Abstract

Recent developments in pre-trained language models advanced the field of natural language processing (NLP). The introduction of Bidirectional Encoders for Transformers (BERT) has had important impact and increased the relevance of pre-trained models. At the beginning, research in this field was started on English text data. This was followed by multilingual text corpora. Despite of this, there has been a lack of domain-specific language models, that could extract information from texts which belong to a specific field. Some examples of such language models are SciBERT, BioBERT and MatSciBERT. The use of these language models however yields subopotimal results when used in the physics and computer science domain. This work presents new BERT language models, that are pre-trained with text data from the physics and computer science domain obtained from the open-access repository arXiv. These models are then evaluated based on their performance in named entity recognition (NER) as a downstream task. We show that the models pre-trained in this work achieved lower pseudo-perplexities than their original counterparts. Additionally, we show that the models pre-trained in this work improved the micro F1 scores of the original models on the computer science and physics named entity recognition datasets by up to 0.69% and 3.85%, respectively. The addition of a Conditional Random Fields (CRF) layer however did not improve the performances of the models on the named entity recognition tasks in this work. However, the models pre-trained in this work still achieved higher micro F1 scores compared to their original counterparts regardless whether the CRF layer was used.

Keywords $\sim$ Natural Language Processing, Language model, Named Entity Recognition, Conditional Random Fields

Zusammenfassung

Vor Kurzem haben vortrainierte Sprachmodelle den Bereich der Verarbeitung von natürlichen Sprachen (NLP) vorangebracht. Die Einführung von Bidirectional Encoders for Transformers (BERT) hat die Bedeutung von vortrainierten Modellen erhöht. Anfangs wurde im Bereich englischer Textdaten geforscht, gefolgt von Modellen, die mit mehrsprachigen Textkorpora trainiert wurden. Allerdings existieren inzwischen wenige domänenspezifische Sprachmodelle, die sich mit der Semantik und Syntax von Texten eines bestimmten Bereichs befassen. Beispiele für BERT-basierte domänenspezifische Sprachmodelle sind SciBERT, BioBERT und MatSciBERT. Die direkte Anwendung dieser Modelle im Bereich der Physik und der Informatik kann zu suboptimalen Ergebnissen führen, denn die Modelle sind nicht mit den für den Bereich spezifischen Bezeichnungen und Fachausdrücke trainiert. In dieser Arbeit werden neue BERT Sprachmodelle vorgestellt, die mit Hilfe von Texten aus dem Bereich der Physik und Informatik aus dem arXiv-Repository vortrainiert werden. Die neue Sprachmodelle werden anhand ihrer Leistung bei einer nachgelagerten Named-Entity-Recognition (NER) Aufgabe in den Informatik- und Physikdomänen bewertet. Die in dieser Arbeit an unserem Textkorpus vortrainierten Sprachmodelle haben kleinere Pseudo-Perplexitätswerte erreicht. Außerdem sind die Sprachmodelle auch in der Lage, höhere Mikro F1-Werte im Vergleich zu BERT und SciBERT zu erzielen. Eine Verbesserung von bis zu 0,69% und 3,89% wurden jeweils für die NER-Datensätze aus dem Informatik- bzw. Physikbereiche gemessen. Ein zusätzlicher Conditional Random Fields (CRF)-Schicht hat allerdings keine Verbesserungen zu der NER-Leistung der Modelle gebracht.

Schlagwörter $\sim$ Natural Language Processing, Sprachmodell, Named Entity Recognition, Conditional Random Fields