NER for acknowledgements
Project Report for NLP Course, Winter 2023

Sebastian Deregowski, Dawid Janus,
Bartosz Jamroży, Klaudia Gruszkowska
Warsaw University 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. This paper presents 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 (1) and train NER models using Flair Embeddings, BERT and XLNet models. According to our hypotheses, the described project showed that advanced deep-learning models perform better than basic machine-learning techniques. Moreover, a small dataset does not perform well during learning; however, if suplemented with additional data, the model's performance increases significantly. The final improvement in the models after using the silver set was most noticeable with the smallest original set, when it increased the F1 score from 0 to 51%. In other cases, the score also increased but not so dramatically. Flair Embeddings proved to be the fastest learner and achiever of all models taken for experimentation. However, models using transformers are very promising, although it takes several (4-6) hours to train them. Our research can contribute to the development of NLP and scientific text analysis.

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

1 Introduction

Named entity recognition is a fundamental task in natural language processing (NLP), aiming to identify and classify named entities into predefined categories such as people, organisations and many others. In various fields, such as healthcare, finance, law and science, correctly recognising 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.

Acknowledgements in scientific papers are a section where authors express gratitude and recognition to individuals, organizations, or funding sources that contributed to the research but may not be directly involved in the writing or analysis. This section typically appears near the end of the paper, before the references. Authors use acknowledgements to appreciate the support, guidance, technical assistance, or financial assistance received during the course of the study. It is a way for researchers to acknowledge the collaborative nature of scientific endeavors and show appreciation for those who played a role in the project's success. The importance of this section in papers has been growing for some time (2).

Our project aims to develop and evaluate named entity recognition (NER) models for identifying and classifying entities in acknowledgements. The plan was 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 evaluation. In addition, we wanted to try a different approach, namely test different models, and compare what results they can achieve. Lastly we created a silver standard set, a corpus made on top of 243 articles with automatic annotations provided by our newly 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 standard 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 dataset and further training models on it will improve performance.
- Proper names of corporations, especially with specific words, will be the most difficult to recognise.

1.3 Significance of the project

In our project, we used advanced NER models, especially those based on transformers, using the Flair library. We focused on identifying and classifying entities in scientific acknowledgements, which is a significant challenge in 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 area of NLP is crucial to improve the identification of relationships between scientists and assess the impact of financial and technical support on research outcomes.

We expect our findings to significantly impact 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 understanding key figures in the field of science. In addition, our research can influence how scientists collaborate, enabling a more efficient and organised knowledge exchange. This combination of advanced NER technologies with the analysis of acknowledgements is a step towards novel solutions in natural language processing and scientific research.

2 Related works

To this day, many different approaches have been developed in the NER field. In 2018, Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever presented a Generative Pre-Training model (12) that later turned out to be one of the state-of-the-art models in the NLP field including NER. GPT is an advanced language model based on the Transformer architecture, pre-trained on vast text datasets, enabling it to generate high-quality, contextually dependent text sequences. While not specifically tailored for Named Entity Recognition (NER), its adaptability allows for effective utilization in various language tasks, including NER, either through direct application or fine-tuning for specific tasks, showcasing its prominence in the field of natural language processing.

As the number of articles containing an acknowledgement section has been growing for a few years now (2), researchers have begun to look at methods to exploit the information they contain. Smirnova and Mayr (1) developed this theme in their work using tools from the Flair library (5). However, a few years earlier Wu, Jian & Wang, Pei & Wei, Xin & Rajtmajer, Sarah & Giles, C. & Griffin, Christopher, in their paper (3), have created a system to extract acknowledgement entities from articles in the database CORD-19 and classifies extraction results from open source NER packages that recognize people and organizations. This work has achieved a giant step compared to the initial work, which began with manual extraction of information. In such a way (4) Blaise Cronin, Gail McKenzie, Lourdes Rubio, and Sherrill Weaver-Wozniak extracted a total of 9561 peer interactive communication (PIC) names from a total of 4200 research sociology articles.

What is equally important as the selected model is to pay attention to the dataset quality on which the training is based. Creating a new one, however, is not easy and most often requires extensive domain knowledge. For the NER problem, the CoNLL-2003 (8) dataset is most often used as a benchmark. The dataset comprises four types of named entities: person, location, organisation, and miscellaneous. However, it does not meet all the needs we want to include in our work.

