วิธีการสำหรับการสร้างหุ่นยนต์สนทนาไทยโดยใช้หน่วยความจำระยะสั้นแบบยาวแบบสยามและการแต่งเติมข้อมูลเชิงข้อความ

น.ส.ธนัญญา พีรพัฒนาการ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

An Approach for Thai Chatbot Construction Using Siamese Long Short-Term Memory and Text Data-Augmentation

Miss Thananya Phreeraphattanakarn

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2019

หัวข้อวิทยานิพนธ์	วิธีการสำหรับการสร้างหุ่นยนต์สนทนาไทยโดยใช้หน่วยความจำระยะสั้นแบบยาวแบบสยามและการแต่งเติมข้อมูลเชิงข้อความ
โดย	น.ส.ธนัญญา พีรพัฒนาการ
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

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

คณะกรรมการสอบวิทยานิพนธ์

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

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

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ฑิตยา หวานวารี)

..... กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.ชลวิช นัทธี)

ธนัญญา พีรพัฒนาการ : วิธีการสำหรับการสร้างหุ่นยนต์สนทนาไทยโดยใช้หน่วยความจำระยะสั้นแบบยาวแบบสยามและการแต่งเติมข้อมูลเชิงข้อความ. (An Approach for Thai Chatbot Construction Using Siamese Long Short-Term Memory and Text Data-Augmentation) อ.ที่ปรึกษาหลัก : ศ. ดร.บุญเสริม กิจศิริกุล

แนวคิดการนำหุ่นยนต์สนทนามาช่วยในการตอบคำถามปัญหาที่พบบ่อยให้กับผู้รับบริการ เช่น การสอบถามข้อมูลทั่วไปเกี่ยวกับผู้ให้บริการ เป็นต้น เริ่มเป็นที่นิยมมากขึ้นในยุคปัจจุบัน อีกทั้งในการเรียนรู้ของเครื่องสำหรับสร้างหุ่นยนต์สนทนานั้น ชุดข้อมูลที่ใช้สำหรับการเรียนรู้ของแบบจำลอง ถือเป็นอีกหนึ่งสิ่งสำคัญที่จะช่วยให้แบบจำลองให้สามารถทำงานได้อย่างมีประสิทธิภาพ ในงานวิจัยนี้ได้รับการสนับสนุนข้อมูลจากการไฟฟ้านครหลวงแห่งประเทศไทยที่ได้รวบรวมข้อมูลการให้บริการการตอบปัญหาลูกค้าผ่านช่องทางสื่อสังคมออนไลน์ โดยจำนวนของชุดคำถามที่ได้นั้นมีปริมาณน้อยกว่า 1,500 คำถาม ทำให้จำนวนและความหลากหลายของข้อมูลที่มีนั้นส่งผลกับการเรียนรู้ของเครื่องโดยตรง งานวิจัยนี้จึงนำเสนอแนวคิดในการแต่งเติมข้อมูลด้วยวิธีการแทนที่คำด้วยคำที่มีความหมายคล้ายกันด้วยการวัดระยะห่างระหว่างเวกเตอร์น้อยที่สุดเมื่อเทียบกับคำที่ต้องการจะนำไปแทนที่ในประโยคเดิม เพื่อเพิ่มจำนวนและความหลากหลายของข้อมูล จากนั้นจึงนำชุดข้อมูลที่ได้ไปประยุกต์ใช้กับแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ที่ใช้ร่วมกับการหาระยะทางร่วมกับการทดลองหาระยะทางของเวกเตอร์ทั้ง 3 แบบ ได้แก่ การหาระยะทางแบบยุคลิด (Euclidean Distance) การหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) และ การหาค่าความคล้ายโคไซน์ (Cosine Similarity) เพื่อนำไปใช้ในการค้นคืนคำตอบของคำถามที่ได้รับมาจากผู้ใช้งาน ซึ่งผลการทดลองแสดงให้เห็นว่าชุดข้อมูลที่ปรับปรุงด้วยวิธีการแต่งเติมข้อมูลเชิงข้อความที่นำเสนอนั้นสามารถเพิ่มประสิทธิภาพของแบบจำลองได้ดีกว่าชุดข้อมูลตั้งต้น

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2562 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6170930921 : MAJOR COMPUTER SCIENCE

KEYWORD: Text-Data Augmentation, Siamese LSTM

Thananya Phreeraphattanakarn : An Approach for Thai Chatbot Construction Using Siamese Long Short-Term Memory and Text Data-Augmentation. Advisor: Prof. BOONSERM KIJSIRIKUL, Ph.D.

The idea of using a dialogue bot is to provide answers to common questions. For training chatbot, the training dataset is also an important part, which helps machines to learn and accurately make the predictions. In this research, the question-answering dataset used for training and evaluating the system is from กฟน. The dataset is less than 1,500 sentences, which is a small size dataset. The size of a dataset is often responsible for poor performances in the training model. This paper presents a method called Text Data-Augmentation for increasing the textual data. Our approach creates new diverse questions by using cosine similarity for finding a similar word and replacing it in the same sequence. This research used the Siamese Long Short-Term Memory and distance similarity approach for the training model. For the evaluation, we used three distance similarity approaches such as Euclidean Distance, Manhattan Distance, and Cosine Similarity to get the most effective model. The experimental results show that the dataset using Text Data-Augmentation is able to improve the performance of the learned model.

Field of Study: Computer Science Student's Signature .....

Academic Year: 2019 Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี จากแรงสนับสนุน คำแนะนำ ความช่วยเหลือ และกำลังใจจากบุคคลหลายฝ่าย ผู้วิจัยจึงใคร่ขอใช้เนื้อหาในส่วนของกิตติกรรมประกาศเพื่อขอขอบพระคุณเป็นอย่างสูงมา ณ ที่นี้

ขอขอบพระคุณ ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษา ที่ได้เสียสละเวลา ให้ความรู้ คอยให้คำแนะนำแนวทางต่างๆ ในการทำวิจัย ช่วยเหลือทั้งในด้านองค์ความรู้และทรัพยากรสำหรับใช้ในงานวิจัย รวมไปถึงกำลังใจและแรงผลักดันที่สำคัญในการจัดทำวิทยานิพนธ์ตลอดที่ผ่านมา ให้สำเร็จลุล่วงไปได้ด้วยดี ดิฉันรู้สึกเป็นเกียรติอย่างสูงที่ได้เป็นลูกศิษย์ของอาจารย์

ขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ ผู้ช่วยศาสตราจารย์ ดร.ฑิตยา หวานวารี และรองศาสตราจารย์ ดร. ชลวิช นัทธี กรรมการภายนอกมหาวิทยาลัย ที่ได้สละเวลาในการให้ข้อเสนอแนะเพื่อให้วิทยานิพนธ์นี้เป็นประโยชน์ต่อการพัฒนาสังคมอย่างแท้จริง

ขอขอบพระคุณการไฟฟ้านครหลวงที่ให้ความช่วยเหลือและการสนับสนุนชุดข้อมูลที่เป็นประโยชน์อย่างยิ่งสำหรับการวิจัย

ขอขอบพระคุณคุณพ่อ-คุณแม่และครอบครัว ที่สนับสนุนและให้กำลังใจเป็นอย่างมากในการศึกษาต่อระดับปริญญาโทในครั้งนี้

สุดท้ายนี้ ขอบคุณเพื่อนๆ ทุกคน "โจ, พี่ฟาง, น้องเม, พี่หนุย, เบส" และเพื่อนๆ ที่เรียนปริญญาโท คณะวิศวกรรมคอมพิวเตอร์ด้วยกัน ที่คอยให้ความช่วยเหลือ คำแนะนำต่างๆ รวมถึงกำลังใจจนวิทยานิพนธ์เล่มนี้เสร็จสมบูรณ์

ธนัญญา พีรพัฒนาการ

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ค

บทคัดย่อภาษาอังกฤษ ..... ง

กิตติกรรมประกาศ ..... จ

สารบัญ ..... ฉ

สารบัญตาราง ..... ฌ

สารบัญรูปภาพ ..... ฎ

บทที่ 1 บทนำ 1

1.1 ที่มาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของงานวิจัย 4

1.3 ขอบเขตการวิจัย 4

1.4 ประโยชน์ที่ได้รับ 5

1.5 ขั้นตอนการดำเนินงาน 5

1.6 ผลงานวิจัยที่ตีพิมพ์ 5

บทที่ 2 ทฤษฎีที่เกี่ยวข้อง 6

2.1 การเตรียมข้อมูล (preprocessing) 6

2.1.1 การทำความสะอาดข้อมูล (Data Cleaning) 6

2.1.2 การตัดคำ (Word Segmentation) 6

2.1.3 การกำจัดคำหยุด (Stop-Word Removal) 7

2.1.4 คำฝังตัว (Word Embedding) 7

2.2 การเรียนรู้ของเครื่อง (Machine Learning) 9

2.2.1 หน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) 9

2.2.2 โครงข่ายสยาม (Siamese Neural Network) 13
2.3 การหาระยะทางระหว่างเวกเตอร์ของคำ (Distance between vector of two words) ... 14
2.3.1 การหาระยะทางแบบยุคลิด (Euclidean Distance) 15
2.3.2 การหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) 15
2.3.3 การหาค่าความคล้ายโคไซน์ (Cosine Similarity) 16
บทที่ 3 งานวิจัยที่เกี่ยวข้อง 17
3.1 กลุ่มงานวิจัยสำหรับแก้ปัญหาในการตอบคำถามของปัญหาที่พบบ่อย 17
3.1.1 งานวิจัยของ Yichao Lu และคณะ 17
3.1.2 งานวิจัยของ Panitan Muangkammuen และคณะ 19
3.2 กลุ่มงานวิจัยที่เกี่ยวกับโครงข่ายสยามและการหาความคล้ายของข้อความ 22
3.2.1 งานวิจัยของ Jonas Mueller และ Aditya Thyagarajan 22
3.2.2 งานวิจัยของ Nouha Othman และคณะ 25
3.3 กลุ่มงานวิจัยที่เกี่ยวข้องกับการแต่งเติมข้อมูล 27
3.3.1 งานวิจัยของ Anna V. Mosolova และคณะ 27
บทที่ 4 แนวคิดและวิธีการดำเนินงาน 30
4.1 แนวทางการประยุกต์ใช้แบบจำลองหน่วยความจำระยะสั้นแบบยาวแบบสยามร่วมกับการแต่งเติมข้อมูลเชิงข้อความ 30
4.2 แนวทางการปรับปรุงคำฝังตัว 35
4.3 ชุดข้อมูล (Data Set) 35
4.3.1 รูปแบบข้อมูล 36
4.3.2 แนวทางการจัดกลุ่มคำถาม 37
4.3.3 แนวทางการสร้างชุดข้อมูลประโยคที่ถูกจับคู่ 38
4.4 แนวทางการค้นคืนคำตอบแก่ผู้ใช้งาน 40
บทที่ 5 วิธีการทดลอง 41

5.1 ชุดข้อมูลที่ใช้ทดสอบ 41
5.1.1 ชุดข้อมูลตั้งต้น 42
5.1.2 ชุดข้อมูลแต่งเติม 42
5.2 แบบจำลองที่นำมาทดลอง 43
5.3 วิธีการประเมินผล 43
บทที่ 6 ผลการทดลอง 44
6.1 ค่าความแม่นยำในการตอบข้อความ (Recall) 44
6.2 ค่าความเที่ยงตรงในการตอบข้อความ (Precision) 44
6.3 ค่าประสิทธิภาพโดยรวมของระบบ (F1-score) 45
6.4 ค่าความเที่ยงตรงในการตอบข้อความ 5 อันดับแรก (Precision at 5) 46
บทที่ 7 สรุปผลการวิจัยและแนวทางการวิจัยในขั้นถัดไป 47
7.1 สรุปผลการวิจัย 47
7.2 แนวทางการวิจัยในขั้นถัดไป 47
รายการอ้างอิง 49
บรรณานุกรม 51
ประวัติผู้เขียน 53

สารบัญตาราง

หน้า

ตารางที่ 1 ตัวอย่างค่าคำฝังตัว 8

ตารางที่ 2 ตัวอย่างข้อมูลฝึกที่อยู่ในรูปแบบคู่คำถาม-คำตอบ ที่ประกอบไปด้วยตัวอย่างที่ถูกต้อง (Label: '1') และตัวอย่างที่ไม่ถูกต้อง (Label: '0') (อ้างอิงจากตารางที่ 1 ใน [5]) 17

ตารางที่ 3 ตัวอย่างคำถามจากลูกค้าและรูปแบบของคำตอบที่แนะนำจากแบบจำลองที่นำเสนอในงานวิจัย (อ้างอิงจากตารางที่ 4 ใน [5]) 19

ตารางที่ 4 ผลการทดสอบด้วยการคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson's correlation : r), ค่าสัมประสิทธิ์สหสัมพันธ์แบบสเปียร์แมน (Spearman Rank Correlation : ρ) และค่าเฉลี่ยความผิดพลาดกำลังสอง (Mean Square Error: MSE) สำหรับชุดข้อมูล SICK ที่เป็นชุดข้อมูล 24

ตารางที่ 5 ประสิทธิภาพของแบบจำลองต่างๆ ในการค้นคืนคำถามของชุดข้อมูลภาษาอังกฤษ (อ้างอิงจากตารางที่ 3 ใน [7]) 27

ตารางที่ 6 จำนวนของข้อมูลในแต่ละประเภท (อ้างอิงจากตารางที่ 1 ใน [9]) 28

ตารางที่ 7 ตารางแสดงผลการทดสอบแบบจำลอง (อ้างอิงจากตารางที่ 2 ใน [9]) 29

ตารางที่ 8 ตัวอย่างการสร้างรายการคำศัพท์ที่ได้จากการค้นหาคำที่มีความหมายคล้ายกันด้วยเกณฑ์ของค่าความเหมือนโคไซน์ที่มีค่ามากกว่า 0.5 ซึ่งมีค่าสูงสุด 3 อันดับแรก 32

ตารางที่ 9 ตัวอย่างลักษณะชุดข้อมูลของคำถามที่พบบ่อย 36

ตารางที่ 10 ตัวอย่างของการจัดกลุ่มคำถาม 38

ตารางที่ 11 ตัวอย่างของการจับคู่ประโยคที่ทำให้เกิดการซ้ำกัน 39

ตารางที่ 12 ตัวอย่างชุดข้อมูลของประโยคที่ถูกจับคู่และจำแนกประเภท 39

ตารางที่ 13 ค่าความแม่นยำในการตอบข้อความของชุดข้อมูลตั้งต้นเปรียบเทียบกับชุดข้อมูลแต่งเติมเมื่อใช้ร่วมกับคำฝังตัวทั่วไปและคำฝังตัวเฉพาะ 44

ตารางที่ 14 ค่าความเที่ยงตรงในการตอบข้อความของชุดข้อมูลตั้งต้นเปรียบเทียบกับชุดข้อมูลแต่งเติม เมื่อใช้ร่วมกับคำฝังตัวทั่วไปและคำฝังตัวเฉพาะ 45

ตารางที่ 15 ค่าประสิทธิภาพโดยรวมของระบบของชุดข้อมูลตั้งต้นเปรียบเทียบกับชุดข้อมูลแต่งเติม เมื่อใช้ร่วมกับคำฝังตัวทั่วไปและคำฝังตัวเฉพาะ 45

ตารางที่ 16 ค่าความเที่ยงตรงในการตอบข้อความ 5 อันดับแรกของชุดข้อมูลตั้งต้นเปรียบเทียบกับชุดข้อมูลแต่งเติม เมื่อใช้ร่วมกับคำฝังตัวทั่วไปและคำฝังตัวเฉพาะ 46

สารบัญรูปภาพ

หน้า

รูปที่ 1 ตัวอย่างปริภูมิสองมิติที่แสดงว่าคำที่ความหมายคล้ายกันจะอยู่ในตำแหน่งที่ใกล้เคียงกัน [แหล่งอ้างอิง http://suriyadeepan.github.io] 3

รูปที่ 2 ภาพรวมของหุ่นยนต์สนทนาสำหรับการตอบปัญหาของคำถามที่พบบ่อยโดยใช้หน่วยความจำระยะสั้นแบบยาวแบบสยามและการแต่งเติมข้อมูลเชิงข้อความที่นำเสนอในงานวิจัยนี้ 4

รูปที่ 3 ตัวอย่างข้อความก่อนการทำความสะอาดข้อมูล 6

รูปที่ 4 ตัวอย่างข้อความก่อนการทำความสะอาดข้อมูล 6

รูปที่ 5 ตัวอย่างปริภูมิสองมิติที่แสดงตำแหน่งของแต่คำละคำ โดยคำที่ความหมายคล้ายกันจะอยู่ในตำแหน่งที่ใกล้เคียงกัน [แหล่งอ้างอิง http://suriyadeepan.github.io] 9

รูปที่ 6 โครงสร้างของหน่วยความจำระยะสั้นแบบยาว (แหล่งอ้างอิง https://medium.com/@divyanshu132/lstm-and-its-equations-5ee9246d04af) 10

รูปที่ 7 ลักษณะของสถานะเซลล์ภายในหน่วยความจำระยะสั้นแบบยาว (แหล่งอ้างอิง https://colah.github.io/posts/2015-08-Understanding-LSTMs) 10

รูปที่ 8 ประตูลืมภายในหน่วยความจำระยะสั้นแบบยาว (แหล่งอ้างอิง https://colah.github.io/posts/2015-08-Understanding-LSTMs) 11

รูปที่ 9 ประตูนำเข้าภายในหน่วยความจำระยะสั้นแบบยาว (แหล่งอ้างอิง https://colah.github.io/posts/2015-08-Understanding-LSTMs) 12

รูปที่ 10 การคำนวณค่าสถานะเซลล์ใหม่ในประตูนำเข้า (แหล่งอ้างอิง https://colah.github.io/posts/2015-08-Understanding-LSTMs) 12

รูปที่ 11 ประตูนำออกภายในหน่วยความจำระยะสั้นแบบยาว (แหล่งอ้างอิง https://colah.github.io/posts/2015-08-Understanding-LSTMs) 13

รูปที่ 12 ตัวอย่างการใช้โครงข่ายสยามร่วมกันกับแบบจำลองหน่วยความจำระยะสั้นแบบยาว (LSTM) และการหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) เพื่อหาค่าความคล้ายของทั้งสองประโยค (อ้างอิงจาก Fig.1 ใน [6]) 14

รูปที่ 13 ตัวอย่างการหาระยะทางแบบต่างๆ บนปริภูมิสองมิติ ได้แก่ การระยะทางแบบยุคลิด (Euclidean Distance) การหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) หรือ การหาค่าความคล้ายโคไซน์ (Cosine Similarity) (แหล่งอ้างอิง https://dh2016.adho.org/abstracts/253) 14

รูปที่ 14 ตัวอย่างรูปแบบเส้นทางที่สามารถคำนวณระยะทางแบบแมนฮัตตัน (สีแดง สีน้ำเงิน และสีเหลือง) และระยะทางในแนวเส้นตรง (สีเขียว) 15

รูปที่ 15 ตัวอย่างโครงสร้างของการเข้ารหัสประโยคคำถามและประโยคคำตอบด้วยแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) และส่งข้อมูลนำเข้าไปยังโครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron: MLP) (อ้างอิงจาก Fig.1 ใน [5]) 18

รูปที่ 16 ภาพรวมของระบบหุ่นยนต์สนทนาของคำถามที่พบบ่อยแบบอัตโนมัติที่นำเสนอ 19

รูปที่ 17 ตัวอย่างข้อมูลเชิงอักษรที่เป็นคู่ของคำถาม-คำตอบ (อ้างอิงจาก Fig.2 ใน [8]) 20

รูปที่ 18 ตัวอย่างค่าความน่าจะเป็นสำหรับการทำนายของแบบจำลองที่จำแนกประเภทของข้อมูลกลุ่มที่ 5 โดยมีค่าความน่าจะเป็นสูงสุดคือ 0.97 (อ้างอิงจาก Fig.4 ใน [8]) 21

รูปที่ 19 ตัวอย่างค่าความน่าจะเป็นสำหรับการทำนายของแบบจำลองที่จำแนกประเภทของข้อมูลกลุ่มที่ 5 โดยมีค่าความน่าจะเป็นต่ำสุดคือ 0.45 (อ้างอิงจาก Fig.5 ใน [8]) 21

รูปที่ 20 โครงสร้างในแต่ละชั้นและมิติของโครงข่ายประสาทเทียมที่ใช้ส่งผ่านข้อมูล 22

รูปที่ 21 ตัวอย่างแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ที่ทำงานร่วมกับการความคล้ายของประโยคด้วยการหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) โดยใช้ภาพรวมของโครงสร้างแบบโครงข่ายสยาม (Siamese Neural Network) (อ้างอิงจาก Fig.1 [5]) 23

รูปที่ 22 ตัวอย่างภาพรวมของแบบจำลองที่นำเสนอ (LSTMQR) สำหรับการค้นคืนคำถาม 25

รูปที่ 23 ลักษณะโครงสร้างโดยทั่วไปของแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ที่ใช้ร่วมกับการหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) ที่เรียกว่า แบบจำลอง MaLSTM 26

รูปที่ 24 อัลกอริทึมของการแต่งเติมข้อมูล (อ้างอิงจาก Fig.2 ใน [9]) 28

รูปที่ 25 ตัวอย่างของการแต่งเติมข้อมูลทั้งหมด 7 ครั้ง โดยมีการเปลี่ยนแปลงของข้อมูลในประโยคทั้งหมด 25 เปอร์เซ็นต์ (อ้างอิงจาก Fig.1 ใน [9]) 28

รูปที่ 26 โครงสร้างแบบจำลองหน่วยความจำระยะสั้นแบบยาวแบบสยามสำหรับงานวิจัยที่นำเสนอ 31

รูปที่ 27 ภาพรวมของวิธีการแต่งเติมข้อมูลเชิงตัวอักษรด้วยการหาความคล้ายจากเวกเตอร์ของคำ. 34

รูปที่ 28 ตัวอย่างบทสนทนาระหว่างผู้ใช้งานกับหุ่นยนต์สนทนาที่พัฒนาด้วยวิธีการที่นำเสนอ เมื่อพบคำถามที่มีค่าความคล้ายมากกว่าเกณฑ์ที่กำหนดไว้และมีค่าความคล้ายสูงสุดกับคำถามที่เป็นข้อมูลนำเข้า..... 34

รูปที่ 29 ตัวอย่างบทสนทนาระหว่างผู้ใช้งานกับหุ่นยนต์สนทนาที่พัฒนาด้วยวิธีการที่นำเสนอ เมื่อไม่สามารถหาคำถามที่มีความคล้ายสูงสุดกับคำถามที่เป็นข้อมูลนำเข้าที่มีค่ามากกว่าเกณฑ์ที่กำหนดได้ 35

