การพัฒนากระบวนการของการถาม-ตอบแบบนอนแฟคทอยด์

วิทยานิพนธ์

ของ

ณัฐกิตติ์ ศรีกาญจนเพริศ

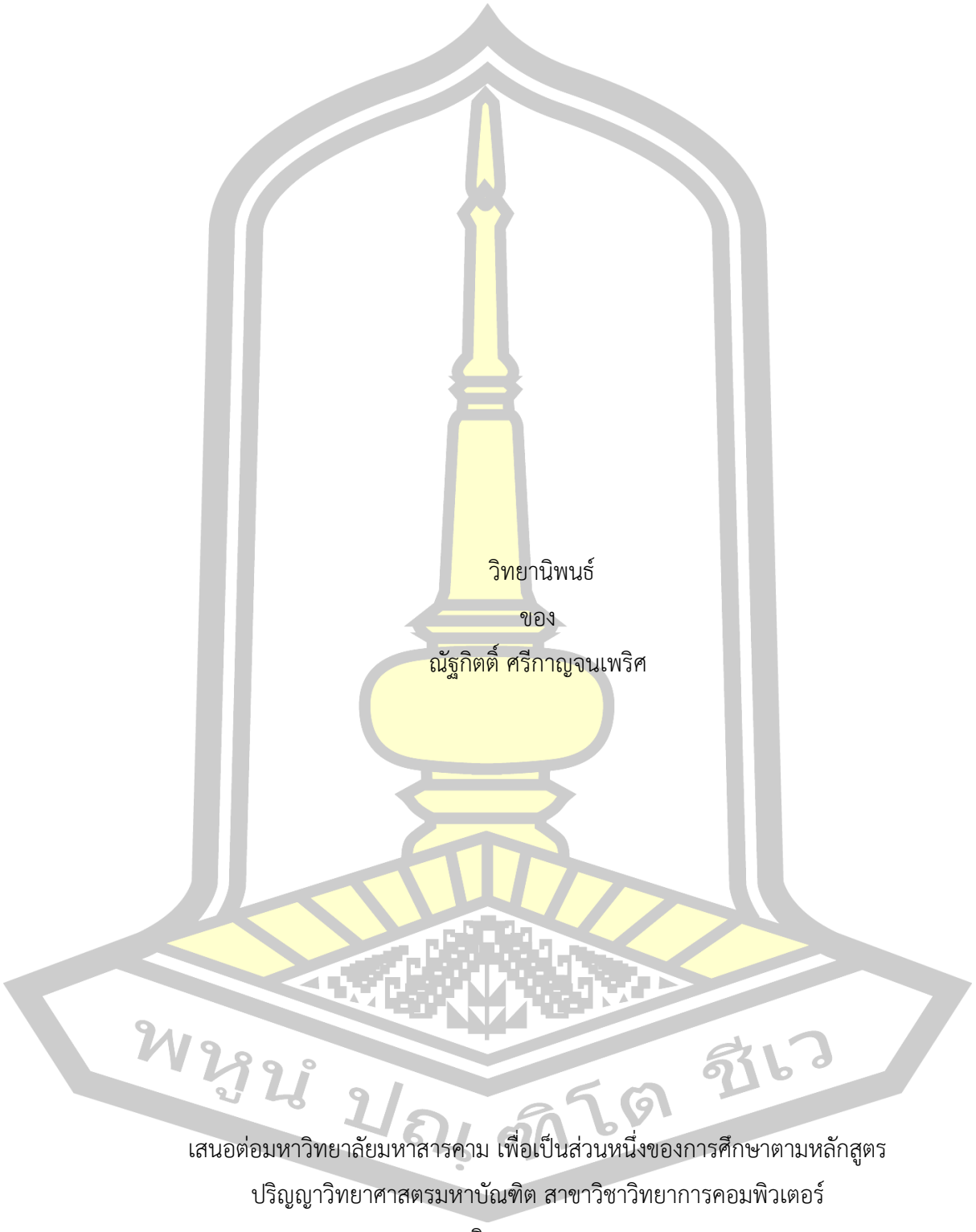
เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

มิถุนายน 2562

สงวนลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

การพัฒนากระบวนการของการถาม-ตอบแบบนอนแฟคทอยด์



วิทยานิพนธ์

ของ

ณัฐกิตติ์ ศรีกาญจนเพริศ

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

มิถุนายน 2562

สงวนลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

Developing Method of Non-Factoid Question Answering

Natthakit Srikanjanapers

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Computer Science)

June 2019

Copyright of Mahasarakham University

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายณัฐกิตติ์ ศรีกาญจนเพริศ แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผศ. ดร. วรารัตน์ สงฆ์แป้น)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. จันทิมา พลพินิจ)

.....กรรมการ
(ผศ. ดร. พัฒนพงษ์ ชมภูวิเศษ)

.....กรรมการ
(ผศ. ดร. พนิดา ทรงรัมย์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

.....

(ผศ. ศศิธร แก้วมั่น) (ผศ. ดร. กริสน์ ชัยมูล)
คณบดีคณะวิทยาการสารสนเทศ คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การพัฒนากระบวนการของการถาม-ตอบแบบนอนแฟคทอยด์
ผู้วิจัย ณัฐกิตติ์ ศรีกาญจนเพริศ
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. จันทิมา พลพินิจ
ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** วิทยาการคอมพิวเตอร์
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2562

บทคัดย่อ

การถาม-ตอบ (Question-answering: QA) มีวัตถุประสงค์เพื่อการค้นหาและสร้างคำตอบที่เกี่ยวข้องกับคำถามที่มนุษย์เขียนไว้ในลักษณะของภาษาธรรมชาติแบบอัตโนมัติ ปัจจุบันระบบการถาม-ตอบที่ได้รับความสนใจมากขึ้นคือการถาม-ตอบที่ต้องการคำตอบที่เป็นเหตุผลเรียกว่าการถาม-ตอบ แบบนอนแฟคทอยด์ (Non-factoid QA) ไม่ใช่คำตอบในลักษณะสั้นๆ (Factoid QA) ดังที่ผ่านมา ดังนั้นปัญหาที่กำลังได้รับการศึกษาในปัจจุบันของการถาม-ตอบ แบบนอนแฟคทอยด์ ก็คือการหาทางส่วนที่เหมาะสมที่สุดในเอกสารที่จะใช้เป็นคำตอบ ซึ่งงานวิจัยฉบับนี้นำเสนอวิธีการค้นหาคำตอบสำหรับการถาม-ตอบแบบนอนแฟคทอยด์ โดยกระบวนการที่นำเสนอมี 2 ขั้นตอนหลัก ขั้นตอนแรกคือการสร้างรูปแบบคำถามและรูปแบบคำตอบด้วยแบบจำลองมาร์คอฟ (Markov Model: MM) โดยรูปแบบคำถามและรูปแบบคำตอบที่ได้จะถูกใช้ในขั้นตอนที่ 2 ที่เรียกว่าขั้นตอนการประมวลผลการถาม-ตอบ ในที่สุดหลังจากที่ประเมินการค้นหาคำตอบในการถาม-ตอบแบบนอนแฟคทอยด์ด้วย BLEU-1 และ ROUGE-L ซึ่งให้ผลลัพธ์ที่เท่ากับ 0.43 และ 0.41 ตามลำดับ และจากการศึกษาที่ผ่านมาพบว่าผลลัพธ์ดังกล่าวเป็นผลลัพธ์ที่ยอมรับได้ในงานวิจัยด้านการถาม-ตอบแบบนอนแฟคทอยด์

คำสำคัญ : ระบบถาม-ตอบแบบนอนแฟคทอยด์, การประมวลผลภาษาธรรมชาติ, การค้นคืนสารสนเทศ, แบบจำลองมาร์คอฟ

TITLE Developing Method of Non-Factoid Question Answering

AUTHOR Natthakit Srikanjanapers

ADVISORS Assistant Professor, Jantima Polpinij , Ph.D.

DEGREE Master of Science **MAJOR** Computer Science

UNIVERSITY Mahasarakham University **YEAR** 2019

ABSTRACT

Question answering (QA) aims to automatically find and generate an answer that is relevant to a question posed by humans in a natural language. Today, the most popular QA systems that are being given more attention are non-factoid. Therefore, the current problem found in non-factoid QA is to find the most appropriate passage in the candidate documents that is considered as the answer. This work presents a method of answer retrieval for non-factoid QA. The proposed method consists of two main stages. The first stage is preliminary that concentrates to provide the question and answer patterns based on Markov Model (MM). These patterns will used for the next stage, called question-answer processing stage. After evaluating of answer retrieval performance by Bilingual Evaluation Understudy (BLEU-1) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L), they are 0.43, and 0.41 respectively. To the best of our knowledge, these results are acceptable.

Keyword : Non-factoid Question Answering, Natural Language Processing, Information Extraction, Markov Model

กิตติกรรมประกาศ

วิทยานิพนธ์ เรื่อง “การถาม-ตอบเชิงการแพทย์เพื่อการสนับสนุนการตัดสินใจ” ฉบับนี้สามารถพัฒนาจนสำเร็จลุล่วงไปได้ด้วยดีโดยได้รับการดูแลเป็นอย่างดีจาก ผศ. ดร. จันทิมา พลพินิจ ที่เป็นทั้งที่ปรึกษาโครงการและอาจารย์ที่กรุณาสั่งสอน ให้คำแนะนำ คำปรึกษาและช่วยชี้แนะแนวทางตรวจทานแก้ไขข้อบกพร่องในการค้นคว้าหาความรู้เพื่อนำมาใช้ในการเขียนโครงงานฉบับนี้ ทางคณะผู้จัดทำรู้สึกซาบซึ้งในความอนุเคราะห์ จากท่านอาจารย์และขอกราบพระคุณไว้เป็นอย่างสูง

ขอขอบคุณ คุณพ่อ คุณแม่ ญาติพี่น้อง รวมถึงพี่ๆ และเพื่อนๆ โดยเฉพาะในหน่วยวิจัยทางปัญญา (Intellec Laboratory) ที่คอยให้ความช่วยเหลือ ให้กำลังใจ และส่งเสริมในทุกๆ ด้าน นอกจากนี้ข้าพเจ้าขอขอบพระคุณอาจารย์ในภาควิชาวิทยาการคอมพิวเตอร์ทุกท่าน ที่ได้ให้ความรู้ที่สามารถใช้ในการศึกษาเล่าเรียนในระดับที่สูงขึ้น ตลอดจนการใช้ประโยชน์จากความรู้เพื่อการทำงานต่อไป

โดยคุณค่าและประโยชน์อันพึงมีมาจากโครงงานฉบับนี้ผู้จัดทำขอมอบแด่ผู้มีพระคุณทุกท่าน

ณัฐกิตติ์ ศรีกาญจนเพริศ

พหูน ปณฺฑิโต ชีเว

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญภาพ ..... ฏ

สารบัญตาราง ..... ฑ

บทที่ 1 บทนำ 1

1.1 หลักการและเหตุผล 1

1.2 วัตถุประสงค์ของการวิจัย 2

1.3 ความสำคัญของการวิจัย 2

1.4 ขอบเขตของการวิจัย 2

1.5 นิยามศัพท์เฉพาะ 3

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 4

2.1 การถาม-ตอบ (Question Answering: QA) 4

2.2 สถาปัตยกรรมโดยทั่วไปของการถาม-ตอบ 5

2.2.1 ขั้นตอนการประมวลผลคำถาม (Question Processing Module) 6

2.2.1.1 การขยายของเขตของคำค้น (Query Expansion) 7

2.2.1.2 การกำหนดคำค้นใหม่จากคำถาม (Question Reformulation) 8

2.2.2 ขั้นตอนการประมวลผลเอกสาร (Document Processing Module) 8

2.2.3 ขั้นตอนการสกัดช่วงของเอกสาร (Paragraph Processing Module) 9

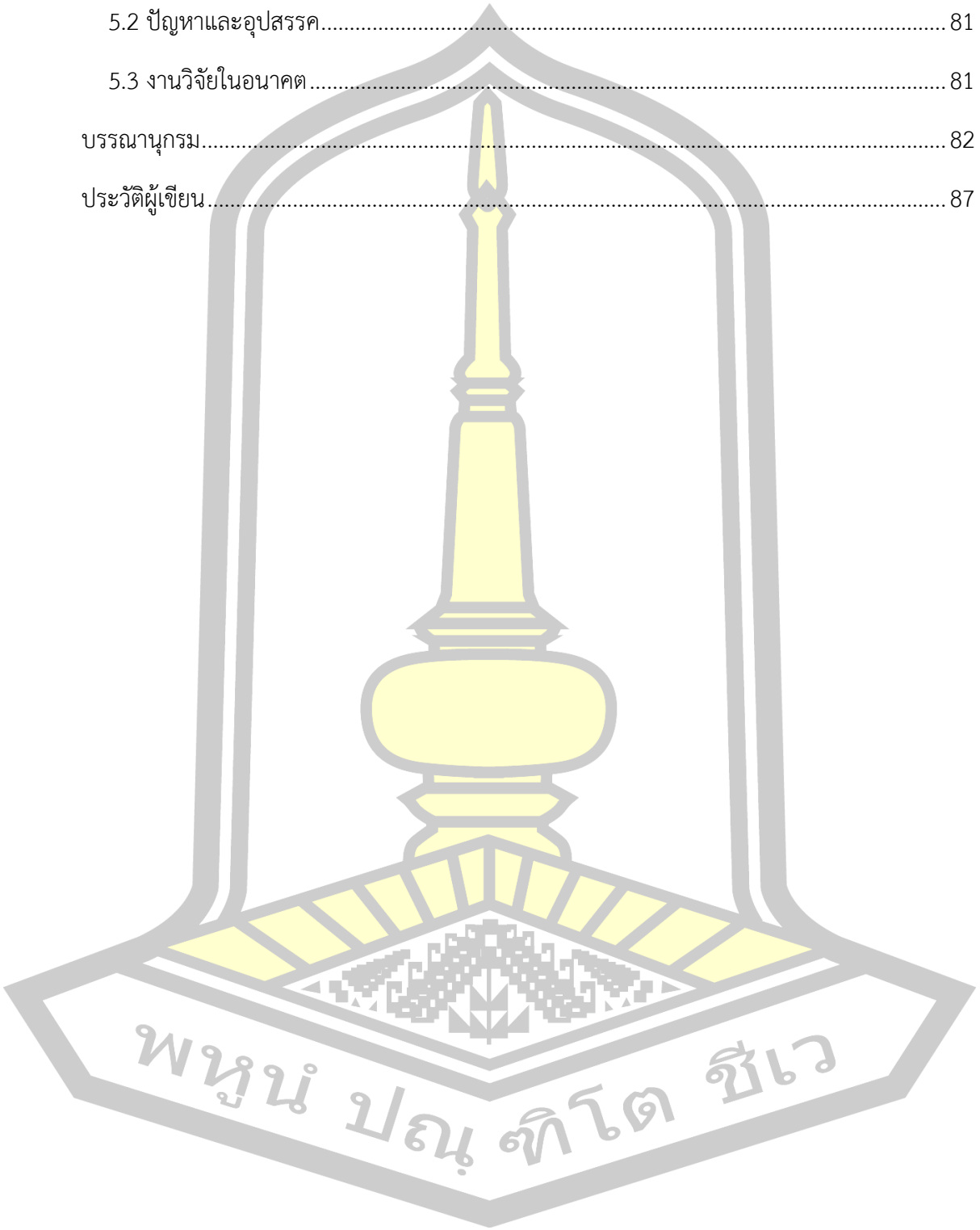
2.2.4 ขั้นตอนการสกัดเอาคำตอบ (Answer Extraction Module) 9

2.3 ประเภทของระบบถาม-ตอบ 9
2.3.1 การจำแนกประเภทของระบบถาม-ตอบ ตามโดเมน 10
2.3.1.1 ระบบถาม-ตอบในโดเมนแบบเปิด 10
2.3.1.2 ระบบถาม-ตอบในโดเมนแบบปิด 10
2.3.2 การจำแนกประเภทของระบบถาม-ตอบ ตามเทคนิคที่ใช้ 11
2.3.2.1 การวิเคราะห์โดยใช้หลักการทางภาษา (Linguistic Analysis) 11
2.3.2.2 การวิเคราะห์ทางสถิติ (Statistical Analysis) 11
2.3.2.3 การเปรียบเทียบรูปแบบ (Pattern Matching) 11
2.3.3 การจำแนกประเภทของระบบถาม-ตอบ ตามลักษณะของคำตอบ 12
2.3.3.1 การถาม-ตอบ แบบแฟคทอยด์ 13
2.3.3.2 ระบบถาม-ตอบ แบบนอนแฟคทอยด์ 13
2.4 เทคนิค แนวคิด และกระบวนการที่เกี่ยวข้อง 14
2.4.1 การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) 14
2.4.1.1 การตัดคำ (Word Segmentation) 15
2.4.1.2 การตัดคำหยุด (Stop words) 16
2.4.1.3 Vector Space Model 17
2.4.1.4 การให้น้ำหนัก (Term Weighting) 17
2.4.1.5 การวิเคราะห์ความคล้ายคลึง (Similarity Analysis) 19
2.4.1.6 การจัดอันดับเอกสารที่เกี่ยวข้อง (Document Ranking) 20
2.4.1.7 การติดแท็กให้กับชนิดของคำ (Part of Speech Tagging) 21
2.4.2 การค้นคืนสารสนเทศ (Information Retrieval : IR) 22
2.4.3 การประเมินกระบวนการของระบบถาม-ตอบ (QA Evaluation) 23
2.4.3.1 Precision at K 23
2.4.3.2 Mean Reciprocal Rank 23

2.4.3.3 Bilingual Evaluation Understudy 24
2.4.3.4 Recall-Oriented Understudy for Gisting Evaluation 25
2.5 งานวิจัยที่เกี่ยวข้อง 25
บทที่ 3 ขั้นตอนการดำเนินงาน 29
3.1 ชุดข้อมูล (Data Set) 29
3.1.1 รูปแบบการจัดเก็บชุดข้อมูล (Dataset Format) 30
3.2 กรอบการดำเนินงาน (Research Methodology) 31
3.3 กระบวนการการประมวลผลขั้นต้น (Preliminary Processing) 32
3.3.1 การเรียนรู้รูปแบบของคำถาม (Learning of Question Patterns) 33
3.3.2 การเรียนรู้รูปแบบคำตอบด้วยการใช้คำหยุด (Learning Answer Patterns using Stop Words) 37
3.4 การประมวลผลในกระบวนการถาม-ตอบ (Question Answering Method) 40
3.4.1 การประมวลผลคำถาม (Question Processing) 41
3.4.1.1 การประมวลผลสำหรับการวิเคราะห์คำถามเพื่อสกัดเอา “คำสำคัญ (Keyword)” 41
3.4.1.2 การประมวลผลคำถามเพื่อประเมินรูปแบบของคำถามสำหรับการค้นหารูปแบบ คำตอบที่เหมาะสม 41
3.4.2 การค้นคืนเอกสารที่เกี่ยวข้อง (Document Candidate Retrieval) 43
3.4.2.1 การจัดอันดับเอกสารที่คาดว่าจะมีคำตอบด้วย BM25 44
3.4.2.2 การจัดอันดับเอกสารที่คาดว่าจะมีคำตอบด้วย MATF 46
3.4.3 การค้นคืนส่วนของเอกสารที่เกี่ยวข้อง (Passage Candidate Retrieval) 48
3.4.3.1 การหาค่าความคล้ายคลึงระหว่างคำสำคัญจากคำถามและส่วนของเอกสารที่คาดว่าจะเป็นคำตอบด้วย BM25 51
3.4.3.2 การหาค่าความคล้ายคลึงระหว่างคำสำคัญจากคำถามและส่วนของเอกสารที่คาดว่าจะเป็นคำตอบด้วย MATF 53

3.5 การวัดประสิทธิภาพของกระบวนการที่นำเสนอ (Evaluation Method) 56
3.5.1 การวัดประสิทธิภาพการค้นคืนเอกสารด้วย Precision at K (P@K) 56
3.5.2 การวัดประสิทธิภาพการค้นคืนเอกสารด้วย Mean Reciprocal Rank (MRR) 57
3.5.3 การวัดประสิทธิภาพคำตอบด้วย Bilingual Evaluation Understudy (BLEU) 58
3.5.4 การวัดประสิทธิภาพคำตอบด้วย Recall-Oriented Understudy for Gisting Evaluation (ROUGE) 60
บทที่ 4 การวัดประสิทธิภาพ 62
4.1 ชุดข้อมูลที่ใช้ในการทดสอบ (Dataset) 62
4.2 ผลการทดสอบการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ (Document Candidate Retrieval Experimental results) 62
4.2.1 การประมวลผลคำถาม 63
4.2.2 การประมวลผลเอกสารที่คาดว่าจะมีคำตอบ 64
4.2.3 การให้น้ำหนักคำและค้นคืนเอกสาร 65
4.2.4 การวัดประสิทธิภาพของการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ 66
4.3 ผลการทดสอบ (The Experimental Results of Answer Retrieval) 69
4.3.1 การทดสอบการค้นคืนส่วนที่คาดว่าจะเป็นคำตอบด้วยชุดข้อมูลที่ใช้ในการเรียนรู้ 70
4.3.1.1 การทดสอบรูปแบบคำตอบที่เรียนรู้ในการค้นคืน “ส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ” โดยใช้ชุดข้อมูลที่ใช้ในการเรียนรู้รูปแบบคำตอบ 70
4.3.1.2 การทดสอบการค้นคืนคำตอบด้วยชุดข้อมูลที่ใช้ในการเรียนรู้รูปแบบคำตอบ 71
4.3.2 การทดสอบการค้นคืนส่วนที่คาดว่าจะเป็นคำตอบด้วยชุดข้อมูลเตรียมไว้ในการทดสอบ 74
4.3.2.1 การค้นคืนส่วนที่คาดว่าจะเป็นคำตอบ 74
4.3.2.2 การจัดอันดับส่วนที่คาดว่าจะเป็นคำตอบ 75
4.3.2.3 การวัดประสิทธิภาพคำตอบที่ได้จากกระบวนการ 76
บทที่ 5 สรุปผลการทดลอง 78

5.1 สรุปการดำเนินงานวิจัย 78
5.2 ปัญหาและอุปสรรค 81
5.3 งานวิจัยในอนาคต 81
บรรณานุกรม 82
ประวัติผู้เขียน 87



สารบัญภาพ

หน้า

รูปที่ 2.1 ภาพรวมของระบบถาม-ตอบแบบทั่วไป (Generic Process of QA) 6
รูปที่ 2.2 แสดงตัวอย่างกลุ่มของคำค้นที่ได้หลังจากการกำหนดใหม่ 8
รูปที่ 2.3 แสดงตัวอย่าง Tags ของ Penn Treebank 22
รูปที่ 3.1 โครงสร้างการจัดเก็บข้อมูลของ MS MARCO เวอร์ชัน 2.1 31
รูปที่ 3.2 กรอบการดำเนินงานของกระบวนการที่นำเสนอ 32
รูปที่ 3.3 การสร้างรูปแบบของกลุ่มคำหยุด 33
รูปที่ 3.4 ตัวอย่างแบบจำลองมาร์คอฟที่ได้ 36
รูปที่ 3.5 กรอบการดำเนินงานเพื่อการสร้างรูปแบบคำตอบด้วยคำหยุด 37
รูปที่ 3.6 ภาพรวมของกระบวนการการถาม-ตอบ 40
รูปที่ 3.7 ตัวอย่างการประมวลผลคำถามเพื่อสกัดเอาคำสำคัญ 41
รูปที่ 3.8 ตัวอย่างการประมวลผลคำถามเพื่อตรวจสอบรูปแบบคำถามและค้นคืนรูปแบบคำตอบ 41
รูปที่ 3.9 กรอบการดำเนินงานเพื่อการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ 43
รูปที่ 3.10 ตัวอย่างการนำเสนอการจัดเก็บดัชนีย์เอกสาร 44
รูปที่ 3.11 ตัวอย่างการนำเสนอการเปรียบเทียบคำค้นกับเอกสาร 44
รูปที่ 3.12 ภาพรวมการค้นคืนส่วนของเอกสารที่เกี่ยวข้อง 49
รูปที่ 3.13 ตัวอย่างการค้นคืนส่วนของเอกสารที่คาดว่าจะมีคำตอบ 50
รูปที่ 3.14 การระบุตำแหน่งของคำค้นและกลุ่มของคำหยุดในเอกสาร 51
รูปที่ 3.15 เอกสารที่สนใจจำนวน K อันดับ 57
รูปที่ 3.16 เอกสารที่ค้นคืนและถูกต้องใน K อันดับ 57
รูปที่ 3.17 ตัวอย่างการจัดอันดับเพื่อวัดประสิทธิภาพด้วย MRR ใน K อันดับ 58
รูปที่ 3.18 ตัวอย่างการนับความยาวของ LCS 60

สารบัญภาพ

หน้า

รูปที่ 4.1 กระบวนการการทดสอบการค้นคืนเอกสาร 63
รูปที่ 4.2 ลักษณะการจัดเก็บผลลัพธ์ของการประมวลผลคำถาม 64
รูปที่ 4.3 ลักษณะการจัดเก็บผลลัพธ์ของการประมวลผลเอกสาร 65
รูปที่ 4.4 ลักษณะการจัดเก็บผลลัพธ์ของการค้นคืนเอกสารที่คาดว่าจะมคำตอบ 66
รูปที่ 4.5 คำตอบที่ได้จากกระบวนการถามตอบ ในชุดข้อมูลชุดทดสอบ 73
รูปที่ 4.6 แสดงคำตอบที่ได้จากชุดข้อมูลชุดทดสอบ 73
รูปที่ 4.9 ส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ 75
รูปที่ 4.10 ผลลัพธ์ที่ได้จากการจัดอันดับคำตอบด้วย MATF 75

พหุน ปณ ฑิโต ชีเว

สารบัญตาราง

หน้า

ตารางที่ 2.1 แสดงกลุ่มของคำถาม 5
ตารางที่ 2.2 แสดงตัวอย่างการขยายขอบเขตคำค้น 7
ตารางที่ 2.3 แสดงความเกี่ยวสัมพันธ์ประเภทระบบถาม-ตอบ 14
ตารางที่ 2.4 แสดงตัวอย่างการตัดคำในภาษาอังกฤษ 16
ตารางที่ 2.5 แสดงการตัดคำหยุดในภาษาอังกฤษ 16
ตารางที่ 3.1 แสดงจำนวนข้อมูลในชุดข้อมูลมาตรฐาน MS MARCO 29
ตารางที่ 3.2 แสดงโครงสร้างของชุดข้อมูลมาตรฐาน MS MARCO 30
ตารางที่ 3.3 ตัวอย่างเมทริกซ์ของจำลองมาร์คอฟ 35
ตารางที่ 3.4 ตัวอย่างรูปแบบของคำถามที่เรียนรู้ได้ด้วยแบบจำลองมาร์คอฟ 36
ตารางที่ 3.5 จำนวนรูปแบบของคำถาม 37
ตารางที่ 3.6 ตัวอย่างการทำ POS tagging ของเอกสารคำตอบ 38
ตารางที่ 3.7 ตัวอย่างการระบุตำแหน่งของคำหยุดในคำตอบ 38
ตารางที่ 3.8 ตัวอย่างรูปแบบคำตอบที่สร้างจากกลุ่มคำหยุดและสรุปจำนวนรูปแบบคำตอบที่ได้ 40
ตารางที่ 3.9 ผลลัพธ์ของการตัดคำหยุดและการหาคำคล้ายคลึง 43
ตารางที่ 3.10 ตัวอย่างข้อมูลสำหรับจัดอันดับเอกสารที่คาดว่าจะมีคำตอบด้วย BM25 45
ตารางที่ 3.11 ผลลัพธ์ของเอกสารที่คาดว่าจะมีคำตอบที่ถูกให้คะแนนด้วย BM25 46
ตารางที่ 3.12 ตัวอย่างข้อมูลสำหรับจัดอันดับเอกสารที่คาดว่าจะมีคำตอบด้วย MATF 46
ตารางที่ 3.13 ตัวอย่างข้อมูลสำหรับจัดอันดับเอกสารที่คาดว่าจะมีคำตอบด้วย MATF 48
ตารางที่ 3.14 ตัวอย่างข้อมูลสำหรับจัดอันดับด้วย BM25 52
ตารางที่ 3.15 ผลลัพธ์ข้อมูลที่ถูกให้คะแนนด้วย BM25 53
ตารางที่ 3.16 ตัวอย่างข้อมูลสำหรับจัดอันดับส่วนของเอกสารที่คาดว่าจะมีคำตอบด้วย MATF 54

สารบัญตาราง

หน้า

ตารางที่ 3.17 ตัวอย่างการจัดอันดับส่วนของเอกสารที่คาดว่าจะเป็นคำตอบด้วย MATF 56
ตารางที่ 3.18 ตัวอย่างข้อมูลการวัดประสิทธิภาพด้วย BLEU 59
ตารางที่ 4.1 สรุปข้อมูลที่ใช้ในการทดสอบ 62
ตารางที่ 4.2 ผลลัพธ์วัดประสิทธิภาพในการค้นคืนเอกสารที่คาดว่าจะมีคำตอบประเภท How 67
ตารางที่ 4.3 ผลลัพธ์วัดประสิทธิภาพในการค้นคืนเอกสารที่คาดว่าจะมีคำตอบประเภท Why 68
ตารางที่ 4.4 ผลลัพธ์วัดประสิทธิภาพด้วยค่า MRR 68
ตารางที่ 4.5 แสดงผลลัพธ์ที่ได้จากการนำ “รูปแบบคำตอบ” ที่เรียนรู้ได้ 70
ตารางที่ 4.6 ตัวอย่างคำตอบของคำถามที่จะทำการประมวลผล 72
ตารางที่ 4.7 ผลการค้นคืนคำตอบด้วยรูปแบบคำถาม “How” ด้วยรูปแบบคำถาม 74
ตารางที่ 4.8 ผลการค้นคืนคำตอบจากคำถามประเภท “Why” ด้วยรูปแบบคำถามที่ 74
ตารางที่ 4.9 ผลลัพธ์วัดประสิทธิภาพคำตอบที่ได้จากกระบวนการ 76
ตารางที่ 4.10 ผลลัพธ์วัดประสิทธิภาพคำตอบที่ได้จากกระบวนการ 76
ตารางที่ 4.11 ผลลัพธ์วัดประสิทธิภาพคำตอบที่ได้จากกระบวนการ 77

พหูน ปณฺฑิโต ชีเว

บทที่ 1

บทนำ

1.1 หลักการและเหตุผล

การถาม-ตอบ (Question-Answering: QA) [1, 2] คือการแสดงคำตอบที่กะทัดรัดและแม่นยำ (concise and precise answers) ที่ตรงกับคำถามของผู้ใช้งาน (user's question) ในรูปแบบภาษาธรรมชาติ โดยพื้นฐานของกระบวนการถามตอบแบบอัตโนมัตินั้นเพื่อที่จะช่วยให้ผู้ใช้งาน ได้รับคำตอบของคำถามที่ตรงตามความต้องการและมีความกระชับได้ใจความ ซึ่งโดยทั่วไปแล้วรูปแบบ "คำตอบ" ที่นำเสนอนั้นจะขึ้นกับประเภทคำถาม สำหรับ QA ในระยะเริ่มต้น ประเภทคำถามจะอยู่ในกลุ่ม who, when, why, what, where, และ how ต่อมามีการขยายประเภทของคำถามให้มีความหลากหลายขึ้น เช่น ข้อเท็จจริง (fact), ลำดับรายการ (list), นิยาม (definition), สมมติฐาน (hypo-thesis), ข้อจำกัดเชิงความหมาย (semantic constrain), และคำถามแบบข้ามภาษา (cross-lingual questions)

หากจำแนกกระบวนการถาม-ตอบ ตามโดเมนของการประมวลผลจะสามารถจำแนกได้ 3 แบบใหญ่ ๆ คือ การถามตอบที่จำแนกประเภทของกระบวนการถาม-ตอบตามโดเมน ประกอบไปด้วย 2 ประเภทย่อยได้แก่ กระบวนการถาม-ตอบแบบปิด (Closed-domain QA) และกระบวนการถาม-ตอบแบบเปิด (Opened-domain QA) สำหรับการถาม-ตอบแบบที่สองคือการถาม-ตอบตามเทคนิคที่ใช้ ซึ่งจะประกอบไปด้วย 3 ประเภทย่อยได้แก่ การวิเคราะห์โดยใช้หลักการทางภาษา, การวิเคราะห์ทางสถิติ และการเปรียบเทียบด้วยรูปแบบที่เตรียมไว้ และสุดท้ายคือการถาม-ตอบแบบที่จำแนกประเภทของกระบวนการถาม-ตอบตามลักษณะของคำตอบ จะประกอบไปด้วย 2 ประเภทย่อยได้แก่ กระบวนการถาม-ตอบแบบแฟคทอยด์ (Factoid QA) และ กระบวนการถาม-ตอบแบบนอนแฟคทอยด์ (Non-Factoid QA)

ซึ่งแต่ละประเภทของกระบวนการถาม-ตอบนั้นจะมีความเกี่ยวเนื่องกันอยู่ ยกตัวอย่างเช่น กระบวนการถาม-ตอบที่ถูกจำแนกประเภทตามเทคนิคที่ใช้ให้อยู่ในกลุ่มที่เป็นกระบวนการถาม-ตอบที่ใช้หลักการทางภาษาเข้ามาช่วย การถาม-ตอบประเภทนี้มักจะเป็นการถาม-ตอบที่ใช้ข้อมูลอยู่ในโดเมนแบบเปิด และมีลักษณะของคำตอบเป็นแบบนอนแฟคทอยด์ เป็นต้น

อย่างไรก็ตามกระบวนการถาม-ตอบ ส่วนใหญ่จึงมักจะอยู่ในลักษณะของโดเมนแบบปิดและมีลักษณะของคำตอบที่เป็นแฟคทอยด์ เพราะมีโครงสร้างคำถามที่สามารถคาดการณ์ได้มากกว่า

ในโดเมนแบบเปิด ทำให้อัตราการได้คำตอบที่ถูกต้องและตรงกับคำถามจะสูงกว่ามาก [3] ดังนั้นการนำกระบวนการถาม-ตอบ ไปใช้ มักจะนำไปใช้ในกระบวนการดังกล่าวหรืองานเฉพาะ

ในการประยุกต์ใช้งานของกระบวนการถาม-ตอบ เกือบจะทั้งหมดจะอยู่ในรูปแบบของกระบวนการถาม-ตอบแบบแฟคทอยด์ คือ การถาม-ตอบที่ต้องการคำตอบแบบสั้นๆ ด้วยการอาศัยกฎ (rule) หรือแม่แบบ (template) ที่มีการเรียนรู้เอาไว้ [3] ในการค้นหาและสกัดคำตอบ แต่ในคำถามบางประเภท ได้แก่ “ทำไม (why)” และ “อย่างไร (how)” [4] คำตอบสั้นๆ อาจจะไม่เพียงพอที่จะเป็นคำตอบ เพราะเป็นลักษณะคำถามที่ต้องการเหตุผล (Reason) หรือคำอธิบาย (Explanation) หรือการขยายความ (Expansion) ดังนั้น ในปัจจุบันการถาม-ตอบที่เป็นแบบนอนแฟคทอยด์ยังคงเป็นงานวิจัยที่มีความยากทั้งในเรื่องของการเลือกเอกสาร การค้นหาคำตอบ รวมไปถึงการประมวลผลคำตอบ [4, 5]

จากสาเหตุข้างต้น ในงานวิจัยฉบับนี้จึงนำเสนอการศึกษาและวิจัยเพื่อนำเสนอกระบวนการในการวิจัย เพื่อการหาคำตอบในการถาม-ตอบแบบนอนแฟคทอยด์สำหรับภาษาอังกฤษ โดยการอาศัยเทคนิคหลักๆ ในการประมวลผลคำถามและคำตอบ คือ การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) การค้นคืนเอกสาร (Information Retrieval: IR) และการสกัดสาร-สนเทศ (Information Extraction: IE)

1.2 วัตถุประสงค์ของการวิจัย

การวิจัยเพื่อนำเสนอกระบวนการของวิเคราะห์คำถามประเภท “Why” และ “How” และการหาคำตอบที่สอดคล้องกับคำถามประเภทเหล่านี้ ในการถาม-ตอบแบบนอนแฟคทอยด์สำหรับภาษาอังกฤษ

1.3 ความสำคัญของการวิจัย

1) เพื่อให้ได้กระบวนการของวิเคราะห์คำถามและการหาคำตอบที่สอดคล้องกับคำถามในกระบวนการของการถาม-ตอบแบบนอนแฟคทอยด์สำหรับภาษาอังกฤษ
2) เพื่อให้ได้กระบวนการต้นแบบของการถาม-ตอบแบบนอนแฟคทอยด์

1.4 ขอบเขตของการวิจัย

1) นำเสนอกระบวนการวิจัยเพื่อการหาวิเคราะห์คำถามและการหาคำตอบที่สอดคล้องกับคำถาม ในกระบวนการของการถาม-ตอบแบบนอนแฟคทอยด์สำหรับภาษาอังกฤษ
2) ประเภทของคำถามที่ใช้ในการศึกษาคือ “ทำไม (why)” และ “อย่างไร (how)”

3) กระบวนการที่นำเสนอมีส่วนประกอบสำคัญ 3 ส่วนได้แก่
   3.1) ส่วนของการประมวลผลคำถาม
   3.2) ส่วนของการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ
   3.3) ส่วนของการสกัดคำตอบที่สอดคล้องกับเอกสาร
4) เทคนิคในการวิเคราะห์คำถามและการค้นหาคำตอบหลักๆ คือ การประมวลผลภาษา-ธรรมชาติ การค้นคืนเอกสาร และการสกัดสารสนเทศ
5) ชุดข้อมูลมาตรฐานในการศึกษานี้คือ MS MARCO เวอร์ชัน 2.1 โดยมีชุดข้อมูลคำถามและคำตอบของ “*Why*” และ “*How*” จำนวน 88,491 ชุด และ 34,548 ชุด ตามลำดับ
6) การประเมินประสิทธิภาพจะทำใน 2 ส่วนคือ
   6.1) การประเมินการค้นหาเอกสารที่มีคำตอบจะทำด้วยค่าค่าความเฉลี่ยทั้งหมดของความถูกต้อง (Mean Average Precision: MAP) โดยสนใจค่าความถูกต้องใน *k* เอกสารแรกที่ค้นคืนได้ (Precision at K: P@k) โดย *k* ในที่นี้คือ 10
   6.2) การประเมินการค้นหาคำตอบในการถาม-ตอบแบบนอนแฟคทอยด์ จะประยุกต์ใช้อัลกอริทึมการวัดประสิทธิภาพของ การประเมินผลแบบสองภาษา (Bilingual Evaluation Understudy : BLEU-1) และ การประเมินของกิสติ้งโดยมุ่งเน้นที่ค่าควมละลึก (Recall-Oriented Understudy for Gisting Evaluation : ROUGE-L)

