Hashing based Answer Selection Literature Review

This paper focuses on the answer selection, a subfield of question answering. It proposes the idea of learning binary matrix representation by using hashing strategy in answer selection. With experiments, it verified the excellence of the strategy compared with other methods.

Question answering is a research direction of NLP. A typical question answering task can be defined as follows.

Definition 1. *Given a question* q *and a candidate set* $\mathrm{A} = \{a_1, a_2, .., a_n\}$ *of answers, the question answering task expects to find the best answer* a_i *or return a sorted sequence of answers in the form of* $\mathrm{A}' = [\mathrm{a}'_1, \mathrm{a}'_2, .., \mathrm{a}'_k], i, k \in \{1,2, .., n\}$.

Figure 1 depicts a typical question answering pipeline [1]. Among these four steps, answer selection is a necessary one. It is used to select the most relevant sentence to the question in the retrieved documents. However, there are several literatures regarding answer selection and extraction as the same step, or simply ignoring the answer extraction step [2]. According to the chosen paper [3], this article will ignore the last step and focus mainly on the answer selection itself, especially the efficiency issue of answer selection task.

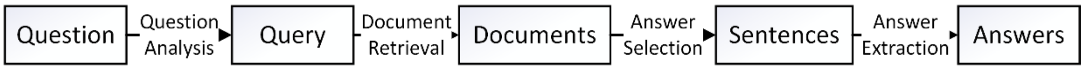

Figure 1. A typical question answering pipeline architecture

The most original question answering task is based on the information retrieval systems. After that, expert (knowledge) systems, search engines and community-based question answering came into existence [4]. However, how to efficiently select the correct answer is always one of the core issues in the field of question answering despite the changing of technology. There are shallow and deep models to tackle this problem. For shallow methods, such as bag-of-words, term frequency, manually designed rules and syntactic trees, they can't achieve high accuracy because they only utilize the surface features instead of semantic information. Therefore, people began to apply deep models into this problem. In recent years, deep models including CNN, LSTM, BERT have shown us their extraordinary talents, but they also brought us huge computation costs. Hence, people started to think about how to reduce the cost of using it while enjoying the high precision brought by the deep model. In general, this phenomenon illustrates the importance of studying the efficiency of answer selection.

Recent studies tried to solve the computational efficiency problem of answer selection task from different aspects. They can be roughly divided into two categories including data abstraction and model modification or simplification.

Methods based on data abstraction often focus on how to simplify the representation of data in different domains. For example, [5] constructed a system that is feasible to answer questions over 400 million of entities of any domain based on the features of linked data and technology of pure NLP. In addition, [6] proposes a method that can significantly reduce training time using knowledge base (KB), whereby a single triple is retrieved from a KB for a given natural language query.

For model modification, [3] tries to learn a binary matrix representation using hashing strategy for each answer, reduce the computation costs and adopt complex encoders like BERT. For model simplification, [7] summarizes three improved technologies for BERT, namely, quantization, pruning and knowledge distillation. [8] introduces factorized embedding parameterization and cross-layer parameter sharing technology. The first technology makes it easier to grow the hidden size without significantly increasing the parameter size of the vocabulary embeddings, while the second one prevents the parameter from growing with the depth of the network.

In the future, researches in the efficiency of answer selection task may still develop in these two directions (data abstraction and model modification or simplification). The former one lacks the feature extraction ability similar to deep models, so it is not as accurate as the deep models. Therefore, we may try to introduce deep models to some subtasks of data abstraction to promote its performance in accuracy. At the same time, because the latter one lacks the ability of large-scale feature storage and online data processing, it is not as efficient as the data abstraction method. Therefore, we may try to utilize abstract data such as knowledge and relationship to boost the efficiency of deep models.

Besides, it is foreseeable that these two directions may even combine with each other to be developed into a real time question answering system, which is based on deep models and applicable to deal with large-scale data. Meanwhile, there are scholars studying question answering system based on transfer learning. They have shown that through a basic transfer learning technique from SQuAD-T, the SOTA result in the WikiQA dataset can be improved [1]. This discovery demonstrates the potential of transfer learning techniques for the answer selection task.

References

[1] T. M. a. B. T. a. L. S. Lai, "A Review on Deep Learning Techniques Applied to Answer Selection," *Association for Computational Linguistics,* pp. 2132-2144, 2018.

[2] A. F. M. A. N. Jamshid Mozafari1, “BAS: An Answer Selection Method Using BERT Language Model,” Cornell University, Apr 11 2019. [联机]. Available: https://arxiv.org/ftp/arxiv/papers/1911/1911.01528.pdf. [访问日期: 23 June 2020].

[3] D. X. a. W.-J. Li, “Hashing based Answer Selection,” *Association for the Advancement of Artificial,* 2020.

[4] C. D. R. P. &. S. H. Manning, Introduction to Information Retrieval, Cambridge, UK: Cambridge University Press., 2008.

[5] K. S. M. M. Y. T. E. Dimitrakis, “Enabling Efficient Question Answering over Hundreds of Linked Datasets,,” *Information Search, Integration, and Personalization,* 9-10 May 2019.

[6] T. A. Happy Buzaaba, “A Modular Approach for Efficient Simple Question Answering Over Knowledge Base,” *Database and Expert Systems Applications,* pp. 237-246, 6 Aug 2019.

[7] S. Sucik, “Compressing BERT for faster prediction,” RASA, 8 Aug 2019. [联机]. Available: https://blog.rasa.com/compressing-bert-for-faster-prediction-2/. [访问日期: 23 June 2020].

[8] “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations,” *International Conference on Learning Representations,* 2020.