An evaluation of Named Entity Recognition tools for detecting person names in philosophical text

Ruben Weijers
Utrecht University

Jelke Bloem
University of Amsterdam
Institute for Logic, Language and Computation

Abstract

For philosophers, mentions of the names of other philosophers and scientists are an important indicator of relevance and influence. However, they don't always come in neat citations, especially in older works. We evaluate various approaches to named entity recognition for person names in 20th century, English-language philosophical texts. We use part of a digitized corpus of the works of W.V. Quine, manually annotated for person names, to compare the performance of several systems: the rule-based *edhiphy*, *spaCy*'s CNN-based system, FLAIR's BiLSTM-based system, and SpanBERT, ERNIE-v2 and ModernBERT's transformer-based approaches. We also experiment with enhancing the smaller models with domain-specific embedding vectors. We find that both *spaCy* and FLAIR outperform transformer-based models, perhaps due to the small dataset sizes involved.

1 Introduction

Named Entity Recognition (NER) tools have the ability to quickly and accurately extract named entities that can then be used to form connections between people. This has clear applications in Digital Humanities research (Ehrmann et al., 2023). A user study on the national library of France's portal *Gallica* showed that 80% of search queries contain a proper name (Chardonnens et al., 2018). In philosophy, and particularly in histories of ideas research, tracing names in digitized texts can reveal how concepts have spread and what influence authors had. Modern citation practices only emerged around the beginning of the 20th century, which makes citation analysis unsuited to study most of the history of science and philosophy. Referencing by mentioning names was the main type of referencing before citation conventions developed, and can thus be used to trace histories of philosophical ideas, as outlined by Petrovich et al. (2024).

However, in the field of philosophy, remarkably little attention has been paid to even simple computational tools that could help with quantitative analysis (Betti et al., 2019) and supplement the traditional close reading of texts. Petrovich et al. (2024) perform mention detection on late 19th century and early 20th century Anglophone philosophical texts using a rule-based gazetteer approach, after finding mistakes in applying *spaCy*'s NER model to some of these texts. However, a downside of this approach is a lack of out-of-domain coverage — such a system can only be expected to identify mentions of philosophers, missing out on e.g. other scientists, politicians or family members.

To support these interests, we perform a quantitative evaluation of a diverse range of NER approaches for philosophical text, from a rule-based approach to the recent ModernBERT LLM.[1] Of course, extensive literature on NER systems and their general performance already exists (e.g. this survey by Hu et al., 2024), but that does not necessarily translate to equal performance in the domain of philosophical texts. Obtaining state of the art results in NER relies heavily on domain-specific knowledge (Lample et al., 2016) and large annotated datasets, and many single-domain systems have been developed (Kormilitzin et al., 2021; Settles, 2004; Leaman and Gonzalez, 2008; Wei et al., 2019; Giorgi and Bader, 2020). Philosophy is certainly a specific domain. Even for philosophical texts that are in English and from the 20th century, there are significant differences between such text and Wikipedia text in terms of lexical semantics and word frequencies (Bloem et al., 2019). The frequent mention of low-frequency philosophical terms and capitalized German nouns may throw off NER systems. Similarly, the types of names mentioned are likely to be different than those in

[1] Our model and evaluation code can be found in the accompanying GitHub repository at https://github.com/bloemj/NERphilosophy.

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 418–425

general-purpose NER training datasets. Lastly, textual corpora in this domain are smaller.

Therefore, in this study, we manually annotated part of the QUINE corpus (Betti et al., 2020), consisting of the works of W. V. Quine, for person names, for the purpose of tuning and evaluating Named Entity Recognition systems in the domain of philosophical text. We use this data to train and evaluate state-of-the-art approaches as well as approaches that are more accessible to humanities researchers by being packaged in text processing tools. For NER tools, we evaluate the CNN-based NER (Lample et al., 2016) as implemented in *spaCy* (Neumann et al., 2019) and BiLSTM-CRF-based NER as implemented in FLAIR (Akbik et al., 2019a). Furthermore, we evaluate the recent ModernBERT LLM (Warner et al., 2024), an updated approach to bidirectional stacked encoders with modern optimizations. We also include ERNIE-v2 (Sun et al., 2020), a model that incorporates entity-level and phrase-level masking strategies into its pre-training objectives, and SpanBERT (Joshi et al., 2020), a model with a span-based version of the BERT masked language modelling training objective. These BERT variants are potentially more suited to performing the NER task compared to base BERT (Devlin et al., 2019). We also include a gazetteer baseline and a rule-based system based on *edhiphy*, Petrovich et al.'s (2024) database of philosophical names for mention detection.

2 Background

Even though some studies regarding the philosophical textual domain have been conducted (Muis et al., 2006; Mazzocchi and Tiberi, 2009), only Petrovich et al. (2024) cover the task of recognising named entities or person name mentions. Nevertheless, names that are frequently referred to in philosophical texts, especially names of philosophers, are often relevant and important to the writer. NER can aid in instantiating a web of relevance in philosophical texts.

There is some work on domain adaptation of word embeddings for philosophical text (Bloem et al., 2019; Zhou and Bloem, 2021). These studies evaluate the performance of embedding models with different types of domain adaptation, focusing on the challenge of having a small amount of in-domain data for philosophy. They test concatenation of in-domain and general-domain data using the Hyperwords (Levy et al., 2015) implementation of a count-based model with SVD dimension reduction, in-domain pretraining from scratch with Word2Vec (Mikolov et al., 2013), continued pretraining on the target domain with Word2Vec, tuning on target in-domain terms with Nonce2Vec (Herbelot et al., 2017) and ELMo contextual embeddings (Peters et al., 2018) with in-domain pretraining as well as general pretraining and in-domain finetuning.

These studies show that models benefit from a combination of in-domain pretraining and general-domain tuning, and that older modeling approaches are competitive with contextual embeddings in small data settings. Based on these findings, we experiment with incorporating domain-specific embeddings into the *spaCy* and FLAIR models.

2.1 *spaCy*

spaCy is a library for NLP in Python originally released in 2005[2] and updated in 2021[3]. The library is a very popular and robust framework that achieved state of the art results on NER and other NLP tasks (Kleinberg et al., 2018; Neumann et al., 2019; Partalidou et al., 2019). Lample et al. (2016) describe the CNN-based deep neural network that *spaCy* is based on. CNNs are shown to have strong generalisation ability, which *spaCy* uses to obtain high accuracy (Wang et al., 2021). In the medical domain, a F1-score of 94% is achieved predicting drug names (Kormilitzin et al., 2021).

2.2 FLAIR

FLAIR is a NLP framework that achieved state of the art results at the time of its release (Akbik et al., 2019a). It implements a BiLSTM-CRF sequence labeling architecture and contains multiple pre-trained contexual word embeddings (Akbik et al., 2019b; Huang et al., 2015). In these embeddings, words are represented as vectors that are derived from training methods similar to neural networks (Levy and Goldberg, 2014). Eldin et al. (2021) show that FLAIR is able to achieve a 95% F1-score on medical information extraction, additionally, Weber et al. (2021) show a 90.57% F1-score, compared to a 83.92% SciSpay (*spaCy* for biomedical text) F1-score.

[2] https://explosion.ai/blog/introducing-spacy
[3] https://spacy.io/usage/v3

3 Data

We use the QUINE corpus, version 0.5 (Betti et al., 2020), which consists of 228 documents, philosophical articles, books and letters; all written by the 20th century American philosopher Willard Van Orman Quine. Topics range from mathematics to formula-heavy logical writing to philosophical theories and concepts. The corpus contains 2,150,356 word tokens in the Format for Linguistic Annotation (FoLiA-XML, van Gompel and Reynaert, 2013), originating from printed texts written by Quine that were digitised using optical character recognition and semi-automatically corrected.

We randomly select 6800 sentences from the corpus (8.8% of the corpus) for manual person name annotation. Random selection ensures sentences from throughout Quine's bibliography are included. Sentences containing formulae were not considered for random selection. Annotation was performed using a web-based annotation tool[4] that yields character-based indices. Entities were annotated by a single annotator for maximum coverage. This annotator received domain instructions from a Quine expert. Names with spelling or OCR errors were also annotated, and in hyphen-linked entities, such as *"The Einstein-Boole theory"*, both separate entities are labelled. Incorrect capitalization was also included in labeling — for example, Alonzo Church (a 20th century mathematician), often referred to as *Church*, is frequently written in lowercase. The annotated data is split into a 70% training, 20% test and 10% validation split.

