Levergeing Character embedding to Chinese Textual Entailment Recognition Based on Attention[1]

Pengcheng Liu, Lingling Mu, Hongying Zan
College of information Engineering, Zhengzhou University,
Zhengzhou, China
E-mail: liupengcheng2016@163.com; iellmu@zzu.edu.cn

Abstract—**Textual Entailment Recognition is a common task in the field of Natural Language Processing，which aims to find the semantic inference relationship between two sentences. In this paper we proposed an attention based method for the task of Chinese textual entailment recognition, this method used a common framework which contains of three parts: encoding layer, attention layer, prediction layer. After training and fine-tuning, it reaches 81.52% in the test set of CCL2018 textual entailment task, which outperformed the benchmark models of the previous evaluation. Experiments also show that character embedding and CNN is useful for Chinese textual entailment recognition.**

Keywords- textual entailment ; attention mechanism; word embedding; character embedding

I. INTRODUCTION

Recognizing Textual Entailment (RTE) is also known as Natural Language Inference (NLI), refers to determine whether there is a one-way semantic inferential relationship between two sentences [1]. As an important task in the field of natural language processing, textual entailment recognition is the basic of machine translation and reading comprehension. Its characteristic lies in that the reasoning relation between two sentences is one-way, and the sentences often have no too long context information, therefore, inference only according to the semantic knowledge and expression of the sentence itself.

Textual entailment recognition is generally considered as a classification problem: each sample consists of two sentences, the Premise and Hypothesis, which require a model to determine the relational categories of entailment, contradiction, and neutral. The Seventeenth China National Conference on Computational Linguistics (CCL 2018) first appeared in the Chinese natural language inference task[2], and the evaluation dataset is currently the largest Chinese textual entailment recognition dataset.

On the basis of decomposable attention model, this paper proposed a combination neural network model of attention mechanism and character embedding. Compared with English, Chinese word segmentation may have some errors, and Chinese characters contain a lot of semantic information. This model can be divided into encoding layer, attention layer, compare layer and prediction layer. Main contributions of the method are: (a), The neural network of attention and compare layer uses fully connected neural network instead of simple multi-layer perceptron; (b), The model used char embedding and CNN encoder to improve performance. After training and adjustment, the attention method achieved 81.52% accuracy in the test set.

II. RELATED WORK

In recent years, the research on textual entailment recognition has developed rapidly, which benefits from the construction of various international evaluation tasks and large-scale corpus. At present, the most authoritative textual entailment corpus is Stanford Natural Language Inference Corpus (SNLI) [2]. SNLI has 570,000 pairs of sentences. It is constructed by describing the same picture to different people. In addition to the three categories of entailment, neutral and contradiction, there are still some undetermined categories. Similarly, there are MultiNLI dataset proposed by Williams *et al.* [3], SciTail proposed by Khot *et al.* [4]. CCL2018 Chinese textual entailment recognition dataset is translated from English datasets such as SNLI, MultiNLI and SciTail, etc. Machine translation is the main part of the construction process, and manual translation is the supplement. Due to the linguistic differences between Chinese and English, some samples have changed their categories in the process of translation.

There are some models which often used as benchmark models in textual entailment recognition. Such as Decomp-Att (Decomposable Attention model)[5] and ESIM (Enhanced LSTM Sentence Inference Model)[6]. Decomp-Att is mainly based on attention alignment mechanism, which combines alignment-based textual entailment recognition with attention mechanism. ESIM consists of two parts. One part uses sequence model to collect context information of words, the other part uses tree model to collect clause information, which improves the accuracy of SNLI to above 88% for the first time. In addition, the SWEMs model proposed by Shen et al. [7] is based on maximum and average pooling of word vectors. It uses only multi-layer perceptron, but does not use more complex neural network structure. It is evaluated on 17 datasets including SNLI and MultiNLI, and most of them achieve the best results. Similar to SWEMs, Tao Shen *et al.* [8] proposed a model DiSAN, which does not use RNN and CNN. It uses three attention matrices to extract features based on self-attention mechanism, and achieves the best accuracy on MultiNLI. Recently, in the field of natural language processing, a new pre-training language model has been proposed, such as EMLo, proposed by Matthew *et al.*[9], which combines context representation word vectors, Experiments on six tasks, including natural language inference, have yielded optimal results. At present, the best

[1] The authors were supported financially by the National Social Science Fund of China (18ZDA315), Programs for Science and Technology Development in Henan province (No.192102210260) and the Key Scientific Research Program of Higher Education of Henan (No.20A520038).

