A Category Detection Method for Evidence-Based Medicine

Jingyan Wang[1], Shenggen Ju[1]✉, Xi Xiong[2], Rui Zhang[1], and Ningning Liu[1]

[1] Sichuan University, Chengdu 610065, China
jsg@scu.edu.cn
[2] Chengdu University of Information Technology, Chengdu 610225, China

Abstract. Evidence-Based Medicine (EBM) gathers evidence by analyzing large databases of medical literatures and retrieving relevant clinical thematic texts. However, the abstracts of medical articles generally show the themes of clinical practice, populations, research methods and experimental results of the thesis in an unstructurized manner, rendering inefficient retrieval of medical evidence. Abstract sentences contain contextual information, and there are complex semantic and grammatical correlations between them, making its classification different from that of independent sentences. This paper proposes a category detection algorithm based on Hierarchical Multi-connected Network (HMcN), regarding the category detection of EBM as a matter of classification of sequential sentences. The algorithm contains multiple structures: (1) The underlying layer produces a sentence vector by combining the pre-trained language model with Bi-directional Long Short Term Memory Network (Bi-LSTM), and applies a multi-layered self-attention structure to the sentence vector so as to work out the internal dependencies of the sentences. (2) The upper layer uses the multi-connected Bi-LSTMs model to directly read the original input sequence to add the contextual information for the sentence vector in the abstract. (3) The top layer optimizes the tag sequence by means of the conditional random field (CRF) model. The extensive experiments on public datasets have demonstrated that the performance of the HMcN model in medical category detection is superior to that of the state-of-the-art text classification method, and the F1 value has increased by 0.4%-0.9%.

Keywords: Evidence-Based medicine · Category detection · Hierarchical multi-connected network · Self-attention · Language model.

1 Introduce

Evidence-Based Medicine (EBM) is a method of clinical practice, which obtains evidence by analyzing large databases of medical literatures such as PubMeb[3] and by retrieving relevant clinical thematic texts. EBM begins with a thesis and continues with human judgement by further extracting the evidential basis of

[3] https://www.ncbi.nlm.nih.gov/pubmeb

specific problems. The definition of clinical practice in the field of EBM usually follows the PICO principle, which is Population(P), Intervention(I), Comparison(C), Outcome(O) [1].

In order to convert articles to medical evidence, the abstracts of articles can be exploited thoroughly since abstracts are short statements without annotations or comments. The abstracts of biomedical articles generally show the themes of clinical practice in thesis research, populations, research methods and experimental results, et al.Due to the lack of effective automatic identification techniques, it has become inefficient for doctors to retrieve medical evidence. When the content of the abstract appears in a structurized form, reading the abstract can be simpler, more convenient and more efficient.

The category detection of the medical abstract can be converted into a classification task of the sentence sequence in abstracts. The sentences of the abstract contain contextual information. In addition, there are complex semantic and grammatical correlations between sentences, which makes the classification of a medical abstract different from that of its independent sentences. This paper focuses on the representation of abstract textual information and processing of sentence characteristics. The goal is to build an automatic labeling method for medical abstracts. In particular the paper proposes a Hierarchical Multi-connected Network (HMcN)-based category detection algorithm, which includes the following mechanisms: (1) The underlying layer produces a sentence vector by combining the pre-trained language model with Bi-directional Long Short Term Memory Network (Bi-LSTM), and applies a multi-layered self-attention structure to the sentence vector so as to work out the internal dependencies of the sentences. (2) The upper layer uses to multi-connected Bi-LSTMs model to directly read the original input sequence to add the contextual information for the sentence vector in the abstract. (3) The last layer optimizes the tag sequence by means of the conditional random field (CRF) model. Experiments on public datasets demonstrates that the performance of the proposed HMcN model in medical category detection is superior to that of the mainstream text classification methods, and the F1 value increases by 0.4%-0.9%. Simulation code and pre-training results can be accessed through https://github.com/pumpkinduo/HMcN. This paper contains five chapters. The second chapter introduces relevant works, the third chapter describes the category labeling method of medical articles' abstracts, the fourth chapter compares the relevant models through experiments, and the fifth chapter discusses the research results and the prospect for future work.

