Long Short-Term Memory for Non-Factoid Answer Selection in Indonesian Question Answering System for Health Information

Retno Kusumaningrum*, Alfi F. Hanifah, Khadijah Khadijah, Sukmawati N. Endah, Priyo S. Sasongko
Department of Informatics, Universitas Diponegoro, Semarang, Indonesia

Abstract—Providing reliable health information to a community can help raise awareness of the dangers of diseases, their causes, methods of prevention, and treatment. Indonesians are facing various health problems partly due to the lack of health information; hence, the community needs media that can effectively provide reliable health information, namely a question answering (QA) system. The frequently asked questions are non-factoid questions. The development of answer selection based on the classical approach requires distinctive engineering features, linguistic tools, or external resources. It can be solved using deep learning approach such as Convolutional Neural Networks (CNN). However, this model cannot capture the sequence of words in both questions and answers. Therefore, this study aims to implement a long short-term memory (LSTM) model to effectively exploit long-range sequential context information for an answer selection task. In addition, this study analyses various hyper-parameters of Word2Vec and LSTM, such as the dimension, context window, dropout, hidden unit, learning rate, and margin; the corresponding values that yield the best mean reciprocal rank (MRR) and mean average precision (MAP) are found to be 300, 15, 0.25, 100, 0.01, and 0.1, respectively. The best model yields MAP and MRR values of 82.05% and 91.58%, respectively. These results experienced an increase in MAP and MRR of 18.68% and 46.11%, respectively, compared to CNN as the baseline model.

Keywords—Answer selection; health information; long short-term memory; LSTM; question answering

I. INTRODUCTION

Providing reliable health information to a community can help raise awareness of the dangers of diseases, their causes, methods of prevention, and treatment. Indonesians are facing various health problems partly due to the lack of health information, including the dangers of smoking, nutritional problems (stunting and obesity), and serious diseases such as heart disease, cancer, and diabetes. Therefore, the community requires media that can provide health information appropriately, namely a question answering (QA) system.

The QA system is a natural language processing (NLP) application that provides specific answers to the questions/queries posed by the user. The QA system is different from a search engine in that the latter will return a set of documents that may contain answers, and users are required to read the documents and search for the exact answers or infer from the set of documents presented. Therefore, the process of finding answers in a QA system is more complex than the process of finding documents presented by a search engine.

Various QA systems have been developed for both the non-Indonesian QA system and the Indonesian QA system. The following QA systems have been developed for non-Indonesian documents: the English QA system [1]–[3], Chinese QA system [4], Spanish QA system [5]–[7], and French QA system [8], [9]. The Indonesian QA system includes QA statistical and linguistic knowledge systems [10], syntactic-semantic processing QA systems [11], [12], QA systems based on machine learning cross-language QA systems [13], pattern matching QA-based systems [14], and pipeline-based cross-language QA systems [15]. In addition, the Indonesian language QA system has been developed for closed-domain QA [16]–[18].

QA systems are differentiated on the basis of the type of questions handled, which are divided into five categories: factoid, non-factoid, yes-no, list, and opinion [19]. Factoid questions have answers in the form of date, quantity, location, person, organisation, and name (in the form of nouns) in addition to the location, person, and organisation categories [13]. Non-factoid questions are those whose answers are generally used to understand something. Non-factoid questions have six categories: question definitions, reasons, methods, degrees, changes, and details [20]. Overall, the Indonesian QA system is still limited to factoid questions, with hardly any non-factoid questions. Related to health information, the types of questions that are commonly encountered are non-factoid questions.

Several studies have been conducted on non-factoid Indonesian QA systems but for non-health data domains. Moreover, these studies generally used a classical approach such as pattern matching and semantic analysis [21], case-based reasoning [16], and similarity score technique [19]. They provide a good performance only when all the patterns of the answer pairs have been defined, making it appropriate only for certain knowledge domains. In addition, the studies were generally implemented for non-factoid questions related to definitions, reasons, and method categories.

