Generative Cross-Domain Data Augmentation for Aspect and Opinion Co-Extraction

Junjie Li, Jianfei Yu[*], and Rui Xia[*]
School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
{jj_li, jfyu, rxia}@njust.edu.cn

Abstract

As a fundamental task in opinion mining, aspect and opinion co-extraction aims to identify the aspect terms and opinion terms in reviews. However, due to the lack of fine-grained annotated resources, it is hard to train a robust model for many domains. To alleviate this issue, unsupervised domain adaptation is proposed to transfer knowledge from a labeled source domain to an unlabeled target domain. In this paper, we propose a new Generative Cross-Domain Data Augmentation framework for unsupervised domain adaptation. The proposed framework is aimed to generate target-domain data with fine-grained annotation by exploiting the labeled data in the source domain. Specifically, we remove the domain-specific segments in a source-domain labeled sentence, and then use this as input to a pre-trained sequence-to-sequence model BART to simultaneously generate a target-domain sentence and predict the corresponding label for each word. Experimental results on three datasets demonstrate that our approach is more effective than previous domain adaptation methods. The source code is publicly released at https://github.com/NUSTM/GCDDA.

1 Introduction

Aspect and opinion co-extraction is a fundamental task in opinion mining. It aims to extract aspect terms and opinion terms from review sentences. Given a review "*the keyboard is great*", the aspect term is *keyboard* and the opinion term is *great*. Most existing works formulate aspect and opinion co-extraction as a sequence labeling task. With the support of deep learning, many supervised methods have achieved remarkable results (Liu et al., 2015; Yin et al., 2016; Wang et al., 2017; Wu et al., 2020). However, owing to the high cost of fine-grained annotations, the scarcity of labeled data in many new domains makes them fail to obtain a robust performance.

To alleviate the deficiency of labeled data, unsupervised domain adaptation is proposed to transfer knowledge from a labeled source domain to an unlabeled target domain. The main difficulty of unsupervised domain adaptation lies in the distribution discrepancy between data from different domains. Specifically, reviews from different domains may have distinct aspect terms, opinion terms, and expression patterns. To tackle the distribution discrepancy problem, many domain adaptation methods were proposed for coarse-grained sentiment classification (Blitzer et al., 2007; Yu and Jiang, 2016; Ziser and Reichart, 2018; Ghosal et al., 2020). However, due to the complexity of fine-grained domain adaptation for aspect and opinion extraction, only a handful of studies attempt to address this issue. Most of them are based on the following three paradigms:

- **rule-based adaptation**, which iteratively extracts aspects and opinions in sentences based on domain-independent syntactic rules and opinion dictionary (Li et al., 2012; Ding et al., 2017).
- **feature-based adaptation**, which incorporates the general syntactic information to learn a domain-invariant word representation across domains (Wang and Pan, 2018; Li et al., 2019; Pereg et al., 2020; Chen and Qian, 2021).
- **data augmentation-based adaptation**, which utilizes the labeled data in the source domain to directly generate high-quality target-domain data with fine-grained annotation (Yu et al., 2021).

However, the paradigms mentioned above suffer from the following problems. For rule-based adaptation methods, it is difficult to design high-quality manual rules and opinion set, which may bring low precision results. For feature-based adaptation methods, although they can bridge the domain gap by using the unlabeled data from both domains, the

[*] Corresponding authors.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4219 - 4229
July 10-15, 2022 ©2022 Association for Computational Linguistics

main task classifier is only trained by the source labeled data, which fails to exploit the important supervision signals in the target domain. Although our recent work (Yu et al., 2021) has shown the superiority of data augmentation-based adaptation methods, its main limitation lies in that it only replaces source-specific aspects and opinions in the source domain review with target-specific aspects and opinions, but ignoring other domain-specific attributes such as collocations and expression styles, which limits the quality and diversity of generated target-domain data.

