Don't Eclipse Your Arts Due to Small Discrepancies: Boundary Repositioning with a Pointer Network for Aspect Extraction

Zhenkai Wei[1] **Yu Hong**[1] ✉ **Bowei Zou**[2] **Meng Cheng**[1] **Jianmin Yao**[1]
[1]School of Computer Science and Technology, Soochow University, 1 Shizi, Suzhou, CHN
[2]Institute for Infocomm Research, 1 Fusionopolis Way, 08-01 Connexis, Singapore
{airzkwei, tianxianer✉}@gmail.com, zou_bowei@i2r.astar.edu.sg,
dcdream@outlook.com, jyao@suda.edu.cn

Abstract

The current aspect extraction methods suffer from boundary errors. These errors lead to a relatively minor difference between the extracted aspects and the ground-truth. However, they hurt the performance severely. In this paper, we propose to utilize a pointer network for repositioning the boundaries. Recycling mechanism is used which enables the training data to be collected without manual intervention. We conduct the experiments on the benchmark datasets SE14 of laptop and SE14-16 of restaurant. Experimental results show that our method achieves substantial improvements over the baseline, and outperforms state-of-the-art methods.

1 Introduction

Aspect extraction (Hu and Liu, 2004) is a crucial task in the field of real-world aspect-oriented sentiment analysis, where an aspect stands for a sequence of tokens which adhere to a specific sentiment word, in general, serving as the target on which people express their views. For example, the tokens "*twist on pizza*" is the aspect of the opinion "*healthy*" in 1). In this paper, we concentrate on the study of aspect extraction conditioned on the unawareness of sentiment words.

1) *Their twist on pizza is healthy.*
Ground-truth: *twist on pizza*
Predicted: [BOUND] *pizza* [BOUND]

2) *Buy the separate RAM memory and you will have a rocket.*
Ground-truth: *RAM memory*
Predicted: [BOUND] *separate RAM memory* [BOUND]

What is undoubtedly true is that the existing neural aspect extraction methods (Section 5.3) have achieved remarkable success to some extent. The peak performance on the benchmark datasets, to our best knowledge, is up to 85.61% F1-score (Li et al., 2018). We suggest that further improvements can be made by fine-tuning the boundaries of the extracted aspects. It is so because some incorrectly-extracted aspects result from minor boundary errors, where the boundaries refer to the start and end positions of a token sequence. For example, reinstating the omitted words "*twist on*" and trimming the redundant word "*separate*" in 1) and 2) by changing the start positions contributes to the recall of the correct aspects.

We propose to utilize a pointer network for repositioning the boundaries (Section 2). The pointer network is separately trained, and it is only used to post-process the resultant aspects output by a certain extractor (Section 3). Supervised learning is pre-requisite for obtaining a well-trained pointer network. However, so far, there is a lack of boundary-misspecified negative examples to construct the training set. Instead of manually labeling negative examples, we recycle those occurring during the time when the extractor is trained (Section 4). Our contributions in this paper are as follows:

- By means of a pointer network, we refine the boundary-misspecified aspects.
- The separately-trained pointer network serves as a post-processor and therefore can be easily coupled with different aspect extractors.
- The use of recycling mechanism facilitates the process of constructing the training set.

2 Pointer Network Based Boundary Repositioning

We train a pointer network to predict the start and end positions of the correct aspect. What we feed into the network include a candidate aspect and the sentence which contains the candidate (herein called source sentence). The candidate may be a

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3678–3684

boundary-misspecified aspect, truly-correct aspect or other text span. The network outputs two words w^s and w^e, one of which is predicted to be the start position, the other the end:

$$\begin{cases} w^s = \arg\max . p_s(w^s) \\ w^e = \arg\max . p_e(w^e) \end{cases} \quad (1)$$

where, P(*) denotes the probability that a word serves as the start or end position, and $\arg\max$ refers to the maximum likelihood estimation. The text span which lies between the start and end positions w^s and w^e will be eventually selected as the boundary-repositioned aspect.

It is noteworthy that, during testing, the status (boundary-misspecified, truly-correct or other) of the candidate aspect is assumed to be unknown. This is derived from the consideration of the practical situation in which the status of the pre-extracted aspect is unforeseeable.