[2] http://www.cips-cl.org/static/CCL2018/call-evaluation.html

result on SNLI is the XLNet model based on attention mechanism and pre-train language model proposed by Yang *et al.* [10].

Chinese textual entailment recognition research started relatively late because of the lack of large-scale corpus corresponding to the language. Wang *et al.* [11] used textual entailment, sentence similarity and other methods to implement experiments on the data of college entrance examination choice questions, which proved the validity of the textual entailment method in answer selection. Based on reading comprehension text, Chen Qian *et al.* [12] proposed a Chinese textual entailment recognition method based on hierarchical LSTM. On the Chinese corpus of RITE 2014, Tan *et al.* [13] proposed a textual entailment recognition method combining CNN and Bi-LSTM.

The above models have been widely used in the English textual entailment recognition, while the Chinese textual entailment has few applications. Based on the attention model Decomp-Att and ESIM, and inspired by the advantages of the neural model SWEM, this paper proposed a Chinese textual entailment model based on attention. The model is characterized by the combination of the attention and character embedding, and the structure of the deep neural network model is simple. The experiment demonstrate that the pooling mechanism functions well ul in the Chinese textual entailment, and 81.52% is obtained on the CCL2018 dataset.

III. OUR MODEL

The proposed model's structure is shown in Figure 1. From bottom to top are the encoding layer, the co-attention layer, the compare layer and the prediction layer. In the Figure, **P** and **H** represent the premise sentence and the hypothesis sentence respectively. *Dense* stands for fully connected neural network, *Duplicate* stands for copy operation, and ⊕ stands for multiple vectors. The splicing operation, ⊙ represents the point multiplication operation of two vectors.

3.1. Encoding Layer

The purpose of the encoding layer is to encode the premise sentence and the hypothetical sentence respectively. The premise **P** and the hypothesis **H** are converted into a sequence of word embedding $\mathbf{p}=(p_1, p_2,\ldots p_i,\ldots p_m)$, $\mathbf{h}=(h_1, h_2,\ldots h_j,\ldots h_n)$,where m and n are the lengths of **p** and **h** respectively. Where p_i is the vector representation of the ith word in **p**, h_j is the vector representation of the jth word in **h**, p_i, $h_j \in \boldsymbol{R^d}$, i∈[1,m]， j∈[1,n], d is the dimension of the word embedding. After obtaining word embedding sequences p and h, the features are extracted by a multi-layer perceptron fully connected neural network to obtain $\overline{\mathbf{p}}$ and $\overline{\mathbf{h}}$.

3.2. Attention Layer

The purpose of attention layer is to obtain the interactive information of the sentence through the calculation of the premise and hypothesis attention[14]. Firstly, the words in the sentence are aligned, and then the attention weights of one sentence are calculated respectively. The alignment is to use the words or sub-sentence in the $\overline{\mathbf{p}}$, $\overline{\mathbf{h}}$ sequences as

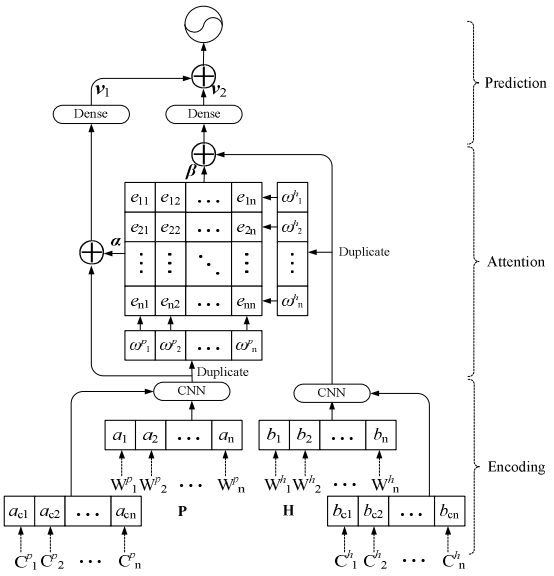

Figure 1. Attention model structure.

