Attention Modeling for Targeted Sentiment

Jiangming Liu and **Yue Zhang**
Singapore University of Technology and Design,
8 Somapah Road, Singapore, 487372
{jiangming_liu, yue_zhang}@sutd.edu.sg

Abstract

Neural network models have been used for target-dependent sentiment analysis. Previous work focus on learning a target specific representation for a given input sentence which is used for classification. However, they do not explicitly model the contribution of each word in a sentence with respect to targeted sentiment polarities. We investigate an attention model to this end. In particular, a vanilla LSTM model is used to induce an attention value of the whole sentence. The model is further extended to differentiate left and right contexts given a certain target following previous work. Results show that by using attention to model the contribution of each word with respect to the target, our model gives significantly improved results over two standard benchmarks. We report the best accuracy for this task.

1 Introduction

Targeted sentiment analysis investigates the classification of opinions polarities towards specific target entity mentions in given sentences (Jiang et al., 2011; Dong et al., 2014; Vo and Zhang, 2015; Tang et al., 2016; Zhang et al., 2016). The input is a sentence with given target entity mentions, and the output consists of two-way or three-way sentimental classes on each target mention. For example, the sentence "*She began to love **miley ray cyrus** since 2013 :)*" is marked with a positive sentiment label on the target "*miley ray cyrus*".

One important problem of targeted sentiment classification is how to model the relation between targets and their context. Earlier methods defined rich features by exploiting POS tags and syntactic structures (Jiang et al., 2011; Dong et al., 2014). Compared with discrete manual features, embedding features are less sparse, and can be learnt from large raw texts, capturing distributional syntactic and semantic information. Dong et al. (2014) use a target-specific recurrent neural network to represent a sentence. Vo and Zhang (2015) use the rich pooling functions to extract the feature vector for a given target.

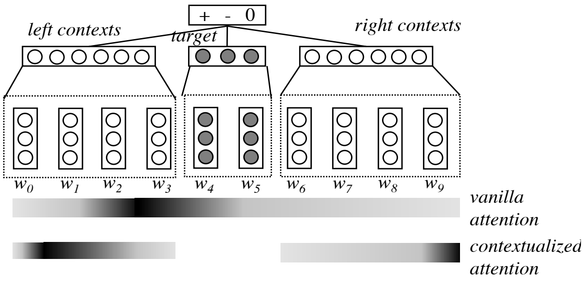


Figure 1: Structures of modeling target, left and right contexts and the attention over words.

One important contribution of Vo and Zhang (2015) is that they split a sentence into three sections including the target, its left contexts and its right contexts, as shown in Figure 1. Zhang et al. (2016) represent words in the input using a bidirectional gated recurrent neural network, and then use three-way gated neural network structure to model the interaction between the target and its left and right contexts. Tang et al. (2016) learn target-specific sentence representation by combining word embeddings with the corresponding targeted embeddings, and then using two recurrent neural networks to encode the left context and the right context, respectively.

The above methods use the different neural network structures to model the relation between contexts and targets, but they did not explicitly model the importance of each word in contributing to the sentiment polarity of the target. For example,

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 572–577,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

the sentence "*#nowplaying [lady gaga]*$_0$ *- let love down*" is neural for the target "*lady gaga*", where the contribution of "*love"* is little, despite that the word "*love*" is a positive word.

To address this, we utilize the attention mechanism to calculate the contribution of each word towards targeted sentiment classes, as shown in Figure 1, where the gray level in the spectrum means the contribution of words. In particular, we build a vanilla model using a bidirectional LSTM to extract word embeddings over the sentence and then apply attention over the hidden nodes to estimate the importance of each word. Furthermore, following Vo and Zhang (2015), Tang et al. (2016) and Zhang et al. (2016), we differentiate the left and right contexts given a target. Our final models give significantly improved results on two standard benchmarks compared to previous methods, resulting in best reported accuracy so far. Our source code is released at `https://github.com/LeonCrashCode/AttentionTargetSentiment`.

2 Related Work

Traditional sentiment classification methods rely on manual discrete features (Pang et al., 2002; Go et al., 2009; Mohammad et al., 2013). Recently, distributed word representation (Socher et al., 2013; Tang et al., 2014; Zhang et al., 2015) and neural network methods (Irsoy and Cardie, 2013; dos Santos and Gatti, 2014; Dong et al., 2014; Zhou et al., 2014; Zhang et al., 2016; Teng et al., 2016; Ren et al., 2016) have shown promising results on this task. The success of such work suggests that using word embeddings and deep neural network structures can automatically exploit the syntactic and semantic structures. Our work is in line with these methods.

The seminal work using the attention mechanism is neural machine translation (Bahdanau et al., 2015), where different weights are assigned to source words to implicitly learn alignments for translation. Subsequently, the attention mechanism has been applied into various other natural language processing tasks including parsing (Vinyals et al., 2015; Kuncoro et al., 2016; Liu and Zhang, 2017), document classification (Yang et al., 2016), question answering (He and Golub, 2016) and text understanding (Kadlec et al., 2016).

For sentiment analysis, the attention mechanism has been applied to cross-lingual sentiment (Zhou et al., 2016), aspect-level sentiment (Wang et al., 2016) and user-oriented sentiment (Chen et al., 2016). To our knowledge, we are the first to use the attention mechanism to model sentences with respect to targeted sentiments.

3 Models

We use a bidirectional LSTM to represent the input word sequence $w_0, w_1, ..., w_n$ as hidden nodes $h_0, h_1, ..., h_n$:

$$[h_0; ...; h_n] = \text{BILSTM}([w_0; ...; w_n]),$$

where the target is denoted as h_t, which is the average of word embeddings in the target phrase $[h_{t_0}; ...; h_{t_m}]$. We propose three variants of attention to model the relation between context words and targets.

3.1 Vanilla Model

We build a vanilla attention model by calculating a weighted value α over each word in sentences. The final representation of the sentence s is then given by[1]:

$$s = attention([h_0; ...; h_n], h_t) = \sum_i^n \alpha_i h_i,$$

where

$$\alpha_i = \frac{exp(\beta_i)}{\sum_j^n exp(\beta_j)}$$

and the weight scores β are calculated by using the target representation and the context word representation,

$$\beta_i = U^T tanh(W_1 \cdot [h_i; h_t] + b_1).$$

The sentence representation s is then used to predict the probability distribution p of sentiment labels on the target by:

$$p = softmax(W_2 s + b_2).$$

We refer to this vanilla model as BILSTM-ATT.

3.2 Contextualized Attention

We make two extensions to the vanilla attention method. The first is a contextualized attention model (BILSTM-ATT-C), where the sentence is divided into two segments with respect to the target, namely left context and right context (Vo and

[1] We only apply attention to non-target words.

Zhang, 2015; Tang et al., 2016; Zhang et al., 2016). Attention is applied on left and right contexts, respectively. In particular, the representation of the left context is:

$$s_l = attention([h_0; ...; h_{t_0-1}], h_t),$$

and the representation of the right context is:

$$s_r = attention([h_{t_m+1}; ...; h_n], h_t).$$

Together with the vanilla representation s, the distribution of sentiment labels is predicted by:

$$p = softmax(W_1 s + W_l s_l + W_r s_r + b_1).$$

3.3 Contextualized Attention with Gates

A second extension is to add gates to control the flow of context information (BILSTM-ATT-G). This is motivated by the fact that sentiment signals can be dominated by the left context, the right context or the entire sentence (Zhang et al., 2016). The three gates, z, z_l and z_r, controlled by the target and the corresponding context, are used.

$$\begin{aligned} z &\propto exp(W_1 s + U_1 h_t + b_1),\\ z_l &\propto exp(W_2 s_l + U_2 h_t + b_2),\\ z_r &\propto exp(W_3 s_r + U_3 h_t + b_3), \end{aligned}$$

where $z + z_l + z_r = \vec{1}$. The linear interpolation among s, s_l and s_r is formulated as

$$\tilde{s} = z \odot s + z_l \odot s_l + z_r \odot s_r.$$

Then the probability distribution of sentiment labels is predicted by:

$$p = softmax(W_4 \tilde{s} + b_4).$$

Training our models are trained to minimize a cross-entropy loss object with a l_2 regularization term, defined by

$$L(\theta) = -\sum_i log\, p_{t_i} + \frac{\lambda}{2}||\theta||^2,$$

where θ is the set of parameters, p_t is the probability of the ith training example given by the model and λ is a regularization hyper-parameter, $\lambda = 10^{-6}$. We use momentum stochastic gradient descent (Sutskever et al., 2013) with a learning rate of $\eta = 0.01$ for optimization.

T-Dataset	#target	#positive	#negative	#neutral
training	6248	1561	1560	3127
test	692	173	173	346

Z-Dataset	#target	#positive	#negative	#neutral
training	9489	2416	2384	4689
development	1036	255	272	509
test	1170	294	295	581

Table 1: Experimental corpus statistics.

