Enhancing Lexical-Based Approach With External Knowledge for Vietnamese Multiple-Choice Machine Reading Comprehension

KIET VAN NGUYEN, KHIEM VINH TRAN, SON T. LUU, ANH GIA-TUAN NGUYEN, AND NGAN LUU-THUY NGUYEN
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
Corresponding author: Ngan Luu-Thuy Nguyen (ngannlt@uit.edu.vn)

ABSTRACT Although Vietnamese is the 17th most popular native-speaker language in the world, there are not many research studies on Vietnamese machine reading comprehension (MRC), the task of understanding a text and answering questions about it. One of the reasons is because of the lack of high-quality benchmark datasets for this task. In this work, we construct a dataset which consists of 2,783 pairs of multiple-choice questions and answers based on 417 Vietnamese texts which are commonly used for teaching reading comprehension for elementary school pupils. In addition, we propose a lexical-based MRC method that utilizes semantic similarity measures and external knowledge sources to analyze questions and extract answers from the given text. We compare the performance of the proposed model with several baseline lexical-based and neural network-based models. Our proposed method achieves 61.81% by accuracy, which is 5.51% higher than the best baseline model. We also measure human performance on our dataset and find that there is a big gap between machine-model and human performances. This indicates that significant progress can be made on this task. The dataset is freely available on our website for research purposes.

INDEX TERMS Machine reading comprehension, multiple choice question, lexical-based, external knowledge.

I. INTRODUCTION
A primary goal of computational linguistics or natural language processing is to make computers able to understand natural language texts, as well as human beings, do. One of the standard tests of natural language understanding ability requires computers to read documents and answer any questions related to their contents, resulting in different research problem settings of machine reading comprehension [1]–[5]. MRC can also be the extended task of question answering (QA). There are many studies on QA [6]–[9], which are also the foundation for the development of MRC. Findings of this research field are implemented into various artificial intelligence applications such as next-generation search engines, AI agents, chatbots, and robots.

One common method for evaluating someone’s understanding of texts is by giving them a multiple-choice reading comprehension test. This type of test can measure abilities such as causal or counterfactual reasoning, inference among relations, or basic understanding of the world in a set of reading texts. In the past ten years, there have been many study works [10]–[16] in this field. In addition to researching MRC in each language, one of the current trends in MRC is cross-lingual studies such as [17], [18]. Hence, the first important thing is the contribution of MRC datasets in each language. Besides, there have been research results in lexical-based approaches [10], [12] and machine-learning-based approaches [4], [12], [19]–[22]. Depending on the characteristics and size of datasets, we propose the appropriate methods to achieve better performances.

English and Chinese are regarded as resource-rich languages when it comes to the accessibility of the tools needed to carry out communication. Still, many other languages are deemed resource-poor, and Vietnamese is one of them. Machine reading comprehension for the Vietnamese language is vital as for other languages because it is useful for
Section IV presents our proposed method for Vietnamese multiple-choice machine reading comprehension. Section V shows experiments and results on the dataset. Section VI describes the result analysis for these experimental results. Finally, Section VII concludes the paper and discusses future work.

II. LITERATURE REVIEW
In this section, we aim to review recent datasets and techniques in machine reading comprehension. In particular, the typical MRC datasets and methods are described as follows.

A. MRC DATASETS
In the last decade, we have witnessed a fast growth of research interest in machine reading comprehension (MRC) and an explosion of datasets for MRC studies for popular languages like English [1], [2], [28]–[35] and Chinese [36]–[38].

In terms of types of answers, MRC datasets are divided into three categories, including extractive, abstractive, and multiple-choice.

- **Extractive MRC** requires computers to locate the correct segment in a provided reading text that answers a specific question related to that text. Recently, there has been a significant increase in the construction of extractive MRC datasets with formal written texts such as SQuAD [2], CNN/Daily Mail [1], CBT [28], NewsQA [29], TriviaQA [31], WIKIHOP [32], DRCD [37], and CMRC2018 [38].

- In contrast to extractive MRC, **abstractive MRC** requires computers to generate answers or synthetic summaries because answers to such questions in abstractive MRC are usually not spans in the reading text. Datasets for abstractive MRC include MS MARCO [39], SearchQA [40], NarrativeQA [41], and DuReader [36].

- **Multiple-choice MRC** includes both extractive and abstractive MRCs; however, the correct answer options are primarily abstractive. Most of the multiple-choice MRC datasets are created using crowdsourcing methods in major steps of dataset construction including generating questions, correct answer options and distractors. MCTest [10], ROCStories [11], MultiRC [13], MCScript [14], and COSMOS QA [42] are typical datasets of this type. The crowd workers also assign to each question the reasoning mechanism that is needed to figure out the answer. Apart from the basic reasoning mechanism - the matching type, a dramatic number of questions require complex reasoning mechanisms which are based on multiple sentences and require external knowledge. Other datasets are collected from examinations designed by educational experts QALD [43], NTCIR-11 QA-Lab [44], dataset from TOEFL exams [45], dataset from NY Regents 4th Grade
Science exams [46], and RACE [12], which aim to evaluate learners.

Until now, there is not yet any dataset available for Vietnamese machine reading comprehension, which is one of the primary reasons that we would like to collect and build a dataset for the Vietnamese language processing community.

### B. MRC METHODS

In this paper, we focus on two main types of MRC method, lexical-based approaches and neural network-based approaches. Therefore, we review the previous study works in these methodologies as follows.

1) **LEXICAL-BASED APPROACHES**

The first method implemented into multiple-choice reading comprehension is the Sliding Window algorithm, a lexical-based approach developed by [10], as our first baseline model. This method was also used as a baseline in other studies [2], [12], [14]. Sliding Window finds an answer based on simple lexical information. Motivated by TF-IDF, this algorithm uses inverse word count as a weight of each lexical unit, and maximizes the bag-of-word similarity between the answer option and lexical units in the given reading text in a window size.

