MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-634

An Approach for Automatic Classification of Radiology Reports in Spanish

Viviana Cotik[a], Darío Filippo[b], José Castaño[a]

[a] *Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina*
[b] *Hospital De Pediatría, Prof. Dr. Juan Pedro Garrahan, Argentina*

Abstract

Automatic detection of relevant terms in medical reports is useful for educational purposes and for clinical research. Natural language processing (NLP) techniques can be applied in order to identify them.

In this work we present an approach to classify radiology reports written in Spanish into two sets: the ones that indicate pathological findings and the ones that do not. In addition, the entities corresponding to pathological findings are identified in the reports.

We use RadLex, a lexicon of English radiology terms, and NLP techniques to identify the occurrence of pathological findings. Reports are classified using a simple algorithm based on the presence of pathological findings, negation and hedge terms.

The implemented algorithms were tested with a test set of 248 reports annotated by an expert, obtaining a best result of 0.72 F1 measure. The output of the classification task can be used to look for specific occurrences of pathological findings.

Keywords:

Natural language processing; Radiology reports; Pathological findings, Negation detection; Text classification.

Introduction

Automatic detection of relevant terms in medical reports is useful for educational purposes, for clinical research and for comparison of findings between institutions.

According to [1], 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) [2,3].

We describe here an approach to identify reports containing pathological findings. We work on a set of medical reports of imaging studies (usually called *radiology reports*) in Spanish. Identifying which reports contain pathological findings will allow the indexing of relevant documents only and discard those which are not relevant (do not contain pathological findings).

In order to test the results of our classification algorithm we use a Test Set annotated by a radiology physician, one of the authors of this paper. The Test Set consists of 248 ultrasonography reports that are annotated indicating medical findings and their anatomical location. We obtain an F1 of 0.72, a recall of 0.83 and a precision of 0.63.

We use NLP tools and techniques such as lemmatization, frequency of bigrams and trigrams, part-of-speech tagging (POS tagging), and hedge and negation tagging, in order to process our data. We also used a radiology ontology. Then we tested some simple algorithms to determine whether there is a factual pathological finding in a report.

There exist different ontologies, terminologies and coding systems in the medical domain such as SNOMED CT[1], MeSH[2], ICD-10[3], LOINC[4], UMLS[5] and RadLex[6]. The latter has specifically been developed to satisfy standardized indexing and retrieval of radiology information. It satisfies the needs in this domain by adopting features of existing terminology systems as well as producing new terms to fill critical gaps. It unifies and supplements other lexicons while it also has mappings to them. However, there is no radiology ontology or machine readable dictionary data that can be used to identify terms that denote pathological findings in Spanish. Using a machine translation from an English ontology presents a number of difficulties:

- Some terms frequently used in Spanish with synonyms are less frequently used in English. For example *arteria mamaria interna* for *internal mammary artery* is commonly used in Spanish, while in English it is referred to as *internal thoracic artery*.
- Sometimes adjectives are preferred to nouns in Spanish. For example, *folículo ovárico* for *ovarian follicle* is commonly used, while in English *follicle of ovary* is the preferred term.
- Terms of interest can be composed of more than one word, which often leads to problems in the order of the translated words.

We use RadLex as the main source of information to detect pathological findings.

Given the amount of annotated text is small, it is not possible to use machine learning (ML) techniques to improve the classification algorithm.

Current results enable physicians to quickly detect diagnoses in the reports and, in the future, images related to them. These

[1] Systematized Nomenclature of Medicine Clinical Terms - SNOMED CT, http://www.ihtsdo.org/snomed-ct/

[2] Medical Subject Headings http://www.ncbi.nlm.nih.gov/mesh/

[3] ICD10: International Statistical Classification of Diseases and Related Health Problems 10th Revision, http://apps.who.int/classifications/icd10/browse/2010/en

[4] LOINC. Logical Observation Identifiers Names and Codes. Database and universal standard for identifying medical laboratory observations, http://loinc.org/

[5] Unified Medical Language System, http://www.nlm.nih.gov/research/umls/

[6] Radiology Lexicon: http://rsna.org/RadLex.aspx

results are planned to be used by physicians in a public hospital in Argentina.

Related work

There are several works addressing related problems. There are existing systems process texts in English and there is some work performed for German.

Khresmoi project[7] uses information extraction from unstructured biomedical texts in a cross-lingual environment. They used RadLex.

MoSearch [4], RADTF [5] and Render [6] allow searching for terms in radiology reports taking into account negation and modality information and using NLP techniques. In the last two, results are linked with images from a picture archiving and communication system (PACS). In RADTF, if the user searches for a RadLex term, it returns its RadLex id. They are mainly used for education and research.

RadMiner [7] retrieves images in radiology reports based on NLP techniques. Bretschneider et al. [8] use a grammar-based sentence classifier to distinguish 'pathological' and 'non-pathological' classes. They report 0.74 recall and 0.54 precision measures. Both are implemented for German and use a German available version of RadLex as a linguistic resource. RadMiner adds new terms taken from the annotation performed by a specialist.

MetaMap [9] recognizes UMLS concepts in medical texts written in English.

Bioportal[8], 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. A UMLS semantic type can be selected.

LEXIMER [10] uses information theory to classify English radiology reports on the basis of the presence or absence of positive findings. They report precision of 0.98 and recall of 0.99.

Negex[1] is a simple algorithm to identify negations in medical texts written in English. It has been implemented in several languages [3,11,12]. Diverse techniques, such as pattern matching, machine learning and a combination of techniques have been applied to negation identification [2,13,14]. Some challenges have been performed: 2010 i2B2/VA Challenge for clinical text[9], ConLL 2010 for biomedical texts[10], and BioNLP 2009 for biological texts[11].

As far as we are aware of, there are no available systems that identify RadLex terms in Spanish radiology reports.

The rest of the paper is organized as follows. The Methods section presents the approach used in this work as well as data, specific techniques and tools used. The Results section explains metrics and shows current results. Final sections are Discussion and Conclusions.

Methods

The proposed solution is composed of several interconnected but independent modules (see Figure 1).

The *syntactic analysis* module does segmentation, lemmatization, normalization and POS tagging as well as parsing. The *entity recognition* module does dictionary lookup, non-exact recognition, and the *hedge* module identifies hedges and negations. Finally, the classification module classifies texts based on the results of previous modules.

The output of these modules is used with the available data to identify pathological findings that might be of interest for physicians.

In the rest of this section we explain in detail the components of each module and the data used. RadLex data has to be obtained, filtered and translated to Spanish. We explain how we filtered the data, and the translation methods used. We also present the annotation process performed by a specialist, and finally, the methods and techniques used to perform each test.

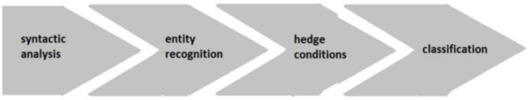

Figure 1–Modules of the proposed solution

Data

We have about 130,000 medical reports from three different studies: ultrasonography (US), computed tomography (CT) and magnetic resonance imaging (MRI). Table 1 shows the number of available reports of each type.

Table 1– Number of reports available of each type of study

Type of Study	Number of Reports
MRI	14635
CT	29327
US	85621

Reports are in non-structured format (the first part is semi-structured). They are brief (approximately 5 lines each) and they state what was found in the study performed on the patient. An example of an annotated ultrasonography report can be seen in the Annotation section.

RadLex, has different versions[12]. We decided to use version 3.6 since it has improvements over the previous ones and it has been used in other works, such as Bioportal, which allows us to compare results. Furthermore, it is being translated by physicians. Version 3.6 has more than 30,000 terms, that are classified[13], among others, by *imaging modality*, *procedure*, *object*, *imaging observation*, *non-anatomical substance*, *anatomical entity* and *clinical finding*. We selected the terms corresponding to *clinical findings* (what we call pathological findings) and *anatomical entity* type (see the section Technical Details).

Translation

As far as we know, there is no complete RadLex translation to Spanish (the translation mentioned in RadLex reference[14] is partial and not every term is precise). In order to be able to use RadLex with Spanish text we had to obtain a translated version.

All RadLex terms were translated to Spanish with Google Translate[15]. We also used 1) a mapping of RadLex and UMLS terms and through UMLS we obtained the corresponding translation of RadLex terms and 2) a mapping

[7] Khresmoi project: www.khresmoi.eu
[8] Bioportal annotator: http://bioportal.bioontology.org/annotator
[9] https://www.i2b2.org/NLP/Relations/
[10] http://www.clips.ua.ac.be/conll2010/
[11] http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/

[12] RadLex Release Notes: https://docs.google.com/document/d/10zRIBkXyj1eLt3LS_A7w3gSGRedXXoYvucH6H4trCZY/edit?hl=en
[13] Radiology Lexicon: http://rsna.org/RadLex.aspx
[14] Who is using RadLex? https://www.rsna.org/Who_Uses_RadLex.aspx
[15] Google Translate, https://translate.google.com/

of English-Spanish Wikipedia[16] terms. Table 2 shows the number of terms translated using different types of translation sources.

Table 2– Number of English-Spanish RadLex translated terms. The second column refers to the number of RadLex terms translated, and the third column to the number of RadLex terms translated of pathological and anatomical type.

Source of translation	Number of RadLex terms	Anat. and pathological terms
Google Translate	30,000	10,357
UMLS	1304	857
Wikipedia	1620	896

In the Entity Recognition section we explain how we used the translations obtained from different sources.

Preprocessing and syntactic analysis

In the syntactic analysis module all the words of radiology reports were normalized. Freeling[17] was used to perform tokenization and lemmatization. We also used Python[18] to process all radiology reports to obtain frequency of words (unigrams), bigrams, and trigrams.

Entity recognition

In order to detect anatomical and pathological entities, we identified in the reports words that are part of some RadLex term. For example: *vessel* does not appear as a RadLex term but is part of more than 100 RadLex terms (as in *blood vessel*), so the term is included as a *term of interest* to be identified in reports.

Each word appearing in a RadLex term was indexed using an inverted index. Each word in the inverted index points to a set of RadLex terms in which the word occurs. Each RadLex term in this set contains their entity class information, i.e. whether they are anatomical or pathological. Using this information, the entity class is assigned to the indexed word. Then, for each report a decision is automatically made to decide whether the word is a single or multi-word term, and the resulting term is tagged with its entity class. In this step stop words are not considered.

The output of this module is the radiological report with the anatomical and pathological terms automatically annotated according to RadLex terms. A set of common pathological terms compiled by the radiologist is also used to identify *interesting* terms in reports.

All the pathological terms identified in the set of 129,583 reports were stored and the most frequent ones were analyzed. Some of them did not appear to be pathological, so we analyzed bigrams and trigrams containing them and the inverted index in order to check if they were incorrectly tagged as pathological findings. We compiled another dictionary with those terms that we considered to be non-pathological, and we used this dictionary to filter out these terms in the tagging process.

Negation and hedge detection

To detect negated terms and hedge signals we compiled a set of negations and a set of hedges (based on a translation to Spanish of RADTF negations and hedges). These two sets of words were used in a simple dictionary lookup to tag these words in the reports. This is very similar to the approach used by NegEx[1]. If one term is contained in another we get the largest of the two terms, for example if *no* and *no se encontró* are in the negation dictionary and *no se encontró* is in the report we will tag this phrase, rather than the phrase *no*.

Classification

Reports are tagged with pathological entities, negations and hedges. Only those reports that contain positive findings are considered relevant. We defined three simple algorithms in order to determine it.

- Algorithm 1. If there is some pathological finding in the report we identify the report as pathological. It is not taken into into account whether or not there are negations in the text.
- Algorithm 2. If there is a pathological finding identified in the report and there is a negation or hedge somewhere in the report (might be in another sentence), the report is identified as non-pathological.
- Algorithm 3. A report is identified as pathological only if it has at least one sentence with a term indicating a pathological finding and no negation or hedge (in the same sentence).

Annotation

In order to test the results of our classification algorithms, we needed some annotations performed by an expert. We elaborated annotation guidelines stating the criteria to be used for the annotations. A number of annotation-revision iterations were performed until the annotations were as expected. Three experts annotated two sets of 17 and 12 reports. The F-measure of the annotations agreement was 0.7. Once the final version of the annotation guidelines was defined, a radiologist annotated 248 ultrasonography reports with the Callisto[19] annotation tool. These 248 reports were used as a Test Set to evaluate the strategies we implemented. Each report was automatically searched for the presence of *pathological findings* annotations, and based on this it was classified as pathological (if there was at least a pathological finding annotated in the report) or non-pathological (if there was no pathological finding annotated in the report).

