Original Paper

German Medical Named Entity Recognition Model and Data Set Creation Using Machine Translation and Word Alignment: Algorithm Development and Validation

Johann Frei, MSc; Frank Kramer, Prof Dr

IT Infrastructure for Translational Medical Research, University of Augsburg, Augsburg, Germany

Corresponding Author:
Johann Frei, MSc
IT Infrastructure for Translational Medical Research
University of Augsburg
Alter Postweg 101
Augsburg, 86159
Germany
Phone: 49 17691464136
Email: johann.frei@informatik.uni-augsburg.de

Abstract

Background: Data mining in the field of medical data analysis often needs to rely solely on the processing of unstructured data to retrieve relevant data. For German natural language processing, few open medical neural named entity recognition (NER) models have been published before this work. A major issue can be attributed to the lack of German training data.

Objective: We developed a synthetic data set and a novel German medical NER model for public access to demonstrate the feasibility of our approach. In order to bypass legal restrictions due to potential data leaks through model analysis, we did not make use of internal, proprietary data sets, which is a frequent veto factor for data set publication.

Methods: The underlying German data set was retrieved by translation and word alignment of a public English data set. The data set served as a foundation for model training and evaluation. For demonstration purposes, our NER model follows a simple network architecture that is designed for low computational requirements.

Results: The obtained data set consisted of 8599 sentences including 30,233 annotations. The model achieved a class frequency–averaged F_1 score of 0.82 on the test set after training across 7 different NER types. Artifacts in the synthesized data set with regard to translation and alignment induced by the proposed method were exposed. The annotation performance was evaluated on an external data set and measured in comparison with an existing baseline model that has been trained on a dedicated German data set in a traditional fashion. We discussed the drop in annotation performance on an external data set for our simple NER model. Our model is publicly available.

Conclusions: We demonstrated the feasibility of obtaining a data set and training a German medical NER model by the exclusive use of public training data through our suggested method. The discussion on the limitations of our approach includes ways to further mitigate remaining problems in future work.

(JMIR Form Res 2023;7:e39077) doi: 10.2196/39077

KEYWORDS

natural language processing; named entity recognition; information extraction

Introduction

Overview

Despite continuous efforts to transform the storage and processing of medical data in health care systems into a framework of machine-readable highly structured data, implementation designs that aim to fulfill such requirements are only slowly gaining traction in the clinical health care environment. In addition to common technical challenges, physicians tend to bypass or completely avoid inconvenient data input interfaces, which enforce structured data formats, by encoding relevant information as free-form unstructured text [1,2].

Electronic data capturing systems are developed to improve the situation of structured data capturing. Yet their primary focus lies on clinical studies. The involvement of these systems needs to be designed in early stages and requires active software management and maintenance. Such electronic data capturing solutions are commonly considered in the context of clinical research but are largely omitted in non–research-centric health care services, and paper-based solutions are preferred [1-4].

Because of the rise of data mining and big data analysis, finding and understanding novel relationships of disease, disease-indicating biomarkers, drug effects, and other input variables require large-scale data acquisition and collection. This induces additional pressure on finding and exploring new possible data sources.

Although new data sets can be designed and created for specific use cases, the amount of obtained data might be very limited and not sufficient for modern data-driven methods. Furthermore, such data collection efforts can turn out as rather inefficient in terms of time and work involved in creating new data sets with respect to the number of acquired data samples.

In contrast, unstructured data of sources from legacy systems and non–research-centric health care, referred to as second use, offer a potential alternative. However, techniques for information extraction and retrieval, mainly from the natural language processing (NLP) domain, need to be applied to transform raw data into structured information.

While the availability of existing NLP models in English, and other non–NLP-based techniques, for medical use cases is the focus of active research, the situation of medical NLP models for non-English languages is less satisfying. As the performance of an NLP model often depends on its dedicated target language, most models cannot be shared and reused easily in different languages but require retraining on new data from the desired target language.

In particular, for the case of detection of entities like prescribed drugs and level or frequency of dosage from German medical documents like doctoral letters, few open and publicly available models have been published. We attribute this to two main contributing factors:

1. Lack of public German data sets: Most open public data sets are designed for English data only. Until 2020, no such dedicated German data set has been published. Specifically in the context of clinical data, legal restrictions and privacy policies prevent the collection and publication of German data sets. Data-driven NLP research for medical applications uses largely internal data for training and evaluation. In addition to the data set itself, to model relevant text features with supervised learning, high-quality annotations of the data set are essential for robust model performance.
2. Protection of sensitive data and privacy concerns: Although few works have been published that present data-driven models for German texts, the weights of these models have not been openly published. Because respective training data have been used in a nonanonymized or pseudonymized fashion, the publication of the model weights inherently comes at the risk of possible data leakage issues through training data extraction [5] from the model, potentially exposing sensitive information like patient names or ID numbers.

In this paper, we aim to tackle the scarcity issue of anonymous training data and publicly available medical German NLP models. Our main contributions are as follows:

- Automated retrieval of German data set: We propose a method to create a custom data set for our target language, based on a public English data set. In addition, we apply a strategy to preserve relevant annotation information across languages.
- Training of medical German NLP model component: We trained and built a named entity recognition (NER) component on the custom data set. The model pipeline supports multiple types of medical entities.
- Evaluation and publication of the NLP component: The retrieved data set and the NER model were evaluated as part of an NLP pipeline. The trained model is publicly available for further use by third parties.

Related Work

In recent years, substantial progress has been made in the area of NLP, which can mostly be attributed to the joint use of large amounts of data and their processing through large language models like BERT (Bidirectional Encoder Representations from Transformers) [6] and its (bio)medical-specific models [7-12]. Such elements display a straightforward way to encode representations of semantic information for further processing in downstream tasks like text classification or text segmentation. These works mostly focus on the English language because of available language corpora like scientific texts from PubMed or specifically designed corpora such as *n2c2* [13] (with annotations) and *MIMIC-III* [14]. For German, only a few works such as *GGPONC* [15] and *BRONCO* [16] have been published in recent years as data sets that carry annotation information. Other German data sets [17,18] lack annotation information. Moreover, the *Technical-Laymen* [19] corpus provides an annotated corpus, yet it is based on crawled texts from nonprofessional online forums. Various other German medical text corpora exist [20-31] as a basis for certain NLP and information extraction use cases but are inaccessible for public distribution.

