

# Overview of CLEF eHealth Task 1 - SpRadIE: A challenge on information extraction from Spanish Radiology Reports

Viviana Cotik<sup>1,2</sup>, Laura Alonso Alemany<sup>3</sup>, Darío Filippo<sup>4</sup>, Franco Luque<sup>3,5</sup>, Roland Roller<sup>6</sup>, Jorge Vivaldi<sup>7</sup>, Ammer Ayach<sup>6</sup>, Fernando Carranza<sup>8</sup>, Lucas Defrancesca<sup>3</sup>, Antonella Dellanzo<sup>1</sup> and Macarena Fernández Urquiza<sup>9</sup>

<sup>1</sup>*Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina*

<sup>2</sup>*Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina*

<sup>3</sup>*Universidad Nacional de Córdoba, Argentina*

<sup>4</sup>*Hospital de Pediatría 'Prof. Dr. Juan P. Garrahan', Argentina*

<sup>5</sup>*CONICET, Argentina*

<sup>6</sup>*German Research Center for Artificial Intelligence (DFKI), Germany*

<sup>7</sup>*Institut de Lingüística Aplicada, Universitat Pompeu Fabra, Spain*

<sup>8</sup>*Instituto de Filología y Literaturas Hispánicas "Dr. Amado Alonso", Universidad de Buenos Aires, CONICET, Argentina*

<sup>9</sup>*FFyL, Universidad de Buenos Aires, Argentina*

## Abstract

This paper provides an overview of SpRadIE, the Multilingual Information Extraction Task of CLEF eHealth 2021 evaluation lab. The challenge targets information extraction from Spanish radiology reports, and aims at providing a standard evaluation framework to contribute to the advancement in the field of clinical natural language processing in Spanish.

Overall seven different teams participated, trying to detect seven named entities and hedge cues. Information extraction from radiology reports has particular challenges, such as domain specific language, telegraphic style, abundance of non-standard abbreviations and a large number of discontinuous, as well as overlapping entities. Participants addressed these challenges using a variety of different classifiers and introduced multiple solutions. The most successful approaches rely on multiple neural classifiers in order to deal with overlapping entities.

As a result of the challenge, a manually annotated dataset of radiology reports in Spanish has been made available. To our knowledge this is the first public challenge for named entity recognition and hedge cue detection for radiology reports in Spanish.

## Keywords

Spanish Information Extraction, BioNLP, eHealth, radiology reports

---

*CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania*

✉ [vcotik@dc.uba.ar](mailto:vcotik@dc.uba.ar) (V. Cotik); [lauraalonsoalemany@unc.edu.ar](mailto:lauraalonsoalemany@unc.edu.ar) (L. Alonso Alemany); [dfilippo@gmail.com](mailto:dfilippo@gmail.com) (D. Filippo); [francolq@unc.edu.ar](mailto:francolq@unc.edu.ar) (F. Luque); [roland.roller@dfki.de](mailto:roland.roller@dfki.de) (R. Roller); [jorge.vivaldi@upf.edu](mailto:jorge.vivaldi@upf.edu) (J. Vivaldi)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

# 1. Introduction and Motivation

In the last years, the volume of digitally available information of the medical domain has been in constant growth. This is due especially to the widespread adoption of clinical information systems and electronic health records (EHRs). Consequently, this is leading to the progressive adoption of natural language processing applications in healthcare because of its recognized potential to search, analyze and interpret patient datasets. Physicians spend a lot of time inputting patients data into EHR systems, most of it stored as narratives of free text. The extraction of information contained in these texts is useful for many purposes from which some of the most relevant are: diagnostic surveillance through automated detection of critical observations; query based case retrieval; quality assessment of radiologic practice; automatic content analysis of report databases; and clinical support services integrated in the clinical workflow [1].

There are many types of medical reports within an electronic health record, such as chart notes, case notes, progress notes, radiology reports and discharge reports. Some of them are written in highly specialized and local vocabulary and in the special case of radiology reports, they may have non-standard abbreviations, typos and ill-formed sentences. Because of the particularities of the medical domain, clinical corpora are difficult to obtain. Clinical records are of sensitive nature, so they are usually not published, and, if done so, they have to be anonymized. Moreover, the highly specialized and local vocabulary makes the annotation a difficult and expensive task.

Most of the currently available resources on clinical report processing are for English. For Spanish, the availability of resources is much more limited, despite being one of the languages with more native speakers in the world. In particular, there are very few available annotated corpora (see Section 2).

In this context, we publish a novel corpus through the organization of the SpRadIE challenge, a task proposed in the context of the CLEF eHealth 2021 evaluation lab [2]<sup>1</sup>. This corpus is a reviewed version of a previously annotated and anonymized corpus of Spanish radiology reports [3, 4]. With SpRadIE we intend to collaborate to the advancement in the automatic processing of medical texts in Spanish, while offering participants the opportunity to submit novel systems and compare their results using the same dataset and a standard evaluation framework. To our knowledge, SpRadIE is the first information extraction challenge on Spanish radiology reports.

More concretely, the SpRadIE challenge dataset consists of a corpus of pediatric ultrasound reports from an Argentinian public hospital. These reports are generally written within a hospital information system by direct typing into a computer a single section of plain text, where the most relevant findings are described. The reports are written using standard boilerplate that guide physicians on the structuring when there are no anomalous findings. However, most of the times they are written in free text to be able to describe the findings discovered in anomalous studies. The fact that input is free text and that anomalies are often found and reported results in great variations in the content of the reports and in their size, ranging from 8 to 193 words.

SpRadIE offers multiple challenges that need to be addressed with creative solutions:

**Low resources:** The availability of linguistic resources in Spanish is greatly limited in com-

---

<sup>1</sup><https://sites.google.com/view/spradie-2020/>

parison to high-resource languages such as English. In particular, there are no specific terminologies for the radiology domain in Spanish.

**Domain-specific language:** The vocabulary used in radiology reports is specific to the radiology domain. The Radiological Society of North America (RSNA) produced Radlex (Radiology Lexicon)<sup>2</sup> an extensive, dedicated and comprehensive set of radiology terms in English, for use in radiology reporting, decision support, data mining, data registries, education and research. Besides, SNOMED CT (Systematized Nomenclature of Medicine, Clinical Terms)<sup>3</sup> is considered to be the most comprehensive, multilingual clinical health-care terminology (it includes Spanish) for medicine and has mappings to Radlex.

