Bidirectional Generative Framework for Cross-domain Aspect-based Sentiment Analysis

Yue Deng [*1,2] **Wenxuan Zhang** [†1] **Sinno Jialin Pan**[2,3] **Lidong Bing**[1]
[1]DAMO Academy, Alibaba Group [2] Nanyang Technological University, Singapore
[3]Chinese University of Hong Kong
{yue.deng, saike.zwx, l.bing}@alibaba-inc.com
sinnopan@cuhk.edu.hk

Abstract

Cross-domain aspect-based sentiment analysis (ABSA) aims to perform various fine-grained sentiment analysis tasks on a target domain by transferring knowledge from a source domain. Since labeled data only exists in the source domain, a model is expected to bridge the domain gap for tackling cross-domain ABSA. Though domain adaptation methods have proven to be effective, most of them are based on a discriminative model, which needs to be specifically designed for different ABSA tasks. To offer a more general solution, we propose a unified bidirectional generative framework to tackle various cross-domain ABSA tasks. Specifically, our framework trains a generative model in both text-to-label and label-to-text directions. The former transforms each task into a unified format to learn domain-agnostic features, and the latter generates natural sentences from noisy labels for data augmentation, with which a more accurate model can be trained. To investigate the effectiveness and generality of our framework, we conduct extensive experiments on four cross-domain ABSA tasks and present new state-of-the-art results on all tasks. Our data and code are publicly available at https://github.com/DAMO-NLP-SG/BGCA.

1 Introduction

Aspect-based sentiment analysis (ABSA) is the task of analyzing people's sentiments at the aspect level. It often involves several sentiment elements, including aspects, opinions, and sentiments (Liu, 2012; Zhang et al., 2022). For instance, given the sentence *"The apple is sweet."*, the aspect is *apple*, its opinion is *sweet*, and the corresponding sentiment polarity is *Positive*. ABSA has attracted increasing attention in the last decade, and various tasks have been proposed to extract either single or multiple sentiment elements under different scenarios. For example, aspect sentiment classification (ASC) predicts the sentiment polarity of a given aspect target (Chen et al., 2017; Li et al., 2018a; Xu et al., 2020a) and aspect term extraction (ATE) extracts aspects given the sentence (Li et al., 2018b; Liu et al., 2015), while aspect sentiment triplet extraction (ASTE) predicts all three elements in the triplet format (Peng et al., 2020; Xu et al., 2021).

The main research line of ABSA focuses on solving various tasks within a specific domain. However, in real-world applications, such as E-commerce websites, there often exist a wide variety of domains. Existing methods often struggle when applying models trained in one domain to unseen domains, due to the variability of aspect and opinion expressions across different domains (Ding et al., 2017; Wang and Pan, 2018, 2019). Moreover, manually labeling data for each domain can be costly and time-consuming, particularly for ABSA requiring fine-grained aspect-level annotation. This motivates the task of cross-domain ABSA, where only labeled data in the source domain is available and the knowledge is expected to be transferable to the target domain that only has unlabeled data.

To enable effective cross-domain ABSA, domain adaptation techniques (Blitzer et al., 2006; Pan and Yang, 2010) are employed to transfer learnt knowledge from the labeled source domain to the unlabeled target domain. They either focus on learning domain-agnostic features (Ding et al., 2017; Wang and Pan, 2018; Li et al., 2019c), or adapt the training distribution to the target domain (Gong et al., 2020; Yu et al., 2021; Li et al., 2022). However, the majority of these works are based on discriminative models and need task-specific designs, making a cross-domain model designed for one ABSA task difficult to be extended for other tasks (Ding et al., 2017; Wang and Pan, 2018; Li et al., 2019c; Gong et al., 2020). In addition, some methods further require external resources, such as domain-specific

[*] Yue Deng is under the Joint PhD Program between Alibaba and Nanyang Technological University.
[†] Wenxuan Zhang is the corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 12272–12285
July 9-14, 2023

opinion lexicons (Yu et al., 2021), or extra models for augmenting pseudo-labeled target domain data (Yu et al., 2021; Li et al., 2022), which narrows their application scenarios.

In a recent research line, pre-trained generative models like BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) have demonstrated impressive power in unifying various ABSA tasks without any task-specific design and external resources. By formulating each task as a sequence-to-sequence problem and producing the desired label words, *i.e.*, the desired sentiment elements, they achieve substantial improvements on various ABSA tasks (Zhang et al., 2021b,c; Yan et al., 2021; Mao et al., 2022). Despite their success in supervised in-domain settings, their effectiveness has yet to be verified in the cross-domain setting. Moreover, unlabeled data of the target domain, which is usually easy to collect, has shown to be of great importance for bringing in domain-specific knowledge (Pan and Yang, 2010). How to exploit such data with the generative formulation remains a challenge.

Towards this end, we propose a **B**idirectional **G**enerative **C**ross-domain **A**BSA (BGCA) framework to fully exploit generative methods for various cross-domain ABSA tasks. BGCA employs a unified sequence-to-sequence format but contains two reverse directions: text-to-label and label-to-text. The text-to-label direction converts an ABSA task into a text generation problem, using the original sentence as input and a sequence of sentiment tuples as output. After training on the source labeled data $\mathcal{D}^{\mathcal{S}}$, the model can then directly conduct inference on the unlabeled data $\mathbf{x}^{\mathcal{T}}$ of the target domain $\mathcal{D}^{\mathcal{T}}$ to get the prediction $\hat{\mathbf{y}}^{\mathcal{T}}$. The prediction can be used as pseudo-labeled data to continue-train the text-to-label model. However, $\hat{\mathbf{y}}^{\mathcal{T}}$ is inevitably less accurate due to the domain gap between the source and target domains. This is where the reverse direction, i.e., label-to-text, plays its role.

