การสร้างตัวแบบตัดคำในภาษาบาลีไทยด้วยเทคนิคแบบผสมผสาน

โดย

นายกลางใจ ธรรมนาม

โครงร่างวิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต

ภาควิชาคอมพิวเตอร์

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ปีการศึกษา 2564

การสร้างตัวแบบตัดคำในภาษาบาลีไทยด้วยเทคนิคแบบผสมผสาน

โดย
นายกลางใจ ธรรมนาม

โครงร่างวิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
ภาควิชาคอมพิวเตอร์
บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร
ปีการศึกษา 2564

A HYBIRD APPROACH FOR PALI COMPOUNDS SPLITTING USING DEEP LEARNING AND RULE BASE

By

MR. Klangjai TAMMANAM

A Thesis Proposal Submitted in Partial Fulfillment of the Requirements

for Master of Science (INFORMATION TECHNOLOGY)

Department of COMPUTER SCIENCE

Graduate School, Silpakorn University

Academic Year 2021

หัวข้อ	การสร้างตัวแบบตัดคำในภาษาบาลีไทยด้วยเทคนิคแบบผสมผสาน
โดย	กลางใจ ธรรมนาม
สาขาวิชา	เทคโนโลยีสารสนเทศ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
อาจารย์ที่ปรึกษาหลัก	ดร. ณัฐโชติ พรหมฤทธิ์

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต

---------- คณบดีบัณฑิตวิทยาลัย
(รองศาสตราจารย์ ดร.จุไรรัตน์ นันทานิช)

พิจารณาเห็นชอบโดย

---------- ประธานกรรมการ
(รองศาสตราจารย์ ดร.อนิราช มิ่งขวัญ)

---------- อาจารย์ที่ปรึกษาหลัก
(ดร.ณัฐโชติ พรหมฤทธิ์)

---------- อาจารย์ที่ปรึกษาร่วม
(ดร.สัจจาภรณ์ ไวจรรยา)

---------- ผู้ทรงคุณวุฒิภายใน
(ผู้ช่วยศาสตราจารย์ ดร.อรวรรณ เชาวลิต)

สารบัญ

หน้า

สารบัญ ..... จ

สารบัญตาราง ..... ซ

สารบัญรูปภาพ ..... ญ

บทคัดย่อภาษาอังกฤษ ..... ฐ

กิตติกรรมประกาศ ..... ฑ

บทที่ 1 บทนำ 1

1.1 ที่มาและความสำคัญ 1

1.2 วัตถุประสงค์ของการวิจัย 4

1.3 ขอบเขตการวิจัย 4

1.4 ประโยชน์ที่คาดว่าจะได้รับ 5

บทที่ 2 ภาษาบาลีและวรรณกรรมที่เกี่ยวข้อง 6

2.1 การอ่านและออกเสียงในภาษาบาลี 6

2.2 วจีวิภาค: ประเภท หน้าที่ และความหมายของคำ 9

2.2.1 คำนาม (Noun) [8] 9

2.2.2 คำกิริยา (Verb) 10

2.2.3 อัพยยศัพท์ (Prefix, Suffix and Particle) 10

2.2.4 การประกอบคำ [8] 11

2.2.5 กติปยศัพท์ 20

2.2.6 สังขยา (Numeral) 20

2.2.7 คำสมาส (Samasa) 20

2.2.8 คำสนธิ (Sandhi) 21

2.3. วรรณกรรมที่เกี่ยวข้อง 21
2.3.1 การประมวลภาษาธรรมชาติด้านภาษาบาลีและสันสฤต 22
2.3.2 การเรียนรู้เชิงลึกสำหรับงานด้านการประมวลผลภาษา 24
2.3.2 การเรียนรู้เชิงลึกสำหรับงานด้านการตัดคำสนธิภาษาสันสกฤต 25
2.3.3 สรุปงานวิจัยที่เกี่ยวข้อง 25
บทที่ 3 ทฤษฎีที่เกี่ยวข้องและความรู้ที่เกี่ยวข้อง 27
3.1 โครงข่ายประสาทเทียมแบบย้อนกลับ 27
3.2 โครงข่ายแอลเอสทีเอ็ม (Long Short-Term Network) 28
3.3 โครงข่ายแอลเอสทีเอ็มแบบสองทิศทาง (Bidirectional Long Short-Term Network) 29
3.4 การวัดประสิทธิภาพ 30
บทที่ 4 วิธีดำเนินการวิจัย 31
4.1 การเตรียมข้อมูลและการรวบรวมชุดข้อมูล 31
4.2 การตัดคำสนธิ 35
4.2.1 การวิเคราะห์รูปแบบการแยกคำสนธิ 36
4.2.2 การทำนายตำแหน่งและรูปแบบตัดคำสนธิ 38
4.2.3 โมเดลทำนายตำแหน่งและรูปแบบตัดคำสนธิ 40
4.2.4 การวิเคราะห์กฎสำหรับรูปแบบการตัดคำสนธิ 43
4.3 การตัดคำสมาสภาษาบาลีอักษรไทย 55
4.3.1 การวิเคราะห์รูปแบบการแยกคำสมาส 55
4.3.2 การทำนายตำแหน่งและรูปแบบตัดคำคำสมาส 57
4.3.3 โมเดลทำนายตำแหน่งตัดคำสมาส 58
4.3.4 การแยกกลุ่มอักษร 59
4.3.5 การแยกเสียงพยางค์ 60
4.3.6 การแปลงรูปคำกลับ 62

บทที่ 5 ผลการดำเนินงาน 63
5.1 ผลการวิจัยการตัดคำสนธิ 63
5.1.1 การทำนายตำแหน่งและรูปแบบตัดคำสนธิ 63
5.1.2 การวัดประสิทธิภาพการตัดคำสนธิ 65
5.1.3 เปรียบเทียบการตัดคำสนธิด้วยโมเดลการตัดคำภาษาไทย 67
5.1.4 เปรียบเทียบกับงานวิจัยด้านการตัดคำในภาษาสันสกฤต 68
5.2 ผลการวิจัยการตัดคำสมาส 70
5.2.1 การทำนายตำแหน่งและรูปแบบตัดคำสมาส 70
5.1.2 การวัดประสิทธิภาพการตัดคำสนธิ 72
5.2.3 เปรียบเทียบการตัดคำสนธิด้วยโมเดลการตัดคำภาษาไทย 73
บทที่ 6 สรุปผลการดำเนินงาน และข้อเสนอแนะ 74
6.1 สรุปผลวิจัยการตัดคำสนธิ 75
6.1 สรุปผลวิจัยการตัดคำสมาส 75
6.2 ข้อเสนอแนะ 75
รายการอ้างอิง 76
ประวัติผู้เขียน 81

สารบัญตาราง

หน้า

ตารางที่ 1 หน่วยเสียงพยัญชนะในภาษาบาลี 6
ตารางที่ 2 เสียงสระในภาษาบาลี 7
ตารางที่ 3 เสียงพยัญชนะท้าย 8
ตารางที่ 4 เสียงพยัญชนะต้นควบกล้ำ 8
ตารางที่ 5 เสียงพยัญชนะท้ายควบกล้ำ 8
ตารางที่ 6 นามนามซึ่งมีต้นเค้าศัพท์เป็นเพศเดียว 12
ตารางที่ 7 นามนามที่มีต้นเค้าศัพท์เหมือนกัน เป็นได้ทั้ง 2 เพศ 12
ตารางที่ 8 คุณนามเป็นได้ทั้ง 3 เพศ 13
ตารางที่ 9 วิภัตตินามและหน้าที่ของคำ 13
ตารางที่ 10 ความหมายบอกเนื้อความของวิภัตตินามในภาษาบาลี [8] 13
ตารางที่ 11 การเปลี่ยนแปลงท้ายศัพท์ของคำที่มีต้นเค้าศัพท์เป็นเพศชาย 15
ตารางที่ 12 ตัวอย่างการผันคำของต้นเค้าศัพท์ที่เป็นเพศชาย 15
ตารางที่ 13 การเปลี่ยนแปลงท้ายศัพท์ของคำที่มีต้นเค้าศัพท์เป็นเพศหญิง 16
ตารางที่ 14 ตัวอย่างการผันคำของต้นเค้าศัพท์ที่เป็นเพศหญิง 16
ตารางที่ 15 การเปลี่ยนแปลงท้ายศัพท์ของคำที่มีต้นเค้าศัพท์มิใช่เพศหญิงและเพศชาย 17
ตารางที่ 16 ตัวอย่างการผันคำของต้นเค้าศัพท์มิใช่เพศหญิงและเพศชาย 17
ตารางที่ 17 การแจงแจงวัตตมานาวิภัตติ [8] 18
ตารางที่ 18 การประกอบคำกิริยาจากรากศัพท์ ปจฺ (หุง, ต้ม) 18
ตารางที่ 19 สรุปงานวิจัยที่เกี่ยวข้อง 26
ตารางที่ 20 ตัวอย่าง Confusion matrix 30
ตารางที่ 21 หนังสือธมฺมปทฏฺฐกถาที่ใช้รวบรวมคำสนธิและคำสมาส 31

ตารางที่ 22 ตัวอย่างข้อมูลคำสนธิและผลเฉลย 34
ตารางที่ 23 ตัวอย่างข้อมูลคำสมาสและผลเฉลย 35
ตารางที่ 24 คำอธิบายตำแหน่งและรูปแบบตัดคำสนธิ 39
ตารางที่ 25 ข้อมูลและผลเฉลยประเภทตำแหน่งตัดคำสนธิ 39
ตารางที่ 26 ตัวอักษรทั้งหมดที่ใช้ 41
ตารางที่ 27 ประเภทตำแหน่งตัดคำสมาส 57
ตารางที่ 28 กฎการแปลงรูปคำกลับ 62
ตารางที่ 29 จำนวนข้อมูลฝึกสอน ข้อมูลตรวจสอบ และข้อมูลทดสอบ 63
ตารางที่ 30 พารามิเตอร์สำหรับโมเดลการทำนายประเภทตำแหน่งตัดคำ 63
ตารางที่ 31 Confusion Matrix 65
ตารางที่ 32 Precision Recall และ F1-score 66
ตารางที่ 33 ตัวอย่างการทำนายตำแหน่งและรูปแบบตัดคำที่ไม่ถูกต้อง 66
ตารางที่ 34 การนับจำนวนการตัดคำสนธิเทียบกับชุดข้อมูลของผู้เชี่ยวชาญ 67
ตารางที่ 35 เปรียบเทียบการตัดคำสนธิกับตัวตัดคำภาษาไทย 67
ตารางที่ 36 แสดงผลลัพธ์เปรียบเทียบระหว่างวิธีที่นำเสนอกับตัวตัดคำภาษาไทย 68
ตารางที่ 37 เทียบเทียบการตัดคำสนธิระหว่างภาษาบาลีและภาษาสันสกฤต 69
ตารางที่ 38 จำนวนข้อมูลฝึกสอน ข้อมูลตรวจสอบ และข้อมูลทดสอบ 70
ตารางที่ 39 พารามิเตอร์สำหรับโมเดลการทำนายประเภทตำแหน่งตัดคำ 70
ตารางที่ 40 Confusion Matrix 72
ตารางที่ 41 Precision Recall และ F1-score 72
ตารางที่ 42 เปรียบเทียบการตัดคำสมาสกับโมเดลตัดคำภาษาไทย 73

สารบัญรูปภาพ

หน้า

รูปที่ 1 ประเภทของคำนาม 10
รูปที่ 2 ส่วนประกอบของคำ 11
รูปที่ 3 แผนภาพงานวิจัยการประมวลผลภาษาบาลี 23
รูปที่ 4 ตัวอย่างโครงข่ายประสาทเทียมแบบย้อนกลับ 27
รูปที่ 5 ตัวอย่างโครงข่ายแบบสองทิศทาง 29
รูปที่ 6 ตัวอย่างหนังสือ 32
รูปที่ 7 ตัวอย่างหนังสือ (ต่อ) 33
รูปที่ 8 ไฟล์ข้อความจากหนังสือธมฺมปทฏฺฐกถา (ปฐโม ภาโค) หน้าที่ 6 34
รูปที่ 9 ภาพรวมการวิจัยการตัดคำสนธิ 35
รูปที่ 10 การแยกคำสนธิรูปแบบที่ 1 37
รูปที่ 11 การแยกคำสนธิรูปแบบที่ 2 37
รูปที่ 12 การแยกคำสนธิรูปแบบที่ 3 37
รูปที่ 13 การแยกคำสนธิรูปแบบที่ 4 38
รูปที่ 14 ตัวอย่างการเตรียมผลเฉลยตำแหน่งและรูปแบบตัดคำสนธิ 39
รูปที่ 15 การเข้ารหัสคำสนธิก่อนป้อนเข้าโมเดลทำนายประเภทตำแหน่งตัดคำ 40
รูปที่ 16 นำผลเฉลยตำแหน่งมาแปลงเป็น one-hot vector 40
รูปที่ 17 โมเดลทำนายตำแหน่งและรูปแบบตัดคำสนธิ 42
รูปที่ 18 ภาพรวมการตัดคำสนธิ 43
รูปที่ 19 ตัวอย่างการตัดคำสนธิโดยใช้ผลการทำนายตำแหน่งและรูปแบบการตัดคำ 43
รูปที่ 20 สรุปการแบ่งคำและการแก้ไขคำสนธิ 44
รูปที่ 21 กฎการตัดคำสนธิรูปแบบที่ 2 45

รูปที่ 22 การแก้ไขคำด้วยกฎการตัดคำสนธิรูปแบบที่ 2 46
รูปที่ 23 กฎ Dictionary Lookup สำหรับการตัดคำสนธิรูปแบบที่ 2 47
รูปที่ 24 กฎการตัดคำสนธิรูปแบบที่ 3 และตัวอย่าง 48
รูปที่ 25 การแบ่งกลุ่มของกฎการตัดคำสนธิรูปแบบที่ 4 49
รูปที่ 26 กฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่ 1 49
รูปที่ 27 กฎการตัดคำสนธิที่ใช้อักษรก่อนหน้า (1) 50
รูปที่ 28 กฎการตัดคำสนธิที่ใช้อักษรก่อนหน้า (2) 50
รูปที่ 29 กฎการตัดคำสนธิที่ใช้อักษรแรกเป็นสระหน้า (1) 51
รูปที่ 30 กฎการตัดคำสนธิที่ใช้อักษรแรกเป็นสระหน้า (2) 52
รูปที่ 31 กฎการตัดคำสนธิที่ใช้อักษรแรกเป็นสระหน้า (3) 52
รูปที่ 32 กฎการตัดคำสนธิที่ใช้อักษรแรกเป็นสระหน้า (4) 53
รูปที่ 33 กฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่ 3 53
รูปที่ 34 กฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่ 4 (1) 54
รูปที่ 35 กฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่ 4 (2) 54
รูปที่ 36 กฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่ 5 และกลุ่มที่ 6 55
รูปที่ 37 การแยกคำสมาสรูปแบบที่ 1 56
รูปที่ 38 การแยกคำสมาสรูปแบบที่ 2 56
รูปที่ 39 การแยกคำสมาสรูปแบบที่ 3 56
รูปที่ 40 การแยกคำสมาสรูปแบบที่ 4 57
รูปที่ 41 คำสมาสที่มีอักษรที่ต้องลบ 58
รูปที่ 42 คำสมาสที่มีอักษรที่ต้องแยกเสียง 58
รูปที่ 43 การแยกกลุ่มอักษร 60
รูปที่ 44 กลุ่มอักษรที่ต้องนำไปแยกเสียงพยางค์ 60
รูปที่ 45 กฎการแยกเสียงพยางค์ 61

รูปที่ 42 การเปลี่ยนรูปคำกลับ 62
รูปที่ 47 ค่าความแม่นยำของข้อมูลตรวจสอบ.....64
รูปที่ 48 ค่าความคลาดเคลื่อนของข้อมูลตรวจสอบ 64
รูปที่ 49 การวัดประสิทธิภาพโดยเปรียบเทียบผลลัพธ์การตัดคำสนธิ.....67
รูปที่ 50 ค่าความแม่นยำของข้อมูลตรวจสอบ.....71
รูปที่ 51 ค่าความคลาดเคลื่อนของข้อมูลตรวจสอบ 71

59309202 : Major (INFORMATION TECHNOLOGY)

Keyword : BiLSTM, Sandhi Thai, Rule base, Sandhi splitting

MR.KLANGJAI TAMMANAM : A HYBRID APPROACH FOR PALI COMPOUNDS SPLITTING USING DEEP LEARNING AND RULE BASE THESIS ADVISOR: NUTTACHOT PROMRIT, Ph.D. CO-THESIS ADVISOR : SAJJAPORN WAIJANYA, Ph.D.

Pali Sandhi is a phonetic transformation from two words into a new word. The phonemes of the neighboring words are changed and merged. Pali Sandhi words segmentation is more challenging than Thai words segmentation because Pali is a high inflected language. This article describes a novel approach which predicts splitting locations by classifying the sample Sandhi words into 5 classes with BiLSTM model. We then applied the classified rules to rectify the words from the splitting locations. We identified 6 ,3 4 5 words of Pali Sandhi words from Dhammapada Atthakatha. We evaluated the performance of our model by the accuracy of the splitting locations and compared the results with the dataset. There were 9 2 . 2 0 percent of the correct results, 1.10 percent of Pali Sandhi words were predicted as non-splitting location words, and 5.83 percent were not matched with the answers (incomplete segmented).

กิตติกรรมประกาศ

ขอกราบขอบพระคุณ อาจารย์ ดร.ณัฐโชติ พรหมฤทธิ์ และ อาจารย์ ดร.สัจจาภรณ์ ไวจรรยา ที่ได้ให้ความเมตตากรุณามาเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ คอยให้ความรู้ คำแนะนำ ข้อคิด ในการทำวิทยานิพนธ์ตั้งแต่ต้นจนสามารถสำเร็จลุล่วงไปได้

ขอกราบขอบพระคุณ พระมหาฉัตรชัย ถาวโร (ฉัตรชัย อภิชล) เปรียญธรรม ๘ ประโยค วัดประดู่ จังหวัดสมุทรสงคราม ที่สละเวลามาจัดเตรียมคำสนธิและคำสมาส พร้อมทั้งผลเฉลยการตัดคำ ในขณะที่เตรียมตัวสอบไล่บาลีประโยค ๙ จนสามารถนำข้อมูลมาวิจัยในการทำวิทยานิพนธ์

ขอขอบพระคุณ รองศาสตราจารย์ ดร.อนิราช มิ่งขวัญ ประธานกรรมการสอบ และผู้ช่วยศาสตราจารย์ ดร.อรวรรณ เชาวลิต ผู้ทรงคุณวุฒิภายใน ที่สละเวลามาสอบวิทยานิพนธ์ พร้อมทั้งให้คำแนะนำ ข้อคิด ตั้งแต่การสอบนำเสนอหัวข้อและวันสอบรอบสุดท้าย

ขอขอบคุณ คณาจารย์ภาควิชาคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาศิลปากร ที่ประสิทธิ์ประสาทให้ความรู้ตั้งแต่เรียนปริญญาตรีรวมถึงปริญญาโท และ ภาควิชาคอมพิวเตอร์ที่มอบทุนการศึกษา

ขอขอบคุณ นักศึกษาภาควิชาคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิยาลัยศิลปากร ทุกท่านที่ให้ความช่วยเหลือ คำแนะนำ ในการวิทยานิพนธ์

ขอกราบพระคุณบุพการีและครอบครัว ที่คอยสนับสนุนและให้กำลังใจ

กลางใจ ธรรมนาม

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญ

ปัจจุบันพุทธศาสนิกชนในประเทศไทยให้ความสำคัญกับการศึกษาภาษาบาลีมากดังจะเห็นได้จากการขยายผลการจัดการเรียนการสอนและการทดสอบความรู้ด้านภาษาบาลี ซึ่งในการจัดการเรียนการสอนและการทดสอบความรู้ด้านปริยัติบาลีที่ควบคุมโดยสำนักงานแม่กองธรรมสนามหลวงได้เปิดโอกาสให้แม่ชีและสามเณราสสามารถเข้าเรียนและเข้าร่วมสอบได้นอกเหนือจากพระสงฆ์ รวมทั้งยังมีการจัดการเรียนการสอนด้านสายสามัญในมหาวิทยาลัยสงฆ์และมหาวิทยาลัยทั่วไปอีกด้วย ทำให้เห็นได้ว่าประเทศไทยให้ความสำคัญกับภาษาบาลีเป็นอย่างยิ่ง และไม่เพียงแต่ในประเทศไทยเท่านั้น กลุ่มประเทศที่นับถือพระพุทธศาสนานิกายเถรวาทก็ล้วนให้ความสำคัญกับภาษาบาลี เพราะพระธรรมและพระวินัยดั้งเดิมจากพุทธวจนะได้ถูกถ่ายทอดสืบต่อกันมาด้วยการท่องจำให้ขึ้นใจซึ่งเรียกว่าวิธีมุขปาฐะ และมีการจารึกเป็นลายลักษณ์อักษรด้วยภาษาบาลีเป็นครั้งแรกที่ประเทศศรีลังกา จากนั้นประเทศที่นับถือพระพุทธศาสนานิกายเถรวาทได้คัดลอกพระธรรมและพระวินัยด้วยอักษรในประเทศของตนสืบต่อมา ถึงแม้ว่าพระธรรมและพระวินัยได้รับการแปลเป็นภาษาของประเทศของตน แต่ก็ไม่มีการละทิ้งต้นฉบับภาษาบาลีไป ทำให้สามารถตรวจสอบและเทียบเคียงการแปลกับต้นฉบับได้เสมอ นอกจากนี้หากพระพุทธศาสนาในประเทศใดเสื่อมถอยลงจนเป็นเหตุให้คัมภีร์สำคัญทางพระพุทธศาสนาสูญหายหรือถูกทำลายไป ภาษาบาลีที่เป็นต้นฉบับนี้สามารถนำมาคัดลอกและส่งต่อได้อยู่เสมอ ดังจะเห็นได้จากวิกฤติการณ์ในอดีตที่เคยเกิดขึ้นในพุทธศาสนา [1] และการเผยแพร่พระพุทธศาสนาของพระธรรมทูตซึ่งใช้คำบาลี พร้อมทั้งคำแปลและคำอธิบายด้วยภาษาของประเทศเจ้าถิ่น เพื่อเผยแพร่และเสนอหลักธรรมคำสอนอันก่อให้เกิดประโยชน์เกิดประโยชน์เกื้อกูลแก่สังคม [2] ดังนั้นภาษาบาลีจึงมีความสำคัญอย่างยิ่งเพราะเป็นเครื่องมือที่ใช้รักษาคำสั่งสอนของพระพุทธเจ้า

เนื่องจากได้มีการจารึกพระไตรปิฎกเป็นภาษาบาลี ทำให้คุณค่าของภาษาบาลียิ่งเด่นชัด ดังจะเห็นได้จากหนังสือวรรณคดีบาลี [3] ได้แสดงความสำคัญและคุณค่าของการค้นคว้าพระไตรปิฎกว่ามีประโยชน์ทางวิชาการมากมายและเป็นแหล่งรวมความรู้ของศาสตร์หลายแขนง เช่น ศาสนศาสตร์ สังคมศาสตร์ ภาษาศาสตร์ และมานุษยวิทยา เป็นต้น เนื้อหาในพระไตรปิฎกและคำภีร์ที่ใช้อธิบายความในพระไตรปิฎกถูกเรียบเรียงด้วยภาษาบาลีมีทั้งรูปแบบร้อยกรองและร้อยแก้ว มีความ

งดงามของการใช้อักษร เสียง และจังหวะได้อย่างไพเราะสละสลวย ทั้งยังใช้ศัพท์กับความหมายได้อย่างสอดคล้องกลมกลืน แม้กระทั่งในภาษาไทยยังรับอิทธิพลทางภาษาของภาษาบาลีมาบางส่วน เช่นคำว่า “ภัยอันตราย” สามารถพูดให้ไพเราะเป็น ภยันตราย

การศึกษาบาลีเป็นการศึกษาขั้นพื้นฐานเพื่อนำไปสู่การศึกษาข้อมูลทางพระพุทธศาสนาหรือวิชาการที่ปรากฏในพระไตรปิฎกรวมถึงคัมภีร์ที่สำคัญทางพระพุทธศาสนาให้เข้าใจอย่างถ่องแท้ โดยเฉพาะอย่างยิ่งควรมีความเข้าใจในไวยากรณ์ภาษาบาลีเพราะคำนามในภาษาบาลีสามารถนำมาสร้างเป็นคำใหม่ได้จำนวนมาก และคำในภาษาบาลีทุกคำสามารถย้อนกลับไปหาความหมายที่แท้จริงได้เสมอ คำทับศัพท์จากภาษาบาลีที่ใช้กันอย่างคุ้นชิน เช่น คำว่าบุรุษ ซึ่งคำแปลคือผู้ชาย แต่ยังความหมายที่ซ่อนเร้นอยู่ในคำศัพท์นี้คือ “เป็นผู้ยังใจของบิดามารดาให้เอิบอาบ” คำว่าภริยา คำแปลคือภรรยา แต่ความหมายที่ซ่อนเร้นอยู่หมายถึง “ผู้ที่บุรุษพึงเลี้ยง” คำว่าโอรสมีความหมายซ่อนเร้นหมายถึงผู้ที่เกิดจากอก แม้กระทั่งคำว่าภิกษุยังมีความหมายซ่อนเร้นอยู่สองนัยด้วยกัน มีทั้งหมายถึงผู้เห็นภัยในสังสารวัฏ และสามารถแปลได้ว่าผู้เที่ยวไปเพื่อขอเป็นต้น คำศัพท์ทางภาษาบาลีนี่เองทำให้ผู้แปลสามารถเข้าใจวัฒนธรรมของสังคมในสมัยที่มีการใช้ภาษาบาลีเพื่อเผยแพร่พระพุทธศาสนา รวมไปถึงทำให้ผู้แปลสามารถเข้าใจความหมายหรืออารมณ์จากเนื้อความได้มากกว่าผู้อ่านคำแปล ในประเทศไทยมีหน่วยงานบริหารการทดสอบภาษาบาลีเพื่อวัดความรู้โดยเฉพาะ การศึกษาภาษาบาลีในประเทศไทยมีทั้งหมดเก้าระดับชั้น ใช้หนังสือธัมมปทัฏฐกถาเป็นแบบเรียนพื้นฐานสำหรับผู้แรกเริ่ม โดยทั่วไปแล้วนิยมท่องบาลีไวยากรณ์ในปีแรกให้จำขึ้นใจเสียก่อน หลังจากปีที่สองจึงเริ่มเรียนการแปลภาษาบาลี ถึงแม้ว่าผู้เริ่มเรียนสามารถจำไวยากรณ์ได้แล้ว แต่การเริ่มต้นเรียนแปลภาษาบาลีใช้เวลาชั่วโมงละหนึ่งถึงสองหน้า ทั้งนี้ขึ้นอยู่กับทักษะของผู้เรียนเป็นสำคัญ แต่เมื่อเริ่มมีทักษะหรือสอบผ่านระดับชั้นแรกไปแล้วสามารถแปลได้เร็วขึ้นเป็นเท่าตัว สาเหตุที่ทำให้ผู้เริ่มเรียนที่ท่องจำไวยากรณ์จนขึ้นใจใช้ระยะเวลาถึงเพียงนั้นเพราะไม่ทราบความหมายของศัพท์ แยกคำสนธิไม่ได้ และแปลความหมายของคำสมาสไม่ออก

คำสนธิเพียงคำเดียวสามารถแยกคำที่ถูกต้องตามไวยากรณ์ได้หลายแบบ นอกจากแยกคำสนธิได้อย่างถูกต้องตามไวยากรณ์แล้วยังต้องสัมพันธ์กับคำอื่นในประโยค รวมถึงสอดคล้องกับเนื้อความที่ผ่านมาหรือประโยคก่อนหน้าด้วย ไวยากรณ์ภาษาบาลีสามารถนำคำศัพท์พื้นมาต่อรวมกันให้เป็นคำศัพท์ใหม่ได้ ถ้าหากไม่ทราบศัพท์ใดเพียงศัพท์เดียวจะไม่สามารถแปลศัพท์ทั้งก้อนได้ ถึงแม้ในประเทศไทยมีอภิธานศัพท์ภาษาบาลีแต่ถ้าหากไม่ทราบไวยากรณ์การสนธิและการสมาสแล้วจะไม่สามารถแปลได้อย่างถูกต้อง เพราะคำสนธิและคำสมาสเป็นไวยากรณ์ที่มีลักษณะ

เป็นรูปแบบเปิด สามารถหาคำมาใช้ทดแทนได้เสมอ จึงเป็นไปได้ยากที่จะแสดงรายการของคำศัพท์ทั้งหมด

ในพุทธศักราช 2531 มหาวิทยาลัยมหิดลได้จัดทำพระไตรปิฎกภาษาบาลีอักษรฉบับคอมพิวเตอร์อยู่ในรูปแบบโปรแกรมคอมพิวเตอร์เรียกว่า BUDSIR (Buddhist Scriptures Information Retrieval) เพื่อใช้ในการค้นหาข้อมูล ซึ่งสามารถค้นหาคำ คำศัพท์ พุทธพจน์ และพุทธภาษิตได้อย่างครบถ้วนและสมบูรณ์ ทั้งสามารถแปลงเป็นอักษรอื่นได้แก่ ไทย โรมัน พม่า เขมร ล้านนา ลาว เทวนาครี และสิงหล ถือว่าเป็นครั้งแรกของประเทศไทยและของโลก ปัจจุบันมีรากศัพท์ประมาณ 17,675 คำ

นอกจากนี้ในประเทศไทยยังมีงานวิจัยเพื่อแปลภาษาบาลีเป็นภาษาไทย [4] โดยใช้พจนานุกรมของคำศัพท์จากโปรแกรม BUDSIR และกฎไวยากรณ์ภาษาบาลีซึ่งถูกจัดเก็บอยู่ในรูปไวยากรณ์ไม่พึ่งบริบทเพื่อให้สามารถแปลประโยคความเดียวและประโยคความซ้อนได้ เช่น การแปลงข้อความภาษาบาลีอักษรไทยเป็นสัทอักษร [5] โปรแกรมจำลองการอ่านภาษาบาลี [6] ระบบค้นคืนพระคาถาธรรมบทภาษาบาลีอักษรไทย [7] อย่างไรก็ตามจากการค้นคว้า ผู้วิจัยยังไม่พบงานวิจัยอื่นทางด้านการประมวลผลภาษาธรรมชาติที่เกี่ยวข้องกับการตัดคำภาษาบาลีอักษรไทย

ถึงแม้ภาษาบาลีเป็นภาษาที่ใช้การเว้นวรรคระหว่างคำ ทำให้สังเกตคำได้ง่าย การงานตัดคำในภาษาบาลีจะไม่สามารถตัดคำโดยการแยกคำจากกันโดยตรง เช่น คำว่าธนู กับ อาคม สนธิกันเป็นคำว่า ธันวาคม การตัดคำต้องตัดให้เป็นคำเดิมก่อนนำมาสนธิกัน และนอกจากแยกสนธิได้ถูกต้องแล้วยังต้องแสดงสัมพันธ์ในประโยคได้ด้วย

ดังนั้นผู้วิจัยจึงได้ศึกษาและการสร้างตัวแบบการตัดคำในภาษาบาลีอักษรไทยให้สามารถแยกคำสนธิหรือแบ่งคำสมาสได้ โดยคงความหมายของคำไว้แบบเหมือนเดิม รวมทั้งมุ่งเน้นตัดคำข้อความประเภทร้อยแก้วที่มาจากหนังสือแบบเรียนภาษาบาลีตามหลักสูตรบาลีสนามหลวง ซึ่งเป็นข้อความที่มีประโยคค่อนข้างสมบูรณ์สมบูรณ์และเหมาะสมต่อผู้เริ่มแปลภาษาบาลี เพื่อให้ผู้สนใจสามารถนำภาษาบาลีอักษรไทยที่ผ่านกระบวนการตัดคำไปต่อยอดและประยุกต์ใช้ในงานที่เหมาะสมได้

วิทยานิพนธ์นี้นำเสนอวิธีการตัดคำบาลีสนธิและคำบาลีสมาส โดยประยุกต์ใช้โครงข่ายประสาทเทียมแบบแอลเอสทีเอ็มแบบสองทิศทางมาใช้เพื่อทำนายตำแหน่งตัดคำและรูปแบบ จากนั้นนำคำสนธิบาลีคำสมาสที่ทราบตำแหน่งและรูปแบบมาแก้ไขคำโดยกฎที่ผู้วิจัยได้วิเคราะห์และถอดรูปแบบมาจากไวยากรณ์ภาษาบาลีไว้เพื่อให้สามารถแก้ไขคำให้ถูกต้องและมีความหมาย

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาวิธีการตัดคำสนธิหรือคำสมาสภาษาบาลีอักษรไทยด้วยวิธีผสมผสานด้วยการเรียนรู้เชิงลึกร่วมกับการใช้กฎ
2. พัฒนาตัวแบบที่ใช้ในการแยกคำสนธิหรือแบ่งคำสมาสให้สามารถทำงานได้อย่างถูกต้อง โดยสามารถวัดผลความถูกต้องและรายงานผลได้

1.3 ขอบเขตการวิจัย

1. ใช้ข้อความภาษาบาลีที่จากหนังสือธมฺมปทฏฺฐกถา เนื่องจากเป็นหนังสือที่ใช้ศึกษาของผู้เริ่มแปลภาษาบาลีเป็นไทยตามหลักสูตรบาลีสนามหลวง และโครงสร้างประโยคเรียบง่ายไม่ซับซ้อนและค่อนข้างสมบูรณ์
2. หนังสือธมฺมปทฏฺฐกถามีทั้งหมด 8 เล่ม มีเนื้อหารวมทั้งหมด 305 เรื่อง ในแต่ละเรื่องเลือกเฉพาะเนื้อความส่วนที่เป็นร้อยแก้ว ไม่รวมร้อยกรองและคำอธิบายศัพท์จากร้อยกรอง
3. ค่าความถูกต้องของการตัดคำสนธิหรือคำสมาสไม่ต่ำกว่า 80%
4. งานวิจัยนี้ตัดคำผสม 2 ประเภท ประกอบด้วย
   - คำสนธิ คือ คำสองคำที่มาเชื่อมกันเพื่อให้ออกเสียงได้อย่างไพเราะ เช่น

     ประโยคบาลี: อถสฺส ภริยาย กุจฺฉิยํ คพฺโภ ปติฏฺฐาสิ

     ประโยคแปล: ครั้งนั้น สัตว์ผู้เกิดในครรภ์ ตั้งอยู่แล้ว ในท้อง ของภรรยา ของเขา

     คำสนธิในประโยคนี้คือ อถสฺส แยกออกจากกันเป็น อถ ทำหน้าที่บอกเวลา กับ อสฺส ทำหน้าแสดงความเป็นเจ้าของ ของคำว่า ภริยา
   - คำสมาส คือ ศัพท์ตั้งแต่ 2 ศัพท์มาต่อกัน ก่อนประกอบวิภัตติให้กลายเป็นคำ เช่น