**Ill-formed texts:** Reports usually present ill-formed sentences, misspellings, inconsistencies in the usage of abbreviations, and lack of punctuation and line breaks, as can be seen in Figure 3.

**Semantic Split:** Training, development and test sets cover different semantic fields, so that various topics and their corresponding entities that occur in the test dataset have not been previously seen in the training dataset.

**Small data:** To approach realistic deployment conditions, only a small amount of annotated reports has been available during training, and the rest has been used for evaluation.

**Complex entities:** The linguistic form of entities presents some particular difficulties: lengthier entities with inner structure, embedded entities and discontinuities. Examples can be found in Section 3.

In the past, some challenges have been organized for information extraction in the medical domain in Spanish (see Section 2). However, our proposal covers a part of the domain spectrum that was not covered by previous work: actual short reports written in haste with mistakes and variability.

In this article we present the SpRadIE challenge and its results. The challenge aims at the detection of seven different named entities as well as hedge cues from ultrasounds reports. Targeted entities include anatomical entities and findings that describe a pathological or anomalous event. Also negations and indicators of probability or future outcomes are to be detected. We provide training, development and test datasets, and evaluate the participating systems using metrics based on lenient and exact match.

Overall seven different teams participated in the task, with participants from Spain, Italy, United Kingdom, and Colombia. In more than 70% of them, there is at least one Spanish native speaker. Most teams experimented with different variations of neural networks, particularly BERT-based approaches. However, there were also submissions based on Conditional Random Fields (CRFs) and pattern rules. The presence of overlapping and discontinuous entities was one of the biggest challenges of the task. In order to overcome this problem, several teams developed multiple classifiers running in parallel, together with pre and post-processing steps for the input/output of the classifiers.

---

<sup>2</sup>RadLex <http://radlex.org/>

<sup>3</sup>SNOMED CT <https://www.snomed.org/>

The remainder of the paper is structured as follows. In the following Section we present previous work for Spanish medical information extraction and Spanish corpora in the medical domain. In Section 3 we describe the target of the annotation, detailing types of entities, their distribution in the annotated dataset, some of their most prominent features and the difficulty of the task as measured by human inter-annotator agreement. Then, Section 4 presents the Evaluation setting. Participating systems and baselines are described in Section 5, while results are discussed in Section 6. We finish with some conclusions and a hope for the advancement in the automatic treatment of medical text in Spanish.

## 2. Previous work

In clinical care, many important patient related information is stored in textual format, supplementing the structured information of electronic health records. To automatically access and make use of this valuable information, methods of natural language processing (NLP), like named entity recognition, relation extraction and negation detection, can be applied. In order to train such methods, domain-related corpora have to be available. Medicine has many sub-domains, such as radiology. The availability of specific corpora for handling them is of utmost importance for the advancement of the BioNLP area.

Given the sensitive nature of medical data and the difficulty of its annotation process, only very few corpora are available, and most of them are in English (e.g., CheXpert [5] and MIMIC-CXR [6]). Moreover, most existing tools to process clinical text are also developed for English. Nevertheless, in recent years the interest and need for processing non-English clinical text has been increasing. In particular for Spanish, one of the languages with more native speakers in the world.

Annotation of medical texts is a difficult task, particularly for clinical records. Wilbur et al. [7] defined annotation guidelines to categorize segments of scientific sentences in research articles of the biomedical domain. The first published guideline for the annotation of radiology reports that we are aware of [3] has been reviewed and enhanced for the annotation of the dataset provided in this challenge. Besides, there are some corpora of negations in Spanish clinical reports [8, 9]. Finally, recently, PadChest, a corpora of 27,593 Spanish annotated radiology reports has been published [10].

In the past, several challenges have been organized for information extraction in the medical domain in Spanish. The CodiEsp shared task in CLEF-2020 [11] addressed clinical cases (longer sentences and paragraphs, more consistent use of vocabulary and less typos than radiology reports). The target of CodiEsp was to assign tags at a document level, and to identify text spans that support the assignation of the tags. The eHealth-KD challenge<sup>4</sup> and the CANTEMIST shared task,<sup>5</sup> both part of IberLEF-SEPLN 2020, targeted the identification of named entities and relations (at different levels of granularity in the types of entities) but in medical research papers instead of clinical reports. Other challenges that targeted information extraction from Spanish biomedical texts were PharmaCoNER [12] (detection of drug and chemical entities) MEDDOCAN [13] (anonymization), TASS eHealth-KD 2018 Task 3 [14] and IberLEF eHealth-KD

---

<sup>4</sup>eHealth-KD: <https://knowledge-learning.github.io/ehealthkd-2020/>

<sup>5</sup>CANTEMIST: <https://temu.bsc.es/cantemist/>

2019 and 2020 [15, 16].

Spanish negation detection in the biomedical domain is also a current subject of interest (see NEGES 2018 Workshop on Negation in Spanish)<sup>6</sup>, that has some works for the medical domain and [17, 18, 19, 20, 21].

Besides the approaches used in previously mentioned challenges, not much work has been done for NER in Spanish clinical reports so far. Focusing in the radiology domain, only a few publications target NER in the context of Spanish radiology reports [22, 23]. General overviews about NLP in radiology can be seen in [1, 24].

### 3. Target of the challenge

The target of the task is Named Entity Recognition and Classification. As mentioned above, these entities present several challenges. We describe the types of entities we are targeting and then exemplify some of the challenges.

#### 3.1. Classes of entities

Seven different classes of concepts in the radiology domain are distinguished. Since these entities refer to very precise, complex concepts, they are realized by correspondingly complex textual forms. Entities may be very long, sometimes even spanning over sentence boundaries, embedded within other entities of different types and may be discontinuous. Moreover, different text strings may be used to refer to the same entity, including abbreviations and typos.

Entities are formed by a word or a sequence of words, not necessarily continuous, and entities can be embedded within other entities. The following entities are distinguished: Anatomical Entity, Finding, Location, Measure, Type of Measure, Degree, and Abbreviation. Hedge cues are also identified, distinguishing: Negation, Uncertainty, and Conditional-Temporal. Examples can be seen in Figure 1.

As mentioned before, these entities present several challenges. Examples of longer, discontinuous and overlapping entities can be found in Figure 2.

