Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 221 (2023) 557–564

Procedia Computer Science

www.elsevier.com/locate/procedia

10th International Conference on Information Technology and Quantitative Management

Identifying key elements for evidence-base medicine using pretrained model and graph convolution network

Fengchun Yang[a, b,+], Xiaowei Xu[a, b,+], Meng Wu[a, b], Xuwen Wang[a, b], Liu Shen[a, b], Qing Qian[a], Jiao Li[a, b] *

[a]*Institute of Medical Information, Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing 100020, China*
[b]*Key Laboratory of Medical Information Intelligent Technology, Chinese Academy of Medical Sciences, Beijing 100020, China*

Abstract

Objective: Evidence-based medicine (EBM) provides a framework to support clinicians' decision-making processes using the best evidence currently available in the field. The key elements of clinical research can be defined by a framework called PICO, which identifies the sentences in a medical literature text that belong to the four key elements reported in clinical trials: Participants/Problem (P), Intervention (I), Comparison (C) and Outcome (O). This study aims to establish an effective detection method for key elements of evidence-based medical research.
Methods: Based on the text features of key elements framework, we propose a deep learning model BERTGCN which combined large-scale pretraining model and graph convolution network (GCN) for detecting key elements of evidence-based medicine. In this model, the sentences which are initialized with pre-trained BERT representations and the words in the EBM evidence were recognized as the sentence nodes and word of the graph which were used to train GCN model. At the same time, the sentences were used to fine tune the pretraining model.
Results: We tested our proposed approach over PubMed-PICO dataset is a data set containing tens of thousands of EBM key elements extracted from PubMed. The F1-score of P/I/O in the model we proposed reached 91.3%, 85.8% and 90.0% respectively. Experimental results show that our model outperforms the current optimal model.
Conclusions: BERTGCN is able to leverage the advantages of both worlds: large-scale pretraining and transudative learning to improve the efficiency of detecting evidence-based medical evidence from research publications.

Keywords: evidence-based medicine, natural language process, graph convolution network, evidence identification

1. Introduction

Evidence-Based Medicine (EBM) enables medical practitioners to form treatment plans based on the complete available evidence [1]. Typically, the analyses that underlie EBM begin by selecting a set of potentially relevant papers,

* Corresponding author. Tel.: +86-10-52328740; *E-mail address:* li.jiao@imicams.ac.cn

10.1016/j.procs.2023.08.022

which are then further refined by human judgment to form the evidence base on which the answer to a specific question depends [2,3]. These evidence are the products of research studies, usually in the forms of Randomized Control Trials (RCTs) or Clinical Trials (CTs) that investigate the effects of a treatment on a specific group of patients and present their findings [4,5].

In practice, successful EBM applications rely on answering clinical questions via analysis of large medical literature databases such as PubMed[4]. As vast evidence bases such as PubMed grows exponentially and rapidly, it becomes more difficult to obtain appropriate evidence-based evidence from the literature library. The increasingly growing number of medical publications is making it extremely difficult for healthcare staff and medical practitioners to stay updated with the latest research and guidelines. Research questions in clinical study were usually defined by the PICO framework, the question were decomposes into four parts: Participants/Problem (P: the characteristics of the study population); Intervention (I: the primary intervention considered); Comparison (C: comparison for the intervention); Outcome (O: the anticipated measures, improvements, or effects). SO, well-formulated and structured research key elements can increase productivity in developing or updating clinical practice guidelines and medical knowledge bases. It is important to automate the extraction of PICO elements from clinical RCT literature for clinicians and investigators to review to reduce the time required to search for key elements of clinical research. An example abstract from the study [6] and it's structured evidence-based medical evidence are shown in Figure 1.