     ประโยคบาลี: สา ทสมาสจฺจเยน กาเลน ปุตฺตํ วิชายิ

     ประโยคแปล: นาง คลอดแล้ว ซึ่งบุตร (คนเดียว) โดยกาล อันล่วงไปแห่งเดือนสิบ

     คำสมาสในประโยคนี้คือ ทสมาสจฺจเยน ศัพท์เดิมก่อนประกอบวิภัตติคือ ทสมาสจฺจย แยกออกเป็น ทส มาส และ อจฺจย แปลว่า สิบ, เดือน, และอันล่วงไป, ทสมาส เป็นสมาสแปลว่า เดือนสิบ (ไม่ได้หมายถึงเดือนที่สิบ แต่หมายถึงสิบเดือน) และนำ ทสมาส ที่เป็นคำสมาสมาสมาสซ้ำกับคำว่า อจฺจย เป็นคำว่า ทสมาสจฺจย แปลว่า ล่วงไปแห่งเดือนสิบ (หมายถึงผ่านพ้นไปสิบเดือน)

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ตัวแบบการแยกคำสนธิและการแบ่งคำสมาสภาษาบาลีอักษรไทย ซึ่งยังไม่เคยมีในงานวิจัยด้านการประมวลผลภาษาธรรมชาติที่ประยุกต์ใช้กับภาษาบาลีอักษรไทยและโรมัน
2. เป็นเครื่องมือที่ใช้ต่อยอดในการประมวลผลภาษาธรรมชาติกับภาษาบาลีอักษรไทยให้มีประสิทธิภาพมากยิ่งขึ้น

บทที่ 2

ภาษาบาลีและวรรณกรรมที่เกี่ยวข้อง

ภาษาบาลีเป็นภาษาที่มีความเก่าแก่ภาษาหนึ่ง มีความสำคัญเป็นอย่างมากต่อพุทธศาสนิกชนนิกายเถรวาทเพราะเป็นภาษาที่ใช้บันทึกคัมภีร์ที่สำคัญในพระพุทธศาสนา เช่น พระไตรปิฎกและคัมภีร์อรรถกถาเป็นต้น นักวิชาการคาดว่าภาษาบาลีเป็นภาษาพูดจึงไม่มีอักษรใช้เฉพาะ เมื่อมีการเผยแพร่พระพุทธศาสนาไปถึงที่ใดมักใช้อักษรของถิ่นนั้นโดยใช้เครื่องหมายเพิ่มเติม เช่นการเขียนภาษาบาลีในประเทศไทยใช้อักษรไทยโดยปรับเพิ่มหรือลดเครื่องหมายได้แก่ การใช้พินทุ นิคหิต และลบเชิงใต้ ญ และ ฐ ปัจจุบันได้ถือว่าเป็นภาษาที่ตายแล้วเพราะไม่มีการพูดหรือสื่อสารกันในชีวิตประจำวัน เพียงแต่มีการศึกษาในประเทศที่นับถือพระพุทธศาสนานิกายเถรวาทเท่านั้น

ในประเทศไทยมีการศึกษาภาษาบาลีมาตั้งแต่สมัยสุโขทัย นอกจากการศึกษาภาษาบาลีของคณะสงฆ์ที่อยู่ภายใต้การควบคุมดูแลของสำนักงานแม่กองบาลีสนามหลวงแล้วยังมีการจัดการเรียนการสอนที่มหาวิทยาสงฆ์ทั้งสองแห่งของประเทศไทย นอกจากนี้สถาบันการทดสอบการศึกษาแห่งชาติยังจัดสอบความถนัดทางทางภาษาบาลีอีกด้วย

2.1 การอ่านและออกเสียงในภาษาบาลี

อักษรไทยที่ใช้ในภาษาบาลีมี 41 [8, 9] ตัวแบ่งเป็นสระ 8 ตัวคือ อ อา อิ อี อุ อู เอ โอ และพยัญชนะมี 33 ตัว แบ่งกลุ่มตามแหล่งการเกิดเสียงเรียกว่าวรรค ๆ ละ 5 ตัว มีทั้งหมด 5 วรรค รวมกับพยัญชนะที่ไม่สามารถจัดเข้ากลุ่มได้อีก 8 ตัวเรียกว่าอวรรค ดังแสดงในตารางที่ 1 จะเห็นว่ามีการใช้พินทุหรือจุดใต้ตัวพยัญชนะเพื่อให้ทราบว่าไม่มีสระอาศัย ไม่สามารถออกเสียงได้

ตารางที่ 1 หน่วยเสียงพยัญชนะในภาษาบาลี

วรรค ก	กฺ k	ขฺ kh	คฺ g	ฆฺ gh	งฺ ṅ
วรรค จ	จฺ c	ฉฺ ch	ชฺ j	ฌฺ jh	ญฺ ñ
วรรค ฏ	ฏฺ ṭ	ฐฺ ṭh	ฑฺ ḍ	ฒฺ ḍh	ณฺ ṇ
วรรค ต	ตฺ t	ถฺ th	ทฺ d	ธฺ dh	นฺ n
วรรค ป	ปฺ p	ผฺ ph	พฺ b	ภฺ bh	มฺ m
อวรรค	ย (y) ร (r) ล (l) ว (v) ส (s) ห (h) ฬ (ḷ) อ (ṃ)				

เมื่อนำพยัญชนะผสมกับสระจะทำให้พินทุใต้พยัญชนะหายไปและปรากฏรูปสระแทน ยกเว้นสระอะ จะไม่มีรูปสระอะปรากฏให้เห็น เช่นคำว่า **วร** อ่านว่า วะ-ระ และ **ปุริส** อ่านว่า ปุ-ริ-สะ เป็นต้น

ส่วนเสียงสระในภาษาบาลีมี 11 [8, 9] เสียงแบ่งเป็นสระเสียงทั่วไป 8 เสียง และสระเสียงนาสิก 3 เสียงคือ อำ (อ่านว่า อัง) อิ ๎ (อ่านว่า อิง เพื่อความสะดวกในการพิมพ์นิยมใช้สระอึ แทนการใช้สระอิและนิคหิต จึงกลายเป็น อึ) และอุ ๎ (อ่านว่า อุง) ดังแสดงในตารางที่ 2

ตารางที่ 2 เสียงสระในภาษาบาลี

เสียงสระ			ตัวอย่างคำบาลี		
ไทย	โรมัน	สากล	อักษรไทย	อักษรโรมัน	สัทอักษรสากล
อะ	a	a	วร	vara	ʋaɻa
อา	ā	aː	วาร	vāra	ʋaːɻa
อิ	i	i	อิติ	iti	iʈi
อี	ī	iː	อีติ	īti	iːʈi
อุ	u	u	กุล	kula	kuɭa
อู	ū	uː	กูล	kūla	kuːɭa
เอ	e	eː	เย	ye	jeː
โอ	o	oː	โย	yo	joː
อำ	aṃ	ã	กำ	kaṃ	kã
อิ ๎	iṃ	ĩ	กิ ๎	kiṃ	kĩ
อุ ๎	uṃ	ũ	กุ ๎	kuṃ	kũ

เมื่อพบพยัญชนะต้นคำอยู่รวมกับสระแล้วตามด้วยพยัญชนะที่มีพินทุต่อท้าย พยัญชนะที่มีพินทุต่อนั้นจะออกเสียงเป็นเสียงกัก (Stop consonants) หรือตัวสะกดในภาษาไทย เช่น ทุกฺขํ, โสตฺถิ, นารึ, กาตุ ๎ ดังแสดงในตารางที่ 3

ตารางที่ 3 เสียงพยัญชนะท้าย

คำบาลี	พยางค์	พยัญชนะต้น	สระ	พยัญชนะท้าย	คำอ่านไทย
ทุกฺขํ	ทุกฺ	ทฺ	สระอุ	กฺ	ทุก
	ขํ	ขฺ	สระอะ	อ ◌ं (นิคหิต)	ขัง
โสตฺถิ	โสตฺ	สฺ	สระโอ	ตฺ	โสด
	ถิ	ถฺ	สระอิ	-	ถิ
นารึ	นา	นฺ	สระอา	-	นา
	รึ	รฺ	สระอิ	อ ◌ं (นิคหิต)	ริง
กาตุ	กา	กฺ	สระอา	-	กา
	ตุ	ตฺ	สระอุ	อ ◌ं (นิคหิต)	ตุง

นอกจากนี้ในภาษาบาลียังปรากฏลักษณะการออกเสียงพยัญชนะควบกล้ำ ซึ่งอาจปรากฏทั้งที่พยัญชนะต้นหรือพยัญชนะท้าย ดังแสดงตัวอย่างในตารางที่ 4 แต่ถ้าเสียงพยัญชนะสะกดควบกล้ำเมื่ออ่านจะทำหน้าที่เป็นทั้งตัวสะกดที่พยางค์แรกและเป็นพยัญชนะต้นของพยางค์ถัดไปดังแสดงตัวอย่างตารางที่ 5

ตารางที่ 4 เสียงพยัญชนะต้นควบกล้ำ

คำบาลี	พยางค์	พยัญชนะต้น	สระ	พยัญชนะท้าย	คำอ่านไทย
พฺยคฺฆ	พฺยคฺ	พฺยฺ	สระอะ	คฺ	พยัก (ไม่ใช่ พะ-ยัก)
	ฆ	ฆฺ	สระอะ	-	ฆะ
ทฺวาร	ทฺวา	ทฺวฺ	สระอา	-	ทวา (ไม่ใช่ ทะ-วา)
	ร	รฺ	สระอะ	-	ระ

ตารางที่ 5 เสียงพยัญชนะท้ายควบกล้ำ

คำบาลี	พยางค์	พยัญชนะต้น	สระ	พยัญชนะท้าย	คำอ่านไทย
กตฺวา	กตฺว	กฺ	สระอะ	ตฺวฺ	กัตว (ไม่ใช่ กัด-ตะ-วะ)
	ตฺวา	ตฺวฺ	สระอา		ตวา (ไม่ใช่ ตะ-วา)
กลฺยา	กฺ	กฺ	สระอะ	ลฺยฺ	กัลย (ไม่ใช่ กัน-ละ-ยะ)
	ลฺยา	ลฺยฺ	สระอา	-	ลยา (ไม่ใช่ ละ-ยา)

ถึงแม้ในประเทศไทยจะมีให้ความสำคัญกับการเรียนการสอนภาษาบาลีของคณะสงฆ์มาก แต่ปัจจุบันไม่ได้เคร่งครัดเรื่องการออกเสียงให้ถูกต้องมากนัก เช่นคำว่า กตฺวา และ กลฺยา ถึงแม้ไม่ได้อ่านว่า กัด-ตะ-วา และ กัล-ละ-ยา แต่ยังคงอ่านเช่นนั้นเพราะความคุ้นเคย

2.2 วจีวิภาค: ประเภท หน้าที่ และความหมายของคำ

ประเภทของคำสามารถแบ่งออกเป็น 3 ประเภท ได้แก่ คำนาม คำกริยา และอัพยยศัพท์ ในภาษาบาลีไม่มีบุพบท แต่ใช้รูปของพยางค์ท้ายคำแสดงความหมายของคำ คำนามและคำกิริยาสามารถเปลี่ยนรูปได้หลายแบบ ทั้งนี้คำนามแต่รูปอาจมีหลายหน้าที่ เช่น ปุริสานํ สามารถทำหน้าที่เป็นกรรมรองและสามารถแสดงความเป็นเจ้าของได้ การหาความหมายที่แท้จริงของคำว่า ปุริสานํ นี้จึงต้องพิจารณาว่าคำนี้มีความสัมพันธ์กับคำใด เช่น ถ้ามีความสัมพันธ์กับคำกริยาจึงทำหน้าที่เป็นกรรมรอง แต่ถ้าสัมพันธ์กับคำนามจะทำหน้าที่แสดงความเป็นเจ้าของ ส่วนอัพพยศัพท์เป็นศัพท์กลุ่มเดียวที่ไม่เปลี่ยนรูป

2.2.1 คำนาม (Noun) [8]

คำนามมี 3 ประเภทได้แก่นามนาม คุณนาม และสัพพนาม คำนามแต่ละประเภทสามารถแบ่งย่อยได้อีกดังแสดงไว้ในรูปที่ 1 นามนามเป็นคำนามที่ใช้เรียกชื่อ คน สัตว์ ที่ และสิ่งของ มีทั้งใช้เรียกแบบเจาะจงเรียกว่าสาธารณนาม และใช้เรียกแบบไม่เจาะจงเรียกว่าอสาธารณนาม ส่วนคุณนามเป็นคำนามที่ใช้แสดงลักษณะของนามนามมี 3 ระดับคือปกติ วิเสส และอติวิเสส เช่น สูง สูงกว่าและสูงที่สุด ตามลำดับ ส่วนคำนามประเภทสุดท้ายคือสัพพนาม เป็นคำนามที่ใช้เรียกชื่อนามนามที่เคยกล่าวถึงแล้วมี 2 ประเภทได้แก่ปุริสสัพพนาม และวิเสสนสัพพนาม

ปุริสสัพพนาม (Pronoun) ตรงกับบุรุษสรรพนามในภาษาไทยใช้กล่าวถึงบุคคลที่สามหรือผู้อื่นเรียกว่าปฐมบุรุษ กล่าวถึงผู้ฟังเรียกว่ามัธยมบุรุษ และกล่าวแทนผู้พูดเรียกว่าอุตตมบุรุษ ส่วนวิเสสนสัพพนามถูกจัดเป็น 2 กลุ่ม

1. บอกความใกล้หรือไกลตรงกับนิยมสรรพนามในภาษาไทย (Demonstrative), เพียงแต่มี 4 คำศัพท์ (นั้น, นี่, นั่น, โน้น)

2. บอกความไม่แน่นอน ไม่เฉพาะเจาะจง (Indefinite) และคำสัพพนามที่เป็นประธานซึ่งใช้เป็นคำถาม (Interrogative) ตรงกับอนิยมสรรพนามและปฤจฉาสรรพนามตามลำดับ มี 13 คำศัพท์

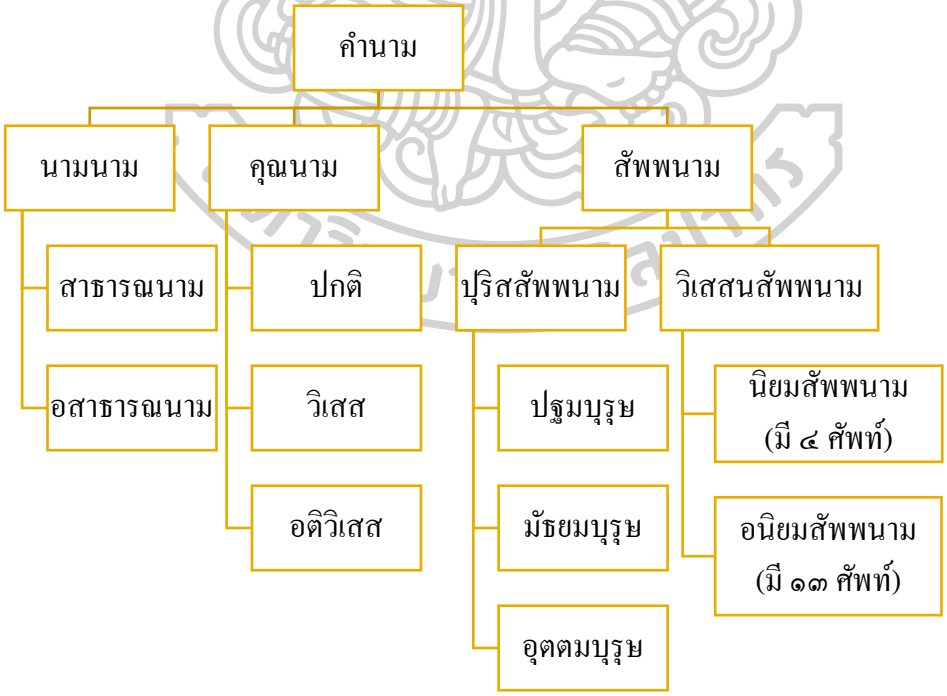
2.2.2 คำกิริยา (Verb)

คำกิริยามี 2 ประเภทคือกิริยาอาขยาตเป็นกิริยากิตก์ กิริยาอาขยาตทำหน้าที่เป็นกิริยาหลักหรือกิริยาแท้ของประโยค ส่วนกิริยากิตก์ทำหน้าที่เป็นคุณนามหรือกิริยาย่อย ในบางครั้งกิริยากิตก์ทำหน้าที่เป็นกิริยาแท้เมื่อไม่มีกิริยาอาขยาต

2.2.3 อัพยยศัพท์ (Prefix, Suffix and Particle)

คำนามและคำกิริยาในภาษาบาลีใช้รูปของคำแสดงความหมาย เช่น ต้นเค้าศัพท์หมายถึง "**เด็ก**" สามารถทราบได้จากคำนามที่ปรากฏในประโยคว่าหมายถึงเด็กคนเดียวและเด็กหลายคน ในทำนองเดียวกันรากศัพท์ที่หมายถึง "**ไป**" สามารถทราบได้จากคำกิริยาที่ปรากฏว่าหมายถึงไปคนเดียวหรือไปหลายคน, ไปในอดีต ปัจจุบัน หรืออนาคต และไปด้วยตนเองหรือถูกผู้อื่นนำไป แต่ในภาษาบาลีมีกลุ่มหน่วยศัพท์ชนิดหนึ่งที่ไม่เปลี่ยนรูปคำ หน่วยศัพท์นี้เรียกว่าอัพยยศัพท์ มี 3 ประเภท ได้แก่

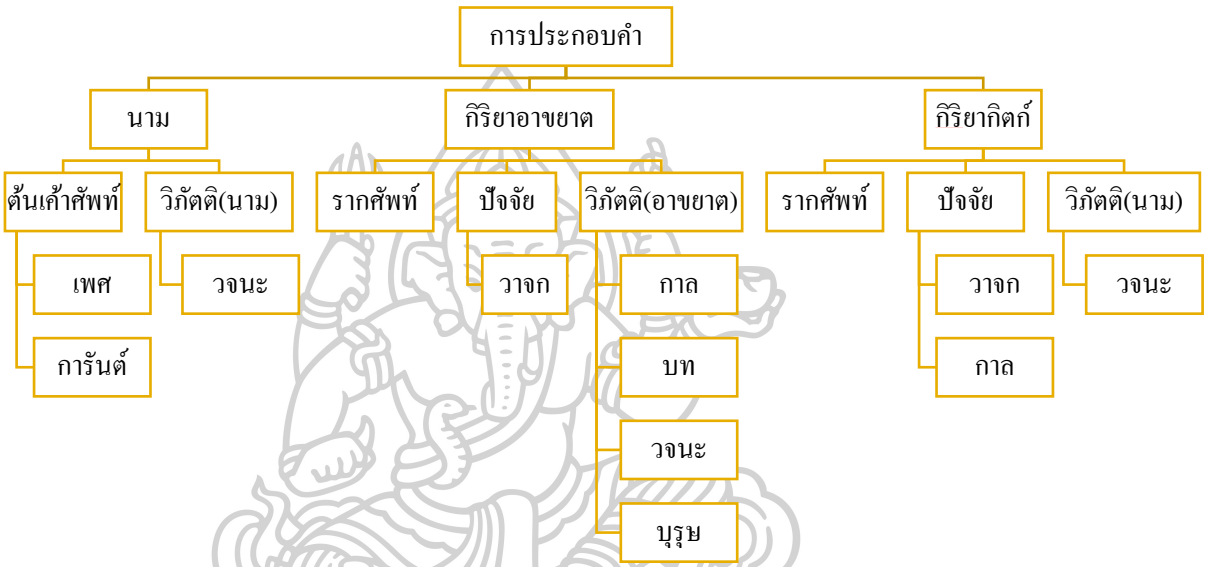
1. อุปสัค เป็นคำศัพท์ที่ใช้นำคำนามหรือคำกิริยา ทำให้มีความชัดเจนหรือเปลี่ยนแปลงไป
2. นิบาต เป็นคำศัพท์ที่ใช้เชื่อมความ ใช้ร้องเรียก ใช้บอกเวลาเป็นต้น
3. ปัจจัย เป็นคำศัพท์ที่ใช้ต่อท้ายคำนามหรือคำกิริยา



รูปที่ 1 ประเภทของคำนาม

2.2.4 การประกอบคำ [8]

การประกอบคำ คือการประกอบหน่วยศัพท์ ให้เป็นหน่วยคำที่มีความหมายในประโยค กล่าวคือคำนาม วิทยานิพนธ์เล่มนี้เรียกหน่วยศัพท์ที่ใช้ประกอบเป็นคำนามว่าต้นเค้าศัพท์ และเรียกหน่วยศัพท์ที่ใช้ประกอบเป็นคำกิริยาว่ารากศัพท์ และหน่วยศัพท์อื่น ๆ เรียกตามชนิดเช่นนิบาตศัพท์ ภาพรวมของการประกอบคำสามารถสรุปรวมไว้ในรูปที่ 2 ยกตัวอย่างอธิบายตามลำดับได้ดังนี้



รูปที่ 2 ส่วนประกอบของคำ

2.2.4.1 การประกอบคำนาม

คำนามเกิดขึ้นต้นเค้าศัพท์และวิภัตติ แล้วแปลงรูป เช่นต้นเค้าศัพท์คือ **ปุริส** (บุรุษ) รวมกับวิภัตติแล้วแปลงรูป สามารถอธิบายเป็นขั้นตอนได้ดังนี้

1. ต้นเค้าศัพท์คือ **ปุริส** (บุรุษ) เป็นคำศัพท์เพศชาย มีสระท้ายศัพท์ (การันต์) คือ **อ**

2. วิภัตติที่ทำให้คำทำหน้าที่เป็นประธานคือ ปฐมาวิภัตติ ฝ่ายเอกวจนะคือ **สิ** รวมกันจะได้

ปุริส + สิ

3. แปลง **อ** ที่ท้ายศัพท์ของ **ปุริส** กับ **สิ** วิภัตติ เป็น **โอ** (ปุริส เปลี่ยนเป็น ปุริสฺ เพราะไม่มีสระอาศัย) จะได้

ปุริส + สิ = ปุริสฺ + อ + สิ = ปุริสฺ + โอ

4. รวม **ปุริสฺ + โอ** เป็น **ปุริโส**

ต้นเค้าศัพท์ในภาษาบาลีสามารถระบุเพศของคำศัพท์ (Gender) ได้ มี 3 เพศ [8, 9] คือ คำศัพท์เพศชาย (Masculine Noun) คำศัพท์เพศหญิง (Feminine Noun) และคำศัพท์ที่ไม่ใช่ทั้งเพศหญิงและเพศชาย(Neuter Nouns) ก่อนแปลงสระท้ายศัพท์กับวิภัตติให้เป็นคำ ต้องทราบเพศและการันต์ของต้นเค้าศัพท์ก่อน

การันต์หมายถึงสระท้ายศัพท์ [8] เช่น สระท้ายศัพท์ของ **ปุริส** คือสระอะ หรือสามารถกล่าวได้ว่า **ปุริสศัพท์** เป็นอการันต์ เพศชายมีการันต์ 5 ตัวคือ อ อิ อี อุ และอูตามลำดับ เพศหญิงมีการันต์ 5 ตัวคือ อา อิ อี อุ และอูตามลำดับ ส่วนต้นเค้าศัพท์ที่ไม่ใช่เพศหญิงและชายมีการันต์ 3 ตัว คือ อ อิ และอุ

ต้นเค้าศัพท์บางศัพท์มีเพศเดียว ในตารางที่ 6 แสดงคำนามที่มีต้นเค้าศัพท์ระบุชัดเพศเดียว บางศัพท์มีสองเพศ บางครั้งต้นเค้าศัพท์เหมือนกันสามารถมีได้ 2 เพศ แต่ใช้สระท้ายศัพท์แตกต่างกันดังตัวอย่างในตารางที่ 7 และคุณนามสามารถเป็นได้ 3 เพศดังแสดงในตารางที่ 8

ตารางที่ 6 นามนามซึ่งมีต้นเค้าศัพท์เป็นเพศเดียว

เพศชาย		**เพศหญิง**		**มิใช่เพศชายและเพศหญิง**	
คำนาม	**คำแปล**	**คำนาม**	**คำแปล**	**คำนาม**	**คำแปล**
อมโร	เทวดา	อัจฉรา	นางอัปสร	องฺคำ	องค์
อาทิจฺโจ	พระอาทิตย์	อาภา	รัศมี	อารมฺมณำ	อารมณ์
อินฺโท	พระอินทร์	อิทฺธิ	ฤทธิ์	ปณฺณำ	ใบไม้
ปพฺพโต	ภูเขา	ปภา	รัศมี	จกฺขำ	นัยน์ตา

ตารางที่ 7 นามนามที่มีต้นเค้าศัพท์เหมือนกัน เป็นได้ทั้ง 2 เพศ

ปุงลิงค์	**อิตถีลิงค์**	**คำแปล**
อรหา หรือ อรหำ	อรหนฺตี	พระอรหันต์
อุปาสโก	อุปาสิกา	อุบาสก, อุบาสิกา
เสฏฺฐี	เสฏฺฐินี	เศรษฐี
ยกฺโข	ยกฺขินี	ยักษ์

ตารางที่ 8 คุณนามเป็นได้ทั้ง 3 เพศ

ปุงลิงค์	อิตถีลิงค์	นปุงสกลิงค์	คำแปล
นาโถ	นาถา	นาถำ	ที่พึ่ง
เชฏฺโฐ	เชฏฺฐา	เชฏฺฐำ	เจริญที่สุด
สทฺโธ	สทฺธา	สทฺธำ	มีศรัทธา
ธมฺมิโก	ธมฺมิกา	ธมฺมิกำ	ตั้งในธรรม

วิภัตติในภาษาบาลี มี 14 ตัว แบ่งเป็นเอกวจนะ 7 และพหุวจนะ 7 ส่วนอาลปนนั้นไม่ถูกจัดว่าเป็นวิภัตติจึงไม่นับรวมด้วย วิภัตติในภาษาบาลีเป็นเครื่องมือสำคัญที่ช่วยให้ทราบความหมายของคำในประโยค ในตารางที่ 9 แสดงวิภัตติของภาษาบาลีพร้อมหน้าที่ของวิภัตติแต่ละตัว ส่วนตารางที่ 10 แสดงคำแปลที่ใช้แปลความหมายของคำ

ตารางที่ 9 วิภัตตินามและหน้าที่ของคำ

วิภัตติ	เอกพจน์	พหุพจน์	หน้าที่ในประโยค
ปฐมา	สิ	โย	ประธาน
ทุติยา	อำ	โย	กรรมตรง
ตติยา	นา	หิ	เครื่องมือในการกระทำ
จตุตถี	ส	นำ	กรรมรอง
ปัญจมี	สฺมา	หิ	แหล่งหรือแดนเกิด
ฉัฏฐี	ส	นำ	แสดงความเป็นเจ้าของ
สัตตมี	สฺมึ	สุ	บอกสถานที่หรือตำแหน่ง
อาลปนะ	สิ	โย	คำเรียก

ตารางที่ 10 ความหมายบอกเนื้อความของวิภัตตินามในภาษาบาลี [8]

วิภัตติ	เอกพจน์	พหุพจน์
ปฐมา	อันว่า... (ใช้ย่อว่า อ.)	อันว่า...ทั้งหลาย (คำว่าทั้งหลาย ใช้คำย่อว่า ท.)
ทุติยา	ซึ่ง, สู่, ยัง, สิ้น, กะ, ตลอด, เฉพาะ	ซึ่ง...ทั้งหลาย, สู่...ทั้งหลาย, ยัง...ทั้งหลาย, สิ้น...ทั้งหลาย, กะ...ทั้งหลาย, ตลอด...ทั้งหลาย, เฉพาะ...ทั้งหลาย
ตติยา	ด้วย, โดย, อัน, ตาม, เพราะ, มี	ด้วย...ทั้งหลาย, โดย...ทั้งหลาย, อัน...ทั้งหลาย, ตาม...ทั้งหลาย, เพราะ...ทั้งหลาย, มี...ทั้งหลาย
จตุตถี	แก่, เพื่อ, ต่อ	แก่...ทั้งหลาย, เพื่อ...ทั้งหลาย, ต่อ...ทั้งหลาย

วิภัตติ	เอกพจน์	พหุพจน์
ปัญจมี	แต่, จาก, กว่า, เหตุ	แต่...ทั้งหลาย, จาก...ทั้งหลาย, กว่า...ทั้งหลาย, เหตุ...ทั้งหลาย
ฉัฏฐี	แห่ง, ของ, เมื่อ	แห่ง...ทั้งหลาย, ของ...ทั้งหลาย, เมื่อ...ทั้งหลาย
สัตตมี	ใน, ใกล้, ที่, ครั้นเมื่อ, ใน เพราะ, เหนือ, บน, ณ	ใน...ทั้งหลาย, ใกล้...ทั้งหลาย, ที่...ทั้งหลาย, ครั้นเมื่อ...ทั้งหลาย, ใน เพราะ...ทั้งหลาย, เหนือ...ทั้งหลาย, บน...ทั้งหลาย, ณ...ทั้งหลาย
อาลปนะ	แน่ะ, ดูก่อน, ข้าแต่	แน่ะ...ทั้งหลาย, ดูก่อน...ทั้งหลาย, ข้าแต่...ทั้งหลาย

เมื่อทราบเพศและสระท้ายศัพท์ของต้นเค้าศัพท์แล้ว และทราบว่าต้องการประกอบต้นเค้าศัพท์ให้ทำหน้าที่ใด จึงจะสามารถแปลงสระท้ายศัพท์กับวิภัตตินามได้ถูกต้องตามไวยากรณ์ เช่น แปลง**อการันต์ของต้นเค้าศัพท์เพศชาย** กับ **สิ วิภัตติ** ซึ่งทำหน้าที่เป็นประธานในประโยค เป็น **โอ** แต่ **อการันต์ในต้นเค้าศัพท์ที่ไม่ใช่เพศหญิงและชาย** กับ **สิ วิภัตติ** ให้แปลง **สิ** เป็น **อำ**

ในตารางที่ 11 - ตารางที่ 16 แสดงรูปสำเร็จที่เกิดจากการแปลงสระท้ายศัพท์กับวิภัตติตามเพศของต้นเค้าศัพท์พร้อมทั้งยกตัวอย่าง ในวิทยานิพนธ์เล่มนี้ขอละคำอธิบายวิธีการแปลงไว้เพียงแต่แสดงรูปสำเร็จไว้เท่านั้น กรณีรูปสำเร็จที่ผ่านการแปลงแล้วมีให้เลือกใช้หลายคำ เช่นตารางที่ 10 จะพบว่าในจตุตถีวิภัตติฝ่ายเอกวจนะของอการันต์มีรูปสำเร็จเป็น **อสฺส อาย** และ**อตฺถ์** เมื่อนำไปรวมกับต้นเค้าศัพท์ เช่น **ปุริส** จะกลายเป็น **ปุริสสฺส ปุริสาย** และ**ปุริสตฺถ์** ตามลำดับ ดังนั้นเมื่อต้องการประกอบคำนามจากต้นเค้าศัพท์ที่เป็นเพศชายและมีอการันต์เป็นกรรมรองให้เลือกมาเพียงคำใดคำหนึ่งเท่านั้น

ตารางที่ 11 การเปลี่ยนแปลงท้ายศัพท์ของคำที่มีต้นเค้าศัพท์เป็นเพศชาย

วิภัตติ	อการันต์		อิการันต์		อีการันต์		อุการันต์		อูการันต์	
	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ
ปฐมา	โอ	อา	อิ	อนโย อี	อี	อิโน อี	อุ	อโว อู	อู	อุโน อู
ทุติยา	อำ	เอ	อึ	อนโย อี	อึ อินำ	อิโน อี	อุำ	อโว อู	อูำ	อุโน อู
ตติยา	เอน	เอหิ เอภิ	อินา	อีหิ อีภิ	อินา	อีหิ อีภิ	อุนา	อูหิ อูภิ	อูนา	อูหิ อูภิ
จตุตถี	อสฺส อาย อตฺถำ	อานำ	อิสฺส อิโน	อีนำ	อิสฺส อิโน	อีนำ	อุสฺส อุโน	อูนำ	อูสฺส อูโน	อูนำ
ปัญจมี	อสฺมา อมฺหา อา	เอหิ เอภิ	อิสฺมา อิมฺหา	อีหิ อีภิ	อิสฺมา อิมฺหา	อีหิ อีภิ	อุสฺมา อุมฺหา	อูหิ อูภิ	อูสฺมา อูมฺหา	อูหิ อูภิ
ฉัฏฐี	อสฺส	อานำ	อิสฺส อิโน	อีนำ	อิสฺส อิโน	อีนำ	อุสฺส อุโน	อูนำ	อูสฺส อูโน	อูนำ
สัตตมี	อสฺมึ อมฺหิ เอ	เอสุ	อิสฺมึ อมฺหิ	อีสุ	อิสฺมึ อมฺหิ	อีสุ	อุสฺมึ อุมฺหิ	อูสุ	อูสฺมึ อูมฺหิ	อูสุ
อาลปนะ	อ	อา	อิ	อนโย อี	อิ	อิโน อี	อุ	อเว อโว	อุ	อุโน อู

