Towards the integration of WordNet into ClinIDMap

Elena Zotova
SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA) / Mikeletegi Pasealekua 57, 20009, Donostia/San-Sebastián, Spain
and
Department of Languages and Computer Systems, University of the Basque Country (UPV-EHU) / Paseo Manuel de Lardizábal, 1, 20018, Donostia/San-Sebastián, Spain
ezotova@vicomtech.org

Montse Cuadros
SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA) / Mikeletegi Pasealekua 57, 20009, Donostia/San-Sebastián, Spain
mcuadros@vicomtech.org

German Rigau
Department of Languages and Computer Systems, University of the Basque Country (UPV-EHU)
and
HiTZ Basque Center for Language Technologies
Paseo Manuel de Lardizábal, 1, 20018, Donostia/San-Sebastián, Spain
german.rigau@ehu.eus

Abstract

This paper presents the integration of WordNet knowledge resource into ClinIDMap tool, which aims to map identifiers between clinical ontologies and lexical resources. ClinIDMap interlinks identifiers from UMLS, SMOMED-CT, ICD-10 and the corresponding Wikidata and Wikipedia articles for concepts from the UMLS Metathesaurus. The main goal of the tool 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. In this new release, we add WordNet 3.0 and 3.1 synsets using the available mappings through Wikidata. Thanks to cross-lingual links in MCR we also include the corresponding synsets in other languages and also, extend further ClinIDMap with different domain information. Finally, the final resource helps in the task of enriching of already annotated clinical corpora with additional semantic annotations.

1 Introduction

The main goal of the ClinIDMap mapping tool (Zotova et al., 2022) 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), SNOMED-CT (Donnelly et al., 2006) and others. The alignment uses the actual IDs of the KBs from the official mapping resources developed by the authors of SNOMED-CT and UMLS. The alignment allows 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 with ICD-10 annotations. 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, ClinIDMap enriches the annotated concepts with multilingual terms and descriptions of its available Wikidata and Wikipedia articles, allowing to expand brief code descriptions to detailed information in multiple languages.

Now, we introduce the functionality of mapping those clinical concepts to WordNet (Miller, 1998). WordNet (WN) is a widely used lexical knowledge resource, which contains information about lexical relations, such as synonyny and super-subordinate relation (hyperonymy, hyponymy). In addition, WordNet is used as a backbone of many other lexical resources. The alignment allows us to enrich manually annotated corpora with extra clinical codes and to obtain multilingual inter-operable cor-

pora 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. And vise versa, a corpus annotated with ICD-10 codes can be used to derive automatically new corpora annotated with UMLS codes, semantic types or groups.

Thus, this paper focuses on two tasks: (1) extending ClinIDMap to include WordNet information, and (2) annotating automatically clinical corpora with new labels related to information associated to WordNet. Concretely, we present the integration of WordNet mapping with clinical identifiers such as UMLS, SNOMED-CT, ICD-10, MeSH a for Spanish, English and other languages. Using this tool, we derive multiple datasets annotated with different coding systems on the base of existing annotated corpora. The previous version of the tool is described in detail in Zotova et al. (2022) and the tool is publicly available[1].

For instance, a Spanish sentence from E3C corpus (Magnini et al., 2020) annotated with a UMLS code is given below.

> Durante los 5 años que permaneció en DP sufrió 10 **peritonitis** [C0031154], 8 por Staphilococcus aureus.
>
> Translation: *During the 5 years on PD he suffered 10* ***peritonitis*** *[C0031154], 8 of which were because of Staphylococcus aureus.*

The code *C0031154* corresponding to the Spanish term *peritonitis* can be mapped to the SNOMED CT code *235983003*, to the ICD-10-CM code *K65*, to the corresponding Wikipedia articles in 48 languages, to synset *14376092-n* in WordNet 3.1. and synset *14352687-n* in WordNet 3.0.

The paper is organized as follows. Section 2 describes previous attempts of mapping clinical codes and also using WordNet in the clinical domain; Section 3 is dedicated to the databases used to develop ClinIDMap—clinical ontologies, mapping schema and general purpose lexical resources; in Section 4 we describe (1) the method of aligning WordNet synsets to clinical codes, WordNet Domains and WordNets in different languages (Subsection 4.1) and (2) semi-automatic method of annotating of the clinical corpora (Subsection 4.2).Finally, Section 5 concludes the paper and presents the future work.

