Application and Effectiveness of BERT in Question and Answer Modelling

Haofeng Weng*

Department of Computer Engineering, Fuzhou University Zhicheng College, Fujian, 350000, China

Abstract. In the field of Artificial Intelligence, chat Question and Answer (Q&A) systems represent a core application that simulates human dialogue capabilities. Given the quick development of natural language processing (NLP) technology, chat Q&A models according to Bidirectional Encoder Representations from Transformers (BERT) have emerged as a significant research focus. The BERT model, known for its deep bi-directional representations, enhances Q&A systems with a level of semantic comprehension that previous models struggled to achieve. This review aims to explore recent research progress in BERT-based chat Q&A models and analyze key issues and challenges encountered in their practical applications. This paper begins by introducing the background and principles of the BERT model, highlighting its importance in natural language processing. Then, the paper reviews in detail the key techniques and approaches of BERT-based chat Q&A models. The purpose of this thorough summary is to provide readers a clear grasp of BERT's role in chat Q&A systems, and the challenges it faces, and ultimately advance research and application development in this domain.

1 Introduction

In Artificial Intelligence, chat Question and Answer (Q&A) systems, as the core application of simulating human dialogue ability, are dedicated to understanding and accurately answering users' natural language queries. With the rapid development of Natural Language Processing technology, chat Q&A models based on Bidirectional Encoder Representations from Transformers (BERT) have emerged as a popular area of study. BERT models, with their deep bi-directional representations, provide Q&A systems with powerful semantic understanding, which is difficult to achieve in previous models.

Before the advent of BERT, research on chat Q&A systems relied heavily on traditional approaches that consisted of steps such as question parsing, information retrieval, template matching, and rule-based systems. For example, early work such as the Q&A system suggested by Liu et al. is mainly based on retrieval and template matching approaches, which are effective in specific domains but limited when dealing with complex and open-domain dialogues [1]. A key limitation of traditional approaches is that they usually lack sufficient semantic understanding, which leads to difficulties in understanding user intent and context.

* Corresponding author: yugangmaple@ldy.edu.rs

In addition, these systems often require extensive manual feature engineering, which limits their scalability and adaptability.

The emergence of the BERT model marks a breakthrough in the study of question-and-answer systems. Instead of manually designing features, BERT models can automatically learn deep features of a language compared to traditional methods. This end-to-end training approach simplifies the building process of Q&A systems and achieves unprecedented accuracy on public datasets such as SQuAD. Although BERT models show great potential in the Q&A domain, when applying them to real-world chat Q&A systems, researchers are faced with new challenges such as model interpretability, domain adaptation, and optimization issues in specific application scenarios. Therefore, an in-depth study of BERT-based chat Q&A models can not only promote the progress of dialogue system technology but also provide new ideas and methods for the innovative application of intelligent information services. Future lookup will center of attention on how to similarly enhance the efficiency and accuracy of the models, and how to better integrate these models into practical applications. This paper discusses the background and principle of the BERT model, the direction of feature representation based on the BERT model, the direction of network structure optimization based on the BERT model, the direction of relationship extraction based on the BERT model, the advantages and limitations of BERT model as well as the significance of the research and the future outlook on these points respectively. This review aims to explore the application and development of the BERT model in question-and-answer systems, with the intention of supplying valuable insights for future lookup and practical applications.

2 BERT model

BERT's ability to capture rich language representations in multi-task learning through its innovative pre-training approach has enabled BERT-based question-and-answer models to achieve significant performance gains on multiple Natural Language Processing (NLP) tasks. A major challenge that the NLP field has long faced is how to enable machines to not only understand the surface form of language but also to capture the deeper semantic and contextual relationships. Before the advent of BERT models, most language models were based on unidirectional language understanding, which restricted their ability to seize the full that means of language.

The BERT model was once proposed by Devlin et al. in 2018, which is primarily based on the Transformer architecture and revolutionizes the introduction of a deep bi-directional pre-training method. This model can take into account the contextual information before and after words simultaneously, thus achieving a qualitative leap in semantic understanding [2]. Pre-training for BERT consists of two key tasks: the Masked Language Model (MLM) and Next Sentence Prediction (NSP) [2]. The MLM forces the model by randomly masking phrases in the enter and predicting them to learn comprehensive contextual information, and NSP trains the model to understand relationships between sentences.