To this end, we propose a Generative Cross-Domain Data Augmentation framework for unsupervised domain adaptation, which generates target labeled data from source labeled data based on a pre-trained sequence-to-sequence model, i.e., BART (Lewis et al., 2020). Specifically, given the labeled samples in the source domain, we first train a classifier to assign pseudo labels for each unlabeled sample in the target domain. With the labeled and pseudo labeled samples from both domains, we then remove their domain-specific features to obtain domain-independent samples. Next, we employ the BART model by feeding the domain-independent samples and their token labels to the encoder and reconstructing their original texts and corresponding labels in the decoder. In the inference stage, given a source-domain sentence in which the domain-specific features are removed, the model can generate a target-domain sentence by integrating target-specific features into the source context while predicting its token-level labels. Compared with the previous method (Yu et al., 2021), our new approach generates high-quality target-domain data with more flexible syntactic patterns by making full use of the knowledge from the domain-invariant source contexts. With the label prediction, our approach can easily obtain the adaptive token labels for each generated target-domain sample.

Our main contributions in this work can be summarized as follows:

- We propose a Generative Cross-Domain Data Augmentation framework for unsupervised domain adaptation, which exploits the pre-trained sequence-to-sequence model BART to generate the target-domain data with fine-grained annotation based on the labeled source-domain data.
- Extensive experiments on three datasets of aspect and opinion co-extraction demonstrate the effectiveness of our generative framework. Further analysis shows that our framework is able to generate more fluent and diversified target-domain samples than previous approaches.

2 Related Work

2.1 Aspect and Opinion Extraction

Aspect and opinion extraction is an fundamental task in opinion mining, which is widely studied in the literature. Unsupervised learning methods have been proposed to extract aspect and opinion terms by association rule and extended opinions (Hu and Liu, 2004) or by manual syntactic rules and predefined opinion dictionary (Qiu et al., 2011). With the constant attention on deep learning, many supervised approaches have achieved remarkable results (Liu et al., 2015; Zhang et al., 2015; Yin et al., 2016; Wang et al., 2017; Yu et al., 2018; Wu et al., 2020). However, the deficiency of annotated data in many domains make them fail to achieve desired results.

2.2 Domain Adaptation

To solve the issue mentioned above, many domain adaptation methods proposed for coarse-grained sentiment classification either align the domain-specific feature space with domain-independent pivot words (Blitzer et al., 2007; Pan et al., 2010; Yu and Jiang, 2016) or learn a domain-invariant representation based on auto-encoder (Glorot et al., 2011; Yin et al., 2016) and domain adversarial learning (Li et al., 2018b). However, only a few fine-grained domain adaptation approaches are proposed for ABSA (Li et al., 2019; Gong et al., 2020) or aspect and opinion extraction. Most of them can be categorized into two types: 1) rule-based methods (Li et al., 2012; Ding et al., 2017) exploit manual syntactic rules and opinion seeds to extract aspects and opinions, which usually obtains less satisfactory results due to the quality of rules; 2) feature representation-based methods focus on learning a domain-invariant word representation (Wang and Pan, 2018; Pereg et al., 2020; Chen and Qian, 2021) based on the unlabeled data. However, the main task classifier fails to exploit the important supervision signals in the target domain.

2.3 Data Augmentation

Data augmentation is another solution for the scarcity of annotation data. Many previous works mainly focus on the sentence-level task (Guo

