Span Classification Based Model For Clinical Concept Extraction

Yongtao Tang[1], Jie Yu, Shasha Li, Bin Ji, Yusong Tan, and Qingbo Wu

College of Computer, National University of Defense Technology, Changsha, China
tyt941016@163.com

Abstract. 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. However, the mainstream method, sequence labeling model has some shortcomings. The clinical concept extraction method based on sequence labeling does not conform to the human cognitive model of language. At the same time, the extraction results produced by this method are difficult to couple with downstream tasks, which will cause error propagation and affect the performance of downstream tasks. To deal with these problems, we propose a span classification based method to improves the performance of clinical concept extraction tasks by considering the overall semantics of the token sequence instead of the semantics of each token. We call this model as span classification model.
Experiments show that the span classification model achieves the best micro-average F1 score(81.22%) on the corpora of the 2012 i2b2 NLP challenges, and obtained an F1 score(89.25%) comparable to SOTA in the 2010 i2b2 NLP challenges. Furthermore, the performance of our approach is always better than the sequence labeling model such as BiLSTM-CRF model and softmax classifier.

Keywords: Clinical Concept Extraction, Information Extraction, Natural Language Processing, Span-based Model, CNN

1 Introduction

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. [23] Clinical concept extraction is a precursor to downstream tasks such as relation extraction [20], and co-reference [13]. Multiple clinical corpora for different tasks such as 2010 i2b2, 2012 i2b2 NLP challenges [23, 24], contain the clinical concept extraction task as a sub-task.

The clinical concept extraction task is essentially a named entity recognition task. The mainstream solution for clinical concept extraction is modeling the clinical concept extraction task as a sequence labeling problem [17]. A common practice is employing a specific set of labels (such as 'BIO' or '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.

However, the clinical concept extraction method based on sequence annotation does not conform to human language cognitive habits. Humans determine the connotation of a phrase (such as the category of a phrase) mainly through the structure and overall semantics of the phrase [6], instead of judging the category of each component. Research has shown that the named entity recognition model based on the sequence labeling method is difficult to naturally couple with downstream tasks. [15] In downstream tasks that rely on named entity recognition, sequence labeling methods need to adopt a two-stage model [21]or introduce external language features (such as syntactic structure) [18] to solve the task. This leads to error propagation. [5] The error of named entity recognition will affect the performance of downstream tasks, but it cannot be corrected by the reverse gradient of downstream task training.

On the other hand, we have observed that compared with the task of named entity recognition in the general field, the task of clinical concept extraction has the characteristics: more standardized entity structure, and the longer entity may contain several shorter entities. For example: "reduction of joint dislocation" and "repair of dura defect", or "alcoholic liver cirrhosis" and "acute blood loss". Although these two groups of words have different types (surgery and disease), they are all composed of shorter clinical concepts, according to a specific structure. Therefore, it is helpful to improve the performance of clinical concept extraction tasks by using structural information and considering the semantic of shorter clinical concepts.

Based on the above two points, we propose a clinical concept extraction model based on span classification. In the proposed model, we capture the structural features in the span by using a convolutional neural network (CNN), and reuse the semantic of shorter span to generate the overall semantic representation. We call this model as a span classification model, and experimented on the corpora of the 2012 i2b2 NLP challenges and the 2010 i2b2 NLP challenges. Experiments show that the span classification model achieves the best micro-average F1 score(81.22%) on the corpora of the 2012 i2b2 NLP challenges, and obtained a an F1 score(89.25%) comparable to SOTA in the 2010 i2b2 NLP challenges. Our main contributions include:

- We use a modeling method that is different from the mainstream sequence labeling model and is more in line with human language cognitive habits to solve the clinical concept extraction task.
- We propose a new span vector representation method, which improves the performance of clinical concept extraction tasks by capturing the structural features of span and using the semantic information of the shorter span.
- We achieved state-of-the-art performance on the 2012 i2b2 task.

2 Related Work

2.1 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, feature-based approaches, and neural network-based approaches. The feature-based approaches manually design features according to clinical domain knowledge, and extract clinical concept by machine learning-based model (e.g., HMM, CRF) [2, 25]. The neural network-based approaches generally model clinical concept extraction problems as sequence 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 [3, 7, 8]. As pre-trained language models make significant advances in the NLP field [4, 19], more researchers applied pre-trained context-sensitive word embeddings to medical NLP. For instance, Elmo has shown excellent performance in clinical concept extraction [26]. Some recent studies have attempted to pre-train medical-specific language models using unlabeled medical field texts [1, 9, 14, 22] 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.

2.2 Span-based Method

The span-based method was first proposed by Lee et al. [15] and was used in the task of coreference resolution. They generate the semantic representation of spans through attention mechanism, and establish the co-referential link between spans by calculating the semantic similarity between spans. Dixit et al. [5] applied this method to relation extraction tasks and achieved SOTA results on multiple data sets. Jiang et al. [10] further extended their work to multiple tasks and verified the feasibility of the span-based model as a general information extraction model.

3 Model

We solve the clinical concept extraction problem with three steps which we explain in detail in the next subsection:

1) **Token Representation Generation**
   Use the pre-trained language model, such as BERT, to create the task-agnostic token embeddings for each token. The task-agnostic token embeddings are used to generate the task-specific embeddings for each possible span.
2) **Span Representation Generation**
   We propose a CNN-based model to generate the task-specific span embeddings. the task-specific span embeddings are used as the basis for the classification of clinical concepts.
3) **Span Classification**
   The span embeddings ars used to obtain a vector of clinical concept type scores for each span. Each span is assigned the clinical concept (Contains an additional 'NONE' type we defined) corresponding to its highest clinical concept type score.

3.1 Step 1: Token Representation Generation

For given document D with T tokens, we use x_t to represent the task-agnostic token embeddings of token t with $1 <= t <= T$. Inspired by the Elmo model, x_t is a concatenation of the raw word embedding and the contextual word embedding.

In this paper, we use Bidirectional Encoder Representation from Transformers (BERT) as the pre-trained language model. The embedding table has been trained in BERT is used to generate the raw word embedding x_t^r. And the final output of the BERT is represented as the contextual word embedding x_t^c

$$x_t = \{x_t^r, x_t^c\}$$

3.2 Step 2: Span Representation Generation

Our model classifies all possible spans. $span_{i,j}$ is defined by all the tokens from i to j inclusive. There are $N = \frac{T(T+1)}{2}$ possible text spans in D. The aim is to obtain a span representation $s_{i,j}$ for each $span_{i,j}$.

We propose a CNN-based network structure to generate span representations. The proposed structure uses shorter spans representation as features map, rather than instead of recalculating the underlying features of each span.

For the span with a length of 1, we use its corresponding token representation as the span representation.

$$s_{i,i} = x_i$$

For the spans with a length greater than 1, we use the following equation to calculate its span representation:

$$s_{i,j} = CNN(s_{i,j-1}, s_{i+1,j})$$

Where CNN is a convolution operator of length 2, and $1 <= i < j <= T$. The model design is based on a simple assumption that the convolutional neural network can retain the important features of the low-level when generating high-level features.

3.3 Step 3: Span Classification

In this step, we predict the clinical concept type for each span. This prediction is performed independently on each span. We compute the vector of clinical concept type scores for each span. The dimensions of the vector are the number of clinical concept types, including the 'NONE' type which represents the non-clinical concept span. We apply the softmax function to obtain a distribution over the clinical concept types. For $span_{i,j}$,

$$score_{i,j} = MLP(s_{i,j})$$
$$p_{i,j} = softmax(score_{i,j})$$

Where MLP is a Multi-Layer Perceptron. The output size of MLP and hence the size of p_i is equal to the number of clinical concept types. We predicte the span as the clinical type corresponding to its highest clinical concept type score. Unlike token-level models, overlapping spans can be extracted since the classification for each span is independent of other spans. For tasks that do not allow entity nesting, we greedily remain the span with the largest p_i among overlapping clinical concepts as the final result.

3.4 Loss

Considering that the number of possible spans far outweigh the number of clinical concepts contained in each sentence. the focal loss [16] is applied to solve the problem of unbalanced positive and negative samples in this model.

$$\hat{p}_{i,j} = \begin{cases} p_{i,j} & \text{if} \quad y = 1 \\ 1 - p_{i,j} & \text{if} \quad y = -1 \end{cases} \tag{1}$$

Where $y \in \{\pm 1\}$ specifies the ground-truth class and $p_{i,j} \in [0, 1]$ is the model's estimated probability for the clinical concept types.

$$loss(\hat{p}_{i,j}) = -(1 - \hat{p}_{i,j})^{\gamma} log(\hat{p}_{i,j})$$

Wher γ is a hyperparameter used to adjust the model's attention to the misclassified samples. We follow Lin et al. and simply set γ to 2.

4 Experiments

4.1 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. The 2012 i2b2 challenge data contains 190 training and 120 testing discharge summaries, with six clinical concept types. The database used in pre-training is MIMIC III [11], which is a public database and consists of almost $2million$ clinical notes.

4.2 Baseline

In this paper, we select two currently mainstream sequence labeling methods to compare with our method. One is the BiLSTM-CRF model, which is a widely used sequence labeling model. we implement this model follows Lample 2016 et al. [12] and use BERT to replace word2vector to generate contextual word embedding. The other is a method proposed by Google in 2018 [4] that only uses the softmax classifier to classify the token representation vectors generated by the pre-trained language model. In the following text, we use 'BiLSTM-CRF' and 'Softmax' to represent the BiLSTM-CRF model, softmax classification method respectively.

In addition, we also compared proposed method with the current SOTA results. The SOTA results in 2012 i2b2, and 2010 i2b2 were obtained by si et al. [22] Their work use the MIMIC III database to train pre-trained language models further, and use the BiLSTM-CRF model for clinical concept extraction. For comparison, we follow they work and train a domain-based pre-training language model using the MIMIC III database. 'SOTA' is used to represent the results they report.

4.3 Domain-based Pre-trained

The medical-specific model is initialized with base-sized BERT (BERT_{base}) and pre-trained using the MIMIC III database [11]. The model is represented by $\text{BERT}_{base-mimic}$. 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 $\text{BERT}_{base-mimic}$. These settings are the same as Si et al. [22]. 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.

4.4 Evaluation

We evaluated the overall performance of the model under strict standards. strict standard means that 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 specific performance metrics are precision, recall, and F1 score, and the micro 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.

4.5 Result

We conducted two sets of experiments to compare the performance of the proposed model with the mainstream sequence annotation model and the SOTA model.

In the first experiment, a base-sized BERT model was used as the pre-trained model to ensure that the experimental results can be reproduced easily. In addition, the dimensions of the representation vector are set to 768, which is consistent with the hidden size of BERT_{base}, in order to reduce the influence of hyperparameters on the experimental results.

Table 1 shows the performance of different models under two clinical concept extraction tasks. The performance is evaluated in the F1 score. In general, our 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 1. The micro-F1 scores of deifferent models using BERT_{base} in i2b2 2010 and i2b2 2012 datasets(%)

Method	**2010 i2b2**	**2012 i2b2**
Softmax [4]	84.70	76.85
BiLSTM-CRF [12]	84.94	76.59
Our model	**85.83**	**78.45**

In the second experiment, we re-implement the work of Si et al. and trained a domain-based pre-trained language model. Because pre-training the language model using external resources is not the focus of this paper, we do not make detailed adjustments to the hyperparameters. Therefore, the performance of the reproduced domain-based pre-trained language model is not optimal(As shown in Table 2) . We use a reproduced language model to replace BERT_{base} for experiments.

Table 2. The results of i2b2 2010 and i2b2 2012 datasets with the domain-based pre-trained language model (%)

Method	**2010 i2b2**	**2012 i2b2**
SOTA [22]	**89.55**	80.34
Re-implement	88.18	79.32
Our model	89.25	**81.22**

Table 2 shows the performance comparison of our model with the SOTA results in 2012 i2b2 and 2012 i2b2. For the 2012 i2b2 corpus, the best performance is achieved by the proposed model with a micro-F1 score of 81.22%. It

improves the performance by 0.88% over the previous SOTA result achieved by Si 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 the reproduction model, we still improved the F1 score by 1.07%.

What's more, we analyzed the performance of the model on different types of clinical concepts to understand the effect of our model on different types of clinical concepts. Take the performance of the model using the BERT_{base} in each clinical concept category in the i2b2 2012 task as an example.

Table 3. Performance of our model in each type of clinical concept in the 2012 i2b2 task.(%)

	Softmax	**BiLSTM-CRF**	**Our model**
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

As shown in **Table 3**, our model has improved micro-F1 scores in all types of clinical concepts compared with the sequence labeling model. In particular, the model has obvious performance improvements in the clinical concepts of the'TEST' and 'TREATMENT' types, and the micro-F1 increased by 1.22% and 0.82% respectively. This is in line with expectations Because these two types have regular syntactic structures. Interestingly, our model has achieved the greatest performance improvement on the 'OCCURRENCE' type, up to 3.89%. This may be because the 'OCCURRENCE' type generally has the structure of a verb phrase.

5 Conclusion

In this paper, we investigate the performance of the span-classification based model on clinical concept extraction and proposed a CNN-based span represent method. Experiments on the 2010 i2b2, 2012 i2b2 corpora prove that 1) Using a model based on span classification can effectively solve the clinical concept extraction task; 2) By considering the structure information of span and reuse the features of shorter spans, the performance of clinical concept extraction tasks can be effectively improved; 3) Based on the same pre-trained language model, our model is better than the current mainstream clinical concept extraction methods (such as BiLSTM-CRF model and softmax classifier).

References

1. 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, p. arXiv:1904.03323 (2019)
2. Bruijin, 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)
3. Chalapathy, R., Borzeshi, E.Z., Picardi, M.: Bidirectional lstm-crf for clinical concept extraction. In: arXiv preprint, p. arXiv:1611.08373 (2016)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT 2019 (2018)
5. Dixit, K., Al-Onaizan, Y.: Span-level model for relation extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5308–5314. Association for Computational Linguistics, Florence, Italy (2019). DOI 10.18653/v1/P19-1525. URL https://www.aclweb.org/anthology/P19-1525
6. Finkel, J.R., Manning, C.D.: Nested named entity recognition. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 141–150. Association for Computational Linguistics, Singapore (2009). URL https://www.aclweb.org/anthology/D09-1015
7. Florez, E., Precioso, F., Riveill, M., Pighetti, R.: Named entity recognition using neural networks for clinical notes. In: In International Workshop on Medication and adverse Drug Event Detection, pp. 7–15 (2018)
8. Habibi, M., Weber, L., Neves, M., Wiegandt, D.L., Leser, U.: Deep learn with word embeddings improves biomedical named entity recognition. Bioinformatics **33(14)**, i37–i48 (2017)
9. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission. In: arXiv preprint, p. arXiv:1904.05342 (2019)
10. Jiang, Z., Xu, W., Araki, J., Neubig, G.: Generalizing natural language analysis through span-relation representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2120–2133. Association for Computational Linguistics, Online (2020). DOI 10.18653/v1/2020.acl-main.192. URL https://www.aclweb.org/anthology/2020.acl-main.192
11. 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**, 160,035 (2016)
12. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: arXiv preprint, p. arXiv:1603.01360 (2016)
13. 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)
14. 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, p. arXiv:1901.08746 (2019)
15. Lee, K., He, L., Lewis, M., Zettlemoyer, L.: End-to-end neural coreference resolution. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 188–197. Association for Computational Linguis-

tics, Copenhagen, Denmark (2017). DOI 10.18653/v1/D17-1018. URL https://www.aclweb.org/anthology/D17-1018
16. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 318–327 (2020)
17. 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)
18. Miwa, M., Bansal, M.: End-to-end relation extraction using LSTMs on sequences and tree structures. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1105–1116. Association for Computational Linguistics, Berlin, Germany (2016). DOI 10.18653/v1/P16-1105. URL https://www.aclweb.org/anthology/P16-1105
19. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: NAACL 2018 (2018)
20. 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)
21. dos Santos, C., Xiang, B., Zhou, B.: Classifying relations by ranking with convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 626–634. Association for Computational Linguistics, Beijing, China (2015). DOI 10.3115/v1/P15-1061. URL https://www.aclweb.org/anthology/P15-1061
22. Si, Y., Wang, J., Xu, H., Roberts, K.: Enhancing clinical concept extraction with contextual embeddings. In: arXiv preprint, p. arXiv:1902.08691 (2019)
23. 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)
24. 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)
25. 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)
26. Zhu, H., Paschalidis, I.C., Tahmasebi, A.: Clinical concept extraction with contextual word embedding. In: arXiv preprint, p. arXiv:1810.10566 (2018)