Attention Aware Semi-supervised Framework for Sentiment Analysis

Jingshuang Liu[1], Wenge Rong[1(✉)], Chuan Tian[1], Min Gao[2], and Zhang Xiong[1]

[1] School of Computer Science and Engineering, Beihang University, Beijing, China
{jingshuangliu,w.rong,chuantian,xiongz}@buaa.edu.cn
[2] School of Software Engineering, Chonqing University, Chongqing, China
gaomin@cqu.edu.cn

Abstract. Using sentiment analysis methods to retrieve useful information from the accumulated documents in the Internet has become an important research subject. In this paper, we proposed a semi-supervised framework, which uses the unlabeled data to promote the learning ability of the long short memory (LSTM) network. It is composed of an unsupervised attention aware long short term memory (LSTM) encoder-decoder and a single LSTM model used for feature extraction and classification. Experimental study on commonly used datasets has demonstrated our framework's good potential for sentiment classification tasks. And it has shown that the unsupervised learning part can improve the LSTM network's learning ability.

Keywords: Sentiment analysis · Semi-supervised learning · Attention · Long short term memory · Encoder-decoder

1 Introduction

Nowadays, people tend to publish opinions and comments on goods, movies, and etc. through the Internet based services. As a result a large number of such documents have been collected, which makes it an important task to retrieve valuable information beneath these online collections [13]. Sentiment analysis is an effective method to investigate the polarity of online posted messages [13].

Many deep models have been introduced to the sentiment analysis tasks, among which Recurrent Neural Network (RNN) has shown its great power since it can learn the underlying relationships between the words [12]. However, RNN had the vanishing gradient problem [6]. To overcome this shortcoming, long short term memory (LSTM) with gates and cell mechanism has been proposed in the literature. It has proven its potential in maintaining the advantages of RNN while overcoming the problems of vanishing gradient [8].

Besides the more and more advanced model, for the classification tasks it is also believed that importing external knowledge from different domains can help to improve the model's performance [3]. Under this assumption, we proposed a framework which employs an unsupervised model for pre-training the parameters in the supervised model. At present, the encoder-decoder is a useful structure for

A. Lintas et al. (Eds.): ICANN 2017, Part II, LNCS 10614, pp. 208–215, 2017.
https://doi.org/10.1007/978-3-319-68612-7_24

unsupervised learning in natural language process (NLP) tasks [5], and attention mechanism is another valid promotion to the behaviour of unsupervised learning [4]. Therefore, we proposed to use an attention mechanism aware LSTM encoder-decoder to pre-train parameters in a unsupervised step. In this part, we applied unlabelled data for training. Then the supervised learning LSTM network is implemented for feature extraction and classification.

The rest of the paper is organized as follows: Sect. 2 will present the background knowledge about our research. The pipeline of our proposed semi-supervised model will be illustrated in Sect. 3. Finally, the experiment study and discussion will be discussed in Sect. 4 and Sect. 5 will conclude this paper and point out possible future directions.

2 Background

Currently sentiment analysis has gained much attention in both academic and industrial community [13]. Since a large amount of information can be collected through the social media network, sentiment analysis has witnessed a great development in analysing the information beneath the online documents. Due to its capability of analysing polarity of online documents, it has become a fundamental technique in a lot of applications [13].

For sentiment analysis, supervised learning has proven its success in many tasks [3]. However, in order to enhance the generalisation ability of the learning model, this kind of method need a large number of labelled corpus. At present, it is convenient to acquire abundant unlabelled corpus from the website. But the dataset are usually limited in quantity, quality, and coverage [2]. Meanwhile, in some cases, the supervised learning model was only randomly initialised and the parameters was likely to get into poor local minimums [3]. If the model can be initialised properly, the generalisation ability can be improved [3].

In machine learning, semi-supervised learning is a useful technique to use both the unlabelled and labelled data [2]. The most common way for semi-supervised learning is to split the training process into two parts [11]. The first part is the unsupervised learning using large unlabelled corpus. In this step, a set of parameters can be obtained. The second is using the retrieved parameters to initialise a supervised learning model. In this way, the parameters are fine tuned with the model globally using labelled data. In this research, we used the data crawled from website for unsupervised learning. For instance, we used the Amazon Review data which has high relevance to our experimental dataset. Therefore, we can acquire high quality parameters with these unlabelled dataset.

The encoder-decoder model has achieved great success in the tasks for NLP lately, such as machine translation, speech recognition, slot filling and text parsing [5]. In this approach, sequence models can be applied as an encoder to encode the input data into a state. Afterwards the state is used for the input of a decoder, which is also variable, to predict the output sequence. In our work, we used LSTM as both the encoder and decoder.

