VSP at MEDDOCAN 2019

De-Identification of Medical Documents in Spanish with Recurrent Neural Networks

Víctor Suárez-Paniagua

Computer Science Department, Carlos III University of Madrid.
Leganés 28911, Madrid, Spain. vspaniag@inf.uc3m.es
http://hulat.inf.uc3m.es/en/nosotros/miembros/vsuarez

Abstract. This work presents the participation in the MEDDOCAN Task of the VSP team with a neural model for the Named Entity Recognition of medical documents in Spanish. The Neural Network consists of a two-layer model that creates a feature vector for each word of the sentences. The first layer uses the character information of each word and the output is aggregated to the second layer together with its word embedding in order to create the feature vector of the word. Both layers are implemented with a bidirectional Recurrent Neural Network with LSTM cells. Moreover, a Conditional Random Field layer classifies the word vectors in one of the 29 types of Protected Health Information (PHI). The system obtains a performance of 86.01%, 87.03%, and 89,12% in F1 for the classification of the entity types, the sensitive span detection, and both tasks merged, respectively. The model shows very high and promising results being a basic approach without using pretrained word embeddings or any hand-crafted feature.

Keywords: Named Entity Recognition, Deep Learning, Recurrent Neural Network, Medical Documents

1 Introduction

Nowadays, healthcare professionals deal with a high amount of unstructured documents that makes very difficult the task of finding the essential data in medical documents. Decreasing the time-consuming task of retrieving the most relevant information can help the fastness of generating a diagnosis for patients by doctors. Instead the vast of information are available as Electronic Health Record (EHR), the manual annotation of them is impracticable because the highly increasing number of generated documents per day and also because they contain sensitive data and Protected Health Information (PHI). For this reason, the development of an automatic system that identifies sensitive information from medical documents is vital for helping doctors and preserving patient confidentiality.

The i2b2 shared task was the first Natural Language Processing (NLP) challenge for identifying PHI in the clinical narratives [13]. The second edition of the

i2b2 shared task Track 1 [12] created a gold standard dataset with annotations of the PHI categories from 1,304 medical records in English. In this competition, the highest ranking system used the Conditional Random Field (CRF) classifier together with hand-written rules for the de-identification of clinical narratives obtaining very promising results with 97.68% in F1 [14].

The goal of the Iberian Languages Evaluation Forum (IberLEF) 2019, which includes the TASS and IberEval workshops, is to create NLP challenges using corpora written in one of the Iberian languages (Spanish, Portuguese, Catalan, Basque or Galician). Following the i2b2 de-identification task, the Medical Document Anonymization task (MEDDOCAN) encourages the research community to design NLP systems for the identification of PHI from clinical texts in Spanish [9]. For this purpose, a corpus of 1,000 clinical case studies with PHI phrases was manually annotated by health documentalists.

Currently, Deep Learning approaches overcome traditional machine learning systems on the majority of NLP tasks, such as text classification [6], language modeling [10] and machine translation [1]. Moreover, these models have the advantage of automatically learn the most relevant features without defining rules by hand. Concretely, the state-of-the-art performance for Named Entity Recognition (NER) task is an LSTM-CRF Model proposed by [8]. The main idea of this system is to create a word vector representation using a bidirectional Recurrent Neural Network with LSTM cells (BiLSTM) with character information encoded in another BiLSTM layer in order to classify the tag of each word in the sentences with a CRF classifier. Following this approach, the system proposed in [3] uses a BiLSTM-CRF Model with character and word levels for the de-identification of patient notes using the i2b2 dataset. This approach overcomes the top ranking system in this task reaching to 97.88% in F1.

This paper presents the participation of the VSP team at the tasks proposed by MEDDOCAN about the classification of PHI types and the sensitive span detection from medical documents in Spanish. The proposed system follows the same approaches of [8] and [3] with some modifications for the Spanish language implemented with NeuroNER tool [2].

2 Dataset

The corpus of the MEDDOCAN task contains 1,000 clinical cases with PHI entities manually annotated by health documentalists. The documents are randomly divided into the training, validation and test sets for creating, developing and ranking the different systems, respectively.

Similarly to the annotation schema of the i2b2 de-identification tasks, the named entities are annotated according to their offsets and their type for each detection and classification (see Figure 1). The 29 types of the annotated PHI mentions follow the Health Insurance Portability and Accountability Act (HIPAA) guidelines for Spanish the health records aggregating some PHI entities.

