CliniRes: Publicly Available Mapping of Clinical Lexical Resources

Elena Zotova[*†], **Montse Cuadros**[*], **German Rigau**[†‡]
[*]SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Mikeletegi Pasealekua 57, 20009, Donostia/San-Sebastián, Spain
{ezotova, mcuadros}@vicomtech.org
[†]Department of Languages and Computer Systems, University of the Basque Country (UPV-EHU)
Paseo Manuel de Lardizábal, 1, 20018, Donostia/San-Sebastián, Spain
[‡] HiTZ Basque Center for Language Technologies
german.rigau@ehu.eus

Abstract

This paper presents a human-readable resource for mapping identifiers from various clinical knowledge bases. This resource is a version of UMLS Metathesaurus enriched with WordNet 3.0 and 3.1 synsets, Wikidata items with their clinical identifiers, SNOMED CT to ICD-10 mapping and Spanish ICD-10 codes description. The main goal of the presented resource is to provide semantic interoperability across the clinical concepts from various knowledge bases and facilitate its integration into mapping tools. As a side effect, the mapping enriches already annotated medical corpora for entity recognition or entity linking tasks with new labels. We experiment with entity linking task, using a corpus annotated both manually and with the mapping method and demonstrate that a semi-automatic way of annotation may be used to create new labels. The resource is available in English and Spanish, although all languages of UMLS may be extracted. The new lexical resource is publicly available.

Keywords: clinical coding, entity linking, data interoperability, lexical resource, clinical NLP

1. Introduction

Annotation of training corpora for clinical coding, clinical concepts detection, entity disambiguation and entity linking tasks is very expensive in expertise and time. Considering that most clinical concepts are transferable across various knowledge bases, terminologies, lexicons and languages, we hypothesise that we can transfer one type of annotated code to another. For this purpose, we create CliniRes—a mapping human-readable resource to get related synonyms in various clinical lexicons so that target entities or concepts can be annotated in different clinical notations. This resource permits to align different types of clinical identifiers (IDs, codes) from different knowledge bases (KB) such as UMLS (Bodenreider, 2004), ICD-10 (World Health Organization (WHO), 2004), SNOMED CT (Donnelly et al., 2006) and others. Also, we enrich the resource with lexical resources, such as Wikidata items (Vrandečić and Krötzsch, 2014) and Wordnet synsets (Fellbaum, 2005). This allows to make clinical codes inter-operable, to use it in data annotation or other applications where clinical codes are involved. Moreover, it allows us to enrich manually annotated corpora with extra clinical codes and to obtain multilingual inter-operable corpora annotated with various coding notations. For instance, if we have a corpus annotated in UMLS codes we can map each code to SNOMED CT codes in order to derive automatically a new version of the corpus with SNOMED CT annotations. And vice versa, corpus annotated with SNOMED CT codes can be used to derive automatically new corpora annotated with UMLS codes, semantic types or groups.

This research is an extension of previously published works (Zotova et al., 2022, 2023a) where we described ClinIDMap[1], a clinical IDs mapping tool with the presented database integration. The functionality of the mapping application includes mapping of a source code (may be UMLS CUI, SNOMED CT, ICD-10-CM and ICD-10-PCS) to the clinical IDs and lexical resources such as Wikidata, Wikipedia and WordNet, including WordNet domains. It also allows updating of the database, as new versions of the ontologies are released yearly and Wikidata annotations are added regularly. The application is developed as REST API, accepts queries in JSON format, the database is indexed in Elasticsearch (Lucene). The source code is Dockerized, so it can be easily deployed.

The main contribution of this work is CliniRes— an alignment resource for mapping of clinical identifiers based on UMLS Metathesaurus, enriched with Spanish version of ICD-10, WordNet, and Wikidata items and annotations. It is ready to be integrated into a mapping application or be processed for synonym and annotation generation. This database is avalable under the licence of UMLS and SNOMED CT [2]. Also, this paper contributes to the methods of semi-automatic corpus annotation in clinical cor-

[1] https://github.com/Vicomtech/ClinIDMap

[2] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/license_agreement_snomed.html

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 163–172
20 May, 2024.

pora showing that the entity linking systems trained with the corpora annotated with this method, perform with the same accuracy as the systems trained with gold-standard corpora.

