Tumor Entity Recognition and Coding for Spanish Electronic Health Records

Fadi Hassan[a,b], David Sanchez[a] and Josep Domingo-Ferrer[a]

[a] CYBERCAT-Center for Cybersecurity Research of Catalonia. UNESCO Chair in Data Privacy.), Department of Computer Science and Mathematics Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia
[b]Department of Computer Science, Hodeidah University, Hodeidah 1821, Yemen

Abstract

This paper describes a two-stage system to solve tumor entity detection and coding in Spanish health records. This system is submitted to the CANcer TExt Mining Shared Task (CANTEMIST), a challenge in the IberLEF 2020 Workshop. We include a comparison between two kinds of systems to tackle this problem. The first kind employ feature-based Conditional Random Fields (CRF), and the second kind is based on deep learning models. The reported experiments show that our proposals and their combination achieve a micro-F1 of 83.1% and 78.6% on the test data set for the first and second sub-tasks, respectively, and a MAP of 79.7% on the third sub-task.

Keywords

Electronic Health Records, Deep Learning, Convolution Neural Networks, Conditional Random Fields, Named entity recognition

1. Introduction

Electronic health records (EHRs) are systematized collections of patient electronically-stored health information. Usually, EHRs come in different formats, the most popular being free-text form. Text is a rich source of information for healthcare research and, in particular, to machine learning tasks. So far, most of the healthcare data available for research is written in English, so there is a need of tagged data for other languages like Spanish.

On the basis of this need, the *Plan de Impulso de las Tecnologías del Lenguaje* (Plan TL) organizes the CANcer TExt Mining Shared Task (CANTEMIST) [1]. This task aims to provide the medical and machine learning fields with Spanish labeled data. Since accessing EHRs is tricky due to privacy issues, the provided notes in this task were clinical case reports (CCRs), which they are as close as possible to EHRs.

CANTEMIST is a shared task in the IberLEF 2020 workshop, which focuses on extracting named entities of critical concepts related to cancer in EHRs. This contest includes three independent sub-tasks: 1) CANTEMIST-NER track, which requires finding tumor morphology mentions automatically; 2) CANTEMIST-NORM track, which is a named entity normalization

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: fadi.hassan@urv.cat (F. Hassan); david.sanchez@urv.cat (D. Sanchez); josep.domingo@urv.cat (J. Domingo-Ferrer)
ORCID:

Dataset	# Tokens	# Sentences	# Named Entities
Train	661,853	27,662	9,609
Dev 1	219,377	9,097	3,287
Dev 2	177,663	8,346	2,624
Test	240,666	10,745	3,569

Table 1
Comparison between the datasets.

task requiring all tumor morphology mentions being returned with their corresponding eCIE-O-3.1 codes; and 3) CANTEMIST-CODING track, which requires returning, for each of document, a ranked list of its corresponding ICD-O-3 codes [2] (Spanish version: eCIE-O-3.1[3]).

As participants in the contest, we tackled the problem by designing several systems based on Conditional Random Fields (CRFs) and deep learning models. We also proposed a combination of these systems to improve the results.

For the first task (NER task), we design two systems (1-BiLSTM and 2-BiLSTM) based on Artificial Neural Networks (ANNs) with Bi-directional Long Short-Term Memory neural networks (Bi-LSTM), and a third one relying on feature-based Conditional Random Fields (CRF). Then, we combined the results of the three systems using voting.

For the second and third tasks (normalization and coding tasks), we have designed a system based on Convolution Neural Networks (CNNs) and Long Short-Term Memory neural networks (LSTMS).

The remainder of the paper is organized as follows. In Section 2, we briefly describe the data. Section 3 describes the systems we propose. Results and discussions are presented in Section 4. Section 5 presents the conclusions and depicts some lines of future work.

2. Data Description

The proposed systems are trained and validated on train and development datasets provided by the organizers of the contest. The corpora released for the tasks consists of 1,300 Spanish medical records, divided into 500 as training data, 250 for each dev1 and dev2 data, and 300 as test data. Table 1 shows general statistics about the four datasets.

3. Systems Description

We developed several systems to detect and code tumor named entities in Spanish health records. The next subsections describe the steps we followed to train and use these systems. Figure 1 describes the general data flow of the system.

3.1. Text Tokenization

In this step we performed a sentence splitting and text tokenization of the health records. For this, we used the spaCy pre-trained model for the Spanish language [4].