Now-a-days, deep learning models have been widely developed for solving several problems using various types of datasets, such as those containing images, signals, and text. Some examples of deep learning implementation using textual data include sentiment analysis [22], [23], machine translation [24], [25], summarisation [26], [27], and QA. A deep learning...
model can be implemented in a QA system as a model for selecting the exact answer from a set of candidate answers, also known as the answer pool. The deep learning model does not require feature engineering, linguistic tools, or external resources \cite{28}. Feature engineering is the stage wherein representative features, such as term frequency–inverse document frequency (TF-IDF) and bag-of-words, are determined. The linguistic tools are linguistic rules and syntax. The implementation of deep learning in a QA system requires a convolutional neural network (CNN). However, this model cannot capture the sequence of words in both questions and answers. This can be overcome by implementing long short-term memory (LSTM).

Therefore, this study aims to implement an LSTM as a model for selecting non-factoid answers in the Indonesian question answering system (IQAS) for Health Information. As mentioned earlier, the LSTM model has never been implemented for answer selection in the IQAS, neither in a specific data domain nor in the general data domain. Hence, the first step in this approach is to train the word2vec model on a health information corpus obtained from various popular health websites written in the Indonesian language. In addition, this study empirically analyses the effect of Word2Vec hyper-parameters, such as the context window size, on the performance of the LSTM model in selecting the right answer to a question. Furthermore, the effect of varying the LSTM hyper-parameters on the performance of the LSTM model as a model for selecting exact answers from an answer candidate pool was studied; thus, we established the best answer selection model.

The contributions of this paper are summarised as follows:

- A pre-trained Word2Vec model for the Indonesian language, specifically on health information.
- An investigation related to the influences of the dimensions and context window size of Word2Vec on the performance of the LSTM model in selecting answers.
- An analysis of the influences of the hyper-parameters on the LSTM model, including the dropout, number of hidden units, learning rate, and margin size, on the performance of the LSTM model for answer selection.
- A pre-trained LSTM model for non-factoid answer selection in the IQAS for health information. Subsequently, it was implemented as a web-based application.

The rest of this paper is organised as follows. Section II describes related work, including a general description of the answer selection task and LSTM in detail. A detailed explanation of the proposed framework is presented in Section III, including descriptions of data collection, training process of Word2Vec, generation of the answer selection model based on the LSTM, and model evaluation. Section IV presents the experimental results. Finally, in Section V, we draw some conclusions from the results.

II. RELATED WORKS

A. Answer Selection Task

Answer selection is a subtask of the QA system that performs the process of selecting sentences containing the required information from a set of candidate answers \cite{29}. Answer selection involves not only matching the terms in the question and answer but also finding the same semantic meaning from both the question and answer. Formally, the answer selection problem can be described as follows:

- There is a question \( q \) and an answer candidate pool \( \{a_1, a_2, ..., a_k\} \) that contains a set of answer candidates for a particular question.
- The aim of answer selection is to select the best answer candidates from the answer candidate pool.

Therefore, the answer selection task can be formulated as a ranking problem, giving better ranks to answers that are more relevant to the respective question. Some of the ranking function approaches include pointwise, pairwise, and list wise \cite{30}. This study implements a pairwise approach to train the ranking function to give higher scores for correct answers and lower scores for wrong ones.

B. Long Short-Term Memory

The LSTM model is a popular variation of the recurrent neural network (RNN) method. The RNN method is widely used to solve data problems whose order requires attention. The LSTM model overcomes the gradient vanishing problem of the RNN method. In addition, LSTM model is more capable of dealing with the context of long and sequential information. The LSTM model used in this study is the one introduced in \cite{31}.

