De-identification of Clinical Text via Bi-LSTM-CRF with Neural Language Models

Buzhou Tang, PhD[1]*, Dehuan Jiang, MS[1], Qingcai Chen, PhD[1], Xiaolong Wang, phD[1], Jun Yan, PhD[2], Ying Shen, PhD[3]

[1]Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology, Shenzhen, China

[2]Yidu Cloud (Beijing) Technology Co., Ltd, Beijing, China

[3]Ying Shen, Peking University, Shenzhen Graduate School, Shenzhen, China

***Corresponding author: tangbuzhou@gmail.com**

Abstract

De-identification of clinical text, the prerequisite of electronic clinical data reuse, is a typical named entity recognition (NER) problem. A number of state-of-the-art deep learning methods for NER, such as Bi-LSTM-CRF (bidirectional long-short-term-memory conditional random fields), have been applied for de-identification. Neural language models used for language representation bring great improvement in lots of NLP tasks when they are integrated with other deep learning methods. In this paper, we introduce Bi-LSTM-CRF with neural language models for de-identification of clinical text, and evaluate it on the de-identification datasets of the i2b2 2014 and the CEGS N-GRID 2016 challenges. Four neural language models of three types individually integrated with Bi-LSTM-CRF are compared in this study. Bi-LSTM-CRF with neural language models achieves the highest "strict" micro-averaged F1-score of 95.50% on the i2b2 2014 dataset and 91.82% on the CEGS N-GRID 2016 dataset, becoming new benchmark results on these two datasets respectively.

Keywords: De-identification, Named entity recognition, Bidirectional long-short-term-memory, Conditional random fields, Neural language models.

Introduction

Clinical records as an important information source for medical research and investigations have been attracting more and more attention of medical professionals. A prerequisite for clinical records sharing is removing individually identifiable health information, known as protected health information (PHI). In the United States, the Health Insurance Portability and Accountability Act (HIPAA)[1]defines 18 different types of PHI. In clinical records, there is plenty of PHI embedded in clinical text, which cannot be identified directly like PHI recorded in structured table according to table fields.

The process of finding and removing PHI is called de-identification. As removing PHI is easy to implement, de-identification mainly focuses on finding PHI, and is usually recognized as a named entity recognition task. In the last decades, a variety of methods have been proposed for de-identification. These methods can be classified into two categories: rule-based methods and machine learning methods. Rule-based methods rely on manually constructed regular expressions and domain dictionaries, which are challenging and time-consuming to develop. They are easy to implement and do not require any labeled data. The main disadvantage of rule-based methods lies in that they are not easy to replant from one domain to another. To prevent disadvantages of rule-based methods, researchers have attempted machine learning methods for de-identification of clinical text, especially since several challenges about de-identification were organized, such as the i2b2 2014 challenge [2],[3] and the CEGS N-GRID 2016 challenge[4]. Classical statistical learning methods, such as decision trees [5], support vector machines [6], conditional random fields (CRF)[7] and structured support vector machines [8], have been applied for de-identification. Among them, CRF is one of the most state-of-the-art methods. All these methods require time-consuming feature engineering.

