Technology Review

Investigation of BERT applications with a focus on association rule learning

Chaochao Zhou (cz76@illinois.edu)

Introduction

Extraction of entities and relations is one of the most important but challenging tasks in text mining. Recently, a new language representation implementation called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, has been proposed [1]. The BERT model was pre-trained in two unsupervised text mining tasks, including masked language modeling and next sentence prediction, using a big text dataset - English Wikipedia (2,500M words). Further transfer learning by fine-tuning of BERT on specific applications can achieve state-of-the-art performance outperforming many other architectures. Based on the power of BERT, it can be hypothesized that the pre-trained BERT has encoded enriched semantic relationships in a large-scale general text corpus. Therefore, it would be also beneficial to extend BERT to extraction of entities and relations. In this review, relevant applications using BERT for the learning of entity semantic relationships and associations was briefly reviewed.

BERT Applications

Entities and their relations are often detected and extracted simultaneously from unstructured text to recognize the semantic relationships between entities mentioned [2]. A key technique in extraction of entities and relations is appropriate tagging. As shown in **Fig. 1**, the input sentence was tagged to represent two relations of "Country-President" and "Company-Founder", while other irrelevant words were tagged as "O" (other words). Based on the tagging scheme, the output represented by the tags can exactly correspond to each word in the input sentence and therefore enables the training of a variety of neural network models in an end-to-end manner. For example, recently, pre-trained BERT has been fine-tuned for jointly extracting entities and relations based on the tagging scheme [3]. In particular, a softmax layer was added to BERT as the output layer to predict the tags, as shown in **Fig. 2**.

In another application, BERT was fine-tuned to predict if there is a reasonable semantic relationship between two entities in a sentence [4]. In the application, a sentence template describing a relation between two entities was given, as represented by φ(X, Y). For example, a template φ(X, Y) could be "X is the capital of Y". By respectively assigning "Rome" and "Italy" to X and Y in the temple, we can be obtained an instance, i.e., "Rome is the capital of Italy", which can be considered as a positive relation. In contrast, "Rome is the capital of France" and "Trump is the capital of Obama" are not the facts, so both of them are negative relations. Therefore, the problem can be formulated as binary classification. Given different sentence templates and different pairs of entities, a BERT model can be implemented to predicted if there are positive and negative relations. Correspondingly, the output layer of the pre-trained BERT model should be modified to a classification layer with linear activation and a binary-cross entropy loss [4].

Furthermore, the exploration of the relationship between numeral and target entities has also gained increasing attention [5]. In this type of learning, studies have been attempted to predict the numeric distribution of a target entity, given a sentence where the accurate numeral value was masked [6]. The prediction of numeric values can be formulated as learning of distribution of the magnitude of a target value, using either regression or classification (e.g., the values can be mapped to uniformly spaced buckets). For example, given a sentence "The size of the dog is 80 cm", learning can be performed to predict how large a dog is, using a template "the size of the A is X cm". Similarly, the pre-trained BERT can be adapted to tackle the problem. However, for the purpose of model training, numerals should be tokenized to facilitate parsing the natural representation as Arabic numbers. To that end, every instance of a number in the training example can be represented by the scientific notation, i.e., a combination of an exponent and mantissa [6]. For example, 314.1 is represented as 3141[EXP]2, where [EXP] is a new token introduced into the vocabulary.

Conclusion

In summary, BERT was pre-trained using a large general text corpus. For specific implementation, the pre-trained BERT model can be fine-tuned with minimal task-specific modifications (e.g., adding just one additional output layer) to create state-of-the-art models. In terms of jointly learning entities and relations, BERT has proved feasible, as it encoded a tremendous number of semantic relations between entities. However, for such application, appropriate tagging and embedding schemes are still very critical to achieve high performance.

References

[1] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding 2018.

[2] Zheng S, Wang F, Bao H, Hao Y, Zhou P, Xu B. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme 2017.

[3] Qiao B, Zou Z, Huang Y, Fang K, Zhu X, Chen Y. A joint model for entity and relation extraction based on BERT. Neural Comput Appl 2022;34:3471–81. https://doi.org/10.1007/s00521-021-05815-z.

[4] Bouraoui Z, Camacho-Collados J, Schockaert S. Inducing Relational Knowledge from BERT. 2020.

[5] Yoshida M, Kita K. Mining Numbers in Text: A Survey. Information Systems - Intelligent Information Processing Systems, Natural Language Processing, Affective Computing and Artificial Intelligence, and an Attempt to Build a Conversational Nursing Robot, IntechOpen; 2021. https://doi.org/10.5772/intechopen.98540.

[6] Zhang X, Ramachandran D, Tenney I, Elazar Y, Roth D. Do Language Embeddings Capture Scales? 2020.

Figures

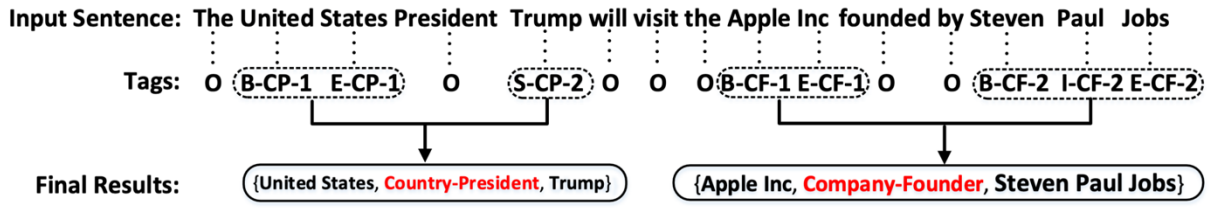

Fig. 1*: A tagging scheme, where "CP" is short for "Country-President" and "CF" is short for "Company-Founder". "1" and "2" denotes two entities in a relationship. If an entity only includes a single word, it is denoted by "S"; otherwise, it is marked by "B" = Begin, "I" = intermediate, and "E" = End. Additionally, "O" represents other words.* [2]

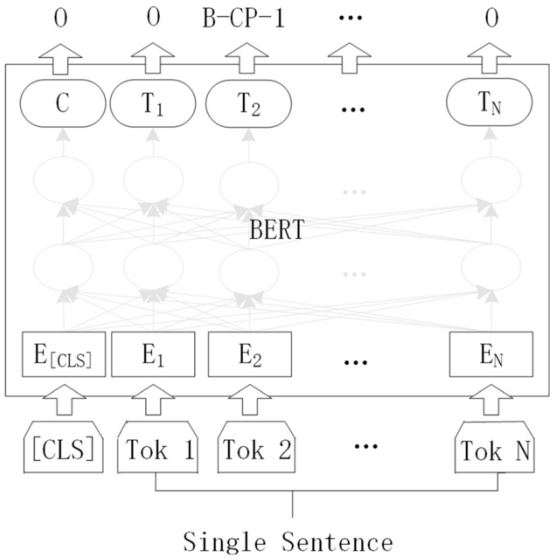

Fig. 2: *A BERT model that is used for the joint extraction of entities and relations* [3]