

NATURAL LANGUAGE PROCESSING FOR QUESTION-ANSWERING SYSTEM

Chalermpol Tapsai¹ & Wilailuk Rakbumrung²

*College of Innovation and Management, Suan Sunandha Rajabhat University, Thailand,
Email: chalerm.pol.ta@ssru.ac.th¹, wilailuk.ra@ssru.ac.th²*

ABSTRACT

Information is essential to create understanding and knowledge among stakeholders for organization management. The Question-Answering system is an important service that provides accurate information for users and stakeholders that can drive the organization to achieve its objectives. This research study on Natural language processing to analyze question sentences and predict the appropriate answer for each question using the Deep learning technique. The experiment was conducted to compare the prediction results between the analysis with and without data pre-processing. The results show that the accuracy of predictions using pre-processed data is higher than raw data. In addition, some techniques, such as ontology and similarity, can be applied to infer the meaning of words and sentences, which improves the results and presents more precise predictions.

Keywords: Natural language processing, Question-Answering, Deep learning.

INTRODUCTION

Information management is very important in business management. Dissemination of accurate and up-to-date information is an essential factor that helps organizations develop their business to be successful. With good information presented and a question-answering system (Q/A system), the organization can offer essential information about the organization, products, and services. Managing an organization to meet its business objectives requires an efficient system that facilitates users to access the information they need easily and quickly. Accurate and up-to-date information will help create a good image for the organization. It also improves marketing efficiency, service quality, and customer satisfaction.

With online technologies, information management has been improved to facilitate users to access the information they need from anywhere and anytime. Unfortunately, to get the relevant information on the website, users still have to manually search for information by typing keywords, clicking menus, or sending a message to ask the staff who manages the website. Most of these processes may take a long time to search and wait for a response. Even though some websites have provided a Q&A system that allows users to make a real-time conversation with their staff. The quality of service and information may differ based on the ability to provide information and answer questions of each staff. And sometimes users have to wait for a long time due to a large number of people using the service. To overcome this problem, the development of an automated Q&A system that analyzes questions using text analysis and Natural language processing (Tapsai, C., et al., 2021) and responds with the answers in a short time is an important tool for communicating and disseminating corporate information to customers and stakeholders.

LITERATURE REVIEW

Currently, the chatbot is the most popular Q/A system that has been widely researched for many years. Some examples of research on chatbot applications and their impact on business are, the application of chatbots in customer compliance (M. Adam, et al., 2021), the impact of chatbots application on customer service (Misischia, C. V., et al., 2022), (Rossmann, A., et al., 2020), application of chatbots on providing personalized customer care (Chung, M., et al., 2020), etc. Some techniques such as Similarity and Machine learning were applied to creating chatbot Q/A systems. For the similarity technique, the inputted questions and a dataset of question-answer are analyzed and compared to define the answer that has the closest similarity value to the inputted question (Shrivastava, R., et al., 2023). Machine learning is the technique that builds the answer-prediction models by processing the learning dataset of question-with-answer (Mohana, R Madana, et al., 2021). Even though the result of using these techniques is good, there are still limitations in terms of these techniques that need human operators to analyze and define key features (features selection) of the learning dataset that is used to create a predictive model. This is an important process that greatly affects the accuracy of the prediction results.

In this research, we present the development of the chatbot using the Deep Learning method, which is an advanced machine learning in which computers will process the learning dataset to create predictive models without human intervention. We will present the detail of the research methodology in the following section.

RESEARCH METHODOLOGY

This research is experimental research. By using the Deep Learning method, two prediction models were created by learning datasets with and without data pre-processing. Then, both models were evaluated to compare the accuracy of the prediction results. We divided the research process into 3 steps:

Step 1 was data collection, which was divided into 2 parts: The first part is collecting data from a website to be used as experimental data. We randomly selected a tourist attraction website from the list of tourist attractions websites in Thailand. Then collected the information about tourist attractions presented on that website, such as general information, characteristics of the tourist attraction, background, travel, and other related information, etc. The second part is the data collection for the learning dataset and the test dataset. The sample group of this research was 50 people who voluntarily participated in the research. The researcher asked each sample to read the experimental data and then write 50 questions and write an answer to each question based on the experimental data. Then, all data with a total of 2500 questions were divided into 2 parts. The first part, with a total of 2250 questions (90% of the total data), was the learning dataset used for creating predictive models. The remaining 250 questions (10%) were used as the test dataset used for model evaluation.

