Journal of Biomedical Informatics 147 (2023) 104513

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

GERNERMED++: Semantic annotation in German medical NLP through transfer-learning, translation and word alignment

Johann Frei [a,*], Ludwig Frei-Stuber [b], Frank Kramer [a]

[a] IT-Infrastructure for Translational Medical Research, University of Augsburg, Alter Postweg 101, 86159 Augsburg, Germany

[b] Institute and Outpatient Clinic for Occupational, Social and Environmental Medicine, 80336 Munich, Germany

ARTICLE INFO

Keywords:
Natural language processing
Medical NLP
Medical named entity recognition
Transfer learning
German NLP
Artificial intelligence

ABSTRACT

We present a statistical model, GERNERMED++, for German medical natural language processing trained for named entity recognition (NER) as an open, publicly available model. We demonstrate the effectiveness of combining multiple techniques in order to achieve strong results in entity recognition performance by the means of transfer-learning on pre-trained deep language models (LM), word-alignment and neural machine translation, outperforming a pre-existing baseline model on several datasets. Due to the sparse situation of open, public medical entity recognition models for German texts, this work offers benefits to the German research community on medical NLP as a baseline model. The work serves as a refined successor to our first GERNERMED model. Similar to our previous work, our trained model is publicly available to other researchers. The sample code and the statistical model is available at: https://github.com/frankkramer-lab/GERNERMED-pp.

1. Introduction

Extraction and processing of key information from medical notes and doctors' letters pose a common challenge in the advanced digitization of healthcare systems. In particular, research-oriented data mining of non-research-centric data sources (often referred to as *second use*) often requires expensive data harmonization processes in order to transform unstructured or semi-structured data into strictly structured, uniform data representations such as HL7 or FHIR. While manually solving these processes can be carried out for document analysis on certain studies, it is rendered impractical for large-scale text analysis on legacy data or processing day-to-day clinical data [1,2].

Handling heterogeneous data from text-based documents is a central subject of natural language processing. In recent years deep learning-inspired approaches have been applied successfully to tackle various NLP tasks effectively. However, training deep language models requires proper datasets in regard to aspects like corpus size, annotation work, data diversity and overall dataset quality, in order to retrieve well-performing models. In medical NLP, obtaining such annotated datasets remains rather difficult for various reasons [3]. For instance, the use and publication of medical data is highly restricted for the reasons of privacy and country-dependent data protection legislation [3]. Even though medical datasets have been published in English, such datasets for German texts in contrast are still frequently unavailable to external researchers [1].

In this paper, we propose an approach of combining multiple ideas to obtain a German medical NLP model, which we refer to as *GERNERMED++* and which serves as a successor to our previous *GERNERMED* [4] model:

- **Translation**: The state of German medical corpora is limited and the use of internal datasets for training and publication of such models is legally unclear. In contrast, medical datasets in English have already been published and therefore, neural machine translation (NMT) can be applied to obtain German data from English datasets.
- **Annotation Projection**: Annotation of large corpora is crucial for supervised learning and determines the quality of the final performance of the model. However the cost of obtaining gold-standard annotations from scratch is prohibitively expensive. Given our set of NMT-based German data, word alignment estimation can be used to project token-level annotations from English data to German data without manual intervention.
- **Transfer-Learning through Model Fine-Tuning**: To further improve the downstream performance of the NLP model under the constraints of our small, task-specific dataset, a larger, pre-trained German LM is used for advanced semantic, context-aware feature extraction and further fine-tuning.

* Corresponding author.
E-mail addresses: johann.frei@informatik.uni-augsburg.de (J. Frei), ludwig.freistuber@med.uni-muenchen.de (L. Frei-Stuber), frank.kramer@informatik.uni-augsburg.de (F. Kramer).

https://doi.org/10.1016/j.jbi.2023.104513
Received 14 November 2022; Received in revised form 27 September 2023; Accepted 4 October 2023
Available online 13 October 2023
1532-0464/

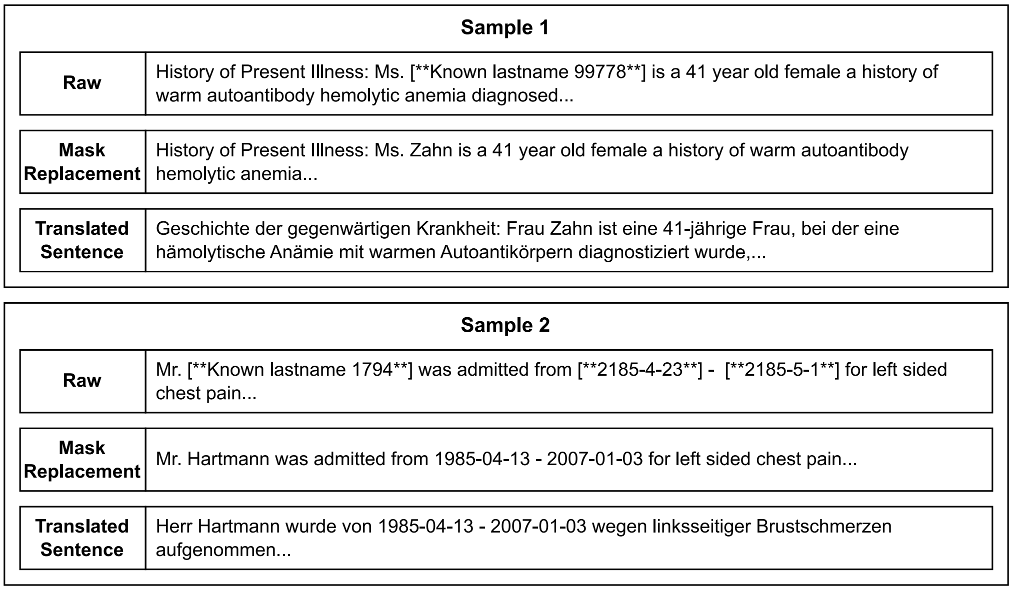

Fig. 1. Effect of mask replacements on the English and German sentences for two exemplary samples.