2) **NEURAL NETWORK-BASED APPROACHES**

With the popularity of the neural network approach, end-to-end models such as Stanford AR [19], GA Reader [20], HAF [21], and Co-Match [47] have produced promising results on multiple-choice MRC. Recently, pre-trained language models have also been added [48], [49]. These models do not rely on complex manually-devised features as in traditional machine learning approaches, but are able to outperform them. In this paper, we employ an end-to-end model called Co-match [47] with different pre-trained word embeddings as another baseline model.

Regarding to the Vietnamese language processing, there are quite a number of research works on other tasks such as parsing [50]–[52], part-of-speech [53], [54], named entity recognition [55]–[57], sentiment analysis [58]–[60], and question answering [61]–[63]. However, to the extent of our knowledge, there are no research publications on multiple-choice machine reading comprehension. Therefore, we decide to build a new dataset of Vietnamese multiple-choice reading comprehension for the research community and evaluate MRC state-of-the-art models on our dataset.

### C. SEMANTIC SIMILARITY MEASUREMENT AND WORD EMBEDDINGS

Recently, the semantic similarity measures between texts have been studied in many natural language processing applications. A range of researchers have used these measures to improve their study works [64]–[66]. These methods proposed for estimating the similarity between two documents include three different types, i.e., lexical matching, linguistic analysis, and semantic features. Lexical matching is not sufficiently strong and linguistic analysis also have limitations. In semantic feature approaches, a word is represented by a vector as semantic meaning before estimating similarity. These study works [67], [68] utilized external knowledge sources to estimate the similarity of two texts. These approaches are only effective when external knowledge sources such as WordNet, word embeddings or other datasets are available for the tested domain or applications.

Word embeddings also play a significant role in machine reading comprehension. Rumelhart *et al.* (1986) [69] proposed word embedding, a technique that maps each word to a vector space and can accurately capture a large proportion of syntactic and semantic relationships in text. Using pre-trained word embedding [70], [71], there are two most common methods to represent words in machine reading comprehension models: word-level embedding and character-level embedding. However, these methods seem to be insufficient because it simply concatenates word-level and character-level embeddings; generated vectors stay the same in different contexts. To tackle these problems, Peters *et al.* (2018) [72] proposed deep contextualized word representations called ELMo which is pre-trained by language model first and fine-tuned according to the learning task. Devlin *et al.* (2018) [49] introduced BERT, which utilizes bidirectional transformer to encode both left and right contexts to the representations. In this article, we take advantage of semantic similarity and word embeddings for enhancing external knowledge to improve the performance of multiple-choice reading comprehension in Vietnamese.

### III. DATASET

#### A. DATASET CREATION

The process of constructing the ViMMRC dataset includes three different phases: reading-text collection, multiple-choice question creation, and dataset validation. These phases are described in detail as follows.

1) **READING-TEXT COLLECTION**

We decide to focus on the reading comprehension levels at primary schools because they only require general knowledge, not too specific knowledge. We collect the Vietnamese reading texts suitable for the 1st to 5th graders from the subject named Vietnamese. In addition, we collect reading comprehension tests from two reliable websites where all reading comprehension tests from 1st to 5th grades are made public for free of charge. As a result, 417 reading texts are gathered.

2) **MULTIPLE-CHOICE QUESTION COLLECTION**

Questions, answer options, and correct answers are created by primary-school teachers. These questions are intended to test the reading comprehension ability of elementary learners. The teachers are asked to create at least five questions per
TABLE 1. Several examples of multiple-choice reading comprehension of Vietnamese texts are taken from our dataset. Each question has four answer options and there is only one correct answer in them. Besides, these Vietnamese examples are translated into English.

| Reading Text | Question | Answer |
|--------------|----------|--------|
| **Vietnamese:** Ngày giữa sông trong, sùng sùng một cây bàng. Mùa đông, cây vươn dài những cánh không khử, trái là. Xuân sang, cánh trên cánh dưới chỉ chỉ những lộc non non non. Hè về, những tàn là xanh um che mát một khoảng sân trường. Thu đến, tung chum quai chua vàng trong kể là. | Cây bàng được trồng ở đâu? (Where is the tropical almond tree planted?) | A. Ngày giữa sông trong, (In the middle of the schoolyard.) |
| **(English translation:** In the middle of the schoolyard stood a towering tropical almond tree. In winter, the tree stretches out its slender, leafless branches. As spring arrives, its branches on the branches below are spangled with young buds. Summer approaches and its green foliage shades the yard. Autumn comes, revealing bunches of gold ripe fruits dangleing in its leaves.) | **Question:** | **Answer:** |
| | A. Trồng ở ngoại đường. (Planted on the street.) | B. Trên ngoài đường. (Planted on the street.) |
| | C. Gần sông. (Near the river.) | D. Dưới mái hiên trường. (Under the porch.) |
| | **Answer:** | A. |

| Vietnamese: | Cạnh và quả. (Bunches, leaves, buds, foliage and fruit.) | D. Quả và thân cây. (Buds, fruit and foliage.) |
| Vietnamese: | Cạnh, lát, lộc, bàn và quả. (Branches, leaves, buds, foliage and fruit.) | C. |
| Vietnamese: | Cạnh, lát, lộc, bàn và quả. (Branches, leaves, buds, foliage and fruit.) | B. Lá và quả. (Leaves and fruit.) |

3) VALIDATION

During this phase, primary-school teachers review the multiple-choice questions, their answer options, and their correct options again to ensure there are no mistakes. Finally, we obtain a highly-qualified dataset for research purposes for the computer multiple-choice reading comprehension mechanism. Table 1 demonstrates some of the examples of Vietnamese multiple-choice MRC questions. In the following section, we analyze the characteristics of the dataset.

