vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Towards the integration of WordNet into ClinIDMap

Elena Zotova, Montse Cuadros and German Rigau
Vicomtech & University of Basque Country

Nombre

January 23-27, 2023
Donostia, Spain

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Elena ZOTOVA

NLP Researcher, PhD Student
Tecnological Centre Vicomtech,
University of the Basque Country
ezotova@vicomtech.org

Nombr

Dr Montse CUADROS

Senior NLP Researcher
Tecnological Centre Vicomtech

mcuadros@vicomtech.org

Dr German RIGAU

Deputy Director, Associate Professor
HiTZ Basque Center for Language Technologies,
University of the Basque Country
german.rigau@ehu.eus

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

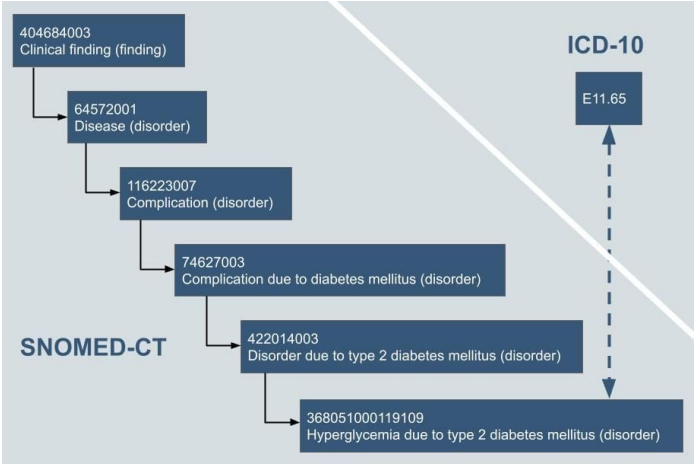
Index

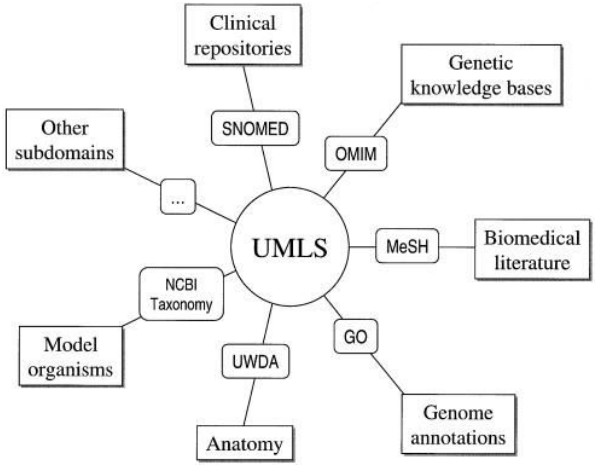
vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Index

Introduction

- Terminology processing is an important part of **clinical NLP**
- There are dictionaries, taxonomies and ontologies of different structure
- Similar concepts, synonyms, hierarchy
- Unique identifier - short description of a term / concept





Introduction

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

- **ClinIDMap** - a tool for mapping identifiers between clinical ontologies and lexical resources (Zotova E et al, 2022)
- Interlinks identifiers from UMLS (CUI), SMOMED-CT, ICD-10 and Wikipedia/Wikidata items and WordNet
- The goal is to provide semantic interoperability across the clinical concepts from various knowledge bases
- Focus on Spanish language but the method is valid for any language
- Explore WordNet for clinical domain

Example from E3C Corpus

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.

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

Example from E3C Corpus

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 (UMLS CUI)** corresponding to the Spanish term *peritonitis* can be mapped to:

- SNOMED CT code **235983003**;
- ICD-10-CM code **K65**;
- Wikipedia articles in **48 languages**;
- **synset 14376092-n** (WordNet 3.1);
- **synset 14352687-n** (WordNet 3.0).