In the field of NLP systems for German medical texts, *medSynDiKATe* [32,33] approaches information extraction on pathological finding reports by parsing and mapping text elements to (semi)automatically build knowledge representation structures. Processing of pathological findings in German has also been applied to the tasks of sentence classification [22].

In the context of patient records, a hybrid relation extraction (RE) and NER parsing approach using the *SProUT* [34] parser has been proposed [35]; however, the entity tags lack medical relevance. A similar general NER for nonmedical entity tags has been applied to enable the deidentification of clinical records [36] using statistical and regex-based models through the *StanfordNLP* parser [37].

Neural methods have been shown to perform well on certain NLP tasks. In particular, convolutional neural network (CNN) approaches for RE [38-40] have become popular in recent years. For German texts, the performance of various methods has been investigated for medical NER tasks [41], such as CNN, long short-term memory, or support vector machine–based models. In this context, the text processing platform *mEx* [42] uses CNN-based methods for solving medical NER in German texts. Similar to our work, *mEx* is built on *SpaCy* [43] but provides custom models for other NLP tasks such as RE. However, the platform has been partially trained on internal clinical data, and thus, its statistical models have not been openly published and may only be used under certain legal restrictions on request. An updated version has been published [44], yet the models can only be retrieved on request under a usage agreement. As additional work, *GGPONC* (release 2.0) [45] provides a baseline model on request. For a more exhaustive survey on non-English clinical NLP in general, we point to [46].

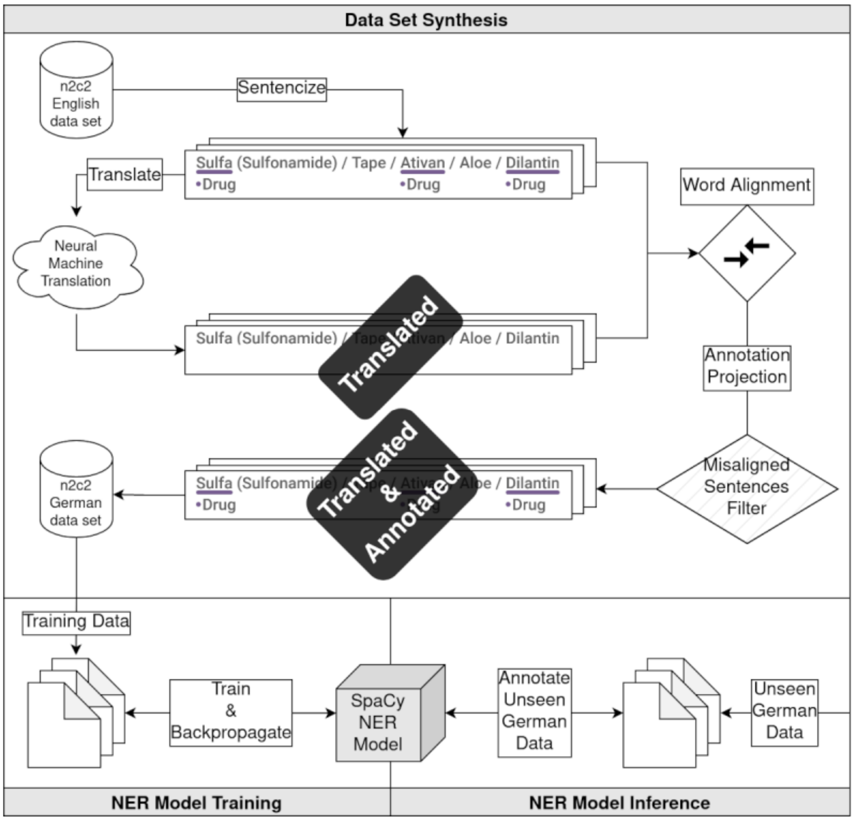
With respect to obtaining cross-lingual annotation information, the basic concept of projecting label data in language pairs via word alignment has been discussed in various NLP contexts [47-57]. For medical use cases, little research exists [49], with focus on English and Chinese data and models. Yet German medical contexts remain largely unexplored.

Methods

Overview

In this section, we first describe our method to synthesize the data set and then describe the used NER model for German text tagging. The entire pipeline is illustrated in Figure 1.

Figure 1. Illustration of the data set creation and natural language processing (NLP) model training process. The initial English *n2c2* data set is transformed into a synthetic German data set. The data set is used for training an NLP NER model. NER: named entity recognition.



Custom Data Set Creation

We relied on the publicly available training data from the *n2c2 NLP 2018 Track 2* [13] data set (adverse drug event [ADE] and medication extraction challenge) as our initial source data set. The data were composed of 303 annotated text documents that have been post processed by the editor for anonymization purposes to explicitly mask sensitive privacy-concerning information. They featured the annotation labels *Drug*, *Route*, *Reason*, *Strength*, *Frequency*, *Duration*, *Form*, *Dosage*, and *ADE*.

To transform the data into a semantically plausible text, we identified the type and text span of text masks such that we were able to replace the text masks by sampling type-compatible data randomly from a set of sample entries. During the sampling stage, depending on the type of mask, text samples for entities like dates, names, years, or phone numbers were generated and inserted into the text. Because every replacement step might affect the location of the text annotation labels as provided by the character-wise start and stop indices, these label annotation indices must be updated accordingly. For further preprocessing, we split up the text into single sentences such that we could omit all sentences with no associated annotation labels.

For automated translation, we made use of the open source *fairseq* (version 0.10.2) [58] model architecture. *fairseq* is an implementation of a neural machine translation model that supports the automatic translation of sequential text data using pretrained models. For our purposes, we ran the *transformer.wmt19.en-de* pretrained model to translate our set of English sentences into German because the model shows a strong BLEU (BiLingual Evaluation Understudy) translation score for English-German translation tasks [59] while maintaining its simplicity for deployment.

The reconstructive mapping of the annotation labels from the English source text to the German target text was tackled by *fast_align* [60]. *fast_align* is an unsupervised method for aligning words from 2 sentences of source and target language. The choice for *fast_align* was reasoned by its low-resource footprint, and it can align sentences fast through its simple statistical model. We projected the annotation labels onto the translated German sentences using the word-level mapping between the corresponding English and German sentence to obtain new annotation label indices in the German sentence. In nonmedical contexts, similar work on non-German target languages exists (eg, [57]).

The word alignment mapping tends to induce errors in situations of sentences with irregular structures such as tabular or itemized text sections. We mitigated the issue and potential subsequent error propagation by inspecting the structure of the word mapping matrix *A*:

$$A_{\text{regular}} = \begin{array}{r c} & \begin{array}{cccccc} \textit{The} & \textit{cat} & \textit{sat} & \textit{on} & \textit{the} & \textit{mat.} \end{array} \\ \begin{array}{r} \textit{Die} \\ \textit{Katze} \\ \textit{saß} \\ \textit{auf} \\ \textit{der} \\ \textit{Matte.} \end{array} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

In situations where *fast_align* fails to establish a meaningful mapping between the source and target sentence, it can be observed that the resulting mapping table collapses to a highly nondiagonal matrix structure, as illustrated by the following example:

$$A_{\text{irregular}} = \begin{array}{r c} & \begin{array}{cccccc} \textit{The} & \textit{cat} & \textit{sat} & \textit{on} & \textit{the} & \textit{mat.} \end{array} \\ \begin{array}{r} \textit{Die} \\ \textit{Katze} \\ \textit{saß} \\ \textit{auf} \\ \textit{der} \\ \textit{Matte.} \end{array} & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

Severely ill-aligned word mapping matrices can be detected and removed from the final set of sentences by applying the simple filter decision rule:

$$\frac{1}{\max(w_{\text{en}}, w_{\text{de}})} \sum_{i=1}^{w_{\text{de}}} \sum_{j=1}^{w_{\text{en}}} A_{ij} \frac{|w_{\text{en}} - i * w_{\text{en}} + i - w_{\text{de}} + j * w_{\text{de}} - j|}{\sqrt{(w_{\text{en}} - 1)^2 + (w_{\text{de}} - 1)^2}} > t$$

where the average distance between a nonzero entry and the diagonal line from $A_{1,1}$ to $A_{w_{\text{de}}, w_{\text{en}}}$ is evaluated, given w_{en} as the number of words in the English sentence and w_{de} as the number of words in the German sentence. If the value exceeds the threshold *t*, the sentence pair is disregarded for the final set of sentences.

The word mapping matrices describe a nonsymmetric cross-correspondence between 2 language-dependent token sets, which enables the projection of tokens within the English annotation span onto the semantically corresponding tokens in the German translation text. Therefore, the annotation label indices for the English text can be resolved to the actual indices for the translated German text at a character level.

NER Model Architecture

For the buildup of our NER model as part of an NLP pipeline, we use *SpaCy* as an NLP framework for training and inference. In comparison with state-of-the-art models, we select a lightweight non–transformer-based model because it serves primarily as a demonstration model and can be trained without significant compute costs.

- Embedding: The word tokens are embedded by Bloom embeddings [61] where different linguistic features are concatenated into a single vector and passed through n_{embed} separate dense layers, followed by a final max pooling and layer norm step. This step enables the model to learn meaningful linear combinations of single input feature embeddings while reducing the number of dimensions.
- Context-aware token encoding: To extract context-aware features that are able to capture larger token window sizes, the final token embedding is passed through a multilayered convolutional network. Each convolution step consists of the convolution itself and the following max-pooling operation to keep the dimensions constrained. For each convolution step, a residual (skip) connection is added to allow the model to pass intermediate data representations from previous layers to subsequent layers.
- NER parsing: For each encoded token, a corresponding feature-token vector is precomputed in advance by a dense layer. For parsing, the document is processed token-wise in a stateful manner. For NER, the state at a given position consists of the current token, the first token of the last entity, and the previous token by index. Given the state, the feature-position vectors are retrieved by indexing the values from the precomputed data and summed up. A dense layer is applied to predict the next action. Depending on the action, the current token is annotated and the next state is generated until the entire document has been parsed.

Ethical Considerations

Because of the nature of our proposed method, our work does not involve data or human subject research, which could potentially violate basic human ethics in a narrow sense.

The public data approach shifts the responsibility of privacy-preserving measures to the data set publisher. We assume that the *n2c2* data set has been deidentified correctly and no privacy-related information can be retrieved anymore.

Results

Data Set Synthesis

The source data set consists of 303 documents from the *n2c2* training data set. As an initial preprocessing step, we needed to replace the anonymization masks with meaningful regular text data to reconstruct the natural appearance of the text and alleviate a potential data set bias that leads to gaps between the data set and real-world data. For numerical data, we could retrieve mask replacements by random sampling. Similar to numerical data, dates and years are sampled and formatted to common date formats. For semantically relevant data types, we used the Python package *Faker*. The package maintains lists of plausible data of various types such as first names, last names, addresses, or phone numbers. We made use of these data entries for certain types of anonymization masks.

To obtain our custom data set, we split the texts from the original data set into single sentences using the sentence splitting algorithm from *SpaCy*. The English sentences were translated into German by the *fairseq* library with beam search (*b*=5). The sentence-wise word alignments were obtained by *fast_align* and cleaned up by our filter decision rule (*t*=1.8). To determine this particular hyperparameter, we sampled 10 ill-aligned samples without applying the filter and gradually lowered the threshold *t* until all 10 samples were detected by the decision rule.

The labels *Reason* and *ADE* were removed from the data set because of the fact that their definitions are rather ambiguous in general contexts beyond the scope of the initial source data set.

Our final custom data set consisted of 8599 sentence pairs, annotated with 30,233 annotations of 7 different class labels. The different class labels and their corresponding frequency in absolute numbers are shown in Table 1. The German sentences consisted of 172,695 tokens in total.

