Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3661–3669
Marseille, 20-25 June 2022

ClinIDMap: Towards a Clinical ID Mapping for Data Interoperability

Elena Zotova[12], **Montse Cuadros**[1], **German Rigau**[23]
[1] SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA),
Mikeletegi Pasealekua 57, Donostia/San-Sebastián, 20009, Spain,
[2] Department of Languages and Computer Systems, University of the Basque Country (UPV-EHU)
[3] HiTZ Basque Center for Language Technologies
Paseo Manuel de Lardizábal, 1, 20018, Donostia/San-Sebastián, Spain
{ezotova, mcuadros}@vicomtech.org, german.rigau@ehu.eus

Abstract

This paper presents ClinIDMap, a tool for mapping identifiers between clinical ontologies and lexical resources. ClinIDMap interlinks identifiers from UMLS, SMOMED-CT, ICD-10 and the corresponding Wikipedia articles for concepts from the UMLS Metathesaurus. Our main goal is to provide semantic interoperability across the clinical concepts from various knowledge bases. As a side effect, the mapping enriches already annotated medical corpora in multiple languages with new labels. For instance, spans manually annotated with IDs from UMLS can be annotated with Semantic Types and Groups, and its corresponding SNOMED-CT and ICD-10 IDs. We also experiment with sequence labelling models for detecting Diagnosis and Procedures concepts and for detecting UMLS Semantic Groups trained on Spanish, English, and bilingual corpora obtained with the new mapping procedure. The ClinIDMap tool is publicly available.

Keywords: Clinical Coding, Concept Classification, Sequence Labeling, Interoperability

1. Introduction

Sequence labeling (SL)—a task which assigns a class or label to each span of text in a given input sequence—is a crucial first step to a number of natural language processing (NLP) tasks in the clinical domain, such as relation extraction, entity linking, and automatic clinical coding. In particular, the automatic clinical coding task consists of detecting an entity span, classifying the span, and linking it to a clinical knowledge base. This paper focuses on the first part of a clinical coding pipeline, entity detection and recognition.

Building supervised machine learning and deep learning models for this task require a large amount of manually annotated data. Manual annotation of a particular textual corpora is one of the most valuable and expensive part when building a new NLP system. The data in the biomedical domain is especially difficult to obtain because of two main reasons. First, clinical reports and other medical documents usually have privacy issues. Even after de-identification, it is complicated to obtain free access to this kind of medical data and, for this reason, there are very few clinical corpora freely available for research. Second, the process of manual annotation requires high level expertise which makes the use of crowd-sourcing platforms almost impossible, making it more expensive than general purpose NLP-corpora. The problem is even more difficult in multilingual setting since very few resources are available for languages other than English.

Fortunately, most clinical concepts are transferable across the various knowledge bases and languages. The goal of our mapping tool is 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) and SNOMED-CT (Donnelly and others, 2006). The alignment uses the actual IDs of the KBs from the official mapping resources developed by SNOMED-CT and UMLS authors.

The alignment allows us to enrich manually annotated corpora with extra clinical codes and to obtain multilingual inter-operable corpora annotated with various coding systems. For instance, if we have a corpus annotated in UMLS codes we can map each code to ICD-10-CM and ICD-10-PSC codes in order to derive automatically a new version of the corpus for training a new SL system. And vise versa, corpus annotated with ICD-10 codes can be used to derive automatically new corpora annotated with UMLS codes, semantic types or groups. Moreover, our tool enriches the annotated concepts with multilingual terms and descriptions of its available Wikipedia articles, which allows us to expand brief code descriptions to detailed information in multiple languages. For instance, a Spanish sentence annotated with a UMLS code is given in Example 1. The code C0011860 corresponding to the Spanish term *diabetes tipo 2* can be mapped to the SNOMED CT code 44054006, to the ICD-10-CM E11.9 code, and to the corresponding Wikipedia articles in 51 languages.

Example 1. La paciente presentaba como antecedentes personales hipertensión y **diabetes tipo 2** (C0011860) controladas mediante tratamiento médico convencional.

Translation: *The patient had a history of hypertension and* ***type 2 diabetes*** *controlled by conventional medical* *treatment.*

Concretely, in this paper we present a tool for mapping UMLS, SNOMED-CT, ICD-10 and Wikidata for Spanish and English. Using this tool, we derive multiple datasets annotated with different coding systems on the base of existing corpora (see Subsection 3.3). The resulting annotated corpus is prepared for training SL models to classify various concepts: semantic groups (high-level categories from UMLS notation), diagnosis and procedures (categories from ICD-10 notation), or unique codes. Finally, we provide a comparative study of classification models for clinical concept recognition trained on the gold-standard corpus and corpus annotated with the mapping method. The tool is publicly available[1].

