุว่าคำที่ไม่ปรากฏในคลังข้อมูลไตรแกรมเพียงอย่างเดียวโดยไม่ได้ผ่านขั้นตอนการคัดกรองหรือคัดเลือกเอาเฉพาะคำที่ต้องสงสัยว่าสะกดผิดจริงก่อน เป็นปัจจัยหลักที่ทำให้ระบบตรวจจับการสะกดผิดอีกครึ่งหนึ่งไม่สำเร็จและไม่ถูกต้อง

ตารางที่ 5.3 แสดงค่าประสิทธิภาพการตรวจจับการสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรม

	Detection	Correction
Recall (R)	0.465	0.85
Precision (P)	0.466	0.85
F-measure (F-1)	0.466	0.85

ในส่วนของประสิทธิภาพในการแก้ไขการสะกดผิดนั้น จากจำนวนข้อความที่ระบบสามารถตรวจจับได้ถูกต้องทั้งหมดจำนวน 465 ข้อความ ระบบได้ทำการแก้ไขข้อความเหล่านั้นได้ถูกต้องจำนวน 394 ข้อความ และแก้ไขผิดพลาดไปทั้งหมด 71 ข้อความ จากตารางที่ 5.3 แสดงให้เห็นว่าเมื่อระบบตรวจจับคำที่สะกดผิดได้ถูกต้อง สามารถทำการแก้คำที่สะกดผิดเหล่านั้นได้เกือบทั้งหมดและถูกต้อง เมื่อผู้วิจัยนำผลการทดสอบมาวิเคราะห์ถึงปัจจัยที่ทำให้ระบบทำการแก้ไขการสะกดผิดได้ไม่ถูกต้องพบว่าปัจจัยส่วนหนึ่งที่ทำให้ระบบแก้ไขคำสะกดผิดไม่ได้นั้นเป็นเพราะข้อจำกัดในการ

ปรับแก้การสะกดผิดของตัวระบบ ซึ่งระบบจะสามารถปรับแก้คำที่สะกดผิดแต่ละคำได้เพียงแค่หนึ่งตัวอักษรเท่านั้น ตัวอย่างเช่น ระบบสามารถปรับแก้คำว่า “เมิล” ซึ่งเป็นคำที่สะกดผิดให้ถูกต้องได้ด้วยการแทนที่ ล ด้วย น ได้เป็น “เมิน” แต่ระบบจะไม่สามารถแก้ไขคำสะกดผิดที่ต้องการการปรับแก้ที่มากกว่าหนึ่งตัวอักษรได้ ตัวอย่างเช่นคำว่า “เสียสูญ” หากต้องการปรับแก้ให้เป็นคำที่สะกดถูกต้องระบบต้องทำการปรับแก้มากกว่าหนึ่งครั้ง คือ ต้องแทนที่ ส ด้วย ศ นับเป็นหนึ่งครั้ง และแทนที่ ญ ด้วย น นับเป็นอีกหนึ่งครั้ง จากนั้นยังต้องเติม ย และตัวการันต์อีกจึงจะได้คำว่า “เสียศูนย์” ซึ่งเมื่อนับจำนวนครั้งที่ใช้ในการปรับแก้คำว่า “เสียสูญ” ให้ถูกต้องนั้นต้องปรับแก้ถึง 4 ครั้ง ซึ่งคำเหล่านี้อาจจะไม่เหมาะที่จะใช้วิธีในการปรับแก้น้อยสุดในการแก้ไข นอกจากปัจจัยเรื่องข้อจำกัดของจำนวนในการปรับแก้คำที่สะกดผิดแล้ว ค่าความถี่ในการปรากฏของคำที่สะกดถูกต้องก็เป็นอีกปัจจัยหนึ่งที่ทำให้ระบบเลือกคำมาใช้แก้ไขการสะกดผิดผิดพลาดไป เนื่องจากระบบจะเลือกคำที่ทำให้ความน่าจะเป็นของข้อความสูงสุดเพียงคำเดียวเท่านั้น เพราะฉะนั้นถ้าหากว่าคำที่ระบบควรเลือกมาแก้ไขการสะกดผิดที่ตรวจพบในข้อความ แต่คำๆ นั้นเป็นคำที่มีความถี่ในการปรากฏน้อยจึงทำให้ค่าความน่าจะเป็นของข้อความน้อยตามไปด้วย ทำให้ระบบเลือกคำที่ให้ค่าความน่าจะเป็นสูงกว่า เป็นต้น

5.3 การตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองยูนิแกรม

ระบบตรวจแก้การสะกดผิดแบบที่สองนี้ เป็นวิธีการพื้นฐานของงานด้านการตรวจแก้การสะกดผิดด้วยคอมพิวเตอร์ ซึ่งมักจะถูกใช้เพื่อนำผลการทำงานหรือค่าประสิทธิภาพในการทำงานของวิธีการพื้นฐานมาเปรียบเทียบกับค่าประสิทธิภาพในการทำงานของวิธีการที่ผู้วิจัยแต่ละท่านนำเสนอ และด้วยเหตุผลเดียวกันนี้ผู้วิจัยจึงได้เลือกที่จะใช้แบบจำลองยูนิแกรมเป็นวิธีการพื้นฐานเพื่อนำผลการทำงานมาเปรียบเทียบกับผลการทำงานของวิธีการตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรม แต่เนื่องจากงานวิจัยนี้สนใจเฉพาะการตรวจแก้การสะกดผิดแบบเป็นคำจริงเท่านั้น ซึ่งหากใช้เฉพาะคลังข้อมูลยูนิแกรมในการตรวจจับและแก้ไขการสะกดผิดแบบเป็นคำจริงคงจะไม่สามารถนำผลการทำงานที่ได้มาเปรียบเทียบหรือประเมินประสิทธิภาพของการตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรมได้ เนื่องจากแบบจำลองยูนิแกรมนั้นสามารถตรวจจับเฉพาะการสะกดผิดแบบไม่เป็นคำ ซึ่งไม่อยู่ในขอบเขตและความสนใจของงานวิจัยนี้ ดังนั้นการตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมในงานวิจัยนี้จะแตกต่างจากการใช้แบบจำลองไตรแกรมเฉพาะในส่วนของการคำนวณค่าความน่าจะเป็นของข้อความเพื่อให้ระบบเลือกทำการแก้ไขการสะกดผิดในข้อความที่เหมาะสมที่สุดหรือเลือกใช้คำที่ทำให้ค่าความน่าจะเป็นของประโยคสูงที่สุดในการแก้ไขการสะกดผิด

5.3.1 หลักการทำงานของระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรม

หลักการทำงานของระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองยูนิแกรมนี้ค่อนข้างคล้ายคลึงกับหลักการทำงานของวิธีการตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรม ซึ่งแบ่งออกเป็น 3 ขั้นตอนหลัก ได้แก่ หาคำที่ต้องสงสัย ปรับแก้คำที่ต้องสงสัย และเลือกคำที่เหมาะสมเพื่อแก้ไขการสะกดผิด ดังที่ผู้วิจัยได้กล่าวถึงการปรับวิธีการตรวจแก้การสะกดผิดของแบบจำลองยูนิแกรมให้เหมาะสมกับงานตรวจแก้การสะกดผิดแบบเป็นคำจริงของงานวิจัยนี้เอาไว้ในตอนต้น ในขั้นตอนการตรวจหาคำต้องสงสัยด้วยแบบจำลองยูนิแกรมนี้จะดำเนินการตรวจจับคำที่ต้องสงสัยว่าสะกดผิดในข้อความเหมือนกระบวนการตรวจจับการสะกดผิดด้วยแบบจำลองไตรแกรมทุกประการ เพราะแบบจำลองยูนิแกรมนั้นพิจารณาเฉพาะคำเดี่ยวๆ จึงไม่สามารถนำมาใช้ตรวจจับการสะกดผิดแบบเป็นจริงที่ต้องอาศัยคำข้างเคียงในการตรวจจับได้ ในส่วนของขั้นตอนการปรับแก้การคำต้องสงสัยเพื่อให้ได้คำที่จะนำมาใช้แก้ไขการสะกดผิดนั้นก็มีขั้นตอนเหมือนกันกับกระบวนการปรับแก้การสะกดผิดของแบบจำลองไตรแกรม แต่ขั้นตอนที่ทำให้การตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมนี้ต่างไปจากการตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรมก็คือ การเลือกคำที่คำที่เหมาะสมมาแก้ไขการสะกดผิดที่ตรวจจับได้ภายในข้อความ เพื่อไม่เป็นการนำเสนอข้อมูลที่ซ้ำซ้อนผู้วิจัยจะนำเสนอขั้นตอนในการทำงานของระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมเฉพาะในส่วนของขั้นตอนการเลือกคำที่ถูกต้องเพื่อใช้แก้ไขคำที่สะกดผิดที่แตกต่างไปจากระบบตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรม

5.3.1.1 ขั้นตอนการเลือกคำที่ถูกต้องเพื่อใช้แก้ไขการสะกดผิด

สำหรับขั้นตอนในการเลือกคำที่ถูกต้องเพื่อนำไปใช้แก้ไขการสะกดผิดด้วยแบบจำลองยูนิแกรมนั้น ก็คล้ายคลึงกับการแก้ไขด้วยแบบจำลองไตรแกรม คือ ขั้นแรกนั้น สายคำเรียงสามที่ได้รับการปรับแก้การสะกดผิดแต่ละสายจะถูกนำไปแทนที่ส่วนที่สะกดผิดในข้อความ แล้วจึงนำข้อความที่ได้รับการแก้ไขแล้วไปผ่านกระบวนการตัดคำเพื่อระบุขอบเขตของคำแต่ละคำในข้อความ ต่อจากนั้นก็นำข้อความแต่ละข้อความที่ได้รับการแก้ไขเรียบร้อยแล้วมาคำนวณหาค่าความน่าจะเป็นของข้อความ ซึ่งในส่วนนี้เองที่ทำให้ระบบตรวจแก้การสะกดผิดแบบยูนิแกรมแตกต่างไปจากระบบตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรม คือ เปลี่ยนจากที่จะคำนวณหาค่าความน่าจะเป็นของข้อความด้วยการใช้คลังข้อมูลไตรแกรมไปใช้เป็นคลังข้อมูลยูนิแกรมแทน แล้วจึงเลือกสายคำที่ให้ค่าความน่าจะเป็นของข้อความสูงสุดเป็นสายคำที่สะกดถูกต้องและใช้แก้ไขข้อความ และเมื่อระบบเลือกสายคำที่นำไปแก้ไขการสะกดผิดในข้อความได้แล้วสายคำต้องสงสัยก่อนที่จะถูกปรับแก้และนำไปใช้แก้ไขคำที่สะกดผิดนั่นเองคือสายคำที่สะกดผิดจริง

5.3.2 ประสิทธิภาพการตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรม

ถึงแม้ว่าระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมจะเหมือนกับระบบตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรมในเกือบทุกขั้นตอนยกเว้นเฉพาะขั้นตอนการเลือกคำที่เหมาะสมเพื่อใช้แก้ไขคำที่สะกดผิด แต่เนื่องจากระบบตรวจแก้การสะกดผิดทั้งสองแบบนั้นไม่ได้ทำการระบุว่าสายคำต้องสงสัยคำใดเป็นสายที่สะกดผิดจริงก่อนจะส่งไปทำการปรับแก้ เพราะฉะนั้นสายคำต้องสงสัยทุกสายจะถูกนำไปปรับแก้ และระบบจะทำการเลือกสายคำปรับแก้แล้วให้ค่าความน่าจะเป็นของข้อความสูงที่สุดเป็นสายคำที่ถูกต้องและสายคำต้องสงสัยที่ถูกปรับแก้แล้วเป็นสายคำที่ระบบเลือกก็คือสายคำที่สะกดผิด เพราะฉะนั้นทั้งคำที่สะกดถูกและผิดของทั้งระบบไตรแกรมและยูนิแกรมนั้นได้มาจากการเลือกข้อความที่มีค่าความน่าจะเป็นสูงสุด แต่คลังข้อมูลที่นำมาใช้คำนวณค่าความน่าจะเป็นของข้อความในขั้นตอนสุดของของระบบทั้งสองนั้นเป็นคนละคลังข้อมูลกันจึงทำให้ผลการทดสอบระบบออกมาแตกต่างกัน ซึ่งจากการทดสอบระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมตรวจแก้การสะกดผิดในข้อความทดสอบ 1,000 ข้อความ พบว่าระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมใช้เวลาในการประมวลผลในส่วนของการอ่านคลังข้อมูลต่างๆ เพื่อเตรียมระบบให้พร้อมสำหรับตรวจแก้การสะกดผิด ซึ่งใช้เวลาใกล้เคียงกับระบบตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรม คือประมาณ 50 วินาที และใช้เวลาอีกประมาณ 46 วินาทีในการตรวจแก้ข้อมูลทดสอบ รวมใช้เวลาในการประมวลผลทั้งหมด 96 วินาที ในส่วนของผลการตรวจแก้การสะกดผิดในข้อความทดสอบพบว่า ระบบสามารถตรวจจับการสะกดผิดในข้อความทดสอบได้ถูกต้องจำนวน 490 ข้อความ ตรวจจับคำที่สะกดผิดผิดพลาดไป 507 ข้อความ และมี 3 ข้อความที่ระบบตรวจไม่พบคำที่สะกดผิดใดๆ ในข้อความ ส่วนผลการแก้ไขการสะกดผิดของระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมพบว่า จากข้อความที่สามารถตรวจจับการสะกดผิดได้ถูกต้องจำนวน 490 ข้อความ ระบบสามารถแก้ไขการสะกดผิดในข้อความเหล่านี้ได้ 426 ข้อความ ซึ่งแปลว่าจำนวนข้อความที่ระบบแก้ไขได้ไม่ถูกต้องนั้นก็คือ 64 ซึ่งเมื่อนำข้อมูลตัวเลขเหล่านี้ไปคำนวณแล้วได้ค่าความครบถ้วนและค่าความแม่นยำในการตรวจจับและแก้ไขการสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองยูนิแกรมดังที่ปรากฏในตารางที่ 5.4 ด้านล่าง

ตารางที่ 5.4 แสดงค่าประสิทธิภาพในการเลือกคำมาแก้ไขการสะกดผิดได้ถูกต้องด้วยแบบจำลองยูนิแกรม

	Detection	Correction
Recall (R)	*0.49*	0.87
Precision (P)	*0.49*	0.87
F-measure (F-1)	*0.49*	0.87

จากตารางที่ 5.4 จะเห็นว่าการใช้คลังข้อมูลยูนิแกรมในการเลือกคำมาแก้ไขการสะกดผิดในข้อความนั้นสามารถทำได้ค่อนข้างแม่นยำ สังเกตได้จากค่าความแม่นยำ (P) ที่ 0.87 จากการวิเคราะห์ข้อมูลพบว่า ปัญหาเรื่องของความถี่ในการปรากฏของคำที่ถูกต้องลดน้อยลง แต่ยังคงมีอยู่ตัวอย่างเช่น ข้อความ “เหตุการณ์|ไฟ|ไหม้|ครั้ง|แรก|ที่|เกิด|ขึ้น|ใน|ช่วง|บ่าย|วัน|ศุกร์|ที่ผ่านมา|” และคำที่จะนำมาใช้แก้คือคำว่า “ไหม้” และ “ใหม่” ซึ่งเมื่อนำไปแทนที่คำที่สะกดผิดในข้อความแล้วได้ค่าความน่าจะเป็นของข้อความเท่ากับ 0.00875 กับ 0.7758 ตามลำดับ คำที่ถูกเลือกมาแก้คำที่สะกดผิดในข้อความนี้คือคำที่ให้ค่าความน่าจะเป็นของข้อความสูงสุด ซึ่งก็คือคำว่า “ใหม่” เพราะฉะนั้นถึงแม้ว่าการใช้คลังข้อมูลยูนิแกรมอาจจะช่วยแก้ไขปัญหาในเรื่องของความถี่ในการปรากฏของคำได้ในบางกรณี แต่วิธีการนี้จะใช้ได้จริงกับเฉพาะเมื่อคำที่มีความถี่ในการปรากฏสูงเป็นคำที่สะกดถูกต้องเท่านั้น

5.4 การตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยชุดคำสับสน

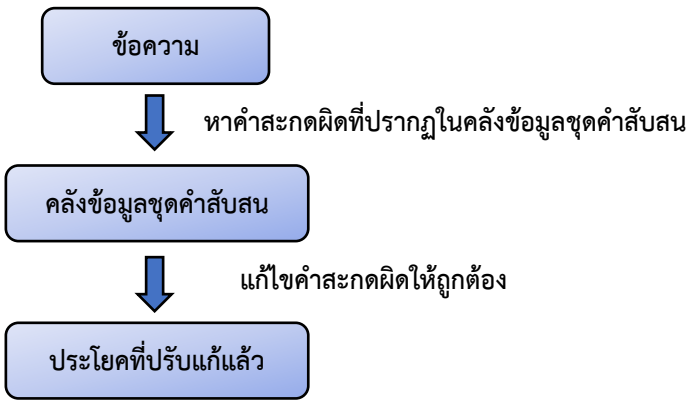
ระบบตรวจแก้การสะกดผิดในแบบที่สามนี้เป็นการนำเอาคลังข้อมูลชุดคำสับสนมาใช้ในการตรวจจับและแก้ไขการสะกดผิด เหตุผลที่ผู้วิจัยเลือกใช้วิธีการนี้ในการตรวจแก้การสะกดผิดแบบเป็นคำจริงคือเพื่อทดสอบสมมติฐานของงานวิจัยนี้ที่คาดว่าการใช้ชุดคำสับสนในการแก้ไขการสะกดผิดนั้นจะสามารถประมวลผลได้รวดเร็วกว่าผลการทำงานของวิธีการแก้ไขแก้น้อยสุด และผู้วิจัยต้องการทราบว่าการตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยการใช้ชุดคำสับสนนั้นจะสามารถทำได้มีประสิทธิภาพมากน้อยเพียงใด และประสิทธิภาพในการตรวจแก้การสะกดผิดด้วยวิธีการนี้จะถูกนำไปเปรียบเทียบกับอีกทั้งสองวิธีที่กล่าวมาแล้วด้วยเช่นกัน