On the other hand, Vajjala and Balasubramaniam in their work (11) perform a broad evaluation of NER state-of-the-art models using different benchmark dataset - Ontonotes, that takes into consideration various text genres and sources constituting the dataset at hand. Three models taken into consideration - Stanza, Spacy, SparkNLP, performed very differently across various NE categories, with huge performance variation observed among the entity types in the dataset. Additionally, all models were very sensitive to small input perturbations and models performed poorly on genres unseen during training.

2.1 Dataset

Therefore, based on Smirnova and Mayr (1) example, we used their base dataset. It recognizes 6 types of entities:

1. funding agency,
2. grant number,
3. individuals,
4. university,
5. corporation,
6. miscellaneous.

The dataset has been divided into 4 different subsets, called corpora, that differ in size (see Table 2.1). The differences in corpora sizes would significantly affect the results obtained (more on that in Section 4).

Corpus	Training Samples
1	832
2	12,621
3	26,981
4	36,213

Table 1: Comparison of size of each corpus

As for the first classification task (entity/no entity), the division was quite even - 42.6% of words were entities, and 57.4% of words were not. Regardless of the corpus, the most popular entity types in acknowledgements were funding agency (FUND), grant number (GRNB) and individual (IND) (see: Figure 1). This will also have an impact on the results.

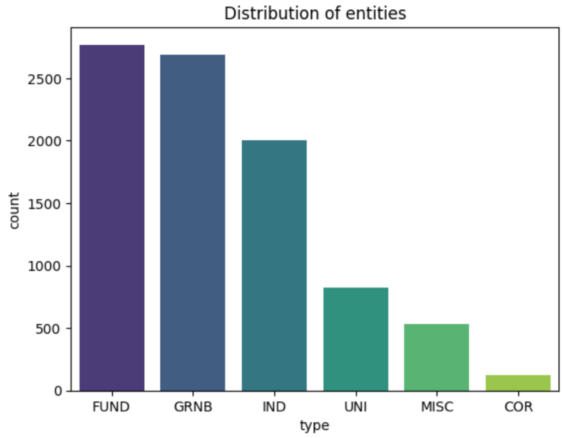

Figure 1: Distribution of entities appearances in corpora

Regarding differentiation within and between entity types, we checked cosine similarity between words by creating vector embeddings of each (see Figure 2). As we can see, words in the given entity group are not so similar to each other (the maximum value in the cosine similarity matrix is 0.27). Still, it appears that the words within a given group are more similar than words from other groups (although the differences in many cases are not significant). This is mainly because of the use of common words shared between some of the groups. For instance, the word *of* is among the top 5 most frequent words in three out of six entity groups. This makes the entity recognition task

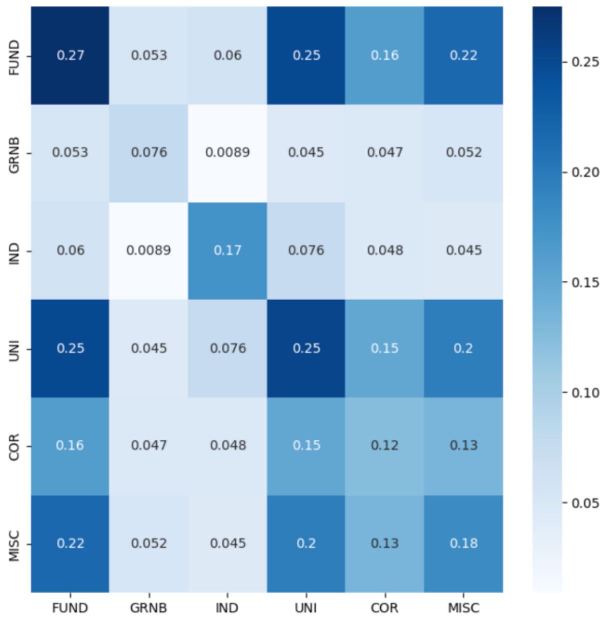

Figure 2: Heatmap of cosine similarity between words with regards to entity type

harder, as models need to evaluate the context of all entities and not just look at single words.

Speaking of words within a given entity, we also checked the average length of a given entity to see whether this may become a clue for the model (see Figure 3). Not surprisingly individuals have, on average, two words (name and surname). The remaining entity types have similar distributions, with funding agency and miscellaneous having the longest tails.