Besides personal names, other named entities that NER typically covers such as organization names and location names were not included in the annotation effort due to resource constraints. We did examine the potential relevance of location mentions in this corpus but found that most of them were related to holidays, travel or unrelated examples rather than e.g. universities, academic events and publisher locations. We also examined instances where names of philosophers sometimes occurred in the text without referring to the actual person, such as *plationian, copernican, boolean*, referring to *Plato, Copernicus, Boole*. These instances, in which philosophers' ideas or groups were mentioned, were initially given the NORP (Nationalities or Religious or Political groups) label. However, this entity group was too strongly dominated by other NORP entries such as *English, French, Greek* for classifiers to learn any domain-specific associations, so we excluded it.

[4]To be found at http://agateteam.org/spacynerannotate/

4 Models

Rule-based baseline As rule-based systems can perform well in narrow domains, we include such a baseline that draws on a gazetteer of 1117 names of philosophers drawn from the Britannica list of philosophers[5] and the website famousscientists.org[6]. We include a partial match baseline that considers last names as a true match, or first names if only a first name occurs, and an exact match baseline that only considers firstname-lastname occurrences as a true match. This baseline has shallower coverage of the philosophy domain than Petrovich et al.'s (2024) approach, who include far more philosophers in their database, but our baseline makes up for it by also including scientists.

edhiphy We also include a rule-based system using a gazetteer of all the philosopher names in Petrovich et al.'s (2024) *edhiphy* database. This includes 10,276 philosopher names. With this database, we exclude all names of 3 or fewer characters to reduce false positives. Again, we try a partial match and an exact match version.

spaCy We include the English *spaCy* pretrained models, as well as three models trained on our training split. One model is trained from scratch, and two models are trained with custom Word2Vec vectors (hyperparameters in Appendix A, following Sienčnik 2015). One of these has vectors from the QUINE corpus (2.2M tokens, 34712 vectors), the other has vectors from QUINE corpus merged with a 4.2M token domain-general corpus consisting of the Brown corpus (Francis and Kucera, 1979), Project Gutenberg corpus[7] and the NLTK Webtext corpus[8], yielding 28093 vectors. The small size is to avoid drowning out the domain-specific data.

FLAIR We use Akbik et al.'s (2019a) hyperparameters, shown in Appendix A, and default domain-general GloVe (Pennington et al., 2014) embeddings for English to train a BiLSTM-CRF model on top of using our training split.

[5]https://www.britannica.com/topic/list-of-philosophers-2027173

[6]https://www.famousscientists.org/

[7]https://github.com/RichardLitt/natural-gutenberg

[8]https://www.nltk.org/nltk_data/

Model	*P*	*R*	*F1*
Rule-based *partial match*	**.97**	.60	.74
Rule-based *exact match*	.80	.63	.70
edhiphy *partial match*	.78	.80	.79
edhiphy *exact match*	.94	.10	.18
en_core_web_sm *spaCy small*	.91	.31	.47
en_core_web_lg *spaCy large*	.90	.56	.69
spaCy trained-base	.90	.84	.87
spaCy trained-Quine W2V	.86	.90	.88
spaCy trained-Merged W2V	.93	.88	.90
FLAIR *trained*	.94	.89	**.91**
SpanBERT *base-tuned*	.83	.93	.87
SpanBERT *large-tuned*	.83	.92	.87
ModernBERT *base-tuned*	.78	.78	.78
ModernBERT *large-tuned*	.77	.83	.80
ModernBERT *CoNLL*	.50	.53	.51
ModernBERT *CoNLL-tuned*	.77	.87	.82
ERNIE *v2, base-tuned*	.76	**.94**	.84
ERNIE *v2, large-tuned*	.82	.90	.86

Table 1: Performance metrics for all models

LLMs For the transformer-based models (ModernBERT, SpanBERT, ERNIE-v2), we use the hyperparameters in Appendix A. For ModernBERT, we tune the base model as well as a model that has been tuned on the CoNLL-2003 NER shared task dataset (general-domain, Tjong Kim Sang, 2003). The models we tuned are used with a token classifier head, tuned on our training split.

