Protected Health Information Recognition byBiLSTM-CRF

Cristóbal Colón-Ruiz[0000−0002−9167−809X] and Isabel Segura-Bedmar[0000−0002−7810−2360]

Universidad Carlos III de Madrid, Leganés, Spain
{ccolon,isegura}@inf.uc3m.es
http://hulat.inf.uc3m.es/

Abstract. Medical records contain relevant information about patients, which can be beneficial to improving healthcare and research in the clinical domain. Due to this, there is a growing interest in developing automatic methods to extract and exploit the information from medical records. However, medical records also contain protected health information about patients. To protect the confidentiality and privacy of patients, this sensitive information should be removed prior to any processing of these documents. In this paper, we describe an architecture for the detection and identification of protected health information from medical records. The architecture is composed of two bidirectional Long Short-Term Memory layers and a final layer based on Conditional Random Fields. Our system participated in the Meddocan shared task, obtaining a micro-F1 of 93.22%.

Keywords: Anonymization· De-identification · Deep Learning · Long Short-Term Memory

1 Introduction

In recent years, the number of electronic medical records (EHRs) has been increasing massively. These registries are very useful resources to perform studies focused on detection/prevention of diseases and medical decision making, among others. However, health records contain protected health information (PHI). For instance, information about family history, treatment tracking, and data that may help to identify a given patient (for example, patient's name, address, telephone number, zip code, etc). Due to this protected information, medical records cannot be shared without a previous de-identification process. This process consists of detecting and subsequently replacing or removing all protected information from the records.

The interest in addressing de-identification problems motivated the proposal of two tasks, the 2006 [12] and 2014 [10] de-identification tracks, organized by Integrating Biology and the Bedside (i2b2). These tracks have significantly influenced the Natural Language Processing (NLP) community in the medical field, and in particular, for the task of automated text anonymization. Nevertheless,

the tasks only focused on records written in English. The Meddocan 2019 has proposed the first task for the anonymization of medical records in Spanish [6].

The identification of PHI inside medical documents can be addressed as a Named Entity Recognition (NER) problem. This problem has been widely studied and the approaches normally used can be divided into several main categories: dictionaries and rule-based systems, machine learning, deep learning, and hybrid systems. Dictionary-based methods are limited by the size of the dictionaries themselves, in addition to the constant growth of vocabulary and spelling errors. Rule-based approaches usually provide high precision, however, they do not usually contemplate all existing cases as a result of the complexity of the language. Furthermore, rule-based and machine learning methods require a previous generation of syntactic and semantic features, as well as domain-specific information. Approaches based on deep learning methods automatically learn relevant patterns, allowing a certain grade of independence of language and domain. Moreover, these approaches have been shown to achieve better results than the best hybrid systems in i2b2 tasks. The system described in [2], which was based on Long-short Term Memory (LSTM) layers combined with Conditional Random Field (CRF) layers, scored 97.87% of F1 surpassing the winning ib2b 2014 system [13] with 96.11%, which was based on a hybrid model combining conditional random fields with keyword and rule-based approaches.

Considering the above, in this paper, we propose the use of an adaptation of the NeuroNer tool [1] for the sub-tasks 1 and 2 of MEDDOCAN-2019 on Spanish records. This tool uses the combination of two bidirectional Long-short Term Memory (BiLSTM) layers with a final Conditional Random Fields layer. The rest of the paper is organized as follows. Section 2 briefly describes the datasets provided for the MEDDOCAN-2019 task. In Section 3, we describe the architecture of our system. Section 4 presents the results obtained for our system. In Section 5, we provide the conclusions.

2 Dataset

The training and development sets provided in the MEDDOCAN-2019 task are composed of 500 and 250 clinical cases respectively and contain 22 different types of PHI entities listed in Table 5. In both sets, the representation of the different types of entity is proportional and unbalanced.

Table 1. Number of clinical cases and PHI entities.

	Number of clinical cases	Number of PHI entities
Training set	500	23618
Development set	250	12180

The clinical cases are initially provided in BRAT format[1], a standoff format where the different annotations are stored separately from the original text in a similar way to the BioNLP Shared Task standoff format[2].

