Fine-tune Bert for DocRED with Two-step Process

Hong Wang[†], **Christfried Focke**[‡], **Rob Sylvester**[‡], **Nilesh Mishra**[‡], **William Wang**[†]
[†] University of California, Santa Barbara
[‡] LogMeIn
{hongwang600, william}@cs.ucsb.edu,
{Christfried.Focke, Rob.Sylvester, Nilesh.Mishra}@logmein.com

Abstract

Modelling relations between multiple entities has attracted increasing attention recently, and a new dataset called DocRED has been collected in order to accelerate the research on the document-level relation extraction. Current baselines for this task uses BiLSTM to encode the whole document and are trained from scratch. We argue that such simple baselines are not strong enough to model to complex interaction between entities. In this paper, we further apply a pre-trained language model (BERT) to provide a stronger baseline for this task. We also find that solving this task in phases can further improve the performance. The first step is to predict whether or not two entities have a relation, the second step is to predict the specific relation [1].

1 Introduction

The task of relation extraction aims to automatically identify relationships between entities. It has been proven to be essential for many downstream applications such as question answering (Yih et al., 2015; Yu et al., 2017). Previous research (Socher et al., 2012; Zeng et al., 2014a, 2015; dos Santos et al., 2015; Xiao and Liu, 2016; Cai et al., 2016; Lin et al., 2016; Wu et al., 2017; Qin et al., 2018; Han et al., 2018; Wang et al., 2019) on relation extraction mainly focuses on sentence-level, i.e., predicting the relation for entities in a given sentence. Recently the large-scale document-level relation extraction dataset DocRED (Yao et al., 2019) was published, which requires the model to predict a relation for every pair of entities in a document. This setting is more challenging, since a large number of relational facts are expressed across multiple sentences, and modeling of complex interactions between entities is required.

In (Yao et al., 2019) several baselines for the document-level Relation Extraction (RE) task are presented. The best model uses a BiLSTM (Hochreiter and Schmidhuber, 1997) to encode the whole document, entities are represented by their average word embedding. A BiLinear layer is then applied to predict the relation for a given entity pair. However, we argue that a pre-trained language model, such as BERT (Devlin et al., 2019), can provide a further boost in performance, since it already captures important language features and may capture some common sense knowledge.

In this paper, we use BERT to encode the document. A BiLinear layer is applied to predict the relation between entity pairs. We fine-tune the whole model using annotated data in the DocRED dataset, which increases the F1 score by about 2%. We also found that modeling the document-level relation extraction through a two-step process can further improve the performance. The first step is to predict whether a pair of entities has a relation or not. The second step is to predict the specific relation for a given entity pair. Note that the model we use in the second step is trained with pairs that have relations annotated in DocRED.

2 Model

In this section, we will first introduce the BERT model for document-level RE, then we will explain how to use the two-steps training process to further improve the performance.

2.1 BERT Model

Let $[x_1, x_2, \cdots, x_n]$ denote the document input, and $[e_1, e_2, \cdots, e_m]$ denote the m entities in the document. We use BERT to encode the document

[1] Code can be found in https://github.com/hongwang600/DocRed

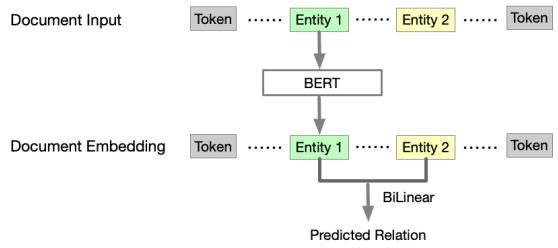

Figure 1: BERT model for DocRED.

as follows:

$$[h_1, h_2, \cdots, h_n] = \text{BERT}([x_1, x_2, \cdots, x_n]),$$

from which we can get the embeddings $[h_{e_1}, h_{e_2}, \cdots, h_{e_m}]$. Then for each pair of entities (e_i, e_j), we use BiLinear layer to predict its relation:

$$r_{i,j} = \text{BiLinear}(h_{e_i}, h_{e_j}).$$

The whole model structure is represented in Figure 1. We use BERT-base in our experiments.

2.2 Two-step Training Process

In the DocRED dataset, there is no relation for most entity pairs, which causes a large label imbalance, i.e., most entity pairs belong to the N/A relation. To alleviate this problem, we use a two-step training process.

In the first step, we only identify whether or not there exists a relation between a given entity pair, i.e., we simplify the problem to a binary classification problem. We use BERT for this step as mentioned above, where all of the annotated data is used to train the model. Sub-sampling is applied to balance relational and N/A pairs in each batch.

In the second step, we learn a model to identify the specific relationship between a given pair of entities. The model structure is the same BERT model as in the first step. The difference lies in the training data and labels: We only use these relational facts (i.e., entity pairs with relations) to train the model, so that the model can learn to distinguish between these different relations. Empirically we found that the second step is relatively easy, as we achieve about 90% accuracy. The bottleneck of the problem lies in the first step, which is to distinguish whether there is a relation or not.

After the two-step training, the testing process is straight forward. For a given pair of entities, the model from the first step is first applied to predict whether there is a relation between them. If it predicts a relation, then the model from the second step is applied to predict a specific relation.

Setting	# Doc	# Rel	# Inst	#Fact
Train	3,053	96	38,269	34,715
Dev	1,000	96	12,332	11,790
Test	1,000	96	12,842	12,101

Table 1: Statistics of the dataset. # Doc, # Rel, # Inst, #Fact denote the number of documents, the number of relations, the number of relation instances and the number of relational facts respectively.