The paper is organized as follows. Section 2 describes previous efforts mapping clinical codes. Section 3 describes the knowledge bases and mapping schemes used in our study; Section 4 presents the mapping algorithm and the experiments of sequence labeling; Section 5 report the results of the experiments and conclude that the mapping method can be useful for interoperable sequence labeling tasks in the clinical domain. Finally, in Section 6 we summarize our main contributions and future work.

2. Background

There are two main parts of clinical codes mapping: (1) concept alignment, or ontology alignment (also known as ontology matching); (2) applications which use the resulting concept mapping to process biomedical text.

Ontology matching is usually performed to find semantically related entities in different knowledge bases (KB). For instance, OAEI Campaign (Ontology Alignment Evaluation Initiative) [2] organizes a every year an ontology matching evaluation shared task. The applied methods combine multiple strategies such as lexical matching, structural matching and logical reasoning (Ochieng and Kyanda, 2018). Novel machine learning and deep learning methods are also applied to ontology alignment (Chen et al., 2021). In this research we use already aligned clinical KBs.

The majority of applications are designed to enrich clinical text with clinical concepts and relations. MetaMap (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.

I-MAGIC is an application, implemented by US National Library of Medicine, that visualises clinical IDs mappings. A demo version of the application is also available[3]. 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 to search a term in SNOMED-CT. However, it is limited to a literal search. The tool does not consider synonyms, nor other language than English.

(Rahimi et al., 2020) proposes to match UMLS concepts to wikidata using a cross-lingual neural reranking model which is based on a pretrained contextual encoding. 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.

Instead, our approach exploits available manual mappings between the different lexical resources.

3. Resources and Data

3.1. Knowledge Bases

The following medical knowledge bases are used to build ClinIDMap. Each of them consists of a set of IDs in alphanumeric format and a brief description.

The UMLS, or Unified Medical Language System[4], is a set of files and software that brings together 102 health and biomedical vocabularies and standards and includes 4 million terms to enable interoperability between computer systems. UMLS consists of three parts: the Metathesaurus, a Semantic Network and the SPECIALIST Lexicon. This database is our main source of mapping information.

Spanish SNOMED-CT[5] is the Spanish translation of SNOMED-CT. It includes the National Extension for Spain, updated and maintained by the SNOMED CT National Reference Centre for Spain, Ministry of Health, Consumer Affairs and Social Welfare. Spanish SNOMED-CT contains 199,961 unique codes.

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. It consists of 99.000 codes which are organized hierarchically. The corresponding Spanish version is

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

[2] `http://oaei.ontologymatching.org/2021/`

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

[4] `https://www.nlm.nih.gov/research/umls/index.html`

[5] `https://www.mscbs.gob.es/profesionales/hcdsns/areaRecursosSem/snomed-ct/areaDescarga.htm`

called CIE-10-ES. We use the official Spanish version of the CIE-10 from July 2020[6].

ICD-10-PCS (Procedure Coding System)[7] is an international system of medical classification used for procedural coding, it consists of 80.000 codes, organized hierarchically. ICD-10-PCS is a result of separation of a chapter from ICD-9 which contained procedures codification. We use the official Spanish version of the ICD-10-PCS from January 2020.

Wikidata[8] (Vrandečić and Krötzsch, 2014) is a free and open knowledge base that can be consulted and edited by both humans and machines. Wikidata acts as 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 one having a label, a description and a number of aliases. Wikidata items related to clinical concepts are annotated with UMLS ID (CUI), Medical Subject Headings (MeSH) (Rogers, 1963) and other clinical taxonomies, so we can search items in Wikidata by these identifiers and extract the corresponding articles in all available languages.

Wikipedia[9] is used in our application as a multilingual online encyclopedia of clinical concepts. Wikipedia provides extensive description of clinical concepts in many languages.

3.2. Mappings

To interconnect the different identifiers from the knowledge bases of interest we use already existing mappings created by clinical experts. The mapping schemes are the following:

UMLS Metathesaurus[10]. The main database of our tool has been derived from the 2021AB UMLS Metathesaurus Files which contains approximately 4.54 million concepts from 220 source vocabularies, including ICD-10-CM, 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 all of the source vocabularies that have the same meaning (synonyms). A single CUI can have several definitions in different languages. The Metathesaurus assigns several types of unique, permanent identifiers to the concepts and concept names it contains, in addition to retaining all identifiers that are 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 entire concept structure appears in a single file in the Rich Release Format (MRCONSO.RRF).

