Hadoken: a BERT-CRF Model for Medical Document Anonymization

Jihang Mao[1], Wanli Liu[2]

[1] Montgomery Blair High School,
51 University Blvd E, Silver Spring, MD 20901, USA

[2] TAJ Technologies, Inc.,
7910 Woodmont Ave #1214, Bethesda, MD 20814, USA

jim-blair@hotmail.com

Abstract. In this paper we report our participation in the 2019 MEDDOCAN task where participants were provided with a synthetic corpus of clinical cases and required to de-identify the protected health information (PHI) in the corpus. Our system Hadoken utilizes BERT, a multi-layer bidirectional transformer encoder which can help learn deep bi-directional representations, and fine-tunes the pre-trained BERT model on training data for MEDDOCAN task. It then feeds the representation of a clinical case into a CRF output layer for token-level classification. To make the results more accurate, we apply some post-processing techniques based on a thorough error analysis. Through utilizing linguistic resources, we were able to achieve results of a higher recall. Our best F-Scores on the test set for Task1 and Task2 are 0.9375, 0.9428 [strict], and 0.9480 [merged] respectively. We find that Hadoken is a robust and competitive system and is applicable to multilingual NER tasks.

Keywords: BERT; Conditional Random Field; Named Entity Recognition; Multilingual Model; Evaluation.

1 Introduction

Recently, there has been a rapid growth of using medical documents across a wide variety of health domains. With the application of AI techniques, clinical records can be used for aiding clinical decision making, expanding knowledge about diseases, improving patients' healthcare experiences, et al. [1]. However, there are growing concerns about patient privacy due to the widespread practice of health information sharing [2]. Clinical records with protected health information (PHI) cannot be directly shared "as is", due to privacy constraints. Despite of the abundance of research in privacy-preserving for structured data [3-5] , it is particularly cumbersome to carry out NLP research in the medical domain.

A necessary precondition for accessing clinical records outside of hospitals is their de-identification, i.e., the exhaustive removal, or replacement, of all mentioned PHI

phrases. Studies have shown that such a simple de-identification strategy lacks the flexibility to adequately meet the diverse needs of data users [6, 7]. The practical relevance of anonymization or de-identification of clinical texts motivated the proposal of two shared tasks, the 2006 [8] and 2014 [9] de-identification tracks, organized under the umbrella of the i2b2 (i2b2.org) community evaluation effort. The i2b2 project has collected multiple sets of clinical documents from different healthcare organizations and made them available for research, but was focused on documents in English and covering characteristics of US-healthcare data providers [10].

MEDDOCAN (Medical Document Anonymization) [11] is the first community challenge task specifically devoted to the anonymization of medical documents in Spanish. The goal of MEDDOCAN is to set new state-of-the-art results for automatically recognizing sensitive entity mentions. The MEDDOCAN challenge task comprises two subtasks: NER (Named Entity Recognition) offset and entity type classification (Sub-Track 1) and sensitive token detection (Sub-Track 2). We participated in both sub-tasks this year. Participating teams were provided with a synthetic corpus of clinical cases enriched with PHI expressions, named the MEDDOCAN corpus. It was selected manually by a practicing physician and augmented with PHI information from discharge summaries and medical genetics clinical records.

A brief description of our method for MEDDOCAN task is presented in Section 2. In Section 3 we show the results of our method on the official MEDDOCAN test datasets. In Section 4 we present a discussion of the results and conclusions of our participation in this challenge.

