Attention Transfer Network for Aspect-Level Sentiment Classification

Fei Zhao Zhen Wu Xinyu Dai
Natural Language Processing Laboratory
Nanjing University

November 2, 2020

Outline

Outline

Aspect-Level sentiment Classification

- Aspect-level sentiment classification(ASC) aims to infer the sentiment polarity (e.g. *positive*, *neutral*, *negative*) of the given target (also called aspect term) in a sentence.

Aspect-Level sentiment Classification

- Aspect-level sentiment classification(ASC) aims to infer the sentiment polarity (e.g. *positive*, *neutral*, *negative*) of the given target (also called aspect term) in a sentence.

Sentence: *I like the service but the food is bad.*
Target: service
Polarity: *positive*
Target: food
Polarity: *negative*

Figure 1: An examples in Aspect-Level sentiment dataset.

Aspect-Level sentiment Classification

Figure 2: An example of sentence containing multiple targets

- One main challenge of ASC is to separate different opinion contexts for different opinion targets.

Outline

Motivation & Approach

Motivation

- Most works employ the attention mechanism(Bahdanau et al., 2014) to capture the corresponding sentiment words of the opinion target, then aggregate them as evidence to infer the sentiment of the target.

Motivation & Approach

Motivation

- Most works employ the attention mechanism(Bahdanau et al., 2014) to capture the corresponding sentiment words of the opinion target, then aggregate them as evidence to infer the sentiment of the target.
- Aspect-level datasets are all relatively small-scale due to the complexity of annotation. Data scarcity causes the attention mechanism sometimes to fail to focus on the corresponding sentiment words of the target.

Motivation & Approach

Motivation

- Most works employ the attention mechanism(Bahdanau et al., 2014) to capture the corresponding sentiment words of the opinion target, then aggregate them as evidence to infer the sentiment of the target.
- Aspect-level datasets are all relatively small-scale due to the complexity of annotation. Data scarcity causes the attention mechanism sometimes to fail to focus on the corresponding sentiment words of the target.

Approach

- Despite the lack of ASC data, enormous labeled data of document-level sentiment classification (DSC) are available at online review sites such as Amazon and Yelp. These reviews contain substantial sentiment knowledge and semantic patterns.

Motivation & Approach

Motivation

- Most works employ the attention mechanism(Bahdanau et al., 2014) to capture the corresponding sentiment words of the opinion target, then aggregate them as evidence to infer the sentiment of the target.
- Aspect-level datasets are all relatively small-scale due to the complexity of annotation. Data scarcity causes the attention mechanism sometimes to fail to focus on the corresponding sentiment words of the target.

Approach

- Despite the lack of ASC data, enormous labeled data of document-level sentiment classification (DSC) are available at online review sites such as Amazon and Yelp. These reviews contain substantial sentiment knowledge and semantic patterns.
- we exploit attention knowledge from big-scale review sentiment classification datasets to assist attention process of aspect-level sentiment classification.

Outline

Attention Transfer Network

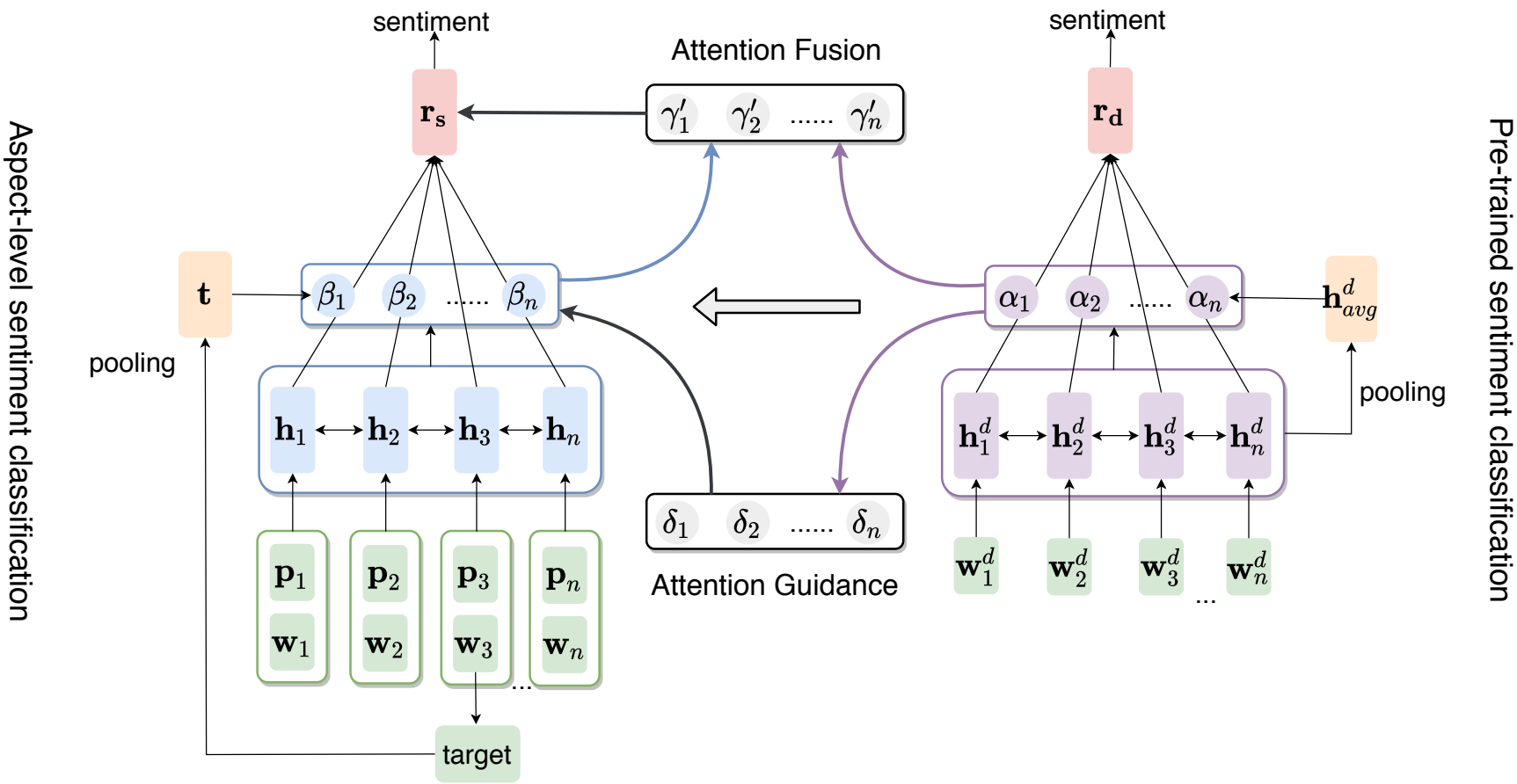

Figure 3: An illustration of our attention transfer network. The left one is the aspect-level sentiment classification, the right one is the pre-trained DSC module, and the middle part presents two proposed attention transfer approaches.

Task Formalization

ASC Formalization

- Given a sample $< s, t >$ from the ASC dataset $\mathcal{A}$, $s = \{w_1, w_2, ..., w_n\}$ is a review sentence consisting of n words and $t = \{w_l, w_{l+1}, ..., w_r\}$ is a given opinion target containing $|r - l|$ words. The opinion target t is a continuous subsequence of s.
- The goal of ASC is to predict the sentiment polarity (i.e., positive, neutral and negative) of the opinion target t in the sentence s.

DSC Formalization

- For a review document d from the DSC dataset $\mathcal{D}$, we regard it as a special long sentence $\{w_1^d, w_2^d, ..., w_n^d\}$ consisting of n words.
- DSC aims to determine the overall sentiment polarity of the review document d.