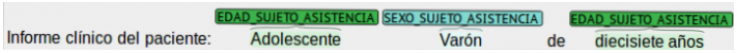

Fig. 1. The annotated offsets and types of the PHI entities in the sentence *'Informe clínico del paciente: Adolescente Varón de diecisiete años.'*. English translation: *'Clinical report of the patient: male teenager of seventeen years.'*

3 Neural model

This section presents the Neural architecture for the classification of the PHI entity types and the sensitive span detection using medical documents in Spanish. Figure 2 shows the entire process of the model using two BiLSTMs for the character and token levels in order to create each word representation until its classification by a CRF.

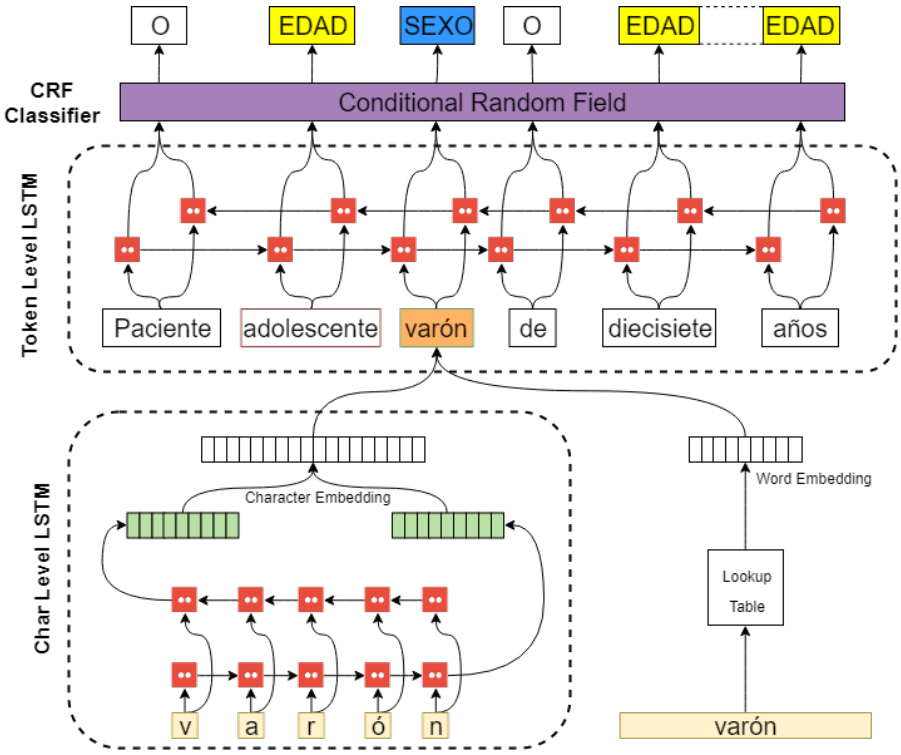

Fig. 2. Neural model for the de-identification of Medical Documents in Spanish using the MEDDOCAN task 2019 corpus.

3.1 Data preprocessing

Before using the system, the documents of the corpus are preprocessed in order to prepare the inputs for the Neural model. Firstly, the clinical cases are separated into sentences using a sentence splitter and the words of these sentences are extracted by a tokenizer, both were adapted for the Spanish language. Once the sentences are divided into word, the BIOES tag schema encodes each token with an entity type. The tag B defines the beginning token of a mention, the I tag defines the inside token of a mention, the E tag defines the ending token of a mention, the S tag indicates that the mention has a single token and the O tag indicates the outside tokens that do not belong to any mention. In many previous NER tasks, using this codification is better than the BIO tag scheme [11], but the number of labels increases because there are two additional tags for each class. Thus, the number of possible classes are the 4 tags times the 29 PHI classes and the O tag for the MEDDOCAN corpus. For the experiments, all the previous processes are performed by the spaCy tool in Python [4].

3.2 BiLSTM layers

RNNs are very effective in feature learning when the inputs are sequences. This Deep Learning model uses two different weights for the input and for the previous output as:

$$h(t) = f(\mathbf{W}x(t) + \mathbf{U}h(t-1) + b)$$

where $h(t)$ is the output at t time of the input x, f is a non-linear function, $\mathbf{W}$ are the weights for the current input, $\mathbf{U}$ are the weights for the previous output, and b the bias term of the Neural Network. However, the basic RNN cannot capture the long dependencies because it loses the information of the gradients as long as the back-propagation is applied to the previous states. For this reason, the incorporation of cell units into the RNN computation solves the long propagation of the gradient problem.

