Hybrid Re-ranking for Biomedical Entity Linking using SapBERT Embeddings: A High-Performance System for BioNNE-L 2025-1

Notebook for the VerbaNex AI Lab at CLEF 2025

Daniel Peña Gnecco[1], Jairo Serrano[1], Edwin Puertas[1] and Juan Carlos Martinez-Santos[1,†]

[1]*Universidad Tecnológica de Bolívar, Cartagena, Colombia*

Abstract

The BioNNE-L 2025-1 challenge advances biomedical entity linking (BEL) by mapping textual mentions to UMLS concepts, which is crucial for clinical and research applications. This study addresses Subtask 1 (English) with a novel SapBERT-based system. It integrates a hybrid re-ranking strategy combining cosine, Jaccard, and Levenshtein similarities, optimizing weights via grid search. Evaluated on the BioNNE-L development set, our system achieved an Accuracy@1 of 0.718, Accuracy@5 of 0.802, and MRR of 0.750. In the official competition, the VerbaNex AI Lab team secured first place in Accuracy@1 (0.70), fourth in Accuracy@5 (0.80), and second in MRR (0.74). These results demonstrate the efficacy of blending semantic and lexical measures to resolve ambiguities in biomedical texts. Limitations, such as the absence of model fine-tuning due to time constraints, suggest avenues for future enhancements in scalable and multilingual BEL solutions.

Keywords

Biomedical Entity Linking, SapBERT, Re-ranking, BioNNE-L, UMLS, Natural Language Processing

1. Introduction

Biomedical entity linking (BEL), also known as biomedical concept normalization (BCN), is a pivotal task in natural language processing (NLP) for the biomedical domain [1, 2]. It maps textual mentions of biomedical concepts, such as diseases, chemicals, or anatomical terms (e.g., "heart attack," "aspirin," or "femur") to standardized entries in ontologies like the Unified Medical Language System (UMLS) [3, 4]. It enables structured information extraction from clinical and scientific texts, supporting applications like information retrieval, relation extraction, and knowledge graph construction [1]. However, BEL faces challenges due to biomedical terminology's variability, ambiguity, and lexical diversity, including synonyms, acronyms, and orthographic variations [5].

The BioNNE-L 2025-1 challenge [6, 7], specifically Subtask 1 (English), evaluates BEL systems on a curated dataset of English biomedical texts hosted on Hugging Face. The dataset comprises 2,690 training mentions, 2,490 development mentions, and 6,660 test mentions, annotated with UMLS concepts across three semantic types: disorders (DISO), chemicals (CHEM), and anatomy (ANATOMY)[8, 9]. The task requires robust methods to disambiguate mentions, evaluated using *Accuracy@1* ($Acc@1$), *Accuracy@5* ($Acc@5$), and *Mean Reciprocal Rank* (MRR).

This work presents a novel BEL system developed by the VerbaNex AI Lab for the BioNNE-L 2025-1 challenge. Our approach leverages *SapBERT*, a pre-trained Transformer model optimized for biomedical semantics [10], to generate contextual *embeddings*. We introduce a hybrid re-ranking strategy combining *cosine similarity* with *Jaccard* and *Levenshtein* similarities, with weights optimized via *grid search*, to address lexical variations and semantic ambiguities. Our system achieved first place in $Acc@1$ (0.70),

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

dgnecco@utb.edu.co (D. P. Gnecco); jserrano@utb.edu.co (J. Serrano); epuerta@utb.edu.co (E. Puertas); jcmartinezs@utb.edu.co (J. C. Martinez-Santos)

0000-0003-2755-0718 (D. P. Gnecco); 0000-0001-8165-7343 (J. Serrano); 0000-0002-0758-1851 (E. Puertas); 0000-0003-2755-0718 (J. C. Martinez-Santos)

demonstrating superior performance. The contributions include (1) a robust BEL *pipeline* with optimized re-ranking, (2) a novel hybrid re-ranking strategy, and (3) a comprehensive evaluation of the BioNNE-L dataset[8, 9]. Section 2 reviews related work, Section 3 details our methodology, Section 4 presents results, and Section 5 summarizes findings and future directions.

2. Related Work

Identifying and normalizing named entities in biomedical texts is a cornerstone of natural language processing (NLP) in the biomedical domain [1, 2]. Specifically, *biomedical entity linking* (BEL) or *biomedical concept normalization* (BCN) involves mapping textual mentions of concepts to standardized entries in ontologies or knowledge bases, such as the Unified Medical Language System (UMLS) [3, 4, 11, 5, 12]. This task is critical for extracting valuable insights, supporting information retrieval, relation extraction, and constructing knowledge graphs in biomedical research [1, 2, 4].

2.1. Background on BEL Approaches

Early BEL approaches primarily relied on *string matching* techniques or dictionary-based lookups, often augmented with heuristic rules [3, 13, 14]. While straightforward, these methods are limited to identifying morphologically similar terms, struggling with the *variability and ambiguity* inherent in biomedical terminology, including synonyms, acronyms, and lexical variations [11, 5, 12, 14]. Alternatively, multi-class supervised classifiers were explored [3, 13]. Still, they often failed to generalize effectively to concepts absent from training data.