Specifically, we first reverse the order of input and output from the text-to-label stage of the source domain to train a label-to-text model. Then this model takes the prediction $\hat{\mathbf{y}}^{\mathcal{T}}$ as input and generates a coherent natural language text $\hat{\mathbf{x}}^{\mathcal{T}}$ that contains the label words of $\hat{\mathbf{y}}^{\mathcal{T}}$. Note that even though the prediction $\hat{\mathbf{y}}^{\mathcal{T}}$ could be inaccurate regarding the original unlabeled data $\mathbf{x}^{\mathcal{T}}$, the generated sentence $\hat{\mathbf{x}}^{\mathcal{T}}$ can plausibly well match with $\hat{\mathbf{y}}^{\mathcal{T}}$. This is because the label-to-text model was trained to generate an output text that can appropriately describe the input labels. Consequently, $\hat{\mathbf{y}}^{\mathcal{T}}$, drawn from the target domain, is able to introduce in-domain knowledge, thereby enhancing the overall understanding of the domain-specific information. In addition, $\hat{\mathbf{x}}^{\mathcal{T}}$ aligns more closely with $\hat{\mathbf{y}}^{\mathcal{T}}$ compared to $\mathbf{x}^{\mathcal{T}}$, which effectively minimizes the prediction noise. As such, they can be paired together to create a more accurate and reliable generated dataset. Finally, the generated target data $\mathcal{D}^{\mathcal{G}}$ and the labeled source data $\mathcal{D}^{\mathcal{S}}$ can be combined to train the model in the text-to-label direction, which effectively enriches the model knowledge in the target domain.

Our proposed BGCA framework exhibits some unique advantages. Firstly, it effectively utilizes the unlabeled target domain data by capturing important domain-specific words (i.e., sentiment elements) of the target domain in the first text-to-label stage. In the meantime, it bypasses the issue from the domain gap since it takes the noisy prediction as input and obtains more accurate text-label pairs in the label-to-text stage. Secondly, we fully leverage generative models' encoding and generating capabilities to predict labels and generate natural sentences within a unified framework, which is infeasible for discriminative models. This allows the model to seamlessly switch between the roles of predictor and generator. Finally, BGCA utilizes a shared model to perform training in both directions, allowing for a more comprehensive understanding of the association between sentences and labels.

In summary, our main contributions are: (1) We evaluate generative methods on four cross-domain ABSA tasks, including aspect term extraction (ATE), unified ABSA (UABSA), aspect opinion pair extraction (AOPE), and aspect sentiment triplet extraction (ASTE), and find that the generative approach is an effective solution. Without any unlabeled target domain data, it can already achieve better performance than previous discriminative methods. (2) We propose a novel BGCA framework to effectively utilize unlabeled target domain data and train a shared model in reverse directions. It can provide high-quality augmented data by generating coherent sentences given noisy labels and a unified solution to learn the association between sentences and labels thoroughly. (3) Our proposed method achieves new state-of-the-art results on all tasks, which validate the effectiveness and generality of our framework.

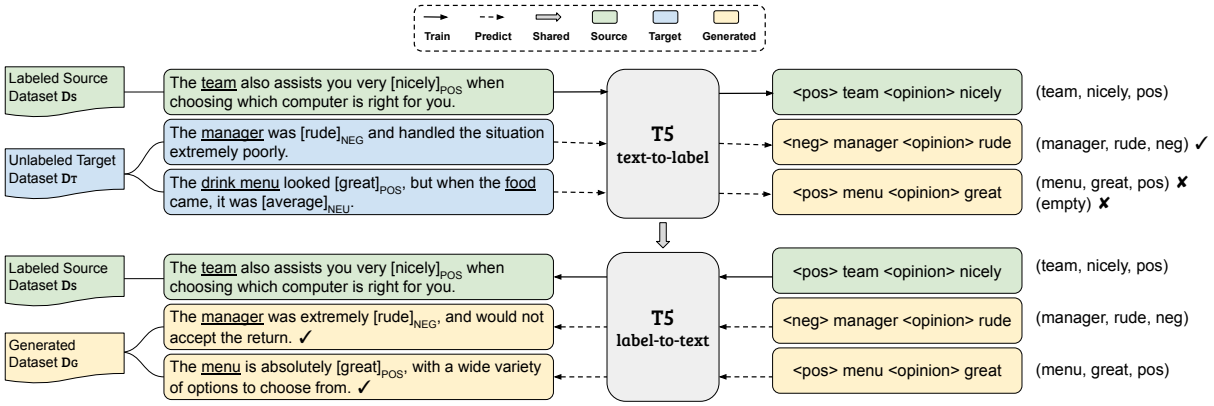

Figure 1: Overview of our proposed BGCA framework, which includes text-to-label and label-to-text directions. We take examples from the ASTE task for illustration. Underlining and square brackets indicate gold aspects and gold opinions, respectively. The gold labels for the target domain are shown for demonstration only. The generated dataset will be combined with the labeled source dataset to conduct final training in a text-to-label manner.

2 Related Work

Cross-domain ABSA Cross-domain ABSA aims to utilize labeled data from a source domain to gain knowledge that can be applied to a target domain where only unlabeled data is available. The main research line of cross-domain ABSA involves two paradigms: feature-based adaptation and data-based adaptation (Zhang et al., 2022). Feature-based adaptation focus on learning domain-invariant features. Some have utilized domain-independent syntactic rules to minimize domain gap (Jakob and Gurevych, 2010; Chernyshevich, 2014; Ding et al., 2017; Wang and Pan, 2018, 2019), while others have employed domain discriminators to encourage the learning of universal features (Li et al., 2019c; Yang et al., 2021; Zhou et al., 2021; Zhang et al., 2021a). On the other hand, data-based adaptation aims to adapt the training data distribution to the target domain. They either adjust the importance of individual training instances through re-weighting (Xia et al., 2014; Gong et al., 2020), or generate additional training data using another pre-trained model (Yu et al., 2021; Li et al., 2022). Despite their effectiveness, most of these works require task-specific design or external resources, preventing easy extensions to other cross-domain ABSA tasks.

Generative ABSA Recently, generative models have obtained remarkable results in unifying various ABSA tasks. By formulating each ABSA task as a sequence-to-sequence problem, generative models can output the desired sentiment element words (Zhang et al., 2021c; Mao et al., 2022) or their indexes (Yan et al., 2021) directly. In addition, some works successfully adopt the generative model on single ABSA tasks by converting the task to a natural language generation or paraphrase generation problem (Liu et al., 2021; Zhang et al., 2021b). Nevertheless, their potential is not explored under the cross-domain setting.

Task	Output Tuple	Example Output
ATE	(a)	(apple)
UABSA	(a, s)	(apple, positive)
AOPE	(a, o)	(apple, sweet)
ASTE	(a, o, s)	(apple, sweet, positive)

Table 1: Output tuple of various ABSA tasks, and example output given the sentence *"The apple is sweet."*, where a, o and s denote aspect, opinion and sentiment.

3 Problem Formulation

To examine the generality of our proposed framework, we consider four ABSA tasks, including ATE, UABSA, AOPE, and ASTE. Given a sentence $\mathbf{x} = [w_1, w_2, ..., w_n]$ with n words, the task is to predict a set of sentiment tuples denoted as $\mathbf{y} = \{t_i\}_{i=1}^{|t|}$, where each tuple t_i may include a single element from aspect (a), opinion (o), and sentiment (s), or multiple elements in pair or triplet format. The element within each tuple depends on the specific ABSA task, detailed in Table 1.

Under the cross-domain ABSA setting, the training dataset consists of a set of labeled sentences from a source domain $\mathcal{D}_\mathcal{S} = \{\mathbf{x}_i^\mathcal{S}, \mathbf{y}_i^\mathcal{S}\}_{i=1}^{N_\mathcal{S}}$ and a set of unlabeled sentences from a target domain $\mathcal{D}_\mathcal{T} = \{\mathbf{x}_j^\mathcal{T}\}_{j=1}^{N_\mathcal{T}}$. The goal is to leverage both $\mathcal{D}_\mathcal{S}$

and $\mathcal{D}_\mathcal{T}$ to train a model, which can predict the label of test data from the target domain.

4 Methodology

We introduce our **B**idirectional **G**enerative **C**ross-domain **A**BSA (BGCA) framework in this section. As shown in Figure 1, it contains two sequential stages, namely text-to-label, and label-to-text, to obtain high-quality augmented data. The text-to-label direction (on the top part) converts various tasks into a unified format and can produce noisy predictions on the unlabeled target data, whereas the label-to-text direction (on the bottom part) utilizes such noisy predictions to generate natural sentences containing the given labels so as to augment high-quality training data and enriches model knowledge of the target domain.

4.1 Text-to-label

The text-to-label direction unifies different ABSA tasks into a sequence-to-sequence format. It takes a sentence as input and outputs a sequence of sentiment tuples extracted from the sentence. We annotate the output sequence with predefined tagger tokens to ensure a valid format, which can prevent decoding ambiguity. The tagger tokens are k continuous tokens $\{\langle m_j \rangle\}_{j=1}^{k}$ initialized by embedding of the words $\{m_j\}_{j=1}^{k}$. Specifically, we use $\langle aspect \rangle$, $\langle opinion \rangle$ to mark aspect and opinion terms, and $\langle pos \rangle$, $\langle neu \rangle$, $\langle neg \rangle$ to annotate positive, neutral and negative sentiments. The output formats with the continuous taggers for different tasks are:

$$\begin{array}{rl} \text{ATE}: & \mathbf{x} \Rightarrow \langle aspect \rangle\, a \\ \text{UABSA}: & \mathbf{x} \Rightarrow \langle pos \rangle\, a \\ \text{AOPE}: & \mathbf{x} \Rightarrow \langle aspect \rangle\, a\, \langle opinion \rangle\, o \\ \text{ASTE}: & \mathbf{x} \Rightarrow \langle pos \rangle\, a\, \langle opinion \rangle\, o \end{array} \tag{1}$$

where a and o denote the aspect and the opinion terms, respectively. Taking ASTE as an example, we use the format of $\langle pos \rangle$ followed by the extracted aspect word(s), and $\langle opinion \rangle$ followed by the extracted opinion word(s) to annotate the positive opinion term expressed on the corresponding aspect term in a sentence. Based on this format, we are able to extract the aspect, opinion, and sentiment from the output sequence to form a complete sentiment tuple through simple regular expressions.

The text-to-label direction is trained on $\{\mathbf{x}, \mathbf{y}\}$ pairs from $\mathcal{D}_\mathcal{S}$ by minimizing the standard maximum likelihood loss:

$$\mathcal{L} = -\sum_{i=-1}^{l} \log p\left(y_i \mid \mathbf{x}; y_{\leq i-1}\right), \tag{2}$$

where l denotes the sequence length.

After training on the source labeled data $\mathcal{D}_\mathcal{S}$, we can directly conduct inference on the target domain $\mathcal{D}_\mathcal{T}$ to extract the sentiment tuples $\hat{\mathbf{y}}^\mathcal{T}$. During the inference, we employ constrained decoding (Cao et al., 2021) to ensure each generated token $\hat{y}_i^\mathcal{T}$ of the output sequence is selected from the input sentence or the predefined tagger tokens, in order to prevent invalid output sequences and ensure that the output is relevant to the specific domain:

$$\hat{y}_i^\mathcal{T} = \underset{y_j \in \mathcal{U}}{\operatorname{argmax}}\, p\left(y_j \mid \mathbf{x}^\mathcal{T}; \hat{y}_{\leq i-1}^\mathcal{T}\right), \tag{3}$$

where $\mathcal{U} = \{w_i\}_{i=1}^{n} \cup \{\langle m_j \rangle\}_{j=1}^{k}$.

4.2 Label-to-text

Although the text-to-label model can be directly applied for prediction on the target domain, it does not exploit the unlabeled target domain data in the training process, which has been proven to be crucial for incorporating target-domain knowledge (Pan and Yang, 2010). One straightforward way to eliminate this problem is to use $(\mathbf{x}^\mathcal{T}, \hat{\mathbf{y}}^\mathcal{T})$ as pseudo-labeled data to continue training the above text-to-label model. However, such naive self-training suffers from the noise of $\hat{\mathbf{y}}^\mathcal{T}$. Our label-to-text stage alleviates this weakness by pairing the label $\hat{\mathbf{y}}^\mathcal{T}$ with a new sentence that matches this label better.

Specifically, we continue to train the above model using the labeled dataset from $\mathcal{D}_\mathcal{S}$. Nevertheless, the training pairs are reversed into the label-to-text direction, where the input is now the sequence $\mathbf{y}$ with sentiment tuples, and the output is the original sentence $\mathbf{x}$:

$$\begin{array}{rl} \text{ATE}: & \langle aspect \rangle\, a \Rightarrow \mathbf{x} \\ \text{UABSA}: & \langle pos \rangle\, a \Rightarrow \mathbf{x} \\ \text{AOPE}: & \langle aspect \rangle\, a\, \langle opinion \rangle\, o \Rightarrow \mathbf{x} \\ \text{ASTE}: & \langle pos \rangle\, a\, \langle opinion \rangle\, o \Rightarrow \mathbf{x} \end{array} \tag{4}$$

Similarly, the label-to-text direction is trained on $\{\mathbf{y}, \mathbf{x}\}$ pairs from $\mathcal{D}_\mathcal{S}$ by minimizing the standard maximum likelihood loss:

$$\mathcal{L} = -\sum_{i=-1}^{l'} \log p\left(x_i \mid \mathbf{y}; x_{\leq i-1}\right), \tag{5}$$

and l' refers to the sequence length.

After training, we use the sentiment tuples $\hat{\mathbf{y}}^\mathcal{T}$, extracted from a target domain unlabeled data $\mathbf{x}^\mathcal{T}$,

Task	ATE&UABSA				AOPE				ASTE			
	L	R	D	S	L14	R14	R15	R16	L14	R14	R15	R16
Train	3045	3877	2557	1492	1035	1462	678	971	906	1266	605	857
Dev	304	387	255	149	116	163	76	108	219	310	148	210
Test	800	2158	1279	747	343	500	325	328	328	492	322	326

Table 2: The statistics of ATE, UABSA, AOPE and ASTE tasks

to generate a natural sentence $\hat{\mathbf{x}}^{\mathcal{T}}$ incorporating the sentiment information in $\hat{\mathbf{y}}^{\mathcal{T}}$. To ensure fluency and naturalness, we decode the whole vocabulary set:

$$\hat{x}_i^{\mathcal{T}} = \underset{x_j \in \mathcal{V}}{\operatorname{argmax}}\, p\left(x_j \mid \hat{\mathbf{y}}^{\mathcal{T}}; \hat{x}_{\leq i-1}^{\mathcal{T}}\right), \quad (6)$$

where $\mathcal{V}$ denotes the vocabulary of the model.

The label-to-text stage thus augments a generated dataset $\mathcal{D}_{\mathcal{G}} = \left\{\hat{\mathbf{x}}_i^{\mathcal{T}}, \hat{\mathbf{y}}_i^{\mathcal{T}}\right\}_{i=1}^{N_{\mathcal{T}}}$. By considering each natural sentence as a combination of context and sentiment elements, we can find that the generated sentence's context is produced by a model pre-trained on large-scale corpora and fine-tuned on the labeled source domain, while its sentiment elements such as aspects and opinions come from the target domain. Therefore, $\mathcal{D}_{\mathcal{G}}$ can play the role of an intermediary which connects the source and target domains through the generated sentences.

As previously mentioned, due to the gap between source and target domains, the text-to-label model's prediction on unlabeled target data is noisy. Instead of improving the text-to-label model, which may be difficult, our label-to-text stage creates a sentence $\hat{\mathbf{x}}^{\mathcal{T}}$ that is generated specifically for describing $\hat{\mathbf{y}}^{\mathcal{T}}$. Thus, even with the presence of noise in the extracted labels $\hat{\mathbf{y}}^{\mathcal{T}}$, the label-to-text stage offers a means of minimizing the negative impact and ultimately yields a more accurate pseudo-training sample. Finally, since these two stages train a shared model based on sentences and labels from two directions, it gives the model a more comprehensive understanding of the association between sentences and labels, leading to a more accurate prediction of labels for given sentences.

