An approach to automatic answering for English reading comprehension tests

Phat Tien Bui[1], Hieu Chi Tran[1], Thanh Huu Duong[1*]

[1]Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
[*]Corresponding author: thanh.dh@ou.edu.vn

ARTICLE INFO

DOI:10.46223/HCMCOUJS.tech.en.14.2.2917.2024

Received: August 20th, 2023
Revised: February 20th, 2024
Accepted: February 26th, 2024

Keywords:

BERT; multiple choice; Masked Language Model (MLM); Question Answering (QA); tokenization; transformer; word embeddings

ABSTRACT

This study focuses on the reading comprehension problem with multiple-choice answers, using the BERT model to achieve the highest performance. The ultimate goal is to create a solution to help solve reading comprehension problems without any reasoning or knowledge, suitable for the level of students in grades six and seven. The model will solve factoid questions from a given text. Our research topic will use a deep learning model-based approach to create a model that automatically answers the English reading comprehension question. We obtain promising results to give an accuracy of 78 percent.

1. Introduction

In the field of Natural Language Processing (NLP), automatic Question and Answer (QA) problems are one of the top concerns. QA requires the model to be able to understand and analyze the content of the text, and then answer the questions posed related to that content. The main goal of QA response model research is to develop Machine Learning or Deep Learning models that are capable of answering questions related to the content of a piece of text.

These models need to be able to understand contexts and natural language representations correctly, in order to give accurate and logical responses. Researching QA response models also sets the goal of improving the efficiency and accuracy of existing models, and providing new approaches to solving QA challenges, such as handling multifaceted words, dealing with complex questions and requiring extensive knowledge. The goal of research on QA response models is also to bring a lot of potential for practical applications such as chatbots, automated answers to questions on websites, or automated customer support systems.

Theoretically, research on QA response models contributes to the development of methods for natural language processing and understanding, and at the same time solves one of the important problems in the field of natural language processing course. These studies also contribute to the development of Machine Learning and Deep Learning models, which improve the efficiency and accuracy of QA response models. In practice, studying the QA response model contributes to the development of practical applications such as chatbots, automatic answers to English reading comprehension questions. In addition, the model can also be developed to address enrollment and academic problems.

The field of QA model research is facing many difficult challenges, requiring researchers to have novelty and creativity to solve these problems. One of the remarkable novelties and innovations is the creation of new datasets to test the model. Creating new datasets can help test and evaluate new QA response models more accurately and comprehensively, help present new challenges to solve, and develop new methods for collecting data and natural language processing.

2. Theoretical basis

In the study of the Reading Comprehension multiple choice problem, there are many fundamental theories used to develop explanations and research methods.

2.1. Natural language processing

Natural language processing theory plays an important role in understanding and interpreting questions and reading passages in the Reading Comprehension multiple choice problem. This theory includes methods and techniques for natural language processing such as parsing, word splitting, data normalization, and information extraction.

2.2. Neural network theory

Neural network theory has special significance for building machine learning models to solve the problem of Reading Comprehension multiple choice. This theory provides methods and techniques for training and using neural network models to predict correct answers to questions in reading passages.

2.3. Deep learning

Deep learning theory also plays an important role in building machine learning models to solve the Reading Comprehension multiple choice problem. This theory provides methods and techniques to train deep neural network models to extract information from readings and predict answers to related questions. Applying the above theories to the Reading Comprehension multiple choice study will help increase the accuracy and efficiency of the model, and at the same time clarify the content of each theory and apply them to problem solving specific math.

2.4. Related works

There have been many studies put forth to solve QA problems in reading comprehension problems such as:

– Chen, Bolton, and Manning (2016) investigated the question-answering problem in CNN/Daily Mail news reporting corpus using different deep learning models.

– Dodge et al. (2020) presented the problems associated with continuing to train pre-training language models like Bert to improve their performance in the question-answering domain.

– Liu and Lane (2016) proposed to use the sequential neural network model with attention mechanism to solve the problem of intent classification and gap filling in question answering.

– Zhang et al. (2017) proposed to use the neural network model to understand and answer the questions. The paper presents the use of the BERT language representation model to classify intents and fill in the blanks in the text (Chen, Zhuo, & Wang, 2019). The author emphasizes that these tasks often require correctly labeled training data and that the use of BERT can improve the generalizability of the model.