#### 3.2. Annotated dataset

The data consists of 513 ultrasonography reports provided by a public pediatric hospital in Argentina. Reports are semi-structured and have orthographic and grammatical errors. They have been anonymized in order to remove patient IDs, names and the enrollment numbers of the physicians [3]. An example of a report can be seen in Figure 3. The annotated training and development partitions of the dataset, and the unannotated test partition are available at the webpage of the SpRadIE challenge <https://sites.google.com/view/spradie-2020/>.

Reports were annotated by clinical experts and then revised by linguists, using the brat annotation tool [25]. Annotation guidelines and training were provided for both rounds of annotations (see [3] for the first round). An example of an annotated excerpt can be seen in Figure 4.

---

<sup>6</sup>NEGES 2018: <https://aclweb.org/portal/content/neges-2018-workshop-negation-spanish>

**Anatomical Entity** Entities corresponding to an anatomical part, for example "breast" (*pecho*), "liver" (*hígado*), "right thyroid lobe" (*lóbulo tiroideo derecho*).

Anatomical Entity  
 vejiga llena  
*full bladder*

**Finding** A pathological finding or diagnosis, for example: "cyst", "cyanosis".

Finding  
 No se detectaron adenomegalias  
*No adenomegalias were detected*

**Location** It refers to a location in the body. The location could by itself indicate of which part of the body it is being talked about or it could have a relation to an anatomical entity. Examples of locations are: "walls", "cavity".

Location  
 quistes en región biliar  
*cysts in biliary region*

**Measure** Expression of measure.

Measure  
 Diametro longitudinal: 8.1 cm.  
*Longitudinal diameter: 8.1 cm.*

**Type of measure** Expression indicating a kind of measure.

Type of Measure  
 Diametro longitudinal: 8.1 cm.  
*Longitudinal diameter: 8.1 cm.*

**Degree** It indicates the degree of a finding or some other property of an entity, for example, "leve", "levemente" (*slight*), "mínimo" (*minimal*).

Degree  
 ligera esplenomegalia  
*slight splenomegaly*

Three subtypes of hedge cues are identified:

**Negation**

Negation  
 No se detectaron adenomegalias  
*No adenomegalias were detected*

**Conditional - Temporal** Hedge cues indicating that something occurred in the past or may occur in the future. Also indicating a conditional form.

Conditional-Temporal  
 antecedentes de atresia  
*history of atresia*

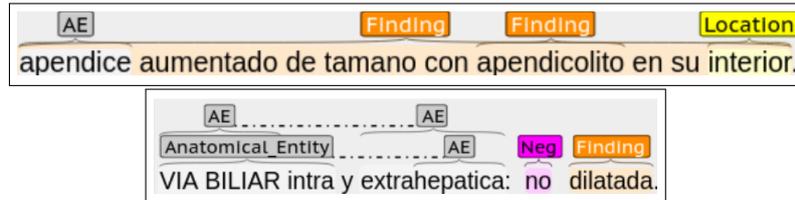
**Uncertainty** Hedge cues indicating a probability (not a certainty) that some finding may be present in a given patient.

Uncertainty  
 compatible con hipertrofia pilórica  
*compatible with pyloric hypertrophy*

Figure 1: Classes of entities distinguished in radiological reports.

### 3.2.1. Distribution of entities

The distribution of entities is shown in Figure 5. The most frequent type, *Anatomical Entity*, has more than 2,000 occurrences, and there are almost 1,500 *Findings*, but there are only 163 hedges, and only 15 *Conditional-Temporal* hedges. It can be expected that performance of automatic systems is poor in types of entities with such few examples.



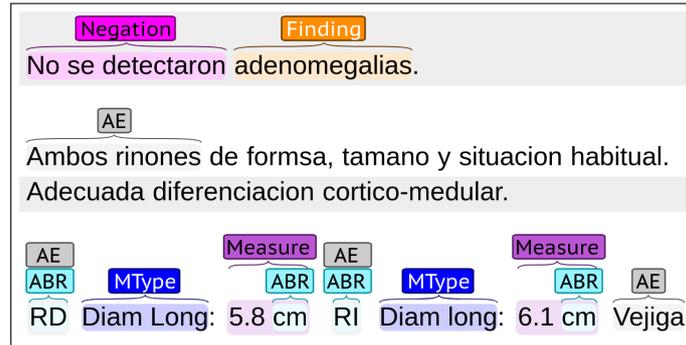
**Figure 2:** Examples of longer, discontinuous and overlapping entities.

2a.  
 HIGADO de forma, tamaño y ecoestructura normal.  
 VIA BILIAR intra y extrahepatica: no dilatada.  
 VESICULA BILIAR: de paredes finas sin imágenes endoluminales.  
 BAZO: tamaño y ecoestructura normal.  
 Diametro longitudinal: 6.89 ( cm ) RETROPERITONEO VASCULAR: sin alteraciones.  
 No se detectaron adenomegalias.  
 Ambos rinones de forma, tamaño y situación habitual.  
 Adecuada diferenciación cortico-medular.  
 RD Diam Long: 5.8 cm RI Diam long: 6.1 cm Vejiga de características normales.  
 No se observó líquido libre en cavidad abdomino-pelviana.

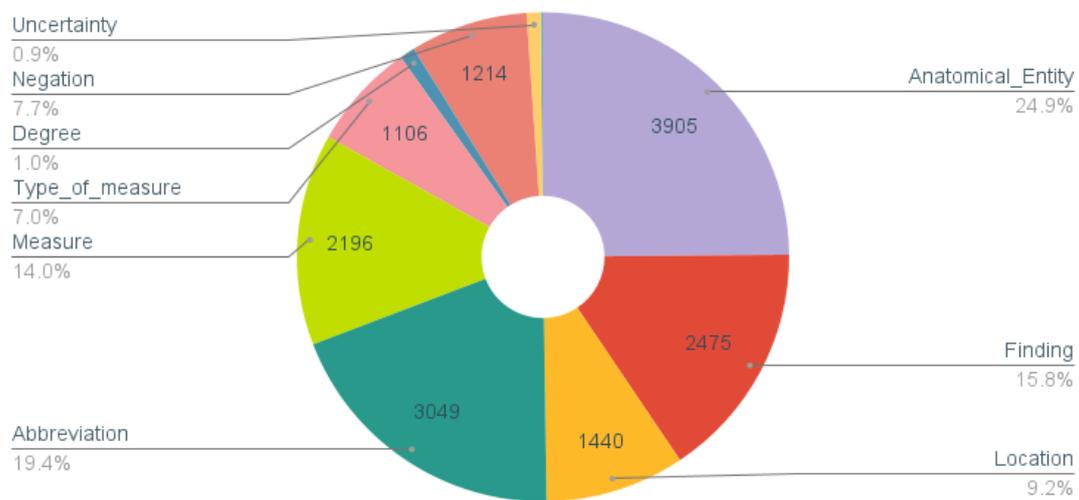
2y.  
*LIVER of regular form, size and echostructure.*  
*Intra and extrahepatic BILE DUCT: non-dilated.*  
*GALLBLADDER: thin walls and no endoluminal images.*  
*SPLEEN: regular size and echostructure.*  
*Longitudinal diameter: 6.89 ( cm ) VASCULAR RETROPERITONEAL: no alterations.*  
*No adenomegalies were found.*  
*Both kidneys of regular form, size and location.*  
*Adequate corticomedullary differentiation.*  
*RK Long diam: 5.8 cm LK Long diam: 6.1 cm Bladder of regular characteristics.*  
*No free liquid was observed within the abdomino-pelvic cavity.*

**Figure 3:** A sample report, with its translation to English. It shows abbreviations (“RD” for right kidney, “RI” for left kidney, “Diam” for diameter), typos (“formsa” for “forma”), and inconsistencies (capitalization of “Vejiga” because of start of sentence without a full stop.)

Moreover, the different types of entities differ a lot among themselves. While entities of the *Finding* type have an average length of 2.35 words and *Anatomical Entities* are in average 1.9 words long, which is not a big difference. However, we can see a big difference in the number of times words are repeated within each type of entity. In Figure 6 we can see that most of the words in *Type of Measure* and *Negation* occur at least 10 times, as is shown by the long box, meaning that the majority of words occur up to 10 times or even more. With quite a tall box, we can see that words in *Anatomical Entities* also tend to occur a high number of times. In contrast, most of the words in *Findings* or *Locations* occur less than 2 or 3 times. If entities are more repetitive in their wording, it is easier for an automatic classifier to identify them.



**Figure 4:** A snippet of the report in Figure 1, with manual annotations. Abbreviations: AE – Anatomical Entity, ABR – Abbreviation, MType – Type of Measure.



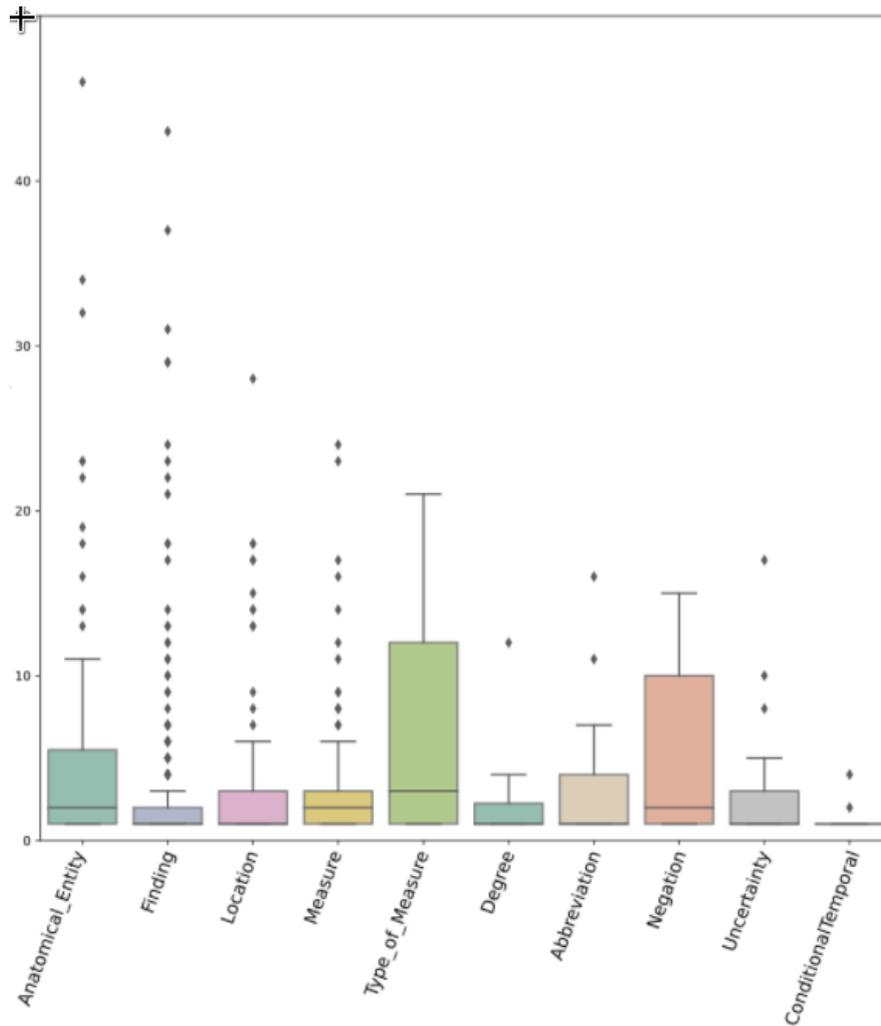
**Figure 5:** Distribution of entities by type in the annotated dataset.

### 3.2.2. Inter-annotator Agreement

Automatic classifiers will be expected to perform well in those cases where human annotators have strong agreement, and worse in cases that are difficult for human annotators to identify consistently.

We carried out a small study of inter-annotator agreement to assess the difficulty of the task for trained human experts. Three trained linguists independently annotated 20 reports (totalling 2,000 words and over 1,700 annotated entities) after reading the annotation guidelines and sharing the annotations for two reports. The mean inter-annotator agreement was  $\kappa = .85$ .

