Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Enhancing cross-encoders using knowledge graph hierarchy for medical entity linking in zero- and few-shot scenarios

Fernando Gallego [a,b,*], Pedro Ruas [c], Francisco M. Couto [c], Francisco J. Veredas [a,b]

[a] Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Málaga, Spain
[b] Research Institute of Multilingual Language Technologies, Universidad de Málaga, Málaga, Spain
[c] LASIGE, Faculdade de Ciências, Universidade de Lisboa, 1749-016, Lisbon, Portugal

ARTICLE INFO

Keywords:
Knowledge enrichment
Medical entity linking
Contrastive-learning
Knowledge bases
Candidate reranking
Zero-few shot

ABSTRACT

Medical Entity Linking (MEL) is a common task in natural language processing, focusing on the normalization of recognized entities from clinical texts using large knowledge bases (KBs). This task presents significant challenges, especially when working with electronic health records that often lack annotated clinical notes, even in languages like English. The difficulty increases in few-shot or zero-shot scenarios, where models must operate with minimal or no training data, a common issue when dealing with less-documented languages such as Spanish. Existing solutions that combine contrastive learning with external sources, like the Unified Medical Language System (UMLS), have shown competitive results. However, most of these methods focus on individual concepts from the KBs, ignoring relationships such as synonymy or hierarchical links between concepts. In this paper, we propose leveraging these relationships to enrich the training triplets used for contrastive learning, improving performance in MEL tasks. Specifically, we fine-tune several BERT-based cross-encoders using enriched triplets on three clinical corpora in Spanish : DisTEMIST, MedProcNER, and SympTEMIST. Our approach addresses the complexity of real-world data, where unseen mentions and concepts are frequent. The results show a notable improvement in lower top-k accuracies, surpassing the state-of-the-art by up to 5.5 percentage points for unseen mentions and by up to 5.9 points for unseen concepts. This improvement reduces the number of candidate concepts required for cross-encoders, enabling more efficient semi-automatic annotation and decreasing human effort. Additionally, our findings underscore the importance of leveraging not only the concept-level information in KBs but also the relationships between those concepts.

1. Introduction

The adoption of digital solutions in healthcare, through electronic health records (EHR), represents one of the most significant paradigm shifts in the management of medical information. This transition from traditional paper to electronic records offers the opportunity to improve patient care, allowing for more personalized, adapted, and precise medical services. To leverage this transformation, it is essential to develop techniques that enable the efficient extraction and utilization of the available information [1,2] existing in these documents. In this context, since a significant part of the information contained in the EHRs is in text format and written in natural language, natural language processing (NLP) techniques emerge as the most popular solutions for clinical information extraction and management. Within the realm of NLP, two areas stand out significantly: named entity recognition (NER) and medical entity linking (MEL). While the former involves the recognition of clinical entities present in medical texts, the latter refers to the standardization of such entities by mapping them to normalized concepts present in specialized medical ontologies. However, these techniques face the inherent complexity of natural language and, more importantly, the scarcity of annotated data, which is even more pronounced when dealing with less-documented languages such as Spanish. Moreover, these challenges are compounded by linguistic phenomena such as synonymy—where the same information can be expressed in various ways—, acronymy—involving the abbreviation of terms—and polysemy—where a single word can have multiple meanings [3].

The traditional NER+MEL pipeline starts by recognizing the clinical entities (such as symptoms, diseases, findings, etc.) present in the analyzed texts, and then goes through a MEL stage in which these entities are normalized one by one through the use of controlled medical ontologies, in which a unique identifier/code is assigned to each of the detected entities. In the clinical context, meta-ontologies

[*] Corresponding author at: Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Málaga, Spain.
E-mail addresses: fgallegodonoso@uma.es (F. Gallego), franveredas@uma.es (F.J. Veredas).

https://doi.org/10.1016/j.knosys.2025.113211
Received 2 October 2024; Received in revised form 17 February 2025; Accepted 18 February 2025
Available online 1 March 2025

such as the Unified Medical Language System (UMLS[1]) or SNOMED-CT[2] are commonly used for this purpose. This normalization process not only enhances the detailed information extracted from these EHRs but also facilitates the development of semi-automatic and automatic annotation tools—thus reducing significantly the human effort required for these tasks—as well as the creation of in-domain knowledge graphs (KGs). The latter, KGs, are essential for a growing trend like retrieval-augmented generation (RAG) with generative large language models (LLMs), which is recently being used to minimize potential hallucinations produced by these models [4]. Moreover, these KGs obtained as a by-product of MEL tasks can be also useful for personalized medicine [5], allowing to discover and represent new drug–drug [6], drug–disease [7] or protein–protein interactions [8].

The advancement of neural networks and, especially, the Transformer architecture [9] has led to significant development in the MEL field. Recently, given the aforementioned scarcity of annotations, the information present in the enormous KBs or meta-ontologies of medical concepts such as UMLS [10], is being leveraged to improve the spatial representations of these concepts—known as concept embeddings—, by adapting transformer-based models through contrastive-learning strategies [11], thus achieving a more precise linking of medical entities found in clinical notes to standardized codes.

However, most of these developments have been carried out on English clinical texts, posing a great challenge for less-documented languages or for their application in multilingual contexts. In Spanish, there is recently a trend to promote the progress of these systems through evaluation campaigns such as DisTEMIST [12], MedProcNER [13], or SympTEMIST [14], which focus on recognizing diseases, procedures, and symptoms.

In our previous developments, we proposed the ClinLinker [15] pipeline focused on MEL in Spanish, adapting various language models to the clinical domain within a contrastive-learning strategy. For this purpose, in our previous work we followed an information-retrieval approach, according to the two characteristic stages of this approach: a first candidate-retrieval phase in which a model—usually a bi-encoder—is trained to propose a set of candidate concepts for the normalization of a given entity, followed by a second candidate-reranking phase in which another model—usually a cross-encoder—is trained to reorder the candidates proposed in the candidate-retrieval phase, thus obtaining a final list of candidate concepts sorted by the likelihood of being the correct candidate. Following this research line, we explored different strategies for enriching the candidate-retrieval phase models (bi-encoder models) with information from medical KGs, specifically by using the synonymy and hierarchy relationships existing between the concepts stored in these KGs [16]. This method achieved high performance compared to the state of the art (SOTA), and managed to shift the main challenge of MEL models to the candidate-reranking phase.

Leveraging the findings from this research work, in this article we have examined the impact of enriching this candidate-reranking phase with the information from KGs. Our approach highlights the richness of ontologies by exploiting both concept-level information and hierarchical relationships between concepts. Specifically, we have generated training triples that incorporate not only synonymous terms but also closely related concepts, thereby enhancing the accuracy of MEL in complex real-world scenarios, and providing clinicians with more reliable automatic and semi-automatic annotation. In contrast, current methods primarily use synonyms or semantically related concepts extracted from the text to generate these triples, without leveraging the similarities between closely related concepts within an ontology. These advancements represent a significant step forward—potentially adaptable to other languages—in the healthcare sector, reducing the substantial manual annotation effort currently required of clinicians. Code is available at: https://github.com/ICB-UMA/KnowledgeGraph.

2. Related work

The growing relevance of the MEL field has prompted the development of various approaches to address its major challenges. Additionally, being a critical sector, it is essential to promote these developments that maximize text comprehension and its subsequent leveraging. These approaches are highly diverse, ranging from more traditional techniques based on heuristics, machine learning, or the more recent deep learning-centered architectures. Initially, the developments consisted of direct matching, with text string-based algorithms, of mentions extracted through regular expressions [17], with entities from a KB that attempted to cover different possibilities [18,19]. However, despite not requiring high computational resources and being easily interpretable and explainable, these solutions exhibited insufficient performance as they were unable to leverage contextual information [20]. Since these approaches rely on structured resources, it is worth noting that the terms "knowledge base" (KB) and "ontology" are often used interchangeably to describe a structured database designed for organizing and retrieving knowledge. In contrast, the concept "KG" refers to a graph-based representation within a KB that encodes relationships between concepts [21,22]— note that some researchers conceptualize KBs as more closely aligned with KGs than with ontologies, due to their focus on graph-based structures and complex semantic relationships [23].

