Automatic Extraction of Nested Entities in Clinical Referrals in Spanish

PABLO BÁEZ, Center for Medical Informatics and Telemedicine, University of Chile, Chile
FELIPE BRAVO-MARQUEZ, Department of Computer Science, University of Chile and IMFD, Chile
JOCELYN DUNSTAN*, Initiative for Data & Artificial Intelligence and Center for Mathematical Modeling - CNRS IRL 2807, University of Chile, Chile
MATÍAS ROJAS, Department of Computer Science, University of Chile, Chile
FABIÁN VILLENA, Center for Mathematical Modeling - CNRS IRL 2807, University of Chile, Chile

Here we describe a new clinical corpus rich in nested entities and a series of neural models to identify them. The corpus comprises de-identified referrals from the waiting list in Chilean public hospitals. A subset of 5,000 referrals (58.6% medical and 41.4% dental) was manually annotated with ten types of entities, six attributes, and pairs of relations with clinical relevance. In total, there are 110,771 annotated tokens. A trained medical doctor or dentist annotated these referrals, and then together with three other researchers, consolidated each of the annotations. The annotated corpus has 48.17% of entities embedded in other entities or containing another one. We use this corpus to build models for Named Entity Recognition (NER). The best results were achieved using a Multiple Single-entity architecture with clinical word embeddings stacked with character and Flair contextual embeddings. The entity with the best performance is *abbreviation*, and the hardest to recognize is *finding*. NER models applied to this corpus can leverage statistics of diseases and pending procedures. This work constitutes the first annotated corpus using clinical narratives from Chile and one of the few in Spanish. The annotated corpus, clinical word embeddings, annotation guidelines, and neural models are freely released to the community.

CCS Concepts: • **Computing methodologies** → **Language resources**; *Information extraction*; • **Applied computing** → Health care information systems.

Additional Key Words and Phrases: natural language processing, supervised machine learning, data mining, data curation, named entity recognition, clinical text mining

ACM Reference Format:
Pablo Báez, Felipe Bravo-Marquez, Jocelyn Dunstan, Matías Rojas, and Fabián Villena. 2021. Automatic Extraction of Nested Entities in Clinical Referrals in Spanish. *ACM Trans. Comput. Healthcare* XX, XX, Article XX (XX 2021), 23 pages. https://doi.org/XXX

*Corresponding author.

Authors' addresses: Pablo Báez, pablobaez@ug.uchile.cl, Center for Medical Informatics and Telemedicine, University of Chile, Av. Independencia 1027, Santiago, RM, Chile, ; Felipe Bravo-Marquez, Department of Computer Science, University of Chile and IMFD, Beauchef 851, Santiago, Chile, fbravo@dcc.uchile.cl; Jocelyn Dunstan, Initiative for Data & Artificial Intelligence and Center for Mathematical Modeling - CNRS IRL 2807, University of Chile, Beauchef 851, Santiago, RM, Chile, jdunstan@uchile.cl; Matías Rojas, matias.rojas.g@ug.uchile.cl, Department of Computer Science, University of Chile, Beauchef 851, Santiago, Chile; Fabián Villena, fabian.villena@uchile.cl, Center for Mathematical Modeling - CNRS IRL 2807, University of Chile, Beauchef 851, Santiago, Chile.

2637-8051/2021/XX-ARTXX $15.00
https://doi.org/XXX

1 Introduction

The analysis of clinical text has particular challenges due to the extensive use of non-standardized abbreviations, the variability of the clinical language across medical specialties and health professionals, and its restricted availability for privacy reasons, to mention a few [14]. Given that most text resources are available for the English language [58], focusing on clinical text in Spanish represents an opportunity to gather efforts on its development.

Human-annotated corpora are the essential resources for supervised learning methods [22], and although they are scarce and costly, they are necessary for at least three reasons: 1) the annotation procedure focuses and clarifies the requirements of a computational algorithm, 2) it provides data for training Natural Language Processing (NLP) systems, and 3) it provides a benchmark against which to evaluate the results obtained by computational models [64].

There is a general lack of annotated corpora for the Spanish language in the clinical domain, particularly in some specialties such as dentistry. There are also a few corpora potentially applicable in several clinical NLP tasks with heterogeneous documents in terms of clinical specialties and institutional origin [63].

A common task in NLP is Named Entity Recognition (NER), which aims to automatically identify essential pieces of information (entities) in a text written in natural language. In the general domain, NER was first defined to identify personal names, organizations, and locations [11], to then be extended to a variety of entities depending on the particular application. Nowadays, the best results for the original 2003 NER task [72] are self-attention networks [4], differentiable neural architecture search methods [28], and LSTM-CRF enriched with ELMo, BERT, and Flair contextual embeddings [71].

In the context of clinical NLP, NER is commonly used for the identification of diseases, body parts, or medications [14]. The automatic extraction of this information allows, for example, the detection of risk factors on discharge records [73], personal information [36], frequencies and doses of drugs [74], or the leverage of epidemiological information on the existence of diseases [42]. Nested NER is a particular case of NER where entities are nested within each other [18]. An example of that is "colon cancer", where a body part (colon) is contained in a disease. Due to this, some authors have reduced the problem of nested entities by keeping only the most external entity and removing inner ones, which is called flat NER. However, removing part of these entities could be a problem in model performance due to the loss of relevant information and inner dependencies.

We previously published a short version of this work at the clinical workshop of EMNLP 2020 [5]. In this manuscript, we show the results of an enlarged corpus made of 5,000 annotated referrals from diverse clinical specialties and institutions. In addition, we treat in depth methodological aspects of our work, giving details of the text processing, the neural methods for automatic detection of nested entities, and error analysis.

In summary, the main contributions of this work are as follows:

- There is a need to create language resources for the clinical Spanish language, and we are contributing by making the corpus, codes[1], and clinical word embeddings freely available.
- We work with de-identified clinical referrals. Many authors do not have access to clinical notes and work with corpora made from scientific articles, which is very different from the writing of health professionals attending patients. In addition, we have referrals from medical and dental specialties. Dental text in Spanish is a very scarce resource.
- Our corpus is one of the richest freely available resources to study nested NER, with almost half of the entity mentions nested. Therefore it can contribute substantially to future research.
- This paper presents a simple yet powerful model that performs competitively in identifying nested entities in this corpus. We also give a deep look at the error analysis in the context of nested NER. We explain the types of errors with a clinical example and summarize the results we get using our model.

[1] https://github.com/plncmm/acm_health_msen

In the following sections, we describe our corpus and its clinical relevance together with a summary of models that address nested entities.

1.1 The Chilean waiting list as the case study

In Chile, the public healthcare system covers 75% of the population [19]. The high demand for a visit to a specialist within this system, which requires a referral from a general practitioner, is handled by a Waiting List (WL) [51]. This is divided into "GES" (acronyms in Spanish for Explicit Health Guarantees), which covers 80 prioritized health conditions [49], and the "non-GES", which covers the remaining consultations. During 2016, about 22,500 patients died while waiting for their first consultation with a specialist, and 2,358 died before the surgery they needed. In 2017, there were 1,661,826 persons in the non-GES WL pending for an appointment with a specialist, with an average waiting time above 400 days [17].

Under this scenario, it is essential to develop automated systems that allow the analysis of this non-GES WL, to both improve the management of patients that should be prioritized as well as the secondary use of the information [6, 78]. Tasks that can be achieved with a working NER model include the prioritization of patients, the selection of cases that can be solved by telemedicine, the estimated number of people who present more than one disease (comorbidity), or that take more than one medication (polypharmacy), statistics of the pending procedures, or the family background of diseases when mentioned.

Every public health institution in Chile uploads weekly spreadsheets with non-GES WL cases. The referrals contain the personal information of the patient, the referring and admitting healthcare providers, the medical specialty, and in the form of unstructured text the suspected diagnosis [51]. Villena and Dunstan [77] examined the unstructured data in this WL, using word clouds to visualize the weighted word frequency by medical specialty. Although this methodology is informative, it is necessary to advance in the automatic detection of diseases within these referrals to improve their clinical management and support epidemiological studies, which is also one of the main motivations to create an annotated *corpus*. Apart from the clinical relevance, choosing the non-GES WL is also practical since it can be accessed through Transparency Law, a country-wide initiative for better access to data [52]. Data comes de-identified from the origin and does not require ethics committee approval as it is public information [47]. The public character of these referrals makes it possible to use them in shared tasks or to share them with the research community.

1.2 Related annotated corpora

In terms of linguistic resources using clinical text in Spanish, publications from Spain are predominant, such as the work by Oronoz et al. [61] that annotated disease, drug, and substance entities in medical records. The same group published a corpus afterward for adverse drug reactions [62]. For negation, there are the works of Cruz Diaz et al. [13] using anamnesis and radiology reports, Marimon et al. [45] using clinical reports from a hospital in Barcelona, and Lima-López et al. [39] who released a biomedical corpus annotated with negation and uncertainty. From Spanish-speaking countries besides Spain, and to the best of our knowledge, the only published work is by Cotik et al. [12] in Argentina for the annotation of clinical findings, body parts, negation, temporal terms, and abbreviations in radiology reports.