Pre-trainig DSC Module

- We employ a BiLSTM network to capture the contextual information for each word, and outputs a sequence of hidden vectors $\{h_1^d, h_2^d, \cdots, h_m^d\}$.
- The attention mechanism is employed to capture the global opinion words that are significant to sentiment classification.

$$\alpha_i = \frac{\exp(f(h_i^d, h_{avg}^d))}{\sum_{j=1}^{n} \exp(f(h_j^d, h_{avg}^d))}, \tag{1}$$

$$f(h_i^d, h_{avg}^d) = h_i^d \cdot W_d \cdot h_{avg}^d + b_d, \tag{2}$$

Base ASC Module: Attention-based BiLSTM

- Given a sentence s and an opinion target t in s, we employ a BiLSTM to generate target-aware context representations $\{h_1, h_2, \cdots, h_n\}$.
- We use the opinion target represenation $t = \sum_{i=l}^{r} h_i/(r-l)$ as query in the ASC task to extract target-dependent sentiment clues:

$$f(h_i, t) = h_i \cdot W_s \cdot t + b_s, \tag{3}$$

$$\beta_i = \frac{\exp(f(h_i, t))}{\sum_{j=1}^{n} \exp(f(h_j, t))}, \tag{4}$$

$$r_s = \sum_{i=1}^{n} \beta_i h_i, \tag{5}$$

- The target-specific representation r_s as the final feature for sentiment prediction and the model is trained by minimizing the cross entropy:

$$y_i = softmax(W_o v + b_o), \tag{6}$$

$$\mathcal{L}_o = -\sum_{i \in D} \hat{y}_i log(y_i). \tag{7}$$

Attention Guidance

- The attention mechanism of the ASC module cannot reach full potential due to limited training data, which means that the attention weights β_i may fail to capture target-relevant sentiment words.
- Sufficient DSC data enables the DSC module to extract sentiment words more accurately.

Attention Guidance

Nevertheless, there is a tiny gap between the attention weights α_i and β_i. Since the DSC task only detects the overall sentiment of a review, the sentiment words captured by α_i are global and target-irrelevant.

- To make up the gap, we use a heuristic method to transform target-irrelevant attention weight α_i into target-relevant weight δ_i:

$$\delta_i = \frac{1}{2^{(l_i-1)}}\alpha_i, \tag{8}$$

- We aim to approximate the distribution of attention β_i to δ_i. Intuitively, we apply KL (Kullback–Leibler divergence) as the optimized function, which describes the differences between distributions:

$$KL(\delta||\beta) = \sum_{i=1}^{n} \delta_i log\frac{\delta_i}{\beta_i}, \tag{9}$$

Attention Guidance

- What is more, the Eq. (9) can be further reduced as following:

$$\mathcal{L}_a = \sum_{i=1}^{n} \delta_i log \frac{\delta_i}{\beta_i}, \tag{10}$$

$$= \sum_{i=1}^{n} (\delta_i log \delta_i - \delta_i log \beta_i). \tag{11}$$

- where $\mathcal{L}_a$ represents the loss of attention guidance. Because δ_i contains fixed values, the equation is equal to

$$\mathcal{L}_a = \sum -\delta_i log \beta_i, \tag{12}$$

- Therefore, we adapt $\mathcal{L}_a$ into classification loss $\mathcal{L}_o$ to guide attention learning. Thus the final loss is defined as follows:

$$L = \mathcal{L}_o + \lambda^a \mathcal{L}_a. \tag{13}$$

where $\lambda^a \in (0,1]$ is hyperparameters that control the weight of $\mathcal{L}_a$.

Attention Fusion

Attention Guidance cannot leverage the attention weights from the DSC module during the testing stage and wastes the pre-trained knowledge.

- We design a fusion gate to integrate the global attention weight α_i and the target-dependent attention weight β_i, thereby generating more comprehensive and accurate attention weight γ'_i:

$$g = \sigma(W_g[\alpha_i; \beta_i]), \tag{14}$$

$$\gamma_i = g\alpha_i + (1 - g)\beta_i, \tag{15}$$

$$\gamma'_i = \frac{e^{\gamma_i}}{\sum_{i=1}^{n} e^{\gamma_i}}, \tag{16}$$

- Finally, we feed γ_i to Eq. (5) rather than β_i to obtain the target-specific representation of the sentence s.

Outline

Datasets

- We conduct experiments on two datasets, as shown in Table 4. They are from the SemEval 2014 Task 4 (Pontiki et al., 2014), which contains the reviews in laptop and restaurant, respectively.
- To pre-train the DSC module, we use the two datasets respectively from Yelp Review and Amazon Review, which are released by(Li et al., 2018a), each example is a sentence and is labeled as having either positive or negative sentiment.

Dataset	#Pos	#Neg	#Neu
Restaurant-Train	2164	807	637
Restaurant-Test	728	196	196
Laptop-Train	994	870	464
Laptop-Test	341	128	169

Figure 4: Statistics of the datasets.

Datasets	#positive	#negative	#total
Yelp Review	266,041	177,218	443,259
Amazon Review	277,228	277,769	554,997

Figure 5: Statistics of the two datasets Amazon Review and Yelp Review.

Overall Performance Comparison

Method	Restaurant		Laptop	
	Acc.	Macro-F1	Acc.	Macro-F1
Majority	65.00	33.33	53.50	33.33
Feature-SVM (Kiritchenko et al., 2014)	80.16	N/A	70.49	N/A
ATAE-LSTM (Wang et al., 2016)	77.20	N/A	68.70	N/A
TD-LSTM (Tang et al., 2016a)	78.00	66.73	71.83	68.43
IAN (Ma et al., 2017)	78.60	N/A	72.10	N/A
MemNet (Tang et al., 2016b)	80.32	N/A	72.37	N/A
RAM (Chen et al., 2017)	80.23	70.80	74.49	71.35
IARM (Majumder et al., 2018)	80.00	N/A	73.80	N/A
MGAN (Fan et al., 2018)	81.25	71.94	75.39	72.47
GCAE (Xue and Li, 2018)	77.43	66.24	71.03	64.43
TNet (Li et al., 2018b)	80.79	70.84	76.01	71.47
PRET+MULT (He et al., 2018)	79.98	69.39	74.14	69.14
TransCap (Chen and Qian, 2019)	80.72	71.98	74.92	70.21
Base ASC model	80.38	70.69	73.52	70.78
ATN-AG	81.39†	72.44†	76.41†	72.59†
ATN-AF	**82.36**†	**74.00**†	**76.48**†	**72.60**†

Figure 6: Main experiment results (%). The base ASC model is attention-based BiLSTM enhanced with position embedding. AT-AG and ATN-AF respectively refer to ATN model using *Attention Guidance* and *Attention Fusion*.

Effects of Attention Guidance and Attention Fusion

- **ATN−AG:** This method only incorporates the attention weights from Pre-trainig DSC Module into Base ASC Module via auxiliary learning signal.
- **ATN−AF:** We only fuse the attention of Pre-trainig DSC Module and Base ASC Module by a *merge gate*.

Summary

- Our attention transfer models ATN-AG and ATN-AF respectively achieve about 1% and 2% improvements in accuracy on the restaurant dataset, and over 2.8% improvements on the laptop dataset. These comparisons demonstrate the effectiveness of our proposal of explicitly transferring attention knowledge from resource-rich DSC data to the ASC task.
- Compared with ATN-AG, ATN-AF achieves better performance on the restaurant dataset. It is reasonable because ATN-AG cannot leverage the attention weights of the DSC module during the testing stage. Nevertheless, ATN-AG still obtains comparable results on the laptop dataset and has a faster inference speed than ATN-AF.

