

# Multilingual bi-encoder models for biomedical entity linking

Zekeriya Anil Guven<sup>1</sup> | Andre Lamurias<sup>2</sup>

<sup>1</sup>Department of Computer Engineering, Izmir Bakircay University, Izmir, Turkey

<sup>2</sup>Department of Computer Science, NOVA School of Science and Technology, Lisbon, Portugal

## Correspondence

Zekeriya Anil Guven, Department of Computer Engineering, Izmir Bakircay University, Izmir, Menemen 35100, Turkey.  
Email: [zekeriyaanil.guven@bakircay.edu.tr](mailto:zekeriyaanil.guven@bakircay.edu.tr)

## Abstract

Natural language processing (NLP) is a field of study that focuses on data analysis on texts with certain methods. NLP includes tasks such as sentiment analysis, spam detection, entity linking, and question answering, to name a few. Entity linking is an NLP task that is used to map mentions specified in the text to the entities of a Knowledge Base. In this study, we analysed the efficacy of bi-encoder entity linking models for multilingual biomedical texts. Using surface-based, approximate nearest neighbour search and embedding approaches during the candidate generation phase, accuracy, and recall values were measured on language representation models such as BERT, SapBERT, BioBERT, and RoBERTa according to language and domain. The proposed entity linking framework was analysed on the BC5CDR and Cantemist datasets for English and Spanish, respectively. The framework achieved 76.75% accuracy for the BC5CDR and 60.19% for the Cantemist. In addition, the proposed framework was compared with previous studies. The results highlight the challenges that come with domain-specific multilingual datasets.

## KEYWORDS

*biomedical entity linking*, data analysis, entity linking, language model, multilingual analysis, natural language processing

## 1 | INTRODUCTION

Today, there is a lot of text data that can be processed in many areas. Since the amount of this data is constantly increasing, it becomes more and more difficult to process it. Therefore, there are Natural Language Processing (NLP) methods to perform data analysis on texts. NLP is a sub-field of artificial intelligence that helps computers understand, interpret and process human language. NLP has many application areas such as text classification and categorization, Named Entity Recognition, Part-of-Speech Tagging, Question Answering, Entity Linking (EL), and Machine Translation.

EL, one of the application areas of NLP, is a technique used to map the mentions specified in the text to the entities stored in a Knowledge Base (KB). It is also known as Entity Disambiguation and Entity Normalization. In EL, it is important to understand the text, but it is difficult to disambiguate entity mentions because differently named entities can have the same form or the same entity may have different aliases. EL methods fall into two categories. Local models are concerned with the contextual words surrounding independently ambiguous mentions. Global models take into account the coherence value between entities referred to in the same document, so the full context of the document is taken into account (Fang et al. 2019).

EL can eliminate inconsistencies such as entity conflicts and ambiguous references, so machines can understand incoming data and build a unified knowledge base with that data (Yin et al. 2019). EL can be applied in many areas such as biomedicine, social media, and question-

---

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Expert Systems* published by John Wiley & Sons Ltd.

answering, to eliminate inconsistencies in data. Biomedical EL is the task of matching expressions from biomedical texts, such as clinical notes or biomedical literature, to medical entities in standard ontology (UMLS, MeSH, etc.). An example of biomedical EL is given in Figure 1. In the healthcare domain, it is very important to eliminate the uncertainty of biomedical entities. If the biomedical entity is misunderstood, the whole context will be misinterpreted. This means high risks in decision-making in terms of medicine (Zhu et al. 2020). To avoid this problem, there are two important steps in EL; candidate generation (CG) and disambiguation. CG is in the process of creating a list of the desired number of KB entities for each mention. Disambiguation is the case of selecting the most likely KB entity in the candidate list (Zhou et al. 2020).

In this study, we compare deep learning approaches for Biomedical EL. Most existing work on EL focuses on English language text, so we explore how these approaches perform in non-English documents. To perform biomedical EL in other languages, a multilingual Bi-encoder EL framework has been implemented. Analysis of the EL framework was performed with the Cantemist (Miranda-Escalada et al. 2020) corpus, which is the Spanish tumour morphology dataset, for Spanish and BC5DCR (Li et al. 2016), containing chemical and disease mentions, in English. Before the EL task, the candidates for the relevant mention were determined by the CG process. Surface-based and Approximate Nearest Neighbour (ANN) search methods are used for CG. While EL models were trained with these candidates, pre-trained BERT, RoBERTa, SapBERT, and BioBERT language models (LMs) were used in the analysis stage. The contributions of this research can be summarized as follows:

- Multilingual Bi-encoder EL framework using pre-trained LMs with surface-based and ANN-search;
- The success of the EL framework with pre-trained LMs was analysed for Spanish, which is the second most spoken language in the world, and English, for which more NLP resources are available.

The remainder of the article is structured as follows. Literature research on EL and biomedical EL is explained under the title “Related Work” in Section 2. Section 3 describes the methodology of this study including datasets, candidate generation, and disambiguation. Experiments and results for each dataset with LMs are analysed in Section 4. Finally, the conclusions of this research and future works are discussed.

## 2 | RELATED WORK

This section presents an overview of studies done for EL in different languages and domains. There are certain challenges with EL studies in multiple languages. One of the challenges of EL is resolving the ambiguity of natural language. Another challenge is to correctly find the feature representation of mentions and entities in the candidate entity ranking for the EL system (Yin et al. 2019).

### 2.1 | Entity linking

De Cao et al. (2020) proposed the GENRE (Generative Entity Retrieval) model to generate autoregressive entity names using a sequence-to-sequence architecture. The GENRE model uses a pre-trained model on BART (Lewis et al. 2019) weights and fine-tuned transformers-based architecture. Due to the problem that the output is not always a valid entity name, the model forces each generated name to be in a predefined set of candidates using a constrained decoding strategy. They analysed the GENRE model in the stages of entity disambiguation, EL, and document retrieval. Empirically, they have shown that the model leads to state-of-art performance on multiple entity retrieval datasets for these phases.

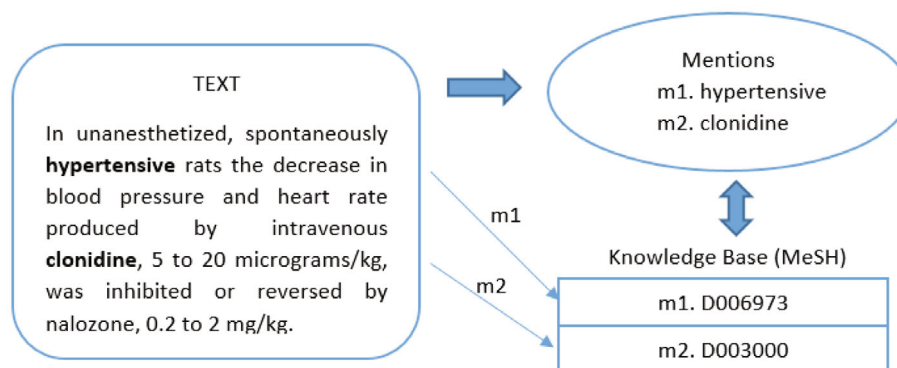


FIGURE 1 An example for entity linking.

Wu et al. (2020) proposed, namely BLINK, a conceptually scalable, and highly effective two-stage approach for EL. In the first stage, they applied retrieval with a pair of encoders that included the mention context and the entity descriptions. After the retrieval process, each candidate is analysed with a cross-encoder that combines the mention and entity text. With the model they developed based on BERT, their model performed better than IR methods for entity retrieval in the zero-shot EL dataset.

Ravi et al. (2021) proposed an approach called heuristic CHOLAN, which includes a modular architecture of two BERT models, for the EL task. While the first BERT model defines entity mentions in a given text, a second BERT model is used to classify the target entity for each mention among a predefined list of candidates. At the analysis stage, they showed that CHOLAN outperformed state-of-the-art approaches on standard datasets such as CoNLL-AIDA, MSNBC, AQUAINT, ACE2004, and T-REx.

Kolitsas et al. (2018) propose a neural end-to-end EL system that can detect and link entities in a text document. The model first generates all possible spans with at least one possible entity candidate. Then, a compatibility score based on word and entity embeddings, along with neural attention and global voting mechanisms, is calculated for each mention-candidate pair. Their purpose is to evaluate all possible ranges as potential mentions and to learn useful contextual similarity scores from among entity candidates. Examining the results of their models, they outperformed the popular systems on the Gerbil<sup>1</sup> platform significantly.

Labusch and Neudecker (2020) proposed a named entity disambiguation and linking (NED, NEL) system. Their system first searches for possible candidates on an ANN directory in BERT embedding. To evaluate candidates, Wikipedia text passages are compared by a trained BERT model. As a result, they ranked the candidates using the previously collected information. They evaluated the system for English, German, and French datasets and showed that NEL approaches are competitive in precision.

Logeswaran et al. (2020) proposed a new zero-shot EL system. In their system, they did not limit mention statements to the named entity. With this system, they showed that the attention between mention in context and entity definitions, which was not present in previous studies, is important for EL systems.

Sil et al. (2018) proposed a neural EL model that analyzes many aspects of fine-grained similarities and differences between the query and the candidate document. The analysis is performed by training the models together with convolution and tensor networks. They have shown that in their English-trained systems, it can be applied to different languages in zero-shot learning as a result of the effective use of multilingual embedding. Their proposed system produced very strong results in both English and cross-lingual Spanish and Chinese languages.

Fang et al. (2019) presented a reinforcement learning-based model for EL. This model learns the policy of sequentially selecting entities and utilizes both local context and global coherence. It makes full use of previously selected entity information to resolve global coherence. In experimental studies, they analysed their models in many benchmark datasets and showed significant improvements in the results. There are also many studies for EL domain, such as text classification (Yamada and Shindo 2019), zero-shot approach (Wu et al. 2020), latent entity type research (Chen, Wang, et al. 2020; Chen, Varoquaux, et al. 2020), and tweet classification (Feng et al. 2018; Chong et al. 2017).

## 2.2 | Biomedical entity linking

In this subsection, we present an overview of the state-of-the-art EL approaches using a knowledge base for the biomedical domain.

### 2.2.1 | CANTEMIST

For Cantemist corpus, Xiong et al. (2020) proposed a joint learning method, which applied medical named entity recognition (MER) and medical named entity normalization (linking) (MEN). They used a machine reading comprehension model to solve the MER task and multiple sequence labeling models to solve the MEN task. Although the models were different for each task, they used the same word representations in both tasks. In the analysis phase, the model has been quite successful. Garca-Pablos et al. (2020) developed a system based on end-to-end deep learning. They obtained semantic representations of texts with pre-trained BERT models. With this semantic representation, ICD-O-3 codes were calculated for each token and this information was used separately for entity recognition, normalization, and coding subtasks of Cantemist. They experimented with different pre-trained BERT models (SciBERT, BETO) and different ways to combine these models. When the results were examined, they showed that the ensemble models were more successful than the individual models. Lopez-Ubeda et al. (2020) presented an automated neural network-based model for the extraction of entities for Cantemist corpus. In the model, they used a two-way variant by utilizing the Conditional Random Fields (CRF) layer and the Long Short-Term Memory (BiLSTM). Then, they used a dictionary-based system to subsequently assign a morphology code to each annotated entity, that is, the EL process. Lange et al. (2020) applied the NER and EL tasks for the Cantemist dataset. The first sub-task utilized neural sequence labeling and parsing approach to address the NER. For the EL task, they used rule-based methods due to the small size of the training data. They normalized the extracted entities for EL using string matching and Levenshtein distance. Jabreel (2020) proposed a neural network-based system, namely E2ENC, which provides a data-driven and end-to-end solution. The system does not need any parser or other language resources. They used this system for MER and MEL tasks. For the analysis of MER and MEL tasks, they

transferred the output of the neural network layers to the four Conditional Random Field (CRF) models that were learned jointly. Many papers also analysed EL for Cantemist corpus (Hassan et al. 2020; Ruas, Neves, et al. 2020; Ruas, Lamurias, et al. 2020; Rivera-Zavala and Martinez 2020).

### 2.2.2 | BC5CDR

For the BC5CDR corpus, Wiatrak and Iso-Sipila (2020) propose a simpler model for evaluating biomedical data, using a neural end-to-end EL model that leverages a pre-trained LM. They focused on the per-entity and per-token classification problems for the entire dictionary. They evaluated the BC5CDR and different datasets to measure the effect of the model on EL and NER with hierarchical multitasking. Ujiie, Iso, Yada, et al. (2021) and Ujiie, Iso, and Aramaki (2021) propose a neural network-based model that takes advantage of dictionary-matching features for biomedical entity normalization. After the models calculate all the span representations for the input sentence, these representations combine the dictionary-matching features. This proposed model was evaluated for the BC5CDR corpus and given better performance for normalization. Angell et al. (2021) proposed an EL model that provides common entity linkage predictions. They developed a cluster-based inference procedure. According to their hypothesis, in co-clustered promises, the entire set can be classified successfully if the word connects with the correct entity. They evaluated their models on BC5CDR and MedMentions datasets. Ruas, Neves, et al. (2020) and Ruas, Lamurias, et al. (2020) propose a framework to improve the precision of graph-based Named EL models. With this graph-based model, they applied relation extraction for entity linking. Relation extraction is used to resolve the uncertainties of biomedical entities. They evaluated the proposed models with BC5CDR and different datasets.

### 2.2.3 | Other studies

Other biomedical studies can be found for EL. Chen, Wang, et al. (2020) and Chen, Varoquaux, et al. (2020) proposed a lightweight neural method for biomedical EL. The method only needs some of the parameters of the BERT model. This model used an alignment layer with attention mechanisms to capture the variations between candidates and mention names. The results of this method analysed on the ShARE/CLEF, NCBI, and ADR benchmark datasets showed that the model was very effective. Vashishth et al. (2020) added a word disambiguation step for EL and proposed a new method called MEDTYPE based on deep learning. In this method, they assigned a semantic type to words defined based on context and used it to develop the list of candidate concepts. By analysing their newly created datasets on the method, they showed that MEDTYPE pruned a significant number of irrelevant candidate concepts and provided state-of-the-art performance for medical EL. Since most of the texts for the biomedical literature are in English, Andrade et al. (2021) proposed the public deep learning NERL (Named Entity Recognition and Linking) system, which offers a parallel corpus in the oncological domain for Spanish and Portuguese. They have developed the Iberian Cancer-Related Entity Recognition and Linking (ICERL) system, the NERL system, and the parallel corpus ICR (Iberian Cancer-related) to validate the possibility of both languages using similar text mining tools and to transfer annotations. As a result of the experiments, they confirmed the similarity between the annotation statistics of the two languages. Zhu et al. (2020) proposed a neural network-based LATTE model for biomedical EL, which models latent types of words and entities. The latent types in the model represent the implicit attributes of each entity. While the model uses additional coarse-grained entities to aid the training process, they also utilized the attention-based mechanism to provide the link mentions with the entities from the knowledge base. As a result of the analysis, they showed that latent-type modelling improves EL performance. There are also neural representations and dictionary matching (Loureiro and Jorge 2020), contextualized neural representations (Schumacher et al. 2020) studies for biomedical EL.

## 3 | METHODOLOGY

In this study, we explore biomedical EL for Spanish and English datasets. A multilingual EL framework is proposed for the analysis of both datasets. In our methodology, first, CG is applied for each dataset. MeSH<sup>2</sup> for English and the eCIE-O<sup>3</sup> for Spanish knowledge bases are used in the CG process. During the training of the model, Bi-encoder methods based on Surface-based and ANN-Search methods are compared with BLINK (Wu et al. 2020) developed by Facebook. Finally, the success of these methods on the generated candidates is measured via pre-trained LMs. The general operation of the methodology is shown in Figure 2. All stages are explained in detail under the following sections.

### 3.1 | Datasets

BC5CDR<sup>4</sup> for English and Cantemist<sup>5</sup> for Spanish were used as biomedical datasets. These datasets are publicly accessible.

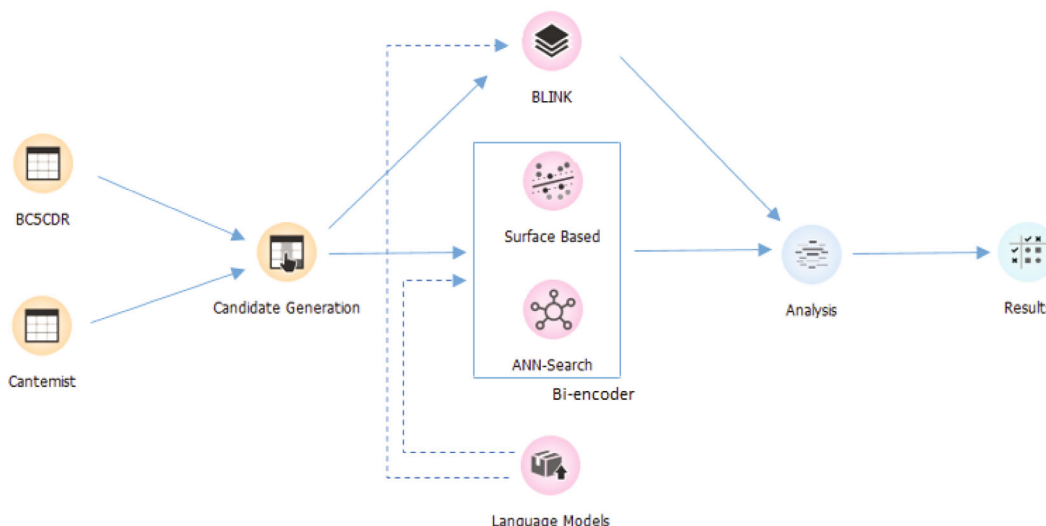


FIGURE 2 Methodology following in this study.

TABLE 1 Statistics of Cantemist corpus.

	Annotations	Unique codes	Sentences	Tokens
Training	6396	493	25,144	447,903
Development	6301	520	23,513	401,994
Test	3633	386	14,359	243,604
Total	16,030	850	63,016	1,093,501

### 3.1.1 | Cantemist corpus

Cantemist is a tumour morphology dataset created from Spanish oncological clinical cases. Cantemist cases contain the patient's clinical signs and symptoms, current illnesses, blood tests, family medical history, diagnosis, and treatments. It includes 1301 case reports, including 501 training, 500 development, and 300 test documents. All documents in the dataset were annotated manually by the experts with tumour morphology. English ICD-O was used for tumour morphology codes (Spanish: eCIE-O) (Miranda-Escalada et al. 2020). The statistics of the Cantemist corpus are given in Table 1.

The Cantemist corpus is used for three sub-tasks in different studies. The achievements of the studies for these tasks are compared (Miranda-Escalada et al. 2020). The sub-tasks of Cantemist can be explained as follows:

- Cantemist-NER: It provides automatic detection of content expressing tumour morphology.
- Cantemist-NORM: It aims to find all tumour morphology entities with morphology codes. This is the clinical concept normalization process.
- Cantemist-CODING: It aims to return the correct eCIE-O codes for documents in sorted order.

### 3.1.2 | BC5CDR corpus

The BC5CDR corpus is an English dataset of PubMed articles that contain annotated chemicals, diseases, and chemical-disease interactions. Each description includes text spans and associated concept identifiers from MeSH. The 1500 PubMed articles in the dataset are split equally for the training, development, and test sets (Li et al. 2016). The statistics of the dataset are given in Table 2.

The BC5CDR corpus is used for two sub-tasks in other studies (Li et al. 2016). The sub-tasks of BC5CDR can be explained as follows:

- DNER: It is the task of named entity recognition and normalization for diseases.
- CID: It is the task applied for the relation extraction of chemical-induced diseases.

**TABLE 2** Statistics of BC5CDR corpus.

	Chemical mention	Disease mention	CID relation
Training	5203	4182	1965
Development	5347	4244	1865
Test	5385	4424	1988

## 3.2 | Candidate generation

CG processes ensure that the most suitable candidates are selected among all candidates. A certain number of candidates are sorted according to their probabilities through the algorithms and libraries used and these candidates are selected for the models to find the correct answer. Thus, analysis is carried out with a subset of candidates instead of all candidates.

