A New Modeling Method Base on Candidate Window for Clinical Concept Extraction

Yongtao Tang (✉ tyt941016@163.com)

Nacional'nyj issledovatel'skij universitet Moskovskij institut elektronnoj tehniki Fakul'tet Intellektual'nyh tehniceskikh sistem <https://orcid.org/0000-0002-7378-8949>

Jie Yu

National University of Defense Technology

Shasha Li

National University of Defense Technology

Bin Ji

National university of Defense Technology

Yusong Tan

National University of Defense Technology

Qingbo Wu

National University of Defense Technology

Technical advance

Keywords: Clinical Concept Extraction, Natural Language Processing, Slide Window, Convolutional Neural Network

Posted Date: June 23rd, 2020

DOI: <https://doi.org/10.21203/rs.2.22462/v2>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

TECHNICAL ADVANCE ARTICLE

A New Modeling Method Base on Candidate Window for Clinical Concept Extraction

Yongtao Tang[†], Jie Yu[†], Shasha Li^{*}, Bin Ji, Yusong Tan and Qingbo Wu

^{*}Correspondence:

lishasha198211@163.com

College of Computer, National
University of Defense Technology,
Changsha, China

Full list of author information is
available at the end of the article

[†]These authors contributed
equally to this work.

Abstract

Background: Recently, how to structuralize electronic medical records (EMRs) has attracted considerable attention from researchers. Extracting clinical concepts from EMRs is a critical part of EMR structuralization. The performance of clinical concept extraction will directly affect the performance of the downstream tasks related to EMR structuralization. We propose a new modeling method based on candidate window classification, which is different from mainstream sequence labeling models, to improves the performance of clinical concept extraction tasks under strict standards by considering the overall semantics of the token sequence instead of the semantics of each token. We call this model as slide window model.

Method: In this paper, we comprehensively study the performance of the slide window model in clinical concept extraction tasks. We model the clinical concept extraction task as the task of classifying each candidate window, which was extracted by the slide window. The proposed model mainly consists of four parts. First, the pre-trained language model is used to generate the context-sensitive token representation. Second, a convolutional neural network (CNN) is used to generate all representation vector of the candidate windows in the sentence. Third, every candidate window is classified by a Softmax classifier to obtain concept type probability distribution. Finally, the knapsack algorithm is used as a post-process to maximize the sum of disjoint clinical concepts scores and filter the clinical concepts.

Results: Experiments show that the slide window model achieves the best micro-average F1 score(81.22%) on the corpora of the 2012 i2b2 NLP challenges and achieves 89.25% F1 score on the 2010 i2b2 NLP challenges under the strict standard. Furthermore, the performance of our approach is always better than the BiLSTM-CRF model and softmax classifier with the same pre-trained language model.

Conclusions: The slide window model shows a new modeling method for solving clinical concept extraction tasks. It models clinical concept extraction as a problem for classifying candidate windows and extracts clinical concepts by considering the semantics of the entire candidate window. Experiments show that this method of considering the overall semantics of the candidate window can improve the performance of clinical concept extraction tasks under strict standards.

Keywords: Clinical Concept Extraction; Natural Language Processing; Slide Window; Convolutional Neural Network

Background

Electronic medical records (EMRs) contains substantial clinical data of patients during hospitalization, such as symptoms, test, and treatment, which is the largest

source of empirical data in medical research. Most of this information exists in the form of natural language, and this unstructured data is hard to be directly used. With the rapid development of EMR systems, how to effectively extract structuralized information from EMRs wrote by natural language has become a research hot-spot.

Clinical concept extraction from free-text is a crucial part of EMRs structuralization. The clinical concept includes both name entity such as problems, tests, treatment, and clinical departments, and events relevant to the patient’s clinical timeline, such as admission, transfer between the department, etc. [1] Clinical concept extraction is a precursor to downstream tasks such as relation extraction [2], and co-reference [3]. Multiple clinical corpora for different tasks such as 2010 i2b2, 2012 i2b2 NLP challenges [1,4], contain the clinical concept extraction task as sub-task.