Effect of DSC Data Size

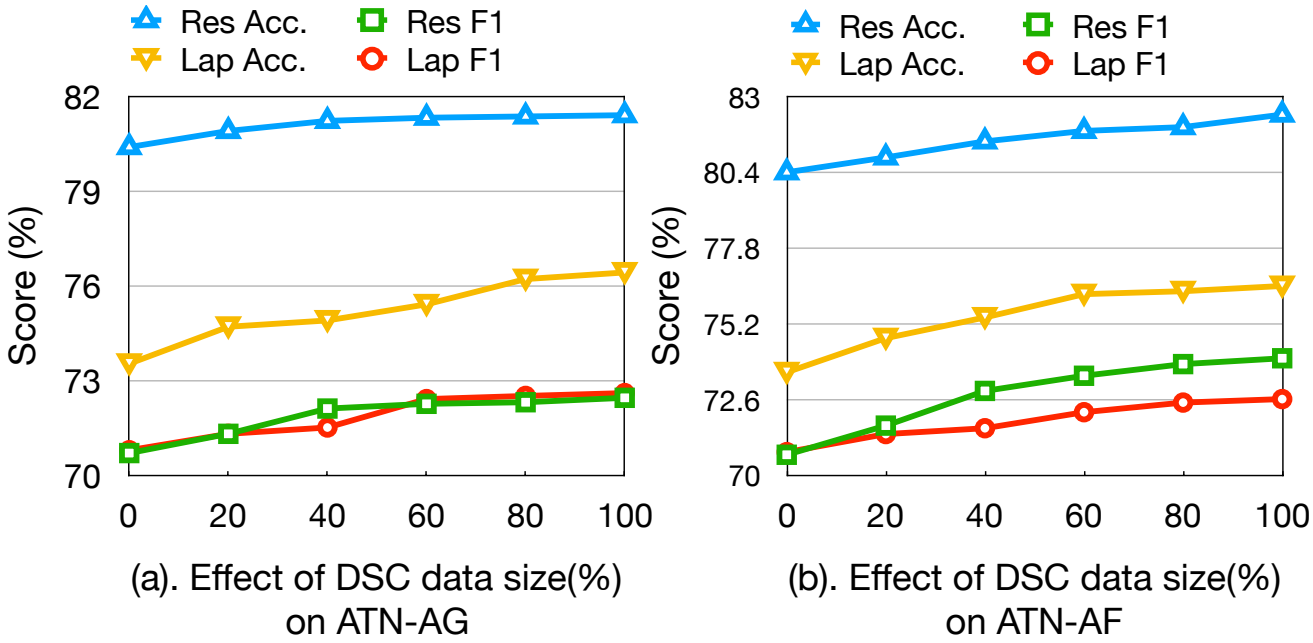

Figure 7: Performance of ATN-AG and ATN-AF with different percentages of DSC data.

Summary

- The changes in Accuracy and Macro-F1 on both datasets are shown in Figure 7. The improvements on Accuracy and Macro-F1 score with increasing number of training examples are stable across all datasets.

Effect of Hyper-parameter λ_a

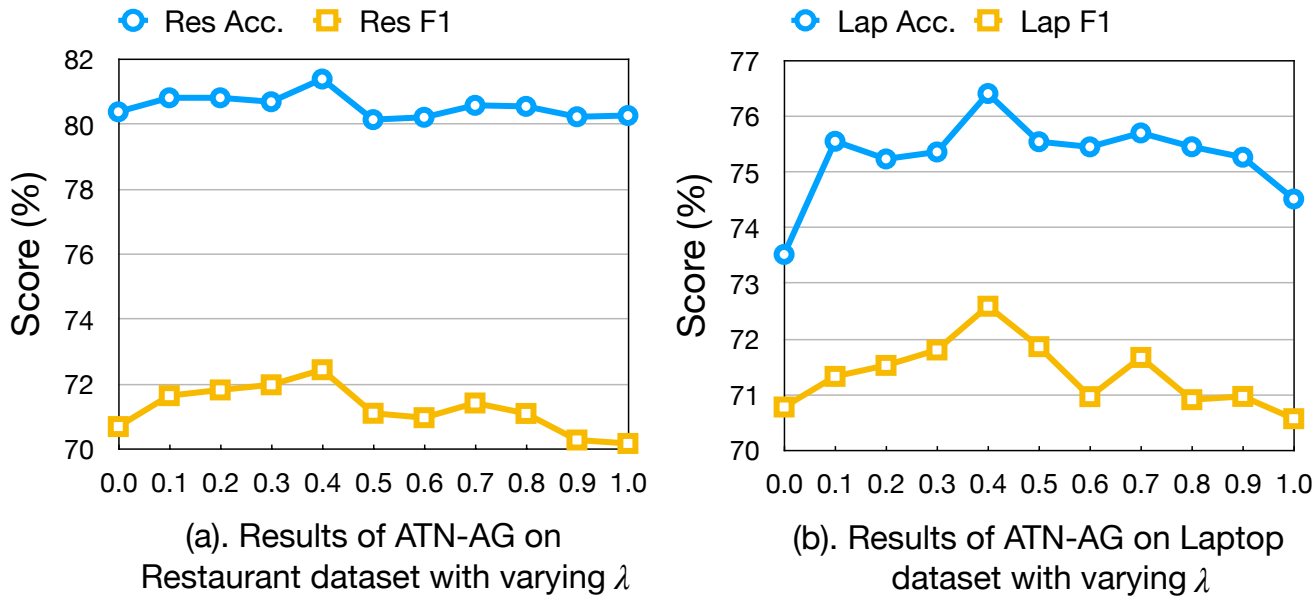

Figure 8: Effect of hyper-parameter λ on ATN-AG.

Summary

- To analyze the importance of the Attention Guidance(AG), we adjust λ_a in $[0, 1]$ to conduct experiments and the step is 0.1. Figure 8 shows the results of ATN-AG with different λ_a on laptop and restaurant datasets.

Case Study

Base model	I use it mostly for **[content creation]** (Audio , video , photo editing) and its reliable .	Neagtive✘
ATN-AG	I use it mostly for **[content creation]** (Audio , video , photo editing) and its reliable .	Positive✔
ATN-AF	I use it mostly for **[content creation]** (Audio , video , photo editing) and its reliable .	Positive✔
Base model	Did not enjoy the new Windows 8 and **[touchscreen functions]**	Positive✘
ATN-AG	Did not enjoy the new Windows 8 and **[touchscreen functions]**	Negative✔
ATN-AF	Did not enjoy the new Windows 8 and **[touchscreen functions]**	Negative✔

Figure 9: Visualization of ABN+AG and ABN+AF, the target is included in []

Summary

- In the first example, the base ASC model mainly focuses on the adverb "*mostly*". According to the statistics, the word "*reliable*" only appears five times in the training set. This indicates that the base model is not good at catching low-frequency sentiment words, thus makes wrong sentiment predictions.
- From the second example, we can see that the base ASC model mainly focuses on the word "*enjoy*" rather than the sentiment negator "*not*". It is hard for the base model to learn the negation with the insufficient labeled dataset.

Outline

Conclusion

- Insufficient labeled data limits the effectiveness of attention-based models for the ASC task.
- we propose a novel attention transfer framework, in which two different attention transfer methods are designed to exploit attention knowledge from resource-rich document-level sentiment classification corpus to enhance the attention process of resource-poor aspect-level sentiment classification.
- Experimental results indicate that our approaches outperform the state-of-the-art works. Further analysis validates the effectiveness and benefits of transferring the attention knowledge from DSC data for the ASC task.

End.

Thanks!

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.

Zhuang Chen and Tieyun Qian. 2019. Transfer capsule network for aspect level sentiment classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 547–556.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Exploiting document knowledge for aspect-level sentiment classification. *arXiv preprint arXiv:1806.04346*.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer

reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018a. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018b. Transformation networks for target-oriented sentiment classification. *arXiv preprint arXiv:1805.01086*.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, Md Shad Akhtar, Erik Cambria, and Asif Ekbal. 2018. Iarm: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3402–3411.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *COLING 2014*.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *COLING*.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *EMNLP*.

Yequan Wang, Minlie Huang, Li Zhao, et al. 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.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. *arXiv preprint arXiv:1805.07043*.