Automated Thai-FAQ Chatbot using RNN-LSTM

Panitan Muangkammuen
Department of Computer Engineering
Faculty of Engineering
Khon Kaen University
Khon Kaen, Thailand
panitan_m@kkumail.com

Narong Intiruk
T2P Co., Ltd.
Bangkok, Thailand
narong@t2pco.com

Kanda Runapongsa Saikaew*
Department of Computer Engineering
Faculty of Engineering
Khon Kaen University
Khon Kaen, Thailand
krunapon@kku.ac.th

***Abstract*—In the e-commerce model that has online customer service, such as email or live chat, customers mostly use live chat because it is fast and comfortable. Thus, a company needs to hire and pay for admins. However, this incurs the problem that admins need to spend an extensive amount of time for writing an answer and customers have to wait for the answers. Several chatbots are available, but they require users to set up key phrases manually. In this article, we propose and develop a Frequently Asked Questions (FAQs) Chatbot which automatically responds to customers by using a Recurrent Neural Network (RNN) in the form of Long Short-Term Memory (LSTM) for text classification. The experimental results have shown that chatbot could recognize 86.36% of the questions and answer with 93.2% accuracy.**

Keywords— chatbot, FAQs, LSTM, text classification

I. INTRODUCTION

Currently, the number of e-commerce customers has increased rapidly. In 2017, the number of digital buyers was over 1.66 billion people worldwide up from 1.32 billion in 2014 [1]. Online shops often require services, such as live chat for customers support. However, such live chat needs admins to stand by to wait to chat with customers. Alternatively, if shops provide online customers that admins work for only certain hours, then customers have to wait for an answer for a long time. As the number of customers has increased by about 10% per year, the demand for customer service also increases. Excellent online customer services will lead to higher customer satisfaction and growing profit. One of the optimal and efficient online customer support approaches is to provide an automated chatbot to answer customers' questions automatically. Chatbot will handle customer problem reports and reply the same solutions for the same type of problems. There are existing chatbots such as Chatfuel [2] that use AI to learn what kind of question by setting up key phrases. However, it cannot be used in our case because we have several key phrases and setting up key phrases of Chatfuel is not automated. This article investigated how to design and develop a chatbot to answer FAQs in a specific domain. In particular, we used a deep-learning AI model that was capable of learning from a massive data and also used LSTM for dealing with a sequence in language. The AI model will classify phrases of questions with each class having their answers.

II. RELATED WORK

There are several ways of implementing a chatbot. First, we will discuss a difference between rule-oriented chatbot and data-oriented chatbot. For rule-oriented chatbots, such as ELIZA [3] or ALICE [4], the degree of intelligent behavior depends on the knowledge base size and quality that chatbot knows by manually coding. Trusted knowledge bases may require years to be created, depending on the domain. Data-oriented chatbot base on learning models from samples of dialogues by using a machine learning approach. Thus we do not have to code the knowledge manually, but we need a corpus for training a model.

Machine learning approach has two models for implementing a chatbot, retrieval-based and generative models. Retrieval-based models use a repository of predefined responses and learn to pick an appropriate response. Generative models' architectures like Sequence to Sequence [5] are typically based on Machine Translation [6] techniques, that will generate new responses from scratch. Therefore they are more efficient than retrieval-based models which only can reply from predefined responses. However, these models are quite likely to make grammar mistakes and require a significant amount of training data.

In our case, we chose a machine learning approach because we do not want to spend much time on development because we also have FAQ corpus for machine learning. This article focus on developing a chatbot that is not necessary to generate new responses, thus the retrieval-based model is a proper choice. Several techniques have been proposed for implementing the retrieval-based model, such as Dual Encoder [7] which is RNN that attempts to match questions and right responses. In practice, we have to check a single question with all predefined responses, and this is ineffective if there are several responses. Therefore, this article chose another technique, which was called LSTM [8]. It was a classification that was based on RNN and capable of learning long-term dependencies for handling time series data like question texts.

III. DESIGN AND DEVELOPMENT

Fig. 1 shows the chatbot system overview. In this figure, chatbot takes the Thai text as an input of a question from a user and processes it into a list of an integer for computing in the classification model. The classification model would generate an integer labeled by the group of questions that have been used to create a model. By recognising the right group of the question, then the system can reply an appropriate answer.

A. Preparing data

In our case, we have 2,636 pairs of questions and answers. We then manually categorized such pairs into 80 classes (according to the number FAQ types) and labeled them with an integer. Then split questions and answers. Questions were used for training AI while the answer would be prepared for replying to customers. Example pairs of questions and answers are shown in Fig 2.

