Semiautomatic Data Generation for Academic Named Entity Recognition in German Text Corpora

Pia Schwarz
Leibniz Institute for the German Language (IDS)
R5 6-13, 68161 Mannheim, Germany
`schwarz@ids-mannheim.de`

Abstract

An NER model is trained to recognize three types of entities in academic contexts: person, organization, and research area. Training data is generated semiautomatically from newspaper articles with the help of word lists for the individual entity types, an off-the-shelf NE recognizer, and an LLM. Experiments fine-tuning a BERT model with different strategies of post-processing the automatically generated data result in several NER models achieving overall F1 scores of up to 92.45%.

1 Introduction

The Leibniz Institute for the German Language (IDS) hosts the German Reference Corpus DeReKo (Kupietz and Keibel, 2009; Kupietz et al., 2010, 2018), the largest German collection of texts available for research, consisting of 57 billion tokens as of March 2024 (Leibniz-Institut für Deutsche Sprache, 2024). The corpus contains texts from the 18th century to the present, including many press releases. Linguistic annotation for DeReKo is provided on a syntactic level (e.g. parts of speech, lemmata, dependency relations), however, no semantic annotation has been added yet. This work concentrates on the annotation of three types of named entities, in particular persons in academia, academic institutions, and academic disciplines. In order to fine-tune a BERT model (Devlin et al., 2019), training data is collected in a semiautomatic manner from DeReKo itself[1].

2 Motivation

DeReKo is searchable via the corpus analysis platform KorAP (Diewald et al., 2016), making it possible to retrieve linguistic annotations as well as descriptive catalog metadata. These include specifications about the title, creation date, author, license, corpus sigle, and text sigle. A sigle is a unique identifier to reference parts of the corpus, in the case of newspaper texts, a text sigle refers to a single newspaper article. This level of granularity makes it possible to enrich DeReKo with semantic metadata such as named entities on the level of individual texts. Finding mentions of academic named entities in newspaper texts might serve as a starting point to investigate the impact or perception of academics beyond research. Moreover, these entities can also serve as links to external knowledge bases such as Wikidata (Vrandečić and Krötzsch, 2014), the Research Organisation Registry (Lammey, 2020) or the German National Library's Integrated Authority File (Behrens-Neumann and Pfeifer, 2011). Links to such external knowledge bases would provide more context to the data in DeReKo.

Creating a model for the task of academic NER requires training data, namely sentences tagged with the three given types. To our knowledge, no such data set exists, so a new one is generated from scratch. Having DeReKo at hand as a high-standard text collection, which at the same time constitutes the real-world data that should be processed by the resulting named entity recognizer, it is an obvious choice to collect sentences from the corpus as training data. The academic NER model should be able to tag literal mentions of the three entity types independent of whether researchers work in academia or in the industry, for example:

(1) *...sagt [Heitzer]$_{\text{PER-RES}}$, Professorin am Lehr- und Forschungsgebiet [Didaktik der Mathematik]$_{\text{AREA-RES}}$ an der [RWTH]$_{\text{ORG-RES}}$.*
'...says [Heitzer]$_{\text{PER-RES}}$, professor of

[1] We release best scoring NER model via WebLicht (Hinrichs et al., 2010) at `https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tools_in_Detail#Named_Entity_Recognition`
Due to strict copyright agreements with our text providers we can provide the data for scientific and non-commercial purposes only after signing a license agreement (free of charge, upon request via E-Mail).

Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), pages 173–181
September 10-13, 2024 ©2024 Association for Computational Linguistics

the teaching and research department [Didactics of Mathematics]$_{\text{AREA-RES}}$ at [RWTH]$_{\text{ORG-RES}}$.'

(2) *Ein paar Stockwerke höher wartet [Astrid Kiermaier]$_{\text{PER-RES}}$ auf uns, die Molekularbiologin arbeitet bei Roche im Bereich [Krebsforschung]$_{\text{AREA-RES}}$ ...*

'A few floors up, [Astrid Kiermaier]$_{\text{PER-RES}}$ is waiting for us, the molecular biologist works in the area of [cancer research]$_{\text{AREA-RES}}$ at Roche ...'

The entity type PER-RES should include the academic title of a person if it precedes the name. However, the model is not expected to resolve coreferences, so neither pronouns referring to an entity nor a noun phrase that does not literally mention the person's name should be tagged, as the following two examples illustrate:

(3) *Mitte März begann ein Team von Forschern der [Universität Hirosaki]$_{\text{ORG-RES}}$ damit, sodass sie im Norden Japans bereits Messungen vor Ort durchführten.*

'In mid-March, a team of researchers from [Hirosaki University]$_{\text{ORG-RES}}$ began with that, such that they already conduced on-site measurements in northern Japan.'

(4) *Der Physiker erfand nicht nur die Luftpumpe, sondern befaßte sich auch mit...*

'The physicist not only invented the air pump but also engaged in...'

Researchers are not always mentioned within the context of research, in example (5), the model is supposed to tag the person as the academic title provides enough context to identify someone who is or was a researcher. The opposite holds for example (6), where a literal mention of the exact same person is not supposed to be tagged as neither an academic title nor the rest of the sentence indicate any academic context. This is also the case for example (7), where the model is not expected to tag researcher Jane Goodall due to the lack of context information.

(5) *[Dr. Frank-Walter Steinmeier]$_{\text{PER-RES}}$, Chef des Bundeskanzleramtes, ist dafür verantwortlich...*

'[Dr. Frank-Walter Steinmeier]$_{\text{PER-RES}}$, head of the Federal Chancellery, is responsible for...'.

(6) *Außenminister Frank-Walter Steinmeier gab sich weiter diplomatisch.*

'Foreign Minister Frank-Walter Steinmeier continued to maintain a diplomatic stance.'

(7) *Schütze, was du liebst - So lautet das Prinzip der Umweltikone Jane Goodall.*

'Protect what you love – This is the principle of environmental icon Jane Goodall.'

The question remains as to how to tag the data without spending too much human resources on annotation but at the same time not compromising on quality either. The goal is to collect enough training data to fine-tune a BERT model in an at least partly automated manner through a rule-based method with word lists and then to improve the model by generating more training data through a deep learning approach using a large language model (LLM).

3 Background

Named entity recognition is a crucial method in NLP and forms part of many downstream tasks. Standard models typically comprise at least the entity types person, location, and organization, but there is also quite some research about domain-specific NER models, dealing for example with biomedical entities such as proteins or chemicals (Lee et al., 2020; Sun et al., 2021). Relevant for the present work are standard NER models and frameworks, especially the spaCy library (Honnibal et al., 2020) for model fine-tuning, and Stanza (Qi et al., 2020) for data preprocessing. Although spaCy and Stanza both provide state of the art NER models, they do have weaknesses once they are more thoroughly evaluated, e.g. regarding unseen text genres during inference or random train/dev/test splits during training (Vajjala and Balasubramaniam, 2022). However, Schmitt et al. (2019) compared the five NER frameworks StanfordNLP, NLTK, OpenNLP, spaCy, and Gate with the result of StanfordNLP scoring best. The Stanza NLP package builds on the Stanford NLP framework and gives access to NER models for multiple languages which is why its German model was used for data preprocessing.

To our knowledge neither a German data set nor a readily trained model is available for the domain of academic entity recognition covering the entity types academic person, institution, and research area. The only data set that comes close to the present task is CrossNER (Liu et al., 2021), which contains 14 entity types, including labels for universities, scientists, and scientific disciplines. However, CrossNER does not contain any German data, and the part of the data set containing the relevant labels is very small, containing a few hundred samples only, in addition to being extracted from a single specific domain of Wikipedia articles about Artificial Intelligence, which might be insufficient for applying the task of NER to the broad domain of newspaper texts. Peng et al. (2020) propose an approach for adapting existing NER models such that they recognize additional entity types. Their partially supervised training algorithm makes use of word lists with prototypical examples for the new entity type to be added. Although the evaluation for some of their data sets looks promising, their method of introducing new types of named entities is not really applicable to the present task. Only in the case of research area, a *new* entity type would be added, whereas the entity types academic person and institution depict an *adaption*, as persons and institutions in academics are a subset of the more general entity types person and organization that most existing NER models have. However, the idea of bootstrapping training data with word lists was indeed inspired by their work. Gilardi et al. (2023) conduct experiments where they let ChatGPT annotate data sets and compare the results to the annotation performance of human crowd workers. The humans receive the same instructions as the LLM (in a zero-shot setting) for the text annotation tasks comprising binary and multi-class classification of sentences. Results show that the LLM outperforms the crowd workers by approximately 25 percentage points in average accuracy. Under the aspect of labeling cost reduction, Wang et al. (2021) experiment with distinct strategies of applying GPT-3 to label various NLP data sets. They use labels generated by the LLM to train smaller and thus more specialized transformer language models and compare these to the raw GPT-3 model as well as to human labeling performance. It turns out that the combination of letting humans adjust low-confidence labels of GPT-3 works best.