1.5 นิยามศัพท์เฉพาะ

1) การถาม-ตอบแบบแฟคทอยด์ (Factoid Question Answering) คือ การถาม-ตอบที่ต้องการคำตอบสั้นๆ ยกตัวอย่างเช่น การถามถึงชื่อสถานที่ การถามชื่อบุคคล เป็นต้น
2) การถาม-ตอบแบบนอนแฟคทอยด์ (Non-Factoid Question Answering) คือ การถาม-ตอบ ที่ต้องการคำตอบในลักษณะของการอธิบายหรือการให้เหตุผล ดังนั้นคำตอบที่ได้มักจะอยู่ในรูปแบบประโยคยาวๆ โดยทั่วไปคำถามที่ต้องการคำตอบในลักษณะนี้คือคำถามที่ขึ้นต้นด้วย “Why” และ “How”

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะเป็นการกล่าวถึงแนวคิด วิธีการ และอัลกอริทึมที่เกี่ยวข้อง ต่อการวิจัยและพัฒนากระบวนการการวิเคราะห์ความรู้สึกจากข้อความ โดยแนวคิดวิธีการและอัลกอริทึมที่เกี่ยวข้องต่องานนี้ ซึ่งมีดังต่อไปนี้

2.1 การถาม-ตอบ (Question Answering: QA)

การถาม-ตอบ [6] คล้ายกระบวนการการค้นคืนสารสนเทศ แต่ระบบถาม-ตอบนี้เป็นระบบที่สามารถให้คำตอบแบบอัตโนมัติให้กับผู้ใช้งานได้ด้วยคำตอบที่ได้จะมีความกระชับและสั้น [7] แทนที่จะเป็นการแสดงด้วยรายการเอกสาร ระบบถาม-ตอบเป็นสิ่งที่มีความสำคัญเพิ่มขึ้นในการค้นหาสารสนเทศ เพราะการค้นหาคำตอบที่มีความกระชับด้วยเซิร์สเอนจิ้นมาตรฐาน (Standard Search Engine) เป็นเรื่องที่ยากมาก

การถาม-ตอบเป็นงานในสาขาของการประมวลผลภาษาธรรมชาติ ที่สามารถประยุกต์ใช้ในของการประมวลผลภาษาธรรมชาติอื่นๆ ได้แก่ การค้นคืนสารสนเทศ และการสกัดสารสนเทศ โดยในส่วนของการค้นคืนสารสนเทศและระบบถาม-ตอบนั้น ถูกมองว่ามีความคล้ายคลึงกันอย่างมาก แต่ระบบถาม-ตอบ เป็นการค้นคืนสารสนเทศแบบข้อความสั้น (Short Passage Retrieval) และใช้เทคนิคการค้นคืนสารสนเทศและการสกัดสารสนเทศเข้ามาช่วยแทน ส่วนใหญ่แล้วนั้น QA จะถูกประยุกต์ในการสกัดสารสนเทศในระบบเปิด (Open-domain Information Extraction) เนื่องจากปัจุบันการเข้ามาของอินเตอร์เน็ตเป็นสิ่งสำคัญ อีกทั้งในโลกที่เปิดกว้างทำให้ขอบเขตในการเรียนรู้มีความท้าทาย และเปิดกว้างมากยิ่งขึ้น

นอกจากการประยุกต์ใช้ในการค้นคืนสารสนเทศและการสกัดสารสนเทศแล้ว ระบบถาม-ตอบ ยังสามารถพบในงานอื่นๆ ของการประมวลผลภาษาธรรมชาติ เช่น ระบบโต้ตอบ (Dialog systems) เพื่อใช้ในการสร้างระบบโต้ตอบแบบอัตโนมัติ ไปยังผู้ใช้งาน ระบบการอ่านเพื่อทำความเข้าใจ (Reading Comprehension Systems) เป็นต้น

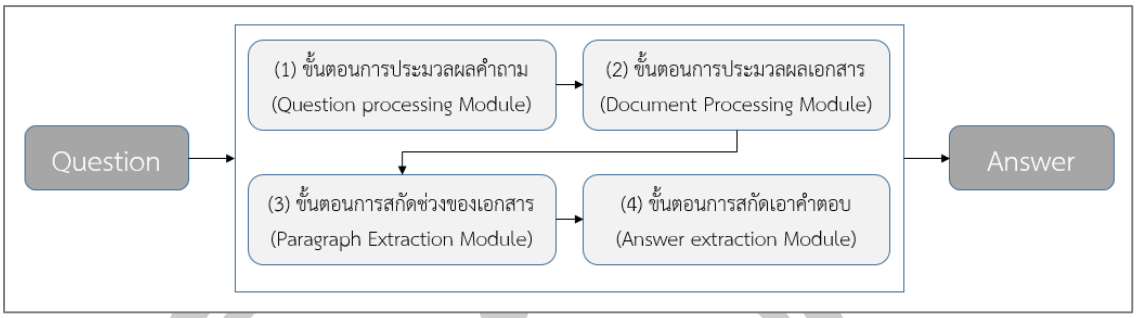
โดยทั่วไป คำถามพื้นฐานในการถาม-ตอบจะประกอบด้วยกลุ่มคำถาม 6 กลุ่ม คือ ใคร (who), ทำไม (why), เมื่อไหร่ (when), อันไหน (which), อะไร (what) และคำถามในกลุ่ม อย่างไร (how) ดังแสดงในตารางที่ 2.1

ตารางที่ 2.1 ประเภทของคำถาม [8]

กลุ่มของคำถาม	กลุ่มย่อยของคำถาม	ประเภทของคำตอบ	ตัวอย่าง
How	-	คำจำกัดความ	How did it happen?
Who	-	บุคคล	Who won the
Why	-	เหตุผล	Why don't we have enough this year?
When	-	วัน / เวลา	When this rain came yesterday?
Which	Which-who	บุคคล	Which person did invent the instrument of aerology?
	Which-where	สถานที่	Which city has the min temperature?
	Which-when	วัน / เวลา	Which month has max rain?
What	What	จำนวน, หัวข้อ, คำจำกัดความ, อื่นๆ	What is the temperature of Tehran?
	What-who	บุคคล	What is the best meteorologist in Iran?
	What-where	สถานที่	What is the capital of Iran?
	What-when	วัน / เวลา	What year do we have max rain?

2.2 สถาปัตยกรรมโดยทั่วไปของการถาม-ตอบ

โดยทั่วไปแล้วนั้นระบบถาม-ตอบ ไม่ว่าจะเป็นระบบถาม-ตอบในโดเมนแบบเปิด หรือระบบถาม-ตอบในโดเมนแบบปิด จะมีกรอบความคิด (Framework) หรือกระบวนการที่สำคัญอยู่ 4 ขั้นตอน [9] ประกอบไปด้วย ขั้นตอนการประมวลผลคำถาม (Question processing Module), ขั้นตอนการประมวลผลเอกสาร (Document Processing Module), ขั้นตอนการสกัดช่วงของเอกสาร (Paragraph Extraction Module) และขั้นตอนการสกัดเอาคำตอบ (Answer extraction Module) เพื่อที่จะนำเอาคำตอบที่ได้กลับไปยังผู้ใช้งาน [2] ดังแสดงใน รูปที่ 2.1



รูปที่ 2.1 ภาพรวมของระบบถาม-ตอบแบบทั่วไป (Generic Process of QA)

จากรูปที่ 2.1 จะเห็นว่าภาพรวมของระบบถาม-ตอบ มีกรอบความคิดที่จะประมวลผลคำถามเพื่อให้ได้ออกมาเป็นคำตอบของคำถามที่เป็นภาษาธรรมชาติเข้ามาสามารถอธิบายเป็นขั้นตอนได้ดังนี้

2.2.1 ขั้นตอนการประมวลผลคำถาม (Question Processing Module)

ในขั้นตอนนี้เป็นขั้นตอนในการประมวลผลคำถามที่เข้ามาในรูปแบบของภาษาธรรมชาติเพื่อแปลงคำถามที่เข้ามาให้เป็นคำถามที่คอมพิวเตอร์สามารถเข้าใจได้ หรือที่เรียกว่า “คำขอ (Query)” [1, 8] ระบบถาม-ตอบ ในขั้นตอนนี้มีกระบวนการหลักคือ กระบวนการระบุประเภทของคำตอบ [10] คือ กระบวนการที่จะระบุประเภทของคำตอบ โดยดูจากประเภทของคำถามที่เข้ามา [8] นอกจากนี้ยังเป็นการระบุกระบวนการหรือขั้นตอนต่อไปในการประมวลผล [10, 11] ตามกระบวนการที่กล่าวมานั้น สามารถแยกประเภทของคำถาม

จากการศึกษาพบว่าในขั้นตอนของการประมวลผลคำถามเพื่อให้ได้มาซึ่ง “คำขอ” นั้นยังคงมีปัญหาอยู่ซึ่งก็คือคำขอที่ได้นั้นไม่เพียงพอต่อการค้นคืนเอกสารที่เกี่ยวข้องกับสิ่งที่ต้องการเท่าที่ควร [8] จึงมีความจำเป็นที่จะต้องทำการขยายของเขตของคำค้น (Query Expansion) หรือการทำการกำหนดคำค้นใหม่จากคำถาม (Question Reformulation) [8, 10–12] เพื่อให้คำค้นที่ได้ในขั้นตอนการประมวลผลคำถามนั้นสามารถค้นคืนเอกสารให้มีความครอบคลุมมากยิ่งขึ้น [10] หรือก็เพื่อการเพิ่มค่าความละลึก (Recall) ให้มากขึ้น ยกตัวอย่างเช่น

Why the sky is blue ?

จากคำถามนี้จะเห็นว่าเมื่อนำมาทำการเตรียมข้อมูลด้วยการตัดคำและตัดคำหยุดออกไปแล้วนั้นคำค้นที่ได้จากคำถามเพื่อที่จะนำไปใช้ในการค้นคืนเอกสารที่ต้องการจากคำถามนี้นั้น สั้นเป็นอย่างมาก ทั้งยังสื่อความหมายได้กว้างจนเกินไป ทำให้การค้นคืนเอกสารนั้นมีจำนวนมากเกินความจำเป็นทำให้ระยะเวลาในการประมวลผลเอกสารเหล่านั้นมีความล่าช้าตามไปด้วย และเพื่อให้ได้

เอกสารที่เกี่ยวข้องและตรงตามความต้องการ จึงต้องทำการขยายขอบเขตของคำค้นด้วยวิธีการต่าง ๆ ดังนี้

2.2.1.1 *การขยายของเขตของคำค้น (*Query Expansion*)*

การขยายของเขตของคำค้น หมายถึง การสร้างคำค้นขึ้นมาใหม่โดยการใช้การขยายจากคำค้นเดิมเป็นฐาน [13] โดยทั่วไปแล้วนั้นการขยายขอบเขตของคำค้นนิยมใช้การขยายขอบเขตบนพื้นฐานของพจนานุกรม (Thesaurus-based Query Expansion) โดยพจนานุกรมที่ว่านั้นจะเป็นพจนานุกรมที่มีการรวบรวมเอาไว้อยู่แล้ว หรือสร้างขึ้นมาใหม่จากเอกสารในคลังก็ได้ ขึ้นอยู่กับการนำไปใช้งาน โดยพจนานุกรมที่ว่านี้ คือการนำคำค้นเดิมมาหาคำที่มีความหมายคล้ายคลึงกัน (Synonyms) เพื่อให้ได้มาซึ่งคำค้นที่มากขึ้น แต่ยังคงมีความหมายที่อยู่ในขอบเขตของคำค้นเดิม ต่อมาก็คือการใช้พจนานุกรมที่สร้างขึ้น ด้วยเทคนิคการวิเคราะห์ความสัมพันธ์ หรือการใช้หลักการทางภาษา รวมไปถึงการวิเคราะห์หน่วยของคำ (Morphological Analysis) เพื่อช่วยในการขยายขอบเขตของคำค้น และในขั้นตอนนี้ยังมีการใช้วิธีการแก้ไขคำผิดจากคำค้น ให้มีความถูกต้องจากคำค้นเดิมอีกด้วย

ในการถาม-ตอบนั้น การขยายขอบเขตของคำค้นจำเป็นที่จะต้องผ่านขั้นตอนการเตรียมข้อมูล (Data Preprocessing) เสียก่อน เพื่อให้ได้มาซึ่งคำค้น เพราะเนื่องจากเราไม่สามารถนำคำถามทั้งประโยคไปใช้ในการค้นหาเอกสารได้ทันที โดยทั่วไปการเตรียมข้อมูลในขั้นตอนนี้ ประกอบไปด้วย การตัดคำ การตัดคำหยุด เป็นต้น

จากประโยคตัวอย่าง : Why the sky is blue ?

คำค้นที่ได้หลังจากผ่านขั้นตอนการเตรียมข้อมูลแล้วนั้นมีเพียงคำว่า Sky และคำว่า Blue เท่านั้น ทำให้ค่าความละลึกที่ไม่สูงมากจึงจำเป็นที่จะต้องทำการเพิ่มค่าความละลึกด้วยการขยายขอบเขตของคำค้น ในที่นี้จะยกตัวอย่างการขยายขอบเขตของคำค้นด้วยการใช้พจนานุกรมความคล้ายคลึง ก็จะได้กลุ่มของคำค้นที่มากขึ้นกว่าเดิม

ตารางที่ 2.2 แสดงตัวอย่างการขยายขอบเขตคำค้น

Query	Synonym
Sky	Azure, Empyrean, Firmament
Blue	Sky-blue, Azure

จากตารางที่ 2.2 จะเห็นว่าคำค้นที่ได้ใหม่ประกอบไปด้วย Sky, Blue, Azure, Empyrean, Firmament และ Sky-blue โดยจำนวนของคำค้นที่ได้นั้นมากขึ้นกว่าเดิม เป็นที่แน่นอนว่าการทำการขยายขอบเขตของคำค้นนั้นจะสามารถเพิ่มค่าความละลึกได้ แต่ทั้งนี้ประสิทธิภาพและปริมาณ

ของคำค้นที่ได้นั้นก็ขึ้นอยู่กับพจนานุกรมที่ใช้ด้วย และยังมีข้อเสียคือประสิทธิภาพอาจจะลดลงในกรณีของคำที่มีความกำกวมทำให้การขยายขอบเขตนั้น ไปนำเอาเอกสารที่ไม่เกี่ยวข้องและไม่ต้องการขึ้นมาด้วย

2.2.1.2 *การกำหนดคำค้นใหม่จากคำถาม (*Question Reformulation*)*

การกำหนดคำค้นใหม่จากคำถาม หมายถึง การใช้เทคนิคการประมวลผลภาษาธรรมชาติในการประมวลผล ไม่ว่าจะเป็นการติดแทคให้กับคำถาม เพื่อวิเคราะห์โครงสร้างทางภาษา จากนั้นทำการตีความจากการวิเคราะห์เหล่านั้นเพื่อให้ได้ คำค้นที่มากกว่าเดิม ทำให้สามารถยกระดับคำค้นให้มีประสิทธิภาพมากยิ่งขึ้นในการค้นคืนเอกสารที่มีความเกี่ยวข้องผ่านกระบวนการการกำหนดใหม่ [14]

เทคนิคการกำหนดคำค้นใหม่จากคำถามนั้น มีความคล้ายคลึงกับการขยายขอบเขตของคำค้นแตกต่างกันตรงที่การประมวลผลไม่จำเป็นที่จะต้องทำการเตรียมข้อมูลมาก่อนสามารถนำคำถามที่เข้ามาไปประมวลผลได้เลยโดยใช้หลักการประมวลผลภาษาธรรมชาติต่าง ๆ เข้ามาช่วยในการวิเคราะห์ความสัมพันธ์ของคำที่อยู่ในคำถาม ยกตัวอย่างเช่นงานของ Agichtein และคณะ [12] ที่ได้นำเสนอระบบถาม-ตอบ ที่เรียกว่า Tritus โดยทำการศึกษาการทำการกำหนดใหม่ของคำค้น ด้วยการใช้การเรียนรู้คำถามเพื่อให้ได้มาซึ่งความสัมพันธ์ของรูปแบบในการตอบคำถามที่เกี่ยวเนื่องกับคำถามเพื่อให้ได้มาซึ่งกลุ่มของคำค้น ที่จะนำไปใช้ในการค้นคืนเอกสารในขั้นตอนถัดไป

Question Phrase	*Candidate Transforms*
"what is a"	"the term" "component" "ans" "a computer" "telephone" "collection of" "stands for" "unit"

รูปที่ 2.2 ตัวอย่างกลุ่มของคำค้นที่ได้หลังจากการกำหนดใหม่ [12]

จากนั้นทำการหาคำถามที่อยู่ในประโยคคำถามเพื่อดูว่าผู้ใช้งานต้องการถามถึงเรื่องอะไร และนำเอากลุ่มของคำค้นที่ได้จากการประมวลผลไปค้นคืนเอกสารร่วมกับคำนามที่ได้เพื่อให้ได้มาซึ่งเอกสารที่เกี่ยวข้อง ในการประมวลผลมากยิ่งขึ้น

2.2.2 ขั้นตอนการประมวลผลเอกสาร (Document Processing Module)

ขั้นตอนนี้เป็นการค้นคืนเอกสารที่มีแนวโน้มว่าจะมีคำตอบ [2] จากคลังข้อมูล โดยดูจากคำถามที่เป็นภาษาธรรมชาติของผู้ใช้ ในขั้นตอนนี้ประกอบด้วยอัลกอริทึมการสร้าง Query และเครื่องมือค้นหาข้อความ [15] ในการสกัดคำสำคัญจากคำถามออกมาจากเอกสารที่คาดว่าจะมีคำตอบ

เพื่อใช้ในการสร้าง Query โดยอาศัยเทคนิคการค้นคืนเอกสาร [10, 16] ในขั้นตอนนี้อาจจะมีการเพิ่มการขยายความคำสำคัญที่ได้มาเพื่อที่จะใช้ในการค้นหาบริบทที่ต้องการ [8]

2.2.3 ขั้นตอนการสกัดช่วงของเอกสาร (Paragraph Processing Module)

การสกัดย่อหน้าหรือการสกัดช่วงของเอกสาร [2] ในขั้นตอนนี้จะใช้เทคนิคการค้นหาช่วงของเอกสารในเอกสารที่คาดว่าจะมีคำตอบ จากนั้นทำการแบ่งเอกสารออกเป็นช่วงๆ และทำการคำนวณคะแนนของแต่ละช่วงเพื่อหาช่วงที่มีคะแนนสูงที่สุด เพื่อนำไปประมวลผลต่อไป [15, 17]

กลไกหลักที่ใช้ในขั้นตอนนี้จะใช้ทั้งเทคนิคที่เป็นเทคนิคการแบ่งเอกสาร (Tokenization Algorithm) และเทคนิคการให้คะแนนต่างๆ ยกตัวอย่างเช่น เทคนิคการหาความคล้ายคลึง (Similarity Algorithm) เป็นต้น [17]

2.2.4 ขั้นตอนการสกัดเอาคำตอบ (Answer Extraction Module)

ในขั้นตอนการสกัดคำตอบนี้ จะนำเอาช่วงของเอกสารที่ได้รับคะแนนสูงสุดจากขั้นตอนที่แล้วมาประมวลผลต่อ โดยการใช้เทคนิคที่เรียกว่าเทคนิคการสกัดคำตอบ (Answer Extraction Algorithms) [2] ในการพยายามสกัดเอาวลีหรือกลุ่มคำภายหลังจากที่เรานำเข้าประโยคจากช่วงของเอกสารที่ได้ เพื่อนำไปเทียบกับคำถามที่ป้อนเข้ามา เพิ่มความละเอียดในการวิเคราะห์คำตอบให้มีความถูกต้องแม่นยำที่สูงขึ้น [17] จากนั้นทำการคัดเลือกประโยคที่มีคะแนนสูงที่สุดมาประมวลผล [8]

ขั้นตอนสุดท้ายเป็นขั้นตอนในการระบุคำตอบ หลังจากที่เราได้ประโยคที่เราต้องการแล้วนั้นก็จะทำการตัดคำที่ไม่เกี่ยวข้องออกจนกว่าข้อความที่มีจะอยู่ในกรอบที่เราตั้งไว้ให้เป็นคำตอบ [2, 9]

กระบวนการหรือขั้นตอนเหล่านี้เป็นขั้นตอนหลักของการทำระบบถาม-ตอบแบบอัตโนมัติ ในบางงานวิจัยจะกล่าวถึงขั้นตอนการทำระบบถาม-ตอบ โดยจัดกลุ่มของขั้นตอนเหล่านี้เป็น 3 ขั้นตอน [8, 18] โดยจะรวม ขั้นตอนการประมวลผลเอกสาร และ ขั้นตอนการสกัดช่วงของเอกสารเข้าด้วยกันเนื่องจากสองขั้นตอนนี้มีความคล้ายคลึงกัน และทำงานอยู่กับข้อมูลชุดเดียวกัน

2.3 ประเภทของระบบถาม-ตอบ

ในการจำแนกประเภทของระบบถาม-ตอบ สามารถแบ่งออกได้เป็น 2 วิธี โดยวิธีแรกสามารถดูได้จากข้อมูลที่ใช้หรือแยกตามโดเมนของระบบถาม-ตอบแบ่งออกเป็นโดเมนแบบเปิดและโดเมนแบบปิด อีกหนึ่งวิธีในการจำแนกประเภทของระบบถาม-ตอบคือการใช้เทคนิคที่ใช้ในการสร้าง

ระบบถาม-ตอบ ในการจำแนกประเภทสามารถ แบ่งออกเป็น 3 กลุ่ม จากการจำแนกประเภทของระบบถาม-ตอบสามารถแสดงได้ดังหัวข้อต่อไปนี้

2.3.1 การจำแนกประเภทของระบบถาม-ตอบ ตามโดเมน

ด้วยการใช้เทคนิคที่เกี่ยวข้องมากมาย โดยการป้อนคำถามเข้าไปไม่ว่าจะเป็นคำถามที่ซับซ้อนหรือไม่ก็ตาม ขอบเขตของระบบถาม-ตอบ ในงานวิจัยส่วนใหญ่สามารถแบ่งออกเป็นสองกลุ่ม ได้แก่ ระบบถาม-ตอบในโดเมนแบบปิด (Closed-Domain Question Answering) และ ระบบถาม-ตอบ ในโดเมนแบบเปิด (Open-Domain Question Answering) ซึ่งโดยทั่วไปแล้วการถาม-ตอบถูกนำไปใช้เพื่อแยกคำตอบจากภาษาที่เป็นธรรมชาติ หรือภาษามนุษย์ [19] และระบบถาม-ตอบยังสามารถที่จะสกัดคำตอบจากข้อมูลที่อยู่ในรูปแบบอื่น ที่ไม่ได้อยู่ในรูปแบบที่เป็นข้อความได้ ยกตัวอย่างเช่นข้อมูลที่เป็นรูปภาพ ข้อมูลเสียง เป็นต้น [8] โดยอาศัยเทคนิคอื่น เข้ามาช่วยในการวิเคราะห์ข้อมูลเหล่านั้น

2.3.1.1 *ระบบถาม-ตอบในโดเมนแบบเปิด*

ระบบถาม-ตอบในโดเมนแบบเปิด [20] เป็นระบบที่ครอบคลุมหลักการพื้นฐานของระบบถาม-ตอบ ที่มีการนำเอาเทคนิคมากมายมาใช้ในการสร้างระบบโดยใช้ฐานข้อมูลที่เป็นข้อมูลทั่วไป แทนที่จะเป็นฐานข้อมูลที่มีโครงสร้าง ด้วยการเข้ามามีบทบาทของ WWW ทำให้ข้อมูลในโดเมนแบบเปิดนั้นมีจำนวนมหาศาล มีความหลากหลาย มีความซับซ้อนเป็นสาเหตุให้งานวิจัยในโดเมนแบบเปิดนั้นมีความยาก และมีความท้าทายเป็นอย่างมาก [1, 20] ในการสกัดเอาคำตอบเพื่อตอบกลับไปยังผู้ใช้งานจากข้อมูลเหล่านี้ ยกตัวอย่างเช่นการทำระบบถาม-ตอบบนเว็บ เนื่องจากขนาดของข้อมูลที่มีขนาดใหญ่นี้ทำให้ข้อเสียหลักของระบบถาม-ตอบในโดเมนแบบเปิดก็คือการที่เราต้องใช้เวลาในการประมวลผลมาก หากไม่มีกระบวนการหรือขั้นตอนการดำเนินงานที่ดีพอ และให้ค่าความถูกต้องต่ำ

2.3.1.2 *ระบบถาม-ตอบในโดเมนแบบปิด*

ระบบถาม-ตอบในโดเมนแบบปิด [21] คือลักษณะของโดเมนที่จำกัด ด้วยโดเมนที่จำกัดนี้ทำให้มีผลกระทบต่อประเภทคำถามที่ถามและคำตอบที่สามารถคาดหวังได้ ดังนั้นจะเห็นได้ว่าในโดเมนที่จำกัดต่างๆ มีประสิทธิภาพที่ต่างกันแม้ว่าจะใช้เทคนิคเดียวกัน [22] ระบบถาม-ตอบในโดเมนแบบปิดนั้นจำเป็นที่จะต้องรวบรวมข้อมูลที่จะใช้เป็นฐานข้อมูลในการค้นคืนคำตอบจากคำถามของผู้ใช้งาน ที่แตกต่างกันตามแต่ละงาน ข้อดีของการถาม-ตอบในโดเมนแบบปิดนี้คือการที่ระบบถาม-ตอบให้ประสิทธิภาพที่ดีกว่าโดเมนแบบเปิด เนื่องจากลักษณะของข้อมูลที่มีความจำเพาะเจาะจงไม่หลายหลาย และมีความซับซ้อนน้อยกว่าข้อมูลระบบถาม-ตอบในโดเมนแบบเปิด

2.3.2 การจำแนกประเภทของระบบถาม-ตอบ ตามเทคนิคที่ใช้

หากพิจารณาประเภทของการถาม-ตอบตาม เทคนิคที่ใช้เพื่อการประมวลผล จะสามารถแบ่งประเภทของการถาม-ตอบได้ 3 ประเภทคือ

2.3.2.1 *การวิเคราะห์โดยใช้หลักการทางภาษา (*Linguistic Analysis*)*

การวิเคราะห์โดยใช้หลักการทางภาษา [18] เป็นการวิเคราะห์ภาษาธรรมชาติด้วยหลักการทางภาษาเช่นการ ตัดคำ (Tokenization), การติดแท็กให้กับชนิดของคำ (Part-of-Speech Tagging: POS), การทำออนโทโลจี้ (Ontology), การทำโครงข่ายความสัมพันธ์เชิงความหมาย (Semantic Network) [2, 18] เป็นต้น เทคนิคเหล่านี้จะนำมาใช้สร้างเป็นคำขอที่มีความแม่นยำมากยิ่งขึ้น เพื่อที่จะนำคำขอนี้ไปค้นคืนเอกสารที่คาดว่าจะมีคำตอบจากฐานข้อมูลแบบมีโครงสร้าง [18]

2.3.2.2 *การวิเคราะห์ทางสถิติ (*Statistical Analysis*)*

การวิเคราะห์ทางสถิติ [33] นั้นมีความสำคัญในการดำเนินงานวิจัยเป็นอย่างมาก โดยมีการจัดการกับปัญหาในงานวิจัยต่างๆ เพื่อให้ได้ข้อสรุปที่มีความหมายสามารถนำไปอธิบายหรือให้เหตุผลได้ จากการดำเนินการในงานวิจัยเชิงประจักษ์ [23] มีหลายงานวิจัยที่ใช้เทคนิคการวิเคราะห์ข้อมูลในหลายส่วนของขั้นตอนการทำระบบถาม-ตอบ ด้วยเทคนิคต่างๆ ไม่ว่าจะเป็นเทคนิคการจำแนกข้อมูลด้วยซัพพอร์ทเวกเตอร์แมทชีน (Support Vector Machine), ตัวจำแนกโดยใช้ทฤษฎีของเบย์ (Bayesian Classifiers), ต้นแบบเอนโทรปีสูงสุด (Maximum Entropy Models) ยกตัวอย่างเช่น ขั้นตอนการวิเคราะห์คำถาม, ขั้นตอนการสกัดคำตอบ เป็นต้น [2, 24]

2.3.2.3 *การเปรียบเทียบรูปแบบ (*Pattern Matching*)*

การเปรียบเทียบรูปแบบ [35] นั้นเป็นเทคนิคที่ให้ประสิทธิภาพที่ดีกว่าเทคนิคอื่นๆ อีกทั้งเทคนิคการเปรียบเทียบรูปแบบนั้นเป็นเทคนิคที่ไม่ซับซ้อนง่ายต่อการเข้าใจ ยกตัวอย่างเช่น “ใครเป็นผู้ชนะเลิศรายการหน้ากากนักร้องซีซั่นที่ 1” จะเห็นว่าจากคำถามนี้เป็นการถามโดยอ้างอิงถึงบุคคล โดยระบบถาม-ตอบจะทำการแปลงคำถามให้เป็นรูปแบบของประโยคด้วย “ใครเป็นผู้ชนะเลิศ <Event Name>” หลังจากนั้นจะนำเอารูปแบบที่ได้มาประมวลผลเพื่อวิเคราะห์คำถาม และนำไปสกัดหาคำตอบเพื่อตอบกลับไปยังผู้ใช้งาน พร้อมทั้งเปลี่ยนรูปแบบให้อยู่ในรูปแบบของประโยคบอกเล่า “ผู้ชนะเลิศ <Event Name> คือ <Person>” ปัจจุบันงานวิจัยที่เกี่ยวกับระบบถาม-ตอบส่วนมากนั้น เรียนรู้ที่จะสร้างรูปแบบจากประโยคโดยใช้หลักการทางภาษาศาสตร์หรือเครื่องมือต่างๆ เช่น เทคนิคการรู้จำชื่อเฉพาะ (Named-Entity Recognizer), ภววิทยาหรือออนโทโลจี้ (Ontology), พจนานุกรมเวิร์ดเน็ท (WordNet Dictionary) และยังมีเครื่องมืออีกมากมายที่ไม่ได้กล่าวถึง ทั้งนี้

เพื่อที่จะใช้ในการวิเคราะห์และประมวลผลสำหรับการค้นคืนคำตอบ [18] จากที่กล่าวมาข้างต้นจะเห็นได้ว่าการเปรียบเทียบด้วยรูปแบบนั้นมีความง่ายในการใช้งานเป็นอย่างมาก จึงเหมาะสมที่จะนำไปใช้กับข้อมูลขนาดเล็กไปจนถึงข้อมูลขนาดกลาง แต่เทคนิคนี้ยังมีข้อเสียอยู่ที่ไม่สามารถแก้ปัญหาที่มีความซับซ้อนมากๆ ได้

เทคนิคการเปรียบเทียบด้วยรูปแบบนั้นมีเทคนิคที่ใช้กันอยู่หลักๆ อยู่ 2 เทคนิคด้วยกันคือ เทคนิคการใช้รูปแบบเบื้องต้น, เทคนิคการเปรียบเทียบแบบมีแม่แบบ โดยส่วนใหญ่แล้วเทคนิคการใช้รูปแบบเบื้องต้นนั้นจะถูกใช้ในการสร้างรูปแบบของคำถามเพื่อใช้ในการค้นคืนเอกสารมาเป็นองค์ความรู้ และเทคนิคการเปรียบเทียบแบบมีแม่แบบจะใช้สำหรับการสร้างรูปแบบของคำตอบ จากองค์ความรู้ที่ได้และตอบกลับไปยังผู้ใช้งาน [24]

เทคนิคที่ 1: *เทคนิคการใช้รูปแบบเบื้องต้น* (*Surface based patterns*)

เทคนิคการใช้รูปแบบเบื้องต้นนั้น เป็นกระบวนการในการสร้างรูปแบบของโครงสร้างเก็บไว้เพื่อที่จะนำไปใช้ในการค้นคืนเอกสาร รูปแบบเหล่านี้สามารถสร้างโดยมนุษย์หรือสร้างโดยอัตโนมัติได้ [24] จากนั้นนำเอกสารที่ค้นคืนได้มาเป็นองค์ความรู้สำหรับสกัดเอาคำตอบกลับไปตอบคำถามที่ผู้ใช้ป้อนเข้ามา โดยอาศัยความคล้ายคลึงของรูปแบบในองค์ความรู้ ที่มีความคล้ายคลึงกับรูปแบบของคำถามแบบมีนัยสำคัญต่อกัน เทคนิคนี้ถึงแม้ว่าจะเป็นเทคนิคที่ง่ายแต่หากต้องการความแม่นยำที่มากขึ้น ก็จะต้องใช้เวลาและความสามารถของผู้ออกแบบรูปแบบเป็นอย่างมากเช่นกัน [18]

เทคนิคที่ 2: *เทคนิคการเปรียบเทียบแบบมีแม่แบบ* (*Template based patterns*)

เทคนิคการเปรียบเทียบแบบมีแม่แบบ นั้นเป็นเทคนิคที่มีประสิทธิภาพสูงที่สุด โดยมีจุดเด่นอยู่ที่การสามารถอธิบายที่มาที่ไปของคำถามและคำตอบได้ [2] เทคนิคนี้นั้นเป็นการสร้างรูปแบบจากฐานข้อมูลนำมาสร้างเป็นรูปแบบ เพื่อใช้ในการค้นคืนเอกสารขึ้นมาโดยมีงานวิจัยอีกมากที่ประสบความสำเร็จในการนำเอาการใช้เทคนิคการเปรียบเทียบแบบมีแม่แบบนี้ไปใช้งาน และขยายขอบเขตการใช้งานให้กว้างขึ้น ยกตัวอย่างเช่นการทำ Search Engine เป็นต้น [25]

2.3.3 การจำแนกประเภทของระบบถาม-ตอบ ตามลักษณะของคำตอบ

การถามตอบโดยแบ่งตามลักษณะของคำตอบการตอบคำถามนั้น สามารถแบ่งออกได้เป็น 2 ประเภทคือการถาม-ตอบแบบแฟคทอยด์ และการถาม-ตอบแบบนอนแฟคทอยด์ดังแสดงในหัวข้อต่อไปนี้

2.3.3.1 *การถาม-ตอบ แบบแฟคทอยด์*

ลักษณะของระบบถาม-ตอบ ประเภทนี้เป็นระบบถาม-ตอบ ที่เป็นที่นิยมมากที่สุดตั้งแต่อดีตจนถึงปัจุบัน โดยจะมีลักษณะของคำตอบที่เฉพาะเจาะจง หรือเป็นชื่อเฉพาะ ยกตัวอย่างเช่น

Question: *Where is the capital city of Thailand*?

Answer: *Bangkok*.

โดยคำตอบที่ตอบกลับไปยังผู้ใช้งานอาจจะอยู่ในรูปของคำตอบ หรืออาจจะสร้างเป็นรูปแบบของคำตอบก็ได้ โดยคำตอบที่ได้จะขึ้นอยู่กับคำถามที่ถามเข้ามาว่าคำถามที่ได้เข้ามาเกี่ยวกับเรื่องใดเป็นต้น

2.3.3.2 *ระบบถาม-ตอบ แบบนอนแฟคทอยด์*

สำหรับระบบถาม-ตอบ ประเภทนี้เป็นการถาม-ตอบที่มีลักษณะของคำตอบในรูปแบบการอธิบายหรือการให้เหตุผล ซึ่งโดยส่วนใหญ่แล้วคำถามประเภทนี้จะเป็นคำถามประเภท “ทำไม (Why)” หรือ “อย่างไร (How)” ระบบถาม-ตอบ ประเภทนี้มีความยากกว่าระบบถาม-ตอบแบบแฟคทอยด์ เป็นอย่างมากตั้งแต่กระบวนการสกัดคำตอบและกระบวนการวัดประสิทธิภาพ เนื่องจากระบบถาม-ตอบ แบบนอนแฟคทอยด์ นั้นมีความยาวของคำตอบที่ไม่มีข้อจำกัดทำให้ไม่ทราบขนาดของคำตอบที่ชัดเจน และยังรวมไปถึงการค้นคืนบางส่วนของเอกสารขึ้นมาเพื่อสกัดเอาคำตอบก็ทำได้ยาก [4, 5] อย่างไรก็ตามปัญหาดังกล่าวมีความใกล้เคียงกับปัญหาของการทำระบบค้นคืนสารสนเทศทั้งกระบวนการในส่วนนี้ยังคงต้องการความรวดเร็วในการทำงาน [5] นอกจากความยากที่กล่าวมาข้างต้นนี้แล้วนั้นปัจุบันยังมีงานวิจัยที่ทำการวิจัยเกี่ยวกับระบบถาม-ตอบประเภทนี้น้อยมากจึงเป็นความท้าทายอย่างมาก

จากตารางที่ 2.3 จะเห็นว่าประเภทของคำถามทั้ง 6 กลุ่ม จะพบว่าในกลุ่มของประเภทคำถามที่เป็นนอนแฟคทอยด์ มีอยู่เพียง 3 กลุ่ม ได้แก่ How, Why และ What ในประเภทของคำถามที่เป็น What นั้นสามารถเป็นได้ทั้งแบบที่เป็นแฟคทอยด์ และนอนแฟคทอยด์ได้โดยสามารถยกตัวอย่างได้ดังต่อไปนี้

Question: *What does "buck naked" mean*?