2 Related Works

Traditional machine learning methods used in sentence classification of clinical medical sequences mainly include naive Bayes, support vector machine [2], and conditional random field [3] and so forth. However, these methods often require a huge number of manually built features, such as features of grammar, semantics and structure, et cetera.

Deep neural text classification, mainly perform the feature extraction through convolutional neural network (CNN) and then do the classification via recurrent neural network (RNN) [4–6]. The self-attention mechanism [7] directly calculates words' dependency, and learns the internal structure of the sentence. The pre-training language model based on ELMo [8] and BERT [9] can fine-tune the generated word vector for specific tasks and achieve the best results in multiple natural language processing tasks. However, none of the above models have been directly applied to the medical domain. Jin et al [10] use deep learning for category detection task in evidence-based medicine for the first time, revealing that the deep learning model can tremendously improve the effect on the classification task of sequential sentences, but the model overlooks the connection between sentences within the abstract when generating the sentence vector.

When the existing work is used for the category detection of clinical medicine, the sentences are often classified separately, and the dependency between words and sentences is considered on the level of textual expression, which will lead to the poor effect on classification. Song et al [11] splice the entire contextual encoding of the sentence with the sentence vectors to be classified for drug classification, lacking internal reliance of the sentence. When Lee and Dernoncourt et al [12] classifing multiple rounds of dialogues, the statements in the preceding text are used for classification of the current sentence, incorporating contextual information. Bidirectional artificial neural networks (Bi-ANN) [13] are used with character information for sentence classification of biomedical abstracts and the classification results are optimized via CRFs.

3 Proposed Model

The HMcN model is comprised of three parts: single-sentence encoding, text information embedding, and tag optimization. As shown in Figure 1, each sentence in the abstract is processed by ELMo and Bi-LSTM [14] in the single-sentence encoding layer to obtain the internal semantic information of the sentence. The obtained sentence vector is fed into the text information embedding layer in units of abstract, and the dependent relationship between the sentence vectors is extracted through the multi-connected Bi-LSTMs network. Finally, the label optimization layer uses a CRF model to deal with the categories.

In this paper, lowercase letters are used to denote scalars, such as x_1;lowercase letters with arrows indicare vectors, such as $\overrightarrow{s}_1$; bold uppercase letters demonstrate matrices, such as H ; The scalar sequences such as $\{x_1, x_2, ..., x_j\}$ and the vector sequence $\{\overrightarrow{s}_1, \overrightarrow{s}_2, ..., \overrightarrow{s}_j\}$ are represented by $x_{1:j}$ and $\overrightarrow{s}_{1:j}$ respectively.

3.1 Single Sentence Encoding

Each sentence is processed differently by ELMo and Bi-LSTM to obtain a sentence vector. Then the sentence vector is used as the input. These two processing methods can be described as:

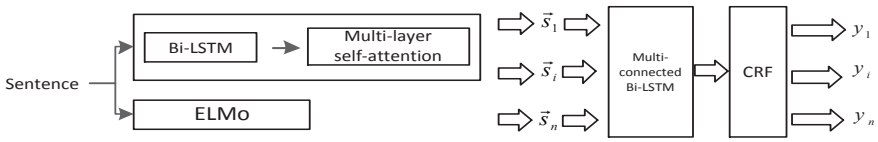

Fig. 1. Hierarchical Multi-connected Network structrue.

1) In order to address the polysemy issue, the sequence is input into ELMo, a pre-training language model. The final sentence vector $\overrightarrow{s}_i^e$ is obtained by ELMo and an average pooling layer for the sequence $\{w_1, w_2, ..., w_t\}$, where t is the length of the sentence.

