# การระบุอาการและอาการแสดงจากข้อความบอกเล่าอาการสำคัญภาษาไทยตามมาตรฐาน ICD-10 ด้วยขั้นตอนการประมวลผลภาษาธรรมชาติ

# ICD-10 Symptoms and Sign Identification from Thai Chief Complaints using Natural Language Processing


**ภวินท์ แซ่คู**

**Pawint Saeku**


**วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา**
**วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์**
**มหาวิทยาลัยสงขลานครินทร์**

**A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science**
**Prince of Songkla University**
**2561**

# การระบุอาการและอาการแสดงจากข้อความบอกเล่าอาการสำคัญภาษาไทยตามมาตรฐาน ICD-10 ด้วยขั้นตอนการประมวลผลภาษาธรรมชาติ

# ICD-10 Symptoms and Sign Identification from Thai Chief Complaints using Natural Language Processing

**ภวินท์ แซ่คู**

**Pawint Saeku**

**วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา**
**วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์**
**มหาวิทยาลัยสงขลานครินทร์**

**A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of**
**Master of Science in Computer Science**
**Prince of Songkla University**
**2561**

**ชื่อวิทยานิพนธ์** การระบุอาการและอาการแสดงจากข้อความบอกเล่าอาการสำคัญภาษาไทยตามมาตรฐาน ICD-10 ด้วยขั้นตอนการประมวลผลภาษาธรรมชาติ

**ผู้เขียน** นาย ภวินท์ แซ่คู

**สาขาวิชา** คอมพิวเตอร์

---

**อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก**

.....

(ดร.จารุณี ดวงสุวรรณ)

**คณะกรรมการสอบ**

.....ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ฐิมาพร เพชรแก้ว)

.....กรรมการ

(ดร.จารุณี ดวงสุวรรณ)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ลัดดา ปรีชาวีรกุล)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.วิภาดา เวทย์ประสิทธิ์)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

.....

(ศาสตราจารย์ ดร.ดำรงศักดิ์ ฟ้ารุ่งสาง)

คณบดีบัณฑิตวิทยาลัย

ขอรับรองว่า ผลงานวิจัยนี้มาจากการศึกษาวิจัยของนักศึกษาเอง และได้แสดงความขอบคุณบุคคลที่มีส่วนช่วยเหลือแล้ว

ลงชื่อ.....
(ดร.จารุณี ดวงสุวรรณ)
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

ลงชื่อ.....
(นายภวินท์ แซ่คู)
นักศึกษา

ข้าพเจ้าขอรับรองว่า ผลงานวิจัยนี้ไม่เคยเป็นส่วนหนึ่งในการอนุมัติปริญญาในระดับใดมาก่อน และไม่ได้ถูกใช้ในการยื่นขออนุมัติปริญญาในขณะนี้

ลงชื่อ .....
(นายภวินท์ แซ่คู)
นักศึกษา

ชื่อวิทยานิพนธ์ การระบุอาการและอาการแสดงจากข้อความบอกเล่าอาการสำคัญภาษาไทย ตามมาตรฐาน ICD-10 ด้วยขั้นตอนการประมวลผลภาษาธรรมชาติ
ผู้เขียน นาย ภวินท์ แซ่คู
สาขาวิชา วิทยาการคอมพิวเตอร์
ปีการศึกษา 2560

## บทคัดย่อ

ข้อความบอกเล่าอาการสำคัญ (Chief Complaint : CC) แสดงข้อมูลของอาการเจ็บป่วยที่นำผู้ป่วยมายังสถานพยาบาล โดยการจำแนก CC ออกเป็นกลุ่มอาการสามารถทำได้โดยใช้เครื่องมือสำหรับการเตรียม CC เพื่อทำให้มีความเหมาะสมและสอดคล้องสำหรับการนำไปจำแนกกลุ่มอาการ อย่างไรก็ตามเครื่องมือที่มีอยู่ในปัจจุบันส่วนใหญ่รองรับเฉพาะภาษาอังกฤษจึงทำให้เกิดข้อจำกัดสำหรับภาษาอื่นๆในการเข้าถึงเครื่องมือที่มีอยู่

งานวิจัยนี้เสนอวิธีจำแนกอาการจาก CC ภาษาไทยให้อยู่ในรูปของรหัสมาตรฐานที่เรียกว่า รหัส ICD-10 โดยอาศัยข้อความอาการและอาการแสดงจากเอกสาร ICD-10 ฉบับภาษาไทยเป็นข้อความตัวอย่าง โดยการเตรียมข้อมูลอาศัยวิธีการตัดคำแบบสองระดับ(Two-Level Tokenization : 2LT) และการตัดคำด้วยพจนานุกรม แล้วจึงนำผลลัพธ์ที่ได้เข้าสู่กระบวนการจำแนกข้อความด้วยวิธีการเรียนรู้ของเครื่องแบบต้นไม้ตัดสินใจ และสามารถปรับปรุงผลการจำแนกอาการและอาการแสดงให้มีประสิทธิภาพเพิ่มขึ้นโดยการเพิ่มค่า Precision จากวิธีรวมผลจำแนกของชุดข้อมูลต่างโครงสร้าง ในขณะที่การเพิ่มค่า Recall ทำได้จากการตัดผลการจำแนกที่คาดว่าจะเป็น false negative โดยอาศัยเกณฑ์ที่คำนวณจากไวยากรณ์ไม่พึ่งบริบท เพื่อกำหนดเกณฑ์ความไม่สอดคล้องระหว่างโครงสร้างของชุดข้อมูล CC และ ICD ที่มีรหัส ICD-10 เดียวกัน

ผลการทดลองจากชุดข้อมูลสองประเภทจากวิธีการตัดคำแบบ 2LT และ LM โดยชุดข้อมูล LM และ 2LT ได้ค่าเฉลี่ย Precision ของการจำแนกคือ 0.85 และ 0.88 ค่าเฉลี่ย recall ที่ 0.69 และ 0.67 ตามลำดับ หลังจากปรับใช้การรวมผลลัพธ์และตัดผลการจำแนกที่คาดว่าจะเป็น false negative โดยอาศัยชุดข้อมูล 2LT สำหรับตรวจสอบความสอดคล้องของข้อความ CC และ ICD ทำให้สามารถเพิ่มค่า Precision เป็น 0.93 และ Recall เป็น 0.83 ซึ่งแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถจำแนกอาการและอาการแสดง และเพิ่มประสิทธิภาพของผลลัพธ์การจำแนกได้

| | |
|---|---|
| Thesis Title | ICD-10 Symptoms and Sign Identification from Thai Chief Complaints using Natural Language Processing |
| Author | Mr. Pawint Saeku |
| Major Program | Computer Science |
| Academic Year | 2017 |

## Abstract

The free text chief complaint (CC) containing the information of patient symptoms could be classified into the symptom groups by using preprocessor to prepare the free text CC which is applied to the existing CC classified tools. However the most available tools support only English language, the other languages cannot use such tools for preparing and classifying.

This work proposes the tokenization algorithm which performs Thai CC text before classifying into ICD codes. The algorithm uses the symptom phrases from ICD-10 Thai modification document as a training data set. Using our proposed Two-level tokenization (2LT) to segment CC in preparing process, and a machine learning-based on decision tree classifier are applied to perform classification task. To improve the precision of experiment, we applied a union operation to classify results in order to increase true positive. Additionally, to increase the recall a context-free grammar technique is employed to calculate the criteria for eliminating a possible false negative based on the inconsistency in structure between ICD phrases and CC text.

The experiment has been done by using two different datasets including LM dataset based on longest matching segmentation and 2LT dataset based on two-level tokenization. The average Precision score of LM and 2LT is 0.85 and 0.88, and the average recall score of both are 0.69 and 0.67 respectively. After improving the precision and recall, the results show that the precision and recall change to 0.93 and 0.83 in orderly. These changing claims that the proposed method could perform and efficiency improve Thai CC text classification into symptom based on ICD standard.

# กิตติกรรมประกาศ

วิทยานิพนธ์นี้สามารถลุล่วงไปได้ด้วยความช่วยเหลือและสนับสนุนจากบุคคลหลายฝ่าย ผู้วิจัยรู้สึกซาบซึ้งในความกรุณาและขอกราบขอบพระคุณอย่างสูงมา ณ โอกาสนี้ คือ

ดร.จารุณี ดวงสุวรรณ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำปรึกษาและข้อเสนอแนะต่าง ๆ รวมทั้งช่วยตรวจทานและแก้ไขบทความวิจัยและวิทยานิพนธ์ให้แก่ผู้วิจัย

ผศ.ดร.ลัดดา ปรีชาวีรกุล อาจารย์ที่ปรึกษาร่วมวิทยานิพนธ์ ที่กรุณาให้คำปรึกษาและข้อเสนอแนะต่าง ๆ รวมทั้งช่วยตรวจทานและแก้ไขวิทยานิพนธ์ให้แก่ผู้วิจัย

อาจารย์และเจ้าหน้าที่ภาควิชาวิทยาการคอมพิวเตอร์ทุกท่านผู้ประสิทธิ์ประศาสตร์วิชาความรู้และอำนวยความสะดวกให้ผู้วิจัยทำให้งานวิจัยดำเนินไปได้ด้วยดี

บิดา มารดา และครอบครัวที่คอยเคียงข้างเป็นกำลังใจสนับสนุนทุกสิ่งทุกอย่างมาโดยตลอด

ผู้วิจัยรู้สึกซาบซึ้งและขอขอบพระคุณทุกท่านไว้ ณ ที่นี้ด้วย

ภวินท์ แซ่คู

# สารบัญ


**หน้า**

สารบัญ (8)

รายการตาราง (11)

รายการภาพประกอบ (12)

บทที่ 1 บทนำ ..... 1

1.1 ความสำคัญและที่มาของงานวิจัย ..... 1

1.2 วัตถุประสงค์ของงานวิจัย ..... 2

1.3 ขอบเขตการดำเนินงานของการวิจัย ..... 2

1.4 ขั้นตอนการดำเนินงาน ..... 3

1.5 ระยะเวลาดำเนินงาน ..... 3

1.6 สถานที่และเครื่องมือที่ใช้ในการดำเนินการวิจัย ..... 5

1.6.1 สถานที่ในการดำเนินงานวิจัย ..... 5

1.6.2 เครื่องมือและอุปกรณ์ที่ใช้ ..... 5

1.7 ประโยชน์ที่ได้รับ ..... 5

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง ..... 6

2.1 ข้อความบอกเล่าอาการสำคัญ (Chief Complaint : CC) ..... 6

2.2 บัญชีการจำแนกทางสถิติระหว่างประเทศและปัญหาสุขภาพที่เกี่ยวข้อง (ICD) ..... 8

2.3 โครงสร้างข้อมูลต้นไม้ (Tree data structure) ..... 10

2.3.1 ประเภทของโครงสร้างต้นไม้ ..... 11

2.4 การเรียนรู้ของเครื่อง (Machine Learning : ML) ..... 12

2.4.1 วิธีการจำแนกด้วยต้นไม้ช่วยตัดสินใจ (Decision Tree Classifier) ..... 14

2.4.2 กระบวนการสร้างโครงสร้างต้นไม้จากข้อมูลชุดสอน ..... 15

2.5 ข้อมูลอภิพันธุ์ (Metadata) ..... 17

2.6 คลังข้อความ (Corpus) ..... 19

2.7 การสกัดข้อมูลสารสนเทศ (Information Extraction :IE) ..... 21

2.8 N-gram ..... 22

2.9 รูปแบบและโครงสร้างภาษาไทย ..... 24

# สารบัญ (ต่อ)

หน้า

2.10 การตัดคำภาษาไทย (Thai Word Segmentation)..... 27
2.11 ไวยากรณ์ไม่พึ่งบริบท (Context-Free Grammars : CFG)..... 30
2.11.1 การสร้างโครงสร้างต้นไม้จาก CFG..... 31
2.12 Stop words..... 33
2.13 ประเมินผลการจำแนกข้อมูล..... 34
2.14 งานวิจัยที่เกี่ยวข้อง..... 35
บทที่ 3 การวิเคราะห์และออกแบบกระบวนการวิธีการจำแนกอาการและอาการแสดง..... 40
3.1 ที่มาของปัญหา..... 40
3.1.1 ขาดคลังข้อความที่เหมาะสม..... 40
3.1.2 มาตรฐานในการสื่อความหมาย..... 41
3.2 ข้อเสนอสำหรับแก้ปัญหา..... 41
3.2.1 การนำรหัส ICD มาใช้ในการระบุอาการและอาการแสดง..... 41
3.2.2 การใช้ข้อความหรือวลีจากเอกสาร ICD มาเป็นข้อมูลชุดสอน..... 41
3.3 การออกแบบกระบวนการจำแนกอาการและอาการแสดง..... 42
3.3.1 ภาพรวมของกระบวนการจำแนกอาการและอาการแสดง..... 42
3.4 กระบวนการเตรียมชุดข้อมูล ICD (ICD Dataset Preparation)..... 44
3.4.1 การศึกษาและวิเคราะห์ข้อความ..... 45
3.4.2 การตัดคำภาษาไทย..... 46
3.4.3 การหาความขัดแย้งกันของข้อความ (Conflict Element Finding)..... 47
3.4.4 การตรวจสอบความเกี่ยวข้องทางการแพทย์ (Medical Context Checking)..... 48
3.5 การเตรียมชุดข้อมูล CC..... 50
3.5.1 การตัดคำแบบสองระดับ (2 Level Tokenization : 2LT)..... 50
3.6 การเพิ่มประสิทธิภาพชุดข้อมูล (Corpus Enhancement)..... 52
3.6.1 การแก้ปัญหาความกำกวมด้วยวิธีทางสถิติ..... 53
บทที่ 4 ขั้นตอนการทดลองจำแนกอาการและอาการแสดงตามมาตรฐานรหัส ICD..... 55
4.1 ข้อมูลชุดสอนเครื่องและชุดทดสอบ..... 55

# สารบัญ (ต่อ)

หน้า

4.1.1 การเลือกชุดข้อมูลสำหรับการทดลอง ..... 55
4.2 การเตรียมข้อมูลสำหรับการทดสอบ ..... 56
4.2.1 การเตรียมชุดข้อมูล ICD ..... 57
4.2.1.1 Conflict Element Finding ..... 57
4.2.1.2 Medical Context Checking ..... 60
4.3 การเปรียบเทียบประสิทธิภาพของวิธีการเรียนรู้และและวิธีการจำแนก ..... 61
บทที่ 5 การเพิ่มประสิทธิภาพของการจำแนกอาการและอาการแสดง ..... 66
5.1 การพิจารณาค่าที่เกี่ยวข้องกับผลการจำแนก ..... 67
5.2 การเพิ่มปริมาณ *tp* ..... 68
5.2.1 การเพิ่มค่า *tp* จากผลการจำแนกจากข้อมูลประเภทเดียวกัน ..... 68
5.2.2 การเพิ่มปริมาณ *tp* โดยอาศัยข้อมูลจากผลจำแนกทั้งสองประเภท ..... 68
5.3 การลดปริมาณ *fn* ..... 70
5.3.1 เกณฑ์การตัดตัวเลือก ..... 71
5.3.2 การทดลองตัด *fn* ตามเกณฑ์ที่กำหนด ..... 77
บทที่ 6 การสรุปผลและงานในอนาคต ..... 80
6.1 การสกัดข้อมูลทางการแพทย์ที่ไม่ใช่ภาษาอังกฤษ ..... 81
6.2 วิธีการแก้ปัญหา ..... 82
6.3 การทดลองและการเพิ่มคุณภาพของผลลัพธ์ ..... 83
6.4 งานในอนาคต ..... 84
บรรณานุกรม ..... 86
ภาคผนวก ..... 92
ภาคผนวก ก ..... 93
ก.1 ชุดข้อมูลที่ในการทดลอง ..... 93
ก.2 ผลการจำแนกอาการภายในขั้นตอนการทดลอง ..... 114
ภาคผนวก ข ..... 127
ผลงานตีพิมพ์ในงานประชุมวิชาการ ICSEC 2016 ..... 127

# สารบัญ (ต่อ)

ประวัติผู้เขียน..... 133

## รายการตาราง

หน้า

2.1 ตัวอย่างชุดข้อมูลชุดสอน..... 15
2.2 ตัวอย่าง Part-of-Speech Tagging..... 19
2.3 ตัวอย่างข้อมูลในคลังข้อความภาษาไทย Orchid..... 20
2.4 การหาค่าความเป็นไปได้จากข้อความ "ฉันกินข้าวมันไก่" แบบ Bi-gram..... 23
2.5 โครงสร้างของนามวลี..... 25
2.6 โครงสร้างของกริยาวลี..... 26
2.7 แสดงตัวอย่างการติดป้ายสำหรับกลุ่มคำ ”ท่านผู้ฟังคะ” ที่แยกประเภทตามตัวอักษร..... 29
2.8 การเปรียบเทียบประสิทธิภาพของการตัดคำแต่ละรูปแบบ..... 30
3.1 ผลของการตัดคำด้วยวิธี LM..... 48
3.2 ผลการหา CE และ RE..... 48
4.1 ICD10token..... 58
4.2 Lexitron..... 58
4.3 TmpTable..... 59
4.4 ผลลัพธ์จากการเชื่อมตาราง TmpTable 2 ตาราง..... 60
4.5 ผลการหาความเกี่ยวของทางการแพทย์ของ CE และ RE..... 61
4.6 ข้อมูลที่ถูกแปลงด้วยวิธีการ Bag of Word..... 63
4.7 ตัวอย่างข้อมูลที่นำไปใช้ในการจำแนกด้วยชุด Classifier..... 63
4.8 การเปรียบเทียบระหว่างผลการทดลองของแต่ละการเรียนรู้ของเครื่อง..... 64
4.9 รายระเอียดของค่าการทำนายผลการจำแนกด้วยวิธี K-Nearest neighbor..... 64
5.1 เปรียบเทียบค่าเฉลี่ยระหว่างผลการทดลองชุดข้อมูล LM และ 2LT..... 66
5.2 แจกแจงการรวมผลลัพธ์ของสองชุดข้อมูล..... 69
5.3 ตารางเปรียบเทียบระหว่าง u3 และ ผลค่าเฉลี่ยของ 2LT*100 และ LM*100..... 70

## รายการตาราง (ต่อ)

5.4 ตัวอย่างการแจกแจงรูปแบบตามฟังก์ชัน $f_1(n)$..... 75
5.5 ตัวอย่างการแจกแจงรูปแบบตามฟังก์ชัน $f_2(n)$..... 75
5.6 การปรับสมการให้สอดคล้องกัน..... 76
5.7 ผลการอินเตอร์เซคแต่ละเกณฑ์จากโครงสร้าง 2LT เปรียบเทียบกับผลจำแนก u3..... 78
5.8 ผลการอินเตอร์เซคแต่ละเกณฑ์จากโครงสร้าง LM เปรียบเทียบกับผลจำแนก u3..... 78

## รายการภาพประกอบ

**หน้า**

2.1 ตัวอย่างโครงข้อมูลสร้างต้นไม้..... 10
2.2 ตัวอย่างโครงสร้าง General Tree..... 11
2.3 Full Binary Tree..... 12
2.4 Complete Binary Tree..... 12
2.5 ตัวอย่างต้นไม้ตัดสินใจ..... 14
2.6 ตัวอย่างโครงสร้างต้นไม้จากไวยากรณ์ CFG..... 31
2.7 การเริ่มสร้างต้นไม้จากบนลงล่าง..... 32
2.8 การเริ่มสร้างต้นไม้จากล่างขึ้นบน..... 32
2.9 ตัวอย่างการสร้างต้นไม้แบบบนลงล่างและขยายโหนดจากทางซ้ายสุด..... 33
3.1 กระบวนการจำแนกอาการและอาการแสดง..... 43
4.1 กระบวนการเปรียบเทียบผลระหว่างชุดข้อมูลทดลองและวิธีการจำแนก..... 62
5.1 โครงสร้างต้นไม้ที่มี c มากที่สุดที่เป็นไปได้จำนวนคี่..... 72
5.2 โครงสร้างต้นไม้ที่มี c มากที่สุดที่เป็นไปได้จำนวนคู่..... 73
5.3 การเปรียบเทียบความสอดคล้องข้อมูล ICD และ ข้อมูล CC..... 77

# บทที่ 1

# บทนำ

## 1.1 ความสำคัญและที่มาของงานวิจัย

การระบุโรคหรือความผิดปกตินั้นมีความจำเป็นต้องอาศัยข้อมูลของผู้ป่วยที่มีนัยสำคัญต่อการวินิจฉัยอาการเจ็บป่วย เช่น อายุ เพศ ประวัติการเจ็บป่วย ข้อความบอกเล่าอาการสำคัญเป็นต้น โดยข้อมูลดังกล่าวเป็นส่วนหนึ่งของเวชระเบียนหรือเอกสารทางการแพทย์ ซึ่งเวชระเบียนและสถิติ (Medical Record and Statistics) ที่มีหน้าที่ในการจัดการข้อมูลเวชระเบียน จะเป็นผู้ดูแลและจัดการข้อมูลและหลักฐานทางการแพทย์ทั้งเชิงปริมาณและเชิงคุณภาพ สำหรับข้อเสียของการจัดเก็บข้อมูลในรูปแบบกระดาษหรือเอกสารคือ การพิจารณาและบันทึกข้อมูลต้องอาศัยผู้ชำนาญการและเวลาในการดำเนินการ รวมทั้งการค้นหาข้อมูลเพื่อนำไปใช้ไม่สะดวกและมีความยุ่งยาก นอกจากนี้แล้วยังมีการเก็บข้อมูลเวชระเบียนในรูปแบบดิจิทัลซึ่งเรียกว่า เวชระเบียนอิเล็กทรอนิกส์ (Electronic Health Record : EHR) [1] สำหรับ EHR เป็นข้อมูลที่ถูกจัดการด้วยระบบที่ถูกออกแบบที่มีจุดประสงค์เพื่อแบ่งปันข้อมูลผู้ป่วยดิจิทัลระหว่างเครือข่ายของโรงพยาบาล คลินิก หรือ สถาบันทางการแพทย์ การเก็บข้อมูลในรูปแบบดิจิทัลสามารถช่วยในการแก้ไขปัญหาเรื่องความยากของการเข้าถึงข้อมูลแบบเอกสาร โดยอาศัยวิธีการค้นหาด้วยคำสำคัญหรือ key word ทำให้การเข้าถึงข้อมูลสามารถทำได้ง่ายขึ้น อย่างไรก็ตามการบันทึกข้อมูลแบบดิจิทัลก็ไม่ได้แก้ปัญหาเรื่องการพิจารณาและกรอกข้อมูล ยังคงต้องอาศัยคนและเวลาในกระบวนการเช่นเดิม

การจำแนกและการระบุโรคหรือความผิดปกติจำเป็นต้องอาศัยปัจจัยที่สำคัญ 2 ปัจจัย คือ บุคลากร และ เวลา หากสามารถนำคอมพิวเตอร์เข้ามาช่วยในการประมวลผลกระบวนการดังกล่าวจะสามารถลดปริมาณปัจจัยที่จำเป็นในกระบวนการดังกล่าวได้ จึงเกิดปัญหาขึ้นว่าทำอย่างไรให้สามารถนำการประมวลผลทางคอมพิวเตอร์มาประยุกต์ใช้ในการจำแนกและระบุโรคหรือความผิดปกติได้ โดยเงื่อนไขสำคัญในการวินิจฉัยอาการและความผิด ปกติจำเป็นต้องอาศัยข้อมูลส่วนตัวของผู้ป่วยที่ผ่านกระบวนการวิเคราะห์เพื่อนำไปประกอบการพิจารณา ได้แก่ ข้อมูลอาการและอาการแสดงจากข้อความบอกเล่าอาการสำคัญ อาการเจ็บป่วยที่เคยเกิดขึ้นในประวัติการเจ็บป่วย อายุ และ เพศ แต่ข้อมูล อายุ และ เพศ รูปแบบการเก็บข้อมูลมีความคงเส้นคงวา

ทำให้ไม่ประสบปัญหาเมื่อต้องนำมาใช้ในการประมวลผลทางคอมพิวเตอร์ แต่ข้อมูลจำพวกประวัติการเจ็บป่วยและข้อความบอกเล่าอาการสำคัญมักจะถูกเก็บอยู่ในรูปแบบภาษาธรรมชาติที่มีความแตกต่างกันไปตามภาษาประจำชาติและโครงสร้างของภาษานั้นๆ เป็นผลให้การประยุกต์ใช้การประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) มีข้อจำกัดแตกต่างกันไปตามโครงสร้างของภาษา นอกจากการประมวลผลภาษาแล้ว การเรียนรู้ของเครื่อง (Machine Learning : ML) ก็เป็นอีกเทคนิคหนึ่งที่นำมาประยุกต์ใช้สำหรับการจำแนกข้อมูล แต่เทคนิคนี้มีความต้องการข้อมูลชุดสอนที่เหมาะสมจำนวนมากเพื่อให้ได้ผลลัพธ์ที่มีความแม่นยำสูง โดยปัจจุบันข้อมูลชุดสอนทางการแพทย์ภาษาไทยยังไม่เป็นที่แพร่หลาย และยังมีงานวิจัยที่เกี่ยวข้องน้อยมาก ดังนั้นจึงไม่สามารถนำเทคนิค ML มาประยุกต์ใช้สำหรับการพัฒนาระบบ NLP จนสามารถนำไปใช้ได้อย่างมีประสิทธิภาพได้

## 1.2 วัตถุประสงค์ของงานวิจัย

เพื่อวิเคราะห์และออกแบบขั้นตอนวิธี (Algorithm) เพื่อนำไปใช้พัฒนาระบบสำหรับการแยกแยะอาการและการแสดงอาการจากข้อมูลข้อความ บอกเล่าอาการสำคัญ (Chief Complaint) ในรูปแบบภาษาไทยตามมาตรฐาน ICD-10

## 1.3 ขอบเขตการดำเนินงานของการวิจัย

- ศึกษาและออกแบบกระบวนการเตรียมชุดข้อมูลเพื่อใช้สำหรับจำแนกอาการและอาการแสดง
- ศึกษาและออกแบบกระบวนการจำแนกอาการและอาการแสดงตามมาตรฐาน ICD-10 กลุ่ม R ในช่วงรหัสตั้งแต่ R00 ถึง R69

## 1.4 ขั้นตอนการดำเนินงาน

1. ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง
2. ศึกษาเทคโนโลยีและเครื่องมือสำหรับงานวิจัย
3. วิเคราะห์และออกแบบขั้นตอนวิธีการสกัดอาการ
4. พัฒนาและทดสอบขั้นตอนวิธีที่ได้ออกแบบไว้
5. เขียนบทความวิจัย
6. จัดทำเอกสารวิทยานิพนธ์

## 1.5 ระยะเวลาดำเนินงาน

| ขั้นตอนการดำเนินงาน | พ.ศ. 2559 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1. ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง | ← | — | — | — | — | — | — | — | — | — | — | → |
| 2. ศึกษาเทคโนโลยีและเครื่องมือสำหรับงานวิจัย | | | ← | — | — | — | — | — | — | — | → | |
| 3. วิเคราะห์และออกแบบขั้นตอนวิธีการสกัดอาการ | | | ← | — | — | — | — | — | — | — | → | |
| 4. พัฒนาและทดสอบประสิทธิภาพของขั้นตอนวิธี | | | | | | | | | | | | |
| 5. เขียนบทความวิจัย | | | | | | | | | | | | |
| 6. จัดทำเอกสารวิทยานิพนธ์ | | | | | | | | | | | | |

| ขั้นตอนการดำเนินงาน | พ.ศ. 2560 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1. ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง | ← | — | — | — | → | | | | | | | |
| 2. ศึกษาเทคโนโลยีและเครื่องมือสำหรับงานวิจัย | ← | — | — | — | → | | | | | | | |
| 3. วิเคราะห์และออกแบบขั้นตอนวิธีการสกัดอาการ | ← | — | — | — | → | | | | | | | |
| 4. พัฒนาและทดสอบประสิทธิภาพของขั้นตอนวิธี | ← | — | — | — | — | — | — | — | — | — | → | |
| 5. เขียนบทความวิจัย | ← | — | — | — | — | — | — | — | — | — | → | |
| 6. จัดทำเอกสารวิทยานิพนธ์ | | | | ← | — | — | — | — | — | — | — | → |

| ขั้นตอนการดำเนินงาน | พ.ศ. 2561 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1. ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง | | | | | | | | | | | | |
| 2. ศึกษาเทคโนโลยีและเครื่องมือสำหรับงานวิจัย | | | | | | | | | | | | |
| 3. วิเคราะห์และออกแบบขั้นตอนวิธีการสกัดอาการ | | | | | | | | | | | | |
| 4. พัฒนาและทดสอบประสิทธิภาพของขั้นตอนวิธี | | | | | | | | | | | | |
| 5. เขียนบทความวิจัย | | | | | | | | | | | | |
| 6. จัดทำเอกสารวิทยานิพนธ์ | ← | — | — | → | | | | | | | | |

←——→ แผนการดำเนินงาน

## 1.6 สถานที่และเครื่องมือที่ใช้ในการดำเนินการวิจัย

### 1.6.1 สถานที่ในการดำเนินงานวิจัย

ห้องปฏิบัติการ CS207 ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์

### 1.6.2 เครื่องมือและอุปกรณ์ที่ใช้

1. เครื่องคอมพิวเตอร์ส่วนบุคคลที่มีคุณลักษณะดังนี้
   - CPU : Intel Core I5 2.60 GHz
   - Hard disk : 1 TB
   - RAM : 8 GB
2. ด้านซอฟต์แวร์
   - ระบบปฏิบัติการ Windows 7 64 bit
   - Python 3.5
   - Java development kit 6
   - MySQL database
   - PHP version 5.6.28

## 1.7 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ขั้นตอนวิธี(Algorithm)สำหรับการแยกแยะอาการและการแสดงอาการจากข้อมูล Chief Complaint ในรูปแบบภาษาไทยตามมาตรฐาน ICD-10
2. ได้ระบบสำหรับการแยกแยะอาการและการแสดงอาการจากข้อมูล Chief Complaint ในรูปแบบภาษาไทยตามมาตรฐาน ICD-10
3. ผู้ใช้งานได้รับรหัสอาการที่จำแนกตามมาตรฐานรหัส ICD-10 จากข้อความ Chief Complaint

# บทที่ 2

# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงแนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวกับการดำเนินงานวิจัยโดยจะกล่าวถึง ข้อความบอกเล่าอาการสำคัญ บัญชีการจำแนกทางสถิติระหว่างประเทศและปัญหาสุขภาพที่เกี่ยวข้อง กระบวนการประมวลผลภาษา วิธีการจำแนกประเภทของข้อมูล และงานวิจัยที่เกี่ยวข้อง

## 2.1 ข้อความบอกเล่าอาการสำคัญ (Chief Complaint : CC)

ข้อความบอกเล่าอาการสำคัญ (Chief Complaint : CC) เป็นเป็นหนึ่งของประวัติผู้ป่วยโดยการเก็บข้อมูลประวัติผู้ป่วยประกอบด้วยรายละเอียดดังนี้ [2]

- ข้อมูลทั่วไปเกี่ยวกับผู้ป่วย (Patient Profile) ได้แก่ ชื่อ เพศ อายุ น้ำหนัก อาชีพ เป็นต้น
- อาการนำ (Chief Complaint) คืออาการนำที่ทำให้ผู้ป่วยต้องมาพบแพทย์และระยะเวลาที่เกิดอาการ
- ประวัติปัจจุบัน (Present Illness) คือรายละเอียดของอาการนำ ต้องเกี่ยวข้องกับ CC เท่านั้น
- ประวัติอดีต (Past History) คือประวัติเจ็บป่วยในอดีตที่อาจเกี่ยวข้องกับการเจ็บป่วยปัจจุบัน นับรวมถึงประวัติของการเข้ารับการผ่าตัด เช่น เคยรับการผ่าตัดช่องท้องเมื่อ 2 ปีก่อนหรือ โรคประจำตัวต่างๆ เช่น เบาหวาน
- ประวัติส่วนตัว (Personal History) ส่วนนี้มักรวมถึงประวัติแพ้ (allergy) ทั้งแพ้ยาและสารอื่นๆ, ประวัติสารเสพติด สุรา บุหรี่
- ประวัติครอบครัว (Family History) คือประวัติเกี่ยวกับสุขภาพของครอบครัว โรคประจำตัวในครอบครัว โดยเฉพาะที่อาจเกี่ยวข้องกับการเจ็บป่วยปัจจุบัน
- ประวัติสังคมเศรษฐกิจ (Socioeconomic History)
- ประวัติทางเพศ (Sexual History) เช่น ประจำเดือน เพศสัมพันธ์ การป้องกัน การคุมกำเนิด เป็นต้น

- การทบทวนตามระบบ (Review of Systems) คือการเจ็บป่วยและสุขภาพทั่วไป โดยเป็นสอบถามเกี่ยวกับระบบอวัยวะต่างๆ เช่น ระบบทั่วไป ระบบผิวหนัง ระบบไหลเวียนโลหิต ฯลฯ

ข้อความบอกเล่าอาการสำคัญ (CC) เป็นข้อความที่บอกเล่าอาการสำคัญที่นำผู้ป่วยมายังสถานพยาบาล โดยภายในข้อความ CC ประกอบไปด้วย อาการ หรือ อาการแสดง ซึ่ง ไม่มีชื่อของ โรค หรือ ความผิดปกติ อยู่ภายในข้อความ CC โดยทั่วไปรูปแบบของข้อความ CC จะเริ่มด้วยลักษณะของอาการหรืออาการแสดง และตามด้วยระยะเวลาที่เกิดขึ้น ตัวอย่างข้อความบอกเล่าอาการสำคัญ เช่น

"มีอาการปวดหัว เป็นไข้ เป็นมา 3 วัน"

"มีอาการปวดท้องมาเป็นเวลา 5 ชั่วโมง"

โดยข้อมูลในข้อความ CC มีความจำเป็นต่อการนำไปวินิจฉัยโรคหรือความผิดปกติเพื่อนำไปสู่การหาวิธีการบำบัดหรือรักษาผู้ป่วย การจำแนกและระบุข้อมูล CC จะถูกจำแนกเป็น *โรค* หรือ *ความผิดปกติ* เท่านั้น หากไม่สามารถจำแนกเป็น *โรค* หรือ *ความผิดปกติ* ได้จึงจะถูกระบุเป็น อาการ หรือ อาการแสดงแทน โดยสามารถให้คำจำกัดความของ "โรค" "ความผิดปกติ" "อาการ" และ "อาการแสดง" ได้ดังนี้

| | | |
|---|---|---|
| ● โรค | (Disease) | หมายถึงโรคทั่วไปทั้งมีสาเหตุจากเชื้อโรคและไม่ได้มาจากเชื้อโรค |
| ● ความผิดปกติ | (Disorders) | หมายถึงการแสดงกลุ่มอาการรวมกัน เช่น ความผิดปกติทางจิต |
| ● อาการแสดง | (Sign) | หมายถึงความผิดปกติที่เกิดขึ้นกับผู้ป่วยและแพทย์ผู้ตรวจสามารถตรวจพบได้ เช่น ผื่น เป็นอาการแสดงที่สามารถเห็นอย่างชัดเจน เป็นอาการระคายเคืองมีลักษณะบวมแดงตามผิวหนัง |
| ● อาการ | (Symptoms) | หมายถึงอาการที่มาจากความรู้สึกของผู้ป่วยที่สามารถทราบได้จากการบอกเล่าของผู้ป่วยเท่านั้น เช่น อาการปวด หรือ เจ็บ ไม่สามารถทราบหากผู้ป่วยไม่ได้แจ้งให้ทราบ |

## 2.2 บัญชีการจำแนกทางสถิติระหว่างประเทศและปัญหาสุขภาพที่เกี่ยวข้อง (ICD)

การระบุข้อมูลเกี่ยวกับทางการแพทย์เริ่มขึ้นในปี ค.ศ. 1860 เมื่อได้มีการเสนอให้จัดทำระบบการจัดเก็บข้อมูลของโรงพยาบาลในงานคองเกรสสถิตินานาชาติที่กรุงลอนดอน ประเทศอังกฤษและใน 1893 นายแพทย์ชาวฝรั่งเศสชื่อ Jacques Bertillon ได้เสนอการระบุสาเหตุการตายหรือ “Bertillon Classification of Causes of Death” [3] ในงานคองเกรสของ International Statistical Institute ที่เมืองชิคาโก ภายหลังได้มีหลายประเทศนำไปปรับใช้และเพิ่มจำนวนของรายชื่อสาเหตุมากขึ้น ในปี 1898 American Public Health Association (APHA) องค์สาธารณสุขในประเทศสหรัฐอเมริกาได้เสนอให้มีการแก้ไขทุกสิบปีเพื่อให้สอดคล้องกับการพัฒนาทางการแพทย์ ฉบับแก้ไขฉบับแรกมีชื่อว่า “International Classification of Causes of Death” จัดทำขึ้นในปี ค.ศ. 1900 และได้มีการเปลี่ยนชื่อเป็น International Statistical Classification of Diseases (ICD) ในฉบับแก้ไขครั้งที่ 6 (ICD-6) ได้ถูกจัดทำภายใต้การรวมกันของคณะกรรมาธิการของ the International Statistical Institute และองค์กรอนามัยของ League of Nations และตั้งแต่ฉบับแก้ไขที่ 7 เป็นต้นไป องค์การอนามัยโลก (World Health Organization : WHO) ได้เข้ามามีบทบาทในการรับหน้าที่เตรียมและเผยแพร่ ICD โดยฉบับแก้ไขสมบูรณ์ปัจจุบันเป็นฉบับที่ 10 มีชื่อเต็มอย่างเป็นทางการว่า “บัญชีการจำแนกทางสถิติระหว่างประเทศและปัญหาสุขภาพที่เกี่ยวข้อง” (International Statistical Classification of Diseases and Related Health Problems (ICD-10) ในขณะที่ ICD-11 ถูกวางแผนให้เสร็จสมบูรณ์ในปี ค.ศ. 2017 แต่ไม่สามารถทำเสร็จสมบูรณ์ได้ทันเวลาจึงถูกเลื่อนออกไปเป็นปี 2018

ICD-10 มีจุดประสงค์สำหรับการจำแนกโรค อาการ อาการแสดง ความผิดปกติที่ตรวจพบ อาการนำ สภาพสังคม หรือ สาเหตุภายนอกของการบาดเจ็บหรือโรคออกมาในรูปแบบของรหัส หรือ code ICD-10 ฉบับมาตรฐานมีรหัสที่แตกต่างกัน 14,400 รหัส นอกจากนี้ยังมีฉบับแก้ไขที่นำมาปรับใช้สำหรับบางประเทศ เช่น ICD-10 Clinical Modification (ICD-10-CM) ของสหรัฐอเมริกา หรือ ICD-10 Thai Modification (ICD-10-TM) เป็นต้น รหัสของ ICD-10 ถูกออกแบบเป็นตัวอักษรภาษาอังกฤษตามด้วยตัวเลขและทำการแบ่งรหัสออกเป็นกลุ่มได้ดังนี้

- A00-B99 โรคติดเชื้อและโรคปรสิตบางโรค
- C00-D48 เนื้องอก
- D50-D89 โรคของเลือดและอวัยวะสร้างเลือดและความผิดปกติบางอย่างของกลไกภูมิคุ้มกัน

- E00-E90 โรคของต่อมไร้ท่อ โภชนาการ และเมตะบอลิซึม
- F00-F99 ความผิดปกติทางจิตและพฤติกรรม
- G00-G99 โรคของระบบประสาท
- H00-H59 โรคของตาและอวัยวะเคียงลูกตา
- H60-H95 โรคของหูและปุ่มกระดูกกกหู
- I00-I99 โรคของระบบไหลเวียนโลหิต
- J00-J99 โรคของระบบหายใจ
- K00-K93 โรคของระบบย่อยอาหาร
- L00-L99 โรคของผิวหนังและเนื้อเยื่อใต้ผิวหนัง
- M00-M99 โรคของระบบกล้ามเนื้อโครงร่าง และเนื้อเยื่อเกี่ยวพัน
- N00-N99 โรคของระบบสืบพันธุ์และระบบปัสสาวะ
- O00-O99 การตั้งครรภ์ การคลอด และระยะหลังคลอด
- P00-P96 ภาวะบางอย่างที่เริ่มต้นในระยะปริกำเนิด
- Q00-Q99 รูปผิดปกติแต่กำเนิด รูปพิการ และความผิดปกติของโครโมโซม
- R00-R99 อาการ อาการแสดง และความผิดปกติที่พบจากการตรวจทางคลินิกและทางห้องปฏิบัติการ มิได้จำแนกไว้ที่ใด
- S00-T98 การบาดเจ็บ การเป็นพิษ และผลสืบเนื่องบางอย่างจากสาเหตุภายนอก
- V01-Y98 สาเหตุภายนอกของการเจ็บป่วยและการตาย
- Z00-Z99 ปัจจัยที่มีผลต่อสถานะสุขภาพและการรับบริการสุขภาพ
- U00-U99 รหัสเพื่อวัตถุประสงค์พิเศษ

ในรหัสของ ICD-10 ประกอบด้วยรหัสของหัวข้อหลักและหัวข้อย่อย โดยหัวข้อย่อยจะมีหรือไม่ก็ได้ อย่างไรก็ตามหากมีหัวข้อย่อยจะไม่นำหัวข้อหลักมาใช้ในการระบุแต่จะระบุหัวข้อย่อยแทน เช่น

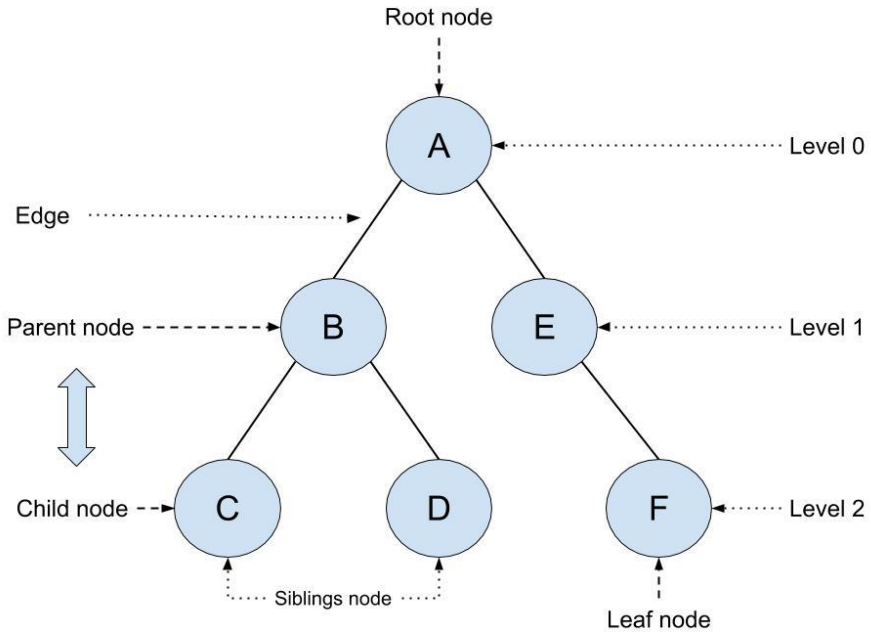
- R10 ปวดท้องและปวดอุ้งเชิงกราน
  - R10.0 ปวดท้องเฉียบพลัน
  - R10.1 ปวดเฉพาะท้องส่วนบน
  - R10.2 ปวดอุ้งเชิงกรานและฝีเย็บ
  - R10.3 ปวดท้องน้อยส่วนอื่น
  - R10.4 ปวดท้องแบบอื่นและไม่ระบุรายละเอียด
- R05 ไอ

จากตัวอย่างข้างต้นสำหรับรหัส R10 “ปวดท้องและปวดอุ้งเชิงกราน” จะไม่สามารถนำมาใช้ได้ เนื่องจากเป็นรหัสที่มีหัวข้อย่อยสำหรับจำแนกรายละเอียดแต่ R05 “ไอ” ซึ่งเป็นหัวข้อหลัก สามารถนำมาใช้เนื่องจากไม่มีหัวข้อย่อย

## 2.3 โครงสร้างต้นไม้ (Tree Data Structure)

โครงสร้างต้นไม้ [4] เป็นโครงสร้างของข้อมูลที่เชื่อมต่อกันด้วยเส้นทางทั้งแบบมีทิศทางและไม่มีทิศทาง โครงสร้างต้นไม้สามารถช่วยในการแก้ปัญหาต่างๆ เช่น การหาค่าใช้จ่ายที่น้อยที่สุด การช่วยตัดสินใจ หรือการเรียงลำดับ และ การจัดกลุ่มข้อมูลเป็นต้น ลักษณะของโครงสร้างต้นไม้ [5] แสดงอยู่ในรูปที่ 2.1 โดยหน่วยที่อยู่ในโครงสร้างต้นไม้ถูกเรียกว่า “โหนด” (Node) เชื่อมต่อกันด้วย “เส้นเชื่อม” (Edge) สามารถแยกประเภทโหนดและองค์ประกอบได้ดังนี้

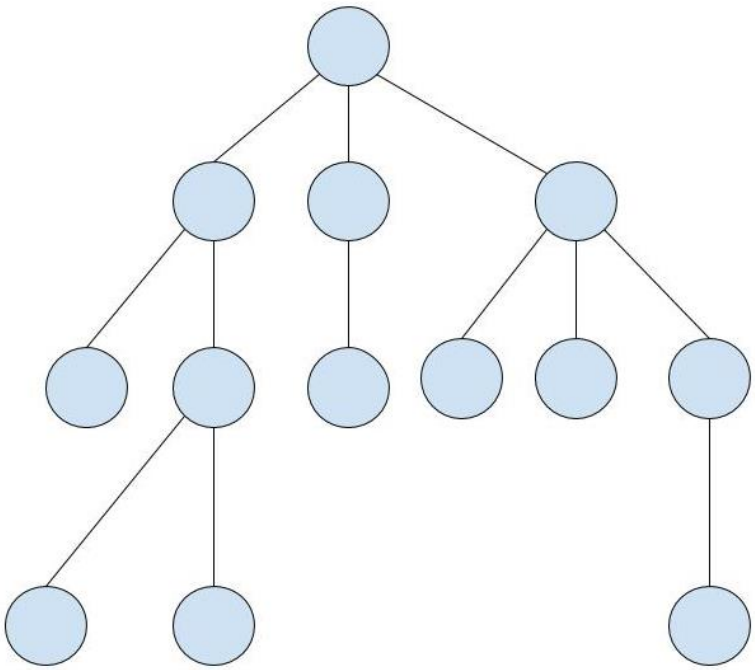
- โหนดราก (Root node) คือโหนดที่อยู่บนสุดของโครงสร้างต้นไม้
- โหนดพ่อ (Parent node) คือโหนดที่อยู่ในระดับที่สูงกว่าของโหนดลูก
- โหนดลูก (Child node) คือโหนดที่อยู่ในระดับที่ต่ำกว่าโหนดพ่อ
- โหนดใบ (Leaf node) คือโหนดที่ไม่มีโหนดลูก
- โหนดพี่น้อง (Siblings node) คือโหนดในระดับเดียวกันที่เกิดจากโหนดพ่อเดียวกัน
- ระดับชั้นความลึก (Level) คือ ระดับชั้นของต้นไม้
- เส้นเชื่อม (Edge) คือ เส้นทางที่เชื่อมต่อระหว่างโหนด




รูปที่ 2.1 ตัวอย่างโครงข้อมูลสร้างต้นไม้ (ดัดแปลงมาจาก [5])

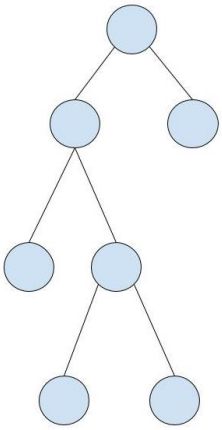
**2.3.1 ประเภทของโครงสร้างต้นไม้**

ประเภทของโครงสร้างต้นไม้สามารถแบ่งเป็นประเภทตามองค์ประกอบของโหนด General Tree หรือโครงสร้างต้นไม้ทั่วไป เป็นโครงสร้างต้นไม้ที่แต่ละโหนดสามารถมีโหนดลูกหรือไม่มีก็ได้และไม่จำกัดปริมาณของโหนดลูก โดยส่วนใหญ่โครงสร้างต้นไม้ทั่วไปถูกใช้สำหรับระบบการจัดการไฟล์

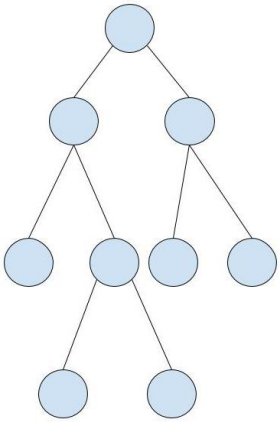


รูปที่ 2.2 ตัวอย่างโครงสร้าง General Tree

- Binary Tree (ต้นไม้ทวิภาค) เป็นโครงสร้างต้นไม้ที่มีโหนดลูกได้ไม่เกิน 2 โหนด โครงสร้างต้นไม้ประเภทนี้ สามารถใช้ประยุกต์ในกระบวนการแก้ปัญหาประเภท การช่วยตัดสินใจ หรือ การค้นหาข้อมูลได้ โดยแบ่งลักษณะเป็นสองแบบ
    - Full Binary Tree คือ ต้นไม้ทวิภาคที่ทุกโหนดพ่อมีโหนดลูก 2 โหนด ดังรูปที่ 2.3
    - Complete Binary Tree คือ ต้นไม้ทวิภาคที่ระดับชั้นความลึกที่ลึกที่สุด ห่างจากลำดับชั้นที่อยู่ติดกันไม่มากกว่า 1 ระดับ ดังรูปที่ 2.4



รูปที่ 2.3 Full Binary Tree



รูปที่ 2.4 Complete Binary Tree

## 2.4 การเรียนรู้ของเครื่อง (Machine Learning : ML)

การเรียนรู้ของเครื่องเป็นกระบวนการที่ทำให้คอมพิวเตอร์สามารถเรียนรู้จากข้อมูลที่มีอยู่ โดย ทอม เอ็ม มิตเชลล์ (Tom M. Mitchell) [6] ได้ให้คำจำกัดความอย่างเป็นทางการว่า “เราจะเรียกคอมพิวเตอร์โปรแกรมว่าได้เรียนรู้จากประสบการณ์ E เพื่อทำงาน T ได้โดยมีประสิทธิผล P เมื่อโปรแกรมนั้นสามารถทำงาน T ที่วัดผลด้วย P แล้วพัฒนาขึ้นจากประสบการณ์ E” หรือกล่าวได้ว่า หากการทำงานของโปรแกรมคอมพิวเตอร์สามารถทำงานได้ดีขึ้นจากข้อมูลที่มีจุดประสงค์เพื่อใช้สำหรับการสอนเครื่องจึงสามารถเรียกว่าเป็นการเรียนรู้ของเครื่องได้ ML ถูกนำมาประยุกต์ใช้งานในหลายด้านเช่น การช่วยทำนายผล การช่วยตัดสินใจ หรือ จำแนกประเภทของข้อมูล ในขณะเดียวกันการเรียนรู้ของเครื่องก็เป็นที่นิยมในการประยุกต์ใช้กับงานทางการแพทย์ เช่นการระบุหรือจำแนก ชื่อโรค ยา หรือ เอกลักษณ์อื่นๆทางการแพทย์ โดยเรียกงานจำพวกนี้ว่าการจดจำเอกลักษณ์ (Name Entity Recognition : NER) ภายในเอกสารซึ่งมีปริมาณข้อมูลจำนวนมากทำให้ใช้เวลาและจำนวนคนในการระบุ หรือ การให้ข้อมูลประกอบการวินิจฉัยโรคจึงทำให้การใช้การเรียนรู้ของเครื่องเหมาะกับการนำมาประยุกต์ใช้กับปัญหาที่ต้องจัดการกับข้อมูลจำนวนมาก โดยสามารถแบ่งประเภทของการเรียนรู้ของเครื่องได้จากงานและการจัดการปัญหาได้ดังนี้ [7], [8]

1. การเรียนรู้ของเครื่องแบ่งตามประเภทงานของข้อมูลนำเข้า

- การเรียนรู้แบบมีผู้สอน (Supervised Learning) การเรียนรู้แบบชุดข้อมูลสอนมีตัวอย่างชุดข้อมูลนำเข้าและผลลัพธ์ โดยมีจุดประสงค์ของการเรียนรู้คือต้องการให้เครื่องสามารถเรียนรู้รูปแบบหรือกฎในการจับคู่ระหว่างข้อมูลขาเข้าและข้อมูลขาออก
- การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ไม่มีการติดป้ายในข้อมูลชุดสอนให้เครื่องหาโครงสร้างของข้อมูลขาเข้าเอง
- การเรียนรู้แบบเสริมกำลัง (Reinforcement Learning) คอมพิวเตอร์มีปฏิสัมพันธ์กับสิ่งแวดล้อมที่เปลี่ยนไปตลอดเวลาโดยคอมพิวเตอร์จะต้องทำงานบางอย่างที่มีเป้าหมายชัดเจน
- การเรียนรู้แบบกึ่งมีผู้สอน (Semi Supervised Learning) เป็นการสอนแบบมีการบอกข้อมูลผลลัพธ์สำหรับข้อมูลนำเข้าบางชุด โดยที่ไม่ใช่ทุกชุดข้อมูลขาเข้าจะมีการระบุผลลัพธ์จึงถูกเรียกว่าการเรียนรู้แบบกึ่งมีผู้สอน
    - ทรานสดักชัน (Transduction) เป็นกรณีพิเศษของการเรียนรู้แบบกึ่งมีผู้สอน ที่ในข้อมูลชุดสอนมีทั้งชุดข้อมูลเข้าทั้งที่ถูกติดและไม่ถูกติดป้ายผลลัพธ์ โดยมีจุดประสงค์ในการติดป้ายให้กับข้อมูลขาเข้าที่ไม่มีการติดป้ายผลลัพธ์ ไม่ใช่การสร้างแบบจำลองการเรียนรู้
- การเรียนวิธีการเรียน (Learning to Learn, Meta-Learning) เป็นวิธีการเรียนรู้วิธีเรียนของตัวเอง โดยเป็นการปรับปรุงอัลกอริทึมที่ใช้ในการสมมติฐานผลลัพธ์จากข้อมูลชุดสอนหรือประสบการณ์ที่ผ่านมาในการทำนายผลลัพธ์ของชุดข้อมูลที่ยังไม่เคยเจอ

2. การเรียนรู้ของเครื่องแบ่งตามประเภทงานตามข้อมูลผลลัพธ์

- การแบ่งประเภทข้อมูล (Classification) คือการสร้างแบบจำลองที่สามารถจำแนกกลุ่มข้อมูลที่ไม่เคยเจอมาก่อนออกมาเป็นประเภท (class) โดยปกติแล้วจะทำด้วยวิธีการเรียนรู้แบบมีผู้สอน
- การวิเคราะห์การถดถอย (Regression) ใช้หลักการเดียวกับการแบ่งประเภทข้อมูล แต่ข้อมูลผลลัพธ์เป็นลักษณะต่อเนื่องมากกว่าเป็นประเภทแยกกัน
- การแบ่งกลุ่มข้อมูล (Clustering) การแบ่งกลุ่มข้อมูลเป็นกลุ่ม โดยกลุ่มของข้อมูลขึ้นอยู่กับค่าของข้อมูลเข้า และเครื่องไม่ทราบล่วงหน้าถึงกลุ่มของข้อมูล โดยปกติเป็นการเรียนรู้แบบไม่มีผู้สอน

- การประเมินความหนาแน่น (Density Estimation) เป็นการหาการกระจายของข้อมูลในมิติบางมิติ
- การลดขนาดของมิติ (Dimensionality Reduction) เป็นการเชื่อมโยงข้อมูลหลายมิติไปสู่ปรภูมิที่มิติต่ำกว่า

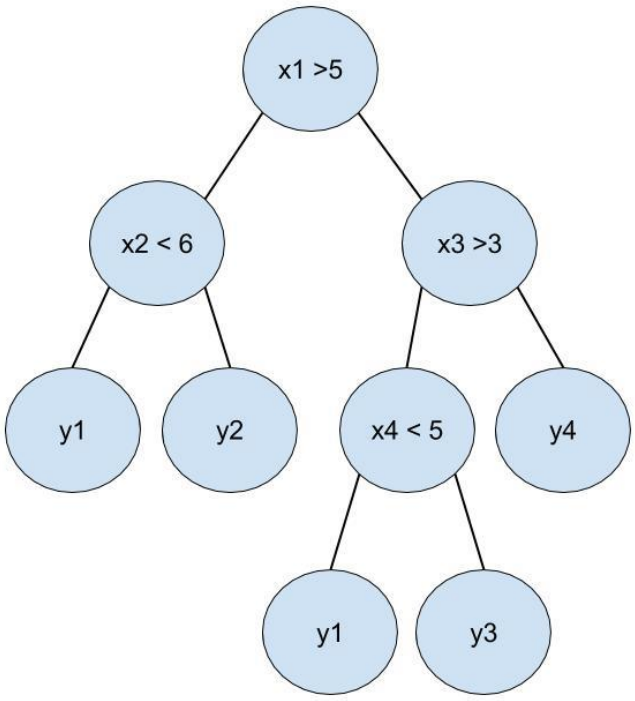
**2.4.1 วิธีการจำแนกด้วยต้นไม้ช่วยตัดสินใจ (Decision Tree Classifier)**

เป็นวิธีการจำแนกแบบแบ่งประเภทข้อมูล (Classification) และเรียนรู้แบบมีผู้สอน โดยอาศัยการวิธีการใช้ต้นไม้ช่วยตัดสินใจเข้ามาประยุกต์ใช้กับวิธีการเรียนรู้ของเครื่อง รูปที่ 2.5 แสดงตัวอย่างของต้นไม้ตัดสินใจในการตัดสินใจจำแนกประเภทของชุดข้อมูล เมื่อ X { x | x เป็นattributes ของข้อมูลนำเข้า} และ Y { y | y เป็นชุดของประเภทข้อมูลผลลัพธ์} ข้อดีของการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจคือสามารถทำการจำแนกหลายผลลัพธ์ได้

ตัวอย่างการจำแนกเมื่อกำหนดให้

X = { x1 = 5, x2 = 6, x3 = 1, x4 = 2, x5 =3}

Y = { y1 = “y1”, y2 = “y2”, y3 = “y3”, y4 = “y4”}




รูปที่ 2.5 ตัวอย่างต้นไม้ตัดสินใจ

จากตัวอย่างดังรูปที่ 2.5 จะได้ว่าชุดข้อมูล X เมื่อผ่านการจำแนกด้วยต้นไม้ตัดสินใจเป็นชุดข้อมูลประเภท “y2” หรือเขียนในรูปทั่วไปได้ว่า Tree(X1) = “y2” เมื่อกำหนดให้ต้นไม้ตัดสินใจรูปที่ 2.5 แทนค่าด้วย Tree() โครงสร้างของต้นไม้ตัดสินใจจะเป็นแบบต้นไม้ทวิภาค เพื่อให้สามารถทำการช่วยตัดสินใจได้ เนื่องจากรูปแบบการตัดสินใจใช้ผลลัพธ์ของนิพจน์ตรรกศาสตร์ให้ผลลัพธ์ในการเลือกสองแบบคือเป็น “จริง” และ “เท็จ” นอกจากนี้ยังได้มีการปรับใช้วิธีการอื่นกับวิธีการจำแนกด้วยต้นไม้ตัดสินใจ เช่นวิธีการ Ensemble [9] ที่ใช้วิธีการอาศัยหลายอัลกอริทึมการเรียนรู้เข้ามาช่วยในการพัฒนาประสิทธิภาพของการทำนายผล ด้วยวิธีการสร้างหลายโมเดลจำแนกและหาค่าเฉลี่ยของผลการทำนาย ได้แก่จำแนกแบบ Random Forests และ Extremely Randomized Trees เป็นต้น

**2.4.2 กระบวนการสร้างโครงสร้างต้นไม้จากข้อมูลชุดสอน**

กระบวนการสร้างโครงสร้างต้นไม้ตัดสินใจ [10] ด้วยการเรียนรู้ของเครื่องสามารถทำได้ด้วยการอาศัยชุดข้อมูลชุดสอนในการสร้างโครงสร้างต้นไม้ ตัวอย่างโครงสร้างของชุดข้อมูลแสดงในรูปแบบตารางที่ 2.1 ในรูปตารางสองมิติในชุดข้อมูล (Data Set) จะประกอบไปด้วยแถวที่เป็นข้อมูลหนึ่งชุดและข้อมูลหนึ่งชุดประกอบไปด้วยคุณลักษณะหรือแอตทริบิวต์ (Attribute) คือ x1 x2 x3 และ x4 สำหรับแอตทริบิวต์ Y เป็นข้อมูลของการแยกประเภทหรือแบ่งกลุ่มของชุดข้อมูลข้อมูลในระเบียนเดียวกัน เช่น ชุดข้อมูลที่ 1 ตามรางที่ 2.1 มีข้อมูลคุณลักษณะ {x1, x2, x3, x4} คือ {0, 2, 2, 3} และเก็บข้อมูลการจำแนกประเภทของกลุ่มข้อมูลเป็นประเภท “A” ใน แอตทริบิวต์ Y

ตารางที่ 2.1 ตัวอย่างชุดข้อมูลชุดสอน

| *Data set* | *Attribute* | | | | |
|---|---|---|---|---|---|
| | *x1* | *x2* | *x3* | *x4* | *Y* |
| *1* | 0 | 2 | 2 | 3 | A |
| *2* | 1 | 2 | 3 | 4 | A |
| *3* | 1 | 3 | 4 | 5 | B |
| *4* | 1 | 6 | 7 | 8 | B |

**การเลือกค่าแอตทริบิวต์ในการสร้างโครงสร้างต้นไม้**

การเลือกค่าแอตทริบิวต์ [11] ในการสร้างโหนดเพื่อแบ่งข้อมูลออกเป็นสองกลุ่ม ของโครงสร้างต้นไม้สามารถแบ่งได้ 2 วิธี คือ

1. การแบ่งกลุ่มข้อมูลด้วย Gini Index เป็นวิธีการในการแบ่งกลุ่มข้อมูล จาการหาคำนวณหาความถี่ในการระบุผลจำแนกของแอตทริบิวต์ ไม่ถูกต้องหากทำการสุ่มจำแนกประเภท ตามการกระจายการติดป้ายของแอตทริบิวต์นั้นๆ ความเหมาะสมในการเลือกแอตทริบิวต์สอดคล้องตามความถี่ในการระบุผลจำแนกผิดพลาดต่ำหรือเข้าใกล้ศูนย์ โดยเหมาะสำหรับการแบ่งข้อมูลที่มีขนาดใหญ่ และข้อมูลมีความซ้ำซ้อนกันเป็นปริมาณมาก การหาค่า Gini Index สามารถคำนวณได้จากสมการดังต่อไปนี้เมื่อกำหนดให้

$$Gini = 1 - \sum_{i=1}^{c}(p_i)^2$$
$$C = cardinality\ of\ class\ values$$
$$p_i = \frac{count(class\_value)}{count(class\_row)}$$

ตัวอย่าง Gini Index ของ Attibute $x_1 = 1$ ตามจากค่าในตารางที่ 2.1

$$x_1 = 1\ and\ y = 'A' \rightarrow p_i = \frac{1}{3}$$
$$x_1 = 1\ and\ y = 'B' \rightarrow p_{i+1} = \frac{2}{3}$$
$$Gini_{x1=1} = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) = 0.4444$$

กำหนดให้ $cost_l$แทนค่า $Gini_{x1=1}$หรือ 0.4444

$$x_1 \neq 1\ and\ y = 'A' \rightarrow p_i = \frac{1}{1}$$
$$x_1 \neq 1\ and\ y = 'B' \rightarrow p_{i+1} = \frac{0}{1}$$
$$Gini_{x1\neq1} = 1 - \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right) = 0.0$$

กำหนดให้ $cost_r$แทนค่า $Gini_{x1\neq1}$ หรือ 0.0

จะได้ค่า Gini Index ของข้อมูลสองกลุ่มคือ $x_1 == 1$ และ $x_1! = 1$ คือ 0.4444 ทั้งคู่ และ 0.0 ตามลำดับจากนั้นทำการหาค่าน้ำหนักของการแบ่งกลุ่ม (split point) ด้วยการคูณอัตราส่วนต่อกลุ่มและนำผลลัพธ์สองผลลัพธ์ที่ได้ โดยค่าที่ดีที่สุดคือ 0.0 และ แย่ที่สุดคือ 0.5

$$split\ point = cost_l * \frac{group\_size_l}{total_{samples}} + cost_r * \frac{group\_size_r}{total_{samples}}$$
$$\left(cost_l * \frac{3}{4}\right) + \left(cost_r * \frac{1}{4}\right) = \left(0.4444 * \frac{3}{4}\right) + \left(0.0 * \frac{1}{4}\right) = 0.3333$$
$$Best\ Classified = 0.0, Worst\ Classified = 0.5$$

2. การแบ่งกลุ่มข้อมูลด้วย Information Gain ตามพื้นฐานการคำนวณ Entropy ซึ่งอาศัยข้อมูลสำหรับการคำนวณความเป็นไปได้ต่อเหตุการณ์ที่จะเกิดขึ้นแต่ละเหตุการณ์ เพื่อหาค่าสำหรับการแบ่งกลุ่มข้อมูลจากเหตุการณ์ที่มีความกำกวมต่ำที่สุด เหมาะกับชุดข้อมูลที่มีความซ้ำซ้อนของข้อมูลน้อยแต่ความแตกต่างของข้อมูลสูง การหาค่า Information Gain สามารถคำนวณได้จากสมการดังต่อไปนี้

$$Entropy = \sum_{i=1}^{c} -p_i \times log_2(p_i)$$
$$C = cardinality\ of\ class\ values$$
$$p_i = \frac{count(class\_value)}{count(class\_row)}$$
$$Best\ Classified = 0.0, Worst\ Classified = 1.0$$
$$Information\ Gain = Entropy_l \times \frac{group\_size_l}{total_{samples}} + Entropy_r \times \frac{group\_size_r}{total_{samples}}$$

ตัวอย่างการหา Information Gain ของ x2 > 2

Entropy ของ x2 <= 2 คือ

$$x_2 <= 2 \text{ and y } = \text{A} \rightarrow p_i = \frac{2}{2}$$
$$x_2 <= 2 \text{ and y } = \text{B} \rightarrow p_{i+1} = \frac{0}{2}$$
$$Entropy_l = -1 \times \left(\frac{2}{2}\right) \times \log\left(\frac{2}{2}\right) + \left(\frac{0}{2}\right) \times \log\left(\frac{0}{2}\right) = 0$$

Entropy ของ x2 > 2 คือ

$$x_2 > 2 \text{ and y } = \text{A} \rightarrow p_i = \frac{0}{2}$$
$$x_2 > 2 \text{ and y } = \text{B} \rightarrow p_{i+1} = \frac{2}{2}$$
$$Entropy_r = -1 \times \left(\frac{2}{2}\right) \times \log\left(\frac{2}{2}\right) + \left(\frac{0}{2}\right) \times \log\left(\frac{0}{2}\right) = 0$$
$$Information\ Gain = 0 * \frac{2}{4} + 0 * \frac{2}{4} = 0$$

จากตัวอย่างการหาค่า Information Gain ของ x2 > 2 มีค่าเท่ากับ 0.0 หรือค่าดีที่สุดที่เป็นไปได้ในการในการแบ่งกลุ่มข้อมูลเพื่อใช้สร้างโครงสร้างต้นไม้ต่อไป

## 2.5 ข้อมูลอภิพันธุ์ (Metadata)

ข้อมูลอภิพันธ์ หรือ Metadata เป็นข้อมูลที่ใช้สำหรับอธิบายข้อมูลอื่น โดย NISO แบ่งข้อมูลอภิพันธุ์ [12] ออกเป็นสามประเภทคือ Structural Metadata, Descriptive Metadata และ Administrative Metadata

- **Structural Metadata** เป็นประเภทของข้อมูลอภิพันธุ์ที่ใช้ในการแสดงถึงรูปแบบของโครงสร้างของข้อมูล เช่น ภาษา XML ตัวอย่างภาษา XML จาก www.w3schools.com/xml

```
<note>
    <to>Tove</to>
    <from>Jani</from>
    <heading>Reminder</heading>
    <body>Don't forget me this weekend!</body>
</note>
```

จากตัวอย่างข้อมูลภาษา XML เป็นข้อมูลของข้อความโดยสามารถแจกแจงข้อมูลตามโครงสร้างได้ว่า ถึงบุคคลชื่อ “Tove” จาก “Jani” มีหัวเรื่องว่า “Reminder” และเนื้อหาคือ “Don't forget me this weekend!” จะเห็นได้ว่าข้อมูลมีโครงสร้างที่ชัดเจนทำให้ง่ายต่อการเข้าใจลักษณะและรูปแบบของข้อมูลส่งผลให้การง่ายต่อการนำไปใช้งานในลำดับถัดไป

- **Descriptive Metadata** ใช้ในการอธิบายและระบุข้อมูลสารสนเทศของทรัพยากร เช่น HTML Metadata ตัวอย่าง

  <meta name="description" content="Free Web tutorials">

  (http://www.w3schools.com/tags/tag_meta.asp)

  สามารถอธิบายตัวอย่างได้ว่า เป็นข้อมูลคำอธิบายที่มีคำสำคัญหรือ keyword คือ “description” มีคำอธิบายหรือเนื้อหาว่า “Free Web tutorials” ทำให้สามารถเข้าถึงข้อมูลได้อย่างสะดวกโดยวิธีการค้นหาด้วยคำสำคัญ

- **Administrative Metadata** เป็นข้อมูลที่อำนวยความสะดวกสำหรับการจัดการทั้งระยะสั้นและระยะยาวสำหรับการเก็บทรัพยากรในรูปแบบดิจิทัล เช่น สร้างอย่างไร และ เมื่อไร, ประเภทของไฟล์ หรือ ใครสามารถเข้าถึงได้บ้าง เป็นต้น

รูปแบบของข้อมูลข้อมูลอภิพันธ์ได้ถูกนำไปประยุกต์ใช้หลายด้าน เนื่องจากความสามารถในการแสดงลักษณะโครงสร้างและการอธิบายข้อมูล เช่น ในการประมวลผลภาษาธรรมชาติได้นำการติดป้าย (Tagging หรือ Annotation) มาทำการอธิบายหน้าที่หรือบทบาทของคำในโครงสร้างของไวยากรณ์ทางภาษา ที่เรียกว่า Part-of-Speech Tagging (POST) [13] เช่น ประโยค “A boy reading a book” เมื่อทำการ POST จะทำให้สามารถแจกแจงและอธิบายโครงสร้าง

ของประโยคได้ตามตารางที่ 2.2 ที่มีการจำแนกบทบาทของแต่ละคำในประโยคทำให้แสดงได้อย่างชัดเจนว่าคำใดมีบทบาทใดในประโยค และลดความกำกวมเช่นคำว่า “book” ในประโยค “A boy is reading a book” สามารถเป็นได้ทั้งคำนามที่แปลว่า “หนังสือ” หรือคำกริยาที่แปลว่า “จอง” หากไม่มีป้ายกำกับจะทำให้ไม่สามารถสรุปได้ว่า “book” ทำหน้าที่หรือหมายถึงความหมายใด

ตารางที่ 2.2 ตัวอย่าง Part-of-Speech Tagging

| *คำศัพท์ในประโยค* | *ป้าย* | *ความหมายของป้าย* |
|---|---|---|
| A | det (determiner) | คำนำหน้านาม |
| boy | noun | คำนาม |
| is | verb | คำกริยา |
| reading | verb | คำกริยา |
| a | det (determiner) | คำนำหน้านาม |
| book | noun | คำนาม |

## 2.6 คลังข้อความ (Corpus)

คลังข้อความคือเอกสารอิเล็กทรอนิกส์หรือไฟล์ที่บรรจุข้อความหรือคำในรูปแบบข้อมูลเชิงโครงสร้างโดยมีโครงสร้างแตกต่างกันไปตามบริบทของหัวข้อเรื่องของคลังข้อความนั้นๆ ในขณะเดียวกันข้อมูลอภิพันธ์มีส่วนเกี่ยวข้องต่อคลังข้อความอย่างเด่นชัดเนื่องจากสามารถแสดงเอกลักษณ์หรือคุณลักษณะของบริบทของคลังข้อความได้อย่างชัดเจน ซึ่งเป็นส่วนสำคัญในการนำไปประยุกต์ใช้กับกระบวนการอื่นที่เกี่ยวข้อง เช่น คลังข้อความภาษาไทย ORCHID ที่มีการติดป้าย part-of-speech สามารถนำมาใช้เป็นข้อมูลชุดสอนกระบวนการ POST ได้ โดยรูปแบบของโครงสร้างคลังข้อความ ORCHID [14] แสดงในตารางที่ 2.3 นอกเหนือจากคลังข้อความที่ติดป้าย part-of-speech แล้ว BEST [15] ยังเป็นโครงสร้างของคลังข้อความ BEST มีการแบ่งขอบเขตของคำด้วยสัญลักษณ์ "|" และมีการติดป้ายออกเป็นสามประเภทคือ

1. <EN> นามเอกลักษณ์ (Name Entity)
2. <AB> อักษรย่อ (Abbreviation)
3. <POEM> บทกลอน บทกวี (Poem)

โดยมีการแบ่งประเภทของข้อความออกเป็น 4 ประเภทคือ

1 บทความ (Article)

2. สารานุกรม (Encyclopedia)

3. ข่าว (News)

4. นิยาย (Novel)

ตัวอย่างข้อความประเภทบทความจากคลังข้อความ BEST

กฎหมาย|กับ|การ|เบียดบัง|คน|จน|

จาก|ต้นฉบับ|เรื่อง| |"|บท|นำ|:| |คน|จน|ภาย|ใต้|ความ|สัมพันธ์|ทาง|กฎหมาย|"|

<NE>ไพสิฐ พาณิชย์กุล</NE>|

เนื่องจากมีการแบ่งขอบเขตคำอย่างชัดเจนและมีการติดป้ายระบุ Name Entity จึงสามารถใช้เป็นข้อมูลชุดสอนในการประมวลผลการตัดคำหรือจดจำเอกลักษณ์ตามประเภทของข้อความได้

ตารางที่ 2.3 ตัวอย่างข้อมูลในคลังข้อความภาษาไทย Orchid

| ***เนื้อหา*** | ***โครงสร้าง*** | ***คำอธิบาย*** |
|---|---|---|
| #P1 | #Pตามด้วยเลขหน้า | เลขที่หน้า |
| #1 | #ตามด้วยเลขที่ข้อความ | เลขที่ข้อความ |
| การประชุมทางวิชาการ ครั้งที่ 1// | เนื้อหาของข้อความลงท้ายด้วย// | ข้อความ |
| การ/FIXN | หน่วยคำ/บทบาท | หน่วยคำและบทบาท |
| ประชุม/VACT | | |
| ทาง/NCMN | | |
| วิชาการ/NCMN | | |
| <space>/PUNC | | |
| ครั้ง/CFQC | | |
| ที่ 1/DONM | | |

## 2.7 การสกัดข้อมูลสารสนเทศ (Information Extraction :IE)

การสกัดนามเอกลักษณ์เป็นประเภทหนึ่งของงานด้านการสกัดข้อมูลสารสนเทศ (Information Extraction : IE) ในงานประเภทการสกัดเอกลักษณ์สามารถแบ่งออกเป็นกลุ่มย่อยได้ 3 กลุ่มคือ การจดจำนามเอกลักษณ์ การอ้างอิงเอกลักษณ์ และ การสกัดความสัมพันธ์ [16], [17]

- การจดจำนามเอกลักษณ์ (Name-Entity Recognition : NER) การจดจำ จำแนก หรือแยกแยะนามเอกลักษณ์ที่อยู่บนข้อความหรือเอกสารออกเป็น บุคคล สถานที่ วัน เวลา องค์กร ฯลฯ โดยสามารถจำแนกวิธีการออกได้เป็น 3 วิธี ดังนี้
  1. Rule-based NER ใช้กฎในการระบุนามเอกลักษณ์ (Name Entity : NE) รูปแบบของกฎสำหรับ NER
     - กฎสำหรับระบุเอกลักษณ์ (Rules to Identify a Single Entity)
       - กฎสำหรับการจับรูปแบบของเนื้อหาที่นำหน้านามเอกลักษณ์
       - กฎสำหรับการจับคู่องค์ประกอบในนามเอกลักษณ์
       - กฎสำหรับการจับรูปแบบของเนื้อหาที่ตามหลังหน้านามเอกลักษณ์
     - กฎสำหรับการกำหนดขอบเขตของนามเอกลักษณ์
     - กฎสำหรับหลายเอกลักษณ์
     - ตัวอย่างกฎในการจดจำเอกลักษณ์
       - กฎ : สถานที่ในท้องถิ่น
         หาความเกี่ยวข้อง : สถานีตำรวจ หาดใหญ่
         {สถานีตำรวจ : พจนานุกรมค้นหา = องค์กร | พจนานุกรมค้นหา = สถานที่ }
         {หาดใหญ่ : พจนานุกรมค้นหา = สถานที่ | พจนานุกรมค้นหา = อำเภอ}
         $\longrightarrow$ องค์กรส่วนอำเภอ
  2. Statistic-based NER อาศัยข้อมูลทางสถิติในการระบุ NE จำเป็นต้องอาศัยข้อมูลที่มีการแบ่งหน่วยของคำและติดป้ายจำแนกหรืออธิบายเพื่อให้สามารถนำข้อมูลไปใช้กับกระบวนการเรียนรู้ของเครื่อง หรือ Machine Learning ได้ เช่น โปรแกรมประยุกต์ Stanford Named Entity Recognizer ที่พัฒนาบนภาษา JAVA ได้อาศัยกระบวนการเรียนรู้ที่มีชื่อว่า Condition Random Field (CRF) สำหรับงานจดจำนามเอกลักษณ์

3. Dictionary-based NER ใช้พจนานุกรรมเฉพาะทางในการระบุ NE เป็นวิธีที่ง่ายและมีประสิทธิภาพหากมีรายการของ NE ที่ต้องการหรือเกี่ยวข้องทั้งหมด เหมาะกับการจดจำชื่อเฉพาะ

- การอ้างอิงเอกลักษณ์ (Coreference Resolution) การอ้างอิงคำสรรพนามหรือ NE อื่นที่เป็นเอกลักษณ์เดียวกันแต่รูปต่างกัน เช่น
  - “มานะชอบกินไข่ไก่ แต่***เขา***ไม่ชอบไข่เป็ด” เขาในที่นี้หมายถึง “มานะ”
  - “มหาวิทยาลัยสงขลานครินทร์” หรือ “ม.อ.” หมายถึงกลุ่มสถาบันเดียวกัน

- การสกัดความสัมพันธ์ (Relation Extraction) คือการหาความสัมพันธ์ระหว่าง EN เช่น บุคคล ทำงานให้ บริษัท หรือ บุคคล อาศัยอยู่ที่ ตำแหน่ง เป็นต้นโดยสามารถจำแนกวิธีการออกได้เป็นสามวิธี
  1. Feature-based เป็นวิธีการหาจุดเด่นของข้อมูลเข้าหรือข้อมูลสอน (training data) จากนั้นอาศัยการเรียนรู้ของเครื่องแบบมีผู้สอนเป็นการหาความสัมพันธ์ระหว่าง NE
  2. Kernel-based เป็นวิธีการในการออกแบบกระบวนการวิธีเฉพาะในการหาเปรียบเทียบความเหมือนกันทางโครงสร้าง ได้แก่ โครงสร้างต้นไม้และกราฟ
  3. Pattern or rule อาศัยรูปแบบหรือกฎในการหาความสัมพันธ์ระหว่าง NE โดยเลือกกฎที่มีลำดับสูงสุดที่สามารถสร้างโครงสร้างจาก NE ที่ต้องการหาความสัมพันธ์ได้
  4. Co-occurrence การหาความสัมพันธ์จากค่าสถิติที่เคยเกิดขึ้นในคลังชุดสอนที่ใช้เป็นตัวอย่าง เช่น ความสัมพันธ์ระหว่าง “นักเรียน” และ “โรงเรียน” เมื่ออยู่บนข้อความเดียวกัน และ มีข้อมูลทางสถิติของความสัมพันธ์ระหว่าง "นักเรียน” และ "โรงเรียน” ผ่านเกณฑ์ที่กำหนด ก็จะระบุว่า “นักเรียน” และ “โรงเรียน” เป็นเอกลักษณ์ที่มีความสัมพันธ์กัน

## 2.8 N-gram

N-gram [18] เป็นหนึ่งในวิธีของการประมวลผลภาษาที่หาความเป็นไปได้ของชุดลำดับอักขระหรือคำศัพท์ โดยอาศัยข้อมูลทางสถิติในการคำนวณความเป็นไปได้ของประโยค เมื่อ N คือขนาดของหน่วยย่อยของชุดอักขระที่ทำการหาความเป็นไปได้ และมีการระบุชื่อตามหน่วยของขนาดของชุดลำดับย่อย ได้แก่ 1 หน่วยย่อย เรียก Uni-gram 2 หน่วยย่อย เรียก Bi-gram และ 3

หน่วยย่อยเรียก Tri-gram โดยเมื่อจำนวนหน่วยย่อยมีมากกว่า 3 จะระบุตามขนาด N เช่น 4-gram หรือ 5-gram เป็นต้น สามารถหาค่าความเป็นไปได้ของประโยคจากสมการดังนี้
เมื่อกำหนดให้

| | |
|---|---|
| $P(w_1^n)$ | คือความเป็นไปได้ของประโยคขนาด n เมื่อ n คือจำนวนใดๆ และ $w$ คือหน่วยตัวอักษรหรือคำศัพท์ |
| $P(w_i \mid w_{i-N+1}^{i-1})$ | คือความเป็นไปได้ที่ $w_i$ ตามหลัง $w_{i-N+1}^{i-1}$ เมื่อ $1 \leq i \leq n$ และ $N$คือขนาดของ N-gram |
| $C(w_{i-N+1}^{i-1} \cdot w_i)$ | คือจำนวนนับของเหตุการณ์ที่ $w_i$ ตามหลัง $w_{i-N+1}^{i-1}$ |
| $C(w_{n-N+1}^{n-1})$ | คือจำนวนนับของ $w_{i-N+1}^{i-1}$ |

$$P(w_1^n) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-N+1}^{i-1}) \quad (1)$$

$$P\left(w_i \mid w_{i-N+1}^{i-1}\right) \approx \frac{C(w_{n-N+1}^{n-1} \cdot w_n)}{C(W_{n-N+1}^{n-1})} \quad (2)$$

$$P(w_1^n) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-1}) \quad (3)$$

$$P(w_1^n) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-2} w_{i-1}) \quad (4)$$

ตัวอย่างการหาค่าความเป็นไปได้ของประโยค “ฉันกินข้าวมันไก่” แบบ Bi-gram จากข้อมูลของตารางที่ 2.4 ได้ค่าความเป็นไปได้ของประโยคอยู่ที่ 0.00006 อย่างไรก็ตามค่าความเป็นไปได้ขึ้นอยู่กับแหล่งที่มาของข้อมูลสถิติ ซึ่งแตกต่างกันไปตามปริมาณและจุดประสงค์ของการเก็บข้อมูล

ตารางที่ 2.4 การหาค่าความเป็นไปได้จากข้อความ “ฉันกินข้าวมันไก่” แบบ Bi-gram

| จำนวนนับคำเดี่ยว | | จำนวนนับคำคู่ | |
|---|---|---|---|
| Ø (ค่าว่าง) | 10000 | Ø·ฉัน | 650 |
| ฉัน | 1000 | ฉัน·กิน | 15 |
| กิน | 30 | กิน·ข้าว | 20 |
| ข้าว | 60 | ข้าว·มัน | 10 |
| มัน | 17 | มัน·ไก่ | 10 |
| ไก่ | 55 | | |

$$P(\text{"ฉันกินข้าวมันไก่"}) = P(\text{ฉัน}|\text{ø}) \text{ x } P(\text{กิน}|\text{ฉัน}) \text{ x } P(\text{ข้าว}|\text{กิน}) \text{ x } P(\text{มัน}|\text{ข้าว}) \text{ x } P(\text{ไก่}|\text{มัน})$$
$$= \frac{650}{10000} \times \frac{15}{1000} \times \frac{20}{30} \times \frac{10}{60} \times \frac{10}{17}$$
$$= 0.00006$$

## 2.9 รูปแบบและโครงสร้างภาษาไทย

การวิเคราะห์ประโยคธรรมชาติของภาษาไทย จำเป็นต้องเข้าใจคุณลักษณะของภาษาไทยก่อน ในงานวิจัย [19] ได้อธิบายถึงคุณลักษณะที่สำคัญของภาษาไทยไว้ว่า เป็นภาษาที่มีรูปแบบโครงสร้างในลักษณะของ ประธาน-กริยา-กรรม หรือ Subject-Verb-Object (SVO) ซึ่งมีลักษณะทั่วไปเช่นเดียวกับภาษาอังกฤษที่เป็นภาษาหลักหรือภาษาธรรมชาติที่ใช้ภาษาหลัก แต่ในขณะเดียวกันข้อความภาษาไทยแตกแต่งจากภาษาอังกฤษคือภาษาไทยเป็นภาษาที่ไม่มีสัญลักษณ์ที่ใช้กำหนดขอบเขตคำหรือขอบเขตประโยค นอกจากนั้นในส่วนโครงสร้างประโยคของภาษาไทยสามารถมีส่วนขยายในหน่วยต่างๆของประโยค และประโยคอาจเป็นประโยคความรวมที่ประกอบขึ้นจากประโยคมากกว่าหนึ่งประโยค ดังนั้นจึงทำให้ประโยคความรวมมีความซับซ้อนมากขึ้น โดยในองค์ประกอบของประโยคสามารถประกอบด้วยคำศัพท์ที่เป็นรากศัพท์เดี่ยว หรืออาจเป็นวลีที่เกิดจากการประกอบกันของคำตั้งแต่หนึ่งคำขึ้นไปแต่ไม่เป็นประโยคที่สมบูรณ์ ในงานวิจัย [20] และ [21] ได้อธิบายถึงการแบ่งประเภทวลีตามคำนำหน้าของวลี แต่หลักการดังกล่าวสามารถยกเว้นได้ในกรณีของอุทานวลี ซึ่งเป็นวลีที่มีคำอุทานเป็นองค์ประกอบแต่คำอุทานไม่จำเป็นต้องนำหน้าเสมอไป โดยสามารถแบ่งประเภทวลีได้ต่อไปนี้

| | | | |
|---|---|---|---|
| ●นามวลี | นำหน้าด้วยคำนาม<br>เช่น *ข้าว*สุก *เก้าอี้*หัก | ●วิเศษณ์วลี | นำหน้าด้วยคำวิเศษณ์<br>เช่น *อร่อย*มาก *เค็ม*เกลือ |
| ●สรรพนามวลี | นำหน้าด้วยคำสรรพนามวลี<br>เช่น *เขา*วิ่งช้า *เธอ*นั่งอยู่ | ● บุพบทวลี | นำหน้าด้วยคำบุพบท<br>เช่น ที่สูง บนหลังคา |
| ●กริยาวลี | นำหน้าด้วยคำกริยา<br>เช่น วิ่งเร็ว *นั่ง*เก้าอี้ | ● สันธานวลี | นำหน้าด้วยคำสันธาน<br>เช่น *เมื่อ*วาน *ตั้งแต่*ตอนนี้ |
| | | ● อุทานวลี | นำหน้าด้วยคำอุทาน<br>เช่น *เฮ้อ!* รอดแล้ว |

นอกจากวิธีการระบุประเภทของวลีตามชนิดของคำนำหน้าแล้ว ในเอกสาร [22] ได้มีการให้รายละเอียดเพิ่มเติมของ โครงสร้าง นามวลี และ กริยาวลี และทำการรวมหัวข้อของสรรพนามให้อยู่หัวเดียวกับนามวลีและได้แจกโครงสร้างของนามวลีและกริยาวลีไว้ดังนี้

**1. โครงสร้างของนามวลี**

นามวลีและสรรพนามวลี สามารถเกิดขึ้นได้จาก คำนามและคำขยายนามโดยคำขยายนามจะมีหรือไม่ก็ได้ ซึ่งบางครั้ง วลี และ ประโยค ก็สามารถเป็นหน่วยของคำนามของประโยคอื่นได้ โดยองค์ประกอบของนามวลีมี 5 ชนิดคือ

1. หน่วยหลักคำนามและสรรพนาม (ล)
2. หน่วยคำคุณศัพท์ (ค)
3. หน่วยจำนวน (จ)
4. หน่วยกำหนด (ก)
5. หน่วยคำขยายเสริม (ขส)

ตารางที่ 2.5 โครงสร้างของนามวลี

| *ลำดับ* | *โครงสร้าง* | *ตัวอย่าง* |
|---|---|---|
| 1 | ล | สุนัข, แมว, ฉัน, เก้าอี้ |
| 2 | ล + ค | รถ + เก่า |
| 3 | ล + จ | ดินสอ + หนึ่งด้าม |
| 4 | ล + ก | หนังสือ + นั่น |
| 5 | ล + ค + จ | รถ + ใหม่ + สามคัน |
| 6 | ล + จ + ค | สมุด + สามเล่ม + ใหม่ |
| 7 | ล + ค + ก | สมุด + เก่า + นั่น |
| 8 | ล + จ + ก | ปากกา + สองด้าม + นั้น |
| 9 | ล + ก + จ | ไข่ไก่ + นี้ + สองใบ |
| 10 | ล + ค + จ + ก | รถ + ใหม่ + สามคัน + นั้น |
| 11 | ล + ค + ก + จ | ปากกา + สีแดง + นี้ + สามด้าม |
| 12 | ล + จ + ค + ก | เนื้อหมู + สามชิ้น + ใหญ่ + นี้ |
| 13 | ล + ขส | หนังสือ + ส่วนมาก |
| 14 | ล + ค + ขส | รถ + ใหญ่ + ส่วนน้อย |
| 15 | ล + ขส + จ | ของ + ส่วนน้อย + สามชิ้น |
| 16 | ล + ค + จ + ขส | แมว + สีขาว + สองตัว + อยู่ใต้โต๊ะ |
| 17 | ล + จ + ค + ขส | แมว + สองตัว + สีเหลือง + อยู่ใต้โต๊ะ |

**2. โครงสร้างของกริยาวลี**

กริยาวลี สามารถประกอบขึ้นจาก คำกริยาและองค์ประกอบของกริยาวลี โดยองค์ประกอบของกริยาวลีมี 4 ชนิดคือ

1. หน่วยกริยา (ก)
2. หน่วยคำช่วยกริยาหน้า (ช1)
3. หน่วยคำช่วยกริยาหลัง (ช2)
4. หน่วยขยาย (ข)

ตารางที่ 2.6 โครงสร้างของกริยาวลี

| *ลำดับ* | *โครงสร้าง* | *ตัวอย่าง* |
|---|---|---|
| 1 | ก | วิ่ง, เดิน, กระโดด, ว่ายน้ำ |
| 2 | ก + ข2 | นั่ง + อยู่ |
| 3 | ก + ข | ชอบ + จัง |
| 4 | ก + ข2 + ข | หัก + อยู่ + ก่อน |
| 5 | ก + ข + ข2 | ปรึกษา + กัน + อยู่ |
| 6 | ข1 + ก | คง + สวย |
| 7 | ข1 + ก + ข2 | กำลัง + นั่ง + อยู่ |
| 8 | ข1 + ก + ข | คง + มา + บ่อย |
| 9 | ข1 + ก + ข2 + ข | น่าจะ + หัก + อยู่ + ก่อน |
| 10 | ข1 + ก + ข + ข2 | ยัง + วิ่ง + กัน + นั้น |

จากคุณลักษณะของภาษาไทยข้างต้น จะส่งผลให้เกิดปัญหาในการวิเคราะห์ข้อความที่มีโครงสร้างประโยคเป็นประโยคบอกเล่าออกเป็น 3 รูปแบบ คือ คำพ้องความหมาย คำขยาย และ ส่วนขยายที่ใช้คำหลักร่วม

**1. คำพ้องความหมาย** คำพ้องความหมายสร้างความกำกวมในการประมวลผลภาษาเนื่องจาก รูปแบบโครงสร้างประโยคหรือวลีสามารถประกอบได้มากกว่าหนึ่งรูปแบบแต่มีความหมายเดียวกันเช่น “เขารักสุนัข” หรือ “เขารักหมา” ต่างมีความหมายเดียวกัน เช่นเดียวกับข้อความอาการและอาการแสดงที่ข้อความส่วนใหญ่อยู่ในรูปวลี เช่น “ปวดหัว” หรือ “ปวดศีรษะ” ทั้งสองวลีมีความหมายเหมือนกันแต่มีการประกอบต่างกัน

**2. คำขยาย** นอกจากคำพ้องความหมายแล้วคำขยายก็เป็นหนึ่งในองค์ประกอบที่ทำให้เกิดความกำกวมในการประมวลผลภาษา โดยคำขยายทำให้ประโยคหรือวลีมีความสมบูรณ์มากขึ้นเช่น “เจ็บหน้าอกตอนหายใจ” มี “ตอน” เป็นส่วนขยายและเมื่อตัดออกไปจะกลายเป็น “เจ็บหน้าอกหายใจ” จะทำให้ไม่ได้ใจความที่สมบูรณ์ นอกจากนี้คำขยายเองก็ได้รับผลกระทบจากคำพ้องความหมายเช่นกันเช่นข้อความว่า “เจ็บหน้าอกขณะหายใจ” ก็มีความหมายเช่นเดียวกันกับ “เจ็บหน้าอกเวลาหายใจ” นอกจากคำขยายที่มีความจำเป็นในการเติมความสมบูรณ์ของข้อความแล้วคำขยายสามารถใส่เพื่อความสละสลวยของข้อความโดยจะมีหรือไม่มีก็ได้เช่น “เจ็บหน้าอก” และ “เจ็บบริเวณหน้าอก” มีความหมายเดียวกันโดยมีคำว่า “บริเวณ” เป็นคำขยายที่จะมีหรือไม่มีก็สามารถสื่อใจความครบถ้วนได้เช่นเดียวกัน

**3.** ส่วนขยายที่ใช้คำหลักร่วม ปัญหาถัดไปที่เกิดขึ้นจากคำขยายคือการทำให้ในหนึ่งวลีสามารถสื่อได้มากกว่าหนึ่งอาการ เช่น “มีอาการปวดหัวและท้อง” จะเห็นได้ว่าวลี “ปวดหัวและท้อง” สามารถสื่อได้สองอาการคือ “ปวดหัว” และ “ปวดท้อง” โดยอาศัยกริยาเดียวกันในการประกอบวลี ทำให้การระบุอาการและอาการแสดงมีความซับซ้อนมากขึ้น

## 2.10 การตัดคำภาษาไทย (Thai Word Segmentation)

ปัญหาของการวิเคราะห์ภาษาไทยที่มาจากการไม่มีตัวอักษรหรือสัญลักษณ์แบ่งแยกคำแต่ละคำภายในข้อความหรือประโยค ซึ่งการแก้ปัญหานี้ได้ถูกนำเสนอในงานวิจัย [23-26] โดยได้เสนอให้รูปแบบการตัดคำภาษาไทยโดยใช้รูปแบบของการตัดคำออกเป็น 3 วิธีการคือ การตัดคำโดยการใช้กฎ (Rule-Based Tokenization : RBT) การตัดคำโดยการใช้พจนานุกรม (Dictionary-Based Tokenization : DBT) และการตัดคำโดยการใช้การเรียนรู้ของเครื่อง (Machine Learning-Based Tokenization : MBT)

งานวิจัย [23] ได้เสนอการตัดพยางค์เพื่อหาขอบเขตหน้าและขอบเขตหลังของประโยคโดยอาศัยกฎที่สร้างขึ้นตามคุณลักษณะของอักขระภาษาไทยที่แบ่งกลุ่มอักขระเป็น 5 กลุ่มดังต่อไปนี้

1. กลุ่มอักขระที่ไม่ทำเกิดการขยับตำแหน่ง เช่น ◌ั ◌ิ ◌ี ◌ึ ◌ื
2. กลุ่มอักขระที่นาหน้าเสมอ เช่น เ แ โ ใ
3. กลุ่มอักขระที่อยู่หลังเสมอ เช่น ะ า ◌ั า

4. กลุ่มอักขระที่มีตัวการันต์อยู่ข้างบน เช่น ์
5. กลุ่มอักขระอื่นๆที่ไม่สามารถระบุได้ในกลุ่มที่ 1-4

งานวิจัย [24], [25] ได้นำเสนอวิธีการตัดคำด้วยพจนานุกรม โดยงานวิจัย [24] ได้เสนอวิธีเรียกว่า Longest Matching (LM) เป็นวิธีเลือกตัดคำจากคำที่ยาวที่สุดที่พบในพจนานุกรม ซึ่งหากคำที่ยาวที่สุดที่ถูกเลือก เป็นคำที่ไม่มีอยู่ในพจนานุกรม ระบบก็จะทำการเลือกคำใหม่ ซึ่งจะเป็นคำที่มีความยาวที่สุดรองลงมา อย่างไรก็ตามวิธีการ LM ยังคงมีจุดบกพร่องจากการเลือกคำที่ยาวที่เป็นอันดับแรกทำให้เกิดความผิดพลาดในการตัดคำได้ โดยในงานวิจัย [25] ได้เสนอวิธีชื่อ Maximum Matching (MM) โดยวิธีนี้จะใช้เทคนิค LM ก่อนจากนั้นจะกลับไปทำ LM ซ้ำอีกครั้งในแต่ละคำที่ถูกตัด เพื่อหารูปแบบที่สามารถเป็นได้ทั้งหมดก่อน จากนั้นจึงเลือกรูปแบบที่มีจำนวนคำน้อยที่สุดเป็นรูปแบบผลลัพธ์ของการตัดคำ ข้อดีของการตัดคำโดยใช้พจนานุกรมคือการประมวลผลที่รวดเร็วและง่าย และสามารถเพิ่มประสิทธิภาพของวิธีการดังกล่าวได้ด้วยการเพิ่มคำศัพท์ใหม่ที่ไม่ปรากฏในพจนานุกรมหรือคำศัพท์เฉพาะได้

