Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260320

Auto Ontology: Towards Automated Term-to-Concept Assignment in Microbiology Analytics

Moritz GROB[a,b,1], Leonhard HAUPTFELD[b], Julia LIEPOLD[b,c], Jakob KAINZ[b], Andreas CSARMANN[b], Daniel RODRIGUES[d], Andrea RAPPELSBERGER[a], and Klaus-Peter ADLASSNIG[a,b]

[a]*Medical University of Vienna, Center for Medical Data Science, Institute of Artificial Intelligence, Spitalgasse 23, 1090 Vienna, Austria*

[b]*Medexter Healthcare, Borschkegasse 7/5, 1090 Vienna, Austria*

[c]*TU Wien, Institute for Logic and Computation, 1040 Vienna, Austria*

[d]*VirtualCare, Porto, Portugal*

ORCiD ID: Moritz Grob 0009-0006-7867-9465, Leonhard Hauptfeld 0009-0006-5902-4666, Julia Liepold 0009-0004-5416-6417, Daniel Rodrigues 0000-0002-0777-1540, Andrea Rappelsberger 0009-0001-0345-4011, Klaus-Peter Adlassnig 0009-0008-5046-1549

Abstract. Maintaining clinical ontologies requires continuous addition of new synonyms from laboratory information systems, which is currently a manual and time-consuming process. This study evaluates whether text embedding models can support semi-automated synonym-to-concept mapping for ontology maintenance. Using the Momo microbiology ontology, we indexed term embeddings locally with Qdrant and compared semantic similarity for assigning unseen terms to known concepts. Performance was assessed using Top-k recall, Mean Reciprocal Rank, and F1-score. Results show that dense embeddings, particularly BioLORD-2023-M (Top-1 recall = 0.68, F1 = 0.67), can retrieve relevant concept suggestions with sufficient accuracy for human-validated workflows, albeit not yet for automatic assignment. This approach demonstrates the potential of embedding-based semantic retrieval for supporting practical ontology curation in resource-constrained clinical settings.

Keywords. Momo – Monitoring of Microorganisms, Semantic Search, Natural Language Processing, Ontology Maintenance, Automation

1. Introduction

Accurate automated interpretation of clinical microbiology data is hindered by the variability of free-text laboratory reports [1]. The Momo microbiology analytics suite [2], in use at the University Hospital of Vienna (UHV) since 2013, addresses this challenge with an ontology-based service that consolidates synonymous and misspelled terms into unified concepts [3]. While this ontology substantially improves data processing, it requires continuous manual maintenance: new textual variants of terms must be

[1] Corresponding Author: Moritz Grob; E-mail: mg@medexter.com

manually associated with existing concepts each week. This process remains a significant bottleneck for scalability and timeliness [4]. The problem of manual maintenance not only affects UHV, but also large-scale international ontologies like SNOMED CT [5].

Recent advances in language models have made it possible to represent text passages as dense semantic vectors (embeddings). By comparing embeddings in a high-dimensional space, semantically similar terms can be retrieved efficiently [6]. Unlike traditional string-matching or rule-based methods, dense embeddings can capture semantic similarity across linguistic and spelling variants, which makes them particularly suited to biomedical text. Applying this to ontology maintenance could reduce manual work to a confirmation step, turning ontology curation into a semi-automated process.

The objective of this study was to evaluate whether dense text embeddings can support synonym-to-concept matching in the Momo ontology to reduce manual curation effort.

2. Methods

2.1. Data

The Momo ontology in use at UHV [3, 4] consists of a relational database with the following relevant tables: *term*, *concept*, and *concept mapping*. However, for this study only *term* and *concept* were used, since *concept mapping* describes the relationship between concepts and this study focuses on the correct classification of synonymous terms to the right concept rather than hierarchical reasoning.

The initial plan was to use two consecutive ontology exports from July and September 2025, with the latter including manually added synonyms, for evaluation purposes. However, since there were only 34 new synonyms for existing concepts in the more recent export, the evaluation was based solely on this export, with an 80/20 split of synonymous terms per concept.

The used ontology export contained 83,787 terms related to 73,045 concepts in total. Since some concepts are associated with multiple synonymous terms and misspellings, this means that the ontology contains a high number of concepts represented by only one term, limiting the ability of embeddings to learn stable concept representations. For example, the concept of *S. aureus* was not only associated with the term “Staphylococcus aureus”, but also with “Staphylococcus aureus Komplex”, “Staphylococcus aureus Kommplex”, “Staphylococcus aureus Komrplex”, and “wtStaphylococcus aureus”.

2.2. Preprocessing

Only German-language terms were included, reducing the total number of terms to 68,828. Term strings were normalized using Unicode NFKC, lowercased, and stripped of whitespace. Duplicate synonyms per concept—i.e., the same normalized string—were removed, leaving 64,135 terms to be split into training and evaluation datasets.

The split was performed with the following strategy: if a concept had multiple associated terms, 80% of those terms were embedded into the vector database right away while the remaining 20% were used to evaluate automated term-to-concept assignment by the respective embedding model. This led to a final 61,153 terms used for training and 2,982 terms for evaluation.

2.3. Embedding Models

The BAAI/bge-m3 model [7] was used as baseline embedding model due to its competitive performance in clinical retrieval benchmarks [8] as a general-purpose embedding model with multilingual support. Its performance was then compared with intfloat/multilingual-e5-base [9], another general-purpose embedding model, and FremyCompany/BioLORD-2023-M [10], an embedding model which is pre-trained for biomedical concepts.

Intentionally, lightweight models deployable on standard CPUs were selected to ensure accessibility for hospitals with limited IT infrastructure, acknowledging that larger models like BioBERT or ModernBERT might offer higher accuracy at greater computational cost.

2.4. Similarity Search

Embeddings were L2-normalized to enable cosine similarity during retrieval. The embeddings were stored in a locally hosted Qdrant vector database [11] due to the possibility of personal health data remaining in the dataset.

Cosine similarity was used as the distance metric. To establish a reproducible baseline without extensive tuning, HNSW indexing was applied with default parameters (M=16, ef_construct=100), facilitating fast nearest-neighbor search [12].

Each new term from the evaluation dataset was embedded using the same model and queried against the index using k=50 neighbors. Since multiple neighbors may belong to the same concept, we collapsed neighbors by concept by keeping the highest-ranked match per concept. This produced a ranked list of concept suggestions for each new term.

2.5. Evaluation Metrics

Retrieval effectiveness was evaluated using standard metrics in semantic search, top-k recall [13] and mean reciprocal rank (MRR) [14], as well as the weighted F1-score to account for the high number of concepts with relatively few synonyms each. These metrics were chosen as they reflect ranking quality in retrieval settings, which directly corresponds to how many relevant concept suggestions appear among the top system outputs.

To evaluate the usability of the semi-automated ontology curation process, we also calculated an auto-accept metric. A suggestion was considered to be automatically accepted if the cosine similarity margin between the top-1 and top-2 candidates exceeded a threshold (Δ=0.05). For these cases, we report the auto-accept rate (% of new terms confidently assigned) and the auto-accept precision (% of correct auto-assignments).

3. Results and Error Analysis

The described pipeline was executed for all three embedding models, producing the evaluation metrics summarized in Table 1. BioLORD-2023-M consistently outperformed the general-purpose models, particularly in recall-oriented measures.

An error analysis revealed systematic failure patterns across all embedding models. Most errors occurred in complex, domain-specific expressions rather than abbreviations.

Long multi-word expressions containing numeric resistance markers (e.g., “E. coli 3MRGN urin”, “ESBL positiv rektal”) accounted for 32–43% of all misclassifications, depending on the model. In contrast, short abbreviations (e.g., “MRSA”, “VRE”) rarely caused errors (5–10 cases per model), indicating that general biomedical abbreviations are well represented in embedding spaces.

However, all models produced a notable number of high-confidence but wrong predictions (multilingual-e5-base: 1785, bge-m3: 1357, BioLORD: 902 cases), highlighting a potential risk for automated ontology assignment. These errors justify the need for a similarity margin threshold for safe auto-acceptance and human-in-the-loop validation.

Table 1. Results of the comparative evaluation of the three chosen embedding models with the approximate 80/20-split as detailed in Section 2 and an auto-accept Δ=0.05. The total number of tested terms were N=2982.

Embedding Model	bge-m3	multilingual-e5-base	BioLORD-2023-M
Vector dimensionality	1024	768	768
Top-1 recall	0.5295	0.4014	**0.6841**
Top-3 recall	0.7103	0.5557	**0.8276**
Top-5 recall	0.7773	0.6321	**0.8605**
Top-10 recall	0.8400	0.7492	**0.9051**
MRR	0.6998	0.5584	**0.8006**
Weighted F1 score	0.5102	0.3850	**0.6690**
Auto-accept rate	**0.2492**	0.0349	0.1657
Auto-accept accuracy	0.8237	**0.9135**	0.8097

4. Discussion

As shown in Table 1, BioLORD-2023-M achieved the best overall retrieval performance, with a Top-1 recall of 0.68 and weighted F1 of 0.67, outperforming bge-m3 and multilingual-e5-base. Its biomedical training likely improved semantic clustering and reduced false matches, while the general models showed weaker alignment for microbiology-specific and numeric resistance terms.