The LSTM model is designed to solve the gradient vanishing problem using a gate mechanism. Its architecture has three gates, namely an input gate \( i_t \), a forget gate \( f_t \), and an output gate \( o_t \), and a memory cell \( c_t \). The LSTM can add or reduce information into the cell state, which is regulated by the gate. The input gate is responsible for determining new information to be added to the memory cell. The forget gate determines which information will be saved or deleted. Finally, the output gate is responsible for determining the information that will be used as output. Fig. 1 shows the LSTM cells.

The hidden state \( h_t \) is calculated on the basis of the three LSTM gates. The size of the hidden state is determined by a parameter called the hidden unit. The hidden unit is a parameter in the LSTM that shows the vector dimension of the hidden state \( h_t \) for each time step. Mathematically, the LSTM model is defined as follows:

\[
\begin{align*}
    i_t &= \sigma(W_i x(t) + U_i h(t-1) + b_i) \\
    f_t &= \sigma(W_f x(t) + U_f h(t-1) + b_f) \\
    o_t &= \sigma(W_o x(t) + U_o h(t-1) + b_o)
\end{align*}
\]
The LSTM architecture has three gates (input, forget, and output) and a cell memory vector. $\tilde{C}_t$ is the sigmoid function. $i_t$, $f_t$, and $o_t$ are the network parameters. The LSTM architecture has three gates (input, forget, and output) and a cell memory vector.

III. METHODOLOGY

This section describes the proposed framework used in this study, comprising four main processes. Fig. 2 shows its general description.

The research framework comprises four main processes: data collection, training process of Word2Vec, generating an answer selection model based on the LSTM, and model evaluation. The detailed explanations for each process are given in the following subsections.

A. Data Collection

In this process, two types of datasets are formed: a QA dataset (pair of question-and-answer datasets) and a health article dataset. The QA dataset was created by collecting question and answer pairs from popular health sites in Indonesia, namely hellosehat.com, alodokter.com, and halodoc.com. Non-factoid questions on topics of diseases and medicines are used as questions. The categories of the questions are definitions, reasons, and methods. In total, 750 pairs of questions and answers are formed, consisting of 355 pairs for definitions, 145 pairs for reasons, and 250 pairs for methods. The article dataset is established using all the articles from the three websites through data scraping.

B. Training Process of Word2Vec Model

The Word2Vec model is a word embedding algorithm proposed in [32] to learn vector representations. Vector representations can efficiently capture the semantic meaning of the words represented. The word vector tends to obey the laws of analogy and describe intuition. Words known as synonyms have the same vector in the cosine equation, whereas antonyms have different vectors. Therefore, the representation of words in the vector space is useful for achieving better performance on NLP problems by grouping similar words.

The dataset used in Word2Vec training is the article dataset. The article dataset contains articles on diseases and medicines found on the three sites previously described. The number of vocabularies formed was 44,700. The Word2Vec model used is skip-gram, and the evaluation method is hierarchical softmax. Fig. 3 illustrates the skip-gram architecture.

Figure 1: LSTM cell

\[
\tilde{C}_t = \tanh(W_c x(t) + U_c h(t-1) + b_c) \\
C_t = i_t \cdot \tilde{C}_t + f_t \cdot C_{t-1} \\
h_t = o_t \cdot \tanh(C_t)
\]

Fig. 1. LSTM cell

Fig. 2. Framework of this study comprising four main processes: data collection, word2vec training, LSTM-based answer selection modelling, and model evaluation

Fig. 3. Illustration of skip-gram architecture of Word2Vec model
C. Generating Answer Selection Model based on LSTM

Modelling for answer selection uses a Siamese architecture. This type of architecture can be used to measure the relevance of candidate answers to a question. Fig. 4 shows the Siamese architecture of the LSTM-based answer selection model. In the embedding layer, the inputted sentences (i.e., the candidate answer and the question) are converted into vector representations generated by Word2Vec training. Thereafter, in the encoding layer, the same encoder is used to create distributed vector representations for the input sentences separately. The encoding layer adopts the QA-LSTM using a bidirectional LSTM (biLSTM) model. During the encoding process, the questions and answers do not have explicit interactions.