2) We also use a pre-trained word vector matrix obtained by joint training of texts from Wikipedia, PubMeb and PMC [16], which contains information of medical entities. A Bi-LSTM model is then built upon the pre-trained word vectors. Using the sentence vector to calculate the self-attention value can discover the internal dependency of the sentence, and the multiple calculation of the self-attention value allows the model to learn the relevant knowledge in different subspaces. Concatenating multiple results can obtain a sentence vector $\overrightarrow{s}_i^a$:

$$\overrightarrow{a} = soft\max(\overrightarrow{v}_2 \tanh(W_1 H_s^T) \tag{1}$$

$$\overrightarrow{s}_i^a = concat(\overrightarrow{a}_1 H_s, \overrightarrow{a}_2 H_s, ..., \overrightarrow{a}_{l_{att}} H_s) \tag{2}$$

Equation (1) represents one self-attention head, where $\mathbf{H}_s^T$ represents the transpose of hidden layer vector matrix of the sentence, $\overrightarrow{v}_2 \in R^{1\times da}$, where the hyperparameter da is the self-attention hidden size, $\mathbf{W}_1 \in R^{da\times 2u}$ and u is the dimension of the hidden layer. Each obtained attention weights are multiplied by the hidden layer representation matrix, and l_{att} is the number of self-attention heads, $\overrightarrow{s}_i^e$ is the concatenation of all heads. At last, each sentence vector $\overrightarrow{s}_i$ is the concatenation of $\overrightarrow{s}_i^e$ and $\overrightarrow{s}_i^a$.

3.2 Textual information embedding

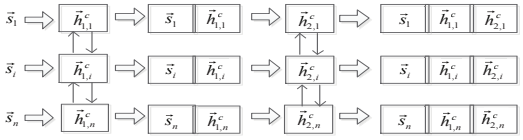

Fig. 2. Multi-connected Bi-LSTMs model.

The textual information embedding layer encodes the abstract's content to a representation vector.

The single-sentence encoding layer produces the sentence vectors S $=\{\overrightarrow{s}_1, \overrightarrow{s}_2, ..., \overrightarrow{s}_n\}$ for n independent sentence in a given abstract,S is the used as the input to the multi-connected Bi-LSTMs. The multi-connected Bi-LSTMs module in HMcN is built on the basis of DC-Bi-LSTM architecture [17]. The structure is shown in Figure 2, the input of all the layers is the concatenation of the output of the previous layers to form a multi-connected Bi-LSTMs network. It outputs a series of new sentence encoding vectors, which contain contextual information. The output of last muti-connected Bi-LSTM layer is averaged out through an average pooling layer. The above processing method can be represented by equation (3)-(4):

$$\overrightarrow{h}^c_{l,i} = lstm(\overrightarrow{h}^c_{l,i-1}, M^c_{l-1,i}), \overleftarrow{h}^c_{l,i} = lstm(\overleftarrow{h}^c_{l,i+1}, M^c_{l-1,i}) \tag{3}$$

$$M^c_{l-1,i} = concat(h^c_{0,i}, h^c_{1,i}, ..., h^c_{l-1,i}), h^c_{0,i} = s_i \tag{4}$$

In the equation (4) $M^c_{l-1,i}$ is the concatenation of the vector representation $h^c_{l,i}$, which is obtained by concatenating the forward hidden layer vector $\overrightarrow{h}^c_{l,i}$ and the reverse hidden layer vector $\overleftarrow{h}^c_{l,i}$ in equation (3). These vectors are input into a single-layer feed forward neural network, and each sentence vector $\overrightarrow{r}_i \in R^d$ output represents the probability that the sentence belongs to each label, where d is the number of labels.

Compared with the traditional Recurrent Neural Networks (RNNs) or deep RNNs, for each RNN layer, the multi-connected Bi-LSTMs network can directly read the original input sequence, namely the ELMo and Bi-LSTM encoded sentence vectors in this paper's technique, which doesn't need to pass all the useful information through the network. This paper employs very few numbers of network neurons to avoid excessive complexity of the module.

3.3 Tag optimization

The CRF model can improve the performance of sentence sequence classification. The sentence to be classified and the sentence label respectively serve as the observation sequence and the state sequence of the CRF model. The labeling probability of a given sentence is acquired by the sentence related feature extracted by the lower layer network.