Some of the work done on biomedical texts is also noteworthy; Moreno-Sandoval and Campillos-Llanos [54] annotated Part-of-Speech in biomedical documents written in Spanish, Japanese, and Arabic, and Krallinger et al. [34] annotated PubMed abstracts in Spanish with chemicals and drugs. Several works have created resources in Spanish for entity recognition and clinical coding to internationally recognized classification systems; Kors et al. [33] created a multilingual corpus for biomedical concept recognition, Campillos-Llanos [8] created a medical lexicon and a clinical trials corpus [9] with words and entities mapped to the Unified Medical Language System (UMLS) [41] identifiers, while Intxaurrondo et al. [26] manually annotated abbreviation mentions and their definitions from clinical case studies and mapped them to control vocabulary resources such as the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT). Finally, there is the work of Miranda-Escalada

et al. [53] who published resources and methods for automatic clinical coding to the International Statistical Classification of diseases and Related Health Problems (ICD-10) on medical documents. Another important contribution was carried out by Marimon et al. [44] on anonymization of medical documents in Spanish using a synthetic corpus of clinical case documents.

Spanish is one of the most widely spoken languages globally, but there is a lack of language resources. Machine understanding of clinical texts requires dealing with a non-standardized use of the language, mainly due to the heavy use of abbreviations, local jargon, and a large presence of spelling errors. Creating clinical resources from different Spanish-speaking countries will allow us to estimate the variability of medical language. This comparison is especially useful when measured over real clinical narratives compared to biomedical literature due to its significantly different properties [21].

1.3 Related neural models for nested NER

Nested NER has received increasing attention from the research community since the publication of the biomedical corpus GENIA (in English) [60], which is rich in nested entities. In recent years, several methods have been proposed to handle this problem, which can be grouped into the following three approaches:

1.3.1 Region-based models: These approaches divide the problem into two stages, first the detection of entity boundaries, and second the assignment of entity labels to these regions. Sohrab and Miwa [68] enumerated all possible spans and took advantage of the representations of boundary tokens and the average of the internal token representations to predict the labels. Zheng et al. [81] used a BiLSTM to detect regions in a fixed-size window and then classified them into their categorical types in a single layer. Yu et al. [80] employed a biaffine model to assign scores to each pair of start and end tokens within a sentence, which are then used to predict the corresponding entity labels. Although the region-based approach is simple and effective, it has a high computational cost and cannot capture dependencies among entities.

1.3.2 Structure-based models: The aim is to capture the dependencies between internal and external entities through complex structures. The standard approach consists of creating these representations with hypergraphs [30, 43, 56, 79] or using a constituency graph [18]. However, as pointed out in Lin et al. [40], their major drawback is that they suffer from strong feature engineering, structure ambiguity, or spurious structures.

1.3.3 Sequence labeling-based models: Some studies show that sequence labeling-based methods can also perform well on the nested NER task. Alex et al. [3] proposed three CRF-based methods to reduce nested NER as several BIO tagging problems. The best performance was obtained using a cascading approach in which separate CRFs were used for each entity type, using the output of the previous CRF as the input features of the current CRF. Similarly, Ju et al. [29] took advantage of inner entity information to encourage outer entity recognition. They dynamically stacked LSTM-CRF layers predicting entities in an inside-to-outside way until no entities were extracted.

In addition to these approaches, recent studies reveal that incorporating BERT contextual embeddings [15] improves nested NER results. For instance, Li et al. [38] formulated nested NER as a machine reading comprehension task (MRC) using BERT as the model backbone. Another method, proposed by Straková et al. [71], uses sequence-to-sequence models leveraging the use of contextual embeddings from BERT and Flair [2], which is one of the elements to be incorporated in our proposed models.

As far as we know, little research exists on the use of various single-entity flat NER models for this task. We explore this approach in this work with successful results.

2 The Waiting List Corpus

During 2018, we requested the non-GES WL from the 29 health services in the country through Transparency Law [52]. These requests were answered positively by 23 of the health services and sent WL datasets for years between 2008 and 2018.

As a result, the group has 5,157,902 referrals, originated from the 40 medical and 11 dental specialties defined in the Chilean regulation [51]. The specialties with more referrals are ophthalmology (14.99%), traumatology (10.07%), and otorhinolaryngology (8.22%). The distribution between medical and dental referrals is 88% versus 12%.

Considering only the reasons for referral (written in free-text), we have 994,946 different diagnoses. A random subset of these diagnoses was selected for annotation, with the criterion of selecting those with more than 100 characters. Using this condition, we reduce the corpus to 107,235 unique candidates. Moreover, we removed diagnoses with text imperfections (such as a clear cut at the end of the referral or a text encoding error) or without extra text information (an exact copy of an ICD-10 diagnosis). After filtering, one of the managers inspected each of the remaining diagnoses to ensure that they fully met the conditions. Even though the referrals come de-identified from the source, this person also checked that no personal information was disclosed.

2.1 Annotation procedure

Five annotators (three medical students, one medical doctor, and one dentist) were trained for the annotation process, and were permanently supported by three project managers. The choice of annotators and their background could be a significant factor. Roberts et al. [65] describe how clinically trained annotators are better than linguists and computer scientists at annotating clinical text with semantic relationships. However, it has also been shown that annotators without medical education achieve high agreement in the semantic annotation of a clinical corpus [7]. It is also common to collect annotations from workers with advanced medical training, either as general practitioners, researchers with training on general medicine, or final-year medical students [32]. Recruiting medical doctors to invest time on the annotation task is a challenge. Therefore, the option of training non-expert annotators such as students, is often considered. We worked here with three third-year medical students, whose annotations contributed to the improvement of the annotation guidelines.

The annotation process involved two stages as shown in Figure 1. In the first stage, a test version of the annotation guidelines was written, with an in-depth study of other available guidelines for similar entities, such as those published by Mota et al. [55] and Intxaurrondo et al. [26]. These guidelines were evaluated during the annotation of 25 referrals, followed by the curation of a reference.

In the second stage, the three medical students annotated the same 50 referrals in weekly annotation rounds for three weeks. In an iterative improvement process, the medical students were retrained after each round of annotation. At this point, the guidelines were further modified to clarify the task and improve consistency. At the end of this stage, the first accepted version of the guidelines was established and released.

A medical doctor or dentist (namely a *senior annotator*) was asked to annotate the same 150 referrals done by the students independently. Each referral was compared with the previous annotations, with the expectation that the analysis and discussion process of finding consensus on annotations helped strengthen the senior annotator's training. We decided to implement a manual pre-annotation stage of the simplest entities, such as abbreviations and body parts, done by medical students. Thus, the senior annotators could focus on entities, attributes, and relations that required higher clinical expertise.

For the consolidation process, we decided to have each annotation revised by a team of four researchers, including at least one senior annotator, a dentist, the manager who worked on the annotation guidelines, and the principal investigator. This means that once a batch of 150 referrals was fully annotated, the three managers and the senior annotator analyzed and discussed the annotations one by one until an agreement was reached. When consolidated, the referrals became part of the ground truth. In the beginning, the consolidation of 150

referrals took around 6 hours, but by round 4, the time was reduced to approximately 3 hours. It is important to note that we did not use automatic pre-annotation methods, each of the referrals was manually annotated from scratch. We used this time-consuming approach to compensate for the absence of a second senior medical or dental annotator.

2.2 Annotation scheme

The annotation scheme includes entities, attributes, and relations between annotated entities, and was designed based on a literature review and our particular interest in this corpus. The UMLS semantic network was used to define 11 medically relevant entities; Finding (with two children entities: Laboratory or Test Result and Sign or Symptom), Procedure (with three children entities: Laboratory, Diagnostic and Therapeutic Procedures), Disease, Family Member, Body Part, and Medication.

We also created the category Abbreviation as an entity, although it could be considered an entity's attribute rather than as an entity per se. In this regard, Névéol et al. [59] considered and annotated abbreviations as an independent entity, while Miñarro-Giménez et al. [48] and Savkov et al. [66] annotated only the long (full) forms of acronyms and abbreviations. On the other hand, the Cantemist [53] and PharmaCoNER [1] tasks did not consider abbreviations as independent entities or attributes; only abbreviated mentions of those entities of importance to the task were annotated considering their long-form. We have two main reasons to justify our decision on Abbreviation as an entity. First, identifying and resolving clinical abbreviations is a complex and interesting problem [27] that we want to address in the future. No study has captured and described the diversity of abbreviations in Chilean clinical text, nor has it focused on their resolution. This task is relevant because the abbreviations used are not always standardized, especially in settings under time pressure [14], such as primary health care. Second, because although abbreviated forms of entities such as Disease, Procedure or Medication may be found (in which case it would be logical to consider these forms as an attribute), not all abbreviations have an *a priori* assigned entity, and yet they may be clinically relevant. Therefore, we could lose that information by considering the abbreviation as an attribute because they would not be annotated.

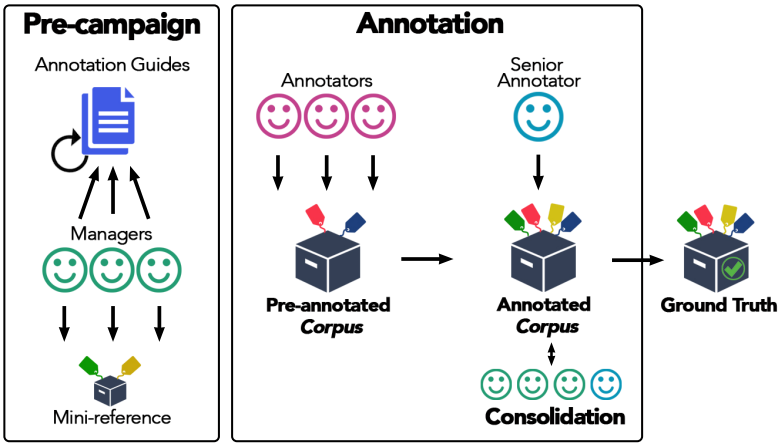


Fig. 1. Annotation stages for the ground truth creation. During the pre-campaign, the annotation guides were developed and tested. The annotation stage included: pre-annotation done by medical students, annotation done by senior annotators, and consolidation done by managers and at least one senior annotator. At the end of the annotation stage, 5,000 referrals were consolidated as ground truth. Figure adapted from Fort [20].

The annotation scheme also comprises the following four attributes: Negated (for Sign or Symptom, Disease, and Procedures) for entities accompanied by negation modifiers, Pending (for Procedures) because the referrals are from a waiting list, and we were interested in describing how many procedures were pending, Maternal or Paternal (for Family Member), and Implicit Family Background (for Sign or Symptom and Disease) because we are also interested in mining the family history of diseases. The corresponding entities were annotated with the label attribute, without including the token(s) used to express the attribute (e.g., in “waiting for surgery", the entity “surgery" is annotated as a Therapeutic Procedure with the Pending attribute). We included Implicit Family Background to consider expressions such as *there is a family history of cancer* without specifying which family member(s) present the disease. Figure 2 summarizes the entities and attributes covered by our annotation scheme.

Finally, the *has* relation was used to connect certain entities, such as Family Member with Disease or Diagnostic Procedure with Laboratory or Test Result. Following the previous example, the entity *cancer* should be connected to the entity that carries family member information when corresponding.

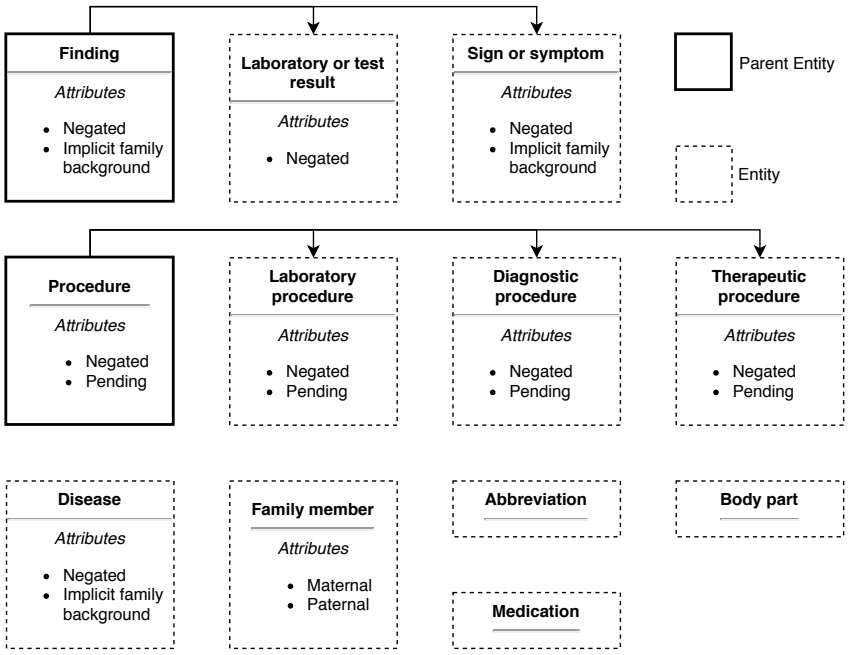


Fig. 2. Description of the entities and attributes we are annotating in the corpus.

2.3 Annotation guidelines

A document with the guidelines for annotators was created by the managers, which was a result of a literature review and their annotation during Stage I [26, 55, 59, 67, 75, 76]. As previously mentioned, UMLS was used to define the entity names and dependencies, while ICD-10 was used to resolve disagreements and uncertainties related, for instance, to the scope of the annotations.

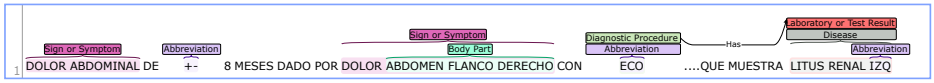
The guidelines were initially designed to instruct medical students and were later improved by the feedback given by the senior annotators. In the current version, the guidelines start with a brief introduction to clinical NLP and instructions to initiate a session in the platform and perform the annotation using BRAT [70] (BRAT Rapid Annotation Tool). This is always complemented with face-to-face meetings with the annotators.

The guidelines are under constant update when the need for clarification or further example cases emerge from the consolidation process. The current version of the annotation guidelines (in Spanish) is freely available[2].

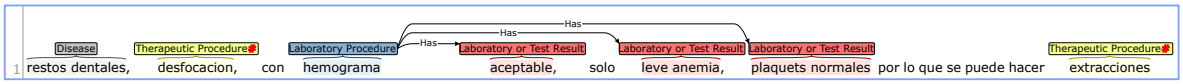
According to the guidelines, the rules for annotation are classified into four types: (i) general, which are suitable for positive and negative rules, (ii) positive, what has to be annotated, (iii) negative, what should not be annotated and (iv) multi-word, when to consider multiple tokens in an entity.

Two general rules were then explained, which are not to include punctuation and white spaces at the end of entities and to annotate even if grammatical errors are found, as long as the meaning is understood. Afterward, the entity was briefly defined, followed by positive, negative and multi-word rules - each of them supported by several examples. The text was complemented with illustrative diagrams of the annotations.

An example of a dental and medical referral is shown in Fig. 3 where a translation to English is provided in the caption.



(a) Medical annotated referral



(b) Dental annotated referral

Fig. 3. Examples of annotated referrals using BRAT Rapid Annotation Tool. These can be translated into English as (a) “abdominal pain of +- 8 months given by right flank abdominal pain with usg .... showing left kidney stone” and (b) “dental remnants, dental extraction, with acceptable hemogram, only slight anemia, normal platelets so extractions can be done" (in Spanish, the word for platelets has a typo. USG: Ultrasonography).

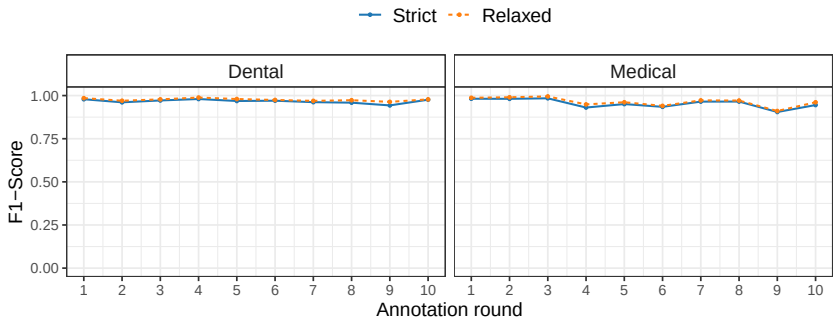
2.4 Inter-annotator agreement

The difficulty of achieving consistent annotations was assessed by calculating the inter-annotator agreement during first and second stage [20]. In particular, we used the F1-Score to compare pairs of annotations [25]. The F1-Scores can be “strict” and “relaxed”. In the strict case, the pair is required to match exactly in entity and tokens selected, while in the relaxed case, the annotations are required to have the same class. However, there may be a partial match in the entity length, with an overlap of tokens. As an example, for the expression “breast cancer,” if an annotator A marks only “cancer” as a disease, and annotator B decides to select the full expression “breast cancer” as a disease, using the strict metric there would be no agreement between A and B. In contrast, with the relaxed metric there would be agreement since both annotators include the word “cancer”.

