The Value of Pre-training for Scientific Text Similarity: Evidence from Matching Grant Proposals to Reviewers

Gabriel Okasa and Anne Jorstad
gabriel.okasa@snf.ch anne.jorstad@snf.ch

Swiss National Science Foundation

Introduction

Matching grant proposals to reviewers is a core task for research funding agencies. Such matching is a time-consuming task which requires scientific officers to screen available reviewers and assess their suitability. We approach this task as a text similarity problem to allow pre-filtering of a relevant subset of potential matches using pre-trained language models. Given the scientific nature of our English text corpus, we investigate the value of targeted pre-training of BERT models towards scientific documents for the matching task based on the text similarity. We benchmark the performance of BERT models with a classical bag-of-words approach using TF-IDF.

Motivation

- matching grant proposals to reviewers is very time-consuming
- matching procedure can be defined as a text similarity task
- choice of the **text vectorization method** is not *a priori* clear

Setting

- each proposal gets 2 **reviewers from a pre-defined pool** of experts
- assigned reviewers need sufficient expertise in the fields of proposals
- final assignment is adjusted and validated by scientific officers

Data

- text similarity is unobserved - human annotation as a proxy
- Postdoc.Mobility funding scheme August 2021 call for proposals
- English texts: **320 grant proposals** and **125 potential reviewers**
- all research areas included:
  - Social Sciences and Humanities (**SSH**)
  - Life Sciences (**LS**)
  - Mathematics, Informatics, Natural Sciences and Technology (**MINT**)

Research Area	# Proposals (%)	# Reviewers (%)
SSH	50 (15.6)	20 (16.0)
LS	123 (38.4)	43 (34.4)
MINT	147 (46.0)	62 (49.6)

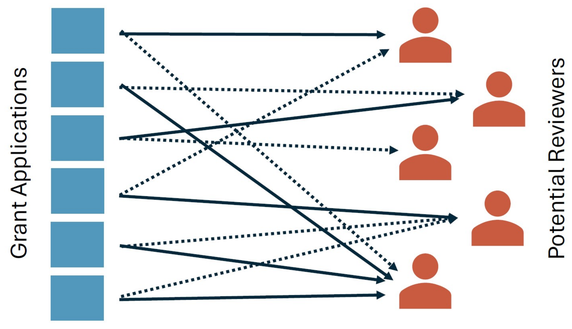
- titles, abstracts and concatenation of both
- last 5 years vs. last 10 years of publications for reviewers

Average Publication Count	SSH	LS	MINT
last 5 years	22.6	44.4	52.1
last 10 years	40.6	81.2	96.0

Methods

- comparison of word embeddings (BERT) vs. bag-of-words (TF-IDF)
- **BERT** models pre-trained specifically on scientific texts:
  - BERT: `google-bert/bert-base-uncased`
  - SciBERT: `allenai/scibert_scivocab_uncased`
  - SPECTER2: `allenai/specter2_base`
- **TF-IDF** as a benchmark method without context awareness
- text vectorization:
  - BERT embeddings via CLS token and mean pooling
  - TF-IDF vectors via uni-grams and 3-grams

Matching



- text vectorization of proposal texts and reviewer publications' texts
- estimation of **text similarities via cosine distance**
- proposal-reviewer similarity via averaging top publication similarities
- rank-order the reviewers according to the average similarities
- for a proposal select top R reviewers as a matching recommendation

Code

- GitHub: `https://github.com/snsf-data/snsf-grant-similarity`

Evaluation

- **Mean Average Precision** (MAP) as a distance-sensitive metric
- penalization of recommendations realized at lower ranks
- approximation of the average area under the precision-recall curve

$$MAP = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{P_i}\sum_{r=1}^{R}\mu(r)\cdot\frac{n\left(\tilde{J}_i^P \cap \hat{J}_i^R(r)\right)}{r}\right)$$

- P_i - true positives; $\mu(\cdot)$ - relevance function; J_i - recommendations.

Results

- **SPECTER2 best performing model** across all scenarios
- TF-IDF comparable to BERT and SciBERT despite its simplicity

Text	BERT	SciBERT	SPECTER2	TF-IDF
title	0.3175 (0.1054)	0.3316 (0.1093)	0.3842 (0.1141)	0.2585 (0.0893)
abstract	0.3675 (0.1067)	0.4205 (0.1053)	**0.4687 (0.1136)**	0.4000 (0.1106)
title + abstract	0.3696 (0.1071)	0.4184 (0.1052)	0.4619 (0.1161)	0.4033 (0.1101)

MAP at R=5: mean pooling / 3-gram; 10 years of publications

Heterogeneity

- heterogeneity across research areas, SSH particularly challenging

Research Area	BERT	SciBERT	SPECTER2	TF-IDF
SSH	0.2915 (0.1095)	0.3150 (0.0960)	**0.4070 (0.1270)**	0.3312 (0.0844)
LS	0.3411 (0.0960)	0.4239 (0.0972)	**0.4583 (0.0998)**	0.3686 (0.0963)
MINT	0.4154 (0.1112)	0.4536 (0.1118)	**0.4985 (0.1199)**	0.4497 (0.1279)

MAP at R=5 by Research Area:
mean pooling / 3-gram; 10 years of publications; abstracts

Conclusion

The results reveal a clear benefit from pre-training BERT on scientific texts and additionally augmenting by citation graphs. In particular, the SPECTER2 model unanimously out-performs all competing methods across all testing scenarios. Interestingly, the TF-IDF approach is on par with most of the BERT models, despite its simplicity and lack of context awareness. The results contain substantial heterogeneity with respect to research areas: the overall best matching performance is recorded for MINT, followed by LS and lastly SSH, for which the matching turns out to be particularly challenging. These results are robust to text inputs and modelling choices.