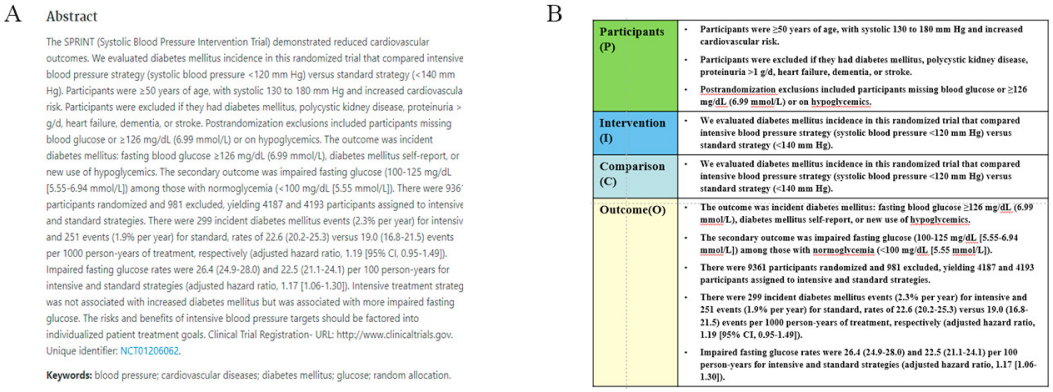

Figure 1 There is a need to transform free-text research abstracts into structured evidence-based text fragments to accelerate the evidence synthesis process. (A) Research abstracts in free text form; (B) Structured text fragments of evidence-based evidence

In previous studies, the generalized use of the PICO framework or similar schema by clinicians was validated for its performance improvement on searching literature for clinical questions [7]. Natural language processing (NLP) -based methods were used to solve evidence-based medicine problems [8,9]. The PICO element detection was formalized as a segment classification task [10]. These studies used machine learning techniques such as naive Bayes [11], random forest [12], support vector machine [13], conditional random field [14]. However, these methods relied on the quality of features of collected from the literature, such as the lexical features, the semantic, syntactic, sequential features. With the development of the natural language process and deep artificial neural network, there were studies used the bidirectional long-short-term memory (bi-LSTM) model [15] and pre-trained language model trained from the large-scale biomedical literature corpus [16] to improve the performance of the classification task. Graph neural network (GNN) is an effective transduction learning method. Some researches [17,18] regard the text classification task as the node classification task of graph neural network, and have achieved good results. Compared with other neural network models, graph neural network models can better utilize the global co-occurrence information of words in natural language processing tasks. And pretraining models have performed very well in natural language processing tasks. So in this work, we propose to boost the PICO element detection accuracy for deep learning models by exploiting two directions in order to make possible the automated identify key elements for evidence-base medicine.

