Long-tail dataset entity recognition based on Data Augmentation

Qikai Liu, Pengcheng Li,Wei Lu,Qikai Cheng
School of information management, WHU

Index

- Introduction
- Model
- Training corpus construction
- Data Augmentation
- Results
- Conclusion
- Limitation

Introduction

Our work focused on a domain-specific named entity recognition task, that is, dataset e
recognition. More specifically, long-tail dataset entity recognition.

Datasets play an important role in today's scientific research. Good datasets can improve experimental results. Commonly used datasets in CS field are, for example, Wordnet, DBpedia, MovieLens, etc.

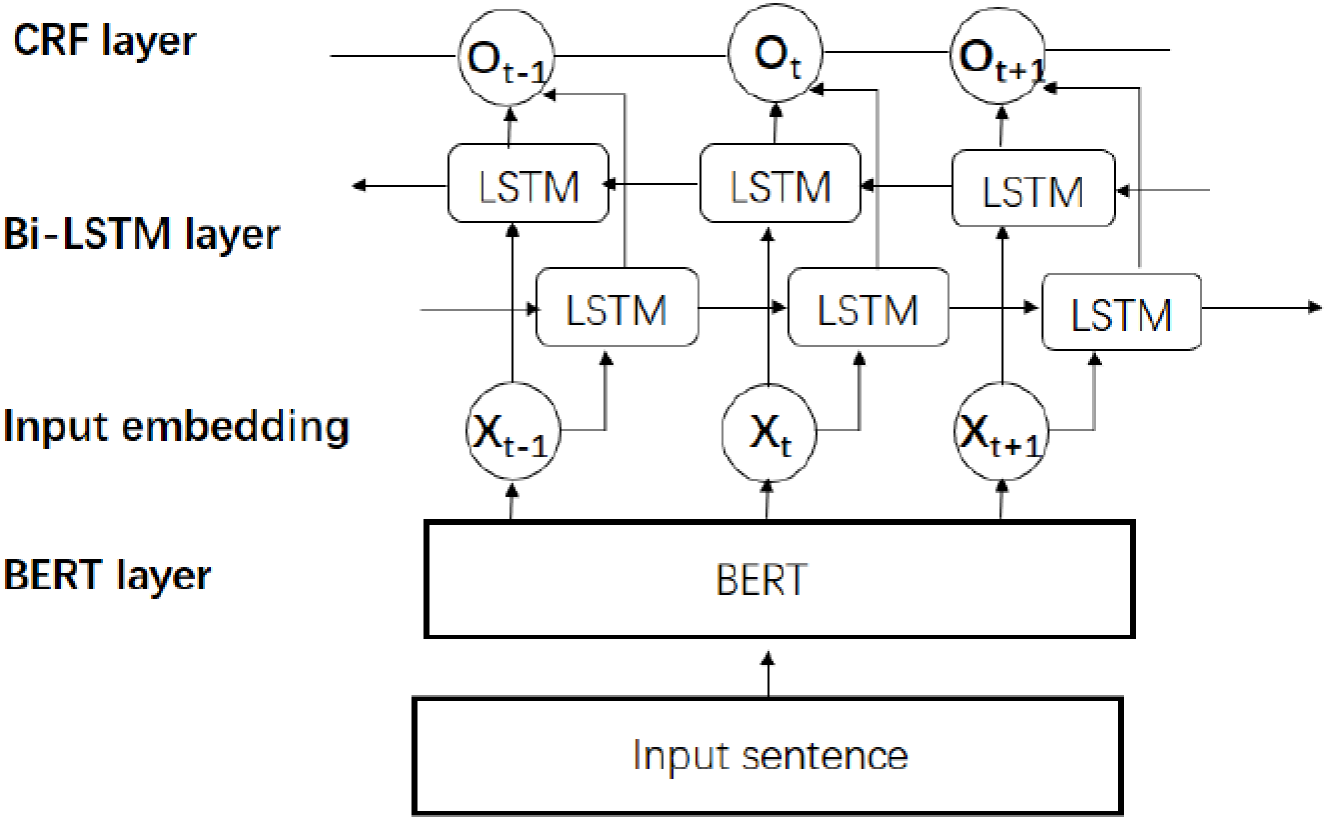
Longtail entities are entities that have a **low frequency** in the document collections and usually have no reference in existing Knowledge Bases, like TrimBot2020 Dataset.
Long-tail entities are important for retrieval and exploration purposes.

Introduction

High quality training data is extremely important yet hard to build for domain-specific named entity recognition task. Our work focused on dataset entity **training data construction**.

We proposed a dataset long-tail entity recognition model based on distant supervision method along with two data augmentation ways to expand the training corpus.

Model



CRF can capture context tagging information and improve the final annotation performance

The context information of sentence

Word embedding vectors trained by BERT model contains context, syntax and semantic information, carried by a dynamic vector

Training corpus construction

Raw sentences : we collected full-text academic papers from ACL and ACM websites and used NLTK to segment sentences, 10747988 sentences is obtained.

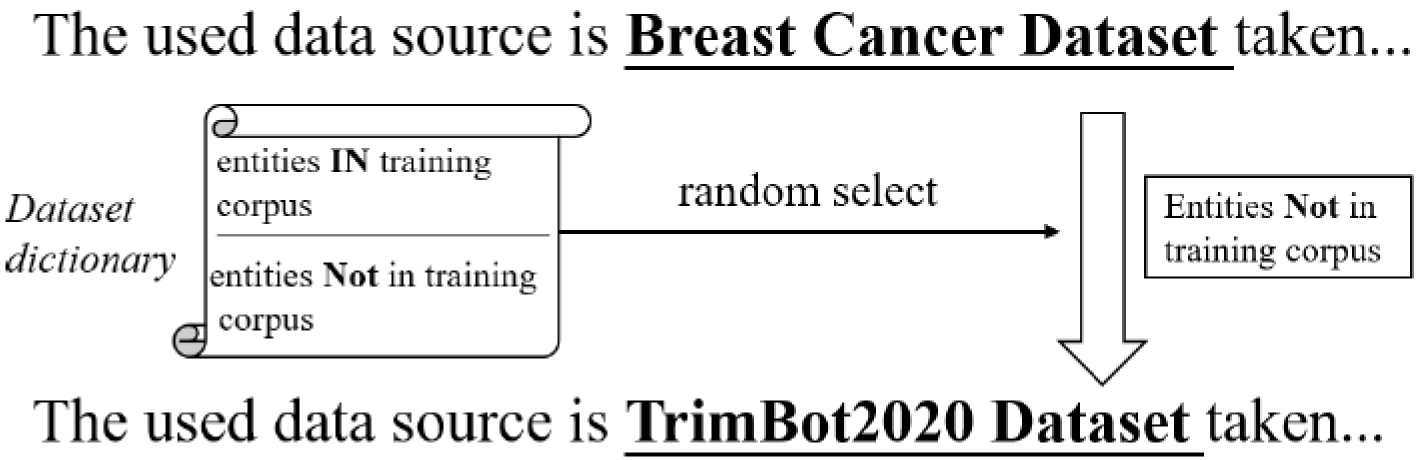
Dictionary construction : we created a dataset entity dictionary of 10873 dataset entity words by crawling commonly used datasets from Kaggle and other websites

Original training corpus : By applying a distant supervision matching method, we obtained a training corpus with 70313 annotated sentences.

Testing corpus: there is no gold standard testing corpus for our work, so we 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.

Data Augmentation

Entity replacement : 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.



The used data source is **Breast Cancer Dataset** taken...

Dataset dictionary

entities **IN** training corpus

entities **Not** in training corpus

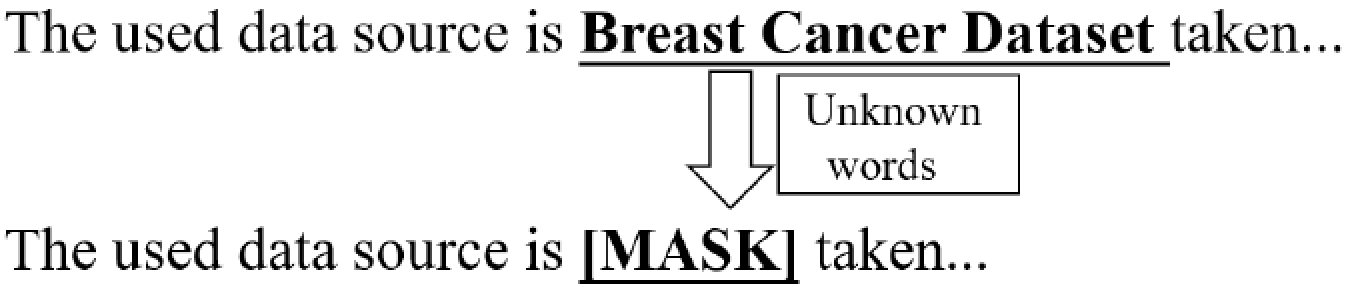
random select

Entities **Not** in training corpus

The used data source is **TrimBot2020 Dataset** taken...

Data Augmentation

Entity mask: the entity words in the original training corpus are replaced with “unknown words” which are generated randomly.

