Automated customer consultation system for Pastry Shops

Trung Quoc Nguyen[1*]

[1]FPT University, Ho Chi Minh City, Vietnam
[*]Corresponding author: trungnq46@fpt.edu.vn

ARTICLE INFO

DOI:10.46223/HCMCOUJS.tech.en.15.2.4409.2025

Received: May 19th, 2025

Revised: July 25th, 2025

Accepted: July 29th, 2025

Keywords:

BiLSTM; BLEU Score; Chatbot; LSTM; MaLSTM; Natural Language Processing; Q-A systems

ABSTRACT

Automated customer consulting is a form of automated customer care and consulting that utilizes texting and chat functions to replace human interaction. This research improves the Bi-LSTM language model. We aim to enhance the accuracy and applicability of an automated customer consultation system, which may impact enterprises and traders. Our question-answer system uses querying the entity and model textual similarity to match models. Automated customer care systems utilize computers or other technologies to assist customers. It empowers clients to address problems without human assistance in customer care. Human resources can address complex requests or high-value consumers, as automation handles many repetitive and straightforward activities. Many firms utilize it, especially fast-growing ones that need to arrange support.

1. Introduction

In some instances, the sales staff at the pastry shops cannot respond quickly to a large number of customers at the same time; in other cases, customers need advice but do not have time to go to the bakery to learn about the types of cakes; and in other cases, when the sales team at the bakery cannot satisfy many customers at once. Our team recognized the need to develop a solution to the problem described above, and we conceived the idea of a Question Answering System (QAS) technique that can be applied in pastry shops. Question-Answering Systems (QAS) are highly beneficial because most deep-learning problems can be thought of as question-answering problems. As a result, this topic is now one of the most researched in computer science. The system can also be trained to understand multiple languages, enabling it to reach a broader audience. One of the most challenging aspects of designing a QA system is accurately understanding user queries or questions. Users can ask questions in various ways and use informal language that may or may not adhere to conventional grammar rules. The system must be able to determine the underlying intent of the question and receive appropriate responses from the knowledge base. NLP algorithms are used to analyze the user's words and choose the main topic of the question.

Our proposed Hierarchical Bi-LSTM Attention Model (HBAM) significantly surpasses conventional designs, such as MaLSTM and Bi-LSTM, by incorporating Bi-LSTM and attention mechanisms within a Siamese framework. It attains superior accuracy, precision, and BLEU scores on an extensive Vietnamese cake consultation dataset. This improved performance demonstrates the effectiveness of our algorithm in understanding diverse consumer inquiries and providing accurate automated responses.

Together, these elements process queries and documents at multiple levels until the correct response is found. For example, if the results of analyzing the questions are poor, then analyzing the answers will certainly yield poor results. Similarly, a good question analysis result does not guarantee a good answer analysis result, and vice versa. As a result, many scientists focus solely on one component of QAS.

We used a hybrid quality assurance technique that combines Query the entity (Hasibi et al., 2017) with an NLP model, which was novel. The entity will be queried in response to a user query. If it fails, a text similarity model will be used to discover answers from a large pastry shop QA dataset. Deep learning and similarity comparison are used to represent the text in this article, which is its best work. The model is HBAM, which has two layers: a Bi-LSTM and a word attention layer. The Bi-LSTM layer collects sentence data in both forward and backward orientations. The attention layer identifies keywords in a phrase.

In business pastry consulting, the Siamese frame and Manhattan distance are used to calculate semantic similarity. The Siamese framework is popular in metric learning (Chen & Salman, 2011; Yih et al., 2011). Comparing the text cosine similarity index (Yih et al., 2011) to the Manhattan city distance. Our HBAM outperformed MaLSTM (Mueller & Thyagarajan, 2016) and Bi-LSTM (Huang et al., 2015) in various testing methods and data sets.

2. Related works

In the following, we define two problems that are at the center of the chatbot system. Realizing the capacity to understand natural language, or creating the required mechanisms to enable a software system to comprehend natural language queries in the same way that a person would, is the first challenge. The second challenge seeks to get pertinent data from a domain-specific database to provide solutions that may be returned to the user:

– Understanding user questions (Intent Detection): To comprehend and handle a user's inquiry, Natural Language Processing (NLP) and Natural Language Understanding (NLU) are used.

– Knowledge base retrieval and storage: The ability to store and query medical queries and answers using a domain knowledge database.

We examine previous research that addresses the two issues mentioned above Chatbots. Joseph (1966) created Eliza, the world's first chatbot, at the MIT Artificial Intelligence Laboratory in 1966. But Eliza doesn't grasp the user's inquiry. Psychiatry professor Kenneth Colby's Turing Test was first completed by Parry in 1972 (Cerf, 1973). However, just 48% of psychiatrists can accurately identify the actual patient based only on their discussion. The multiple-turn dialog decision tree was attempted to be used by Ni and Liu (2018) and Ni et al. (2017) to make decisions on behalf of a patient. According to Zhao and Liu (2018), the accuracy of the model on shorter context talks may be successfully increased by using transfer learning to transfer typical cases from SQuAD to Bible QA. Using N-gram approaches, which effectively minimize the data noise, Dai et al. (2016) developed a "focused pruning method" to limit the candidate result space and make some improvements. "APVA" was introduced by Wang et al. (2018) to forecast the relationship between question and response entities precisely. A novel framework for semantic analysis was suggested by Yih et al. (2015); after the question has been translated and examined in query language, the new inquiry will be connected to the knowledge base. To increase performance in 2017, Yu et al. (2017) developed a hierarchical RNN network that employed residual learning. It can

identify the relationship inside the knowledge base when an input query is present. Additionally, they created a straightforward KBQS system that incorporates connection detection and entity linking.

2.1. Siamese-based semantic sentence similarity

A Siamese Long Short-Term Memory (LSTM) network has been suggested by Mueller and Thyagarajan (2016) to calculate the semantic similarity between two variable-length phrases. LSTM, however, is unable to identify keywords inside a phrase. A Siamese design with bidirectional Long Short-Term Memory (LSTM) networks and an attention mechanism was suggested by Baziotis et al. (2017). To capture both two-directional contexts, the model utilizes bidirectional long short-term memory. To classify, they take into account the fully connected (tanh) in the last layer, which may lead to overfitting.

2.2. Word Embedding

One-hot encoding: Bagui et al. (2021) encode categorical data to enhance the predictiveness of machine learning algorithms. Categorical values are numerical representations of dataset categories. The one-hot encoder displays categories as binary, with 0 indicating the absence of a feature and 01 indicating its presence. The model is trained using this binary feature vector. A one-hot encoding often represents a machine situation. A decoder determines the machine's binary or gray code status. One-hot machines don't need a decoder to identify their state; hence, this doesn't apply. A bit set to a high value defines the n-th state of an element.

A one-hot vector in the corpus or dictionary is comparable to a feature vector where each feature is assigned a value of 0 or 1, indicating the presence or absence of a word based on its dictionary word number. The feature vector contains all dictionary terms and their indexes. Words and indices are retained at the same feature vector index. Dictionary embedding of feature vectors yields a binary vector.

This project uses the skip-gram model (Mikolov et al., 2013) from the word2vec method. Data was used to train Gensim's word2vec Python package. Specific hyperparameters were focused on to improve word embeddings. The parameters include training method, dimensionality, context window, and subsampling. Negative sampling was employed in this project because it is more computationally efficient than hierarchical softmax. The hidden layer of the neural network was expanded to 300, improving word embeddings. A sub-sampling rate of 1e-3 was used to balance the sample's uncommon and common words.

Given a large amount of data, a minimum count of 01 was established to ensure that every word in the corpus was accounted for throughout the training process. The word2vec model underwent training using the processed data, employing the hyperparameter values indicated earlier. The resulting model was then stored in a file format. The word vectors generated by the word embedding model possess a dimension of m*n, where m represents the size of the dictionary and n represents the size of the hidden layer.

3. Automated customer consultation system for pastry shops

3.1. Materials

Virtual Sales-Assistant Systems: In general, a virtual sales assistant system serves as a guide for navigating various online resources. The content of the current website or catalog is presented in response to the input sample.