2 Related Work

2.1 Aligning Clinical Codes

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

Ontology matching finds semantically related entities in different knowledge bases (KB). For instance, the OAEI Campaign (Ontology Alignment Evaluation Initiative) [2] organizes 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). ClinIDMap uses already aligned clinical KBs.

Most 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 neu-

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

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

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

ral re-ranking model which is based on a pre-trained 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, ClinIDMap exploits available manual mappings between the different lexical resources.

2.2 WordNets for the Clinical Domain

There were various attempts to create domain specific WordNets such as the Medical WordNet (Smith and Fellbaum, 2004) with the goal of linking different terms, both professional terminology and general language. These resources should also be ready for NLP automatic applications such as relation extraction, entity linking, and automatic clinical coding.

WordNet was proposed as a method for giving patients interpretative support when annotating foreign word-meanings with the corresponding Norwegian synset (Ingvaldsen and Veres, 2004). This was supposed to be an add-on for the electronic medical record systems that will help regular patients in getting insight to their diagnoses. The add-on service is based annotating polysemous and foreign terms with WordNet synsets and then use the relationships established in WordNet to return definitions and hypernymy, meronymy and entailment meanings of a term.

WordNet was used to improve the direct mapping of data elements during the integration of biomedical resources in the study of Mougin et al. (2006). WordNet contributes external information useful for disambiguation and validation of UMLS direct mappings. WordNet can also help identify indirect mappings of DEs to the UMLS. Also, WordNet synsets help identify indirect mappings to the UMLS when no direct UMLS mapping was found.

There were also studies of how to align WordNet domains and Wikipedia categories to obtain domain specific corpora (Gella et al., 2014). The authors expected that the multilingual, and comparable, domain-specific corpora have the potential to enhance research in word-sense disambiguation and terminology extraction in different languages, which could enhance the performance of various NLP tasks.

3 Background

This section describes the resources and databases used to build ClinIDMap. It includes a brief information about the clinical and general knowledge bases used and the resources exploited for mapping the different codes.

3.1 Clinical Knowledge Bases

The following medical knowledge bases are used to build ClinIDMap. Each of them consists of a set of identifiers (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.

MeSH[5] stands for Medical Subject Headings (MeSH) thesaurus which is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine. It is used for indexing, cataloging, and searching of biomedical and health-related information. MeSH includes the subject headings appearing in MEDLINE/PubMed, the National Library of Medicine[6] (NLM) Catalog, and other NLM databases.

Spanish SNOMED-CT[7] 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 called CIE-10-ES.

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

[5] https://meshb.nlm.nih.gov/

[6] https://www.nlm.nih.gov/

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

ClinIDMap uses the official Spanish version of the CIE-10 from July 2020[8].

ICD-10-PCS (Procedure Coding System)[9] 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. ClinIDMap uses the official Spanish version of the ICD-10-PCS from January 2020.

3.2 Clinical Codes Mapping Resources

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

UMLS Metathesaurus[10]. This database 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).

The Semantic Network from UMLS is used for grouping CUIs. 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. About a third part 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 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 Lexical Resources

ClinIDMap has been enriched with general purpose lexical resources in order to include terminology descriptions in different languages. The following lexical resources are included.

Wikidata[12] (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 Wikidata can be used to extract the corresponding articles in all available languages.

Wikipedia[13] is used as a multilingual online encyclopedia of clinical concepts. Wikipedia provides extensive description of clinical concepts in many languages.

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

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

[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

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

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

WordNet 3.1[14] (Fellbaum, 2005) is the latest version 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 by means of conceptual-semantic and lexical relations. This version contains 155,327 words organized in 175,979 synsets for a total of 207,016 word-sense pairs.

WordNet 3.0 [15] (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 sense. 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.

WordNet Domains[16] (Magnini and Cavaglià, 2000) is a lexical resource created in a semi-automatic way by augmenting WordNet with domain labels. WordNet synsets have been annotated with at least one semantic domain label, selected from a set of 170 labels structured according the WordNet Domain Hierarchy. There are various domains related to health and medicine. It is unclear what type of relations among the relevant domains is established. For instance, arguably, surgery and pharmacy may be included in the broader domain of medicine or health. We manually select a set of domains:

```
medicine, anatomy, pharmacy,
health, biochemistry, surgery,
physiology, genetics,
psychological_features, psychology,
radiology, genetics, dentistry,
psychiatry, optics, chemistry
```

We use these domains for semi-automatic data annotation.