Encoding Assume C=$\{w_1, ..., w_n\}$ represents the candidate aspect, where $w_i^c \in \mathbb{R}^l$ stands for the combination of the word, position and segment embeddings of the i-th token in C. The source sentence is represented in the same way and denoted by U=$\{w_1,..., w_m\}$. We concatenate C and U to construct the input representation:

$$\mathbb{W}_{C \oplus U} = [\text{CLS}, C, \text{SEP}, U, \text{SEP}] \quad (2)$$

where, CLS denotes the embedding of a dummy variable, while SEP is that of a separator (Devlin et al., 2019). In our experiments, WordPiece embeddings are used which can be obtained from the lookup table of Wu et al. (2016). The embeddings of position, segment, separator and dummy variable are initialized randomly.

We encode each element w_i in the input representation $\mathbb{W}_{C \oplus U}$ by fine-tuning BERT (Devlin et al., 2019): h_i=BERT(w_i), $i \in$[1, n+m+3].

Decoding Due to the use of the multi-head self-attention mechanism (Vaswani et al., 2017), BERT is able to perceive and more heavily weight the attentive words in the source sentence U, according to the information in the candidate aspect C, and vice versa. This property allows the attention-worthy words out of C to be salvaged and meanwhile enables the attention-unworthy words in C to be laid aside. On the other hand, a trainable decoder tends to learn the consistency between the ground-truth aspect and the attentive words. Therefore, we suppose that the decoder is able to leave the boundaries of C unchanged if C aligns with the ground-truth aspect, otherwise redefine the boundaries in U in terms of the attentive words.

Following the practice in prior research (Vinyals et al., 2015), we decode the representation h_i with a linear layer and the softmax function, where $W \in \mathbb{R}^{2 \times l}$ and $b \in \mathbb{R}^2$ are trainable parameters:

$$\begin{bmatrix} p_s(w_i) \\ p_e(w_i) \end{bmatrix} = \text{softmax}(W h_i + b) \quad (3)$$

Training Our goal is to assign higher probabilities to the start and end positions $\hat{w}^s$ and $\hat{w}^e$ for all the ground-truth aspects in the training set. Therefore, we measure loss by calculating the average negative log-likelihood for all pairs of $\hat{w}^s$ and $\hat{w}^e$:

$$L_B = -\frac{1}{N_B} \sum_{i=1}^{N_B} \left[\frac{\log p_s(\hat{w}_i^s) + \log p_e(\hat{w}_i^e)}{2} \right] \quad (4)$$

where, N_B is the number of ground-truth aspects. During training, we obtain the parameters W and b in equation (3) by minimizing the loss L_B.

3 BiLSTM-CRF based Pre-Extraction

We use the pointer network to post-process the pre-extracted aspects (which are referred to the candidate aspects in section 2). In our experiments, we employ a BiLSTM-CRF model to obtain the candidate aspects.

In this case, we solve aspect pre-extraction as a sequence labeling task. BIO labeling space y=$\{B$, I, $O\}$ (Xu et al., 2018) is specified as the output for each token in the source sentence, in which B, I and O respectively signal the beginning of an aspect, inside of an aspect and non-aspect word.

First of all, we represent the tokens in the source sentence using GloVe embeddings (Pennington et al., 2014). On the basis, we use a bidirectional recurrent neural network with Long-Short Term Memory (BiLSTM for short) (Liu et al., 2015) to encode each token, so as to obtain the initial hidden state vector h_i^{lstm}. Self-attention mechanism (Vaswani et al., 2017) is utilized for the resolution of long-distance dependency, by which we obtain the attention-weighted hidden state h_i^{att}. We concatenate h_i^{lstm} and h_i^{att} to produce the final feature vector for the i-th token: $\hat{h}_i$=$h_i^{\text{lstm}} \oplus h_i^{\text{att}}$.

Conditioned on the feature vector $\hat{h}_i$ emitted by BiLSTM with attention, we estimate the emission probabilities that the i-th token may serve as B, I and O respectively. The fully-connected dense

layer is used to map $\hat{h}_i$ to the BIO labeling space: $p_i(BIO) = f_{den}(\hat{h}_i)$. Over the emission probabilities of all the tokens in the source sentence, we utilized a linear-chain Conditional Random Field (CRF) (Wang et al., 2016) to predict the optimum label sequence of BIO. Eventually, the tokens labeled with B and I will be taken as the aspects.

