Telehealth Ecosystems in Practice
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI230743

Clinical Acronym Disambiguation via ChatGPT and BING

Amila KUGIC[a,1], Markus KREUZTHALER[a] and Stefan SCHULZ[a]
[a] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*
ORCiD ID: Amila Kugic https://orcid.org/0000-0003-4674-0146,
Markus Kreuzthaler https://orcid.org/0000-0001-9824-9004,
Stefan Schulz https://orcid.org/0000-0001-7222-3287

Abstract. Clinical texts are written with acronyms, abbreviations and medical jargon expressions to save time. This hinders full comprehension not just for medical experts but also laypeople. This paper attempts to disambiguate acronyms with their given context by comparing a web mining approach via the search engine BING and a conversational agent approach using ChatGPT with the aim to see, if these methods can supply a viable resolution for the input acronym. Both approaches are automated via application programming interfaces. Possible term candidates are extracted using natural language processing-oriented functionality. The conversational agent approach surpasses the baseline for web mining without plausibility thresholds in precision, recall and F1-measure, while scoring similarly only in precision for high threshold values.

Keywords. Natural Language Processing, Electronic Health Records, Artificial Intelligence

1. Introduction

The language of clinical narratives is overly compact; original texts such as in findings reports, clinical and nursing notes as well as summaries like discharge letters are hastily written and not copy-edited like scientific publications. Besides clinical duties, documentation has a merely supportive role and no more effort than needed is spent on it. This means that clinical narratives are understood by those clinicians, who share knowledge on clinical domains and processes as well as the local jargon with the author, which makes them suited to communication and documentation within similar clinical settings. This is much different when texts are to be understood beyond this context. This may happen between hospitals, specialties, jurisdictions and particularly when patients or their family want to understand clinical texts.

A characteristic feature of many technical languages, particularly noteworthy in clinical narratives, is the prevalence of short forms, i.e. acronyms and abbreviations [1], [2]. Whereas medical publishers require short forms to be explicitly introduced, this rarely ever happens in clinical texts. Readers who are not very familiar with the writer's context are thus left alone. Lexicon look-up is often painful or ineffective as long as acronyms are ambiguous or understandable in a local scope only.

[1] Corresponding Author: Amila Kugic, amila.kugic@medunigraz.at

This ambiguity of short-form content is discussed in depth by Schwarz et al. [2], which is why we limit our scope to German medical acronyms, as it is the type of short form that causes most difficulties when understanding German language clinical texts, both by humans and machines. With this paper, we wanted to compare the performance of a web mining and conversational agent approach for clinical acronym resolution. For web mining, the content of SERPs (search engine result pages) are processed as proposed by Menaha and Jayanthi [3]. In comparison, ChatGPT, a conversational agent created by the company OpenAI, is utilized for the generation of possible term candidates with conversational prompts. Finally, possible resolutions for each method are generated with the help of natural language processing (NLP) and rule-based approaches, which are manually annotated and evaluated.

2. Background and Related Work

Various clinical acronym sense disambiguation approaches have been implemented to resolve the high ambiguity present in clinical notes. In 2019, the shared task on concept normalization [4] showed that most ambiguities and wrong resolutions across all participating teams stemmed from acronyms, abbreviations, misspellings present in medical notes. In this case, the most prevalent approach for acronyms seemed to be dictionary matching from online resources. This method though many times did not account for multiple senses of an acronym being present.

Link et al. [5] carried out an analysis of binary acronym disambiguation with context information from clinical notes, although the investigation was somewhat limited by the fact that the target sense needed to be identified with disease specific information prior to performing the acronym sense resolution. Menaha and Jayanthi [3] performed a survey of automatic acronym disambiguation approaches from text and web documents. Existent approaches covered the application of text mining and machine learning approaches, such as heuristic web-based acronym expansion, support vector machines, neural network models, etc.

With the introduction of large language models (LLMs) for natural language generation, new avenues of research have opened up for NLP. Examples are the synthetic generation of discharge summaries [6] or automatic retrieval of symptom mentions [7], etc. This work compares the application of text mining with ChatGPT, a chatbot powered by generative pretrained transformer. The latter is utilized to generate an acronym expansion based on a text-based input prompt.