Expanding on these advances, given the high performance originally demonstrated by machine-learning-based systems, these rule-based or direct matching solutions were gradually replaced. For instance, LINDEN [24] enhanced entity linking by integrating taxonomic relations from YAGO [25] and semantic associations from Wikipedia. By constructing a semantic network and leveraging multiple disambiguation features, including link probability and global coherence, this framework demonstrated the benefits of incorporating structured knowledge to improve accuracy in tasks requiring complex semantic reasoning. Leaman and coworkers [26] presented DNorm, a machine-learning-based system that used TF–IDF representations to rerank *(mention, entity)* pairs by comparing each mention with different concepts from a KB. This solution represented a significant advancement but was very computationally expensive, especially when working with large ontologies. Building on the idea of semantic similarity, Zhu and Iglesias [27] developed SCSNED, which combines corpus-based and ontology-based methods to improve entity disambiguation. While effective for short texts, this approach does not take advantage of hierarchical relationships within knowledge graphs, limiting its applicability in more complex contexts such as medical entity linking. In contrast, our method leverages these relationships to enhance candidate reranking, providing significant benefits in scenarios that require handling extensive clinical documents and intricate contextual information. The rise of deep learning and models based on the Transformer architecture [28] caused a paradigm shift in MEL [29]. Thus, Guo and collaborators [30] proposed what they called *Walking Named Entity Disambiguation* (WNED), which, through random walks, generated dense graphs, allowing the exploitation of indirect relationships between mentions within the same document based on semantics.

The most recent advancements have applied a contrastive-learning strategy to align the vector representations of recognized entities from clinical texts with concepts stored in large KBs, thus leveraging available information about medical terms. Following this approach, Miftahutdinov and coworkers [31] proposed DILBERT, a bi-encoder based on the Transformer architecture which the authors fine-tuned by means of a strategy that used triplets composed of "positive" and "negative" concepts obtained from the UMLS. Meanwhile, Ledell and collaborators [32] proposed a two-phase pipeline consisting of a bi-encoder that retrieves candidates through the context of a mention and the

[1] Unified Medical Language System®(UMLS®): https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

[2] SNOMED-CT®: https://www.nlm.nih.gov/healthit/snomedct/index.html

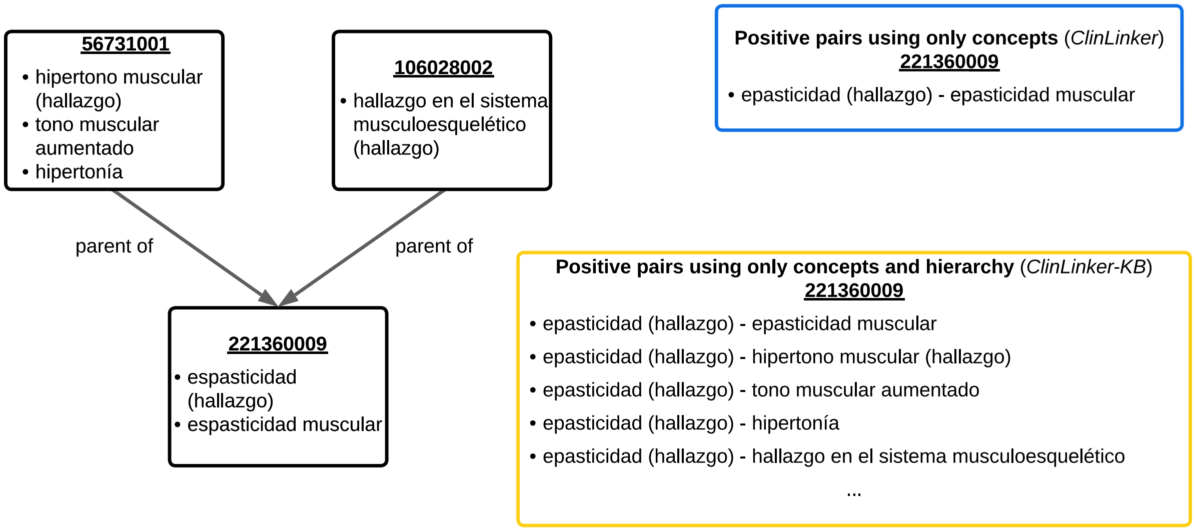

56731001
• hipertono muscular (hallazgo)
• tono muscular aumentado
• hipertonía
106028002
• hallazgo en el sistema musculoesquelético (hallazgo)
parent of
parent of
221360009
• espasticidad (hallazgo)
• espasticidad muscular
Positive pairs using only concepts (ClinLinker)
221360009
• epasticidad (hallazgo) - epasticidad muscular
Positive pairs using only concepts and hierarchy (ClinLinker-KB)
221360009
• epasticidad (hallazgo) - epasticidad muscular
• epasticidad (hallazgo) - hipertono muscular (hallazgo)
• epasticidad (hallazgo) - tono muscular aumentado
• epasticidad (hallazgo) - hipertonía
• epasticidad (hallazgo) - hallazgo en el sistema musculoesquelético
...

Fig. 1. Leveraging hierarchical and semantic relationships in knowledge bases enhances performance in medical entity linking.

descriptions of entities, and a cross-encoder that reranks the candidates given by the bi-encoder. The use of contextual information allows the model to adapt to the domain, although it does not enrich its knowledge with a source of medical information. In [33], Sheng and collaborators presented KRISSBERT, which utilized the UMLS ontology to generate self-supervised mention examples from unlabeled texts such as PubMed and trained a contextual encoder using contrastive-learning to improve the accuracy in linking biomedical entities. Another highly influential contribution is the work of Liu and coworkers [11], in which they adapted a BERT-based model using contrastive-learning and the different descriptions of each concept, leading to the development of SapBERT. On the other hand, Tutubalina [34] exploited the Transformer architecture to capture contextual relationships between these mentions.

Moreover, these latest developments have mainly focused on retrieving a set of medical concept candidates for the normalization of a given recognized entity on the basis of on the similarity between the entity and concept embeddings [35]. However, because text annotations are scarce in the clinical domain and the range of possible concepts for standardization is extremely large and varied, candidate-retrieval systems usually fail to accurately distinguish between the right candidate and just the “close” ones. For example, the term “lesion of lung” is closely related to the term “benign neoplasm of bronchus of left lower lobe”, the former being higher in the SNOMED-CT hierarchy than the latter. Although their vector representations (embeddings) could be very similar—since they share the location of the disorder—, the latter contains more specific and detailed information and could be a more precise concept for the right normalization of the entity “benign lesion of bronchus of left lung”. Recently, solutions have emerged that focus on maximizing the candidate-reranking phase [36]. Thus, Borchert and collaborators [37], in their xMEN framework, proposed a pipeline composed of candidate-retrieval and subsequent re-ranking of concepts by using cross-encoder-based models, which is an approach similar to the one presented in our work ClinLinker [15]. However, only the concepts and their different descriptions—synonyms—are used. The hierarchical information of the ontology, which is a very valuable resource in the face of the scarcity of annotations, is not exploited. Continuing this line of research, we further developed ClinLinker-KB [16], where we explored enriching the candidate retrieval phase within the bi-encoder by generating new positive pairs based on the relationships present in the UMLS graph (see Fig. 1).

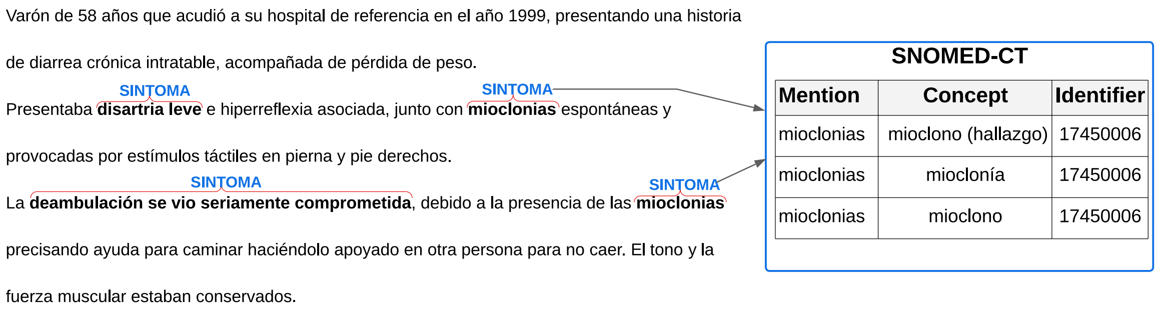
On the other hand, the revolution driven by LLMs in numerous sectors has led many solutions to be based on them [38]. Recently, several studies have been published that aim to address entity linking tasks using this type of models [39] for zero-shot or few-shot scenarios [40], exploiting information from KBs. However, they still suffer from hallucinations, and there is no clear strategy for prompt formulation. Borchert and collaborators [41] have successfully integrated these LLMs into their entity linking pipeline to accurately retrieve candidates, achieving significant results in this task. Although all these developments are fundamental to our work, none exploit the information present in ontology relationships, as they leverage the concepts and definitions of entities but not the relationships between them.

Despite significant advancements in MEL, current approaches often overlook the potential of exploiting hierarchical and semantic relationships within KB beyond simple concept embeddings. While recent studies have improved candidate retrieval and reranking phases, the integration of structured relationships such as parent–child links remains underexplored, particularly in complex, multilingual, or low-resource settings. This gap underscores the need for methodologies that fully leverage the richness of these ontologies to enhance performance in challenging real-world scenarios.

In this paper, we showcase how leveraging relationships within KGs can enrich the candidate re-ranking phase through the generation of new pairs. This approach is particularly beneficial in complex, real-world scenarios, where the hierarchical and semantic structure of ontologies provides crucial contextual information for more accurate in MEL.

3. Material and methods

Entity linking in the medical domain is a complex task where solutions proposing a single concept for entity normalization still lack enough confidence. On the contrary, a set of potential concept candidates is usually considered in the process of standardization of each entity recognized in a clinical text. In this study, we work with health records from the main available corpora for MEL in Spanish—supplied in connection with the shared tasks DisTEMIST [12], MedProcNER [13], and SympTEMIST [14]—which are detailed in the following sections. The methodology described in this section shows significant progress for the efficient re-ranking of close and intricate candidates for accurate entity normalization and, consequently, it has a remarkable impact on the field of MEL, through the exploitation of the richness of the relationships between medical concepts stored in well-known in-domain KBs such as UMLS or SNOMED-CT.



Varón de 58 años que acudió a su hospital de referencia en el año 1999, presentando una historia de diarrea crónica intratable, acompañada de pérdida de peso. Presentaba **disartria leve** e hiperreflexia asociada, junto con **mioclonias** espontáneas y provocadas por estímulos táctiles en pierna y pie derechos. La **deambulación se vio seriamente comprometida**, debido a la presencia de las **mioclonias** precisando ayuda para caminar haciéndolo apoyado en otra persona para no caer. El tono y la fuerza muscular estaban conservados.

SNOMED-CT

Mention	Concept	Identifier
mioclonias	mioclono (hallazgo)	17450006
mioclonias	mioclonía	17450006
mioclonias	mioclono	17450006

Fig. 2. Recognition and linking of medical entities from a clinical record using SNOMED-CT. This example shows a fragment from the SympTEMIST dataset, where entities are first identified within the clinical text and then linked to standardized SNOMED-CT concepts with unique identifier.

3.1. Description of the corpora

To rigorously evaluating the performance of the models presented in this article, we analyzed their behavior on the three most representative Spanish language MEL corpora. These three shared tasks involved the manual annotation of 1,000 clinical cases, each focusing on different types of entities documented. The first one, *DISease TExt Mining Shared Task*[3] [12]—DisTEMIST—, was a shared task created for the track of BioASQ in CLEF 2022. This task focused on the automatic recognition and subsequent linking of disease mentions present in Spanish medical records to SNOMED-CT codes. In contrast, *MEDical PROCedure Named Entity Recognition*—MedProcNER[4] [13]—, was a shared task focused on the detection, normalization, and indexing mentions of clinical procedures in Spanish. Finally, *SYMPtoms, signs and findings TExt MIning Shared Task*[5]—SympTEMIST [14]—was designed to recognize and normalize entities such as symptoms, signs, and findings (see Fig. 2 for an example of entity recognition and normalization on a text fragment extracted from SympTEMIST corpus).

Additionally, in order to approximate the performance that would be achieved with real-world data (RWD)—where the mentions in the texts do not perfectly match the text of the entities in the KBs or do not exist in the training set—we did not evaluate the models on the gold standard (GS) version of the corpora, as it includes a significant number of mentions that exhibit an exact textual match with concepts in the KB, a condition that rarely occurs in real-world applications—composed of 2,599, 3,618, and 3,104 annotations respectively. Instead, we work with two subsets: unseen mentions (UM) and unseen codes (UC). UM refers to mentions in the test sets that do not have exact textual matches in the training sets or the gazetteers provided by the shared task organizers, representing cases where the model encounters novel surface forms that require generalization beyond previously seen examples. UC, on the other hand, corresponds to concepts present in the test sets but entirely absent from the training sets, meaning the model must infer mappings to new concepts not encountered during training. The absence of a direct match between text mentions and entities in the KB makes the linking process more complex, decreasing the performance of systems based on string matching and for contrastive-learning approaches that rely on similarity. It is important to note that for all three tasks, we used the gazetteer—a subset of the SNOMED-CT KB that includes all possible codes required for normalization—provided by the shared-task's organizers. As can be seen in Table 1, the UM set contains 1375, 1730, and 1573, while the UC set consists of 1115, 878, and 763 mentions from the test set, for DisTEMIST, MedProcNER and SympTEMIST corpora, respectively.

Table 1
Distribution of the datasets used in this study, highlighting the number of annotated instances for two subsets of the test data.

Corpora	Unseen mentions	Unseen codes
DisTEMIST	1,375	1,115
MedProcNER	1,730	878
SympTEMIST	1,573	763

3.2. Candidate re-ranking using cross-encoders for enhanced entity disambiguation

Given the limited number of annotations that are usually available for MEL tasks, it is crucial to explore alternative information sources that can enhance learning for MEL models [42]. One effective strategy is to leverage large ontologies, such as UMLS—comprising over 4.5 million medical concepts and more than 4.4 million relationships—or SNOMED-CT. These KBs offer not only relevant medical terminology but also synonyms and alternative descriptions, thus allowing to shift MEL tasks from zero-shot to stratified-shot learning problems. Zero-shot learning involves making predictions for tasks or classes that the model has never seen during training. In contrast, few-shot learning, in general, trains models on a small number of labeled examples, enabling them to generalize effectively to new, unseen data. Within this paradigm, stratified-shot learning, a specialized subset of few-shot learning, is represented in our case by unseen mentions, which are carefully designed to reflect real-world scenarios where no string matching occurs between mentions and concepts.

Our previous works [15,16,43] demonstrated the benefits of incorporating information from ontologies to the bi-encoders used for candidate-retrieval in MEL tasks, showing initial improvements with the addition of synonym-based descriptions and further gains with the inclusion of hierarchical relationships, such as "parent–child" relationships between medical concepts within the ontology. The findings of those studies show that the primary challenge of candidate-retrieval has now shifted to the phase of candidate-reranking, since achieving high accuracy rates was only possible when evaluating a large set of candidates (specifically, with top-200 accuracy).

To date, most systems have approached MEL tasks through a single candidate retrieval phase, selecting a set of candidates based on traditional classification or similarity-based techniques. However, these classical approaches tackle only one stage of the complete MEL pipeline, which ultimately requires reranking the retrieved concept candidates to sort them by their probability of being the correct term for entity normalization. In this study, we focus on examining the impact of integrating KG information derived from an ontology like UMLS in this candidate-reranking phase. Thus, we have trained various cross-encoder models to optimize them for dealing accurately with the task of concept reranking. These cross-encoders, based on the Transformer architecture—specifically on BERT-based LLMs—, are designed to calculate similarity scores between pairs of sentences—in this context,

[3] DISease TExt Mining Shared Task®(DisTEMIST®): https://zenodo.org/records/7614764

[4] MEDical PROCedure Named Entity Recognition®(MedProcNER®): https://zenodo.org/records/8224056

[5] Symptoms TExt MIning Shared Task®(SympTEMIST®): https://zenodo.org/records/10635215

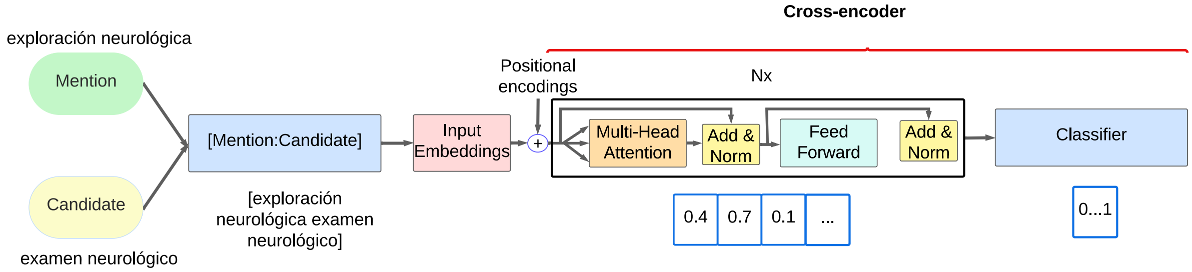


Fig. 3. Pairwise scoring of *(mention, candidate)* combinations using a BERT-based cross-encoder model to compute semantic similarity for entity linking tasks. *Note: The candidates were initially retrieved using a bi-encoder model during the candidate-retrieval phase.*

(mention, candidate) pairs. Training these models involves generating training triplets—discussed in detail in Section 3.3—consisting of these pairs and an associated label: 0 for a "negative" candidate concept and 1 for a "positive" candidate. For each input pair—composed of two terms, a mention and a candidate concept for normalization—, the cross-encoder concatenates both terms and generates a joint spatial vector representation. This vector is then processed by a classification layer, which calculates a score between 0 and 1, reflecting the similarity between the two concepts (see Fig. 3). Each one of the cross-encoders analyzed in this study is initialized with the weights of a bi-encoder—that has been previously trained by fine-tuning a BERT-based model for candidate-retrieval, using contrastive-learning and the UMLS hierarchy (see [16] for a detailed description of these bi-encoders)—and then trained with that set of training triplets.

The objective during the training of these cross-encoders for candidate-reranking is to accurately distinguish between positive (i.e., "ground-truth" concepts for correct entity normalization) and negative candidates (i.e., wrong concept candidates but semantically close to the "ground-truth" concept) in relation to the mentions from the text. It is important to note again that these cross-encoders have previously been initialized with the weights from a trained bi-encoder used during the candidate-retrieval phase. We use binary cross-entropy as the loss function to train the cross-encoders, that way tuning the weights of the network to adjust the vector representation of each training triplet to approximate the similarity function that enables the precise reranking of the candidates retrieved by the bi-encoders.

Despite these advancements, the problem remains complex. While bi-encoders used in the candidate-retrieval phase compare diverse entities from an ontology, in the reranking phase, candidate concepts are even more similar to each other since they have already been retrieved as the most relevant by the bi-encoder, having closer vector representations—embeddings. We refer to these candidates as "hard negatives" due to their similarity with the right concept (i.e., "positive" concept) for entity normalization. An example of these "hard negatives", are "disnea", "disnea espiratoria", "disnea, clase IV", "disnea postprandial" o "disnea crónica", where all are very similar terms, but each corresponds to a different concept with its own unique identifier. This similarity makes the construction of effective triplets for training the models a critical aspect of the reranking process.

3.3. Triplets definition through candidate enrichment using knowledge graph hierarchy

Defining triplets to distinguish between positive and negative candidates is one of the most important steps in training cross-encoders. In our preliminary studies, we experimented with generating random candidates as negative samples to train cross-encoders for candidate-reranking in several MEL tasks, which resulted in reduced performance. Subsequently, for this study we examined the impact of using "hard negatives" to compose the triplets used to train the cross-encoders with contrastive-learning. This latter approach has enhanced the performance of the candidate-reranking phase, thus contributing to improve the efficacy, yielding higher top-k-accuracy values with lower values of k.

These findings, combined with the wealth of information available in ontologies like UMLS and SNOMED-CT, highlight the need to incorporate more refined data into the triplets to enable the model to effectively distinguish between these candidates with close representations. To address this, we define two strategies of triplets definition using hierarchy information from the KG. We name these strategies *knowledge graph ("kg")* and *bidirectional knowledge graph ("bkg")*—in addition to "sim", which is the "classical" strategy consisting of a list of candidates supplied by a bi-encoder that uses a similarity loss-function with threshold parameter t. The "kg" strategy uses as negative samples those candidates directly related with each retrieved negative candidate through a "parenthood" relationship, ensuring that none of the new candidates are the correct one. This approach leverages the hierarchical structure of the knowledge graph to introduce more challenging negatives, thereby enhancing the model's ability to distinguish between the positive term and hard negatives. On the other hand, the "bkg" strategy includes not only the parent concepts of the negative candidates but also their child concepts as additional negative samples. These additional samples are generated by traversing the hierarchical relationships in both directions—parent-to-child and child-to-parent—while ensuring that none of these newly added candidates corresponds to the correct concept. Additionally, we introduce a parameter called "depth", which sets the level of exploration within the hierarchy of the knowledge graph. To implement this, we created triplets containing these pairs—*(mention, candidate)*—along with a label: 0 for a negative candidate and 1 for a positive candidate.

For triplet creation, we considered each mention $m \in M$ available in the training sets provided by the organizers of the shared tasks, obtaining list of candidates C_m for each of these mentions using a bi-encoder, in our case, 200 candidates. After obtaining C_m, we examined the corresponding mention m to see if the concept unique identifier (CUI) associated with it matched any candidate CUI for the input mention. If a match was found, it was labeled as a positive term—i.e., the triplet $(m, c, 1)$ is added to the training set—, while the others were marked as negative samples—i.e., the triplet $(m, c, 0)$ is included in the training set. Finally, for the triplet strategy "kg"' or "bkg", the ancestors, or both the ancestors and descendants in the KG hierarchy, respectively, of these candidates are obtained up to the specified "depth" (see Algorithm 1). This is a procedure similar to that used in ClinLinker-KB [16], but instead of enriching the triplets with positive pairs, in this case, we use them to create negative pairs, providing the cross-encoder with more examples to distinguish between the correct candidate and those that are closely related—often, these closely related candidates are the ancestors of the correct concept, due to their high semantic similarity. It is important to highlight that the set of triplets does not contain duplicate triplets, and in case the correct concept is accessed through ancestor or descendant relationships, it will never be added to the training set as a negative sample, thus avoiding any inconsistency. For example, for a mention linked to a concept such as "Degeneration of acoustic nerve (disorder)", one of the candidates might be “Degeneration of adrenal gland (disorder)”, which is similar but incorrect. Using the ‘kg’ strategy, negative candidates like “Degenerative disorder (disorder)” and “Disorder of adrenal gland (disorder)” would be added as ancestors. In contrast, with the ‘bkg’ strategy, in addition to these two negatives as ancestor nodes, “Adrenal calcification (disorder)” and “Atrophy of adrenal cortex (disorder)” would be included as descendant samples with a depth of 1.