– Bordes, Chopra, and Weston (2014) presented a system that learns to answer questions from a knowledge base using manual features. The model learns to embed low-dimensional knowledge base words and components and use them to evaluate candidate questions and answers.

All of the above works show the use of different machine learning models, which can be based on techniques such as CNN, RNN, LSTM, Attention, BERT, or a combination of many different techniques for the problem of Reading Comprehension. It is also shown that the use of BERT and natural language processing models will significantly improve the accuracy of the model on the Reading Comprehension multiple choice problem. Different works focus on various aspects of the question-answering task, including model architectures, fine-tuning strategies, and the utilization of attention mechanisms. Each work contributes to the understanding and advancement of question-answering techniques, with a mix of classical and modern deep learning approaches. However, there are still many challenges and problems to be solved to improve the efficiency of the model, such as enhancing inference skills and making correct predictions for outliers.

3. The approach

We have used BERT (Devlin, Chang, Lee, & Toutanova, 2019) or the QA (Question Answering) system, BERT (Bidirectional Encoder Representations from Transformers) is a very popular deep learning model to solve automatic question-answering problems. BERT is a neural network architecture based on a Transformer model, capable of understanding the semantics of words and sentences in a corpus. To use BERT for QA, we need to train the model on a dataset of questions and answers. This training was done by us using datasets such as SQuAD v2 (Stanford Question Answer Dataset). Once we have obtained the best QA model trained from the SQuAD v2 dataset, we use that model to generate answers from the given text and questions. The answers generated from the model will be compared with the multiple-choice answers using the cosine similarity method after we conduct sentence embedding of the multiple-choice questions and the answers generated from the QA model using the sentence-transformers/all-MiniLM-L6-v2 model (Hugging Face, n.d.).

3.1. Question answering language model

A QA Language model (Question-Answering Language Model) is a type of modeling language that is trained to answer questions based on knowledge sources such as text or structured knowledge. To answer a question, the input to the QA Language model includes the question and the documentation associated with the question. These documents can be structured documents or knowledge, and are often represented as vectors or matrices. The question and document are then fed into a deep learning model that learns to analyze the question and document and find the most appropriate answer.

This model will generate a probability distribution (softmax) - the softmax function is commonly used in machine learning to compute the probability distribution over classes, for all possible answers to the question - for all possible answers to the question. To find the final answer, one can choose the answer with the highest probability or use other methods such as keyword search or using decoding techniques. The formula of the softmax function is as follows:

Given an input vector **z** with k elements **z** = [z_1, z_2, ..., z_k], the softmax function computes the probability distribution for each element in the input vector, denoted **p** = [p_1, p_2, ..., p_k], by applying the following formula:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad (1)$$

Where *(z)*$_i$ is the ith element of the input vector z having k elements, and $\sigma(z)_i$ is the probability allocation for that element. The exponent e is the natural exponent and$\sum_{j=1}^{k} e^{z_j}$ is the sum of the exponential values of all the elements in the input vector.

3.2. Transformer

Transformer is a neural network architecture used in the BERT (Bidirectional Encoder Representations from Transformers) model (Vaswani et al., 2023). The BERT model is a natural language model trained on large amounts of unlabeled text data, and it can be used to perform a variety of NLP tasks, such as text classification. text, machine translation, text summaries, and more.

Transformer is a neural network architecture that allows the model to learn non-linear relationships between words in a sentence, without having to use traditional techniques like LSTM or GRU. Transformer is designed to learn self representations of words and the dependencies between words in the same sentence or text.

In the BERT model, the Transformer is used to build a multidirectional coding model, allowing the model to learn the dependencies between words in both directions. Specifically, the BERT model uses a series of Transformer encoding layers to represent words in a sentence as vectors, and then uses these vectors to perform various NLP tasks. The use of Transformer in the BERT model has significantly contributed to the model's success in many NLP tasks.

BERT achieves a deep and flexible understanding of language representation by combining two important methods in the training process: Masked Language Model (MLM) and Next Sentence Prediction (NSP). During the MLM process, BERT mainly focuses on understanding the meaning of each word in its context. This helps BERT learn representations of words in a way that is natural, flexible, and independent of the direction of the sentence. NSP, on the contrary, focuses on the relationships between sentences. BERT is trained to predict whether a sentence is a continuation of another sentence. Thanks to this process, the model is able to understand the context between consecutive sentences, helping to create a more flexible linguistic representation. Combining both methods, BERT can deeply understand the meaning of each word in a sentence and also recognize the relationship between sentences in a paragraph. This makes BERT a powerful and multitasking language model, suitable for many natural language processing applications.