รูปที่ 30 การแบ่งสัดส่วนของชุดข้อมูลที่ใช้ทดสอบแบบจำลอง 42

บทที่ 1
บทนำ

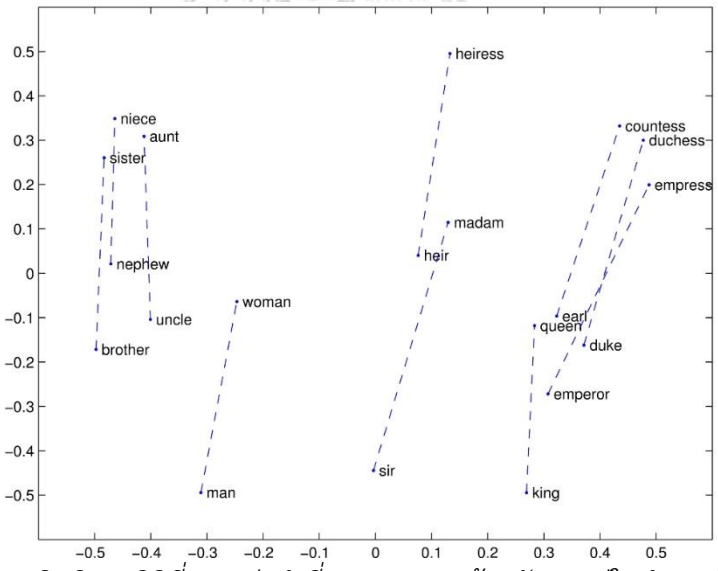
1.1 ที่มาและความสำคัญของปัญหา

การใช้สื่อสังคมออนไลน์ในปัจจุบันได้เข้ามามีบทบาทในการเป็นช่องทางในการติดต่อสื่อสาร ซึ่งกำลังเติบโตและเป็นกระแสนิยมอย่างต่อเนื่อง ด้วยความสะดวก รวดเร็ว และทั่วถึงเพียงแค่เชื่อมต่ออินเทอร์เน็ต ทำให้มีบริษัทและผู้ประกอบการมากมายให้ความสนใจและหันมาทำกิจกรรมต่างๆ กับลูกค้าบนสื่อสังคมออนไลน์มากยิ่งขึ้น โดยผลจากการรายงานของ “Hootsuite” ซึ่งเป็นหนึ่งในผู้ให้บริการระบบจัดการสื่อสังคมออนไลน์และที่ปรึกษาทางการตลาด [1] ได้กล่าวว่าในปี 2019 บริษัทชั้นนำหลายแห่งมีการปรับตัวกับกระแสความต้องการของผู้บริโภคที่นิยมการรับบริการแบบรายบุคคลมากขึ้น จากการสำรวจข้อมูลจากบริษัทที่เป็นลูกค้าของ Hootsuite จำนวน 3,255 ราย พบว่า ห้าปัจจัยหลักสำคัญของกระแสสื่อสังคมออนไลน์ในปี 2019 ได้แก่ (1) การสร้างความเชื่อมั่นผ่านสื่อสังคมออนไลน์อย่างต่อเนื่อง (2) การนำเสนอเรื่องราวของผลิตภัณฑ์ผ่านการเล่าเรื่องในรูปแบบใหม่ๆ (3) การลดช่องว่างในการโฆษณาเพื่อให้เข้าถึงลูกค้าได้มากขึ้น เพราะบริษัทชั้นนำหลายแห่งต่างแข่งขันกันประชาสัมพันธ์โฆษณาบนสื่อสังคมออนไลน์ (4) ประยุกต์การใช้งานเทคโนโลยีเพื่อให้สำเร็จในการขายบนโลกออนไลน์ได้ (5) ความต้องการของลูกค้าที่อยากได้รับประสบการณ์แบบ 1:1 บนสื่อสังคมออนไลน์ จากปัจจัยข้อที่ 5 ยังมีข้อมูลเพิ่มเติมอีกว่า แอปพลิเคชันที่ใช้ในการสื่อสารยอดนิยมอย่าง WhatsApp, Facebook Messenger, WeChat, QQ, และ Skype ต่างมีจำนวนผู้ใช้งานระบบในแต่ละเดือนประมาณ 5 ล้านบัญชี [2] และมีแนวโน้มว่าผู้ใช้งานสื่อสังคมออนไลน์นั้นจะใช้เวลากับการสนทนาในแอปพลิเคชันสำหรับส่งข้อความ (Messenger) มากกว่าการแบ่งปันข่าวสารผ่านบัญชีของตน [3] นอกจากนี้ยังมีการสำรวจผู้ใช้งานแอปพลิเคชัน Facebook จำนวน 6,000 คน พบว่า 9 ใน 10 คน นิยมใช้การส่งข้อความในการพูดคุยทางธุรกิจ [4] และในประเทศสหรัฐอเมริกา การส่งข้อความเป็นช่องทางที่ได้รับความนิยมสูงสุดสำหรับการติดต่อในการให้บริการลูกค้า (Customer Service) โดยในปี 2018 Facebook ได้ทำการสำรวจผู้ใช้งานจำนวน 8,000 คน พบว่าร้อยละ 69 ของผู้ใช้งานรู้สึกว่าการได้ติดต่อสื่อสารด้วยการส่งข้อความโดยตรงไปยังบริษัทช่วยให้พวกเขามีความเชื่อมั่นในผลิตภัณฑ์นั้นๆ มากขึ้น ซึ่งแนวโน้มเหล่านี้แสดงให้เห็นว่าเหล่าผู้รับบริการนิยมที่จะใช้ช่องทางของสังคมออนไลน์ในการติดต่อกับบริษัทเพื่อสอบถามข้อมูล รวมถึงแจ้งปัญหาต่างๆ มากกว่าการโทรศัพท์ติดต่อเพื่อสอบถามข้อมูลผ่านทางศูนย์บริการข้อมูลลูกค้า (Call Center) ซึ่งใช้เวลาค่อนข้างนาน เนื่องจากมีผู้ใช้บริการโทรศัพท์เข้ามาเป็นจำนวนมากและในบางครั้งเจ้าหน้าที่ผู้ให้บริการก็ไม่เพียงพอต่อความต้องการ ทำให้ผู้รับบริการรู้สึกไม่ได้รับการตอบสนองเมื่อเปรียบเทียบกับการส่งข้อความไปหาบริษัทโดยตรง นอกจากนี้การตอบปัญหาจากผู้รับบริการที่ติดต่อ

เข้ามายังเป็นอีกหนึ่งปัญหาสำคัญที่มีมาก่อนที่จะเกิดกระแสความนิยมในการส่งข้อความ ทำให้บริษัทชั้นนำอย่าง Amazon ที่เป็นผู้ให้บริการในการซื้อขายสินค้าออนไลน์ ได้พัฒนาและวิจัย [5] ระบบตอบบทสนทนากับปัญหาที่มีขอบเขตจำกัด (Close Domain) ขึ้นมาในปี 2017 จากการศึกษางานวิจัยพบว่าร้อยละ 70 ของคำถามที่มาจากผู้รับบริการมักจะเป็นคำถามที่ถูกถามบ่อยครั้ง (Frequently asked questions: FAQ) ในงานวิจัยได้นำเอาคำถามจากผู้รับบริการไปทำการค้นหาว่าคำถามนั้นมีความคล้ายกับคำถามใดในรายการของคำถามที่เคยถูกถามเข้ามาและระบบได้ทำการตอบคำถามเหล่านั้นไปแล้ว โดยใช้การเรียนรู้ของเครื่องแบบหน่วยความจำระยะสั้นแบบยาว (Long-Short Term Memory: LSTM) ร่วมกับการใช้โครงข่ายสยาม (Siamese Neural Network) ซึ่งเป็นวิธีที่ได้รับความนิยมและได้รับการยอมรับในการหาความคล้ายจากงานวิจัยของ Jonas Mueller และคณะในปี 2016 [6] และเมื่อระบบพบว่าคำถามที่ผู้รับบริการถามนั้นคล้ายกับคำถามใด ระบบจะทำการตอบสนองด้วยการให้คำตอบของคำถามที่คล้ายกันออกไปเพื่อแก้ไขปัญหาเบื้องต้นให้กับของผู้รับบริการโดยจะมุ่งเน้นไปที่การตอบปัญหาปลายปิด ซึ่งประกอบไปด้วย 2 ส่วนหลักคือ (1) การคำนวณหาความคล้ายกันของคำถาม (2) ตอบคำถามด้วยคำตอบของคำถามที่คล้ายกัน งานวิจัยส่วนมากมักจะถูกพัฒนาเพื่อใช้สำหรับภาษาอังกฤษ แต่จากการศึกษาพบว่ามีงานวิจัยที่เกี่ยวข้องกับระบบหุ่นยนต์สนทนาเพื่อใช้ตอบปัญหาของคำถามที่พบบ่อยสำหรับภาษาอื่น เช่น งานวิจัยของ Nouha Othman และคณะ [7] ที่เสนอแนวคิดเกี่ยวกับการค้นคืนคำถามโดยใช้แบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ที่ใช้ร่วมกับการหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) ที่มีโครงสร้างแบบโครงข่ายสยาม (Siamese Neural Network) เพื่อจะค้นคืนคำถามที่เคยถูกถามแล้วและทำการหาความคล้ายกับคำถามใหม่ที่ได้รับเป็นข้อมูลนำเข้าที่เป็นภาษาอาราบิกและงานวิจัยของ Panitan Muangkammuen และคณะ [8] ที่นำเสนอระบบหุ่นยนต์สนทนาเพื่อใช้ตอบปัญหาของคำถามที่พบบ่อย โดยนำเสนอการใช้แบบจำลองนิวรอลเน็ตเวิร์กแบบวกกลับ (Recurrent Neural Network: RNN) ชนิดรูปแบบหน่วยความจำระยะสั้นแบบยาว (Long-Short Term Memory: LSTM) มาใช้ในการหาคำตอบ

ในการพัฒนาระบบหุ่นยนต์สนทนาเพื่อใช้ตอบปัญหาของคำถามที่พบบ่อย (FAQ Chatbot) ด้วยวิธีการเรียนรู้ของเครื่องนั้น ผู้วิจัยพบว่าการเรียนรู้ของเครื่องจำเป็นที่จะต้องมีข้อมูลสำหรับฝึกและทดสอบการเรียนรู้ของเครื่อง ซึ่งผู้วิจัยได้รับการสนับสนุนข้อมูลจากการไฟฟ้านครหลวงแห่งประเทศไทยที่ได้รวบรวมข้อมูลการให้บริการการตอบปัญหาลูกค้าผ่านช่องทางสื่อสังคมออนไลน์ เช่น Facebook Messenger, Twitter และ Line เพื่อใช้เป็นข้อมูลสำหรับการเรียนรู้ของเครื่อง โดยข้อมูลที่ได้นั้นเป็นข้อมูลที่มีทั้งคำถามแบบปลายเปิดและคำถามแบบปลายปิด หลังจากที่ผู้วิจัยได้ทำการทำความสะอาดข้อมูล (Data Cleaning) เพื่อให้ได้คำถามเฉพาะปลายปิดและเป็นคำถามที่พบบ่อยเรียบร้อยแล้ว จำนวนของชุดคำถามที่ได้นั้นมีปริมาณน้อยกว่า 1,500 คำถาม ซึ่งจำนวนและ

ความหลากหลายของข้อมูลนั้นส่งผลกับการเรียนรู้ของเครื่องโดยตรง ดังนั้นในการแก้ปัญหาเรื่องจำนวนและความหลากหลายของข้อมูล ผู้วิจัยจึงมีแนวคิดที่จะพัฒนาวิธีการเพิ่มชุดข้อมูลคำถามให้มีจำนวนมากขึ้นโดยใช้วิธีการสร้างชุดข้อมูลเพิ่มจากข้อมูลเดิมด้วยการปรับคุณลักษณะบางอย่างของข้อมูลเดิม (Data Augmentation) ที่นิยมนำมาใช้ในงานประมวลผลภาพ แต่จากการศึกษาพบว่า ในปี 2018 Anna V. Mosolova และคณะ [9] ได้นำเสนอวิธีการแต่งเติมข้อมูลที่ได้แนวคิดมาจากการแต่งเติมข้อมูลรูปและเสียง มาประยุกต์ใช้กับงานประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) โดยนำเสนอหลักการในการแต่งเติมข้อมูลด้วยวิธีการ แทนที่คำด้วยคำที่มีความหมายคล้ายกัน (Synonymy) โดยผลลัพธ์ที่ได้จากงานวิจัยพบว่า การแต่งเติมข้อมูล (Data Augmentation) ที่นำเสนอส่งผลให้ประสิทธิการเรียนรู้ของเครื่องมีผลลัพธ์ที่ดีขึ้น ผู้วิจัยจึงมีแนวคิดที่จะประยุกต์ใช้วิธีการแต่งเติมข้อมูลด้วยการแทนคำที่มีความคล้ายกันด้วยการวัดระยะห่างระหว่างเวกเตอร์น้อยที่สุดเมื่อเทียบกับคำที่ต้องการจะนำไปแทนที่ในประโยคเดิม ซึ่งกระบวนการนี้จะทำให้ได้ข้อมูลใหม่ที่ได้ยังคงมีความคล้ายคลึงกับชุดข้อมูลเดิม



รูปที่ 1 ตัวอย่างปริภูมิสองมิติที่แสดงว่าคำที่ความหมายคล้ายกันจะอยู่ในตำแหน่งที่ใกล้เคียงกัน [แหล่งอ้างอิง http://suriyadeepan.github.io]

ในงานวิจัยนี้จะนำแบบจำลองจากงานวิจัยของ Jonas Mueller และคณะ [6] มาเป็นแบบจำลองที่ใช้ร่วมกันกับแนวคิดการแต่งเติมข้อมูลเชิงตัวข้อความที่ผู้วิจัยนำเสนอเพื่อใช้ประเมินผลลัพธ์ที่ได้ และมีภาพรวมของวิธีการที่นำเสนอดังรูปที่ 2

รูปที่ 2 ภาพรวมของหุ่นยนต์สนทนาสำหรับการตอบปัญหาของคำถามที่พบบ่อยโดยใช้หน่วยความจำระยะสั้นแบบยาวแบบสยามและการแต่งเติมข้อมูลเชิงข้อความที่นำเสนอในงานวิจัยนี้

1.2 วัตถุประสงค์ของงานวิจัย

นำเสนอแนวทางการสร้างหุ่นยนต์สนทนาไทยโดยใช้หน่วยความจำระยะสั้นแบบยาวแบบสยามและการแต่งเติมข้อมูลเชิงข้อความเพื่อพัฒนาให้แนวทางที่นำเสนอสามารถทำงานได้อย่างมีประสิทธิภาพมากขึ้นและทำงานได้ดีกับข้อมูลที่มีปริมาณน้อย

1.3 ขอบเขตการวิจัย

1. ข้อมูลเชิงตัวอักษรที่ใช้ในงานวิจัยนี้จะเป็นข้อมูลภาษาไทยเท่านั้น
2. ข้อมูลเชิงตัวอักษรที่ใช้ในงานวิจัยนี้ จะเป็นคำถามปลายปิดที่เป็นกลุ่มคำถามเกี่ยวกับการสอบถามงานด้านบริการผู้ใช้ไฟฟ้า, สอบถามงานด้านวิธีการชำระเงินค่าบริการไฟฟ้า, สอบถามข้อมูลประกาศดับไฟ และ สอบถามเรื่องทั่วไปเกี่ยวกับการไฟฟ้านครหลวง
3. ทำการวัดประสิทธิภาพของวิธีการเพิ่มข้อมูลที่นำเสนอด้วยการวัดประสิทธิภาพแบบจำลองโดยพิจารณาจาก ค่าประสิทธิภาพโดยรวมของระบบ (F1-score) และค่าความเที่ยงตรงในการตอบข้อความ 5 อันดับแรก (Precision at 5)
4. เปรียบเทียบการเพิ่มประสิทธิภาพของแบบจำลอง [5] ด้วยการให้แบบจำลองเรียนรู้กับข้อมูลที่ผ่านกระบวนการแต่งเติมข้อมูลด้วยวิธีที่นำเสนอด้วยเทียบกับการเรียนรู้กับข้อมูลปกติ
5. ข้อมูลนำออกที่ได้จะเป็นคำตอบของคำถามที่มีความคล้ายสูงสุดเมื่อเทียบกับคำถามที่เป็นข้อมูลนำเข้า

1.4 ประโยชน์ที่ได้รับ

1. สามารถแต่งเติมข้อมูลเพื่อใช้ในการเพิ่มคุณภาพของแบบจำลอง
2. สามารถเพิ่มคุณภาพของแบบจำลองเพื่อรองรับการใช้งานสำหรับภาษาไทย
3. สามารถนำวิธีการแต่งเติมข้อมูลที่นำเสนอไปประยุกต์ใช้กับการแต่งเติมข้อมูลในภาษาอื่นๆ
4. สามารถนำกรอบงานวิจัยนี้ไปประยุกต์ใช้กับปัญหาขอบเขตอื่น
5. สามารถนำข้อมูลจำลองบทสนทนาภาษาไทยไปพัฒนาแบบจำลองอื่นๆ ต่อไป

1.5 ขั้นตอนการดำเนินงาน

1. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง
2. ดำเนินการทดสอบงานวิจัยที่เกี่ยวข้อง
3. วิเคราะห์ผลจากการทดสอบ
4. นำผลจากการวิเคราะห์มาปรับปรุงการเตรียมข้อมูลเพื่อใช้กับแบบจำลอง
5. ทดสอบเบื้องต้นกับวิธีการเตรียมข้อมูลที่นำเสนอ
6. สอบหัวข้อวิทยานิพนธ์
7. เขียนบทความเพื่อตีพิมพ์ผลงานทางวิชาการ
8. ทดสอบแบบจำลองพร้อมปรับปรุงแบบจำลองเพิ่มเติม
9. สรุปผลและเขียนวิทยานิพนธ์
10. สอบวิทยานิพนธ์

1.6 ผลงานวิจัยที่ตีพิมพ์

“Text Data Augmentation Using Text Similarity with Manhattan Siamese Long Short Term Memory for Thai Language” โดย ธนัญญา พีรพัฒนาการ และ บุญเสริม กิจศิริกุล ในงานประชุมวิชาการ “2020 International Conference on Computational Linguistics and Natural Language Processing (CLNLP 2020)” จัดขึ้น ณ สาธารณรัฐเกาหลี ระหว่างวันที่ 20 ถึง 21 กรกฎาคม 2563

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 การเตรียมข้อมูล (preprocessing)

การนำข้อมูลประเภทตัวอักษรมาใช้งานร่วมกับกับแบบจำลองนั้น มีความจำเป็นที่จะต้องผ่านกระบวนการเตรียมข้อมูลเพื่อแปลงภาษาธรรมชาติ (Natural Language) หรือภาษาที่มนุษย์ใช้สื่อสารให้กลายเป็นข้อมูลประเภทตัวเลข (Numerical) ก่อนที่จะนำไปใช้ในการเรียนรู้ด้วยคอมพิวเตอร์ ซึ่งประกอบด้วยขั้นตอนต่างๆ ดังต่อไปนี้

2.1.1 การทำความสะอาดข้อมูล (Data Cleaning)

เนื่องจากข้อมูลประเภทตัวอักษรที่ได้มาจากบทสนทนาระหว่างผู้รับบริการและผู้ให้บริการนั้น อาจมีเครื่องหมายหรือสัญลักษณ์ต่างๆ ที่ไม่ต้องการนำมาใช้ในการฝึกสอน เช่น !, €, @, %, #, *, &, ~ รวมถึงข้อมูลที่เป็นตัวเลข จึงต้องทำความสะอาดข้อมูลเหล่านั้นก่อนที่จะนำมาทำเป็นตัวแทนข้อความต่อไป ตัวอย่างข้อมูลก่อนและหลังการทำความสะอาดข้อมูลแสดงดังรูปที่ 3 และ 4 ตามลำดับ

```
[1]  text_list = ['''เฟ็ดเฟ่อย่างกับเดอะค็อป 999 (คลิป) <br> #ข่าวสด #เด่นออนไลน์ <br>รีบดูก่อนโดนลบ! เฟ็ดเฟ่สร้างความฮึกเฮิม นั่งหลังคารถชูถ้วยจุดพลุ มั่นใจมาแน่นอน
                  <br>บิวต์กันเข้าไป ก่อนเกมการแข่งขัน ยูฟ่าแชมเปียนส์ลีก รอบ … <br><img src='https://xxx'>''',

                  '''แอดขอให้คุณผู้อ่านใช้วิจารณญาณและดุลยพินิจในการอ่าน คำพูดที่สอบถามคุณยายเป็นเพียงความข้างเดียวที่ออกมาจากใจและปากของคุณยาย ยังไม่มีโอกาสไปสอบถามความ
                  อีกมุมจากตัวลูกชายคนเล็ก #แม่เฒ่าถูกไล่ #แม่เฒ่าขาพิการ #พาหลานชายพิการนอนข้างทาง <br>'อย่าว่าเขาๆคงมีเหตุผล' 'ยาย'ถูก'ลูก'ไล่-ขอร้องแม้ปวดใจ <br>
                  ประชาชนที่รู้ข่าวต่างนำอาหารกับน้ำดื่ม ไปให้ 2 ยายหลานผู้พิการได้ดื่มกิน หลังถูกลูกชายไล่จนต้องนอนศาลาร.... <br><img src='https://xxx'>''']
```

รูปที่ 3 ตัวอย่างข้อความก่อนการทำความสะอาดข้อมูล

(แหล่งอ้างอิง https://link.medium.com/snawwVvJ43)

```
clean message:
 เฟ็ดเฟ่อย่างกับเดอะค็อป 999 คลิป ข่าวสด เด่นออนไลน์ รีบดูก่อนโดนลบ เฟ็ดเฟ่สร้างความฮึกเฮิม นั่งหลังคารถชูถ้วยจุดพลุ มั่นใจมาแน่นอน บิวต์กันเข้าไป ก่อนเกมการแข่งขัน
ยูฟ่าแชมเปียนส์ลีก รอบ …
clean message:
 แอดขอให้คุณผู้อ่านใช้วิจารณญาณและดุลยพินิจในการอ่าน คำพูดที่สอบถามคุณยายเป็นเพียงความข้างเดียวที่ออกมาจากใจและปากของคุณยาย ยังไม่มีโอกาสไปสอบถามความ อีกมุมจากตัวลู
เล็ก แม่เฒ่าถูกไล่ แม่เฒ่าขาพิการ พาหลานชายพิการนอนข้างทาง อย่าว่าเขาๆคงมีเหตุผล ยายถูกลูกไล่ขอร้องแม้ปวดใจ ประชาชนที่รู้ข่าวต่างนำอาหารกับน้ำดื่ม ไปให้ 2 ยายหลาน
ผู้พิการได้ดื่มกิน หลังถูกลูกชายไล่จนต้องนอนศาลาร
```

รูปที่ 4 ตัวอย่างข้อความก่อนการทำความสะอาดข้อมูล

(แหล่งอ้างอิง https://link.medium.com/snawwVvJ43)

2.1.2 การตัดคำ (Word Segmentation)

