A Simple Terminology-Based Approach to Clinical Entity Recognition

José Castaño[1], Laura Gambarte[1], Carlos Otero[1] and Daniel Luna[1]

[1]Departamento de Informática en Salud, Hospital Italiano, Buenos Aires, Argentina

Abstract

We describe how we use terminology resources as a basic approach to entity recognition and normalization in Spanish. In particular we use a proprietary large vocabulary and thesaurus that extends SNOMED CT, SNOMED CT itself and UMLS. The proprietary terminology uses historical data of clinical terms used in the EHR problem list. Clinical terms are noisy descriptions typed by healthcare professionals, in Spanish, in the electronic health record system (EHR) and contain clinical findings and suspected diseases, among other categories of concepts. Descriptions are very short texts presenting high lexical variability containing synonymy, acronyms, abbreviations, and typographical errors. Each term is mapped to SNOMED CT concepts. This approach was evaluated using the DisTEMIST corpus in the entity recognition and entity linking tasks.

Keywords

Terminology resources, Named entity recognition, Entity linking, DisTEMIST

1. Introduction

Text mining and Natural Language Processing (NLP) techniques have been used to extract and access information in clinical documents to obtain valuable clinical information. Recently many approaches have been tested in languages other than English. The use of manually labeled clinical texts annotated by professional experts is a standard tradition to promote and evaluate the use of different techniques for a set of tasks. Usually, those tasks are Named Entity Recognition, Entity Linking or Entity Normalization. Techniques used are dictionary or gazetteer based, rule-based or machine learning, and any combination of them. Deep learning and transformation-based learning technologies have been used a lot in recent years, yielding very good results. The DisTEMIST challenge (Disease Text Mining Shared Task) proposes two tasks, DISTEMIST-entities (Named entity recognition) and DISTEMIST-linking (Entity linking), over a broad category of entities, covering, diseases, disorders, and anomalies, as is understood from the DisTEMIST homepage. The cover name used is Diseases, but it does not correspond to a category in a given ontology. The DisTEMIST task follows such previous efforts asPharmaCONER [1], SpRadIE [2] and in particular the CanTEMIST (CANcer TExt Mining Shared Task) track at IberLEF 2020 [3] which had very good results using deep learning algorithms and language models.

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

jose.castano@hospitalitaliano.org.ar (J. Castaño); laura.gambarte@hospitalitaliano.org.ar (L. Gambarte); carlos.otero@hospitalitaliano.org.ar (C. Otero); daniel.luna@hospitalitaliano.org.ar (D. Luna)

Table 1
DisTEMIST Training corpus

Task	clinical cases (files)	Entity Mentions	Unique Entity Mentions
DISTEMIST-entities	750	8065	5349
DISTEMIST-linking	584	5136	3453

Table 2
DisTEMIST Training corpus word length

Instances	Word length
2387	1
2333	2
1714	3
490	4
317	5
236	6
588	≥ 6

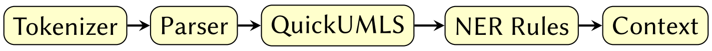
2. The DisTEMIST Dataset

The DisTEMIST corpus data [4] was distributed including training sets, multilingual resources, a so-called DisTEMIST dictionary, and additional concept information (crossmappings to other ontological resourses). The DisTEMIST corpus itself is a collection of 1000 files corresponding to clinical cases. It has been randomly divided into a training set, 750 clinical cases, and a test set of 250 cases. The test set was released inside 3000 clinical cases, the so-called test background, so as the participants would not know which were the cases used for the performance evaluation. There were two interdependent subtasks, the first sub-task, named entity recognition, requires identifyng the named entities corresponding to the cover category Disease (*ENFERMEDAD*). The second sub-task, entity linking required to provide for each named entity recognized the corresponding SNOMED CT [5] concept. The training files also provided information about whether the corresponding named entity was an EXACT description for the SNOMED concept (3803 instances) or a NARROW description (1121 instances), when the entity mention is not a direct mapping to the SNOMED code. COMPOSITE was used when there were two concepts associated to a given entity mention (211 instances). However, these parameters were not used nor required to be submitted. Table 1 shows the correspondence in the training corpus for named entities and unique entity mentions.

The DisTEMIST dictionary [6] contains 134697 terms, with 103154 concepts. Most of the terms were labeled as disorder (132332), 1626 were labeled as finding, and 738 as morphologic abnormality. These labels correspond to the SNOMED CT hierarchy. Also there is an indication on the FAQ section that the SNOMED CT codes to be returned should correspond to the subset in the dictionary, other SNOMED CT codes were not considered.

The named entity terms had a word length distribution very similar to the one available in our terminology resources (see Table 2).

