NER for acknowledgements
Project Proposal for NLP Course, Winter 2023

Sebastian Deregowski, Dawid Janus,
Bartosz Jamroży, Klaudia Gruszkowska
Warsaw Univeristy of Technology
sebastian.deregowski.stud@pw.edu.pl
klaudia.gruszkowska.stud@pw.edu.pl
bartosz.jamrozy.stud@pw.edu.pl
dawid.janus.stud@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP), aimed at identifying and classifying named entities in text. In this paper, we present our project's goal and methodology, which revolves around developing and evaluating NER models for recognizing and classifying entities in scientific acknowledgements. We build upon the work of Smirnova and Mayr and train NER models using Flair embeddings and Transformer models. Our research questions address the effectiveness of NLP techniques, the impact of training data size, and the use of text normalization techniques in NER. We hypothesize that deep learning models, increased training data diversity, and text normalization improve NER performance. Our project's scientific goal is to improve the understanding of scientific acknowledgements, aiding the identification of relationships and collaborations among scientists. We anticipate that our findings will contribute to the development of NLP and scientific text analysis. We provide insights into the project's work plan, research methodology, and datasets used. Additionally, we discuss alternative NER approaches, including CycleNER and Question Answering, which could be valuable for future research in NER.

Keywords: Named Entity Recognition (NER), Natural Language Processing (NLP), acknowledgements

1 Goal of the project

Named entity recognition is a fundamental task in natural language processing (NLP), the aim of which is to identify and classify named entities into predefined categories such as people, organisations and many others. In a variety of fields such as healthcare, finance, law and science, the correct recognition of these entities is crucial for making meaningful inferences, facilitating information retrieval and enhancing text comprehensibility. Incorrect or incomplete recognition of units can lead to misinformation, misinterpretations and impede accurate decision-making. The aim of our project is to develop and evaluate named entity recognition (NER) models for identifying and classifying entities in acknowledgements. We plan to build our work on the foundations of the paper written by Smirnova and Mayr (1) by training the models presented in the paper on the provided data and conducting their own evaluation. In addition, we want to try a different approach and compare what results LLM models can achieve and create a silver set, a corpus of articles with automatic annotations provided by our new trained models.

1.1 Research questions

- What techniques and models in the field of NLP are most effective for recognising named individuals in a specific area?
- How does the performance of NER models change with different types and amounts of training data?
- Can the use of different text normalisation techniques, such as lemmatization or stop word removal, improve the effectiveness of NER models in identifying entities in scientific acknowledgement texts?
- Can preparing a silver-set and further training models on it improve their effectiveness?
- Which types of entities are the most difficult to classify?

1.2 Hypotheses

- Deep learning models, especially those using transform architectures, will outperform traditional machine learning approaches in recognising named entities, especially for complex and context-dependent entities.
- Increasing the size and diversity of the training dataset will improve the accuracy and generality of NER models, providing better recognition of actors
- The use of text normalisation techniques, such as lemmatization or stop word removal, combined with NER models, leads to improved accuracy and precision in the identification of entities in scientific acknowledgement texts.
- Preparing a silver-set and further training models on it will improve performance.
- Proper names of corporations, especially with specific words, will be the most difficult to recognise.

2 Significance of the project

In our project, we will use advanced NER models, especially those based on transformers, using the Flair library. We focus on the identification and classification of entities in scientific acknowledgements, which is a significant challenge in the field of natural language processing (NLP). With the growth of scientific data, understanding the implicit relationships and collaborations of scientists in acknowledgements becomes particularly important. Our research in this specific area of NLP is crucial to improve the identification of relationships between scientists and to assess the impact of financial and technical support on research outcomes.
We expect our findings to have a significant impact on the development of the field of NLP and scientific text analysis. By identifying key individuals in scientific acknowledgements and analysing their interrelationships, our findings may facilitate the understanding of key figures in the field of science. In addition, our research will influence the way scientists collaborate, enabling a more efficient and organised exchange of knowledge. This combination of advanced NER technologies with the analysis of acknowledgements a step towards novel solutions in natural language processing and scientific research.

3 Work plan

During first two weeks we are going to work on Proof of Concept. As our result for that phase, we'd like to present a detailed analysis of the datasets provided, as well as the preprocessing steps and feature engineering. We believe it's crucial to prepare the data as precisely as possible before moving to modelling phase. Next in line, we want to start working on NLP models (described in more details in the next section).