	Unique codes
SNOMED-CT to ICD-10	125,823
SNOMED-CT Spanish to ICD-10	57,393
SNOMED-CT International	485,977

Table 1: Number of SNOMED-CT codes in SNOMED-CT to ICD-10CM mapping and the total number of concepts in SNOMED-CT.

The Semantic Network from UMLS is used to map semantic groups of each CUI. Examples of the semantic groups are Organisms, Anatomical structures, Biologic function, Chemicals, Events, Physical objects, Concepts or Ideas. These types are suitable for corpus annotation and training sequence labeling models and further linking to UMLS.

SNOMED-CT to ICD-10-CM Mapping[11]. 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. As shown in Table 1), about a third part of all active SNOMED-CT concepts are within the scope of the mapping, which means that 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 organizing 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. In addition, not all ICD-10-CM codes will appear as targets.

3.3. Corpora

We apply our mapping approach to various corpora. The following clinical datasets annotated with ICD-10 and UMLS CUI codes in Spanish and English have been used.

CodiEsp. In 2020, Task 1 of the CLEF e-Health included the CodiEsp track, which was the first shared task consisting of the automatic coding of clinical cases

[6] https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_10_mc.html

[7] https://www.cms.gov/Medicare/Coding/ICD10/2020-ICD-10-PCS

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

[9] https://www.wikipedia.org/

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

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

	Diagnosis	Procedures	Sentences
Train	17,408	6,235	9,762
Dev	1,584	630	931
Test	6,131	2,289	3,599
Total	24,125	8,777	14,292

Table 2: Number of tokens annotated in CodiEsp corpus.

in Spanish (Miranda-Escalada et al., 2020). CodiEsp is a corpus with 1.000 samples of clinical cases, manually curated by the organizers of the task. The CodiEsp-X subtask required the prediction of both ICD-10-CM-ES and ICD-10-PCS-ES codes together with the exact reference to the text segment that served as justification for the assignment of such codes. This last subtask became a sequence labeling and normalization task. In this task, the systems are expected to detect the span and predict if the detected term is a Diagnosis (Spanish *diagnóstico*) or a Proceeding (Spanish *procedimiento*).

We use the CodiEsp original train-test split: 500 documents for train set, and 250 documents for development and test sets respectively. We keep the test set intact for reference evaluation, but we combine train and development sets, obtaining 750 documents, 10% of this combined set is reserved for development. In Table 2 the dataset split and annotated tokens are shown.

E3C Corpus. E3C (Magnini et al., 2020) is a freely available multilingual corpus in English, French, Italian, Spanish, and Basque of semantically annotated clinical narratives. It consists of two types of annotations: clinical entities (e.g., pathologies, drugs, anatomy, etc.) and temporal information and factuality (e.g., events). The E3C corpus is organised into three layers: Layer 1 is manually annotated corpus which consists of about 25,000 tokens per language, Layer 2 is an automatically annotated corpus (silver standard) and Layer 3 is a collection of non-annotated clinical texts. We use the manually annotated Layer 1 with CUIs and semantic types in Spanish only.

CT-EBM-SP (Campillos-Llanos, 2019) is a collection of 1,200 texts in Spanish about clinical trials studies and clinical trials announcements: 500 abstracts from journals published under a Creative Commons license and 700 clinical trials. The corpus is annotated with UMLS CUI, semantic types and semantic groups.

MANTRA corpus (Kors et al., 2015) consists of text segments from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. The Spanish part does not contain patient claims, it consists of 100 MedLine titles and 100 drug labels. Is is manually annotated with the biomedical concepts and CUIs, based on a subset of the UMLS and covering a wide range of semantic groups. We use the Spanish part of the corpus.

Ontology	Unique IDs	Codes per CUI
SNOMED-CT	489,141	0.84
ICD-10-CM	95,671	1.08
ICD-10-PCS	190,673	1.00
MeSH	347,565	1.30
UMLS CUI	1,106,486	

Table 3: Number of mapped identifiers in English UMLS Metathesaurus.

MedMentions (Mohan and Li, 2019) is a large English dataset identifying and linking entity mentions in PubMed abstracts to specific UMLS concepts. It consists of over 4,000 abstracts and over 350,000 linked mentions in English.