ตารางที่ 12 ตัวอย่างการผันคำของต้นเค้าศัพท์ที่เป็นเพศชาย

วิภัตติ	ปุริส (บุรุษ)		มุนิ (มุนี)		เสฏฺฐี (เศรษฐี)		ครุ (ครู)		วิญฺญู (ผู้รู้)	
	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ
ปฐมา	ปุริโส	ปุริสา	มุนิ	มุนโย มุนี	เสฏฺฐี	เสฏฺฐิโน เสฏฺฐี	ครุ	ครโว ครู	วิญฺญู	วิญฺญุโน วิญฺญู
ทุติยา	ปุริสำ	ปุริเส	มุนึ	มุนโย มุนี	เสฏฺฐึ เสฏฺฐินำ	เสฏฺฐิโน เสฏฺฐี	ครุำ	ครโว ครู	วิญฺญุำ	วิญฺญุโน วิญฺญู
ตติยา	ปุริเสน	ปุริเสหิ ปุริเสภิ	มุนินา	มุนีหิ มุนีภิ	เสฏฺฐินา	เสฏฺฐีหิ เสฏฺฐีภิ	ครุนา	ครูหิ ครูภิ	วิญฺญุนา	วิญฺญูหิ วิญฺญูภิ
จตุตถี	ปุริสสฺส ปุริสาย ปุริสตฺถำ	ปุริสานำ	มุนิสฺส มุนิโน	มุนีนำ	เสฏฺฐิสฺส เสฏฺฐิโน	เสฏฺฐีนำ	ครุสฺส ครุโน	ครูนำ	วิญฺญุสฺส วิญฺญุโน	วิญฺญูนำ
ปัญจมี	ปุริสสฺมา ปุริสมฺหา ปุริสา	ปุริเสหิ ปุริเสภิ	มุนิสฺมา มุนิมฺหา	มุนีหิ มุนีภิ	เสฏฺฐิสฺมา เสฏฺฐิมฺหา	เสฏฺฐีหิ เสฏฺฐีภิ	ครุสฺมา ครุมฺหา	ครูหิ ครูภิ	วิญฺญุสฺมา วิญฺญุมฺหา	วิญฺญูหิ วิญฺญูภิ
ฉัฏฐี	ปุริสสฺส	ปุริสานำ	มุนิสฺส มุนิโน	มุนีนำ	เสฏฺฐิสฺส เสฏฺฐิโน	เสฏฺฐีนำ	ครุสฺส ครุโน	ครูนำ	วิญฺญุสฺส วิญฺญุโน	วิญฺญูนำ
สัตตมี	ปุริสสฺมึ ปุริสมฺหิ ปุริเส	ปุริเสสุ	มุนิสฺมึ มุนิมฺหิ	มุนีสุ	เสฏฺฐิสฺมึ เสฏฺฐิมฺหิ	เสฏฺฐีสุ	ครุสฺมึ ครุมฺหิ	ครูสุ	วิญฺญุสฺมึ วิญฺญุมฺหิ	วิญฺญูสุ
อาลปนะ	ปุริส	ปุริสา	มุนิ	มุนโย มุนี	เสฏฺฐิ	เสฏฺฐิโน เสฏฺฐี	ครุ	ครเว ครโว	วิญฺญุ	วิญฺญุโน วิญฺญู

ตารางที่ 13 การเปลี่ยนแปลงท้ายศัพท์ของคำที่มีต้นเค้าศัพท์เป็นเพศหญิง

วิภัตติ	อาการันต์		อิการันต์		อีการันต์		อุการันต์		อูการันต์	
	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ
ปฐมา	อา	อาโย อา	อิ	อิโย อี	อี	อิโย อี	อุ	อุโย อู	อู	อุโย อู
ทุติยา	อำ	อาโย อา	อึ	อิโย อี	อึ อิยำ	อิโย อี	อุำ	อุโย อู	อูำ	อุโย อู
ตติยา	อาย	อาหิ อาภิ	อิยา	อีหิ อีภิ	อิยา	อีหิ อีภิ	อุยา	อูหิ อูภิ	อุยา	อูหิ อูภิ
จตุตถี	อาย	อานำ	อิยา	อีนำ	อิยา	อีนำ	อุยา	อูนำ	อุยา	อูนำ
ปัญจมี	อาย	อาหิ อาภิ	อิยา	อีหิ อีภิ	อิยา	อีหิ อีภิ	อุยา	อูหิ อูภิ	อุยา	อูหิ อูภิ
ฉัฏฐี	อาย	อานำ	อิยา	อีนำ	อิยา	อีนำ	อุยา	อูนำ	อุยา	อูนำ
สัตตมี	อาย อายำ	อาสุ	อิยา อิยำ	อีสุ	อิยา อิยำ	อีสุ	อุยา อุยำ	อูสุ	อุยา อุยำ	อูสุ
อาลปนะ	เอ	อาโย อา	อิ	อิโย อี	อิ	อิโย อี	อุ	อุโย อู	อุ	อุโย อู

ตารางที่ 14 ตัวอย่างการผันคำของต้นเค้าศัพท์ที่เป็นเพศหญิง

วิภัตติ	อาการันต์		อิการันต์		อีการันต์		อุการันต์		อูการันต์	
	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ
ปฐมา	กญฺญา	กญฺญาโย กญฺญา	รตฺติ	รตฺติโย รตฺตี	นารี	นาริโย นารี	รชฺชุ	รชฺชุโย รชฺชู	วธู	วธุโย วธู
ทุติยา	กญฺญำ	กญฺญาโย กญฺญา	รตฺตึ	รตฺติโย รตฺตี	นารึ นาริ โย	นาริโย นารี	รชฺชุำ	รชฺชุโย รชฺชู	วธูำ	วธุโย วธู
ตติยา	กญฺญาย	กญฺญาหิ กญฺญาภิ	รตฺติยา	รตฺตีหิ รตฺตีภิ	นาริยา	นารีหิ นารีภิ	รชฺชุยา	รชฺชูหิ รชฺชูภิ	วธุยา	วธูหิ วธูภิ
จตุตถี	กญฺญาย	กญฺญานำ	รตฺติยา	รตฺตีนำ	นาริยา	นารีนำ	รชฺชุยา	รชฺชูนำ	วธุยา	วธูนำ
ปัญจมี	กญฺญาย	กญฺญาหิ กญฺญาภิ	รตฺติยา รตฺยา	รตฺตีหิ รตฺตีภิ	นาริยา	นารีหิ นารีภิ	รชฺชุยา	รชฺชูหิ รชฺชูภิ	วธุยา	วธูหิ วธูภิ
ฉัฏฐี	กญฺญาย	กญฺญานำ	รตฺติยา	รตฺตีนำ	นาริยา	นารีนำ	รชฺชุยา	รชฺชูนำ	วธุยา	วธูนำ
สัตตมี	กญฺญาย กญฺญายำ	กญฺญาสุ	รตฺติยา รตฺติยำ รตฺยำ	รตฺตีสุ	นาริยา นาริยำ	นารีสุ	รชฺชุยา รชฺชุยำ	รชฺชูสุ	วธุยา วธุยำ	วธูสุ
อาลปนะ	กญฺญเ	กญฺญาโย กญฺญา	รตฺติ	รตฺติโย รตฺตี	นาริ	นาริโย นารี	รชฺชุ	รชฺชุโย รชฺชู	วธุ	วธุโย วธู

ตารางที่ 15 การเปลี่ยนแปลงท้ายศัพท์ของคำที่มีต้นเค้าศัพท์มิใช่เพศหญิงและเพศชาย

วิภัตติ	อการันต์		อิการันต์		อุการันต์	
	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ
ปฐมา	อำ	อานิ	อิ	อีนิ อี	อุ	อูนิ อู
ทุติยา	อำ	อานิ	อึ	อีนิ อี	อุ๋	อูนิ อู
ตติยา	เอน	เอหิ เอภิ	อินา	อีหิ อีภิ	อุนา	อูหิ อูภิ
จตุตถี	อสฺส อาย อตฺถำ	อานำ	อิสฺส อิโน	อีนำ	อุสฺส อุโน	อูนำ
ปัญจมี	อสฺมา อมฺหา อา	เอหิ เอภิ	อิสฺมา อิมฺหา	อีหิ อีภิ	อุสฺมา อุมฺหา	อูหิ อูภิ
ฉัฏฐี	อสฺส	อานำ	อิสฺส อิโน	อีนำ	อุสฺส อุโน	อูนำ
สัตตมี	อสฺมึ อมฺหิ เอ	เอสุ	อิสฺมึ อิมฺหิ	อีสุ	อุสฺมึ อุมฺหิ	อูสุ
อาลปนะ	อ	อานิ	อิ	อีนิ อี	อุ	อูนิ อู

ตารางที่ 16 ตัวอย่างการผันคำของต้นเค้าศัพท์มิใช่เพศหญิงและเพศชาย

วิภัตติ	อการันต์		อิการันต์		อุการันต์	
	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ	เอกวจนะ	พหุวจนะ
ปฐมา	กุลำ	กุลานิ	อกฺขิ	อกฺขีนิ อกฺขี	วตฺถุ	วตฺถูนิ วตฺถู
ทุติยา	กุลำ	กุลานิ	อกฺขึ	อกฺขีนิ อกฺขี	วตฺถุ๋	วตฺถูนิ วตฺถู
ตติยา	กุเลน	กุเลหิ กุเลภิ	อกฺขินา	อกฺขีหิ อกฺขีภิ	วตฺถุนา	วตฺถูหิ วตฺถูภิ
จตุตถี	กุลสฺส กุลาย กุลตฺถำ	กุลานำ	อกฺขิสฺส อกฺขิโน	อกฺขีนำ	วตฺถุสฺส วตฺถุโน	วตฺถูนำ
ปัญจมี	กุลสฺมา กุลมฺหา กุลา	กุเลหิ กุเลภิ	อกฺขิสฺมา อกฺขิมฺหา	อกฺขีหิ อกฺขีภิ	วตฺถุสฺมา วตฺถุมฺหา	วตฺถูหิ วตฺถูภิ
ฉัฏฐี	กุลสฺส	กุลานำ	อกฺขิสฺส อกฺขิโน	อกฺขีนำ	วตฺถุสฺส วตฺถุโน	วตฺถูนำ
สัตตมี	ลสฺมึ ลมฺหิ เล	เลสุ	อกฺขิสฺมึ อกฺขิมฺหิ	อกฺขีสุ	วตฺถุสฺมึ วตฺถุมฺหิ	วตฺถูสุ
อาลปนะ	กุล	กุลานิ	อกฺขิ	อกฺขีนิ อกฺขี	วตฺถุ	วตฺถูนิ วตฺถู

2.2.4.2 การประกอบคำกิริยาอาขยาต

คำกิริยาที่ใช้เป็นกิริยาหลักในประโยคเรียกว่ากิริยาอาขยาต ประกอบด้วยรากศัพท์ ปัจจัย และวิภัตติอาขยาต ปัจจัยสามารถทำให้เครื่องหมายให้ทราบวาจก และวิภัตติอาขยาตที่ใช้เป็นเครื่องหมายให้ทราบกาล บท วจนะ และบุรุษสรรพนาม

วิภัตติในอาขยาตมี 8 วิภัตติ แต่ละวิภัตติประกอบด้วย 2 บท 2 วจนะและ 3 บุรุษสรรพนาม เช่นวัตตมานาวิภัตติ เป็นวิภัตติที่ใช้กล่าวถึงปัจจุบันกาล แปลว่า “...อยู่, ย่อม... หรือ จะ...” ตามลำดับดังแสดงในตารางที่ 17

ตารางที่ 17 การแจงแจงวัตตมานาวิภัตติ [8]

บุรุษ	**ปรัสสบท**		**อัตตโนบท**	
	เอกวจนะ	**พหุวจนะ**	**เอกวจนะ**	**พหุวจนะ**
ปฐม	ติ	อนฺติ	เต	อนฺเต
มัธยม	สิ	ถ	เส	วฺเห
อุตตม	มิ	ม	เอ	มฺห

การประกอบคำกิริยาเช่น รากศัพท์ **ปจฺ** ซึ่งมีความหมายว่า หุงหรือต้ม เมื่อลงปัจจัย กล่าวคือ รากศัพท์ **ปจฺ + อ** ปัจจัย กลายเป็น **ปจ** แล้วนำ **ปจ** ไปประกอบวัตตมานาวิภัตติ จะกลายเป็นคำกิริยาที่มีความหมายดังแสดงในตารางที่ 18 แต่เนื่องจากกฎไวยากรณ์จึงต้องทีฆะสระท้ายปัจจัย คือแปลง อ เป็น อา ก่อนแจกกับ **มิ** หรือ **ม** ดังนั้นจะได้ **ปจา** ก่อนนำประกอบกับ **มิ** และ **ม** ทำให้ได้คำกิริยาเป็น **ปจามิ** และ **ปจาม** แทน **ปจมิ** และ **ปจม** ตามลำดับ

ตารางที่ 18 การประกอบคำกิริยาจากรากศัพท์ ปจฺ (หุง, ต้ม)

บุรุษ	**ปรัสสบท**		**อัตตโนบท**	
	เอกวจนะ	**พหุวจนะ**	**เอกวจนะ**	**พหุวจนะ**
ปฐม	ปจติ	ปจนฺติ	ปจเต	ปจนฺเต
มัธยม	ปจสิ	ปจถ	ปจเส	ปจวฺเห
อุตตม	ปจามิ	ปจาม	ปเจ	ปจามฺห

สำหรับวิภัตติอาขยาตฝ่ายปรัสสบทส่วนมากนิยมใช้ในประโยคที่ประธานเป็นผู้ทำ หรือประธานเป็นผู้ใช้ให้ผู้อื่นทำ ส่วนวิภัตติอาขยาตฝ่ายอัตตโนบทนิยมใช้ในประโยคที่ประธานเป็นผู้ถูกกระทำ แต่ไม่สามารถกำหนดเป็นกฎที่ชัดเจนเพราะสามารถใช้สลับกันได้

การประกอบกิริยาอาขยาต ซึ่งเป็นกิริยาหลักของประโยค ต้องประกอบวจนะให้ตรงตามประธาน เช่น ถ้าประธานเป็นบุคคลที่ 3 และเป็นเอกวจนะ กล่าวคือมีคนเดียว ต้องเลือกใช้ ปจติ แต่ถ้ากล่าวถึงบุคคลที่ 3 ตั้งแต่สองคนขึ้นไป ต้องเลือกใช้ ปจนฺติ เป็นคำกิริยา

2.2.4.3 การประกอบคำกิริยากิตก์

กิริยากิตก์ สามารถทำหน้าที่เป็นคุณนามหรือเป็นกิริยาย่อยในประโยคได้ ประกอบด้วยรากศัพท์ ปัจจัยในกิริยากิตก์ และวิภัตตินาม ซึ่งวิภัตตินามนี้จะต้องตรงคำนามที่มีความสัมพันธ์กัน เช่น **คมฺ** แปลว่า ไป ลง **ต** ปัจจัยในกิริยากิตก์ สำเร็จรูปเป็น **คต** สามารถอธิบายได้ดังนี้

1. ลบพยัญชนะตัวสุดท้ายของรากศัพท์ที่ลงท้ายด้วย ม, น และ ร ดังนั้น **คมฺ** จึงกลายเป็น **ค** เพราะถูกลบพยัญชนะตัวท้าย ด้วยกฎของ **ต** ปัจจัยในกิริยากิตก์

2. ลง **ต** ปัจจัยต่อจากรากศัพท์ที่ได้จากข้อ 1 กลายเป็น **คต**

3. นำ **คต** ไปแจกตามแบบวิภัตตินามได้ทั้ง 3 เพศ เช่น **ปุริโส คโต** แปลว่า อ.บุรุษไปแล้ว, **กญฺญา คตา** แปลว่า อ.หญิงสาว ไปแล้ว เป็นต้น

คำคุณนามในภาษาบาลีเมื่อนำไปขยายคำใด จะต้องประกอบลิงค์ วจนะและวิภัตติให้ตรงกับคำที่ต้องการขยาย คำที่ใช้ขยายและคำที่ถูกขยายอาจมีการันต์เหมือนหรือแตกต่างกันก็ได้ แต่สำนวนแปลของคำที่เป็นคุณนามหรือคำขยายจะใช้เป็น **ผู้, ที่, ซึ่ง** หรือ **อัน** เท่านั้น ยกตัวอย่างเช่น

บุรุษผู้กล้าหาญ คำว่าบุรุษ ตรงกับ **ปุริส** และผู้กล้า ตรงกับ **วีร** ในที่นี้ **วีร** ทำหน้าที่เป็นวิเสสนะของ **ปุริส** เมื่อนำไปไปใช้ทำหน้าที่เป็นภาคประธานจะต้องประกอบรูปเป็นปฐมาวิภัตติและผันรูปเป็นปฐมาวิภัตติทั้งคู่ **ปุริโส วีโร** แปลว่า อ.บุรุษ ผู้กล้า (ไม่ได้แปลว่า อ.บุรุษ อ.ผู้กล้า)

หนังสือของบุรุษผู้ฉลาด คำว่าหนังสือตรงกับ **ปณฺณ** เป็นอการันต์ในนปุงสกลิงค์ (ศัพท์ที่ไม่ใช่เพศหญิงหรือเพศชาย) ทำหน้าที่เป็นประธาน ประกอบรูปเป็นปฐมาวิภัตติได้เป็น **ปณฺณำ** (ถ้าใช้ **ปณฺณานิ** จะหมายถึงหนังสือหลายเล่ม) ส่วนบุรุษในประโยคนี้แสดงความเป็นเจ้าจึงต้องประกอบรูปเป็นฉัฏฐีวิภัตติตามอการันต์ในปุงลิงค์ได้เป็น **ปุริสสฺส** ผู้ฉลาด ส่วนต้นเค้าศัพท์ที่แปลว่าผู้ฉลาดคือ **ปณฺฑิต** ซึ่งทำหน้าที่เป็นคุณนามสามารถเป็นได้ทั้ง 3 ลิงค์ เมื่อขยาย **ปุริส** จึงต้องประกอบรูปที่มีลิงค์ วจนะและวิภัตติตาม กลายเป็น **ปณฺฑิตสฺส** ดังนั้นข้อความว่า “หนังสือของบุรุษ

ผู้ฉลาด" ที่ตรงกับภาษาบาลีคือ "**ปณฺฑิตสฺส ปุริสสฺส ปณฺณ**" หรือ **ปณฺฑิตสฺส ปณฺณ** (สามารถละ **ปุริสสฺส** ได้เพราะ ผู้ฉลาดแสดงชัดอยู่แล้วว่าขยายคน ก็จะตรงกับสำนวนว่า หนังสือของผู้ฉลาด) นอกจากนี้ **ปณฺฑิต** เป็นคุณนามที่สามารถแปลเป็นทับศัพท์แบบตรงตัวได้ จะกลายเป็น "หนังสือของบัณฑิต"

2.2.5 กติปยศัพท์

การประกอบคำนาม ต้องประกอบตามการันต์และเพศของต้นเค้าศัพท์ โดยปกติแล้วต้นเค้าศัพท์ที่จะใช้เป็นนามนามหรือคุณนามล้วนต้องแปลงสระท้ายศัพท์กับวิภัตติตามที่กล่าวไว้ข้างต้นทั้งสิ้น แต่ในภาษาบาลียังมีกลุ่มคำศัพท์ที่ไม่แปลงรูปเหมือนวิภัตตินามทั่วไป เช่น **ราช** เป็นอการันต์ในปุงลิงค์ โดยทั่วไปจะต้องแปลงรูปแบบ **ปุริส** แต่คำศัพท์ที่อยู่ในกลุ่มนี้เรียกว่ากติปยศัพท์นี้มีวิธีแจกเฉพาะตน มีทั้งหมด 12 ศัพท์ [8] ได้แก่ **อตฺต** (ตน) **พฺรหฺม** (พรหม) **ราช** (พระราชา) **ภควนฺตุ** (พระผู้มีพระภาคเจ้า) **อรหนฺต** (พระอรหันต์) **ภวนฺตุ** (ผู้เจริญ) **สตฺถุ** (ผู้สอน, พระศาสดา) **ปิตุ** (บิดา) **มาตุ** (มารดา) **มน** (ใจ) **กมฺม** (การงาน, กรรม) และ**โค** (โค)

2.2.6 สังขยา (Numeral)

คำพูดที่ใช้อธิบายหรือบอกจำนวนเชิงตัวเลขภาษาบาลีเรียกว่าสังขยา ๆ แปลว่าการนับ แบ่งออกเป็น 2 ประเภทได้แก่

1. ปกติสังขยา (Cardinal Number) คือคำที่ใช้บอกให้ทราบจำนวนนับเช่น **เอโก ปุริโส** (บุรุษ 1 คน), **เอกำ คามำ** (บ้าน 1 หลัง) หรือ **เอกา อิตฺถี** (หญิง 1 คน) เป็นต้น

2. ปูรณสังขยา (Ordinal Number) คือคำที่ใช้บอกให้ทราบลำดับที่เช่น ปฐโม **ภาโค** (ภาคที่หนึ่ง), **ปฐมำ คามำ** (บ้านหลังแรก) เป็นต้น

2.2.7 คำสมาส (Samasa)

การย่อหน่วยศัพท์ตั้งแต่ 2 ศัพท์ขึ้นไปเข้าไปบทเดียวเรียกว่าสมาส สมาสบางอย่างคล้ายคลึงกับการย่อคำพูดในภาษาไทย เช่น ยาสำหรับคนไข้ หรือยาที่เตรียมไว้ให้คนไข้ เมื่อพูดสั้น ๆ คือยาคนไข้ ภาษาบาลีคือ **คิลานสฺส เภสชฺช** เมื่อเข้าสมาสจะลบวิภัตติออกให้กลายเป็นคำศัพท์ดั้งเดิม **คิลานสฺส** (ของคนไข้) เมื่อลบวิภัตติออกกลายเป็น **คิลาน** จากนั้นนำไปต่อกันกลายเป็น **คิลานเภสชฺช** หรือดอกไม้ที่อยู่ในป่าหรือเกิดจากในป่า คนไทยนิยมเรียกสั้น ๆ ว่าดอกไม้ป่า ตรงกับภาษาบาลีคือ **วเน** (ในป่า) และ **ปุปฺผ** (ดอกไม้) กลายเป็น **วนปุปฺผ** (ดอกไม้ป่า)

สมาสบางอย่างย่อคำให้สั้นลงคล้าย พ่อและแม่ กลายเป็นพ่อแม่ คำว่า “แม่และพ่อ” ใช้คำเชื่อมวางไว้ระหว่างคำที่ต้องการเชื่อมเนื้อความเข้าด้วยกัน แต่ในภาษาบาลีจะใช้คำศัพท์สำหรับเชื่อมความวางไว้ด้านหลัง ได้แก่ มาตา จ ปิตา จ แปลว่า อ.มารดาด้วย อ.บิดาด้วย คำสำเร็จรูป มาตา และ ปิตา ในที่นี้เป็นปฐมาวิภัตติเอกวจนะ เมื่อเข้าสมาสจะกลายเป็น มาตาปิตโร แปลว่า อ.มารดาและบิดาทั้งหลาย (เปลี่ยนคำสำเร็จรูปของ ปิตา เป็น ปิตโร รูปพหูพจน์ เพราะถือว่าเป็น 2 คน) เมื่อย่อคำศัพท์จนกลายเป็นสมาสแล้วจะถือว่าเป็นคำเดียว ไม่สามารถแยกออกจากกันได้อีก

การนำคำศัพท์มารวมกันด้วยคำสมาสนี้ด้วยคำที่ทราบความหมายอยู่แล้ว แต่เมื่อเป็นคำสมาสแล้วสามารถมีคำแปลได้หลายแบบ เช่น เทวราชา สามารถแปลว่า อ.เทวดาผู้พระราชา หรือ อ.ราชาของเทวดา โดยทั้งสองคำแปลนี้ล้วนมีความหมายไปในทิศทางเดียวกันคือหัวหน้าของเทวดา ดังนั้นคำสมาสเหล่านี้มักไม่ปรากฏคำแปลหรือความหมายในพจนานุกรม เพราะเมื่อผู้เรียนสามารถหาคำแปลหรือความหมายได้จากการตั้งรูปวิเคราะห์ทางภาษาบาลี

โดยส่วนมากแล้วคำสมาสจะปรากฏคำศัพท์เดิมให้เห็น ทำให้ผู้แปลสามารถคาดเดาความหมายได้ง่าย แต่มีบางครั้งที่ใช้เพียงเนื้อความของคำศัพท์แรกเพียงแค่อักษรหรือพยางค์เดียวไปต่อรวมกับคำศัพท์หลัง เช่น กุจฺฉิตา ทิฏฺฐิ แปลว่า อ.ทิฐิ อันบัณฑิต รังเกียจแล้ว (ความเห็นที่ถูกรังเกียจ)

2.2.8 คำสนธิ (Sandhi)

การนำคำที่มีความหมายและอยู่ติดกันตั้งแต่สองคำขึ้นไป มาเปลี่ยนรูปให้กลายเป็นคำเดียวเพื่อให้สามารถออกเสียงได้อย่างไพเราะ เรียกว่าคำสนธิ ๆ นี้ใช้ในคำประพันธ์ประเภทร้อยกรองและร้อยแก้ว แต่การนำคำมาเชื่อมเป็นคำสนธินี้จะทำให้ไม่สามารถแสดงคำทั้งหมดในพจนานุกรมได้ ต้องตัดคำสนธิให้เป็นรูปเดิมเสียก่อน คำสนธินี้คล้ายคลึงกับการเขียนคำย่อในภาษาอังกฤษ เช่น You are ย่อเป็น You’re เป็นต้น การสนธินี้อาจทำรูปอักษรของคำหน้าหรือคำหลังหายไป ดังนั้นการตัดสนธิออกแล้วคืนรูปคำให้เหมือนเพื่อให้สามารถสื่อความหมายได้เหมือนเดิม

2.3. วรรณกรรมที่เกี่ยวข้อง

วรรณกรรมที่เกี่ยวข้องกับวิทยานิพนธ์นี้ ผู้วิจัยศึกษางานประมวลผลภาษาด้านการตัดคำในภาษาบาลีสันสฤต และงานวิจัยที่นำการเรียนรู้เชิงลึกมาแก้ปัญหางานด้านการประมวลผลภาษาธรรมชาติดังนี้

2.3.1 การประมวลภาษาธรรมชาติด้านภาษาบาลีและสันสฤต

การประมวลผลธรรมชาติเป็นกระบวนการวิเคราะห์และประมวลผลด้านภาษา โดยมีจุดมุ่งหมายเพื่อให้คอมพิวเตอร์สามารถเข้าใจภาษาของมนุษย์ มีการประยุกต์อย่างกว้างขวาง เช่น การวิเคราะห์อารมณ์จากข้อความ การแปลภาษา และการสังเคราะห์เสียงพูดเป็นต้น

การตัดคำเป็นการประมวลผลเพื่อทำให้ทราบขอบเขตของคำก่อนนำไปประมวลผล การตัดคำนั้นเกิดขึ้นกับหลายภาษามีทั้งภาษาที่ไม่มีมีเครื่องหมายเว้นวรรคระหว่างเช่นภาษาไทย และภาษาที่มีใช้เครื่องหมายเว้นวรรคเช่นภาษาอังกฤษ ภาษาเยอรมัน และภาษาฝรั่งเศษเป็นต้น งานวิจัยด้านตัดคำในภาษาที่ใช้เครื่องหมายเว้นวรรคระหว่างคำมุ่งเน้นไปที่การตัดคำผสม (Compounds word) ซึ่งเกิดจากการรวมหน่วยศัพท์ตั้งแต่สองศัพท์ขึ้นไปเข้าเป็นคำเดียว โดยเฉพาะอย่างยิ่งในภาษาที่สามารถเปลี่ยนรูปคำได้หลากหลาย (Rich morphology)

ภาษาบาลีเป็นภาษาหนึ่งที่สามารถเปลี่ยนรูปคำได้หลากหลาย สามารถสร้างคำใหม่ได้อย่างไม่รู้จบ และที่น่าสนใจคือในประเทศไทยมักบัญญัติศัพท์ใหม่มาจากคำที่ถูกสร้างขึ้นด้วยภาษาบาลีสันสฤต แต่ในงานประมวลผลภาษาธรรมชาติที่ประยุกต์ใช้กับภาษาบาลีมีไม่มากนัก ในประเทศไทยระบบการแปลภาษาบาลีอักษรไทยเป็นภาษาไทยด้วยคอมพิวเตอร์ [4] ซึ่งใช้พจนานุกรมภาษาบาลี-ไทยรวมทั้งจัดเก็บไวยากรณ์โครงสร้างภาษาบาลีด้วยไวยากรณ์แบบไม่พึ่งบริบท (Context-free grammar) เป็นแหล่งความรู้สำหรับการประมวลผลภาษาธรรมชาติเพื่อแปลประโยคความเดียวและประโยคความซ้อน นักวิจัยได้ทดลองกับประโยคข้อความเดียว 20 ตัวอย่าง และประโยคความซ้อน 32 ตัวอย่าง

การแปลภาษาบาลีเป็นภาษาไทยเป็นเรื่องที่ท้าท้ายมากแต่พบงานวิจัยได้ไม่มากนัก เพราะคำ (Word) เพียงคำเดียวไม่ว่าจะเป็นคำนามหรือคำกิริยาสามารถบอกเนื้อความได้หลายอย่างตั้งแต่เพศ วจนะบุพบท หน้าที่ของคำ ดังนั้นการแปลภาษาบาลีเป็นไทยในงานวิจัยนี้จึงต้องจัดเก็บหน่วยศัพท์ (Lexeme) รวมถึงการเปลี่ยนรูปทุกแบบเพื่อให้ครอบคลุม กล่าวคือ 1 หน่วยศัพท์สามารถแจกแจงคำได้มากสูงสุดถึง 120 คำ มาจาก 3 (3 เพศ) × 8 (7 วิภัตติ + อาลปนะ) × 5 (เอกวจนะเปลี่ยนรูปได้มาก 3 รูป + พหุวจนะเปลี่ยนได้มากสุด 2 รูป)

ปัจจุบันมีการวิจัยในงานด้านการประมวลผลภาษาธรรมชาติและภาษาศาสตร์เชิงคำนวณที่เกี่ยวข้องกับภาษาบาลีและสันกฤต ในรูปที่ 3 แสดงแผนภาพงานวิจัยการประมวลผลภาษาธรรมชาติในภาษาบาลีที่เขียนด้วยอักษรต่าง ๆ ได้แก่อักษรไทย อักษรพม่า อักษรโรมัน และอักษรเทวนาครี

รูปที่ 3 แผนภาพงานวิจัยการประมวลผลภาษาบาลี

งานวิจัยที่เกี่ยวข้องกับภาษาบาลีอักษรไทย ได้แก่เครื่องแปลภาษาจากภาษาบาลีอักษรไทยเป็นภาษาไทย [4] การแปลภาษาบาลีอักษรไทยเป็นอังกฤษ [10] การแปลงอักษรบาลีเป็นสัทอักษร [5] การค้นคืนพระคาถาในธรรมบท [7] ระบบจำลองการอ่านภาษาบาลี [6]และการตรวจจับคำยืมจากภาษาบาลีสันสกฤตที่มาใช้ในภาษาไทย [11]

งานวิจัยที่เกี่ยวข้องกับภาษาบาลีอักษรพม่าพบ 2 งาน คือ เครื่องแปลภาษาบาลีอักษรพม่าเป็นภาษาพม่า [12] และการระบุคำพม่าที่แปลงมาจากภาษาบาลี [13]

งานวิจัยที่เกี่ยวข้องกับภาษาบาลีอักษรเทวนาครี เช่น งานด้านการสังเคราะห์เสียง [14] การตรวจสอบชนิดของคำ [15] การสร้างพจนานุกรมภาษาบาลีสำหรับสืบค้นข้ามภาษา [16] และการพัฒนากฎการสนธิคำตามรูปแบบไวยากรณ์ของนักไวยากรณ์ชื่อปาณินิ [17]

ส่วนงานวิจัยเกี่ยวกับภาษาบาลีอักษรโรมันได้แก่ การพัฒนาเครื่องมือที่แสดงการเปลี่ยนรูปต้นเค้าศัพท์ที่เป็นไปได้ทั้งหมด การวิเคราะห์หน้าที่ของคำ การแยกหน่วยคำ รวมไปถึงการเชื่อมและการตัดคำสนธิ ในส่วนของการตัดคำสนธิใช้วิธีการพิจารณาทีละตัวอักษรจากซ้ายไปขวาจนสิ้นสุดคำร่วมกับกฎที่สร้างขึ้นจากทุกความเป็นได้ของการผสมอักษร เช่น อ+อ = อ เป็นกฎที่ 1 และ อ+อ = อา เป็นกฎที่ 2 เป็นต้น ดังนั้นการตัดสนธิในสามารถช่วยแสดงรายการการตัดคำสนธิทั้งหมดที่เป็นไปได้ แต่ผู้ใช้ต้องเลือกรายการที่สอดคล้องกับประโยคเอง [18]

ภาษาสันสกฤตเป็นภาษาที่มีความใกล้เคียงกับภาษาบาลีมากเพราะมีแหล่งกำเนิดและโครงสร้างภาษาที่ใกล้เคียงกัน ปัจจุบันภาษาสันสกฤตไม่ได้ใช้สื่อสารในชีวิตประจำวันเช่นเดียวกับ

ภาษาบาลี แต่ยังคงมีการสอนภาษาสันสกฤตในประเทศอินเดียและประเทศใกล้เคียง ซึ่งอาจเป็นพราะเป็นภาษาที่ใช้บันทึกคัมภีร์ที่สำคัญของศาสนาพราหมณ์ฮินดู ส่วนภาษาบาลียังคงมีสอนในประเทศที่นับถือพุทธศาสนานิการเถรวาทเพื่อศึกษาคำสั่งสอนของพระพุทธเจ้า มีการวิจัยการมวลผลภาษาธรรมชาติด้านภาษาสันสกฤตที่หลากหลาย เช่น การตัดคำสนธิ [19-21] ประยุกต์ใช้การเรียนรู้เชิงลึกจำนวน 2 งาน ซึ่งจะกล่าวในหัวข้อถัดไป และตัดสนธิโดยเปรียบเทียบเครื่องมือและชุดข้อมูล ถึงแม้ว่าจะมีแหล่งข้อมูลและงานวิจัยจำนวนมาก แต่ยังไม่พบชุดข้อมูลที่เป็นมาตรฐานกลาง และไม่มีการเปรียบเทียบประสิทธิภาพของเครื่องมือดังนั้นงานวิจัยนี้จึงได้เปรียบเทียบและรายงานผล การแปลความหมายของคำสมาส [22] ซึ่งเริ่มหาความหมายของคำสมาสจากการแยกต้นเค้าศัพท์ออกจากกันเป็น N หน่วย จากนั้นพิจารณาต้นเค้าศัพท์ 2 หน่วยที่ติดกัน เช่น คำสมาส abc สามารถแยกต้นเค้าศัพท์ได้เป็น a-b-c ดังนั้น ในขั้นตอนนี้สามารถพิจารณาได้เป็น <a-<bc> และ <<ab>-c> เป็นต้น หลังจากจับคู่สมาสเรียบร้อยแล้วจะวิเคราะห์ว่าเป็นสมาสชนิดใด เมื่อทราบชนิดของสมาสแล้วจะสามารถทราบคำแปลหรือความหมายได้ แต่ในขั้นตอนการวิเคราะห์ชนิดของสมาสจำเป็นต้องทราบชนิดของคำ (POS) เสียก่อน การสร้างคำสมาสในภาษาบาลีนั้นแต่ละครั้งสามารถรวมต้นเค้าศัพท์ตั้งแต่สองศัพท์ขึ้นไปโดยใช้การสมาสครั้งเดียว และสามารถนำคำสมาสมาสร้างคำสมาสซ้ำได้อีกซึ่งเรียกว่าสมาสหลายชั้น