The mainstream solution for clinical concept extraction is modeling the clinical concept extraction task as a sequence labeling problem [5]. A common practice is employing a specific set of labels (such as ‘BIO’ and ‘BEMO’ , etc.) to label each token in the sentence separately, and then extract the corresponding clinical concepts from the sentence according to the results of the labeling. A ‘BIO’ labeling case study is shown below:

	The	patient	coughs	with	fever	.
Correct:	O	O	B-PROBLEM	I-PROBLEM	I-PROBLEM	O
Incorrect:	O	O	B-PROBLEM	O	B-PROBLEM	O

Where the ‘B-PROBLEM’ tag represents the starting token of a ‘PROBLEM’ entity, the ‘I-PROBLEM’ tag represents a successor token of the ‘PROBLEM’ entity, and the ‘O’ represents a non-entity token. One problem which exists with this modeling method is that the wrong prediction of any token label will lead to the wrong extraction of the entire clinical concept. For example, in the above case, ‘coughs with fever’ should be extracted as a complete clinical concept. However, if the label of ‘with’ is incorrectly predicted as ‘O’, the clinical concepts extracted from the labels will become two clinical concepts of ‘coughs’ and ‘fever’. Please note that the label of ‘fever’ has been changed to ‘B-PROBLEM’ in consideration of the legality of the label.

We call this type of extraction error a boundary segmentation error. The impact of such errors on the performance of the clinical concept extraction model can be estimated by observing the performance difference of the model under two different evaluation criteria. Under the relax standard that only requires the intersection of the extracted clinical concept and the correct clinical concept, the concept of such an error is still considered correct. However, the concepts with such error will be regarded as wrong concepts under the strict standard that require extracted clinical concepts and correct clinical concepts split boundary is completely consistent. **Table 1** shows the performance of the two sequence labeling models on the 2010 i2b2 and 2012 i2b2 tasks.

Where ‘BERT_{base}+Softmax’ and ‘BERT_{base}+BiLSTM-CRF’ respectively represent the softmax classifier [6] and the BiLSTM-CRF model [7] which use Bidirectional Encoder Representation from Transformers (BERT) [6] of base size as

Table 1 The performances of two systems on the 2010 i2b2 and 2012 i2b2 tasks(%)

Method	2010 i2b2		2012 i2b2	
	Strict	Relax	Strict	Relax
BERT _{base} +Softmax	84.70	92.55	76.85	85.51
BERT _{base} +BiLSTM-CRF	84.94	92.62	76.59	85.15

embedding layer. As can be seen from **Table 1**, boundary segmentation errors caused about 7.7% and 8.6% performance degradation in the 2010 i2b2 and 2012 i2b2 tasks, respectively, under the strict standard. It can be seen that the error of boundary segmentation greatly affects the performance of the clinical concept extraction model under strict standards.

Through in-depth analysis of such problems, we found that the main reason why some tokens are difficult to accurately predict is their uneven probability of appearing as different labels in the dataset. The probability of such tokens appearing as a non-clinical concept component in the dataset is much higher than their probability as a clinical concept component. When the model predicts such tokens, it tends to predict such tokens as a non-clinical concept component, which leads to incorrect segmentation of clinical concept boundaries. One of the feature of this kind of token is that its own semantics are not much related to the clinical concepts that need to be extracted. It only appears as a grammatical structure of complex clinical concepts, such as modifiers, conjunctions, and symbols. Take conjunction as an example, in the training set of the 2012 i2b2 task, 'with' appeared 1392 times in total, of which only 17 times appeared as part of the 'problem' type entity (labeled I-PROBLEM or B-PROBLEM), which accounted for about the total number of samples 1.22%. The uneven distribution of labels leads a sequence labeling model hard to correctly predict the label of 'with' as 'I-PROBLEM' or 'B-PROBLEM'. Similar ones include 'and', 'of', '/' and so on. **Figure 1** shows the distribution of the label of these tokens in the 2012 i2b2 task training set.

[width=12cm]Figure 1.pdf

Figure 1 The labels distribution of tokens in the 2012 i2b2 task training set

In order to improve the performance of clinical concept extraction tasks under strict standards, we propose a modeling method different from the sequence labeling problem. The clinical concept extraction task is modeled as a problem that classifies all possible token sequences. The token sequence to be classified is called a candidate window. For example, 'The patient', 'coughs', and 'coughs with fever' are candidate windows.

