Trading Syntax Trees for Wordpieces: Target-oriented Opinion Words Extraction with Wordpieces and Aspect Enhancement

Samuel Mensah
Computer Science Department
University of Sheffield, UK
s.mensah@sheffield.ac.uk

Kai Sun
BDBC and SKLSDE
Beihang University, China
sunkai@buaa.edu.cn

Nikolaos Aletras
Computer Science Department
University of Sheffield, UK
n.aletras@sheffield.ac.uk

Abstract

State-of-the-art target-oriented opinion word extraction (TOWE) models typically use BERT-based text encoders that operate on the word level, along with graph convolutional networks (GCNs) that incorporate syntactic information extracted from syntax trees. These methods achieve limited gains with GCNs and have difficulty using BERT wordpieces. Meanwhile, BERT wordpieces are known to be effective at representing rare words or words with insufficient context information. To address this issue, this work trades syntax trees for BERT wordpieces by entirely removing the GCN component from the methods' architectures. To enhance TOWE performance, we tackle the issue of aspect representation loss during encoding. Instead of solely utilizing a sentence as the input, we use a sentence-aspect pair. Our relatively simple approach achieves state-of-the-art results on benchmark datasets and should serve as a strong baseline for further research.

1 Introduction

Target-oriented opinion word extraction (TOWE; Fan et al. (2019)) is a subtask in aspect-based sentiment analysis (ABSA; Pontiki et al. (2014b)), which aims to identify words that express an opinion about a specific target (or aspect) in a sentence. For instance, in the sentence "Such an *awesome* **surfboard**.", a TOWE model should identify *"awesome"* as the opinion word for the given aspect **surfboard**. TOWE provides explicit aspect-opinion pairs which can be used to improve results in downstream tasks such as opinion summarization (Kim et al., 2011) and information extraction (Pontiki et al., 2014b; Tang et al., 2016; Sun et al., 2023).

1. **Sentence**:	Such an *awesome* **surfboard**
Wordpieces:	'such', 'an', *'awesome'*, **'surf'**, **'##board'**
2. **Sentence**:	A *great* **snowboard** which holds edges well when riding on snow.
Wordpieces:	'A', *'great'*, **'snow'**, **'##board'**, 'which', 'holds', 'edges', 'well', 'when', 'riding', 'on', 'snow'.

Table 1: Sentences demonstrating contextual understanding through shared wordpieces. The table shows each sentence and its corresponding BERT wordpiece sequence. Aspect words are bold-typed and opinion words are italicized. The shared wordpiece ’##board’ helps in decoding the meaning of "surfboard".

Currently, many TOWE methods (Veyseh et al., 2020; Chen et al., 2020; Jiang et al., 2021; Feng et al., 2021; Mensah et al., 2021) use pretrained BERT (Devlin et al., 2018), to encode the input sentence. BERT has the ability to effectively capture context, which can improve TOWE performance. However, many of these methods are rather complex, as they often incorporate syntax tree information using a graph convolutional network (GCN) (Kipf and Welling, 2017). For instance, Veyseh et al. (2020) uses an ordered-neuron LSTM (Shen et al., 2018) encoder with a GCN while Jiang et al. (2021) applies an attention-based relational GCN on the syntax tree. Mensah et al. (2021) applies a BiLSTM (Hochreiter and Schmidhuber, 1997) on BERT embeddings and incoporate syntax information via a GCN.

While incorporating syntax information through GCNs has been shown to provide some performance gains in TOWE, these are usually limited (Mensah et al., 2021). Moreover, modeling subword tokens with a GCN can be challenging because the syntax tree consists of whole words rather than subword tokens like wordpieces (Schuster and Nakajima, 2012; Devlin et al., 2018). Models based on subword tokens strike a good balance between character- and word-based encoders. They are able to effectively learn representations of rare words or words with insufficient context information. Consider the example in Table 1. The context

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 999–1007
July 9-14, 2023 ©2023 Association for Computational Linguistics

information for “surfboard” is limited, making it difficult to understand its meaning without additional context. However, both aspects share the wordpiece “##board”, which allows the meaning of “surfboard” to be partially understood by using information from the context of “snowboard”. In this case, “riding” is related to both aspects through the shared wordpiece, enabling the representation of “surfboard” to be improved.

