

Institut für Informatik

Reproducibility of SciBert

Jan-Niklas Weder

Modul : Data Mining

Contents

1	Introduction	1
2	State of the art	2
2.1	Bert	2
2.2	Datasets	2
2.2.1	Chemprot	2
3	SciBert	3
3.1	Corpus	3
3.2	Vocabulary	3
3.2.1	SentencePiece	3
4	Experiments	4
4.1	NER	4
4.2	PICO	4
4.3	CLS	4
4.4	REL	4
4.5	DEP	4
4.6	Finetuning	4
4.7	Frozen embeddings	4
5	Discussion	5
6	Further development	6

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 Wissenschaftliche texte unterscheiden 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

mat the Chemprot copus consists of a develop, test, train of which the develop, test and train folder correspond to the identical named files inside the chemprot folder provided on the GitHub site of scibert. The diffrence arises from database like structure in which the chemprot corpus is original provided, in contrast to those subdivided information sets where for example the text itself is in another file than the postitions and annotiations. Those devided information where combined and are provided in a single file in the already mentioned format. [1, 4]

2.2 Datasets

zum Beispiel NCBI-Disease (versuch einen goldstandard für corpora zu erstellen)
-> sehr günstig um darauf entsprechende modelle zu trainieren [2] -> SciERC /sciie im repo [3] Due to the availability of the original used Datasets from the Authors. We will use their prepeard Datasets, which relay on the original Datasets but is already prepeard so that it is easier to use for training of neural nets. In the following paragraph we will take a closer look at the original data and the individual changes that have been made to use those Datasets for the training process.

2.2.1 Chemprot

Chemprot is in a json lines file format provided. More precise every line consists of a text and the corresponding label. A field for metadata exists as well but is most of the time not used. In its original for-

SciBert

3.1 Corpus

Comparison

3.2 Vocabulary

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

3.2.1 SentencePiece

Experiments

kurze einföhrung in die test fälle mit einer
erklärung, was f1 scores sind

Alles NLP Aufgaben bei denen Bert "überraschend"
gut abschneidet

Jetzt mit der erweiterung zu scibert erneut betrach-
tet

Einföuss des vocabulars und des corpus genau
gegenübergestellt

all got dropout of 0.1 loss cross entropy optemizer
adam finetuning for 2 to 5 epochs

4.6 Finetuning

4.7 Frozen embeddings

4.1 NER

Pretrained model -, linear classification layer with
softmax output

4.2 PICO

4.3 CLS

4.4 REL

4.5 DEP

dependency tag and arc embedding of size 100 and
biaffine matrix attention

Discussion

Further development

Bibliography

- [1] Iz Beltagy, Kyle Lo, and Arman Cohan. “SciBERT: A Pretrained Language Model for Scientific Text”. In: *EMNLP 2019* (Mar. 26, 2019). arXiv: 1903.10676 [cs.CL].
- [2] 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.
- [3] Yi Luan et al. “Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction”. In: *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*. 2018.
- [4] Qinghua Wang et al. “Overview of the interactive task in BioCreative V”. In: *Database 2016* (2016), baw119. DOI: 10.1093/database/baw119.