Customer input is a customer inquiry. The semantic similarity computation uses the word embedding from the consumer's inquiry. It will discover the highest K rating index and deliver the results. No results will be returned if the customer's question cannot be determined or the rating index is too low. New product data will be updated via inquiries. It enables customers to retrieve bakery product data using concise queries that include pricing, images, and key information. The customer will be routed back to the original site to reenter the search query if it cannot be resolved. Catalog extensions provide additional product information and recommendations to help buyers make informed decisions.

Automated customer consulting for pastry shops aims to minimize human involvement in customer service and consultation by utilizing automated texting and chat services. This project will use the Bi-LSTM language model to enhance word production by providing accurate contextual meanings. We aim to improve the precision and applicability of an effective automated customer consultation system, which could impact commercial organizations and trade firms. Our chatbot system uses Query the entity and model textual similarity to match models. We also used F1 Score, BLEU Score, Precision, and Recall to evaluate our technique. The Question_Duplicated data collection contains almost 337,103 similar questions collected via data augmentation.

3.2. Methods

Processing data: In light of the absence of publicly accessible datasets in the pastry sector, we developed a bespoke Vietnamese-language dataset for cake consultation. Initially, 5,230 question-answer pairs were meticulously generated based on prevalent consumer inquiries concerning the purchase, sale, and consultation of cake items. All responses were meticulously crafted to guarantee semantic precision and cultural appropriateness. To mitigate the restricted dimensions of the initial dataset, we employed data augmentation methods, producing five semantically analogous variations for each original question. This led to an augmented training set including 337,103 question-answer pairs, designated as the Question_Duplicated Data. Furthermore, we extracted product data from the official website of Thien Thuan Phat Bakery, encompassing 17 varieties of cakes, including their names, prices, descriptions, and photographs. This dataset, tailored for the store, accurately represents sales and consulting scenarios and has been meticulously examined to conform to the Vietnamese language and environment.

Data Augmentation (DA) generates multiple copies of the current dataset to increase the training data size without requiring additional data collection. To improve classification performance, the data must be modified to preserve class categories. Computer vision and Natural Language Processing (NLP) utilize data augmentation technologies to address issues related to data availability and diversity. Augmented photos are easy to create, but Natural Language Processing (NLP) is challenging due to the complexity of language. Since it would change context, replacing every word with its synonym is not an option. Data augmentation increases the training dataset, improving the model's performance. An enhanced data distribution should strike a balance between similarity and dissimilarity to the original data. Data augmentation methods should strike a balance to avoid overfitting and poor performance.

Manhattan LSTM Model

The structure of the proposed Manhattan LSTM (MaLSTM) model is depicted in Figure 1. In this research, we focus solely on Siamese architectures with linked weights, specifically LSTMa and LSTMb. These two networks individually analyze one phrase from a

given pair of phrases. It is important to note that LSTMa and LSTMb are identical in this context, meaning that they have the same weights. However, the overall unbound version of this paradigm may be more advantageous for applications with asymmetrical domains, such as information retrieval, where search queries have distinct stylistic differences from stored texts. To calculate the semantic similarity between two sentences using MaLSTM, we use the Manhattan distance between their hidden representations as follows:

$$Similarity(Q_1, Q_2) = exp(-\| h_{Q_1} - h_{Q_2} \|_1) \quad (1)$$

Where h_{Q_1} and h_{Q_2} *these* are the output vectors from the LSTM encoder for the two input sentences.

Hierarchical Bi-LSTM Attention model: The schematic illustrating our newly suggested hierarchical Bi-LSTM Attention model (Bao et al., 2020). Its purpose is to facilitate the comparison of semantic similarities. The entire architecture is built upon a Siamese LSTM framework (Mueller & Thyagarajan, 2016). The Siamese structure incorporates a single Bi-LSTM layer and a single word attention layer. The sentences on the bottom left and right show the user's input query and the question from the QA dataset. The two inquiries will be demonstrated by employing word embedding first, followed by utilizing Bi-LSTM (Schuster & Paliwal, 1997) to construct the whole phrase embedding, taking into account the surrounding context. The attention weight αt for each time step t is computed using the following formulation:

$$e_t = v^T \tanh(Wh_t + b) \quad (2)$$

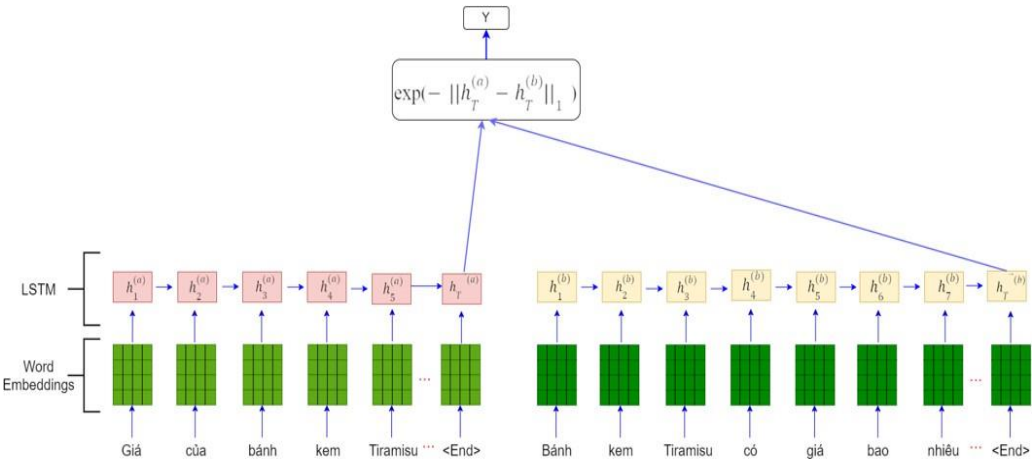
$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \quad (3)$$

Where h_t is the hidden state at time step *t*, *W*, *v*, and *b* are learnable parameters.

This research presents a novel hybrid framework, the HBAM model, which integrates attention mechanisms designed explicitly for brief, domain-specific Vietnamese consultation inquiries, while building upon established architectures such as MaLSTM and Bi-LSTM. Furthermore, by developing a tailored dataset for pastry-related customer interactions and implementing data augmentation, we ensure the model is refined for practical application. This integration of architectural adaptation and contextual training differentiates our methodology from previous research.

Figure 1

Overview of the Hierarchical Bi-LSTM Attention Model Architecture



Note. The researcher's data analysis

The HBAM model architecture was selected to effectively capture both sequential dependencies and key semantic elements in user queries, which are challenges that traditional models often struggle with, particularly in short, domain-specific sentences. HBAM addresses this by combining a bidirectional LSTM layer, which encodes contextual information in both directions, with an attention mechanism that emphasizes the most informative words in each phrase.

4. Experiments and results

4.1. Dataset

The Cake Question and Answer dataset was manually constructed from 5,230 Vietnamese question-answer pairs, reflecting typical customer interactions related to purchasing and consulting about cakes. All entries were carefully selected and reviewed to ensure semantic clarity and cultural relevance for Vietnamese users. After preprocessing and deduplication, the final dataset retained 5,230 high-quality consultation pairs suitable for model training. Question: A duplicated dataset is a set of sentence pairs in the open domain. It has 337,103 pairs of sentences tagged with a format like "text1 text2 is duplicate," which means whether the two sentences are semantically similar. If they have equal semantic meaning, then the tag will be "1"; otherwise, it will be "0".

We collected product data from the official website of Thien Thuan Phat Bakery (https://thienthuanphatbakery.com/), including 17 types of cakes along with their names, prices, images, and descriptions. This information was used to create a domain-specific dataset that reflects authentic sales and consultation scenarios. All data entries were manually reviewed to ensure relevance, consistency, and alignment with the Vietnamese cultural and linguistic context.

