

# Extracting Neoplasms Morphology Mentions in Spanish Clinical Cases through Word Embeddings

Pilar López-Úbeda<sup>a</sup>, Manuel C. Díaz-Galiano<sup>a</sup>, M. Teresa Martín-Valdivia<sup>a</sup> and L. Alfonso Ureña-López<sup>a</sup>

<sup>a</sup>SINAI research group, University of Jaén, Spain

## Abstract

Biomedicine is an ideal environment for the use of Natural Language Processing (NLP), due to the huge amount of information processed and stored in electronic format. This information can be managed in different ways by NLP tasks such as the Named Entity Recognition (NER). To address this task, CANTEMIST is the first challenge specifically focusing on NER and named entity normalization of a critical type of concept related to cancer. For the entity standardization, the challenge proposes to use the ICD-O codes (International Classification of Diseases for Oncology, 3rd Edition – ICD-O-3).

In this paper, we present an automated system based on neural networks for the extraction of tumor morphology mentions in Spanish clinical cases. In particular, we use a Bidirectional variant of Long Short Term Memory (BiLSTM) neural network with a Conditional Random Fields (CRF) layer. The input to this network is a combination of different word embeddings. In the NER task we achieved encouraging results obtaining 85.5% of F1-score. Moreover, a dictionary-based system is used to subsequently assign an ICD-O code to each annotated entity. For this subtask our group achieved 75.9% of F1-score.

## Keywords

Named Entity Recognition, named entity normalization, tumor morphology, word embeddings, neural networks

## 1. Introduction

Named Entity Recognition (NER) is the task of identifying and categorizing key entities in the text. The NER task is part of Natural Language Processing (NLP) and text mining that studies natural language. The NER task makes it possible to generate knowledge and improve health services by processing unstructured information. This information is an important part of the data collected on a daily basis in clinical practice. NER uses health-related information and supports the new challenges of recording, structuring, and exploring information.

Understanding diseases requires the extraction of certain key entities like diseases, treatments or symptoms and their attributes from textual data. The recognition of these entities in different languages is a hard task that needs to be addressed by the NLP community. To accomplish this task, the CANTEMIST [1] challenge at the IberLEF 2020 (Iberian Languages Evaluation Forum) is proposed.

---


*Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*

EMAIL: plubeda@ujaen.es (P. López-Úbeda); mcdiaz@ujaen.es (M.C. Díaz-Galiano); maite@ujaen.es (M.T. Martín-Valdivia); laurena@ujaen.es (L.A. Ureña-López)

ORCID: 0000-0003-0478-743X (P. López-Úbeda); 0000-0001-9298-1376 (M.C. Díaz-Galiano); 0000-0002-2874-0401 (M.T. Martín-Valdivia); 0000-0001-7540-4059 (L.A. Ureña-López)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

CANTEMIST is the first shared task specifically focusing on named entity recognition of a critical type of concept related to cancer, namely tumor morphology. The CANTEMIST task is structured into three independent subtasks, each taking into account a particular important use case scenario:

- CANTEMIST-NER track: This subtask consists of automatically finding tumor morphology mentions.
- CANTEMIST-NORM track: The second subtask requires returning all tumor morphology entity mentions together with their corresponding ICD-O codes (Spanish version: eCIE-O-3.1), i.e. finding and normalizing tumor morphology mentions. This subtask is also known as clinical concept normalization or named entity normalization.
- CANTEMIST-CODING track: The last subtask requires returning for each of document a ranked list of its corresponding ICD-O-3 codes.

This paper describes the system presented by the SINAI team for the first two subtasks proposed by the organizers of the CANTEMIST challenge: NER track and named entity normalization track.

## 2. Related work

In the biomedical domain, NER systems identify entities from clinical patient reports. Several NER systems have been developed using NLP-based systems such as MedLEE [2], MetaMap [3] and cTAKES [4]). Most of these are rule-based systems that use extensive medical vocabularies.

Approaches for NER can be classified into different categories [5]: dictionary-based, rule-based and machine learning-based. Specifically, our group has experience in the NER task in the biomedical domain using different methodologies such as traditional machine learning [6], Recurrent Neural Networks (RNNs) [7] and unsupervised machine learning [8, 9], among others.

