A Hybrid Retrieval System for Adverse Event Concept Normalization Integrating Contextual Scoring, Lexical Augmentation, and Semantic Fine-Tuning

Saipriya Dipika Vaidyanathan
Faculty of Engineering, Architecture and Information Technology
The University of Queensland, Australia
s.vaidyanathan@student.uq.edu.au

Abstract

This paper presents a fully automated pipeline for normalizing adverse drug event (ADE) mentions identified in user-generated medical texts, to MedDRA concepts. The core approach here is a hybrid retrieval architecture combining domain-specific phrase normalization, synonym augmentation, and explicit mappings for key symptoms, thereby improving coverage of lexical variants. For candidate generation, the system employs a blend of exact dictionary lookups and fuzzy matching, supplemented by drug-specific contextual scoring. A sentence-transformer model (distilroberta-v1) was fine-tuned on augmented phrases, with reciprocal rank fusion unifying multiple retrieval signals.

1 Introduction

The accurate identification and standardization of Adverse Event (AE) mentions within unstructured, user-generated medical texts are critical tasks for pharmacovigilance and drug safety monitoring. This process, known as Adverse Event Concept Normalization (AECN), requires mapping patient-reported phrases to controlled medical terminologies, such as the Medical Dictionary for Regulatory Activities (MedDRA) (Combi et al., 2019). Effective normalization ensures consistency and enables accurate statistical tracking of reported events.

A significant challenge in AECN arises from the noisy and varied nature of the input data, which often contains misspellings, colloquialisms, and synonyms. To overcome this challenge, our system utilizes a hybrid retrieval architecture that combines domain-specific lexical knowledge, statistical contextualization, and semantic learning.

The core contributions of this work are demonstrated in a competitive shared task environment hosted by ALTA 2025 (Mollá et al., 2025) . The objective of this shared task was to normalise Adverse Drug Events (ADE) to standardized medical terminology such as MedDRA. Further details of the shared task can be viewed at the ALTA 2025 shared task description.

We develop a comprehensive toolkit that: (1) Implements extensive lexical normalization and augmentation to maximize coverage of variant terms.; (2) Constructs a contextual knowledge base that incorporates drug-specific co-occurrence statistics; (3) Employs a fine-tuned semantic model to enhance understanding of phrase meanings. and (4) Utilizes Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) to synthesize multiple retrieval signals into a final, optimized prediction ranking.

Our system was placed 9th on the leaderboard[1]. Although it performed weaker than the baseline model during the test phase, this outcome provided valuable learning insights. In particular, we found that the absence of explicit drug information likely contributed to the limited performance of our system despite the fragility of its contextual component when faced with missing metadata on the test phase, highlighting the need for systems resilient to input constraints.

2 Related work

The normalization of adverse drug event (ADE) mentions in unstructured user-generated or clinical texts remains a crucial focus for pharmacovigilance research and drug safety monitoring (Jeetu and Anusha, 2010; Wang et al., 2014; Beninger, 2018; Liu et al., 2019). Foundational work in this domain established the importance of standardizing clinical narratives to controlled terminologies, particularly MedDRA, which underpins consistent reporting and analysis across global regulatory bodies (Combi et al., 2018, 2019) . Mechanisms for direct mapping of patient-reported events to MedDRA have evolved from rule-based systems to neural sequence modeling and hybrid architectures.

[1] https://www.codabench.org/competitions/9717/#/results-tab

Recent studies reveal that user-generated medical text introduces unique vocabulary, substantial spelling variations, and non-standard synonyms, posing well-documented challenges to automated normalization efforts (Dirkson et al., 2019; Lee and Uzuner, 2020; Luo et al., 2019) . Lexical normalization, including spelling correction and synonym expansion, has proven essential for improving concept coverage and recall in adverse event mining (Dirkson et al., 2019; Lee and Uzuner, 2020) . Dirkson et al.(2019) specifically demonstrated that spelling correction paired with explicit mapping tables substantially increases exact match rates in normalization tasks.

Hybrid systems leveraging both lexical and semantic signals are increasingly prevalent. Luo et al.(2019) and Chen et al.(2020) show that supplementing dictionary-based candidate generation with distributed semantic representations, fine-tuned on medical pairs, yields best-in-class performance. This is especially important where mentions are ambiguous, polysemous, or not present in the dictionary.

