Clinical Named Entity Recognition Using Deep Learning Models

Yonghui Wu, PhD, Min Jiang, MS, Jun Xu, PhD, Degui Zhi, PhD, Hua Xu, PhD
School of Biomedical Informatics, the University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

Clinical Named Entity Recognition (NER) is a critical natural language processing (NLP) task to extract important concepts (named entities) from clinical narratives. Researchers have extensively investigated machine learning models for clinical NER. Recently, there have been increasing efforts to apply deep learning models to improve the performance of current clinical NER systems. This study examined two popular deep learning architectures, the Convolutional Neural Network (CNN) and the Recurrent Neural Network (RNN), to extract concepts from clinical texts. We compared the two deep neural network architectures with three baseline Conditional Random Fields (CRFs) models and two state-of-the-art clinical NER systems using the i2b2 2010 clinical concept extraction corpus. The evaluation results showed that the RNN model trained with the word embeddings achieved a new state-of-the-art performance (a strict F1 score of 85.94%) for the defined clinical NER task, outperforming the best-reported system that used both manually defined and unsupervised learning features. This study demonstrates the advantage of using deep neural network architectures for clinical concept extraction, including distributed feature representation, automatic feature learning, and long-term dependencies capture. This is one of the first studies to compare the two widely used deep learning models and demonstrate the superior performance of the RNN model for clinical NER.

Introduction

Clinical studies often require detailed patients' information documented in clinical narratives. Named Entity Recognition (NER)[1] is a fundamental Natural Language Processing (NLP) task to extract entities of interest (e.g., disease names, medication names and lab tests) from clinical narratives, thus to support clinical and translational research.[2,3] Researchers have developed computational models and applied them in general clinical NLP systems. Most of the general clinical NLP systems such as MetaMap[4], MedLEE[5], and KnowledgeMap[6], applied rule-based methods that rely on existing medical vocabularies for NER. The clinical NLP community has organized several challenges to examine the performances of state-of-the-art methods.[7-10] Most of the top-performed systems[11-13] are primarily based on supervised machine learning models with manually defined features. To further improve the performance, researchers have explored various strategies within the current infrastructure of conventional machine learning models, including ensemble models that combine multiple machine learning methods[14,15], hybrid systems that combine machine learning with high-confidence rules,[16] unsupervised features generated using clustering algorithms [17,18] (e.g., Brown clustering[19]), and domain adaptation[20,21] to leverage labeled corpora from other domains.

Machine learning methods formulate the clinical NER task as a sequence labeling problem that aims to find the best label sequence (e.g., BIO format labels) for a given input sequence (individual words from clinical text). Researchers have applied many machine learning models, including Conditional Random Fields (CRFs)[22], Maximum Entropy (ME), and Structured Support Vector Machines (SSVMs)[23]. Many top-ranked NER systems applied the CRFs model, which is the most popular solution among conventional machine learning algorithms. A typical state-of-the-art clinical NER system usually utilizes features from different linguistic levels, including orthographic information (e.g., capitalization of letters, prefix and suffix), syntactic information (e.g. POS tags), word n-grams, and semantic information (e.g., the UMLS concept unique identifier). Some hybrid models[16] further leverage the concepts and semantic types from the existing clinical NLP systems such as MetaMap, cTAKES[24]. To further improve the performance, researchers have also utilized ensemble methods to combine different machine learning models, such as re-ranking[14,15]. More recently, researchers[17,25] also start to examine the unsupervised features derived from large volumes of unlabeled corpora, such as the word clusters generated using Brown clustering[19] and random indexing. The continuous and intensive hard work from the clinical NLP community have boosted the performance of clinical NER, while also identified several bottlenecks that impede further improvement, including:

1) *Fragile feature representation*.[26,27] Initially, the dominant feature representation in clinical NER is the bag-of-word model, which is a simplified representation of a piece of text based on the presence/absence of its words

irrespective of the orders between words, grammatical relation, and semantic information. The bag-of-word model is fragile for clinical NER due to the sparsity problem. For example, the following two similar clinical entities: "mildly dilated right atrium" and "somewhat enlarged left ventricle" have nothing in common using the bag-of-word feature representation. However, they are two related concepts. There is a need for more robust feature representations.