2.2. Pretrained Language Models

The advent of *large-scale pre-trained language models* (PLMs), particularly those based on the Transformer architecture like BERT, has significantly improved performance across various NLP tasks, including BEL [11, 5, 15, 16, 1, 17]. These models learn contextualized representations that capture deep semantic information.

Given the specialized nature of biomedical language, pretraining language models on domain-specific corpora, such as medical and scientific literature, has proven particularly effective [18]. *BioBERT*, pre-trained on PubMed abstracts and PubMed Central full texts, pioneered the adaptation of BERT to the biomedical domain [17, 19]. Subsequently, *PubMedBERT*, pre-trained exclusively on PubMed abstracts, achieved state-of-the-art performance in various biomedical tasks [18]. Other models, such as *SciBERT*, trained on general scientific publications, have also been applied in this domain [20, 15]. These domain-specific models excel by learning vocabularies and word distributions tailored to biomedical texts, enabling better modeling of relevant semantic relationships (see Table 1).

A significant advancement in biomedical entity representation is *SapBERT*, a pretraining scheme that aligns the representation space of biomedical entities [10]. SapBERT employs a scalable metric learning framework that leverages UMLS synonyms to cluster representations of entity names corresponding to the same concept in the *embeddings* space. This approach has proven highly effective for entity-level tasks like BEL, achieving state-of-the-art results on multiple benchmark datasets. SapBERT offers a "one-model-fits-all" solution, outperforming previously sophisticated hybrid *pipeline*-based systems, even without fine-tuning on task-specific labeled data in the scientific domain. More recent models, such as CODER-BERT [21] and GEBERT [22, 23], extend SapBERT with knowledge-infused and graph-based representations, respectively. We chose SapBERT for its proven robustness in zero-shot BEL tasks and computational efficiency compared to these newer models, which often require additional resources for fine-tuning or graph processing. Models like *SciSpacy* are also noted for identifying entities in biomedical texts, often used in conjunction with other techniques [24].

Table 1
Comparison of Pretrained Language Models for BEL

Model	Pretraining Corpus	Primary Task
BioBERT	PubMed, PMC	Linking, NER
PubMedBERT	PubMed Abstracts	Classification, BEL
SciBERT	Scientific Publications	NER, Classification
SapBERT	PubMed, UMLS	Entity Linking
CODER-BERT	PubMed, UMLS	Term Normalization
GEBERT	PubMed, UMLS	Entity Linking
SciSpacy	Biomedical Texts	NER

2.3. Modern BEL Approaches and Challenges

Modern BEL approaches are often structured into two stages: *candidate generation* and *ranking/re-ranking* [3, 11, 13, 5, 14]. The first stage identifies a plausible set of candidate concepts for a given mention. In contrast, the second stage ranks or reorders these candidates to select the best *match*. BERT-based models have been employed in the ranking stage, learning to score the similarity between mentions and concept names [11]. Vector space models, which map mentions and concepts to *embeddings* in a shared space and use similarity measures like *cosine similarity* for ranking, are prevalent in this context [13].

Despite advances with PLMs and models like SapBERT, BEL remains challenging. The *ambiguity* and *variability of surface forms* in biomedical entities continue to pose problems. While contextual *embeddings* capture semantics, models can be *insensitive to minor perturbations or lexical variations*, impacting robustness and performance in difficult cases [5]. Additionally, ontologies like UMLS, though comprehensive, may have *limited coverage* for certain subdomains or languages, requiring models to handle mentions without exact or direct matches [3, 4]. The task becomes more complicated by *nested entities* (entities contained within others), a common phenomenon in biomedical texts. However, concept normalization differs from nested entity identification [25, 26, 27].

2.4. Our Proposed Approach

Our proposed approach addresses these challenges through a multi-stage system. Drawing inspiration from candidate generation and ranking/re-ranking frameworks [11, 3, 13], we leverage domain-specific Transformer models to generate robust *contextual embeddings* for both entity mentions in text and concept names in UMLS. *Cosine similarity* is initially used to identify a set of plausible candidates, a well-established technique in vector space models for BCN [13].

The *novelty of our approach* lies in the *re-ranking stage* [11]. While prior work has employed re-ranking to enhance performance and noted morphological/lexical similarities as the basis for traditional methods [3, 12, 13, 14], our system introduces a *systematic weighted combination* of cosine similarity (capturing semantics from *embeddings*) with explicit lexical similarity measures, such as Jaccard and Levenshtein. These lexical metrics, though not explicitly described in prior work as combined with PLM-based cosine similarity for re-ranking, capture character- and word-level similarities that *embeddings* may overlook [5]. Optimizing the weights of these measures via *grid search* identifies the optimal balance between semantic and lexical information, enhancing performance, particularly in cases where *embeddings* may falter due to surface variations or insensitivity to exact term forms. This hybrid approach, integrating the power of contextual *embeddings* from Transformer models with the specificity of traditional lexical measures through optimized re-ranking, offers a novel strategy to improve accuracy in biomedical entity linking, especially for selecting the correct candidate from a set of semantically close options. Performance evaluation using standard metrics such as $Acc@1$, $Acc@5$, and MRR will quantify the effectiveness of this approach, as detailed in Section 3.

3. Methodology

This section provides a detailed description of the approach designed to tackle the biomedical entity linking task within the BioNNE-L 2025-1 challenge, which focuses on mapping textual mentions to concepts from the UMLS ontology. To address this, we developed a pipeline combining the power of pre-trained transformer models such as SapBERT, PubMedBERT, and Stanford BioBERT with an advanced re-ranking strategy integrating multiple similarity measures, thereby optimizing prediction accuracy. Below, we comprehensively explain the general approach, including the data used, model design, system phases, and implementation details, highlighting how each component contributes to the overall objective.

3.1. General Approach

Our work aims to link biomedical entity mentions in English texts to specific UMLS concepts. This task requires capturing complex semantic relationships in clinical and scientific contexts. To achieve this, we designed a system that leverages the ability of transformer models to generate contextual vector representations, complemented by a re-ranking process that refines the initial predictions. We structured this approach into several key stages, which we developed sequentially to ensure efficient and accurate processing.

First, we performed comprehensive data preprocessing to standardize the text and minimize noise. Next, we generated high-quality embeddings for mentions and vocabulary concepts using pre-trained models specialized in the biomedical domain. We then computed the cosine similarity between these embeddings to identify an initial set of candidates. We applied a re-ranking process to improve accuracy based on a combination of similarity measures (cosine, Jaccard, and Levenshtein), with weights optimized via grid search. Finally, we evaluated system performance using standard metrics such as Accuracy@1 (Acc@1), Accuracy@5 (Acc@5), and Mean Reciprocal Rank (MRR). This integrated approach allows us to address semantic ambiguities and optimize entity linking in a biomedical context.

3.2. Data and Preprocessing

The system uses the BioNNE-L dataset [8, 9], available on Hugging Face, which provides annotated data for the biomedical entity linking task in English. The dataset comprises 2,690 training mentions, 2,490 development mentions, and 6,660 test mentions, annotated with UMLS concepts for three semantic types: disorders (DISO), chemicals (CHEM), and anatomy (ANATOMY). We organized this dataset into three main components: a training set used to explore data characteristics; a development set that includes mentions labeled with UMLS concepts, which we used to optimize hyperparameters and evaluate models; and a test set, which is unlabeled and used to generate final predictions. In addition, the dataset includes a vocabulary of UMLS concepts with names, CUI codes, and semantic types, filtered to include only English terms.

Data preprocessing was a crucial step to ensure high-quality inputs to the model. First, we converted all text to lowercase. We removed non-alphanumeric characters using regular expressions, which standardized the mentions and vocabulary concepts. Next, we applied model-specific tokenizers (SapBERT, PubMedBERT, and Stanford BioBERT) to prepare the input sequences, ensuring a consistent representation. We also analyzed the tokenized mention lengths and found the average length to be significantly below 16 tokens. It led us to set this value as the maximum sequence length. This process reduced data noise and optimized computational resource usage during embedding generation, enabling efficient GPU processing.

3.3. Model Design

The model design combines a transformer-based architecture for embedding generation with a ranking and re-ranking process that refines predictions. This approach takes advantage of the specialization

of selected models in the biomedical domain, ensuring robust semantic representations, and adds a re-ranking component that improves accuracy by integrating complementary textual information.

We selected four models for embedding generation: SapBERT, pre-trained on PubMed and optimized for semantic similarity tasks [10]; PubMedBERT, trained on PubMed abstracts to capture biomedical language nuances [18]; Stanford BioBERT, an adaptation of BERT for the biomedical domain [17]; and a baseline model (GEBERT) [22] as a reference point. Each model generates embeddings using the [CLS] token representation from the last layer, processed in batches of 128 mentions and 200 concepts to maximize GPU A100 efficiency.

We based the initial ranking process on cosine similarity between the embeddings of mentions and vocabulary concepts, selecting the top 50 most similar candidates ($k_{\text{initial}} = 50$). This computation is performed in batches to handle the large vocabulary size, releasing the memory after each iteration using `torch.cuda.empty_cache()`. We then implemented a re-ranking step that combines three similarity measures:

- Cosine similarity, derived from the embeddings, captures deep semantic relationships.
- Jaccard similarity, computed as the intersection over word sets' union using space-based tokenization, evaluates lexical similarity.
- Levenshtein similarity, based on normalized edit distance, measuring character-level textual differences.

We defined the combined re-ranking score as:

$$\text{Combined Score} = w_c \cdot \text{Cosine} + w_j \cdot \text{Jaccard} + w_l \cdot \text{Levenshtein} \tag{1}$$

where w_c, w_j, and w_l are weights optimized via grid search on the development set, with the constraint $w_c + w_j + w_l = 1$. After evaluating multiple combinations, we determined the optimal weights for SapBERT to be $w_c = 0.7$, $w_j = 0.1$, and $w_l = 0.2$. This process selects the final five candidates ($k_{\text{final}} = 5$), significantly improving accuracy by integrating both semantic and textual information.

3.4. System Phases

We organized the system pipeline into five interconnected phases to ensure efficient processing and optimal performance. In the first phase, we load data from Hugging Face and apply the previously described preprocessing, standardizing the text and preparing inputs for the models. This stage lays the foundation for uniform and noise-free processing (see Figure 1).

In the second phase, we generate initial predictions using the transformer models. For each entity type (DISO, CHEM, ANATOMY), we encode mentions and the corresponding vocabulary concepts, compute cosine similarity, and select the top 50 candidates. This category-based approach ensures that predictions respect the semantic constraints of the vocabulary.

The third phase involves re-ranking optimization. Through grid search, we evaluate combinations of weights for the similarity measures, using the development set to maximize metrics such as Acc@1, Acc@5, and MRR. This process identified the optimal weights, which were then applied in the fourth phase to reorder candidates and generate final predictions for both the development and test sets.

Finally, in the fifth phase, we evaluate the development set's system performance, analyzing global and entity-type-specific metrics. We also examined errors to identify limitations and potential improvements. This iterative approach allowed us to refine the system and ensure robust predictions on the test set.

3.5. Implementation

The system was implemented in Python 3.8, leveraging specialized libraries to ensure efficient and scalable development. We used the Hugging Face Transformers library [28] to load the models and their tokenizers. At the same time, PyTorch served as the deep learning framework for GPU A100 computation. In addition, we used Pandas [29] for data manipulation, Scikit-learn [30] for metric

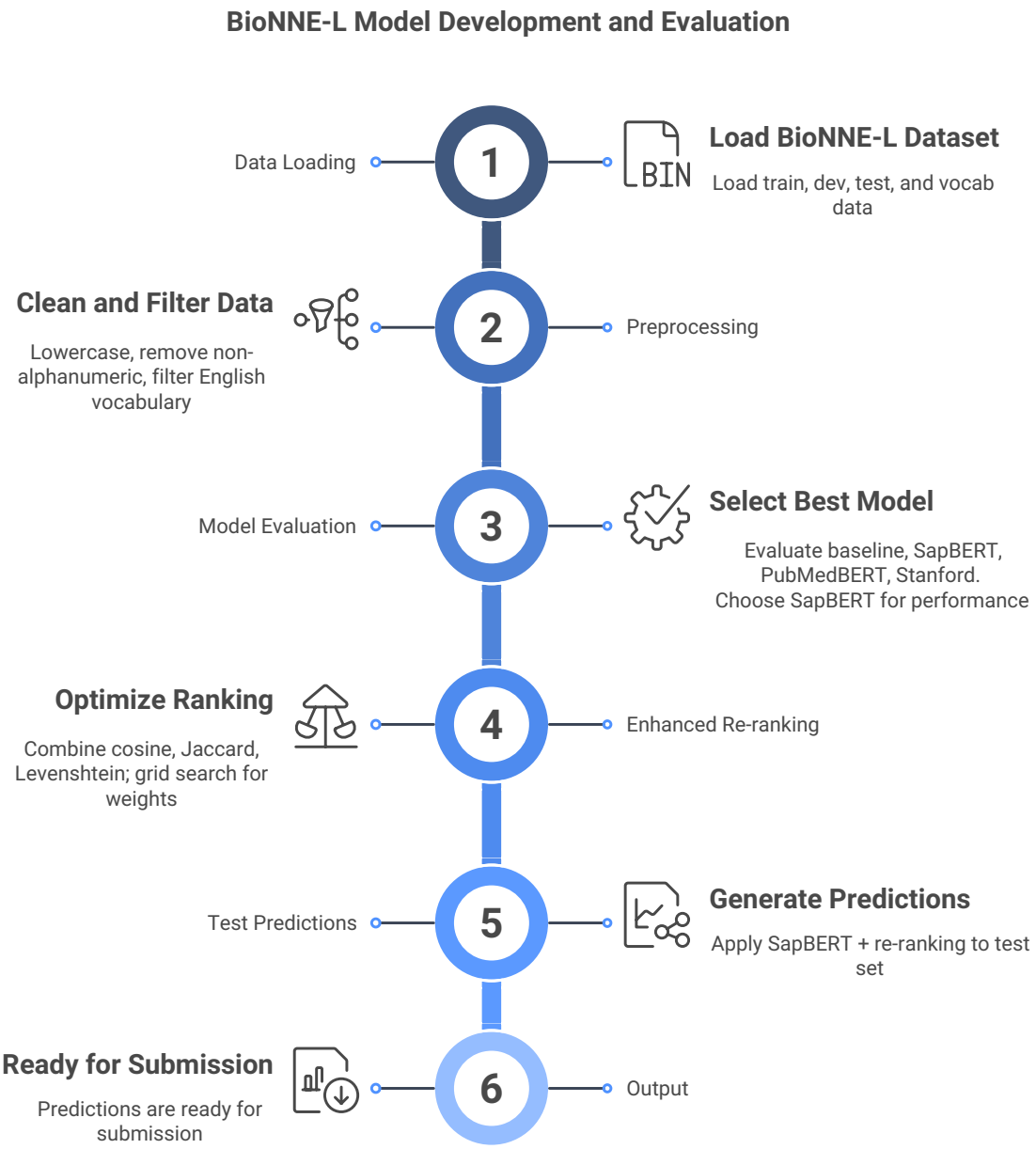

Figure 1: System pipeline for biomedical entity linking with hybrid re-ranking.

