Embedding-Based Acronym Disambiguation Supported by Large Language Models in German Clinical Narratives

Amila Kugic[1,*], Stefan Schulz[1] and Markus Kreuzthaler[1]

[1] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*

Abstract

An embedding-based approach for acronym disambiguation in German was developed using a combination of large language model (LLM) prompts and the MedBERT.de model without human-annotated training data. Both zero-shot and few-shot prompting techniques were employed with the Generative Pretrained Transformer model, GPT-4o, to generate training examples for the creation of embedding spaces, which were then indexed using Faiss (Facebook AI Similarity Search) for a nearest-neighbor search. Each embeddings-based search for three distinct acronyms in context identified the closest long-form resolutions based on distances between embeddings. Acronym disambiguation achieved a maximum accuracy of 0.69 [0.64-0.73], which was comparable to the baseline accuracy of 0.65 [0.61-0.69] obtained using LLMs. However, synthetic training examples characterized by zero-shot-prompting to build the embedding spaces resulted in a lower accuracy of 0.46 [0.41-0.50], in comparison to few-shot prompting of synthetic clinical narratives. The results underscore the challenge of accurately disambiguating acronyms in real-world clinical narratives without human-labeled data and highlight the contextual complexity involved.

Keywords

Natural Language Processing, Electronic Health Records, Machine Learning

1. Introduction

Short forms, i.e., abbreviations and acronyms, are typically found in technical language, such as in narrative content of electronic health records. Compact language expressions are often preferred by clinicians to quickly and concisely communicate and document information about patients. The drawbacks of short forms are a decrease in readability and an increase in ambiguity [1]. A systematic review covering the past 30 years, conducted on the readability of patient information, such as discharge instructions or educational health information in various clinical specialties, showed that the reading level of clinical information was too high for patients [2]. Consequently, resolving short forms could be one of the steps taken to increase readability for patients. Correct disambiguation can only be achieved using natural language processing (NLP) methods that are sensitive to the surrounding context. In 2024, a systematic scoping review on the processing of short forms in clinical narratives with NLP [3] illustrated the need for further research on this topic in languages other than English. Additionally, embeddings-based methods have demonstrated state-of-the-art performance in disambiguating short forms in English clinical texts. These methods leverage embedding representations to position semantically similar n-grams close to one another, facilitating information retrieval. However, building effective machine learning (ML) models for this task is expensive, as it relies on datasets annotated by domain experts.

The aim of this paper was to investigate the use of large language models (LLMs) to generate silver-standard training examples for a special form of short forms, *viz.* acronyms, and create embedding spaces with these examples. Silver-standard annotations are defined as automatically generated training labels that approximate gold-standard annotations. A clinical text corpus, used as a test set, was applied to gauge the applicability of this method. Three research questions guided the focus of this

SLM4Health: Improving healthcare with small language models, June 26, 2025, Pavia, Italy

[*] Corresponding author.

amila.kugic@medunigraz.at (A. Kugic); stefan.schulz@medunigraz.at (S. Schulz); markus.kreuzthaler@medunigraz.at (M. Kreuzthaler)

0000-0003-4674-0146 (A. Kugic); 0000-0001-7222-3287 (S. Schulz); 0000-0001-9824-9004 (M. Kreuzthaler)

investigation: (i) Is the application of embeddings with an LLM-generated training set feasible for acronym disambiguation? (ii) What difference does a zero-shot vs. few-shot application of LLM prompts have on the training set, and consequently on the performance results? (iii) How does the performance of embeddings for acronym resolution compare to a straightforward LLM-based disambiguation approach?