2.3.2 การเรียนรู้เชิงลึกสำหรับงานด้านการประมวลผลภาษา

การเรียนรู้เชิงลึก เป็นการเรียนรู้ที่ถูกพัฒนาขึ้นมาการโครงข่ายประสาทเทียม แต่มีจำนวนชั้นมากกว่า มีชื่อเรียกหลากหลายตามลักษณะการทำงานและโครงสร้าง มีการทำมาประยุกต์ใช้อย่างแพร่หลาย สำหรับการประยุกต์ใช้ที่เกี่ยวข้องกับการประมวลผลภาษาได้แก่โครงข่ายประสาทประสาทเทียมแบบสังวัตนาการ (CNN: Convolutional Neural Networks) ซึ่งมีจุดเริ่มต้นมาจากงานวิจัยด้านการประมวลผลภาพด้านการรู้จำตัวอักษร แต่มีงานวิจัยใช้ CNN เพื่อสกัดลักษณะเฉพาะของตัวอักษรไปแก้ปัญหาด้านการติดแท็ก [23] และการประเมินอารมณ์จากข้อความ [24] นอกจากนี้ยังนำมาประยุกต์ใช้กับการจำแนกประเภทข้อความภาษาไทยโดยสามารถละขั้นตอนการตัดคำได้ [25] รวมถึงนำมาใช้ในบทประพันธ์เช่น การสอนคำกลอนด้านความรัก [26] การวิเคราะห์ชนิดของคำในบทร้อยกรอง [27] เป็นต้น ส่วนโครงข่ายประสาทเทียมที่นำมาใช้กับงานด้านประมวลผลภาษาธรรมชาติและมีประสิทธิภาพที่ดีคือโครงข่ายประสาทเทียมแบบหมุนกลับ (RNN: Recurrent Neural Network) จากการเปรียบเทียบประสิทธิภาพและตรวจสอบ

พารามิเตอร์ที่เหมาะสมกับงานด้านการประมวลผลภาษา [28, 29] พบว่าการเลือกใช้ RNN จะทำเหมาะสมมากกว่า CNN

2.3.2 การเรียนรู้เชิงลึกสำหรับงานด้านการตัดคำสนธิภาษาสันสกฤต

ภาษาสันสกฤตมีงานวิจัยด้านการตัดคำสนธิที่ใช้โครงข่ายประสาทเทียมแบบ 2 ทิศทาง [30] ซึ่งดำเนินการตัดคำสองระดับ ในระดับแรกทำนายตำแหน่งของการตัดสนธิ และส่งต่อให้ระดับที่สองทำนายการตัดคำสนธิ งานวิจัยนี้มีความแม่นยำในการทำนายตำแหน่งที่ใช้ตัดคำสนธิ 95% และทำนายการตัดคำสนธิ 79.5% ตามลำดับ ในงานวิจัยนี้ใช้ข้อมูลเข้าเป็นคำสนธิและผลลัพธ์คือคำที่ถูกแยกออกจาก เช่นข้อมูลเข้าคือ "**เอตมฺมงฺคลมุตฺตมำ**" และผลลัพธ์คือ "**เอตำ+มงฺคลำ+อุตฺตมำ**" นอกจากนี้ยังพบงานที่ใช้แนวคิดแบ่งการตัดคำสนธิออกเป็นสองส่วนที่คลายคลึงกัน [20] โดยทำงานตำแหน่งตัดคำก่อน แล้วค่อยตัดคำโดยใช้ข้อมูลจากตำแหน่งตัดคำที่ทำนายได้

2.3.3 สรุปงานวิจัยที่เกี่ยวข้อง

จากการค้นคว้ายังไม่พบงานตัดคำภาษาบาลีอักษรไทย มีเพียงงานแปลภาษาบาลีอักษรไทยที่จัดเก็บไวยากรณ์ภาษาบาลีอยู่ในรูปไวยากรณ์ไม่พึ่งบริบทเพื่อให้ทราบหน้าที่ของคำและนำไปสู่การแปลภาษา

ในภาษาบาลีอักษรโรมันที่พบเครื่องมือที่เกี่ยวข้องนี้กับการตัดสนธินี้มีกฎที่ได้จากการรวบรวมการผสมอักษรแล้วแปลงคำเพื่อให้สามารถนำคำสนธิสองคำมาเชื่อมเป็นคำเดียวกัน กฎการแยกสนธิซึ่งเป็นกฎการแปลงย้อนกลับของกฎการสนธิ สำหรับงานนี้เพียงช่วยให้สามารถสนธิและแยกสนธิได้ตามหลักไวยากรณ์ สำหรับกฎการสนธิและแยกคำสนธินี้ถูกจัดเก็บอยู่ในรูปของนิพจน์ปกติ แต่ข้อเสียที่ชัดคือคำบางคู่อาจเชื่อมสนธิด้วยกฎที่ผู้คนไม่นิยมใช้กัน

เนื่องจากงานวิจัยด้านการประมวลผลภาษาธรรมชาติมีน้อยมาก ดังนั้นจึงได้ศึกษาภาษาที่ใกล้เคียงกับภาษาบาลีนั่นคือภาษาสันสกฤตจึงพบงานวิจัยด้านการตัดคำสมาสและตัดคำสนธิ ในส่วนของงานตัดคำสมาสนี้แบ่งเป็นหลายขึ้นตอนแต่ขึ้นตอนที่สำคัญที่สุดคือขั้นตอนการแบ่งคำสมาสซึ่งถูกประมวลผลด้วยยูนิแกรม ส่วนการตัดคำสนธิใช้โครงข่ายประสาทเทียมแบบวกกลับ ดังสรุปข้อมูลไว้ในตารางที่ 19

ตารางที่ 19 สรุปงานวิจัยที่เกี่ยวข้อง

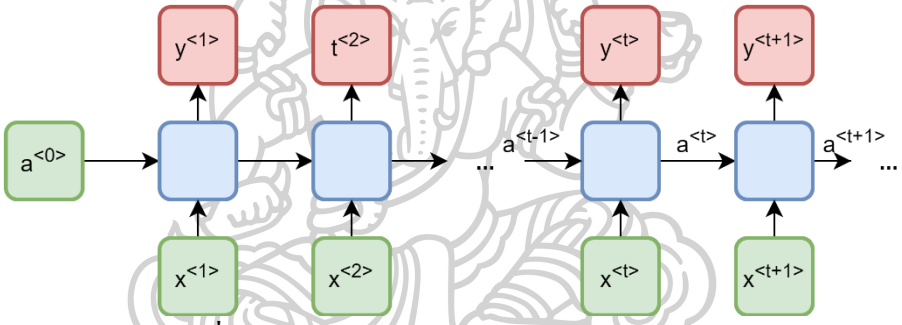
ลำดับ	งานวิจัย	ลักษณะงาน	เทคนิค
1	Morphological analyzer and generator for Pali [18]	แจกแจงการแปลงรูปคำและตัดคำสนธิบาลี	RE
2	Sanskrit Sandhi Splitting using seq2(seq)^2 [19]	ตัดคำสนธิสันสกฤต	RNN ENCODE DECODER
3	Neural Compound-Word (Sandhi) Generation and Splitting in Sanskrit Language [20]	ตัดคำสนธิสันสกฤต	RNN+LSTM
4	Sandhikosh: A benchmark corpus for evaluating sanskrit sandhi tools [21]	เปรียบเครื่องมือการตัดคำสนธิสันสกฤต	Tools: JNU, INRIA, UoH
5	Sanskrit Compound Processor [22]	แปลคำสมาสสันสกฤต	แยกศัพท์จากสมาสด้วย UNIGRAM

บทที่ 3
ทฤษฎีที่เกี่ยวข้องและความรู้ที่เกี่ยวข้อง

ในงานวิจัยนี้ศึกษาทฤษฎีและความรู้ที่เกี่ยวข้องกับการตัดคำ และโครงข่ายแอลเอสทีเอ็มแบบสองทิศทาง (Bidirectional Long Short Term Network - BiLSTM) คำสนธิ คำสมาส และวิธีการวัดประสิทธิภาพ

3.1 โครงข่ายประสาทเทียมแบบย้อนกลับ

โครงข่ายประสาทเทียมแบบย้อนกลับ (Recurrent neural network : RNN) เป็นโครงข่ายประสาทเทียมที่ยอมให้ใช้ผลลัพธ์ของโหนดก่อนหน้าเป็นข้อมูลเข้าในโหลดถัดไป ดังแสดงตัวอย่างในรูปที่ 4



รูปที่ 4 ตัวอย่างโครงข่ายประสาทเทียมแบบย้อนกลับ

กำหนดให้ ณ เวลา t สามารถหาผลลัพธ์ที่คำนวณได้ $a^{<t>}$ จากฟังก์ชันกระตุ้น g_1 และผลลัพธ์ $y^{<t>}$ จากฟังก์ชันกระตุ้น g_2 ได้ดังสมการที่ 3.1 และ 3.2

$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a) \tag{3.1}$$

$$y^{<t>} = g_2(W_{ya}a^{<t>} + b_y) \tag{3.2}$$

W_{aa} คือค่าน้ำหนักบนเส้นเชื่อมระหว่างผลลัพธ์เก่าและใหม่ $a^{<t-1>}$ และ $a^{<t>}$

W_{ax} คือ ค่าน้ำหนักบนเส้นเชื่อมระหว่างชั้นข้อมูลเข้า $x^{<t>}$และชั้นซ่อน $a^{<t>}$

W_{ya} คือ ค่าน้ำหนักบนเส้นเชื่อมระหว่างชั้นซ่อน $a^{<t>}$ และชั้นผลลัพธ์ $y^{<t>}$

b_a, b_y คือ bias

ในระหว่างการรับข้อมูลค่าของเข้าข้อมูลเข้า $x^{<1>}$ จะถูกคูณกับค่าน้ำหนักบนเส้นเชื่อมทุกครั้งและทำให้ค่าของข้อมูลเข้าลดลงจนเข้าใกล้ศูนย์ ส่งผลให้ค่าน้ำหนักไม่ถูกปรับปรุง เรียกว่าปัญหาวานิชชิงเกรเดียน จึงได้มีการพัฒนาโครงข่ายแอลเอสทีเอ็มถูกพัฒนามาเพื่อแก้ไขปัญหาวานิชชิงเกรเดียน

3.2 โครงข่ายแอลเอสทีเอ็ม (Long Short-Term Network)

โครงข่ายแอลเอสทีเอ็มถูกได้ออกแบบมาเพื่อแก้ไขปัญหานิชชิงเกรเดียน มีแนวคิดในการนำข้อมูลก่อนหน้ามาใช้ทำนายร่วมด้วย โดยข้อมูลที่ถูกเก็บก่อนหน้า เรียกว่า state และใช้แนวคิดของ gate แทนชั้น (layer) ซึ่งประกอบด้วย gate จำนวน 3 อันได้แก่ forget gate, input gate และ output gate

เมื่อโครงข่ายรับข้อมูล $x^{<t>}$ และรับค่าผลลัพธ์ก่อนหน้าที่คำนวณได้ $a^{<t-1>}$ เข้ามา และให้ forget gate พิจารณาว่าควรลบข้อมูลหรือเก็บไว้ด้วย sigmoid function ดังสมการที่ 3.3

$$\Gamma_{\mathrm{f}} = \sigma(W_f x^{<t>} + U_f a^{<t-1>} + b_f) \tag{3.3}$$

W_f คือค่าน้ำหนักบนเส้นเชื่อมข้อมูลเข้า $x^{<t>}$ ของ forget gate

U_f คือ ค่าน้ำหนักบนเส้นเชื่อมจากผลลัพธ์ก่อนหน้า $a^{<t-1>}$ มายัง forget gate

b_f คือ bias ของ forget gate

โดยข้อมูลที่รับเข้ามา จะมี input gate พิจารณาว่าควรเก็บค่าข้อมูลใดไว้ใน cell sate โดยคำนวณผ่าน sigmoid function ดังสมการที่ 3.4

$$\Gamma_i = \sigma(W_i x^{<t>} + U_i a^{<t-1>} + b_i) \tag{3.4}$$

W_i คือค่าน้ำหนักบนเส้นเชื่อมข้อมูลเข้า $x^{<t>}$ ของ input gate

U_i คือ ค่าน้ำหนักบนเส้นเชื่อมจากผลลัพธ์ก่อนหน้า $a^{<t-1>}$ มายัง input gate

b_i คือ bias ของ input gate

จากนั้นรวบรวมข้อมูลปัจจุบัน $\tilde{C}^{<t>}$ กับข้อมูลก่อนหน้าสมการดังสมการที่ 3.5

$$\tilde{C}^{<t>} = tanh(W_c x^{<t>} + U_c a^{<t-1>} + b_c) \tag{3.5}$$

W_c คือค่าน้ำหนักบนเส้นเชื่อมข้อมูลเข้า $x^{<t>}$ ของข้อมูลปัจจุบัน

U_c คือ ค่าน้ำหนักบนเส้นเชื่อมจากผลลัพธ์ก่อนหน้า $a^{<t-1>}$ มายังข้อมูลปัจจุบัน

b_c คือ bias ของข้อมูลปัจจุบัน

โดยข้อมูลปัจจุบัน $\tilde{C}^{<t>}$ เกิดจากการรวมข้อมูลจาก forget gate และ input gate เข้าด้วยกัน ดังสมการที่ 3.6

$$\tilde{C}^{<t>} = \Gamma_f \tilde{C}^{<t-1>} + \Gamma_i \tilde{C}^{<t>} \tag{3.6}$$

ขั้นสุดท้ายจะได้ค่าของ output gate และผลลัพธ์ $a^{<t>}$ ดังสมการที่ 3.7 และ 3.8

$$\Gamma_o = \sigma(W_o x^{<t>} + U_o a^{<t-1>} + b_o) \tag{3.7}$$

$$a^{<t>} = \Gamma_o \tanh(\tilde{C}^{<t>}) \tag{3.8}$$

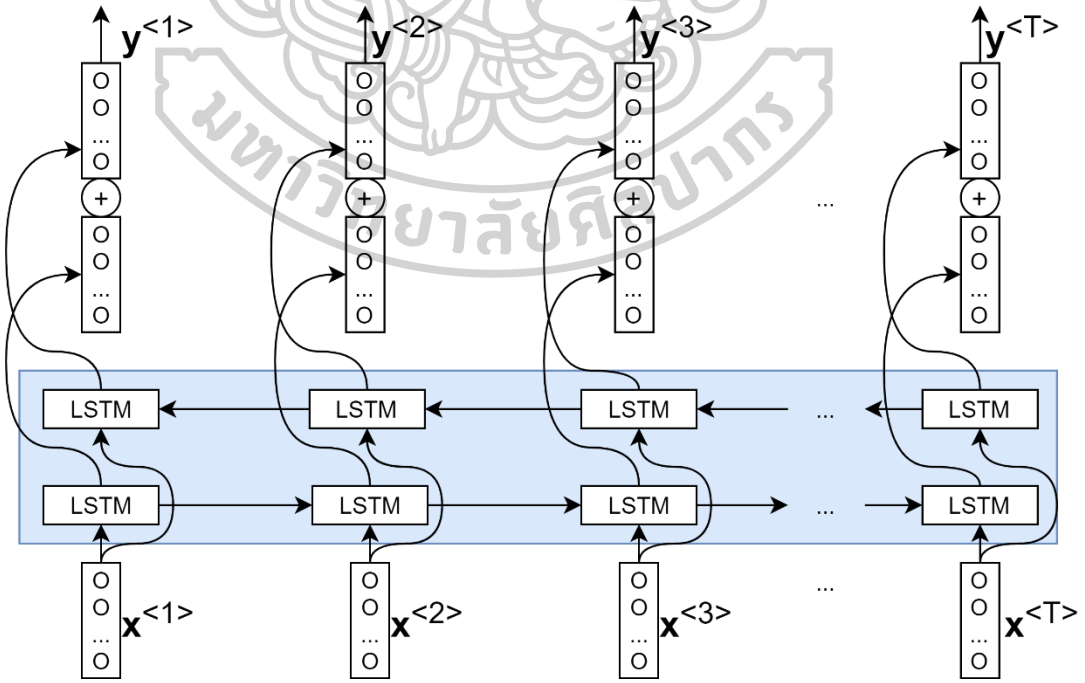
W_o คือค่าน้ำหนักบนเส้นเชื่อมข้อมูลเข้า $x^{<t>}$ ของ output gate

U_o คือ ค่าน้ำหนักบนเส้นเชื่อมจากผลลัพธ์ก่อนหน้า $a^{<t-1>}$ มายัง output gate

b_o คือ bias ของ output gate

3.3 โครงข่ายแอลเอสทีเอ็มแบบสองทิศทาง (Bidirectional Long Short-Term Network)

โครงข่ายแอลเอสทีเอ็มมีแบบสองทิศทางมีแนวคิดการใช้ข้อมูลจากซ้ายไปขวา และจากขวาไปซ้าย ช่วยให้สามารถนำข้อมูลก่อนหน้ามาทำนายข้อมูลถัดไปในการเก็บข้อมูลก่อนหน้า โดยนำ cell ของโครงข่ายแอลเอสทีเอ็มมาเรียกต่อกันสองชั้นดังรูปที่ 5



รูปที่ 5 ตัวอย่างโครงข่ายแบบสองทิศทาง

3.4 การวัดประสิทธิภาพ

Confusion matrix เป็นวิธีการวัดประสิทธิภาพโมเดลการจำแนกประเภทที่นิยมวิธีหนึ่ง มีลักษณะเป็นตารางที่ใช้แสดงจำนวนเปรียบเทียบระหว่างผลฉลยจริงและผลเฉลยที่ทำนายได้ ใจตารางที่ 20 แสดงตัวตัวอย่าง Confusion matrix ที่มีการทำนายแบบ 2 คลาสที่ทำนายเป็นจริงและเท็จ โดยค่าในตารางประกอบด้วย

1. True Positive (TP): ทำนายว่าคำตอบเป็นจริงและผลเฉลยเป็นจริง (ทำนายถูก)
2. False Negative (FN): ทำนายว่าคำตอบเป็นเท็จ แต่ผลเฉลยเป็นจริง (ทำนายผิด)
3. False Positive (FP): ทำนายว่าเป็นเท็จ แต่ผลเฉลยเป็นจริง (ทำนายผิด)
4. True Negative (FN): ทำนายว่าเป็นเท็จ และผลเฉลยเป็นเท็จ (ทำนายถูก)

ตารางที่ 20 ตัวอย่าง Confusion matrix

		ทำนาย	
		จริง	เท็จ
ผลเฉลย	จริง	TP	FN
	เท็จ	FP	TN

จากค่าที่ปรากฏใน confusion matrix สามารถนำมาคำนวณหาค่า Precision, Recall และ F1-score ได้ด้วยสมการต่อไปนี้

$$Precision = \frac{TP}{TP + FP} \tag{3.9}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.10}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.11}$$

บทที่ 4

วิธีดำเนินการวิจัย

งานวิจัยนี้ศึกษาเรื่องการตัดคำในภาษาบาลีอักษรไทยใช้ข้อมูลจากหนังสือธมฺมปทฏฺฐกถา (ธัมมปทัฏฐากถา – ทำ-มะ-ปะ-ทัด-ถะ-กะ-ถา) รวบรวมคำสนธิได้จำนวน 6,385 คำ และรวบรวมคำสมาสได้จำนวน 4,478 คำ เนื้อหาในบทนี้ประกอบด้วยการรวบรวมและเตรียมหนังสือในรูปแบบไฟล์ข้อความ (Plan text), การตัดคำสนธิ และการตัดคำสมาส

4.1 การเตรียมข้อมูลและการรวบรวมชุดข้อมูล

การวิจัยนี้รวบรวมคำสนธิและคำสมาสจากหนังสือภาษาบาลีอักษรไทยจากหนังสือธมฺมปทฏฺฐกถา เพราะเป็นหนังสือที่ใช้เป็นแบบเรียนตามหลักสูตรการเรียนภาษาบาลีในประเทศไทย มีทั้งหมด 8 ภาค ดังแสดงข้อมูลการรวบรวมคำสนธิและสมาสในตารางที่ 21

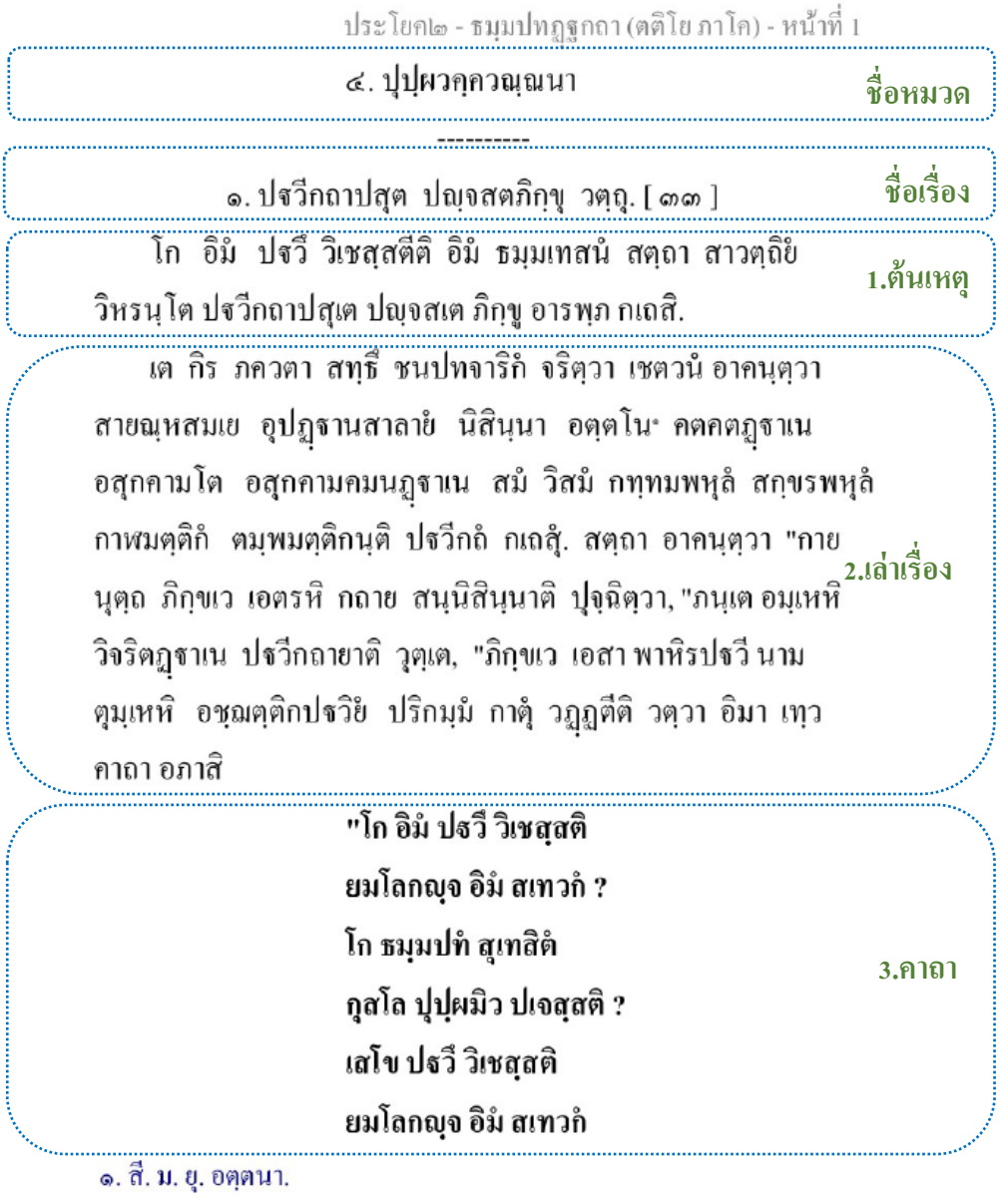
ตารางที่ 21 หนังสือธมฺมปทฏฺฐกถาที่ใช้รวบรวมคำสนธิและคำสมาส

ชื่อ	ชื่อภาษาไทย	รวบรวมคำสนธิ	รวบรวมคำสมาส
ปฐโม ภาโค	ภาคที่ 1 (เล่มที่ 1)	✓	✓
ทุติโย ภาโค	ภาคที่ 2 (เล่มที่ 2)	✓	✓
ตติโย ภาโค	ภาคที่ 3 (เล่มที่ 3)	✓	✓
จตุตฺโถ ภาโค	ภาคที่ 4 (เล่มที่ 4)	✓	✓
ปญฺจโม ภาโค	ภาคที่ 5 (เล่มที่ 5)	✓	
ฉฏฺโฐ ภาโค	ภาคที่ 6 (เล่มที่ 6)	✓	
สตฺตโม ภาโค	ภาคที่ 7 (เล่มที่ 7)	✓	
อฏฺฐโม ภาโค	ภาคที่ 8 (เล่มที่ 8)	✓	

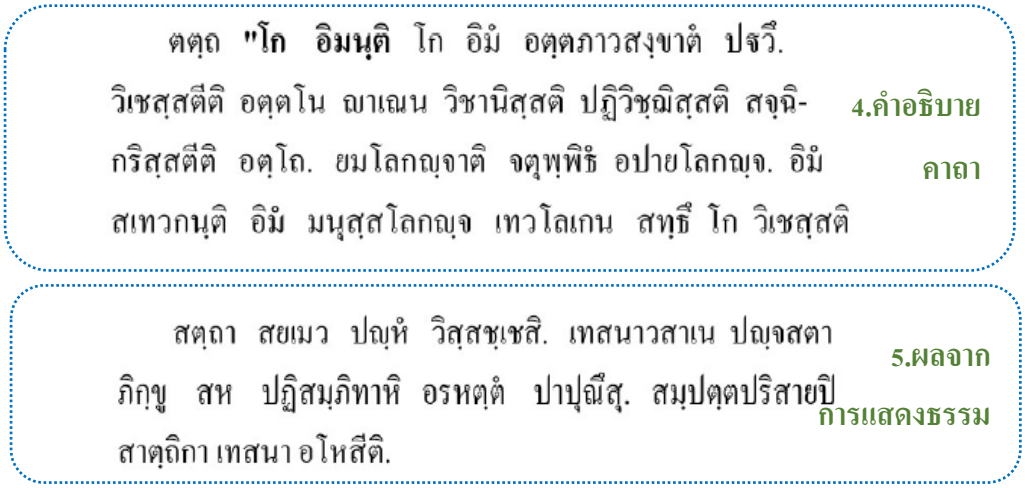
ภาษาบาลีใช้การเว้นวรรคระหว่างคำ ช่วยให้สังเกตคำได้ง่าย หนังสือธมฺมปทฏฺฐกถาใช้เครื่องหมายจุดหรือมหัพภาค (.) เพื่อแบ่งประโยค แต่หนังสือบาลีอักษรไทยบางเล่มใช้เครื่องหมายคั่นเดี่ยว (ฯ) เป็นเครื่องหมายจบประโยค รูปแบบการประพันธ์เนื้อหาทั้งร้อยแก้วและร้อยกรอง โดยบทประพันธ์ส่วนร้อยกรองจะถูกเน้นตัวหน้าเอาไว้ ลักษณะการเรียงในหนังสือธมฺมปทฏฺฐกถาประกอบด้วย 5 ส่วนดังต่อไปนี้ ดังแสดงในรูปที่ 6 และรูปที่ 7 ได้แก่

1. การสนทนาปรารภต้นเหตุการเทศนา
2. ต้นเหตุการเทศนาของพระพุทธเจ้า

3. คาถาหรือบทประพันธ์ร้อยกรอง
4. ส่วนคำอธิบายของคำที่ปรากฏคาถาหรือบทประพันธ์ร้อยกรอง
5. ผลประโยชน์ของธรรมเทศนา



รูปที่ 6 ตัวอย่างหนังสือ

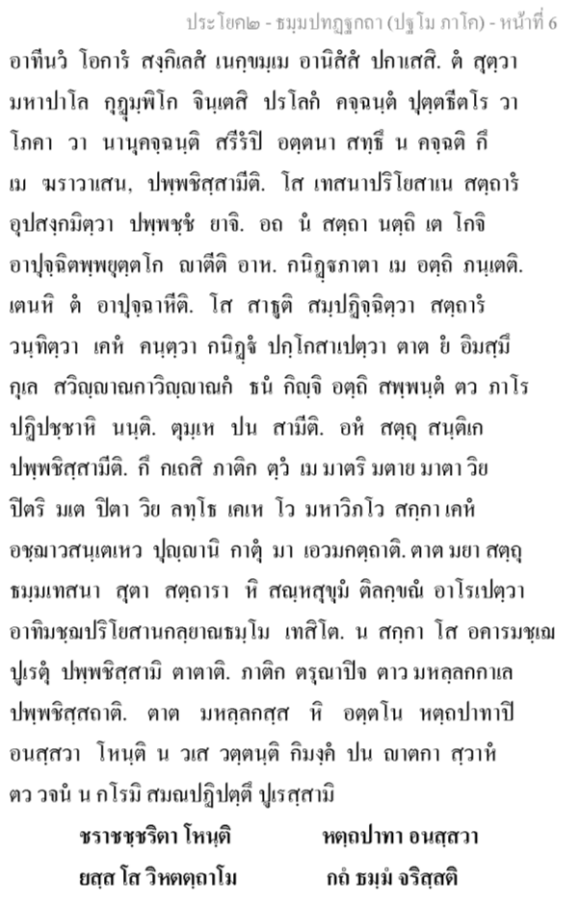


รูปที่ 7 ตัวอย่างหนังสือ (ต่อ)

ขั้นตอนการรวบรวมหนังสือที่อยู่ในรูปแบบไฟล์ข้อความ ดังต่อไปนี้

1. ดาวน์โหลดข้อมูลจาก http://www.learntripitaka.com/ โดยเลือกที่เมนู “Download พระไตรปิฎก” และ “หนังสือหมวดเปรียญประโยค ๑ – ๙” ตามลำดับ
2. แยกไฟล์แล้วจะพบไฟล์ข้อความเท่ากับจำนวนหน้าของหนังสือในเล่มนั้น ๆ เช่น ธมฺมปทฏฺฐกถา (ปฐโม ภาโค) มีทั้งหมด 148 หน้าจะมีไฟล์ข้อความ 148 ไฟล์
3. ลบข้อมูลที่ไม่เกี่ยวข้องของแต่ละไฟล์ออกได้แก่ และ ส่วน <B></B> ไม่ได้ถูกลบออกเพราะใช้บอกขอบเขตของเนื้อความที่ถูกประพันธ์เป็นร้อยกรอง
4. ลบรหัสอักขระที่ต่าง ๆ ที่ไม่พบในการเข้ารหัสแบบ TIS620 หรือ Window874 และแปลงตัวอักษร ฌ และ ฯ ที่ไม่มีเชิง ตามที่ปรากฏในภาษาบาลีอักษรไทย ให้กลายเป็น ญ และ ฐ ตามรูปแบบพยัญชนะที่มีเชิง เพราะอักษรสองตัวนั้นมีรหัสตัวอักษรที่ไม่ได้ใช้จึงแสดงผลไม่ได้
5. แก้ไขคำขาด ซึ่งปรากฏเครื่องหมาย – ที่ท้ายบรรทัดเนื่องจากถูกตัดขึ้นบรรทัดใหม่ และถูกเติม \n หลัง – เช่น อุปติสฺส-\nคาโม แก้ไขเป็น อุปติสฺสคาโม

เนื่องจากคำสมาสและสนธิ สามารถตัดคำให้กลายเป็นคำที่มีความหมายได้หลากหลายรูปแบบ ดังนั้นจึงได้รวบรวมเป็นไฟล์ข้อความก่อนนำส่งให้ผู้เชี่ยวชาญค้นหาคำสนธิและตัดคำสนธิโดยพิจารณาจากเนื้อความ ดังแสดงตัวอย่างไฟล์ข้อความในรูปที่ 8



ประโยค๒ - ธมฺมปทฏฐกถา (ปฐโม ภาโค) - หน้าที่ 6

อาทีนวๆ โอการๆ สงฺกิเลสๆ เนกฺขมฺเม อานิสๆสๆ ปกาเสสิ. ตๆ สุตฺวา มหาปาโล กุฏุมฺพิโก จินฺเตสิ ปรโลกๆ คจฺฉนฺตๆ ปุตฺตธีตโร วา โภคา วา นานุคจฺฉนฺติ สรีรๆปิ อตฺตนา สทฺธึ น คจฺฉติ กึ เม ฆราวาเสน, ปพฺพชิสฺสามีติ. โส เทสนาปริโยสาเน สตฺถารๆ อุปสงฺกมิตฺวา ปพฺพชฺช ยาจิ. อถ นๆ สตฺถา นตฺถิ เต โกจิ อาปุจฺฉิตพฺพยุตฺตโก ญาตีติ อาห. กนิฏฺฐภาตา เม อตฺถิ ภนฺเตติ. เตนหิ ตๆ อาปุจฺฉาหีติ. โส สาธูติ สมฺปฏิจฺฉิตฺวา สตฺถารๆ วนฺทิตฺวา เคหๆ คนฺตฺวา กนิฏฺฐๆ ปกฺโกสาเปตฺวา ตาต ยๆ อิมสฺมึ กุเล สวิญฺญาณกาวิญฺญาณกๆ ธนๆ กิญฺจิ อตฺถิ สพฺพนฺตๆ ตว ภาโร ปฏิปชฺชาหิ นนฺติ. ตุมฺเห ปน สามีติ. อหๆ สตฺถุ สนฺติเก ปพฺพชิสฺสามีติ. กึ กเถสิ ภาติก ตฺวๆ เม มาตริ มตาย มาตา วิย ปิตริ มเต ปิตา วิย ลทฺโธ เคเห โว มหาวิภโว สกฺกา เคหๆ อชฺฌาวสนฺเตเหว ปุญฺญานิ กาตุๆ มา เอวมกตฺถาติ. ตาต มยา สตฺถุ ธมฺมเทสนา สุตา สตฺถารา หิ สญฺหสุขุมๆ ปิลกฺขณๆ อาโรเปตฺวา อาทิมชฺฌปริโยสานกลฺยาณธมฺโม เทสิโต. น สกฺกา โส อคารมชฺเฌ ปูเรตุๆ ปพฺพชิสฺสามิ ตาตาติ. ภาติก ตรุณาปิจ ตาว มหลฺลกกาเล ปพฺพชิสฺสถาติ. ตาต มหลฺลกสฺส หิ อตฺตโน หตฺถปาทาปิ อนสฺสวา โหนฺติ น วเส วตฺตนฺติ กิมงฺคๆ ปน ญาตกา สุวาหๆ ตว วจนๆ น กโรมิ สมณปฏิปตฺตึ ปูเรสฺสามิ