ScispaCy<sup>6</sup> from AllenNLP is used to generate candidates for the BC5CDR corpus. ScispaCy is an entity span detection model trained on biomedical data by adding new token rules to the rule-based tokenizer of the spaCy library. The model can apply the CG process to biomedical data for English by using knowledge bases such as MeSH, and UMLS (Neumann et al. 2019). Therefore, this library was chosen for CG for BC5CDR and the analysis was performed by selecting 5, 10, and 50 candidates.

The fuzzywuzzy<sup>7</sup> library is used for CG in the Cantemist dataset. This library is a Fuzzy text-matching technique for finding sequences that partially match the sequence. Thus, it also determines how close the two sequences are. Fuzzywuzzy uses Levenshtein distance to calculate the proximity between sequences. Analysis was performed with 50, 100, and 250 candidates selected for the Cantemist corpus.

## 3.3 | Methods

### 3.3.1 | Blink

BLINK<sup>8</sup> is an EL library that uses the Wikipedia knowledge base. The library was developed by the Facebook research team. It is based on the fine-tuned BERT architecture and uses a two-stage approach. Firstly, it performs a retrieval in the related space defined by bi-encoders that independently embed the mentioned context and entity definitions. Afterward, each candidate is analyzed in more detail with a cross-encoder that combines the word and entity text (Wu et al. 2020). Laskar, Chen, Martsinovich, et al. (2022), Laskar, Chen, Johnston, et al. (2022), Bhargav et al. (2022) have also employed this method in other studies.

### 3.3.2 | Surface-based CG

If an alias table does not exist or is of limited scope, CG can be performed by exploiting the surface similarity between mentions and entities. This method only uses entity mentions, it has no semantic understanding while generating candidates for each entity mention. A surface-based bi-encoder model was used in this study. The operation of this model is shown in Figure 3. First, the entity mention is sent to the CG stage with the mentioned encoder. In this process, CG is performed for each mention by utilizing the entity's definitions and structures. The candidates are given to the entity encoder and the score of each candidate entity is obtained. The sorted candidates create as many candidate lists as desired. Murty et al. (2018) and Zhu et al. (2020) used surface similarity between mentions and entities.

### 3.3.3 | Approximate nearest neighbour search

Gillick et al. (2019) proposed the ANN search method to remove the dependency on alias tables. This method requires a wide range of contextual entity-mention pairs for training. It calculates the similarity between the mention and all other mentions using the distance metric. In ANN search, new index structures are obtained by narrowing the search space. The high dimension problem is also reduced with the implicit neighbourhoods in the indices (Gillick et al. 2019). A bi-encoder EL model was used by the ANN search method. The operation of the model is given in Figure 4. First, the canonical names and descriptions of all entities are encoded for entity mentions. All entities are encoded with these properties. Afterward, entity estimation is performed with the ANN search and score function of the entities. With these sorted scores, a certain number of candidate lists can be created. Wu et al. (2020) and Logeswaran et al. (2020) utilized this approach together with the Zero-shot approach.

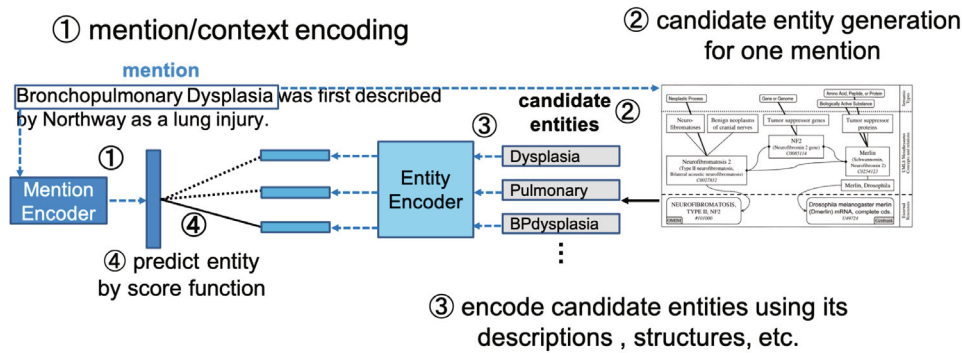


FIGURE 3 The operations of surface-based CG izu (Building Transformer-Based Entity Linking System, n.d.)

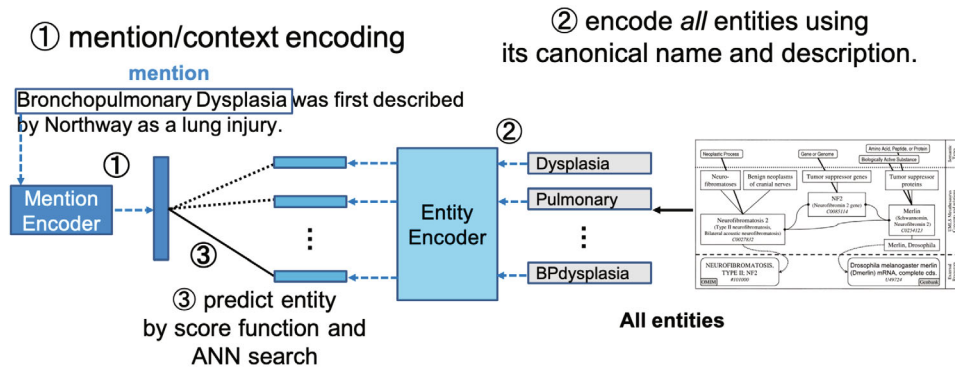


FIGURE 4 The operations of ANN search izu (Building Transformer-Based Entity Linking System, n.d.)

### 3.4 | Language models

#### 3.4.1 | BERT

BERT is a transformer-based bidirectional encoder representation model. It produces contextual word representations, trained on large dataset collection. It randomly masks some tokens in the input with a novel method called the masked language model (MLM). The model aims to predict this masked word based on its context only. Masking is applied once during the data preprocessing phase. The BERT model can be used pre-trained for different tasks such as text classification, text summarization, text generation, question answering, and machine translation Devlin et al. (2019).

#### 3.4.2 | RoBERTa

The training phase of BERT is computationally costly because it takes a long time for high-dimensional data. The selection of hyperparameters during the training phase gains importance in terms of cost. Therefore, the Robustly Optimized BERT (RoBERTa) model has been developed by applying changes in the pre-training phase of the BERT model. RoBERTa creates a masking pattern every time a sequence is fed to the model with dynamic masking during the training phase. It uses larger mini-batches than BERT. In addition, while BERT implementation uses a character-level BPE vocabulary of size 30 K, RoBERTa uses a larger byte-level BPE vocabulary containing 50 K (Liu et al. 2019).

#### 3.4.3 | SapBERT

SapBERT is proposed to pre-train the Transformer-based language model on the biomedical infographic of UMLS, which consists of an extensive collection of biomedical synonyms. The model includes a BERT-based framework that learns to self-align synonymous biomedical entities. This metric learning framework is used for pre-training on UMLS and fine-tuning custom datasets (Liu et al. 2020).

### 3.4.4 | BioBERT

BioBERT is a language representation model pre-trained on biomedical data (PubMed summaries and PMC full-text articles). The model is first initialized with the weight in the BERT model using the general corpus and then pre-trained with biomedical data. BioBERT has been fine-tuned and performance measured on three popular NER, Relation Extraction, and Question Answering biomedical text mining tasks (Lee et al. 2019).

In this study, BERT-Base<sup>9</sup>, BioBERT<sup>10</sup>, and SapBERT<sup>11</sup> pre-trained models were used for English, and Roberta-base-biomedical-es<sup>12</sup> (RBB-Es), Roberta-base-biomedical-clinical-es<sup>13</sup> (RBBC-Es), BioBert-base<sup>14</sup> (BB-Es) pre-trained models for Spanish. The information of these LMs is given in Table 3.

## 4 | EXPERIMENTS AND RESULTS

### 4.1 | Dataset preparation

Preprocessing techniques were applied to make Cantemist and BC5CDR datasets suitable for the EL models. These datasets contain annotations (id, text, definition, offset, ICD-O code) and masked word (Neuromielitis óptica paraneoplásica secundaria a < objetivo > carcinoma microcítico). These datasets were split into development, training, and test sets in JSON format. By utilizing the information in these files, the file names of the corpus have been stored in a separate txt format. Afterward, files were created with the data (concept\_id, aliases, canonical\_name, definition) of the ICD-O and MeSH knowledge bases used for Spanish and English, respectively. Finally, EL analysis can be performed with all files obtained from datasets and knowledge bases. The procedure for the preparation of datasets is indicated in Figure 5.

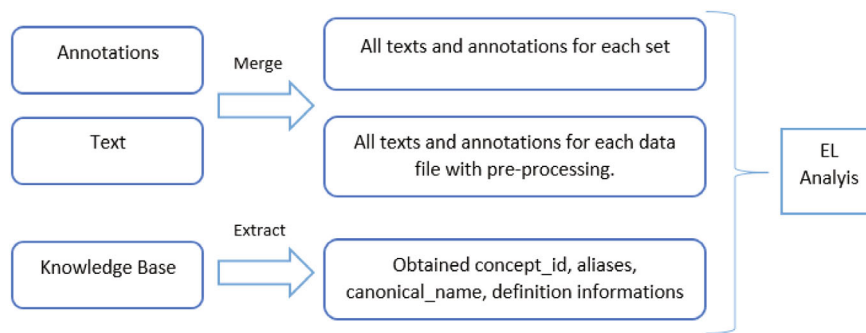
The CG process was carried out by using the obtained datasets and the morphology codes in the knowledge base. The morphology codes record the type of tumour and how it behaves. MeSH knowledge base and morphology codes in ICD-O were used for the BC5CDR and Cantemist datasets, respectively. For the BC5CDR corpus, lists of 5, 10, and 50 candidates were obtained through the ScispaCy library for the CG process. For the Cantemist corpus, lists of 50, 100, and 250 candidates were produced using the fuzzy-wuzzy string matching library, which uses the Levenshtein Distance algorithm. The reason why the number of candidates is different for both datasets is to search for the most suitable number of candidates that can be accessed for each language. In these datasets, statistics on whether there is an actual morphology code among the candidates were also extracted. Information about this statistic is given in Table 4. When the statistics in the table are examined, approximately 75.55% of the actual candidates are reached for the Cantemist corpus, while this rate is approximately 74% for the BC5CDR corpus.

### 4.2 | Entity linking analysis

In this study, the success of the EL models with pre-trained LMs was analysed. The parameters used in the analysis phase are given in Table 5. The training was performed for 16 and 48-batch sizes. The dataset, bert\_name, model\_for\_training, candidates\_dataset, max\_candidates\_num, language parameters change according to the datasets.

**TABLE 3** Pre-trained LMs used in this study and their descriptions.

Models	Descriptions	Language
BERT-Base	The BERT model was pre-trained on BookCorpus and English Wikipedia. The model was trained on 4 cloud TPUs in Pod configuration (16 TPU chips total) with a batch size of 256. The sequence length was limited to 128 tokens for 90% of the steps and 512 for the remaining 10%.	English
BioBERT	BioBERT was pre-trained on English Wikipedia, BooksCorpus, and PubMed biomedical articles. It is trained in the same way as BioBERT-Base v1.1 but includes LM head, which can be useful for probing.	English
SapBERT	Self-align synonymous biomedical entities are learned by this model. SapBERT trained with UMLS 2020AA (English only), using microsoft/ BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext as the base model.	English
RBB-Es	This model is a RoBERTa-based model trained on a biomedical corpus in Spanish collected from several sources. The training lasted a total of 48 hours with 16 NVIDIA V100 GPUs of 16GB DDRAM, using Adam optimizer with a peak learning rate of 0.0005 and an effective batch size of 2048 sentences.	Spanish
RBBC-Es	This model is a RoBERTa-based model trained on a biomedical-clinical corpus in Spanish collected from several sources. Train phase same as with Roberta-base-biomedical.	Spanish
BB-Es	BERT masked language model fine-tuned from dccuchile/bert-base-Spanish-wwm-uncased over clinical text in Spanish. This model was fine-tuned over a clinical corpus comprised of 5,157,902 free-text diagnostic suspicions extracted from Chilean waiting list referrals.	Spanish



**FIGURE 5** The procedure for the preparation of datasets.

**TABLE 4** Statistics of finding the actual morphology code within the candidates.

Cantemist candidates	Rates (%)	BC5CDR candidates	Rates (%)
50	61.14	5	67.82
100	66.80	10	70.87
250	75.55	50	74.29

**TABLE 5** Important parameters of EL model.

Name	Value
learning_rate	1.00E-05
beta1	0.9
beta2	0.999
epsilon	1.00E-08
word_embedding_dropout	0.1
num_epochs	10
model_for_training	Blink, surface-biencoder, ANN-biencoder
max_candidates_num	50
language	(english or spanish)

First, the success of the BC5CDR dataset in the EL approaches together with the language models was analysed. BERT-Base, BioBERT, and SapBERT as LMs, 5, 10, and 50 as the candidate number was used for analysis. The results of the analysis are given in Table 6 for 16 batch sizes and in Table 7 for 48 batch sizes. Recall and accuracy are used as evaluation metrics. While the recall is calculated as the number of true positives (TP) divided by the total number of TP and false negatives (FN), the accuracy gives the ratio of TP and TN to all data (TP, TN, FP, and false negatives (FN)).

Then, the success of the Cantemist corpus in the EL framework together with the pre-trained language models was analysed. RoBERTa-biomedical, RoBERTa-biomedical-clinical, and BioBERT as language models, 50, 100, and 250 as the candidate number was used for analysis. The results of the analysis are given in Table 8 for 16 batch sizes and in Table 9 for 48 batch sizes.

### 4.3 | Results comparison with previous studies

In this section, the results of the previous studies and the proposed EL framework are compared. Models using the entire BC5CDR dataset were chosen to compare accuracy, recall, or F1-measure values. Some studies do not include accuracy or recall values. Therefore, the F1-measure value of the most successful model in the BC5CDR dataset was also calculated. A comparison of studies is shown in Table 10. When the results are analysed, our proposed EL framework is more successful than some studies.

**TABLE 6** Analysis statistics of the EL framework for the BC5CDR corpus (batch size: 16).

Batch size: 16		BLINK		Surface-Biencoder		ANN-Biencoder	
Models	Candidate	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy
BERT-Base	5	74.36	63.57	74.36	65.43	74.36	62.2
	10	77.14	54.84	77.14	52.94	77.14	56.37
	50	78.98	46.37	78.96	48.91	78.97	49.33
BioBERT	5	74.36	68.71	74.36	67.86	74.36	65.8
	10	77.14	63.04	77.14	62.48	77.14	63
	50	78.98	52.55	78.97	57.83	78.97	58.63
SapBERT	5	74.36	76.75	74.36	75.34	74.36	76.05
	10	77.14	73.19	77.14	72.41	77.14	71.9
	50	78.97	71.95	78.97	69.68	78.97	71.11

**TABLE 7** Analysis statistics of the EL framework for the BC5CDR corpus (batch size: 48).

Batch size: 48		BLINK		Surface-Biencoder		ANN-Biencoder	
Models	Candidate	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy
BERT-Base	5	74.36	54.58	77.14	42.26	74.36	47.56
	10	77.14	40.4	77.14	40.1	77.14	42.6
	50	78.96	35.83	78.98	27.62	78.96	35.6
BioBERT	5	74.36	67.82	74.36	68.6	74.36	67.86
	10	77.14	62.88	77.14	60.54	77.14	62.88
	50	78.99	56.27	78.97	58.01	78.96	58.35
SapBERT	5	74.36	76.21	74.36	75.98	74.36	75.24
	10	77.14	73.62	77.14	73.11	77.14	71.85
	50	78.96	71.27	78.96	70.75	78.99	71.82

**TABLE 8** Analysis statistics of the EL framework for the Cantemist corpus (batch size: 16).

Batch size: 16		BLINK		Surface-Biencoder		ANN-Biencoder	
Models	Candidate	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy
RBB-Es	50	52.94	52.85	52.94	53.05	52.94	52.64
	100	55.26	53.25	55.26	51.21	55.26	60.19
	250	60.58	56.68	60.58	52.78	60.58	57.62
RBBC-Es	50	52.94	49.09	52.94	49.39	52.94	49.91
	100	55.26	48.44	55.26	49.33	55.26	50.83
	250	60.58	53.29	60.58	53.5	60.58	55.76
BB-Es	50	52.94	41.17	52.94	41.3	52.94	41.46
	100	55.26	42.92	55.26	39.85	55.26	42.47
	250	60.58	43.86	60.58	33.13	60.58	41.94

For the Cantemist dataset, the results of the previous studies and our proposed EL framework were compared. To make the comparison consistent, the F1-measure value of the model with the highest accuracy value was calculated in Cantemist. The results are given in Table 11. When the results are analysed, the proposed multilingual bi-encoder EL framework outperforms some previous studies.

**TABLE 9** Analysis statistics of the EL framework for the Cantemist corpus (batch size: 48).

Batch size: 48		BLINK		Surface-Biencoder		ANN-Biencoder	
Models	Candidate	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy
RBB-Es	50	52.94	53.02	52.94	52.01	52.94	51.58
	100	55.26	52.49	55.26	54.73	55.26	57.69
	250	60.58	54.79	60.58	55.84	60.58	55.16
RBBC-Es	50	52.94	49.5	52.94	51.54	52.94	51.54
	100	55.26	51.23	55.26	52.59	55.26	51.75
	250	60.58	55.23	60.58	53.67	60.58	54.57
BB-Es	50	52.94	34.56	52.94	38.48	52.94	41.49
	100	55.26	39.74	55.26	39.23	55.26	35.32
	250	60.58	39.42	60.58	36.16	60.58	36.54

**TABLE 10** Comparison for BC5CDR dataset with previous studies.

	Accuracy	Recall
Ujii, Iso, Yada, et al. (2021) and Ujii, Iso, and Aramaki (2021)	69	N/A
Bhowmik et al. (2021)	N/A	74.4
Wiatrak and Iso-Sipila (2020)	91.65	59.27
Angell et al. (2021)	91.4	N/A
Ours	76.75	78.99

**TABLE 11** Comparison for Cantemist dataset with previous studies.

	Recall	F1-measure
Xiong et al. (2020)	82.6	82.5
García-Pablos et al. (2020)	82.1	82.1
López-Úbeda et al. (2020)	75.5	75.9
Lange et al. (2020)	76.6	76.7
Jabreel Jabreel (2020)	77.9	77.7
Hassan et al. (2020)	77.4	78.6
Ours	60.58	62.63
Episource Miranda-Escalada et al. (2020)	61	58.2
Bigbyte Miranda-Escalada et al. (2020)	46.7	54.2
Rivera-Zavala and Martinez (2020)	14	16.5
Ruas, Neves, et al. (2020) and Ruas, Lamurias, et al. (2020)	5.8	6.1

## 5 | DISCUSSION

The evaluation and discussion of the results in terms of theoretical and practical/application are presented under two subsections.

### 5.1 | Theoretical

The research developed leads the way to the use of multilingual LMs for existing studies. This study presents also a comparison of these methods for EL using different methods.

For the EL task, three different methods and LMs were used to explore the state-of-the-art in terms of multilingual NLP. The number of studies using such different methods in the candidate generation phase and using LMs in the training phase is very low, even though LMs have been

rising in popularity in the last few years. Providing this diversity and performing analysis on different languages for the biomedical field contributes positively to the biomedical text-mining community.

## 5.2 | Practical

The success of our proposed EL framework was analysed for both Spanish and English in this study. For this, recall and accuracy values are calculated with pre-trained LMs by using Surface-based, ANN-search, and BLINK methods during the training phase. As a result of the evaluation of the model for both languages, an accuracy value of 76.75% for BC5CDR and 60.19% for Cantemist was obtained.

Compared to previous studies, the use of surface-based and ANN-search methods had a positive effect on our framework. In addition, performance comparisons of LMs were made with the use of many pre-trained LMs in our EL framework. When compared with previous studies, our framework outperformed some approaches but did not achieve state-of-the-art. The reason why the success here is lower than some approaches is that our framework has been developed to support multilingual data. The studies compared were proposed for only one language such as English and Spanish, while we used the same framework (with different LMs) for both languages.

## 6 | CONCLUSION AND FUTURE WORKS

In this study, a multilingual Bi-encoder EL framework is proposed. Analysis was performed for the BC5CDR and Cantemist datasets with the EL model. Before the analysis, the CG process was applied to both datasets. We explored the effect of candidate lists of different sizes. Surface-based, ANN-search, and BLINK methods were used in the training phase. In addition to these methods, pre-trained language models were used in the training phase. In the training process, the batch size was operated with two options, 16 and 48. When the test results are examined, the SapBERT model and BLINK method obtained the best results for the BC5CDR dataset, and the RBB-Es model and ANN-search method gave the most successful accuracy values for the Cantemist. Thus, the utilization of surface-based, ANN-search, and BLINK methods in candidate generation has been shown to have a positive effect on the success of LMs. For this, the proposed framework is also compared with previous studies. For both datasets, the framework performed better than some of the previous studies.

The use of the proposed biomedical EL framework for different languages will be analysed in future studies. In addition, the implementation of biomedical EL model with knowledge bases in other fields rather than biomedical tasks will be investigated.

### AUTHOR CONTRIBUTIONS

All authors have equally contributed to the development of this work.

### ACKNOWLEDGEMENTS

The authors would like to thank Prof. Dr. Murat Osman Unalir and Prof. Dr. Katja Hose.

### DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article

### ENDNOTES

- <sup>1</sup> <http://gerbil.aksu.org/gerbil/>
- <sup>2</sup> <https://www.ncbi.nlm.nih.gov/mesh/>
- <sup>3</sup> [https://eciemaps.mscbs.gob.es/ecieMaps/browser/index\\_o\\_3.html](https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_o_3.html)
- <sup>4</sup> <https://biocreative.bioinformatics.udel.edu/resources/corpora/biocreative-v-cdr-corpus/>
- <sup>5</sup> <https://zenodo.org/record/3978041#.Yc7NhWjP1Pa>
- <sup>6</sup> <https://github.com/allenai/scispacy>
- <sup>7</sup> <https://github.com/seatgeek/fuzzywuzzy>
- <sup>8</sup> <https://github.com/facebookresearch/BLINK>
- <sup>9</sup> <https://huggingface.co/bert-base-uncased>
- <sup>10</sup> <https://huggingface.co/dmis-lab/biobert-base-cased-v1.2>
- <sup>11</sup> <https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext>
- <sup>12</sup> <https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-es>
- <sup>13</sup> <https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es>
- <sup>14</sup> <https://huggingface.co/fvillena/bio-bert-base-spanish-wwm-uncased>

## REFERENCES

- Andrade, V. D. T., Ruas, P., & Couto, F. M. (2021). Named entity recognition and linking: A Portuguese and Spanish oncological parallel corpus. *bioRxiv*. <https://doi.org/10.1101/2021.09.16.460605>
- Angell, R., Monath, N., Mohan, S., Yadav, N., & McCallum, A. (2021). Clustering-based inference for biomedical entity linking. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2598–2608.
- Bhargav, G. P. S., Khandelwal, D., Dana, S., Garg, D., Kapanipathi, P., Roukos, S., Gray, A., & Subramaniam, L. V. (2022). Zero-shot entity linking with less data. In: *Findings of the Association for Computational Linguistics: NAACL 2022 Seattle, United States: Association for Computational Linguistics*, 1681–1697. <https://aclanthology.org/2022.findings-naacl.127>
- Bhowmik, R., Stratos, K., & de Melo, G. (2021). Fast and effective biomedical entity linking using a dual encoder. *arXiv Preprint arXiv:210305028*.
- Building Transformer-Based Entity Linking System izuna385 | Nerd For Tech. *Medium*. <https://medium.com/nerd-for-tech/building-bi-encoder-based-entity-linking-system-with-transformer-6c111d86500>
- Chen, L., Varoquaux, G., & Suchanek, F. M. (2020). A lightweight neural model for biomedical entity linking. *arXiv*. [www.aaii.org](http://www.aaii.org)
- Chen, S., Wang, J., Jiang, F., & Lin, C. Y. (2020). Improving entity linking by modeling latent entity type information. In: *AAAI 2020—34th AAAI Conference on Artificial Intelligence*.
- Chong, W. H., Lim, E. P., & Cohen, W. (2017). Collective entity linking in tweets over space and time. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.
- De Cao, N., Izcard, G., Riedel, S., & Petroni, F. (2020). Autoregressive entity retrieval. *arXiv*. <https://arxiv.org/abs/2010.00904v3>
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- Fang, Z., Li, Q., Cao, Y., Zhang, Z., Zhang, D., & Liu, Y. (2019). Joint entity linking with deep reinforcement learning. In: *The web conference 2019—proceedings of the world wide web conference, WWW*.
- Feng, Y., Zarrinkalam, F., Bagheri, E., Fani, H., & Al-Obeidat, F. (2018). Entity linking of tweets bas dominant entity candidates. *Social Network Analysis and Mining*, 8, 1–16.
- García-Pablos, A., Perez, N., & Cuadros, M. (2020). Vicomtech at cantemist 2020. In: *CEUR Workshop Proceedings*.
- Gillick, D., Kulkarni, S., Lansing, L., Presta, A., Baldridge, J., Ie, E., & Garcia-Olano, D. (2019). Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. <http://arxiv.org/abs/1909.10506>
- Hassan, F., Sanchez, D., & Domingo-Ferrer, J. (2020). Tumor entity recognition and coding for Spanish electronic health records. In: *CEUR Workshop Proceedings*.
- Jabreel, M. (2020). End-to-end neural coder for tumor named entity recognition. In: *CEUR Workshop Proceedings*.
- Kolitsas, N., Ganea, O. E., & Hofmann, T. (2018). End-to-end neural entity linking. In: *CoNLL 2018—22nd Conference on Computational Natural Language Learning, Proceedings*.
- Labusch, K., & Neudecker, C. (2020). Named entity disambiguation and linking on historic newspaper OCR with BERT. *CLEF*. <https://qurator.ai>
- Lange, L., Dai, X., Adel, H., & Strötgen, J. (2020). Nlnde at cantemist: Neural sequence labeling and parsing approaches for clinical concept extraction. In: *CEUR Workshop Proceedings*.
- Laskar, M. T. R., Chen, C., Johnston, J., Fu, X. Y., Bhushan, T. N. S., & Corston-Oliver, S. (2022). An auto encoder-based dimensionality reduction technique for efficient entity linking in business phone conversations. *SIGIR 22, New York, NY, USA: Association for Computing Machinery*, 3363–3367. <https://doi.org/10.1145/3477495.3536322>
- Laskar, M. T. R., Chen, C., Martsinovich, A., Johnston, J., Fu, X. Y., Shashi Bhushan, T. N., & Corston-Oliver, S. (2022). BLINK with elasticsearch for efficient entity linking in business conversations.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv Preprint arXiv:191013461*.
- Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., Davis, A. P., Mattingly, C. J., Wieggers, T. C., & Lu, Z. (2016). BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database*, 2016.
- Liu, F., Shareghi, E., Meng, Z., Basaldella, M., & Collier, N. (2020). Self-alignment pre-training for biomedical entity representations. <https://arxiv.org/abs/2010.11784>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pre-training approach. <http://arxiv.org/abs/1907.11692>
- Logeswaran, L., Chang, M. W., Lee, K., Toutanova, K., Devlin, J., & Lee, H. (2020). Zero-shot entity linking by reading entity descriptions. In: *ACL 2019—57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*.
- López-Úbeda, P., Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Urená-López, L. A. (2020). Extracting neoplasms morphology mentions in Spanish clinical cases throughword embeddings. In: *CEUR Workshop Proceedings*.
- Loureiro, D., & Jorge, A. M. (2020). MedLinker: Medical entity linking with neural representations and dictionary matching. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.
- Miranda-Escalada, A., Farré, E., & Krallinger, M. (2020). Named entity recognition, concept normalization and clinical coding: overview of the Cantemist track for cancer text Mining in Spanish, corpus, guidelines, methods and results. In: *IberLEF@ SEPLN*, 303–323.
- Murty, S., Verga, P., Vilnis, L., Radovanovic, I., & McCallum, A. (2018). Hierarchical losses and new resources for fine-grained entity typing and linking. In: *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 97–109). Association for Computational Linguistics. <https://aclanthology.org/P18-1010>
- Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: Fast and robust models for biomedical natural language processing. In: *Proceedings of the 18th BioNLP Workshop and Shared Task Florence, Italy: Association for Computational Linguistics*, 319–327. <https://www.aclweb.org/anthology/W19-5034>

- Ravi, M. P. K., Singh, K., Mulang, I. O., Shekarpour, S., Hoffart, J., & Lehmann, J. (2021). CHOLAN: A modular approach for neural entity linking on wikipedia and wikidata. In: EACL 2021—16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference.
- Rivera-Zavala, R., & Martinez, P. (2020). Deep neural model with contextualized-word embeddings for named entity recognition in Spanish clinical text. In: CEUR Workshop Proceedings.
- Ruas, P., Lamurias, A., & Couto, F. M. (2020). Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature. *Journal of Cheminformatics*, 12(1), 1–11.
- Ruas, P., Neves, A., Andrade, V. D. T., & Couto, F. M. (2020). Lasigebiotm at cantemist: Named entity recognition and normalization of tumour morphology entities and clinical coding of Spanish health-related documents. In: CEUR Workshop Proceedings.
- Schumacher, E., Mulyar, A., & Dredze, M. (2020). Clinical concept linking with contextualized neural representations. In: In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8585–8592.
- Sil, A., Kundu, G., Florian, R., & Hamza, W. (2018). Neural cross-lingual entity linking. In: 32nd AAAI Conference on Artificial Intelligence, AAAI 2018.
- Ujii, S., Iso, H., & Aramaki, E. (2021). Biomedical entity linking with contrastive context matching. arXiv Preprint arXiv:210607583.
- Ujii, S., Iso, H., Yada, S., Wakamiya, S., & Aramaki, E. (2021). End-to-end biomedical entity linking with span-based dictionary matching. arXiv Preprint arXiv:210410493.
- Vashishth, S., Joshi, R., Dutt, R., Newman-Griffis, D., & Rosé, C. P. (2020). MedType: Improving medical entity linking with semantic type prediction. <https://arxiv.org/abs/2005.00460>
- Wiatrak, M., & Iso-Sipila, J. (2020). Simple hierarchical multi-task neural end-to-end entity linking for biomedical text. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, 12–17.
- Wu, L., Petroni, F., Josifoski, M., Riedel, S., & Zettlemoyer, L. (2020). Scalable zero-shot entity linking with dense entity retrieval. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 6397–6407.
- Xiong, Y., Huang, Y., Chen, Q., Wang, X., Ni, Y., & Tang, B. (2020). A joint model for medical named entity recognition and normalization. In: CEUR Workshop Proceedings.
- Yamada, I., & Shindo, H. (2019). Neural attentive bag-of-entities model for text classification. In: CoNLL 2019—23rd Conference on Computational Natural Language Learning, Proceedings of the Conference.
- Yin, X., Huang, Y., Zhou, B., Li, A., Lan, L., & Jia, Y. (2019). Deep entity linking via eliminating semantic ambiguity with BERT. *IEEE Access*, 7, 169434–169445.
- Zhou, S., Rijnwani, S., Wieting, J., Carbonell, J., & Neubig, G. (2020). Improving candidate generation for low-resource cross-lingual entity linking. *Transactions of the Association for Computational Linguistics*, 1(8), 109–124. [https://doi.org/10.1162/tacl\\_a\\_00303](https://doi.org/10.1162/tacl_a_00303)
- Zhu, M., Celikkaya, B., Bhatia, P., & Reddy, C. K. (2020). LATTE: Latent type modeling for biomedical entity linking. In: AAAI 2020—34th AAAI Conference on Artificial Intelligence.

## AUTHOR BIOGRAPHIES

**Zekeriya Anil Guven** is currently an Assistant Professor at Izmir Bakircay University, Department of Computer Engineering. He obtained his Ph.D. in Computer Engineering at Ege University (Turkey). He worked as a research assistant at Ege University in Turkey for more than 4 years. He decided to start a Ph.D. in 2018 on creating new approaches within the question-answering and semantic domain with his supervisor Prof. Dr. Murat Osman Unalir. During his Ph.D., he got familiar with Machine Learning and Language Models and he published a good number of papers on topics related to how to apply language models to question-answering.

**Andre Lamurias** is an Assistant Professor at NOVA School Science and Technology, Department of Computer Science. His supervisors were Katja Hose and Thomas D. Nielsen at Aalborg University during the post-doc research. He's part of the DKWE group. Currently, he is working on the Microbial Dark Matter meets Data Science project, where they develop computer science methods to improve the recovery of genomes from environmental samples. In the past, He's worked at Priberam Labs and did my Ph.D. at LASIGE, Faculdade de Ciencias, Universidade de Lisboa.

**How to cite this article:** Guven, Z. A., & Lamurias, A. (2023). Multilingual bi-encoder models for biomedical entity linking. *Expert Systems*, 40(9), e13388. <https://doi.org/10.1111/exsy.13388>