the rows and columns of a matrix, respectively. The matrix as in equation (3), it is an m×n matrix $\boldsymbol{E}_{mn}$, its elements is $\boldsymbol{e}_{ij} \in \mathbf{R^{m\times n}}$. Next, we perform attention through $\boldsymbol{e}_{ij}$ to calculate the attention weight β_i of $\overline{\mathbf{h}}$ relative to $\overline{\mathbf{p}}$, and the attention weight α_j of $\overline{\mathbf{p}}$ relative to $\overline{\mathbf{h}}$. Finally we obtain the closer part of the relationship between words or sub-sentences. Due to the use of the attention mechanism, more interactive information is obtained relative to the self-attention calculation of the sentence itself, which is helpful for the recognition of the inference relationship. The formal representation of the process for calculating the weight matrix is as shown in equations (1) to (3).

$$\mathrm{E_{mn}} = \overline{\mathbf{p}}^T\ \overline{\mathbf{h}} \tag{1}$$

$$\beta_i = \sum_1^n \frac{\exp(\boldsymbol{e_{ij}})}{\sum_{k=1}^j \exp(\boldsymbol{e_{ik}})} h_j \tag{2}$$

$$\alpha_j = \sum_1^m \frac{\exp(\boldsymbol{e_{ij}})}{\sum_{k=1}^j \exp(\boldsymbol{e_{jk}})} p_i \tag{3}$$

Where β_i represents the $\overline{\mathbf{h}}$ clause and $\overline{\mathbf{p}}$ aligned attention weight; α_j represents the $\overline{\mathbf{p}}$ clause and $\overline{\mathbf{h}}$ aligned attention weight.

3.3. Compare Layer

The purpose of compare layer is to aggregate the feature obtained by the attention layer with the feature of the encoding layer to obtain the relationship between the two sentences. The vectors p_i, h_j in the two sequences $\overline{\mathbf{p}}$, $\overline{\mathbf{h}}$ and the attention weights β_i and α_j obtained in the attention layer are respectively connected, subtracted and multiplied. Finally, this layer realize aggregation and compare information of the encoding layer and the result of the comparison of the attention weight sequence and the text sequence. The pair of vectors p_i and β_i, $h_j$and α_j are compared by a full connected network layer G, and the weight vectors $\boldsymbol{v_{1,i}}$ and $\boldsymbol{v_{2,j}}$ corresponding to the sub-

sentences of each sequence are obtained. The formal representation of the comparison operation is as shown in equations (4) to (5).

$$\boldsymbol{v_{1,i}} = \mathrm{G}([\beta_i , a_i]) \tag{4}$$

$$\boldsymbol{v_{2,j}} = \mathrm{G}([\alpha_j , b_j]) \tag{5}$$

Where G represents the operation of full connected neural network; [] represents the splicing operation of vectors; − represents the absolute value of the difference; • represents the dot product between vectors.

3.4. Prediction Layer

The purpose of prediction layer is to further extract and classify the features obtained by the compare layer. That is, the weight vectors $\boldsymbol{v_{1,i}}$ and $\boldsymbol{v_{2,j}}$ of the words are first aggregated into weight vectors $\boldsymbol{v_1}$, $\boldsymbol{v_2}$ representing the entire sentence, respectively. Then $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ are connected, and a fully connected neural network H is used for classification to obtain a vector $\mathbf{v} \in \boldsymbol{R^3}$, and finally $\mathbf{v}$ is output to the final label $\tilde{l}$ by argmax function. The formal representation of the prediction layer is as shown in equations (6) to (9).

$$\boldsymbol{v_1} = \sum_1^m \boldsymbol{v_{1,i}} \tag{6}$$

$$\boldsymbol{v_2} = \sum_1^n \boldsymbol{v_{2,j}} \tag{7}$$

$$\mathbf{v} = \mathbf{H}([\boldsymbol{v_1}; \boldsymbol{v_2}]) \tag{8}$$

$$\tilde{\boldsymbol{l}} = \mathbf{argmax}(\mathbf{v}) \tag{9}$$

Where $\boldsymbol{v_3}$ and $\boldsymbol{v_4}$ represent the sequence obtained by premise and hypothesis, iϵ[1, m], j$\in$[1,n]. H represents a classifier consisting of fully connected networks; [] represents the connection of vector; $\tilde{l}$ represents the predicted label.