We train the extractor by maximizing the log-likelihood of sequence labeling (Luo et al., 2019):

$$L_E = \sum_{i=1}^{N_E} \log P(y|f_{den}(\hat{h}_i), \hat{W}, \hat{b}) \quad (5)$$

where, N_E denotes the number of tokens in the training set, $\hat{W}$ is a trainable parameter which plays a role of transition matrix in CRF and $\hat{b}$ is the bias.

4 Recycling Mechanism

The extractor can be trained on the benchmark datasets provided by the SemEval tasks (Pontiki et al., 2016). However, it is impractical to separately train the positioner because there is a lack of boundary-misspecified negative examples. To solve the problem, we recycle the negative examples occurring during the training of the extractor.

We define a negative example to be a text span which partially overlaps with the ground-truth aspect. The text spans which are completely inconsistent with the ground-truth are not considered. For example, "*Fresh ingrediants*" in 3) is an eligible negative example, but "*super tasty*" is ineligible.

3) *Fresh ingrediants and super tasty*.
Ground-truth: *ingrediants*
Eligible: *Fresh ingrediants*
Ineligible: *super tasty*

We maintain a table that maps each ground-truth aspect to a list of negative examples. We initialize the mapping table by taking ground-truth aspects as entries and assigning an empty list to each of them. For each entry, we traverse the results output by the extractor in each training epoch and pick up the eligible negative examples. The newly-observed negative examples will be added to the list of the entry only if they have not yet been included in the list. We perform recycling in the first 20 epochs. Few examples can be found in the subsequent epochs.

5 Experimentation

5.1 Datasets

We evaluate the proposed methods on the laptop and restaurant datasets provided by SemEval 2014-2016 aspect-based sentiment analysis tasks (SE14-16 for short) (Pontiki et al., 2014, 2015, 2016). For comparison purpose, we follow the previous work to randomly select 20% of the official training data to form the validation set.

Table 1 shows the sample statistics in the training, validation and test sets as well as that of the recycled negative examples (denoted by Neg).

Dataset	Training		Validation	Test
	Aspect	Neg	Aspect	Aspect
SE14-L	1,853	2,008	505	654
SE14-R	2,961	3,208	733	1,134
SE15-R	966	1,050	234	542
SE16-R	1,398	1,424	346	612

Table 1: Sample statistics for SE14-16. "*L*" indicates the *laptop* domain and "*R*" the *restaurant*.

5.2 Hyperparameter Settings

For the aspect pre-extraction model, we initialize all word embeddings by 100-dimensional GloVe word embeddings (Pennington et al., 2014). Each of BiLSTM units is of 100 dimensions and the number of hidden states in the self-attention layer is set to 200. We employ dropout on the output layer of BiLSTM (i.e., penultimate layer) and the dropout rate is set to 0.5. The learning rate for parameter updating is set to 1e-3.

For the boundary reposition model, we employ basic BERT (Devlin et al., 2019) as the encoder which contains 12 transformer encoding blocks. Each block holds 768 hidden units and 12 self-attention heads. During training, the maximum length of the input sequence is set to 180 and the batch size is set to 10. The learning rate is set to 3e-5 and the number of training epochs is set to 5.

5.3 Compared Models

We compare with the state-of-the-art models. By taking learning framework as the criterion, we divide the models into two classes:

Single-task Learning In the family of aspect-oriented single-task learning, the traditional **CRF**[1] is used at the earliest time which is based on feature engineering. On the basis, **HIS_RD** (Chernyshevich, 2014) additionally utilizes the part-of-speech and named entity features. **NLANGP** (Toh and Su, 2016) first incorporates syntactic features and word embeddings. **HIS_RD** and **NLANGP** top

[1] https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html

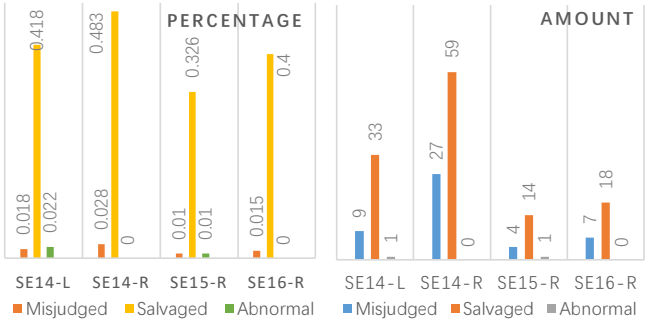

Figure 1: Amounts of the salvaged and misjudged aspects, and the percentages in all the samples