Algorithm 1 Triplet Creation Process

```
function CONCEPT($terms$)
    return List of unique concept identifiers for $terms$
end function
Require: List of mentions $M$, similarity threshold $t$, strategy $str$, depth $d$
Ensure: Set of triplets $T$
$T \leftarrow \emptyset$
for each mention $m$ in $M$ do
    $C_m \leftarrow$ candidates for mention $m$
    $C_m^{conc} \leftarrow$ CONCEPT($C_m$)
    $m^{conc} \leftarrow$ CONCEPT($\{m\}$)
    if $m^{conc} \notin C_m^{conc}$ then
        continue
    end if
    for each candidate $c_i$ in $C_m$ do
        if $str = \text{sim}$ then
            if $\text{sim}(m, c_i) > t$ then
                if $m^{conc} = C_m^{conc}[i]$ then
                    $T \leftarrow T \cup \{(m, c_i, 1)\}$
                else
                    $T \leftarrow T \cup \{(m, c_i, 0)\}$
                end if
            end if
        else if $str = \text{kg}$ or $str = \text{bkg}$ then
            if $m^{conc} = C_m^{conc}[i]$ then
                $T \leftarrow T \cup \{(m, c_i, 1)\}$
            else
                $T \leftarrow T \cup \{(m, c_i, 0)\}$
            end if
            Retrieve ancestors $A_c$ up to depth $d$
            if $str = \text{bkg}$ then
                Retrieve descendants $D_c$ up to depth $d$
            end if
            for each $a$ in $A_c$ or $D_c$ do
                if $m^{conc} \neq$ CONCEPT($a$) then
                    $T \leftarrow T \cup \{(m, a, 0)\}$
                end if
            end for
        end if
    end for
end for
return $T$
```

We have explored depths 1 and 2. It is important to mention that training cross-encoder models require considerable computational resources, and the amount of ontology-based information (and, as a consequence, the number of triplets used to train the model) increases significantly as the analyzed depth grows, a difference that becomes even more pronounced if the triplet strategy is bidirectional. This can be observed in Table 2, where the number of training triplets from the DisTEMIST corpus increases from more than 400K when using the “sim” strategy, to more than 5M and 7M unique triplets when the graph is explored unidirectionally with depth 1 and 2, respectively, and up to more than 120M triplets when working bidirectionally and depth 2. For the other two corpora analyzed herein, a similar behavior can be observed in Table 2.

This strategy has demonstrated to have clear performance improvement over the SOTA, using either the unidirectional or the bidirectional approaches. We will provide a detailed explanation of these improvements in the following sections, underscoring their practical utility, and uncovering the scenarios in which these strategies could be applicable.

4. Results and discussion

For a comprehensive evaluation of the methodology and models for MEL presented in this paper, we have analyzed their performance on three leading Spanish language medical corpora: DisTEMIST, MedProcNER, and SympTEMIST, focussing in the more challenging scenarios, targeting stratified-shot and zero-shot settings.

We have used top-k accuracy as our evaluation metric, which determines whether the correct candidate CUI is among the top “k” predicted values. We have selected values of “k” at 1, 5, 25, and 50 to assess the model’s performance across different scenarios, ranging from the most challenging—automatic annotation with a single candidate—to cases involving multiple candidates, demonstrating the potential of our approach. It is important to note that fully automatic clinical coding remains an elusive goal, mainly due to the complexity of mentions, the high similarity between potential candidates, the inherent difficulties of the medical domain, and the scarcity of annotated data.

To evaluate the effectiveness of the different triplet generation strategies analyzed herein and the corresponding efficacy of the cross-encoders for candidate re-ranking trained with these triplet sets, we have employed two different bi-encoders for the candidate-retrieval phase: the first one, ClinLinker-KB [16], was developed as a consequence of our previous work and was trained on the entire UMLS dataset in Spanish, utilizing KGs to exploit parent–child relationships between concepts; while the second, SapBERT-XLM-R-large [11], is a multilingual model known for its strong performance in similar tasks, positioning it as one of the leading solutions for entity linking. Additionally, for comparison with a base model, we have also analyzed the efficacy of the Personalized PageRank (PPR) [44–46] algorithm for candidate re-ranking, starting from the candidate list retrieved by ClinLinker-KB [16]. The PPR algorithm assigns a score to each candidate in a graph, which contains candidates for every entity mention in a given document. The score is determined by the connections between each candidate in the graph. In this way, PPR establishes the coherence of each candidate within the graph. The highest-scoring candidate for each entity mention is then associated with that mention. However, this approach is not designed to precisely distinguish between nearly identical concepts—i.e., hard negatives—especially in ontologies with a large variety of concepts and numerous relationships between them, as in the case of UMLS.

Furthermore, we have included TEMUNormalizer,[6] a baseline system developed by the task organizers, to provide a benchmark for our proposed approaches. TEMUNormalizer employs a sequential procedure for entity normalization: it first applies string matching to identify exact matches between entity mentions and concepts in the knowledge base. If no exact match is found, it proceeds with fuzzy matching to capture approximate matches based on lexical similarity. Finally, it applies sentence similarity techniques, using as base model the same used for training in ClinLinker-KB—*PlanTL-GOB-ES/roberta-base-biomedical-clinical-es*. This multi-step approach ensures a robust baseline performance, enabling a meaningful comparison with the methods proposed in this study.

It is important to note that the results presented in this paper are not directly comparable to those obtained in [15,16], as the versions of the datasets used are not the same. In this case, the training sets have been expanded, which has led to a decrease in the number of unseen mentions and unseen codes, making the evaluation scenarios even more challenging. For this reason, we have reevaluated ClinLinker-KB using

[6] TEMUNormalizer: https://github.com/TeMU-BSC/TEMUNormalizer

Table 2
Size of triplet sets used for the cross-encoders training, based on different generation strategies across multiple corpora. *Note: Candidates were generated using the version of the bi-encoder ClinLinker-KB [16] that performed best for each corpus.*

	Strategy				
Corpus	sim	kg		bkg	
		1	2	1	2
DisTEMIST	426,646	5, 484, 631	7, 645, 351	16, 131, 267	121, 865, 655
MedProcNER	514,424	7, 252, 631	9, 321, 604	24, 825, 237	139, 258, 253
SympTEMIST	986,499	12, 063, 769	15, 989, 653	20, 892, 503	181, 462, 527

Table 3
Performance (top-k accuracy) of cross-encoders with different triplet-generation strategies (`str`) and KG exploration `depth` for MEL on the DisTEMIST corpus, based on different bi-encoders for candidate-retrieval. The accuracy of the bi-encoders alone without candidate re-ranking (*bi* strategy) is also shown, along with the performance of the PPR re-ranking algorithm (base model) when using ClinLinker-KB for candidate-retrieval. *Note: The best model for each evaluation set—UM or UC—is highlighted in* ***bold****, and the second best is underlined.*

			top-k accuracy			
Candidate-retrieval	str	depth	1	5	25	50
			Unseen Mentions (UM)			
ClinLinker-KB	bi	–	*.352*	*.558*	*.714*	*.768*
	sim	–	**.389**	.581	**.753**	**.807**
	kg	1	.368	.588	.739	.784
	bkg	1	.371	**.592**	.732	.786
	kg	2	.364	.590	.737	.783
	bkg	2	.124	.276	.527	.668
SapBERT-XLM-R-large	bi	–	*.357*	*.529*	*.643*	*.713*
	sim	–	.374	.583	.689	.721
	kg	1	.284	.449	.582	.634
	bkg	1	.088	.128	.342	.444
	kg	2	.051	.145	.332	.454
	bkg	2	.358	.529	.644	.680
ClinLinker-KB + PPR	–	–	.007	.041	.201	.405
TEMUNormalizer	–	–	.001	.020	.108	.188
			Unseen Concepts (UC)			
ClinLinker-KB	bi	–	*.406*	*.570*	*.709*	*.760*
	sim	–	**.423**	.602	**.745**	**.802**
	kg	1	.413	.609	.738	.781
	bkg	1	.417	.605	.736	.782
	kg	2	.413	**.611**	.739	.783
	bkg	2	.120	.259	.502	.656
SapBERT-XLM-R-large	bi	–	*.404*	*.561*	*.651*	*.683*
	sim	–	.408	.591	.691	.719
	kg	1	.346	.488	.594	.641
	bkg	1	.115	.146	.356	.457
	kg	2	.068	.172	.351	.461
	bkg	2	.404	.561	.651	.683
ClinLinker-KB + PPR	–	–	.004	.024	.163	.320
TEMUNormalizer	–	–	.005	.021	.109	.234

the latest version to provide an accurate comparison under these new conditions.

The results on the DisTEMIST corpus demonstrate the strong performance of our ClinLinker-KB-based cross-encoders with respect to the other two analyzed models (see Table 3). Starting with this bi-encoder, which was specifically adapted to the clinical domain, leads to superior performance in most triplet generation strategies analyzed, compared to the cross-encoders fed with SapBERT-XLM-R-large's retrieved candidates. Additionally, we observe that the graph-based triplet generation strategy provides minimal improvement, except when working with top-5 candidates, where the highest values are achieved by incorporating KG information on both the UM (0.592) and UC (0.611) datasets. In contrast, the performance of the cross-encoders based on SapBERT-XLM-R-large decreases when hierarchical ontology information is included in the triplet generation process. This may be because the model is trained for multiple languages, and such adaptation may negatively impact its performance. Finally, the challenging nature of the task is evident when comparing the results of cross-encoder-based strategies to PPR or TEMUNormalizer, which only achieves top-1 accuracy scores of 0.007 and 0.004 or 0.001 and 0.005 on UM and UC datasets, respectively. In these scenarios, when the number of candidates for each mention is high, PPR has reduced its performance. It should also be noted that increasing the "depth" parameter does not necessarily lead to better results and can even drastically degrade performance when both directions of the graph are explored at a depth of 2. The large number of triplets—121, 865, 655 samples—appears to cause the model to overfit (see Table 3).

Table 4, which presents the performance of the different models on the MedProcNER corpus, shows a similar trend, with ClinLinker-KB-based cross-encoders dominating. However, in this case, the use of graph information outperforms the "sim" candidate generation strategy for the lower top-k accuracies—k = 1, 5, 25—achieving values of 0.446 in top-1 for UM and 0.404 for UC. Once again, PPR and TEMUNormalizer demonstrate significantly lower performance compared to cross-encoder-based approaches.

Finally, for the SympTEMIST corpus the results observed in MedProcNER are confirmed. As with the DisTEMIST and MedProcNER corpora, cross-encoders based on the ClinLinker-KB bi-encoder for candidate-retrieval dominate, showing a significant gap compared to SapBERT-XLM-R-large-based approaches and even more so with PPR or TEMUNormalizer, which both struggle to handle tasks with such similar candidates. In this case, the graph-based strategy demonstrates consistent and robust performance, outperforming the similarity-based approach across all top-k accuracies on both the UM and UC test sets, achieving improvements such as 0.446 versus 0.387 in top-1 accuracy and 0.690 versus 0.635 in top-5 accuracy. As with the other corpora, the use of SapBERT-XLM-R-large as the bi-encoder leads to a considerable drop in performance, although it still remains superior to PPR in all cases (see Table 5).

The results obtained highlight both the importance of leveraging hierarchical information from medical ontologies and the level of detail in task definition, achieving significant improvements in more recent tasks such as MedProcNER and SympTEMIST, with a more refined gazetteer and greater variability of concepts—this can be observed in the size differences shown in the UC and UM sets for each corpus in Table 2—, although in the latter case, the bi-encoder already provided excellent performance, achieving over 96% top-k accuracy for high values of k. It can also be noted that the inclusion of more information or a higher level of detail in the triplet generation process for multilingual models does not yield the desired behavior, significantly reducing their performance. This is particularly evident when increasing the "depth" parameter in the KG-based triplet generation strategy. Although the addition of the first negative candidates with low depth enhances model performance by improving the distinction between correct and incorrect candidates, a substantial increase in depth causes a significant imbalance, ultimately reducing model performance due to the overabundance of hard negatives. Finally, the values obtained

Table 4
Performance (top-*k* accuracy) of cross-encoders with different triplet-generation strategies (`str`) and KG exploration `depth` for MEL on the MedProcNER corpus, based on different bi-encoders for candidate-retrieval. The accuracy of the bi-encoders alone without candidate re-ranking (*bi* strategy) is also shown, along with the performance of the PPR re-ranking algorithm (base model) when using ClinLinker-KB for candidate-retrieval. *Note: The best model for each evaluation set—UM or UC—is highlighted in* ***bold****, and the second best is underlined.*

			top-*k* accuracy			
Candidate-retrieval	str	depth	1	5	25	50
Unseen Mentions (UM)						
ClinLinker-KB	bi	–	*.386*	*.610*	*.753*	*.809*
	sim	–	.413	.658	.808	.849
	kg	1	**.446**	.676	**.817**	.850
	bkg	1	**.446**	.676	.814	**.853**
	kg	2	.434	**.677**	.808	.845
	bkg	2	.328	.530	.672	.727
SapBERT-XLM-R-large	bi	–	*.400*	*.583*	*.706*	*.757*
	sim	–	.397	.622	.747	.798
	kg	1	.205	.296	.381	.431
	bkg	1	.113	.236	.407	.519
	kg	2	.312	.539	.698	.738
	bkg	2	.399	.582	.705	.756
ClinLinker-KB + PPR	–	–	.008	.040	.237	.443
TEMUNormalizer	–	–	.002	.017	.091	.173
Unseen Concepts (UC)						
ClinLinker-KB	bi	–	*.317*	*.513*	*.678*	*.738*
	sim	–	.359	.590	.767	**.811**
	kg	1	.395	**.617**	**.769**	.806
	bkg	1	**.404**	.607	.757	.800
	kg	2	.383	.608	.752	.792
	bkg	2	.251	.438	.607	.663
SapBERT-XLM-R-large	bi	–	*.315*	*.513*	*.650*	*.705*
	sim	–	.325	.558	.692	.755
	kg	1	.175	.239	.311	.347
	bkg	1	.116	.223	.384	.497
	kg	2	.249	.459	.638	.688
	bkg	2	.313	.510	.649	.703
ClinLinker-KB + PPR	–	–	.005	.034	.166	.336
TEMUNormalizer	–	–	.002	.024	.120	.227

Table 5
Performance (top-*k* accuracy) of cross-encoders with different triplet-generation strategies (`str`) and KG exploration `depth` for MEL on the SympTEMIST corpus, based on different bi-encoders for candidate-retrieval. The accuracy of the bi-encoders alone without candidate re-ranking (*bi* strategy) is also shown, along with the performance of the PPR re-ranking algorithm (base model) when using ClinLinker-KB for candidate-retrieval. *Note: The best model for each evaluation set—UM or UC—is highlighted in* ***bold****, and the second best is underlined.*

			top-*k* accuracy			
Candidate-retrieval	str	depth	1	5	25	50
Unseen Mentions (UM)						
ClinLinker-KB	bi	–	*.435*	*.708*	*.855*	*.895*
	sim	–	.387	.635	.847	.902
	kg	1	.433	.690	**.865**	**.915**
	bkg	1	**.446**	.676	.814	.853
	kg	2	.433	.690	**.865**	**.915**
	bkg	2	.156	.350	.572	.702
SapBERT-XLM-R-large	bi	–	*.390*	*.572*	*.700*	*.744*
	sim	–	.378	.589	.744	.787
	kg	1	.181	.357	.554	.655
	bkg	1	.238	.417	.592	.663
	kg	2	.181	.357	.554	.655
	bkg	2	.057	.181	.437	.587
ClinLinker-KB + PPR	–	–	.008	.036	.193	.380
TEMUNormalizer	–	–	.003	.019	.089	.198
Unseen Concepts (UC)						
ClinLinker-KB	bi	–	*.317*	*.591*	*.790*	*.837*
	sim	–	.315	.533	.790	.856
	kg	1	**.345**	**.592**	**.802**	**.870**
	bkg	1	**.345**	**.592**	**.802**	**.870**
	kg	2	.337	556	.765	.837
	bkg	2	.151	.321	.524	.658
SapBERT-XLM-R-large	bi	–	*.312*	*.460*	*.596*	*.649*
	sim	–	.303	.491	.651	.700
	kg	1	.148	.284	.474	.577
	bkg	1	.193	.334	.509	.579
	kg	2	.148	.284	.474	.577
	bkg	2	.045	.169	.376	.511
ClinLinker-KB + PPR	–	–	.005	.024	.149	.318
TEMUNormalizer	–	–	.007	.024	.093	.201

for the different top-*k* accuracy metrics of the PPR algorithm and TEMUNormalizer emphasize the complexity of the task, with nearly identical terms and very few examples—zero-shot or stratified-shot settings.

5. Conclusions

In this work, we have introduced a novel approach for MEL, by enriching BERT-based cross-encoders through the hierarchical structure of a large medical ontology like UMLS, applied to the three most representative MEL corpora existing in Spanish. The results show a clear improvement in the performance of the analyzed cross-encoders during the candidate re-ranking phase that follows a candidate-retrieval stage provided by an in-domain BERT-based bi-encoder pre-trained on a large corpus in Spanish (ClinLinker-KB). We have proposed strategies for the generation of training triplets on the basis of the hierarchical relationships between the medical concepts used for the normalization of the medical entities present in the clinical texts. When working with a first level of depth in the extraction of these training triplets from the KG, we are able to increase the number of training triplets by a factor of ten, which positively impacts the efficacy of the resulting MEL models. However, this improvement is not observed when working with multilingual candidate-retrieval bi-encoders, such as SapBERT, or when using a greater level of depth for the extraction of training triples. In these last cases, the performance of the resulting cross-encoders for MEL can be significantly reduced.

In general, the efficacy in MEL tasks in Spanish of these KG-enriched models surpasses those that set the current SOTA in the most complex stratified-shot and zero-shot scenarios analyzed, i.e., those that focus on subsets composed of mentions and codes not previously seen during training, which could be considered as the closest case to what can be found when working with real-world medical data. The results presented herein demonstrate the robustness of our MEL models across several medical corpora in Spanish, which are capable of coping effectively with the high-complexity task of entity normalization by precisely re-ranking sets of candidate concepts exhibiting high similarity among them. The complexity of this task is even greater if we take into account that we have very few (or no examples) of normalization of test set entities and concepts during model training. On the other hand, the enrichment of these MEL models with information from in-domain KGs has allowed for a reduction in the number of concept candidates that the cross-encoders need to supply to ensure sufficient confidence for entity normalization—thus achieving a top-25 accuracy of around 75% across all corpora, and in some cases, even higher. This could allow for the transformation of what is usually a manual annotation process into a semi-automatic labeling task. In this way, it would only be necessary to start from a reduced set of candidate concepts for standardization provided by our MEL model, instead of having to consider the thousands of possible concepts that appear in an ontology such as UMLS and that could be considered, a priori, possible candidates for the standardization of an entity detected in a clinical text. It is important to notice that this methodology is also applicable to any other language or research area for which this kind of KBs is available. Based on the findings, we recommend using triplets with unidirectional hierarchy and low depth values, as increasing the depth significantly raises computational costs and introduces a considerable imbalance in the number of negative samples.

For future work, after exploring the enrichment of the candidate retrieval and candidate re-ranking phases through ontologies, our main goal is to effectively address the task of MEL using generative LLMs, with particular attention to integrating knowledge from ontologies both during fine-tuning and inference stages. This aims to improve the response of these models and reduce the number of hallucinations they suffer in domains that deviate from their training data. We will focus on works such as [47,48], which employ RAG for biomedical applications. These approaches would enhance the performance of generative LLMs, providing robustness and reliability—critical aspects within the medical domain. Moreover, the inclusion of information from these ontologies when applying RAG could result in more effective, precise, and disambiguated information for tasks such as medical entity linking.

CRediT authorship contribution statement

Fernando Gallego: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Pedro Ruas:** Writing – review & editing, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Francisco M. Couto:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Francisco J. Veredas:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the support from the Ministerio de Ciencia e Innovación (MICINN), Spain under projects AEI/10.13039/501100011033 and PID2020-119266RA-I00 and supported by FCT (Fundação para a Ciência e a Tecnologia), Portugal through funding of the PhD Scholarship with ref. 2020.05393.BD, and the LASIGE Research Unit, ref. UIDB/00408/2020 (https://doi.org/10.54499/UIDB/00408/2020) and ref. UIDP/00408/2020 (https://doi.org/10.54499/UIDP/00408/2020).

Appendix. Hyperparameters settings

The hyperparameters used for training these models are detailed here to ensure reproducibility. A uniform learning rate of 2×10^{-5} and a weight decay of 0.01 were applied across all experiments. The batch size was set to 128 for XLM-R base models—*ClinLinker* and *ClinLinker-KB*—, while it was reduced to 64 for XLM-R large models due to their higher computational requirements—*SapBERT-XLM-R-large*—. For these models, a similarity-based threshold of 0.35 was chosen after an initial exploratory analysis. These hyperparameters were consistently applied across all experiments involving similarity-based triplets to ensure comparable and robust results.

Data availability

No data was used for the research described in the article.

References

[1] Mohamed AlShuweihi, Said A. Salloum, Khaled Shaalan, Biomedical corpora and natural language processing on clinical text in languages other than english: A systematic review, in: Mostafa Al-Emran, Khaled Shaalan, Aboul Ella Hassanien (Eds.), Recent Advances in Intelligent Systems and Smart Applications, Springer International Publishing, Cham, 2021, pp. 491–509.

[2] Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F. Jones, Richard Forshee, Mark Walderhaug, Taxiarchis Botsis, Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review, J. Biomed. Inf. 73 (2017) 14–29.

[3] Wei Shen, Jianyong Wang, Jiawei Han, Entity linking with a knowledge base: Issues, techniques, and solutions, IEEE Trans. Knowl. Data Eng. 27 (2) (2015) 443–460.

[4] Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, Yasha Wang, Hykge: A hypothesis knowledge graph enhanced framework for accurate and reliable medical llms responses, 2024.

[5] Ping Zhang, Fei Wang, Jianying Hu, Robert Sorrentino, Towards personalized medicine: leveraging patient similarity and drug similarity analytics, AMIA Summits Transl. Sci. Proc. 2014 (2014) 132.

[6] Meng Wang, Haofen Wang, Xing Liu, Xinyu Ma, Beilun Wang, Drug-drug interaction predictions via knowledge graph and text embedding: instrument validation study, JMIR Med. Inform. 9 (6) (2021) e28277.

[7] Richard M. Goldberg, John Mabee, Linda Chan, Sandra Wong, Drug-drug and drug-disease interactions in the ed: analysis of a high-risk population, Am. J. Emerg. Med. 14 (5) (1996) 447–450.

[8] Haiying Lu, Qiaodan Zhou, Jun He, Zhongliang Jiang, Cheng Peng, Rongsheng Tong, Jianyou Shi, Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials, Signal Transduct. Target. Ther. 5 (1) (2020) 213.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.

[10] Olivier Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic Acids Res. 32 (suppl_1) (2004) D267–D270.