Bidirectional LSTM utilises both the previous and future contexts by processing in two directions and generates two independent sequences of LSTM output vectors. The two output vectors are concatenated as follows:

The implementation of max pooling was used to generate representations for the questions and answers based on the word-level biLSTM outputs. The relevance scores of the candidate answers to a question are obtained based on pooled vectors. Subsequently, using the cosine similarity measures the distance between the candidate’s answer and the question.

D. Model Evaluation

The evaluation techniques used are the mean reciprocal rank (MRR) and mean average precision (MAP), which are the standard metrics for information retrieval and QA. The MRR can be calculated as follows:

\[
MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{7}
\]

The MAP can be calculated as follows:

\[
MAP = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{precision}(R_{jk}) \tag{8}
\]

IV. RESULTS AND DISCUSSION

A. Experimental Setup

The data used in this research are in the form of 750 question–answer pairs. There are 1564 unique answers collected in the answer space. With regard to the distribution ratio of the training and test data, 70% is for training and 30% is for testing. Following the data distribution, we have 525 pairs as training data and 225 pairs as test data. The pool size is 50. It was generated by sending the ground-truth answers to the pool and randomly sampling negative answers from the answer space until the pool size reached 50.

The experiment employs several hyperparameters of the Word2Vec model and LSTM. Each model is trained for 100 epochs. The Word2Vec hyperparameters are dimension (100, 200, and 300) and context window (5, 10, and 15). At the same time, the LSTM hyperparameters are dropout (0.25, 0.5, and 0.75), number of hidden units (50, 75, and 100), learning rate (0.00001, 0.0001, 0.001, and 0.01), and margin (0.05, 0.1, and 0.15).

B. Experimental Scenarios

Several scenarios are established to determine the impacts of the various parameters tested on the performance of the proposed model; scenarios 1, 2, 3, 4, 5, and 6 are for the Word2Vec dimension, context window, dropout, hidden unit, learning rate, and margin, respectively. Fig. 5 shows the overview of these scenarios.

C. Experimental Results and Analysis

Scenario 1 is aimed at studying the impact of Word2Vec dimensions on the MRR and MAP results. Table I shows that the model yields the best averages of MRR (78.75%) and MAP (63.70%) when the Word2Vec dimension is 300. The MRR and MAP values are directly proportional to the dimensions of Word2Vec; therefore, the higher the dimensions of Word2Vec, the higher the MRR and MAP...
values. The Word2Vec dimension represents the size of the learned word vector, or it can be referred to as the features of each word. A higher dimension tends to capture more information and better word representations.

Scenario 2 is aimed at studying the impacts of context window on the MRR and MAP results. The best averages of the MRR and MAP values are obtained when the context window is 15, as shown in Table II. From the table, it can be concluded that the averages of the MRR and MAP are directly proportional to the context window, which means that, the larger the context window size, the higher the average MRR and MAP values. The size of the context window defines the range of words to be included as the context of a target word. For instance, a window size of 5 takes five words before and after a target word as its context for training. A larger context window is required to answer non-factoid questions on health information because this type of question requires a longer answer. Moreover, answers related to health information typically have a long explanation.

Scenario 3 is aimed at studying the impacts of dropout rate on the MRR and MAP results. The best MRR and MAP values are 81.25% and 66.58% when the dropout value is set to 0.25. From the average MRR and MAP obtained for all the tested dropout values, it can be concluded that the dropout value is inversely proportional to the average MRR and MAP, which means that, the lower the dropout value, the higher the MRR and MAP. Dropout refers to ignoring units (i.e. neurons) during the training phase of a certain set of neurons. A higher dropout value indicates that more neurons are ignored, and this will cause the model to lose its ability to learn. Moreover, the dropout performed on the LSTM model can make the model to be more limited in keeping the memory. Therefore, lower dropouts are considered better for storing memory in the LSTM model. Table II lists the results of scenario 3.