In Figure 7 it can be seen that, among the frequent entities, *Location* is the one with lowest agreement and variation in agreement. *Degree* also has low agreement, and the *Uncertainty* hedge. No figures for *Conditional-Temporal* were obtained because there were few cases in the



**Figure 6:** Number of times that words are repeated within each type of entity.

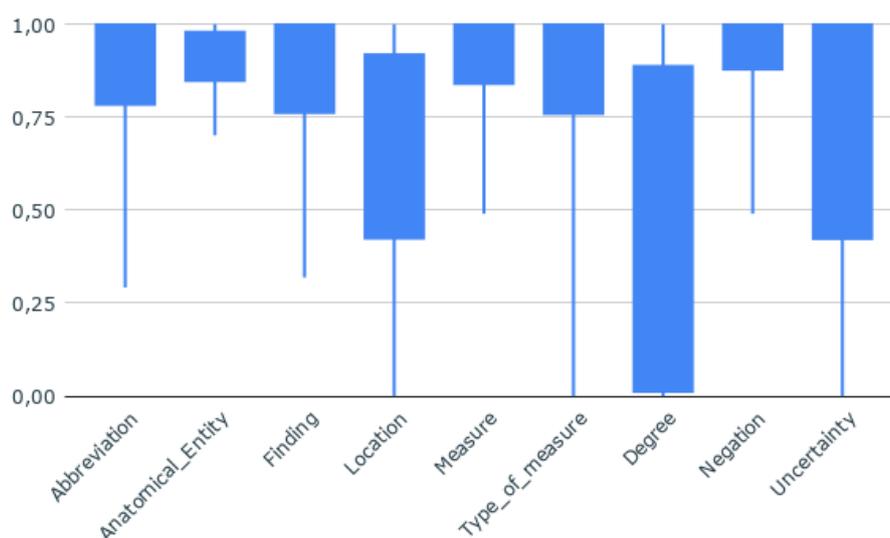
dataset.

Thus, it can be expected that automatic classifiers perform worse in these categories than in other that are more easily identified by humans, like *Abbreviation*, *Anatomical Entity of Finding*.

## 4. Evaluation setting

### 4.1. Dataset Partitions

Since reports are highly repetitive, almost half of the annotated corpus (207 reports) was used as test set for evaluation. The test set was created by identifying terms belonging to a given semantic field within the reports, and selecting all reports containing those terms. Thus, the test set was guaranteed to contain words not in the training corpus, which was useful to assess



**Figure 7:** Variations in inter-annotator agreement across different types of entities, using Cohen's kappa coefficient. In the graphics, boxes range from the mean minus standard deviation to the mean plus standard deviation, whiskers range from minimum to maximum.

**Table 1**

Words occurring only in test and development partitions.

|             |  |
|-------------|--|
| test        | ovario utero endometrio premicc postmicc pielocalicial ureter pielica<br>hepatopeto vci hepatofug suprahepatic peristaltismo ganglio tiroid<br>maxil lobul parotid traquea glandula cervical |
| development | hipertension epiplon portal cardiac aorta corona   |

portability to (slightly) different domains. An additional development partition (45 reports) was created with reports containing terms not in the training set or in the test set. The words occurring only in development and test partitions can be seen in Table 1.

The remaining part of the dataset consisted of a training partition (175 reports), a development partition (47 reports) and a test partition (45 reports).

## 4.2. Metrics

Submissions were evaluated using precision, recall and F1 scores, using both an exact and a lenient matching scheme. Metrics were computed separately for each entity type. Therefore, no credit is given for predictions with correct span but incorrect type. Global results were obtained by micro-averaging. This way, the influence of each entity type in the global results is proportional to its frequency in the test corpus.

We believe that small variations in the span of named entities do not severely affect the quality of the results. Minor differences in the spans still provide useful information for possible applications. Lenient matching can be used to give credit to partial matches in named entities.

For this challenge, we used as the main metric a micro-averaged F1 based on lenient match. We also computed scores based on exact match as secondary metrics.

The lenient scores are calculated using the Jaccard index and based on the metrics used in the Bacteria Biotope task of BioNLP 2013 [26]. The Jaccard index is used as a similarity measure between a reference and a predicted entity. It is defined as the ratio of intersection over union as follows:

$$J(ref, pred) = \frac{overlap(ref, pred)}{length(ref) + length(pred) - overlap(ref, pred)}$$

where  $ref$  represents the reference string of the gold standard and  $pred$  the corresponding string which was predicted. Both, overlap and length are measured in characters. An exact match has a value of 1.

To compute the lenient metrics, we first match reference and predicted entities pairwise. The Jaccard index is used as the point-wise similarity to be optimized in the matching process. To guarantee a global optimal matching in the general case, a bipartite graph matching algorithm would be required. Instead, for simplicity, we implemented a greedy matching algorithm that iterates over the ordered predicted entities and chooses the best matching reference entity. This approach was tested using hand-crafted test cases specifically designed for complex situations, and it always gave the expected matchings.

The matching process returns a set  $M$  of matching pairs of reference and predicted entities. Then, lenient precision and recall are computed as follows:

$$PREC_{lenient} = \frac{\sum_{(ref, pred) \in M} J(ref, pred)}{P}$$

$$REC_{lenient} = \frac{\sum_{(ref, pred) \in M} J(ref, pred)}{R}$$

where  $P$  and  $R$  are the total number of predicted and reference entities respectively.

Exact precision and recall are computed using only exact matches, this is, those matches in  $M$  with a similarity value of 1:

$$PREC_{exact} = \frac{|\{(ref, pred) \in M : J(ref, pred) = 1\}|}{P}$$

$$REC_{exact} = \frac{|\{(ref, pred) \in M : J(ref, pred) = 1\}|}{R}$$

Our official scripts to compute the metrics were offered to the participants before evaluation and are published in a public repository.<sup>7</sup>

## 5. Participating systems

Overall seven different teams participated in the shared task, with participants belonging to institutions from Spain (4), Italy (2), UK (1) and Colombia (1). Most participating teams were

<sup>7</sup><https://github.com/franco1q/spradie>.

experimenting with different variations of neural networks, particularly transformer-based approaches.

Team EdIE [27] (University of Edinburgh and Health Data Research, UK) and SINAI [28] (Universidad de Jaén, Spain) rely on a pre-trained BERT model for Spanish, namely BETO [29]. EdIE uses an ensemble method, combining multiple BERT classifiers, with a dictionary, while SINAI uses a single multiclass model for all entities. SWAP [30] (Università di Bari Aldo Moro, Italy) instead relies on XLM-RoBERTa [31], and CTB [32] (Universidad Politécnica de Madrid, Spain and Universidad del Valle, Colombia) on a multilingual version of BERT.

As an alternative to transformer-based models, team LSI [33] (Universidad Nacional de Educación a Distancia and Escuela Nacional de Sanidad, Spain) uses a neural architecture with a Bi-LSTM followed by a CRF layer.