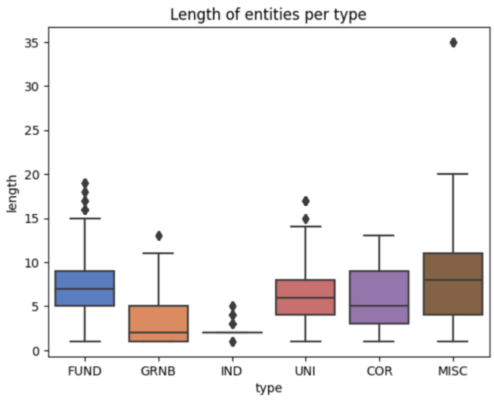

Figure 3: Distribution of length of entities with regards to entity type

3 Approach & research methodology

3.1 Models

We built our work upon three different models that are considered to be state-of-the-art solutions for named entity recognition task.

3.1.1 Flair Embeddings

NER model with Flair Embeddings is a method which create contextual word embeddings for text data (6). 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 is achieved through the use of two-way Long Short-Term Memory (LSTM) models, enabling Flair to capture complex relationships between words, which makes it 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.

3.1.2 BERT

Apart from Flair Embeddings, Smirnowa and Mayr, in their work, used also another Flair model, this time the one based on transformers approach (1). In our work, we tried a different transformer-based model, BERT (Bidirectional Encoder Representations from Transformers) (10). This model use attention mechanism, which is a way for a model to assign weight to input features based on their importance to some task. Before feeding word sequences into BERT, 15% of the words in each sequence are replaced with a mask token. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence.

3.1.3 XLNet

Lastly, we proposed a different language model, XLNet (7). Introduced in 2020, it also belongs to the family of transformer-based models. 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 named entity recognition. For the purpose of the project we used two versions of XLNet - base (the lighter version) and large. Large, although obtaining better results, trained very long compared to the other methods, therefore, base was also used in order to compare the time/results tradeoff – more on that in section 4.

3.2 Evaluation methods

Named entity recognition is a task that can be divided into two steps - extracting entities from plain text and providing a category for given entity. This can be considered as two classification tasks - first 0-1 classification (entity, no entity) and then multiclass classification (entity A, entity B, etc.). Similarly to Smyrnova and Mayr's approach, we can treat these two classification tasks as one - if we consider no entity as a separate class in multiclass classification.

Therefore, we evaluated the models' performance with an F1 score, as it's a very good metric for multiclass classification tasks.

3.3 Silver set

One of the research questions was about the impact of adding new data on the model results. Because of the specific setup of the dataset, we were not able to find anywhere on the Internet yet another dataset to enlarge our training set. Instead, we decided to enrich the gold standard set provided by Smyrnova, with a silver standard set. The key difference is that the goal standard is made or approved by some domain expert, while the silver standard set allows automatic or semi-automatic ways to create labels, in this example labels of entities for text in acknowledgement sections.

Therefore, our plan was first to train a model based on a gold standard set provided. Additionally, we manually collected acknowledgement sections of 243 papers from the science domain. Next in line, our plan was to create labels for those 243 acknowledgement sections based on predictions from the model we trained before. Lastly, we wanted to verify the hypothesis that having more data (even not only in the gold standard, but also in silver) can be beneficial in terms of the model's efficiency. This is described in more detail in section 4.

3.4 Equipment and devices

Due to the use of models requiring high computing resources, we were not able to carry out experiments and model fine-tuning processes on our own hardware. Instead, all work was carried out on a dedicated project in the Google Cloud Platform. There, we set up a virtual machine (16 vCPUs, 64 GB RAM) with the Jupyter Notebook environment.

4 Experiments and Results

The goal of the first set of experiments was to train the Flair Embeddings model on different corpora and compare the results. As mentioned in section 3, corpus 1 contains the least data, while corpus 4 the most. Each corpus has been split into training, validation and test sets. The setup is shown in Figure 4. The general conclusion is - the smaller the corpus used for training, the worse the results of the model. The F1 scores for all the corpora can be found in Table 2.

Corpus	F1 Score
1	0%
2	81%
3	83%
4	83%

Table 2: Comparison of results of F1 scores for each corpus

As the amount of data in the corpus increases, the model achieves a better result. However, for corpus 3 and 4 there is no significant difference in the model performance.

The Flair Embeddings model trained on corpus 4 was then used to create the silver standard corpus. The intention was to select the model trained on the most numerous corpus in order to create a silver corpus of the best possible quality. This setup can also be seen in Figure 4.

