ICIC Express Letters
Volume **16**, Number **9**, September **2022**

ICIC International ©2022 ISSN 1881-803X
pp. 1011–1017

LEARNING CONTEXTUAL MEANING FOR QUESTION RETRIEVAL USING SIAMESE LSTM ON ISLAMIC QUESTION ANSWERING SYSTEM

Nurjayanti, Adiwijaya* and Said Al Faraby

School of Computing
Telkom University
Jalan Telekomunikasi No. 1, Bandung 40257, Indonesia
{jayanti; saidalfaraby}@telkomuniversity.ac.id
*Corresponding author: adiwijaya@telkomuniversity.ac.id

Received August 2021; accepted November 2021

Abstract. *One of the breakthroughs in Question Answering (QA) development is the usage of neural networks to solve question retrieval task. The neural network model typically takes as input dense, low-dimensional vectors that model the context. We proposed to use a pre-trained word embedding and fed them into the Siamese Long Short-Term Memory (LSTM) model to understand the contextual meaning in the questions. The model predicts the question similarity in the final layer using the Manhattan function. The proposed QA achieved precision@5, recall, and Mean Average Precision (MAP): 0.4820, 0.9508, and 0.8463, respectively.*

Keywords: Question answering, Question retrieval, Siamese LSTM, Word embedding

1. **Introduction.** For Muslims learning about Islam has never been easier in today's Digital Age. Nevertheless, not all available on the Internet can be said to be from credible sources. Muslims need to be careful about where they learn their religion. Though the Holy Quran provides complete and thorough teaching in Islam [1], most people prefer getting the answer to their religious questions from the Internet even when most search method cannot satisfy their need for exact and specific information. They bear the burden of tedious searching through all the documents to find the answer [2]. Building an Islamic QA system to provide a brief and precise answer to a question in natural language has been a long-standing research problem for its apparent practical value. Islamic QA research is carried out on various sources of knowledge including the Holy Quran [1, 2, 3, 4] and hadith [5]. The architecture of QA systems, in general, has three components: question analyzer, document retrieval, and answer analyzer [6]. The recent approach is to use question retrieval rather than documents collection [1, 7]. The question retrieval aims to rank all the questions and return the top similar to the query given a query question and all the stored questions in a knowledge base. While the question retrieval commonly used in community QA systems (cQA) [8, 9], the effectiveness of question retrieval for closed-domain QA is relatively unstudied.

Conventional techniques, such as the pattern match [1], and average of vectors [11], are commonly used in question retrieval due to their fast and easy implementation. However, the neural networks technique promises more robust results and better performance on this task [7, 9], due to its ability to proceed beyond word matching and omit the feature engineering phase. Motivated by these neural networks models tremendous success, we proposed a question retrieval using neural networks model to predict the similarity between the user questions and knowledge base. Our neural networks model for question retrieval tasks is inspired by previous work [13, 14] that achieves impressive performance in

DOI: 10.24507/icicel.16.09.1011

Semantic Textual Similarity (STS). The architecture of neural networks has two identical, Siamese Long Short-Term Memory (LSTM) networks [12, 13, 14]. These networks learn contextual meaning from question representations, a fixed-length dense vector created by mapping each word and its vector using word embedding.

The remainder of this paper is organized as follows. Section 2 describes in detail our proposed QA system. Section 3 presents experimental results on datasets and discussions. Finally, some conclusions are stated and discussed in Section 4.

2. **Proposed Method.** The proposed question retrieval is utilizing pre-trained word embedding to understand the contextual meaning in the questions. The neural networks model is used to measure the similarity score for the user question and relevant questions found in the knowledge base. As shown in Figure 1, the architecture of the proposed QA system has three components: 1) question preprocessing, this phase uses word embedding to transform questions into a representation of fixed size vector; 2) pre-trained ManLSTM model where the similarity score is calculated using a pre-trained Manhattan Siamese LSTM (ManLSTM) model; 3) question selection and answer extraction.

