Software Impacts 11 (2022) 100212

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

GERNERMED: An open German medical NER model

Johann Frei [*], Frank Kramer

Faculty of Applied Computer Science, University of Augsburg, Alter Postweg 101, 86159 Augsburg, Germany

ARTICLE INFO

Keywords:
Named entity recognition
Natural language processing
Clinical text mining
Machine learning

ABSTRACT

Recent advancements in natural language processing (NLP) have been achieved by the use of increasingly complex neural networks. In clinical context, NLP is a key technique to access highly relevant information from unstructured texts such as clinical notes. We evaluate the feasibility of training our neural model GERNERMED on annotated German training data generated by automated translation from a public English dataset. The work guides other researchers about the use of machine-translation methods for dataset acquisition. Due to the public origin of the dataset, our trained software can be used by fellow researchers without any legal access restrictions.

Code metadata

Current code version	*v1.0*
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2021-181
Permanent link to reproducible capsule	https://codeocean.com/capsule/0396930/tree/v1
Legal code license	*MIT License*
Code versioning system used	*none*
Software code languages, tools and services used	*Python, C++, pytorch/fairseq, clab/fast_align, explosion/SpaCy.*
Compilation requirements, operating environments and dependencies	*Python 3, SpaCy library*
If available, link to developer documentation/manual	*Readme page:* https://github.com/frankkramer-lab/GERNERMED/blob/main/README.md
Support email for questions	johann.frei@informatik.uni-augsburg.de

1. Introduction

Recent advancements in natural language processing (NLP) have been achieved by the extensive use of increasingly complex neural networks. For example, large general purpose language models from the kind of BERT [1]- or GPT [2,3]-inspired architectures are commonly trained on large corpora such as Common Crawl [4] or The Pile [5] that are composed of 320 TiB (Common Crawl) or 825 GiB (The Pile) raw text data. Since any kind of such large-scale data is infeasible to annotate, these datasets are mainly purposed for unsupervised methods such as pretraining [6]. However, when facing case-specific downstream tasks, well-suited datasets are used for fine-tuning in a supervised fashion [6]. In this context the dataset is required to be annotated for a certain task accordingly. The dataset plays a key role since the quality of such NLP models highly correlates with the quantity and qualitye of the training dataset that governs the model's learned parameters.

While public datasets have been used for training NLP models for specific tasks, the availability of these datasets falls short when it comes to non-English text data. For instance, in the case of NLP for clinical application, several public English datasets are accessible to the research community [7,8]. However, for clinical NLP in German, only limited data is available [9] to open research due to GDPR and other privacy protection concerns as well as the frequent lack of gold standard annotations.

Processing of unstructured German clinical data remains an ongoing area of research. Common tasks in NLP, such as named entity recognition (NER), are used for determining key elements from texts like medication information and various related information like dosage and duration [6].

In this work we present the GERNERMED software component, which was trained on a custom dataset of clinical notes for German

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

[*] Corresponding author.
E-mail addresses: johann.frei@informatik.uni-augsburg.de (J. Frei), frank.kramer@informatik.uni-augsburg.de (F. Kramer).

https://doi.org/10.1016/j.simpa.2021.100212
Received 10 December 2021; Received in revised form 22 December 2021; Accepted 22 December 2021

