The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish

Pablo Báez[1], Fabián Villena[1,2], Matías Rojas[3], Manuel Durán[1], and Jocelyn Dunstan[1,2]

[1]Center for Medical Informatics and Telemedicine, University of Chile.
[2]Center for Mathematical Modeling, University of Chile.
[3]Department of Computer Sciences, University of Chile.
{pablobaez,manuel.duran,matias.rojas.g}@ug.uchile.cl
{fabian.villena,jdunstan}@uchile.cl

Abstract

In this work we describe the Waiting List Corpus consisting of de-identified referrals for several specialty consultations from the waiting list in Chilean public hospitals. A subset of 900 referrals was manually annotated with 9,029 entities, 385 attributes, and 284 pairs of relations with clinical relevance. A trained medical doctor annotated these referrals, and then together with other three researchers, consolidated each of the annotations. The annotated corpus has nested entities, with 32.2% of entities embedded in other entities. We use this annotated corpus to obtain preliminary results for Named Entity Recognition (NER). The best results were achieved by using a biLSTM-CRF architecture using word embeddings trained over Spanish Wikipedia together with clinical embeddings computed by the group. NER models applied to this corpus can leverage statistics of diseases and pending procedures within this waiting list. This work constitutes the first annotated corpus using clinical narratives from Chile, and one of the few for the Spanish language. The annotated corpus, the clinical word embeddings, and the annotation guidelines are freely released to the research community.

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 some (Dalianis, 2018). Given that most text resources are available for the English language (Névéol et al., 2018), focusing on clinical text in Spanish represents an opportunity to gather efforts on its development.

A common task in Natural Language Processing (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 (Chinchor and Robinson, 1997), to then be extended to a variety of entities depending on the particular application. Nowadays, the best results for the original 2003 NER task (Sang and De Meulder, 2003) are self-attention networks (Baevski et al., 2019), differentiable neural architecture search methods (Jiang et al., 2019), and LSTM-CRF enriched with ELMo, BERT, and Flair contextual embeddings (Straková et al., 2019).

In the context of clinical NLP, NER is commonly used for the identification of diseases, body parts, or medications (Dalianis, 2018). The automatic extraction of this information allows, for example, the detection of risk factors on discharge records (Uzuner et al., 2008), personal information (Lange et al., 2019), frequencies and doses of drugs (Uzuner et al., 2010a), or the leverage of epidemiological information on the existence of diseases (Lott et al., 2018).

Human-annotated clinical corpora are costly, but 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 resolving NLP tasks, and 3) it provides a benchmark against which to evaluate the results obtained by computational models (Roberts et al., 2007).

In practice, the manual annotation process implies that the annotator, with expertise in a subject previously discussed and defined as appropriate, reviews the corpus. Following guidelines and an annotation scheme, he or she selects a text segment in the document and assigns it to an entity type and, if appropriate, adds an attribute or relation to connect the segment to another entity.

1.1 The Chilean waiting list as the case study

In Chile, the public healthcare system covers 75% of the population (Fondo Nacional de Salud, 2013).

Proceedings of the 3rd Clinical Natural Language Processing Workshop, pages 291–300
November 19, 2020. ©2020 Association for Computational Linguistics

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) (Ministerio de Salud de Chile, 2011b). This is divided into “GES” (acronyms in Spanish for Explicit Health Guarantees), which covers 80 prioritized health conditions (Ministerio de Salud de Chile, 2004), 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 a specialist’s appointment, with an average waiting time above 400 days (Estay et al., 2017).

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. 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 que contiene informacion sobre las interconsultas. 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 (Ministerio de Salud de Chile, 2011b). Villena and Dunstan (2019) 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 (Ministerio Secretaría General de la Presidencia, 2008). Data comes de-identified from the origin and does not require ethics committee approval as it is public information (Martinez et al., 2019). 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 of Oronoz et al. (2013) that annotated disease, drug, and substance in medical records. The same group published a corpus afterward for adverse drug reactions (Oronoz et al., 2015). For negation, there are the works of Cruz Diaz et al. (2017) using anamnesis and radiology reports, Marimon et al. (2017) using clinical reports from a hospital in Barcelona, and Lima et al. (2020) 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. (2017) 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 (2013) annotated Part-of-Speech in biomedical documents written in Spanish, Japanese, and Arabic,Krallinger et al. (2015) annotated PubMed abstracts in Spanish with chemicals and drugs. More recently, Campillos-Llanos (2019) created a medical lexicon by mapping words to the Unified Medical Language System (UMLS) identifiers. 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.

2 The Waiting List Corpus

During 2018, we requested the non-GES WL from the 29 health services in the country through Transparency Law (Ministerio Secretaría General de la Presidencia, 2008). 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 5,176,858 referrals, originated at the 40 medical and 11 dental specialties defined in the Chilean regulation (Ministerio de Salud de Chile, 2011b). The specialties with more referrals are ophthalmology (14.49%), traumatology (9.44%), and otorhinolaryngology (7.53%). The distribution between medical and dental referrals is 83% versus 11%, and 6% of the referrals have missing values in the specialty attribute.

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 for any personal information.

3 Annotation scheme

Four annotators (three medical students and one medical doctor) were selected for the initial stage of the annotation process, who were permanently supported by three project managers. The choice of annotators and their background is a significant factor. Roberts et al. (2009) describe how clinically trained annotators are better than linguists and computer scientists at annotating clinical text with semantic relationships. It is 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 (Koeling et al., 2011). We worked here with three third-year medical students, whose annotations contributed to the improvement of the annotation guidelines.

The annotation process involved three 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. (2018) and Intxaurrondo et al. (2018). 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 50 identical 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.

In stage three, a medical doctor joined the group (namely *senior annotator*) and was asked to annotate the same 150 referrals done by the students independently. Each referral was compared with the previous annotations, with the aim that the analysis and discussion process to find consensus on annotations helped strengthen the senior annotator's training. Recruiting medical doctors to invest time

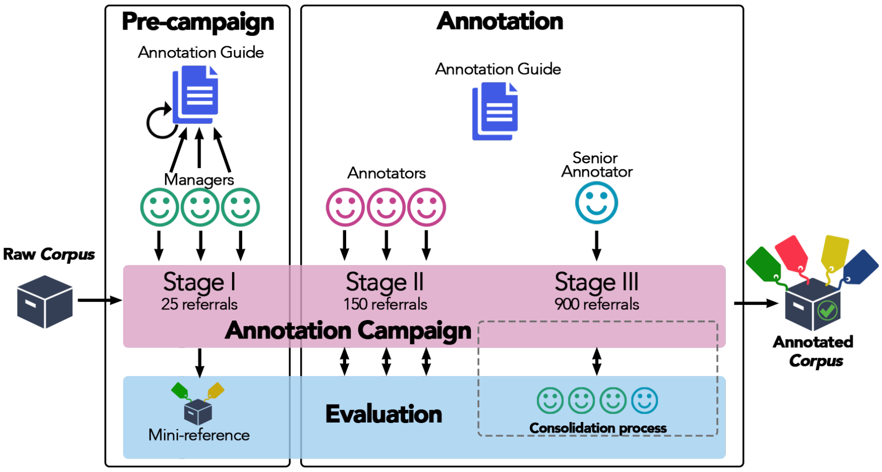


Figure 1: Annotation stages for the creation of annotation guidelines, the training of the senior annotator, and the production stage where 900 referrals were consolidated. Figure adapted from Fort (2016).

on the annotation task is a challenge. Therefore, the option of training non-expert annotators such as students, is often considered. However, the annotation of some complex entities, by definition or extension, may lead to low agreement among non-expert annotators impacting the overall agreement as well (Lewinski et al., 2017). We addressed this situation by implementing a pre-annotation stage of straightforward entities, such as abbreviations and body parts, done by medical students. Thus, the senior annotator 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 the senior annotator, a dentist, the postdoc that created 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 annotator.

3.1 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 (Mota et al., 2018; Uzuner et al., 2010b, 2011; Névéol et al., 2011; Intxaurrondo et al., 2018; Skeppstedt et al., 2014). The Unified Medical Language System (UMLS) was used to define the entity names and dependencies and resolve disagreements and uncertainties.

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

The guidelines are under constant update when the need for clarification or further example cases emerges from the consolidation process. The current version of the annotation guidelines (in Spanish) is freely available [1].

The annotated entities and attributes are described in Table 1. The choice of entities was based on literature revision and our interest within this corpus. For example, the referrals are from a waiting list, and we were interested in describing how many procedures were pending. Moreover, it was important for us to distinguish between laboratory, diagnostic or therapeutic procedures. We are also interested in mining the family history of diseases, and therefore, we included entities, attributes, and relations *ad hoc* with this goal. The corpus was for example, enriched with the *has* relation between family members and disease, and to connect diagnostic procedures and laboratory or test results.