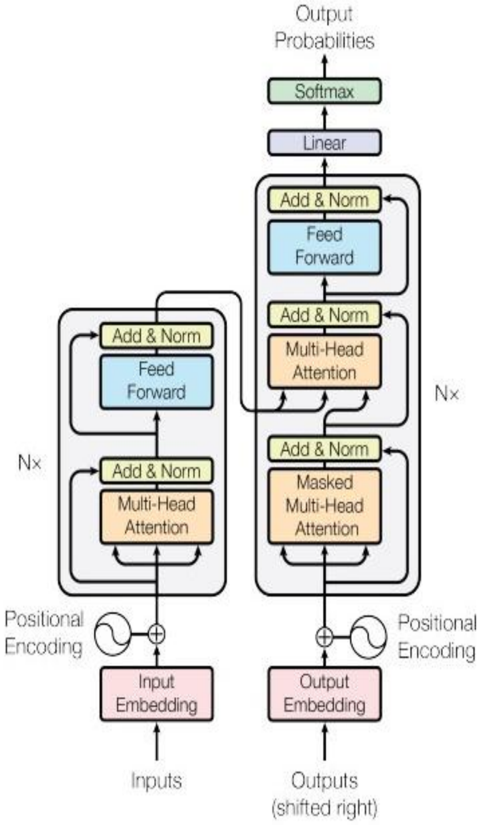

Figure 1. Transformer model architecture

3.3. Tokenization

Tokenization is the process of converting a piece of text into a sequence of “tokens” - the smallest unit in natural language processing. In BERT, the Tokenization process is done using a special encryption called WordPiece. WordPiece is a natural language encoding method in which a text string is divided into chunks and each chunk is encoded into a set of tokens. With the WordPiece method, words can be split into smaller parts and encoded into tokens.

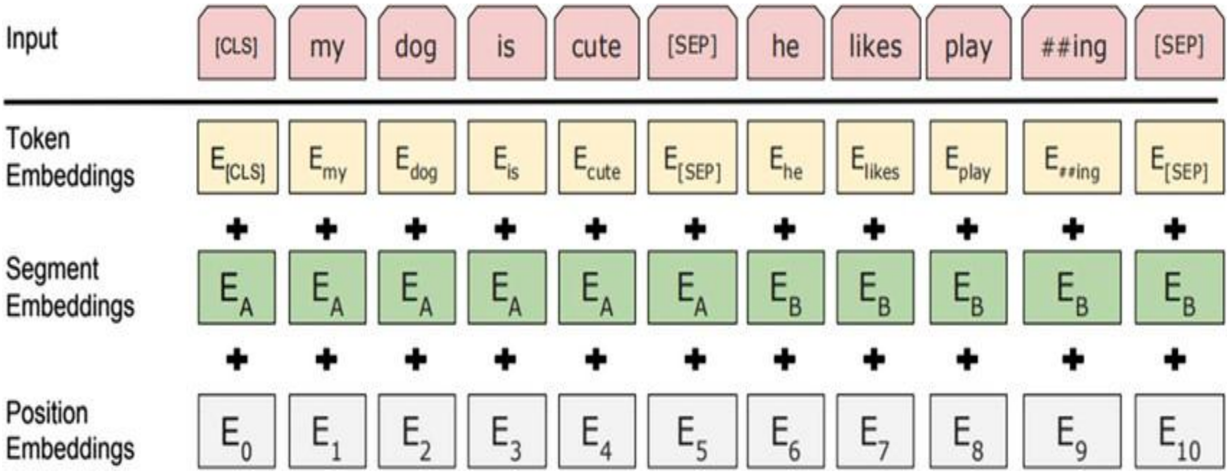

Figure 2. BERT tokenizer

3.4. Masked language modeling

Masked Language Modeling (MLM) is a method to train a natural language processing model, in which some words in the text are masked (marked as “mask”) and the model has to predict the masked word using the information about the rest of the words in the sentence.