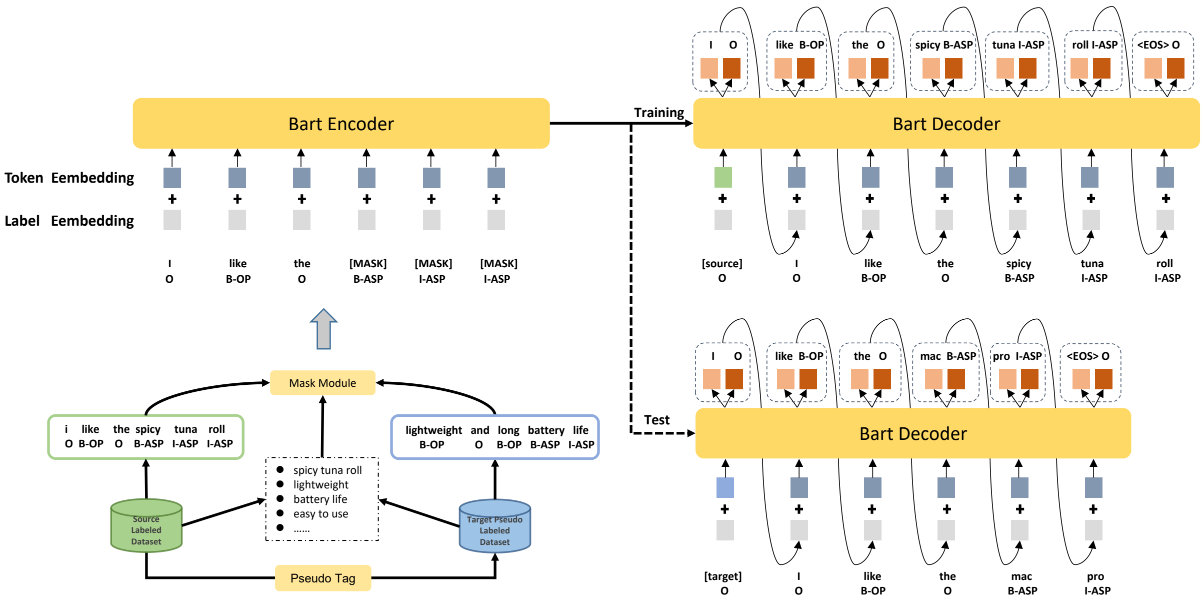

Figure 1: Overview of our proposed Generative Cross-Domain Data Augmentation framework.

et al., 2019; Min et al., 2020) and the token-level task (Gao et al., 2019; Chen et al., 2020) . For aspect/opinion extraction, Ding et al. (2020) designed a data augmentation method with language models trained on the linearized labeled sentences to generate considerable labeled data. Hsu et al. (2021) exploited a masked language model BERT to substitute unimportant words in sentences to enhance the data diversity. Li et al. (2020) proposed to generate a new review while preserving its original labels and aspect terms with a masked sequence-to-sequence model. However, all the studies above follow the in-domain setting, which fails to exploit the rich contexts from other domains. Our recent work (Yu et al., 2021) proposed a cross-domain review generation method, which replaces source-specific aspects and opinions in labeled source-domain reviews with target-specific aspect and opinions based on the masked language model BERT. However, it fails to consider other domain-specific attributes such as collocations and expression styles in cross-domain review generation. Based on our previous work, we propose a Generative Cross-Domain Data Augmentation framework for unsupervised domain adaptation, which can generate more flexible target data with fine-grained annotation in this paper.

3 Problem Formulation

In this paper, we focus on the aspect and opinion co-extraction task and formulate it as a sequence labeling problem. We denote a review as a sequence of tokens $x = [x_1, x_2, ..., x_n]$, and the task aims to predict the label sequence of the review $y = [y_1, y_2, ..., y_n]$, where $y_i \in \{B{-}ASP, I{-}ASP, B{-}OP, I{-}OP, O\}$. For unsupervised domain adaptation, we are given a set of source-domain labeled reviews $D_S = \{(x_i^s, y_i^s)\}_{i=1}^{N^s}$ and a set of unlabeled data from the target domain $D_U = \{x_i^u\}_{i=1}^{N^u}$. The goal is to predict token-level labels on the test set from the target domain $D_T = \{x_i^t\}_{i=1}^{N^t}$.

4 Methodology

In this paper, we propose a Generative Cross-Domain Data Augmentation framework, which generates target-domain reviews with fine-grained annotation from source-domain labeled reviews. Figure 1 illustrates the overall architecture of our generative framework. Specifically, we leverage the labeled samples in the source domain to train a classifier, and then assign pseudo labels for each unlabeled sample in the target domain. Given the labeled and pseudo labeled sentences from both domains, we employ a segment mask module to remove their domain-specific features while preserving their context and original labels. Next, a pre-trained sequence-to-sequence model BART concatenates the domain-invariant context and original labels as the input of the encoder, followed by recovering the original sentences and predicting their token-level labels via the decoder. During the inference stage, given a labeled source-domain sentence, the model removes its domain-specific features, and then employs the fine-tuned BART model to generate a target-domain sentence with its token-level labels.