4. Methodology

In this section we describe (1) the methodology of the mapping tool and corpus enrichment and (2) the methodology of the classification models trained with this corpus.

4.1. Code Mapping

Our main goal is to exploit the existing mappings to enrich the annotated corpus with additional annotations from different lexical resources. The reference ID (concept) is usually annotated manually.

We combine the following mapping schemes manually created: UMLS Metathesaurus and SNOMED-CT to ICD-10 (see Subsection 3.2). Our tool allows searching in the following lexicons: SNOMED-CT, ICD-10-CM, ICD-10-PCS and UMLS CUIs.

The following steps are executed by our application.

1) Extract all CUIs mapped to SNOMED-CT, ICD-10-CM and ICD-10-PCS from the UMLS Metathesaurus. We use the English concepts of UMLS as it contains the largest representation of concepts. Table 3 presents the number of unique concepts present within the Metathesaurus. Every CUI also has synonyms of different types. For example, PN (Metathesaurus Preferred Name), or CD (Clinical Drug), or AC (Activities), etc. On average a unique CUI has 2.79 synonyms. In addition, one CUI can be mapped to various SNOMED-CT and ICD-10 codes, and vise versa, one ICD-10 code can be mapped to multiple CUIs and various SNOMED-CT IDs.

2) Extract the ICD-10-CM codes from the SNOMED CT to ICD-10-CM Mapping.

3) Extract the definitions of the ICD-10-CM codes from the Spanish version of ICD-10-CM. The ICD-10-PSC codes are extracted from the UMLS Metathesaurus as the ICD-10-PCS codes are not present in the SNOMED-CT to ICD-10 mapping. Finally, the definitions are extracted from the Spanish version of the ICD-10-PSC.

4) Extract all Wikidata items that contain the given CUI and corresponding MeSH codes by using the Wikidata Query Service. We use the Wikibase API to retrieve the Wikidata page of each CUI and the Wikipedia articles corresponding to these Wikidata items. As Wikidata is constantly being updated, by December 2020, we gathered 27.847 Wikidata items annotated with a CUI and 36.832 items annotated with MeSH codes. The tool also has the functionality of updating the list of Wikidata items with CUIs by demand. Wikipedia and Wikidata alignment may help for code description enriching, search and entity linking algorithms, or expand the language support for the UMLS terminology, which allows to extend the term descriptions in various languages.

The output of the mapping application is a JSON formatted list of IDs from the lexicons and the corresponding descriptions of each code in English and Spanish. An example of the output of final mapping is shown in Figure 1. In this example the source ID is UMLS CUI C0153458, which has six synonymous alias in UMLS ontology, two SNOMED-CT codes in Spanish version SNOMED-CT, two ICD-10-CM codes, and one Wikidata item annotated with this CUI, which links to 55 language specific Wikipedia articles.

We apply the mapping method to the corpora presented in Section 3.3 to derive new versions also annotated with ICD-10 codes. For instance, the CodiEsp corpus originally annotated with CUIs and semantic types from UMLS is used to derive a new corpora annotated with both, ICD-10 and UMLS CUI concepts. Although not all CUIs are mapped to ICD-10 identifiers, all ICD-10 codes from the corpora are mapped to UMLS CUIs. Table 4 shows that depending on the corpora, it is possible to map from 10% to 50% of the CUIs to ICD-10 codes.

Table 5 shows an example from the CodiEsp corpus annotated with the mapping method. This corpus is manually annotated with ICD-10 Diagnosis / Proceeding. ICD-10 codes are mapped to UMLS CUIs and their associated Semantic Types and Semantic Groups from the Semantic Network.

All code have been implemented in Python 3 and it is publicly available.

4.2. Sequence Labelling Models

We train two types of models: (1) the first one for the classification of Diagnosis and Procedures (SL-DP), according to ICD-10-CM and ICD-10-PSC notation, and (2) a second model for labelling the UMLS Semantic Groups[12] (SL-SG), such as Anatomy, Disorder, Procedure, Chemical, etc. We train the sequence labeling models to predict the clinical concepts using BIO labels.

We fine-tuned a multilingual BERT-base for Spanish and the bilingual corpora, and BERT-base for the English corpora (Devlin et al., 2019). We perform the experiments with the mentioned datasets (Subsection 3.3). All models have been developed using HuggingFace 4.4 library[13] and trained on a single GPU Nvidia GeForce RTX 2080 Ti with 11 Gb RAM, with learning rate 2e-5, batch size 8, and 100 epochs.

