SINAI at CLEF eHealth 2020: testing different pre-trained word embeddings for clinical coding in Spanish

José M. Perea-Ortega[1][0000−0002−7929−3963], Pilar López-Úbeda[2][0000−0003−0478−743X], Manuel C. Díaz-Galiano[2][0000−0001−9298−1376], M. Teresa Martín-Valdivia[2][0000−0002−2874−0401], and L. Alfonso Ureña-López[2][0000−0001−7540−4059]

[1] University of Extremadura, Badajoz, Spain
jmperea@unex.es
[2] University of Jaén, Jaén, Spain
{plubeda,mcdiaz,maite,laurena}@ujaen.es

Abstract. This paper describes the system presented by the SINAI team for the Multilingual Information Extraction task of the CLEF eHealth Lab 2020. This task focuses on the automatic assignment of the International Classification of Diseases (ICD) codes to health-related texts in Spanish. Our proposal follows a deep learning-based approach where we have used the bidirectional variant of a Long Short Term Memory (LSTM) network along with a stacked Conditional Random Fields (CRF) decoding layer (BiLSTM+CRF). The aim of the experiments carried out was to test the performance of different pre-trained word embeddings for recognizing diagnoses and procedures in clinical text. The main finding was that combining word embeddings could be a useful strategy to apply for deep learning-based approaches, even though the combined embeddings do not belong to the medical domain. The best MAP scores achieved were 0.314 and 0.293 for the CodiEsp-D and CodiEsp-P subtasks, respectively.

1 Introduction

Within health organizations, clinical coding can be seen as a task in which information from Electronic Health Records (EHR) is translated into alphanumeric codes by using internationally recognized terminologies or classifications. For example, acute appendicitis is represented by code 'K35.8' using the International Classification of Diseases (ICD). In hospitals, these data are critical for clinical professionals, research, and other purposes, such as statistical analysis

and decision-making. However, this task is often performed manually by clinical coders, where the effort required for information abstraction is extremely laborious, time-consuming, and prone to human errors.

To alleviate this problem, the research community has to lead to the organization of challenges and shared tasks to promote automatic clinical coding systems. Over the past years, CLEF eHealth offered challenges addressing several aspects of related information access, providing researchers with datasets to work with and validate the outcomes [18, 17, 7]. In 2020, they continue to offer two shared tasks: i) Multilingual Information Extraction (IE), which focuses on ICD coding for clinical textual data in Spanish, and ii) Consumer Health Search, which follows a standard information retrieval shared challenge paradigm.

This paper describes the system presented by the SINAI team for the Multilingual IE subtask of the CLEF eHealth Lab 2020. Automatic assignment of ICD codes for health-related texts can be considered a special case of multilabel text classification, which may be approached either from a Natural Language Processing (NLP) perspective by using syntactic and/or semantic decision rules, or a machine learning perspective. For this purpose, machine learning algorithms have been successfully applied, particularly those that have focused on deep learning-based methods. In this paper, we mainly focus on Recurrent Neural Network (RNN), specifically on the bidirectional variant of Long Short Term Memory along with a stacked Conditional Random Fields decoding layer (BiLSTM+CRF) [8, 15]. For training the network, our approach proposes the use of different types of vectors representing word meanings (word embeddings) by using only the training data provided by the organizers.

In the next section, we briefly present the background. Section 3 describes the architecture of our system presented to the Multilingual IE task of the CLEF eHealth lab. Section 4 reports the results obtained for the different experiments carried out and, finally, conclusions and future work are presented in Section 5.

2 Background

Clinical coding can be approached as a Named Entity Recognition (NER) task where medical concepts should be firstly detected within the text. Then, they should be mapped to a specific code related to that concept. In recent years, deep learning approaches have been used for NER, leading to state-of-the-art results [9, 5, 14]. Our group has experience in clinical NER by using different methodologies such as traditional machine learning [11], Recurrent Neural Networks (RNNs) [12] and unsupervised machine learning [10, 13].

Clinical NER is being commonly approached as a sequence labelling problem, where the text is treated as a sequence of words to be labeled with linguistic tags. Current state-of-the-art approaches for sequence labeling propose the use of RNNs to learn useful representations automatically, since they facilitate modeling long-distance dependencies between the words in a sentence. These networks usually rely on word embeddings, that are commonly pre-trained over very large corpora to capture latent syntactic and semantic similarities between words. A

novel type of word embeddings called *contextual string embeddings* is proposed by Akbik et al. [2], which essentially model words as sequences of characters, thus contextualizing a word by their surrounding text and allowing the same word to have different embeddings depending on its contextual use.