3 Experiments

In this section, we will introduce the DocRED dataset, our implementation details of the BERT model, and the experimental results.

3.1 DocRED Dataset

The DocRED dataset is collected through distant supervision (Mintz et al., 2009) on Wikipedia documents and Wikidata. The named entity recognition is performed first on each document. Then the identified entities are linked to Wikidata items, and entities with same Knowledge Base (KB) ID are merged. Finally, the relations for entity pairs are obtained by querying Wikidata. Further processing, like named entity and coreference annotation, entity linking, and relation and supporting evidence collection, is conducted based on the collected distantly supervised data. We refer the readers to the original paper (Yao et al., 2019) for more details.

The DocRED dataset has wide coverage over a variety of topics. The entity types include person, location, organization, time, number and miscellaneous entity names. The relation types include science, art, time, personal life, etc. In order to be successful on this dataset, a variety of reasoning types are required, including pattern recognition, logical reasoning, coreference reasoning, and common-sense reasoning. The DocRED dataset provides both annotated training data (sampled from collected distantly supervised and humanly labeled data) and distantly supervised data. In our experiments, we only use the annotated data. Statistics about the dataset are listed in Table 1.

Model	Dev	Test
CNN	43.45	42.26
LSTM	50.68	50.07
BiLSTM	50.94	51.06
Context-Aware	51.09	50.70
BERT	54.16	53.20
BERT-Two-Step	**54.42**	**53.92**

Table 2: Comparison of the BERT model with other baselines. We report F_1 score on the Dev and Test set.

3.2 Implementation Details

We use BERT-base in our experiments. The learning rate is set to 10^{-5}. The embedding size of BERT model is 768. A transformation layer is used to project the BERT embedding into a low-dimensional space of size 128. In the low-dimension space, a BiLinear layer is applied to predict the relation for a given entity pair.

In the first step, we set the relation label for all relational instances to be 1, while the label for all N/A relations to be 0. We randomly sample N/A relations at a ratio $3:1$ within a batch. In the second step, we train a new model using only relational instances, and the specific relation label is kept in this step.

3.3 Results

We compare the BERT model with several baselines presented in (Yao et al., 2019) including a CNN (Zeng et al., 2014b), LSTM (Hochreiter and Schmidhuber, 1997), bidirectional LSTM (BiLSTM) (Cai et al., 2016) and Context-Aware models. The first three models differ from the BERT model in the encoder, i.e., they use CNN, LSTM, and BiLSTM as encoder respectively. Details about Context-Aware model can be found in (Sorokin and Gurevych, 2017).

The main results are presented in Table 2. We can see that we obtain a 2% F_1 improvement by using the BERT encoder, which indicates that it may contain useful information such as commonsense knowledge in order to solve this task. By using the two-step training process, performance is improved further improve. In our experiments, we find that the accuracy for the second step is above 90%, which means the bottleneck lies in the first step, e.g., predicting whether a relation exists for a given entity pair.

Model	F1	AUC
BiLSTM	50.94	50.26
SentModel	50.97	49.31

Table 3: Comparison of the BiLSTM baseline with SentModel which encode the document sentence by sentence. We report the F1 score and AUC on the Dev set here.

3.4 Complex interaction modeling

To test whether current model can capture the complex interaction between entities, we use a SentModel which encodes the document sentence by sentence. Then we locate each entity within a specific sentence and compute its embedding by averaging the word embedding of the entity name. In this way, there will be no interaction between sentences since we encode the whole document sentence by sentence. We present the results in Table 3. Surprisingly, the SentModel can achieve very similar performance compared to the BiLSTM model which encodes the whole document as a sequence. Therefore, the current model fails to capture complex interactions among entities, and only local information around each entity is used to predict a relation.

4 Conclusion & Discussion

In this paper, we investigate the usage of BERT for document-level RE. We find that BERT can improve the performance significantly, which we think may benefit from the common sense knowledge learned during pre-training. We also find that using a two-step training process can further improve the performance. The difficulty of this dataset is to distinguish whether there exists a relation between a pair of entities, while identifying a specific relation seems to be less challenging. Another discovery is that current models fail to model complex interaction between entities, which we think is the key to solve the problem of document-level RE.

References

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019*, pages 4171–4186. Association for Computational Linguistics.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2236–2245. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 1003–1011. The Association for Computer Linguistics.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2137–2147. Association for Computational Linguistics.

Cícero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 626–634. The Association for Computer Linguistics.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 1201–1211. ACL.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1784–1789. Association for Computational Linguistics.

Haoyu Wang, Ming Tan, Mo Yu, Shiyu Chang, Dakuo Wang, Kun Xu, Xiaoxiao Guo, and Saloni Potdar. 2019. Extracting multiple-relations in one-pass with pre-trained transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1371–1377, Florence, Italy. Association for Computational Linguistics.

Yi Wu, David Bamman, and Stuart J. Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1778–1783. Association for Computational Linguistics.

Minguang Xiao and Cong Liu. 2016. Semantic relation classification via hierarchical recurrent neural network with attention. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1254–1263. ACL.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of ACL 2019*.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1321–1331. Association for Computational Linguistics.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cícero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 571–581.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via

piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1753–1762. The Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014a. Relation classification via convolutional deep neural network. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2335–2344. ACL.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014b. Relation classification via convolutional deep neural network. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2335–2344. ACL.