DIAGÑOZA: a Natural Language Processing Tool for Automatic Annotation of Clinical Free Text with SNOMED-CT

Matic Bernik[a], Robert Tovornik[a] , Borut Fabjan[a] and Luis Marco-Ruiz[b]

[a] *Better Ltd, Štukljeva cesta 48 1000, Ljubljana, Slovenia*
[b] *Norwegian Centre for E-health Research, Forskningsparken i Breivika, Sykehusveien 23 9019 , Tromsø, Norway*

Abstract

Secondary use of clinical data needs to deal with large amounts of free text present in Electronic Health Records. DIAGÑOZA is a tool that combines the advantages of statistical Natural Language Processing with the high performance of a Vector Database (VDB) for query answering. This paper presents the approach followed in the BioASQ-DisTEMIST subtasks for disease identification in free text and their annotation with the SNOMED-CT terminology. DIAGÑOZA relies on pretrained Spanish language models for tokenization, Part-of-speech tagging, and partially entity extraction. The text entities resulting from the NLP pipeline are vectorized by FastText. In addition, vector embeddings of SNOMED-CT fully specified names, definitions and synonyms are normalized and stored in the VDB. This allows the VDB to estimate similarities between the entity embeddings and SNOMED-CT concept embeddings to perform matching between them.

Keywords [1]

SNOMED-CT, Natural Language Processing, Computational Linguistics, Spanish, Terminologies, Semantic indexing, Ontology linking, Named entity recognition

1. Introduction

Secondary use of clinical data stored in health information systems during clinical practice is on the agenda of most digitized economies [1–3]. These data can play a crucial role in the discovery of disruptive treatments, assessment of long-term effects of drugs, epidemiology surveillance, and health quality measures to name a few [4]. To that end, data must be made available to analytical methods such as data-driven AI which allow us to test hypotheses and unveil hidden patterns that would otherwise remain hidden to researchers. In the last decade, AI, understood as Machine Learning (ML), has been pointed to unleash the power of clinical data [5,6]. Countries such as the US and the EU have specific strategies for the development of AI methods [3,7] aiming to provide more cost-effective healthcare. The most prominent area within biomedical AI that has shown benefits for clinical users is Medical Imaging [8]. Algorithms for breast cancer assessment, brain tumors, pathology and so on have overcome the research stage and are now commercialized as FDA approved and CE marked products.

[1]CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
EMAIL: matic.bernik@better.care (A.1); robert.tovornik@better.care (A.2); borut.fabjan@better.care (A.3); luis.marco.ruiz@ehealthresearch.no (A.4)
ORCID: 0000-0001-6349-3162 (A.4)

CEUR Workshop Proceedings (CEUR-WS.org) Proceedings

However, medical imaging is the exception rather than the norm when it comes to success in the use of AI in healthcare. Other areas such as internal medicine, psychiatry or pediatrics that require processing and analyzing clinical notes lag behind medical images due to a significant lack of high-quality data which is mostly expressed as free text. Some exams such as echocardiography results, blood tests and so on can be structured without a negative impact on clinicians' workflow [9]. However, in others such as internal medicine, where differential diagnoses are common, this becomes an important barrier for clinicians to document their findings in a way that expresses their human reasoning. Thus, although some parts of EHRs are being structured and standardized, it becomes clear that natural language in the form of free text will always remain to some extent because it is an intrinsic part of human reasoning and it is particularly useful for expressing uncertainty about clinical findings. However, when it comes to enabling secondary use of information for research, free text poses a significant demand for preprocessing tasks in order to make information available to data-driven AI methods. Examples are the table-like structures required by libraries such as R and Python. This lays a need for analyzing free text notes and extracting knowledge from them utilizing Natural Language Processing (NLP) methods.

NLP has been available for many decades and it encompasses several common tasks such as record linkage, ontology learning, entity recognition, information retrieval etc. Early efforts relied on grammar and rule-based systems [10]. In the 90s and 2000s stochastics automata, Markov models and entropy models were used [11]. Building on these methods and using the current availability of high computational power has allowed us to apply more computationally expensive ML methods such as Deep Learning and Convolutional Neural Networks [12]. These developments have crystallized in widely used libraries such as the StanfordNLP, OpenNLP or spaCy. However, while the performance in online businesses such as online stores is approaching the human ability to understand the text; complex domains such as medicine and biology still present a sizable gap towards performing NLP reliably with minimum human supervision. Aware of this challenge, several initiatives organize yearly competitions to apply NLP to biomedical datasets and compare different methods. One of them is BioASQ which is an annual challenge for entity recognition on free text that aims to advance the field of Natural Language Processing (NLP) in healthcare. BioASQ is divided into several subtasks each of them aiming for a different objective. In 2022 BioASQ encompasses the DisTEMIST task which aims for the identification of diseases in a corpus of Spanish free text. DisTEMIST is divided in two subtasks: (a) disease identification in clinical text notes (sub-task-1); and (b) disease identification and linkage into the medical terminology SNOMED-CT (sub-task-2). This paper presents the methods used by the DIAGÑOZA system to approach both subtasks of the DisTEMIST challenge.

2. Methods

While the basis for DisTEMIST sub-task-1 and sub-task-2 is the same, we performed different configurations described in the following. Figure 1 depicts the general architecture which is explained in detail below.

2.1. Sub-task-1 disease identification in clinical text notes

To solve sub-task-1 of the DisTEMIST challenge, we trained a new Named Entity Recognition (NER) model on top of spaCy's Spanish transformer (see green and yellow sections in the figure). With regards to training, we relied only on the DisTEMIST training texts and the annotations for sub-task-1. Our framework relies on spaCy (https://spacy.io/) open source Natural Language Processing library for Python. spaCy provides pre-trained text processing pipelines for several languages, as well as the means to easily modify them and train new NLP components. We used spaCy Spanish pipeline to perform text preprocessing steps. The Spanish spaCy pipeline together with the newly trained Spanish model dealt with splitting documents to sentences, text tokenization, punctuation, stop word removal, and entity extraction from text spans (see yellow boxes in Figure 1 with the stages involved).

Statistical morphology analysis was used for tagging and identifying several types of words (verb, noun, prepositions etc.). Statistical NER was used at the latest stage of the preprocessing pipeline.This makes the use of dependency parsing not required since we used a trained NER recognizer.

NER was done by means of scaPy using transition-based NER components. For parsing the translation-based parsing transforms the text into a numeric representation which is structured as a projective tree whose leaves and branches represent the precedence of words. TTracking of the position in the sentence is done with a stack to keep control on which words have been processed. The system uses Idle speculation. In this regard, to decide which part-of-speech to process, average perceptron is used to map the state to a set of features and decide on the best valid move in the process.

2.2. Sub-task-2 disease identification in clinical text notes

Extracted entity spans were fed into the next pipeline stage which is the component responsible for linking text mentions to their corresponding SNOMED-CT concepts.

Spanish SNOMED-CT distribution used for linking was 20220430T120000Z. For sub-task-2 the system was constrained to use only concepts from hierarchies “trastorno” (disease), “hallazgo” (disorder) and “anomalia morfologica” (morphologic abnormality) (represented in the bottom right of Figure1). We implemented a linking step as an Approximate Nearest Neighbor (ANN) lookup based on the similarity between entity text span and all the SNOMED-CT descriptions embedded in the trained FastText vector space.

We chose fastText over BERT-like embedding techniques as it is much simpler in terms of number of free parameters and hence easier to train. That advantage enabled us to train the model from scratch on a training data set that was more representative of the DisTEMIST semantic linking task despite time constraints. We chose fastText over Word2Vec embedding technique, as it operates on a sub-word level (meaning it deals with character ngrams instead of only whole words) and therefore implicitly covers not only contextual but also morphological similarity. It also addresses the problem of how to vectorize the words which were not previously seen by the model during the training phase.

Milvus database was used for storing FastText vector embeddings of SNOMED-CT descriptions and unique identifiers of their corresponding SNOMED-CT concepts.

Milvus is a vector database and is as such designed to store, index and manage embedding vectors converted from unstructured data. It offers high performance vector similarity search as it uses Approximate nearest neighbour algorithms. Few of the parameters that Milvus allows user to set when creating a new database are a type of index being built for vector search, a type of metric used for calculating distances between vectors and a number of cluster units that the vector data is being divided into. In our case we set the index type to IVF_FLAT meaning that the distances between the query vector and centroids of each cluster are being calculated first in order to identify the most similar cluster. Only vectors from the most similar cluster are then being used for similarity search results. We set the metric to the inner product and number of clusters to 128.

All SNOMED-CT description vectors were normalized before inserting them into Milvus database, as well as the query vector of the entity text aforementioned. Once we transformed entity spans into FastText vector embeddings, it was also the Milvus engine that performed the ANN lookup and returned the resulting list of candidates. The distance measure used was the inner product between vectors which, since vectors were normalized, is equivalent to the cosine distance.

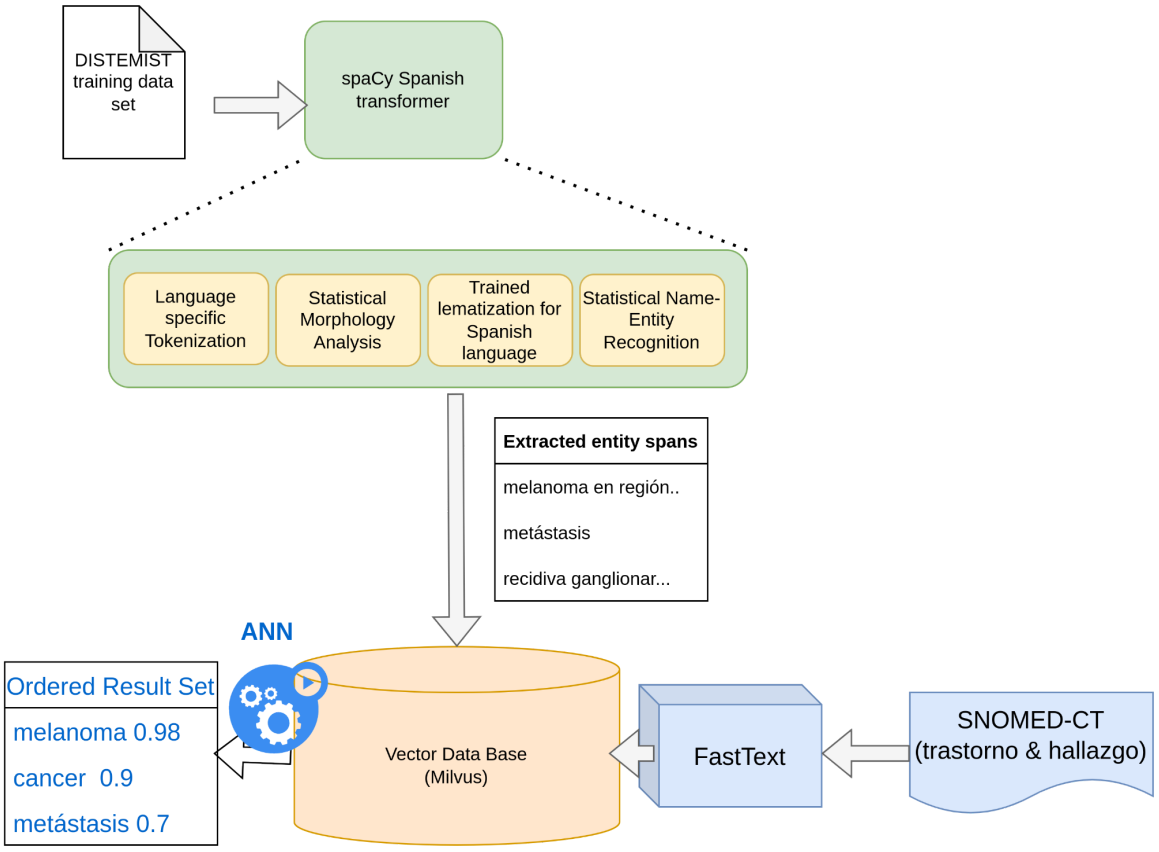

Figure 1: General architecture of the system.

Each ANN lookup returns one or more candidate SNOMED-CT concepts and their respective similarity scores. DisTEMIST task allows for only one SNOMED-CT concept prediction per entity span. To determine the final concept for linking into SNOMED-CT the following steps are followed:

- Candidate concept matches having similarity scores below 0.72 are discarded.
- If there are no candidates left in the Milvus ANN result set, the entity span is also excluded from the results.
- In cases where multiple candidates have similarity scores that differ at most 2%, points from the highest similarity score recorded in the returned result-set, candidate concepts are being prioritized based on their semantic tags. “trastorno“ (disorder) has the highest priority; followed by “hallazgo” (finding); and finally, “anomalia morfologica” (morphologic abnormality) or any other semantic tag.