ชราชชฺชริตา โหนฺติ หตฺถปาทา อนสฺสวา
ยสฺส โส วิหตตฺถาโม กถๆ ธมฺมๆ จริสฺสติ

รูปที่ 8 ไฟล์ข้อความจากหนังสือธมฺมปทฏฐกถา (ปฐโม ภาโค) หน้าที่ 6

ชุดข้อมูลคำสนธิที่ไม่ซ้ำกันจำนวน 6,845 คำ โดยรวบรวมจากผู้เชี่ยวชาญ สามารถแยกออกเป็นคำเดี่ยวได้สูงสุด 5 คำ แสดงตัวอย่างในตารางที่ 22

ตารางที่ 22 ตัวอย่างข้อมูลคำสนธิและผลเฉลย

คำสนธิ	**ผลเฉลย (คำตัด)**				
	คำที่ 1	**คำที่ 2**	**คำที่ 3**	**คำที่ 4**	**คำที่ 5**
คณฺหิตุนฺติ	คณฺหิตุๆ	อิติ			
อิทญฺจิทญฺจ	อิทๆ	จ	อิทๆ	จ	
อิทญฺจิทญฺจาติ	อิทๆ	จ	อิทๆ	จ	อิติ
ปจฺเจกพุทฺโธปิสฺส	ปจฺเจกพุทฺโธ	ปิ	อสฺส		

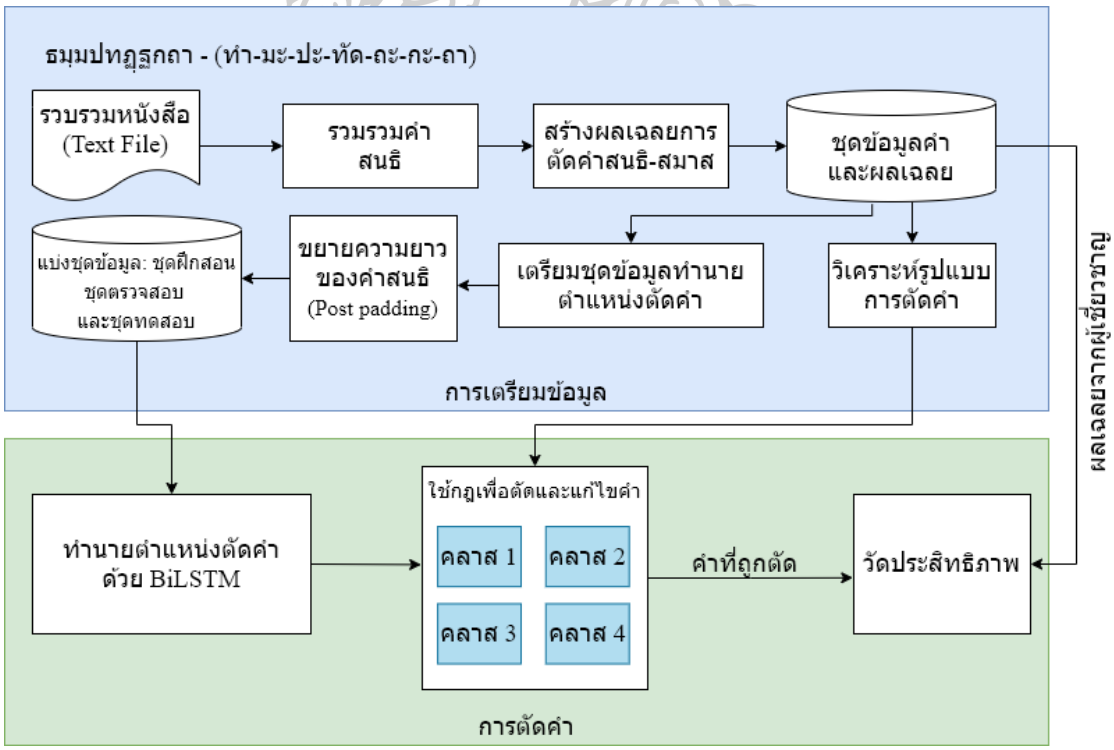
ชุดข้อมูลคำสมาสที่ไม่ซ้ำกันจำนวน 4,478 คำ โดยรวบรวมจากผู้เชี่ยวชาญ ซึ่งพบคำสนธิที่สามารถเป็นคำเดี่ยวในประโยคสูงสุด 7 คำ แสดงตัวอย่างในตารางที่ 23

ตารางที่ 23 ตัวอย่างข้อมูลคำสมาสและผลเฉลย

คำสมาส	ผลเฉลย (คำตัด)						
	ศัพท์ที่ 1	ศัพท์ที่ 2	ศัพท์ที่ 3	ศัพท์ที่ 4	ศัพท์ที่ 5	ศัพท์ที่ 6	ศัพท์ที่ 7
วตฺถเภสชฺชปานกาทิ	วตฺถ	เภสชฺช	ปานก	อาทิ			
สปริวาร	สห	ปริวาร					
ธมฺมสฺสามิ	ธมฺม	สามิ					
ธมฺมกถาทิ	ธมม	กถา	อาทิ				

4.2 การตัดคำสนธิ

ภาพรวมการวิจัยตัดคำสนธิดังแสดงในรูปที่ 9 เริ่มตั้งแต่การรวบรวมหนังสือเป็นไฟล์ดิจิทัล และขอความอนุเคราะห์จากผู้เชี่ยวชาญค้นหาคำสนธิพร้อมทั้งตัดคำสนธิเพื่อสร้างเป็นชุดข้อมูลคำสนธิและผลเฉลย ซึ่งคำสนธิสามารถตัดได้หลายแบบ แต่แบบที่ถูกต้องจะต้องถูกตัดโดยพิจารณาจากเนื้อความ จึงได้วิเคราะห์รูปแบบการตัดคำพบว่าคำสนธิที่แยกจากกันโดยมีทั้งแก้ไขและไม่แก้ไข และจึงได้สร้างกฎเพื่อใช้สำหรับปรับปรุงคำที่ต้องแก้ไข



รูปที่ 9 ภาพรวมการวิจัยการตัดคำสนธิ

เมื่อออกแบบรูปแบบการตัดคำและกฎที่ใช้แก้ไขเรียบร้อยแล้ว จึงได้พิจารณาแนวคิดที่สามารถตัดคำสนธิที่เกิดจากการเชื่อมหลายคำเข้าด้วยกัน โดยปกตินักเรียนบาลีจะตัดครั้งเดียวเช่น คำสนธิว่า “อิทญฺจิทญฺจิ” (อิ-ทัน-จิ-ทัน-จะ) ตัดเป็น “อิทำ จ อิทำ จ” (อิ-ทัง จะ อิ-ทัง จะ) แต่กฎที่ได้ออกแบบไว้ก่อนหน้าสำหรับตัดคำสนธิแล้วได้ผลลัพธ์จากการคำตัดสองคำ ดังนั้นจึงต้องพิจารณาว่าจะตัดคำสนธิจากแบบที่ 1 หรือแบบที่ 2 ซึ่งในวิทยานิพนธ์เล่มนี้ได้ใช้รูปแบบที่ 1

1. แบบที่ 1 คำสนธิ → คำสนธิ + คำ | คำ + คำ
2. แบบที่ 2 คำสนธิ → คำ + คำสนธิ | คำ + คำ

เนื่องจากกฎการแก้ไขที่จะกล่าวถึงในหัวข้อ 4.2.4 การวิเคราะห์กฎสำหรับรูปแบบการตัดคำสนธิ มีทั้งส่วนการแก้ไขคำที่เพิ่มอักษรด้านหน้าและเพิ่มอักษรด้านหลัง กรณีที่แก้ไขโดยเพิ่มอักษรด้านหน้าจะทำให้ตำแหน่งของอักษรขยับไปทางขวา และถ้าอักษรตัวนั้นเป็นตำแหน่งตัดคำจะทำให้พบข้อผิดพลาดดังนั้นแนวทางการตัดคำสนธิในเล่มนี้จึงเป็น

แบบที่ 1 คำสนธิ → คำสนธิ + คำ + คำ | คำสนธิ + คำ | คำ + คำ + คำ | คำ + คำ

หลังจากที่ทราบรูปแบบการวิเคราะห์และออกแบบที่ใช้สำหรับแก้ไขคำแล้ว จึงได้เตรียมชุดข้อมูลทำนายตำแหน่ง โดยนำคำสนธิแต่ละคำมาระบุตำแหน่งและรูปแบบตัดคำ เพื่อฝึกสอนโมเดลทำนายตำแหน่งตัดคำ จากนั้นจึงนำผลลัพธ์จากการทำนายดังกล่าวไปใช้กฎเพื่อแก้ไขคำและวัดประสิทธิภาพโดยนำคำที่ถูกตัดมาเปรียบเทียบกับชุดข้อมูลคำสนธิและผลเฉลยของผู้เชี่ยวชาญ

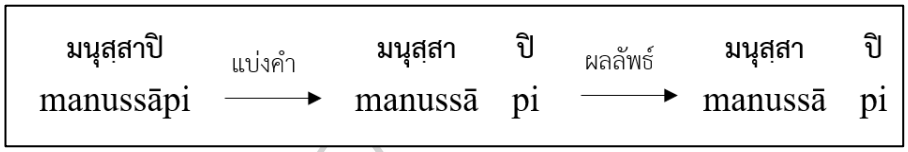
4.2.1 การวิเคราะห์รูปแบบการแยกคำสนธิ

ความแตกต่างของระบบการเขียนภาษาบาลีอักษรไทยและอักษรโรมัน ส่งผลให้รูปแบบการตัดคำสนธิมีความซับซ้อนมากยิ่งขึ้นดังนี้

- ภาษาบาลีอักษรโรมัน มีรูปสระเพียงอักษรเดียว (a, ā, i, ī, u, ū, e, o) แต่ภาษาบาลีอักษรไทยใช้ อ เข้ามาแทนการกำกับสระในพยางค์ที่ไม่มีเสียงพยัญชนะต้น (อ, อา, อิ, อี, อุ, อู, เอ, โอ) และถ้ามีเสียงพยัญชนะต้น จะใช้รูปของพยัญชนะต้นแทนอักษร อ
- ภาษาบาลีอักษรโรมัน มีรูปแบบการเขียนตามลำดับการถ่ายเสียง คือ พยัญชนะต้น สระ และตัวสะกด แต่ภาษาบาลีอักษรไทยมีรูปแบบการเขียนคำไทยคือไม่ตรงกับการถ่ายเสียง เพราะมีสระหน้า (สระเอ และ สระโอ) เช่น เทโว - Devo

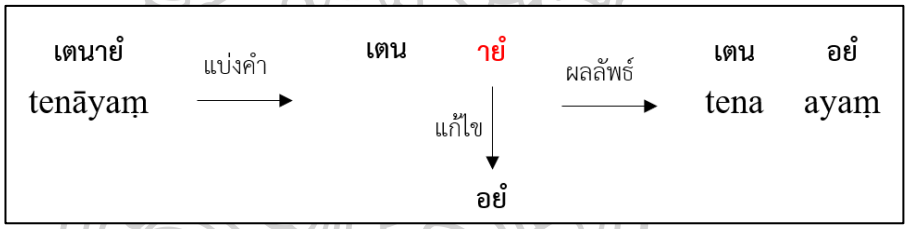
จากการวิเคราะห์ชุดข้อมูลสนธิและผลเฉลยตารางที่ 22 จึงกำหนดรูปแบบการตัดคำสนธิเป็น 4 ประเภท โดยจะแสดงอักษรบาลีโรมันกำกับไว้แทนคำอ่าน เพื่อให้เห็นชัดถึงความแตกต่างของรูปแบบการเขียน ซึ่งทำให้การวิเคราะห์การตัดคำภาษาบาลีไทยมีความซับซ้อนดังต่อไปนี

1. คำสนธิมีตำแหน่งตัดคำที่สามารถแยกออกเป็นสองคำ โดยได้คำที่ถูกต้องและมีความหมายทั้งสองคำ (รูปที่ 10)



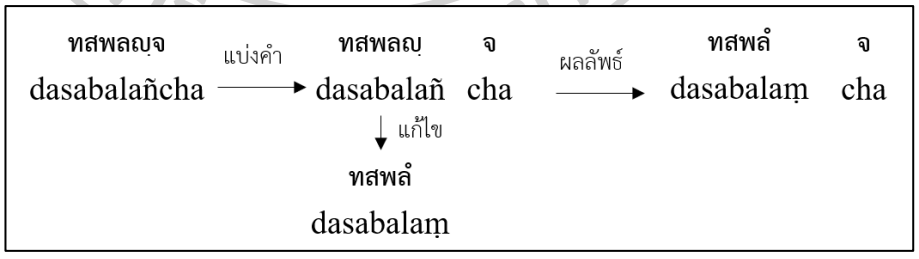
รูปที่ 10 การแยกคำสนธิรูปแบบที่ 1

2. คำสนธิมีตำแหน่งตัดคำที่สามารถแยกออกเป็นสองคำ คำแรกเป็นคำที่ถูกต้องและมีความหมาย แต่คำหลังต้องแก้ไข (รูปที่ 11)



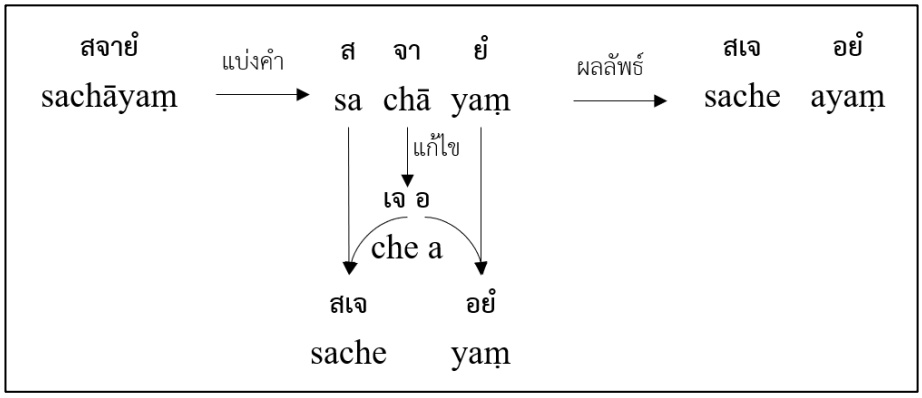
รูปที่ 11 การแยกคำสนธิรูปแบบที่ 2

3. คำสนธิมีตำแหน่งตัดคำที่สามารถแยกออกเป็นสองคำ คำหลังเป็นคำที่ถูกต้องและมีความหมาย แต่คำแรกต้องแก้ไข (รูปที่ 12)



รูปที่ 12 การแยกคำสนธิรูปแบบที่ 3

4. คำสนธิไม่มีตำแหน่งตัดคำที่สามารถแยกคำที่มีความหมายได้ จึงกำหนดให้มีตำแหน่งตัดคำสองจุด และปรับปรุงข้อความภายในตำแหน่งตัดคำ (รูปที่ 13)



รูปที่ 13 การแยกคำสนธิรูปแบบที่ 4

4.2.2 การทำนายตำแหน่งและรูปแบบตัดคำสนธิ

การทำนายตำแหน่งและรูปแบบตัดคำสนธิ คือ การทำนายตัวอักษรในคำสนธิว่าตัวอักษรใดที่เป็นตำแหน่งตัดคำ และตัวอักษรที่เป็นตำแหน่งตัดคำนั้นใช้รูปแบบตัดคำใด รูปแบบตัดคำประเภทที่ 1 – 3 มีตำแหน่งตัดคำเพียงจุดเดียว แต่รูปแบบตัดคำประเภทที่ 4 มีตำแหน่งตัดคำสองจุด ดังนั้นจะแทนการตัดคำครั้งที่ i ของคำสนธิ s ด้วย $(r,p,l)_i$ โดย r คือรูปแบบการแยกคำ, p คือ ตำแหน่งตัดคำ และ l คือ ความยาวของข้อความในพื้นที่ตัดคำของรูปแบบการแยกคำประเภทที่ 4 แต่รูปแบบประเภทที่ 1 – 3 ไม่ต้องระบุค่านี้

คำสนธิ s สามารถตัดคำได้มากกว่า 1 ครั้ง ดังนั้นจะมีตำแหน่งตัดคำเป็น $[(r,p,l)_n,\dots(r,p,l)_1]$ ในรูปที่ 14 (ก) และรูปที่ 14 (ข) แสดงตัวอย่างคำสนธิที่ต้องตัดคำตามรูปแบบประเภทที่ 1 และ 4 ส่วนรูปที่ 14 (ค) แสดงตัวอย่างคำสนธิที่ตัดคำสนธิมากกว่า 1 ครั้ง

คำสนธิ	ม	น	ุ	ส	◌	ส	า	ป	ิ
ตำแหน่ง	0	1	2	3	4	5	6	7	8
กฎ								1	
รูปอย่างง่าย	[(1,7)]								

รูปที่ 14 (ก) ตำแหน่งตัดคำรูปแบบที่ 1

คำสนธิ	ช	า	เ	ม	ต	◌
ตำแหน่ง	0	1	2	3	4	5
กฎ			4	4		
รูปอย่างง่าย	[(4,2,2)]					

รูปที่ 14 (ข) ตำแหน่งตัดคำรูปแบบที่ 4

คำสนธิ	ส	ห	า	ย	โ	ก	ป	ิ	ส	.	ส
ตำแหน่ง	0	1	2	3	4	5	6	7	8	9	10
กฎ							1		2		
รูปอย่างง่าย	[(2,8) , (1,6)]										

รูปที่ 14 (ค) คำสนธิที่มีตำแหน่งตัดคำมากกว่า 1 ครั้ง

รูปที่ 14 ตัวอย่างการเตรียมผลเฉลยตำแหน่งและรูปแบบตัดคำสนธิ

การทำนายตำแหน่งและรูปแบบการตัดคำสนธิ s และมีผลเฉลย o ที่ $|s| = |o|$ และ $o_i \in [0,5]$ ซึ่งแสดงคำอธิบายไว้ในตารางที่ 24 ส่วนตารางที่ 25 แสดงตัวอย่างคำสนธิและผลเฉลยตำแหน่งและรูปแบบตัดคำสนธิ

ตารางที่ 24 คำอธิบายตำแหน่งและรูปแบบตัดคำสนธิ

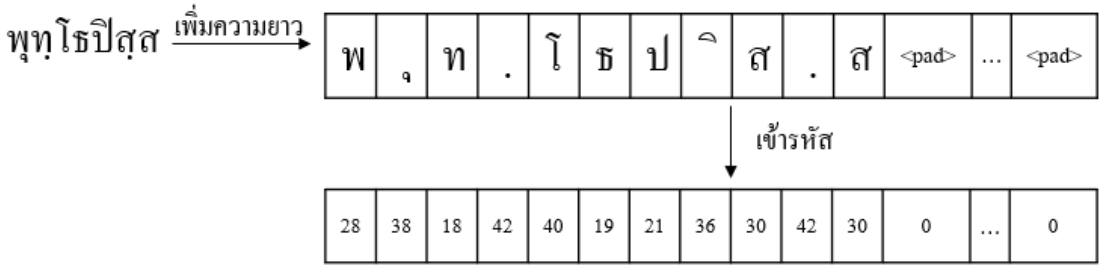
ผลเฉลย	**คำอธิบาย**
0	ไม่ใช่ตำแหน่งตัดคำ
1	เป็นตำแหน่งตัดคำ และตัดคำสนธิด้วยรูปแบบที่ 1
2	เป็นตำแหน่งตัดคำ และตัดคำสนธิด้วยรูปแบบที่ 2
3	เป็นตำแหน่งตัดคำ และตัดคำสนธิด้วยรูปแบบที่ 3
4	เป็นตำแหน่งตัดคำ และตัดคำสนธิด้วยรูปแบบที่ 4
5	ส่วนขยายความยาว (Post Padding)

ตารางที่ 25 ข้อมูลและผลเฉลยประเภทตำแหน่งตัดคำสนธิ

คำสนธิ	**ผลเฉลยรูปย่อ**	**ผลเฉลย (ความยาว = 57)**
มนุสฺสาปิ (มะ-นุด-สา-ปิ)	[(1,7)]	[0, 0, 0, 0, 0, 0, 0, 1, 0, 5, 5, 5, 5, 5, …, 5]
อถสฺสาหํ (อะ-ถัด-สา-หัง)	[(2,5), (2,2)]	[0, 0, 2, 0, 0, 2, 0, 0, 5, 5, 5, 5, 5, 5, …, 5]
อิทญฺจิทญฺจ (อิ-ทัน-จิ-ทัน-จะ)	[(3,10), (2,6), (3,5)]	[0, 0, 0, 0, 0, 3, 2, 0, 0, 0, 3, 5, 5, 5, …, 5]
อิทญฺจิทญฺจาติ (อิ-ทัน-จิ-ทัน-จา-ติ)	[(2,11), (3,10) , (2,6), (3,5)]	[0, 0, 0, 0, 0, 3, 2, 0, 0, 0, 3, 2, 0, 0, …, 5]
พุทฺโธปิสฺส (พุด-โธ-ปิด-สะ)	[(2, 8), (1, 6)]	[0, 0, 0, 0, 0, 0, 1, 0, 2, 5, 5, 5, 5, 5, …, 5]

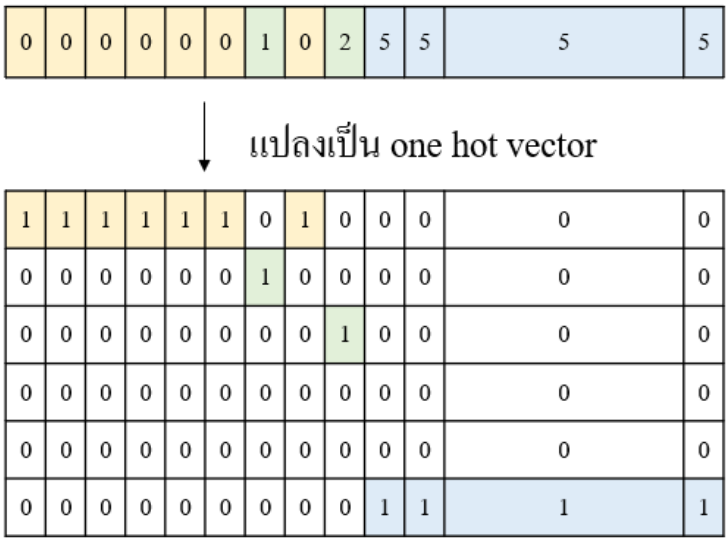
4.2.3 โมเดลทำนายตำแหน่งและรูปแบบตัดคำสนธิ

จากชุดข้อมูลที่เตรียมไว้สำหรับทำนายประเภทของตำแหน่งตัดคำ เริ่มจากปรับคำสนธิให้มีความยาวเท่ากับคำสนธิที่ยาวที่สุด (Post Padding) จากนั้นเข้ารหัสตัวอักษรในคำสนธิที่อยู่ในรูปของจำนวนเต็ม ดังรูปที่ 15



รูปที่ 15 การเข้ารหัสคำสนธิก่อนป้อนเข้าโมเดลทำนายประเภทตำแหน่งตัดคำ

ผลเฉลยจะถูกขยายด้วยเลขศูนย์ (Zero Padding) แล้วปรับให้ผลเฉลยตำแหน่งตัดคำ ที่อยู่ในรูปแบบ scalar ให้กลายเป็นรูปแบบ one-hot vector ดังรูปที่ 16 แสดงตัวอย่างการแปลงผลเฉลยตำแหน่งตัดคำของคำว่า พุทฺโธปิสฺส (พุด-โธ-ปิด-สะ)



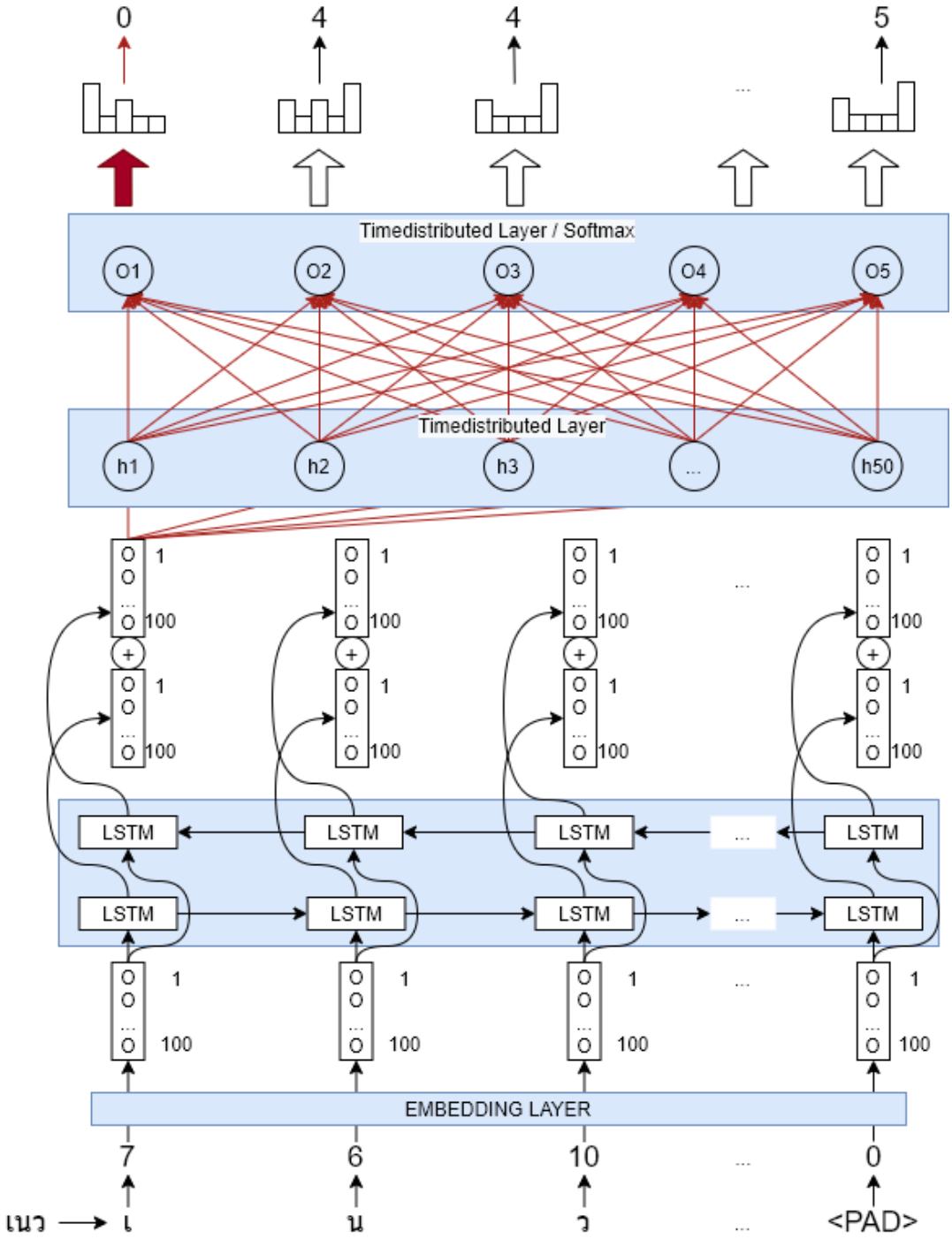
รูปที่ 16 นำผลเฉลยตำแหน่งมาแปลงเป็น one-hot vector

ตัวอักษรที่ใช้ได้ทั้งหมดมีจำนวน 44 ตัว ได้แก่ รูปพยัญชนะจำนวน 32 ตัว รูปสระ 7 ตัว (ภาษาบาลีมีเสียงสระ 8 เสียง แต่มีเพียง 7 รูป เพราะไม่มีรูปสระอะ และใช้รูป อ แทนเสียงพยางค์ที่ไม่มีพยัญชนะต้น) รูปพินทุ 1 ตัว และอักษรพิเศษจำนวน 3 ตัว ดังแสดงในตารางที่ 26

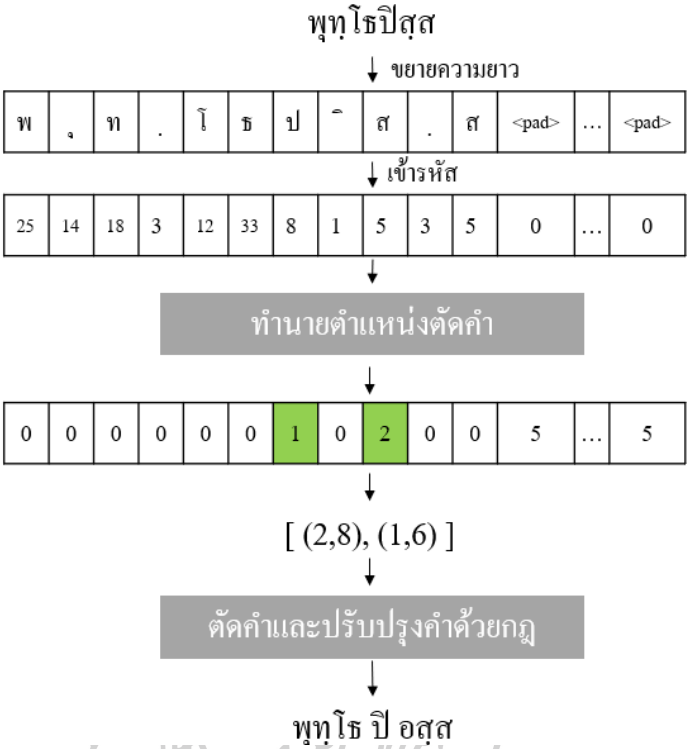
ตารางที่ 26 ตัวอักษรทั้งหมดที่ใช้

รูปพยัญชนะ	ก ข ค ฆ ง จ ฉ ช ฌ ญ ฏ ฐ ฑ ฒ ณ ต ถ ท ธ น ป ผ พ ภ ม ย ร ล ว ส ห ฬ
รูปสระ	า ิ ี ุ ู เ โ
รูปนิคหิต	◌ं
รูปพินทุ	◌ฺ
อักษรพิเศษ 1	อ (ใช้แทนพยางค์ที่ไม่มีเสียงพยัญชนะต้น)
อักษรพิเศษ 2	◌ึ (ใช้แทน ◌ิ◌ं)
อักษรพิเศษ 3	(ใช้เพื่อเป็นสัญลักษณ์ส่วนขยาย)

งานวิจัยนี้ทำนายตำแหน่งและรูปแบบตัดคำสนธิในระดับตัวอักษร ภาพโครงสร้างของโมเดลแสดงในรูปที่ 17 เมื่อทำนายตำแหน่งและรูปแบบตัดคำสนธิจากโครงสร้างโมเดลที่แสดงในรูปที่ 17 แล้วจะใช้กฎของแต่ละรูปแบบเพื่อตัดคำสนธิและปรับปรุงคำให้มีความหมาย ดังแสดงขั้นตอนการทำงานรวมของการตัดคำสนธิได้ดังแสดงในรูปที่ 18



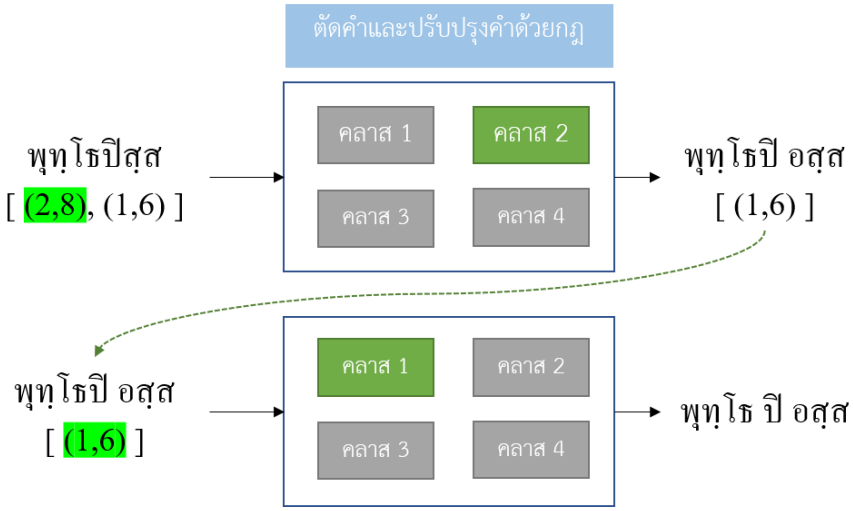
รูปที่ 17 โมเดลทำนายตำแหน่งและรูปแบบตัดคำสนธิ



รูปที่ 18 ภาพรวมการตัดคำสนธิ

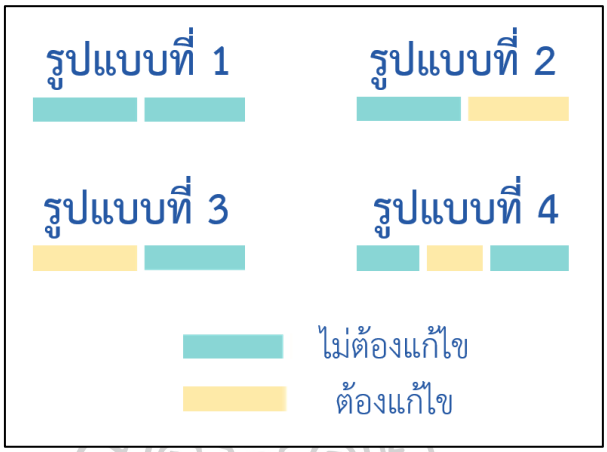
4.2.4 การวิเคราะห์กฎสำหรับรูปแบบการตัดคำสนธิ

จากรูปที่ 18 เมื่อทำนายตำแหน่งและรูปแบบตัดคำสนธิแล้ว จะนำคำสนธิ “**พุทฺโธปิสฺส**” ที่มีผลการทำนายตัดคำเป็น [(2,8), (1,6)] ซึ่งต้องใช้กฎเพื่อตัดคำสองครั้ง โดยครั้งแรกได้ผลลัพธ์เป็น “**พุทฺโธปิ อสฺส**” และเหลือตำแหน่งตัดคำคือ [(1,6)] จากนั้นเมื่อตัดคำด้วยกฎซ้ำอีกครั้งจะได้ผลลัพธ์เป็น “**พุทฺโธ ปิ อสฺส**” ไปใช้กฎที่เตรียมเอาไว้ดังรูปที่ 19