B. DATASET ANALYSIS

We randomly divide our dataset into train, development, and test sets of 292 (70%), 42 (10%), and 83 (20%) texts, respectively. The statistics of the training, development, and test sets are summarized in Table 2. In the table, the number of questions, the average words of reading texts, questions, answer options, correct answers, and vocabulary sizes are also listed.

In this section, we present analysis of our dataset from different aspects. Table 3 shows statistics of our dataset with different grades. Vocabulary size, text length, question length, answer option length, and correct answer length are calculated in words. We used the word segmentation pyvi. We found that the number of reading texts for the 1st grade is small, which is obvious because the 1st grade focuses on developing basic language skills rather than reading comprehension skill. It can be inferred that the vocabulary sizes are correlated with the difficulty level of the reading comprehension task.

The types of reasoning required to solve the multiple-choice machine reading comprehension (MMRC) task directly influence the performance of MMRC models. In this paper, we classify the questions in our dataset following the same reasoning types as used in the analysis of the well-known dataset RACE [12]. These types are shown as follows, in ascending order of the difficulty level:

- **Word matching (WM):** Important tokens in the question exactly match tokens in the reading text. Thus, it is easy to use a keyword search algorithm for finding the correct answer of this question based on the reading text.
- **Paraphrasing (PP):** The question is paraphrased from a single sentence in the reading text. In particular, we may use synonymy and world knowledge to create the question.
- **Single-sentence reasoning (SSR):** The answer is inferred from a single sentence in the reading text. Such answers could be created by extracting incomplete information or conceptual overlap.
- **Multi-sentence reasoning (MSR):** The answer is inferred from multiple sentences in the reading text by information synthesis techniques.
- **Ambiguous or insufficient (AoI):** The question has many answers or answers are not found in the reading text.

2Vietnamese word segmentation tool: https://pypi.org/project/pyvi/
We manually annotate all questions in our dataset according to these types. Examples and percentages of these type are listed in Table 8. It can be seen from the table that single-sentence reasoning and ambiguous-or-insufficient make up the lowest proportions in our dataset (7.35% for single-sentence reasoning and 6.12% for ambiguous-or-insufficient). Meanwhile, word matching and multiple-sentence reasoning types account for the largest percentage, at 25.85% and 36.73% respectively. This demonstrates that ViMMRC is a challenging dataset for evaluating reading comprehension models for the Vietnamese language.

C. COMPARISON WITH THE MCTest DATASET

In this section, we compare our dataset with the MCTest dataset. The size of the MCTest dataset is approximately the same as our dataset. Table 4 shows differences between our dataset and the MCTest dataset. As can be seen from the table, although the number of reading texts in our dataset is less than that of the MCTest dataset, the number of questions of our dataset is greater. Besides, the average numbers of words per reading text, per question and per answer in our dataset are also higher than those of the MCTest dataset.

IV. METHODOLOGY

In this section, we introduce our proposed approach for the Vietnamese multiple-choice machine reading comprehension corpus. Because deep learning methods require a large dataset, so we only focus on the development of lexical-based methods on our dataset. Fig. 1 presents our proposed model by integrating semantic similarity and external knowledge sources into the lexical-based approach. This method is briefly described as follows. First of all, we pre-process the texts. Next, we calculate the sliding windows scores, the distance scores, and the external knowledge scores, respectively. Last, we combine those three scores for calculating the final score. The final score is used for predicting the correct answer in our approach. We implement this system through the algorithms described in detail as follows.

A. PRE-PROCESSING TECHNIQUES

Pre-processing techniques play an important role in many applications of NLP. These techniques help to get rid of meaningless and confusing words, so we clean this data by following the steps shown in Algorithm 1 and Algorithm 2. There are many techniques in natural language processing which are implemented in the pre-processing phase. In particular, Algorithm 1 pre-processes for a sentence, applied to sentence processing in the reading text, questions and answer options.

| Algorithm 1 Pre-Processing a Raw Vietnamese Sentence S |
|------------------------------------------------------|
| **Input:** A raw Vietnamese sentence S.              |
| **Output:** A list of Vietnamese words S.           |
| **procedure** Pre-processing a Vietnamese sentence  |
| X = tokenizing S into a list of tokens.             |
| Removing punctuations in X.                         |
| Removing Vietnamese stop words in X.                |
| S' = converting X into a lower-case sentence.       |
| L = segmenting S' into a list of words by the Viet-  |
| name word segmentation.                             |
| **return** L.                                       |

| Algorithm 2 Pre-Processing a Vietnamese Reading Text T |
|-------------------------------------------------------|
| **Input:** A Vietnamese reading text T.               |
| **Output:** A pre-processed reading text T'           |
| **procedure** Pre-processing a Vietnamese reading text |
| L = splitting T into a list of single sentences.       |
| for i = 1 to len(L) do                                |
| L_i = Pre-processing for a raw Vietnamese sentence(L_i). |
| T' = a pre-processed reading text converted from the  |
| list L.                                               |
| **return** T'.                                       |

In Algorithm 1, firstly we use the tokenizer to break a sentence into a list of Vietnamese tokens X. In our work, this step performs in three steps, removing punctuation marks, stop words and noise words (short vowels) in the list X. After that, we convert the list X into a lower-case sentence S'. Lastly, we use the Vietnamese word segmentation tool to parse the sentence S' into a list of Vietnamese words L which is the output of this algorithm. We also apply Algorithm 1 to both questions and answer options. We use the tool pyvi3 for word segmentation in this algorithm.

In Algorithm 2, first of all, we split an input reading text into a list of sentences L. Then, we run the Pre-processing function (see Algorithm 1) for each sentence on all items of the list L. The output of this algorithm is a pre-processed reading text T' converted from the list L. Algorithm 1 and Algorithm 2 are implemented in reading texts and multiple-choice questions on MMRC models.