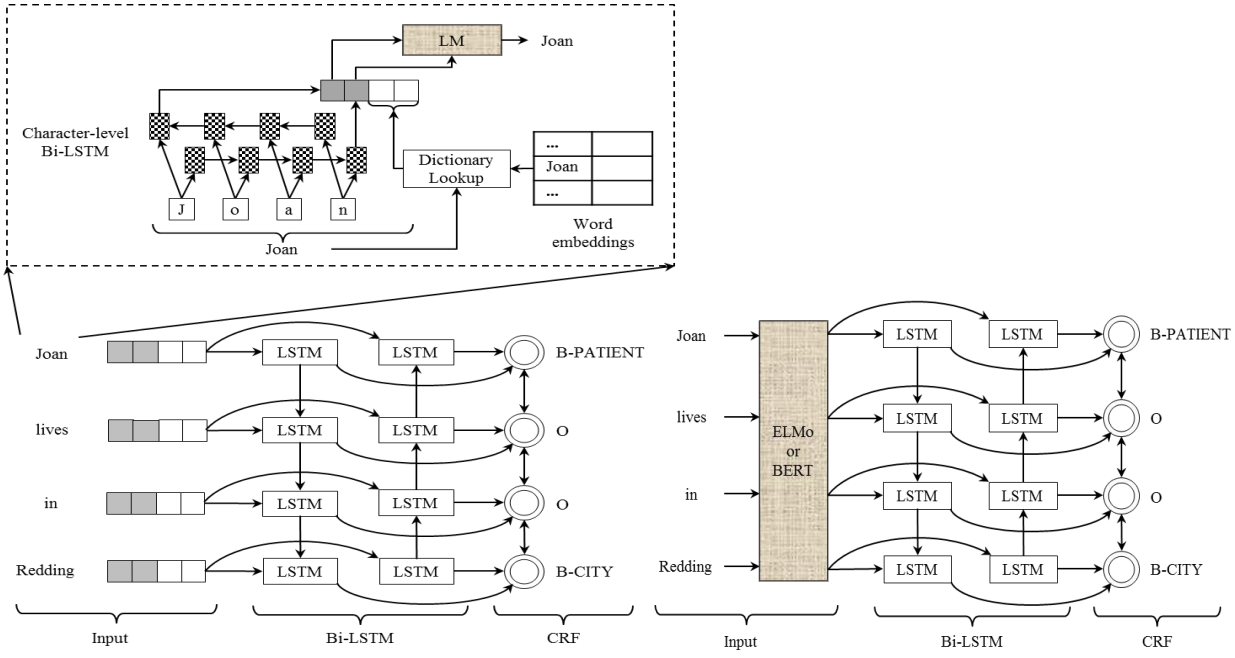
In recent years, deep neural networks, which have the ability of learning effective features from large-scale unlabeled data instead of feature engineering in traditional statistical learning methods, have been widely used in various tasks in natural language processing, such as language representation, NER, parsing, text classification, question answering and machine translation, and have shown promising results. For de-identification, deep neural networks also achieve state-of-the-art results [9],[10] . The representative method is Bi-LSTM-CRF (bidirectional long-short-term-memory conditional random fields)[10-15]. Bi-LSTM-CRF takes as input the sequence of word embeddings

and uses Bi-LSTM [16] for sentence representation and CRF for label sequence prediction. The input sequence of word embeddings learnt by neural language models on large-scale unlabeled data is an important influencing factor of Bi-LSTM-CRF. The neural language models may be classified into the following three types: 1) The common neural language models such as CBOW[17], Skip-gram[18] and GloVe[19] are usually used to obtain representations of words. Each word learnt by these models has only one representation, denoted by word embedding. However, a great percentage of words have different meanings when appearing in different context. 2) Recently, several neural language models that determine the embeddings of a word according to its context at each time dynamically, such as ELMo[20], GPT[21] and BERT[22], have been proposed and have brought significant improvement to existing methods. 3) Another way to improve existing methods using neural language models is optimizing neural language models and existing methods simultaneously. For example, Liu L et al.[23]. introduced a character-level Bi-LSTM to predict each word of the input sentence, which was optimized together with Bi-LSTM-CRF. In this way, Bi-LSTM-CRF obtained improvement on a number of NLP tasks.

In this paper, we investigate Bi-LSTM-CRF with neural language models of the above-mentioned three types on de-identification and compare four neural language models, that is, Skip-gram[18], ELMo[20], BERT[22] and LM[23] . Experiments conducted on the de-identification datasets of the i2b2 2014 challenge and the CEGS N-GRID 2016 challenge show that Bi-LSTM-CRF with neural language models achieves state-of-the-art performance. Bi-LSTM-CRF(BERT) is the best one with a "strict" micro-averaged F1-score of 95.50% on the i2b2 2014 de-identification dataset and 91.82% on the CEGS N-GRID 2016 challenge dataset, which is higher than that of the current best system[9] by 0.39%.

Methods

We start with the basic Bi-LSTM-CRF, and then the variants when different neural language models are added separately. Figure 1 shows their overall architecture.



(a)Skip-gram and LM (b) ELMo and BERT

Figure 1. Overview architecture of Bi-LSTM-CRF with neural language models

Bi-LSTM-CRF

It consists of three main components, called "layers" here, that are input layer, Bi-LSTM layer and CRF-layer. The input layer converts an input sentence $s=w_1w_2\ldots w_n$ into a sequence of word embeddings $x=x_1x_2\ldots x_n$. The Bi-LSTM layer further obtains two sequences $h_f=h_{f1}\ h_{f2}\ldots\ h_{fn}$ and $h_b=h_{b1}\ h_{b2}\ldots\ h_{bn}$, which represent context information of each

word of interest at every position from forward and backward directions respectively. The CRF layer finally predicts a label sequence $y=y_1y_2\ldots y_n$. For detailed information, please refer to [13].