5.4.1 หลักการทำงานของระบบตรวจแก้การสะกดผิดด้วยชุดคำสับสน

การตรวจแก้การสะกดผิดด้วยการใช้ชุดคำสับสนนั้นมีขั้นตอนในการประมวลผลที่ไม่ซับซ้อนอย่างสองวิธีการแรก ซึ่งหลักการทำงานของวิธีการนี้ก็คือการนำคำที่สะกดถูกต้องไปแทนที่คำที่สะกด

ผิด โดยอาศัยข้อมูลในคลังข้อมูลชุดคำสับสน ขั้นตอนในการทำงานของการตรวจการแก้สะกดผิดแบบเป็นคำจริงด้วยชุดคำสับสนมีดังนี้

เมื่อข้อความที่ต้องการให้ตรวจแก้การสะกดผิดได้ถูกป้อนให้กับระบบแล้ว ระบบจะอ่านข้อความทีละข้อความและเอาเครื่องหมาย | ที่ใช้บอกขอบเขตของคำในข้อความออก และเหตุผลที่ต้องนำเครื่องหมายแบ่งคำ | ออกก่อนที่จะเริ่มทำการตรวจจับคำที่สะกดผิดนั้นเป็นเพราะว่าในช่วงฝึกฝนระบบผู้วิจัยพบว่าเครื่องบอกขอบเขตคำนั้นมีผลต่อการตรวจแก้การสะกดผิด คือ คำสะกดผิดบางคำเมื่อปรากฏร่วมกับคำอื่นแล้วระบบตัดคำจะตัดคำเปลี่ยนไปทำให้ใช้ชุดคำสับสนตรวจจับไม่พบ ตัวอย่างเช่น คำว่า “|ชตา|” และ “|ชะตา|” เป็นชุดคู่คำสะกดผิดและสะกดถูกต้องที่อยู่ภายในชุดคำสับสน นั่นหมายความว่าเมื่อไหร่ก็ตามที่พบคำที่สะกดผิด คำนั้นจะถูกแก้ไขด้วยคำที่สะกดถูกต้อง แต่ในกรณีนี้ ผู้วิจัยพบว่าเมื่อคำว่า “ชตา” ปรากฏพร้อมกับคำว่า “โชค” เมื่อนำไปผ่านระบบตัดคำ จะถูกตัดคำออกเป็น “|โช|คช|ตา|” หากใช้ชุดคำสับสนในการตรวจจับคำที่สะกดผิดในลักษณะเช่นนี้ก็จะทำให้ตรวจไม่พบคำที่สะกดผิด จากนั้นระบบจะนำคำที่สะกดผิดในชุดคำสับสนมาเทียบดูว่าในข้อความมีคำสะกดผิดที่ตรงกับข้อมูลในชุดคำสับสนปนอยู่ด้วยหรือไม่ โดยคำสะกดผิดในชุดคำสับสนที่ยาวที่สุดจะถูกนำมาเทียบกับข้อความทดสอบก่อน และเรียงลำดับไปหาคำที่สั้นที่สุด หรือก็คือคำที่มีจำนวนตัวอักษรมากที่สุดจะถูกนำมาเทียบก่อน หากว่ามีความยาวเท่ากันก็เรียงลำดับตามตัวอักษรเริ่มจากตัว ก ไปหาตัว ฮ ซึ่งภายในคลังข้อมูลชุดคำสับสนนั้นได้รวบรวมคำที่มักเขียนผิดเก็บเอาไว้เป็นรายการคำ ซึ่งคำที่สะกดผิดแต่ละคำในคลังข้อมูลชุดคำสับสนจะถูกเก็บไว้คู่กับคำที่สะกดถูกต้อง ดังนั้นถ้าหากระบบตรวจพบว่าในข้อความมีคำสะกดผิดที่ตรงกับข้อมูลในชุดคำสับสนอยู่ ระบบก็จะแทนที่คำที่สะกดผิดนั้นด้วยคำสะกดถูกต้อง แต่ถ้าหากไม่พบก็แสดงว่าภายในข้อความนั้นไม่มีคำที่สะกดผิดที่ปรากฏในคลังข้อมูลชุดคำสับสน หรือระบบตรวจไม่พบคำที่สะกดผิดนั้น



รูปภาพที่ 5.4 กระบวนการตรวจแก้การสะกดผิดด้วยชุดคำสับสน

5.4.2 ประสิทธิภาพการตรวจแก้การสะกดผิดด้วยชุดคำสับสน

หลังจากที่ผู้วิจัยได้ทดสอบวิธีการตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยชุดคำสับสน พบว่าด้วยคลังข้อมูลชุดคำสับสนที่มีขนาดเล็กกว่าคลังข้อมูลไตรแกรมและยูนิแกรมมาก ทำให้วิธีการนี้ใช้เวลาในการประมวลผลน้อยที่สุด คือ 3 วินาที เนื่องจากเป็นวิธีการตรวจแก้การสะกดผิดที่ไม่ซับซ้อนและไม่จำเป็นต้องใช้คลังข้อมูลขนาดใหญ่ในการตรวจแก้การสะกดผิด ซึ่งถือเป็นข้อดีของการตรวจแก้ด้วยวิธีนี้ ในการตรวจแก้การสะกดผิดในข้อมูลทดสอบ ซึ่งมีข้อความภาษาไทยอยู่ทั้งหมด 1,000 ข้อความ พบว่าวิธีการนี้สามารถตรวจจับการสะกดผิดในข้อความได้ถูกต้องเกือบทั้งหมด คือ 978 ข้อความจากทั้งหมด 1,000 ข้อความ มีเพียง 22 ข้อความเท่านั้นที่ระบบตรวจแก้การสะกดผิดด้วยชุดคำสับสนตรวจจับได้ผิดพลาด และในส่วนของการแก้ไขการสะกดผิดที่ตรวจพบถูกต้องทั้งหมด 978 ข้อความนั้น ระบบตรวจแก้การสะกดผิดด้วยชุดคำสับสนนี้สามารถแก้ไขได้ถูกต้องทุกข้อ ความ ซึ่งข้อมูลในตารางที่ 5.5 แสดงค่าความครบถ้วน ค่าความแม่นยำ และค่าประสิทธิภาพในการตรวจจับและการแก้ไขการสะกดผิดด้วยชุดคำสับสน

ตารางที่ 5.5 แสดงประสิทธิภาพในการตรวจจับการสะกดผิดด้วยชุดคำสับสน

	Detection	Correction
Recall (R)	0.978	1
Precision (P)	0.978	1
F-measure (F-1)	0.978	1

สำหรับประสิทธิภาพในการตรวจแก้การสะกดผิดด้วยการใช้ชุดคำสับสนนั้น ผู้วิจัยคาดว่าวิธีการนี้ควรจะสามารถตรวจจับการสะกดผิดและแก้ไขการสะกดผิดได้อย่างถูกต้องและครบถ้วนทั้งหมด หลังจากที่ผู้วิจัยได้นำผลการทดสอบมาวิเคราะห์ถึงสาเหตุที่ทำให้เกิดความผิดพลาดในการตรวจจับการสะกดผิดที่เกิดขึ้นพบว่า ความสั้นยาวของคำในชุดคำสับสนที่นำมาใช้ตรวจจับการสะกดผิดนั้นเป็นสาเหตุที่ทำให้ระบบตรวจจับการสะกดผิดได้ไม่ถูกต้อง ตัวอย่างเช่น ในข้อความนี้ “เกิดความคิดสร้างสรรค์สร้างแรงบันดานใจให้เด็ก” ซึ่งคำที่สะกดผิดก็คือ คำว่า “บันดาน” ที่ควรถูกแก้ไขเป็น “บันดาล” แต่คำว่า “สร้างสรร” เป็นคำสะกดผิดที่ปรากฏภายในชุดคำสับสนเช่นกัน และเนื่องจากในการเทียบหาคำที่สะกดผิดในข้อความนั้น ระบบจะนำคำสะกดผิดในชุดคำสับสนที่มียาวที่สุดหรือมีจำนวนตัวอักษรมากที่สุดมาเทียบก่อน ดังนั้นเมื่อระบบทำการตรวจหาคำที่สะกดใน

