TAM IMIA
Theme: Healthcare Smart x Medicine Deep
Med TAIPEI TAIWAN 2025 Info
The 20th World Congress on Medical and Health Informatics
9-13 AUGUST 2025

Improving MedDRA/J Coding Accuracy with a Fine-Tuned Text Embedding Model

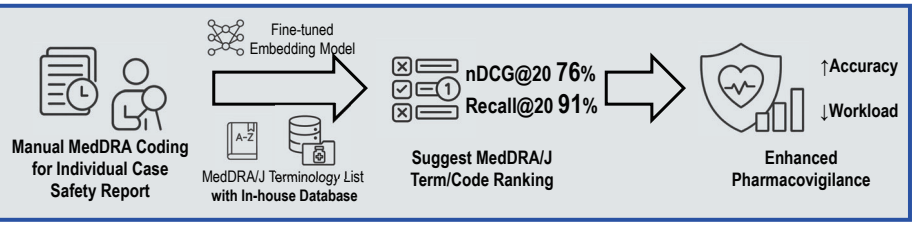
The University of Osaka

Shoya WADA[a,b], Masaharu OKAMOTO[b], Kento SUGIMOTO[b], Yasushi MATSUMURA[b,c], Katsuki OKADA[b], Shozo KONISHI[b], Takuya HARA[d], Jun KATO[d], Tadashi MATSUNO[e], Saki NAKANO[e], Toshihiro TAKEDA[b]

[a] *Department of Transformative System for Medical Information, Graduate School of Medicine, The University of Osaka*
[b] *Department of Medical Informatics, Graduate School of Medicine, The University of Osaka*
[c] *National Hospital Organization Osaka National Hospital*
[d] *Pharmacovigilance Department, Drug Development and Regulatory Science Division, Shionogi & Co., Ltd.*
[e] *Data Science Department, DX Promotion Division, Shionogi & Co., Ltd.*

Highlights

- **Aim**: Validate context-aware embeddings for MedDRA coding.
- **Impact**: Streamline MedDRA coding and strengthen pharmacovigilance.



Introduction

- ✓ **Legal context (Japan)**: Under the Act on Pharmaceuticals and Medical Devices, companies must collect and report post-marketing adverse events (AE), infections, and defects.
- ✓ **Bottleneck:** MedDRA coding—utilizing MedDRA, an international regulatory terminology—requires domain expertise but is a straightforward process. By contrast, identifying AEs from free-text, colloquial reports (e.g., call-center narratives) and coding them is a labor-intensive workflow. Because the existing MedDRA/J search tool mainly supports exact or partial string matching, coders (safety registrants) must perform query reformulation and iterative search to map the identified AEs to the correct Lowest Level Terms.
- ✓ **Prior work**: NLP and neural embeddings have improved terminology search in SNOMED CT and ICD-10(-CM), yet few studies target MedDRA—especially in Japanese.
- ✓ **Opportunity**: Transformer-based, context-aware sentence embeddings (e.g., Sentence-BERT) are promising for higher-accuracy retrieval.
- ✓ **This study**: We compare embedding-based MedDRA searches with text matching and Word2Vec baselines, and quantify gains from leveraging an in-house AE expression database.

Methods

Terminology Scope & Data Preparation

- ✓ **Coding level:** MedDRA **Lowest Level Terms (LLT)**, the most granular concept, was used for all evaluations.
- ✓ **MedDRA/J v27.0 (Mar 2024):** Removed entries flagged **"Non-current (Japanese)"**, thereby retaining only dictionary headword entries intended for current coding → **71,339 LLTs including synonyms**.
- ✓ **In-house Pharmacovigilance DB (Shionogi & Co., Ltd.)**: 244,438 AE records; after excluding pairs already in MedDRA/J, **71,813 unique AE–LLT combinations** remained, covering **45,395 AE expressions** hard to capture by exact match.

Table 1. Dataset Splits and Intended Use.

Split	Pairs (n)	Purpose
Training	50,667	Fine-tuning the text embedding and terminology expansion
Development	10,482	Hyper-parameter selection
Test	10,664	Final evaluation

Table 2. Baseline & Fine-tuned Models Overview.