IV. EXPERIMENT

4.1. Dataset

We using the CCL2018 Chinese Natural Language Inference evaluation dataset. The training set is 90000, the development set is 10,000, and the test set is 10,000. The statistical results of the categories are shown in Table 1. And some samples are shown in Table 2. It can be seen that on each dataset, the three categories of entailment, contradiction, neutral are evenly distributed.

The following Equation(10) serves to evaluate the accuracy.

$$Acc = \frac{\tilde{l}_{correct}}{l_{ture}} \tag{10}$$

Where $\tilde{l}_{correct}$ represents the number of labels that are correctly classified by all categories; l_{true} is the number of true labels of raw dataset.

TABLE I. CATEGORY STATISTICS OF DATASETS

	Neutral	Entailment	Contradiction	Total
Train	31325	29738	28937	90000
Dev	3098	3485	3417	10000
Test	3182	3475	3343	10000

TABLE II. TEXTUAL ENTAILMENT DATASET SAMPLES

Premise	Hypothesis	Label
孩子从秋千上倒挂下来。 (The child hung upside down from the swing)	孩子坐直，用脚推地。 (The child sat upright and pushed the floor with his feet.)	Contradiction
两名女子走在拥挤的街道上。 (Two women walk in the crowded street)	两名女子走在街上。 (Two women walking in the street)	Entailment
楼梯上挤满了人。 (The stairs were crowded with people.)	人们在大楼里面。 (People are in the building.)	Neutral

4.2. Experimental Setting

In the process of data preprocessing, we use the jieba [3] word segmentation tool, and use Chinese word embedding obtained from the People's Daily and other corpuses[15]. Word embedding dimension is 300.

Experiment uses the Tensorflow deep learning framework, batchsize is 32, MLP hidden layer nodes are set to 300, learning rate is 0.0001, and dropout rate is 0.3. Optimization function uses the Adam function, and loss function uses the cross entropy function. In addition, early stop was used to prevent over-fitting. Label $l \in \boldsymbol{R^3}$, corresponds to three categories: neutral N, entailment E, contradiction C. The model is tested on test set after training on validation set for best performance.

4.3. Experiments Results

Experiment uses Decomp-Att as the baseline model, using the Model proposed in this paper and other attention models ESIM, Decomp_Att, and the first place LSTM+CNN model of the evaluation task [4], which is current best accuracy 82.38% for evaluation. The experimental results are shown in Table 3. In the table, N, E, C represent the accuracy rates in Neural, Entailment, Contradiction categories respectively. The reason that LSTM+CNN without detail categories accuracy is it was the first place in the evaluation.

TABLE III. MODEL COMPARISON RESULTS OF CCL2018 DATASET

Models	Train (%)	Test (%)
ESIM	76.91	72.22
Decomp_Att	78.64	78.18
LSTM + CNN	-	82.38
BiLSTM	85.73	78.33
Ours	86.44	81.52
Ours-Att	75.22	69.26
Ours-char	84.96	77.71
Ours-CNN	83.23	79.12
Ours-CNN-char	80.63	74.38

[3] https://github.com/fxsjy/jieba
[4] https://competitions.codalab.org/competitions/19911

The LSTM + CNN [16] method is the best public achievement on the test set in 2018, reaching 82.38%. ESIM is the baseline model of evaluation, with an accuracy of 72.22%. Decomp-Att is a decomposable attention model with an accuracy of 78.18%. Our model achieved 81.52% after combined character embedding and CNN encoder and outperformed the baseline models Decomp-Att and ESIM.