Bi-LSTM-CRF with Neural Language Models

There are two ways to integrate neural language models into Bi-LSTM-CRF: at the input layer and at the CRF layer. At the input layer, there are two types of representations for input sentences: word-level and sentence-level. At the CRF-layer, the input sentences should be also produced by neural networks like autoencoding.

Skip-gram is a model for word-level representations, the idea of which is to use a word of interest to predict words around it. Given large-scale unlabeled data, Skip-gram produces one embedding for each word. Then any input sentence can be represented by replacing each word by its word embedding, which is a process of dictionary lookup.

ELMo is a model for sentence-level representation. It uses stacked bidirectional LSTMs to encode input sentences, each word of which is predicted by the forward LSTM and the backward LSTM based on the two subsentences before and after it respectively. The same word in different sentences may have different meanings and should have different representations.

Similar to ELMo, BERT is also a model sentence-level representation. It adopts neural networks based on transformer to encode input sentences, where some of the words from input sentences are randomly masked, and are predicted based only on its context. Other than that, it further introduces a "next sentence prediction" task for text pair representation.

LM is a task-aware neural language model, which predicts each word based on character-level context and is optimized together with a label sequence task in a multi-task framework. When LM is integrated into Bi-LSTM-CRF at the CRF layer, another neural language model such as Skip-gram is also integrated at the input layer. In this study, we only consider the case of Bi-LSTM-CRF(Skip-gram+LM).

Experimental Settings and Results

Datasets

We compare different Bi-LSTM-CRF variants on the de-identification datasets of the i2b2 2014 challenge and the CEGS N-GRID 2016 challenge, which are publicly available. The i2b2 2014 dataset includes a training set of 790 records with 17,045 PHI instances and a test set of 514 records with 11,462 PHI instances. The CEGS N-GRID 2016 dataset includes of a training set of 600 records with 20,845 PHI instances and a test set of 400 records with 13,521 PHI instances. All PHI instances are classified into seven main categories with subcategories. The number of PHI instances in each category is listed in Table 1, where NA denotes no subcategory, and categories defined in HIPAA are marked by *.

Experimental Settings

Evaluation: The performance of all models is measured by standard micro-averaged precision (P), recall (R) and F1-score (F1) under different criteria, namely "token", "strict", "relaxed", "HIPAA token", "HIPAA strict", "HIPAA relaxed", "binary token", "binary strict", "binary HIPAA token" and "binary HIPAA strict", where "token" and "strict" correspond to exact matching at token-level and instance-level respectively, "relaxed" allows at most two characters mismatched at the end, "HIPAA" corresponds HIPAA-defined categories, and "binary" only considers the boundaries of PHI instances no matter their categories. "strict" is the primary criterion. We use the same preprocessing as our previous work [9] in this study.

Parameters: We adopt the same settings as [9] for Bi-LSTM-CRF(Skip-gram) except that all categories of PHI instances are recognized by Bi-LSTM-CRF(Skip-gram) in this study, while some categories of PHI instances are recognized by rules in [9]. We use the pretrained ELMo (https://s3-us-west-2.amazonaws.com/allennlp/models/elmo/2x4096_512_2048cnn_2xhighway/elmo_2x4096_512_2048cnn_2xhighway_weights.hdf5) and BERT-Base (https://storage.googleapis.com/bert_models/2018_11_03/multilingual_L-12_H-768_A-12.zip) for neural language model initialization and they are fine-tuned during Bi-LSTM model training. The hyperparameters used in our study are listed in Table 2 in detail.

Table 1. Number of PHI instances in each category in the two de-identification datasets.

