Original Paper

# Combining Contextualized Embeddings and Prior Knowledge for Clinical Named Entity Recognition: Evaluation Study

Min Jiang, PhD; Todd Sanger, PhD; Xiong Liu, PhD

Eli Lilly and Company, Indianapolis, IN, United States

**Corresponding Author:**
Min Jiang, PhD
Eli Lilly and Company
893 Delaware St
Indianapolis, IN,
United States
Phone: 1 615 926 8277
Email: jiang_min@lilly.com

## Abstract

**Background:** Named entity recognition (NER) is a key step in clinical natural language processing (NLP). Traditionally, rule-based systems leverage prior knowledge to define rules to identify named entities. Recently, deep learning–based NER systems have become more and more popular. Contextualized word embedding, as a new type of representation of the word, has been proposed to dynamically capture word sense using context information and has proven successful in many deep learning–based systems in either general domain or medical domain. However, there are very few studies that investigate the effects of combining multiple contextualized embeddings and prior knowledge on the clinical NER task.

**Objective:** This study aims to improve the performance of NER in clinical text by combining multiple contextual embeddings and prior knowledge.

**Methods:** In this study, we investigate the effects of combining multiple contextualized word embeddings with classic word embedding in deep neural networks to predict named entities in clinical text. We also investigate whether using a semantic lexicon could further improve the performance of the clinical NER system.

**Results:** By combining contextualized embeddings such as ELMo and Flair, our system achieves the F-1 score of 87.30% when only training based on a portion of the 2010 Informatics for Integrating Biology and the Bedside NER task dataset. After incorporating the medical lexicon into the word embedding, the F-1 score was further increased to 87.44%. Another finding was that our system still could achieve an F-1 score of 85.36% when the size of the training data was reduced to 40%.

**Conclusions:** Combined contextualized embedding could be beneficial for the clinical NER task. Moreover, the semantic lexicon could be used to further improve the performance of the clinical NER system.

***(JMIR Med Inform 2019;7(4):e14850)*** doi: 10.2196/14850

**KEYWORDS**

natural language processing; named entity recognition; deep learning; contextualized word embedding; semantic embedding; prior knowledge

## Introduction

### History of Clinical Named Entity Recognition

Clinical named entity recognition (NER), an important clinical natural language processing (NLP) task, has been explored for several decades. In the early stage, most NER systems leverage rules and dictionaries to represent linguistic features and domain knowledge to identify clinical entities, such as MedLEE [1], SymText/MPlus [2,3], MetaMap [4], KnowledgeMap [5], cTAKES [6], and HiTEX [7]. To promote the development of machine learning–based system, many publicly available corpora have been developed by organizers of some clinical NLP challenges such as the Informatics for Integrating Biology and the Bedside (i2b2) 2009 [8], 2010 [9-13], 2012 [14-18], 2014 [19-23], ShARe/CLEF eHealth Evaluation Lab 2013 dataset [24], and Semantic Evaluation 2014 task 7 [25], 2015 task 6 [26], 2015 task 14 [27], and 2016 task 12 [28] datasets. Many machine learning–based clinical NER systems have been proposed, and they greatly improved performance compared with the early rule-based systems [13,29,30]. Most systems are implemented based on two types of supervised machine learning

algorithms: (1) classification algorithms such as support vector machines (SVMs) and (2) sequence labeling algorithms such as conditional random fields (CRFs), hidden Markov models (HMMs), and structural support vector machines (SSVMs). Among all of the algorithms, CRFs play the leading roles due to the advantage of the sequence labeling algorithms over classification algorithms in considering context information when making the prediction; CRFs, as one type of discriminative model, tend to achieve better performance for the same source of testing data compared with generative model-based algorithms such as HMMs. Even though CRFs have achieved a huge success in the clinical NER area, they have some obvious limitations: CRF-based systems lie in manually crafted features, which are time consuming, and their ability to capture context in a large window is limited.

## Deep Neural Network–Based Named Entity Recognition Algorithms

In recent years, deep neural network–based NER algorithms have been extensively studied, and many deep learning–based clinical NER systems have been proposed. They have an obvious advantage over traditional machine learning algorithms since they do not require feature engineering, which is the most difficult part of designing machine learning–based systems. They also improve the ability to leverage the context information. Initially, word embedding [31] is proposed as a method to represent the word in a continuous way to better support neural network structure. Then several new neural network structures including recurrent neural networks (RNNs) and long short-term memory (LSTM) [32] have been introduced to better represent sequence-based input and overcome long-term dependency issues. Recently, contextual word representations generated from pretrained bidirectional language models (biLMs) have been shown to significantly improve the performance of state-of-the-art NER systems [33].

In biLMs, the language model (LM) can be described as: given a sequence of N tokens, ($t_1$, $t_2$, ..., $t_N$), the probability of token $t_k$ can be calculated given the history ($t_1$, ..., $t_{k-1}$), and the sequence probability can be computed as seen in Figure 1.

Recent neural LMs usually include one layer of token input, which is represented by word embedding or a CNN over characters, followed by L layers of forward LSTMs. On the top layer, the SoftMax layer is added to generate a prediction score for the next token [33]. The biLM combines two such neural LMs: the forward LM and backward LM; the backward LM is similar to the forward LM, except it runs over the reverse sequence. As a whole, the biLM tries to maximize the log-likelihood of the forward and backward directions as seen in Figure 2.