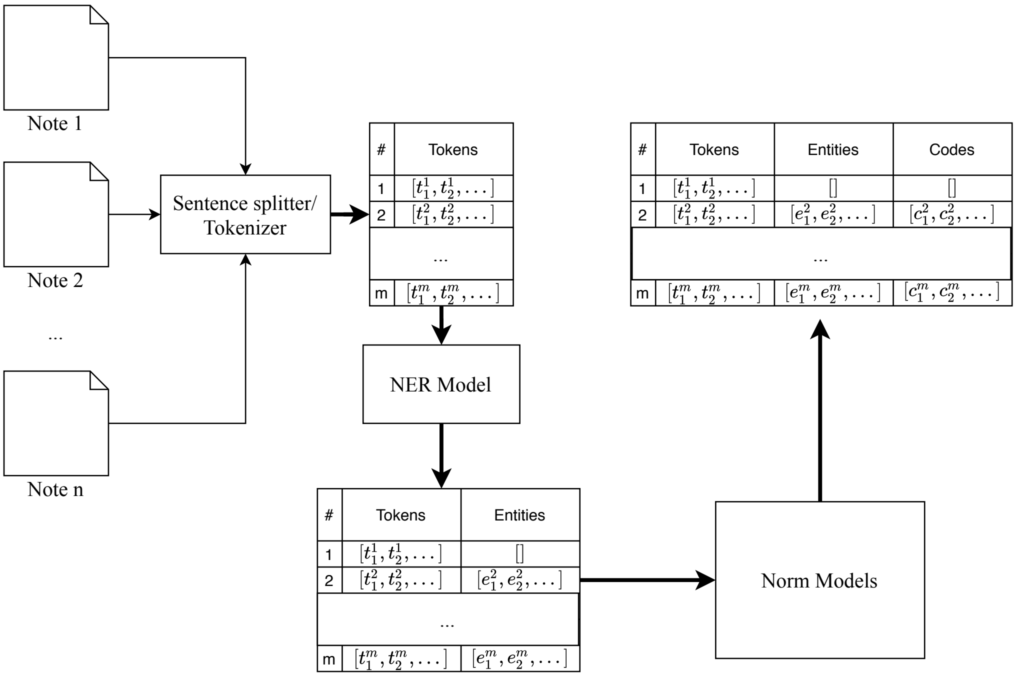

Figure 1: Data flow diagram for the three sub-tasks.

3.2. Word Embedding

After we got all the tokens from the previous step, we used word embedding to map these tokens to meaningful n-dimensional vectors. The word embedding model has been generated from Spanish medical corpora [5]. The model was built with fasttext and the dataset used to generate the model consisted of two data sources: (i) the SciELO database, which contains full-text articles primarily in English, Spanish and Portuguese, and (ii) a subset of the Wikipedia, which we call Wikipedia Health, consisting on the articles under the following categories: Pharmacology, Pharmacy, Medicine and Biology.

3.3. Data Augmentation

In addition to the available sentences in the training dataset, we increased the number of sentences with named entities by using the list of tumor entities in the file "valid-codes.txt" from the extra resources provided by the organizers. This file contains 4,203 records of tumor named entities with their corresponding eCIE-O-3.1 codes. This data augmentation was performed by choosing a random sentence containing NEs from the training dataset and replacing the existing entity by a random entity from the list of entities in the "valid-code.txt" file.