	Entity	**Attribute**
Finding	Laboratory or Test Result	
	Sign or Symptom	Negated
Procedure	Laboratory Procedure	Pending
	Diagnostic Procedure	
	Therapeutic Procedure	
	Family Member	Maternal Paternal
	Disease	Negated IFB
	Body Part	
	Medication	
	Abbreviation	

Table 1: Description of the entities and attributes we are annotating in the corpus. IFB: Implicit Family Background

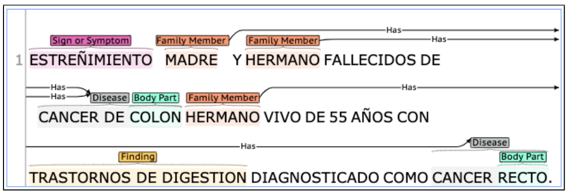
For all the entities, the rules were classified into four types: (i) general, which are suitable for positive and negative rules, (ii) positive, what has to be annotated, (iii) negative, what should not be annotated and (iv) multi-word, when to consider multiple tokens in an entity. Two general rules 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 diagrams constructed from screenshots of the platform.

For the case of attributes, there were four types: negated (for sign/symptoms and diseases), pend-

[1] https://fvillena.github.io/annodoc/

ing (for procedures), maternal or paternal (for family member), and implicit family background (for diseases). The corresponding entities were annotated with the label, 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. Finally, relations were used to connect certain entities. Following the previous example, the entity *cancer* should be connected to the entity that carries family member information when corresponding.

An example of an annotated referral is shown in Fig. 2. In this referral, one can see three relations between diseases and family members. Additionally, we observe nested entities in *cancer de colon* (colon cancer) and *cancer de recto* (rectal cancer). In both cases, there is a body part contained in a longer disease entity. Section 4 describes nested entities within this corpus.



1 ESTREÑIMIENTO MADRE Y HERMANO FALLECIDOS DE
CANCER DE COLON HERMANO VIVO DE 55 AÑOS CON
TRASTORNOS DE DIGESTION DIAGNOSTICADO COMO CANCER RECTO.

Figure 2: An example of an annotated referral using BRAT Rapid Annotation Tool. This can be translated into English as follows: "Constipation deceased mother and brother of colon cancer living brother aged 55 with digestive disorders diagnosed as rectal cancer"

3.2 Inter-annotator agreement

The difficulty of the task was assessed by calculating the inter-annotator agreement during Stages I and II (Fort, 2016). In particular, we used the F1-Score to compare pairs of annotations (Hripcsak and Rothschild, 2005). The F1-Scores can be "strict" and "relaxed". In the strict case, the annotation is required to match exactly in entity and tokens selected, while in the relaxed case, the annotation is 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 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 calculated the inter-annotator agreement between pairs, considering the three medical students and the ground truth. Figure 3 shows the F1 strict and relaxed for every pair in 150 referrals.

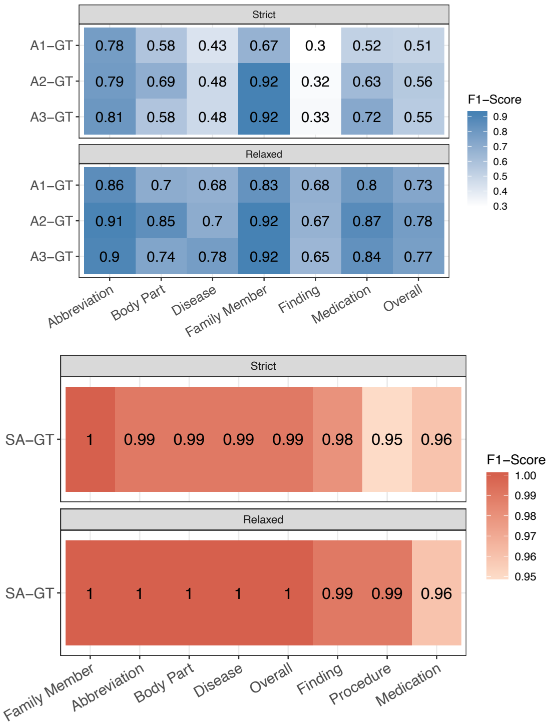


Figure 3: F1-score (strict and relaxed). Top: Every medical student (A1, A2 and A3) compared with the ground truth and calculated over the first 150 referrals. Bottom: Comparison between the senior annotator and the ground truth for the referrals 1-150.