Corpus		i2b2 2014			CEGS N-GRID 2016		
Category		**Training**	**Test**	**Total**	**Training**	**Test**	**Total**
NAME	**PATIENT***	1317	881	2198	1270	837	2107
	DOCTOR	2894	1913	4807	2396	1567	3963
	USERNAME	264	92	356	25	0	25
PROFESSION	**NA**	234	179	413	1471	1010	2481
LOCATION	**COUNTRY**	66	117	183	666	376	1042
	STATE	312	190	502	662	481	1143
	CITY*	394	260	654	1394	820	2214
	STREET*	216	136	352	46	34	80
	ZIP*	212	140	352	23	17	40
	HOSPITAL	1437	875	2312	2197	1327	3524
	ORGANIZATION*	124	82	206	1113	698	1811
	ROOM	0	0	0	0	0	0
	DEPARTMENT	0	0	0	0	0	0
	OTHER	4	13	17	25	19	44
AGE*	**NA**	1233	764	1997	3636	2354	5990
DATE*	**NA**	7488	4985	12473	5723	3822	9545
CONTACT	**PHONE***	309	215	524	143	113	256
	FAX*	8	2	10	4	5	9
	EMAIL*	4	1	5	2	5	7
	URL	2	0	2	5	3	8
	IPADDR	0	0	0	0	0	0
ID	**MEDICALRECORD***	611	422	1033	4	2	6
	SSN*	0	0	0	0	0	0
	DEVICE*	7	8	15	0	0	0
	IDNUM*	261	195	456	2	8	10
	BIOID*	1	0	1	0	0	0
	HEALTHPLAN*	1	0	1	0	2	2
	VEHICLE*	0	0	0	0	0	0
	ACCOUNT*	0	0	0	0	0	0
	LICENSE*	0	0	0	38	21	59

Table 2. Hyperparameters used in our study.

Hyperparameter	Value
Dimension of token-level word embeddings	Skip-gram: 50, ELMo: 1024, BERT: 768
Dimension of character-level word embeddings	Skip-gram:25
Character-level LSTM size	Skip-gram:25
Token-level LSTM size	100
Dropout probability	0.5
Learning rate	0.005
Training epochs	80

Experimental Results

The comparison results of the Bi-LSTM-CRF with different neural language models on the two datasets are shown in Table 3 and Table 4 respectively, where the highest P, R and F1 are highlighted in bold. The Bi-LSTM-CRF with the two neural language models, namely ELMo and BERT, for sentence-level representation achieves higher F1-score than the Bi-LSTM-CRF with the one neural language model, namely Skip-gram, for word-level representation. Bi-LSTM-CRF(BERT) achieves the highest "strict" F1-score of 95.50% on the i2b2 2014 dataset and 91.82% on the CEGS N-GRID 2016 dataset, outperforming LSTM-CRF(ELMo) by 1.54% and 1.12% repectively. LM

brings an improvement of 0.72% in "strict" F1-score on the i2b2 2014 dataset and 1.19% on the CEGS N-GRID 2016 dataset, when it is added into Bi-LSTM-CRF(Skip-gram). LSTM-CRF(ELMo) and Bi-LSTM-CRF(Skip-gram+LM) shows comparative performance.

Table 3. Comparison of Bi-LSTM-CRF variants with different neural language models on the i2b2 2014 dataset.

Criterion	Bi-LSTM-CRF (Skip-gram)	Bi-LSTM-CRF (ELMo)	Bi-LSTM-CRF (BERT)	Bi-LSTM-CRF (Skip-gram+LM)
	P/R/F1(%)	P/R/F1(%)	P/R/F1(%)	P/R/F1(%)
Token	96.79/95.91/96.35	96.96/96.76/96.86	**97.79/97.16/97.48**	97.49/96.55/97.02
Strict	94.58/93.35/93.96	94.76/94.14/94.45	**95.99/95.02/95.5**	95.56/93.81/94.68
Relaxed	94.74/93.51/94.12	94.29/94.3/94.61	**96.17/95.2/95.69**	95.78/94.03/94.9
Binary token	98.81/97.91/98.36	98.57/98.37/98.47	**99.02/98.38/98.7**	99.06/98.1/98.58
Binary strict	96.63/95.38/96.0	96.59/95.96/96.28	**97.35/96.36/96.85**	97.28/95.51/96.39
HIPAA token	97.71/97.57/97.64	98.11/98.12/98.12	**98.8/98.28/98.54**	98.46/98.02/98.24
HIPAA strict	96.13/95.66/95.9	96.42/96.24/96.33	**97.52/96.74/97.13**	97.08/96.25/96.66
HIPAA relaxed	96.26/95.79/96.02	96.57/96.39/96.48	**97.66/96.88/97.27**	97.2/96.36/96.78
HIPAA binary token	98.08/97.94/98.01	98.35/98.36/98.36	**98.98/98.45/98.71**	98.68/98.25/98.47
HIPAA binary strict	96.50/96.03/96.26	96.67/96.49/96.58	**97.73/96.95/97.33**	97.28/96.44/96.86

Table 4. Comparison of Bi-LSTM-CRF variants with different neural language models on the CEGS N-GRID 2016 dataset.