Our method and our results highlight the effectiveness of non-German data sources for training a German NER model for medical semantic annotation such as medication detection. Our model can surpass the performance of the prior German NLP model GGPONC 2 [5] which is traditionally trained on German text data. In principle, the method is not inherently limited to German because NMT and word alignment techniques also exist for several other languages and therefore, it could be applied to other languages as well.

1.1. Related work

In the recent decade, in particular in the last five years, the field of natural language processing has been radically transformed by the use of data-driven, neural methods that are able to surpass previous state-of-the-art performances [2,6]. This development is likewise reflected by several empirical facts such as quantity of published research or project funding [2]. The introduction of the attention-based transformer model [7] in the field of NLP led to various follow-up works such as BERT [8] and similar deep language models that are trained and applied on domain-specific contexts [9–14]. All these domain-specific works share in common that their research focus lies primarily on English application and use.

The training of novel transformer-based German NLP models requires large, well-suited datasets with respect to size and quality. In purely supervised scenarios, this also includes the need for gold-standard annotation labels. While several works with internal datasets exist, their datasets are not shared among the research community and remain undisclosed [15–26], and thus this presents major hurdles for open research and independent reproducibility. The situation on public, English datasets is more convenient and several large datasets like MIMIC-III [27] or the i2b2 challenges with datasets such as the n2c2 2018 dataset [28] have been published, as well as the multilingual Mantra GSC [29] dataset from the biomedical domain. Only in recent years has the German medical NLP research community addressed this issue and developed novel German medical datasets that are publicly accessible as foundation for future NLP work [30,31]. Regarding the GGPONC [30], an updated iteration has been presented [5].

With regards to novel German medical NLP systems, commercial software like *Averbis Health Discovery* [32][1] and *German Spark NLP for Healthcare* [33][2] are proprietary and require licenses. As an exception, *mEx* [34] is freely available, but the model weights can only be requested and used under data use agreement. An updated iteration has been presented as well [35]. For German medical NER tasks, only few public, open neural models are available to the best of our knowledge, such as *GGPONC* [5] and *GERNERMED* [4].

Statement of significance

Summary	Description
Problem or issue	Training data for NLP annotation models is a major limiting factor for successful model training.
What is already known	For several reasons, matching datasets are often not available in a certain target language.
What this paper adds	We combine multiple techniques to utilize data from outside of the target language to obtain a annotation model for our selected target language. Our results show the model's ability to surpass the performance of the baseline model trained traditionally with internal data. Consequently, our work highlights a way to utilize datasets of nontarget languages for a certain target language. We apply our method in the context of medical semantic text annotation in German which is a novel contribution to the field.

2. Methods

2.1. Dataset acquisition

The dataset retrieval pipeline for German texts follows the approach proposed in GERNERMED [4]: As a starting point, the *2018 n2c2 shared task on ADE and medication extraction in EHR* dataset serves as an English source dataset of medical entities from anonymized electronic health records. The English source dataset is decomposed into sentences as the initial preprocessing step. During that process,

[1] https://averbis.com/de/health-discovery/.

[2] https://nlp.johnsnowlabs.com/2021/03/31/ner_healthcare_de.html.

text spans that have been replaced with an anonymized identifier text bracket by the editors of the source dataset are detected and replaced with randomized synthetic data from the *Faker* Python module in order to reduce the number of irregular text occurrences while updating the initial annotation span indices accordingly. For instance, this includes text entities like first and family name, dates and postal addresses. For illustration purposes, two samples from the corpus are shown in Fig. 1.

We apply the publicly available FAIRseq *transformer.wmt19.en-de* [36] NMT model for sentence-wise automatic translation, which features a transformer-based neural model for translating sentences from English to German. Since the annotation information from the English source dataset cannot be directly preserved for German sentences, the reconstruction of the annotation spans for the translated German sentences can be estimated by the means of a bitext word alignment as a postprocessing step. Artifacts in translation and alignment have been discussed for GERNERMED [37]. In contrast to the approach in GERNERMED, we refine the word alignment estimation step in regard to the following aspects:

- **Improved Tokenization**: The tokenization of sentences for the word alignment differs from modern tokenizers that generate sub-word-level tokens optimized through techniques such as byte pair encoding schemes. Most word alignment methods operate on word-level tokenization with whitespace-based token splitting. In order to reduce the number of misaligned words, we further refined the word-level tokenization by separating punctuation from words instead of only relying on tokenization splits on whitespace characters. In our previous work [4], the projected German label spans often included trailing punctuation because a whitespace-based tokenization does not separate trailing punctuation from words and therefore, the label span reconstruction algorithm is unable to differentiate between words and punctuation within a token. This effect impedes subsequent model training but is countered by the improved, punctuation-aware tokenization.
- **Word Alignment Technique**: In NLP bitext word alignment is the task of determining the semantic correspondence between words from a bilingual sentence pair consisting of the source and translated sentence. In previous work, the *Fast_Align* [38] implementation has been used for establishing such correspondences. It uses the IBM 2 alignment model for alignment estimation in a purely unsupervised fashion. While there are also other models inspired by statistical machine translation [39,40], recent work has been done towards neural approaches [41,42]. For this work, we use the pre-trained model from *Awesome-Align* [42]. In short, the model tackles the task by encoding both sentences through a pre-trained cross-lingual language model in order to obtain contextualized word vector embeddings. Although the words of the sentence pairs largely differ with respect to their syntactic and linguistic features, the implementation makes use of the assumption that corresponding words are similar in terms of their word vectors in embedding space in order to find the word correlations in each sentence.

