CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Representations for Learning and Language

Centro de Modelamiento Matemático

Nested named entity recognition in diagnoses from the Chilean Waiting List in public hospitals

Author: Matías Rojas
Supervisors: Felipe Bravo and Jocelyn Dunstan

Outline.

01.

Context

A brief introduction to the Waiting List system.

02.

Chilean Waiting List corpus

Description of the dataset used in our study.

03.

Problem

Major challenges to be solved according to the nested NER task.

04.

Objectives

Description of the thesis objectives.

05.

Nested NER in referrals

Experiments carried out using the MLC approach in our corpus.

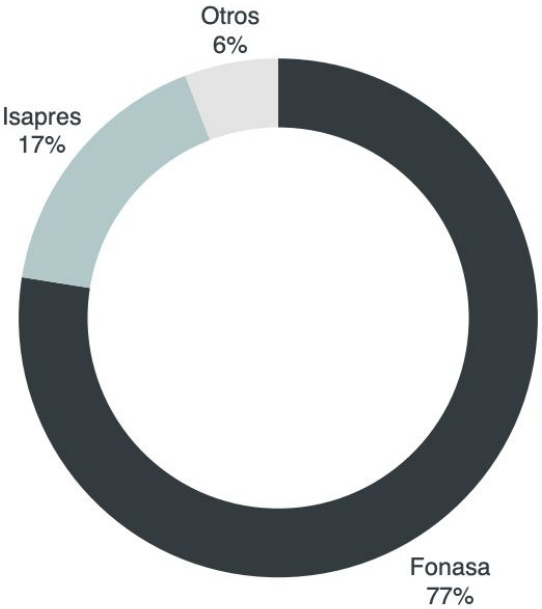
06.

Empirical study

Empirical study of the nested NER task.

The Waiting List in Chilean public hospitals

- 77% of the Chilean population are in the public healthcare system [1].
- To see an specialist you have to go first to primary care physician.
- He/she puts you in a Waiting List (WL) for specialty consultation.



High demand problem

- Currently, the average waiting time is 543 days.
- In 2020-01, 15,665 patients died while waiting for their first consultation.
- In 2021, 1,965,653 patients are waiting for their first consultation [2].

Can we improve the management of the Chilean Waiting List using NLP?

Can we have a secondary use of the information?

National registry of Waiting List

Sex	Age	Specialty	Reason for referral

Written in free-text

Named Entity Recognition (NER)

NER is an important task in NLP that seeks to identify sequences of words (entities) expressing references to predefined categories (entity types).

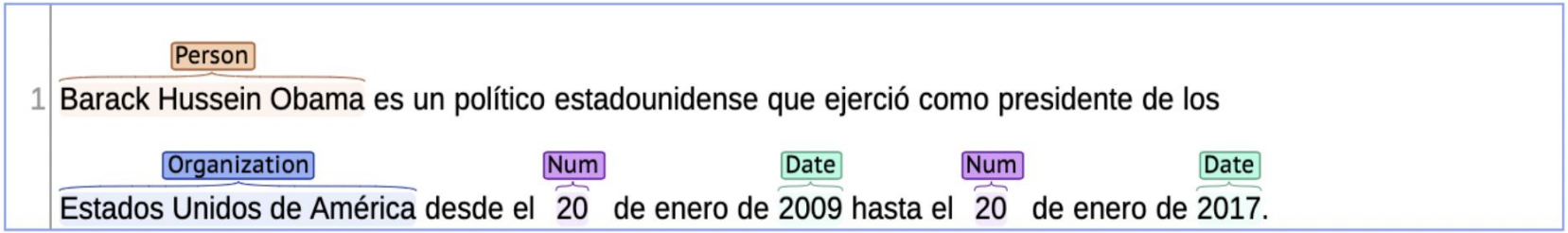


Figure 1: Example of named entities extracted using the Stanford NER system [3].

NER in the Chilean Waiting List

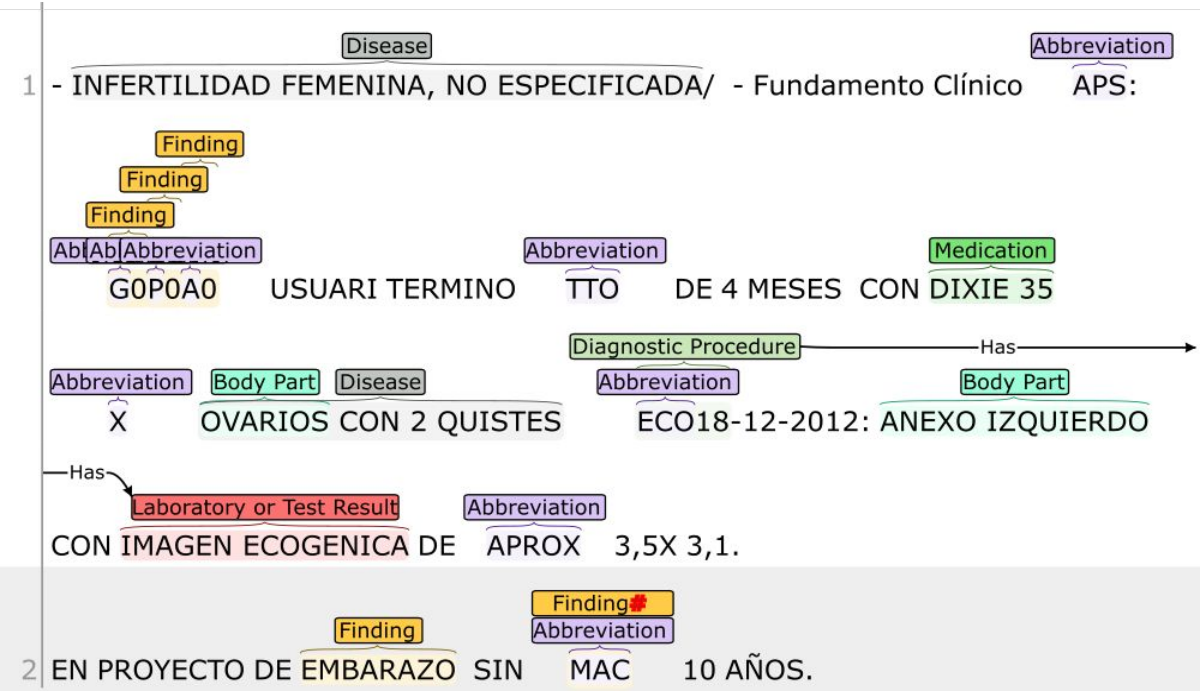
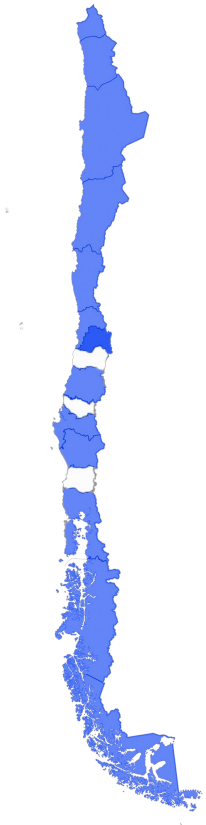


Figure 2: An example of an annotation in the Chilean Waiting List corpus.