In BERT, the MLM method is used to train the model. When training, some words in the input text are randomly selected and replaced with a special character “[MASK]”. Then, the model is trained to predict the masked word using the information about the remaining words in the text. For example, suppose there is the following sentence: “I want to [MASK] this book.” When training the BERT model with the MLM method, the word “buy” can be chosen to be masked and the sentence becomes “I want to [MASK] this book.” The model will be trained to predict the masked word “buy” using information about the rest of the words in the sentence, such as “I”, “want”, “to” and “this book”.

However, in BERT QA (BERT for Question Answering), the MLM method is used differently. Instead of masking words in a sentence, BERT QA uses MLM to generate question-answer pairs, where a word in the answer is hidden and the model has to predict that word. For example, suppose the answer is “Ha Noi is the capital of Vietnam” and the question asked to answer is “Where is the capital of Vietnam?” During the training, a word from the answer, for example such as “capital”, can be selected to be masked and the answer becomes “Ha Noi is the [MASK] of Viet Nam.” The model must predict the masked word “capital” using information about the remaining words in the answer and question.

The MLM method in BERT QA is needed to help the model understand the complex relationships between questions and answers and learn natural language features in question-answering tasks. In addition, this method also helps the model learn how to handle new or rare words in questions and answers without having to have available training data.

3.5. Next sentence prediction

Next Sentence Prediction (NSP) is a method to train a model in BERT (Bidirectional Encoder Representations from Transformers) to help the model understand the relationship between sentences in the text and identify relevant answers to the question or not. NSP works by feeding the model a mock question-answer pair and asking the model to predict whether the answer will match the question. If the answer matches the question, the model is given a label "*IsNext*" (a pair of adjacent sentences) and if it doesn't, the model is given a label of "*NotNext*" (a pair of non-adjacent sentences).

For example, give a mock question-answer pair like this: "*Ha Noi is the capital of Vietnam.*" and "*Viet Nam is located in Asia.*" An NSP model is trained to predict whether the answer ("*Viet Nam is located in Asia*") is related to the previous question ("*Viet Nam is located in Asia*").

Using NSP in training the BERT QA model is very important because it helps the model understand the grammatical structure and relationships between sentences in the text, thereby helping the model give more accurate answers.

3.6. Cosine similarity

After the model has given us the answer results, we need to choose the best option among the multiple-choice options, we use Cosine similarity to evaluate the similarity between the answers given by the model compared with multiple-choice answers. Cosine similarity is a measure of the similarity between two vectors in a multidimensional space. This measurement measures the cosine of the angle between two vectors and allows the similarity between them to be calculated based on their direction and magnitude. The cosine similarity of two vectors u and v in multidimensional space is calculated by computing the cosine of the angle between them. The formula for cosine similarity is:

$$similarity(u, v) = \frac{u.v}{|u||v|} = cos(\theta) \quad (2)$$

Where, u.v is the dot product of the two vectors u and v, ‖u‖ and ‖v‖ are the lengths of those two vectors, and theta is the angle between the two vectors.

This formula gives us a value between -1 and 1, called the cosine similarity score. The closer this value is to 1, the more similar the two vectors are, and vice versa, if the value is close to -1, the two vectors are relatively opposite. If the value is 0, then the two vectors are perpendicular to each other.

4. Results and discussion

To conduct model training, we used the SQuAD v2 (Rajpurkar, Jia, & Liang, 2018) - the Stanford Question Answering Dataset. It is a crucial resource in natural language processing research, specifically for question-answering systems - a dataset consisting of more than 100,000 question-answer pairs annotated with context snippets from Wikipedia articles. The data set is divided into a training set of more than 87,000 examples and a development set of more than 10,000 examples. Then, we performed a test on a dataset of 1,000 questions taken from online English tests, textbook exams, to measure the results and achieved accuracy that is 78% correct (Table 2). An example of test data is as follows (Table 1):

Table 1

Some samples for test data

Question	A	B	C	D	Answer
What are coral reefs made of?	Sand and shells	Calcium carbonate skeletons	Fish and crustaceans	Algae and plankton	Calcium carbonate skeletons
What is the primary role of coral reefs in coastal areas?	Providing a home for marine life	Attracting tourists for scuba diving	Protecting coastlines from erosion	Supporting fishing commu-nities	Protecting coastlines from erosion
What is one significant threat to coral reefs mentioned in the passage?	Decreased ocean salinity	Overfishing of coral polyps	Global cooling	Ocean acidification	Ocean acidification
Which statement is NOT true about coral reefs?	Coral reefs support a wide variety of marine life	They are formed by tiny marine creatures called coral polyps	Coral reefs are immune to all environmental threats	Coastal communities benefit economically from coral reefs	Coral reefs are immune to all environmental threats