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Index

Motivation

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

The task of clinical coding:

- Detect a span with clinical terminology (Sequence labelling model)
- Assign a high-level class: Diagnosis, Procedure, Anatomy etc.
- Link classified span with clinical taxonomy and assign an ID

Motivation

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Data for machine learning models in the clinical domain is especially difficult:

- Clinical information is private
- Manual annotation requires high-level expertise in medicine
- Few data is available for languages other than English
- If we have a corpus annotated with UMLS codes we transfer it to the corpus with ICD-10/SNOMED-CT/other codes

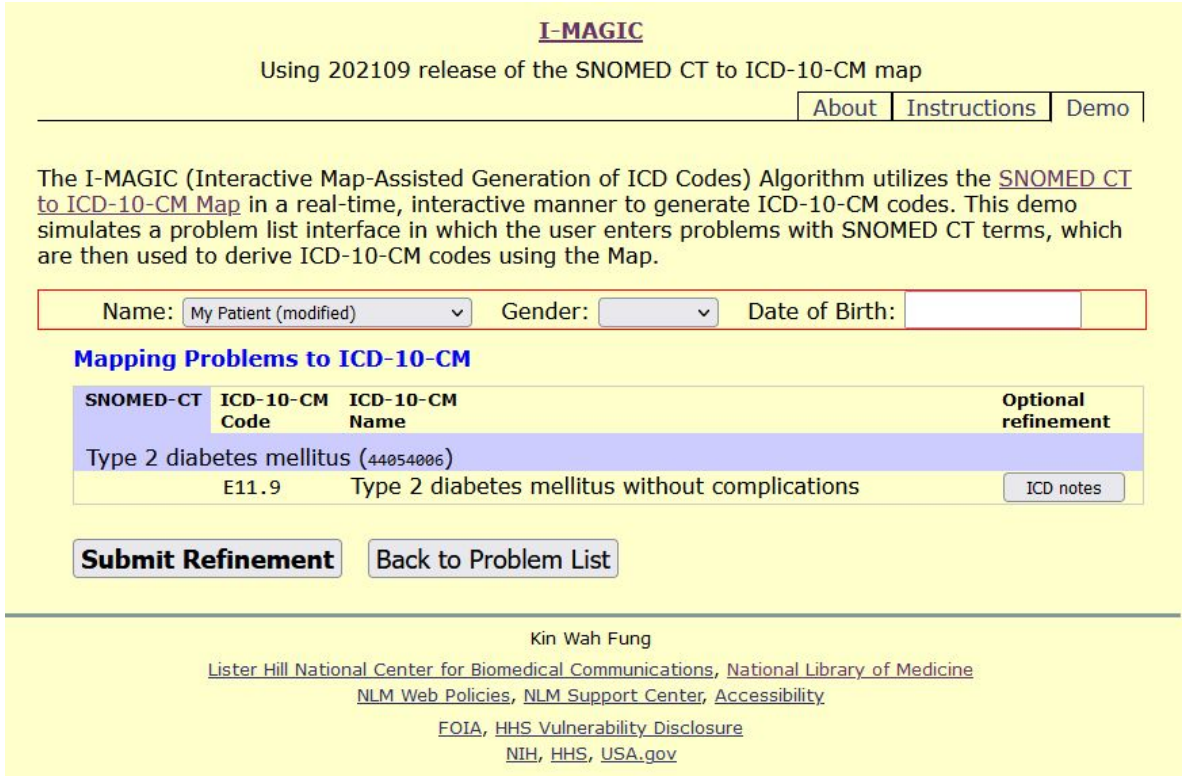
vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Index

Background

Concept alignment, or ontology alignment (also known as ontology matching): lexical matching, structural matching and logical reasoning.

Applications which use the resulting concept mapping to process biomedical text: MetaMap, I-MAGIC