**Figure 1.** Sequence probability in bidirectional language models.

$$(1) \qquad p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1})$$

**Figure 2.** Log-likelihood of the forward and backward directions language models.

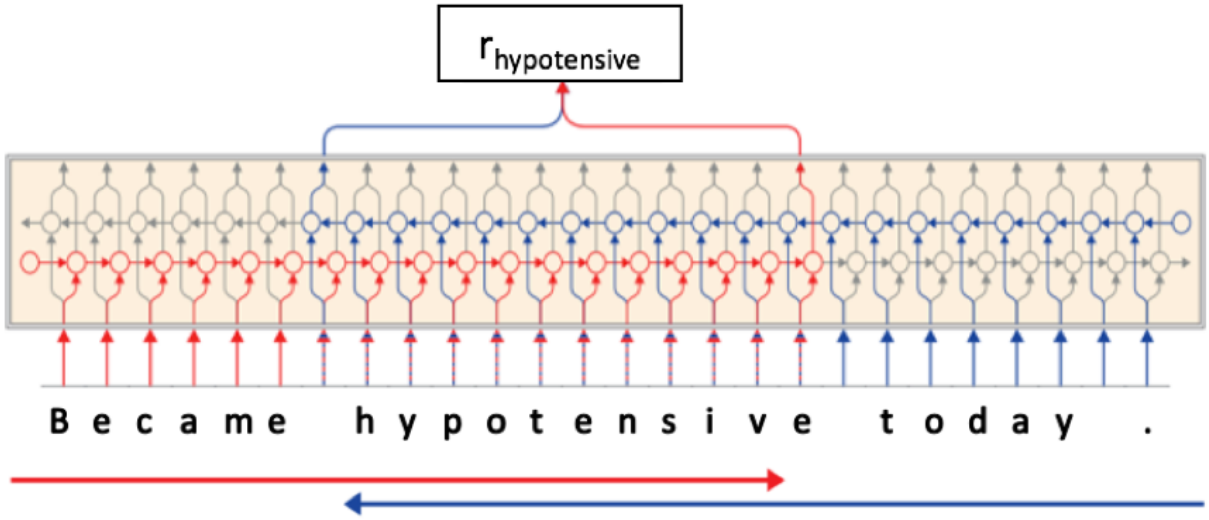
$$(2) \qquad \sum_{k=1}^{N} (\log p(t_k \mid t_1, \ldots, t_{k-1}; \theta_x, \vec{\theta}_{LSTM}, \theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \theta_x, \overleftarrow{\theta}_{LSTM}, \theta_s))$$

Where $\theta_x$ represents the token representation layer, $\theta_s$ represents the Softmax layer, and $\vec{\theta}_{LSTM}$ and $\overleftarrow{\theta}_{LSTM}$ represent the forward and backward directions of the LSTM layer.

In 2017, Peters et al [34] introduced a sequence tagger called TagLM that combines pretrained word embeddings and biLM embeddings as the representation of the word to improve the performance of the NER system. Since the output of each layer of the biLM represents a different type of contextual information [35], Peters et al [33] proposed another embedding, a deep contexualized word representation, ELMo, by concatenating all the biLM layer outputs into the biLM embedding with a weighted average pooling operation. The ELMo embedding adds CNN and highway networks over the character for each token as the input. ELMo has been proven to enhance the performance of different NLP tasks such as semantic role labeling and question answering [33].

Similar to Peters' ELMo, Akbik et al [36] introduced contextual string embeddings for sequence labeling, which leverages neural character-level language modeling to generate a contextualized embedding for each word input within a sentence. The principle of the character-level LM is that it is the same as biLMs except that it runs on the sequences of characters instead of tokens. Figure 3 shows the architecture of extracting a contextual string embedding for the word "hypotensive" in a sentence. We can see that instead of generating a fixed representation of the embedding for each word, the embedding of each token is composed of pretrained character embeddings from surrounding text, meaning the same token has dynamic representation depending on its context.

**Figure 3.** Architecture of extracting a contextual string embedding.




### Deep Neural Network–Based Clinical Named Entity Recognition Systems

In the clinical domain, researchers investigated the performance of clinical NER tasks on various types of deep neural network structures. In 2015, researchers showed it is beneficial to use the large clinical corpus to generate word embeddings for clinical NER systems, and they comparatively investigated the different ways of generating word embeddings in the clinical domain [37]. In 2017, Wu et al [38] produced state-of-the-art results on the i2b2 2010 NER task dataset by employing the LSTM-CRF structure. Liu et al [39] investigated the effects of two types of character word embeddings on LSTM-based systems on multiple i2b2/Veterans Administration (VA) NER task datasets. In 2018, Zhu et al [40] employed a contextualized LM embedding on clinical data and boosted the state-of-the-art performance by 3.4% on the i2b2/VA 2010 NER dataset. The above studies show that, with the development of methods in text representation learning, especially contextual word embedding, more and more hidden knowledge can be learned from a large unannotated clinical corpus, which is beneficial for clinical NER tasks. According to the study by Peters et al [35], contextual word representations derived from pretrained biLMs can learn different levels of information that vary with the depth of the network, from local syntactic information to long-range dependent semantic information. Even without leveraging traditional domain knowledge such as lexicon and ontology, deep learning–based NER systems can achieve better performance than traditional machine learning–based systems.

