Spanish Medical Document Anonymization with Three-channel Convolutional Neural Networks

Jordi Porta-Zamorano

Centro de Estudios de la Real Academia Española
c/ Serrano 187-189, Madrid 28002
Tel.: +34 91 745 55 33
Fax: +34 91 745 55 34
porta@rae.es

Abstract. This paper describes the system presented at the MEDDOCAN (Medical Document Anonymization) task. The system consists of a candidate generator which uses a PoS-tagger, and a candidate classifier based on a convolutional neural network which uses three channels and pretrained word embeddings to represent the sequence of words to be classified and its left and right context. On the MEDDOCAN Test Set, the systems achieved an F_1 score of 0.9184 for subtask 1 (NER offset and entity type classification) and 0.9345 for subtask 2 (sensitive token detection).

Keywords: Medical Document Anonymization · Convolutional neural networks (CNN) · Three-channel CNN (TCCNN)

1 Introduction

This paper describes the system presented at the MEDDOCAN (Medical Document Anonymization) task within the IberLEF 2019 initiative. MEDDOCAN is the first community challenge task specifically devoted to the anonymization of medical documents in Spanish. Although the MEDDOCAN task is structured in two subtasks: (*i*) NER offset and entity type classification and (*ii*) sensitive token detection, these two subtasks are approached simultaneously providing the same output for both. A more detailed description of MEDDOCAN and the results of systems submitted can be found in [7].

The anonymization or de-identification task can be classified as a named entity recognition and classification problem, which has been extensively studied from a variety of approaches ranging from rule-based systems to machine learning algorithms. Several systems using different kinds of recurrent neural networks can be found in [6], [3], and [4] reporting continuous improvements in the state of the art for English in similar anonymization tasks.

The use of multichannel convolutional neural networks for sentence classification was first described in [5], where channels were used to represent different sets of word vectors. In [1], the standard deep learning model for text classification and sentiment analysis using a word embedding layer and a one-dimensional

convolutional neural network was expanded by multiple parallel convolutional neural networks with different kernel sizes over the same text. Here, we will use channels to represent the sequence of words to be classified and its left and right context.

2 MEDDOCAN Corpus Description

The MEDDOCAN corpus is composed of 1000 clinical cases distributed over a Training Set (500 cases), Development Set (250 cases), and a Test Set (250 cases). Each clinical case came in raw text format with its corresponding BRAT annotation file [10].

One of these texts is shown in Fig. 1, where MEDDOCAN entities in the corresponding annotation file in Fig. 2 have been highlighted in red and section titles in blue in order to show the underlying structure of these clinical cases and the co-occurrence of entities with labels that introduce many of them (e.g., *Apellidos: Garcia Prieto*). Note also that, although these texts have been annotated blindly by two different expert annotators with 98% of inter-annotator agreement, the text contains a false positive (*Testes*, wrongly classified as `FAMILIARES_SUJETO_ASISTENCIA`) and a false negative (*08022*, not identified and classified as `TERRITORIO`).

3 System Description

The system built for MEDDOCAN consists of a candidate generator and a candidate classifier, which are depicted in Fig. 3 and described in the following paragraphs.

The candidate generation is performed in two steps: (*i*) Texts are PoS-tagged with a general in-house Spanish PoS-tagger which performs sentence segmentation, tokenization, and morphosyntactic analysis. The tokenization subcomponent has been slightly modified to split textual elements containing hyphens and some textual elements consisting of unbroken sequences of words, numbers and punctuation without proper spacing (e.g., *Edad:68* $\rightarrow$ *Edad : 68*); and (*ii*) PoS-Tagged texts are then given to an n-gram generator to provide candidates along with their textual context for all the types of entities defined by the MEDDOCAN task. Since most of the n-grams do not correspond to entities from a linguistic point of view (they should be noun phrases), a filter is applied to reduce the number of false candidates. This filter removes n-grams with non-allowed PoS tags in n-gram initial and final positions (punctuation, prepositions, adverbs, articles, conjunctions, etc.), n-grams with internal unpaired punctuation, and n-grams containing verbs (except past participle forms) or any of the words in a blacklist of the 15,000 most frequent non-allowed words within entities manually extracted from the MEDDOCAN texts in the Training Set.

Candidate n-grams with their context are classified in two steps: (*i*) n-grams are initially classified by a three-channel convolutional neural network (TCCNN) with precomputed word embeddings (channels represent left and right context,

Nombre: Marc.
Apellidos: Garcia Prieto.
CIPA: 270058.
NASS: 27 2226350 05.
Domicilio: Av. Litoral, 30, 1C .
Localidad/ Provincia: Barcelona.
CP: 08005.
NHC: 270058.
Datos asistenciales.
Fecha de nacimiento: 27-10-1968.
País: España.
Edad: 47 Sexo: H.
Fecha de Ingreso: 13-12-2015.
Especialidad: andrología.
Médico: Anna Bujons Tur NºCol: 08 08 76541.
Antecedentes: sin antecedentes patológicos de interés ni hábitos tóxicos.
Historia Actual: Paciente varón de 47 años de edad, acude a la consulta de andrología por presentar erecciones prolongadas no dolorosas de aproximadamente 4 años de evolución tras traumatismo perineal cerrado con el manillar de una bicicleta.
Exploración física:En la exploración física se observan cuerpos cavernosos aumentados de consistencia, no dolorosos a la palpación, sin palpar pulsos anómalos. Sensibilidad peneana conservada. Testes móviles en ambas bolsas escrotales y sin alteraciones.
Resumen de pruebas complementarias: Como exploraciones complementarias se le realiza ecodoppler penenano: vascularización cavernosa derecha aparentemente conservada; en la porción más proximal del cuerpo cavernoso izquierdo se observa formación anecoica (2x1.8x1.5cm) con flujo turbulento en su interior compatible con fístula arteriovenosa (FAV) de larga duración.
Evolución y comentarios: Con la orientación diagnóstica de Priapismo de alto flujo se decide realización de Arteriografía pudenda con anestesia local confirmándose FAV y posterior embolización de la misma mediante 2 coil 3x5. Tras la embolización el paciente e voluciona favorablemente con detumecencia peneana completa y erecciones normales. Actualmente está asintomático.
Diagnóstico Principal: Priapismo
Remitido por: Anna Bujons Tur Calle Joaquim Folguera, 3 - 5º 2ª 08022 Barcelona. abujons@gmail.com

Fig. 1. MEDDOCAN Clinical Case S0004-06142006000600015-1: Raw Text

```
T1  NOMBRE_SUJETO_ASISTENCIA 9 13  Marc
T2  NOMBRE_SUJETO_ASISTENCIA 27 40  Garcia Prieto
T3  ID_SUJETO_ASISTENCIA 48 54  270058
T4  ID_ASEGURAMIENTO 62 75  27 2226350 05
T5  CALLE 88 107  Av. Litoral, 30, 1C
T6  TERRITORIO 132 141  Barcelona
T7  TERRITORIO 147 152  08005
T8  ID_SUJETO_ASISTENCIA 159 165  270058
T9  FECHAS 209 219  27-10-1968
T10 PAIS 227 233  España
T11 EDAD_SUJETO_ASISTENCIA 241 243  47
T12 SEXO_SUJETO_ASISTENCIA 250 251  H
T13 FECHAS 271 281  13-12-2015
T14 NOMBRE_PERSONAL_SANITARIO 318 333  Anna Bujons Tur
T15 ID_TITULACION_PERSONAL_SANITARIO 347 358  08 08 76541
T16 SEXO_SUJETO_ASISTENCIA 460 465  varón
T17 EDAD_SUJETO_ASISTENCIA 469 484  47 años de edad
T18 FAMILIARES_SUJETO_ASISTENCIA 871 877  Testes
T19 NOMBRE_PERSONAL_SANITARIO 1719 1734  Anna Bujons Tur
T20 CALLE 1735 1760  Calle Joaquim Folguera, 3
T21 TERRITORIO 1775 1784  Barcelona
T22 CORREO_ELECTRONICO 1786 1803  abujons@gmail.com
```

Fig. 2. MEDDOCAN Clinical Case S0004-06142006000600015-1: BRAT File

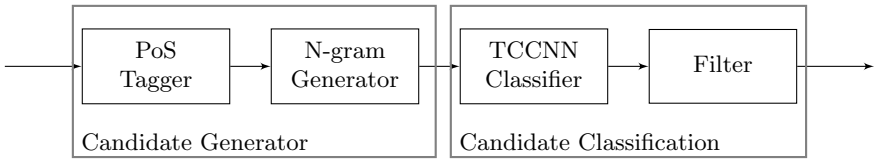

Fig. 3. System Arquitecture

and the n-gram to be classified); and (*ii*) a final filter is applied to discard misclassifications and to solve overlapped classified candidates selecting left-most longest n-grams. This final filter applies constraints at word and character levels on classified n-grams. For example, emails must contain the @ character, streets have no street type information, etc. These constraints have been created on the basis of observation of misclassifications in the Development Set.

Despite the initial filter applied to reduce the number of candidates, 89% of the candidates in the MEDDOCAN corpus are still non-entities, resulting in training sets with a very unbalanced label distribution. To avoid bias to the machine learning algorithm, labels are weighted to make them equally important to the algorithm. Word embeddings have been trained using Word2Vec [8] from Spanish newspaper archives, the Iberia Corpus [9], and the MEDDOCAN Training Set, with a total of 1,355,054,567 tokens and a vocabulary size of 940,576.

The vectors have a dimensionality of 100 and were trained using the continuous bag-of-words architecture. Words out of vocabulary are initialized to zero.

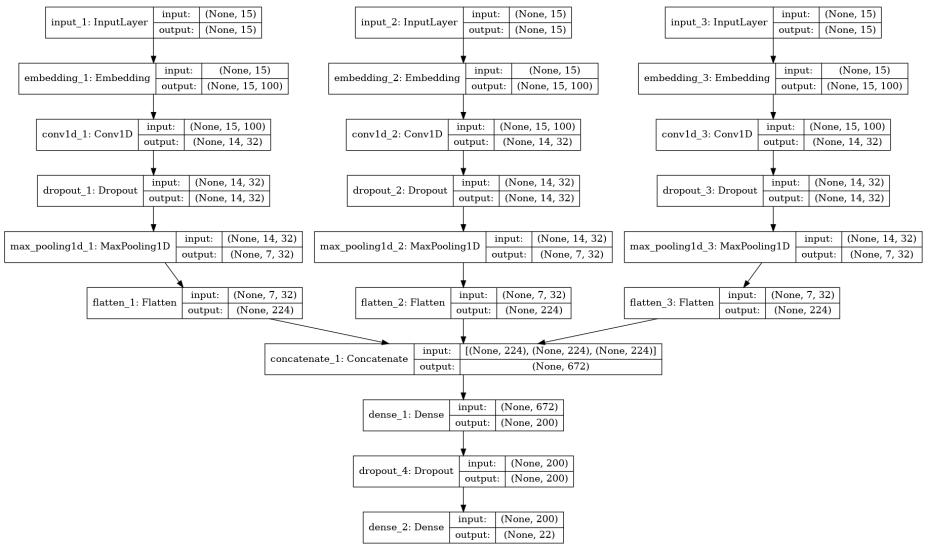

Fig. 4. TCCNN for Candidate Classification

The topology of the TCCNN is depicted in Fig. 4. Three input channels are defined for processing the sequence of words to be classified and its left and right context. Each channel has an input layer, limiting to 15 the length of the input sequence, followed by an embedding layer and a one-dimensional convolutional layer. Max pooling layers down-sample the output from the convolutional layers and a flatten layer reduces its dimension for concatenation. The concatenated output from the three channels is processed by a dense layer and an output layer. The code used for implementing the TCCNN is an adaptation of the code published in [1], implemented in Keras [2].

Training and development sets were created from MEDDOCAN datasets using the candidate generator to train and evaluate the TCCNN. For each entity type, its weight was computed as n/n_i where $n = \sum_i n_i$ and n_i is the number of examples of type i. Hyperparameters of the neural network were manually explored until a peak of 0.965 accuracy was reached on the development set (corresponding to 0.992 on the training set). Channels have a capacity of 15 words, dropout rates are set to 0.005, filter windows to 32, kernel and pool sizes to 2; dense layers contain 200 units with L_2 kernel regularizers set to 0.001; activation functions are ReLU except for the output layer, which is softmax. The mini-batch was 156 and the optimizer was Adam with a learning rate of 0.001. The learning curve of Fig. 5 indicates that the model fits quite well with a training-development accuracy gap of 3%. The performance of the neural net-

work classifier (without the final filter) on every entity type in terms of precision and recall, ordered by F_1, is shown in Table 1. There is no apparent correlation between performance and the number of examples of each entity type.

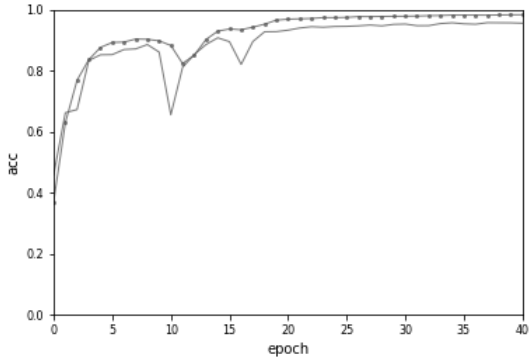

Fig. 5. TCCNN Learning Rate on Training and Development Sets

4 Results

The two systems submitted differ in the training corpus: System II is trained with the Training Set and System III is trained with all available data, i.e., the Training and Development Sets.

The results on Development and Test Sets in subtasks 1 and 2 are shown in Tables 2.1 and 2.2, where the effect of the final filter is also measured (in all cases left-most longest matches are selected in case of overlapping). Results for System III on the Development Set are not given since this set has been seen during the training. For all the experiments, the effect of the filter is positive, increasing all figures by 0.8-3.41%. On the Test Set, Systems II and III have slightly different precision and recall but a similar F_1 of about 0.918 for subtask 1 and 0.93 for subtask 2.

