Improving De-identification of Clinical Text with Contextualized Embeddings

Youngjun Kim, PhD[1], Stéphane M. Meystre, MD, PhD[1]
[1]Medical University of South Carolina, Charleston, South Carolina, USA

Introduction: De-identification is an important task for natural language processing (NLP) applied to clinical narratives and has recently been addressed with deep neural network approaches. As in other NLP tasks, pre-trained word representations have been employed to train more accurate de-identification models[1,2]. Furthermore, in recent years, context-dependent representations such as ELMo[3] (Embeddings from Language Models) have been proposed to overcome limitations of context-independent word embeddings. We create ELMo-enhanced Bi-LSTM (bidirectional long short-term memory networks) models trained on three publicly available corpora (the 2006 i2b2[4], 2014 i2b2[1], and 2016 CEGS N-GRID[2] shared tasks). Our de-identification models outperform previous studies[5,6] that attempted to use contextualized word representations and obtain performance comparable to state-of-the-art results.

Methods: We tackle text de-identification as a sequential labeling problem to assign a class label to each word in a sequence. We created LSTM models, a widely used RNN (recurrent neural network) variant, as a function that maps a multi-dimensional vector to a label output vector. The bidirectional LSTM (Bi-LSTM) processes a sequence as-is and also in reverse order. We also trained another LTSM model for each data set that incorporated ELMo layers (Bi-LSTM+ELMo). We used the ELMo model trained on a dataset of 5.5 billion tokens consisting of Wikipedia and the news crawl corpus. Bi-LSTM and Bi-LSTM+ELMo models were trained for 50 epochs with 25% and 50% dropouts, respectively. We calculated the average value between the ten trials of each Bi-LSTM model because of its non-deterministic results due to random weight initialization and random shuffling of training data. For comparison, we also trained CRF (conditional random fields) models, one of the shallow learning algorithms that have been widely applied to solving structured prediction problems.

Results: We used the 2016 CEGS N-GRID shared task[2] evaluation script to calculate performance metrics with strict entity matching and binary token matching. For the latter, each identifier term is evaluated by token, regardless of the identifier category. Overall, Bi-LSTM variant models outperformed CRF models on all data sets. With strict entity evaluation on the 2014 and 2016 test sets, Bi-LSTM+ELMo models achieved significantly higher F_1-scores (95.08% and 92.40%) than other methods at the 95% significance level. The model trained on 2016 N-GRID data benefited the most from the contextualized embeddings, with a 2.03% recall gain (= 92.63% – 90.60%).

Table. Accuracy of each de-identification model (Pre: Precision, Rec: Recall, F_1: F_1-score).

	2006 i2b2 Test set			2014 i2b2 Test set			2016 N-GRID Test set		
Strict	**Pre**	**Rec**	**F_1**	**Pre**	**Rec**	**F_1**	**Pre**	**Rec**	**F_1**
CRF	96.79	95.46	96.12	95.55	91.31	93.38	91.61	84.84	88.09
Bi-LSTM	**97.12**	96.65	**96.89**	95.12	93.66	94.38	92.13	90.60	91.36
Bi-LSTM+ELMo	96.46	**96.71**	96.58	**95.68**	**94.49**	**95.08**	**92.18**	**92.63**	**92.40**
Binary Token	**Pre**	**Rec**	**F_1**	**Pre**	**Rec**	**F_1**	**Pre**	**Rec**	**F_1**
CRF	99.28	97.98	98.62	**99.27**	96.31	97.77	97.64	91.26	94.34
Bi-LSTM	99.36	98.86	99.11	98.87	97.96	98.41	**96.98**	95.43	96.19
Bi-LSTM+ELMo	**99.41**	**98.99**	**99.20**	99.09	**98.18**	**98.63**	96.91	**96.93**	**96.92**

Conclusion: Our results showed that the Bi-LSTM model exploiting bidirectional language model layers achieved substantial performance improvement for clinical text de-identification. Our future efforts will focus on improving generalization across different corpora by model adaptation or combination.

References

1. 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. J. Biomed. Inform. 2015;58:S11–S19.
2. Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks track 1. J. Biomed. Inform. 2017;75:S4–S18.
3. Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. 2018 NAACL-HLT, 2018:2227–37.
4. Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. JAMIA 2007;14(5):550–63.
5. Khin K, Burckhardt P, Padman R. A deep learning architecture for de-identification of patient notes: Implementation and evaluation. arXiv preprint arXiv:1810.01570 2018.
6. Lee K, Filannino M, Uzuner Ö. An empirical test of GRUs and deep contextualized word representations on de-identification. Stud. Health Technol. Inform. 2019;264:218–22.