Annotation of Entities and Relations in Spanish Radiology Reports

Viviana Cotik□, Darío Filippo△, Roland Roller⋆, Hans Uszkoreit⋆ and Feiyu Xu⋆
□Departamento de Computación, FCEyN, UBA, Argentina
`vcotik@dc.uba.ar`
△Hospital De Pediatría, Prof. Dr. Juan P. Garrahan, Argentina
`dfilippo@garrahan.gov.ar`
⋆Language Technology Lab, DFKI, Berlin, Germany
`{firstname.surname}@dfki.de`

Abstract

Radiology reports express the results of a radiology study and contain information about anatomical entities, findings, measures and impressions of the medical doctor. The use of information extraction techniques can help physicians to access this information in order to understand data and to infer further knowledge.

Supervised machine learning methods are very popular to address information extraction, but are usually domain and language dependent. To train new classification models, annotated data is required. Moreover, annotated data is also required as an evaluation resource of information extraction algorithms. However, one major drawback of processing clinical data is the low availability of annotated datasets. For this reason we performed a manual annotation of radiology reports written in Spanish. This paper presents the corpus, the annotation schema, the annotation guidelines and further insight of the data.

1 Introduction

Annotated data is required to evaluate information extraction algorithms and for training supervised machine learning methods. There is a scarcity of annotated corpora from the biomedical domain, in particular for non-English texts. There are two main reasons for that: first, the generation of new annotated data is expensive due to the need of expert knowledge, and, second, the ownership of the data is very discussed, specially when it refers to information that might identify the patient. Each country and institution has different regulations and some tasks -eg. anonymization-have to be performed before publishing the data.

So, although the availability of annotated data is a highly valuable asset for the research community, it is very difficult to access it. Furthermore, annotation guidelines have to be carefully designed and reviewed in an iterative process. They have to be clear enough so as to be followed by different annotators with a high annotation agreement. The important amount of information in short texts with complex terminology of the medical domain makes the guidelines definition difficult.

We are interested in supporting physicians with automatic text processing methods, such as named entity recognition (NER), relation extraction (RE) and negation and uncertainty detection in Spanish radiology reports. This could help to detect the main illnesses present among the patients, the patients evolution and to detect problems not expressed in an explicit way. See Sevenster et al. (2012) for an example of possible applications. To the best of our knowledge, there are no annotated datasets of Spanish medical reports which would be appropriate for our use case. For this reason, this work focuses on creating an annotated corpus of Spanish radiology reports.

The radiology reports used for this work are very short and are written by physicians after doing the examination of the patient. They show similarities to other clinical data in the fact that sentences are not always well formed and that many of them have a telegraphic style. There are also spelling mistakes and the use of non-standard abbreviations (that may include some named entities or negation markers, eg. *RD* for *riñon derecho - right kidney-*) is frequent. This, added to the use of specialized language of the medical domain, makes the annotation task difficult.

After describing briefly some previous guidelines definitions and annotated data (Section 2), this work describes the dataset used, the annotation schema and guidelines (Section 3). Then an

analysis of the resulting annotated dataset is presented, that includes the number of entities and relations discovered by the annotators and the inter-annotator agreement among them (Section 4). Finally, Section 5 discusses the results of the dataset analysis and Section 6 presents conclusions.

2 Previous Works

The definition of guidelines is a time consuming and difficult task. There exist some previous definitions for more generic entity types (eg. persons, organizations and geographical locations). For example, MUC-7 and ACE competitions defined guidelines for the named entity recognition tasks organized by them in the past[1], [2]. The annotation criteria is not easy to establish. For example, both guidelines differ in the way that the name of a Saint has to be annotated.

ISO space[3] and ISO TimeML standards[4] establish guidelines of space-related features and of temporal relations.[5]

Wilbur et al. (2006) defined annotation guidelines to categorize segments of scientific sentences in research articles of the biomedical domain (see Shatkay et al. (2005)).

There is usually a scarcity of available data for the biomedical domain. The department of Radiology Informatics of Stanford University own a large dataset of radiology reports, that is not annotated, nor publicly available, as far as we know.[6] There are some annotated datasets available for languages different to Spanish in the clinical domain, eg. for English (Uzuner et al., 2011; Pradhan et al., 2013, 2014), for Swedish (Skeppstedt et al., 2014), for French (Névéol et al., 2015), for Polish (Mykowiecka et al., 2009) and for German (Roller et al., 2016). Oronoz et al. (2015) presented an annotated dataset in Spanish for adverse drug reactions analysis. Although the dataset is in

Spanish and addresses the biomedical domain, it concerns a different use case and covers different information. Recently, Cruz et al. (2017) and Marimon et al. (2017) annotated negations in Spanish clinical reports.

3 Annotation Process

Our annotation guideline was improved within three iterations consisting of annotation and revision. In order to decrease the annotation time, entities, negation and uncertainty terms were pre-annotated automatically. Then, based on the annotation guideline, two native speakers of Spanish annotated the pre-annotated reports, making corrections and adding relations.

In this section we describe how we selected and anonymized the reports to be annotated, we present the annotation schema and guidelines, and the automatic and manual annotation process.

3.1 Data

A set of different kinds of ultrasound reports (e.g. kidney, abdominal, small parts, Doppler) provided by a hospital in Argentina were selected. They contain only one section, that includes findings, conclusions and suggestions.

3.1.1 Selection of the Dataset

Since we are interested in examining the existence of different health problems, we performed a selection of the reports to be annotated defining four sets. The first, called **hyperthrophic pyloric stenosis**, involves reports containing information about the pyloric muscle and pyloric canal, that might refer to pyloric stenosis (pyloric obstruction); the second, called **splenomegaly**, contains reports referring to the spleen, whether of normal size or enlarged; the third, called **appendicitis**, has reports that mention the appendix and that might or might not refer to appendicitis and the fourth, called **generic**, comprises a set of ultrasound reports corresponding to different body parts and possibly involving different findings or diseases not including the previous cases.

The first three sets are particularly interesting because the extraction of entities and relations among them could suggest possible medical problems, that might lead to surgical interventions. The first set is useful for studying entities and relation extraction in general terms.

For instance, taking into account the age of the patient and the size of the spleen it is possible to

---

[1] http://itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html

[2] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf

[3] https://www.iso.org/standard/60779.html

[4] https://www.iso.org/standard/37331.html

[5] They have relation with our 'type of measure' entities (longitudinal, transversal, etc) and temporal relations (in the past, etc.). See Subsection 3.2.

[6] http://langlotzlab.stanford.edu/nlp-datasets/

determine whether the patient has splenomegaly (enlargement of the spleen) or not according to normal reference values. Furthermore, the visibility of the appendix, its maximal outer diameter not exceeding 6 mm and its non-compressibility are the most reliable criteria used in the diagnosis of acute appendicitis. The automatic detection of critical issues, such as appendicitis and pyloric stenosis, is of interest and is being studied (see e.g. Do et al. (2013) and Morioka et al. (2016)) and could allow their communication by pager or alternatives methods, as Lakhani and Langlotz (2009) describe.

Table 1 presents the number of files processed by each annotator for each report type. Overall 513 different files have been annotated.

report type	ann. 1	ann. 2	both	total
pyloric sten.	7	51	7	51
splenom.	123	31	13	141
append.	58	45	12	91
generic	176	83	29	230
Total	417	262	61	513

Table 1: Overview of annotated files.

3.1.2 Anonymization

Reports contain a report number, a patient identification number, the date of the study and the age of the patient. In some cases they also have information about the doctor or doctors who performed the ultrasonography and their medical license number. Names might be preceded with the title *Dr.* or *Dra.* (male or female doctor). In occasions, the medical license number is written after the doctor's name, sometimes it can appear without the doctors name. The license number can have been issued at a national (*MN*) or at a state level (*MP*).

Before performing the automatic and manual annotations reports had to be anonymized. Therefore, regular expressions were used considering the different ways of writing the title of the doctors (e.g. *DR*, *Dr.*, *doctor*, *Dra.*), the doctor's names, the enrollment numbers and the order among them. Also names of the doctors appearing with titles or enrollments were searched to see if they appeared without titles and without enrollments and were removed. Patient and report identification were changed in a way that it is not possible to identify a patient. The date of the study was removed.

3.2 Annotation Schema

The following entities and characteristics were considered for the annotation:

findings (FI): entities corresponding to a pathological finding or diagnosis, eg. cyst, gallstone, abscess,

anatomical entities -or body parts- (AE): eg. breast, right thyroid lobe, liver,

location (LO): location in the body or in the body part, eg. medial, distal, peripheral, unilateral, apical, adjacent,

measure (ME): eg. 0.3 mm, 0.5 cc, 2 cm., 0.8 (cm.), large, small, scarce, minimum,

type of measure (TM): indication of the kind of measure that a number is referring to. eg. in *longitudinal 3 (cm) and transversal 1 (cm)*, *longitudinal* and *transversal* will be annotated as two type of measures and *3 (cm)* and *1 (cm)* as two measures, and

texture (TE): e.g. homogeneous or heterogeneous.

Other annotated concepts are the following ones:

negation (NT) and uncertainty terms (UT): We call them modifiers. e.g. were not detected and might correspond to,

abbreviations: e.g. *RI* for *riñon izquierdo* (*right kidney*), and

temporal terms (TT): terms that denote mentions to the past or that express conditionals. eg. previous, residual, old, preoperative, persistent and *had* and *if the patient has* for the phrases *the patient had fever (...)* and *if the patient has fever again (...).*

The following binary relations were annotated:

occurs in: among findings and the part of the body where they occur (AE or LO). eg. in *vescícula biliar de paredes engrosadas*, -*thick(FI)-walled(LO) gallbladder(AE)*-, the finding (*engrosadas*, -*thick*-) occurs in the wall (LO),

located in: between location and an anatomical entity. The goal is to know where in an anatomical entity a finding is located. eg. in the example shown above, the walls (LO) are located in the gallbladder,

area of: associates an anatomical entity with a location. eg. in *kidneys without enlargement of the excretory pathway*, there is an *area of* relation among excretory pathway (LO) and kidney (AE),

has measure type: associates a type of measure with a measure (eg. in *longitudinal 3 (cm), anteroposterior 0.54 (cm)*, the measure *3 (cm)* has measure type longitudinal),

measure of: associates a measure or a type of measure with an anatomical entity, a location or a finding. Also in *pyloric muscle thickness: 3.5 cm.*, there is a relation *measure of* from thickness (TM) to the pyloric muscle (AE) and a relation *has type* from thickness to 3.5 cm. (ME),

texture of: associates an entity of *texture* type to an anatomical entity, a finding or a location. eg. in *[kidneys](AE) of [conserved](TE) echotexture*, the conserved echotexture is related to AE kidneys,

negates: relates a negation term with a finding. eg. in *without enlargement*, the NT *without* is related with the FI *enlargement*,

speculates: relation among an uncertainty term and a finding. eg. in *compatible (UT) with [fatty liver] (FI)* the uncertainty term compatible is related with the fatty liver, and

not present: relates terms referring to the past or conditional terms and a finding. eg. in *gallblader(AE): history(TT) of cholecystectomy (FI)*, the cholecystectomy does not necessarily exist at the present moment and is related as not present with the temporal term.

Entities and relations to annotate were selected based on the named entities and relations that are interesting for physicians.

3.3 Annotation Guidelines

During the annotation process and discussion rounds with the annotators, the original annotation guideline was adjusted. The main annotation guidelines were following:

Largest possible term: as in MUC NER task definition[7] (see Appendix A.1.3), the largest possible term of a particular entity type (that contains as substring terms of the same entity type) has to be annotated (eg [[[retro[peritoneo]] vascular][8] should be annotated as retroperitoneo vascular - *vascular retroperitoneum-*).

Use of lexicons as resources: doubts about the category of an entity (sometimes it is not clear whether a term is, for example, an anatomical entity or its boundaries are not clear) should be

solved using RadLex[9] (it needs a translation into English) or UMLS,[10] (that exists for Spanish)[11]. **Existence of spelling errors:** terms with spelling errors should also be annotated.

Multi-name expressions/terms: unlike MUC-7 NER task definition, when there is elision of the head of one conjunct, the expression should be annotated as different terms (discontinuous expressions can be annotated by our annotation tool, eg. in the construction *intra and extrahepatic*, annotators were asked to annotate *intrahepatic* and *extrahepatic* as entities). The decision to annotate the different terms that form multi-word expressions was taken -after some discussion- because we want the gold standard to be correct and representative of the entities existing in the real world. However, to avoid too complex annotations, cases with more than three terms were annotated as a single *term*.

Relations across sentences: they have to be annotated (eg. in *orthotopic left kidney. Size diminished and (. . .)*, *size* refers to the *orthotopic left kidney* and a relation among them has to be annotated).

Abbreviations: the abbreviations of some entity types (eg. AE or FI) should not only be annotated as abbreviations, but also as entities (eg. in *RD* -riñon derecho, *right kidney*-, *RD* should be annotated as abbreviation and as anatomical entity as well).

Segmentation of annotations: if it is not clear if part of the term corresponds to a LO and part to an AE, if from the term it is clear where the AE is located, the whole term should be labeled as anatomical entity, else a segment might be labeled as LO and another as AE (eg. lymph node should be labeled as an AE, since it is possible to identify the location of lymph nodes in the body. The same occurs with *right iliac fossa*).[12] In *upper part of the head* it is not clear what exactly the upper part is. So *upper part* should be annotated as location and *head* as AE and *the tumor is located in the upper left part of the liver* should be annotated as follows: *the [tumor](FI) is located in*

---

[7]http://itl.nist.gov/iaui/894.02/ related_projects/muc/proceedings/ne_ task.html

[8]Words between brackets show valid anatomical entities.

[9]http://radlex.org/

[10]https://www.nlm.nih.gov/research/ umls/

[11]RadLex is more appropriate for the radiology domain, but has the disadvantage of not being translated into Spanish.

[12]RadLex should be used as a source to detect which is the largest possible concept corresponding to an AE (i.e. if right iliac fossa exists as AE in RadLex then it should be annotated as an AE)
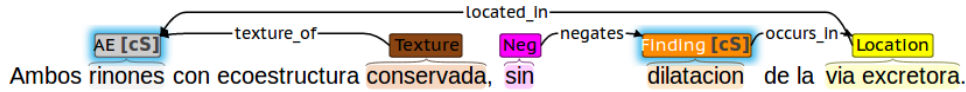
Figure 1: Both kidneys of conserved ecostructure, without enlargement of the excretory pathway.
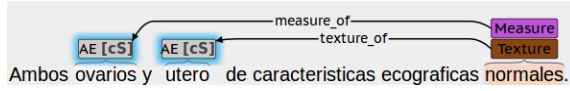

Figure 2: Both ovarys and uterus of normal eco-graphic signs.

the [upper](LO) [left part of the liver](AE).

Prioritize AE over LO: if there is a doubt as whether a term corresponds to an anatomical entity or to a location, it should be annotated as anatomical entity.

Prioritize findings over locations: if a concept refering to a finding includes a location, the largest possible concept that refers to a finding has to be annotated (eg. *pyloric stenosis* refers to a FI, that includes a location. Therefore, [pyloric stenosis](FI) should be chosen over [pyloric](AE) [stenosis](FI)).

Negation and uncertainty terms: negations and uncertainty terms should be annotated only if there is a relation among them and a finding.

Anatomical entities: anatomical entities have to be annotated although there is no relation among them and a finding (eg. in *right lobe of the liver has the usual size*, *right lobe of the liver* should be annotated, although it is not associated to any finding).

Some other decisions that had to be taken were how to annotate certain frequently occurring concepts in the best way. For example, we decided to always annotate *kidney implant* as an AE. Furthermore, *ovarian cyst* and *cyst in the ovary* should be annotated as [ovarian cyst](FI) instead of [ovarian](AE) [cyst](FI), and as cyst[FI] in the ovary[AE].

Figures 1 and 2 show examples of annotations of some sentences.

3.4 Automatic Pre-Annotation

In order to decrease the annotation time, entities were pre-annotated automatically. For this purpose, regular expressions, UMLS and a manually-created dictionary were used.

Regular expressions were used to detect the concept *measure*. *Anatomical entities* and *findings* were detected by the use of some semantic types (STY) of UMLS (see mapping among our concepts and UMLS STYs in Table 2).

entity	UMLS STY
AE	Body Part, Organ, or Organ Component; Body Space or Junction; Body System; Tissue
FI	Anatomical Abnormality; Congenital Abnormality; Acquired Abnormality; Finding; Sign or Symptom; Pathologic Function; Disease or Syndrome; Mental or Behavioral Dysfunction; Neoplastic Process; Injury or Poisoning

Table 2: Mapping of UMLS Semantic Types to our annotation schema.

First, concepts of the Spanish UMLS were mapped to the radiology reports. If one of the corresponding semantic types corresponded to an anatomical entity or to a finding, then the concept was pre-annotated. Finally, a manually-created dictionary, that contains terms and their corresponding entity type -such as negation and uncertainty terms, locations or textures- was used. For example *no puede descartarse* (*cannot be discarded*) is mapped to *Uncertainty*. Many of the concepts in this dictionary have been included within an iterative process followed from the annotations performed by the annotators. A dictionary lookup algorithm was used to look for terms of UMLS and of the terms of the dictionary in the texts. Therefore, terms were stemmed using the Spanish Snowball implementation of NLTK.

After applying automatic pre-annotation, data was processed by human annotators. Annotations wrongly made by the tool were removed or corrected and missing concepts included.

3.5 Manual Annotation

The manual annotation was carried out by two Spanish native speakers: an advanced medicine student (6th year of the career) (Annotator 1) and a professional of a technical field (Annotator 2), that were not trained in medical document annotation. brat annotation tool (Stenetorp et al., 2012) was used for this purpose.

Many meetings were held with the annotators in order to solve doubts. After having annotated a first dataset (Annotation iteration 1 in Table 5) doubts and differences in criteria were reviewed

and the annotation guidelines (described in Subsection 3.3) were written by the computational linguits with more detail. After two annotation-revision iterations, the final guidelines were defined and annotations were performed (in what we call iteration 3). Disagreements were solved by a computational linguist with expertise in the biomedical domain together with a physician.

4 Dataset Analysis

Once the annotation was performed, the annotated dataset was analyzed in order to know how many entities and relations of each type were found. The analysis of the annotations is calculated for the whole set of 513 annotated reports (not for each of the datasets described in section 3.1.1). For those reports annotated by both annotators the annotation done by the medical student was chosen.

type	total	different
anatomical	4398	405
finding	2637	745
location	722	201
measure	3210	975
texture	1890	74
type of measure	1127	72
negation	1489	51
uncertainty	109	26
abbreviations	880	105
temporal expressions	35	15
multi-name terms	788	210

Table 3: Type and amount of entities, modifiers and other characteristics.

Table 3 shows the number of entities, modifiers of entities and other characteristics (abbreviations, temporal expressions and multi-name terms). In all cases the total number of concepts and the number of different concepts is shown. It can be seen that there are a total of 880 abbreviations. 470 of them correspond to anatomical entities and 7 to findings. The rest correspond to type of measures (266), locations (20) and 117 have no associated entity type. Table 4 shows for each type of relation, the entities related by them, the total number of relations and the number of different relations appearing in the annotated texts.

There are 867 relations across sentences and a total of 10987 relations in the 513 reports. 7.89% of the relations are across sentence relations.

The most frequent multi-name terms are *via biliar extrahepática* (*extrahepatic bile duct*) (232) and *via biliar intra hepática* (*intrahepatic bile duct*) (219).

relation	entities	total	different
occurs in	FI-AE	2161	750
	FI-LO	233	218
located in	LO-AE	538	165
area of	AE-LO	65	53
measure of	TM-AE	1007	154
	TM-LO	46	36
	TM-FI	59	56
	ME-AE	1651	578
	ME-LO	74	48
	ME-FI	407	346
has meas. type	ME-TM	1123	831
texture	TE-AE	1495	192
	TE-LO	387	54
	TE-FI	90	37
negates	NG-FI	1478	164
speculates	UT-FI	96	86
not present	CT-FI	33	33

Table 4: Relations with more than five occurrences annotated among entities in iteration 3.

4.1 Inter-Annotator Agreement

To evaluate the consistency among the annotations performed by both annotators, the inter-annotator agreement (IAA) was calculated using the Cohen's Kappa coefficient (κ) (Cohen, 1960). κ was calculated with the scikit-learn toolkit on a token level.

Multiple annotations per token are possible (and are frequently used) due to various meanings and to the existence of nested or overlapping concepts, e.g. *normal* can be labeled as measure, as texture or as both of them (consider, for example, the phrase *normal size and echotexture*). For the calculation of the IAA we decided to consider that a token is labeled in the same way by both annotators if and only if both annotators assigned to it the exact same set of labels (or no label at all).

Table 5 shows the IAA for each of the annotation datasets. It can be appreciated that it improves in each annotation iteration step. Annotation dataset 3 was the final one, and was annotated once the annotation schema and criteria (see Subsections 3.2 and 3.3) were stabilized.

ann. iter.	# reports	# processed by both annotators	κ
1	16	16	0.5883
2	20	20	0.8577
3	513	61	0.8893

Table 5: IAA (κ) and number of annotated reports in different annotation iterations.

The subset of dataset 3 annotated by both annotators contains 427 tokens with more than one annotation of a total of 5894 tokens. That is 7.24% tokens belonged to more than one entity according

to one of the annotators.

5 Discussion

As reported before, abbreviations are often used in our dataset. Considering that there are 7035 anatomical entities and findings (see Table 3), and that there are 477 abbreviations of anatomical entities and findings, we can think that about 6% of the anatomical entities and findings are written in an abbreviated way. Also, there is a total of 105 different abbreviations in 513 reports (see Table 3), which we consider a high number.

The other results of Table 3 are useful for the later use of this annotated dataset. Multi-name terms will probably not be easily recognized by standard entity recognition algorithms. The relation of findings with temporal expressions (see Table 4) should be taken into account to determine the factuality of a finding. The same occurs with terms that denote negation and uncertainty. Relations between sentences will be also difficult to discover.

As depicted in Table 5, the inter-annotator agreement improves in each annotation iteration step. This makes sense, since after each annotation iteration many meetings were held with both annotators to solve doubts and the annotation criteria was changed according to new questions the annotators asked until it stabilized. With this stabilized annotation guidelines, annotators performed the annotation of dataset 3, which had an inter-annotator agreement of 0.89.

We do not have an objective measure related to how the annotation easiness increased after the pre-annotation process. Eventhough, annotators reported that after some improvements of our manually built dictionary (occured after some 'annotation-automatic improvement of the dictionary' iterations) pre-annotations were much more accurate and that their annotation was much easier. We also noticed an increase of the reports annotated per hour.

Considering Tables 3 and 4, we can see that 56% of the findings are negated (1478 out of 2637). This might lead to future implementation of methods to detect negated findings in reports (see Chapman et al. (2001), and Cotik et al. (2015) for Spanish). Only 1.25% of the findings are reported as a past issue or as a conditional issue in the future (33 *not present* relations out of 2637 *findings*).

The development of the annotation criteria has not been an easy task. New entities (eg. location) had to be added to the initial annotation schema. The need to add these entities came from the actual annotation process and the questions that the annotators had. The initial set of relations grew also much more than expected due to the complexity of some of the sentences that revealed the existence of relations that were not considered initially.

In many cases it was not easy to determine if a concept belonged to an entity type or to another. In particular we found that a location can be referring to an anatomical entity and an anatomical entity to a location. Many doubts of this kind arose and helped us to define the definitive annotation guidelines (Section 3.3).

6 Conclusions

In this work we presented an annotation criteria developed for a set of radiology reports written in Spanish with the goal to be able to use the annotated corpus as an evaluation resource for name entity recognition and relation extraction and as input for the training of supervised learning methods to solve these tasks. We divided the total available set in four subsets in order to be able to extract in the future relations among the data that give further information (for instance, the existence of appendicitis), that might be useful for physicians and patients. We anonymized data and trained the annotators to do the annotation task.

The shortness of the texts, the abundance of acronyms and the specificity of the medical language made the annotation task difficult. Furthermore, it was not easy to keep up with the goal to achieve a simple annotation criteria.

The analysis of the annotated dataset shows some interesting characteristics, as the abundance of negated findings. That might lead to the development of negation detection algorithms. This annotated dataset is useful for its evaluation.

We noticed the importance of having annotators with expertise in the annotation task and in the medical domain and we consider that in this particular domain it is even more difficult than in others to obtain annotations from experts.

References

Wendy Chapman, W. Bridewell, P. Hanbury, GF. Cooper, and BG. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform.* 34(5):301–310.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Viviana Cotik, Vanesa Stricker, Jorge Vivaldi, and Horacio Rodriguez. 2015. Syntactic methods for negation detection in radiology reports in Spanish. In *ACL - Workshop on Replicability and Reproducibility in Natural Language Processing: adaptative methods, resources and software*. Buenos Aires, Argentina.

Noa Cruz, Roser Morante, Manuel J. Maña López, Jacinto Mata Vázquez, and Carlos L. Parra Calderón. 2017. Annotating negation in spanish clinical texts. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*. Association for Computational Linguistics, Valencia, Spain, pages 53–58.

B. Do, A.S. Wu, J. Maley, S., and Biswal. 2013. Automatic retrieval of bone fracture knowledge using natural language processing. *J Digit Imaging* 26(4):709–13.

Paras Lakhani and Curtis P. Langlotz. 2009. Automated detection of radiology reports that document non-routine communication of critical or significant results. *J Digit Imaging* 23(6):647–57.

Montserrat Marimon, Jorge Vivaldi, and Núria Bel. 2017. Annotation of negation in the iula spanish clinical record corpus. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*. Association for Computational Linguistics, Valencia, Spain, pages 43–52.

C. Morioka, F. Meng, R. Taira, J. Sayre, P. Zimmerman, D. Ishimitsu, J. Huang, L. Shen, and S. El-Saden. 2016. Automatic classification of ultrasound screening examinations of the abdominal aorta. *J Digit Imaging* 29(6):742–48.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. 2009. Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics* 42(5):923 – 936. Biomedical Natural Language Processing.

Aurélie Névéol, Cyril Grouin, Xavier Tannier, Thierry Hamon, Liadh Kelly, Lorraine Goeuriot, and Pierre Zweigenbaum. 2015. CLEF ehealth evaluation lab 2015 task 1b: Clinical named entity recognition. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*.

Maite Oronoz, Koldo Gojenola, Alicia Pérez, Arantza Díaz de Ilarraza, and Arantza Casillas. 2015. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics* 56:318 – 332.

Sameer Pradhan, Noémie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. 2014. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association* .

Sameer Pradhan, Noémie Elhadad, Brett R. South, David Martínez, Lee M. Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana K. Savova. 2013. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*.

Roland Roller, Feiyu Xu Hans Uszkoreit, Laura Seiffe, Michael Mikhailov, Oliver Staeck, Klemens Budde, Fabian Halleck, and Danilo Schmidt. 2016. A fine-grained corpus annotation schema of German nephrology records. *Proceedings of the Clinical Natural Language Processing Workshop* 28(1):69–77.

M. Sevenster, R. van Ommering, and Y. Qian. 2012. Automatically correlating clinical findings and body locations in radiology reports using MedLEE. *J Digit Imaging* 25(2):240–249.

Hagit Shatkay, W. John Wilbur, and Andrey Rzhetsky. 2005. Annotation guidelines [Online; accessed 28-04-2017].

Maria Skeppstedt, Maria Kvist, Gunnar H. Nilsson, and Hercules Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics* 49:148 – 158.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. Association for Computational Linguistics, Avignon, France.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 18(5):552–556.

W John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics* 356(7).