B. SLIDING WINDOW AND DISTANCE SCORES

We present how to calculate sliding window scores (see Algorithm 3) and distance scores (see Algorithm 4) in the original sliding window algorithm (SW), a lexical-based approach developed by [10]. This approach matches a bag of words, constructed from a question Q and an answer option O_i, with a given reading text, and calculates a TF-IDF style matching score for each answer option. The two algorithms are important components in our proposed model. To under-
FIGURE 1. System overview of our proposed method.

Algorithm 3 Calculating the Sliding-Widow Scores

Input: Reading text $T$, set of words in question $Q$, and set of words in answer options $O_1..4$.

Output: Returning the score of the best answer option.

\[
\text{procedure Calculating sliding-window scores}
\]

\[
C(w) = \text{Count}(w, T)
\]

\[
\text{Initialize a list } sw\text{ of sliding-window scores for answer options.}
\]

\[
\text{for } i = 1 \text{ to } \text{len}(O) \text{ do}
\]

\[
S = O_i \cup Q
\]

\[
sw_i = \max_{j=1}^{\text{|T|}} \sum_{l=1}^{\text{|S|}} \left\{ \log(1 + \frac{1}{C(T_{j+l})}) \right\}, \text{if } T_{j+l} \in S, \text{Otherwise}
\]

\[
\text{return } sw
\]

Algorithm 4 Calculating the Distance Scores

Input: Text $T$, set of reading-text words $T_W$, set of words in question $Q$, and set of words in answer options $O_1..4$.

Output: Returning the distance score for answer options of the question.

\[
\text{procedure Calculating distance scores}
\]

\[
\text{Initialize a list } d\text{ of distance score for answer options.}
\]

\[
\text{for } i = 1 \text{ to } \text{len}(O) \text{ do}
\]

\[
S_Q = Q \cap T_W
\]

\[
S_{O_i} = O_i \cap T_W
\]

\[
\text{if } |S_Q| = 0 \text{ or } |S_{O_i}| = 0 \text{ then}
\]

\[
d_i = 1
\]

\[
\text{else}
\]

\[
d_i = \frac{1}{|T|-1} \max_{q \in S_Q, a \in S_{O_i}} d(T, q, a)
\]

\[
\text{where } d(T, q, a) \text{ is the minimum number of words an occurrence of } q \text{ and an occurrence of } a \text{ in } T, \text{ increase } 1
\]

\[
\text{return } d
\]

In particular, we add a boosted score (denoted by $web_i$) to the final score of each answer option. Algorithm 5 presents how to calculate the boosted score. To understand Algorithm 5, we introduce two notations $V_T$ and $V_{O_i}$ to denote the ordered sets of words in the reading text $T$ and in the answer option $O_i$, respectively. We calculate $web_i$, the maximum cosine similarity between $V_{O_i}$ and span words $X$ of the same length in $V_T$. $\bar{v}$ is the average of the word embeddings of the lexical units in $v$. Fig. 2 shows semantic similarity architecture estimating the boosted core between an answer option and a span in $T$. The semantic similarity of the two vectors $V_{O_i}$ and $X$ is
Algorithm 5 Calculating the Boosted Score for Answer Options of a Question

**Input:** Reading text $T$, set of reading-text words $TW$, set of words in question $Q$, and set of words in answer options $O_{1\ldots4}$. $V^T$ and $V^O_i$ to denote the ordered sets of words in the reading text $T$ and in the answer option $O_i$, respectively. $\overline{v}$ is the average of the word embeddings of the lexical units in $v$. Note that $T$, $Q$, and $O_{1\ldots4}$ are pre-processed by Algorithm 1 and Algorithm 2.

**Output:** Returning the boosted score for answer options of a question.

**procedure** Calculating the boosted scores

1. Initialize a list $web_i$ of boosted score for answer options.
2. for $i = 1$ to $\text{len}(O)$ do
   
   $web_i = \max_{j \in \{T\}} \text{similarity} \left( V^{O_i}, \frac{V^T_{j-|V^O_i|-1}}{|V^O_i|} \right)$

3. return $web$

A. BASELINE MODELS

We evaluate random, lexical-based approaches (Sliding Window and Distance-based Sliding Window [10]) and a neural network based model (Co-match [47]) as baseline models. Sliding Window and Distance-based Sliding Window were used in the baseline methods of various datasets such as MCTest [10], SQuAD [2] and RACE [12]. Co-match is a strong neural network based model in multiple choice machine reading comprehension. Co-match achieve the positive results on RACE [12] and also was chosen to be the first baseline models on COSMOS QA [42] and C3 [27]. Despite the limited data size of our dataset, we verify to evaluate how well the Co-match method do and then analyze the need for increasing training data for neural network based models presented in Sub-section VI-E.

B. EVALUATION METRIC AND EXPERIMENTAL SETTINGS

We use accuracy as the primary evaluation metric which is computed as follows:

$$\text{Accuracy} = \frac{\text{Number of questions correctly answered}}{\text{Total number of questions}}$$

In all experiments, we use the word segmentation tool \texttt{pyvi}\(^4\) and six different pre-trained word embeddings proposed by [76]. The training, development and test sets are divided as shown in Table 2. Besides, we implement three methods such as Random, Sliding Window and Distance-based Sliding Window as baseline models on our dataset.

For the model Co-match, we fine-tune the model parameters suitable for the Vietnamese multiple-choice MRC. In particular, we use a mini-batch size of 32, and the hidden memory size of 10. The number of epochs is set to a number of 30. Adamax optimizer is used for optimization with a starting learning rate of 0.002. In this model, we also in turn test the same word embeddings used on our proposed model.