Statistical contextualization using drug-event co-occurrence data further refines predictions, aiding in disambiguation of mentions whose context implies specific medical codes (Chopard et al., 2021). Reciprocal Rank Fusion (RRF)—introduced by Cormack et al.(2009) has emerged as an effective technique for merging ranked retrieval lists from multiple candidate generators, boosting the top-k accuracy of normalization systems by pooling diverse signals.

The ALTA series, including the ALTA 2025 Shared Task (Molla et al., 2025) , provides rigorous benchmarks for concept mapping models and exposes the limits of current contextual techniques when applied to noisy, metadata-poor user submissions.

3 Data

The primary training (`train.json`) and development (`dev.json`) files both consist of line-delimited JSON records. Each record corresponds to a user-generated medical post annotated with document-level information and a list of mention spans. Every mention includes its character offsets and the gold-standard MedDRA concept(s). These files are used to train and tune models for concept normalization.

Test data contain the same structure as the training and dev files, but without gold-standard concept annotations and drug labels (drug_id).

Dataset Type	No. of Records
Train	773
Development	161
Test	83

Table 1: Dataset information

Another file called `meddra.json` was also given. This provides the official MedDRA dictionary used for normalization, including thousands of controlled vocabulary entries encompassing AE concepts, codes, and preferred terms.

4 System Description

4.1 System Configuration and Knowledge Base Construction

The system begins by loading the training data and the complete MedDRA dictionary. A crucial preprocessing pipeline is applied to the training data to build three essential knowledge dictionaries. The process is as follows:

- **Lexical preprocessing:** Mentions undergo rigorous normalization, using a rule based approach and regular expressions for converting text to lowercase and correcting numerous domain-specific typographical errors (e.g., ’vomitting’ ’vomiting’) and common variants. Complex mentions are automatically split into sub-phrases for multi-concept querying.

- **Augmentation:** A rule-based approach, employing an explicit synonyms dictionary to generate paraphrases of mention texts (e.g., ’pain’ to ’ache’, ’discomfort’) for richer training and lookup coverage (e.g., pain, fatigue, nausea).

- **Explicit mappings:** A set of high-precision, explicit mappings for critical symptoms (e.g., nausea, vomiting, dizziness) ensures immediate high-confidence assignment for common variants.

- **Statistical dictionaries:** The processed data populates statistical dictionaries linking normalized phrases to concepts and, critically, linking the document’s drug to observed concepts to capture contextual co-occurrence statistics.

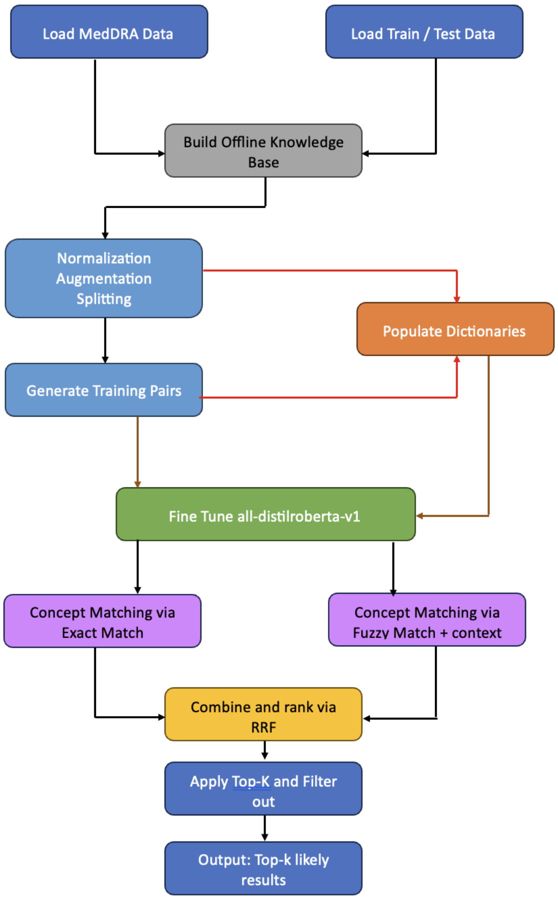
Load MedDRA Data
Load Train / Test Data
Build Offline Knowledge Base
Normalization
Augmentation
Splitting
Populate Dictionaries
Generate Training Pairs
Fine Tune all-distilroberta-v1
Concept Matching via Exact Match
Concept Matching via Fuzzy Match + context
Combine and rank via RRF
Apply Top-K and Filter out
Output: Top-k likely results

Figure 1: High level workflow of system.