Answer: *Dating from the 1920's the expression buck naked commonly means completely or totally naked or without a “stitch” (as opposed to partially naked). Synonyms include “bare naked”, “buck-arse naked”, and “butt naked” (also spelled, facetiously, “butt nekkid”). In the South, “buck naked” or “butt*

naked" means not in your house, out looking for trouble on Saturday night and wearing no clothes.

ตารางที่ 2.3 แสดงความเกี่ยวสัมพันธ์ประเภทระบบถาม-ตอบ

กลุ่มของคำถาม	กลุ่มย่อยของคำถาม	ประเภทของคำตอบ	Factoid	Non-Factoid
How	-	คำจำกัดความ	/	/
Who	-	บุคคล	/	-
Why	-	เหตุผล	-	/
When	-	วัน / เวลา	/	-
Which	Which-who	บุคคล	/	-
	Which-where	สถานที่	/	-
	Which-when	วัน / เวลา	/	-
What	What	จำนวน, หัวข้อ, คำจำกัดความ, อื่นๆ	/	/
	What-who	บุคคล	/	-
	What-where	สถานที่	/	-
	What-when	วัน / เวลา	/	-

2.4 เทคนิค แนวคิด และกระบวนการที่เกี่ยวข้อง

2.4.1 การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP)

การประมวลผลภาษาธรรมชาติ [26] คือ สาขาย่อยของปัญญาประดิษฐ์และภาษาศาสตร์ที่ศึกษาปัญหาในการประมวลผลและใช้งานภาษาธรรมชาติ รวมทั้งการทำความเข้าใจภาษาธรรมชาติ ทั้งนี้เพื่อให้คอมพิวเตอร์สามารถเข้าใจภาษามนุษย์ได้ โดยการจำลองความสามารถของมนุษย์ยกตัวอย่างเช่น การแก้ปัญหาโดยใช้เหตุผล, การเรียนรู้วิธีการแก้ปัญหาด้วยตนเอง รวมไปถึงการสื่อสารด้วยภาษามนุษย์ ในภาษาธรรมชาตินี้แบ่งออกเป็น 2 รูปแบบคือ รูปแบบเสียงหรือภาษาพูด (Speech) และ รูปแบบอักษรหรือภาษาเขียน (Text) [27] โดยการแบ่งแยกหน่วยทางภาษานั้นจะประกอบไปด้วย

1. ตัวอักษร (Alphabet) คือสัญลักษณ์ที่ใช้แทนเสียง เป็นกลุ่มของสัญลักษณ์ที่จำกัดกลุ่ม
2. คำ (Word) คือ กลุ่มของตัวอักษรที่มาเรียงต่อกันเป็นคำ เช่น ฉัน, กิน เป็นต้น
3. วลี (Phase) คือ กลุ่มของคำที่เป็นการนำคำตั้งแต่สองคำขึ้นไปมาเรียงต่อกันทำให้เกิดความหมายเพิ่มขึ้น เช่น กินข้าว, ไปเที่ยว เป็นต้น
4. ประโยค (Sentence) คือ กลุ่มของคำที่นำมาเรียงต่อกันเพื่อแทนความหมาย ประโยคจึงอยู่ในรูปของข้อความ เช่น การมาเรียนเป็นเรื่องสนุก เป็นต้น

ระดับของการประมวลผลภาษาธรรมชาติ มีทั้งสิ้น 5 ระดับ [26] คือ

1. Morphological Analysis เป็นการวิเคราะห์หน่วยคำ โดยจะวิเคราะห์ว่า คำหนึ่งๆ สามารถแยกย่อยได้อะไรบ้าง และคำๆ นั้นมีหน้าที่อะไร
2. Syntactic Analysis เป็นการวิเคราะห์ทางไวยากรณ์ เพื่อให้รู้ว่าประโยคหนึ่งๆ มีโครงสร้างเชิงวากยสัมพันธ์อย่างไร
3. Semantic Analysis เป็นการวิเคราะห์ความหมายของประโยคนั้น ๆ
4. Discourse Integration โดยการพิจารณาความหมายของประโยค โดยพิจารณาจากประโยคข้างเคียง
5. Pragmatic Analysis เป็นการแปลความหมายของประโยค เพื่อดูความตั้งใจในการสื่อสารของผู้สื่อสาร ว่าจุดประสงค์จะกล่าวถึงอะไร

ในระบบถาม-ตอบ นั้นจำเป็นที่จะต้องใช้การประมวลผลภาษาธรรมชาติเข้ามาช่วยในการประมวลผลทั้งเรื่องของการตีความหมายของคำถามไปจนถึงขั้นตอนของการประมวลผลคำตอบกลับไปยังผู้ใช้ ประกอบไปด้วยเทคนิคที่ใช้ในการประมวลผลภาษาธรรมชาติมากมาย เพื่อให้เครื่องจักร (Machine) นั้นสามารถที่จะทำความเข้าใจ เรียนรู้ และประมวลผลภาษาธรรมชาติดังกล่าวได้ยกตัวอย่างเช่น

2.4.1.1 *การตัดคำ* (Word Segmentation)

เนื่องจากการประมวลผลภาษาธรรมชาติ ที่ใช้ข้อมูลที่เกี่ยวข้องกับข้อความจำเป็นอย่างยิ่งที่จะต้องหาขอบเขตของแต่ละคำ เพื่อที่จะสามารถทำการประมวลผลกับข้อความในเอกสารได้อย่างสะดวกและจากลักษณะของการเขียนภาษาไทยนั้นไม่มีการใช้ตัวอักษรหรือสัญลักษณ์ที่นำมาใช้คั่นระหว่างคำหรือว่ามีการวรรคระหว่างคำเหมือนภาษาอังกฤษ โดยการตัดคำ [28, 29] คือ การแบ่งตัวอักษรจากข้อความ (String) เพื่อหาขอบเขตของแต่ละหน่วยคำ (Morpheme) ในภาษาอังกฤษจะ

มีการใช้ช่องว่าง (Space) คั่นระหว่างคำเพื่อเป็นตัวแบ่งคำ ซึ่งทำให้ง่ายต่อการกำหนดขอบเขตของคำ ดังตารางที่ 2.4

ตารางที่ 2.4 แสดงตัวอย่างการตัดคำในภาษาอังกฤษ

ประโยคเดิม	ประโยคที่ผ่านการตัดคำ
Who has won The Mask single season 1	Who / has / won / The / Mask / single / season / 1
Where is the capital of Thailand	Where / is / the / capital / of / Thailand

2.4.1.2 *การตัดคำหยุด (*Stop words*)*

การตัดคำหยุด [30] คือ กระบวนการตัดคำหรือสัญลักษณ์ที่พบบ่อยมากในเอกสาร แต่คำหรือสัญลักษณ์เหล่านั้นไม่ได้ส่งผลต่อการประมวลผลภาษาธรรมชาติ ดังนั้นเมื่อทำการตัดออกแล้วไม่ทำให้ใจความในเอกสารนั้นๆ เปลี่ยนไป เป็นการนำคำที่ไม่มีนัยสำคัญออกโดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง คำที่ไม่มีนัยสำคัญ ในที่นี้หมายถึงคำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจาก เอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง ตัวอย่างเช่น คำบุพบทเป็นคำที่ใช้เชื่อมคำหรือ กลุ่มคำให้สัมพันธ์กัน คำสันธานเป็นคำที่ทำหน้าที่เชื่อมคำกับคำ คำสรรพนามเป็นคำที่ใช้แทน คำนามที่กล่าวถึงมาแล้วในประโยค เป็นต้น คำหยุดมักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเอกสาร และปรากฏในเอกสารเกือบทุกฉบับ จึงถือได้ว่าคำหยุดเป็นคำที่ไม่มีประโยชน์หรือไม่เกี่ยวข้องใน การค้นคืนดังนั้นการกำจัดคำหยุดจึงเป็นมีความจำเป็นที่จะทำก่อนการจัดทำดัชนีเพื่อจัดการกับคุณลักษณะที่ไม่เป็นประโยชน์พร้อมทั้งลดขนาดหรือจำนวนของดัชนีลง ส่งผลให้สามารถประหยัดทั้งพื้นที่และเวลาในการประมวลผล ตัวอย่างคำหยุด เช่น ที่ (At), ใน (In), และ (And), จะ (Will) เป็นต้น

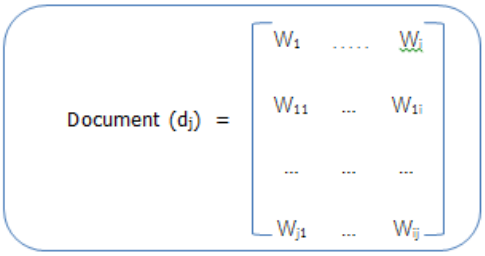
การตัดคำหยุด มีความจำเป็นอย่างมากในการประมวลผลภาษาธรรมชาติ เพราะจะช่วยลดระยะเวลาในการประมวลผลลงได้เป็นอย่างมาก เนื่องจากระบบฯ จะไม่เสียเวลาในการประมวลผลคำเหล่านี้ ในกรณีภาษาอังกฤษ อาจจะหมายถึง กลุ่มคำจำพวก Article หรือ กลุ่มคำที่ใช้แทนชื่อ คน สัตว์ สิ่งของ เป็นต้น ตารางที่ 2.5

ตารางที่ 2.5 แสดงการตัดคำหยุดในภาษาอังกฤษ

ประโยคที่ผ่านการตัดคำ	คำที่ได้จากประโยค
Pond / in / the / morning / time	won / Mask / single / season / 1
Where / is / the / capital / of / Thailand	capital / Thailand

2.4.1.3 Vector Space Model

Vector Space Model (VSM) [31] คือการนำเสนอเอกสาร (Document Representation) เป็นการแปลงข้อมูลจาก Text ให้อยู่ในรูปแบบของ VSM โดยค่าที่อยู่ใน VSM นี้จะเป็นค่าที่ถูกถ่วงน้ำหนักหรือไม่ก็ได้ จากนั้นจัดเก็บเป็นส่วนหนึ่งในถุงของคำ หรือที่เรียกกันว่า Bag of Word (BOW) ซึ่งวิธีการนี้เป็นวิธีการเตรียมข้อมูล Text ในรูปเวกเตอร์ของคำ ก่อนเข้าสู่การประมวลผลโดยเฉพาะทางด้านการประมวลผลภาษาธรรมชาติ หรือการค้นคืนสารสนเทศ การแทนเอกสารด้วยรูปแบบ BOW เป็นการกำหนด “คุณลักษณะ (Feature)” ในเอกสารด้วย W_{ij} ดังนั้นเอกสารลำดับที่ j ใดๆ สามารถเขียนแทนได้ด้วย $d_j = (w_{11}, w_{12}, w_{13}, ..., w_{ij})$



รูปที่ 2.2 ภาพแสดงการนำเสนอเอกสารในรูปแบบของ VSM

2.4.1.4 *การให้น้ำหนัก* (Term Weighting)

การให้น้ำหนัก [32] คือการแสดงความสัมพันธ์ระหว่าง คุณลักษณะ (Feature) และเอกสาร ที่มีอยู่ ในรูปแบบของเวกเตอร์ การกำหนดค่าน้ำหนัก “คุณลักษณะ” เพื่อการเรียนรู้และทำการทดสอบ การให้น้ำหนักเป็นหลักการให้ความสำคัญกับคุณลักษณะเหล่านั้นในด้านของค่าสถิติศาสตร์ เช่น การให้ค่าสถิติความถี่ (frequency) การเกิดขึ้นของคุณลักษณะนั้นๆ ในบางงานคุณลักษณะที่ได้จะเป็นคำ หรือวลี ในการแทนคุณลักษณะของเอกสาร จากหลักไวยากรณ์ภาษาในการใช้คำหรือวลีเดิมซ้ำๆ [26, 32] เพื่อเป็นการเน้นย้ำถึงความสำคัญของคุณลักษณะที่พบในเอกสาร ส่วนการนำเสนอเอกสาร เป็นการจัดรูปแบบเอกสารที่แสดงความสัมพันธ์ระหว่างเอกสารและคุณลักษณะที่พบในเอกสารนั้น ในรูปแบบของเมตริกซ์แบบ 2 มิติที่เรียกว่า Vector Space Model หรือ ถุงของคำ อย่างไรก็ตาม เพื่อให้ความสัมพันธ์ระหว่างเอกสารและคุณลักษณะมีความน่าเชื่อถือ ซึ่งรูปแบบนี้ เป็นรูปแบบที่พร้อมต่อการนำเอาเอกสารเหล่านี้เข้าสู่กระบวนการเรียนรู้เพื่อสร้างโมเดลด้วยอัลกอริทึมการเรียนรู้ของเครื่อง (Machine Learning)

การให้น้ำหนักคุณลักษณะดังกล่าวนั้นมีความสำคัญเป็นอย่างมาก เนื่องจากวิธีการนี้จะทำให้ศักยภาพของการจัดทำดัชนี ในการสร้างระบบค้นคืนสารสนเทศ หรือแม้แต่การสร้างดัชนีสำหรับระบบถาม-ตอบ [33] วิธีการให้น้ำหนักคุณลักษณะมีหลายวิธี เช่นการให้น้ำหนักคำด้วยความถี่ของ

คุณลักษณะที่พบในเอกสาร (Term frequency: TF) การให้น้ำหนักคุณลักษณะด้วยส่วนกลับของคุณลักษณะดังกล่าวพี่พบในเอกสาร (Inverse document frequency: IDF) หรือการนำเอาวิธีการดังกล่าวมาไว้ด้วยกัน (Term frequency–Inverse document frequency: TF-IDF)

วิธีการให้น้ำหนักด้วยความถี่

ในวิธีการให้น้ำหนักคำด้วยความถี่นี้ [34] สำหรับวิธีการนี้ง่ายที่สุดคือการนับคำที่พบในเอกสาร ยกตัวอย่างเช่นคุณลักษณะแทนด้วย t ที่พบในเอกสาร เกิดขึ้นในเอกสารแทนด้วย d จากนั้นนับคำที่เกิดขึ้นในเอกสารสามารถเขียนความเป็นไปได้ เช่น การให้น้ำหนักแบบเทียบกับความยาวของเอกสารโดยการนำเอาความถี่ของคุณลักษณะที่พบในเอกสารส่วนด้วยคุณลักษณะทั้งหมดที่พบในเอกสาร

$$tf(t,d) = \frac{f_{t,d}}{n} \tag{1}$$

วิธีการให้น้ำหนักด้วยการหาส่วนกลับของเอกสาร

วิธีการนี้เปรียบเสมือนการหาความสำคัญของคุณลักษณะว่าคุณลักษณะที่พบนั้นมีความสำคัญมากน้อยเพียงใดต่อเอกสารทั้งหมด [34] โดยการใช้ลอการิทึมในการหาส่วนผกผันระหว่างคุณลักษณะกับเอกสารด้วยการหาร เอกสารทั้งหมดที่พบคุณลักษณะนั้นอยู่ จากนั้นนำค่าลอการิทึมที่ได้มาเป็นค่าน้ำหนัก ในวิธีการนี้สามารถหาค่าส่วนกลับของเอกสารด้วยสมการดังต่อไปนี้

$$idf(t,D) = log\frac{N}{|d \in D : t \ \in d|} \tag{2}$$

เมื่อ N แทนด้วยจำนวนของเอกสารทั้งหมด

วิธีการให้น้ำหนักด้วยการหาความถี่และส่วนกลับของเอกสาร

เทคนิคการให้น้ำหนักด้วยการหาความถี่และส่วนกลับของเอกสารเป็นเทคนิคที่ใช้ในการกรองคุณลักษณะที่พบบ่อยในเอกสาร และให้น้ำหนักคุณลักษณะเหล่านั้นด้วยสมการลอการิทึม [34] หากว่าคุณลักษณะใดพบบ่อยในเอกสาร ค่าที่ได้จากสมการจะมีค่าเข้าใกล้ 0 สามารถหาได้จากสมการดังต่อไปนี้

$$tf - idf(t,d,D) = tf(t,d) \times idf(t,D) \tag{3}$$

2.4.1.5 *การวิเคราะห์ความคล้ายคลึง (*Similarity Analysis*)*

การวิเคราะห์ความคล้ายคลึง [35] เป็นการวิเคราะห์เพื่อหาความสอดคล้องหรือความเหมือนของเอกสารที่เป็นคำขอและเอกสารที่ต้องการตรวจสอบ ซึ่งจะเป็นการวัดความคล้ายคลึงระหว่าง คำ, โครงสร้างประโยค ไปจนถึงการวัดความคล้ายคลึงของประโยคก็ได้ หรือบางครั้งอาจจะเป็นการวิเคราะห์เพื่อตรวจสอบทั้งเอกสาร [36] โดยในการวิเคราะห์ความคล้ายคลึงนี้มีฟังก์ชั่นที่ใช้ในการให้คะแนนความคล้ายคลึงอยู่หลากหลายวิธี แต่จากวิธีการทั้งหมดนั้นมีแนวคิดที่ใช้ในการคำนวณคะแนนความคล้ายคลึง [37] โดยมีข้อมูลอยู่ทั้งสิ้นสองชุดคือ ชุดข้อมูลที่เป็นคำขอ กับชุดของข้อมูลที่เป็นข้อมูลที่อยู่ในประโยคหรือเอกสารที่ต้องการนำไปเปรียบเทียบ สามารถเขียนได้เป็น $X = \{x_1, x_2, \dots, x_n\}$ โดยที่ x_i เป็นแอทริบิวต์ที่พบเจอในคำขอ และ $Y = \{y_1, y_2, \dots, y_n\}$ โดยที่ y_n เป็นแอทริบิวต์ที่พบเจอในเอกสารที่จะทำการเปรียบเทียบ X และ Y หมายถึงการแสดงเวกเตอร์แบบ 2 กลุ่ม (Binary) ของคำขอของผู้ใช้และในเอกสารโดยระบุว่ามีหรือไม่มีแอทริบิวต์ในกลุ่ม ซึ่งในงานวิจัยนี้จะใช้การวัดค่าความคล้ายคลึงด้วย Multi Aspect TF-IDF (MATF) ในการให้คะแนนความคล้ายคลึงระหว่างคำค้นและเอกสาร โดยสารมารถแสดงได้ด้วยสมการดังต่อไปนี้

$$MATF_{(Q,D)} = \sum_{i=1}^{|Q|} TFF_{(q_i,D)} \times TDF_{(q_i,C)} \tag{4}$$

กำหนดให้ Q หมายถึงจำนวนคำค้นทั้งหมด D หมายถึงเอกสารที่จะทำการเปรียบเทียบ และ C แทนเอกสารทั้งหมดในคลังข้อมูล

สมการของ MATF จะแบ่งออกเป็น 2 ส่วนใหญ่ ๆ คือส่วนของการให้น้ำหนักภายในเอกสารหรือที่เรียกว่า (Local Weight) และส่วนของการให้น้ำหนักกับเอกสารทั้งหมดหรือที่เรียกว่า (Global Weight) นั่นก็คือค่า TFF และ TDF ตามลำดับโดยสามารถหาได้จากสมการดังต่อไปนี้

$$TFF_{(q,D)} = w \times BRITF_{(t,D)} + (1 - w) \times BLRTF_{(t,D)} \tag{5}$$

จากสมการของ TFF จะเป็นการให้ค่าน้ำหนักกับความยาวของเอกสาร ว่าเราสนใจที่จะให้ค่าน้ำหนักกับเอกสารที่มีความยาวหรือสั้นมากกว่ากัน ประกอบไปด้วยสมการของ BRITF และ BLRTF โดยสามารถแจกแจงออกมาได้ดังสมการต่อไปนี้

$$BRITF_{(q,D)} = \frac{RITF_{(t,D)}}{1 + RITF_{(t,D)}} \tag{6}$$

$$RITF_{(q,D)} = \frac{TF_{(t,D)}}{Avg.TF_{(D)}} \tag{7}$$

เมื่อพิจารณาจากสมการของ BRITF จะพบว่าเป็นการนำค่าความถี่ของคำที่เกิดในเอกสารมาทำให้ค่ามีความสมดุลมากขึ้นหรือการทำนอมอลไรซ์ (Normalize) นั่นเอง

$$BLRTF_{(q,D)} = \frac{LRTF_{(t,D)}}{1 + LRTF_{(t,D)}} \quad (8)$$

$$LRTF_{(q,D)} = TF_{(t,D)} \times log_2(1 + \frac{ADL_{(C)}}{length_{(D)}}) \quad (9)$$

สมการของ BLRTF นั้นจะเป็นการดูค่าความยาวของเอกสารที่กำลังพิจารณาว่ามี ความสอดคล้องกับความยาวของเอกสารที่เหลือในคลังข้อมูลหรือไม่ และคูณกับค่าน้ำหนักที่เราให้ความสนใจ โดยใช้ค่าความยาวเฉลี่ยของเอกสารที่อยู่ในคลังข้อมูล (Average Document Length : ADL) เทียบกับความยาวของเอกสารปัจจุบัน

$$TDF_{(a,C)} = IDF_{(q,C)} \times \frac{AEF_{(q,C)}}{1 + AEF_{(q,C)}} \quad (10)$$

ในสมการของ TDF เป็นการพิจารณาค่าน้ำหนักของคำค้นภายกับเอกสารปัจุบัน ว่ามีผลกระทบต่อคลังเอกสารโดยรวมเป็นเท่าใด โดยค่า AEF สามารถหาได้จากสมการดังต่อไปนี้

$$AEF_{(q,\text{ฉ})} = \frac{CTF_{(q,C)}}{DF_{(q,C)}} \quad (11)$$

โดยกำหนดให้ CTF เป็นจำนวนความถี่ของคำค้นที่พบในเอกสารทั้งหมด และค่า DF คือจำนวนเอกสารที่ปรากฏคำค้นนั้น ๆ อยู่ในเอกสาร

2.4.1.6 *การจัดอันดับเอกสารที่เกี่ยวข้อง* (Document Ranking)

ในการจัดอันดับเอกสารจะมีเทคนิคในการหาข้อความที่เกี่ยวข้องหรือสอดคล้องหรือตรงกับคำถาม งานวิจัยนี้จะทำการประยุกต์ใช้เทคนิคของการจัดอันดับเอกสาร (Ranking) ด้วยเทคนิคของ Okapi BM25 ในการหาความคล้ายคลึงของข้อความที่เกี่ยวข้องกับคำถาม โดยใช้คำสำคัญ และกลุ่มของคำหยุดที่ได้หาไว้ข้างต้นในการให้คะแนนส่วนหนึ่งของเอกสารที่คาดว่าจะมีคำตอบอยู่ โดยในขั้นตอนนี้จะทำการค้นคืนย่อหน้าของเอกสารขึ้นมาก่อนจากนั้นจึงทำการจัดอันดับย่อหน้าที่มีคะแนนมากที่สุดเพื่อนำไปหาส่วนที่คาดว่าจะเป็นคำตอบภายในย่อหน้านั้น ๆ โดยใช้สมการดังต่อไปนี้

$$BM25(Q,P) = \sum_{i=1}^{|Q|} \log\left(\frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)}\right) \times \frac{(k_1 + 1) \cdot f_i}{K + f_i} \times \frac{(k_2 + 1) \cdot qf_i}{k_2 + qf_i} \quad (12)$$

เมื่อค่า k_1 และ k_2 นั้นเป็นค่าพารามิเตอร์ที่สามารถปรับได้ (Hyper-Parameter) และค่า K สามารถหาได้จาก

$$K = k_1 \cdot ((1-b) + b \cdot \frac{doc_length}{avg_doc_length}) \quad (13)$$

โดยที่ปกติแล้วค่า k_1 จะอยู่ที่ 1.2 ค่า k_2 อยู่ระหว่าง 0 ถึง 1,000 และค่าอคติ (Bias : b) นั้นจะมีค่าอยู่ระหว่าง 0 ถึง 1 โดยส่วนใหญ่อยู่ที่ 0.75 ซึ่งเป็นค่าที่ไม่เกี่ยวข้องกับชุดข้อมูลใด ๆ การใส่ค่าดังกล่าวเข้าไปเพื่อที่จะเป็นการปรับระดับของการปรับสมดุลให้กับอัลกอริทึมการจัดอันดับอ้างอิงจากงานประชุมวิชาการที่เป็นองค์กรณ์ในการกำหนดมาตรฐานสำหรับการค้นคืนสารสนเทศรวมไปถึงการจัดอันดับเอกสารที่ค้นคืนได้ [38, 39]

ต่อมาคือค่า r_i หมายถึงจำนวนเอกสารที่เกี่ยวข้องและพบคำค้นที่ i โดยมีเงื่อนไขว่าหากเอกสารที่นำมาจัดอันดับนั้นมีความเกี่ยวข้องทั้งหมด ให้ระบุค่า $r_i = 0$

ค่า n_i หมายถึงจำนวนของเอกสารหรือส่วนของเอกสารที่พบคำค้นที่ i

ค่า N หมายถึงจำนวนเอกสารหรือส่วนของเอกสารทั้งหมด

ค่า R หมายถึงจำนวนของเอกสารหรือส่วนของเอกสารที่เกี่ยวข้องทั้งหมด และมีเงื่อนไขว่าหากเอกสารหรือส่วนของเอกสารเป็นเอกสารที่เกี่ยวข้องทั้งหมดนั้นให้ระบุเป็น 0

ค่า f_i หมายถึงค่าความถี่ของคำค้นที่พบในเอกสารที่มีคำค้นนั้นปรากฏอยู่ทั้งหมด

ค่า qf_i หมายถึงความถี่ของคำค้นนั้น ๆ ในเซ็ตของคำค้นทั้งหมด

2.4.1.7 *การติดแท็กให้กับชนิดของคำ* (Part of Speech Tagging)

การติดแท็กให้กับชนิดของคำ หรือ POS Tagging [26, 40] เป็นกระบวนการเพื่อกำหนดขอบเขตของคำ เพื่อระบุชนิดของคำในประโยคการกำหนดขอบเขตของคำ เพื่อกำกับชนิดของคำโดยอัตโนมัติเรียกว่า “Tagging” ในขณะที่ตัวกำกับคำเรียกว่า “tag” [26]

ภาษาธรรมชาติสามารถแบ่งไวยากรณ์ของคำศัพท์ได้เป็น 8 ประเภทได้แก่ คำกริยา (verb) คำนาม (noun) คำสันธาน (conjunction) คำสรรพนาม (pronoun) คำคุณศัพท์ (adjective) คำวิเศษณ์ (adverb) คำบุพบท (preposition) และคำอุทาน (interjection)

ในการระบุ POS ในขั้นตอนแรกนั้นควรกำหนดชุดมาตรฐานของ tags ในการทำ POS tagging ในปัจจุบันมีชุดของ Tag สำหรับภาษาอังกฤษสามมาตรฐานที่นิยมใช้กันได้แก่ Penn Treebank [40], Brown Corpus Tag-set และ British national corpus ซึ่งโดย Penn treebank

กำหนดไว้ที่ 45 tag (รวมสัญลักษณ์พิเศษ) ในขณะที่ Brown Corpus และ British national corpus กำหนดไว้ที่ 192 tag และ 61 tag ตามลำดับ

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

รูปที่ 2.3 แสดงตัวอย่าง Tags ของ Penn Treebank

2.4.2 การค้นคืนสารสนเทศ (Information Retrieval : IR)

การค้นคืนสารสนเทศ [41] นั้นมีการพัฒนาตั้งแต่ช่วงปี ค.ศ. 1940 การค้นคืนสารสนเทศเป็นการค้นคืนเอกสารจากฐานข้อมูลโดยใช้คำขอหาของผู้ใช้งาน (User's queries) บ่อยครั้งที่ฐานข้อมูลของระบบค้นคืนสารสนเทศไม่จำเป็นที่จะต้องเก็บข้อมูลทั้งหมดเพื่อใช้ในการค้นหา แต่จะเก็บแค่ส่วนที่สามารถระบุที่มาของเอกสารได้ เช่น ชื่อเรื่อง, บทคัดย่อ, ผู้แต่ง เป็นต้น เพื่อที่จะลดขนาดของฐานข้อมูลและเพิ่มความเร็วในการค้นคืนสารสนเทศ

ระบบถาม-ตอบแบบอัตโนมัตินั้นมีความใกล้เคียงกับการค้นคืนสารสนเทศเป็นอย่างมาก โดยดูจากการที่ผู้ใช้งานป้อนคำขอ เพื่อที่จะหาคำตอบจากคำถามที่คิดและป้อนเข้ามา [19] ระบบการค้นคืนสารสนเทศ เป็นระบบการเรียกค้นเอกสารโดยระบบจะส่งคืนเอกสารที่เกี่ยวข้องกับความต้องการข้อมูลของผู้ใช้ แต่ไม่ได้ให้คำตอบโดยตรง ซึ่งต่างจากระบบถาม-ตอบที่มุ่งเน้นในการค้นคืนคำตอบที่กระชับและตรงตามความต้องการของผู้ใช้งานแทนที่จะเป็นเอกสาร [1]

งานวิจัยต่างๆ ในกลุ่มของการทำระบบถาม-ตอบ แบบอัตโนมัตินั้นใช้เทคนิคการประมวลผลภาษาธรรมชาติเข้ามาช่วยในการประมวลผลในหลายขั้นตอนไม่ว่าจะเป็นขั้นตอนการ

วิเคราะห์คำถาม ขั้นตอนการวิเคราะห์เอกสารเพื่อทำการค้นคืนสารสนเทศ ขั้นตอนการสกัดองค์ความรู้จากเอกสารที่ค้นคืนเพื่อที่จะสกัดเอาคำตอบ เพื่อตอบกลับไปยังผู้ใช้งาน [6]

2.4.3 การประเมินกระบวนการของระบบถาม-ตอบ (QA Evaluation)

การวัดประสิทธิภาพโดยทั่วไปในการประเมินกระบวนการถามตอบนั้นคล้ายคลึงกับการวัดประสิทธิภาพกระบวนการค้นคืนสารสนเทศ ซึ่งมีมาตรฐานในการวัดประสิทธิภาพที่แตกต่างกันส่วนใหญ่ขึ้นอยู่กับโดเมนที่ศึกษา ที่ถูกกำหนดโดย TREC ยกตัวอย่างเช่นระบบการค้นคืนสารสนเทศในโดเมนแบบปิด และการค้นคืนสารสนเทศในโดเมนแบบเปิดหรือการค้นคืนบนโลกอินเตอร์เน็ตนั่นเอง [42] ดังนั้นในงานวิจัยนี้จึงเสนอกระบวนการในการวัดประสิทธิภาพของกระบวนการถามตอบ โดยทำการวัดประสิทธิภาพด้วยกัน 2 วิธี โดยวิธีการแรกคือการดูว่าเอกสารที่คาดว่าจะมีคำตอบที่ได้จากขั้นตอนของการค้นคืนเอกสารนั้นถูกต้องหรือไม่จากจำนวนเอกสารที่เราจะให้ความสนใจทั้งหมด และการวัดประสิทธิภาพคำตอบที่ได้จากกระบวนการถามตอบ ประกอบไปด้วยการวัดประสิทธิภาพดังต่อไปนี้

2.4.3.1 Precision at K

ขั้นตอนในการวัดประสิทธิภาพของกระบวนการการถาม-ตอบนั้น จะไม่ใช้คอนฟิวชั่นเมทริกซ์ในการวัดประสิทธิภาพ แต่จะใช้การวัดประสิทธิภาพจากการหาความเกี่ยวข้องกันกับคำสำคัญหรือคำค้น กับเอกสารหรือส่วนของเอกสารใด ๆ ที่ได้คืนจากกระบวนการการถาม-ตอบ [43, 44] ซึ่งหลายครั้งที่กระบวนการเหล่านี้ใช้การค้นคืนเอกสารจากแหล่งที่มาที่ใหญ่ขึ้นบนเว็บไซต์หรือเวิร์ดไวด์เว็บ ทำให้การคำนวณหาค่าความละลึกไม่สามารถทำได้หรือการที่ไม่สามารถค้นคืนเอกสารที่มีความเกี่ยวข้องขึ้นมาได้ทั้งหมด [34] ดังนั้นเราจึงมีวิธีการในการวัดประสิทธิภาพด้วยการสนใจผลลัพธ์ของการค้นคืนที่ K อันดับแรก เรียกว่า Precision at K (P@K) สามารถคำนวณได้จากสมการดังต่อไปนี้

$$P@K = \frac{Correct\ Doc}{|K\ Relevant\ Doc|} \tag{14}$$

2.4.3.2 Mean Reciprocal Rank

ค่าเฉลี่ยการจัดอันดับแบบอิงกลุ่ม (Mean Reciprocal Rank : MRR) [38, 39] เป็นการคำนวณทางสถิติเพื่อหาค่าการตอบสนองที่เป็นไปได้จากผลลัพธ์ที่ได้หลังจากการเปรียบเทียบกับคำค้น เพื่อวัดประสิทธิภาพของการจัดอันดับด้วยการประยุกต์ใช้หลักการทางสถิติที่เรียกว่าส่วนกลับของสมการการคูณ (Multiplicative Inverse) หรือเรียกอีกอย่างหนึ่งว่าส่วนกลับ (Reciprocal) ของการจัดอันดับ โดยสามารถหาได้จากสมการดังต่อไปนี้

$$MRR = \frac{1}{|K\ Relevant\ Doc|} \times \sum_{i=1}^{K} \frac{1}{rank_i} \quad (15)$$

โดยที่ค่า K คือค่าที่เป็นจำนวนเอกสารที่เป็นผลลัพธ์กระบวนการถามตอบ จากนั้นหาค่าเฉลี่ยของส่วนกลับในแต่ละเอกสาร

จากกระเทคนิคในการวัดประสิทธิภาพที่กล่าวมาข้างต้นนั้น กล่าวอีกทางหนึ่งก็คือกระบวนการดังกล่าวไม่สามารถที่จะค้นคืนเอกสารที่มีความเกี่ยวเนื่องกับคำค้นได้ทั้งหมดหรือไม่สามารถวัดได้ ดังนั้นการวัดประสิทธิภาพจึงจำเป็นที่จะต้องใช้การวัดประสิทธิภาพกับเอกสารที่สืบค้นได้ในปัจุบันเท่านั้น [44]

2.4.3.3 Bilingual Evaluation Understudy

Bilingual Evaluation Understudy (BLEU-1) [45] เป็นการวัดประสิทธิภาพคำตอบที่ได้จากกระบวนการถาม-ตอบ จากการประยุกต์ใช้เทคนิคการวัดประสิทธิภาพของการแปลภาษา ด้วยค่าเฉลี่ยเรขาคณิต (Geometric Mean) โดยนำเอาข้อความมาเปรียบเทียบกัน โดยมีข้อความเฉลยเป็นตัวอ้างอิงความถูกต้องของข้อความ ในที่นี้เราจะมีจำนวนข้อความที่ใช้สำหรับการเปรียบเทียบอยู่ทั้งสิ้น 10 ข้อความต่อการเปรียบเทียบคำตอบสองคำตอบ และยังมีการประยุกต์ใช้สมการในการวัดค่าความแม่นยำให้กับค่าแกรม (Modified n-gram Precision) โดยสามารถคำนวณได้จากสมการดังต่อไปนี้

$$\log(BLUE - N) = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n \cdot \log(p_n) \quad (16)$$

เมื่อ c คือ ความยาวของคำตอบที่จะทำการประเมิน และ r เป็นค่าความยาวของเอกสารที่ทำการเปรียบเทียบโดยที่ค่า p_n สามารถคำนวณได้จากสมการดังต่อไปนี้

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count(n-gram)} \quad (17)$$

โดยที่กำหนดค่า N เป็นจำนวนแกรมที่เราต้องการ และค่า C ในสมการหมายถึงคอเล็คชั่นของเทอมที่เราสนใจในคำตอบนั้น ๆ

2.4.3.4 Recall-Oriented Understudy for Gisting Evaluation

Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) [46] เป็นการใช้พื้นฐานจากการประเมินด้วยค่าเอฟซึ่งคำนวณได้จากค่าความละลึก (Recall) และค่าความแม่นยำ (Precision) และทำการหาค่าเฉลี่ยของทั้งสองออกมา แต่ใน ROUGE จะเป็นการหาค่าเฉลี่ยจากประโยคทั้งสองประโยค และอาศัยการคำนวณโดยอาศัยการเปรียบเทียบแบบที่มีลำดับที่ยาวที่สุด (Longest Common Subsequence : LCS) ดังนั้น ROUGE จึงมีการวิเคราะห์ในหลายระดับขึ้นอยู่กับว่าเราจะสนใจในการวัดประสิทธิภาพที่ระดับใด ซึ่งในงานวิจัยนี้จะให้ความสนใจอยู่ที่ระดับโครงสร้าง จะถูกเรียกว่า ROUGE-L สามารถคำนวณได้จากสมการดังต่อไปนี้

$$R_{lcs} = \frac{LCS(X,Y)}{m} \tag{18}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \tag{19}$$

$$ROUGE - L = \frac{2 \cdot R_{lcs} \cdot P_{lcs}}{R_{lcs} + P_{lcs}} \tag{20}$$

โดยที่ค่า *X* หมายถึงคำตอบที่ได้จากระบบถามตอบ ค่า *Y* หมายถึงประโยคที่ต้องการทำการเปรียบเทียบ ค่า *m* คือจำนวนของเทอมที่อยู่ในคำตอบ *X* และค่า *n* คือค่าที่เป็นจำนวนของเทอมที่อยู่ในประโยคที่เราต้องการทำการเปรียบเทียบ

2.5 งานวิจัยที่เกี่ยวข้อง

ระบบถาม-ตอบเริ่มมีการพัฒนาตั้งแต่ช่วงปี ค.ศ. 1960 เช่นงานของ Bert Green และคณะ [47] ที่ทำระบบถาม-ตอบเกี่ยวกับการแข่งขันกีฬาเบสบอล (Baseball) โดยใช้หลักการทางภาษาในการประมวลผลเป็นหลักเทียบกับพจนานุกรมคำถามที่อยู่ในรูปแบบโครงสร้างของวลี เรียกว่า รายการข้อมูลจำเพาะ (Specification List) ที่เตรียมไว้อยู่ก่อนแล้ว และคำตอบที่ได้มีเพียงคำตอบที่เป็น ใช่ หรือ ไม่ใช่ เท่านั้น