ในงานวิจัย [26] ได้มีการเปรียบเทียบวิธีการตัดคำภาษาไทยแบบ ใช้พจนานุกรม และ ใช้การเรียนรู้ของเครื่อง การตัดคำแบบใช้พจนานุกรม ประกอบไปด้วย LM และ MM ในขณะที่การตัดคำด้วยวิธีการเรียนรู้ของเครื่อง ประกอบไปด้วย Naive Bayes (NB), Decision Tree, Support Vector Machine (SVM) และ Conditional Random Field (CRF) ในการสอนได้ทำการแบ่งตัวอักษรออกเป็นกลุ่ม (type) และ ประเภท (class) ตามที่แสดงตัวอย่างในที่ 2.11

ตัวอย่างกลุ่มของตัวอักษร

| | |
|---|---|
| กลุ่ม c ตัวอักษรที่สามารถเป็นพยัญชนะตัวสุดท้าย | เช่น ก ข ค ฆ ง จ |
| กลุ่ม n ตัวอักษรที่ไม่สามารถเป็นพยัญชนะตัวสุดท้าย | เช่น ต ฉ ผ ฝ ห ฮ |
| กลุ่ม v สระที่ไม่สามารถขึ้นต้นคำได้ | เช่น ะ า ั ิ ี |
| กลุ่ม w สระที่สามารถขึ้นต้นคำได้ | เช่น เ แ โ ใ ไ |
| กลุ่ม t ตัวอักษรวรรณยุกต์ | เช่น ่ ้ ๊ ๋ |

จำแนกตัวอักขระออกเป็นสองประเภทคือ

1) ตัวอักษรที่ขึ้นต้นคำ (Word-beginning Character : B)

2) ตัวอักษรที่อยู่ภายในคำ (Word-Inning Character : I)

ตารางที่ 2.7 แสดงตัวอย่างการติดป้ายสำหรับกลุ่มคำ “ท่านผู้ฟังคะ” ที่แยกประเภทตามตัวอักษร

| ***Character*** | ***Type*** | ***Class*** |
|---|---|---|
| ท | C | B |
| ่ | T | I |
| า | V | I |
| น | C | I |
| ผ | N | B |
| ู | V | I |
| ้ | T | I |
| ฟ | C | I |
| ั | V | I |
| ง | C | I |
| ค | C | B |
| ะ | C | I |

สำหรับการทดสอบการตัดคำโดยอาศัยพจนานุกรมด้วยวิธี LM และ MM นั้น จะใช้พจนานุกรม 2 ประเภทคือ พจนานุกรมสำหรับคำทั่วไปชื่อ Lexitron และ พจนานุกรมตามบริบทของข้อความที่ต้องการตัดกำหนดให้ชื่อว่า Domain

การทดสอบการตัดคำแบบใช้การเรียนรู้ของเครื่องใช้คลังข้อความ ORCHID และใช้ n-gram ในการหาขอบเขตของคำจากกลุ่มของตัวอักษรแบ่งออกเป็น 3, 5, 7, 9 และ 11 gram ในการประเมินผลของศักยภาพของทุกแบบทำการประเมินด้วยวิธี 10-fold Cross Validation และวัดผลความแม่นยำในการตัดคำด้วยคะแนน F1 ด้วยโดยผลของการทดลองแสดงของการเปรียบเทียบระหว่างการตัดคำแบบ DBT และ MBT จาก พบว่า การตัดคำด้วยการเรียนรู้แบบ CRF ได้คะแนน F1 ที่ใช้สำหรับประเมินความแม่นยำสูงที่สุดในการตัดคำ

ตารางที่ 2.8 การเปรียบเทียบประสิทธิภาพของการตัดคำแต่ละรูปแบบ

| *Approach* | *Algorithm* | *Precision* | *Recall* | *F1-score* |
|---|---|---|---|---|
| DBT | LM-Lexitron | 88.21 | 86.91 | 87.56 |
| | LM-Domain | 95.20 | 88.25 | 91.75 |
| | MM-Lexitron | 88.34 | 87.39 | 87.86 |
| | LM-Domain | 95.27 | 88.92 | 91.98 |
| MBT | NB | 69.70 | 60.60 | 64.90 |
| | DecisionTree-J48 | 80.81 | 75.10 | 77.50 |
| | SVM | 92.87 | 88.71 | 90.74 |
| | CRF | 95.79 | 94.98 | 95.38 |

## 2.11 ไวยากรณ์ไม่พึ่งบริบท (Context-Free Grammars : CFG)

ไวยากรณ์ไม่พึ่งบริบท หรือ CFG [27], [28] เป็นไวยากรณ์ของโครงสร้างภาษาที่มีพื้นฐานมาจากเซตของกฎที่ให้ผลลัพธ์เป็นเซตของอักขระที่เป็นไปได้ทั้งหมดของภาษานั้นๆ โดยสามารถเขียนไวยากรณ์ทางภาษาให้อยู่ในรูปมาตรฐาน CFG ได้ดังนี้

$$G = (V, \Sigma, R, S)$$

เมื่อ $G$ คือ ไวยากรณ์ของภาษาที่ถูกกำหนดด้วยมาตรฐานของ CFG

$V$ คือ เซตของ NonTerminal หรือตัวแปรที่แสดงถึงประเภทของวลีหรือประโยคแต่ไม่ใช่เนื้อหาจริงๆของวลีหรือประโยค

$\Sigma$ คือ เซตของ Terminal หรือชุดของอักขระในภาษาที่เป็นเนื้อหาของประโยค

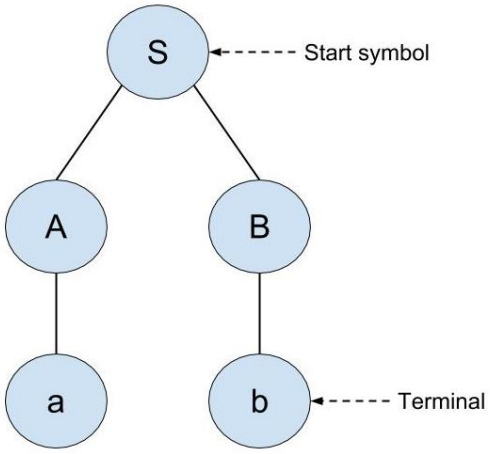
$R$ คือ กฎ (Rule) หรือ ผล (Product) ของ NonTerminal ที่แสดงถึงโครงสร้างของภาษา

$S$ คือ ตัวแปรเริ่มต้นของประโยค ที่ใช้แสดงถึงความสอดคล้องระหว่างไวยากรณ์และชุดของอักขระ เมื่อสามารถสร้างโครงสร้างต้นไม้จากกฎของ CFG ที่มีโหนดรากเป็น S และทุกอักขระเป็นโหนดใบ

**2.11.1 การสร้างโครงสร้างต้นไม้จาก CFG**

การแสดงให้เห็นว่าประโยคหรือชุดอักขระดังกล่าวอยู่ในกฎของไวยากรณ์ของภาษาตามมาตรฐานของ CFG หรือไม่ สามารถอาศัยวิธีการสร้างโครงสร้างต้นไม้ [29] ในการพิสูจน์ความสมบูรณ์ของประโยค หากข้อความดังกล่าวสามารถสร้างโครงสร้างต้นไม้ที่มีโหนดรากเป็น S ทุกโหนดพ่อเป็น NonTerminal และมีโหนดใบเป็นอักขระในข้อความที่อยู่ในชุดของ Terminal ในไวยากรณ์ CFG ให้ถือว่าเป็นข้อความสมบูรณ์ตามไวยากรณ์ CFG ตามรูปตัวอย่างที่ 2.6 เมื่อกำหนดให้

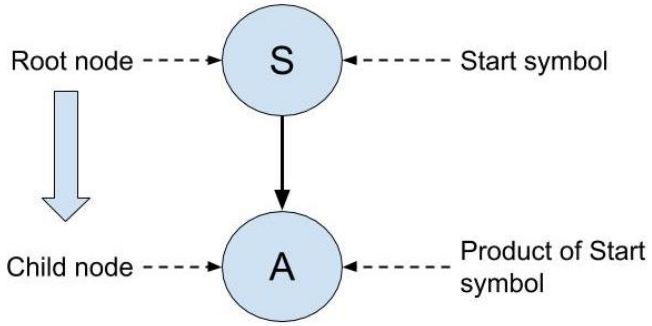
$$
\begin{aligned}
G &= (V, \Sigma, R, S) \\
V &= A, B \\
\Sigma &= a, b \\
R &= S \rightarrow A \\
&\quad\ S \rightarrow B \\
&\quad\ A \rightarrow a, aA, aB \\
&\quad\ B \rightarrow b, Bb, Ab
\end{aligned}
$$




รูปที่ 2.6 ตัวอย่างโครงสร้างต้นไม้ของ $ab$ จากไวยากรณ์ CFG

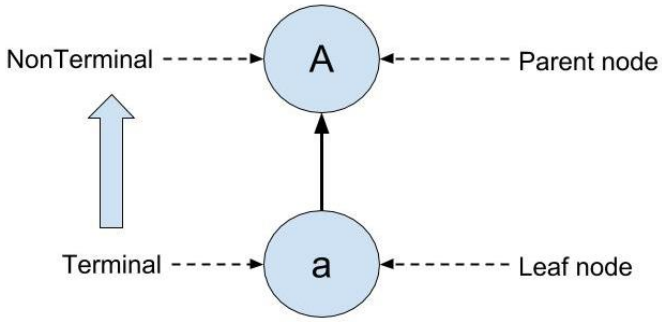
วิธีการเริ่มต้นสร้างโครงสร้างต้นไม้สามารถแบ่งประเภทออกได้เป็นสองประเภทคือ

- เริ่มสร้างต้นไม้จากบนลงล่าง Top-down — เริ่มสร้างต้นไม้จาก Root node ถูกกำหนดให้เป็น S ที่เป็นจุดเริ่มต้นของประโยคและอาศัยกฎในการแจกแจงโหนดลูกลงไปจบที่โหนดใบ หรือ Terminal ของไวยากรณ์แต่ละภาษา




รูปที่ 2.7 การเริ่มสร้างต้นไม้จากบนลงล่าง

- เริ่มสร้างต้นไม้จากล่างขึ้นบน Bottom-up — เริ่มสร้างต้นไม้จากโหนดใบที่เป็น Terminal และจับคู่กับ NonTerminal ที่สามารถสร้างชุดของ Terminal ตามกฎของไวยากรณ์ที่กำหนดและกำหนดให้ NonTerminal ดังกล่าวเป็นโหนดพ่อของชุด Terminal



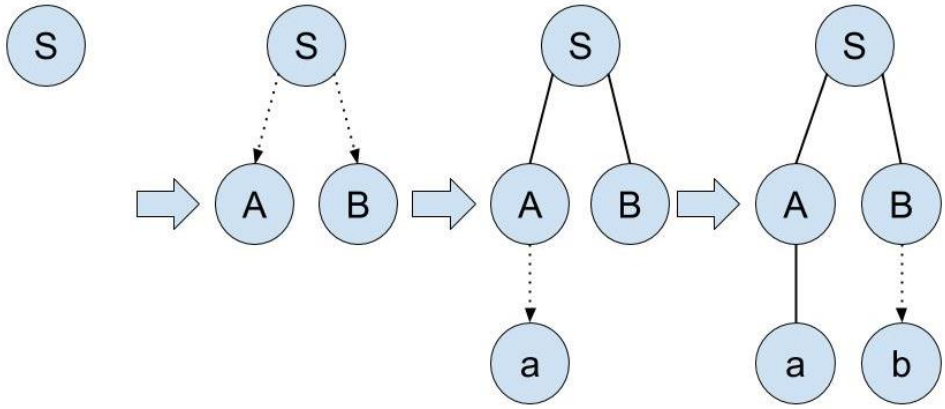

รูปที่ 2.8 การเริ่มสร้างต้นไม้จากล่างขึ้นบน

การขยาย node หลังจากการเริ่มต้นสร้างโครงสร้างต้นไม้ก็เป็นปัจจัยที่ต้องมีการพิจารณาเนื่องจากถ้ามี NonTerminal ในระดับความลึกเดียวกันมากกว่าหนึ่งโหนดจำเป็นต้องมีวิธีการเลือกว่าจะเริ่มทำจากโหนดใดก่อน โดยสามารถแบ่งออกได้เป็นสองวิธี คือ

- เริ่มทำจากซ้ายสุด Leftmost derivation
- เริ่มทำจากขวาสุด Rightmost derivation

การเริ่มสร้างโครงสร้างต้นไม้ด้วยวิธีจากบนลงล่างและเริ่มจากซ้ายสุด จะเริ่มสร้างจาก root node และทำการสร้างจากโหนดลูกทางซ้าย จนกว่าผลลัพธ์ทางซ้ายสุดจะเป็น Terminal จึงกลับไปสร้างโหนดลูกต่อจาก NonTerminal โหนดทางซ้ายสุด ที่สามารถสร้างโหนดลูกได้ และทำเช่นเดิมซ้ำจนกว่าจะได้ Terminal ที่เป็นองค์ประกอบของประโยคเป็นโหนดใบทั้งหมด สำหรับการสร้างต้นไม้จากล่างขึ้นบน จะเริ่มสร้างโหนดใบจากทางซ้ายสุดที่สามารถ ตามกฎของ CFG ได้ หลังจากนั้นจะทำการสร้างโหนดพ่อโดยพิจารณาจากตำแหน่งของโหนดลูกที่อยู่

ตำแหน่งซ้ายที่สุดจาก node ทั้งหมดที่สามารถสร้างโหนดพ่อได้ จนกว่าโหนดใบที่เป็น Terminal และ องค์ประกอบของประโยคจะถูกใช้ในสร้างโครงสร้างต้นไม้ทั้งหมด และมีโหนดที่อยู่ชั้นบนสุดเป็นโหนดราก สำหรับการเริ่มสร้างโครงสร้างต้นไม้จากทางขวา ทำเช่นเดียวกับทางซ้าย เพียงแต่เปลี่ยนทิศทางการสร้างโครงสร้างต้นไม้เป็นทางขวา




รูปที่ 2.9 ตัวอย่างการสร้างต้นไม้แบบบนลงล่างและขยายโหนดจากทางซ้ายสุด

## 2.12 Stop words

Stop words เป็นคำที่ถูกพิจารณาว่ามีนัยสำคัญต่ำ หรือ ไม่มีความสำคัญต่อการนำไปประมวลผลในกระบวนการทางคอมพิวเตอร์ เช่น การค้นหาข้อมูลด้วยคำสำคัญ หรือการประมวลภาษาธรรมชาติ เป็นต้น โดย [31] ได้พัฒนาเครื่องมือสำหรับประมวลผลความหนาแน่นคำสำคัญ สำหรับเพิ่มประสิทธิภาพของเครื่องมือค้นหา หรือ กระบวนการอื่นที่เกี่ยวข้องกับการใช้คำสำคัญ โดยเก็บชุดข้อมูล Stop words มากกว่า 40 ภาษา รวมถึง Stop words ของภาษาไทย เช่น

| | | | |
|---|---|---|---|
| ไว้ | เฉพาะ | ยัง | ต่อ |
| ไม่ | เคย | มี | ตาม |
| ไป | เข้า | มาก | ตั้งแต่ |
| ได้ | เขา | มา | ตั้ง |
| ให้ | อีก | พร้อม | ด้าน |
| ใน | อาจ | พบ | ด้วย |
| โดย | อะไร | ผ่าน | ดัง |

| | | | |
|---|---|---|---|
| แห่ง | ออก | ผล | ซึ่ง |
| แล้ว | อย่าง | บาง | ช่วง |
| และ | อยู่ | น่า | จึง |
| แรก | อยาก | นี้ | จาก |
| แบบ | หาก | นำ | จัด |
| แต่ | หลาย | นั้น | จะ |
| เอง | หลังจาก | นัก | คือ |
| เห็น | หลัง | นอกจาก | ความ |
| เลย | หรือ | ทุก | ครั้ง |
| เริ่ม | หนึ่ง | ที่สุด | คง |
| เรา | ส่วน | ที่ | ขึ้น |
| เมื่อ | ส่ง | ทำให้ | ของ |
| เพื่อ | สุด | ทำ | ขอ |
| เพราะ | สำหรับ | ทาง | ขณะ |
| เป็นการ | ว่า | ทั้งนี้ | ก่อน |
| เป็น | วัน | ั้ง | ก็ |
| เปิดเผย | ลง | ถ้า | การ |
| เปิด | ร่วม | ถูก | กับ |
| เนื่องจาก | ราย | ถึง | กัน |
| เดียวกัน | รับ | ต้อง | กว่า |
| เดียว | ระหว่าง | ต่างๆ | กล่าว |
| เช่น | รวม | ต่าง | |

## 2.13 ประเมินผลการจำแนกข้อมูล

การประเมินผลการจำแนกจากกระบวนการทางสถิติ สามารถประเมินได้ด้วย F1 score (F-score หรือ F-measure) [31] คือการประเมินความแม่นยำของผลการจำแนกด้วยการอาศัยค่า precision และ recall โดยสามารถแจกแจงประเภทของผลการจำแนกและค่าการประเมินผลได้ดังนี้

ประเภทของผลการจำแนก

true positive (*tp*) = ผลการจำแนกประเภทที่ต้องการและถูกจำแนก

false negative (*fn*) = ผลการจำแนกประเภทที่ไม่ต้องการและถูกจำแนก

true negative (*tn*) = ผลการจำแนกประเภทที่ไม่ต้องการและไม่ถูกจำแนก

false positive (*fp*) = ผลการจำแนกประเภทที่ต้องการและไม่ถูกจำแนก

ประเภทของค่าการประเมินผลและวิธีการคำนวณ

Precision $= \frac{tp}{tp+fp}$ คือ ค่าที่สะท้อนปริมาณผลการจำแนกที่เกี่ยวข้อง

Recall $= \frac{tp}{tp+fn}$ คือ ค่าสะท้อนความถูกต้องของผลการจำแนก

F-1 score $= 2 \cdot \frac{precision \cdot recall}{precision+recall}$ คือ ค่าเฉลี่ยของน้ำหนักระหว่าง Precision และ Recall

## 2.14 งานวิจัยที่เกี่ยวข้อง

ในการจำแนกหรือการสกัดข้อมูลความเหมาะสมของข้อมูลตัวอย่างและข้อมูลทดลองก็มีผลต่อการจำแนกหรือการทำนายผลลัพธ์ เช่น ในงานวิจัย [32] ได้ทำการเปรียบเทียบประสิทธิภาพของวิธีการเตรียมข้อความบอกเล่าอาการสำคัญ 2 วิธี ระหว่างกระบวนการ CCP และกระบวนการ EMP-P เพื่อนำผลที่ได้ไปใช้ต่อในกระบวนการจำแนกข้อความ CC และประเมินผลการเตรียมข้อมูลด้วย Statistic-based และ Keyword-based Classifier นอกจากนี้ในการวิจัย [33] ได้เสนอการเตรียม CC ด้วย กระบวนการ EMP-P และเปรียบเทียบความเหมือนทางความหมายของ CC กับ ตารางกลุ่มอาการและ UMLS Ontology (คือมาตรฐานของโปรแกรมทางการแพทย์) ด้วยวิธีการใช้กฎก่อนจะนำ CC ที่ผ่านการเตรียมแล้วไปจำแนกด้วยโปรแกรมประยุกต์ rule-based

งานวิจัย [34] ได้อธิบายให้เห็นว่างานกระบวนการสกัดข้อมูลได้รับความนิยมในการนำไปประยุกต์ใช้กับงานวิจัยทางด้านชีวการแพทย์ (Biomedical) เนื่องจากวิธีดังกล่าวสามารถนำไปสู่การ ค้นหาข้อมูล การได้ความองค์ความรู้ใหม่ และ การสร้างสมมติฐาน อย่างมีประสิทธิภาพตาม รูปแบบของการสกัดข้อมูล และ ปริมาณและความสมบูรณ์ของข้อมูล เช่น งานวิจัย [35] ได้เสนอวิธีการจดจำเอกลักษณ์ทางการแพทย์ เช่น ปัญหาสุขภาพ การทดสอบ และ

บำบัด เป็นต้นโดยวิธีที่ใช้ในการจดจำเอกลักษณ์เป็นวิธีการที่ผสมระหว่างการเรียนรู้ของเครื่องและวิธีการใช้กฎ โดยมีขั้นตอนการทำงานดังนี้

ขั้นตอนที่ 1 การประมวลผลเบื้องต้นการจดจำเอกลักษณ์ด้วยการเรียนรู้ด้วยเครื่องด้วยวิธี Condition Random Field-based NER

ขั้นตอนที่ 2 ทำการแก้ไขการทำนายที่ผิดพลาดและเพิ่มประสิทธิภาพของการทำงานในขั้นตอนถัดไปด้วยการใช้กฎสามแบบคือ

2.1 Part of Speech : กฎที่พิจารณาจากหน้าที่ของคำ

2.2 Orthographic : กฎที่พิจารณาจากรูปแบบของของคำ เช่น การขึ้นต้นขึ้นด้วยตัวพิมพ์ใหญ่ ลงท้ายด้วยตัวห้อยเฉพาะ หรือ มีอักขระพิเศษประกอบ

2.3 prefix and suffix : กฎที่พิจารณาจากคำขึ้นต้นและลงท้าย

ขั้นตอนที่ 3 การรวมโมดูล เป็นการสร้างโมดูลจากที่รวมผลลัพธ์จาก NER Classifiers 4 ตัว โดย Classifier ตัวที่ 1 ตัวที่ 2 และตัวที่ 3 มาจากระบบ NLP 4 ระบบที่แตกต่างกันตามลำดับดังนี้ คือ MedLEE KnowledgeMap และ DST สำหรับ Classifier ตัวที่ 4 นั้นเป็นการรวมคุณสมบัติและข้อมูลเข้าจากทั้งสาม Classifiers ของระบบ NLP ข้างต้น โดยขั้นตอนการรวมผลลัพธ์ของ Classifiers 4 ตัวประกอบไปด้วย 2 ขั้นตอนคือ

3.1 อินเตอร์เซคผลลัพธ์ของเอกลักษณ์ทางการแพทย์จากการสกัดด้วย classifier ตัวที่ 1 ตัวที่ 2 และตัวที่ 3 หรือหมายถึงเอกลักษณ์ที่ทั้ง 4 classifiers สามารถสกัดได้ตรงกัน

3.2 ทำการยูเนียนเอกลักษณ์ทางการแพทย์ที่ได้จากการสกัดด้วย classifier ตัวที่ 4 กับผลลัพธ์ของขั้นตอนที่ 3.1

ในงานวิจัยที่ [36] ได้มีการนำงานวิจัย [35] มาทำการทดลองด้วยวิธีการเรียนรู้แบบเชิงรุก (Active learning) เป็นการเรียนรู้แบบเลือกตัวอย่างโดยแบ่งกระบวนการวิธีออกได้กลุ่มได้ 3 กลุ่มคือ

1. วิธีการสอบถามบนพื้นฐานของความไม่แน่นอน (Uncertainty-based Querying Algorithms) เป็นการเลือกแบบประโยคที่ไม่มีความแน่นอน เนื่องจากตั้งสมมุติฐานว่าประโยคที่มีการติดป้ายที่ไม่แน่นอนมีสามารถให้ข้อมูลสำหรับการเรียนรู้ของเครื่องได้มากที่สุด มากกว่า

ประโยคที่มีการติดป้ายแน่นอน ทำให้ง่ายต่อการระบุเอกลักษณ์ซึ่งทำให้เครื่องไม่ได้เรียนรู้รูปแบบที่มีความกำกวมหรือความซับซ้อนมากกว่า

2. วิธีการสอบถามบนพื้นฐานของความหลากหลาย (Diversity-based querying algorithms) เป็นการเลือกประโยคที่ไม่มีความซ้ำซ้อนกันของประโยคที่มีการติดป้ายเดียวกันแต่มีรูปประโยคคล้ายคลึงกัน

3. วิธีการเลือกจากประโยค (Baseline algorithms) เป็นการเลือกโดยพิจารณาจากความยาวของประโยคหรือจำนวนของคอนเซ็บต์ทางคลินิกโดยให้สมมุติฐานว่าจำนวนที่มากกว่าให้ข้อมูลที่มากกว่า

การทดลองด้วยวิธีการเรียนรู้แบบเชิงรุก ทำโดยอาศัยข้อมูลที่แบ่งออกเป็น 2 กลุ่มตามประเภทของการติดป้าย คือ การติดป้ายตามประโยค (ALC1) และติดป้ายตามคำ (ALC2) จากทั้ง 3 วิธีพบว่าการสอบถามบนพื้นฐานของความไม่แน่นอนด้วยชุดข้อความ ALC2 สามารถให้ผลลัพธ์ตาม F1-score ได้สูงกว่าแบบไม่มีการเลือกตัวอย่าง(Passive Learning) เล็กน้อย ทำให้สามารถสรุปได้ว่าวิธีการใช้การเรียนรู้แบบเชิงรุกสามารถลดปริมาณของข้อมูลตัวอย่างโดยไม่ส่งผลกระทบให้กับการจดจำเอกลักษณ์ทางการแพทย์ได้ และ งานวิจัย [34] เป็นงานวิจัยสกัดความรู้ใหม่จากสื่อตีพิมพ์ที่ใช้สื่อตีพิมพ์ชีวการแพทย์เป็นกรณีศึกษาและมีเป้าหมายในการค้นหาความสัมพันธ์ระหว่างเอกลักษณ์ โดยจดจำเอกลักษณ์ด้วยวิธีการใช้พจนานุกรมและค้นหาความสัมพันธ์ด้วยวิธีการใช้กฎ

ความแตกต่างระหว่างวิธีการจำแนกด้วยการเรียนรู้ของเครื่องและการใช้กฎคือวิธีการเรียนรู้ของเครื่องจำเป็นต้องมีข้อมูลชุดสอนหรือข้อมูลตัวอย่างจำนวนมากและมีเนื้อหาของข้อมูลและโครงสร้างที่เหมาะสม เช่น งานวิจัย [37] ได้ทำการพัฒนาคลังข้อความทางด้านการแพทย์เพื่อนำไปใช้สำหรับงานประมวลผลภาษาทาง Evidence-Based Medicine (EBM) ด้วยปัญหามาจากขาดคลังข้อความที่เหมาะสมสำหรับงานประเภท EBM นอกจากคลังข้อความสำหรับงานเฉพาะด้านแล้ว ในงานวิจัย [38] ได้ทำการพัฒนาคลังข้อความทางการแพทย์มาตรฐานเพื่อใช้สำหรับงานทางด้านการประมวลผลภาษาธรรมชาติ ทำให้เห็นว่าชุดข้อมูลการสอนที่เหมาะสมมีความจำเป็นต่องานแต่ละประเภทที่มีคุณลักษณะที่แตกต่างกันไป รวมไปถึงปริมาณก็เป็นตัวแปรสำคัญต่อผลของการจำแนกด้วย ตัวอย่างที่มากขึ้นสามารถให้ข้อมูลเพื่อใช้ในการจำแนกหรือทำนายได้มากขึ้น แต่ในขณะเดียวกันการรวบรวมข้อมูลเพื่อและทำให้ข้อมูลที่รวบรวมได้อยู่ในโครงสร้างเดียวกันก็เป็นเรื่องที่ต้องใช้เวลาและทรัพยากรบุคคลเช่นกัน อย่างที่สามารถเห็นได้ในงานวิจัย [36] ที่มีความ

พยายามลดจำนวนของข้อมูลตัวอย่างเพื่อลดค่าใช้จ่ายในการเตรียมข้อมูลตัวอย่างที่มีการติดป้ายเพื่อใช้สำหรับสอนเครื่อง ต่างกับวิธีการใช้กฎซึ่งอาศัยข้อมูลตัวอย่างปริมาณน้อยกว่าหรือไม่ใช้เลยแต่ใช้งานได้กับการจำแนกที่มีความซับซ้อนในข้อมูลน้อยและหารูปแบบของโครงสร้างได้ไม่ยาก เมื่อพิจารณาถึงคลังข้อความที่มีอยู่ในภาษาไทย National Electrics and Computer Technology Center (NECTEC) มีบทบาทสำคัญในการเริ่มต้นการค้นคว้าและวิจัยในการสร้างคลังข้อความเพื่อใช้ในงานทาง NLP ภาษาไทย เริ่มต้นจากคลังข้อความ Orchid [14] ที่เป็นคลังข้อความการประชุมทางวิชาการทางอิเล็กทรอนิกส์ มีการแบ่งขอบเขตคำและติดป้ายบทบาทของแต่ละคำอย่างชัดเจน ถัดมาเป็นคลังข้อความมาตรฐานในการวัดผลสำหรับงานสำหรับงาน NLP ภาษาไทย [15] ที่เป็นคลังข้อความมาตรฐานสำหรับการประมวลผลภาษาไทย แต่อย่างไรก็ตามคลังข้อความทางด้านการแพทย์ยังคงไม่ได้รับความสนใจมากพอจะได้รับการสนับสนุนในการพัฒนาคลังข้อความที่มีความเหมาะสมสำหรับการนำไปใช้การประมวลผลในระบบทางการแพทย์ได้

นอกเหนือจากการสกัดข้อมูลทางการแพทย์จากข้อมูลด้วยภาษาอังกฤษแล้ว ได้มีความพยายามในการสกัดข้อมูลบนฐานของภาษาทางตะวันออก เช่นงานวิจัย [39] ได้เสนอวิธีการจำแนก CC ออกเป็นกลุ่มอาการที่รองรับหลายภาษา โดยใช้ภาษาจีนเป็นกรณีทดสอบ โดยใช้วิธีการแปลภาษาจากวลีสำคัญที่ได้จากการสกัดจาก CC ภาษาจีนด้วยวิธีทางสถิติและนำข้อความที่ผ่านการแปลเป็นภาษาอังกฤษแล้วไปจำแนกกลุ่มอาการด้วยวิธีการจำแนก CC ภาษาอังกฤษที่มีอยู่ ชุดข้อมูลที่ใช้ในการทดสอบมีจำนวน 939,024 ระเบียน โดยวิธีการเลือกตัวเลือกวลีสำคัญของงานวิจัยดังกล่าวมีชื่อว่า Extended Mutual Information(EMI) เมื่อกำหนดให้ p : $\{p_1,p_2, \dots , p_n\}$ เป็นรูปแบบวลีที่สนใจ และ f(p) เป็นความถี่ที่เกิดขึ้นของ p และ r : $\{p_2, p_3, \dots , p_n\}$ และ l : $\{p_1, p_2, \dots , p_{n-1}\}$ เป็นรูปแบบรองที่ยาวที่สุดขวาและซ้ายตามลำดับ ค่า EMI สามารถคำนวณได้ตามสมการที่ (1)

$$EMI = \frac{f(p)}{f(l)+f(r)-f(p)} \quad (1)$$

เมื่อค่า EMI ตามสมการที่ (1) เข้าใกล้ 1 เท่าไรก็มีความมีความเป็นไปได้ว่ามีความสำคัญเท่านั้น เช่น $p$ = “ABC” จะได้ว่า $r$ และ $l$ เป็น “BC” และ “AB” ตามลำดับ เมื่อทำการทดลองกับชุดข้อมูลพบว่าได้ตัวเลือกวลีสำคัญจำนวน 2533 วลีจากนั้นทำการเลือกวลีสำคัญจากตัวเลือกทั้งหมดด้วยคน โดยทำการตัดตัวเลือกเมื่อ

1. ตัวเลือกไม่ใช่วลีที่มีความหมาย
2. ตัวเลือกไม่ได้บรรจุข้อมูลที่เกี่ยวข้องกับ Syndromic Surveillance
3. ความหมายของตัวเลือกสามารถสื่อได้จากวลีที่สั้นกว่าที่อยู่ในตัวเลือก

ตัวเลือกทั้งหมด 2533 วลีผ่านเกณฑ์จำนวน 415 วลีและเพิ่มจากพจนานุกรมสองภาษา จีน-อังกฤษ โดยกำหนดเกณฑ์ว่าหากวลีดังกล่าวปรากฏใน CC ของชุดข้อมูลทดลองเกิน 5 ครั้งให้ทำการเพิ่มวลีดังกล่าวลงในเซตของวลีสำคัญ โดยได้วลีจากพจนานุกรมอีก 55 วลี ร่วมวลีสำคัญทั้งสิ้น 470 วลี จากนั้นได้ให้นายแพทย์จำนวน 3 คนทำการแปลวลีสำคัญเป็นภาษาอังกฤษ และ ทำการทบทวนซ้ำเพื่อความแน่ใจในความคงเส้นคงวา

ในการแปล CC เป็นภาษาอังกฤษ วลีสำคัญถูกเพิ่มลงไปในพจนานุกรม เพื่อใช้ในการตัดคำภายในข้อความ CC โดยใช้วิธีตัดจากคำที่ยาวที่สุดในพจนานุกรมก่อนและแปลคำที่ถูกตัดบนฐานของวลีสำคัญ และใช้วิธีการจำแนกจากงานวิจัย [33] ในการจำแนก CC ออกเป็นกลุ่มอาการ

จากงานวิจัยที่ผ่านมาทำให้เห็นว่าปัญหาสำคัญที่เกิดขึ้นคือขาดคลังข้อความที่เหมาะสมต่อการนำมาประยุกต์ใช้ในงาน NLP ทางการแพทย์ ในการวิจัย [39] ก็ได้ให้เหตุผลเช่นเดียวกันว่าการตัดคำโดยใช้คลังข้อความที่เหมาะสมหรือประเภทเดียวกันกับชุดทดสอบสามารถทำให้ปฏิบัติงานการตัดคำได้อย่างมีประสิทธิภาพ แต่ขาดคลังข้อความที่เหมาะสมจึงจำเป็นต้องตัดคำด้วยวิธีใช้พจนานุกรมแทน โดยจะเห็นได้ว่าถึงแม้จะมีข้อมูลเป็นปริมาณมากแต่ก็เป็นข้อมูลดิบที่ต้องทำการสกัดข้อมูลเพื่อสามารถนำไปใช้ในการช่วยจำแนกกลุ่มอาการหรือก็คือวลีสำคัญ จะเห็นได้จากปัญหาที่เกิดขึ้นจากขั้นตอนดังกล่าวคือจำเป็นต้องอาศัยผู้เชี่ยวชาญและเวลาในกระบวนการเช่นเดียวกับความคงเส้นคงวาที่ต้องตรวจสอบด้วยบุคคลทำให้มีความเป็นไปได้ที่จะเกิดความไม่คงเส้นคงวาในการตรวจสอบตามปริมาณของวลีสำคัญและวิจารณญาณของผู้ตรวจสอบ นอกจากนี้ยังสามารถเกิดปัญหาได้จากการมีส่วนขยายเป็นคำเชื่อมระหว่างวลีทำให้ไม่ถูกแปลเป็นภาษาอังกฤษเนื่องจากแปลเฉพาะวลีสำคัญจึงทำให้วลีต้องมีความครอบคลุมในความเป็นไปได้ที่แตกต่างสามารถเห็นได้ว่าในงานวิจัย [39] ใช้ค่าเกณฑ์ EMI ที่ต่ำ (แต่ไม่ได้ระบุชัดเจนว่าเป็นค่าเท่าไร) เพื่อให้สามารถได้ตัวเลือกในปริมาณที่มากพอจะครอบคลุมวลีที่มีความเป็นไปได้ ที่จะให้ข้อมูลในการจำแนกกลุ่มและนำมาตัดตัวเลือกในภายหลัง

# บทที่ 3

## การวิเคราะห์และออกแบบกระบวนการวิธีการจำแนกอาการและอาการแสดง

ในการจำแนกประเภทของกลุ่มข้อมูล ปัจจัยที่มีความจำเป็นต่อการจำแนกหรือแยกประเภทประกอบด้วย ชุดข้อมูลชุดสอน และ วิธีการจำแนก เช่นเดียวกันกับการจำแนกอาการและอาการแสดงก็มีความจำเป็นที่จะต้องมีคลังข้อความอาการและอาการแสดงที่เหมาะสมและกระบวนการจำแนกที่สามารถจัดการจำแนกข้อมูลออกเป็นประเภทได้อย่างมีประสิทธิภาพ ดังนั้นการเลือกข้อมูลชุดสอนและวิเคราะห์ทำความเข้าใจเพื่อออกแบบวิธีการเตรียมชุดข้อมูลและวิธีการจำแนกอาการเจ็บป่วยจึงเป็นสิ่งจำเป็น

## 3.1 ที่มาของปัญหา

ปัญหาของการวิจัยเกี่ยวกับการจำแนกประเภทข้อมูลที่เป็นข้อความหรือเอกสารที่อยู่ในรูปแบบภาษาธรรมชาติ คือความกำกวม (Ambiguity) โดยเฉพาะอย่างยิ่งในภาษาที่ไม่มีสัญลักษณ์สำหรับการแบ่งหรือกำหนดขอบเขตของคำที่ชัดเจน ต่างจากภาษาอังกฤษที่ถือเป็นภาษาธรรมชาติที่เป็นภาษาสากลของโลกและมีงานค้นคว้าวิจัยอย่างต่อเนื่อง เพื่อสามารถนำไปใช้กับกระบวนการจำแนกหรือสกัดขอมูลจากข้อความได้ ทำให้การวิจัยการสกัดข้อมูลบนเอกสารทางการแพทย์ในภาษาอื่นๆยังคงถือเป็นเรื่องที่ท้าทายอยู่ โดยปัญหาหลักในการวิจัยการจำแนกอาการและอาการแสดงภาษาไทยแจกแจงได้ดังนี้

### 3.1.1 ขาดคลังข้อความที่เหมาะสม

จากที่ได้กล่าวถึงคลังข้อความภาษาไทยในหัวข้อที่ 2.4 จะเห็นได้ว่าปัจจุบันยังคงไม่มีคลังข้อความทางการแพทย์ภาษาไทยที่พร้อมใช้งานได้ แต่ในการวิจัย [39] ที่ทำการวิจัยจำแนก CC โดยใช้ภาษาจีนเป็นกรณีศึกษาโดยใช้วลีสำคัญในการแปลภาษาจีนเป็นภาษาอังกฤษและจำแนกด้วยวิธีการจำแนกทางภาษาอังกฤษที่มีอยู่ แต่การหาวลีสำคัญด้วยวิธีการทางสถิติจำเป็นที่จะต้องมีปริมาณข้อมูลมากพอที่จะแสดงให้เห็นถึงความแตกต่างระหว่างความถี่ที่เกิดขึ้นของแต่ละชุดอักขระหรือวลีและครอบคลุมความเป็นไปได้ทั้งหมด แม้วิธีดังกล่าวจะไม่ได้จำเป็นต้องใช้ข้อมูลที่มีโครงสร้างเหมาะสมแค่มีข้อความที่มีอยู่ในหัวข้อที่สนใจก็สามารถหาวลีออกมาได้

### 3.1.2 มาตรฐานในการสื่อความหมาย

ความแตกต่างทางภาษายังคงเป็นอุปสรรคสำคัญในการแบ่งปันข้อมูลระหว่างภูมิภาคที่มีภาษาต่างกัน เช่นการแปลวลีสำคัญในงานวิจัยที่ [39] ที่ทำการแปลโดยแพทย์และทำการตรวจสอบความคงเส้นคงวาด้วยตัวผู้ทำวิจัย อย่างไรก็ตามวลีหรือคำแปลดังกล่าวไม่มีอะไรยืนยันหรือไม่มีมาตรฐานในการรองรับว่าจะเป็นคำเดียวกันหรือรูปแบบวลีเดียวกันในภาษาอื่น ทำให้การแบ่งปันข้อมูลหรือการเก็บข้อมูลเพื่อนำไปใช้ในการทำนายผลต่อไปทำได้ยาก

## 3.2 ข้อเสนอสำหรับแก้ปัญหา

จากปัญหาที่ได้กล่าวไว้ข้างต้นได้แก่การขาดคลังข้อความที่เหมาะสมและขาดมาตรฐานกลางในการสื่อความหมาย รวมถึงการใช้วิธีอาศัยข้อมูลทางสถิติก็จำเป็นต้องอาศัยข้อมูลปริมาณมากในการนำไปปฏิบัติ อย่างไรก็ตามการจะนำข้อมูลปริมาณมากมาใช้ในการทดลองจำเป็นต้องได้รับความร่วมมือจากหน่วยงานที่เกี่ยวข้องหรือมีข้อมูลมาตรฐานสำหรับการทดสอบที่มีการเผยแพร่สาธารณะซึ่งในปัจจุบันยังคงมีข้อมูลหรือเอกสารทางการแพทย์ที่สามารถใช้ในการทดลองหรือวิจัยการสกัดข้อมูลหรือประมวลผลภาษาไม่เพียงพอที่จะสามารถนำมาใช้ในการกระบวนการทางสถิติได้ งานวิจัยนี้จึงได้เสนอวิธีการเตรียมข้อมูลจากข้อความ ICD ออกเป็นหน่วยย่อยเพื่อสามารถนำไปจำแนกอาการและอาการแสดงบนพื้นฐานเชิงโครงสร้างได้ เนื่องจากในหลายประเทศมีเอกสาร ICD เป็นภาษาประจำชาติของตนเอง

### 3.2.1 การนำรหัส ICD มาใช้ในการระบุอาการและอาการแสดง

ในเอกสาร [49] ได้ระบุถึงการใช้รหัส ICD สำหรับแสดงในบิลเรียกเก็บเงินสำหรับการสรุปการรักษาของผู้ป่วย อย่างไรก็ตามรหัส ICD ก็ถือเป็นมาตรฐานที่ใช้กันทั่วโลกและมีหลายประเทศที่มีเอกสารรหัส ICD เป็นภาษาของตัวเอง ดังนั้นการใช้รหัส ICD แทนข้อมูลที่ในปรากฏในข้อความสามารถแก้ไขความแตกต่างทางภาษาให้สามารถเข้าใจถึงข้อมูลที่บรรจุอยู่ในข้อความได้

### 3.2.2 การใช้ข้อความหรือวลีจากเอกสาร ICD มาเป็นข้อมูลชุดสอน

ในแต่ละรหัสของ ICD จะมีข้อความกำกับถึงรายละเอียดของอาการหรืออาการแสดงในแต่ละภาษาของประเทศนั้น อย่างไรก็ตามข้อความในเอกสาร ICD ถือเป็นข้อความมาตรฐาน สำหรับรูปแบบของข้อความหรือวลีที่ปรากฏในการข้อความ CC อาจไม่ได้มีรูปแบบหรือโครงสร้างเดียวกับวลีตามรหัส ICD โดยมีความเป็นไปได้ตามตัวอย่างต่อไปนี้

กรณีที่มีคำเชื่อมเป็นส่วนขยาย เช่น

"ปวดศีรษะ" ⟶ "ปวดบริเวณศีรษะ"

"เจ็บหน้าอก" ⟶ "เจ็บบริเวณหน้าอก"

กรณีที่มีลำดับคำที่แตกต่าง เช่น

"เจ็บหน้าอกเวลาหายใจ" ⟶ "เวลาหายใจเจ็บหน้าอก"

กรณีที่บางคำในวลีหายไปหรือถูกแทนด้วยคำอื่น เช่น

"เจ็บหน้าอกเวลาหายใจ" ⟶ "เจ็บหน้าอกตอนหายใจ"

จากกรณีตัวอย่างขั้นต้นทำให้การจะใช้วิธีเปรียบเทียบตัวอักษรของข้อความต่อทุกกรณีเป็นไปได้ยาก เนื่องจากไม่มีข้อความตัวอย่างมากพอจะครอบคลุมทุกข้อความที่เป็นไปได้ จึงจำเป็นต้องหาวิธีอื่นเข้ามาช่วยในการจำแนกอาการและอาการแสดง

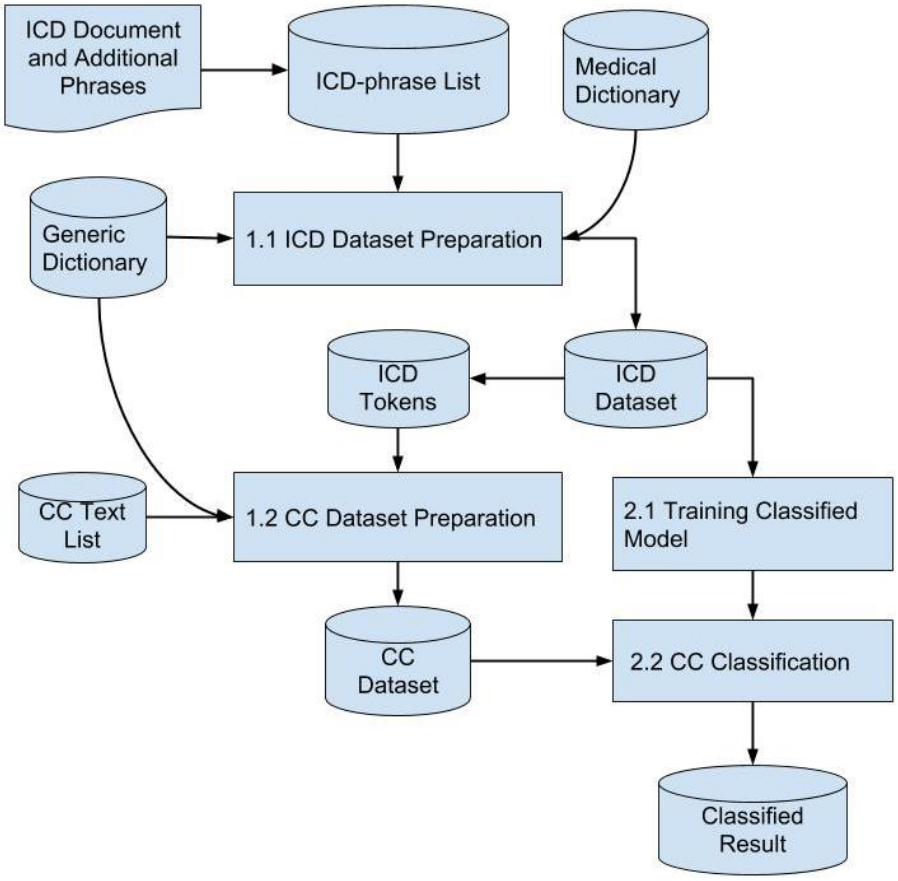
## 3.3 การออกแบบกระบวนการจำแนกอาการและอาการแสดง

ในการออกแบบกระบวนการจำแนก CC ออกเป็นอาการและอาการแสดงตามมาตรฐานรหัส ICD ออกแบบโดยอาศัยพื้นฐานของการเตรียมข้อมูลโดยยึดหลักจากโครงสร้างของวลีจาก ICD เป็นหลัก เนื่องจากวลีของ ICD ถูกใช้เป็นส่วนสำคัญในการระบุอาการและอาการแสดง และเมื่อสามารถเตรียมชุดข้อมูลวลี ICD เสร็จสิ้น ก็จะทำการเตรียมชุดข้อมูล CC โดยใช้โครงสร้างของวลีภายในชุดข้อมูลวลี ICD ที่ผ่านการเตรียมแล้วเป็นต้นแบบในการเตรียมชุดข้อมูล CC เพื่อให้ข้อมูลทั้งสองชุดมีความสอดคล้องกันและสามารถนำไปจำแนกโดยใช้มาตรฐานเดียวกันได้

### 3.3.1 ภาพรวมของกระบวนการจำแนกอาการและอาการแสดง

กระบวนการจำแนกอาการและอาการแสดงเริ่มต้นจากกระบวนการเตรียมชุดข้อมูล ICD โดยก่อนจะเริ่มกระบวนการ จะทำการเพิ่มวลีที่เป็นไปได้บนฐานของคำพ้องความหมาย เช่นคำว่า "ปัสสาวะ" สามารถแทนหรือมีความหมายเช่นเดียวกับคำว่า "ฉี่" แต่คำว่า "ปัสสาวะ" จะเป็นคำที่อยู่ในรูปแบบทางการมากกว่าคำว่า "ฉี่" แต่คำว่า "ฉี่" กลับเป็นคำที่ถูกใช้อย่างทั่วไปบ่อยกว่าคำว่า "ปัสสาวะ" ถึงแม้จะคงยังไม่ครอบคลุมทุกข้อความที่เป็นไปได้ที่เกิดขึ้นแต่ทำให้สามารถแก้ไขปัญหาคำพ้องความหมายได้

เมื่อทำการเพิ่มวลีที่เป็นไปได้จากพื้นฐานของคำพ้องความหมาย ก็จะทำการจัดเก็บข้อมูลในฐานข้อมูลเพื่อใช้ในกระบวนต่างๆ ตามรูปที่ 3.1 โดยมีลำดับของกระบวนการดังต่อไปนี้




รูปที่ 3.1 กระบวนการจำแนกอาการและอาการแสดง

1. การเตรียมข้อมูล (Data Preparation)
   1.1. การเตรียมชุดข้อมูล ICD (ICD Dataset Preparation) โดยใช้วิธีการประมวลผลภาษาธรรมชาติและพจนานุกรมภาษาไทยทั่วไป และ พจนานุกรมทางการแพทย์ภาษาอังกฤษในการทำงานของกระบวนการนี้
   1.2. การเตรียมชุดข้อมูล CC (CC Dataset Preparation) เป็นการเตรียมข้อมูล CC โดยอาศัยส่วนประกอบของโครงสร้างข้อความ ICD ที่ผ่านการเตรียมแล้ว (ICD tokens) และพจนานุกรมภาษาไทยทั่วไปในการเตรียมชุดข้อมูล CC
2. การจำแนก CC ออกเป็นรหัส ICD (CC Classification based on ICD code)
   2.1. การสร้างโมเดลสำหรับการจำแนก (Training Classified Model) คือการอาศัยชุดข้อมูลที่ผ่านการเตรียมแล้วมาเป็นข้อมูลตัวอย่างสำหรับการสร้างโมเดลการจำแนกอาการและอาการแสดง