* Corresponding author

78-1-5386-8164-0/18/$31.00 ©2018 IEEE

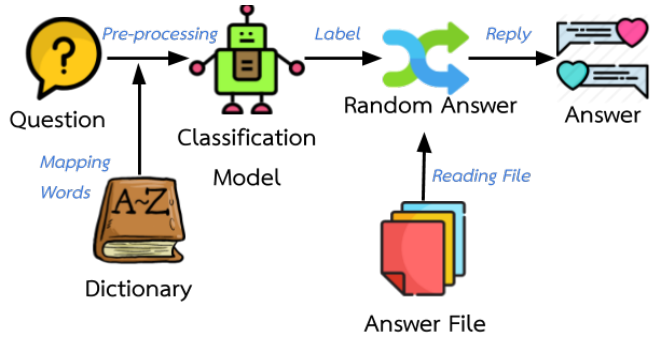

Fig. 1. System overview

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)

Fig. 2. Example pairs of questions and answers

B. *Pre-processing*

Pre-processing consists of three main components: tokenization, mapping dictionary, and zero padding. Tokenization or word segmentation is an essential task in natural language processing (NLP) for the Thai language that does not have word boundaries. After the text was segmented into words, each word would be mapped to an integer by dictionary index for processing. We would get the length of the list of integers equal to the number of words of the text. However, the classification model needed a fixed length of the input, so we used zero padding to make every input have the same length.

- Question: "ขอสอบถามได้ไหมครับ" (Can I ask you a question?)
- Tokenization: ['ขอ', 'สอบถาม', 'ได้', 'ไหม', 'ครับ'] ('Please', 'ask', 'ok', 'right', 'krub')
- Mapping dictionary: [1967, 1335, 814, 1364, 1351]
- Zero padding: [0, 0, 0, ..., 1967, 1335, 814, 1364, 1351]

C. *Classification Model*

The classification model is a neural network that takes an input from pre-processing for learning to categorize the questions. It consists of three layers. First, the embedding layer is the NLP module where words (an integer) are mapped to vectors of real numbers that learn representation for predefined fixed sized vocabulary from a corpus of text. The embedding visualization in two dimensions after training is shown in Fig. 3. Second, the long short-term memory (LSTM) layer is a particular kind of recurrent neural network (RNN), which is capable of learning sequential data such as text and video. LSTM enables RNN to remember inputs over a long period. Third, Dense layer (Output layer) with softmax activation function is used in various multiclass classification methods. The softmax activation function in the output layer represents a categorical distribution over class labels and obtaining the probabilities of each input belonging to a label. Because of softmax activation function is used at the output layer, we have to encode the label of questions to one-hot format for the learning process of the model. Fig. 4 and Fig 5. show the example of high and low confident prediction of the learned model. And the neural network layers and dimensions of data passing out each layer is shown in Fig 6.

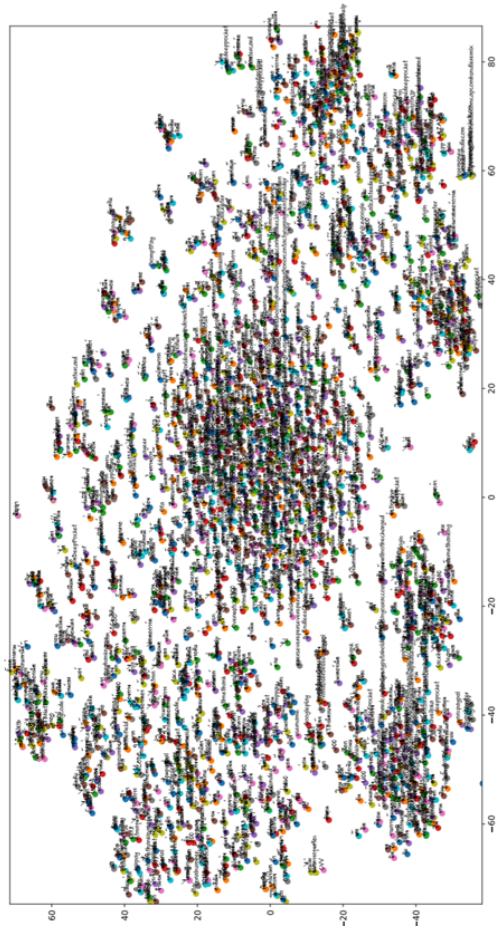

Fig. 3. Embedding Visualization

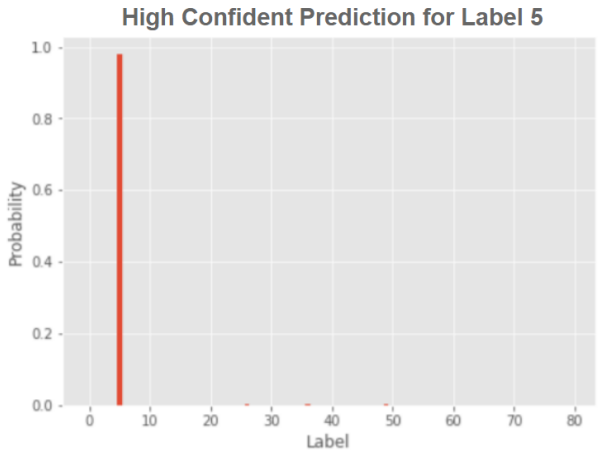

Fig. 4. Example of high confident prediction (max probability is 0.97)