Figure 1: NLP Pipeline



3. Terminology Resources and the Distemist Corpus

Some electronic health records (EHR) implementations allow free text descriptions in structured data entries. Free text descriptions enable more expressiveness, ease of use, and flexibility for physicians. Descriptions are short texts, mostly 3 to 5 words long. Those descriptions must be encoded according to their meaning to allow information interoperability. They have to be mapped to concepts in a controlled vocabulary according to the meaning, and usually, SNOMED CT is used. SNOMED CT is a controlled reference terminology and coding medical ontology that allows storage and retrieval of healthcare information. It is a standard for electronic health records. It can be used in clinical decision support systems. The Hospital Italiano of Buenos Aires (henceforth HIBA), has a Spanish interface terminology [7, 8] where each term is mapped via a direct relation or using compositional expressions to SNOMED CT as its reference vocabulary. The HIBA interface vocabulary was implemented many years ago and has more than 2 million description terms in its terminology system. It was implemented using those description terms typed by the healthcare professionals in structured textual data. A major benefit of the local interface vocabulary is its size and coverage, but it is also the biggest obstacle to its use and maintenance.

We implemented a simple named entity recognition and entity linking system in Python. We used open-source libraries, such as spaCy[9], MedspaCy [10], Quickumls [11], to use our HIBA terminology. It uses spaCy components in a standard pipeline of tokenization, parsing, NER and context rules for NEGEX[12] (See Figure 1). The system allows to recognize those terms that exist in the controlled vocabulary. It also allows to select terms using UMLS types, or HIBA terminology codes. HIBA terminology codes are mapped to SNOMED CT codes (and it also has crossmaping to other terminologies, such as ICD-10).

Our first approach was to run our system directly on the DisTEMIST entity mention terms that are distributed to evaluate the results of the training corpus. The results using the HIBA terminology were very poor only 3490 from the 8065 entity mentions (43%) had at least a partial correspondence with terms in the HIBA controlled vocabulary. Therefore we also added terms from SNOMED CT and UMLS, and obtained a better correspondence 6677 (82%) had at least a partial correspondence with terms in the vocabulary. Table 3 shows the distribution of UMLS types corresponding to the DisTEMIST entity mention terms. The UMLS types provide more detailed information than the SNOMED CT hierarchy types supplied by the DisTEMIST gazetteer. This set of UMLS types was used to select those matched terms in our HIBA terminology that should be identified by the cover term Disease. Type T061 Therapeutic or Preventive Procedure presented a problem because it does not fit in a cover term of Disease, either there was a specific interpretation in a particular context or a misinterpretation of ambiguous terms. Types T033 Finding, and T184 Sign or Symptom, also looked problematic. There is also an important number of entity mentions in the training set (18%) that do not match even partially to a term in any of the terminological resources.

Table 3
UMLS labels for the entity mentions at the DisTEMIST Training corpus

UMLS TYPE	Label	HIBA	HIBA+SNOMED CT+UMLS
T047	Disease or Syndrome	1780	2968
T191	Neoplastic Process	616	949
T033	Finding	322	997
T046	Pathologic Function	318	706
T184	Sign or Symptom	166	268
T037	Injury or Poisoning	157	519
T048	Mental or Behavioral Dysfunction	101	202
T061	Therapeutic or Preventive Procedure		41
T041	Mental Process		19
T049	Cell or Molecular Dysfunction	4	6
T042	Organ or Tissue Function	1	2
Total		3490	6677

4. Experimentation and Results

We performed several experiments to see what was the performance of the system. Precision and recall measures were obtained using the DisTEMIST evaluation tool. They are presented in Table 4. We were surprised by the low precision we obtained using the HIBA terminology resources. We did not expect high recall, but better precision. A number of variations were used, in particular adding SNOMED CT and UMLS description terms. This increased Recall a little bit but lowered Precision. Then we filtered terms that were in the T033 category and considered only a subset of them, using HIBA and SNOMED CT concepts that were in the training set. Only UMLS categories T047, T033, T037, T048, T041, T191 and T046 were considered. This increased precision to reach 0.633. Adding contextual rules, the use of NEGEX, did not change significantly the results. In some cases, the use of the most simple NEGEX rules lowered the precision.

We also modified the evaluation script to find out how many initial offsets, were correct. The results are depicted in the last line (Training set start only). In this case, the precision was significantly high 0.936. In other words, it means that most predicted initial spans of named entities were correct, and that the problem was to identify correctly the right boundary of the named entity mentions.

We submitted only one set of predictions for the test set, the results were a little lower than in the training set.

We used the HIBA concept codes and their mapping to SNOMED CT codes, as well as the SNOMED CT codes themselves, when the recognized entity mention was using a SNOMED CT term. The results were lower (see Table 5), given the task was dependent on the entity mention recognition.

Table 4
Disease Recognition results in Training and Test sets

Terminology	Precision	Recall	F-score
HIBA Training set	0.3586	0.3883	0.3729
HIBA+SNOMED+UMLS Training set	0.271	0.463	0.342
HIBA+SNOMED+UMLS+Filter Training set	0.633	0.409	0.497
HIBA+SNOMED+UMLS+Fiter+Rules Training set	0.6381	0.4079	0.4977
HIBA+SNOMED+UMLS+Filter+Rules Test set	0.5622	0.3772	0.4515
HIBA+SNOMED+UMLS+Filter+Rules Training set start only	0.936	0.5984	0.7301

Table 5
Disease Linking results on Training and Test sets

Terminology	Precision	Recall	F-score
HIBA+SNOMED Training 1	0.366	0.2593	0.3035
HIBA+SNOMED Training 1+2	0.3519	0.2446	0.2886
HIBA+SNOMED Test set	0.4795	0.2292	0.3102

5. Conclusions and Future Work

It is well known that dictionary lookup and regular expressions, are very limited approaches for NER tasks. They are useful on limited scope tasks and as a preliminary baseline approach. They can also be combined to be used with machine learning approaches. We had very limited time and human resources to test machinery that was not mature nor tested previously. This experience allowed us to find unexpected outcomes on some terms from the HIBA terminology which were not in the system. The DisTEMIST corpus presented a particular challenge on some general terms like *lesión, tumor, herida*, which produced many false positives and negatives. It might be questionable if such general terms are valuable for some of the tasks NER would serve, such as document indexing. This is a problem of granularity, which has also the opposite side: terms with too detailed information which might not be relevant. Also, the DisTEMIST corpus seems quite different from user-generated texts in a healthcare institution, which is a major source of our HIBA terminology. The UMLS types were good for restricting some of the target terms but there were two problematic categories (T033 and T184). In future work, we will try to solve some of the flaws we found in our approach and work with machine learning techniques. We will also look into error and corpus analysis.

References

[1] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, M. Krallinger, Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019, pp. 1–10.

[2] V. Cotik, L. A. Alemany, D. Filippo, F. Luque, R. Roller, J. Vivaldi, A. Ayach, F. Carranza,

L. Francesca, A. Dellanzo, et al., Overview of clef ehealth task 1-spradie: A challenge on information extraction from spanish radiology reports, in: CLEF 2021 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS, 2021.

[3] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results., IberLEF@ SEPLN (2020) 303–323.

[4] A. Miranda-Escalada, L. Gascó, S. Lima-López, , D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, E. Farré, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2022.

[5] D. Lee, R. Cornet, F. Lau, N. de Keizer, A survey of snomed ct implementations, Journal of Biomedical Informatics 46 (2013) 87 – 96. URL: http://www.sciencedirect.com/science/article/pii/S1532046412001530. doi:https://doi.org/10.1016/j.jbi.2012.09.006.

[6] L. Gascó, M. Krallinger, Distemist gazetteer, 2022. URL: https://doi.org/10.5281/zenodo.6505583. doi:10.5281/zenodo.6505583, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

[7] H. Navas, A. Lopez Osornio, A. Baum, A. Gomez, D. Luna, F. Gonzalez Bernaldo de Quiros, et al., Creation and evaluation of a terminology server for the interactive coding of discharge summaries, in: Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems, IOS Press, 2007, p. 650.

[8] D. Luna, G. Lopez, C. Otero, A. Mauro, C. T. Casanelli, F. G. B. de Quirós, Implementation of interinstitutional and transnational remote terminology services, in: AMIA Annual Symposium Proceedings, volume 2010, American Medical Informatics Association, 2010, p. 482.

[9] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spacy: Industrial-strength natural language processing in python, 2020. doi:10.5281/zenodo.1212303.

[10] H. Eyre, A. Chapman, K. Peterson, J. Shi, P. Alba, M. Jones, T. Box, S. DuVall, O. Patterson, Launching into clinical space with medspacy: a new clinical text processing toolkit in python, AMIA ... Annual Symposium proceedings. AMIA Symposium 2021 (2022) 438–447.

[11] L. Soldaini, N. Goharian, Quickumls: a fast, unsupervised approach for medical concept extraction (2016).

[12] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, B. G. Buchanan, A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries, Journal of Biomedical Informatics 34 (2001) 301–310. URL: http://dx.doi.org/10.1006/jbin.2001.1029. doi:10.1006/jbin.2001.1029.