In this paper, we propose a substantial simplification for syntax-aware TOWE models (Veyseh et al., 2020; Jiang et al., 2021; Mensah et al., 2021) by replacing the syntax tree with subword information while maintaining good prediction performance. This is accomplished by removing the GCN from these architectures and using BERT wordpieces instead. Additionally, we address the issue of aspect representation degradation during encoding. This degradation negatively affects TOWE performance by reducing the availability of semantic information about the aspect for determining the opinion words to extract. To solve this problem, we propose using a sentence-aspect pair as input rather than just a sentence, similar to the approach used by Tian et al. (2021) for aspect-based sentiment classification. Through extensive experimentation, we found that our simple approach achieves state-of-the-art (SOTA) results by outperforming the method proposed by Mensah et al. (2021) without the need of a GCN component.

2 Task Formalization

The TOWE task aims to identify an opinion word in a sentence $S = \{w_1, \ldots, w_{n_s}\}$ with respect to an aspect $w_a \in S$. The sentence is typically tokenized into a sequence of tokens at different levels of granularity (e.g. subwords or whole words), $T = \{t_1, \ldots, t_{n_t}\}$, with $t_a \in T$ denoting a subsequence of the aspect w_a and $n_s \leq n_t$. The goal is to assign one of three tags (I, O, or B) to each token using the IOB format (Ramshaw and Marcus, 1995), which indicates whether the word is at the Inside, Outside or Beginning of the opinion word relative to the aspect.

3 Syntax-aware Approaches to TOWE

Typically, syntax-aware approaches to TOWE (Veyseh et al., 2020; Jiang et al., 2021; Mensah et al., 2021) employ a text encoder that utilizes pretrained BERT (Devlin et al., 2018) and position embeddings (Zeng et al., 2014) (or category embeddings (Jiang et al., 2021)) to learn whole word representations that are aware of the aspect’s location in text. These approaches also include a GCN that operates on a syntax tree in order to incorporate syntactic information into the model.

Ordered-Neuron LSTM GCN (ONG): Veyseh et al. (2020) combine an ordered neuron LSTM (ON-LSTM; Shen et al. (2018)) and a GCN for TOWE. The ON-LSTM layer is an LSTM variant that considers the order of elements in the input sequence (including BERT and position embeddings) when modeling dependencies between them. The GCN encodes syntactic structural information into the representations obtained by the ON-LSTM layer.

BERT+BiLSTM+GCN: Mensah et al. (2021) replaces the ON-LSTM of the ONG model with a BiLSTM to better capture short-term dependencies between aspect and opinion words.

Attention-based Relational GCN (ARGCN): Jiang et al. (2021) combine contextualized embedding obtained using BERT with a category embedding (i.e., IOB tag embedding) to incorporate aspect information. They subsequently use a relational GCN (Schlichtkrull et al., 2018) and BiLSTM to respectively incorporate syntactic and sequential information for TOWE classification.

4 Trading Syntax Trees for Wordpieces

Mensah et al. (2021) have recently demonstrated that the use of a GCN to incorporate syntax tree information has little impact in TOWE model performance. Meanwhile, the GCN presents challenges when using subword tokens, as previously mentioned. Therefore, we propose a simplified version of the TOWE model that omits the GCN component from syntax-aware approaches and instead uses subword tokens as the input to the BERT component. In this work, we use BERT’s Wordpieces (Devlin et al., 2018) as the subword representation because they are highly informative, having been derived from the BERT pretraining process. However, methods such as Byte-Pair Encoding (BPE) (Sennrich et al., 2016) can also be used, as we will see later in the experiments.

4.1 Formatting BERT Input

Given sentence S, the BERT wordpiece tokenizer segments S into a sequence of wordpieces $T =$

Models	Granularity	Lap14	Res14	Res15	Res16	Avg
ONG	word	75.77	82.33	78.81	86.01	80.73
ONG w/o GCN	word	74.17	84.10	78.33	84.87	80.37
ONG(S) w/o GCN	wordpiece	79.79	86.63	80.72	88.30	83.86
ONG(S,A) w/o GCN	wordpiece	81.70	**88.70**	**82.55**	91.18	86.03
ARGCN	word	76.36	85.42	78.24	86.69	81.68
ARGCN w/o R-GCN	word	76.38	84.36	78.41	84.61	80.94
ARGCN(S) w/o R-GCN	wordpiece	80.08	85.92	81.36	89.72	84.27
ARGCN(S,A) w/o R-GCN	wordpiece	81.37	88.18	82.49	90.82	85.72
BERT+BiLSTM+GCN	word	78.82	85.74	80.54	87.35	83.11
BERT+BiLSTM	word	78.25	85.60	80.41	86.94	82.80
BERT+BiLSTM(S)	wordpiece	80.45	86.27	80.89	89.80	84.35
BERT+BiLSTM(S,A)	wordpiece	**82.59**	88.60	82.37	**91.25**	**86.20**

Table 2: F1 performance of syntax-aware methods and their variants. "Avg" refers to the average F1 score calculated across all of the datasets. “Granularity” highlights the level of granularity at which input tokens are represented.

