ICRC-HIT: A Deep Learning based Comment Sequence Labeling System for Answer Selection Challenge

Xiaoqiang Zhou   Baotian Hu   Jiaxin Lin   Yang Xiang   Xiaolong Wang
Intelligence Computing Research Center
Department of Computer Science &Technology
Harbin Institute of Technology, Shenzhen Graduate School
{xiaoqiang.jeseph,baotianchina,dongshanjx,xiangyang.hitsz}@gmail.com
wangxl@insun.hit.edu.cn

Abstract

In this paper, we present a comment labeling system based on a deep learning strategy. We treat the answer selection task as a sequence labeling problem and propose recurrent convolution neural networks to recognize good comments. In the recurrent architecture of our system, our approach uses 2-dimensional convolutional neural networks to learn the distributed representation for question-comment pair, and assigns the labels to the comment sequence with a recurrent neural network over CNN. Compared with the conditional random fields based method, our approach performs better performance on Macro-F1 (53.82%), and achieves the highest accuracy (73.18%), F1-value (79.76%) on predicting the *Good* class in this answer selection challenge.

1 Introduction

The community question answering site or system (CQA) is one kind of common platforms where people can freely ask questions, deliver comments and participate in discussions. The high-quality comments given a question are the important resources to generate useful question-answer pairs, which are of great value for knowledge base construction and information retrieval (IR). However, due to the unrestricted expressions in CQA, it still one problem to recognize the high-quality comments from the open domain data, which are involve in a large of noise information. Nevertheless, the semantic relevance between question and comment makes sense to predict the quality of comment by modeling the semantic matching for question-comment pair.

Prior work on predicting the class of comment (or answer) mainly attempted to measure the semantic similarity between question and comment with typical classification approaches, such as LR and SVM. To achieve the semantic relevance matching for question-comment pair, a large number of works focus on constructing feature-engineering to extract the features of question and comment as the input of models. Beyond typical textual feature, some works integrate the structural information (Wang et al., 2009; Huang et al., 2007) into the discrete representations of question-comment pairs to improve the performances of comment classifiers. Another option is extracting user metadata (Chen and Nayak, 2008; Shah and Pomerantz, 2010) from the question answering portal for enriching the feature-engineering. Empirically the approaches above have been shown to improve performances on recognizing positive answers, but they rely on large numbers of hand-crafted features, and require various external resources which may be difficult to obtain. Furthermore, they suffer from the limitation of requiring task-specific feature extraction for new domain.

Recently the works about neural network-based distributed sentence models (Socher et al., 2012; Kalchbrenner et al., 2014) have achieved successes in natural language processing (NLP). As a consequence of this success, it appears natural to attempt to solve question answering using similar techniques. To recognize the high-quality answers, Hu et al. (2013) learned the joint representation for each question-answer pair

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 210–214,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

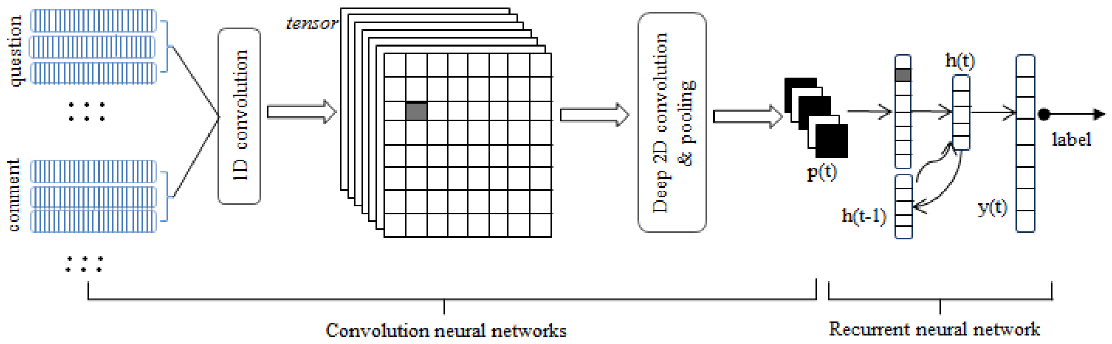

Figure 1. The architecture of comment labeling system based on deep learning

by taking both of the textual and non-textual features as the input of multi-DBN model. To achieve the answer sentence selection, Yu et al. (2014) proposed convolution neural networks based models to represent the question and answer sentences. For the semantic matching between question and answer, the methods based on deep learning generally exploit to learn the distributed representation of question-answer pair as the input. Instead of extracting a variety of features, these approaches learn the semantic features to represent question and answer. However, these approaches only focus on modeling the semantic relevance between question and answer, ignoring the semantic correlations in answer sequence.