After the translation of the sentences, applying the word alignment estimation on the set of sentence pairs given the refinements for tokenizer and word alignment yields essential information on the relationship between the annotation spans of the English entity labels and their German counterparts. This step is crucial because potentially misaligned labels are further propagated and impede the quality of the dataset and NER scores of the final model. The process is illustrated by Fig. 2.

As a minor disadvantage of the common *Pharaoh* alignment format, the difference in annotation granularity cannot be preserved completely on character level. Even though the annotation spans of the source dataset are provided as character-level indices, the word-level tokenization restricts the ability to reconstruct sub-word-level annotation spans in the German target data when the backprojection of the word-level indices from the word alignment estimation onto the character-level indices of the target sentence text string is evaluated.

Table 1
The distribution of annotations in the (raw) synthesized German dataset in absolute numbers. Note that a single tag sample count may include multiple tokens. The dataset consists of 16 632 sentences. Abbreviations: named entity recognition (NER).

NER tag	Count
Drug	26 003
Route	8 560
Reason	6 244
Strength	10 546
Frequency	9 794
Duration	956
Form	10 546
Dosage	6 700
ADE	1 557

2.2. Entity recognition training

The training of our entity recognition model employs the entity recognition parser from the *SpaCy* library which follows a transducer-based parsing approach [43] with a BILOU [44] scheme (***Begin***, ***Inside***, ***Last***, ***Outside***, ***Unit***; an extension to the IOB [45] scheme) instead of a state-agnostic token tagging approach.

Slim model: Without the use of a transfer-learning-based approach, in *SpaCy* the transformation from discrete tokens into a dense vector representation is implemented by a model that is usually trained from scratch. Such model includes the embedding of the tokens into vectors via Bloom [46] embeddings and further uses convolutional and dense layers to establish context-awareness and feature abstraction.

Transfer-learning: Inspired by the success of transformer-based neural networks and their effectiveness on language modeling through pre-training on large-scale text corpora, transfer-learning-based methods using deep transformer models can also contribute to stronger entity recognition performance by providing contextualized token embeddings through earlier pre-training without the need to train such large models from scratch. As one instance, the masked language model BERT and several descendants have been released with pre-trained weights for various different languages including German, making it well-suited for transfer-learning.

Entity Parsing: The entity parser from the SpaCy implementation is strongly influenced by the state-based text chunking algorithm from Lample et al. [43]. The parser uses the feature vectors from previous stages (such as from the slim model or the transfer-learning approach) and aggregates a feature vector from the current parsing state to predict the next valid action which likewise annotates the current token during NER parsing. The whole process is shown in Fig. 3.

3. Results

3.1. Dataset acquisition

The English source dataset from the *2018 n2c2 shared task on ADE and medication extraction in EHR* consists of 404 annotated text documents. The annotation includes the labels *Strength*, *Form*, *Dosage*, *Route*, *Frequency*, *Drug*, *Duration*, *Reason*, *ADE*. The documents are split into sentences using the SpaCy sentencizer for English texts. After the sentence-wise translation we apply the word alignment step. During this process we discard sentences whenever an annotation label cannot be reconstructed due to incomplete word alignment mappings. We obtain our raw German dataset with 17 938 sentences. The annotation distribution of the raw German dataset is shown in Table 1.

For further clean-up of the raw dataset, sentences that do not contain any entity label at all are discarded from the set of sentences, resulting in a total of 16 632 sentence samples.

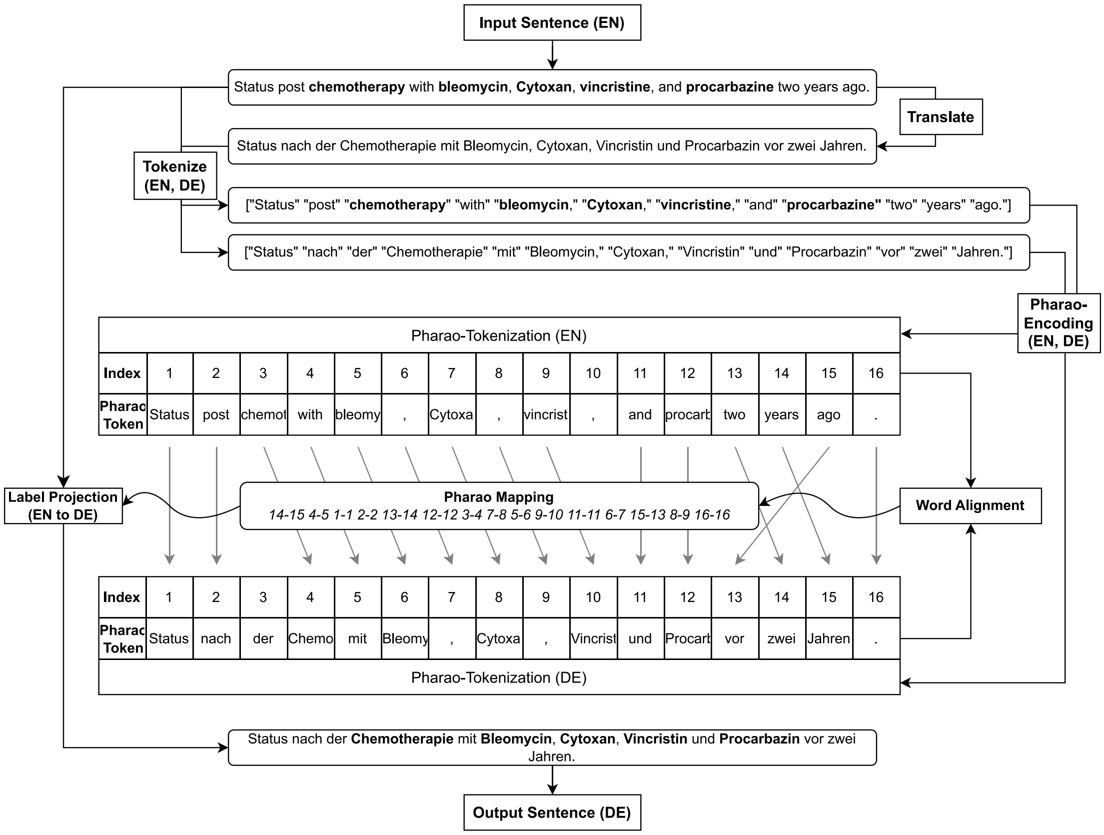

Fig. 2. Whitespace-based tokenization and additional Pharao-based tokenization for word alignment with subsequent annotation projection. Annotations in the text samples are highlighted by bold font. Only *Drug* annotations are shown in this example.