2. Methods and Materials

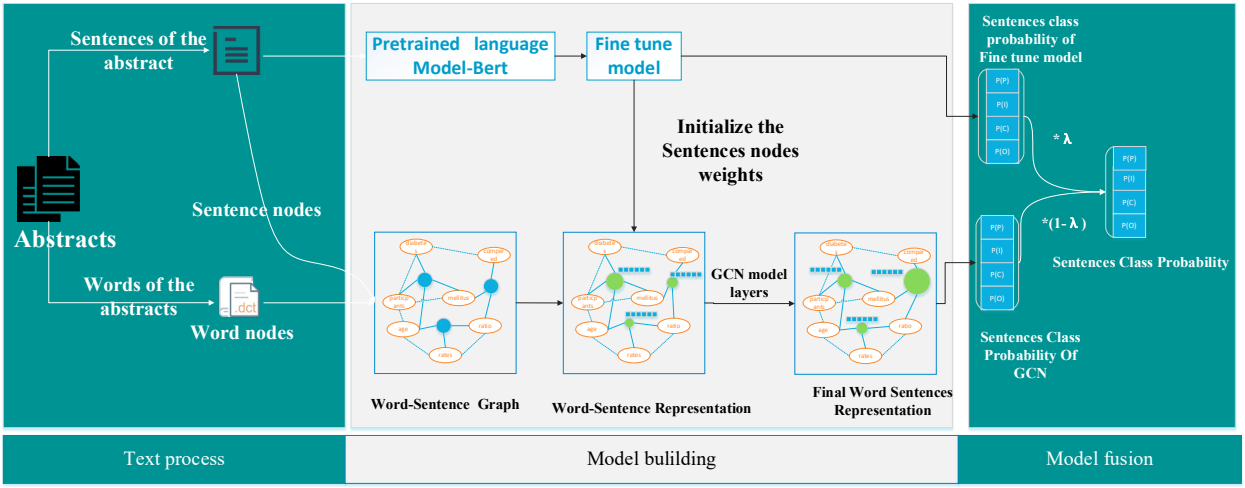

Figure 2 The overall workflow includes three parts: text processing, model construction and model fusion. In the text processing stage, we obtain the text for model training and the set of sentences and words for constructing the text graph. Fine-tuning and constructing GCN models for text classification; then the output probabilities of the two models are fused.

In this study, we regard key element recognition as a node classification task in graph, which is essentially a text classification task. Raw text cannot be directly used for text classification, we convert the text into a data format that can be recognized by the model. In this study, we use the pre-training model to represent the text. In general, the sentences in the abstract and the words that make up the sentences are regarded as nodes of the heterogeneous graph, and sentences embeddings are initialized with pre-trained BERT representations, then a graph convolutional networks (GCN) model was used for sentences classification. Meanwhile, we use a direct BERT embedded auxiliary classifier to optimize the BERTGCN classifier. We use the dataset to train a classifier by fine-tuning the model on a pre-trained model.

Graph Convolution Networks： We use the classic multiple-layer GCN for document classification. The main characteristic of GCN is a multilayer neural network that operates directly on a graph and induces embedding vectors of nodes based on properties of their neighborhoods. Formally, consider a graph $G = (V, E)$, where $V(|V| = n)$ and E are sets of nodes and edges, respectively. Let $X \in R^{n*m}$ be a matrix containing all n nodes with their features, where m is the dimension of the feature vectors, each row $x_v \in R^m$ is the feature vector for v. We introduce an adjacency matrix A of G and its degree matrix D, where $D_{ii} = \sum_j A_{ij}$. For a one-layer GCN, the new k-dimensional node feature matrix$L^1 \in R^{n*k}$ is computed as:

$$\boldsymbol{L}^{(1)} = \boldsymbol{\rho}\left(\tilde{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}_0\right) \quad \mathbf{1}$$

Where is the normalized symmetric adjacency matrix and ρ is an activation function. The model can only calculate the neighbor node directly connected to the node, only when multiple GCN layers are stacked, information about larger neighborhoods are integrated. The $j-$ th layer is calculated as follows:

$$\boldsymbol{L}^{(j+1)} = \boldsymbol{\rho}\left(\tilde{\boldsymbol{A}}\boldsymbol{L}^{j}\boldsymbol{W}_0\right) \quad \mathbf{2}$$

Large-scale pretraining model: In the step of text representation, we use a pre-training model based on large-scale corpus for text representation. Compared with word embedding model, like Word2Vec which is a context-free text representation, BERT is a contextualized representation model, which can combine contextual information of text. BERT extracts the contextualized embeddings by training a deep bidirectional encoder from transformers [19] on the BooksCorpus and English Wikipedia. The transformer structure mainly consists of identical blocks, and each block contains sub-modules based on multi-head self-attention and a feed-forward neural network. It dispenses with

recurrence and convolutions, and achieves state-of-the-art performance on NLP tasks. The pre-trained BERT can be fine-tuned with a simple additional output layer for downstream tasks.