เป็นการนำข้อมูลนำเข้าที่เป็นประโยคมาทำการตัดแบ่งประโยคให้ออกมาเป็นคำและสร้างเป็นรายการของคำในแต่ละประโยค ซึ่งในการประมวลผลภาษาธรรมชาตินั้น ภาษาไทยถูกจัดอยู่ในประเภทของภาษาที่ไม่ตัดคำ (Unsegmented Language) เนื่องจากในภาษาไทยจะไม่มีการใช้ตัวอักขระใดๆ ในการบ่งบอกขอบเขตของคำอย่างชัดเจน ทำให้วิธีการตัดคำมีความแตกต่างจากภาษาอื่นๆ เช่น ภาษาอังกฤษที่สามารถใช้ช่องว่าง (space) ที่คั่นอยู่ระหว่างคำมาใช้ในการบ่งบอกขอบเขตของคำทำให้สะดวกต่อการตัดคำ ดังนั้นการตัดคำในภาษาไทยจึงมีการพัฒนาวิธีการตัดคำโดยเฉพาะสามารถแบ่งออกได้ 3 วิธี ได้แก่ (1) หลักการตัดคำด้วยการใช้กฎไวยากรณ์ทางภาษา (Rule-based) (2) หลักการตัดคำด้วยการอ้างอิงคำจากพจนานุกรม (Dictionary-based) และ (3) หลักการตัดคำด้วยการสร้างแบบจำลองการเรียนรู้จากคลังข้อความขนาดใหญ่ (Machine Learning or Corpus

based) โดยแต่ละวิธีนั้นให้ผลลัพธ์ในด้านของความถูกต้อง ความรวดเร็วในการทำงานและปริมาณทรัพยากรที่ใช้แตกต่างกัน ในงานวิจัยนี้ได้เลือกใช้ PyThaiNLP ซึ่งเป็นไลบารีของภาษาไพธอน (Python) สำหรับตัดคำภาษาไทย [15] และเลือกใช้หลักการตัดคำด้วยการอ้างอิงคำจากพจนานุกรม (Dictionary-based) ร่วมกับการตัดคำแบบตรงมากที่สุด (Maximum Matching) โดยจะใช้วิธีการหารูปแบบในการตัดคำที่สามารถเป็นไปได้ทั้งหมด จากนั้นจะเลือกรูปแบบที่สามารถตัดคำแล้วได้จำนวนคำที่น้อยที่สุดเป็นผลลัพธ์ ดังตัวอย่างต่อไปนี้

ประโยคตัวอย่าง : [“จะเปลี่ยนชื่อผู้ขอใช้ไฟฟ้าจากชื่อคนที่เป็นเจ้าของบ้านมาเป็นชื่อผมซึ่งเป็นหลานสามารถทำได้หรือเปล่าครับ”]

การตัดคำแบบสอดคล้องมากที่สุด (Maximum Matching) ให้ผลลัพธ์เป็น : ['จะ', 'เปลี่ยน', 'ชื่อ', 'ผู้', 'ขอ', 'ใช้', 'ไฟฟ้า', 'จาก', 'ชื่อ', 'คน', 'ที่', 'เป็น', 'เจ้าของบ้าน', 'มา', 'เป็น', 'ชื่อ', 'ผม', 'ซึ่ง', 'เป็น', 'หลาน', 'สามารถ', 'ทำได้', 'หรือเปล่า', 'ครับ']

2.1.3 การกำจัดคำหยุด (Stop-Word Removal)

เป็นการนำคำที่ไม่มีความหมายเป็นนัยสำคัญออกจากประโยคโดยที่ไม่ทำให้ความหมายของทั้งประโยคเปลี่ยนไป ซึ่งคำที่ไม่มีความหมายที่เป็นนัยสำคัญเหล่านี้เป็นคำที่สามารถใช้ได้ในความหมายทั่วไปและสามารถพบได้บ่อยในหลายประโยคและไม่ได้มีความหมายเฉพาะในประโยคนั้นๆ ทำให้เมื่อนำคำเหล่านี้ออกจากประโยคแล้วก็ไม่ได้ทำให้ใจความสำคัญของประโยคนั้นเปลี่ยนไป ตัวอย่างเช่น 'ทั้งหมด', 'พอเหมาะ', 'เมื่อไร', 'ประการหนึ่ง', 'เพื่อให้', 'อย่างไรเสีย', 'ไร', 'ตลอดถึง', 'เป็นต้นไป', 'ครา', 'พวกนู้น', 'ไม่', 'ตามๆ', 'ทันใดนั้น', 'สมัยโน้น', 'แค่ไหน', 'มอง', 'เช่นเดียวกัน', 'ช่วงระหว่าง', 'คุณ', 'ที่ละ', 'หนึ่ง', 'ก็', 'บางที', 'นั้นไว', 'ได้แต่', 'พวกแก', 'แค่ว่า', 'คราวละ', 'นี้เอง', 'ฝ่ายใด', 'นานๆ', 'ปฏิบัติ', 'นาน', 'ฯลฯ', 'กัน', 'ครบ', 'เหตุไร', 'เป็นเพื่อ', 'เสีย', 'ถ้า', 'ถึงแก่', 'สำคัญ', 'จริง', 'รวม', 'หลัง', 'นางสาว', 'เผื่อ', 'ยิ่งกว่า', 'แหละ', 'สั้นๆ', 'เปิดเผย', 'ทั้งเป็น', 'ก่อนๆ', 'ต่างหาก', 'พอกัน', 'ข้างล่าง', 'แค่นั้น', 'ในระหว่าง', 'เพียงเพื่อ', 'จากนี้', 'คราวหนึ่ง', 'ไหน', 'ใดๆ', 'เสียนั่นเอง', 'ที', 'ด้วยเหตุที่' เป็นต้น

ซึ่งการกำจัดคำหยุดนั้นทำให้ประหยัดเวลาในการประมวลผลและลดการสร้างคุณลักษณะของคำศัพท์ที่ไม่จำเป็น

2.1.4 คำฝังตัว (Word Embedding)

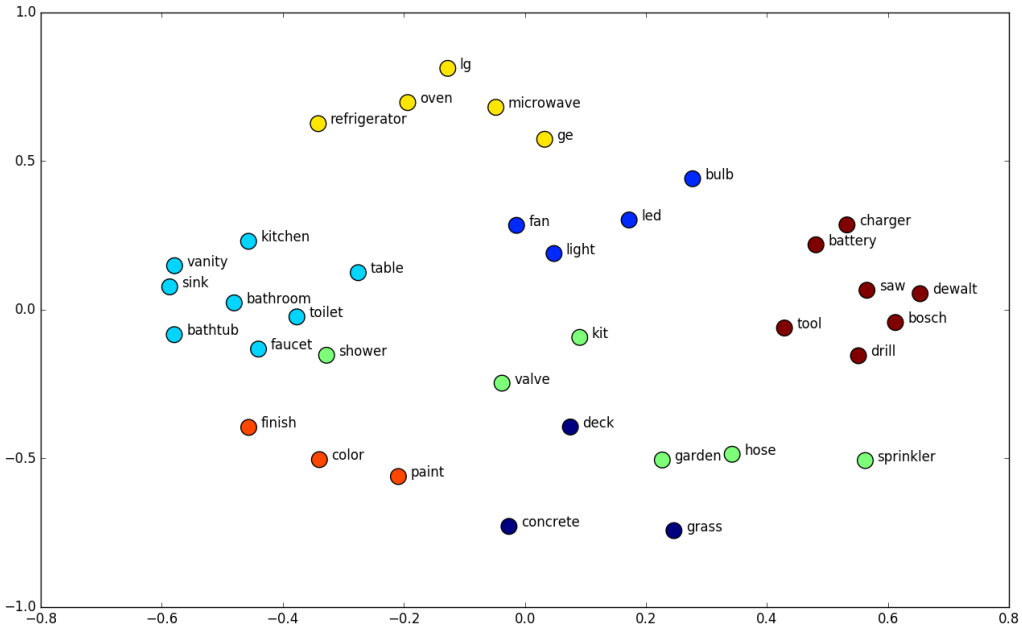
คำฝังตัวเป็นการแปลงภาษาธรรมชาติ (Natural Language) หรือภาษาที่มนุษย์ใช้สื่อสารให้กลายเป็นข้อมูลประเภทตัวเลข (Numerical) ที่อยู่ในรูปแบบของเวกเตอร์เพื่อใช้แสดงถึงความสัมพันธ์หรือความหมายที่ใกล้เคียงกันของคำ เนื่องจากภาษาธรรมชาติ เช่น คำว่า “กิน” กับ “วิ่ง” นั้น เราไม่สามารถทราบได้ว่า สองคำนี้มีความสัมพันธ์หรือมีความหมายที่ใกล้เคียงกันหรือไม่

Mikolov และคณะ [12] ได้ทำการพัฒนาการทำคำฝังตัวสำหรับแก้ปัญหาในการหาคุณลักษณะของคำที่ไม่คำนึงถึงไวยากรณ์ ด้วยการนำข้อมูลมาสร้างเป็นเวกเตอร์ของคำที่ได้จากการคำนวณตัวเลขจากบริบทของคำนั้นๆ แล้วเลือกเฉพาะคุณลักษณะที่สำคัญหรือเหมาะสมต่อต่อคลังคำศัพท์ อีกทั้งวิธีนี้จะทำให้ขนาดของเวกเตอร์ลดลง คำฝังตัวที่ได้จึงเป็นตัวแทนของเวกเตอร์แบบหนาแน่น (Dense Vector Representation) โดยในงานวิจัยนี้ได้ใช้คำฝังตัวที่ถูกฝึกสอนแล้วจากข้อมูลวิกิพีเดียภาษาไทยของ Thai2fit [14]

ตารางที่ 1 ตัวอย่างค่าคำฝังตัว

	ผู้ชาย	ผู้หญิง	ราชา	ราชินี	แตงโม	ส้ม
เพศ	-1	1	-0.95	0.97	0.00	0.01
ราชวงศ์	0.01	0.02	0.93	0.95	-0.01	0.00
อายุ	0.03	0.02	0.7	0.69	0.03	-0.02
อาหาร	0.04	0.01	0.02	0.01	0.95	0.97

คำฝังตัวสามารถสร้างได้จากการสร้างการเมทริกซ์ฝังตัว (Matrix Embedding) โดยใช้วิธีการสร้างเวกเตอร์ของคุณลักษณะ (Feature Vector) เช่น การสร้างเวกเตอร์ที่มีจำนวนของคุณลักษณะทั้งหมด 300 คุณลักษณะกับทุกๆ คำในคลังของคำศัพท์ ตัวอย่างจากตารางที่ 1 ได้แสดงค่าน้ำหนักของแต่ละคุณลักษณะในแต่ละคำศัพท์ สังเกตได้ว่าหากเป็นคุณลักษณะที่เกี่ยวกับเพศ เช่นคำว่า “ผู้ชาย” และ “ผู้หญิง” จะมีค่าน้ำหนักต่อคุณลักษณะเพศสูงกว่าคำอื่นๆ เช่นเดียวกันกับ“แตงโม” และ “ส้ม” ที่มีค่าน้ำหนักต่อคุณลักษณะอาหารสูงกว่าคำอื่นๆ เมื่อนำค่าน้ำหนักเหล่านี้วาดลงไปในปริภูมิที่มิติเท่ากับจำนวนเวกเตอร์ ดังรูปที่ 5 จะพบว่าคำที่มีความหมายคล้ายกันจะอยู่ในตำแหน่งใกล้เคียงกัน ซึ่งการวัดระยะทางสามารถวัดได้จากการหาระยะทางแบบยุคลิด (Euclidean Distance) การหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) หรือ การค่าความเหมือนโคไซน์ (Cosine Similarity)

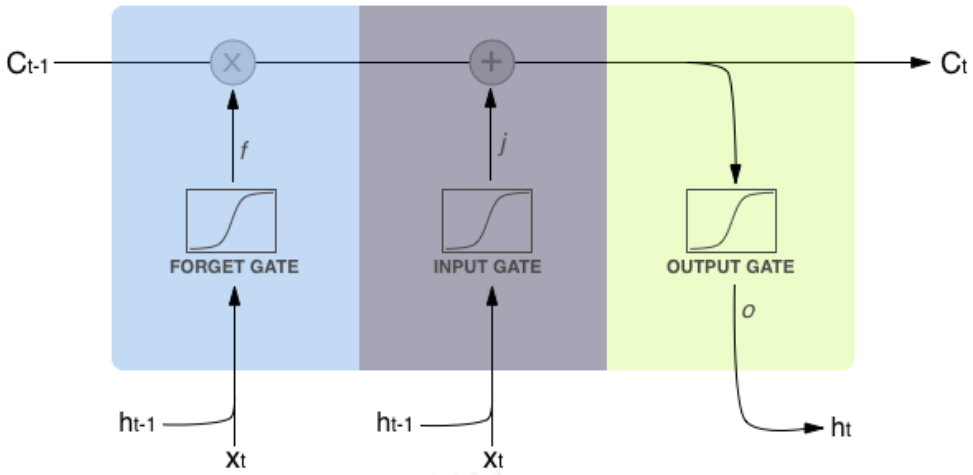


รูปที่ 5 ตัวอย่างปริภูมิสองมิติที่แสดงตำแหน่งของแต่คำละคำ โดยคำที่ความหมายคล้ายกันจะอยู่ในตำแหน่งที่ใกล้เคียงกัน [แหล่งอ้างอิง http://suriyadeepan.github.io]

2.2 การเรียนรู้ของเครื่อง (Machine Learning)

2.2.1 หน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM)

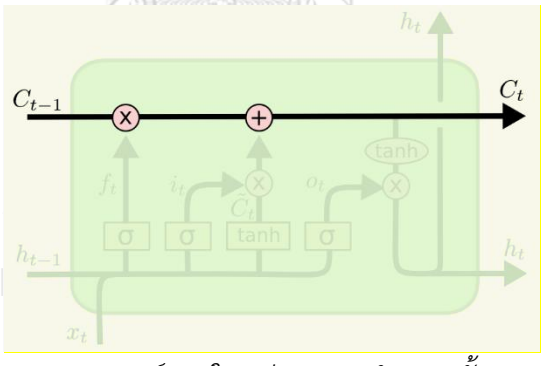
งานวิจัยของ Sepp Hochreiter และ Juergen Schmidhuber [13] ได้นำเสนอหน่วยความจำระยะสั้นแบบยาวเพื่อแก้ไขปัญหาวานิชชิงเกรเดียนต์ โดยความสามารถของหน่วยความจำระยะสั้นแบบยาวที่โดดเด่นกว่าเมื่อเปรียบเทียบกับนิวรอลเน็ตเวิร์กแบบวกกลับคือหน่วยความจำระยะสั้นแบบยาวสามารถเรียนรู้ได้ว่า เมื่อใดที่ควรจะเขียน (Write), ลืม (Forget) หรืออนุญาตให้อ่าน (Read) ได้สำหรับข้อมูลนำเข้า ทำให้สามารถเก็บข้อมูลในปริมาณมากขึ้น หน่วยความจำระยะสั้นแบบยาวมีโครงสร้างดังรูปที่ 6



รูปที่ 6 โครงสร้างของหน่วยความจำระยะสั้นแบบยาว (แหล่งอ้างอิง https://medium.com/@divyanshu132/lstm-and-its-equations-5ee9246d04af)

ส่วนประกอบหลักของหน่วยความจำระยะสั้นแบบยาวได้แก่

2.2.1.1 สถานะเซลล์ (Cell State) ทำหน้าที่เหมือนสายพานลำเลียงเพื่อแจกจ่ายข้อมูลและยังตัวเก็บสถานะของเซลล์ความจำ (Memory Cell) ในหน่วยความจำระยะสั้นแบบยาว มีโครงสร้างดังรูปที่ 7

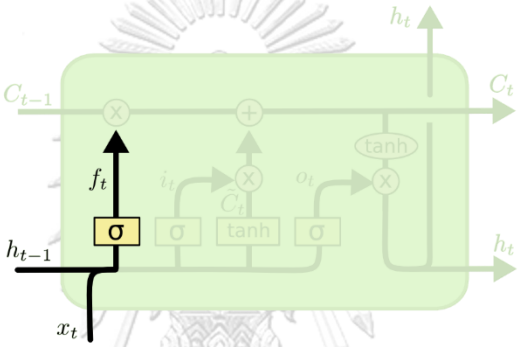


รูปที่ 7 ลักษณะของสถานะเซลล์ภายในหน่วยความจำระยะสั้นแบบยาว (แหล่งอ้างอิง https://colah.github.io/posts/2015-08-Understanding-LSTMs)

- ประตูควบคุมการทำงาน (Gate) ทำหน้าที่ควบคุมการทำงานของแบบจำลองที่จะกระทำกับข้อมูลนำเข้าที่ประกอบด้วย ประตูลืม (forget gate) ประตูนำเข้า (input gate) และประตูนำออก (output gate) ซึ่งแต่ละประตูจะมีฟังก์ชันกระตุ้น ได้แก่ (1) ฟังก์ชันซิกมอยด์ (Sigmoid Function) สามารถเขียนแทนด้วย σ โดยผลลัพธ์ที่ได้จากฟังก์ชันนี้จะมีค่าอยู่ระหว่าง 0 และ 1 เท่านั้น โดยค่า 0 หมายถึง ไม่มีข้อมูลใดผ่านออกจากประตู และ ค่า 1 หมายถึง ประตูจะปล่อยให้ข้อมูลถูกส่งออกไปได้ (2) ฟังก์ชันไฮเพอร์โบลิกแทนเจนต์

(Hyperbolic Tangent: tanh) หรือ ฟังก์ชันแทน ผลลัพธ์ที่ได้จากฟังก์ชันนี้จะมีค่าอยู่ระหว่าง -1 ถึง 1 ซึ่งฟังก์ชันนี้สามารถปรับปรุงข้อเสียของฟังก์ชันซิกมอยด์ได้ หากไม่มีการใช้ฟังก์ชันแทนร่วมกับฟังก์ชันซิกมอยด์แล้ว ค่าของผลลัพธ์ที่ได้จากการฝึกแบบจำลองจะเพิ่มสูงมากขึ้นตามจำนวนรอบในการฝึก ซึ่งประตูควบคุมการทำงานนั้นประกอบไปด้วย

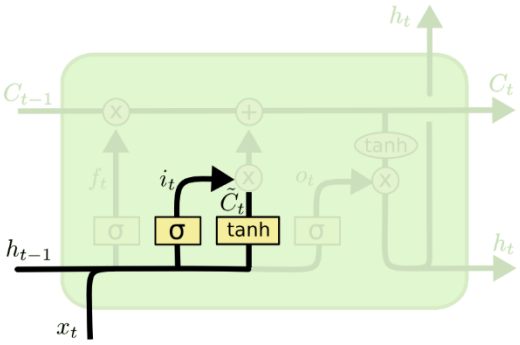
1. ประตูลืม (forget gate) ทำหน้าที่ตัดสินใจว่าควรเก็บหรือลบข้อมูลนำเข้านี้ และเมื่อข้อมูลนำเข้าปัจจุบันและค่าจากสถานะก่อนหน้าผ่านฟังก์ชันซิกมอยด์แล้วได้ผลลัพธ์เป็นค่าที่เข้าใกล้ 0 ประตูลืมจะลบค่าสถานะเซลล์เดิมออกไป หากค่าเข้าใกล้ 1 ประตูลืมจะเก็บค่าสถานะเซลล์นี้ไว้ โครงสร้างภายในของประตูลืมเป็นดังรูป 8 และมีสมการที่ใช้ในการคำนวณดังสมการที่ (1)



รูปที่ 8 ประตูลืมภายในหน่วยความจำระยะสั้นแบบยาว (แหล่งอ้างอิง https://colah.github.io/posts/2015-08-Understanding-LSTMs)

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (1)$$

2. ประตูนำเข้า (Input Gate) ทำหน้านี้คำนวณว่าจะทำการปรับสถานะของเซลล์เมื่อมีข้อมูลนำเข้าใหม่ให้เป็นปัจจุบันหรือไม่ หากมีการปรับจะใช้ฟังก์ชันซิกมอยด์ในการคำนวณค่าที่จะใช้ในการปรับ โดยใช้ทั้งข้อมูลนำเข้าปัจจุบันและค่าสถานะซ่อนก่อนการคำนวณผลลัพธ์ที่ได้เมื่อผ่านฟังก์ชันแล้วจะมีค่าอยู่ระหว่าง 0 ถึง 1 หากค่าที่ได้เป็น 1 จะแสดงถึงการที่ข้อมูลนำเข้าใหม่มีความสำคัญและควรทำการปรับสถานะเซลล์ให้เป็นปัจจุบัน หากค่าที่ได้เป็น 0 ก็ไม่มีความจำเป็นต้องปรับสถานะเซลล์ให้เป็นปัจจุบัน จากนั้นฟังก์ชันแทนจะทำหน้าที่จัดการกับผลลัพธ์ที่ได้หรือค่า $\tilde{C}_t$ ซึ่งเป็นตัวแทนของการนำผลลัพธ์ที่ได้ไปดัดแปลงอีกครั้ง ตามสถานะที่ถูกคำนวณมาก่อนหน้าเพื่อส่งเป็นข้อมูลนำออกต่อไป โครงสร้างภายในประตูนำเข้า แสดงที่รูป 9 และมีสมการที่ใช้ในการคำนวณดังสมการที่ (2) และ (3)



รูปที่ 9 ประตูนำเข้าภายในหน่วยความจำระยะสั้นแบบยาว (แหล่งอ้างอิง https://colah.github.io/posts/2015-08-Understanding-LSTMs)

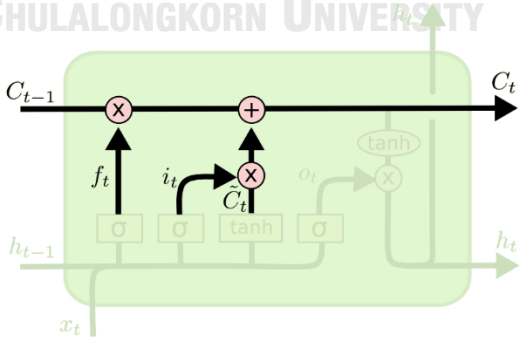
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

เมื่อได้ข้อมูลนำเข้าจากประตูลืมและประตูนำเข้าแล้ว ก็จะมีข้อมูลเพียงพอสำหรับการปรับค่าสถานะเซลล์ให้เป็นปัจจุบัน ซึ่งเขียนเป็นสมการคำนวณได้ตามสมการที่ (4)

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

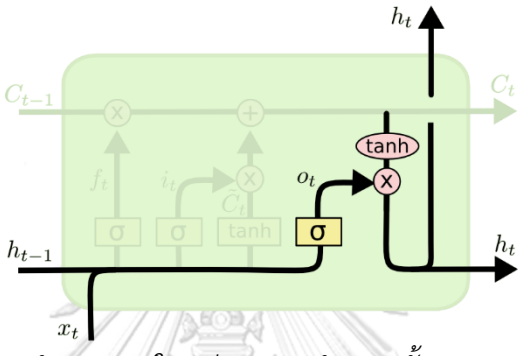
การคำนวณค่าสถานะเซลล์ประกอบด้วย 2 ส่วนดังรูปที่ (10) ส่วนที่ 1) หากค่าจากประตูลืมมีค่าเป็น 0 ค่าเซลล์ความจำก่อนหน้าหรือ $\mathbf{C_{t-1}}$ จะไม่ถูกเอามาพิจารณา แต่หากมีค่าเป็น 1 ค่า $\mathbf{C_{t-1}}$ จะถูกนำมาพิจารณาด้วย ส่วนที่ 2) การปรับค่าสถานะเซลล์จากข้อมูลนำเข้าใหม่ให้เป็นปัจจุบัน โดยหาก $\mathbf{i_t}$ มีค่าเป็น 1 แล้วค่า $\tilde{\boldsymbol{C}}_{\boldsymbol{t}}$ จะถูกนำมาปรับให้เป็นค่าปัจจุบัน เมื่อได้ค่าจากทั้งสองส่วนแล้ว ค่า $\mathbf{C_t}$ ที่ได้จะเป็นค่าที่ถูกปรับเป็นปัจจุบัน



