Mining Acknowledgement Texts in Web of Science (MinAck)

Nina Smirnova*[1,2], Philipp Mayr[1]

[1] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany; [2] University of Bonn, Germany

* Presenting author: ninasmirnova@web.de

Background and objectives

The focus of our project MinAck is the **detection and quantitative analysis of acknowledged entities**, i.e., named entity recognition (**NER**) task in a larger corpus of Web of Science (**WoS**) articles, which include **acknowledgements**.

Why Acknowledgments?

- may give an insight on reward systems, collaboration structures, and hidden research trends in scientific community (Giles & Councill, 2004)
- can help the reader to better understand the set-up and framing of a given scientific text
- their automatic analysis poses interesting research and methodological problems

NLP tool:

FLAIR framework for state-of-the-art NLP (Akbik et al., 2019).

Project steps:

Step 1. Create datasets: define the disciplines and gather the acknowledgement texts from WoS.

- **2 training corpora** (50 and 300 entries)
- **1 acknowledgments corpus** (200,000 entries, i.e., 50,000 from each of the 4 scientific domains)

Records from:

- **four** different scientific disciplines (**sociology, economics**, **oceanography, computer science)**
- published from **2014** to **2019**
- WoS records types "***article***" and "***review***"

Step 2. Annotate the training data for the FLAIR NLP Framework.

6 entity types were defined:

IND : person
FUND : funding organization
GRNB : grant number
UNI : university
COR : corporation
MISC : miscellaneous

(1) Jan De Houwer is supported by Methusalem Grant BOF09/01M00209 of Ghent University and by the Interuniversity Attraction Poles Program initiated by the Belgian Science Policy Office (IUAPVII/33).
(2) Data on Anthem Blue Cross PPO enrollees were provided by Anthem, Inc.

Annotated corpus:

id	txt	GRNB	FUND	IND	UNI	COR	MISC
3121934	Scott Griffiths is supported by an Australian National Health and Medical Research Council Early Career Fellowship (grant number: 1121538).	1121538	Australian National Health and Medical Research Council Early Career Fellowship	Scott Griffiths			
4935696	This research was supported by the Oesterreichische Nationalbank, Anniversary Fund (Project No. 13042).	13042				Oesterreichische Nationalbank, Anniversary Fund	
1365702	The author would like to express her gratitude to the Center of Excellence in Teaching and Learning (CETaL), UTP for awarding the Scholarship of Teaching and Learning (SoTL: 0152AA-A09) research grant for this study.	SoTL: 0152AA-A09			Center of Excellence in Teaching and Learning;CETaL;UTP		Scholarship of Teaching and Learning

Flair corpus format

```
Scott B-IND
Griffiths I-IND
is O
supported O
by O
an O
Australian B-FUND
National I-FUND
Health I-FUND
and I-FUND
Medical I-FUND
```

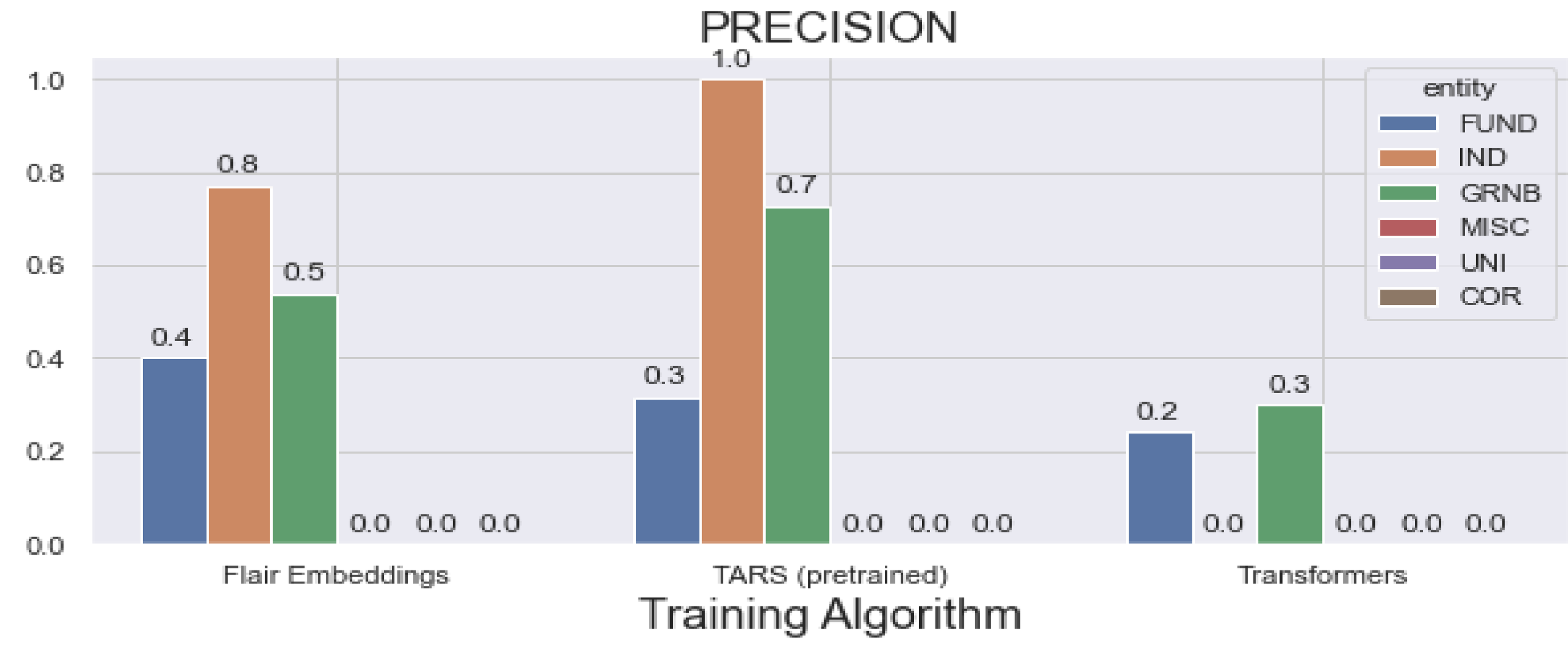
A semi-automated annotation approach was developed.
Two training corpora, containing possibly equal amount of entities of all types were created.
The effectiveness of corpuses of different sized will be tested on different training algorithms.