Chilean Transparency Law



5 million referrals

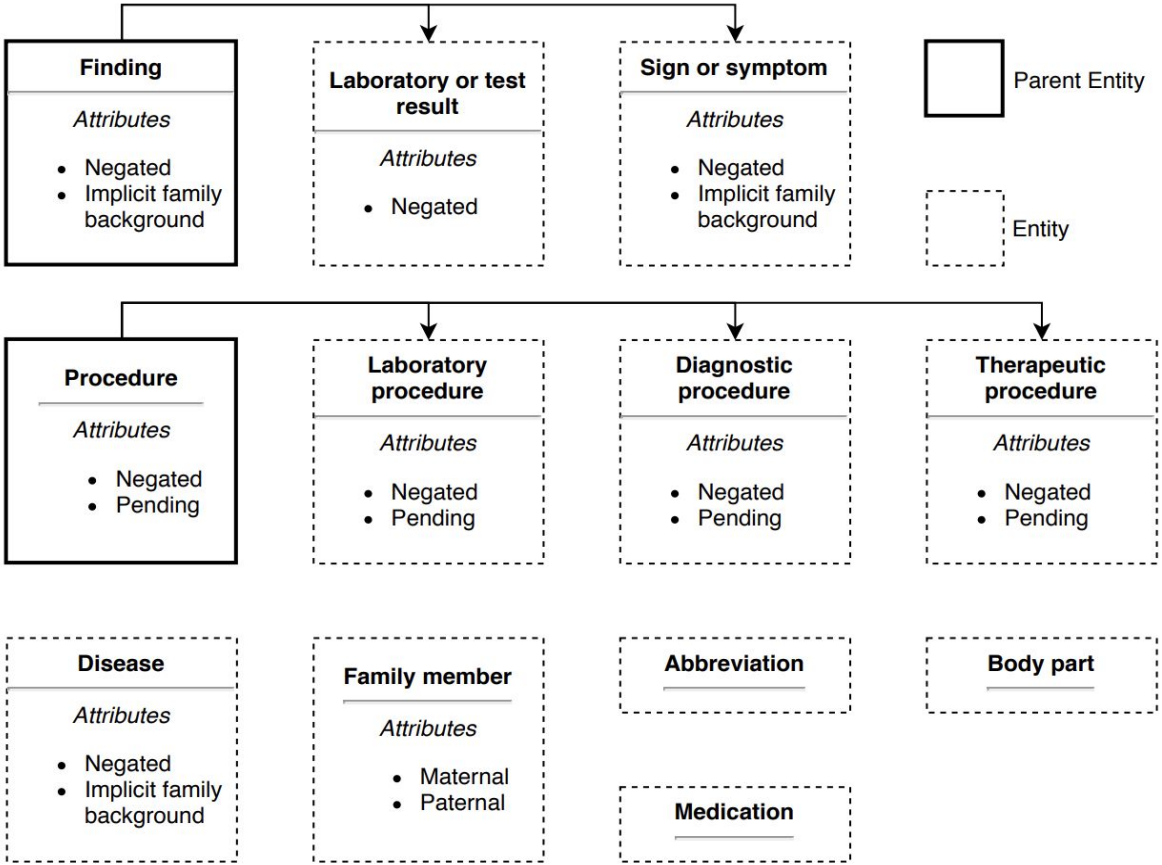


Figure 3: List of entity types (in bold) in the Chilean Waiting List [4].

Problem statement

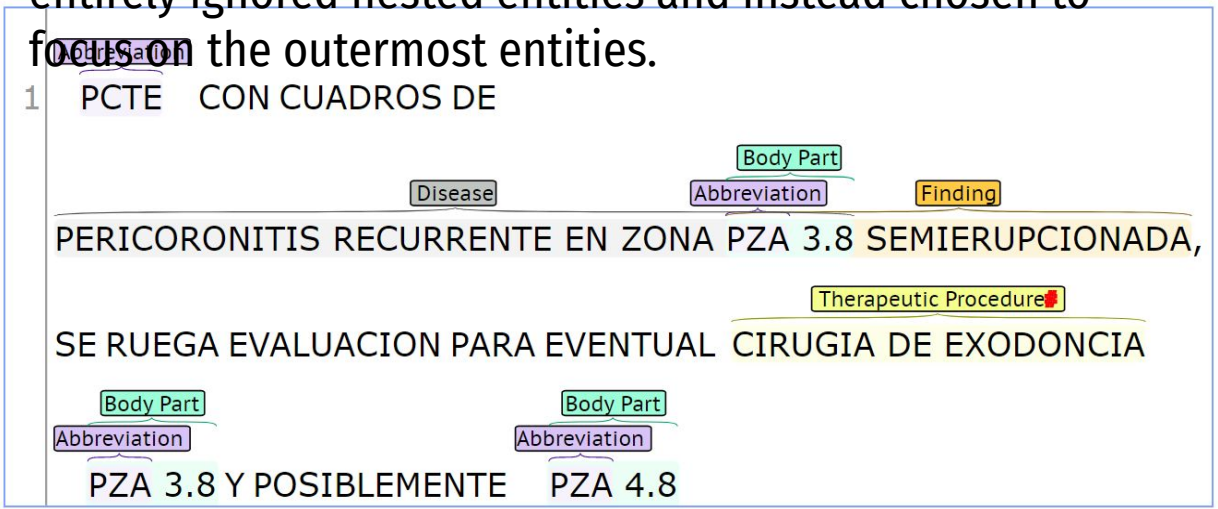
Flat NER

Traditional NER approach, where each word can be tagged with at most one label.

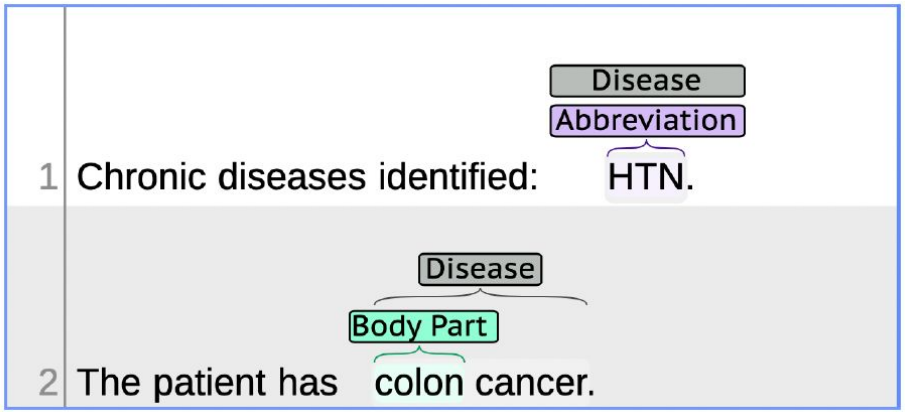
Nested NER

Due to the properties of natural language, named entities can be nested in other entities. Under this approach, a word could be tagged with multiple entities.

Most of the work on named entity recognition has almost entirely ignored nested entities and instead chosen to focus on the outermost entities.



The Chilean Waiting List corpus



- 48.12% of the entities are nested in other entities.

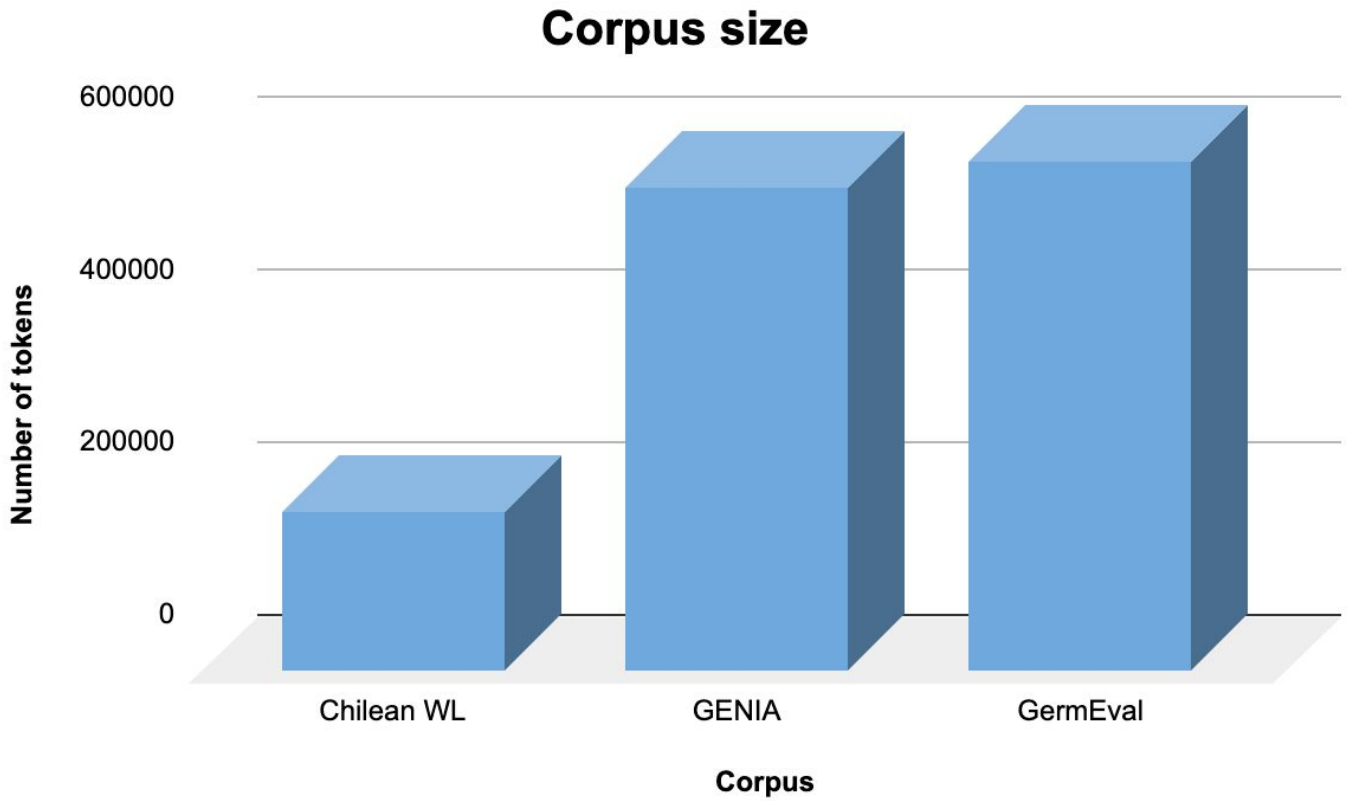


Figure 4: Comparison of corpus size in nested NER corpora.

An overlooked architecture for nested NER

Most of the models used to solve flat NER tasks are based on deep learning architectures such as LSTM-CRF approach, which belongs to the *sequence labeling* category.

However, little research has been conducted on adapting this architecture to the nested NER task by using independent flat NER models for each entity type.

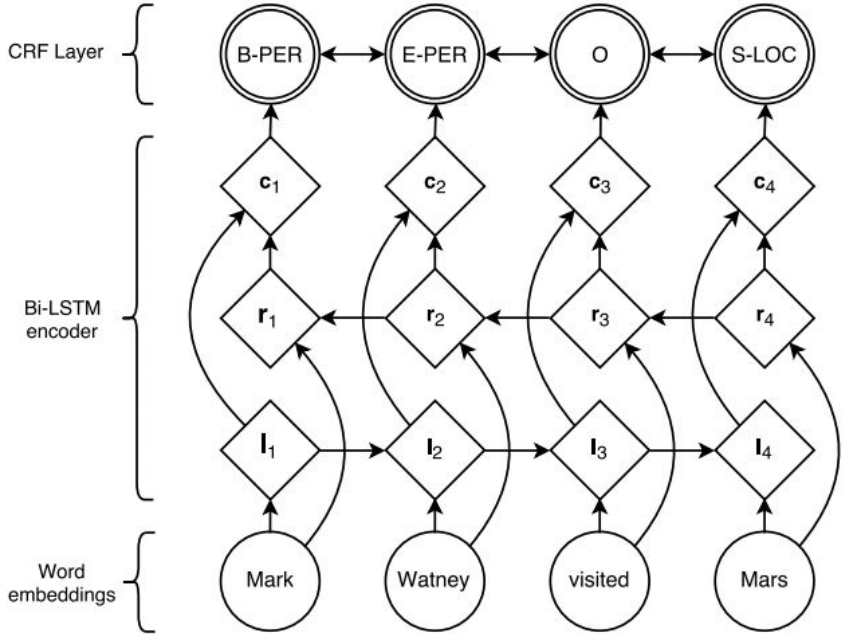


Figure 5: LSTM-CRF architecture [5].

That said, it is not clear whether we can achieve good performance on recognizing nested entities in our corpus by making simple modifications to sequence labeling-based architectures.

Objectives

The main objectives of our research are the following:

1. Propose and develop deep neural architectures for solving the nested NER task in the Chilean Waiting List corpus.
2. Provide an empirical study comparing the proposed models with other state-of-the-art architectures in the nested NER task and testing these models on other related corpora to validate their effectiveness.
3. Propose new task-specific evaluation metrics that adequately measure the model's performance on nesting.
4. Integrate the proposed models in a test environment allowing health professionals to test them.