The Long Short-Term Memory cell (LSTM) [5] defines four gates for creating a word representation taking the information of the current and previous cells. The input gate i_t, the forget gate f_t and the output gate o_t for the current t step transform the input vector x_t taking the previous output h_{t-1} using its corresponding weights and bias computed with a sigmoid function. The cell state c_t takes the information given from the previous cell state c_{t-1} regulated by the forget cell and the information given from the current cell c'_t regulated by the input cell using the element-wise represented as:

$$
\begin{aligned}
f_t &= \sigma(\mathbf{W}_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(\mathbf{W}_i \cdot [h_{t-1}, x_t] + b_i) \\
c'_t &= tanh(\mathbf{W}_c \cdot [h_{t-1}, x_t] + b_c) \\
c_t &= f_t * c_{t-1} + i_t * c'_t \\
o_t &= \sigma(\mathbf{W}_o \cdot [h_{t-1}, x_t] + b_o) \\
h_t &= o_t * tanh(c_t)
\end{aligned}
$$

Finally, the current output h_t is represented with the hyperbolic function of the cell state and controlled by the output gate. Furthermore, another LSTM can be applied in the other direction from the end of the sequence to the start. Computing the two representations is beneficial for extracting the relevant features of each word because they have dependencies in both directions.

Character level The first layer takes each word of the sentences individually. These tokens are decomposed into characters that are the input of the BiLSTM. Once all the inputs are computed by the network, the last output vectors of both directions are concatenated in order to create the vector representation of the word according to its characters.

Token level The second layer takes the embedding of each word in the sentence and concatenates them with the outputs of the first BiLSTM with the character representation. In addition, a Dropout layer is applied to the word representation in order to prevent overfitting in the training phase. In this case, the outputs of each direction in one token are concatenated for the classification layer.

3.3 Contional Random Field Classifier

CRF [7] is the sequential version of the Softmax that aggregates the label predicted in the previous output as part of the input. In NER tasks, CRF shows better results than Softmax because it adds a higher probability to the correct labelled sequence. For instance, the I tag cannot be before a B tag or after a E tag by definition. For the proposed system, the CRF classifies the output vector of the BiLSTM layer with the token information in one of the classes.

4 Results and Discussion

The architecture was trained over the training set during 25 epochs with shuffled mini-batches and choosing the best performance over the validation set. The values of the two BiLSTM and CRF parameters for generating the prediction of the test set are presented in Table 1. The embeddings of the characters and words are randomly initialized and learned during the training of the network.

Table 1. The parameters of the Neural model and their values used for the MEDDOCAN results.

Parameter	**Value**
Character embeddings dimension	25
Character-level LSTM hidden units	25
Word embeddings dimension	300
Word-level LSTM hidden units	256
Optimizer	Adam
Learning rate	0.001
Dropout rate	0.5
Gradient clipping	5

Additionally, a gradient clipping keeps the weight of the network in a low range preventing the exploding gradient problem.

The results were measured with precision (P), recall (R) and F-measure (F1) using the True Positives (TP), False Positives (FP) and False Negatives (FN) for its calculation. Table 2 presents the results of the Neural Model with the two BiLSTM levels and the CRF classifier over the test set of the MEDDOCAN task. The performance over the NER offset and entity type classification (Task 1) shows an 86,01% in F1 and the performance over the sensitive token detection (Task 2) shows an 87,03% in F1 taking into consideration only if the entities have exact boundary match and entity type (Strict). Thus, the results for both tasks merged reach to 89,12% in F1.

From the table, it can be observed that the number of FN and FP are very similar giving very similar Precision and Recall results in all the classes. On the one hand, there are classes with very high performance, such as *CORREO_ELECTRONICO*, *EDAD_SUJETO_ASISTENCIA*, *FECHAS*, *NOMBRE_SUJETO_ASISTENCIA* and *PAIS* that are greater than the 95% in F1 because of the data is presented in the same location between documents and they are easy to disambiguate from the remaining classes. On the other hand, the classes of *OTROS_SUJETO_ASISTENCIA* and *PROFESION* shows a very low performance because they have a very small number of instances in the training set making hard the learning of their representation in the network. In order to alleviate this problem, the use of oversampling techniques is proposed to increase the number of instances of the less representative classes and making more balanced this dataset.

5 Conclusions and Future work

This work proposes a Neural model for the detection and classification of PHI from clinical texts in Spanish. The architecture is based on RNNs in both direction of the sentences using LSTM for the computation of the outputs. Finally, a CRF classifier performs the classification for tagging the PHI entity types. The results shows a performance of 86.01% and 87.03% in F1 for the classifi-

Table 2. Results of the Neural Model over the test set of the MEDDOCAN.