Table 1. The model performance scores per named entity recognition (NER) tag and the annotation distribution in the custom data set in absolute numbers.[a]

NER tag	Precision (%)	Recall (%)	F_1 score (%)	Label tags,n
Drug	67.33	66.17	66.74	8305
Strength	92.34	90.99	91.66	4071
Route	89.93	90.14	90.04	4549
Form	91.94	89.24	90.57	4238
Dosage	87.83	87.57	87.70	409
Frequency	79.14	76.92	78.01	5242
Duration	67.86	52.78	59.37	3419
Total	82.31	80.79	81.54	30,233

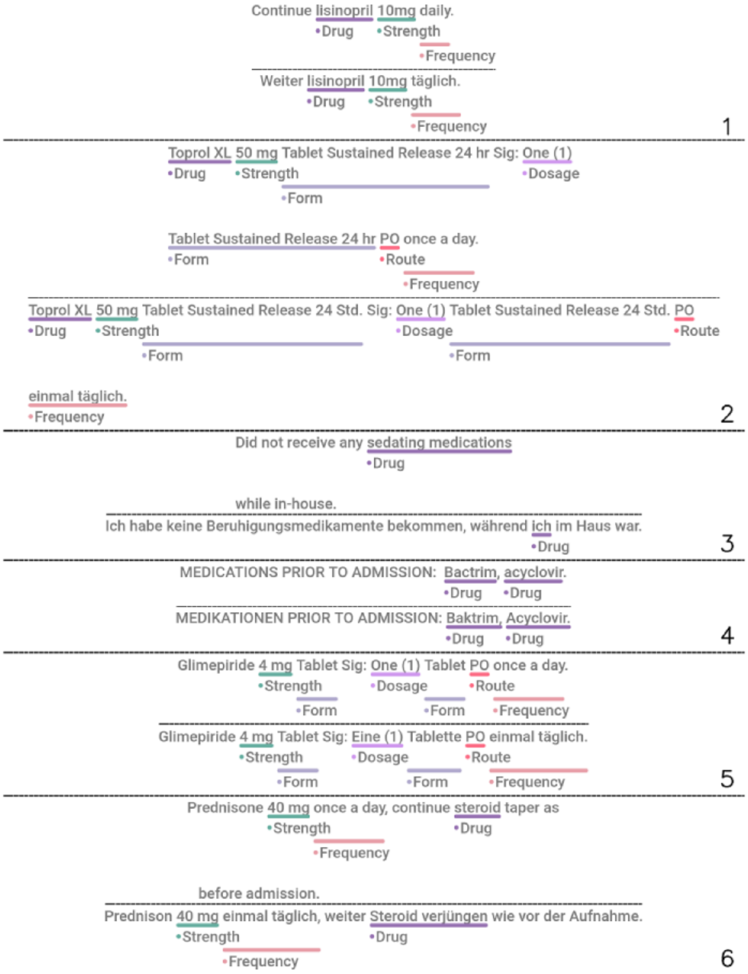
[a]The evaluation is based on the separated test set. Total scores are aggregated by label-frequency-weighted averaging. The total data set consists of 8599 sentence samples (172,695 tokens). A single-tag sample may span multiple tokens.

Translation and Alignment Artifacts

We sampled and selected a set of sentence pairs to investigate and illustrate the artifacts that we could observe in the synthesized data set with regard to translation as well as word alignment. The selection of samples is presented in Figure 2. Overall, we found the alignment and translation quality acceptable in sentences of simple structure and semantics (sample 1). However, the translation tended to fail in abbreviations such as PO (samples 2 and 5) as well as in text with uncommon syntax such as uppercase text (sample 4) or domain-dependent context (samples 2 and 6). To our surprise, the translation model was able to translate the sequence “One (1)” correctly in sample 5 but failed for the same term in sample 2. We attribute this to the context-sensitive, neural black-box model of the translation engine. In terms of alignment, most tokens were well aligned in sentence pairs of simple syntax and structure. Alignment errors could be found in sentence pairs with different sentence structures in English and German (sample 3) where our filter rule does not apply.

Because of parsing and alignment issues, we found that annotations were discarded (samples 5 and 6) in cases of single-token annotations that start as the first word of the sentence. This artifact affected primarily the label class *Drug*.

Figure 2. Selection of sampled sentence pairs from the synthetic data set. Most samples show correct outputs. The translation and word alignment artifacts occur in unusual syntactical contexts (translation) or complex sentence structures (alignment). Both English sentence (top) and its translated sentence (bottom) are depicted. Annotations with failed German correspondence resolution are not shown in the English sentence.



Continue lisinopril 10mg daily.

Weiter lisinopril 10mg täglich.

Toprol XL 50 mg Tablet Sustained Release 24 hr Sig: One (1) Tablet Sustained Release 24 hr PO once a day.

Toprol XL 50 mg Tablet Sustained Release 24 Std. Sig: One (1) Tablet Sustained Release 24 Std. PO einmal täglich.

Did not receive any sedating medications while in-house.

Ich habe keine Beruhigungsmedikamente bekommen, während ich im Haus war.

MEDICATIONS PRIOR TO ADMISSION: Bactrim, acyclovir.

MEDIKATIONEN PRIOR TO ADMISSION: Baktrim, Acyclovir.

Glimepiride 4 mg Tablet Sig: One (1) Tablet PO once a day.

Glimepiride 4 mg Tablet Sig: Eine (1) Tablette PO einmal täglich.

Prednisone 40 mg once a day, continue steroid taper as before admission.

Prednison 40 mg einmal täglich, weiter Steroid verjüngen wie vor der Aufnahme.