ข้อจำกัดรวมถึงข้อดีของงานนี้ที่ผู้เขียนได้ทำการสรุปไว้พร้อมทั้งบอกว่า ข้อดีของข้อมูลจำเพาะที่เตรียมไว้นั้นเพื่อที่จะเป็นตัวระบุกลุ่มหรือประเภทของคำถามที่ระบบถาม-ตอบนี้สามารถตอบได้ แต่อย่างไรก็ตามงานวิจัยนี้ยังมีข้อจำกัดอยู่มากทั้งในเรื่องของประเภทของคำถามที่ไม่สามารถถามนอกเหนือจากโครงสร้างวลี ที่กำหนดไว้ได้และคำตอบที่ได้ยังคงจำกัดแค่ ใช่ หรือ ไม่ใช่เท่านั้น

ต่อมาในปี 2001 Agichtein และคณะ [12] ได้นำเสนอกระบวนการการถาม-ตอบแบบนอนแฟคทอยด์ โดยใช้ชื่อว่า “Tritus” ซึ่งเป็นกระบวนการ การถาม-ตอบ ที่เน้นไปยังการวิเคราะห์

คำถาม ด้วยการแปลงประโยคคำถามที่อยู่ในรูปแบบที่เป็นภาษาธรรมชาติ ให้เปลี่ยนเป็นเทอม (Term) หรือวลี (Phrase) ที่คาดว่าจะปรากฏในเอกสารที่มีคำตอบ เพื่อให้สอดคล้องกับลักษณะของคำถามที่มีการถามเข้ามา เนื่องจากผู้วิจัยตั้งสมมุติฐานไว้ว่าเอกสารที่คาดว่าจะมีคำตอบ หรือบางครั้งเอกสารที่อาจจะมีคำตอบที่ดีกว่า อาจจะไม่ถูกดึงขึ้นไปจากกระบวนการปกติเพราะมีคำสำคัญปรากฏไม่เพียงพอ ดังนั้นผู้วิจัยจึงนำเสนอกระบวนการเรียนรู้คำถาม และคำตอบจากชุดข้อมูลเพื่อที่จะหารูปแบบการตอบจากคำถามที่เข้ามา เพื่อแปลงเป็นคำค้นผสมกับคำสำคัญที่อยู่ในคำถาม โดยเทคนิคที่ใช้ในการเรียนรู้รูปแบบดังกล่าวนั้น ผู้วิจัยใช้เทคนิค n-grams ในการหารูปแบบที่เป็นไปได้ และใช้กระบวนการการค้นหาเอกสารที่คาดว่าจะมีคำตอบด้วยเครื่องมือการค้นหา (Search Engine) สามตัวเพื่อเปรียบเทียบคำตอบที่ได้รับคือ Google, AltaVista และ AskJeeves หลังจากที่ทำการขยายขอบเขตของคำค้นเป็นที่เรียบร้อยแล้ว ผลลัพธ์ที่ได้นั้นผู้วิจัยใช้การหาค่าความถูกต้องเฉลี่ยที่ได้รับจากจำนวนคำตอบที่ได้รับสูงสุด (Precision at K) สามารถเขียนแทนได้ด้วย P@K โดยที่ค่า *K* หมายถึงจำนวนผลลัพธ์ที่ได้สูงสุด K อันดับแรก ปรากฏว่าตัวที่ให้ประสิทธิภาพดีที่สุดนั้นคือกระบวนการที่ใช้ Tritus ร่วมกับ ระบบการค้นคืนสารสนเทศของ Google

ในปี 2007 ได้มีงานวิจัยของ Fukumoto [4] ได้นำเสนอเกี่ยวกับวิธีการวัดประสิทธิภาพแบบอัตโนมัติของระบบถาม-ตอบ แบบนอนแฟคทอยด์ ด้วยการประยุกต์ใช้เทคนิคของการวัดประสิทธิภาพของการย่อความด้วยเทคนิคที่เรียกว่า (Basic Element Method : BE) ประกอบกับระบบค้นคืนคำตอบที่เรียกว่า RitQ และทำการประมวลผลคำตอบเหล่านั้นเพื่อสกัดหาคำตอบกลับไปยังผู้ใช้งาน โดยมีขอบเขตของงานอยู่ที่การประมวลผลคำถามประเภทนอนแฟคทอยด์ เฉพาะคำถามประเภท Why, How และ Definition ด้วยภาษาญี่ปุ่นจากนั้นทำการวัดประสิทธิภาพของคำตอบที่ได้ภายหลังจากการทำ BE เปรียบเทียบกับเฉลยด้วยการหาค่าเอฟ (F-measurement)

ภายหลังจากการค้นคืนคำตอบด้วย RitQ นั้นพบว่าได้รับคำตอบทั้งสิ้น 167 คำตอบจากคำถามทั้งหมด 100 คำถาม จากนั้นจะให้บุคคลทั่วไปเป็นผู้ประเมินและให้คะแนนคำตอบเหล่านั้นเป็นเกรดทั้งสิ้น 4 เกรดเรียงจากคำตอบที่ดีที่สุดไปยังคำตอบที่ดีน้อยที่สุด แบ่งออกเป็น A, B, C และ D หลังจากนั้นจึงให้เทคนิค BE ในการประมวลผลคำตอบเหล่านี้จากนั้นนำไปวัดประสิทธิภาพด้วยค่าเอฟ ได้ผลลัพธ์สำหรับคำตอบระดับ A มากที่สุดอยู่ที่ 0.69 และคำตอบระดับ D อยู่ที่ 0.94

ในปี 2010 ได้มีนักวิจัยจากบริษัทใหญ่หลายที่อันได้แก่ Surdeanu และคณะซึ่งเป็นการรวมตัวกันของนักวิจัยจาก Stanford, Google และ Yahoo! [5] ได้ร่วมกัน เขียนงานวิจัยเกี่ยวกับวิธีการการคัดเลือกคำตอบที่ได้จากการค้นคืนเอกสารเพื่อทำการประมวลผลและตอบกลับไปยังผู้ใช้งาน โดยในงานวิจัยนี้ผู้วิจัยได้ทำการมุ่งเน้นไปที่กระบวนการในการค้นหาคำตอบ จากข้อมูลบนเครือข่ายสังคมออนไลน์ของ Yahoo Answer โดยสนใจเฉพาะคำถามประเภท How เท่านั้นโดยมี

คำถามที่ใช้ทั้งสิ้น 364,419 คำถามจากนั้นแบ่งข้อมูลออกเป็น 3 กลุ่ม ได้แก่ 60% สำหรับชุดสอน 20% สำหรับการพัฒนา 20% สำหรับการทดสอบประสิทธิภาพโดยผลลัพธ์ที่ได้จะแสดงในรูปแบบของการหาผลลัพธ์ที่ได้ว่ามีประสิทธิภาพดีขึ้นเป็นเท่าใดจากกระบวนการที่นำเสนอ ผลลัพธ์ที่ได้จากการวัดค่าด้วย MRR มีค่ามากที่สุดอยู่ที่ 64.88 น้อยที่สุดอยู่ที่ 56.09 และ P@1 มากที่สุดอยู่ที่ 50.91% ต่ำที่สุดอยู่ที่ 41.12%

เนื่องจากภายหลังจากช่วงปี 2010 พบว่าการเรียนรู้เชิงลึก (Deep Learning) เข้ามามีบทบาทสำคัญในงานวิจัยหลากหลายสาขา พบว่า ปี 2011 นักวิจัย Merugu [48] และคณะได้ทำการนำเสนอแนวคิดที่จะวิเคราะห์คำถามและคำตอบด้วยการใช้โครงข่ายประสาทเทียมแบบรีเคอร์ (Recursive Nueral Network : ReNN) ในการเรียนรู้เชิงลึกเกี่ยวกับการตีความเชิงความหมายของลักษณะคำถามและคำตอบ โดยใช้หลักการของการตรวจสอบการเปลี่ยนแปลงความหมาย (Paraphrase Detection) จากนั้นคำนวณหาความคล้ายคลึงของคำถามและคำตอบที่ค้นหา โดยใช้คุณลักษณะของคำ, วลี และลักษณะของวลีที่อยู่ในรูปของกลุ่มคำ จัดเก็บในรูปแบบของเวกเตอร์สเปซโมเดล เพื่อคำนวณหาค่าความคล้ายคลึงด้วยการวัดระยะทางแบบ ยูคลีเดียน (Euclidean Distance) จากนั้นวัดประสิทธิภาพด้วยค่าความถูกต้อง และค่าเฉลี่ยการจัดอันดับแบบอิงกลุ่ม (Mean Reciprocal Rank : MRR) โดยมีผลเฉลยด้วยการเฉลยด้วยคน ให้ค่าอยู่ที่ 0.41 และ 0.56 ตามลำดับ

แต่ทั้งนี้งานวิจัยส่วนใหญ่ในการพัฒนากระบวนการการถาม-ตอบ ยังคงไม่ได้รับผลกระทบจากการเรียนรู้เชิงลึกมากเท่ากับโดเมนของงานวิจัยอื่น ซึ่งในปี 2015 มีงานวิจัยกระบวนการการถาม-ตอบ ในโดเมนทางการแพทย์นำเสนอโดย Hristovski และคณะ [49] นำเสนอกระบวนการการถามตอบด้วยการใช้ความสัมพันธ์เชิงความหมาย ผู้วิจัยได้กล่าวไว้ว่าการหาคำตอบและการสกัดองค์ความรู้จากวรรณกรรมทางการแพทย์นั้นยังคงเป็นไปได้ยาก โดยทั่วไปแล้วการค้นหาวรรณกรรมดังกล่าวยังคงกระทำโดยการใช้การค้นคืนเอกสารทั่วไป ซึ่งทำให้ได้เอกสารที่มีความเกี่ยวข้องขึ้นมา แต่ยังคงต้องใช้ผู้เชี่ยวชาญในการสกัดองค์ความรู้ออกมา พร้อมทั้งยังใช้เวลานานในการสกัดองค์ความรู้ดังกล่าว ซึ่งกระบวนการเหล่านี้สามารถลดระยะเวลา และช่วยผู้เชี่ยวชาญในการค้นหาองค์ความรู้จากวรรณกรรมทางการแพทย์ได้รวดเร็วยิ่งขึ้น โดยการใช้กระบวนการการถาม-ตอบ โดยกระบวนการถาม-ตอบดังกล่าวมีเป้าหมายหลักคือการหาคำตอบที่คาดว่าจะถูกต้องที่สุดกลับไปยังผู้ใช้งาน โดยกระบวนการที่นำเสนอนี้เป็นการเรียนรู้จากวรรณกรรมทางการแพทย์ โดยอาศัยการเรียนรู้ความสัมพันธ์เชิงความหมายจากประโยคต่าง ๆ ที่ปรากฏอยู่ในวรรณกรรมทางการแพทย์ โดยใช้เครื่องมือที่เรียกว่า SemRep ในการหาองค์ประกอบดังกล่าว และอาศัยกระบวนการการถามตอบที่เรียกว่า SemBT ในการค้นหาคำตอบที่คาดว่าจะเป็นคำตอบที่ถูกต้อง โดยใช้ผู้เชี่ยวชาญในการสร้าง

ผลเฉลยจำนวนทั้งสิ้น 80 คน พบว่าด้วยการใช้ความสัมพธ์เชิงความหมายที่สกัดได้จำนวนทั้งสิ้น 2,675 ความสัมพันธ์ และใช้ความสัมพัธ์ดังกล่าวในการถาม-ตอบ จำนวนทั้งสิ้น 12,083 ครั้ง พบว่าคำตอบที่ได้ถูกต้องมีจำนวน 8,228 ครั้ง คิดเป็น 68%

ล่าสุดในปี 2016 Liu Yang และคณะได้ทำการวิจัยระบบถาม-ตอบ สำหรับการหาคำตอบด้วยการค้นหาคำตอบบนเว็บแบบนอนแฟคทอยด์ สำหรับการทดลอง Liu Yang และคณะ ได้ทำการใช้การประมวลผลระดับความหมาย (Semantic) และบริบท (Context) ในการสกัดคุณลักษณะเพื่อทำการหาคำตอบตอบกลับไปยังผู้ใช้งานต่อไป และทำการประเมินประสิทธิภาพด้วยการหาค่าการลดหลั่นสะสมแบบปรับปรุง (Normalized Discounted Cumulative Gain : NDCG), P@10 และ MRR อยู่ที่ 0.1864, 0.2024 และ 0.4512 ตามลำดับ

จากงานวิจัยที่ได้กล่าวมาข้างต้นนั้น จะพบว่าการถามตอบเป็นงานวิจัยสาขาหนึ่ง ที่มีการนำเทคนิคการประมวลผลภาษาธรรมชาติเข้ามาช่วยในการวิเคราะห์ในส่วนต่าง ๆ เพื่อที่จะสามารถให้กระบวนการดังกล่าว สามารถค้นหาคำตอบกลับไปยังผู้ใช้งานได้อย่างถูกต้อง โดยที่ผู้ใช้งานไม่จำเป็นที่จะต้องทำการค้นหาข้อมูลด้วยตนเอง โดยงานวิจัยที่ได้รับความนิยมในช่วงปี 2010 เป็นต้นมานั้นเป็นการให้ความสนใจการถามตอบที่มีความซับซ้อนมากยิ่งขึ้นเพื่อให้คอมพิวเตอร์สามารถตอบคำถามที่มีความยากมากยิ่งขึ้นได้ ไม่ว่าจะเป็นทางด้านการแพทย์ หรือการนำไปใช้ในด้านอื่น ๆ อีกทั้งปัญหาในกระบวนการถามตอบ ยังคงเป็นปัญหาที่ยากต่อการแก้ไข จึงทำให้มีความท้าทายในการแก้ปัญหาดังกล่าว และด้วยทิศทางในการดำเนินงานวิจัยในช่วงปี 2010 มานี้เป็นเหตุให้งานวิจัยที่นำเสนอนี้ จะเป็นการนำเสนอกระบวนการถามตอบ และคำตอบที่ได้จากกระบวนการจะเป็นคำตอบที่อยู่ในลักษณะของคำอธิบาย โดยจะทำการค้นคืนส่วนของเอกสารจากเอกสารที่เกี่ยวข้อง ขึ้นมาเป็นคำตอบสามารถแสดงขั้นตอนและกระบวนการที่นำเสนอได้ในบทต่อไป

พหุน ปณ ฑิโต ชีเว

บทที่ 3
ขั้นตอนการดำเนินงาน

3.1 ชุดข้อมูล (Data Set)

ชุดข้อมูลมาตรฐานที่ใช้ในงานวิจัยนี้คือ Microsoft Machine Reading Comprehension Dataset (MS MARCO) เวอร์ชัน 2.1 ซึ่งเป็นชุดข้อมูลมาตรฐานของ Microsoft [50, 51] สามารถดาวน์โหลดได้จาก msmarco.org โดยมีชุดข้อมูลที่ประกอบไปด้วย คำถาม, หมายเลขคำถาม, ประเภทของคำถาม, คำตอบ และเอกสารที่เป็นที่มาของคำตอบจำนวนทั้งสิ้น 1,010,916 ชุด โดยชุดข้อมูลมาตรฐานนี้แบ่งออกเป็น 3 กลุ่มได้แก่กลุ่มของชุดคำถามที่ใช้ในการเรียนรู้ (Training Set) กลุ่มของชุดคำถามที่ใช้ในการทดสอบประสิทธิภาพ (Evaluation Set) และชุดคำถามที่ใช้ในขั้นตอนการพัฒนากระบวนการ (Development Set) ชุดข้อมูลมาตรฐานนี้บริษัท Microsoft ได้รวบรวมจากการใช้งานระบบค้นคืนสารสนเทศที่ชื่อว่า Bing ในการรวบรวมคำถาม โดยคำตอบที่เป็นผลเฉลยนั้นมาจากผู้ใช้งานจริงที่ได้เข้ามาใช้งานระบบค้นคืนสารสนเทศนี้ และเก็บรวบรวมออกมาเป็นชุดข้อมูลคำถามดังที่อธิบายไปข้างต้น

ตารางที่ 3.1 จำนวนข้อมูลในชุดข้อมูลมาตรฐาน MS MARCO เวอร์ชัน 2.1

Dataset	จำนวนทั้งหมด	แบ่งออกเป็น				
		How		Why		Other
		Description	Other	**Description**	Other	
Training	808,726	**30,205**	136,736	**83,456**	915	557,414
Evaluation	101,102	**4,932**	18,531	**3,379**	254	74,006
Development	101,088	**3,411**	12,823	**1,656**	124	83,074
รวมทั้งสิ้น	1,010,916	**38,548**	168,090	**88,491**	1,293	714,494

ในชุดคำถามของ MS MARCO เวอร์ชัน 2.1 นั้นแบ่งออกเป็นชุดคำถามที่ใช้ในการเรียนรู้จำนวน 808,726 ชุด ชุดคำถามที่ใช้ในการทดสอบประสิทธิภาพจำนวน 101,102 ชุด และชุดคำถามที่ใช้ในขั้นตอนการการพัฒนากระบวนการจำนวน 101,088 ชุด ความแตกต่างของชุดข้อมูลทั้ง 3 กลุ่มนั้นได้แก่ชุดข้อมูลในกลุ่มที่ใช้สำหรับเรียนรู้นั้นจะมีคำถามพร้อมทั้งคำตอบและที่มาของคำตอบแต่ละคำตอบเพื่อใช้เป็นผลเฉลย โดยในแต่ละชุดคำถามนั้นมีการแบ่งประเภทของคำถามออกเป็น 5 ประเภทคำถามได้แก่ สถานที่ (Location), จำนวน (Numeric), บุคคล (Person), คำเฉพาะ (Entity)

และ ประเภทพรรณา (Description) โดยในงานวิจัยนี้ผู้วิจัยสนใจเฉพาะชุดคำถามที่เป็นคำถามประเภทพรรณา หรืออยู่ในลักษณะของการอธิบายซึ่งจัดอยู่ในกลุ่มของคำถามที่เป็นนอนแฟคทอยด์

ดังนั้นในงานวิจัยนี้จะใช้ชุดข้อมูลจากชุดข้อมูลมาตรฐาน MS MARCO เวอร์ชัน 2.1 จำนวนทั้งสิ้น 127,039 ชุด แบ่งออกเป็นคำถามประเภท How จำนวน 34,548 ชุดคำถาม และ Why จำนวน 88,491 ชุดคำถาม

3.1.1 รูปแบบการจัดเก็บชุดข้อมูล (Dataset Format)

ชุดข้อมูลมาตรฐาน MS MARCO เวอร์ชัน 2.1 นั้นจัดเก็บอยู่ในรูปแบบของ Jason Lines (JSONL) ซึ่งเป็นการจัดเก็บข้อมูลเชิงโครงสร้างที่ทันสมัย และมีความเหมาะสมกับข้อมูลที่มีขนาดใหญ่ ทำให้ง่ายต่อการค้นหาและเรียกใช้งานข้อมูล

ลักษณะของโครงสร้างในการจัดเก็บข้อมูลในหนึ่งชุดนั้นประกอบไปด้วยข้อมูลจำนวน 6 ฟิลด์ ได้แก่ ไอดีของคำถามจะเรียกว่า Query ID, ประเภทของคำถาม หรือ Query Type, คำถามในที่นี้จะเรียกว่า Query, ชุดคำตอบพร้อม Passage, คำตอบหรือ Answers และคำตอบที่ดีที่สุดที่ได้รับการยืนยันหรือ Well Formed Answer โดยแต่ละฟิลด์นั้นสามารถอธิบายได้ดังตารางที่ 3.2

ตารางที่ 3.2 โครงสร้างของชุดข้อมูลมาตรฐาน MS MARCO เวอร์ชัน 2.1

ฟิลด์ในการจัดเก็บข้อมูล	คำอธิบาย
Query ID	เป็นการจัดเก็บรหัสของชุดคำถามนั้น ๆ
Query Type	ประเภทของคำถาม ประกอบไปด้วย สถานที่, จำนวน , บุคคล, คำเฉพาะ และประเภทพรรณา
Passages	เป็นเซ็ตของคำตอบจำนวนทั้งสิ้น 10 คำตอบที่เป็นคำตอบของคำถามในชุดคำถามนี้ พร้อมทั้ง URLs ที่เป็นแหล่งที่มาของคำตอบดังกล่าว ภายในประกอบไปด้วย ฟิลด์ย่อย ๆ จำนวน 3 ฟิลด์ ได้แก่ is_selected, answer และ URL โดยที่ is_selected มีค่าเป็น 0 และ 1 ถ้าหากมีค่าเป็น 1 นั่นหมายถึงคำตอบดังกล่าวได้รับการพิจารณาจากผู้ใช้งานให้เป็นคำตอบที่ถูกเลือกจากผู้ใช้งาน
Query	เป็นคำถามของชุดคำถามดังกล่าว
Answers	คำตอบที่ถูกเลือกมาเป็นคำตอบจากชุดคำตอบทั้งหมด มาเป็นคำตอบของชุดคำถามดังกล่าว
Well Formed Answers	ในกรณีที่คำตอบที่ได้ มีมากกว่า 1 คำตอบ ฟิลด์นี้จะเป็นฟิลด์ที่ใช้สำหรับจัดเก็บคำตอบที่ดีที่สุด

ในการจัดเก็บข้อมูลมาตรฐานนี้สามารถแสดงรูปแบบการจัดเก็บข้อมูลแบบ JSON ได้ดังแสดงในรูปที่ 3.1 ต่อไปนี้

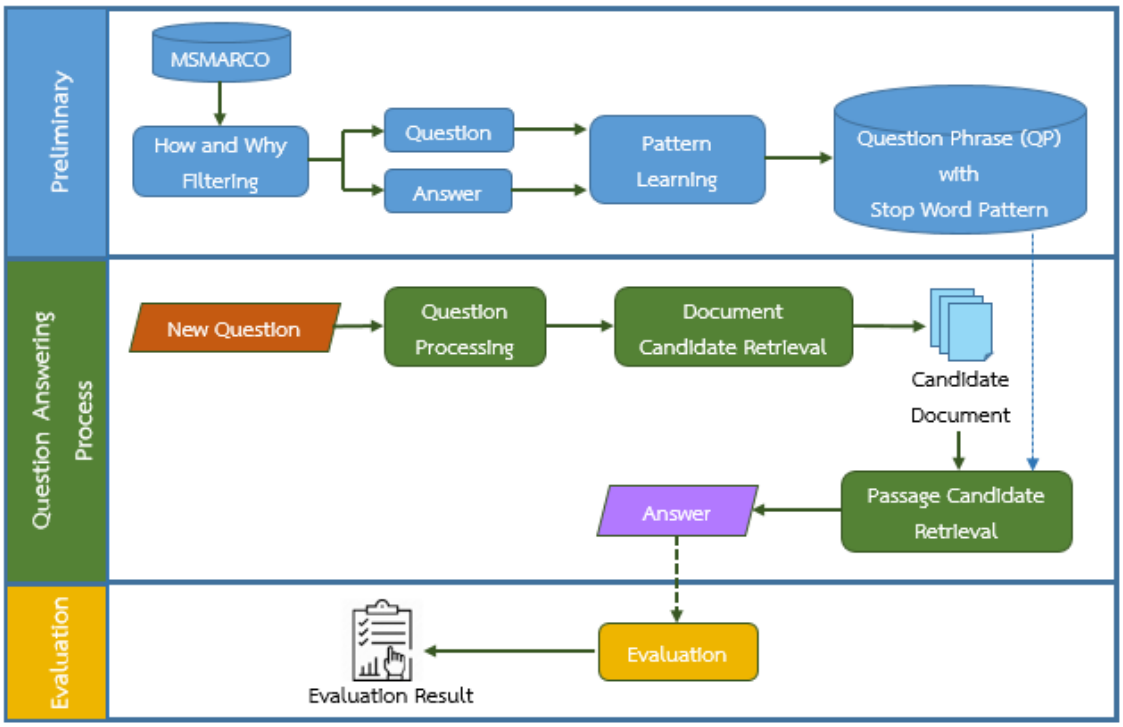

```
{
        "answers":["A corporation is a company or group of people authorized to act as a s
        "passages":[
                {
                        "is_selected":0,
                        "url":"http:\/\/www.wisegeek.com\/what-is-a-corporation.htm",
                        "passage_text":"A company is incorporated in a specific nation, of
                ...
                }],
        "query":". what is a corporation?",
        "query_id":1102432,
        "query_type":"DESCRIPTION",
        "wellFormedAnswers":"[]"
}
```


รูปที่ 3.1 โครงสร้างการจัดเก็บข้อมูลของ MS MARCO เวอร์ชัน 2.1

3.2 กรอบการดำเนินงาน (Research Methodology)

ในงานวิจัยนี้มีแนวคิดในการนำเสนอกระบวนการเพื่อสกัดความรู้บนพื้นฐานของการถาม-ตอบ เพื่อแสดงความรู้ในรูปแบบคำตอบที่สกัดได้ โดยกรอบแนวคิดของกระบวนการถาม-ตอบที่จะนำเสนอ สามารถแสดงได้ดังรูปที่ 3.2

กรอบการดำเนินงานที่นำเสนอประกอบด้วยส่วนการทำงาน 3 ส่วนหลัก ๆ ได้แก่ ส่วนที่หนึ่งการทำการประมวลผลเบื้องต้น (Preliminary Processing) เพื่อทำการเรียนรู้รูปแบบการตอบคำถามของคำตอบที่มี โดยเน้นไปที่การหากลุ่มของคำหยุดเพื่อที่จะนำไปใช้ประกอบการวิเคราะห์การหาคำตอบจากเอกสารต่อไป ขั้นตอนถัดมาส่วนที่สองเป็นส่วนของกระบวนการการถามตอบ ซึ่งประกอบไปด้วยกระบวนการย่อยอีก 3 กระบวนการ โดยเริ่มจากกระบวนการการประมวลผลคำถาม การค้นคืนเอกสารที่คาดว่าน่าจะมีคำตอบ และกระบวนการสุดท้ายคือการค้นคืนบางส่วนของเอกสารที่คาดว่าจะมีคำตอบจากเอกสารที่ค้นคืนได้ จากนั้นทำการตอบกลับไปยังคำถามที่รับเข้ามา

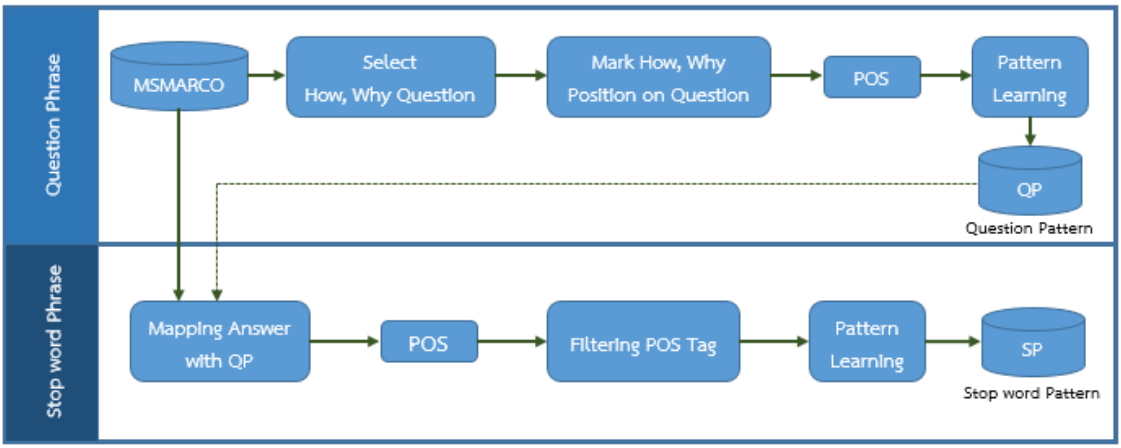


รูปที่ 3.2 กรอบการดำเนินงานของกระบวนการที่นำเสนอ

3.3 กระบวนการการประมวลผลขั้นต้น (Preliminary Processing)

ในงานวิจัยนี้มีแนวคิดในการนำเสนอกระบวนการเพื่อสกัดความรู้บนพื้นฐานของการถาม-ตอบ เพื่อแสดงความรู้ในรูปแบบคำตอบที่สกัดได้ กรอบแนวคิดของกระบวนการถาม-ตอบที่จะนำเสนอ โดยขั้นตอนแรกของกระบวนการการถาม-ตอบนั้น ประกอบไปด้วยการกรองข้อมูลให้เหลือเพียงประเภทของคำถามที่เป็น “*Why*” และ *How*” ตามที่ได้กล่าวไว้ข้างต้นจากนั้นทำการเรียนรู้รูปแบบของคำถามและคำตอบเพื่อที่จะสามารถนำไปใช้ในการหาคำตอบที่ได้ สามารถแสดงได้ดังรูปที่ 3.3

จากรูปที่ 3.3 จะเห็นว่าจากชุดข้อมูล MS MARCO เวอร์ชัน 2.1 นั้นในขั้นตอนแรกจะเป็นการทำความสะอาดข้อมูล (Data Cleaning) เบื้องต้นเพื่อทำการกรองเฉพาะคำถามที่มีประเภทของคำถามที่เป็น “*Why*” และ “*How*” จากนั้นจะพบว่ามีขั้นตอนหลักอยู่ 2 ขั้นตอน ขั้นตอนแรกได้แก่ ขั้นตอนการเรียนรู้วลีคำถาม (Question Phrase : QP) โดยการนำคำถามเหล่านี้จะถูกนำมาวิเคราะห์คู่ของคำต่าง ๆ ภายในประโยคคำถาม เพื่อที่จะหาส่วนที่เป็นลักษณะวลีคำถาม ว่ามีวลีคำถามใดบ้าง จากนั้นขั้นตอนที่สองจะทำการนำวลีคำถามที่ได้ ไปเรียนรู้คำที่จะเป็นองค์ประกอบของคำตอบในแต่ละรูปแบบของคำถาม ด้วยการใช้ประโยชน์จากรูปแบบของการเกิดคำหยุดในคำตอบ ในงานวิจัยนี้จะเรียกว่า “*รูปแบบคำหยุด* (*Stop word Pattern*: *SP*)”



รูปที่ 3.3 การสร้างรูปแบบของกลุ่มคำหยุด

3.3.1 การเรียนรู้รูปแบบของคำถาม (Learning of Question Patterns)

ตามที่ได้กล่าวไว้ข้างต้นว่าขั้นตอนนี้จะทำการหาวลีของคำถามเพื่อใช้ในการเรียนรู้คำที่จะเป็นองค์ประกอบของคำตอบในแต่ละรูปแบบของคำถาม ที่มีวลีของคำถามตรงกัน โดยเลือกประโยคคำถามที่ขึ้นต้นด้วย “*How*” และ “*Why*” มาใช้ในการเรียนรู้รูปแบบของคำถาม ยกตัวอย่างเช่น ประโยคคำถามดังต่อไปนี้

ขั้นตอนที่ 1 : ขั้นตอนการระบุตำแหน่งของ “How” และ “Why”

ในขั้นตอนนี้จะเป็นการระบุตำแหน่งของ “*How*” และ “*Why*” เพื่อแยกส่วนที่เป็นคำขึ้นต้นคำถามดังกล่าว และส่วนที่เหลือของประโยคออกจากกัน จากนั้นจะทำการเรียนรู้รูปแบบของประโยคคำถามว่าประโยคคำถามที่ขึ้นต้นด้วย “*How*” หรือ “*Why*” นั้นมีส่วนที่เหลือของประโยคเป็นรูปแบบใดได้บ้าง ยกตัวอย่างเช่น

Q1 : How to have an intelligent baby?

Q2 : How do you like your eggs?

Q3 : Why does a hard drive make a clicking sound?

Q4 : Why is a Desktop background called a Wallpaper?

ในขั้นตอนแรกเมื่อได้ประโยคคำถามที่ต้องการแล้วนั้นให้ทำการ ระบุตำแหน่งของ “*How*” และ “*Why*” ในประโยคคำถามและทำการแยกส่วนที่เหลือของประโยคออกจากกัน สามารถแสดงได้ดังต่อไปนี้

Q1 : <How>, <to have an intelligent baby?>

Q2 : <How>, <do you like your eggs?>

Q3 : <Why>, <does a hard drive make a clicking sound?>

Q4 : <Why>, <is a Desktop background called a Wallpaper?>

เมื่อทำการระบุตำแหน่งและทำการแยกส่วนที่เหลือในประโยคออกจากกันแล้วนั้น ขั้นตอนถัดไปจะเป็นการระบุหน้าที่ของคำให้กับส่วนที่เหลือในประโยค

ขั้นตอนที่ 2 : ขั้นตอนการระบุหน้าที่ของคำให้ส่วนที่เหลือในประโยค

หลังจากที่ได้ผลลัพธ์จากขั้นตอนที่หนึ่งเป็นที่แล้ว จากนั้นจะเป็นการระบุหน้าที่ของคำด้วยการติดแทกซ์ (Tags) ให้กับส่วนที่เหลือของประโยคคำถาม โดยเทคนิค POS tagging ของ Stanford ที่ใช้ชุดข้อมูลสำหรับระบุหน้าที่ของคำ (Tag set) ที่เรียกว่า Penn Tree Bank ในการระบุหน้าที่ของคำให้กับส่วนที่เหลือในประโยค สามารถแสดงผลลัพธ์ได้ดังต่อไปนี้

Q1 : <How>, <TO VB DT JJ NN .>

Q2 : <How>, <VBP PRP VB PRP$ NNS .>

Q3 : <Why>, <VBZ DT JJ NN VB DT VBG NN .>

Q4 : <Why>, <VBZ DT JJ NN VBN DT NN .>

จากผลลัพธ์ข้างต้น จะเห็นว่าในแต่ละประโยคคำถาม จะถูกปรับเปลี่ยนให้อยู่ในรูปแบบของลำดับของชนิดคำในประโยคคำถาม (Sequence of tags: SOP)

ขั้นตอนที่ 3 : ขั้นตอนการเรียนรู้รูปแบบของคำถามด้วยแบบจำลองมาร์คอฟ

ในขั้นตอนการเรียนรู้รูปแบบของคำถามนั้นจะใช้หลักการของความน่าจะเป็นที่เกิดขึ้นเป็นลำดับโดยใช้เทคนิคที่เรียกว่าแบบจำลองมาร์คอฟ (Markov Model) เพื่อให้ทราบรูปแบบที่เป็นไปได้ทั้งหมดของประโยคคำถาม โดยอาศัยการประมาณค่าความน่าจะเป็นจากลำดับการเกิดขึ้นของแทกซ์ที่ระบุชนิดของคำในแบบไบแกรม (Bigrams) และกำหนดให้ตำแหน่งของ “*How*” และ “*Why*” ที่ระบุไว้ในขั้นตอนที่หนึ่งเป็นโหนด (node) ที่ต่อเชื่อมกับโหนดเริ่มต้น (Start node)

จากนั้นทุก SOP ทุกประโยคจะเข้าสู่การเรียนรู้รูปแบบประโยคคำถามด้วยแบบจำลองมาร์คอฟ ซึ่งมีสมการได้ดังต่อไปนี้

$$P(Tag_i, Tag_{i-1}) = \frac{count(Tag_{i-1}, Tag_i)}{count(Tag_{i-1})} \qquad (21)$$

ในการเรียนรู้รูปแบบประโยคคำถามนั้น จะเป็นการเรียนรู้เฉพาะส่วนที่เป็นกลุ่มคำกริยา (Verb Phrase: VP) ของประโยคเท่านั้น โดยวิธีการระบุขอบเขตการเรียนรู้ในส่วนที่เหลือของประโยคคำถามจะเป็นการนำลำดับการเกิดขึ้นของแทกซ์ ที่ถูกระบุไว้มาทีละ 2 ตำแหน่ง และเมื่อตำแหน่งที่ทำการพิจารณาเป็นแทกซ์ที่ระบุหน้าที่ของคำให้เป็นคำกริยาตัวแรกในประโยค จึงจะหยุดการเรียนรู้ลำดับของการเกิดแทกซ์ในประโยคคำถามนั้น ๆ โดยแทกซ์ที่ระบุคำกริยาในประโยคได้แก่ VB, VBD, VBG, VBN, VBP และ VBZ ซึ่งแทกซ์ระบุหน้าที่เหล่านี้เป็นตัวแทนคำกริยาทั้งหมด และส่วนที่เหลือจะถูกมองเป็นวัตถุ (Object) ซึ่งสามารถเป็นคำหรือกลุ่มคำอะไรก็ได้ จากนั้นนำเข้าสู่ขั้นตอนการเรียนรู้เพื่อสร้างแบบจำลอง จากตัวอย่างประโยคคำถามที่มีสามารถแสดงวิธีการเรียนรู้ได้ดังต่อไปนี้

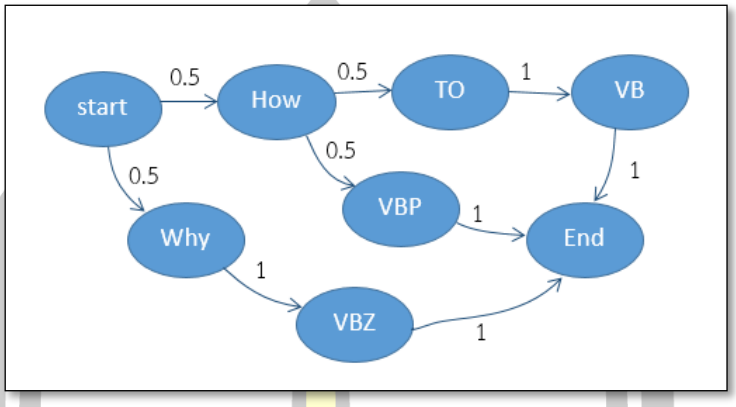
P(<How> | <Start>) = 2/4 = 0.5
P(<Why> | <Start>) = 2/4 = 0.5
P(TO | <How>) = 1/2 = 0.5
P(VB | TO) = 1/1 = 1
P(VBP | <How>) = 1/2 = 0.5
P(VBZ | <Why>) = 2/2 = 1
P(<End> | VB) = 1/1 = 1
P(<End> | VBZ) = 2/2 = 1
P(<End> | VBP) = 1/1 = 1

จากนั้นเมื่อนำค่าที่ได้เข้าสู่เมทริกซ์การเปลี่ยนแปลง (Transition Matrix) เพื่อสร้างเป็นแบบจำลองมาร์คอฟในขั้นตอนต่อไป สามารถแสดงได้ดังตารางที่ 3.3 และ รูปที่ 3.4 ดังนี้