On the basis of training set and development set, the hyper-parameters of our model is experimented to determine the hyper-parameters on the test set. For example, the dropout rate of super parameters is used in all fully connected neural network layers in the model, it refers to a part of the neural network nodes are randomly lost to prevent over-learning of training data. Figure 2 shows the effect of different dropout rates on the accuracy as shown.

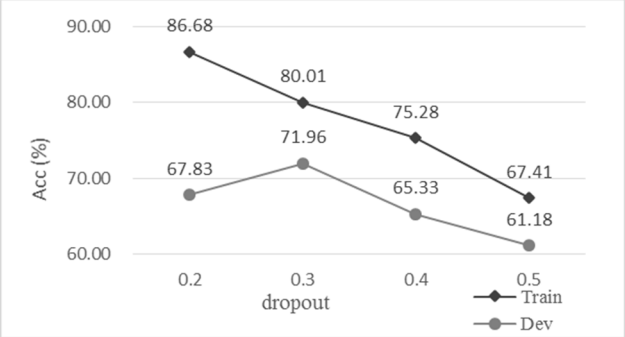

Figure 2. Effect of dropout on Accuracy.

It can be seen that in the case of lower dropout rate, the accuracy rate is higher in the training set, while the model on the test set tends to be over-fitting; when the dropout rate is higher, the accuracy rate on both datasets decreases, and the model produces under-fitting. Therefore, according to the experimental results on the development set, the dropout rate used in the test set is 0.3.

4.4. Analysis

In this paper, the instances of classification errors in the test set of the model are counted. The results are shown in Table 4, which N denotes the error rate of neutral types, E denotes the error rate of entailment categories, and C denotes the error rate of contradiction types.

Table 4 is the confusion matrix of the specific classification results on the test set. From Table 4, it can be seen that the identification of contradiction relations is the most difficult, and the error percentage of contradiction relations reaches 44.01%. The error rates of neutral and entailment are lower. Relatively speaking, the accuracy rate of recognizing entailment relation is higher than that of

TABLE IV. TEST SET CONFUSION MATRIX

		Prediction			Total
		N	E	C	
True	N	2535	465	182	3182
	E	409	2897	169	3475
	C	470	193	2680	3343

TABLE V. ERROR SAMPLES

Sentences	True	Prediction
P：滑板运动员在电线杆前做空中跳跃。 (Skateboarders jump in the air in front of the pole.) H：一个人在玩电脑。 (A man is playing with a computer.)	N	E
P：红线是倾斜的。 (The red line is tilted.) H：黑色线条曲折。 (The black lines are tortuous.)	C	E
P：一个女孩骑在秋千上。 (A girl was riding on a swing.) H：一位女性正坐在外面。 (A woman was sitting outside.)	E	C

neutral relation, because entailment relation is a one-way reasoning inference relation, but neutral and contradiction is two-way inference relationship.

Table 5 shown some error samples of our model for analysis. From example, in the first pair of error samples, neutral is misclassified as entailment. It can be seen that the similarity of function words or verbs such as “在(in)” and “跳跃(jump)” leads to errors in recognition, and further experiments can be carried out with the combination of function word relations or the removal of function words. The second and third pairs of examples are that contradiction are misclassified into entailment and entailment into contradiction, respectively. The reason is that the similarities between noun vectors such as “线(line)”, “红色(red)”, “黑色(black)”, “一个(A)” and “女孩(girl)” are high, and the attention weight between two sentences is high, which leads it more difficult for the model to distinguish contradictions from entailments. The recognition of contradiction relation is also a difficult point in textual entailment recognition [17], which accords with the above analysis of confusion matrix.

Experimental results shown that the pooling operation is important for reducing error in textual entailment recognition. We analyzed the effect of pooling operation in our model by counted the number of samples of removing the pooling operation model and all the error instances on the original model, as shown in Figure 3. There are six categories of errors, in which N2E represents error that the neutral category be predicted as entailment incorrectly, and so on.