Multiple LSTM-CRF (MLC)

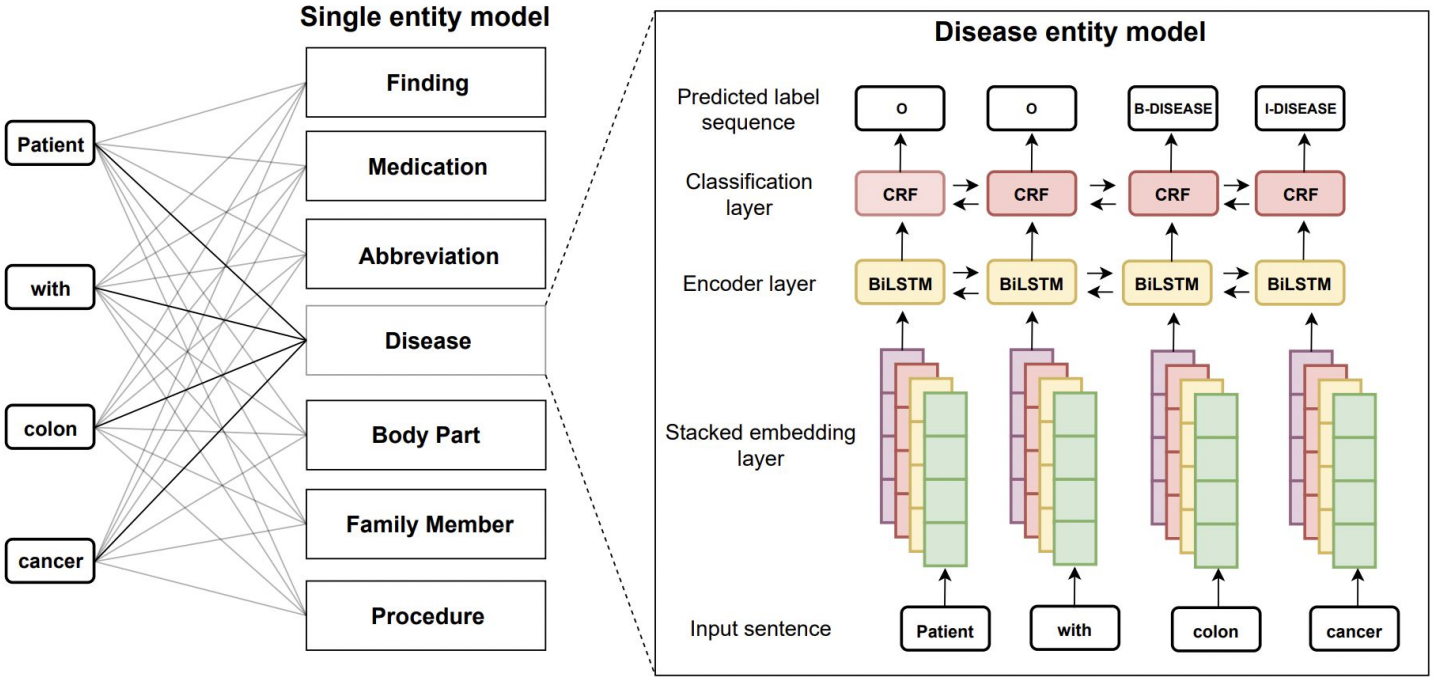


Figure 6: Overview of the MLC architecture, where each entity type has an associated flat NER model [4].

Embedding Layer

- Pre-trained word embeddings in clinical domain[1].
- Character-level embeddings retrieved from a BiLSTM [5].
- Contextual word embeddings obtained from Flair [6] and BERT [7].

[1] https://zenodo.org/record/3924799

Evaluation Metric

The official nested NER metric [8] consists of calculating the micro F1-score using a strict evaluation approach. This metric considers an entity correct when both entity types and boundaries are predicted correctly.

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F1 = \frac{2 \cdot P \cdot R}{P + R}$$

Model validation and hyperparameters

- As a baseline, we chose the Layered Architecture [9].
- To select the best hyperparameters, we performed a random search over a given range of values, measuring the performance on the validation set.
- We used the statistical test: *k-fold cross-validated paired t-test* [10].
- Finally, an error analysis of the MLC model was performed [11].

Overall results

Model	Precision	Recall	F1-score
Neural Layered Model [48] (baseline)	77.0	72.12	74.48
SML	76.6	72.7	74.60
MLC [Word]	76.59	74.84	75.71
MLC [Word+Char]	77.75	78.29	78.02
MLC [Word+Char+BERT]	79.72	78.83	79.27
MLC [Word+Char+Flair]	**80.24**	**80.30**	**80.27**
MLC [Word+Char+Flair+BERT]	79.90	78.13	79.01

Table 1: Results obtained with different models and settings on the Chilean Waiting List corpus.

- Interestingly, each of the proposed models (MLC and SML) outperforms the layered baseline by a wide margin according to the F1 measure.

- The model with the best performance (MLC) was achieved in the MLC setting trained on the word embeddings concatenated with character and Flair embeddings, achieving a micro F1-score of 80.27.

MLC results for each entity type

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 2: Results for each entity type using the best MLC setting in the test subset.

- The entity type with the best results was Abbreviation, which is not surprising since it is usually easy to recognize from the morphological point of view.
- The entity type with the worst results was Finding, which is also expected since it is usually one of the most challenging ones.

Statistical Test Results

	Neural Layered Model [48] (baseline)	MLC [Word + Char + Flair]	P value
Mean	73.20	79.81	$8.8e^{-9}$
SD	0.752	0.469	
Min	72.16	79.16	
Max	74.65	80.66	

Table 3: Results of the 10-fold cross-validation on the best MLC setting and the baseline.

- The cross-validation process demonstrated the efficacy and high level of generalization of the MLC model on unseen data, significantly outperforming the baseline in all measurements, consistent with the overall results.