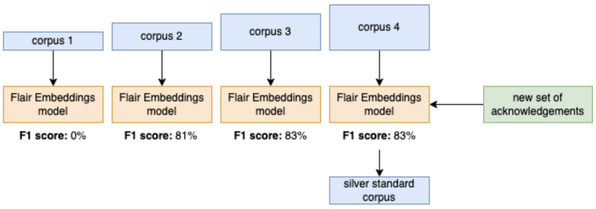

Figure 4: Flair Embeddings training setup

The goal of the second set of experiments was to see how the silver standard set affects the results. For two different corpora, we trained the model with and without a silver standard set. The silver set was added to the train set of the given corpus, the validation and test sets remained the same for each corpus. For corpus 1, we used the Flair Embeddings model, while for corpus 4 - XLNet-large. The setup and results are shown in Figure 5 and Table 4, respectively.

Corpus	Silver set	Model	F1 score
1	no	F. Embeddings	0%
1	yes	F. Embeddings	52%
4	no	XLNet-large	79%
4	yes	XLNet-large	81%

Table 3: Comparison of the results obtained by training models directly on the corpus and combined with the silver standard set

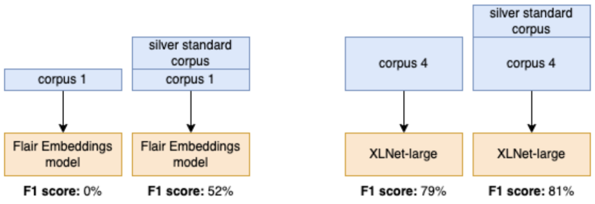

Figure 5: Comparison of the effectiveness of models trained directly on the corpus vs. trained on the corpus combined with the silver standard set

The use of the silver standard set in training models significantly improves results. For corpus 1, the model finally started to learn. F1 score increased from 0% to 52%. For corpus 4, there was also an increase from 79% to 81%.

The results of training history (both losses and F1 scores) can be seen in Figure 6. As we can see, the training with a silver set gives better results starting from the very first epoch.

The third set of experiments was performed in order to compare the performance of the Flair Embedding model wiht respect to entities. The worst results were always recorded for the classes *Miscellaneous* and *Corporation*; this is due to the fact that these are the two least numerous classes (compare with Figure 1). The results are shown in Figure 7. Flair Embeddings trained directly on Corpus 1 is not included in the graph, as its F1 score was 0.

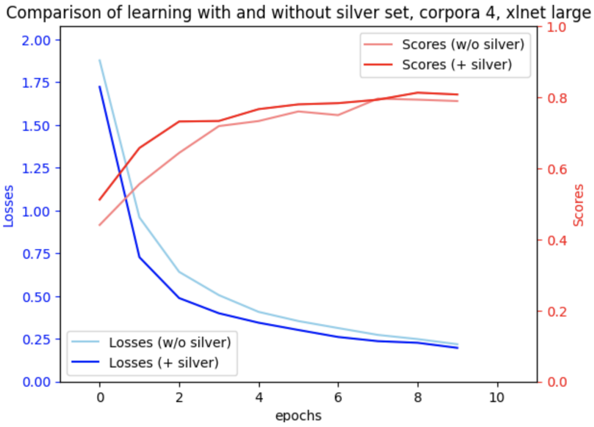

Figure 6: Graph of the loss decline and score growth of the xlnet large model as iterates over the training epochs directly on corpus and combined with the silver standard set

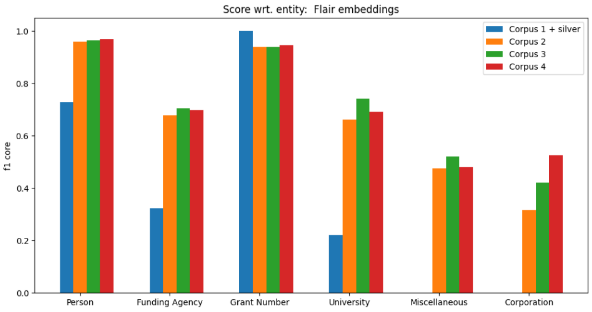

Figure 7: Comparison of F1 scores for Flair Embedding model for each class and corpus

The fourth set of experiments shows a comparison of different models, Flair Embeddings, XLNet-large, XLnet-base and BERT on the same corpus 4. The setup and results can be found in Figure 8 and Table 4, respectively.

