

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
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

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

SciBert

3.1 Corpus

Comparison

3.2 Vocabulary

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

3.2.1 SentencePiece

Experiments

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