4.2.1. Sequence Labeling: Diagnosis and Procedures

Four SL-DP models are trained with the following corpora:

- CodiEsp is a manually annotated corpus with Diagnosis and Procedures. The model trained in this corpus is our reference model.
- Combined Corpus (Combined-es). After applying the mapping on the Spanish corpora (CodiEsp, MANTRA, E3C, CT-EBM-SP, see Subsection 3.3), we obtain 3,500 extra sentences and about 11,000 annotated tokens in Spanish to add to the reference corpus. We annotate all the corpora with ICD-10-CM and ICD-10-PCS and derive whether it is a Diagnosis or a Procedure concept. 80% of the corpus is used for training 10% development and 10% for test. We also test the model on the CodiEsp test set.
- MedMentions corpus is already annotated with CUIs. We map them to ICD-10-CM and ICD-10-PSC and obtain a corpus annotated with Diagnosis and Procedures: 48,435 sentences and 39,199 tokens annotated as Diagnosis and 565 tokens annotated as Procedures. The corpus is highly unbalanced because the purpose of this corpus was not to annotate Procedures.
- Bilingual Corpus. We also experiment training a bilingual model combining the Spanish and English corpora (using Combined Corpus and MedMentions corpus excluding the CodiEsp and MedMentions test sets for monolingual testing). The resulting bilingual corpus has 62,689 sentences. 10% of the corpus is reserved for development and 10% for the test set; 68,277 tokens are annotated as Diagnosis and 7,156 tokens are annotated as Procedures. The model is also tested on CodiEsp and MedMentions test sets.

4.2.2. Sequence Labeling: Semantic Groups

We also experiment with the corpus annotated by mapping the ICD-10 codes to CUIs. We train a SL model for labelling the UMLS Semantic Groups (SL-SG). After mapping the codes from the CodiEsp corpus to the UMLS CUIs, the semantic groups of each CUI are derived, using the Semantic Network from UMLS. As one ICD-10 code belongs to various semantic groups, we select the CUIs that occur most frequently in the

[12] https://www.nlm.nih.gov/research/umls/new_users/online_learning/SEM_003.html

[13] https://huggingface.co/

Corpus	Lang	Source	Annotation Type	Tokens	Tokens CUI	Tokens ICD-10
CT-EBM-SP	es	Clinical trials studies and announcements	CUI, semantic group, semantic type, PoS	141.158	23.264	9.222
MANTRA	es	Medline abstract titles, drug labels	CUI, semantic types	3.492	1.058	392
E3C-Corpus (Layer 1)	es	Clinical narratives	CUI, semantic types, temporal information and factuality	28.815	2.268	1.573
CodiEsp	es	Clinical narratives	ICD-10, Diagnosis, Procedure	401.010	32.902	32.902
MedMentions	en	Biomedical papers	CUI, semantic type	1.258.847	540.138	39.764

Table 4: Corpora annotated with CUIs and ICD-10 (both CM and PCS) using our mapping procedure.

Token	ICD-10	Diagnosis / Procedures	CUI	Semantic Type	Semantic Group
Se					
realiza					
una					
radiografía	BT04ZZZ	B-PROCEDIMIENTO	C2456752	T060	B-PROC
reno-vesical	BT04ZZZ	I-PROCEDIMIENTO	C2456752	T060	I-PROC
mostrando					
múltiples					
cálculos	N20.0	B-DIAGNOSTICO	C0156257 C0268722	T047	B-DISO
radiopacos					
localizados					
en					
el					
riñón	N20.0	B-DIAGNOSTICO	C0156257 C0268722	T047	B-DISO
derecho					
.					

Table 5: Example of annotated sentence from CodiEsp corpus. Translation: *A reno-vesical X-ray shows multiple radio-opaque stones in the right kidney*. ICD-10 codes are mapped to UMLS CUIs and their associated Semantic Types and Semantic Groups from the Semantic Network. PROC stands for Procedure, DISO stands for Disorder semantic group.

corpus. Mantra, E3C, CT-EBM-SP and MedMentions corpora are already annotated with semantic types and semantic groups manually. The distribution of the classes across the corpora is shown in Table 6. Note that the resulting bilingual corpora is highly unbalanced, and the number of classes changes across the languages and corpora.