\(^4\)Vietnamese word segmentation tool: https://pypi.org/project/pyvi/
TABLE 5. Experimental results of different models with various pre-trained word embeddings on our dataset ViMMRC.

| Model               | Method                      | Dev (%) | Test (%) |
|---------------------|-----------------------------|---------|----------|
| Baselines           | Random                       | 24.49   | 24.80    |
|                     | Sliding Window (SW)          | 58.50   | 56.30    |
|                     | Distance-based SW            | 60.54   | 56.30    |
|                     | Co-match + W2V               | 43.97   | 41.49    |
|                     | Co-match + W2VC2V            | 43.77   | 43.87    |
|                     | Co-match + fastText          | 43.39   | 41.84    |
|                     | Co-match + Bert-base         | 42.61   | 43.88    |
|                     | Co-match + ELMo              | 45.38   | 44.94    |
|                     | Co-match + Multi             | 43.00   | 43.23    |
| Our proposed approach | Boosted score with W2V      | 61.91   | 60.04    |
|                     | Boosted score with W2VC2V    | 61.91   | 60.04    |
|                     | Boosted score with fastText  | 63.27   | 60.04    |
|                     | Boosted score with Bert-base | 63.27   | 61.24    |
|                     | Boosted score with ELMo      | 65.99   | 61.81    |
|                     | Boosted score with Multi     | 63.61   | 60.24    |
| Human performance   |                             | 91.20   | 91.10    |

C. HUMAN PERFORMANCE

We randomly take 100 questions from the test set and 100 questions from the development set. We conduct the tests on ten students. As a result, human performance reaches 91.20% in accuracy on the development set and 91.10% on the test set. These results are much higher than our best model. To overcome human performance is challenging to explore a new machine reading comprehension model suitable for this dataset in the future.

D. MODEL PERFORMANCE

We report the performances of the baseline models and our proposed model in Table 5. Sliding Window and Distance-based Sliding Window achieve different performances, 58.50% and 60.55%, on the development set but they have the same accuracy of 56.30% on the test set. Our proposed method achieves the accuracies over 60% on the test set and over 61% on the development set. Specifically, this method with the ELMo word embedding achieves the highest results on both of the test and development sets, 65.99% and 61.81%, respectively. This proves that our proposed method is more effective than the baseline models for the Vietnamese MMRC task at present with improvements of 5.45% and 5.51% on the development and test sets, respectively. However, these results are much lower than the human performance of 29.29% on the test set. This is a great challenge in the study of Vietnamese multiple-choice machine reading comprehension.

Comparing the experimental results of the Co-match model with different word embeddings, we can see that ELMo only achieves the best accuracy of 45.58% and 44.94% on development and test sets. However, ELMo is still the best word embedding on both lexical-based and neural-based approaches. In addition, the best performance of the Co-match model on the test set is 16.87% lower than that of our proposed model. It is also much lower than the human performance of 46.16%. Because the data size is not large enough, we evaluate this model on different sizes of training data in Sub-section VI-E, which helps us to make a decision whether to continue increasing the data size in future work.

VI. RESULT ANALYSIS

To gain insights into the best model (our proposed method with the ELMo embedding), we analyze the experimental results in terms of different aspects such as question length, reading-text level, reasoning type, and word embedding. Besides, we aim to evaluate how the size of our training set has an impact on the neural network-based method.

A. EFFECTS OF THE QUESTION LENGTH

To verify whether the length of question is a reason for the poor performance of our best model, we measure the performances of the best model according to the question length. In particular, we divide the development set into five groups corresponding to the following question lengths: \( \leq 10 \), \( 11 - 15 \), \( 16 - 20 \), \( 21 - 25 \), and \( \geq 26 \) words. The accuracies are analyzed and visualized in Fig. 3. As can be seen from the figure, questions of the 16 – 20 word length result in better performance than questions of other lengths. For short
questions, our method predicts less effective. This may be because short questions contain less information beneficial to search for the correct answer. In particular, the performances on shorter questions (64.15% for the ≤ 10-word questions and 65.18% for 10 – 15-word questions) are lower than the performances on longer questions which are over 66% in accuracy. Fig. 4 shows performance comparison between the best baseline and our proposed method with different groups of the question length. The accuracy of our proposed model has an improvement on all question lengths (except the question lengths over 25 words), of which the three groups (11 – 15, 16 – 20, 21 – 25) have significant increases.

B. EFFECTS OF THE READING TEXT LEVEL

Fig. 5 shows the accuracies of the best model according to different levels of reading text - the first to fifth grades. We can observe that the difficulty of the reading comprehension task increases together with the level of reading text. The system could answer questions of the 2nd grade well, over 78% in accuracy. It is more challenging to predict correct answers for questions of the 3rd to 5th grades (less than 68%). The performance on 1st grade questions is not as high as that on the 2nd grade questions because the amount of questions of the 1st grade is much fewer than those of other grades. Fig. 6 shows performance comparison between the best baseline and our proposed method with different reading text levels. The accuracy of our proposed model has an improvement on all reading text levels (except the first grade level), of which the three types (2-grade, 4-grade, and 5-grade) have significant increases.

C. EFFECTS OF THE REASONING TYPE

We also perform analysis to see how the reasoning types influence the best MMRC model. Fig. 7 shows the analysis results. We found that the system determines answers more efficiently for the of the word matching and the paraphrasing reasoning types (WM and PP), 92.11% and 82.93% in accuracy, respectively. In contrast, complex forms of reasoning result in lower performances. They include single-sentence reasoning, multi-sentence reasoning, and ambiguous-or-insufficient. Fig. 8 shows performance comparison between the best baseline and our proposed method
TABLE 6. Statistics of the performance and improvement of our proposed model according to different lengths of question.

| Question Length (words) | Ratio (%) | Accuracy (%) | Improvement (%) |
|-------------------------|-----------|--------------|-----------------|
|                         |           | Best Baseline| Our Proposed Method |                     |
| 10                      | 36.95     | 62.38        | 64.36           | +1.98              |
| 11−15                   | 38.10     | 58.68        | 64.46           | +5.78              |
| 16−20                   | 16.67     | 71.74        | 84.78           | +13.04             |
| 21−25                   | 6.12      | 58.82        | 64.71           | +5.89              |
| ≥26                     | 3.06      | 11.11        | 11.11           | 0.00               |

TABLE 7. Statistics of the performance and improvement of our proposed model according to different types of reasoning.

| Reasoning Type | Ratio (%) | Accuracy (%) | Improvement (%) |
|----------------|-----------|--------------|-----------------|
|                |           | Best Baseline| Our Proposed Method |                     |
| WM            | 25.85     | 84.21        | 92.11           | +7.90              |
| PP            | 13.95     | 70.73        | 82.93           | +12.20             |
| SSR           | 17.35     | 50.98        | 52.94           | +1.96              |
| MSR           | 36.73     | 48.15        | 50.00           | +1.85              |
| AoI           | 6.12      | 38.89        | 50.00           | +11.11             |

FIGURE 9. Performances of Co-match with various word embeddings according to different sizes of training data.