We have a context like this: coral reefs are incredible underwater ecosystems known for their biodiversity and vibrant beauty. These marine environments are formed by the accumulation of calcium carbonate skeletons produced by tiny coral polyps over many years. Coral reefs support a vast array of marine life, including fish, crustaceans, and various invertebrates. Additionally, they play a crucial role in protecting coastlines from erosion and serve as a valuable source of income through tourism and fishing for many coastal communities. However, coral reefs are facing numerous threats, such as ocean acidification, pollution, and global warming, which endanger their existence.

In Table 2, we have summarized the accuracy of the model for different parameter values. We tested on five different learning rates (1e - 4, 2e - 4, 3e - 4, 4e - 4, 5e - 4), with three batch sizes (8, 16, 32) and two max sequence length values (128, 484).

Table 2

Accuracy of BERT using different parameters

Learning Rate	Batch size	Max Length	Accuracy
1e - 5	8	128	**78**
1e - 4	8	128	71.1
		484	76.73
	16	128	74.11
		484	77.54
	32	128	75.73
2e - 4	8	128	27.59
		484	23.46
	16	128	74.92

Learning Rate	Batch size	Max Length	Accuracy
		484	74.22
	32	128	71.9
3e - 4	8	128	21.3
		484	36.85
	16	128	33.13
		484	26.69
	32	128	29
4e - 4	8	128	21.3
		484	27.49
	16	128	33.83
		484	23.36
	32	128	26.38
5e - 4	8	128	20.05
		484	21.2
	16	128	19.8
		484	23.46
	32	128	27.49

In Figure 3, we have plotted the accuracy test results for different parameters when the max length is set to 128. The results show that at a learning rate of 1e - 5 and a batch size of 8, the model achieves the highest accuracy of 78%.

In Figure 4, we have plotted the accuracy test results for different parameters when the max length is set to 484. The results show that at a learning rate of 1e - 4 and a batch size of 16, the model achieves the highest accuracy of 77.54%.

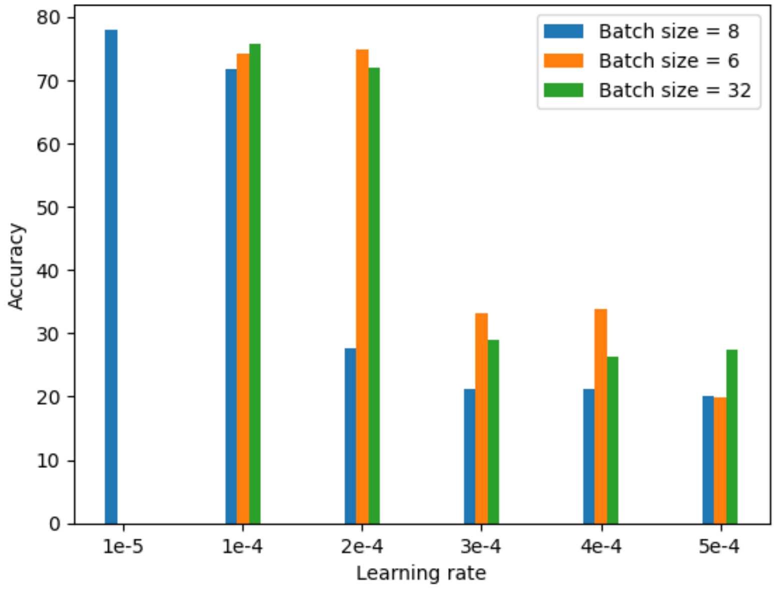

Figure 3. Accuracy of BERT (max length = 128)