Error Analysis

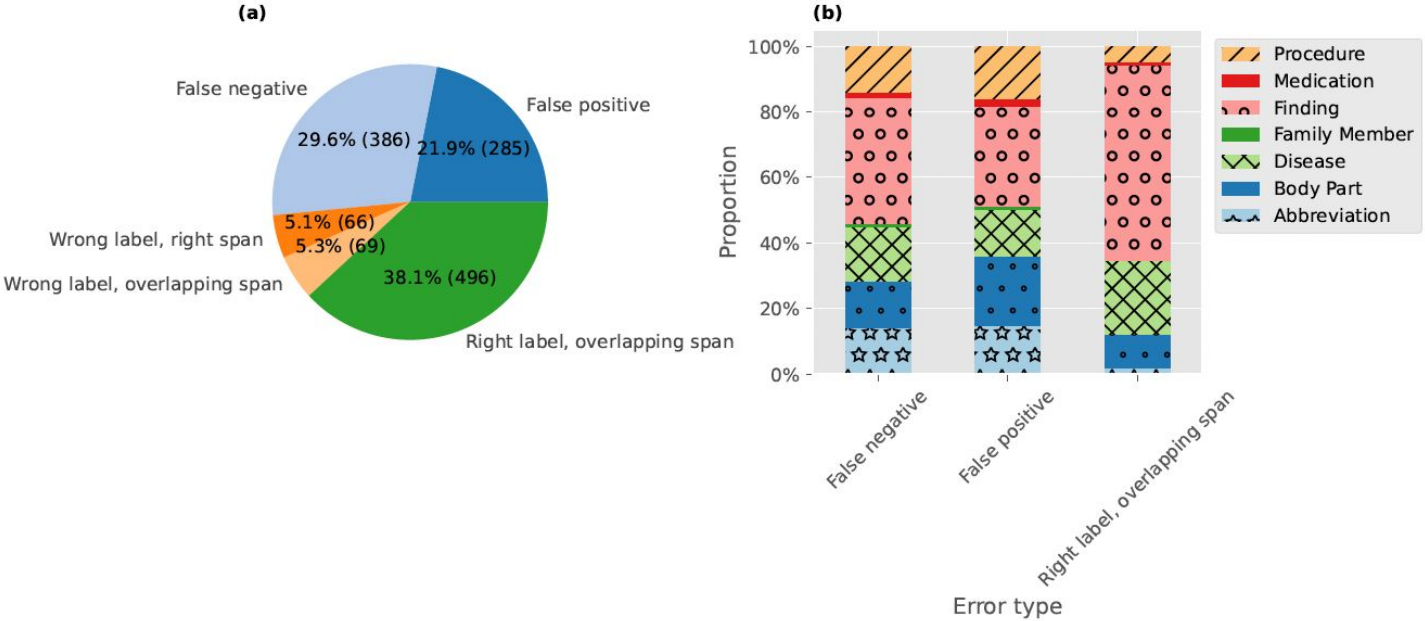


Figure 7: Distribution of the errors types found by the error analysis.

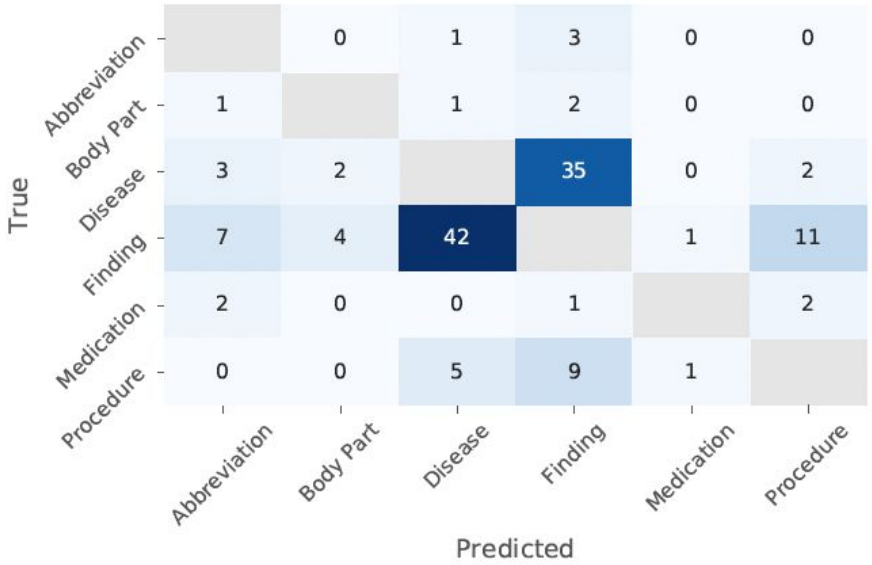


Figure 8: Confusion matrix for the wrong label errors found by the error analysis.

- Given the promising results, we would be interested to know whether the MLC architecture is the most suitable approach for solving the nested NER in our corpus or there are models with better performance.

- Additionally, we wonder if this model can obtain good results on other nested NER corpora from different domains and languages.

- On the other hand, we wonder if we are correctly measuring model performance using the standard evaluation metric in nested NER.

Related corpora

- GENIA [12]: Biomedical corpus collected from 2,000 MEDLINE abstracts. It comprises five entity types and 55,740 entity mentions, of which 17.3% are involved in nesting.
- GermEval [13]: Corpus sampled from German Wikipedia and online news. This dataset consists of 41,124 entity mentions, where 14.9% of them are involved in nesting.

	GENIA			GermEval			Chilean Waiting List		
	Train	Test	Dev	Train	Test	Dev	Train	Test	Dev
tokens	454,882	57,021	48,932	452,853	96,499	41,653	149,574	18,436	16,754
sentences	15,023	1,854	1,669	24,000	5,100	2,200	8,014	990	890
avg sent len	30.3	30.8	29.3	18.9	18.9	18.9	18.7	18.6	18.8
entities	45,929	5,474	4,337	31,545	6,693	2,886	35,480	4,289	3,971
avg entity len	2.9	2.9	3.1	1.4	1.4	1.5	2.6	2.7	2.6
nested entities (%)	17.0	20.6	16.8	15.0	14.7	14.1	46.4	45.9	46.7
nested entities	7,795	1,130	727	4,721	986	407	16,456	1,969	1,856
- different type	3,712	589	369	4,230	892	366	12,635	1,555	1,398
- same type	4,132	547	358	536	93	44	0	0	0
- multi-label entities	0	0	0	2	2	0	4,241	470	502

Table 4: Statistics of the datasets involved in our study.

Baselines

- ➔ Layered [Ju et al., 2018].
- ➔ Boundary [Zheng et al., 2019]
- ➔ Exhaustive [Sohrab and Miwa, 2018].
- ➔ Recursive-CRF [Shibuya and Hovy, 2020].
- ➔ Pyramid [Wang et al., 2020].
- ➔ Biaffine [Yu et al., 2020].

Overall results using the standard metric

	GENIA			GermEval			Chilean Waiting List		
Model	P	R	F1	P	R	F1	P	R	F1
Layered	73.9	68.7	71.2	71.8	64.1	67.7	75.0	72.8	73.9
Exhaustive	74.1	69.7	71.8	78.6	64.6	70.9	76.3	71.7	68.2
Boundary	76.7	71.8	74.2	74.4	65.5	69.7	74.0	67.6	70.7
Pyramid	78.1	72.8	75.3	77.8	66.9	71.9	79.6	75.4	77.5
Biaffine	79.1	73.7	76.3	89.0	77.4	82.8	81.5	67.1	73.6
Recursive-CRF	75.8	75.2	75.5	85.1	78.2	81.5	75.1	77.2	76.1
MLC	77.6	74.2	75.8	86.8	77.2	81.7	77.7	78.3	78.0
LM-based									
Biaffine [BERT]	79.9	76.5	**78.1**	88.3	85.0	86.6	78.7	70.8	74.5
Recursive-CRF									
- Flair	77.1	78.0	77.6	83.4	82.9	83.2	78.0	79.9	78.9
- BERT	76.4	77.4	76.9	84.3	83.0	83.6	76.6	77.8	77.2
- Flair + BERT	77.4	76.8	77.1	84.8	82.1	83.4	77.1	77.9	77.5
Pyramid									
- Flair	77.8	75.6	76.7	83.4	80.0	81.7	80.1	77.2	78.6
- BERT	79.1	76.9	78.0	87.7	85.8	**86.7**	78.0	73.6	75.7
- Flair + BERT	80.4	75.0	77.6	87.7	84.4	86.0	78.5	77.2	77.9
MLC									
- Flair	80.1	75.2	77.6	85.3	82.4	83.8	80.6	80.5	**80.5**
- BERT	79.4	74.3	76.8	85.1	80.3	82.6	79.7	78.8	79.3
- Flair + BERT	78.8	75.2	75.5	84.7	80.1	82.3	79.9	78.1	79.0

Table 5: Overall results on three nested NER corpora, including ours.

Proposed evaluation metrics

Revisited metrics.

- m_{flat}
- m_{nested}
- m_{inner}
- m_{outer}

However, none of these existing metrics capture the ability of the models to recognize both inner and outer entities simultaneously

Proposed metrics.

- $m_{nesting}$
- m_{NST}
- m_{NDT}
- m_{ME}

Results using the nesting metrics

GENIA				
Model	m_{flat}	m_{nested}	m_{inner}	m_{outer}
Layered	73.2	62.3	42.9	79.8
Exhaustive	76.6	55.0	42.6	67.9
Boundary	77.4	59.5	42.0	75.6
Biaffine [BERT]	81.2	**65.8**	**49.3**	80.5
Pyramid [BERT]	81.1	65.2	46.1	**82.4**
Recursive-CRF [Flair]	**81.5**	62.3	46.9	77.4
MLC [Flair]	80.7	63.8	41.7	82.2

GermEval				
Model	m_{flat}	m_{nested}	m_{inner}	m_{outer}
Layered	68.8	60.9	62.0	59.7
Exhaustive	73.4	56.1	65.7	45.7
Boundary	70.9	54.5	54.1	55.0
Biaffine [BERT]	88.4	76.6	**78.1**	75.0
Pyramid [BERT]	**88.5**	**76.7**	77.3	**76.1**
Recursive-CRF [BERT]	85.5	73.0	74.9	71.0
MLC [Flair]	86.0	71.6	74.5	68.4

Chilean Waiting List				
Model	m_{flat}	m_{nested}	m_{inner}	m_{outer}
Layered	73.4	74.5	82.4	64.5
Exhaustive	71.7	63.8	71.5	53.4
Boundary	73.4	61.1	65.5	55.4
Biaffine [BERT]	76.2	72.5	75.2	69.2
Pyramid [Flair]	79.0	78.1	84.7	69.3
Recursive-CRF [Flair]	80.3	77.4	82.8	70.4
MLC [Flair]	**80.9**	**80.1**	**86.2**	**72.5**

Table 6: Results on nested and non-nested entities.

GENIA				
Model	$m_{nesting}$	m_{ME}	m_{NDT}	m_{NST}
Layered	26.2	-	41.7	9.7
Exhaustive	25.8	-	41.2	17.7
Boundary	26.6	-	40.5	17.8
Biaffine [BERT]	**34.5**	-	**51.9**	**22.9**
Pyramid [BERT]	33.4	-	49.5	20.9
Recursive-CRF [Flair]	31.5	-	49.1	19.4
MLC [Flair]	27.9	-	47.8	0

GermEval				
Model	$m_{nesting}$	m_{ME}	m_{NDT}	m_{NST}
Layered	37.3	-	40.4	16.2
Exhaustive	27.8	-	38.2	9.7
Boundary	21.2	-	25.5	7.8
Biaffine [BERT]	55.7	-	**64.3**	20.8
Pyramid [BERT]	**56.5**	-	63.8	21.4
Recursive-CRF [BERT]	51.1	-	58.9	**23.9**
MLC [Flair]	49.1	-	59.3	0

Chilean Waiting List				
Model	$m_{nesting}$	m_{ME}	m_{NDT}	m_{NST}
Layered	51.6	71.1	49.5	-
Exhaustive	28.4	0	41.7	-
Boundary	28.2	0	35.4	-
Biaffine [BERT]	41.8	0	55.1	-
Pyramid [Flair]	54.9	**73.7**	57.9	-
Recursive-CRF [Flair]	56.0	71.7	58.8	-
MLC [Flair]	**60.6**	72.5	**60.0**	-

Table 7: Our task-specific metrics.

Detección Automática de Entidades Médicas en Textos Clínicos

Este es un modelo predictivo en fase de desarrollo. Las respuestas retornadas por el modelo no deben ser utilizadas para la toma de decisiones.

Texto a anotar

HTA DM CA COLON OPERADO ANEMIA TROMBOSIS HPB MARCAPASOS ULTIMO CONTROL DE TELEMETRIA ABRIL15 HISTOGRAMA SIN EVENTOS MCP CON BUEN SENSADO Y CAPTURA, TVP VENA AXILAR IZQUIERDA EN TACO LE DETECTARON GLAUCOMA EN TTO

Anotar

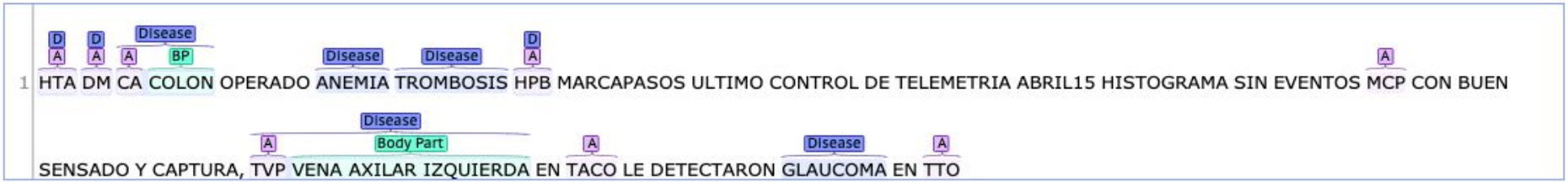


Figure 9: Web application created by the research group to test our model.

Conclusions

- Extensive experiments with three nested NER corpora show that, regardless of the simplicity of the MLC model, its performance is better or at least as well as more sophisticated methods.

- We demonstrated that standard NER metrics do not measure well the ability of a model to detect nested entities, while our task-specific metrics provide new evidence on how existing approaches handle the task.

- The results obtained suggest that the MLC architecture is the model that best suits the nested NER task in our corpus. This model can be used for many studies to understand the high demand present in the Waiting List system.