WordNet extended Domains (Gonzalez-Agirre et al., 2012b) is a resource aiming to improve WordNet Domains. The original domain labels have been projected to WordNet 3.0 using automatic mappings across WordNet versions (Daude et al., 2003). Since the automatic mapping is not complete due to new synsets, changes in the structure, etc., many synsets were left unlabeled. The extended WordNet domains were elaborated by an expansion process through the graph of WordNet. This resource consists of 170 files, one for each of the original WordNet Domains. Each file contains a vector of 117,536 synsets sorted by weight, from highest lo lowest. Thus, the most representative synsets for a given domain are at the top positions. For instance, the first four lines of the file *health.ppv* correspond to the first synset-weight pairs (we have added the variants):

```
00624738-n 0.00771299 exercise_1
14049711-n 0.00561771 good_health_1
01017738-a 0.00504791 unfit_2
05216365-n 0.00492294 body_1
```

Multilingual Central Repository (MCR) 3.0 (Gonzalez-Agirre et al., 2012a) integrates using the EuroWordNet framework, WordNets from six different languages: English, Spanish, Catalan, Basque, Galician and Portuguese. The Inter-Lingual-Index (ILI) allows the connection from words in one language to equivalent translations in any of the other languages thanks to the automatically generated mappings among WordNet versions. The current ILI version corresponds to WordNet 3.0.

Coarse Sense Inventory (CSI) (Lacerra et al., 2020) is a coarse-grained sense inventory where semantic labels are shared across the lexicon of WordNet. There are 46 labels in total and we select the class *HEALTH_AND_MEDICINE_* to filter clinical identifiers.

4 Methodology

4.1 WordNets Mapping

Items in Wikidata are annotated manually by Wikidata experts. However, there may be variations and mismatches with respect the UMLS or SNOMED CT to ICD-10 mappings described in Subsection 3.2.

Step 1. Collect all Wikidata items. First of all, we need to gather all the Wikidata items including WordNet 3.1 synsets, optionally adding their corresponding clinical IDs, such as UMLS CUI, SNOMED CT, MeSH and ICD-10.

Step 2. WordNets 3.1 and WordNet 3.0 mapping. Resources such as WordNet Domains, CSI and MCR are aligned with WordNet 3.0, while Wikidata items use WordNet 3.1. To obtain the corresponding domains and CSI codes we need to map WordNet 3.1 offsets to those of WordNet 3.0. We use the sense key index for this mapping. According to Kafe (2018), 99,4% of sense keys from WordNet 3.0 persist in WordNet 3.1–716 KSI were

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

[15] https://wordnetcode.princeton.edu/3.0/WordNet-3.0.tar.gz

[16] https://wndomains.fbk.eu/

added and 1,304 KSI were removed. Each version of WordNet distribution contains a file *index.sense* which includes all senses with their corresponding offsets. These sense keys are coded as follows. For instance, the sense key "adenoma%1:26:00::" contains a lemma of the synset "adenoma". The first number refers to the part of speech (1 is noun, 2 is verb, 3 is adjective, 4 is adverb, 5 is adjective satellite). The second two-digit code is representing the name of the lexicographer file (e.g. part of speech and its attribute, such as time, person, body—44 names in total). The third two-digit code refers to ID in lexicographical file. We use the whole sense key to match senses across the different WordNet versions.

Step 3. WordNet 3.0 to WNDomains, CSI and MCR mapping. Once having the WordNet 3.0 synsets, we can easily access the rest of the KBs –Domains, CSI and MCR. The resulting table is used as establish the mapping between the clinical codes and the WordNet synsets.

For instance, below is an example of 14235793-n synset (*adenoma*) and its five most probable WNDomains.

```
14235793-n  0.00010198  medicine
14235793-n  0.00005412  veterinary
14235793-n  0.00003494  anatomy
14235793-n  0.00001745  radiology
14235793-n  0.00001649  cycling
```

There are about 27,500 Wikidata items annotated with WordNet 3.1 synsets. As we see in Table 1, only a small part of Wikidata items (approximately 1 to 10%) annotated with WordNet synsets is also annotated with clinical codes. Some of the items are annotated with multiple synsets, the distribution of the multiple synsets across the Wikidata items is shown in Table 2. Table 5 shows some examples of some Wikidata items connected to various clinical identifiers. This database can be used to connect clinical codes to WordNet synsets. +

Database	Unique items
Wikidata items	27,516
WordNet 3.1	26,953
WordNet 3.0 (mapped)	26,938
UMLS CUI	2,076
ICD-10	833
SNOMED CT	282

Table 1: Numbers of Wikidata items annotated with both WordNet synsets and clinical IDs.