4 Approach

The steps to obtain a custom NER model recognizing academic entities comprise the following: (i) For each entity type, create lists of prototypical entities or words that form part of candidate entities. Detect candidate entities in the corpus text by applying a German off-the-shelf NER model and the word lists. (ii) Manually post-process enough sentences to obtain sufficient training data for an initial data set and fine-tune a German BERT$_{\text{BASE}}$ model to obtain a custom NER model. (iii) Generate more training data by applying the custom NER model and an LLM on unseen data in order to again fine-tune the German BERT$_{\text{BASE}}$ model with the initial plus the additional data.

At the last step, various experiments with the additional data – post-processed in different ways – show possible uses of this extra data and evaluate how well they work. These different variants of data post-processing result in three additional data sets for retraining: One data set containing only the extra sentences tagged by the initial custom NER model, a second one with only the tags on which the LLM and the initial custom NER model agree, and a third one being the manually post-processed version of the second data set. Each of the additional training data sets results in a new fine-tuned custom NER model, respectively. Finally, we compare the three additional custom NER models and the initial custom NER model from step (ii).

5 Data

In order to filter DeReKo for a suitable initial data set, a few preprocessing steps are necessary. The word lists are created as a starting point to find sentences that contain one of the three relevant entity types. The first list, used to search for potential academic persons, contains words or abbreviations representing academic titles such as *Dr.*, *Professor* or *PhD*. The second list contains names of academic institutions, mainly based on a list provided by the German Federal Report of Research and Innovation (Bundesbericht Forschung und Innovation, 2023). The third and last one lists areas of research, inspired by the German Research Foundation's classification of research fields (Deutsche Forschungsgemeinschaft, 2023). The word list with academic titles further serves in a previous step to filter DeReKo for potentially relevant texts, which becomes necessary due to the sheer size of the corpus. We use this word list assuming that texts in which academic titles appear might contain mentions of academic institutions and research areas as well. Whereas all three word lists are used to find candidate entities through string matches, the candidate entities for the entity type academic person were detected with the additional help of an off-the-shelf NER model from the Stanza NLP package, applying the condition that only a named entity of type person having a preceding or succeeding academic title becomes an actual candidate entity.

Out of more than 340,000 filtered texts, 10,000 are randomly selected to automatically find candidate entities. A subsequent manual post-processing[2] with the deletion or correction of wrong entities and the insertion of missed entities, yields a total of 4,928 sentences with 4,223 tags for academic persons (PER-RES), 2,300 tags for academic institutions (ORG-RES), and 676 tags for research areas (AREA-RES). The manual review of all three entity types comes with some challenges. Regarding candidate persons, for example, there are many cases in which schoolteachers (teaching in secondary but not tertiary education) were erroneously tagged as academics because of their preceding title of professor in the sentence. This happens in Austrian newspaper texts, where the convention holds to use this kind of title for schoolteachers who studied at university. Similar are the cases of detected academic persons from fiction or pen names such as *Dr. Seuss*. A weakness of the Stanza NER model is the incorrect recognition of first and last names with hyphens, which are both quite common for German names, e.g. *Prof. DDr. Franz-Josef Radermacher* or *Prof. Barbara Städtler-Mach*. Another problem is that academic persons sometimes stay undetected in sentences in which their academic title does not occur, even when the context is unambiguously academic, e.g.:

(8) *...der Neurobiologe Mathias Jucker vom Hertie-Institut der Universität Tübingen...*

‘...the neurobiologist Mathias Jucker from the Hertie Institute of the University of Tübingen...’