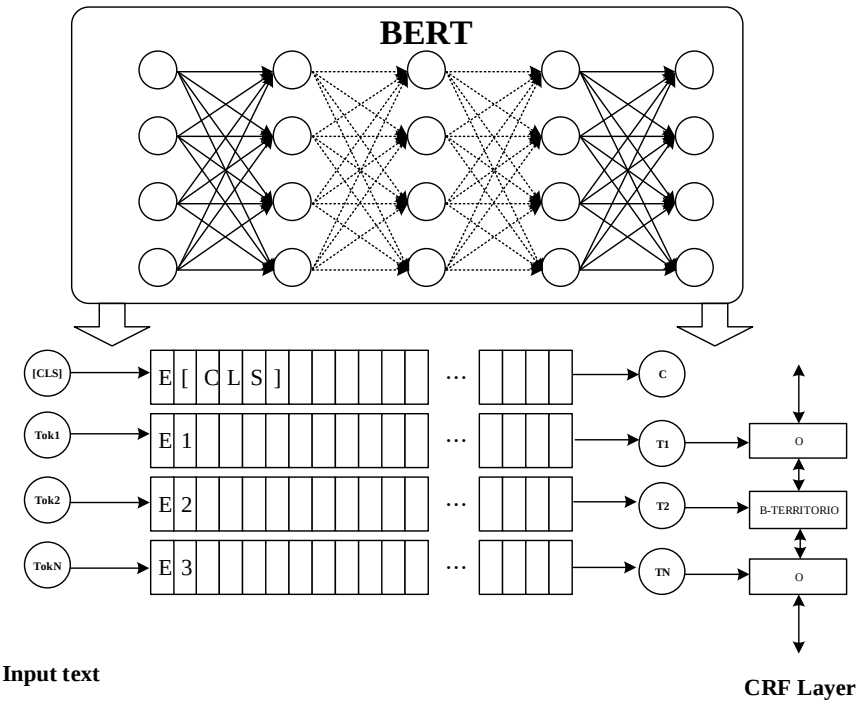
2 Methods

For MEDDOCAN Task, our system Hadoken builds on BERT-NER (github.com/kyzhouhzau/BERT-NER) developed based on BERT, which has obtained state-of-the-art performance on most NLP tasks recently [12]. BERT utilizes a multilayer bidirectional transformer encoder which can learn deep bi-directional representations and can be later fine-tuned for a variety of tasks such as NER. Before BERT, deep learning models, such as Long Short-Term Memory (LSTM) and Conditional Random Field (CRF) have greatly improved the performance in NER over the last few years [13]. And ELMo [14] has proved the effectiveness of contextualized representations from a deeper structure for transfer learning.

Given a clinical case, our method first obtains its token representation from the pre-trained BERT model using a case-preserving WordPiece model, including the maximal document context provided by the data. Next we formulate this as a tagging task by feeding the representation into a CRF [15] output layer, which is a token-level classifier over the NER label set. Finally, we use some post-processing techniques to generate the result.

The pre-trained BERT models are trained on a large corpus (Wikipedia + BookCorpus). There are several pre-trained models release. In MEDDOCAN, we chose BERT-Base, multilingual cased model for following reasons: First, multilingual model is better for Spanish documents in MEDDOCAN because the English-only model splits tokens not available in its vocabulary into sub-tokens, which will affect the accuracy of the NER task. Second, although BERT-Large generally outperforms BERT-Base in English NLP tasks, BERT-Large versions of multilingual models haven't been released. Third, the cased model is better because the case information is important for the NER task. In MEDDOCAN task, we represent the input passage as a single packed sequence using BERT embedding, then use a CRF layer as the tag decoder (Figure 1).

Fig. 1. Architecture of the Hadoken model for NER. Following [12], we denote input embedding as *E*, the final hidden vector of the special [CLS] token, and the final hidden vector for the *i*th input token as T_i



Next, we conducted an error analysis on the development set to find out which types of PHI entity were not predicted well with the above model. Since high precision (0.97) and low recall (0.75) were observed, we investigated all types of PHI entity that were not recognized by NER. A list of PHI entity types with the top number of false negatives is shown in Table 1. We then use some post-processing techniques to improve our results. For "TERRITORIO", "CALLE" and "HOSPITAL", we utilize the information in the Gazetteer of MEDDOCAN related entities provided by organizers as a linguistic

resource, to identify some hospitals and different types of locations (provinces, cities, towns, …). For "NOMBRE_PERSONAL_SANITARIO", we perform an exhaustive search to label the names in the clinical case that match the tokens already recognized as "NOMBRE_PERSONAL_SANITARIO" by the model. For "CORREO_ELECTRÓNICO", we use regular expression to find email addresses in the clinical case. After post-processing, the recall can be greatly improved (e.g., all "CORREO_ELECTRÓNICO" entities in Table 1 can be identified) while the precision remains stable.

Table 1: PHI entity types with the top number of false negatives on the development dataset

PHI entity types	Number of false negatives	Total number of entities
TERRITORIO	359	987
NOMBRE_PERSONAL_SANITARIO	253	497
CORREO_ELECTRÓNICO	241	241
CALLE	189	434
HOSPITAL	133	140

3 Results & Discussion

The MEDDOCAN corpus has been randomly sampled into three subsets: the train, the development, and the test set. The training set contains 500 clinical cases, and the development and test set 250 clinical cases each.