4.3 Training

Ideally, the generated dataset $\mathcal{D}_{\mathcal{G}}$ should fulfil the following requirements: 1) the natural sentence should exclusively contain sentiment elements that are labeled in the sentiment tuples, and should not include any additional sentiment elements; 2) the natural sentence should accurately convey all the sentiment elements as specified in the sentiment tuples without any omissions; 3) the sentiment tuples should be in a valid format and can be mapped back to the original labels; Therefore, we post-process $\{\hat{\mathbf{x}}^t, \hat{\mathbf{y}}^t\}$ pairs from $\mathcal{D}_{\mathcal{G}}$ by: 1) filtering out pairs with $\hat{\mathbf{y}}^t$ in invalid format or contains words not present in $\hat{\mathbf{x}}^t$; 2) utilizing the text-to-label model to eliminate pairs where $\hat{\mathbf{y}}^t$ is different from the model's prediction on $\hat{\mathbf{x}}^t$. In the end, we combine the source domain $\mathcal{D}_{\mathcal{S}}$, and the generated dataset $\mathcal{D}_{\mathcal{G}}$ as the ultimate training dataset and continue to train the same model in a text-to-label manner as outlined in Section 4.1.

5 Experiments

5.1 Experimental Setup

Datasets We evaluate the proposed framework on four cross-domain ABSA tasks, including ATE, UABSA, AOPE, and ASTE. Datasets of these tasks mainly consist of four different domains, which are Laptop (L), Restaurant (R), Device (D), and Service (S). L, also referred to as L14, contains laptop reviews from SemEval ABSA challenge 2014 (Pontiki et al., 2014). R is a set of restaurant reviews based on SemEval ABSA challenges 2014, 2015, and 2016 (Pontiki et al., 2014, 2015, 2016), denoted as R14, R15, R16 for the AOPE and ASTE tasks. D contains digital device reviews provided by Toprak et al. (2010). S includes reviews from web service, introduced by Hu and Liu (2004). Specifically, we can perform the ATE and UABSA tasks on all four domains, whereas the AOPE and ASTE tasks can be conducted on L and R domains, with R being further divided into R14, R15, and R16. We follow the dataset setting provided by Yu et al. (2021) for the ATE and UABSA task, and Fan et al. (2019), Xu et al. (2020b) for the AOPE, ASTE task respectively. We show the statistics in Table 2.

Settings We consider all possible transfers between each pair of domains for each task. Following previous work (Li et al., 2019a,b; Gong et al., 2020; Yu et al., 2021), we remove D→L and L→D for the ATE and UABSA tasks due to their domain similarity. Additionally, we exclude transfer pairs

Methods	S→R	L→R	D→R	R→S	L→S	D→S	R→L	S→L	R→D	S→D	**Avg.**
ATE											
Hier-Joint†	46.39	48.61	42.96	27.18	25.22	29.28	34.11	33.02	34.81	35.00	35.66
RNSCN†	48.89	52.19	50.39	30.41	31.21	35.50	47.23	34.03	**46.16**	32.41	40.84
AD-SAL†	52.05	56.12	51.55	39.02	38.26	36.11	45.01	35.99	43.76	**41.21**	43.91
BERT_B-UDA†	56.08	51.91	50.54	34.62	32.49	34.52	46.87	43.98	40.34	38.36	42.97
BERT_B-CDRG†	56.26	60.03	52.71	42.36	47.08	41.85	46.65	39.51	32.60	36.97	45.60
GAS	61.24	53.02	56.44	31.19	32.14	35.72	52.24	43.76	42.24	37.77	44.58
BERT_E-UDA†*	59.07	55.24	56.40	34.21	30.68	38.25	54.00	44.25	42.40	40.83	45.53
BERT_E-CDRG†*	59.17	68.62	58.85	**47.61**	**54.29**	42.20	55.56	41.77	35.43	36.53	50.00
$\textbf{BGCA}_{\text{text-to-label}}$	60.03	55.39	55.83	36.02	35.43	37.73	54.18	43.45	42.49	37.89	45.84
$\textbf{BGCA}_{\text{label-to-text}}$	**63.20**	**69.53**	**65.33**	45.86	44.85	**54.07**	**57.13**	**46.15**	37.15	38.24	**52.15**
UABSA											
Hier-Joint†	31.10	33.54	32.87	15.56	13.90	19.04	20.72	22.65	24.53	23.24	23.72
RNSCN†	33.21	35.65	34.60	20.04	16.59	20.03	26.63	18.87	33.26	22.00	26.09
AD-SAL†	41.03	43.04	41.01	28.01	27.20	26.62	34.13	27.04	35.44	33.56	33.71
AHF	46.55	43.49	44.57	33.23	33.05	34.96	34.89	29.01	37.33	**39.61**	37.67
BERT_B-UDA†	47.09	45.46	42.68	33.12	27.89	28.03	33.68	34.77	34.93	32.10	35.98
BERT_B-CDRG†	47.92	49.79	47.64	35.14	38.14	37.22	38.68	33.69	27.46	34.08	38.98
GAS	54.61	49.06	53.40	30.99	29.64	33.34	43.50	35.12	39.29	35.81	40.48
BERT_E-UDA†*	53.97	49.52	51.84	30.67	27.78	34.41	43.95	35.76	40.35	38.05	40.63
BERT_E-CDRG†*	53.09	57.96	54.39	40.85	**42.96**	38.83	**45.66**	35.06	31.62	34.22	43.46
$\textbf{BGCA}_{\text{text-to-label}}$	54.12	48.08	52.65	33.26	30.67	35.26	44.57	36.01	**41.19**	36.55	41.24
$\textbf{BGCA}_{\text{label-to-text}}$	**56.39**	**61.69**	**59.12**	**43.20**	39.76	**47.94**	45.52	**36.40**	34.16	36.57	**46.07**