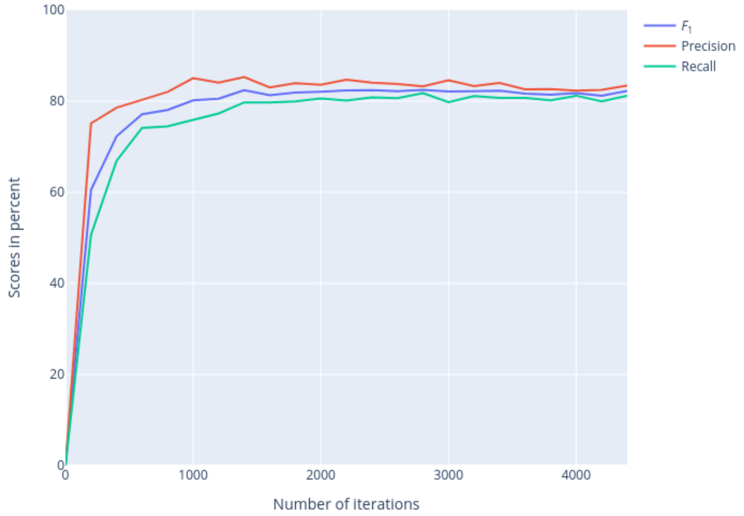
NER Model Training and Evaluation

For training, we used our custom German data set as our training data and split the data set into a training set (80%, 6879 sentence samples), validation set, and test set (both 10%, 860 samples). The training setup followed the default NER setup of *SpaCy*; the Adam optimizer with a learning rate of 0.001 with decay (β_1=.9, β_2=.999) was used. The training took 10 minutes on an Intel i7-8665U CPU.

The model performance during training is shown in Figure 3. The corresponding performance scores were evaluated on the validation set (as part of the training set).

We selected the final model based on the highest F_1 score on the validation set. The performance of the selected model was evaluated on the test set per NER tag as well as in total. The evaluation concerns the token-wise IOB (Inside, Outside, Begin)-action prediction. The results are shown in Table 1.

Figure 3. Training scores on validation set (as part of the training set): evaluation scores are computed at every 200th iteration.



Comparison to the English Baseline Data Set

To empirically quantify the error propagation through translation and word alignment, we retrained an equivalent model with all English sentences from our sentence pairs. The evaluation strategy remained similar to the strategy for the scores from Table 1. The scores are reported in Table 2. The results from the English model show comparable results to the German model for all labels except for *Drug*.

Table 2. Test set performance scores per named entity recognition (NER) tag of the model trained on the English sentences from the obtained data set in absolute numbers.[a]

NER tag	Precision (%)	Recall (%)	F_1 score (%)	German F_1 score (%)
Drug	80.94	82.02	81.47	−14.73 (66.74)
Strength	89.02	90.12	89.57	2.09 (91.66)
Route	85.55	95.08	90.06	−0.02 (90.04)
Form	94.36	87.12	90.60	−0.03 (90.57)
Dosage	89.41	89.97	89.69	−1.99 (87.70)
Frequency	80.55	80.15	80.35	−2.34 (78.01)
Duration	62.50	51.02	56.18	3.19 (59.37)
Total	85.14	85.82	85.48	−3.94 (81.54)

[a] The evaluation similar to the results from Table 1. Total scores are aggregated by label-frequency-weighted averaging. For comparison, the F_1 score differences of the German model to the English model are provided.

Advanced Evaluation and Model Comparison on a Separated Data Set

To further estimate the performance scores on a separated data set, we evaluated the model on a custom out-of-distribution (OoD) data set. The data set was created internally by clinical physicians by manually writing down and annotating 30 fake sentences (*Internal Gold*). For model comparison, we used the baseline model from *GGPONC* (release 2.0) [45] and evaluated its annotation performance on the label class we considered equivalent to our *Drug* label class. In comparison with our model (approximately 5 MB), the *GGPONC* model is orders of magnitudes larger due to its use of pretrained transformers (approximately 500 MB). The results are given in Table 3. The F_1 score corresponds to the character-wise label classification performance.

Table 3. Evaluation on our out-of-distribution data set with the related GGPONC baseline model for reference: the model performance drops significantly for certain infrequent label classes.[a]

Data set and GERNERMED	Sample, n	F_1 score (%)	*GGPONC* baseline	F_1 score (%)
Internal Gold (30 sentences)				
Drug	36	54.48	Chemicals_Drug	56.07
Strength	37	67.70	No equivalent	N/A[b]
Form	19	23.83	No equivalent	N/A
Dosage	4	02.47	No equivalent	N/A
Frequency	20	48.14	No equivalent	N/A
Duration	3	0	No equivalent	N/A

[a] The F_1 scores are evaluated as performance scores of character-wise label classifications. The label classes Dosage and Duration occur less frequently and therefore their scores are less reliable.

[b] N/A: not applicable.

Discussion

Principal Findings

We were able to obtain a synthetic German data set for medical purposes from an English data set by the method proposed in previous sections. As expected, the translation and alignment method introduced artifacts into the output data, but our model was still able to yield proper performance on the test set after training on the training set from our synthetic data set.

Separate training on the English sentence pairs yielded similar results for all label classes except for the *Drug* label class. Because it can be assumed that the inherent structure and vocabulary bias from the data set are preserved through translation, the drop for *Drug* can be explained by 2 joint reasons. First, the lexical properties of a translated *Drug* word to its source word can differ frequently and severely. More basic tokens like from the label *Strength* lack language-dependent elements and can be aligned in a robust manner. In cases of other label classes, phrases are often less diverse or are used repeatedly because of the data set bias. This enables robust alignments from *fast_align* due to their statistically frequent correlations.

When evaluating and comparing our model with another model on an OoD data set, we observed a drop in performance scores across labels. Aside from the label classes of low sample size, we attributed the gaps to the data set shift, which was not captured well by the underlying model architecture. The model cannot rely on high-level semantic embeddings but relies on basic structural patterns, and thus, it works well on the test set but yields less accurate results on independent data sets. The model is intentionally kept primitive as it is meant to serve as a demonstration of feasibility and does not make use of a pretrained transformer.

Limitations

Because non–drug-related label classes are not available as annotation data in most external data sets, we cannot independently quantify the drop in performance on these label classes. In the context of this work, it further remains unclear how impactful the use of pretrained transformer networks will be in terms of annotation performance on external data sets if it is trained on the synthesized data set. In this work, the choice of the statistical model and the slim neural model architecture, in particular, is attributed to its small computational footprint while being able to achieve satisfying results. In addition, the NER pipeline of *SpaCy* explicitly induces inductive bias through hand-crafted feature extraction during the token embedding stage. However, the focus of our work lies on the presentation of the translation and alignment method for data set synthesis and its demonstration data for training purposes in the German medical context. We consider an exhaustive hyperparameter optimization as well as the use of a transformer-based model as future work.

