ClinIDMap: Towards a Clinical ID Mapping for Data Interoperability Mapping Tool

INTRODUCTION & MOTIVATION

ClinIDMap: a tool for mapping identifiers between clinical ontologies and lexical resources: UMLS Metathesaurus, SMOMED-CT, ICD-10, Wikipedia

GitHub: https://github.com/Vicomtech/ClinIDMap

Goals:
- semantic interoperability across the clinical concepts
- enrich already annotated medical corpora in multiple languages with new labels
- create new datasets for machine learning models

Experiments with sequence labelling (SL) models for detecting:
- Diagnosis and Procedures
- UMLS Semantic Groups

Languages: Spanish, English, bilingual.

What for:
- Detect and categorize a span with clinical terminology with a high-level class: Diagnosis, Procedure, Anatomy, Chemical etc.
- Link classified span with clinical taxonomy and assign a unique ID (code)

Data for machine learning models in the clinical domain is especially difficult:
- Clinical information is private
- Manual annotation requires a high level of expertise in medicine
- Few data is available for languages other than English

EXPERIMENTS AND RESULTS

Annotate corpora with mapped codes:
- CodiEsp (es) [8], E3C Corpus (es) [6], CT-EBM-SP (es) [2], MANTRA (es) [5], MedMentions (en) [9]. Corpora is also combined in order to augment the training data and perform bilingual experiments.
- Corpora annotated in Diagnosis and Procedures was mapped with Semantic Groups (UMLS)and vice versa the corpora annotated with Semantic Groups were mapped with Diagnosis and Procedures (ICD-10).

Sequence labelling models:
1. Classification of Diagnosis and Procedures (SL-DP), according to ICD-10-CM and ICD-10-PSC notation,
2. Labelling the UMLS Semantic Groups12 (SL-SG), such as Anatomy, Disorder, Procedure, Chemical, etc.

All models testes on two test sets: their own and gold-standard Spanish set from Codiesp corpus.

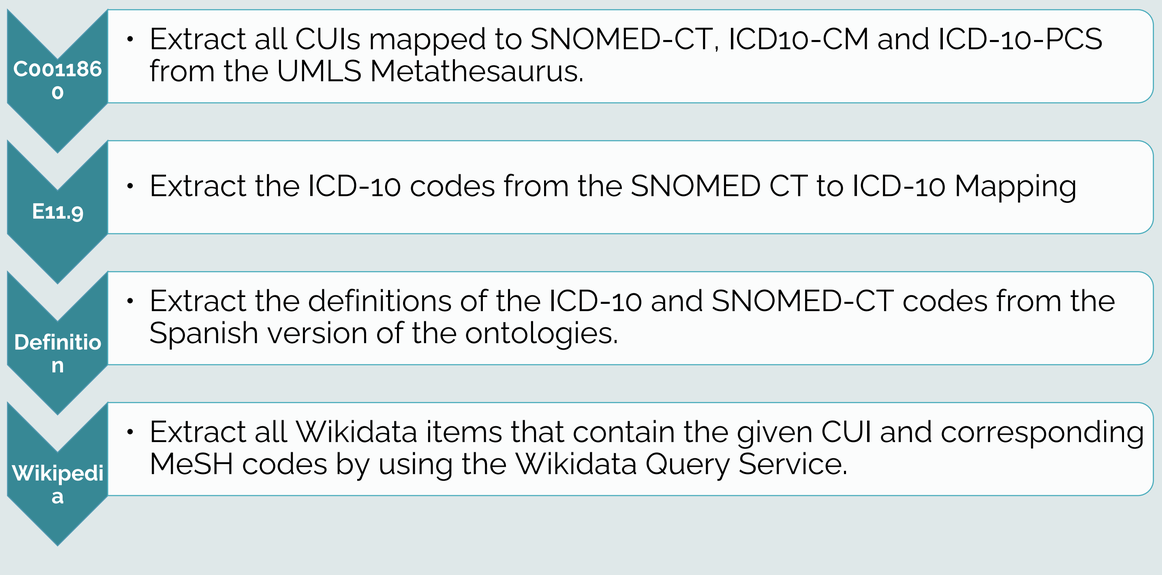
Architecture: BERT [3], Nvidia GeForce RTX 2080Ti 11 Gb RAM, 100 epochs, batch size 8

Results: The F1 score of the models trained on the corpus annotated with the mapping method is very similar to the gold corpus, annotation with mapping transfer knowledge across the lexical resources.

Corpus		SL-DP (2 classes) P	R	F1		SL-SG (15 classes) 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.

METHODOLOGY



- Mapping based on IDs and codes is used
- Any of KBs IDs may be mapped if there are lexically and semantically aligned ontologies
- Some codes are aligned with various codes in the other ontologies.
- The mapping output format is JSON
- Implemented with Python and Elasticsearch


```
{
    "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/[illegible]",
                ...
                "zhwiki": "https://zh.wikipedia.org/wiki/[illegible]"
            }
        ]
    ]
```


CONCLUSIONS

- The models and corpora are quite **interoperable** with respect to different coding systems and languages
- The mapping tool is **scalable** for different languages

- **Future work**:
  - experiments with more languages
  - Annotate more corpora with the methos and do additional experiments
  - Add new taxonomies and ontologies (NCBI, BIOS, MESH)
  - link CUI descriptions with Wikidata/Wikipedia items with deep learning. Only less than 1% of the one million UMLS CUIs can be found in Wikidata.

REFERENCES

[1] Campillos-Llanos, L. (2019). First Steps towards Building a Medical Lexicon for Spanish with Linguistic and Semantic Information. pages 152–164, August
[2] 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 4171–418
[3] Donnelly, K. et al. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics, 121:279.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] Mohan, S. and Li, D. (2019). MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. ArXiv, abs/1902.09476.

Elena Zotova (1,2), Montse Cuadros (1), German Rigau (2,3)
{ezotova, mcuadros}@vicomtech.org, german.rigau@ehu.eus
1 SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
2 Department of Languages and Computer Systems, University of the Basque Country (UPV-EHU)
3 HiTZ Basque Center for Language Technologies
San Sebastian, Basque Country, Spain

Universidad del País Vasco · Euskal Herriko Unibertsitatea
HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology
vicomtech MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE