A Multi-sentiment-resource Enhanced Attention Network for Sentiment Classification

Zeyang Lei^{1,2}, Yujiu Yang^{1}, Min Yang^{3}, and Yi Liu^{2}
Graduate School at Shenzhen, Tsinghua University^{1}
Peking University Shenzhen Institute^{2}
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences^{3}
leizy16@mails.tsinghua.edu.cn, yang.yujiu@sz.tsinghua.edu.cn,
min.yang1129@gmail.com, eeyliu@gmail.com

Abstract

Deep learning approaches for sentiment classification do not fully exploit sentiment linguistic knowledge. In this paper, we propose a Multi-sentiment-resource Enhanced Attention Network (MEAN) to alleviate the problem by integrating three kinds of sentiment linguistic knowledge (e.g., sentiment lexicon, negation words, intensity words) into the deep neural network via attention mechanisms. By using various types of sentiment resources, MEAN utilizes sentiment-relevant information from different representation subspaces, which makes it more effective to capture the overall semantics of the sentiment, negation and intensity words for sentiment prediction. The experimental results demonstrate that MEAN has robust superiority over strong competitors.

1 Introduction

Sentiment classification is an important task of natural language processing (NLP), aiming to classify the sentiment polarity of a given text as positive, negative, or more fine-grained classes. It has obtained considerable attention due to its broad applications in natural language processing (Hao et al., 2012; Gui et al., 2017). Most existing studies set up sentiment classifiers using supervised machine learning approaches, such as support vector machine (SVM) (Pang et al., 2002), convolutional neural network (CNN) (Kim, 2014; Bonggun et al., 2017), long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997; Qian et al., 2017), Tree-LSTM (Tai et al., 2015), and attention-based methods (Zhou et al., 2016; Yang et al., 2016; Lin et al., 2017; Du et al., 2017).

Despite the remarkable progress made by the previous work, we argue that sentiment analysis still remains a challenge. Sentiment resources including sentiment lexicon, negation words, intensity words play a crucial role in traditional sentiment classification approaches (Maks and Vossen, 2012; Duyu et al., 2014). Despite its usefulness, to date, the sentiment linguistic knowledge has been underutilized in most recent deep neural network models (e.g., CNNs and LSTMs).

In this work, we propose a Multi-sentiment-resource Enhanced Attention Network (MEAN) for sentence-level sentiment classification to integrate many kinds of sentiment linguistic knowledge into deep neural networks via multi-path attention mechanism. Specifically, we first design a coupled word embedding module to model the word representation from character-level and word-level semantics. This can help to capture the morphological information such as prefixes and suffixes of words. Then, we propose a multi-sentiment-resource attention module to learn more comprehensive and meaningful sentiment-specific sentence representation by using the three types of sentiment resource words as attention sources attending to the context words respectively. In this way, we can attend to different sentiment-relevant information from different representation subspaces implied by different types of sentiment sources and capture the overall semantics of the sentiment, negation and intensity words for sentiment prediction.

The main contributions of this paper are summarized as follows. First, we design a coupled word embedding obtained from character-level embedding and word-level embedding to capture both the character-level morphological information and word-level semantics. Second, we propose a multi-sentiment-resource attention module to learn more comprehensive sentiment-specific sentence representation from multiply subspaces

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 758–763
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

implied by three kinds of sentiment resources including sentiment lexicon, intensity words, negation words. Finally, the experimental results show that MEAN consistently outperforms competitive methods.

2 Model

Our proposed MEAN model consists of three key components: coupled word embedding module, multi-sentiment-resource attention module, sentence classifier module. In the rest of this section, we will elaborate these three parts in details.

2.1 Coupled Word Embedding

To exploit the sentiment-related morphological information implied by some prefixes and suffixes of words (such as "Non-", "In-", "Im-"), we design a coupled word embedding learned from character-level embedding and word-level embedding. We first design a character-level convolution neural network (Char-CNN) to obtain character-level embedding (Zhang et al., 2015). Different from (Zhang et al., 2015), the designed Char-CNN is a fully convolutional network without max-pooling layer to capture better semantic information in character chunk. Specifically, we first input one-hot-encoding character sequences to a 1×1 convolution layer to enhance the semantic nonlinear representation ability of our model (Long et al., 2015), and the output is then fed into a multi-gram (i.e. different window sizes) convolution layer to capture different local character chunk information. For word-level embedding, we use pre-trained word vectors, GloVe (Pennington et al., 2014), to map each word to a low-dimensional vector space. Finally, each word is represented as a concatenation of the character-level embedding and word-level embedding. This is performed on the context words and the three types of sentiment resource words [1], resulting in four final coupled word embedding matrices: the $W^c = [w^c_1, ..., w^c_t] \in \mathbb{R}^{d \times t}$ for context words, the $W^s = [w^s_1, ..., w^s_m] \in \mathbb{R}^{d \times m}$ for sentiment words, the $W^i = [w^i_1, ..., w^i_k] \in \mathbb{R}^{d \times k}$ for intensity words, the $W^n = [w^n_1, ..., w^n_p] \in \mathbb{R}^{d \times p}$ for negation words. Here, t, m, k, p are the length of the corresponding items respectively, and d is the embedding dimension. Each W is normalized to better calculate the following word correlation.

2.2 Multi-sentiment-resource Attention Module

After obtaining the coupled word embedding, we propose a multi-sentiment-resource attention mechanism to help select the crucial sentiment-resource-relevant context words to build the sentiment-specific sentence representation. Concretely, we use the three kinds of sentiment resource words as attention sources to attend to the context words respectively, which is beneficial to capture different sentiment-relevant context words corresponding to different types of sentiment sources. For example, using sentiment words as attention source attending to the context words helps form the sentiment-word-enhanced sentence representation. Then, we combine the three kinds of sentiment-resource-enhanced sentence representations to learn the final sentiment-specific sentence representation. We design three types of attention mechanisms: sentiment attention, intensity attention, negation attention to model the three kinds of sentiment resources, respectively. In the following, we will elaborate the three types of attention mechanisms in details.

First, inspired by (Xiong et al.), we expect to establish the word-level relationship between the context words and different kinds of sentiment resource words. To be specific, we define the dot products among the context words and the three kinds of sentiment resource words as correlation matrices. Mathematically, the detailed formulation is described as follows.

$$M^s = (W^c)^T \cdot W^s \in \mathbb{R}^{t \times m} \tag{1}$$

$$M^i = (W^c)^T \cdot W^i \in \mathbb{R}^{t \times k} \tag{2}$$

$$M^n = (W^c)^T \cdot W^n \in \mathbb{R}^{t \times p} \tag{3}$$

where M^s, M^i, M^n are the correlation matrices to measure the relationship among the context words and the three kinds of sentiment resource words, representing the relevance between the context words and the sentiment resource word.

After obtaining the correlation matrices, we can compute the sentiment-resource-relevant context word representations X^c_s, X^c_i, X^c_n by the dot products among the context words and different types of corresponding correlation matrices. Meanwhile, we can also obtain the context-word-relevant sentiment word representation matrix X^s by the dot product between the correlation matrix M^s and the sentiment words W^s, the context-

[1] To be precise, sentiment resource words include sentiment words, negation words and intensity words.

word-relevant intensity word representation matrix X^i by the dot product between the intensity words W^i and the correlation matrix M^i, the context-word-relevant negation word representation matrix X^n by the dot product between the negation words W^n and the correlation matrix M^n. The detailed formulas are presented as follows:

$$X_s^c = W^c M^s, X^s = W^s(M^s)^T \tag{4}$$

$$X_i^c = W^c M^i, X^i = W^i(M^i)^T \tag{5}$$

$$X_n^c = W^c M^n, X^n = W^n(M^n)^T \tag{6}$$

The final enhanced context word representation matrix is computed as:

$$X^c = X_s^c + X_i^c + X_n^c. \tag{7}$$

Next, we employ four independent GRU networks (Chung et al., 2015) to encode hidden states of the context words and the three types of sentiment resource words, respectively. Formally, given the word embedding X^c, X^s, X^i, X^n, the hidden state matrices H^c, H^s, H^i, H^n can be obtained as follows:

$$H^c = GRU(X^c) \tag{8}$$

$$H^s = GRU(X^s) \tag{9}$$

$$H^i = GRU(X^i) \tag{10}$$

$$H^n = GRU(X^n) \tag{11}$$

After obtaining the hidden state matrices, the sentiment-word-enhanced sentence representation o_1 can be computed as:

$$o_1 = \sum_{i=1}^{t} \alpha_i h_i^c, q^s = \sum_{i=1}^{m} h_i^s / m \tag{12}$$

$$\beta([h_i^c; q_s]) = u_s^T tanh(W_s[h_i^c; q_s]) \tag{13}$$

$$\alpha_i = \frac{exp(\beta([h_i^c; q_s]))}{\sum_{i=1}^{t} exp(\beta([h_i^c; q_s]))} \tag{14}$$

where q^s denotes the mean-pooling operation towards H^s, β is the attention function that calculates the importance of the i-th word h_i^c in the context and α_i indicates the importance of the i-th word in the context, u_s and W_s are learnable parameters.

Similarly, with the hidden states H^i and H^n for the intensity words and the negation words as attention sources, we can obtain the intensity-word-enhanced sentence representation o_2 and the negation-word-enhanced sentence representation o_3. The final comprehensive sentiment-specific sentence representation $\tilde{\mathbf{o}}$ is the composition of the above three sentiment-resource-specific sentence representations o_1, o_2, o_3:

$$\tilde{\mathbf{o}} = [o_1, o_2, o_3] \tag{15}$$

2.3 Sentence Classifier

After obtaining the final sentence representation $\tilde{\mathbf{o}}$, we feed it to a softmax layer to predict the sentiment label distribution of a sentence:

$$\hat{y} = \frac{exp(\tilde{W_o}^T \tilde{\mathbf{o}} + \tilde{b_o})}{\sum_{i=1}^{C} exp(\tilde{W_o}^T \tilde{\mathbf{o}} + \tilde{b_o})} \tag{16}$$

where $\hat{y}$ is the predicted sentiment distribution of the sentence, C is the number of sentiment labels, $\tilde{W_o}$ and $\tilde{b_o}$ are parameters to be learned.

For model training, our goal is to minimize the cross entropy between the ground truth and predicted results for all sentences. Meanwhile, in order to avoid overfitting, we use dropout strategy to randomly omit parts of the parameters on each training case. Inspired by (Lin et al., 2017), we also design a penalization term to ensure the diversity of semantics from different sentiment-resource-specific sentence representations, which reduces information redundancy from different sentiment resources attention. Specifically, the final loss function is presented as follows:

$$L(\hat{y}, y) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^j log(\hat{y}_i^j) + \lambda(\sum_{\theta\in\Theta}\theta^2) \tag{17}$$

$$+ \mu||\tilde{O}\tilde{O}^T - \psi I||_F^2$$

$$\tilde{O} = [o_1; o_2; o_3] \tag{18}$$

where y_i^j is the target sentiment distribution of the sentence, $\hat{y}_i^j$ is the prediction probabilities, θ denotes each parameter to be regularized, Θ is parameter set, λ is the coefficient for L_2 regularization, μ is a hyper-parameter to balance the three terms, ψ is the weight parameter, I denotes the the identity matrix and $||.||_F$ denotes the Frobenius norm of a matrix. Here, the first two terms of the loss function are cross-entropy function of the predicted and true distributions and L_2 regularization respectively, and the final term is a penalization term to encourage the diversity of sentiment sources.

3 Experiments

3.1 Datasets and Sentiment Resources

Movie Review (MR)[2] and Stanford Sentiment Treebank (SST)[3] are used to evaluate our model. MR dataset has 5,331 positive samples and 5,331 negative samples. We adopt the same data split as in (Qian et al., 2017). SST consists of 8,545 training samples, 1,101 validation samples, 2210 test samples. Each sample is marked as very negative, negative, neutral, positive, or very positive. Sentiment lexicon combines the sentiment words from both (Qian et al., 2017) and (Hu and Liu, 2004), resulting in 10,899 sentiment words in total. We collect negation and intensity words manually as the number of these words is limited.

3.2 Baselines

In order to comprehensively evaluate the performance of our model, we list several baselines for sentence-level sentiment classification.

RNTN: Recursive Tensor Neural Network (Socher et al., 2013) is used to model correlations between different dimensions of child nodes vectors.

LSTM/Bi-LSTM: Cho et al. (2014) employs Long Short-Term Memory and the bidirectional variant to capture sequential information.

Tree-LSTM: Memory cells was introduced by Tree-Structured Long Short-Term Memory (Tai et al., 2015) and gates into tree-structured neural network, which is beneficial to capture semantic relatedness by parsing syntax trees.

CNN: Convolutional Neural Networks (Kim, 2014) is applied to generate task-specific sentence representation.

NCSL: Teng et al. (2016) designs a Neural Context-Sensitive Lexicon (NSCL) to obtain prior sentiment scores of words in the sentence.

LR-Bi-LSTM: Qian et al. (2017) imposes linguistic roles into neural networks by applying linguistic regularization on intermediate outputs with KL divergence.

Self-attention: Lin et al. (2017) proposes a self-attention mechanism to learn structured sentence embedding.

ID-LSTM: (Tianyang et al., 2018) uses reinforcement learning to learn structured sentence representation for sentiment classification.

3.3 Implementation Details

In our experiments, the dimensions of character-level embedding and word embedding (GloVe) are both set to 300. Kernel sizes of multi-gram convolution for Char-CNN are set to 2, 3, respectively. All the weight matrices are initialized as random orthogonal matrices, and we set all the bias vectors as zero vectors. We optimize the proposed model with RMSprop algorithm, using mini-batch training. The size of mini-batch is 60. The dropout rate is 0.5, and the coefficient λ of L_2 normalization is set to 10^{-5}. μ is set to 10^{-4}. ψ is set to 0.9. When there are not sentiment resource words in the sentences, all the context words are treated as sentiment resource words to implement the multi-path self-attention strategy.

3.4 Experiment Results

In our experiments, to be consistent with the recent baseline methods, we adopt classification accuracy as evaluation metric. We summarize the experimental results in Table 1. Our model has robust superiority over competitors and sets state-of-the-art on MR and SST datasets. First, our model brings a substantial improvement over the methods that do not leverage sentiment linguistic knowledge (e.g., RNTN, LSTM, BiLSTM, CNN and ID-LSTM) on both datasets. This verifies the effectiveness of leveraging sentiment linguistic resource with the deep learning algorithms. Second, our model also consistently outperforms LR-Bi-LSTM which integrates linguistic roles of sentiment, negation and intensity words into neural networks via the linguistic regularization. For example, our model achieves 2.4% improvements over the MR dataset and 0.8% improvements over the SST dataset compared to LR-Bi-LSTM. This is because that MEAN designs attention mechanisms to leverage sentiment resources efficiently, which utilizes the interactive information between context words and sentiment resource words.

In order to analyze the effectiveness of each component of MEAN, we also report the ablation test in terms of discarding character-level embedding (denoted as MEAN w/o CharCNN) and sentiment words/negation words/intensity words (denoted as MEAN w/o sentiment words/negation words/intensity words). All the tested factors con-

[2] http://www.cs.cornell.edu/people/pabo/movie-review-data/

[3] https://nlp.stanford.edu/sentiment/we train the model on both phrases and sentences but only test on sentences

tribute greatly to the improvement of the MEAN. In particular, the accuracy decreases sharply when discarding the sentiment words. This is within our expectation since sentiment words are vital when classifying the polarity of the sentences.

Methods	MR	SST
RNTN	75.9%#	45.7%
LSTM	77.4%#	46.4%
BiLSTM	79.3%#	49.1%
Tree-LSTM	80.7%#	51.0%
CNN	81.5%	48.0%
NSCL	82.9%	51.1%
LR-Bi-LSTM	82.1%	50.6%
Self-attention	81.7%*	48.9%*
ID-LSTM	81.6%	50.0%
MEAN(our model)	**84.5%**	**51.4%**
MEAN-CharCNN	83.2%	50.0%
MEAN-sentiment words	82.1%	48.4%
MEAN-negation words	82.9%	49.5%
MEAN-intensity words	83.5%	49.3%

Table 1: Evaluation results. The best result for each dataset is in bold. The result marked with # are retrieved from (Qian et al., 2017), and the results marked with * denote the results are obtained by our implementation.

4 Conclusion

In this paper, we propose a novel Multi-sentiment-resource Enhanced Attention Network (MEAN) to enhance the performance of sentence-level sentiment analysis, which integrates the sentiment linguistic knowledge into the deep neural network.

Acknowledgements

This work was supported in part by the Research Fund for the development of strategic emerging industries by ShenZhen city (No.JCYJ20160301151844537 and No. JCYJ20160331104524983).

References

Shin Bonggun, Lee Timothy, and D. Choi Jinho. 2017. Lexicon integrated cnn models with attention for sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, ACL 2017*.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *Proceedings of ICML 2015*.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. In *Proceedings of IJCAI 2017*.

Tang Duyu, Wei Furu, Qin Bing, Liu Ting, and Zhou Ming. 2014. Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014*.

Lin Gui, Yu Zhou, Ruifeng Xu, Yulan He, and Qin Lu. 2017. Learning representations from heterogeneous network for sentiment classification of product reviews. *Knowledge-Based Systems*, 124:34–45.

Li Hao, Chen Yu, Ji Heng, Muresan Smaranda, and Zheng Dequan. 2012. Combining social cognitive theories with linguistic features for multi-genre sentiment analysis. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, pages 1735–1780.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of SIGKDD 2004*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP 2014*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of ICLR 2017*.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of CVPR 2015*.

Isa Maks and Piek Vossen. 2012. A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, pages 680–688.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of ACL 2002*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*.

Qiao Qian, Minlie Huang, Jinhao Lei, and Xiaoyan Zhu. 2017. Linguistically regularized LSTM for sentiment classification. In *Proceedings of ACL 2017*, pages 1679–1689.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP 2013*.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of ACL 2015*.

Zhiyang Teng, Duy-Tin Vo, and Yue Zhang. 2016. Context-sensitive lexicon features for neural sentiment analysis. In *Proceedings of EMNLP 2016*.

Zhang Tianyang, Huang Minlie, and Li Zhao. 2018. Learning structured representation for text classification via reinforcement learning. In *Proceedings of AAAI 2018*.

Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. In *ICLR*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL 2016*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of NIPS 2015*.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Attention-based lstm network for cross-lingual sentiment classification. In *Proceedings of EMNLP 2016*.