After PoC, we plan to spend the next three weeks on model's implementation and tuning. Based on PoC findings we are going to choose several methods to implement and compare. The final solution should consist of a reproducible data processing pipeline, group of techniques and models (exact number to be discussed after PoC) that we'll consider the most relevant and effective, as well as a detailed summary of their comparison. We'd like to pay attention to the advantages and drawbacks of all methods provided and describe how data affects the performance of each. The last week (after submitting the solution) we plan to spend on preparing a report summarizing all the work done.

As for risks, the most probable problems that we may encounter are: (i) lack of reproducible code for modeling, (ii) inadequate methods that may force looking for new ones, (iii) poor results, (iv) lack of time to deeply investigate whole domain.

4 Approach & research methodology

4.1 Datasets

Creating a dataset for NLP tasks and especially for this type of NER task is very hard due to the need to tag a lot of data, which should be done by experts. Therefore, we plan to use a ready-made training dataset provided. It consists of 4 corpus. Each corpus consists of a different amount of data. Starting with the first one whose dataset is the smallest the amount of data in each subsequent one is increased. Each corpus has been divided into training, testing and validation collections and figure 1 shows an example of how the data has been labeled.
IND: person, FUND: funding organization,

GRNB: grant number, UNI: university, COR: corporation, MISC: miscellaneous

```
This O
work O
was O
supported O
by O
National B-FUND
Social I-FUND
Science I-FUND
Foundation I-FUND
of I-FUND
China I-FUND
( O
17BTJ019 B-GRNB
) O
. O
```

Figure 1: Schema of dataset

4.2 Flair Embeddings

NER model with Flair Embeddings is a method which create contextual word embeddings for text data (5). Unlike traditional word embeddings like Word2Vec or GloVe, Flair Embeddings consider the surrounding words and the context in which a word appears. It thus gives different embeddings for the same word depending on it's surrounding text. This makes them particularly useful for tasks like named entity recognition, part-of-speech tagging, sentiment analysis, and other sequence labeling tasks, where the meaning of a word can vary depending on its context.

4.3 Flair Transformers

Transformers models (or FLERT - Document-Level Features for Named Entity Recognition) is an advanced approach to entity recognition in natural language processing (NLP) using transformer-based models. Transformers, such as BERT, GPT-3, and their variants, have revolutionized NLP tasks, including NER, by achieving state-of-the-art results due to their ability to capture contextual information effectively. Using a transformers model is adapting pretrained transformer model, such as XLM-RoBERTa (4), as your base model and fine-tuning it on NER dataset. During fine-tuning, it adapt to recognize entities by updating the model's weights with labeled data.

4.4 TARS

The TARS (task-aware representation of sentences) is a transformer-based model that enables training without the need for extensive training data, making it suitable for zero-shot learning or few-shot learning scenarios. What sets the TARS approach apart from traditional transfer learning methods is its consideration of semantic information inherent in the class labels. For instance, when analyzing acknowledgments, class labels such as "funding organization" or "university" inherently contain valuable semantic information.

4.5 XLNet

XLNet is a pre-trained language model that belongs to the family of transformer-based models, similar to BERT (Bidirectional Encoder Representations from Transformers). It was introduced by researchers in 2020 (6). XLNet combines the best of both worlds by utilizing autoregressive modeling (like in traditional language models) and autoencoder modeling. It maximizes the likelihood of predicting the next word in a sequence while considering all possible permutations of the input sequence. It aims to capture bidirectional context effectively, leading to improved performance on a wide range of NLP tasks, including text classification, question answering, and named entity recognition.

4.6 Evaluation methods

The models are evaluated in two steps:

1) F1-measure score, to check what percentage of entities have been delimited from the text.

2) Accuracy score, to check what number of named entities have been correctly assigned to categories. At this stage, if entities have been correctly identified it will be considered a success. If entities have been wrongly assigned it will be considered a failure.

A combination of the above measures can be used for the final evaluation of the models(for example an average or a weighted average).

5 Other literature review

Apart from joint work made by Smirnova and Mayr (1), which is going to be a main focus of our project, a proposal of the project consisted of two additional papers that introduced different approcaches to the NER problem. Because of lack of accessible code for each, we're not going to implement them. However, we found both of the papers very interesting and we believe it's worth to mention the existence of different techniques that can be used in NER.

5.1 CycleNER

The first one is an unsupervised approaches to NER problem. Among them there is CycleNER, introduced by Iovine et al. (2). This approach is based on cycle-consistency training that uses two functions: sentence-to-entity (S2E) and entity-to-sentence (E2S). Annotations are not required anymore as the model is trained on two independent sets - set of sentences (S) and set of example entities (Q) that can be labelled.