For converting text spans into vectors, we trained FastText embeddings. We used text data from the following corpora: MeSpEn Parallel Corpora, DisTEMIST corpora text files and descriptions from the Spanish version of SNOMED-CT. Documents were split into training sentences (using spaCy Spanish pipeline) - each sentence was additionally preprocessed. FastText *skipgram* model [13] was trained using the following parameters:

- vector embedding dimension = 700;
- size of context window = 4 characters;
- only words that occurred 5 or more times in the learning corpora were kept;
- max length of word *ngrams* = 2;
- number of negatives sampled was set to 10;
- length of character *ngrams* in vocabulary should be at least 3 characters and at most 8.

All the training text sentences, SNOMED-CT descriptions and entity spans were preprocessed using the same steps: stripping text of accents, replacing new lines with spaces, surrounding special character sequences with spaces (“/”, ”.-”, “.”, ““”), lowercasing text, removing stop words and punctuations. Additionally, semantic tags were removed from the end of SNOMED-CT description of type Fully Specified Name (FSN).

2.3. Semantic relations

Sub-task-2 asked to determine for each link between a text span a SNOMED-CT code and its semantic relationship. To that end, mappings should be tagged with “EXACT” when an exact match between the text span and the SNOMED-CT concept was found. Conversely, mappings should be tagged with “NARROW” when the concept identified had no exact match in SNOMED-CT and it had to be linked to its immediate parent in the SNOMED-CT hierarchy (i.e., normalized to a more generic concept). When more than one code was associated with a text span, that was considered a “COMPOSITE” relationship and it should be concatenated with “+”. We used a similarity score as the basis to determine the type of the match, which can be either “EXACT”, “NARROW” or “COMPOSITE". At this stage the Jaro-Winkler similarity score was used to select among the set of candidate results following these rules:

- All the candidates having similarity scores below 0.72 are being discarded. If there are no candidates left in the Milvus ANN result set, entity text mention is also excluded from the results.
- Matches with similarity scores from 0.72 to 0.75 were marked as COMPOSITE type.
- Matches with similarity scores in the range from 0.75 to 0.9 were marked as NARROW type.
- Matches with similarities either being higher than 0.9 or the entity span text having a Jaro-Winkler similarity higher than 0.9 with any of the SNOMED-CT descriptions, were marked as EXACT type.