Table 3: Results on cross-domain ATE and UABSA tasks. The best results are in bold. Results are the average F1 scores over 5 runs. † denotes results from Yu et al. (2021), and the others are based on our implementation. ∗ represents methods that utilize external resources.

between R14, R15, and R16 for the AOPE and ASTE tasks since they come from the same restaurant domain. As a result, there are ten transfer pairs for the ATE and UABSA tasks, and six transfer pairs for the AOPE and ASTE tasks, detailed in Table 3 and 4. We denote our proposed framework as $\textbf{BGCA}_{\text{label-to-text}}$, which includes the bidirectional augmentation and utilizes the augmented data for training the final model. To investigate the effectiveness of the generative framework for cross-domain ABSA tasks, we also report the results with a single text-to-label direction, denoted as $\textbf{BGCA}_{\text{text-to-label}}$, which is essentially a zero-shot cross-domain method.

Metrics We choose the Micro-F1 score as the evaluation metric for all tasks. A prediction is counted as correct if and only if all the predicted elements are exactly matched with gold labels.

Implementation Details We choose T5 (Raffel et al., 2020) as our backbone model and use T5-base checkpoint from *huggingface*[1]. It is a transformer model (Vaswani et al., 2017) that utilizes the encoder-decoder architecture where all the pre-training tasks are in sequence-to-sequence format. For simplicity, we use the Adam optimizer with a learning rate of 3e-4, a fixed batch size of 16, and a fixed gradient accumulation step of 2 for all tasks. Regarding training epochs for text-to-label, label-to-text, and final training, we search within a range in {15, 20, 25, 30} using the validation set of the source domain for selection. We train our model on a single NVIDIA V100 GPU.

[1] https://github.com/huggingface/

5.2 Baselines

For cross-domain ATE and UABSA tasks, we follow previous works to compare with established baselines including Hier-Joint (Ding et al., 2017), RNSCN (Wang and Pan, 2018), AD-SAL (Li et al., 2019c), AHF (Zhou et al., 2021), $\text{BERT}_{\text{B/E}}$-UDA (Gong et al., 2020), and $\text{BERT}_{\text{B/E}}$-CDRG (Yu et al., 2021) where BERT_B and BERT_E refer to models based on the original BERT and the continually trained BERT on large-scale E-commerce data containing around 3.8 million reviews (Xu et al., 2019). All of these methods utilize unlabeled target data, and $\text{BERT}_{\text{B/E}}$-CDRG are trained in a self-training manner, which generates pseudo labels and retrain a new model with such labels.

Methods	R14→L14	R15→L14	R16→L14	L14→R14	L14→R15	L14→R16	**Avg.**
AOPE							
SDRN	45.39	37.45	38.66	47.63	41.34	46.36	42.81
RoBMRC	52.36	46.44	43.61	54.70	48.68	55.97	50.29
SpanASTE	51.90	48.15	47.30	61.97	55.58	63.26	54.69
GAS	57.58	53.23	52.17	64.60	60.26	66.69	59.09
BGCA$_{\text{text-to-label}}$	58.54	54.06	51.99	64.61	58.74	67.19	59.19
BGCA$_{\text{label-to-text}}$	**60.82**	**55.22**	**54.48**	**68.04**	**65.31**	**70.34**	**62.37**
ASTE							
RoBMRC	43.90	40.19	37.81	57.13	45.62	52.05	46.12
SpanASTE	45.83	42.50	40.57	57.24	49.02	55.77	48.49
GAS	49.57	43.78	45.24	64.40	56.26	63.14	53.73
BGCA$_{\text{text-to-label}}$	52.55	**45.85**	46.86	61.52	55.43	61.15	53.89
BGCA$_{\text{label-to-text}}$	**53.64**	45.69	**47.28**	**65.27**	**58.95**	**64.00**	**55.80**

Table 4: Results on cross-domain AOPE and ASTE tasks. The best results are in bold. Results are the average F1 scores over 5 runs.

Methods	**ATE**	**UABSA**	**AOPE**	**ASTE**	**Avg.**
BGCA†	**52.15**	**46.07**	**62.37**	55.80	**54.10**
- self-training*	46.13	41.56	61.33	**55.99**	51.25
- continue*	46.63	42.22	58.56	54.70	50.53
- w/o sharing	52.08	44.72	61.64	55.76	53.55

Table 5: Ablation Study. BGCA† represents our BGCA$_{\text{label-to-text}}$ setting. * denotes replacing the label-to-text stage with the corresponding training method.

For cross-domain AOPE and ASTE tasks, since there is no existing work on these two tasks under the cross-domain setting, we leverage the in-domain state-of-the-art models in a zero-shot manner for comparisons, including SDRN (Chen et al., 2020) for AOPE, and RoBMRC (Liu et al., 2022), SpanASTE (Xu et al., 2021) for ASTE task. In addition, we also refine RoBMRC and SpanASTE to work for the AOPE task by simply omitting the prediction of sentiment polarity.