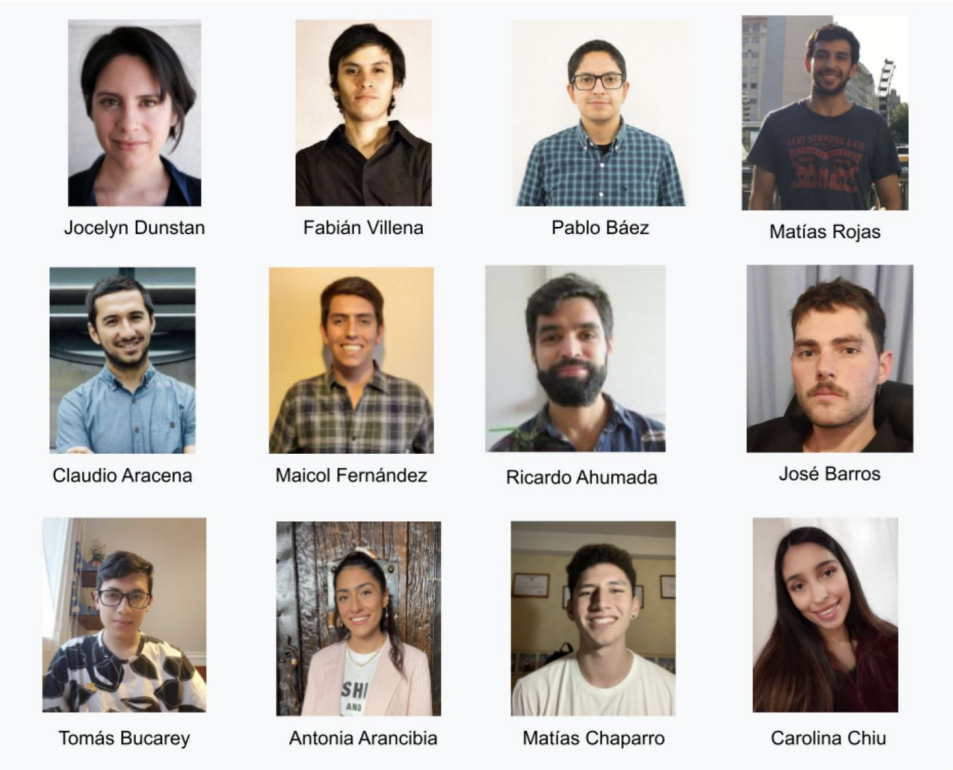
Future Work

- Hosting a shared task using the Chilean Waiting List corpus.
- Improve the MLC model to identify nested entities of the same type.
- Improve the web interface of the prototype under development.

Contributions

- The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish [19].
- Automatic Extraction of Nested Entities in Clinical Referrals in Spanish [4].
- Simple yet Powerful: An Overlooked Architecture for Nested Named Entity Recognition.

Natural Language Processing Group at Center for Mathematical Modeling - University of Chile.