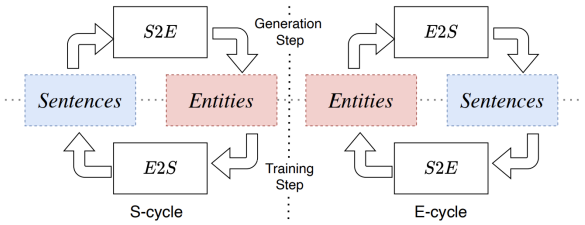

Figure 2: CycleNER training.

CycleNER training can be seen on Figure 2. It can be divided into two parts (so called cycles) and each of the parts consists of two steps. First cycle is S-cycle. From given sentence $s \in S$, we obtain q - a list of entities extracted from the sentence. Then in the second step we generate sentence s' out of q. The training based is on minimising cross-entropy between original sentence s and the recreated one s'. Similarly, the second part of training is E-cycle. Here, starting from q - a list of entities, we generate a sentence s. Next, we extract entities q' from s and compare q with q' in terms of cross-entropy.

The authors of the CycleNER demonstrates that their approach achieves competitive results to supervised methods on some of the most popular benchmark datasets. Even though it generally has a slightly lower accuracy, the gain is that we don't have to label datasets. Iovine et al. believe that in many cases (e.g. lack of domain knowledge or difficult access to the data) one may benefit from this the unsupervised approach.

5.2 Question Answering

The second approach has been introduced by Borst et al. in 2022 and is related to question answering (QA) (3). It focuses on the automatic recognition of entities that fund research work

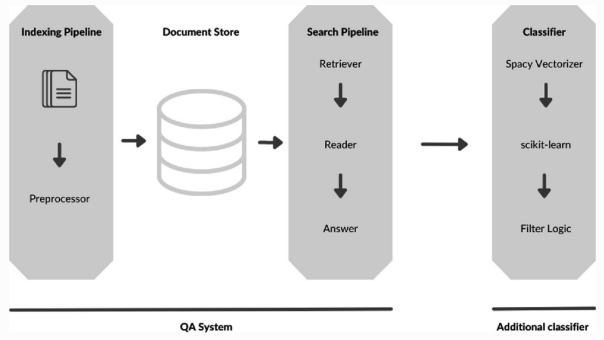

Figure 3: QA overview.

in economics as expressed in scientific publications. The primary goal is to identify funding information using a QA approach. The paper discusses several challenges, including the need to confirm that the QA approach outperforms manual indexing, disambiguation of funding organizations by linking their names to authority data, and integrating the generated metadata into a digital library application.

The authors focus on recognizing funding entities explicitly mentioned in acknowledgment phrases or sections of scientific papers. They employ a processing pipeline that includes extracting text from PDF documents, enriching it with metadata, and using pre-trained language models for question-answering. The architecture diagram can be found on Figure 3.

The study uses a dataset of open-access documents from the EconStor repository, associated DOIs, and funder information obtained via the Crossref and DataCite APIs. The F-scores of the examination, with and without the classifier, show promising results, with F-scores close to 0.8 for some language models. The classifier helps reduce false positives.

The paper highlights the feasibility of automatically extracting funding entities but mentions that the sample size was small. They acknowledge the need for a gold standard of manually checked funder information and the challenge of identifying and excluding open-access acknowledgments.

References

1. Smirnova, N., Mayr, P. Embedding models for supervised automatic extraction and classification of named entities in scientific acknowledgements. *Scientometrics* (2023)

2. Iovine, A., Anjie, F., Fetahu, B., Rokhlenko, O., Malmasi, S. CycleNER: An unsupervised training approach for named entity recognition. In Proceedings of the ACM Web Conference 2022 (WWW '22). Association for Computing Machinery, New York, NY, USA, 2916-2924.

3. Borst, T., Mielck, J., Nannt, M., Riese, W. (2022) Extracting Funder Information from Scientific Papers - Experiences with Question Answering. In G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, A. Poggi (Eds.), Linking Theory and Practice of Digital Libraries (Vol. 13541, pp. 289–296). Springer International Publishing.

4. Schweter, S., Akbik, A. (2020). FLERT: Document-level features for named entity recognition. ArXiv. 10.48550/arXiv.2011.06993

5. Akbik, A., Blythe, D., Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In 2018, 27th International Conference on Computational Linguistics (pp. 1638–1649).

6. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q., (2020) XLNet: Generalized Autoregressive Pretraining for Language Understanding