Background

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

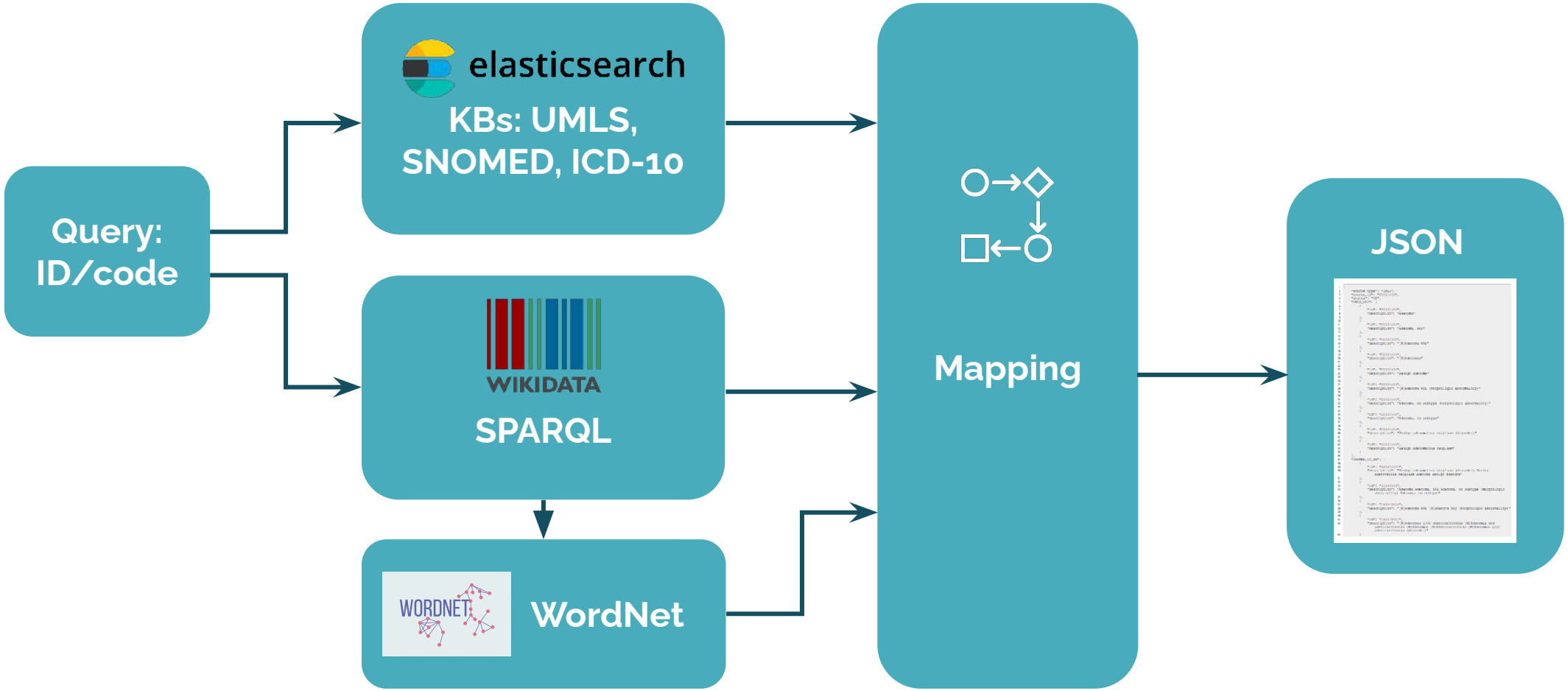
Attempts to use WordNets for clinical domain

- Medical WordNet had the goal of linking different terms, both professional terminology and general language. (Smith and Fellbaum, 2004)
- Explaining foreign clinical terms in Norwegian (Ingvaldsen and Veres, 2004)
- Disambiguation of UMLS mappings (Mougin et al., 2006).
- Aligning WordNet domains and Wikipedia categories to obtain domain corpora (Gella et al., 2014)

