An Approach for Multiple Choice Question Answering System

Dinh-Huy Vo[1], Anh-Khoa Do-Vo[1], Tram-Anh Nguyen-Thi[2], and Huu-Thanh Duong[1(✉)]

[1] Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
{1851010052huy,1851010057khoa,thanh.dh}@ou.edu.vn
[2] Department of Fundamental Studies, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
tramanh.nguyen@ou.edu.vn

Abstract. Multiple choice question (MCQ) system, which is a form of question answering system, includes a question, a set of choices and the correct answers from these choices. The rapidly growth of linguistics models improves natural language understanding and motivates the automatic systems in natural language processing. This paper proposes an approach based on the Language Model for MCQ system as incomplete questions in English tests. This not only reduces the burden of human experts to teach their students, but also is useful for self-studying the students. We perform many experiments to evaluate the effectiveness and achieve 85.45% accuracy for our proposal.

Keywords: MCQ · BERT · MLM · Multiple choice · Language model · Unigram · Bigram · Masked Language Model

1 Introduction

Question answering (QA), which is an essential task in natural language processing (NLP), gives the succinct and short answers for the user's query in natural language, MCQ system is a subdomain of QA system where a question and the limited multiple choices are given and we need to choose the correct answers among these choices in particular context without user intervention.

MCQ plays a major role in educational assessment. With abundant resources and rapid growth of the pre-trained models in English, the machine totally answers automatically the multiple choice questions in the specific context. Using the automatic system not only reduces the burden of human experts to teach and evaluate their students, but also is useful for self-studying. In this paper, we develop the MCQ system for the forms of incomplete questions in English exams in any subjects such as grammar, nouns, pronouns, vocabularies, verb tenses based on our training corpus which is collected from the well-known websites.

P. Cong Vinh and N. Huu Nhan (Eds.): ICTCC 2021, LNICST 408, pp. 76–82, 2021.
https://doi.org/10.1007/978-3-030-92942-8_7

Our main contribution has performed the existing methods for multiple choice question problems such as Language Model (unigram, bigram) and Masked Language Model (MLM), proposed and evaluated an approach as an ensemble of above methods to improve the accuracies of multiple choice question problems.

The rest of this paper is organized as follows. Section 2 presents the related works. Section 3 presents methods that are used for the automated English multiple-choice test system. Next, Sect. 4 shows the experimental results. Finally, the conclusion and our future direction of system development will be presented in Sect. 5.

2 Related Works

QA systems have been concerned and has many proposals for it such as Duong and Ho [1] used Language Model for QA system in legal documents, Duong and Hoang [2] combined of machine learning classifiers in news classification, Y. Sharma et al. [3] applied the deep learning approaches for QA system, W. Yu et al. [4] used transfer learning with ALBERT to build QA system in technical domains.

As a form of QA system, MCQ has emerged and attracted many research teams to resolve the multiple choice questions in recent years. A. Chaturvedi et al. [5] proposed to use CNN model to answer the multiple choice questions in particular domains. The accuracies are better than the baseline results as LSTM model in two datasets as Textbook question answering (TQA) and SciQ datasets. K. Moholkar et al. [6] proposed an ensemble of models approach, including LSTM, hybrid LSTM-CNN and multiplayer perceptron models to predict the correct answer. LSTM and hybrid LSTM-CNN models are firstly trained parallelly, then the multilayer perceptron model predicts the correct answer. This approach obtains higher accuracies than using the models separately. R. Chitta and A.K. Hudek [7] developed the question answering system for multiple choice questions in legal contracts. They have extracted the relevance text in contracts and used a multi-class classifier to choose the correct answer based on the extracted text.

Language Model (LM) is a statistical approach learning the probability distribution over a sequence of words based on the existing corpus. This model is widely used in information retrieval. This refers to an effective solution to predict the correct choice of the multiple choice question in incomplete form based on its probability distribution.

J. Devlin et al. at Google API Language published BERT (Bidirectional Encoder Representations from Transformers) in [8], this applies the bidirectional training of a transformer instead of only looking at the single directional training like the previous efforts. In this paper, they also introduced a novel technique named Masked Language Model (MLM). MLM is an important model of BERT, here we give a sentence to which some words are attached [MASK] and then predict these words. The benefit of BERT is that it predicts the hidden word

through the context of words on the right and left so when we cover a word, it can predict the hidden word effectively because based on the context of the whole question. Thus, MLM is a suitable method to resolve multiple choice questions in incomplete form.

3 The Approach

We use a statistical approach including LM (unigram and bigram) and MLM. These models are conducted separately to evaluate the effectiveness. Then we propose an ensemble of two models to improve the performances of MCQ problems. The experiments present this approach achieves better accuracy significantly than applying the models separately.

3.1 Language Model

The main purpose of LM is to calculate probability distribution over a sequence of words: w_1, w_2, w_3, $\ldots w_n$ based on the Naïve Bayes formula:

$$P(w_1 w_2 \ldots w_n) = P(w_1) \times P(w_2|w_1) \times \ldots \times P(w_n|w_1 w_2 \ldots w_{n-1}) \quad (1)$$

According to this formula, LM needs to use a large amount of memory to store the results of the string. This would be impossible if the length of the string is a paragraph or even more because the length of a natural text is infinite. The n-gram models (or n-level Markov Models) will solve this problem by calculating the probability of the sequence with lower memory. The occurrence probability of (w_m) will depend on n consecutive words preceding it, not on the entire sequence. Thus we have the probability of word (w_m) calculated as follows:

$$P(w_1 w_2 \ldots w_m) = P(w_1) \times P(w_2|w_1) \times \ldots \times P(w_m|w_{m-n-1} w_{m-n} \ldots w_{m-1}) \quad (2)$$

For $n = 1$ calls unigram which is the simplest model in n-gram. It evaluates the probability of each word in a sentence independently as below formula.

$$P(w_i) = \frac{C(w_i)}{N} \quad (3)$$

Where $P(w_i)$ is the probability of the word w_i, $C(w_i)$ is the number of occurrences of w_i, N is the total number of words in the training corpus.

For $n = 2$ calls bigram, this will consider the probability of occurrence of two consecutive words in a sentence according to the formula:

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1} w_i)}{C(w_{i-1})} \quad (4)$$

3.2 Masked Language Model

Masked Language Model is the most important model structure of BERT. MLM combines Transformer encoder and masked tokens can predict a missing token in a sentence. The output of MLM is the word embeddings of corresponding tokens which feed to a simple softmax classifier to get the final prediction.

The missing token is represented by the [MASK] symbol in a sentence and MLM predicts the suitable tokens for replacement. BERT can predict through the context of words in its right and left directions. It assumes the missing token is at position t (w_t) , the context tokens to predict includes w_1, w_2, ..., w_{t-1}, $w_t + 1$, ..., w_n (n is the number of tokens in a sentence). This considers the context of the whole sentence, so this can predict the missing token.

3.3 Our Proposal

MLM uses bidirectional context words of the hidden word and considers the context of the whole sentence, so this is a good approach for incomplete forms. However, for multiple choice questions, the correct answer only limits specific choices. As a result, MLM cannot sometimes give the final answer since the predicted words do not match any choices. Otherwise, n-gram models always obtain the final answer, but they only consider the previous words of the hidden word. Thus, our proposal gives the combination of n-gram (unigram and bigram) models and MLM. In order to determine the correct answer, we use a formula based on a linear equation:

$$f(x, y) = \lambda x + (1 - \lambda)y \tag{5}$$

Where λ is a coefficient between [0–1], x is the score of choosing the answer by MLM, y is the score of choosing the answer of the unigram or bigram.

4 The Experiments

For the training corpus, we have extracted 207,319 documents (magazines, articles, etc.) from the well-known websites in English and prepared 502 multiple choice questions with answers for testing to evaluate our approach (Table 1 shows some samples of testing data). The experiments are firstly conducted on unigram, bigram, MLM separately, the selected answer has the highest score from these models. Afterward, we combine these models, including (unigram and MLM) and (bigram and MLM), the scores are determined by formula (5).

Table 2 presents the accuracies of various experiments, where unigram and bigram always give the answer and obtain 29.48% and 57.37% accuracies respectively. The multiple choice questions demand the formal structures and grammars, so the order of words is essential in text. The unigram does not concern this, leading to this model obtaining a low accuracy in this problem. The bigram model calculates the probability of a word based on the context of its previous word, it improves accuracy significantly compared to the unigram model.

For MLM, the selected answer is suggested by MLM and matches one of the choices of the question, its accuracy is 44.02%. This model predicts the hidden word based on its bidirectional words. However, there are many cases in which it is impossible to produce the results since some suggested words of MLM not matching any choices of the question.

The combination of unigram and MLM obtains 82.86% accuracy with λ in (0.2, 0.3, 0.4, 0.5). The highest result, which is the combination of bigram and MLM, is 85.45 % accuracy with λ as 0.99. In order to select a good λ value, we perform many experiments with various λ values, the Fig. 1 and Fig. 2 show the results.

Table 1. The samples of testing data.