Aside from neural approaches, a classical CRF approach was used by team HULAT [34] (Universidad Carlos III de Madrid, Spain), and team IMS [35] (Università di Padova, Italy) applied a pattern based approach. Moreover, most teams also explored the usage of different techniques, and different models. Each team was allowed to submit up to four different runs.

Various teams opted for combinations of specialized classifiers instead of a single multiclass model. This is the case of EdIE, combining multiple BETOs. CTB trained separate instances of the same model to predict up to three overlapping entity types on the same token. The CRF layer of the architecture implemented by the LSI team was actually a combination of parallel CRFs specialized for different entity types. For negation, they used a separate model based on transfer learning.

Most teams put much effort into pre- and particularly in post-processing, making most of the differences within the four submissions for a team. Others submitted different architectures or parameterizations of their neural architectures. This is the case for the SWAP team, which experimented with architectures that are partially specialized for clinical text, partially optimized for Spanish, and also multilingual approaches.

More detail on the particulars of each system can be seen in the individual papers in the proceedings of the challenge.

## 5.1. Baselines

Two baselines were constructed: an instance-based learner based on string matching and an off-the-shelf neural learner.

As previously mentioned, the annotated entities in SpRadIE dataset are very repetitive. For this reason, the first simple baseline is an instance-based learner that relies on a simple string matching approach. For each entity in the training set, we extract the different annotated strings with a minimum length of two characters. Whenever a match is found in the test set, it is classified just as it had been seen during training.

A second baseline system uses the Flair framework [36]. Very limited effort was put into pre- and post-processing. Only spans of text tagged with overlapping entities were simplified to a single entity, the most frequent one. The data was fed to a neural sequence tagger with 256 hidden layers and 0.3 locked dropout probability, including a CRF decoder appended at the end. The model also utilizes a stack of Spanish fastText embeddings as well as contextual string embeddings [37]. The model trained for a maximum of 20 epochs with a mini-batch size of 40,

**Table 2**

Overall results for the best performing system for each team on the SpRadIE task, sorted by lenient micro-averaged F1.

| Team                            | lenient      |              |              | exact |       |       |
|---------------------------------|--------------|--------------|--------------|-------|-------|-------|
|                                 | PREC         | REC          | F1           | PREC  | REC   | F1    |
| EdIE (UK) – run2                | 87.24        | <b>83.85</b> | <b>85.51</b> | 81.88 | 78.70 | 80.26 |
| LSI (Spain) – run1              | <b>90.28</b> | 78.33        | 83.88        | 86.17 | 74.76 | 80.07 |
| CTB (Spain, Colombia) – run3    | 78.62        | 78.32        | 78.47        | 73.27 | 72.99 | 73.13 |
| HULAT (Spain) – run1            | 78.38        | 73.08        | 75.64        | 67.28 | 62.73 | 64.92 |
| SINAI (Spain) – run2            | 86.07        | 64.43        | 73.70        | 79.37 | 59.42 | 67.96 |
| SWAP (Italy) – run1             | 70.18        | 51.14        | 59.17        | 56.75 | 41.35 | 47.84 |
| IMS (Italy) – run1              | 9.29         | 57.62        | 16.00        | 5.45  | 33.77 | 9.38  |
| <b>String Matching Baseline</b> | 38.61        | 50.66        | 43.82        | 27.98 | 36.71 | 31.76 |
| <b>Flair Baseline</b>           | 80.72        | 55.34        | 65.66        | 48.60 | 33.32 | 39.53 |

resulting in a 1,5 GB NER model.

## 6. Analysis of Results

In this section we present the results obtained by the seven teams in the task, together with the baselines. Each team could submit up to four runs, however we just report the results of the best performing run of each team in Table 2. Details for the rest of the runs can be seen in the individual papers for each learner. Best scores for lenient precision, recall and F1 are highlighted in bold. Table 3 shows the detail of performance for the 5 most frequent entity types, which cover more than 80% of all entity mentions.

**Baselines** The string matching baseline provides a reference for a very naïve approach to the task, without any kind of generalization. The Flair baseline is a reference of the performance that can be obtained with a more sophisticated learning architecture but without putting effort into optimization, pre- or post-processing. This second baseline already shows that for our problem machine learning quickly outperforms the simple string match approach. While resulting in a similar recall, the precision of the machine learning approach clearly improves.

**Participating Teams** Overall, **EdIE** achieved the best results in the challenge, for both lenient and exact F1 score and recall. EdIE achieves the best results for all five most frequent entity types. However, its performance is worse for less frequent concepts, such as *Degree* (54%), *Uncertainty* (34%) or *ContionalTemporal* (0%).

The overall outcome of **LSI** is very close to EdIE in terms of F1, particularly for exact match. In contrast to EdIE, LSI achieves a higher precision. Overall results are solid across entity types. Similarly to EdIE, LSI has problems dealing with *ContionalTemporal*, with a performance of 0%.

**Table 3**

Detail of scores obtained by each team for lenient precision, recall and F1 across the 5 most frequent types of entities, covering more than 80% of all entity mentions. Best scores for entity type in bold, best F1 also shadowed in grey.

|       | Finding    |            |            | Anatomical Entity |            |            | Location   |            |            | Abbreviation |            |            | Measure    |            |            |
|-------|------------|------------|------------|-------------------|------------|------------|------------|------------|------------|--------------|------------|------------|------------|------------|------------|
|       | Prec       | Rec        | F1         | Prec              | Rec        | F1         | Prec       | Rec        | F1         | Prec         | Rec        | F1         | Prec       | Rec        | F1         |
| EdIE  | 73%        | <b>76%</b> | <b>74%</b> | 88%               | <b>84%</b> | <b>86%</b> | 76%        | 63%        | <b>69%</b> | <b>96%</b>   | <b>95%</b> | <b>95%</b> | 90%        | <b>86%</b> | <b>88%</b> |
| LSI   | <b>82%</b> | 66%        | 73%        | <b>89%</b>        | 78%        | 83%        | <b>84%</b> | 55%        | 67%        | 94%          | 88%        | 91%        | <b>92%</b> | 85%        | <b>88%</b> |
| CTB   | 65%        | 74%        | 69%        | 84%               | <b>84%</b> | 84%        | 68%        | <b>64%</b> | 66%        | 85%          | 81%        | 83%        | 74%        | 76%        | 75%        |
| HULAT | 71%        | 67%        | 69%        | 88%               | 71%        | 79%        | 76%        | 62%        | 68%        | 91%          | 77%        | 84%        | 54%        | 73%        | 62%        |
| SINAI | 74%        | 74%        | <b>74%</b> | 91%               | 79%        | 84%        | 81%        | 59%        | <b>69%</b> | 87%          | 15%        | 26%        | 84%        | 81%        | 82%        |
| SWAP  | 64%        | 52%        | 57%        | 78%               | 56%        | 65%        | 36%        | 46%        | 41%        | 78%          | 40%        | 53%        | 81%        | 55%        | 66%        |
| IMS   | 60%        | 58%        | 59%        | 25%               | 69%        | 37%        | 25%        | 61%        | 35%        | 4%           | 55%        | 7%         | 24%        | 49%        | 32%        |

In contrast, **CTB**, the third ranked system, performs particularly well for *ContionalTemporal* (67%), as well as for *Anatomical Entities* and *Location*. For the rest of concepts, results tend to be about 5-10 points below the best system.