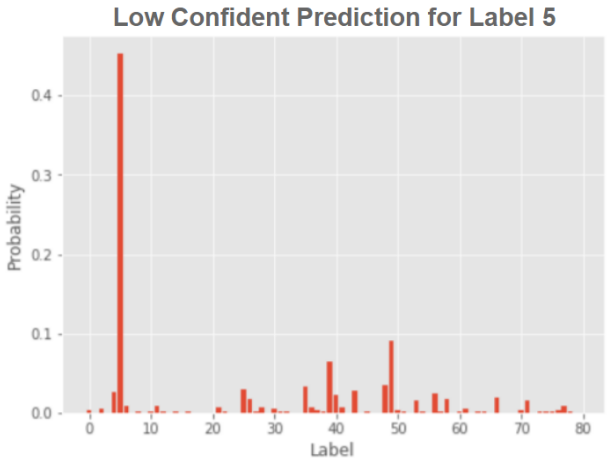

Fig 5. Example of low confident prediction (max probability is 0.45)

D. *Data splitting*

We split 2,636 labeled question samples into three sets for training, evaluating, and testing a model. First, the training set had 60% (1,581 samples) for training the model directly. Second, the validation set had 20% (527 samples) for evaluating the model during the training and also use for tuning hyperparameters that will be described in the next section. Third, the test set had 20% (528 samples) for evaluating the model after training.

E. *Hyperparameters tuning*

Hyperparameters are tuned for optimizing the model, to find closely value of hyperparameter that yields the highest performance. In this article, we tuned four hyperparameters which include 1) the number of hidden units, 2) embedding dimension, 3) regularization value, and 4) learning rate. We used a random search for the tuning process because several hyperparameters were available to tune and some of them were much more important than others. Then we keep rescaling and tuning until the performance is not improved.

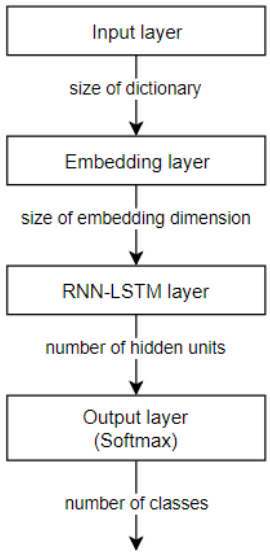

Fig. 6. Neural network layers and dimensions of data through the network

IV. Experiment and Evaluation

In the previous section, we described the process of preparing data for training the AI, designed the model of AI and chose hyperparameters for tuning. In this section, we will explain how to train and find the value of hyperparameters which optimize the model. Then, we will evaluate the optimized model with the test set of data and set a probability threshold for picking the right answers.

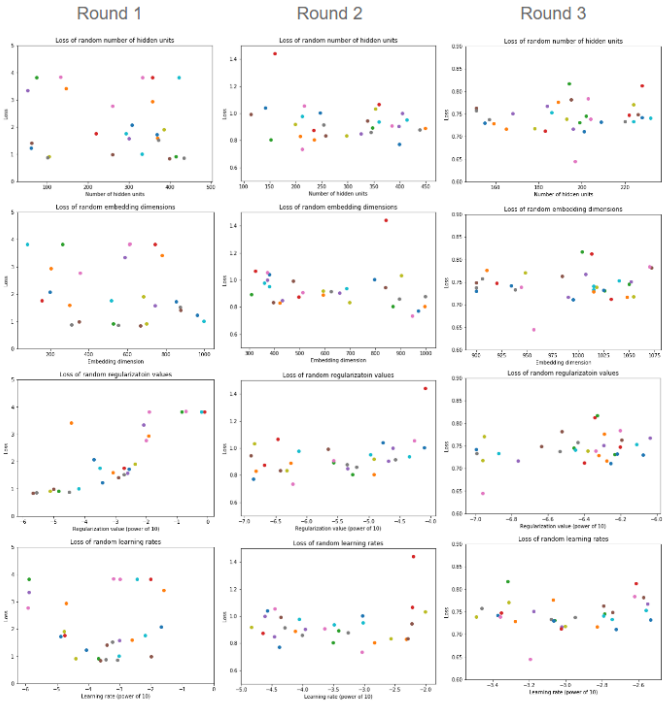

Fig 7. Graphs between hyperparameters and loss in the tuning process

A. *The result of tuning hyperparameters*

After we tuned hyperparameters for three rounds, with 30 random samples each round as shown in Fig 7., we then achieved optimized hyperparameters. There are 197 hidden units, 956 embedding dimensions, 10-6.96 for regularization value and 10-3.19 for learning rate. Using this hyperparameters value, the training model had 83.4% accuracy with the validation set of data