Most of the above baselines are discriminative methods based on the pre-trained BERT model. To enable a fair comparison, we also employ GAS (Zhang et al., 2021c) for all four ABSA tasks, which is a strong unified generation method based on the same pre-trained generative model, i.e., T5-base, as our proposed BGCA method.

5.3 Main Results

We report the main results for the ATE and UABSA tasks in Table 3 and the AOPE and ASTE tasks in Table 4. We have the following observations: 1) Our method with a single text-to-label direction (**BGCA**$_{\text{text-to-label}}$) establishes a strong baseline for cross-domain ABSA tasks. Compared to discriminative baseline methods without external resources, it shows an improvement of 0.24%, 2.26%, 4.5%, and 5.4% on the cross-domain ATE, UABSA, AOPE, and ASTE tasks, respectively. This demonstrates that generative models can actually generalize well across different domains with our designed continuous tagger to indicate the desired sentiment elements. 2) Our proposed framework **BGCA**$_{\text{label-to-text}}$ with bidiretional augmentations achieves new state-of-the-art results on all four cross-domain ABSA tasks. It outperforms the previous best models by 2.15% and 2.61% on the ATE and UABSA tasks and by 3.28% and 2.07% on AOPE and ASTE. Notably, it requires no external resources and can be seamlessly applied to all cross-domain ABSA tasks. This verifies the generalizability and effectiveness of our proposed bidirectional generation-based augmentation method. 3) Compared to other generation-based methods such as GAS and **BGCA**$_{\text{text-to-label}}$, **BGCA**$_{\text{label-to-text}}$ outperforms all of them on four tasks, indicating that the label-to-text direction can effectively utilize the unlabeled target data and leverage the potential of generative models.

5.4 Ablation Study

We conduct ablation studies to analyze the effectiveness of each component in BGCA. Results of different model variants are reported in Table 5.

Ablation on label-to-text generation To investigate the effectiveness of the label-to-text direction, and verify our assumption that it can fix the noisy prediction issue, we replace it with the self-training method and denote it as "self-training" in Table 5. Specifically, we use the pseudo labels of the unlabeled target domain data extracted by the text-to-label stage to replace our augmented data. As shown in Table 5, the performance drops about

Sentence from R	Prediction	Label-to-text Generation
The [service]$_{\text{POS}}$ was good to excellent along with the [attitude]$_{\text{POS}}$.	(service, POS)	The [service]$_{\text{POS}}$ I received from Toshiba was excellent.
[Bottles of wine]$_{\text{POS}}$ are cheap and good.	(bottles, POS)	I love the [bottles]$_{\text{POS}}$ they are made out of.
Our [waitress]$_{\text{NEU}}$ wasn't mean, but not especially warm or attentive either.	(waitress, NEG)	The [waitress]$_{\text{NEG}}$ didn't even answer my question.

Table 6: Examples on L→R from the UABSA task. Gold aspects are marked by square brackets. POS, NEU and NEG denote positive, neutral and negative sentiment.

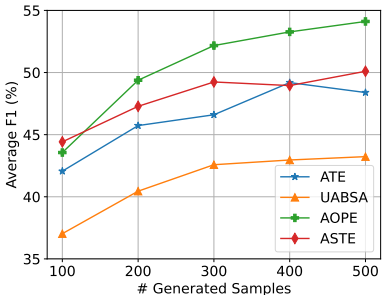

Figure 2: Comparison results of our method with a different number of generations.

Group	ATE		UABSA	
	text→label	label→text	text→label	label→text
Zero	45.31	36.48	50.02	39.18
Single	41.53	47.99	35.02	43.17
Multiple	26.61	37.20	21.99	29.59

Table 7: Comparison results on cross-domain ATE and UABSA tasks over different sentence groups containing zero, single, or multiple aspects respectively.

three points on average for four tasks. This indicates that the pseudo-labeled samples from the text-to-label model contain more noise. Adding label-to-text generation could effectively address this issue by generating pseudo-training data with less noise. To further investigate the effectiveness of generated samples, we manually check some samples on L→R from the UABSA task and show some representative samples in Table 6. Note that the gold labels for the target domain are not available during training, and we display them here for investigation only. The first two example's predictions either omit an aspect or gives an incomplete aspect, while the third example's prediction gives the wrong sentiment. However, the label-to-text model can generate a correct sentence that appropriately describes the prediction, although it is inaccurate regarding to the original input sentence. These examples demonstrate how the label-to-text stage can resolve noisy prediction issues and produce high-quality target domain data.

Ablation on unlabeled data utilization Continue training has shown to be an effective method to leverage unlabeled data by conducting pre-training tasks on relevant corpora to capture domain-specific knowledge (Xu et al., 2019; Gong et al., 2020; Yu et al., 2021). We compare it with our method to discuss how to utilize unlabeled data for generative cross-domain ABSA and denote it as "continue" in Table 5. Specifically, we replace the label-to-text stage with conducting continue-train on the unlabeled data of the target domain, with the span reconstruction objective as original T5 pre-training (Raffel et al., 2020). The results show that continue training lies behind our proposed method and demonstrate that our framework can effectively utilize unlabeled target domain data. The possible reason may be that continue training requires many training samples, which is infeasible in cross-domain ABSA scenarios.

Ablation on model sharing To demonstrate the advantages of training a shared model in both directions, we compare it to a method where a model is newly initialized before each stage of training and denote it as "w/o sharing" in Table 5. Results on four tasks show that our approach outperforms the non-shared method by an average of 0.6%, suggesting that a shared model owns a better understanding of the association between sentences and labels.