In both MEDDOCAN sub-tracks, the official evaluation and the ranking of the submitted systems will be based exclusively on the F-score (F1) measure (labeled as "SubTrack 1 [NER]" and "SubTrack 2 [strict]" in the evaluation script). Here we present the results on the test set. The background set is composed of 3,751 clinical cases, in which the test set with Gold Standard annotations is composed of 250 clinical cases. We submitted five prediction results of our system. For "Submission1", "Submission2", and "Submission3", the BERT-CRF model was fine-tuned using the hyperparameter values suggested in [11]: learning rate=2e-5, number of epochs=3, max sequence length=256, 384, 512, and batch size=16, 14, 10, respectively. For "Submission4" and "Submission5", the BERT model is further pre-trained with the background corpora (the unlabeled 3,751 clinical cases), and the maximum sequence length and batch size were set to 320 and 14 respectively. However, it seems the new model doesn't make any improvement, probably due to the size of the background corpora is too small compared to the original training corpora (Wikipedia + BookCorpus).

As shown in Table 2, "Submission1" outperformed all other submissions in precision and F-measure for Task1, while the choice of maximum sequence length of 512 in "Submission3" resulted in the highest performance in leak and recall for Task1. We also note that "Submission3" consistently achieved best performance in F1[strict] and F1[merged] for Task2, suggesting that positional embeddings are very useful for specific token detection tasks.

Table 2. Official results for our submissions on MEDDOCAN test set. The best results among all submissions are highlighted in bold.

Systems	Task1 Leak	Task1 Precision	Task1 Recall	Task1 F1	Task2 F1[strict]	Task2 F1[merged]
Submission1	0.06617	**0.96451**	0.91203	**0.93753**	0.94098	0.94627
Submission2	0.06604	0.96164	0.91221	0.93627	0.9398	0.94524
Submission3	**0.05395**	0.93306	**0.92828**	0.93067	**0.94283**	**0.94798**
Submission4	0.05567	0.93125	0.92598	0.92861	0.9406	0.94589
Submission5	0.05594	0.92547	0.92563	0.92555	0.93727	0.94458

4 Conclusion

We described our approach Hadoken that participated in the MEDDOCAN Medical Document Anonymization task in IberLEF 2019. Compared to previous methods, Hadoken has several significant differences from system architecture to the processing flow. It is a general and robust framework and showed competitive performance during the MEDDOCAN evaluations. The performance of the proposed methodology can be improved by additional Spanish Linguistic resources, such as the MEDOCAN-Gazetteer. However, in our experiments, the accuracy for the 'HOSPITAL' entities is still low after utilizing the information of hospitals in the gazetteer, which is interesting and worth further research. With more and more training corpora available, we also plan to explore the application of Hadoken in other NER task in future work.

Acknowledgements

The authors would like to thank Dr. Yutao Zhang for providing Jihang Mao the intern opportunity at George Mason University and valuable suggestions and comments on the manuscript. The authors would also like to thank the MEDDOCAN task organizers for providing the task data and tools.

References

1. Jensen, P.B., Jensen, L.J., Brunak, S. Mining electronic health records: Towards better research applications and clinical care. Nature Rev Genetics 13(6), 395–405. (2012).
2. Safran, C., Bloomrosen, M., Hammond, W.E., Labkoff, S., Markel-Fox, S., Tang, P.C., Detmer, D.E. Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper. J Amer Medical Informatics Assoc. 14(1):1–9. (2007).
3. Aggarwal, C.C., Yu, P.S. Privacy-Preserving Data Mining: Models and Algorithms. Springer; New York (2008).
4. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S. Privacy-preserving data publishing: A survey of recent developments. ACM Comput Surveys 42(4) Article 14. (2010).
5. Melville, N., McQuaid, M. Generating shareable statistical databases for business value: Multiple imputation with multimodal perturbation. Inform Systems Res.23(2):559–574. (2012).
6. Uzuner, Ö., Sibanda, T., Luo, Y., Szolovits, P. A De-identifier for Medical Discharge Summaries. International Journal Artificial Intelligence in Medicine. 42(1): 13-35. (2008).
7. Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S., Samore, M.H. Automatic de-identification of textual documents in the electronic health record: A review of recent research. BMC Medical Res Methodology. 10: Article 70. (2010).
8. Uzuner, Ö., Juo, Y., Szolovits, P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc. 14(5):550-63. (2007).
9. 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).
10. Stubbs, A., Uzuner, Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. Journal of biomedical informatics 58 (2015): S20-S29.
11. Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodríguez, H., Lopez Martin, J., 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), CEUR Workshop Proceedings (CEUR-WS.org), (2019).
12. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805. (2018).
13. Huang, Z., Xu, W., & Yu, K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991. (2015).
14. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. Deep contextualized word representations. arXiv preprint arXiv: 1802.05365. (2018).

15. Lafferty J., McCallum A., and Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the International Conference on Machine Learning(ICML), pages 282–289. (2001).