3 System Overview

3.1 Dataset

The corpus provided for the Multilingual IE task of the CLEF eHealth lab consisted of 1,000 clinical case comprising 16,504 sentences and 396,988 words, with an average of 396.2 words per clinical case [16]. The corpus had 18,483 annotated codes, of which, 3,427 were unique. These were divided into two groups:

- ICD10-CM codes (CIE10 *Diagnóstico* in Spanish). They are codes belonging to the International Classification of Diseases, 10th revision, Clinical Modification, and they are tagged as `DIAGNOSTICO`.
- ICD10-PCS codes (CIE10 *Procedimiento* in Spanish). They are codes belonging to the International Classification of Diseases, 10th revision, Procedure codes (related to procedures performed in hospitals), and they are tagged as `PROCEDIMIENTO`.

The entire corpus was randomly sampled into three subsets: training, development and test. The training set comprised 500 clinical cases, and the development and test sets 250 clinical cases each. Together with the test set, the organizers released an additional collection of more than 2,000 documents (background set) to make sure that participating teams were not be able to do manual corrections.

We performed a preliminar preprocessing phase to the train and dev data sets provided for the task, considering `DIAGNOSTICO` and `PROCEDIMIENTO` annotations in a separate way. First, we used Freeling [19] to tokenize the text and get the Part-Of-Speech (POS) tag of each word. Then, we generated the training corpus with the following features: original form of the word, POS tag and NER tag. For performing the NER tagging, the provided annotations were encoded by using the BIO tagging scheme, which represents that a token is at the beginning of an entity (B-ENT), inside of an entity (I-ENT), or outside (O) of an entity. Finally, only the sentences with BIO tags were considered to generate the training corpus. Figure 1 shows an example of the generated training corpus for the assignment of `DIAGNOSTICO` codes (left) and `PROCEDIMIENTO` codes (right).

3.2 BiLSTM+CRF architecture

Our proposal follows a deep learning-based approach where a Recurrent Neural Network (RNN) is used to generate different learning models. Specifically, we have used the bidirectional variant of Long Short Term Memory along with a stacked Conditional Random Fields decoding layer (BiLSTM+CRF) [8, 15]. This specialized architecture is chosen to approach NER because it facilitates

```
No RN O
antecedentes NC O
de SP O
nefrolitiasis NC B-ENT
ni CC O
hematuria NC B-ENT
ni CC O
infecciones NC B-ENT
del SP I-ENT
tracto NC I-ENT
urinario AQ I-ENT
. Fp O
```

```
Mujer NC O
de SP O
42 Z O
años NC O
en SP O
el DA O
momento NC O
de SP O
someterse VM O
a SP O
trasplante NC B-ENT
hepático AQ I-ENT
. Fp O
```

Fig. 1. Example of the generated training corpus for the assignement of `DIAGNOSTICO` codes (left) and `PROCEDIMIENTO` codes (right).

the processing of arbitrary length input sequences and enables the learning of long-distance dependencies, which is particularly advantageous in the case of clinical coding to detect medical concepts. Moreover, our approach proposes the combination of different types of pre-trained word embeddings by concatenating each embedding vector to form the final word vectors. In this way, the probability of recognizing a specific medical concept in a text should be increased since different types of word representation are combined. For the case of contextual string embeddings, since they are robust in face of misspelled words, we suppose they could be highly suitable for clinical NER.

As shown in Figure 2, the proposed architecture gets a context of each word on the clinical case using BiLSTM (encoding layer), and then makes word predictions simultaneously on the CRF layer (decoding layer). It should be noted that diagnoses and procedures were managed independently, i.e., we generated learning models to predict diagnoses exclusively, and other different models to predict procedures. We have used Flair Library[3] [1] to apply this architecture. Flair is an open source NLP library developed by Zalando Research. It is built on Pytorch[4] and has fairly good GPU support.