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Index

High-level scheme



Mapping: Knowledge Bases

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

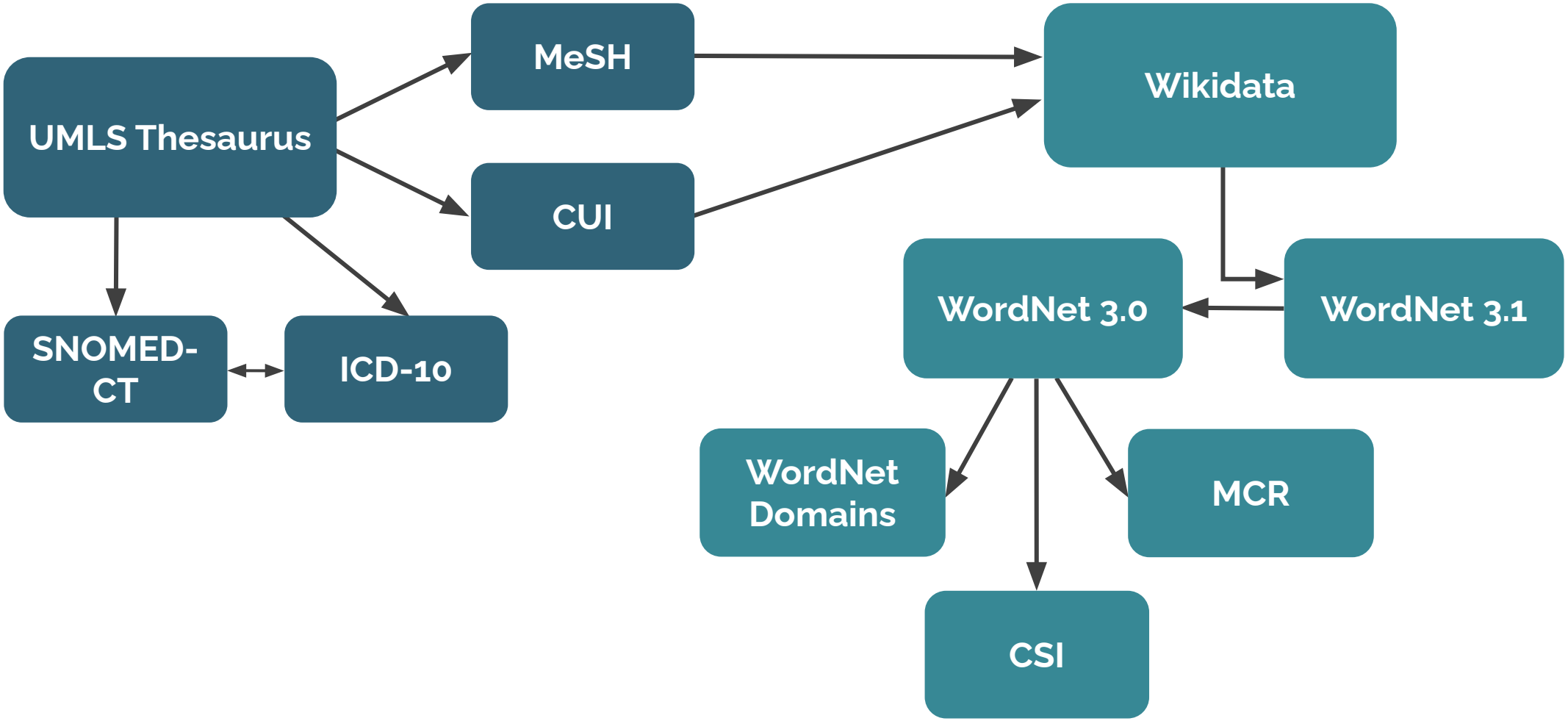
- UMLS Metathesaurus: mappings with different clinical ontologies
- SNOMED-CT (es, en)
- ICD-10-CM (es, en)
- ICD-10-PCS (es, en)
- SNOMED-CT to ICD-10 mapping (en)
- MeSH
- Wikidata
- Wikipedia
- WordNet 3.1 and 3.0

