Institut für Informatik

Reproducibility of SciBert

Jan-Niklas Weder

Modul : Data Mining

Contents

Introduction

Als pretrained model hängt bert stark von dem korpus ab
Genauso ist das vocabular sehr wichtig
Andere arbeiten zeigten den einfluss eines erweiterten trainings/besser passenden korpus? referenz suchen
=¿ Roberta zum Beispiel zeigte, dass allein weiteres training die ergebnisse verbessern kann Wissenschaftlöiche texte unterschieden sich allgemein sehr stark von "normalen"
=¿ Somit ist ein auf ähnliche weise trainiertes modell als bessere grundlagen für NLP aufgaben im wissenschaftlichen bereich sinnvoll
Gab es in dieser form noch nicht
besitzt daher potential.
insbesondere als grundbaustein für unterschiedlichste aufgaben im !!! wiss. bereich
Allgemein stellt sich das Problem von Datensätzen insbesondere da diese annotiert werden müssen (im wiss. bereich teuer da hochqualifizierte experten notwendig sind)

State of the art

2.1 Bert

Bert as revolution
pretrained-models
usefull even without finetuning
=¿ unexpected precision

nowadays used for many different NLP tasks
Architecture of bert
extensions of bert like roberta

2.2 Datasets

zum Beispiel NCBI-Disease (versuch einen goldstandard für corpora zu erstellen)
-¿ sehr günstig um darauf entsprechende modelle zu trainieren [1] -¿ SciERC /sciie im repo [2]

SciBert

3.1 Corpus

Comparison

3.2 Vocabulary

BaseVocab vs. SCIVocab
! $\approx 42\%$ overlap

3.2.1 SentencePiece

Experiments

kurze einführung in die test fälle mit einer erklärung, was f1 scores sind
Alles NLP Aufgaben bei denen Bert "überraschend" gut abschneided
Jetzt mit der erweiterung zu scibert erneut betrachtet
Einföuss des vocabulars und des corpus genau gegenübergestellt

4.1 NER

4.2 PICO

4.3 CLS

4.4 REL

4.5 DEP

4.6 Finetuning

4.7 Frozen embeddings

Discussion

Further development

Bibliography

[1] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. "NCBI disease corpus: A resource for disease name recognition and concept normalization". In: *Journal of Biomedical Informatics* 47 (Feb. 2014), pp. 1–10. DOI: 10.1016/j.jbi.2013.12.006.

[2] Yi Luan et al. "Multi-Task Identification of Entities, Relations, and Coreferencefor Scientific Knowledge Graph Construction". In: *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*. 2018.