3. Materials and Methods

3.1. Dataset

A collection of de-identified discharge summaries from different disciplines with a focus on cardiology, dermatology and oncology were used as initial dataset. From a random sample of this corpus, 143 occurrences of acronyms were extracted with their context, which is defined as a window of 70 characters. 43 acronyms with context were used for manual training and analysis, and 100 for testing. Acronyms were defined as sequences of two to seven characters, two of which need to be upper case letters.

3.2. Pre-processing

The snippets were normalized by the removal of stop words, symbols and tokens containing digits, particularly dates and lab values, which are known to heavily influence the search results. Each token in the snippet was lowercased with exception of the acronym in question and separated by whitespace characters.

3.3. Web mining

For each acronym in context, a specific SERP corpus with 400 hits was acquired. This was done via a BING application programming interface (API) query containing the acronym in its syntactic context, pre-processed as described above. The API settings included German as target language, 50 hits per query, enhanced by n additional queries with respective offsets of n x 50. The accumulated SERPs were then downloaded as text and cleaned (suppression of tokens with non-Latin characters) and analyzed as follows: Token n-gram count of the whole corpus (maximum n given by paragraph length), were ordered by decreasing frequency. For each n-gram occurring more than twice, the acronym candidate score was computed.

The rating of medical acronym-expansion pairs was done by a domain expert. The goal was to assign a plausibility score for each acronym candidate (AC) - expansion candidate (EC), with a possible maximum score of 1. This score was decreased by several features: compression, EC length imbalance, Levenshtein edit distance, casing, stop word occurrence, placement of neighboring tokens. The weight of each feature was determined by the domain expert.

The list was then ordered by decreasing scores and truncated after a threshold of 50 lines. The same was done for a corpus extract of paragraphs containing acronym-expansion pairs, such as “EKG (Elektrokardiogramm)”, or “Elektrokardiogramm (EKG)”. Two AC - EC lists were generated, for both of which the acronym candidate score multiplied by the decadic logarithm of the n-gram count was computed and ranked in decreasing order. If both lists were empty, the AC itself was taken as its EC.

3.4. Conversational agent

For each acronym and its context, a ChatGPT API query using the model *gpt-3.5-turbo* was performed, with further instructions on system behavior and resolution requirements. The system was instructed to act as an acronym disambiguation tool for medical acronyms. Each query was phrased as a question in German, with the following structure, where the capitalized words are replaced with the pre-processed content in quotation from the dataset: “*What does ACRONYM mean in context of this clinical narrative CONTEXT?*” Further instructions dealt with output requirements, i.e. JSON formatting. Per acronym in every answer, the initial acronym and its expansion were found, which were parsed and processed with rules.

3.5. Term candidates

Valid acronym expansions are sequences of one or more words that can be meaningfully separated by the characters that constitute the acronym. Non-canonical term variants (e.g. “c” instead of “k”, or “oe” instead of “ö”, as common in German clinical language) were allowed in expansions, as well as English or Latin words, but not their German translations as long as they did not match the acronym letter sequence. Pseudoacronyms are never expanded, due to these tokens only structurally resembling acronyms, e.g. Roman numbers.

3.6. Evaluation

A domain expert annotated the test set with one or more correct expansions per acronym in context. German language exhibits much morphological variability and spelling variation, therefore one single reference resolution would not be sufficient in many cases. E.g., "*Elektrokardiogramm*" (electrocardiogram) or "*elektrokardiografisch"* (electrocardiographic) would be valid resolutions for "EKG" (ECG). Each methodology delivered only one expansion candidate per acronym for final evaluation. For the evaluation of this information extraction task, the well-established performance metrics, precision, recall and F1-score were utilized for comparison. Statistical significance is determined with Chi-square hypothesis testing with the assumption for the null hypothesis being that there is no difference in performance between both methodologies. In case of statistical significance, the null hypothesis is rejected.

4. Results

In Table 1, performance metrics for both the web mining approach and the conversational agent approach are listed. From the final results, by utilizing both BING and ChatGPT for acronym resolution and disambiguation, we can see that with the application of a physician inspired scale and introducing thresholds, text mining can deliver high precision results, similar to the conversational agent approach, but recall and F1-score are significantly higher with ChatGPT, i.e. the difference is statistically significant (*$p>0.05$*). Without thresholds in place, ChatGPT supersedes the results from text mining via BING in every metric.