2. Related Work

For the generation of silver-standard annotation, Li et al. [4] compared both zero-shot and few-shot synthetic data generation for a variety of openly available datasets for text classification tasks, such as news and reviews, to gauge the classification performance. With BERT [5] and RoBERTa [6] models, the trained models in few-shot scenarios seem to always outperform zero-shot labels, and real-world data would almost always outperform models trained on synthetic datasets. Kruschwitz and Schmidhuber [7] analyzed the possibility of creating synthetic datasets in English for online toxicity detection. The authors summarized that while the method seems promising for further research, it did not improve the classification task compared to applying original (human annotated) data. For acronym disambiguation in English clinical narratives, Adams et al. [8] used contextualized word representations, Gaussian embeddings, and a Bayesian skip-gram model to improve short-form expansion, resulting in a performance of weighted mean F1-scores of 0.69 for the MIMIC-III [9] dataset, and 0.51 for the CASI (Clinical Abbreviation Sense Inventory) [10] dataset. Jaber and Martínez [11] outperformed Adams et al. [8] on the CASI dataset by using a masked language modeling approach with three pretrained BERT models, and incorporating context and expansions without fine-tuning, achieving 0.99 in accuracy.

3. Data

3.1. Clinical Narrative Dataset

Clinical narratives from cardiology, dermatology, and oncology departments of KAGes, an Austrian hospital network, were used to create training and test sets for acronym disambiguation. For the creation of the dataset, a rule-based approach was applied, `[\s[A-Z][A-Z0-9]{2}\s]`, to extract ambiguous two-letter acronyms with a context width of 100 characters, with the matching acronym placed in the middle. The dataset comprised three two-letter acronyms ("AP", "HT", "VA") with multiple senses per acronym. The created training set, applied for the creation of the synthetic dataset in the next step, consisted of two examples of acronyms in context from clinical narratives per possible target sense. The test set consisting of 500 contextual examples had already been applied for acronym disambiguation with LLMs [12], denoted as the *German 3A* dataset.

3.2. Synthetic Training Dataset

Two training datasets were created by prompting the language model GPT-4o (release: gpt-4o-2024-08-06) from OpenAI with the application programming interface (API). The choice of language model was informed by a prior study on acronym resolution, which had found GPT-4 to be the best-performing among four LLMs [12]. We created two separate training datasets in order to investigate the effect of zero-shot vs. few-shot prompting on the results, i.e., prompting without examples vs. prompting with examples. The prompts for both datasets were identical except for the addition of examples in the few-shot (two-shot) prompt, e.g., "*Generate 50 snippets that are 100 characters long each, in German. Each snippet should use the acronym "AP" in the middle, and correspond to the long form "Alkalische Phosphatase".*" The prompts were executed six times for each long form, due to LLM response length constraints, so that a total of 300 examples per long form could be obtained. Each prompt required the creation of 50 examples with a length of approximately 100 characters each with the acronym appearing in the middle of the example, approximately after 50 characters. The target sense for the acronym was included in the prompt for accurate context creation. The context of the 100 characters was required to be similar to clinical narrative documentation practices in German, with incomplete sentences, use of short forms, laboratory results, etc. This aimed to descriptively create similar contexts in comparison to clinical narratives for the synthetic datasets. The few-shot prompt included two examples per long form from the training dataset. Each row of the dataset consisted of a unique row number, the acronym, the target sense, and the synthetic example context.

4. Methodology

4.1. Pre-processing

Input texts processed for embedding generation and embedding search were uniformly processed. Any characters outside of `[a-zA-Z0-9üäöÄÖÜß-]` were replaced with whitespace characters, and whitespaces collapsed, so that consecutive whitespaces, created by pre-processing, were removed.

4.2. Embedding Space Generation

Prior to building the index, each short form was replaced with the corresponding long form in each synthetic example via reverse substitution to facilitate the resolution of acronyms during an embedding-based search. For 5-gram and 10-gram decomposition of the synthetic training dataset, two 768-dimensional embedding spaces were built, each for zero-shot and few-shot created examples, and the resulting vectors were indexed using Faiss [13]. The language model MedBERT.de [14] was applied to create the embedding spaces, because the language model was created with a large corpus of German medical documents and achieved state-of-the-art results in a variety of NLP tasks[1].

