Annotated dataset creation through large language models for non-english medical NLP

Johann Frei, Frank Kramer

Angaben zur Veröffentlichung / Publication details:

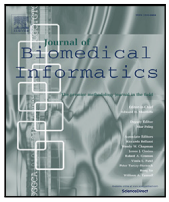
Frei, Johann, and Frank Kramer. 2023. "Annotated dataset creation through large language models for non-english medical NLP." *Journal of Biomedical Informatics* 145 (August): 104478. https://doi.org/10.1016/j.jbi.2023.104478.

Journal of Biomedical Informatics 145 (2023) 104478

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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



Annotated dataset creation through large language models for non-english medical NLP

Johann Frei *, Frank Kramer

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

ARTICLE INFO

Keywords:
Natural language processing
Information extraction
Named entity recognition
Data augmentation
Knowledge distillation
Medication detection

ABSTRACT

Obtaining text datasets with semantic annotations is an effortful process, yet crucial for supervised training in natural language processing (NLP). In general, developing and applying new NLP pipelines in domain-specific contexts for tasks often requires custom-designed datasets to address NLP tasks in a supervised machine learning fashion. When operating in non-English languages for medical data processing, this exposes several minor and major, interconnected problems such as the lack of task-matching datasets as well as task-specific pre-trained models.

In our work, we suggest to leverage pre-trained large language models for training data acquisition in order to retrieve sufficiently large datasets for training smaller and more efficient models for use-case-specific tasks. To demonstrate the effectiveness of your approach, we create a custom dataset that we use to train a medical NER model for German texts, GPTNERMED, yet our method remains language-independent in principle. Our obtained dataset as well as our pre-trained models are publicly available at https://github.com/frankkramer-lab/GPTNERMED.

1. Introduction

In situations of low-resource languages, neural baseline techniques for specific tasks in natural language processing (NLP) are often difficult to be applied successfully due to the lack of sufficient and adequately annotated training data. While English can be perceived as the most relevant language in the field of NLP research as being a high-resource language, effectively any other language can be considered as a rather low-resource language in contrast. Yet the abundance of plain textual resources is no uniquely decisive factor when it comes to dealing with embedded NLP problems in real-life applications. In this regard, a domain-specific dataset needs to be obtained to match the applied context, and the underlying data acquisition process can involve access to highly restricted data, manual engagements from domain experts or time- and cost-intensive data gathering. Another concern relates to the actual NLP objective of the use case and usually heavily determines the final design of the obtained dataset and its collection of task-related annotations.

We study the use case to annotate certain medical entity classes in German throughout this paper since it is an instance that suffers from all formerly mentioned challenges. In this work, we demonstrate an effective method for synthesizing a custom, domain-aligned dataset with annotation information in an unsupervised fashion. Furthermore, we show evidence of its effectiveness by training a generic medical model for German medical named entity recognition (NER) by finetuning a pre-trained BERT language model along with a classification head for NER. Due to the inherently generic nature of our work, we do not see fundamental obstacles in applying the approach to related entity classes in medical or even non-medical tasks, or for different non-English languages of similar quantitative levels of resource abundance.

2. Background and related work

2.1. Medical datasets

In NLP, deep learning-based methods have been proven highly effective in order to tackle frequent tasks, most notably the self-attention-mechanism-based transformer architecture [1]. One fundamental problem of deep learning-based methods remains to be the need for vast amounts of data for training, including corresponding annotations for supervised learning.

In English medical NLP, these challenges have been addressed to a certain extent by the availability of annotated datasets, such as the MIMIC-III [2] and MIMIC-IV [3] datasets or n2c2 datasets from the i2b2 challenges [4]. In general, multilingual textual datasets are

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

[1] UFAL Medical Corpus (accessed at 22.08.2022): https://ufal.mff.cuni.cz/ufal_medical_corpus.

https://doi.org/10.1016/j.jbi.2023.104478
Received 5 January 2023; Received in revised form 1 August 2023; Accepted 21 August 2023
Available online 23 August 2023

available that carry medical texts from multiple languages. The datasets often entail parallel corpora for translation tasks and lack semantic annotation like the *UFAL Medical Corpus*[1] for the WMT'17 biomedical challenge [5]. Driven by manual annotation work, Mantra GSC [6] is a public gold-standard annotated corpus with multilingual texts based on prior parallel corpora and provides limited UMLS information.

For German medical NLP, the field has made notable advances in terms of available datasets. While work in this field of NLP has been published, internal and proprietary datasets are frequently used as underlying datasets [7–17]. In recent years, semi-publicly available datasets like BRONCO [18] and GGPONC 1.0 [19] and 2.0 [19] have been made available. While BRONCO is advertised as based on real discharge letters with annotations, other datasets like GGPONC originate from non-clinical or synthetic data sources like clinical practice guidelines or are assembled from multiple, diverse sources or crawled data from the web. If annotation data is provided, such metadata differ in terms of entity types, entity type definitions or their overall task objectives. Hence, a direct comparison of datasets and corresponding models cannot be made directly with respect to NER F1/tagging scores, or entity linking to different ontologies. Only metrics of rather limited interest such as test set performance of trained models, or token size and number of entities for a dataset are directly derivable for comparison. For an extensive overview of the recent state of German medical datasets, we point to Borchert et al. [19].

2.2. Medical models and applications

We restrict our focus on models and applications to items of general interest and practical applicability. Most works from the presented dataset section develop accompanied models to the datasets and publish internally evaluated scores. However, in many cases, the reproducibility of the described results is not possible since models are not made publicly available along with the paper. Furthermore, some models or systems are designed for narrow NLP tasks and are not of interest for general application in the field, like cardiography texts [11]. Since models are trained on sensitive training data, privacy concerns arise from the fact that potential training data extraction attacks could uncover patient-related data. This concern is amplified by the increasing use of fine-tuning larger language models that are susceptible to such attacks [20]. In the German domain, the neural German model GERNERMED [21] avoids this issue by using public data from English in combination with neural machine translation to be the first publicly available model with unrestricted access and further improved their method for stronger models [22]. Authors from GGPONC [19] and BRONCO [18] provide access to their own models after registration or signed user agreement. On a broader perspective, the software mEx [23] provides an entire stack of different models and dockerized software layers to serve an integrated text processing system, their models can be obtained on request through signed user agreement [24]. Commercial applications from Health Discovery (Averbis)[2] and SparkNLP (John Snow labs)[3] are available but are purely proprietary applications. Contrary to perceptions of domain experts and reviewers, Amazon Comprehend Medical[4] does *not* support German texts at the time of writing. Popular, open solutions like Apache cTAKES [25] and MetaMaps [26] do not exist for the German community. Due to the rapid change in the field, we do not consider this list of available models and software as conclusive. We point to Roller et al. [24] and Borchert et al. [19] for a more exhaustive enumeration of available models and systems.

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

[3] https://nlp.johnsnowlabs.com/analyze_medical_text_german

[4] https://docs.aws.amazon.com/comprehend-medical/latest/dev/comprehendmedical-welcome.html

2.3. Language model-based dataset generation

Data augmentation is a popular technique in the Machine Learning community, in which the objective is to sample new data points from the manifold that models the set of known data points. In computer vision, semantic invariance applies to basic image transformation in many situations [27]. However in NLP, such basic techniques cannot always be applied if semantic information of sentences needs to be preserved, but more sophisticated approaches are used such as back-translation [28] of words or phrases through translation, yet failures in translation can jeopardize the augmentation method [29]. The idea to use pre-trained language models for data augmentation has been proposed as an effective method for augmenting small datasets [30,31] or even creating datasets nearly from scratch [32,33]. With the increasing popularity of large, prompt-based language models like GPT-2/3 [34,35] and open source counterparts [36,37], methods with various objectives have been developed to improve the quality and usefulness of the models in different contexts such as sentence similarity estimation [33]. In addition to classical few-shot text generation, task instruction-driven zero-shot methods are likewise an active field of research [33,38,39]. For medical NLP purposes, text generation has been shown for synthesizing EHR reports [40] and its application for downstream tasks [41] using a GPT-2 model. To the best of our knowledge, we are the first team to expand the general idea to the field of German medical NLP.

3. Methods

In this work, we leverage the capabilities of pre-trained language models in regard to their example-driven few-shot learning for text generation. The method follows the basic idea implemented in various related contexts [30,33,40]. We apply the GPT NeoX language model from EleutherAI [37] for input processing and text generation. The model implementation is kept close to the GPT-2/3 architecture, an autoregressive model which is closely related to the vanilla Transformer architecture [1] with decoder-only blocks. Note that we do not perform gradient-based fine-tuning of the model on novel data, but the model is only used for inference. In difference to other models like GPT-3 [35], the internal model weights are publicly available similar to its smaller GPT-J [36] model. We decided to use the NeoX model over GPT-J due to its larger size[5] which has been shown to exceed the performance of GPT-J on several tasks [37] yet being sufficiently small to run on our local instance. In addition, large multilingual language models are able to improve task performance on low-resource languages (e.g. German) by the multilingual knowledge transfer from a high-resource language (e.g. English) [42].

As previously discussed, LM-based text generation models are used to generate their respective text output by conditioning on an input text sequence, highlighting two main aspects of the input sequence design. First, the sequence can carry a task description in natural language to advise the model on its task objective. While writing an obvious prompt command seems obvious to a normal person, the performance of language models varies between different semantically equivalent task instructions [33]. Second, the input can inject information on the task during the prediction of the next word by providing text examples in the input sequence.

In this work, we do not focus on tuning task instructions in natural language but rather demonstrate that straightforward few-shot-learning-based example prompting as model input suffices for synthetic dataset generation within the scope of our use case. To avoid the issue of only generating plain natural text without valuable annotation metadata, we design our input prompt in the style of a simple markup language, where the language model reads the data as a collection of sentences. Each sentence is enclosed by `<s>` and `</s>` signs and

[5] Model parameter size (billions): GPT-J: 6B, GPT NeoX: 20B, GPT-3: 175B.

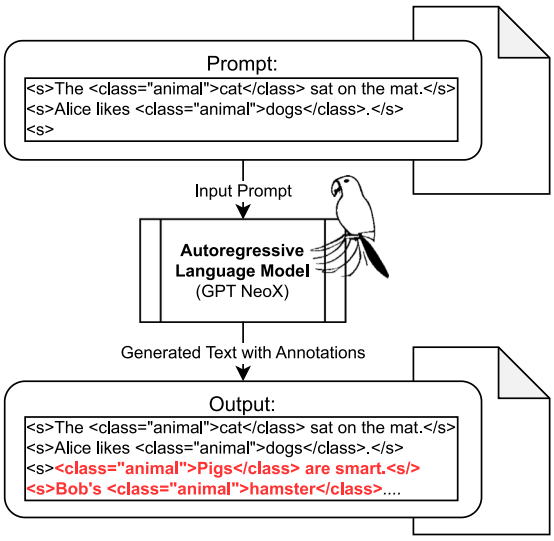


Fig. 1. Synthesis of markup-based text with annotation information: The input prompt consists of the markup-encoded text of a few pre-written sentences. The set of sentences is augmented by the language model that generates new data (red) token-wise in an autoregressive fashion.

separated by a line break. For each sentence, every word from a certain label class l is enclosed by `<class="l">` and `</class>` respectively. We select a small set of exemplary sentences, encode them according to the basic markup rules and append the opening sentence tag `<s>` to the prompt to indicate the start of an additional sentence. The whole process is illustrated in Fig. 1.

In language generation, the unnormalized probabilities over tokens, referred to as logits, are normalized and smoothed by the last softmax layer in the network

$$softmax(l_i) = \frac{e^{l_i/\tau}}{\sum_j^n e^{l_j/\tau}} \tag{1}$$

where n is the number of tokens in the vocabulary, l_i is the unnormalized predicted probability for token i. The temperature parameter τ is used for smoothing the normalized probability distribution. In this, higher values of τ increase the probabilities for less probable tokens at the expense of highly probable tokens. We can utilize the parameter to reduce the risk of generating invalid markup-based text data by setting the temperature to $0 < \tau < 1.0$ in combination with $top-p < 1.0$ prior to token sampling.

After collecting the output data, we parse the markup text to obtain a synthetic, silver-standard corpus along with its corresponding annotations. For further data cleansing, we only keep sentences that fulfill the following requirements: First, the sentence needs to have a closing `</s>` tag. Second, the parsing of the sentence can succeed and the annotations are provided by valid `<class="l">` and `</class>` tags. Third, the sentence has at least one annotation. Fourth, all annotation labels are part of the pre-defined set of label classes. Fifth, duplicate sentences are reduced to unique occurrences (deduplication).

The synthesis of annotated sentences from a large language model and its transfer to a smaller, more efficient model can be considered a high-level form of knowledge distillation: For the very purpose of developing a German NER model for medical entities, we are able to transform the implicit knowledge of the 20B parameter model about this very context into a dedicated NER model with a faster, less resource-intensive computational footprint. In fact, these properties align well with the aim of the practical applicability of our method and its resulting model in dedicated domain contexts. For the development of a robust NER model, we train a neural-based NER parser from the open-source SpaCy NLP library on our dataset. While the NER parser component is trained from scratch, its input vectors are generated through a pre-trained BERT-based encoder model to improve the performance of the final model through transfer learning and contextualization. The BERT-based encoder is fine-tuned to the data by gradient update during the training procedure.

The data acquisition process and model training are shown in Fig. 2.

Table 1
Iterative data cleansing: About half of the predicted sentences have been removed. The majority of sentence removals are due to the duplicate removal filter. All percentage numbers are rounded.

Applied Filter	#Sentences	% of Baseline	Impact
Baseline	17776	100%	
↪ No `</s>` tag	16603	93%	15%
↪ Duplicates removal	11328	64%	66%
↪ Invalid syntax removal	11326	64%	0%
↪ Invalid or no labels	9845	55%	18%
⟹ Final	9845	55%	

Table 2
Annotation statistics: Number of counts and tokens of each label class. The SpaCy tokenizer is used for tokenization.

Label	Count	#Tokens
Medikation	9868	10138
Dosis	7547	15845
Diagnose	5996	7656

4. Results

4.1. Dataset sampling

We provide the GPT-based NeoX model with an input sequence of twelve sentences in German language, encoded in the described markup style. The sentences are pre-annotated with the label classes *Medikation* (medication/drug), *Dosis*(dosage/strength) and *Diagnose* (diagnosis). The label classes are chosen to match the focus of the national core dataset of the German Medizininformatik-Initiative [43] on treatment (medication and strength) and diagnosis topics. The prompt which we use for the dataset sampling is displayed in Fig. 3.

During inference, we set τ to 0.8 and $top-p$ to 0.9 for language generation and sample 1000 different outputs with a maximum length of 768 tokens each, and additional 100 outputs with an increased temperature τ set to 0.9. Given the parameters, we obtain a raw *baseline* dataset of 17776 sentences which we reduce to 9845 sentences after the different filters were applied, as shown in Table 1. The final dataset consists of 121027 tokens according to the SpaCy tokenizer (245107 tokens according to the GPT tokenizer) with annotations for *Dosis* (# 7547), *Medikation* (# 9868) and *Diagnose* (# 5996) as shown in Table 2.

The inference was computed on an NVIDIA DGX workstation with two NeoX models running in parallel on different A100 GPUs. Regarding the ecological impact, the inference took a total of 118 h of compute time, which results in an estimated GPU power consumption of 35,400 Wh and about 15 kg of carbon emissions.[6]

We randomly sampled ten sentences from the final dataset to be presented in Fig. 4 for a demonstration of the actual sentences and annotation structure. In general, the text structure of the ten sampled sentences appears genuine and no obviously broken text structure is generated, yet its medical validity is limited in some instances (e.g. *Trimethoprim* with unusual *1600 mg* strength). The annotation of the drug label class (Medikation) is correct except for one false-positive

[6] According to the United States Environmental Protection Agency: https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator.

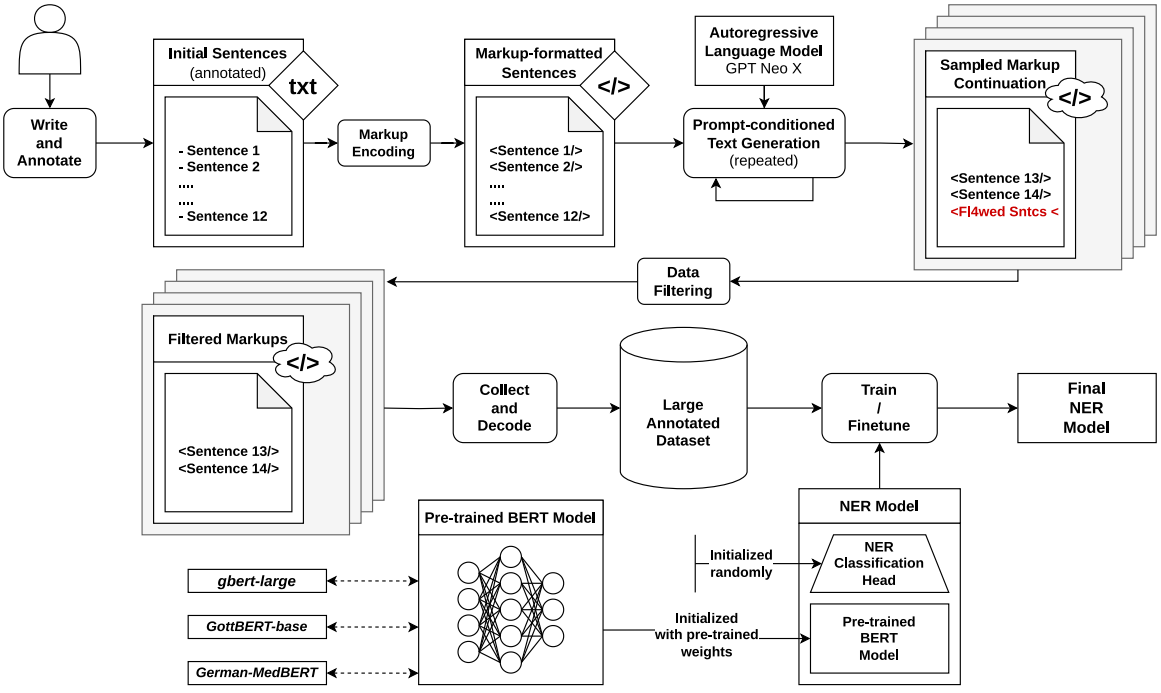


Fig. 2. Illustration of the data acquisition and model training process: The process starts with twelve annotated sentences written by a user and ends with a final NER model. Abbreviations: named entity recognition (NER).

```
<s>Zur weiteren Bekämpfung des <class="Diagnose">Juckreiz</class> wird die Einnahme von tä
    ↪ glich <class="Dosis">100mg</class> <class="Medikation">Cortison</class> empfohlen.</
    ↪ s>
<s>Bei wiederkehrender Infektion wie einer <class="Diagnose">Sepsis</class> oder schweren <
    ↪ class="Diagnose">Pneumonien</class> wird eine Überwachung erforderlich sein.</s>
<s><class="Medikation">Valsartan</class>/<class="Medikation">HCT</class> <class="Dosis
    ↪ ">160</class>/<class="Dosis">12,5 mg</class> 1-0-0</s>
<s><class="Medikation">Pantoprazol</class> <class="Dosis">40 mg</class> p.o.</s>
<s>Die feingewebliche histopathologische Untersuchung ergab den Befund einer <class="
    ↪ Diagnose">Metastase</class> des bekannten malignen <class="Diagnose">Melanoms</class
    ↪ >.</s>
<s><class="Diagnose">Diabetes Typ 2</class>-Patienten müssen regelmäßig <class="Medikation">
    ↪ Insulin</class> (mindestens mit <class="Dosis">12ml</class> dosiert) spritzen.</s>
<s>Ich nehme <class="Medikation">Antibiotika</class> seit Tagen. Seitdem ist die <class="
    ↪ Diagnose">Mandelentzündung</class> deutlich besser geworden.</s>
<s>Entlassung: <class="Dosis">40mg</class> <class="Medikation">Lidocain</class> wegen <class
    ↪ ="Diagnose">Kopfschmerzen</class></s>
<s>Zusammenfassende D: Zervix-PE bei 11 und 2 Uhr mit ausgeprägter <class="Diagnose">
    ↪ chronisch-florider Zervizitis<class="Diagnose">.</s>
<s>Die Verschreibung von <class="Medikation">Hämatokrin</class> <class="Dosis">43mg</class>
    ↪ war unnötig.</s>
<s>Der Patient klagt über <class="Diagnose">Karditiden</class> und nimmt täglich <class="
    ↪ Medikation">Nifedipin</class> ein.</s>
<s>D: PE-Material der Portio bei 1 Uhr mit Nachweis einer schwergradigen <class="Diagnose">
    ↪ squamösen intraepithelialen Läsion</class> (<class="Diagnose">HSIL</class>; hier
    ↪ noch <class="Diagnose">CIN II</class>).</s>
<s>
```

Fig. 3. Input prompt: The sentences are encoded according to the markup scheme. The trailing <s> indicates the beginning of a new sentence to the model. Only the shown sentences are used for the sentence sampling as this input prompt is directly used as input for the GPT model.

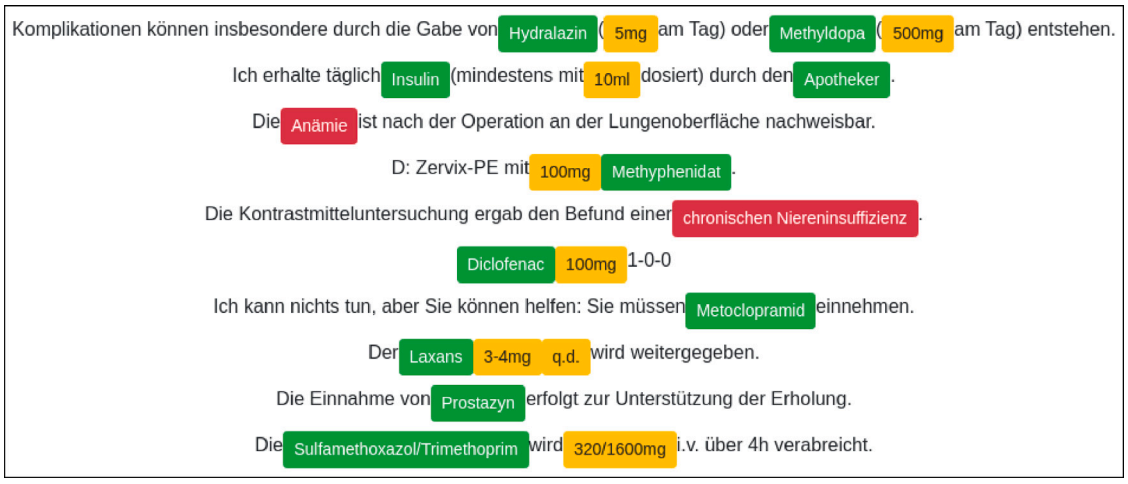


Fig. 4. Sampled sentences: Ten randomly sampled sentences from the final dataset. The annotations are provided (green: Medikation/drug, orange: Dosis/strength, red: Diagnose/diagnosis). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

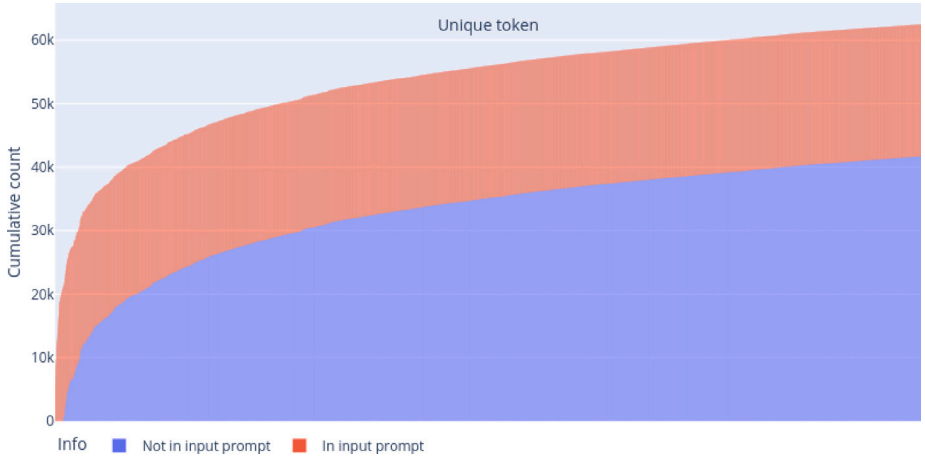


Fig. 5. Cumulative distribution of (filtered) tokens: 0.9% of all unique tokens are part of the initial prompt but account for 33.3% of all tokens by frequency. The tokens are counted case-insensitively and have been filtered for stop words and punctuation.

occurrence (*Apotheker*). Entities from the strength label class (Dosis) occur with different units (*ml* and *mg*) and are correctly annotated except for one false-positive item (*q.d.*) as well. For diagnosis, only two entities are present in the set of ten sentences but their annotations are correct. The detection of entity boundaries is not correctly identified in two cases when the entities should be split (*Sulfamethoxazol/Trimethoprim* and *320/1600 mg*). However, generic phrases from the input prompt are repeated in some instances (*ergab den Befund einer*, *die Einnahme von*), but specific terms from medications or diagnoses are newly generated as they are clearly not drawn from the input prompt (except for the common *Insulin* term).

For the investigation of the (case-insensitive) token distribution over the dataset, we temporarily filtered the dataset to exclude all stop word and punctuation tokens, resulting in 62520 total tokens, and 8794 unique tokens after a token deduplication. We found that 20842 total tokens and 76 unique tokens can be also found in the input prompt while 41678 total tokens and 8718 unique tokens are not present in the input prompt. This implies that 0.9% of the unique tokens, yet 33,3% of the total tokens are also found in the prompt. The cumulative histogram of the dataset tokens is shown in Fig. 5.

Given that 99% of the unique tokens are not found in the input prompt, it suggests that the dataset synthesis process is able to effectively augment the vocabulary from the input prompt.

However, the share of 33% total tokens is contributed by tokens that are also present in the input prompt and could potentially stem from trivial token repetition from the input prompt, but also be a well-justified use of domain-specific or use-case-specific vocabulary. In the case of trivial token repetition, we expect the topmost tokens to almost exclusively consist of tokens already present in the prompt and to strictly follow the token distribution of the input prompt regardless of the actual domain-dependent token distribution. To investigate this, we extracted the most frequent tokens from the dataset. As a reference, we also extracted the topmost tokens from the input prompt. The most frequent tokens are given in Fig. 6.

From the list of top tokens, the first nine tokens can be found in the input prompt and occur more than 500 times, yet they are rather generic terms (e.g. unit *mg* or *patient*) and are expected to be mentioned frequently. Still, we could identify two tokens (*zervix-pe* and *hsil*), which are expected to be less frequent in generic medical texts, that occur frequently in the dataset (18th and 30th most frequent token) and are also mentioned in one of the twelve sentences from the input prompt. Yet both of these instances occur less frequently than 500 times. In general, the token distribution from the synthetic dataset does not strictly match the distribution from the input prompt as we can observe that new, frequent tokens are introduced instead (e.g. common drugs *aspirin*, *propranolol* or new numeric values *10*, *25*, *50*).

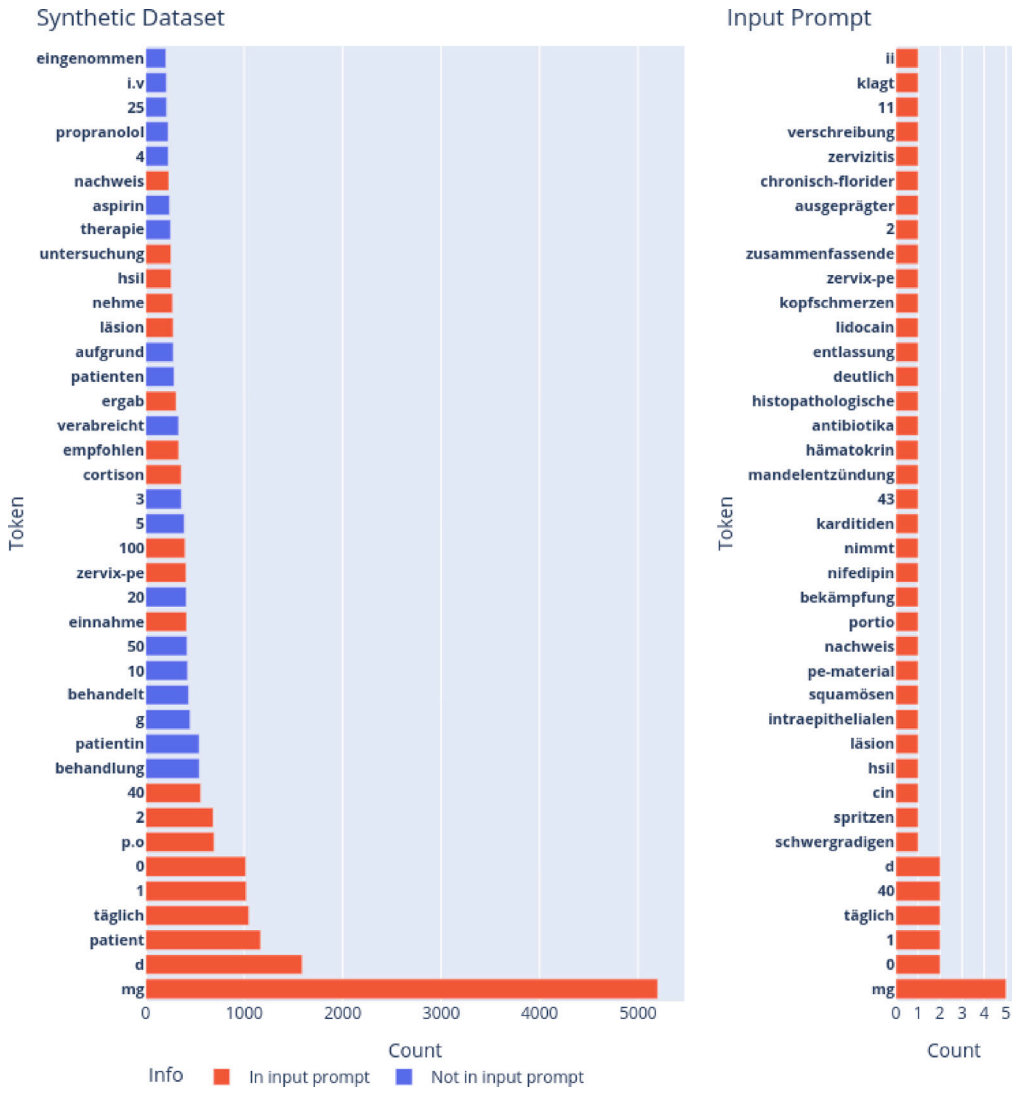


Fig. 6. Top tokens (filtered): The most frequent tokens from the final dataset (left) and input prompt (right). Several top tokens from the final dataset are not part of the initial prompt. The shown tokens are counted case-insensitively and have been filtered for stop words and punctuation.

4.2. NER training

As a follow-up step, we train three NER models on the synthesized dataset with the pre-trained **gbert-large** [44], **GottBERT-base** [45] and **German-MedBERT**[7] models retrieved from the HuggingFace platform as contextualized feature encoders. We split the dataset randomly into (80%, 10%, 10%) sets for training, validation and test. The Adam optimizer with an initial learning rate $5e^{-5}$ and a batch size of 128 are used, as we stick close to the default hyperparameters from SpaCy for training. We select the final model based on the highest F1-score on the validation set. The training iterations took 55 m (gbert), 25 m (GottBERT), 48 m (German-MedBERT).

We evaluate the performance of the respective models on precision, recall and F1-score on the testset. The evaluation is computed in strict mode as a character-wise classification task, meaning that exact overlaps and label classes are considered. The results are shown in Table 3. The results indicate strong performance of the models on all label classes, with gbert and GottBert as the models with the best average F1-scores. As a significant caveat, while the dataset is split into training, validation and test set and no samples are shared across these sets, the synthesized dataset contains structurally similar sentences and it allows the models to potentially overfit implicitly by learning syntax and structure of such homogeneous sentences instead of overfit to certain words directly. The homogeneity could be reduced by various techniques including increasing the temperature τ at the expense of increasing the probability of generating invalid sentences.

We further evaluate the models on a small gold-standard German dataset proposed in [22] as an out-of-distribution (OoD) dataset. Since the dataset contains label annotations largely compatible with the n2c2 2018 ADE dataset [4], we cannot directly compare all label classes, yet in the interest of an OoD performance evaluation, we assume that the label class *Drug* shares significant semantic overlap with the label class *Medikation*. The results are provided in Table 4. Beyond the expected drop in terms of the *Medikation* scores across all models, the gbert-based and GottBERT-based models are identified as the models with the best F1-scores, with GottBERT surpassing gbert by 2.6% in F1-score (test set reference: −0, 6%).

To put our results in a broader perspective, we evaluate our method on three related external datasets and compare the obtained scores to a pre-existing baseline model on shared drug-like entity classes. The datasets consist of the Medline dataset (from Mantra GSC [6]), the GGPONC [19] dataset and the BRONCO [18] ontology dataset. Since

[7] German MedBERT on Huggingface (accessed 22.08.2022): https://huggingface.co/smanjil/German-MedBERT.

Table 3
Results on the test set: The total results are based on the labels' frequency-weighted average. The label annotations are evaluated character-wise by **Pr**ecision, **Re**call and **F1**-scores Abbreviations: named entity recognition (NER).

Scores on test set		NER Tags			
Model		Medikation	Diagnose	Dosis	Total
gbert-large	Pr	0.870	0.870	0.883	0.918
	Re	**0.936**	**0.895**	**0.921**	**0.919**
	F1	**0.949**	**0.882**	**0.901**	**0.918**
GottBERT-base	Pr	0.979	0.896	**0.887**	**0.936**
	Re	0.910	0.844	0.907	0.886
	F1	0.943	0.870	0.897	0.910
German-MedBERT	Pr	**0.980**	**0.910**	0.829	0.932
	Re	0.905	0.730	0.890	0.842
	F1	0.941	0.810	0.858	0.883

Table 4
Results on the out-of-distribution dataset: As caveat, the label definitions of *Medikation* (ours) and *Drug*(from the 2018 n2c2 ADE dataset [4]) is inaccurately assumed to be equivalent for comparison. The label annotations are evaluated character-wise by **Pr**ecision, **Re**call and **F1**-scores. Abbreviations: named entity recognition (NER).

Scores on OoD set		NER Tag
Model		Drug = Medikation
gbert-large	Pr	0.707
	Re	**0.979**
	F1	0.821
GottBERT-base	Pr	**0.800**
	Re	0.899
	F1	**0.847**
German-MedBERT	Pr	0.727
	Re	0.818
	F1	0.770

the entity classes vary across the external datasets, we only consider the entity classes which are most similar to the drug/Medikation entity class as already discussed in the OoD evaluation stage. The performance scores are shown in Table 5. The evaluation remains similar to the OoD evaluation strategy but also includes a token-wise label evaluation. It should be noted that the tokens are obtained by the SpaCy tokenizer which mainly implements a whitespace-based tokenization strategy in contrast to other byte-pair-based tokenization strategies. In the case of the GGPONC dataset, some annotation spans do not exactly align with the spans of the tokens and are therefore omitted.

Our results indicate a notable drop in performance scores for the *Medikation/drug* entity class across all models including the GGPONC reference model on all three datasets. Since the GGPONC model originates from the work on the GGPONC dataset, it performs best on the dataset and is closely followed by the German-MedBERT-based model. However, to our surprise, the GottBERT-based and gbert-based models are able to clearly beat the reference model on the independent datasets Medline and BRONCO respectively, surpassing the 72% character-wise F1-score. As expected, the dataset-dependent variations in performance scores highlight the general difficulty for all NLP models to adapt to dataset shifts and biases as well as ill-defined entity class definitions.

5. Discussion

We demonstrate the effectiveness of our method for utilizing pre-trained large language models for dataset synthesis by training a neural NER model on this synthesized dataset, yet the limited availability of annotated German medical NLP datasets with ill-defined or even dissimilar label classes remains a major obstacle when it comes to a more exhaustive, yet reliable evaluation of the trained NER model for all label classes. Given the evaluation scores on the *drug/Medikation* labels it must be considered that our method achieves these results based on twelve initial sentences. Aside from the evaluation, we did

Table 5
Evaluation of all models on external datasets. Only drug-related label classes are considered. The GGPONC reference model is evaluated for comparison. **Pr**ecision, **Re**call and **F1**-scores are evaluated. Annotations from the GGPONC dataset do not align with the tokens from the SpaCy tokenizer and are therefore omitted.

Scores on Related Datasets		Performance Scores	
Model/Dataset		Drug (char-wise)	Drug (token-wise)
Medline Dataset [6]		CHEM = Medikation	
GPTNERMED (gbert-large)	Pr	0.749	0.760
	Re	**0.711**	**0.745**
	F1	**0.729**	**0.752**
GPTNERMED (GottBERT-base)	Pr	**0.919**	**0.900**
	Re	0.468	0.529
	F1	0.620	0.667
GPTNERMED (German-MedBERT)	Pr	0.725	0.788
	Re	0.471	0.510
	F1	0.571	0.619
GGPONC [19]	Pr	0.822	0.771
	Re	0.488	0.529
	F1	0.612	0.628
GGPONC Dataset [46]		Chemicals_Drugs = Medikation	
GPTNERMED (gbert-large)	Pr	0.460	n/a
	Re	0.789	n/a
	F1	0.581	n/a
GPTNERMED (GottBERT-base)	Pr	0.301	n/a
	Re	**0.854**	n/a
	F1	0.445	n/a
GPTNERMED (German-MedBERT)	Pr	0.569	n/a
	Re	0.681	n/a
	F1	0.620	n/a
GGPONC [19]	Pr	**0.636**	n/a
	Re	0.737	n/a
	F1	**0.683**	n/a
BRONCO Dataset [18]		MEDICATION = Medikation	
GPTNERMED (gbert-large)	Pr	0.462	0.465
	Re	**0.911**	**0.864**
	F1	0.613	0.605
GPTNERMED (GottBERT-base)	Pr	**0.655**	**0.678**
	Re	0.809	0.750
	F1	**0.724**	**0.712**
GPTNERMED (German-MedBERT)	Pr	0.617	0.575
	Re	0.705	0.662
	F1	0.658	0.615
GGPONC [19]	Pr	0.573	0.346
	Re	0.449	0.430
	F1	0.504	0.384

not further perform hyperparameter search for dataset synthesis on parameters like temperature τ or top-k/top-p sampling or beam search due to the high computational costs of running the NeoX model as well as due to limited access to GPU resources.

Even though the initial need for computational resources is a major downside of our method, we believe that this factor becomes negligible with respect to the fact that the method can operate without input from costly human annotators. For very domain-specific contexts, such as German medical texts, this not only provides an opportunity to work on NLP approaches independent of external monopoly-like data sources and medical institutions that also constitute a severe asymmetry in academic competition. Yet it also allows the further use of the dataset without additional efforts in pseudonymization and legal ramifications that are usually unavoidable when working with datasets originating from real patient data. Therefore, we are able to publicly provide the synthesized corpus and the trained models for third-party use without further access restrictions.

While our NER model exhibits strong performance in general and proves the dataset to comprise useful and valid data for text and corresponding annotation, the dataset remains synthetic in nature and thus

cannot be considered as gold standard-level dataset. The question to which degree the corpus carries additional domain knowledge remains open for future work.

This also applies to the need for further validation of the synthetic corpus in general. The investigation of the token distribution indicates that the text augmentation contributes additional tokens and sentences that are not composed of trivial token repetitions from the initial prompt. In similar ways, the NER results from the trained models indicate semantically meaningful annotation information of the corpus because quantitatively substantial annotation flaws impede the models' scores on external datasets. However, a rigorous and manually conducted assessment of the corpus, including a manual annotation task, would be required to further minimize uncertainties about the qualitative properties of the corpus.

In practical terms, using large language models for data augmentation can be a powerful tool to obtain synthetic, annotated datasets from low-resource domains and languages that enables the development of new NLP models for clinical applications and allows open model and dataset sharing between community members, including instances where no real-world data for training exist. Shortcomings of published models and datasets on real-world clinical data should be fed back to further improve these datasets and models towards more general applicability. However, actual employments of such models beyond research-related use cases remain subject to regulatory processes in the clinical domain.

6. Conclusion

In this work, we leveraged the few-shot ability of the pre-trained language model GPT NeoX to generate an annotated dataset for German medical texts without the need for manual annotations by introducing few annotated text samples to the language model in a simple markup format. We further used the dataset to train NER models by fine-tuning three pre-trained BERT encoder models combined with a classification head for NER. Our evaluation on testset as well as OoD set indicates a robust performance of the NER models even for smaller shifts in the dataset. The evaluation on related datasets demonstrates the ability of our approach to outperform a pre-existing reference model on external datasets, yet our method remains highly data-efficient. We discussed the disadvantages and advantages of our method as well as its potential implications for the German medical NLP research community and beyond.

The corpus and the trained models are publicly available on GitHub at: https://github.com/frankkramer-lab/GPTNERMED

CRediT authorship contribution statement

Johann Frei: Conceptualization, Methodology, Software, Investigation, Validation, Formal analysis, Writing – original draft, Resources, Data curation. **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.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[2] T.J. Pollard, A.E. Johnson, The MIMIC-III clinical database, 2016, http://dx.doi.org/10.13026/C2XW26.

[3] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L.A. Celi, R. Mark, MIMIC-IV, 2021, http://dx.doi.org/10.13026/s6n6-xd98, https://physionet.org/content/mimiciv/1.0/.

[4] S. Henry, K. Buchan, M. Filannino, A. Stubbs, O. Uzuner, 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records, J. Am. Med. Inform. Assoc. 27 (1) (2020) 3–12, http://dx.doi.org/10.1093/jamia/ocz166, https://europepmc.org/articles/PMC7489085.

[5] A.J. Yepes, A. Névéol, M. Neves, K. Verspoor, O. Bojar, A. Boyer, C. Grozea, B. Haddow, M. Kittner, Y. Lichtblau, Findings of the wmt 2017 biomedical translation shared task, in: Proceedings of the Second Conference on Machine Translation, 2017, pp. 234–247.

[6] J.A. Kors, S. Clematide, S.A. Akhondi, E.M. van Mulligen, D. Rebholz-Schuhmann, A multilingual gold-standard corpus for biomedical concept recognition: The mantra GSC, J. Am. Med. Inform. Assoc. 22 (5) (2015) 948–956, http://dx.doi.org/10.1093/jamia/ocv037.

[7] U. Hahn, F. Matthies, C. Lohr, M. Löffler, 3000Pa-towards a national reference corpus of german clinical language, in: MIE, 2018, pp. 26–30.

[8] R. Roller, H. Uszkoreit, F. Xu, L. Seiffe, M. Mikhailov, O. Staeck, K. Budde, F. Halleck, D. 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.

[9] J. Wermter, U. Hahn, An annotated german-language medical text corpus as language resource, in: LREC, Citeseer, 2004.

[10] M. Kreuzthaler, S. 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.

[11] M. Toepfer, H. Corovic, G. Fette, P. Klügl, S. Störk, F. Puppe, Fine-grained information extraction from german transthoracic echocardiography reports, BMC Med. Inform. Decis. Mak. 15 (1) (2015) 1–16.

[12] C. Bretschneider, S. Zillner, M. 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.

[13] G. Fette, M. Ertl, A. Wörner, P. Kluegl, S. Störk, F. Puppe, Information extraction from unstructured electronic health records and integration into a data warehouse, in: INFORMATIK, Gesellschaft für Informatik eV, 2012.

[14] M. König, A. Sander, I. Demuth, D. Diekmann, E. 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) e0224916.

[15] V. Cotik, R. Roller, F. Xu, H. Uszkoreit, K. Budde, D. 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.

[16] J.A. Miñarro-Giménez, R. Cornet, M.-C. Jaulent, H. Dewenter, S. Thun, K.R. Gøeg, D. Karlsson, S. Schulz, Quantitative analysis of manual annotation of clinical text samples, Int. J. Med. Inf. 123 (2019) 37–48.

[17] J. Krebs, H. Corovic, G. Dietrich, M. Ertl, G. Fette, M. Kaspar, M. Krug, S. Störk, F. Puppe, Semi-automatic terminology generation for information extraction from german chest x-ray reports, GMDS 243 (2017) 80–84.

[18] M. Kittner, M. Lamping, J. Götze D.T. Rieke, B. Bajwa, I. Jelas, G. Rüter, H. Hautow, M. Sänger, M. Habibi, M. Zettwitz, T. d Bortoli, L. Ostermann, J. Åeva, J. Starlinger, O. Kohlbacher, N.P. Malek, U. Keilholz, U. Leser, Annotation and initial evaluation of a large annotated german oncological corpus, JAMIA Open 4 (2) (2021) ooab025, http://dx.doi.org/10.1093/jamiaopen/ooab025.

[19] F. Borchert, C. Lohr, L. Modersohn, J. Witt, T. Langer, M. Follmann, M. Gietzelt, B. Arnrich, U. Hahn, M.-P. Schapranow, GGPONC 2.0 - the german clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER taggers, European Language Resources Association, 2022, pp. 3650–3660, https://aclanthology.org/2022.lrec-1.389.

[20] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, Extracting training data from large language models, in: 30th USENIX Security Symposium, USENIX Security 21, 2021, pp. 2633–2650.

[21] J. Frei, F. Kramer, GERNERMED: An open german medical NER model, Softw. Impact. 11 (2022) 100212, http://dx.doi.org/10.1016/j.simpa.2021.100212, https://www.sciencedirect.com/science/article/pii/S2665963821000944.

[22] J. Frei, L. Frei-Stuber, F. Kramer, GERNERMED++: Transfer learning in german medical NLP, 2022, http://dx.doi.org/10.48550/arXiv.2206.14504.

[23] R. Roller, C. Alt, L. Seiffe, H. 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, December 3-5, Semantic Web Applications and Tools for Healthcare and Life Sciences, Antwerp, Belgium, 2018.

[24] R. Roller, L. Seiffe, A. Ayach, S. Möller, O. Marten, M. Mikhailov, C. Alt, D. Schmidt, F. Halleck, M. Naik, W. Duettmann, K. Budde, A medical information extraction workbench to process german clinical text, 2022, http://dx.doi.org/10.48550/arXiv.2207.03885.

[25] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5) (2010) 507–513, http://dx.doi.org/10.1136/jamia.2009.001560, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995668.

[26] A.R. Aronson, F.-M. Lang, An overview of MetaMap: Historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (3) (2010) 229–236.

[27] P.Y. Simard, Y.A. LeCun, J.S. Denker, B. Victorri, Transformation invariance in pattern recognition—tangent distance and tangent propagation, Neural networks: Tricks of the trade 23 (1998) 239–274.

[28] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, Long Papers, 2016, pp. 86–96.

[29] M. Artetxe, G. Labaka, E. Agirre, Translation artifacts in cross-lingual transfer learning, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 7674–7684.

[30] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, N. Zwerdling, Do not have enough data? Deep learning to the rescue!, Proc. AAAI Conf. Artif. Intell. 34 (5) (2020) 7383–7390, http://dx.doi.org/10.1609/aaai.v34i05.6233.

[31] G. Raille, S. Djambazovska, C. Musat, Fast cross-domain data augmentation through neural sentence editing, 2020, 10254, http://dx.doi.org/10.48550/arXiv.2003.10254.

[32] T. Schick, H. Schütze, Few-shot text generation with natural language instructions, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021, pp. 390–402, http://dx.doi.org/10.18653/v1/2021.emnlp-main.32.

[33] T. Schick, H. Schütze, Generating datasets with pretrained language models, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021, pp. 6943–6951, http://dx.doi.org/10.18653/v1/2021.emnlp-main.555.

[34] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, in: OpenAI Blog, 2019, p. 24.

[35] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901, https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[36] B. Wang, A. Komatsuzaki, GPT-J-6b: A 6 Billion Parameter Autoregressive Language Model, 2021.

[37] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U.S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, S. Weinbach, GPT-NeoX-20b: An open-source autoregressive language model, in: Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, Association for Computational Linguistics, 2022, pp. 95–136, http://dx.doi.org/10.18653/v1/2022.bigscience-1.9.

[38] R. Puri, B. Catanzaro, Zero-shot text classification with generative language models, 2019, arXiv:1912.10165 [cs].

[39] Y. Meng, J. Huang, Y. Zhang, J. Han, Generating training data with language models: Towards zero-shot language understanding, 2022, http://dx.doi.org/10.48550/arXiv.2202.04538.

[40] C.A. Libbi, J. Trienes, D. Trieschnigg, C. Seifert, Generating synthetic training data for supervised de-identification of electronic health records, Future Internet 13 (5) (2021) 136, http://dx.doi.org/10.3390/fi13050136.

[41] A. Amin-Nejad, J. Ive, S. Velupillai, Exploring transformer text generation for medical dataset augmentation, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 4699–4708.

[42] S. Schuster, S. Gupta, R. Shah, M. Lewis, Cross-lingual transfer learning for multilingual task oriented dialog, 2019, http://arxiv.org/abs/1810.13327.

[43] T. Ganslandt, M. Boeker, M. Löbe, F. Prasser, J. Schepers, S.C. Semler, S. Thun, U. Sax, Der kerndatensatz der medizininformatik-initiative: Ein schritt zur sekundärnutzung von versorgungsdaten auf nationaler ebene, in: Forum Der Medizin-Dokumentation Und Medizin-Informatik, vol. 20, (1) 2018, p. 17.

[44] B. Chan, S. Schweter, T. Möller, German's next language model, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, 2020, pp. 6788–6796, http://dx.doi.org/10.18653/v1/2020.coling-main.598.

[45] R. Scheible, F. Thomczyk, P. Tippmann, V. Jaravine, M. Boeker, GottBERT: A pure german language model, 2020, ArXiv.

[46] F. Borchert, C. Lohr, L. Modersohn, T. Langer, M. Follmann, J.P. Sachs, U. Hahn, M.-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.