4.1 Results

Table 1 shows all of our model results. Overall, we observe that the pre-transformer deep learning models outperform the transformer models in our setup, with the highest F1 score for labeling person names being achieved by FLAIR, trained on our training split. All models with good performance are dependent on domain-specific labeled data. We find that the rule-based baselines indeed outperform the general-domain *spaCy* models, as was also anecdotally found by Petrovich et al. (2024), mainly due to poor recall of *spaCy*. Presumably, it hasn't been trained on many philosopher names. Partial matches of gazetteer names in the rule-based setups are fairly successful for this specific domain, achieving the highest precision and a reasonable F1 score of .79 thanks to Quine's frequent mentioning of fairly famous philosophers and scientists that are included in the list. The rule-based *edhiphy* approach outperforms our rule-based baseline, achieving lower precision due to the larger list of names including some false-positives-inducing names like 'English', but higher recall, while still failing to recognize the names of some non-philosophers such as (Pierre de) Fermat.

Still, training *spaCy* on our labeled data leads to clearly better performance, and incorporating pre-trained vectors enhances this further. The best *spaCy* result (F1 = .91) is achieved with vectors trained on a combination of in-domain and out-of-domain data, which is in line with previous findings for this domain (Zhou and Bloem, 2021). In-domain word embeddings appear to lower precision, while increasing recall. FLAIR slightly outperforms trained *spaCy* with a more recent architecture and access to larger GloVe embeddings, achieving the best overall performance.

Among LLMs, we also observe the need for in-domain data. ModernBERT tuned on the CoNLL-2003 shared task NER-labeled data does not outperform the baseline (.51). Tuning it on our training data leads to far better results (.78), with slightly higher performance if the CoNLL-tuned model is used as a base (.82). Despite being smaller than ModernBERT (139M vs 103M parameters), ERNIE-v2 outperforms it, perhaps due to more relevant pre-training objectives. This includes a knowledge masking task, which requires the model to learn to predict masked spans and masked named entities rather than just tokens, forming a suitable base for the NER task. This model also achieves the highest recall of all models. SpanBERT-base is also smaller than ModernBERT (110M parameters), but has a more relevant pre-training objective of span masking. With this, it achieves the highest F1-score of the transformer-based approaches (.87). Lastly, we observe negligible differences between base and large versions of models. This suggests that the transformer models are mainly limited by the tuning of their classifier heads, for which we have limited labeled data available.

Some examples of errors made by the best-performing LLM, SpanBERT-large, include identifying "Ibid." as a name (when used to refer to an earlier reference), not identifying Cantor in "Cantor's principle", identifying "Oklahoma" as a person name, and not identifying Aristotle as a person name. In the sentence "Tom believes Cicero denounced Catiline", used as an example sentence, only Catiline is identified while the other names are not. In "Church cites examples from Ayer and

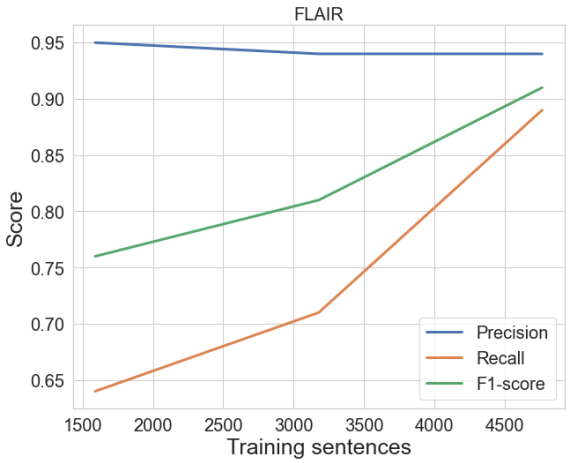

Figure 1: The influence of training data on FLAIR NER

Ryle", only Church is identified, while the others are not. Due to the black box nature of these LLMs, we can only speculate on why these errors might occur. The Aristotle error may be due to the occurrence of "Aristotelian" (not labeled as a person name) in the tuning data. Not identifying names like Tom in example sentences may not be a bad thing in this context, as the name does not refer to a philosopher. We would need to perform a structured error analysis on a larger dataset to identify patterns in the errors made.

5 Discussion

Our results show that in-domain labeled data is essential for successfully performing the NER task in the domain of philosophy, even if the amount of labeled data is fairly small. With limited data, simpler and older model architectures occasionally outperform state-of-the-art ones, an observation that has also been made by Ehrmanntraut et al. (2021) in the digital humanities context of language modelling for literary text.

To investigate the data size issue, we performed an ablation study with FLAIR, shown in Figure 1. We observe that FLAIR starts as a conservative model with high precision and relatively low recall, and then seemingly learns domain-specific names during training to increase recall and therefore the F1-score. More than half of our total training split is necessary to beat the LLM's performance. This suggests that, with the LLMs requiring more data to tune a larger number of parameters, the size of the labeled dataset is the bottleneck that causes older architectures to outperform recent LLMs in our philosophical domain corpus.

