A Deep Learning-Based System for the MEDDOCAN Task

Dehuan Jiang[1], Yedan Shen[1], Shuai Chen[1], Buzhou Tang[*1], Xiaolong Wang[1], Qingcai Chen [1], Ruifeng Xu[1], Jun Yan[2], Yi Zhou[*3]

[1]Department of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China, 518055
[2]Yidu Cloud (Beijing) Technology Co., Ltd, Beijing
[3]Sun YAT-SEN UNIVERSITY

jiangdehuan@stu.hit.edu.cn, shenyedan@stu.hit.edu.cn, chenshuai726@gmail.com, tangbuzhou@gmail.com, wangxl@insun.hit.edu.cn, qingcai.chen@gmail.com, xuruifeng@hit.edu.cn, Jun.YAN@Yiducloud.cn, zhouyi@sysu.edu.cn

Abstract. Due to privacy constraints, de-identification, identifying and removing all PHI mentions, is a prerequisite for accessing and sharing clinical records outside of hospitals. Large quantities of studies on de-identification have been conducted in recent years, especially with the efforts of i2b2 (the Center of Informatics for Integrating Biology and Bedside). The i2b2 community has organized challenges about de-identification for clinical text in English many times. In 2019, Martin Krallinger et al. organized a challenge task specifically devoted to the anonymization of medical documents in Spanish, called the MEDDOCAN (Medical Document Anonymization) task. We participated in and developed a deep learning-based system for the MEDDOCAN task. Our system was developed on a training set of 500 records and a development set of 250 records. Evaluation on a test set of 250 shows that our system achieved a "strict" F1-score of 0.9646 at entity level, a "strict" F1-score of 0.97 at span level and a "merged" F1-score of 0.9821 at span level.

Keywords: De-identification, Protected Health Information, medical document anonymization, deep learning

1 Introduction

De-identification is a prerequisite of clinical record accessing and sharing outside of hospitals, which is very important for secondary use of clinical data. In the past few years, de-identification had attracted plenty of attention and a large number of efforts had been made for de-identification, especially for clinical documents in English. The

[*] Corresponding author at: Department of Computer Science, Harbin Institute of Technology, Shenzhen, China, 518055. Zhongshan School of Medicine, Sun YAT-SEN UNIVERSITY, Guangzhou, China, 510080

representative works are natural language processing (NLP) challenges including the de-identification task of clinical text, such as the i2b2 (the Center of Informatics for Integrating Biology and Bedside) 2006 [1] and 2014 [2-4], and the N-GRID (the Centers of Excellence in Genomic Science (CEGS) Neuropsychiatric Genome-scale and RDOC Individualized Domains) 2016 [5]. As these challenges are public and provide manually annotated corpora for de-identification, they attract lots of research teams to participate in and develop various kinds of systems [6-9]. According to the overview report of the N-GRID 2016 NLP challenge [5], the best system is a hybrid system based on deep learning methods [10].

In 2019, Martin Krallinger et al. organized a challenge task special for the de-identification of medical documents in Spanish, called the MEDDOCAN (Medical Document Anonymization) task [11]. The organizers provided a training set of 500 clinical records, a development set of 250 clinical records and a test set of 250 clinical records embedded in synthetic corpus 3751 clinical records. We participated in this challenge task and developed a system based on latest deep learning methods such as BERT (Bidirectional Encoder Representations from Transformers) (https://github.com/google-research/bert) and flair (https://github.com/zalandoresearch/flair). The system developed on the training set and development set achieved a “strict” F1-score of 0.9646 at entity level, a “strict” F1-score of 0.97 at span level and a “merged” F1-score of 0.9821 at span level. It should be noted that the results reported here were the new results after we added a post-processing module to fix tokenization errors when testing.

2 Material and Methods

The overview architecture of our system for the MEDDOCAN task is shown Fig.1. We first tokenized raw clinical texts in Spanish, and then deployed two individual deep learning methods (i.e., BERT+CRF and flair) for de-identification respectively. Our system was described below in detail.

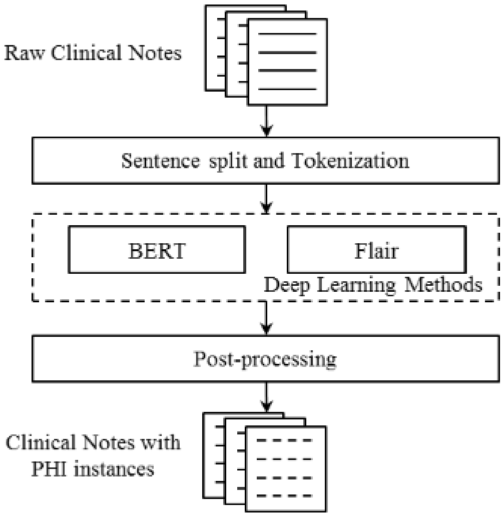
Raw Clinical Notes
Sentence split and Tokenization
BERT
Flair
Deep Learning Methods
Post-processing
Clinical Notes with PHI instances

Fig. 1. Overview architecture of our de-identification system.