This paper is organized as follows. In Section 2 we briefly describe the background of clinical IDs mapping. Section 3 is dedicated to the knowledge bases and the mapping method. In Section 4 we give the details about the experiment with entity linking task done with the code mapping method. Finally, Section 5 concludes the work and discusses the future work in this topic.

2. Related Work

Two main parts of clinical codes mapping exist: (1) concept alignment, or ontology alignment (also known as ontology matching); (2) applications which use the concept mapping to enrich biomedical text or extract these concepts.

Ontology matching. The aim of ontology matching is to find semantically related entities in knowledge bases of different notations. For instance, the OAEI Campaign (Ontology Alignment Evaluation Initiative) [3] organizes every year an ontology matching evaluation shared task. The applied methods combine multiple strategies such as lexical matching, structural matching, logical reasoning, using background knowledge such as general purpose lexical resources, automatic translation and pretrained language models (Portisch et al., 2022; Wang et al., 2021). For instance, WordNet graphs were broadly used to clinical ontology matching (Lin and Sandkuhl, 2008) for measuring semantic similarity between the concepts (Pedersen et al., 2007). Some attempts to integrate WordNet to the clinical knowledge bases (Smith and Fellbaum, 2004) were made. Nevertheless, we should admit that the most of the studies are done with the resources in English. Novel machine learning and deep learning methods, such as generative adversarial networks, are also applied to ontology alignment (Chen et al., 2021; Kim et al., 2017).

Concept mapping applications. To our knowledge, there are not many open-source applications for concept mapping, especially for languages different from English. One of them is I-MAGIC, an application, implemented by US National Library of Medicine, that visualises clinical ID mappings. A demo version of the application is also available[4]. Using the rule-based SNOMED-CT to ICD-10-CM Mapping (Fung and Xu, 2012), the algorithm determines whether a valid ICD-10-CM code can be found based on the SNOMED-CT term and patient context information (age and gender). The application allows one to search a term in SNOMED-CT vocabulary, however, it is limited to a lexical match. The tool does not consider synonyms, nor other languages other than English, and its code is not open-source.

Most applications for clinical coding are designed to enrich clinical text with clinical concepts and relations. MetaMap [5](Aronson and Lang, 2010; Aronson, 2001) is an application for mapping biomedical text to the UMLS Metathesaurus or, equivalently, to discover UMLS concepts referred in the text. MetaMap uses a knowledge-intensive approach based on symbolic, NLP and computational-linguistic techniques to provide a link between the text of biomedical literature and the KB, including synonymy relationships, embedded in the Metathesaurus. The input of the application is English text. It is based on a lexical lookup of input words. Another example is CLAMP (Soysal et al., 2017), which takes two approaches: a machine learning using Conditional Random Field and a dictionary-based approach, which maps mentions to standardised ontologies. Apache cTAKES (Bodenreider, 2004) uses a dictionary look-up in unstructured clinical text, detects maned entities and each mention is mapped to a UMLS concept.

Some applications are also private, as they are developed by big tech companies. Spark NLP[6] and Amazon Comprehend Medical[7] offer service for mapping clinical findings to ICD-10-CM, SNOMED CT and other codes, in addition to entities and relations extraction.

There are also studies in topic of UMLS and Wikipedia connection, for instance, Rahimi et al. (2020) proposes to match UMLS concepts to Wikidata using a cross-lingual neural re-ranking model which is fine-tuned as a pair binary classification model aimed to categorize if a pair of texts is similar or not. As the UMLS descriptions are brief and the medical entity pages in Wikipedia provide detailed descriptions (also enriched with the Wikidata knowledge graph), they use the UMLS concept description to query the Wikidata entity aliases to retrieve the best matching Wikipedia pages.

3. Method

This section describes the knowledge bases and lexical resources used to create CliniRes and the method of mapping of clinical identifiers.

[3] http://oaei.ontologymatching.org/2023/

[4] https://imagic.nlm.nih.gov/imagic/code/map

[5] https://www.nlm.nih.gov/research/umls/implementation_resources/metamap.html

[6] https://demo.johnsnowlabs.com/healthcare/ER_ICD10_CM/

[7] https://aws.amazon.com/es/comprehend/medical/

3.1. Knowledge Bases

To interconnect the different identifiers from the knowledge bases of interest, we use the following existing KBs and mappings created by clinical experts.