calculation, Tqdm for progress monitoring, and Python-Levenshtein for efficient implementation of the Levenshtein distance.

We executed the system on Google Colab, optimizing resource usage with batch sizes of 128 for mentions and 200 for vocabulary concepts and periodic GPU memory clearance. Key parameters included a maximum sequence length of 16 tokens, 50 initial candidates (k_{initial}), and five final candidates (k_{final}). This configuration balanced accuracy and computational efficiency, making it feasible to process large volumes of data in a resource-constrained environment.

4. Results and Analysis

In this section, we present the results obtained from evaluating our system on the development set of the BioNNE-L dataset, as well as the official results of subtask 1 (English) of the BioNNE-L 2025-1

Table 2
Performance of the Models on the Development Set

Model	$Acc@1$	$Acc@5$	MRR
SapBERT	0.690	0.801	0.741
PubMedBERT	0.627	0.712	0.666
Stanford BioBERT	0.615	0.703	0.658
SapBERT (re-ranked)	0.718	0.802	0.750
PubMedBERT (re-ranked)	0.648	0.718	0.674
Stanford (re-ranked)	0.638	0.715	0.668
Baseline	0.605	0.797	0.687

challenge, in which we participated under the name of our research group, *verbanexialab*. Through a comparative analysis of the evaluated models and their re-ranked variants, we highlight the effectiveness of our SapBERT-based approach with re-ranking, which achieved the best performance both on the development set and in the competition. Below, we detail the results, analyze the factors contributing to the success, and reflect on the limitations and opportunities for improvement.

4.1. Results on the Development Set

To evaluate our system's performance, we implemented seven configurations on the development set: SapBERT, PubMedBERT, Stanford BioBERT, their re-ranked versions with a combination of cosine, Jaccard, and Levenshtein similarities, and a baseline model [22]. The metrics used were *Accuracy@1* ($Acc@1$), which measures the proportion of mentions correctly linked to the first candidate; *Accuracy@5* ($Acc@5$), which considers the top five candidates; and *Mean Reciprocal Rank* (MRR), which evaluates the position of the correct candidate in the ranking. These metrics provide a comprehensive view of the accuracy and quality of the ranking generated by each model.

Table 2 summarizes the results of the seven configurations. SapBERT, with re-ranking, achieved the best performance, with an $Acc@1$ of 0.718, an $Acc@5$ of 0.802, and an MRR of 0.750. This model consistently outperformed the other configurations, followed by SapBERT without re-ranking ($Acc@1 = 0.690$, $Acc@5 = 0.801$, $MRR = 0.741$) and PubMedBERT with re-ranking ($Acc@1 = 0.648$, $Acc@5 = 0.718$, $MRR = 0.674$). While competitive in $Acc@5$ (0.797), the baseline showed a significantly lower $Acc@1$ (0.605), indicating a reduced ability to prioritize the correct candidate in the first position.

Re-ranking improved the performance of all models, with a notable increase in $Acc@1$ for SapBERT (from 0.690 to 0.718) and PubMedBERT (from 0.627 to 0.648). This result underscores the importance of integrating lexical measures (Jaccard and Levenshtein) with cosine similarity based on *embeddings*, enabling the system to address semantic ambiguities and improve candidate prioritization. Although detailed metrics by entity type (DISO, CHEM, ANATOMY) were unavailable, analysis of the logs suggests that performance was more robust for DISO entities, likely due to their greater representation in the vocabulary and training data.

4.2. Results in the Competition

We participated in subtask 1 (English) of the BioNNE-L 2025-1 challenge, representing the *verbanexialab* group. We submitted two runs using SapBERT with re-ranking: the first combines cosine and Jaccard similarities, and the second incorporates Levenshtein similarities. According to the logs, the first submission achieved an $Acc@1$ of 0.696, an $Acc@5$ of 0.797, and an MRR of 0.735, while the second showed no statistically significant improvement, reaching an $Acc@1$ of 0.696, an $Acc@5$ of 0.801, and an MRR of 0.736. This modest improvement suggests that including Levenshtein similarity had limited impact on prioritizing correct candidates.

