Information Extraction of Texts in the Biomedical Domain

Viviana Cotik
Computer Science Department, FCEyN, Universidad de Buenos Aires
Buenos Aires, Argentina
vcotik@dc.uba.ar

Abstract

Automatic detection of relevant terms in medical reports is useful for educational purposes and for clinical research. Natural language processing techniques can be applied in order to identify them. The main goal of this research is to develop a method to identify whether medical reports of imaging studies (usually called radiology reports) written in Spanish are important (in the sense that they have non-negated pathological findings) or not. We also try to identify which finding is present and if possible its relationship with anatomical entities.

1 Problem Description

Automatic identification of relevant entities in medical reports is useful for clinical research. According to [Chapman *et al.*, 2001], approximately half of the medical conditions described in the medical domain are negated. There also exist *hedges* (uncertain facts). Being able to differentiate which conditions are present and which are absent in a medical report is a current topic in the area of natural language processing (NLP) [Wu *et al.*, 2011; Chapman *et al.*, 2013].

The use of information extraction from unstructured radiology reports allows to improve aspects of diagnosis and patient care within an institution by identifying findings and diagnose frequency in different imaging modalities.

2 Background & Related work

There are several works addressing related problems. Most existing systems process texts in English, and there is some work done for German. Khresmoi project[1] uses information extraction from unstructured biomedical texts in a cross-lingual environment. MoSearch [Ramaswamy *et al.*, 1996], RADTF [Do *et al.*, 2010] and Render [Dang *et al.*, 2009] search terms in radiology reports taking into account negation and modality information and using NLP techniques. RadMiner [Gerstmair *et al.*, 2012] retrieves images in radiology reports and Bretschneider et al. [2013] use a grammar-based sentence classifier to distinguish *pathological* and *non-pathological* classes. Both are implemented for German and use a German available version of RadLex as a linguistic resource. MetaMap [Aronson, 2001] recognizes UMLS concepts in medical texts written in English. BioPortal, a repository of biomedical ontologies, provides a tool that tags text based on an ontology selected by the user (there are no Spanish ontologies available). LEXIMER [Dang *et al.*, 2005] uses information theory to classify English radiology reports on the basis of the presence or absence of positive findings. Negex [Chapman *et al.*, 2001] is a simple algorithm to identify negations in medical texts written in English. It has been implemented in several languages [Wu *et al.*, 2011; Skeppstedt, 2011; Chapman *et al.*, 2013]. Diverse techniques such as pattern matching, machine learning (ML) and a combination of them, have been applied for this problem. Some challenges have been performed for clinical and biological texts: 2010 i2B2/VA[2], ConLL 2010[3], BioNLP 2009[4], BioCreAtIvE[5] and currently CLEF 2015[6].

3 Research Objectives

Specifically, the main goal of the proposed work is to perform entity and relationship recognition in the biomedical domain for texts written in Spanish. Entity and relationship recognition is a current research subject in the BioNLP area and there is not much work done for Spanish.

Associated with this problem are *negation and hedge detection* (an event or relationship has to be distinguished by its factual information -i.e. whether a fact is identified, or mere possibility or non existence are presented-), *slot filling* (the storage of the extracted information in structured format), and *acronym detection* and the determination of their expansion. Other related issues are *report anonymization* and the ability to relate reports with imaging studies. Negation detection in the biomedical domain is also a current research subject and has to take into account many issues, such as the scope of negated terms. We also plan to work in the main associated problems mentioned above.

[1] http://www.khresmoi.eu/

[2] https://www.i2b2.org/NLP/Relations/

[3] http://www.clips.ua.ac.be/conll2010/

[4] http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/

[5] http://biocreative.sourceforge.net/

[6] http://clef2015.clef-initiative.eu/CLEF2015/cfl.php

4 Research Approach & Methods

To achieve these goals several tools and techniques will be used.

In order to recognize radiological entities, RadLex[7], an ontology for the radiological domain, that is not available in Spanish, will be used. Although there are a number of ontologies, such as SNOMED CT[8] and ICD-10[9], RadLex is the most appropriate for this domain.

NLP tools and techniques, such as tokenization, labeling, entity normalization, lemmatization, frequency of bigrams and trigrams, and part-of-speech tagging (POS tagging) are applied to analyze Spanish reports.

Information extraction techniques, such as indexing, and machine learning are used to detect entities of interest in the reports and to classify reports as *interesting* (with positive and certain *findings*) or not.

Hedges and negations are planned to be detected with the use of dictionaries, regular expressions, dependency parsing and machine learning.

5 Progress

RadLex was translated into Spanish using Google Translate[10] and its translation was improved using mappings from English to Spanish Wikipedia and from RadLex to UMLS terms. A direct automatic translation from English ontologies present a number of difficulties, among others: some terms are frequently used in Spanish with synonyms that are less frequently used in English, and sometimes terms in Spanish are preferred in an adjectival way rather than as a noun.

We filtered RadLex terms in order to obtain only those that correspond to anatomical and pathological entities. These terms were searched in radiology reports using an inverted index.

Based on the pathological entities identified in the reports, and in the negation detection, a classification algorithm has been implemented in order to determine if a report has positive and certain findings or not.

A Test Set annotated by a physician of the radiology area has been used to test the results of our classification algorithm. A portion of the Test Set has been annotated by more than one physician, with an Inter Annotator Agreement of 0,7. Given the amount of annotated text is small, it is not possible to use machine learning techniques to improve the classification algorithm.

6 Next Steps

Next steps include: 1) improvement of translations (performed by radiologists). This might provide a resource for achieving better entity recognition, 2) enlargement of manually annotated Test Set in order to be able to use ML techniques to improve our classification algorithm, 3) detection of scope of negation to improve classification (i.e. knowing what is actually being negated) and 4) evaluating and improving the detection of findings. We plan to compare the results of our algorithm with the use of additional resources, such as SNOMED CT and ICD-10, both available in Spanish.

[7]Radiological Lexicon: http://www.rsna.org/radlex.aspx

[8]Systematized Nomenclature of Medicine Clinical Terms - SNOMED CT

[9]International Statistical Classification of Diseases and Related Health Problems 10th Revision

[10]https://translate.google.com/

References

[Aronson, 2001] A. Aronson. A effective mapping of biomedical text to the umls metathesaurus: The metamap program. In *Proc AMIA Symp*, pages 17–21, 2001.

[Bretschneider *et al.*, 2013] C. Bretschneider, S. Zillner, and M. Hammon. Identifying pathological findings in german radiology reports using a syntacto-semantic parsing approach. In *Proc of Workshop on Biomedical Natural Language Processing*, pages 27–35, 2013.

[Chapman *et al.*, 2001] W.W. Chapman, W. Wridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Biomedical Informatics*, 34(5):301–310, 2001.

[Chapman *et al.*, 2013] W.W. Chapman, D. Hilert, S. Velupillai, M. Kvist, M. Skeppstedt, BE Chapman, M. Conway, M. Tharp, DL. Mowery, and L. Deleger. Extending the negex lexicon for multiple languages. *Stud Health Technol Inform*, 192:677–681, 2013.

[Dang *et al.*, 2005] P.A. Dang, M.K. Kalra, T.J. Schultz, S.A. Graham, and K.J. Dreyer. Abbreviations: Application of recently developed computer algorithm for automatic classification of unstructured radiology reports. *Radiology*, 234:323–329, 2005.

[Dang *et al.*, 2009] P.A. Dang, M.K. Kalra, T.J. Schultz, S.A. Graham, and K.J. Dreyer. Informatics in radiology: Render: an online searchable radiology study repository. *Radiographics*, 29(5):1233–1246, 2009.

[Do *et al.*, 2010] B.H. Do, A. Wu, S. Biswal, A. Kamaya, and D.L. Rubin. Informatics in radiology: Radtf: a semantic search-enabled, natu-ral language processor-generated radiology teaching file. *Radiographics*, 30(7):2039–2048, 2010.

[Gerstmair *et al.*, 2012] A. Gerstmair, P. Daumke, K. Simon, M. Langer, and E. Kotter. Intelligent image retrieval based on radiology reports. *European Radiology*, 22(12):2750–2758, 2012.

[Ramaswamy *et al.*, 1996] M.R Ramaswamy, D.S. Patterson, L. Yin, and B.W. Goodacre. Mosearch: a radiologist-friendly tool for finding-based di-agnostic report and image retrieval. *Radiographics*, 16(4):923–933, 1996.

[Skeppstedt, 2011] M. Skeppstedt. Negation detection in swedish clinical text: An adaption of negex to swedish. *Journal of bio-medical semantics*, 2(3):1–12, 2011.

[Wu *et al.*, 2011] A.S. Wu, B.H. Do, J. Kim, and D.L. Rubin. Evaluation of negation and uncertainty detection and its impact on precision and recall in search. *Digital Imaging*, 24(2):234–242, 2011.