[11] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, Nigel Collier, Self-Alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 4228–4238, Online.

[12] Antonio Miranda-Escalada, Luis Gasco, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, Martin Krallinger, Overview of disTEMIST at bioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2022.

[13] Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, Martin Krallinger, Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023, in: Conference and Labs of the Evaluation Forum, 2023.

[14] Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, Martin Krallinger, Overview of Symptemist at Biocreative VIII: Corpus, Guidelines and Evaluation of Systems for the Detection and Normalization of Symptoms, Signs and Findings from Text, vol. 11, Zenodo, 2023.

[15] Fernando Gallego, Guillermo López-García, Luis Gasco-Sánchez, Martin Krallinger, Francisco J. Veredas, ClinLinker: Medical entity linking of clinical concept mentions in Spanish, 2024.

[16] Fernando Gallego, Guillermo López-García, Luis Gasco, Martin Krallinger, Francisco J. Veredas, ClinLinker-KB: clinical entity linking in spanish with knowledge-graph enhanced biencoders, 2024, Available at SSRN 4939986.

[17] André Leal, Bruno Martins, Francisco M. Couto, Ulisboa: Recognition and normalization of medical concepts, in: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval 2015, 2015, pp. 406–411.

[18] Hongfang Liu, Stephen T. Wu, Dingcheng Li, Siddhartha Jonnalagadda, Sunghwan Sohn, Kavishwar Wagholikar, Peter J. Haug, Stanley M. Huff, Christopher G. Chute, Towards a semantic lexicon for clinical natural language processing, in: AMIA Annual Symposium Proceedings, Vol. 2012, American Medical Informatics Association, 2012, p. 568.

[19] Jennifer D'Souza, Vincent Ng, Sieve-based entity linking for the biomedical domain, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 297–302.

[20] Evan French, Bridget T. McInnes, An overview of biomedical entity linking throughout the years, J. Biomed. Informatics 137 (2023) 104252.

[21] Stefan Schulz, Udo Hahn, Medical knowledge reengineering—converting major portions of the UMLS into a terminological knowledge base, Int. J. Med. Informatics 64 (2–3) (2001) 207–221.

[22] James J. Cimino, George Hripcsak, Stephen B. Johnson, Carol Friedman, Daniel J. Fink, Paul D. Clayton, R.A. Miller, UMLS as knowledge base—a rule-based expert system approach to controlled medical vocabulary management, in: Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, 1990, pp. 4–7.

[23] Wei Shen, Jianyong Wang, Jiawei Han, Entity linking with a knowledge base: issues, techniques, and solutions, IEEE Trans. Knowl. Data Eng. 27 (2) (2014) 443–460.

[24] Wei Shen, Jianyong Wang, Ping Luo, Min Wang, LINDEN: Linking named entities with knowledge base via semantic knowledge, in: Proceedings of the 21st International Conference on World Wide Web, ACM, 2012, pp. 449–458, http://dx.doi.org/10.1145/2187836.2187898.

[25] Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum, YAGO: A core of semantic knowledge unifying WordNet and wikipedia, in: Proceedings of the 16th International World Wide Web Conference, WWW, 2007, pp. 697–706.

[26] Robert Leaman, Rezarta Islamaj Doğan, Zhiyong Lu, Dnorm: disease name normalization with pairwise learning to rank, Bioinformatics 29 (22) (2013) 2909–2917.

[27] Ganggao Zhu, Carlos A. Iglesias, Exploiting semantic similarity for named entity disambiguation in knowledge graphs, Expert Syst. Appl. 101 (2018) 8–24.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[29] Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, Valentin Malykh, Medical concept normalization in social media posts with recurrent neural networks, J. Biomed. Informatics 84 (2018) 93–102.

[30] Zhaochen Guo, Denilson Barbosa, Robust named entity disambiguation with random walks, Semant. Web 9 (4) (2018) 459–479.

[31] Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, Elena Tutubalina, Medical concept normalization in clinical trials with drug and disease representation learning, Bioinformatics 37 (21) (2021) 3856–3864.

[32] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, Luke Zettlemoyer, Scalable zero-shot entity linking with dense entity retrieval, in: Bonnie Webber, Trevor Cohn, Yulan He, Yang Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 6397–6407, Online.

[33] Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, Hoifung Poon, Knowledge-rich self-supervision for biomedical entity linking, 2021, arXiv preprint arXiv:2112.07887.

[34] Elena Tutubalina, Artur Kadurin, Zulfat Miftahutdinov, Fair evaluation in concept normalization: a large-scale comparative analysis for BERT-based models, in: Donia Scott, Nuria Bel, Chengqing Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain, 2020, pp. 6710–6716, (Online).

[35] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, Jaewoo Kang, Biomedical entity representations with synonym marginalization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 3641–3650, http://dx.doi.org/10.18653/v1/2020.acl-main.335.

[36] Andre Lamurias, Pedro Ruas, Francisco M. Couto, Ppr-ssm: personalized pagerank and semantic similarity measures for entity linking, BMC Bioinformatics 20 (2019) 1–12.

[37] Florian Borchert, Ignacio Llorca, Roland Roller, Bert Arnrich, Matthieu-P. Schapranow, Xmen: A modular toolkit for cross-lingual medical entity normalization, 2023, arXiv preprint arXiv:2310.11275.

[38] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, Sheng Yu, Biobart: Pretraining and evaluation of a biomedical generative language model, 2022, arXiv preprint arXiv:2204.03905.

[39] Hongyi Yuan, Zheng Yuan, Sheng Yu, Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning, 2022, arXiv preprint arXiv:2204.05164.

[40] Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Haitao Zheng, Zhiyuan Liu, Juanzi Li, Hong-Gee Kim, Prompt-learning for fine-grained entity typing, in: Yoav Goldberg, Zornitsa Kozareva, Yue Zhang (Eds.), Findings of the Association for Computational Linguistics, EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6888–6901.

[41] Florian Borchert, Ignacio Llorca, Matthieu-P. Schapranow, Improving biomedical entity linking for complex entity mentions with llm-based text simplification, Database 2024 (2024) baae067.

[42] Bilal Abu-Salih, Muhammad Al-Qurishi, Mohammed Alweshah, Mohammad Al-Smadi, Reem Alfayez, Heba Saadeh, Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities, J. Big Data 10 (1) (2023) 81.

[43] Fernando Gallego, Francisco J. Veredas, ICB-UMA at BioCreative VIII @ AMIA 2023 task 2 SYMPTEMIST (symptom TExt mining shared task), in: Rezarta Islamaj, Cecilia Arighi, Ian Campbell, Graciela Gonzalez-Hernandez, Lynette Hirschman, Martin Krallinger, Salvador Lima-López, Davy Weissenbacher, Zhiyong Lu (Eds.), Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the Era of Generative Models, 2023.

[44] Maria Pershina, Yifan He, Ralph Grishman, Personalized page rank for named entity disambiguation, in: Human Language Technologies: The 2015 Annual Conference Ofthe North American Chapter Ofthe ACL, number Section 4, Denver, Colorado, May 31 – June 5 2015, Association for Computational Linguistics, 2015, pp. 238–243.

[45] Andre Lamurias, Pedro Ruas, Francisco M. Couto, PPR-SSM: Personalized PageRank and semantic similarity measures for entity linking, BMC Bioinformatics 20 (1) (2019) 1–12.

[46] Pedro Ruas, Andre Lamurias, Francisco M. Couto, Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature, J. Cheminformatics 12 (1) (2020) 1–11.

[47] Mingchen Li, Halil Kilicoglu, Hua Xu, Rui Zhang, BiomedRAG: A retrieval augmented large language model for biomedicine, 2024, arXiv preprint arXiv:2405.00465.

[48] Mohammad Alkhalaf, Ping Yu, Mengyang Yin, Chao Deng, Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records, J. Biomed. Inf. 156 (104662) (2024) 104662.