Overview of BERT and its Application in Address Parsing

Xing (Iris) Fu

Introduction

Address parsing might seem to be a simple task on a first look, however, addresses don’t always appear in a specific format and might not follow a specific structure. Creating a map containing all possible locations could be one solution, maintaining and updating the mapping would be labor intensive. Identifying the addresses mentioned in the text documents is critical to provide more context and insights and further enriches our ability to answer the “where” question. It can also be used to display related text information to the location the user is close to or interested in. Historically Named Entity Recognition (NER) tags were used to identify entities, but they were not able to extract the precise locations mentioned from the text document. Thus, we need to find a novel way to parse the address from the text document.

Body

Bidirectional Encoder Representations from Transformers (BERT) is a fine-tuning based presentation model that combines masked language model’s prediction of randomly masked words based only on its context and “next sentence prediction” task that jointly pretrains text-pair representations. A fine tuning based model introduces minimal task-specific parameters and trains based on fine-tuning all pretrained parameters. [1] As a pre-trained and fine-tuned general purpose language model, BERT was pre-trained using BookCorpus (800M words) and English Wikipedia (2,500M words) which provides a great coverage similar to text documents available online.

The dataset is first tagged, labeled, tokenized and transformed into I-B-O notation to train the BERT model. Then the BERT model can be used to combine tokens in the new articles to extract the location. Then with reverse geocoding, it could be stored as latitude and longitude, or used for other purposes.

Three heuristics are used to identify the sentence that could contain geolocation. Such filtering makes the labeling task cost-effective, but could decrease the overall precision of the learned result. The resulting tokens are assigned labels using

Inside-outside-beginning (I-O-B) notation to identify the beginning of the location entity and subsequently which words are related to the location entity and which are not.

After that it becomes a classification problem and with the ready made BERT implementations, precision, recall and F1 scores can be calculated by training BERT models of different sizes. It was able to achieve 70+ precision and recall and produces more accurate and relevant results than the standard NER tools.

Conclusion

BERT models are much better than traditional NER taggers to extract locations in their entirety. However, the model itself could be further fine-tuned or trained by providing more comprehensive training datasets and using more granular labels. If more context can be provided by the model output, it could also be useful for the reverse geocoding process.

References

[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[2] Gupta, Sarang, and Kumari Nishu. "Mapping Local News Coverage: Precise location extraction in textual news content using fine-tuned BERT based language model." *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*. 2020.