- CodiEsp. After mapping ICD-10 codes to UMLS semantic groups we obtain a corpus annotated with three classes: Concept, Disorder and Procedure. The corpus is unbalanced in the same proportion as the original dataset: Disorder labels mainly correspond to Diagnosis labels from the original dataset and Procedure labels correspond to Procedures annotation.
- The combined Spanish Corpus is labelled with 11 semantic groups. 80% of the corpus is reserved for training, 10% for development and 10% for test. We test the model on the corpus test and on the original CodiEsp test. When testing this model on CodiEsp test set, we calculate the performance of three classes in CodiEsp (Concept, Disorder and Procedure) which are present.
- MedMentions is manually annotated with 15 UMLS semantic types, which are mapped to semantic groups with the Semantic Network.
- The bilingual Corpus is a combination of the Spanish and English corpora, without including the original test sets. In total, the corpus is annotated in 15 semantic groups from UMLS. We test the bilingual model on Spanish test set (CodiEsp) and on the English test set (MedMentions).

5. Results

The performance of the models is shown in Table 7. It is worth mentioning that two classes are used in the SL-DP models and 15 classes in the SL-SG models.

Label	Description	CodiEsp	Combined-es	MedMentions	Bilingual
CONC	Concepts & Ideas	37	40	151,526	151,119
DISO	Disorders	23,590	35,146	73,360	108,506
CHEM	Chemicals & Drugs	-	5,190	68,917	74,011
PROC	Procedures	8,639	22,009	67,290	89,274
LIVB	Living Beings	-	99	44,190	44,289
PHYS	Physiology	-	41	38,532	38,573
ANAT	Anatomy	-	3,651	30,720	34,371
ACTI	Activities & Behaviors	-	-	15,135	15,175
PHEN	Phenomena	-	18	10,851	10,869
GENE	Genes & Molecular Sequences	-	-	10,717	10,813
OBJC	Objects	-	11	9,986	10,429
DEVI	Devices	-	11	5,329	5,340
GEOG	Geographic Areas	-	6	4,295	4,301
ORGA	Organizations	-	-	2,256	2,256
OCCU	Occupations	-	-	2,125	2,125

Table 6: Tokens annotated with UMLS semantic groups for the SL-SG models

		SL-DP (2 classes)				SL-SG (15 classes)		
Corpus		P	R	F1		P	R	F1
CodiEsp (es)	gold	76.76	71.45	74.01	map	73.19	73.82	73.50
Combined-es (es)	gold+map	89.61	88.15	88.87	gold+map	88.91	88.17	88.54
Combined-es Test CodiEsp (es)	gold+map	74.42	68.53	71.35	gold+map	71.05	70.05	70.55
MedMentions (en)	map	92.69	86.53	89.50	gold	84.51	86.29	85.39
Bilingual (es+en)	gold+map	87.85	87.19	87.52	gold+map	86.19	87.00	86.59
Bilingual Test CodiEsp (es)	gold+map	71.57	70.19	70.87	gold+map	71.16	69.68	70.41
Bilingual Test MedMentions (en)	gold+map	89.10	85.34	87.18	gold+map	85.38	86.81	86.09

Table 7: Performance of the SL models on the test sets.

In addition, the Combined-es test set contains Spanish corpora of different genres (clinical reports, scientific papers of the biomedical domain), while CodiEsp contains clinical reports only. Moreover, depending on the selected training and evaluation setting, the model can predict labels not represented in the test sets. For instance, the bilingual SL-SG models are learning to annotate 15 classes while only three classes appear in CodiEsp (es) and 11 classes in Combined-es (es). Also note that these results have not been averaged on several runs.

On SL-DP (2 classes), it seems that the additional examples from the rest of Spanish corpora are not helping to improve the results on the CodiEsp test (74.01 vs. 71.35 F1 score). The same phenomena seem to occurs when including the additional examples from the English corpora (74.01 vs. 70.87 F1 score). Similarly, the additional Spanish examples are not helping to improve the results in MedMentions (89.50 vs. 87.18 F1 score). However, it seems that the results obtained when using the Combined-es corpora and MedMentions (en) using BERT are much higher and quite similar (88.87 vs. 89.50 f1 score).

A similar behavior seems to appear on SL-SG models when evaluating the different settings. Interestingly, we observe a small increase of performance when adding the Spanish corpus (Bilingual es+en with 86.59 F1 score) with respect to the English one (MedMentions with F1 85.39).

In summary, we can conclude that the new models and corpora are quite interoperable with respect the different coding systems and languages. None of these experiments across languages and corpora would have been possible without the new ClinIDMap resource.

