Developing QA algorithm using Machine Reading Comprehension

Wootae Jeong(wtjeong@42maru.com), Hyelim Do
42maru

Abstract

Introduction: According to Gartner, by 2020, 30% of web browsing sessions will be done without a screen by 2020. Another Internet marketing research company, ComScore, predicted that 50% of all searches will be done by voice by 2020. The search paradigm has changed with the change of the input method. In the case of the previous character-based search, the search by one or three keyword input is mainly performed. However, the search by voice is much closer to the natural language used by the person, so the search query is also longer and is closer to the question in the daily conversation ought. In addition, in the case of character-based Web search, it is necessary for the user to take additional action in order to find out the desired result among the presented candidates by listing the web documents or the like which are likely to include the result to be searched by the user in the form of a list. However, in the voice-based search, it usually presents about one to four correct answers directly to the user, taking into consideration the general human memory limit, memory span, and the like.

Machine Reading comprehension: Machine Reading Comprehension, which is becoming an important technology in the field of Question Answering, is a general term referring to a technique in which a machine uses an artificial intelligence network to derive an answer corresponding to a question in a given text. According to IDC's forecast, by 2020, more than 80% of the data will be unstructured in 2020, such as plain text documents. And these unstructured data increase by more than 60% faster than regular data. Considering the fact that more data is being generated by the Web, social networks, IOT, etc., it is reasonably expected. In order to derive the user's desired answer from unstructured data, which is generated more and more, it is more important to have the ability of reading by a machine rather than a person. To create an artificial intelligence model for Machine Reading Comprehension, a data set for learning is needed. SQuAD, provided by Stanford University, is the most representative of these data sets. In 2016, they released a version 1.1 that provided 100,000 Q&A pairs of the Wikipedia article, and in that version, there is always an answer to every question, which is a weakness that can be learned to give a similar answer even if the model is uncertain. Last year, they released version 2.0 with the addition of 50,000 unanswerable questions. The leaderboard of SQuAD is also active because SQuAD is a powerful front runner among MRC-related data sets and academics. The ability of the MRC model created using SQuAD 1.1 data surpasses human ability. In the case of the 2.0 leaderboard, the MRC model is less than a year old but has already reached a human level.

BERT: BERT means Bidirectional Encoder Representation from Transformers, a language representation model released by Google in October 2018. The BERT model basically has a pre-training task of non-geographic type for the general language and a transfer learning form consisting of an additional map-based fine-tuning process for each task. It is no exaggeration to say that it has become an inflection point in the field of NLP, and a great change occurs before and after the appearance of BERT. According to GLUE, a site that benchmarks common tasks for natural language understanding, the BERT model, as of December 2018, outperforms other models in most of the tasks. The BERT is composed of three parts. First, pre-training is conducted on the corpus in the way of Masked LM, Next Sentence Prediction to understand the whole language. It also introduces the concept of the Transformer, which performs better than the conventional Attention mechanism and solves the long-term dependency problem that arises when a general language sequence becomes long. Finally, with respect to word embedding, by using the WordPiece model, the tokens used in common are divided into subword units rather than word units. It is possible to solve the Out Of Vocabulary(OOV) problem. At the same time, in order to solve the OOV problem in general, the number of learning parameters and learning time Increase the problem, such as through the WordPiece model to prevent. BERT implements the Attention mechanism using the encoder part of the Transformer and learns by learning the contents of both sides of the sentence at the same time due to the Masked LM learning, predicting the final masked contents and reducing the loss accordingly. In the GPA model of Open AI, which learns the language model by using the decoder part of the Transformer, learning of the language is done only from left to right. In the case of the ELMo model, the LSMT is used to learn by using the sequence on the left and the sequence on the right separately. Because of these differences, the BERT model, which learns both sides of the sentence at the same

time, has a better understanding of the language. Second, the fact that the model of the BERT itself is much larger also has some effect on its high performance. In the case of the BERT model, the total number of parameters increases as the number of heads and layers of the Transformer encoder increases. In the case of Base model, it is a very large model, which is about 110 million pieces, and large models, about 340 million pieces. As the performance change depending on the number of layers and heads and the size of the hidden dimension, we can say that the performance is improved as the size increases. Therefore, it takes about two weeks to learn BASE model even TPU that performs better than GPU when pretraining from the beginning.

Conclusion: Recent empirical improvements due to transfer learning with language models have demonstrated that rich, unsupervised pre-training is an integral part of many language understanding systems. In particular, these results enable even low-resource tasks to benefit from very deep unidirectional architectures. Our major contribution is further generalizing these findings to deep bidirectional architectures, allowing the same pre-trained model to successfully tackle a broad set of NLP tasks. While the empirical results are strong, in some cases surpassing human performance, important future work is to investigate the linguistic phenomena that may or may not be captured by BERT.

Future work: The existing QA dataset is a simple form of learning where the machine imitates basic human-level reading capabilities. In the case of SQuAD, it is designed to search for answers in one paragraph, and the reasoning ability of large-scale context is poor. TriviaQA & Search QA extract the correct answer from several documents, but it is also single-hop reasoning with just the prolonged length of the document. At a stage where the ability of artificial intelligence has improved at the level of humans, more challenging goals are needed in the future. Multi-hop inference based on the single-hop method is not a method of finding an answer directly in one document, but a method of obtaining a final answer by additionally deducing the first clue found in the document. For example, if there is a question such as "What is the law school of the United States that produced the first African American president?", after first finding who is the first African American president, you can solve the problem by finding the law school in America where he graduated. It is a way to get the final result through two or more multi-hops rather than seeking the correct answer with one inference.
The services that can be provided using the QA platform based on artificial intelligence include chatbot, smartwatch, and QA service through a smart home. Furthermore, the service can be utilized outside of the home, in a hotel, accommodation, office space, and the like. In the case of B2B, considering the case of a customer center, when the agent responds to the customer on a one-to-one basis, the customer center itself may cause customer complaints due to the limited working time and long waiting time. On the other hand, if the service is provided through the AI-based QA platform, it can be supported 24 hours a day, 365 days a year. In particular, many of the questions asked by the customer center, in particular, have a lot of general inquiries that can be answered in common, and in such cases, it is more efficient to respond immediately through the QA platform. Based on the semantic QA search, if you can find the information you need accurately and quickly from the various legacy systems, documents, etc. inside the company, you can expect not only an increase in individual productivity but also an improvement in the productivity of the whole company.

Keywords: *Machine Reading Comprehension,*

References

[1] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

Biography

■ QA Engineer in 42MARU

■ 2018 the Talent Award of Korea (the Deputy Prime Minister and Minister of Education of Korea)

- in recognition for Korea's future leaders who have performed exemplary talents or outstanding meritorious service.

■ M.Sc. in Computational Science and Technology, College of Engineering, Seoul National University
■ STEM(SNU Tomorrow's Engineers Membership) honor member
■ Bachelor of Science in Energy Resource Engineering