ข้อความข้างก็จะตรวจจับว่า “สร้างสรร” เป็นคำที่สะกดผิดในข้อความข้างต้น เพราะสร้างสรรค์มีจำนวนตัวอักษรมากกว่า

5.5 ผลการเปรียบเทียบประสิทธิภาพในการตรวจแก้การสะกดผิดของระบบทั้งสาม

ในส่วนนี้ผู้วิจัยจะอภิปรายถึงประสิทธิภาพในการตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมซึ่งเป็นวิธีที่นำเสนอเปรียบเทียบกับประสิทธิภาพในการตรวจแก้การสะกดผิดด้วยวิธีที่นำมาใช้เปรียบเทียบอีกสองวิธี โดยจะพิจารณาถึงประสิทธิภาพของระบบแต่ละระบบใน 3 ด้าน ได้แก่ ด้านระยะเวลาที่ใช้ในการประมวลผล ด้านการตรวจกับการสะกดผิด และด้านการแก้ไขการสะกดผิด

5.5.1 ด้านระยะเวลาที่ใช้ในการประมวลผล

สำหรับประสิทธิภาพด้านระยะเวลาที่ระบบใช้ในการตรวจแก้ข้อความภาษาไทยจำนวน 1,000 ข้อความหรือประมาณ 33,000 คำ นั้นพบว่าระบบทั้งสามแบบใช้ระยะเวลาในการประมวลผลที่แตกต่างกัน ซึ่งระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมเป็นระบบที่ใช้เวลานานมากที่สุด คือ 128 วินาที ส่วนระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมเป็นระบบที่ใช้ระยะเวลาในการประมวลผลนานน้อยรองลงมา คือ 96 วินาที ส่วนระบบที่พบว่าใช้เวลาในการประมวลผลการตรวจแก้การสะกดผิดในข้อความทดสอบน้อยที่สุดคือ ระบบตรวจแก้การสะกดผิดด้วยชุดคำสับสน ซึ่งใช้เวลาประมาณ 3 วินาที ดังที่ปรากฏในตารางที่ 5.6 โดยผลการเปรียบเทียบด้านระยะเวลาที่ใช้ในการประมวลผลที่ปรากฏนั้นเป็นไปตามสมมติฐานข้อที่ 3 ของงานวิจัยนี้ ที่ผู้วิจัยได้คาดเดาไว้ว่าการตรวจแก้การสะกดผิดแบบเป็นคำจริงที่ใช้ชุดคำสับสนในการตรวจแก้การสะกดผิดจะใช้เวลาในการประมวลผลน้อยกว่าการใช้แบบจำลองไตรแกรมคำหรือแบบจำลองยูนิแกรมคำซึ่งใช้วิธีปรับแก้น้อยสุดในการตรวจแก้การสะกดผิด เนื่องจากวิธีการใช้ชุดคำสับสนในการตรวจแก้การสะกดผิดนั้น ได้กำหนดคำที่มักสะกดผิดและคำที่สะกดถูกต้องเอาไว้ล่วงหน้าแล้วเพราะฉะนั้นจึงไม่จำเป็นต้องเสียเวลาในการตรวจจับว่าคำใดในข้อความเป็นคำต้องสงสัยว่าสะกดผิดหรือปรับแก้คำที่สะกดผิดให้ถูกต้องด้วยวิธีการปรับแก้น้อยสุด เพื่อนำคำที่ถูกต้องมาแก้ไขการสะกดผิดในข้อความเหมือนวิธีการทางสถิติอย่างไตรแกรมละยูนิแกรม

ตารางที่ 5.6 แสดงผลการเปรียบเทียบระยะเวลาที่ใช้ในการประมวลของแต่ละระบบ

อันดับที่	ระบบที่ใช้ในการตรวจแก้การสะกดผิด	ระยะเวลาที่ใช้ในการประมวลผล (วินาที)
1	ชุดคำสับสน	3
2	ยูนิแกรม	96
3	ไตรแกรม	128

5.5.2 ด้านการตรวจจับการสะกดผิดแบบเป็นคำจริง

การประเมินประสิทธิภาพด้านการตรวจจับการสะกดผิดแบบเป็นคำจริงในงานวิจัยนี้จะพิจารณาจากค่าความแม่นยำ (precision) ในการตรวจจับคำที่สะกดผิดว่าระบบแต่ละระบบสามารถตรวจจับคำที่สะกดผิดในข้อความทดสอบได้ถูกต้องมากน้อยเพียงใด และพิจารณาจากค่าความระลึก (recall) ว่าระบบสามารถตรวจจับคำที่สะกดผิดได้ถูกต้องครบทุกคำหรือไม่ ดังที่ปรากฏในตารางที่ 5.7

ตารางที่ 5.7 แสดงค่าประสิทธิภาพในการตรวจจับการสะกดผิดด้วยระบบที่แตกต่างกัน

Detection	trigram	unigram	confusion set
recall (R)	0.465	0.49	0.978
precision (P)	0.466	0.49	0.978
F-measure (F-1)	0.466	0.49	0.978

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จากข้อมูลในตารางที่ 5.7 แสดงให้เห็นว่าระบบที่สามารถตรวจจับการสะกดผิดแบบเป็นคำจริงได้อย่างถูกต้องและครบถ้วนมากที่สุดคือระบบตรวจแก้สะกดผิดด้วยชุดคำสับสน รองลงมาเป็นระบบยูนิแกรมและไตรแกรม แต่ในความเป็นจริงแล้วค่าประสิทธิภาพในการตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยชุดคำสับสนในงานวิจัยนี้ขึ้นอยู่กับคำทั้งหมดภายในชุดคำสับสนเพียงอย่างเดียวเท่านั้นซึ่งมีทั้งหมด 1,674 คู่คำสะกดผิดและคำสะกดถูก ค่าประสิทธิภาพในการตรวจจับการสะกดผิดด้วยชุดคำสับสนอาจจะเท่ากับ 0 ก็ได้ถ้าหากว่าภายในข้อความที่ตรวจแก้นั้นไม่มีคำสะกดผิดที่ปรากฏอยู่ในชุดคำสับสนเลย เพราะฉะนั้นเมื่อพิจารณาถึงการนำไปใช้ตรวจแก้การสะกดผิดที่เกิดขึ้นในความเป็นจริง วิธีการนี้ถือเป็นวิธีการที่ไม่เหมาะสมนัก ในส่วนของประสิทธิภาพในการตรวจจับคำที่สะกดผิดด้วยการใช้ข้อมูลเชิงสถิติทั้งสองวิธีคือด้วยแบบจำลองไตรแกรมและแบบจำลองยูนิแกรมนั้นพบว่าสามารถตรวจจับการสะกดผิดได้ถูกต้องประมาณครึ่งต่อครึ่ง ซึ่งยังถือว่าเป็นจำนวนที่ยังไม่สามารถนำไปใช้ได้ในชีวิตจริง เพราะยังคงต้องพึ่งแรงงานคนในการตรวจแก้การสะกดผิดอีกครึ่งที่เหลือ

ปัจจัยที่ทำให้ผลของการตรวจแก้คำที่สะกดผิดด้วยแบบจำลองไตรแกรมและด้วยแบบจำลองยูนิแกรมมีค่าต่างกันถึงแม้ว่าการทำงานของทั้งสองระบบจะต่างกันแค่ในขั้นตอนการเลือกคำเพื่อใช้แก้ไขการสะกดผิดนั้นเป็นเพราะระบบที่ผู้วิจัยพัฒนาขึ้นนั้นไม่ได้ทำการระบุว่าคำต้องสงสัยใดเป็นคำที่สะกดผิดจริงก่อนจะส่งไปปรับแก้ให้ถูกต้อง กล่าวคือ ระบบนำคำที่ต้องสงสัยทั้งหมดไปปรับแก้แล้วเลือกคำที่ปรับแก้แล้วและเป็นคำที่ให้ค่าความน่าจะเป็นของข้อความสูงสุดหนึ่งคำมาใช้แก้ไขการสะกดผิด ซึ่งคำสะกดผิดที่ถูกแก้ไขแล้วให้ค่าความน่าจะเป็นของข้อความสูงสุดนั่นเองเป็นคำสะกดผิดที่ระบบตรวจจับได้ เพราะฉะนั้นอาจกล่าวได้ว่าระบบที่ผู้วิจัยพัฒนาขึ้นนั้นใช้ความน่าจะเป็นของข้อความทั้งในการตรวจจับและแก้ไขการสะกดผิด ด้วยเหตุนี้ค่าประสิทธิภาพในการตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรมและแบบจำลองยูนิแกรมจึงไม่เท่ากัน