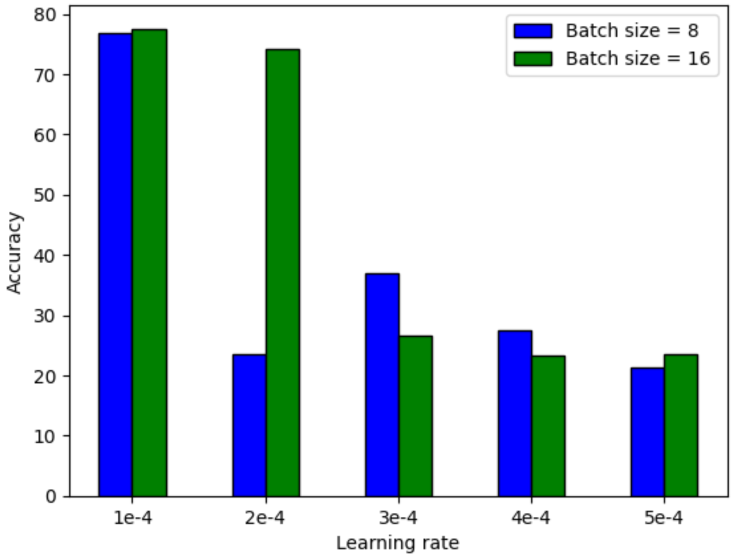

Figure 4. Accuracy of BERT (max length = 484)

With quite remarkable accuracy ≈ 78%, the model can be integrated into chatbot or virtual assistant systems to answer questions from users. This can be applied in many areas, from customer support to support in healthcare, finance, travel, and many more. In the field of education, we can create automated scoring applications, teachers can use this model to create and score multiple-choice reading comprehension tests more effectively. In the field of content summarization, this model can be used to create multiple-choice questions based on a given text, helping to summarize and extract key information. Additionally, language learners can benefit from the model by practicing comprehension skills through multiple-choice exercises and receiving immediate feedback on how well they understand the material. In a chatbot or virtual assistant, the model can help understand and respond to customer queries or complaints by extracting information from relevant documents and presenting them in the form of multiple-choice options. And many other areas can be applied.

5. Conclusions

In this paper, we developed learning models, which improve the efficiency and accuracy of QA response models. We have trained a good answering model for the question-answering task by combining many methods. First, we train a QA model that answers reading comprehension questions from a given text and questions, using the bert-base-cased pre-train model and the SQuAD v2 dataset. We trained and tested many hyperparameters to find the best model. Then, we compared the answers generated from the model with multiple-choice answers in a multiple-answer reading comprehension question using the cosine similarity method. In the future, we will try to further improve the performance of the model, as well as study and train the models for other tasks such as fill in the blanks, text synthesis problems, or translation problems, etc.

References

Bordes, A., Chopra, S., & Weston, J. (2014). *Question answering with subgraph embeddings*. Retrieved May 10, 2023 from https://arxiv.org/pdf/1406.3676.pdf

Chen, D., Bolton, J., & Manning, C. D. (2016). *A thorough examination of the CNN/Daily mail reading comprehension task.* Retrieved May 10, 2023 from https://arxiv.org/pdf/1606.02858.pdf

Chen, Q., Zhuo, Z., & Wang, W. (2019). *BERT for joint intent classification and slot filling*. Retrieved May 10, 2023 from https://arxiv.org/pdf/1902.10909.pdf

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Retrieved May 10, 2023 from https://arxiv.org/pdf/1810.04805.pdf

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. A. (2020). *Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.* Retrieved May 10, 2023 from https://arxiv.org/pdf/2002.06305.pdf

Hugging Face. (n.d.). *Sentence transformers*. Retrieved May 10, 2023 from https://huggingface.co/sentence-transformers

Liu, B., & Lane, I. (2016). *Attention-based recurrent neural network models for joint intent detection and slot filling.* Retrieved May 10, 2023 from https://arxiv.org/pdf/1609.01454.pdf

Rajpurkar, P., Jia, R., & Liang, P. (2018). *Know what you don't know: Unanswerable questions for SQuAD.* Retrieved May 10, 2023 from https://arxiv.org/pdf/1806.03822.pdf

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2023). *Attention is all you need.* Retrieved May 10, 2023 from https://arxiv.org/pdf/1706.03762.pdf

Zhang, J., Zhu, X., Chen, Q., Dai, L., Wei, S., & Jiang, H. (2017). *Exploring question understanding and adaptation in neural-network-based question answering*. Retrieved May 10, 2023 from https://arxiv.org/pdf/1703.04617.pdf