This example also illustrates the problem of how to deal with hierarchical relations between academic institutions – in this case whether to tag both the *Hertie-Institut* as well as the *Universität Tübingen* or only the latter. Both were tagged eventually as *Hertie-Institut* unambiguously refers to the subordinate organization, which is not the case for mentions such as *Faculty for Computer Science*. Instead, *Computer Science* would be tagged as an entity of the type research area. Another tagging decision for research areas is to handle two areas as a single entity when they appear in one compound expression connected with a hyphen, e.g. *Wirtschafts- und Sozialwissenschaft* (‘economic and social science’). Although a good amount of the work can already be done automatically, these edge cases illustrate that manual post-processing remains an essential step to obtain data of good quality.

Data Set	A	B	C	Initial
PER-RES	5,421	4,157	3,774	2,942
ORG-RES	2,826	2,076	2,136	1,624
AREA-RES	1,157	726	749	450
# Sentences	6,768	5,089	4,533	3,449

Table 1: Training data statistics of the initial and the three additional data sets. Note that the number of sentences of the initial data set was originally 4,928 but is reduced by the development and test data.

5.1 Additional Data Sets

To further improve the custom NER model, we generate more training data with the help of the initial custom NER model and an LLM, both applied to tag additional sentences from 1,000 unseen DeReKo texts. The few-shot prompt for the LLM is provided in Appendix A.1. The decision as to which LLM to use is made in favor of *Llama-2-13B-chat* after experimenting with different instructional prompts as input to compare the two models *Llama-2-13B-chat* (Touvron et al., 2023) and *OpenOrca-Platypus2-13B* (Lee et al., 2023). See Appendix A.2 for further details. The three additional data sets created with the initial custom NER model and *Llama-2-13B-chat* all contain the training data from the initial data set *plus* the newly generated data. They differ from each other with respect to the newly generated data as follows:

A) contains sentences with tags detected by the initial custom NER model

B) contains sentences with tags on which the initial custom NER model and *Llama-2-13B-chat* agree

C) contains sentences from B) with manually reviewed tags (deleted, inserted or corrected)

Table 1 provides an overview of the different training data set sizes and the distribution of the three entity types. The biggest data set is data set A, followed by B and finally C, corresponding to the increasingly stricter measures of quality assurance.

[2] Only the author of this paper reviewed the data manually due to practical considerations. While acknowledging the importance of inter-annotator agreement as a measure of reliability, involving external annotators was not feasible within the given time frame and budget constraints for this research.

6 Experiments

The German cased $\text{BERT}_{\text{BASE}}$ model *de_dep_news_trf* consisting of 12 layers with 12 attention heads each and a total of 768 hidden states is fine-tuned separately with each of the four data sets using the spaCy transformer library on a single Tesla P4 GPU. To obtain the same development and test data for the four passes of fine-tuning $\text{BERT}_{\text{BASE}}$, the initial data set is split into train/dev/test portions with a ratio of 70/20/10. For better comparability, all hyperparameters for model training are kept identical and correspond to spaCy's default configuration with a batch size of 128, a dropout rate of 0.1, the Adam optimizer with an initial learning rate of 10^{-5}, and early stopping based on the F1 score.

7 Results and Discussion

We evaluate all four models on the test split consisting of 489 sentences containing 423 PER-RES, 192 ORG-RES, and 79 AREA-RES tags. Table 2 shows that there are only few differences between the model performances, all ranging within overall F1 scores of 91.32% and 92.45%. The initial custom NER model reaches the best score, which is slightly surprising as it is trained on the smallest data set. Intuitively, the expectation would be that model C (trained on data set C) would yield the best score as it comprises roughly 30% more sentences that are, on top of that, manually reviewed. However, model C is only ranked third, even slightly behind model B without the manual review but trained on more sentences. Model A with the strategy to augment the data only using the initial custom NER model yields the worst scores, not only regarding the overall F1 score but also the F1 scores for the individual entity types. A possible explanation might be the missing quality checks for the data, as training data is neither double checked by an LLM nor by a human. It seems to be an insufficient strategy to only increase the amount of data without any measures of quality assurance.