3. Results

3.1. Internal testing results

Annotated spans (sub-task-1 of the DisTEMIST competition) from 95% of documents (712 documents - 7523 annotations) were used to train the NER model. Model performance was evaluated on text spans from the remaining 5% of documents (38 documents - 369 annotations). This training/test split was motivated purely by time constraints. According to this evaluation approach, our NER component was able to correctly extract relevant text spans from medical documents with an F1 score of 0.75 (precision being 0.77 and recall 0.73). The entity linking component relies on the quality of text spans extracted in the first step. Therefore, the performance measured for the linking component also reflects the shortcomings of the NER model. For this stage we made a random 80:20 document split between training and test sets. In this case we chose a more standard split ratio because the linking task was our main interest. In the evaluation phase, only the text spans from the training set were combined with the SNOMED-CT descriptions for the nearest neighbor search. When submitting final results, we also used the text spans from the test set. Table 1 shows the evaluation results. We compare the fasttext vector embedding model against the 300 dimension skip-gram biomedical fasttext embeddings from Spanish Biomedical Word Embeddings in FastText [14]. We also compared the performance with and without the inclusion of annotated text spans in the set of SNOMED-CT descriptions for the nearest neighbor search.

Table 1
Comparison of performance by NER model and inclusion/exclusion of SNOMED-CT descriptions

	Only SNOMED-CT Fully Specified Name			With all SNOMED-CT descriptions		
	MiP	MiR	MiF1	MiP	MiR	MiF1
Spanish Biomedical Word Embeddings in FastText	0.4477	0.3963	0.4205	0.5444	0.4956	0.5189
DIAGÑOZA FastText model	0.5204	0.4090	0.4580	0.6240	0.5365	0.5770

3.2. DisTEMIST challenge results

Our approach achieved a MiF1 score of 0.7303 for disease recognition task and ranked 7th in the competition. For the disease linking task our approach achieved MiF1 of 0.4987 placing it 2nd.

Table 2
Comparison of DIAGÑOZA performance on Disease recognition and Disease linking DisTEMIST subtasks.

	Disease recognition task			Disease linking task		
	MiP	MiR	MiF1	MiP	MiR	MiF1
DIAGÑOZA	0.7724	0.6925	0.7303	0.5478	0.4577	0.4987

3.3. Web application

In addition to the linking task, DIAGÑOZA displays a graphical user interface and the visualization of the NLP pipeline results. Figure 2 displays the system GUI detecting diseases only (the one used for the competition). Figure 3 displays entities recognized using the combination of the new NER model with a POS tags detector respectively. On the bottom left the system displays the entities that were detected in the analyzed text and the list of SNOMED-CT concepts (on the right) ordered using the cosine similarity score. The system allows filtering by the different types of hierarchies in SNOMED-CT (upper right), thus allowing constraint results to one or more hierarchies and increasing performance.

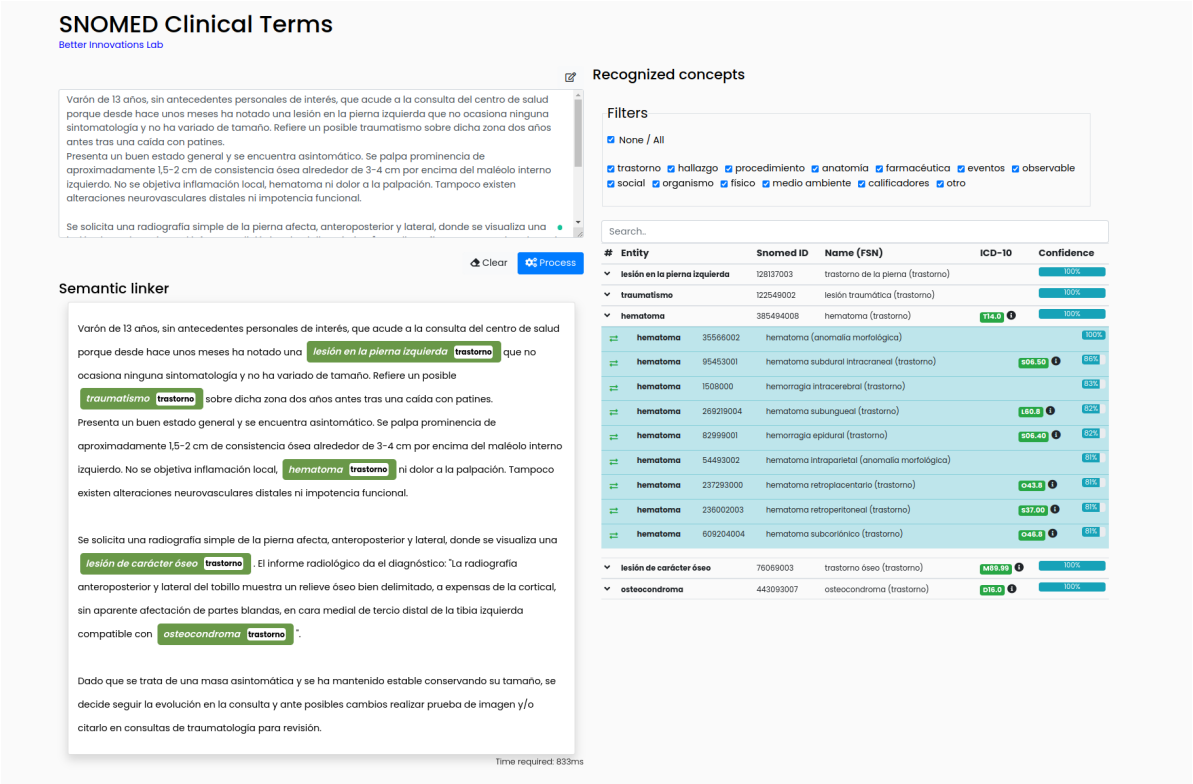

Figure 2: Screenshot of the system identifying SNOMED-CT candidate concepts in the text and scoring the potential matches of the terminology over diseases ("trastorno") only.

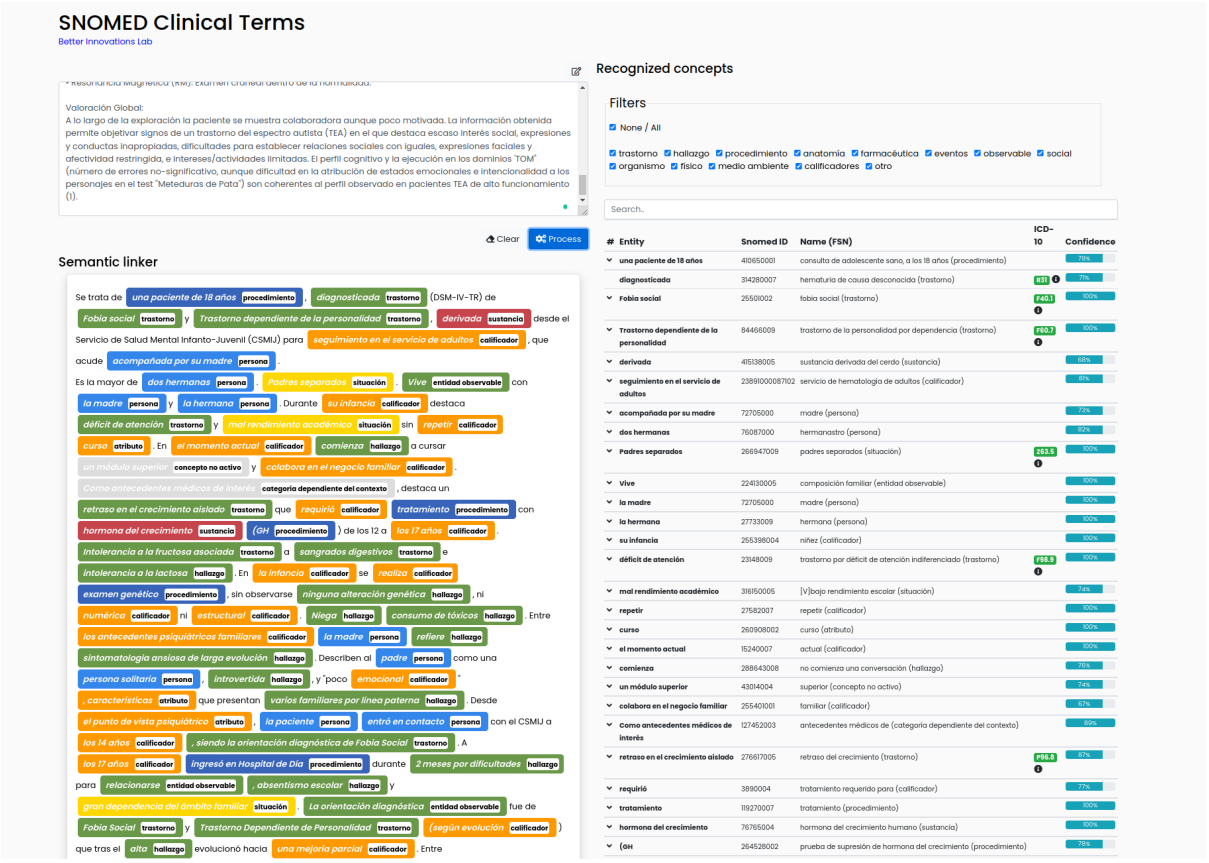

Figure 3: Screenshot of the system identifying SNOMED-CT candidate concepts in the text and scoring the potential matches of the terminology over all entities identified.

4. Discussion

Our initial developments used rule-based morphology and dependency tree structure with POS tags to extract relevant entities, but after some experiments, we concluded that the gold-standard spans are highly context-dependent and that it would be more efficient to rely on statistical morphology models to tag the distinct parts of the text. Another aspect that was significant for performance improvement was the inclusion of more synonyms in the VDB. Originally results without the inclusion of SNOMED-CT synonyms had a poorer performance. The inclusion of all Spanish SNOMED-CT synonyms from the terminology distribution raised this performance as shown in Table 1. This shows a need not only to rely on the training dataset but also to include other sources of known synonyms in the corpora. In this regard, there are several possibilities to improve the performance that remains as future work. Currently, many terminologies such as ICD-10 have been mapped to SNOMED-CT. This can be used to automatically enrich any corpora to enhance the performance of entity matching not only to diseases, but to other clinical concepts such as anatomical locations, procedures, or substances.

Beyond corpora enrichment, there are other possible improvements. Firstly, regarding sub-task-1, we aimed for exploring the performance by substituting spaCy's default Spanish transformer for the one that has been pretrained on biomedical texts. Secondly, regarding sub-task-2, we foresee several lines of action that may improve performance. Examples are: (a) to experiment with FastText hyperparameters and optimize their configuration; (b) to train a context-aware classifier for choosing the final match from the list of nearest neighbors; and (c) to address composite types of matches by splitting text spans into (possibly overlapping) sub-sections and linking them separately. These future works may help in dealing with text spans that contain anatomical locations and help the classifier with situations where diseases are linked to anatomical parts. An example is the text span *"melanoma en región dorso lumbar"* (which means melanoma in the lumbar spine). The former text span was linked to the SNOMED-CT code *1119216000 |Pain in right lumbar region of back (finding)|*. Here our system prioritized the anatomical location (*1119216000 |Pain in right lumbar region of back (finding)|*) rather than the disease (melanoma). Ideally, we should control this behavior by, for example, creating our own parser so it links to the correct term in SNOMED-CT (i.e., *372244006 |melanoma maligno (trastorno)|*).

