Long-tail dataset entity recognition based on Data Augmentation

Qikai Liu[1], Pengcheng Li[1],Wei Lu[1],Qikai Cheng[1]
[1] School of Information Management, Wuhan University, Wuhan, China
liuqikai67k@126.com,fantasticlpc@whu.edu.cn,weilu@whu.edu.cn,chengqikai0806@163.com

ABSTRACT

Datasets play an important role in data-driven scientific research. With the development of scientific research, more and more new datasets have been constructed. It is important to recognize dataset entities correctly, especially when it comes to unusual long-tail dataset entities. However, it is very difficult to obtain high quality training corpus in named entity recognition. We obtained our data based on a distant supervision method along with two data augmentation methods. We then use a BERT-BiLSTM-CRF model to predict long-tail dataset entity. By applying data augmentation methods, we achieve a highest F1-score of 0.7471.

KEYWORDS

entity recognition, long-tail entity, data augmentation, BERT

1 INTRODUCTION

With the development and rise of artificial intelligence, data-driven research has become a new paradigm. Therefore, the value of datasets has been paid more and more attention. Commonly used datasets are, for example, Wordnet, DBpedia, MovieLens, etc. Currently, dataset entity recognition research is still in the exploratory stage, especially when it comes to long-tail entity recognition. Long-tail entities are entities that have a low frequency in the document collections and usually have no reference in existing Knowledge Bases[1] . Current state-of-art NER models or tools usually perform badly on long-tail entities of specific domains. Therefore it is important to build models for long-tail entity recognition. However, high quality training data is extremely important yet hard to build for domain-specific named entity recognition task. Data augmentation is usually used in computer vision field, but more researchers are trying to apply data augmentation in NLP filed. Wei J W et al.[2] proposed EDA(easy data augmentation) techniques for boosting performance on text classification tasks.

In this paper, we propose a dataset long-tail entity recognition model based on distant supervision method along with two data enhancement ways to expand the training corpus. We finally achieve a best F1-score of 0.7471 using data augmentation methods.

2 RELATED WORK

There are many researchers focusing on domain entity recognition. Aguilar G et al.[3] proposed a multi-task approach to recognize named entity in social media data. Their approach obtained an F1-score of 0.4186 in WNUT-2017. As for dataset entity recognition, Duck G et al.[4] explored recognition of database and software entities and compared dictionary and machine learning approaches to each identification. Their machine learning approach achieved an F1-score of 0.63. Long-tail entities are named entities that are rare, often relevant only in specific knowledge domains, yet important for retrieval and exploration purposes. Mesbah S et al.[5] presented an iterative approach for training NER classifiers in scientific publications, focusing on the long-tail entities types Datasets, Methods in computer science publications, and Proteins in biomedical publications.

3 METHODS

3.1 Model for Named Entity Recognition

In this paper, BERT + BiLSTM + CRF model structure is adopted, as shown in figure 1. The training corpus is firstly pre-trained by BERT layer to get the word embedding vectors. Word embedding vectors trained by BERT model contains context, syntax and semantic information, carried by a dynamic vector. In different contexts, the embedding for the same word may be different, which is capable of carrying the context information of sentence. Then the embedding vectors are fed into the BiLSTM layer. We use vanilla LSTM and two way BiLSTM. Finally, the CRF layer takes the BiLSTM outputs and decode them into labels we need for final annotation. Unlike softmax, CRF can capture context tagging information and improve the final annotation performance.

During training, we feed the ground truth annotations to decoder and minimize the loss in sequence level. We use adam optimizer and the learning rate is scheduled using warming-up mechanism following Vaswani et al. [6].

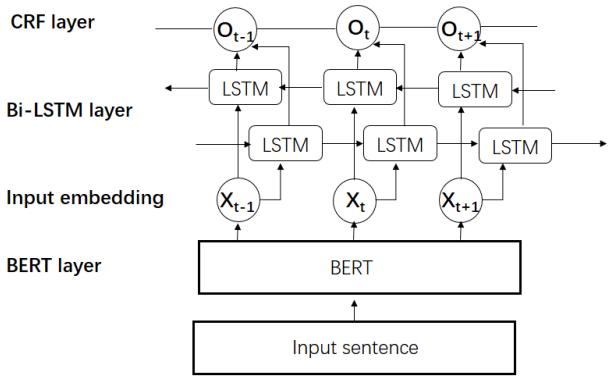

Figure 1: Dataset long-tail entity recognition model

3.2 Data Augmentation

The size and diversity of training data can make a huge difference when training a deep learning model. Therefore, to improve

long-tail dataset entity recognition performance, we not only used BERT model to capture potential semantic features of sentences but also adopted two data augmentation methods in NLP field: entity replacement and entity mask.