รูปที่ 10 การคำนวณค่าสถานะเซลล์ใหม่ในประตูนำเข้า (แหล่งอ้างอิง https://colah.github.io/posts/2015-08-Understanding-LSTMs)

3. ประตูนำออก (output gate) ทำหน้าที่ตัดสินใจว่าสถานะซ่อนถัดไปควรมีลักษณะอย่างไร มีวิธีการได้แก่ นำค่าของสถานะซ่อนก่อนหน้าและข้อมูลนำเข้าปัจจุบันผ่านฟังก์ชัน

ซิกมอยด์ เพื่อให้ได้ค่าข้อมูลนำออก จึงนำค่า $\mathbf{C_t}$ ของสถานะเซลล์อันใหม่ที่มีการคำนวณตามสมการที่ (4) ผ่านฟังก์ชันแทน แล้วจึงนำมาคูณกับค่าของสมการที่ (5) หากค่าจากประตูนำออกหรือค่า $\mathbf{o_t}$ มีค่าเป็น 0 แล้ว ค่าของ $\mathbf{h_t}$ ก็จะมีค่าเป็น 0 เช่นกัน ซึ่งจะไม่มีการส่งค่าใดๆ ออกไป และในทางตรงข้าม หาก $\mathbf{o_t}$ มีค่าเป็น 1 ก็จะคำนวณค่า $\mathbf{h_t}$ ตามสมการที่ (5) ผลลัพธ์สุดท้ายที่ได้คือค่าใหม่ของสถานะเซลล์และสถานะซ่อน เพื่อใช้กับหน่วยความจำระยะสั้นแบบยาวในลำดับถัดไป โครงสร้างภายในแสดงดังรูปที่ 11



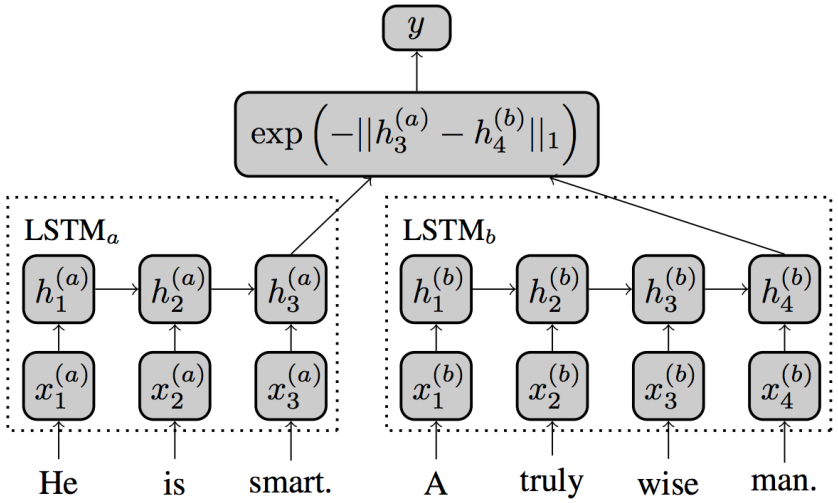
รูปที่ 11 ประตูนำออกภายในหน่วยความจำระยะสั้นแบบยาว (แหล่งอ้างอิง https://colah.github.io/posts/2015-08-Understanding-LSTMs)

$$h_t = o_t * tanh(C_t) \quad (5)$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \quad (6)$$

2.2.2 โครงข่ายสยาม (Siamese Neural Network)

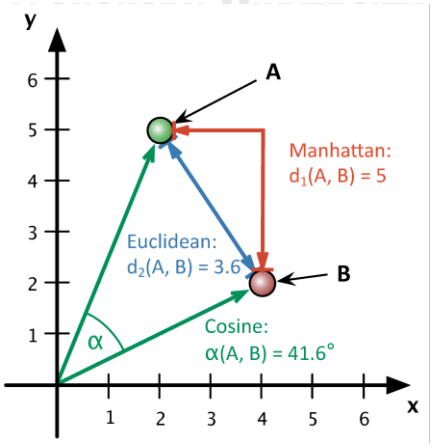
หมายถึงสถาปัตยกรรมของการเรียนรู้เชิงลึกที่มีโครงข่ายย่อยของแบบจำลองการเรียนรู้ของเครื่องจำนวนสองโครงข่ายที่เป็นคู่แฝดกัน คล้ายกับฝาแฝด อิน-จัน ซึ่งเป็นแฝดสยามที่มีชื่อเสียง โครงข่ายย่อยทั้งสองจะมีค่าน้ำหนัก (weight) และค่าความเอนเอียง (bias) ของแบบจำลองเหมือนกัน ในแต่ละรอบของการฝึกสอนนั้นแบบจำลองของทั้งสองโครงข่ายจะถูกปรับค่าน้ำหนักและค่าความเอนเอียงเหมือนๆ กัน จากนั้นผลลัพธ์ที่ได้จากโครงข่ายย่อยจะถูกนำไปคำนวณเพื่อหาระยะทางระหว่างเวกเตอร์ของทั้ง 2 ประโยค ได้แก่ การระยะทางแบบยุคลิด (Euclidean Distance) การหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) หรือ การค่าความคล้ายโคไซน์ (Cosine Similarity) ผลลัพธ์ที่ได้คือค่าความคล้ายกันของทั้งสองประโยค



รูปที่ 12 ตัวอย่างการใช้โครงข่ายสยามร่วมกันกับแบบจำลองหน่วยความจำระยะสั้นแบบยาว (LSTM) และการหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) เพื่อหาค่าความคล้ายของทั้งสองประโยค (อ้างอิงจาก Fig.1 ใน [6])

2.3 การหาระยะทางระหว่างเวกเตอร์ของคำ (Distance between vector of two words)

เป็นการคำนวณหาระยะทางของเวกเตอร์ของคำบนปริภูมิสองมิติ โดยคำที่มีความหมายคล้ายกันจะมีตำแหน่งบนปริภูมิสองมิติใกล้กัน ส่วนคำที่มีความหมายต่างกันจะมีระยะทางระหว่างเวกเตอร์ที่เพิ่มขึ้น ทำให้ระยะทางระหว่างเวกเตอร์ของคำสองคำนั้นมีผลโดยตรงกับค่าความคล้ายของคำ การวัดระยะทางสามารถวัดได้จากการหาระยะทางแบบยุคลิด (Euclidean Distance) การหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) หรือ การหาค่าความคล้ายโคไซน์ (Cosine Similarity)



รูปที่ 13 ตัวอย่างการหาระยะทางแบบต่างๆ บนปริภูมิสองมิติ ได้แก่ การระยะทางแบบยุคลิด (Euclidean Distance) การหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) หรือ การหาค่าความคล้ายโคไซน์ (Cosine Similarity) (แหล่งอ้างอิง https://dh2016.adho.org/abstracts/253)

2.3.1 การหาระยะทางแบบยุคลิด (Euclidean Distance)

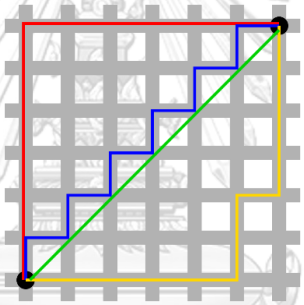
เป็นการหาระยะทางระหว่างจุดสองจุดในแนวเส้นตรงบนบนปริภูมิสองมิติ ยิ่งระยะทางที่ได้มีค่าน้อยมากเท่าไหร่ แสดงว่าเวกเตอร์ของคำทั้งสองคำนั้นมีความคล้ายกันมากขึ้นเท่านั้น สามารถเขียนสมการคำนวณได้ตามสมการที่ (7)

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \quad (7)$$

โดย d คือ ระยะทางระหว่างเวกเตอร์ A และ B

2.3.2 การหาระยะทางแบบแมนฮัตตัน (Manhattan Distance)

เป็นการหาผลรวมของระยะทางในแนวตั้งและแนวนอนระหว่างจุดสองจุดบนปริภูมิสองมิติ รูปแบบเส้นทางในการคำนวณระยะทางแบบแมนฮัตตันเป็นดังรูปที่ 14



รูปที่ 14 ตัวอย่างรูปแบบเส้นทางที่สามารถคำนวณระยะทางแบบแมนฮัตตัน (สีแดง สีน้ำเงิน และสีเหลือง) และระยะทางในแนวเส้นตรง (สีเขียว)
(แหล่งอ้างอิง https://en.wiktionary.org/wiki/Manhattan_distance)

ซึ่งในการหาระยะทางแบบแมนฮันตัน หากระยะทางที่คำนวณได้นั้นยิ่งมีค่าน้อยยิ่งแสดงถึงความคล้ายของเวกเตอร์ของคำทั้งสองคำที่มากขึ้น โดยสามารถเขียนเป็นสมการคำนวณได้ตามสมการที่ (8)

$$d(A,B) = \sum_{i=1}^{n}|A_i - B_i| \quad (8)$$

โดย d คือ ระยะทางระหว่างเวกเตอร์ A และ B

2.3.3 การหาค่าความคล้ายโคไซน์ (Cosine Similarity)

เป็นการคำนวณหาความคล้ายด้วยการคำนวณองศาระหว่างเวกเตอร์ของทั้งสองคำบนปริภูมิสองมิติ หากคำทั้งสองคำมีองศาของมุมยิ่งน้อยยิ่งแสดงถึงความคล้ายกันของเวกเตอร์ของคำทั้งสองที่มากขึ้น สามารถคำนวณได้ตามสมการที่ (9)

$$Similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (9)$$

โดย Similarity จะมีค่าอยู่ระหว่าง 0 ถึง 1

บทที่ 3
งานวิจัยที่เกี่ยวข้อง

หุ่นยนต์สนทนาสำหรับการตอบคำถามของปัญหาที่พบบ่อยมุ่งเน้นการจัดการกับปัญหาที่มีขอบเขตจำกัด (Closed Domain) ที่เกี่ยวกับการตอบคำถามที่พบบ่อยให้แก่ผู้สอบถาม โดยในหัวข้อนี้จะแบ่งงานวิจัยที่เกี่ยวข้องออกเป็น 2 กลุ่ม ได้แก่ การเข้าใจภาษาธรรมชาติและวิธีการตัดสินใจของระบบ

3.1 กลุ่มงานวิจัยสำหรับแก้ปัญหาในการตอบคำถามของปัญหาที่พบบ่อย

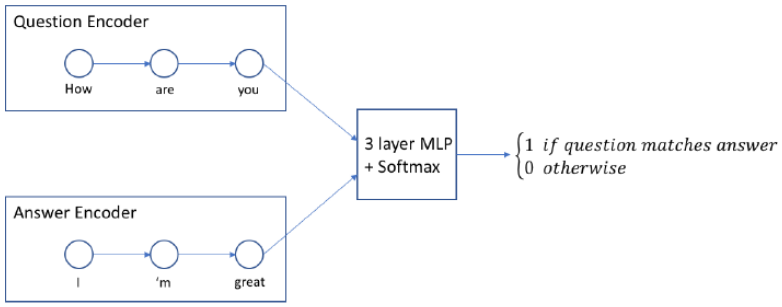
3.1.1 งานวิจัยของ Yichao Lu และคณะ

งานวิจัยนี้ [5] เกิดขึ้นในปี 2017 โดยมีเป้าหมายที่จะนำเสนอแบบจำลองเพื่อการโต้ตอบบทสนาทนาสำหรับการติดต่อศูนย์บริการลูกค้า (Customer Service) ของ Amazon เนื่องจากในทุกๆ ปี จะมีลูกค้านับล้านรายที่ติดต่อเข้ามายังศูนย์บริการลูกค้า ถึงแม้ว่า Amazon จะมีช่องทางในการติดต่อหลากหลายช่องทาง เช่น ทางโทรศัพท์ การสนทนาผ่านแอปพลิเคชัน หรือ อีเมล ดังนั้นการพัฒนาแบบจำลองในงานวิจัยนี้จะช่วยพัฒนาการให้บริการลูกค้าได้ดียิ่งขึ้น โดยแบบจำลองที่ใช้จะถูกฝึกสอนด้วยข้อมูลบทสนทนาระหว่างเจ้าหน้าที่ของ Amazon กับลูกค้าในปัญหาที่เกี่ยวข้องกับการขนส่งสินค้า โดยข้อมูลของผู้รับบริการจะถูกนำไปกรองข้อมูลสำคัญออก เช่น ข้อมูลเกี่ยวกับชื่อ หรือบัตรเครดิต เป็นต้น ข้อมูลฝึกสอนสำหรับแบบจำลองที่มีโครงสร้างแสดงตามรูปที่ 15 จะอยู่ในรูปแบบของคู่คำถาม-คำตอบโดยสามารถแบ่งได้เป็นสองประเภท ได้แก่ ‘0’ สำหรับคู่คำถาม-คำตอบที่ไม่ถูกต้องและ ‘1’ สำหรับคู่คำถาม-คำตอบที่ถูกต้อง ซึ่งข้อมูลของคู่คำถาม-คำตอบที่ไม่ถูกต้องนั้น ได้ใช้วิธีการนำคำถามมาจับคู่กับคำตอบของคำถามอื่นที่สุ่มมา จากนั้นจึงกำหนดประเภทให้เป็น ‘0’ ดังตัวอย่างในตารางที่ 2

ตารางที่ 2 ตัวอย่างข้อมูลฝึกที่อยู่ในรูปแบบคู่คำถาม-คำตอบ ที่ประกอบไปด้วยตัวอย่างที่ถูกต้อง (Label: ‘1’) และตัวอย่างที่ไม่ถูกต้อง (Label: ‘0’) (อ้างอิงจากตารางที่ 1 ใน [5])

Customer Inquiry	Agent Response	Label
and will i be sent an email ?	yes , NAME .	1
can the ship speed be changed ?	yes , i ’ve already upgraded .	1
ok so what i have to do now ?	it ’s a good company to work for	0
can I ask for a resend ?	both the orders will be delivered to you today .	0

จากนั้นจะข้อมูลคู่คำถาม-คำตอบที่เตรียมไว้ไปใช้กับแบบจำลองที่มีโครงสร้างตามรูปที่ 15 ซึ่งได้แนวคิดมาจากโครงข่ายสยาม (Siamese Neural Network)



รูปที่ 15 ตัวอย่างโครงสร้างของการเข้ารหัสประโยคคำถามและประโยคคำตอบด้วยแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) และส่งข้อมูลนำเข้าไปยังโครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron: MLP) (อ้างอิงจาก Fig.1 ใน [5])

โดยแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ทั้งสองตามรูปที่ 15 จะมีข้อมูลนำเข้าเป็นคำถามจากลูกค้า เช่น “Will I receive a new tracking number?” และคำตอบ เช่น “Yes we’ll have it emailed to you.” แบบจำลองทั้งสองจะเข้ารหัสและสร้างคำฝังตัวใหม่ที่มิติขนาดเล็ก จากนั้นคำฝังตัวทั้งสองจะถูกนำมาต่อกันแล้วส่งต่อไปเป็นข้อมูลนำเข้าของโครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron: MLP) และข้อมูลนำออกที่ได้จะเป็นค่าความน่าจะเป็นว่าคู่คำถาม-คำตอบที่เป็นข้อมูลนำเข้านั้นถูกต้องหรือไม่ โดยอัตราส่วนข้อมูลของจำนวนคู่คำถาม-คำตอบที่ถูกต้องต่อจำนวนคู่คำถาม-คำตอบที่ไม่ถูกต้องนั้น มีอัตราส่วนที่ 1:2 สำหรับชุดข้อมูลพัฒนา (Validation set) สามารถทำให้แบบจำลองมีความถูกต้องสูงสุดถึง 81%

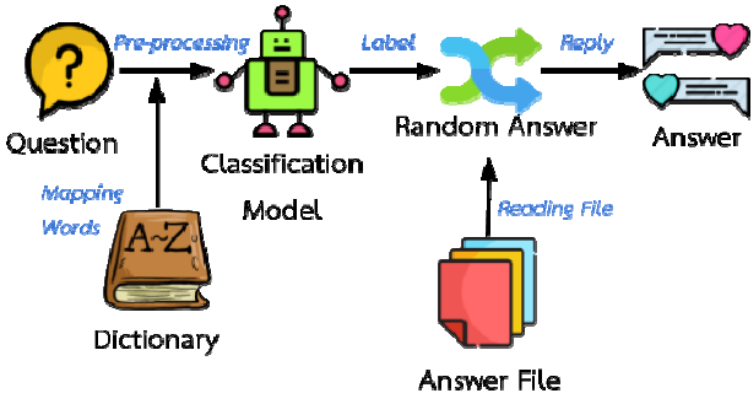
ตารางที่ 3 เป็นตารางแสดงผลลัพธ์จากแบบจำลองที่นำเสนอในงานวิจัยนี้ โดยแบบจำลองจะให้ผลลัพธ์เป็นคำตอบของคำถามที่ผู้รับบริการได้ทำการสอบถาม โดยมีความน่าจะเป็นที่คำตอบนั้นจะเป็นคำตอบของคำถามที่มีค่าสูงสุด 3 อันดับแรก ซึ่ง Yichao Lu และคณะมีแผนที่จะพัฒนาโดยการนำผลลัพธ์ที่ได้จากแบบจำลองนี้ ไปใช้กับระบบให้บริการข้อมูลแก่ลูกค้า เช่น การสนทนาออนไลน์บนแอปพลิเคชัน หรือ การสนทนาทางโทรศัพท์เพื่อให้บทสนทนานั้นมีความเป็นธรรมชาติมากขึ้น

ตารางที่ 3 ตัวอย่างคำถามจากลูกค้าและรูปแบบของคำตอบที่แนะนำจากแบบจำลองที่นำเสนอในงานวิจัย (อ้างอิงจากตารางที่ 4 ใน [5])

Question	Top 3 recommended answers
when will i receive my shoes ?	it will be delivered DATE you will get the items on DATE you 'll receive the package within 24 hours .
how can i use the gift card balance ?	you can use it on your next purchase . you can use after 2 hours . because it will take only 1-2 hours to credit in your account . the refund will be reflected in your gift card balance in the next 1-3 hour
hi are you there ?	yes I 'm here . yes , i 'm checking it . sorry for the delay in responding
can i cancel the order ?	i can cancel it for you . i 've cancelled it . which items you need to cancel ?
why it has n't been shipped yet ?	i am glad to check the status of your order . your order is already entered to the shipping process . it is out of stock.

3.1.2 งานวิจัยของ Panitan Muangkammuen และคณะ

งานวิจัยนี้ [8] เกิดขึ้นในปี 2018 โดยมีเป้าหมายที่จะนำเสนอหุ่นยนต์สนทนาของคำถามที่พบบ่อยแบบอัตโนมัติที่จะช่วยพัฒนาการให้บริการในการตอบคำถามแก่ผู้รับบริการที่ติดต่อเข้ามาในหลากหลายช่องทาง เช่น การสนทนาออนไลน์บนแอปพลิเคชัน หรือ การส่งอีเมล ผู้รับบริการส่วนใหญ่มักจะถามคำถามผ่านการสนทนาออนไลน์บนแอปพลิเคชันเพราะเป็นช่องทางที่สะดวกและรวดเร็ว ดังนั้นบริษัทจึงต้องจ้างผู้ดูแลสำหรับการบริการตอบคำถามผ่านช่องทางเหล่านี้ แต่ผู้ดูแลก็ต้องใช้เวลาในการหาคำตอบและผู้รับบริการก็ต้องใช้เวลาในการรอคำตอบเช่นเดียวกัน ดังนั้นหุ่นยนต์สนทนาของคำถามที่พบบ่อยแบบอัตโนมัติจึงสามารถช่วยให้คำตอบแก่ผู้รับบริการทันทีที่ได้รับคำถาม ซึ่งในงานวิจัยนี้ได้ใช้แบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) มาใช้ในการจำแนกประเภทของข้อมูลเชิงตัวอักษรที่เป็นภาษาไทย โดยมีภาพรวมของวิธีการที่นำเสนอตามรูปที่ 16



รูปที่ 16 ภาพรวมของระบบหุ่นยนต์สนทนาของคำถามที่พบบ่อยแบบอัตโนมัติที่นำเสนอ (อ้างอิงจาก Fig.1 ใน [8])

ในงานวิจัยนี้ใช้จำนวนข้อมูลเชิงตัวอักษรที่มีรูปแบบเป็นคู่ของคำถาม-คำตอบทั้งหมด 2,636 คู่และถูกจำแนกออกเป็นประเภทต่างๆ รวมทั้งสิ้น 80 ประเภท โดยจำแนกตามกลุ่มของคำถามที่พบบ่อยและแต่ละกลุ่มจะถูกกำหนดประเภทด้วยตัวเลข จากนั้นจะแยกคำถามและคำตอบออกจากกัน ในส่วนของคำถามจะถูกนำไปใช้ในการเรียนรู้ของแบบจำลองและคำตอบจะถูกนำไปเตรียมไว้สำหรับเป็นคำตอบให้แก่ลูกค้า ตัวอย่างของคู่คำถาม-คำตอบแสดงดังรูปที่ 17

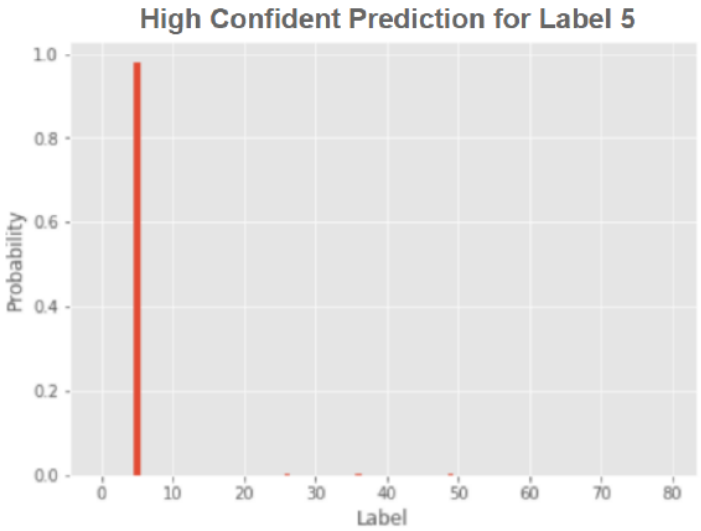
Q1: พี่ค่ะเราสามารถเปลี่ยนเลขหน้าบัตรได้ไหมคะ
(Can I change the card number?)
A1: ไม่สามารถเปลี่ยนเลขได้ยกเว้นกรณีสมัครใหม่เท่านั้นครับ
(Cannot change the number unless you register a new one)

Q2: โอนเงินจาก แอพ ไป ธนาคารได้ไหมคับ
(Can I transfer money from the application to a bank account?)
A2: สามารถทำได้ครับ โดยเข้าไปที่ Setting และเลือกที่ถอนเงินครับผม
(It can be done by going to the setting and select the withdraw menu)