The emergence of the BERT model has had a profound impact on the NLP field. Its bidirectional deep learning strategy not only improves the machine's comprehensive understanding of language but also achieves breakthrough performance on a wide range of NLP tasks, such as query and reply systems, text classification, named entity recognition, etc. BERT's strong generalization ability and its ability to adapt to different downstream tasks through simple fine-tuning have led to its tremendous use in both academic research and industrial applications.

In addition, the success of BERT has also facilitated the development of a series of subsequent pre-trained language models, such as Generative Pre-trained Transformer, Robustly optimized BERT approach, etc., which have shown excellent performance in

different tasks and domains. BERT has not only become an important milestone in NLP BERT has not only become an important milestone in NLP research, but additionally gives robust technical assistance for performance improvement of purposes such as sensible assistants, search engines, and suggestion systems.

3 BERT research direction based on Q&A systems

3.1 Characterisation

The feature representation of BERT refers to the deep and enriched vector representation of text data learned by the BERT model through its pre-training process. These feature representations can seize the semantic facts and contextual relationships of words, phrases, sentences, and even entire text sequences. The feature representations of the BERT model mainly include the following aspects: dynamic sequence unfolding, contextual relevance, multilayer representations, self-attention mechanism, cross-language pre-training, and fine-tuning. Among them, one of the most representative studies and improvements of the BERT model is the introduction of cross-language pre-training and fine-tuning. This strategy enhances the model's potential to understand and represent features in exclusive languages by way of pre-training the mannequin on a multilingual corpus. Examples include the use of the es-BERT model pre-trained and tuned for the Spanish language which performs significantly better than most Spanish models in processing Spanish and the RochBert model which uses a combination of phonological and graphemic features in the fine-tuning phase to BERT generation [3, 4]. Another representative improvement is the multi-layer Transformer encoder, based on GPT and Embeddings from Language Models, the BERT model adopts a multi-layer Transformer architecture, a design that not only integrates the advantages of both but also endows BERT with superior feature recognition techniques. Meanwhile, its bidirectional structure design also ensures that the model can deeply understand the contextual environment of the text [5, 6]. In addition, knowledge-enhanced pre-training is also an important direction for BERT feature representation, which enhances the model's potential to represent domain-specific understanding by incorporating structured external from external know-how bases into the model's pre-training process. Finally, the design of self-supervised gaining knowledge of tasks is also key to enhancing BERT feature representation, e.g., by constructing more complex prediction tasks that force the model to learn deeper linguistic features, resulting in richer word vector representations.

3.2 Network structure

The network structure of BERT is based on Transformer Encoder, which mainly deals with the long-distance dependency problem through the self-attention mechanism. With the deepening of the research, the optimization of the network structure focuses on improving the efficiency and expressiveness of the model. With the deepening of the research, various methods for optimizing the network structure emerged. The network structure of the BERT model mainly includes the following aspects: the mask language model, the self-attention mechanism, the model variant RoBERTa, and the word embedding layer. Among them, the masked language model is also an important direction for BERT feature representation. In the BERT model in the MLM task, the BERT model randomly masks some words of the input text and predicts the masked words by combining the contextual information, and expresses the masked word vectors in a way more consistent with the human mind, which also enables the BERT model to continuously learn the meaning of each word in different contexts [7, 8]. Meaning in different contexts. Secondly, the optimization of the self-attention