Suppose that the sentence vector sequence $\overrightarrow{r}_{1:n}$ output by texual information embedding layer is known. This layer outputs a tag sequence $y_{1:n}$, where y_i represents the prediction tag assigned to the i-th sentence. $\mathbf{T}$[i:j] is defined as the probability with the sentence with the label i which is followed by the sentence with the label j. The score of $y_{1:n}$ is defined as the sum of the predicted probability of the label and the transition probability [13]:

$$score(y_{1:n}) = \sum_{i=1}^{n} r_i\left[y_i\right] + \sum_{i=2}^{n} T\left[y_{i-1}, y_i\right] \tag{5}$$

The correct tag sequence probability can be acquired by the *softmax* function, and the tag sequence earning the highest score through the Viterbi algorithm serves as the predicted outcome.

4 Experiments

4.1 Experimental settings

Datasets In order to quantitatively analyze the detection performance of the HMcN model on the sentence category detection in the medical abstract, we perform classification experiments on two standard medical abstract datasets. The datasets are described separately as follows:

NICTA-PIBOSO dataset [19] (*NP* dataset): This dataset is shared on the ALTA 2012 Shared Task, and its main purpose is to apply the biomedical abstract sentence classification task to evidence-based medicine. The label include "Population", "Intervention", "Outcome", "Study Design", "Background", and "Other".

PubMeb 20k RCT dataset [20] (*PubMeb* dataset): The data is derived from PubMeb-the largest database of biomedical articles. The class labels include "Objectives", "Background", "Methods", "Results" and "Conclusions".

The specific information of the dataset is shown in Table 1:

Table 1. Statistics of experimental dataset.

Dataset	$\|C\|$	$\|V\|$	Train	Validation	Test
NICTA-PIBOSO	6	17k	720(8k)	80(0.9k)	80(2k)
PubMeb 20k PCT	5	68k	15k(195k)	2.5k(20k)	2.5k(19k)

In Table 1, $|C|$ and $|V|$represent the total number of class labels and the vocabulary size respectively. For training datasets, validation datasets, and test datasets, the numbers outside the parentheses show the number of abstracts, and the numbers in parentheses indicate the number of sentences. Each abstracted sentence has merely one unique label.

Parameter settings The sentence vector is obtained using the open source pre-training model ELMo, and the hidden layer dimension of sentence vector is 1024. The parameters including the Bi-LSTM network and the multi-layer self-attention module are updated by Adam [21]. At each level, Dropout [22] is used to solve the overfitting problem, and L2 regularization [23] is utilized to further narrow the gap between the results of training dataset and validation dataset. The parameter settings are as follow: the self-attention hidden size da is set to 150, single sentence encoding layer hidden size u is set to 150 and 200, multi-connected Bi-LSTMs last layer dimension u_l^a as 50 and 100, multi-connected Bi-LSTMs other layer dimension u_o^a is set to 13, the number of tags R^d is set to

6 and 5, the number of the multi-connected Bi-LSTMs layer l is set to 6, learning rate lr is set to 0.001, dropout do is set to 0.5, batch size bz is set to 30 and 40, and the number of the multi-layer self-attention layer l_{att} is set to 3.

Comparison algorithm LR [13]: Logistic regression classifier, which utilizes the n-gram feature extracted from the current sentence without using any information from surrounding sentences. CRF [3]: The conditional random field classifier, as the input of the classification sentence vector, each output variable corresponds to the label of a sentence, and the sentence sequence considered by the CRF is the entire abstract. Therefore, when classifying the current sentence, the CRF baseline uses the preceding and following sentences at the same time. Best Published: A method proposed by Lui in 2012 [24], based on a variety of feature sets, introduces feature stacking and performs best on NP dataset. Bi-ANN: An annotated model proposed by Dernoncourt et al.in 2017 [13] which optimizes classification results by CRF and character vectors.

4.2 Experimental results

Comparison of the entire results The experimental results are measured by Precision, Recall and F1 values. The experimental results are shown in Table 2. As displayed in Table 2, the F1 value of the HMcN model increases by 0.4%-

Table 2. Main results.