4.1 Pseudo Label Annotation

First, we train a base classifier on the labeled data from the source domain D_S, which employs a pre-trained BERT model (Devlin et al., 2019) to obtain the contextualized word representation and a Conditional Random Field (CRF) layer for sequence labeling. The trained classifier is then applied to assign pseudo labels on the unlabeled data from the target domain D_U to achieve the pseudo labeled target-domain data D_{TP}. However, the label quality of D_{TP} tends to be relatively low due to the distribution discrepancy between the source and target domains. Therefore, it is crucial to generate high-quality target-domain data by combining the contexts from the labeled source-domain data D_S with the weak supervision in the pseudo labeled target-domain data D_{TP}.

4.2 Domain-Specific Feature Mask and Reconstruction

To automatically generate the target-domain data with fine-grained annotation, we propose a two-step approach as follows. First, a domain-specific segment set is extracted from the labeled and pseudo labeled datasets in both domains. Next, given a labeled sentence, we replace the domain-specific segments with $[mask]$ tokens, followed by feeding the masked domain-independent review to a pre-trained BART model (Lewis et al., 2020) to reconstruct its original text, domain prompt, and token-level labels.

4.2.1 Domain-Specific Segment Mask

Because reviews in different domains contain different aspects, opinions, and expression patterns, we define the text segments occurring more frequently in one domain as domain-specific segments or features. To obtain these domain-specific features, we introduce a frequency-ratio method based on (Li et al., 2018a). Specifically, we segment all sentences into word segments of different lengths, and then calculate the relative frequency of the n-gram segment w in the dataset D_v as follows:

$$s(w, D_v) = \frac{count(w, D_v) + \lambda}{\left(\sum_{v' \in V, v' \neq v} count(w, D_{v'})\right) + \lambda}, \tag{1}$$

where $count(w, D_v)$ denotes the frequency of an n-gram w in D_v, $v \in V$, $V = \{S, TP\}$, and λ is the smoothing parameter.

Next, we filter these n-gram w based on the relative frequency as follows:

$$s(w, D_v) \geq \delta, \tag{2}$$

where δ is a specified threshold.

We regard the filtered n-gram segments as the domain-specific segment set M. Given a review, we exploit the Forward Maximum Matching algorithm to match the segments that appear in M, and then replace each matched word with a special token $[mask]$. For example, in Figure 1, given a sentence "*i like the spicy tuna roll*", if the matched segment is "*spicy tuna roll*", we obtain a corresponding masked sentence "*i like the* $[mask]$ $[mask]$ $[mask]$". It is worth noting that as long as one word of a domain-specific aspect phrase or opinion phrase is masked, we will mask the whole aspect or opinion phrase.

4.2.2 Reconstruction and Label Generation

According to the above steps, for each sample $(X, L) \in D_S \cup D_{TP}$, a corresponding masked tuple $(\tilde{X}, L)$ can be obtained as the model input, where the masked sentence is $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n]$ and the label sequence for each word is $L = [l_1, l_2, ..., l_n]$. To reconstruct the original text as well as its corresponding labels, we implement several modification on BART.

Encoder: In addition to the word embedding and position embedding layers in BART, we also establish a label embedding layer:

$$E_x = TokenEmb([\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n]), \tag{3}$$

$$E_l = LabelEmb([l_1, l_2, ..., l_n]), \tag{4}$$

where $E_x \in R^{n \times d}$, $E_l \in R^{n \times d}$, and d is the dimension of the embedding. The output of the hidden state can be formulated as:

$$H = BartEncoder(E_x + E_l), \tag{5}$$

where $H \in R^{n \times d'}$, d' denotes the hidden dimension.

Decoder: To inform the BART model to distinguish the features from different domains, we insert a domain label tuple $([source], O)$ or $([target], O)$ at the beginning of the decoder, which can be regarded as a domain prompt. For each time step t, the decoder takes the previous decoder predictions

Algorithm 1 Training Procedure

```
Require: $D_S$: source labeled dataset; $D_U$: target unlabeled dataset; $K$: number of training iterations for BART;
Train a base classifier on dataset $D_S$
Get the target pseudo labeled dataset $D_{TP}$ by assigning pseudo labels on $D_U$
Extract and construct domain-specific n-gram segments set $M$ from $D_S \cup D_{TP}$ using Eq.(1),Eq.(2)
for $i \leftarrow 1$ to $K$ do
    Select an example$(X, L)$,$d$ from $D_S \cup D_{TP}$,where $d \in \{[source], [target]\}$
    $(\tilde{X}, L) \leftarrow SegmentMask(X, L, M)$
    $(X^{'}, L^{'}) \leftarrow ([d] + X, [O] + L)$
    $\theta \leftarrow BART((\tilde{X}, L), (X^{'}, L^{'}))$
end for
Return $\theta$
```