As mentioned before, the senior annotator carries out the first version of the annotations. The referrals are then consolidated by the three managers and the senior annotator. Both the time required and the number of editions during the consolidation process decrease as several rounds of annotation are achieved. Figure 3 (bottom) shows the comparison between the senior annotator and the ground truth over 150 referrals (referrals 1-150 in the corpus).

4 Results

4.1 Annotated corpus statistics

The corpus consists of 900 referrals, with 1,912 sentences, 36,157 tokens, and a vocabulary size of 7,980 tokens. Each diagnosis has a mean of 40 [37 - 42 CI 95 %] tokens, normally distributed across

the diagnoses. The medical specialties more often annotated are traumatology (16.64%), gynecology (8.85%) and pediatrics (7.02%). The ratio between medical and dental specialties is 88:12. The annotated corpus is freely available[2].
A total of 9,029 entities were annotated and the distribution per entity type and document (referral) is shown in Fig. 4. In terms of the annotated attributes and relations, they are much less in number than entities. For the attributes, we have 256 negated, 126 pending, 2 implicit family background, and 1 maternal. For relations, we have 284 pairs of relations.

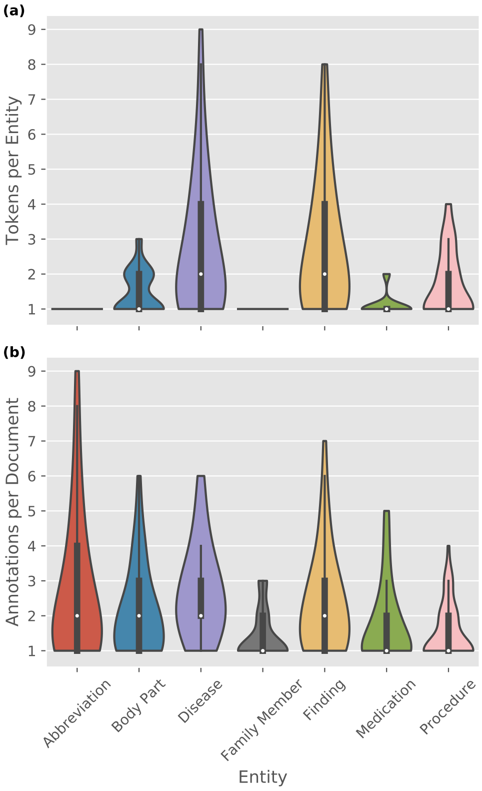
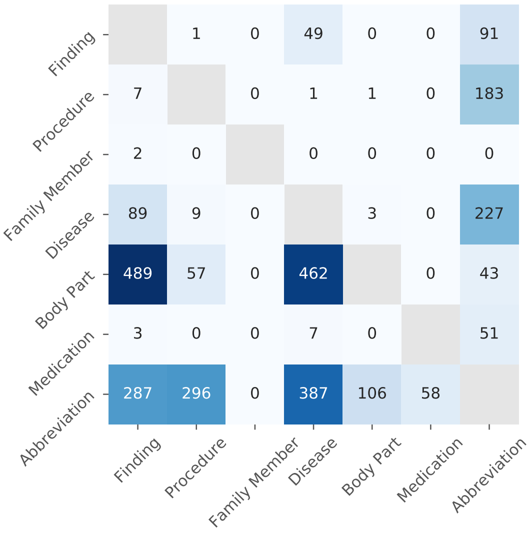


Figure 4: Frequency distribution and median (white point) of (a) tokens per entity across the corpus, and (b) annotated entities per document.

As previously mentioned, this corpus has nested entities, which are entities embedded in other entities (Finkel and Manning, 2009). For example, in Figure 2, the body part *colon* is nested inside the disease entity *cancer de colon*. Figure 5 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 (287 times) than a finding in an abbreviation (91). Besides, when nested annotations have the same length, we count them as embedded one into each other for both entities. An example of that is *HTA* (*high blood pressure*), which is both a disease and an abbreviation.



	Finding	Procedure	Family Member	Disease	Body Part	Medication	Abbreviation
Finding		1	0	49	0	0	91
Procedure	7		0	1	1	0	183
Family Member	2	0		0	0	0	0
Disease	89	9	0		3	0	227
Body Part	489	57	0	462		0	43
Medication	3	0	0	7	0		51
Abbreviation	287	296	0	387	106	58	

Figure 5: Characterization of nested entities. The numbers indicate how many times the entity in the row is embedded in the entity in the column.