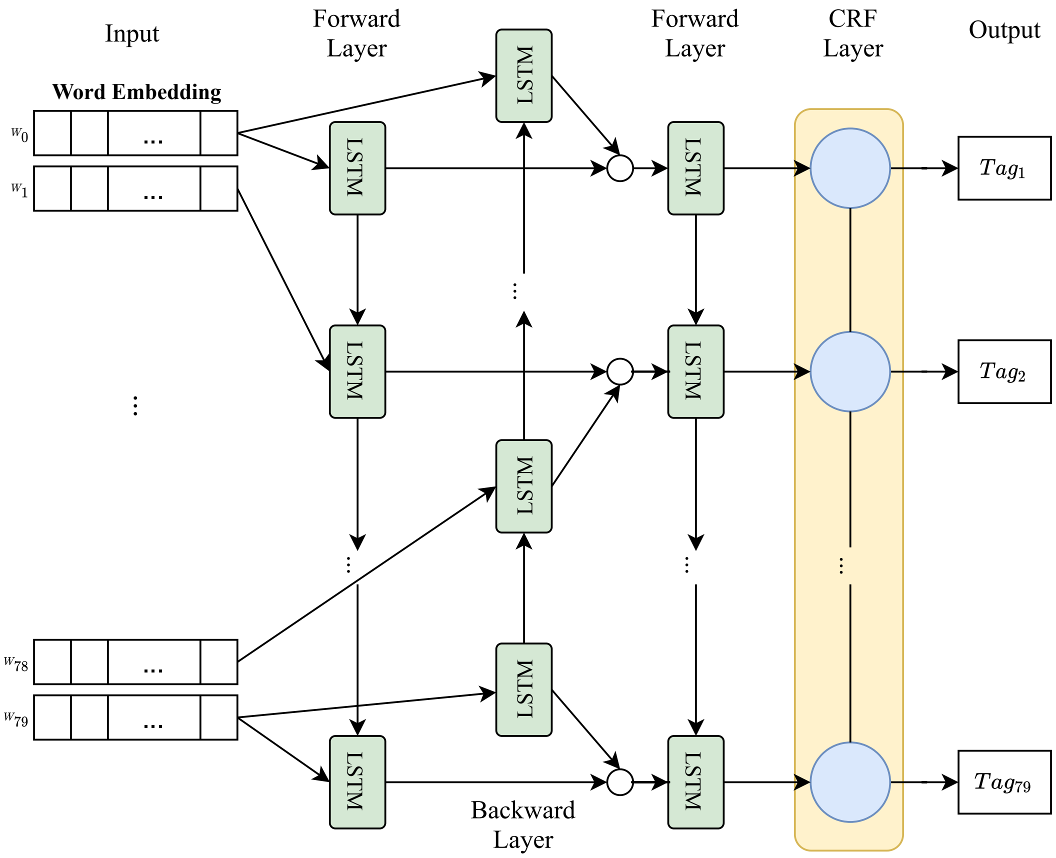

Figure 2: The architecture for the Bi-LSTM with a CRF Layer model.

3.4. Tumor Named Entity Recognition

This section describes the proposed systems for handling the tumor named entity recognition. The first subsection explains the feature-based NER using CRF, while the second subsection explains the deep learning model using Bi-LSTM neural networks.

3.4.1. Feature-Based CRF

In principle, we included this system only for comparison purposes. However, its results were comparable to those of the deep learning model, so we took advantage of that to improve the final results by combining the outcomes of the two systems.

This system was implemented using Python 3.7 with the sklearn-crfsuite package [6]. The input for this system was the tokenized sentences from the text tokenizer. The system extracts some features for every word token like the Part-of-Speach tag (POS), word prefix, word suffix, word length, etc. This is similar to what we did in our previous work in the MEDDOCAN competition [7, 8, 9]. Similarly, we also used the BIO tagging scheme to set the labels of the tokens [10]. As a result, each word token in the medical record is labeled using one of three possible tags: B, I, or O, which indicate if the word is at the beginning, middle, or outside of a Tumor entity.

3.4.2. Bi-LSTM with CRF Layer

As shown in Figure 1, the first stage in our system is the NER, which takes as input the sequence of tokens of a sentence and predicts labels using BIO tags. We choose to use Bi-LSTM as our encoder because of its ability to take all the information sequentially and pass it to the classifier. After that, we have a CRF classifier. CRF does sequence labeling for every token considering the label of neighboring tokens.

These kind of systems take as input a list of tokens with a fixed size. In our case, we choose the maximum size of 80 tokens per sentence; extra tokens were truncated as a separated sentence. If the sentence's length is less than 80, we added pre-defined padding to reach the maximum size.

As introduced before, by means of word embedding, each token is mapped to its corresponding word embedding vector. After that, all vectors are passed to the LSTM layers and, finally, the output of the LSTM is passed to the CRF layer to do the classification. The CRF layer classifies every token to one of the BIO tags.

3.5. Clinical Concept Normalization

The second and the third tasks in the competition focus on tumor named entity normalization, where every tumor NE should be mapped to their corresponding CIE-O-3.1 codes. In the following section, we describe our proposed system for these tasks.

3.5.1. CNN with LSTM Layer

The second and third sub-tasks in the competition (Norm and coding tasks) are clinical concept normalization. The input is a list of entities that are detected by the NER system, which should be mapped to the corresponding eCIE-O-3.1 codes. eCIE-O-3.1 is the Spanish version of the International Classification of Diseases for Oncology (ICD-O). This ontology aims to standardize the tumor named entities in health records and make them understandable internationally.