Step 2 was modeling. We created 2 prediction models for the Q/A system (chatbots) using the Deep Learning method. The first model was created from the learning dataset without data pre-processing. The second model was created from the learning dataset with data pre-processing that deletes all stop words from the question sentence in the learning dataset.

Step 3 is the evaluation of models. In this step, both models were evaluated by inputting the test dataset and collecting the prediction results for calculating the accuracy value. Then, we compare the accuracy values between the two models.

RESULTS

The prediction results of both models are shown in Table 1.

Table 1 The Experimental result of model evaluation

Prediction models >	Without data pre-processing	With data pre-processing
True Positive (a)	216	228
False Negative (b)	0	0
False Positive (c)	34	22
True Negative (d)	0	0
Accuracy	0.86	0.91

The experimental results show that the model1 created from the learning dataset with data pre-processing that removed stop words from the question sentence could predict the answer with an accuracy of 0.91 higher than the model2 created from data without data pre-processing with an accuracy of 0.86. The number of incorrect predictions of model1 and model2 was 34 and 22 respectively. The overall incorrect prediction of two models after eliminating duplicate questions was 36 questions.

CONCLUSION AND DISCUSSION

The prediction results of the model created from the learning dataset that eliminates stop words from the question sentence provide higher accuracy prediction because stop words are the basic words that are used in most sentences and cannot represent the difference between questions. Therefore, when used in learning datasets, it results in more inaccurate prediction results.

When considering the cause of the error, we found 12 questions with misspelled words, 5 questions with synonyms or words that do not appear in the learning dataset, and 4 questions with words that need to infer the meaning, such as the word “more than”, “less than”, etc. For future research, we will cover these errors to improve the accuracy of prediction by applying many techniques such as Misspelling correction will be applied to cover various types of misspellings (Tapsai, C., 2018). Ontology and similarity will be applied to convert synonymous words to basic words and analyzed words to infer the meaning before inputting to the model for prediction.

ACKNOWLEDGEMENTS

The author would like to thank the Research and Development Institute, Suan Sunandha Rajabhat University, Bangkok, Thailand for research support.

REFERENCES

- Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance”, *Electronic Markets*. vol. 31, no. 2, pp. 427-445, 2021.
- Chung, M., et al. (2020). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research* 117: 587-595.

- Misichia, C. V., et al. (2022). Chatbots in customer service: Their relevance and impact on service quality. *Procedia Computer Science* 201: 421-428.
- Mohana, R., Nagarjuna, P., & Kumari, P L. (2021). Customer Support Chatbot Using Machine Learning. 10.1007/978-981-15-5679-1_42.
- Rossmann, A., Zimmermann, A., & Hertweck, D. (2020). The Impact of Chatbots on Customer Service Performance. *Advances in the Human Side of Service Engineering*, 237–243.
- Shrivastava, R., Pujahari, A., Singh, S.P., & Bhowmik, T. (2023). Efficient Question Answering in Chatbot Using TF-IDF and Cosine Similarity. In: *Innovations in Information and Communication Technologies. Algorithms for Intelligent Systems*. Springer, Singapore. https://doi.org/10.1007/978-981-19-3796-5_3
- Tapsai, C. (2018). Analysis of patterns and causes of misspelling and slang words for natural language processing. In *Proceedings of 135th The IRES International Conference*, Moscow, Russia, pp. 25-30.
- Tapsai, C., Unger, H., & Meesad, P. (2021). *Thai Natural Language Processing, Word Segmentation, Semantic Analysis, and Application*. Cham, Switzerland: Springer International Publishing. 10.1007/978-3-030-56235-9.