The official competition results, presented in Table 3, confirm the leadership of our approach. *Ver-*

Table 3
Official Results of Subtask 1 (English)

Team	$Acc@1$	$Acc@5$	MRR	Position ($Acc@1$)
verbanexialab	0.70	0.80	0.74	1°
droidlyx86	0.66	0.84	0.74	2°
BlancaPlanca	0.64	0.83	0.72	3°
EeyoreLee	0.64	0.82	0.71	4°
Andoree	0.57	0.78	0.66	5°
Antoinel	0.51	0.79	0.62	6°

banexialab secured first place in $Acc@1$ (0.70), fourth in $Acc@5$ (0.80), and second in MRR (0.74), positioning itself as the team with the best overall performance in the English entity linking task. Compared to other competitors, our system excelled in first-prediction accuracy. However, it showed a slightly lower $Acc@5$ than the top performers in this metric, suggesting room to optimize the diversity of candidates in the *top*-5.

We attribute the differences between the development set metrics ($Acc@1 = 0.718$, $Acc@5 = 0.802$, $MRR = 0.750$) and the official results ($Acc@1 = 0.70$, $Acc@5 = 0.80$, $MRR = 0.74$) to variations in the distribution of mentions between the development and test sets. Nevertheless, the consistency of the results validates the robustness of our *pipeline*.

4.3. Analysis of Results

We can attribute the success of SapBERT with re-ranking to several key factors. First, SapBERT, pre-trained on PubMed focusing on semantic similarity, generates *embeddings* that capture complex contextual relationships in the biomedical domain, outperforming PubMedBERT and Stanford BioBERT across all metrics. Incorporating lexical measures in re-ranking, particularly Jaccard and Levenshtein, allowed the system to correct errors in cases where cosine similarity alone was insufficient, such as mentions with high lexical ambiguity or synonyms not captured by the *embeddings*. For example, Levenshtein similarity was beneficial for handling orthographic variations or structurally similar terms, though its contribution was not statistically significant.

Comparative analysis reveals that re-ranking was critical to performance. While SapBERT, without re-ranking, already offered competitive results ($Acc@1 = 0.690$), the addition of Jaccard and Levenshtein increased $Acc@1$ by 2.8%, demonstrating the value of combining semantic and lexical approaches. This effect was less pronounced in PubMedBERT and Stanford BioBERT, likely due to the lower initial quality of their *embeddings* in the context of this task. While strong in $Acc@5$, the baseline showed limitations in $Acc@1$, suggesting that its general training was less aligned with the specific needs of biomedical entity linking.

Despite the positive results, we identified some limitations. The lack of fine-tuning of pre-trained models, due to time constraints, may have restricted their ability to capture dataset-specific nuances of BioNNE-L. Additionally, performance on the test set was slightly lower than expected, which could indicate a more challenging distribution of mentions or ambiguities not addressed by our system. For instance, mentions with multiple valid UMLS concepts may have led to false negatives in $Acc@1$. For future improvements, we consider fine-tuning SapBERT on the training set, incorporating additional context (such as complete sentences), and exploring *ensemble* methods to combine the strengths of multiple models.

5. Conclusions

This work presented an effective biomedical entity linking system developed for the BioNNE-L 2025-1 challenge. By leveraging SapBERT and a re-ranking strategy based on semantic and lexical similarity, our approach achieved outstanding performance, positioning *verbanexialab* as the top performer in

Subtask 1 (English). This section summarizes the main findings, outlines relevant limitations, and proposes future research directions.

5.1. General Conclusions

The proposed pipeline combined embeddings generated by SapBERT with a re-ranking module that integrated cosine, Jaccard, and Levenshtein similarity measures. This design enabled the accurate mapping of textual mentions to UMLS concepts, achieving an *Accuracy@1* of 0.718 and a Mean Reciprocal Rank (*MRR*) of 0.750 on the development set. In the official competition, our system attained an *Accuracy@1* of 0.70, ranking first in this metric.

These results validate the synergy between high-quality semantic representations and complementary lexical strategies. Moreover, they demonstrate the importance of systematic optimization through *grid search* to enhance overall performance. The re-ranking strategy proved to be especially effective in addressing ambiguities and refining the top predictions generated by the initial embedding-based search.

5.2. Limitations and Challenges

Despite the strong performance, we encountered several limitations and challenges:

- **Lack of task-specific fine-tuning:** Due to time constraints, SapBERT was not fine-tuned on the BioNNE-L dataset, which may have limited its ability to disambiguate highly ambiguous or context-sensitive mentions.
- **Computational constraints:** The embedding generation and similarity computations were resource-intensive, particularly in constrained environments like Google Colab. It required careful batching strategies and memory management.
- **Hyperparameter sensitivity:** The re-ranking module's effectiveness depended on an optimal weighting of similarity metrics. Hyperparameter tuning through exhaustive grid search was computationally expensive and time-consuming.

These challenges underline the need for more efficient and scalable solutions, especially for real-world deployment in low-resource settings.