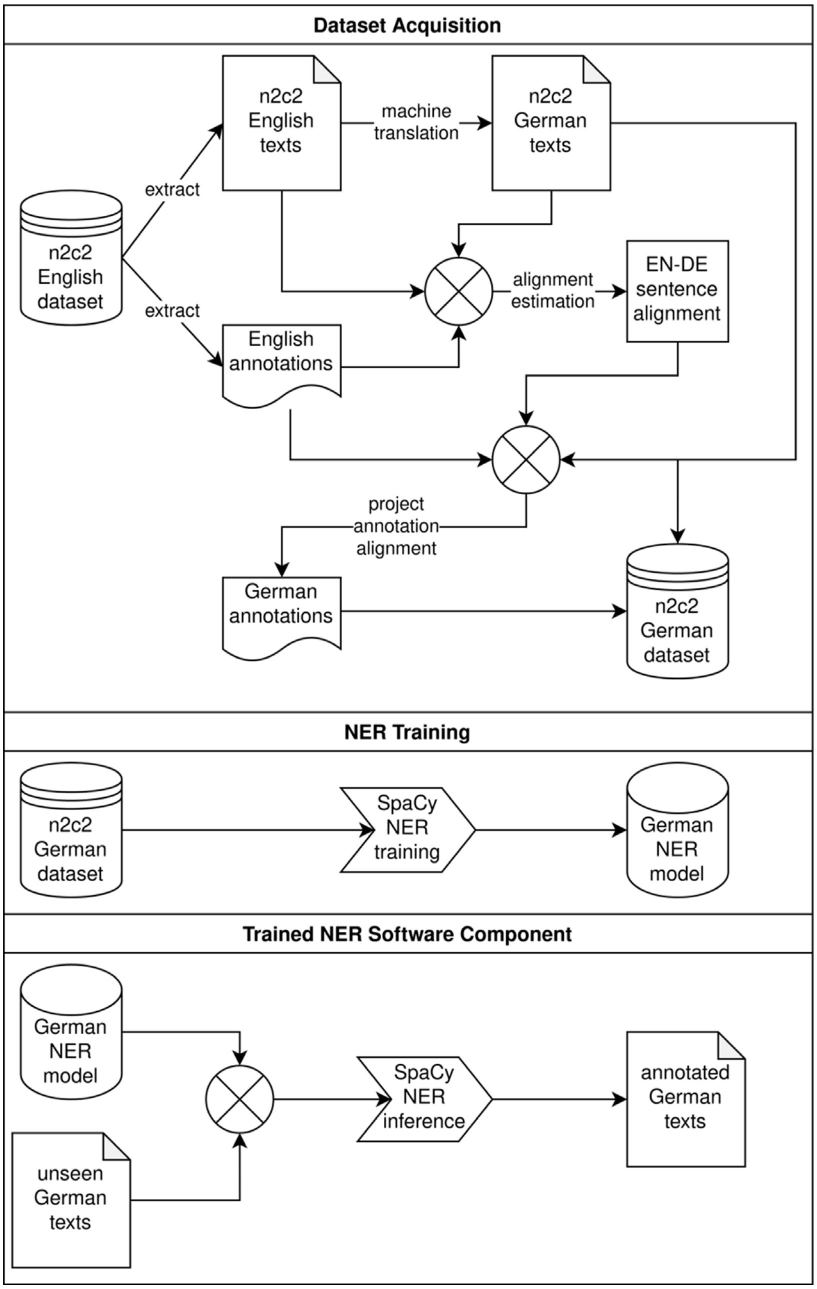

Fig. 1. Pipeline illustration: workflow for data synthesis, NER training and inference.

texts and can be easily deployed as independent part or used as a part of a larger NLP pipeline. As foundational work, the underlying dataset was automatically synthesized from publicly available English-based clinical data. The work also aims to guide other researchers about the use of machine-translation methods for similar silver-labeled dataset acquisition without manual quality control.

2. Materials and methods

The n2c2 2018 [8] challenge often provides the basis for English NLP work. The dataset consists of 303 annotated training documents and 202 gold standard-annotated test documents. We extracted and parsed the text and annotation labels from the dataset in order to translate the text from English into German using a pretrained neural machine translation model from Fairseq [10].

It cannot be assumed that the translated text does not differ from the structure of the original text due to inherent differences in syntax for English and German. For instance, it is not guaranteed that a translation-wise correspondence between exactly the fourth word in English and the fourth word in German exists.

In order to establish a word-to-word correspondence, we build upon the FastAlign [11] software that estimates a word-to-word alignments based on an expectation maximization-based algorithm given the pairs of input and output sentences. Because of the simplification of the statistical model for sentence alignment we expect the alignment estimation results to exhibit flaws in outlier samples that do not follow ordinary sentence structures in the original dataset. In order to filter these misalignment artifacts, we encode the assumption that successful alignment estimation approximately follows the word order of an English and German sentence pair. We discard samples from the dataset if average distance from the entries of the alignment matrix to its diagonal axis is exceeded by a certain threshold value. Given the alignment for each sentence pair, the annotation information for the

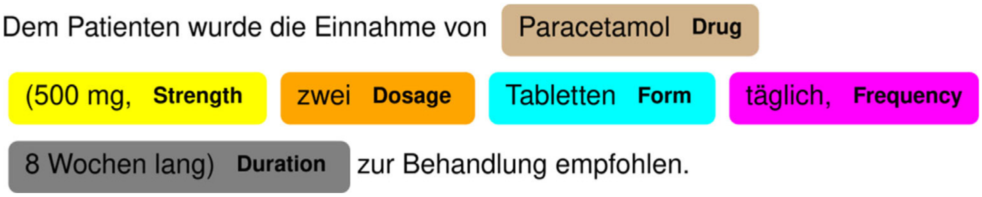

Fig. 2. NER tagging: successful processing of a German demo sentence.

English sentence can be propagated to the corresponding tokens in the German sentence.

Using our synthesized dataset, we can train a custom named entity recognizer component for the clinical application use case. For the implementation of the neural component and sentence parsing, we use the SpaCy [12] software for training and inference. The workflow is illustrated in Fig. 1.

3. Results

Here, we present a named entity recognizer component, which enables fellow researchers to directly integrate an annotation component into their research software systems. It was trained given the default SpaCy parameters for named entity recognition components. Our obtained dataset consists of 8599 sentences with a total number of 172695 tokens. The dataset was conventionally split into training (80%), validation (10%) and test set (10%) in order to measure the learning behavior as well as the final model performance.