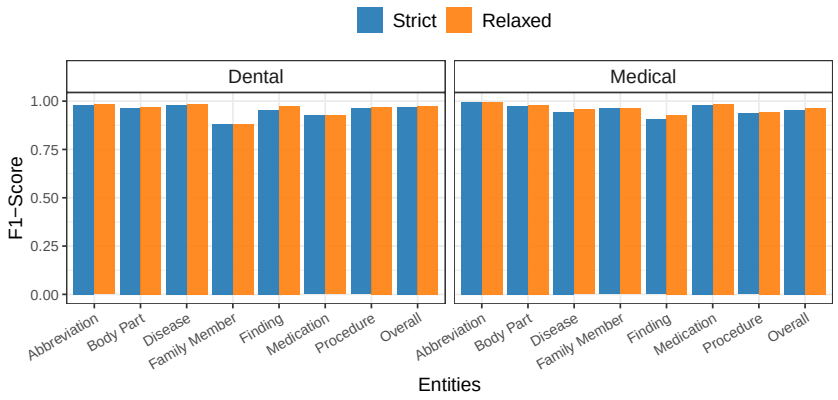
We followed up the agreement between medical students until an agreement above 0.7 was reached on entities such as body part, abbreviation, and medication (data not shown). We did not follow up their agreements beyond the first 150 pre-annotated referrals, as each senior annotator verified and corrected any potential errors identified during their annotation rounds. Instead, we focused on comparisons between each senior annotator’s annotations and the ground truth obtained after each round of consolidation. As mentioned, once the referrals are pre-annotated, the senior annotator (medical doctor or dentist) carries out the annotations’ first full version. The referrals are then consolidated by the three managers and the senior annotator. Although this method is not traditional, we consider it quite robust, as four people at the same time evaluated each of the annotated referrals on the whole corpus. The time required during the consolidation process decreased as several rounds of annotation were performed. Figure 4a shows the agreement trend across the consolidation rounds, with some

[2]https://plncmm.github.io/annodoc/

oscillations of the F1-Score but with excellent values, consistently above 0.9, and minor differences between strict and relaxed forms. Figure 4b shows the high agreement scores for each annotated entity and the small differences between strict and relaxed forms. The entities Family Member and Medication obtained the lowest F1-scores (0.88 and 0.92 respectively) in the dental subcorpus, which could be explained by the fact that they are low-frequency entities and, therefore, a missing annotation (false negative) impacts the agreement considerably. On the other hand, in the medical subcorpus, Finding was the entity with the lowest strict F1-Score (0.9), reaching a slight increase in the relaxed form (0.93), suggesting a conflict at the scope annotation level for this entity in both subcorpora.



(a)



(b)

Fig. 4. F1-score (strict and relaxed) for (a) the ten annotation rounds and (b) the annotated entities across the the dental and medical subcorpus.

3 Automatic Entity Recognition

We tackle the task of Named Entity Recognition on the corpus described above. In particular, we had to deal with nested entities, for which we proposed an approach based on single entities. In the following, we first define

what we will consider a nested entity and the nested NER task, the preprocessing of the corpus, the experimental design to evaluate the performance of different models, and a baseline. Finally, we explain the different types of errors that can be found in this nested NER problem.

3.1 Formalization of the task

We consider that the formalization of the nested NER task has not been addressed in-depth, and clarification of the different borderlines cases is needed. Moreover, for the particular type of NER we are trying to solve, we found a gap of knowledge related to spans associated with multiple entity types, which was first noticed by Alex et al. [3]. We therefore proceed to define what we understand by nested entities.

Definition 1 (Nested entities). *Given an input sequence $X = \{x_1, x_2, ..., x_n\}$ of words, an entity Q is defined by a tuple (S_q, E_q, T_q), where S_q and $E_q \in [1, n]$ represents entity boundaries in X, and T_q in $\mathcal{E}$ (the entity space) corresponds to entity type. Given two entities Q and R, we say that Q is anidated in R if $S_r \leq S_q$ and $E_q \leq E_r$ (with $T_r \neq T_q$). The particular case of $S_q = S_r$ and $E_q = E_r$ corresponds to an entity with multiple labels, in our statistics we consider these cases as nested entities for both entity types.*

Please note that under this definition, *HTA*, which is an abbreviation for hypertension, will be both a disease and an abbreviation as well as be counted as one example for both entity types. Similarly, we define the problem of nested NER as follows:

Definition 2 (Nested NER). *Given an input sequence $X = \{x_1, x_2, ..., x_n\}$, the aim of nested NER is to correctly identify the boundaries for every entity Q in X and assign it the correct entity type from a predefined list of categories. This identification must be done for cases where nested entities are involved and when not (flat NER).*

Section 3.6 describes the error analysis by which the possible assignment of an incorrect entity type, wrong entity boundaries, and other types of errors were evaluated.

3.2 Data preprocessing

BRAT annotation software generates a file in *standoff* format[3] for each referral. This file follows a basic structure with columns containing the following: an ID per annotation and its consecutive order of appearance, the entity type with the indexes of the beginning and end characters of the annotation, and the character string that constitutes the entity. We implemented a pre-processing script in our repository to transform these files into the format requested by each architecture tested. In all cases, we converted our BRAT annotated files into a CoNLL-like format [72], which is a standard input format for several NLP libraries. We followed the IOB tagging format for NER (short for inside, outside, beginning). Finally, for the tokenization, we used *esnewslg*, which is a Spanish statistical tokenizer trained with the Spanish AnCora and WikiNER [23, 24]. We further enhanced it by adding a tokenizer based on regular expressions.

3.3 Multiple Single-entity NER models

We refer to our approach to nested NER as Multiple Single-entity NER (MSEN), which consists of independently training multiple flat NER models, one for each entity type. Specifically, to create each of these models, we followed the LSTM-CRF approach proposed by Lample et al. [35], one of the most widely used architectures for sequence labeling models. The output of each model is merged with the others to obtain all the entities, including the nested ones. As we have different models, the advantage is that we can run them in parallel. Figure 5 shows an overview of our architecture. To encode the input sentences, we produced different combinations of embeddings. We first trained domain-specific embeddings concatenated with character embeddings using a

[3]http://2011.bionlp-st.org/home/file-formats

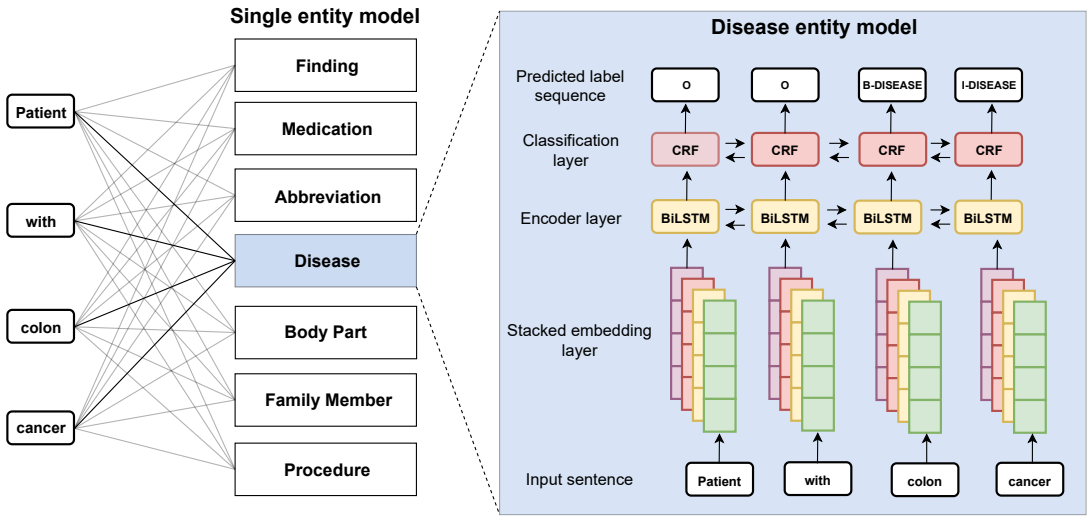


Fig. 5. Overview of the MSEN architecture, where each entity type has a flat NER model associated with it. The right side of the figure shows, as an example, the detailed flat NER architecture of the disease entity. Labels of the input sentences correspond to the union of the outputs of each of these models, thus capturing the nested entities.

BiLSTM character-level language model suggested by Lample et al. [35]. We also enrich the embedding layer by adding contextualized embeddings from Flair [2] and BERT [15]. The output is fed into a BiLSTM encoding layer to obtain long-contextual information. Finally, we use a CRF and Viterbi algorithm to decode the most likely tag sequence using the IOB tagging format.

3.4 Experimental design

3.4.1 Baseline: We chose as a baseline the Neural Layered model proposed by Ju et al. [29]. It works by dynamically stacking LSTM-CRF layers to predict entities in an inside-to-outside way until no entities are extracted. Both this architecture and our proposed model belong to the sequence labeling-based category described in Section 1.3. The decision was based on how this architecture treats nested entities and the ease of adapting the code to our corpus. In the Waiting List Corpus, 10.75% of the entities are involved in spans of text tagged with multiple entity types, which is a problem little addressed in the literature, and this architecture, like the MSEN model, can deal with these cases. Moreover, the Neural Layered model is inspired by Lample's architecture, which facilitates the comparison of hyperparameters with our proposed model since both have similar components. The source code is freely available[4] to reproduce the experiments, and input files can be obtained using our preprocessing toolkit described above.