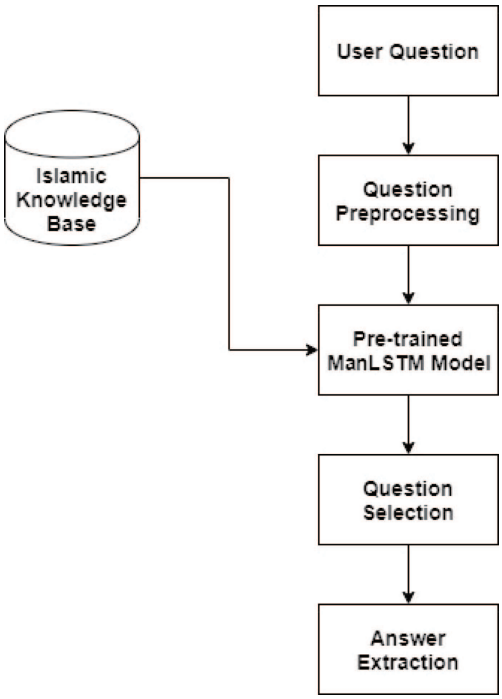

FIGURE 1. Proposed QA system architecture

2.1. **Question processing.** The first step in question preprocessing module is to transform user questions and knowledge base into queries. This module encompasses operations such as normalization, tokenization, and removing punctuation. A set of queries is defined as $q = t_1, t_2, \ldots, t_i$ where the query q has i number of separated terms t. The next step is to build word embedding from the set of queries. Word embeddings are representations of words in lower-dimensional space of dense vector. It can capture the context of a word in a document, semantic-syntactic similarity, and word relationship. In this step, we map each term from the queries into a fix-sized vector using pre-trained Word2Vec [15].

2.2. **Pre-trained ManLSTM model.** One of the essential tasks for language understanding is modeling the underlying similarity between words, phrases, or sentences. However, a problem remains hard because having labeled data is scarce and understanding complex data structure is complicated. Traditionally, Term Frequency-Inverse Document

Frequency (TF-IDF) model was ruled over several years in NLP but was limited to understanding the context by its inherent term-specificity. In the Manhattan LSTM (ManLSTM) model proposed by Mueller and Thyagarajan [13] shown in Figure 2, there is a twin Long Short-Term Memory (LSTM) network which was trained on paired sentences to learn contextual meaning from sentence representations. LSTM is a variation of Recurrent Neural Networks (RNNs) used in deep learning [16]. LSTM is suitable to model sequential data, learn long-term dependencies and prevent the vanishing gradient problem. LSTM maintains its state over time using memory and gates to regulate the information flow. Given input vector x_t, hidden state h_t, and memory state c_t, LSTM performed weight updates as follows:

$$i_t = sigmoid\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{1}$$

$$f_t = sigmoid\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{2}$$

$$\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \tag{3}$$

$$c_t = i_t \odot \tilde{c}_t + f_t \odot c_{t-1} \tag{4}$$

$$o_t = sigmoid\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

where i_t, f_t, o_t are input, forget, and output gates at time t, respectively. W_i, W_f, W_c, W_o, U_i, U_f, U_c, and U_o are LSTM parameterized weight matrices. The bias vectors b_i, b_f, b_c, b_o and $\odot$ denote the Hadamard product, an entry wise multiplication.

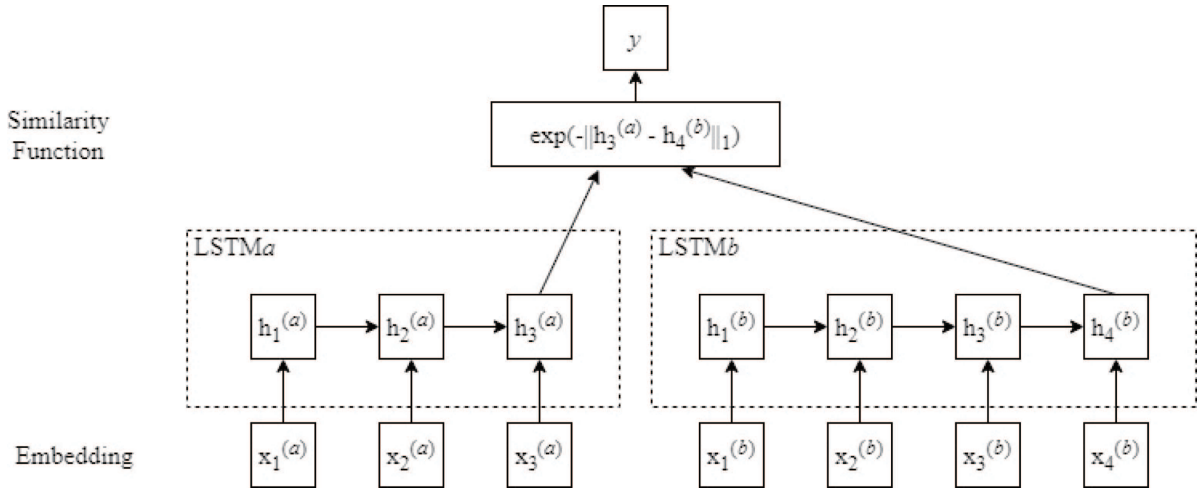

FIGURE 2. Manhattan Siamese LSTM model

The inputs x_1 and x_2 are two fixed-length vectors representing a pair of questions fed into the embedding layer. The embedding layer will look up the embedding for each word and encapsulate them into a vector. This vector represents the given questions as a set of embedding, and the hidden state of the final hidden layer is a 50-dimensional vector for each question. Both hidden states are then compared to compute a similarity score in the output layer as depicted in Figure 2. Similarities in the question representations are employed to infer the similarity of context or semantic between pair of questions. The similarity function used is Manhattan function y defined as