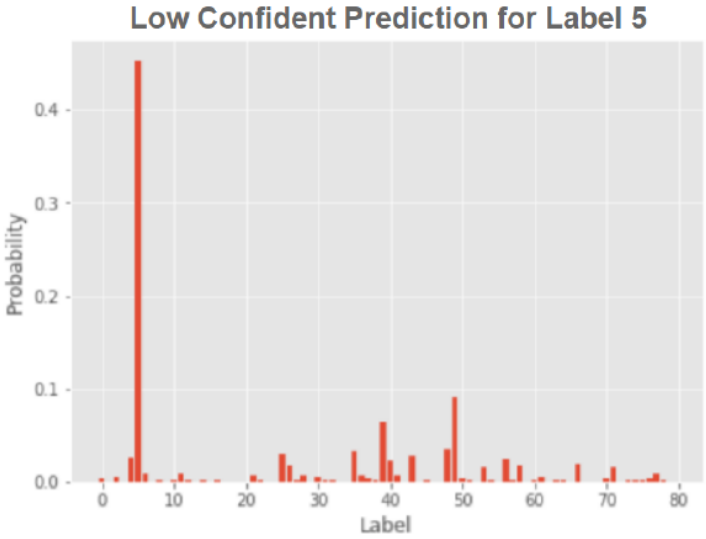
รูปที่ 17 ตัวอย่างข้อมูลเชิงอักษรที่เป็นคู่ของคำถาม-คำตอบ (อ้างอิงจาก Fig.2 ใน [8])

หลังจากข้อมูลคำถามถูกจำแนกเป็นประเภทต่างๆ แล้ว ข้อมูลคำถามจำนวน 2,636 คำถาม จะถูกแบ่งออกเป็นชุดข้อมูล 3 ชุด ประกอบด้วย (1) ชุดข้อมูลฝึกสอนจำนวน 60% (1,581 คำถาม) (2) ชุดข้อมูลพัฒนา (Validation set) จำนวน 20% (527 คำถาม) และ (3) ชุดข้อมูลทดสอบจำนวน 20% (528 คำถาม) ส่วนข้อมูลนำออกที่ได้จากแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ที่ใช้ในงานวิจัยนี้ จะถูกคำนวณด้วยฟังก์ชันค่าสูงสุดอย่างอ่อน (Softmax Activation Function) และได้ผลลัพธ์ออกมาเป็นค่าของความน่าจะเป็น (Probability) ในแต่ละประเภทของข้อมูลคำถามที่ได้แบ่งไว้ เพราะในกระบวนการเรียนรู้ของแบบจำลองนั้น Panitan Muangkammuen และคณะได้ทำการเข้ารหัสตัวเลขที่บ่งบอกถึงประเภทของข้อมูลคำถามไว้อยู่ในรูปแบบของเวกเตอร์วันฮอท (One-hot Vector) และใช้ในกระบวนการเรียนรู้ของแบบจำลองด้วย ตัวอย่างของค่าความน่าจะเป็นที่ได้ในแต่ละประเภทของกลุ่มคำถามหลังผ่านการ

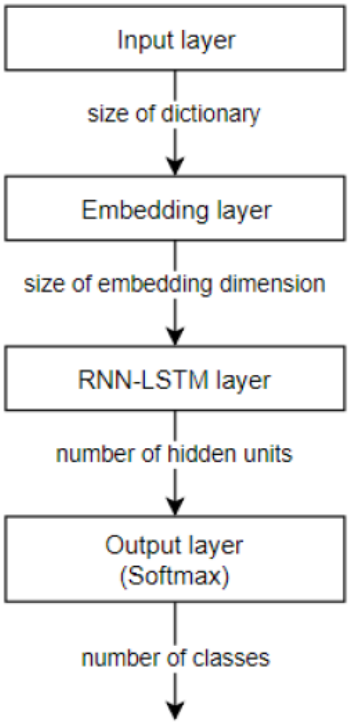
จำแนกประเภทด้วยแบบจำลองที่นำเสนอในงานวิจัยนี้ เป็นดังรูปที่ 18 และ 19 และโครงสร้างในแต่ละชั้นและมิติของโครงข่ายประสาทเทียมที่ใช้ส่งผ่านแสดงดังรูปที่ 20



รูปที่ 18 ตัวอย่างค่าความน่าจะเป็นสำหรับการทำนายของแบบจำลองที่จำแนกประเภทของข้อมูลกลุ่มที่ 5 โดยมีค่าความน่าจะเป็นสูงสุดคือ 0.97 (อ้างอิงจาก Fig.4 ใน [8])



รูปที่ 19 ตัวอย่างค่าความน่าจะเป็นสำหรับการทำนายของแบบจำลองที่จำแนกประเภทของข้อมูลกลุ่มที่ 5 โดยมีค่าความน่าจะเป็นต่ำสุดคือ 0.45 (อ้างอิงจาก Fig.5 ใน [8])



รูปที่ 20 โครงสร้างในแต่ละชั้นและมิติของโครงข่ายประสาทเทียมที่ใช้ส่งผ่านข้อมูล (อ้างอิงจาก Fig.6 ใน [8])

ประสิทธิภาพที่ได้จากแบบจำลองที่นำเสนอในงานวิจัยนี้เมื่อประเมินด้วยชุดข้อมูลทดสอบที่เตรียมไว้นั้น พบว่ามีค่าความถูกต้อง 83.9% โดยข้อมูลนำออกที่ได้จากชั้นสุดท้ายของโครงข่ายประสาทเทียมจะเป็นค่าความน่าจะเป็นที่กระจายไปตามประเภทที่ได้จำแนกไว้และประเภทที่จะถูกเลือกให้เป็นข้อมูลนำออกนั้นจะเป็นประเภทที่มีค่าความน่าจะเป็น (Probability) สูงที่สุด จากการทดลองพบว่าค่าเฉลี่ยของความน่าจะเป็นสูงสุดของประเภทของข้อมูลที่ถูกต้องนั้นอยู่ที่ 0.92 และค่าเฉลี่ยของความน่าจะเป็นของประเภทของข้อมูลที่ไม่ถูกต้องอยู่ที่ 0.48 ดังนั้นในงานวิจัยนี้จึงกำหนดค่าแบ่งของความน่าจะเป็นในการเลือกข้อมูลนำออกที่ 0.50 จากนั้นทำการประเมินผลแบบจำลองอีกครั้งด้วยข้อมูลทดสอบชุดเดิมและใช้ค่าแบ่งของความน่าจะเป็นด้วยค่า 0.5 พบว่า 13.64% ของคำถามจะไม่มีคำตอบและ 86.36% ของคำถามนั้นจะพบคำตอบและได้คำตอบที่ถูกต้องถึง 93.2%

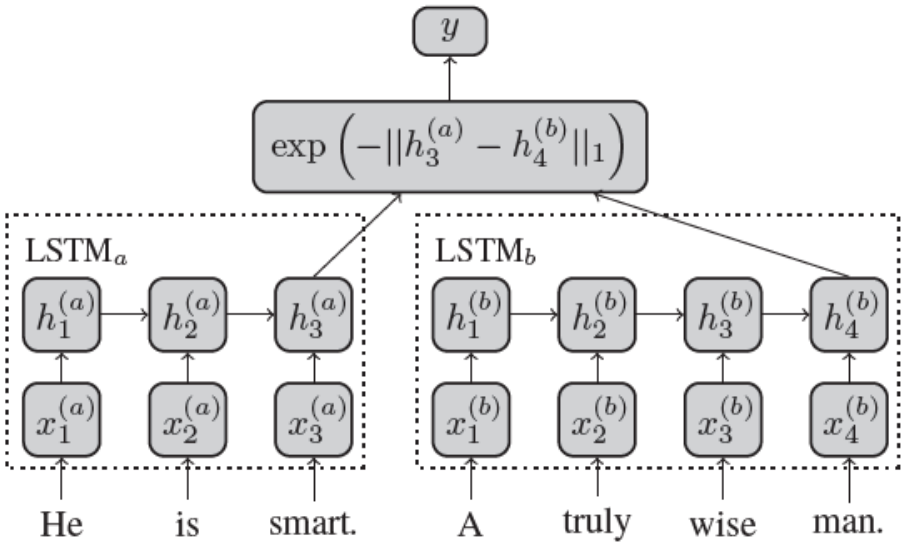
3.2 กลุ่มงานวิจัยที่เกี่ยวกับโครงข่ายสยามและการหาความคล้ายของข้อความ

3.2.1 งานวิจัยของ Jonas Mueller และ Aditya Thyagarajan

งานวิจัยนี้ [5] เกิดขึ้นในปี 2016 โดยนำเสนอแนวความคิดในการนำโครงข่ายสยาม (Siamese Neural Network) มาประยุกต์ใช้กับแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long

Short-Term Memory: LSTM) เพื่อใช้กับข้อมูลที่มีคำตอบ (Labeled Data) ที่เป็นประโยคที่คู่กันและมีความยาวของประโยคที่หลากหลาย ซึ่งแนวคิดในงานวิจัยนี้เป็นวิธีการล้ำสมัย (State of the art) ในการหาความคล้ายกันระหว่างประโยคสองประโยค โดยจะใช้เวกเตอร์ของคำฝังตัวร่วมกับข้อมูลที่บ่งบอกความเหมือนในการเป็นข้อมูลนำเข้าของแบบจำลองหน่วยความจำระยะสั้นแบบยาวที่มีข้อจำกัดในเรื่องของขนาดของเวกเตอร์ที่จะเข้ารหัสเพื่อแสดงถึงลักษณะพื้นฐานที่ประโยคนั้นๆ (โดยไม่คำนึงถึงคำเฉพาะ) และใช้การหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) มาใช้ในการคำนวณหาความคล้ายจากข้อมูลนำออกของตัวแทนประโยคที่ได้จากแบบจำลอง

โดยแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ที่ใช้ร่วมกับการหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) ในงานวิจัยนี้จะเรียกว่า Manhattan LSTM (MaLSTM) และมีภาพรวมของโครงสร้างดังรูปที่ 21 ซึ่งประกอบด้วยแบบจำลองหน่วยความจำระยะสั้นแบบยาวจำนวน 2 แบบ ได้แก่ $LSTM_a$ และ $LSTM_b$ โดยแต่ละแบบจำลองจะได้รับข้อมูลนำเข้าเป็นประโยคที่ถูกกำหนดให้เป็นคู่กัน และจะใช้ค่าน้ำหนักร่วมกันตามรูปแบบของโครงข่ายสยาม ทำให้ $LSTM_a$ และ $LSTM_b$ มีค่าน้ำหนักเท่ากัน ซึ่งแบบจำลองทั้งสองจะถูกใช้เพื่ออ่านเวกเตอร์ของคำที่เป็นตัวแทนของประโยคข้อมูลนำเข้าและได้ข้อมูลนำออกจากแบบจำลองทั้งสองเป็นเวกเตอร์ที่เป็นตัวแทนของประโยคนั้นๆ จากนั้นจะถูกส่งต่อไปเป็นข้อมูลนำเข้าในการหาความคล้ายระหว่างตัวแทนเวกเตอร์ทั้งสองและได้ข้อมูลนำออกเป็นค่าความคล้าย



รูปที่ 21 ตัวอย่างแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ที่ทำงานร่วมกับการความคล้ายของประโยคด้วยการหาระยะทางแบบแมนฮัตตัน (Manhattan

Distance) โดยใช้ภาพรวมของโครงสร้างแบบโครงข่ายสยาม (Siamese Neural Network) (อ้างอิงจาก Fig.1 [5])

แบบจำลอง Manhattan LSTM (MaLSTM) ที่นำเสนอในงานวิจัยนี้นั้นมีผลการทดสอบด้วยหน่วยประเมินผลทั้ง 3 แบบดังตารางที่ 4 ซึ่งแบบจำลองที่นำเสนอสามารถหาความสัมพันธ์ระหว่างข้อมูลนำเข้าได้ดี จึงทำให้มีค่าของค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson's correlation : r) และ ค่าสัมประสิทธิ์สหสัมพันธ์แบบสเปียร์แมน (Spearman Rank Correlation : ρ) สูงกว่าแบบจำลองอื่นๆ

ตารางที่ 4 ผลการทดสอบด้วยการคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson's correlation : r), ค่าสัมประสิทธิ์สหสัมพันธ์แบบสเปียร์แมน (Spearman Rank Correlation : ρ) และค่าเฉลี่ยความผิดพลาดกำลังสอง (Mean Square Error: MSE) สำหรับชุดข้อมูล SICK ที่เป็นชุดข้อมูล

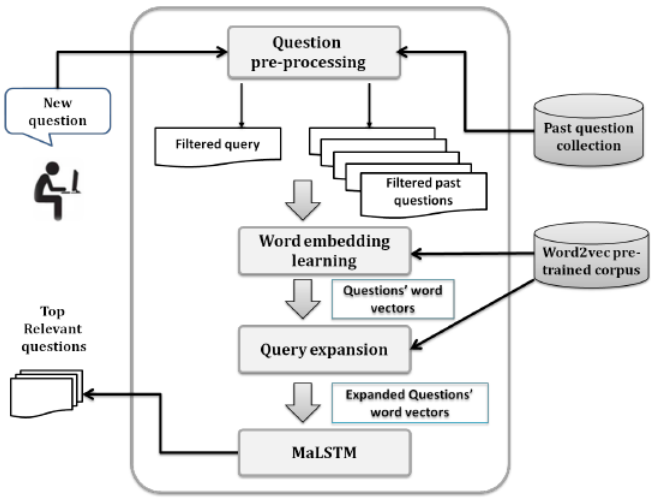
Method	r	ρ	**MSE**
Illinois-LH (Lai and Hockenmaier 2014)	0.7993	0.7538	0.3692
UNAL-NLP (Jimenez et al. 2014)	0.8070	0.7489	0.3550
Meaning Factory (Bjerva et al. 2014)	0.8268	0.7721	0.3224
ECNU (Zhao, Zhu, and Lan 2014)	0.8414	–	–
Skip-thought+COCO (Kiros et al. 2015)	0.8655	0.7995	0.2561
Dependency Tree-LSTM (Tai, Socher, and Manning 2015)	0.8676	0.8083	0.2532
ConvNet (He, Gimpel, and Lin 2015)	0.8686	0.8047	0.2606
MaLSTM	**0.8822**	**0.8345**	**0.2286**

ในงานวิจัยนี้ได้เสนอแนวคิดการใช้แบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) อย่างง่ายมาใช้ในการแก้ปัญหาที่ซับซ้อนอย่างการหาความหมายของคำ โดยกำหนดให้ตัวแทนของคำเหล่านั้นถูกจำแนกประเภทอย่างถูกต้อง จากการวิเคราะห์การเรียนรู้ของแบบจำลองทำให้พบว่าประโยคที่มีลักษณะต่างกันก็จะถูกเข้ารหัสไม่เหมือนกันทำให้มีความหลากหลายเกิดขึ้นในชั้นซ่อนของแบบจำลอง อีกทั้งประสิทธิภาพของแบบจำลองที่นำเสนอในเรื่องของเวลาที่ใช้นั้นเพียงพอที่จะนำไปใช้เป็นแอปพลิเคชันได้จริง ซึ่งแนวคิดที่นำเสนอในงานวิจัยนี้ขึ้นอยู่

กับเวกเตอร์ของคำที่ถูกฝึกสอนไว้แล้วที่เป็นข้อมูลนำเข้าของแบบจำลองที่เป็นผลจากงานวิจัยที่เกี่ยวข้องกับคำฝังตัว

3.2.2 งานวิจัยของ Nouha Othman และคณะ

งานวิจัยนี้ [7] เกิดขึ้นในปี 2019 เพื่อนำเสนอแนวคิดเกี่ยวกับการค้นคืนคำถามโดยใช้แบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ที่ใช้ร่วมกับการหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) ที่มีโครงสร้างแบบโครงข่ายสยาม (Siamese Neural Network) โดยจะค้นคืนคำถามที่เคยถูกถามแล้วและทำการหาความคล้ายกับคำถามใหม่ที่ได้รับเป็นข้อมูลนำเข้า ซึ่งคำตอบของคำถามที่เคยถูกถามแล้วจะสามารถนำมาใช้กับคำถามใหม่ได้หากพบว่าคำถามทั้งสองมีความคล้ายกัน อีกทั้งความท้าทายของงานวิจัยนี้คือ ปัญหาความสั้นของคำถามและคำศัพท์ที่ไม่เหมือนกัน ซึ่งผู้ใช้สามารถถามคำถามใหม่ที่มีความหมายคล้ายกันแต่ใช้คำที่แตกต่างกันได้ โดยนำเสนอการใช้คำฝังตัวที่สามารถจะจับลักษณะของคำที่มีความหมายคล้ายกันในบริบทนั้นๆ ได้แล้วทำให้เป็นเวกเตอร์ของประโยคคำถาม จากนั้นนำเวกเตอร์ของประโยคคำถามไปเป็นข้อมูลนำเข้าของแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ที่มีโครงสร้างแบบโครงข่ายสยาม (Siamese Neural Network) และวัดความคล้ายระหว่างคำถามด้วยการหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) โดยภาพรวมของระบบที่นำเสนอนั้นเป็นไปตามรูปที่ 22



รูปที่ 22 ตัวอย่างภาพรวมของแบบจำลองที่นำเสนอ (LSTMQR) สำหรับการค้นคืนคำถาม

ในงานวิจัยนี้ยังได้นำเสนอแนวคิดในการแก้ไขปัญหาประโยคคำถามที่สั้นจนทำให้ความหลากหลายของคำในประโยคนั้นน้อยลงและนำไปสู่การค้นคืนข้อมูลที่ไม่ถูกต้องเนื่องจากไม่สามารถ

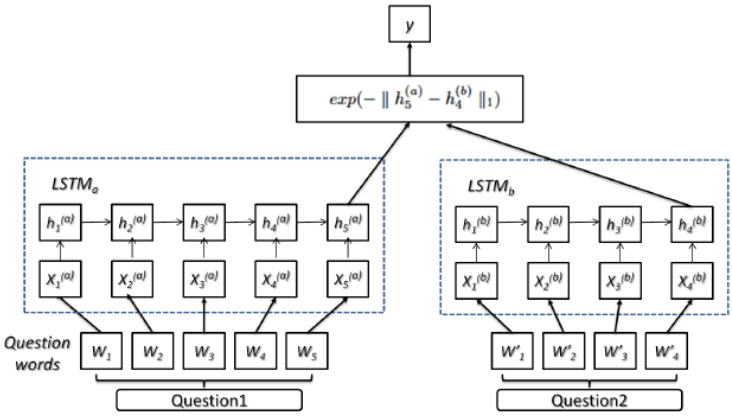
หาคำที่เหมือนกันได้ ซึ่งถือเป็นอีกหนึ่งความท้าทายที่งานวิจัยนี้ได้กล่าวถึง โดยการนำเสนอวิธีขยายความยาวของประโยค (Question Expansion) ด้วยการเพิ่มคำที่มีเวกเตอร์ของคำที่คล้ายกัน จำนวนคำที่จะเพิ่มเข้าไปในประโยคนั้นจะถูกกำหนดให้เป็นตัวแปร N_{sw} และจะเป็นจำนวนของคำที่มีเวกเตอร์ของคำที่คล้ายกันจากรายการคำศัพท์ จากนั้นคำที่มีเวกเตอร์คล้ายกันจะถูกเพิ่มด้วยการนำไปต่อท้ายที่ประโยคคำถามเดิมโดยไม่ทำให้ลำดับของคำเปลี่ยนแปลงไป เช่น

ประโยคคำถาม : Do chocolate really kill my dog?

จำนวนคำศัพท์ที่เพิ่ม (N_{sw}) : 3

ประโยคหลังจากขยายความยาว : chocolate kill dog eat death bitch candy toxic puppy food sick animal

โดยรูปแบบของโครงสร้างและประเภทของแบบจำลองที่ใช้ในงานวิจัยนี้ อ้างอิงจากงานวิจัยของ Jonas Mueller และ Aditya Thyagarajan [5] มีลักษณะโครงสร้างโดยทั่วไปดังรูปที่ 23 และใช้การหาความคล้ายของประโยคด้วยการหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) สามารถเขียนเป็นสมการคำนวณได้ตามสมการที่ (10)



รูปที่ 23 ลักษณะโครงสร้างโดยทั่วไปของแบบจำลองหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ที่ใช้ร่วมกับการหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) ที่เรียกว่า แบบจำลอง MaLSTM

$$y = exp\left(-\left\|h^{(left)} - h^{(right)}\right\|_1\right) \qquad (10)$$

แบบจำลองที่นำเสนอถูกนำไปเปรียบเทียบประสิทธิภาพกับแบบจำลองอื่นๆ ดังตารางที่ 5 ด้วยการวัดประสิทธิภาพของแบบจำลองทั้งหมด 3 แบบได้แก่ ค่าความแม่นยำเฉลี่ย (Mean average

precision : MAP) , ค่าความแม่นยำที่ 5 อันดับแรก (Precision@5 : P@5) และ ค่าความแม่นยำที่ 10 อันดับแรก (Precision@10 : P@10) พบว่าแบบจำลอง (LSTMQR) ที่ Nouha Othman และคณะนำเสนอนั้น มีประสิทธิภาพดีกว่าแบบจำลองอื่นๆ ในทุกการวัดประสิทธิภาพ

ตารางที่ 5 ประสิทธิภาพของแบบจำลองต่างๆ ในการค้นคืนคำถามของชุดข้อมูลภาษาอังกฤษ (อ้างอิงจากตารางที่ 3 ใน [7])

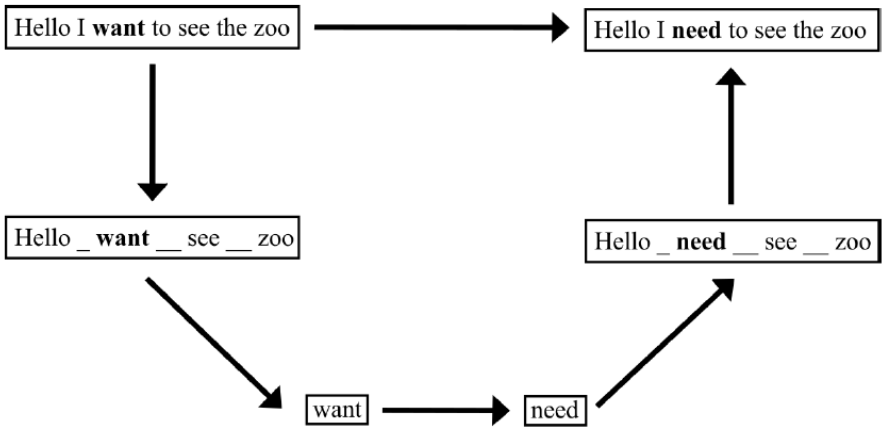
	TLM	ETLM	PBTM	WKM	M-NET	ParaKCM	WEKOS	LSTMQR
P@5	0.3238	0.3314	0.3318	0.3413	0.3686	0.3722	0.4338	**0.5023**
P@10	0.2548	0.2603	0.2603	0.2715	0.2848	0.2889	0.3647	**0.4188**
MAP	0.3957	0.4073	0.4095	0.4116	0.4507	0.4578	0.5036	**0.5739**

3.3 กลุ่มงานวิจัยที่เกี่ยวข้องกับการแต่งเติมข้อมูล

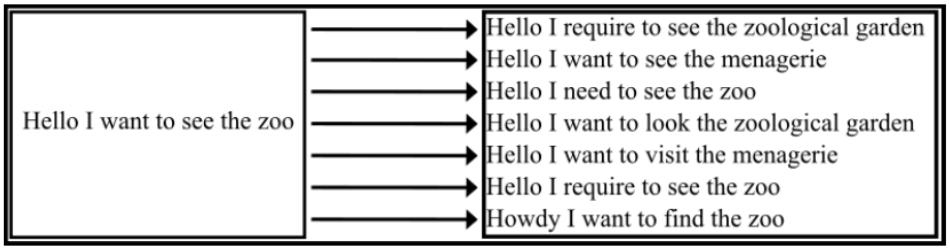
3.3.1 งานวิจัยของ Anna V. Mosolova และคณะ