**HULAT**, opposed to the other three systems, uses a CRF instead of a neural architecture. It achieves very similar results regarding lenient F1 in comparison to CTB, but has a drop in exact F1. Overall the system performs quite well regarding *Location*, *Finding*, *Negation*, *Uncertainty* and *ContionalTemporal*. In case of *Measure*, the system has a strong drop of performance in comparison to the best system. More focus on this concept, would have certainly boosted the performance further.

The **SINAI** team, while fifth in overall performance, achieved highest scoring results for two of the most challenging entity types, namely, *Finding* and *Location*. Location was one of the concepts where human annotators showed less consistency. Conversely, the system has got a strong drop in performance regarding *Abbreviations* - about 70 points in comparison to the lenient F1 of the best system. This might have strongly influenced the overall performance of the system, as abbreviations occurs very frequently.

**SWAP** achieved mostly fair performance for all concepts. It performs well above the string matching baseline and better than the Flair baseline for exact match. **IMS** provides a simple pattern based approach, similar to our string matching baseline. It shows how such a simple approach can easily obtain a lenient recall around 60%, which may be useful for applications like information retrieval.

**Performance across entity types** It can be seen that the performance across entity types has some correlation with repetitiveness of strings within entities of a given type and with the consistency of human annotators.

Indeed, entities where annotators were less consistent, mainly *Location* (see Figure 7), overall performance was lower, with a drop of more than 10% with respect to overall performance in most cases, and 15% in the best performing systems. We believe this may be due to these entities having a less defined reference than others, like *Anatomical Entities*.

The other major entity type with lower performance is *Finding*. In this case, we believe

less defined semantics may be a cause for difficulty, but also the form of these entities itself; as described in Figure 6, words in *Findings* occur much less frequently than in other kinds of entities.

**Discussion** Participating systems can be differentiated in three coarse groups with respect to performance. The first group, consisting of the first two teams, both provide quite similar results and have got some distance to the group in the middle field, consisting of the next three teams. The remaining two teams show lower performance. While IMS describes an easy and quick system to start with, and achieves therefore baseline performance, SWAP might have chosen an inadequate architecture for the task. The authors finally submitted a run with XLM-RoBERTa, although other systems performed better during development of their system. However, while multilingual BERT performed best, BETO might have been a better option, as it is solely trained on Spanish language data.

With respect to best performing systems, what seems to have the biggest impact in performance is an architecture based on multiple classifiers, instead of a single multiclass model. The only exception to this would be CTB, scoring third with a multiclass model. This seems to be related to the phenomenon of overlapping entities, which is pervasive in the dataset.

Pre- and post-processing also made slight differences in performance, but less than differences in the architecture of learners.

## 7. Conclusions

We have presented the results of the SpRadIE challenge for detection and classification of named entities and hedge cues in radiology reports in Spanish.

Seven teams participated in the challenge, achieving good performance, with results well above baselines. Although challenging entity types, like *Finding*, *Location* or hedge cues like *Conditional-Temporal* barely reach 74% F1, *Anatomical Entities*, *Measure* or *Abbreviation* can be recognized at almost 90% F1. This shows promising performance for integration within productive workflows.

Among the different approaches to the problem, we have found that combinations of multiple classifiers clearly outperform single multiclass models. Neural approaches specifically trained for Spanish also tend to perform better than generic or multilingual approaches. Also pre- and post-processing have a positive impact in performance.

Although Spanish has hundreds of millions of native speakers worldwide, not much work has been done in information extraction from Spanish medical reports. It is important to note that at least five of the seven participating teams have at least a Spanish native speaker. With this challenge, we provided a standard evaluation framework for a domain of Spanish medical text processing, namely radiology reports, that had not been previously addressed in this kind of effort.

We hope that this challenge and the promising results obtained by participating systems encourage other institutions to make resources publicly available, and thus contribute to the advancement in the automatic processing of medical texts, specially in Spanish.

## Acknowledgments

We want to thank Mariana Neves, for helping us shape the construction of the challenge.

## Author Contribution

VC & RR conceived the idea of the challenge. With equal contribution, VC & LAA co-led the Task, and with JV they reviewed the previously created annotation guidelines. LAA re-annotated reports. VC, LAA, FL & RR organized the task. RR & FL proposed the evaluation method and FL implemented the evaluation scripts, DF solved annotation criteria issues, AD calculated inter-annotator agreement, AA & LDF implemented the baselines, FC and MFU annotated for inter-annotator agreement metrics. LAA, RR, FL, JV, DF & VC discussed the results and contributed to the final manuscript.