4.2 Semantic Model Integration

A Sentence Transformer model based on the (all-distilroberta-v1)[2] (Reimers and Gurevych, 2019) architecture is integrated to learn semantic relationships.

- **Fine-Tuning Objective:** The model is fine-tuned for a predetermined number of epochs using a ranking-based contrastive loss function, specifically `MultipleNegativesRankingLoss` [3] (Reimers and Gurevych, 2019) This process maximizes the similarity between positive phrase/concept pairs (generated from the augmented training data) while minimizing similarity to negative examples within the training batch.

4.3 Hybrid Concept Retrieval

The concept prediction relies on combining multiple signals via Reciprocal Rank Fusion (RRF) (Cormack et al., 2009).

- **Lexical Retrieval:** Initial candidates are generated via high-confidence lookups in the statistical dictionaries and explicit mappings.
- **Fuzzy Matching[4]:** If lexical matches are insufficient, the system performs fuzzy candidate retrieval against all MedDRA terms using a token-based similarity metric.
- **Contextual Prioritization:** The fuzzy matches are prioritized using the drug context. Concepts previously associated with the document's drug are identified from the dictionary. Their fuzzy match score is boosted by a tuned, fixed weight. This ensures that contextually relevant concepts are ranked higher, even if their raw textual match score is slightly lower than a competitor.
- **Rank Fusion:** The ranked lists derived from the initial lexical matches and the contextualized fuzzy matches are unified using RRF (Cormack et al., 2009) . RRF combines the positions of concepts across the different signal lists, generating a final, optimized ranking. The system then ensures exactly predictions are returned as a json file.

5 Results

Performance was evaluated using Acc@K on the competition's test set, where the Accuracy@1 metric was used by the organizers for ranking submissions.

5.1 Official Competition Metrics

The system achieved the following accuracy scores on the training phase, and the test phase:

Metric	Accuracy	Rank
Accuracy@1	60.47%	8
Accuracy@5	66.65%	8
Accuracy@10	68.69%	8

Table 2: System Performance during Training Phase

[2] https://huggingface.co/sentence-transformers/all-distilroberta-v1

[3] https://sbert.net/docs/package_reference/sentence_transformer/losses.html#multiplenegativesrankingloss

[4] https://github.com/rapidfuzz

Metric	Accuracy	Rank
Accuracy@1	10.84%	9
Accuracy@5	14.46%	8
Accuracy@10	19.28%	8

Table 3: System Performance during Testing Phase

6 Conclusion

The results show that the correct MedDRA concept was ranked as the top prediction (A@1) in 60.47% of the training cases and just under 11% of the test cases, validating the precision of the high-confidence lexical lookups and the contextual prioritization. While the model showed promising results on training samples, its performance was more limited on new, unseen examples. Although, The significant lift to nearly 20% by A@10 confirms that the hybrid approach successfully retrieves the correct concept into the top tier in many more instances but indicates a bottleneck in accurately discriminating the single best candidate. These findings suggest that while the approach can identify learned patterns within the training data, further refinement of the ranking and mapping methodologies or expanded data coverage may be needed for consistent generalization in broader clinical concept extraction tasks, yielding more promising results in the future.

7 Limitations

- The system achieves high precision when the necessary features (like drug context, available in the training data via the doc_id and used to populate the drug to concept mapping dictionary) were present and used by the ranking system (as evidenced by the 60% training A@1). However, during the test phase, the drug-specific context boost—failed due to the missing drug label in the hidden test data, and the remaining signals (lexical and fuzzy match) were insufficient to maintain high precision on unseen examples.

- The fallback mechanism of the system when no technique produced appropriate results, was to use a random prediction of the drug label. Improvement in this technology may improve system performance and accuracy in the future.

References

Paul Beninger. 2018. Pharmacovigilance: An overview. *Clinical Therapeutics*, 40(12):1991–2004.