All the codes in the eCIE-O-3.1 come in the form of three codes separated by /. The regular expression for these codes is dddd/dd?(/H)?. The first idea that comes to the mind is to build a system that takes the NER output as input and predicts the corresponding code. However, implementing this system gave us bad results. After we analyzed why this straight forward system did not perform well, our findings were: 1) the three parts of the code vary between different tumor name entities; and 2) some of these combinations appear few times in the

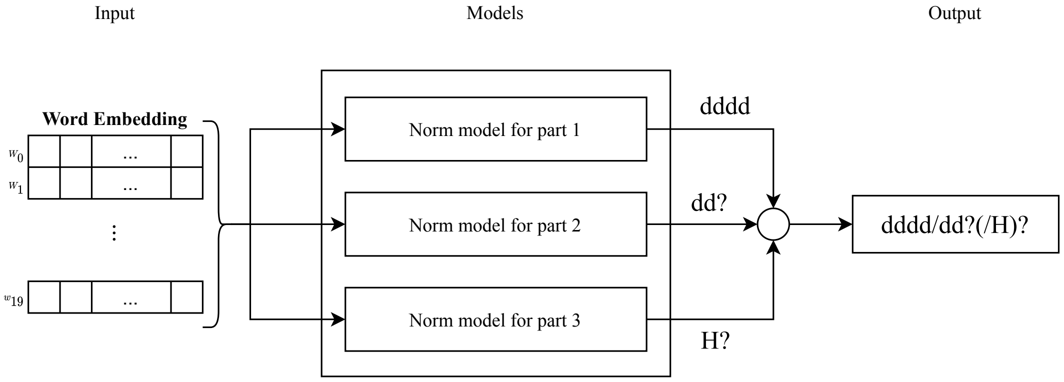

Figure 3: Overall overview of the clinical concept normalization system.

training dataset, which produces too many combinations. Because of that, we decided to treat these three parts separately, i.e. build a separate classifier for every part.

Our system contains three identical sub-systems as we mentioned before (see Figure 3 and Figure 4). The input for these three models is a list of 20 tokens (because the number of tokens in the tumor named entities is between 1 and 20). We used CNNs because these kinds of networks reduce the computation by exploiting local correlation of the input data. In our case, the tumor named entities contain in average three words, so we used CNNs with 512 filters of the size three. We do 2-max pooling to get the max two numbers for every filter response, which gives us two vectors of size 512. These two vectors were passed to the LSTM to combine them. Finally, the output was passed to the classifier, which is a fully connected layer.

4. Results and Discussion

We have proposed four systems. Table 2 provides details about the training and evaluation of each of them. NER models were trained by feeding all the sentences without exception, while Norm models were trained only on those sentences that contain NEs.

4.1. Evaluation Metrics

The standard metrics to evaluate the systems participating in this competition are the F1-score, for both NER and Norm sub-tasks, and the Mean Average Precision MAP, for the coding sub-task. Precision and Recall also are included to give more insights about the performance of the sub-models.

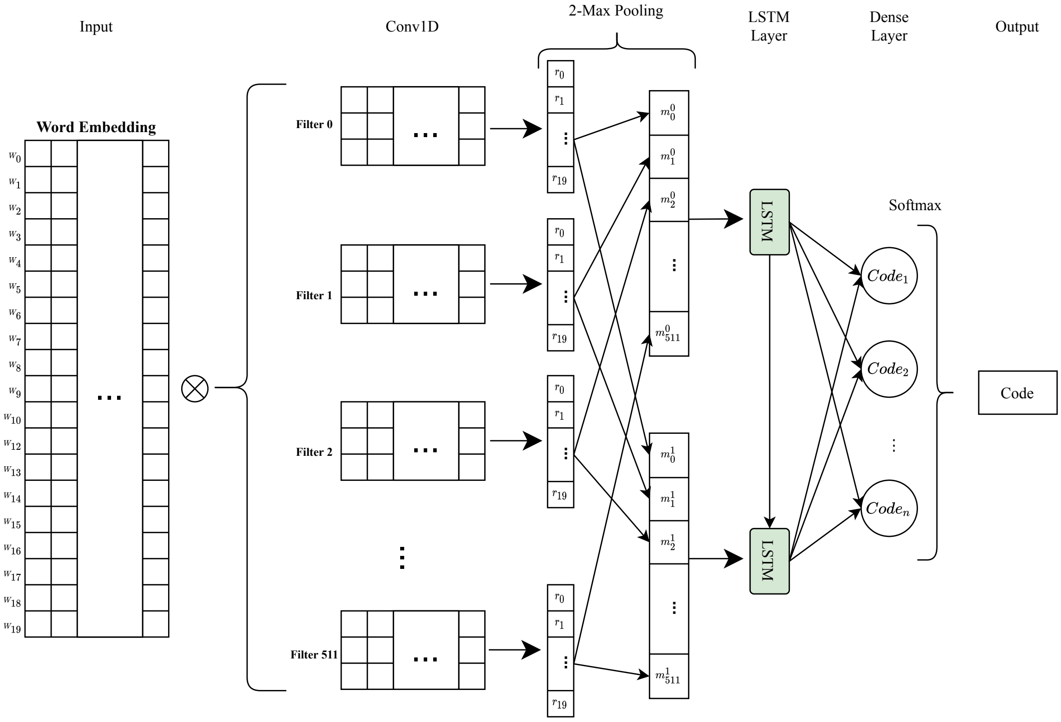

Figure 4: The architecture for sub-code models for clinical concept normalization.

System name	Tasks	Training sets	Validation
1-BILSTM	NER	train and dev2	dev1 set
	Norm & coding		
2-BILSTM	NER	train and dev1	dev2 set
	Norm & coding		
CRF	NER	train, dev1 and dev2	20%
	Norm & coding	train and dev1	dev2 set
CRF+BILSTM	NER	Voting (BILSTMs and CRF)	-
	Norm & coding	train and dev1	dev2 set

Table 2
Training and validation datasets used to train our systems.

4.2. Result

Table 3 provides a detailed comparison of the performance of the baseline system and our systems. The baseline system was provided by the organizers, which is a dictionary lookup based system. It looks for the NEs that found in train and development sets in the test set.

In the three sub-tasks, combining the three systems using voting gave better results on F1-score. However, it gave worse results on the third sub-task for the MAP metric. This happens

System name	NER task			Norm task			Coding task
	P	R	F1	P	R	F1	MAP
Baseline	18.1	73.3	29.1	18.0	73.0	28.8	58.4
Our systems							
1-BILSTM	80,7	**83,0**	81,8	76,5	**78,6**	77,6	78,3
2-BILSTM	82,4	82,4	82,4	77,9	78,0	77,9	**79,7**
CRF	80,6	77,6	79,1	77,5	74,6	76,0	77,9
CRF+BILSTM	**84,4**	81,8	**83,1**	**79,8**	77,4	**78,6**	78.7

Table 3
Performance comparison between the baseline system and our systems on the test set.

because of the code "8000/6", which appears more than the other codes. Combining the three systems helped to improve the wrongly classified samples for that code but, at the same time, caused several wrongly classified samples for the other codes.

5. Conclusion and Future Work

In this paper we describe the implementation of several systems to solve the problem of tumor named entity recognition and normalization. We compared the performance of our systems with a feature-based system. Results show that a combination of several of our systems provided the best results in most cases.

As future work, we plan to use transformer-based models like BERT [11] or XLNet [12], which they are the state of the art in the NLP field nowadays. We expect these models will give us better results on this task.

Acknowledgments

Partial support to this work has been received from the European Commission (project H2020-871042 “SoBigData++”), the Government of Catalonia (ICREA Acadèmia Prize to J.Domingo-Ferrer and grant 2017 SGR 705), the Spanish Government (projects RTI2018-095094-B-C21 “Consent” and TIN2016-80250-R “Sec-MCloud”) and the Norwegian Research Council (project no. 308904 “CLEANUP”). The authors are with the UNESCO Chair in Data Privacy, but the views in this paper are their own and are not necessarily shared by UNESCO.

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] Steward/Custodian, International classification of diseases for oncology, 3rd edition

(icd-o-3), https://www.who.int/classifications/icd/adaptations/oncology/en/, (accessed: 17.08.2020).

[3] Portal estadístico del ministerio de sanidad servicios sociales e igualdad, tabla de códigos de morfología de las neoplasias (cie-o-3.1) válidos, https://datosabiertos.castillalamancha.es/dataset/registro-de-actividad-de-atenci%C3%B3n-sanitaria-especializada-de-castilla-la-mancha-rae-clm-11, (accessed: 17.08.2020).

[4] M. Honnibal, I. Montani, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, To appear 7 (2017).

[5] 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, 2019, pp. 124–133.

[6] M. Korobov, sklearn-crfsuite, https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html, (accessed: 17.08.2020).

[7] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodriguez, J. L. Martin, M. Villegas, M. Krallinger, Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results., in: IberLEF@ SEPLN, 2019, pp. 618–638.

[8] F. Hassan, M. Jabreel, N. Maaroof, D. Sánchez, J. Domingo-Ferrer, A. Moreno, Recrf: Spanish medical document anonymization using automatically-crafted rules and crf., 2019.

[9] M. Jabreel, F. Hassan, D. Sánchez, J. Domingo-Ferrer, A. Moreno, E2ej: Anonymization of spanish medical records using end-to-end joint neural networks., 2019.

[10] E. F. Sang, J. Veenstra, Representing text chunks, arXiv preprint cs/9907006 (1999).

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[12] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: Advances in neural information processing systems, 2019, pp. 5753–5763.