Label	**TP**	**FN**	**FP**	**R**	**P**	**F1**
CALLE	226	187	225	54,72%	50,11%	52,31%
CENTRO_SALUD	4	2	3	66,67%	57,14%	61,54%
CORREO_ELECTRONICO	244	5	7	97,99%	97,21%	97,6%
EDAD_SUJETO_ASISTENCIA	504	14	37	97,3%	93,16%	95,18%
FAMILIARES_SUJETO_ASISTENCIA	55	26	44	67,9%	55,56%	61,11%
FECHAS	585	26	25	95,74%	95,9%	95,82%
HOSPITAL	102	28	35	78,46%	74,45%	76,4%
ID_ASEGURAMIENTO	184	14	22	92,93%	89,32%	91,09%
ID_CONTACTO_ASISTENCIAL	35	4	4	89,74%	89,74%	89,74%
ID_SUJETO_ASISTENCIA	251	32	37	88,69%	87,15%	87,92%
ID_TITULACION_PERSONAL_SANITARIO	217	17	19	92,74%	91,95%	92,34%
INSTITUCION	29	38	31	43,28%	48,33%	45,67%
NOMBRE_PERSONAL_SANITARIO	468	33	26	93,41%	94,74%	94,07%
NOMBRE_SUJETO_ASISTENCIA	502	0	4	100%	99,21%	99,6%
NUMERO_FAX	5	2	2	71,43%	71,43%	71,43%
NUMERO_TELEFONO	20	6	3	76,92%	86,96%	81,63%
OTROS_SUJETO_ASISTENCIA	0	7	1	0%	0%	0%
PAIS	350	13	5	96,42%	98,59%	97,49%
PROFESION	2	7	5	22,22%	28,57%	25%
SEXO_SUJETO_ASISTENCIA	228	233	232	49,46%	49,57%	49,51%
TERRITORIO	885	71	61	92,57%	93,55%	93,06%
Task 1	4896	765	828	86,49%	85,53%	86,01%
Task 2 (Strict)	-	-	-	87,51%	86,55%	87,03%
Task 2 (Merged)	-	-	-	89,36%	88,88%	89,12%

cation of the entity types and the sensitive span detection over the MEDDOCAN corpus giving 89,12% in F1 for the merged tasks as the official result. The results are very similar in Precision and Recall for all the classes giving a low performance in the less representative classes and a higher performance in the well-structured PHI entities, such as *NOMBRE_SUJETO_ASISTENCIA* *EDAD_SUJETO_ASISTENCIA*, *CORREO_ELECTRONICO*, *FECHAS*, and *PAIS*.

As future work, exploring the contribution of each representation individually and fine-tuning the parameters of the model will be useful in order to increase the performance. In addition, the aggregation of embeddings from different external information, such as Part-of-Speech tags, syntactic parse trees or semantic tags, could increase the representation of each word for improving its classification. Moreover, the sentence splitter of spaCy seems to divide sentences when some acronyms appear, such as 'Dr.', 'Dra.', 'Sr.' or 'Sra.' (Spanish honorific prefix). For this reason, the creation of simple rules in order to avoid these cases could be beneficial for increasing the performance. Furthermore, adding more layers to each BiLSTM is proposed to be included in the architecture.

References

1. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1179, https://www.aclweb.org/anthology/D14-1179
2. Dernoncourt, F., Lee, J.Y., Szolovits, P.: NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 97–102. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-2017, https://www.aclweb.org/anthology/D17-2017
3. Dernoncourt, F., Young Lee, J., Uzuner, O., Szolovits, P.: De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association : JAMIA **24** (06 2016). https://doi.org/10.1093/jamia/ocw156
4. Explosion AI: spaCy - Industrial-strength Natural Language Processing in Python (2017), https://spacy.io/
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
6. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751 (2014)
7. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data pp. 282–289 (2001), http://dl.acm.org/citation.cfm?id=645530.655813
8. 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
9. Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodríguez, H., Lopez Martin, J.A., Villegas, M., Krallinger, M.: 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). vol. TBA, p. TBA. CEUR Workshop Proceedings (CEUR-WS.org), Bilbao, Spain (Sep 2019), TBA
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
11. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009). pp. 147–155. Association for Computational Linguistics (2009), http://aclweb.org/anthology/W09-1119
12. Stubbs, A., Kotfila, C., Uzuner, O.: Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. Journal of biomedical informatics **58** (07 2015). https://doi.org/10.1016/j.jbi.2015.06.007

13. Özlem Uzuner, Luo, Y., Szolovits, P.: Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association **14**(5), 550 – 563 (2007). https://doi.org/https://doi.org/10.1197/jamia.M2444, http://www.sciencedirect.com/science/article/pii/S106750270700179X
14. Yang, H., Garibaldi, J.M.: Automatic detection of protected health information from clinic narratives. Journal of Biomedical Informatics **58**, S30–S38 (2015)