In this work, we present a novel comment labeling system based on deep learning. We propose the recurrent convolutional neural networks (R&CNN) approach to assign the labels to comments given a question. Based on the distributed representations learned form 2-dimensional CNN (2D-CNN) matching, our approach achieves to comment sequence learning and predict the classes of comments. Using the word embedding trained by provided Qatar Living data, R&CNN not only models the semantic relevance for question and comment, but also captures the correlative context in comment sequence for predicting the class of comment. The experimental results show that our system performs better performances than the CRF based method (Ding et al., 2008) on recognizing good comments, and performs more adaptive on the development and test dataset.

2 System Description

The architecture of our comment labeling system is a recurrent architecture (shown in Figure 1) with a recurrent neural network over the convolutional neural networks. Given a question, our approach achieves to learn the semantic relevance between question and comment by 2D-CNN matching and generate the distributed representation of each question-comment pair. After that, our approach uses the RNN to model the semantic correlations in comment sequence, and makes the quality predictions for the comment sequence with the captured context.

2.1 Convolutional Neural Networks for question-comment matching

Convolutional neural networks are a natural extension of neural networks for treating image. Hu et al. (2014) proposed the 2D-CNN model to do semantic matching between two sentences. In our work, we use 2D-CNN to learn the distributed representations for question-comment pairs. Unlike 1D-CNN, executing the interaction between question and answer in final multi-layer perception (MLP) with their individual representations, 2D-CNN maps question and comment into a common space for learning the representation of question-comment pair and captures the rich matching patterns between question and answer by layer-by-layer convolution and pooling.

The first layer is 1D-convolution layer, whose role is converting word embedding of question and comment into one common space with the sliding window, whose size k is (3×3). For the word i on question q and word j on comment c, 1D-convolution can formulated as:

$$\hat{z}^{(0)}_{i,j} = [q^T_{i:i+k-1}, c^T_{i:i+k-1}] \qquad (1)$$

where $\hat{z}^{(0)}_{i,j}$ simply concatenates the vectors of sentence segments in question q and comment c;

The 1D-convolution converts the concatenated matrix H_0 of question and comment into the real-value matrix H_1 . After that, 2D-CNN executes deep 2D-convolution and pooling, similar to that of traditional image input. The output of the *m^th^* hidden layer is computed as:

$$H_m = \sigma\big(pool(w_m H_{m-1} + b_m)\big) \qquad (2)$$

Here, w_m is the parameter matrix for the feature maps on *m*th hidden layer and b_m is the bias vector. $\sigma(.)$ is the sigmoid activation function. The final distributed representation p_t of question-comment pair learned from 2D-CNN represents the semantic relevance between question and comment, and provides the reliable evidences to make a quality prediction for the corresponding comment.

2.2 Recurrent Neural Network for comment sequence labeling

Recurrent neural network is a straightforward adaptation of the standard feed-forward neural network (Bengio et al., 2012) to allow it to model sequential data. The recurrent neural network in our work has one input layer X, one hidden layer H for updating the hidden state, and the output layer Y. For the time step t, the input to RNN includes the learned representation $p(t)$ and the previous hidden state $h(t-1)$. The output is denoted as $y(t)$. The output of input, hidden and output layers are computed as:

$$x(t) = w_i p(t) + w_h h(t-1) + b_h \qquad (3)$$

$$h(t) = \sigma\big(x(t)\big) \qquad (4)$$

$$y(t) = g\big(w_y h(t) + b_y\big) \qquad (5)$$

where w_i is the matrix of connection between CNN and the input layer of RNN; w_h plays role in updating network state or context; and w_y is the matrix of connection between hidden layer and output layer. Both of b_h and b_y are bias vectors. Here, $\sigma(.)$ is the sigmoid activation function; $g(.)$ is the softmax function. $x(t)$ is the joint representation of current pair and context. Our approach is able to capture the context by updating the hidden state $h(t)$.

To train the networks proposed here, we use the backpropagation through time with stochastic gradient descent (SGD) algorithm. At each training step, error vector is computed according to cross entropy criterion, weights are updated as:

$$Error(t;\theta) = R(t) - y(t) \qquad (6)$$

where $y(t)$ is the result from our system, and $R(t)$ is the true class; and θ includes all the parameters of CNN and RNN.

3 Experiments

3.1 Experimental setup

We evaluate our approach (R&CNN) on both the development and test data of this answer selection challenge. The statistics of experimental dataset are summarized in Table 1. In this dataset, there are 3,229 questions and 21,062 answers, and the percentage of good comments is about 50%. The average length of comment sequence is 6.

data	#question	#comment	#average	% good
Train	2600	16541	6.36	48.78
Devel	300	1645	5.48	53.19
Test	329	1976	6.00	50.46

Table 1. Statistics of experimental dataset

In our approach, we use 100-dimensional word embedding trained on the provided Qatar Living data with Word2vec (Mikolov et al., 2013). The maximum size of coding the sentences with word embedding is set to be 100, and we use 3-words sliding window for 1D-convolution. The learning rate is initialized to be 0.01 and adapted dynamically using *ADADELTA* Method (Matthew, 2012). Based on the results on development set, all the hyperparameters of our approach are optimized on train set.

Table 2 lists the experimental methods and the corresponding official results. The baselines of comment sequence labeling include the method based on CRF and the approach CRF+V, which integrates distributed representation learnt from our approach (R&CNN). In addition, we illustrate the best result achieved by the supervised feature-rich approach *SFR*[1].

Results	Methods
ICRC-HIT-primary	CRF+V
ICRC-HIT-contrastive1	R&CNN
ICRC-HIT-contrastive2	CRF
JAIST-contrasive1	SFR

Table 2. The official results and experimental methods

[1] It is the approach of JAIST team in subtask-A English.

3.2 Results and analysis

Table 3 and Table 4 illustrate the results in development and test dataset respectively. As can be seen, our proposed R&CNN outperforms CRF and CRF+V on whole performances. Specifically, R&CNN achieves the state-of-the-art with the accuracy 73.18%, and 79.76% in F1-value of predicting *Good* class while performs 53.82% in Macro-F1 on the test dataset.

Methods	Macro.	Acc.	P	R	F1
CRF	50.56	59.82	72.41	**77.37**	74.81
CRF+V	**52.14**	**61.03**	74.80	76.00	**75.40**
R&CNN	52.10	60.85	**75.09**	75.09	75.09

Table 3. Performances on development dataset (%)

Methods	Macro.	Acc.	P	R	F1
CRF	40.54	60.12	57.90	**95.89**	72.21
CRF+V	49.50	67.86	65.99	91.68	76.74
R&CNN	53.82	**73.18**	74.39	85.96	**79.76**
SFR	**57.29**	72.67	**80.51**	78.03	79.11

Table 4. Performances on test dataset (%)

Compared to CRF and CRF+V, our approach outperforms them in evolution metrics. There are several reasons for the unsatisfying performances of CRF and CRF+V. First, it is sparse to extract semantic features of question-comment pairs from short contents in baselines. In contrast, the distributed representation learned from our model is able to capture semantic relationship between words of question-comment pairs based on deep convolution and pooling. Secondly, there are large amount of noise information involved in CQA, such as various emotional symbols and the abbreviated words. The feature-engineering of CRF based method generally suffers from the quality of dataset. Besides of that, the divergences of class distribution between the development and test influence the effectiveness directly. Hence, our approach performs more powerful and adaptive to different dataset or new domain. We also can demonstrate this point by comparing the experimental results of CRF and CRF+V on the test (shown in Table 4). By integrating the distributed representation from our R&CNN, CRF+V improves 9% on Macro-F1, 7.74% on accuracy over CRF, and 4.53% in F1-value of predicting *Good* class.

Taking only word embedding as the original features, our approach has achieved 53.82% in Macro-F1. In contrast, the supervised feature-rich (SFR) approach performs 57.29% in Macro-F1 by integrating multi-type features, such as word embedding, features from topic models and user metadata etc. The main reason for that is the low performance of our approach on predicting the answers of *Potential* class, which has a major import on Macro-F1 due to the effect of marcoaveraging. There are several factors for that result. The first is the imbalance distribution in training data, which is lacking of the train samples of *Potential* class. So the distributed models based purely on word embedding are not very well equipped to learn the meaningful representations for question and potential comments. Secondly, *Potential* class is an intermediate category (Màrquez et al., 2015) that was quite hard to human annotators. Hence, surface-form matching between the words of question-comment pair is hard to identify its correct class merely using word embedding.

In addition, when considering the heavy reliance of feature-engineer of SFR in comparison to the simplicity of our approach, the Macro-F1 our approach obtained is highly encouraging. What's more, our model achieves the start-of-the-art in accuracy and F1-value of *Good* class. These promising results indicate the effectiveness of our approach in predicting the high-quality comments in CQA.

4 Conclusion

In this paper, we present a comment labeling system based on the deep learning architecture. Without the complicated feature-engineering and external semantic resources, the recurrent convolutional neural networks (R&CNN) approach proposed by us not only is able to capture semantic matching patterns between question and comments, but also learn the meaningful context in the comment sequence. In this answer selection task, our approach achieves the state-of-the-art on recognizing good comments, and performs better accuracy than baselines while obtains powerful results in Macro-F1.

In the future, we would like to investigate the methods of training the imbalance data (e.g. the *Potential* class) to improve the performances of our approach, such as the typical oversampling and undersampling methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61272383, 61173075 and 61203378), the Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20120613151940045).

References

Baotian Hu, Zhengdong Lu, Hang Li, Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In Proceedings of Neural Information Processing Systems (NIPS), Montreal, Quebec, Canada. 2014.

Baoxun Wang, Bingquan Liu, Chengjie Sun, Xiaolong Wang, and Lin Sun. 2009. Extracting Chinese Question-Answer Pairs from Online Forums. IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 1159-1164. 2009.

Haifeng Hu, Bingquan Liu, Baoxun Wang, Ming Liu, Xiaolong Wang. 2013. Multimodal DBN for predicting high-quality answers in cQA portals. In Proceedings of Association for Computational Linguistics (ACL), pages 843–847, Sofia, *Bulgaria*. 2013.

Jizhou Huang, Ming Zhou, and Dan Yang. 2007. Extracting chatbot knowledge from online discussion forums. In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), pages 423–428, Hyderabad, India. 2007.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, Stephen Pulman. 2014. Deep learning for answer sentence selection. In Proceeding of Neural Information Processing Systems (NIPS): Deep Learning and Representation Learning Workshop, Montreal, Quebec, Canada. 2014.

Lin Chen, Richi Nayak. 2008. Expertise Analysis in a Question Answer Portal for Author Ranking. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pages 134-140. 2008.

Lluís Màrquez, James Glass, Walid Magdy, Alessandro Moschitti, Preslav Nakov, and Bilal Randeree. 2015. SemEval-2015 Task 3: Answer Selection in Community Question Answering. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015). 2015

Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. CoRR abs/1212.5701. 2012

Nai Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In Proceedings of the Association for Computational Linguistics (ACL), pages 655-665, Baltimore, USA. 2014.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In Proceedings of Empirical Methods on Natural Language Processing and Computational Natural Language Learning (EMNLP), pages 1201-1211, Jeju Island, Korea. 2012.

Chirag Shah and Jefferey Pomerantz. 2010. Evaluating and predicting answer quality in community QA. In the 33rd International Conference on Research and development information retrieval on Research and Development in Information Retrieval (SIGIR'10), pages 411-418, NewYork, USA. 2010.

Shilin Ding, Gao Cong, Chin-Yew Lin, and Xiaoyan Zhu. 2008. Using conditional random fields to extract contexts and answers of questions from online forums. In Proceedings of Association for Computational Linguistics (ACL), pages 710-718, Columbus, Ohis, USA. 2008.

Tom Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781. 2013

Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. 2012. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In Proceedings of the 29th International Conference on Machine Learning (ICML), Edinburgh, Scotland, UK. 2012.