$$y = \exp(-||h_1 - h_2||_1) \in [0, 1] \tag{7}$$

The vector from the final hidden states of two LSTM networks denoted as h_1, h_2, and the model output is a similarity score between 0 and 1. In this study, we also build a Manhattan Bidirectional LSTM (ManBiLSTM) model to compare its performance with previous ManLSTM. In ManBiLSTM shown in Figure 3, the input x_1, x_2 is fed into

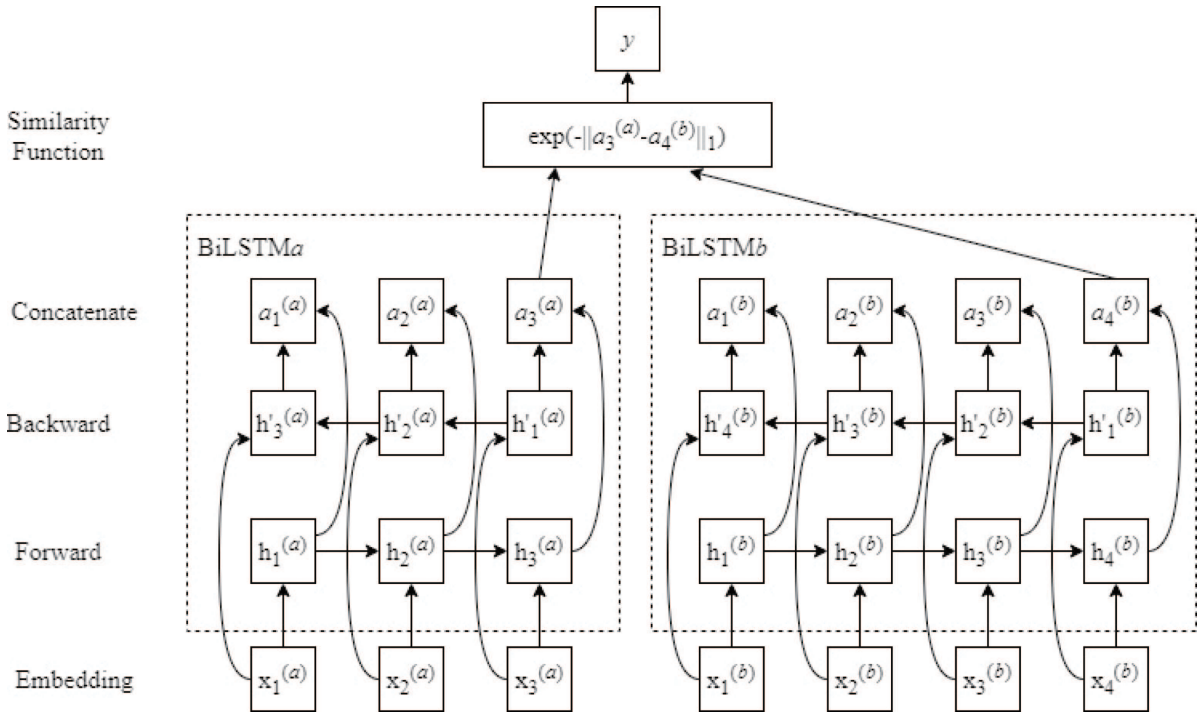

FIGURE 3. Manhattan Siamese BiLSTM model

Siamese LSTM networks once from start to the end (forward) and once from end to start (backward).

2.3. **Question selection and answer extraction.** The question selection module's main target is to provide a list of questions relevant to the user question. This module includes two operations: filtering relevant questions and ranking them to determine the most related question matching the user question. The final step in the answer extraction is to select the best relevant question according to their scores and then extract its answer from corresponding questions.

3. **Results and Discussions.** We used an English Islamic knowledge base from the Quranic Arabic-English Question and Answer Corpus (QAEQ&AC), which contains 590 pairs of question-answer for English corpus [1]. To train the Siamese LSTM model, we build the dataset from questions on Islamic knowledge by pairing each question within datasets. The relevant questions were created by paraphrasing the semantically equivalent to the original question. The collection contains 1213 question pairs, where each pair has an annotated similarity target (1: similar, 0: dissimilar).