4.2 Preliminary NER models

BRAT annotation generates a file in *standoff* format[3] for each referral. This file follows a basic structure with three columns containing: an ID per annotation and its consecutive order of appearance, the entity type with the indexes for the beginning and end characters of the annotation, and the character string that constitute that entity.

These files can be converted to the *CoNLL* format[4] (Furrer et al., 2019), which is widely used in the NLP community. Unfortunately, this format does not support nested entities, therefore we have to choose which nested entity to use (commonly the longest). When a token is annotated with two entities, *HTA* for example, to translate it to *CoNLL* format we have to keep one of the two arbitrarily.

For the preliminary results shown here, we decided to focus on three specific entities. Disease, because its recognition is a task of enormous clinical relevance. Medication, since medical doc-

[2] http://doi.org/10.5281/zenodo.3926705

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

[4] https://github.com/spyysalo/standoff2conll

tors sometimes prescribe the active component and other the commercial brand, we wanted to explore how well the different tested models deal with that. Finally, abbreviations were chosen since they are widespread in the corpus and are morphologically distinctive.

We compared the performance of a multiclass model (where nested entities are lost) with three models for each entity (where all entities are retained, no matter if they are nested or nesting an entity). As a baseline, we used the Flair Framework, a biLSTM-CRF architecture that creates contextual embeddings for each word. This approach was the state-of-the-art for the NER CoNNL03 task in English and German (Akbik et al., 2018). This architecture is easy to implement as code and pre-trained language models are available to the community[5].

For the embedding layer, we compared Flair embeddings pre-trained over Wikipedia in Spanish with those enhanced by domain-specific embeddings. The latest were trained over a clinical corpus composed by the unannotated Waiting List Corpus described in Section 2 plus referrals collected by the group for another project. The vocabulary size of this corpus is 57,112 tokens. These clinical embeddings can be downloaded from here[6]. Furthermore, the two embeddings were not left static, so the weights were updated during the training stage.

The grid-search method from Flair was used to tune the hyperparameters: a learning rate of 0.1, batch size of 32, 100 epochs, LSTM hidden size of 256 and dropout 0.1864. The models were tested ten times with different initialization parameters. The results were expressed as mean and standard deviation (SD). Table 2 shows the results for the multiclass model where nested entities are lost, while Table 3 shows the results of three different NER models, one for each of the entities.

As expected, multiple models outperform a multiclass model where nested entities are lost. In terms of the embedding layer's choice, models with added clinical embeddings have a better performance than those using the Spanish Wikipedia Flair embeddings alone. Nevertheless, it is interesting the small difference between the two for the recognition of abbreviations. This is probably due to the different sizes of both corpora, where the training corpus for the Flair embedding is significantly larger, but also the fact that the embeddings are dynamically tuned during training. The most significant improvement of adding clinical embeddings is observed in the medication entity, which is also the one with fewer training examples. Finally, abbreviations are indeed the most manageable entities to learn, with an F1-score of 0.92.

5 Conclusions and future work

There is a lack of language resources for the Spanish language in the clinical domain, and the work presented here constitutes the first annotated corpus using Chilean narratives. We believe that projects like ours help filling the gap with respect to clinical NLP done in English. This paper shares 900 annotated referrals, the annotation guidelines, clinical word embeddings, and the code used to generate the results presented here.

In terms of developing NER models trained on this corpus, future work includes improving the recall for disease and medication, due to the importance of identifying these entities, and dealing with nested NER, for which there is a variety

Entity	# of test entities	Pre-trained embbedings			Pre-trained + clinical embeddings		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
Abbreviations	457	0.85 (0.012)	0.91 (0.014)	0.88 (0.004)	0.86 (0.01)	0.92 (0.013)	0.89 (0.005)
Disease	403	0.73 (0.034)	0.65 (0.023)	0.69 (0.008)	0.75 (0.013)	0.71 (0.015)	0.73 (0.01)
Medication	44	0.68 (0.05)	0.59 (0.032)	0.63 (0.021)	0.74 (0.047)	0.72 (0.036)	0.73 (0.036)

Table 2: Multiclass model where nested entities are lost. Data shown are mean (SD).