4.3. Search Procedure

For each acronym, a context-based search in the embeddings spaces was performed. All possible long forms found as possible senses in the corresponding clinical narrative dataset were recorded in a lookup table prior to starting the search. For disambiguation, the complete 100-character context around the indicated acronym was used for the search. Distances recorded for the nearest neighbors for each example were grouped to calculate the mean distance per possible acronym resolution. The mean embedding distance values were ordered in ascending order. The shortest mean distance resulted in the long form classification, i.e., the long form was assigned as the resolution candidate.

4.4. Evaluation

Two domain experts annotated the long forms assigned for correctness, i.e., the labels “correct” or “incorrect” were allocated for all acronym resolutions part of the *German 3A* dataset. The evaluation of the results was performed with the metric accuracy and a 95% confidence interval (CI). The metric accuracy was calculated by dividing the number of correctly labeled long forms by the total number of annotations in the dataset, while the 95% CI provided a measure of statistical significance for comparisons with other baselines, i.e., previously published works.

4.5. Baseline Comparison

The best-performing run in previous results [12], i.e., prompting for the resolution of acronyms with GPT-4, was used as a baseline comparison. The baseline aimed to resolve acronyms in a single step, where the placeholders “*ACRONYM*” and “*CONTEXT*” were substituted with the corresponding information from clinical narratives. The complete baseline zero-shot prompt: “*What is the resolution of the acronym ACRONYM in the following clinical context: CONTEXT. The answer should be kept short and concise. The acronym resolution should be given out in the following format: short form, long form. The answer should not contain any further explanations.*”

[1] https://huggingface.co/GerMedBERT/medbert-512

5. Results

In Table 1, the performance results for this acronym disambiguation task were listed. The inter-rater agreement was calculated for the test set, i.e., a Cohen's kappa κ of above 0.9 indicates a high agreement between the annotators [15, 16]. The various ways to generate silver-standard labels significantly impacted the performance between zero-shot and few-shot prompting techniques, based on the confidence intervals. Few-shot prompting for silver-standard labels generally resulted in higher accuracy compared to zero-shot prompting. Across prompt types and n-gram decompositions, the maximum accuracy of 0.69 was recorded for a 5-gram embedding space with few-shot synthetic training examples. However, this did not result in a statistically significant improvement compared to the baseline of 0.65, which consisted of prompting GPT-4 in a zero-shot manner directly for the disambiguation of the acronyms.

Table 1
Accuracy and 95% confidence interval (CI) for acronym resolution of the *German 3A* dataset via training data generated with zero-shot and few-shot prompting techniques.

Method	Prompt Type	n-gram	Accuracy	95% CI
Embeddings	zero-shot	5-gram	0.45	$[0.41 - 0.50]$
	few-shot	**5-gram**	**0.69**	$[0.64 - 0.73]$
	zero-shot	10-gram	0.47	$[0.43 - 0.52]$
	few-shot	10-gram	0.67	$[0.63 - 0.72]$
LLM baseline [12]	zero-shot		0.65	$[0.61 - 0.69]$

6. Discussion

6.1. Research Questions

6.1.1. Is the application of embeddings with an LLM-generated training set feasible for abbreviation disambiguation?

While the application of embeddings with an LLM-generated training set is feasible for abbreviation disambiguation, the results do not yet yield performance levels adequate for deployment in a clinical context , e.g., for the disambiguation of acronyms in discharge summaries. The performance results achieved similar results to Adams et al. [8], although Jaber and Martínez [11] have shown that even better acronym disambiguation would be possible. However, both publications used real-world datasets in English, i.e., they did not use synthetically created training datasets. Consequently, similarly to related works in other domains by Li et al. [4] and Kruschwitz and Schmidhuber [7], the use of on-premise datasets, labeled by annotators, would probably offer higher performance results, but at the cost of human annotation hours.