Model	NICTA-PIBOSO			PubMeb 20k RCT		
	Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
LR	73.8	69.5	71.6	82.7	82.5	82.6
CRF	83.0	79.5	80.0	86.1	84.5	85.3
Best Published	-	-	82.0	-	-	-
Bi-ANN	-	-	82.7	-	-	90.0
HMcN	82.4	83.8	83.1	91.2	91.0	90.9

8.3% respectively compared with the other models. The LR method performs better on the PubMed dataset than on the NP dataset, which reveals that the dependencies between the tags in the NP dataset are closer. The indicators of HMcN model are all superior to the CRF model, demonstrating that the model optimizes sentence-level features. HMcN outperforms the Best Published method in the NP dataset, indicating that the HMcN model can acquire deeper feature information. HMcN model is better than that Bi-ANN, which shows that HMcN incorporates multi-granularity information of words, sentences and paragraphs for textual representation, taking note of the internal dependence of the sentence while sentences are being encoded, which helps optimize the category detection results.

Comparison of single-label predicted effects Table 3 and Table 4 respectively demonstrate the confusion matrix [25] and predicted effects [26, 27] while running single-label prediction on the PubMeb dataset. The columns in Table 3 reveal real tags and the rows represent predicted tags. For instance, 476 sentences labeled as "Background" are predicted as "Objectives". It can be told that differentiating between "Background" and "Objectives" tags is the most tremendous problem the classifier encountered. The main reason is that there is confusion in "Background" as well as "Objectives" per se, furthermore, when the sentences tagged as "Objectives" tags are compared with those of other categories in the abstract, their semantics and characteristics are not obvious.

Table 3. Confusion matrix.

	Background	Conclusions	Methods	Objectives	Results
Background	2964	30	147	476	0
Conclusions	2	2437	25	0	190
Methods	39	13	5580	21	110
Objectives	600	0	47	706	0
Results	0	60	244	0	5432

Table 4. Single category detection results.

Label	Precision(%)	Recall(%)	F1(%)	Count
Background	69.4	86.7	77.1	3627
Conclusions	97.1	91.8	94.4	2654
Methods	93.9	97.1	95.5	5744
Objectives	79.9	52.8	63.1	1353
Results	94.8	94.7	94.7	5736
Total	91.2	91.0	90.9	19114

Comparison of ablation experiments In order to verify the effect of each step in the model, HMcN-multiLSTM, HMcN-attention, HMcN-multiattention, HMcN-ELMo, and HMcN-CRF, which respectively represent the ablation model of removing the multi-connected Bi-LSTMs architecture, removing the multi-layer self-attention mechanism, replacing multi-layer self-attention with single-layer self-attention, removing ELMo and removing the CRF layer. Table 5 demonstrates the experimental results on the PubMeb dataset. It can be seen that each module is conducive to the effect of category detection, and the multi-connected Bi-LSTMs architecture is the most important component of the HMcN model.

In order to verify that the multi-connected LSTMs model can achieve better results with less parameter quantity in comparison with the ordinary LSTMs,

Table 5. Ablation study of HMcN.

Model	F1(%)
HMcN-multiLSTM	87.5
HMcN-attention	90.3
HMcN-multiattention	90.6
HMcN-ELMo	90.0
HMcN-CRF	89.1
Full	90.9

Table 6. Parameter quantity of multi-connected Bi-LSTMs and the ordinary LSTMs.

l	u_l^a	u_o^a	Parameter	F1(%)
2	200	5	$1.40*10^6$	88.9
6	100	13	$1.31*10^6$	90.9

this paper carries out an analytical and comparative experiment on the parameter size. In Table 6 the first row is ordinary LSTM, and the second one is our model.Compared with the second model, the increase on PubMeb dataset are 1%, with the parameters decreased.

5 Conclusion

This paper constructs a hierarchical multi-connected network model for abstract's category detection in evidence-based medicine. The model uses the multi-connected Bi-LSTMs network to better capture complete dependencies and contextual information between sentences. Combined with a multi-layer self-attention mechanism, this model promotes the overall quality of sentence encoding, and achieves good results in the public datasets of medical abstracts. For further study, the HMcN model can be applied to tackle specific problems relevant to evidence-based medicine, such as the exploration of medical texts, document retrieval and so forth, to achieve the goal of assisting in medical treatment.