Entity	# of test entities	Pre-trained embbedings			Pre-trained + clinical embeddings		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
Abbreviations	507	0.92 (0.004)	0.92 (0.007)	0.92 (0.004)	0.91 (0.002)	0.93 (0.003)	0.92 (0.002)
Disease	456	0.76 (0.008)	0.65 (0.009)	0.70 (0.004)	0.79 (0.004)	0.75 (0.009)	0.77 (0.005)
Medication	53	0.71 (0.02)	0.50 (0.032)	0.58 (0.026)	0.79 (0.038)	0.71 (0.016)	0.75 (0.024)

Table 3: Multiple models for each entity. All entities are retained. Data shown are mean (SD).

[5] https://github.com/zalandoresearch/flair

[6] http://doi.org/10.5281/zenodo.3924799

of approaches as summarized in (Dadas and Protasiewicz, 2020). Besides, our annotated corpus has hierarchical entities (for example, test result and sign/symptom are part of the entity finding). We plan to investigate the hierarchical nested NER using architectures as in Marinho *et al.*(Marinho et al., 2019). 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.

One of our goals working on this corpus and training NER models is to recognize diseases within this waiting list automatically. In particular, telemedicine has been posed as one of the solutions to decrease the waiting times in the Chilean public healthcare sector (Ministerio de Salud de Chile, 2011a). To correctly estimate the effect, one needs to summarize the suspected diagnoses and check which of them are eligible for telemedicine consultations. Furthermore, diseases that need to be examined rapidly could be prioritized using an automatic detection of diseases.

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.

Acknowledgements

This work was funded by CMM-ANID AFB 170001. Also, PB receives support from U-INICIA VID 2019 UI-004/19 and ICM P09-015F, and JD from the cost center 570111 - CIMT-CORFO. We want to thank Leonardo Campillos-Llanos for his constructive comments on the manuscript and overall support. We also acknowledge the help received from Martin Krallinger, Aureliè Neveol, Felipe Bravo, and Karen Fört. Finally, we are very grateful to reviewers for their helpful comments.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

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

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*, pages 152–164.

Nancy Chinchor and Patricia Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, volume 29, pages 1–21.

Viviana Cotik, Darío Filippo, Roland Roller, Hans Uszkoreit, and Feiyu Xu. 2017. Annotation of entities and relations in spanish radiology reports. In *RANLP*, pages 177–184.

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*, pages 53–58.

Sławomir Dadas and Jarosław Protasiewicz. 2020. A bidirectional iterative algorithm for nested named entity recognition. *IEEE Access*, 8:135091–135102.

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

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).

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*, pages 141–150.

Fondo Nacional de Salud. 2013. Población Inscrita en FONASA, https://public.tableau.com/views/Poblacion2002-2020/INEeInscritos. Technical report.

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

Lenz Furrer, Joseph Cornelius, and Fabio Rinaldi. 2019. Uzh@ craft-st: a sequence-labeling approach to concept recognition. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 185–195.

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):296–298.

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*.

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)*, pages 3576–3581.

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*, pages 43–50.

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):1–17.

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*, 2421:671–678.

Nastassja A Lewinski, Ivan Jimenez, and Bridget T McInnes. 2017. An annotated corpus with nanomedicine and pharmacokinetic parameters. *International journal of nanomedicine*, 12:7519.

Salvador Lima, Naiara Perez, Montse Cuadros, and German Rigau. 2020. Nubes: A corpus of negation and uncertainty in spanish clinical texts. *arXiv preprint arXiv:2004.01092*.

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):24–29.

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.*

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*, pages 28–34.

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):233.

Ministerio de Salud de Chile. 2004. Ley 19.966, `https://www.leychile.cl/navegar?idnorma=229834`.

Ministerio de Salud de Chile. 2011a. Estrategia Nacional de Salud para el cumplimiento de los Objetivos Sanitarios de la Década 2010-2020.

Ministerio de Salud de Chile. 2011b. Norma Técnica Para El Registro De Las Listas De Espera, `www.supersalud.gob.cl/664/w3-propertyvalue-6249.html`.

Ministerio Secretaría General de la Presidencia. 2008. Ley 20.285, `https://www.leychile.cl/navegar?idnorma=276363&idparte=`.

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*, 95:33–39.

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.

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):12.

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):310–318.

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*, pages 536–543. Springer.

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

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*, volume 2007, page 625. American Medical Informatics Association.

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):950–966.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

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

Jana Straková, Milan Straka, and Jan Hajič. 2019. Neural architectures for nested ner through linearization. *arXiv preprint arXiv:1908.06926*.

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):14–24.

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

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

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

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):1229–1238.