We call this model the slide window model (SW). The slide window model classifies the candidate windows instead of predicting the labels of each token that composes the clinical concept, which avoids the boundary segmentation errors caused by the wrong prediction of individual token labels and improves the performance of clinical concept extraction tasks under the strict standard. Our main contributions include:

- We modeled the clinical concept extraction task as a candidate window classification problem for the first time.

- We proposed a slide window model that suggest possible segmentations first and then measure how possible a segmentation to be extracted with features extracted by CNN.
- We achieved state-of-the-art performance on the 2012 i2b2 task.

Clinical concept extraction

Clinical concept extraction aims to extract clinical concepts (e.g., problem, test, and treatment) from EMRs. The solution to the clinical concept extraction task can be roughly divided into two categories. The first category consists of feature-based approaches, which extract clinical concept by machine learning-based model (e.g., HMM, CRF) using manually designed clinical domain knowledge as features [8, 9]. The other category mainly contains neural network-based approaches, which generally model clinical concept extraction problems as sence labeling problems and clinical concepts are extracted by predicting the label for each token. Up to now, the most mainstream model for clinical concept extraction tasks is a bidirectional Long Short-Term Memory with Conditional Random Field (BiLSTM-CRF) model [10–12]. The BiLSTM-CRF model captures forward and backward context-sensitive information in the sentence by the BiLSTM layer and considers the correlations of the label sequence by the CRF layer using the Viterbi algorithm. As pre-trained language models make significant advances in the NLP field [6, 13], more researchers applied pre-trained context-sensitive word embeddings to medical NLP. For instance, Elmo has shown excellent performance in clinical concept extraction [14]. Some recent studies have attempted to pre-train medical-specific language models using unlabeled medical field texts [15–18] to improve the performance of the BiLSTM-CRF model in the medical NLP task. These studies do demonstrate the enormous potential of domain-based pre-training in medical NLP tasks.

Method

We solve the clinical concept extraction problem in two-stages. The first stage is called candidate concept generation, in which we classify candidate windows into different concept types and calculate the classification confidence using features of candidate windows extracted by CNN. If the confidence of a candidate window exceeds the threshold, it is added to the list of candidate concepts. Here, a candidate window means a sequence of tokens in a sentence within the maximum length. The second stage is called the concept score maximization. Considering that in the candidate concept set generated in the first stage, there may be overlap between the two candidate concepts, we need to post-process the candidate concept set to obtain the final extraction result of the clinical concepts without overlap. In this paper, we use the knapsack algorithm as a post-processing method, which considers classification confidence of each candidate concept to generate its clinical concept score and optimize the final clinical concept set in a sentence by maximizing the sum of the clinical concept score of the sentence. The details of each stage are described as follows.

Candidate concept Generation

The overview architecture of candidate concept generation is shown in **Figure 2**, which primarily involves the following three components: i) Context-sensitive token

representation. It generates a semantic representation of each token of a sentence using bidirectional encoder representation from Transformers (BERT). ii) CNN-based representation of candidate window. It convolves all token representations in a candidate window to obtain a feature representation of each candidate window. iii) Candidate window classification. It predicts the concept type of each candidate window and measures the classification confidences by their feature representations. Details are shown in the following sections.

[width=12cm]Figure 2.pdf

Figure 2 The overview architecture of the candidate concept generation

Context-sensitive token representation

In natural language, the semantic of a token varies with its context. Recent studies show that pre-trained language models that consider contextual information can more accurately represent the semantic information of each token. In this paper, we used bidirectional encoder representation from Transformers (BERT) to generate context-sensitive embedding of each token, and adopted a labeling strategy consistent with the vanilla BERT model [6]. Inspired by the Elmo model, we spliced the context-sensitive embedding (blue rectangles in **Figure 2**) from the last layer of the BERT model and the token embedding (green rectangles in **Figure 2**) from the embedding layer of BERT model as input to the downstream task. This splicing method retains the semantic information of tokens at different levels, which can provide more information for downstream tasks.

In order to compare with the SOTA results on the 2010 i2b2 and 2012 i2b2 tasks, we followed Si 2019 et al. [18] and used the MIMIC III database to pre-train the BERT model on the domain. The parameters used in training are set according to the authors' detailed instructions. MIMIC III [19] is a public database of intensive care unit (ICU) patients.

