De-Identification through Named Entity Recognition for Medical Document Anonymization

Hermenegildo Fabregat[1], **Andres Duque**[2,3], **Juan Martinez-Romo**[1,3], and **Lourdes Araujo**[1,3]

[1]NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos
[2]Departamento de Sistemas de Comunicación y Control
[1,2]Universidad Nacional de Educación a Distancia (UNED)
[3]Instituto Mixto de Investigación - Escuela Nacional de Sanidad (IMIENS)
gildo.fabregat@lsi.uned.es, aduque@scc.uned.es, juaner@lsi.uned.es, lurdes@lsi.uned.es

Abstract. This paper introduces the system developed by the NLP_UNED team participating in MEDDOCAN (Medical Document Anonymization) task, framed in the IberLEF 2019 evaluation workshop. The system DINER (De-Identification through Named Entity Recognition) consists of a deep neural network based on a core BI-LSTM structure. Input features have been modeled in order to suit the particular characteristics of medical texts, and especially medical reports, which can combine short semi-structured information with long free text paragraphs. The first results of the system on a synthetic test corpus corpus of 1000 clinical cases, manually annotated by health documentalists, indicate the potential of the DINER system.

Keywords: named entities recognition · deep learning · medical document anonymization · electronic health record · natural language processing

1 Introduction

Nowadays, the use of digitized medical records of patients has allowed progress in biomedical research in different areas of interest through the use of natural language processing techniques. However, one of the main problems in distributing these records is the personal information that appears in them. These records are stored with great security measures to prevent them from being public and their anonymization is a challenge that still has a long way to go. In response to this need, the MEDDOCAN task [8] has been organized, oriented to the anonymization of medical documents with protected health information (PHI) within IberLEF 2019 (Iberian Languages Evaluation Forum).

This task has distributed a corpus of 1000 documents with medical records artificially created. This corpus was manually selected by a practicing physician and augmented with information from discharge reports and genetic medical records. MEDDOCAN in turn is composed of two subtasks; one of them whose objective is the identification and classification of named entities; and the other subtask that focuses on the detection of sensitive tokens.

This paper describes the participation of the NLP_ UNED team using the DINER system in the MEDDOCAN task.

The importance of anonymization or de-identification of clinical texts has been addressed in the past in two shared tasks, the 2006 uzuner2007evaluating and 2014 stubbs2015automated de-identification tracks, organized by the i2b2 tranSMART Foundation[1].

In addition, the problem of de-identification has been addressed recently by neural networks as in the work by [2] where this kind of systems is used for the first time on the corpus i2b2 2014, outperforming the state of the art systems. The work by [6] presents another approach based on LSTM (long-short term memory). The LSTM model consists of three layers: input layer — generates a vectorial representation of each word of a sentence; LSTM layer — outputs another word representation sequence that captures the context information of each word in this sentence; Inference layer — makes tagging decisions according to the output of LSTM layer, that is, outputting a label sequence. In the work by [5] have developed a hybrid system composed of four individual subsystems, that is, a subsystem based on bidirectional LSTM, a subsystem-based on bidirectional LSTM with features, a subsystem based on conditional random fields (CRF) and a rule-based subsystem, are used to identify PHI instances. Then, an ensemble learning-based classifier was deployed to combine all PHI instances predicted by the above three machine learning-based subsystems. Finally, the results of the ensemble learning-based classifier and the rule-based subsystem are merged together. Finally in [10] authors propose an algorithm based on a deep learning architecture. Authors implement and compare different variants of the RNN architecture, including Elman and Jordan and a CRF based model with the traditional features. They observe that the variants of the RNN architecture outperform the baseline built using a popular CRF based model.

2 System description

The proposed system for the detection of PHI consists of: 1) a pre-processing phase where data is adapted and prepared, 2) a supervised learning model based on deep learning and 3) a phase of application of rules for the correcting of recurrent errors.

2.1 Pre-processing

The corpus has been pre-processed and re-annotated following the BILOU annotation schema [9]. The BILOU scheme challenges classifiers to learn the Start, Inside and Last token of the different annotations, as well as the Unit length segments.

[1] www.i2b2.org

Regarding the tokenization of each document, a sentence splitter was tested using CoreNLP [7], but having obtained worse results and lack of coherence in the BILOU format this splitter was discarded. Instead, basic split of the documents was performed, by just taking into account the line breaks and a maximum sentence size of 150 words. In the case of larger sentences only the first 150 words are labelled and those of shorter length are filled in using a padding approach.

After the pre-processing phase, a total of 79 classes have been obtained (BILOU Annotation + Type of entity).

2.2 Supervised Learning Model

In this section we present the supervised learning model and the different attributes considered to be the input of the deep learning stack. Four attributes have been used and are as follows:

Words A representation based on pre-trained word embeddings has been used. The word vectors presented in [1] have been selected due to the richness of the sources from which they were generated and to their high recall. These vectors have a total of 300 dimensions and gather around 1,000,653 unique tokens.