Regarding the best model, the picture changes a bit when taking a look at the entity type F1 scores. While the best score for the entity type PER-RES of 95.4 is still achieved with the initial custom NER model, model C achieves the best score for the entity type ORG-RES, and model B does so for AREA-RES. Thus, with the experiments in this work it cannot be stated that there is clearly one single data augmentation strategy for all entity types.

Model		A	B	C	Initial
PER-RES	P	91.56	90.61	92.39	93.68
	R	96.49	97.19	96.72	97.19
	F1	93.96	93.79	94.51	**95.40**
ORG-RES	P	90.91	92.06	92.15	89.58
	R	87.63	89.69	90.72	88.66
	F1	89.24	90.86	**91.43**	89.12
AREA-RES	P	82.35	86.59	79.76	89.47
	R	82.35	83.53	78.82	80.00
	F1	82.35	**85.03**	79.29	84.47
Overall	P	90.30	90.53	90.86	92.12
	R	92.35	93.48	92.92	92.78
	F1	91.32	91.99	91.88	**92.45**

Table 2: Precision, recall and F1 scores (in percent) for the individual entity types and overall.

Except for the entity AREA-RES in example (2) all entities listed in section 2 are all correctly recognized by the best (initial) model. To test a few cases that are presumably more difficult for the model, we modify the examples (1) and (2) by cutting off the second half of the sentence after the last comma. In both cases the person entities should not be tagged anymore due to the lack of context. The model does so for example (2) but not for (1) where *Heitzer* keeps beeing tagged as PER-RES. For sentence (7) we replace the tokens *Umweltikone* ('environmental icon') for *Primatenforscherin* ('primatologist'), which changes the model's behavior as it now tags *Jane Goodall*.[3]

8 Conclusion

This work shows different strategies of generating training data to obtain a custom NER model through fine-tuning. For the sake of obtaining high-quality data, suitable data is augmented semiautomatically, with some amount of sentences undergoing manual review. The results show that there is no single best data generation strategy for all entity types, such that a combination for the three best-scoring models might be considered for future applications with the specific domain of academic named entity recognition. With the small differences of the resulting F1 scores in mind, a careful conclusion that can be drawn is that LLMs like *Llama-2-13B-chat* are beneficial to ensure data quality at a low cost whereas it might not be worth to invest too much into manual data review.

[3] See Appendix A.3 for all examples and their variations.

9 Limitations

There are several possible improvements for future model fine-tuning, one of which is to see whether a different train/dev split of the three additional data sets would lead to better results and how other/newer LLMs like GPT-4 or Llama-3 might show improvements for data preprocessing. Another idea is to qualitatively evaluate the results of the best model more thoroughly and investigate if wrong model predictions follow certain patterns (e.g. research areas composed of many words are often not well recognized) and if so, generate more training sentences targeted to eliminate these error patterns. Finally, it would be interesting to know by how much the initial data set can be reduced without compromising much on model performance in order to find a good balance between the amount of manually annotated and automatically generated data to further reduce manual annotation cost.

10 Ethical Considerations

For the purpose of this contribution, the authors received access to data files from DeReKo. Due to copyright restrictions the sampled data set can only be made available under certain conditions, for further details see section 1. However, interested parties can easily register for the corpus analysis platform KorAP[4], which allows to query DeReKo as a whole. We do not see any data privacy issues as the texts from which the training data is sampled have all been previously made available by (newspaper) publishers.

11 Acknowledgements

This publication was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

[4] The corpus can be queried either through KorAP's API or its web client: https://korap.ids-mannheim.de/

References

Renate Behrens-Neumann and Barbara Pfeifer. 2011. Die Gemeinsame Normdatei: ein Kooperationsprojekt. *Dialog mit Bibliotheken*, 23(1):37–40.