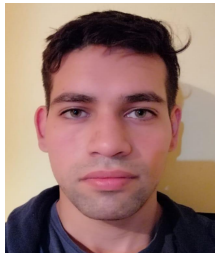
http://pln.cmm.uchile.cl

Thanks for your attention! 🌳

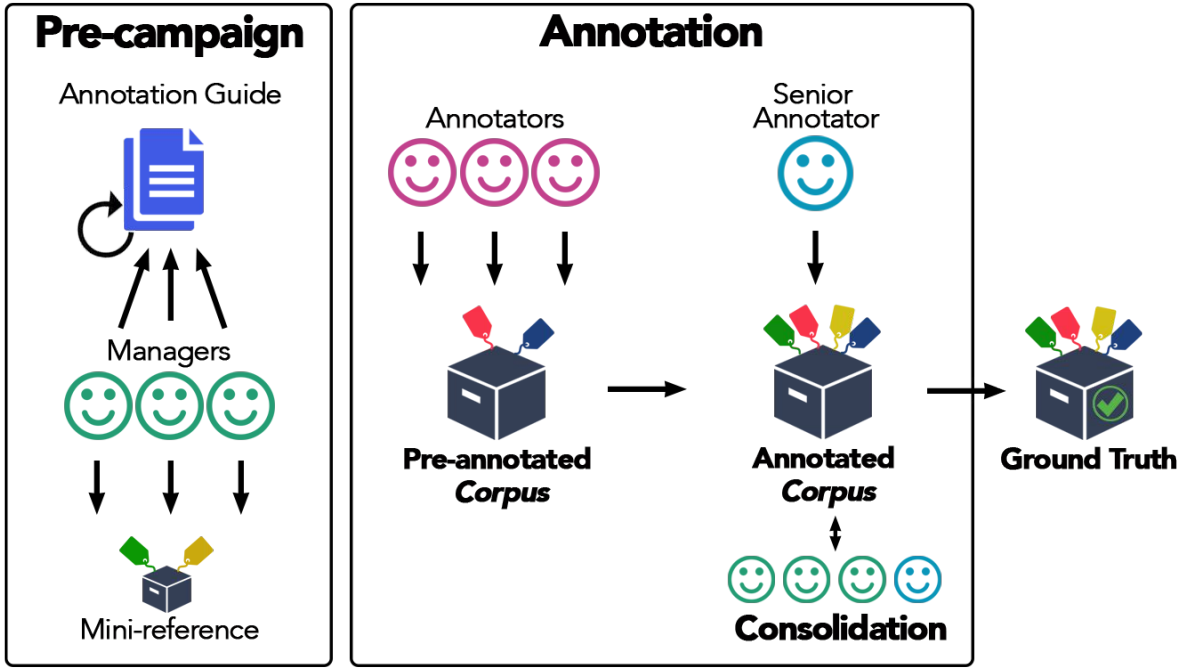
Pablo Báez, PhD

Maicol Fernández, DDS

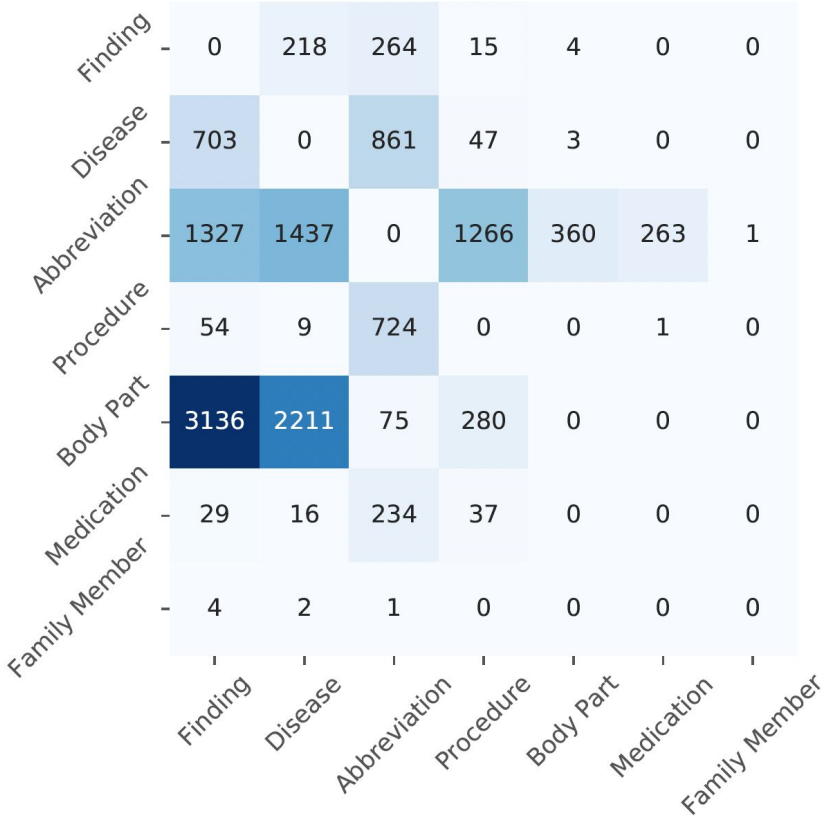
Fabián Villena, DDS



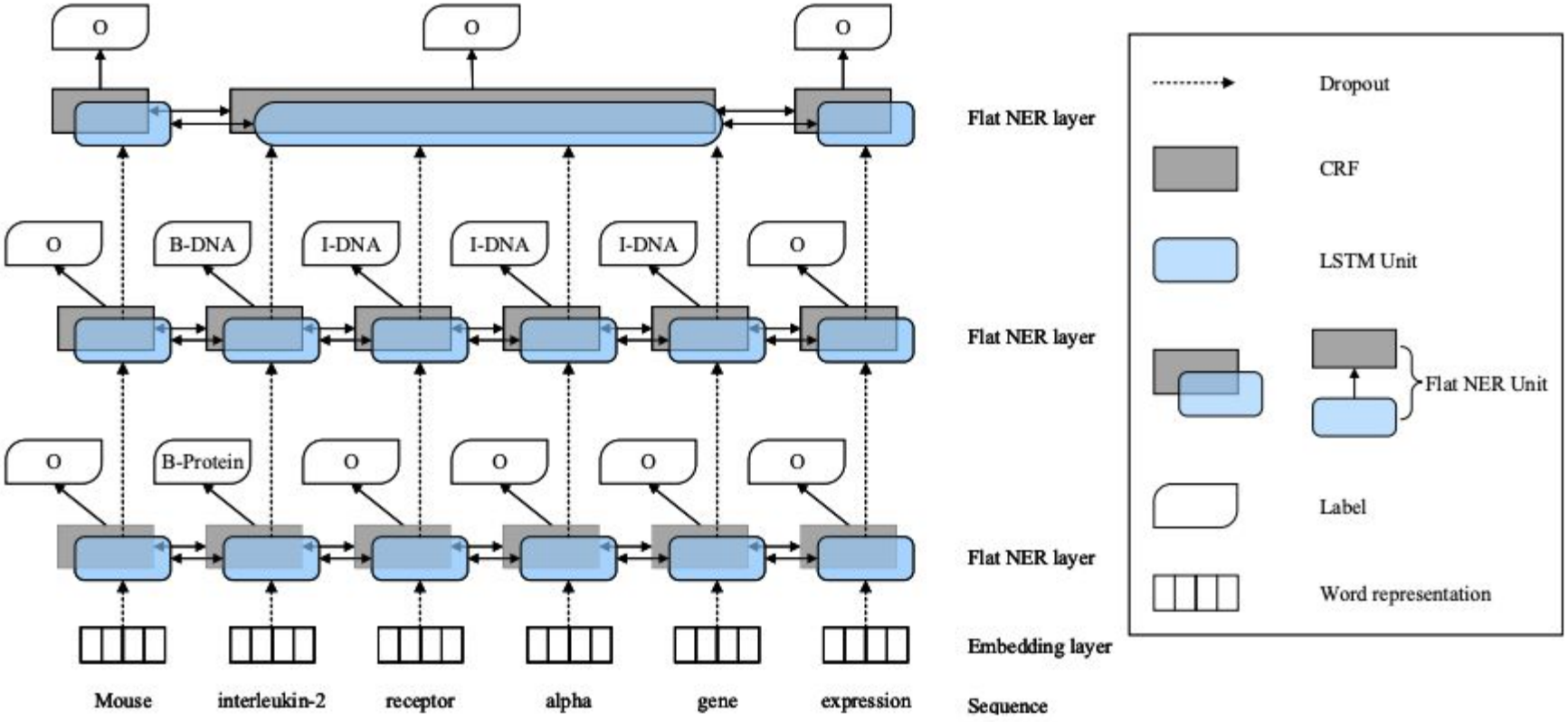
Manuel Durán, MD



Annotation stages for the creation of annotation guidelines [3].



Characterization of nested entities. The numbers in each cell indicate how many times the entity in the row is nested in the entity in the column.


O
O
O
O
B-DNA
I-DNA
I-DNA
I-DNA
O
O
B-Protein
O
O
O
O
Mouse
interleukin-2
receptor
alpha
gene
expression
Flat NER layer
Flat NER layer
Flat NER layer
Embedding layer
Sequence
Dropout
CRF
LSTM Unit
Flat NER Unit
Label
Word representation

Task Formalization

Definition 2 (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 nested in R if $S_r \leq S_q$ and $E_q \leq E_r$. The particular case of $S_q = S_r$ and $E_q = E_r$ corresponds to an entity with multiple labels. Note that under this definition we consider the three types of nesting described above.

Definition 3 (Nested NER) Given an input sequence $X = \{x_1, x_2, ..., x_n\}$, nested NER aims 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 made for cases where nested entities are involved and when not.

References I

[1] FONASA. Cuenta pública 2020.

[2] Minsal. Glosa 06: Lista de Espera No Ges y Garantías de Oportunidad GES Retrasadas.

[3] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling.

[4] Pablo Báez, Felipe Bravo-Marquez, Jocelyn Dunstan, Matías Rojas, and Fabián Villena. Automatic Extraction of Nested Entities in Clinical Referrals in Spanish.

[5] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition.

[6] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding.

[8] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition.

[9] Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. Pyramid: A layered model for nested named entity recognition.

References II

[10] Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms.

[11] Isar Nejadgholi, Kathleen C. Fraser, and Berry de Bruijn. Extensive error analysis and a learning-based evaluation of medical entity recognition systems to approximate user experience.

[12] Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. Genia corpus—a semantically annotated corpus for bio-textmining.

[13] Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pado. Germeval 2014 named entity recognition shared task.

[14] Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung, and Guandong Xu. A boundary-aware neural model for nested named entity recognition.

[15] Mohammad Golam Sohrab and Makoto Miwa. Deep exhaustive model for nested named entity recognition.

[16] Takashi Shibuya and Eduard Hovy. Nested named entity recognition via second-best sequence learning and decoding.

[17] Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. Pyramid: A layered model for nested named entity recognition.

[18] Juntao Yu, Bernd Bohnet, and Massimo Poesio. Named entity recognition as dependency parsing.

[19]] Pablo Baez, Fabian Villena, Matıas Rojas, Manuel Duran, and Jocelyn Dunstan. The Chilean waiting list corpus: a new resource for clinical named entity recognition in Spanish.