2.1. Model structure

As is shown in Figure2, the model contains two parts, one of the part is an auxiliary classifier that directly operates on bidirectional encoder representation from transformers (BERT) [20] embeddings leads to faster convergence and better performances, the other part of the model is a GCN model [21]. The final training objective is the linear interpolation of the prediction from BERTGCN [22] and the prediction from BERT, which is given by:

$$\boldsymbol{Z} = \boldsymbol{\lambda}\boldsymbol{Z}_{\boldsymbol{GCN}} + \left(1-\boldsymbol{\lambda}\right)\boldsymbol{Z}_{\boldsymbol{BERT}} \quad \textbf{3}$$

Where the λ was used to controls the tradeoff between the two parts of the whole model. When λ=1, it means that we use the full GCN model, λ=0 means that we only use the BERT model. When λ ∈ (0,1), we are able to balance the predictions from both of the models, and the BERTGCN model can be better optimized.

The inputs of Z_{GCN} model is a heterogeneous graph constructed following the TextGCN [18], which contains two kinds of node and two kinds of edges : sentence nodes and word nodes, word-sentence edges and word-word edges. An identity matrix $X = I_{n_{doc}+n_{word}}$ is used as initial node features, where n_{doc} is the number of sentence nodes n_{word} is the number of the word nodes, both of the training and test set included. We initialize representations for sentence nodes in a text graph using a BERT-style model (e.g., BERT, RoBERTa [23]). In this study we used BioBERT, which is initialized by BERT and further trained on biomedical literature including PubMed abstracts and PMC full-text articles [24], to initialize the representations of the sentences of the abstracts.

Sentence node embeddings are denoted by $X_{doc} \in R^{n_{doc}*d}$, where d is the dimension of the embeddings of the representations. The initial node feature matrix is given by:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{\boldsymbol{doc}} \\ 0 \end{pmatrix}_{(\boldsymbol{n}_{\boldsymbol{doc}}+\boldsymbol{n}_{\boldsymbol{word}})*\boldsymbol{d}} \quad \textbf{4}$$

The edges of the graph were defined based on the term frequency-inverse sentence frequency (TF-IDF) [25] and positive point-wise mutual information (PPMI) which is a popular measure for word associations, to calculate weights between two word nodes. The weight of the edge between a sentence node and a word node is the term frequency-inverse sentence frequency (TF-IDF) of the word in the sentence, where term frequency is the number of times the word appears in the sentence, inverse sentence frequency is the logarithmically scaled inverse fraction of the number of sentence that contain the word. The weight of an edge between two nodes i and j is defined as:

$$\mathbf{E}_{i,j} = \begin{cases} PPMI\left(\boldsymbol{i},\boldsymbol{j}\right), ?\, i, j ?are\, words\, and\, i \neq j \\ TF-IDF\left(\boldsymbol{i},\boldsymbol{j}\right) ?, ?is\, document, j ?s\, word \\ 1, ? = j \\ 0, otherwise \end{cases} \quad \textbf{5}$$

Then X was feed into a GCN model which iteratively propagates messages across training and test examples. Details of the GCN model are as follows: the input of the i-th layer of the GCN is computed as the formula in $L^{(i)}$:

$$\boldsymbol{L}^{(i)}\left(\boldsymbol{A},\boldsymbol{L}^{(i-1)}\right) = \boldsymbol{\rho}\left(\boldsymbol{A}\boldsymbol{L}^{(i-1)}\boldsymbol{W}^{i}\right) \quad \textbf{6}$$

Where ρ is an activation function, A is the normalized adjacency matrix and W^i is the input weight matrix of the i— th layer of the GCN, $L^0 = X$ is the input feature matrix of the model. The output of last layer of the model was regarded as the representations of the sentences, which is fed to the softmax layer for classification:

$$\boldsymbol{Z}_{\boldsymbol{GCN}} = \boldsymbol{softmax}(\boldsymbol{g}\left(\boldsymbol{X},\boldsymbol{A}\right)) \quad \textbf{7}$$

Where g represents the GCN model. Specifically, we construct an auxiliary classifier by directly feeding sentence embeddings (denoted by X) to a dense layer with softmax activation:

$$Z_{BERT} = softmax(WX) \quad (8)$$

We use the cross entropy loss over labeled sentence nodes to jointly optimize parameters for BERT and GCN.