Current state-of-the-art approaches to the NER task propose the use of RNNs to learn useful representations automatically because they facilitate the modeling of long-distance dependencies between words in a sentence. These networks usually rely on word embeddings, which represent words as vectors of numbers. There are different types of word embeddings: classical [10, 11], character-level [12] and contextualized [13] which are commonly pre-trained over very large corpora to capture latent syntactic and semantic similarities between words.

Based on the success of the previous challenges focused on the NER task in the biomedical domain (PharmaCoNER [14], eHealth-KD [15], eHealth CLEF [16], MEDDOCAN [17], CHEMDNER [18], i2b2 [19]), CANTEMIST is the first shared task specifically focusing on the extraction of a critical type of concept related to cancer, namely tumor morphology or neoplasms morphology.

Following the neural network proposed by Huang et al [20], our work uses the Bidirectional variant of Long Short Term Memory along with a stacked Conditional Random Fields decoding layer (BiLSTM-CRF) to extract the tumor morphology mentions in Spanish biomedical literature.

**Table 1**  
CANTEMIST corpus statistics.

	Train	Dev 1	Dev 2	Test gold
No. docs	500	250	250	300
No. annotated entities	6,396	3,341	2,660	3,633
No. annotated unique entities	1,978	1,189	960	1,210
No. annotated ICD-0 codes	493	338	334	386
No. sentences	22,022	10,847	9,917	12,739
No. tokens	441,993	219,172	177,574	240,562
No. unique tokens	22,280	15,391	13,921	16,551

We also evaluate the usefulness of some word embedding in two different ways: independently and in combination. Subsequently, we use an unsupervised dictionary-based method in order to automatically assign a ICD-O code to each entity detected.

### 3. Dataset

The CANTEMIS corpus was manually annotated by clinical experts following the guidelines<sup>1</sup>. The CANTEMIST corpus is composed of clinical cases and is divided into different sets: training, development and test set. The task organizers have also provided two development sets (dev 1 and dev 2) so that participants can train their systems more accurately. Some statistics of the CANTEMIST corpus can be found in Table 1. In this table you can see information related to the number of annotated entities, number of ICD-O codes, information related to sentences and vocabularies, among others.

## 4. Methodology

The workflow to address the proposed task in CANTEMIST challenge consists of two sequential steps, first detecting tumor morphology mentions in Spanish clinical documents, and subsequently the extracted entities must be assigned to a unique identifier code using ICD-O terminology. This section is organized according to each subtask carried out by our team: NER track and named entity normalization track.

### 4.1. Named entity recognition

The approach used to address in the first subtask of the CANTEMIST challenge is based on deep learning by implementing the RNN proposed by Huang et al. [20]. Specifically, we have used a BiLSTM with a sequential CRF.

---

<sup>1</sup><https://zenodo.org/record/3878179>

#### 4.1.1. Word embedding

RNNs generally use an embedding layer as an input, which makes it possible to represent words and documents using a dense vector representation. Word embeddings are a type of word representation that allows words with similar meanings to have a similar representation.

Our first approach involves combining different word embeddings to form the input layer to the proposed deep neural network. Each word representation used is explained in detail below:

- **FastText embeddings trained over Wikipedia.** This type of word embeddings are considered classic because they are static and word-level, meaning that each distinct word receives exactly one pre-computed embedding. Our experiments use FastText<sup>2</sup> embeddings trained over Spanish Wikipedia and size 100.

- **Spanish Medical Embeddings (SME).** Although there are available biomedical word embeddings for Spanish [21, 22, 23], we have tried to generate new ones from existing corpora related to the biomedical domain in Spanish. For this purpose, firstly we extracted the Spanish corpus from MeSpEN [24]. Later, extra information in Spanish from different clinical information websites such as Mayo Clinic [25], World Health Organization [26], and WebMD [27] was added to the former corpus. The pre-processing carried out to train the word embeddings consisted of converting the text to lowercase, removing the URLs, and removing the multi-lines. Finally, FastText [28] was used to perform the training by applying the following setup: skip-gram model, 0.05 for the learning rate, size of 300 for the word vectors, 10 for the number of epochs and 5 for the minimal number of word occurrences.

- **Contextual word embedding (CWE).** This word embeddings capture latent syntactic-semantic information that goes beyond standard word embeddings [29]. This representation treats text as distributions over characters and is capable of generating embeddings for any string of characters within any textual context, in other words, the same word will have different embeddings depending on its contextual use. For our experiments, we used the *pooled contextualized embeddings* proposed by Akbik et al. [30] to help with the recognition of tumor morphology mentions.

#### 4.1.2. BiLSTM-CRF model

For the entity recognition task, the annotations provided were encoded by using the BIO tagging scheme. Thus each token in a sentence was labeled with O (non-entity), B (beginning token of an entity), or I (inside token of an entity). This scheme is the most popular in the NER task.

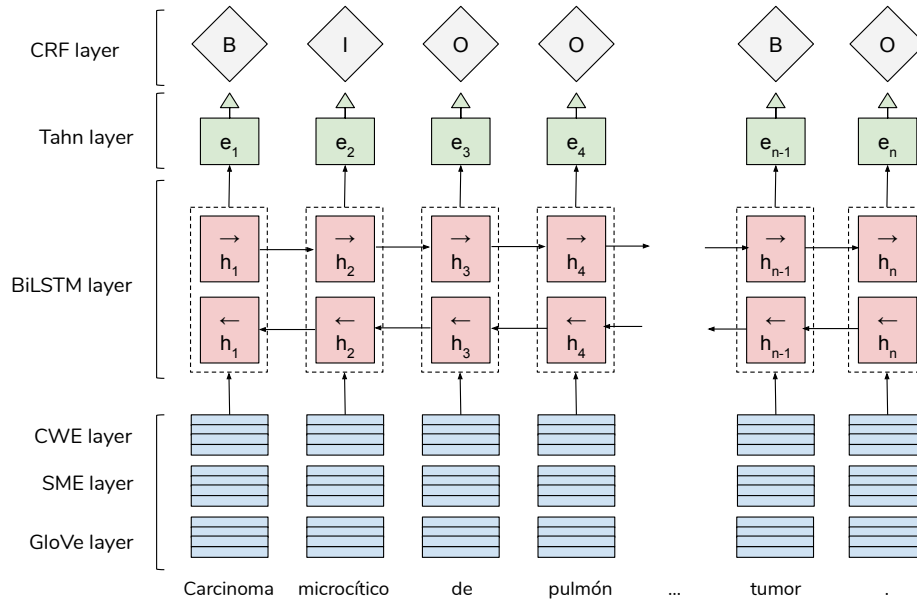
The architectures of BiLSTM-CRF model is illustrated in Figure 1. This architecture is similar to the proposal by Huang et al. [20], Lample et al. [31], Ma and Hovy [32]. This model is described in a layered way:

- 1 Embedding combination layer. Firstly, each sentence is divided by tokens using the SPACCC POS Tagger tool<sup>3</sup>. The sentence is represented as a sequence of words  $S = (w_1, w_2, w_3, \dots, w_n)$ , where  $n$  is the length of the sentence. In addition, each word is represented as a vector of concatenated embeddings. In our approach we use the word embeddings mentioned previously: FastText, SME and CWE.

---

<sup>2</sup><https://fasttext.cc>

<sup>3</sup>DOI: 10.5281/zenodo.2560344



**Figure 1:** Proposed model based on a BiLSTM-CRF neural network that uses a combination of different word embeddings as an input layer.

- 2 BiLSTM layer. The embeddings are given as input to a BiLSTM layer. In this layer, BiLSTM layer is to split the neurons of a regular LSTM into two directions, a forward state computes a representation  $\vec{h}_n$  of the sequence from left to right at every word  $n$ , and another for negative time direction (backward states) computes a representation  $\overleftarrow{h}_n$ . Using two time directions, input information from the past and future of the current time frame can be used to better understand the context of the medical report. The representation of a word  $h_n = [\vec{h}_n; \overleftarrow{h}_n]$  is obtained by concatenating its forward and backward states[31].
- 3 Tahn layer. Next, tahn layer is used to predict confidence scores for each word according to the existing labels in the corpus.
- 4 CRF layer. Finally, CRF layer combines the advantage of graphical modeling to predict multivariate output, in other words, this layer decodes the best label in all possible labels [33].