3.4.2 Settings: The MSEN model was implemented using the Flair framework[5]. To encode sentences, we used word embeddings previously trained over a clinical corpus built by the combination of the unannotated Waiting List Corpus described in Section 2 and referrals collected by the group for another project [78], comprising 56,079,828 tokens and 57,112 types. These 300-dimensional clinical embeddings can be downloaded from here[6]. For the experiments with contextual embeddings, we used the cased version of Spanish BERT (BETO) [10] without fine-tuning. Since BERT uses word-piece tokenization, we decided to compute the word embeddings

[4]https://github.com/meizhiju/layered-bilstm-crf
[5]https://github.com/zalandoresearch/flair
[6]http://doi.org/10.5281/zenodo.3924799

using the average of subtokens embeddings. For Flair embeddings [2], we concatenated the Spanish-forward and Spanish-backward embeddings created using a character-level language model.

We chose the best hyperparameters via random search over the range of parameters shown in Table 1. We tried to use hyperparameters that are as similar as possible between our model and the baseline. During the training process, pre-trained embeddings were not left static, and out-of-vocabulary words were initialized using a zero vector.

After preprocessing the referrals, we obtained 9,894 sentences. 8,014 were used for training, 890 for validation, and 990 for testing, leading to a ratio of 0.81: 0.09: 0.1, the same ratio used in similar nested NER corpora, such as GENIA [31]. The MSEN model is trained up to a maximum of 100 epochs. For optimization, we used SGD with a mini-batch size of 16 and an initial learning rate of 0.1. We used a learning rate scheduler and an early stopping strategy based on the performance of the development partition. We also applied dropout regularization [69] after the embedding layer and BiLSTM.

Table 1. Hyperparameter search space and best values found for the baseline and our model (MSEN).

Parameter	Range	Baseline	MSEN
max epochs	[20, 100]	20	100
optimizer	[SGD, Adam, AdamW]	Adam	SGD
batch size	[8, 32]	16	16
learning rate	[0.0001, 0.1]	0.001	0.1
char emb dim	[20, 50]	25	25
dropout	[0.2, 0.8]	0.3	0.3
BiLSTM depth	[1, 3]	3	3
BiLSTM hidden size	[128, 512]	128	128

3.5 Evaluation metrics.

Our system was evaluated using precision, recall, and micro F1-score. An entity is considered correct when both entity type and boundaries are predicted correctly, which is the official evaluation standard for NER used in Conll-03 [72]. Note that this metric is analogous to the strict inter-annotator agreement calculated in Section 2.4.

Reporting a single performance score is insufficient to compare non-deterministic approaches since results might change when using different subsets. Here we report a k-fold cross-validation process using $k = 10$. To determine whether the observed differences between the performance of the best MSEN model and the baseline were statistically significant, we performed a k-fold cross-validated paired t-test [16], setting the α level at P <0.05.

Since neural network models are stochastic processes, it is important to mention that replicating these experiments may lead to slightly different results in different runs. To ensure the reproducibility of our experiments, we made public in the repository the partitions used for this process and the original subsets on which the experiments were tested.

3.6 Error analysis.

For a better comprehension and explainability of our entity recognition model, an analysis of the errors was performed using the work proposed by Nejadgholi et al. [57] but modified for nested entities. The output from an entity recognition model can be incorrect because either the span is wrong, or the label is wrong (or both). Based on these principles, we distinguish five types of errors listed below and exemplified in Figure 6

(1) False-positive: the model predicts one or more entities that are not annotated in the test subset.

True annotation	abdominal pain of +- 8 months given by right flank abdominal pain with usg .... showing left kidney stone
Error type	Predicted annotation
False-positive	abdominal pain of +- 8 months given by right flank abdominal pain with usg .... showing **left kidney stone**
False-negative	abdominal pain of +- 8 months given by **right flank abdominal pain** with usg .... showing left kidney stone
Wrong label, right span	abdominal pain of +- 8 months given by right flank abdominal pain with **usg** .... showing left kidney stone
Wrong label, overlapping span	abdominal pain of +- 8 months given by right flank abdominal pain **with usg** .... showing left kidney stone
Right label, overlapping span	abdominal pain of +- 8 months given by right flank **ab dominal pain** with usg .... showing left kidney stone

Fig. 6. Example annotations for each error type. A correctly annotated span of text is described in the head, and malformed annotations are described below. For illustrative purposes, we are only showing annotations for the Sign or Symptom (in light purple) and Diagnostic Procedure (in dark green). Malformed annotations are shown in bold. Note that we are using the first referral showed in Figure 3.

(2) False-negative: the model predicts no entities for a given span, but the test subset contains no entities. This malformed addition can be complete (the model predicted no entities for the span) or partial (the model predicted an incomplete list of entities for the given span)
(3) Wrong label, right span: an annotated entity in the test subset and the predicted entity have the exact spans but different entities.
(4) Wrong label, overlapping span: the annotated entity in the test subset and the predicted entity have overlapping spans and different entity classes.
(5) Right label, overlapping span: the annotated entity in the test subset and the predicted entity have the same entity classes but overlapping spans.

4 Results

4.1 Annotated corpus statistics

The corpus is a collection of 5,000 referrals divided in two: 2,067 dental, and 2,933 medical. The statistics of the full corpus and subcorpus documents and entities are described in Table 2 and 3. The document distribution among the dental and medical specialties are described in Tables 4 and 5, respectively. The annotated corpus is freely available for non-commercial use[7]. From the total of annotated entities, the distribution of tokens per entity type and entity types per document (referral) is shown in Fig. 7. In terms of the annotated attributes and relations, they are much less in number than entities. For the attributes, we have 1,068 negated, 973 pending, 26 implicit family background, 1 paternal and 3 maternal. For relations, we have 1,114 pairs of relations.

As previously mentioned, this corpus has nested entities. Figure 8 illustrates this fact, with numbers indicating how many times the entity in the row is nested in the entity in the column. Please note that this matrix is not symmetric, as it is much more common to find, for example, an abbreviation in a finding than the other way

[7]https://doi.org/10.5281/zenodo.4287459

around. In fact, in the medical subcorpus, abbreviations are nested 1323 times in findings, while findings are 263 times part of an abbreviation. Besides, when nested annotations have the same length, we count them as embedded into each other for both entities. An example of that is *HTA* (hypertension), which is both a disease and an abbreviation. In summary, 48.17% of the entities in the corpus are embedded in or contain other entities, and the maximum nesting depth is three.

Table 2. Corpus statistics divided by medical and dental documents.

Metric	Total	Medical	Dental
Documents	5,000	2,933	2,067
Tokens	181,706	123,162	58,544
Vocabulary	22,165	18,675	6,543
Lexical diversity[a]	12.2 %	15.2 %	11.2 %
Mean tokens per document[b]	36.3 (29.2)	42.0 (35.5)	28.3 (13.0)
Mean entities per document[b]	9.0 (7.2)	10.5 (8.7)	6.9 (3.2)
Annotated tokens	110,771	71,384	39,387
Entities	44,941	30,766	14,175

[a]Ratio of different unique tokens to the total number of tokens.
[b]Standard deviation in parentheses.

Table 3. Quantity of entities by entity class and subcorpus.

Subcorpus	Total	Medical	Dental
Abbreviation	10,961	8,514	2,447
Body Part	7,265	4,226	3,039
Disease	11,103	7,775	3,328
Family Member	298	276	22
Finding	10,677	6,890	3,787
Medication	1,050	846	204
Procedure	3,587	2,239	1,348

4.2 Nested Named Entity Recognition

As shown in Table 6, the approach proposed by Ju et al. [29] was used as a reference to compare with our models. As expected, each setting of the MSEN architecture outperforms the baseline because it focuses on only one entity type per model. We also noticed that by adding new elements to the embeddings layer, such as contextual embeddings, the results improve even more. The model with the best result (highlighted in bold) is the one that used clinical word embeddings concatenated with character and Flair embeddings, achieving a micro F1-score of 80.27.

Regarding the best model, Table 7 shows precision, recall and F1-score per entity, as well as the number of examples in the test partition. The entity with the best results was abbreviation, which is expected since it is easy to recognize from the morphological point of view. This entity is usually one token long; therefore, the chances of being mistaken due to wrong boundaries are low. The opposite occurs with the entity finding, which is four

Table 4. Documents distribution by dental specialty.