L. Chen, W. Fu, Y. Gu, Z. Sun, H. Li, E. Li, L. Jiang, Y. Gao, and Y. Huang. 2020. Clinical concept normalization with a hybrid natural language processing system combining multilevel matching and machine learning ranking. *J Am Med Inform Assoc*, 27(10):1576–1584. 1527-974x Chen, Long Fu, Wenbo Gu, Yu Sun, Zhiyong Li, Haodan Li, Enyu Jiang, Li Gao, Yuan Huang, Yang Journal Article England 2020/10/09 J Am Med Inform Assoc. 2020 Oct 1;27(10):1576-1584. doi: 10.1093/jamia/ocaa155.

Daphne Chopard, Matthias S Treder, Padraig Corcoran, Nagheen Ahmed, Claire Johnson, Monica Busse, and Irena Spasic. 2021. Text mining of adverse events in clinical trials: Deep learning approach. *JMIR Med Inform*, 9(12):e28632.

Carlo Combi, Margherita Zorzi, Gabriele Pozzani, Elena Arzenton, and Ugo Moretti. 2019. Normalizing spontaneous reports into meddra: Some experiments with MagiCoder. *IEEE Journal of Biomedical and Health Informatics*, 23(1):95–102.

Carlo Combi, Margherita Zorzi, Gabriele Pozzani, Ugo Moretti, and Elena Arzenton. 2018. From narrative descriptions to meddra: automagically encoding adverse drug reactions. *Journal of Biomedical Informatics*, 84:184–199.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Anne Dirkson, Suzan Verberne, Gerard van Oortmerssen, Hans van Gelderblom, and Wessel Kraaij. 2019. Lexical normalization of user-generated medical forum data. In *Proceedings of 2019 ACL workshop Social Media Mining*, volume 4.

G. Jeetu and G. Anusha. 2010. Pharmacovigilance: a worldwide master key for drug safety monitoring. *J Young Pharm*, 2(3):315–20. 0975-1505 Jeetu, G Anusha, G Journal Article India 2010/11/03 J Young Pharm. 2010 Jul;2(3):315-20. doi: 10.4103/0975-1483.66802.

K. Lee and Ö Uzuner. 2020. Normalizing adverse events using recurrent neural networks with attention. *AMIA Jt Summits Transl Sci Proc*, 2020:345–354. 2153-4063 Lee, Kahyun Uzuner, Özlem Journal Article United States 2020/06/02 AMIA Jt Summits Transl Sci Proc. 2020 May 30;2020:345-354. eCollection 2020.

F. Liu, A. Jagannatha, and H. Yu. 2019. Towards drug safety surveillance and pharmacovigilance: Current progress in detecting medication and adverse drug events from electronic health records. *Drug Saf*, 42(1):95–97. 1179-1942 Liu, Feifan Jagannatha, Abhyuday Yu, Hong R01 HL125089/HL/NHLBI NIH HHS/United States Editorial Introductory Journal Article Research Support, N.I.H., Extramural New Zealand 2019/01/17 Drug Saf. 2019 Jan;42(1):95-97. doi: 10.1007/s40264-018-0766-8.

Y. F. Luo, W. Sun, and A. Rumshisky. 2019. A hybrid normalization method for medical concepts in clinical narrative using semantic matching. *AMIA Jt Summits Transl Sci Proc*, 2019:732–740. 2153-4063 Luo, Yen-Fu Sun, Weiyi Rumshisky, Anna Journal Article United States 2019/07/02 AMIA Jt Summits Transl Sci Proc. 2019 May 6;2019:732-740. eCollection 2019.

Diego Mollá, Xiang Dai, Sarvnaz Karimi, and Cécile Paris. 2025. Overview of the 2025 alta shared task: Normalise adverse drug events. In *Proceedings of the 2025 Australasian Language Technology Workshop (ALTA 2025)*, Sydney, Australia.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

L. Wang, G. Jiang, D. Li, and H. Liu. 2014. Standardizing adverse drug event reporting data. *J Biomed Semantics*, 5:36. 2041-1480 Wang, Liwei Jiang, Guoqian Li, Dingcheng Liu, Hongfang Journal Article England 2014/08/27 J Biomed Semantics. 2014 Aug 12;5:36. doi: 10.1186/2041-1480-5-36. eCollection 2014.