For the implementation, we employed Flair [34]. Flair is a simple framework for NLP tasks including NER. Flair is used with the following configuration: learning rate as 0.1, dropout as 0.5, maximum epoch as 150, 300 neurons with *tahn* activation function and a batch size of 32.

## 4.2. Named entity normalization

The second subtask consists of assigning a unique ICD-O-10 identifier to each entity annotated in subtask one. For this subtask we use an unsupervised dictionary-based method. The methodology followed for the selection of an identifier is shown below:

### 1 ICD-O dictionary creation.

In order to generate the dictionary, we use the descriptions and ICD-0 codes obtained from the official website<sup>4</sup> and the training set provided by the organizers. The main purpose was to generate a tuple containing the ICD-O code and the description as shown above:

```
< 8000/6, neoplasia metastásica >  
< 8000/6, émbolo tumoral >  
< 8000/6, tumor metastásico >  
< 8000/3, neoplasia maligna >  
< 8000/3, cáncer >  
< 8000/3, malignidad >
```

As we can see, the same code corresponds to several descriptions, which means that these are synonymous. All descriptions are in lowercase in order to obtain a better future match.

### 2 Levenshtein distance.

The second step is to compute the Levenshtein distance between the recognized entity and each dictionary description. Levenshtein distance is the minimum number of operations required to transform one character string into another. For instance, for the entity *cáncer*, the output provided by this step would be as shown below::

```
< 8000/6, neoplasia metastásica, 20 >  
< 8000/6, émbolo tumoral, 13 >  
< 8000/6, tumor metastásico, 16 >  
< 8000/3, neoplasia maligna, 17 >  
< 8000/3, cáncer, 0 >  
< 8000/3, malignidad, 9 >
```

According to our previous example, *cáncer* gets a value of 0 with the code *8000/3 - cáncer*. However, we can see that the entity *cáncer* gets other values for each description, for example, the distance between *cáncer* and *émbolo tumoral* is 13.

### 3 Code ranking.

The last step involves choosing from all the calculated distance values. To do this, we sort the dictionary by Levenshtein distance and choose the first ICD-O code from the list. In this way we assign a unique identifier to the entity.

## 5. Experimental setup and results

The task organizers provided several datasets (training, development and test) to allow the participants to train their systems properly. For all scenarios, we used the training set and the number 1 development set (22,022 + 10,847 sentences respectively) for training, while the number 2 development set (9,917 sentences) was used to validate our system. We decided to include the number 1 development set in the training since it has a greater number of annotated entities than the number 2 development set.

---

<sup>4</sup>[https://eciemaps.mscbs.gob.es/ecieMaps/browser/index\\_o\\_3.html](https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_o_3.html)

**Table 2**

Evaluation results obtained by the SINAI team in NER subtask.

System	Precision (%)	Recall (%)	F1-score (%)
FastText	80.9	81.7	81.3
SME	83.5	85.5	84.5
CWE	85.9	82.8	84.3
SME + CWE	85.9	85.1	85.5
SME + CWE + FastText	85.8	84.7	85.2

**Table 3**

Evaluation results obtained by the SINAI team in named entity normalization subtask. P: Precision, R: Recall, F: F1-score, No-Met: non-metastasis

System	P (%)	R (%)	F (%)	P-No-Met (%)	R-No-Met (%)	F1-No-Met (%)
FastText	72.8	73.5	73.2	73.0	70.7	71.8
SME	74.7	76.6	75.6	74.8	73.2	74.0
CWE	76.9	74.1	75.5	75.1	73.0	74.0
SME + CWE	76.3	75.5	75.9	74.9	73.2	74.0
SME + CWE + FastText	76.4	75.4	75.9	75.3	73.5	74.4

The metrics defined by the CANTEMIST challenge to evaluate the submitted experiments are those commonly used for some NLP tasks such as NER or text classification, namely precision, recall, and F1-score. Table 2 and 3 shows the results obtained by the SINAI team for the first and the second subtask respectively.

As we can see in Table 2, the results are encouraging. The combination of different word embeddings improves the use of each one of them separately. This means that each word embeddings used provides a meaningful representation of each word, which helps the neural network to learn to detect the entity.

