Borst, Timo; Mielck, Jonas; Nannt, Matthias; Riese, Wolfgang

Conference Paper — Published Version

Extracting Funder Information from Scientific Papers - Experiences with Question Answering

Suggested Citation: Borst, Timo; Mielck, Jonas; Nannt, Matthias; Riese, Wolfgang (2022) : Extracting Funder Information from Scientific Papers - Experiences with Question Answering, In: Silvello, Gianmaria et al. (Ed.): Linking Theory and Practice of Digital Libraries. Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, ISBN 978-3-031-16802-4, Springer, Cham, pp. 289-296, https://doi.org/10.1007/978-3-031-16802-4_24

This Version is available at:
http://hdl.handle.net/11108/533

Extracting Funder Information from Scientific Papers - Experiences with Question Answering

Timo Borst[1], Jonas Mielck[2], Matthias Nannt[2], and Wolfgang Riese[1(✉)]

[1] Leibniz Information Center for Economics, Kiel, Germany
w.riese@zbw.eu
[2] stackOcean GmbH, Kiel, Germany

Abstract. This paper is about automatic recognition of entities that funded a research work in economics as being expressed in a publication. While many works apply rules and/or regular expressions to candidate sections within the text, we follow a question answering (QA) based approach to identify those passages that are most likely to inform us about funding. With regard to a digital library scenario, we are dealing with three more challenges: confirming that our approach at least outperforms manual indexing, disambiguation of funding organizations by linking their names to authority data, and integrating the generated metadata into a digital library application. Our computational results by means of machine learning techniques show that our QA performs similar to a previous work (AckNER), although we operated on rather small sets of training and test data. While manual indexing is still needed for a gold standard of reliable metadata, the identification of funding entities only worked for a subset of funder names.

Keywords: Funder recognition · Question answering · Metadata

1 Introduction

Named-entity recognition (NER) has become one of the most important fields of research in text analysis that has yielded some impressive results with regard to identifying almost any kind of 'thing' or entity a text is about [12]. However, despite of some undisputed progress in adopting and fine-tuning linguistic and computational methods for extracting entities, we still rarely see those techniques being adopted within digital library scenarios and applications. This may be symptomatic: first, it still might mean quite a step even for digital libraries and their workflows to build trust in automatic metadata extraction [7], and secondly it requires some long-term commitment and technical expertise not only to engage with these approaches, but also to support and maintain them in a productive setting.

This paper is about automatic and trained recognition of research funding agencies (FA) that are explicitly mentioned in parts or sections of scientific

G. Silvello et al. (Eds.): TPDL 2022, LNCS 13541, pp. 289–296, 2022.
https://doi.org/10.1007/978-3-031-16802-4_24

publications, which are commonly known as acknowledgement phrase (AP) or acknowledgement text (AT). While this might appear as a simple and straightforward task to be perfectly handled by some NER framework or (pre-)trained language model, we were more curious about finding if recent question answering approaches can be applied to meet three basic requirements: first, we aim at performing as automatically as possible by generating metadata that looked particularly suited for that purpose by implying an essentially binary decision ('funder/grant no. or not?'). Automatic text mining of funder information generally outperforms manual curation particularly with respect to recent papers that are to be indexed yet, with 90% of grants (almost correctly, according to the authors) found by text mining [4]. Taking this for granted, although a text mining approach will miss around 10% recall, it suggests providing information much more timely than a manual indexing process. Secondly, by becoming productive we require the generated metadata to be as flawless as possible, in particular by preventing false-positives. And thirdly, we strive for a productive setting in terms of an existing search application indexing the funding information, e.g. as a metadata facet.

In the following section, we discuss some related work treating NER. In section three, we depict our corpus and textual data together with some subsidiary data sets to accomplish our own NER approach. We explain our technical approach and framework more into detail including different language models and parameters we used. In section four, we delineate and compare the results of our test runs. We conclude by relating the main outcomes to our three basic requirements.