2.2. Dataset Preparation

In this study, we evaluate its effectiveness of the BERTGCN architecture and present results a shared benchmark datasets for text comprehension for the medical literature, **PubMed-PICO**. This dataset is the benchmark dataset from the study (https://github.com/jind11/PubMed-PICO-Detection). It was a free access database on medical articles, which was derived from PubMed. Each sentence of an abstract is annotated into one of the seven labels automatically based on the section headings in this dataset: Aim (A), Participants (P), Intervention (I), Outcome (O), Method (M), Results (R) and Conclusion (C). In this study, we only care about the performance of P/I/O labels and report, because the intervention and the control group usually appear together, the text contents of the two categories are grouped into the same category represented by I in this data set. There are 24 668 abstracts in total, each of which contains at least one of the P/I/O labels. In detail, there are 21 198 abstracts with P labels, 13 712 with I labels and 20 473 with O labels. We used the NLTK package [26] to remove stopwords from all abstracts text, the rest of the words are used to build the word sets.

2.3. Experiments settings

In this study, the training process of the model has two steps, in the first step, we use the pre-training model to fine-tune the Z_{BERT} model on the training set and obtain a fine-tuned model which was used to initialize representations the node of sentences, we experiment with different parameters for batch sizes, learning rates [1e-4, 1e-5], epochs=30. The second step is the training process of the whole combain model. The parameters of Z_{BERT} model in BERTGCN use the parameters sets which get the best performance in the first step. Studies have shown that domain corpora used for pre-training can affect the performance of downstream tasks [27]. In order to explore the effect of different pre-trained on model performance, we used three kind of pre-trained model [BERT-base-uncased, roBERT-base-uncased, bioBERT]. A two-layer GCN was used in this Z_{GCN} model, the learning rate of GCN is 1e-3. The hyper-parameter λ which was used to controls the tradeoff between the two parts of the whole model was set as[0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1]. The sets of λ was also used to explore the effect of Z_{BERT} and Z_{GCN} in the whole model.

For the final models presented bellow, we used batch size 128, trained for 10 epochs and with a learning rate of 1e-4 of Z_{BERT} . At the same time, we also tested other possible influencing factors of the model, and explored the performance of the model under different sample sizes and sample lengths. We test the performance of the model under different sample size [1000, 5000, 10000, 50000]. We test the performance of the model under different lengths range of the whole model [5-10, 10-15, 15-20, 20-25, 25-30, 30-] with a sample size of 10000.

In this study, five-fold cross-validation was used to report the performance of the model, where we divided the full dataset into 5-folds and iteratively used 1-fold as the development set, one as the test set and the rest as the train set. We report the results using the standard class-based (or micro) precision, recall and $F1$ scores:The test set was evaluated at the highest development set performance. This enables us to provide a clear view of the behavior of the classifier in each class, in addition to comparing our results to prior approaches.

3. Experimental results

We compare the performance of our BERTGCN model in two parts. Firstly we compared our model to the previously published methods for comparison include logistic regression (LR), Multilayer Perceptron (MLP), Conditional Random Field (CRF) and Bi-directional Long Short-Term Memory (Bi-LSTM)+ CRF, CPED-BioBERT, which are all from Jin and Szolovits. The CPED-BioBERT was the state-of-the-art of the PICO element detection. Second, because there is a variant in our proposed model, the pre-trained model. We explore the effect of different types of pre-training models on overall performance.Table1 summarizes the performance of our proposed model for the PubMed-PICO dataset by comparing with previous results. In this table, we use bold to identify the best overall values

for models predicting all PICO elements simultaneously. Our model, BERTGCN, achieves the best overall performance, outperforming single entity model scores in Recall for Population and Intervention entities and in Precision and F1 for Outcome entities. Our model BERTGCN improves by a large margin over all previous methods for all three P/I/O elements compared with the-state-of-art.