The use of the classic word embeddings trained on wikipedia (FastText) obtains the worst value of precision, recall and F1. On the other hand, the use of the word embeddings trained on a corpus related to medicine improves and achieves 84.5 of F1 and gets the best recall (85.5%). SME word embeddings also improve CWE results, although in this case the difference is not very significant. These results demonstrate that using resources trained on a specific domain helps the task to be solved. In our case, training some word embeddings related to the biomedical domain helps the recognition of named entities.

Our best result (taking into account the F1 measure) was obtained by combining two types of word embeddings: SME + CWE. With this combination we achieved an 85.5% of F1, an 85.9% of precision and an 85.1% of recall.

In addition, we were also able to check the use of FastText word embeddings trained with Wikipedia in their combination (SME + CWE + FastText) but they reached worse values than without them.

Table 3 summarizes the results obtained for the second subtask (named entity normalization) by the SINAI team. In this subtask, we use the output of the previous task to standardize the

recognized entities. Taking into account the F1-score for the extraction of neoplastic tumours, we found that the combination of word embeddings obtains the same percentage (95.9%). In contrast, if we consider the measure F1 for non-metastasis recognition, the best value of F1 was obtained using the combination of the three word representations (SME + CWE + FastText).

## 6. Conclusion and future work

This paper presents the participation of the SINAI group in the CANTEMIST challenge. Our group has participated in two subtasks proposed by the challenge: NER track and named entity normalization track. Both subtasks are considered relevant in the NLP because they are related to information extraction and standardization in the biomedical domain through an oncology dictionary. More specifically, the first subtask addresses the task of extracting mentions of neoplasms morphology mentions, and the second subtask involves assigning an ICD-O code to each previously recognized entity.

In the first subtask our proposal follows a deep learning-based approach for NER in Spanish clinical cases. This methodology is focused on the use of a BiLSTM-CRF where different word embeddings are combined as input to the architecture. Our main goal was to prove the performance of different types of word embeddings for the NER task in the medical domain: own-generated medical embeddings (SME), contextualized embeddings (CWE) and FastText embeddings trained over Wikipedia. The obtained results are encouraging for the NER task achieving 85.5% of F1-score, 85.9% of precision and 85.1% of recall. On the other hand, in the second subtask we use an unsupervised dictionary-based method. In this scenario, the combination of different word representations also obtained the best value achieving 75.9% of F1-score, 76.4% of precision and 75.4% of recall. For future work we plan to improve our entity detection system using new transfer learning techniques. In addition, there are available pre-trained models for the biomedical domain such as BioBERT that could be taken into consideration. Although BioBERT is in English, an ideal scenario would be the generation of a new model for Spanish through a large biomedical corpus.

## Acknowledgments

This work has been partially supported by LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government and Fondo Europeo de Desarrollo Regional (FEDER).

## References

- [1] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [2] C. Friedman, Towards a comprehensive medical language processing system: methods and issues., in: Proceedings of the AMIA annual fall symposium, American Medical Informatics Association, 1997, p. 595.



- [3] A. R. Aronson, F.-M. Lang, An overview of metamap: historical perspective and recent advances, *Journal of the American Medical Informatics Association* 17 (2010) 229–236.
- [4] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications, *Journal of the American Medical Informatics Association* 17 (2010) 507–513.
- [5] D. Campos, S. Matos, J. L. Oliveira, Biomedical named entity recognition: a survey of machine-learning tools, *Theory and Applications for Advanced Text Mining* (2012) 175–195.
- [6] P. L. Úbeda, M. C. D. Galiano, M. T. Martín-Valdivia, L. A. U. Lopez, Using machine learning and deep learning methods to find mentions of adverse drug reactions in social media, in: *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, 2019, pp. 102–106.
- [7] P. L. Úbeda, M. C. D. Galiano, L. A. U. Lopez, M. T. Martín-Valdivia, Using Snomed to recognize and index chemical and drug mentions, in: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, 2019, pp. 115–120.
- [8] P. López-Ubeda, M. C. Díaz-Galiano, M. T. Martín-Valdivia, L. A. Urena-López, Sinai en tass 2018 task 3. clasificando acciones y conceptos con umls en medline, *Proceedings of TASS 2172* (2018).
- [9] P. López-Úbeda, M. C. Díaz-Galiano, A. Montejo-Ráez, M.-T. Martín-Valdivia, L. A. Ureña-López, An Integrated Approach to Biomedical Term Identification Systems, *Applied Sciences* 10 (2020) 1726.
- [10] J. Pennington, R. Socher, C. D. Manning, GloVe: Global vectors for word representation, in: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014. doi:10.3115/v1/d14-1162.
- [11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: *Advances in Neural Information Processing Systems*, 2013. arXiv:1310.4546.
- [12] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 2016. doi:10.18653/v1/n16-1030. arXiv:1603.01360.
- [13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2018. doi:10.18653/v1/n18-1202. arXiv:1802.05365.
- [14] A. G. Agirre, M. Marimon, A. Intxaurreondo, O. Rabal, M. Villegas, M. Krallinger, Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track, in: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, 2019, pp. 1–10.
- [15] A. Piad-Morffis, Y. Gutiérrez, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, 2020.