#Wikidata items	#synsets
5	6
10	5
38	4
265	3
1,663	2
25,535	1

Table 2: Number of Wikidata items annotated with various WordNet synsets.

We also map all the Wikidata items to extended WordNet domains and to the CSI domains. For each synset, we select the 5 most probable domains from the extended WordNet domains that contain a clinical domain. Table 3 shows the number of Wikidata items with clinical codes from extended WordNet domains and from CSI, and its overlapping.

Database	Wikidata items
CSI	3,133
WordNet clinical domains	3,396
Total clinical domain only	2,398

Table 3: Number of Wikidata items annotated with clinical domains (from CSI and Extended WordNet Domains).

WordNet 3.0 offsets are also used for gathering the non-English synsets included into the MCR.

4.2 Corpora annotation with WordNet synsets

After building the new version of ClinIDMap, now integrating WordNet synsets, we study how many clinical IDs from the domain corpora (see the description of the used corpora in (Zotova et al., 2022)) can be mapped to the WordNet synsets and its corresponding domains. Four corpora of various types were selected for the experiments: CodiEsp 2020 (clinical narratives in Spanish, annotated with ICD-10 codes), E3C (clinical narratives in Spanish), CT-EBM-SP (clinical trials in Spanish annotated with CUI), MedMentions (biomedical papers in English annotated with CUI). Then, we annotate the corpora with two types of labels: (1) WordNet domains; (2) CSI labels.

As shown in the Table 4, about 5-20% percent of the clinical annotations are mapped to WordNet synsets, possibly not only from the clinical domain. The variety of the unique synsets in the corpus depends, first, on its size, and on the na-

Corpus	Tokens	Annotated CUI	Mapped WN	Unique WN
E3C ES (Magnini et al., 2020)	28,815	2,268	422	107
MedMentions (Mohan and Li, 2019)	1,258,847	540,138	24,754	841
Mantra (Kors et al., 2015)	3,492	1,058	117	62
CT-EBM-SP (Campillos-Llanos, 2019)	141,158	23.264	5,786	431
CodiEsp 2020 (Miranda-Escalada et al., 2020)	401,010	32,902	11,464	399

Table 4: Number of tokens annotated with both WN synsets and clinical IDs using mapping of UMLS CUI to WN synsets.

item	label	MESH	CUI	ICD-10	SNOMED-CT	WN 3.1	WN 3.0	sense	domain	CSI
Q272741	adenoma	D000236	C0001430	D35.0	32048006	14259275-n	14235793-n	adenoma%1:26:00::	medicine	HEALTH_AND_MEDICINE_
Q272741	adenoma	D000236	C0334389	D35.2	32048006	14259275-n	14235793-n	adenoma%1:26:00::	medicine	HEALTH_AND_MEDICINE_
Q7365	muscle organ	D009132	C0026845			05296796-n	05289297-n	musculus%1:08:00::	health	BIOLOGY_
Q84133	myocardium	D009206	C0027061			05398343-n	05391000-n	myocardium%1:08:00::	anatomy	HEALTH_AND_MEDICINE_
Q223102	peritonitis	D010538	C0029823	K65		14376092-n	14352687-n	peritonitis%1:26:00::	medicine	HEALTH_AND_MEDICINE_

Table 5: Examples of WordNets mapped with clinical IDs, WordNet domains and CSI.

ture of the data. Here, the corpus MedMEntions compiled from English biomedical papers has the largest number of mappings to WordNet synsets, but the Spanish part of E3C has in proportion the largest number of distinct mappings.

Using the new version of ClinIDMap, now including WordNet synsets we can also project all these annotations to other resources associated to WordNet such as WordNet Domains and CSI domains. Table 6 presents the distribution of medical WordNet Domain labels as there are also entities annotated with CUIs not belonging to the medical domain. Now, with the new version of ClinIDMap we can select those annotations belonging to the clinical domain. As we can see in the number of domains differs from corpus to corpus and is also related to the data type—clinical narratives contain less labels than scientific papers or trials.

We also derive a new corpora annotated with CSI labels. Table 7 shows the distribution of CSI labels across the different corpora. If various CSI domains are assigned to a token, the most frequent one is selected. Again, the distributions of the labels across the tokens is not balanced. The larger corpus (MedMentions) is annotated with 23 labels while the E3C is annotated with only four. As expected, the prevalence of health-related labels is high. Nevertheless, the texts also contain labels not related to the medical domain.