Dental specialty	Documents	Percentage
Oral rehabilitation: Removable dentures	515	10.30 %
Endodontics	501	10.02 %
Orthodontics	343	6.86 %
Periodontology	343	6.86 %
Maxillofacial surgery	142	2.84 %
Oral Surgery	114	2.28 %
Oral rehabilitation: Crowns	51	1.02 %
Operative dentistry	23	0.46 %
Temporomandibular disorders and orofacial pain	3	0.06 %
General dentistry	3	0.06 %

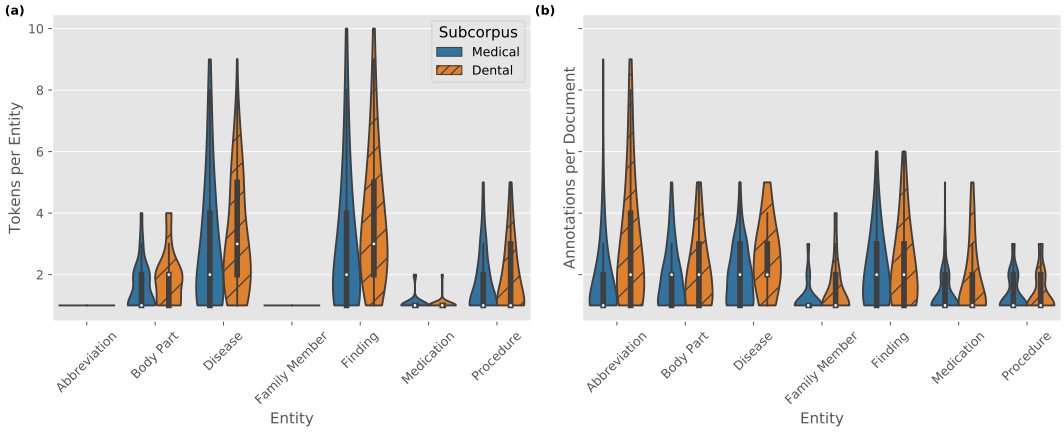


Fig. 7. Frequency distribution and median (white point) of (a) tokens per entity across the subcorpus, and (b) annotated entities per document by subcorpus.

tokens long on average, thus very easy to have it wrong in the limits. Moreover, clinical findings are the hardest to have consistently annotated by humans, hence is not surprising that it is also hard to be identified by the model.

The cross-validation process demonstrated the efficacy and high level of generalization of the MSEN model on unseen data, significantly outperforming the baseline in all measurements (Table 8), consistent with the results in Table 6. In practical terms, the statistical results and the k-fold cross-validation provide convincing evidence that the MSEN and Layered models perform differently.

4.3 Error analysis

Our best MSEN model made 1,302 errors on the test subset. The highest proportion of errors corresponds to *right label, overlapping span* (38%), followed by false negatives (29.6%) and false positives (22%) (Figure 9a). The finding entity class is the entity with the highest proportion of these three types of errors, covering almost 60% of *right label, overlapping span* error, 40% of false negatives and 35% of false positives (Figure 9b). It has previously

Table 5. Documents distribution by medical specialty.

	Documents	Percentage
Traumatology	489	9.78 %
Gynecology	277	5.54 %
Otorhinolaryngology	223	4.46 %
Ophthalmology	216	4.32 %
Neurology	197	3.94 %
Internal medicine	174	3.48 %
Surgery	168	3.36 %
Pediatrics	158	3.16 %
Cardiology	150	3.00 %
Gastroenterology	131	2.62 %
Dermatology	105	2.10 %
Urology	96	1.92 %
Psychiatry	80	1.60 %
Vascular surgery	64	1.28 %
Endocrinology	56	1.12 %
Pediatric surgery	53	1.06 %
Nephrology	53	1.06 %
Pulmonology	43	0.86 %
Obstetrics	43	0.86 %
Neurosurgery	38	0.76 %
Abdominal surgery	23	0.46 %
Rheumatology	20	0.40 %
Hematology	15	0.3 %
Physical medicine and rehabilitation	13	0.26 %
Infectology	10	0.20 %
Oncology	9	0.18 %
Genetics	9	0.18 %
Colorectal surgery	7	0.14 %
Breast Surgery	6	0.12 %
Plastic Surgery	3	0.06 %
Geriatrics	2	0.04 %
Cardiothoracic Surgery	1	0.02 %
Anesthesiology	1	0.02 %

been reported that better NER models generate more *right label, overlapping span* errors, suggesting that it could be because the span information may be more vague in the representation resulting from contextualized embeddings by combining the meaning of words through an attention mechanism. Consequently, proper treatment of this type of error is essential in the comparison of modern NER systems [57].

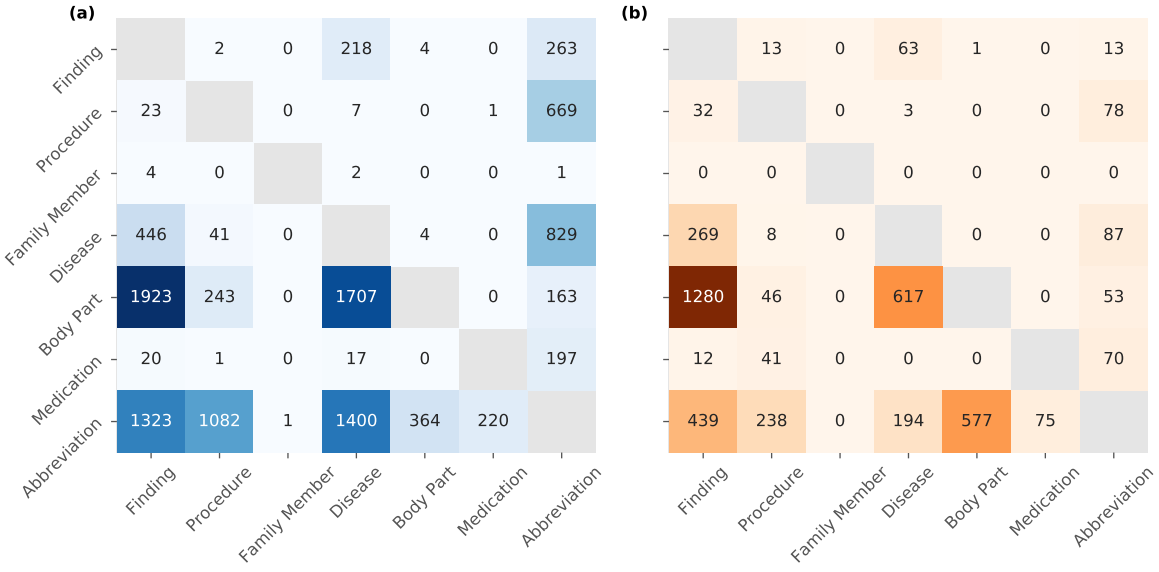


Fig. 8. Characterization of nested entities. The numbers indicate how many times the entity in the row is embedded in the entity in the column. (a) represents the medical subcorpus and (b) represents the dental subcorpus.

Table 6. Results obtained with different models on the Waiting List Corpus. Here *Word* stands for word embedding, *Char* is character embedding, and the Flair and BERT models were implemented as described in the text.

Model	Precision	Recall	F1-Score
Neural Layered Model [29] (baseline)	77.0	72.12	74.48
MSEN [Word]	76.59	74.84	75.71
MSEN [Word+Char]	77.75	78.29	78.02
MSEN [Word+Char+BERT]	79.72	78.83	79.27
MSEN [Word+Char+Flair]	**80.24**	**80.30**	**80.27**
MSEN [Word+Char+Flair+BERT]	79.90	78.13	79.01

Regarding the *wrong label* errors, the confusion matrix (Figure 10) shows that the finding and disease entity classes are more often confused by the model, and this could be mainly because of the close semantic relatedness of both entity classes; these classes are often subject to discussion even by the expert annotators.

5 Conclusions and future work

In this paper we have presented a new entity-annotated corpus of medical and dental referrals in Spanish as well as an in-depth experimental study of neural NER models trained on it. The entities we chose are clinically relevant, and include nested entities. Detecting these entities was addressed using a Multiple Single-entity NER approach, which obtains better results than a state-of-the-art baseline. As well, adding contextualized embeddings led to a significant improvement in metrics, especially in recall. We also share the 5,000 annotated referrals, the annotation guidelines, clinical word embeddings, and the code used to generate the results presented here.

Table 7. Results for each entity type using the best model in the test dataset.

Entity	Precision	Recall	F1-Score	Support
Abbreviations	93.65	95.07	94.35	993
Disease	82.65	83.19	82.92	1,071
Medication	87.21	81.52	84.27	92
Finding	62.31	62.13	62.22	1,059
Body Part	85.91	87.01	86.46	708
Family Member	96.55	87.50	91.80	32
Procedure	72.96	69.46	71.17	334

Table 8. Results of the 10-fold cross-validation on the best MSEN model and the baseline. Results are calculated based on the micro F1-score metric.

	Neural Layered Model [29] (baseline)	MSEN [Word + Char + Flair]	P value[a]
Mean	73.20	79.81	$8.8e^{-9}$
SD	0.752	0.469	
Min	72.16	79.16	
Max	74.65	80.66	

[a]10-fold cross-validated paired t-test.