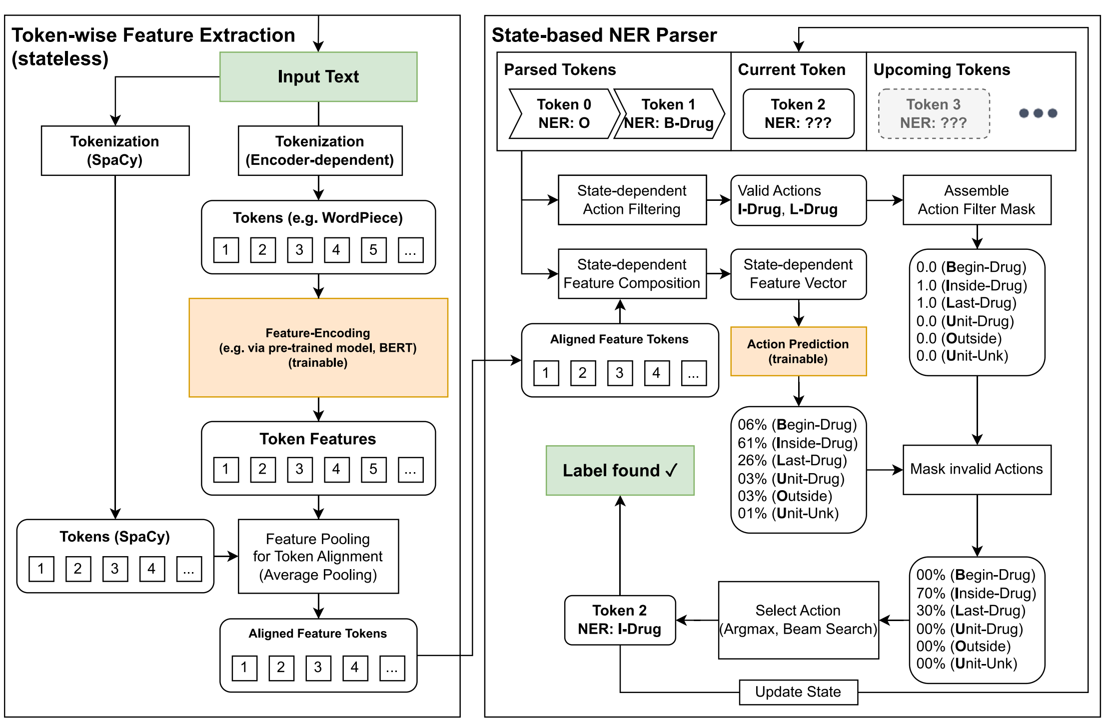

Fig. 3. Logical text processing steps for text encoding and entity parsing in SpaCy. The feature encoding can utilize pre-trained deep embeddings via transfer-learning or SpaCy's native Bloom embeddings [46]. Abbreviations: named entity recognition (NER).

Table 2
Information on the filtered German dataset. Overlapping annotation spans were removed. The following named entity recognition (NER) tags were omitted: Route, Reason, ADE.

Dataset	Split	# Tokens	# Entities	# Sentences
Train set	0.8	293 693	50 955	13 306
Validation set	0.1	37 218	6 420	1 663
Test set	0.1	36 168	6 064	1 663
Total	1.0	367 079	63 439	16 632

3.2. Entity recognition training

For the training of the NER model, we ignore the following annotation labels for the following reasons[3]:

- ***ADE***: The scope of the English source dataset covers the analysis of medical texts with respect to adverse drug effects. We consider the task of detecting adverse drug effects in texts as of lesser general interest and observed low scores in preliminary experiments when we trained a NER model on all labels including *ADE*. In general, the decision on text phrases in the *ADE* class is complex and context-dependent across datasets.
- ***Reason***: Similar to *ADE*, its usefulness depends on the nature of the dataset and the context, and in preliminary experiments the label class yielded low scores.
- ***Route***: While we consider *Route* to be of potential general interest, we found that the label diversity in the English source dataset is quite low. For instance, 5356 times (out of 8560 total *Route* annotations) the phrases' value is "PO". The second most frequent value is "IV" (874 times). We decided to refrain from including the *Route* label class because its lack of diversity yields to high scores on the test set and could lead readers to draw misleading conclusions about the actual annotation capabilities of a model for this label class.

Before the entity recognition model is trained, we split the previously described, filtered German dataset into training, validation and test set (80%,10%,10%). The split statistics are provided in Table 2. Since the IOB-based entity recognition parser requires the annotated dataset to contain only non-overlapping annotation spans, annotation overlaps are resolved by removing the annotation span of shorter length while only preserving the longest span.

We investigate the ability of improving the entity recognition performance by the means of transfer-learning on deep language models on the basis of two German models:

- **German BERT** [47][4] (*bert-base-german-cased*): The model from Deepset AI follows the default architecture of BERT and has been specifically pre-trained on German data. The pre-training dataset stems from German Wikipedia, OpenLegalData, and German news articles.
- **GottBERT** [48]: The model is based on the RoBERTa architecture and has been trained on the OSCAR dataset using the fairseq implementation. OSCAR is a German subset of CommonCrawl.