Bundesbericht Forschung und Innovation. 2023. Wissenschaftseinrichtungen: Liste der Einrichtungen. https://www.bundesbericht-forschung-innovation.de/de/Liste-der-Einrichtungen-1790.html. Accessed: 19.04.2024.

Deutsche Forschungsgemeinschaft. 2023. Fächerstruktur der DFG. https://www.dfg.de/de/foerderung/antrag-foerderprozess/interdisziplinaritaet/faecherstruktur. Accessed: 19.04.2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nils Diewald, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt. 2016. KorAP Architecture - Diving in the Deep Sea of Corpus Data. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3586–3591. Paris: European Language Resources Association (ELRA) 2016, Portoroz, Slovenia.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *arXiv preprint arXiv:2303.15056*.

Erhard W. Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. SpaCy: Industrial-strength Natural Language Processing in Python. Zenodo (online).

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marc Kupietz and Holger Keibel. 2009. The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In Makoto/Kawaguchi Minegishi, editor, *Workings Papers in Corpus-based Linguistics and Language Education*, volume 3, pages 53–59. Tokyo University of Foreign Studies 2009, Tokyo.

Marc Kupietz, Harald Lüngen, Pawel Kamocki, and Andreas Witt. 2018. The German Reference Corpus DeReKo: New Developments – New Opportunities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4353–4360, Miyazaki, Japan. European Language Resources Association (ELRA).

Rachael Lammey. 2020. Solutions for Identification Problems: A Look at the Research Organization Registry. *Science Editing*, 7(1):65–69.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, Cheap, and Powerful Refinement of LLMs. *Preprint*, arXiv:2308.07317.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, 36(4):1234–1240.

Leibniz-Institut für Deutsche Sprache. 2024. Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2024-I (Release of 13.03.2024). Mannheim: Leibniz-Institut für Deutsche Sprache. `https://www.ids-mannheim.de/digspra/kl/projekte/korpora/releases/`.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. CrossNER: Evaluating Cross-Domain Named Entity Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13452–13460.

Minlong Peng, Ruotian Ma, Qi Zhang, Lujun Zhao, Mengxi Wei, Changlong Sun, and Xuanjing Huang. 2020. Toward Recognizing More Entity Types in NER: An Efficient Implementation using Only Entity Lexicons. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 678–688, Online. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Xavier Schmitt, Sylvain Kubler, Jérémy Robert, Mike Papadakis, and Yves LeTraon. 2019. A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343.

Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. 2021. Biomedical Named Entity Recognition Using BERT in the Machine Reading Comprehension Framework. *Journal of Biomedical Informatics*, 118:103799.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

Sowmya Vajjala and Ramya Balasubramaniam. 2022. What do we Really Know about State of the Art NER? *Preprint*, arXiv:2205.00034.

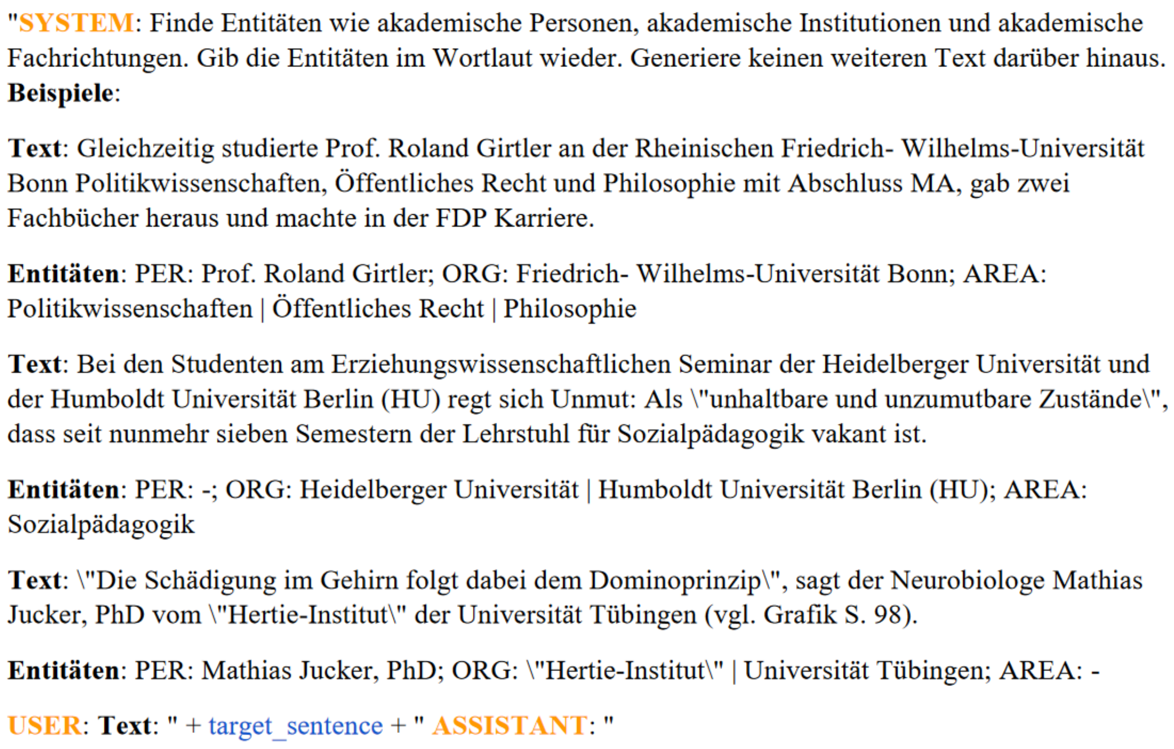
Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78—-85.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to Reduce Labeling Cost? GPT-3 Can Help. *arXiv preprint arXiv:2108.13487*.

A Appendix

A.1 Few-Shot Prompt

The model generated the most useful output with few-shot prompting, i.e. when providing three examples of correctly tagged sentences as the desired output. The actual target sentence required to be tagged by the LLM is then attached at the end of the prompt, see Figure 1. The challenge was to select examples as diverse as possible that are also short enough to not exceed the model's context window size of 512 tokens. Sometimes the target sentence was too long and maxed out the context window size, which led to an error and therefore no output was returned from the LLM. Other challenges consisted in the unexpected output formatting done by the model: No separation of entities of the same



"SYSTEM: Finde Entitäten wie akademische Personen, akademische Institutionen und akademische Fachrichtungen. Gib die Entitäten im Wortlaut wieder. Generiere keinen weiteren Text darüber hinaus.
Beispiele:

Text: Gleichzeitig studierte Prof. Roland Girtler an der Rheinischen Friedrich- Wilhelms-Universität Bonn Politikwissenschaften, Öffentliches Recht und Philosophie mit Abschluss MA, gab zwei Fachbücher heraus und machte in der FDP Karriere.

Entitäten: PER: Prof. Roland Girtler; ORG: Friedrich- Wilhelms-Universität Bonn; AREA: Politikwissenschaften | Öffentliches Recht | Philosophie

Text: Bei den Studenten am Erziehungswissenschaftlichen Seminar der Heidelberger Universität und der Humboldt Universität Berlin (HU) regt sich Unmut: Als \"unhaltbare und unzumutbare Zustände\", dass seit nunmehr sieben Semestern der Lehrstuhl für Sozialpädagogik vakant ist.

Entitäten: PER: -; ORG: Heidelberger Universität | Humboldt Universität Berlin (HU); AREA: Sozialpädagogik

Text: \"Die Schädigung im Gehirn folgt dabei dem Dominoprinzip\", sagt der Neurobiologe Mathias Jucker, PhD vom \"Hertie-Institut\" der Universität Tübingen (vgl. Grafik S. 98).