2) *Task-specific and time-consuming human feature engineering*.[27,28] The typical machine learning based clinical NER is composed of two steps: feature extraction and parameter optimization, where feature extraction is the most critical but time-consuming step. In the conventional machine learning solutions, the feature extraction heavily depends on humans while the machine can only handle the parameter optimization supervised by the gold-standard annotations. Researchers manually screen the positive/negative samples to identify possible features (e.g., tokens containing capitalized letters are more likely to have special meanings) and design feature combinations (e.g., body location followed by a disorder mention), commonly referred to as "human feature engineering". The human feature engineering has several problems. First, the features extracted by human are either incomplete – human cannot enumerate all possible features, or over specified – the same information is repeated in many complex features and feature combinations. Second, researchers must engineer the features again for a different task or different data source. There is an increasing need for automatic feature learning algorithms to release researchers from the time-consuming manual feature engineering.

3) *Lack of long-term dependencies*.[28-30] The typical CRFs based NER systems usually require the set up of a word window for the input sequence of tokens. Many studies that examined the system prediction errors have reported false negatives caused by the lack of long-term dependencies. However, simply increasing the window size cannot solve this problem as it may dilute the signal with more noise into the feature space and greatly increase the training time. The clinical NER need a better architecture to capture long-term dependencies from clinical texts.

Recently, there have been increasing efforts to explore a new emerging technology, deep learning[31] (or deep neural networks), to improve the current clinical NLP systems. Deep learning is a sub-domain of machine learning that uses deep architectures to learn high-level feature representations. Currently, deep neural networks are commonly used as the unique deep architecture for high-level feature learning. Deep learning models introduced word embedding[28,32] as a critical technique to train densely-valued vector representation of words to replace the fragile bag-of-word representation. Each row of the matrix is associated with a word in the vocabulary and each column of the matrix represents a latent feature. The input word sequence can be transformed into a vector by concatenating the corresponding word vectors from the embedding matrix. Deep neural network architectures can learn high-level features automatically to release researchers from time-consuming human feature engineering. To capture long-term dependencies in a word sequence, researchers designed two popular deep architectures, including the Convolutional Neural Networks (CNN)[28,33] and the Recurrent Neural Networks (RNN)[34]. Recent research from the general NLP domain reported that the CNN and RNN developed using only the word embeddings achieved comparable performance as the state-of-the-art CRFs with human engineered features and knowledge from dictionaries.[28,34] In the clinical NLP domain, in one of our previous works[35], we examined the word embedding trained using a neural network and applied the embeddings to an SVMs model for word sense disambiguation (WSD)[36] and a CRFs model for clinical NER[35]. The evaluation results showed that the word embeddings could be used as useful features to improve the state-of-the-art performances of conventional machine learning models for WSD and NER. Later, we developed a CNN based NER method[37] and applied it to the Chinese clinical notes, which outperformed a state-of-the-art CRFs model. Jagannatha et al.[38] applied RNN for medical event detection from clinical notes and later compared several RNN models with different loss functions. The experimental results using annotated cancer patient notes showed that the RNN outperformed a baseline CRFs model with context features.

However, the previous study by Jagannatha et al.[38] compared RNN with a baseline CRFs with only context word features and the evaluation corpus is not openly accessible. Our previous CNN study[37] was evaluated using Chinese clinical corpus. It is still not clear which deep architecture is better and whether they could outperform the state-of-the-art clinical NER systems for English clinical corpora. Therefore, it is necessary to further examine the two popular deep architectures, CNN and RNN, using an open challenge corpus and compare them with the most state-of-the-art clinical NER systems. This study is one of the first studies to compare the two widely used deep learning models using an open clinical corpus – the i2b2 2010 clinical concept extraction dataset. We compared the two deep learning models with the state-of-the-art clinical NER systems with human designed features and demonstrated the superior performance of RNN model.

Methods

Data sets

In this study, we reused the word embeddings developed in our previous study[35] using the unlabeled Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II corpus[39]. The MIMIC II corpus is composed of 403,871 notes from four different note types including discharge notes, radiology notes, ECG and ECHO notes. Table 1 shows the detailed information about the MIMIC II corpus. We used the labeled corpus developed for the Concept Extraction task of the 2010 i2b2 NLP challenge for model training and evaluation. We used the same training and test data sets as in the challenge, which consisted of 349 notes for training and 477 notes for testing. For each note, annotators manually extracted entities about Problem, Treatment, and Test. Table 1 shows detailed information for both the training dataset and the test data set.

Table 1. Descriptive statistics of MIMIC II corpus and i2b2 dataset.

Data set		Notes	Entities	Entity types
i2b2 2010	Training	349	27,837	Problem, Treatment
	Test	477	45,009	Test
MIMIC II	N/A	403,871	N/A	N/A

Machine learning models

Baseline CRFs systems

The CRFs model approaches clinical NER as a sequence labeling problem that tries to find the best label sequence $Y^*=y_1\ y_2 \ldots y_N$ for a given input sequence $X=x_1\ x_2 \ldots\ x_N$. The clinical concept extraction task can be converted into a sequence labeling problem by assigning the annotated entities with appropriate tag representations. This study used the standard "BIO" schema to convert the named entity annotations into sequence labeling tags, in which each word is assigned to a label as following: B = beginning of an entity, I = inside an entity, and O = outside of an entity. As the i2b2 2010 concept extraction challenge focused on three types of clinical concepts, we used a total number of seven different tags (B-problem, B-test, B-treatment, I-problem, I-test, I-treatment, O). For comparison, we constructed three CRFs baseline systems using features extracted at different levels. The first system used only the context word and n-gram features. Based on the basic features, the second systems further added linguistic features and document level features (section names). Similarly, the third system further added knowledge features derived from general clinical NLP systems (MedLEE, MetaMap and KnowledgeMap) and dictionary match features from existing vocabularies (UMLS). All the baseline systems were developed using the CRFs implemented in the CRF++ package.

State-of-the-art NER systems

To challenge the deep learning models, we selected two state-of-the-art systems among all studies regarding the i2b2 2010 clinical concept extraction challenge dataset. The first system is the best-performed system developed by Debruijn et al.[11]. This system explored many typical NER features including word features from different linguistic levels, knowledge from existing clinical NLP systems, and Brown clustering. Authors proposed a semi-supervised Markov model solution and achieved the best F1 score during this challenge. The second system[25] applied SSVMs model and further examined the distributional word representation feature generated by Random Indexing algorithm. To the best of our knowledge, this system achieved the best after-challenge performance on the i2b2 2010 dataset up until now.

Deep learning based NER system

Similar to the CRFs model, deep learning models formulate the clinical concept extraction task into a sequence labeling problem using the same “BIO” tagging schema. The input of deep learning models is quite different with conventional machine learning models. The input of CRFs model is human designed features represented in bag-of-word style vector. Whereas, the input for deep learning model is the raw sequence of words in the sentence without human engineering. We applied a word embedding layer to transform the sequence of words into densely-valued vectors, where most of the values are non-zero. Next, we train a deep neural network to learn high-level feature

representations, capture long-term dependencies, and global features to help identify clinical entities. This study explored two deep architectures including CNN and RNN.

CNN architecture: This study adopted a popular CNN architecture using sentence level log-likelihood approach proposed by Dr. Ronan Collobert.[28] This architecture consists of a convolutional layer, a non-linear layer using the hard version of the hyperbolic tangent (HardTanh), and several standard linear layers. This architecture achieved state-of-the-art NER performance in the general English domain and later was applied to many other NLP tasks later. Using word embeddings, each word in the input window can be mapped to an N-dimension vector (N is the embedding dimension). Then, a convolution layer captures the long-term dependencies (or global features) in the hidden nodes. Both the local features and the global features are then concatenated together and fed into a standard neural network trained using stochastic gradient descending. Finally, the classification layer utilized a sentence level log-likelihood approach to calculate the loss supervised by gold-standard annotations. Detailed information for this CNN architecture can be found from our previous study.

RNN architecture: RNN is an emerging new deep architecture for sequence data. Recent studies have shown that RNNs have good ability to capture long-term dependencies for sequence data. In this study, we adopted a RNN architecture implemented using the Long Short Term-Memory (LSTM) by Lample et al.[34] The LSTM is the most popular implementation of RNN architecture. A basic LSTM unit is composed of three multiplicative gates, including an input gate to control the proportion of input information transferred to a memory cell; a forget gate to control the proportion of historical information from the previous state; and an output gate to control the proportion of output information to pass on to the next step. We also applied several standard deep learning techniques including character embedding and dropout. For the input layer, we combined the word embeddings and the character embeddings in an input vector. The word embeddings were pre-trained from the MIMIC II corpus and the character embeddings were initiated with random values. The final classification layer of the RNN used a CRFs loss function, which is similar to the CNN architecture examined in this study.