Mapping: Knowledge Bases

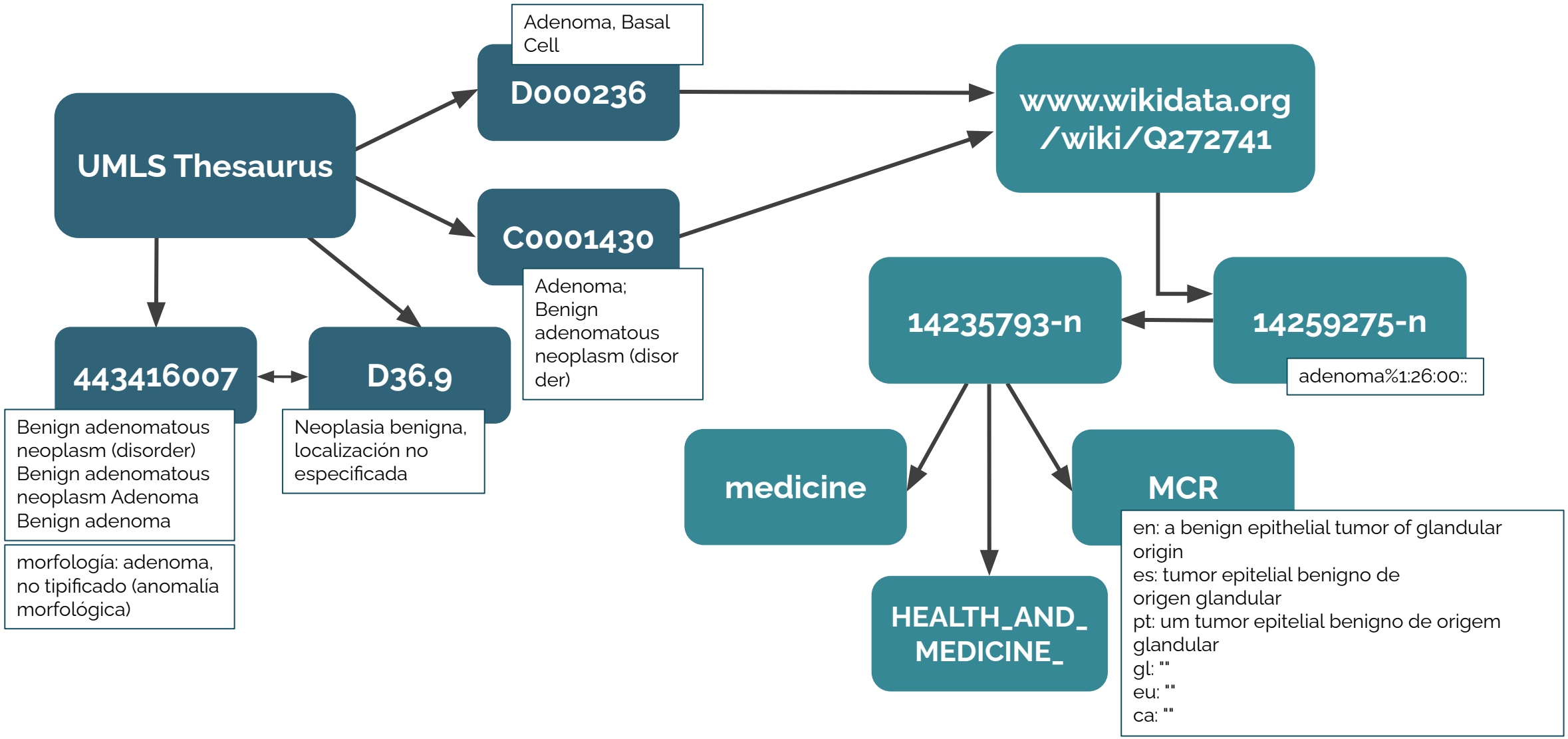
vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Knowledge base	Unique IDs
SNOMED-CT	489,141
ICD-10-CM	95,671
ICD-10-PCS	90,673
MeSH	347,565
UMLS CUI	1,106,486
SNOMED-CT 2 ICD-10	125,823
SNOMED-CT-ES 2 ICD-10	57,393
Wikidata	28,113
Wikidata+WordNet 3.1	26,953