3 Methods and system description

3.1 Pre-processing

We pre-process the text of the clinical cases taking into account different steps. First, sentences are split using the Spacy (Spacy.io), an open-source library that provides support for texts in several languages, including Spanish. Due to the nature of the language used in this type of text, we have defined a set of rules to avoid detecting dots corresponding to acronyms or abbreviations as sentence separators.

Afterward, the resulting sentences are tokenized by using Spacy. However, some tokens meet the following pattern "field:value". For instance, the token "cp:28007", which refers to a postal code, should be split into the corresponding field ("cp") and its value ("28007"). Due to this, we have included a set of rules to correctly split this kind of tokens.

Finally, the text and its annotations are transformed into the CoNLL-2003 format[3] using the BIOES schema [9]. In this schema, tokens are annotated using the following tags: The B tag represents a token that is the beginning of an entity, The I tag indicates that the token belongs to an entity, the O tag represents that the token does not belong to any entity, the E tag marks a token as the end of a given entity, and the S tag indicates that an entity is comprised of a single token.

3.2 Network description

As we can appreciate in section 1, approaches composed of Bidirectional LSTMs layers in conjunction with CRF, provide good results in NER tasks [5, 2]. Bidirectional LSTMs are a type of recurrent neural network (RNN) that takes into account the context of words in the sentence by capturing past (previous words) and future (next words) information. In addition, to improve the accuracy of predictions provided by the BiLSTM layer, the CRF layer uses information from the neighbor tags (at sentence level) in order to predict current tags. Considering the above, in order to address the de-identification problem described in the MEDDOCAN-2019 task, we propose to use the NeuroNer tool [1], which is composed of three main layers:

1. Token representation with character-enhanced and token embedding layer.
2. BiLSTM prediction layer

[1] http://brat.nlplab.org/standoff.html
[2] http://2011.bionlp-st.org/home/file-formats
[3] https://www.clips.uantwerpen.be/conll2003/ner/

3. CRF sequence optimization layer

The first layer aims to generate vector representations of the tokens that conform the input sequences. The direct representation of token to vector (word embedding) can be pre-trained or can be learned in conjunction with the rest of the model by adjusting its weights. Pre-trained models can be obtained from a large amount of unlabeled data with methods such as word2vec or GloVe [7, 8]. However, the different word embedding models do not contain representation for those tokens not included in their vocabularies. The first layer addresses this problem by incorporating a representation of tokens based on their characters (character embeddings). Each token character is represented by its own vector, allowing the network to learn morphological information even from tokens that are not included in the vocabulary of the word embedding model [4].

The character embedding sequence of each token is passed as input to a BiLSTM to obtain character-based word embedding as output. Finally, the representation of word embeddings and character-based word embedding are concatenated for each token, which will be the input for the second BiLSTM layer. This BiLSTM layer aims to obtain the sequence of probabilities for each token to pertain to a given label using the BIOES coding. The label for each token will be the one with the highest probability.

The last layer consists of a conditional random fields layer. This layer receives as input the sequence of probabilities of the previous layer in order to improve predictions. This is due to the ability of the layer to take into account the dependencies between the different labels. The output of this layer provides the most probable sequence of labels.

The parameters of the embedding and hyperparameters of our model used for the MEDDOCAN-2019 task are listed below:

- **Word Embeddings**: randomly initialized and adjusted during training. The dimension of the vectors is 100.
- **Character Embeddings dimension**: randomly initialized and adjusted during training. The dimension of the vectors is 25.
- **First BiLSTM hidden state dimension**: 25 for the forward and backward layers
- **Second BiLSTM hidden state dimension**: 100 for the forward and backward layers
- **Optimizer**: Stochastic gradient descent (SGD), learning rate: 0.01
- **Dropout**: 0.5
- **Number of Epochs**: 100

4 Results

The MEDDOCAN-2019 evaluation process consists of two different scenarios. The first scenario consists of detecting exactly the location in the text of each PHI, as well as the type of entity. The second scenario consists of identifying