- [16] L. Kelly, H. Suominen, L. Goeuriot, M. Neves, E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, G. Zuccon, H. Scells, et al., Overview of the clef ehealth evaluation lab 2019, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 322–339.
- [17] M. Marimon, A. Gonzalez-Agirre, A. Intxaurreondo, H. Rodríguez, J. A. Lopez Martin, M. Villegas, M. Krallinger, Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), volume TBA, CEUR Workshop Proceedings (CEUR-WS.org), Bilbao, Spain, 2019, p. TBA. URL: TBA.
- [18] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, A. Valencia, Chemdner: The drugs and chemical names extraction challenge, *Journal of cheminformatics* 7 (2015) S1.
- [19] W. Sun, A. Rumshisky, O. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 Challenge, *Journal of the American Medical Informatics Association* 20 (2013) 806–813. URL: <https://doi.org/10.1136/amiajnl-2013-001628>. doi:10.1136/amiajnl-2013-001628. arXiv:<https://academic.oup.com/jamia/article-pdf/20/5/806/17374624/20-5-806.pdf>.
- [20] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, arXiv preprint arXiv:1508.01991 (2015).
- [21] F. Soares, M. Villegas, A. Gonzalez-Agirre, M. Krallinger, J. Armengol-Estapé, Medical word embeddings for Spanish: Development and evaluation, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 124–133. URL: <https://www.aclweb.org/anthology/W19-1916>. doi:10.18653/v1/W19-1916.
- [22] S. Santiso, A. Casillas, A. Pérez, M. Oronoz, Word embeddings for negation detection in health records written in Spanish, *Soft Computing* (2019). doi:10.1007/s00500-018-3650-7.
- [23] I. Segura-Bedmar, P. Martínez, Simplifying drug package leaflets written in Spanish by using word embedding, *Journal of Biomedical Semantics* (2017). doi:10.1186/s13326-017-0156-7.
- [24] M. Villegas, A. Intxaurreondo, A. Gonzalez-Agirre, M. Marimon, M. Krallinger, The MeSpEN resource for English-Spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations, *LREC MultilingualBIO: Multilingual Biomedical Text Processing* (Malero M, Krallinger M, Gonzalez-Agirre A, eds.) (2018).
- [25] Mayo clinic, 1998-2020. URL: <https://www.mayoclinic.org/es-es>.
- [26] Organización mundial de la salud, 2020. URL: <https://www.who.int/es>.
- [27] Webmd - better information. better health., 2005-2020. URL: <https://www.webmd.com/>.
- [28] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics* 5 (2017) 135–146.
- [29] A. Akbik, D. Blythe, R. Vollgraf, Contextual string embeddings for sequence labeling, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.
- [30] A. Akbik, T. Bergmann, R. Vollgraf, Pooled contextualized embeddings for named entity recognition, 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, 2019, pp. 724–728. URL: <https://www.aclweb.org/anthology/N19-1078>. doi:10.18653/v1/N19-1078.

- [31] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, arXiv preprint arXiv:1603.01360 (2016).
- [32] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf, arXiv preprint arXiv:1603.01354 (2016).
- [33] J. Lafferty, A. McCallum, F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).
- [34] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 54–59. URL: <https://www.aclweb.org/anthology/N19-4010>. doi:10.18653/v1/N19-4010.