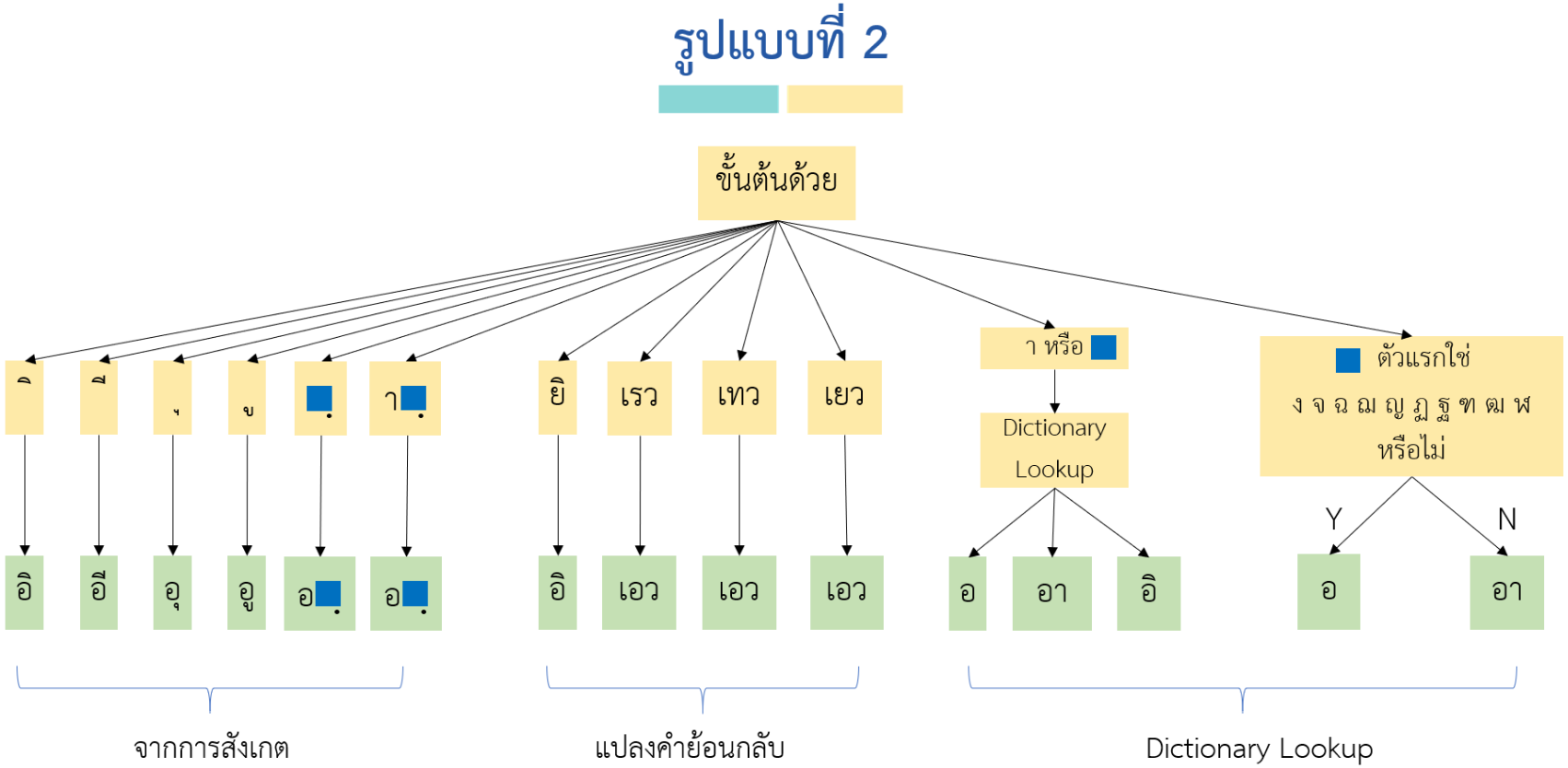
รูปที่ 19 ตัวอย่างการตัดคำสนธิโดยใช้ผลการทำนายตำแหน่งและรูปแบบการตัดคำ

จากรูปแบบการตัดคำที่วิเคราะห์ไว้ในหัวข้อ 4.2.1 สามารถแสดงภาพสรุปรูปแบบการตัดคำได้ดังรูปที่ 20 โดยแต่ละรูปแบบมีกฎการตัดและปรับปรุงคำดังต่อไปนี้



รูปที่ 20 สรุปการแบ่งคำและการแก้ไขคำสนธิ

1. รูปแบบการตัดคำสนธิประเภทที่ 1 ไม่มีกฎ เนื่องจากรูปแบบนี้สามารถแบ่งคำที่ถูกต้องและมีความหมายได้ทั้งสองคำ
2. กฎสำหรับรูปแบบการตัดคำสนธิประเภทที่ 2 เป็นกฎสำหรับปรับปรุงคำหลังให้ถูกต้องและมีความหมาย ดังแสดงกฎในรูปที่ 21



รูปที่ 21 กฎการตัดคำสนธิรูปแบบที่ 2

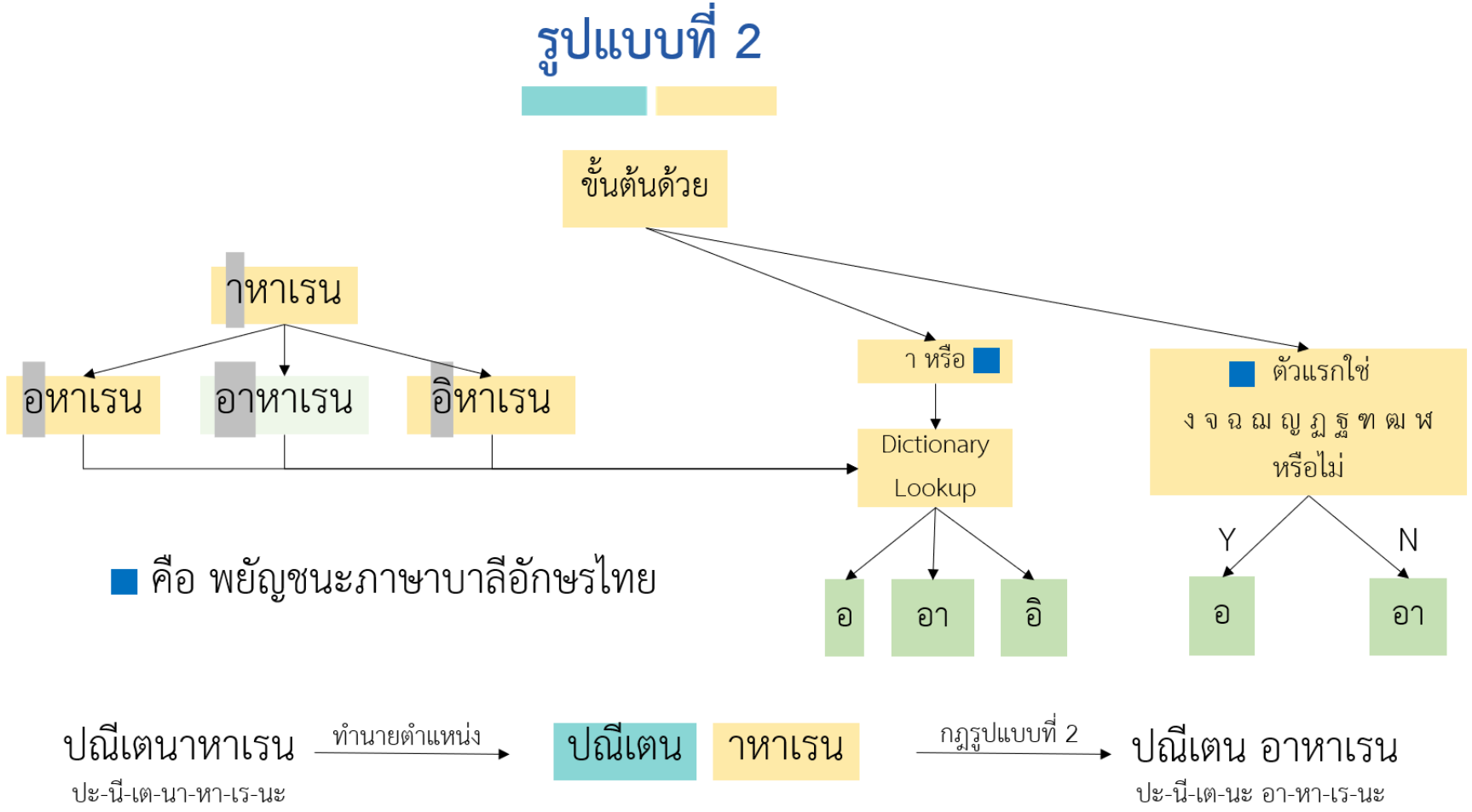
รูปที่ 22 แสดงตัวอย่างการตัดคำสนธิ “**นยิทำ**” ซึ่งสามารถแบ่งคำจากโมเดลทำนายตำแหน่งตัดคำได้เป็น “**น**” กับ “**ยิทำ**” ซึ่งคำส่วนหลัง “**ยิทำ**” ขึ้นต้นด้วย “**ยิ**” สอดคล้องกับกฎที่แสดงไว้ในรูปที่ 21 จึงเปลี่ยน “**ยิ**” เป็น “**อิ**” ดังนั้นจะได้ผลลัพธ์จากการตัดคำสนธิเป็น “**น ยิทำ**”

รูปที่ 22 การแก้ไขคำด้วยกฎการตัดคำสนธิรูปแบบที่ 2

จากรูปที่ 21 สำหรับกฎ Dictionary Lookup ที่นำมาใช้กับกฎการตัดคำสนธิรูปแบบที่ 2 จะต้องค้นหาว่าควรแก้ไขเสียงพยางค์แรกเป็น อ อา หรือ อิ โดยทดลองแก้ไขแล้วนำไปเปรียบเทียบกับกลุ่มที่ไม่เปลี่ยนรูปและกลุ่มคำที่เปลี่ยนรูป

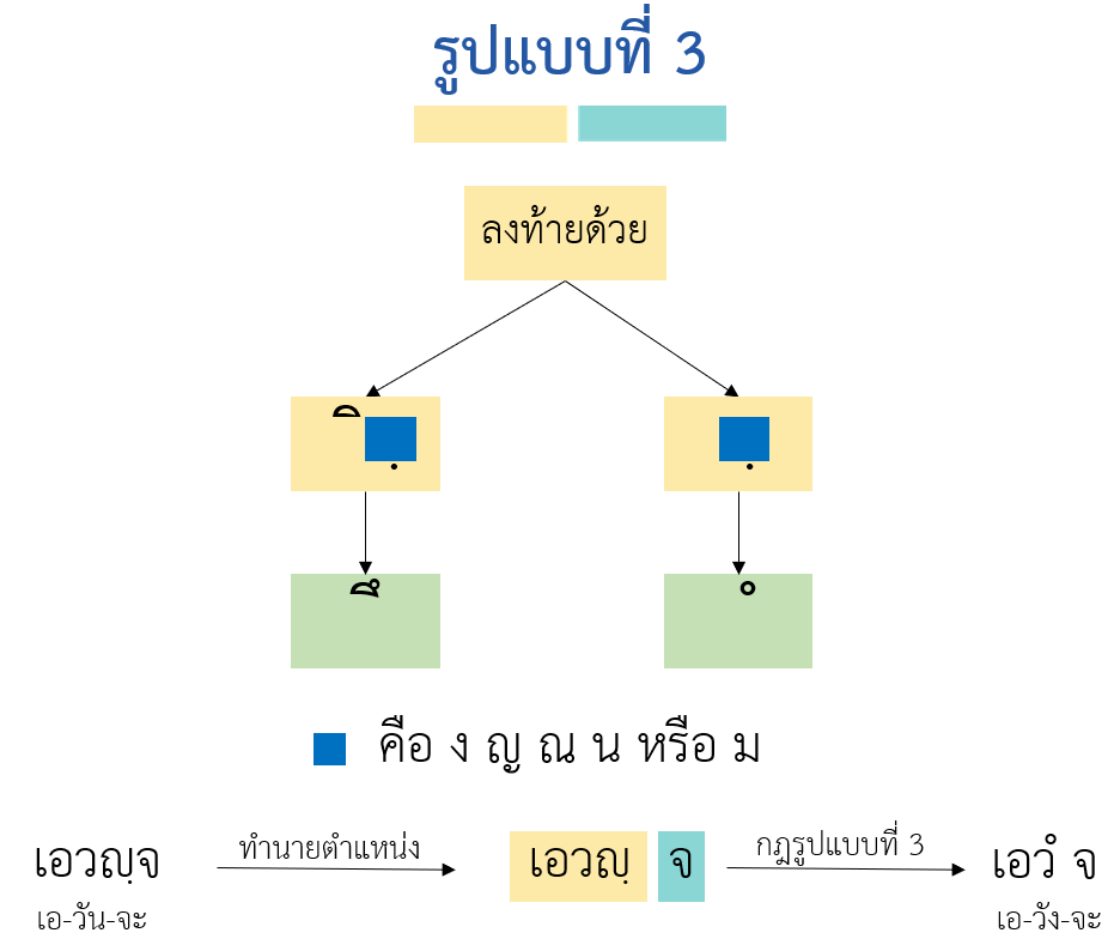
กลุ่มคำที่ไม่เปลี่ยนรูปได้แก่ นิบาต ปัจจัยในอัพยยศัพท์ คำสรรพนาม และคำกริยาอาขยาตของ อสฺ ธาตุ ส่วนกลุ่มคำที่เปลี่ยนรูปนั้นพิจารณาเฉพาะคำที่เสียงพยางค์แรกขึ้นต้น อ อา และ อิ จากพจนานุกรม (วิทยานิพนธ์นี้ใช้พจนานุกรมไทย - บาลี ฉบับภูมิพโล)

รูปที่ 23 แสดงตัวอย่างการตัดคำสนธิ “**ปณีเตนาหาเรน**” ซึ่งสามารถแบ่งคำจากโมเดลทำนายตำแหน่งได้คำเป็น “**ปณีเตน**” กับ “**าหาเรน**” จะเห็นได้ว่าคำส่วนหลัง “**าหาเรน**” นั้นขึ้นต้นด้วยรูปสระ า จึงต้องแก้ไขเสียงพยางค์แรกเป็น “**อหาเรน**” “**อาหาเรน**” และ “**อิหาเรน**” เมื่อนำไปเปรียบเทียบกับกลุ่มคำที่เปลี่ยนรูปและกลุ่มคำไม่เปลี่ยนรูปจะพบว่ามีเพียง “**อาหาเรน**” เท่านั้น ดังนั้นจะได้ผลลัพธ์จากการตัดคำสนธิ “**ปณีเตนาหาเรน**” เป็น “**ปณีเตน อาหาเรน**”



รูปที่ 23 กฎ Dictionary Lookup สำหรับการตัดคำสนธิรูปแบบที่ 2

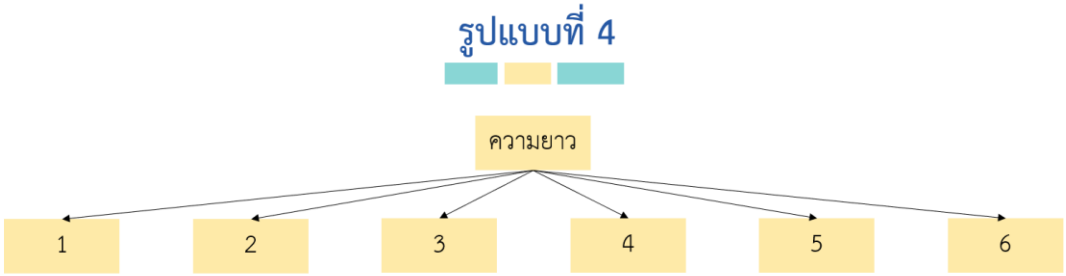
3. กฎสำหรับรูปแบบการตัดคำสนธิประเภทที่ 3 เป็นกฎสำหรับปรับปรุงคำแรกให้ถูกต้อง และมีความหมายดังแสดงกฎในรูปที่ 24



รูปที่ 24 กฎการตัดคำสนธิรูปแบบที่ 3 และตัวอย่าง

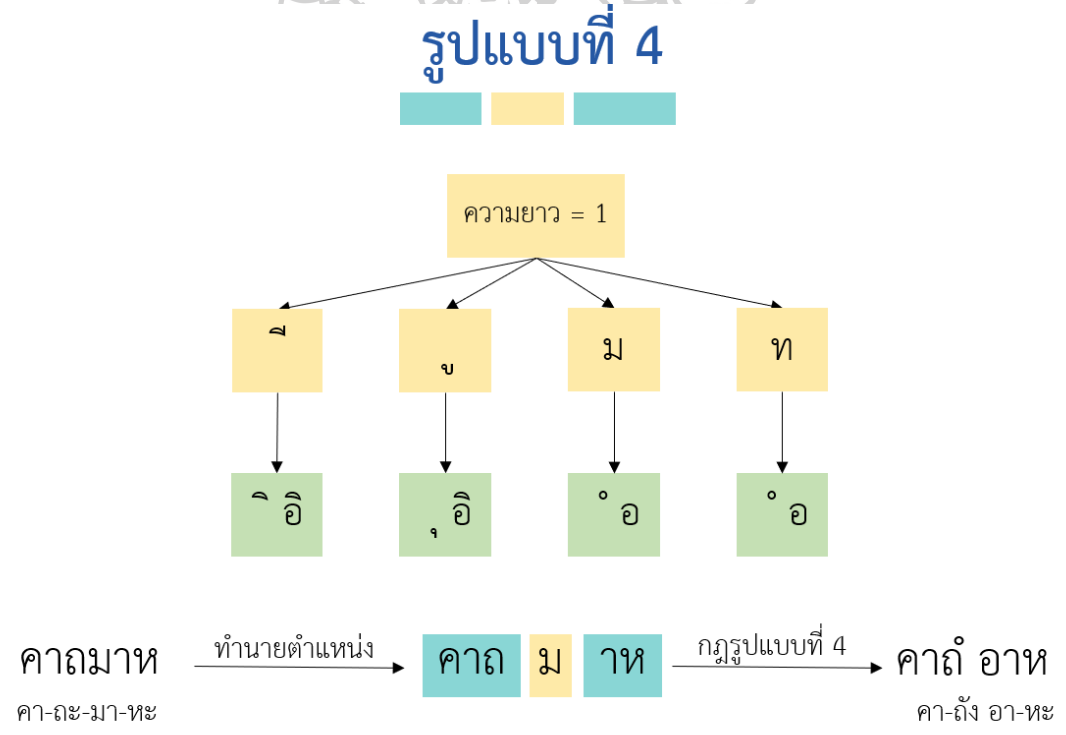
จากรูปที่ 24 แสดงตัวอย่างการตัดคำสนธิ “**เอวญ์จ**” ซึ่งสามารถแบ่งคำจากโมเดลทำนายตำแหน่งได้คำเป็น “**เอวญ์**” กับ “**จ**” ซึ่งคำส่วนแรก “**เอวญ์**” นั้นลงท้ายด้วย พยัญชนะตามเงื่อนไข คือ ญ และพินทุ จึงต้องเปลี่ยนอักษร 2 ตัวสุดท้ายเครื่องหมายนิคหิตกลายเป็น“**เอวำ**” ดังนั้นจะได้ผลลัพธ์จากการตัดคำสนธิ “**เอวญ์จ**” เป็น “**เอวำ จ**”

4. กฎสำหรับรูปแบบการตัดคำสนธิประเภทที่ 4 คือการปรับปรุงคำที่อยู่ภายในจุดตัด โดยได้แบ่งกลุ่มตามความยาวของอักษรที่อยู่ภายในจัดตัดเป็น 1 – 6 ดังรูปที่ 25



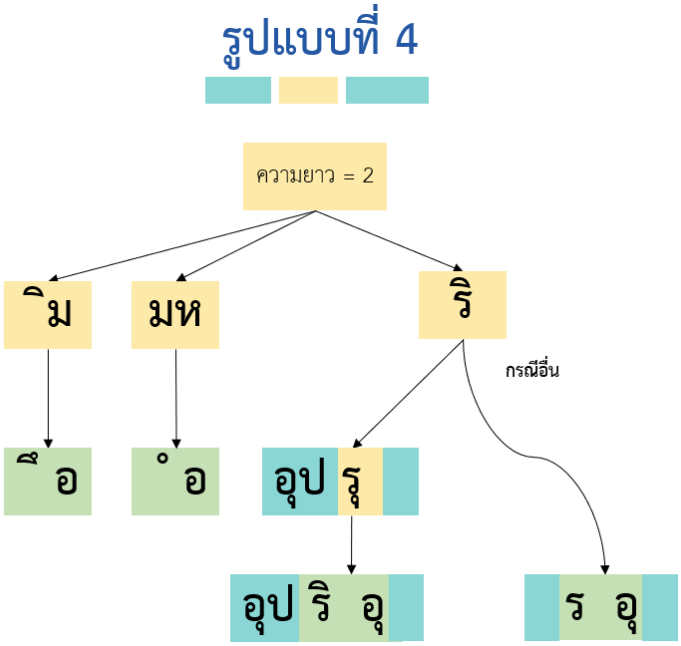
รูปที่ 25 การแบ่งกลุ่มของกฎการตัดคำสนธิรูปแบบที่ 4

รูปที่ 26 แสดงกฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่ 1 ได้แก่ 1) การเปลี่ยน “ี” เป็น “ิ อิ” 2) การเปลี่ยน “ู” เป็น “ุ อิ” และ 3) การเปลี่ยน “ม” หรือ “ท” เป็น “ำ อ” โดยยกตัวอย่างการตัดคำสนธิ “คาถมาห” ที่แบ่งคำเป็น 3 ส่วนคือ “**คาถ**” “**ม**” และ “**าห**” ซึ่งส่วนที่ 2 คือ “**ม**” จึงเปลี่ยนเป็น “ำ อ” เมื่อนำไปร่วมกับส่วนที่ 1 และส่วนที่ 3 จะได้ผลลัพธ์จากการตัดคำสนธิเป็น “**คาถำ อาห**”

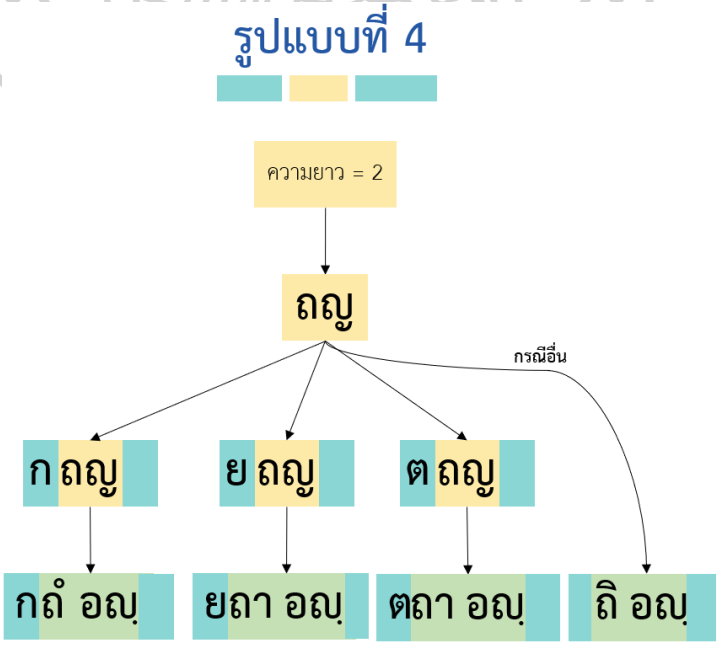


รูปที่ 26 กฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่ 1

รูปที่ 27 - รูปที่ 28 แสดงกฎการตัดสนธิที่ใช้อักษรส่วนแรกเป็นเงื่อนไขการแก้ไข จากรูปที่ 27 อักษรส่วนที่ 2 เป็น “**ริ**” และอักษรส่วนแรกเป็น “**อุป**” จะแก้ไข “**ริ**” เป็น “**ริ อุ**” แต่ถ้าอักษรแรกไม่ได้เป็น “**อุป**” จะแก้ไข “**ริ**” เป็น “**ร อุ**”

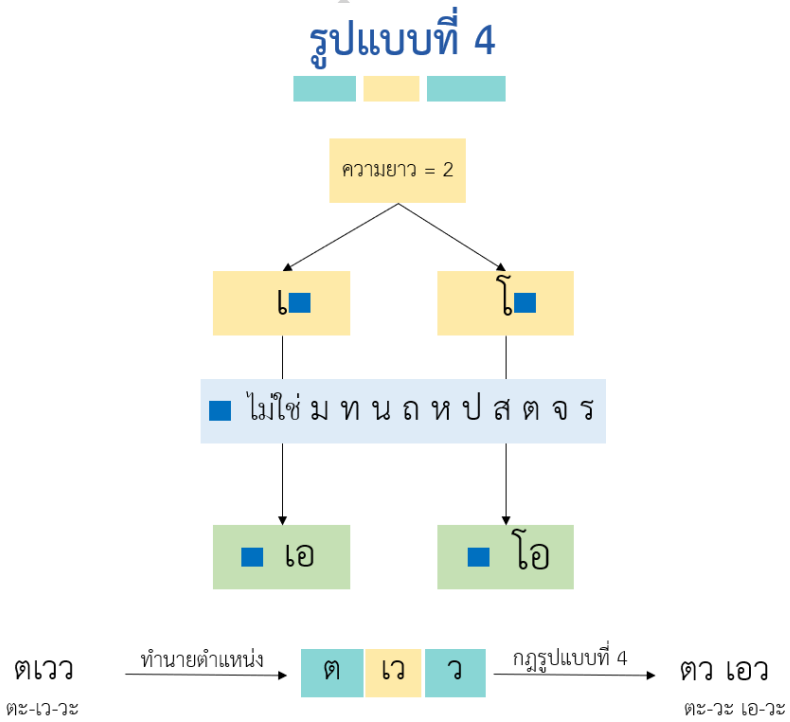


รูปที่ 27 กฎการตัดคำสนธิที่ใช้อักษรก่อนหน้า (1)



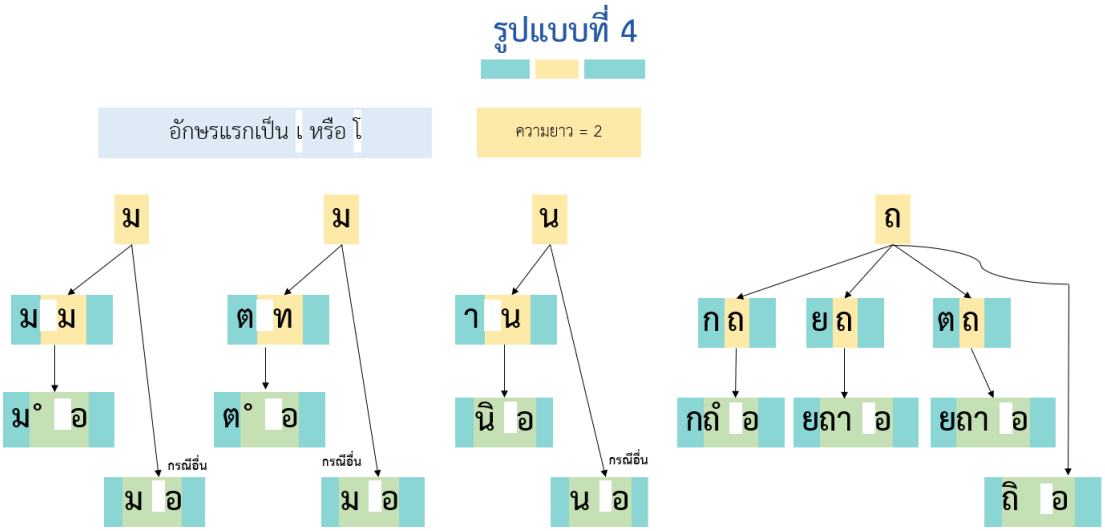
รูปที่ 28 กฎการตัดคำสนธิที่ใช้อักษรก่อนหน้า (2)

รูปที่ 29 แสดงกฎการตัดสนธิที่อักษรส่วนแก้ไขมีอักษรตัวแรกเป็นรูปสระหน้าและอักษรตัวที่สองไม่ใช่ ม ท น ถ ห ป ส ต จ หรือ ร โดยยกตัวอย่างการตัดคำสนธิ “**ตเวว**” ที่แบ่งคำเป็น 3 ส่วนคือ “**ต**” “**เว**” และ “**ว**” ซึ่งส่วนที่ 2 คือ “**เว**” มีอักษรตัวแรกเป็นสระรูปสระ เ และอักษรที่สองไม่ใช่ ม ท น ถ ห ป ส ต จ หรือ ร ดังนั้นจึงแยกเสียงพยัญชนะและสระออกจากกัน กลายเป็น “**ว เ**” โดยระบบการเขียนภาษาบาลีอักษรไทยใช้อักษร อ กำกับในพยางค์ที่ไม่มีพยัญชนะ ดังนั้นจึงจะกลายเป็น “**ว เอ**” เมื่อนำไปร่วมกับส่วนที่ 1 และส่วนที่ 3 จะได้ผลลัพธ์จากการตัดคำสนธิเป็น “**ตว เอว**”

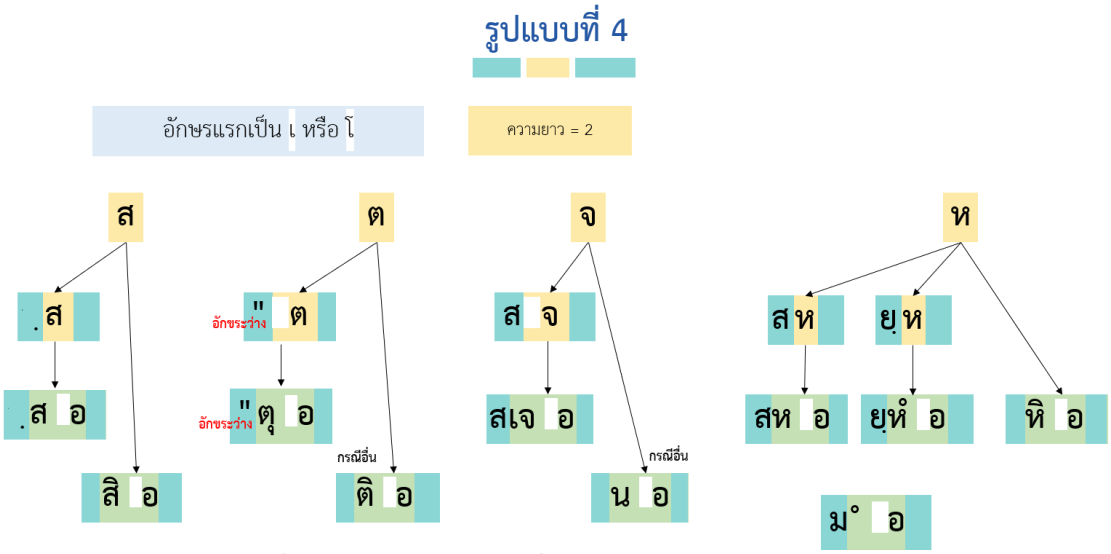


รูปที่ 29 กฎการตัดคำสนธิที่ใช้อักษรแรกเป็นสระหน้า (1)

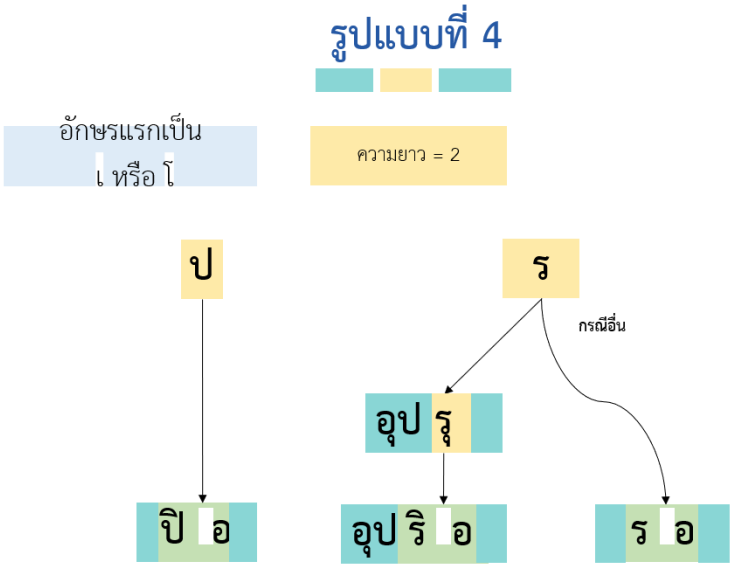
สำหรับกรณีที่อักษรส่วนแก้ไขมีอักษรตัวแรกเป็นรูปสระหน้าและอักษรตัวที่สองเป็น ม ท น ถ ห ป ส ต จ หรือ ร จะแก้ไขโดยใช้กฎดังที่ได้แสดงไว้ในรูปที่ รูปที่ 30 ถึงรูปที่ 32



รูปที่ 30 กฎการตัดคำสนธิที่ใช้อักษรแรกเป็นสระหน้า (2)

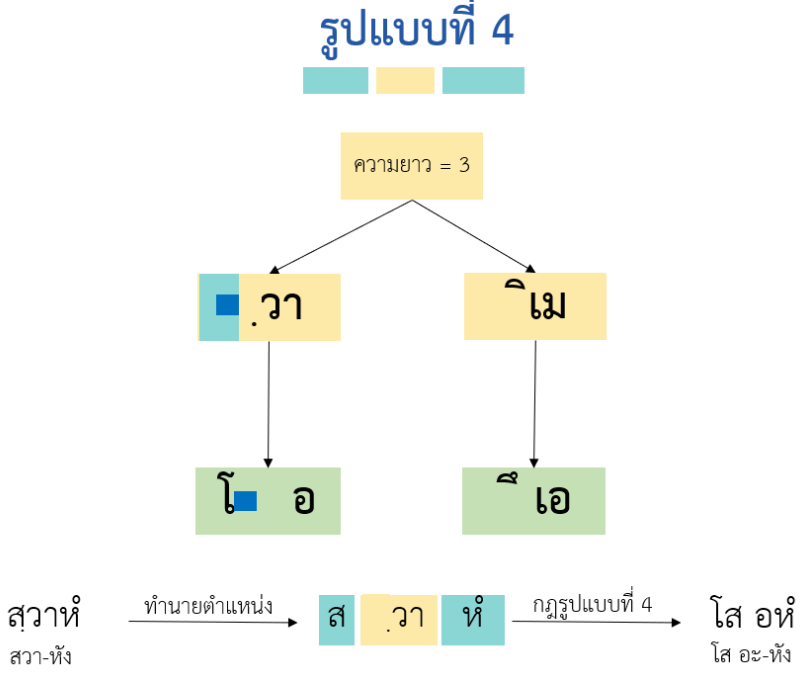


รูปที่ 31 กฎการตัดคำสนธิที่ใช้อักษรแรกเป็นสระหน้า (3)