In general, the availability of German NER models and methods for medical and clinical domains still leaves much to be desired as described in previous sections. German data sets in this domain have been largely kept unpublished in the past. However, its implications are significantly broader. In the case of unpublished NLP models, it renders independent reproduction of results and fair comparisons impossible. In the case of lacking data sets or inconsistent annotations, novel competitive data-driven techniques cannot be developed or validated easily.

Conclusions

In this paper, we presented our method for obtaining a synthesized data set and neural NER model for German medical text as an open, publicly available model. We trained the model on our custom German data set from a publicly available English data set. We described the method to extract and postprocess texts from the masked English texts and generate German texts by translating and cross-lingual token aligning. In addition, the NER model architecture was described and the final model performance was evaluated for single NER tags as well as its performance in total. We discussed the observed issues with the synthesized data set and the performance drop through data set shifts. The advanced evaluation was done on an independent OoD data set. We believe that our method is a well-suited foundation for future work in the context of German medical entity recognition and natural language processing. In particular, the use of primitive NER model architecture remains an important point for future work. The need for independent data sets to further improve the situation for the research community on this matter has been highlighted.

The model and the test set corpus are available on GitHub [62].

Acknowledgments

This work is a part of the Data Integration for Future Medicine (DIFUTURE) project funded by the German Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) grant FKZ01ZZ1804E.

Conflicts of Interest

None declared.

References

1. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
2. Percha B. Modern clinical text mining: a guide and review. Annu Rev Biomed Data Sci 2021 Jul 20;4:165-187. [doi: 10.1146/annurev-biodatasci-030421-030931] [Medline: 34465177]
3. Krüger-Brand H, Osterloh F. Elektronische Patientenakte: Viele Modelle - noch keine Strategie. Dtsch Arztebl Int 2017;114(43):A1960-A1966.
4. Pohlmann S, Kunz A, Ose D, Winkler EC, Brandner A, Poss-Doering R, et al. Digitalizing health services by implementing a personal electronic health record in Germany: qualitative analysis of fundamental prerequisites from the perspective of selected experts. J Med Internet Res 2020;22(1):e15102 [FREE Full text] [doi: 10.2196/15102] [Medline: 32012060]
5. Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting training data from large language models. Usenix Association. 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting [accessed 2023-01-27]
6. Devlin J, Chang MW, Lee K, Toutanova K. 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). 2019 Jun Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
7. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: 18th BioNLP Workshop and Shared Task; August 2019; Florence, Italy p. 58-65. [doi: 10.18653/v1/w19-5006]
8. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021 May 20;4(1):86. [doi: 10.1038/s41746-021-00455-y] [Medline: 34017034]
9. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
10. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June 2019; Minneapolis, MN p. 72-78 URL: https://www.aclweb.org/anthology/W19-1909 [doi: 10.18653/v1/w19-1909]
11. Beltagy I, Lo K, Cohan A. SciBERT: pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019 Presented at: EMNLP-IJCNLP; November 2019; Hong Kong, China p. 3615-3620. [doi: 10.18653/v1/d19-1371]
12. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-tuning Bidirectional Encoder Representations From Transformers (BERT)-based models on large-scale electronic health record notes: an empirical study. JMIR Med Inform 2019 Sep 12;7(3):e14830 [FREE Full text] [doi: 10.2196/14830] [Medline: 31516126]
13. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc 2020 Jan 01;27(1):3-12 [FREE Full text] [doi: 10.1093/jamia/ocz166] [Medline: 31584655]
14. Pollard T, Johnson A, Mark R. The MIMIC-III clinical database. PhysioNet. 2016. URL: https://physionet.org/content/mimiciii/1.4/ [accessed 2023-01-27]
15. Borchert F, Lohr C, Modersohn L, Langer T, Follmann M, Sachs JP, et al. 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 Presented at: 11th International Workshop on Health Text Mining and Information Analysis; November 2020; Online p. 38-48. [doi: 10.18653/v1/2020.louhi-1.5]

16. Kittner M, Lamping M, Rieke DT, Götze J, Bajwa B, Jelas I, et al. Annotation and initial evaluation of a large annotated German oncological corpus. JAMIA Open 2021 Apr;4(2):ooab025 [FREE Full text] [doi: 10.1093/jamiaopen/ooab025] [Medline: 33898938]
17. Suominen H, Kelly L, Goeuriot L, Krallinger M. CLEF eHealth Evaluation Lab 2020. In: Jose JM, Yilmaz E, Magalhães J, Castells P, Ferro N, Silva MJ, et al, editors. Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II. Cham: Springer; 2020:587-594.
18. Lohr C, Buechel S, Hahn U. Sharing copies of synthetic clinical corpora without physical distribution: a case study to get around IPRs and privacy constraints featuring the German JSYNCC corpus. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018 Presented at: LREC 2018; May 2018; Miyazaki, Japan p. 7-12.
19. Seiffe L, Marten O, Mikhailov M, Schmeier S, Möller S, Roller R. From witch's shot to music making bones: resources for medical laymen to technical language and vice versa. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020 Presented at: Twelfth Language Resources and Evaluation Conference; May 2020; Marseille, France p. 6185-6192.
20. Wermter J, Hahn U. An annotated German-language medical text corpus as language resource. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation. 2004 Presented at: LREC; May 2004; Lisbon, Portugal.
21. Fette G, Ertl M, Wörner A, Kluegl P, Störk S, Puppe F. Information extraction from unstructured electronic health records and integration into a data warehouse. In: Goltz U, Magnor M, Appelrath HJ, Matthies HK, Balke WT, Wolf L, editors. INFORMATIK. Bonn: Gesellschaft für Informatik e.V; 2012:1237-1251.
22. Bretschneider C, Zillner S, Hammon M. Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach. In: Proceedings of the 2013 Workshop on Biomedical Natural Language Processing. 2013 Presented at: Workshop on Biomedical Natural Language Processing; August 2013; Sofia, Bulgaria p. 27-35.
23. Roller R, Uszkoreit H, Xu F, Seiffe L, Mikhailov M, Staeck O, et al. A fine-grained corpus annotation schema of German nephrology records. In: Proceedings of the Clinical Natural Language Processing Workshop. 2016 Presented at: ClinicalNLP; December 2016; Osaka, Japan p. 69-77.
24. Lohr JM, McDevitt DT, Lutter KS, Roedersheimer LR, Sampson MG. Operative management of greater saphenous thrombophlebitis involving the saphenofemoral junction. Am J Surg 1992 Sep;164(3):269-275. [doi: 10.1016/s0002-9610(05)81084-0] [Medline: 1415928]
25. Kreuzthaler M, Schulz S. Detection of sentence boundaries and abbreviations in clinical narratives. BMC Med Inform Decis Mak 2015;15 Suppl 2(Suppl 2):S4 [FREE Full text] [doi: 10.1186/1472-6947-15-S2-S4] [Medline: 26099994]
26. Cotik V, Roller R, Xu F, Uszkoreit H, Budde K, Schmidt D. Negation detection in clinical reports written in German. In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining. 2016 Presented at: BioTxtM2016; December 2016; Osaka, Japan p. 115-124.
27. Krebs J, Corovic H, Dietrich G, Ertl M, Fette G, Kaspar M, et al. Semi-automatic terminology generation for information extraction from german chest x-ray reports. Stud Health Technol Inform 2017;243:80-84. [Medline: 28883175]
28. Hahn U, Matthies F, Lohr C, Löffler M. 3000PA-towards a national reference corpus of German clinical language. Stud Health Technol Inform 2018;247:26-30. [Medline: 29677916]
29. Miñarro-Giménez JA, Cornet R, Jaulent MC, Dewenter H, Thun S, Gøeg KR, et al. Quantitative analysis of manual annotation of clinical text samples. Int J Med Inform 2019 Mar;123:37-48 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.12.011] [Medline: 30654902]
30. König M, Sander A, Demuth I, Diekmann D, Steinhagen-Thiessen E. Knowledge-based best of breed approach for automated detection of clinical events based on German free text digital hospital discharge letters. PLoS One 2019;14(11):e0224916 [FREE Full text] [doi: 10.1371/journal.pone.0224916] [Medline: 31774830]
31. Toepfer M, Corovic H, Fette G, Klügl P, Störk S, Puppe F. Fine-grained information extraction from German transthoracic echocardiography reports. BMC Med Inform Decis Mak 2015 Nov 12;15:91 [FREE Full text] [doi: 10.1186/s12911-015-0215-x] [Medline: 26563260]
32. Hahn U, Romacker M, Schulz S. MEDSYNDIKATE--a natural language system for the extraction of medical information from findings reports. Int J Med Inform 2002 Dec 04;67(1-3):63-74. [doi: 10.1016/s1386-5056(02)00053-9] [Medline: 12460632]
33. Hahn U, Romacker M, Schulz S. How knowledge drives understanding--matching medical ontologies with the needs of medical language processing. Artif Intell Med 1999 Jan;15(1):25-51. [doi: 10.1016/s0933-3657(98)00044-x] [Medline: 9930615]
34. Piskorski J, Homola P, Marciniak M, Mykowiecka A, Przepiórkowski A, Wolinski M. Information extraction for Polish using the SProUT platform. In: Kłopotek MA, Wierzchoń ST, Trojanowski K, editors. Intelligent Information Processing and Web Mining Proceedings of the International IIS: IIPWM‘04 Conference held in Zakopane, Poland, May 17–20, 2004. Berlin, Heidelberg: Springer; 2004.
35. Krieger HU, Spurk C, Uszkoreit H, Xu F, Zhang Y, Müller F, et al. Information extraction from German patient records via hybrid parsing and relation extraction strategies. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. 2014 Presented at: LREC14; May 2014; Reykjavik, Iceland p. 2043-2048.