Both language models are publicly available. We retrieve both models from the Huggingface platform. For fine-tuning the entity recognizer on top of the language model, we utilize SpaCy for training. In this context, the model-specific tokenizer is inherited from the language model.

[3] Experimental results on NER model training for all label classes as well the visualization of class-specific label text distributions are provided as supplementary data.

[4] https://www.deepset.ai/german-bert.

The training was performed on a single Nvidia Titan RTX. The training took 8–47 min (*German BERT*: 47 m, *GottBERT*: 26 m, *Slim*: 8 m). Due to our observations from the preliminary hyperparameter search, we chose to stick to the default hyperparameters from SpaCy (Adam with weight decay, $\alpha = 0.00005$ (GottBERT, GermanBERT)/0.001 (Slim), $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch size = 128 (GottBERT, GermanBERT)/1000 (Slim)) as we did not find major score-wise improvements. In order to measure the differences in performance scores, we also compare the SpaCy Slim model using the same training and test set as baseline model, as well as the publicly available GERNERMED model as static model evaluated on the test set. It should be noted that the GERNERMED model scores must be considered as tainted because its weights are trained on a dataset that might partially contain samples from our test set. For evaluation, the NER procedure is considered as a token-wise multi-class classification problem. We computed the precision (*Pr*), recall (*Re*) and F1 score (*F1*) for each individual label class as well as its respective (class-frequency-weighted) average score (*Total*). The final results on the test set are depicted in Table 3.

Both transfer-learning-based approaches exhibit strong performance in absolute numbers. Though to our surprise, German BERT achieves notably inferior performance scores in direct comparison to GottBERT by 0.7% total F1 score difference. We attribute this performance gap to the differences in pre-training dataset sizes for German BERT (12 GB) and GottBERT (145 GB) and the use of the RoBERTa architecture as for NER such observation and conclusion have been reported and drawn by the authors of GottBERT as well for monolingual models [48].

To verify the robustness of our observations and estimate the degree of a test set selection bias, we re-trained the GottBERT model using 10-fold cross-validation on the dataset. GottBERT was chosen due to its strongest total F1 score in Table 3. The mean and standard deviation of the 10-fold models are provided as well as the distance to the GottBERT results from Table 3 to the mean scores. The results are shown in Table 4.

3.3. Out-of-distribution evaluation

The evaluation on the test set does not provide valuable information on how a model can maintain its scores beyond the scope of the train and test set. A known property of neural networks as statistical models is their ability to overfit to the training dataset. While strong performance on the test set indicates the ability to abstract from individual samples without blunt sample memorization, it cannot measure the model's reliance on the inherent bias of the dataset and its ability to generalize to *out-of-distribution*(OoD) samples. To investigate the OoD generalization ability, we retrieved 30 text samples provided by independent physicians annotated with equivalent labels to our dataset and evaluated the models' performance on this separated dataset. Since the physicians were instructed to use the class labels from our initial dataset, the OoD samples are annotated with matching label classes and can be directly used for full evaluation of our models. The results are shown in Table 5.

The results display the impact of the transfer-learning-based NER models in order to preserve strong performance on OoD data samples. However similar to the results on the test set, German BERT performs inferior to the GottBERT-based model by an increased margin according to the weighted F1 score. In contrast, the baseline models suffer from substantially degraded scores in comparison to their scores on the test set.

Due to the sparseness and independent origin of the OoD dataset, the number of labels is imbalanced across individual class labels and explains that the evaluation scores can yield *1.0* or *0.0* in several situations. While the reliability of the scores in these cases remains a major limitation, the scores still indicate the degree of abstraction beyond the in-distribution bias in other cases, because the evaluation on the test set is unable to quantify such in-distribution biases.

Table 3
Evaluation of models' performance scores on test set for the labels **Str**ength, **Dur**ation, **Form**, **Dos**age, **Drug** and **Freq**uency. **Pr**ecision, **Re**call and **F1**-scores are evaluated. Abbreviations: named entity recognition (NER).

Scores on test set		NER tags						
Model		Str	Dur	Form	Dos	Drug	Freq	Total
GERNERMED++ (GottBERT)	Pr	**0.971**	0.806	0.947	**0.967**	**0.969**	**0.880**	**0.942**
	Re	0.964	**0.825**	**0.969**	**0.971**	0.923	0.953	**0.950**
	F1	**0.967**	**0.815**	0.958	0.969	0.945	**0.915**	**0.946**
GERNERMED++ (GermanBERT)	Pr	0.944	0.791	0.956	0.963	**0.969**	0.859	0.932
	Re	**0.973**	**0.825**	0.962	**0.971**	**0.933**	0.924	0.947
	F1	0.958	0.807	**0.959**	0.967	**0.951**	0.890	0.939
GERNERMED++ (SpaCy Slim)	Pr	0.965	**0.823**	**0.965**	0.958	0.929	0.855	0.926
	Re	0.967	0.749	0.950	**0.971**	0.884	**0.966**	0.941
	F1	0.966	0.784	0.957	0.964	0.906	0.907	0.932
GERNERMED [4][a]	Pr	0.916	0.613	0.842	0.915	0.644	0.739	0.790
	Re	0.917	0.697	0.882	0.959	0.634	0.901	0.841
	F1	0.917	0.652	0.861	0.937	0.639	0.812	0.814

Note:
[a] Specific training set might be tainted by samples from the test set.

Table 4
Averaged scores of test folds from 10-fold cross-validation for labels **Str**ength, **Dur**ation, **Form**, **Dos**age, **Drug** and **Freq**uency. All fold-wisely trained models are based on GottBERT. For reference, the score differences to the presented GottBERT model from Table 6 are given. Abbreviations: named entity recognition (NER), standard deviation (std dev), difference to reference (diff to ref).