Model	Epochs	F1 score
F. Embeddings	78	83%
XLNet-large	10	77%
XLNet-base	20	80%
BERT	20	79%

Table 4: Comparison of results of training different models on corpus 4

As we can see, the best result was obtained by Flair Embeddings model, with XLNet-large achieving the lowest F1 score. However, it should be noted that from the above results, we cannot make any conclusions, as they are biased by the

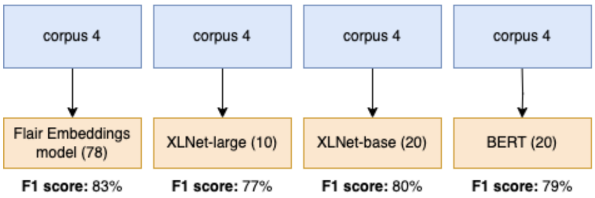

Figure 8: Setup of fourth experiment

different number of epochs. The original intention was to train the models on the same number of 20 epochs. However, XLNet-Large is a more complex model, requires more computational resources and takes longer to train (around half an hour per epoch), which is why it was trained on 10 epochs.

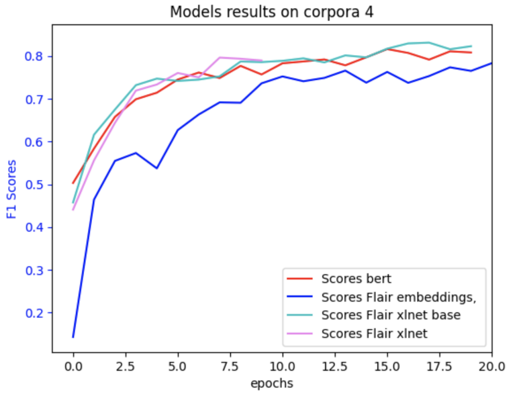

Figure 9: Graph of score growth during training models on corpus 4, iterates over the training epochs

The FLair Embeddings model, on the other hand, is the lightest model with very fast training (2-3 minutes per epoch), so the number of epochs was chosen as the number after which the model did not improve in the next iteration (the learning rate decreased from 0.1 to 0.00001 by an order of magnitude after 3 epochs without improvement). For this experiment, the last epoch was the 78th one. The scores history while training can be found in Figure 9 and the losses history are in Figure 10.

As we can see, the XLNet-large stops after the 10th epoch, while Flair Embeddings run further than 20 epochs but still, we are able to see that the latter one, although in the end achieving the highest result, definitely learns the slowest. On the other hand, on the 10th epoch, XLNet-large obtained the highest F1 score on the validation set. This may be a hint that if given a required amount of time, this model may outperform the others. It's also worth noticing that BERT always had the lowest loss out of all four models.

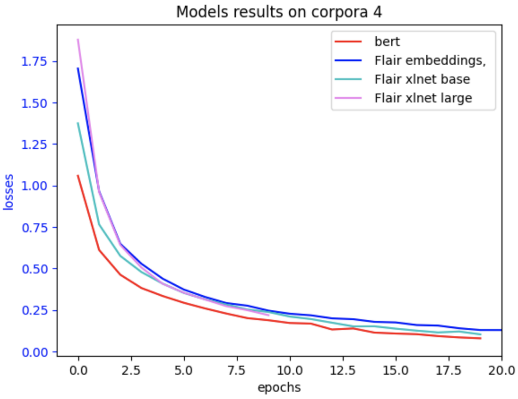

Figure 10: Graph of loss decline during training models on corpus 4, iterates over the training epochs

Lastly, 5th experiment shows the results of the models from the perspective of specific entities on corpus 4. The worst results usually occur again for the classes the least populated classes (see Figure 11). The comparison also includes separately the XLNet-large models trained on corpus 4 with and without the silver standard set. As we can see, XLNet-large model with silver standard included in training performs the best when it comes to *Funding agency* and *Corporation* classes, while performing a bit poorer than others in *Grant number*.

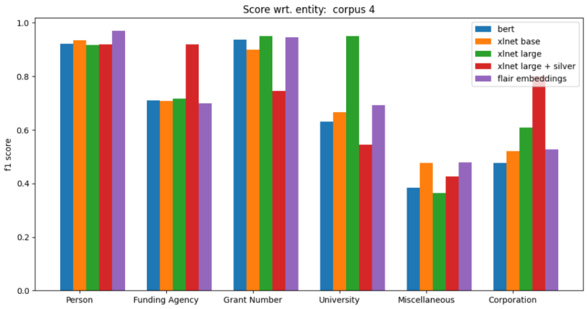

Figure 11: Comparison of F1 scores for each class and model

5 Discussion on your results

We believe that we have achieved most of the initial objectives of the project and can answer the research questions posed:

- What techniques and models in the field of NLP are most effective for recognising named individuals in a specific area?