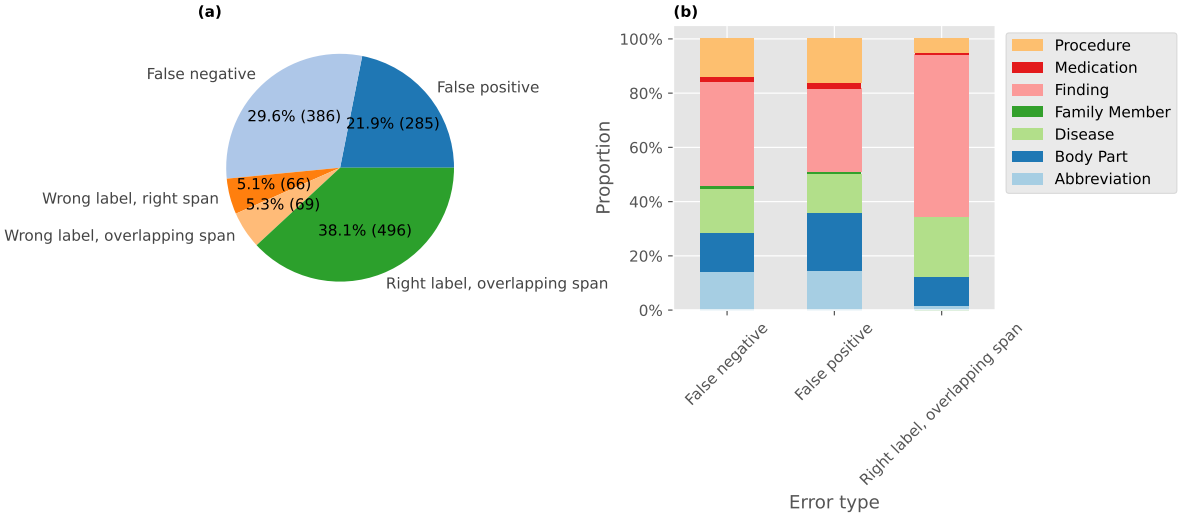


Fig. 9. Distribution of the errors types found by the error analysis on the incorrect best models' predictions on the test subset. Panel (a) shows the overall distribution of the error types, and panel (b) shows the distribution of entities inside error types.

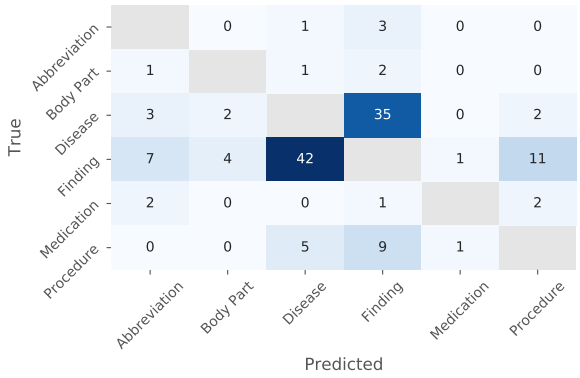


Fig. 10. Confusion matrix for the wrong label errors found by the error analysis on the incorrect best models' predictions on the test subset.

In terms of developing NER models trained on this corpus, future work includes improving the recall for procedure and finding due to the importance of identifying these entities. The low values in the metrics for finding can be explained mainly by a lack of agreement in the boundaries. It is interesting to notice that the model analysis has also helped us to identify annotation inconsistencies. We plan to homogenize the annotations, tracking inconsistencies through an automatic harmonization process similar to Campillos et al. [7].

Besides, our annotated corpus has hierarchical entities (for example, test result and sign/symptom are part of the entity finding), and we plan to investigate the hierarchical nested NER using architectures as in Marinho et al. [46]. Finally, our corpus has attributes and relations which we have not addressed yet. Once we have a higher amount of annotated referrals, we plan to host a shared task to advance this corpus's multiple challenges.

To the best of our knowledge, this is the first annotated dental corpus for the Spanish language, and one of the strengths in this work is having a medical doctor and a dentist on the team. The lack of a second medical or dental annotator can be perceived as a weakness of the study, however we consider that our annotation scheme was very rigorous, requiring the simultaneous evaluation of four researchers during consolidation to obtain high quality annotations.

One of our goals working on this corpus and training NER models is to recognize diseases within this waiting list automatically. An example of that is the recognition of new cases of psoriasis within this waiting list [37], which could be extended for the detection of all diseases. Besides, telemedicine has been posed as one of the solutions to decrease the waiting times in the Chilean public healthcare sector [50]. To correctly estimate the effect, one needs to summarize the suspected diagnoses and check which of them are eligible for telemedicine consultations. We believe our work can speed up this task.

Part of the group's expertise is in the genetic components of diseases. For that reason, we want to explore the possible risk factors (genetic or environmental), which could be obtained from the mentions of the patients' family history and habits. In this regard, we pay special attention to identifying relations between family members and diseases, with maternal and paternal components labeled. This corpus is not particularly rich in those entities. However, we are starting to collaborate with a cancer center, and we plan to translate the know-how from this annotated corpus to future projects in that direction.

Acknowledgments

This work was funded by Centro de Modelamiento Matemático (CMM), ACE210010, AFB170001, and FB21005 Basal Funds for Center of Excellence from ANID-Chile. In addition, we got funding from U-INICIA VID UI-004/19 and UI-004/20, FONDECYT 11201250 and 11200290, CIMT-CORFO cost center 570111, ICM P09-015F, Postdoctoral FONDECYT 3210395, and ANID - Millennium Science Initiative Program - Code ICN17_002. We thank Maicol Fernández, Manuel Durán, and Esteban Galindo for performing annotations and consolidations on this corpus, and Ren Cerro for English proofreading. This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM- 02).

References

[1] Aitor Gonzalez Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. 1–10.

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1638–1649. https://www.aclweb.org/anthology/C18-1139

[3] Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising Nested Named Entities in Biomedical Text. In *Biological, translational, and clinical language processing*. Association for Computational Linguistics, Prague, Czech Republic, 65–72. https://www.aclweb.org/anthology/W07-1009

[4] Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785* (2019).

[5] Pablo Báez, Fabián Villena, Matías Rojas, Manuel Durán, and Jocelyn Dunstan. 2020. The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Online, 291–300. https://doi.org/10.18653/v1/2020.clinicalnlp-1.32

[6] Pablo Báez, Fabián Villena, Karen Zúñiga, Natalia Jones, Gustavo Fernández, Manuel Durán, and Jocelyn Dunstan. 2021. Construcción de recursos de texto para la identificación automática de información clínica en narrativas no estructuradas. *Revista médica de Chile* 149, 7 (2021), 1014–1022.

[7] Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. 2018. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). *Language Resources and Evaluation* 52, 2 (2018), 571–601.

[8] Leonardo Campillos-Llanos. 2019. First Steps towards Building a Medical Lexicon for Spanish with Linguistic and Semantic Information. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. 152–164.

[9] Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. 2021. A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC Medical Informatics and Decision Making* 21, 1 (2021), 1–19.

[10] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

[11] Nancy Chinchor and Patricia Robinson. 1997. MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, Vol. 29. 1–21.

[12] Viviana Cotik, Darío Filippo, Roland Roller, Hans Uszkoreit, and Feiyu Xu. 2017. Annotation of Entities and Relations in Spanish Radiology Reports.. In *RANLP*. 177–184.

[13] Noa P Cruz Diaz, Roser Morante, Manuel J Mana 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*. 53–58.

[14] Hercules Dalianis. 2018. *Clinical text mining: Secondary use of electronic patient records*. Springer Nature.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[16] Thomas G. Dietterich. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation* 10, 7 (1998), 1895–1923. https://doi.org/10.1162/089976698300017197

[17] Roberto Estay, Cristóbal Cuadrado, Francisca Crispi, Fernando González, Francisco Alvarado, and Natalia Cabrera. 2017. Desde el conflicto de listas de espera, hacia el fortalecimiento de los prestadores públicos de salud: Una propuesta para Chile. *Cuadernos Médico*

Sociales 57, 1 (2017).

[18] Jenny Rose Finkel and Christopher D Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 conference on empirical methods in natural language processing*. 141–150.

[19] Fondo Nacional de Salud. 2013. *Población Inscrita en FONASA, https://public.tableau.com/views/Poblacion2002-2020/INEeInscritos*. Technical Report.

[20] Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

[21] Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of biomedical informatics* 35, 4 (2002), 222–235.

[22] Archana Goyal, Vishal Gupta, and Manish Kumar. 2018. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review* 29 (2018), 21–43.

[23] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[24] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303

[25] George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association* 12, 3 (2005), 296–298.

[26] Ander Intxaurrondo, Juan Carlos de la Torre, H Rodriguez Betanco, Montserrat Marimon, Jose Antonio Lopez-Martin, Aitor Gonzalez-Agirre, J Santamarıa, Marta Villegas, and Martin Krallinger. 2018. Resources, guidelines and annotations for the recognition, definition resolution and concept normalization of Spanish clinical abbreviations: the BARR2 corpus. In *SEPLN*.

[27] Ander Intxaurrondo, Montserrat Marimon, Aitor Gonzalez-Agirre, Jose Antonio Lopez-Martin, Heidy Rodriguez, Jesus Santamaria, Marta Villegas, and Martin Krallinger. 2018. Finding Mentions of Abbreviations and Their Definitions in Spanish Clinical Cases: The BARR2 Shared Task Evaluation Results.. In *IberEval@ SEPLN*. 280–289.