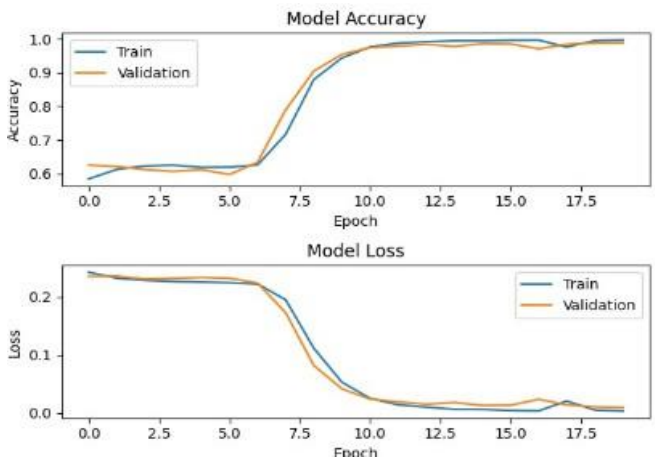
4.2. Train HBAM, MaLSTM, and Bi_LSTM models

In our experiments, we trained three models on the Question_Duplicated dataset, which contains 337,103 pairs of questions labeled as 1 (same) and 0 (different). However, we adjust the batch size, epoch, and max_seq_length to be compatible with the Question_ Duplicated dataset.

Figure 2's Siamese-based HBAM model utilizes a bi-directional long short-term memory (Bi-LSTM) layer and word-level attention. Bi-LSTM contextual embeddings and semantic emphasis are used to process the user query and candidate response. A distance measure, such as the cosine or Manhattan distance, is used to calculate a similarity score between sentence representations for accurate semantic matching.

Figure 2

Training and Validation Performance of the Hierarchical Bi-LSTM Attention Model over Epochs

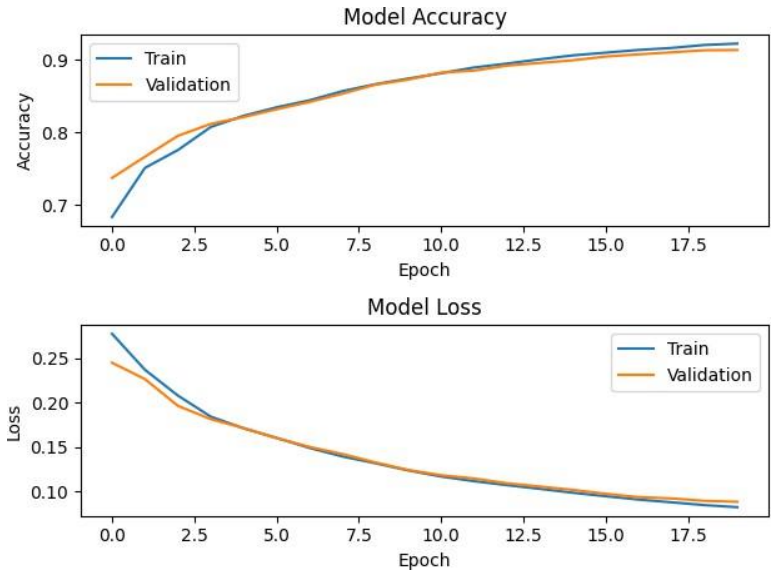


Note. Data analysis results of the research

Figure 3's MaLSTM model processes two input sentences with shared weights using a Siamese LSTM architecture. LSTM outputs are evaluated using Manhattan distance and processed exponentially to create similarity scores. This method captures sentence-level semantic similarity but does not weight key phrases.