While modern transformer architectures like ModernBERT or BioBERT might offer superior semantic reasoning, they require GPU resources often unavailable in clinical back-offices. Conversely, classical rule-based systems like MetaMap excel at standardized entity extraction but struggle with the noisy, informal vernacular of local lab reports [15]. Our approach occupies a necessary middle ground: it offers greater flexibility than strict rule-based mapping while remaining deployable on standard CPU hardware, unlike large language models. The performance ceiling of 0.68 Top-1 recall reflects this trade-off, confirming that while lightweight embeddings cannot fully automate curation, they successfully filter the search space to reduce the burden of human review.

Embedding similarity provides deterministic ranking, but it lacks contextual reasoning. LLM-assisted curation frameworks can exploit biomedical context and generate explanations but are currently limited by computational cost and privacy constraints [16]. All used models remained computationally feasible on standard CPU hardware, suggesting local deployment is practical for hospital environments. However, future work could assess larger GPU-based models (e.g., SFR-Embedding-Mistral or LLM2Vec-Llama-3-8B, as they performed well in [8]) and incorporate hybrid retrieval strategies as well as hierarchical concept mapping to evaluate near-miss predictions within ontology branches.

5. Conclusion

Embedding-based retrieval can meaningfully reduce manual effort in ontology maintenance by narrowing the candidate space to a ranked subset for human review. BioLORD-2023-M performed best overall, whereas general models offered safer high-confidence predictions. These results suggest that compact, CPU-friendly models can enable reliable semi-automatic curation, bridging the gap between fully manual work and large-scale automation. Future work should focus on hybrid retrieval strategies to handle complex microbiology expressions.

References

[1] de Quirós FGB, Otero C, Luna D. Terminology Services: Standard Terminologies to Control Health Vocabulary. Yearb Med Inform. 2018;27(01):227–33. doi: 10.1055/s-0038-1641200

[2] Medexter Healthcare GmbH. Momo [computer software]. Vienna: Medexter Healthcare; 2025 [cited 2025 Oct 30]. Available from: https://www.medexter.com/products-services/monitoring-of-microorganisms/

[3] Grob M, Kainz J, Csarmann A, Rappelsberger A, Adlassnig K-P. Enhancing Arden-Syntax-Based Clinical Reasoning with Ontologies. Stud Health Technol Inform. 2024;321:210–4. doi: 10.3233/shti241094

[4] Koller W, Kleinoscheg G, Willinger B, Rappelsberger A, Adlassnig K-P. Augmenting Analytics Software for Clinical Microbiology by Man-Machine Interaction. Stud Health Technol Inform. 2019;264:1243–7. doi: 10.3233/shti190425

[5] Ehrsam J, Gaudet-Blavignac C, Mattei M, Baumann M, Lovis C. Semantics in action: a guide for representing clinical data elements with SNOMED CT. J Biomed Semant. 2025;16:7. doi: 10.1186/s13326-025-00326-5

[6] Jing Z, Su Y, Han Y, Yuan B, Xu H, Liu C, et al. When Large Language Models Meet Vector Databases: A Survey. arXiv [Preprint]. 2025 Jun 23 [cited 2025 Oct 10]. doi: 10.48550/arXiv.2402.01763

[7] Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv [Preprint]. 2024 Jun 28 [cited 2025 Oct 15]. doi: 10.48550/arXiv.2402.03216

[8] Myers S, Miller TA, Gao Y, Churpek MM, Mayampurath A, Dligach D, et al. Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. J Am Med Inform Assoc. 2024;32(2):357–64. doi: 10.1093/jamia/ocae308

[9] Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F. Multilingual E5 Text Embeddings: A Technical Report. arXiv [Preprint]. 2024 Feb 8 [cited 2025 Oct 10]. doi: 10.48550/arXiv.2402.05672

[10] Remy F, Demuynck K, Demeester T. BioLORD-2023: Semantic Textual Representations Fusing LLM and Clinical Knowledge Graph Insights. arXiv [Preprint]. 2023 Nov 27 [cited 2025 Oct 10]. doi: 10.48550/arXiv.2311.16075

[11] Qdrant Solutions GmbH. Qdrant – Vector Database (version 1.15.5) [computer software]. Berlin: Qdrant Solutions; 2025 [cited 2025 Oct 28]. Available from: https://qdrant.tech

[12] Malkov YA, Yashunin DA. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. IEEE Trans Pattern Anal Mach Intell. 2020;42(4):824–36. doi: 10.1109/tpami.2018.2889473

[13] Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008. doi: 10.1017/cbo9780511809071

[14] Voorhees EM, Tice DM. The TREC-8 Question Answering Track Evaluation. Proc Eighth Text REtrieval Conference (TREC-8). 1999. doi: 10.6028/nist.sp.500-246.qa-overview

[15] Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Xu H. A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries. AMIA Annu Symp Proc. 2012;2012:997-1006

[16] Toro S, Anagnostopoulos AV, Bello SM, Blumberg K, Cameron R, Carmody L, et al. Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI). J Biomed Semant. 2024;15:19. doi: 10.1186/s13326-024-00320-3