[28] Yufan Jiang, Chi Hu, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2019. Improved differentiable architecture search for language modeling and named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3576–3581.

[29] Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A Neural Layered Model for Nested Named Entity Recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1446–1459. https://doi.org/10.18653/v1/N18-1131

[30] Arzoo Katiyar and Claire Cardie. 2018. Nested Named Entity Recognition Revisited. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 861–871. https://doi.org/10.18653/v1/N18-1079

[31] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. *GENIA corpus—a semantically annotated corpus for bio-textmining*. Vol. 19. i180–i182 pages.

[32] Rob Koeling, John Carroll, Rosemary Tate, and Amanda Nicholson. 2011. Annotating a corpus of clinical text records for learning to recognize symptoms automatically. In *Proceedings of LOUHI 2011 Third International Workshop on Health Document Text Mining and Information Analysis. CEUR Workshop Proceedings*. 43–50.

[33] Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M Van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association* 22, 5 (2015), 948–956.

[34] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics* 7, 1 (2015), 1–17.

[35] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 260–270. https://doi.org/10.18653/v1/N16-1030

[36] Lukas Lange, Heike Adel, and Jannik Strötgen. 2019. NLNDE: The neither-language-nor-domain-experts' way of Spanish medical document de-identification. *CEUR Workshop Proceedings* 2421 (2019), 671–678.

[37] C Lecaros, J Dunstan, F Villena, DM Ashcroft, R Parisi, CEM Griffiths, S Härtel, JT Maul, and C De la Cruz. 2021. The incidence of psoriasis in Chile: an analysis of the national Waiting List Repository. *Clinical and Experimental Dermatology* (2021). https://doi.org/doi:10.1111/ced.14713

[38] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A Unified MRC Framework for Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5849–5859. https://doi.org/10.18653/v1/2020.acl-main.519

[39] Salvador Lima-López, Naiara Pérez, Montse Cuadros, and German Rigau. 2020. Nubes: A corpus of negation and uncertainty in spanish clinical texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 5772–5781.
[40] Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2019. Sequence-to-Nuggets: Nested Entity Mention Detection via Anchor-Region Networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5182–5192. https://doi.org/10.18653/v1/P19-1511
[41] Donald A. Lindberg, Betsy L. Humphreys, and Alexa T. McCray. 1993. The unified medical language system. *Methods of information in medicine* 32, 4 (1993), 281.
[42] Jason P Lott, Denise M Boudreau, Ray L Barnhill, Martin A Weinstock, Eleanor Knopp, Michael W Piepkorn, David E Elder, Steven R Knezevich, Andrew Baer, Anna NA Tosteson, et al. 2018. Population-based analysis of histologically confirmed melanocytic proliferations using natural language processing. *JAMA dermatology* 154, 1 (2018), 24–29.
[43] Wei Lu and Dan Roth. 2015. Joint Mention Extraction and Classification with Mention Hypergraphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 857–867. https://doi.org/10.18653/v1/D15-1102
[44] Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results.. In *IberLEF@ SEPLN*. 618–638.
[45] Montserrat Marimon, Jorge Vivaldi, and Núria Bel Rafecas. 2017. 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).
[46] Zita Marinho, Alfonso Mendes, Sebastiao Miranda, and David Nogueira. 2019. Hierarchical nested named entity recognition. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 28–34.
[47] Diego A Martinez, Haoxiang Zhang, Magdalena Bastias, Felipe Feijoo, Jeremiah Hinson, Rodrigo Martinez, Jocelyn Dunstan, Scott Levin, and Diana Prieto. 2019. Prolonged wait time is associated with increased mortality for Chilean waiting list patients with non-prioritized conditions. *BMC public health* 19, 1 (2019), 233.
[48] Jose A Miñarro-Giménez, Ronald Cornet, Marie-Christine Jaulent, Heike Dewenter, Sylvia Thun, Kirstine Rosenbeck Gøeg, Daniel Karlsson, and Stefan Schulz. 2019. Quantitative analysis of manual annotation of clinical text samples. *International journal of medical informatics* 123 (2019), 37–48.
[49] Ministerio de Salud de Chile. 2004. Ley 19.966, https://www.leychile.cl/Navegar?idNorma=229834.
[50] Ministerio de Salud de Chile. 2011. Estrategia Nacional de Salud para el cumplimiento de los Objetivos Sanitarios de la Década 2010-2020.
[51] Ministerio de Salud de Chile. 2011. Norma Técnica Para El Registro De Las Listas De Espera, www.minsal.cl/wp-content/uploads/2016/03/Norma-Tecnica-118.pdf.
[52] Ministerio Secretaría General de la Presidencia. 2008. Ley 20.285, https://www.leychile.cl/Navegar?idNorma=276363&idParte=.
[53] Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. 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*.
[54] Antonio Moreno-Sandoval and Leonardo Campillos-Llanos. 2013. Design and Annotation of MultiMedica–A Multilingual Text Corpus of the Biomedical Domain. *Procedia-Social and Behavioral Sciences* 95 (2013), 33–39.
[55] Enrique Mota, Nelson Martín, Ángel Moreno, Elvira Ferrete, Jesús Santamaría, Montserrat Marimon, Ander Intxaurrondo, Aitor González-Agirre, Marta Villegas, and Martin Krallinger. 2018. Guías de anotación de información de salud protegida.
[56] Aldrian Obaja Muis and Wei Lu. 2017. Labeling Gaps Between Words: Recognizing Overlapping Mentions with Mention Separators. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2608–2618. https://doi.org/10.18653/v1/D17-1276
[57] Isar Nejadgholi, Kathleen C. Fraser, and Berry de Bruijn. 2020. Extensive Error Analysis and a Learning-Based Evaluation of Medical Entity Recognition Systems to Approximate User Experience. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Online, 177–186. https://doi.org/10.18653/v1/2020.bionlp-1.19
[58] Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics* 9, 1 (2018), 12.
[59] Aurélie Névéol, Rezarta Islamaj Doğan, and Zhiyong Lu. 2011. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. *Journal of biomedical informatics* 44, 2 (2011), 310–318.
[60] Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, Hideki Mima, and Junichi Tsujii. 2002. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the second international conference on Human Language Technology Research*. 82–86.
[61] Maite Oronoz, Arantza Casillas, Koldo Gojenola, and Alicia Perez. 2013. Automatic annotation of medical records in Spanish with disease, drug and substance names. In *Iberoamerican Congress on Pattern Recognition*. Springer, 536–543.

[62] 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 (2015), 318–332.

[63] Ana Carolina Peters, Adalniza Moura Pucca da Silva, Caroline P Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid A Hasan, Claudia Maria Cabral Moro, et al. 2020. SemClinBr–a multi institutional and multi specialty semantically annotated corpus for Portuguese clinical NLP tasks. *arXiv preprint arXiv:2001.10071* (2020).

[64] Angus Roberts, Robert Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay Subbarao Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, et al. 2007. The CLEF corpus: semantic annotation of clinical text. In *AMIA Annual Symposium Proceedings*, Vol. 2007. American Medical Informatics Association, 625.

[65] Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics* 42, 5 (2009), 950–966.

[66] Aleksandar Savkov, John Carroll, Rob Koeling, and Jackie Cassell. 2016. Annotating patient clinical records with syntactic chunks and named entities: the Harvey Corpus. *Language resources and evaluation* 50, 3 (2016), 523–548.

[67] 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 (2014), 148–158.

[68] Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep Exhaustive Model for Nested Named Entity Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2843–2849. https://doi.org/10.18653/v1/D18-1309

[69] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

[70] 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 at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 102–107.

[71] Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural Architectures for Nested NER through Linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5326–5331. https://doi.org/10.18653/v1/P19-1527

[72] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 142–147. https://www.aclweb.org/anthology/W03-0419

[73] Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association* 15, 1 (2008), 14–24.

[74] Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association* 17, 5 (2010), 514–518.

[75] Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association* 17, 5 (2010), 519–523.

[76] 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 (2011), 552–556.

[77] Fabián Villena and Jocelyn Dunstan. 2019. Obtención automática de palabras clave en textos clínicos: una aplicación de procesamiento del lenguaje natural a datos masivos de sospecha diagnóstica en Chile. *Revista médica de Chile* 147, 10 (2019), 1229–1238.

[78] Fabián Villena, Jorge Perez, René Lagos, and Jocelyn Dunstan. 2021. Supporting the classification of patients in public hospitals in Chile by designing, deploying and validating a system based on natural language processing. *BMC Medical Informatics and Decision Making* 21, 1 (2021), 208. https://doi.org/10.1186/s12911-021-01565-z

[79] Bailin Wang and Wei Lu. 2018. Neural Segmental Hypergraphs for Overlapping Mention Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 204–214. https://doi.org/10.18653/v1/D18-1019

[80] Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6470–6476. https://doi.org/10.18653/v1/2020.acl-main.577

[81] Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung, and Guandong Xu. 2019. A Boundary-aware Neural Model for Nested Named Entity Recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 357–366. https://doi.org/10.18653/v1/D19-1034