the list for aspect extraction in 2014 and 2016 SemEval challenges. During the period, **WDEmb** (Yin et al., 2016) enhances word embeddings using the linear context. And Liu et al. (2015)'s work may be the first attempt to directly use vanilla **LSTM** for aspect analysis. Soon afterwards, Xu et al. (2018) construct a multi-layer Convolution Neural Network (**DE-CNN**) which integrates GloVe and domain-specific embeddings. Ma et al. (2019) first use Sequence-to-Sequence learning (**Seq2Seq4ATE**) with GRUs and the position-aware attention mechanism this year.

Multi-task Learning For aspect-oriented multi-task learning, Li and Lam (2017) design a triple-LSTM model (**MIN**) to share the features which are generated toward extraction and classification tasks. **CMLA** (Wang et al., 2017) uses a multi-layer attention mechanism for the joint extraction of aspect terms and sentiment words. **HAST** (Li et al., 2018) strengthens the joint model using truncated history-attention and selective transformation network. **RINANTE** (Dai and Song, 2019) shares features in the bottom layer of BiLSTM-CRF and uses distant supervision to expand the training data.

Similar to RINANTE, our aspect pre-extraction model (Baseline) is based on BiLSTM-CRF. However, we force it to work in the single-task learning framework. More importantly, instead of distant supervision, we use recycling mechanism to acquire local boundary-misspecified examples, and instead of retraining BiLSTM-CRF for use, we only reposition the boundaries of the resultant aspects.

5.4 Main Results

We show the performance difference over test sets in Table 2. It can be observed that the single-task BiLSTM-CRF based extractor either achieves a comparable performance to some of the current state-of-the-art methods, or performs worse than others. Nevertheless, refining the pre-extracted aspects by boundary repositioning yields substantial improvements and achieves the best performance.

Method	SE14-L	SE14-R	SE15-R	SE16-R
CRF	72.77	79.72	62.67	66.96
HIS-RD (2014)	74.55	79.62	-	-
LSTM (2015)	75.71	82.01	68.26	70.35
NLANGP (2016)	-	-	67.12	72.34
WDEmb (2016)	75.16	84.97	69.73	-
DE-CNN (2018)	81.59	85.20	68.28	74.37
Seq2Seq (2019)	80.31	-	-	75.14
MIN (2017)	77.58	-	-	73.44
CMLA (2017)	77.80	85.29	70.73	-
HAST (2018)	79.52	85.61	71.46	73.61
RINANTE (2019)	73.47	84.06	66.17	-
BiSELF-CRF (ours)	78.15	83.73	68.81	73.49
+Repositioning	**81.90**	**86.58**	**71.72**	**75.56**

Table 2: Performance (F-scores) comparison

Figure 1 provides further insight into the test results. It shows that there are 41% of boundary-misspecified aspects in average can be successfully salvaged. On the contrary, there are only 1.7% of correctly-extracted aspects in average have been misjudged. Besides, there are few completely erroneous extraction results can be rectified.

5.5 Adaptation to BERT

In a separate experiment, we examine the adaptation performance of boundary repositioning. The original pre-extraction model is replaced by the fine-tuning BERT and a more sophisticated model. The former is coupled with a dense layer and a softmax layer. The latter is constructed by coupling the fine-tuning BERT and the BiSELF-CRF network. On the contrary, the set of negative examples which are recycled in the earlier experiment remains unchanged. Table 3 shows the test results. It can be observed that boundary repositioning still achieves considerable improvements in performance. This demonstrates the robust adaptation ability.

Method	SE14-L	SE14-R	SE15-R	SE16-R
BERT (fine-tuning)	78.48	85.49	69.49	74.98
+Repositioning	**81.43**	**87.10**	**72.68**	**77.71**
BERT+BiSELF-CRF	80.15	85.60	66.64	75.64
+Repositioning	**82.68**	**87.11**	**70.23**	**77.51**

Table 3: Test results (F-scores) for adaptation analysis

Method	SE14-L	SE14-R	SE15-R	SE16-R
DE-CNN(reported)	81.59	85.20	68.28	74.37
DE-CNN(retrained)	82.09	80.07	66.40	74.09
+repositioning	**84.17**	**84.55**	**72.03**	**75.40**

Table 4: Performance (F-scores) achieved by coupling the retrained DE-CNN with boundary repositioning

5.6 Cooperation with the State-Of-The-Art

We tend to verify whether boundary repositioning can cooperate with the existing methods. Considering that DE-CNN (Xu et al., 2018) has a competitive advantage, we take it in this case study. We utilize DE-CNN for pre-extracting aspects and conduct boundary repositioning over the resultant aspects. The following notes needs to be considered if one tends to conduct a similar experiment.