Criterion	Bi-LSTM-CRF (Skip-gram)	Bi-LSTM-CRF (ELMo)	Bi-LSTM-CRF (BERT)	Bi-LSTM-CRF (Skip-gram+LM)
	P/R/F1(%)	P/R/F1(%)	P/R/F1(%)	P/R/F1(%)
Token	91.37/91.78/91.57	93.82/91.97/92.89	**95.36/92.32/93.81**	94.10/91.55/92.81
Strict	89.58/89.19/89.38	91.37/90.03/90.70	**93.39/90.31/91.82**	92.19/89.01/90.57
Relaxed	89.71/89.31/89.51	91.54/90.19/90.86	**93.56/90.47/91.99**	92.36/89.18/90.74
Binary token	95.19/**95.62**/95.4	96.83/94.92/95.87	**97.70**/94.58/**96.11**	97.15/94.52/95.81
Binary strict	92.63/92.22/92.42	93.75/**92.37**/93.06	**95.26**/92.11/**93.66**	94.52/91.26/92.86
HIPAA token	93.39/93.30/93.34	94.48/93.61/94.04	**96.48/93.91/95.18**	94.96/93.44/94.20
HIPAA strict	91.41/91.34/91.37	92.28/91.73/92.00	**94.71/92.31/93.50**	93.26/91.26/92.25
HIPAA relaxed	91.58/91.51/91.55	92.49/91.93/92.21	**94.92/92.52/93.70**	93.44/91.44/92.43
HIPAA binary token	94.54/94.44/94.49	95.47/**94.59**/95.03	**96.96**/94.39/**95.66**	96.02/94.49/95.25
HIPAA binary strict	92.35/92.28/92.31	92.95/92.39/92.67	**95.13/92.72/93.91**	93.97/91.96/92.95

The results of the best Bi-LSTM-CRF variant, Bi-LSTM-CRF(BERT), on the main seven categories are shown in Table 5. On the i2b2 challenge dataset, Bi-LSTM-CRF(BERT) performs best on DATE with a "strict" F1-score of 98.75% and worst on PROFESSION with a "strict" F1-score of 82.78%, while on the CEGS N-GRID 2016 dataset, Bi-LSTM-CRF(BERT) performs best on DATE with a "strict" F1-score of 97.03% and worst on ID with a "strict" F1-score of 59.70%.

Table 5. Detailed results of the best Bi-LSTM-CRF variant on the main seven categories ("strict").

Corpus	i2b2 2014			CEGS N-GRID 2016		
Category	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
NAME	95.26	94.83	95.05	94.99	92.30	93.63
PROFESSION	82.32	83.24	82.78	85.03	75.94	80.23
LOCATION	90.47	86.93	88.66	88.06	82.76	85.33
AGE	97.51	97.38	97.45	97.26	96.52	96.89
DATE	98.97	98.53	98.75	97.34	96.73	97.03
CONTACT	96.36	97.25	96.80	91.94	90.48	91.20
ID	92.97	91.04	92.00	58.82	60.61	59.70

We also compare other state-of-the-art methods with Bi-LSTM-CRF(BERT) as shown in Table 6, where only "binary token" P, R and F1 are reported as the P, R and F1 under other criteria are not complete. "Bi-LSTM-CRF(BERT)" shows better performance than the other methods for comparison.

Table 6. Comparison of other state-of-the-art methods with Bi-LSTM-CRF(BERT) ("binary token").

Corpus	i2b2 2014			CEGS N-GRID 2016		
Method	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Liu et al. [9]	97.94	96.04	96.98	95.56	90.70	93.07
Zhao et al. [14]	98.89	97.23	98.05	NA	NA	NA
Dernoncourt et al. [10]	98.34	98.53	98.44	NA	NA	NA
Kim et al. [15]	99.16	98.06	98.61	NA	NA	NA
Bi-LSTM-CRF(BERT)	99.02	98.38	98.70	97.70	94.58	96.11

Discussion

It is easy to understand that Bi-LSTM-CRF(ELMo) and Bi-LSTM-CRF(BERT) perform better than Bi-LSTM-CRF(Skip-gram) since ELMo and BERT can dynamically give more detailed representation of each word according to its context. For example, "Brighman Young" is usually a name, but an organization in "Father is supervisor. College (Brighman Young, community college in TX, UCSD, Baylor), still working on bachelor's, major sociology.", which is correctly recognized by Bi-LSTM-CRF(ELMo) and Bi-LSTM-CRF(BERT), but wrongly recognized as a doctor's name by Bi-LSTM-CRF(Skip-gram). Similar to previous studies in other domains[19], Bi-LSTM-CRF(BERT) shows better performance than Bi-LSTM-CRF(ELMo). As LM has ability to predict sentences regarding a specific task, Bi-LSTM-CRF(Skip-gram+LM) can fit de-identification here.