Besides using pretrained representation from large unlabeled corpora, researchers started to integrate prior knowledge into deep learning frameworks to improve the performance of the NER system. For example, in the general domain, Yu and Dredze [41] created a semantic word embedding based on WordNet and evaluated the performance on language modeling, semantic similarity, and human judgment prediction. In another example, Weston et al [42] leveraged a CNN to generate a semantic embedding based on hashtags to improve the performance of the document recommendation task. In the clinical domain, Wu et al [43] compared two types of methods to inject medical knowledge into deep learning–based clinical NER solutions and found that the RNN-based system combining medical knowledge as embeddings achieved the best performance on the i2b2 2010 dataset. In 2019, Wang et al [44] explored two different architectures that extend the bidirectional LSTM (biLSTM) neural network and five different feature representation schemes to incorporate the medical dictionaries. In addition, other studies also use prior knowledge to generate embeddings [45-49].

To date, no detailed analysis has been published to investigate the value of combining different types of word embeddings and prior knowledge for clinical NER. In this study, we made the following contributions: (1) we proposed an innovative method to combine two types of contextualized embeddings to study their effects on the clinical NLP challenge dataset, (2) we incorporated prior knowledge from semantic resources such as medical lexicon to evaluate if it could further improve the performance of the clinical NER system, and (3) we conducted a thorough evaluation on our models with different sizes of data to gain knowledge on how much data are needed to train a high-performance clinical NER system.

## Methods

### Datasets

For this study, we used two datasets, the 2010 i2b2/VA concept extraction track dataset and the Medical Information Mart for Intensive Care III (MIMIC-III) corpus. The 2010 i2b2/VA challenge dataset is annotated with named entities, while the MIMIC-III corpus is unannotated data.

#### *2010 i2b2/VA Concept Extraction Track Dataset*

The goal of the 2010 i2b2/VA concept extraction task is to identify three types of clinical named entities including problem, treatment, and test from clinical notes. The original dataset includes 349 notes in the training set and 477 notes in the testing set, which include discharge summaries and progress notes from three institutions: Partners HealthCare, Beth Israel Deaconess

Medical Center, and University of Pittsburgh Medical Center. Since the University of Pittsburgh Medical Center's data have been removed from the original data set, the portion of discharge summaries that is available contains 170 notes for training and 256 for testing. In total, the training set contains 16,523 concepts including 7073 problems, 4844 treatments, and 4606 tests. The test set contains 31,161 concepts including 12,592 problems, 9344 treatments, and 9225 tests.

#### *Medical Information Mart for Intensive Care III Corpus*

The MIMIC-III corpus [50] is from MIMIC-III database, which is a large, freely available de-identified health-related dataset that integrates de-identified, comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts.

The dataset comprises 2,083,180 notes from 15 different note types including "rehab services," "case management," "general," "discharge summary," "consult," "radiology," "electrocardiography," "nutrition," "social work," "pharmacy," "echocardiography," "physician," "nursing," "nursing/other," and "respiratory."

### Embedding Generation

In order to fit our text input into the deep neural network structure, we generated three types of embeddings: classic word embeddings, (2) contextualized LM–based word embeddings, and semantic word embeddings.

#### *Training Classic Word Embeddings*

We generated two types of word embeddings based on the MIMIC-III corpus and a medical lexicon: MIMIC-III corpus-based embeddings and tagged MIMIC-III corpus-based embeddings. We adopted the Word2Vec implementation database from Github [51] to train word embeddings based on the MIMIC-III corpus. We used a continuous bag-of-words architecture with negative sampling. In accordance with the results from the study by Xu et al [52], we set the dimension of embedding as 50.

#### *Training Contextual Language Model–Based Embeddings*

Besides the word embeddings, we employed two recently proposed methods to generate contextual LM-based embeddings: ELMo embeddings and (2) contextual string embeddings for sequence labeling (Flair).

### Training ELMo Embeddings

We followed the method introduced by Zhu et al [40] that uses a partial MIMIC-III corpus combined with a certain portion of Wikipedia pages as a training corpus to train the ELMo contextual LM in the clinical domain. In more detail, it combines discharge summaries and radiology reports from the MIMIC-III corpus and all the Wikipedia pages with titles that are items from the Systematized Nomenclature of Medicine–Clinical Terms. Such a corpus is trained on a deep neural network that contains a character-based CNN embedding layer followed by a two-layer biLSTM. Details have been published elsewhere [40].

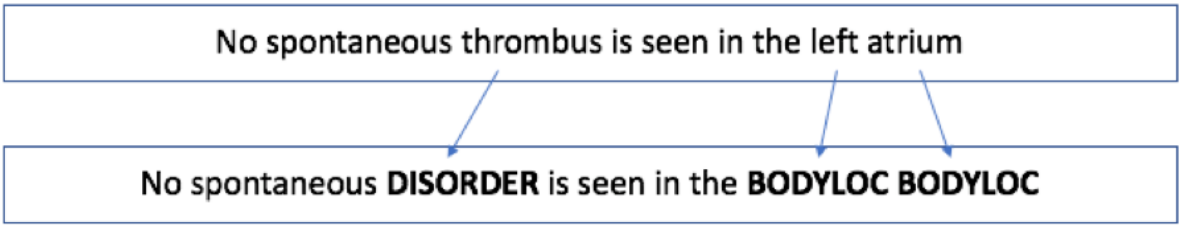
### Training Contextual String Embeddings for Sequence Labeling