Based on our findings, it seems possible to achieve better performance on our dataset in future work by augmenting FLAIR with domain-adapted embeddings, or higher quality embeddings in general. Annotating a larger portion of our corpus would lead to better NER performance and potentially allow the state-of-the-art LLMs to reach their full potential. One open question is to what extent models tuned on our data would generalize to other domains of philosophical text, such as authors writing about the same topics in an earlier time period or working in different traditions than analytic philosophy. Historical data is very relevant to philosophical research and there are BERT models pre-trained on historical text that could be used for NER, but a study on NER for Dutch historical texts has shown that models pretrained on historical text do not necessarily outperform modern models at person name identification, even on 17th and 18th century data (Provatorova et al., 2024). Most importantly, future work will have to demonstrate whether NER for philosophical text can be combined with bibliometric analysis or other downstream tasks to gain more detailed insight into networks of authors and the history of ideas.

6 Limitations

Our experiments are limited in scope — although representative for philosophical text where target domains are often narrow, the corpus we used only covers a single author writing in a single language. We only cover the initial stages of a pipeline for bibliometric analysis, and do not experiment with automated entity linking, which would be the next step for incorporating mentions into bibliometric analysis. The use of a single annotator means that we don't have an inter-annotator agreement score to quantify the difficulty of the task, although annotating person names isn't the most difficult of tasks. In annotating their NER dataset for the archaeology domain, Brandsen et al. (2020) observed an inter-annotator agreement rate of 0.95.

The applicability of our described methods is limited by the fact that the most successful ones require thousands of in-domain labeled sentences. This limits the extent to which our method can be applied in other linguistic contexts and areas of philosophy. To facilitate comparison between architectures and data domains, we haven't fully optimized all our model conditions. Performance would benefit from model-specific hyperparameter

tuning, although this would also involve the use of more computational resources, and some of the top-performing models could be equipped with better general-domain or domain-adapted embeddings.

Acknowledgements

We are grateful to Floris Eskens and Arianna Betti for their input as Quine domain experts, as well as Martin Reynaert for support in the use of the QUINE corpus, and to Hein van den Berg for finding important additional related work. This research was supported by VICI grant *e-Ideas* (277-20-007), financed by the Dutch Research Council (NWO).

References

Akbik et al. 2019a. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019b. Pooled contextualized embeddings for named entity recognition. 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 724–728.

Arianna Betti, Martin Reynaert, Thijs Ossenkoppele, Yvette Oortwijn, Andrew Salway, and Jelke Bloem. 2020. Expert concept-modeling ground truth construction for word embeddings evaluation in concept-focused domains. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6690–6702, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Arianna Betti, Hein Van Den Berg, Yvette Oortwijn, and Caspar Treijtel. 2019. History of philosophy in ones and zeros. *Methodological advances in experimental philosophy*, pages 295–332.

Jelke Bloem, Antske Fokkens, and Aurélie Herbelot. 2019. Evaluating the consistency of word embeddings from small data. In *Natural Language Processing in a Deep Learning World*, International Conference Recent Advances in Natural Language Processing, RANLP, pages 132–141. Incoma Ltd. 12th International Conference on Recent Advances in Natural Language Processing, RANLP 2019 ; Conference date: 02-09-2019 Through 04-09-2019.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577.

Anne Chardonnens, Ettore Rizza, Mathias Coeckelbergs, and Seth Van Hooland. 2018. Mining user queries with information extraction methods and linked data. *Journal of Documentation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47.

Anton Ehrmanntraut, Thora Hagen, Leonard Konle, and Fotis Jannidis. 2021. Type-and token-based word embeddings in the digital humanities. In *Proceedings of the Conference on Computational Humanities Research 2021*.

Heba Gamal Eldin, Mustafa AbdulRazek, Muhammad Abdelshafi, and Ahmed T Sahlol. 2021. Med-Flair: medical named entity recognition for diseases and medications based on flair embedding. *Procedia Computer Science*, 189:67–75.

W. N. Francis and H. Kucera. 1979. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

John M Giorgi and Gary D Bader. 2020. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, 36(1):280–286.

Aurelie Herbelot, Marco Baroni, et al. 2017. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2017)*, pages 304–309. EastStroudsburg PA: ACL.