Acknowledgements

The work was partially supported by the China Postdoctoral Science Foundation under Grant No. 2019M653400; the Sichuan Science and Technology Program under Grant Nos. 2018GZ0253, 2019YFS0236, 2018GZ0182, 2018GZ0093 and 2018GZDZX0039.

References

1. Richardson, W.S.,Wilson, M.C., Nishikawa, J., Hayward, R.S.: The well-built clinical question: a key to evidence-based decisions. ACP journal, **123**(3), A12 (1995)
2. Wang, S., Manning, C D.: Baselines and bigrams: simple, good sentiment and topic classification. ACL: ACM, 90-94 (2012)
3. Hassanzadeh, et al.:Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. J of biomedical informatics. 49,159-170 (2014)

4. Kim, Y.: Convolutional neural networks for sentence classification. In: EMNLP (2014)
5. Conneau, A., Schwenk, H., Barrault, L., Lecun, Y.: Very deep convolutional networks for text classification. In: EACL, Volume 1,Long Papers, pp. 1107–1116 (2017)
6. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: AAAI,volume 333, pp. 2267–2273 (2015)
7. Lin, Z., Feng, M., Santos, C.N.D, Mo Yu, Xiang, B., Zhou B., Bengio, Y.:A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130 (2017)
8. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: NAACL (2018)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805 (2018)
10. Jin, D., Szolovits, P.: Pico element detection in medical text via long short-term memory neural networks. In: BioNLP, pp. 67–75 (2018)
11. Song, X., Petra, J.,Roberts,A.:A Deep Neural Network Sentence Level Classification Method with Context Information.arXiv:1809.00934v,2018.(2018)
12. Lee, J.Y., Dernoncourt, F.: Sequential short-text classification with recurrent and convolutional neural networks. arXiv preprint arXiv:1603.03827.(2016)
13. Dernoncourt, F., Lee, J.Y., Szolovits, P.: Neural networks for joint sentence classification in medical paper abstracts. In: EACL, 2, 694–700 (2017)
14. Graves A, Schmidhuber J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw, **18**(5), 602–610 (2005)
15. Gu, J., Lu, Z.,Li H: Incorporating copying mechanism in sequence-to-sequence learning. arXiv preprint arXiv:1603.06393 (2016)
16. Moen, S., Ananiadou, T.S.S.: Distributional semantics resources for biomedical text processing. In: LBM, pp. 39–43. Tokyo,Japan (2013)
17. Ding Z, Xia R, Yu J, et al.: Densely Connected Bidirectional LSTM with Applications to Sentence Classification. In: NLPCC (2018)
18. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: EMNLP, pp. 1532–1543 (2014)
19. Amini, I., Martinez, D., Molla, D., et al.:Overview of the alta 2012 shared task (2012)
20. Dernoncourt, F. Lee, J.Y.: Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. arXiv preprint arXiv:1710.06071 (2017)
21. Kingma, D.P., JimmyBa, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Srivastava, N., et al.: Dropout: A simple way to prevent neural networks from overfifitting. The Journal of Machine Learning Research, **15**(1), 1929–1958 (2014)
23. Ma, X., Gao, Y., Hu, Z., Yu, Y., Deng, Y., Hovy, E.: Dropout with expectation-linear regularization. arXiv preprint arXiv:1609.08017 (2016)
24. Liu, M.: Feature Stacking for Sentence Classification in Evidence-Based Medicine. Australasian Language Technology Association Workshop, pp. 134–138.(2012)
25. Xiong, X., et al.: ADPDF: A Hybrid Attribute Discrimination Method for Psychometric Data with Fuzziness. IEEE Transactions on SMC: Systems: 49,265-278.(2019)
26. Xiong, X.,Li, Y.,Qiao, S.: An Emotional Contagion Model for Heterogeneous Social Media with Multiple Behaviors. Physica A: 490,185-202.(2018)
27. Xiong, X., et al.: Affective Impression: Sentiment-awareness POI Suggestion via Embedding in Heterogeneous LBSNs. IEEE T on Affective Computing: 1-1.(2019)