WND	MM	CT	E3C	CE
NULL	1,239,202	136,287	28,480	393,580
medicine	4,349	1,450	285	5,185
anatomy	3,001	928		72
biochemistry	3,821	807		21
pharmacy	1,922	842	29	569
radiology	529	408	3	308
psychiatry	1,678	257	37	361
optics	380	161	8	130
physiology	230	134	9	322
surgery	254	81	8	43
health	394	62	18	175
genetics	1,018	49	3	97
chemistry	974	24		30
dentistry	159	15	2	80
psychology		14	6	34

Table 6: Number of tokens annotated with WordNet domains (WN-D) using the mapping method from MedMentions (MM), CT-EBM-SP (CT), E3C, CodiEsp 2020 (CE).

5 Conclusions

In this paper we present an extension of ClinIDMap now integrating WordNet synsets in different languages and its domain information. We also use the new medical resource to provide different prespectives to the annotated data. As a future work we plan to experiment with the annotated corpora and train deep learning models for sequence labeling of WN domains and CSI labels. We also plan to use other WordNet relations and associated knowledge. We would also like to add new clinical and lexical resources to ClinIDMap such as additional knowledge from different Wikipedia.

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

CSI Label	MM	CT	E3C	CE
NULL	1,234,315	135,402	28,396	393,252
HEALTH_AND_MEDICINE_	15,796	4,379	370	7,758
BIOLOGY_	5,907	1,231	30	
CHEMISTRY_AND_MINERALOGY_	1,743	119		
MATHEMATICS_	470			
FOOD_DRINK_AND_TASTE_	112	2		57
EVALUATION_	100	17	17	
TIME_	92			
BUSINESS_ECONOMICS_AND_FINANCE_	47			
POLITICS_GOVERNMENT_AND_NOBILITY_	46			
SEX_	39	1		48
METEOROLOGY_	37			
PHILOSOPHY_PSYCHOLOGY_AND_BEHAVIOR_	36	2		305
EDUCATION_AND_SCIENCE_	26			
CULTURE_ANTHROPOLOGY_AND_SOCIETY_	20			
PHYSICS_AND_ASTRONOMY_	14	3		11
GEOGRAPHY_AND_PLACES_	11			
FARMING_	9			
COMPUTING_	7			
CRAFT_ENGINEERING_AND_TECHNOLOGY_	6			
VISUAL_	5		2	59
LANGUAGE_AND_LINGUISTICS_	5			
ART_ARCHITECTURE_AND_ARCHAEOLOGY_	2			
EMOTIONS_	1			15
WARFARE_DEFENSE_AND_VIOLENCE	1	2		

Table 7: Number of tokens annotated with CSI labels using the mapping method using the mapping method from MedMentions (MM), CT-EBM-SP (CT), E3C, CodiEsp 2020 (CE).

Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Leonardo Campillos-Llanos. 2019. First steps towards building a medical lexicon for Spanish with linguistic and semantic information. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 152–164, Florence, Italy. Association for Computational Linguistics.

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.

Jordi Daude, Lluis Padro, and German Rigau. 2003. Validation and tuning of wordnet mapping techniques. In *Proceedings of RANLP*, pages 117–123.

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.

Spandana Gella, Carlo Strapparava, and Vivi Nastase. 2014. Mapping WordNet domains, WordNet topics and Wikipedia categories to generate multilingual domain specific resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1117–1121, Reykjavik, Iceland. European Language Resources Association (ELRA).

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. *Proceedings of the 6th Global WordNet Conference (GWC'12) ISBN 978-80-263-0244-5.*

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012a. Multilingual central repository version 3.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2525–2529. European Language Resources Association (ELRA).

Aitor Gonzalez-Agirre, German Rigau, and Mauro Castillo. 2012b. A graph-based method to improve wordnet domains. In *Computational Linguistics and Intelligent Text Processing - 13th International Conference, CICLing 2012, New Delhi, India, March*

11-17, 2012, Proceedings, Part I, volume 7181 of *Lecture Notes in Computer Science*, pages 17–28. Springer.

Jon Espen Ingvaldsen and Csaba Veres. 2004. Using the wordnet ontology for interpreting medical records. In *CAiSE Workshops*.

Eric Kafe. 2018. Persistent semantic identity in wordnet. *Cognitive Studies*, 2018.