Part-of-speech This feature has been used due to its importance in different natural language processing tasks. The PoS-Tagging model used was the one provided by the CoreNLP [7] library for Spanish. This feature has been represented in the model by embeddings generated during training. The resulting embeddings consist of 25 dimensions.

Casing This feature satisfies the need to minimize the impact of the simplification process applied to complex expressions found in the different instances. This is achieved by modeling each term with an additional 8-position one-hot vector which represents different cases: term ending in comma or in dot, uppercased first letter or uppercased term, digits within the term, etc.

Chars Each term has been modeled by means of a representation based on character embeddings since a complex tokenization has not been applied (only a space splitting, trying to respect the offset in every case). The aim of this technique is to increase the recall of the word embeddings also used as feature. As can be seen in the following example, making use of this feature, cases without a char embedding are represented.

Example:
“Nombre:Pedro Garca-Lopez” ⇒
“nombrepedro” “garcialopez”

As you can see in the previous example, after tokenizing by spaces and removing alphanumeric characters, the resulting token sequence is not interpretable in most cases by word embeddings. These char embeddings have been trained on the corpus.

Deep Learning model The model implemented for PHI detection, as shown in the Figure 1, consists of a Bi-LSTM (Bidirectional Long short-term memory) followed by two Dense layers. The inputs of the architecture are the vectors C, P, W, and CH that represent the information of Casing, Pos-Tag, word, and characters per word respectively. The last dense layer corresponds to the output layer.

The convolutional model proposed by [11] has been used for character sequence processing.

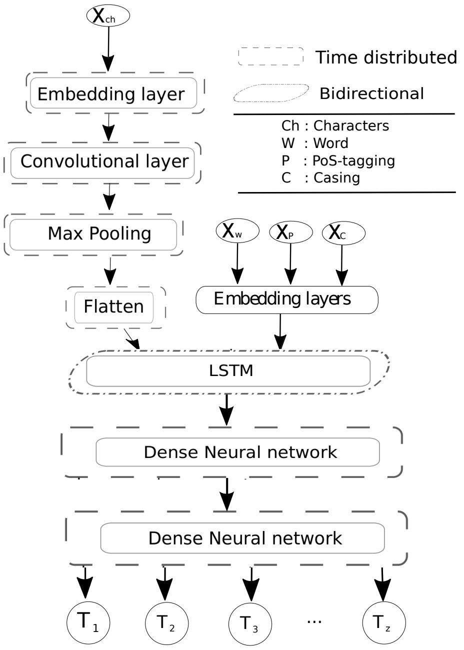

Fig. 1. Architecture of the proposed model for the extraction of PHIs

Convolutional layer As Figure 1 shows this model applies a series of convolutions and stacks in order to extract the most important characteristics of the sequence of characters and reduce the dimensionality of the resulting vector. As a result, a summary vector is obtained with the information obtained for each word.

Bi-LSTM. LSTMs [4] are proven to offer good performance in sequential NLP tasks. This layer responds to the need to process each term according to its context. Each LSTM is configured with 150 neurons and a ReLU [3] as an activation function. In order to avoid over-adjustment, dropouts of 0.5 and recurrent dropouts of 0.3 have been applied.

Dense (middle). This layer has been added in order to simplify the information generated by the previous layers, thus reducing the solution space in subsequent layers.

Dense (output). The output layer is configured with 79 neurons and a softmax activation function.

Hyper-parameters The hyper-parameters used in the system are the following:

- Maximum size of instance and word: 150 words and 50 characters.
- Char embedding of 50 dimensions.
- Postag embedding of 25 dimensions.
- Word embedding of 300 dimensions.
- The Bi-LSTM layer is composed of two LSTM of 150 dimensions.
- Each LSTM layer uses a dropout of 0.25 and a recurrent-dropout of 0.15.
- Dense layer of 50 dimensions.

2.3 Rules

Two types of rules have been applied to the output of the DL architecture, both for error correction. On the one hand, the first set of rules is oriented to the correction of frequent errors. The vast majority of rules of this type that have been applied aim to increase accuracy by filtering out those cases that do not meet a certain format. As an example, there is the format that telephone numbers must comply with (minimum 9 numbers, country codes being optional) and the format that e-mails must comply with (at least it must contain an @).

On the other hand, the second set of rules aims to ensure that the final output of the system correctly follows the output BILOU format.

Results on the Development Set We have carried out a set of experiments in order to analyze the effect of the use of the implemented rules. Tables 1, 2, and 2.3 show the results on the development set using the system without and with the use of rules.

As the tables show, the use of rules improves the system performance in the two tasks proposed in MEDDOCAN. This improvement is reflected in an increase between 0.023 and 0.025 points in the F1-measure. Analyzing the results in more depth, one can also appreciate how the effect of the rules produces a higher improvement in the precision than in the recall.

Error analysis After an in-depth analysis of the system's performance, we detected a couple of errors that the system usually makes. The errors are as follows:

- The system makes labeling errors when confusing the hospital address with the name of the hospital itself.
- The system is not able to identify some of the expressions of the class "Familiar_sujeto_asistencia".

