A Study of Convolutional Neural Networks for Clinical Document Classification in Systematic Reviews: SysReview at CLEF eHealth 2017

Grace Eunkyung Lee

School of Computer Science and Engineering
Nanyang Technological University, Singapore
leee0020@e.ntu.edu.sg

Abstract. Identifying eligible documents for systematic reviews is one of the most time-consuming steps in writing the reviews. From retrieving numerous clinical documents to manually checking the documents with detailed criteria requires a tremendous amount of time and skilled workforce. In this paper, to increase the efficiency of the process we examine the role of convolutional neural networks for classifying medical documents for systematic reviews. The analysis is carried out in the context of the CLEF 2017 eHealth Task 2 as a participant. The evaluation demonstrates that the suggested methods show slightly better performance for full document screening than abstract screening.

Keywords: document classification, systematic review, convolutional neural network

1 Introduction

Recognizing relevant documents out of thousands of documents is one of the most time-consuming yet important steps in writing systematic reviews. Systematic reviews analyze and appraise all pertinent literature that meets a set of pre-defined eligibility criteria. Before analyzing selected literature for a review, systematic review authors need to filter related documents by manually investigating numerous documents for their eligibility. Since missing out relevant documents is critical, researchers initially collect thousands of documents from several databases which might be eligible for a review. The collected documents are thoroughly examined for eligibility through two steps of abstract and full document screenings.

There have been several studies to automatic the laborious screening process. However, imbalanced data and different levels of complexity for eligibility criteria make automating the process a challenging task. Specifically, among 50 Cochrane systematic reviews more than 5,000 documents are initially collected on average, and only around 20 documents are turned out to be eligible for the review as indicated in Table 1. Furthermore, systematic reviews have a broad range of topics from education for health professionals to heart disease and blood circulation. The review topic and its scope lead to manifold eligibility criteria [2].

The approaches toward solving the issues have been proposed for the past years. Regarding to the imbalanced data, negative undersampling and weighting schemes are used and, especially, active learning showed promising performance to settle the limited number of positive data [6]. Moreover, the majority of existing work for improving screening process applied feature selections and conventional machine learning algorithms such as SVM to train classifiers. In addition, in [5], systematic reviews from two different domains are evaluated and show different characteristics, but the number of reviews is limited and diversity of review topics can be further expanded.

In this paper, we examine the efficacy of convolutional neural networks(CNN) for medical document classification in systematic reviews. The analysis of the approach is carried out in the context of CLEF 2017 eHealth Task 2: Technologically Assisted Reviews in Empirical Medicine [1, 3]. The contribution of this approach is studying a modern machine learning algorithm, CNN, on the task of identifying eligible clinical documents, despite the challenges of imbalanced data distribution. To resolve the imbalance of data with the small number of positive cases, we train the model on sentence-level context, rather than document-level, with undersampling. We also concatenate context of systematic reviews criteria with sentences of the collected documents. This provides a hint of anchor information of to which systematic review each sentence is related. We evaluate various combinations of contexts from reviews and documents with CNN.

The remainder of the paper is laid out as follows. In Section 2 we present data description and Section 3 provides detailed description of our approach on the task and variations of models. In Section 5 we evaluate our models and analyze the results. Finally, we conclude the paper by summarizing the major results in Section 5.

2 Data description

In this work, we use the CLEF 2017 eHealth Task 2 dataset [3]. The dataset consists of 50 diagnostic test accuracy (DTA) Cochrane systematic reviews. The reviews include a title, boolean queries, and PMIDs retrieved from the queries. Besides, PMIDs are indicated for eligibility results after two-stage screening: abstract screening, and full document screening.

Table 1 demonstrates statistics of medical documents collected from 50 systematic reviews and the number of documents as a result of examining title and abstract, and full document screening, respectively. From the Table 1, we can see that the initial collection contains numerous medical research papers. In contrast, the number of positive documents after abstract screening is a small fraction of the entire collection. Even further, after full document screening, the final number of documents to be included in the reviews is dramatically reduced from the initial collection of documents. Hence, the collection of documents retrieved via boolean queries are noisy and contain many irrelevant documents for reviews.

Table 1: Statistics of clinical documents from 50 systematic reviews

	Total # of documents	# of documents after title and abstract screening	# of documents after full document screening
Min	64	2	0
Max	43411	619	99
Average	5389.26	93.22	21.86
SD	7040.28	123.88	22.24

3 Approach

Different from common approaches to document categorization or sentiment analysis, several inherent characteristics of systematic reviews make the current task unique and challenging. One characteristic of the task is scarcity of positive data. The number of final positive documents is not enough to train a model for a review since it is often less than 50. In spite of adopting techniques of reducing the imbalance, the absolute number of positive documents is still not sufficient. In order to overcome data sparsity, we combine all documents in training dataset and utilize sentences as a training unit to build one general classifier for DTA systematic reviews.

Training a general classifier leads to face another challenge. In this task, each document can be classified either positive or negative, depending on eligibility criteria, and the eligibility criteria vary over systematic reviews. For instance, a medical document is positive in review A and negative in review B because of different criteria, even though the document is retrieved in both the reviews. As a result, one document labeled positive and negative becomes training inputs for one classifier. Thus, a document or sentence itself is not able to be a stand-alone input as training data.

To resolve the challenge, we provide eligibility criteria with each sentence by concatenating a title of reviews and sentences from clinical documents. Since titles of reviews contain imperative elements of reviews in a brief format, we believe that titles of reviews would provide a snippet of eligibility criteria of reviews. The detailed description of concatenating eligibility criteria and sentences are demonstrated in Section 4.3.

3.1 CNN model

In this section, we explain a simple CNN with one layer of convolution on top of word vectors. The model is a slight variant of the CNN architecture proposed in [4]. Let $x_i \in \mathbb{R}^k$ be the k-dimensional word embedding vector corresponding to the i-th word in the sentence. A sentence s is represented as

$$s = x_1 \oplus x_2 \oplus \cdots \oplus x_n, \tag{1}$$

where $\oplus$ is the concatenation operator and n is the maximum length of sentences. Sentences are padded to the maximum length if necessary. Likewise, review context is represented as equation 1. A convolution layer involves a filter $w \in \mathbb{R}^{hk}$, which is applied with a window size, h, to grasp information from surrounding words. A new feature, c_i capturing the context with a window of h words, is generated by

$$c_i = f\left(w \cdot x_{i:i+h-1} + b\right), 1 \leq i \leq (n-h+1) \tag{2}$$

where b is a bias term and function f is a hyperbolic tangent. As a result, a sentence is represented by multiple feature vectors

$$c = [c_1, c_2, \ldots, c_{n-h+1}] \tag{3}$$

with $c \in \mathbb{R}^{n-h+1}$. Next, we max-pool the result of the convolutional layer into a long feature vector, $\widehat{c} = \max\{c\}$, which is to merge the results into the most representative feature vector. We then incorporated the common regularization method, dropout, to prevent feature vectors from co-adapting and force them to learn useful features in a independent manner. For more details on dropout we refer the reader to [4]. After regularization we classify the result using a softmax layer. Finally, predicted results of sentences are combined for document classification, which is our ultimate task, and derive the document classification as follows

$$D = \frac{1}{|D|} \sum_{s \in D} p\left(s\right), \tag{4}$$

where $|D|$ is the total number of sentences in a document D and $p\left(s\right)$ is the prediction probability of sentence s derived from the CNN model.

4 Experimental Setup

In this section, we discuss our experimental setup to evaluate the effectiveness of the proposed approach for modeling a medical document classifier. In particular, Section 4.1 describes the preprocessing and normalization according to characteristics of biomedical text, while Section 4.2 presents the hyperparameters for the CNN model. Section 4.3 discusses the variants of our approaches used in our experiments.

Undersampling is a common way to deal with imbalanced data. We used negative undersampling when training classifiers because of the limited number of eligible documents compared to irrelevant documents.

Given ids for PubMed documents collected for 50 systematic reviews, we used a title and abstract of clinical documents from PubMed. Even though the goal of a task is to improve both of a title and abstract screening and full document screening, we solely exploit only a title and abstract of the documents as input data. We believe they contain the most important content of documents like a summary.

Table 2: Statistics of training data. The number of relevant sentences are the total number of sentences of relevant documents from 20 systematic reviews

	# of relevant sentences	# of irrelevant sentences	**Total**
Training	4435	4437	**8872**

4.1 Normalization

Prior to classification, sentences from documents undergo normalization in which a script using regular expressions simplifies complex numerical and mathematical notation into a canonical form. All integers, real numbers, and percentage are mapped to INT, FLOAT and PERCENT, respectively. Acronyms are appeared with parenthesis when they are mentioned for the first time, so the parenthesis are eliminated and the acronyms are considered as single words. Lastly, measurements such as dosages, 100g/d, are normalized by MEASUREMENT.

4.2 Hyperparameters and Training

After normalization word tokens are represented by pre-trained word embedding. In order to reflect characteristics of biomedical text, we leverage the pre-trained Word2Vec vectors with PubMed and PubMed Central dumps [1]. Since the word embeddings are trained on the entire available biomedical literature, we believe that it can effectively capture semantics for the biomedical domain. The vector representations has the dimensionality of 200. If words are not present in the set of pre-trained words, they are initialized with all zeros.

We use rectified linear units and filter size (h) are set to 3, 4, 5, and 6. The dropout rate is set to 0.5, mini batch size is 50, and $L2$ norm in regularization is not used for the purpose of simplicity. For training, 20 systematic reviews are employed and the training is conducted through stochastic gradient descent over shuffled mini-batches. The rest 30 reviews are allocated for testing which is identical with the set up of CLEF eHealth 2017 Task 2. The statistics of training data for relevant and irrelevant sentences is presented in Table 2.

4.3 Variations of Model

In this work, we try three variants of data concatenation between eligibility criteria and documents to be evaluate. Eligibility criteria have various elements for reviews and they are often described in a document so-called protocols. Rather than accessing long and descriptive protocols about criteria, we consider a title of systematic review as criteria, since a review title represents vital elements of documents in a brief format. By providing a hint of eligibility criteria, each sentence is differentiated from which criteria it is evaluated on. The variations of concatenation of criteria and sentence information are as follows.

[1] http://bio.nlplab.org/

- **Cri-Titlesent**: A model where a title of systematic review, a title of medical document are concatenated to a sentence of abstract of the medical document as prefix and utilized as input data.
- **Cri-Sent**: Same as the model above but a title of systematic review is used concatenated except a title of document.
- **Cri-Title**: A model where a title of systematic review and a title of clinical documents are combined. Sentences from abstract are not used in this model. Thus, compared to the previous models, it is built with less input data.

5 Results and Discussion

In this section, we present and discuss the results obtained by our models on the test data of the task.

Table 3 shows results of the three models on different evaluation measurements. A wss@N indicates Work Saved over Sampling @ Recall and measures how much a model reduces workload of reviewers [6]. Measuring reduced workload has been one of the common evaluation approaches for the task of automating screening process in systematic reviews. A norm area represents area under the cumulative recall curve normalized by the optimal area. More details on evaluation measures used in the task, we refer to [3].

Since CNN architecture requires massive amount of training data to achieve reasonable performance, the suggested models show poor performances. This indicates that the models meed more consistent labeled data even though the number of training data has been increased in this models. Compared to the two models, Cri-Titlesent and Cri-Sent, the model Cri-Title displays lower performance because it utilizes the fewer number of data for training.

From the results of wss@100 and wss@95 presented in Table 3, the proposed models have slightly better evaluation results on full document screening than abstract screening. The relevance results of abstract screening include not only relevant cases but also cases which cannot be judged because of the lack of information in the currently given data. Hence, the results from abstract screening might be less consistent.

Besides, further investigations on the limited performances revealed that the model fails to make right predictions when there is no abstract text. Some relevant documents do not have abstract text in PubMed, only their titles. Therefore, low-ranked relevant studies deteriorate the overall performances.

We believe that performances of the models have room for improvement. Handling the process of collecting abstract text of relevant studies from various biomedical literature databases as well as PubMed, the increased training data, and fine tuning on CNN architecture will lead to enhanced results. We leave this part as future work for improvement.

6 Conclusion

In this work, we have presented simple CNN models for improving the laborious task of identifying eligible documents for systematic reviews. The suggested

Table 3: Evaluation results of the three models

	Cri-Titlesent		Cri-Sent		Cri-Title	
	abstract	full_doc	abstract	full_doc	abstract	full_doc
wss@100	0.089	0.204	0.117	0.204	0.091	0.148
wss@95	0.108	0.217	0.131	0.197	0.075	0.141
norm area	0.612	0.647	0.595	0.618	0.538	0.545
total cost uniform	3936.481	3606.368	3936.481	3606.368	3937.347	3607.195
total cost weighted	4130.067	3557.586	4130.067	3557.586	4130.933	3558.414
average precision	0.078	0.05	0.06	0.039	0.052	0.024
reliability	0.548	0.717	0.548	0.717	0.549	0.718
loss_r	0.01	0	0.01	0	0.01	0
loss_e	0.538	0.717	0.538	0.717	0.539	0.717
recall	0.982	0.996	0.982	0.996	0.982	0.996

models are designed for any DTA systematic reviews even though every systematic review is accompanied with different complexities of eligibility criteria. The models take advantage of concatenated context from criteria and clinical documents. The evaluation results show that while the performance of the proposed approaches has room for improvement, they have higher performance in full document screening than abstract screening. This work is a step towards applying deep neural networks to improve the screening process despite the scarcity of labeled documents and the data imbalance.

References

[1] Lorraine Goeuriot et al. "CLEF 2017 eHealth Evaluation Lab Overview". In: *CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS)*. Springer. 2017.

[2] Julian PT Higgins and Sally Green. *Cochrane handbook for systematic reviews of interventions*. Vol. 4. John Wiley & Sons, 2011.

[3] Evangelos Kanoulas et al. "CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview". In: *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum*. CEUR Workshop Proceedings. 2017.

[4] Yoon Kim. "Convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1408.5882* (2014).

[5] Makoto Miwa et al. "Reducing systematic review workload through certainty-based screening". In: *Journal of biomedical informatics* 51 (2014), pp. 242–253.

[6] Alison ÓMara-Eves et al. "Using text mining for study identification in systematic reviews: a systematic review of current approaches". In: *Systematic reviews* 4.1 (2015).