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

1. Motivation and introduction to the topic

Acknowledgements in scientific papers are short texts where the author(s) "*identify those who made special intellectual or technical contribution to a study that are not sufficient to qualify them for authorship*" (Kassirer & Angell, 1991, p. 1511). The focus of our project MinAck[1] is the detection and quantitative analysis of acknowledged entities, i.e., the named entity recognition (NER) task in a larger corpus of Web of Science (WoS) articles, which include acknowledgements. An *acknowledged entity* is an object in the acknowledgment which can consist of e.g. names and surnames of individuals (also abbreviations), names of institutions and organizations, numbers, or identifiers of grants.

The analysis of acknowledgments is particularly interesting as acknowledgments may give an insight on such aspects of the scientific community as reward systems, collaboration structures, and hidden research trends (Giles & Councill, 2004). In addition, acknowledgements can help the reader to better understand the set-up and framing of a given scientific text. From the linguistic point of view, acknowledgements are unstructured text data, which automatic analysis poses interesting research and methodological problems like data cleaning, tokenization, word embedding.

The present project aims to create a method for automatic extraction and classification of acknowledged entities from acknowledgment texts and examine the correlation between the acknowledged entity category and scientific domain.

2. Materials and method

The aim of the project is to conduct a large-scale analysis of acknowledgment texts from WoS using the FLAIR NLP Framework (Akbik et al., 2019).

As WoS contains millions of metadata records, the data chosen for the present study was restricted by year and scientific discipline. Records from four different scientific disciplines published from 2014 to 2019 are considered: two disciplines from the social sciences (**sociology** and **economics**) and **oceanography** and **computer science** for comparison. Only WoS records types "*article*" and "*review*", published in a scientific journal, were selected. The entire acknowledgments dataset for our study contains approx. 200,000 entries, i.e. 50,000 from each of the four scientific domains.

Two of the aims of the present project are to extract acknowledged entities from the acknowledgments corpus and ascribe them to different categories. Categories were chosen according to Giles and Councill (2004, p. 17601) classification: funding agencies (FUND), corporations (COR), universities (UNI), individuals (IND) and grant numbers (GRNB). For the present project this classification was enhanced with the MISC (miscellaneous) category. In this

[1] https://kalawinka.github.io/minack/

category fall entities, which could provide useful information, but can not be ascribed to the other categories, e.g., names of the ships, names of projects, names of conferences.

FLAIR provides the possibility to create a custom NER tagger (Chauhan, 2020). Creating a custom NER tagger will allow us to accomplish these two aims (a) acknowledged entity recognition and (b) acknowledged entity classification in one step.

In order to train the FLAIR custom NER tagger model, two corpora containing 50 and 300 acknowledgments texts were created. The effectiveness of corpora of different sizes has to be tested in order to find out the most efficient training corpus size. The last FLAIR release (0.9[2]) has the possibility to conduct NER without any training data or with a small dataset (Zero-shot Named Entity Recognition (NER) with TARS) (Halder et al., 2020). NER without training data did not work properly for some categories. We are planning to test TARS using a training set with 50 sentences and compare the results using the usual training algorithm with a set of 300 sentences. Corpus annotation was produced by the authors using a semi-automated approach developed by the authors.

3. Work in progress report

The presentation will be a work in process paper. We will present the created acknowledgments corpus (with approx. 200,000 entries) and four training corpora, together with the approaches for creating the corpora.

FLAIR has three default training methods. These methods will be described and discussed. At the end of the project we are planning to quantitatively analyze the extracted entities: e.g. the most cited entities will be detected. The appearance and frequency of entities of different categories will be analyzed and compared between different scientific domains (sociology and economics against oceanography and computer science). Our preliminary results of the study will be presented and discussed with the workshop participants.

Keywords: natural language processing, named entity recognition, FLAIR, acknowledgments from Web of Science

[2] https://github.com/flairNLP/flair/releases

References

1. 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
2. Chauhan, A. (2020). Training Custom NER Model Using Flair. *TheCyPhy*. https://medium.com/thecyphy/training-custom-ner-model-using-flair-df1f9ea9c762
3. Diaz-Faes, A. A. & Bordons, M. (2017). Making visible the invisible through the analysis of acknowledgements in the humanities. *Aslib Journal of Information Management*, *69*(5), 576–590. https://doi.org/10.1108/AJIM-01-2017-0008
4. 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
5. Halder, K., Akbik, A., Krapac, J. & Vollgraf, R. (2020). Task Aware Representation of Sentences for Generic Text Classification. *28th International Conference on Computational Linguistics*. COLING 2020.
6. Kassirer, J. P. & Angell, M. (1991). On authorship and acknowledgments. *The New England Journal of Medicine*, *325*(21), 1510–1512. https://doi.org/10.1056/NEJM199111213252112