$\{t_1, t_2, \ldots, t_{n_t}\}$. The BERT input for S is then formatted as follows:

$$T^{(S)} = \{[\text{CLS}], T, [\text{SEP}]\} \qquad (1)$$

where [CLS] and [SEP] are special tokens that mark the boundaries of the sentence.

While this format may be adequate for some NLP tasks, it can be problematic for learning good aspect representations in aspect-based sentiment classifica- tion (Tian et al., 2021). To mitigate this issue, we adopt the approach of Tian et al. (2021) and reformat the BERT input by using a sentence-aspect pair $T^{(S,A)}$, which combines $T^{(S)}$ and t_a (i.e. the aspect subsequence) along with special tokens.

$$T^{(S,A)} = \{[\text{CLS}], T, [\text{SEP}], t_a, [\text{SEP}]\} \qquad (2)$$

4.2 Classification and Optimization

The input $T^{(S,A)}$ consists of two parts: $T^{(S)}$ and t_a. Since t_a only serves to enhance the aspect representation in $T^{(S)}$, sequence labeling is done on $T^{(S)}$ only. During sequence labeling, we follow the common approach of predicting based on the first wordpiece representation of a word. For instance, given the word “surfboard” that consists of the wordpieces “surf” and “##board” which both are learned during encoding, only the representation of “surf” is fed to a softmax classifier to predict the tag for the whole word. The cross-entropy function is minimized for each word in the training set.

5 Experiments and Results

We experiment with the following baselines: ARGCN, BERT+BiLSTM+GCN and ONG. We use the suffixes (S) or (S,A) to indicate whether the modified versions of these methods uses a wordpiece sentence or wordpiece sentence-aspect pair as input, respectively. We used the publicly available code and optimal hyperparameter settings from the authors of ARGCN[1] and BERT+BiLSTM+GCN.[2] We have implemented ONG model variants ourselves using the suggested hyperparameter configurations from the authors.[3] Following previous work (Fan et al., 2019), we use the same experimental setup and evaluate on the Laptop dataset (Lap14) and the Restaurant datasets (Res14, Res15, Res16) (Pontiki et al., 2014a, 2015, 2016). The result reported for each dataset is the average over Micro F1 scores obtained from five different runs. Each run uses a different random seed to ensure the stability of our results.

5.1 F1 Performance Comparison

The results, shown in Table 2, indicate that removing the GCN component from syntax-aware approaches does not substantially impact their performance, with average decreases in performance of 0.36, 0.74, and 0.31, respectively. However, we observed a large improvement in model performance when using wordpieces, as indicated by the models with the (S) suffix. It is possible that BERT captures enough syntax information already (Clark et al., 2019) and, therefore, using GCNs to exploit syntax trees does not substantially improve

[1] https://github.com/samensah/encoders_towe_emnlp2021

[2] https://github.com/wcwowwwww/towe-eacl

[3] https://github.com/samensah/Towe-TradeSyntax4WP

Model	Lap14	Res14	Res15	Res16	Avg
BERT-BiLSTM(S)	80.45	86.27	80.89	89.80	84.35
-Mask Aspect	80.01	86.11	80.42	88.59	83.78

Table 3: F1 performance of BERT-BiLSTM(S) with and without aspect masking.

performance on the task. This suggests that it may be beneficial to prioritize wordpieces over syntax trees to allow BERT to fully utilize rare and out-of-vocabulary words. We also discovered that using a sentence-aspect pair as input resulted in better performance than using only the sentence for the models, as indicated by the results of models with the (S,A) suffix. We believe that this may be due to the aspect information being lost or degraded during the encoding process for models with the (S) suffix. Among the methods, BERT+BiLSTM(S,A) had the highest average F1 score of 86.2.

5.2 Influence of Aspect Representation

To determine if the aspect representation is degraded during encoding, we evaluate BERT+BiLSTM(S) with and without aspect masking. The results, shown in Table 3, show that masking the aspect representation had only a minimal impact on performance, with a decrease in performance of 0.44 (Lap14), 0.16 (Res14), 0.47 (Res15), and 1.2 (Res16). These findings suggest that the aspect information has limited contribution and requires enhancement to improve performance, as demonstrated by the improved results of BERT+BiLSTM(S,A).

5.3 Qualitative Analysis

We examined the performance of BERT+BiLSTM, BERT+BiLSTM(S), and BERT+BiLSTM(S,A) on three case examples, as shown in Table 4. The results show that the BERT+BiLSTM and BERT+BiLSTM(S) models struggled to identify opinion words that were farther away from the aspect, particularly in the first and second cases where the opinion words "beautiful" and "fresh" were missed. Upon further investigation, we discovered that these opinion words were closer to the aspect's co-referential term "it". The model struggled to determine what "it" referred to due to degradation of the aspect representation, leading to the missed identification of the opinion words. However, BERT+BiLSTM(S,A) was able to recover these opinion words due to its ability to enhance the aspect representation. In the third case example, the use of wordpieces was beneficial as the opinion word "minimally" was not present in the training set, but its wordpiece "##ly," was associated with 15 opinion words in the training set. BERT+BiLSTM(S) and BERT+BiLSTM(S,A) were able to identify the opinion word "minimally" in the test set by leveraging the context of "##ly,".

6 Impact of BPE Subword Representations

We previously examined the use of wordpiece representations derived from pretrained BERT for TOWE models. In this section, we look into using Byte Pair Encoding (BPE) (Sennrich et al., 2016) as an alternative method for subword representation, which is inspired by data compression techniques (Gage, 1994). It is worth noting that BPE representations are generally not obtained from pretrained BERT. However, since RoBERTa is pretrained using BPE, and RoBERTa is a variant of BERT, we can still explore the impact of using BPE representations in TOWE models. To do this, we replace the BERT component in our best model, BERT+BiLSTM(S,A), with RoBERTa, developing the model RoBERTa+BiLSTM(S,A). The results of RoBERTa+BiLSTM(S,A) and its variations are shown in Table 5.

Note, while RoBERTa+BiLSTM(S,A) and RoBERTa+BiLSTM(S) use BPE subword token representations as input, RoBERTa+BiLSTM and RoBERTa+BiLSTM+GCN operate on the word-level. Our findings support the notion that GCNs have a limited impact on performance, as demonstrated by a relatively small decrease in average F1 score when comparing RoBERTa+BiLSTM+GCN to RoBERTa+BiLSTM. On the other hand, using BPE representations instead of GCN resulted in a substantial improvement in model performance of +5.27 when comparing RoBERTa+BiLSTM and RoBERTa+BiLSTM(S). The results indicate that syntax trees via GCNs may not be necessary and can be replaced by subword representations such as BPE for better performance in TOWE. Additionally, the performance of RoBERTa+BiLSTM(S) can be further improved by using BPE-based sentence-aspect pairs, as seen by the +1.75 performance gain in RoBERTa+BiLSTM(S,A).

6.1 State-of-the-art Models

Finally, we compare the performance of BERT+BiLSTM(S,A) with recent methods,

Sentence	**BERT+BiLSTM**	**BERT+BiLSTM(S)**	**BERT+BiLSTM(S,A)**
The **OS** is *fast* and *fluid*, everything is organized and it's just *beautiful*.	*fast, fluid*	*fast, fluid*	*fast, fluid, beautiful*
Certainly *not the best* **sushi** in new york, however, it is always *fresh*, and the place is very clean, sterile.	*fresh*	*not the best*	*not the best, fresh*
Although somewhat load, the **noise** was *minimally intrusive*	*loud, intrusive*	*loud, minimally intrusive*	*loud, minimally intrusive.*

Table 4: Case Study: Evaluating the model performance on different case examples. Aspect words are bold-typed and opinion words are italicized.

Model	**Lap14**	**Res14**	**Res15**	**Res16**	**Avg**
RoBERTa-BiLSTM(S,A)	82.77	88.27	83.84	91.06	86.49
RoBERTa-BiLSTM(S)	81.10	86.95	82.21	88.70	84.74
RoBERTa-BiLSTM	75.87	81.38	75.94	84.70	79.47
RoBERTa-BiLSTM+GCN	77.57	82.09	77.85	85.37	80.72

Table 5: F1 Performance of RoBERTa models to investigate the use of BPE subword representations.