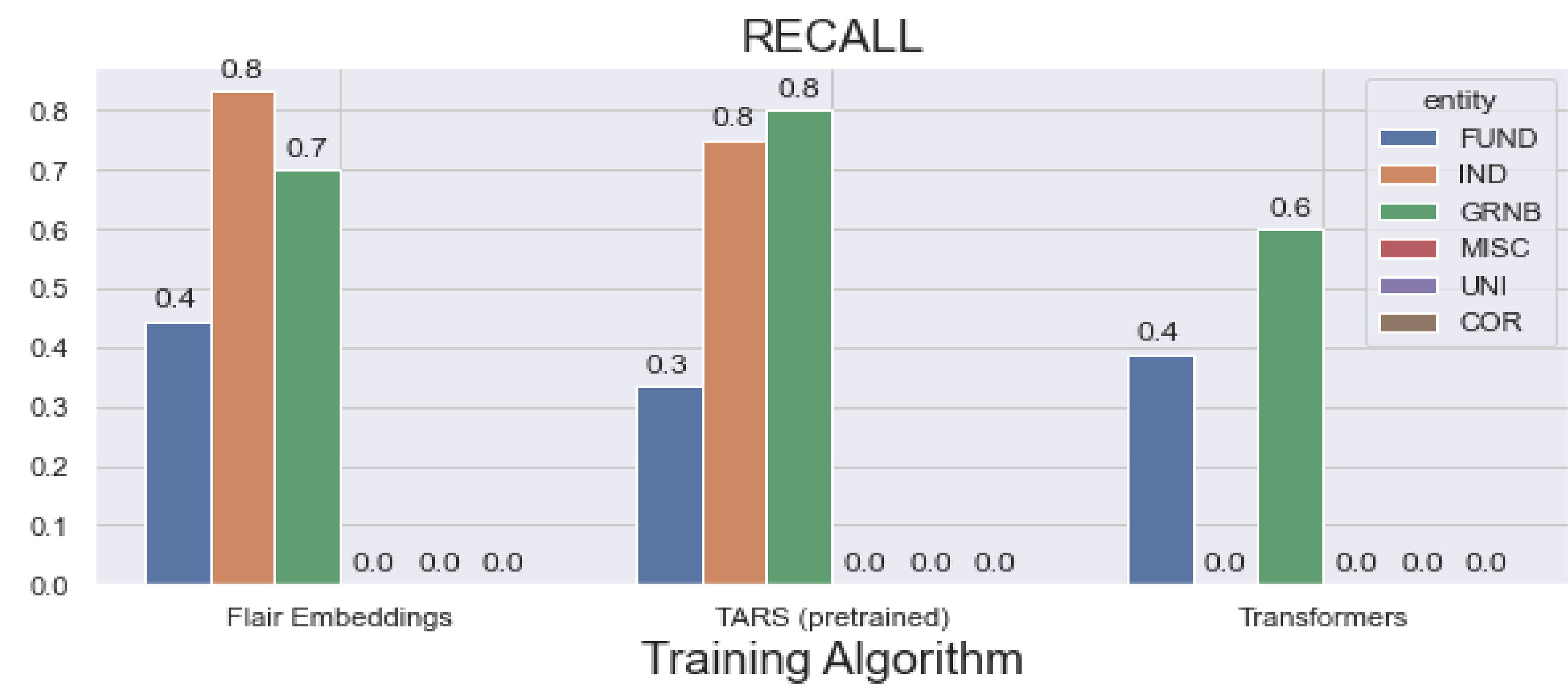
Step 3. (in progress) Train the FLAIR with the training datasets and define the best model and training algorithm.