10-fold Cross-validation		NER tags						
(GottBERT model)		Str	Dur	Form	Dos	Drug	Freq	Total
Precision	μ (mean)	0.967	0.798	0.964	0.962	0.938	0.961	0.950
	σ (std dev)	0.008	0.043	0.012	0.015	0.012	0.009	0.004
	Δ (diff to ref)	−0.004	−0.008	0.017	−0.005	−0.031	0.081	0.008
Recall	μ (mean)	0.967	0.841	0.953	0.958	0.958	0.863	0.939
	σ (std dev)	0.010	0.066	0.010	0.010	0.010	0.014	0.008
	Δ (diff to ref)	0.003	0.016	−0.016	−0.013	0.035	−0.09	−0.011
F1	μ (mean)	0.967	0.817	0.958	0.960	0.948	0.909	0.944
	σ (std dev)	0.006	0.033	0.006	0.010	0.008	0.008	0.004
	Δ (diff to ref)	0.000	0.002	0.000	−0.009	−0.003	0.003	−0.002

Table 5
Evaluation of models' performance scores on separated out-of-distribution (OoD) dataset for the labels **Str**ength, **Dur**ation, **Form**, **Dos**age, **Drug** and **Freq**uency. **Pr**ecision, **Re**call and **F1**-scores are evaluated. Abbreviations: named entity recognition (NER).

Scores on OoD Dataset		NER tags						
Model		Str	Dur	Form	Dos	Drug	Freq	Total
GERNERMED++ (GottBERT)	Pr	0.866	**1.000**	**1.000**	**0.125**	**0.891**	**0.923**	**0.883**
	Re	**0.960**	0.400	**0.632**	**0.250**	0.932	0.615	**0.835**
	F1	**0.911**	0.571	**0.774**	**0.167**	**0.911**	**0.738**	**0.845**
GERNERMED++ (GermanBERT)	Pr	**0.955**	**1.000**	0.909	0.077	0.830	0.456	0.817
	Re	0.832	**0.800**	0.526	**0.250**	**1.000**	**0.667**	0.797
	F1	0.889	**0.889**	0.667	0.118	0.907	0.542	0.794
GERNERMED++ (SpaCy Slim)	Pr	0.951	0.000	**1.000**	0.111	0.690	0.486	0.778
	Re	0.772	0.000	0.316	**0.250**	0.659	0.462	0.623
	F1	0.852	0.000	0.480	0.154	0.674	0.474	0.679
GERNERMED	Pr	0.851	0.000	0.500	0.045	0.460	0.390	0.619
	Re	0.624	0.000	0.158	0.250	0.523	0.410	0.500
	F1	0.720	0.000	0.240	0.077	0.489	0.400	0.541
#Labels		37	3	19	4	36	20	119

3.4. Related datasets

We select three relevant datasets in order to further evaluate our models. To put our results in perspective, we also evaluate the reference model from GGPONC [5] on these datasets. The entity labels from the datasets differ from the labels of our training dataset and our OoD dataset. This limits our ability to perform a complete comparison of our model with respect to all label classes. All related datasets provide annotation information on entities that we consider to be semantically strongly related to the class label *Drug*, although the datasets commonly lack clear and homogeneous definitions on their label classes. We evaluate the scores as a classification task on token- and character-level. The results are shown in Table 6.

To no surprise, the GGPONC reference model archives better performance on its native GGPONC dataset [30], yet all our models with transfer-learning-based, pre-trained BERT encoder outperform the reference model, our slim model and the baseline GERNERMED model. Considering that the baseline GGPONC model was developed in traditional fashion using a manually crafted German dataset, the archived performance margins from both GottBERT- and GermanBERT-based

Table 6
Evaluation of models' F1 scores on related dataset. The GGPONC reference model [5] is evaluated for comparison. To allow fair comparison, only Drug-related label classes are selected. Annotations from the GGPONC [30] dataset do not align onto the tokens from the SpaCy tokenizer and are therefore omitted. **Pr**ecision, **Re**call and **F1**-scores are evaluated.

Scores on related datasets		F1 scores	
Model/Dataset		Drug (char-wise)	Drug (token-wise)
Medline Dataset [29]		Drug = CHEM	
GERNERMED++ (GottBERT)	Pr	0.858	0.837
	Re	**0.701**	**0.706**
	F1	**0.772**	0.766
GERNERMED++ (GermanBERT)	Pr	**0.885**	**0.875**
	Re	0.638	0.686
	F1	0.742	**0.769**
GERNERMED++ (SpaCy Slim)	Pr	0.437	0.500
	Re	0.182	0.216
	F1	0.257	0.301
GERNERMED	Pr	0.477	0.414
	Re	0.207	0.235
	F1	0.288	0.300
GGPONC [5]	Pr	0.822	0.771
	Re	0.488	0.529
	F1	0.612	0.628
GGPONC Dataset [30]		Drug = Chemicals_Drugs	
GERNERMED++ (GottBERT)	Pr	0.535	n/a
	Re	0.664	n/a
	F1	0.592	n/a
GERNERMED++ (GermanBERT)	Pr	0.522	n/a
	Re	0.645	n/a
	F1	0.577	n/a
GERNERMED++ (SpaCy Slim)	Pr	0.185	n/a
	Re	0.433	n/a
	F1	0.260	n/a
GERNERMED	Pr	0.089	n/a
	Re	0.303	n/a
	F1	0.138	n/a
GGPONC [5]	Pr	**0.636**	n/a
	Re	**0.737**	n/a
	F1	**0.683**	n/a
BRONCO Dataset [31]		Drug = MEDICATION	
GERNERMED++ (GottBERT)	Pr	0.673	0.726
	Re	**0.789**	**0.752**
	F1	**0.726**	**0.739**
GERNERMED++ (GermanBERT)	Pr	**0.684**	**0.730**
	Re	0.677	0.637
	F1	0.680	0.680
GERNERMED++ (SpaCy Slim)	Pr	0.320	0.378
	Re	0.512	0.486
	F1	0.394	0.425
GERNERMED	Pr	0.155	0.148
	Re	0.478	0.482
	F1	0.234	0.227
GGPONC [5]	Pr	0.573	0.346
	Re	0.449	0.430
	F1	0.504	0.384

models are unexpected. Throughout the tasks, the GottBERT-based model beats the GermanBERT-based model which is consistent with previous observations.

4. Discussion

Our results indicate strong performance of all models on the test set, however our evaluation on the OoD dataset as well as on external, related datasets shows the impact of using the transfer-learning abilities of pre-trained BERT-based feature encoders to solidify the robust performance on such external datasets. Considering the fact that our models were developed without additional manual work of annotating datasets and only a public non-German dataset was used, the obtained models compete surprisingly well with the pre-existing reference model and are able to outperform it on independent datasets. The lack of more independent annotated datasets, lacking matching annotation labels and unclear label class definitions still limit the possibility to deeper evaluate and compare novel models and methods. In this context, the small sample size of our OoD dataset remains a major limitation of our work and emphasizes the continuous need for German medical corpora with diverse label annotations.

In general, considering the current poor availability of open medical NLP systems for non-English natural languages as well as for German in particular, our refined approach demonstrates a powerful opportunity to build a strong medical NER model solely by the use of a public English dataset.

5. Conclusion

In this work, we presented a fine-tuned German NER model for semantic medical entity annotation using deep pre-trained language models by the means of transfer-learning. We demonstrated its ability to outperform the basic baseline model on the test set and on an out-of-distribution dataset. In comparison to the existing GGPONC reference model, we showed competitive results on external datasets and outperformed the reference model on all independent datasets. Furthermore, we described the process and its relevant improvements to obtain a medical-specific German dataset without the use of internal data. Our open NER model is publicly available for third-party use on GitHub.

CRediT authorship contribution statement

Johann Frei: Conceptualization, Methodology, Software, Investigation, Validation, Formal analysis, Writing – original draft. **Ludwig Frei-Stuber:** Resources, Data curation, Clinical partner. **Frank Kramer:** Supervision, Project administration, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors have declared that no competing interests exist.

Acknowledgment

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

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104513.

References

[1] Johannes Starlinger, Madeleine Kittner, Oliver Blankenstein, Ulf Leser, How to improve information extraction from German medical records, IT - Inf. Technol. 59 (4) (2017) 171–179, Publisher: De Gruyter Oldenbourg.

[2] Honghan Wu, Minhong Wang, Jinge Wu, Farah Francis, Yun-Hsuan Chang, Alex Shavick, Hang Dong, Michael T.C. Poon, Natalie Fitzpatrick, Adam P. Levine, Luke T. Slater, Alex Handy, Andreas Karwath, Georgios V. Gkoutos, Claude Chelala, Anoop Dinesh Shah, Robert Stewart, Nigel Collier, Beatrice Alex, William Whiteley, Cathie Sudlow, Angus Roberts, Richard J.B. Dobson, A survey on clinical natural language processing in the United Kingdom from 2007 to 2022, NPJ Digit. Med. 5 (1) (2022) 1–15, Number: 1 Publisher: Nature Publishing Group.

[3] Wendy W. Chapman, Prakash M. Nadkarni, Lynette Hirschman, Leonard W. D'Avolio, Guergana K. Savova, Ozlem Uzuner, Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions, J. Am. Med. Inform. Assoc.: JAMIA 18 (5) (2011) 540–543.

[4] Johann Frei, Frank Kramer, GERNERMED: An open German medical NER model, Softw. Impacts 11 (2022) 100212.
[5] Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, Matthieu-P. Schapranow, GGPONC 2.0 - the German clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER taggers, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, 2022, pp. 3650–3660.
[6] Tian-Xiang Sun, Xiang-Yang Liu, Xi-Peng Qiu, Xuan-Jing Huang, Paradigm shift in natural language processing, Mach. Intell. Res. 19 (3) (2022) 169–183.
[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, CoRR, abs/1810.04805. _eprint: 1810.04805.
[9] Yifan Peng, Shankai Yan, Zhiyong Lu, 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, pp. 58–65.
[10] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, Degui Zhi, Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, NPJ Digit. Med. 4 (1) (2021) 1–13, Publisher: Nature Publishing Group.
[11] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2019).
[12] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, Matthew McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, 2019, pp. 72–78.
[13] Iz Beltagy, Kyle Lo, Arman Cohan, SciBERT: Pretrained language model for scientific text, in: EMNLP, 2019, _eprint: arXiv:1903.10676.
[14] Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, Hong Yu, Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: An empirical study, JMIR Med. Inform. 7 (3) (2019) e14830.
[15] Joachim Wermter, Udo Hahn, An annotated German-language medical text corpus as language resource, in: LREC, Citeseer, 2004.
[16] Georg Fette, Maximilian Ertl, Anja Wörner, Peter Kluegl, Stefan Störk, Frank Puppe, Information extraction from unstructured electronic health records and integration into a data warehouse, in: INFORMATIK 2012, Gesellschaft für Informatik eV, 2012.
[17] Claudia Bretschneider, Sonja Zillner, Matthias Hammon, 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, pp. 27–35.
[18] Martin Toepfer, Hamo Corovic, Georg Fette, Peter Klügl, Stefan Störk, Frank Puppe, Fine-grained information extraction from German transthoracic echocardiography reports, BMC Med. Inform. Decis. Mak. 15 (1) (2015) 1–16, Publisher: Springer.
[19] Markus Kreuzthaler, Stefan Schulz, Detection of sentence boundaries and abbreviations in clinical narratives, in: BMC Medical Informatics and Decision Making, Vol. 15, BioMed Central, 2015, pp. 1–13.
[20] Roland Roller, Hans Uszkoreit, Feiyu Xu, Laura Seiffe, Michael Mikhailov, Oliver Staeck, Klemens Budde, Fabian Halleck, Danilo Schmidt, A fine-grained corpus annotation schema of German nephrology records, in: Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), 2016, pp. 69–77.
[21] Viviana Cotik, Roland Roller, Feiyu Xu, Hans Uszkoreit, Klemens Budde, Danilo Schmidt, Negation detection in clinical reports written in German, in: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), 2016, pp. 115–124.
[22] Jonathan Krebs, Hamo Corovic, Georg Dietrich, Max Ertl, Georg Fette, Mathias Kaspar, Markus Krug, Stefan Störk, Frank Puppe, Semi-automatic terminology generation for information extraction from German chest x-ray reports, GMDS 243 (2017) 80–84.
[23] Udo Hahn, Franz Matthies, Christina Lohr, Markus Löffler, 3000Pa-towards a national reference corpus of German clinical language, in: MIE, 2018, pp. 26–30.
[24] Jose A. Miñarro-Giménez, Ronald Cornet, Marie-Christine Jaulent, Heike Dewenter, Sylvia Thun, Kirstine Rosenbeck Gøeg, Daniel Karlsson, Stefan Schulz, Quantitative analysis of manual annotation of clinical text samples, Int. J. Med. Inform. 123 (2019) 37–48, Publisher: Elsevier.
[25] Maximilian König, André Sander, Ilja Demuth, Daniel Diekmann, Elisabeth Steinhagen-Thiessen, Knowledge-based best of breed approach for automated detection of clinical events based on German free text digital hospital discharge letters, PLoS One 14 (11) (2019) 0224916, Publisher: Public Library of Science San Francisco, CA USA.
[26] Anton Schäfer, Nils Blach, Oliver Rausch, Maximilian Warm, Nils Krüger, Towards automated anamnesis summarization: BERT-based models for symptom extraction, 2020, arXiv:2011.01696 [cs].
[27] Tom J. Pollard, Alistair E.W. Johnson, The MIMIC-III clinical database, 2016.
[28] 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.
[29] Jan A. Kors, Simon Clematide, Saber A. Akhondi, Erik M. van Mulligen, Dietrich Rebholz-Schuhmann, A multilingual gold-standard corpus for biomedical concept recognition: the mantra GSC, J. Am. Med. Inform. Assoc. 22 (5) (2015) 948–956.
[30] 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.
[31] Madeleine Kittner, Mario Lamping, Damian T. Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, Marit Zettwitz, Till de Bortoli, Leonie Ostermann, Jurica Ševa, Johannes Starlinger, Oliver Kohlbacher, Nisar P. Malek, Ulrich Keilholz, Ulf Leser, Annotation and initial evaluation of a large annotated German oncological corpus, JAMIA Open 4 (2) (2021) ooab025.
[32] Averbis Health Discovery - Analyse Von Patienten Daten, Averbis Gmbh.
[33] Detect Symptoms, Treatments and Other Entities in German- Spark NLP Model, John Snow Labs Inc., 2021.
[34] Roland Roller, Christoph Alt, Laura Seiffe, He Wang, 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 (SWAT4HCLS'2018). Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4HCLS-2018), December 3-5, Antwerp, Belgium, 2018.
[35] Roland Roller, Laura Seiffe, Ammer Ayach, Sebastian Möller, Oliver Marten, Michael Mikhailov, Christoph Alt, Danilo Schmidt, Fabian Halleck, Marcel Naik, Wiebke Duettmann, Klemens Budde, A medical information extraction workbench to process German clinical text, 2022.
[36] 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.
[37] Johann Frei, Frank Kramer, German medical named entity recognition model and data set creation using machine translation and word alignment: Algorithm development and validation, JMIR Form. Res. 7 (1) (2023) e39077, Company: JMIR Formative Research Distributor: JMIR Formative Research Institution: JMIR Formative Research Label: JMIR Formative Research Publisher: JMIR Publications Inc. Toronto, Canada.
[38] Chris Dyer, Victor Chahuneau, Noah A. Smith, A simple, fast, and effective reparameterization of IBM model 2, in: Lucy Vanderwende, Hal Daumé, Katrin Kirchhoff (Eds.), Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, The Association for Computational Linguistics, 2013, pp. 644–648.
[39] Franz Josef Och, Hermann Ney, A systematic comparison of various statistical alignment models, Comput. Linguist. 29 (1) (2003) 19–51.
[40] Robert Östling, J. Tiedemann, Efficient word alignment with Markov chain Monte Carlo, Prague Bull. Math. Linguist. (2016).
[41] Masoud Jalili Sabet, Philipp Dufter, François Yvon, Hinrich Schütze, SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 1627–1643.
[42] Zi-Yi Dou, Graham Neubig, Word alignment by fine-tuning embeddings on parallel corpora, 2021, arXiv:2101.08231 [cs].
[43] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer, Neural architectures for named entity recognition, 2016.
[44] R. Grishman, Andrew Borthwick, A maximum entropy approach to named entity recognition, 1999.
[45] Lance Ramshaw, Mitch Marcus, Text chunking using transformation-based learning, in: Third Workshop on Very Large Corpora, 1995.
[46] Lester.James Miranda, Ákos Kádár, Adriane Boyd, Sofie Van Landeghem, Anders Søgaard, Matthew Honnibal, Multi hash embeddings in spacy, 2022.
[47] Branden Chan, Timo Möller, Malte Pietsch, Tanay. Soni, German BERT - state of the art language model for German NLP, 2019.
[48] Raphael Schreible, Fabian Thomczyk, P. Tippmann, V. Jaravine, M. Boeker, GottBERT: a pure German language model, 2020, arXiv.