TABLE I. PERFORMANCE COMPARISON WHEN VARYING THE WORD2VEC DIMENSIONS

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Average of MAP (%)</th>
<th>Average of MRR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>56.76</td>
<td>72.91</td>
</tr>
<tr>
<td>200</td>
<td>61.77</td>
<td>77.23</td>
</tr>
<tr>
<td>300</td>
<td>63.70</td>
<td>78.75</td>
</tr>
</tbody>
</table>

TABLE II. PERFORMANCE COMPARISON WHEN VARYING THE CONTEXT WINDOWS

<table>
<thead>
<tr>
<th>Context Window</th>
<th>Average of MAP (%)</th>
<th>Average of MRR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>58.24</td>
<td>74.30</td>
</tr>
<tr>
<td>10</td>
<td>61.15</td>
<td>76.63</td>
</tr>
<tr>
<td>15</td>
<td>62.84</td>
<td>77.96</td>
</tr>
</tbody>
</table>

TABLE III. PERFORMANCE COMPARISON WHEN VARYING THE DROPOUT RATES

<table>
<thead>
<tr>
<th>Dropout Rate</th>
<th>Average of MAP (%)</th>
<th>Average of MRR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>66.58</td>
<td>81.25</td>
</tr>
<tr>
<td>0.5</td>
<td>62.14</td>
<td>77.50</td>
</tr>
<tr>
<td>0.75</td>
<td>53.51</td>
<td>70.14</td>
</tr>
</tbody>
</table>

Scenario 4 is aimed to study the impacts of hidden units on the MRR and MAP results. As mentioned before, this study applies different numbers of hidden units: 50, 75, and 100. From Table IV, it can be concluded that the number of hidden units is directly proportional to the average MRR and MAP. The output dimension determines the number of dimensions for each word in the input sequence. Dimension implies the number of features to be remembered. The best averages of MRR and MAP are obtained under a hidden unit value of 100. This is because using more features provides a better representation than using fewer features.

Scenario 5 is aimed at studying the impacts of the learning rate on the MRR and MAP results. Several learning rates were set: 0.01, 0.001, 0.0001, and 0.00001. Based on Table V, it can be concluded that the learning rate is directly proportional to the averages of MRR and MAP. The best averages of MRR and MAP are obtained under a learning rate of 0.01. As explained in the experimental results section, all the models are trained for 100 epochs. The learning rate is a hyperparameter that helps control the degree of model change. A low learning rate may result in a long training process that could get stuck, making it difficult to converge. These results can be obtained because the epoch used tends to be small; therefore, a high learning rate will decrease the MRR and MAP values.

Scenario 6 is aimed at studying the impact of margin on the MRR and MAP results. As previously explained, there are three different margin values: 0.05, 0.1, and 0.15. The highest average MRR and MAP were obtained under a margin of 0.1, as listed in Table VI. No specific pattern is generated between the margins with the average MRR and MAP. Margin is a variable in the hinge loss function. The hinge loss function is an employed loss function that was minimised in this research. If the ground-truth answer has a score higher than the negative answer by at least a margin, the expression has a zero loss. Condition here implies margins as the optimum distance that can be produced between the ground-truth answer and negative answers. If the margin value is too low, the ground-truth answer and the negative answer will not be separated appropriately. The lower the margin, the smaller the distance between the ground-truth and negative answers. This condition can make relevant answers irretrievable. Meanwhile, if the margin is too high, the distance between the correct answer and the wrong answer will be even greater. This makes irrelevant answers be incorrectly taken as correct answers.
TABLE V. PERFORMANCE COMPARISON WHEN VARYING THE LEARNING RATES