Table 1 Performance of the PubMed-PICO dataset in terms of precision (p), recall (r) and F_1 on the test set

Models	P elements(%)			I element(%)			O element(%)		
	P	R	F_1	P	R	F_1	P	R	F_1
LR	66.9	68.5	67.7	55.6	55.0	55.3	65.4	67.0	66.2
MLP	77.8	74.1	75.8	64.3	65.9	64.9	73.8	77.9	75.8
CRF	82.2	77.5	79.8	67.8	70.3	69.0	76.0	76.3	76.2
Bi-LSTM+CRF	87.8	83.4	85.5	72.7	81.3	76.7	81.1	85.3	83.1
CPED − BioBERT	92.8	**89.2**	91.0	84.1	85.0	84.6	88.0	89.8	88.9
BERTGCN-BioBERT	**94.1**	88.8	**91.3**	**85.5**	**86.2**	**85.8**	87.8	**92.3**	**90.0**

We compared the model performance of BERTGCN in combination with three different pre-training models, the result are as shown in follow table. The model which combine with BioBERT gets the best performance of $F1$ scores in all elements, and best performance of precision in P/O, best performance of recall in I/O. Overall, the model pre-training based on domain knowledge is more likely to achieve better predictive performance than other types of training models

Table 2 Performance of the BERTGCN using different pre-trained models

Models	P elements(%)			I element(%)			O element(%)		
	P	R	F_1	P	R	F_1	P	R	F_1
BERTGCN-BioBERT	94.1	88.8	91.3	85.5	86.2	85.8	87.8	92.3	90.0
BERTGCN-BERT	91.2	89.2	90.2	84.8	83.1	83.8	83.9	92.1	87.8
BERTGCN-RoBERTa	93.9	89.1	91.0	86.8	84.7	85.7	85.9	91.6	89.2

The performance of BERTGCN-BioBERT of under different lengths range with a sample size of 10000 are shown in flow table. The change of the macro average precision of the three elements of the length of the text are shown in the figure below. It shown that longer text length our proposed model is more likely to achieve a better text classification performance.

Table 3 Performance of BERTGCN-BioBERT with different text lengths

Sentence length	P elements(%)			I element(%)			O element(%)		
	P	R	F_1	P	R	F_1	P	R	F_1
5-10	91.8	91.7	91.6	79.0	72.2	75.5	85.2	91.4	88.2
10-15	93.3	89.4	91.3	75.1	77.5	76.3	86.8	89.4	88.1
15-20	92.5	91.0	91.7	82.0	73.8	77.7	85.3	94.0	89.4
20-25	91.7	87.9	89.8	81.0	77.4	79.2	84.4	91.3	87.7
25-30	93.0	87.7	90.3	85.9	87.3	86.6	87.5	91.6	89.5
30-	93.5	87.5	90.5	85.6	88.4	87.0	88.3	91.6	90.0

4. Discussion

In this work, we introduced a state-of-the-art evidence-based medicine research PICO element detection model BERTGCN, which takes the advantages from both large-scale pretraining models and transductive learning models. Experiments demonstrate the power of the proposed BERTGCN model. Specifically, when sentences embeddings are initialized with pre-trained BioBERT representations, our proposed model has achieved the best performance in all

three kinds of task of evidence-based medicine research key elements detection. For the PubMed-PICO dataset, our model achieves the prior state of the art, the F1-score of P/I/O in the model we proposed reached 91.3%, 85.8% and 90.0% respectively. This model is able to leverage the advantages of both kind of models: BERT, large-scale pre-trained model, takes the advantage of the massive amount of raw data, then the GCN which is a kind of transductive learning model jointly learns representations for both training data and unlabeled test by propagating label influence through graph edges. Transductive learning is a particular method for text classification which makes use of both labeled and unlabeled examples in the training process. The merits of GNNs and transductive learning are as follows: (1) the decision for an element sentence (both training and test) does not depend merely on itself, but also its neighbors. This makes the model more immune to data outliers; (2) at the training time, since the model propagates influence from supervised labels across both training and test instances through graph edges, unlabeled data also contributes to the process of representation learning, and consequently higher performances.

In this study, every potential research key elements is regarded as a node in the graph neural network, and it is embedded in the graph with the link of the words that constitute it. This means that in the process of information transmission, the longer the text length of research key elements, the more information can be obtained from other nodes. Therefore, we tested the variation of the performance of our proposed model in the task of detecting research key elements with different text lengths. As shown in Figure 3, F1 scores presents an upward trend with the increase of text length, which reflects that text length is an important factor affecting evidence detection in evidence-based medicine. Therefore, we should carry out the task of research key elements detection under the longer text length.

There are limitations in this study. We only used document statistics to build the graph, which might be sub-optimal compared to models that are able to automatically construct edges between nodes. The relationships between words can be constructed on basis of medical terminology such as MeSH [28] and UMLS [29]. It might improve the performance when the texts are insufficient.

5. Conclusions

In this study, we present a novel method that combines the graph neural network and pretrained model, BERTGCN, for the detection of medical evidence from literature. This method not only takes advantage of the text representation function of the pre-trained model, but also makes use of the global co-occurrence information of words. The evaluation on the PubMed-PICO dataset results showed that our method can successfully identified PICO elements with high accuracy performance. It is proved that the text classification model based on graph neural network can improve the recognition of evidence-based evidence. At the same time, graph neural network provides a new method for the use of evidence-based evidence, such as the fusion of evidence with knowledge graph.

The capability of these tools is such that they can be deployed in an automated approach to monitor the latest evidence-based medicine evidence, assist in system review writing and clinical practice guideline development, clinical knowledge base development, and so on.

Acknowledgements

This research is supported by Chinese Academy of Medical Sciences (Grant No. 2021-I2M-1-056).

References

[1] Sackett DL. Evidence-based medicine. Seminars in perinatology. Feb 1997;21(1):3-5. doi:10.1016/s0146-0005(97)80013-4

[2]. Cohen AM, Adams CE, Davis JM, et al. Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. presented at: Proceedings of the 1st ACM International Health Informatics Symposium; 2010; Arlington, Virginia, USA. https://doi.org/10.1145/1882992.1883046

[3]. Levay P, Heath A, Tuvey D. Efficient searching for NICE Public Health Guidelines: would using fewer sources still find the evidence? Research synthesis methods. Jun 3 2022;doi:10.1002/jrsm.1577

[4]. Goldstein A, Venker E, Weng C. Evidence appraisal: a scoping review, conceptual framework, and research agenda. J Am Med Inform Assoc. Nov 1 2017;24(6):1192-1203. doi:10.1093/jamia/ocx050

[5]. Stylianou N, Razis G, Goulis DG, Vlahavas I. EBM+: Advancing Evidence-Based Medicine via two level automatic identification of Populations, Interventions, Outcomes in medical literature. Artif Intell Med. Aug 2020;108:101949. doi:10.1016/j.artmed.2020.101949

[6]. Roumie CL, Hung AM, Russell GB, et al. Blood Pressure Control and the Association With Diabetes Mellitus Incidence: Results From SPRINT Randomized Trial. Hypertension. Feb 2020;75(2):331-338. doi:10.1161/hypertensionaha.118.12572

[7]. Hassanzadeh H, Groza T, Hunter J. Identifying scientific artefacts in biomedical literature: the Evidence Based Medicine use case. J Biomed Inform. Jun 2014;49:159-70. doi:10.1016/j.jbi.2014.02.006

[8]. Lange T, Schwarzer G, Datzmann T, Binder H. Machine learning for identifying relevant publications in updates of systematic reviews of diagnostic test studies. Research synthesis methods. Jul 2021;12(4):506-515. doi:10.1002/jrsm.1486

[9]. Wang Q, Liao J, Lapata M, Macleod M. Risk of bias assessment in preclinical literature using natural language processing. Research synthesis methods. May 2022;13(3):368-380. doi:10.1002/jrsm.1533

[10].Hassanzadeh H, Kholghi M, Nguyen A, Chu K. Clinical Document Classification Using Labeled and Unlabeled Data Across Hospitals. AMIA Annual Symposium proceedings AMIA Symposium. 2018;2018:545-554.

[11].Huang KC, Chiang IJ, Xiao F, Liao CC, Liu CC, Wong JM. PICO element detection in medical text without metadata: are first sentences enough? J Biomed Inform. Oct 2013;46(5):940-6. doi:10.1016/j.jbi.2013.07.009

[12].Boudin F, Nie JY, Bartlett JC, Grad R, Pluye P, Dawes M. Combining classifiers for robust PICO element detection. BMC Med Inform Decis Mak. May 15 2010;10:29. doi:10.1186/1472-6947-10-29

[13].Hansen M J, Rasmussen N Ø, Chung G. A method of extracting the number of trial participants from abstracts describing randomized controlled trials[J]. Journal of Telemedicine and Telecare, 2008, 14(7): 354-358. doi.org/10.1258/jtt.2008.007007

[14].Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support Evidence Based Medicine. BMC Bioinformatics. Mar 29 2011;12 Suppl 2(Suppl 2):S5. doi:10.1186/1471-2105-12-s2-s5

[15].Jin, D., & Szolovits, P. (2018, July). Pico element detection in medical text via long short-term memory neural networks. In Proceedings of the BioNLP 2018 workshop(pp. 67-75). doi.org/10.18653/v1/W18-2308

[16].Jin D, Szolovits P. Advancing PICO element detection in biomedical text via deep neural networks. Bioinformatics. Jun 15 2020;36(12):3856-3862. doi:10.1093/bioinformatics/btaa256

[17].Liu X E, You X X, Zhang X, et al. Tensor Graph Convolutional Networks for Text Classification[C]. 34th AAAI Conference on Artificial Intelligence / 32nd Innovative Applications of Artificial Intelligence Conference / 10th AAAI Symposium on Educational Advances in Artificial Intelligence, 2020: 8409-8416.

[18]. Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. In Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 7370-7377. doi.org/10.1609/aaai.v33i01.33017370

[19]. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017: 6000–10.

[20]. Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[21]. Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

[22]. Lin Y , Meng Y , Sun X , et al. BERTGCN: Transductive Text Classification by Combining GNN and BERT. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021.

[23]. Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[24]. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. Feb 15 2020;36(4):1234-1240. doi:10.1093/bioinformatics/btz682

[25]. Yuan X, Xiaoli L, Shilei L, Qinwen S, Ke L. Extracting PICO Elements From RCT Abstracts Using 1-2gram Analysis And Multitask Classification. presented at: Proceedings of the third International Conference on Medical and Health Informatics 2019; 2019; Xiamen, China. https://doi.org/10.1145/3340037.3340043

[26]. Loper E, Bird S. Nltk: The natural language toolkit[J]. arXiv preprint cs/0205028, 2002.

[27]. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676, 2019.

[28] Névéol A, Shooshan SE, Mork JG, Aronson AR. Fine-grained indexing of the biomedical literature: MeSH subheading attachment for a MEDLINE indexing tool. *AMIA Annu Symp Proc.* 2007:553–557.

[29]. Kang T, Perotte A, Tang Y, Ta C, Weng C. UMLS-based data augmentation for natural language processing of clinical research literature. J Am Med Inform Assoc. Mar 18 2021;28(4):812-823. doi:10.1093/jamia/ocaa309