Towards a multilingual corpus for Named Entity Linking evaluation in the clinical domain *

Pedro Ruas[1][0000−0002−1293−4199], André Lamúrias[1][0000−0001−7965−6536], and Francisco M Couto[3][0000−0003−0627−1496]

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon 1749-016, Portugal
ps_ruas@fc.ul.pt

Abstract. We propose a new multilingual, parallel corpus for Named Entity Linking benchmarking which comprises English, Portuguese and Spanish clinical case reports[1]. The medical diagnostic entities in the reports were annotated with the respective code of the International Classification of Diseases 10 - Clinical Modification (ICD10-CM) terminology and its Portuguese and Spanish versions. The result is a preliminary annotation set, which will be further validated and expanded by humans. Additionally, the ICD10-CM codes in the annotations will be mapped to the respective Medical Subject Headings (MeSH) identifiers when possible.

Keywords: Text Mining · Multilingual clinical case reports · Named Entity Linking · Information retrieval · Named Entity Recognition

1 Introduction

In Text Mining pipelines, Named Entity Linking (NEL) systems are applied after the Named Entity Recognition (NER) step and before the Relation Extraction step. The goal of NEL systems is to map the entity mentions in text with the respective concept identifier in a Knowledge Base (KB). Currently, most NEL approaches are still being developed with English text in mind, but there is a growing interest in the development of tools able to process non-English text. However, the main challenge to the development of new tools is the scarcity of multilingual NEL datasets containing clinical text. In addition, building a gold standard from scratch is time-consuming and demands high expertise, and for non-English languages, there is a lack of controlled vocabularies.

In this work, we applied a pipeline of Information Retrieval, NER and NEL tools to build a multilingual NEL corpus. The goal of the pipeline is to obtain preliminary annotations of medical diagnostic entities in clinical case reports, which will facilitate and speed up the further task of human validation.

* Supported by FCT through the DeST: Deep SemanticTagger project, ref. PTDC/CCI-BIO/28685/2017, LASIGE ResearchUnit, UIDB/00408/2020

[1] https://github.com/lasigeBioTM/MultiNEL-corpus

2 Building the corpus

2.1 Abstract Retrieval

SciELO[2] is a digital library for scientific articles, its majority written in Spanish, Portuguese and English. One of the main advantages is that many articles have versions in different languages. We extracted the abstract of clinical case reports (search filters: **AND subject_area:("Health Sciences") AND type:("case-report") AND la:("es" OR "pt" OR "en")*), and only considered those with the three versions simultaneously available (English, Portuguese and Spanish). We obtained 1917 abstracts in the three languages, corresponding to 639 clinical case reports, which we considered enough to test the annotation approach described below.

2.2 Annotation of medical diagnostic entities using NER and NEL

We used the python interface of MER [1], which recognises entity mentions in the text according to a given lexicon, i.e., a list of terms that represent the concepts of a vocabulary or a KB. In this work, we used as target KB the International Classification of Diseases 10th Revision - Clinical Modification (ICD10-CM), since it is available in several languages. This vocabulary contains codes relative to medical diagnostics and an hierarchy defining subsumption relations between them. For each language, we used the most recent available edition: the 2020 edition for the English ICD10-CM provided by the Center for Disease Control and Prevention (CDC)[3]; the 2020 edition for the Spanish *Classificación Internacional de Enfermedades - 10^a Revisión - Modificación Clínica* (CIE10-CM), provided by the Spanish Ministry of Health[4]; the 2017 edition for the Portuguese *Classificação Internacional de Doenças - 10^a Revisão - Modificação Clínica* (CID10-CM), provided by the Portuguese Ministry of Health[5]. MER recognised the entity mentions related with medical diagnostics in the clinical case reports, and then linked each mention to the respective code in the ICD10-CM (or the respective language version). The resulting annotations were converted to the brat Standoff format.

3 Discussion

The overall statistics pertaining the annotation process are available in Table 1. MER was able to recognise entity mentions in the text expressed in the three languages, but its NER performance was slightly higher in English text than in other languages as expected. Surprisingly, the NEL performance was higher in Portuguese text. As example, a sentence from a retrieved abstract is expressed in the three languages: (1 - English) "Among the identified nursing diagnosis was

[2] https://scielo.org/

[3] www.cdc.gov/nchs/icd/icd10cm.htm

[4] www.mscbs.gob.es/estadEstudios/estadisticas/normalizacion/home.htm

[5] www.ctc.min-saude.pt/category/catalogos/

included: acute confusion, *constipation* and knowledge deficit."; (2 - Portuguese) "Entre os diagnósticos de enfermagem identificados incluíram-se confusão aguda, *constipação* e conhecimento deficiente."; (3 - Spanish) "Los resultados del estudio permitieron identificar los seguientes diagnósticos de enfermería: confusión aguda, *constipación* e conocimiento deficiente.". MER was able to identify the italicised entity "constipation" in the English sentence (1) because there is a ICD10-CM term with the same designation: "Constipation" (code K59.0). However, the Portuguese and Spanish equivalents "constipação" (sentence 2) and "constipación" (sentence 3) were not recognised nor linked because the respective terms in the ICD10-CM have a different designation: "Obstipação" and "Estreñimiento" (code K59.0).

Table 1. Statistics for the annotation of medical diagnostic entities in the clinical case reports

	English	Portuguese	Spanish
Abstracts retrieved	639	639	639
Abstracts with annotations	217	197	199
Ratio of annotated abstracts	0.340	0.308	0.314
Entity mentions	533	432	465
Entity mentions per annotated abstract	2.456	2.193	2.340
Linked entity mentions	463	432	389
Linked entity mentions per annotated abstract	2.134	2.193	1.955
Ratio of linked entity mentions	0.867	1.000	0.837

The resulting corpus is available at `https://github.com/lasigeBioTM/MultiNEL-corpus`. The future work consists in the human validation of the annotation set, as well as its expansion with new annotations. This validation will be performed either by expert analysis or by crowd-sourcing, a less expensive approach that has shown comparable results to the expert analysis [2]. Additionally, the ICD10-CM codes present in the annotations will be further mapped to the respective Medical Subject Headings (MeSH) concepts using the MeSDiCon subset for CodiEsp [3], which will improve the cross-linking evaluation capability.

References

1. Couto, Francisco M. and Lamurias, Andre: MER: a shell script and annotation server for minimal named entity recognition and linking. Journal of Cheminformatics **10**(1), 58 (2018)
2. Campos, Luis F, Lamurias, Andre and Couto, Francisco M: Can the Wisdom of the Crowd Be Used to Improve the Creation of Gold-standard for Text Mining applications?. In: 9th INForum - Simpósio de Informática (INForum 2017), Aveiro, Portugal (2017)
3. Miranda, Antonio and Krallinger, Martin. (2020). MeSDiCon subset for CodiEsp: MESH terms in MeSDiCon mapped to ICD10 CM and ICD10 PCS (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3657429