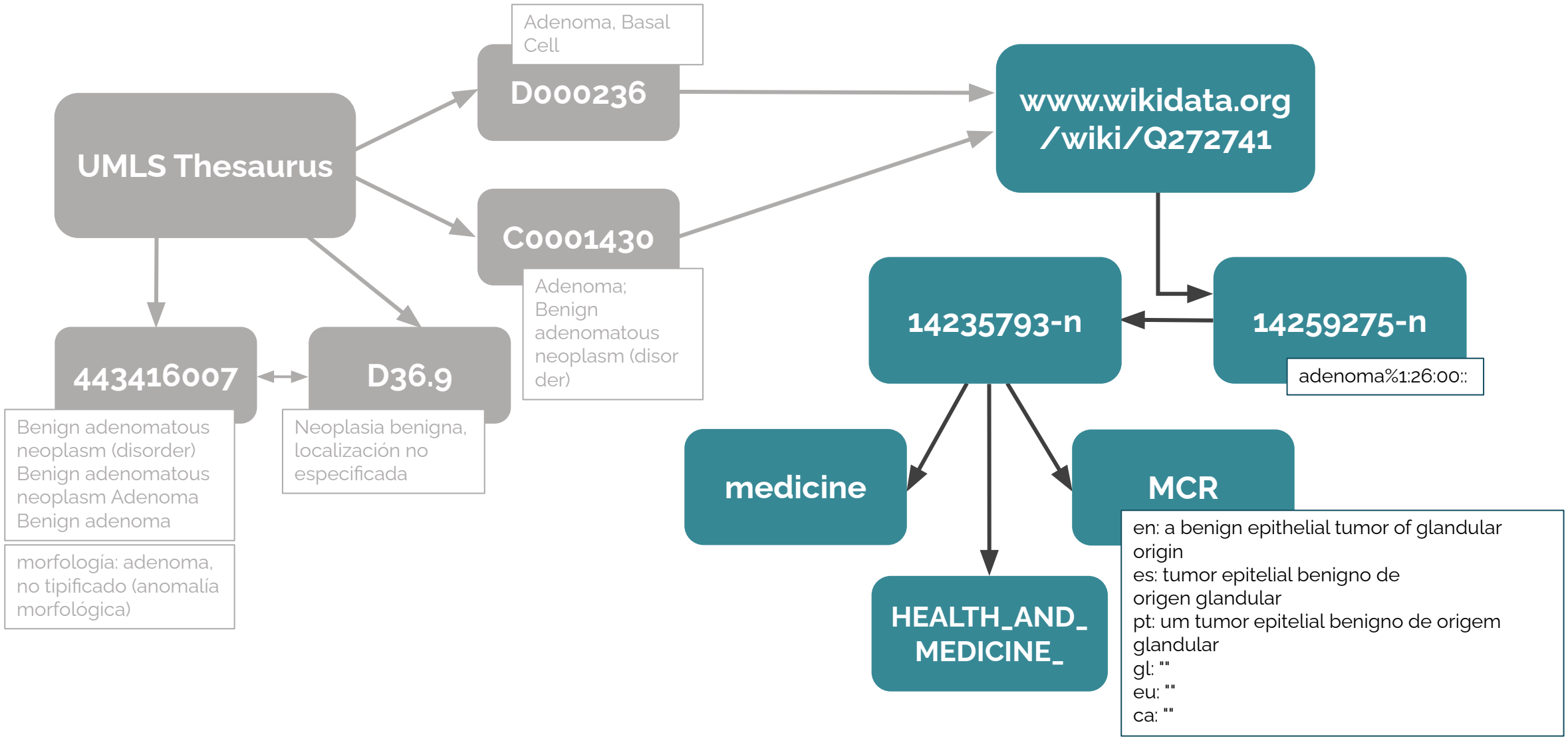
Mapping


UMLS Thesaurus
MeSH
CUI
Wikidata
SNOMED-CT
ICD-10
WordNet 3.0
WordNet 3.1
WordNet Domains
MCR
CSI

Mapping


UMLS Thesaurus
D000236
Adenoma, Basal Cell
C0001430
Adenoma; Benign adenomatous neoplasm (disorder)
443416007
Benign adenomatous neoplasm (disorder) Benign adenomatous neoplasm Adenoma Benign adenoma
morfología: adenoma, no tipificado (anomalía morfológica)
D36.9
Neoplasia benigna, localización no especificada
www.wikidata.org/wiki/Q272741
14235793-n
14259275-n
adenoma%1:26:00::
medicine
MCR
HEALTH_AND_MEDICINE_
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: ""

Mapping


Adenoma, Basal Cell
D000236
UMLS Thesaurus
www.wikidata.org /wiki/Q272741
C0001430
Adenoma; Benign adenomatous neoplasm (disor der)
14235793-n
14259275-n
adenoma%1:26:00::
443416007
D36.9
Benign adenomatous neoplasm (disorder) Benign adenomatous neoplasm Adenoma Benign adenoma
Neoplasia benigna, localización no especificada
morfología: adenoma, no tipificado (anomalía morfológica)
medicine
MCR
HEALTH_AND_ MEDICINE_
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: ""

Steps to integrate WordNets

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

1. Collect Wikidata items with SPARQL
2. Map WordNet 3.1. to WordNet 3.0 with index.sense
3. Map WordNet 3.0 to WN Domains and
4. Map to CSI (Coarse Sense Inventory)
5. Map to Multilingual WordNet (MCR)