ตารางที่ 3.3 ตัวอย่างเมทริกซ์ของจำลองมาร์คอฟ

	<Start>	<How>	TO	VB	VBP	VBZ	<Why>	<End>
<Start>	0	**2/4**	0	0	0	0	**2/4**	0
<How>	0	0	**1/2**	0	**1/2**	0	0	0
TO	0	0	0	**1/1**	0	0	0	0
VB	0	0	0	0	0	0	0	**1/1**
VBP	0	0	0	0	0	0	0	**1/1**
VBZ	0	0	0	0	0	0	0	**2/2**
<Why>	0	0	0	0	0	**2/2**	0	0
<End>	0	0	0	0	0	0	0	0

จากตาราง Transition matrix ที่ได้สามารถแสดงให้อยู่ในรูปของกราฟได้ดังต่อไปนี้



รูปที่ 3.4 ตัวอย่างแบบจำลองมาร์คอฟที่ได้

เมื่อนำแบบจำลองมาร์คอฟที่ได้มาหาค่าความน่าจะเป็นรวมหรือการทำการประมาณค่ามาร์คอฟ (Markov Assumption) เพื่อให้ได้ค่าความน่าจะเป็นโดยรวมออกมาของแต่ละรูปแบบที่สามารถเป็นไปได้ โดยใช้สมการดังต่อไปนี้

$$P(Tag_1, Tag_2, Tag_3, \dots, Tag_n) = \prod_{i}^{n} P(Tag_i | Tag_1, Tag_2, \dots, Tag_{i-1}) \qquad (22)$$

ดังนั้นค่าความน่าจะเป็นที่ได้คือค่าความน่าจะเป็นที่เกิดขึ้นของลำดับการเกิดแทกซ์ จากประโยคคำถาม โดยค่าที่ได้จะเป็นการบ่งบอกถึงโอกาสในการเกิดขึ้นของลำดับของแทกซ์ที่ระบุประเภทของคำนั่นเอง สามารถแสดงได้ดังตารางที่ 3.4

ตารางที่ 3.4 ตัวอย่างรูปแบบของคำถามที่เรียนรู้ได้ด้วยแบบจำลองมาร์คอฟ

ประเภทคำถาม	รูปแบบที่ได้	ความหมาย
How	TO VB	How + TO + Verb (Base)
	VBP	How + Verb (Singular present)
Why	VBZ	Why + Verb (3rd Person singular present)

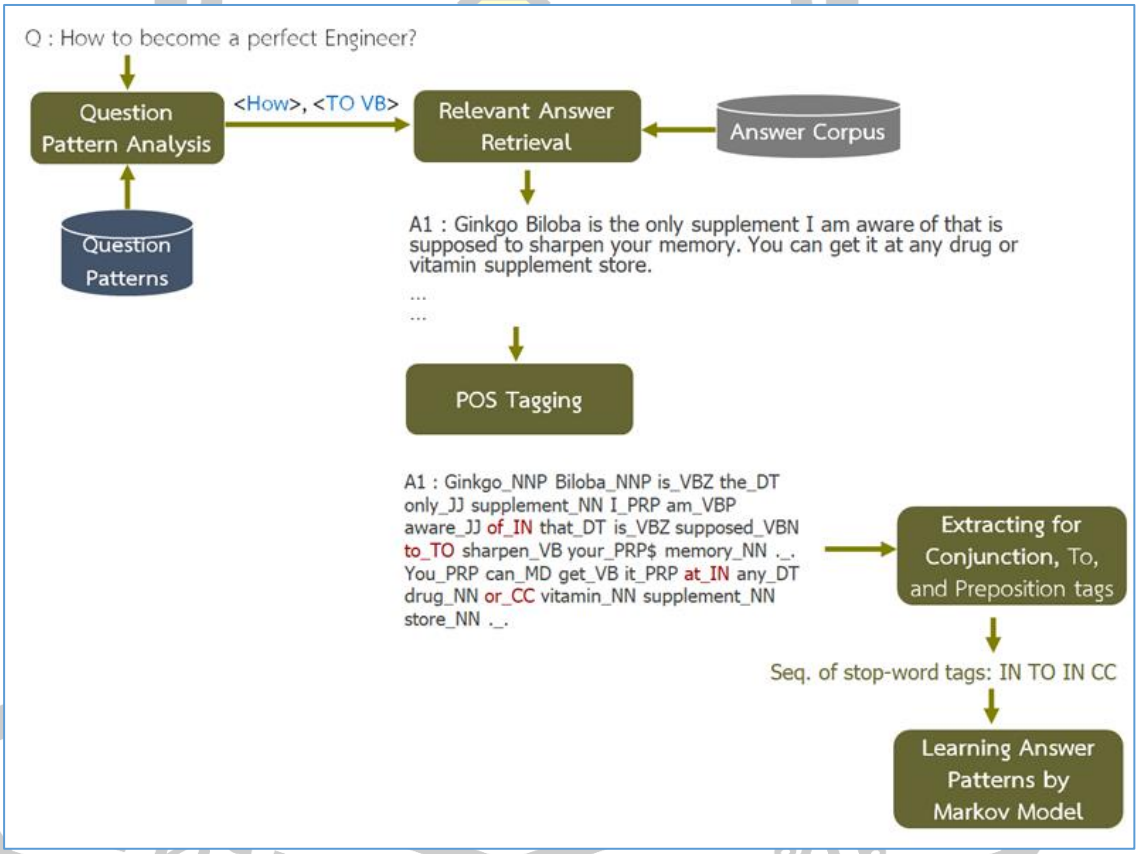
ในการเรียนรู้รูปแบบคำถามจากชุดข้อมูล MS MARCO เวอร์ชัน 2.1 ทั้งสิ้นจำนวน 400 ชุดคำถามต่อประเภทคำถาม รวมเป็นจำนวนทั้งสิ้น 800 ชุดคำถาม ได้รูปแบบของคำถามทั้งสิ้นจำนวน 96 รูปแบบ แบ่งออกเป็นรูปแบบคำถามของคำถามประเภท “*How*” และ “*Why*” ดังแสดงในตารางที่ 3.5

ตารางที่ 3.5 จำนวนรูปแบบของคำถาม

ประเภทคำถาม	จำนวนรูปแบบที่ได้
How	54
Why	42

3.3.2 การเรียนรู้รูปแบบคำตอบด้วยการใช้คำหยุด (Learning Answer Patterns using Stop Words)

ในส่วนของการเรียนรู้รูปแบบคำตอบ ในงานวิจัยนี้จะใช้ประโยชน์จากคำหยุด โดยกรอบการดำเนินงานเพื่อให้ได้รู้แบบของคำตอบด้วยการใช้คำหยุดสามารถแสดงได้ดังรูปที่ 3.5



รูปที่ 3.5 กรอบการดำเนินงานเพื่อการสร้างรูปแบบคำตอบด้วยคำหยุด

ขั้นตอนที่ 1: การค้นคืนคำตอบที่สอดคล้องกับรูปแบบคำถาม

ในขั้นตอนนี้จะทำการค้นคืนคำตอบทั้งหมดภายใต้รูปแบบคำถามนั้นๆ จากภาพที่ 3.5 จะเห็นว่าประโยคคำถามที่นำเข้ามาคือ “*How to become a perfect engineer?*” ซึ่งเมื่อนำเอา

ประโยคคำถามนี้ไปตรวจสอบกับรูปแบบคำถามที่ได้จากหัวข้อ 3.3.1 พบว่าคำถามนี้ตรงกับรูปแบบของคำถามคือ “<How> + <TO VB>” ดังนั้นคำตอบในคลังข้อมูลที่ตรงกับคำถามรูปแบบนี้จะถูกค้นขึ้นมา เพื่อใช้ในการเรียนรู้รูปแบบคำตอบ

ขั้นตอนที่ 2: การทำ POS tagging

เมื่อคำตอบที่เกี่ยวข้องกับคำถามนั้นๆ ถูกค้นคืนขึ้นมา จะเข้าสู่การทำ POS tagging ที่ดำเนินการผ่าน POS tagger ของ Stanford และใช้ Penn Treebank เป็นเซตของแทกซ์ที่ระบุประเภทหรือหน้าที่ของคำ ผลลัพธ์ที่ได้สามารถแสดงได้ดังตารางที่ 3.6

ตารางที่ 3.6 ตัวอย่างการทำ POS tagging ของเอกสารคำตอบ

คำตอบต้นฉบับ	คำตอบที่ผ่านการทำ POS tagging
Ginkgo Biloba is the only supple-ment I am aware of that is supposed to sharpen your memory. You can get it at any drug or vitamin supplement store.	Ginkgo_NNP Biloba_NNP is_VBZ the_DT only_JJ supplement_NN I_PRP am_VBP aware_JJ of_IN that_DT is_VBZ supposed_VBN to_TO sharpen_VB your_PRP$ memory_NN ._. You_PRP can_MD get_VB it_PRP at_IN any_DT drug_NN or_CC vitamin_NN supplement_NN store_NN ._.

ขั้นตอนที่ 3: การสกัดคำหยุดจากคำตอบ

ขั้นตอนนี้คือการระบุตำแหน่งของคำหยุดในคำตอบที่นำมาสร้างรูปแบบของคำตอบ โดยประเภทของคำที่สนใจได้แก่ Conjunction, To, และ Preposition ซึ่งก็คือ tag .ในกลุ่มของ CC, IN, และ TO นั่นเอง ผลลัพธ์ที่ได้สามารถแสดงได้ดังตารางที่ 3.7

ตารางที่ 3.7 ตัวอย่างการระบุตำแหน่งของคำหยุดในคำตอบ

คำตอบที่ผ่านการทำ POS tagging	ตำแหน่งของคำหยุดที่ตรวจจับได้
Ginkgo_NNP Biloba_NNP is_VBZ the_DT only_JJ supplement_NN I_PRP am_VBP aware_JJ of_IN that_DT is_VBZ supposed_VBN to_TO sharpen_VB your_PRP$ memory_NN ._. You_PRP can_MD get_VB it_PRP at_IN any_DT drug_NN or_CC vitamin_NN supplement_NN store_NN ._.	Ginkgo_NNP Biloba_NNP is_VBZ the_DT only_JJ supplement_NN I_PRP am_VBP aware_JJ ***of_IN*** that_DT is_VBZ supposed_VBN ***to_TO*** sharpen_VB your_PRP$ memory_NN ._. You_PRP can_MD get_VB it_PRP ***at_IN*** any_DT drug_NN ***or_CC*** vitamin_NN supplement_NN store_NN ._.

เมื่อพิจารณาจากตารางที่ 3.7 จะเห็นว่าคำตอบนี้ ภายหลังจากการระบุตำแหน่งของคำหยุดจะให้ผลลัพธ์คือ <IN TO IN CC>

ขั้นตอนที่ 4: การเรียนรู้รูปแบบคำตอบจากกลุ่มคำหยุดด้วยแบบจำลองมาร์คอฟ

ภายหลังจากที่ได้กลุ่มคำหยุดที่พบในชุดคำตอบแล้ว ตัวแทนของชุดคำตอบจะอยู่ในรูปแบบลำดับของกลุ่มคำหยุด (Sequence of stop-words: SSW) ซึ่งในงานวิจัยนี้จะใช้กลุ่มคำหยุดเหล่านี้ในการสร้างรูปแบบคำตอบด้วยแบบจำลองมาร์คอฟ

สมมติว่ามีลำดับของกลุ่มคำหยุด ดังต่อไปนี้

A1: IN -> TO -> IN -> CC

A2: IN -> IN -> -> TO -> IN -> IN -> IN

A3: IN -> CC -> TO -> IN -> IN -> IN

A4: TO -> IN -> CC -> IN -> IN

ลำดับของกลุ่มคำหยุด แต่ละชุดจะถูกติดสัญลักษณ์พิเศษคือ <start> และ <end> เพื่อแสดงถึงตำแหน่งของการเริ่มต้นและการสิ้นสุดของข้อมูลแต่ละกลุ่มตามลำดับ ดังแสดงได้ต่อไปนี้

<start> IN -> TO -> IN -> CC <end>

<start> IN -> IN -> -> TO -> IN -> IN -> IN <end>

<start> IN -> CC -> TO -> IN -> IN -> IN <end>

<start> TO -> IN -> CC -> IN -> IN <end>

จากนั้นข้อมูลเหล่านี้จะถูกนำมาเรียนรู้ด้วยแบบจำลองมาร์คอฟ

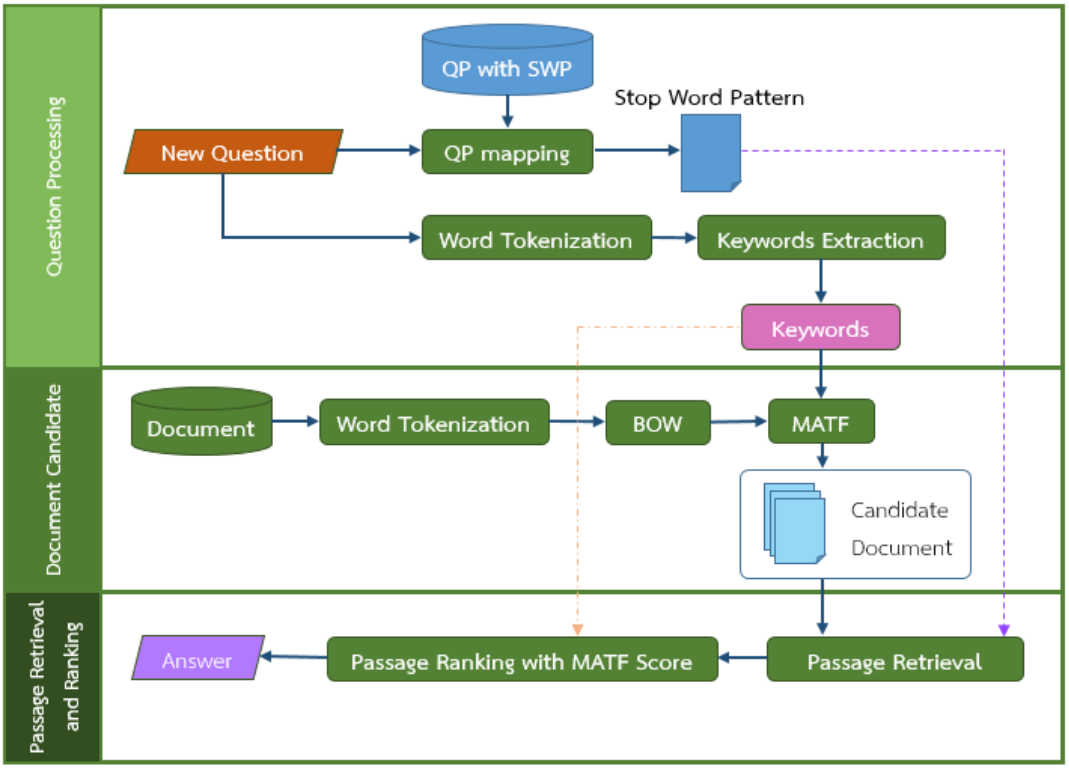
สำหรับการเรียนรู้รูปแบบคำตอบ ในงานวิจัยนี้ใช้ข้อมูลคำตอบแต่ละรูปแบบคำถามคือ 40 คำตอบ นั่นหมายความว่า เมื่อมีรูปแบบคำถาม “*How*” จำนวน 54 รูปแบบก็จะใช้จำนวนคำตอบในการเรียนรู้รูปแบบทั้งสิ้น 2,160 คำตอบ ที่เป็นรูปแบบของคำถามประเภท “*How*” ขณะที่รูปแบบคำถาม “*Why*” ที่มีจำนวน 42 รูปแบบ ที่เป็นรูปแบบของคำถามประเภท “*Why*” หลังจากนั้นก็จะใช้จำนวนคำตอบในการเรียนรู้รูปแบบทั้งสิ้น 1,680 คำตอบ และภายหลังจากการเรียนรู้ งานวิจัยนี้จะได้รูปแบบคำตอบ และตัวอย่างรูปแบบที่ได้ดังที่สรุปในตารางที่ 3.8

ตารางที่ 3.8 ตัวอย่างรูปแบบคำตอบที่สร้างจากกลุ่มคำหยุดและสรุปจำนวนรูปแบบคำตอบที่ได้

ประเภทคำถาม	ตัวอย่างรูปแบบคำตอบที่สร้างจากคำหยุด	จำนวนรูปแบบคำตอบของแต่ละประเภทคำถาม
How	IN -> TO -> IN -> CC	3,106
	IN -> IN -> -> TO -> IN -> IN -> IN	
Why	IN -> CC -> TO -> IN -> IN -> IN	3,049
	TO -> IN -> CC -> IN -> IN	

3.4 การประมวลผลในกระบวนการถาม-ตอบ (Question Answering Method)

จากภาพรวมของกระบวนการที่นำเสนอนั้นจะเห็นว่าในการหาคำตอบของการถาม-ตอบแบบนอนแฟคทอยด์ จากกระบวนการหลัก 3 กระบวนการ โดยเริ่มจาก ขั้นตอนที่หนึ่งการประมวลผลคำถาม เพื่อให้ได้คำค้นในการนำไปสืบค้นเอกสารในขั้นตอนที่สองที่เกี่ยวข้องกับคำถาม จากนั้นขั้นตอนที่สามจะเป็นการหาส่วนที่คาดว่าจะมีคำตอบของคำถาม ในเอกสารนั้น ๆ อยู่ และขั้นตอนสุดท้ายคือการสกัดเอาคำตอบจากส่วนที่คาดว่าจะมีคำตอบขั้นตอนที่สามนั้น สรุปเป็นคำตอบสุดท้ายกลับไปยังผู้ใช้งาน โดยรายละเอียดในแต่ละขั้นตอนของการหาคำตอบนั้นสามารถแสดงได้ตามขั้นตอนดังต่อไปนี้



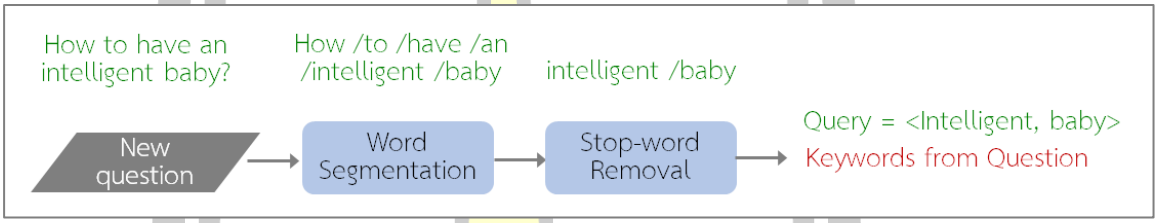
รูปที่ 3.6 ภาพรวมของกระบวนการการถาม-ตอบ

3.4.1 การประมวลผลคำถาม (Question Processing)

ในขั้นตอนนี้จะเป็นการประมวลผลคำถาม โดยมีวัตถุประสงค์ 2 ประการคือ

3.4.1.1 *การประมวลผลสำหรับการวิเคราะห์คำถามเพื่อสกัดเอา "คำสำคัญ (*Keyword*)"*

สำหรับขั้นตอนนี้ คือการประมวลผลเพื่อการวิเคราะห์คำถามและสกัดเอา *"คำสำคัญ (Keyword)"* จากคำถามนั้น ๆ โดยคำสำคัญดังกล่าวจะใช้ในขั้นตอนของการค้นคืนเอกสารที่คาดว่าจะมีคำตอบจากคลังเอกสาร จะมีการดำเนินดังที่แสดงในรูปที่ 3.7

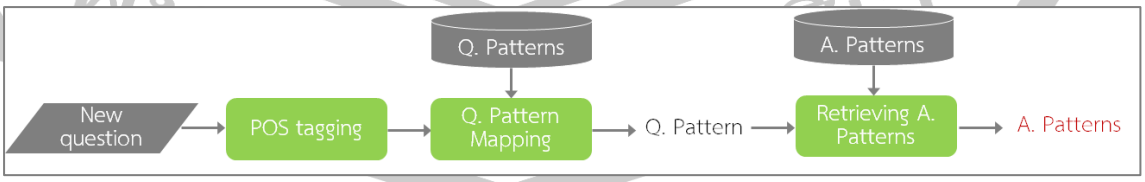


รูปที่ 3.7 ตัวอย่างการประมวลผลคำถามเพื่อสกัดเอา "คำสำคัญ"

จากรูปที่ 3.7 จะเห็นว่าเมื่อมีการป้อนคำถามเข้ามา คำถามนั้นๆ จะถูกประมวลผล โดยเริ่มจากการตัดคำ จากนั้นกำจัดคำหยุด คำที่เหลือจะเป็น *"คำสำคัญ"* ที่สามารถใช้ในการค้นคืนเอกสารที่คาดว่าจะมีคำตอบจากคลังเอกสาร

3.4.1.2 *การประมวลผลคำถามเพื่อประเมินรูปแบบของคำถามสำหรับการค้นหารูปแบบคำตอบที่เหมาะสม*

ในส่วนนี้จะเป็นส่วนของการวิเคราะห์คำถามเพื่อตรวจสอบว่า คำถามที่ป้อนเข้ามานั้นมีตรงกับรูปแบบคำถามใดที่ได้เรียนรู้ไว้ก่อนหน้า เมื่อทราบรูปแบบคำถามแล้ว ก็จะนำรูปแบบคำถามที่ได้ไปตรวจสอบว่า รูปแบบคำตอบที่เป็นไปได้หรือสอดคล้องกับรูปแบบคำถามนั้นๆ มีรูปแบบใดบ้าง เพราะรูปแบบคำตอบจะเป็นเครื่องมือที่ใช้ในการกำหนดขอบเขตของคำตอบที่ค้นคืนได้จากเอกสารที่คาดว่าจะมีคำตอบ ซึ่งกรอบการดำเนินงานสามารถแสดงได้ดังรูปที่ 3.8



รูปที่ 3.8 ตัวอย่างการประมวลผลคำถามเพื่อตรวจสอบรูปแบบคำถามและค้นคืนรูปแบบคำตอบที่เหมาะสม

จากรูปที่ 3.8 จะเห็นได้ว่า คำถามที่ป้อนเข้ามานั้นจะถูกประมวลผลด้วย POS tagging เป็นอันดับแรก ผลลัพธ์ที่ได้จะถูกนำไปเปรียบเทียบกับรูปแบบคำถามที่เรียนรู้ได้ เมื่อได้รูปแบบคำถามที่ตรงกับคำถามที่ป้อนเข้ามา รูปแบบคำถามนั้นจะถูกใช้ในการค้นหารูปแบบคำตอบที่สอดคล้องหรือเหมาะสมกับรูปแบบคำถามนั้น ๆ

ในขั้นตอนนี้จะเป็นการประมวลผลคำถามที่เป็นภาษาธรรมชาติจากผู้ใช้งาน โดยขั้นตอนนี้จุดประสงค์ก็เพื่อให้ทราบว่า คำถามนี้เกี่ยวข้องกับอะไร เป็นคำถามประเภทไหน (Question Type) และมีวลีของคำถามชนิดใด เพื่อที่จะสามารถระบุกลุ่มของคำหยุดเพื่อนำมาวิเคราะห์หาคำตอบในขั้นตอนต่อไป

ยกตัวอย่างเช่นคำถามที่เข้ามาใหม่นั้นคือ “How to have an intelligent baby?” จะมีการประมวลผลอยู่ 2 ส่วนหลัก ได้แก่ ส่วนที่หนึ่ง คือ การนำคำถามที่เข้ามาใหม่เข้าไปหาส่วนของคำถาม เพื่อแยกประเภทของคำถามจากนั้นจับคู่กับคู่ของประเภทคำถามในคลังข้อมูลของกลุ่มคำหยุดที่ได้ทำการเรียนรู้ไว้ ส่วนที่สองคือการหาคำสำคัญจากประโยคคำถามที่ถามเข้ามา โดยวิธีการในส่วนที่หนึ่งจะคล้ายคลึงกับการเรียนรู้รูปแบบของคำถามประกอบไปด้วยขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 : ขั้นตอนการระบุตำแหน่งของ “How” และ “Why”

จากตัวอย่างประโยคคำถามที่ยกตัวอย่างมานั้นจะทำการระบุตำแหน่งของทั้ง “How” และ “Why” และแยกส่วนที่เหลือออกจากตำแหน่งที่ระบุจะสามารถแสดงได้ดังต่อไปนี้

Q : <How>, <to have an intelligent baby ?>

ขั้นตอนที่ 2 : ขั้นตอนการระบุหน้าที่ของคำให้ส่วนที่เหลือในประโยค

เมื่อทราบตำแหน่งของ “*How*”, “*Why*” และส่วนที่เหลือในประโยคคำถามแล้วนั้นจะทำการระบุหน้าที่ของคำให้กับส่วนที่เหลือในประโยคคำถามเพื่อที่จะค้นหารูปแบบของประโยคคำถามว่าตรงกับรูปแบบใดจากประโยคคำถามที่ได้ทำการเรียนรู้มา โดยสามารถแสดงหน้าที่ของคำภายในส่วนที่เหลือของประโยคคำถามได้ดังต่อไปนี้

Q : <How>, <TO VB DT JJ NN .>

และเมื่อได้ระบุหน้าที่ของคำให้กับส่วนที่เหลือในประโยคคำถามเป็นที่เรียบร้อยแล้วนั้นจะนำหน้าที่ของคำเหล่านี้ไปเปรียบเทียบกับรูปแบบที่ได้ทำการเรียนรู้ไว้เพื่อนำคำหยุดที่ได้เรียนรู้ในขั้นตอนก่อนหน้ามาเก็บไว้เพื่อรอการประมวลผล

ขั้นตอนที่ 3 : ขั้นตอนการสกัดคำสำคัญในประโยคคำถาม

เมื่อทำการจับคู่เรียบร้อยแล้วจะทำการนำคำถามที่เข้ามาไปประมวลผลเพื่อที่จะหาคำสำคัญไว้สำหรับสืบค้นเอกสารที่เกี่ยวข้องและคาดว่าจะมีคำตอบ โดยเริ่มจากการตัดคำด้วยช่องว่าง (White Space) และตัดคำหยุด (Stop Word) ที่อยู่ในคำถาม

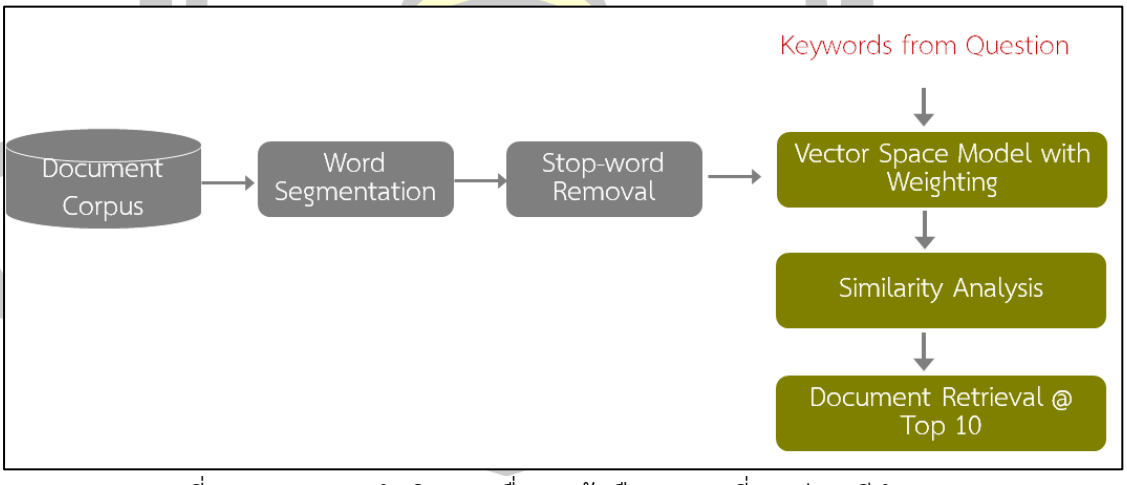
ตารางที่ 3.9 ผลลัพธ์ของการตัดคำหยุดและการหาคำคล้ายคลึง

ประโยคคำถาม	How to have an intelligent baby?
ตัดคำ / ตัดคำหยุด	How / to / have / an / intelligent / baby / ?
คำสำคัญ	intelligent / baby

เมื่อได้คำสำคัญเรียบร้อยแล้วนั้น จะเห็นว่าคำสัญนั้นถูกแบ่งเป็นกลุ่มย่อยตามจำนวนของคำที่ได้ภายหลังจากตัดคำหยุด

3.4.2 การค้นคืนเอกสารที่เกี่ยวข้อง (Document Candidate Retrieval)

ภายหลังจากที่ได้ “คำสำคัญ” จากการประมวลผลคำถาม แล้ว คำขอนั้น ๆ จะถูกนำมาสืบค้นเอกสาร (document) ที่เกี่ยวข้องทั้งหมดด้วยการประยุกต์ใช้เทคนิคการค้นคืนสารสนเทศ [19, 41] เพื่อให้ได้เอกสารที่เกี่ยวข้อง โดยการนำคำสำคัญเหล่านี้ค้นหาด้วยการค้นหาแบบหลายคำสำคัญ (Multiple Keyword Search) ด้วยวิธีการเปรียบเทียบความคล้ายคลึงโดยอาศัยหลักการของการเรียงคุณลักษณะของเอกสารแบบเมทริกซ์ หรือที่เรียกว่า Vector Space Model โดยกรอบการดำเนินงานเพื่อการค้นคืนเอกสารที่คาดว่าจะมีคำตอบสามารถแสดงได้ดังรูปที่ 3.9



รูปที่ 3.9 กรอบการดำเนินงานเพื่อการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ

โดยรายละเอียดของการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ สามารถอธิบายได้ดังนี้

ขั้นตอนที่ 1 : การเตรียมเอกสารในคลังเอกสาร

เอกสารในคลังข้อมูลทั้งหมดจะเข้าสู่กระบวนการตัดคำ และกำจัดคำหยุด จากนั้นนำเอาคำที่ได้หลังจากวิธีการดังกล่าว นำมาเตรียมไว้เป็นการทำเป็นดัชนีย์ให้กับเอกสารในคลังข้อมูลทุกเอกสารในคลังข้อมูล

$$\begin{array}{l} \text{Doc 1} \\ \text{Doc 2} \\ \text{Doc 3} \end{array} \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix}$$

รูปที่ 3.10 ตัวอย่างการนำเสนอการจัดเก็บดัชนีย์เอกสาร

ขั้นตอนที่ 2 : การแสดงเอกสารในรูปแบบ Vector Space Model

จากคำถาม “*How to have an intelligent baby?*” เมื่อผ่านการประมวลผลคำถามด้วยการตัดคำและการตัดคำหยุดก็จะได้คำสำคัญคือ *Intelligent* และ *baby* ซึ่งจะถูกใช้เป็นคำค้นหรือคำขอ (Query) นั่นเอง ขณะเดียวกัน สมมติมีเอกสารในคลังเอกสารจำนวน 3 เอกสาร โดยค่าคำค้นที่ได้จะถูกนำไปเทียบกับแต่ละเอกสารเพื่อหาค่าความคล้ายคลึง สามารถแสดงตัวอย่างการเปรียบเทียบการหาค่าความคล้ายคลึงได้ดังรูปที่ 3.11

$$\begin{array}{l} \text{Query} \\ \\ \text{Doc 1} \\ \text{Doc 2} \\ \text{Doc 3} \end{array} \begin{bmatrix} q_1 & q_2 & q_3 & q_4 \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix}$$

รูปที่ 3.11 ตัวอย่างการนำเสนอการเปรียบเทียบคำค้นกับเอกสาร

3.4.2.1 *การจัดอันดับเอกสารที่คาดว่าจะมีคำตอบด้วย* BM25

การหาความคล้ายคลึงระหว่างคำค้นและเอกสารที่คาดว่าจะมีคำตอบจากคลังข้อมูลนั้น สามารถคำนวณหาค่าความคล้ายคลึงจากสมการการหาค่าความคล้ายคลึงด้วย BM25 และ MATF เนื่องจากเป็นวิธีการที่เหมาะแก่การค้นคืนเอกสารที่มีความยาวของเอกสารต่างกันมาก สามารถแสดงการคำนวณได้จากสมการดังต่อไปนี้

$$BM25(Q,D) = \sum_{i=1}^{|Q|} \log\left(\frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \right) \times \frac{(k_1+1)\times tf_i}{K + tf_i} \times \frac{(k_2+1)\times qf_i}{k_2 + qf_i}$$

จากข้อมูลในขั้นตอนที่สองจะพบว่าในคลังข้อมูลมีทั้งสิ้น 3 เอกสาร ดังนั้นจะได้ว่า N = 3 และสมมุติให้ *Intelligent* ปรากฏอยู่ในเอกสารดังกล่าวจำนวน 2 เอกสาร จะได้ว่า n_1 = 2 ซึ่งคำว่า *Intelligent* ในเอกสารดังกล่าวเป็นจำนวน 15 ครั้ง จะได้ว่า tf_1 = 15

สำหรับคำว่า *baby* สมมุติให้ *baby* ปรากฏอยู่ในเอกสารดังกล่าวจำนวน 1 เอกสาร จะได้ว่า n_2 = 1 ซึ่งพบคำว่า *baby* ในเอกสารดังกล่าวเป็นจำนวน 25 ครั้ง จะได้ว่า tf_2 = 25 โดยความถี่ของคำค้นนั้น ๆ ในเซ็ตของคำค้นทั้งหมดในที่นี้ จะได้ว่า *qfi* = 1

หากกำหนดให้ค่า k_1 =1.2, k_2 = 100 และค่า b = 0.75 ตามมารตฐานที่นิยมใช้ในงาวิจัยก่อนหน้า ดังนั้นสามารถคำนวณหาค่า K ได้ดังนี้

$$K = k_1 \cdot ((1-b) + b \cdot \frac{doc_length}{avg_doc_length})$$

K = 1.2 x (0.25 + 0.75 x 0.9) ดังนั้นจะได้ค่า K = 1.11

ซึ่งเป็นค่าที่ไม่เกี่ยวข้องกับชุดข้อมูลใด ๆ การใส่ค่าดังกล่าวเข้าไปเพื่อที่จะเป็นการปรับระดับของการปรับสมดุลให้กับอัลกอริทึมการจัดอันดับ อ้างอิงจากงานประชุมวิชาการที่เป็นองค์กรในการกำหนดมาตรฐานสำหรับการค้นคืนสารสนเทศ รวมไปถึงการจัดอันดับเอกสารที่ค้นคืนได้ [38, 39]

เนื่องด้วยเอกสารหรือส่วนของเอกสารที่จะทำการจัดอันดับนั้นไม่ทราบข้อมูลความเกี่ยวข้องทำให้ค่า r_i และ R มีค่าเท่ากับ 0

ตารางที่ 3.10 ตัวอย่างข้อมูลสำหรับจัดอันดับเอกสารที่คาดว่าจะมีคำตอบด้วย BM25

	จำนวน Intelligent	จำนวน baby
Document 1	15	25
Document 2	15	1
Document 3	1	0

$$BM25(Q,D) = \left[\log\left(\frac{(0+0.5)\div(0-0+0.5)}{(2-0+0.5)\div(3-2-0-0+0.5)}\right) \times \frac{(1.2+1)\cdot 15}{1.11+15} \times \frac{(100+1)\cdot 1}{100+1}\right] + \left[\log\left(\frac{(0+0.5)\div(0-0+0.5)}{(1-0+0.5)\div(3-1-0-0+0.5)}\right) \times \frac{(1.2+1)\cdot 25}{1.11+25} \times \frac{(100+1)\cdot 1}{100+1}\right]$$

$$= -0.38204$$

ตารางที่ 3.11 ผลลัพธ์ของเอกสารที่คาดว่าจะมีคำตอบที่ถูกให้คะแนนด้วย BM25

	BM25(Intelligent)	BM25(baby)	BM25 Score
Document 1	-0.38204	0.31386	-0.06817
Document 2	-0.23131	0.39940	0.16808
Document 3	-0.31386	0.35625	0.04238

3.4.2.2 *การจัดอันดับเอกสารที่คาดว่าจะมีคำตอบด้วย* MATF

ในการค้นคืนเอกสารที่คาดว่าจะมีคำตอบด้วย MATF นั้นจะทำการยกตัวอย่างโดยที่กำหนดให้มีคำสำคัญในการจัดอันดับเอกสารที่คาดว่าจะมีคำตอบอยู่ทั้งสิ้นจำนวน 2 คำได้แก่ Q = {intelligent, baby} จำนวนเอกสารทั้งหมด 50,000 เอกสาร ความยาวเฉลี่ยของเอกสารทั้งหมดในคลังข้อมูลอยู่ที่ 3,000 คำ เอกสารในคลังข้อมูลที่ปรากฏคำสำคัญทั้งสองคำเท่ากับ 300 เอกสาร และ 120 เอกสารตามลำดับ พร้อมทั้งจำนวนความถี่ของคำสำคัญที่ไปปรากฏอยู่ในทุกเอกสารในคลังข้อมูลคือ 900 และ 460 ตามลำดับ จากนั้นข้อมูลของเอกสารในแต่ละเอกสารที่พบเจอคำสำคัญทั้งสองคำสามารถแสดงได้ดังตารางดังต่อไปนี้

ตารางที่ 3.12 ตัวอย่างข้อมูลสำหรับจัดอันดับเอกสารที่คาดว่าจะมีคำตอบด้วย MATF

	ความยาวเอกสาร	จำนวนคำที่ไม่ซ้ำกัน	คำสำคัญ	
			intelligent	baby
Document 1	2300	1400	18	2
Document 2	3100	2200	1	8
Document 3	800	450	16	7
Document 4	1600	900	11	4