6. Conclusions and future work

We have presented a mapping approach for clinical data interoperability, which allows to map codes from UMLS, SNOMED-CT and ICD-10 lexicons, in English and Spanish. We also experiment with Spanish, English and bilingual Sequence Labelling (SL) models trained on the corpora annotated with different coding systems. The SL models are trained for the detection and classification of Diagnosis and Procedures in ICD-10 notation, and for Semantic Groups in UMLS notation.

As a future work we plan to align CUI descriptions with Wikidata/Wikipedia items. Only less than 1% of the one million UMLS CUIs can be found in Wikidata. We plan to apply different deep learning methods for aligning UMLS concepts and Wikipedia pages (Rahimi et al., 2020). This will provide access to the Wikipedia knowledge base and increase the number of CUIs aligned with Wikipedia/Wikidata items.

We focus on Spanish, although our method is applicable for any language having the appropriate resources, and may help to translate and enrich the knowledge bases in specific languages. Having an annotated corpus in a language, it is possible to map these annotations with codes from the other knowledge bases.

In addition, it is possible to add more clinical taxonomies, such as NCBI[14], or MeSH[15] or BIOS[16]. Finally, we plan to develop new approaches for deriving more fine grained and interoperable annotations.

7. Acknowledgements

This work has been partially funded by the projects DeepText (KK-2020-00088, SPRI, Basque Government) and DeepReading (RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE).

8. Bibliographical References

Aronson, A. R. and Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 05.

Aronson, A. (2001). Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 2001:17–21, 02.

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

Campillos-Llanos, L. (2019). First Steps towards Building a Medical Lexicon for Spanish with Linguistic and Semantic Information. pages 152–164, August.

Chen, J., Jiménez-Ruiz, E., Horrocks, I., Antonyrajah, D., Hadian, A., and Lee, J. (2021). Augmenting ontology alignment by semantic embedding and distant supervision. In *European Semantic Web Conference*, pages 392–408. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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

Fung, K. W. and Xu, J. (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.

Kors, J. A., Clematide, S., Akhondi, S. A., van Mulligen, E. M., and Rebholz-Schuhmann, D. (2015). A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, 22(5):948–956, 05.

Magnini, B., Altuna, B., Lavelli, A., Speranza, M., and Zanoli, R. (2020). The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases.

Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., and Krallinger, M. (2020). Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020.

Mohan, S. and Li, D. (2019). MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. *ArXiv*, abs/1902.09476.

Ochieng, P. and Kyanda, S. (2018). Large-scale ontology matching: State-of-the-art analysis. *ACM Comput. Surv.*, 51(4), jul.

Rahimi, A., Baldwin, T., and Verspoor, K. (2020). WikiUMLS: Aligning UMLS to Wikipedia via Cross-lingual Neural Ranking. *arXiv preprint arXiv:2005.01281*.

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

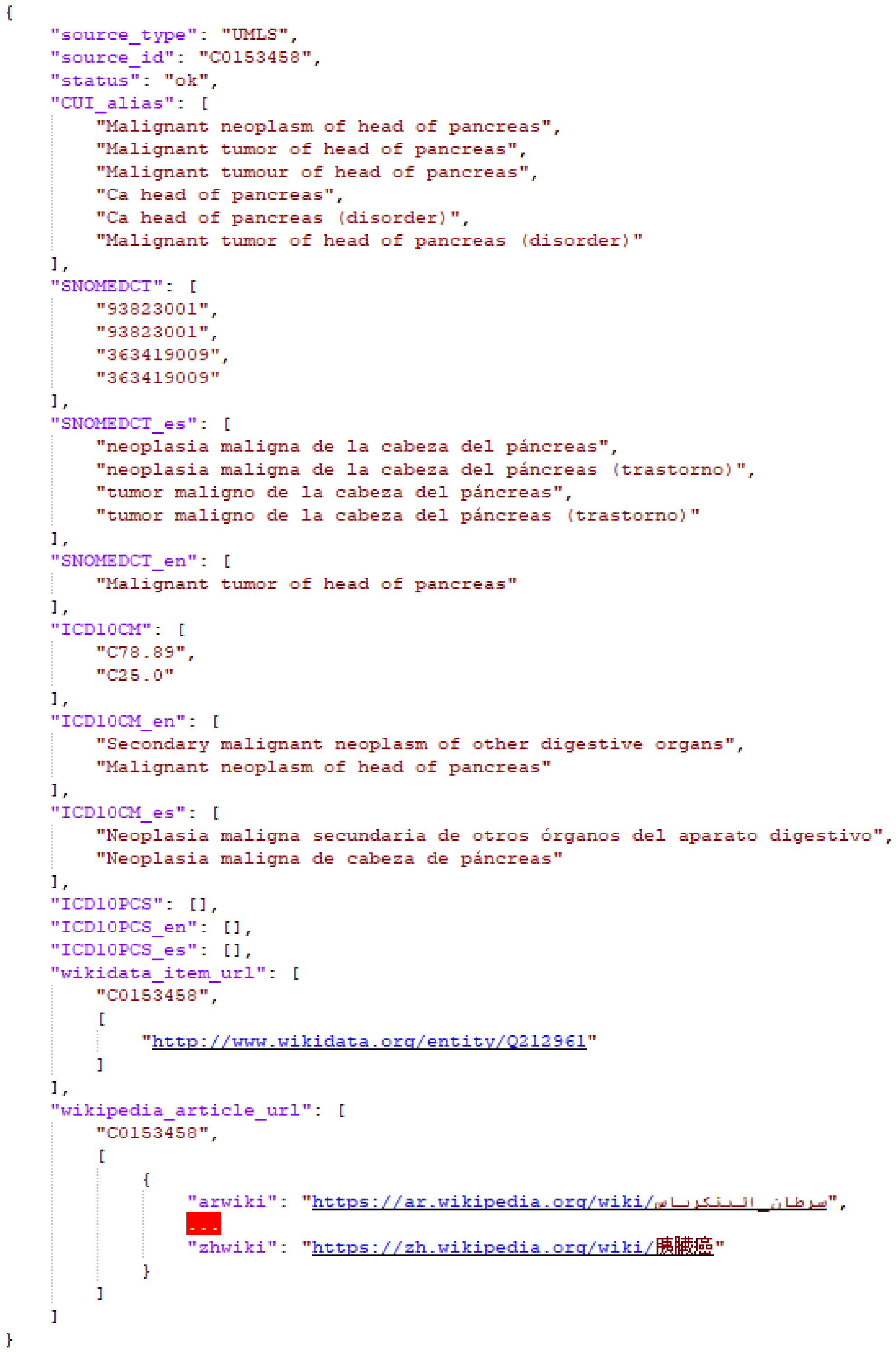
Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10):78–85, sep.

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

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

[15] https://www.nlm.nih.gov/mesh/meshhome.html

[16] https://bios.idea.edu.cn/



```
{
    "source_type": "UMLS",
    "source_id": "C0153458",
    "status": "ok",
    "CUI_alias": [
        "Malignant neoplasm of head of pancreas",
        "Malignant tumor of head of pancreas",
        "Malignant tumour of head of pancreas",
        "Ca head of pancreas",
        "Ca head of pancreas (disorder)",
        "Malignant tumor of head of pancreas (disorder)"
    ],
    "SNOMEDCT": [
        "93823001",
        "93823001",
        "363419009",
        "363419009"
    ],
    "SNOMEDCT_es": [
        "neoplasia maligna de la cabeza del páncreas",
        "neoplasia maligna de la cabeza del páncreas (trastorno)",
        "tumor maligno de la cabeza del páncreas",
        "tumor maligno de la cabeza del páncreas (trastorno)"
    ],
    "SNOMEDCT_en": [
        "Malignant tumor of head of pancreas"
    ],
    "ICD10CM": [
        "C78.89",
        "C25.0"
    ],
    "ICD10CM_en": [
        "Secondary malignant neoplasm of other digestive organs",
        "Malignant neoplasm of head of pancreas"
    ],
    "ICD10CM_es": [
        "Neoplasia maligna secundaria de otros órganos del aparato digestivo",
        "Neoplasia maligna de cabeza de páncreas"
    ],
    "ICD10PCS": [],
    "ICD10PCS_en": [],
    "ICD10PCS_es": [],
    "wikidata_item_url": [
        "C0153458",
        [
            "http://www.wikidata.org/entity/Q212961"
        ]
    ],
    "wikipedia_article_url": [
        "C0153458",
        [
            {
                "arwiki": "https://ar.wikipedia.org/wiki/سرطان_البنكرياس",
                ...
                "zhwiki": "https://zh.wikipedia.org/wiki/胰腺癌"
            }
        ]
    ]
}
```

Figure 1: Example of the mapping for C0153458 UMLS CUI.