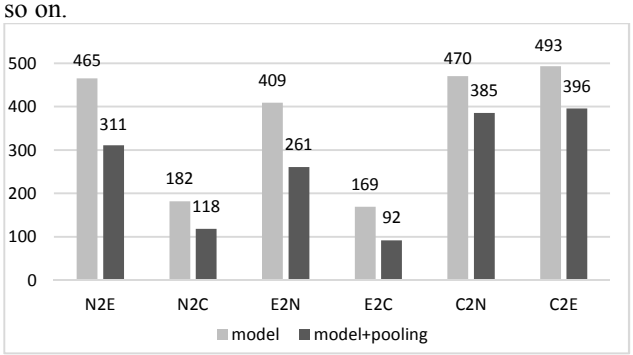

Figure 3. Error prediction numbers of models.

As can be seen from Figure 3, the number of error samples for each category decreases in the model with pooling operation. The pooling operation of the decoding part of the model can further extract text features, so the accuracy has been improved to a certain extent, but the contradiction categories (C2N, C2E) are still the category with the largest number of error samples. The obvious improvement effect is the error categories of N2E and E2N, which accords with the assumption that one-way entailment relationship is easier to judge. The N2C category has the worst improvement effect, but it is one of the lowest error samples, so the improvement effect on this category is not obvious.

V. CONCLUSION

This paper proposed a method based on attention mechanism for Chinese textual entailment recognition task. In this method, attention alignment operation is used to extract features of sentence pairs and calculate attention weights between them. Benefit from combining the character embedding and CNN encoder, the model's ability that recognizing inference relation of two sentences be improved. The accuracy of our model is 81.52%, higher than the benchmark models.

At present, the method still has a high recognition error rate for contradiction categories. The next step can focus on improving the classification effect of contradiction relations, such as adding lexical features or syntactic dependencies, and further improving the model in order to solve the task of Chinese textual entailment recognition better.

REFERENCES

[1] Bos J, Markert K. Recognising textual entailment with logical inference[C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005: 628-635.

[2] Bowman, S. R., Angeli, G., Potts, C.: A large annotated corpus for learning natural language inference. Computer Science (2015).

[3] Williams A, Nangia N, Bowman S R. A broad-coverage challenge corpus for sentence understanding through inference[J]. arXiv preprint arXiv:1704.05426, 2017.

[4] Khot T, Sabharwal A, Clark P. SciTail: A textual entailment dataset from science question answering[C]//Proceedings of AAAI. 2018.

[5] Parikh A P, Täckström O, Das D, et al. A Decomposable Attention Model for Natural Language Inference [J]. 2016:2249-2255.

[6] Chen Q, Zhu X, Ling Z, et al. Enhanced LSTM for Natural Language Inference [J]. 2016:1657-1668.

[7] Shen D , Wang G , Wang W , et al. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms[J]. 2018.

[8] Shen T, Zhou T, Long G, et al. Disan: Directional self-attention network for rnn/cnn-free language understanding[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[9] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. 2018.

[10] Yang, Zhilin, et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." arXiv preprint arXiv:1906.08237 (2019).

[11] Wang B, Zheng D , Wang X, Zhao S, Zhao T. Multiple-choice Question Answering Based on Textual Entailment [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1):134-140.

[12] Chen, Q., Chen, X., Guo, X.: Multiple-to-One Chinese Textual Entailment for Reading Comprehension. Journal of Chinese information processing, 32(4): 87-94 (2018).

[13] Tan, Y., Liu, Z., Lv, X.: CNN and BiLSTM Based Chinese Textual Entailment Recognition. Journal of Chinese information processing, 32(7): 11-19 (2018).

[14] Guo M, Zhang Y, Liu T. Research Advances and Prospect of Recognizing Textual Entailment and Knowledge Acquisition[J].Chinese Journal of Computers, 2017, 40(4):889-910.

[15] Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). Analogical Reasoning on Chinese Morphological and Semantic Relations. arXiv preprint arXiv:1805.06504

[16] Cheng, J., Dong, L., & Lapata, M. (2016). Long short-term memory-networks for machine reading. arXiv preprint arXiv:1601.06733.

[17] MacCartney B, Manning C D. Natural language inference [M]. Stanford: Stanford University, 2009.