งานวิจัยนี้ [9] เกิดขึ้นในปี 2018 เพื่อนำเสนอแนวทางในการแก้ปัญหาการเรียนรู้ของโครงข่ายประสาทเทียมด้วยการใช้ข้อมูลเชิงตัวอักษรที่มีขนาดเล็ก ด้วยการนำเอาแนวคิดจากวิธีการแต่งเติมข้อมูลในข้อมูลรูปภาพและข้อมูลเสียงให้มีจำนวนเพิ่มขึ้นเพื่อช่วยในการเพิ่มประสิทธิภาพในการเรียนรู้ของแบบจำลอง มาประยุกต์ใช้กับข้อมูลเชิงอักษรด้วยการแต่งเติมข้อมูลเชิงอักษรด้วยการใช้คำที่มีความหมายเหมือนกัน

วิธีการแต่งเติมข้อมูลเชิงตัวอักษรที่นำเสนอด้วยใช้การแทนคำที่มีความหมายเหมือนกันเข้าไปในคำเดิมที่มีอยู่แล้วในประโยคนั้นๆ โดยที่ยังไม่สูญเสียความหมายของประโยค จะเป็นการแทนคำที่ไม่ใช่คำจำพวกคำลักษณะนาม คำเชื่อม คำบุพบท หรือ คุณศัพท์ และคำที่มีความหมายเหมือนกันที่จะถูกนำมาแทนในประโยคนั้นจะถูกสุ่มเลือกซึ่งขึ้นอยู่กับการตั้งค่าเปอร์เซ็นต์ของจำนวนคำที่จะถูกแทน ตัวอย่างเช่น ประโยคที่ประกอบไปด้วยคำทั้งหมด 10 คำและการตั้งค่าเปอร์เซ็นต์ของจำนวนคำที่จะถูกแทนในการแต่งเติมข้อมูลไว้ที่ 25 เปอร์เซ็นต์ อัลกอริทึมก็จะแทนคำจำนวน 2 คำสำหรับทุกคำที่สามารถหาคำที่มีความหมายเหมือนกันในประโยคนี้ และคำที่สุ่มมานั้นจะถูกนำไปแทนกับคำที่มีความหมายเหมือนกันที่ตำแหน่งเดิม ดังตัวอย่างในรูปที่ 24 และในอัลกอริทึมจะมีตัวแปรที่กำหนดจำนวนรวมของประโยคที่ขยายได้ เช่นตัวอย่างในรูปที่ 25 ที่กำหนดค่าตัวแปรของจำนวนรวมของประโยคที่ได้จากการแต่งเติมข้อมูลมีค่าเท่ากับ 7 เป็นต้น



รูปที่ 24 อัลกอริทึมของการแต่งเติมข้อมูล (อ้างอิงจาก Fig.2 ใน [9])



รูปที่ 25 ตัวอย่างของการแต่งเติมข้อมูลทั้งหมด 7 ครั้ง โดยมีการเปลี่ยนแปลงของข้อมูลในประโยคทั้งหมด 25 เปอร์เซ็นต์ (อ้างอิงจาก Fig.1 ใน [9])

การค้นหาคำที่มีความหมายเหมือนกันจะทำการค้นหาจากรายการของ WordNet ซึ่งเป็นงานวิจัยของ George A. Miller [10] ที่รวบรวมรายการของคำศัพท์ที่มีความหมายเหมือนกันและใช้ชุดข้อมูลจาก [11] ที่ประกอบไปด้วยชุดข้อมูลฝึกสอนจำนวน 159,571 ตัวอย่าง ที่ถูกจำแนกออกเป็น 6 ประเภทและชุดข้อมูลทดสอบจำนวน 153,164 ตัวอย่าง ที่ถูกจำแนกออกเป็น 6 ประเภท ตามกลุ่มของข้อมูลที่แบ่งออกเป็น toxic, severe toxic, obscene comments, threats, insults และ identity hate โดยมีสัดส่วนของข้อมูลในกลุ่มต่างๆ ดังตารางที่ 6

ตารางที่ 6 จำนวนของข้อมูลในแต่ละประเภท (อ้างอิงจากตารางที่ 1 ใน [9])

Type	Samples	Percentage
Toxic	15249	9,6%
Severe toxic	1959	1,2%
Obscene	8449	5,3%
Threat	478	0,3%
Insults	7877	4,9%
Identity hate	1405	0,9%
Overall	159571	100%

ในการประเมินผลประสิทธิภาพของแนวคิดการแต่งเติมข้อมูลนั้น Anna V. Mosolova และคณะ ใช้แบบจำลองโครงข่ายประสาทเทียมคอนโวลูชัน (Convolutional Neural Network หรือ CNN) และประเมินผลด้วยการให้คะแนนอัลกอรึทึมจากค่าเฉลี่ยของพื้นที่ใต้กราฟ (Area under the curve: AUC) สำหรับแต่ละรอบของการทำนายประเภทของข้อมูล โดยมีผลการทดสอบดังตารางที่ 7

ตารางที่ 7 ตารางแสดงผลการทดสอบแบบจำลอง (อ้างอิงจากตารางที่ 2 ใน [9])

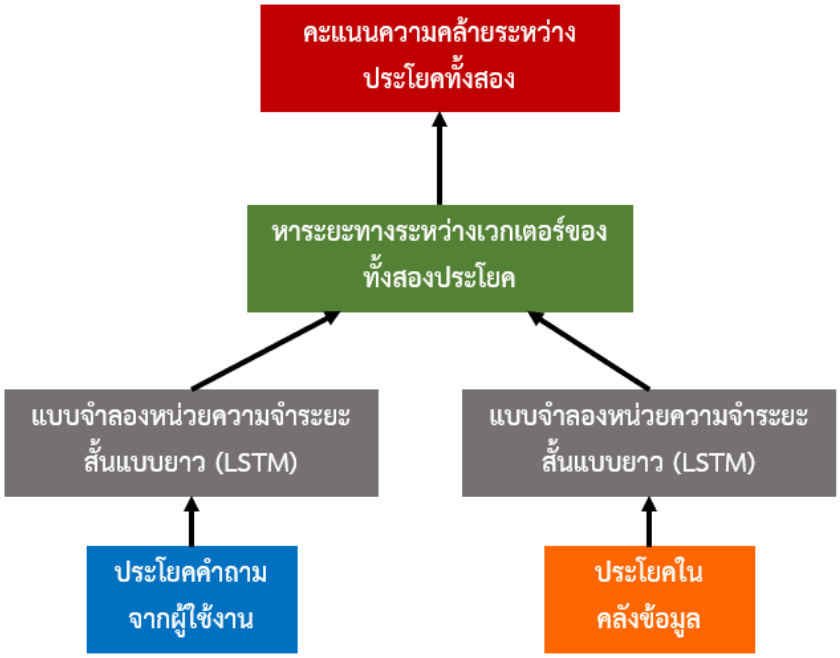
Model	**Public score**	**Private Score**
CNN with character embeddings	0.9065	0.8933
CNN with character embeddings and with a 6 times augmentation for 25% of all words	0.9436	0.9446
CNN with word embeddings	0.9752	0.9742
CNN with word embeddings and with a 6 times augmentation for 25% of all words	0.9743	0.9721

บทที่ 4
แนวคิดและวิธีการดำเนินงาน

วิธีการแต่งเติมข้อมูลเชิงข้อความในงานวิทยานิพนธ์นี้มีวัตถุประสงค์เพื่อเพิ่มจำนวนชุดข้อมูลที่ใช้สำหรับการเรียนรู้ขอแบบจำลองหน่วยความจำระยะสั้นแบบยาวแบบสยามให้มีประสิทธิภาพมากขึ้นและทำงานได้ดีกับข้อมูลที่มีปริมาณน้อย เพื่อนำไปประยุกต์ใช้กับการสร้างหุ่นยนต์สนทนาไทยสำหรับการตอบปัญหาของคำถามที่พบบ่อย เพียงแค่ผู้ใช้งานระบุคำถามที่ต้องการ ระบบจะช่วยค้นหาคำตอบของคำถามจากคลังข้อมูลที่มีเพื่อตอบปัญหาที่ได้รับจากผู้ใช้งาน

4.1 แนวทางการประยุกต์ใช้แบบจำลองหน่วยความจำระยะสั้นแบบยาวแบบสยามร่วมกับการแต่งเติมข้อมูลเชิงข้อความ

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่า แบบจำลองหน่วยความจำระยะสั้นแบบยาวแบบสยามเหมาะสมที่จะนำมาพัฒนาต่อ เนื่องจากเป็นวิธีที่ให้ผลลัพธ์ดีและมีประสิทธิภาพ เช่น งานวิจัยของ Yichao Lu และคณะ [5] ที่ได้นำเสนอแนวคิดในการใช้แบบจำลองมาแก้ไขปัญหาในการตอบคำถามที่พบบ่อย โดยใช้โครงสร้างของแบบจำลองในลักษณะเดียวกันกับงานวิจัยของ Jonas Mueller และ Aditya Thyagarajan [6] ซึ่งทั้งสองงานวิจัยได้ใช้โครงข่ายสยามในการหาความคล้ายระหว่างประโยคเหมือนกัน รวมไปถึงงานวิจัยของ Nouha Othman และคณะ [7] ที่เสนอแนวคิดการใช้แบบจำลองลักษณะเดียวกันมาใช้ในการหาความคล้ายกันของคำถามที่ได้รับจากผู้รับบริการและให้ผลลัพธ์เป็นคำตอบที่ได้จากการหาความคล้ายของประโยคคำถาม นอกจากนี้จากการศึกษายังพบว่า งานวิจัยของ Panitan Muangkammuen และคณะ [8] ได้นำเสนอหุ่นยนต์สนทนาของคำถามที่พบบ่อยแบบอัตโนมัติที่ใช้กับข้อมูลภาษาไทยจำนวน 2,636 คำถามที่ให้ผลลัพธ์ที่ดี แต่การใช้แบบจำลองหน่วยความจำระยะสั้นแบบยาวในการจำแนกประเภทของปัญหาและสุ่มคำตอบที่อยู่ในปัญหาประเภทนั้นๆ มาเป็นข้อมูลนำออก อาจทำให้ผู้ใช้งานได้รับคำตอบที่ไม่เหมือนเดิมแม้ว่าจะถามปัญหาเดิมก็ตาม ดังนั้นในงานวิจัยนี้จึงนำเสนอการใช้งานแบบจำลองหน่วยความจำระยะสั้นแบบยาวแบบสยามที่มีลักษณะของโครงสร้างดังรูปที่ 26



รูปที่ 26 โครงสร้างแบบจำลองหน่วยความจำระยะสั้นแบบยาวแบบสยามสำหรับงานวิจัยที่นำเสนอ

เมื่อได้คะแนนความคล้ายระหว่างประโยคคำถามจากผู้ใช้งานกับประโยคทั้งหมดในคลังข้อมูล เราจะทราบว่าประโยคใดในคลังข้อมูลที่มีคะแนนความคล้ายสูงที่สุด คำตอบของคำถามนั้นจะถูกนำมาเป็นข้อมูลนำออกให้แก่ผู้ใช้งาน และผู้วิจัยพบว่าการเพิ่มประสิทธิภาพของแบบจำลองหน่วยความจำระยะสั้นแบบยาวนั้นคือการให้แบบจำลองได้ผ่านการเรียนรู้ด้วยข้อมูลฝึกที่หลากหลาย สามารถทำให้ประสิทธิภาพของแบบจำลองดีขึ้นได้ ผู้วิจัยจึงได้ทำการศึกษางานวิจัยที่เกี่ยวข้องกับการแต่งเติมข้อมูลเชิงข้อความ (Text Augmentation) เช่น งานวิจัยของ Anna V. Mosolova และคณะ [9] ที่นำเสนอวิธีการแต่งเติมข้อมูลเชิงข้อความโดยใช้คำที่มีความหมายเหมือนกันโดยได้ประยุกต์แนวคิดมาจากการแต่งเติมข้อมูลประเภทรูปภาพ และได้ผลลัพธ์ที่ดีขึ้นทั้งในเรื่องของการเพิ่มจำนวนข้อมูล ซึ่งส่งผลให้ข้อมูลที่ได้มีความหลากหลายมากขึ้นในการนำไปใช้กับการเรียนรู้ของแบบจำลอง ดังนั้นในงานวิจัยนี้จะทำการประยุกต์การใช้โครงสร้างแบบจำลองหน่วยความจำระยะสั้นแบบยาวแบบสยามร่วมกับการแต่งเติมข้อมูลเชิงข้อความด้วยวิธีการสร้างประโยคคำถามใหม่จากการแทนคำที่มีระยะทางของเวกเตอร์ใกล้กัน ซึ่งกระบวนการนี้ทำให้ข้อมูลหลังจากผ่านการเติมแต่งมีความหลากหลายของคำศัพท์ในชุดข้อมูลเพิ่มขึ้นและสามารถนำไปใช้ฝึกฝนแบบจำลองหุ่นยนต์สนทนาที่สร้างขึ้นสำหรับตอบคำถามของปัญหาที่พบ โดยภาพรวมของวิธีการแต่งเติมข้อมูลเชิงข้อความที่นำเสนอ ดังรูปที่ 27 และมีรายละเอียดดังนี้

- หุ่นยนต์สนทนาจะเริ่มบทสนทนาด้วยการกล่าวคำทักทาย เช่น สวัสดีค่ะ วันนี้มีอะไรให้ช่วยคะ
- รับข้อมูลในรูปแบบของข้อมูลเชิงตัวอักษรที่มีลักษณะเป็นประโยคจากผู้ใช้งานและนำข้อมูลไปผ่านการเตรียมข้อมูลในขั้นตอนของการทำความสะอาดข้อมูล การตัดคำ และการกำจัดคำหยุด
- คำแต่ละคำในประโยคของข้อมูลนำเข้าที่แบ่งเรียบร้อยแล้ว จะถูกนำไปสร้างเป็นรายการคำศัพท์โดยไม่นับรวมคำที่ซ้ำกับคำที่มีในรายการคำศัพท์
- คำทุกคำในรายการคำศัพท์ที่สร้างขึ้นจะถูกแปลงให้เป็นคำฝังตัว
- นำคำศัพท์ทั้งหมดที่เป็นคำฝังตัว ไปค้นหาในคลังข้อมูลที่รวบรวมเวกเตอร์ของคำต่างๆ โดยจะค้นหาคำที่มีระยะทางของเวกเตอร์ใกล้กับเวกเตอร์ของคำศัพท์ในรายการที่สร้างขึ้นเอง โดยใช้การคำนวณหาค่าความเหมือนโคไซน์ (Cosine Similarity) ระหว่างเวกเตอร์ของคำทั้งสอง ถ้าพบว่ามีคำที่มีค่าความเหมือนโคไซน์มากกว่า 0.97 และเป็นคำที่มีค่าความเหมือนโคไซน์สูงที่สุดอันดับแรก จะถือว่าคำสองคำนั้นมีความคล้ายกันและคำศัพท์ในคลังคำนั้นจะถูกเพิ่มเข้าไปเป็นคอลัมน์ใหม่ในรายการคำศัพท์ เช่น ประโยค “ตอนเย็นจะไปเที่ยวกับครอบครัว” จะถูกนำไปแบ่งคำและกำจัดคำหยุด และได้เป็นรายการของคำ “['ตอนเย็น', 'ไปเที่ยว', 'ครอบครัว']” จากนั้นนำคำในรายการไปค้นหาคำที่คล้ายกันและสร้างเป็นรายการคำศัพท์แสดงดังตารางที่ 8

ตารางที่ 8 ตัวอย่างการสร้างรายการคำศัพท์ที่ได้จากการค้นหาคำที่มีความหมายคล้ายกันด้วยเกณฑ์ของค่าความเหมือนโคไซน์ที่มีค่ามากกว่า 0.5 ซึ่งมีค่าสูงสุด 3 อันดับแรก

คำศัพท์	คำที่มีความหมายคล้ายกัน
ตอนเย็น	ตอนเช้า, ตอนกลางวัน, กลางดึก
ไปเที่ยว	มาหา, ไปดู, เดินสาย
ครอบครัว	-

- คำแต่ละคำในแต่ละประโยคของข้อมูลนำเข้า จะถูกแทนที่ด้วยคำที่มีความคล้ายกันตามรายการคำศัพท์ที่ทำขึ้น โดยการแทนคำไปในประโยคนั้นจะแทนคำไปที่คำแหน่งเดิมเพื่อรักษาลำดับของคำในประโยคนั้นๆ ให้เหมือนเดิม
- ทำการแทนที่คำทีละตำแหน่งเข้าไปที่ตำแหน่งเดิมในประโยค และไล่ลำดับของคำที่จะแทนที่ละหนึ่งคำจนครบทั้งประโยค หากคำไหนไม่มีรายการของคำที่คล้ายก็จะใช้

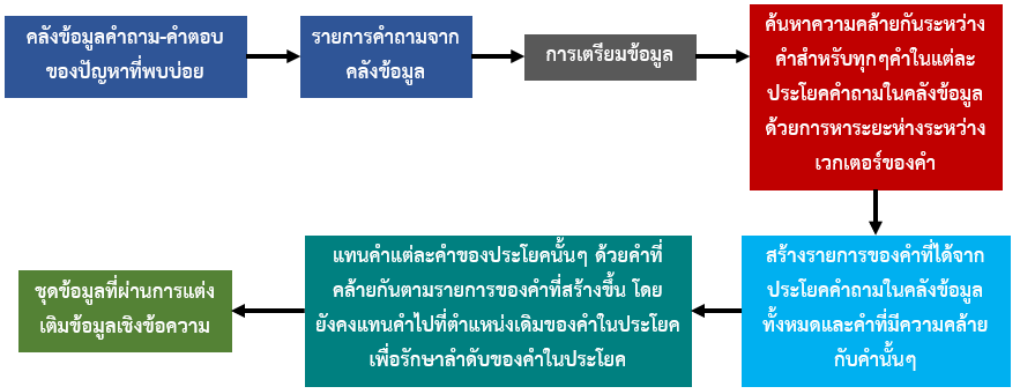
คำศัพท์เดิมที่ใช้ในข้อมูลนำเข้า เช่น รายการของคำคือ “['ตอนเย็น', 'ไปเที่ยว', 'ครอบครัว']” เมื่อนำคำที่คล้ายกันในตารางที่ 8 มาใช้ จากนั้นทำการแต่งเติมข้อมูลจะได้ผลลัพธ์ดังต่อไปนี้

1) ['ตอนเย็น', 'ไปเที่ยว', 'ครอบครัว']
2) ['ตอนเย็น', 'มาหา', 'ครอบครัว']
3) ['ตอนเย็น', 'ไปดู', 'ครอบครัว']
4) ['ตอนเย็น', 'เดินสาย', 'ครอบครัว']
5) ['ตอนเช้า', 'ไปเที่ยว', 'ครอบครัว']
6) ['ตอนเช้า', 'มาหา', 'ครอบครัว']
7) ['ตอนเช้า', 'ไปดู', 'ครอบครัว']
8) ['ตอนเช้า', 'เดินสาย', 'ครอบครัว']
9) ['ตอนกลางวัน', 'ไปเที่ยว', 'ครอบครัว']
10) ['ตอนกลางวัน, 'มาหา', 'ครอบครัว']
11) ['ตอนกลางวัน', 'ไปดู', 'ครอบครัว']
12) ['ตอนกลางวัน, 'เดินสาย', 'ครอบครัว']
13) ['กลางดึก', 'ไปเที่ยว', 'ครอบครัว']
14) ['กลางดึก, 'มาหา', 'ครอบครัว']
15) ['กลางดึก', 'ไปดู', 'ครอบครัว']
16) ['กลางดึก, 'เดินสาย', 'ครอบครัว']

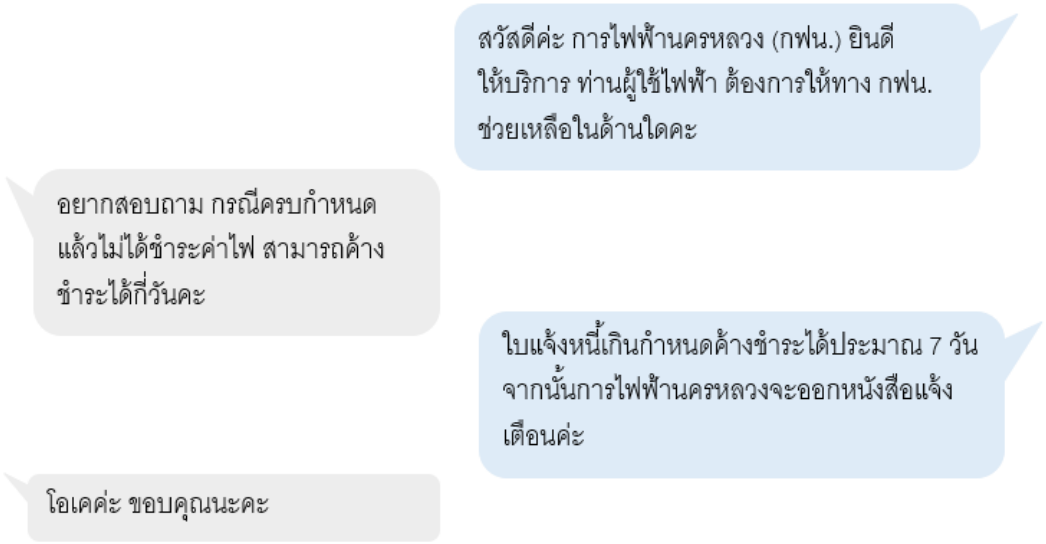
- เมื่อทำการแต่งเติมข้อมูลเรียบร้อยแล้ว จะนำข้อมูลทั้งหมดไปใช้กับแบบจำลองหน่วยความจำระยะสั้นแบบยาว (LSTM) ที่มีลักษณะของโครงสร้างแบบโครงข่ายสยาม (Siamese Neural Network) ร่วมกับการหาความคล้ายเวกเตอร์ระหว่างประโยคทั้งสอง ด้วยการหาระยะทางแบบต่างๆ ดังนี้
    1. ระยะทางแบบยุคลิด (Euclidean Distance)
    2. การหาระยะทางแบบแมนฮัตตัน (Manhattan Distance)
    3. การค่าความเหมือนโคไซน์ (Cosine Similarity)
- เมื่อได้ค่าความคล้ายระหว่างคำถามจากผู้ใช้งานกับคำถามทั้งหมดในคลังข้อมูลแล้วนั้น จะนำคำตอบของคำถามที่มีค่าความคล้ายมากกว่าเกณฑ์ที่กำหนดไว้และมีค่าความคล้ายสูงสุดกับคำถามที่เป็นข้อมูลนำเข้ามาใช้เป็นข้อมูลนำออกซึ่งแสดงดังรูปที่ 28 หากไม่สามารถหาคำถามที่มีความคล้ายสูงสุดที่มีค่ามากกว่าเกณฑ์ที่กำหนดได้ ก็จะใช้ประโยค

ที่เตรียมไว้เป็นข้อมูลนำออกแทนการใช้คำตอบจากคำถามในคลังข้อมูลเพื่อแจ้งให้ผู้ใช้งานสอบถามข้อมูลใหม่อีกครั้ง ดังรูปที่ 29