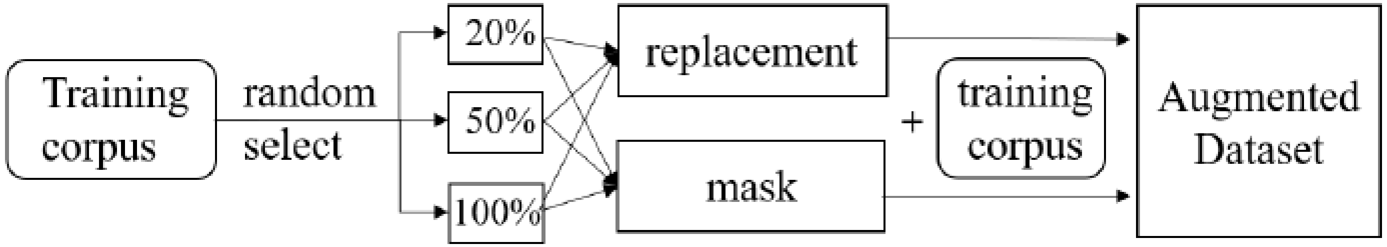


The used data source is **Breast Cancer Dataset** taken...

The used data source is **[MASK]** taken...

Data Augmentation

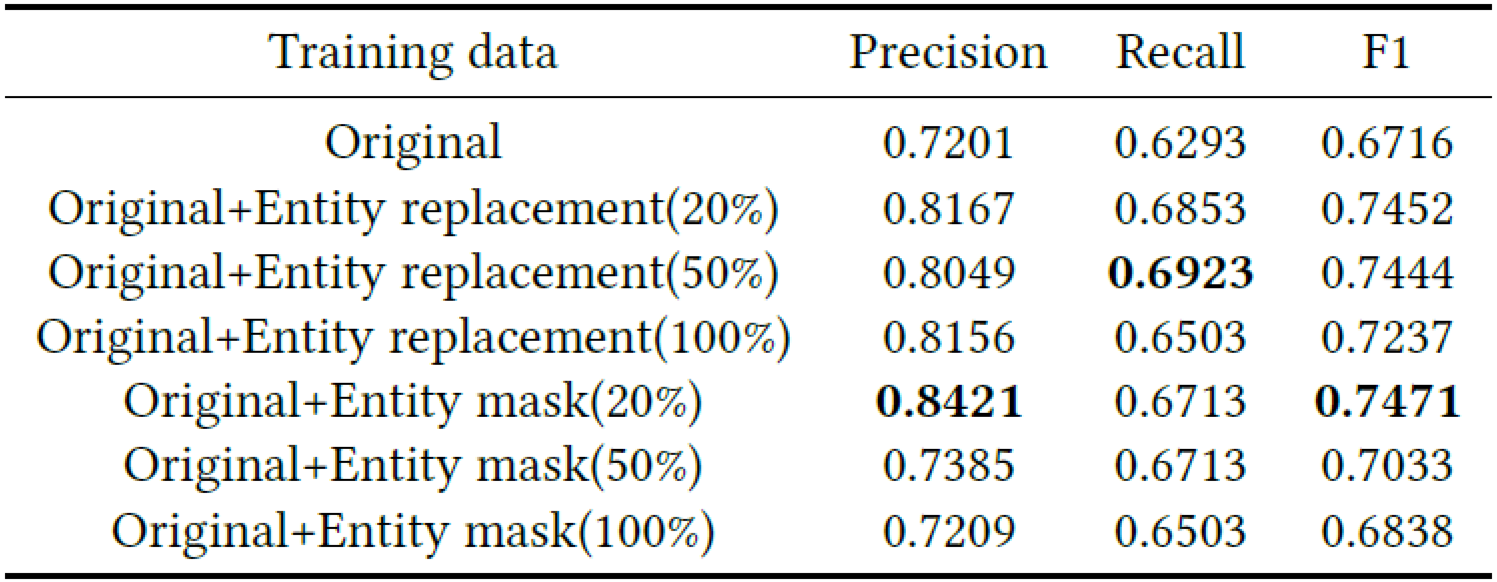
We randomly take 20%, 50% and 100% of our original training corpus and applied the above two data augmentation methods respectively



Through data augmentation, we got 6 more training corpus.

Results

We conducted total seven experiments. 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.



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

Conclusion

We use a distant supervision method 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. Data augmentation shows great potential in improving the performance of long tail entity recognition.

Limitation

- We didn’t compare our results with the state-of-art models
- More experiments can be conducted to find the most effective data augmentation amount
- Our testing corpus is small and may not be very representative

Thanks for watching