mechanism allows the BERT model to take into account all words in the input sequence when processing each word. The self-attention mechanism is the key to its ability to understand and process complex linguistic structures, and it allows the BERT model to deal with long-distance dependencies, which enhances its sensitivity to the sequence ordering and provides the model with a powerful ability to capture and exploit the richness of information in the text information [9, 10]. In addition, model variants are designed by Roberta to optimize model performance by expanding the training dataset, not employing the next sentence prediction task, supporting longer input sequences, and implementing dynamic masking techniques. This strategy allows the model to learn in an ever-changing masking environment, which enhances the uncertainty of the data inputs and thus improves the learning efficiency of the model. Roberta is dedicated to training language models and analyzing and processing semantic connections between sentences through its three-layer structure of outputs, encodings, and inputs [11, 12]. Finally, the word embedding layer of BERT is also an integral part of the network structure; the word embedding layer consists of three parts, namely Token Embeddings, Position Embeddings, and Segment Embeddings, all of which provide the model with rich expressive language techniques [13]. Word Embeddings help BERT understand word meanings, Position Embeddings help BERT understand sentence structures, and Segment Embeddings help understand overall text semantic relationships. All of these embeddings incorporate self-attention mechanisms and feed-forward neural networks to learn deeper linguistic features.

3.3 Relationship extraction

Relational extraction by BERT refers to the relational extraction using BERT, through which the interrelationships between entities within a text are identified by this model. Relational extraction by the BERT model mainly includes the following aspects: joint entity-relationship extraction, special labeling, multi-task learning framework, data enhancement and pre-training, and fine-grained entity recognition. Among them, joint entity-relationship extraction is some models identify entities and their interrelationships in the text simultaneously by joint learning, since the traditional method of joint extraction can be achieved by serial labeling, which is highly accurate but cannot solve the entity overlapping problem, entity-relationship combining based on Span-BERT pre-training model was proposed in 2024, which is a new approach [14]. This approach is usually more effective than the serial approach (entity recognition followed by relationship classification) and effectively puts the problem of entity overlapping to rest. Secondly, the utilisation of special markers within the input sequence serves to facilitate the clear distinction and localisation of entities, thereby enabling the model to more effectively comprehend the boundaries and relationships between said entities [7]. In the BERT model, Classification token (CLS) and Separator token (SEP) markers play an important role, CLS expresses the main idea of the sentence. SEP is used to space out different sentences or sequences. The application of a multi-task learning framework to the relational extraction task enables the model to learn multiple related linguistic tasks concurrently, which enhances the model's comprehensive understanding of the text. Data augmentation and pre-training refer to pre-training using a large amount of unlabelled data, as well as expanding the training set through data augmentation techniques, this enhances the model's capacity to generalise to disparate types of relations. Finally, fine-grained entity recognition is widely applied to BERT, some models cannot correctly handle rare words due to multiple meanings of the word, and the nesting phenomenon after fine-grained classification so that the entities are cut off greatly reduces the recognition accuracy of the model, but the combination of the BERT model's ability to understand the context and constraints can improve the model's correctness for discriminating entities under the preservation of the original meaning [15].

4 Strengths and limitations of the BERT model

The application of the BERT model to question-and-answer systems demonstrates its significant advantages in understanding and processing natural language. Firstly, the BERT model learns a rich linguistic representation through Prior to training, a substantial corpus of textual data was made available for examination, which makes it excel in understanding natural language. This capability is especially important in Q&A systems, it is imperative that the system is able to accurately comprehend the user's query and subsequently retrieve the appropriate response from the knowledge base. The BERT model's bi-directional contextual understanding capability, through the bi-directional Transformer architecture, can incorporate contextual semantic information at the same time, which can be used in Q&A systems to find the key information of the inputs quickly and to make quick matching with the information in the knowledge graph. Secondly, the pre-training nature of the BERT model means that it can learn rich linguistic representations from large-scale textual data, which provides a powerful foundation for semantic understanding in Q&A systems. In addition, the BERT model can be integrated with other models, such as Bidirectional Long Short-Term Memory Conditional Random Field, to facilitate fine-grained entity recognition and attribute extraction. This is a crucial process in accurate answer retrieval in Q&A systems. This fine-grained processing capability enables the model to better understand the specific requirements of the question and retrieve more accurate answers from the knowledge base.