Method	Brief Description
Jaccard Index	Token-set similarity; simulates unordered keyword (partial-match) search.
Word2Vec	Japanese Wikipedia Entity Vectors from Tohoku University; sentence embedding by mean-pooling token vectors.
GLuCoSE v1 / v2	Pre-trained, publicly available Japanese sentence-embedding models (PKSHA Technology Inc.); v2 applies large-scale distillation and multi-stage contrastive learning.
Fine-tuned Model (ours)	GLuCoSE v2 further fine-tuned on MedDRA/J + in-house pairs. Triplet loss: anchor = Preferred Term (PT), positive = query expression, negative = Lowest Level Term from a different PT but same High-Level Term (HLT). Hyper-parameters (learning rate, epoch, batch size) were optimised with Optuna (100 trials).

Table 3. Retrieval Metrics.

Metrics	Brief Description
Mean Average Precision (MAP)	overall ranking fidelity
nDCG@20	graded relevance within the top 20 candidates
Recall@20 and Recall@100	coverage of all correct LLTs among the first 20 and 100 results, respectively

Results

Table 4. Performance metrics of different methods in MedDRA/J terminology search.

Method	MAP	nDCG@20	Recall@20	Recall@100
Word2Vec	20.6% (19.8-21.4)	24.9% (24.1-25.7)	34.3% (33.2-35.3)	45.1% (44.1-46.2)
Jaccard index	33.8% (33.0-34.6)	40.1% (39.3-40.9)	51.5% (50.6-52.5)	59.6% (58.6-60.5)
GLuCoSEv1	32.0% (31.2-32.9)	39.5% (38.7-40.3)	52.8% (51.9-53.8)	62.8% (61.9-63.7)
GLuCoSEv2	**45.0% (44.1-45.8)**	**53.1% (52.3-53.9)**	**63.6% (62.7-64.5)**	**71.2% (70.4-72.0)**

Table 1 summarizes the results **without using the in-house database** (corresponding to the point "*Number of Database Records = 0*" in Figure 1).

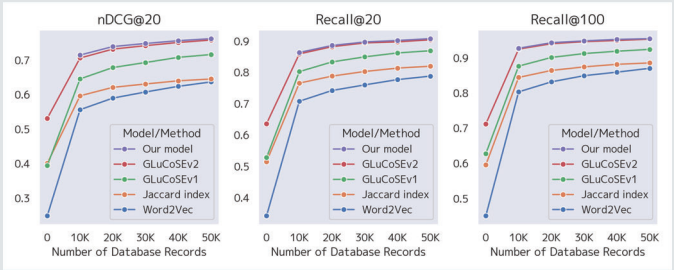


Figure 1. Effect of in-house database volume on model performance metrics.

- ✓ Compared to GLuCoSE v2, our fine-tuned model achieved a **0.4–0.9 %** improvement in nDCG@20.
- ✓ With 50K entries from the in-house dataset, our fine-tuned model reached nDCG@20 of **76.2 %**, Recall@20 of **90.8 %**, and Recall@100 of **95.4 %**.

Discussion

- ✓ **Embedding addresses out-of-dictionary phrases**: Advanced text embedding models suggested appropriate LLTs even when terms were absent from MedDRA/J terminology.
- ✓ **Practicality**: 130 M params → CPU runnable, feasible for local/on-premise use.
- ✓ **Data > Model upgrade**: Incorporating the in-house AE database improved recall/ranking more than adopting a newer embedding model; even Word2Vec + 10k in-house pairs outperformed GLuCoSE v2 alone.
- ✓ **Why gains were modest**: Our triplet-margin was limited to the HLT–PT–LLT levels; hierarchy-aware or listwise ranking losses may yield larger gains.
- ✓ **Real-world variation & generalizability:** This approach mapped colloquial input (e.g., "胃が痛い", lit. "my stomach hurts") to the correct LLT ("胃痛", lit. "Gastralgia"). The same pipeline can be generalized to other ontologies (e.g., ICD, SNOMED), provided comparable training data is available.
- ✓ **Limitations**: (1) Queries may reflect registrants' paraphrases, not raw text; (2) no external validation yet; (3) evaluation was in Japanese; however, the workflow is language-agnostic (no Japanese-specific processing), suggesting broader applicability.
- ✓ **Next steps**: External validation and hierarchy-aware/listwise ranking losses to further enhance performance and generalizability.

Conclusion

- ✓ **Embedding + in-house data significantly improved MedDRA/J search accuracy** (nDCG@20, Recall@K), reducing manual coding workload.
- ✓ **Organizational corpora are crucial:** Leveraging local AE expressions substantially boosts retrieval quality beyond model upgrades alone.
- ✓ **Implication:** Advanced NLP can streamline MedDRA coding and strengthen pharmacovigilance; future work will extend to other terminologies and explore stronger ranking objectives (including generative re-ranks).