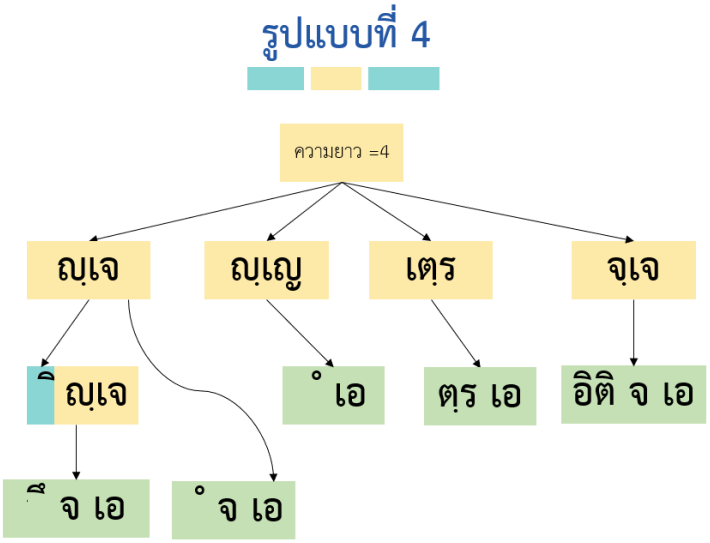
รูปที่ 32 กฎการตัดคำสนธิที่ใช้อักษรแรกเป็นสระหน้า (4)

รูปที่ 33 แสดงกฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่มีความยาวเท่ากับ 3 เมื่อเครื่องหมาย**พินทุ** (จุดใต้อักษร) นำหน้า “**วา**” จะเปลี่ยนอักษรสามตัวนี้เป็น “**โ อ**” โดยนำสระโอไปวางหน้าพยัญชนะตัวสุดท้ายของกลุ่มอักษรส่วนแรก และเมื่ออักษรส่วนที่ 2 เป็น “ิ เม” จะเปลี่ยนอักษรสามตัวนี้เป็น “ึ เอ”

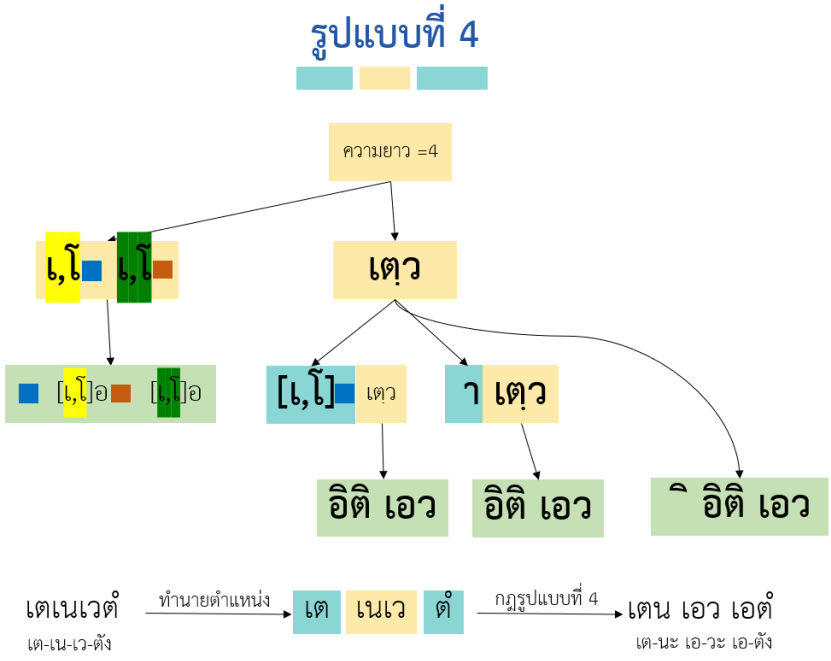


รูปที่ 33 กฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่ 3

กฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่มีความยาวส่วนที่ต้องแก้ไขเป็น 4 แสดงในรูปที่ 34 และ รูปที่ 35 พร้อมยกตัวอย่างการตัดคำสนธิ “**เตเนเวตำ**” ที่ได้ผลจากการทำนายตำแหน่งเป็น “**เต**” “**เนเว**” และ “**ตำ**” โดยแยกเสียงพยัญชนะซึ่งเป็นอักษรตัวที่ 2 และ 4 พร้อมทั้งใช้อักษร อ กำกับในพยางค์สระจะได้ “น เอว เอ” เมื่อนำไปรวมกับอักษรส่วนที่หนึ่งและสามจะได้ผลลัพธ์จากการตัดสนธิเป็น “**เตน เอว เอตำ**”

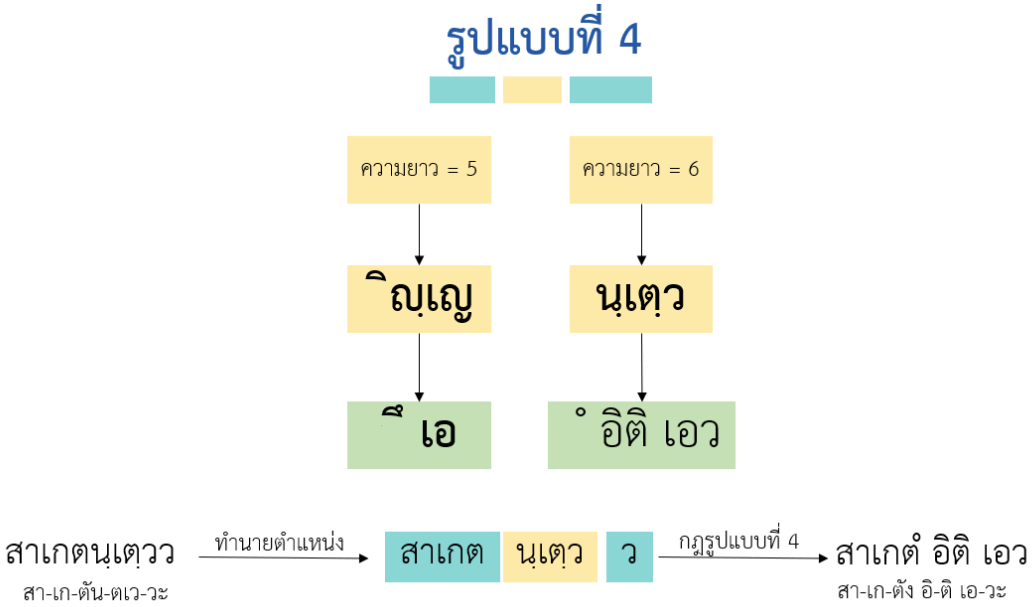


รูปที่ 34 กฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่ 4 (1)



รูปที่ 35 กฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่ 4 (2)

รูปที่ 36 แสดงกฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่มีความยาวเท่ากับ 5 และ 6 ซึ่งมีกฎอย่างละหนึ่งข้อได้แก่การเปลี่ยน “ิญเญ” เป็น “ึ เอ” และการเปลี่ยน “นเตฺว” เป็น “ ิติ เอว”



รูปที่ 36 กฎการตัดคำสนธิรูปแบบที่ 4 กลุ่มที่ 5 และกลุ่มที่ 6

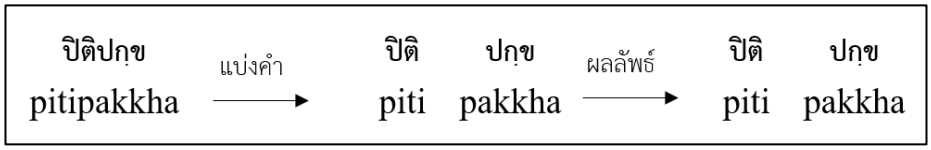
4.3 การตัดคำสมาสภาษาบาลีอักษรไทย

การวิจัยการตัดคำสมาสประกอบด้วย 1) วิเคราะห์รูปแบบการแยกคำสมาส 2) การทำนายตำแหน่งและรูปแบบตัดคำสมาส 3) โมเดลทำนายตำแหน่งตัดคำสมาส 4) การแยกกลุ่มอักษร 5) การแยกเสียงพยางค์ และ 6) การแปลงรูปคำกลับ

4.3.1 การวิเคราะห์รูปแบบการแยกคำสมาส

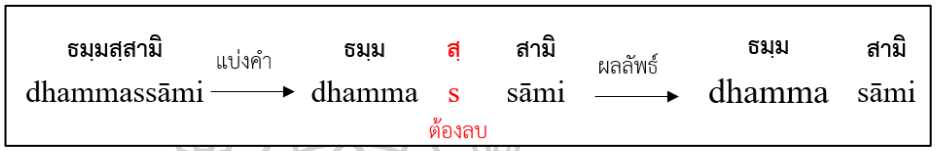
จากการวิเคราะห์รูปแบบการตัดคำสมาส จึงกำหนดรูปแบบการตัดคำสมาสเป็น 4 ประเภท ดังต่อไปนี้

1. คำสมาสมีตำแหน่งตัดคำที่สามารถแยกคำศัพท์ออกเป็นสองคำ ได้คำศัพท์ที่ถูกต้องและมีความหมายทั้งสองคำ (รูปที่ 37)



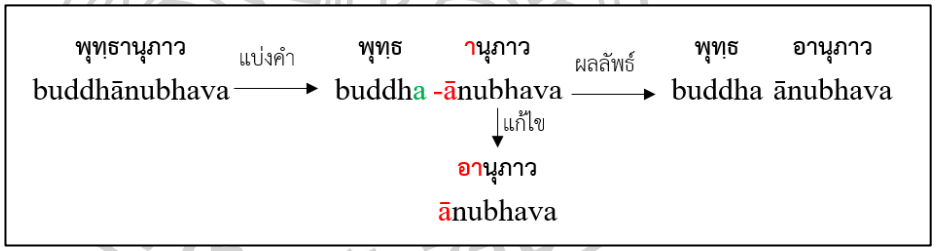
รูปที่ 37 การแยกคำสมาสรูปแบบที่ 1

2. คำสมาสมีตำแหน่งตัดคำที่สามารถแยกคำศัพท์ออกเป็นสองคำ ได้คำศัพท์ที่ถูกต้องและมีความหมายทั้งสองคำ แต่มีกลุ่มอักษรต้องลบออก เพราะกลุ่มอักษรเหล่านี้เกิดจากหลักการซ้อนพยัญชนะในขั้นตอนการสร้างคำสมาส (รูปที่ 38)



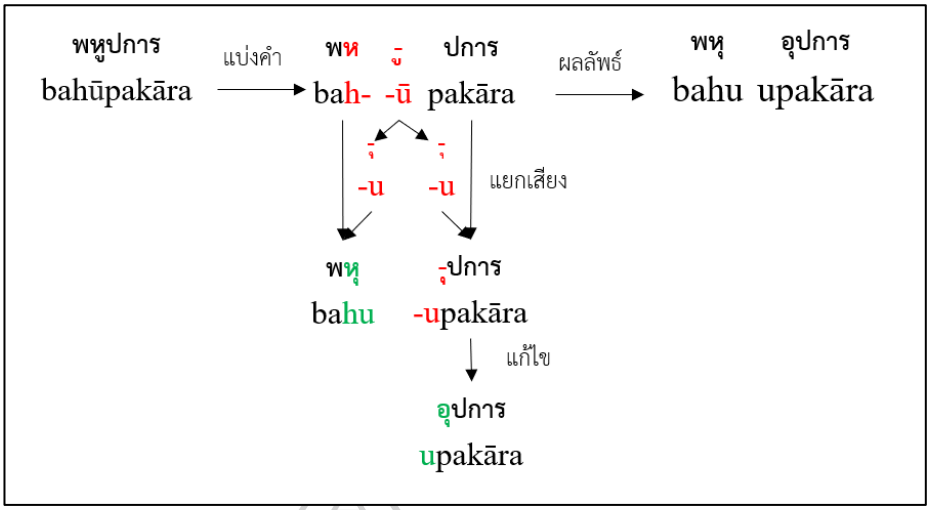
รูปที่ 38 การแยกคำสมาสรูปแบบที่ 2

3. คำสมาสมีตำแหน่งตัดคำที่สามารถแยกคำศัพท์ออกเป็นสองคำ พบว่ามีรูปพยัญชนะต้นกับรูปสระถูกแยกออกจากกัน แล้วคำหลังขึ้นต้นด้วยรูปสระจะไม่สามารถอ่านเป็นคำได้ เพราะไม่ถูกต้องตามหลักการเขียน (รูปที่ 39) ดังนั้นจึงต้องใช้กฎเพื่อแก้ไขรูปคำให้ถูกต้อง



รูปที่ 39 การแยกคำสมาสรูปแบบที่ 3

4. คำสมาสมีตำแหน่งตัดคำที่สามารถแยกคำศัพท์ออกเป็นสองคำ พบคำศัพท์ที่ไม่มีความหมายสองคำ เพราะเกิดการเปลี่ยนรูปสระของคำศัพท์ทั้งสองเป็นรูปสระตัวอื่นตามหลักการเชื่อมเสียงของภาษาบาลี ดังนั้นจึงต้องหารูปสระที่จะต้องแปลงกลับ ซึ่งตรงกับรูปสระอู ในรูปที่ 40



รูปที่ 40 การแยกคำสมาสรูปแบบที่ 4

4.3.2 การทำนายตำแหน่งและรูปแบบตัดคำคำสมาส

หลังจากวิเคราะห์รูปแบบการแยกคำสมาส จึงได้เตรียมข้อมูลสำหรับทำนายตำแหน่งตัดคำและประเภทตัดคำสมาส โดยผลการทำนายตำแหน่งตัดคำสมาส s และมีผลเฉลย o ที่ $|s| = |o|$ และ $o_i \in [0,4]$ ซึ่งแสดงคำอธิบายตารางที่ 27

ตารางที่ 27 ประเภทตำแหน่งตัดคำสมาส

คลาส	คำอธิบาย
0	ไม่ใช่ตำแหน่งตัดคำ
1	ตำแหน่งของอักษรท้ายคำ
2	ตำแหน่งของอักษรที่ต้องแปลงรูปคืน
3	ตัวอักษรที่ต้องลบออก
4	ส่วนขยายความยาว (Post Padding)

รูปที่ 41 แสดงตัวอย่างการเตรียมข้อมูลทำนายตำแหน่งตัดคำสมาสของคำว่า ธมฺมสฺสามิ (ทำ-มัด-สา-มิ) ซึ่งมีกลุ่มอักษรที่ต้องลบคือ สฺ เพราะถูกเพิ่มเข้ามาตามหลักการซ้อนพยัญชนะ

คำสนธิ	ธ	ม	.	ม	ส	.	ส	า	ม	ิ
ตำแหน่ง	0	1	2	3	4	5	6	7	8	9
อักษรท้ายคำ				✓						✓
อักษรที่ต้องแยกเสียง										
อักษรที่ต้องลบ					✓	✓				
คลาส	0	0	0	1	3	3	0	0	0	1

รูปที่ 41 คำสมาสที่มีอักษรที่ต้องลบ

รูปที่ 42 แสดงตัวอย่างการเตรียมข้อมูลทำนายตำแหน่งตัดคำสมาสของคำว่า ธมฺมกถาทิ (ทำ-มะ-กะ-ถา-ทิ) ซึ่งมีกลุ่มอักษรแยกเสียงพยางค์คือรูปสระ า โดยจะใช้กฎมาช่วยเพื่อแปลงรูปสระ า กลับเป็น อ อา หรือ อ อ

คำสนธิ	ธ	ม	.	ม	ก	ถ	า	ท	ิ
ตำแหน่ง	0	1	2	3	4	5	6	7	8
อักษรท้ายคำ				✓					✓
อักษรที่ต้องแยกเสียง							✓		
อักษรที่ต้องลบ									
คลาส	0	0	0	1	0	0	2	0	1

รูปที่ 42 คำสมาสที่มีอักษรที่ต้องแยกเสียง

4.3.3 โมเดลทำนายตำแหน่งตัดคำสมาส

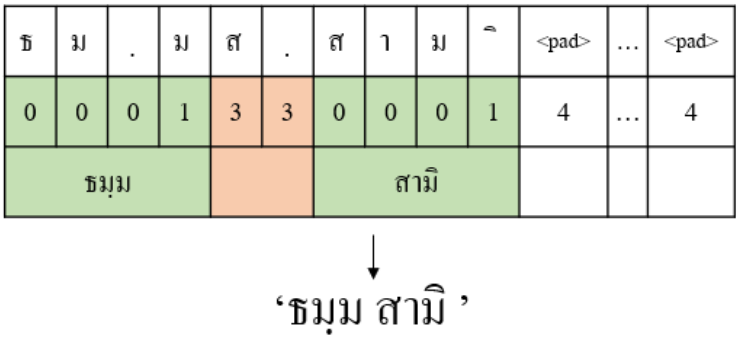
จากชุดข้อมูลที่เตรียมไว้สำหรับทำนายประเภทของตำแหน่งตัดคำ เริ่มจากปรับคำสมาสให้มีความยาวเท่ากับคำสมาสที่ยาวที่สุด (Post Padding) จากนั้นเข้ารหัสตัวอักษรในคำสมาสที่อยู่ในรูปแบบของจำนวนเต็ม และนำผลเฉลยไปขยายด้วยเลขศูนย์ (Zero Padding) ก่อนแปลงเป็น one-hot vector รวมถึงภาพโครงสร้างโมเดลการทำนายตำแหน่งตัดคำสมาส เป็นเช่นเดียวกันกับโมเดลทำนายตำแหน่งตัดคำสนธิที่ได้กล่าวไว้แล้วในหัวข้อ 4.2.3 โมเดลทำนายตำแหน่งและรูปแบบตัดคำสนธิ

4.3.4 การแยกกลุ่มอักษร

หลังจากป้อนคำสมาสและทำนายตัวอักษรในคำสมาสเป็นประเภทตำแหน่งตัดคำจากโมเดลแล้ว จากนั้นจะแยกกลุ่มอักษร เพื่อพิจารณาการเว้นวรรค การไม่แสดงผล และใช้กฎเพื่อช่วยแยกเสียงพยางค์ โดยในหัวข้อนี้กล่าวถึงกลุ่มอักษร 4 ประเภทดังนี้

1. กลุ่มอักษรที่มีอักษรสิ้นสุดคำ สามารถเขียนแทนด้วยนิพจน์ปกติ (Regular Expression) คือ $0^{*}1$ หมายถึง มีอักษรที่ถูกทำนายเป็นคลาส 0 จำนวน 0 ตัวหรือมากกว่า และต้องมีอักษรที่ถูกทำนายเป็นคลาส 1 ปิดท้าย กลุ่มอักษรประเภทนี้เมื่อนำไปแสดงผลจะมีเครื่องหมายเว้นวรรคตามหลัง
2. กลุ่มอักษรไม่มีอักษรสิ้นสุดคำ สามารถเขียนแทนด้วยนิพจน์ปกติ (Regular Expression) คือ 0^{+}[^23] หมายถึง มีอักษรที่ถูกทำนายเป็นคลาส 0 อย่างน้อย 1 ตัว และตัวถัดไปต้องไม่มีอักษรที่ถูกทำนายเป็นคลาส 2 หรือ 3 กล่าวอีกนัยหนึ่งคือ กลุ่มอักษรนี้สนใจเฉพาะอักษรที่ถูกทำนายเป็นคลาส 0 ล้วน และไม่รวมอักษรที่ถูกทำนายเป็นคลาส 2 หรือ 3 เมื่อนำไปแสดงผลจะไม่เพิ่มเครื่องหมายเว้นวรรค
3. กลุ่มอักษรที่ต้องแยกเสียง สามารถเขียนแทนด้วยนิพจน์ปกติ (Regular Expression) คือ 2^{+} หมายถึง มีอักษรที่ถูกทำนายเป็นคลาส 2 ติดกันอย่างน้อย 1 ตัว กรณีที่ทำนายตำแหน่งตัดคำสมาสและพบกลุ่มอักษรนี้ จะใช้กฎเพื่อแยกเสียงด้วย
4. กลุ่มอักษรที่ต้องลบ สามารถเขียนแทนด้วยนิพจน์ปกติ (Regular Expression) คือ 3^{+} หมายถึง มีอักษรที่ถูกทำนายเป็นคลาส 3 ติดกันอย่างน้อย 1 ตัว และอักษรกลุ่มนี้จะไม่ถูกนำไปแสดงผล

รูปที่ 43 แสดงตัวอย่างผลลัพธ์การแยกกลุ่มอักษรของคำสมาส ธมฺมสฺสามิ ที่ผ่านการทำนายตำแหน่งตัดคำแล้วจะได้กลุ่มอักษรที่มีอักษรสิ้นสุดคำ คือ ธมฺม ตามด้วยกลุ่มอักษรต้องลบ คือ สฺ และกลุ่มอักษรที่มีอักษรสิ้นสุดคำคือ สามิ



รูปที่ 43 การแยกกลุ่มอักษร

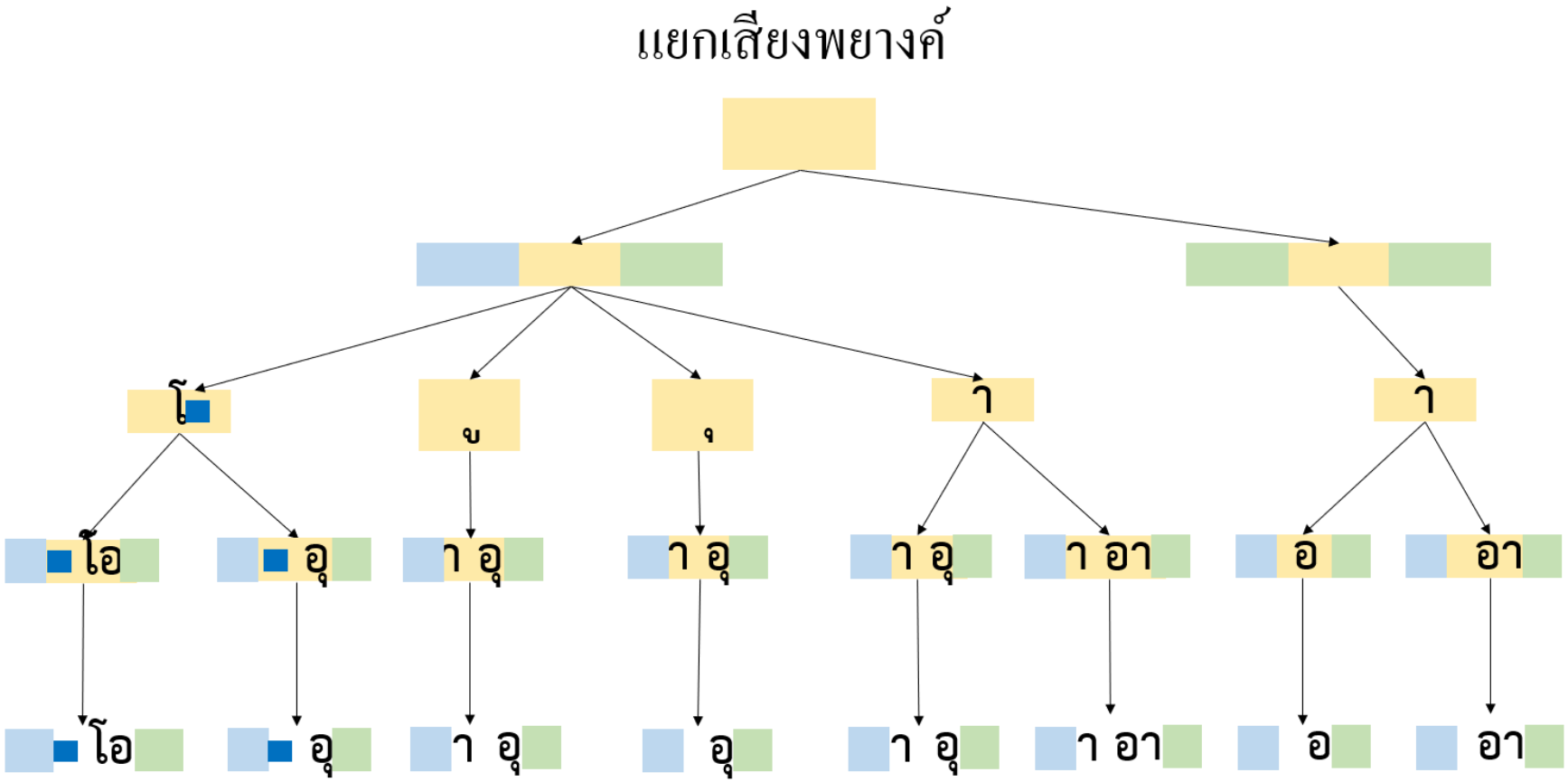
4.3.5 การแยกเสียงพยางค์

กรณีที่แยกกลุ่มอักษรแล้วพบกลุ่มอักษรที่ต้องแยกเสียงเช่นรูปสระ า ในรูปที่ 44 จะต้องใช้กฎเพื่อแยกเสียง โดยใช้ข้อมูลจากกลุ่มอักษรหน้าและหลังของกลุ่มอักษรที่ต้องเปลี่ยนรูปพิจารณาการแยกเสียงพยางค์ด้วย

รูปที่ 44 กลุ่มอักษรที่ต้องนำไปแยกเสียงพยางค์

รูปที่ 45 แสดงกฎการแยกเสียงพยางค์ กรณีที่มีกลุ่มอักษรไม่มีอักษรสิ้นสุดคำอยู่หน้า (แทนด้วยสีฟ้า) จะใช้เงื่อนไขจากกลุ่มอักษรที่ต้องแยกเสียง (แทนด้วยสีเหลือง) ได้แก่

1. ถ้ากลุ่มอักษรที่ต้องแยกเสียงเป็น สระโอตามด้วยพยัญชนะ จะนำพยัญชนะตัวนี้เป็นอักษรสิ้นสุดคำของกลุ่มอักษรด้านหน้า และเปลี่ยนรูปสระ โ เป็น **อุ** หรือ **โอ** ไว้นำหน้ากลุ่มอักษรด้านหลัง โดยใช้การตรวจสอบจากกลุ่มคำไม่เปลี่ยนรูปและกลุ่มคำเปลี่ยนรูปที่พบในพจนานุกรม
2. ถ้ากลุ่มอักษรที่ต้องแยกเสียงเป็น รูปสระ อู จะเปลี่ยนเป็น า อุ โดยนำ า ไปเป็นอักษรสิ้นสุดคำของกลุ่มอักษรด้านหน้า และนำอุไปนำหน้ากลุ่มอักษรด้านหลัง
3. ถ้ากลุ่มอักษรที่ต้องแยกเสียงเป็น รูปสระ อุ จะเปลี่ยนเป็น า อุ โดยนำ า ไปเป็นอักษรสิ้นสุดคำของกลุ่มอักษรด้านหน้า และนำอุไปนำหน้ากลุ่มอักษรด้านหลัง
4. ถ้ากลุ่มอักษรที่ต้องแยกเสียงเป็น รูปสระ า จะเปลี่ยนเป็น า อา หรือ า อ โดยนำ า ไปเป็นอักษรสิ้นสุดคำของกลุ่มอักษรด้านหน้า และเพิ่ม อ หรือ อา ให้คำหลังพร้อมทั้งตรวจสอบจากกลุ่มคำไม่เปลี่ยนรูปและกลุ่มคำเปลี่ยนรูปที่พบในพจนานุกรม



รูปที่ 45 กฎการแยกเสียงพยางค์

จากรูปที่ 49 แสดงกฎการแยกเสียงพยางค์ กรณีที่มีกลุ่มอักษรมีอักษรสิ้นสุดคำอยู่หน้า (แทนด้วยสีเขียว) โดยมีเงื่อนไขถ้ากลุ่มอักษรที่ต้องแยกเสียงเป็นรูปสระ า จะตรวจสอบจากกลุ่มคำไม่เปลี่ยนรูปและกลุ่มคำเปลี่ยนรูปที่พบในพจนานุกรมว่าควรแก้ไขโดยแปลง รูปสระ า นี้เป็น อ หรือ เป็น อา

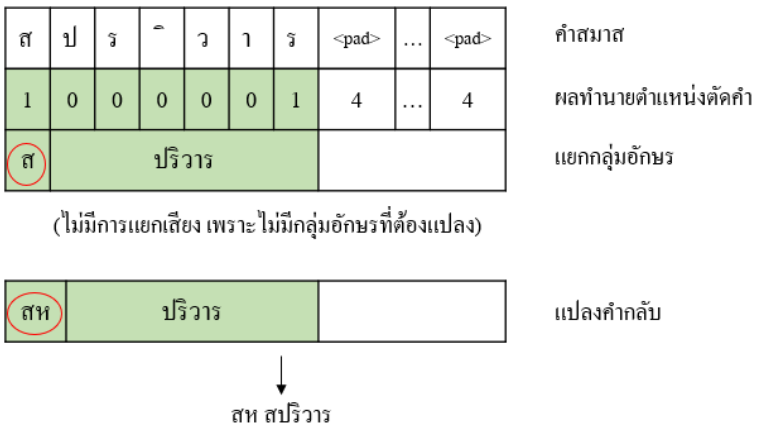
4.3.6 การแปลงรูปคำกลับ

เมื่อผ่านขั้นตอนการแยกกลุ่มอักษรและไม่พบกลุ่มอักษรต้องแยกเสียง หรือผ่านขั้นตอนการแยกเสียงพยางค์กรณีที่พบกลุ่มอักษรที่ต้องแยกเสียง จากนั้นจะนำหน่วยคำมาตรวจสอบเพิ่มเติมว่าตรงเงื่อนไขตามตารางที่ 28 เพื่อแปลงรูปคำกลับให้เป็นคำเดิมก่อนที่ถูกเปลี่ยนรูปจากกระบวนการสมาส

ตารางที่ 28 กฎการแปลงรูปคำกลับ

เงื่อนไข	ผลลัพธ์
มห	มหนฺต
อน	น
อ	น
ส	สห
ขึ้นต้นด้วย นุ ภิ ติ ป หรือ น	เพิ่ม อ นำหน้า
ขึ้นต้นด้วยรูปสระ	เพิ่ม อ นำหน้า

รูปที่ 46 แสดงตัวอย่างการป้อนคำสมาส สปริวาร และแยกกลุ่มอักษรได้เป็น ส ปริวาร แต่ ส ตรงกับเงื่อนไขการแปลงรูปคำกลับจึงเปลี่ยนรูปเป็น สห ดังนั้นจะได้ผลลัพธ์จากการตัดคำสมาสเป็น สห ปริวาร



รูปที่ 46 การเปลี่ยนรูปคำกลับ

บทที่ 5

ผลการดำเนินงาน

5.1 ผลการวิจัยการตัดคำสนธิ

ผลการวิจัยในหัวข้อนี้ประกอบด้วย 1) ผลการทำนายตำแหน่งและรูปแบบตัดคำสนธิ 2) ประสิทธิภาพการตัดคำสนธิ 3) เปรียบเทียบการตัดคำสนธิกับโมเดลการตัดคำภาษาไทย 4) เปรียบเทียบกับงานวิจัยด้านการตัดคำในภาษาสันสกฤต

5.1.1 การทำนายตำแหน่งและรูปแบบตัดคำสนธิ

การทำนายตำแหน่งและรูปแบบตัดคำสนธิด้วยชุดข้อมูลคำสนธิที่จัดเตรียมโดยผู้เชี่ยวชาญ พบคำสนธิที่ยาวที่สุดจำนวน 57 ตัวอักษร ดังนั้นจึงแสดงจำนวนข้อมูลที่นำมาใช้กับการทำนายตำแหน่งและรูปแบบตัดคำสนธิไว้ในในตารางที่ 29 และกำหนดพารามิเตอร์ของโมเดลทำนายประเภทตำแหน่งตัดคำดังแสดงตารางที่ 30

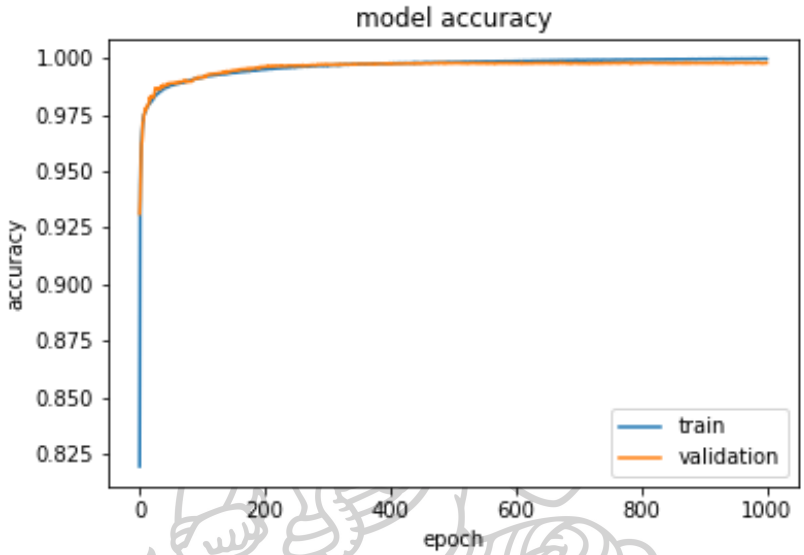
ตารางที่ 29 จำนวนข้อมูลฝึกสอน ข้อมูลตรวจสอบ และข้อมูลทดสอบ

	จำนวนคำสนธิ	จำนวนตัวอักษร
ข้อมูลฝึกสอน	4,597	4,597 x 57 = 262,029
ข้อมูลตรวจสอบ	511	511 x 57 = 29,127
ข้อมูลทดสอบ (20%)	1,277	1,277 x 57 = 72,789
รวม	6,385	363,945

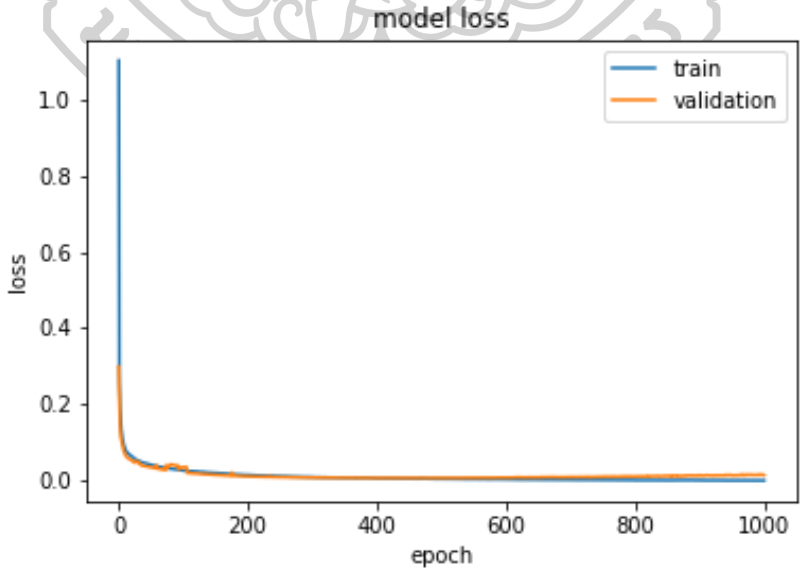
ตารางที่ 30 พารามิเตอร์สำหรับโมเดลการทำนายประเภทตำแหน่งตัดคำ

Embedding size	100
LSTM unit	100+100
Dropout	0.5
Hidden layer node	50
Hidden activation function	ReLU
Activation function (Output layer)	Softmax
Learning rate	0.0001
Epoch	1,000
Optimizer	Adam

จากการฝึกสอนโมเดลทำนายตำแหน่งและรูปแบบตัดคำ และวัดประสิทธิภาพด้วยชุดการฝึกสอนด้วยข้อมูลตรวจสอบ ได้ค่าความแม่นยำ 0.9999 และ 0.9978 ตามลำดับ ดังแสดงในรูปที่ 47 และมีค่าคลาดเคลื่อน 0.0001 และ 0.0022 ตามลำดับ ดังแสดงในรูปที่ 48 ทำให้เห็นได้ว่าโมเดลสามารถทำนายได้อย่างแม่นยำกับข้อมูลตรวจสอบเพราะระยะห่างระหว่างเส้นค่อนข้างน้อย



รูปที่ 47 ค่าความแม่นยำของข้อมูลตรวจสอบ



รูปที่ 48 ค่าความคลาดเคลื่อนของข้อมูลตรวจสอบ

5.1.2 การวัดประสิทธิภาพการตัดคำสนธิ

การวัดประสิทธิภาพการตัดคำสนธิ แบ่งออกเป็น 2 ส่วนตามขั้นตอนการทำงาน โดยนำชุดข้อมูลทดสอบมีจำนวน 1,277 คำ มาวัดประสิทธิภาพ

1. วัดประสิทธิภาพการทำนายตำแหน่งและรูปแบบตัดคำสนธิด้วย Confusion matrix, Precision, Recall และ F1-score โดยใช้ตำแหน่งตัดคำจำนวน 72,789 ตำแหน่งจากคำสนธิ 1,277 คำ ซึ่งแต่ละคำขยายความความเป็น 57 (1,277 x 57 = 72,789)
2. วัดประสิทธิภาพด้วยเปรียบเทียบการตัดคำสนธิ ซึ่งใช้ข้อมูลตำแหน่งและรูปแบบที่ได้จากโมเดลร่วมกับกฎที่เตรียมไว้ และนำไปเปรียบเทียบกับชุดข้อมูลคำสนธิที่ผู้เชี่ยวชาญจัดเตรียมให้

ตารางที่ 31 แสดง Confusion matrix ของโมเดลทำนายตำแหน่งและรูปแบบตัดคำทั้ง 5 คลาส โดยแนวหลักแสดงคำตอบจริง ส่วนแนวแถวคำตอบที่โมเดลทำนาย

ตารางที่ 31 Confusion Matrix

		ทำนาย					
		0	**1**	**2**	**3**	**4**	**5**
ผลเฉลย	**0**	11,406	6	41	2	31	0
	1	3	261	0	0	1	0
	2	43	0	435	0	4	0
	3	0	0	0	103	0	0
	4	22	0	6	0	804	0
	5	0	0	0	0	0	59,621

ตารางที่ 32 แสดงประสิทธิภาพของโมเดลด้วย Precision Recall และ F1-score ของข้อมูลทั้ง 6 คลาส จะเห็นว่าจำนวนของข้อมูลแต่ละคลาสมีจำนวนไม่เท่ากัน ดังนั้นจะใช้ค่า Weighted – average เป็นค่าความแม่นยำ ซึ่งมีค่า 0.9978

ตารางที่ 32 Precision Recall และ F1-score

คลาส	Precision	Recall	F1-Score	จำนวนข้อมูล
0	0.9941	0.9930	0.9936	11,486
1	0.9775	0.9849	0.9812	268
2	0.9025	0.9025	0.9025	482
3	0.9810	1.0000	0.9904	103
4	0.9571	0.9663	0.9617	832
5	1.0000	1.0000	1.0000	59,621
Macro - average	0.9687	0.9745	0.9716	72,789
Weighted - average	0.9978	0.9978	0.9978	72,789

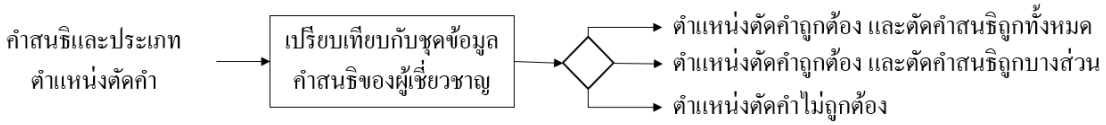
จากตารางที่ 32 โมเดลทำนายตำแหน่งและรูปแบบตัดคำของคลาส 2 มีค่าความแม่นยำน้อยที่สุด จากการตรวจสอบเพิ่มเติมพบว่าตัวอักษรที่ถูกระบุว่าเป็นตำแหน่งตัดคำของรูปแบบที่ 2 นั้นมักเป็นสระ (า, ิ, ี, ุ, ู, เ และ โ) ซึ่งมีจำนวนสระไม่เกิน 482 ตัวที่เป็นตำแหน่งตัดคำของคลาส 2 จากสระทั้งหมด และจากข้อมูลทดสอบจำนวน 72,789 ตำแหน่ง มีคลาส 2 เพียง 482 ตำแหน่ง ดังนั้นโอกาสที่สระจะไม่ใช่ตำแหน่งตัดคำนั้นมีสูง จึงทำให้คลาส 2 ทำนายคลาดเคลื่อนสูงกว่าคลาสอื่น

ตารางที่ 33 แสดงตัวอย่างการทำนายประเภทตำแหน่งตัดคำที่ไม่ถูกต้อง พร้อมทั้งแสดงผลลัพธ์การใช้ตำแหน่งตัดคำที่ไม่ถูกต้อง พร้อมทั้งแสดงคำอธิบาย

ตารางที่ 33 ตัวอย่างการทำนายตำแหน่งและรูปแบบตัดคำที่ไม่ถูกต้อง

คำสนธิ	คำตอบจริง		คำตอบจากที่ทำนาย		คำอธิบาย
	ตำแหน่งตัดคำและรูปแบบ	แยกคำ	ตำแหน่งตัดคำและรูปแบบ	แยกคำ	
เทมาติ (เท-มา-ติ)	0 0 0 **2** 0 0 5 5 5 5 5 5 …	เทม อิติ	0 0 0 **0** 0 0 5 5 5 5 5 5 …	เทมาติ	ไม่มีตำแหน่งตัดคำ
วตฺวาติ (วัด-ตฺวา-ติ)	0 0 0 0 **0 2** 0 5 5 5 5 5 …	วตฺวา อิติ	0 0 0 0 **2 2** 0 5 5 5 5 5 …		ตำแหน่งตัดคำเกิน
อิทญฺจิทญฺจ (อิ-ทัน-จิ-ทัน-จะ)	0 0 0 0 0 3 **2** 0 0 0 3 5 …	อิทำ จ อิทำ จ	0 0 0 0 0 3 **0** 0 0 0 3 5 …	อิทำ จิทำ จ	ตำแหน่งตัดคำไม่ครบ
สยถาติ (สะ-ยะ-ถา-ติ)	0 0 0 **2 0** 0 5 5 5 5 5 5 …	สยถ อิติ	0 0 0 **0 2** 0 5 5 5 5 5 5 …	สยถา อิติ	ตำแหน่งตัดคำผิด

การวัดประสิทธิภาพโดยเปรียบเทียบระหว่างผลลัพธ์การตัดคำสนธิและชุดข้อมูลคำสนธิที่เตรียมไว้โดยผู้เชี่ยวชาญ สามารถจัดแบ่งออกเป็น 3 กลุ่ม ดังรูปที่ 49 ได้แก่ 1) ทำนายประเภทตำแหน่งตัดคำถูกต้องและตัดคำสนธิถูกต้องทุกตัวอักษร 2) ทำนายประเภทตำแหน่งตัดคำถูกต้องและตัดคำสนธิถูกบางส่วน 3) ทำนายประเภทตำแหน่งตัดคำไม่ถูกต้อง และแสดงจำนวนของแต่ละกลุ่มไว้ในตารางที่ 34



รูปที่ 49 การวัดประสิทธิภาพโดยเปรียบเทียบผลลัพธ์การตัดคำสนธิ

ตารางที่ 34 การนับจำนวนการตัดคำสนธิเทียบกับชุดข้อมูลของผู้เชี่ยวชาญ

ตำแหน่งและรูปแบบตัดคำสนธิถูก		**ตำแหน่งและรูปแบบตัดคำสนธิไม่ถูก**
ตัดคำสนธิถูกทั้งหมด	**ตัดคำสนธิถูกต้องบางส่วน**	**ตำแหน่งตัดคำไม่ถูกต้อง**
1,144 คำ (89.58%)	31 คำ (2.43%)	102 คำ (7.99%)

จากคำสนธิที่ตัดคำได้ถูกต้องมี 1,144 คำ (89.58%) และคำสนธิที่ตัดคำและปรับปรุงคำได้ถูกต้องบางส่วนมี 31 คำ (2.43%) คำสนธิที่ทำนายประเภทตำแหน่งตัดคำไม่ถูกต้องมี 102 คำ (7.99%)

5.1.3 เปรียบเทียบการตัดคำสนธิด้วยโมเดลการตัดคำภาษาไทย

จากรูปแบบการตัดคำคำสนธิประเภทที่ 1 สามารถแยกคำออกจากกันโดยที่ไม่มีการเปลี่ยนแปลงตัวอักษร ดังนั้นจึงทำได้นำชุดข้อมูลสนไปทดลองตัดคำกับโมเดลตัดคำภาษาไทยที่มีความแม่นยำและนิยมใช้กับการตัดภาษาไทยได้แก่ DeepCut และ AttaCut พบว่าตัวตัดคำภาษาไทยยังไม่เหมาะสมกับการนำมาใช้ตัดคำสนธิภาษาบาลีอักษรไทย เนื่องจากสามารถตัดคำได้ถูกต้องน้อย ดังแสดงในตารางที่ 35 และตารางที่ 36

ตารางที่ 35 เปรียบเทียบการตัดคำสนธิกับตัวตัดคำภาษาไทย

วิธีที่นำเสนอ	**โมเดลตัดคำภาษาไทย**	
	DeepCut	**AttaCut**
1144 คำ (89. 58%)	25 คำ (2.04%)	10 คำ (0.97%)

ตารางที่ 36 แสดงผลลัพธ์เปรียบเทียบระหว่างวิธีที่นำเสนอกับตัวตัดคำภาษาไทย

คำสนธิ	**ผลเฉลย**	**ผลลัพธ์การตัดคำ**		
		วิธีที่นำเสนอ	**DeepCut**	**AttaCut**
มาตาปีตูน์ปี (มา-ตา-ปี-ตู-นัง-ปี)	มาตาปีตูน์ ปี	มาตาปีตูน์ ปี ✓	มาตาปีตูน ◌ ปี ×	มาตาปีตูน์ปี ×
ปูชีปี (ปู-ชัง-ปี)	ปูชี์ ปี	ปูชี์ ปี ✓	ปูชีปี ×	ปู ชีปี ×
กาญจนรูปกโตปี (กาน-จะ-นะ-รู-ปะ-กะ-โต-ปี)	กาญจนรูปกโต ปี	กาญจนรูปกโต ปี ✓	กาญจนรูปกโตปี ×	กาญจน รูป กโตปี ×
วีสติปี (วิ-สะ-ติ-ปี)	วีสติ ปี	วีสติ ปี ✓	วีสติปี ×	วีสติ ปี ✓
วิชชมานาปี (วิด-ชะ-มา-นา-ปี)	วิชชมานา ปี	วิชชมานา ปี ✓	วิชชมานา ปี ✓	วิชชมา นาปี ×

5.1.4 เปรียบเทียบกับงานวิจัยด้านการตัดคำในภาษาสันสกฤต

เนื่องจากงานวิจัยที่เกี่ยวข้องกับตัดคำสนธิมีค่อนข้างน้อย และงานตัดคำสนธิในภาษาบาลีที่พบไม่ได้ทดสอบการตัดคำ จึงได้นำการตัดคำสนธิในภาษาบาลีที่นำเสนอไปเปรียบเทียบกับงานวิจัยการตัดคำสนธิในภาษาสันกฤตในตารางที่ 37 เพราะภาษาทั้งสองมีแหล่งกำเนิดภาษาและโครงสร้างใกล้เคียงกัน

ตารางที่ 37 เทียบเทียบการตัดคำสนธิระหว่างภาษาบาลีและภาษาสันสกฤต

ชื่อบทความ	ขอบเขต	ภาษา	ชุดข้อมูล	จำนวนคำ	วิธีการ	ความแม่นยำ
SandhiKosh: A Benchmark Corpus for Evaluating Sanskrit Sandhi Tools [21]	สร้างชุดข้อมูล, สำรวจเครื่องมือและชุดข้อมูล	สันสกฤต (เทวนาครี)	UoH + SandhiKosh	13,648 (ชุดทดสอบ 100%)	JNU Tools, UoH Tools, INRIA Tools	JNU 7.64%, UoH 53.69%, INRIA 58.18%
Sanskrit Sandhi Splitting using seq2(seq)^2 [19]	ตัดคำสนธิด้วยโมเดล	สันสกฤต (เทวนาครี)	UoH + SandhiKosh	71,747 (ชุดฝึก 80%, ชุดทดสอบ 20%)	Double Decoder RNN	ทำนายตำแหน่ง 95.0% และตัดคำถูก 79.5%
Neural Compound-Word (Sandhi) Generation and Splitting in Sanskrit Language [20]	ตัดคำสนธิด้วยโมเดล	สันสกฤต (เทวนาครี)	UoH + SandhiKosh	77,842 (ชุดฝึก 80%,ชุดทดสอบ 20%)	ทำนายตำแหน่งด้วย RNN และตัดคำด้วย BiLSTM	ทำนายตำแหน่ง 92.3% และตัดคำถูก 86.8%
งานวิจัยที่นำเสนอ	ตัดคำสนธิด้วยโมเดล	บาลี (อักษรไทย)	ธมฺมปทฏฺฐกถา 8 เล่ม	6,385 (ชุดฝึก 80%,ชุดทดสอบ 20%)	ทำนายตำแหน่งด้วย BiLSTM และตัดคำด้วยกฎ	ทำนายตำแหน่ง 92.01% และตัดคำถูก 89.58%

5.2 ผลการวิจัยการตัดคำสมาส

ผลการวิจัยในหัวข้อนี้ประกอบด้วย 1) การทำนายตำแหน่งและรูปแบบตัดคำสมาส 2) การวัดประสิทธิภาพการตัดคำสมาส 3) เปรียบเทียบการตัดคำสมาสด้วยโมเดลการตัดคำภาษาไทย ดังต่อไปนี้

5.2.1 การทำนายตำแหน่งและรูปแบบตัดคำสมาส

การทำนายตำแหน่งและรูปแบบตัดคำสนธิด้วยชุดข้อมูลคำสนธิที่จัดเตรียมโดยผู้เชี่ยวชาญ พบคำสนธิที่ยาวที่สุดจำนวน 34 ตัวอักษร ดังนั้นจึงแสดงจำนวนข้อมูลที่นำมาใช้กับการทำนายตำแหน่งและรูปแบบตัดสมาสไว้ในในตารางที่ 38 และกำหนดพารามิเตอร์ของโมเดลทำนายประเภทตำแหน่งตัดคำดังแสดงตารางที่ 39

ตารางที่ 38 จำนวนข้อมูลฝึกสอน ข้อมูลตรวจสอบ และข้อมูลทดสอบ

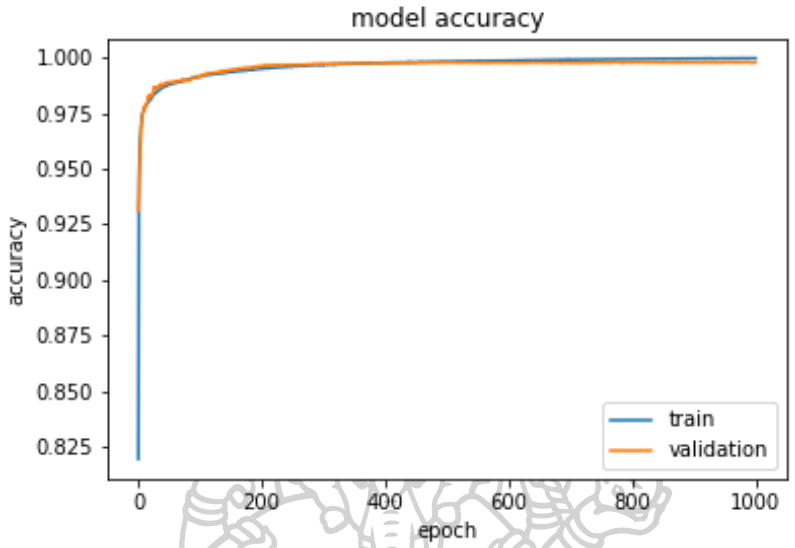
	จำนวนคำสมาส	จำนวนตัวอักษร
ข้อมูลฝึกสอน	3,223	3,223 x 34 = 109,582
ข้อมูลตรวจสอบ	359	359 x 34 = 12,206
ข้อมูลทดสอบ (20%)	896	896 x 34 = 30,464
รวม	4,478	152,252

ตารางที่ 39 พารามิเตอร์สำหรับโมเดลการทำนายประเภทตำแหน่งตัดคำ

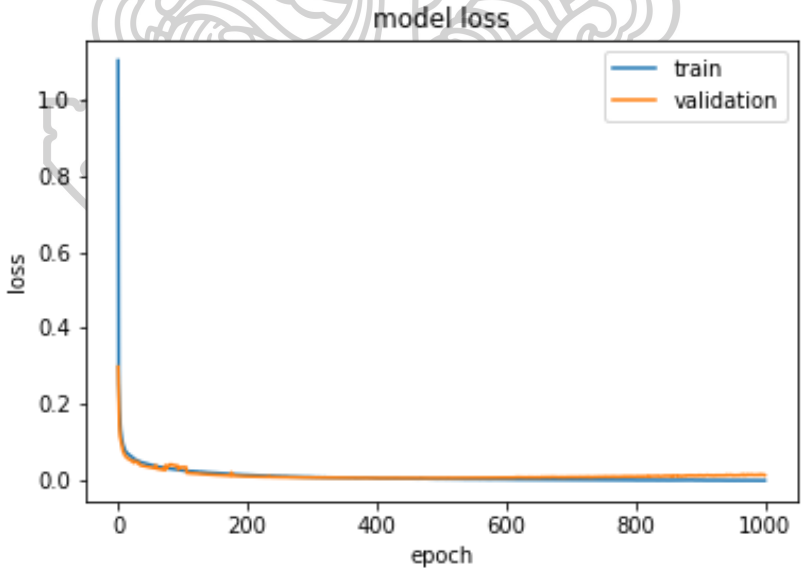
Embedding size	100
LSTM unit	100+100
Dropout	0.5
Hidden layer node	50
Hidden activation function	ReLU
Activation function (Output layer)	Softmax
Learning rate	0.0001
Epoch	1,000
Optimizer	Adam

จากการฝึกสอนโมเดลทำนายตำแหน่งและรูปแบบตัดคำและวัดประสิทธิภาพด้วยชุดการฝึกสอนด้วยข้อมูลตรวจสอบได้ค่าความแม่นยำ 0.9921 และ 0.9879 ตามลำดับ ดังแสดงในรูปที่ 47

และมีค่าคลาดเคลื่อน 0.0079 และ 0.0073 ตามลำดับ ดังแสดงในรูปที่ 48 ทำให้เห็นได้ว่าโมเดลสามารถทำนายได้อย่างแม่นยำกับข้อมูลตรวจสอบเพราะระยะห่างระหว่างเส้นค่อนข้างน้อย



รูปที่ 50 ค่าความแม่นยำของข้อมูลตรวจสอบ



รูปที่ 51 ค่าความคลาดเคลื่อนของข้อมูลตรวจสอบ

5.1.2 การวัดประสิทธิภาพการตัดคำสนธิ

การวัดประสิทธิภาพการตัดคำสนธิ แบ่งออกเป็น 2 ส่วนตามขั้นตอนการทำงาน โดยนำชุดข้อมูลทดสอบมีจำนวน 896 คำ มาวัดประสิทธิภาพ

1. วัดประสิทธิภาพการทำนายตำแหน่งและรูปแบบตัดคำสนธิด้วย Confusion matrix, Precision, Recall และ F1-score โดยใช้ตำแหน่งตัดคำจำนวน 30,464 ตำแหน่งจากคำสนธิ 896 คำ ซึ่งแต่ละคำขยายความความเป็น 34 (896 x 34 = 30,464)
2. วัดประสิทธิภาพด้วยเปรียบเทียบการตัดคำสนธิ ซึ่งใช้ข้อมูลตำแหน่งและรูปแบบที่ได้จากโมเดลร่วมกับกฎที่เตรียมไว้ และนำไปเปรียบเทียบกับชุดข้อมูลคำสนธิที่ผู้เชี่ยวชาญจัดเตรียมให้

ตารางที่ 40 แสดง Confusion matrix ของโมเดลทำนายตำแหน่งและรูปแบบตัดคำทั้ง 4 คลาส โดยแนวหลักแสดงคำตอบจริง ส่วนแนวแถวคำตอบที่โมเดลทำนาย

ตารางที่ 40 Confusion Matrix

		ทำนาย				
		0	**1**	**2**	**3**	**4**
ผลเฉลย	**0**	6,636	84	14	10	0
	1	93	1,571	8	2	0
	2	3	1	583	0	0
	3	23	1	0	71	0
	4	0	2	0	0	21,362

ตารางที่ 41 แสดงประสิทธิภาพของโมเดลด้วย Precision Recall และ F1-score ของข้อมูลทั้ง 5 คลาส จะเห็นว่าจำนวนของข้อมูลแต่ละคลาสมีจำนวนไม่เท่ากัน ดังนั้นจะใช้ค่า Weighted – average เป็นค่าความแม่นยำ ซึ่งมีค่า 0.9920

ตารางที่ 41 Precision Recall และ F1-score

คลาส	**Precision**	**Recall**	**F1-Score**	**จำนวนข้อมูล**
0	0.9824	0.9840	0.9832	6,744
1	0.9470	0.9385	0.9427	1,674

คลาส	Precision	Recall	F1-Score	จำนวนข้อมูล
2	0.9636	0.9932	0.9782	587
3	0.8554	0.7474	0.7978	95
4	1.000	0.9999	1.0000	21,364
Micro - average	0.9497	0.9326	0.9404	30,464
Weighted - average	0.9920	0.9921	0.9920	30,464

5.2.3 เปรียบเทียบการตัดคำสนธิด้วยโมเดลการตัดคำภาษาไทย

จากรูปแบบการตัดคำคำสมาสประเภทที่ 1 สามารถแยกคำออกจากกันโดยที่ไม่มีการเปลี่ยนแปลงตัวอักษร ดังนั้นจึงทำได้นำชุดข้อมูลสนไปทดลองตัดคำกับโมเดลตัดคำภาษาไทยที่มีความแม่นยำและนิยมใช้กับการตัดภาษาไทยได้แก่ DeepCut และ AttaCut ดังแสดงใตารางที่ 42

ตารางที่ 42 เปรียบเทียบการตัดคำสมาสกับโมเดลตัดคำภาษาไทย

วิธีที่นำเสนอ	โมเดลตัดคำภาษาไทย	
	DeepCut	AttaCut
896 คำ (81.91%)	22 คำ (0.0421%)	22 คำ (0.0421%)

บทที่ 6

สรุปผลการดำเนินงาน และข้อเสนอแนะ

การสร้างตัวแบบการตัดคำภาษาบาลีอักษรไทยให้สามารถแยกคำสนธิและแบ่งคำสมาสได้ โดยประยุกต์ใช้โครงข่ายประสาทเทียมแบบแอลเอสทีเอ็มแบบสองทิศทางมาใช้ เพื่อทำนายตำแหน่งตัดคำและรูปแบบ จากนั้นนำคำสนธิหรือคำสมาสที่ทราบตำแหน่งและรูปแบบมาแก้ไขคำด้วยกฎที่ผู้วิจัยได้วิเคราะห์และถอดรูปแบบมาจากไวยากรณ์ภาษาบาลีโดยมีวัตถุประสงค์หลักเพื่อศึกษาวิธีการตัดคำสนธิและคำสมาสภาษาบาลีอักษรไทย และสามารถนำไปใช้ได้อย่างถูกต้อง

ชุดข้อมูลคำสนธิจำนวน 6,854 และ ข้อมูลคำสมาสจำนวน 4,478 ได้รับความอนุเคราะห์จากพระมหาฉัตรชัย ถาวโร (ฉัตรชัย อภิชล) เปรียญธรรม ๘ ประโยค วัดประดู่ จังหวัดสมุทรสงคราม สละเวลามาจัดเตรียมคำสนธิและคำสมาส พร้อมทั้งผลเฉลยการตัดคำ จากนั้นผู้วิจัยได้วิเคราะห์และถอดรูปแบบมาจากไวยากรณ์ภาษาบาลีและสร้างเป็นกฎเพื่อให้สามารถแก้ไขคำให้ถูกต้องและมีความหมาย

จากการค้นคว้าเพื่อศึกษาการประมวลผลภาษาธรรมชาติที่เกี่ยวข้องกับภาษาบาลีและสันสกฤต พบงานวิจัยที่ใช้อักษรในภาษาอื่น ๆ ได้แก่ อักษรไทย อักษรพม่า อักษรโรมัน และอักษรเทวนาครี โดยพบงานวิจัยการตัดคำสนธิภาษาสันสกฤตใช้โครงข่ายประสาทเทียมจำนวน 2 งาน ซึ่งประกอบด้วยกระบวนการทำนายตำแหน่งตัดคำและกระบวนการแก้ไขคำ และพบงานตัดคำสนธิภาษาบาลีอักษรโรมันจำนวน 1 งาน ซึ่งใช้กฎที่สร้างด้วยนิพจน์ปกติ

ส่วนผู้วิจัยได้นำเสนอวิธีการตัดคำสมาสและคำสนธิโดยใช้โครงข่ายประสาทเทียมร่วมกับกฎ เนื่องจากไม่พบงานวิจัยที่เกี่ยวข้องกับการตัดคำสนธิในภาษาบาลีอักษรไทย จึงได้ประเมินผลโดยพิจารณาความแม่นยำของการทำนายตำแหน่งตัดคำที่ได้จากโครงข่ายประสาทเทียมแบบแอลเอสทีเอ็มแบบสองทิศทาง และวัดประสิทธิภาพการตัดคำโดยนำผลลัพธ์ที่ผ่านการทำนายตำแหน่งและแก้ไขคำด้วยกฎไปเปรียบเทียบกับชุดข้อมูลที่ถูกจัดเตรียมโดยผู้เชี่ยวชาญ

จากขั้นตอนดังกล่าวข้างต้น ในบทนี้จึงได้แบ่งเป็น 3 ส่วนดังนี้

1. การสรุปผลวิจัยการตัดคำสนธิ
2. การสรุปผลวิจัยการตัดคำสนธิ
3. ข้อเสนอแนะ

6.1 สรุปผลวิจัยการตัดคำสนธิ

จากบทที่ 4 วิธีดำเนินการวิจัย และ บทที่ 5 ผลการดำเนินงาน ประเมินผลโมเดลทำนายตำแหน่งตัดด้วย Precision Recall และ F1-score ได้ค่าความแม่นยำ 0.9978 ทำนายตำแหน่งแบบเทียบทั้งคำถูกต้อง 0.9201 และวัดประสิทธิภาพโดยเปรียบเทียบระหว่างผลลัพธ์การตัดคำสนธิและชุดข้อมูลคำสนธิที่เตรียมไว้โดยผู้เชี่ยวชาญได้ถูกต้อง 89.58%

6.1 สรุปผลวิจัยการตัดคำสมาส

จากบทที่ 4 วิธีดำเนินการวิจัย และ บทที่ 5 ผลการดำเนินงาน ประเมินผลโมเดลทำนายตำแหน่งตัดด้วย Precision Recall และ F1-score ได้ค่าความแม่นยำ 0.9921ทำนายตำแหน่งแบบเทียบทั้งคำถูกต้อง 0.83059 และวัดประสิทธิภาพโดยเปรียบเทียบระหว่างผลลัพธ์การตัดสมาสและชุดข้อมูลคำสมาสที่เตรียมไว้โดยผู้เชี่ยวชาญได้ถูกต้องทั้งคำ 81.91%

6.2 ข้อเสนอแนะ

1. เนื่องจากคำชุดข้อมูลคำสนและคำสมาสที่พบในหนังสือที่ใช้รวบรวมปรากฏพบ อีกทั้งคำสมาสและคำสนธิที่มักเกิดจากหลายศัพท์รวมกัน ดังนั้นสามารถเพิ่มชุดข้อมูลด้วยตัดคำสมาสหรือคำสนธิที่เกิดจากการรวมศัพท์มากกว่าสองศัพท์ขึ้นไปได้
2. ปัจจุบันทำกฎของรูปแบบตัดคำโดยใช้การสังเกต การตรวจสอบกฎและเปลี่ยนแปลงกฎเมื่อได้รับคำสนธิใหม่จะช่วยให้โมเดลสามารถเรียนรู้ได้เอง
3. การแบ่งคำสมาสเพื่อตัดคำ ควรแบ่งคำศัพท์ให้เป็นคำที่สามารถค้นหาความหมายได้ในพจนานุกรม

รายการอ้างอิง

[1] เดือน คำดี, "วิกฤตศาสนา : ปัญหาและทางออก," วารสารภาษาไทยและวัฒนธรรมไทย, 2551.

[2] พระมหาสมาน ชาตวิริโย, "การศึกษาคุณลักษณะที่พึงประสงค์ของพระธรรมทูตในการเผยแผ่พระพุทธศาสนาในประเทศสหรัฐอเมริกา," วารสารสังคมศาสตร์และมานุษยวิทยาเชิงพุทธ, 2563.

[3] คณาจารย์มหาวิทยาลัยมหาจุฬาลงกรราชวิทยาลัย, วรรณคดีภาษาบาลี. กรุงเทพมหานคร: มหาวิทยาลัยจุฬาลงกรณราชวิทยาลัย, 2555.

[4] N. Phonson, "The rule-based machine translation system from Pali to Thai," Master's thesis Mahidol University, Bangkok, 2001.

[5] W. MALEELAI, "GRAPHEME TO PHONEME TRANSLATION FOR PALI-THAI," Master, 2013, Khon Kaen University.

[6] A. Wiansaow, "Simulate Program for Reading Pali Language of Conditional Conditional Random Fields," Master Master, King Mongkut's University of Technology North Bangkok, 2014.

[7] R. Chumkaew, "Retrieval System For Pali Dhammabot In Thai Alphabet," Master, King Mongkut's University of Technology North Bangkok, 2008.

[8] วัดพระธรรมกาย, สูตรสำเร็จ บาลีไวยากรณ์. เลี่ยงเชียง.

[9] C. Duroiselle, *A Practical grammar of the pali language* 3ed., 1997.

[10] B.-O. Kornwirat, "A PROGRAM FOR THE MACHINE TRANSLATION OF PALI INTO ENGLISH (PALI MT)," Master, Mahidol University. Bangkok (Thailand). Graduate School., 2003.

[11] B. Wanglem and N. Tongtep, "Pattern-Sensitive Loanword Estimation for Thai Text Clustering," 2017.

[12] P. P. Khaing and K. Z. Thwe, "Proposed Framework for Pali Words to Myanmar Text Translation," presented at the Thirteenth International Conferences on Computer Applications (ICCA 2015), 2015, 2015.

[13] Z. M. Maung and Y. Mikami, "Identification of Adopted Pali Words in Myanmar Text,"

2012.

[14] S. Mache and C. Mahender, "Development of Text-to-Speech Synthesizer for Pali Language," 2016.

[15] Y. Haribhakta and L. Nadageri, "Parts of Speech Tagger for Pali Language," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology,* 2017.

[16] J. Knauth and D. Alfter, "A Dictionary Data Processing Environment and Its Application in Algorithmic Processing of Pali Dictionary Data for Future NLP Tasks," in *Proceedings of the Fifth Workshop on South and Southeast Asian Natural Language Processing*, Dublin, Ireland, aug 2014: Association for Computational Linguistics and Dublin City University, pp. 65-73.

[17] F. Elwert, S. Sellmer, S. Wortmann, M. Pachurka, J. Knauth, and D. Alfter, "Toiling with the Pali Canon," 2016.

[18] D. Alfter, "Morphological analyzer and generator for Pali," Bachelor, 2014.

[19] R. Aralikatte, N. Gantayat, N. Panwar, A. Sankaran, and S. Mani, "Sanskrit Sandhi Splitting using seq2(seq)^2," in *the 2018 Conference on Empirical Methods in Natural Language Processing.*, Brussels, 2018: Belgium, pp. 4909–4914.

[20] S. Dave, A. K. Singh, P. A. P., and B. Lall, "Neural Compound-Word (Sandhi) Generation and Splitting in Sanskrit Language," 2020.

[21] S. Bhardwaj, N. Gantayat, N. Chaturvedi, R. Garg, and S. Agarwal, "Sandhikosh: A benchmark corpus for evaluating sanskrit sandhi tools," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018 May.

[22] A. Kumar, V. Mittal, and A. Kulkarni, "Sanskrit Compound Processor," Berlin, Heidelberg, 2010: Springer Berlin Heidelberg, in Sanskrit Computational Linguistics, pp. 57-69.

[23] C. N. D. S. B. Zadrozny, "Learning character-level representations for part-of-speech tagging," in *Proceedings of the 31st International Conference on International Conference on Machine Learning*, 2014, vol. 32.

[24] C. a. G. d. B. Dos Santos, Maira, "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts," 2014.

[25] T. Koomsubha and P. Vateekul, "A character-level convolutional neural network with dynamic input length for Thai text categorization," *2017 9th International Conference on Knowledge and Smart Technology (KST),* pp. 101-105, 2017.

[26] O. Khongtum, N. Promrit, and S. Waijanya, "Text-based LSTM Networks for Automatic Thai Love Quotes Generation on Twitter," *Information Technology Journal,* 2019.

[27] O. Khongtum, N. Promrit, and S. Waijanya, "The Entity Recognition of Thai Poem Compose by Sunthorn Phu by Using the Bidirectional Long Short Term Memory Technique," Cham, 2019: Springer International Publishing, in Multi-disciplinary Trends in Artificial Intelligence, pp. 97-108.

[28] I. G. Nils Reimers, "Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks," 2017.

[29] W. Y. a. K. K. a. M. Y. a. H. Schütze, "Comparative Study of CNN and RNN for Natural Language Processing," 2017.

[30] N. G. Rahul Aralikatte, Naveen Panwar, Anush Sankaran, Senthil Mani, "Sanskrit Sandhi Splitting using seq2(seq)^2," 2018.

มหาวิทยาลัยศิลปากร

มหาวิทยาลัยศิลปากร

ประวัติผู้เขียน

ชื่อ-สกุล	กลางใจ ธรรมนาม
วัน เดือน ปี เกิด	9 กุมภาพันธ์ 2537
สถานที่เกิด	กรุงเทพมหานคร