Akbik et al [36] proposed a new method to generate a neural character-level LM. The paper shows the state-of-the-art performance on the Conference on Computational Natural Language Learning 2003 NER task dataset. The LM for the general domain is publicly accessible. The author also integrates all the codes into an NLP framework called Flair. It achieved great success on the data in the general domain. However, according to the research by Friedman et al [53], clinical language has unique linguistic characteristics compared with general English, which make models generated from the public domain poorly adaptable to clinical narratives. It is demanding to train the LM on the clinical corpus to better support the clinical NER task. For training corpus preparation, we first did sentence segmentation on the entire corpus, then we randomly selected 1500 sentences as the testing set and another 1500 sentences for the validation set. The remaining part serves as the training set. For the hyperparameters, we kept the default setting: learning rate as 20.0, batch size as 32, anneal factor as 0.25, patience as 10, clip as 0.25, and hidden size as 1024.

### Training Semantic Word Embeddings

Injecting domain knowledge into the deep learning model is a potential way to further improve the performance of the NER system. According to the results by Wu et al [43], combining medical knowledge into the embedding outperforms the method of representing it as a one-hot vector. Therefore, we similarly created the embedding to represent medical lexicon and fed it into the deep learning framework in our study. More specifically, we initially generated a lexicon dictionary based on a subset of semantic categories in the Unified Medical Language System. We then identified all the lexicon occurrences in the corpus using the dictionary and replaced them with semantic categories. Figure 4 shows an example of the conversion. In the example sentence of "No spontaneous thrombus is seen in the left atrium," "thrombus" is replaced with the tag "DISORDER" and "left atrium" is replaced with two "BODYLOC" tags. In this way, we can integrate semantic information into the word embeddings. For the embedding generation, we use the same setting as in the previous section.

**Figure 4.** One example of converting the sentence into the tagged sentence.

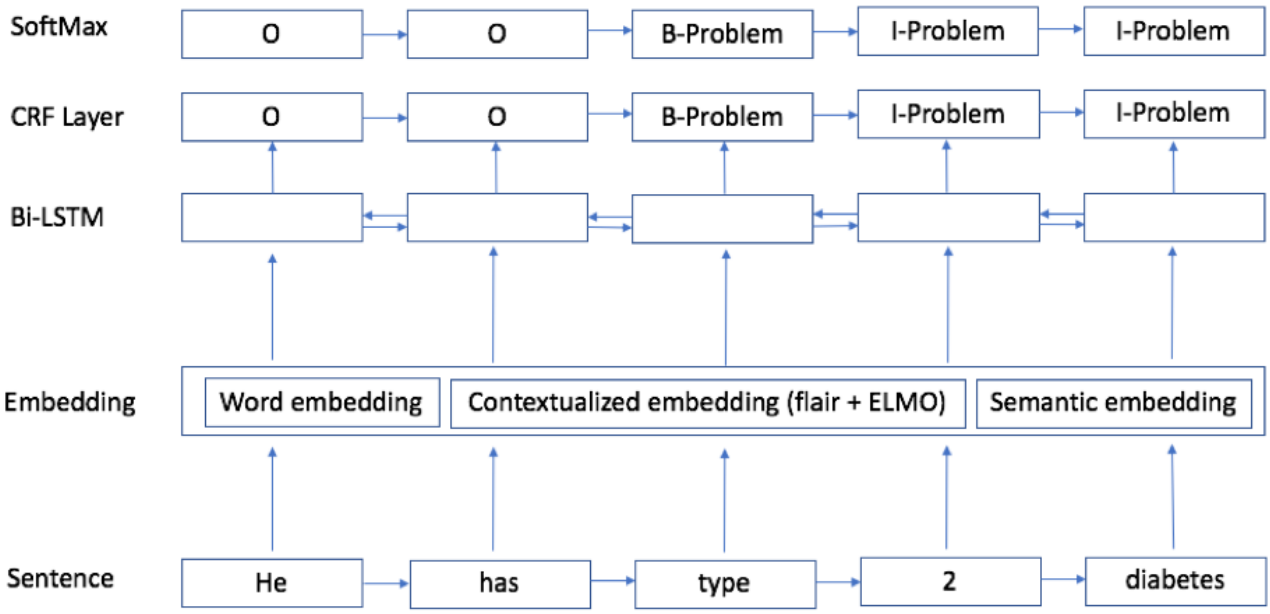



## Deep Neural Network Architecture

After we generated all the embeddings, we started to fit them as the input into our deep neural network for the supervised training stage. Since each type of embedding is generated using one method, meaning each represents different aspects of knowledge from the large corpus, combining them is an obvious solution to potentially further improve the performance, which has also been proven by clinical NER studies [40,43]. Although there are many options to combine multiple embeddings in the deep neural network system such as weighting [54] and ensemble [55], in this study, we adopted the most straightforward way, which is simply concatenating them as the input.

We used the biLSTM-CRF sequence labeling module proposed by Huang et al [56]. Figure 5 shows the architecture of the whole deep neural network structure; the input is the embedding layer, which is concatenated by different types of embeddings as described in the previous section. Before we extracted embeddings for tagged word embedding, we used the same medical lexicon–based tagger to replace the tokens with the semantic tags. All the embedding inputs went through the biLSTM layer to generate forward and backward output, which was used to calculate the probability score by CRF layers. On the top, the prediction was given by a SoftMax layer.