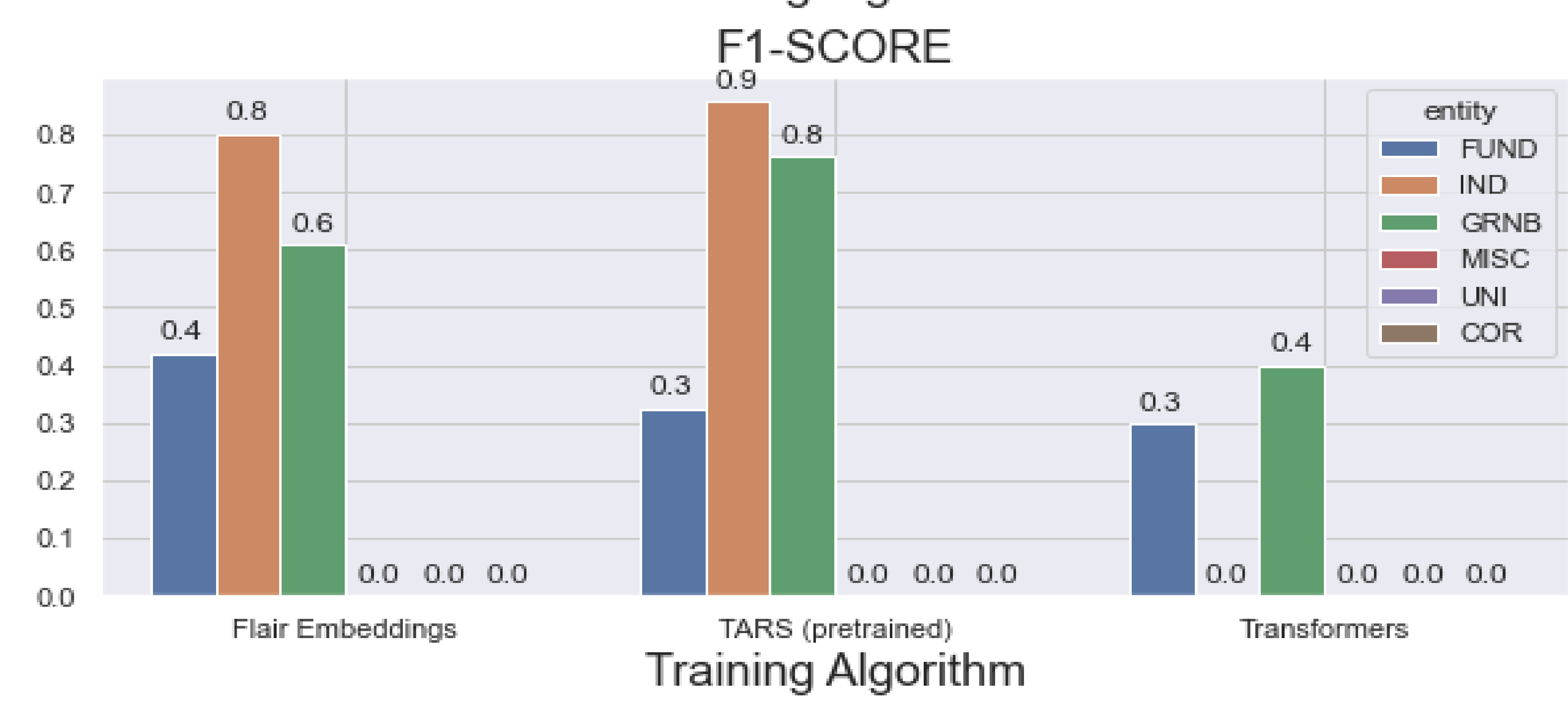
A small dataset (50 sentences) was tested with three FLAIR training algorithms:

- NER Model with Flair Embeddings
- NER Model with Transformers
- Zero-shot NER Model (TARS)

Training results:







Step 4. Analysis with the best FLAIR model.

Step 5. Aggregating the results.

References: Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S. & Vollgraf, R. (2019). *FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP* (W. Ammar, A. Louis, & N. Mostafazadeh, Eds.; pp. 54–59). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4010

Giles, C. L. & Councill, I. G. (2004). Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences*, *101*(51), 17599–17604. https://doi.org/10.1073/pnas.0407743101