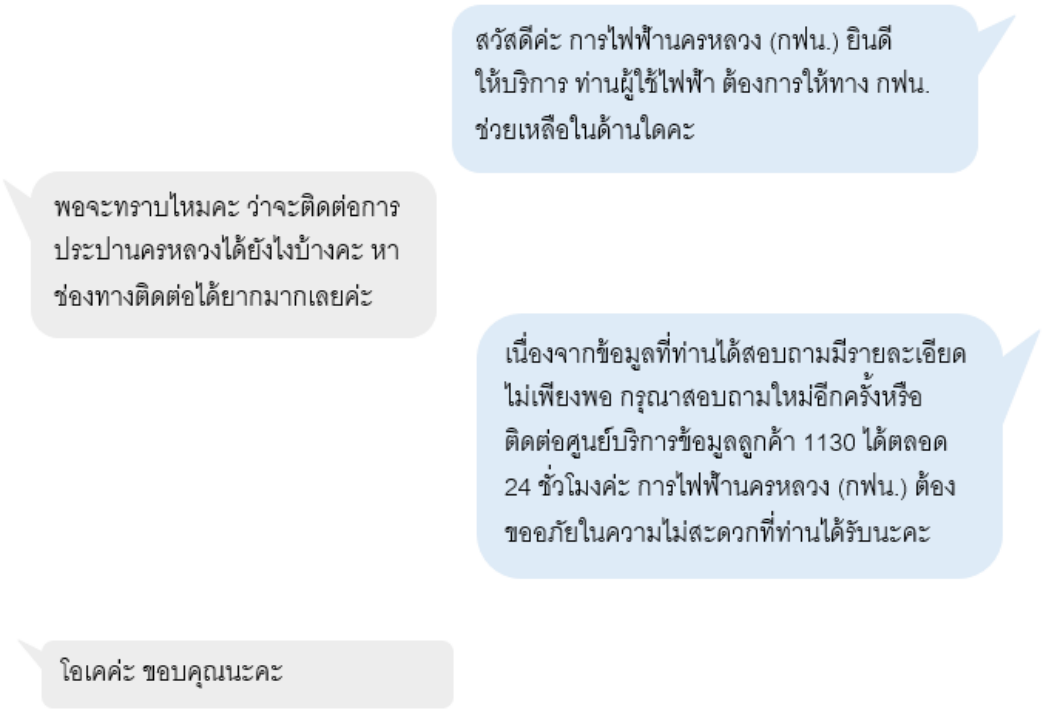
- ข้อมูลที่ใช้กับแบบจำลอง ประกอบด้วยชุดข้อมูล จำนวน 2 ชุด ได้แก่
    1. ข้อมูลเชิงตัวอักษรที่ได้จากบทสนทนาในการให้บริการตอบคำถามผู้ใช้ไฟฟ้า
    2. ข้อมูลเชิงตัวอักษรที่ได้จากบทสนทนาในการให้บริการตอบคำถามผู้ใช้ไฟฟ้าที่ผ่านกระบวนการแต่งเติมข้อมูลด้วยวิธีการที่นำเสนอในงานวิจัยนี้



รูปที่ 27 ภาพรวมของวิธีการแต่งเติมข้อมูลเชิงตัวอักษรด้วยการหาความคล้ายจากเวกเตอร์ของคำ



รูปที่ 28 ตัวอย่างบทสนทนาระหว่างผู้ใช้งานกับหุ่นยนต์สนทนาที่พัฒนาด้วยวิธีการที่นำเสนอ เมื่อพบคำถามที่มีค่าความคล้ายมากกว่าเกณฑ์ที่กำหนดไว้และมีค่าความคล้ายสูงสุดกับคำถามที่เป็นข้อมูลนำเข้า



รูปที่ 29 ตัวอย่างบทสนทนาระหว่างผู้ใช้งานกับหุ่นยนต์สนทนาที่พัฒนาด้วยวิธีการที่นำเสนอ เมื่อไม่สามารถหาคำถามที่มีความคล้ายสูงสุดกับคำถามที่เป็นข้อมูลนำเข้าที่มีค่ามากกว่าเกณฑ์ที่กำหนดได้

4.2 แนวทางการปรับปรุงคำฝังตัว

เนื่องจากข้อมูลเชิงตัวอักษรที่ใช้ในงานวิจัยนี้ จะเป็นคำถามปลายปิดที่เป็นกลุ่มคำถามของกลุ่มผู้ใช้บริการไฟฟ้าของการไฟฟ้านครหลวง ผู้วิจัยจึงมีแนวคิดที่จะปรับปรุงคำฝังตัว ด้วยการนำแบบจำลองของคำฝังตัวที่เรียนรู้ด้วยคำทั่วไป มาเรียนรู้ต่อด้วยข้อมูลที่เป็นประโยคที่เกี่ยวข้องกับการไฟฟ้านครหลวงจำนวนทั้งสิ้น 25,543 ประโยค เพื่อปรับค่าน้ำหนักของแบบจำลองและส่งผลให้ประสิทธิภาพของคำฝังตัวให้สามารถแปลงภาษาธรรมชาติ (Natural Language) ให้กลายเป็นข้อมูลประเภทตัวเลข (Numerical) ที่อยู่ในรูปแบบของเวกเตอร์เพื่อใช้แสดงถึงความสัมพันธ์หรือความหมายที่ใกล้เคียงกันของคำได้ดียิ่งขึ้นเมื่อนำมาใช้กับข้อมูลที่มีลักษณะเฉพาะทาง

4.3 ชุดข้อมูล (Data Set)

ผู้วิจัยต้องการทดสอบแบบจำลองว่าสามารถค้นคืนคำถามจากคลังข้อมูลที่มีความคล้ายสูงสุดเมื่อเทียบกับคำถามที่เป็นข้อมูลนำเข้าได้ดีเพียงใด เพื่อนำคำตอบของคำถามในคลังข้อมูลนั้นเป็นข้อมูลนำออกให้แก่ผู้ใช้งาน โดยชุดข้อมูลที่ใช้นั้นมีรายละเอียดดังนี้

4.3.1 รูปแบบข้อมูล

ชุดข้อมูลคำถามและคำตอบที่ใช้ในงานวิจัยนี้จะเป็นกลุ่มของคำถามที่พบบ่อยในกลุ่มที่เกี่ยวข้องกับการสอบถามงานด้านบริการผู้ใช้ไฟฟ้า, สอบถามงานด้านวิธีการชำระเงินค่าบริการไฟฟ้า, สอบถามข้อมูลประกาศดับไฟ และ สอบถามเรื่องทั่วไปเกี่ยวกับการไฟฟ้านครหลวง โดยมีจำนวนคำถามทั้งหมด 1,169 ประโยค ตารางที่ 9 แสดงตัวอย่างของกลุ่มคำถามและคำตอบที่ใช้ในงานวิจัยนี้

ตารางที่ 9 ตัวอย่างลักษณะชุดข้อมูลของคำถามที่พบบ่อย

คำถาม	คำตอบ
ค่าไฟค้างได้กี่เดือนคะ	1 เดือนครับ
ค่าไฟค้างจ่ายได้กี่เดือนครับ	1 เดือนครับ
อยากทราบว่าเดี๋ยวนี้ค่าไฟค้างได้กี่เดือนคะ	1 เดือนครับ
เสียค่าติดหม้อใหม่เท่าไหร่คะ	40 บาทครับ
ค่าธรรมเนียมถ้าจะให้ติดหม้อใหม่นี่กี่บาทคะ	40 บาทครับ
ถ้าจะเอารถ EV ไปชาร์จไฟ นี่ กฟน คิดค่าบริการชาร์จไฟรถยังไงบ้างคะ	เข้ามาชาร์จได้เลย ไม่มีค่าบริการครับ
ที่ กฟน. มีที่ชาร์จไฟของรถ EV ถ้าเข้าไปชาร์จจะมีค่าใช้จ่ายอะไรไหมคะ	เข้ามาชาร์จได้เลย ไม่มีค่าบริการครับ
คือหนูดูมิเตอร์ที่ห้อง มันมีเลขอะไรขึ้นมาไม่รู้เต็มไปหมดเลยค่ะ	เครื่องวัดไฟฟ้าขนาด 15 (45) A โดยทั่วไปจะอ่านเครื่องวัดไฟฟ้า 4 ตัวหน้า ตัวสุดท้ายไม่นำมาคิดครับ
มิตเตอร์ที่ห้องมันมีเลขอะไรไม่รู้เต็มไปหมดเลย ผมจะอ่านเลขที่มิตเตอร์เองได้ยังไงครับ	เครื่องวัดไฟฟ้าขนาด 15 (45) A โดยทั่วไปจะอ่านเครื่องวัดไฟฟ้า 4 ตัวหน้า ตัวสุดท้ายไม่นำมาคิดครับ
ขอรบกวนหน่อยนะครับ มิตเตอร์แบบนี้ต้องอ่านเลขครบทุกตัวใช่หรือเปล่าครับ	เครื่องวัดไฟฟ้าขนาด 15 (45) A โดยทั่วไปจะอ่านเครื่องวัดไฟฟ้า 4 ตัวหน้า ตัวสุดท้ายไม่นำมาคิดครับ
อาคารที่พักผมไฟดับครับ	เบื้องต้นแนะนำให้ท่านติดต่อทางนิติบุคคลของอาคารที่พักอาศัย เพราะอาจจากเกิดจากระบบภายในของอาคารค่ะ
สวัสดีค่ะ ขอสอบถามค่ะ ไฟในห้องพักดับ	เบื้องต้นแนะนำให้ท่านติดต่อทางนิติบุคคลของ

คำถาม	คำตอบ
อันดับแรกต้องแจ้งใครก่อนคะ	อาคารที่พักอาศัย เพราะอาจจากเกิดจากระบบภายในของอาคารค่ะ
ถ้าผมสมัคร E-service ใบแจ้งหนี้จะเข้ามาทางเมลผมใช่ไหมครับ ขอบคุณครับ	เมื่อท่านสมัคร E-service แล้ว ท่านสามารถดำเนินการสมัคร E-invoice เพื่อขอรับใบแจ้งค่าไฟฟ้าทาง Email ได้ค่ะ
มีแจ้งใบแจ้งหนี้ค่าไฟทางเมลไหมคะ	เมื่อท่านสมัคร E-service แล้ว ท่านสามารถดำเนินการสมัคร E-invoice เพื่อขอรับใบแจ้งค่าไฟฟ้าทาง Email ได้ค่ะ
สอบถามเรื่องการแจ้งใบแจ้งหนี้ทางเมลค่ะ	เมื่อท่านสมัคร E-service แล้ว ท่านสามารถดำเนินการสมัคร E-invoice เพื่อขอรับใบแจ้งค่าไฟฟ้าทาง Email ได้ค่ะ
ขอโทษนะคะ เลขสัญญาคือเลขกี่หลักคะ	เลขบัญชีแสดงสัญญามีทั้งหมด 9 หลักครับ
บัญชีแสดงสัญญา ต้องมี 8 ตัว หรือ 9 ตัว ครับ	เลขบัญชีแสดงสัญญามีทั้งหมด 9 หลักครับ
จะขอไฟใหม่นี่ต้องเตรียมเอกสารอะไรบ้างครับ	เอกสารที่ใช้ถ้าท่านเป็นเจ้าของเดิมและไปติดต่อด้วยตัวเอง 1.บัตรประชาชน 2.ใบแจ้งหนี้ค่าไฟฟ้าเดือนล่าสุดครับ
สวัสดีค่ะ ขอหม้อแปลงไฟฟ้าใหม่ ใช้เอกสารอะไรบ้างคะ?	เอกสารที่ใช้ถ้าท่านเป็นเจ้าของเดิมและไปติดต่อด้วยตัวเอง 1.บัตรประชาชน 2.ใบแจ้งหนี้ค่าไฟฟ้าเดือนล่าสุดครับ
ถ้าเราเปิดตู้แช่แข็งเฉพาะกลางวันกับเปิดเฉพาะกลางคืนอันไหนสิ้นเปลืองกว่ากันครับ	แนะนำเปิดเวลาใช้งานครับและหลังไม่ใช้งานควรปิด คอมเพลสเซอร์จะได้ทำงานน้อยลง
ถ้าเราเปิดปิดตู้แช่แข็งบ่อยๆ การกระชากไฟตอนเปิดมีผมให้ค่าไฟแพงมากไหมครับ	แนะนำเปิดเวลาใช้งานครับและหลังไม่ใช้งานควรปิด คอมเพลสเซอร์จะได้ทำงานน้อยลง

4.3.2 แนวทางการจัดกลุ่มคำถาม

ผู้วิจัยได้นำคำถามจำนวนทั้งหมด 1,169 ประโยคมาทำการจัดกลุ่มของคำถามที่มีคำตอบคล้ายกัน สามารถจัดกลุ่มได้ทั้งหมด 120 กลุ่ม โดยตารางที่ 10 แสดงตัวอย่างของการจัดกลุ่มคำถามที่มีคำตอบเหมือนกันในงานวิจัยนี้

ตารางที่ 10 ตัวอย่างของการจัดกลุ่มคำถาม

กลุ่ม	ประโยค
1	ค่าไฟค้างได้กี่เดือนคะ ค่าไฟค้างจ่ายได้กี่เดือนครับ อยากทราบว่าเดี๋ยวนี้ค่าไฟค้างได้กี่เดือนคะ
2	ที่บ้านไฟกระชากบ่อยมากเลยค่ะ แก้ไขได้ยังไงบ้างคะ สอบถามหน่อยครับ เกี่ยวกับระบบไฟฟ้ากระชากหรือไฟฟ้าเกินหน่อยครับ
3	อยากทราบว่าปกติจ่ายค่าไฟของบริษัท สามารถจ่ายบัตรเครดิตได้ไหมคะ ต้องการยกเลิกการจ่ายค่าไฟฟ้าผ่านบัตรเครดิตค่ะ ต้องการยกเลิกตัดค่าไฟผ่านบัตรครับ ยกเลิกชำระผ่านบัตรผมต้องแจ้งทางนี้หรือธนาคารครับ สอบถามหน่อยครับ จะทำเรื่องหักค่าไฟให้ตัดผ่านบัตรเครดิตมีแบบฟอร์มให้โหลดไหมครับ เข้าดูที่เว็บมีแต่หักผ่านบัญชีธนาคาร ดิฉันต้องการเปลี่ยนการตัดค่าไฟผ่านบัตรค่ะ ถ้าต้องการเปลี่ยนการตัดค่าไฟผ่านบัตรเครดิตใบใหม่ ต้องทำไงคะ
4	เขาดูค่าไฟกันยังไงคะ การคิดหน่วยไฟฟ้ามิเตอร์ไฟเราดูแค่สี่ตัวเท่านั้นถูกต้องไหมคะ หนูจะดูค่าตรงมิเตอร์ยังไงคะ
5	ถ้าเราจะย้ายเสาไฟราคาจะอยู่ที่เท่าไรคะ สอบถามเรื่องการย้ายเสาสาย เราขอแจ้งย้ายเสาไฟได้ใช่ไหมคะ เสาไฟที่บดบังวิสัยทัศน์ เราสามารถย้ายได้ไหมคะ? ถ้าต้องการย้ายเสาไฟฟ้าที่อยู่หน้าร้านต้องทำยังไงคะ ใช้เอกสารอะไรบ้าง

4.3.3 แนวทางการสร้างชุดข้อมูลประโยคที่ถูกจับคู่

ผู้วิจัยได้นำคำถามในชุดข้อมูลที่มีทั้งหมด 1,169 ประโยค ไปผ่านกระบวนการเตรียมข้อมูล จะได้จำนวนประโยคหลังผ่านกระบวนการนี้ออกมาทั้งหมด 1,131 ประโยค เนื่องจากมีการกำจัดคำหยุดและนำคำที่ไม่สามารถแปลงเป็นคำฝังตัวได้ออกและรวมถึงหลังจากผ่านกระบวนนี้พบว่า มีประโยคบางส่วนเหลือความยาวของประโยคเพียงแค่ 1 คำ เช่น คำว่า “ไฟ” และยังพบว่า หลายๆ ประโยคนั้นซ้ำกันกับประโยคอื่นที่อยู่ในชุดข้อมูล ซึ่งจะถูกคัดทิ้งให้เหลือเพียงแค่ประโยคเดียว และทำ

การจัดกลุ่มของคำถามที่มีคำตอบเหมือนกันออกเป็นกลุ่มต่างๆ ได้ทั้งหมด 120 กลุ่ม จากนั้นจะถูกนำมาสร้างเป็นชุดข้อมูลที่มีลักษณะเป็นคู่ของประโยคคำถามและจำแนกประเภทออกเป็นข้อมูลด้วยการแบ่งออกคู่ของประโยคที่สร้างขึ้นออกเป็น 2 กลุ่ม ได้แก่

1) คู่ของประโยคที่มีคล้ายกันจะถูกกำกับด้วย “1” จะเกิดจากการนำประโยคที่ถูกจัดอยู่ในกลุ่มเดียวกันทั้งหมดมาจับคู่ระหว่างกัน

2) คู่ของประโยคที่แตกต่างกันจะถูกกำกับด้วย “0” จะเกิดจากการนำประโยคที่อยู่คนละกลุ่มทั้งหมดมาจับคู่ระหว่างกัน

โดยทั้งสองกลุ่มที่ถูกจำแนกนั้นจะไม่มีคู่ของประโยคที่ซ้ำกัน ตารางที่ 11 แสดงตัวอย่างของประโยคที่จับคู่แล้วเป็นคู่ที่มีลักษณะซ้ำกัน และตารางที่ 12 แสดงตัวอย่างของชุดข้อมูลคู่ของประโยคคำถามที่สร้างขึ้นในขั้นตอนนี้

ตารางที่ 11 ตัวอย่างของการจับคู่ประโยคที่ทำให้เกิดการซ้ำกัน

ประโยค 1	ประโยค 2	ป้ายกำกับ
อาคารที่พักผมไฟดับครับ	คอนโดผมไฟดับสักพักแล้วครับ	1
คอนโดผมไฟดับสักพักแล้วครับ	อาคารที่พักผมไฟดับครับ	1
อาคารที่พักผมไฟดับครับ	ขอปรึกษาเรื่องการผ่อนชำระค่าไฟครับ	0
ขอปรึกษาเรื่องการผ่อนชำระค่าไฟครับ	อาคารที่พักผมไฟดับครับ	0

ตารางที่ 12 ตัวอย่างชุดข้อมูลของประโยคที่ถูกจับคู่และจำแนกประเภท

ประโยค 1	ประโยค 2	ป้ายกำกับ
อาคารที่พักผมไฟดับครับ	คอนโดผมไฟดับสักพักแล้วครับ	1
อาคารที่พักผมไฟดับครับ	บางนาคอนโดไฟดับค่ะ	1
อาคารที่พักผมไฟดับครับ	ขอปรึกษาเรื่องการผ่อนชำระค่าไฟครับ	0
บางนาคอนโดไฟดับค่ะ	ขอสอบถามช่องทางการชำระเงินนอกเวลาทำการ	0
ผมอยากขอเงินประกันคืนครับ แต่ชื่อเจ้าของเป็นชื่อคุณพ่อ ผมต้องทำยังไงบ้างครับ	ขอสอบถามเรื่องการคืนเงินประกันค่ะ จะได้คืนเมื่อยกเลิกเท่านั้นใช่ไหมคะ	1

ผมอยากขอเงินประกันคืนครับ แต่ชื่อเจ้าของเป็นชื่อคุณพ่อ ผมต้องทำยังไงบ้างครับ	ดิฉันอยากจะทำเรื่องขอคืนเงินประกัน ตื่อเจ้าของเป็นชื่อคุณแม่ ดิฉันสามารถทำเรื่องขอคืนเองได้มั้ยคะ	1
ผมอยากขอเงินประกันคืนครับ แต่ชื่อเจ้าของเป็นชื่อคุณพ่อ ผมต้องทำยังไงบ้างครับ	หม้อแปลงกี่แอมป์คะถึงติดแอร์ได้	0
ผมอยากขอเงินประกันคืนครับ แต่ชื่อเจ้าของเป็นชื่อคุณพ่อ ผมต้องทำยังไงบ้างครับ	การอบรมของ กฟน.มีค่าใช้จ่ายไหมครับ	0

4.4 แนวทางการค้นคืนคำตอบแก่ผู้ใช้งาน

จากแนวทางการจัดกลุ่มคำถามที่ได้เสนอไปในข้างต้นนั้น เมื่อแบบจำลองได้ทำการค้นคืนคำถามที่มีความคล้ายสูงสุดกับคำถามที่ได้รับจากผู้ใช้งานแล้วนั้น คำตอบของกลุ่มคำถามนั้นจะเป็นข้อมูลนำออกเพื่อตอบคำถามแก่ผู้ใช้งาน หากแบบจำลองไม่สามารถค้นคืนคำถามที่มีความคล้ายมากกว่าค่าที่กำหนด ระบบจะคืนคำตอบแก่ผู้ใช้งานด้วยคำตอบเริ่มต้นที่ได้เตรียมไว้ โดยในงานวิจัยนี้มีแบบจำลองที่ถูกนำมาปรับปรุงการหาระยะทางระหว่างเวกเตอร์ของคำทั้งหมด 3 แบบ ซึ่งแต่ละแบบนั้นมีเกณฑ์ของค่าความคล้ายที่กำหนดไว้แตกต่างกัน เพื่อใช้ในการจำแนกข้อมูลนำออกของแบบจำลอง ได้แก่ ค่าความคล้ายเท่ากับ 0.07 สำหรับการหาระยะทางระหว่างเวกเตอร์ของคำแบบยุคลิด ค่าความคล้ายเท่ากับ 0.03 สำหรับการหาระยะทางระหว่างเวกเตอร์ของคำแบบแมนฮัตตัน และ ค่าความคล้ายเท่ากับ 0.06 สำหรับการหาค่าความคล้ายโคไซน์

บทที่ 5
วิธีการทดลอง

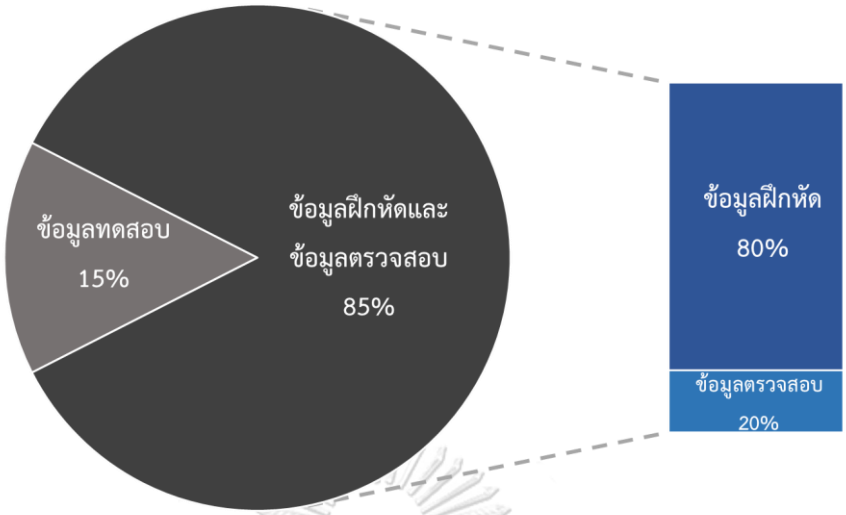
5.1 ชุดข้อมูลที่ใช้ทดสอบ

ชุดข้อมูลที่ได้จากจัดกลุ่มของคำถามหลังผ่านกระบวนการเตรียมข้อมูลที่ประกอบไปด้วยประโยคจำนวนทั้งหมด 1,131 ประโยค จะเป็นชุดข้อมูลที่นำมาใช้ทดสอบในงานวิจัยนี้ โดยจะถูกว่าเรียกว่า ชุดข้อมูลต้นฉบับ ซึ่งถูกแบ่งออกเป็นสัดส่วนโดยประมาณได้ดังนี้