($x_{<t}$, $l_{<t}$) and the encoder output H as inputs to get the possibility of the next token and the next token label with two separate linear layers:

$$P(x_t|x_{<t}, l_{<t}, H) = \text{Softmax}(W_x z_t + b_x), \quad (6)$$

$$P(l_t|l_{<t}, x_{<t}, H) = \text{Softmax}(W_l z_t + b_l), \quad (7)$$

where $W_x \in R^{|V_x| \times d}$, $W_l \in R^{|V_l| \times d}$, and $|V_x|$ and $|V_l|$ refer to the dictionary size and the number of label types, respectively. The hidden layer vector z_t for the time step t is as follows:

$$z_t = BartDecoder(E_t), \quad (8)$$

$$E_t = TokenEmb(x_{t-1}) + LabelEmb(l_{t-1}). \quad (9)$$

For each sample, we calculate the negative log-likelihood loss for tokens and the label sequence, respectively:

$$\mathcal{L}_x = -\sum_{t=1}^{n+1} \log(P(x_t|x_{<t}, l_{<t}, H)), \quad (10)$$

$$\mathcal{L}_l = -\sum_{t=1}^{n+1} \log(P(l_t|l_{<t}, x_{<t}, H)). \quad (11)$$

The final training loss consists of the addition of two parts:

$$\mathcal{L} = \mathcal{L}_x + \mathcal{L}_l. \quad (12)$$

Algorithm 1 details the training procedure of our method.

4.3 Cross-Domain Data Generation

During the inference stage, given a source-domain labeled sentence $(X, L) \in D_S$, we employ the same mask strategy to remove its domain-specific features. Then, the masked tuple $(\tilde{X}, L)$ is fed to the BART encoder. Different from the decoder in the training phase, we only provide $([target], O)$ as the domain prompt to decode a target-domain sentence based on the auto-regressive way and jointly predict its token-level labels. To generate more samples and further increase the diversity of generated samples, we introduce several strategies as follows. For each sample $(X, L) \in D_S$, we repeat the mask step three times to get three masked tuple $(\tilde{X}, L)$ with different seeds. Specifically for the matched segments without any aspects or opinions, we randomly mask them with a probability of 60%. Thus, we can obtain three different target reviews from one source labeled review.

Dataset	Domain	Sentence	Train	Test
R	Restaurant	5841	4381	1460
L	Laptop	3845	2884	961
D	Device	3836	2877	959

Table 1: The statistics of our datasets.

4.4 Post-Processing

We post-process our generated target labeled data with the following steps: 1) Delete sentences with incorrect labels that do not follow the BIO schema; 2) Use a base classifier trained on the labeled source-domain data to assign labels on the generated target-domain data. Delete sentences whose assigned tags are inconsistent with generated ones.

4.5 Training for the Main Task

Finally, we can obtain the high-quality target-domain data with fine-grained annotation generated from our approach. We then use these generated samples for the main task of aspect and opinion co-extraction. Similar to Section 4.1, we employ a pre-trained BERT model (Devlin et al., 2019) with a CRF layer as the main task classifier, and train it on the generated data only. Based on the trained classifier, we evaluate its performance on the test set of the target domain.

5 Experiments

5.1 Datasets

We use three benchmark datasets from different domains, namely Restaurant (R), Laptop (L) and Device (D). R and L are two combination datasets from SemEval 2014 and 2015 (Pontiki et al., 2014, 2015). D consists of reviews from digital device collected by (Hu and Liu, 2004). Following previous works (Wang and Pan, 2018; Pereg et al., 2020; Chen and Qian, 2021), we use three different seeds to split the dataset from each domain into training set and testing set by a ratio of 3:1. The statistics are shown in Table 1. We report the average results on different splits.

5.2 Experimental Setting

5.2.1 Cross-Domain Data Augmentation

For domain-specific segment mask module, we set the length of n-gram segment w to $n \in [1, 4]$ and the relative frequency threshold δ to 10. A pre-trained sequence-to-sequence model BART is fine-tuned on $D_S \cup D_{TP}$ for 3 training epochs with batch size set to 8. We use Adam as the optimizer with a learning rate of 5e-5. The fine-tuned BART is used to generate triple target-domain labeled data conditioned on the source-domain labeled data.

5.2.2 Aspect and Opinion Co-Extraction

As mentioned in Section 3, we formulate aspect and opinion co-extraction as a sequence labeling task. We use BERT$_{\mathrm{E}}$+CRF as the base classifier to assign pseudo labels in Section 4.1 and Section 4.4, which consists of the uncased BERT$_{base}$ (Devlin et al., 2019) model post-trained on a combined data of E-commerce reviews from the Amazon Electronics dataset and the Yelp Challenge (Xu et al., 2019) and a CRF layer (Lafferty et al., 2001). For the main task, we also use the BERT$_{\mathrm{E}}$+CRF classifier and train it on our generated data. We use the Adam optimizer fro both BERT and CRF with different learning rates of 3e-5 and 0.02, respectively. Moreover, we employ the early stopping strategy, and stop the training procedure of the BERT$_{\mathrm{E}}$+CRF classifier after 2 epochs in both pseudo label annotation and main task training stages. We finally report the average Micro-F1 of three different splits for aspect and opinion co-extraction.

5.3 Compared Methods

We divide all comparison systems into four parts:

Part 1 denotes hand-built feature-based methods. CrossCRF (Jakob and Gurevych, 2010) uses a linear-chain CRF with some manual features including POS tags and dependency relations to bridge the domain gap. RAP (Li et al., 2012) proposes a Relational Adaptive bootstraPping method to extract aspects and opinions based on source labeled data and pre-defined syntactic rules.

Part 2 denotes word2vec-based methods. Hier-Joint (Ding et al., 2017) is a cross-domain RNN model, which exploits auxiliary labels consisting of aspect terms extracted by manually designed syntactic rules and opinion seeds. RNSCN (Wang and Pan, 2018) and TRNN-GRU (Wang and Pan, 2020) integrate syntactic relations by constructing the dependency tree. TIMN (Wang and Pan, 2019) proposes a Transferable Interactive Memory Network that can learn shared representations across domains effectively. SemBridge (Chen and Qian, 2021) is a CNN-based model, which links the source and target domains with semantic bridges by retrieving transferable semantic prototypes based on syntactic roles.

Part 3 denotes pre-trained model-based methods. BERT$_{\mathrm{B}}$+CRF consists of the uncased BERT$_{base}$ model and a CRF layer. SA-EXAL (Pereg et al., 2020) achieves domain adaptation by combining the pre-trained BERT model with a syntactic-aware attention mechanism. BERT$_{\mathrm{E}}$+CRF is the base classifier introduced in Section 5.2.2. Both the BERT$_{\mathrm{B}}$+CRF and BERT$_{\mathrm{E}}$+CRF are only trained on the labeled source-domain data. CDRG (Yu et al., 2021) is a data augmentation-based adaptation method, which generates labeled target-domain data by replacing source-specific aspects and opinions in source domain reviews with target-specific aspects and opinions. CDRG-Merge trains the BERT$_{\mathrm{E}}$+CRF classifier on the combination of labeled source-domain data and generated target-domain data for aspect and opinion co-extraction.

Part 4 denotes the proposed Generative Cross-Domain Data Augmentation Framework, i.e., GCDDA, which designs a BART-based generative model (Lewis et al., 2020) to convert the labeled source-domain review to a target-domain review while predicting its corresponding token-level labels. We train the BERT$_{\mathrm{E}}$+CRF classifier on the generated target-domain data for aspect and opinion co-extraction.