Model	**Lap14**	**Res14**	**Res15**	**Res16**	**Avg**
IOG	71.35	80.02	73.25	81.69	76.58
LOTN	72.02	82.21	73.29	83.62	77.79
SDRN+BERT*	73.69	83.10	76.38	85.40	79.64
ONG	75.77	82.33	78.81	86.01	80.73
ARGCN	76.36	85.42	78.24	86.69	81.68
BERT+BiLSTM+GCN	78.82	85.74	80.54	87.35	83.11
QD-OWSE	80.35	87.23	80.71	88.14	84.11
TSMSA	82.18	86.37	81.64	89.20	84.85
BERT-BiLSTM (S,A)	**82.59**	**88.60**	**82.37**	**91.25**	**86.20**

Table 6: Performance of TOWE methods. Results for the method marked "*" are from (Jiang et al., 2021).

including IOG (Fan et al., 2019), LOTN (Wu et al., 2020), SDRN+BERT (Chen et al., 2020), BERT+BiLSTM+GCN (Mensah et al., 2021), QD-OWSE (Gao et al., 2021), TSMSA (Feng et al., 2021). The results of this comparison are shown in Table 6. Among these methods, the recent proposed methods QD-OWSE and TSMSA, which both use BERT as a basis for their approach, achieved competitive results with ours. QD-OWSE uses a generated question-answer pair as BERT input, while TSMSA uses multi-head attention to identify opinion words. These methods go on to demonstrate that BERT can capture sufficient syntax information for this task, even without the use of syntax trees. However, BERT+BiLSTM(S,A) achieved the best results, with F1 scores 82.59 (Lap14), 88.6 (Res14), 82.37 (Res15) and 91.25 (Res16), setting a new SOTA for the task.

7 Conclusion

We demonstrated that replacing GCNs with BERT wordpieces while enhancing the aspect representation achieves SOTA results in syntax-aware TOWE approaches. The aspect enhancement method serves as a "prompt" for the model. We intend to explore prompt-based learning (Brown et al., 2020) to further improve the aspect representation.

8 Limitations

Currently, our approach does not effectively leverage syntax tree information via GCNs, a commonly used method for incorporating syntax trees in this task. Further research is required to determine the most effective way to integrate syntax tree information into TOWE models.

Acknowledgements

This work was supported by the Leverhulme Trust under Grant Number: RPG#2020#148.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi. 2020. Synchronous double-channel recurrent network for aspect-opinion pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6515–6524.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2018. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Zhifang Fan, Zhen Wu, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2509–2518.

Yuhao Feng, Yanghui Rao, Yuyao Tang, Ninghua Wang, and He Liu. 2021. Target-specified sequence labeling with multi-head self-attention for target-oriented opinion words extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1805–1815.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Lei Gao, Yulong Wang, Tongcun Liu, Jingyu Wang, Lei Zhang, and Jianxin Liao. 2021. Question-driven span labeling model for aspect–opinion pair extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12875–12883.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Junfeng Jiang, An Wang, and Akiko Aizawa. 2021. Attention-based relational graph convolutional network for target-oriented opinion words extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1986–1997.

Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, and ChengXiang Zhai. 2011. Comprehensive review of opinion summarization.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Samuel Mensah, Kai Sun, and Nikolaos Aletras. 2021. An empirical study on leveraging position embeddings for target-oriented opinion words extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9174–9179, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Maria Pontiki, Dimitrios Galanis, Harris 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 2015)*, pages 486–495.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014a. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, page 27–35.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014b. 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. The Association for Computer Linguistics.

Lance A. Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora, VLC@ACL 1995, Cambridge, Massachusetts, USA, June 30, 1995*.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web*, pages 593–607, Cham. Springer International Publishing.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron C. Courville. 2018. Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*.

Kai Sun, Richong Zhang, Mensah Samuel, Aletras Nikolaos, Yongyi Mao, and Xudong Liu. 2023. Self-training through classifier disagreement for cross-domain opinion target extraction. In *Proceedings of the ACM Web Conference 2023*, pages 1594–1603.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224.

Yuanhe Tian, Guimin Chen, and Yan Song. 2021. Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2910–2922.

Amir Pouran Ben Veyseh, Nasim Nouri, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020. Introducing syntactic structures into target opinion word extraction with deep learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8947–8956. Association for Computational Linguistics.

Zhen Wu, Fei Zhao, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2020. Latent opinions transfer network for target-oriented opinion words extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9298–9305. AAAI Press.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☒ A2. Did you discuss any potential risks of your work?
There are no risks

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

5

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

5

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.