CNN-based representation of candidate window

In this component, we apply a convolutional neural network(CNN) to extract the features of each candidate window. Considering that If the length of the candidate window is not limited, the number of candidate windows is too large, which is $O(N^2)$, where N is the number of tokens in the sentence. We set a maximum window length and select all token sequences whose length is less than the maximum window length as candidate windows. This not only reduces the number of candidate windows but also ensures that the model does not miss any clinical concepts within the maximum window length. The setting of the maximum window length depends on the statistical results of the concept length in the data set.

Using CNN to extract the features of each candidate window, one straightforward method is to define K CNNs to perform feature extraction on different candidate windows, where K is the total number of candidate windows in the sentence. Generally, the number of K is large, about $O(NM)$. Where M is the maximum window length, this will cause a problem, because the number of candidate windows is extensive, defining a CNN for each candidate window separately will cause unacceptable

space overhead. Another direct idea is to use a two-stage model. After obtaining a context-based token representation, use CNN to extract features for each candidate window once. There is no doubt that this method is very time-consuming because this method requires K convolution operations for each sentence. In this paper, we propose a new method that extracts the feature vectors of all candidate windows in the sentence using CNN only once.

[width=12cm]Figure 3.pdf

Figure 3 Perform 1D-convolution operation of length 2 on the candidate window

Figure 3 shows the process of convolving the candidate window "coughs with fever" using a 1-dimensional convolution operator of length 2. Generally, we only use the final result (represented by the orange region) generated by CNN as the feature vector of the candidate window, without paying attention to the output of the lower convolution layer (represented by the yellow region). However, by observing **Figure 3**, it can be found that the feature vector generated by the lower convolution layer contains the feature information of the shorter candidate window. For example, the feature vector represented by the yellow region in **Figure 3** can be considered as the feature representation of the candidate window "coughs with".

Based on these observations, we use feature vectors obtained from different convolutional layers to represent the feature information of candidate windows of different lengths and achieve the purpose of extracting all candidate window feature vectors, using CNN only once.

As shown in **Figure 4**, based on the context-based token representation, we use a multi-layer convolutional neural network to obtain a vector representation of candidate windows whose length does not exceed the maximum window length. Each convolutional layer corresponds to a candidate window of a specific length. The context-sensitive token representation act as the feature representation of all candidate windows with length 1. In the convolutional layer, we convolve the input with a 1D convolution operates of length 2. After passing through a convolutional layer, the perception field will be expanded by 1 length. The outputs of the first convolutional layer act as the representations of all candidate windows with length 2, the outputs of the second convolutional layer act as the representations of all candidate windows with length 3, and so on. Considering that during the convolution process, vector representations of different layers may have different dimensions, we project these vector representations into a vector space of the same dimension to serve as the final feature representation of each candidate window. By collecting the output of different convolutional layers, we obtain a vector representation of all candidate windows whose length does not exceed the maximum window length.

[width=12cm]Figure 4.pdf

Figure 4 Perform 1D-convolution operation of length 2 on the candidate window

Candidate windows classification

In this component, a candidate concept list can be generated. First, we classify each candidate window as a type of clinical concept (such as PROBLEM, TEST or TREATMENT) or “non-clinical concept” using a Softmax classifier. Then, we add all candidate windows, which are predicted as a clinical concept, and the confidence exceeds the threshold, into the candidate concept list. We set the threshold manually based on experience. In this paper, we set the threshold to 0.5 simply.

We use a one-hot vector to represent the type of candidate window and cross-entropy loss to measure the correctness of the classification. The window is classified as a clinical concept is called a positive case (such as ‘coughs with fever’ in **Figure 4**), and the window that is non-clinical concept is called a negative case (such as ‘with’, ‘coughs with’ and ‘with fever’ in **Figure 4**). We noticed that the positive case is very sparse in all candidate windows. For example, in the i2b2 data set, each sentence contains no more than ten positive cases. However, the number of candidate windows it generates is much larger than the number of positive cases. For example, in a slide window model with a maximum sequence length of 128 and a maximum candidate window of 10, the number of candidate windows will be as high as 1,650. To solve the problem of data imbalance, we introduce a weighting factor in the loss function to increase the weight of the positive case. As shown in **Equation 1**.

$$L = \alpha \times l_p + (1 - \alpha) \times l_n \quad (1)$$

Where l_p is the mean of the loss of all positive examples, while l_n corresponds to the mean of loss of all negative examples, and α is the weighting factor to increase the weight of the positive case in the loss function, L represents the weighted loss.

concept score maximization

Since the generated candidate concepts may collide, such as boundaries coinciding with two candidate concepts, we need to select a set of candidate concepts without conflicts from the candidate concept list. To measure the quality of different candidate concepts, we define a score for each candidate concept, and our goal is to find a set of candidate concepts with the most significant concept score. We call this process concept score maximization.

In this paper, we take the $\text{logit} = \log(\frac{p}{1-p})$ generated by the softmax classifier as the concept score, where p represents the classification confidence corresponding to the candidate concept. And then, we use the knapsack algorithm to find the sequence of concepts with the highest score. **Equation 2** shows the state transition equation of the knapsack algorithm.

$$S(n+1) = \max_{i=0}^n [S(i) + F(i+1, n+1)] \quad (2)$$

Where $S(n)$ is the most significant sum of concept score on the first n token sequences, and $F(i, j)$ is the concept score of the candidate concept starting with the i th token and ending with the j th token. If the token sequence is non-candidate concept, the concept score is 0. Each time $S(n)$ is updated, we record the value of i as $Pre(n)$, which is used to get the last concept $E(i+1, n)$ and the previous point. The specific algorithm is shown in **Algorithm 1**.

Algorithm 1 Concept Score Maximization Based Knapsack Algorithm

Input: The concept score $F(i, j)$, Sentence length N

Output: A set of the selected clinical concept *ConceptSet*

```

1: for  $i = 0 \rightarrow N$  do
2:    $S[i] = 0$ 
3: for  $i = 1 \rightarrow N$  do
4:   for  $j = 0 \rightarrow i - 1$  do
5:     if  $S(i) \leq S(j) + F(j + 1, i)$  then
6:        $S(i) = S(j) + F(j + 1, i)$ 
7:        $Pre(i) = j$ 
8:  $t = N$ 
9: while  $t! = 0$  do
10:  if  $CandidateWindow(Pre(t), t)$  is a concept then
11:    add  $CandidateWindow(Pre(t), t)$  into ConceptSet
12:     $t = Pre(t)$ 
13: return ConceptSet

```

Result

Dataset

Our experiments are performed on two widely studied public available datasets, 2010 i2b2, and 2012 i2b2. The 2010 i2b2 challenge data contains a total of 170 training and 256 testings EMRs with three clinical concept types: PROBLEM, TEST, and TREATMENT. The 2012 i2b2 challenge data contains 190 training and 120 testing discharge summaries, with six clinical concept types: PROBLEM, TEST, TREATMENT, EVIDENTIAL, CLINICAL DEPARTMENT, and OCCURRENCE. There are two different evaluation criteria on These two datasets: relax standard and strict standard. Under the relax standard, if the extracted clinical concept overlaps with the correct clinical concept and has the same type, the extracted clinical concept is considered to be correct. However, under strict standards, only when the extracted clinical concept is exactly the same as the correct clinical concept (including the boundary and the type), it is considered to be correct. The pre-training database is MIMIC III [19], which is a public database and consists of almost 2million clinical notes.

Baseline

In this paper, we conducted two experiments. In order to illustrate the performance improvement of the proposed method compared with the sequence labeling-based method under strict standards in the clinical concept extraction task, we have selected two currently mainstream sequence labeling methods for comparison in the first experiment. One is the BiLSTM-CRF model, which is a widely used sequence labeling model. The structure of the model follows Lample 2016 et al. [7] The other is a method proposed by Google in 2018 that only uses the softmax classifier to classify the token representation vectors generated by the pre-trained language model. In **Table 2**, we use ‘BiLSTM-CRF’, ‘Softmax’, and ‘SW’ to represent the BiLSTM-CRF model, softmax classifier, and slide window model, respectively.

On the other hand, we compare the performance between the proposed method and the current SOTA results in the second experiment. The SOTA results in 2012 i2b2, and 2010 i2b2 were obtained by si 2019 et al. Their work used the MIMIC III database to train pre-trained language models further. For comparison, we have followed their work and used the MIMIC III database to pre-train domain models.

In **Table 3**, we use ‘SOTA’ and ‘SW’ to represent the SOTA result and our results, respectively.

Pre-training

The medical-specific model is initialized with base-sized BERT ($BERT_{base}$) and pre-trained using the MIMIC III database [19]. The model is represented by $BERT_{base-mimic}$. Unless specified, we follow the original detailed instruction, which Google proposed, to set up the pre-training parameters. The vocabulary list consisting of 30522 word-pieced tokens applied in $BERT_{base}$ is adopted, and all words are set to lowercase, as is standard practice. We performed 700000 pre-training steps and took the latest preserved model as the $BERT_{base-mimic}$. These settings are the same as Si 2019 et al. [18]. The maximum sequence length of the BERT model is set to 128. For all sentences that exceed this length, we use the default tokenize method of BERT to truncate it.

Experiment

In the first experiment, a base-sized BERT model was used in these three models, the BiLSTM-CRF model, softmax classifier, and slide window model, to generate a context-based token vector representation in order to reduce the performance difference caused by the different token embedding vectors. On the other hand, in order to limit the difference in vector expression ability caused by the different representation vector length, the length of the vector representation output by our three models is set to 768, which is consistent with the vector representation length output by $BERT_{base}$. The implementation of the BiLSTM-CRF model and softmax classifier follows Lample 2016 et al. [7] and Google 2018 [6], respectively. As for the hyperparameter settings of the slide window model, we set the maximum window length to 10 based on the statistical results on the 2010 i2b2 and 2012 i2b2 task training sets. This window length can cover more than 98.8% of clinical concepts without generating too many candidate windows. The smaller maximum window length will lose performance because it will lose overmany clinical concepts. And the larger maximum window length will consume substantial computing resources. **Figure 5** shows the number and proportion of clinical concepts that can be covered by different maximum window lengths in the 2012 i2b2 task. Meanwhile, we set the weighting factor in the loss function to 0.6 and the threshold for generating a candidate concept to 0.5 based on our experience. The convolutional neural network uses 1-dimensional convolution operators of length 2 in the slide window model. For the activation function of the convolution operator, we tried Sigmoid and Relu, and achieve similar performance.

[width=12cm]Figure 5.pdf

Figure 5 Clinical concept coverage of different maximum window lengths on 2012 i2b2 tasks

In the second experiment, we used the results reported by Si 2019 et al. [18] as the baseline, which is the state-of-the-art on the 2010 i2b2 and 2012 i2b2 corpora, so far. The baseline system used the external database MIMIC III to further pre-train the BERT model. After that, the BiLSTM-CRF model was used to label the

token sequence of the sentence to extract the clinical concept. For comparison, we followed the work of Si 2019 et al., And further pre-trained the BERT_{base} model using the MIMIC III database to obtain the BERT_{base-mimic} model. In this experiment, we use the BERT_{base-mimic} model instead of the BERT_{base} model to generate the context-based token vector representation in the slide window model. The settings of other hyperparameters in the slide window model are the same as the first experiment. Because pre-training the language model using external resources is not the focus of this paper, our pre-training process follows si 2019 et al. Without making more detailed adjustments to the hyperparameters. The final BERT_{base-mimic} model performance is slightly worse than that used by the baseline system Language model. The reproduction result of the SOTA method using the BERT_{base-mimic} model is shown in **Table 3** and represented by BiLSTM-CRF.

Evaluation

We evaluated the overall performance of the model under strict standards. The specific performance metrics are precision, recall, and F1 score, and the F1 score is used as the final evaluation criteria.

Training data is divided into ten times cross-validation. We train the model using the training set and evaluate the effect of the model training using the development dataset. Finally, the performance of the model on the test dataset is reported.

The checkpoint of the model is saved every 1000 steps during training. The one with the best performance from the last five saved checkpoint is selected as the final result.

Performance comparison

Table 2 shows the performance of different models under two clinical concept extraction tasks. The performance is evaluated in the F1 score. In general, the slide window model performs better than the BiLSTM-CRF model and the Softmax model under the strict standard when the same pre-training language model is used.

Table 2 Comparison of the performances of deifferent models on the two tasks(%)

Method	2010 i2b2	2012 i2b2
Softmax	84.70	76.85
BiLSTM-CRF	84.94	76.59
SW	85.83	78.45

Table 3 shows the performance comparison of our proposed slide window model with the SOTA results in 2012 i2b2 and 2012 i2b2. For the 2012 i2b2 corpus, the best performance under the strict standard is achieved by the slide window model with an F1 score of 81.22%. It improves the performance by 0.88% over the previous SOTA result achieved by Si 2019 et al. with an F1 score of 80.34%. On the other hand, the performance of our model in 2010 i2b2 is slightly worse than the current SOTA results, with F1 scores differing by 0.3%. However, compared to the results of our reproduction of SOTA, we still improved the F1 score by 1.07%.

In order to further analyze the impact of different categories of clinical concepts on the performance of the slide window model, we separately calculated the performance of the slide window model on different types of clinical concepts in the 2012

Table 3 Comparing the performances of proposed model with SOTA on the two tasks(%)

Method	2010 i2b2	2012 i2b2
SOTA	89.55	80.34
BiLSTM-CRF	88.18	79.32
SW	89.25	81.22

i2b2 task. The experimental setup was the same as the first experiment, and the F1 score was used as the final evaluation criterion. The experimental results are shown in **Table 4**.

Table 4 Performance of each type of clinical concept on 2012 i2b2 task.

	Softmax	BiLSTM+CRF	SW
OCCURRENCE	60.31	60.72	64.20
CLINICAL_DEPT	79.43	79.17	79.48
TEST	82.25	81.32	83.47
PROBLEM	82.10	81.15	82.16
EVIDENTIAL	69.94	71.78	72.88
TREATMENT	80.38	78.76	81.20
TOTAL	76.85	76.59	78.45

Discussion

In this study, we investigate the performance of the slide window model on the clinical concept extraction task. Compared with the BiLSTM-CRF model, the slide window model has improved F1 scores in both the 2010 i2b2 and 2012 i2b2 corpora under the strict standard. However, we noticed that the slide window model compared with the BiLSTM-CRF model on the 2010 i2b2 and 2012 i2b2 tasks did not have the same improvement. This is mainly because the types of clinical concepts included in the 2010 i2b2 and 2012 i2b2 tasks are different. As can be seen from **Table 4**, compared with the two methods based on sequence labeling, the slide window model has improved performance in all categories of clinical concepts, but the magnitude of the improvement is different. Among them, the clinical concept of the 'OCCURRENCE' category has improved the most.

Through further research, we found that the slide window model has a relatively small improvement in relatively standardized clinical concepts compared to the sequence labeling model, such as clinical departments. This type of clinical concept contains a more uniform structure of named entities and higher similarity between named entities. Models based on sequence annotation are more comfortable to learn such clinical concepts, and it is easier to predict token labels in clinical concepts correctly. Conversely, for clinical concepts that have a relatively vague structure and a relatively sparse distribution of named entities, such as emergencies, the slide window model can achieve much better performance than the sequence labeling model.

Conclusion

In this paper, we comprehensively investigate the performance of the slide window model on clinical concept extraction. Experiments on the 2010 i2b2, 2012 i2b2 corpora prove that 1) Proposed modeling method based on candidate window classification is better than the model based on sequence labeling; 2) Based on the same pre-trained language model, the slide window model is better than the current mainstream clinical concept extraction methods (such as BiLSTM-CRF model

and softmax classifier); 3) By considering the overall semantics of clinical concepts instead of the semantics of each token, the performance of clinical concept extraction tasks can be effectively improved.

List of abbreviations

SW: slide window model

0 CNN: convolutional neural network

ICU: intensive care unit

NLP: natural language processing

BERT: bidirectional encoder representation from transformers

BiLSTM-CRF: bidirectional long short-term memory with conditional random field

Acknowledgements

Not applicable.

Funding

Not applicable.

Availability of data and materials

The datasets that support the findings of this study are available from i2b2 challenges, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Data are however available from i2b2's website: <https://portal.dbmi.hms.harvard.edu/data-sets/> upon reasonable application with a signed "data use and confidentiality agreement" to the program manager: Barbara Mawn (E-mail: Barbara.Mawn@hms.harvard.edu). The ethics approval would not been required as all the data have been De-Identified within the meaning of the Health Insurance Portability and Accountability Act of 1996 privacy regulations (HIPAA).

The datasets that support the findings of this study are available from MIMIC, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from MIMIC's website: <https://mimic.mit.edu/gettingstarted/access/> upon reasonable request and completing the CITI "Data or Specimens Only Research" course. The contents of restricted-access clinical databases maintained by PhysioNet were derived from original data that contained protected health information (PHI), as defined by HIPAA. The providers of the data have given scrupulous attention to the task of locating and removing all PHI, so that the remaining data can be considered de-identified and therefore not subject to the HIPAA Privacy Rule restrictions on sharing PHI.

Authors' contributions

The work presented here was carried out in collaboration between all authors. Y.T., S.L. and B.J. designed the methods and experiments, and contributed to the writing of manuscript. J.Y., Y.T. and Q.W. provided guidance and reviewed the manuscript critically. All authors have approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

References

1. Sun, W., Rumshisky, A., Uzuner, O.: Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association* **20**(5), 806–813 (2013)
2. Rink, B., Harabagiu, S., Roberts, K.: Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association* **18**(5), 594–600 (2011)
3. Lee, H., Peirsman, Y., andNathanael Chamberts, A.C., Surdeanu, M., Jurafsky, D.: Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In: *In Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 28–34 (2011)
4. Uzuner, O., South, B.R., Shen, S., Piccardi, M.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* **18**(5), 552–556 (2011)
5. Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z., Xu, H.: Entity recognition from clinical texts via recurrent neural network. *BMC Medical Informatics and Decision Making* **17**(supple), 67 (2017)
6. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *NAACL-HLT 2019* (2018)
7. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: *arXiv Preprint*, pp. 1603–01360 (2016)

8. Wang, Y., Wang, L., Moon, M.R.-M.S., Shen, F., Affza, N., Liu, S., Zeng, Y., saeed Mehrabi, Sohn, S.: Clinical information extraction applications: a literature review. *Journal of biomedical informatics* **77**, 34–49 (2018)
9. Bruijijn, B.D., Cherry, C., Kirichenko, S., Martin, J., Xhu, 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* **18**(5), 557–562 (2011)
10. Habibi, M., Weber, L., Neves, M., Wiegandt, D.L., Leser, U.: Deep learn with word embeddings improves biomedical named entity recognition. *Bioinformatics* **33**(14), 37–48 (2017)
11. Florez, E., Precioso, F., Riveill, M., Pighetti, R.: Named entity recognition using neural networks for clinical notes. In: *International Workshop on Medication and Adverse Drug Event Detection*, pp. 7–15 (2018)
12. Chalapathy, R., Borzeshi, E.Z., Picardi, M.: Bidirectional lstm-crf for clinical concept extraction. In: *arXiv Preprint*, pp. 1611–08373 (2016)
13. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: *NAAACL 2018* (2018)
14. Zhu, H., Paschalidis, I.C., Tahmasebi, A.: Clinical concept extraction with contextual word embedding. In: *arXiv Preprint*, pp. 1810–10566 (2018)
15. Lee, J., Yoon, W., Kim, S.K.D., Kim, S., So, C.H., Kang, J.: Biobert:pre-trained biomedical language representation model for biomedical text mining. In: *arXiv Preprint*, pp. 1901–08746 (2019)
16. Alsentzer, E., Murphy, J.R., Boag, W., weiHung Weng, Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. *arxiv preprint*. In: *arXiv Preprint*, pp. 1904–03323 (2019)
17. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission. In: *arXiv Preprint*, pp. 1904–05342 (2019)
18. Si, Y., Wang, J., Xu, H., Roberts, K.: Enhancing clinical concept extraction with contextual embeddings. In: *arXiv Preprint*, pp. 1902–08691 (2019)
19. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. *Scientific data* **3**, 160035 (2016)

Figures

Figure 1

Clinical concept coverage of different maximum window lengths on 2012 i2b2 tasks

Figure 2

Perform1D-convolution operation of length 2 on the candidate window

Figure 3

Perform 1D-convolution operation of length 2 on the candidate window

Figure 4

The overview architecture of the candidate concept generation

Figure 5

The labels distribution of tokens in the 2012 i2b2 task training set