3.3 Pre-trained Word Embeddings

RNNs generally use an embedding layer as an input, which makes it possible to represent words using a dense vector representation. In order to fit the text input into the BiLSTM+CRF architecture, we have combined different types of pre-trained word embeddings:

- **Classic Word Embeddings**. Classic word embeddings are static and word-level, meaning that each distinct word gets exactly one pre-computed embed-

[3] `http://github.com/flairNLP/flair`
[4] `http://pytorch.org`

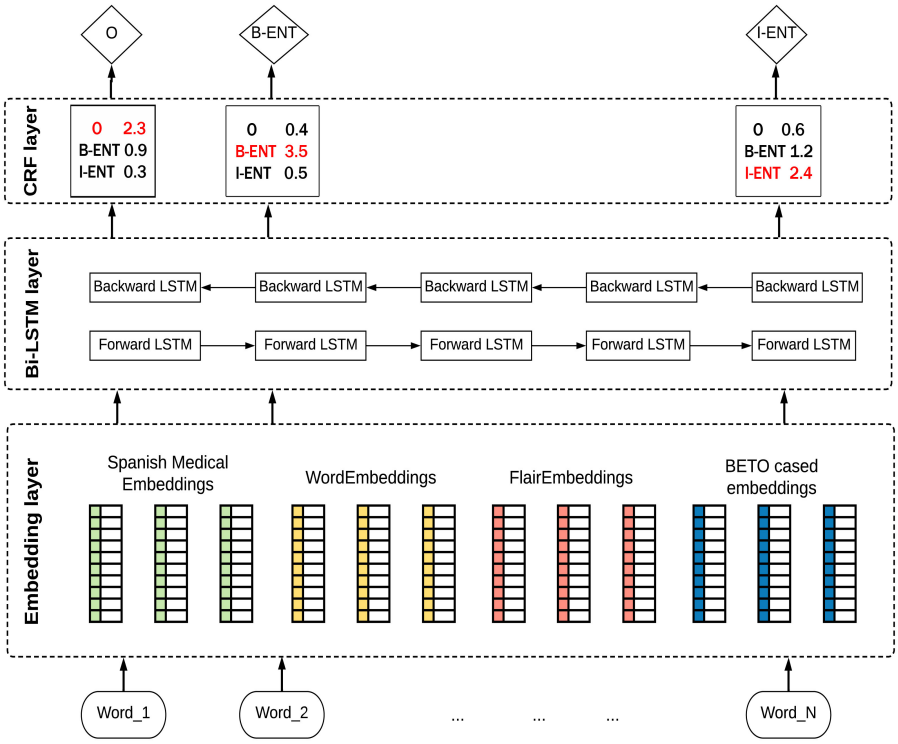

Fig. 2. BiLSTM+CRF architecture that uses different word embeddings as an input layer.

ding. For our experiments we have used the *WordEmbeddings* class provided by the Flair Library [1] that was initialized with fastText[5] embeddings pre-trained over Spanish Wikipedia.

– **Contextual Word Embeddings**. Contextual word embeddings are considered powerful embeddings because they capture latent syntactic-semantic information that goes beyond standard word embeddings [2]. These embeddings are based on character-level language modeling and their use is particularly advantageous when the NER task is approached as a sequential labeling problem. For our experiments we have used the *FlairEmbeddings* class provided by the Flair Library. These contextual string embeddings were pre-trained over Spanish Wikipedia.
– **Word Embeddings based on *Transformers***. Bidirectional Encoder Representations from Transformers (BERT) [6] is based on a multilayer bidirectional transformer-encoder, where the transformer neural network uses parallel attention layers rather than sequential recurrence [21]. This kind of embeddings are commonly pre-trained over very large corpora to capture la-

[5] `https://fasttext.cc`

tent syntactic and semantic similarities between words. For our experiments we have used BETO cased embeddings [3], which follows a BERT model trained on a big corpus composed of text portions extracted from different web sources in Spanish.
– **In-domain Word Embeddings**. Most of the available word embeddings are focused on general-domain texts, and their uses not necessarily apply well to clinical text analysis [4]. In order to test biomedical word embeddings for our experiments, we have used the first version of Spanish Medical Embeddings[6] [20], which are based on the fastText model and were developed from two data sources: (i) the SciELO database, and (ii) Wikipedia Health, comprised by the categories of Pharmacology, Pharmacy, Medicine and Biology.