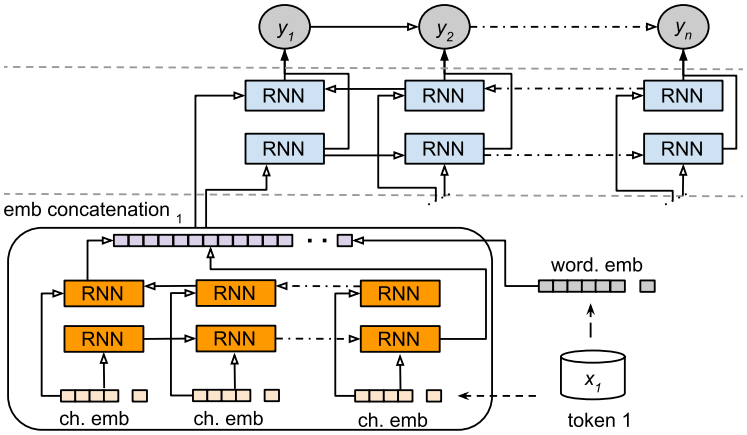

Fig. 1. Overview architecture of our system

sensitive data in order to be replaced, regardless of the type of entity. In the last scenario, two different types of evaluations are performed: (1) Strict evaluation of the spans of PHI that belong to sensitive phrases. (2) Merge evaluation where the spans of PHI connected by non-alphanumerical characters are merged. To evaluate both scenarios, the metrics proposed by the organizers are micro-averaged precision, recall, and F1-score.

To evaluate the trained models, as well as their hyperparameters, we performed a set of experiments with the development dataset provided by the MEDDOCAN-2019 organizers. We used grid search to adjust the word embeddings dimension, the number of units in the BiLSTM hidden layer, the optimizer and the learning rate.

We can observe in Tables 2 and 3 that our best results on the overall development set do not present significant differences. The run0 model was trained using the hyperparameters mentioned in section 3.2. The run1 model was trained using ADAM [3] as optimizer, with a word embeddings dimension of 300 and 200 units in the second layer of the BiLSTM. The run2 model was trained using ADAM employing a word embeddings dimension of 200 and 100 units in the second layer of the BiLSTM.

Table 2. Results of Subtask 1 on development set . The top scores are bold.

	Subtask 1		
	Precision	Recall	F1
run 0	0.9435	0.9279	**0.9356**
run 1	0.9391	0.9284	0.9337
run 2	0.9364	0.9267	0.9315

Table 3. Results of Subtask 2 on development set . The top scores are bold.

	Subtask 2 Strict			Subtask 2 Merged		
	Precision	Recall	F1	Precision	Recall	F1
run 0	0.9507	0.9350	**0.9428**	0.9635	0.9465	0.9549
run 1	0.9461	0.9353	0.9407	0.9649	0.9519	**0.9584**
run 2	0.9449	0.9351	0.9400	0.9590	0.9463	0.9526

Considering that both the models run0 and run1 achieved the best results in the two subtasks, these models were used to process the test set provided by MEDDOCAN-2019 in scenarios 1 and 2. The results obtained in both tasks can be seen in Table 4. Moreover, we can verify that the run0 model is the one that provides the best results in all scenarios (F1 of 93.22% in sub-task 1, F1 of 94.26% in sub-task 2 for strict evaluation and F1 of 95.77% for merged evaluation).

Table 4. Results of all subtasks on the test set . The top scores are bold.

	run0			run1		
	Precision	Recall	F1	Precision	Recall	F1
Task 1	0.9365	0.9279	**0.9322**	0.927	0.9309	0.9289
Task 2 Strict	0.9470	0.9383	**0.9426**	0.9365	0.9404	0.9384
Task 2 Merged	0.9630	0.9524	**0.9577**	0.9564	0.9563	0.9563

Once the run 0 model has been selected as our best model, it is interesting to perform a more detailed study for each of the PHI due to the unbalance of the problem. As we can see in Table 5, not all types of entities are classified so easily. For example, entities such as ID_EMPLEO_PERSONAL_SANITARIO and OTROS_SUJETO_ASISTENCIA are never correctly classified. This may be due to their low representation in the training set, as well as the confusion it may cause with other similar types of entities such as FAMILIARES_SUJETO_ASISTENCIA. On the other hand, the types of entity with the best results are those with the highest representation in the training set or those with a specific structure such as ID_ASEGURAMIENTO and ID_CONTACTO_ASISTENCIAL.

5 Conclusions

Medical records are sources of a wide range of studies to detect and prevent diseases, as well as to provide information for medical decision making. The growing volume of these types of records, in addition to the studies associated with them, increases the importance of de-identification. This type of process

Table 5. Results of subtask 1 on the test set. Micro precision, recall and F1 for entity.