Steps to integrate WordNets

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Step 1. Collect Wikidata items with SPARQL

```
SELECT ?item ?itemLabel ?identificador_de_synset_de_WordNet_3_1 ?UMLS_CUI
?identificador_MeSH ?ICD_10 ?Snomed_CT WHERE {
  SERVICE wikibase:label { bd:serviceParam wikibase:language
"[AUTO_LANGUAGE],en". }
  OPTIONAL { ?item wdt:P8814 ?identificador_de_synset_de_WordNet_3_1. }
  OPTIONAL { ?item wdt:P2892 ?UMLS_CUI. }
  OPTIONAL { ?item wdt:P486 ?identificador_MeSH. }
  OPTIONAL { ?item wdt:P494 ?ICD_10. }
  OPTIONAL { ?item wdt:P4229 ?ICD_10_CM. }
  OPTIONAL { ?item wdt:P1690 ?ICD_10_PCS. }
  OPTIONAL { ?item wdt:P5806 ?Snomed_CT. }
}
```

Steps to integrate WordNets

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Step 2. Map WordNet 3.1. to WordNet 3.0 with index.sense

WN 3.0 adenoma%1:26:00:: 14235793

WN 3.1. adenoma%1:26:00:: 14259275

Steps to integrate WordNet

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Step 3. Domain Labels.

170 domain labels, clinical domain:

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

WordNet extended Domains, 117,536 synset-weight pairs
5 most probable for adenoma

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

A. Gonzalez-Agirre, G. Rigau, and M. 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.

Step to integrate WordNet

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Step 4. Map to CSI (Coarse Sense Inventory)

- 45 high-level semantic labels (Coarse Sense Inventory)
- High-level semantics is domain-based
- Manually crafted labels

We select label **HEALTH_AND_MEDICINE_**

C. Lacerra, M. Bevilacqua, T. Pasini, and R, 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.

Synsets Clinical Domain

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Total Wikidata items	26,953
CSI	3,133
WordNet clinical domains	3,396
Total clinical domain only	2,398

Multilingual Central Repository

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Step 5. Map to Multilingual WordNet (MCR)

- English
- Spanish
- Catalan
- Basque
- Galician
- Portuguese


```
"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": ""
            }
        ]
```

Result

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE


```
{
    "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": ""
                }
            ]
        }
    ]
}
```

Clinical concepts mapped

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

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.

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Index

Corpora

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

- **CodiEsp** ES: clinical narratives
- **E3C Corpus** ES: clinical narratives
- **CT-EBM-SP** ES: clinical trials
- **Mantra** ES: clinical narratives
- **MedMentions** EN: biomedical papers

Corpora annotated with WordNet synsets

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

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.

Corpora annotated with CSI

Obliteration of the upper extraconal fat leads to inferior displacement and ocular proptosis.

La obliteración de la grasa extraconal superior condiciona desplazamiento inferior y proptosis ocular.

Corpora annotated with CSI

Obliteration of the upper extraconal fat leads to inferior displacement and ocular proptosis.

La obliteración de la grasa extraconal superior condiciona desplazamiento inferior y proptosis ocular.

Corpora annotated with CSI

Obliteration of the upper extraconal fat leads to inferior displacement and ocular proptosis.

La obliteración de la grasa extraconal superior condiciona desplazamiento inferior y proptosis ocular.

CUI C0015668

SNOMED 238888007 Fat necrosis

MESH D005218 Necrosis, Fat

CUI C0015300

WN 3.1 14313017-n

exophthalmos%1:26:00::

WN 3.0 14336444-n

MESH D005094 exophthalmos

ICD-10 H05: Constant exophthalmos

optics

HEALTH_AND_MEDICINE_

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Index

Conclusions

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

- WordNets in various languages are integrated into ClinIDMap
- The mapping tool is scalable for different languages and datasets
- Annotated corpora for sequence labelling models

GITHUB

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

https://github.com/Vicomtech/ClinIDMap

We release

- Source code
- The resulting table of Wikidata to WordNets mapping

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

THANK YOU
ESKERRIK ASKO
GRACIAS

www. vicomtech.org

ezotova@vicomtech.org

REFERENCES

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

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. (2019). MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. ArXiv, abs/1902.09476.

Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estape, 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.

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

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.

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.

C. Lacerra, M. Bevilacqua, T. Pasini, and R, 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.