5.5.3 ด้านการแก้ไขการสะกดผิดแบบเป็นคำจริง

ตารางที่ 5.8 แสดงค่าประสิทธิภาพในการแก้ไขการสะกดผิดด้วยระบบที่แตกต่างกัน

Correction	trigram	unigram	confusion set
recall (R)	0.85	0.87	1
precision (P)	0.85	0.87	1
F-measure (F-1)	0.85	0.87	1

จากข้อมูลที่ปรากฏในตารางที่ 5.8 จะเห็นได้ว่าทั้งสามระบบสามารถแก้ไขคำที่สะกดผิดแบบเป็นคำจริงได้ถูกต้องเป็นส่วนใหญ่ซึ่งค่า F-1 ของระบบตรวจแก้การสะกดผิดด้วยชุดคำสับสนเท่ากับ 1 นั้นหมายความคำสะกดผิดที่ระบบชุดคำสับสนสามารถตรวจจับได้ถูกต้องทั้งหมดนั้นได้รับการแก้ไขได้อย่างถูกต้องทั้งหมดเช่นกัน ส่วนประสิทธิภาพในการแก้ไขการสะกดผิดด้วยแบบจำลองไตรแกรมและแบบจำลองยูนิแกรมนั้นอยู่ในระดับที่ค่อนข้างน่าพอใจ เพราะทั้งสองระบบสามารถแก้ไขคำสะกดผิดได้ถูกต้องมากกว่า 80% แต่อย่างไรก็ตามค่าประสิทธิภาพในระดับนี้ยังหมายความว่า มีช่องว่างอีกประมาณ 20% เพื่อปรับปรุงและพัฒนาระบบให้สามารถแก้ไขการสะกดผิดแบบเป็นคำจริงได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

จากข้อมูลที่ปรากฏในตารางที่ 5.7 และ 5.8 พบว่ามีจุดที่น่าสนใจคือค่าประสิทธิภาพในการตรวจจับและการแก้ไขคำที่สะกดผิดของแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมนั้นต่ำกว่าการตรวจแก้ด้วยแบบจำลองยูนิแกรมอยู่เล็กน้อย ซึ่งเป็นผลที่อยู่นอกเหนือความคาดหมายของผู้วิจัยและทำให้สมมติฐานข้อที่ 2 ของงานวิจัยนี้ที่ผู้วิจัยได้ตั้งเอาว่าวิธีการตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรมจะสามารถตรวจสอบได้ถูกต้องมากกว่าวิธีการตรวจแก้ด้วยแบบจำลองยูนิแกรมนั้นเป็นเท็จ

ผู้วิจัยจึงได้วิเคราะห์ผลการทดสอบระบบทั้งสองระบบและพบว่าปัจจัยที่ทำให้การเลือกคำที่เหมาะสมเพื่อใช้แก้ไขคำที่สะกดผิดจากการใช้ความน่าจะเป็นยูนิแกรมคำนวณหาคำที่ให้ค่าความน่าจะเป็นของข้อความสูงสุดนั้นดีกว่าการใช้ความน่าจะเป็นไตรแกรมก็คือ ความถี่ในการปรากฏของคำที่เหมาะสมเนื่องจากแบบจำลองยูนิแกรมนั้นใช้ความถี่ในการปรากฏของคำหนึ่งคำเดี่ยวๆ จึงทำให้คำทุกคำในข้อความที่ถูกนำไปคำนวณหาค่าความน่าจะเป็นนั้นไม่มีปัญหาในเรื่องความถี่ในการปรากฏ ซึ่งอาจเป็นความบังเอิญที่ความถี่ในการปรากฏของคำที่เหมาะสมที่สุดนั้นมากกว่าคำอื่นๆ ระบบจึงสามารถเลือกมาแก้ไขได้ถูก ซึ่งในทางกลับกันหากว่าคำที่สะกดถูกต้องนั้นเป็นคำที่มีความถี่ในการปรากฏน้อยผลในการใช้ค่าความน่าจะเป็นยูนิแกรมในการเลือกคำที่เหมาะสมที่สุดนั้นลดลงได้

ปัญหาในเรื่องของความถี่ในการปรากฏนั้นยังเป็นปัจจัยสำคัญที่ทำให้ระบบตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรมไม่ได้เลือกคำที่ถูกต้องเพราะมีค่าความน่าจะเป็นน้อยกว่า ซึ่งผู้วิจัยยังพบว่าปัญหาเรื่องความถี่ในการปรากฏนี้ทำให้ระบบไตรแกรมไม่เลือกคำที่ระบบสามารถแก้ไขได้ถูกต้องถึง 82 ข้อความ และอาจจะเป็นสาเหตุสำคัญที่ทำให้ประสิทธิภาพในการตรวจจับและแก้ไขการสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมต่ำกว่าการใช้แบบจำลองยูนิแกรม

จากผลการวิเคราะห์ข้างต้นสามารถกล่าวได้ว่าในงานวิจัยนี้ ระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมนั้นยังคงเป็นระบบที่เหมาะสมในการนำมาใช้ตรวจแก้การสะกดผิดแบบเป็นคำจริงมากกว่าระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมและด้วยชุดคำสับสน แม้ว่าผลการประเมินประสิทธิภาพในการตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยระบบคอมพิวเตอร์ทั้งสามแบบในงานวิจัยนี้จะบ่งชี้ว่าวิธีการใช้ชุดคำสับสนในการตรวจแก้การสะกดผิดแบบเป็นคำจริงนั้นดีกว่าอีกสองแบบซึ่งเป็นเพราะคำที่สะกดผิดแบบเป็นคำจริงในข้อความทดสอบนั้นล้วนเป็นคำที่นำมาจากชุดคำสับสน ซึ่งถ้าหากว่าในประโยคทดสอบไม่มีคำในชุดคำสบสนระบบก็จะไม่สามารถตรวจจับการสะกดผิดใดๆ ได้ เพราะฉะนั้นในทางปฏิบัติวิธีการใช้ชุดคำสับสนเพียงอย่างเดียวในการตรวจแก้การสะกดผิดอาจไม่ใช่ทางเลือกที่เหมาะสมเท่าใดนัก และในบทต่อไปผู้วิจัยจะสรุปผลการวิจัย และกล่าวถึงปัญหาที่พบในงานวิจัยครั้งนี้ รวมถึงข้อเสนอแนะสำหรับผู้ที่สนใจจะทำงานวิจัยทางด้านนี้ในอนาคต

บทที่ 6

สรุปผลการวิจัย ปัญหาและข้อเสนอแนะ

ในบทนี้ผู้วิจัยจะกล่าวถึงผลการวิจัยในงานวิจัยนี้ทั้งหมดโดยสรุป และจะกล่าวถึงปัญหาที่พบในงานวิจัยนี้รวมถึงข้อเสนอแนะที่จะเป็นประโยชน์ต่องานวินัยอื่นๆ ต่อไปต่อไป

6.1 สรุปผลการวิจัย

ในส่วนนี้ผู้วิจัยจะนำเสนอผลการศึกษาวิเคราะห์เกี่ยวกับการตรวจแก้การสะกดผิดแบบเป็นคำจริงแยกเป็นสองส่วน คือ ในส่วนแรกจะเป็นการสรุปผลการศึกษาวิเคราะห์คำไทยที่มักสะกดผิด และในส่วนหลังจะเป็นการสรุปผลการทดสอบประสิทธิภาพในการทำงานของระบบตรวจแก้การสะกดผิดแบบเป็นคำจริง

6.1.1 สรุปผลการศึกษาวิเคราะห์คำไทยที่มักสะกดผิด

ในการศึกษาวิเคราะห์นี้ ผู้วิจัยได้รวบรวมเอาคำมักเขียนผิดในภาษาไทยจากสื่อแหล่งต่างๆ ทั้งหนังสือและบนอินเตอร์เน็ตมาศึกษาวิเคราะห์ โดยคัดเลือกเอาเฉพาะการสะกดผิดแบบเป็นคำจริงด้วยการนำคำที่สะกดผิดที่รวบรวมมาได้ทั้งหมดไปตัดคำด้วยระบบตัดคำ ThaiSegmentation พัฒนาโดย ภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ซึ่งสำหรับงานทางด้านการประมวลผลภาษาไทยนั้น เฉพาะคำที่มีความหมายและปรากฏในพจนานุกรมเท่านั้นที่จะตัดคำได้สำเร็จ ดังนั้นคำสะกดผิดที่ตัดคำได้สำเร็จก็คือคำที่สะกดผิดแบบเป็นคำจริง จากนั้นก็นำคำที่สะกดผิดแบบเป็นคำจริงทั้งหมดเหล่านี้ไปศึกษาวิเคราะห์ ซึ่งผู้วิจัยได้ตั้งสมมติฐานไว้ว่าคำที่สะกดผิดแบบเป็นคำจริงนั้นมักจะมีการสะกดผิดที่ตำแหน่งตัวสะกดมากที่สุดและจะยังคงออกเสียงเหมือนเดิม