3.1. **Model evaluations.** In the process building input model, there are 110 word-vectors not found in the training set from pre-trained Word2Vec. These words were called unknown words or Out-of-Vocabulary (OOV). Most of the unknown words were Islamic terms, while typos, misspelling, and number caused the rest. The subword embeddings are useful to handle OOV words [17]. Therefore, we built question representation using subword embedding called FastText, which reduced 36% OOV in the training set. The comparison of total OOV for Word2Vec and FastText is shown in Table 1 below.

TABLE 1. Question representations

Embeddings	OOV	Islamic terms	Typos	Other
Word2Vec	110	72	18	20
FastText	70	56	13	1

To understand how the Siamese LSTM model learns the contextual meaning, we compared the different structures of neural networks models, namely ManLSTM and ManBiLSTM. We also compared the use of word embedding in the input model.

The results in Table 2 show that ManLSTM performed better than ManBiLSTM in P@N and recall. Meanwhile, in MAP, BiLSTM gives a better result. According to MAP, around 87% of relevant questions from BiLSTM have a good ranking, slightly higher than the questions retrieved by the ManLSTM model. Therefore, in our study, the backward process improves the model performance to learn and predict the similarity score between user questions and relevant questions. We also found that word embedding enabled the model to capture the contextual meaning and semantics better. Both models achieved around 25% higher number of good ranking relevant questions when using word embedding in the input layer.

Table 2. Siamese LSTM model evaluation

Models	P@5	P@10	Recall	MAP
ManLSTM	0.2426	0.1393	0.6557	0.5889
ManBiLSTM	0.2393	0.1410	0.6721	0.6126
ManLSTM + Word2Vec	0.4820	0.2820	0.9508	0.8463
ManBiLSTM + Word2Vec	0.4689	0.2656	0.9344	0.8729

We tried to compare different similarity measurements during the training of Siamese LSTM model. The Manhattan and Cosine outperform the Euclidean when used as a similarity function as shown in Table 3, demonstrating that it is better to catch the questions' semantics and contextual meaning. These results are the same as conclusions found by Mueller and Thyagarajan [13]. Considering that the context meaning is defined as its point directions, questions with similar meanings will have a similarity score near 1. The Manhattan function summits the absolute differences of two vectors while Cosine computes the dot product and each vector's magnitude. Both functions are only concerned about the relevant words of the two vectors.

Table 3. Comparison similarity measurements

Functions	P@5	P@10	Recall
Manhattan	0.4820	0.2820	0.9508
Cosine	0.4885	0.2770	0.9508
Euclidean	0.4689	0.2639	0.8853

3.2. **Proposed QA evaluation.** In the question retrieval evaluation of the proposed QA system, we tested our proposed system against the QAEQAS by Hamoud and Atwell [1]. We used a set of questions which was previously used to test QAEQAS in the evaluation phase. We used precision and recall as the metrics for performance evaluation. Precision returns the proportion of retrieved questions relevant to the user question. Meanwhile, recall returns the proportion of relevant questions over total questions asked by the user.

Table 4 shows that our proposed model outperforms QAEQAS in English data test by returning more relevant questions than QAEQAS. The proposed model achieved a better result in retrieving relevant questions with similar contexts, such as *What Muslims think about purgatory?* and *Do Muslims believe in Purgatory?* which word think has the same context with belief. Interestingly, our proposed model also worked better on short questions such as *Wtat is Satan?* and successfully returned *Who is Satan?*, displaying how the ManLSTM model produced the question representations and learned the contextual meaning between pairs of questions. The proposed model processed questions with typos

TABLE 4. Performance evaluation on proposed model and base-line QA

QA systems	Precision	Recall
QAEQAS	75%	73%
Proposed QA	90%	87%

TABLE 5. Examples of irrelevant questions returned by ManLSTM model

User questions	Irrelevant questions	Cause
Are there any festivals in Islam? What are they?	Are there any other sacred sources? What are they?	Different context
Do you believe that prophet Muhammed is the last messenger?	What is the name of the last messenger?	No relevant question
To whom did God send Muhammad?	Why the messengers were sent by Allah?	Different context
What are the believes about angles in Islam?	What is the meaning of Islam?	Misspelling word
What do you think about Jesus is he a God?	Whatever of blessings and good things you have, is it from Allah?	Different context

and returned its relevant questions, such as *Wtat* instead of *What*. Besides, the problem of similar ranking in pattern matching approach is reduced by understanding contextual meaning in user questions. Table 5 shows some examples of questions that were incorrectly answered and the reason behind that.