## References

- [1] E. Pons, L. M. Braun, M. M. Hunink, J. A. Kors, Natural language processing in radiology: a systematic review, *Radiology* 279 (2016) 329–343.
- [2] L. Goeuriot, H. Suominen, L. Kelly, L. A. Alemany, N. Brew-Sam, V. Cotik, D. Filippo, G. G. Saez, F. Luque, P. Mulhem, G. Pasi, R. Roller, S. Seneviratne, J. Vivaldi, M. Viviani, C. Xu, CLEF eHealth 2021 Evaluation Lab, in: *Advances in Information Retrieval – 43st European Conference on IR Research*, Springer, Heidelberg, Germany, 2021.
- [3] V. Cotik, D. Filippo, R. Roller, H. Uszkoreit, F. Xu, Annotation of Entities and Relations in Spanish Radiology Reports, in: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 2017, pp. 177–184.
- [4] V. Cotik, Information extraction from Spanish radiology reports, in: *PhD Thesis*, 2018.
- [5] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 2019, pp. 590–597.
- [6] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports, *Scientific data* 6 (2019) 1–8.
- [7] W. J. Wilbur, A. Rzhetsky, H. Shatkay, New directions in biomedical text annotation: definitions, guidelines and corpus construction, *BMC bioinformatics* 7 (2006) 1–10.
- [8] M. Marimon, J. Vivaldi, N. Bel Rafecas, Annotation of negation in the iula spanish clinical record corpus, Blanco E, Morante R, Saurí R, editors. *SemBEaR 2017. Computational Semantics Beyond Events and Roles*; 2017 Apr 4; Valencia, Spain. Stroudsburg (PA): ACL; 2017. p. 43-52. (2017).
- [9] S. Lima Lopez, N. Perez, M. Cuadros, G. Rigau, NUBes: A corpus of negation and uncertainty in Spanish clinical texts, in: *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2020, pp. 5772–5781. URL: <https://aclanthology.org/2020.lrec-1.708>.

- [10] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, Padchest: A large chest x-ray image dataset with multi-label annotated reports, *Medical image analysis* 66 (2020) 101797.
- [11] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020, in: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, 2020.
- [12] A. Gonzalez-Agirre, M. Marimon, A. Intxaurreondo, O. Rabal, M. Villegas, M. Krallinger, PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track, in: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019*, pp. 1–10. URL: <https://www.aclweb.org/anthology/D19-5701>. doi:10.18653/v1/D19-5701.
- [13] M. Marimon, A. Gonzalez-Agirre, A. Intxaurreondo, H. Rodriguez, J. L. Martin, M. Villegas, M. Krallinger, Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results, in: *IberLEF@SEPLN*, 2019.
- [14] E. M. Cámara, M. D. Galiano, S. Estévez-Velarde, M. A. G. Cumbreiras, M. G. Vega, Y. Gutiérrez, A. M. Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, J. V. R. (eds.), Overview of tass 2018: Opinions, health and emotions, in: *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018), Sun SITE Central Europe, 2018*, pp. 13–27.
- [15] A. Piad-Morffis, Y. Gutierrez, P. Consuegra-Ayala, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the ehealth knowledge discovery challenge at iberlef 2019s, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, 2019, pp. 1–16.
- [16] A. Piad-Morffis, Y. Gutiérrez, H. Canizares-Diaz, S. Estevez-Velarde, R. Muñoz, A. Montoyo, Y. Almeida-Cruz, et al., Overview of the ehealth knowledge discovery challenge at iberlef 2020 (2020).
- [17] R. Costumero, F. López, C. Gonzalo-Martín, M. Millan, E. Menasalvas, An approach to detect negation on medical documents in spanish, in: *International Conference on Brain Informatics and Health, Springer*, 2014, pp. 366–375.
- [18] V. Cotik, V. Stricker, J. Vivaldi, H. Rodriguez, 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*, 2015.
- [19] W. Koza, D. Filippo, V. Cotik, V. Stricker, M. Muñoz, N. Godoy, N. Rivas, R. Martínez-Gamboa, Automatic detection of negated findings in radiological reports for spanish language: Methodology based on lexicon-grammatical information processing, *Journal of digital imaging* 32 (2019) 19–29.
- [20] S. Santiso, A. Pérez, A. Casillas, M. Oronoz, Neural negated entity recognition in spanish electronic health records, *Journal of biomedical informatics* 105 (2020) 103419.
- [21] R. R. Zavala, P. Martinez, The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: Comparative study, *JMIR Medical Informatics* 8 (2020) e18953.
- [22] V. Cotik, D. Filippo, J. Castaño, An approach for automatic classification of radiology reports in Spanish., in: *MedInfo*, 2015, pp. 634–638.
- [23] V. Cotik, H. Rodríguez, J. Vivaldi, Spanish named entity recognition in the biomedical

- domain, in: Annual International Symposium on Information Management and Big Data, Springer, 2018, pp. 233–248.
- [24] A. Casey, E. Davidson, M. Poon, H. Dong, D. Duma, A. Grivas, C. Grover, V. Suárez-Paniagua, R. Tobin, W. Whiteley, et al., A systematic review of natural language processing applied to radiology reports, *BMC medical informatics and decision making* 21 (2021) 1–18.
- [25] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: a web-based tool for nlp-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107.
- [26] R. Bossy, W. Golik, Z. Ratkovic, P. Bessières, C. Nédellec, BioNLP shared task 2013 — an overview of the Bacteria Biotope Task, in: Proceedings of the BioNLP Shared Task 2013 Workshop, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 161–169. URL: <https://www.aclweb.org/anthology/W13-2024>.
- [27] V. Suárez-Paniagua, H. Dong, A. Casey, A multi-BERT hybrid system for Named Entity Recognition in Spanish radiology reports, in: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2021.
- [28] P. López-Úbeda, M. C. Díaz-Galiano, L. A. Ureña-López, M. T. Martín-Valdivia, Pre-trained language models to extract information from radiological reports, in: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2021.
- [29] J. Canete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, PML4DC at ICLR 2020 2020 (2020).
- [30] M. Polignano, M. de Gemmis, G. Semeraro, Comparing Transformer-based NER approaches for analysing textual medical diagnoses, in: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2021.
- [31] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.
- [32] O. Solarte-Pabón, O. Montenegro, A. Blazquez-Herranz, H. Saputro, A. Rodriguez-González, E. Menasalvas, Information Extraction from Spanish Radiology Reports using multilingual BERT, in: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2021.
- [33] H. Fabregat, A. Duque, L. Araujo, J. Martinez-Romo, LSI\_UNED at CLEF eHealth2021: Exploring the effects of transfer learning in negation detection and entity recognition in clinical texts, in: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2021.
- [34] M. Ángel Martín-Caro García-Largo, I. S. Bedmar, Extracting information from radiology reports by Natural Language Processing and Deep Learning, in: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2021.
- [35] G. M. D. Nunzio, IMS-UNIPD @ CLEF eHealth Task 1: A Memory Based Reproducible Baseline, in: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2021.
- [36] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, Flair: An easy-to-use

framework for state-of-the-art nlp, in: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 2019, pp. 54–59.

- [37] A. Akbik, D. Blythe, R. Vollgraf, Contextual string embeddings for sequence labeling, in: COLING 2018, 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.