Another important challenge that we detected in our study is the management of syndromes. We believe that this affects all NLP tools. We detected that the text span "prostatic syndrome" had been linked to the generic concept Disease (*64572001 |Disease (disorder)*) rather than its closest semantic concept "*21173002 |adenoma benigno de la próstata (trastorno)|*". This situation is of particular interest not to our system alone, but any NLP system dealing with SNOMED-CT. SNOMED-CT contains diseases and symptoms separately, but it does not contain the syndromes that characterize them (i.e., the groups of symptoms and signs that together characterize a condition). This means that if NLP systems aim for dealing with text spans referring to syndromes, which is common in medical jargon, and link them to possible underlying diseases, developers of these NLP systems will need to develop a knowledge base that links both types of concepts and account for their causal relationships.

5. Conclusion

DIAGÑOZA is a NLP pipeline that relies on a combination of spaCy NLP pipeline and a VDB. Firstly, this allows for using the latest advances in NLP libraries such as statistical NER and language-specific pre-trained models. Secondly, this allows for deriving vector embeddings from the text spans identified in the NLP pipeline and storing them in a VDB which allows for high performance and agile query answering. The results of this combination were satisfactory, but we believe that further improvements can be achieved by using knowledge of the structure of

SNOMED-CT to customize the NLP pipeline for dealing with the specific structure of text embeddings referring to SNOMED-CT concepts.

6. Acknowledgements

This work was funded by Better Innovations Lab and the Norwegian Center For E-health Research.

7. References

[1]. Committee on the Learning Health Care System in America, Institute of Medicine. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America [Internet]. Smith M, Saunders R, Stuckhardt L, McGinnis JM, editors. Washington (DC): National Academies Press (US); 2013 [cited 2022 Jun 1]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK207225/

[2]. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med. 2018 May;57(S 1):e50–6.

[3]. Zoi K, Kalra D, Wilson P, Martins H. DigitalHealthEurope recommendations on the European Health Data Space Supporting responsible health data sharing and use through governance, policy and practice [Internet]. 2021 Jul p. 18. Available from: https://digitalhealtheurope.eu/wp-content/uploads/DHE_recommendations_on_EHDS_July_2021.pdf

[4]. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform. 2017 Aug;26(1):38–52.

[5]. Elliott JH, Grimshaw J, Altman R, Bero L, Goodman SN, Henry D, et al. Informatics: Make sense of health data. Nature. 2015 Nov;527(7576):31–2.

[6]. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016 Sep 29;375(13):1216–9.

[7]. Forrest CB, McTigue KM, Hernandez AF, Cohen LW, Cruz H, Haynes K, et al. PCORnet® 2020: current state, accomplishments, and future directions. J Clin Epidemiol. 2021 Jan;129:60–7.

[8]. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018 Aug;18(8):500–10.

[9]. Barbisan CC, Andres MP, Torres LR, Libânio BB, Torres US, D'Ippolito G, et al. Structured MRI reporting increases completeness of radiological reports and requesting physicians' satisfaction in the diagnostic workup for pelvic endometriosis. Abdom Radiol. 2021 Jul 1;46(7):3342–53.

[10]. Advances in natural language processing | Science [Internet]. [cited 2016 Sep 29]. Available from: http://science.sciencemag.org/content/349/6245/261

[11]. Daniel J, James M. Speech and Language Processing: An Introduction to Natural Language Processing , Computational Linguistics, and Speech Recognition. Upper Saddle River, N.J; 2008. 1032 p.

[12]. Hobson L, Cole H, Hannes H. Natural Language Processing in Action: Understanding, analyzing, and generating text with Python. Shelter Island, NY; 2019. 544 p.

[13]. Python module · fastText [Internet]. [cited 2022 Jun 2]. Available from: https://fasttext.cc/index.html

[14]. Gutiérrez-Fandiño A, Armengol-Estapé J, Carrino CP, De Gibert O, Gonzalez-Agirre A, Villegas M. Spanish Biomedical Word Embeddings in FastText [Internet]. Zenodo; 2021 [cited 2022 Jun 2]. Available from: https://zenodo.org/record/4543236