Development Set			
Task 1			
System	P	R	F1
No Rules	0.933	0.911	0.922
With Rules	0.964	0.930	0.947

Table 1. Results from Task 1 on the development set. F1 stands for F1 measure, P for precision and R for recall.

Development Set			
Task 2 (Strict)			
System	P	R	F1
No Rules	0.938	0.916	0.927
With Rules	0.968	0.934	0.951

Table 2. Results from Task 2 (Strict) on the development set. F1 stands for F1 measure, P for precision and R for recall.

3 Results

This section shows the official results of the DINER system in the two tasks organised in MEDDOCAN. At the time of writing this paper there were no baselines or results available from other participants, so only the results of participation of the DINER system are shown.

Regarding the task, 18 teams participated, submitting a total number of 63 system runs.

Taking into account the results on the development set shown in the previous section, our system has used rules for all the MEDDOCAN sub-tasks. Our team only sent one run for each task and therefore only this information is shown in the results.

Tables 4, 5, and 6 show the official results on the test set.

As the tables show, the results between the development and the test set are very similar. On the one hand, this fact reflects the correct split of the corpus into different sets and on the other hand the robustness of the DINER system.

The system was trained for the task 1, so such optimal results in task 2 are the result of a balanced system. The fact that the results in the two evaluations

Development Set			
Task 2 (Merged)			
System	P	R	F1
No Rules	0.948	0.925	0.936
With Rules	0.976	0.941	0.959

Table 3. Results from Task 2 (Merged) on the development set. F1 stands for F1 measure, P for precision and R for recall.

of the task 2 (Strict and Merged) are so similar to each other, means that the implemented rules have provided a great coherence to the system.

On the other hand, the differences between these two evaluations could be due to the fact that in the documents of the corpus some entities could be written in a free format. In this way, spaces between the digits of a telephone number or the format of an e-mail could be the reason for the difference between these two evaluations of the task 2.

Official Results				
Task 1				
System	**Leak**	**P**	**R**	**F1**
NLP_UNED	0.054	0.959	0.928	0.943

Table 4. Results from Task 1. F1 stands for F1 measure, P for precision and R for recall.

Official Results			
Task 2 (Strict)			
System	**P**	**R**	**F1**
NLP_UNED	0.964	0.934	0.949

Table 5. Results from Task 2 (Strict). F1 stands for F1 measure, P for precision and R for recall.

Official Results			
Task 2 (Merged)			
System	**P**	**R**	**F1**
NLP_UNED	0.973	0.942	0.957

Table 6. Results from Task 2 (Merged). F1 stands for F1 measure, P for precision and R for recall.

4 Conclusions

In this paper we have described our system DINER, and its performance in the MEDDOCAN task of the IberLEF 2019 competition. The proposed system is divided into two phases, the first of them making use of deep neural networks, and the second one using hand-crafted rules.

The DINER system has obtained a score of 0.943 in the F1-measure for the task 1 and 0.949 in the F-measure for task 2 (strict evaluation).

In spite of not knowing the performance of other systems or a baseline, we could say that at least the performance of the DINER system is optimal. One of the characteristics of the system could be its robustness since its performance has been very similar between the development and test sets. On the other hand, we would like to highlight the coherence provided to the system by the use of rules.

We plan to address improvements in the PHI extraction as a future line of work, especially by studying how more valuable syntactic and semantic information can be added to the network that performs PHI identification, and also how systematic post-processing rules can be automatically extracted from the obtained results.

Agreements

This work has been partially supported by the Spanish Ministry of Science and Innovation within the projects PROSA-MED (TIN2016-77820-C3-2-R) and EXTRAE (IMIENS 2017).

References

1. Cardellino, C.: Spanish billion words corpus and embeddings (march 2016). URL http://crscardellino. me/SBWCE (2016)
2. Dernoncourt, F., Lee, J.Y., Uzuner, O., Szolovits, P.: De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association **24**(3), 596–606 (2017)
3. Hahnloser, R.H., Sarpeshkar, R., Mahowald, M.A., Douglas, R.J., Seung, H.S.: Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature **405**(6789), 947 (2000)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
5. Liu, Z., Tang, B., Wang, X., Chen, Q.: De-identification of clinical notes via recurrent neural network and conditional random field. Journal of biomedical informatics **75**, S34–S42 (2017)
6. Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z., Xu, H.: Entity recognition from clinical texts via recurrent neural network. BMC medical informatics and decision making **17**(2), 67 (2017)
7. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)
8. Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodrguez, 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
9. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the thirteenth conference on computational natural language learning. pp. 147–155. Association for Computational Linguistics (2009)
10. Yadav, S., Ekbal, A., Saha, S., Bhattacharyya, P.: Deep learning architecture for patient data de-identification in clinical records. In: Proceedings of the clinical natural language processing workshop (ClinicalNLP). pp. 32–41 (2016)
11. Zhang, X., Zhao, J.J., LeCun, Y.: Character-level convolutional networks for text classification. CoRR **abs/1509.01626** (2015), http://arxiv.org/abs/1509.01626