Table 1. Performance metrics for acronym-expansion via BING and ChatGPT

Experiment	Precision	Recall	F1-score
BING: no threshold	0.535	0.449	0.488
BING: threshold >0.1	0.530	0.440	0.481
BING: threshold >0.5	0.750	0.101	0.179
ChatGPT – gpt-3.5-turbo	0.740	0.627	0.679

5. Discussion

The results show that the web mining approach is only able to resolve clinical acronyms in an acceptable way when choosing high threshold values. High precision would be mandatory because the primary goal is to avoid false-positive resolutions. The low performance values are explainable by the fact that many clinical acronyms are only common in clinical documents and rarely on the web. A summary error analysis revealed for web mining particularly problems in cases where context tokens were not domain-specific, i.e. this yielded a SERP corpus from which mostly non-medical expansion candidates were retrieved. For the conversational agent approach, the context seemed to be unclear when examining incorrect resolutions, such as "[...] vgl VU keine [...]" (compare VU no) being resolved as "Verkehrsunfall" (traffic accident) rather than

"Voruntersuchung" (preliminary assessment). With these results, we can assume that the application of the LLM is a better alternative in comparison to text mining via BING. This does not just take performance metrics into account but also the resources involved in performing API calls, i.e. comparing the processing time, as well as the cost one of API call for ChatGPT acronym resolution versus up to 20 API calls to create a cleaned SERP corpus for each entity for web mining.

6. Conclusion and Outlook

This paper reported on the use of small corpora created out of SERPs (search engine results pages) for each acronym and its context, and compared this to a conversational agent via ChatGPT, with the application of more traditional processing techniques based on rules and corpus statistics. While the plausibility values and thresholds in web mining reduced the false positive rate by scoring acronym-expansion fitting as well as multiple other factors, this came at the cost of the suppression of many true positives. With the conversational agent, a simple solution with prompts delivered a result that improved upon the text mining baseline considerably. Future work for both approaches will consist of adjusting the window size for the context around the acronym, enhancing resolutions by including context words in queries for more specific results, and optimizing SERP corpora by suppressing non-medical content via text-genre classifiers.

References

[1] Soyer P. Acronyms, initialisms, and abbreviations. Diagnostic and Interventional Imaging. 2018 Oct;99(10):589–90. doi: 10.1016/j.diii.2018.10.002.

[2] Schwarz CM, Hoffmann M, Smolle C, Eiber M, Stoiser B, Pregartner G, et al. Structure, content, unsafe abbreviations, and completeness of discharge summaries: A retrospective analysis in a University Hospital in Austria. J Eval Clin Pract. 2021 Dec;27(6):1243–51. doi: 10.1111/jep.13533.

[3] Menaha R, Jayanthi VE. A Survey on Acronym–Expansion Mining Approaches from Text and Web. In: Satapathy SC, Bhateja V, Das S, editors. Smart Intelligent Computing and Applications. Singapore: Springer; 2019. p. 121–33. (Smart Innovation, Systems and Technologies). doi: 10.1007/978-981-13-1921-1_12.

[4] Henry S, Wang Y, Shen F, Uzuner O. The 2019 National Natural language processing (NLP) Clinical Challenges (n2c2)/Open Health NLP (OHNLP) shared task on clinical concept normalization for clinical records. J Am Med Inform Assoc. 2020 Oct 1;27(10):1529–37. doi: 10.1093/jamia/ocaa106.

[5] Link NB, Huang S, Cai T, Sun J, Dahal K, Costa L, et al. Binary acronym disambiguation in clinical notes from electronic health records with an application in computational phenotyping. International Journal of Medical Informatics. 2022 Jun 1;162:104753. doi: 10.1016/j.ijmedinf.2022.104753.

[6] Patel SB, Lam K. ChatGPT: the future of discharge summaries? The Lancet Digital Health. 2023 Mar 1;5(3):e107–8. doi: 10.1016/S2589-7500(23)00021-3.

[7] Jiang K, Mujtaba MM, Bernard GR. Large Language Model as Unsupervised Health Information Retriever. Stud Health Technol Inform. 2023 May 18;302:833–4. doi: 10.3233/SHTI230282.