B. *Performance evaluation*

The performance evaluation with the test set of data had 83.9% accuracy. As we used the softmax activation function at the output layer of the model, we obtained the outputs as a probability distribution of categories and a predicted category is a category which has the highest probability. Then we have analysed the correct and incorrect outputs shown in Fig. 8 and Fig. 9. The average max probability of correct output is 0.92, and incorrect output is 0.48. Therefore, we set the probability threshold at 0.5 then evaluated the model with that threshold again with the same test set of data. As a result, 13.64% of questions were ignored because correct answers for such questions were not found. On the other hand, 86.36% of questions were processed with 93.2% accuracy for being the correct answers.

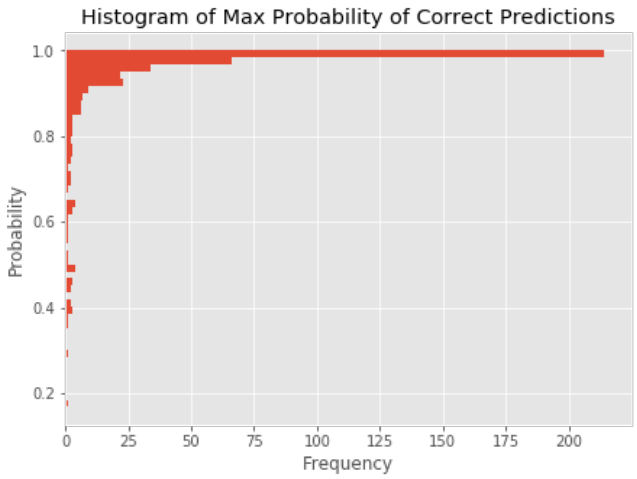

Fig. 8. Correct outputs analysis

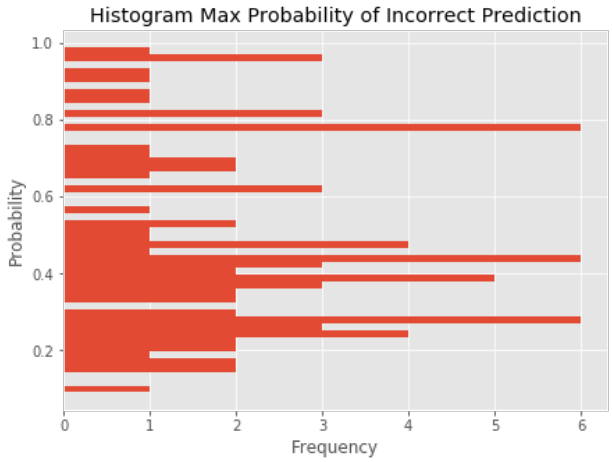

Fig. 9. Incorrect outputs analysis

V. Conclusion

We have developed an FAQ chatbot with LSTM in an artificial neural network model. Our chatbot takes Thai text question and provides answers if the output probability more than the threshold value to give the correct answers. We have evaluated the system performance and found that the developed chatbot system could process 86.36% of questions with 93.2% accuracy of correct answers. For future work, we are interested in improving the accuracy of the classification model.

References

[1] Statista Inc, "Global number of digital buyers 2014-2021", https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/.

[2] Chatfuel Inc, "Chatfuel", https://chatfuel.com/.

[3] Joseph Weizenbaum, "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine", Association for Computing Machinery (ACM), 1966.

[4] Bayan AbuShawar, Eric Atwell, "ALICE Chatbot: Trials and Outputs", Computación y Sistemas, Vol. 19, No. 4, 2015, pp. 625–632.

[5] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, "Sequence to Sequence Learning with Neural Networks", arXiv preprint arXiv:1409.3215v3, 2014.

[6] Minh-Thang Luong, Hieu Pham, Christopher D. Manning, "Effective Approaches to Attention-based Neural Machine Translation", arXiv preprint arXiv:1508.04025v5, 2015.

[7] Ryan Lowe, Nissan Pow, Iulian V. Serban, Joelle Pineau, "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems", arXiv preprint arXiv:1506.08909v3, 2016.

[8] Christopher Olah, "Understanding LSTM Networks", http://colah.github.io/posts/2015-08-Understanding-LSTMs/.