D. EFFECTS OF THE WORD EMBEDDINGS

Table 5 shows the experimental results with external knowledge sources as pre-trained word embeddings. It can be seen that the results are influenced by the methods when combined with these word embeddings. In particular, our lexical-based approach achieves better results when using word embeddings, approximately 5% higher. The experimental results show that ELMo is the best among the other word embeddings.

In addition, we conduct the detailed analysis of the effect of external knowledge integrated to our proposed model compared with the best baseline model according to different aspects such as the question length and reasoning type. In particular, Table 6 shows statistics of the performance and improvement of our proposed model according to different types of length. Our model improves the results of short questions (≤10 words) with an increasing accuracy of 1.98% and average-length questions with an improvement of 5.78% for 11−15-word questions and the one of 13.04% for 16−20-word questions. For longer questions, this model does not improve its performance. However, this number is not significant because the number of long questions accounts for low percentage. Table 7 shows statistics of the performance and improvement of our proposed model according to different types of reasoning. We found that our proposed model is a right solution for three types of reasoning, word matching, paraphrasing and ambiguous or insufficient, increasing 7.90%, 12.20% and 11.11% of the total number of solved questions, respectively. However, the number of questions
TABLE 8. Statistics of different reasoning types in the development set of our dataset ViMMRC. The correct answer in each question is the answer option in bold. The options which are exact answers are italicized in the examples.

| Reasoning Type              | Example                                                                 | Ratio (%) |
|-----------------------------|-------------------------------------------------------------------------|-----------|
| Word Matching               | Reading text: Vũa sáp sặc với ra bánh. Tướng bống nghe có tiếng chuẩn điện thoại. (Just putting the books on the table, tuong suddenly heard a phone ringing.)
Question: Việc gì đã xảy ra khi Tướng vừa sáp sặc với ra bánh? (What happened when tuong just put his books on the table?)
A. Mẹ nhờ Tướng đi chắt. (Mom asked tuong to go to the market.)
B. Có tiếng chuẩn điện thoại. (There is a phone ringing.)
C. Bạn rủ Tướng đi chắt. (Tuong’s friends invite him to go out.)
D. Nghe tiếng ai đó bên ngoài. (Hearing someone’s voice outside.) | 25.85     |
| Paraphrasing                | Reading text: Tôi đang nằm một chiếc cũ Cò-rét-ti đang khâu vải vào topi, làm cho cây bút nguyệt ra một đường rạng rỡ.
(When I was sharpening the letters word for word, coretti touched my elbows, making the pen scribble very badly.)
Question: Khi nhận vật tôi đang nằm một chiếc có-rết-ti. (The character ‘I’ made coretti’s written characters really ugly.)
A. Nhận vật tôi làm ngược chủ trong việc của Cò-rét-ti. (The character ‘T’ was writing the lesson, what happened?)
B. Cò-rét-ti cãi nhau với một chiếc viết ngược. (Coretti quarreled over a scribble.)
C. Cò-rét-ti chăm khâu tay làm lúi của ngược chủ. (Coretti touched the elbows, making me scribble.)
D. Nhận vật tôi và Cò-rét-ti làm trang nhâu đó động. (The character ‘T’ and Coretti competed together for getting something.) | 13.95     |
| Single-sentence Reasoning   | Reading text: Khi tiếng đàn, tiếng hát của A ri-ôn vang lên, có một đêm cai heo đã bỏ để vay quanh tầu, say sưa thuong troi tiếng hát của neghe si tại tại.
(As the sound of Arion’s musical playing and singing started, a group of dolphins swam around the ship, passionately enjoying the singing of the talented artist.)
Question: Dùu kí là gì đã xảy ra khi nghe si A ri-ôn cài tiếng hát giát biệt cười đối? (What strange thing happened when the artist Arion sang goodbye to life?)
A. Dàn ca heo đã an thứ ông. (The dolphins swallowed him.)
B. Dàn ca heo đã bố chây đi mất. (The dolphins ran away.)
C. Dàn ca heo đã nhận chim ông xuống biển. (The dolphins drop him to the sea.)
D. Dàn ca heo đã bố để vay quanh tầu. (The dolphins swam around the boat.) | 7.35      |
| Multiple-sentence Reasoning | Reading text: Chim dừng hạt nữa, bá ếp em rồi, lạng cho bá ngoài. Bàn tay bể nhỏ, vậy quất thật đau. Ngân nàng thích thú, dưới trên trong trắng. Cau nhà đã vàng, có chén nằm im. Đợi mất lìm dim, ngủ ngon bá nhẹ.
(Bird! Please don’t sing, my grandma is sick, keep silent for her to sleep. Tiny hands are waving fans evenly. Sunlight stale parked on the white wall. The house is empty. The cup lies still. Eyes dim sleep. Sleep well, my grandma.)
Question: Bạn nhớ đang làm gì? (What was the young boy doing?)
A. Ngâm cây cô trong vườn. (Viewing the trees in the garden.)
B. Nói chuyện với chim cúc chở. (Talking with the warbler.)
C. Đơn đẹp nhà cau. (Cleaning his house.)
D. Quat cho bá ngoại. (Waving fans for his grandma’s sleep.) | 36.73     |
| Ambiguous or insufficient   | Reading text: Câu bé nhìn bá, suy ngợi một chút rồi thì thẩm: những cặp nhất, bá ai!.
(The boy looked at his grandma, thought for a while and whispered: "The wrinkle, grandma!")
Question: Câu trả lời cuối cùng của câu bé muốn nói lên điều gì? (What is the meaning of the boy’s last answer?)
A. Câu trả thì những người có cặp nhất. (The boy likes people with wrinkles very much.)
B. Câu thay những cặp nhất vây đẹp. (The boy thinks that wrinkles are very beautiful.)
C. Trong đợi mặt câu, những cặp nhất của bà rất đẹp và câu rất yêu những cặp nhất này. (In the boy’s eyes, wrinkles are very beautiful and he loves these wrinkles.)
D. Trong đợi mạt câu, hiện ra những vật nhất của cô giả. (In the boy’s eyes, there are the girl’s wrinkle.) | 6.12      |