UMLS Metathesaurus[8]. This database has been derived from the 2023AB UMLS Metathesaurus Files which contains approximately 3.15 million concepts from 220 source vocabularies, including ICD-10, MeSH, and SNOMED-CT, Hierarchies, definitions, and other relationships and attributes. The Metathesaurus is the biggest component of the UMLS. It is organised as a set of Concept Unique Identifiers (CUI), which links all the names from the source vocabularies with the same meaning (synonyms) in various languages. The Metathesaurus assigns several types of unique, permanent identifiers to the concepts and concept names it contains, in addition to retaining all identifiers present in the source vocabularies. The Metathesaurus concept structure includes concept names, their identifiers, and key characteristics of these concept names (e.g., language, vocabulary source, name type). The majority of the concept descriptions are short, less than one sentence. The entire concept structure appears in a single file in the Rich Release Format (MRCONSO.RRF). The distribution across the non-English languages is not proportional, as we can see in Table 1 there are significantly less concepts and synonyms in Spanish than in English.

The Semantic Network and Semantic Groups from UMLS is used to map semantic groups of each CUI. The Semantic Network consists of a set of broad subject categories, or Semantic Types, that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus. The concepts are also grouped according to the semantic types assigned to them. For certain purposes, however, an even smaller and coarser-grained set of semantic type groupings may be desirable. The following principles were used to design the groupings: semantic validity, parsimony, completeness, exclusivity, naturalness, and utility. The semantic groups provide a partition of the UMLS Metathesaurus for 99.5% of the concepts. Examples of semantic groups are Organisms, Anatomical structures, Biological functions, Chemicals, Events, Physical objects, Concepts or Ideas. These types are suitable for corpus annotation and training sequence labelling models and further linking to UMLS.

SNOMED-CT to ICD-10-CM Mapping[9]. The main purpose of the SNOMED-CT to ICD-10-CM mapping is to support semi-automated generation of ICD-10-CM codes from clinical data encoded in SNOMED-CT for reimbursement and statistical purposes. It is designed as a directed set of relationships from SNOMED-CT source concepts to ICD-10-CM target classification codes. This mapping is curated by trained terminology specialists, and it is more comprehensive than the Metathesaurus CUI linking. About a third of all active SNOMED-CT concepts are within the scope of the mapping, about 125,000 SNOMED-CT codes from the international version are mapped to ICD-10-CM codes. About 57,000 codes from the Spanish SNOMED-CT are included in the mapping (around 30% of all Spanish SNOMED-CT codes). Due to the differences in granularity, emphasis and organising principles between SNOMED-CT and ICD-10-CM, it is not always possible to have one-to-one mappings between a SNOMED-CT concept and an ICD-10-CM code, moreover, not all ICD-10-CM codes will appear as targets.

ICD-10-CM (International Statistical Classification of Diseases and Related Health Problems) establishes a standardized coding that allows the statistical analysis of mortality and morbidity of patients in healthcare services. The corresponding Spanish version is called CIE-10-ES and it consists of 100,158 codes, which are organised hierarchically. We use the official Spanish version of the CIE-10 from January 2022.

ICD-10-PCS (Procedure Coding System) is an international system of medical classification used for procedural coding, it consists of 80,266 codes, organised hierarchically. We use the official Spanish version of the ICD-10-PCS from March, 2022[10].

Wikidata[11] (Vrandečić and Krötzsch, 2014) is a free and open knowledge base that can be consulted and edited by both humans and machines. Wikidata is a central repository for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others. The Wikidata repository consists mainly of items, each with a label, a description and several aliases. Wikidata items related to clinical concepts are manually annotated with UMLS ID (CUI), Medical Subject Headings (MeSH) (Rogers, 1963), NCBI[12] (biomedical and genomic database) and other clinical taxonomies, so we can search items in Wikidata by these identifiers and extract the corresponding articles in all available languages.

WordNet 3.1[13] (Fellbaum, 2005) is the latest ver-

[8] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

[9] https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html

[10] https://www.sanidad.gob.es/fr/estadEstudios/estadisticas/normalizacion/home.htm

[11] https://www.wikidata.org

[12] https://www.ncbi.nlm.nih.gov/

[13] https://wordnet.princeton.edu/

Code	Num
Total rows	13,501,908
Unique CUIs	3,145,136
ENG CUIs	8,510,801
ENG Unique	3,144,365
SPA CUIs	1,371,376
SPA unique	491,713
CUIs with SNOMED mapping	359,757
SNOMED codes with CUI mapping	367,700

Table 1: Number of concepts in UMLS

sion of a lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked using conceptual-semantic and lexical relations. The WordNet also contains senses which are discrete representations of each aspect of the meaning of words. In the database, each sense has its unique sense key index (SKI) which provides a method for accessing synsets and word senses in the WordNet database. This version contains 155,327 words organised in 175,979 synsets for 207,016 word-sense pairs.

WordNet 3.0 [14] (Fellbaum, 2005) is the previous release of the lexical database. The WordNet 3.0 release has 117,798 nouns, 11,529 verbs, 22,479 adjectives, and 4,481 adverbs. The average noun has 1.23 senses, and the average verb has 2.16 senses. In total, there are 206,941 sense keys. As far as we know, no direct mapping between WN 3.0 and WN 3.1. exists, so we map the WordNet 3.1 to the WordNet 3.0 on the base sense key index. There are about 1,000 senses in all WordNet graph that cannot be transferred.

3.2. Code Mapping

To generate the enriched version of UMLS Thesaurus, we extracted all the Wikidata items annotated with UMLS CUI, NCBI, Wordnet 3.1, ICD-10 and SNOMED CT identifiers (updated on October 5, 2023). The Wikidata items are manually annotated by Wikidata experts. As shown in Table 2, there are about 860,000 items labelled with clinical IDs, the largest number is for UMLS CUI (about 86%), followed by NCBI IDs, and a smaller proportion, about 4% of items is annotated with WordNet synsets. Some of the Wikidata items are annotated with multiple WordNet synsets, up to six per item, in the table they are separated with a blank space. The less present identifiers in Wikidata items are SNOMED CT and ICD-10-PCS.

We merge all the tables databases described in Subsection 3.1 to the UMLS Thesaurus based on the CUI, SNOMED CT and ICD-10 codes. As a result, we obtain a large matrix of 37 columns and 15,945,228 rows where the first 18 columns are from the original UMLS table, and the rest of the columns are added through the SNOMED CT to ICD-10 mapping and through the Wikidata annotations. Figure 1 schematically depicts the method of mapping, where we can see how the identifiers are connected. Wikidata annotations and Semantic groups and types are connected through the UMLS CUI, the SNOMED CT to ICD-10 mapping is based on SNOMED CT codes, and ICD-10 codes presented in UMLS are extended to their Spanish definitions.

SNOMED CT to ICD-10 mapping adds more mappings between CUI and ICD-10, because the Spanish version of ICD-10 is not presented in UMLS, while SNOMED CT is presented. Spanish descriptions of ICD-10 codes are added, too. The codes extracted from Wikidata are marked as _WIKI, and this code may be different to the UMLS mapping, because of the manual expertise of the Wikidata editors. As the table is large, the definitions of the columns are detailed in Appendix A, Table 6. We also encourage the reader to see the sample of the resulting table in the GitHub repository[15].

This large matrix allows us to extract all related information based on any ID, WordNet sense or Wikidata item and then extend to more details. Wiki-

Source	Num
UMLS CUI	742,537
NCBI	623,397
Wordnet 3.1	31,897
Wordnet 3.0	31,884
MeSH	46,023
ICD10	7,650
ICD10CM	15,618
ICD10PCS	74
SNOMED_CT	1,579
Total Wikidata items	860,245

Table 2: Number of Wikidata items annotated with clinical codes extracted from Wikidata database.

[14] https://wordnetcode.princeton.edu/3.0

[15] https://github.com/Vicomtech/ClinIDMap/tree/master/LREC2024/samples

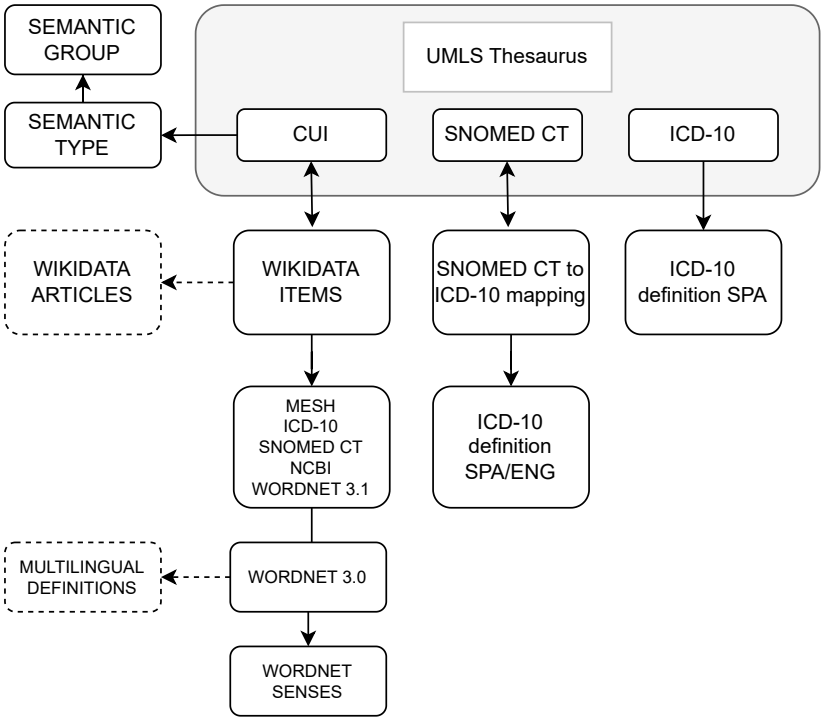

Figure 1: Scheme of clinical resources mapping, where they are connected by CUI, SNOMED CT or ICD-10 IDs. Dotted lines show a possible connection to more information about each concept.

data annotations allow us to derive further details, such as extracting more information about the item: its Wikipedia articles in all available languages, aliases, synonyms and other annotations. With that, short KB descriptions are extended to lexical resources, encyclopedia definitions and contexts. Moreover, Wikidata items exist, annotated both with CUI and ICD-10, so that we can consider it to be new code mapping. This provides 8,698 ICD-10 and 5,746 codes present in Wikidata annotations but not in UMLS mapping.

4. Entity Linking Experiment

To show that the codes in the clinical KBs are interoperable, we have already experimented with named entity recognition task to detect diagnosis and procedures or semantic types from UMLS notation in previous study, the work is described in our previous publication (Zotova et al., 2022). Now, we experiment with the entity linking task. Entity linking, or entity normalisation, is the key technology enabling semantic applications and informatics pipelines in the biomedical domain. This task aims to assign an identifier from clinical KB to the text span from clinical text written in natural language. In our case, both texts and KBs are in Spanish.

For the experiment, we use two annotated datasets of similar nature. These datasets consist of clinical case reports—a type of textual genre in medicine that describes a patient's medical history, symptoms, diagnosis, and treatment in detail. Both datasets are prepared for the entity linking task. Short descriptions of each corpus are below.

- MedProcNER (Lima-Lopez et al., 2023) is a collection of 1,000 clinical case reports written in Spanish, from which 750 documents are prepared for system training and 250 are for testing. In the train set, 4,857 text spans are manually annotated with SNOMED CT codes; 1,829 are unique, and some of the codes are composite, where two or more codes overlap. All codes are also annotated as procedures.
- CodiEsp 2020 (Miranda-Escalada et al., 2020) is a collection of 1,000 clinical case reports written in Spanish, where 750 documents are prepared for training purposes and 250 documents are reserved for testing. All documents were manually annotated by professional clinical coders with codes from the Spanish version of ICD-10 (procedure and diagnosis), and contain 3,427 unique codes, 2,557 of them are diagnoses and 870 codes are procedures. There are overlapping codes, too. The train set consists of 13,658 annotated text spans.

We evaluate unsupervised systems and use the whole training subset, without splitting it to the development subset. With the mapping tool, we transfer the gold-standard annotations (SNOMED CT

and ICD-10) to UMLS CUI and obtain corpora annotated with new codes. There are limitations of exact mapping because of the granularity of the ontologies and the annotations guides; some of SNOMED CT or ICD-10 codes have no direct mapping to another vocabulary. In MedProcNER corpus, from 4,857 entities there are 176 codes which cannot be transferred to CUI, in CodiEsp corpus 100% of ICD-10 codes are transferable to CUIs. At the same time, one SNOMED CT or ICD-10 code may be mapped to various CUIs, to simplify the experiment and make it comparable with the single SNOMED CT annotations we take only one CUI, the first in the database.
The example from MedProcNER corpus below shows a case of codes mapping from SNOMED CT to UMLS CUI. The term "Serologías específicas para Brucella" (*Specific serologies for Brucella*) is annotated with SNOMED CT code 104279004 "prueba de anticuerpos anti-Brucella" (*anti-Brucella antibody test*) and mapped to the CUI C0523269.

> Durante el ingreso se solicitan Hemocultivos: positivo para Brucella y **Serologías específicas para Brucella**: Rosa de Bengala +++; <...>
>
> *During admission, blood cultures are requested: positive for Brucella and **specific serologies for Brucella**: Rosa de Bengala +++;* <...>

Our approach to entity linking task is based on Semantic Text Similarity (STS) techniques. STS determines how similar two textual documents are by measuring their degree of semantic closeness. Semantic search is based on STS, allowing retrieval of relevant text results beyond mere lexical matching. The main concepts of semantic search are query, collection of documents (database), and degree of relevance between a query and retrieved documents. There are different methods of measuring the degree of relevance and relatedness of two pieces of text-—cosine distance, inner product, etc. We implement the following two unsupervised approaches.

- Statistical method with BM25 algorithm (Robertson et al., 1998). This function ranks a set of documents based on the query terms appearing in each document, regardless of their proximity, and it works on the concept of bag-of-words and TF-IDF. In a search time all the documents and a query are tokenized by white space and lower-cased.
- Transformer-based Semantic Search based on pre-trained Transformer models (Vaswani et al., 2017) to obtain the corresponding embeddings (multidimensional vectors) and compute the score using a similarity metric, in this case it is normalised inner product. This type of approach is implemented with HuggingFace (Wolf et al., 2020) and FAISS framework (Johnson et al., 2021).

The semantic search involves embedding all entries (sentences, documents, or, in this case, KB code descriptions) into a single vector space. At search time, the query, represented by a texts span from a clinical narrative, is also embedded into the same vector space. This allows a direct comparison of vectors using cosine distance between the vectors. The closest document, in our case, CUI, ICD-10 or SNOMED CT description, is linked to our query and the code assigned to this document is returned as a prediction of linked identifier. In this case we encode the texts with SapBERT-XLM-R-large model (Liu et al., 2021), as it is a XLM-RoBERTa-large model (Conneau et al., 2019) trained on the descriptions from UMLS Thesaurus and brings the domain knowledge to the entity linking system. An embedding dimension of 1024 is enough to encode all the terminology and corpus entities without truncation. [CLS] token of the transformer's architecture is used for the vector representation of a text.
To reduce the search space and make it comparable to the original task, developed with clinical experts (search in SNOMED CT and ICD-10 databases in Spanish), we filter UMLS Thesaurus to Spanish terms only, lowercase the descriptions, and obtain about 1.28 million Spanish synonyms from different vocabularies related to approximately 490,000 unique CUIs, still being the largest of three collection of documents to search. Table 3 shows the exact number of the KBs for entity linking, where we can see that the size of SNOMED CT and ICD-10 is comparable, but UMLS is much larger.

KB	Num
UMLS CUI	1,283,535
SNOMED_CT	242,228
ICD10CM+ICD10PCS	180,424

Table 3: Size of knowledge bases for entity linking.

Annotation	Method	Accuracy
SNOMED CT (gold)	SapBERT	43.44
	BM25	19.96
UMLS CUI (map)	SapBERT	34.78
	BM25	24.27

Table 4: Performance of semantic search approach on MedProcNER corpus.

The results of the search methods, as depicted in Tables 4 for MedProcNer corpus and 5 for CodiEsp corpus, are comparable across the gold-standard and mapped corpus, regardless of different size

Annotation	Method	Accuracy
ICD-10 (gold)	SapBERT	29.62
	BM25	10.57
UMLS CUI (map)	SapBERT	28.76
	BM25	25.75

Table 5: Performance of semantic search approach on CodiEsp-2020 corpus.

of the vector space and different coding systems. Transformer based system in case of MedProcNer performs 9 points better on the gold-standard corpus, but we should admit that the database to search in is much bigger— 1,3 millions entries in UMLS versus 242,000 in SNOMED CT. The BM25 model perform 4 point better in case of mapped labels (CUI). It can be explained with the fact that UMLS contains various vocabularies which could be closer lexically to the corpus. As the STS method based on distance measure, it highly depends on the number of documents in the collection to search in, the less is the collection, the easier is the retrieving task.

Entity linking on CodiEsp also shows very similar result in both systems, we observe less than one point difference in transformer-based semantic search with SapBERT model. The BM25 method, which performs in the gold-standard corpus worse. It can be explained by the broader variety of synonyms in UMLS, which represents better the lexical content of the corpus.

These STS models do not perform with the state of the art accuracy scores, and we do not compete with these scores, moreover the original task designed for these corpora, based first on the named entity recognition task, and then, the recognised entities should be linked to the KBs. We skip the named entity recognition step and experiment with entity linking only, where the exact text span is already known and manually annotated. That is why our results are nor comparable to the previous studies. In summary, we can conclude that the new models and corpora are quite inter-operable with respect the different coding systems.

5. Conclusions and Future Wok

In this paper we described a human-readable database for interoperability between clinical concepts of various knowledge bases. For this, we explained how we enriched the UMLS Thesaurus with Wikidata items, WordNet senses and SNOMED CT to ICD-10 mappings, we also added Spanish definitions of ICD-10 codes present in this resource. This resource is ready to be integrated into any application or be used for clinical synonyms generated. We demonstrate the use of the resulting resource in the mapping tool which is publicly available as opens-source, both the code and the demo-version of the API.

We experimented with entity linking task on the corpora annotated with different coding systems, showing that the labels obtained with the mapping method can be used to build new entity linking or information retrieval systems, as the results of the entity linking systems are comparable.

As future work we see the experimentation on deep learning methods and large language models for mapping between English and multilingual concepts, paying special attention to underrepresented in UMLS languages. As we mentioned in Subsection 3.1, the distribution of the UMLS concepts and vocabularies are disproportional for non-English languages. We see the opportunity to use the novel approaches to contribute to creation of clinical terminologies and create background knowledge for concept matching in multilingual setting. We also plan to experiment with more methods for entity linking in clinical databases and lexical resources.

6. Bibliographical References

Alan Aronson. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 2001:17–21.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Jiaoyan Chen, Ernesto Jiménez-Ruiz, Ian Horrocks, Denvar Antonyrajah, Ali Hadian, and Jaehun Lee. 2021. Augmenting ontology alignment by semantic embedding and distant supervision. In *European Semantic Web Conference*, pages 392–408. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Kevin Donnelly et al. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279.

Christiane Fellbaum. 2005. Wordnet and wordnets. In *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford.

Kin Wah Fung and Junchuan Xu. 2012. Synergism between the Mapping Projects from SNOMED CT to ICD-10 and ICD-10-CM. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:218–227.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Sun Kim, Nicolas Fiorini, W. John Wilbur, and Zhiyong Lu. 2017. Bridging the Gap: Incorporating a Semantic Similarity Measure for Effectively Mapping PubMed Queries to Documents. *Journal of Biomedical Informatics*, 75:122–127.

Salvador Lima-Lopez, Eulaia Farre-Maduell, Luis Gasco, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2023. Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. In *Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum*.

Feiyu Lin and Kurt Sandkuhl. 2008. A survey of exploiting wordnet in ontology matching. *International Federation for Information Processing Digital Library; ARTIFICIAL INTELLIGENCE IN THEORY AND PRACTICE II ;*, 276:341–350.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 565–574. Association for Computational Linguistics.

George A Miller. 1998. *WordNet: An Electronic Lexical Database*. MIT press.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. 2007. Measures of Semantic Similarity and Relatedness in The Biomedical FDomain. *Journal of biomedical informatics*, 40(3):288–299.

Jan Portisch, Michael Hladik, and Heiko Paulheim. 2022. Background knowledge in ontology matching: A survey. *Semantic Web*, pages 1–55.

Afshin Rahimi, Timothy Baldwin, and Karin Verspoor. 2020. WikiUMLS: Aligning UMLS to Wikipedia via cross-lingual neural ranking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5957–5962, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. 1998. Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proceedings of The Seventh Text REtrieval Conference, TREC 1998, Gaithersburg, Maryland, USA, November 9-11, 1998*, volume 500-242 of *NIST Special Publication*, pages 199–210. National Institute of Standards and Technology (NIST).

Frank Rogers. 1963. Medical subject headings. *Bulletin of the Medical Library Association*, 51:114–116.

Barry Smith and Christiane Fellbaum. 2004. Medical WordNet: A New Methodology for the Construction and Validation of Information Resources for Consumer Health. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, page 371–es, USA. Association for Computational Linguistics.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei V. S. Pakhomov, Hongfang Liu, and Hua Xu. 2017. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association : JAMIA*, 25:331 – 336.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10):78–85.

Peng Wang, Yun Zhong Hu, Shaochen Bai, and Shiyi Zou. 2021. Matching biomedical ontologies: Construction of matching clues and systematic evaluation of different combinations of matchers. *JMIR Medical Informatics*, 9.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger,

Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

World Health Organization (WHO). 2004. *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*, 2nd ed edition. World Health Organization.

Elena Zotova, Montse Cuadros, and German Rigau. 2022. ClinIDMap: Towards a clinical IDs mapping for data interoperability. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3661–3669, Marseille, France. European Language Resources Association.

Elena Zotova, Montse Cuadros, and German Rigau. 2023a. Towards the integration of wordnet into clinidmap. In *Proceedings of the 12th Global Wordnet Conference*, pages 352–362.

Elena Zotova, Aitor Garcia-Pablos, Montse Cuadros, and German Rigau. 2023b. VICOMTECH at MedProcNER 2023: Transformers-based Sequence-labelling and Cross-encoding for Entity Detection and Normalisation in Spanish Clinical Texts. In *Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum*.

A. Appendix A

Column Name	Description
CUI	Concept unique identifier
LAT	Language of terms
TS	Term status
LUI	Lexical (term) Unique Identifiers
STT	String Type
SUI	String Unique Identifiers
ISPREF	Atom status - preferred (Y) or not (N)
AUI	Atom identifier
SAUI	Source asserted atom identifier
SCUI	Source asserted concept identifier
SDUI	Source asserted descriptor identifier
SAB	Abbreviated source name, for example, SNOMEDCT_US or ICD10CM
TTY	Abbreviation for term type in source vocabulary
CODE	Most useful source asserted identifier
STR	String
SRL	Source restriction level
SUPPRESS	Suppressible flag
CVF	Content View Flag
ICD10CM_SPA	ICD-10-CM definition in Spanish
ICD10PCS_SPA	ICD-10-PCS definition in Spanish
SNOMEDCT2ICD10	ICD-10 identifier, mapped to SNOMED CT identifier
SNOMEDCT2ICD10_ENG	ICD-10 definition in English from SNOMED CT to ICD-10 mapping
WIKIDATA	Wikidata item identifier
MESH_WIKI	MeSH identifier extracted from WIkidata
SNOMED_CT_WIKI	SNOMED CT identifier extracted from WIkidata
ICD10_WIKI	ICD-10 identifier extracted from WIkidata
ICD10CM_WIKI	ICD-10-CM identifier extracted from WIkidata
ICD10PCS_WIKI	ICD-10-PCS identifier extracted from WIkidata
NCBI_WIKI	NCBI identifier extracted from WIkidata
WN31	WordNet 3.1 identifiers, blank space separated
WN30	WordNet 3.0 identifiers, blank space separated
WN_SENSE	WordNet synsets, blank space separated
TUI	Semantic Type Unique Identifier
SEMTYPE	Name of the Semantic Type
SEMGROUP	Semantic group abbreviation
DEF	Definition of the Semantic Group

Table 6: Description of the columns in the lexical resource.