2 Related Work

In recent years, the analysis and extraction of FAs and/or grant numbers (GN) has become subject of both experimental data analyses and retrieval applications. Many works rely on the assumption that acknowledgements are a broader concept in scientific communication, e.g. by distinguishing between moral, financial, editorial, presentational, instrumental, technical, and conceptual support [5]. Therefore, most approaches follow a two-stage process by first identifying a potential textual area, before analysing this 'candidate section' more into detail by distinguishing between FAs represented by their names or acronyms and grants represented by their numbers or codes [1,3,4,6,8,13]. During the second stage, these two metadata are constitutive for extracting FAs. Some works differ in applying either regular expressions, rule-based and/or machine learning based approaches, or a combination of them. While [5] and [10] use regular expressions to identify name variations of selected FAs they are interested in, more inclusive approaches such as [4] or [13] apply a 'rule-based section tagger' to identify the most significant parts of an acknowledgement phrase including candidates for funding entities. Classifiers for calculating and weighing the acknowledgement phrase and its constituents rely on popular language models, such as Stanford-CRF [6], spaCy [1] or SVMs [1,5,13]. Despite benefiting from these pre-trained

models, only [13] and [3] made active use of supervised machine learning by organizing and tuning their runs with different training data and continuously adapted classifiers. Only a few works tackle the challenge of normalizing a FA's name by mapping it to a canonical notation [10] or an authority record from a funder database, such as the funder registry from Crossref [6].

Apart from providing manually indexed funding data with databases such as Web of Science[1] or information portals[2], some works at least temporarily integrated the results of their runs into productive bibliographic online databases, e.g. PMCEurope [4] or DBLP[3]. Even if these information services are not propagating the retrieval of FAs, stakeholders such as research funders may find them as a valuable source for assessing the impact of their fundings, as suggested and depicted in [3].

3 Approach

3.1 Processing Pipeline

We decided to use Haystack[3], a framework written in Python, for NLP based QA through transformer-based models (e.g., RoBERTa [9], MiniLM [11]). To our advantage there already exist several pre-trained models for QA with Haystack on Hugging Face[4], so that there was no need to train a model on our own.

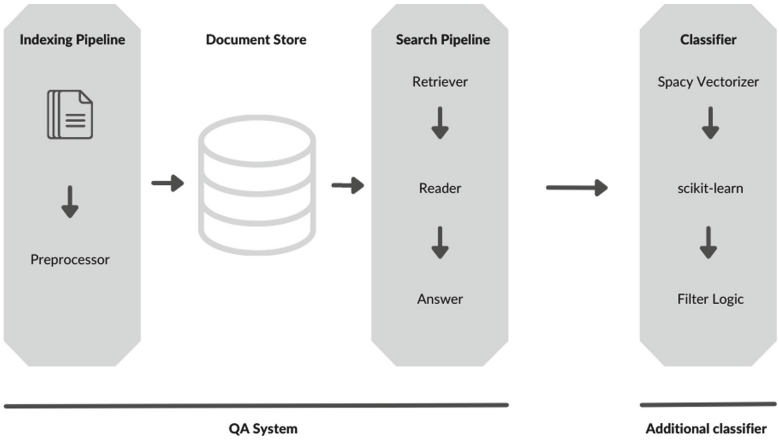

Fig. 1. Processing pipeline

[1] https://www.webofscience.com.
[2] https://explore.openaire.eu/.
[3] https://github.com/deepset-ai/haystack.
[4] https://huggingface.co/models?dataset=dataset:squad_v2&pipeline_tag=question-answering.

By manually going through different papers and analysing the wording in which the funder information is presented, we came up with the following questions we provided to the QA system:

Table 1. Question overview: The questions in bold text are the ones that perform best according to their F-score measures, cf. Table 2.

Number	Question
1	Who funded the article?
2	Who funded the work?
3	Who gives financial support?
4	**By whom was the study funded?**
5	Whose financial support do you acknowledge?
6	**Who provided funding?**
7	Who provided financial support?
8	**By which grant was this research supported?**

In the processing pipeline we first extracted the plain text from a PDF document, before we enriched it with various metadata from the repository and its language, the latter determined via the PYCLD2[5] library. In the subsequent step of pre-processing, for example white spaces and empty lines are removed.