จากการวิเคราะห์คำที่สะกดผิดแบบเป็นคำจริงที่รวบรวมมาได้ทั้งหมดพบว่า คำที่สะกดผิดแบบเป็นคำจริงเหล่านี้สามารถแบ่งออกเป็นสองกลุ่มใหญ่ตามจำนวนการสะกดผิดที่พบในหนึ่งคำ คือ กลุ่มที่สะกดผิดหนึ่งตำแหน่งและกลุ่มที่สะกดผิดหลายตำแหน่ง โดยสัดส่วนของคำที่สะกดผิดสองกลุ่มนี้คือ 80 ต่อ 20 หมายความว่าร้อยละ 80 ของคำที่สะกดผิดแบบเป็นคำจริงที่พบจะสะกดผิดเพียงหนึ่งตำแหน่งและร้อยละ 20 เป็นคำที่สะกดผิดหลายตำแหน่ง ซึ่งส่วนใหญ่คือสองตำแหน่ง เมื่อพิจารณาคำที่สะกดผิดในกลุ่มแรกคือกลุ่มที่พบการสะกดผิดหนึ่งตำแหน่งพบว่าสามารถจำแนกคำใน

กลุ่มนี้ออกตามตำแหน่งที่พบการสะกดผิดได้ 5 กลุ่มย่อยโดยเรียงอันดับตามจำนวนที่พบจากมากสุดไปน้อยสุด ได้ดังนี้

ก. กลุ่มคำสะกดผิดที่พยัญชนะต้น 382 คำ หรือ 28.53%

ข. กลุ่มคำสะกดผิดที่ตัวสะกด 338 คำ หรือ 25.24%

ค. กลุ่มคำสะกดผิดที่สระ 306 คำ หรือ 22.85%

ง. กลุ่มคำสะกดผิดที่ตัวการันต์ 270 คำ หรือ 20.16%

จ.กลุ่มคำสะกดผิดที่วรรณยุกต์ 43 คำ หรือ 3.21%

และเมื่อพิจารณาคำที่สะกดผิดหลายตำแหน่งพบว่าสามารถจำแนกออกเป็นสองกลุ่มเรียงตามจำนวนคำในกลุ่มจากมากสุดไปน้อยสุด ได้ดังนี้

ก. กลุ่มคำสะกดผิดหลายตำแหน่งที่ออกเสียงเหมือนเดิม 239 คำ หรือ 71.34%

ข. กลุ่มคำสะกดผิดหลายตำแหน่งที่ออกเสียงเปลี่ยนไป 96 คำ หรือ 28.66%

จากผลการวิเคราะห์คำที่สะกดผิดในงานวิจัยนี้จะเห็นได้ว่าคำสะกดผิดส่วนใหญ่ที่พบมากที่สุดเป็นคำสะกดผิดหนึ่งตำแหน่งที่พยัญชนะต้น รูปแบบของการสะกดผิดในลักษณะนี้ที่พบบ่อยที่สุด คือ การใช้ ร และ ล สลับกัน เช่น คำว่า “เลิกรา” สะกดผิดเป็น “เลิกลา” ส่วนคำสะกดผิดหนึ่งตำแหน่งที่ตัวสะกดซึ่งพบมากรองลงมานั้นพบว่ารูปแบบของการสะกดผิดที่พบบ่อยคือ การใช้ตัวสะกดมาตราเดียวกันผิดโดยไม่ทำให้การออกเสียงของคำเปลี่ยนไป โดยมาตราตัวสะกดที่ใช้ผิดมากที่สุดคือ แม่ กน เช่นคำว่า “เพิ่มพูน” สะกดผิดเป็น “เพิ่มพูล” ซึ่งจากผลการวิเคราะห์นี้แสดงว่าสมมติฐานข้อแรกที่ผู้วิจัยตั้งไว้นั้นเป็นจริงเพียงส่วนเดียวคือการสะกดผิดที่ตัวสะกดนั้นมักจะออกเสียงเหมือนเดิม

6.1.2 สรุปผลการทดสอบประสิทธิภาพในการทำงานของระบบตรวจแก้การสะกดผิดแบบเป็นคำจริง

ในงานวิจัยนี้นำเสนอระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรม ซึ่งข้อมูลที่นำมาสร้างเป็นคลังข้อมูลไตรแกรมนั้นใช้ข้อมูลจากคลังข้อมูลภาษาไทยแห่งชาติ และข้อมูลที่นำมาใช้ทดสอบนั้นเป็นข้อมูลที่ได้จากการนำเอาคำที่สะกดผิดแบบเป็นคำจริงในคลังข้อมูลชุดคำสับสนที่รวบรวมไว้ โดยผู้วิจัยได้สุ่มเลือกคำที่สะกดผิดในคลังข้อมูลออกมาจำนวน 375 คำ แล้วนำไปค้นหาข้อความตัวอย่างการใช้จริงที่พบบนอินเตอร์เน็ตได้จำนวนทั้งหมด 1,000 ข้อความ ดังนั้นแต่ละประโยคจะมีคำที่สะกดผิดแบบเป็นคำจริงที่ต้องการให้ระบบตรวจแก้ได้อยู่หนึ่งคำ ซึ่งกระบวนการทำงานของระบบนี้คือ แบ่งข้อความออกเป็นกลุ่มสายคำเรียงสามคำ แล้วนำสายคำเรียงสามทีละสาย

ไปตรวจสอบว่ามีสายคำเรียงสามสายใดปรากฏในคลังข้อมูลไตรแกรมหรือไม่ ถ้าหากไม่พบแสดงว่าสายคำเรียงสามนั้นต้องสงสัยว่ามีคำที่สะกดผิด และจะถูกนำไปปรับแก้ให้เป็นคำที่ถูกต้องด้วยวิธีการปรับแก้น้อยที่สุด (minimum edit distance) ซึ่งผู้วิจัยได้นำเอารูปแบบในการแก้ไขการสะกดผิดมาปรับใช้ในกระบวนการนี้ด้วย โดยรูปแบบในการแก้ไขการสะกดผิดที่นำมาใช้นั้นจะช่วยระบุกลุ่มตัวอักษรที่เป็นไปได้ในการปรับแก้ตัวอักษรแต่ละตัวให้ถูกต้อง เพราะโดยส่วนใหญ่แล้วตัวอักษรแต่ละตัวนั้นมีความเป็นไปได้ที่จะถูกใช้สลับกับตัวอักษรเพียงบางตัวเท่านั้น เช่น ร มักถูกแทนที่ด้วย ล ในตำแหน่งพยัญชนะต้น เช่น ราด ลาด เป็นต้น ซึ่งเป็นการช่วยให้ระบบทำการปรับแก้คำต้องสงสัยให้ถูกต้องได้รวดเร็วยิ่งขึ้น หลังจากที่สายคำเรียงสามที่ต้องสงสัยว่าสะกดผิดแต่ละสายได้รับการปรับแก้แล้ว จะถูกนำไปตรวจสอบว่ามีปรากฏอยู่ในคลังข้อมูลไตรแกรม หากไม่พบก็ในไปตรวจสอบกับคลังข้อมูลไบแกรมและยูนิแกรมตามลำดับ ซึ่งเฉพาะสายคำเรียงสามที่ตรวจพบว่ามีปรากฏอยู่ในคลังข้อมูลใดข้อมูลหนึ่งเท่านั้นที่จะถูกส่งต่อไปยังขั้นตอนสุดท้าย นั่นคือการนำสายคำเรียงสามที่ปรับแก้แล้วไปแทนที่คำที่สะกดผิดในข้อความแล้วคำนวณค่าความน่าจะเป็นของข้อความ และสายคำเรียงสามที่ให้ค่าความน่าจะเป็นของข้อความสูงที่สุดจะเป็นสายคำเรียงที่ถูกต้องและนำไปใช้แก้ไขคำที่สะกดผิดในข้อความ นอกจากนี้ผู้วิจัยได้พัฒนาระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงอีกสองแบบได้แก่ ระบบที่ใช้แบบจำลองยูนิแกรมและระบบที่ใช้ชุดคำสับสนในการตรวจแก้การสะกดผิดแบบเป็นคำจริงเพื่อนำผลการทดสอบการตรวจแก้การสะกดผิดด้วยระบบสองระบบนี้ไปใช้เปรียบเทียบประสิทธิภาพของระบบตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรม ส่วนหลักการทำงานของระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมนั้นมีขั้นตอนในการทำงานเหมือนกับระบบแบบจำลองไตรแกรมทั้งหมดยกเว้นหนึ่งขั้นตอนคือ ขั้นตอนเลือกคำที่เหมาะสมเพื่อใช้แก้ไขการสะกดผิดเท่านั้นที่แตกต่างไปจากระบบแบบจำลองไตรแกรม คือเปลี่ยนวิธีการคำนวณค่าความน่าจะเป็นของข้อความจากการใช้ค่าความน่าจะเป็นของไตรแกรม เป็นการใช้ความน่าจะเป็นของยูนิแกรมแทน และในส่วนหลักการทำงานของระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยชุดคำสับสนนั้นไม่ซับซ้อนเหมือนระบบที่ใช้ข้อมูลเชิงสถิติทั้งสองที่กล่าวไป ซึ่งขั้นตอนในการตรวจแก้การสะกดผิดด้วยชุดคำสับสนนั้นเริ่มต้นด้วยการนำคำที่สะกดผิดในชุดคำสับสนมาเทียบว่ามีปรากฏอยู่ในข้อความหรือไม่โดยเริ่มจากคำสะกดผิดที่ยาวที่สุดไปสั้นสุดหรือคำที่มีมีตัวอักษรในคำมากสุดไปน้อยสุด หากตรวจพบว่าในข้อความมีคำที่ตรงกับคำสะกดผิดในชุดคำสับสน ระบบจะทำการแก้ไขการสะกดผิดด้วยการนำเอาคู่คำที่สะกดถูกต้องของคำสะกดผิดในชุดคำสับสนมาแทนที่คำสะกดผิดที่ตรวจพบในข้อความ