Parameters	value
word dimension	200
LSTM hidden dimension	150
attention hidden dimension	100
dropout probability	0.5

Table 2: Hyper-parameter values.

4 Experiments

4.1 Data

We run experiments on two datasets, namely the benchmark training/test dataset of Tang et al. (2016) (T-Dataset) and the training/dev/test dataset of Zhang et al. (2016) (Z-Dataset), which consist of the MPQA corpus[2] and Mitchell et al. (2013)'s corpus[3]. Table 1 shows the corpus statistics. Both dataset are three-way classification data.

4.2 Parameters & Metrics

The hyper-parameters are given in Table 2[4]. We use GloVe vectors (Pennington et al., 2014) with 200 dimensions as pre-trained word embeddings, which are tuned during training. Two metrics are used to evaluate model performance: the classification accuracy and macro F1-measure over the three sentiment classes.

4.3 Development Experiments

We run three variants of targeted sentiment classification models on the development section of Z-Dataset to investigate the effectiveness of attention mechanism. A simple BILSTM without attention is deployed as our baseline. Table 3 shows the development results. We find that BILSTM-C gives a 0.6% accuracy improvement by differentiating the left and right contexts. However, surprisingly, BILSTM-G does not give much improvement despite using gates to control the contexts.

[2] http://mpqa.cs.pitt.edu/corpora/mpqa corpus/

[3] http://www.m-mitchell.com/code/index.html

[4] The hyper-parameters are set following previous works on twitter sentiment analysis.

Model	Accuracy	Macro F1
BILSTM	74.0	71.6
BILSTM-C	74.6	71.4
BILSTM-G	74.3	71.7
BILSTM-ATT	75.1	72.8
BILSTM-ATT-C	75.8	73.3
BILSTM-ATT-G	76.3	74.6

Table 3: Development results (%).

	T-testset		Z-testset	
Model	Acc	F1	Acc	F1
Jiang et al. (2011)	63.4	63.3	/	/
Dong et al. (2014)	66.3	65.9	/	/
Vo and Zhang (2015)	71.1	69.9	69.6	65.6
Tang et al. (2016)	71.5	69.5	/	/
Zhang et al. (2016)	72.0	70.9	71.9	69.6
BILSTM-ATT	72.4	70.5	73.5	70.6
BILSTM-ATT-C	72.5	70.9	74.1	71.3
BILSTM-ATT-G	**73.6**	**72.1**	**75.0**	**72.3**

Table 4: Final results (%).

This is different from the observation of Zhang et al. (2016), who find that gate mechanism improves accuracy without using attention. Finally, compared to baseline models without attention, our models give an average 1.2% accuracy improvement and a 1.8% macro F1 improvement. Our final model (BILSTM-ATT-G) gives a 2.3% accuracy significant improvement ($p < 0.01$ using t-test) and a 3.0% macro F1 improvement over the strongest baseline.

4.4 Final Results

We compare our models with previous work. The final results are shown in Table 4. Our final models outperform both Zhang et al. (2016) and Tang et al. (2016) by achieving 73.55% accuracy and 72.07% macro F1 on T-Dataset, and 75.04% accuracy and 72.29% macro F1 on Z-Dataset, respectively. Compared with Zhang et al. (2016), our final models have significant improvements ($p < 0.05$) on the Z-Dataset.

4.5 Analysis

We compare the performances of various models against OOV rates. In particular, we split the test sentences into two sets, where one contains sentences that have no OOV and the other consist of sentences which have at least one OOV. The results are shown in Figure 2. The BILSTM-ATT-G performs the best, especially on OOV sentences, which shows the robustness of the BILSTM-ATT-G.

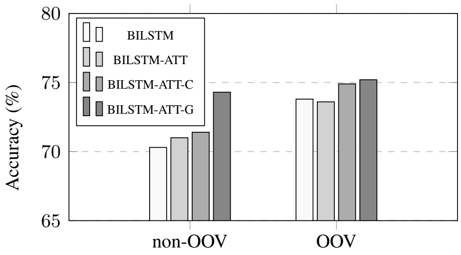


Figure 2: Accuracy against OOV rates.