The pre-processed plain-text documents and their metadata are then placed in the Elasticsearch 'Document Store' from where they are retrieved by the 'Search Pipeline'. This pipeline starts by processing the list of varying questions about the funder of the research work. Then the retriever proceeds by filtering the 'Document Store' and returning the documents that are most likely relevant for each variant question. By using a pre-trained language model, the Reader predicts an answer for each question variant and provides further data, for example an accuracy score.

While processing the APs from the documents and the funder metadata from Crossref[6] and DataCite[7], another problem came to our attention. So far, we had expected that only the funding of the research would be stated. But we noticed, that in some of the acknowledgement sections of the publications, the funding of open access publishing had also been acknowledged. Since only the information on research funding was relevant to us, this posed an unexpected complication, because now we also needed an automatic detection/filtering of these unwanted findings.

To filter them and to lower the number of false positives, we came up with an additional classifier. The complete pipeline including classifier is shown in Fig. 1.

[5] https://pypi.org/project/pycld2/.
[6] https://github.com/CrossRef/rest-api-doc.
[7] https://support.datacite.org/docs/api.

The classifier receives a prediction from the QA model and checks whether it is really a funder. For this purpose, the classifier was trained with the context information found by Haystack in the previous steps. We balanced the dataset before training by splitting it into 80% training data and 20% test data that is unknown to the model. The contexts are then trained in a support vector machine with the support of grid search for hyperparameter tuning.

In the final step of our approach we looked up the preferred labels of the extracted funding information with the help of the authority file of the Crossref Funder Registry v1.34[8] structured in RDF. It is important to note that we set up our pipeline as asynchronous process in order to be more independent from a just-in-time metadata generation probably demanding more powerful hardware.

3.2 Data

For our test sample we first extracted about 7100 open access documents, with a license that would allow us text and data mining, from the EconStor[9] repository, a publication server for scholarly economic literature. In a second step we identified 653 documents - most of them in English - with an associated DOI in that sample. For these documents we extracted the funder information via the Crossref and DataCite APIs. But a quick checkup revealed that not all funders mentioned in the metadata associated with the DOIs could be confirmed in the documents. Therefore we had the APs manually labelled from these 653 documents, so that we had a list with the complete AP for each document that indeed contained such a statement. All this data is combined into a single spreadsheet containing information about all papers including the manually labelled statement whether the paper was funded or not, according to our definition. This definition excludes open access funding which was the case for four of these papers. However, the following attributes have been stored: local repository id, DOI, funder by Crossref API (DOI), funder by Crossref API (plain text) and the manually extracted AP. In order to identify the funders, we used the authority file from the Crossref Funder Registry structured in RDF with 27953 funders and 46390 alternative labels. During our analysis we identified 83 papers to have not been published in English.

4 Results

The F-score relates true positive (TP), false positive (FP) and false negative (FN) values and is commonly used for measuring the quality of NER, cf. [1] or [2]. In this paper, the following formula is used for calculation:

$$F = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{1}$$

[8] https://gitlab.com/crossref/open_funder_registry/-/commits/v1.34.
[9] https://www.econstor.eu.

Table 2. F-score comparison of the different language models and three best performing questions. If there is no F-score shown, the accuracy was below benchmark in data pipeline and the language model was therefore dropped.

Question	RoBERTa	ELECTRA	albert-xxl-v2	minilm	XLM-RoBERTa
Question 4	0.6807	0.6721	0.6867	–	0.6015
Question 6	0.6135	**0.6729**	**0.6868**	–	**0.6248**
Question 8	**0.7836**	–	–	**0.7996**	–

Two of the language models used achieve F-scores of close to 0.8 which is equal to what [1] find.

The F-scores of the examination including the classifier differ slightly from the results without classifier. However, the deviation is not large enough to draw conclusions from it. In order to compare the results with and without classifier, the test data of the model without classifier must be reduced to the test data of the model with classifier. This results in a data set split to about 400 papers for training and about 100 papers for testing. We found the test set too small to make any statements about model performance. To put things into perspective, [1] use 321 articles for testing, [3] train on 800 and test on two data sets of 600 documents and [13] train on 2100 documents which they add up iterative and test iterative up to 1100 papers. But this overview suggests that the F-scores shown in Table 2, based on the 653 papers calculated without the classifier, are calculated from a sample size that is similar to what other researchers have. Hence, the presented results appear to be robust from that perspective. As additional analysis, we looked up the preferred label for the funder names with the help of the Crossref Funder Registry. Our algorithm utilized RapidFuzz[10] text comparison and was able to identify 126 funding entities from the 367 funder names correctly found with question 8 "By which grant was this research supported?" and the "RoBERTa" model. The algorithm identified 17 funders incorrectly.

5 Discussion and Outlook

Our results demonstrate the feasibility to automatically extract funding entities, basically confirming the results from AckNER. Our sample size was too small to evaluate the quality of the self-trained classifier model, to this end we would need a larger corpus. Moreover, we still require a gold standard of manually checked funder information, as the reference data provided through Crossref metadata turned out to be inaccurate. In particular, we could not train our classifier for identifying and excluding open access acknowledgements, which are becoming more frequent. With respect to transfering our results into a service environment in terms of a digital library, we could set up an asychronous data processing pipeline for regular metadata generation that demands maintenance of its components, such as Haystack and some Python libraries.

[10] https://github.com/maxbachmann/RapidFuzz/tree/v1.4.1.

The code and data underlying this paper is available on Github at https://github.com/zbw/Funder-NER.

References

1. Alexander, D., de Vries, A.P.: This research is funded by...: named entity recognition of financial information in research papers (2021). http://ceur-ws.org/Vol-2847/paper-10.pdf, https://repository.ubn.ru.nl/handle/2066/236372
2. Bian, J., Huang, L., Huang, X., Zhou, H., Zhu, S.: Grantrel: grant information extraction via joint entity and relation extraction. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 2674–2685 (2021). https://doi.org/10.18653/v1/2021.findings-acl.236
3. Councill, I.G., Giles, C.L., Han, H., Manavoglu, E.: Automatic acknowledgement indexing: expanding the semantics of contribution in the CiteSeer digital library. In: Proceedings of the 3rd International Conference on Knowledge Capture, pp. 19–26. K-CAP 2005, Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1088622.1088627
4. Europe PMC Consortium: Extracting funding statements from full text research articles in the life sciences (2014). https://cordis.europa.eu/project/id/637529/results
5. Giles, C.L., Councill, I.G.: Who gets acknowledged: measuring scientific contributions through automatic acknowledgment indexing. Proc. Natl. Acad. Sci. **101**(51), 17599–17604 (2004). https://doi.org/10.1073/pnas.0407743101
6. Gregory, M., Kayal, S., Tsatsaronis, G., Afzal, Z.: Systems and methods for extracting funder information from text (2019). https://patents.google.com/patent/US20190005020A1/en
7. Irvin, K.M.: Comparing information retrieval effectiveness of different metadata generation methods (2003). https://doi.org/10.17615/grff-0v98
8. Liu, W., Tang, L., Hu, G.: Funding information in web of science: an updated overview. arXiv:2001.04697 [cs] (2020). http://arxiv.org/abs/2001.04697
9. Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach (2019). http://arxiv.org/abs/1907.11692
10. Sirtes, D., Riechert, M.: A fully automated method for the unification of funding organizations in the web of knowledge (2014). https://doi.org/10.13140/2.1.3086.5285
11. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers (2020). https://doi.org/10.48550/ARXIV.2002.10957
12. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models (2019). https://doi.org/10.48550/arXiv. 1910.11470
13. Zhang, X., Zou, J., Le, D.X., Thoma, G.: A semi-supervised learning method to classify grant support zone in web-based medical articles. In: Berkner, K., Likforman-Sulem, L. (eds.) Document Recognition and Retrieval XVI, vol. 7247, pp. 286–293. International Society for Optics and Photonics, SPIE (2009). https://doi.org/10.1117/12.806076