Incorporating Knowledge Bases into SciBERT and BioBERT pre-trained language models

Abdullah Kiwan Sven Giesselbach Stefan Rüping

Fraunhofer IAIS, Sankt Augustin, Germany

June 19, 2020

1 Introduction

Recent progress in NLP has been driven by the adoption of deep neural models, but training such models often requires large amounts of labeled data. In scientific and biomedical domains, annotated data is difficult and expensive to collect due to the expertise required for quality annotation. Therefore, SciBERT [1] and BioBERT [2] were released. Pre-trained language models based on BERT, trained on a large corpus of scientific and biomedical text respectively, which improved performance on a range of NLP tasks in the scientific and biomedical domains. However SciBERT and BioBERT rely on textual information only, without incorporating knowledge which is stored in structured resources. In this research, we will incorporate knowledge bases into the SciBERT and BioBERT pretrained models using KnowBert [3], which is method to embed multiple knowledge bases (KBs) into large scale models, and fine-tune the models on relation classification tasks for both domains.

2 Method

We are going to train two different models, SciKnowBERT for the computer science domain and BioKnowBERT for the biomedical domain. Each model requires a corpus, an ontology with a graph embedding, a list of the entities extracted from the corpus and their linking to the ontology entities.

2.1 Data Preparation

SciKnowBERT is trained with the computer science papers of SemanticScholar as a corpus, and the Computer Science Ontology (CSO), while BioKnowBERT is trained using PubMed abstracts as a corpus, and the Unified Medical Language System (UMLS).

2.2 Entity Linking

KnowBert gives the possibility to train the entity linker either independently from the language model using annotated data, or jointly with the language model using a small amount of entity linking supervision in a form of a candidate entity list (self-supervision).

To the best of our knowledge, no labelled data for computer science entity linking exists. Therefore, we train the entity linker in a self-supervision way. The candidate entity list is generated using the levenshtein distance algorithm between the SemanticScholar entities (that are provided in the corpus), and the CSO entities. An empirical evaluation showed that the levenshtein threshold, and the coverage of the candidate entity list have a big influence on the model performance. We included this in our experiments.

For the biomedical data, we use MetaMap to extract entities from PubMed abstracts, and link them with the correct UMLS entities.

2.3 Training

For SciKnowBERT where no entity linking (EL) supervision is available, the EL with self-supervision on the candidate entity list is jointly trained with the language model.

For BioKnowBERT where biomedical EL supervision is available, we pretrain the KB specific EL parameters while freezing the remainder of the network, then train the network again initialized with the trained EL parameters.

3 Experiments

Our models are evaluated on several relation classification tasks. SciKnowBERT is fine-tuned on the SemEval 2018 dataset, and compared against SciBERT, whereas BioKnowBERT is fine-tuned on the following datasets: DDI extraction 2013 corpus - ChemProt - GAD - EU-ADR - Appen, and compared against BioBERT. Our experiments showed that SciKnowBERT performed better than SciBERT. On the SemEval 2018 dataset the F1-score is approximately 3% higher. The experiments on BioKnowBERT are still ongoing.

4 Discussion

Incorporating knowledge bases into pre-trained models could provide rich structured knowledge facts for better language understanding, especially in scientific and biomedical domains which contain a lot of structured knowledge in the form of ontologies or knowledge graphs. Future work will involve fine-tuning the pre-trained models on many other language understanding tasks.

5 Acknowledgment

This research has been funded by the Federal Ministry of Education and Research of Germany as part of the competence center for machine learning ML2R (01IS18038B).

References

[1] Arman Cohan Iz Beltagy, Kyle Lo. *SciBERT: A Pretrained Language Model for Scientific Text.* 2019.

[2] Sungdong Kim Donghyeon Kim Sunkyu Kim Chan Ho So Jaewoo Kang Jinhyuk Lee, Wonjin Yoon. *BioBERT: a pre-trained biomedical language representation model for biomedical text mining.* 2019.

[3] Robert L. Logan IV Roy Schwartz Vidur Joshi Sameer Singh Noah A. Smith Matthew E. Peters, Mark Neumann. *Knowledge Enhanced Contextual Word Representations.* 2019.