Entitäten: PER: Mathias Jucker, PhD; ORG: \"Hertie-Institut\" | Universität Tübingen; AREA: -

USER: **Text**: " + target_sentence + " ASSISTANT: "

Figure 1: Few-shot prompt with the prompt template keywords colored in orange and a placeholder for the target sentence in blue. The desired output as indicated in the examples is formatted as follows: PER: entity1 | entity2 | entity3; ORG: entity4 | entity5; AREA: entity6. A dash is inserted if an entity type is not detected at all.

type with the required separator symbol or the unrequested modifications of entities, e.g. the conversion of *Heidelberger Universität* into *Universität Heidelberg*, and halluzinations in the shape of inventing additional sentences. This behavior made the final extraction of entities impossible for some of the target sentences, which were then skipped and not included in the additional data sets. For the sentences where the output generation was successful and where the model kept the desired output format (i.e. designating the entity type followed by the entity values separated with vertical bars), the recognized entities could easily be extracted.

A.2 LLM Comparison

Table 3 shows the results of the LLM evaluation, which is performed on a test set consisting of 489 sentences from the initial data set. For better comparison, both models were instructed with the same few-shot prompt containing three examples of sentences and corresponding entity tags. *Llama-2-13B-chat*[5] achieved an F1 score of 85%, outperforming *OpenOrca-Platypus2-13B*[6] by more than 10 percent.

	Llama 2 Chat	OpenOrca Platypus 2
P	88.53	92.48
R	81.76	62.35
F1	**85.01**	74.49

Table 3: Precision, recall, and F1 scores (in percent) on tagging performance for 489 test set sentences.

A.3 Example Sentences

(1a) *"Aber riesige Zahlen sind immer noch nicht unendlich", sagt Heitzer, Professorin am Lehr- und Forschungsgebiet Didaktik der Mathematik an der RWTH.*

"But huge numbers are still not infinite', says Heitzer, professor of the teaching and research department Didactics of Mathematics at RWTH.'

(1b) *"Aber riesige Zahlen sind immer noch nicht unendlich", sagt Heitzer.*

[5] https://huggingface.co/TheBloke/Llama-2-13B-chat-GGML

[6] https://huggingface.co/TheBloke/OpenOrca-Platypus2-13B-GGML

“But huge numbers are still not infinite’, says Heitzer.’

(2a) *Ein paar Stockwerke höher wartet Astrid Kiermaier auf uns, die Molekularbiologin arbeitet bei Roche im Bereich Krebsforschung und leitet dort ein Team von 14 Mitarbeitern.*

‘A few floors up, Astrid Kiermaier is waiting for us, the molecular biologist works in the area of cancer research at Roche and leads a team of 14 employees there.’

(2b) *Ein paar Stockwerke höher wartet Astrid Kiermaier auf uns.*

‘A few floors up, Astrid Kiermaier is waiting for us.’

(3) *Mitte März begann ein Team von Forschern der Universität Hirosaki damit, sodass sie im Norden Japans bereits Messungen vor Ort durchführten.*

‘In mid-March, a team of researchers from Hirosaki University began with that, such that they already conduced on-site measurements in northern Japan.’

(4) *Der Physiker erfand nicht nur die Luftpumpe, sondern befaßte sich auch mit der barometrischen Erforschung des Luftdrucks.*

‘The physicist not only invented the air pump but also engaged in the barometric study of air pressure.’

(5) *Dr. Frank-Walter Steinmeier, Chef des Bundeskanzleramtes, ist dafür verantwortlich, Streitigkeiten zwischen den Politikern zu schlichten.*

‘Dr. Frank-Walter Steinmeier, head of the Federal Chancellery, is responsible for mediating disputes between politicians.’

(6) *Außenminister Frank-Walter Steinmeier gab sich weiter diplomatisch.*

‘Foreign Minister Frank-Walter Steinmeier continued to maintain a diplomatic stance.’

(7a) *Schütze, was du liebst - So lautet das Prinzip der Umweltikone Jane Goodall.*

‘Protect what you love – This is the principle of environmental icon Jane Goodall.’

(7b) *Schütze, was du liebst - So lautet das Prinzip der Primatenforscherin Jane Goodall.*

‘Protect what you love – This is the principle of primatologist Jane Goodall.’

(8) *“Die Schädigung im Gehirn folgt dabei dem Dominoprinzip”, sagt der Neurobiologe Mathias Jucker vom Hertie-Institut der Universität Tübingen.*

“The damage in the brain follows the domino principle’, says the neurobiologist Mathias Jucker from the Hertie Institute of the University of Tübingen.’