However, even though BERT performs well in Q&A systems, it has some limitations. First, BERT models require huge computational resources for pre-training and inference, which limits its application to some extent. Especially in resource-constrained environments or devices, the deployment and operation of BERT models may be challenging. Secondly, BERT models may require further domain adaptation training in domain-specific Q&A systems. This is because BERT models mainly rely on general-purpose textual data in the pre-training phase, whereas in a specific domain, there may be a large number of jargon and specific contexts, which require additional learning and adaptation of the model. In addition, BERT models may encounter bottlenecks when dealing with long text tasks. Since BERT models usually use a fixed-length context window, this limits the model's ability to process very long texts. When processing long text, the model may not be able to receive information about the entire sentence, this may result in the model being unable to fully comprehend the semantics of the context, which could subsequently impact the accuracy of the response.

5 Conclusions

The BERT model provides powerful semantic understanding support for natural language processing tasks, especially Q&A systems, through its bidirectional contextual understanding capability, which greatly improves the accuracy of user intent and the quality of personalized responses. This article reviews the progress of research on BERT-based chat Q&A models and analyses the advantages of BERT in dealing with complex dialogue scenarios, including its bidirectional context understanding capability and powerful pre-trained language model. The article also explores the key techniques of BERT in Q&A modeling, such as feature representation, network structure optimization, and relation extraction, and discusses the challenges encountered by the model in real-world applications, for example, there is a high consumption of computational resources., domain adaptation issues and bottlenecks in processing long texts. Although the application of BERT models in Q&A systems has achieved remarkable results, their high demand for computational resources, challenges in domain-specific adaptability, and performance limitations in processing long texts are still issues that require further investigation and analysis in order to advance current research and applications. Future research may focus on optimizing the computational efficiency of the

BERT model, improving its domain-specific adaptability and generalization, and improving the model's ability to handle long texts, to better integrate the BERT model into diverse practical applications.

References

1. P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101 (2016)
2. J. Devlin, BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. J. Cañete, G. Chaperon, R. Fuentes, et al., Spanish pre-trained BERT model and evaluation data. arXiv preprint arXiv:2308.02976 (2023)
4. Z. Zhang, J. Li, N. Shi, et al., RochBERT: Towards robust BERT fine-tuning for Chinese. arXiv preprint arXiv:2210.15944 (2022)
5. N. Li, R. Fang, Aspect-level sentiment analysis integrating multi-layer features of BERT. Computer Sci. Appl. **10**, 2147 (2020)
6. S. Yi, Research on knowledge graph question answering based on improved BERT. Computer Sci. Appl. **10**, 2361 (2020)
7. S. Li, Subword-level Chinese text classification method based on BERT. Comput. Sci. Appl. **10**, 12677 (2020)
8. N. Lin, Hierarchical multi-label text classification based on BERT. Adv. Appl. Math. **13**, 2141 (2024)
9. C. Zhao, S. Wang, D. Li, et al., Cross-domain sentiment classification via parameter transferring and attention sharing mechanism. Inf. Sci. **578**, 281-296 (2021)
10. Y.T. Peng, C.-L. Lei, Using Bidirectional Encoder Representations from Transformers (BERT) to predict criminal charges and sentences from Taiwanese court judgments. PeerJ Comput. Sci. **10**, e1841 (2024)
11. Z. Zhao, J. Shan, J. Wang, Knowledge entity recognition in high school mathematics based on RoBERTa-CNN-BiLSTM-CRF. Artif. Intell. Robotics Res. **13**, 121 (2024)
12. R. Xiang, Z. Li, P. Sun, Research on sentiment analysis of scenic area reviews based on RoBERTa-BiGRU-Attention: A case study of Shenyang. Hans J. Data Mining. **13**, 312 (2023)
13. O. Galal, A. H. Abdel-Gawad, M. Farouk, Rethinking of BERT sentence embedding for text classification. Neural Comput. Appl. 1-14 (2024)
14. J. Yu, B. Ji, H. Wu, et al., A span-based method for joint entity and relation extraction. Comput. Eng. Sci. **44**(03), 502-508 (2022)
15. C. Lothritz, K. Allix, L. Veiber, et al., Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition, in Proceedings of the 28th International Conference on Computational Linguistics, (2020)