ในส่วนของการประเมินประสิทธิภาพในการตรวจแก้การสะกดผิดแบบเป็นคำจริงในข้อความภาษาไทยที่นำมาใช้ทดสอบ 1,000 ข้อความของระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงทั้ง 3 แบบ ผู้วิจัยได้ประเมินประสิทธิภาพในการทำงานของระบบแต่ละระบบใน 3 ด้าน ได้แก่ ด้าน

ระยะเวลาที่ใช้ในการประมวลผล ด้านประสิทธิภาพในการตรวจจับการสะกดผิด และด้านประสิทธิภาพในการแก้ไขการสะกดผิด

ก. ด้านระยะเวลาที่ใช้ในการประมวลผล พบว่าระบบตรวจแก้การสะกดผิดด้วยชุดคำสับสนเป็นระบบที่ใช้เวลาในการประมวลผลน้อยที่สุดคือ ประมาณ 3 วินาที รองลงมาเป็นระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรมที่ใช้ระยะเวลาในการประมวลผล 96 วินาที ส่วนระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมคำเป็นระบบที่ใช้เวลานานมากที่สุด คือ 128 วินาที ซึ่งผลการเปรียบเทียบประสิทธิภาพในด้านระยะเวลาในการประมวลผลที่ได้นั้นเป็นไปตามที่ผู้วิจัยได้ตั้งสมมติฐานไว้ คือ การใช้ชุดคำสับสนในการแก้ไขนั้นจะใช้เวลาในการประมวลผลน้อยกว่าการปรับแก้น้อยสุดในการปรับแก้การสะกดผิด

ข. ด้านการตรวจจับการสะกดผิดแบบเป็นคำจริง พบว่าระบบที่ตรวจจับการสะกดผิดแบบเป็นคำจริงได้ดีที่สุดคือระบบที่ใช้ชุดคำสับสน รองลงมาเป็นระบบที่ตรวจจับการสะกดผิดด้วยแบบจำลองยูนิแกรมและแบบจำลองไตรแกรม ซึ่งมีค่าความแม่นยำและค่าความครบถ้วนในการตรวจจับการสะกดผิดเท่ากันคือ 0.978 0.49 และ0.47 ตามลำดับ

ค. ด้านการแก้ไขการสะกดผิดแบบเป็นคำจริง พบว่าระบบที่แก้ไขการสะกดผิดแบบเป็นคำจริงได้ดีที่สุดคือระบบตรวจแก้การสะกดผิดด้วยชุดคำสับสน ด้วยค่าความแม่นยำและค่าความครบถ้วนเท่ากับ 1 รองลงมาเป็นระบบตรวจแก้การสะกดผิดด้วยแบบจำลองยูนิแกรม ซึ่งมีค่าความแม่นยำและความครบถ้วนในการแก้ไขการสะกดผิดแบบเป็นคำจริงเท่ากับ 0.87 และอันดับสุดท้ายคือระบบตรวจแก้การสะกดผิดด้วยแบบจำลองไตรแกรมด้วยค่าความแม่นยำและค่าความครบถ้วนเท่ากับ 0.85

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ผลการทดสอบประสิทธิภาพของระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงสามแบบนี้แสดงให้เห็นว่าวิธีการตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมคำในงานวิจัยนี้นั้นยังมีจุดบกพร่องและต้องการการพัฒนาเพิ่มเติมเพื่อที่จะสามารถนำไปใช้ในการตรวจแก้การสะกดผิดแบบเป็นคำจริงได้จริงและมีประสิทธิภาพมากยิ่งขึ้น

6.2 ปัญหาที่พบ

ปัญหาหลักที่ผู้วิจัยพบในการพัฒนาระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมคำในงานวิจัยนี้ ที่ส่งผลต่อการทำงานของระบบเป็นอันมากมีดังต่อไปนี้

6.2.1 ปัญหาด้านระบบ

1) ระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมที่พัฒนาขึ้นในงานวิจัยนี้ยังขาดวิธีการตรวจจับคำที่สะกดผิดที่มีประสิทธิภาพ ทำให้ระบบต้องทำการแก้ไขคำสะกดผิดและใช้ระยะเวลาในการประมวลผลมากเกินความจำเป็น

2) ระบบตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยแบบจำลองไตรแกรมในการวิจัยนี้สามารถแก้ไขเฉพาะคำสะกดผิดที่ปรับแก้ให้ถูกต้องด้วยจำนวนการปรับแก้หนึ่งครั้งเท่านั้น คือ ระบบไม่สามารถแก้ไขคำสะกดผิดที่ต้องการการปรับแก้มากกว่าหนึ่งครั้งให้ถูกต้องได้ เช่น คำสะกดผิดที่เกิดจากการสลับที่ตัวอักษรสองตัวที่ติด คำว่า แสลง กับ สแลง ซึ่งในการจะแก้คำที่สะกดผิดในลักษณะเช่นนี้ต้องใช้การปรับแก้สองครั้งคือ ลบสระหน้า แ หรือพยัญชนะ ส ออก แล้วจึงค่อยเติมสระ แ หรือ ส กลับเข้าไป เป็นต้น

6.2.2 ปัญหาด้านข้อมูล

เนื่องจากข้อความที่นำมาใช้ฝึกฝนและทดสอบในงานวิจัยครั้งนี้ ผู้วิจัยต้องการข้อความที่เป็นตัวอย่างของการสะกดผิดที่เกิดขึ้นจริง เพื่อให้ผลการทดสอบระบบนั้นแสดงถึงประสิทธิภาพในการตรวจแก้การสะกดผิดที่เกิดขึ้นจริง แต่ปัญหาในการใช้ข้อมูลตัวอย่างที่สืบค้นบนอินเตอร์เน็ตคือข้อความตัวอย่างที่ได้มานั้นมีสิ่งรบกวนภายในข้อความปะปนอยู่เป็นจำนวนมาก ตัวอย่างเช่น การเว้นวรรคระหว่างคำหรือประโยค โครงสร้างประโยค หรือคำศัพท์ที่ใช้ ที่ไม่เป็นไปตามหลักการใช้ภาษา สิ่งเหล่านี้เป็นปัจจัยสำคัญอีกอย่างหนึ่งที่ส่งผลต่อประสิทธิภาพในตรวจแก้การสะกดผิดที่พัฒนาขึ้นในงานวิจัยนี้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

6.3 ข้อเสนอแนะ

เนื่องจากการตรวจแก้การสะกดผิดแบบเป็นคำจริงในภาษาไทยนั้นยังคงเป็นงานที่ท้าทายและต้องการการค้นคว้าวิจัยเพิ่มเติมเพื่อพัฒนาระบบที่สามารถตรวจแก้การสะกดผิดแบบเป็นคำจริงในภาษาไทยได้อย่างมีประสิทธิภาพมากยิ่งขึ้น ซึ่งผู้วิจัยมีข้อเสนอแนะสำหรับผู้ที่สนใจเกี่ยวกับงานทางด้านนี้หรือต้องการจะค้นคว้าวิจัยต่อไปในอนาคตดังนี้

สำหรับผู้ที่สนใจจะทำการวิจัยเกี่ยวกับการตรวจแก้การสะกดผิดแบบเป็นคำจริง ควรมุ่งเน้นการพัฒนาประสิทธิภาพในการตรวจจับการสะกดผิด เพราะหัวใจสำคัญของงานตรวจแก้การสะกดผิดคือ ระบบต้องสามารถระบุคำที่สะกดผิดให้ได้อย่างครบถ้วนและแม่นยำมากที่สุด เพราะการตรวจจับคำที่สะกดผิดแบบเป็นคำจริงให้ได้อย่างครบถ้วนและแม่นยำนั้นเป็นงานที่ยากกว่าการแก้ไขคำสะกด