6.1.2. What difference does a zero-shot vs. few-shot application of LLM prompts have on the training set, and consequently on the performance results?

From the performance results, a statistically significant improvement (95% CI) was achieved for the use of few-shot prompting, i.e., by only introducing the same two examples into each prompt. Synthetic examples for zero-shot prompts often included wording unrepresentative for clinical narratives, including nonsensical hallucinated information that did not adhere to the initial prompt, and was in large proportions dissimilar to real-world clinical text, e.g., "*Herzton haufen alternierende Molmyklene, rhythmische Herzton [...]*" (*heart sound heap alternating Molmyklene, rhythmic heart sound*). Typical errors included hallucinated words and phrases, non-adherence to length requirements (often generating less than the required 100 characters), and while abbreviations were used in a minor number of cases, the majority of information was written out in full sentences, which might have been the largest difference between the zero-shot and few-shot examples. Conversely, the variability in all examples were given, i.e., no examples repeated themselves and were unique for each silver-standard dataset. In the few-shot prompted synthetic dataset, these erroneous effects were still there, although less in comparison to the zero-shot approach, and the examples largely similar to real-world clinical narrative datasets. An example here would be the following: "*Sättigung auf 96%, pulsierender HT, BD leicht erhöht, kein sezernierendes Exanthem, temp. norm*" (*saturation at 96%, pulsating heart sound, blood pressure slightly elevated, no secreting exanthem, temperature normal*).

6.1.3. How does the performance of embeddings for acronym resolution compare to a straightforward LLM-based disambiguation approach?

The straightforward LLM-based approach slightly underperformed compared to the best performing results for embeddings-based acronym resolution, but without statistical significance based on the confidence intervals. With the rapid advancements in LLM technology and considering that the baseline method did in fact reach 0.98 in accuracy for a subset of the CASI dataset in English, future LLMs would probably outperform the *German 3A* dataset baseline. From a resources perspective, the creation of the embedding spaces, indexing and search procedures were far less resource-intensive in comparison to the computational and memory resources needed, if one had to train and prompt LLMs on premise.

6.2. Error Analysis

A summary error analysis of acronym disambiguation revealed that for zero-shot prompted synthetic examples, the context was too dissimilar to real-world clinical narrative datasets. As a result, in 30% of nearest-neighbor searches, no nearest neighbor could be found and therefore no resolution candidate was chosen. For few-shot prompted training examples, all embedding searches found a resolution candidate. Reporting similarities of clinicians in clinical narratives, negation variations, and discontinuous spans made distinctions particularly challenging in the embedding space.

6.3. System Limitation

One limitation of this embeddings-based approach was that prior to starting the disambiguation, any possible long forms of the short forms need to be known, so that the embedding space included representative examples for each long form. To explain, a search with the context around the acronym "HT" for "Herzton" would have been found, because the generated examples contain that sense. If a search was performed for the acronym "HT", but with the resolution "Hydroxytryptamin" as part of "5-HT Rezeptor", an abbreviation for seratonin receptors, this sense would not have been found as this was not part of the list of possible senses. The latter case had no negative impact on model performance, as this sense was not represented in the *German 3A* dataset. Another limitation was the restriction to two-letter abbreviations. Even though these indicate high ambiguity contextually based on previous results [12], a larger and more diverse dataset would have been more representative for the acronym disambiguation capability of the embeddings-based approach.

7. Conclusion and Outlook

We presented a method to create synthetic training datasets for clinical narratives, labeled automatically by LLMs, to be used as input for an embeddings-based acronym disambiguation task. The results demonstrate acceptable performance sufficient for initial deployment testing with German clinical narratives. Embeddings-based acronym resolution shows great promise. In future investigations, other methods for acronym resolution will be investigated for the same dataset, which is interesting due to the high ambiguity of acronyms. One possibility would be the annotation of a subsection of clinical narratives by human annotators to compare the performance, when the embedding space is trained on the same real-world dataset.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools during the preparation of this work.