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

Caterina Lacerra, Michele Bevilacqua, Tommaso Pasini, and Roberto Navigli. 2020. CSI: A coarse sense inventory for 85% word sense disambiguation. In *Proceedings of the 34th Conference on Artificial Intelligence*, pages 8123–8130. AAAI Press.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases.

Bernardo Magnini and Gabriela Cavaglià. 2000. Integrating subject field codes into WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

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.

Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. *ArXiv*, abs/1902.09476.

Fleur Mougin, Anita Burgun, and Olivier Bodenreider. 2006. Using wordnet to improve the mapping of data elements to umls for data sources integration. In *AMIA Annual Symposium Proceedings*, volume 2006, page 574. American Medical Informatics Association.

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

Afshin Rahimi, Timothy Baldwin, and Karin Verspoor. 2020. WikiUMLS: Aligning UMLS to Wikipedia via Cross-lingual Neural Ranking. *arXiv preprint arXiv:2005.01281*.

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 371es, USA. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10):7885.

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.

A Appendix. Output mappings


```
{
    "source_type": "UMLS",
    "source_id": "C0001430",
    "status": "OK",
    "UMLS_CUI": [
        {
            "id": "C0001430",
            "description": "Adenoma"
        },
        {
            "id": "C0001430",
            "description": "Adenoma, NOS"
        },
        {
            "id": "C0001430",
            "description": "[M]Adenoma NOS"
        },
        {
            "id": "C0001430",
            "description": "[M]Adenomas"
        },
        {
            "id": "C0001430",
            "description": "Benign adenoma"
        },
        {
            "id": "C0001430",
            "description": "[M]Adenoma NOS (morphologic abnormality)"
        },
        {
            "id": "C0001430",
            "description": "Adenoma, no subtype (morphologic abnormality)"
        },
        {
            "id": "C0001430",
            "description": "Adenoma, no subtype"
        },
        {
            "id": "C0001430",
            "description": "Benign adenomatous neoplasm (disorder)"
        },
        {
            "id": "C0001430",
            "description": "Benign adenomatous neoplasm"
        }
    ],
    "SNOMED_CT_EN": [
        {
            "id": "443416007",
            "description": "Benign adenomatous neoplasm (disorder) Benign
                adenomatous neoplasm Adenoma Benign adenoma"
        },
        {
            "id": "32048006",
            "description": "Adenoma Adenoma, NOS Adenoma, no subtype (morphologic
                abnormality) Adenoma, no subtype"
        },
        {
            "id": "189579004",
            "description": "[M]Adenoma NOS [M]Adenoma NOS (morphologic abnormality)"
        },
        {
            "id": "189578007",
            "description": "[M]Adenomas &/or adenocarcinomas [M]Adenomas and
                adenocarcinomas [M]Adenomas [M]Adenocarcinomas [M]Adenomas &/or
                adenocarcinomas (disorder)"
        }
```

```
    ],
    "SNOMED_CT_ES": [
        {
            "id": "32048006",
            "description": "adenoma"
        },
        {
            "id": "32048006",
            "description": "morfología: adenoma, no tipificado (anomalía morfológica
                )"
        }
    ],
    "ICD10CM_ES": [
        {
            "id": "D36.9",
            "description": "Neoplasia benigna, localización no especificada"
        }
    ],
    "ICD10PCS_ES": [],
    "MESH": [
        {
            "id": "D000236",
            "description": "Adenoma, Basal Cell"
        },
        {
            "id": "D000236",
            "description": "Adenoma, Follicular"
        },
        {
            "id": "D000236",
            "description": "Adenoma, Microcystic"
        }
    ],
    "wikidata_item_url": [
        "http://www.wikidata.org/entity/Q272741"
    ],

    "wikipedia_article_url": [
        {
            "arwiki": "https://ar.wikipedia.org/wiki/_>
            ...
            "zhwiki": "https://zh.wikipedia.org/wiki/"
        }
    ],
    "WordNet": [
        {
            "WordNet 3.1": "14259275-n",
            "WordNet 3.0": "14235793-n",
            "CSI": "HEALTH_AND_MEDICINE_",
            "WordNet Domain": "medicine",
            "sense": "adenoma%1:26:00::",
            "MCR synset": [
                {
                    "en": "a benign epithelial tumor of glandular origin",
                    "es": "tumor epitelial benigno de origen glandular",
                    "pt": "um tumor epitelial benigno de origem glandular",
                    "gl": "",
                    "eu": "",
                    "ca": ""
                }
            ]
        }
    ]
}
```