Figure 3

Training and Validation Performance of the MaLSTM Model over Epochs

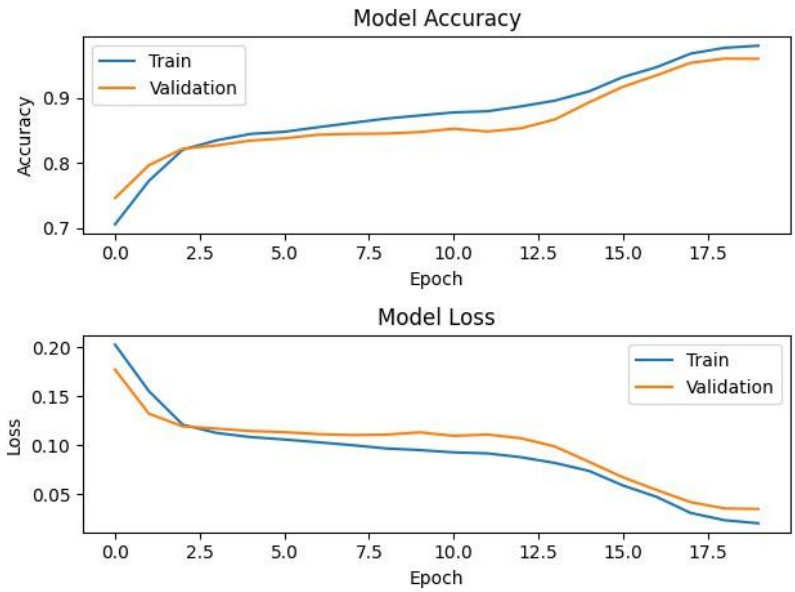


Note. Data analysis results of the research

The Bi-LSTM model, depicted in Figure 4, functions as a standard benchmark in sentence embedding tasks. It analyzes each input sequence bidirectionally to capture contextual information. This model, unlike MaLSTM or HBAM, lacks a Siamese architecture and an attention mechanism. It serves as a standard for evaluating the contributions of more complex systems such as HBAM.

Figure 4

Accuracy and Loss Curves for Training and Validation Phases of the Bi_LSTM Model



Note. Data analysis results of the research

Setting parameters of the training model: When setting up the parameters to fine-tune the HBAM, MaLSTM, and Bi-LSTM models, we choose an Epoch of 20 for training and set the batch size for both evaluation and n_hidden to 50. Additionally, we set the Gradient Accumulation parameter to increase the batch size of the model to 1,024. Instead, we increase the batch size via gradient accumulation, as directly increasing the batch size would result in increased GPU memory usage. In this training session, we utilized a GPU T4 from Google Colab, which has approximately 13 GB of GPU memory. This means we cannot increase the batch size further without running into memory overflow problems. Therefore, the Gradient Cumulative setting ensures that the batch size can increase up to a maximum of 1,024. Additionally, we also want to ensure that the batch size is not too low, as a batch size that is too low will lead to problems with loss of function during training.

4.3. Final results of the similarity comparison

BLEU Score: BLEU Score is an evaluation metric for Machine Translation tasks. It is calculated by comparing the n-grams of machine-translated sentences to the n-grams of human-translated sentences. Usually, it has been observed that the BLEU score decreases as the sentence length increases. This, however, might vary depending on the model used for translation. Here, we use the BLEU Score to compare the answer results between customers and the actual question sets. BP stands for Brevity Penalty, which penalizes the score when the Machine Translation is too short compared to the reference (correct) translations. The BLEU score used to evaluate the similarity between generated and reference sentences is given by:

$BLEU =$ BP*exp($\sum_{n=1}^{N} w_n \, logp_n$(4)

Where BP is the brevity penalty, pn is the precision for n-grams, and wn are the weights (typically uniform).

Comparison results between models: We selected a separate testing dataset from the training dataset to conduct an objective model evaluation process. There are 2,840 pairs of similar and distinct texts in our test data set. Table 1 presents the comparison results generated using the test data set.

Table 1

Precision, Recall, F1-Score, and BLEU-Score of the MaLSTM, BiLSTM, and HBAM Models on the Test Set

Models	Precision	Recall	F1-Score	BLEU-Score
MaLSTM	0.88	0.87	0.86	0.85
BiLSTM	0.96	0.95	0.95	0.95
HBAM	0.97	0.97	0.97	0.96

Note. Data analysis results of the research

Table 1 demonstrates that the HBAM model regularly surpasses both BiLSTM and MaLSTM in all evaluation metrics. HBAM demonstrates exceptional proficiency in comprehending semantically similar client requests, achieving an F1-score of 0.97 and a BLEU score of 0.96. This performance highlights the efficacy of combining attention mechanisms with bidirectional context modeling, which markedly improves sentence-level semantic understanding - especially vital for managing informal and succinct texts, such as those encountered in customer service communications.