2.1 Dataset

The organizers of the MEDDOCAN task provided participants with a synthetic corpus of 1000 discharge summaries and medical genetics clinical records manually annotated by medical experts according to a guideline defining 22 types of PHI. The corpus were divided into three parts: a training set of 500 records with 11,333 PHI mentions, a development set of 250 records with 5801 PHI mentions, and a test set of 250 records with 5661 PHI mentions. The test set was embedded in a background set of 3751 clinical records that have been manually split into sentences. The statistics of the corpus, including number of documents, sentences and PHI mentions are listed in Table 1, where "NA" denotes unknown.

Table 1. Statistics of the MEDDOCAN corpus.

Statistic	Training	Test	Development	Background
Document	500	250	250	3751
Sentence	17,323	5,155	9,037	135,110
PHI mention	11,333	5,661	5,801	NA

2.2 Sentence Split and Tokenization

Sentence split and tokenization are two important preprocessing steps for natural language processing (NLP). We developed a simple rule-based system for sentence split and tokenization. A document was split into sentences by ';', '?', '!', '\n' or '.' not in numbers, and each sentence was tokenized by the method proposed by Liu et al. [10]

2.3 Deep Learning Methods

De-identification is a typical named entity recognition problem, which is usually recognized as a sequence labeling problem. In this study, we deployed two deep learning methods for the MEDDOCAN task, that is, BERT+CRF and flair as follows:

BERT+CRF. a method that appends a condition random field (CRF) layer to BERT. In our study, we compared the cases using different settings.

Flair. a sequence labeling method based on contextual string embeddings.

2.4 Post-processing

As clinical records in the test set have been manually split into sentence, to fixed errors caused by sentence split, we mapped the split sentences back to the gold ones and combined the neighbor PHI mentions of the same type together.

2.5 Evaluation

All system performance was measured by micro-average precisions (P), recalls (R), and F1-scores (F1) under three criteria: "strict" at entity level (track 1), "strict" at span level (track 2), and "merged" at span level, where "strict" at entity level checks whether a recognized PHI mention exactly matches a gold one of the same type, "strict" at span level checks whether a recognized PHI mention has the same span as a gold one no matter their types, and "merged" at span level is a "strict" at span level after merging the spans of PHI mentions connected by non-alphanumerical characters. All evaluations were conducted on the independent test data set, and the measures were calculated by the tool provided by the MEDDOCAN organizers.

2.6 Experiments Setup

In this study, PHI mentions were represented by "BIO" (B-beginning of a PHI mention, I-insider a PHI mention, O-outside a PHI mention). The hyper-parameters and parameter estimation algorithm listed in Table 2 were used in the deep learning methods. The pre-trained neural language models (https://storage.googleapis.com/bert_models/2018_11_03/multilingual_L-12_H-768_A-12.zip and https://github.com/zalandoresearch/flair) were used in BERT+CRF and flair respectively. The other parameters were optimized on the development set, and all models are evaluated on the independent test set.

Table 2. Hyper-parameters and parameter estimation algorithm used in the deep learning methods.

Hyper-parameter	Value
Dimension of word representation	768
Dimension of character representation	256
Dimension of POS representation	30
Dropout probability	0.5
Learning rate	Flair: 0.1, BERT: 1e-5
Training epochs	BERT: 30, Flair: 150
Parameter estimation algorithm	BERT: adam with warmup, Flair: SGD with momentun

3 Results

The micro-average precisions, recalls and F1-scores of our system under the three criteria were listed in Table 3. BERT+CRF outperformed flair by about 0.3% in F1-scores because of higher recalls. When POS features were added, the performance of BERT+CRF decreased a little bit. When we further fine-tuned BERT+CRF on the combination of training and development sets, BERT+CRF did not change very much. Our system achieved the highest "strict" F1-score of 0.9646 at entity level, a "strict" F1-score of 0.97 at span level and a "merged" F1-score of 0.9821 at span level.

Table 3. Results of different deep learning methods.

Method	Track1			Track2			Merged		
	P	R	F1	P	R	F1	P	R	F1
BERT+CRF	0.9624	0.9668	0.9646	0.9675	0.9719	0.9697	0.9813	0.9828	0.982
BERT+CRF1	0.9548	0.9656	0.9601	0.9609	0.9717	0.9663	0.9761	0.9821	0.9791
BERT+CRF2	0.9609	0.9679	0.9644	0.9665	0.9735	0.97	0.9813	0.983	0.9821
Flair	0.9611	0.9604	0.9608	0.9652	0.9645	0.9648	0.9809	0.9766	0.9787

BERT+CRF1 and BERT+CRF2 denote BERT+CRF using POS features and fine-tuned on the combination of the training and development sets 10 epochs more respectively.

4 Discussion

The new results (shown in Table 3) reported here are the results of our first submissions (shown in Table 4) after post-processing. The great differences between "strict"

F1-scores (track 2) and “merged” F1-scores inspired us to find errors caused by sentence split. For example, in sentence ”Domicilio: Av. de Jaén, 28.”, “Av. de Jaén, 28” is a entity of “CALLE”, but was split into two entities of “CALLE”: “Av.” and “de Jaén, 28” as the sentence were split into two sentence “Domicilio: Av.” and “de Jaén, 28.” by ‘.’. The sentence split errors result in an F1-score difference of about 0.4 between “strict” F1-scores and “merged” F1-scores. We can see that the post-processing module brings a “strict” F1-score gain of 0.0245 for track 1 and a “strict” F1-score gain of 0.0243 for track 2. The differences between “strict” F1-scores (track 2) and “merged” F1-scores decrease sharply when the post-processing module is added.

Table 4. The first submitted results of different deep learning methods.

Method	**Track1**			**Track2**			**Merged**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERT+CRF	0.9289	0.9511	0.9399	0.9339	0.9562	0.9449	0.9803	0.9828	0.9816
BERT+CRF1	0.9222	0.9502	0.9360	0.9282	0.9564	0.9421	0.9750	0.9820	0.9785
BERT+CRF2	0.9281	0.9525	0.9401	0.9336	0.9581	0.9457	0.9803	0.9834	0.9818
Flair	0.9287	0.9454	0.9370	0.9329	0.9497	0.9412	0.9796	0.9762	0.9779

BERT+CRF1 and BERT+CRF2 denote BERT+CRF using POS features and fine-tuned on the combination of the training and development sets 10 epochs more respectively.

To analyze errors in our system, we evaluated the performance on each category of entity and found that the F1-scores on “PROFESION” and “INSTITUCION” are much lower than other categories except “OTROS_SUJETO_ASISTENCIA”, on which the F1-score is zero. There are main three reasons why these three categories of entities are not well recognized. Firstly, entities in some categories are too few. For example, there are only 15 entities of “OTROS_SUJETO_ASISTENCIA” in the training set and development set in all, and only 7 in the test set. Secondly, entities of “INSTITUCION” vary greatly in format. Thirdly, there may be some entities wrongly labeled as gold standards. For example, “militar” and “ex-operario de industria textil”, which means “soldier” and “ex-textile industry operator” respectively, are recognized by our system but not labeled as gold standards.

5 Conclusion

In this study, we developed a deep learning-based system for the MEDDOCAN task, a challenge special for de-identification of clinical text in Spanish. The system achieves a promising performance. Besides, “BERT+CRF” outperforms flair. In the future, we will investigate whether BERT and flair can be combined together for further improvement.

Acknowledgements

This paper is supported in part by grants: NSFCs (National Natural Science Foundations of China) (U1813215, 61876052 and 61573118), National Key Research and Development Program of China (2017YFB0802204), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20170307150528934 and JCYJ20180306172232154), Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052).

References

1. Ö. Uzuner, Y. Luo and P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, Journal of the American Medical Informatics Association, vol. 14, no. 5, 2007, pp. 550-563.
2. A. Stubbs and Ö. Uzuner, Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus, Journal of biomedical informatics, vol. 58, 2015, pp. S20-S29.
3. Ö. Uzuner and A. Stubbs, Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks, Journal of biomedical informatics, vol. 58, 2015, pp. S1-S5.
4. A. Stubbs, C. Kotfila and Ö. Uzuner, Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1, Journal of biomedical informatics, vol. 58, 2015, pp. S11-S19.
5. Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1[J]. Journal of biomedical informatics, 2017, 75: S4-S18.
6. S.M. Meystre, F.J. Friedlin, B.R. South, S. Shen and M.H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC medical research methodology, vol. 10, no. 1, 2010, pp. 70.
7. O. Ferrández, B.R. South, S. Shen, F.J. Friedlin, M.H. Samore and S.M. Meystre, Evaluating current automatic de-identification methods with Veteran's health administration clinical documents, BMC medical research methodology, vol. 12, no. 1, 2012, pp. 109.
8. L. Deleger, K. Molnar, G. Savova, F. Xia, T. Lingren, Q. Li, K. Marsolo, A. Jegga, M. Kaiser and L. Stoutenborough, Large-scale evaluation of automated clinical note de-identification and its impact on information extraction, Journal of the American Medical Informatics Association: JAMIA, vol. 20, no. 1, 2013, pp. 84-94.
9. Z. Liu, Y. Chen, B. Tang, X. Wang, Q. Chen, H. Li, J. Wang, Q. Deng and S. Zhu, Automatic de-identification of electronic medical records using token-level and character-level conditional random fields, Journal of Biomedical Informatics, vol. 58, 2015, pp. S47-S52.
10. Liu Z, Tang B, Wang X, et al. De-identification of clinical notes via recurrent neural network and conditional random field[J]. Journal of biomedical informatics, 2017, 75: S34-S42.
11. Marimon, Montserrat, Gonzalez-Agirre, Aitor, et al. Automatic de-identification of medical texts in Spanish: the MEDDOCAN track, corpus, guidelines, methods and evaluation of results,Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)