Experiments and Evaluation

This study used word embeddings with 50 dimensions, which was pre-trained from the MIMIC II corpus in one of our previous studies[35]. We used a neural network with negative sampling to train the embeddings as our previous study showed that this embedding is better or at least comparable to the word2vec algorithm. For CNN, we used a Java implementation developed in one of our previous study[37]. We compared several combinations of network parameters based on our previous study and finally used the following parameters for CNN: learning rate at 0.01, the word embedding dimension at 50, and hidden node number at 300. For RNN, we adopted a Python implementation using Theano package. Based on the parameters reported by Lample et al.[34], we examined the character embedding sizes and learning rates. The final RNN model used the following parameters: word embedding dimension at 50; character embedding dimension at 25; the LSTM layer for word level is 100 and the LSTM layer for the character level is 25; learning rate at 0.005; dropout probability is 0.5. A Nvidia Tesla K40 GPU was used to train the RNN model. The official evaluation scripts provided by the i2b2 organizers were used to calculate the strict micro-averaged precision, recall, and F1-score.

Results

Table 2 compares the performances of the baseline CRFs, the Semi-Markov (best system during the challenge), the SSVMs (current best system developed after the challenge), the CNN and RNN using the i2b2 2010 test dataset. All evaluation scores were based on exact matching. For the baseline CRFs, the linguistic and document level features improved the baseline F1 score from 77.33% to 79.87%. The knowledge base features further boosted the score to 83.60%. For the state-of-the-art systems, the semi-Markov model developed by Debruijn et al. during the challenge further explored a semi-supervised word cluster feature using Brown clusters and applied a semi-supervised Markov model, which achieved the best F1 score at 85.23%. Later, Tang et al. explored the SSVMs model and distributional word representation from the Random Indexing algorithm and further improved the F1 score to 85.82%, which is the best performance score reported using the i2b2 2010 dataset. For deep learning models, the CNN with only word embeddings outperformed the baseline CRFs with word, linguistic and document level features. However, the performance of CNN did not outperform the baseline CRFs further combined with knowledge features. The RNN architecture implemented using bi-direction LSTM neurons outperformed the current best system and achieved the best F1 score (85.91%) on this data set using only the word/character embeddings.

Table 2. Performance comparison of all machine learning models.

Approach	Feature	Precision (%)	Recall (%)	F1 Score (%)
CRFs baselines	Word	82.32	72.92	77.33
	Word+Linguistic+Discourse	83.25	76.75	79.87
	Word+Linguistic+Discourse+ MedLEE+KnowledgeMap+DST	86.52	81.04	83.60
SSVMs by Tang et al. (Current best)	All features in CRF baselines +Brown clustering + Random indexing	**87.38**	84.31	85.82
Semi-Markov (Best in challenge)	Word+context+sentence+section+cTAKES +MetaMap+ConText+Brown clustering	86.88	83.64	85.23
CNN	Word embedding	84.91	80.73	82.77
RNN	Word embedding	85.33	**86.56**	**85.94**

Word: bag-of-word, orthographic such as capitalized letters and special symbols; Discourse: sections, note types; Linguistic: part of speech tags, prefix, and suffix; DST: Dictionary-based Semantic Tagger using UMLS

Discussion

This study examined two deep learning architectures to extract concepts from clinical texts. We constructed three baseline systems using the CRFs model with different levels of features. Two deep learning architectures, including a CNN and an RNN, were developed. We compared the deep learning architectures with the state-of-the-art clinical NER systems using i2b2 2010 corpus. The experimental results using the standard training, test and evaluation showed that the RNN model trained with only word embeddings achieved the new state-of-the-art performance for clinical NER, which outperformed the best system during the challenge and the current best system based on an SSVMs model. This study shows the advantage of using deep neural network architectures for information extraction from clinical texts. To the best of our knowledge, this is the first study comparing two popular deep learning architectures (CNN and RNN) for clinical concept extraction.