of word matching and paraphrasing improved significantly because they account for a high proportion in the dataset.

E. EFFECTS OF THE TRAINING DATA SIZE

To verify whether the size of training data is a reason for the poor accuracy of the machine model, we evaluate Co-match [47] as a neural network-based model on different sizes of training data including 508, 1010 and 1975 human-created questions. In this experiment, we implement Co-match with different pre-trained word embeddings [76]. Experimental results (in accuracy) on the test set are presented in Fig. 9. The figure shows that the model performance is improved when we increase the amount of the training data. These observations suggest that increasing training data size would improve the accuracy. This is also a future direction for addressing this problem.

VII. CONCLUSION AND FUTURE WORK

In this paper, we propose the lexical-based approach utilizing semantic similarity and external knowledge sources and
perform experiments to compare the performances between this method and baseline lexical-based and neural network-based methods. The experimental results show that our proposed method is effective for Vietnamese multiple-choice reading comprehension. The best performance reaches 61.81% in accuracy on our dataset. However, there is still a large gap between the human performance and the best model (a significant difference of 29.29%). We also analyze the best models in different linguistic aspects to gain in-depth insights into the data. These analyses result illustrate that our corpus is a challenging dataset and need further studies. We also contribute this dataset for studies of the multiple-choice machine reading comprehension task for the Vietnamese language. This dataset includes 2,783 multiple-choice questions based on a set of 417 Vietnamese reading texts. This dataset encourages further advances in machine reading comprehension and guides the development of artificial intelligence for the Vietnamese language.

In future, we plan to increase the quantity and quality of the dataset in terms of the number of reading texts. The analysis results also suggest that we should focus on methods to improve the performance on long questions and difficult reasoning types. When the dataset is large enough, we will further research on state-of-the-art methodologies such as deep neural networks and transfer learning to explore suitable models for Vietnamese multiple-choice MRC. In addition, we can use classification of the level of difficulty on multiple-questions to conduct experiments with curriculum learning [77].

APPENDIX
STATISTICS OF DIFFERENT REASONING TYPES
Table 8 shows the ratio of each reasoning type in the development set. Those types of reasoning have been described in Table 8 shows the ratio of each reasoning type in the development set. Those types of reasoning have been described in Section III. Besides, we have an example for each reasoning type and these Vietnamese examples are translated into English.

ACKNOWLEDGMENT
The authors would like to appreciate the editors and anonymous reviewers for their helpful feedback.

REFERENCES
[1] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, “Teaching machines to read and comprehend,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 1693–1701.
[2] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000 questions for machine comprehension of text,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2016, pp. 2383–2392.
[3] E. Choi, H. He, M. Iyer, M. Yatskar, W.-T. Yih, Y. Choi, P. Liang, and L. Zettlemoyer, “QuAC: Question answering in context,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2018, pp. 2174–2184.
[4] C. Park, H. Song, and C. Lee, “S^2-Net: SRU-based sentence and self-matching networks for machine reading comprehension,” ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 19, no. 3, pp. 1–14, Apr. 2020, doi: 10.1145/3365679.
[5] S. Reddy, D. Chen, and C. D. Manning, “CoQA: A conversational question answering challenge,” Trans. Assoc. Comput. Linguistics, vol. 7, pp. 249–266, Nov. 2019.
[6] E. M. Voorhees, “The TREC-8 question answering track report,” in Proc. TREC, vol. 99, 1999, pp. 77–82.
[7] E. M. Voorhees and H. T. Dang, “Overview of the TREC 2002 question answering track,” in Proc. TREC, 2003, pp. 54–68.
[8] P. Nakov, D. Hoogeveen, L. Márquez, A. Moschitti, H. Murphy, T. Baldwin, and K. Verspoor, “SemEval-2017 task 3: Community question answering,” 2019, arXiv:1912.00730. [Online]. Available: http://arxiv.org/abs/1912.00730
[9] D. Gupta, A. Ekbol, and P. Bhattacharyya, “A deep neural network framework for English Hindi question answering,” ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 19, no. 2, pp. 1–22, Nov. 2019, doi: 10.1145/3359988.
[10] M. Richardson, C. J. Burgess, and E. Renshaw, “Mctes: A challenge dataset for the open-domain machine comprehension of text,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2013, pp. 193–203.
[11] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen, “A corpus and cloze evaluation for deeper understanding of commonsense stories,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., 2016, pp. 839–849.
[12] K. V. Nguyen, Q. Xie, H. Liu, Y. Yang, and E. Hovy, “RACE: Large-scale ReAding comprehension dataset from examinations,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2017, pp. 785–794.
[13] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth, “Looking beyond the surface: A challenge set for reading comprehension over multiple sentences,” in Proc. NAACL-HLT, 2018, pp. 252–262.
[14] S. Ostermann, M. Roth, A. Modi, S. Thater, and M. Pinkal, “SemEval-2018 task 11: Machine comprehension using commonsense knowledge,” in Proc. 12th Int. Workshop Semantic Eval., 2018, pp. 747–757.
[15] S. Liu, S. Zhang, X. Zhang, and H. Wang, “R-trans: RNN transformer network for Chinese machine reading comprehension,” IEEE Access, vol. 7, pp. 27736–27745, 2019.
[16] K. Liu, X. Liu, A. Yang, J. Liu, J. Su, S. Li, and Q. She, “A robust adversarial training approach to machine reading comprehension,” in Proc. AAAI, 2020, pp. 8392–8399.
[17] F. Yuan, L. Shou, X. Bai, M. Gong, Y. Liang, N. Duan, Y. Fu, and D. Jiang, “Enhancing answer boundary detection for multilingual machine reading comprehension,” 2020, arXiv:2004.14069. [Online]. Available: http://arxiv.org/abs/2004.14069
[18] P. Lewis, B. Oğuz, R. Rinott, S. Riedel, and H. Schwenk, “MLQA: Evaluating cross-lingual extractive question answering,” 2019, arXiv:1910.07475. [Online]. Available: http://arxiv.org/abs/1910.07475
[19] D. Chen, J. Bolton, and C. D. Manning, “A thorough examination of the CNN/daily mail reading comprehension task,” in Proc. 54th Annu. Meeting Assoc. Comput. Linguistics, 2016, pp. 2358–2367.
[20] B. Dinhra, H. Liu, Z. Yang, W. W. Cohen, and R. Salakhutdinov, “Gated-attention readers for text comprehension,” 2016, arXiv:1606.01549. [Online]. Available: https://arxiv.org/abs/1606.01549
[21] H. Zhu, F. Wei, B. Qin, and T. Liu, “Hierarchical attention flow for multiple-choice reading comprehension,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 6077–6084.
[22] X. Liu, K. Liu, X. Li, J. Su, Y. Ge, B. Wang, and J. Luo, “An iterative multi-source mutual knowledge transfer framework for machine reading comprehension,” in Proc. IJCAI, 2020, pp. 1–7.
[23] B. Bi, C. Wu, M. Yan, W. Wang, J. Xia, and C. Li, “Incorporating external knowledge into machine reading for generative question answering,” 2019, arXiv:1909.02745. [Online]. Available: http://arxiv.org/abs/1909.02745
[24] H. Chen, J. Bolton, and C. D. Manning, “A thorough examination of the CNN/daily mail reading comprehension task,” in Proc. 54th Annu. Meeting Assoc. Comput. Linguistics, 2016, pp. 2358–2367.
[25] Y. Xu, W. Liu, G. Chen, B. Ren, S. Zhang, S. Gao, and J. Guo, “Enhancing machine reading comprehension with position information,” IEEE Access, vol. 7, pp. 141602–141611, 2019.
[26] Y. Chen, S. Kang, and J. Seo, “Improved machine reading comprehension using data validation for weakly labeled data,” IEEE Access, vol. 8, pp. 5667–5677, 2020.
[27] K. Sun, D. Yu, D. Yu, and C. Cardie, “Investigating prior knowledge for challenging Chinese machine reading comprehension,” Trans. Assoc. Comput. Linguistics, vol. 8, pp. 141–155, Jul. 2020.
[28] F. Hill, A. Bordes, S. Chopra, and J. Weston, “The goldilocks principle: Reading children’s books with explicit memory representations,” in Proc. IJCLP, 2016, pp. 1–13.
K. V. Nguyen et al.: Enhancing Lexical-Based Approach With External Knowledge for Vietnamese Multiple-Choice MRC

[72] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., 2018, pp. 2227–2237.

[73] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013, arXiv:1301.3781. [Online]. Available: http://arxiv.org/abs/1301.3781

[74] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, “Character-aware neural language models,” in Proc. 30th AAAI Conf. Artif. Intell., 2016, pp. 1–9.

[75] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Trans. Assoc. Comput. Linguistics, vol. 5, pp. 135–146, Dec. 2017.

[76] X.-S. Vu, T. Vu, S. N. Tran, and L. Jiang, “ETNLP: A visual-aided systematic approach to select pre-trained embeddings for a downstream task,” in Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP), 2019, pp. 1285–1294.

[77] Y. Liang, J. Li, and J. Yin, “A new multi-choice reading comprehension dataset for curriculum learning,” in Proc. Asian Conf. Mach. Learn., 2019, pp. 742–757.

KOIET VAN NGUYEN received the B.S. and M.S. degrees from the University of Information Technology, Ho Chi Minh City, Vietnam, in 2012 and 2017, respectively. He is currently a Lecturer with the Faculty of Information Science and Engineering, University of Information Technology, Vietnam National University, Ho Chi Minh City. His research interests include natural language processing, machine reading comprehension, and deep learning.

KHIEI VINH TRAN is currently a junior student with the University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam. His research interests include text processing, machine reading comprehension, and sentiment analysis. He just took part in the WNUT-2020 Task 2 and ranked the third place in this competition.

SON T. LUU received the B.S. degree from the University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam, in 2019, where he is currently pursuing the master’s degree. He is currently a Research Assistant with the University of Information Technology. His research interests include machine reading comprehension, toxic comment detection, sentiment analysis, and knowledge representation.

ANH GIA-TUAN NGUYEN received the B.S., M.S., and Ph.D. degrees in information technology from the University of Science, Vietnam National University, Ho Chi Minh City, Vietnam, in 1995, 1998, and 2013, respectively. He is currently the Dean of the Faculty of Information Science and Engineering, University of Information Technology, Vietnam National University. His research interests include GIS and intelligent system.

NGAN LUU-THUY NGUYEN received the Ph.D. degree in information science and technology from the University of Tokyo, Japan. She was a Postdoctoral Researcher with the National Institute of Informatics, Japan, from 2012 to 2013. She is currently a Scientist with the University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam. Her research interests include natural language processing and data analysis.

* * *