The trained NER component is capable of detecting the medical-related entity tags Drug, Strength, Route, Form, Dosage, Frequency and Duration on an average F1-score of 81.54%. An example of the text annotation result is provided in Fig. 2.

Since our NER component is based on the component code of the SpaCy NLP pipeline, the component can be easily installed by a single command and included into related clinical text processing research pipelines in two lines of code.

4. Impact overview

Extracting relevant information such as drugs and medications from unstructured text data is a highly relevant use case because it enables other researchers with access to hospital-internal clinical notes to process large amounts of German text data in order to study and track health-related information for further research. In general, unstructured text processing does not only concern current data collection but includes processing of historic and legacy text data. Thus, it features relevance for retrospective study designs and secondary use of health data.

GERNERMED can provide benefits to the mining of patient records for the *DIFUTURE ProVal-MS* study [13] on Multiple Sclerosis in order to extract medication and drug-related information from German clinical notes at the local university hospital. Understanding the drug-disease interactions in multiple sclerosis can contribute to advancements in treatment decision and outcome. The *DIFUTURE* research project ("use case") on Parkinson's Disease [13] faces similar challenges, yet detection and extraction of medication data through our NER model can improve the quality of existing study data for statistical analysis.

Similarly, our model can be used by fellow researchers for other NLP pipelines in clinical research. The main impact in this research field is on automated annotation of non-English clinical documents.

Because the NER model was trained on data derived from publicly available sources instead of highly sensitive internal data from hospitals, we bypass the legal regulations and restrictions on privacy-related health data and are allowed to provide the trained NER model to the public audience. Due to the open nature of our component, the software can be further used for a broad variety of situations including commercial applications within the domain of German clinical NLP, but also be used for potential statistical model analysis since the model weights are publicly accessible.

Due to the novelty of our software component, we aim to receive feedback from upcoming internal as well as external projects and users to provide an updated iteration of the component as part of future work.

5. Discussion

The dataset was automatically generated through translation and alignment, error-inducing translation and alignment estimation are expected to degrade the quality of the dataset in comparison to manually curated datasets. However, the NER performance scores point out the capabilities and limits of such automated data synthetization and therefore, can be also relevant for other researchers from different domains.

We regard the deep analysis of the dataset and the software component as future work. The software can be considered as a baseline for competing open NLP components that will potentially be published in upcoming research work.

6. Conclusion

We presented the GERNERMED software component, an open named entity recognition system for German clinical texts. As a prerequisite for training such component, we described means to fast and effectively obtain a language-specific dataset from datasets of foreign languages for clinical domains.

Applying the method of public datasets allows us to provide the trained components for public use and make it easily accessible for interested users without relying on access restrictions. Furthermore, we supply example code and the performance evaluation script for our software in order to increase reproducibility in this research area.

Our results also provide other researchers general information on the effectiveness of building NLP components through machine translation-based dataset generation as an alternative to time- and cost-intensive manual dataset acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, 2018, CoRR, abs/1810.04805.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Ka-plan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc, 2020, pp. 1877–1901.

[3] Ben Wang, Aran Komatsuzaki, GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model, 2021, https://github.com/kingoflolz/mesh-transformer-jax.

[4] Common crawl blog, http://commoncrawl.org/connect/blog/. (Accessed: 2021-12-10).

[5] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, Connor Leahy, The pile: An 800 gb dataset of diverse text for language modeling, 2020, arXiv preprint arXiv:2101.00027.

[6] Bethany Percha, Modern clinical text mining: A guide and review, Annu. Rev. Biomed. Data Sci. 4 (1) (2021) 165–187, PMID: 34465177.

[7] Tom J. Pollard, Alistair E.W. Johnson, The mimic-iii clinical database, 2016, http://dx.doi.org/10.13026/C2XW26.

[8] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, Ozlem Uzuner, 2018 n2c2 Shared task on adverse drug events and medication extraction in electronic health records, J. Am. Med. Inform. Assoc.: JAMIA 27 (1) (2020) 3—12.

[9] Florian Borchert, Christina Lohr, Luise Modersohn, Thomas Langer, Markus Follmann, Jan.Philipp Sachs, Udo Hahn, Matthieu-P. Schapranow, Ggponc: A corpus of german medical text with rich metadata based on clinical practice guidelines. in: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, 2020, pp. 38–48.

[10] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, Michael Auli, Fairseq: A fast, extensible toolkit for sequence modeling, in: Proceedings of NAACL-HLT 2019, Demonstrations, 2019.

[11] Chris Dyer, Victor Chahuneau, Noah A. Smith, A simple, fast, and effective reparameterization of IBM model 2, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 644–648.

[12] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, spaCy: Industrial-strength Natural Language Processing in Python, 2020.

[13] Fabian Prasser, Oliver Kohlbacher, Ulrich Mansmann, Bernhard Bauer, Klaus A. Kuhn, Data integration for future medicine (difuture), Methods Inf. Med. 57 (S 01) (2018) e57–e65.