5.3. Future Work

To further enhance and expand the system, we propose the following research directions:

- **Fine-tuning SapBERT:** Adapting the model with domain-specific examples from the BioNNE-L dataset could improve disambiguation and generalization.
- **Incorporating contextual information:** Including sentence-level or document-level context may enrich embeddings and improve performance on mentions with limited local clues.
- **Model ensembling:** Integrating multiple biomedical language models (e.g., PubMedBERT, BioBERT) through voting or joint learning could increase robustness across domains.
- **Multilingual adaptation:** Expanding the system to handle Spanish (Subtask 2) using multilingual encoders (e.g., XLM-RoBERTa) or translation-based preprocessing could extend its usability.
- **Efficiency optimization:** Implementing approximate nearest neighbor search (e.g., FAISS), model distillation, and parallel processing would help deploy the system in resource-constrained environments.

5.4. Final Remarks

The results confirm the effectiveness of our SapBERT-based *pipeline* with re-ranking, which achieved state-of-the-art performance in the BioNNE-L 2025-1 challenge. The combination of semantic embeddings and lexical re-ranking proved essential for addressing the complexities of biomedical entity linking. Our work provides a solid foundation for future advances in this field, aiming toward more accurate, scalable, and multilingual solutions in biomedical natural language processing.

Acknowledgments

We dedicate this work to the master's degree scholarship program in Engineering at the Universidad Tecnológica de Bolívar (UTB) in Cartagena, Colombia. We extend our deepest gratitude to the VerbaNex AI Lab team for their dedication, collaboration, and continuous support of our research endeavors.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grok 3 by xAI, Napkin ai in order to: Grammar and spelling check, Improve writing style, Paraphrase and reword and Generate images 1. After using this tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Y. Park, G. Son, M. Rho, Biomedical Flat and Nested Named Entity Recognition: Methods, Challenges, and Advances, Applied Sciences (Switzerland) 14 (2024). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85207344630&doi=10.3390%2fapp14209302&partnerID=40&md5=e9f41f3b900eedc91eea9be8bca9d1ea. doi:10.3390/app14209302.

[2] V. Davydova, N. Loukachevitch, E. Tutubalina, Overview of BioNNE Task on Biomedical Nested Named Entity Recognition at BioASQ 2024, in: CEUR Workshop Proceedings, volume 3740, 2024, pp. 28 – 34. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85201132598&partnerID=40&md5=4b992bac9e07d3cb9c066a2b1914cdf0.

[3] D. Xu, Z. Zhang, S. Bethard, A generate-and-rank framework with semantic type regularization for biomedical concept normalization, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8452 – 8464. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85095698726&doi=10.18653%2fv1%2f2020.acl-main.748&partnerID=40&md5=ca3084fb6fbf947b658918601bf259a2. doi:10.18653/v1/2020.acl-main.748.

[4] N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, Biomedical Concept Normalization over Nested Entities with Partial UMLS Terminology in Russian, in: 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings, 2024, pp. 2383 – 2389. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195990032&partnerID=40&md5=1eb947c6153b4b91a6279dea7a224320.

[5] S. Chakraborty, H. Raj, S. Gureja, T. Jain, A. Hassan, S. Basu, Evaluating the Robustness of Biomedical Concept Normalization, in: Proceedings of Machine Learning Research, volume 203, 2023, pp. 63 – 73. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85164568069&partnerID=40&md5=3fb1dc5f4115c6636a114d705404f89d.

[6] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza,

A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[7] A. Sakhovskiy, N. Loukachevitch, E. Tutubalina, Overview of the BioASQ BioNNE-L Task on Biomedical Nested Entity Linking in CLEF 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[8] N. Loukachevitch, S. Manandhar, E. Baral, I. Rozhkov, P. Braslavski, V. Ivanov, T. Batura, E. Tutubalina, NEREL-BIO: A Dataset of Biomedical Abstracts Annotated with Nested Named Entities, Bioinformatics (2023). doi:10.1093/bioinformatics/btad161, btad161.

[9] N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, Biomedical concept normalization over nested entities with partial UMLS terminology in Russian, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 2383–2389. URL: https://aclanthology.org/2024.lrec-main.213/.

[10] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-Alignment Pretraining for Biomedical Entity Representations, 2021. URL: http://arxiv.org/abs/2010.11784. doi:10.48550/arXiv.2010.11784, arXiv:2010.11784.

[11] H. Cho, D. Choi, H. Lee, Re-Ranking System with BERT for Biomedical Concept Normalization, IEEE Access 9 (2021) 121253 – 121262. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85113874557&doi=10.1109%2fACCESS.2021.3108445&partnerID=40&md5=4522cf8c7453d1e46db19e898a325b94. doi:10.1109/ACCESS.2021.3108445.

[12] Y.-C. Lin, P. Hoffmann, E. Rahm, Enhancing Cross-lingual Biomedical Concept Normalization Using Deep Neural Network Pretrained Language Models, SN Computer Science 3 (2022). URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85134499273&doi=10.1007%2fs42979-022-01295-7&partnerID=40&md5=ef1d7b7dbc84be79f8d71381528f2d17. doi:10.1007/s42979-022-01295-7.