เมื่อนำมาคำนวณจะสามารถหาค่าความคล้ายคลึงของเอกสารกับคำสำคัญดังกล่าวได้ดังต่อไปนี้

$$TFF_{(q,D)} = w \times BRITF_{(t,D)} + (1-w) \times BLRTF_{(t,D)} \quad (23)$$

จากสมการของ TFF จะพบว่ามีการนำค่า w ซึ่งเป็นค่าน้ำหนักที่เราให้ความสนใจในเอกสารที่มีลักษณะใดมากเป็นพิเศษ มาคำนวณร่วมกับค่าของ BRITF และ BLRTF ซึ่งสามารถคำนวณหาค่าได้จากสมการดังต่อไปนี้

$$RITF_{(\text{intelligent,Document 1})} = \frac{18}{2300 \div 1400}$$

$$= 10.9568$$

$$BRITF_{(\text{intelligent,Document 1})} = \frac{10.9568}{1 + 10.95685}$$

$$= 0.91636$$

จากนั้นจะเป็นการคำนวณหาค่า BLRTF จากสมการดังต่อไปนี้

$$LRTF_{(\text{intelligent,Document 1})} = 18 \times log_2(1 + \frac{3000}{2300})$$

$$= 21.67845$$

$$BLRTF_{(\text{intelligent,Document 1})} = \frac{21.67845}{1 + 21.67845}$$

$$= 0.955905$$

เมื่อได้ค่า BRITF และค่า BLRTF แล้วจากนั้นให้ทำการนำค่าทั้งสองไปคำนวณกับน้ำหนักที่เราต้องการ ในที่นี้จะใช้ค่ากึ่งกลางสำหรับค่าทั้งสองตัวคือ 0.5 เพื่อที่จะคำนวณหาค่า TFF จากสมการดังต่อไปนี้

$$TFF_{(\text{intelligent,Document 1})} = 0.5 \times 0.91636 + (1 - 0.5) \times 0.955905$$

$$= 0.9361344$$

เมื่อทำการคำนวณหาค่า TFF ตามสมการเป็นที่เรียบร้อยแล้วนั้น ให้เราทำการหาค่า TDF ที่เป็นการพิจารณาเอกสารที่คาดว่าจะมีคำตอบในปัจจุบันกับคลังเอกสารที่มีทั้งหมด สามารถหาได้จากสมการดังต่อไปนี้

$$TDF_{(q,C)} = IDF_{(q,C)} \times \frac{AEF_{(q,C)}}{1 + AEF_{(q,C)}} \quad (24)$$

ในสมการของ TDF จะนำค่าในส่วนของน้ำหนักที่มองต่อเอกสารทั้งคลังเอกสาร โดยเริ่มจากค่า IDF และค่าความยาวของเอกสารที่คาดว่าจะมีคำตอบในปัจจุบัน ต่อเอกสารทั้งคลังเอกสาร โดยสามารถคำนวณหาค่าได้จากสมการดังต่อไปนี้

$$IDF_{(\text{intelligent},Collection)} = log_{10}(50000/300)$$

$$= 2.22184$$

$$AEF_{(\text{intelligent},Collection)} = \frac{900}{300}$$

$$= 3$$

$$TDF_{(\text{intelligent},Collection)} = 2.22184 \times \frac{3}{1+3}$$

$$= 1.666386$$

เมื่อทำการคำนวณหาค่า TFF และ TDF เป็นที่เรียบร้อยแล้วนั้นจะทำการนำหาทั้งสองมาคำนวณหาค่าความคล้างคลึงด้วย MATF จะสามารถแสดงได้ดังสมการดังต่อไปนี้

$$MATF_{(\text{intelligent},Document\ 1)} = 0.9361344 \times 1.666386$$

$$= 1.559961882$$

$$MATF_{(\text{intelligent},Document\ 1)} = \frac{1.559961882}{1+1.559961882}$$

$$= 0.60936918$$

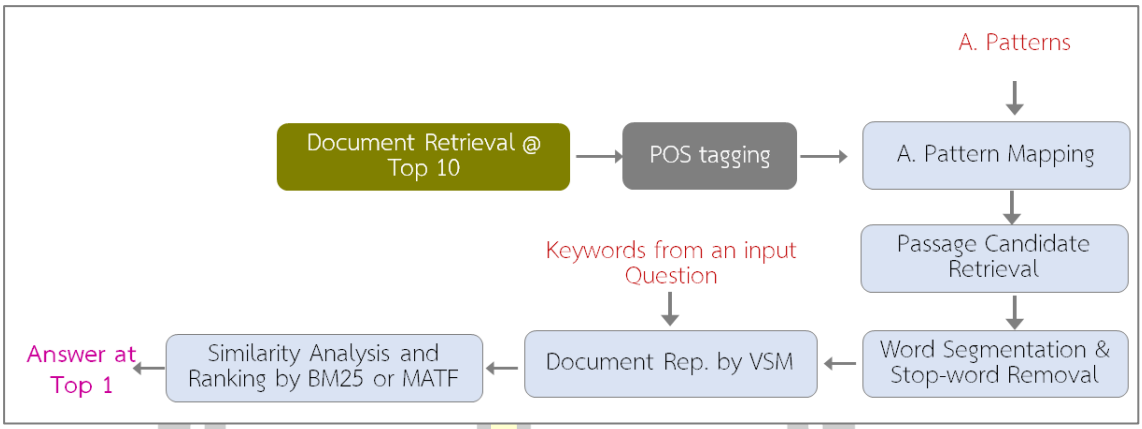
เมื่อทำการคำนวณตามสมการที่มีแล้วสุดท้ายเราจะได้ค่าความคล้ายคลึงระหว่างคำสำคัญและเอกสารที่คาดว่าจะมีคำตอบ โดยจะทำการคำนวณแบบนี้กับทุกคำสำคัญไปทีละเอกสาร จนครบทุกเอกสารที่จัดเก็บอยู่ในคลังข้อมูลนั่นเอง โดยค่าความคล้ายคลึงของเอกสารที่เหลือสามารถคำนวณได้ตามตารางที่ 3.13

ตารางที่ 3.13 ตัวอย่างข้อมูลสำหรับจัดอันดับเอกสารที่คาดว่าจะมีคำตอบด้วย MATF

	MATF(intelligent)	MATF(baby)	MATF Score
Document 1	0.6093	0.5660	1.1754
Document 2	0.4310	0.6434	1.0744
Document 3	0.6094	0.6435	1.2529
Document 4	0.6005	0.6171	1.2176

3.4.3 การค้นคืนส่วนของเอกสารที่เกี่ยวข้อง (Passage Candidate Retrieval)

ภายหลังจากที่ทำการค้นคืนเอกสารที่คาดว่าจะมีคำตอบได้แล้วนั้น ซึ่งจะเป็นการเลือกเอกสารที่มีค่าความคล้ายคลึงที่ดีที่สุด 10 อันดับแรก จากนั้นจะทำการค้นหาส่วนของเอกสารเหล่านี้ที่คาดว่าจะมีคำตอบ โดยสามารถแสดงขั้นตอนการค้นหาตำตอบที่เหมาะสมที่สุดดังรูปที่ 3.12



รูปที่ 3.12 ภาพรวมการค้นคืนส่วนของเอกสารที่เกี่ยวข้อง

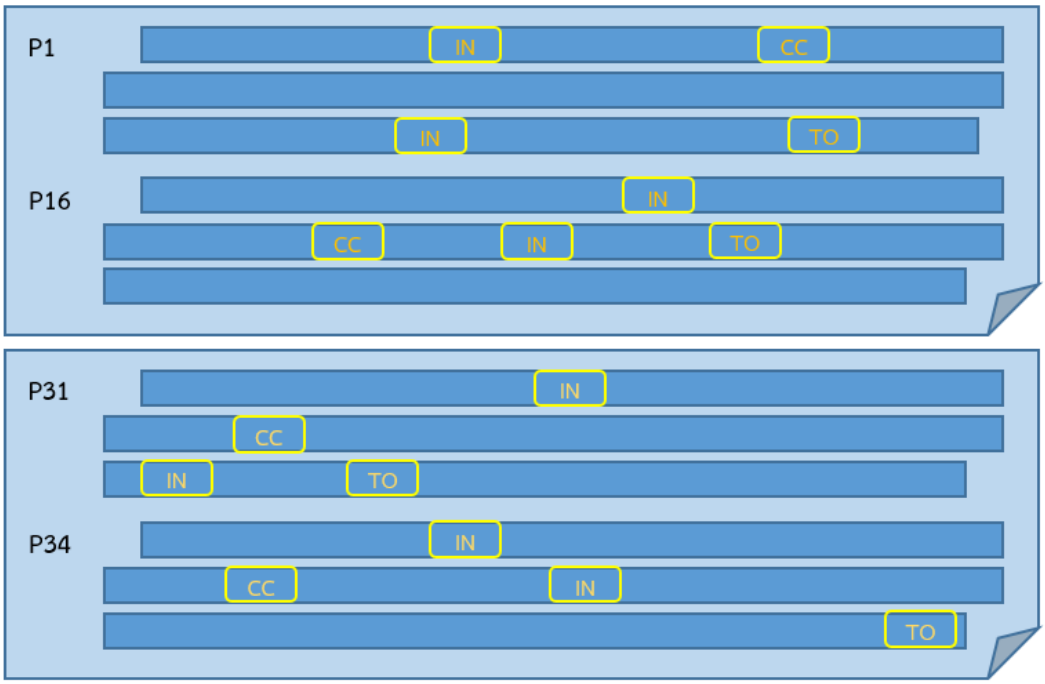
ขั้นตอนที่ 1: การจัดรูปแบบเอกสารที่ค้นคืนได้ให้ง่ายต่อการเทียบกับรูปแบบของคำตอบ

ขั้นตอนนี้จะเริ่มจากการทำ POS tagging จากนั้นจะนำเอารูปแบบคำตอบที่เหมาะสมกับรูปแบบคำถามนั้นๆ มาเทียบกับเอกสาร (Pattern mapping) เพื่อหาว่าส่วนของเอกสาร (passage) ใดของเอกสารที่ตรงกับรูปแบบของคำตอบ หาก “ส่วนของเอกสาร” ใด ตรงกับรูปแบบคำตอบที่เรียนรู้ไว้ “ส่วนของเอกสาร” เหล่านั้นจะถูกค้นคืนมาเป็น กลายเป็น “ส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ (passage candidate)”

ขั้นตอนที่ 2 : การค้นหาคำตอบที่เหมาะสมที่สุด

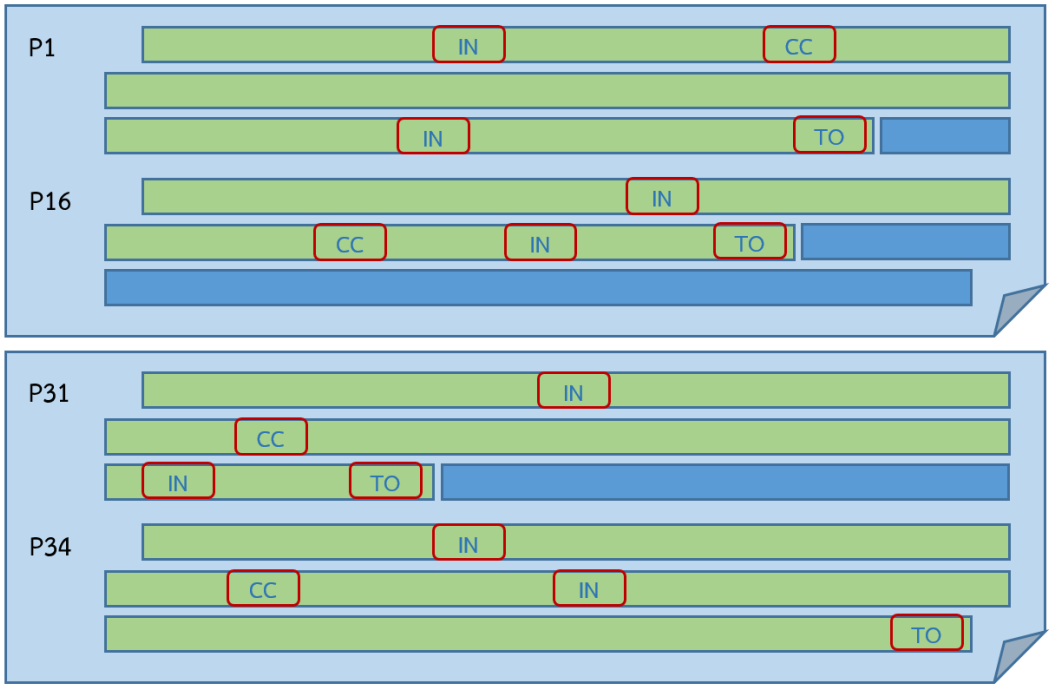
เมื่อได้ “ส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ” แล้ว ส่วนของเอกสารเหล่านี้จะเข้าสู่กระบวนการตัดคำ และกำจัดคำหยุด จากนั้น “ส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ” เหล่านี้จะถูกนำมาเปรียบเทียบกับ “คำสำคัญ” ที่ได้จากคำถามที่ป้อนเข้ามา เพื่อคำนวณหาความสอดคล้องหรือค่าความคล้ายคลึง (Similarity Score) ด้วยวิธี BM25 และ MATF โดยที่ “ส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ” ที่มีค่าความคล้ายคลึงอยู่ที่อันดับที่หนึ่ง (Top 1) จะถือว่าเป็นคำตอบของคำถามที่ป้อนเข้ามา

สำหรับการค้นคืนส่วนของข้อความที่คาดว่าจะมีคำตอบนั้น จะใช้ประโยชน์จาก 2 ส่วนหลักๆ คือ รูปแบบของคำหยุดที่เรียนรู้ได้ที่มีการกล่าวไว้ในหัวข้อที่ 3.3.2 และคำสำคัญที่ได้จากกระบวนการประมวลผลคำถาม โดยสมมุติให้รูปแบบที่ต้องการคือ IN -> CC -> IN -> TO ขั้นตอนในการค้นคืนส่วนของข้อความที่คาดว่าจะมีคำตอบนั้นจะสามารถแสดงได้ดังรูปที่ 3.13



รูปที่ 3.13 ตัวอย่างการค้นคืนส่วนของเอกสารที่คาดว่าจะมีคำตอบ

จากรูปที่ 3.13 เป็นขั้นตอนที่หาย่อหน้าของเอกสารที่คาดว่าจะมี “คำตอบ” ที่สอดคล้องกับคำถามจากข้อความที่สืบค้นได้ในขั้นตอนที่ 2 โดยการระบุตำแหน่งของคำค้นในส่วนของเอกสารนั้น ๆ พร้อมทั้งระบุรูปแบบของคำตอบตามที่ได้เรียนรู้มาด้วย รูปแบบของการเกิดคำหยุดในประเภทของคำถามดังกล่าว โดยสามารถหาคะแนนในแต่ละส่วนของเอกสารโดยเทียบกับคำค้น ซึ่งในงานวิจัยนี้ เราได้ศึกษาการจัดอันดับส่วนของเอกสารที่คาดว่าจะเป็นคำตอบด้วย 2 เทคนิค คือ BM25 และ MATF โดยการประยุกต์ใช้งานของอัลกอริทึมแต่ละตัว



รูปที่ 3.14 การระบุตำแหน่งของคำค้นและกลุ่มของคำหยุดในเอกสาร

3.4.3.1 *การหาค่าความคล้ายคลึงระหว่างคำสำคัญจากคำถามและส่วนของเอกสารที่คาดว่าจะเป็นคำตอบด้วย* BM25

ขั้นตอนนี้เป็นขั้นตอนในการหาข้อความที่เกี่ยวข้องหรือสอดคล้องหรือตรงกับคำถาม ตัวอย่างของเทคนิคในการหาข้อความที่เกี่ยวข้องหรือสอดคล้องหรือตรงกับคำถาม จะทำการประยุกต์ใช้เทคนิคของการจัดอันดับเอกสาร (Ranking) ด้วยเทคนิคของ BM25 และ MATF เช่นเดียวกับการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ ในการหาความคล้ายคลึงของข้อความที่เกี่ยวข้องกับคำถาม จะใช้คำสำคัญ และรูปแบบของคำหยุดที่ได้จากการเรียนรู้ข้างต้น ในการให้คะแนนส่วนของเอกสารที่คาดว่าจะมีคำตอบอยู่ โดยในขั้นตอนนี้จะทำการค้นคืนย่อหน้าของเอกสารขึ้นมาก่อน จากนั้นจึงทำการจัดอันดับย่อหน้าที่มีคะแนนมากที่สุดเพื่อนำไปหาส่วนที่คาดว่าจะเป็นคำตอบภายในย่อหน้านั้น ๆ

เนื่องจากการให้คะแนนส่วนของเอกสารที่คาดว่าจะเป็นคำตอบนั้นใช้วิธีการให้คะแนนแบบเดียวกับการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ แตกต่างกันในส่วนของขนาด และความยาวของข้อความที่ต้องการจัดอันดับ ดังนั้นในหัวข้อนี้จะยกตัวอย่างด้วยจำนวนข้อความและเอกสารที่น้อยลง แต่วิธีการคำนวณยังคงเหมือนเดิม สามารถคำนวณได้ตามข้อมูลดังต่อไปนี้

กำหนดให้มีคำสำคัญในการจัดอันดับอยู่ 2 คำได้แก่ Q = {intelligent, baby} และส่วนของเอกสารที่มีทั้งหมด 1,000 ข้อความ พบว่า intelligent ปรากฏอยู่ในกลุ่มของส่วนของเอกสารดังกล่าวเป็นจำนวน 200 ข้อความ baby จำนวน 90 ข้อความ และส่วนของเอกสารที่คาดว่าจะเป็นคำตอบดังกล่าวพบ intelligent เป็นจำนวน 4 ครั้ง และ baby จำนวน 2 ครั้ง และมีความยาวคิดเป็น 70% ของค่าเฉลี่ยความยาวทั้งหมด

จากข้อมูลในขั้นตอนที่สองจะพบว่าในคลังข้อมูลมีทั้งสิ้น 3 เอกสาร ดังนั้นจะได้ว่า N = 1,000

สมมุติให้ *Intelligent* ปรากฏอยู่ในเอกสารดังกล่าวจำนวน 2 เอกสาร จะได้ว่า n_1 = 200

พบ *Intelligent* ในเอกสารดังกล่าวเป็นจำนวน 15 ครั้ง จะได้ว่า tf_1 = 4

สมมุติให้ *baby* ปรากฏอยู่ในเอกสารดังกล่าวจำนวน 1 เอกสาร จะได้ว่า n_2 = 90

พบ *baby* ในเอกสารดังกล่าวเป็นจำนวน 25 ครั้ง จะได้ว่า tf_2 = 2

ความถี่ของคำค้นนั้น ๆ ในเซ็ตของคำค้นทั้งหมดในที่นี้ จะได้ว่า *qfi* = 1

กำหนดให้ค่า k_1 =1.2, k_2 = 100 และค่า b = 0.75 ดังนั้นสามารถคำนวณหาค่า K ได้ดังนี้

K = 1.2 x (0.25 + 0.75 x 0.9) ดังนั้นจะได้ค่า K = 1.11

และเนื่องด้วยเอกสารหรือส่วนของเอกสารที่จะทำการจัดอันดับนั้นไม่ทราบข้อมูลความเกี่ยวข้องทำให้ r_i และ R มีค่าเท่ากับ 0 สามารถคำนวณได้จากตัวอย่างดังตารางที่ 3.14

ตารางที่ 3.14 ตัวอย่างข้อมูลสำหรับจัดอันดับด้วย BM25

	จำนวน intelligent	จำนวน baby
Passage 1	4	2
Passage 2	1	5
Passage 3	2	3
Passage 4	1	2
Passage 5	2	2

$$BM25(Q,P) = \left[\log\left(\frac{(0+0.5)\div(0-0+0.5)}{(200-0+0.5)\div(1000-200-0-0+0.5)}\right) \times \frac{(1.2+1)\cdot 4}{1.11+4} \times \frac{(100+1)\cdot 1}{100+1}\right] + \left[\log\left(\frac{(0+0.5)\div(0-0+0.5)}{(90-0+0.5)\div(1000-90-0-0+0.5)}\right)\right.$$

$$\times \frac{(1.2+1)\cdot 2}{1.11+2} \times \frac{(100+1)\cdot 1}{100+1}\Big]$$

$$= [0.60124 \times 1.784 \times 1] + [1.0026 \times 1.5017 \times 1]$$

$$= 1.0792 + 1.5056$$

$$= 2.5788$$

ตารางที่ 3.15 ผลลัพธ์ข้อมูลที่ถูกให้คะแนนด้วย BM25

	BM25(Query1)	BM25(Query2)	BM25 Score
Paragraph 1	1.0792	1.5056	2.5788
Paragraph 2	0.6853	1.8598	2.5452
Paragraph 3	0.9028	1.6838	2.5867
Paragraph 4	0.6853	1.5056	2.1910
Paragraph 5	0.9028	1.5056	2.4085

จากนั้นเมื่อทำการประมวลผลส่วนของเอกสารที่คาดว่าจะเป็นคำตอบด้วย BM25 แล้วนั้น จะเลือกขึ้นมาเพียงแค่ 1 คำตอบเพื่อนำไปเป็นคำตอบของคำถามที่ถูกถามเข้ามาโดยผู้ใช้งาน ดังนั้น การหาคำตอบด้วย BM25 ในตัวอย่างนี้นั้น จะได้ว่า ส่วนของเอกสารที่ 3 จะถูกนำไปเป็นคำตอบของคำถามที่ถูกถามเข้ามานั่นเอง

3.4.3.2 *การหาค่าความคล้ายคลึงระหว่างคำสำคัญจากคำถามและส่วนของเอกสารที่คาดว่าจะเป็นคำตอบด้วย* MATF

ในการจัดอันดับส่วนของเอกสารที่คาดว่าจะเป็นคำตอบด้วย MATF นั้นจะคล้ายคลึงกับการให้คะแนนเอกสารที่คาดว่าจะมีคำตอบในขั้นตอนก่อนหน้า แต่จะแตกต่างกันตรงที่ความยาวของข้อความ เนื่องจากข้อความที่ต้องการให้คะแนนนั้น เป็นข้อความที่อยู่ในรูปแบบของกลุ่มของข้อความไม่ใช่เอกสารทั้งเอกสารเหมือนดังขั้นตอนก่อนหน้า

เช่นเดียวกันกับการให้คะแนนส่วนที่คาดว่าจะเป็นคำตอบด้วย BM25 โดยในการให้คะแนนส่วนของเอกสารที่คาดว่าจะเป็นคำตอบด้วย MATF นั้นจะพิจารณาความยาวของส่วนของเอกสารที่คาดว่าจะเป็นคำตอบร่วมด้วย โดยสามารถยกตัวอย่างได้ดังต่อไปนี้

กำหนดให้มีคำสำคัญในการจัดอันดับเอกสารที่คาดว่าจะมีคำตอบอยู่ทั้งสิ้นจำนวน 2 คำ ได้แก่ Q = {intelligent, baby} จำนวนเอกสารทั้งหมด 1,000 ข้อความ ความยาวเฉลี่ยของส่วนของเอกสารที่คาดว่าจะเป็นคำตอบทั้งหมดในคลังข้อมูลอยู่ที่ 350 คำ ส่วนของเอกสารที่ค้นคืนได้และ

ปรากฏคำสำคัญทั้งสองคำเท่ากับ 200 ข้อความ และ 90 ข้อความตามลำดับ พร้อมทั้งจำนวนความถี่ของคำสำคัญที่ไปปรากฏอยู่ในทุก ๆ ส่วนของเอกสารที่ค้นคืนได้คือ 140 และ 60 ตามลำดับ จากนั้นข้อมูลของคำที่อยู่ในส่วนของเอกสารในแต่ละส่วนเอกสาร ที่พบเจอคำสำคัญทั้งสองคำสามารถแสดงได้ดังตารางดังต่อไปนี้

ตารางที่ 3.16 ตัวอย่างข้อมูลสำหรับจัดอันดับส่วนของเอกสารที่คาดว่าจะมีคำตอบด้วย MATF

	ความยาวเอกสาร	จำนวนคำที่ไม่ซ้ำกัน	intelligent	baby
Paragraph 1	230	140	4	2
Paragraph 2	310	220	1	5
Paragraph 3	80	45	2	3
Paragraph 4	160	90	1	2

จากข้อมูลที่กำหนดให้ในตารางที่ 3.16 จะสามารถนำไปเข้าสมการของ MATF เพื่อคำนวณหาค่าความคล้างคลึงระหว่างข้อความที่สนใจกับคำสำคัญได้ดังต่อไปนี้

$$RITF_{(\text{intelligent,Paragraph 1})} = \frac{4}{230 \div 140}$$

$$= 2.43478$$

$$BRITF_{(\text{intelligent,Paragraph 1})} = \frac{2.43478}{1 + 2.43478}$$

$$= 0.70886$$

จากนั้นจะเป็นการคำนวณหาค่า BLRTF จากสมการดังต่อไปนี้

$$LRTF_{(\text{intelligent,Paragraph 1})} = 4 \times log_2(1 + \frac{1000}{230})$$

$$= 5.3376$$

$$BLRTF_{(\text{intelligent,Paragraph 1})} = \frac{5.3376}{1 + 5.3376}$$

$$= 0.842213$$

จากนั้นนำค่าที่คำนวณได้ของทั้งสองสมการ ไปคำนวณหาค่า TFF ได้ดังต่อไปนี้

$$TFF_{(\text{intelligent,Document 1})} = 0.5 \times 0.70886 + (1 - 0.5) \times 0.842213$$

$$= 0.775537$$

จากนั้นทำการคำนวณหาค่า TDF เพื่อที่จะนำค่าที่ได้ไปเข้าสู่สมการของ MATF ในขั้นตอนถัดไป สามารถคำนวณได้จากสมการดังต่อไปนี้

$$IDF_{(\text{intelligent},Collection)} = log_{10}(1000/200)$$

$$= 0.69897$$

$$AEF_{(\text{intelligent},Collection)} = \frac{200}{140}$$

$$= 1.428571$$

$$TDF_{(\text{intelligent},Collection)} = 0.69897 \times \frac{1.428571}{1 + 1.428571}$$

$$= 0.287811$$

เมื่อทำการคำนวณหาค่า TFF และค่า TDF ได้ทั้งสองค่าแล้วจะทำการนำเข้าสู่สมการของ MATF เพื่อวัดค่าความคล้ายคลึงของส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ ได้ดังต่อไปนี้

$$MATF_{(\text{intelligent},Document\ 1)} = 0.775537 \times 0.287811$$

$$= 0.2232082$$

$$MATF_{(\text{intelligent},Document\ 1)} = \frac{0.2232082}{1 + 0.2232082}$$

$$= 0.18247778$$

เมื่อทำการคำนวณหาค่า MATF กับทุกข้อความตัวอย่างที่ค้นคืนได้แล้วนั้น จะสามารถแสดงค่าความคล้ายคลึงได้ดังตารางที่ 3.17

ตารางที่ 3.17 ตัวอย่างการจัดอันดับส่วนของเอกสารที่คาดว่าจะเป็นคำตอบด้วย MATF

	MATF(intelligent)	MATF(baby)	MATF Score
Paragraph 1	0.18247	0.21071	0.393194
Paragraph 2	0.11878	0.25367	0.372455
Paragraph 3	0.163531	0.23966	0.403199
Paragraph 4	0.124238	0.213676	0.337914

จะพบว่าส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ ที่ถูกจัดอันดับด้วย MATF นั้นส่วนของเอกสารที่คาดว่าจะเป็นคำตอบที่ได้คะแนนสูงที่สุดคือส่วนที่ 3 มีคะแนนอยู่ที่ 0.403 โดยส่วนของเอกสารที่คาดว่าจะเป็นคำตอบนั้นจะถูกนำไปเป็นคำตอบให้กับคำถามที่ถูกถามเข้ามา

3.5 การวัดประสิทธิภาพของกระบวนการที่นำเสนอ (Evaluation Method)

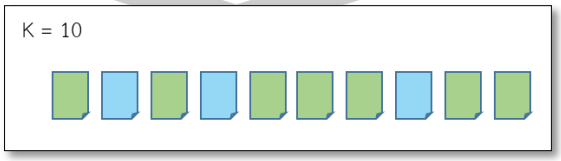
ในส่วนของขั้นตอนการวัดประสิทธิภาพนั้น ในงานวิจัยนี้ได้มีการวัดประสิทธิภาพของกระบวนการในแต่ละขั้นตอน โดยการวัดประสิทธิภาพที่เป็นมาตรฐานหนึ่งของการวัดประสิทธิภาพของกระบวนการประเภทนี้ได้แก่ การวัดค่ากลางของค่าเฉลี่ยความถูกต้อง ซึ่งเป็นการวัดประสิทธิภาพที่ใช้กับการจัดอันดับเอกสารที่มีคำค้นมากกว่า 1 คำค้น

3.5.1 การวัดประสิทธิภาพการค้นคืนเอกสารด้วย Precision at K (P@K)

ขั้นตอนในการวัดประสิทธิภาพการค้นคืนเอกสารด้วย Precision at K นั้นจะเป็นการวัดประสิทธิภาพจากการหาความเกี่ยวข้องกันกับคำสำคัญหรือคำค้น กับเอกสารหรือส่วนของเอกสารใด ๆ ที่ได้คืนจากกระบวนการการถาม-ตอบ ซึ่งเป็นการวัดประสิทธิภาพด้วยการสนใจผลลัพธ์ของการค้นคืนที่ *K* อันดับแรก เรียกว่า Precision at K (P@K) สามารถคำนวณได้จากสมการดังต่อไปนี้

$$P@K = \frac{Correct\ Doc}{|K\ Relevant\ Doc|} \quad (25)$$

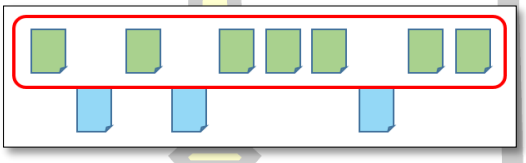
จากสมการดังกล่าวจะสมมุติให้มีเอกสารที่ค้นคืนได้เป็นจำนวนทั้งสิ้น 200 เอกสาร และกำหนดให้สนใจเฉพาะเอกสารจำนวน 10 อันดับแรก



รูปที่ 3.15 เอกสารที่สนใจจำนวน K อันดับ

ปรากฏว่าเอกสารที่ค้นคืนได้ใน 10 อันดับแรกนั้นมีเอกสารที่ถูกต้องเป็นจำนวนทั้งสิ้น 7 เอกสาร ดังนั้นค่า P@10 ที่ได้จะมีค่าเท่ากับ

$$P@10 = \frac{7}{10} = 0.7$$



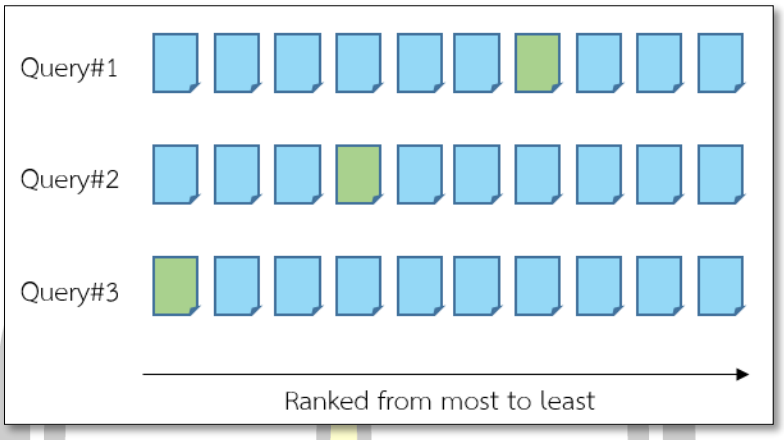
รูปที่ 3.16 เอกสารที่ค้นคืนและถูกต้องใน K อันดับ

3.5.2 การวัดประสิทธิภาพการค้นคืนเอกสารด้วย Mean Reciprocal Rank (MRR)

ในการวัดประสิทธิภาพด้วยค่า MRR นั้นเป็นการวัดประสิทธิภาพของการจัดอันดับ โดยที่จะสนใจใน *K* อันดับแรก โดยที่จะดูว่า นอกจากใน *K* อันดับแรกที่สนใจจะมีความถูกต้องมากน้อยเพียงใดแล้วนั้น ใน *K* อันดับแรกดังกล่าว ยังแสดงด้วยว่าเอกสารที่ถูกค้นคืนได้ถูกต้อง จะถูกค้นคืนขึ้นในอันดับที่เท่าใด และหากค่าที่ได้ มีค่าเข้าใกล้ 1 มากเท่าใด ก็จะแสดงว่ากระบวนการที่นำเสนอมีประสิทธิภาพในการจัดอันดับเอกสารที่คาดว่าจะมีคำตอบสูง โดย MRR สามารถคำนวณได้จากสมการดังต่อไปนี้

$$MRR = \frac{1}{|Q\ of\ Question|} \times \sum_{i=1}^{K} \frac{1}{rank_i} \quad (26)$$

จากสมการข้างต้น สมมติกำหนดให้ *K* ที่สนใจคือ 10 เอกสารและค่า Q คือ 3 โดยระบุให้มีค่า Q = {query1, query2, query3} จากนั้นกำหนดให้ค่าอันดับในแต่ละการจัดอันดับมีค่าเท่ากับ 7, 4 และ อันดับที่ 1 ในแต่ละคำค้นตามลำดับ สามารถแสดงได้ดัง รูปที่ 3.17



รูปที่ 3.17 ตัวอย่างการจัดอันดับเพื่อวัดประสิทธิภาพด้วย MRR ใน K อันดับ

จากนั้นเมื่อเราทราบอันดับของเอกสารที่ถูกต้องในแต่ละคำค้นที่เราสนใจแล้วนั้น ค่าอันดับที่ได้ในแต่ละคำค้นเราจะเรียกมันว่า Reciprocal จะนำค่าที่ได้นี้มาเข้าสมการในการคำนวณหาค่า MRR ได้ดังต่อไปนี้

$$Reciprocal\ of\ Query\#1 = \frac{1}{7}$$
$$Reciprocal\ of\ Query\#2 = \frac{1}{4}$$
$$Reciprocal\ of\ Query\#3 = \frac{1}{1}$$

ภายหลังจากที่ทราบค่า Reciprocal ของคำแต่ละคำที่อยู่ใน “คำค้น” สำหรับที่ใช้ในการจัดอันดับแล้วนั้น ก็จะนำค่าเหล่านั้นเข้าสมการของ MRR เพื่อทำการวัดประสิทธิภาพในการจัดอันดับได้ดังต่อไปนี้

$$MRR = \frac{1}{3} \times \left[\frac{1}{7} + \frac{1}{4} + \frac{1}{1}\right]$$
$$MRR = 0.4642$$

3.5.3 การวัดประสิทธิภาพคำตอบด้วย Bilingual Evaluation Understudy (BLEU)

เป็นการวัดประสิทธิภาพคำตอบที่ได้จากกระบวนการถามตอบ จากการประยุกต์ใช้เทคนิคการวัดประสิทธิภาพของการแปลภาษา โดยอาศัยค่าเฉลี่ยเรขาคณิต (Geometric Mean) โดยนำเอาข้อความมาเปรียบเทียบกัน โดยมีข้อความเฉลยเป็นตัวอ้างอิงความถูกต้องของข้อความ ในที่นี้เราจะมีจำนวนข้อความที่ใช้สำหรับการเปรียบเทียบอยู่ทั้งสิ้น 10 ข้อความต่อการเปรียบเทียบคำตอบสอง

คำตอบ และยังมีการประยุกต์ใช้สมการในการวัดค่าความแม่นยำให้กับค่าแกรม (Modified n-gram Precision) โดยสามารถคำนวณได้จากสมการดังต่อไปนี้

$$\log(BLEU - N) = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n \cdot \log(p_n) \qquad (27)$$

เมื่อ c คือ ความยาวของคำตอบที่จะทำการประเมิน และ r เป็นค่าความยาวของเอกสารที่ทำการเปรียบเทียบโดยที่ค่า p_n สามารถคำนวณได้จากสมการดังต่อไปนี้

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count(n-gram')} \qquad (28)$$

จากสมการของ BLEU ที่แสดงข้างต้นจะพบว่าค่า N ที่อยู่ในสมการหมายถึงจำนวนแกรมที่เราสนใจจะวัดประสิทธิภาพส่วนของเอกสารที่คาดว่าจะเป็นคำตอบโดยในงานวิจัยที่นำเสนอนั้น จะสนใจวัดประสิทธิภาพแบบแกรมเดียว โดยสมมุติให้ประโยคที่คาดว่าจะเป็นคำตอบที่ทำการค้นคืนได้และประโยคคำตอบที่เป็นเฉลยของคำถามสามารถแสดงได้ดังตารางที่ 3.18

ตารางที่ 3.18 ตัวอย่างข้อมูลการวัดประสิทธิภาพด้วย BLEU

คำตอบที่ได้	the	the	cat	the	the	cat	the
ประโยคเฉลย 1	the	cat	it	on	the	mat	
ประโยคเฉลย 2	there	is	a	cat	on	the	mat

จากตัวอย่างที่กำหนดให้จะพบว่าคำตอบที่ได้ มีความยาวทั้งสิ้น 7 คำ ในประโยคที่เป็นเฉลยประโยคแรกประกอบไปด้วยคำว่า “the” 2 ครั้ง และคำว่า “cat” 1 ครั้ง ทำให้สามารถคำนวณหาค่า P_n ได้จากสมการดังต่อไปนี้

$$p_{(the,cat)} = \frac{2+1}{7}$$

$$p_{(the,cat)} = \frac{3}{7}$$

$$= 0.42850$$

จากนั้นเมื่อทำการคำนวณค่า P_n กับประโยคที่เป็นผลเฉลยแล้วนั้นจะทำการนำค่าที่ได้มาเข้าสมการของ BLUE เพื่อวัดประสิทธิภาพสามารถคำนวณได้ดังต่อไปนี้

$$\log(BLEU-1) = min\left(1-\frac{7}{6},0\right)+(1\times\log(0.4285))$$
$$= (-0.166667)+(-0.36797678529)$$
$$= -0.53464345196$$
$$BLEU-1 = 0.30120$$

3.5.4 การวัดประสิทธิภาพคำตอบด้วย Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