Zhentao Hu, Wei Hou, and Xianxing Liu. 2024. Deep learning for named entity recognition: a survey. *Neural Computing and Applications*, 36(16):8995–9022.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *Preprint*, arXiv:1508.01991.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Bennett Kleinberg, Maximilian Mozes, Arnoud Arntz, and Bruno Verschuere. 2018. Using named entities for computer-automated verbal deception detection. *Journal of forensic sciences*, 63(3):714–723.

Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. 2021. Med7: a transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118:102086.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Robert Leaman and Graciela Gonzalez. 2008. Banner: an executable survey of advances in biomedical named entity recognition. In *Biocomputing 2008*, pages 652–663. World Scientific.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Fulvio Mazzocchi and Melissa Tiberi. 2009. Knowledge organization in the philosophical domain: dealing with polysemy in thesaurus building. *KO KNOWLEDGE ORGANIZATION*, 36(2-3):103–112.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop*.

Krista R Muis, Lisa D Bendixen, and Florian C Haerle. 2006. Domain-generality and domain-specificity in personal epistemology research: Philosophical and empirical reflections in the development of a theoretical framework. *Educational Psychology Review*, 18(1):3–54.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. pages 319–327.

Eleni Partalidou, Eleftherios Spyromitros-Xioufis, Stavros Doropoulos, Stavros Vologiannidis, and Konstantinos I Diamantaras. 2019. Design and implementation of an open source Greek POS tagger and entity recognizer using spaCy. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 337–341. IEEE.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*, pages 2227–2237, New Orleans, Louisiana. ACL.

Eugenio Petrovich, Sander Verhaegh, Gregor Bös, Claudia Cristalli, Fons Dewulf, Ties van Gemert, and Nina IJdens. 2024. Bibliometrics beyond citations: introducing mention extraction and analysis. *Scientometrics*, 129(9):5731–5768.

Vera Provatorova, Marieke Van Erp, and Evangelos Kanoulas. 2024. Too young to ner: Improving entity recognition on dutch historical documents. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 30–35.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, pages 107–110.

Scharolta Katharina Sienčnik. 2015. Adapting word2vec to named entity recognition. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 239–243.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding.

Erik Tjong Kim Sang. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003, Edmonton, Canada*, pages 142–147.

Maarten van Gompel and Martin Reynaert. 2013. FoLiA: A practical XML format for linguistic annotation–a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81.

Dalei Wang, Cheng Xiang, Yue Pan, Airong Chen, Xiaoyi Zhou, and Yiquan Zhang. 2021. A deep convolutional neural network for topology optimization with perceptible generalization ability. *Engineering Optimization*, 0(0):1–16.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794.

Hao Wei, Mingyuan Gao, Ai Zhou, Fei Chen, Wen Qu, Chunli Wang, and Mingyu Lu. 2019. Named Entity Recognition from biomedical texts using a Fusion Attention-Based BiLSTM-CRF. *IEEE Access*, 7:73627–73636.

Wei Zhou and Jelke Bloem. 2021. Comparing contextual and static word embeddings with small data. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 253–259.

A Hyperparameters

dim	alpha	wsize	min_c	sample	neg	epoch
500	0.025	**2**	**5**	0.001	5	5

Table 2: Hyperparameters used for Word2Vec embeddings used in *spaCy* models. Bolded values are adapted from default to suit the small data setting.

Emb	hidden	crf	alpha	max_epochs
GloVe	256	true	.1	150

Table 3: Hyperparameters of the FLAIR model. FLAIR trains either until it reaches max_epochs, or until it encounters a series of 4 consecutive "bad epochs", defined by the absence of improvement in F1-score. All models were done after $\sim$60 epochs.

alpha	batch	epoch	optimizer	epsilon
2e-5	8	25	adamw	1e-08

Table 4: Hyperparameters used for tuning the transformer models - ModernBERT, SpanBERT and ERNIE-v2.

B Software specifications

Python: 3.11.4
numpy: 2.2.2
torch: 2.6.0+cu124
transformers: 4.48.2

All models are available on HuggingFace:
SpanBERT/spanbert-large-cased
answerdotai/ModernBERT-base
IsmaelMousa/modernbert-ner-conll2003
nghuyong/ernie-2.0-base-en

C Hardware specifications

GPU: NVidia L4
GPU Memory: 24GB
CPU: AMD 9445P
Total Number of Cores: 64
Memory: 384 GB