5 Conclusion and Future Work

An initial approach to the MEDDOCAN task was to create a pattern-based system using context-free grammars to classify and generate candidates for the kind of entities defined by the task (System I). The development of this approach results in high precision but low recall due to the difficulty of modelling the span and context of some entity types. The course of the development was then changed by the n-gram generation using a blacklist of word forms combined

Table 1. TCCNN Evaluation on the Development Set

Label	**Freq.**	**Prec.**	**Rec.**	F_1
NON_ENTITY	46885	0.992	0.973	0.982
ID_ASEGURAMIENTO	194	0.979	0.985	0.982
SEXO_SUJETO_ASISTENCIA	454	0.951	0.987	0.969
PAIS	347	0.939	0.977	0.958
CORREO_ELECTRONICO	240	0.935	0.954	0.944
EDAD_SUJETO_ASISTENCIA	521	0.917	0.971	0.943
NOMBRE_SUJETO_ASISTENCIA	503	0.899	0.990	0.942
FECHAS	724	0.935	0.930	0.932
ID_CONTACTO_ASISTENCIAL	32	0.865	1.000	0.928
ID_SUJETO_ASISTENCIA	290	0.921	0.928	0.924
TERRITORIO	985	0.811	0.920	0.862
NOMBRE_PERSONAL_SANITARIO	488	0.782	0.951	0.858
CENTRO_SALUD	2	0.667	1.000	0.800
CALLE	417	0.741	0.856	0.794
NUMERO_FAX	5	0.667	0.800	0.727
ID_TITULACION_PERSONAL_SANITARIO	226	0.505	0.982	0.667
HOSPITAL	140	0.540	0.821	0.652
NUMERO_TELEFONO	24	0.475	0.792	0.594
FAMILIARES_SUJETO_ASISTENCIA	92	0.381	0.554	0.451
INSTITUCION	68	0.176	0.574	0.269
PROFESION	4	0.100	0.250	0.143
OTROS_SUJETO_ASISTENCIA	5	0.000	0.000	0.000

Table 2. MEDDOCAN Subtasks System Evaluation

MEDDOCAN Subtask 1					
Sys.	**Filt.**	**Set**	**Prec.**	**Rec.**	F_1
II	−	Dev	0.9235	0.8778	0.9000
II	+	Dev	**0.9440**	**0.9005**	**0.9217**
II	−	Test	0.9013	0.8732	0.8870
II	+	Test	**0.9315**	**0.9057**	**0.9184**
III	−	Test	0.8838	0.8843	0.8841
III	+	Test	**0.9191**	**0.9173**	**0.9182**

MEDDOCAN Subtask 2					
Sys.	**Filt.**	**Set**	**Prec.**	**Rec.**	F_1
II	−	Dev	0.9433	0.9109	0.9268
II	+	Dev	**0.9548**	**0.9156**	**0.9348**
II	−	Test	0.9240	0.9085	0.9162
II	+	Test	**0.9436**	**0.9215**	**0.9324**
III	−	Test	0.9090	0.9276	0.9182
III	+	Test	**0.9327**	**0.9359**	**0.9343**

with a machine learning classifier, which was previously used for named entity classification, resulting in a best F_1 of 0.9184 for subtask 1 and 0.9345 for subtask 2. Part of the success in the tasks can be explained by the distribution of the entities within the texts, many of which are placed in fields and surrounded by labels introducing or delimiting them. A natural path for future work is to incorporate PoS tags and word character level information into new channels of the TCCNN, for ther purpose of eliminating the filter in the n-gram candidate generation and the final filter.

References

1. Brownlee, J.: Deep Learning for Natural Language Processing, chap. Develop an n-gram CNN for Sentiment Analysis (2017)
2. Chollet, F., et al.: Keras. https://keras.io (2015)
3. Dernoncourt, F., Lee, J.Y., Uzuner, Ö., Szolovits, P.: De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association : JAMIA **24**(3), 596–606 (2016)
4. Khin, K., Burckhardt, P., Padman, R.: A Deep Learning Architecture for De-identification of Patient Notes: Implementation and Evaluation (2018), http://arxiv.org/abs/1810.01570
5. 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. Association for Computational Linguistics, Doha, Qatar (2014)
6. 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), a Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry
7. 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 (2019), TBA
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 3111–3119. NIPS'13, USA (2013)
9. Porta-Zamorano, J., del Rosal García, E., Ahumada-Lara, I.: Design and development of Iberia: a corpus of scientific Spanish. Corpora **6**(2), 145–158 (2011)
10. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: A Web-based Tool for NLP-assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. EACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012)