การวัดประสิทธิภาพด้วย ROUGE เป็นการใช้พื้นฐานจากการประเมินด้วยค่า F1 ซึ่งคำนวณได้จากค่าความละลึก (Recall) และค่าความแม่นยำ (Precision) และทำการหาค่าเฉลี่ยของทั้งสองออกมา แต่ใน ROUGE อาศัยการคำนวณโดยอาศัยการเปรียบเทียบแบบที่มีลำดับที่ยากที่สุด (Longest Common Subsequence : LCS) ซึ่งในงานวิจัยนี้จะให้ความสนใจอยู่ที่ระดับโครงสร้างจะถูกเรียกว่า ROUGE-L สามารถคำนวณได้จากสมการดังต่อไปนี้

$$R_{lcs} = \frac{LCS(candidate, reference)}{length\ of\ cadidate\ term} \quad (29)$$

$$P_{lcs} = \frac{LCS(candidate, reference)}{length\ of\ reference\ term} \quad (30)$$

$$ROUGE-L = \frac{2\cdot R_{lcs}\cdot P_{lcs}}{R_{lcs}+P_{lcs}} \quad (31)$$

จากสมการข้างต้นกำหนดให้ใช้ประโยคตัวอย่างดังตารางที่ 3.18 โดยค่า LCS ในทั้งสองสมการ ที่ได้สามารถคำนวณได้ดังสมการดังต่อไปนี้

the	the	cat	the	the	cat	the
the	cat	it	on	the	mat	

รูปที่ 3.18 ตัวอย่างการนับความยาวของ LCS

จากรูปที่ 3.18 จะพบว่าความยาวต่อเนื่องที่ยาวที่สุดระหว่างประโยคที่คาดว่าจะเป็นคำตอบกับประโยคที่เป็นผลเฉลยของประโยคคำถามนั้น มีความยาวสูงสุดอยู่ที่ 2 เทอม ดังนั้นจึงนำค่าดังกล่าวไปใช้ในการวัดประสิทธิภาพประโยคคำตอบที่ได้จากกระบวนการถามตอบที่นำเสนอได้ดังต่อไปนี้

$$R_{lcs} = \frac{2}{7}$$

$$P_{lcs} = \frac{2}{6}$$

เมื่อได้ค่า R_{lcs} และ P_{lcs} ที่เป็นการประยุกต์ใช้ค่าความแม่นยำและค่าความระลึกเป็นที่เรียบร้อยแล้วนั้น ขั้นตอนถัดมาจะเป็นการนำค่าที่ได้ทั้งสองค่าไปทำการหาค่าเฉลี่ยโดยสามารถคำนวณได้ดังต่อไปนี้

$$ROUGE - L = \frac{2 \cdot \frac{2}{7} \cdot \frac{2}{6}}{\frac{2}{7} + \frac{2}{6}}$$

$$= \frac{8}{42} \times \frac{42}{26}$$

$$0.3076923$$

จากการแสดงวิธีการวัดประสิทธิภาพด้วย ROUGE-L นั้นจะพบว่าการวัดประสิทธิภาพนี้จะดูลำดับการเกิดของโครงสร้างประโยคที่เกิดขึ้นในประโยคคำตอบที่ได้จากกระบวนการถามตอบ เทียบกับประโยคคำตอบที่เป็นผลเฉลยของประโยคคำถาม

พหุน ปณ ฑิโต ชีเว

บทที่ 4
การวัดประสิทธิภาพ

ในบทนี้จะเป็นการแสดงขั้นตอนในการทดสอบกระบวนการถามตอบที่นำเสนอ โดยอธิบายถึงขั้นตอนแต่ละขั้นตอนในการทดสอบดังต่อไปนี้

4.1 ชุดข้อมูลที่ใช้ในการทดสอบ (Dataset)

ข้อมูลที่ใช้ในการทดสอบเป็นข้อมูลจากชุดมาตรฐาน MS MARCO เวอร์ชัน 2.1 ซึ่งเป็นข้อมูลชุดเดียวกันกับข้อมูลที่ใช้ในการสร้างรูปแบบคำถามและรูปแบบคำตอบ แต่เป็นกลุ่มข้อมูลที่แตกต่างกัน โดยชุดข้อมูลทดสอบสามารถแสดงได้ดังตารางที่ 4.1

ตารางที่ 4.1 สรุปข้อมูลที่ใช้ในการทดสอบ

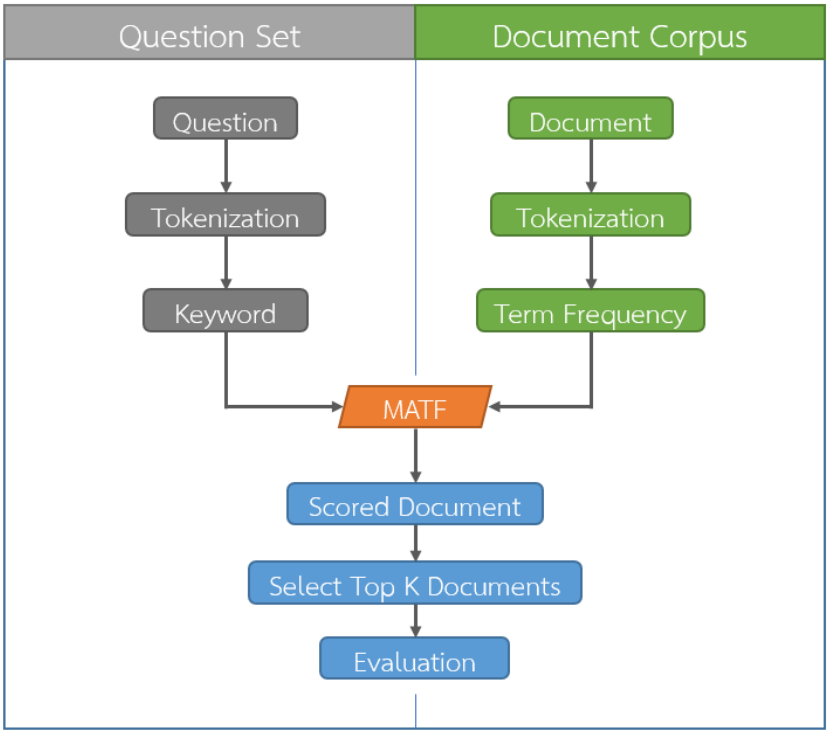
ประเภทคำถาม	จำนวนคำถาม	จำนวนเอกสารที่มีคำตอบ
why	100	1,000
how	100	1,000

จากตารางที่ 4.1 จะเห็นว่าคำถามที่ใช้ในทั้งสองประเภทของคำถาม มีจำนวนคำถามประเภทละ 100 คำถาม ที่ไม่ได้เป็นคำถามเดียวกันกับชุดคำถามที่ใช้ในการเรียนรู้รูปแบบคำถาม และในแต่ละคำถามประกอบไปด้วย 10 เอกสารที่มีคำตอบของคำถามดังกล่าว ทำให้เอกสารที่อยู่ในคลังข้อมูลมีจำนวนทั้งสิ้น 2,000 เอกสาร

4.2 ผลการทดสอบการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ (Document Candidate Retrieval Experimental results)

ในส่วนของผลการทดลองนี้จะแบ่งออกเป็น 3 ส่วนหลัก ได้แก่ ผลลัพธ์ของการเรียนรู้รูปแบบคำถามและคำตอบ ผลการทดลองการค้นคืนเอกสาร และส่วนของผลการทดลองการค้นคืนคำตอบ โดยในส่วนของการค้นคืนเอกสารนี้จะกล่าวถึงกระบวนการในการทดสอบและประเมินผลลัพธ์ที่ได้ จากนั้นจะนำผลลัพธ์ที่ได้จากขั้นตอนการค้นคืนเอกสารไปใช้ในการค้นคืนคำตอบ และทำการวัดประสิทธิภาพคำตอบที่ได้จากขั้นตอนนี้ การวัดประสิทธิภาพของทั้งสองขั้นตอนสามารถอธิบายได้ดังต่อไปนี้

ในส่วนของการค้นคืนเอกสารนั้น เอกสารที่ได้จากขั้นตอนนี้จะถือว่าเป็นเอกสารที่คาดว่าจะมีคำตอบของคำถามที่เข้ามา โดยกระบวนการในการวัดประสิทธิภาพสามารถอธิบายได้ดังรูปที่ 4.1



รูปที่ 4.1 กระบวนการการทดสอบการค้นคืนเอกสาร

จากรูปที่ 4.1 จะเห็นว่ามีขั้นตอนการทดสอบการค้นคืนเอกสารที่ประกอบไปด้วยการประมวลผลคำถามและนำไปค้นคืนเอกสารด้วย MATF ทีละเอกสารในคลังข้อมูล โดยจะทำการทดสอบกระบวนการถามตอบด้วยกัน สองรูปแบบประกอบไปด้วยการทดสอบกระบวนการถามตอบด้วยชุดข้อมูลที่ใช้ในการเรียนรู้ และชุดข้อมูลที่ใช้ในการทดสอบสามารถแสดงรายละเอียดได้ดังนี้

4.2.1 การประมวลผลคำถาม

ในขั้นตอนการประมวลผลคำถามนี้ จะเริ่มจากการนำชุดคำถามที่ใช้ในการทดสอบมาทีละ 1 คำถาม จากนั้นนำคำถามที่ได้ไปทำการตัดคำและตัดคำหยุด ฉะนั้นคำที่เหลือจะจัดว่าเป็นคำสำคัญของคำถามนั้น ๆ

เมื่อได้คำสำคัญจากคำถามแล้วนั้นจะเป็นการนำคำถามที่เข้ามาไปทำการระบุหน้าที่ของคำให้แต่ละคำในประโยคคำถาม และเลือกเอาเฉพาะส่วนที่เป็นการอธิบายถึงการกระทำในประโยค เพื่อระบุว่าคำถามที่ถามเข้ามานั้นอยู่ในรูปแบบใด เพื่อที่จะนำไปใช้ในการค้นรูปแบบของการตอบคำถามจากโมเดลที่ได้ทำการเรียนรู้ไว้ จากนั้นทำการจัดเก็บผลลัพธ์ที่ได้จากการประมวลผลคำถาม ในรูปแบบของ JSON ไฟล์ดังแสดงในรูปที่ 4.2

```
[
    {
        "question": "why does my computer user account need my last name",
        "keyword": [
            "computer",
            "user",
            "account",
            "need",
            "last",
            "name"
        ],
        "keyword_ex": [ ···
        ],
        "qid": "1067400"
    },
    {
        "question": "why did vespasian build the colosseum",
        "keyword": [
            "vespasian",
            "build",
            "colosseum"
        ],
        "keyword_ex": [
            "Amphitheatrum_Flavium",
            "flesh",
            "progress",
            "Titus_Flavius_Sabinus_Vespasianus",
            "physique",
            "soma",
            "form",
            "anatomy",
            "construct",
            "ramp_up",
            "human_body",
```


รูปที่ 4.2 ลักษณะการจัดเก็บผลลัพธ์ของการประมวลผลคำถาม

4.2.2 การประมวลผลเอกสารที่คาดว่าจะมีคำตอบ

ขั้นตอนการประมวลผลเอกสารที่มีคำตอบนี้จะเป็นการเตรียมข้อมูลในคลังเอกสารโดยการนำเอกสารที่เก็บรวบรวมไว้มาทำการตัดโครงสร้างของ HTML ออกจากนั้นทำการตัดคำ และตัดคำหยุดทีละเอกสาร เมื่อดำเนินการตามขั้นตอนข้างต้นแล้วให้ทำการนับคำที่อยู่ในเอกสารนั้น ๆ เพื่อระบุความถี่ของคำแต่ละคำที่อยู่ภายในเอกสาร นอกจากนี้ยังมีการนับความยาวของแต่ละเอกสารเพื่อใช้สำหรับการประมวลผลในขั้นตอนถัดไป สามารถแสดงได้ดังรูปภาพดังต่อไปนี้

```
{
    "doc_count": 50,
    "documents": [
        {
            "doc_file": "1074754_0_1.txt",
            "doc_length": 205,
            "doc_tf": {…
            }
        },
        {
            "doc_file": "1064406_0_8.txt",
            "doc_length": 232,
            "doc_tf": {
                "smell": 7,
                "good": 2,
                "top": 1,
                "legit": 1,
                "health": 2,
                "beauty": 1,
                "baking": 1,
                "alley": 1,
                "losing": 1,
                "weight": 1,
                "got": 2,
                "anatomy": 1,
                "season": 1,
                "autumn": 1,
                "equinox": 1,
                "science": 1,
```

รูปที่ 4.3 ลักษณะการจัดเก็บผลลัพธ์ของการประมวลผลเอกสาร

4.2.3 การให้น้ำหนักคำและค้นคืนเอกสาร

ในการค้นคืนเอกสารนั้นจะเป็นการเปรียบเทียบระหว่างคำสำคัญกับเอกสารทุกเอกสารภายในคลังข้อมูล เพื่อเปรียบเทียบว่าคำสำคัญที่ได้จากการประมวลผลคำถามนั้น แต่ละคำถามจะมีเอกสารใดที่มีค่าความคล้ายคลึงกันมากน้อยเพียงใด

ในการทดสอบนี้จะใช้ผลลัพธ์ที่ได้จากการประมวลผลคำถามและผลลัพธ์จากการประมวลผลเอกสาร โดยเริ่มจากการนำคำสำคัญจากคำถามที่ผ่านการประมวลมาให้คะแนนความคล้ายคลึงระหว่างคำสำคัญและเอกสารทุกเอกสารด้วย MATF

จากนั้นให้เรียงลำดับเอกสารที่มีค่าคะแนนสูงที่สุดออกเป็น 2 แบบ แบบที่หนึ่งคือการเรียงลำดับเอกสารที่มีคะแนนสูงที่สุดเพียงแค่อันดับแรก และการเรียงลำดับเอกสารที่มีคะแนนสูงที่สุด

10 อันดับแรก และทำการจัดเก็บผลลัพธ์ที่ได้จากขั้นตอนนี้ในรูปแบบ JSON ไฟล์ดังแสดงในรูปที่ 4.4 เพื่อนำไปวัดประสิทธิภาพด้วย P@K และ MRR ในขั้นตอนต่อไป

```
[
    {
        "question_id" : "1074754",
        "question": "why does my computer user account need my last name",
        "keyword": [
            {
                "document_id" : "1074754_0_1",
                "ranked_score" : 1.235423333
            },
            {
                "document_id" : "1074754_0_5",
                "ranked_score" : 1.114526766666666
            },
            {
                "document_id" : "80124_0_7",
                "ranked_score" : 0.72472177264
            },
            .....
        ]
    }
]
```

รูปที่ 4.4 ลักษณะการจัดเก็บผลลัพธ์ของการค้นคืนเอกสารที่คาดว่าจะมคำตอบ

4.2.4 การวัดประสิทธิภาพของการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ

ในส่วนของการวัดประสิทธิภาพการค้นคืนเอกสารที่คาดว่าจะมีคำตอบนั้น จะทำการวัดประสิทธิภาพด้วยกัน 2 วิธี โดยวิธีการแรกคือการดูว่าเอกสารที่คาดว่าจะมีคำตอบที่ได้จากขั้นตอนของการค้นคืนเอกสารนั้นถูกต้องหรือไม่จากจำนวนเอกสารที่เราจะให้ความสนใจทั้งหมด ในกรณีนี้จะเรียกว่าเป็นการวัดประสิทธิภาพด้วย P@K โดยผลลัพธ์ที่ได้จากการวัดประสิทธิภาพนั้นสามารถแสดงได้ดังตารางที่ 4.2

พหุน ปณ ฑิโต ชีเว

ตารางที่ 4.2 ผลลัพธ์วัดประสิทธิภาพในการค้นคืนเอกสารที่คาดว่าจะมีคำตอบประเภท How

How				
จำนวนคำถามละเทคนิคที่ใช้		P@1	P@5	P@10
30	TF	0.521	0.533	0.477
	TF-IDF	0.610	0.629	0.559
	MATF	**0.681**	**0.673**	**0.577**
	BM25	0.481	0.483	0.421
50	TF	0.572	0.553	0.496
	TF-IDF	0.656	0.664	0.577
	MATF	**0.691**	**0.693**	**0.591**
	BM25	0.491	0.494	0.427
100	TF	0.616	0.586	0.514
	TF-IDF	0.694	0.688	0.597
	MATF	**0.736**	**0.722**	**0.618**
	BM25	0.512	0.499	0.447

จากตารางที่ 4.2 และตารางที่ 4.3 จะพบว่าค่าที่ให้ความถูกต้องในการค้นคืนเอกสารที่สูงที่สุดคือการค้นคืนเอกสารด้วย MATF และนอกจากนี้ยังได้มีการทดสอบในจำนวนเอกสารที่เราสนใจที่มากขึ้น

จากนั้นวิธีที่สองที่ใช้คือการวิเคราะห์ว่าประสิทธิภาพในการจัดอันดับของกระบวนการเราดีหรือไม่ ในจำนวนเอกสารที่ค้นคืนได้จากจำนวนเอกสารที่เราให้ความสนใจ ในการวัดประสิทธิภาพกรณีนี้จะใช้การวัดประสิทธิภาพด้วยค่า MRR โดยการวัดประสิทธิภาพด้วยค่านี้จะใช้การวิเคราะห์กับแค่ผลลัพธ์ของค่า P@K ที่มีค่ามากกว่า 1 สามารถแสดงได้ดังตารางที่ 4.4

พหุน ปณ ฑิโต ชีเว

ตารางที่ 4.3 ผลลัพธ์วัดประสิทธิภาพในการค้นคืนเอกสารที่คาดว่าจะมีคำตอบประเภท Why

Why				
จำนวนคำถามละเทคนิคที่ใช้		P@1	P@5	P@10
30	TF	0.519	0.531	0.471
	TF-IDF	0.590	0.627	0.553
	MATF	**0.679**	**0.671**	**0.563**
	BM25	0.479	0.477	0.419
50	TF	0.568	0.551	0.494
	TF-IDF	0.644	0.660	0.569
	MATF	**0.689**	**0.687**	**0.587**
	BM25	0.489	0.490	0.423
100	TF	0.611	0.586	0.510
	TF-IDF	0.692	0.682	0.589
	MATF	**0.730**	**0.721**	**0.614**
	BM25	0.502	0.498	0.446

ตารางที่ 4.4 ผลลัพธ์วัดประสิทธิภาพด้วยค่า MRR

		How		Why	
จำนวนคำถามละเทคนิคที่ใช้		P@5	P@10	P@5	P@10
30	TF	0.614	0.629	0.613	0.624
	TF-IDF	0.679	0.691	0.676	0.691
	MATF	**0.732**	**0.744**	**0.731**	**0.741**
	BM25	0.566	0.579	0.563	0.577
50	TF	0.642	0.654	0.639	0.653
	TF-IDF	0.711	0.721	0.709	0.717
	MATF	**0.738**	**0.747**	**0.737**	**0.744**
	BM25	0.572	0.583	0.569	0.582
100	TF	0.681	0.692	0.678	0.689
	TF-IDF	0.752	0.759	0.750	0.755
	MATF	**0.776**	**0.781**	**0.775**	**0.780**
	BM25	0.587	0.599	0.584	0.595

จากผลลัพธ์ที่ได้ทำการทดสอบมาทั้งหมดข้างต้นจะเห็นว่า ค่า P@K ที่ได้จากกระบวนการค้นคืนเอกสารที่คาดว่าจะมีคำตอบนั้น เป็นค่าที่อยู่ในกลุ่มที่ยอมรับได้ ซึ่งเป็นที่แน่ชัดแล้วว่าถ้าหากทำการค้นคืนเอกสารที่ถูกต้องขึ้นมาได้มากเพียงใด นั่นหมายความว่าเรามีโอกาสที่จะได้คำตอบของคำถามที่เข้ามา มีความถูกต้องมากยิ่งขึ้น โดยอัลกอริทึมในการให้คะแนนเอกสารที่ให้ประสิทธิภาพดีที่สุดคือ MATF

นอกจากนี้ยังมีการวัดประสิทธิภาพในการจัดอันดับด้วยค่า MRR ซึ่งจะพิจารณาว่านอกจากจะสนใจความแม่นยำในการค้นคืนเอกสารใน *k* อันดับแรกแล้ว จะต้องพิจารณาด้วยว่าเอกสารที่คาดว่าจะมีคำตอบถูกค้นขึ้นมาในอันดับที่เท่าไหร่ในจำนวน *k* เอกสารที่ถูกค้นคืนขึ้น หากค่า MRR มีค่าเข้าใกล้ 1 แสดงว่าเอกสารที่ค้นคืนได้ใน *k* อันดับแรกยอมรับได้ [48, 52] ซึ่งในงานวิจัยที่เรานำเสนอ พบว่าค่า MRR โดยเฉลี่ยอยู่ที่ 0.67 สำหรับการวัดแบบ P@5 และ 0.68 สำหรับการวัดแบบ P@10 เมื่อนำผลลัพธ์ที่ได้ ไปเทียบกับงานวิจัยก่อนหน้าที่ผ่านมา [12, 48, 53] พบว่าผลลัพธ์จากงานวิจัยนี้ให้ค่า MRR ที่ยอมรับได้ เมื่องานวิจัยเหล่านั้นจะมีค่าเฉลี่ยของ MRR อยู่ระหว่าง 0.45 – 0.65

นอกจากนี้ หากพิจารณาอัลกอริทึมที่ใช้ในการให้น้ำหนักและการประเมินค่าความคล้ายคลึงของเอกสารที่คาดว่าจะมีคำตอบ พบว่า *MATF* ให้ประสิทธิภาพที่สูงกว่าการให้น้ำหนักและการประเมินค่าความคล้ายคลึงที่พิจารณาด้วย *tf*, *tf-idf*, และ *BM25* เนื่องจากเอกสารที่จัดเก็บรวบรวมไว้ในคลังข้อมูลเป็นเอกสารที่อยู่ในรูปแบบของข้อความเว็บหน้าเว็บไซต์ (Web Page Content) ทำให้ความยาวของเอกสารที่จัดเก็บมีความไม่แน่นอน อีกทั้งยังมีความแตกต่างกันเป็นอย่างมาก

ทำให้พบว่าอัลกอริทึม MATF นั้นให้ประสิทธิภาพที่ดีในการให้คะแนนความคล้ายคลึงของเอกสารที่มีขนาดความยาวที่แตกต่างกันมาก ต่างจาก BM25 ที่ให้ความสนใจกับเอกสารเพียงแค่ทางใดทางหนึ่งเท่านั้น โดยสังเกตุได้จากค่า *b* ของสมการ BM25 ที่เป็นการบ่งบอกว่าเรานั้นต้องการที่จะให้คะแนนของเอกสารที่มีความยาวเข้าใกล้ความยาวเฉลี่ยของเอกสารทั้งหมดในคลังข้อมูลมากน้อยเพียงใด และแตกต่างจากการให้คะแนนด้วย tf และ tf-idf ซึ่งไม่มีการนำค่าความยาวของเอกสารมาคำนวณร่วมแต่อย่างใด

4.3 ผลการทดสอบ (The Experimental Results of Answer Retrieval)

ในส่วนของการค้นคืนคำตอบที่ได้จากกระบวนการถามตอบนั้น สามารถแบ่งย่อยได้อีก 2 ขั้นตอนประกอบไปด้วยขั้นตอนการนำรูปแบบที่ทำการเรียนรู้มาค้นคืนบางส่วนของเอกสารที่คาดว่า

จะเป็นคำตอบ จากนั้นนำบางส่วนของเอกสารไปทำการจัดอันดับ เพื่อเลือกตัวที่ดีที่สุดมาเป็นคำตอบให้กับกระบวนการที่นำเสนอ โดยจะทำการแยกกระบวนการการทดสอบออกเป็นสองส่วน ประกอบไปด้วยการทดสอบด้วยการใช้ข้อมูลที่ใช้ในการเรียนรู้ และการทดสอบโดยใช้ชุดข้อมูลที่เตรียมไว้สำหรับทำการทดสอบ สามารถอธิบายการทดสอบได้ดังต่อไปนี้

4.3.1 การทดสอบการค้นคืนส่วนที่คาดว่าจะเป็นคำตอบด้วยชุดข้อมูลที่ใช้ในการเรียนรู้

การทดสอบการค้นคืนส่วนที่คาดว่าจะเป็นคำตอบด้วยชุดข้อมูลที่ใช้ในการเรียนรู้นั้น เพื่อเป็นการตรวจสอบว่ากระบวนการที่นำเสนอสามารถค้นคืนคำตอบได้จริง โดยมีการทดสอบใน 2 รูปแบบคือ

4.3.1.1 *การทดสอบรูปแบบคำตอบที่เรียนรู้ในการค้นคืน “ส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ” โดยใช้ชุดข้อมูลที่ใช้ในการเรียนรู้รูปแบบคำตอบ*

งานวิจัยนี้ได้ทดสอบนำรูปแบบคำตอบทั้งหมดไปค้นคืน “ส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ” จากเอกสารต้นฉบับที่ใช้ในการสร้างรูปแบบคำตอบเหล่านี้ เพื่อประเมินว่ารูปแบบเหล่านี้เป็นรูปแบบคำตอบสามารถใช้ในการค้นคืน “ส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ” ได้จริง ซึ่งทดสอบสามารถแสดงผลลัพธ์ได้ดังตารางที่ 4.5

ตารางที่ 4.5 แสดงผลลัพธ์ที่ได้จากการนำ “รูปแบบคำตอบ” ที่เรียนรู้ได้ไปค้นคืนคำตอบจากเอกสารต้นฉบับที่ใช้ในการเรียนรู้รูปแบบคำตอบ

ประเภทคำถาม	จำนวนรูปแบบคำตอบ	ค่าความระลึกเฉลี่ย
How	3,106	1.00
Why	3,049	1.00

จากตารางที่ 4.5 จะเห็นว่า รูปแบบคำตอบที่ได้จากการเรียนรู้ทั้งหมดสำหรับประโยคคำถาม “How” และ “Why” ได้ถูกทดสอบประมวลผลในการการหาคำตอบ เพื่อค้นคืน “ส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ” จากเอกสารต้นฉบับที่ใช้ในการสร้างรูปแบบคำตอบเอง ปรากฏว่าเมื่อวัดด้วยค่าความระลึก (Recall) พบว่าได้ประสิทธิภาพเฉลี่ยของทุกรูปแบบคำตอบอยู่ที่ 1.00

ด้วยเทคนิคที่ใช้ในการทดสอบชุดข้อมูลคำถามนี้ โดยทั่วไปในการเรียนรู้แบบมีผู้สอน จะหลีกเลี่ยงการสร้างแบบจำลองหรือการพัฒนากระบวนการที่มีจะมีการรู้จำชุดข้อมูลที่ใช้สอนมาก

จนเกินไป ซึ่งจะทำให้ประสิทธิภาพในการนำแบบจำลองที่พัฒนาขึ้นไปใช้กับชุดข้อมูลจริงนั้น มีประสิทธิภาพที่ต่ำลงเราจะเรียกเหตุการณ์ประเภทนี้ว่า Over Fitting

แนวคิดของการทดสอบเช่นนี้คล้ายกับแนวคิดในงานวิจัยนำแนวคิดดังกล่าวนี้ไปใช้ในการช่วยวิเคราะห์แบบจำลองที่ทำการพัฒนาขึ้นได้ว่า มีประสิทธิภาพเพียงพอ ก่อนที่จะนำไปใช้กับข้อมูลจริงหรือไม่ [54] เพื่อใช้ในการทดสอบประสิทธิภาพของแบบจำลองที่ทำการเรียนรู้โดยชุดข้อมูลชุดเดียวกัน นอกจากนี้ในบางเทคนิคหรือบางอัลกอริทึมที่ใช้ในการเรียนรู้ สามารถปรับค่าพารามิเตอร์เพื่อลดความเป็นอคติ (Bias) ของชุดข้อมูลได้ [55]

ทั้งนี้เทคนิคดังกล่าวยังใช้เพื่อเป็นการวิเคราะห์และตรวจสอบความผิดพลาด ของแบบจำลองที่เรียนรู้ขึ้นก่อนนำไปใช้งานจริง เพื่อลดเวลาในการดำเนินงานวิจัยกรณีที่เกิดข้อผิดพลาดได้ [56] นอกจากนี้ชุดข้อมูลทดสอบนี้ยังมีข้อควรระวังในการวิเคราะห์ เนื่องจากถ้าหากว่าให้ประสิทธิภาพที่ดีเกินไปอาจจะเป็นการบ่งบอกถึงแบบจำลองที่พัฒนาขึ้นนั้นมีโอกาสที่จะรู้จำชุดข้อมูลที่ใช้สอนมากเกินไป [56] โดยทั่วไปแล้วการใช้ชุดข้อมูลทดสอบนี้จะใช้เพื่อประเมินว่าแบบจำลองที่พัฒนาขึ้นนั้นสามารถนำไปใช้งานได้หรือไม่ โดยไม่ได้ให้ความสนใจอยู่ที่ประสิทธิภาพของแบบจำลองมากเท่าใด [56]

4.3.1.2 *การทดสอบการค้นคืนคำตอบด้วยชุดข้อมูลที่ใช้ในการเรียนรู้รูปแบบคำตอบ*

นอกจากนี้ ยังได้ทดสอบโดยการนำเอารูปแบบคำตอบที่ได้เทียบกลับเข้าไปยัง “ส่วนของเอกสารที่คาดว่าจะมีคำตอบ” แต่เพิ่มการใช้ “คำสำคัญ” ที่ได้จากคำถามเหล่านั้นร่วมด้วย โดยชุดข้อมูลที่ใช้ทดสอบมีจำนวน 200 คำถาม ประกอบไปด้วยคำถามประเภท “How” จำนวน 100 คำถาม และคำถามประเภท “Why” จำนวน 100 คำถาม ซึ่งคำตอบของคำถามเหล่านี้คือคำตอบที่ใช้ในการสร้างรูปแบบคำตอบนั่นเอง

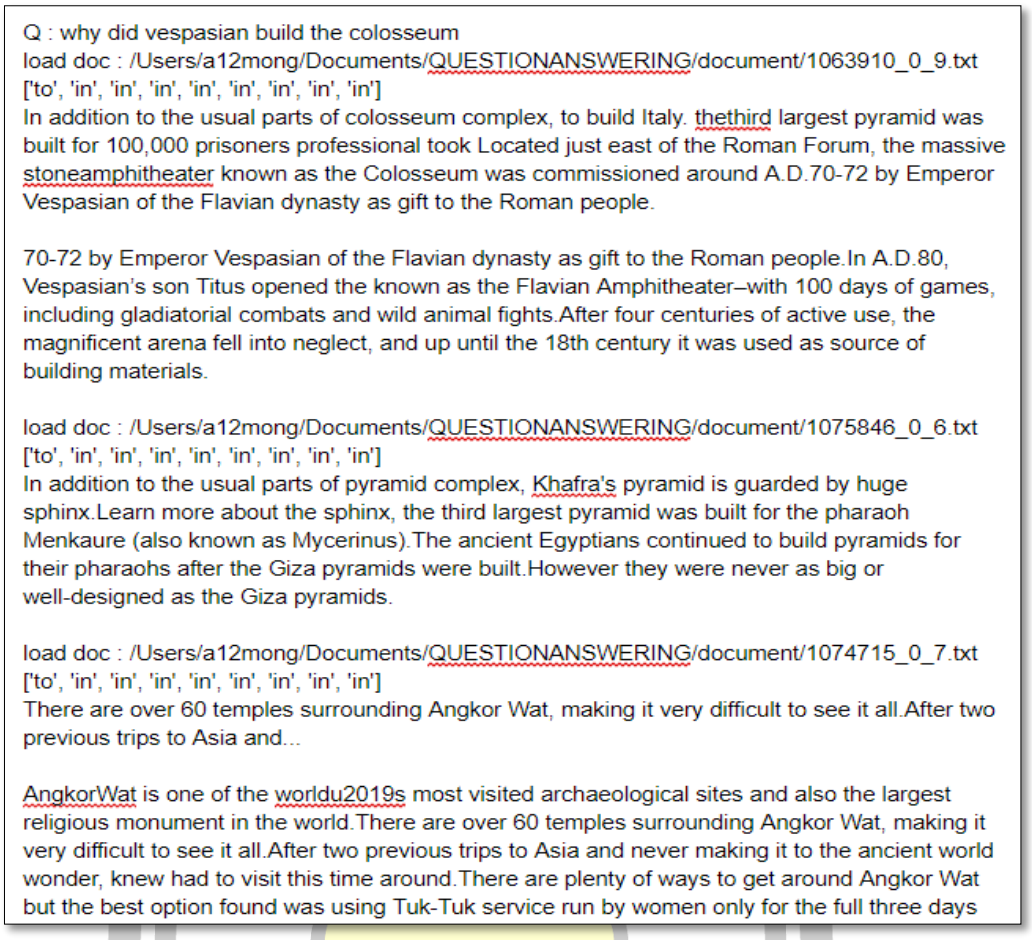
จากตารางที่ 4.6 จะเห็นว่าคำถามที่เข้ามาคือ “Why did Vespasian build the colosseum?” มีคำตอบของคำถามดังกล่าวอยู่ทั้งสิ้นจำนวน 10 คำตอบ แต่ในที่นี้จะนำตัวอย่างคำตอบมาแสดงเพียง 5 คำตอบจาก 10 คำตอบ ประกอบไปด้วยคำตอบ ดังแสดงในตารางที่ 4.6

จากนั้นจะทำการนำเอาคำถามดังกล่าวเข้าสู่กระบวนการถามตอบที่นำเสนอเพื่อดูผลลัพธ์ที่ได้จากกระบวนการถามตอบ โดยสามารถแสดงตัวอย่างได้ดังรูปที่ 4.7

ตารางที่ 4.6 ตัวอย่างคำตอบของคำถามที่จะทำการประมวลผล

คำตอบ	คำตอบ
1	The decision which led to building the Colosseum was made by the Vespasian who was emperor of Rome from 1 July 69 23 June 79 AD. His rule quickly followed the reign of the infamous Nero. During the rule of Nero the Great Fire of Rome of 64AD wrecked the city.
2	It was built in 70 AD and was completed in 82 AD Construction on the Colosseum began under Emperor Vespasian (69-79) and was completed under his son Titus (79-81) in 80 AD.
3	Construction of the Colosseum began under the rule of the Emperor Vespasian in around 70-72 AD. The Colosseum had been completed up to the third story by the time of Vespasian's death in 79.
4	To celebrate the end of Nero's rule, the Emperor Vespasian built the Colosseum on the site of Nero's lake. This was seen as giving back the land to the people of Rome. The Roman's often built monuments to celebrate important events, and the Colosseum is a part of that tradition.
5	70-72 by Emperor Vespasian of the Flavian dynasty as gift to the Roman people.In A.D.80, Vespasian’s son Titus opened the known as the Flavian Amphitheater–with 100 days of games, including gladiatorial combats and wild animal fights.After four centuries of active use, the magnificent arena fell into neglect, and up until the 18th century it was used as source of building materials.

จะเห็นว่าคำตอบที่ได้จากกระบวนการถามตอบจะมีคำตอบที่เป็นคำตอบที่แท้จริงของคำถามดังกล่าวรวมอยู่ด้วย และเมื่อทำการให้คะแนนความคล้ายคลึงคำตอบที่ได้และนำคำตอบที่มีค่าความคล้ายคลึงสูงที่สุดมาเป็นคำตอบของคำถามนั้นสามารถแสดงได้ดังรูปที่ 4.8 ซึ่งตรงกับคำตอบในลำดับที่ 5 จากตารางที่ 4.6


Q : why did vespasian build the colosseum
load doc : /Users/a12mong/Documents/QUESTIONANSWERING/document/1063910_0_9.txt
['to', 'in', 'in', 'in', 'in', 'in', 'in', 'in', 'in']
In addition to the usual parts of colosseum complex, to build Italy. thethird largest pyramid was built for 100,000 prisoners professional took Located just east of the Roman Forum, the massive stoneamphitheater known as the Colosseum was commissioned around A.D.70-72 by Emperor Vespasian of the Flavian dynasty as gift to the Roman people.

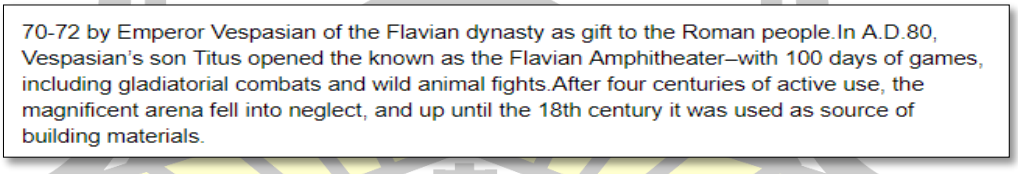
70-72 by Emperor Vespasian of the Flavian dynasty as gift to the Roman people.In A.D.80, Vespasian's son Titus opened the known as the Flavian Amphitheater–with 100 days of games, including gladiatorial combats and wild animal fights.After four centuries of active use, the magnificent arena fell into neglect, and up until the 18th century it was used as source of building materials.

load doc : /Users/a12mong/Documents/QUESTIONANSWERING/document/1075846_0_6.txt
['to', 'in', 'in', 'in', 'in', 'in', 'in', 'in', 'in']
In addition to the usual parts of pyramid complex, Khafra's pyramid is guarded by huge sphinx.Learn more about the sphinx, the third largest pyramid was built for the pharaoh Menkaure (also known as Mycerinus).The ancient Egyptians continued to build pyramids for their pharaohs after the Giza pyramids were built.However they were never as big or well-designed as the Giza pyramids.

load doc : /Users/a12mong/Documents/QUESTIONANSWERING/document/1074715_0_7.txt
['to', 'in', 'in', 'in', 'in', 'in', 'in', 'in', 'in']
There are over 60 temples surrounding Angkor Wat, making it very difficult to see it all.After two previous trips to Asia and...

AngkorWat is one of the worldu2019s most visited archaeological sites and also the largest religious monument in the world.There are over 60 temples surrounding Angkor Wat, making it very difficult to see it all.After two previous trips to Asia and never making it to the ancient world wonder, knew had to visit this time around.There are plenty of ways to get around Angkor Wat but the best option found was using Tuk-Tuk service run by women only for the full three days

รูปที่ 4.5 คำตอบที่ได้จากกระบวนการถามตอบ ในชุดข้อมูลชุดทดสอบ


70-72 by Emperor Vespasian of the Flavian dynasty as gift to the Roman people.In A.D.80, Vespasian's son Titus opened the known as the Flavian Amphitheater–with 100 days of games, including gladiatorial combats and wild animal fights.After four centuries of active use, the magnificent arena fell into neglect, and up until the 18th century it was used as source of building materials.