36. Richter-Pechanski P, Riezler S, Dieterich C. De-identification of German medical admission notes. Stud Health Technol Inform 2018;253:165-169. [Medline: 30147065]
37. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The stanford coreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014 Presented at: 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; June 2014; Baltimore, MD p. 55-60 URL: http://www.aclweb.org/anthology/P/P14/P14-5010 [doi: 10.3115/v1/p14-5010]
38. Nguyen TH, Grishman R. Relation extraction: perspective from convolutional neural networks. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015 Presented at: 1st Workshop on Vector Space Modeling for Natural Language Processing; June 2015; Denver, CO p. 39-48. [doi: 10.3115/v1/w15-1506]
39. Sahu S, Anand A, Oruganty K, Gattu M. Relation extraction from clinical texts using domain invariant convolutional neural network. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. 2016 Presented at: 15th Workshop on Biomedical Natural Language Processing; August 2016; Berlin, Germany p. 206-215. [doi: 10.18653/v1/W16-2928]
40. Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014 Presented at: COLING 2014; August 2014; Dublin, Ireland p. 2335-2344.
41. Roller R, Rethmeier N, Thomas P, Hübner M, Uszkoreit H, Staeck O, et al. Detecting named entities and relations in German clinical reports. In: Rehm G, Declerck T, editors. Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings. Cham: Springer; 2017:146-154.
42. Roller R, Alt C, Seiffe L, Wang H. mEx - an information extraction platform for German medical text. In: Proceedings of the 11th International Conference on Semantic Web Applications and Tools for Healthcare and Life Sciences. 2018 Presented at: SWAT4HCLS-2018; December 3-5, 2018; Antwerp, Belgium p. 3-5.
43. Honnibal M, Montani I, Van Landeghem S, Boyd A. Industrial-strength natural language processing (NLP) with Python. Zenodo. URL: https://zenodo.org/record/3701227#.Y9VgXnZBw2w [accessed 2023-01-28]
44. Roller R, Seiffe L, Ayach A, Möller S, Marten O, Mikhailov M, et al. Information extraction models for German clinical text. 2020 Presented at: IEEE International Conference on Healthcare Informatics; November 30-December 3, 2020; Oldenburg, Germany p. 1-2. [doi: 10.1109/ichi48887.2020.9374385]
45. Borchert F, Lohr C, Modersohn L, Witt J, Langer T, Follmann M, et al. GGPONC 2.0—the German clinical guideline corpus for oncology: curation workflow, annotation policy, baseline NER taggers. In: Proceedings of the 13th Language Resources and Evaluation Conference. 2022 Presented at: 13th Language Resources and Evaluation Conference; November 2020; Online p. 3650-3660 URL: https://aclanthology.org/2022.lrec-1.389 [doi: 10.18653/v1/2020.louhi-1.5]
46. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than english: opportunities and challenges. J Biomed Semantics 2018 Mar 30;9(1):12 [FREE Full text] [doi: 10.1186/s13326-018-0179-8] [Medline: 29602312]
47. Ehrmann M, Turchi M, Steinberger R. Building a multilingual named entity-annotated corpus using annotation projection. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011. 2011 Presented at: International Conference Recent Advances in Natural Language Processing; September 2011; Hissar, Bulgaria p. 118-124 URL: https://aclanthology.org/R11-1017
48. Mayhew S, Tsai CT, Roth D. Cheap translation for cross-lingual named entity recognition. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: Conference on Empirical Methods in Natural Language Processing; September 2017; Copenhagen, Denmark p. 2536-2545 URL: https://aclanthology.org/D17-1269 [doi: 10.18653/v1/d17-1269]
49. Ding P, Wang L, Liang Y, Lu W, Li L, Wang C, et al. Cross-lingual transfer learning for medical named entity recognition. In: Nah Y, Cui B, Lee SW, Yu JX, Moon YS, Whang SE, editors. Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part I. Cham: Springer; 2020:403-418.
50. Xu W, Haider B, Mansour S. End-to-end slot alignment and recognition for cross-lingual NLU. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020 Presented at: EMNLP; November 2020; Online p. 5052-5063 URL: https://aclanthology.org/2020.emnlp-main.410 [doi: 10.18653/v1/2020.emnlp-main.410]
51. Yarowsky D, Ngai G, Wicentowski R. Inducing multilingual text analysis tools via robust projection across aligned corpora. In: Proceedings of the First International Conference on Human Language Technology Research. 2001 Presented at: HLT '01; March 18-21, 2001; San Diego, CA p. 1-8. [doi: 10.3115/1072133.1072187]
52. Zitouni I, Florian R. Mention detection crossing the language barrier. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2008 Presented at: EMNLP '08; October 25-27, 2008; Honolulu, HI p. 600-609. [doi: 10.3115/1613715.1613789]
53. Ni J, Dinu G, Florian R. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long

Papers). 2017 Jul Presented at: 55th Annual Meeting of the Association for Computational Linguistics; July 2017; Vancouver, Canada p. 1470-1480 URL: https://aclanthology.org/P17-1135 [doi: 10.18653/v1/p17-1135]

54. Xie J, Yang Z, Neubig G, Smith NA, Carbonell J. Neural cross-lingual named entity recognition with minimal resources. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: Conference on Empirical Methods in Natural Language Processing; October-November 2018; Brussels, Belgium p. 369-379 URL: https://aclanthology.org/D18-1034 [doi: 10.18653/v1/d18-1034]
55. Jain A, Paranjape B, Lipton ZC. Entity projection via machine translation for cross-lingual NER. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019 Presented at: EMNLP-IJCNLP; November 2019; Hong Kong, China p. 1083-1092 URL: https://www.aclweb.org/anthology/D19-1100 [doi: 10.18653/v1/d19-1100]
56. Schuster S, Gupta S, Shah R, Lewis M. Cross-lingual transfer learning for multilingual task oriented dialog. 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). 2019 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2019; Minneapolis, MN p. 3795-3805 URL: https://aclanthology.org/N19-1380 [doi: 10.18653/v1/n19-1380]
57. Hatami A, Mitkov R, Corpas PG. Cross-lingual named entity recognition via fastAlign: a case study. In: Proceedings of the Translation and Interpreting Technology Online Conference. 2021 Presented at: Translation and Interpreting Technology Online Conference; July 2021; Online p. 85-92 URL: https://aclanthology.org/2021.triton-1.10 [doi: 10.26615/978-954-452-071-7_010]
58. Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, et al. fairseq: a fast, extensible toolkit for sequence modeling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). 2019 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics; June 2019; Minneapolis, MN p. 48-53. [doi: 10.18653/v1/n19-4009]
59. Ng N, Yee K, Baevski A, Ott M, Auli M, Edunov S. Facebook FAIR's WMT19 news translation task submission. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). 2019 Aug Presented at: Fourth Conference on Machine Translation; August 2019; Florence, Italy p. 314-319. [doi: 10.18653/v1/w19-5333]
60. Dyer C, Chahuneau V, Smith NA. 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. 2013 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2013; Atlanta, GA p. 644-648.
61. Svenstrup D, Hansen J, Winther O. Hash embeddings for efficient word representations. 2017 Presented at: 31st Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 4935-4943.
62. GERNERMED—an open German medical NER model. GitHub. 2022. URL: https://github.com/frankkramer-lab/GERNERMED [accessed 2023-01-27]

Abbreviations

ADE: adverse drug event
CNN: convolutional neural network
NER: named entity recognition
NLP: natural language processing
OoD: out-of-distribution
RE: relation extraction

Edited by A Mavragani; submitted 27.04.22; peer-reviewed by R Roller, C Zhao; comments to author 22.08.22; revised version received 11.09.22; accepted 03.11.22; published 28.02.23

Please cite as:
Frei J, Kramer F
German Medical Named Entity Recognition Model and Data Set Creation Using Machine Translation and Word Alignment: Algorithm Development and Validation
JMIR Form Res 2023;7:e39077
URL: https://formative.jmir.org/2023/1/e39077
doi: 10.2196/39077
PMID: