Applicability of Models Trained on Generated Clinical German Datasets on Out-domain Data

Oğuz Şerbetçi[1], Ulf Leser[1]

[1]*Humboldt Universität zu Berlin, Berlin, Germany*

Abstract

Strong privacy constraints heavily constrain the public availability of health record data in German which hinders the development of advanced NLP methods for clinical texts. As a remedy, work by Frei and Kramer [1] has leveraged the recent breakthroughs in generative large language models to generate a synthetic dataset for training Named Entity Recognition (NER) models in the clinical domain. Because the basis is synthetic, both the corpus and the NER model are publicly available. However, as clinical text is highly idiosyncratic, it is not clear how well this approach performs on real data. We evaluate the model on two real-world German clinical datasets from cardiology and oncology departments. Our analysis shows that learning on generated data for NER models do not transfer well to real-world data.

Keywords

german medical natural language processing, large language models„ medical informatics

1. Introduction

Some of the important information in electronic health records is found only in unstructured text and needs to be extracted using natural language processing (NLP). An important task to this end is named entity recognition (NER) [2, 3]. However, this task is difficult in the clinical domain because the goal of medical text is to concisely capture large amounts of factual information, resulting in abbreviations, domain-specific terminology and awkward grammar. Differences between processes, hospitals and practitioners make it difficult to transfer models between different datasets [4]. In addition, privacy concerns limit the sharing of such data and make annotation expensive due to the need for anonymisation [5]. This problem is particularly exacerbated for German-language applications, not only due to strong EU privacy regulations such as GDPR, but also due to stronger national privacy regulations [6].

Recently, the first annotated German clinical datasets BRONCO150 [7] and CARDIO:DE [8] for NER have been published, which are distributable under a data use agreement (DUA). Both use anonymisation. Kittner et al. [7] goes a step further and shuffles and filters sentences such that a clinical document cannot be reconstructed and that parts with frequent personal information are not distributed. Annotation of both datasets was an iterative and time consuming process to improve inter-annotator agreement due to the aforementioned challenges of medical text. In contrast, Frei and Kramer [1] attempts to circumvent the privacy issue and the costly annotation process by using recent advances in large language models. They generate both the clinical text

LWDA'23: Lernen, Wissen, Daten, Analysen. October 09–11, 2023, Marburg, Germany

oguz.serbetci@informatik.hu-berlin.de (O. Şerbetçi); ulf.leser@informatik.hu-berlin.de (U. Leser)

Table 1
Datasets: The GPTNERMED corpus has been used to train a clinical German NER model [1]. We evaluate the GPTNERMED model using BRONCO150 [7] and CARDIO:DE [8].

Name	Description	#Sentences, #Tokens, #Annotations	Entity types
GPTNERMED	Generated clinical sentences using the large language model	9 845, 245 107, 23 411	Diagnosis, Medication, Dosage
BRONCO150	Discharge summaries from two oncology departments	8 976, 70 572, 8 760	Diagnosis, Medication, Treatment
CARDIO:DE	Cardiovascular doctor's letters from clinical routine	96 203, 805 617, 19 345	Active Ingredient, Drug, Dosage, Route, Frequency, Duration, Strength, Reason, Form

and its NER annotations for diagnosis, medication and dosage using prompts with manually modified examples of clinical notes sentences, allowing both the data and the corresponding NER model to be made publicly available without DUA. However, it is not yet clear how such data and models trained on it can deal with the idiosyncrasies and heterogeneity of the real clinical setting, as clinical text varies widely between practitioners, departments and institutions. There are also differences in how the text has been pre-processed between different applications, how anonymisation is performed and which parts of clinical documents are included.

2. Experimental setup

We want to evaluate how the approach proposed by Frei and Kramer [1], models trained on a language model generated annotated clinical text, performs on two real-world clinical NER datasets BRONCO150 and CARDIO:DE, which are described in Table 1.

Frei and Kramer [1] first generate 23,411 clinical sentences that are annotated with diagnosis, medication, and dosage entities using the large language model GPT-NeoX and example clinical sentences that have been hand written based on real clinical text. They then fine-tune different pretrained German medical language models based on the BERT architecture [9] to obtain the GPTNERMED model that can extract diagnosis, medication and dosage entities.

The model is published using the spacy library[1], and we use it to evaluate and train all NER models in our experiments. We always use the best performing public model for GPTNERMED that is based on the German-MedBERT (G-MedBERT) medical language model [10]. For the evaluation on BRONCO150, we use the first predefined 20% split as the test and all the other splits as the training data. For the evaluation on CARDIO:DE, which does not have predefined splits, we use the first 80% of the documents as training and the rest as test data. Entity predictions are evaluated with exact span matches—creating a challenge due to variances in annotation schemes across datasets and even disagreements among human annotators, as seen

[1] https://spacy.io

Table 2
Named Entity Recognition Benchmarks on BRONCO150 and CARDIO:DE datasets using GPTNERMED model, baseline setting with fine-tuning pretrained clinical BERT model G-MedBERT on both datasets, and the in-corpus setting with scores reported by dataset authors. For BRONCO150, we only consider the intersection of entity labels diagnosis and medication. For CARDIO:DE, we map labels strength and frequency to dosage, and active ingredient and drug to medication and report the macro average F1 score for in-corpus setting as the authors do not report entity based precision and recall.

Dataset	Model	Label	Precision	Recall	F1
BRONCO150	GPTNERMED	Diagnosis	0.26	0.44	0.33
		Medication	0.33	0.89	0.50
	Fine-tuned G-MedBERT	Diagnosis	**0.81**	**0.79**	**0.80**
		Medication	0.92	**0.93**	**0.93**
	In-corpus [7]	Diagnosis	**0.81**	0.74	0.77
		Medication	**0.96**	0.87	0.91
CARDIO:DE	GPTNERMED	Medication	0.29	**0.87**	0.43
		Dosage	0.02	0.11	0.03
	Fine-tuned G-MedBERT	Medication	**0.91**	0.86	**0.88**
		Dosage	**0.94**	**0.97**	**0.96**
	In-corpus [8]	Medication	-	-	0.84
		Dosage	-	-	0.94

in lower inter-annotator agreement compared to token based evaluation for both CARDIO:DE and BRONCO150 [8, 7]. Despite the difficulty, correct span prediction is crucial for downstream tasks like entity normalization and linking.

Our first experiment evaluates the GPTNERMED model as it is, i.e. without any finetuning on BRONCO150 and CARDIO:DE. As a second experiment, we fine-tune the best performing GPTNERMED base model G-MedBERT with training configuration published by Frei and Kramer [1] on both datasets, which means resulting models have not seen the GPTNERMED data. We also provide the results provided by the respective dataset papers BRONCO150 and CARDIO:DE for comparison [7, 8]. In a third experiment, we try different fine-tuning strategies to identify how to best utilize the work of Frei and Kramer [1]. We try following setups: (1) finetuning G-MedBERT, (2) fine-tune GPTNERMED model on BRONCO150, and (3) fine-tune G-MedBERT on both BRONCO and GPTNERMED data.

3. Results

Results for the GPTNERMED are in Table 2. We observe unsatisfactory performance of off-the-shelf GPTNERMED model on real-world datasets BRONCO150 and CARDIO:DE. Results from experiment probing the best fine-tuning strategy using the GPTNERMED data are presented in Table 3. We see that none of the strategies bring any considerable improvement.

When we inspect the GPTNERMED predictions, we see that it predicts 74% of tokens with more than one capital letter consecutively as medication or diagnosis, where as only 1% of these tokens are annotated as medication or diagnosis. False positives include BRONCO150's

Table 3
Experiments with different fine-tuning strategies. First, we fine-tune the German-MedBERT (G-MedBERT), the base model of GPTNERMED, with exact configuration on BRONCO150. This is the same setting as the baseline in Table 2. Second, we fine-tune the GPTNERMED model on BRONCO150. Third, we fine-tune the G-MedBERT on BRONCO150 and GPTNERMED. In all cases we evaluate with BRONCO150.

Setting	Label	Precision	Recall	F1
Finetune G-MedBERT on BRONCO150	Diagnosis	0.81	0.79	**0.80**
	Medication	0.92	0.93	0.93
Finetune GPTNERMED model on BRONCO150	Diagnosis	**0.82**	0.77	0.79
	Medication	**0.93**	0.93	0.93
Finetune G-MedBERT on BRONCO150 & GPTNERMED	Diagnosis	0.80	**0.80**	**0.80**
	Medication	**0.93**	**0.95**	**0.94**

anonymisation replacement tokens such as ”PATIENTen” and ”KRANKENHAUS” being predicted as medication and B-SALUTE, B-PERSON being predicted as diagnosis. Some common treatments, e.g. ”CT”, and diagnosis, e.g. ”HCC-suspekten Laesionen” that have capitalized abbreviations are also predicted as medication.

Similar problems also exist with dosage predictions in CARDIO:DE as it includes full doctor’s reports, whereas GPTNERMED dataset only includes sentences with a mention of medicine and potentially its dosage. This results in a lot of non-medical numerical information in CARDIO:DE, e.g. age and birth date of the patient, are confused with dosage label by GPTNERMED. Of all the tokens with at least one digit, GPTNERMED predicts dosage for 61%, whereas only 5% are labeled as Dosage.

We suspect both covariate and label shift occur with both dataset, which can be potentially addressed by different methods. For a good overview we refer the reader to the review by Hupkes et al. [11].

4. Conclusion and Future Work

Generating annotated medical text using large language models can circumvent the privacy issues during dataset curation in clinical domain and address the lack of available training data. However, the evaluations show that current solutions cannot deal with idiosyncrasies of medical text. It is, however, important that published models consider down-stream use-cases and perform according evaluations. For a clinical application of a model it is required that it is able to deal with anonymisation and sentences without any medical named entities. Furthermore generated datasets should consider real world scenarios, where there are typos and wrong punctuation.

Acknowledgments

Funded by Gemeinsame Bundesausschuss (G-BA, 01VSF22041).

References

[1] J. Frei, F. Kramer, Annotated Dataset Creation through General Purpose Language Models for non-English Medical NLP, 2022. arXiv:2208.14493.

[2] S. R. Kundeti, J. Vijayananda, S. Mujjiga, M. Kalyan, Clinical named entity recognition: Challenges and opportunities, in: 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 1937–1945. doi:10.1109/BigData.2016.7840814.

[3] J. Starlinger, S. Pallarz, J. Ševa, D. Rieke, C. Sers, U. Keilholz, U. Leser, Variant information systems for precision oncology, BMC Medical Informatics and Decision Making 18 (2018) 107. doi:10.1186/s12911-018-0665-z.

[4] A. Kormilitzin, N. Vaci, Q. Liu, A. Nevado-Holgado, Med7: A transferable clinical natural language processing model for electronic health records, Artificial Intelligence in Medicine 118 (2021) 102086. doi:10.1016/j.artmed.2021.102086.

[5] I. Spasic, G. Nenadic, Clinical Text Data in Machine Learning: Systematic Review, JMIR Medical Informatics 8 (2020) e17984. doi:10.2196/17984.

[6] T. Kolditz, C. Lohr, J. Hellrich, L. Modersohn, B. Betz, M. Kiehntopf, U. Hahn, Annotating german clinical documents for de-identification., in: MedInfo, 2019, pp. 203–207.

[7] M. Kittner, M. Lamping, D. T. Rieke, J. Götze, B. Bajwa, I. Jelas, G. Rüter, H. Hautow, M. Sänger, M. Habibi, M. Zettwitz, T. de Bortoli, L. Ostermann, J. Ševa, 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 (2021) ooab025. doi:10.1093/jamiaopen/ooab025.

[8] P. Richter-Pechanski, P. Wiesenbach, D. M. Schwab, C. Kiriakou, M. He, M. M. Allers, A. S. Tiefenbacher, N. Kunz, A. Martynova, N. Spiller, J. Mierisch, F. Borchert, C. Schwind, N. Frey, C. Dieterich, N. A. Geis, A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters, Scientific Data 10 (2023) 207. doi:10.1038/s41597-023-02128-9.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[10] M. Shrestha, Development of a Language Model for Medical Domain, Ph.D. thesis, Hochschule Rhein-Waal, 2021.

[11] D. Hupkes, M. Giulianelli, V. Dankers, M. Artetxe, Y. Elazar, T. Pimentel, C. Christodoulopoulos, K. Lasri, N. Saphra, A. Sinclair, D. Ulmer, F. Schottmann, K. Batsuren, K. Sun, K. Sinha, L. Khalatbari, M. Ryskina, R. Frieske, R. Cotterell, Z. Jin, State-of-the-art generalisation research in NLP: A taxonomy and review, 2023. arXiv:2210.03050.