  All the models we tested achieved good results, both transformer-based (BERT and XLnet) and from the flair library (see Table 4). It seems to us that XLnet large trained on more epochs would achieve the best results.

- How does the performance of NER models change with different types and amounts of training data?

  As expected, the more data, the better the training process. (see Tables 2.1 and 2). By adding our silver standard set we not only improved the decreasing of loss function but also improved the F1 score on the test set (see Figure 6).

- 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?

  Because of our lack of knowledge about NLP data preprocessing, we thought that such techniques could improve model performance. However, after the research we made, it turned out that these techniques are not used in NER task, because in English language those words are in their proper noun form. Additionally, removing stop words can also affect the model, because in some entity types stop words are important (e.g. grant number). Finally, they can interfere the context of sentences and as the models can understand the context, it's better to leave them in their original form.

- Can preparing a silver set and further training models on it improve their effectiveness?

  Yes, definitely. Although the silver standard is not as precise as the gold standard (which requires a domain expert), it still can improve the effectiveness of the model (taking into account around 80% F1 score). This provides an opportunity to first train model on relatively small amount of gold standard data and then based on this generate huge amount of silver standard data, which then can be used to train a better model. That approach may be beneficial when we don't have a domain expert or he/she is unable to manually label vast amount of data.

- Which types of entities are the most difficult to classify?

  Not surprisingly, the hardest class is miscellaneous which contains all the entities not classified to any other class (see Figure 7. Models had also struggles with corporations but this is mostly because of the least amount of examples in the training data.

6 Conclusions and future work

Considering the wide field of this work and our current results, the works can surely be continued. The results obtained are satisfactory, however, large models such as XLnet-large have not been trained (due to limited equipment and time) on the optimal number of epochs. Therefore, the first step of improvement would be to train this model on a better virtual machine and re-evaluate the results, also with the silver standard set.

In order to increase the accuracy of the models, it is worth considering how to improve the effectiveness of the worst classes. According to the current results, these entities are Miscellaneous and Corporation. This may be due to the fact that the least data was found for these classes. Taking this into account, the next step would be to try to balance the training dataset and see if the results are better. In addition, one should check which units were misclassified most often and do more work to identify them better.

In the results obtained, the improved performance of the model on the combined dataset with the silver set is evident (from 0 to 0.52 on corpus 1 and from 0.79 to 0.81 on corpus 4 in terms of F1 score). In order to avoid manually extracting acknowledgements from scientific articles, it would be good to prepare a script or application that with the help of LLM model's API (such as GPT-3), automatically reads them and processes them into a file ready for model learning.

The culmination of this work can be the creation of a user interface to load a custom acknowledgements section and it will be returned with recognised and classified entities.

	D.S.	J.B.	J.D.	G.K.
Research	4	4	4	4
EDA	4	4	1	1
Initial modeling	2	2	9	9
Silver set	2	6	2	6
Final models	12	8	11	7
Evaluation	4	4	1	1
Documentation	4	4	4	4
Sum	32	32	32	32

Table 5: Estimated workload per person and phase (in hours).

References

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

2. Cronin, B. Acknowledgement trends in the research literature of information science, Journal of Documentation, Vol. 57 No. 3, pp. 427-433 (2001)

3. Wu, J., Wang, P., Wei, X., Rajtmajer, S., Giles, C., Griffin, C. Acknowledgement Entity Recognition in CORD-19 Papers (2020)

4. Cronin, B., McKenzie, G., Rubio, L., Weaver-Wozniak, S. Accounting for influence: Acknowledgments in contemporary sociology. Journal of the American Society for Information Science (1993)

5. Akbik et al., FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP, NAACL (2019)

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

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

8. Tjong Kim Sang De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition (CoNLL 2003)

9. 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.

10. Devlin, J., Chang, M., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018)

11. Vajjala, S., Balasubramaniam, R. What do we Really Know about State of the Art NER?, *arXiv preprint arXiv:2205.00034*, 2022.

12. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., Improving Language Understanding by Generative Pre-Training (2018)