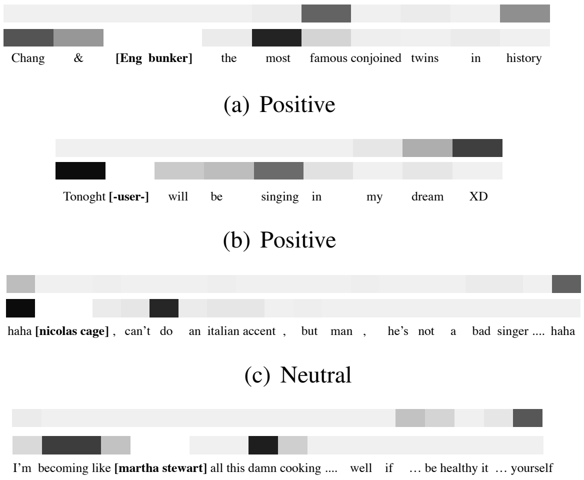
Model	Positive	Negative	Neural
BILSTM	61.0	69.9	79.4
BILSTM-ATT	61.4	71.1	79.7
BILSTM-ATT-C	60.5	73.2	80.2
BILSTM-ATT-G	64.7	70.8	81.4

Table 5: F1 scores (%) of each distinct polarity.

We compare the performances of various models on each distinct polarity. The results are shown in Figure 5. Interestingly, compared to BILSTM-ATT without contextualized attention, BILSTM-ATT-C loses accuracies on positive (-1.1%). However, BILSTM-ATT-G gives large improvements on positive (+4.2%) and neutral (+1.2%) targets but loses accuracy on negative (-2.4%). Overall, both BILSTM-ATT-C and BILSTM-ATT-G outperform BILSTM-ATT on neural cases, which account for 50% of all targets.

4.6 Examples

Figure 3 demonstrates the lexical weights given by BILSTM-ATT-G. The contribution of each word is visualized by the grey level, where high grey level means high contribution. The examples of Figure 3(a), Figure 3(b) and Figure 3(d) are consistent with the institution. The words "*most*", "*famous*", "*history*", "*XD*" lead to a positive label, while the word "*damn*" leads to a negative label. In Figure 3(c), although "*haha*" could be a positive word, here the sentimental class of the target is neutral. This can be explained by the fact that the word "*haha*" shows the happiness of the speaker instead of the target "*Nicolas Cage*". Figure 3(d) shows one example long sentence, where the left context dominates the sentiment. Applying attention mechanism into left and right context of the target is meaningful and beneficial.



(a) Positive

(b) Positive

(c) Neutral

(d) Negative

Figure 3: Attention visualization, where bold phrases are targets.

5 Conclusion

Prior work on targeted sentiment analysis investigates sentence representation that are target-specific but do not explicitly model the contribution of each word towards targeted sentiment. We investigated various attentional neural networks for targeted sentiment classification. Experiments demonstrated that attention over words is highly useful for targeted sentiment analysis. Our model gives the best reported results on two different benchmarks.

Acknowledgments

We thank the anonymous reviewers for their detailed and constructive comments. Yue Zhang is the corresponding author.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1650–1659. Association for Computational Linguistics.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54. Association for Computational Linguistics.

Cicero dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78. Dublin City University and Association for Computational Linguistics.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12.

Xiaodong He and David Golub. 2016. Character-level question answering with attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1598–1607. Association for Computational Linguistics.

Ozan Irsoy and Claire Cardie. 2013. Bidirectional recursive neural networks for token-level labeling with structure. *arXiv preprint arXiv:1312.0493*.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160. Association for Computational Linguistics.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 908–918. Association for Computational Linguistics.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A. Smith. 2016. What do recurrent neural network grammars learn about syntax? In *European Chapter of the Association for Computational Linguistics*.

Jiangming Liu and Yue Zhang. 2017. Shift-reduce constituent parsing with neural lookahead features. *Transactions of the Association of the Computational Linguistics*.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654. Association for Computational Linguistics.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, 2002. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, chapter Thumbs up? Sentiment Classification using Machine Learning Techniques.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Context-sensitive twitter sentiment classification using neural network. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 215–221.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, D. Christopher Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics.

Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. 2013. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1139–1147.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565. Association for Computational Linguistics.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307. The COLING 2016 Organizing Committee.

Zhiyang Teng, Tin Duy Vo, and Yue Zhang. 2016. Context-sensitive lexicon features for neural sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1629–1638. Association for Computational Linguistics.

Oriol Vinyals, Ł ukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2773–2781. Curran Associates, Inc.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 1347–1353.

Yequan Wang, Minlie Huang, xiaoyan zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Tin Duy Vo. 2015. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 3087–3093.

Shusen Zhou, Qingcai Chen, Xiaolong Wang, and Xiaoling Li. 2014. Hybrid deep belief networks for semi-supervised sentiment classification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1341–1349. Dublin City University and Association for Computational Linguistics.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Attention-based lstm network for cross-lingual sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 247–256. Association for Computational Linguistics.