1. ข้อมูลฝึก 70%
2. ข้อมูลตรวจสอบ 15%
3. ข้อมูลทดสอบ 15%

ในการแบ่งข้อมูลนั้น เริ่มจากการแบ่งประโยคทั้งหมดออกเป็นสองส่วน ซึ่งส่วนแรกจะเป็นประโยคสำหรับข้อมูลฝึกและข้อมูลตรวจสอบ และส่วนที่สองจะเป็นประโยคสำหรับข้อมูลทดสอบ โดยจะไม่มีประโยคที่เหมือนกันระหว่างทั้งสองส่วน และผู้วิจัยได้ทำการแบ่งข้อมูลโดยคำนึงถึงการกระจายข้อมูลของกลุ่มคำถามทั้งหมด ข้อมูลทั้งสองส่วนนั้นจะประกอบด้วยกลุ่มของประโยคครบทุกกลุ่ม และเนื่องจากจำนวนของประโยคในแต่ละกลุ่มคำถามนั้นมีจำนวนไม่เท่ากัน โดยจะเลือกนำประโยคเพิ่มไปในส่วนแรกที่ใช้สำหรับข้อมูลฝึกและข้อมูลตรวจสอบ จากนั้นจึงเพิ่มประโยคไปในข้อมูลทดสอบ ส่งผลให้สัดส่วนของข้อมูลที่ถูกแบ่งในงานวิจัยนี้จึงเป็นสัดส่วนโดยประมาณ

ซึ่งในงานวิจัยนี้ใช้ชุดข้อมูลที่มีจำนวนประโยคทั้งหมด 946 ประโยค สำหรับข้อมูลฝึกและข้อมูลตรวจสอบ และจะถูกนำไปสร้างเป็นชุดข้อมูลประโยคที่ถูกจับคู่ ก่อนที่จะถูกแบ่งเป็นข้อมูลฝึกและข้อมูลตรวจสอบออกเป็นสัดส่วนโดยประมาณ 80% สำหรับชุดข้อมูลฝึกและ 20% สำหรับข้อมูลตรวจสอบ ในส่วนของประโยคที่เหลืออีก 185 ประโยค จากจำนวนประโยคทั้งหมดจะนำมาใช้เป็นข้อมูลทดสอบและจะถูกนำไปสร้างเป็นชุดข้อมูลประโยคที่ถูกจับคู่เช่นเดียวกัน ดังรูปที่ 30 แสดงสัดส่วนในภาพรวมของการแบ่งชุดข้อมูลที่ใช้ทดสอบแบบจำลอง



รูปที่ 30 การแบ่งสัดส่วนของชุดข้อมูลที่ใช้ทดสอบแบบจำลอง

จากนั้นผู้วิจัยได้นำข้อมูลฝึกจากข้อมูลชุดต้นฉบับไปทำการแต่งเติมข้อมูลด้วยวิธีการที่นำเสนอในงานวิจัยนี้ ทำให้ประโยคของชุดข้อมูลฝึกสอนนั้นมีจำนวนเพิ่มมากขึ้น ส่วนข้อมูลตรวจสอบและข้อมูลทดสอบนั้นยังคงใช้ชุดข้อมูลเดียวกันกับข้อมูลตั้งต้น ซึ่งข้อมูลชุดนี้จะถูกเรียกว่า ชุดข้อมูลแต่งเติม และจะถูกนำไปสร้างเป็นชุดข้อมูลประโยคที่ถูกจับคู่

ชุดข้อมูลตั้งต้นและข้อมูลแต่งเติมที่อยู่ในรูปแบบชุดข้อมูลประโยคที่ถูกจับคู่เรียบร้อยแล้วนั้นจะถูกนำไปใช้เป็นชุดข้อมูลสำหรับการเรียนรู้ของแบบจำลอง โดยชุดข้อมูลตั้งต้นและชุดข้อมูลแต่งเติมจะมีจำนวนคู่ของประโยคทั้งหมดดังนี้

5.1.1 ชุดข้อมูลตั้งต้น

ชุดข้อมูลตั้งต้นประกอบไปด้วยคู่ของประโยคทั้งหมด 464,005 คู่ โดยแบ่งออกเป็น 3 ส่วนดังนี้

1. ข้อมูลฝึก จำนวน 356,820 คู่
2. ข้อมูลตรวจสอบ จำนวน 90,165 คู่
3. ข้อมูลทดสอบ จำนวน 17,020 คู่

5.1.2 ชุดข้อมูลแต่งเติม

ชุดข้อมูลแต่งเติมประกอบไปด้วยคู่ของประโยคทั้งหมด 1,821,070 คู่ โดยแบ่งออกเป็น 3 ส่วนดังนี้

1. ข้อมูลฝึก จำนวน 1,713,885 คู่
2. ข้อมูลตรวจสอบ จำนวน 90,165 คู่
3. ข้อมูลทดสอบ จำนวน 17,020 คู่

5.2 แบบจำลองที่นำมาทดลอง

ผู้วิจัยใช้แบบจำลองหน่วยความจำระยะสั้นแบบยาวแบบสยามมาทดลองด้วย การนำแบบจำลองมาปรับปรุงการหาระยะทางของเวกเตอร์ทั้งหมด 3 แบบ ได้แก่ การหาระยะทางแบบยุคลิด (Euclidean Distance) การหาระยะทางแบบแมนฮัตตัน (Manhattan Distance) และการหาค่าความคล้ายโคไซน์ (Cosine Similarity) และนำมาเรียนรู้กับชุดข้อมูลตั้งต้นและชุดข้อมูลแต่งเติมเพื่อเปรียบเทียบประสิทธิภาพระหว่างแบบจำลองที่เรียนรู้ด้วยชุดข้อมูลตั้งต้นและชุดข้อมูลแต่งเติม

5.3 วิธีการประเมินผล

เป้าหมายของการแต่งเติมข้อมูลเชิงตัวอักษรคือการเพิ่มจำนวนข้อมูลและนำไปใช้ในการเรียนรู้ของแบบจำลอง เพื่อเพิ่มประสิทธิภาพในการค้นคืนข้อมูลนำออกได้แม่นยำมากยิ่งขึ้น ซึ่งในงานวิจัยนี้ใช้วิธีการประเมินผลทั้งหมด 4 แบบ ได้แก่

1. ค่าความแม่นยำในการตอบข้อความ (Recall)
2. ค่าความเที่ยงตรงในการตอบข้อความ (Precision)
3. ค่าประสิทธิภาพโดยรวมของระบบ (F1-score)
4. ค่าความเที่ยงตรงในการตอบข้อความ 5 อันดับแรก (Precision at 5)

บทที่ 6
ผลการทดลอง

จากแนวทางและวิธีการดำเนินงานที่กล่าวไปในข้างต้น และในขั้นตอนของการหาระยะทางระหว่างเวกเตอร์ของคำมาใช้ร่วมกันกับแบบจำลอง และในส่วนของขั้นตอนการทำคำฝังตัวนั้น จะมีการนำชุดข้อมูลแปลงเป็นคำฝังตัวโดยใช้คำฝังตัวทั้งหมด 2 แบบ ได้แก่ คำฝังตัวทั่วไป และ คำฝังตัวเฉพาะ ซึ่งเกิดจากการปรับปรุงคำฝังตัวด้วยแนวทางที่ได้นำเสนอไปในข้างต้น โดยในบทนี้จะกล่าวถึงผลการทดลองซึ่งจะแบ่งออกตามวิธีการประเมินผลที่นำมาใช้ในการทดลองดังต่อไปนี้

6.1 ค่าความแม่นยำในการตอบข้อความ (Recall)

จากผลการทดลองดังตารางที่ 13 พบว่าการใช้แบบจำลองที่เรียนรู้ด้วยชุดข้อมูลตั้งต้นร่วมกับการหาระยะทางระหว่างเวกเตอร์ของคำแบบแมนฮัตตันที่ใช้คำฝังตัวเฉพาะนั้น ให้ค่าความแม่นยำในการตอบข้อความสูงสุดถึงร้อยละ 42.642

ตารางที่ 13 ค่าความแม่นยำในการตอบข้อความของชุดข้อมูลตั้งต้นเปรียบเทียบกับชุดข้อมูลแต่งเติมเมื่อใช้ร่วมกับคำฝังตัวทั่วไปและคำฝังตัวเฉพาะ

ชุดข้อมูล	ค่าความแม่นยำในการตอบข้อความ (Recall) (ร้อยละ)					
	คำฝังตัวทั่วไป			คำฝังตัวเฉพาะ		
	ระยะทางระหว่างเวกเตอร์ของคำแบบยุคลิด	ระยะทางระหว่างเวกเตอร์ของคำแบบแมนฮัตตัน	ค่าความคล้ายโคไซน์	ระยะทางระหว่างเวกเตอร์ของคำแบบยุคลิด	ระยะทางระหว่างเวกเตอร์ของคำแบบแมนฮัตตัน	ค่าความคล้ายโคไซน์
ข้อมูลตั้งต้น	6.839	10.377	9.811	17.170	42.642	23.206
ข้อมูลแต่งเติม	11.698	14.717	13.208	15.094	20.755	28.679

6.2 ค่าความเที่ยงตรงในการตอบข้อความ (Precision)

จากผลการทดลองดังตารางที่ 14 พบว่าการใช้แบบจำลองที่เรียนรู้ด้วยชุดข้อมูลตั้งต้นร่วมกับการหาระยะทางระหว่างเวกเตอร์ของคำแบบยุคลิดที่ใช้คำฝังตัวทั่วไปนั้น ให้ค่าเที่ยงตรงในการตอบข้อความสูงสุดถึงร้อยละ 6.893

ตารางที่ 14 ค่าความเที่ยงตรงในการตอบข้อความของชุดข้อมูลตั้งต้นเปรียบเทียบกับชุดข้อมูลแต่งเติม เมื่อใช้ร่วมกับคำฝังตัวทั่วไปและคำฝังตัวเฉพาะ

ชุดข้อมูล	ค่าความเที่ยงตรงในการตอบข้อความ (Precision) (ร้อยละ)					
	คำฝังตัวทั่วไป			คำฝังตัวเฉพาะ		
	ระยะทางระหว่างเวกเตอร์ของคำแบบยุคลิด	ระยะทางระหว่างเวกเตอร์ของคำแบบแมนฮัตตัน	ค่าความคล้ายโคไซน์	ระยะทางระหว่างเวกเตอร์ของคำแบบยุคลิด	ระยะทางระหว่างเวกเตอร์ของคำแบบแมนฮัตตัน	ค่าความคล้ายโคไซน์
ข้อมูลตั้งต้น	6.893	3.965	6.288	3.931	3.797	3.916
ข้อมูลแต่งเติม	5.250	6.028	6.055	5.362	4.446	4.104

6.3 ค่าประสิทธิภาพโดยรวมของระบบ (F1-score)

จากผลการทดลองดังตารางที่ 15 พบว่าการใช้แบบจำลองที่เรียนรู้ด้วยชุดข้อมูลแต่งเติมร่วมกับการหาระยะทางระหว่างเวกเตอร์ของคำแบบแมนฮัตตันที่ใช้คำฝังตัวทั่วไปนั้น ให้ค่าความแม่นยำในการตอบข้อความสูงสุดถึงร้อยละ 8.553

ตารางที่ 15 ค่าประสิทธิภาพโดยรวมของระบบของชุดข้อมูลตั้งต้นเปรียบเทียบกับชุดข้อมูลแต่งเติม เมื่อใช้ร่วมกับคำฝังตัวทั่วไปและคำฝังตัวเฉพาะ

ชุดข้อมูล	ค่าประสิทธิภาพโดยรวมของระบบ (F1-score) (ร้อยละ)					
	คำฝังตัวทั่วไป			คำฝังตัวเฉพาะ		
	ระยะทางระหว่างเวกเตอร์ของคำแบบยุคลิด	ระยะทางระหว่างเวกเตอร์ของคำแบบแมนฮัตตัน	ค่าความคล้ายโคไซน์	ระยะทางระหว่างเวกเตอร์ของคำแบบยุคลิด	ระยะทางระหว่างเวกเตอร์ของคำแบบแมนฮัตตัน	ค่าความคล้ายโคไซน์
ข้อมูลตั้งต้น	6.910	5.738	7.664	6.397	6.973	6.701
ข้อมูลแต่งเติม	7.247	8.553	8.304	7.913	7.324	7.180

6.4 ค่าความเที่ยงตรงในการตอบข้อความ 5 อันดับแรก (Precision at 5)

จากผลการทดลองดังตารางที่ 16 พบว่าการใช้แบบจำลองที่เรียนรู้ด้วยชุดข้อมูลแต่งเติมร่วมกับการหาระยะทางระหว่างเวกเตอร์ของคำแบบยุคลิดที่ใช้คำฝังตัวทั่วไปนั้น ให้ค่าความเที่ยงตรงในการตอบข้อความ 5 อันดับแรกสูงสุดถึงร้อยละ 32.865

ตารางที่ 16 ค่าความเที่ยงตรงในการตอบข้อความ 5 อันดับแรกของชุดข้อมูลตั้งต้นเปรียบเทียบกับชุดข้อมูลแต่งเติม เมื่อใช้ร่วมกับคำฝังตัวทั่วไปและคำฝังตัวเฉพาะ

ชุดข้อมูล	ค่าความเที่ยงตรงในการตอบข้อความ 5 อันดับแรก (Precision at 5) (ร้อยละ)					
	คำฝังตัวทั่วไป			คำฝังตัวเฉพาะ		
	ระยะทางระหว่างเวกเตอร์ของคำแบบยุคลิด	ระยะทางระหว่างเวกเตอร์ของคำแบบแมนฮัตตัน	ค่าความคล้ายโคไซน์	ระยะทางระหว่างเวกเตอร์ของคำแบบยุคลิด	ระยะทางระหว่างเวกเตอร์ของคำแบบแมนฮัตตัน	ค่าความคล้ายโคไซน์
ข้อมูลตั้งต้น	11.676	11.784	16.216	11.676	12.432	29.081
ข้อมูลแต่งเติม	26.811	26.702	28.432	30.703	30.811	32.865

บทที่ 7
สรุปผลการวิจัยและแนวทางการวิจัยในขั้นถัดไป

7.1 สรุปผลการวิจัย

วิทยานิพนธ์ฉบับนี้นำเสนอแนวทางการแต่งเติมข้อมูลเชิงข้อความด้วยเทคนิคการหาระยะทางระหว่างเวกเตอร์ของคำสำหรับภาษาไทย และนำมาใช้ในการเรียนรู้ของแบบจำลองหน่วยความจำระยะสั้นแบบยาวแบบสยามเพื่อการค้นคืนคำตอบของคำถามที่พบบ่อย โดยในงานวิจัยนี้ได้นำข้อมูลตั้งต้นมาผ่านกระบวนการทำความสะอาดข้อมูล ตัดคำ และกำจัดคำหยุด และถูกแบ่งออกเป็น ข้อมูลฝึก ข้อมูลตรวจสอบ และ ข้อมูลทดสอบ จากนั้นนำชุดข้อมูลฝึกมาแต่งเติมข้อมูลด้วยวิธีการที่นำเสนอ ก่อนจะแปลงข้อมูลเชิงตัวอักษรไปผ่านกระบวนการทำคำฝังตัว และใช้เป็นข้อมูลนำเข้าของแบบจำลอง

นอกเหนือจากการปรับปรุงชุดข้อมูลแล้ว ภายในโครงสร้างของแบบจำลองหน่วยความจำระยะสั้นแบบยาวแบบสยามนั้น ผู้วิจัยได้นำเสนอการปรับปรุงการหาระยะทางของเวกเตอร์เพื่อทดสอบประสิทธิภาพของแบบจำลองเพิ่มเติม โดยทำการทดลองร่วมกับการหาระยะทางระหว่างเวกเตอร์แบบต่างๆ เพื่อประยุกต์ใช้การหาระยะทางของเวกเตอร์แทนการหาประโยคที่มีความคล้ายกัน โดยจากการทดลองพบว่าเมื่อนำชุดข้อมูลที่ถูกปรับปรุงด้วยวิธีการแต่งเติมข้อมูลเชิงข้อความที่นำเสนอมาใช้ในการเรียนรู้ของแบบจำลองร่วมกันกับการหาระยะทางของเวกเตอร์แบบต่างๆ โดยไม่ว่าจะใช้คำฝังตัวทั่วไปหรือคำฝังตัวเฉพาะ ค่าประสิทธิภาพโดยรวมของระบบและค่าความเที่ยงตรงในการตอบข้อความ 5 อันดับแรก ของแบบจำลองที่เรียนรู้ด้วยชุดข้อมูลแต่งเติมนั้น มีค่าเพิ่มขึ้นเมื่อเทียบกันกับค่าประสิทธิภาพโดยรวมของระบบและค่าความเที่ยงตรงในการตอบข้อความ 5 อันดับแรก ของแบบจำลองที่เรียนรู้ด้วยชุดข้อมูลตั้งต้น ในทุกรูปแบบของการหาระยะทางระหว่างเวกเตอร์ ดังผลการทดลองในตารางที่ 16 และตารางที่ 17 ส่งผลให้แนวคิดการแต่งเติมข้อมูลเชิงข้อความที่นำเสนอนั้นช่วยทำให้แบบจำลองมีประสิทธิภาพเพิ่มขึ้นเมื่อทำงานร่วมกับข้อมูลที่มีปริมาณน้อย อีกทั้งยังแบบจำลองที่ได้สามารถพัฒนาเพื่อนำไปประยุกต์ใช้เป็นส่วนหนึ่งของหุ่นยนต์สนทนาเพื่อตอบปัญหาของคำถามที่พบบ่อยจากผู้ใช้งานได้

7.2 แนวทางการวิจัยในขั้นถัดไป

1) นำแนวคิดการแต่งเติมข้อมูลเชิงข้อความที่นำเสนอไปประยุกต์ใช้กับชุดข้อมูลภาษาอื่นๆ
2) นำข้อมูลจำลองบทสนทนาภาษาไทยที่ได้จากแนวคิดการแต่งเติมข้อมูลที่นำเสนอไปพัฒนาแบบจำลองอื่นๆ ต่อไป

3) จากการศึกษาทฤษฎีที่เกี่ยวข้องพบว่า วิธีการตัดคำในภาษาไทยนั้นมีหลายรูปแบบ ซึ่งในงานวิจัยนี้เลือกใช้หลักการตัดคำด้วยการอ้างอิงคำจากพจนานุกรม (Dictionary-based) ร่วมกับการตัดคำแบบตรงมากที่สุด (Maximum Matching) โดยยังมีหลักการตัดคำด้วยการใช้กฎไวยากรณ์ทางภาษา (Rule-based) และหลักการตัดคำด้วยการสร้างแบบจำลองการเรียนรู้จากคลังข้อความขนาดใหญ่ (Machine Learning or Corpus based) ที่สามารถนำมาประยุกต์ใช้เพื่อปรับปรุงความแม่นยำของการตัดคำที่อาจจะส่งผลให้ประสิทธิภาพความแม่นยำของการแต่งเติมข้อมูลเชิงตัวอักษรที่นำเสนอให้มีประสิทธิภาพเพิ่มขึ้นได้
4) จากผลการทดลองที่พบว่าแบบจำลองที่ใช้คำฝังตัวเฉพาะ มีประสิทธิภาพในการหาความคล้ายขอประโยคได้ดีขึ้น ซึ่งในงานวิจัยนี้ใช้ข้อมูลจำนวน 25,543 ประโยค ในการปรับปรุงคำฝังตัว หากมีข้อมูลจำนวนเพิ่มมาขึ้นมาใช้ในการปรับปรุงคำฝังตัว จะสามารถทำให้คำฝังตัวนั้นสามารถแปลงข้อมูลเชิงตัวอักษรให้รูปแบบของเวกเตอร์ได้ดีขึ้น และเมื่อนำไปใช้ในการหาระยะห่างระหว่างเวกเตอร์จะสามารถให้ผลลัพธ์ที่แม่นยำมากขึ้น

รายการอ้างอิง

[1] Social Media Trends Report 2019. Available from: http://mediakit.nurse.com/wp-content/uploads/2019/02/SocialMediaTrends2019-Report-Hootsuite.pdf [2019, February 2]

[2] Q3 GLOBAL DIGITAL Statshot 2018. Available from: https://www.slideshare.net/wearesocialsg/digital-in-2018-q3-global-digital-statshot [2018, July 24]

[3] Digital News Report 2019. Available from: http://www.digitalnewsreport.org/

[4] How Consumers Use Messaging Today Article. Available from: https://www.twilio.com/learn/commerce-communications/how-consumers-use-messaging

[5] Y. Lu, P. Keung, S. Zhang, J. Sun and V. Bhardwaj. "A practical approach to dialogue response generation in closed domains", arXiv preprint arXiv:1703.09439v1 (2017).

[6] Mueller, J. and A. Thyagarajan (2016). Siamese recurrent architectures for learning sentence similarity. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona, AAAI Press: 2786–2792.

[7] Othman, N., et al. (2019). Manhattan Siamese LSTM for Question Retrieval in Community Question Answering, Cham, Springer International Publishing.

[8] P. Muangkammuen, N. Intiruk and K. R. Saikaew, "Automated Thai-FAQ Chatbot using RNN-LSTM," 2018 22nd International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand, 2018, pp. 1-4.

[9] Mosolova, A., Fomin, V., & Bondarenko, I. (2018), Text Augmentation for Neural Networks, In Supplementary Proceedings of the Seventh International Conference on Analysis of Images, Social Networks and Texts, Moscow, pp. 104-109.

[10] Miller, G. A. (1995). "WordNet: a lexical database for English." Commun. ACM 38(11): 39–41.

[11] Toxic Comment Classification Challenge. Available from: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[12] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[13] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

[14] “ULMFit Language Modeling, Text Feature Extraction and Text Classification in Thai Language. Created as part of pyThaiNLP” https://github.com/cstorm125/thai2fit/ [Accessed: September 15, 2019].

[15] PyThaiNLP. Available from: https://github.com/PyThaiNLP/pythainlp [2019, November 23]

บรรณานุกรม

ประวัติผู้เขียน

ชื่อ-สกุล	ธนัญญา พีรพัฒนาการ
วัน เดือน ปี เกิด	30 มีนาคม 2535
สถานที่เกิด	เขตป้อมปราบศัตรูพ่าย กรุงเทพมหานคร
วุฒิการศึกษา	วิศวกรรมศาสตรบัณฑิต (วศ.บ.) สาขาระบบควบคุมและเครื่องมือวัด คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี พ.ศ.2557
ที่อยู่ปัจจุบัน	9/94 หมู่บ้านอลิชาเพลส2 ซอยครุใน3 แขวงทุ่งครุ เขตทุ่งครุ กทม. 10140
ผลงานตีพิมพ์	T. Phreeraphattanakarn and B. Kijsirikul, "Text Data Augmentation Using Text Similarity with Manhattan Siamese Long Short Term Memory for Thai Language", 2020 International Conference on Computational Linguistics and Natural Language Processing (CLNLP 2020)