รูปที่ 4.6 แสดงคำตอบที่ได้จากชุดข้อมูลชุดทดสอบ

จากรูปที่ 4.5 จะเห็นว่ากระบวนการถามตอบที่นำเสนอนั้นสามารถค้นคืนคำตอบด้วยรูปแบบคำตอบที่สอดคล้องการคำถามที่ป้อนเข้ามาได้จริง โดยชุดข้อมูลชุดทดสอบที่มีคำถาม 200 คำถามนั้น เมื่อทดสอบด้วยค่าความระลึก สามารถแสดงผลลัพธ์ ได้ดังตารางที่ 4.7 และตารางที่ 4.8

ตารางที่ 4.7 ผลการค้นคืนคำตอบด้วยรูปแบบคำถาม “How” ด้วยรูปแบบคำถามที่เรียนรู้ได้ ร่วมกับคำสำคัญในคำถาม ด้วยชุดข้อมูลที่เป็นข้อมูลที่ใช้ในการเรียนรู้รูปแบบคำตอบ

How	
จำนวนคำถาม	ค่าความระลึกเฉลี่ย
100	1.00

ตารางที่ 4.8 ผลการค้นคืนคำตอบจากคำถามประเภท “Why” ด้วยรูปแบบคำถามที่เรียนรู้ได้ ร่วมกับคำสำคัญในคำถาม ด้วยชุดข้อมูลที่เป็นข้อมูลที่ใช้ในการเรียนรู้รูปแบบคำตอบ

Why	
จำนวนคำถาม	ค่าความระลึกเฉลี่ย
100	1.00

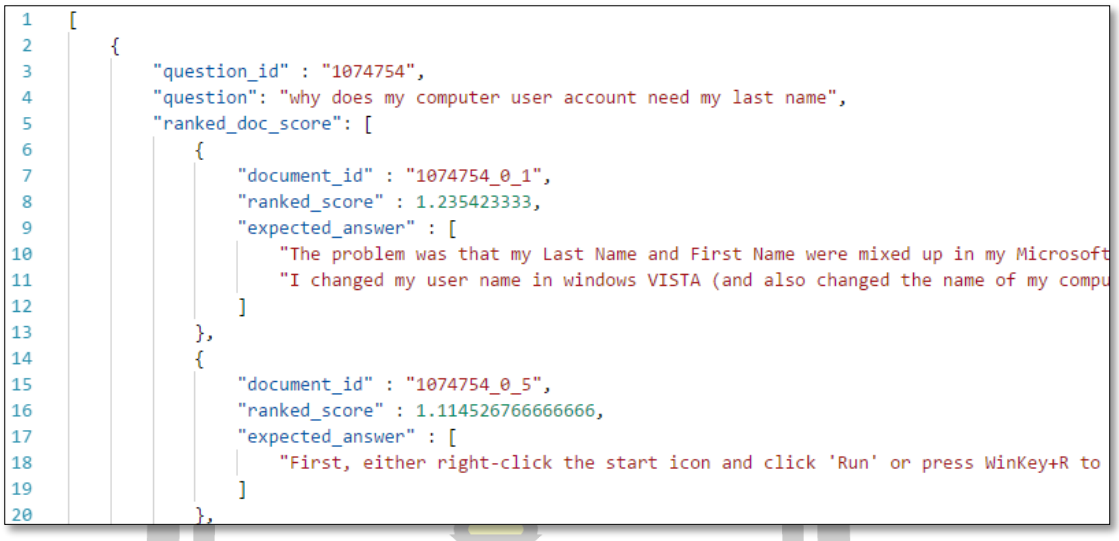
จากตารางที่ 4.7 และตารางที่ 4.8 จะเห็นว่าการทดสอบการค้นคืนคำตอบจากคำถามประเภท “Why” และ “How” ด้วยรูปแบบคำถามที่เรียนรู้ได้ ร่วมกับคำสำคัญในคำถาม ด้วยชุดข้อมูลที่เป็นข้อมูลที่ใช้ในการเรียนรู้รูปแบบคำตอบนั้น ให้ค่าความระลึกเฉลี่ยเท่ากับ 1.0

4.3.2 การทดสอบการค้นคืนส่วนที่คาดว่าจะเป็นคำตอบด้วยชุดข้อมูลเตรียมไว้ในการทดสอบ

การทดสอบการค้นคืนส่วนที่คาดว่าจะเป็นคำตอบด้วยชุดข้อมูลเตรียมไว้ในการทดสอบนั้น จะดำเนินการทดสอบโดยอาศัยกระบวนการถามตอบเป็นหลักในการดำเนินการทดสอบ เพื่อให้ทราบถึงผลลัพธ์ที่เกิดขึ้นในแต่ละขั้นตอนของกระบวนการ สามารถแสดงได้ดังต่อไปนี้

4.3.2.1 *การค้นคืนส่วนที่คาดว่าจะเป็นคำตอบ*

ในขั้นตอนการค้นคืนส่วนที่คาดว่าจะเป็นคำตอบนี้จะใช้ผลลัพธ์ที่ได้จากขั้นตอนก่อนหน้าซึ่งก็คือเอกสารที่ได้รับการจัดอันดับจากกระบวนการ การค้นคืนเอกสารเป็นที่เรียบร้อยแล้ว จากนั้นให้ใช้รูปแบบของคำตอบที่อยู่ภายใต้คำถามที่ได้จากการเรียนรู้ มาใช้ในการค้นคืนส่วนที่คาดว่าจะเป็นคำตอบจากรายการเอกสารที่ถูกจัดอันดับ จากนั้นจัดเก็บเอกสารให้อยู่ในรูปของ JSON ไฟล์ ดังแสดงในรูปที่ 4.7

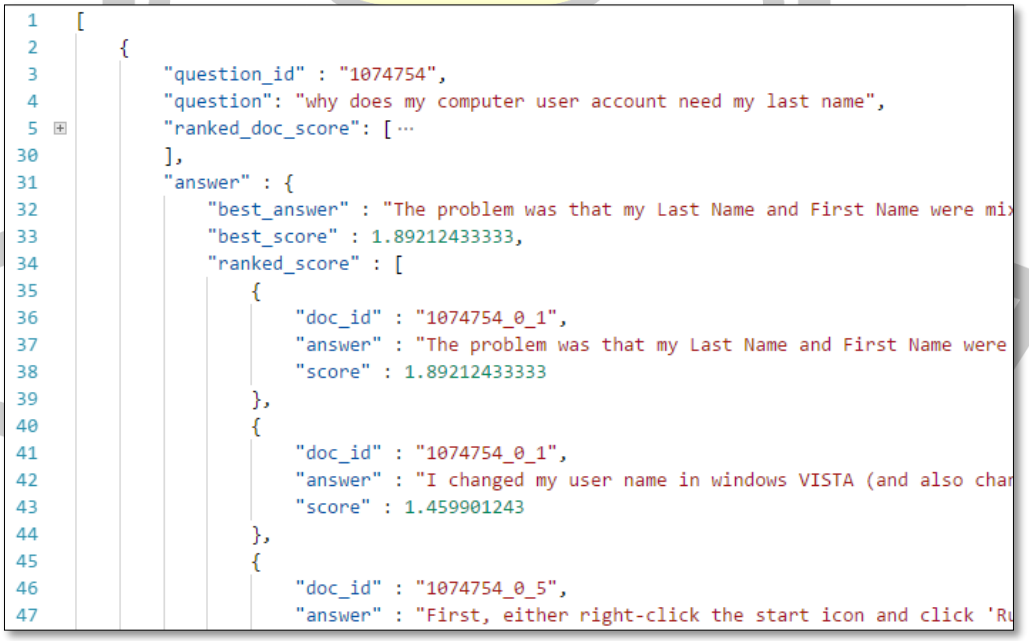


```
[
    {
        "question_id" : "1074754",
        "question": "why does my computer user account need my last name",
        "ranked_doc_score": [
            {
                "document_id" : "1074754_0_1",
                "ranked_score" : 1.235423333,
                "expected_answer" : [
                    "The problem was that my Last Name and First Name were mixed up in my Microsoft
                    "I changed my user name in windows VISTA (and also changed the name of my compu
                ]
            },
            {
                "document_id" : "1074754_0_5",
                "ranked_score" : 1.114526766666666,
                "expected_answer" : [
                    "First, either right-click the start icon and click 'Run' or press WinKey+R to
                ]
            },
```

รูปที่ 4.7 ส่วนของเอกสารที่คาดว่าจะเป็นคำตอบ

4.3.2.2 *การจัดอันดับส่วนที่คาดว่าจะเป็นคำตอบ*

เมื่อเราได้ส่วนที่คาดว่าจะเป็นคำตอบจากเอกสารที่ค้นคืนเป็นที่เรียบร้อยแล้ว ขั้นตอนถัดมาจะเป็นการนำเอาส่วนที่คาดว่าจะเป็นคำตอบนั้น มาทำการจัดอันดับ โดยวัดค่าความคล้ายคลึงระหว่างคำค้นที่ได้จากคำถาม และคำที่อยู่ในส่วนที่คาดว่าจะเป็นคำตอบด้วย MATF จากนั้นจะทำการเลือกคำตอบที่มีคะแนนสูงที่สุดเป็นคำตอบของคำถามที่เข้ามา



```
[
    {
        "question_id" : "1074754",
        "question": "why does my computer user account need my last name",
        "ranked_doc_score": [ ···
        ],
        "answer" : {
            "best_answer" : "The problem was that my Last Name and First Name were mix
            "best_score" : 1.89212433333,
            "ranked_score" : [
                {
                    "doc_id" : "1074754_0_1",
                    "answer" : "The problem was that my Last Name and First Name were
                    "score" : 1.89212433333
                },
                {
                    "doc_id" : "1074754_0_1",
                    "answer" : "I changed my user name in windows VISTA (and also char
                    "score" : 1.459901243
                },
                {
                    "doc_id" : "1074754_0_5",
                    "answer" : "First, either right-click the start icon and click 'Ru
```

รูปที่ 4.8 ผลลัพธ์ที่ได้จากการจัดอันดับคำตอบด้วย MATF

4.3.2.3 *การวัดประสิทธิภาพคำตอบที่ได้จากกระบวนการ*

ในขั้นตอนสุดท้ายจะเป็นการวัดประสิทธิภาพของกระบวนการถามตอบที่นำเสนอ สามารถวัดได้จากคำตอบที่ได้จากกระบวนการถามตอบเทียบกับคำตอบที่เป็นผลเฉลยในชุดข้อมูล โดยจะวัดประสิทธิภาพด้วย BLEU-1 และ ROUGE-L

การวัดประสิทธิภาพทั้งสองแบบนั้นเป็นการวัดประสิทธิภาพมาตรฐานที่ใช้กันในชุดข้อมูลของ MS MARCO เวอร์ชัน 2.1 โดยจะทำการทดสอบกับชุดคำถามทีละชุดแบ่งออกเป็น 30, 50 และ 100 ชุดคำถามในแต่ละประเภทคำถามตามลำดับ เพื่อที่จะดูว่าในกระบวนการถามตอบที่นำเสนอนั้นหากมีข้อมูลเพิ่มขึ้นกระบวนการถามตอบที่นำเสนอ จะมีทิศทางในการประเมินไปในทางใด

ตารางที่ 4.9 ผลลัพธ์วัดประสิทธิภาพคำตอบที่ได้จากกระบวนการ

How		
จำนวนชุดคำถาม	BLUE-1	ROUGE-L
30	0.486	0.462
50	0.449	0.436
100	0.436	0.417

ตารางที่ 4.10 ผลลัพธ์วัดประสิทธิภาพคำตอบที่ได้จากกระบวนการ

Why		
จำนวนชุดคำถาม	BLUE-1	ROUGE-L
30	0.4808	0.4604
50	0.4474	0.4302
100	0.4322	0.4054

จากผลลัพธ์ที่ได้จากตารางที่ 4.9 และตารางที่ 4.10 ที่แสดงถึงการวัดประสิทธิภาพคำตอบที่ได้จากกระบวนการถามตอบจะเห็นว่าค่าที่ได้จากตัววัดประสิทธิภาพทั้ง BLEU-1 และ ROUGE-L นั้นบ่งบอกว่าคำถามประเภท “*How*” นั้นมีประสิทธิภาพที่ดีกว่าคำถามประเภท “*Why*” เพียงเล็กน้อย อันเนื่องมาจากว่าคำตอบที่ใช้ในการทดสอบของชุดข้อมูล MS MARCO เวอร์ชัน 2.1 นั้นมีลักษณะของคำตอบเป็นการอธิบายทำให้ประโยคมีความยาวใกล้เคียงกัน อีกทั้งยังเป็น “ส่วนของเอกสาร” ที่อยู่ในลักษณะของการอธิบาย ทำให้โครงสร้างประโยคที่ออกมานั้นไม่ต่างกันมาก

ภายหลังจากการทดสอบ พบว่าคำถามประเภท “*How*” ให้ประสิทธิภาพดีกว่าคำถามประเภท “*Why*” เนื่องจากจำนวนรูปแบบคำตอบสำหรับคำถามประเภท How มีมากกว่าจำนวนรูปแบบคำตอบสำหรับคำถามประเภท “*Why*” ดังนั้น เมื่อมีการใช้รูปแบบคำตอบสำหรับคำถามประเภท “*How*” ในการค้นคืน “ส่วนของเอกสาร” ที่คาดว่าจะมีคำตอบ จึงสามารถค้นคืนได้ครอบคลุมและดีกว่ารูปแบบคำตอบสำหรับคำถามประเภท “*Why*” อย่างไรก็ตาม แม้ผลลัพธ์ที่ได้สำหรับคำถามประเภท “*How*” จะดีกว่าผลลัพธ์ที่ได้จากคำถามประเภท “*How*” แต่ก็ไม่แตกต่างกันมากนัก

ตารางที่ 4.11 ผลลัพธ์วัดประสิทธิภาพคำตอบที่ได้จากกระบวนการ

How และ Why		
จำนวนชุดคำถาม	Average BLUE-1	Average ROUGE-L
60	0.4834	0.4612
100	0.4482	0.4331
200	0.4341	0.4112

โดยในชุดข้อมูล MS MARCO เวอร์ชัน 2.1 นั้นมีการตั้งตัวอย่างของการวัดประสิทธิภาพด้วย BLEU-1 และ ROUGE-L อยู่ที่ 0.129 และ 0.268 ตามลำดับ [58]

เมื่อพิจาณาจากผลลัพธ์ที่ได้จากกระบวนการถามตอบที่นำเสนอนั้น จะพบว่าผลลัพธ์ที่ได้นั้นอยู่ในช่วงคะแนนที่สูงกว่ามาตรฐาน (Base Line) ของชุดข้อมูลมาตรฐาน MS MARCO

พหุน ปณฺฑิโต ชีเว

บทที่ 5
สรุปผลการทดลอง

ในบทนี้จะเป็นการสรุปการดำเนินงานวิจัยในเรื่อง ตลอดปัญหาและอุปสรรค และงานวิจัยที่น่าจะสามารถดำเนินต่อไปได้ในอนาคต

5.1 สรุปการดำเนินงานวิจัย

จากการศึกษาพบว่า ในการประยุกต์ใช้งานของกระบวนการถาม-ตอบ เกือบจะทั้งหมดจะอยู่ในรูปแบบของกระบวนการถาม-ตอบแบบแฟคทอยด์ คือ การถาม-ตอบที่ต้องการคำตอบแบบสั้นๆ ด้วยการอาศัยกฎ (rule) หรือแม่แบบ (template) ที่มีการเรียนรู้เอาไว้ [3] ในการค้นหาและสกัดคำตอบ แต่ในคำถามบางประเภท ได้แก่ “ทำไม (why)” และ “อย่างไร (how)” คำตอบสั้นๆ อาจจะไม่เพียงพอที่จะเป็นคำตอบ เพราะเป็นลักษณะคำถามที่ต้องการเหตุผล (Reason) หรือคำอธิบาย (Explanation) หรือการขยายความ (Expansion) ดังนั้น ในปัจจุบันการถาม-ตอบที่เป็นแบบนอนแฟคทอยด์ยังคงเป็นงานวิจัยที่มีความยากทั้งในเรื่องของการเลือกเอกสาร การค้นหาคำตอบ รวมไปถึงการประมวลผลคำตอบ

จากสาเหตุข้างต้น ในงานวิจัยฉบับนี้จึงนำเสนอการศึกษาและวิจัยเพื่อนำเสนอกระบวนการในการวิจัย เพื่อการหาคำตอบในการถาม-ตอบแบบนอนแฟคทอยด์สำหรับภาษาอังกฤษ โดยการอาศัยเทคนิคหลักๆ ในการประมวลผลคำถามและคำตอบ คือ การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) การค้นคืนเอกสาร (Information Retrieval: IR) และการสกัดสาร-สนเทศ (Information Extraction: IE) งานวิจัยนี้นำเสนอการค้นคำตอบของการถาม-ตอบแบบนอนแฟคทอยด์

ซึ่งงานวิจัยฉบับนี้นำเสนอกระบวนการในการค้นหาคำตอบสำหรับการถาม-ตอบแบบนอนแฟคทอยด์ โดยกระบวนการที่นำเสนอมีการทำงานใน 2 ส่วนหลักๆ คือ

(1) ส่วนแรกจะเป็นส่วนของการสร้างรูปแบบคำถาม (Question pattern) และรูปแบบคำตอบ (Answer pattern) ด้วยแบบจำลองมาร์คอฟ (Markov Model: MM) โดยในการสร้างรูปแบบคำถามและรูปแบบคำตอบนั้นจะเริ่มจากการสร้างรูปแบบคำถาม จากนั้นจึงทำการสร้างรูปแบบคำตอบ

ในการสร้างรูปแบบคำถาม คำถามประเภท “*How*” และ “*Why*” จากชุดข้อมูลที่ใช้ในการเรียนรู้ (Training set) จะเริ่มจากการติดมาสก์ (mask) ในตำแหน่งของคำว่า “*How*” และ “*Why*”

จากนั้น ส่วนที่เหลือของประโยคคำถามจะประมวลผลด้วยการทำ POS tagging จากนั้นจะได้ชุดข้อมูลในลักษณะของลำดับของ tag ที่บ่งบอกประเภทของคำ ซึ่งชุดข้อมูลดังกล่าวจะถูกนำไปสร้างรูปแบบคำถามด้วยแบบจำลองมาร์คอฟ ต่อไป ซึ่งในงานวิจัยนี้ใช้ข้อมูลคำถาม “*Why*” และ “*How*” อย่างละ 400 ประโยค และเมื่อทำการสร้างรูปแบบคำถามด้วยแบบจำลองมาร์คอฟ จะได้รูปแบบคำถามสำหรับคำถามประเภท “*Why*” จำนวน 42 รูปแบบ และจะได้รูปแบบคำถามสำหรับคำถามประเภท “*How*” จำนวน 54 รูปแบบ

เมื่อได้รูปแบบคำถามแล้ว รูปแบบคำถามที่ได้จะนำไปค้นคืนว่าลักษณะคำถามดังกล่าว มีคำตอบใดที่เป็นไปได้บ้าง ซึ่งข้อมูลคำตอบที่ถูกค้นขึ้นมานี้จะเป็นชุดข้อมูลสอนสำหรับการสร้างรูปแบบคำตอบ ชุดข้อมูลดังกล่าวจะเริ่มจากการทำ POS tagging เช่นกัน จากนั้นจึงสกัดเอาเฉพาะคำหยุด (Stop word) ที่อยู่ในกลุ่มของคำเชื่อม จากนั้น ก็จะได้ชุดข้อมูลที่ใช้ในการเรียนรู้ ที่อยู่ในรูปแบบลำดับของแทกซ์ระบุประเภทของคำที่อยู่ในกลุ่มคำหยุดออกมา โดยกลุ่มคำหยุดที่สนใจในงานวิจัยนี้คือ Conjunction, TO, และ Preposition จากนั้นชุดข้อมูลดังกล่าวจะถูกนำไปสร้างรูปแบบคำตอบด้วยแบบจำลองมาร์คอฟ ซึ่งจะได้รูปแบบคำตอบสำหรับคำถามประเภท “*Why*” จำนวน 1,680 รูปแบบ และจะได้รูปแบบคำตอบสำหรับคำถามประเภท “*How*” จำนวน 2,160 รูปแบบ

ส่วนที่สองนี้จะเป็นส่วนของการประมวลผลการถาม-ตอบ ซึ่งประกอบด้วยการประมวลผลคำถาม การค้นคืนเอกสารที่คาดว่าจะมีคำตอบ และการสกัดส่วนของเอกสารที่คาดว่าจะเป็นคำตอบของคำถามนั้น ๆ

ในส่วนของการประมวลผลคำถามนั้น มี 2 เป้าหมาย นั่นคือ (1) เป็นการประมวลผลเพื่อให้ได้คำสำคัญ เพื่อใช้เป็นคำค้นหรือคำขอ (Query) ในการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ และการส่วนของเอกสารที่คาดว่าจะเป็นคำตอบของคำถามนั้น ๆ และเป้าหมายที่ (2) คือการประมวลผลคำถามเพื่อตรวจสอบหารูปแบบคำถาม เมื่อได้รูปแบบคำถาม ก็จะทำการตรวจสอบต่อไปว่ารูปแบบคำถามนั้น ๆ มีรูปแบบคำตอบแบบใดได้บ้าง

เมื่อได้คำสำคัญจากการประมวลผลคำถาม คำสำคัญนั้นๆ จะถูกนำมาใช้ในการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ ซึ่งเทคนิคที่ใช้ในการศึกษาในงานวิจัยนี้คือ BM25 และ MATF โดยเป็นการพิจารณาเอกสารที่มีความคล้ายคลึงหรือสอดคล้องกับคำสำคัญที่ได้จากคำถามใน 10 อันดับแรก และเอกสารเหล่านี้จะถูกใช้ในขั้นตอนการค้นหาคำตอบ

เมื่อได้เอกสารที่คาดว่าจะมีคำตอบแล้วก็จะเข้าสู่กระบวนการของการสกัดส่วนของเอกสารที่จะเป็นคำตอบ โดยเริ่มจากการทำ POS tagging จากนั้นเอารูปแบบคำตอบที่ได้มาจากขั้นตอนของการประมวลผลคำถามมาเปรียบเทียบเพื่อหาส่วนของเอกสารที่สอดคล้องกับรูปแบบคำตอบ ซึ่งส่วนของเอกสารเหล่านี้จะเรียกว่า passage จากนั้นจะนำ passage ที่ได้เหล่านี้มาตรวจสอบว่า passage ใด ควรจะได้รับการพิจารณาเป็นคำตอบของคำถามนั้น ๆ โดยเป็นการพิจารณา passage ที่มีค่าความคล้ายคลึงสูงที่สุดเพียงอันดับแรก สำหรับการพิจารณาความคล้ายคลึงระหว่างคำถามกับคำตอบที่ได้นั้น จะเป็นการใช้ประโยชน์จากคำสำคัญที่ได้จากคำถามมาเปรียบเทียบกับคำที่พบใน passage ที่ค้นคืนได้ ด้วยเทคนิคการวัดความคล้ายคลึงแบบ BM25 และ MATF

สำหรับชุดข้อมูลที่ใช้ในการศึกษาคือ ชุดมาตรฐาน MS MARCO เวอร์ชัน 2.1 ซึ่งถูกนำเสนอโดย Microsoft ในปี 2018

ในการประเมินผลลัพธ์ที่ได้ในงานวิจัยฉบับนี้ ได้ทำการประเมินใน 2 ส่วนคือ (1) การประเมินผลในส่วนของการค้นคืนเอกสารที่คาดว่าจะมีคำตอบ และ (2) การประเมินการค้นและแสดงคำตอบของคำถามนั้น ๆ

ในการประเมินส่วนแรก คือ การค้นคืนเอกสารที่คาดว่าจะมีคำตอบ จะมีการประเมินด้วย 2 เทคนิค คือ (1) การประเมินด้วย P@K ซึ่งเป็นการวัดค่าความแม่นยำในการค้นคืนเอกสาร จากการค้นคืน *k* เอกสาร และ (2) การประเมินด้วยค่า MRR ซึ่งเป็นการประเมินประสิทธิภาพในการจัดอันดับ โดยทั้งค่า P@K และ MRR ล้วนเป็นการประเมินตามมาตรฐานของการถาม-ตอบ และในงานวิจัยนี้พบว่าค่าเฉลี่ย P@10 และ MRR อยู่ที่ 0.591 และ 0.756 ตามลำดับ ซึ่งเมื่อทำการเปรียบเทียบกับงานวิจัยที่ผ่านมา [48, 52] พบว่าผลลัพธ์ดังกล่าวเป็นผลลัพธ์ที่ยอมรับได้ในงานวิจัยด้านการถาม-ตอบ [12, 48, 53]

สำหรับการประเมินการสกัดคำตอบจะเป็นการประเมินด้วยค่า BLEU-1 และ ROUGE-L ซึ่งเป็นการประเมินตามมาตรฐานของการถาม-ตอบเช่นกัน และการวัดด้วยค่าดังกล่าวยังเป็นการวัดค่ามาตรฐานสำหรับชุดข้อมูล MS MARCO เวอร์ชัน 2.1 ได้กำหนดไว้ ภายหลังจากการทดสอบพบว่าค่า BLEU-1 และ ROUGE-L ในงานวิจัยฉบับนี้มีค่าเท่ากับ 0.43 และ 0.41 ตามลำดับ และจากการศึกษาที่ผ่านมาพบว่าผลลัพธ์ดังกล่าวเป็นผลลัพธ์ที่ยอมรับได้ในงานวิจัยด้านการถาม-ตอบแบบนอนแฟคทอยด์ [57, 59, 60]

5.2 ปัญหาและอุปสรรค

ชุดข้อมูลที่ใช้คือ MS MARCO เวอร์ชัน 2.1 เป็นชุดข้อมูลที่นำเสนอโดย Microsoft โดยเวอร์ชันล่าสุดถูกนำเสนอในปี 2018 แม้ชุดข้อมูลดังกล่าวจะเป็นชุดข้อมูลมาตรฐานที่มีที่ข้อมูลคำถาม คำตอบของแต่ละคำถาม และเอกสารที่มีตอบ แต่เอกสารที่มีคำตอบอยู่ส่วนใหญ่จะเป็นลิงค์ของเว็บ ซึ่งพบว่าหลายๆ ลิงค์นั้นไม่ปรากฏอยู่บนอินเตอร์เน็ตแล้ว ดังนั้นข้อมูลที่ใช้ในการศึกษาจึงมีไม่มากนัก สำหรับในงานวิจัยนี้คำถามที่ขึ้นต้นด้วย “*Why*” และ “*How*” ที่เอกสารที่มีคำตอบอย่างน้อย 10 เอกสารและเหมาะต่อการศึกษาเพียงอย่างละ 400 คำถาม ส่วนคำถามที่เหลือเป็นคำถามที่มีคำตอบแต่ปราศจากเอกสารที่ใช้ในการอ้างอิงแหล่งที่มาของคำตอบ

5.3 งานวิจัยในอนาคต

จากผลการวิจัยและการศึกษาในงานวิจัยก่อนหน้าพบว่า การค้นคำตอบของการถาม-ตอบแบบนอนแฟคทอยด์ยังไม่ผลลัพธ์ที่ควรปรับปรุง แม้ว่าจะเป็นการศึกษาในโดเมนปิด (Closed domain)

พหุน ปณฺฑิโต ชีเว

บรรณานุกรม

บรรณานุกรม

[1] Andrenucci A, Sneiders E. Automated question answering: Review of the main approaches. Proc - 3rd Int Conf Inf Technol Appl ICITA 2005. 2005;I:514–9.

[2] Pundge AM. Question Answering System , Approaches and Techniques : A Review. Int J Comput Appl. 2016;141:34–9.

[3] Cuteri B. Closed domain question answering for cultural heritage. CEUR Workshop Proc. 2016;1769:17–22.

[4] Fukumoto J. Question Answering System for Non-factoid Type Questions and Automatic Evaluation based on BE Method. Proc NTCIR-6 Work Meet. 2007;441–7.

[5] Surdeanu M, Ciaramita M, Zaragoza H. Learning to Rank Answers to Non-Factoid Questions from Web Collections. Comput Linguist. 2011 Jun;37:351–83.

[6] Jayalakshmi S, Sheshasaayee A. Automated question answering system using ontology and semantic role. In: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE; 2017. p. 528–32.

[7] Clark P, Thompson J, Porter B. A knowledge-based approach to question-answering. Proc AAAI. 1999;1–16.

[8] Polpinij J. Ontology-based knowledge discovery from unstructured and semi-structured text. 2014;

[9] Tellex S. Pauchok: A Modular Framework for Question Answering. In: Master Thesis Submitted to the Department of Electrical Engineering and computer science, Maccachusetts institute of Technology. 2003.

[10] Hovy E, Gerber L, Hermjakob U, Junk M, Lin C. Question Answering in Webclopedia. Proc Ninth Text Retr Conf. 2000;655–64.

[11] Hermjakob U, Echihabi A. Natural language based reformulation resource and web exploitation for question answering. Information Science and Systems Engineering The University of Tokushima, Tokushima, Japan. 2002.

[12] Agichtein E, Lawrence S, Gravano L. Learning search engine specific query transformations for question answering. Proc tenth Int Conf World Wide Web - WWW '01. 2001;169–78.

[13] Vechtomova O, Wang Y. A study of the effect of term proximity on query expansion. J Inf Sci. 2006;32:324–33.

[14] Buck C, Bulian J, Ciaramita M, Gajewski W, Gesmundo A, Houlsby N, et al. Ask the Right Questions: Active Question Reformulation with Reinforcement Learning. 2017;

[15] Hu H. A Study on Question Answering System Using Integrated Retrieval Method. Phd Thesis Submitted to Graduate school of engineering at the University of Tokushima; 2006.

[16] Hovy E, Hermjakob U, Lin C, Rey M. The Use of External Knowledge in Factoid QA. Proc TREC-10 Conf NIST. 2001;cited 150.

[17] Woods WA. Progress in natural language understanding. In: Proceedings of the June 4-8, 1973, national computer conference and exposition on - AFIPS '73. New York, New York, USA: ACM Press; 1973. p. 441.

[18] Dwivedi SK, Singh V. Research and Reviews in Question Answering System. Procedia Technol. 2013;10:417–24.

[19] HIRSCHMAN L, GAIZAUSKAS R. Natural language question answering: the view from here. Nat Lang Eng. 2001;7.

[20] Prager J. Open-Domain Question–Answering. Found Trends® Inf Retr. 2006;1:91–231.

[21] Vargas-Vera M, Lytras MD. AQUA: A Closed-Domain Question Answering System. Inf Syst Manag. 2010 Jul 16;27:217–25.

[22] Mollá D, Vicedo JL. Question Answering in Restricted Domains: An Overview. Comput Linguist. 2007;33:41–61.

[23] Newton RR, Rudestam KE. Your Statistical Consultant: Answers to Your Data Analysis Questions - Rae R. Newton, Kjell Erik Rudestam - Google Books. 2013.

[24] Sasikumar U. A Survey of Natural Language Question Answering System. Int J Comput Appl. 2014;108:975–8887.

[25] Andrenucci A, Sneiders E. Automated question answering: Review of the main approaches. Proc - 3rd Int Conf Inf Technol Appl ICITA 2005. 2005;I:514–9.

[26] Dale R, Moisl H, Somers HL. Handbook of natural language processing. Marcel Dekker; 2000. 943 p.

[27] Allen J, James. Natural language understanding. Benjamin/Cummings Pub. Co; 1995. 654 p.

[28] Meknavin S, Charoenpornsawat P, Kijsirikul B. Feature-based Thai Word Segmentation. Proc Nat Lang Process Pacific Rim Symp 1997. 1997;2.

[29] Jusczyk PW, Houston DM, Newsome M. The beginnings of word segmentation in english-learning infants. Cogn Psychol. 1999;39:159–207.

[30] Srivastava A, Sahami M. Text mining: Classification, clustering, and applications. 2009.

[31] Yang CS, Wong A, Salton G. A vector space model for automatic indexing. Commun ACM. 1975;18:613–20.

[32] Jaruskulchai C, Frieder G. An automatic indexing for Thai text retrieval. Vol. 9839343. 1998. 136-136 p. p.

[33] Hinterberger H, Domingo-Ferrer J, Kashyap V, Khatri V, Snodgrass RT, Terenziani P, et al. Term Weighting. In: Encyclopedia of Database Systems. Boston, MA: Springer US; 2009. p. 3037–40.

[34] Manning CD, Raghavan P, Schutze H. Introduction to information retrieval. Cambridge University Press; 2008. 482 p.

[35] Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. 1995;

[36] Lin D. An Information-Theoretic Definition of Similarity. Proc ICML. 1998;296–304.

[37] Lewis J, Ossowski S, Hicks J, Errami M, Garner HR. Text similarity: An alternative way to search MEDLINE. Bioinformatics. 2006;22:2298–304.

[38] Hawking D. Overview of the TREC-9 web track. NIST Spec Publ 500-249 Ninth Text Retr Conf. 2001;54:87–102.

[39] Hawking D, Voorhees EM, Craswell N, Bailey P. Overview of the TREC-8 Web Track. NIST Spec Publ 500-246 Eighth Text Retr Conf (TREC 8). 1999;1–18.

[40] Marcus MP, Santorini B, Marcinkiewicz MA. Building a large annotated corpus of English: The Penn Treebank. Comput Linguist. 1993;19:313–30.

[41] Baeza-Yates R, Frakes WB. Information retrieval: data structures & algorithms. Prentice Hall PTR; 1992. 419-442 p.

[42] Rodrigo A, Peñas A. A study about the future evaluation of Question-Answering systems. Knowledge-Based Syst. 2017;137:83–93.

[43] National Institute of Standards and Technology. Common Evaluation Measures. Eighteenth Text Retr Conf (TREC 2009) Proc. 2009;

[44] Manning CD, Raghavan P, Schütze H. Chapter 8: Evaluation in information retrieval. In: Introduction to information retrieval. 2008.

[45] Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a Method for Automatic Evaluation of Machine Translation.

[46] Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries.

[47] Green B, Wolf A, Chomsky C, Laughery K. Baseball: An automatic Question Answerer. Proc West Comput Conf. 1961;219–24.

[48] Merugu D, Nightingale R, Thobani I. Finding Answers to Non-factoid Questions Using a Recursive Neural Network. :1–4.

[49] Hristovski D, Dinevski D, Kastrin a, Rindflesch TC. Biomedical question answering using semantic relations. BMC Bioinformatics. 2015;16:6.

[50] Nguyen T, Rosenberg M, Song X, Gao J, Tiwary S, Majumder R, et al. MS MARCO: A human generated MAchine reading COmprehension dataset. CEUR Workshop Proc. 2016;1773.

[51] Duan N, Tang D, Chen P, Zhou M. Question Generation for Question Answering. Proc of EMNLP'17. 2017;877–85.

[52] Cairns BL, Nielsen RD, Masanz JJ, Martin JH, Palmer MS, Ward WH, et al. The MiPACQ Clinical Question Answering System. AMIA Annu Symp Proc. 2011;171–180.

[53] Yang L, Ai Q, Spina D, Chen RC, Pang L, Bruce Croft W, et al. Beyond factoid QA: Effective methods for non-factoid answer sentence retrieval. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2016;9626:115–28.

[54] Gareth J, Daniela W, Trevor H, Rober T. An Introduction to Statistical Learning with Applications in R. Curr Med Chem. 2000;7:995–1039.

[55] Ripley BD. Pattern recognition and neural networks. Pattern Recognit Neural Networks. 2014;1–403.

[56] Prechelt L. Early stopping - But when? Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2012;7700 LECTU:53–67.

[57] Tan C, Wei F, Yang N, Du B, Lv W, Zhou M. S-Net: From Answer Extraction to Answer Generation for Machine Reading Comprehension. 2017;

[58] Seo M, Kembhavi A, Farhadi A, Hajishirzi H. Bidirectional Attention Flow for Machine Comprehension. 2016;

[59] Clarke CLA, Terra EL. Passage Retrieval vs . Document Retrieval for Factoid Question Answering. 2002;0–1.

[60] Wang Y, Liu K, Liu J, He W, Lyu Y, Wu H, et al. Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification. 2018;

ประวัติผู้เขียน

ชื่อ	นายณัฐกิตติ์ ศรีกาญจนเพริศ
วันเกิด	วันที่ 21 กันยายน พ.ศ. 2537
สถานที่เกิด	อำเภอเมือง จังหวัดราชบุรี
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 511/80 หมู่ที่ 20 ตำบลขามเรียง อำเภอกันทรวิชัย จังหวัดมหาสารคาม รหัสไปรษณีย์ 44150
ตำแหน่งหน้าที่การงาน	พนักงานประจำ
สถานที่ทำงานปัจจุบัน	บริษัท อัลฟ่าเมตริกส์ จำกัด เลขที่ 109/1 หมู่ที่ 2 ตำบลมิตรภาพ อำเภอมวกเหล็ก จังหวัดสระบุรี รหัสไปรษณีย์ 18180
ประวัติการศึกษา	พ.ศ. 2553 มัธยมศึกษาตอนต้น โรงเรียนร้อยเอ็ดวิทยาลัย อำเภอเมืองร้อยเอ็ด จังหวัดร้อยเอ็ด พ.ศ. 2556 ประกาศนียบัตรวิชาชีพ (ปวช.) สาขาคอมพิวเตอร์ธุรกิจ วิทยาลัยอาชีวศึกษาร้อยเอ็ด อำเภอเมืองร้อยเอ็ด จังหวัดร้อยเอ็ด พ.ศ. 2560 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยมหาสารคาม อำเภอกันทรวิชัย จังหวัดมหาสารคาม พ.ศ. 2562 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยมหาสารคาม อำเภอกันทรวิชัย จังหวัดมหาสารคาม
ผลงานวิจัย	Srikanjanapers N., Polpinij J., Luaphol B., An Answer Retrieval Method for Non-factoid Question Answering, 24th International Symposium on Artificial Life and Robotics, B-Con Plaza, Beppu, 2019

พหุน ปณฺฑิโต ชีเว