**Figure 5.** Deep neural network structure with combined embeddings. Bi-LSTM: bidirectional long short-term memory; CRF: conditional random field.




## Training the Deep Neural Network–Based Sequence Tagger

For the implementation, we employed Flair [57], which is a simple framework for NLP tasks including NER and text classification. We used the default hyperparameter setting in Flair, and we used the following configuration: learning rate as 0.1, batch size as 32, dropout probability as 0.5, and maximum epoch as 500. The learning rate annealing method is basically the same as the default: we half the learning rate if the training loss does not fall for the consecutive "patience" number of epochs. We set the patience number to 12 in this study. A TITAN V (NVIDIA Corporation) graphics processing unit was used to train the model. We took about 4 hours to train our model each time.

## Evaluation

In order to get more reliable results, we ran each model three times. For the measurement of each running, we used precision, recall, and F-1 score.

## Results

Table 1 shows the performance of the challenge winner system and different deep neural network systems. We used four benchmarks as our baseline systems, and then we reported the performance of the systems when adding ELMo embeddings, Flair embeddings, and tagged embeddings one at a time. All evaluation scores were based on exact matching. For the baseline systems, the first one is the semi-Markov model, developed by Debruijn et al [13], which reported an F-1 score of 85.23%. The second and third baselines are both based on the LSTM model, and they reported F-1 scores of 85.78% and 85.94%, respectively. The last baseline is the best result for the nonensemble models from Zhu et al [40], which used ELMo embedding. The three baseline systems used the original corpus (training: 349 notes; test: 477 notes), all other systems are based on the existing modified corpus (training: 170 notes; test: 256 notes). To start, we combined word embeddings with ELMo and Flair embeddings, respectively. Both models achieved an F-1 score of 87.01%, which is a little bit higher than what was reported by Zhu et al [40]. After combining word embeddings with ELMo and Flair embeddings, the F-1 score increased to 87.30%. When the word embedding on the tagged corpus was incorporated, the performance was further improved to 87.44% for the F-1 score.

In order to test if the improvement between different results is statistically significant, we conducted a statistical test based on results from bootstrapping. From the prediction result of the test set, we randomly selected 1000 sentences with replacement for 100 times and generated 100 bootstrap data sets. For each bootstrap data set, we evaluated F-measures for three pairs of results: (1) "biLSTM + ELMo" and "biLSTM + ELMo + Flair," (2) "biLSTM + ELMo + Flair" and "biLSTM + ELMo + Flair + semantic embedding," and (3) "biLSTM + ELMo by Zhu et al [40]" and "biLSTM + ELMo + Flair + semantic embedding." After that, we adopted a Wilcoxon signed rank test [58] to determine if the differences between F-measures from the three pairs were statistically significant. The results show that the improvement of F-measures for all three pairs were statistically significant (*P* values were .01, .02, and .03, respectively).

**Table 1.** Performance of all the models on the 2010 i2b2/VA dataset.

| Model | F-1 (%) | Precision (%) | Recall (%) |
|---|---|---|---|
| Hidden semi-Markov[a] | 85.23 | 86.88 | 83.64 |
| LSTM[b] by Liu et al [39][a] | 85.78 | —[c] | —[c] |
| LSTM by Wu et al [43][a] | 85.94 | 85.33 | 86.56 |
| BiLSTM[d] + ELMo by Zhu et al [40][a] | 86.84 (0.16) | 87.44 (0.27) | 86.25 (0.26) |
| BiLSTM + Flair | 87.01 (0.18) | 87.54 (0.15) | 86.49 (0.21) |
| BiLSTM + ELMo | 87.01 (0.24) | 87.64 (0.19) | 86.40 (0.30) |
| BiLSTM + ELMo + Flair | 87.30 (0.06) | 87.78 (0.09) | 86.85 (0.07) |
| BiLSTM + ELMo + Flair + semantic embedding | 87.44 (0.07) | 88.03 (0.14) | 86.91 (0.10) |

[a]Model is trained using the complete dataset of i2b2 2010, which contains 349 notes in the training set and 477 notes in the test set.

[b]LSTM: long short-term memory.

[c]Not reported.

[d]BiLSTM: bidirectional LSTM.

## Discussion

### Principal Findings

NER is a fundamental task in the clinical NLP domain. In this study, we investigated the effects of combinations of different types of embeddings on the NER task. We also explored how to use medical lexicon to further improve performance. Based on the result, we found that either ELMo or Flair embeddings could boost the system's performance, and combining both embeddings could further improve the performance. Although both ELMo and Flair embeddings use biLM to train the LM on MIMIC-III corpus, they actually generate the contextualized word embeddings in different ways. ELMo concatenates all the biLM layers to represent all different levels of the knowledge, while Flair embedding is generated by a character-level LM. Character-level LM is different from character-aware LM [59] since it actually uses word-level LM while leveraging character-level features through a CNN encoding step. It was composed by the surrounding text's embedding in the character-level. The difference between ELMo and Flair embeddings could explain the reason why they can play complementary roles in the model.