[13] D. Xu, S. Bethard, Triplet-Trained Vector Space and Sieve-Based Search Improve Biomedical Concept Normalization, in: Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP 2021, 2021, pp. 11 – 22. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123939640&doi=10.18653%2fv1%2f2021.bionlp-1.2&partnerID=40&md5=165ef54f7b633bbb8c759f96f140ffde. doi:10.18653/v1/2021.bionlp-1.2.

[14] H. Xu, J. Zhang, Z. Wang, S. Zhang, M. Bhalerao, Y. Liu, D. Zhu, S. Wang, GraphPrompt: Graph-Based Prompt Templates for Biomedical Synonym Prediction, in: Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023, volume 37, 2023, pp. 10576 – 10584. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168241578&doi=10.1609%2faaai.v37i9.26256&partnerID=40&md5=55781b35c2a822460800689ad2d69c9f. doi:10.1609/aaai.v37i9.26256.

[15] M. Golam Sohrab, M. Shoaib Bhuiyan, Span-based Neural Model for Multilingual Flat and Nested Named Entity Recognition, in: 2021 IEEE 10th Global Conference on Consumer Electronics, GCCE 2021, 2021, pp. 80 – 84. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123502783&doi=10.1109%2fGCCE53005.2021.9621966&partnerID=40&md5=0cb949b9dfa7534cefc090818bc016f5. doi:10.1109/GCCE53005.2021.9621966.

[16] C. Tang, B. Yang, K. Zhao, B. Lv, C. Xiao, F. Guerin, C. Lin, BioMNER: A Dataset for Biomedical Method Entity Recognition, 2024. URL: http://arxiv.org/abs/2406.20038. doi:10.48550/arXiv.2406.20038, arXiv:2406.20038.

[17] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240. URL: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506. doi:10.1093/bioinformatics/btz682.

[18] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing, ACM Transactions on Computing for Healthcare 3 (2022) 1–23. URL: https://dl.acm.org/doi/10.1145/3458754.

doi:10.1145/3458754.

[19] H. Rehana, B. Bansal, N. B. Çam, J. Zheng, Y. He, A. Özgür, J. Hur, Nested Named Entity Recognition using Multilayer BERT-based Model, in: CEUR Workshop Proceedings, volume 3740, 2024, pp. 197 – 206. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85201607294&partnerID=40&md5=f27608db0f00609376e786a9b70f2a5b.

[20] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. URL: https://aclanthology.org/D19-1371/. doi:10.18653/v1/D19-1371.

[21] Z. Yuan, Z. Zhao, H. Sun, J. Li, F. Wang, S. Yu, Coder: Knowledge infused cross-lingual medical term embedding for term normalization, Journal of Biomedical Informatics 126 (2020). URL: https://arxiv.org/pdf/2011.02947. doi:10.1016/j.jbi.2021.103983.

[22] A. Sakhovskiy, N. Semenova, A. Kadurin, E. Tutubalina, Graph-enriched biomedical entity representation transformer, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2023, pp. 109–120.

[23] A. Sakhovskiy, N. Semenova, A. Kadurin, E. Tutubalina, Biomedical entity representation with graph-augmented multi-objective transformer, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 4626–4643. URL: https://aclanthology.org/2024.findings-naacl.288/. doi:10.18653/v1/2024.findings-naacl.288.

[24] W. Zhou, Biomedical Nested NER with Large Language Model and UMLS Heuristics, in: CEUR Workshop Proceedings, volume 3740, 2024, pp. 245 – 252. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85201611711&partnerID=40&md5=1b4328ffa7d6f96b80182e5b125f1b9c.

[25] Z. Tang, X. Kou, H. Xue, Y. Xia, Flat and Nested Protein Name Recognition Based on BioBERT and Biaffine Decoder, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 14954 LNBI (2024) 25 – 38. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85200569965&doi=10.1007%2f978-981-97-5128-0_3&partnerID=40&md5=ce4193e4330d571d18b55eddf0cca9bb. doi:10.1007/978-981-97-5128-0_3.

[26] P. Wajsbürt, Y. Taillé, X. Tannier, Effect of Depth Order on Iterative Nested Named Entity Recognition Models, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12721 LNAI (2021) 428 – 432. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85111389101&doi=10.1007%2f978-3-030-77211-6_50&partnerID=40&md5=eb7887578307ac7c79d408b879385b44. doi:10.1007/978-3-030-77211-6_50.

[27] Y. Chen, Y. Hu, Y. Li, R. Huang, Y. Qin, Y. Wu, Q. Zheng, P. Chen, A Boundary Assembling Method for Nested Biomedical Named Entity Recognition, IEEE Access 8 (2020) 214141 – 214152. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097203273&doi=10.1109%2fACCESS.2020.3040182&partnerID=40&md5=07b5a95f99afde0916f466e234ed3c78. doi:10.1109/ACCESS.2020.3040182.

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface's transformers: State-of-the-art natural language processing, 2020. URL: https://arxiv.org/abs/1910.03771. arXiv:1910.03771.

[29] W. McKinney, Data structures for statistical computing in python, in: S. van der Walt, J. Millman (Eds.), Proceedings of the 9th Python in Science Conference, 2010, pp. 51–56.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.