An example of an annotation in Spanish and it's translation to English can be seen below:

33289|16a4m|20070807|A27611 HIGADO:<RADLEX> lobulo caudado aumentado de tamano</RADLEX>, resto de higado de ecoestructura conservada. VIA BILIAR intra y extrahepatica: no dilatada. VESICULA BILIAR: alitiasica. Paredes y contenido normal. PANCREAS: tamano y ecoestructura normal. <RADLEX>BAZO: minimamente aumentado de tamano</RADLEX>. Diametro longitudinal:13.5 (cm) RETROPERITONEO VASCULAR: sin alteraciones. No se detectaron adenomegalias. No se observo liquido libre en cavidad. Ambos rinones de caracteristicas normales.

33289 |16y 4m |20070807|A27611 LIVER: <RADLEX> caudate lobe with increased size </ RADLEX>, the other lobes of the liver appear normal. Intra and extrahepatic BILIARY TREE: not dilated. GALLBLADDER: no gallstones were seen. Wall and content appear normal. PANCREAS: normal size and echotexture. <RADLEX> SPLEEN: minimally increased in size </ RADLEX>.

[16] Wikipedia: https://www.wikipedia.org/

[17] Freeling. An open source suite of language analyzers: http://nlp.lsi.upc.edu/freeling/

[18] https://www.python.org/about/

[19] Callisto annotating tool http://annotation.exmaralda.org/index.php/Callisto

Longitudinal diameter: 13.5 (cm) VASCULAR RETROPERITONEAL COMPARTMENT: unremarkable. No lymphadenopathy was detected. No free fluid in the peritoneal cavity was observed. Both kidneys unremarkable.

Technical details

RadLex was downloaded in Protégé format. The selection of anatomical and pathological RadLex has been performed with the help of the tutorial performed by MantasCode[20].

Freeling was used for the syntactic analysis module and Python for implementing the remaining modules.

Results

Table 3 shows the results of evaluation of the three algorithms used to identify reports containing pathological findings against the Test Set. As a reference we describe formulas of calculated metrics. They are accuracy (acc): (TP+TN)/(TP+FN+FP+TN), precision (prec): (TP/(TP+FP)), recall: (TP/(TP+FN)), F1: 2*(prec * recall)/ (prec + recall).

Table 3 – Results of comparison of three algorithms with the Test Set. Algorithm 1: negations are not taken into account. Algorithm 2: negations are taken into account on a report basis. Algorithm 3: negations are taken into account on a sentence basis. References: acc.: accuracy, prec: precision, alg.: algorithm TP: true positives, FN: false negatives, FP: false positives, TN: true negatives.

Measure	alg. 1	alg. 2	alg. 3
acc.	0.60	0.57	**0.67**
prec.	0.56	**0.74**	0.63
recall	**0.96**	0.25	0.83
F1	0.71	0.38	**0.72**
TP	122	32	106
FN	5	95	21
FP	95	11	62
TN	26	110	59

Discussion

Algorithm 3 is the one with best results, since it has the best F1 (a measure that balances precision (those identified as positive and how many are really positive) and recall (the proportion of the positive findings that were retrieved) and the best accuracy (rate of correctly classified documents)). Algorithm 1 has naturally a greater amount of TP, but also of FP, that are decreased with Algorithm 3. This is consistent with the algorithm used because all the findings were tagged independently of the occurrence of negations or hedges.

These results show that there is room for improvement, in particular regarding precision results, and they are promising considering that we are working with very noisy data given that terms used to identify pathological findings were obtained through automatic machine translation. We can assume that as a first step to identify reports with pathological findings, the results are good.

LEXIMER has better results for English and our work has better results than that of Bretschneider et al. for German, but in both cases the results are incomparable, since they have been obtained with different data and for different languages.

Conclusion

Although there are tools that generate structured radiology reports aiming at easier information retrieval, unstructured text is still preferred by most radiologists. It allows a better formulation of their ideas, and writing the report as a continuum, instead of doing it with check boxes and templates [15]. Given that situation, NLP is incorporated as a promising resource for information extraction in this context.

Identifying the frequency of findings and diagnoses found in the different imaging modalities, through the use of information extraction from unstructured radiology reports, should improve aspects of diagnosis and patient care within an institution.

The possibility of linking these findings with corresponding images through PACS makes the radiologist's task easier when he has to evaluate studies and prepare reports. It allows comparison with previous studies that may have similar findings. This set of images and text provide an excellent support for decision making.

We are not aware of existing solutions for Spanish reports. Once the work is finished it could be used in the reports of other Spanish speaking hospitals.

In terms of NLP, the challenges are the application of existing techniques to Spanish, the non-availability of RadLex in Spanish, and the scarcity of resources (annotations) that do not allow us to use ML techniques to improve the classification algorithm.