ผิดมาก เนื่องจากการตรวจจับคำสะกดผิดแบบเป็นคำจริงจำเป็นต้องอาศัยคำบริบทข้างเคียงหรือความหมายช่วยในการระบุคำที่สะกดผิดนั้น

จากการศึกษาวิเคราะห์คำไทยที่มักเขียนผิดทำให้ผู้วิจัยทราบรูปแบบของการสะกดผิดที่มักจะเกิดขึ้น ผู้วิจัยได้นำข้อมูลที่ได้ในส่วนนี้ไปปรับใช้ในการปรับแก้การสะกดผิด แต่ไม่ได้นำมาปรับใช้ในการตรวจจับคำที่สะกดผิด ซึ่งผู้วิจัยคาดว่ารูปแบบของการสะกดผิดเหล่านี้จะช่วยเพิ่มประสิทธิภาพในการตรวจจับคำที่สะกดผิดให้ได้ครบถ้วนและแม่นยำมากยิ่งขึ้น

ในส่วนของประเด็นที่น่าสนใจและควรนำไปศึกษาค้นคว้าเพิ่มเติมหรือพัฒนานั้น ผู้วิจัยคิดว่าการขยายขอบเขตของงานวิจัยนี้ก็เป็นแนวทางหนึ่งที่สามารถนำไปพัฒนาต่อได้ เพราะผู้วิจัยได้กำหนดขอบเขตของการวิจัยไว้เป็นการตรวจแก้การสะกดผิดแบบเป็นคำจริงในภาษาไทยเท่านั้น ซึ่งในความเป็นจริงแล้วข้อความส่วนใหญ่มักมีคำภาษาอังกฤษปนอยู่ การค้นคว้าวิจัยหาวิธีแก้ไขการสะกดผิดที่อยู่ในข้อความภาษาไทยที่มีภาษาอังกฤษปนอยู่ก็เป็นงานวิจัยที่น่าสนใจ เนื่องจากผลการวิจัยที่ได้จะแสดงถึงความสามารถในการตรวจแก้การสะกดผิดแบบเป็นคำจริงในข้อความที่ใกล้เคียงกับที่พบจริงในชีวิตประจำวันมากขึ้น

งานทางด้านการตรวจแก้การสะกดผิดแบบเป็นคำจริงด้วยระบบคอมพิวเตอร์นั้น คำบริบทแวดล้อมที่ปรากฏร่วมกับคำที่สะกดผิดแบบเป็นคำจริงมีความสำคัญต่อการตรวจจับและแก้ไขคำที่สะกดผิดได้ถูกต้องเป็นอย่างมาก ซึ่งในงานวิจัยนี้ได้ใช้แบบจำลองไตรแกรมในการตรวจแก้การสะกดผิดนั้นเป็นวิธีการที่ใช้คำที่อยู่ติดกับคำที่สะกดผิดในระยะห่างไม่เกินสองคำ ซึ่งคำบริบทที่จะช่วยระบุได้ว่าคำๆ นั้นสะกดผิดอาจจะอยู่ห่างจากคำต้องสงสัยเกินกว่าสองคำ ดังนั้นถ้าหากมีการศึกษาค้นคว้าเกี่ยวกับบริบทของคำแต่ละคำหรือนำข้อมูลที่มีการกำกับข้อมูลชนิดของคำ (POS tagging) เอาไว้มาปรับใช้ในการตรวจแก้การสะกดผิดในลักษณะนี้ ก็อาจจะทำให้ประสิทธิภาพในการตรวจแก้การสะกดผิดแบบเป็นคำจริงเพิ่มมากขึ้นได้

รายการอ้างอิง

ภาษาไทย

ดนัย เมธิตานนท์. (2549). ๒๘๐ *คำไทย ที่มักเขียนและใช้กันผิด*. กรุงเทพฯ: มิติใหม่.

ตระการ เอี่ยมตระกูล. (2554). *คำไทยที่มักใช้ผิด*. กรุงเทพฯ: คลื่นอักษร.

ธนู ทดแทนคุณ. (2550). *ร้อยแปด (๑๐๘) คำไทยที่มักใช้ผิด*. ปทุมธานี: สกายบุ๊กส์.

ฝ่ายวิชาการ พีบีซี. (2553). *ภาษาไทย คำที่มักเขียนผิด*. กรุงเทพฯ: พีบีซี.

ราชบัณฑิตยสถาน. (2557). *อ่านอย่างไรและเขียนอย่างไร ฉบับราชบัณฑิตยสถาน (แก้ไขเพิ่มเติม)*. กรุงเทพฯ: ราชบัณฑิตยสถาน.

สุนันท์ อัญชลีนุกูล. (2552). *ระบบคำภาษาไทย*. กรุงเทพฯ: โครงการเผยแพร่ผลงานวิชาการ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

ภาษาอังกฤษ

Aroonmanakul, W. (2002). *Collocation and Thai Word Segmentation* Paper presented at the The Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop, Pathumthani.

Aroonmanakul, W. (2007). Thai National Corpus. Retrieved from http://www.arts.chula.ac.th/~ling/TNCII/

Baba, Y., & Suzuki, H. (2012, 8-14 July 2012). *How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs.* Paper presented at the the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea.

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 5(2), 179-190.

Bassil, Y. (2012). Parallel Spell-Checking Algorithm Based on Yahoo! N-Grams Dataset. *International Journal of Research and Reviews in Computer Science (IJRRCS),* 3(1).

Bowers, N. (2015). neilb/Text-Levenshtein. https://github.com/neilb/Text-Levenshtein

Dembitz, Š., Gledec, G., & Randić, M. (2009). Spellchecker *Wiley Encyclopedia of Computer Science and Engineering* (pp. 2793–2804).

Golding, A. R. (1995). A Bayesian Hybrid Method for Context-Sensitive Spelling Correction *Proceedings of the Third Workshop on Very Large Corpora*.

Golding, A. R., & Roth, D. (1999). A Winnow Based-Approach to Context Sensitive Spelling Algorithm. *Machine Learning,* 34, 107-130.

Golding, A. R., & Schabes, Y. (1996). *Combining Trigram-based and feature-based methods for context-sensitive spelling correction*. Paper presented at the Proceedings of the 34th annual meeting on Association for Computational Linguistics, Santa Cruz, California.

Hirst, G., & Budanitsky, A. (2005). Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering,* 11(1), 87-111. doi:10.1017/S1351324904003560

Islam, A., & Inkpen, D. (2009). *Real-Word Spelling Correction using Google Web 1T 3-grams.* Paper presented at the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore.

Kaur, J., & Garg, K. (2014). Hybrid Approach for Spell Checker and Grammar Checker for Punjabi. *International Journal of Advanced Research in Computer Science and Software Engineering,* 4(6).

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Comput. Surv.,* 24(4), 377-439. doi:10.1145/146370.146380

Mays, E., Damerau, F. J., & Mercer, R. L. (1991). Context based spelling correction. *Inf. Process. Manage.,* 27(5), 517-522. doi:10.1016/0306-4573(91)90066-u

Mishra, R., & Kaur, N. (2013). A Survey of Spelling Error Detection and Correction Techniques. *International Journal of Computer Trends and Technology,* 4(3), 372-374.

Mitton, R. (1987). Spelling Checkers, Spelling Correctors and the Misspellings of Poor Spellers. *Information Processing and Management,* 23(5), 495-505.

S, R. M., Madi, V., D, S., & P, R. K. (2012). A NON-WORD KANNADA SPELL CHECKER USING MORPHOLOGICAL ANALYZER AND DICTIONARY LOOKUP METHOD. *International Journal of Engineering Sciences & Emerging Technologies,* 2(2), 43-52.

Stehouwer, H. (2011). *Statistical langauge models for alternative sequence selection.* Tilburg University.

Verberne, S. (2002). Context-sensitive spell checking based on word trigram probabilities. *Unpublished master's thesis, University of Nijmegen*.

Wilcox-O'Hearn, L. A. (2014). Detection is the central problem in real-word spelling correction. *CoRR, abs/*1408.3153.

ประวัติผู้เขียนวิทยานิพนธ์

นาย พลวัฒน์ ไหลมนู เกิดที่จังหวัดอุดรธานี สำเร็จการศึกษาในระดับปริญญาศิลปศาสตรบัณฑิต สาขาวิชาภาษาอังกฤษ จากมหาวิทยาลัยเชียงใหม่ ในปีการศึกษา 2552 และเข้าศึกษาต่อในหลักสูตรอักษรศาสตรมหาบัณฑิต ภาควิชาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2556

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University