5.5 Further Analysis

Analysis on number of generated samples Figure 2 shows the comparison results over four tasks with different numbers of generated samples. To better analyze the effect of the number of generations, we exclude the source training data and solely use the generated samples for final training. There is an apparent trend of performance improvement with the increasing number of generated samples, revealing that the generated samples can boost cross-domain ability.

Analysis on improvement types To understand what types of cases our method improved, we cate-

gorize sentences from the test set into three groups: without any aspect, with a single aspect, and with multiple aspects. We conduct our analysis on the cross-domain ATE and UABSA tasks since they contain sentences without any aspect, and evaluate the performance of both the text-to-label and label-to-text settings for each group. We choose sentence-level accuracy as the evaluation metric, *i.e.*, a sentence is counted as correct if and only if all of its sentiment elements are correctly predicted. We present the average accuracy across all transfer pairs in Table 7. The text-to-label model has less knowledge of the target domain and thus tends to predict sentences as no aspect, leading to high accuracy in the group without any aspect. However, it also misses many sentiment elements in the other two groups. On the other hand, although label-to-text lies behind text-to-label in the group without any aspect, it significantly improves the performance of sentences with single or multiple aspects. This indicates that the label-to-text model has obtained more target domain knowledge than the text-to-label setting, and thus can identify more sentiment elements.

6 Conclusions

In this work, we extend the generative method to cross-domain ABSA tasks and propose a novel BGCA framework to boost the generative model's cross-domain ability. Specifically, we train a shared generative model in reverse directions, allowing high-quality target domain augmentation and a unified solution to comprehend sentences and labels fully. Experiments on four cross-domain ABSA tasks verify the effectiveness of our method.

7 Limitations

In this paper, we present a bidirectional generative framework for cross-domain ABSA that has achieved outstanding results on four cross-domain ABSA tasks. Although there is only one stage during inference, our method involves multiple training stages, including text-to-label, label-to-text, and final training. These additional training stages not only lengthen the training time but also require additional computational resources, which may hinder scalability for large-scale data and result in a burden for the environment.

Acknowledgements

S. J. Pan thanks for the support of the Hong Kong Global STEM Professorship. Y. Deng is supported by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore.

References

John Blitzer, Ryan T. McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *EMNLP 2006, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, pages 120–128.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

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, ACL 2020, Online, July 5-10, 2020*, pages 6515–6524.

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.

Ying Ding, Jianfei Yu, and Jing Jiang. 2017. Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3436–3442.

Zhifang Fan, Zhen Wu, Xin-Yu 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.

Chenggong Gong, Jianfei Yu, and Rui Xia. 2020. Unified feature and instance based domain adaptation for aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7035–7045.

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*, pages 168–177.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1035–1045.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Junjie Li, Jianfei Yu, and Rui Xia. 2022. Generative cross-domain data augmentation for aspect and opinion co-extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4219–4229.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018a. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6714–6721.

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

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019b. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41.

Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang. 2019c. Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4590–4600.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Jian Liu, Zhiyang Teng, Leyang Cui, Hanmeng Liu, and Yue Zhang. 2021. Solving aspect category sentiment analysis as a text generation task. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4406–4416.

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*, pages 1433–1443.

Shu Liu, Kaiwen Li, and Zuhe Li. 2022. A robustly optimized BMRC for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 272–278.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2215–2225.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. 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 8600–8607.

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 Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-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 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 2014)*, pages 27–35.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584.

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, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Wenya Wang and Sinno Jialin Pan. 2018. Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2171–2181.

Wenya Wang and Sinno Jialin Pan. 2019. Transferable interactive memory network for domain adaptation in fine-grained opinion extraction. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7192–7199.

Rui Xia, Jianfei Yu, Feng Xu, and Shumei Wang. 2014. Instance-based domain adaptation in NLP via in-target-domain logistic approximation. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pages 1600–1606.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335.

Lu Xu, Lidong Bing, Wei Lu, and Fei Huang. 2020a. Aspect sentiment classification with aspect-specific opinion spans. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3561–3567.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4755–4766.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020b. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2416–2429.

Min Yang, Wenpeng Yin, Qiang Qu, Wenting Tu, Ying Shen, and Xiaojun Chen. 2021. Neural attentive network for cross-domain aspect-level sentiment classification. *IEEE Trans. Affect. Comput.*, 12(3):761–775.

Jianfei Yu, Chenggong Gong, and Rui Xia. 2021. Cross-domain review generation for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4767–4777.

Kai Zhang, Qi Liu, Hao Qian, Biao Xiang, Qing Cui, Jun Zhou, and Enhong Chen. 2021a. Eatn: An efficient adaptive transfer network for aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35:377–389.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021b. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9209–9219.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021c. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 504–510.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *CoRR*, abs/2203.01054.

Yan Zhou, Fuqing Zhu, Pu Song, Jizhong Han, Tao Guo, and Songlin Hu. 2021. An adaptive hybrid framework for cross-domain aspect-based sentiment analysis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021*, pages 14630–14637.

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?
Pure scientific research

☑ A3. Do the abstract and introduction summarize the paper's main claims?
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?
5

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All datasets are free to use for research purpose.

☑ 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)?
5

☒ 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?
The original paper of datasets have already discussed this content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
5

☑ 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.
In Table 2

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?
5

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?
5

☑ 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?
5

☑ 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.)?
5

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.?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.