These enhancements lead to increased customer satisfaction, a decrease in unresolved inquiries, and a more effective distribution of human resources in retail settings. Furthermore, the model operates continuously, making it particularly suitable for online or hybrid pastry enterprises that require scalable and always-accessible customer support solutions.

5. Conclusions

A Vietnamese QA dataset and HBAM, a hybrid deep learning model, are utilized to offer an automated customer consultation system for pastry stores. The 5,230 selected pairings and 337,103 augmented pairs provide a basis for developing Vietnamese language chatbots. The HBAM model, which integrates Bi-LSTM and attention inside a Siamese framework, demonstrates superior precision, recall, F1-score, and BLEU score. Its capacity to address succinct inquiries renders it advantageous for retail guidance. We adapt and enhance existing architectures for the Vietnamese business environment. Enhancements in scalability and user experience will be achieved through expansion into additional regions and the incorporation of multilingual and voice-activated functionalities.

NO CONFLICT OF INTEREST STATEMENT

The author declares that they have no conflict of interest.

References

Bagui, S., Nandi, D., Bagui, S., & White, R. J. (2021). Machine learning and deep learning for phishing email classification using one-hot encoding. *Journal of Computer Science*, *17,* 610-623. https://doi.org/10.3844/jcssp.2021.610.623

Bao, Q., Ni, L., & Liu, J. (2020). Hhh: An online medical chatbot system based on knowledge graph and hierarchical bi-directional attention. In *Proceedings of the australasian computer science week multiconference* (pp. 1-10). ACM Digital Library.

Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017). Datastories at semeval-2017 task 4: Deep lstm with attention for message level and topic based sentiment analysis. In *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* (pp. 747-754). Association for Computational Linguistics.

Cerf, V. (1973). *Parry encounters the doctor*. https://doi.org/10.17487/RFC0439

Chen, K., & Salman, A. (2011). Extracting speaker specific information with a regularized siamese deep network. *Advances in Neural Information Processing Systems*, *24,* 298-306.

Dai, Z., Li, L., & Xu, W. (2016). *CFO: Conditional focused neural question answer- ing with large-scale knowledge bases*. https://arxiv.org/abs/1606.01994

Hasibi, F., Balog, K., & Bratsberg, S. E. (2017). Entity linking in queries: Efficiency vs. effectiveness. *Lecture Notes in Computer Science*, *10193*, 40-53. https://doi.org/10.1007/978-3-319-56608-5_4

Huang, Z., Xu, W., & Yu, K. (2015). *Bidirectional LSTM-CRF models for sequence tagging*. https://arxiv.org/abs/1508.01991

Joseph, W. (1966). Elizaa computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111-3119.

Mueller, J., & Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the aaai conference on artificial intelligence* (vol. 30). ACM Digital Library.

Ni, L., & Liu, J. (2018). A framework for domain specific natural language information brokerage. *Journal of Systems Science and Systems Engineering*, *27*, 559-585.

Ni, L., Lu, C., Liu, N., & Liu, J. (2017). Mandy: Towards a smart primary care chatbot application. In *International symposium on knowledge and systems sciences* (pp. 38-52). Springer.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673-2681.

Wang, Y., Zhang, R., Xu, C., & Mao, Y. (2018). The apvaturbo approach to question answering in knowledge base. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1998-2009). Association for Computational Linguistics.

Yih, S. W., Chang, X. H. M. W., & Gao, J. (2015). Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the joint conference of the 53rd annual meeting of the acl and the 7th international joint conference on natural language processing of the afnlp.* Association for Computational Linguistics.

Yih, W.-T., Toutanova, J. C. P. K., & Meek, C. (2011). Learning discriminative projections for text similarity measures. In *Proceedings of the fifteenth conference on computational natural language learning* (pp. 247-256). Association for Computational Linguistics.

Yu, M., Yin, W., Hasan, K. S., Santos, C. d., Xiang, B., & Zhou, B. (2017). *Improved neural relation detection for knowledge base question answering*. https://arxiv.org/abs/1704.06194

Zhao, H. J., & Liu, J. (2018). Finding answers from the word of God: Domain adaptation for neural networks in biblical question answering. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1-8. https://doi.org/10.1109/IJCNN.2018.8489756