4 Experiments and Results

Our team submitted a total of 10 runs for the Multilingual IE task, 5 for each proposed main subtasks: diagnosis coding (CodiEsp-D) and procedure coding (CodiEsp-P). Besides, other 5 runs were submitted for the exploratory subtask called CodiEsp-X, where systems were required to submit the reference in text to the predicted codes for both diagnosis and procedure.

The aim of the experiments carried out was to test the performance of different pre-trained word embeddings for recognizing diagnoses and procedures in clinical text. Thus, several learning models were generated using the default hyperparameter setting in Flair: 0.1 of learning rate, 32 of batch size, 0.5 of dropout probability, and 150 of maximum epoch. All experiments were performed on a single Tesla-V100 32 GB GPU with 192 GB of RAM. The configuration used for each submitted run is shown below:

– Run 1: Spanish Medical Embeddings (**SME**). In-domain word embeddings generated from two data sources: (i) the SciELO database, and (ii) Wikipedia Health.
– Run 2: WordEmbeddings + FlairEmbeddings (**Word+Flair**). This was performed by using the *StackedEmbeddings* class of Flair, whereby words are embedded in a single vector using a concatenation of the different embeddings combined.
– Run 3: WordEmbeddings (**WordEmbed**).
– Run 4: BETO cased embeddings (**BETO**).
– Run 5: FlairEmbeddings (**Flair**).

The evaluation metrics defined by the organizers were those commonly used for some NLP tasks such as NER or information retrieval, namely Mean Average Precision (MAP), Precision (P), Recall (R), and F1-score. Table 1 and Table 2 shows the results obtained by the SINAI team for the main and exploratory subtasks, respectively.

[6] `http://doi.org/10.5281/zenodo.2542722`

Subtask	Model	MAP	P	R	F1-score
CodiEsp-D	SME	0.301	0.412	0.538	0.467
	Word+Flair	**0.314**	**0.443**	**0.544**	**0.488**
	WordEmbed	0.302	0.418	0.540	0.471
	BETO	0.251	0.450	0.433	0.441
	Flair	0.291	0.402	0.528	0.456
CodiEsp-P	SME	0.280	0.367	0.452	0.405
	Word+Flair	**0.293**	**0.370**	**0.476**	**0.416**
	WordEmbed	0.271	0.342	0.455	0.391
	BETO	0.250	0.343	0.422	0.378
	Flair	0.254	0.318	0.458	0.376

Table 1. Official results obtained by the SINAI team in the main subtasks CodiEsp-D and CodiEsp-P.

Subtask	Model	P	R	F1-score
CodiEsp-X	SME	0.330	0.425	0.371
	Word+Flair	**0.360**	**0.447**	**0.399**
	WordEmbed	0.323	0.420	0.365
	BETO	0.337	0.346	0.342
	Flair	0.313	0.421	0.359

Table 2. Official results obtained by the SINAI team in the exploratory subtask CodiEsp-X.

As shown in Table 1, the results obtained for both main subtasks are relatively low. This behavior may be due to the limited amount of training data used since we have only used the sentences with BIO tags found in the train and dev data sets provided by the organization. Another reason of the poor performance could be the use of embeddings that have not been generated from medical texts. Nevertheless, our best MAP result in both subtasks was achieved when different pre-trained word embeddings were combined (classic and contextual) and used as an input layer to the BiLSTM+CRF architecture. This may lead to the finding that combining word embeddings could be an interesting strategy to consider for the future, even though the combined embeddings do not belong to the medical domain.

5 Conclusions and future work

This paper describes the participation of the SINAI research group in the Multilingual Information Extraction task of the CLEF eHealth Lab 2020. This task focuses on the automatic assignment of codes to clinical textual data in Spanish. The classification proposed to perform the coding is the Spanish version of the International Classification of Diseases, 10th revision, ICD10 (CIE10 in Span-

ish). Two main NLP subtasks were defined: diagnosis coding (CodiEsp-D) and procedure coding (CodiEsp-P).

Our proposal follows a deep learning-based approach for clinical NER. It is focused on the use of a BiLSTM+CRF architecture where different pre-trained word embeddings are used as an input to the neural network. Then, training is performed by using the annotated datasets provided by the organization, which were previously tokenized and NER-tagged by using the BIO scheme. Our main goal was to test the performance of different types of pre-trained word embeddings for detecting and recognizing diagnoses and procedures in medical texts in Spanish. We believe that the poor performance obtained is due to the limited amount of training data, and the use of word embeddings that were not generated from medical texts. Nevertheless, the main finding was that combining word embeddings could be a useful strategy to apply for deep learning-based approaches, even though the combined embeddings do not belong to the medical domain.