Question	A	B	C	D	Answer
The simplest way to reduce your ... footprint is to cycle to school	Carbon	Chemical	Chemistry	Dioxide	Carbon
My brother has to work ... a night shift once a week	In	On	At	By	On
It is hardly possible to ... the right decision all the time	Do	Arrive	Make	Take	Make
Buying organic food is better for the environment because it uses less ...	Fertilizer	Fertilize	Fertilizes	Fertilized	Fertilizer
No one on the plane was alive in the accident last night, ... they?	Wasn't	Weren't	Were	Was	Were
The movie is ... on a true story	Based	Keen	Hang	Let	Based
My father learned to play ... piano when he was five years old	A	An	The	Was	The
The dog was frightened by the sound of the thunder	Belt	Bolt	Bell	Bull	Bolt
It turned out that i ... have bought frank a present after all	Oughtn't	Mustn't	Needn't	Mightn't	Needn't

Table 2. The accuracies of various experiments.

Unigram	Bigram	MLM	Unigram + MLM ($\lambda = 0.50$)	Bigram + MLM ($\lambda = 0.99$)
29.48%	57.37%	44.02%	82.86%	*85.45%*

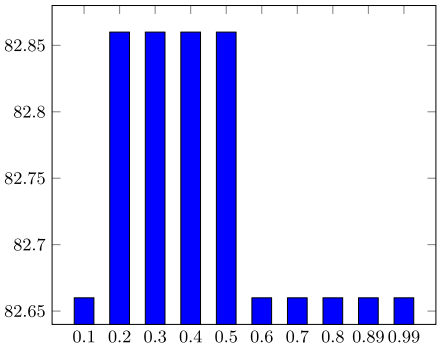

Fig. 1. Unigram combined with MLM with various λ values.

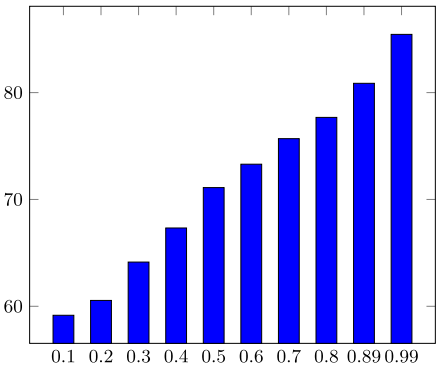

Fig. 2. Bigram combined with MLM with various λ values.

5 Conclusions and Future Works

In this paper, we have built an effective MCQ system for multiple choice questions in incomplete form for the English test. The experimental section conducts several experiments based on LM (unigram, bigram) and MLM. Moreover, we combine the LM and MLM and achieve promising accuracy.

In future works, we will broaden the system with more forms of multiple choice questions, resolve multiple choice questions of reading comprehension form based on the context of a defined paragraph or answer many multiple choice questions depending on an incomplete paragraph (such as incomplete part of the TOEIC test).

References

1. Duong, H.-T., Ho, B.-Q.: A Vietnamese question answering system in Vietnam's legal documents. In: Saeed, K., Snášel, V. (eds.) CISIM 2014. LNCS, vol. 8838, pp. 186–197. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-45237-0_19
2. Duong, H., Truong Hoang, V.: Question answering based on ensemble classifier for university enrolment advising. In: 2019 11th International Conference on Knowledge and Smart Technology (KST), Phuket, Thailand, 2019, pp. 35–39 (2019). https://doi.org/10.1109/KST.2019.8687616
3. Sharma, Y., Gupta, S.: Deep learning approaches for question answering system. Procedia Comput. Sci. **132**, 785–794 (2018). https://doi.org/10.1016/j.procs.2018.05.090
4. Yu, V., et al.: A technical question answering system with transfer learning. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 92–99 (2020) https://doi.org/10.18653/v1/2020.emnlp-demos.13

5. Chaturvedi, A., Pandit, D., Garain, U.: CNN for text-based multiple choice question answering. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers, Melbourne, Australia, 2018, pp. 272–277 (2018). https://doi.org/10.18653/v1/P18-2044
6. Moholkar, K., Patil, S.H.: Multiple choice question answer system using ensemble deep neural network. In: 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 762–766 (2020). https://doi.org/10.1109/ICIMIA48430.2020.9074855
7. Chitta, R., Hudek, A.K.: A reliable and accurate multiple choice question answering system for due diligence. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, Montreal QC Canada, pp. 184–188, June 2019. https://doi.org/10.1145/3322640.3326711
8. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding, May 2019. arXiv:1810.04805 [cs]. Accessed 01 May 2021, arXiv:1810.04805
9. Salazar, J., Liang, D., Nguyen, T.Q., Kirchhoff, K.: Masked language model scoring. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2699–2712 (2020). https://doi.org/10.18653/v1/2020.acl-main.240