- Both the source code of Xu et al (2018)'s DE-CNN and the preprocessed input data in SE14-L and SE16-R are publicly available. Conditioned on the input data, the retrained DE-CNN obtains similar performance to that reported in Xu et al (2018)'s study.

- Dai et al (2019) reported the performance of DE-CNN on SE14-R and SE15-R. However, it wasn't mentioned whether Xu et al (2018)'s open-source DE-CNN was used or it was reproduced. We retrained Xu et al (2018)'s open-source DE-CNN and preprocessed the input data in SE14-R and SE15-R all over again. The obtained performance on the datasets are worse than that reported in Dai et al (2019)'s work.

Table 4 shows the performance of DE-CNN, including the reported performance in Xu et al (2018) and Dai et al (2019)'s work, that of the retrained DE-CNN, as well as the one coupled with boundary repositioning. It can be observed that boundary repositioning yields substantial improvements over the retrained DE-CNN on all the four datasets. Compared to the reported performance, the use of boundary repositioning also results in significant improvements on SE14-L, SE 15-R and SE16-R.

Method	P-value
BiSELF-CRF vs BiSELF-CRF+repositioning	0.0017
DE-CNN vs DE-CNN+repositioning	0.0222

Table 5: Test results for significance analysis

5.7 Statistical Significance

We follow Johnson (1999) to use the sampling-based P-values for examining the significance. Johnson (1999) suggest that the ideal threshold of P-value is 0.05. It indicates that a system achieves significant improvements over others only if P-values are less than 0.05, otherwise insignificant. Besides, it has been proven that the smaller the P-value, the higher the significance (Dror et al., 2018).

We form the updated versions of BiSELF-CRF and DE-CNN by coupling them with boundary repositioning. On the basis, we compute P-values by comparing the extraction results of the two models to that of the updated versions. Table 5 shows the P-values. It can be observed that the P-values are much lower than the threshold. This demonstrates that boundary repositioning produces significant improvements.

In brief, we prove that boundary repositioning can be used as a reliable post-processing method for aspect extraction. The source code of boundary repositioning to reproduce the above experiments has been made publicly available. We submit the source code and instruction along with this paper.

6 Conclusion

Our experimental results demonstrate that boundary repositioning can be used as a simple and robust post-processing method to improve aspect extraction. Our findings reveal that illustrative aspects in scientific literature are generally long-winded. Extracting these aspects suffers more severely from boundary errors. In the future, we will develop a syntax-based multi-scale graph convolutional network to deal with both short and long aspects.

Acknowledgments

We thank the reviewers for their insightful comments. The idea is proposed by the corresponding author (Yu Hong). Yuchen Pan provides an effective assistance for conducting the experiments. We thank the colleagues for their help.

This work was supported by the national Natural Science Foundation of China (NSFC) via Grant Nos. 61672368, 61672367, 61703293.

References

Maryna Chernyshevich. 2014. IHS r&d belarus: Cross-domain extraction of product features using CRF. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 309–313.

Hongliang Dai and Yangqiu Song. 2019. Neural aspect and opinion term extraction with mined rules as weak supervision. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5268–5277.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177.

Douglas H Johnson. 1999. The insignificance of statistical significance testing. *The journal of wildlife management*, pages 763–772.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4194–4200.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2886–2892.

Pengfei Liu, Shafiq R. Joty, and Helen M. Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1433–1443.

Huaishao Luo, Tianrui Li, Bing Liu, and Junbo Zhang. 2019. DOER: dual cross-shared RNN for aspect term-polarity co-extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 591–601.

Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. Exploring sequence-to-sequence learning in aspect term extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3538–3547.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 486–495.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35.

Zhiqiang Toh and Jian Su. 2016. NLANGP at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 282–288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2692–2700.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 616–626.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3316–3322.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 592–598.

Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised word and dependency path embeddings for aspect term extraction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2979–2985.