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

CURRENT STATUS: UNDER REVIEW

BMC Medical Informatics and Decision Making BMC Series

Yongtao Tang
Nacional'nyj issledovatel'skij universitet Moskovskij institut elektronnoj tehniki Fakul'tet Intellektual'nyh tehniceskih sistem

tyt941016@163.com*Corresponding Author*
ORCiD: https://orcid.org/0000-0002-7378-8949

Shasha Li
National University of Defense Technology

Bin Ji
National university of Defense Technology

Jie Yu
National University of Defense Technology

Yusong Tan
National University of Defense Technology

Qingbo Wu
National University of Defense Technology

DOI:

10.21203/rs.2.22462/v1

SUBJECT AREAS

Medical Informatics

KEYWORDS

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

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.

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Figures

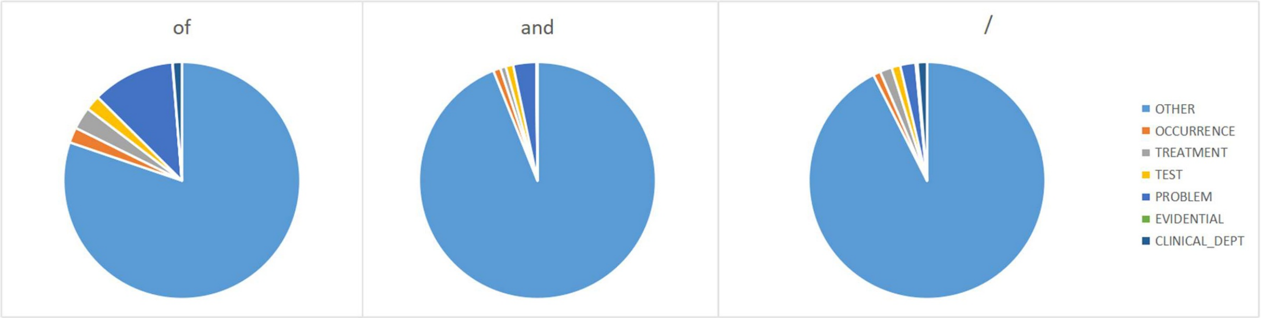

Figure 1

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

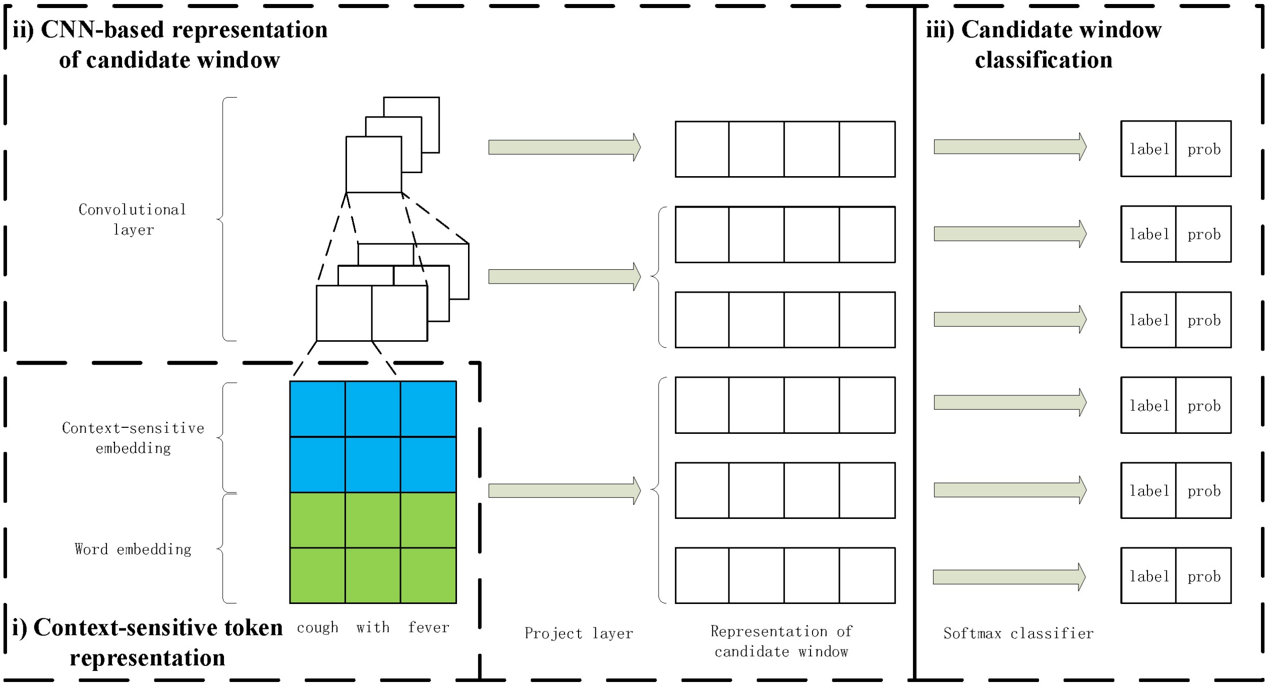
ii) CNN-based representation of candidate window
iii) Candidate window classification
Convolutional layer
Context-sensitive embedding
Word embedding
i) Context-sensitive token representation
cough
with
fever
Project layer
Representation of candidate window
Softmax classifier
label
prob

Figure 2

The overview architecture of the candidate concept generation

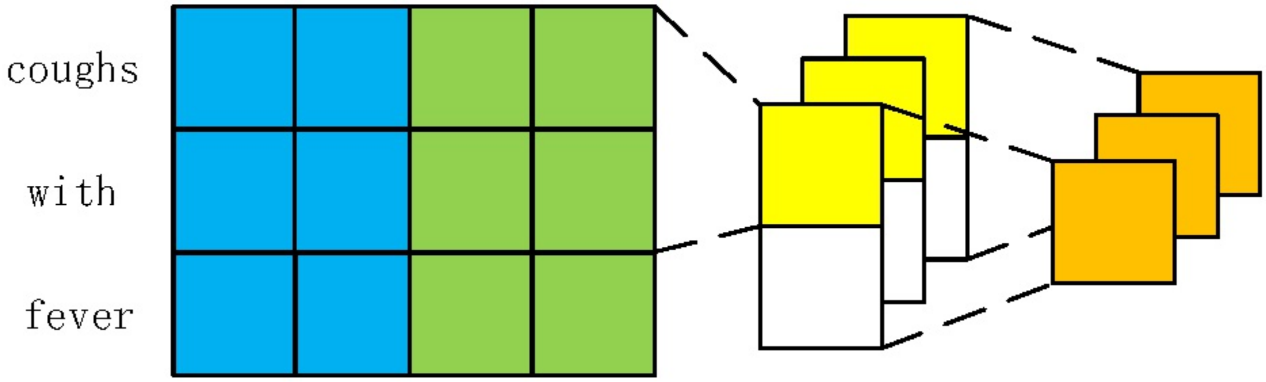

Figure 3

Perform1D-convolution operation of length 2 on the candidate window

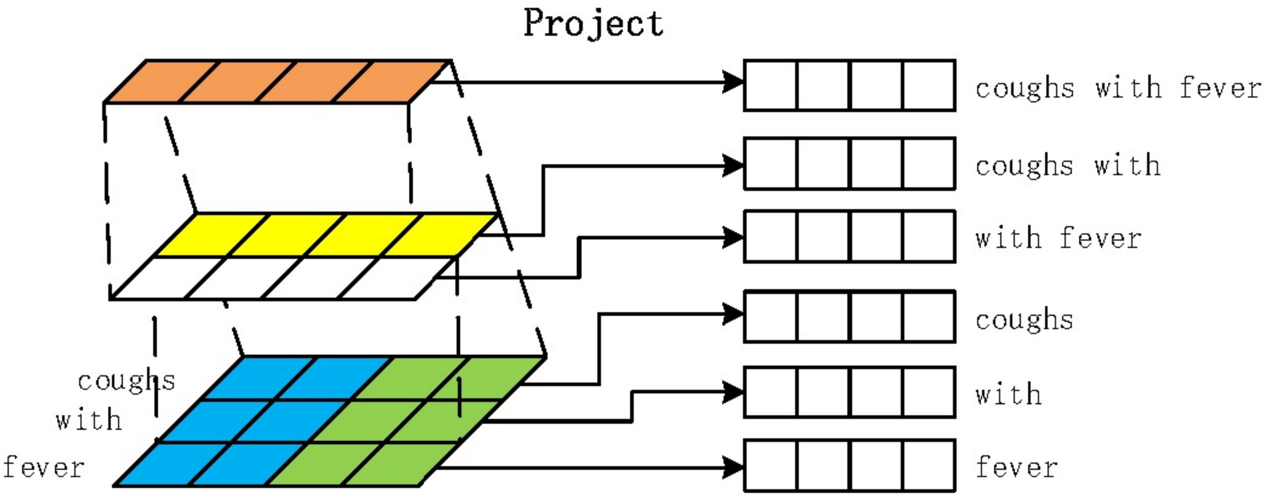

Figure 4

Perform1D-convolution operation of length 2 on the candidate window

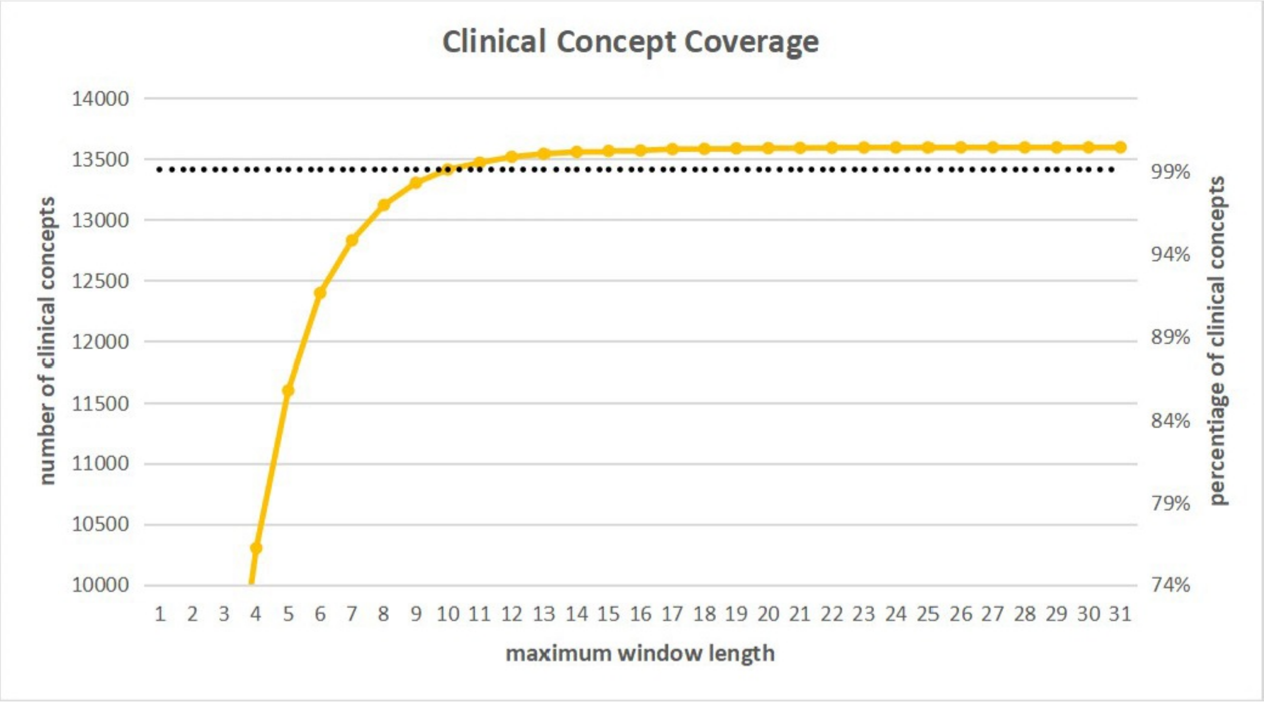

Figure 5

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