4. **Conclusions.** The question retrieval task in QA systems is to retrieve the most relevant questions given on user questions from the knowledge base. The simple question retrieval approach uses a pattern match to measure question pairs similarity using matching terms. This approach failed to retrieve questions that have a similar context or semantic. Therefore, we proposed to use the word embedding and Siamese LSTM model to predict the similarity between the user questions and the knowledge base. Our approach showed a promising result by returning a good ranking relevant questions, 0.4820 on P@5, 0.9508 on Recall, and 0.8463 on the MAP. Interestingly, we showed that Siamese LSTM could model the contextual meaning between user question and knowledge base. Also, question representation using word embedding enabled the model to learn better on question representation.

The Siamese LSTM uses memory cell units to store information across long input sequences to learn dependencies [16]. However, LSTM might fail to compress all necessary information into its representation for long sequences. Therefore, we can extend the LSTM with an attention mechanism to let the model give extra attention when attending all past outputs. In this study, the Siamese LSTM was trained using a specific domain dataset. Hence, it is possible to transfer learning the model using public datasets to increase model knowledge in predicting questions similarities. We also have OOV in our training data which affects the model's performance in predicting questions' similarity. Therefore, we suggest training or transfer learning the word embedding to accommodate the Islamic terms, which are not found in published pre-trained word embeddings.

Acknowledgment. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] B. Hamoud and E. Atwell, Using an Islamic question and answer knowledge base to answer questions about the Holy Quran, *International Journal on Islamic Applications in Computer Science and Technology*, pp.20-29, 2016.

[2] H. Abdelnasser, R. Mohamed, M. Ragab, A. Mohamed, B. Farouk, N. El-Makky and M. Torki, Al-Bayan: An Arabic question answering system for the Holy Quran, *Proc. of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, Doja, Qatar, 2014.

[3] M. A. H. Adany and E. Atwell, Quran question answering system using Arabic number patterns (singular, dual, plural), *International Journal on Islamic Applications in Computer Science and Technology*, vol.5, no.2, pp.1-12, 2017.

[4] N. Puteh, M. ZabidinHusin, H. M. Tahir and A. Hussain, Building a question classification model for a Malay question answering system, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol.8, 2019.

[5] A. Abdi, S. Hasan, M. Arshib, S. M. Shamsuddina and N. Idris, A question answering system in Hadith using linguistic knowledge, *Computer Speech & Language*, vol.60, 2020.

[6] L. Kodra and E. K. Meçe, Question answering systems: A review on present developments, challenges and trends, *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol.8, no.9, pp.217-224, 2017.

[7] N. Othman, R. Faiz and K. Smaili, Manhattan Siamese LSTM for question retrieval in community question answering, *The 18th International Conference on Ontologies, DataBases, and Applications of Semantics*, Rhodès, Greece, 2019.

[8] Z. Chen, C. Zhang, Z. Zhao, C. Yao and D. Cai, Question retrieval for community-based question answering via heterogeneous social influential network, *Neurocomputing*, vol.285, pp.117-124, 2018.

[9] C. Zavou, *Question Retrieval in Community Question Answering Enhanced by Tags Information in a Deep Neural Network Framework*, Master Thesis, University of Amsterdam, 2017.

[10] M. Farouk, Measuring sentences similarity: A survey, *Indian Journal of Science and Technology*, vol.12, 2019.

[11] J. W. G. Putra and T. Tokunaga, Evaluating text coherence based on semantic similarity graph, *Proc. of the Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-11)*, Vancouver, Canada, 2017.

[12] J. P. T. Yusiong and P. C. Naval, Jr., Unsupervised monocular depth estimation of driving scenes using Siamese convolutional LSTM networks, *International Journal of Innovative Computing, Information and Control*, vol.16, no.1, pp.91-106, 2020.

[13] J. Mueller and A. Thyagarajan, Siamese recurrent architectures for learning sentence similarity, *Proc. of the 30th AAAI Conference on Artificial Intelligence*, 2016.

[14] W. Bao, W. Bao, J. Du, Y. Yang and X. Zhao, Attentive Siamese LSTM network for semantic textual similarity measure, *International Conference on Asian Language Processing (IALP)*, Bandung, Indonesia, 2018.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, *NIPS*, pp.3111-3119, 2013.

[16] S. Hochreiter, Long short-term memory, *Neural Computation*, vol.9, no.8, pp.1735-1780, 1997.

[17] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, vol.5, pp.135-146, 2017.

[18] N. Indurkhya and F. J. Damerau, *Handbook of Natural Language Processing*, Chapman & Hall/CRC, 2010.