<table>
<thead>
<tr>
<th>Learning Rates</th>
<th>Average of MAP (%)</th>
<th>Average of MRR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00001</td>
<td>50.74</td>
<td>68.41</td>
</tr>
<tr>
<td>0.0001</td>
<td>54.14</td>
<td>70.95</td>
</tr>
<tr>
<td>0.001</td>
<td>62.56</td>
<td>78.10</td>
</tr>
<tr>
<td>0.01</td>
<td>75.53</td>
<td>87.72</td>
</tr>
</tbody>
</table>

TABLE VI. PERFORMANCE COMPARISON WHEN VARYING THE MARGINS

<table>
<thead>
<tr>
<th>Margins</th>
<th>Average of MAP (%)</th>
<th>Average of MRR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>60.80</td>
<td>76.39</td>
</tr>
<tr>
<td>0.1</td>
<td>61.02</td>
<td>76.55</td>
</tr>
<tr>
<td>0.15</td>
<td>60.41</td>
<td>75.94</td>
</tr>
</tbody>
</table>

Based on the results of scenarios 1 to 6, the best answer selection model is obtained when using the following hyperparameters: word2vec dimension is 300, context window size is 15, dropout rate value is 0.25, number of hidden units is 100, learning rate is 0.01, and margin value is 0.1. This model yields MAP and MRR values of 82.05% and 91.58%, respectively.

Compared with previous research, this study also ran experiments using CNN with an architecture consisting of 4 convolution layers (kernel size in 1, 2, 3, and 5) and one pooling layer. The word2vec dimension used in the test uses the same dimension, namely 300. The best parameter results for the CNN model include margin 0.15, hidden unit 100, dropout 0.25, learning rate 0.01, and context window 15. The MAP and MRR values obtained are 63.37% and 45.47%, respectively. An illustration of the difference between the CNN model and the proposed model can be seen in Fig. 6. It can be seen that the increases in MAP and MRR were 18.68% and 46.11%, respectively.

Subsequently, the best model is implemented for the QA application, which is given the name MediQA. Fig. 7 shows the sample result of the answer selection.

Fig. 7. Siamese architecture of LSTM-based answer selection

As mentioned in the previous section, this study evaluates three questions: definitions, reasons, and methods. Fig. 8 shows a sample of the correct and incorrect answer results given by the MediQA application for the definition question type. Fig. 9 shows the same for the method question type. Both figures consist of two parts, the first part shows a result example of choosing the incorrect answer by the system, and the second part shows a result example of choosing the correct answer by the system. In the answer pool section, sentences in green indicate sentences that should have been selected as the correct answer. Meanwhile, sentences written in red are incorrect answer sentences and are output as answers by the system.
results—first, the larger the dimension of Word2Vec, the better the MRR and MAP values. A dimension of 300 yielded the best MRR and MAP. Second, a context window size of 15 yielded the best MRR and MAP results, indicating that a more extensive context window can yield better MRR and MAP results. Third, a lower dropout value yielded better MRR and MAP values, and the best MRR and MAP were achieved under a dropout value of 0.25. Fourth, the optimal hidden unit value was found to be 100; the higher the number of hidden units, the better the MRR and MAP values. Fifth, a higher learning rate showed significant improvements in the MRR and MAP, given the relatively small number of datasets used in this research. Sixth, a margin of 0.1 produced the best MRR and MAP results. The best model yielded MRR and MAP values of 82.05% and 91.58%, respectively. These results experienced an increase in MRR and MAP of 18.68% and 46.11%, respectively, compared to CNN as the baseline model.

This research is still limited to selecting answers on IQAS for a particular domain (health information), while the need for open-domain QA in Indonesian is still very open. On the other hand, the latest language modelling developments, such as Bidirectional Encoder Representations from Transformers (BERT), have also been developed for Indonesian, commonly known as IndoBERT. Therefore, it provides an opportunity for further research to apply IndoBERT and LSTM as a model for selecting answers in the Indonesian language open-domain QA.

ACKNOWLEDGMENT

This work was supported in part by the Faculty of Science and Mathematics, Universitas Diponegoro under Grant 4866/UN7.5.8/PP/2019.

REFERENCES