For future work, we first should analyze in-depth why the results were low. Then, further research should focus on injecting domain knowledge into the deep learning model. Another future direction would be to explore how the machine translation of Spanish into English performs to use greater availability of existing knowledge resources in English.

Acknowledgements

This work has been partially supported by LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government, Junta de Extremadura (GR18135) and Fondo Europeo de Desarrollo Regional (FEDER).

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: 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). pp. 54–59 (2019)
2. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1638–1649 (2018)
3. Cañete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish Pre-Trained BERT Model and Evaluation Data. In: Practical ML for Developing Countries Workshop (ICLR 2020) (2020)
4. Chiu, B., Crichton, G.K.O., Korhonen, A., Pyysalo, S.: How to Train good Word Embeddings for Biomedical NLP. In: Cohen, K.B., Demner-Fushman, D., Ananiadou, S., Tsujii, J. (eds.) BioNLP@ACL. pp. 166–174. Association for Computational Linguistics (2016)
5. Chokwijitkul, T., Nguyen, A., Hassanzadeh, H., Perez, S.: Identifying risk factors for heart disease in electronic medical records: A deep learning approach. In: Proceedings of the BioNLP 2018 workshop. pp. 18–27. Association for Computational

Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/W18-2303, https://www.aclweb.org/anthology/W18-2303
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Saez Gonzales, G., Viviani, M., Xu, C.: Overview of the CLEF eHealth evaluation lab 2020. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., andNicola Ferro, L.C. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020) . LNCS Volume number: 12260 (2020)
8. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
9. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 260–270. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-1030, https://www.aclweb.org/anthology/N16-1030
10. López-Ubeda, P., Díaz-Galiano, M.C., Martín-Valdivia, M.T., Ureña-López, L.A.: Sinai en TASS 2018 task 3. clasificando acciones y conceptos con UMLS en Medline. Proceedings of TASS (2018)
11. López-Ubeda, P., Díaz-Galiano, M.C., Martín-Valdivia, M.T., Ureña-López, L.A.: 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. pp. 102–106 (2019)
12. López-Ubeda, P., Díaz-Galiano, M.C., Ureña-López, L.A., Martín-Valdivia, M.T.: Using Snomed to recognize and index chemical and drug mentions. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. pp. 115–120 (2019)
13. López-Úbeda, P., Díaz-Galiano, M.C., Montejo-Ráez, A., Martín-Valdivia, M.T., Ureña-López, L.A.: An Integrated Approach to Biomedical Term Identification Systems. Applied Sciences **10**(5), 17–26 (2020)
14. Luu, T.M., Phan, R., Davey, R., Chetty, G.: Clinical name entity recognition based on recurrent neural networks. In: 2018 18th International Conference on Computational Science and Applications (ICCSA). pp. 1–9 (2018)
15. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers (2016). https://doi.org/10.18653/v1/p16-1101
16. Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., Krallinger, M.: Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of eHealth CLEF 2020. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., andNicola Ferro, L.C. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020) . LNCS Volume number: 12260 (2020)
17. Neves, M.L., Butzke, D., Dörendahl, A., Leich, N., Hummel, B., Schönfelder, G., Grune, B.: Overview of the clef ehealth 2019 multilingual information extraction. In: Cappellato, L., Ferro, N., Losada, D.E., Müller, H. (eds.) CLEF (Working

Notes). CEUR Workshop Proceedings, vol. 2380. CEUR-WS.org (2019), http://dblp.uni-trier.de/db/conf/clef/clef2019w.html#NevesBDLHSG19
18. Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikan, L., Ramadier, L., Rey, G., Zweigenbaum, P.: Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF (Working Notes). CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018), http://dblp.uni-trier.de/db/conf/clef/clef2018w.html#NeveolRGMOPRRZ18
19. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA, Istanbul, Turkey (May 2012)
20. Soares, F., Villegas, M., Gonzalez-Agirre, A., Krallinger, M., Armengol-Estapé, J.: Medical word embeddings for Spanish: Development and evaluation. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 124–133. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/W19-1916, https://www.aclweb.org/anthology/W19-1916
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)