The attention mechanism has been proposed in recent years and has made great breakthroughs in fields like machine translation, video and image analysis

[4]. It can free the model from having to encode a whole sequence into a vector, and the decoder only needs to focus on the relevant information [4]. This mechanism can help enhance the accuracy of the unsupervised process greatly.

3 Methodology

The proposed sentiment analysis framework is shown in Fig. 1, where the first section is the unsupervised model to obtain the parameters for initialisation. The second section is a supervised LSTM network with a Softmax layer used to learn the hidden features of the instance and make prediction.

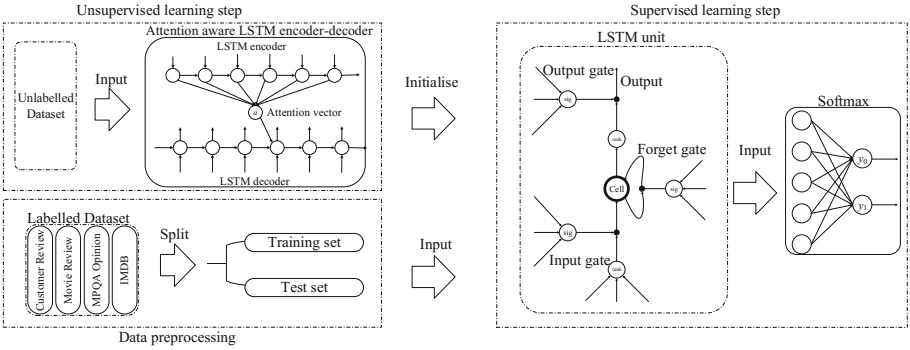

Fig. 1. Attention aware semi-supervised LSTM framework

3.1 Unsupervised Learning

Here we employed an attention aware LSTM encoder-decoder structure in the unsupervised learning step to pre-train the parameters. The LSTM encoder-decoder is used to reconstruct the input word sequence. For the encoding part, at each time step, a single word was inputted into the encoder until the model gets an $\langle EOS \rangle$ symbol. Finally the hidden state of the encoder could be gained. And for the decoding part, the hidden state acquired from former step worked as the initial state of the LSTM decoder. At each time step, the output was applied as the input for next time step until the model outputted an $\langle EOS \rangle$.

The attention mechanism is employed as follows: As shown in the unsupervised learning step in Fig. 1, a context vector a_t is added to the output layer of the LSTM decoder. The vector a_t was computed by:

$$e_{tj} = sim(s_{t-1}, h_j) \tag{1}$$

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{i=1}^{T_x} \exp(e_{ti})} \tag{2}$$

$$a_t = \sum_{j=1}^{T_x} \alpha_{tj} h_j \tag{3}$$

where e_{tj} is the cosine similarity between the state of the memory cell in time step t_1 in the decoder s_{t-1} and the state of the memory cell in time step j of the encoder h_j. In the figure, we only take one attention vector for example for simplicity, and in fact there are a lot more context vectors (For instance, if the input length is 10, then there are 10 context vectors in total). With the context vector added, the values of the gates and cells in the LSTM decoder were computed as follows:

$$i = \sigma(x_t U^i + s_{t-1} W^i + a_t) \tag{4}$$
$$f = \sigma(x_t U^f + s_{t-1} W^f + a_t) \tag{5}$$
$$o = \sigma(x_t U^o + s_{t-1} W^o + a_t) \tag{6}$$
$$\hat{c}_t = \tanh(x_t U^c + s_{t-1} W^c + a_t) \tag{7}$$
$$c_t = c_{t-1} \circ f + \hat{c}_t \circ i \tag{8}$$
$$s_t = \tanh(c_t) \circ o \tag{9}$$

where x_t is the input to the memory cell. U^i, U^f, U^o, U^c, W^i, W^f, W^o, W^c are weight matrices. i, f, o stand for values of the input gate, forget gate, and output gate. $\hat{c}_t$ is the candidate value for states of the memory cell. c_t is the new state of memory cell and s_t is the output of hidden state at time step t. And a_t is added respectively.

In this part, we trained the unsupervised model with unlabelled dataset. After the unsupervised training process, we extracted the parameters of the LSTM encoder and used it in the later supervised training part.

3.2 Supervised Learning

Feature Extraction. We initialised the LSTM network with the parameters we obtained from the last part. In this step, we used the labelled instances as input. In the LSTM, the new state of memory cell c_t and the output of hidden state s_t are computed the same as Eqs. 8 and 9. The other values of gates and cells were computed like those we introduced before in Sect. 3.1 only without the attention vector a_t.

Prediction. In the last part, a Softmax classifier was attached to the LSTM network to predict whether the output of the LSTM network is positive or negative. After Implementing the framework, we trained the two parts respectively. The Cross Entropy was used as the cost function:

$$\text{Cost} = -\frac{1}{m}\sum_{i=1}^{\text{m}}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)] \tag{10}$$

where $\hat{y}_i$ is the predicted result and y_i is the label of the instance i, m is the number of the instances. In this research, we employed stochastic gradient descend (SGD) method to optimise the back propagation through time algorithm. Our

algorithm was realized on the Theano platform. And in order to accelerate our training process, we referenced Bengio et al.'s approach to train compute the i, f, o, a_t in parallel [1].

4 Experiment Study

4.1 Datasets and Evaluation Metrics

In this research, in order to evaluate the proposed model and test the performance of different word embedding strategy, four public datasets are employed, i.e. the non-balanced dataset Customer Review[1], MPQA opinion corpus[2], and the balanced dataset Movie Review[3] and IMDB[4]. The four datasets have 3,772, 10,662, 10624, 50,000 instances respectively and the positive and negative instances rates are 0.64/0.36, 0.31/0.69, 0.5/0.5, 0.5/0.5. In this research, for the first three datasets, we randomly split the datasets into ten sets and adopt the 10-fold cross validation strategy to compute the average accuracy. And for IMDB, since the author has already split it into 50%/50% for training and testing, we just followed this common splitting approach.

To evaluate the proposed model's potential, the widely used measurement accuracy [9] is employed in this research and it is defined as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} 1\{y_i = p_i\}}{\#testdata} \tag{11}$$

where y_i stands for the true value that the instance is labelled, and p_i is the result predicted by our model. By evaluating the accuracy in testing set, we can see our model's performance towards generalised dataset.

To test the performance of the proposed model, several baselines in the literature are employed. For the Customer Review, MPQA Opinion Corpus and Movie Review, Bag-of-Words, Vote by lexicon, Rule-based reversal, Tree-Based CRF, word embedding based CNN, RNN and LSTM are employed [7,8,10,12]. While for IMDB dataset, LSA, LDA, MAAS Semantic, MAAS Full, word embedding based CNN, RNN and LSTM are used [7–9,12]. Also, for comparsion, we implemented a framework with random initialised parameters (without attention aware pre-training step).

4.2 Results and Discussion

The comparison of the proposed model against the baseline methods are displayed in Tables 1 and 2. It is found that the proposed model outperform the

[1] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.
[2] http://mpqa.cs.pitt.edu/.
[3] http://www.cs.cornell.edu/people/pabo/movie-review-data/.
[4] http://ai.stanford.edu/~amaas/data/sentiment/.

baseline methods in all the four datasets. The reason is that other baseline methods cannot learn the underlying relationships between words in great intervals. For instance, word embedding based CNN is not a time series model, Word embedding based RNN has the problem of vanishing gradient and WB+LSTM is not well pre-trained since it is a simple supervised learning model.

Table 1. Accuracy for customer review, MPQA opinion and movie review

Method	Customer review	MPQA opinion	Movie review
Bag-of-word	0.814	0.841	0.764
Voting by lexicon	0.742	0.817	0.631
Rule-based reversal	0.743	0.818	0.629
Tree-CRF	0.814	0.861	0.773
Word embedding based CNN	0.819	0.918	0.778
Word embedding based RNN	0.821	0.867	0.781
Word embedding based LSTM	0.764	0.922	0.774
Our framework (no attention aware)	0.830	0.921	0.784
Our framework (attention aware)	**0.846**	**0.928**	**0.797**

Table 2. Accuracy for IMDB

Method	IMDB
LSA	0.839
LDA	0.674
MAAS semantic	0.873
MAAS full	0.874
Word embedding based CNN	0.884
Word embedding based RNN	0.829
Word embedding based LSTM	0.835
Our framework (no attention aware)	0.891
Our framework (attention aware)	**0.901**

Besides, our framework surpasses the semi-supervised LSTM without attention mechanism. It indicates that the attention mechanism used in our framework can promote the performance of the LSTM encoder-decoder model. The reason is that the attention mechanism makes the LSTM encoder not have to encode a long sequence into a single vector. It helps the LSTM decoder focus on the relevant information and enhances the unsupervised process.

Comparing with other baseline methods, we find the proposed model gained satisfactory performance. The reason is probably mainly three folds: (1) the unsupervised step which pre-training the parameters for the LSTM network

can really help to improve the performance of the LSTM; (2) the attention mechanism makes the LSTM encoder-decoder get better training ability; (3) the LSTM network behaves well in learning long-range dependencies of words as it can catch long sequence information.

Furthermore, we also compared different unlabelled dataset used for the pre-training step and the result is as shown in Fig. 2, where we set a LSTM with random initialised parameters (without pre-training) as the baseline. In this experiment, we found the Amazon review (256,479 sentences) attained the best result. And the model pre-trained by IMDB training set (25,000 sentences) also got better results than the randomly initialised LSTM. The reason is that the unlabelled dataset consists of exterior knowledge. And the Amazon Review includes highly relevant information to our experimental dataset since most of them are all reviews. As a result, the Amazon Review has more useful messages.

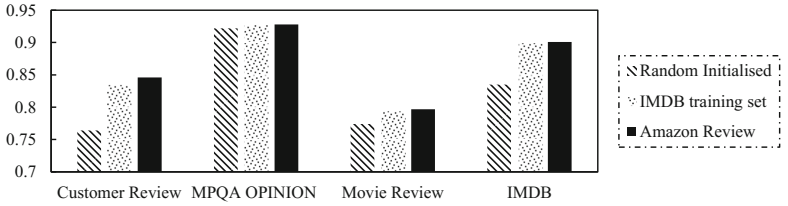

Fig. 2. Results of different datasets used in the unsupervised learning step

5 Conclusion and Future Work

In this paper, we proposed an attention aware semi-supervised LSTM framework. We first introduced the tasks of sentiment analysis and analysed the benefits of unsupervised learning for sentiment analysis. The attention mechanism made the unsupervised encoder-decoder LSTM to focus on the useful information and thus we can obtain the well pre-trained parameters for initialising the supervised LSTM. Afterwards, we presented our model in detail. We first employed an unsupervised learning model for pre-training the parameters. Then we constructed an LSTM network for feature extraction and prediction. The LSTM network was initialised by the parameters obtained in the unsupervised procedure. In our experiments, the proposed framework beat the baseline methods. The experimental results had proven the generalisation ability of framework for sentiment analysis tasks. Concerning to the future work, we plan to use different kinds of unlabelled data for the unsupervised learning process. And we will replace the LSTM encoder-decoder structure with other sequence model.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (No. 61332018), and the Fundamental Research Funds for the Central Universities.

References

1. Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I.J., Bergeron, A., Bouchard, N., Warde-Farley, D., Bengio, Y.: Theano: new features and speed improvements. CoRR abs/1211.5590 (2012)
2. Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M., Liu, Y.: Semi-supervised learning for neural machine translation. In: Proceedings of 54th Annual Meeting of the Association for Computational Linguistics, pp. 1965–1974 (2016)
3. Erhan, D., Bengio, Y., Courville, A.C., Manzagol, P., Vincent, P., Bengio, S.: Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. **11**, 625–660 (2010)
4. Firat, O., Cho, K., Bengio, Y.: Multi-way, multilingual neural machine translation with a shared attention mechanism. In: Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 866–875 (2016)
5. Gu, J., Lu, Z., Li, H., Li, V.O.K.: Incorporating copying mechanism in sequence-to-sequence learning. In: Proceedings of 54th Annual Meeting of the Association for Computational Linguistics, pp. 1631–1640 (2016)
6. Jozefowicz, R., Zaremba, W., Sutskever, I.: An empirical exploration of recurrent network architectures. In: Proceedings of 32nd International Conference on Machine Learning, pp. 2342–2350 (2015)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of 26th Annual Conference on Neural Information Processing Systems, pp. 1106–1114 (2012)
8. Lu, Y., Salem, F.M.: Simplified gating in long short-term memory (LSTM) recurrent neural networks. CoRR abs/1701.03441 (2017)
9. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of 49th Annual Meeting of the Association for Computational Linguistics, pp. 142–150 (2011)
10. Nakagawa, T., Inui, K., Kurohashi, S.: Dependency tree-based sentiment classification using CRFs with hidden variables. In: Proceedings of 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 786–794 (2010)
11. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T.: Semi-supervised learning with ladder networks. In: Proceedings of 29th Annual Conference on Neural Information Processing Systems, pp. 3546–3554 (2015)
12. Raza, K.: Recurrent neural network based hybrid model for reconstructing gene regulatory network. Comput. Biol. Chem. **64**, 322–334 (2016)
13. Zhao, J., Liu, K., Xu, L.: Sentiment analysis: mining opinions, sentiments, and emotions. Comput. Linguist. **43**(3), 595–598 (2016)