All deep learning models developed using only word embeddings outperformed the baseline CRFs with basic word level, linguistic level, and document level features, showing the efficiency of automatic feature learning from large unlabeled corpora. The performance improvements of the deep learning architectures are mainly from the recall (from 72.92% to 86.56% for RNN), showing that the unsupervised feature learning can capture extra features that did not exist in the training dataset to boost the recall. The RNN outperformed other systems with a new state-of-the-art F1 score at 85.94%. To the best of our knowledge, this is the best performance ever reported for the i2b2 2010 clinical concept extraction dataset. The RNN architecture outperformed CNN, another deep neural network architecture designed to learn high-level feature representations, showing that the RNN architecture is more efficient for sequence labeling tasks.

As an emerging technology, deep learning provides distributed word representation to replace the fragile bag-of-word model, automatic high-level feature learning to release researchers from time-consuming feature engineering, and deep architectures to capture long-term dependencies. We do observed differences between the general NLP and clinical NLP when applying deep learning models. For example, the clinical NLP have more knowledge bases (e.g., UMLS) with decent coverage. However, it is very hard to generate such comprehensive knowledge bases with decent coverage in the general NLP domain. Therefore, we can see that most of the top-performing clinical NER systems utilized the knowledge from dictionaries and integrated with other clinical NLP systems, which may make it hard for the deep learning models using only word embeddings to compete with the traditional clinical NER models in the clinical domain. For example, the CNN architecture outperformed the state-of-the-art NER systems on both the general English NER dataset and the Chinese clinical corpus, where there are limited knowledge bases. Whereas, the CNN model did not outperform the state-of-the-art clinical NER systems on the i2b2 corpus, where the top

systems utilized many knowledge bases and were hybrid systems using multiple general clinical NLP systems. The success of the RNN architecture shows the promise of using deep learning architectures for clinical NLP: machines can learn features better than humans, given the correct deep neural network architectures and advanced optimization algorithms.

This study has limitations. We only explored hidden layer dimension and learning rate but used arbitrarily selected network parameters according to previous studies. More robust and efficient parameter tuning strategy is needed. We compared RNN with the current best NER system and reported a new state-of-the-art performance on the i2b2 data set. Although the numeric improvement looks small, this study demonstrated the efficiency of RNN using the minimal feature engineering. Deep learning architectures provide a unified solution for clinical NLP, which may simplify the complex system architecture of current state-of-the-art clinical NLP systems to speed up their application in practical NLP systems. With the distributed word representation and high-level feature abstraction, deep learning may achieve the state-of-the-art performance without the requirement of combining with other models and systems. It worth further investigating on how the deep learning models can be applied to practical clinical NLP systems.

Conclusion

This study compared two deep learning models, including the CNN and RNN, to extract clinical concepts from clinical texts. We compared CNN and RNN with baseline CRFs using different levels of features as well as the state-of-the-art clinical NER systems using the i2b2 2010 clinical concept extraction corpus. The experimental results show that the RNN outperformed the best clinical NER system based on SSVMs with a new state-of-the-art F1 score at 85.94%. The RNN model can be adapted to other clinical NLP tasks to improve their performances and to release researchers from time-consuming feature engineering.

Acknowledgement

This study was supported by grants from the NLM 2R01LM010681, NIGMS 1R01GM103859, NCI U24 CA194215 and 1R01GM102282. We would like to thank the 2010 i2b2/VA challenge organizers for the development of the corpus used in this study. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

1. Nadeau D, Sekine S. A survey of named entity recognition and classification. *Lingvisticae Investigationes.* 2007;30(1):3-26.

2. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics.* 2008:128-144.

3. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *Journal of the American Medical Informatics Association : JAMIA.* Sep-Oct 2011;18(5):544-551.

4. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA.* May-Jun 2010;17(3):229-236.

5. Friedman C. Towards a comprehensive medical language processing system: methods and issues. *Proceedings : a conference of the American Medical Informatics Association. AMIA Fall Symposium.* 1997:595-599.

6. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A, 3rd. The KnowledgeMap project: development of a concept-based medical school curriculum database. *AMIA ... Annual Symposium proceedings. AMIA Symposium.* 2003:195-199.

7. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association : JAMIA.* Sep-Oct 2010;17(5):514-518.

8. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA.* Sep-Oct 2011;18(5):552-556.

9. Pradhan S, Elhadad N, South BR, et al. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association : JAMIA.* Jan 2015;22(1):143-154.

10. Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G. SemEval-2014 Task 7: Analysis of Clinical Text. *SemEval 2014.* 2014;199(99):54.

11. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association : JAMIA.* Sep-Oct 2011;18(5):557-562.

12. Zhang Y, Wang J, Tang B, et al. UTH_CCB: A Report for SemEval 2014–Task 7 Analysis of Clinical Text. *SemEval 2014.* 2014:802.

13. Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association : JAMIA.* Sep-Oct 2013;20(5):828-835.

14. Sil A, Yates A. Re-ranking for joint named-entity recognition and linking. *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. San Francisco, California, USA: ACM; 2013:2369-2374.

15. Yoshida K, Tsujii Ji. Reranking for biomedical named-entity recognition. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Prague, Czech Republic: Association for Computational Linguistics; 2007:209-216.

16. Jiang M, Chen Y, Liu M, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association : JAMIA.* Sep-Oct 2011;18(5):601-606.

17. Tang B, Cao H, Wang X, Chen Q, Xu H. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international.* 2014;2014:240403.

18. Tang B, Wang X, Wu Y, Jiang M, Wang J, Xu H. Recognizing Chemical Entities in Biomedical Literature using Conditional Random Fields and Structured Support Vector Machines. Paper presented at: BioCreative Challenge Evaluation Workshop vol. 22013.

19. Brown PF, Desouza PV, Mercer RL, Pietra VJD, Lai JC. Class-based n-gram models of natural language. *Computational linguistics.* 1992;18(4):467-479.

20. Chiticariu L, Krishnamurthy R, Li Y, Reiss F, Vaithyanathan S. Domain adaptation of rule-based annotators for named-entity recognition tasks. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, Massachusetts: Association for Computational Linguistics; 2010:1002-1012.

21. Guo H, Zhu H, Guo Z, Zhang X, Wu X, Su Z. Domain adaptation with latent semantic association for named entity recognition. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: Association for Computational Linguistics; 2009:281-289.

22. Lafferty JD, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*: Morgan Kaufmann Publishers Inc.; 2001:282-289.

23. Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large Margin Methods for Structured and Interdependent Output Variables. *J. Mach. Learn. Res.* 2005;6:1453-1484.

24. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA.* Sep-Oct 2010;17(5):507-513.

25. Tang B, Cao H, Wu Y, Jiang M, Xu H. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. *BMC Med Inform Decis Mak.* 2013;13 Suppl 1:S1.

26. Wang M, Manning CD. Effect of Non-linear Deep Architecture in Sequence Labeling. *Proceedings of the Sixth International Joint Conference on Natural Language Processing.* 2013:1285-1291.

27. Socher R, Bengio Y, Manning CD. Tutorial at Association of computational linguistics, 2013. 2013; http://www.anthology.aclweb.org/attachments/P/P12/P12-4005.Presentation.pdf. Accessed March, 2017.

28. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* 2011;12:2493-2537.

29. Zhang R, Lee H, Radev D. Dependency sensitive convolutional neural networks for modeling sentences and documents. *arXiv preprint arXiv:1611.02361.* 2016.

30. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems*. Montreal, Canada: MIT Press; 2014:3104-3112.

31. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* May 28 2015;521(7553):436-444.

32. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.* 2013.

33. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers)*: Association for computational linguistics; 2014.

34. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360.* 2016.

35. Wu Y, Xu J, Jiang M, Zhang Y, Xu H. A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text. *AMIA ... Annual Symposium proceedings. AMIA Symposium.* 2015;2015:1326-1333.

36. Wu Y, Xu J, Zhang Y, Xu H. Clinical Abbreviation Disambiguation Using Neural Word Embeddings. *ACL-IJCNLP 2015.* 2015:171.

37. Wu Y, Jiang M, Lei J, Xu H. Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network. *Studies in health technology and informatics.* 2015;216:624-628.

38. Jagannatha AN, Yu H. Bidirectional RNN for Medical Event Detection in Electronic Health Records. *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting.* Jun 2016;2016:473-482.

39. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. *Critical care medicine.* May 2011;39(5):952-960.