Statistical graph matching for indexing Spanish biomedical documents*

Alicia Lara-Clares and Ana Garcia-Serrano

ETSI Informática
Universidad Nacional de Educación a Distancia (UNED)
{alara,agarcia}@lsi.uned.es

Abstract. In this work, we describe a statistical graph matching method for semantic indexing of documents from large-scale biomedical repositories in Spanish language provided at the MESINESP 2020 task (8th BioASQ Workshop [15]). The results obtained show enough accurate behavior, especially with respect to the rest of the results in the task. The execution time and computational requirements have been a priority in our approximation, which has proved to be efficient and robust for tackle further improvements.

Keywords: Biomedical semantic indexing · Knowledge Discovery · Graph matching

1 Introduction

Although the number of medical data is growing at an exponential rate, literature in the medical domain is often found as unstructured or semi-structured data. In these cases, it is necessary to find methods to automatically extract and categorize the data contained in them, using different techniques as, for example, biomedical semantic indexing.

The BioASQ [15] is an EU-funded support action [1] to set up a challenge on biomedical semantic indexing and question answering (QA). The MESINESP task is based on the use of resources such as a structured medical vocabulary DeCS [2] used in two databases for Spanish health content: IBECS [3] and LILACS [7]. The main objective of this task is the development of a semantic indexing tool for Spanish content. Other objectives are: (a) determining the current-state-of-the art, (b) identifying challenges, and (c) comparing the strategies and results to those published for English data[1].

* Supported by the UNED predoctoral grant started in April 2019 (BICI N7, November 19th, 2018)

[1] https://temu.bsc.es/mesinesp/

In this paper, we propose a statistical graph matching method implemented as a module into the HESML framework [9–11]. This method obtains information on the frequency with which DeCS codes are annotated to rank the list of candidates that are extracted from the text following two different methods described in Section 2.1.

The results are encouraging enough, especially when compared to the rest of the experiments and knowing the main difficulties. We will continue working in this task with mixed approaches ([6]), looking forward to obtaining a robust and efficient method capable of correctly indexing DeCS codes. An important feature of this approach is its independence of the language.

The rest of the paper is organized as follows. In section 2, we describe the architecture of the system. Section 3 describes the evaluation process and the results obtained. Finally, section 4 outlines the conclusions and future work.

2 System description

The MESINESP task, is the first task on semantic indexing of Spanish medical texts, provides a dataset divided in training (318.658 documents), development (750 documents) and test (23.509 documents) sets. The average of DeCS codes per document is 8.12, and the document with the maximum number of codes has a total of 53 different ones. At a glance, the training set give us the idea of a scattered distribution of the codes annotated in the documents, as described in Table 1. Our proposed method try to overcome this problem using statistical information about the frequency of a DeCS code annotated in a document.

Total codes that appear more than 10 of documents%	6
Total codes that appear more than 1 of documents%	48
Total codes that appear less than 1 of documents%	33654
Total codes that never appear	22523
Total codes	**33702**

Table 1. Frequencies of the codes in the training set[2]

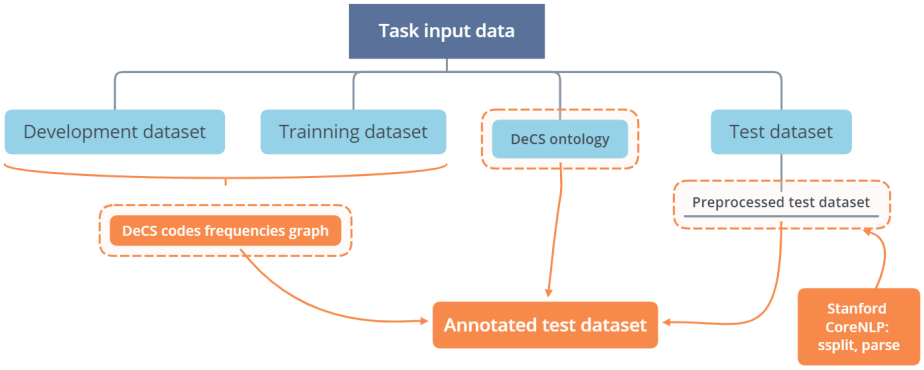
2.1 Proposed method

The method proposed herein is a first approximation focused on the efficiency and robustness of the system. Figure 1 represents the information flow to annotate the test set.

The process stages are the following:

1. Creation of the frequencies graph from the training and development data. In this step, a directed graph is developed, where each DeCS code represent a node, and the edges are the number of times the codes co-occur in each document.

Fig. 1. Information resources.



2. Parse the DeCS ontology and the list of codes and descriptors provided for the competition [3].
3. Split the sentences and identification of the chunks for each sentence using the Stanford CoreNLP library [12].

Once the test dataset is processed, the next step is the alignment of the documents with a list of DeCS code candidates. In this work, there has been carried out using two different methods, (a) exact matching and (b) graph-based matching. In the first one, every possible descriptor is matched with each document. If the descriptor exists in the text, it is selected as a candidate. In the second method, every chunk is compared with the list of possible descriptors, aligning as much DeCS codes as possible.

Other experiments have been planned but could not be carried out due to time constraints. The first one is the alignment of chunks with DeCS codes using a semantic sentence similarity measure, as for example, the Jaccard similarity [13, 4]. The HESML framework provides a set of semantic similarity measures that allows the comparison between every descriptor with all the available chunks. The problem of this approximation is that the annotation of each document takes about 30 seconds, so the method would take more than a week to annotate all the documents.

3 Evaluation and results

MESINESP task has been evaluated using the following measures: (a) Accuracy (Acc.), (b) Example Based Precision (EBP), (c) Example Based Recall (EBR), (d) Example Based F-Measure (EBF), (e) Macro Precision (MaP), (f) Macro Recall (MaR), (g) Macro F-Measure (MaF), (h) Micro Precision (MiP), (i) Micro

[3] Downloaded from https://temu.bsc.es/mesinesp/index.php/resources/

Recall (MiR) and (j) Micro F-Measure (MiF), but we only include in this section the Micro F-measure, since it is the official evaluation measure for this task.

Our results are shown in Table 2 including the first position and the baseline results of the task.

System	MiF	EBF	MaF	Acc	Position (MiF)
X-BERT BioASQ F1	0,1071	0,1051	0,0008	0,0575	1
Graph matching (ours)	0,0836	0,0846	0,001	0,0451	15
Exact matching (ours)	0,0826	0,0829	0,001	0,0442	18
BioASQ_Baseline	0,0161	0,0217	0,0022	0,0116	22

Table 2. This table shows the results or the two methods (Graph matching and Exact matching), as well as the best and the baseline ones

A total of 25 methods have been submitted to the MESINESP competition. Our work has focused on the efficiency and robustness of the method and executes the whole process in less than 30 minutes without requiring a training process. We have used the DeCS ontology and our new hypothesis is that the results will improve using another ontology-based similarity measure to the concept alignment without losing efficiency of the system.

The main difficulty in this task is derived from the use of a purely statistical method that prioritizes the most frequent terms and considers neither the ontology hierarchy for avoiding the annotation of redundant child-parent terms nor the less frequent codes that the experts annotate in the gold standard. For example, the terms "tumor de mediastino" and "mediastino" are annotated using our approximation for the document ID "biblio-1000005", but the experts only annotate the terms "mediastino" and "neoplasias del timo". But, it happens that the term "tumor de mediastino" is explicitly written in the title of the document and, for this reason, it is considered as relevant for our algorithm. On the other hand, the term "pesar" is wrongly considered as relevant for our algorithm in most of the documents, because no semantics is considered in the selection of candidates.

4 Conclusions and Future Work

In this work, we describe a statistical graph matching method for semantically index documents from large-scale biomedical repositories in Spanish language provided at the MESINESP 2020 task [15]). The execution time and computational requirements have been priority factors in our approximation, giving us a first approach that is efficient and sufficiently robust to improve the results in the future.

Addressing the task, we understand that the Spanish language has not been thoroughly studied in a semantic indexing task, and there are only a few available tools. For example, there are some Name Entity Recognizers (NER) that find

UMLS concepts in Spanish biomedical documents, such as QuickUMLS [14] or IXAMedTagger [8]. But, as far as we know, there is not a NER tool for aligning DeCS codes with texts. Even more, the code sets tend to follow biased, unbalanced, and scattered distributions, as shown in a similar task of indexing CIE-10 codes for Spanish clinical documents [5].

In the future work, we are going to focus on the integration of the parser for the DeCS ontology in HESML. We will try to prove that our proposal will overcome the problems with the running time of the experiments based on sentence similarity measures by allowing the use of different ontology-based measures. Finally, we want also to test a new model that recognizes co-occurrence patterns beyond the basic measure of the frequency of occurrence of terms.

References

1. The bioasq challenge — bioasq.org. `http://www.bioasq.org/`, accessed: 2020-6-8
2. DeCS - health sciences descriptors. `http://decses.bvsalud.org/I/homepagei.htm`, accessed: 2020-6-10
3. IBECS. `https://www.isciii.es/QueHacemos/Servicios/Biblioteca/Paginas/IBECS.aspx`, accessed: 2020-6-10
4. . Manning, C.D., Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999)
5. Almagro, M., Unanue, R.M., Fresno, V., Montalvo, S.: ICD-10 coding of spanish electronic discharge summaries: An extreme classification problem. IEEE Access **8**, 100073–100083 (2020)
6. Benavent, J., Benavent, X., de Ves, E., Granados, R., Garcia-Serrano, A.: Experiences at imageclef 2010 using cbir and tbir mixing information approaches. In: CEUR Proceedings. 2010 CLEF September, Padua, Italy. vol. 1176 (2010)
7. BIREME (http://www.bireme.br/). LILACS Unity (http://metodologia.lilacs.bvs.br/): LILACS database) :. `http://metodologia.lilacs.bvsalud.org/php/level.php?&component=19`, accessed: 2020-6-10
8. Gojenola, K., Oronoz, M., Pérez, A., Casillas, A., Taldea, I.X.A.: IxaMed: Applying freeling and a perceptron sequential tagger at the shared task on analyzing clinical texts. In: SemEval@ COLING. pp. 361–365. ixa.eus (2014)
9. Lastra-Diaz, J.J., Garcia-Serrano, A., Batet, M., Fernandez, M., Chirigati, F.: Hesml: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. Information Systems **66**, 97–118 (2017)
10. Lastra-Díaz, J.J., Goikoetxea, J., Taieb, M.A.H., others: A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. Applications of Artificial ... (2019)
11. Lastra-Díaz, J.J., Goikoetxea, J., Hadj Taieb, M.A., García-Serrano, A., Aouicha, M.B., Agirre, E.: Reproducibility dataset for a large experimental survey on word embeddings and ontology-based methods for word similarity. Data Brief **26**, 104432 (Oct 2019)
12. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60. aclweb.org (2014)

13. P. Jaccard: Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. sci. nat. **44**, 223–270 (1908)
14. Soldaini, L., Goharian, N.: Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR workshop, sigir. pp. 1–4. ir.cs.georgetown.edu (2016)
15. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artiéres, T., Ngomo, A.C.N., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., Paliouras, G.: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics **16**, 138 (Apr 2015)