Future Work

Currently we are working on a number of subjects: 1) improvement of translations (performed by radiologists). This might provide a resource for achieving better entity recognition in the future, 2) enlargement of the manually annotated Test Set. This will allow us to use ML techniques to improve our classification algorithm. The use of boosting is being considered, 3) detection of scope of negation to improve classification (i.e. knowing what is actually being negated). Dependency parsers and ML techniques can be used to identify the scope of negation and hedges. We are also working on the implementations of Negex to Spanish, and 4) evaluation and improvement of detection of findings. We will compare the results of our algorithm with the use of additional resources, such as SNOMED CT and ICD-10 instead of RadLex.

As future work it would also be important to do automatic anonymization of radiology reports. Information about the physician who performed the study and medical record number of the patient should be removed. It is an important task because we are working with sensitive information. Nowadays we are not working with image information from PACS, but we have the keys to relate the reports to their corresponding images. A separate project is being carried out by other people in order to relate the information extracted from reports with the associated images.

References

[1] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics, 2001: 34(5):301-10.

[2] Wu AS, Do BH, Kim J, Rubin DL. Evaluation of Negation and Uncertainty Detection and its Impact on Precision and

[20] JAVA: How to programmatically manipulate a Protégé-Frames lexicon/ontology/dictionary using Protege API and Java. http://mantascode.com/java-how-to-programmatically-manipulate-a-protege-frames-lexicon-ontology-dictionary-using-protege-api-and-eclipse/

Recall in Search. Journal of Digital Imaging 2011, 24(2): 234–242.

[3] Chapman WW, Hilert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, Conway M, Tharp M, Mowery DL, Deleger L. Extending the NegEx Lexicon for Multiple Languages. Studies in Health Technology and Informatics, 2013: 192: 677-681.

[4] Ramaswamy MR, Patterson DS, Yin L, Goodacre BW. MoSearch: a radiologist-friendly tool for finding-based diagnostic report and image retrieval. Radiographics, 1996. 16(4): 923-33.

[5] Do BH, Wu A, Biswal S, Kamaya A, Rubin, DL. Informatics in radiology: RADTF: a semantic search-enabled, natural language processor-generated radiology teaching file. Radiographics, 2010: 30(7), 2039-48.

[6] Dang PA, Kalra MK, Schultz TJ, Graham SA, Dreyer KJ. Informatics in radiology: Render: an online searchable radiology study repository. Radiographics, 2009: 29(5):1233-46.

[7] Gerstmair A, Daumke P, Simon K, Langer M, Kotter E. Intelligent image retrieval based on radiology reports. European Radiology, 2012: 22(12): 2750-8.

[8] Bretschneider C, Zillner S, Hammon M. Identifying Pathological Findings in German Radiology Reports Using a Syntacto-semantic Parsing Approach. Proc of Workshop on Biomedical Natural Language Processing, 2013: 27-35.

[9] Aronson A: A Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. Proc AMIA Symp, 2001:17-21.

[10]Dreyer, KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, Halpern EF, Thrall JH. Abbreviations : Application of Recently Developed Computer Algorithm for Automatic Classification of Unstructured Radiology Reports. Radiology, 2005, 234: 323–329.

[11] Skeppstedt M. Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. Journal of biomedical semantics, 2011. 2 (S3):S3 doi:10.1186/2041-1480-2-S3-S3.

[12] Costumero R, Lopez F, Gonzalo-Mart C. An Approach to Detect Negation on Medical Documents in Spanish. Brain Informatics and Health. Lecture Notes in Computer Science, 2014. Volume 8609, pp 366-375.

[13] Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. Journal of Journal of the American Medical Informatics Association, 2007: 14(3):304-11.

[14] Cruz Díaz NP, Maña López MJ, Mata Vázquez J, Pachón Álvarez V. A machine-learning approach to negation and speculation detection in clinical texts. Journal of the American Society for Information Science and Technology, 2012. 63(7): 1398-1410.

[15] Bosmans JM., Peremans L, Menni M, De Schepper AM, Duyck PO, Parizel PM. Insights Imaging. Structured reporting: if, why, when, how-and at what expense? Results of a focus group meeting of radiology professionals from eight countries, 2012: 3(3):295-302.

Address for correspondence

Viviana Cotik: vcotik@dc.uba.ar.