References

[1] C. M. Schwarz, M. Hoffmann, C. Smolle, M. Eiber, B. Stoiser, G. Pregartner, L. Kamolz, G. Sendlhofer, Structure, content, unsafe abbreviations, and completeness of discharge summaries: A retrospective analysis in a University Hospital in Austria, Journal of Evaluation in Clinical Practice 27 (2021) 1243–1251. URL: https://onlinelibrary.wiley.com/doi/10.1111/jep.13533. doi:10.1111/jep.13533.

[2] T. Okuhara, E. Furukawa, H. Okada, R. Yokota, T. Kiuchi, Readability of written information for patients across 30 years: A systematic review of systematic reviews, Patient Education and Counseling (2025) 108656. URL: https://www.sciencedirect.com/science/article/pii/S0738399125000230. doi:10.1016/j.pec.2025.108656.

[3] A. Kugic, I. Martin, L. Modersohn, P. Pallaoro, M. Kreuzthaler, S. Schulz, M. Boeker, Processing of Short-Form Content in Clinical Narratives: Systematic Scoping Review, Journal of Medical Internet Research 26 (2024) e57852. URL: https://www.jmir.org/2024/1/e57852. doi:10.2196/57852, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[4] Z. Li, H. Zhu, Z. Lu, M. Yin, Synthetic data generation with large language models for text classification: Potential and limitations, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 10443–10461. URL: https://aclanthology.org/2023.emnlp-main.647/. doi:10.18653/v1/2023.emnlp-main.647.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[7] U. Kruschwitz, M. Schmidhuber, LLM-based synthetic datasets: Applications and limitations in toxicity detection, in: R. Kumar, A. K. Ojha, S. Malmasi, B. R. Chakravarthi, B. Lahiri, S. Singh, S. Ratan (Eds.), Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024, ELRA and ICCL, Torino, Italia, 2024, pp. 37–51. URL: https://aclanthology.org/2024.trac-1.6/.

[8] G. Adams, M. Ketenci, S. Bhave, A. Perotte, N. Elhadad, Zero-Shot Clinical Acronym Expansion via Latent Meaning Cells, in: Proceedings of Machine Learning Research, 2020.

[9] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016) 160035. URL: http://www.nature.com/articles/sdata201635. doi:10.1038/sdata.2016.35.

[10] S. Moon, S. Pakhomov, G. Melton, Clinical Abbreviation Sense Inventory, 2012. URL: http://conservancy.umn.edu/handle/11299/137703, accepted: 2012-10-31T19:58:41Z.

[11] A. Jaber, P. Martínez, Disambiguating clinical abbreviations using a one-fits-all classifier based on deep learning techniques, Methods of Information in Medicine 61 (2022) e28–e34.

[12] A. Kugic, S. Schulz, M. Kreuzthaler, Disambiguation of acronyms in clinical narratives with large language models, Journal of the American Medical Informatics Association 31 (2024) 2040–2046. URL: https://doi.org/10.1093/jamia/ocae157. doi:10.1093/jamia/ocae157.

[13] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, IEEE Transactions on Big Data 7 (2019) 535–547.

[14] K. K. Bressem, et al., medbert.de: A comprehensive german bert model for the medical domain, Expert Systems with Applications 237 (2024) 121598. URL: https://www.sciencedirect.com/science/article/pii/S0957417423021000. doi:https://doi.org/10.1016/j.eswa.2023.121598.

[15] M. L. McHugh, Interrater reliability: the kappa statistic, Biochemia medica 22 (2012) 276–282.

[16] C. O'Connor, H. Joffe, Intercoder reliability in qualitative research: Debates and practical guidelines, International Journal of Qualitative Methods 19 (2020) 1609406919899220. URL: https://doi.org/10.1177/1609406919899220. doi:10.1177/1609406919899220. arXiv:https://doi.org/10.1177/1609406919899220.