2.2. การจำแนก CC (CC Classification) ในกระบวนการนี้จะทำการนำโมเดลจำแนกอาการและอาการแสดงที่ผ่านการสอนในขั้นตอนที่ 2.1 มาทดลองจำแนกกับชุดข้อมูลทดสอบ

โครงสร้างของกระบวนการจำแนก CC ออกเป็นรหัส ICD มีกระบวนการหลัก 2 กระบวนการคือ การเตรียมข้อมูลและการจำแนกข้อความบอกเล่าอาการสำคัญโดยรายละเอียดของทั้งสองกระบวนการจะทำการอธิบายในหัวข้อถัดไป

## 3.4 กระบวนการเตรียมชุดข้อมูล ICD

วิธีการจำแนกหรือจดจำข้อความที่ไม่ซับซ้อน แต่มีประสิทธิภาพคือวิธีการใช้การเปรียบเทียบชุดอักขระ แต่ในขณะเดียวกันก็เป็นวิธีที่อาศัยทรัพยากรหรือคลังคำศัพท์เฉพาะทางปริมาณมากเช่นกัน เพราะจำเป็นที่จะต้องมีคำศัพท์ครอบคลุมความเป็นไปได้ของชุดอักขระที่จะเกิดขึ้น โดยวิธีการที่จะได้ชุดคำศัพท์โดยไม่อาศัยข้อมูลตัวอย่างสามารถแบ่งออกได้เป็นสองกลุ่มคือ

1. การสกัดชุดคำศัพท์ด้วยวิธีการสถิติ

ในกระบวนการมีความต้องการเอกสารจำนวนมากพอที่จะสามารถใช้วิธีทางสถิติในการสกัดคำศัพท์ที่เป็นไปได้และอาจจำเป็นต้องอาศัยทรัพยากรบุคคลที่มีความเชี่ยวชาญในสาขาดังกล่าว เพื่อทำการทบทวนชุดคำศัพท์เพื่อยืนยันความถูกต้องและความสอดคล้องของข้อมูล

2. การสร้างชุดคำศัพท์ด้วยผู้เชี่ยวชาญ

การสร้างชุดคำศัพท์ด้วยผู้เชี่ยวชาญมีความเหมาะสมอย่างยิ่งกับหัวข้องานวิจัยที่มีขนาดเล็กและลักษณะคำศัพท์ที่ต้องการสกัดเป็นนามเอกลักษณ์ (Name Entities) ที่มีลักษณะเฉพาะหรือรูปแบบในการจดจำที่แน่นอนชัดเจนและใช้เวลาน้อยกว่าวิธีการทางสถิติบนเงื่อนไขที่กล่าวมา

สำหรับงานวิจัยนี้อาศัยข้อความจากเอกสาร ICD เป็นชุดคำศัพท์เพื่อการสอน อย่างไรก็ตามชุดคำศัพท์ดังกล่าวไม่สามารถนำมาใช้กับวิธีการเปรียบเทียบอักขระได้ เนื่องจากลักษณะของข้อความอาการและอาการแสดง เป็นลักษณะของข้อความบอกเล่าที่สามารถแสดงได้หลายรูปแบบที่ให้ความหมายเดียวกัน นอกจากนี้ข้อความจากเอกสาร ICD ไม่สามารถครอบคลุมทุกความเป็นไปได้ที่จะเกิดข้อความอื่นๆในความหมายเดียวกัน ในขณะเดียวกันการนำวิธีทางสถิติมาใช้นอกจากจะต้องการข้อความ CC ปริมาณมากแล้ว ยังจำเป็นที่จะต้องระบุด้วยว่าข้อความดังกล่าวอยู่ในรหัสใดในมาตรฐาน ICD ทำให้จำเป็นต้องอาศัยทรัพยากร บุคลากร และเวลาปริมาณมาก ในขณะเดียวกันการอาศัยทรัพยากรและบุคลากรจำเป็นต้องอาศัยความร่วมมือจากองค์กรหรือ

สถาบันที่มีความเกี่ยวข้อง แต่ในขณะที่ทำการวิจัยยังไม่มีข้อมูล CC ที่เผยแพร่สาธารณะจำนวนมากพอที่จะสามารถใช้วิธีการทางสถิติได้ ประกอบกับการไม่สามารถขอความอนุเคราะห์ข้อมูลเพื่อนำไปใช้ในการวิจัยได้สำเร็จ ทำให้จำเป็นต้องหาวิธีอื่นในการจำแนกหรือจดจำข้อความอาการและอาการแสดงภายใต้ข้อจำกัดดังกล่าว ดังนั้นงานวิจัยนี้จึงอาศัยวิธีการแปลงข้อความให้อยู่ในรูปโครงสร้างตามความเกี่ยวข้องของคำศัพท์ทางการแพทย์ เพื่อให้สามารถอาศัยโครงสร้างของข้อความจำแนกประเภทของข้อมูล แทนการเปรียบเทียบด้วยตัวอักษรได้

การเตรียมชุดข้อมูล ICD มีความสำคัญและต้องทำเป็นขั้นตอนแรกเนื่องจากข้อความจากชุดข้อมูล ICD จะถูกใช้เป็นตัวอย่างในการระบุรหัส ICD และองค์ประกอบของโครงสร้างข้อความจะถูกนำไปใช้ในการสร้างชุดข้อความ CC ในขั้นตอนถัดไปและเนื่องจากไม่สามารถใช้วิธีการเปรียบเทียบทั้งข้อความได้เนื่องจากข้อความที่มีไม่มากพอที่จะครอบคลุมความเป็นไปได้ จึงจำเป็นต้องมีการเตรียมชุดข้อมูลเพื่อพร้อมที่สามารถรองรับวิธีการจำแนกอื่นได้

### 3.4.1 การศึกษาและวิเคราะห์ข้อความ

ความซับซ้อนของภาษาเกิดขึ้นเมื่อคำหนึ่งคำสามารถแทนได้หลายคำและหลายรูปแบบของข้อความโดยสามารถจำแนกปัจจัยที่ทำให้เกิดความกำกวมได้ดังนี้

#### 1. คำพ้องความหมาย

ความกำกวมของคำพ้องความหมายเกิดจากการที่คำมากกว่าหนึ่งคำมีความหมายเดียวกัน เช่น คำว่า "ศีรษะ” และ "หัว” มีความหมายเช่นเดียวกันและสามารถนำมาประกอบเพื่อสร้างข้อความที่มีรูปที่แตกต่างและให้ความหมายเดียวกันได้เช่น "ปวดศีรษะ” และ "ปวดหัว”

#### 2. คำขยาย

คำขยายก่อให้เกิดความกำกวมจากการขยายข้อความเพื่อให้ใจความสมบูรณ์มากขึ้น เช่น "เจ็บหน้าอก" สามารถเพิ่มส่วนขยายด้วยคำว่า "บริเวณ” ในการระบุตำแหน่งเปลี่ยนรูปข้อความจาก "เจ็บหน้าอก" เป็น "เจ็บบริเวณหน้าอก” ที่มีรูปแบบโครงสร้างข้อความต่างออกไปทำให้ยากต่อการจำแนกด้วยการประมวลผลทางคอมพิวเตอร์ว่าทั้งสองข้อความมีความหมายเดียวกัน สำหรับการตัด Stop words สามารถช่วยลดปัญหาความกำกวมจากคำที่ไม่มีนัยสำคัญได้ แต่เนื่องจากในแต่ละบริบทแต่ละคำมีนัยสำคัญแตกต่างกัน เช่น คำว่า “ท้องขึ้น” และ “เลือดกำเดาออก” คำว่า “ขึ้น” และ “ออก” เป็นคำที่ปรากฏอยู่ในรายการ Stop words ดังแสดงในหัวข้อที่ 2.12

จึงทำให้ยังไม่สามารถนำการตัด Stop words มาใช้ได้เนื่องจากอาจทำให้คำที่มีนัยสำคัญสำหรับการสื่ออาการขาดหายไป

**3. โครงสร้างของไวยากรณ์**

ในแต่ละภาษามีโครงสร้างของไวยากรณ์ที่แตกต่างกันไปตามแต่ละภาษาและในขณะเดียวกันข้อความต่างข้อความที่ถึงแม้จะมีคำศัพท์ชุดเดียวกันแต่มีรูปแบบต่างกันก็สามารถให้ความหมายเดียวกันได้เช่น “เจ็บหน้าอกเวลาหายใจ” และ “เวลาหายใจเจ็บหน้าอก” จะเห็นได้ว่าทั้งสองข้อความประกอบไปด้วยชุดคำศัพท์เดียวกันและมีลำดับการจัดวางต่างกันแต่มีความหมายเดียวกันจึงทำให้ไม่สามารถสรุปได้ทันทีว่าการที่มีโครงสร้างข้อความที่ต่างกันจะให้ความหมายที่แตกต่างกัน

**3.4.2 การตัดคำภาษาไทย**

การที่จะสามารถจำแนกข้อความโดยที่ไม่ใช้การเปรียบเทียบความเหมือนของทุกลำดับตัวอักษรหรือการเปรียบเทียบข้อความนั้น จำเป็นที่จะต้องแยกหน่วยของข้อความให้เล็กลงเพื่อจะสามารถเปรียบเทียบในเชิงส่วนประกอบของโครงสร้างข้อความได้ ซึ่งวิธีการตัดคำสามารถเข้ามาจัดการในส่วนของการแยกองค์ประกอบของข้อความได้ อย่างไรก็ตามยังคงมีปัจจัยอื่นเป็นส่วนประกอบในกระบวนการที่มีผลต่อผลลัพธ์ต่อการตัดคำ โดยสามารถแจกแจงได้ตามวิธีการตัดคำซึ่งถูกแบ่งออกเป็นสองประเภทหลักคือ การตัดคำด้วยวิธีการเรียนรู้ของเครื่อง และ การตัดคำโดยอาศัยพจนานุกรม โดยการตัดคำด้วยวิธีการเรียนรู้ของเครื่องมีข้อจำกัดสำหรับการเตรียมข้อความ ICD เพื่อนำไปใช้ในกระบวนการจำแนกอาการและอาการแสดงดังนี้ คือ ขาดคลังข้อความสำหรับสอนที่เหมาะสม ไม่สามารถให้ผลลัพธ์ที่คงเส้นคงวาได้ และ นำไปปฏิบัติได้ยากกว่าวิธีการใช้พจนานุกรม ในทางกลับกันการตัดคำด้วยพจนานุกรมสามารถทำได้ง่ายและมีความคงเส้นคงวามากกว่า และมีข้อจำกัดที่ใกล้เคียงกับการตัดคำด้วยการเรียนของเครื่องคือขาดพจนานุกรมที่เหมาะสม อย่างไรก็ตาม ด้วยเหตุที่สามารถให้ผลลัพธ์ที่คงเส้นคงวาซึ่งมีผลต่อการนำไปใช้ในกระบวนการจำแนก CC ที่อาศัยโครงสร้างของข้อความในการจำแนกซึ่งหากโครงสร้างของข้อความขาดคงเส้นคงวาแล้วจะทำให้การจำแนกด้วยโครงสร้างเป็นไปได้ยากด้วยเช่นกัน

**3.4.3 การหาความขัดแย้งกันของข้อความ (Conflict Element Finding)**

จากการศึกษาและวิเคราะห์ข้อความในหัวข้อ 3.4.1 เพื่อแก้ปัญหาความแตกต่างทางโครงสร้างของข้อความ สิ่งที่ทำให้เกิดความกำกวมประกอบไปด้วย คำพ้องความหมาย คำขยาย และโครงสร้างของไวยากรณ์ การที่จะพิจารณาในเชิงโครงสร้างจำเป็นที่จะต้องอาศัยการตัดคำเข้ามาช่วยในแยกหน่วยของข้อความออกเป็นส่วนย่อย แต่จะมั่นใจได้อย่างไรว่าข้อความที่ตัดสามารถแยกหน่วยของข้อความได้อย่างเหมาะสม และสามารถนำไปใช้ในกระบวนการจำแนกอาการได้ ผลของการตัดคำด้วยพจนานุกรมนั้นขึ้นอยู่กับพจนานุกรมที่ใช้และแต่ละพจนานุกรมมีความแตกต่างกันไปตามจุดประสงค์และปริมาณของคำศัพท์ ดังนั้นจึงไม่มีสิ่งใดสามารถยืนยันได้ว่าผลของการตัดคำด้วยพจนานุกรมจะมีความเหมาะสมสำหรับนำมาใช้ในการจำแนกข้อความอาการและอาการแสดง ดังนั้นจึงจำเป็นที่จะต้องมีกระบวนการในการตรวจสอบโครงสร้างบนพื้นฐานของการแก้ปัญหา คำพ้องความหมาย คำขยาย และ โครงสร้างข้อความด้วยการหาความขัดแย้งที่เกิดขึ้นบนองค์ประกอบของข้อความที่แตกต่างกัน เช่น คำว่า “ปวดหัว” และ “ปวดท้อง” คำทั้งสองคำเมื่อผ่านการตัดด้วยวิธี LM แล้วยังคงอยู่ในรูป “ปวดหัว” และ “ปวดท้อง” เนื่องจากมีคำทั้งสองคำอยู่ในพจนานุกรมจึงไม่ถูกตัดให้อยู่ในหน่วยย่อย ในทางกลับกันหากใช้วิธีเดียวกันคือวิธีการ LM ในการตัดคำว่า “ปวดที่ท้อง” จะถูกแบ่งออกเป็น “ปวด” “ที่” “ท้อง” เมื่อทำการเปรียบในเชิงองค์ประกอบแล้วจะเห็นได้ว่า “ปวดท้อง” และ “ปวด” “ที่” “ท้อง” ไม่มีความคล้ายในเชิงองค์ประกอบแม้จะมีความหมายเดียวกัน แต่ถ้าหากสามารถทำให้ “ปวดท้อง” อยู่ในรูป “ปวด” “ท้อง” ได้จะพบว่าทั้งสองข้อความมีองค์ประกอบที่ใกล้เคียงกัน เมื่อทำการสังเกตจะพบว่าทั้ง “ปวดหัว” และ “ปวดท้อง” มีคำว่า “ปวด” เป็นองค์ประกอบร่วมกันทำให้เห็นว่ามีความเป็นไปได้ที่คำว่า “ปวด” จะเป็นองค์ประกอบที่สามารถแทรกด้วยคำขยายหรือประกอบกับคำอื่นได้ โดยเรียกว่าองค์ประกอบที่ขัดแย้ง (Conflict Element : CE) ซึ่งจำเป็นที่จะต้องแยกองค์ประกอบเพื่อให้สามารถเปรียบเทียบความคล้ายคลึงในองค์ประกอบได้ โดยในงานวิจัยได้เสนอการหา CE ด้วยวิธีการต่อไปนี้

เมื่อกำหนดให้

$P_i$ = วลีใด $T_i = \{t_{i,j}, t_{i,j}$ เป็นองค์ประกอบของ $P_i$ ที่ข้อความ $LM\}$

$CE = t_{i,j} \cap t_{i+c,j+c}$ เมื่อ $c$ เป็นค่าใดๆ

ตัวอย่างข้อความสำหรับแสดงกระบวนการ Conflict Element finding

$P_0$ = ปวดท้องน้อย $T_0$ = {ปวดท้อง, น้อย}

$P_1$ = ปวดฝีเย็บ $T_1$ = {ปวด, ฝีเย็บ}

เมื่อ $P_0$ และ $P_1$ ผ่านกระบวนการตัดคำด้วย LM ก็จะได้ $T_0$ และ $T_1$ ตามในตารางที่ 3.1 จากนั้นก็จะทำการหา CE ด้วยวิธีการอินเตอร์เซคแต่ละ element ของ $T_0$ และ $T_1$ และกำหนดให้ผลของการอินเตอร์เซคเป็น CE และส่วนที่เหลือกำหนดให้เป็น องค์ประกอบที่เหลือ (Remaining Element : RE) ตามที่เห็นในตารางที่ 3.2 โดยจะเห็นว่าในลำดับที่ 1 เกิด Conflict ขึ้นและทำการตรวจว่าทั้ง CE และ RE อยู่ในพจนานุกรมหรือเพื่อให้แน่ใจว่า CE และ RE ไม่ใช่ชุดตัวอักษรที่ไม่มีความหมาย และให้สัญลักษณ์ Ø แทนเซตว่างของผลลัพธ์การอินเตอร์เซคกรณีที่ไม่เกิดความขัดแย้งระหว่าง $t$ ใดๆ หลังจากทำการยืนยันว่า CE และ RE เป็นคำที่มีความหมายทั้งหมดแล้ว จะทำการหาความสัมพันธ์ทางการแพทย์ของ CE และ RE ในหัวข้อถัดไป

ตารางที่ 3.1 ผลของการตัดคำด้วยวิธี LM

| Segmentation of $P_i$ | Token Elements | |
|---|---|---|
| | $t_{i,0}$ | $t_{i,1}$ |
| $T_0$ | ปวดท้อง | น้อย |
| $T_1$ | ปวด | ฝีเย็บ |

ตารางที่ 3.2 ผลการหา CE และ RE

| case | intersect | Thai word | conflict element(CE) | remaining element(RE) |
|---|---|---|---|---|
| **1** | $t_{0,0} \cap t_{1,0}$ | ปวดท้อง ⋂ ปวด | ปวด | ท้อง |
| **2** | $t_{0,0} \cap t_{1,1}$ | ปวดท้อง ⋂ ฝีเย็บ | Ø | ปวดท้อง, ฝีเย็บ |
| **3** | $t_{0,1} \cap t_{1,0}$ | น้อย ⋂ ปวด | Ø | น้อย, ปวด |
| **4** | $t_{0,1} \cap t_{1,1}$ | น้อย ⋂ ฝีเย็บ | Ø | น้อย, ฝีเย็บ |

### 3.4.4 การตรวจสอบความเกี่ยวข้องทางการแพทย์ (Medical Context Checking)

ในขั้นตอนที่ผ่านมาได้เสนอขั้นตอนเพื่อแก้ปัญหาความแตกต่างทางโครงสร้างของข้อความ โดยการหา CE แต่ยังไม่สามารถมั่นใจได้ว่าองค์ประกอบเหล่านั้นอยู่ในบริบททางการแพทย์หรือไม่ เช่น คำว่า “อาการ” และ “อาหาร” หากทำการอินเตอร์เซคจะได้ “อา” เป็น conflict element และ “การ” และ “หาร” เป็น RE โดยจะเห็นได้อย่างชัดเจนว่า “อา” เป็นคำที่

หมายถึงน้องของพ่อ หรือคำว่า “หาร” ที่หมายถึงการหาจำนวนนับของการแบ่งจำนวนออกเป็นขนาดเท่าๆกันนั้นไม่มีความเกี่ยวข้องกับการแพทย์จึงไม่สมควรจะแยกส่วนย่อยเพิ่มขึ้นให้เกิดความซับซ้อนโดยไม่มีความจำเป็น อย่างไรก็ตามการตรวจสอบคำศัพท์ถึงความเกี่ยวข้องในภาษาไทยไม่สามารถทำได้ง่าย เนื่องจากพจนานุกรมทางการแพทย์ภาษาไทยยังไม่ได้รับความสนใจทำให้ยังไม่มีพจนานุกรมทางการแพทย์ภาษาไทยพร้อมสำหรับการใช้งานในขณะทำการวิจัยนี้ ทำให้จำเป็นต้องอาศัยเครื่องมือหรือกระบวนการอื่นในการตรวจสอบความเกี่ยวข้องเช่น การใช้พจนานุกรมทางการแพทย์ภาษาอังกฤษ ถึงแม้การจะหาความสัมพันธ์ทางแพทย์จากพจนานุกรมทางการแพทย์ภาษาอังกฤษจะทำได้ยากกว่าการใช้พจนานุกรมภาษาไทย แต่ก็สามารถทำได้ โดยอาศัยวิธีการเปรียบเทียบชุดคำแปลภาษาอังกฤษของ CE และ RE (CE ∪ RE ⟶ Conflict and Remaining tokens : CRTs) ว่ามีคำแปลใดของแต่ละคำศัพท์ตรงกับพจนานุกรมทางการแพทย์ภาษาอังกฤษ (English Medical dictionary:Medict) หรือไม่ หากพบว่าอยู่ใน Medict ก็จะถือว่าสามารถยอมรับกรณี conflict นั้นๆว่าอยู่ในบริบททางการแพทย์จริงและยอมรับกรณี conflict นั้นโดยสามารถเขียนกระบวนการตรวจสอบความเกี่ยวข้องของคำศัพท์ทางการแพทย์ให้อยู่ในรูปทั่วไปได้ดังนี้

- $CRTs_i$ = $\{ t \mid t$ คือชุดของ conflict และ remaining tokens $\}$
- $E(t_i)$ = $\{ e \mid e$ เป็นชุดของคำแปลภาษาอังกฤษสำหรับ $t_i \}$
- $M$ = $\{ m \mid m$ เป็นชุดของคำศัพท์ทางการแพทย์ภาษาอังกฤษ $\}$
- $if(True)$ = $\prod_{i=0}^{n} E\,(t_i) \cap M$

เมื่อสามารถยืนยันได้ว่า conflict กรณีนั้นๆเป็นจริง แล้วจึงทำการเปลี่ยนองค์ประกอบเดิมด้วย conflict กรณีนั้นๆแทน เช่นตัวอย่างในตารางที่ 3.2 ในกรณีที่ 1 เกิด conflict ระหว่างคำว่า “ปวดท้อง” และ “ปวด” เกิด conflict ที่คำว่า “ปวด” สามารถเขียนด้วยการแทนค่าได้ดังต่อไปนี้

- $CRTs_1$ = {“ปวด”, “ท้อง”}
- $E(t_0 \mid$ "ปวด"$)$ = { “ache” , ”pain”}
- $E(t_1 \mid$ "ท้อง"$)$ = { “abdomen” , “stomach” , “belly”}
- $M$ = {“abdomen”, …, “ache”, …, “belly”, …,“pain”, …, “stomach”}
- $True$ = $E(t_0) \cap M \cdot E(t_1) \cap M \rightarrow$ accepted

ข้อความใดที่ประกอบด้วยคำว่า “ปวดท้อง” ก็จะถูกแทนที่ด้วย “ปวด” และ “ท้อง” แทนข้อความเดิม เพื่อให้สอดคล้องกับการหาความขัดแย้งกันของข้อความจากหัวข้อ 3.4.3

## 3.5 การเตรียมชุดข้อมูล CC

สำหรับการเตรียมข้อมูล CC จำเป็นต้องทำให้ชุดข้อมูลมีความสอดคล้องกับชุดข้อมูล ICD เพื่อให้อยู่ในมาตรฐานเดียวและสามารถนำไปใช้ในการหาความสัมพันธ์หรือความใกล้เคียงสำหรับการจำแนกอาการและอาการแสดงตามรหัส ICD อย่างไรก็ตามการเตรียมข้อมูล CC มีความซับซ้อนน้อยกว่า ICD มาก เนื่องจากสามารถอ้างอิงและใช้ข้อมูลจากชุดข้อมูล ICD ที่ผ่านการประมวลผลมาแล้วโดยเพียงแต่ทำการปรับข้อมูลให้มีโครงสร้างในรูปแบบเดียวกันโดยประยุกต์การตัดคำแบบLM เข้ามาช่วย

### 3.5.1 การตัดคำแบบสองระดับ (Two-Level Tokenization : 2LT)

กระบวนการเตรียมชุดข้อมูล CC ทำได้ด้วยการอาศัยวิธีการตัดคำแบบ LM เข้ามาช่วยโดยทำการตัดคำที่ยาวที่สุดด้วยวิธี LM และใช้คำศัพท์จากชุดข้อมูล ICD ก่อนหากไม่พบก็จะตัดคำโดยอาศัยจากพจนานุกรมภาษาไทยต่อไป โดยสามารถให้อยู่ในรูปทั่วไปได้ดังต่อไปนี้

เมื่อกำหนดให้

ICDtokens คือ หน่วยคำ (token) ของข้อความ ICD ที่ผ่านกระบวนการ Conflict Element finding และ Medical checking โดยเรียงลำดับตามขนาด token จากยาวไปยังสั้นที่สุด

DictTH คือ พจนานุกรมภาษาไทย โดยเรียงลำดับตามขนาดของคำศัพท์ (word) จากยาวไปยังสั้นที่สุด

CClist คือ ชุดข้อความ CC ที่ยังไม่ผ่านการเตรียมข้อมูล

รายละเอียดของขั้นตอนแสดงในฟังก์ชันรหัสเทียม 2_Level_Tokenz (2LTz) โดยมีกระบวนการหลักดังต่อไปนี้

- การลำดับการตัดคำ เป็นให้ความสำคัญของแหล่งคำศัพท์ที่ใช้ในกระบวนการตัดคำโดยลำดับการตัดคำถูกควบคุมด้วยคำสั่งการทำซ้ำ แสดงในบรรทัดที่ 7 ของฟังก์ชัน 2LTz โดยเริ่มจากตัวแปร *ICDtokens* และตามด้วย *DictTH*
- การตัดคำ การตัดคำทำด้วยวิธี LM และถูกแยกเป็น 2 หน่วยย่อยดังนี้
  - ไม่พบคำตรงกับ 2 แหล่งคำ หากไม่พบคำในข้อความ CC จากตำแหน่งปัจจุบันตรงกับคำศัพท์ใน *ICDtokens* หรือ *DictTH* ก็จะทำการตัดคำตัวอักษรตำแหน่งปัจจุบันเก็บไว้ในตัวแปร *unknown* แสดง

ในคำสั่งบรรทัดที่ 15 และ 16

○พบคำตรงกับแหล่งคำ ในกรณีที่ ตัวแปร *unknown* เป็นค่าว่าง จะทำการตัดคำที่พบเก็บลงในตัวแปร *CCset* ซึ่งเป็นตัวแปรสำหรับเก็บชุดข้อความ CC ที่ผ่านการตัดคำแล้ว แสดงในคำสั่งบรรทัดที่ 12 และ 14

ในกรณีที่ ตัวแปร *unknown* ไม่เป็นค่าว่าง จะทำการเพิ่มชุดตัวอักษรจากตัวแปร *unknown* และตัดคำที่พบเก็บลงในตัวแปร *CCset* ซึ่งเป็นตัวแปรสำหรับเก็บชุดข้อความ CC ที่ผ่านการตัดคำแล้ว แสดงในคำสั่งบรรทัดที่ 9 และ 11

โดยรายละเอียดในของฟังก์ชันรหัสเทียม 2LTz ได้ทำการอธิบายต่อจากชุดคำสั่งรหัสเทียมด้านล่าง

2_Level_Tokenz(*CClist*, *ICDtokens, DictTH*):

---

1. set *CCset* = [NULL*size_of(*CClist*)]
2. set *Dict_set* = [*ICDtokens* = sort_desc(*ICDtokens*), *dictTH* = sort_desc(*DictTH*)]
3. set *unknown* = "", *segm_check* = *false*, *i* = 0
4. For ∀ *text* in *CClist*:
5. While *text* != "":
6. set *segm_check* = *true*
7. for *TokenSet* in *Dict_set*:
8. for *token* in *TokenSet* ∈ *text* :
9. if (*token* == *text*[:length(*token*)]) AND (*unknown* != "" AND !*segm_check*) :
10. remove *text*[:length(*token*)] and add *unknown* and *token* to *CCset*[*i*]
11. set *segm_check* = *false* and *unknown* = ""
12. else if (*token* == *text*[:length(*token*)]) AND (!*segm_check*) :
13. remove *text*[:length(*token*)] and add *token* to *CCset*[*i*]
14. *segm_check* = *false*
15. if (*segm_check*) :
16. add *text*[*0*] to *unknown* and remove *text*[*0*] from *text*
17. if (*unknown* != ""):
18. add *unknown* to *CCset*[*i*]

19. *i = i +1*

20. return *CCset*

ฟังก์ชัน 2LTz รับข้อมูลเข้าประกอบด้วย *CClist* *ICDtokens* และ *DictTH* ซึ่งเป็นชุดข้อความ CC องค์ประกอบชุดข้อมูล ICD และพจนานุกรมภาษาไทยตามลำดับ เมื่อข้อมูลเข้าผ่านกระบวนการภายในฟังก์ชัน 2LTz ก็จะสามารถทำให้มีโครงสร้างเดียวกับชุดข้อมู ICD เพื่อสามารถหาความเชื่อมโยงความสัมพันธ์ระหว่างสองชุดข้อมูลได้ โดยบรรทัดที่ 1 ถึง 3 ของฟังก์ชัน 2LTz เป็นการกำหนดค่าเริ่มให้กับตัวแปร ได้แก่ การกำหนดขนาดของตัวแปร *CCset* ที่ใช้สำหรับการเก็บผลลัพธ์การตัดคำ การกำหนดให้ตัวแปร *Dict_set* เก็บคำศัพท์ *ICDtokens* และ *DictTH* โดยเรียงตามขนาดข้อความแบบมากไปหาน้อย กำหนดตัวแปร *unknown* เพื่อเก็บชุดอักขระที่ไม่พบในคำศัพท์จากตัวแปร *Dict_set* กำหนดตัวแปร *segm_check* สำหรับระบุว่ามีการตัดคำการคำศัพท์ภายใน *Dict_set* หรือไม่ และ กำหนดตัวแปร *i* สำหรับระบุตำแหน่งปัจจุบันที่ต้องการเก็บผลการตัดคำลงในตัวแปร *CCset* บรรทัดที่ 4 เป็นการเริ่มต้นการทำซ้ำของการตัดคำของข้อความ CC และ บรรทัดที่ 6 เป็นการกำหนดค่าสำหรับตรวจว่าข้อความในตัวแปร *text* ที่ต้องการตัดคำทุกตัดคำออกหมดแล้วหรือไม่ บรรทัดที่ 7 ถึง 16 เป็นทำซ้ำกระบวนการตัดคำตามลำดับของชุดคำศัพท์ในตัวแปร *Dict_set* คือเริ่มจากคำศัพท์ของ *ICDtokens* และ ตามด้วย *dictTH* หากพบคำศัพท์ตรงกับตำแหน่งซ้ายสุดของ *text* ก็จะทำการตัดคำที่พบเพิ่มลงในตัวแปร *CCset* ตำแหน่ง *i* หากไม่พบคำที่ตรงกับชุดตัวอักษรทางขวาสุดของตัวแปร *text* จะทำการตัดตัวอักษรตำแหน่งซ้ายสุดของตำแหน่งปัจจุบันลงในตัวแปร *unknown* สำหรับเก็บชุดตัวอักษรที่ไม่พบในชุดคำศัพท์ ในกรณีเกิดเหตุการณ์ตามบรรทัดที่ 9 คือพบคำศัพท์ที่สามารถตัดได้ แต่ตัวแปร *unknown* ไม่ใช่ค่าว่างหรือพบชุดตัวอักษรที่ไม่มีอยู่ในชุดคำศัพท์ ให้ทำการเพิ่มชุดตัวอักษรลงในตัวแปร *CCset* ก่อน จากนั้นจึงทำการเพิ่มคำศัพท์ที่พบลงใน *CCset* ต่อไป เมื่อตัวอักษรใน *text* ถูกตัดจนหมด แต่ยังมีชุดตัวอักษรอยู่ในตัวแปร *unknown* ก็จะทำการเพิ่มชุดตัวอักษรดังกล่าวลงใน *CCset* ตำแหน่ง i ก่อนจะทำการเพิ่มค่า *i* เป็น *i+1* และเริ่มทำการตัด *text* ลำดับถัดไป เมื่อทำการตัด *text* ครบจากตัวแปร *CClist* ก็จะทำการ return ค่าของตัวแปร *CCset* ซึ่งเป็นผลลัพธ์การตัดคำที่มีโครงสร้างเดียวกับชุดข้อมูล ICD เพื่อใช้สำหรับทำการทดลองต่อไป

## 3.6 การเพิ่มประสิทธิภาพชุดข้อมูล (Corpus Enhancement)

ในกระบวนการที่ผ่านมาเราเสนอวิธีหาความเป็นไปได้ที่องค์ประกอบของข้อความจะมีความกำกวมเกิดขึ้น และทำการแยกส่วนองค์ประกอบใดที่ทำให้เกิดความกำกวม

เหล่านั้น แต่ในขณะเดียวกันไม่มีอะไรยืนยันได้เลยว่าองค์ประกอบเหล่านั้น จำเป็นต้องแยกองค์ประกอบออกเป็นหน่วยย่อยหรือไม่ เช่น คำว่า “ปวดท้องน้อย” เมื่อผ่านกระบวนการต่างๆจะได้ผลลัพธ์ดังนี้

- การตัดข้อความด้วยวิธี LM : “ปวดท้อง”, “น้อย”
- การตัดข้อความด้วยวิธี 2LT : “ปวด”, “ท้อง”, “น้อย”

เมื่อทำการสังเกตจะพบว่าคำว่า “ท้อง” และ “น้อย” เป็นชื่อเฉพาะที่ควรจะอยู่ติดกันเป็นคำว่า “ท้องน้อย” เสมอแต่ถูกตัดให้พิจารณาแยกว่าเป็นคนละคำนั้นจะสามารถก่อให้เกิดปัญหาความกำกวมได้ เช่น

“ปวดท้อง” ⟶ “ปวด” “ท้อง”
“ปวดท้องน้อย” ⟶ “ปวด” “ท้อง” “น้อย”
⟶ “ปวด” “ท้องน้อย”

จะเห็นได้ว่าหากกำหนดให้ “ท้องน้อย" แยกกันเป็นคนละคำ “ท้อง" และ “น้อย" จะทำให้มีโครงสร้างของคำที่ใกล้เคียงกันกับ “ปวด” “ท้อง” มากกว่าทำให้ยากต่อการระบุหรือยืนยันว่าหมายถึง “ปวดท้อง” หรือ “ปวดท้องน้อย” แต่หากรวม “ท้องน้อย” เป็นคำเดียวกันจะทำให้สามารถจำแนกว่าแตกต่างได้อย่างชัดเจนยิ่งขึ้น แต่ก็มีปัญหาอยู่ที่ว่าจะสามารถทำอย่างไรให้สามารถระบุได้อย่างชัดเจนว่าชุดคำใดควรจะพิจารณาเป็นคำเดียว

### 3.6.1 การแก้ปัญหาความกำกวมด้วยวิธีทางสถิติ

วิธีการที่นำมาใช้เป็นการประยุกต์ใช้ N-gram สำหรับหาความเป็นไปได้ที่ชุดคำศัพท์ชุดหนึ่งจะเกิดขึ้น โดยอาศัยจำนวนนับของเหตุการณ์ที่เกิดขึ้นของชุดข้อมูล CC เนื่องจากข้อมูลในชุดข้อมูล ICD เป็นชุดข้อมูลมาตรฐานซึ่งเป็นไปได้ยากที่จะมีส่วนขยายที่ไม่จำเป็น โดยปรับใช้กับ N-gram ที่มีขนาด N เท่ากับ 2 หรือเรียกว่า Bi-gram มีวิธีการคำนวณคำนวณดังนี้

$P(w_{i-1}, w_i)$ คือ ความเป็นไปได้ที่ $w_i$ จะอยู่หลัง $w_{i-1}$ เมื่อ $w$ คือคำศัพท์ในข้อความและ $i$ คือตำแหน่งของข้อความในประโยค

$C(w_{i-1}, w_i)$ คือ จำนวนนับของเหตุการณ์ที่ $w_i$ อยู่หลัง $w_{i-1}$ ที่เกิดขึ้นทั้งหมด

$N(w_{i-1})$ คือ จำนวนนับของ $w_{i-1}$ที่เกิดขึ้นทั้งหมด

$$P(w_{i-1}, w_i) = \frac{C(w_{i-1}, w_i)}{N(w_{i-1})}$$

จากสมการที่ใช้สำหรับการคำนวณ Bi-gram จำเป็นต้องทำการปรับให้สอดคล้องกับการหาความเป็นไปได้ของ token ที่เป็นองค์ประกอบของข้อความ 2 tokens ที่อยู่ติดกันเสมอใน ICD รหัสเดียวกัน โดยสามารถคำนวณวิธีการหาความเป็นไปได้ดังนี้

เมื่อกำหนดให้

- $CCset$ = { CCp | CCp คือชุดข้อความ CC ที่ผ่านกระบวนการเตรียมด้วย 2LT }
- $CCp$ = { t | t คือองค์ประกอบของข้อความ CCp}
- $ICDcode$ = คือรหัส ICD
- $P(t_{i-1}, t_i, ICDcode)$ = ค่าความเป็นได้ที่ $t_{i-1}$ จะตามด้วย $t_i$ ของรหัส $ICDcode$
- $C(t_{i-1}, t_i, ICDcode)$ = จำนวนนับของ $t_{i-1}$ ตามด้วย $t_i$ ที่มีรหัส ICD เดียวกัน
- $N(t_{i-1}, t_i, ICDcode)$ = จำนวนนับทั้งหมดของ $CCp$ ที่มีรหัส $ICDcode$ และประกอบด้วย $t_{i-1}$ และ $t_i$
- $P(t_i,t_{i+1},ICDcode)$ = $\frac{C(t_i,t_{i+1},ICDcode)}{N(t_i,t_{i+1},ICDcode)}$

การคำนวณหาความเป็นไปได้ของ $t_{i-1}$ ตามด้วย $t_i$ ที่เป็นเป็นสมาชิกของข้อความที่มีรหัส ICD เดียวกัน หากผลการหาความน่าจะเป็นของสมการ $P(t_{i-1}, t_i, ICDcode)$ มีค่าเข้าใกล้ 1.0 เท่าไร ก็หมายความว่าโอกาสที่ $t_{i-1}$ จะตามด้วย $t_i$ มีมากขึ้นเท่านั้น และเมื่อค่าความน่าจะเป็นมีค่าเท่ากัน 1.0 หรือ ผ่านเกณฑ์ขั้นต่ำที่ได้กำหนด จึงทำการทำการแทนที่ค่า $t_{i-1}$ และ $t_i$ ด้วย $t_{i-1} + t_i$ ทั้งในชุดข้อมูล ICD และชุดข้อมูล CC เพื่อลดความกำกวมสำหรับข้อความรหัส ICD อื่นที่ $t_{i-1}$ และ $t_i$ ไม่ได้อยู่ติดกันเสมอ

# บทที่ 4

# ขั้นตอนการทดลองจำแนกอาการและอาการแสดงตามมาตรฐานรหัส ICD

ในการทดลองการจำแนกอาการและอาการแสดง ประกอบไปด้วย 3 ส่วนย่อยคือ ส่วนของชุดข้อมูลการสอนเครื่องและชุดข้อมูลทดสอบ ซึ่งเป็นส่วนที่อธิบายถึงลักษณะของข้อมูลที่นำมาใช้ในการสอนและทดสอบ ส่วนของการเตรียมข้อมูลสำหรับการทดสอบ อธิบายถึงการนำกระบวนการเตรียมข้อมูลที่ออกแบบมาประยุกต์ใช้ และ ส่วนของการเปรียบเทียบประสิทธิภาพ อธิบายถึงการเปรียบเทียบผลลัพธ์ของการนำชุดข้อมูลไปจำแนกออกเป็นอาการ โดยมีการแจกแจงประเภทของการเตรียมข้อมูลและวิธีการจำแนก

## 4.1 ข้อมูลชุดสอนเครื่องและชุดทดสอบ

การทดลองการจำแนกอาการและอาการแสดง แบ่งข้อมูลออกเป็นสองชุด คือ

1. ข้อมูลชุดสอน คือ ข้อความพื้นฐานจาก ICD
2. ข้อมูลชุดทดสอบ คือ ข้อความบอกเล่าอาการสำคัญ

โดยข้อความ ICD เป็นข้อความมาตรฐานที่ใช้การสอนเครื่องเพื่อการจำแนกอาการและข้อความบอกเล่าอาการสำคัญที่มีส่วนขยายหรือมีองค์ประกอบของอาการภายในข้อความมากกว่า 1 อาการ เป็นต้น

### 4.1.1 การเลือกชุดข้อมูลสำหรับการทดลอง

ในข้อความประเภทที่หนึ่งหรือข้อความพื้นฐานของรหัส ICD ถูกเลือกมาจากเอกสาร ICD-10-TM หมวด R00-R69 อาการ อาการแสดง และความผิดปกติที่พบจากการตรวจทางคลินิก โดยเลือกจากอาการที่เป็นข้อความปลายปิดหรือข้อความที่มีการระบุความผิดปกติอย่างชัดเจนไม่สามารถตีความเป็นความหมายอื่นได้ เช่น “R07.4 เจ็บหน้าอกเวลาหายใจ” เป็นการบอกเล่าว่ามีอาการเจ็บเมื่อทำการหายใจ ถึงจะมีการเขียนในรูปแบบอื่นเช่น “ปวดหน้าอกเวลาหายใจ” หรือ “เวลาหายใจเจ็บหน้าอก” ก็ยังสามารถให้ความหมายเดียวกันได้

ข้อความปลายเปิด หรือ ข้อความที่สื่อความหมายไม่เจาะจง ทำให้ไม่สามารถใช้ข้อความดังกล่าวเป็นพื้นฐานในการเป็นข้อความชุดสอนได้ เนื่องจากยากต่อการหาความสัมพันธ์ต่อข้อความบอกเล่าอาการสำคัญได้ เช่น “R20.3 ความรู้สึกที่ผิวหนังผิดเพี้ยน” คำว่า “ผิดเพี้ยน” สามารถสื่อได้หลายความหมายที่ตรงข้ามกับ “ความปกติ” ที่มีต่อ “ผิวหนัง” เช่น “ความรู้สึกไวกว่าปกติ” “รู้สึกเจ็บแปลๆ” “เสียว" หรือ “ชาตามผิวหนัง” ทั้งหมดถือเป็น “ความผิดเพี้ยน” ที่เกิดขึ้นต่อ “ผิวหนัง” มีความหมายแตกต่างกันแต่เป็นกลุ่มความหมายเดียวกันบนพื้นฐานของข้อความปลายเปิด

ข้อมูลที่ใช้ในการทดลองประกอบไปด้วยข้อความที่เลือกจากเอกสาร ICD-10-TM และข้อความที่ถูกเพิ่มเติมจากคำพ้อง มีทั้งหมด 82 ข้อความ โดยแต่ละข้อความติดป้าย 1 รหัส และข้อความบอกเล่าอาการสำคัญที่เขียนโดยผู้วิจัยตามรูปแบบการเขียนข้อความ CC จำนวน 37 ข้อความ และ ข้อความ CC จากอินเทอร์เน็ตจำนวน 50 ข้อความ [40-62] แต่ละข้อความติดป้ายรหัสจำนวน 1 ถึง 3 รหัส ในการทดลอง

## 4.2 การเตรียมข้อมูลสำหรับการทดสอบ

เพื่อที่จะเปรียบเทียบข้อมูลที่ถูกใช้ในการทดสอบ การเตรียมข้อมูลจะแบ่งข้อมูลออกเป็นสามชุด ตามประเภทการเตรียมข้อมูล เพื่อสามารถเปรียบความแตกต่างของการเตรียมข้อมูลได้ ประกอบด้วยข้อมูลดังนี้

1. ข้อมูลที่ไม่ผ่านการเตรียมข้อมูล Non-segmented Text
2. ข้อมูลที่ถูกตัดคำด้วยวิธี LM-based Text
3. ข้อมูลที่ถูกตัดด้วยวิธี 2LT-based Text

ข้อมูลแต่ละชุดจะแบ่งข้อมูลออกเป็นสองส่วนคือ

1. ข้อมูลชุดสอน ข้อมูลชุดสอนจะประกอบด้วย ข้อความจากรหัส ICD ทั้งหมด 82 ข้อความ แต่ละข้อความถูกติดป้ายด้วยรหัส ICD จำนวน 1 ป้าย
2. ข้อมูลชุดทดสอบ ข้อมูลชุดทดสอบประกอบไปด้วยข้อความบอกเล่าอาการสำคัญทั้งหมด 87 ข้อความ แต่ละข้อความถูกติดด้วยป้ายด้วยรหัส ICD จำนวน 1 ถึง 3 ป้าย

**การตัดคำภาษาไทยด้วยวิธี LM**

การตัดข้อความด้วยวิธี LM โดยใช้โปรแกรมรหัสเปิดซึ่งพัฒนาโดย NECTEC ด้วยภาษา JAVA มีชื่อว่า LongLexTo โดยอาศัยพจนานุกรม LEXiTRON ที่ประกอบไปด้วยคำศัพท์จำนวน 42,221 คำศัพท์ซึ่งถูกจัดทำโดย NECTEC เช่นเดียวกันในการตัดข้อความ

**การตัดคำภาษาไทยแบบ 2LT**

การตัดข้อความด้วยวิธี 2LT เป็นการเตรียมชุดข้อมูลของ CC ซึ่งอาศัยแหล่งคำศัพท์ 2 แหล่งในการตัดคำ คือ

1. ชุดคำศัพท์จากชุดข้อมูล ICD เรียกว่า ICD10token
2. คำศัพท์จากพจนานุกรม LEXiTRON

โดยการตัดคำในข้อความ CC จะใช้วิธีการตัดแบบ LM ด้วยโปรแกรม LongLexTo ในการตัด แต่จะทำการประยุกต์ขั้นตอนการตัดคำแบบ 2LT ตามกระบวนการที่ได้เสนอไว้ในหัวข้อ 3.5.1 คือ การตัดคำจะพิจารณาจาก ICD10token ก่อนเมื่อไม่พบจึงดำเนินการต่อด้วยคำศัพท์จาก LEXiTRON อย่างไรก็ตามจำเป็นต้องมีชุดคำศัพท์จากชุดข้อมูล ICD ที่ผ่านขั้นตอนการเตรียมข้อมูล ICD ที่เสนอในหัวข้อที่ 3.4 ก่อน จึงจะสามารถทำการปรับใช้การตัดคำแบบ 2LT ได้ โดยการเตรียมข้อมูล ICD-10 ได้อธิบายรายละเอียดในหัวข้อ 4.2.1 และอาศัย ICD10token จากการทดลองเตรียมข้อมูลมาใช้ในกระบวนการตัดคำแบบ 2LT เพื่อเตรียมชุดข้อมูล CC ต่อไป

งานในอนาคตจะทำการทดลองกระบวนการ Corpus Enhancement ที่ไม่ได้ถูกนำมาใช้ในการทดลองนี้ เนื่องจากปริมาณข้อความ CC มีปริมาณไม่มากพอที่จะสามารถนำวิธีการทางสถิติมาใช้ได้อย่างมีประสิทธิภาพ เพื่อพิสูจน์ผลการที่ได้นำเสนอว่าได้ผลลัพธ์ตามที่คาดหวังหรือไม่

**4.2.1 การเตรียมชุดข้อมูล ICD**

ในหัวข้อนี้กล่าวถึงการเตรียมข้อมูลข้อความ ICD ประกอบด้วย Conflict Element Finding และ Medical Context Checking เพื่อนำไปใช้ในกระบวนการ 2LT ต่อไป

**4.2.1.1 Conflict Element Finding**

กระบวนการ Conflict Element finding ถูกนำไปเก็บด้วยฐานข้อมูล MySQL โดยอาศัยข้อความ ICD-10 และพจนานุกรม Lexitron ในการทดลอง ก่อนนำข้อมูลเข้าสู่ฐานข้อมูลจะทำการตัดคำข้อความ ICD ที่มีการติดป้ายรหัส ICD-10 ด้วยวิธี LM โดยอาศัย LongLexTo หลังจากทำ

การตัดข้อความเรียบร้อยจึง Import ข้อมูล ICD และข้อมูลพจนานุกรม Lexitron เข้าสู่ฐานข้อมูล MySQL ต่อไป โดยกำหนดให้ตารางที่เก็บข้อมูล ICD มีชื่อว่า ICD10token และตารางที่เก็บข้อมูลพจนานุกรมมีชื่อว่า Lexitron โครงสร้างตามตารางที่ 4.1 และ 4.2 ตามลำดับเมื่อกำหนดให้ชื่อคอลัมน์มีความหมายดังต่อไปนี้

- ตาราง ICD10token
  - ICD_ID คือ รหัส ICD-10
  - cases คือ ลำดับของข้อความ
  - orders คือ ลำดับของคำศัพท์ที่ประกอบในข้อความ
  - Token คือ คำศัพท์ที่ประกอบในประโยคตามลำดับของ cases และ orders
- ตาราง Lexitron
  - word คือ คำศัพท์ในพจนานุกรม Lexitron

ตารางที่ 4.1 ICD10token

| ICD_ID | Cases | orders | Token |
|---|---|---|---|
| R00.0 | 0 | 0 | หัวใจ |
| R00.0 | 0 | 1 | เต้น |
| R00.0 | 0 | 2 | เร็ว |
| R00.0 | 1 | 0 | ใจเต้น |
| R00.0 | 1 | 1 | เร็ว |

ตารางที่ 4.2 Lexitron

| word |
|---|
| กงสี |
| กงสุล |
| กงสุลใหญ่ |
| กงเกวียน |
| กงเต็ก |

การหาองค์ประกอบที่ขัดแย้ง (Conflict Element : CE) คือการหาองค์ประกอบที่เหมือนกันของคำศัพท์ที่ต่างกัน และ CE ต้องเป็นชุดตัวอักษรที่มีความหมายเท่านั้น ซึ่งหมายความว่าทุก CE ที่เป็นไปได้ต้องเป็นคำศัพท์ที่อยู่ในพจนานุกรม ดังนั้นทุกคำศัพท์ในตาราง Lexitron ที่เป็นองค์ประกอบของคอลัมน์ Token ในตาราง ICD10token จึงเป็น CE ที่เป็นไปได้ทั้งหมด จากเหตุผลดังกล่าวทำให้สามารถหาชุด CE ที่เป็นไปได้ด้วยการเชื่อมตาราง ICD10token เข้ากับตารางพจนานุกรม Lexitron เมื่อคอลัมน์ word ที่บรรจุคำศัพท์ในพจนานุกรมของ Lexitron เป็น

องค์ประกอบในคอลัมน์ token ของ ICD10token ตามตารางที่ 4.3 ซึ่งเป็นผลลัพธ์ของการรวมระหว่างตาราง ICD10token และ Lexitron

ตารางที่ 4.3 TmpTable

| *ICD_ID* | *cases* | *orders* | *Token* | *word* |
|---|---|---|---|---|
| R10.2 | 0 | 2 | เชิงกราน | กร |
| R10.2 | 1 | 2 | เชิงกราน | กร |
| R25.2 | 0 | 4 | เกร็ง | กร |
| R25.2 | 2 | 2 | เกร็ง | กร |
| R25.2 | 3 | 1 | เกร็ง | กร |

การยืนยัน CE สามารถทำได้โดยการเชื่อมตารางที่ 4.3 จำนวน 2 ตารางเข้าด้วยกัน เพื่อให้ได้คอลัมน์ Token ที่มีองค์ประกอบจากคอลัมน์ word เดียวกันตามตารางที่ 4.4 เมื่อกำหนดให้คอลัมน์ word คือ CE และคอลัมน์ remaining หรือ องค์ประกอบที่เหลือ (Remaining Element : RE) เป็นผลของการตัดอักขระของคอลัมน์ Token ด้วย CE ซึ่งคอลัมน์ remaining เก็บข้อมูลเป็นรูปแบบโครงสร้าง ใช้สัญลักษณ์ “,” ในการแบ่งชุดอักขระ RE สัญลักษณ์ “:” สำหรับการแบ่งชุด RE ของ Token เมื่อชุด RE ที่อยู่ด้านหน้า “:” คือ RE ของตาราง TmpTable #1 และชุด RE ที่อยู่ด้านหลังเป็น RE ของตาราง TmpTable #2 อย่างไรก็ตามถึงแม้ CE เป็นชุดอักขระที่มีความหมายแน่นอนเนื่องจากทุกคำมาจากคำศัพท์ของพจนานุกรม Lexitron แต่ชุดอักขระ RE ไม่ได้มาจากพจนานุกรม ดังนั้นจึงทำการตรวจสอบด้วยวิธีการเปรียบเทียบอักขระว่า RE เป็นชุดอักขระที่มีอยู่ในพจนานุกรม Lexitron หรือไม่ หลังจากสามารถหา CE และ RE ได้และยืนยันความหมายด้วยพจนานุกรม Lexitron แล้วจะนำไปตรวจสอบความสัมพันธ์ทางการแพทย์ของ CE และ RE ในหัวข้อถัดไป

ตารางที่ 4.4 ผลลัพธ์จากการเชื่อมตาราง TmpTable 2 ตาราง

| TmpTable #1 | | | | join | TmpTable #2 | | | | *remaining* |
|---|---|---|---|---|---|---|---|---|---|
| *ICD_ID* | *cases* | *orders* | *Token* | *word* | *ICD_ID* | *cases* | *orders* | *Token* | |
| R25.2 | 0 | 4 | เกร็ง | กร | R10.2 | 1 | 2 | เชิงกราน | เ,็ง:เชิง,าน |
| R25.2 | 2 | 2 | เกร็ง | กร | R10.2 | 1 | 2 | เชิงกราน | เ,็ง:เชิง,าน |
| R25.2 | 3 | 1 | เกร็ง | กร | R10.2 | 1 | 2 | เชิงกราน | เ,็ง:เชิง,าน |
| R10.2 | 0 | 2 | เชิงกราน | กร | R25.2 | 0 | 4 | เกร็ง | เชิง,าน:เ,็ง |
| R32 | 0 | 0 | กล้ัน | กล | R25.2 | 0 | 2 | กล้ามเนื้อ | ,ั้น:,้ามเนื้อ |

**4.2.1.2 Medical Context Checking**

กระบวนการตรวจสอบความเกี่ยวข้องทางการแพทย์ของ CE และ RE ตามที่ได้ออกแบบในหัวข้อ 3.4.4 คือหาชุดคำศัพท์ภาษาอังกฤษของ CE และ RE และเปรียบเทียบกับคำศัพท์จากพจนานุกรมทางการแพทย์ภาษาอังกฤษที่พร้อมใช้งานในปัจจุบันว่าตรงกันหรือไม่ โดยชุดคำศัพท์ภาษาอังกฤษสำหรับ CE และ RE มาจากพจนานุกรมทั่วไปภาษาไทยและอังกฤษ YAiTRON ที่พัฒนามาจากพจนานุกรม LEXiTRON จับคู่กับคำศัพท์จากพจนานุกรมทางการแพทย์ภาษาอังกฤษจากมาตรฐานของ e-MedTools and Raj&Co [63] สำหรับการหาความสัมพันธ์ทางการแพทย์ ตัวอย่างของผลลัพธ์การหาความสัมพันธ์ทางการแพทย์แสดงในตารางที่ 4.5 เมื่อคอลัมน์ Vocab_check แสดงสถานะของการตรวจสอบความหมายของ CE และ RE ว่ามีอยู่ในพจนานุกรมทั่วไปหรือไม่ และ Med_check แสดงสถานะของการตรวจสอบความเกี่ยวข้องทางการแพทย์ของ CE และ RE จากพจนานุกรมทางการแพทย์ของภาษาอังกฤษ โดยทั้งสองคอลัมน์ใช้คำว่า “accepted” สำหรับระบุการยอมรับ และ “rejected” สำหรับระบุการปฏิเสธการตรวจสอบ

ตารางที่ 4.5 ผลการหาความเกี่ยวของทางการแพทย์ของ CE และ RE

| TmpTable #1 | | | | *CE* | TmpTable #2 | | | | *RE* | Vocab_check | Med_check |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *ICD_ID* | *cases* | *orders* | *Token* | *word* | *ICD_ID* | *cases* | *orders* | *Token* | *remaining* | | |
| R56.0 | 0 | 0 | การชัก | การ | R25.1 | 0 | 0 | อาการ | ,ชัก:อา, | accepted | rejected |
| R25.1 | 0 | 0 | อาการ | การ | R56.0 | 0 | 0 | การชัก | อา,:,ชัก | accepted | rejected |
| R10.1 | 0 | 2 | ท้อง | ท้อง | R10.0 | 0 | 0 | ปวดท้อง | ,:ปวด, | accepted | accepted |
| R14 | 0 | 0 | ท้องอืด | ท้อง | R10.0 | 0 | 0 | ปวดท้อง | ,อืด:ปวด, | accepted | accepted |
| R14 | 1 | 0 | ท้อง | ท้อง | R10.0 | 0 | 0 | ปวดท้อง | ,:ปวด, | accepted | accepted |

## 4.3 การเปรียบเทียบประสิทธิภาพของวิธีการเรียนรู้และและวิธีการจำแนก

การทดลองการจำแนกอาการและอาการแสดงทำการทดลองโดยอาศัยแพ็คเกจของ sklearn ในกระบวนการทดลองจำแนกอาการและอาการแสดง ซึ่งเป็นเครื่องมือสำหรับการทำ data mining และ data analysis ด้วยวิธีการเรียนรู้ของเครื่อง โดยการเปรียบเทียบถูกแบ่งออกตามวิธีการเรียนรู้และจำแนกที่สามารถจัดการการจำแนกแบบหลายผลลัพธ์ แต่ละประเภทการจำแนกจะทำการทดสอบด้วยชุดข้อมูล 3 ชุด ตามที่กล่าวในหัวข้อ 4.2 คือ ชุดข้อมูลที่ไม่ได้ผ่านการเตรียมข้อมูล ชุดข้อมูล LM และ ชุดข้อมูล 2LT โดยชุด classifiers จากแพ็คเกจของ sklearn ที่ใช้ในการทดลองมีดังนี้

- DecisionTreeClassifier : CART Optimization

  เป็นการจำแนกด้วยวิธีจำแนกแบบโครงสร้างต้นไม้ตัดสินใจ โดยอาศัยวิธีกระบวนการ CART เป็นวิธีการสร้างโครงสร้างต้นไม้ตัดสินใจแบบต้นไม้ทวิภาค โดยอาศัยข้อมูลคุณลักษณะและเกณฑ์จากการคำนวณด้วยวิธี Information gain หรือ gini-index ในการสร้างโหนด

- K-Nearest Neighbors Classifier (KNN)

  เป็นวิธีการจำแนกกลุ่มข้อมูลแบบการจับกลุ่ม โดยจับกลุ่มข้อมูลจากข้อมูลที่ใกล้ตำแหน่ง N (Neighbors) จำนวน K

- Radius Neighbors Classifier

เป็นวิธีการจำแนกกลุ่มข้อมูลแบบการจับกลุ่ม โดยจับกลุ่มข้อมูลจากรัศมีวงกลมรอบตำแหน่ง N

- RandomForest Classifier

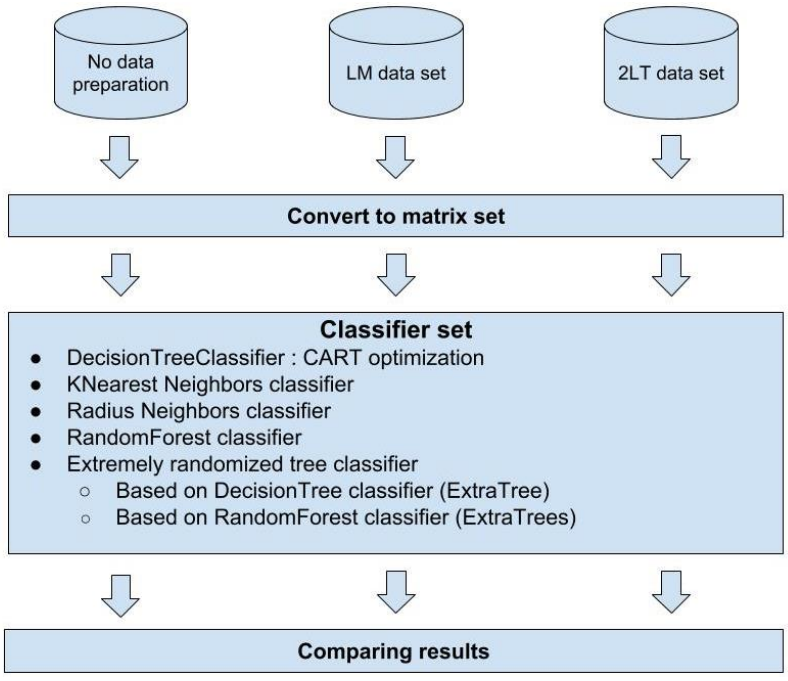
เป็นการจำแนกโดยอาศัยการสร้างต้นไม้ตัดสินใจจำนวนมากด้วยหลายข้อมูลย่อยจากข้อมูลชุดสอนทั้งหมด และใช้ค่าเฉลี่ยในการเพิ่มความแม่นยำในการทำนายผล

- Extremely Randomized Tree Classifier
  - Based on DecisionTree Classifier (ExtraTree)

เป็นการจำแนกแบบต้นไม้ตัดสินใจ โดยแตกต่างจากต้นไม้ตัดสินใจปกติที่วิธีการสร้างต้นไม้ตัดสินใจ ในขั้นตอนการแบ่งกลุ่มข้อมูลออกเป็น 2 กลุ่มเพื่อสร้างโหนดจะทำการสุ่มจุดเด่นหรือคุณลักษณะ (Feature) ที่จะนำมาพิจารณาสำหรับแบ่งกลุ่มข้อมูล

  - Based on RandomForest Classifier (ExtraTrees)

เป็นการจำแนกโดยอาศัยพื้นฐานจาก RandomForest Classifier แต่ต่างกันที่รูปแบบการสร้างโครงสร้างต้นไม้จะเป็นแบบ ExtraTree




รูปที่ 4.1 กระบวนการเปรียบเทียบผลระหว่างชุดข้อมูลทดลองและวิธีการจำแนก

ในการนำข้อความจาก ICD และ CC มาใช้ในการทดลองการจำแนกอาการด้วยวิธีการเรียนรู้ของเครื่องจำเป็นต้องมีการแปลงข้อความให้อยู่ในรูปเมตริกเพื่อให้สามารถนำไปใช้กับเครื่องมือการเรียนรู้ของเครื่องได้ โดยใช้แพ็กเกจของ sklearn ช่วยในการแปลงข้อความให้อยู่ในรูปของเมตริกซ์ด้วยวิธี Bag of Word ตามรูปแบบที่แสดงในตารางที่ 4.6 ที่เป็นรายการของจำนวนนับของคำที่เกิดขึ้นแต่ละประโยค และจัดเก็บตามรูปแบบตารางที่ 4.7 ซึ่งประกอบไปด้วย ตำแหน่งของคำในกระเป๋า และ จำนวนที่ปรากฏในข้อความ เพื่อนำไปใช้ในการเรียนรู้ของเครื่องต่อไป

ตารางที่ 4.6 ข้อมูลที่ถูกแปลงด้วยวิธีการ Bag of Word

| *ข้อความที่* | *คำในกระเป๋า (Bag of Word)* |
|---|---|
| | ['กล', 'การช', 'กำเดา', 'ขณะ', 'ขา', ... , 'ใจ', 'ใน', 'ไข', 'ได', 'ไม', 'ไหล', 'ไอ'] |
| 1 | [[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 … 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]] |
| 2 | [[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 … 0 0 0 0 0 0 0 **1** 0 0 0 0 0 0]] |
| 3 | [[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 … 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]] |
| 4 | [[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 … 0 0 0 0 0 0 **1** 0 0 0 0 0 0 0]] |

ตารางที่ 4.7 ตัวอย่างข้อมูลที่นำไปใช้ในการจำแนกด้วยชุด Classifier

| *ข้อความที่* | *ตำแหน่งในกระเป๋า* | *จำนวนครั้งที่เกิดขึ้น* |
|---|---|---|
| 20 | 15 | 1 |
| | 26 | 1 |
| | 52 | 1 |
| | 33 | 2 |
| | 18 | 1 |
| 59 | 27 | 2 |
| | 74 | 1 |
| | 40 | 1 |
| | 69 | 1 |
| 113 | 69 | 1 |
| | 33 | 2 |
| | 18 | 1 |

ตารางที่ 4.8 การเปรียบเทียบระหว่างผลการทดลองของแต่ละการเรียนรู้ของเครื่อง

| Classifier | Segmentation Method | | | | | |
|---|---|---|---|---|---|---|
| | Non-Segmentation | | Longest matching | | 2LT | |
| | Precision | Recall | Precision | Recall | Precision | Recall |
| CART | **0.59** | 0.71 | **0.88** | 0.69 | **0.88** | 0.69 |
| K-Nearest Neighbors(KNN) | 0.06 | 0.88 | 0.09 | **0.92** | 0.07 | 0.75 |
| Radius Neighbors | 0.06 | 0.88 | 0.09 | **0.92** | 0.07 | 0.75 |
| RandomForest | 0.40 | **0.78** | 0.35 | 0.85 | 0.32 | **0.78** |
| ExtraTree | 0.52 | 0.69 | 0.60 | 0.69 | 0.69 | 0.75 |
| ExtraTrees | 0.59 | 0.70 | 0.60 | 0.88 | 0.66 | 0.76 |

จากผลการทดลองในตารางที่ 4.8 จะเห็นได้ว่าเมื่อมีการใช้ segmentation แบบ LM และ 2LT สามารถให้ผลการจำแนกจากวิธีต้นไม้ตัดสินใจด้วย CART Algorithm ได้ค่า Precision สูงที่สุดเมื่อเปรียบเทียบกับวิธีการอื่น ในขณะที่วิธีการอื่นได้ค่า Recall ที่สูงกว่า แต่เมื่อดูในเชิงรายละเอียดจากในตารางที่ 4.9 จะเห็นได้ว่าค่าของการจำแนกแบบ KNN ได้ค่า Recall ที่สูงที่สุด เนื่องจากปริมาณความเป็นไปได้ของป้ายต่อข้อความรวมกับการที่ classifier ไม่สามารถจำแนกป้ายทำให้ค่าของ *fn* และ *tp* มีค่าใกล้เคียงหรือเท่ากับศูนย์ และ *tn* เข้าใกล้หรือเท่ากับจำนวนประเภทการจำแนกที่ไม่ต้องการทั้งหมด แต่ค่า Precision ต่ำหรือเข้าใกล้ศูนย์ทำให้ไม่สามารถนำผลของการจำแนกวิธีดังกล่าวมาใช้ได้เนื่องจากไม่สามารถทำนายผลป้ายที่ถูกต้องที่สามารถนำไปใช้งานต่อได้เลย จากเหตุข้างต้นจึงให้เหตุผลว่า Precision มีนัยสำคัญกว่า Recall ในการบ่งชี้ประสิทธิภาพของการจำแนกอาการและอาการแสดง

ตารางที่ 4.9 รายระเอียดของค่าการทำนายผลการจำแนกด้วยวิธี K-Nearest neighbor

| **K-Nearest Neighbor** | | | |
|---|---|---|---|
| *tp* | *fp* | *tn* | *fn* |
| 11 | 113 | 3178 | 0 |
| *precise* | *recall* | *F1score* | |
| 0.09 | 0.92 | 0.16 | |

นอกเหนือจากวิธีการจำแนกแล้วผลการจำแนกยังมีความแตกต่างกันไปตามชุดข้อมูลที่ใช้ในการทดลอง จะเห็นได้ว่าชุดข้อมูลที่ไม่ได้มีการเตรียมให้ผลลัพธ์ในการทำนายผลจำแนกอาการให้ค่า Precision อยู่ที่ 0.59 ซึ่งให้ค่าน้อยกว่า LM และ 2LT ที่ให้ค่า Precision อยู่ที่ 0.88 ทั้งคู่อยู่พอสมควรในการจำแนกแบบวิธี CART ทำให้เห็นว่าการเตรียมข้อความก่อนนำไปใช้ในการจำแนกมีผลอย่างมากต่อความแม่นยำของการทำนาย อย่างไรก็ตามผลการจำแนกมีความไม่คงที่จึงจำเป็นต้องมีการพิสูจน์และปรับปรุงวิธีการเพื่อเพิ่มคุณภาพของผลลัพธ์ซึ่งจะกล่าวถึงในหัวข้อถัดไป

# บทที่ 5

## การเพิ่มประสิทธิภาพของการจำแนกอาการและอาการแสดง

จากผลการทดลองการจำแนกอาการตามรหัส ICD ที่ได้กล่าวถึงในบทที่ 4 นั้น แสดงให้เห็นว่าการจำแนกแบบ Decision Tree ด้วยวิธีการ CART สามารถให้ค่า Precision ได้สูงที่สุดและมีความแตกต่างกันเล็กน้อยระหว่างค่า Precision และ Recall ของการตัดคำแบบ LM และ 2LT อย่างไรก็ตามผลลัพธ์ที่ได้จากการจำแนกมีความไม่คงเส้นคงวาอยู่ ทำให้ไม่สามารถระบุได้อย่างชัดเจนว่าการตัดคำแบบไหนสามารถให้ผลลัพธ์ที่ดีกว่า จึงทำการสรุปผลด้วยวิธีการหาค่าเฉลี่ยของผลการจำแนกตามตารางที่ 5.1 โดยคอลัมน์ 2LT*100 และ LM*100 แทนหัวข้อข้อมูลผลลัพธ์จากการจำแนกจำนวน 100 ครั้งของชุดข้อมูล 2LT และ LM ตามลำดับ โดยค่าเฉลี่ยของผลการจำแนกของชุดข้อมูล 2LT และ LM ให้ค่า Precision อยู่ที่ 0.88 และ 0.85 ค่าเฉลี่ย Recall อยู่ที่ 0.67 และ 0.69 ตามลำดับ และให้ค่า F1-score อยู่ที่ 0.76 เท่ากันทั้งคู่

ตารางที่ 5.1 เปรียบเทียบค่าเฉลี่ยระหว่างผลการทดลองชุดข้อมูล LM และ 2LT

| **Score Type** | ***2LT*100*** | | ***LM*100*** | |
|---|---|---|---|---|
| | ***Precision*** | ***Recall*** | ***Precision*** | ***recall*** |
| **Maximum** | 0.91 | 0.71 | 0.89 | 0.75 |
| **Minimum** | 0.84 | 0.62 | 0.82 | 0.63 |
| **Average** | **0.88** | **0.67** | **0.85** | **0.69** |
| **Standard Deviation** | 0.02 | 0.02 | 0.02 | 0.03 |
| **F1-score** | **0.76** | | **0.76** | |

การทดลองในตารางที่ 5.1 พบว่าผลค่าเฉลี่ยของการจำแนกจากข้อมูลชุด 2LT และ LM อยู่ในเกณฑ์ที่ใกล้เคียงกัน แต่เมื่อทำการสังเกตค่า Recall ให้ผลลัพธ์ที่อยู่ในเกณฑ์ที่ไม่สูงพอที่จะทำให้ข้อมูลมีความน่าเชื่อถือเพียงพอต่อการนำไปใช้งาน อย่างไรก็ตามที่ได้ให้เหตุผลในหัวข้อที่ผ่านมาว่านัยสำคัญของ Precision มีสูงกว่า Recall เนื่องจากสะท้อนถึงข้อมูลที่บรรจุอยู่ในผลของการจำแนก ดังนั้นจึงมีการตั้งเป้าหมายตามลำดับความสำคัญว่าทำอย่างไรจึงสามารถเพิ่มค่า Precision ให้สูงขึ้น

และในขณะเดียวกันก็สามารถเพิ่มค่า Recall ให้สูงขึ้นเช่นกัน แต่ส่งผลกระทบต่อค่า Precision น้อยที่สุด

## 5.1 การพิจารณาค่าที่เกี่ยวข้องกับผลการจำแนก

จากการประเมินผลการจำแนกตามหัวข้อที่ 2.12 ทำให้การคำนวณค่า Precision และ Recall สามารถคำนวณได้จากสมการ $\frac{tp}{tp+fp}$ และ $\frac{tp}{tp+fn}$ ตามลำดับ ซึ่งจากสมการดังกล่าวจะมีค่าที่เกี่ยวข้องคือ *tp, fp* และ *fn* โดยผลการจำแนกประกอบด้วย tp และ *fn* เมื่อ tp เป็นผลการจำแนกประเภทที่ต้องการและถูกจำแนก และเมื่อ *fn* เป็นผลการจำแนกประเภทที่ไม่ต้องการและถูกจำแนก สำหรับส่วนของ *fp* คือค่าผลการจำแนกประเภทที่ต้องการและไม่ถูกจำแนก ดังนั้นเมื่อพิจารณาถึงการคำนวณ Precision ที่อาศัยค่า *tp* เป็นตัวตั้งและค่า *tp* + *fp* เป็นตัวหาร จะได้ว่าการที่จะเพิ่มจำนวน *tp* จะสามารถทำให้ Precision สูงขึ้นได้ แต่เนื่องจากค่า *tp* ผกผันกับค่า *fp* จึงทำให้ตัวหารไม่มีการเปลี่ยนแปลงแต่ค่าของตัวตั้งจะสูงขึ้น หรืออาจกล่าวได้ว่าการเพิ่ม *tp* คือการลด *fp* ในเวลาเดียวกันนั่นเอง สำหรับค่า Recall ที่อาศัยค่า *tp* และ *fn* ในการคำนวณ ซึ่งค่า *tp* ถูกใช้เป็นทั้งตัวตั้งและตัวหารแต่ค่า *fn* ไม่ได้ผกผันกับค่า *tp* เช่นเดียวกับค่า *fp* ที่ใช้สำหรับการคำนวณค่า Precision ดังนั้นการเพิ่มค่าหรือลดค่า *fn* จึงมีผลต่อการคำนวณ Recall เนื่องจากค่าของตัวหารที่ลดลง

จากการอธิบายข้างต้นจะได้ว่าการเพิ่ม *tp* เป็นการลด *fp* ในเวลาเดียวกัน ทำให้สามารถเพิ่มค่า Precision จากตัวตั้งที่เพิ่มขึ้นและค่าของตัวหารที่ลดลงได้ สำหรับการคำนวณ Recall ทำได้โดยการเพิ่ม *tp* ซึ่งเป็นทั้งตัวตั้งและองค์ประกอบของตัวหาร ทำให้ผลจากการเพิ่ม *tp* ไม่ชัดเจนเหมือนกับผลที่เกิดขึ้นต่อค่า Precision แต่สำหรับการลดค่า *fn* สามารถแสดงผลได้ชัดเจนกว่าเนื่องจากเป็นลดค่าของตัวหารโดยตรง ดังนั้นการเพิ่มคุณภาพของข้อมูลการจำแนกสามารถทำได้ด้วยการ เพิ่มปริมาณ tp และลดปริมาณ *fn* นั่นเอง สำหรับลำดับขั้นตอนของการเพิ่มคุณภาพของข้อมูลสามารถเป็นไปได้สองรูปแบบคือ เพิ่มปริมาณ tp ก่อนและลดปริมาณ *fn* หรือ ลดปริมาณ *fn* ก่อนและเพิ่มปริมาณ *tp* โดยรูปแบบลำดับกระบวนการแบบลดปริมาณ *fn* ก่อนและเพิ่มปริมาณ *tp* มีข้อเสียคือ ในกระบวนการเพิ่ม *tp* มีความเป็นไปได้ที่จะเพิ่มปริมาณ *fn* เช่นกันจากขึ้นอยู่วิธีในการเพิ่มปริมาณผลการจำแนกที่คาดว่าจะเป็น *tp* ดังนั้นหากทำการลดปริมาณ *fn* ก่อน จะทำให้ข้อมูล *fn* ที่อาจจะถูกเพิ่มเข้ามาจากกระบวนการเพิ่ม *tp* ไม่ถูกตัดออกไป ดังนั้นจึงต้องทำกระบวนการเพิ่ม *tp*

ก่อน และตามด้วยขั้นตอนการลด *fn* เพื่อป้องกันความผิดพลาดของการเพิ่ม *fn* จากขั้นตอนการเพิ่ม *tp*

## 5.2 การเพิ่มปริมาณ *tp*

การเพิ่มค่า *tp* ถูกแบ่งเป็นสองขั้นตอน คือ การเพิ่มค่า *tp* จากข้อมูลชุดเดียวกัน และ การเพิ่มค่า *tp* โดยอาศัยข้อมูลทั้งสองชุด

### 5.2.1 การเพิ่มค่า tp จากผลการจำแนกจากข้อมูลประเภทเดียวกัน

เมื่อพิจารณาจากตารางที่ 5.1 พบว่าค่าการจำแนกโดยอาศัยชุดข้อมูล 2LT และ LM ได้ค่า Precision เฉลี่ยอยู่ที่ 0.88 และ 0.85 ตามลำดับ การหาผลลัพธ์ของเซตการจำแนกที่ให้ค่า Precision สูงสุดสามารถทำได้ โดยอาศัยวิธีการยูเนียนผลการจำแนกจากข้อมูลประเภทเดียวกัน เพื่อรวมผลการจำแนกที่ต้องการหรือ *tp* ที่มีความแตกต่างกันในแต่ละผลลัพธ์การจำแนก อย่างไรก็ตามการกระทำดังกล่าวก็จะเพิ่มปริมาณของผลการจำแนกที่ไม่ถูกต้องหรือ fn เช่นกัน โดยการทดลองยูเนียน ใช้ผลการจำแนกจำนวน 100 ชุดเพื่อให้สอดคล้องกับจำนวนชุดที่หาค่าเฉลี่ยของตารางที่ 5.1 ผลลัพธ์ของการทดลองยูเนียนผลการจำแนกของทั้งสองประเภท แสดงในตาราง 5.2 โดยแทนผลการยูเนียนของการจำแนกชุดข้อมูล 2LT จำนวน 100 ชุด ด้วย Union(2LT*100) และ ชุดข้อมูล LM จำนวน 100 ชุดด้วย Union(LM*100) หรือแทนด้วยสัญลักษณ์ u1 และ u2 ตามลำดับ ผลการทดลองให้ผลลัพธ์ตามที่ตั้งสมมติฐานข้างต้นว่าสามารถทำให้ค่า Precision สูงสุดของการจำแนกด้วยชุดข้อมูล 2LT และ LM ตามตารางที่ 5.1 คือ 0.91 และ 0.89 และค่า Recall ของ 2LT และ LM ลดลงเหลือ 0.62 และ 0.63 ตามลำดับ โดยข้อมูลสถิติเริ่มมีความคงที่หรือหยุดการเปลี่ยนแปลงตั้งแต่รอบที่ 20 โดยประมาณ

### 5.2.2 การเพิ่มปริมาณ *tp* โดยอาศัยข้อมูลจากผลจำแนกทั้งสองประเภท

นอกจากการเพิ่มค่า Precision ของการทดลองทั้งสองชุดข้อมูลด้วยวิธีการรวมหรือการยูเนียนชุดผลลัพธ์การจำแนกเข้าด้วยกันแล้ว ได้อาศัยข้อมูลจากการตั้งข้อสมมติฐานจากโครงสร้างที่แตกต่างกันของชุดข้อมูล 2 ชุด สามารถให้ผลลัพธ์การจำแนกข้อมูลที่แตกต่างกันมารวมกันเพื่อเพิ่มปริมาณของการทำนายผลที่ถูกต้อง หรือ *tp* เช่นการทำนายผลของข้อความที่มีคำขยายเพิ่มเติมเช่น “ปวดที่ศีรษะ” โดยทั้งสองวิธีสามารถตัดออกมาได้ในรูปแบบเดียวกันคือ “ปวด”, “ที่”, “ศีรษะ” แต่เมื่อทำการจำแนกข้อความด้วย CART decision tree โดยอาศัยชุดข้อมูลที่ตัดคำด้วยวิธี LM ไม่สามารถทำนายผล “ปวด”, “ที่”, “ศีรษะ” เป็นอาการ “ปวดหัว” หรือ “ปวดศีรษะ” ได้

ถูกต้อง จากการทดลองจำนวน 1000 ครั้ง เนื่องจากไม่สามารถหาความเชื่อมโยงระหว่างข้อความทดสอบและข้อความชุดสอนที่มีโครงสร้างแตกต่างกันได้ แต่กลุ่มข้อมูลทดสอบที่ตัดข้อความ CC ด้วยวิธี 2LT และแปลงโครงสร้างของข้อความ ICD ที่เป็นข้อมูลชุดทดสอบด้วยวิธี Conflict Element Finding และ วิธี Medical Context Checking สามารถทำนายผลข้อความ “ปวดที่ศีรษะ” ได้ เมื่อสังเกตในข้อมูลชุดสอนจากข้อความบนฐานรหัส ICD ของทั้งสองวิธี วิธี Conflict Element Finding และ Medical Context Checking ได้ทำการแปลงรูปเดิมจากการตัดคำของ LM คือ “ปวดศีรษะ” ให้อยู่ในรูป “ปวด”,“ศีรษะ” ทำให้สามารถใช้เป็นตัวอย่างสำหรับข้อความที่มีคำเชื่อมเป็นคำขยายได้ ในขณะที่ตัวอย่างข้อมูลที่ตัดด้วยวิธี LM สามารถทำนายผลได้ดีในข้อความที่มีความกำกวมน้อยกว่าจากโครงสร้างของข้อมูลที่มีความซับซ้อนน้อยกว่า

จากเหตุผลขั้นต้นทำให้ตั้งข้อสมมติฐานว่าผลลัพธ์ของการทดลองทั้งสองชุดข้อมูลต่างมีผลการจำแนกที่แตกต่างกัน รวมไปถึงผลการจำแนกที่ถูกต้องย่อมมีบางส่วนที่แตกต่างกัน เมื่อนำผลลัพธ์ทั้งสองชุดคือ u1 และ u2 มารวมกันทำให้สามารถเพิ่มปริมาณข้อมูลที่เป็นไปได้มากขึ้น มีผลให้สามารถเพิ่มปริมาณการจำแนกที่ถูกต้องหรือ *tp* แต่ก็เพิ่มปริมาณ *fn* เป็นผลให้เกิดการเพิ่มของค่า Precision และการลดลงของ Recall ขึ้นอยู่กับคุณภาพของข้อมูลชุดสอนและกระบวนการวิธีที่ใช้ในการจำแนก เมื่อพิสูจน์ด้วยการทดลองจะเห็นได้จากตารางที่ 5.2 ว่าค่า u3 ซึ่งเป็นค่ายูเนียนระหว่าง u1 และ u2 ได้ผลลัพธ์ค่า Precision เพิ่มจากค่าสูงสุด Precision ของ u1 และ u2 จาก 0.91 และ 0.89 เป็น เพิ่มขึ้นเป็น 0.94 และ ค่า Recall ของ u1 และ u2 ลดจาก 0.64 และ 0.63 ลดลงเป็น 0.60 ซึ่งได้ผลลัพธ์ตามที่ได้ตั้งข้อสมมติฐานไว้ เมื่อเปรียบเทียบผล u3 กับค่าเฉลี่ยของการจำแนกด้วยวิธีการปกติแสดงในตารางที่ 5.3 พบว่าสามารถเพิ่มค่า Precision ได้มากกว่าค่า Precision 0.91 ของ u1 อย่างไรก็ตามค่า Recall ที่ลดลงก็ต่ำกว่าค่าต่ำสุดของค่าสถิติเช่นกัน โดยในหัวข้อถัดไปจะกล่าวถึงแนวคิดและวิธีการว่าจะทำอย่างไรจึงจะสามารถเพิ่มค่า Recall และส่งผลกระกระทบต่อค่า Precision น้อยที่สุด

ตารางที่ 5.2 แจกแจงการรวมผลลัพธ์ของสองชุดข้อมูล

| | Union(2LT*100) ≃ (u1) | Union(LM*100) ≃ (u2) | Union(u1, u2) ≃ (u3) |
|---|---|---|---|
| Precision | 0.91 | 0.89 | 0.94 |
| Recall | 0.62 | 0.63 | 0.60 |
| F1-score | 0.73 | 0.74 | 0.73 |

ตารางที่ 5.3 ตารางเปรียบเทียบระหว่าง u3 และ ผลค่าเฉลี่ยของ 2LT*100 และ LM*100

| | **u3** | **2LT*100** | | | **LM*100** | | |
|---|---|---|---|---|---|---|---|
| | | *Avg* | *Max* | *Min* | *Avg* | *Max* | *Min* |
| Precision | 0.94 | 0.88 | 0.91 | 0.84 | 0.85 | 0.89 | 0.82 |
| Recall | 0.60 | 0.67 | 0.71 | 0.62 | 0.69 | 0.75 | 0.63 |
| F1- score | 0.73 | 0.76 | 0.79 | 0.71 | 0.76 | 0.81 | 0.71 |

## 5.3 การลดปริมาณ *fn*

จากทดลองในหัวข้อที่ผ่านมาเราสามารถเพิ่ม Precision ของผลการจำแนกให้สูงถึง 0.94 ได้ แต่ค่า Recall ลดลงอยู่ที่ 0.60 ซึ่งเป็นค่าที่ระดับต่ำกว่าค่าต่ำสุดของผลสถิติของการจำแนกด้วยชุดข้อมูล 2LT และ ML คือ 0.62 และ 0.63 ตามลำดับ โดยค่า Recall ในระดับนี้แสดงถึงความไม่น่าเชื่อถือหรือความกำกวมของผลลัพธ์ของการจำแนก ดังนั้นจึงต้องหาวิธีว่าจะทำอย่างไรให้สามารถเพิ่มค่า Recall โดยส่งผลกระทบต่อค่า Precision น้อยที่สุด

ในหัวข้อ 5.2.2 ได้แสดงให้เห็นว่าแม้ว่าผลลัพธ์ u3 มีค่า Recall เท่ากับ 0.59 แต่ค่า Precision อยู่ที่ 0.94 ซึ่งหมายความว่าในผลลัพธ์ดังกล่าวประกอบด้วยผลการจำแนกที่ถูกต้องจำนวนมาก เพียงแต่มีผลการจำแนกที่ไม่ถูกต้องรวมอยู่ด้วยเช่นกัน ดังนั้นหากสามารถหาเกณฑ์ในการตัดผลการจำแนกที่ไม่ถูกต้องออกไปจะทำให้สามารถเพิ่มค่า Recall ให้สูงขึ้นได้ ทั้งนี้ผลการตัดตัวเลือก (Candidates) จากชุดผลลัพธ์ u3 มีความเป็นไปได้ 2 ประเภทคือ

1. ตัวเลือกที่ตัดออกไปเป็น *tp* หรือเป็นผลลัพธ์ที่จำแนกได้อย่างถูกต้อง การตัดตัวเลือกประเภทนี้ออกไปมีผลทำให้ค่า Precision หรือความแม่นยำของผลการจำแนกลดลง
2. ตัวเลือกที่ตัดออกไปเป็น *fn* หรือเป็นผลลัพธ์ที่จำแนกผิดพลาด การตัดตัวเลือกประเภทนี้ออกไปมีผลทำให้ค่า Recall หรือความน่าเชื่อถือของข้อมูลมีสูงขึ้น

ถึงแม้เป้าหมายคือการตัด *fn* แต่เป็นไปได้ยากที่จะสามารถคัดกรอง *fn* ออกไปได้ทั้งหมด หรือไม่เกิดความผิดพลาดเลือกตัด *tp* ออกจากชุดผลลัพธ์ u3 ซึ่งความถูกต้องของผลลัพธ์ขึ้นอยู่กับกระบวนการวิธีที่ใช้และเกณฑ์ในการตัดตัวเลือกที่คาดว่าจะเป็น *fn* ออกจากผลจำแนกที่ใช้เป็นตัวเลือกหรือ u3 ในการทดลองนี้

**5.3.1 เกณฑ์การตัดตัวเลือก**

การที่จะทำการตัดตัวเลือก จำเป็นต้องเข้าใจถึงลักษณะของข้อมูลที่จะทำการตัดหรือคัดออกก่อน โดยการตัดป้ายที่คาดว่าจะเป็นประเภท *fn* ของผลการจำแนกข้อความ CC ถูกจำแนกด้วยวิธีการเรียนรู้ของเครื่องโดยอาศัยข้อความ ICD เป็นข้อมูลชุดสอน เช่นเดียวกันการตัดป้ายที่คาดว่าจะเป็น *fn* ก็จำเป็นต้องเลือกตัดป้ายจากข้อความ CC ที่ไม่ผ่านเกณฑ์ความสอดคล้องกับข้อความ ICD ที่ถือเป็นข้อมูลตัวอย่างและมีป้ายรหัสเดียวกัน แต่การกำหนดเกณฑ์ที่ใช้สำหรับตัด *fn* ก็ต้องจำเป็นที่จะต้องเข้าใจถึงลักษณะของข้อมูลซึ่งหมายถึงข้อความภาษาไทยในการทดลองนี้ จึงสามารถระบุได้ว่าข้อความ CC และ ข้อความ ICD มีความสัมพันธ์หรือความเกี่ยวข้องกันหรือไม่

ข้อความ ICD ถูกกำหนดเป็นข้อความตัวอย่างในการระบุอาการของข้อความ CC ตามเหตุผลที่ได้ให้ในขั้นต้น เมื่อพิจารณาลักษณะของข้อความ ICD ส่วนใหญ่เป็นข้อความที่เป็นประโยคไม่สมบูรณ์ โดยอาจจะเป็นคำเดี่ยวเช่น “ไข้” หรือ “สั่น” หรือ เป็นวลีเช่น “ปวดหัว” เป็นการประกอบกันระหว่างคำว่า “ปวด” และ “หัว” โดยทั้งสองคำเป็นคำหลักที่มีนัยสำคัญในการสื่อความหมายของข้อความ หากปราศจากคำใดคำหนึ่ง จะไม่สามารถให้ความหมายเหมือนเมื่ออยู่พร้อมกันทั้งสองคำได้ แต่อย่างไรก็ตามทดแทนด้วยคำอื่นที่มีความเดียวกันได้เช่น คำว่า “หัว” สามารถแทนด้วยคำว่า “ศีรษะ” โดยไม่ทำให้ความหมายผิดไปจากรูปแบบองค์ประกอบเดิม นอกจากคำหลักที่มีนัยสำคัญในการสื่อความหมายแล้ว ยังมีคำขยายที่ช่วยให้ข้อความมีความสมบูรณ์มากขึ้น แต่ในขณะเดียวกันคำขยายเช่นคำเชื่อมไม่ได้มีนัยสำคัญในการสื่อความหมายเมื่อคำขยายเหล่านั้นเข้ามาอยู่โครงสร้างของข้อความ จะเป็นองค์ประกอบที่ไม่ใช่นัยสำคัญหลักในการสื่อความหมายของข้อความดังกล่าว หมายความว่าองค์ประกอบในข้อความทั้งหมดอาจไม่ได้มีนัยสำคัญในการสื่ออาการ เช่น คำว่า “ปวดท้อง” สามารถเพิ่มคำว่า “บริเวณ” ให้อยู่ในรูป “ปวดบริเวณท้อง” ไม่ว่าจะมีคำว่า “บริเวณ” หรือไม่ยังคงสื่อความเดียวกัน เนื่องจากไม่จำเป็นว่าทุกองค์ประกอบของข้อความ CC และ ข้อความ ICD จะเป็นคำบ่งชี้ความหมาย ดังนั้นจึงต้องมีวิธีการกำหนดเกณฑ์ในการเปรียบเทียบความสอดคล้องของระหว่างองค์ประกอบของข้อความ CC และข้อความ ICD โดยการแจกแจงโครงสร้างของข้อความอาศัยการวิเคราะห์แบบไวยากรณ์ไม่พึ่งบริบทหรือ Context-Free Grammars (CFG) เพื่อหาความเป็นไปได้ที่จะมีคำเชื่อมเป็นองค์ประกอบของข้อความ สามารถเขียนแจกแจงโครงสร้างของข้อความให้อยู่ในรูปของไวยากรณ์ไม่พึ่งบริบทหรือ CFG ดังนี้

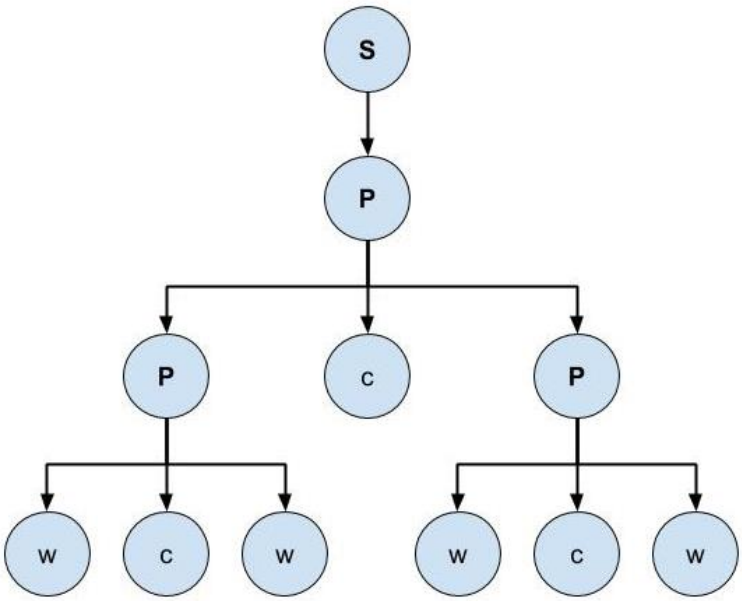
เมื่อ NonTermimal = {S,P,W,C} และ Terminal = {w, c} และกำหนดให้

| | | |
|---|---|---|
| P แทน หน่วยวลี | C | = { c : c เซตคำเชื่อม } |
| S แทน จุดเริ่มต้นของข้อความ | W | = {w : w เซตของคำที่ไม่ใช่คำเชื่อม} |

กฎสำหรับการแจกแจงผลลัพธ์ของ NonTerminal

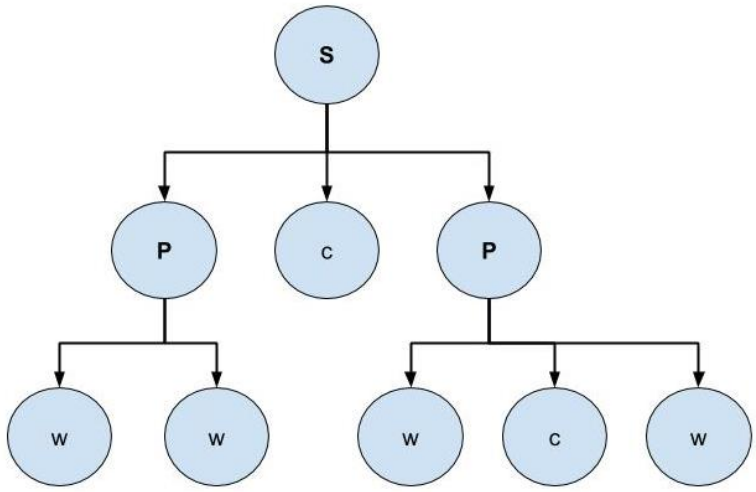
| | |
|---|---|
| $S \longrightarrow P$ | $P \longrightarrow P + c + P$ |
| $P \longrightarrow w + w$ | $P \longrightarrow w + c + w$ |
| $P \longrightarrow P + c + w$ | $P \longrightarrow P + w$ |

จุดประสงค์ของการแจกแจงด้วย CFG เพื่อต้องการจะหาเกณฑ์ในการกำหนดจำนวนขั้นต่ำของ w ในแต่ละข้อความ หรือ ปริมาณสูงสุดที่จะเกิด c เมื่อพิจารณาจากผลของกฎจะเห็นได้ว่ามี 3 กฎที่ P ให้ผลที่มี NonTerminal เป็นองค์ประกอบ คือ P + c + w, P + c + P และ P+w ทั้งสามผลลัพธ์มี P เป็น NonTerminal และกฎที่มี c เป็นองค์ประกอบมี 2 กฎ คือ P + c + w และ P + c + P หาก P ของทั้งสองรูปแบบให้ผลลัพธ์ที่เป็น Terminal ทั้งหมดที่มี c เป็นองค์ประกอบมากที่สุด หรือ w + c + w จะได้ผลลัพธ์ w + c + w + c + w และ w + c + w + c + w + c +w ตามลำดับ




| Terminal set | w | c | w | c | w | c | w |
|---|---|---|---|---|---|---|---|

รูปที่ 5.1 โครงสร้างต้นไม้ที่มี c มากที่สุดที่เป็นไปได้จำนวนคี่




| Terminal | w | w | c | w | c | w |
|---|---|---|---|---|---|---|

รูปที่ 5.2 โครงสร้างต้นไม้ที่มี c มากที่สุดที่เป็นไปได้จำนวนคู่

เมื่อพิจารณาตามกฎที่มี c และ P เป็นองค์ประอบมี 2 กฎ คือ P + c + w และ P + c + P ไม่ว่าทั้งสองกฎจะมีการสร้างผลลัพธ์เป็นรูปแบบที่มี P เป็นองค์ประกอบปริมาณ c สูงสุดที่เป็นไปได้ก็ขึ้นอยู่กับกฎ P ⟶ w + c + w อย่างไรก็ตามผลลัพธ์ที่ได้เป็นปริมาณจำนวนคี่เสมอตามรูปแบบ w + k(c + w) การจะได้โครงสร้างข้อความที่เป็นจำนวนคู่ และมี c เป็นองค์ประกอบของข้อความเยอะที่สุดที่เป็นไปได้ สามารถพิจารณาจากกฎที่ให้ผลลัพธ์เป็นจำนวนคู่คือ P ⟶ w + w, P + w ต้องถูกใช้แจกแจงโครงสร้างหนึ่งครั้งตามรูปที่ 5.2 หรือหมายถึงภายในข้อความมี w องค์ประกอบติดกันโดยไม่มี c หนึ่งคู่ โดยสามารถเขียนแจกแจงได้ว่า w + w + k(c+w) การคำนวณหาจำนวนสูงสุดของ c และจำนวนต่ำสุดของ w ต้องทำการคำนวณหาค่า k ก่อน เพื่อสามารถแจกแจงจำนวนที่เหลือของได้โดยเมื่อกำหนดให้

| | |
|---|---|
| $max(c)$ | เป็นจำนวนสูงสุดที่จะมี c เป็นองค์ประกอบของโครงสร้าง |
| $min(w)$ | เป็นจำนวนต่ำสุดที่จะมี w เป็นองค์ประกอบของโครงสร้าง |
| $n$ | เป็นจำนวน Terminal ปลายทางและ $n = max(c) + min(w)$ |
| $f(n)$ | เป็นรูปแบบของโครงสร้างที่มีจำนวน $c = max(c)$ และ จำนวน $w = max(w)$ เมื่อ $f_1(n)$ เป็นฟังก์ชันเมื่อ $n$ เป็นจำนวนคี่ และ $f_2(n)$ เป็นฟังก์ชันเมื่อ $n$ เป็นจำนวนคู่ |
| k | เป็นจำนวนครั้งที่องค์ประกอบ w ถูกเชื่อมด้วย c |

การคำนวณหาค่า k สำหรับแจกแจงจำนวน c และ w

| เมื่อ $n$ เป็นจำนวนคู่ | เมื่อ $n$ เป็นจำนวนคี่ |
|---|---|
| $n = w + w + k(c + w)$ | $n = w + k(c + w)$ |
| $n = 1 + 1 + k(1 + 1)$ | $n = 1 + k(1 + 1)$ |
| $n = 2 + k(2)$ | $n = 1 + k(2)$ |
| $n - 2 = k(2)$ | $n - 1 = k(2)$ |
| $\frac{n-2}{2} = k$ | $\frac{n-1}{2} = k$ |

หลังจากสามารถหาสมการระบุค่า $k$ ที่แน่นอนได้จึงทำการแจกแจงสมการเพื่อระบุจำนวนค่าสูงสุดของ $c$ ที่เป็นไปได้และ $w$ ต่ำสุดที่เป็นไปได้จากสมการ $f_1(n)$ และ $f_2(n)$

1) $$f_1(n) = \{n > 2 \;\&\; n\%2 = 1 \rightarrow w + (\tfrac{n-1}{2} * (c + w))\}$$
$$\rightarrow \left(\tfrac{(n-1)}{2}c + \tfrac{(n-1)}{2}w + w\right)$$
$$\rightarrow \left(\tfrac{(n-1)}{2}c\right) + \left(\tfrac{(n-1)}{2}w + w\right)$$
$$max_1(c) = \tfrac{(n-1)}{2} = \tfrac{(n)}{2} - \tfrac{(1)}{2}$$
$$min_1(w) = \tfrac{(n-1)}{2} + 1 = \tfrac{(n)}{2} - \tfrac{(1)}{2} + 1 = \tfrac{(n)}{2} + \tfrac{(1)}{2}$$

2) $$f_2(n) = \{n \geq 2 \;\&\; n\%2 = 0 \rightarrow 2w + (\tfrac{n-2}{2} * (c + w))\}$$
$$\rightarrow \left(\tfrac{(n-2)}{2}c + \tfrac{(n-2)}{2}w + 2w\right)$$
$$\rightarrow \left(\tfrac{(n-2)}{2}c\right) + \left(\tfrac{(n-2)}{2}w + 2w\right)$$
$$max_2(c) = \tfrac{(n-2)}{2} = \tfrac{(n)}{2} - \tfrac{(2)}{2} = \tfrac{(n)}{2} - 1$$
$$min_2(w) = \tfrac{(n-2)}{2} + 2 = \tfrac{(n)}{2} - \tfrac{(2)}{2} + 2 = \tfrac{(n)}{2} + 1$$

จากสมการการหาจำนวนสูงสุดของ c และจำนวนต่ำสุดของ w จะเห็นได้ว่าจากทั้ง 2 สมการให้ผลวิธีการคำนวณค่า $max_i(c)$ และ $min_i(w)$ แตกต่างกันเล็กน้อย โดยสามารถพิสูจน์ความสอดคล้องของ $f_i(n)$ กับ $max_i(c)$ และ $min_i(w)$ จากการเปรียบเทียบสมการหา $max_i(c)$ และ $min_i(w)$ กับรูปแบบที่เป็นผลลัพธ์ $f_i(n)$ ตามตารางที่ 5.4 และ 5.5 โดยสามารถพิสูจน์ได้ว่าการคำนวณจำนวนของ $max_i(c)$ และ $min_i(w)$ สามารถให้ผลลัพธ์ที่สอดคล้องกับสมการ$f_i(n)$และค่า $n$ จริง

ตารางที่ 5.4 ตัวอย่างการแจกแจงรูปแบบตามฟังก์ชัน $f_1(n)$

| $n$ | $f_1(n) = w + (\frac{n-1}{2} * (c+w))$ | $max_1(c) = \frac{(n)}{2} - \frac{(1)}{2}$ | $min_1(w) = \frac{(n)}{2} + \frac{(1)}{2}$ |
|---|---|---|---|
| *3* | $w + c + w$ | $\frac{(3)}{2} - \frac{(1)}{2} = 1$ | $\frac{(3)}{2} + \frac{(1)}{2} = 2$ |
| *5* | $w + c + w + c + w$ | $\frac{(5)}{2} - \frac{(1)}{2} = 2$ | $\frac{(5)}{2} + \frac{(1)}{2} = 3$ |
| *7* | $w + c + w + c + w + c + w$ | $\frac{(7)}{2} - \frac{(1)}{2} = 3$ | $\frac{(7)}{2} + \frac{(1)}{2} = 4$ |
| *9* | $w + c + w + c + w + c + w + c + w$ | $\frac{(9)}{2} - \frac{(1)}{2} = 4$ | $\frac{(9)}{2} + \frac{(1)}{2} = 5$ |

ตารางที่ 5.5 ตัวอย่างการแจกแจงรูปแบบตามฟังก์ชัน $f_2(n)$

| $n$ | $f_2(n) = 2w + (\frac{n-2}{2} * (c+w))$ | $max_2(c) = \frac{(n)}{2} - 1$ | $min_2(w) = \frac{(n)}{2} + 1$ |
|---|---|---|---|
| *2* | $w + w$ | $\frac{(2)}{2} - 1 = 0$ | $\frac{(2)}{2} + 1 = 2$ |
| *4* | $w + w + c + w$ | $\frac{(4)}{2} - 1 = 1$ | $\frac{(4)}{2} + 1 = 3$ |
| *6* | $w + w + c + w + c + w$ | $\frac{(6)}{2} - 1 = 2$ | $\frac{(6)}{2} + 1 = 4$ |
| *8* | $w + w + c + w + c + w + c + w$ | $\frac{(8)}{2} - 1 = 3$ | $\frac{(8)}{2} + 1 = 5$ |

อย่างไรก็ตามยังคงไม่สามารถสรุปการกำหนดเกณฑ์เนื่องจากวิธีการคำนวณค่า $max_i(c)$ และ $min_i(w)$ ของสมการ $f_1(n)$ ไม่สอดคล้องกับสมการ $f_2(n)$ จึงจำเป็นต้องมีการปรับสมการทั้งสองชุดเพื่อให้มีความสอดคล้องและสามารถกำหนดที่ชัดเจนได้ โดยการขั้นตอนการปรับสมการทั้งสองชุดให้สอดคล้องกันแสดงในตารางที่ 5.6 ด้วยการปัดเศษของ $\frac{(n)}{2}$ ขึ้นของทุกสมการสามารถแทนได้ว่า $\left\lceil \frac{n}{2} \right\rceil$ โดยการปัดเศษขึ้นไม่ส่งผลกระทบต่อการคำนวณ $max_2(c)$ และ $min_2(w)$ ของสมการ $f_2(n)$ เนื่องจาก $n$ ของ $f_2(n)$ เป็นเลขจำนวนคู่เมื่อทำการหาร 2 จึงได้เศษ 0 เสมอ ต่างจากการหา $max_1(c)$ และ $min_1(w)$ ของสมการ $f_1(n)$ ที่ $n$ เป็นจำนวนคี่ ดังนั้น $\frac{(n)}{2}$ ของ $n$ จำนวนคี่จะมีเศษ 0.5 เสมอ เมื่อปัดเศษขึ้นจะทำให้ผลลัพธ์มีค่าเพิ่มขึ้น 0.5 หน่วยด้วย ดังนั้นทำการลบด้วย 0.5 จะเห็นได้ว่าค่า $max_i(c)$ มีวิธีการคำนวณด้วยสมการเดียวกันจึงสรุปให้ $\max(c)$ มีค่า $\left\lceil \frac{n}{2} \right\rceil - 1$ และ เนื่องจากโครงสร้างมีจำนวนองค์ประกอบ $n$ หน่วยโดยองค์ประกอบมีสองประกอบประเภทคือ c และ w จึงสามารถสรุปได้ว่า $min(w)$ มีค่าเท่ากับ $n - \max(c)$ เสมอ ทำให้สามารถกำหนดเกณฑ์จากสมการ $\max(c) = \left\lceil \frac{n}{2} \right\rceil - 1$ ได้ปริมาณ $c$ สูงสุดมีไม่เกินครึ่งหนึ่งหรือน้อยกว่า 50 เปอร์เซ็นต์ของจำนวนองค์ประกอบทั้งหมด ที่เหลือเป็นองค์ประกอบของ

$w$ หรือคำอื่นที่ไม่ใช่คำเชื่อม แสดงว่าความสอดคล้องขององค์ประกอบข้อมูลตัวอย่างและข้อมูลทดสอบต้องมีความสอดคล้องมากกว่า 50 เปอร์เซ็นต์เป็นต้นไปให้ถือว่าเป็น *tp* และ ข้อมูลตัวอย่างและข้อมูลทดสอบที่มีความสอดคล้องน้อยกว่าหรือเท่ากับ 50 เปอร์เซ็นต์ให้ถือว่าเป็น *fn* และตัดทิ้งจากชุดผลลัพธ์การจำแนก

ตารางที่ 5.6 การปรับสมการให้สอดคล้องกัน

| $f_1(n):\ n>2\ ,n\%2=1$ | | $f_2(n):\ n\geq 2\ ,n\%2=0$ | |
|---|---|---|---|
| $max_1(c)$ | $min_1(w)$ | $max_2(c)$ | $min_2(w)$ |
| $\frac{(n)}{2}-\frac{(1)}{2}$ | $\frac{(n)}{2}+\frac{(1)}{2}$ | $\frac{(n)}{2}-1$ | $\frac{(n)}{2}+1$ |
| $\left\lceil\frac{n}{2}\right\rceil-\frac{(1)}{2}-\frac{(1)}{2}$ | $\left\lceil\frac{n}{2}\right\rceil+\frac{(1)}{2}-\frac{(1)}{2}$ | $\left\lceil\frac{n}{2}\right\rceil-1$ | $\left\lceil\frac{n}{2}\right\rceil+1$ |
| $\left\lceil\frac{n}{2}\right\rceil-1$ | $\left\lceil\frac{n}{2}\right\rceil$ | $\left\lceil\frac{n}{2}\right\rceil-1$ | $\left\lceil\frac{n}{2}\right\rceil+1$ |
| $\left\lceil\frac{n}{2}\right\rceil-1$ | $n-max_1(c)$ | $\left\lceil\frac{n}{2}\right\rceil-1$ | $n-max_2(c)$ |

การสรุปผลสมการสำหรับการคำนวณ

$$
\begin{aligned}
n &= \min(w)+\max(c)\\
\max(c) &= \left\lceil\frac{n}{2}\right\rceil-1\\
\min(w) &= n-\max(c)\\
min_2(w)>min_1(w)>\max(c) &= \left\lceil\frac{n}{2}\right\rceil+1>\left\lceil\frac{n}{2}\right\rceil>\left\lceil\frac{n}{2}\right\rceil-1\\
\therefore\ \min(w)>\max(c)&
\end{aligned}
$$

ตัวอย่าง $n$ เป็นเลขคี่ เมื่อ $n$ มีค่าเท่ากับ 5

$$
\begin{aligned}
f_1(5) &= w+\tfrac{5-1}{2}(c+w)\equiv w+c+w+c+w\\
\max(c) &= \left\lceil\tfrac{5}{2}\right\rceil-1=\ 3-1=2\\
\min(w) &= 5-2=3\\
\therefore \min(w)>\max(c) &= 3>2\ \rightarrow true
\end{aligned}
$$

ตัวอย่าง $n$ เป็นเลขคู่ เมื่อ $n$ มีค่าเท่ากับ 6

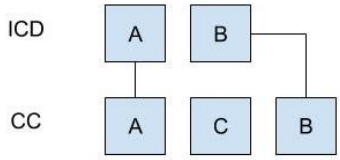
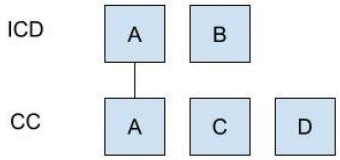
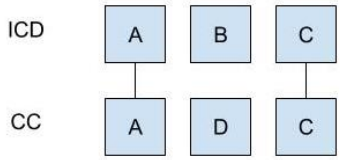
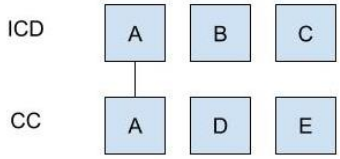
$$
\begin{aligned}
f_2(6) &= 2w+2(c+w)\equiv w+w+c+w+c+w\\
\max(c) &= \left\lceil\tfrac{6}{2}\right\rceil-1=\ 3-1=2\\
\min(w) &= 6-2=4\\
\therefore \min(w)>\max(c) &= 4>2\ \rightarrow true
\end{aligned}
$$

สำหรับการหาความสอดคล้องระหว่างข้อความ ICD และ ข้อความ CC สามารถทำได้โดยการเปรียบเทียบ token จากชุดข้อความที่ผ่านกระบวนการตัดคำด้วยวิธีการเดียวกัน มาเปรียบเทียบความสอดคล้องของสองข้อความ โดยสามารถคำนวณได้เมื่อกำหนดให้

$N(ICD)$ คือจำนวน token ภายในข้อความ ICD

$N(ICD, CC)$ คือจำนวน token ของข้อความ CC ที่สอดคล้องกับข้อความ ICD

$\frac{N(ICD,CC)}{N(ICD)}$ คือความสอดคล้องของข้อความ CC ที่มีต่อข้อความ ICD ในรหัส ICD เดียวกัน

| | | |
|---|---|---|
|  | $N(ICD, CC) = 2$<br>$N(ICD) = 2$<br>$\frac{N(ICD,CC)}{N(ICD)} = \frac{2}{2}$ | $= 1.0$ |
|  | $N(ICD, CC) = 1$<br>$N(ICD) = 2$<br>$\frac{N(ICD,CC)}{N(ICD)} = \frac{1}{2}$ | $= 0.5$ |
|  | $N(ICD, CC) = 2$<br>$N(ICD) = 3$<br>$\frac{N(ICD,CC)}{N(ICD)} = \frac{2}{3}$ | $= 0.67$ |
|  | $N(ICD, CC) = 1$<br>$N(ICD) = 3$<br>$\frac{N(ICD,CC)}{N(ICD)} = \frac{1}{3}$ | $= 0.33$ |

รูปที่ 5.3 การเปรียบเทียบความสอดคล้องข้อมูล ICD และ ข้อมูล CC

### 5.3.2 การทดลองตัด *fn* ตามเกณฑ์ที่กำหนด

การทดลองใช้ผลจำแนก u3 เป็นตัวบ่งชี้ว่าข้อความ CC บรรจุอาการใดบ้างตามรหัส ICD ดังนั้นวิธีการตัดผลการจำแนกที่คาดว่าจะเป็น *fn* ทำด้วยวิธีการหาความสอดคล้องระหว่างองค์ประกอบของข้อความ ICD และ ข้อความ CC ที่ผ่านการตัดคำภาษาไทย และเนื่องจากข้อความ ICD เป็นข้อมูลตัวอย่าง ดังนั้นข้อความ CC บรรจุองค์ประกอบที่สอดคล้องข้อความ ICD > 50 % จึงถือเป็น *tp* และ <= 50 % ถือเป็น *fn* และ ถูกตัดทิ้งออกจากผลการจำแนก ในการทดลอง

ใช้ชุดข้อมูลสองชุดที่มีความแตกต่างกันทางโครงสร้างตามวิธีการตัดคำคือ วิธีการ LM และ 2LT และทำการหาความสอดคล้องขององค์ประกอบด้วยวิธีการอินเตอร์เซคข้อความ ICD และ CC ที่ผ่านการตัดคำจากสองวิธีที่กล่าวมา โดยใช้เกณฑ์ความสอดคล้องที่คำนวณได้คือ > 50 % และเกณฑ์ความสอดคล้องที่ใกล้เคียงคือ >40%, > 49 % และ >60% ในการระบุการจำแนกที่คาดว่าเป็น *tp* และคัดผลการจำแนกที่ไม่ผ่านเกณฑ์ออก ผลการทดลองแสดงในตารางที่ 5.7 โดยอาศัยความสอดคล้องของชุดข้อมูลโครงสร้าง 2LT และ 5.8 อาศัยความสอดคล้องของชุดข้อมูลโครงสร้าง LM จากทั้งสองตารางจะเห็นได้ว่ามีการเปลี่ยนแปลงของคุณภาพของข้อมูลที่ชัดเจนจากเกณฑ์ความสอดคล้อง >0.50% จากตารางที่ 5.7 การอาศัยความสอดคล้องของชุดข้อมูล 2LT สามารถเพิ่มค่า Recall จาก 0.59 ของผลจำแนก u3 เป็น 0.83 โดยที่ยังสามารถรักษาค่า Precision ของ u3 จาก 0.94 ให้อยู่ที่ 0.93 หรือลดลงเพียง 0.01 ได้ ขณะที่การอาศัยความสอดคล้องของชุดข้อมูล LM จากตารางที่ 5.8 สามารถเพิ่มค่า Recall เป็น 0.87 แต่ค่า Precision ลดลงเป็น 0.77

ตารางที่ 5.7 ผลการอินเตอร์เซคแต่ละเกณฑ์จากโครงสร้าง 2LT เปรียบเทียบกับผลจำแนก u3

| **Dataset** | **2LT** | | | | **u3** |
|---|---|---|---|---|---|
| ***Criteria*** | $> 0.40 \approx tp$ | $> 0.49 \approx tp$ | $> 0.50 \approx tp$ | $> 0.60 \approx tp$ | |
| **Precision** | 0.93 | 0.93 | 0.93 | 0.93 | 0.94 |
| **Recall** | 0.71 | 0.71 | 0.83 | 0.83 | 0.90 |
| **F1-score** | 0.80 | 0.80 | 0.88 | 0.88 | 0.73 |

ตารางที่ 5.8 ผลการอินเตอร์เซคแต่ละเกณฑ์จากโครงสร้าง LM เปรียบเทียบกับผลจำแนก u3

| **Dataset** | **LM** | | | | **u3** |
|---|---|---|---|---|---|
| ***Criteria*** | $> 0.40 \approx tp$ | $> 0.49 \approx tp$ | $> 0.50 \approx tp$ | $> 0.60 \approx tp$ | |
| **Precision** | 0.80 | 0.80 | 0.77 | 0.77 | 0.94 |
| **Recall** | 0.74 | 0.74 | 0.87 | 0.87 | 0.60 |
| **F1-score** | 0.77 | 0.77 | 0.82 | 0.82 | 0.73 |

จากผลลัพธ์จากตารางที่ 5.7 และ 5.8 ทำให้เห็นได้ว่าการอาศัยโครงสร้างข้อมูลแบบ 2LT สามารถให้ผลลัพธ์ในการกรองผลการจำแนกที่เป็น *fn* ออกได้โดยส่งผลกระทบต่อค่า Precision น้อยกว่ากว่าการอาศัยโครงสร้างข้อมูลแบบ LM และเมื่อสังเกตผลลัพธ์ที่ได้จากเกณฑ์ที่ใช้สำหรับกรองการจำแนกที่คาดว่าจะเป็น *fn* ออกจากผลการจำแนก พบว่าการเปลี่ยนแปลงของ

คุณภาพข้อมูลเริ่มต้นที่ความสอดคล้องที่ >49% และ >50% ในขณะที่เกณฑ์ในระยะใกล้เคียงกันคือ >40% และ >60% ไม่เกิดความเปลี่ยนแปลงจากเกณฑ์ที่อยู่ใกล้เคียงกัน เมื่อพิจารณาการกำหนดเกณฑ์ตามหัวข้อ 5.3.1 เมื่อ $min_2(w) = \left\lceil \frac{n}{2} \right\rceil + 1$ เป็นฟังก์ชันที่เป็น $n$ จำนวนคู่ และ $min_1(w) = \left\lceil \frac{n}{2} \right\rceil$ เป็นฟังก์ชันที่มี $n$ เป็นจำนวนคี่เมื่อทำการหาร 2 และปัดเศษขึ้นจะมีค่ามากกว่าครึ่งหนึ่ง ทำให้สามารถสรุปได้จากทั้งสองสมการในการเกณฑ์คือ w หรือคำที่คาดว่ามีนัยสำคัญต้องมีปริมาณอย่างน้อยมากกว่า 50% จึงถือว่าเป็น *tp* หากไม่ผ่านเกณฑ์ที่กำหนดถือว่า เป็น *fn* และทำการตัดออกจากผลการจำแนก ทำให้เห็นว่าทั้งรูปแบบโครงสร้าง และ การกำหนดเกณฑ์ความสอดคล้องของข้อมูลมีผลต่อการเพิ่มคุณภาพของผลการจำแนกด้วยวิธีการหาความสอดคล้องของข้อมูล

ในการกรองผลการจำแนกที่คาดว่าจะเป็น *fn* นอกจากผลการจำแนกที่เป็น *fn* ที่ถูกตัดออกแล้วผล *tp* ก็ถูกตัดออกไปเช่นกัน สามารถสังเกตจากค่าที่ลดลงของ Precision หมายถึงปริมาณที่ลดลงของ *tp* เมื่อพบว่าการเปรียบความสอดคล้องขององค์ประกอบด้วยชุดข้อมูล 2LT ให้ค่า Precision มากกว่าชุดข้อมูล LM อยู่ที่ 0.15 คือ 0.93 และ 0.77 ตามลำดับ อย่างไรก็ตามสามารถให้เหตุผลของค่า Precision ที่แตกต่างกันได้จากความแตกต่างของโครงสร้างข้อมูล เนื่องจากชุดข้อมูล 2LT มีโครงสร้างที่มีความละเอียดกว่า LM ทำให้การเปรียบเทียบความสอดคล้องทำได้ยากกว่า ส่งผลให้ *tp* ไม่ผ่านเกณฑ์และถูกตัดออกไป นอกเหนือจากโครงสร้างแล้วเกณฑ์ที่ใช้ก็มีความสำคัญ เช่น กำหนดให้มีความสอดคล้อง 100 % จะเป็นให้เป็นวิธีการที่ไม่ต่างกับเปรียบเทียบข้อความทุกตัวอักษร และเมื่อเกณฑ์เพดานสูงขึ้นก็มีโอกาสที่ *tp* จะไม่ผ่านเกณฑ์เพิ่มขึ้นเช่นกันจึงต้องมีการหาเกณฑ์ขั้นต่ำของความสอดคล้องตามหัวข้อที่ 5.3.1 เพื่อลดโอกาสที่จะตัด *tp* ออกไปแต่ยังคงสามารถตัด *fn* ออกไปได้อย่างมีประสิทธิภาพ

# บทที่ 6

## การสรุปผลและงานในอนาคต

ข้อความบอกเล่าอาการสำคัญ (Chief Complaint : CC) เป็นข้อความที่สื่อถึงอาการสำคัญที่นำผู้ป่วยมายังสถานพยาบาล โดยแพทย์ผู้ตรวจจะเป็นผู้บันทึกข้อมูลข้อความ CC ลงในประวัติผู้ป่วย ภายในข้อความ CC บรรจุข้อมูลที่สามารถช่วยในการนำมาใช้ในการประกอบการวินิจฉัย หรือ ให้ข้อมูลแก่ระบบช่วยตัดสินใจทางการแพทย์ได้ โดยการสกัดข้อมูลจากเอกสารทางการแพทย์หรือชีววิทยา สามารถทำได้โดยอาศัยวิธีการสกัดข้อมูลซึ่งประกอบด้วย การใช้พจนานุกรมที่บรรจุชื่อเฉพาะ การใช้กฎ หรือ การใช้การเรียนรู้ของเครื่อง สำหรับการใช้พจนานุกรมในการสกัดข้อมูลจะเหมาะกับงานที่ข้อมูลหรือนามเอกลักษณ์มีชุดอักขระที่ชัดเจน โดยประสิทธิภาพของการสกัดข้อมูลวิธีนี้ขึ้นอยู่กับความครอบคลุมของพจนานุกรม ในขณะที่การสกัดข้อมูลด้วยกฎเป็นวิธีที่เหมาะกับการสกัดข้อมูลที่รูปแบบขององค์ประกอบของข้อความที่ชัดเจน ซึ่งแตกต่างจากการใช้พจนานุกรมที่ไม่จำเป็นต้องอาศัยข้อความที่เหมือนกันทุกอักขระ แต่อาศัยคำสำคัญหรือสัญลักษณ์พิเศษในการระบุนามเอกลักษณ์ และกำหนดขอบเขตของนามเอกลักษณ์จากตำแหน่งของคำสำคัญ สำหรับการสกัดข้อมูลด้วยการเรียนรู้ของเครื่องต่างจากวิธีการสกัดนามเอกลักษณ์ด้วยการใช้กฎตรงที่ผู้ทำการจำแนกไม่จำเป็นต้องหารูปแบบของนามเอกลักษณ์เอง แต่อาศัยวิธีการเรียนรู้ของเครื่องในการหารูปแบบหรือเกณฑ์จากข้อมูลตัวอย่างในการจำแนกกลุ่มหรือประเภทของข้อมูล อย่างไรก็ตามข้อมูลตัวอย่างหรือข้อมูลชุดสอนจำเป็นต้องมีรูปแบบโครงสร้างหรือการทำสัญลักษณ์ เช่นการติดป้ายเพื่อให้เครื่องสามารถเรียนรู้และจำแนกรูปแบบได้อย่างแม่นยำและถูกต้อง ในขณะเดียวกันการเตรียมข้อมูลตัวอย่างที่เหมาะสมอย่างมีคุณภาพยังคงต้องอาศัย เวลา และ ผู้เชี่ยวชาญในการเตรียมข้อมูลตัวอย่าง อย่างไรก็ตามในชุดข้อมูลเอกสารมาตรฐานภาษาอังกฤษทางชีวการแพทย์รองรับต่อการวิจัยสำหรับการสกัดข้อมูลชีวการแพทย์ภาษาอังกฤษนอกเหนือจากชุดข้อมูลเอกสารมาตรฐานภาษาอังกฤษทางชีวการแพทย์แล้ว ยังมีเครื่องมือสำหรับการช่วยเตรียมข้อความดิบ CC ภาษาอังกฤษ ให้อยู่ในรูปที่พร้อมสำหรับนำไปจำแนกกลุ่มอาการด้วยโปรแกรมประยุกต์ ทำให้เห็นว่าการสกัดหรือจำแนกข้อมูลของเอกสารทางการแพทย์สำหรับภาษาอังกฤษ มีความพร้อมต่อการนำไปประยุกต์ใช้หรือวิจัยเพื่อพัฒนาต่อได้

## 6.1 การสกัดข้อมูลทางการแพทย์ที่ไม่ใช่ภาษาอังกฤษ

ความแตกต่างระหว่างภาษาอังกฤษและภาษาอื่นคือภาษาอังกฤษถือเป็นภาษาสากลของโลก และการทำวิจัยในแต่ละภาษาจำเป็นต้องอาศัย เวลา ทรัพยากร และ ผู้เชี่ยวชาญในพัฒนางานวิจัยในการสกัดข้อมูลทางการแพทย์ของแต่ละภาษา เนื่องจากแต่ละภาษามีอักขระของตัวอักษรและสัญลักษณ์รวมถึงไวยากรณ์ต่างกัน ดังนั้นการกำหนดเป้าหมายการพัฒนาหรือการวิจัยในภาษาเดียว จึงสามารถให้ผลลัพธ์ที่ชัดเจนกว่าการกระจายการวิจัยในหลายภาษา อย่างไรก็ตามการพัฒนาการสกัดข้อมูลส่วนใหญ่จะเป็นภาษาอังกฤษภาษาเดียว ทำให้การที่ภาษาอื่นนอกจากภาษาอังกฤษจะเข้าถึงการใช้งานโปรแกรมประยุกต์ หรือ ประยุกต์ใช้วิธีการสกัดข้อมูลทำให้เกิดความยุ่งยาก ส่งผลทำให้การสกัดข้อมูลสารสนเทศจากเอกสารทางการแพทย์ค่อนข้างจำกัดสำหรับภาษาอื่นๆที่ไม่ใช่ภาษาอังกฤษ และยากต่อการรวบรวมข้อมูลจากเอกสารภาษาต่างๆเพื่อพัฒนาและส่งเสริมการระบบวิเคราะห์ทางการแพทย์

ถึงแม้จะมีข้อจำกัดตามที่ได้กล่าวข้างต้น แต่ยังมีงานวิจัยส่วนหนึ่งที่ทำการทดลองเพื่อเข้าถึงโปรแกรมประยุกต์จำแนกกลุ่มอาการของข้อความ CC ภาษาอังกฤษ ด้วยการแปลข้อความ CC ภาษาจีนเป็นภาษาอังกฤษ ทั้งนี้การแปลข้อความ CC จำเป็นต้องมีความเหมาะสมเพียงพอที่จะสามารถนำไปใช้ในการเตรียมข้อ CC ภาษาอังกฤษ โดยอาศัยวลีสำคัญในการแปลข้อความและซึ่งทำหน้าที่เช่นเดียวกับนามเอกลักษณ์ ในการระบุวลีที่มีนัยสำคัญต่อการจำแนกข้อความ โดยการได้รับวลีสำคัญผ่านขั้นตอนทั้งหมด 2 กระบวนการคือ กระบวนการจำแนกวลีด้วยกฎการหาขอบเขตหน้าและขอบเขตหลังจากข้อมูลสถิติ และกระบวนการเลือกวลีสำคัญและแปลเป็นภาษาอังกฤษด้วยบุคลากรทางการแพทย์ เพื่อนำผลที่ได้ไปพิจารณาถึงวิธีการระบุนามเอกลักษณ์ มีสามวิธีคือ พจนานุกรมที่บรรจุชื่อเฉพาะ การใช้กฎ หรือ การใช้การเรียนรู้ของเครื่อง จะเห็นได้ว่าเนื่องจากขาดชื่อเฉพาะในการระบุนามเอกลักษณ์ จึงต้องอาศัยวิธีการทางสถิติและการใช้กฎในการกรองวลีสำคัญ สำหรับข้อมูลภายในข้อความ CC ส่วนใหญ่นั้นจะเป็นข้อความที่บรรจุอาการที่เป็นการประกอบกันของคำทั่วไปของอาการและตำแหน่ง เช่น คำในภาษาไทย “ปวดแขน” “ปวดขา” หรือ “ปวดหัว” มีส่วนน้อยที่เป็นชื่อระบุ จึงทำให้ยากต่อการระบุขอบเขต หลังจากได้วลีสำคัญการแปลวลีเพื่อนำไปเก็บในพจนานุกรม สำหรับเป็นคำศัพท์ในการแปลข้อความ CC จากที่กล่าวเป็นต้นมา ทำให้สามารถระบุเงื่อนไขที่เกิดขึ้นในการระบุข้อมูลสารสนเทศที่บรรจุอยู่ในข้อความ CC นั่นคือเมื่อต้องอาศัยข้อมูลจำนวนมากเพื่อหาทุกวลีที่เป็นไปได้ ต้องการบุคลากรทางการแพทย์สำหรับการคัดกรองและแปลวลีเป็นภาษาอังกฤษ สำหรับวิธีการใช้การเรียนรู้ของเครื่องจำเป็นต้อง

อาศัยจำนวนบุคลากรและเวลาที่มากกว่า ในการสร้างชุดข้อมูลตัวอย่างที่เหมาะสมและเป็นปริมาณมากพอ และในขณะเดียวกันสามารถนำไปปฏิบัติได้เฉพาะในภาษาที่มีคลังข้อความที่ติดป้ายแล้วเท่านั้น

## 6.2 วิธีการแก้ปัญหา

สำหรับการสกัดข้อมูลภาษาอื่นๆนอกเหนือจากภาษาอังกฤษ หากไม่มีคลังข้อความที่มีการติดป้ายเพื่อใช้สำหรับการสอนเครื่อง ก็จำเป็นต้องหาวิธีในการหาข้อความที่ระบุถึงวลีหรือข้อความที่มีนัยสำคัญ ถึงแม้วิธีการทางสถิติจะสามารถช่วยในการหารูปแบบวลีที่เกิดขึ้นบ่อยได้ แต่ก็จำเป็นต้องอาศัยผู้เชี่ยวชาญในสาขาดังกล่าว ในการจำแนกว่าวลีใดมีนัยสำคัญต่อการสื่อความหมายในบริบทที่สนใจหรือไม่ ในขณะเดียวกันการแปลวลีดังกล่าวถึงแม้จะทำโดยบุคลากรทางการแพทย์ แต่ก็ไม่ได้มีมาตรฐานกลางในการแทนความหมาย จึงทำให้มีข้อด้อยต่อการกระจายข้อมูลต่อระบบทางการแพทย์ที่มีภาษาแตกต่างกัน สำหรับการสกัดข้อมูลจากข้อความ CC ภาษาไทย คลังข้อความทางการแพทย์ที่มีความเหมาะสมต่อการนำไปใช้สำหรับการสอนเครื่องยังคงไม่ได้รับความสนใจมากนัก ทำให้ปัจจุบันยังคงไม่มีคลังข้อความสำหรับการสกัดข้อมูลทางการแพทย์ที่เหมาะสมและนำมาใช้สำหรับการทดลองการสกัดข้อมูลได้ สำหรับการระบุอาการที่บรรจุอยู่ในข้อความ CC การระบุความผิดปกติหรือความเจ็บป่วยตามมาตรฐานรหัส ICD มีรหัสสำหรับการจำแนกอาการหรืออาการแสดงที่เกิดขึ้นต่อผู้ป่วย และ การจำแนกมาตรฐาน ICD ได้ถูกนำไปใช้ในหลายประเทศ และ ประเทศไทยเองก็ได้นำมาตรฐาน ICD มาใช้เช่นกัน โดยปัจจุบันเป็นฉบับที่ 10 หรือ ICD-10 โดยแต่ละประเทศมีการปรับใช้มีเอกสาร ICD เป็นภาษาทางราชการของประเทศนั้น และรหัส ICD-10 สามารถใช้แทนอาการหรือข้อมูลที่บรรจุอยู่ในข้อความ CC ได้โดยไม่มีผลกระทบต่อความแตกต่างทางภาษา อย่างไรก็ตามข้อความจากเอกสาร ICD ก็ไม่ได้ครอบคลุมทุกรูปแบบความเป็นไปได้ของวลีที่สื่ออาการ ดังนั้นจึงจำเป็นต้องมีวิธีที่สามารถนำวลีที่ระบุถึงอาการ ที่มีอยู่ในเอกสาร ICD ไปหาความสอดคล้องกับข้อความ CC ว่าภายในข้อความ CC บรรจุอาการใดตามรหัส ICD บ้าง

การหาความสัมพันธ์ของวลีของ ICD กับข้อความ CC อยู่บนข้อจำกัดของปริมาณรูปแบบที่มีจากเอกสาร ICD จึงทำให้จำเป็นต้องหาวิธีอื่นนอกจากการเปรียบเทียบทุกอักขระ ดังนั้นสำหรับการระบุอาการในข้อความ CC จึงอาศัยวิธีการแยกองค์ประกอบด้วยการตัดคำภาษาไทยด้วยวิธีการอาศัยพจนานุกรม และ อาศัยวิธีอินเตอร์เซคจากทฤษฎีของเซตในการแยกองค์ประกอบที่มีสมาชิกร่วมกัน เนื่องจากมีความเป็นไปได้ที่องค์ประกอบที่มีสมาชิกร่วมกันจะเกิด

รูปแบบอื่นของการประกอบกันเป็นวลีหรือข้อความได้เช่น “ปวดหัว” และ “ปวดท้อง” มีความเป็นไปได้ที่คำว่า “ปวด” จะเป็นสมาชิกร่วม แต่สามารถสร้างรูปแบบองค์ประกอบอื่นที่สามารถให้ความหมายเดิมได้ เช่น “ปวด” “ที่” “ท้อง” หรือ “ปวดศีรษะ” แต่ในขณะเดียวกันการแยกองค์ประกอบให้มีความละเอียดมากขึ้น ก็ทำให้กระบวนการมีความซับซ้อนมากขึ้นเช่นกัน จึงอาศัยพจนานุกรมทางการแพทย์ภาษาอังกฤษในการตรวจสอบว่าองค์ประกอบที่ถูกแยกจากการอินเตอร์เซคอยู่ในบริบททางการแพทย์หรือไม่ เพื่อไม่ให้เกิดความซับซ้อนหรือความกำกวมที่มากขึ้นโดยไม่จำเป็น เช่น คำว่า “อาการ” และ “อาหาร” มี “อา” เป็นองค์ประกอบร่วม แต่ “อา” ไม่อยู่ในบริบททางการแพทย์ จึงไม่มีความจำเป็นต่อการแยกองค์ประกอบ เนื่องจากการแยกองค์ประกอบของข้อความออกเป็นส่วนย่อยถึงแม้จะสามารถทำให้พิจารณาในเชิงโครงสร้างได้ แต่การแยกองค์ประกอบที่มีรายละเอียดมากขึ้น จะส่งผลให้ความซับซ้อนมากขึ้น การลดความซับซ้อนจากปริมาณองค์ประกอบ สามารถประยุกต์วิธีการที่อาศัยข้อมูลเชิงสถิติมาช่วยการหาความเป็นไปได้ที่องค์ประกอบใดตั้งแต่จำนวนสององค์ประกอบขึ้นไปจะอยู่ติดกันเสมอต่อการระบุอาการใดอาการหนึ่ง เช่น วิธีการ N-gram เป็นต้น

## 6.3 การทดลองและการเพิ่มคุณภาพของผลลัพธ์

การทดลองปฏิบัติการบนภาษาไพทอน โดยจำแนกอาการออกเป็นรหัส ICD-10 ด้วยวิธีการเรียนของเครื่อง โดยได้แบ่งชุดข้อมูลเป็น 3 ชุดคือ ข้อมูลที่ไม่ผ่านการเตรียมข้อมูล (Non-segmented dataset) ข้อมูลที่ถูกตัดคำด้วยวิธี Longest Matching (LM-based dataset) และ ข้อมูลที่ถูกตัดด้วยวิธี 2LT (2LT -based dataset) ตามหัวข้อ 5.3.2 โดยในการทดลองไม่ได้นำวิธีทางสถิติมาปรับใช้กับการทดลอง เนื่องจากไม่มีปริมาณข้อมูลมากพอที่จะทำให้วิธีการทางสถิติเห็นผลลัพธ์ได้ โดยผลลัพธ์จากการจำแนกด้วยด้วยวิธีการจำแนกด้วยต้นไม้ตัดสินใจจากอัลกอริทึม CART โดยผลลัพธ์การทดลองจากด้วยชุดข้อมูล 2LT และ LM สามารถให้ค่าเฉลี่ยของ Precision อยู่ที่ 0.88 และ 0.85 ค่า Precision สูงสุดอยู่ที่ 0.90 และ0.89 และค่า Recall เฉลี่ยอยู่ที่ 0.67 และ 0.69 ตามลำดับ โดยผลลัพธ์ของชุดข้อมูลที่ไม่ผ่านการเตรียมข้อมูล ไม่ได้ถูกนำมาพิจารณาเนื่องจากได้ค่า Precision ต่ำกว่า 0.60 ซึ่งน้อยกว่าอีกสองชุดอย่างเห็นได้ชัด สำหรับการเพิ่มค่า Precision ของทั้งสองชุดข้อมูล ทำได้โดยอาศัยการยูเนียนผลลัพธ์การจำแนกของแต่ละชุดข้อมูล ทำให้ได้ชุดผลลัพธ์การจำแนกที่ได้ Precision เป็นค่าสูงสุด และรวมผลลัพธ์ทั้งสองชุดเข้าด้วยกันเพื่อเพิ่มความเป็นไปได้ในการเพิ่มการจำแนกผลที่ถูกต้องหรือ *tp* ทำให้สามารถเพิ่มค่า Precision เป็น 0.94 ได้ อย่างไรก็ตามการผลลัพธ์การจำแนกก็มีความเป็นไปได้ในการเพิ่มผลการทำนายที่ผิดหรือ *fn* เช่นกัน เป็นผล

ให้ค่า Recall ลดลงจากค่าเฉลี่ยของ 2LT และ LM คือ 0.62 และ 0.63 เป็น 0.60 โดยวิธีการเพิ่มค่า Recall อาศัยวิธีการหาความสอดคล้องระหว่างข้อความ ICD และ ข้อความ CC ที่ผ่านการแบ่งองค์ประกอบแล้ว โดยหาเกณฑ์ความสอดคล้องขั้นต่ำที่ส่งผลกระทบต่อ true positive น้อยที่สุดด้วยวิธีการแจกแจงโครงสร้างของภาษาตามหลักของ ไวยากรณ์ไม่พึ่งบริบท (Context-Free grammar) หลังการปรับใช้เกณฑ์เพื่อการคัดกรองผลลัพธ์การจำแนกที่คาดว่าจะเป็น false negative ออกจากชุดจำแนก พบว่าการใช้ชุดข้อมูล 2LT สำหรับการหาความสอดคล้องของข้อมูล ICD และ CC สามารถเพิ่มค่าจาก recall จาก 0.60 เป็น 0.83 และส่งผลกระทบต่อค่า Precision โดยลดลงจาก 0.94 เป็น 0.93 เท่านั้น เมื่อเปรียบเทียบกับการอาศัยชุดข้อมูล LM เพื่อหาความสอดคล้องของข้อมูลที่สามารถเพิ่มค่า Recall เป็น 0.87 แต่ทำให้ค่า Precision ลดลงเป็น 0.77 ทำให้เห็นว่าการพิจารณาในเชิงโครงสร้าง สามารถช่วยเพิ่มประสิทธิภาพของการจำแนกอาการจากข้อจำกัดของวลีที่ไม่ครอบคลุมได้ หากวลีตัวอย่างมีองค์ประกอบที่มีนัยสำคัญในการบ่งชี้อาการ และการจำแนกอาการออกเป็นรหัส ICD ทำให้สามารถการแบ่งปันข้อมูลเพื่อใช้ในการสนับสนุนระบบทางการแพทย์ที่มีภาษาแตกต่างกันได้

## 6.4 งานในอนาคต

ไม่สามารถปฏิเสธได้ว่าข้อมูลทางสถิติสามารถช่วยให้ข้อมูลสำหรับการเพิ่มประสิทธิภาพของการจำแนกข้อมูล ถึงแม้ในวิธีการที่นำเสนอในงานวิจัยจะเป็นการจำแนกโดยการอาศัยโครงสร้างของข้อความบนพื้นฐานของคำศัพท์ทางการแพทย์ แต่ข้อมูลทางสถิติสามารถช่วยในการระบุได้ว่าองค์ประกอบใดในข้อมูลชุดมีผลต่อการระบุอาการหรือไม่มีผลต่อการระบุอาการ หรือใช้ในกระบวนการ Corpus Enhancement ที่ได้เสนอในหัวข้อที่ 3.6 เพื่อหาว่าองค์ประกอบใดที่ควรจะพิจารณาเป็นองค์ประกอบเดียวกันเช่น “ท้อง”,“น้อย” หรือ “ลิ้น”,“ปี่” ในกรณีที่สามารถทำให้องค์ประกอบซ้ำซ้อนลดลงก็จะสามารถทำให้ความกำกวมลดลงด้วยเช่นเดียวกัน เช่น สามารถแยกความแตกต่างระหว่าง “ท้อง” และ “ท้องน้อย” หรือ “ลิ้น” และ “ลิ้นปี่” ได้ นอกเหนือจากการหาว่าองค์ประกอบใดที่ควรพิจารณาเป็นองค์ประกอบเดียวกันหรือไม่ ข้อมูลทางสถิติยังสามารถช่วยในการพิจารณาว่าอาการใดสามารถจำแนกได้ด้วยวิธีการประมวลผลภาษา ตามที่ได้อธิบายถึงข้อความปลายเปิด เช่น “ความผิดปกติทางจิต” เป็นข้อความปลายเปิดที่ไม่มีคำที่เป็นนัยสำคัญสำหรับระบุอาการแน่ชัด เนื่องจากคำว่า “มีอาการกลัวโดยไม่มีสาเหตุ” หรือ “มีอาการตระหนกตลอดเวลา” ก็ถือเป็น “ความผิดปกติทางจิต” จึงทำให้มีความจำเป็นต้องมีการจำแนกว่า อาการใดสามารถ ระบุ

ด้วยวิธีการประมวลผลภาษาได้หรือไม่อย่างชัดเจน โดยในอนาคตจะทำการเพิ่มประสิทธิภาพโดยอาศัยข้อมูลทางสถิติว่าจะสามารถให้ผลลัพธ์การจำแนกที่ดีกว่าการจำโดยไม่อาศัยข้อมูลทางสถิติหรือไม่ นอกจากการปรับใช้วิธีทางสถิติแล้ว ในปี 2561 ได้มีการประกาศที่จะเผยแพร่ ICD ฉบับที่ 11 หรือ ICD-11 โดย World Heath Organization(WHO) โดยมีการตีพิมพ์ทุก 10 ปี ทำให้เห็นว่ามาตรฐาน ICD มีการปรับปรุงและเปลี่ยนแปลงให้ทันกับปัจจุบันเสมอ โดยในอนาคตสามารถนำข้อความจาก ICD-11 มาปรับใช้เพื่อให้สามารถระบุอาการได้ครอบคลุมมากขึ้น ในขณะเดียวกันการเปลี่ยนรหัสจากฉบับหนึ่งไปสู่อีกฉบับสามารถทำได้ เนื่องจากมีโปรแกรมประยุกต์รองรับ เช่น ในปัจจุบันมีเว็บไซต์ที่สามารถแปลงรหัส ICD-9 เป็น ICD-10 แบบออนไลน์ได้ หรือ ระบบทางการแพทย์ภาษาอังกฤษมีที่การทำงานดังกล่าว จึงทำให้เห็นว่าถึงแม้รหัสในการระบุอาการจะมีการเปลี่ยนแปลงไปตามการพัฒนา ยังคงสามารถแก้ไขรหัสในฉบับก่อนหน้า ให้สอดคล้องกับฉบับปัจจุบันได้

# บรรณานุกรม


[1] Rouse, M. 2017. electronic health record (EHR). http://searchhealthit.techtarget.com/definition/electronic-health-record-HER. (accessed March 26, 2018).

[2] healthcarethai. http://www.healthcarethai.com/การซักประวัติผู้ป่วย. (accessed January 5, 2018).

[3] World Health Organization. 2011. History of ICD. www.who.int/classiications/icd/en/. (accessed January 26, 2018).

[4] Gunawardena, A. (Ed.). 2005. DataStructures. http://www.cs.cmu.edu/~clo/www/CMU/DataStructures/. (accessed March 26, 2018).

[5] tutorialspoint. 2018. https://www.tutorialspoint.com/data_structures_algorithms/tree_data_structure.htm. (accessed March 26, 2018).

[6] Mitchell, T. 1997. Machine Learning. McGraw Hill. p. 2

[7] Wikipedia. 2018. การเรียนรู้ของเครื่อง. https://th.wikipedia.org/wiki/การเรียนรู้ของเครื่อง. (accessed March 26, 2018).

[8] James, G., Witten, D., Hastie, T., & Tibshirani, R. 2013. An introduction to statistical learning (Vol. 112). Springer: New York.

[9] scikit-learn. Ensemble methods. http://scikit-learn.org/stable/modules/ensemble.html. (accessed March 26, 2018).

[10] W. 2016. Decision Tree Flavors: Gini Index and Information Gain. http://www.learnbymarketing.com/481/decision-tree-flavors-gini-info-gain/. (accessed January 26, 2018).

[11] Brownlee, J. 2016. How To Implement The Decision Tree Algorithm From Scratch In Python. https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/. (accessed March 26, 2018).

[12] Hodge, G. M. 2004. Understanding metadata. NISO Press: Bethesda MD.

[13] D Nau, D. S. 2010. Part-of-Speech Tagging. https://www.cs.umd.edu/~nau/cmsc421/part-of-speech-tagging.pdf. (accessed January 8, 2018).

[14] Sornlertlamvanich, V., Takahashi, N., & Isahara, H. 1999. Building a Thai part-of-speech tagged corpus (ORCHID). Journal of the Acoustical Society of Japan (E), 20(3): 189-198.

[15] Nectec. 2010. Benchmark for Enhancing the Standard of Thai language processing. https://thailang.nectec.or.th/downloadcenter/. (accessed March 26, 2018).

[16] Wikipedia. 2018. Information extraction. https://en.wikipedia.org/wiki/Information_extraction. (accessed January 26, 2018).

[17] Sarawagi, S. 2008. Information extraction. Foundations and Trends® in Databases, 1(3): 261-377.

[18] Jurafsky, D., & Martin, J. H. 2000. N-gram. In: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, NJ: Prentice Hall, pp 189-232.

[19] Ketui, N., Theeramunkong, T., & Onsuwan, C. 2013. Thai elementary discourse unit analysis and syntactic-based segmentation. International Information Institute (Tokyo). Information, 16(10), 7423.

[20] krupasathaipanmai1920. 2009. วลีและประโยค. https://krupasathaipanmai1920.wordpress.com/2009/09/09/มารู้จักวลีและประโยคกั/. (accessed March 26, 2018).

[21] krusarunya. 2013. วลีและชนิดของวลี. https://krusarunya.wordpress.com/2013/06/26/ประโยค/. (accessed March 26, 2018).

[22] Kanchanawan, N. 1999. ไวยากรณ์ไทยตามทฤษฎีโครงสร้าง. In: Analysis of Thai structure. Ramkhamhaeng University: Thailand, pp 161-206.

[23] Charnyapornpong, S. 1983. A Thai Syllable Separation Algorithm. Asian Institute of Technology, Pathumthani, Thailand.37

[24] Poowarawan, Y. 1986. Dictionary-based Thai syllable separation. In Proceedings of the Ninth Electronics Engineering Conference. Thailand, pp. 409-418.

[25] Sornlertlamvanich, V. 1993. Word segmentation for Thai in machine translation system. Machine Translation, National Electronics and Computer Technology Center, Bangkok: 50-56.

[26] Haruechaiyasak, C., Kongyoung, S., & Dailey, M. 2008, May. A comparative study on Thai word segmentation approaches. In Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on 1, pp 125-128. IEEE.

[27] Wikipedia. 2018. Context-Free grammar. https://en.wikipedia.org/wiki/Context-free_grammar. (accessed March 26, 2018).

[28] Halfeld, M. Context-Free Grammars (CFG). https://www.univ-orleans.fr/lifo/Members/Mirian.Halfeld/Cours/TLComp/l3-CFG.pdf. (accessed March 26, 2018).

[29] tutorialspoint. 2018. Context-Free Grammar Introduction. https://www.tutorialspoint.com/automata_theory/context_free_grammar_introduction.htm. (accessed March 26, 2018).

[30] Doyle, D. 2018. Thai Stopwords. https://www.ranks.nl/stopwords/thai-stopwords. (accessed June 27, 2018).

[31] Yedidia, A. 2016. Against the F-score. https://adamyedidia.files.wordpress.com/2014/11/f_score.pdf. (accessed June 27, 2018).

[32] Dara, J., Dowling, J. N., Travers, D., Cooper, G. F., & Chapman, W. W. (2008). Evaluation of preprocessing techniques for chief complaint classification. Journal of Biomedical Informatics, 41(4): 613-623.

[33] Lu, H. M., Zeng, D., Trujillo, L., Komatsu, K., & Chen, H. 2008. Ontology-enhanced automatic chief complaint classification for syndromic surveillance. Journal of biomedical informatics, 41(2): 340-356.

[34] Song, M., Kim, W. C., Lee, D., Heo, G. E., & Kang, K. Y. 2015. PKDE4J: Entity and relation extraction for public knowledge discovery. Journal of biomedical informatics, 57: 320-332.

[35] Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C., & Xu, H. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. Journal of the American Medical Informatics Association, 18(5): 601-606.

[36] Chen, Y., Lasko, T. A., Mei, Q., Denny, J. C., & Xu, H. 2015. A study of active learning methods for named entity recognition in clinical text. Journal of biomedical informatics, 58: 11-18.

[37] Mollá, D., & Santiago-Martinez, M. E. 2011. Development of a corpus for evidence based medicine summarisation.

[38] Deleger, L., Li, Q., Lingren, T., Kaiser, M., & Molnar, K. 2012. Building gold standard corpora for medical natural language processing tasks. In AMIA Annual Symposium Proceedings, pp 144. American Medical Informatics Association.

[39] Lu, H. M., Chen, H., Zeng, D., King, C. C., Shih, F. Y., Wu, T. S., & Hsiao, J. Y. 2009. Multilingual chief complaint classification for syndromic surveillance: an experiment with Chinese chief complaints. international journal of medical informatics, 78(5): 308-320.

[40] UNC School of Medicine. 2018. Outpatient Medicine Clerkship. https://www.med.unc.edu/medselect/resources/sample-notes/sample-write-up-1. (accessed June 27, 2018).

[41] UNC School of Medicine. 2018. Outpatient Medicine Clerkship. https://www.med.unc.edu/medselect/resources/sample-notes/sample-write-up-1. (accessed June 27, 2018).

[42] UCSD. 2015. A Practical Guide to Clinical Medicine. https://meded.ucsd.edu/clinicalmed/write.htm. (accessed June 27, 2018).

[43] UNC School of Medicine. 2018. Medicine Clerkship Inpatient. https://www.med.unc.edu/medclerk/files/. (accessed June 27, 2018).

[44] MCP med student resources. chief complaint. https://sites.google.com/site/mcpmedstudent/how-to-write-a-progress-note/chief-complaint. (accessed June 28, 2018).

[45] LUC. 2006. Chief Complaint. http://www.lumen.luc.edu/lumen/meded/elective/pulmonary/copd/120699/case2.htm. (accessed June 28, 2018).

[46] ILLINOIS STATE UNIVERSITY. Diagnostic Reasoning for Advanced Nursing Practice 43. http://my.ilstu.edu/~ddwilso2/nur431/analysis_of_symptom.htm. (accessed June 28, 2018).

[47] MARYLAND. 2015. CARE PLAN EXAMPLE. http://hscrc.maryland.gov/Documents/md-maphs/wg-meet/cc/2015-02-12/4-Care-Plan-Example-1.pdf. (accessed June 28, 2018).

[48] SONOMA State University. 2009. Health Maintenance Practicum. http://web.sonoma.edu/users/s/smithwe/549/week2.html. (accessed June 28, 2018).

[49] Wagner, M. M., Hogan, W. R., Chapman, W., & Gesteland, P. 2006. Chief complaints and ICD Codes. Handbook of biosurveillance, 333-360.

[50] Swanson, D., Elnicki, D., Chief Complaint Assessment (CCA) Workshop. www.im.org/d/do/4994. (accessed June 28, 2018).

[51] Tufts University. 2007. Medical Record Review. http://ocw.tufts.edu/data/38/439780.pdf. (accessed June 28, 2018).

[52] Advocate Children's Hospital - Oak Lawn. History and Physical Guide. http://hope-pediatrics.com/hope/node/94. (accessed June 28, 2018).

[53] Wayne State University. H&P – Example – Partial Exam. http://www2 . med.wayne.edu/aesculapians/documents/Year%2 0 Two/ClinMed%2 0 PD/H&P%20-%20Sample%20-%20Partial%20Exam.doc. (accessed June 28, 2018).

[54] Life Line Medical Ambulance. Run Report. http://www.lifelinemedicalambulance.com/run_report.html. (accessed June 28, 2018).

[55] McMasters, M. K. 2010. E/M Coding. http://louisvillesurgery.com/downloads/InpatientEMCodingPres.pdf. (accessed June 28, 2018).

[56] Suanders. 2012. Clinical Procedures Chapter 1 homework. https://www.quia.com/files/quia/users/gerrylandes/Clinical-Procedures-Chapter-1-Homework-8th-Edition.docx. (accessed June 28, 2018).

[57] Gotoknow. 2013. Chief Complaint. https://www.gotoknow.org/posts/402169. (accessed June 29, 2018).

[58] Department of Family Medicine, Chiang Mai University. 2012. ตัวอย่างการเขียนรายงานผู้ป่วย. (accessed June 29, 2018).

[59] Nakhu. 2015. ตัวอย่าง กรณีศึกษา กุมาชเวชกรรม Acute diarrhea, R/O Shigellosis. http://nakhu.com/ตัวอย่าง-กรณีศึกษา-กุมาช/. (accessed June 29, 2018).

[60] Lamphun Provincial Heath office. 2017. แนวทางการดำเนินการคุณภาพเวชระเบียน. https://lamphunhealth.go.th/web_ssj/webmanager/uploads/2017-02-14082156แ น ว ทางการดำเนินการคุณภาพเวชระเบียน.pdf. (accessed June 29, 2018).

[61] เอกชัย-สุธากานต์ เชาว์ดี. ตรวจสอบคุณภาพกาบันทึกเวชระเบียน. http://203.157.213.9/web2014/wp-content/uploads/2016/02/การตรวจสอบคุณภาพกาบันทึกเวชระเบียน.pptx. (accessed June 29, 2018).

[62] Tubsiruk. N., Case Study. https://prezi.com/8myxzjskqz6z/case-study/. (accessed June 29, 2018).

[63] e-medtools and raj&co. 2018. List of English Medical Terms. https://github.com/glutanimate/wordlist-medicalterms-en. (accessed July 22, 2018).

# ภาคผนวก

# ภาคผนวก ก

# ชุดข้อมูล

ในส่วนของภาคผนวก ก จะแสดงชุดข้อมูลที่ใช้ในการทดลอง และ ผลการจำแนกที่เกิดขึ้นในแต่ละขั้นตอน และ ผลการทดลองประกอบด้วย ผลการจำแนกจากการเพิ่มค่า precision และ ผลการจำแนกของการเพิ่มค่า recall

## ก.1 ชุดข้อมูลที่ในการทดลอง

ชุดข้อมูลที่ใช้การทดลอง ประกอบด้วย ชุดข้อมูลที่ไม่ผ่านกระบวนการตัดคำ ชุดข้อมูล LM ที่ตัดคำด้วยวิธีการอาศัยพจนานุกรมแบบ Longest Matching และ ชุดข้อมูล 2LT ที่ผ่านกระบวนการ Conflict Element finding และ Medical context checking ข้อความในชุดข้อมูลมีจำนวนทั้งหมด 169 ข้อความ ข้อความแถวที่ 1 ถึง 82 เป็นข้อความ ICD ซึ่งใช้เป็นข้อมลชุดสอน และ 83 ถึง 169 เป็น ข้อความ CC ที่ใช้สำหรับทดสอบ

| **ชุดข้อมูลที่ไม่ผ่านการตัดคำ (Non-segmentation Dataset)** | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 1 | R00.0 | หัวใจเต้นเร็ว |
| 2 | R00.0 | ใจเต้นเร็ว |
| 3 | R00.1 | หัวใจเต้นช้า |
| 4 | R00.2 | ใจสั่น |
| 5 | R00.2 | ใจเต้น |
| 6 | R04.0 | เลือดกำเดาไหล |
| 7 | R04.0 | เลือดกำเดาออก |
| 8 | R04.2 | ไอเป็นเลือด |
| 9 | R04.2 | เสมหะมีเลือด |
| 10 | R04.2 | เสลดมีเลือด |
| 11 | R05 | ไอ |
| 12 | R06.0 | หายใจลำบาก |
| 13 | R06.0 | หายใจไม่ทัน |

| ชุดข้อมูลที่ไม่ผ่านการตัดคำ (Non-segmentation Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 14 | R06.0 | หายใจสั้น |
| 15 | R06.6 | สะอึก |
| 16 | R06.7 | จาม |
| 17 | R07.0 | เจ็บในคอ |
| 18 | R07.0 | เจ็บในลำคอ |
| 19 | R07.0 | เจ็บคอ |
| 20 | R07.1 | เจ็บหน้าอกเวลาหายใจ |
| 21 | R07.4 | เจ็บหน้าอก |
| 22 | R10.0 | ปวดท้องเฉียบพลัน |
| 23 | R10.1 | ปวดท้องเฉพาะท้องส่วนบน |
| 24 | R10.1 | ปวดท้องเฉพาะส่วนบน |
| 25 | R10.1 | ปวดท้องบริเวณลิ้นปี่ |
| 26 | R10.1 | ปวดท้องตรงส่วนบน |
| 27 | R10.2 | ปวดอุ้งเชิงกรานและฝีเย็บ |
| 28 | R10.2 | ปวดอุ้งเชิงกราน |
| 29 | R10.2 | ปวดฝีเย็บ |
| 30 | R10.3 | ปวดท้องน้อย |
| 31 | R10.4 | ปวดท้อง |
| 32 | R11.9 | คลื่นไส้และอาเจียน |
| 33 | R11.9 | คลื่นไส้ |
| 34 | R11.9 | อาเจียน |
| 35 | R14 | ท้องอืด |
| 36 | R14 | ท้องมีแก๊ส |
| 37 | R14 | ท้องเฟ้อ |
| 38 | R14 | ท้องขึ้น |
| 39 | R25.1 | อาการสั่น |
| 40 | R25.1 | สั่น |

| ชุดข้อมูลที่ไม่ผ่านการตัดคำ (Non-segmentation Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 41 | R25.2 | ตะคริวและกล้ามเนื้อหดเกร็ง |
| 42 | R25.2 | ตะคริว |
| 43 | R25.2 | กล้ามเนื้อหดเกร็ง |
| 44 | R25.2 | กล้ามเนื้อเกร็ง |
| 45 | R26.2 | เดินลำบาก |
| 46 | R26.2 | เดินยาก |
| 47 | R26.3 | เคลื่อนที่ไม่ได้ |
| 48 | R30.0 | ถ่ายลำบาก |
| 49 | R30.0 | ฉี่ลำบาก |
| 50 | R30.1 | ปวดเบ่งปัสสาวะ |
| 51 | R30.1 | ปวดเบ่งฉี่ |
| 52 | R30.9 | ปวดเวลาถ่ายปัสสาวะ |
| 53 | R30.9 | ปวดขณะถ่ายปัสสาวะ |
| 54 | R30.9 | ปวดเวลาฉี่ |
| 55 | R30.9 | ปวดขณะฉี่ |
| 56 | R31 | ปัสสาวะเป็นเลือด |
| 57 | R31 | ฉี่เป็นเลือด |
| 58 | R32 | กลั้นปัสสาวะไม่ได้ |
| 59 | R32 | กลั้นฉี่ไม่ได้ |
| 60 | R33 | ปัสสาวะไม่ออก |
| 61 | R33 | ฉี่ไม่ออก |
| 62 | R34 | ไม่มีปัสสาวะและปัสสาวะน้อย |
| 63 | R34 | ไม่มีปัสสาวะ |
| 64 | R34 | ปัสสาวะน้อย |
| 65 | R34 | ไม่มีฉี่ |
| 66 | R34 | ฉี่น้อย |
| 67 | R35 | ปัสสาวะมาก |

| ชุดข้อมูลที่ไม่ผ่านการตัดคำ (Non-segmentation Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 68 | R35 | ฉี่มาก |
| 69 | R42 | เวียนศีรษะ |
| 70 | R42 | เวียนหัว |
| 71 | R50.9 | ไข้ |
| 72 | R51 | ปวดศีรษะ |
| 73 | R51 | ปวดหัว |
| 74 | R53 | เหนื่อยล้า |
| 75 | R53 | อ่อนแรง |
| 76 | R53 | อิดโรย |
| 77 | R53 | อ่อนเปลี้ย |
| 78 | R53 | อ่อนเพลีย |
| 79 | R55 | เป็นลม |
| 80 | R56.0 | การชักจากไข้สูง |
| 81 | R56.0 | ชักจากไข้สูง |
| 82 | R57.9 | ช็อก |
| 83 | R04.2,R07.0 | ไอเป็นเลือดและเจ็บคอ |
| 84 | R05,R50.9 | ไอและมีไข้ |
| 85 | R55,R57.9 | เป็นลมและช็อก |
| 86 | R00.0 | มีอาการใจเต้น เป็นมา 3 วัน |
| 87 | R04.2 | มีอาการไอออกมาเป็นเลือด |
| 88 | R04.2 | มีอาการไอและมีเลือดในเสลด |
| 89 | R06.0 | หายใจยาก |
| 90 | R07.0 | เจ็บในคอ |
| 91 | R07.1 | เจ็บหน้าอกขณะหายใจ |
| 92 | R07.1 | เจ็บหน้าอกเมื่อหายใจ |
| 93 | R07.4 | เจ็บบริเวณหน้าอก |
| 94 | R07.4 | เจ็บหน้าอก |

| ชุดข้อมูลที่ไม่ผ่านการตัดคำ (Non-segmentation Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 95 | R10.1 | ปวดบริเวณลิ้นปี่ |
| 96 | R10.1 | ปวดลิ้นปี่ |
| 97 | R14 | มีแก๊สในท้อง |
| 98 | R25.2 | เป็นตะคริว |
| 99 | R26.3 | เคลื่อนที่ลำบาก |
| 100 | R31 | มีเลือดในฉี่ |
| 101 | R34 | มีฉี่น้อย |
| 102 | R50.9 | เป็นไข้ |
| 103 | R50.9 | มีไข้ |
| 104 | R51 | มีอาการปวดที่ศีรษะ |
| 105 | R51,R50.9,R06.7 | มีอาการปวดหัวเป็นไข้และจาม |
| 106 | R10.4,R30.0 | มีอาการปวดท้องและฉี่ลำบาก |
| 107 | R00.2,R04.0 | ใจเต้น และ เลือดกำเดาไหล |
| 108 | R25.2,R26.2 | เป็นตะคริวที่ขาและเดินลำบาก |
| 109 | R50.9,R06.7,R04.2 | เป็นไข้ จามและในเสลดมีเลือด |
| 110 | R51,R11.9 | ปวดหัวคลื่นไส้และอาเจียน |
| 111 | R30.9,R34 | ปวดเวลาปัสสาวะและมีฉี่น้อย |
| 112 | R25.1,R50.9,R07.4 | มีอาการสั่นมีไข้และเจ็บหน้าอก |
| 113 | R07.4,R06.0 | เจ็บหน้าอกและหายใจลำบาก |
| 114 | R10.1,R11.9 | ปวดบริเวณลิ้นปี่มีอาการคลื่นไส้และอาเจียน |
| 115 | R07.0,R06.0 | มีอาการเจ็บในลำคอและหายใจลำบาก |
| 116 | R10.4,R14 | ปวดท้องและท้องอืด |
| 117 | R51,R05,R06.7 | มีอาการปวดหัวไอและจาม |
| 118 | R25.1,R57.9 | มีอาการสั่น และ ช็อก |
| 119 | R51,R53,R50.9 | ปวดหัว อ่อนเพลียและมีไข้ |
| 120 | R10.4 | ปวดท้อง |
| 121 | R07.4 | นาย Smith อายุ 70 ปี เข้าพบเพื่อตรวจสอบอาการเจ็บหน้าอกที่ |

| ชุดข้อมูลที่ไม่ผ่านการตัดคำ (Non-segmentation Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| | | มากขึ้น |
| 122 | R06.0 | ลิ้นบวมและหายใจและกลืนลำบาก |
| 123 | R06.0 | หายใจไม่ทัน |
| 124 | R07.4 | ผู้หญิงอายุ 36 มีอาการเจ็บหน้าอกเฉียบพลันหลังประสบประสบอุบัติเหตุทางรถยนต์ |
| 125 | R06.0 | ฉันหายใจไม่ทัน ฉันหยุดไอไม่ได้ |
| 126 | R10.4 | ปวดท้องและท้องร่วงเป็นเวลา 3 สัปดาห์ |
| 127 | R06.0 | ฉันหายใจเสียงดัง และ หายใจไม่ทัน |
| 128 | R50.9,R05 | นาย H อายุ 50 ปี ผู้ป่วยโรค AIDS เข้าพบเพื่อตรวจอาการเป็นไข้ หนาว ไอ เป็นเวลา 3วัน |
| 128 | R06.0 | หายใจลำบาก |
| 130 | R07.4 | เจ็บหน้าอก |
| 131 | R10.4,R11.9 | ปวดท้อง คลื่นไส้ อาเจียน |
| 132 | R50.9 | ไข้ |
| 133 | R11.9,R06.6 | คลื่นไส้ ท้องร่วง แน่นหน้าอก สะอึก |
| 134 | R07.4,R11.9 | เจ็บหน้าอก อาเจียน |
| 135 | R05 | ไอ |
| 136 | R42 | เวียนหัว |
| 137 | R30.0 | ปัสสาวะลำบาก |
| 138 | R53 | อ่อนแรง |
| 139 | R51 | ปวดหัว |
| 140 | R04.2 | ไอเป็นเลือด |
| 141 | R31 | มีเลือดในปัสสาวะ |
| 142 | R42 | อาเจียน |
| 143 | R00.2 | ใจสั่น |
| 144 | R55 | เป็นลม |
| 145 | R32 | กลั้นปัสสาวะไม่ได้ |

| ชุดข้อมูลที่ไม่ผ่านการตัดคำ (Non-segmentation Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 146 | R10.1,R50.9 | ผู้ป่วยบอกเล่าเกี่ยวกับอาการปวดลิ้นปี่พร้อมกับมีไข้และหนาว เป็นช่วงเวลา 3 สัปดาห์ |
| 147 | R10.1,R05 | เด็กผู้ชายอายุ 12 ปี มีอาการไข้ ไอ และ หายใจเสียงฮืดๆ แย่ลง ในช่วงเวลา 3 วันที่ผ่านมา |
| 148 | R42 | เวียนหัวและขาขวาอ่อนแรง |
| 149 | R06.0,R50.9,R05 | หายใจสั้น มีไข้ ไอ |
| 150 | R06.0 | ผู้หญิงอายุ 31 ปี เข้าพบในวันนี้ด้วยอาการหายใจสั้นอย่าง รุนแรง |
| 151 | R07.4 | ผู้ชายอายุ 55 ปีเข้าพบในวันนี้ด้วยอาการเจ็บหน้าอกอย่าง รุนแรง |
| 152 | R05,R50.9,R06.0 | ผู้ชายอายุ 70 ปี ป่วยเป็นโรคหอบหืด กรดไหลย้อน และ ปอดบวม มีอาการไอ มีไข้ และ หายใจสั้น |
| 153 | R50.9 | ปวดหูและมีไข้เป็นเวลา 2 วัน |
| 154 | R30.9 | ปวดขณะปัสสาวะเริ่มเป็นเมื่อวาน |
| 155 | R07.0,R50.9 | เจ็บคอและเป็นไข้เป็นเวลา 24 ชั่วโมง |
| 156 | R07.4 | เจ็บหน้าอกเกิดขึ้นตอนเช้าวันนี้ |
| 157 | R10.4 | ปวดท้อง 4 ชั่วโมง ก่อนมาโรงพยาบาล |
| 158 | R51 | ปวดศีรษะมา 3 วัน ก่อนมาโรงพยาบาล |
| 159 | R50.9,R05 | มีไข้ ไอ เป็นมา 3 วัน ก่อนมาโรงพยาบาล |
| 160 | R07.0 | เจ็บคอ 2 วันก่อนมาโรงพยาบาล |
| 161 | R10.4 | ปวดท้อง ถ่ายเหลว 7 ชั่วโมง ก่อนมาโรงพยาบาล |
| 162 | R10.1 | ปวดท้องจุกแน่นลิ้นปี่ 1 อาทิตย์ก่อนมาโรงพยาบาล |
| 163 | R50.9 | มีไข้ ถ่ายเหลว 6 ครั้ง เป็นก่อนมา 1 วัน |
| 164 | R50.9,R05 | ไข้ ไอ เป็นมา 2 วัน |
| 165 | R05 | ไอ เมื่อคืน |
| 166 | R07.0 | เจ็บคอ ปวดเข่า 2 วัน แน่นท้อง |
| 167 | R11.9 | อาเจียน 3 ครั้ง |
| 168 | R51,R50.9 | ปวดศีรษะ ปวดตามตัว และไข้ ได้ 1 วัน |

| ชุดข้อมูลที่ไม่ผ่านการตัดคำ (Non-segmentation Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 169 | R50.9 | มีแผลที่เท้าซ้าย และมีไข้ 2 วันก่อนมาโรงพยาบาล |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี longest Matching (LM segmentation-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 1 | R00.0 | หัวใจ เต้น เร็ว |
| 2 | R00.0 | ใจเต้น เร็ว |
| 3 | R00.1 | หัวใจ เต้น ช้า |
| 4 | R00.2 | ใจสั่น |
| 5 | R00.2 | ใจเต้น |
| 6 | R04.0 | เลือดกำเดา ไหล |
| 7 | R04.0 | เลือดกำเดา ออก |
| 8 | R04.2 | ไอ เป็น เลือด |
| 9 | R04.2 | เสมหะ มี เลือด |
| 10 | R04.2 | เสลด มี เลือด |
| 11 | R05 | ไอ |
| 12 | R06.0 | หายใจ ลำบาก |
| 13 | R06.0 | หายใจ ไม่ทัน |
| 14 | R06.0 | หายใจ สั้น |
| 15 | R06.6 | สะอึก |
| 16 | R06.7 | จาม |
| 17 | R07.0 | เจ็บ ใน คอ |
| 18 | R07.0 | เจ็บ ใน ลำคอ |
| 19 | R07.0 | เจ็บ คอ |
| 20 | R07.1 | เจ็บ หน้าอก เวลา หายใจ |
| 21 | R07.4 | เจ็บ หน้าอก |
| 22 | R10.0 | ปวดท้อง เฉียบพลัน |
| 23 | R10.1 | ปวดท้อง เฉพาะ ท้อง ส่วน บน |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี longest Matching (LM segmentation-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 24 | R10.1 | ปวดท้อง เฉพาะ ส่วน บน |
| 25 | R10.1 | ปวดท้อง บริเวณ ลิ้น ปี่ |
| 26 | R10.1 | ปวดท้อง ตรง ส่วน บน |
| 27 | R10.2 | ปวด อุ้ง เชิงกราน และ ฝีเย็บ |
| 28 | R10.2 | ปวด อุ้ง เชิงกราน |
| 29 | R10.2 | ปวด ฝีเย็บ |
| 30 | R10.3 | ปวดท้อง น้อย |
| 31 | R10.4 | ปวดท้อง |
| 32 | R11.9 | คลื่นไส้ และ อาเจียน |
| 33 | R11.9 | คลื่นไส้ |
| 34 | R11.9 | อาเจียน |
| 35 | R14 | ท้องอืด |
| 36 | R14 | ท้อง มี แก๊ส |
| 37 | R14 | ท้องเฟ้อ |
| 38 | R14 | ท้องขึ้น |
| 39 | R25.1 | อาการ สั่น |
| 40 | R25.1 | สั่น |
| 41 | R25.2 | ตะคริว และ กล้ามเนื้อ หด เกร็ง |
| 42 | R25.2 | ตะคริว |
| 43 | R25.2 | กล้ามเนื้อ หด เกร็ง |
| 44 | R25.2 | กล้ามเนื้อ เกร็ง |
| 45 | R26.2 | เดิน ลำบาก |
| 46 | R26.2 | เดิน ยาก |
| 47 | R26.3 | เคลื่อนที่ ไม่ได้ |
| 48 | R30.0 | ถ่ายปัสสาวะ ลำบาก |
| 49 | R30.0 | ฉี่ ลำบาก |
| 50 | R30.1 | ปวด เบ่ง ปัสสาวะ |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี longest Matching (LM segmentation-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 51 | R30.1 | ปวด เบ่ง ฉี่ |
| 52 | R30.9 | ปวด เวลา ถ่ายปัสสาวะ |
| 53 | R30.9 | ปวด ขณะ ถ่ายปัสสาวะ |
| 54 | R30.9 | ปวด เวลา ฉี่ |
| 55 | R30.9 | ปวด ขณะ ฉี่ |
| 56 | R31 | ปัสสาวะ เป็น เลือด |
| 57 | R31 | ฉี่ เป็น เลือด |
| 58 | R32 | กลั้น ปัสสาวะ ไม่ได้ |
| 59 | R32 | กลั้น ฉี่ ไม่ได้ |
| 60 | R33 | ปัสสาวะ ไม่ ออก |
| 61 | R33 | ฉี่ ไม่ ออก |
| 62 | R34 | ไม่มี ปัสสาวะ และ ปัสสาวะ น้อย |
| 63 | R34 | ไม่มี ปัสสาวะ |
| 64 | R34 | ปัสสาวะ น้อย |
| 65 | R34 | ไม่มี ฉี่ |
| 66 | R34 | ฉี่ น้อย |
| 67 | R35 | ปัสสาวะ มาก |
| 68 | R35 | ฉี่ มาก |
| 69 | R42 | เวียน ศีรษะ |
| 70 | R42 | เวียนหัว |
| 71 | R50.9 | ไข้ |
| 72 | R51 | ปวดศีรษะ |
| 73 | R51 | ปวดหัว |
| 74 | R53 | เหนื่อยล้า |
| 75 | R53 | อ่อนแรง |
| 76 | R53 | อิดโรย |
| 77 | R53 | อ่อนเปลี้ย |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี longest Matching (LM segmentation-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 78 | R53 | อ่อนเพลีย |
| 79 | R55 | เป็นลม |
| 80 | R56.0 | การชัก จาก ไข้ สูง |
| 81 | R56.0 | ชัก จาก ไข้ สูง |
| 82 | R57.9 | ช็อก |
| 83 | R04.2,R07.0 | ไอ เป็น เลือด และ เจ็บ คอ |
| 84 | R05,R50.9 | ไอ และ มีไข้ |
| 85 | R55,R57.9 | เป็นลม และ ช็อก |
| 86 | R00.0 | มี อาการ ใจเต้น เป็น มา 3 วัน |
| 87 | R04.2 | มี อาการ ไอ ออก มา เป็น เลือด |
| 88 | R04.2 | มี อาการ ไอ และ มี เลือด ใน เสลด |
| 89 | R06.0 | หายใจ ยาก |
| 90 | R07.0 | เจ็บ ใน คอ |
| 91 | R07.1 | เจ็บ หน้าอก ขณะ หายใจ |
| 92 | R07.1 | เจ็บ หน้าอก เมื่อ หายใจ |
| 93 | R07.4 | เจ็บ บริเวณ หน้าอก |
| 94 | R07.4 | เจ็บ หน้าอก |
| 95 | R10.1 | ปวด บริเวณ ลิ้น ปี่ |
| 96 | R10.1 | ปวด ลิ้น ปี่ |
| 97 | R14 | มี แก๊ส ใน ท้อง |
| 98 | R25.2 | เป็น ตะคริว |
| 99 | R26.3 | เคลื่อนที่ ลำบาก |
| 100 | R31 | มี เลือด ใน ฉี่ |
| 101 | R34 | มี ฉี่ น้อย |
| 102 | R50.9 | เป็น ไข้ |
| 103 | R50.9 | มี ไข้ |
| 104 | R51 | มี อาการ ปวด ที่ ศีรษะ |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี longest Matching (LM segmentation-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 105 | R51,R50.9,R06.7 | มี อาการ ปวดหัว เป็นไข้ และ จาม |
| 106 | R10.4,R30.0 | มี อาการ ปวดท้อง และ ฉี่ ลำบาก |
| 107 | R00.2,R04.0 | ใจเต้น และ เลือดกำเดา ไหล |
| 108 | R25.2,R26.2 | เป็น ตะคริว ที่ ขา และ เดิน ลำบาก |
| 109 | R50.9,R06.7,R04.2 | เป็นไข้ จาม และ ใน เสลด มี เลือด |
| 110 | R51,R11.9 | ปวดหัว คลื่นไส้ และ อาเจียน |
| 111 | R30.9,R34 | ปวด เวลา ปัสสาวะ และ มี ฉี่ น้อย |
| 112 | R25.1,R50.9,R07.4 | มี อาการ สั่น มีไข้ และ เจ็บ หน้าอก |
| 113 | R07.4,R06.0 | เจ็บ หน้าอก และ หายใจ ลำบาก |
| 114 | R10.1,R11.9 | ปวด บริเวณ ลิ้น ปี่ มี อาการ คลื่นไส้ และ อาเจียน |
| 115 | R07.0,R06.0 | มี อาการ เจ็บ ใน ลำคอ และ หายใจ ลำบาก |
| 116 | R10.4,R14 | ปวดท้อง และ ท้องอืด |
| 117 | R51,R05,R06.7 | มี อาการ ปวดหัว ไอ และ จาม |
| 118 | R25.1,R57.9 | มี อาการ สั่น และ ช็อก |
| 119 | R51,R53,R50.9 | ปวดหัว อ่อนเพลีย และ มีไข้ |
| 120 | R10.4 | ปวดท้อง |
| 121 | R07.4 | นาย Smith อายุ 70 ปี เข้าพบ เพื่อ ตรวจสอบ อาการ เจ็บ หน้าอก ที่ มากขึ้น |
| 122 | R06.0 | ลิ้น บวม และ หายใจ และ กลืน ลำบาก |
| 123 | R06.0 | หายใจ ไม่ทัน |
| 124 | R07.4 | ผู้หญิง อายุ 3 6 มีอา การ เจ็บ หน้าอก เฉียบพลัน หลัง ประสบ ประสบอุบัติเหตุ ทาง รถยนต์ |
| 125 | R06.0 | ฉัน หายใจ ไม่ทัน ฉัน หยุด ไอ ไม่ได้ |
| 126 | R10.4 | ปวดท้อง และ ท้องร่วง เป็นเวลา 3 สัปดาห์ |
| 127 | R06.0 | ฉัน หายใจ เสียงดัง และ หายใจ ไม่ทัน |
| 128 | R50.9,R05 | นาย H อายุ 50 ปี ผู้ป่วย โรค AI DS เข้าพบ เพื่อ ตรวจ อาการ เป็นไข้ หนาว ไอ เป็นเวลา 3 วัน |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี longest Matching (LM segmentation-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 128 | R06.0 | หายใจ ลำบาก |
| 130 | R07.4 | เจ็บหน้าอก |
| 131 | R10.4,R11.9 | ปวดท้อง คลื่นไส้ อาเจียน |
| 132 | R50.9 | ไข้ |
| 133 | R11.9,R06.6 | คลื่นไส้ ท้องร่วง แน่นหน้าอก สะอึก |
| 134 | R07.4,R11.9 | เจ็บ หน้าอก อาเจียน |
| 135 | R05 | ไอ |
| 136 | R42 | เวียนหัว |
| 137 | R30.0 | ปัสสาวะ ลำบาก |
| 138 | R53 | อ่อนแรง |
| 139 | R51 | ปวดหัว |
| 140 | R04.2 | ไอ เป็น เลือด |
| 141 | R31 | มี เลือด ใน ปัสสาวะ |
| 142 | R42 | อาเจียน |
| 143 | R00.2 | ใจสั่น |
| 144 | R55 | เป็นลม |
| 145 | R32 | กลั้น ปัสสาวะ ไม่ได้ |
| 146 | R10.1,R50.9 | ผู้ป่วย บอกเล่า เกี่ยวกับ อาการ ปวด ลิ้น ปี่ พร้อมกับ มีไข้ และ หนาว เป็นช่วง เวลา 3 สัปดาห์ |
| 147 | R10.1,R05 | เด็กผู้ชาย อายุ 12 ปี มี อาการ ไข้ ไอ และ หายใจ เสียง ฮืด ๆ แย่ ลง ใน ช่วงเวลา 3 วันที่ ผ่าน มา |
| 148 | R42 | เวียนหัว และ ขา ขวา อ่อนแรง |
| 149 | R06.0,R50.9,R05 | หายใจ สั้น มีไข้ ไอ |
| 150 | R06.0 | ผู้หญิง อายุ 31 ปี เข้าพบ ใน วันนี้ ด้วย อาการ หายใจ สั้น อย่างรุนแรง |
| 151 | R07.4 | ผู้ชาย อายุ 55 ปี เข้าพบ ใน วันนี้ ด้วย อาการ เจ็บ หน้าอก อย่างรุนแรง |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี longest Matching (LM segmentation-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 152 | R05,R50.9,R06.0 | ผู้ชาย อายุ 70 ปี ป่วย เป็นโรค หอบหืด กรด ไหล ย้อน และ ปอดบวม มี อาการไอ มีไข้ และ หายใจ สั้น |
| 153 | R50.9 | ปวดหู และ มีไข้ เป็นเวลา 2 วัน |
| 154 | R30.9 | ปวด ขณะ ปัสสาวะ เริ่ม เป็น เมื่อวาน |
| 155 | R07.0,R50.9 | เจ็บ คอ และ เป็น ไข้ เป็นเวลา 24 ชั่วโมง |
| 156 | R07.4 | เจ็บ หน้าอก เกิดขึ้น ตอนเช้า วันนี้ |
| 157 | R10.4 | ปวดท้อง 4 ชั่วโมง ก่อน มา โรงพยาบาล |
| 158 | R51 | ปวดศีรษะ มา 3 วัน ก่อน มา โรงพยาบาล |
| 159 | R50.9,R05 | มีไข้ ไอ เป็นมา 3 วัน ก่อน มา โรงพยาบาล |
| 160 | R07.0 | เจ็บ คอ 2 วันก่อน มา โรงพยาบาล |
| 161 | R10.4 | ปวดท้อง ถ่าย เหลว 7 ชั่วโมง ก่อน มา โรงพยาบาล |
| 162 | R10.1 | ปวดท้อง จุก แน่น ลิ้น ปี่ 1 อาทิตย์ ก่อน มา โรงพยาบาล |
| 163 | R50.9 | มีไข้ ถ่าย เหลว 6 ครั้ง เป็น ก่อน มา 1 วัน |
| 164 | R50.9,R05 | ไข้ ไอ เป็นมา 2 วัน |
| 165 | R05 | ไอ เมื่อคืน |
| 166 | R07.0 | เจ็บ คอ ปวด เข่า 2 วัน แน่นท้อง |
| 167 | R11.9 | อาเจียน 3 ครั้ง |
| 168 | R51,R50.9 | ปวดศีรษะ ปวด ตามตัว และ ไข้ ได้ 1 วัน |
| 169 | R50.9 | มี แผล ที่ เท้า ซ้าย และ มีไข้ 2 วันก่อน มา โรงพยาบาล |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี Two-level tokenization (2LT-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 1 | R00.0 | หัว ใจ เต้น เร็ว |
| 2 | R00.0 | ใจ เต้น เร็ว |
| 3 | R00.1 | หัว ใจ เต้น ช้า |
| 4 | R00.2 | ใจ สั่น |
| 5 | R00.2 | ใจ เต้น |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี Two-level tokenization (2LT-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 6 | R04.0 | เลือด กำเดา ไหล |
| 7 | R04.0 | เลือด กำเดา ออก |
| 8 | R04.2 | ไอ เป็น เลือด |
| 9 | R04.2 | เสมหะ มี เลือด |
| 10 | R04.2 | เสลด มี เลือด |
| 11 | R05 | ไอ |
| 12 | R06.0 | หาย ใจ ลำ บาก |
| 13 | R06.0 | หาย ใจ ไม่ทัน |
| 14 | R06.0 | หาย ใจ สั้น |
| 15 | R06.6 | สะอึก |
| 16 | R06.7 | จาม |
| 17 | R07.0 | เจ็บ ใน คอ |
| 18 | R07.0 | เจ็บ ใน ลำ คอ |
| 19 | R07.0 | เจ็บ คอ |
| 20 | R07.1 | เจ็บ หน้าอก เวลา หาย ใจ |
| 21 | R07.4 | เจ็บ หน้าอก |
| 22 | R10.0 | ปวด ท้อง เฉียบพลัน |
| 23 | R10.1 | ปวด ท้อง เฉพาะ ท้อง ส่วน บน |
| 24 | R10.1 | ปวด ท้อง เฉพาะ ส่วน บน |
| 25 | R10.1 | ปวด ท้อง บริเวณ ลิ้น ปี่ |
| 26 | R10.1 | ปวด ท้อง ตรง ส่วน บน |
| 27 | R10.2 | ปวด อุ้ง เชิงกราน และ ฝีเย็บ |
| 28 | R10.2 | ปวด อุ้ง เชิงกราน |
| 29 | R10.2 | ปวด ฝีเย็บ |
| 30 | R10.3 | ปวด ท้อง น้อย |
| 31 | R10.4 | ปวด ท้อง |
| 32 | R11.9 | คลื่นไส้ และ อาเจียน |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี Two-level tokenization (2LT-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 33 | R11.9 | คลื่นไส้ |
| 34 | R11.9 | อาเจียน |
| 35 | R14 | ท้อง อืด |
| 36 | R14 | ท้อง มี แก๊ส |
| 37 | R14 | ท้องเฟ้อ |
| 38 | R14 | ท้อง ขึ้น |
| 39 | R25.1 | อาการ สั่น |
| 40 | R25.1 | สั่น |
| 41 | R25.2 | ตะคริว และ กล้ามเนื้อ หด เกร็ง |
| 42 | R25.2 | ตะคริว |
| 43 | R25.2 | กล้ามเนื้อ หด เกร็ง |
| 44 | R25.2 | กล้ามเนื้อ เกร็ง |
| 45 | R26.2 | เดิน ลำ บาก |
| 46 | R26.2 | เดิน ยาก |
| 47 | R26.3 | เคลื่อนที่ ไม่ ได้ |
| 48 | R30.0 | ถ่าย ปัสสาวะ ลำ บาก |
| 49 | R30.0 | ฉี่ ลำ บาก |
| 50 | R30.1 | ปวด เบ่ง ปัสสาวะ |
| 51 | R30.1 | ปวด เบ่ง ฉี่ |
| 52 | R30.9 | ปวด เวลา ถ่าย ปัสสาวะ |
| 53 | R30.9 | ปวด ขณะ ถ่าย ปัสสาวะ |
| 54 | R30.9 | ปวด เวลา ฉี่ |
| 55 | R30.9 | ปวด ขณะ ฉี่ |
| 56 | R31 | ปัสสาวะ เป็น เลือด |
| 57 | R31 | ฉี่ เป็น เลือด |
| 58 | R32 | กลั้น ปัสสาวะ ไม่ ได้ |
| 59 | R32 | กลั้น ฉี่ ไม่ ได้ |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี Two-level tokenization (2LT-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 60 | R33 | ปัสสาวะ ไม่ ออก |
| 61 | R33 | ฉี่ ไม่ ออก |
| 62 | R34 | ไม่ มี ปัสสาวะ และ ปัสสาวะ น้อย |
| 63 | R34 | ไม่ มี ปัสสาวะ |
| 64 | R34 | ปัสสาวะ น้อย |
| 65 | R34 | ไม่ มี ฉี่ |
| 66 | R34 | ฉี่ น้อย |
| 67 | R35 | ปัสสาวะ มาก |
| 68 | R35 | ฉี่ มาก |
| 69 | R42 | เวียน ศีรษะ |
| 70 | R42 | เวียน หัว |
| 71 | R50.9 | ไข้ |
| 72 | R51 | ปวด ศีรษะ |
| 73 | R51 | ปวด หัว |
| 74 | R53 | เหนื่อยล้า |
| 75 | R53 | อ่อนแรง |
| 76 | R53 | อิดโรย |
| 77 | R53 | อ่อนเปลี้ย |
| 78 | R53 | อ่อนเพลีย |
| 79 | R55 | เป็นลม |
| 80 | R56.0 | การชัก จาก ไข้ สูง |
| 81 | R56.0 | ชัก จาก ไข้ สูง |
| 82 | R57.9 | ช็อก |
| 83 | R04.2,R07.0 | ไอ เป็น เลือด และ เจ็บ คอ |
| 84 | R05,R50.9 | ไอ และ มี ไข้ |
| 85 | R55,R57.9 | เป็นลม และ ช็อก |
| 86 | R00.0 | มี อาการ ใจ เต้น เป็น มา 3 วัน |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี Two-level tokenization (2LT-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 87 | R04.2 | มี อาการ ไอ ออก มา เป็น เลือด |
| 88 | R04.2 | มี อาการ ไอ และ มี เลือด ใน เสลด |
| 89 | R06.0 | หาย ใจ ยาก |
| 90 | R07.0 | เจ็บ ใน คอ |
| 91 | R07.1 | เจ็บ หน้าอก ขณะ หาย ใจ |
| 92 | R07.1 | เจ็บ หน้าอก เมื่อ หาย ใจ |
| 93 | R07.4 | เจ็บ บริเวณ หน้าอก |
| 94 | R07.4 | เจ็บ หน้าอก |
| 95 | R10.1 | ปวด บริเวณ ลิ้น ปี่ |
| 96 | R10.1 | ปวด ลิ้น ปี่ |
| 97 | R14 | มี แก๊ส ใน ท้อง |
| 98 | R25.2 | เป็น ตะคริว |
| 99 | R26.3 | เคลื่อนที่ ลำ บาก |
| 100 | R31 | มี เลือด ใน ฉี่ |
| 101 | R34 | มี ฉี่ น้อย |
| 102 | R50.9 | เป็น ไข้ |
| 103 | R50.9 | มี ไข้ |
| 104 | R51 | มี อาการ ปวด ที่ ศีรษะ |
| 105 | R51,R50.9,R06.7 | มี อาการ ปวด หัว เป็น ไข้ และ จาม |
| 106 | R10.4,R30.0 | มี อาการ ปวด ท้อง และ ฉี่ ลำ บาก |
| 107 | R00.2,R04.0 | ใจ เต้น และ เลือด กำเดา ไหล |
| 108 | R25.2,R26.2 | เป็น ตะคริว ที่ขา และ เดิน ลำ บาก |
| 109 | R50.9,R06.7,R04.2 | เป็น ไข้ จาม และ ใน เสลด มี เลือด |
| 110 | R51,R11.9 | ปวด หัว คลื่นไส้ และ อาเจียน |
| 111 | R30.9,R34 | ปวด เวลา ปัสสาวะ และ มี ฉี่ น้อย |
| 112 | R25.1,R50.9,R07.4 | มี อาการ สั่น มี ไข้ และ เจ็บ หน้าอก |
| 113 | R07.4,R06.0 | เจ็บ หน้าอก และ หาย ใจ ลำ บาก |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี Two-level tokenization (2LT-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 114 | R10.1,R11.9 | ปวด บริเวณ ลิ้น ปี่ มี อาการ คลื่นไส้ และ อาเจียน |
| 115 | R07.0,R06.0 | มี อาการ เจ็บ ใน ลำ คอ และ หาย ใจ ลำ บาก |
| 116 | R10.4,R14 | ปวด ท้อง และ ท้อง อืด |
| 117 | R51,R05,R06.7 | มี อาการ ปวด หัว ไอ และ จาม |
| 118 | R25.1,R57.9 | มี อาการ สั่น และ ช็อก |
| 119 | R51,R53,R50.9 | ปวด หัว อ่อนเพลีย และ มี ไข้ |
| 120 | R10.4 | ปวด ท้อง |
| 121 | R07.4 | นาย Smith อายุ 70 ปี เข้าพบเพื่อตรวจสอบ อาการ เจ็บ หน้าอก ที่ มาก ขึ้น |
| 122 | R06.0 | ลิ้น บวม และ หาย ใจ และ กลืน ลำ บาก |
| 123 | R06.0 | หาย ใจ ไม่ทัน |
| 124 | R07.4 | ผู้หญิง อายุ 36 มี อาการ เจ็บ หน้าอก เฉียบพลัน หลัง ประสบ ประสบอุบัติเหตุ ทาง รถยนต์ |
| 125 | R06.0 | ฉัน หาย ใจ ไม่ทัน ฉันหยุด ไอ ไม่ ได้ |
| 126 | R10.4 | ปวด ท้อง และ ท้อง ร่วง เป็น เวลา 3 สัปดาห์ |
| 127 | R06.0 | ฉัน หาย ใจ เสียงดัง และ หาย ใจ ไม่ทัน |
| 128 | R50.9,R05 | นาย H อายุ 50 ปี ผู้ป่วยโรค AIDS เข้าพบเพื่อตรวจ อาการ เป็น ไข้ หนาว ไอ เป็น เวลา 3 วัน |
| 128 | R06.0 | หาย ใจ ลำ บาก |
| 130 | R07.4 | เจ็บ หน้าอก |
| 131 | R10.4,R11.9 | ปวด ท้อง คลื่นไส้ อาเจียน |
| 132 | R50.9 | ไข้ |
| 133 | R11.9,R06.6 | คลื่นไส้ ท้อง ร่วง แน่น หน้าอก สะอึก |
| 134 | R07.4,R11.9 | เจ็บ หน้าอก อาเจียน |
| 135 | R05 | ไอ |
| 136 | R42 | เวียน หัว |
| 137 | R30.0 | ปัสสาวะ ลำ บาก |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี Two-level tokenization (2LT-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 138 | R53 | อ่อนแรง |
| 139 | R51 | ปวด หัว |
| 140 | R04.2 | ไอ เป็น เลือด |
| 141 | R31 | มี เลือด ใน ปัสสาวะ |
| 142 | R42 | อาเจียน |
| 143 | R00.2 | ใจ สั่น |
| 144 | R55 | เป็นลม |
| 145 | R32 | กลั้น ปัสสาวะ ไม่ ได้ |
| 146 | R10.1,R50.9 | ผู้ป่วยบอกเล่าเกี่ยวกับ อาการ ปวด ลิ้น ปี่ พร้อมกับ มี ไข้ และ หนาว เป็น ช่วง เวลา 3 สัปดาห์ |
| 147 | R10.1,R05 | เด็กผู้ชายอายุ 12 ปี มี อาการ ไข้ ไอ และ หาย ใจ เสียง ฮืด ๆ แย่ลง ใน ช่วง เวลา 3 วันที่ผ่านมา |
| 148 | R42 | เวียน หัว และ ขาขวา อ่อนแรง |
| 149 | R06.0,R50.9,R05 | หาย ใจ สั้น มี ไข้ ไอ |
| 150 | R06.0 | ผู้หญิงอายุ 31 ปี เข้าพบ ใน วันนี้ด้วย อาการ หาย ใจ สั้น อย่างรุนแรง |
| 151 | R07.4 | ผู้ชายอายุ 55 ปีเข้าพบ ใน วันนี้ด้วย อาการ เจ็บ หน้าอก อย่างรุนแรง |
| 152 | R05,R50.9,R06.0 | ผู้ชายอายุ 70 ปี ป่วย เป็น โรคหอบหืด กรด ไหล ย้อน และ ปอดบวม มี อาการ ไอ มี ไข้ และ หาย ใจ สั้น |
| 153 | R50.9 | ปวด หู และ มี ไข้ เป็น เวลา 2 วัน |
| 154 | R30.9 | ปวด ขณะ ปัสสาวะ เริ่ม เป็น เมื่อวาน |
| 155 | R07.0,R50.9 | เจ็บ คอ และ เป็น ไข้ เป็น เวลา 24 ชั่วโมง |
| 156 | R07.4 | เจ็บ หน้าอก เกิด ขึ้น ตอนเช้าวันนี้ |
| 157 | R10.4 | ปวด ท้อง 4 ชั่วโมง ก่อนมาโรงพยาบาล |
| 158 | R51 | ปวด ศีรษะ มา 3 วัน ก่อนมาโรงพยาบาล |
| 159 | R50.9,R05 | มี ไข้ ไอ เป็น มา 3 วัน ก่อนมาโรงพยาบาล |
| 160 | R07.0 | เจ็บ คอ 2 วันก่อนมาโรงพยาบาล |

| ชุดข้อมูลที่ผ่านการตัดคำด้วยวิธี Two-level tokenization (2LT-based Dataset) | | |
|---|---|---|
| *แถว* | *รหัส ICD* | *ข้อความ* |
| 161 | R10.4 | ปวด ท้อง ถ่าย เหลว 7 ชั่วโมง ก่อนมาโรงพยาบาล |
| 162 | R10.1 | ปวด ท้อง จุกแน่น ลิ้น ปี่ 1 อาทิตย์ก่อนมาโรงพยาบาล |
| 163 | R50.9 | มี ไข้ ถ่าย เหลว 6 ครั้ง เป็น ก่อนมา 1 วัน |
| 164 | R50.9,R05 | ไข้ ไอ เป็น มา 2 วัน |
| 165 | R05 | ไอ เมื่อคืน |
| 166 | R07.0 | เจ็บ คอ ปวด เข่า 2 วัน แน่น ท้อง |
| 167 | R11.9 | อาเจียน 3 ครั้ง |
| 168 | R51,R50.9 | ปวด ศีรษะ ปวด ตามตัว และ ไข้ ได้ 1 วัน |
| 169 | R50.9 | มี แผลที่เท้าซ้าย และ มี ไข้ 2 วันก่อนมาโรงพยาบาล |

**ก.2 ผลการจำแนกอาการจากการทดลอง**

ผลการทดลองประกอบด้วย ผลการจำแนก u1 u2 และ u3 ของการเพิ่มปริมาณ tp จากหัวข้อ 5.2.1 และ ผลการจำแนกของการลดปริมาณ *fn* จากการทดลอง 5.2.2 โดยไม่ได้

| แถว | ผลการจำแนก u2 หัวข้อที่ 5.2.1 |
|---|---|
| 83 | R04.2 R07.0 |
| 84 | R05 R50.9 |
| 85 | R55 R57.9 |
| 86 | R00.2 R25.1 R31 |
| 87 | R04.2 R25.1 R33 |
| 88 | R04.2 R05 R25.1 R07.0 |
| 89 | R06.0 |
| 90 | R07.0 |
| 91 | R30.9 R06.0 R07.1 R07.4 |
| 92 | R06.0 R07.1 R07.4 |
| 93 | R07.4 R10.1 |
| 94 | R07.4 |
| 95 | R10.1 |
| 96 | |
| 97 | R14 R07.0 |
| 98 | R25.2 R31 |
| 99 | R26.3 R30.0 |
| 100 | R04.2 R07.0 |
| 101 | R34 |
| 102 | R50.9 R31 |
| 103 | R50.9 |
| 104 | R25.1 |
| 105 | R06.7 R25.1 R51 R31 |
| 106 | R10.4 R25.1 R30.0 |
| 107 | R00.2 R04.0 |

| แถว | ผลการจำแนก u2 หัวข้อที่ 5.2.1 |
|---|---|
| 108 | R25.2 R26.2 R31 |
| 109 | R06.7 R31 R07.0 |
| 110 | R11.9 R51 |
| 111 | R30.9 R34 |
| 112 | R07.4 R25.1 R50.9 |
| 113 | R06.0 R07.1 R07.4 |
| 114 | R10.1 R11.9 R25.1 |
| 115 | R07.0 R25.1 R06.0 |
| 116 | R10.4 R14 |
| 117 | R05 R06.7 R25.1 R51 |
| 118 | R25.1 R57.9 |
| 119 | R50.9 R51 R53 |
| 120 | R10.4 |
| 121 | R07.4 R25.1 |
| 122 | R06.0 |
| 123 | R06.0 |
| 124 | R07.4 R10.0 |
| 125 | R05 R06.0 R32 |
| 126 | R10.4 R31 |
| 127 | R06.0 |
| 128 | R25.1 R05 |
| 128 | R06.0 |
| 130 | R07.4 |
| 131 | R10.4 R11.9 |
| 132 | R50.9 |
| 133 | R06.6 R07.4 R11.9 |
| 134 | R07.4 R11.9 |
| 135 | R05 |

| แถว | ผลการจำแนก u2 หัวข้อที่ 5.2.1 |
|---|---|
| 136 | R42 |
| 137 | R30.0 |
| 138 | R53 |
| 139 | R51 |
| 140 | R04.2 |
| 141 | R04.2 R07.0 |
| 142 | R11.9 |
| 143 | R00.2 |
| 144 | R55 |
| 145 | R32 |
| 146 | R25.1 R30.9 R50.9 R31 |
| 147 | R05 R06.0 R25.1 R26.2 R50.9 R07.0 |
| 148 | R42 R53 |
| 149 | R05 R06.0 R50.9 |
| 150 | R06.0 R25.1 R34 R07.0 |
| 151 | R07.4 R25.1 R34 R07.0 |
| 152 | R06.0 R50.9 R31 |
| 153 | R50.9 R51 R31 |
| 154 | R00.0 R30.9 R31 |
| 155 | R07.0 R31 |
| 156 | R07.4 |
| 157 | R10.4 |
| 158 | R51 |
| 159 | R50.9 R05 |
| 160 | R07.0 |
| 161 | R10.4 |
| 162 | R10.4 |
| 163 | R50.9 R31 |

| แถว | ผลการจำแนก u2 หัวข้อที่ 5.2.1 |
|---|---|
| 164 | R50.9 R05 |
| 165 | R05 |
| 166 | R07.0 |
| 167 | R11.9 |
| 168 | R32 R50.9 R51 |
| 169 | R50.9 |

| แถว | ผลการจำแนก u1 จากหัวข้อที่ 5.1.1 |
|---|---|
| 83 | R04.2 R07.0 |
| 84 | R05 R50.9 |
| 85 | R55 R57.9 |
| 86 | R25.1 R31 |
| 87 | R04.2 R25.1 R33 |
| 88 | R04.2 R25.1 R05 |
| 89 | R06.0 |
| 90 | R07.0 |
| 91 | R06.0 R07.4 R30.9 R07.1 |
| 92 | R06.0 R07.4 R07.1 |
| 93 | R07.4 R10.1 |
| 94 | R07.4 |
| 95 | R10.1 R51 |
| 96 | R51 |
| 97 | R14 |
| 98 | R25.2 R31 |
| 99 | R26.3 R30.0 |
| 100 | R04.2 |
| 101 | R34 |
| 102 | R31 R50.9 |

| แถว | ผลการจำแนก u1 จากหัวข้อที่ 5.1.1 |
|---|---|
| 103 | R50.9 |
| 104 | R25.1 R51 |
| 105 | R06.7 R25.1 R31 R50.9 R51 |
| 106 | R10.4 R25.1 R30.0 |
| 107 | R04.0 |
| 108 | R25.2 R26.2 R31 |
| 109 | R06.7 R31 R50.9 |
| 110 | R11.9 R51 |
| 111 | R30.9 R10.3 R34 |
| 112 | R07.4 R25.1 R50.9 |
| 113 | R06.0 R07.4 R07.1 |
| 114 | R10.1 R11.9 R25.1 R51 |
| 115 | R06.0 R07.0 R25.1 |
| 116 | R10.4 |
| 117 | R05 R06.7 R25.1 R51 |
| 118 | R25.1 R57.9 |
| 119 | R50.9 R51 R53 |
| 120 | R10.4 |
| 121 | R07.4 R25.1 R35 |
| 122 | R06.0 |
| 123 | R06.0 |
| 124 | R07.4 R10.0 R25.1 |
| 125 | R05 R06.0 R32 |
| 126 | R10.4 R30.9 R31 |
| 127 | R06.0 |
| 128 | R05 R25.1 R50.9 R30.9 |
| 128 | R06.0 |
| 130 | R07.4 |

| แถว | ผลการจำแนก u1 จากหัวข้อที่ 5.1.1 |
|---|---|
| 131 | R10.4 R11.9 |
| 132 | R50.9 |
| 133 | R06.6 R07.4 R11.9 R14 |
| 134 | R07.4 R11.9 |
| 135 | R05 |
| 136 | R42 |
| 137 | R30.0 |
| 138 | R53 |
| 139 | R51 |
| 140 | R04.2 |
| 141 | R04.2 |
| 142 | R11.9 |
| 143 | R00.2 |
| 144 | R55 |
| 145 | R32 |
| 146 | R25.1 R30.9 R31 R50.9 |
| 147 | R05 R25.1 R26.2 R50.9 R06.0 |
| 148 | R42 R53 |
| 149 | R05 R06.0 R50.9 |
| 150 | R06.0 R25.1 R34 |
| 151 | R07.4 R25.1 R34 |
| 152 | R05 R06.0 R25.1 R50.9 |
| 153 | R30.9 R31 R50.9 |
| 154 | R00.0 R30.9 R31 |
| 155 | R07.0 R31 R50.9 R30.9 |
| 156 | R07.4 |
| 157 | R10.4 |
| 158 | R51 |

| แถว | ผลการจำแนก u1 จากหัวข้อที่ 5.1.1 |
|---|---|
| 159 | R05 R50.9 |
| 160 | R07.0 |
| 161 | R10.4 |
| 162 | R10.4 |
| 163 | R31 R50.9 |
| 164 | R05 R50.9 |
| 165 | R05 |
| 166 | R07.0 R10.4 |
| 167 | R11.9 |
| 168 | R50.9 R51 R32 |
| 169 | R50.9 |

| แถว | ผลการจำแนก u3 จากหัวข้อที่ 5.2.2 |
|---|---|
| 83 | R07.0 R04.2 |
| 84 | R50.9 R05 |
| 85 | R57.9 R55 |
| 86 | R25.1 R31 R00.2 |
| 87 | R25.1 R33 R04.2 |
| 88 | R25.1 R07.0 R05 R04.2 |
| 89 | R06.0 |
| 90 | R07.0 |
| 91 | R07.4 R07.1 R06.0 R30.9 |
| 92 | R07.1 R06.0 R07.4 |
| 93 | R10.1 R07.4 |
| 94 | R07.4 |
| 95 | R10.1 R51 |
| 96 | R51 |
| 97 | R07.0 R14 |

| แถว | ผลการจำแนก u3 จากหัวข้อที่ 5.2.2 |
|---|---|
| 98 | R31 R25.2 |
| 99 | R26.3 R30.0 |
| 100 | R07.0 R04.2 |
| 101 | R34 |
| 102 | R50.9 R31 |
| 103 | R50.9 |
| 104 | R25.1 R51 |
| 105 | R50.9 R06.7 R31 R51 R25.1 |
| 106 | R25.1 R10.4 R30.0 |
| 107 | R04.0 R00.2 |
| 108 | R31 R25.2 R26.2 |
| 109 | R50.9 R06.7 R31 R07.0 |
| 110 | R51 R11.9 |
| 111 | R10.3 R34 R30.9 |
| 112 | R50.9 R25.1 R07.4 |
| 113 | R07.1 R06.0 R07.4 |
| 114 | R10.1 R25.1 R51 R11.9 |
| 115 | R07.0 R25.1 R06.0 |
| 116 | R10.4 R14 |
| 117 | R06.7 R51 R25.1 R05 |
| 118 | R57.9 R25.1 |
| 119 | R50.9 R53 R51 |
| 120 | R10.4 |
| 121 | R25.1 R07.4 R35 |
| 122 | R06.0 |
| 123 | R06.0 |
| 124 | R25.1 R07.4 R10.0 |
| 125 | R32 R06.0 R05 |

| แถว | ผลการจำแนก u3 จากหัวข้อที่ 5.2.2 |
|---|---|
| 126 | R31 R10.4 R30.9 |
| 127 | R06.0 |
| 128 | R50.9 R25.1 R05 R30.9 |
| 128 | R06.0 |
| 130 | R07.4 |
| 131 | R10.4 R11.9 |
| 132 | R50.9 |
| 133 | R06.6 R07.4 R11.9 R14 |
| 134 | R07.4 R11.9 |
| 135 | R05 |
| 136 | R42 |
| 137 | R30.0 |
| 138 | R53 |
| 139 | R51 |
| 140 | R04.2 |
| 141 | R07.0 R04.2 |
| 142 | R11.9 |
| 143 | R00.2 |
| 144 | R55 |
| 145 | R32 |
| 146 | R50.9 R25.1 R31 R30.9 |
| 147 | R26.2 R50.9 R25.1 R07.0 R06.0 R30.9 |
| 148 | R42 R53 |
| 149 | R50.9 R06.0 R05 |
| 150 | R25.1 R34 R07.0 R06.0 |
| 151 | R25.1 R07.4 R34 R07.0 |
| 152 | R50.9 R25.1 R31 R05 R06.0 |
| 153 | R50.9 R31 R51 R30.9 |

| แถว | ผลการจำแนก u3 จากหัวข้อที่ 5.2.2 |
|---|---|
| 154 | R00.0 R31 R30.9 |
| 155 | R50.9 R07.0 R31 R30.9 |
| 156 | R07.4 |
| 157 | R10.4 |
| 158 | R51 |
| 159 | R50.9 R05 |
| 160 | R07.0 |
| 161 | R10.4 |
| 162 | R10.4 |
| 163 | R50.9 R31 |
| 164 | R50.9 R05 |
| 165 | R05 |
| 166 | R07.0 R10.4 |
| 167 | R11.9 |
| 168 | R50.9 R32 R51 |
| 169 | R07.0 R04.2 |

| แถว | ผลการจำแนกตามหัวข้อที่ 5.3.2 การตัดผลการจำแนกที่คาดว่าเป็น *fn* |
|---|---|
| 83 | R04.2 R07.0 |
| 84 | R50.9 R05 |
| 85 | R57.9 R55 |
| 86 | R00.2 |
| 87 | R04.2 |
| 88 | R04.2 R05 |
| 89 | R06.0 |
| 90 | R07.0 |
| 91 | R07.4 R07.1 |
| 92 | R07.4 R07.1 |

| แถว | ผลการจำแนกตามหัวข้อที่ 5.3.2 การตัดผลการจำแนกที่คาดว่าเป็น *fn* |
|---|---|
| 93 | R07.4 |
| 94 | R07.4 |
| 95 | R10.1 |
| 96 | |
| 97 | R14 |
| 98 | R25.2 |
| 99 | R30.0 |
| 100 | R04.2 |
| 101 | R34 |
| 102 | R50.9 |
| 103 | R50.9 |
| 104 | R51 |
| 105 | R50.9 R06.7 R51 |
| 106 | R10.4 R30.0 |
| 107 | R04.0 R00.2 |
| 108 | R25.2 R26.2 |
| 109 | R50.9 R06.7 R31 |
| 110 | R51 R11.9 |
| 111 | R10.3 R30.9 R34 |
| 112 | R50.9 R07.4 R25.1 |
| 113 | R07.4 R07.1 R06.0 |
| 114 | R10.1 R11.9 |
| 115 | R07.1 R07.0 R06.0 |
| 116 | R10.4 R14 |
| 117 | R51 R06.7 R05 |
| 118 | R57.9 R25.1 |
| 119 | R50.9 R51 R53 |
| 120 | R10.4 |

| แถว | ผลการจำแนกตามหัวข้อที่ 5.3.2 การตัดผลการจำแนกที่คาดว่าเป็น *fn* |
|---|---|
| 121 | R07.4 |
| 122 | R06.0 |
| 123 | R06.0 |
| 124 | R07.4 |
| 125 | R06.0 R05 |
| 126 | R10.4 R30.9 |
| 127 | R06.0 |
| 128 | R50.9 R05 |
| 128 | R06.0 |
| 130 | R07.4 |
| 131 | R10.4 R11.9 |
| 132 | R50.9 |
| 133 | R06.6 R11.9 |
| 134 | R07.4 R11.9 |
| 135 | R05 |
| 136 | R42 |
| 137 | R30.0 |
| 138 | R53 |
| 139 | R51 |
| 140 | R04.2 |
| 141 | R04.2 |
| 142 | R11.9 |
| 143 | R00.2 |
| 144 | R55 |
| 145 | R32 |
| 146 | R50.9 R30.9 |
| 147 | R50.9 R06.0 R05 |
| 148 | R42 R53 |

| แถว | ผลการจำแนกตามหัวข้อที่ 5.3.2 การตัดผลการจำแนกที่คาดว่าเป็น *fn* |
|---|---|
| 149 | R50.9 R06.0 R05 |
| 150 | R06.0 |
| 151 | R07.4 R07.0 |
| 152 | R50.9 R05 R06.0 |
| 153 | R50.9 R30.9 |
| 154 | R31 R30.9 |
| 155 | R50.9 R07.0 |
| 156 | R07.4 |
| 157 | R10.4 |
| 158 | R51 |
| 159 | R50.9 R05 |
| 160 | R07.0 |
| 161 | R10.4 |
| 162 | R10.4 |
| 163 | R50.9 |
| 164 | R50.9 R05 |
| 165 | R05 |
| 166 | R07.0 R10.4 |
| 167 | R11.9 |
| 168 | R50.9 R51 |
| 169 | R50.9 |

# ภาคผนวก ข.

## ผลงานตีพิมพ์

| | |
|---|---|
| เรื่อง | การระบุอาการและอาการแสดงจากข้อความบอกเล่าอาการสำคัญภาษาไทยตามมาตรฐาน ICD-10 โดยขั้นตอน การประมวลผลภาษาธรรมชาติ |
| งานประชุมวิชาการ | The 20th International Computer Science and Engineering Conference 2016 (ICSEC 2016) |
| สถานที่ | Chiangmai, Thailand |
| วันที่ | 14 - 17 December 2016 |

# การระบุอาการและอาการแสดงจากข้อความบอกเล่าอาการสำคัญภาษาไทยตามมาตรฐาน ICD-10 โดยขั้นตอนการประมวลผลภาษาธรรมชาติ

ภวินท์ แซ่คู[1], จารุณี ดวงสุวรรณ[2], และ ลัดดา ปรีชาวีรกุล[3]
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ 15 ถ.กาญจนวนิช อ.หาดใหญ่ จ.สงขลา 90110
E-mail: tamakosan14@gmail.com[1], jarunee.d@psu.ac.th[2], ladda.p@psu.ac.th [3]

**บทคัดย่อ**

บทความนี้นำเสนอกระบวนเพื่อระบุอาการและอาการแสดงจากการสกัดข้อความบอกเล่าอาการผู้ป่วยด้วยกระบวนการประมวลผลภาษาธรรมชาติ สำหรับข้อความบอกเล่าอาการผู้ป่วยหรือ Chief Complains (CCs) ถูกบันทึกอยู่ในรูปแบบภาษาธรรมชาติด้วยภาษาไทย ดังนั้นการที่จะนำสกัดข้อความดังกล่าว และแปลงให้อยู่ให้อยู่ในรูปแบบที่สามารถนำมาใช้ในประมวลผลทางคอมพิวเตอร์ได้มีความยุ่งยากและซับซ้อน เพราะนอกจากจะต้องเข้าใจถึงลักษณะ ข้อจำกัด วิธีการจัดการกับข้อจำกัดของภาษาไทย แล้วยังต้องกำหดนกระบวนการและขั้นตอนสำหรับสกัดอาการและอาการแสดงให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถนำไปประมวลผลได้เช่นกัน

คำสำคัญ: การประมวลผลภาษาธรรมชาติ, บัญชีจำแนกทางสถิติระหว่างประเทศของโรคและปัญหาสุขภาพที่เกี่ยวข้อง, ข้อความบอกเล่าอาการสำคัญ

**Abstract**

This article describes a signs and symptoms extraction from Chief Complains (CCs) in Thai language. CCs describes about what an illness leading the patient to the hospital. There are useful information contained in CCs, but due the barrier in Thai natural language, to collective information from Thai CCs and preparing to use for computation, there need to understand and overcome the circumstance in Thai language and method to extract signs and symptoms to the form ready to computation has been talk in this article.

Keywords: natural language processing, International Statistical Classification of Diseases and Related Health Problems(ICD), Chief complaints

## 1. คำนำ

ข้อความบอกเล่าอาการสำคัญ หรือ Chief Complaints (CCs) เป็นข้อความที่อธิบายถึงอาการสำคัญที่นำผู้ป่วยมายังโรงพยาบาลเพื่อทำการรักษา โดยแพทย์จะเป็นผู้บันทึกข้อมูล CCs ของผู้ป่วย โดยในข้อมูล CCs จะไม่มีการบันทึกอาการสำคัญในรูปแบบของชื่อโรค แต่จะบันทึกเพียงคำอธิบายที่บ่งบอกอาการที่สำคัญที่ได้จากการพูดคุย สอบถามระหว่างแพทย์และผู้ป่วย สำหรับ CCs นั้นเป็นส่วนประกอบหนึ่งของเวชระเบียน (medical record) ซึ่งโดยทั่วไปข้อมูล CCs ดังกล่าวจะถูกบันทึกอยู่ในรูปภาษาธรรมชาติทั้งในรูปแบบของภาษาพูด หรือภาษาเขียน ซึ่งส่งผลให้การที่จะนำข้อมูล CCsไปใช้นั้นเป็นเรื่องยากและจำเป็นต้องอาศัยผู้เชี่ยวชาญในการจำแนกข้อมูลCCs เพื่อทำให้ได้ข้อมูลที่ถูกจำแนกออกมามีคุณภาพ และสามารถนำไปใช้ในระบบทางการแพทย์ที่เกี่ยวข้องได้อย่างมีประสิทธิภาพ รวมถึงสามารถนำข้อมูลที่ได้ไปใช้ในการประมวลผลในระบบคอมพิวเตอร์ได้เช่น ระบบการเก็บข้อมูลทางการแพทย์อิเล็กทรอนิกส์ มีชื่อเรียกว่าเวชระเบียนอิเล็กทรอนิกส์(Electronic health record : EHR หรือ Electronic medical record : EMR) ซึ่งเป็นระบบของการสนับสนุนให้เกิดการใช้ข้อมูลร่วมกันระหว่างโรงพยาบาลต่างๆ เช่นข้อมูลของโรค อาการ และ อาการแสดงใน EHR โดยข้อมูลเหล่านี้จะถูกบันทึกในรูปแบบรหัสของ International Statistical Classification of Diseases and Related Health Problems(ICD) ซึ่งปัจจุบันเป็นฉบับที่10(ICD-10) ทำให้สามารถนำข้อมูลดังกล่าวไปใช้ประมวลผลทางคอมพิวเตอร์ได้โดยไม่เกิดปัญหาจากผลกระทบทางภาษา

## 2. งานวิจัยและทฤษฎีที่เกี่ยวข้อง

ที่ผ่านมางานวิจัย [1] ได้มีการใช้ข้อมูล ICD-9 จากข้อมูลอิเล็กทรอนิกของ 189 โรงพยาบาลในไต้หวันเพื่อการจำแนกประเภทของ CCs และในงานวิจัย [2] ได้เสนอการจำแนกการเฝ้าระวังการระบาดของโรคที่สนับสนุนหลายๆภาษาโดยใช้กรณีศึกษาเป็นภาษาจีน

เมื่อกล่าวถึงคุณลักษณะของภาษาไทยในงานวิจัย [3] ได้อธิบายถึงคุณลักษณะที่สำคัญของภาษาไทยไว้ว่า เป็นภาษาที่มีรูปแบบโครงสร้างในลักษณะของ ประธาน-กริยา-กรรม หรือ Subject-Verb-Object (SVO) ซึ่งมีลักษณะทั่วไปเช่นเดียวกับภาษาอังกฤษ จากคุณลักษณะในงานวิจัยดังกล่าว เมื่อพิจารณาถึงปัญหาในการสื่อถึงปัญหาของการบันทึกข้อมูล CCs นั้นคือการที่แพทย์สามาถบันทึก ข้อมูล อาการหรืออาการแสดง ในภาษาไทย โดยการที่คำหนึ่งคำสามารถแทนได้หลายความหมาย หรือสามารถใช้คำหลายคำแทนความหมายเดียวกันได้ เช่น ปวดหัว เวียนหัว ปวดศีรษะ เวียนศีรษะ ซึ่งคำทั้งสี่คำนี้เป็นการสื่อถึงอาการเดียวกัน หรืออีกตัวอย่างหนึ่งเช่น เจ็บไหล่ ปวดไหล่ โดยวลีเหล่านี้มีรูปแบบโครงสร้างที่เหมือนกันคือ Verb-Object(VO) หรือ Object-Verb(OV) แต่คำที่นำมาใช้ประกอบมีความแตกต่างกัน อย่างไรก็ตามข้อความที่เกิดขึ้นจริงในภาษาไทยอาจมีองค์ประกอบของวลีที่สื่อถึงความหมายเดียวกันและอาจมีรูปแบบอื่นๆซึ่งมีความหมายเดียวกันได้ การแก้ปัญหาถัดมาคือการแก้ปัญหาการตัดคำเนื่องจากภาษาไทยไม่มีสัญลักษณ์ที่ใช้กำหนดขอบเขตคำ ในการแก้ไขปัญหาดังกล่าวได้เสนอแนวทางในงานวิจัย [4] โดยเสนอวิธีการตัดพยางค์เพื่อหาขอบเขตหน้าและขอบเขตหลังของประโยคโดยอาศัยกฎที่สร้างขึ้นตามคุณลักษณะของอักขระภาษาไทย สำหรับงานวิจัยงาน [5] และ [6] ได้นำเสนอวิธีการตัดคำด้วยพจนานุกรม โดยใช้วิธีที่เรียกว่าLongest matching(LM) ซึ่งเป็นวิธีที่จะเลือกตัดคำจากคำที่ยาวที่สุดที่พบในพจนานุกรม ซึ่งหากคำที่ยาวที่สุดที่ถูกเลือก เป็นคำที่ไม่มีอยู่ในพจนานุกรม ระบบก็จะทำการเลือกคำใหม่ซึ่งยาวรองลงมา โดย Maximum Matching(MM) จะทำ LM ก่อนจากนั้นจะกลับไปทำ LM ซ้ำอีกครั้งในแต่ละคำที่ถูกตัด

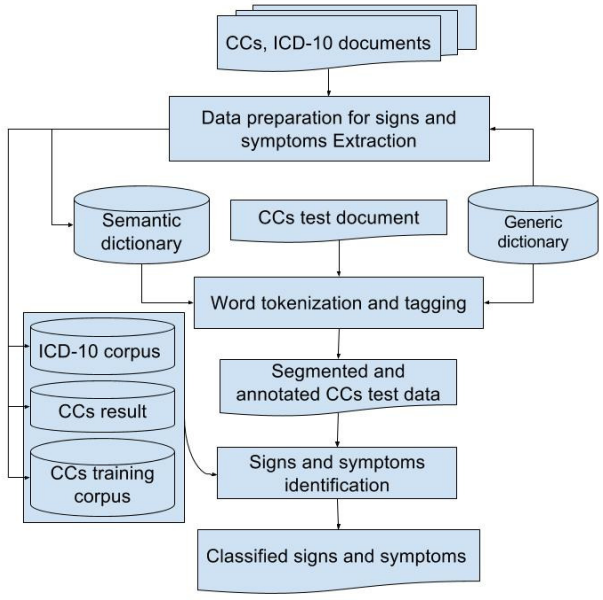
คลังข้อความคือเอกสารอิเล็กทรอนิกส์หรือไฟล์ที่บรรจุข้อความหรือคำในรูปแบบข้อมูลเชิงโครงสร้างโดยมีโครงสร้างแตกต่างกันไปตามบริบทของงานที่นำคลังข้อความดังกล่าวไปใช้ การติดป้าย(tagging) เป็นการกำกับคำศัพท์ลงไปในคลังข้อความซึ่งเป็นวิธีการหนึ่งที่ช่วยในการอธิบายนัยสำคัญของคำนั้นๆต่อบริบทของเรื่องที่สนใจได้อย่างมีประสิทธิภาพ เนื่องจากคำบางคำในบริบทใดบริบทหนึ่งอาจมีความหมายหรือหน้าที่ต่างไปจากการใช้คำดังกล่าวโดยบริบททั่วไป การติดป้ายเพื่ออธิบายคำศัพท์ตามบริบทของเรื่องที่สนใจจึงมีความจำเป็นในการอธิบายถึงความหมายของคำศัพท์ต่อบริบทที่สนใจ

การสกัดข้อมูลสารสนเทศคือการสกัดข้อมูลเชิงโครงสร้างจากข้อมูลที่ไม่เป็นโครงสร้างหรือข้อมูลกึ่งโครงสร้างที่เครื่องคอมพิวเตอร์สามารถเข้าใจได้ ในงานวิจัย [9] ได้แสดงให้เห็นว่างานวิจัยด้านการสกัดข้อมูล (Information Extraction : IE) ได้รับความนิยมในการนำไปประยุกต์ใช้กับงานทางด้านชีวการแพทย์ เนื่องจากงานวิจัยดังกล่าวสามารถนำไปสู่การ ค้นหาข้อมูล การได้ความองค์ความรู้ใหม่ และการสร้างสมมติฐานอย่างมีประสิทธิภาพ ตัวอย่างงานวิจัยด้านการสกัดข้อมูลอื่นๆเช่นงานวิจัย [7] และ [8] ได้นำเสนอการสกัด Name Entity ทางการแพทย์ โดยอาศัยหลักการของ Machine learning ในขณะที่งานวิจัย [9] เป็นงานวิจัยเกี่ยวกับการสกัดข้อมูลโดยอาศัยวิธี dictionary-based ในการสกัด Entity และ rule-based ในการสกัด Relation จากสื่อตีพิมพ์ biomedical ตัวอย่างงานวิจัยด้านการสกัดข้อมูลที่เป็นภาษาไทยคืองานวิจัย [10] ได้กล่าวถึงการสกัด Name Entity ในรูปแบบหลายคำในภาษาไทยบนคลังข้อความข่าวการเมืองโดยอาศัยวิธี Maximum Entropy Model โดยที่ผ่านมาพบว่างานวิจัยภาษาธรรมชาติด้านทางการแพทย์ในภาษาไทยยังไม่ได้รับความสนใจและมีน้อย

## 3. ขั้นตอนวิธีการที่นำเสนอ

ในบทความนี้ได้นำเสนอขั้นตอนวิธีการระบุอาการและอาการแสดงจากข้อความบอกเล่าอาการสำคัญภาษาไทยตามมาตรฐาน ICD-10 โดยขั้นตอนการประมวลผลภาษาธรรมชาติ ดังรูปที่ 1 โดยเริ่มจากการสกัดข้อความสำคัญ ได้แก่ อาการ อาการแสดง จากข้อความ CCs ซึ่งถูกบันทึกด้วยภาษาไทย โดยอาศัยขั้นตอนประมวลภาษาธรรมชาติ ได้แก่ การตัดคำ (tokenization) การทำรากศัพท์ (Word stemming) การตัดคำที่ไม่มีนัยสำคัญ (Stop word removal), การแทนข้อความให้อยู่ในรูปเวกเตอร์สเปซโมเดล (Vector Space Model) และการคำนวณค่าความเหมือน (similarity calculation) แล้วจึงส่งผ่านข้อความสำคัญซึ่งเป็นผลลัพธ์ที่ได้ไปทำการจับคู่กับรหัสอาการตามที่กำหนดโดยมาตรฐาน ICD-10เพื่อให้ได้รหัสของ ICD-10 ที่ตรงกับอาการที่ระบุในข้อมูล CCs ดังนั้นเพื่อเป็นการเตรียมข้อมูลให้พร้อมและเหมาะสมกับหัวข้อหรือบริบทของงานวิจัยซึ่งจะมีผลโดยตรงต่อผลลัพธ์และกระบวนการสกัดข้อมูล จึงจำเป็นที่จะต้องเข้าใจธรรมชาติของลักษณะข้อมูล เพื่อสามารถเตรียมข้อมูลให้ง่ายต่อการสกัดและได้ผลลัพธ์ที่มีความแม่นยำโดยขั้นตอนวิธีประกอบไปด้วย 2 ขั้นตอนหลัก คือ การเตรียมข้อมูลสำหรับการสกัดอาการและอาการแสดง และการสกัดอาการและอาการแสดง รูปที่ 1 แสดงภาพรวมของกระบวนการ

ระบุโรคจากการสกัดข้อความบอกเล่าอาการผู้ป่วยโดยอาศัยเทคนิคการประมวลผลภาษาธรรมชาติ โดยเริ่มจากเอกสาร CCs และ ICD-10 จะถูกส่งผ่านกระบวนการเตรียมข้อมูลเพื่อนาไปใช้ในการระบุอาการและอาการแสดง โดยก่อนที่จะนำเอกสารข้อความ CCs ทดสอบเข้าสู่กระบวนการการระบุอาการและอาการแสดง ระบบจะทำการตัดคำและติดป้ายเชิงความหมายโดยใช้พจนานุกรมทั่วไปและพจนานุกรมเชิงความหมายที่ได้จากการเตรียมข้อมูลเมื่อเตรียมข้อความ CCs ทดสอบก่อนเพื่อให้อยู่ในรูปแบบที่พร้อมสำหรับการระบุอาการและอาการแสดง จากนั้นจะนำเข้าสู่ขั้นตอนการสกัดอาการและอาการแสดงโดยอาศัยข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูล เพื่อให้สามารถจำแนกอาการและอาการแสดงจากข้อความ CCs ให้อยู่ในรหัส ICD-10 ได้ในที่สุด




รูปที่ 1 กระบวนการระบุโรคจากการสกัดข้อความบอกเล่าอาการผู้ป่วยโดยอาศัยเทคนิคการประมวลผลภาษาธรรมชาติ

## 3.1 การเตรียมข้อมูลสำหรับสกัดอาการและอาการแสดง

การเตรียมข้อมูลประกอบไปด้วยขั้นตอนย่อยดังแสดงในรูปที่ 2 โดยผลลัพธ์ของขั้นตอนดังกล่าวประกอบไปด้วยคลังข้อความ CCs (CCs corpus : A), คลังข้อความ ICD-10(ICD-10 corpus : B), ผลลัพธ์ของแต่ละกรณี CCs( CCs result : C) และ พจนานุกรมเชิงความหมาย (semantic dictionary : D)

### 3.1.1การเตรียมคลังข้อความอาการที่มาจาก ICD-10

คลังข้อความอาการและอาการแสดงจาก ICD-10 มีความสำคัญในการระบุหรือจำแนกอาการและอาการแสดงที่ถูกระบุหรืออธิบายอยู่ในข้อความ CCs โดยทำการตัดคำและติดป้ายในการอธิบายแต่ละคำจากเอกสาร ICD-10 ผลลัพธ์ของขั้นตอนนี้คือคลังข้อความ ICD-10 ตัวอย่างรหัส ICD-10 รหัส R51 "ปวดศีรษะ"[headache] ในการตัดคำด้วยโปรแกรม LongLexTo (http://www.sansarn.com/lexto/) ซึ่งเป็นโปรแกรมประยุกต์ที่ถูกพัฒนาบนภาษา Java โดยอาศัยวิธี LM โดยผลลัพธ์ที่จากการโปรแกรม LongLexTo คือ "ปวดศีรษะ" อย่างไรก็ตามเนื่องจากรูปแบบของภาษาไทยทั่วไปคำว่า "ปวดศีรษะ" สามารถแสดงในรูปอื่นๆที่มีความหมายเดียวกันได้ เช่น "ปวดหัว" หรือ "ปวดบริเวณศีรษะ" ซึ่งจะเห็นได้ว่าหากใช้คำว่า "ปวดศีรษะ" จะไม่สามารถใช้ในการระบุ"ปวดบริเวณศีรษะ" ได้ ดังนั้นหลังจากการตัดคำด้วยโปรแกรม LongLexTo แล้วจะต้องมีกระบวนการเพิ่มเติมเพื่อพิจารณาว่าคำดังกล่าวสามารถตัดคำเป็นส่วนย่อยโดยสามารถคงความหมายเดิมไว้ได้หรือไม่เช่น ปวด[ache]|ศีรษะ[head] เมื่อทำการพิจารณาเช่นนี้แล้วจะสามารถแก้ไข้ปัญหาในการระบุ" ปวดบริเวณศีรษะ" ได้

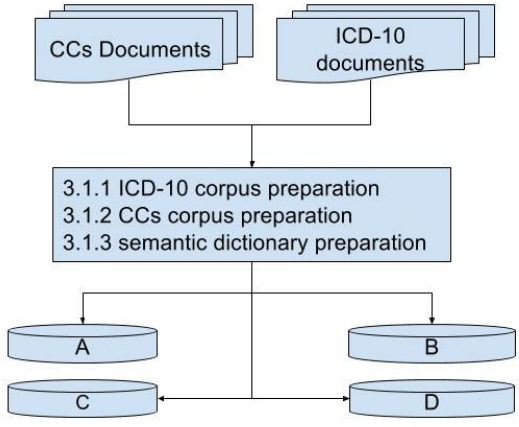
ตารางที่ 1 ตัวอย่างข้อมูลใน ICD-10 corpus

| รหัส ICD-10 | อาการ (ภาษาอังกฤษ) | อาการ (ภาษาไทย) | ผลลัพธ์จาก LongLexto | ตัวอย่างข้อมูล ICD-10 corpus |
|---|---|---|---|---|
| R00.2 | Palpitations | ใจสั่น | ใจสั่น\| | ใจ[heart] สั่น[beat] |
| R04.0 | epistaxis | เลือดกำเดาไหล | เลือด กำเดา\| ไหล\| | เลือดกำเดา [epistaxis]ไหล[discharge] |
| R05 | cough | ไอ | ไอ | ไอ[cough] |
| R07.0 | Pain in throat | เจ็บในคอ | เจ็บ\|ใน\| คอ\| | เจ็บ[pain]ใน [in]คอ [throat] |
| R07.1 | Chest pain on breathing | เจ็บหน้าอกเวลาหายใจ | เจ็บ\| หน้าอก \| เวลา\| หายใจ\| | เจ็บ [pain]หน้าอก [chest]เวลา [when]หายใจ [breathe] |
| R07.4 | Chest pain | เจ้าหน้าอก | เจ็บ\| หน้าอก\| | เจ็บ [pain]หน้าอก [chest] |

| รหัส ICD-10 | อาการ (ภาษาอังกฤษ) | อาการ (ภาษาไทย) | ผลลัพธ์จาก LongLexto | ตัวอย่างข้อมูล ICD-10 corpus |
|---|---|---|---|---|
| R42 | Dizziness and giddiness | เวียนศีรษะ | เวียน\|ศีรษะ\| | เวียน [dizziness]ศีร ษะ [head] |
| R51 | headache | ปวดศีรษะ | ปวด ศีรษะ\| | ปวด [ache]ศีรษะ [head] |

### 3.1.2การเตรียมคลังข้อความบอกเล่าอาการ

ในขั้นตอนนี้จะประกอบไปด้วยส่วนของการตัดคำและการติดป้ายอธิบาย พร้อมทั้งระบุผลลัพธ์ของแต่ละกรณี โครงสร้างของคลัง CCs ประกอบไปด้วย ลำดับกรณี ประโยค CCs และ ประโยค CCs ที่ถูกตัดและติดป้ายแล้วผลลัพธ์ของขั้นตอนนี้คือ คลังข้อความ CCs และ ผลลัพธ์ของแต่ละกรณีของ CCs ตัวอย่างข้อความ CCs ที่ถูกตัดด้วย LongLexTo มีผลลัพธ์ ดังนี้ “|มีอา|การ|เกร็ง|และ|อ่อนแรง|ของ|แขน| |ขา| |ทั้ง| |2| |ข้าง| |มา|ประมาณ| |1| |ปี| |6| |เดือน|” จะเห็นได้ว่าคำว่า "มีอาการ" ถูกตัดออกเป็น "|มีอา|การ|" เป็นการตัดคำที่ไม่ตรงความหมาย หากเป็นคำที่ไม่มีนัยสำคัญแล้วจะไม่สนใจว่าตัดคำได้ถูกต้องหรือไม่ ทั้งนี้เนื่องจากคำดังกล่าวจะไม่ถูกนำมาพิจารณาในการระบุอาการหรืออาการแสดง แต่ในความเป็นจริงแล้วคำว่า "อาการ" หรือ "มีอาการ" เป็นคำที่มักขึ้นต้นในการกล่าวถึงอาการเสมอจึงถูกพิจารณาให้เป็นคำที่มีนัยสำคัญหรือมีส่วนในการระบุอาการหรืออาการแสดงนั่นเอง




รูปที่ 2 การเตรียมข้อมูลสำหรับการสกัดอาการและอาการแสดง

### 3.1.3 การเตรียมพจนานุกรมเชิงความหมาย

พจนานุกรมเชิงความหมายจะเป็นคำที่ถูกนำมาจากคลังข้อความทั้งสองในขั้นต้น ซึ่งจะถูกนำมาใช้ในกระบวนการต่อไปนี้ การตัดคำ การติดป้ายอธิบายคำศัพท์ และ การแก้ไขปัญหาความพ้องความหมาย

การตัดคำ เป็นการตัดคำโดยอาศัยพจนานุกรมมีความจำเป็นในการตัดคำที่จำเป็นที่จะต้องตัดคำสำคัญได้อย่างถูกต้อง เพื่อนำคำสำคัญดังกล่าวมาใช้ในการประมวลผลถัดไป

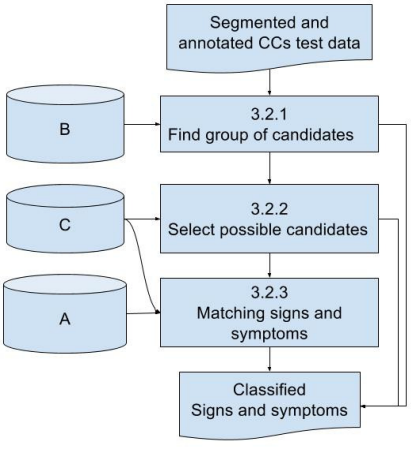
การติดป้ายอธิบาย โดยป้ายอธิบายถูกแบ่งออกเป็นสามประเภท ป้ายเชิงความหมาย ป้ายเชิงรูปแบบ และ ป้ายใดๆ

ในการแก้ไขปัญหาความพ้องความหมาย เช่น หัว/region,head ศีรษะ/region,head จะพบว่าทั้งสองมีรูปประกอบและคำอ่านออกเสียงที่ต่างกันแต่มีความหมายเหมือนกัน หรือ ที่/หน้าอก หรือ บริเวณ/หน้าอก บริเวณ/อก พบว่าการประกอบของวลีมีความแตกต่างกันแต่สื่อความหมายเดียวกัน

- ป้ายเชิงความหมาย คือป้ายที่ระบุถึงอาการสำคัญที่ประกอบอยู่ในคลังข้อความอาการ
- ป้ายเชิงรูปแบบ คือป้ายเชิงความหายที่ไม่ได้สื่อหรือระบุถึงอาการสำคัญแต่มีส่วนในการใช้พิจารณารูปแบบเพื่อหาอาการสำคัญ
- ป้ายใดๆ คือป้ายที่เป็นได้ทั้งป้ายเชิงความหมายและป้ายเชิงรูปแบบหรือไม่เป็นทั้งสองอย่างก็ได้

### 3.2 การสกัดอาการและอาการแสดง

ในการสกัดอาการและอาการแสดงจะใช้คลังข้อความ CCs ที่ตัดโดยพจนานุกรม โดยสามารถแบ่งขั้นตอนการสกัดอาการออกได้เป็น 3 ขั้นตอนดังนี้




รูปที่ 3 กระบวนการการสกัดอาการและอาการแสดง

### 3.2.1 การหาอาการและอาการแสดงที่เป็นไปได้

ในการหาอาการที่เป็นไปได้จากข้อความ CCs ที่ใช้ทดสอบจะพิจารณาจากความสัมพันธ์ระหว่างป้ายเชิงความหมายสำคัญที่ถูกติดบนข้อความ CCs และจากป้ายเชิงความหมายสำคัญที่ถูกติดบนคลังอาการ

### 3.2.2 การเลือกอาการที่เป็นไปได้จากคลังอาการ

ในขั้นตอนที่ผ่านมาทำให้สามารถระบุอาการที่อาจถูกระบุอยู่ในข้อความ CCs โดยอาการที่ยังไม่ได้ถูกยอมรับในขั้นตอนที่ 3.2.1 จะทำการเปรียบเทียบกับรูปแบบที่เกิดขึ้นในกับผลลัพธ์ของคลังข้อความ CCs ที่มีผลลัพธ์ตรงกับอาการที่มีตรงกับอาการที่น่าจะเป็นไปได้ในขั้นตอนที่ 3.2.1 หากเคยเกิดรูปแบบดังกล่าวขึ้นจริงและภายในรูปแบบของป้ายเชิงความสำคัญไม่ได้มีป้ายเชิงความหมายอื่นๆปะปนอยู่ก็สามารถระบุได้ว่าเป็นอาการดังกล่าวจริง

### 3.2.3 การจับคู่อาการและอาการแสดง

จากขั้นตอน 3.2.2 หากพบว่ามีรูปแบบถูกต้องแต่มีป้ายเชิงความหมายอื่นปนอยู่ก็จะทำการเปรียบเทียบรูปแบบจากข้อมูลทั้งผลลัพธ์ของข้อความทดสอบ และ ข้อความทดสอบจากคลังข้อความ CCs

## 4.บทสรุปและงานในอนาคต

บทความนี้เสนอกระบวนการเพื่อระบุชื่อโรคหรืออาการที่เป็นไปได้ในรูปแบบของรหัส ICD-10 โดยอาศัยกระบวนการตามทฤษฎีของการประมวลผลภาษาธรรมชาติเพื่อสกัดข้อความ CCs ภาษาไทย สำหรับงานในอนาคตผู้วิจัยจะทำการสร้างระบบเพื่อสนับสนุนการทำงานตามกระบวนการที่ได้นำเสนอไปแล้วนั้น และนำระบบที่ได้ไปทดสอบกับข้อความ CCs ที่เป็นข้อมูลจริง และจะนำผลลัพธ์ที่ได้มาตรวจสอบความถูกต้องและความแม่นยำของกระบวนการดังกล่าวโดยการเปรียบเทียบกับข้อความ CCs ที่ได้รับการวินิจฉัยและระบุรหัส ICD-10 โดยแพทย์จากโรงพยาบาล เพื่อประเมินประสิทธิภาพของกระบวนการที่นำเสนอในงานวิจัยนี้

## เอกสารอ้างอิง


[1] Wu, Tsung-Shu Joseph, et al. "Establishing a nationwide emergency department-based syndromic surveillance system for better public health responses in Taiwan." BMC public health 8.1 (2008): 1.

[2] Lu, Hsin-Min, et al. "Multilingual chief complaint classification for syndromic surveillance: An experiment with Chinese chief complaints." international journal of medical informatics 78.5 (2009): 308-320.

[3] Ketui, Nongnuch, Thanaruk Theeramunkong, and Chutamanee Onsuwan. "Thai elementary discourse unit analysis and syntactic-based segmentation." International Information Institute (Tokyo). Information 16.10 (2013): 7423.

[4] Chamyapornpong. S. 1983. A Thai Syllable Separation Algorithm. Master thesis, Asian

[5] Poonwarawan. Y. 1986. Dictionary-based Thai Syllable Separation. In proceeding of the 9th Electrical Engineering Conference

[6] Sornlertlamvanich. V. 1993. Word Segmentation for Thai in a Machine Translation System. NECTEC, Bangkok.

[7] Jiang, Min, et al. "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries." Journal of the American Medical Informatics Association 18.5 (2011): 601-606.

[8] Chen, Yukun, et al. "A study of active learning methods for named entity recognition in clinical text." Journal of biomedical informatics 58 (2015): 11-18.

[9] Song, Min, et al. "PKDE4J: Entity and relation extraction for public knowledge discovery." Journal of biomedical informatics 57 (2015): 320-332.

[10] Chanlekha, Hutchatai, and Asanee Kawtrakul. "Thai named entity extraction by incorporating maximum entropy model with simple heuristic information." Proceedings of the IJCNLP. 2004.

# ประวัติผู้เขียน

**ชื่อ สกุล** นายภวินท์ แซ่คู

**รหัสประจำตัวนักศึกษา** 5710220071

**วุฒิการศึกษา**

| วุฒิ | ชื่อสถาบัน | ปีที่สำเร็จการศึกษา |
|---|---|---|
| วิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์) | มหาวิทยาลัยสงขลานครินทร์ | 2554 |

**การตีพิมพ์เผยแพร่ผลงาน**

Saeku, P., Duangsuwan, J. 2016. ICD-10 Symptoms and Signs Identification from Thai Chief Complaints using Natural Language Processing, Im: Proceeding of 2016 International Computer Science and Engineering Conference (ICSEC 2016)., pp 665-669.