The results show that adding semantic embeddings could further improve performance. According to the study by Peters et al [35], the lower biLM layer specializes in local syntactic relationships, while the higher layers focus on modeling longer range relationships. Those relationships are learned from the pure clinical corpus without any resources from outside such as medical lexicons and ontologies. This study shows an effective way to incorporate domain knowledge into the deep neural network–based NER system.

A large amount of training data is required to achieve success when applying deep learning algorithms [60]. Within the general domain, it is more difficult to accumulate a large size of the

annotated corpus for most of the clinical NLP tasks since it usually requires the annotator to have in-depth domain knowledge. Contextualized word embeddings, as an effective way of transferring the knowledge from the large unlabeled corpus, could address the issue of lack of training data. According to the results, by only using the small size of the training corpus (170 notes), contextualized word embedding–based models could achieve better performance than the models that use the large size training corpus (349 notes). To further investigate the effectiveness of transfer learning in our proposed models, we compared the performance of our best model generated from different sizes of the training data. Table 2 shows the F-1 score for the model "biLSTM + ELMo + Flair + semantic embedding" on randomly selected 80%, 60%, 40%, 20%, and 10% of the training data. Surprisingly, we found that using only 40% of the training corpus could achieve comparable performance as the original state-of-the-art traditional machine learning–based system. Even using 20% of the training corpus, the model's F-1 score is still more than 80%. This result indicates that contextualized word representation could potentially be an effective way to reduce the size of the training corpus, which could significantly improve the feasibility of applying deep learning to real practice.

Besides the performance reported in the Results section, we also recorded the change of performance for our proposed models during the fine-tuning stage. Table 3 shows the F-1 score on 1, 20, 40, and 60 epochs for our three models. On epoch 1, comparing to only word embeddings, any contextualized word embedding boosts the F-1 score. This is mostly because pretraining on contextualized word embeddings is very beneficial for the task of named entity recognition. This proves that the LM is a good way for pretraining that can be adapted to different downstream NLP tasks. Another interesting finding is that even though the model ELMo achieved the best performance among our three models, it was surpassed by the other two models on later epochs, which indicates that during the optimization process, the best starting point does not necessarily lead to the best local optimal solution.

**Table 2.** Performance of the best model training, BiLSTM[a] + ELMo + Flair + semantic embedding, on different sizes of the training corpus.

| Amount of training data (%) | F-1 (%) | Prec (%) | Rec (%) |
|---|---|---|---|
| 10 | 71.13 | 69.59 | 72.74 |
| 20 | 82.05 | 81.92 | 82.18 |
| 40 | 85.36 | 85.83 | 84.90 |
| 60 | 86.33 | 86.81 | 85.86 |
| 80 | 86.92 | 87.42 | 86.43 |

[a]BiLSTM: bidirectional long short-term memory.

**Table 3.** F-1 score for our proposed models on different epochs.

| Model | 1 epoch (%) | 20 epochs (%) | 40 epochs (%) | 60 epochs (%) |
|---|---|---|---|---|
| Classic word embedding | 61.23 | 75.67 | 78.11 | 79.52 |
| Classic word embedding + ELMo | 76.18 | 85.64 | 85.68 | 86.63 |
| Classic word embedding + ELMo + Flair | 73.28 | 85.33 | 85.97 | 86.96 |
| Classic word embedding + ELMo + Flair + semantic embedding | 74.38 | 85.85 | 86.46 | 87.13 |

## Limitations

This study has some limitations. For contextualized embedding generation, we followed others' research methods and didn't test different configurations for LM training. For example, for ELMo embeddings, we followed the work of Zhu et al [40] for Flair embedding generation and kept the same configuration as seen in the work by Akbik et al [36]. For the fine-tuning stage, we only fine-tuned a limited set of hyperparameters including learning rate and patience. For domain knowledge integration, there are a lot of options that could be explored to merge the lexicon information into the input of the deep neural network structure. In this study, we only tried one way to represent it in the form of word embeddings. In this paper, we studied two contextualized embeddings: ELMo and Flair. In the future, we plan to test our framework by adding bidirectional encoder representations from transformers, which is another popular contextualized embedding [61].

## Conclusions

In this study, we investigated the effects of the combination of two contextualized word embeddings including ELMo and Flair and clinical knowledge for the clinical NER task. Our evaluation on the 2010 i2b2/VA challenge dataset shows that using both ELMo and Flair embeddings outperforms using only ELMo embeddings, which indicates its great potential for the clinical NLP research. Furthermore, we demonstrate that incorporating the medical lexicon into the word representation could further improve the performance. Finally, we found that adopting our best model would be an effective way to reduce the size of the required training corpus for the clinical NER task.

## Acknowledgments

This research was supported by the Advanced Analytics and Data Science organization at Eli Lilly and Company.

## Conflicts of Interest

None declared.

## References

1. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161-174 [FREE Full text] [Medline: 7719797]
2. Christensen L, Haug P, Fiszman M. MPLUS: a probabilistic medical language understanding system. 2002 Presented at: Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain; 2002; Stroudsburg. [doi: 10.3115/1118149.1118154]
3. Koehler SB. SymText: A Natural Language Understanding System for Encoding Free Text Medical Data. Provo: University of Utah; 1999.
4. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
5. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A. The KnowledgeMap project: development of a concept-based medical school curriculum database. AMIA Annu Symp Proc 2003:195-199 [FREE Full text] [Medline: 14728161]
6. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
7. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006 Jul 26;6(1):30. [doi: 10.1186/1472-6947-6-30]
8. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514-518 [FREE Full text] [doi: 10.1136/jamia.2010.003947] [Medline: 20819854]
9. Kim Y, Riloff E, Hurdle JF. A study of concept extraction across different types of clinical notes. AMIA Annu Symp Proc 2015;2015:737-746 [FREE Full text] [Medline: 26958209]
10. 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 [FREE Full text] [doi: 10.1186/1472-6947-13-S1-S1] [Medline: 23566040]
11. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
12. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
13. 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. J Am Med Inform Assoc 2011;18(5):557-562 [FREE Full text] [doi: 10.1136/amiajnl-2011-000150] [Medline: 21565856]
14. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013 Sep 01;20(5):806-813. [doi: 10.1136/amiajnl-2013-001628]
15. Xu Y, Wang Y, Liu T, Tsujii J, Chang EI. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. J Am Med Inform Assoc 2013;20(5):849-858 [FREE Full text] [doi: 10.1136/amiajnl-2012-001607] [Medline: 23467472]
16. Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. A hybrid system for temporal information extraction from clinical text. J Am Med Inform Assoc 2013;20(5):828-835 [FREE Full text] [doi: 10.1136/amiajnl-2013-001635] [Medline: 23571849]
17. Sohn S, Wagholikar KB, Li D, Jonnalagadda SR, Tao C, Komandur Elayavilli R, et al. Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification. J Am Med Inform Assoc 2013 Sep 01;20(5):836-842. [doi: 10.1136/amiajnl-2013-001622]
18. Kovačević A, Dehghan A, Filannino M, Keane JA, Nenadic G. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. J Am Med Inform Assoc 2013 Sep 01;20(5):859-866. [doi: 10.1136/amiajnl-2013-001625]
19. Stubbs A, Kotfila C, Uzuner O. Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform 2015 Dec;58 Suppl:S11-S19 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.007] [Medline: 26225918]
20. Yang H, Garibaldi JM. Automatic detection of protected health information from clinic narratives. J Biomed Inform 2015 Dec;58 Suppl:S30-S38 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.015] [Medline: 26231070]

21. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. J Biomed Inform 2015 Dec;58 Suppl:S47-S52 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.009] [Medline: 26122526]
22. He B, Guan Y, Cheng J, Cen K, Hua W. CRFs based de-identification of medical records. J Biomed Inform 2015 Dec;58 Suppl:S39-S46 [FREE Full text] [doi: 10.1016/j.jbi.2015.08.012] [Medline: 26315662]
23. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Combining knowledge- and data-driven methods for de-identification of clinical narratives. J Biomed Inform 2015 Dec;58 Suppl:S53-S59 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.029] [Medline: 26210359]
24. Suominen H, Salanterä S, Velupillai S, Chapman W, Savova G, Elhadad N. Overview of the ShARe/CLEF eHealth evaluation lab. 2013 Presented at: International Conference of the Cross-Language Evaluation Forum for European Languages; 2013; Valencia. [doi: 10.1007/978-3-642-40802-1_24]
25. Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G. Semeval-2014 task 7: analysis of clinical text. 2014. URL: http://alt.qcri.org/semeval2014/cdrom/pdf/SemEval2014007.pdf [accessed 2019-10-22]
26. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. Semeval-2015 task 6: clinical tempeval. 2015. URL: http://alt.qcri.org/semeval2015/cdrom/pdf/SemEval136.pdf [accessed 2019-10-22]
27. Elhadad N, Pradhan S, Gorman S, Manandhar S, Chapman W, Savova G. SemEval-2015 task 14: analysis of clinical text. 2015. URL: http://alt.qcri.org/semeval2015/cdrom/pdf/SemEval051.pdf [accessed 2019-10-22]
28. Bethard S, Savova G, Chen W, Derczynski L, Pustejovsky J, Verhagen M. SemEval-2016 task 12: clinical TempEval. 2016. URL: http://alt.qcri.org/semeval2016/task12/ [accessed 2019-10-22]
29. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc 2011;18(5):601-606 [FREE Full text] [doi: 10.1136/amiajnl-2011-000163] [Medline: 21508414]
30. Zhang Y, Wang X, Hou Z, Li J. Clinical named entity recognition from Chinese electronic health records via machine learning methods. JMIR Med Inform 2018 Dec 17;6(4):e50 [FREE Full text] [doi: 10.2196/medinform.9965] [Medline: 30559093]
31. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Info Process Sys 2013.
32. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735]
33. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K. Deep contextualized word representations. arXiv preprint 2018:180205365. [doi: 10.18653/v1/n18-1202]
34. Peters M, Ammar W, Bhagavatula C, Power R. Semi-supervised sequence tagging with bidirectional language models. arXiv preprint 2017:170500108. [doi: 10.18653/v1/p17-1161]
35. Peters M, Neumann M, Zettlemoyer L. Dissecting contextual word embeddings: architecture and representation. arXiv preprint 2018. [doi: 10.18653/v1/d18-1179]
36. Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling. 2018. URL: https://alanakbik.github.io/papers/coling2018.pdf [accessed 2019-10-22]
37. Wu Y, Xu J, Jiang M, Zhang Y, Xu H. A study of neural word embeddings for named entity recognition in clinical text. AMIA Annu Symp Proc 2015;2015:1326-1333 [FREE Full text] [Medline: 26958273]
38. Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical named entity recognition using deep learning models. AMIA Annu Symp Proc 2017;2017:1812-1819 [FREE Full text] [Medline: 29854252]
39. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Mak 2017 Jul 05;17(Suppl 2):67 [FREE Full text] [doi: 10.1186/s12911-017-0468-7] [Medline: 28699566]
40. Zhu H, Paschalidis I, Tahmasebi A. arXiv preprint. 2018. Clinical concept extraction with contextual word embedding. URL: https://arxiv.org/abs/1810.10566 [accessed 2019-10-22]
41. Yu M, Dredze M. Improving lexical embeddings with semantic knowledge. 2014 Presented at: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2); 2014; Baltimore. [doi: 10.3115/v1/p14-2089]
42. Weston J, Chopra S, Adams K. Semantic embeddings from hashtags. 2014 Presented at: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014; Doha. [doi: 10.3115/v1/d14-1194]
43. Wu Y, Yang X, Bian J, Guo Y, Xu H, Hogan W. Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition. AMIA Annu Symp Proc 2018;2018:1110-1117 [FREE Full text] [Medline: 30815153]
44. Wang Q, Zhou Y, Ruan T, Gao D, Xia Y, He P. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. J Biomed Inform 2019 Apr;92:103133. [doi: 10.1016/j.jbi.2019.103133] [Medline: 30818005]
45. Liu Q, Jiang H, Wei S, Ling Z, Hu Y. Learning semantic word embeddings based on ordinal knowledge constraints. 2015 Presented at: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1); 2015; Beijing. [doi: 10.3115/v1/p15-1145]

46. Mencia EL, de Melo G, Nam J. Medical concept embeddings via labeled background corpora. 2016 Presented at: Proceedings of the 10th Language Resources and Evaluation Conference (LREC ); 2016; Portoroz.
47. Peng S, You R, Wang H, Zhai C, Mamitsuka H, Zhu S. DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. Bioinformatics 2016 Jun 15;32(12):i70-i79 [FREE Full text] [doi: 10.1093/bioinformatics/btw294] [Medline: 27307646]
48. Celikyilmaz A, Hakkani-Tur D, Pasupat P, Sarikaya R. Enriching word embeddings using knowledge graph for semantic tagging in conversational dialog systems. URL: https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/download/10333/10034 [accessed 2019-10-22]
49. Passos A, Kumar V, McCallum A. Lexicon infused phrase embeddings for named entity resolution. arXiv preprint 2014:14045367. [doi: 10.3115/v1/w14-1609]
50. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
51. Word2Vec implementation. URL: https://github.com/dav/word2vec [accessed 2019-10-22]
52. Xu J, Zhang Y, Wang J, Wu Y, Jiang M, Soysal E. UTH-CCB: the participation of the SemEval 2015 challenge—Task 14. URL: https://clamp.uth.edu/challenges-publications/UTH-CCB-%20the%20participation%20of%20the%20SemEval%202015%20challenge%E2%80%93Task%2014.pdf [accessed 2019-10-22]
53. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002 Aug;35(4):222-235 [FREE Full text] [Medline: 12755517]
54. Reimers N, Gurevych I. Alternative weighting schemes for ELMo embeddings. arXiv preprint 2019:190402954.
55. Speer R, Chin J. An ensemble method to produce high-quality word embeddings. arXiv preprint 2016:160401692.
56. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint 2015:150801991.
57. Akbik A. Flair implementation. URL: https://github.com/zalandoresearch/flair/graphs/contributors2018 [accessed 2019-10-22]
58. Woolson R. Wilcoxon signed-rank test. Wiley Encyclopedia of Clinical Trials 2007:1-3. [doi: 10.1002/9780471462422.eoct979]
59. Kim Y, Jernite Y, Sontag D, Rush AM. Character-aware neural language models. URL: https://arxiv.org/abs/1508.06615 [accessed 2019-10-22]
60. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 27;521(7553):436-444. [doi: 10.1038/nature14539]
61. Devlin J, Chang M, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint 2018:181004805.


## Abbreviations

**biLM:** bidirectional language model
**biLSTM:** bidirectional long short-term memory
**CNN:** convolutional neural network
**CRF:** conditional random field
**HMM:** hidden Markov model
**i2b2:** Informatics for Integrating Biology and the Bedside
**LM:** language model
**LSTM:** long short-term memory
**MIMIC-III:** Medical Information Mart for Intensive Care III
**NER:** named entity recognition
**NLP:** natural language processing
**RNN:** recurrent neural network
**SSVM:** structural support vector machine
**SVM:** support vector machine
**VA:** Veterans Affairs


*Edited by G Eysenbach; submitted 28.05.19; peer-reviewed by F Li, B Polepalli Ramesh; comments to author 18.06.19; revised version received 16.07.19; accepted 19.10.19; published 13.11.19*

*Please cite as:*
*Jiang M, Sanger T, Liu X*
*Combining Contextualized Embeddings and Prior Knowledge for Clinical Named Entity Recognition: Evaluation Study*
*JMIR Med Inform 2019;7(4):e14850*
*URL: http://medinform.jmir.org/2019/4/e14850/*
*doi: 10.2196/14850*
*PMID: 31719024*