Although Bi-LSTM-CRF(BERT) shows promising overall performance, there are still some errors. Taking the results on the CEGS N-GRID 2016 dataset for example, the effect of boundary errors or category errors on overall F1-score is around 2.0% (the "strict" F1-score of 91.82% vs the "token" F1-score of 93.81% or the "binary strict" F1-score of 93.66%). The "strict" F1-scores of Bi-LSTM-CRF(BERT) on PROFESSION, LOCATION and ID are much lower than other categories as shown in Table 4. We look into the errors in these three categories and find that: 1) Lots of instances in PROFESSION are not detected as they are not mentioned directly like PHI instances in other categories, for example, "danced" is not a word to denote PROFESSION directly, but we can infer that the patient is a dancer from "She danced with MBT for 10 years;". 2) Boundary errors widely exist in LOCATION, for example, in "video editing (DiC Entertainment Affiliate) in broadcasting Financial Stress: Yes", "DiC Entertainment Affiliate" is recognized as an ORGANIZATION instance instead of "DiC Entertainment". The "binary strict" F1-score on LOCATION is 89.13%, higher than the "strict" F1-score on LOCATION (i.e., 85.33%) by 3.80%. Other than that, there are also more category errors on LOCATION than other categories because of more subcategories in LOCATION. 3) The reason why Bi-LSTM-CRF(BERT) performs bad on ID is that the number of ID instances is too small. From Table 1, we can see there are only 44 ID instances in the training set of the CEGS N-GRID 2016 challenge, in particular, only 2 instances in IDNUM.

For the future work, there may be two directions for further improvement: 1) using BERT pretrained on large-scale clinical data; 2) integrating LM and BERT into Bi-LSTM-CRF at the same time since they work at different layer.

Conclusion

In this study, we introduce Bi-LSTM-CRF with neural language models for de-identification of clinical text. The Bi-LSTM-CRF variants achieve the highest "strict" micro-averaged F1-score of 95.50% on the i2b2 2014 de-identification dataset and 91.82% on the CEGS N-GRID 2016 de-identification dataset, constituting new benchmarks for de-identification.

Acknowledgement

This paper is supported in part by grants: NSFCs (National Natural Science Foundations of China) (U1813215, 61876052 and 61573118), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20160531192358466), Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052).

References

1. Act A. Health insurance portability and accountability act of 1996. Public law. 1996;104:191.
2. Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. Journal of biomedical informatics. 2015;58:S20–S29.
3. Uzuner Ö, Stubbs A. Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks. Journal of biomedical informatics. 2015;58(Suppl):S1.
4. Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1. Journal of biomedical informatics. 2017;75:S4–S18.
5. Quinlan JR. Induction of decision trees. Machine learning. 1986;1(1):81–106.
6. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their applications. 1998;13(4):18–28.
7. Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001;
8. Xue H, Chen S, Yang Q. Structural support vector machine. In: International Symposium on Neural Networks. Springer; 2008. p. 501–511.
9. 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. 2017;75:S34–S42.
10. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association. 2017;24(3):596–606.
11. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. arXiv preprint arXiv:160301360. 2016;
12. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:150801991. 2015;
13. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC medical informatics and decision making. 2017;17(2):67.
14. Zhao Y-S, Zhang K-L, Ma H-C, Li K. Leveraging text skeleton for de-identification of electronic medical records. BMC medical informatics and decision making. 2018;18(1):18.
15. Kim Y, Heider P, Meystre S. Ensemble-based Methods to Improve De-identification of Electronic Health Record Narratives. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association; 2018. p. 663.
16. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735–1780.
17. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013;
18. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013. p. 3111–3119.
19. 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). 2014. p. 1532–1543.
20. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv preprint arXiv:180205365. 2018;
21. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding with unsupervised learning. Technical report, OpenAI; 2018.
22. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018;
23. Liu L, Shang J, Ren X, Xu FF, Gui H, Peng J, et al. Empower sequence labeling with task-aware neural language model. In: Thirty-Second AAAI Conference on Artificial Intelligence. 2018.