Entity replacement (see figure 2): the entity words in the original training corpus are randomly replaced with entities that in our dataset entity dictionary but did not appear in the training corpus.

Entity mask (see figure 3): the entity words in the original training corpus are replaced with "unknown words" which are generated randomly.

We randomly take 20%, 50% and 100% of our original training corpus and applied the above two data augmentation methods respectively. Through data augmentation process, see figure 4, we finally obtained six more augmented training corpus to train our model.

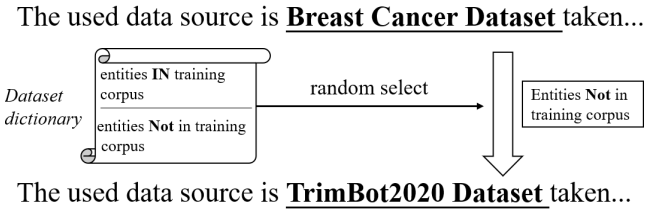

Figure 2: Entity replacement

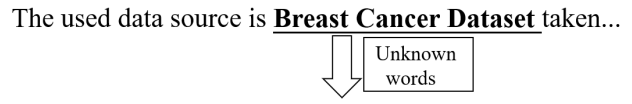

Figure 3: Entity mask

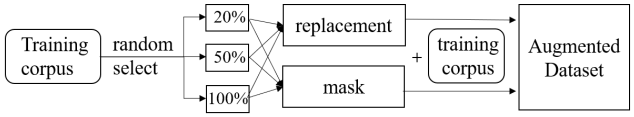

Figure 4: Data augmentation process

4 EXPERIMENTS AND DISCUSSION

We collected full-text academic papers from ACL and ACM. We then parsed all the PDF files into XML files using tools developed in our own lab. We used NLTK to segment sentences and 10747988 sentences is obtained. And we created a dataset entity dictionary of 10873 dataset entity words by crawling commonly used datasets in the computer science field from Kaggle and other websites. By applying a distant supervision matching method, we obtained a training dataset with 70313 annotated sentences. We also had human annotators labeled 200 sentences. The dataset entity mentions in the 200 sentences are those that are infrequent and never appear in our 10873 dataset entity dictionary, therefore it is reasonable to regard these entities as long-tail entity.

Table 1: Long-tail entity recognition results

Training data	Precision	Recall	F1
Original	0.7201	0.6293	0.6716
Original+Entity replacement(20%)	0.8167	0.6853	0.7452
Original+Entity replacement(50%)	0.8049	**0.6923**	0.7444
Original+Entity replacement(100%)	0.8156	0.6503	0.7237
Original+Entity mask(20%)	**0.8421**	0.6713	**0.7471**
Original+Entity mask(50%)	0.7385	0.6713	0.7033
Original+Entity mask(100%)	0.7209	0.6503	0.6838

We conducted total seven experiments including one original corpus training and six data augmentation experiments, see table 1. The experimental results show that the prediction results of the model are greatly improved by adopting data augmentation methods on six experiments. The best F1-score 0.7471 is obtained by using entity mask(20%) and two entity replacement(20%, 50%)also achieve F1-scores above 0.74.

The results demonstrate the usefulness of data augmentation methods on long-tail entity recognition. It also shows that data augmentation doesn't always work better when we augment more data. We achieved best results on both methods by augmenting only 20% of the training corpus. Therefore, we suggest conducting more experiments when applying data augmentation methods as it may have different results when it comes to different tasks.

5 CONCLUSION

In this paper, a distant supervision method is used to obtain a large number of training data, and data augmentation is used to expand the training data. The model performance in long tail entity recognition is considerably improved by adopting data augmentation mechanism, which has great theoretical and practical value. Data augmentation shows great potential in improving the performance of long tail entity recognition. In the future work, we plan to conduct more experiments and employ methods such as active learning to further improve long-tail dataset entity recognition performance.

REFERENCES

[1] José Esquivel, Dyaa Albakour, Miguel Martinez, David Corney, and Samir Moussa. On the long-tail entities in news. In *European Conference on Information Retrieval*, 2017.

[2] Jason W Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 2019.

[3] Gustavo Aguilar, Suraj Maharjan, Adrián Pastor López-Monroy, and Thamar Solorio. A multi-task approach for named entity recognition in social media data. *arXiv preprint arXiv:1906.04135*, 2019.

[4] Geraint Duck, Aleksandar Kovacevic, David L Robertson, Robert Stevens, and Goran Nenadic. Ambiguity and variability of database and software names in bioinformatics. *Journal of biomedical semantics*, 6(1):29, 2015.

[5] Sepideh Mesbah, Christoph Lofi, Manuel Valle Torre, Alessandro Bozzon, and Geert-Jan Houben. Tse-ner: An iterative approach for long-tail entity extraction in scientific publications. In *International Semantic Web Conference*, pages 127–143. Springer, 2018.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.