	Precision	Recall	F1
CALLE	0.825287	0.869249	0.846698
CENTRO_SALUD	0.250000	0.166667	0.200000
CORREO_ELECTRONICO	0.948000	0.951807	0.949900
EDAD_SUJETO_ASISTENCIA	0.980315	0.961390	0.970760
FAMILIARES_SUJETO_ASISTENCIA	0.658537	0.666667	0.662577
FECHAS	0.964111	0.967267	0.965686
HOSPITAL	0.873950	0.800000	0.835341
ID_ASEGURAMIENTO	0.929293	0.929293	0.929293
ID_CONTACTO_ASISTENCIAL	0.795455	0.897436	0.843373
ID_EMPLEO_PERSONAL_SANITARIO	0.000000	0.000000	0.000000
ID_SUJETO_ASISTENCIA	0.964413	0.957597	0.960993
ID_TITULACION_PERSONAL_SANITARIO	0.931330	0.927350	0.929336
INSTITUCION	0.290909	0.238806	0.262295
NOMBRE_PERSONAL_SANITARIO	0.937500	0.928144	0.932798
NOMBRE_SUJETO_ASISTENCIA	0.994024	0.994024	0.994024
NUMERO_FAX	0.666667	0.571429	0.615385
NUMERO_TELEFONO	0.645161	0.769231	0.701754
OTROS_SUJETO_ASISTENCIA	0.000000	0.000000	0.000000
PAIS	0.969014	0.947658	0.958217
PROFESION	0.571429	0.444444	0.500000
SEXO_SUJETO_ASISTENCIA	0.982646	0.982646	0.982646
TERRITORIO	0.966595	0.938285	0.952229

has the goal of eliminating or replacing sensitive data to allow access to medical information without compromising the identification of the patients involved.

Most previous efforts in the task of anonymization medical records have been focused mostly on texts written in English. Meddocan is the first shared task devoted to the anonymization of medical records in Spanish. One of the major challenges of this shared task is that there are a large number of sensitive data categories (22 different types of entities). In addition, these data are often unbalanced in the text. This results in difficulties to classify them correctly.

In this paper, we describe our participating system in this task. It exploits the NeuroNer tool, a tool based on deep learning with bi-directional LSTM and CRF layers for the task of NER. In spite of the challenges above described, our system obtains a micro-F1 of 93.22% on the test set.

For future works, we plan to explore other deep learning architectures as well as exploiting pre-trained word embedding models, as well as other types of embeddings such as sense embeddings [11]. Due to the unbalanced data, we also plan to explore how the weighting of different classes in training can affect the performance, as well as the use of different sampling methods. Furthermore, the identification of certain types of entities (table 5)(such as PROFESION, INSTITUCION, and CENTRO_SALUD) might be improved by using dictionary-based approaches in addition to our system.

Acknowledgements

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain (DeepEMR project TIN2017-87548-C2-1-R).

References

1. Dernoncourt, F., Lee, J.Y., Szolovits, P.: Neuroner: an easy-to-use program for named-entity recognition based on neural networks. arXiv preprint arXiv:1705.05487 (2017)
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. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
4. Ling, W., Luís, T., Marujo, L., Astudillo, R.F., Amir, S., Dyer, C., Black, A.W., Trancoso, I.: Finding function in form: Compositional character models for open vocabulary word representation. arXiv preprint arXiv:1508.02096 (2015)
5. Lyu, C., Chen, B., Ren, Y., Ji, D.: Long short-term memory rnn for biomedical named entity recognition. BMC bioinformatics **18**(1), 462 (2017)
6. 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
7. 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)
8. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
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. Stubbs, A., Kotfila, C., Uzuner, Ö.: Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. Journal of biomedical informatics **58**, S11–S19 (2015)
11. Trask, A., Michalak, P., Liu, J.: sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. arXiv preprint arXiv:1511.06388 (2015)
12. Uzuner, O., Szolovits, P., Kohane, I.: i2b2 workshop on natural language processing challenges for clinical records. In: Proceedings of the Fall Symposium of the American Medical Informatics Association. Citeseer (2006)
13. Yang, H., Garibaldi, J.M.: Automatic detection of protected health information from clinic narratives. Journal of biomedical informatics **58**, S30–S38 (2015)