5.4 Main Results

The comparison results of all methods are shown in Table 2. We can clearly observe that our approach achieves the best performance in terms of the average Micro-F1. Specifically, GCDDA outperforms the baseline approach BERT$_{\mathrm{E}}$+CRF by 4.90% and 0.36% for aspect extraction and opinion extraction, respectively. Compared with the significant improvement on aspect extraction, we achieve a relatively low result on opinion term extraction. We attribute it to the following reasons: 1) the aspect terms across two domains tend to be quite different and have small overlaps; and

Model	R->L		R->D		L->R		L->D		D->R		D->L		AVE	
	AS	OP	AS	OP	AS	OP	AS	OP	AS	OP	AS	OP	AS	OP
CrossCRF	19.72	59.20	21.07	52.05	28.19	65.52	29.96	56.17	6.59	39.38	24.22	46.67	21.63	53.17
RAP	25.92	62.72	22.63	54.44	46.90	67.98	34.54	54.25	45.44	60.67	28.22	59.79	33.94	59.98
Hier-Joint	33.66	-	33.20	-	48.10	-	31.25	-	47.97	-	34.74	-	38.15	-
RNSCN	40.43	65.85	35.10	60.17	52.91	72.51	40.42	61.15	48.36	73.75	51.14	71.18	44.73	67.44
TRNN-GRU	40.15	65.63	37.33	60.32	53.78	73.40	41.19	60.20	51.17	74.37	51.66	68.79	45.88	67.12
TIMN	43.68	68.44	35.45	59.05	54.12	73.69	38.63	62.22	53.82	76.52	52.46	69.32	46.36	68.21
SemBridge	50.67	71.51	43.34	63.46	63.04	80.48	44.91	**64.15**	60.19	80.21	53.02	72.63	52.53	72.08
BERT$_{\mathrm{B}}$+CRF	40.71	74.96	40.28	64.73	40.31	80.74	44.24	59.22	59.35	81.17	53.46	75.05	46.39	72.65
SA-EXAL	47.59	75.79	40.50	63.33	54.67	80.05	42.19	60.19	54.54	71.57	47.72	63.98	47.87	69.15
BERT$_{\mathrm{E}}$+CRF	52.77	75.94	43.65	**66.36**	50.08	82.39	**45.47**	60.29	64.21	81.79	**58.75**	76.15	52.49	73.82
CDRG-Merge	58.23	76.08	37.96	62.19	**72.88**	82.34	40.62	59.04	66.79	82.23	54.26	76.42	55.12	73.05
GCDDA (Ours)	**66.56**	**77.63**	**44.80**	64.86	62.22	**82.67**	45.11	60.72	**68.23**	**82.44**	57.44	**76.75**	**57.39**	**74.18**

Table 2: Comparison results of different methods for aspect and opinion extraction on Micro-F1. The AVE above means averaged scores on all domain pairs. The best scores are in bold. All methods are grouped into four parts. The last part is our approach.

Model	F1		BLEU	Perplexity
	ASP	OP		
Source	52.49	73.82	100.00	27.48
CDDA-MLM	56.66	73.66	90.59	28.49
GCDDA	**57.39**	**74.18**	**89.72**	**28.14**

Table 3: Comparison results of two different Cross-Domain Data Augmentation methods on several evaluation metrics for generated data. Source denotes the source-domain labeled data. CDDA-MLM denotes the data generated by BERT. GCDDA is our method.

2) the opinion terms across two domains tend to have a significant overlap, e.g., great, good, terrible usually occurring in both domains. The similar trend can also be observed in the state-of-the-art approach SemBridge (Chen and Qian, 2021). Moreover, compared with SemBridge which is based on static word vectors, our method GCDDA improves its performance by 4.86% and 2.10%, which proves the effectiveness of the pre-trained language model. Lastly, GCDDA outperforms our recent data augmentation-based method CDRG-Merge by 2.27% and 1.13% on the two tasks, respectively. All these observations show the effectiveness of our proposed approach.

5.5 Evaluation on Generated Samples

In this section, we conduct several experiments to evaluate the quality of our generated target labeled data. For better comparison, we construct another cross-domain data augmentation approach (CDDA-MLM), which can be regarded as a variant of our GCDDA approach based on the masked language model BERT. CDDA-MLM adopts the same mask strategy and post-processing as GCDDA, and we post-train BERT on the unlabeled data from the target domain with the standard MLM objective. Moreover, we employ it to replace the $[mask]$ token with target words. Unlike GCDDA, there is no label embedding and label prediction in CDDA-MLM. In addition, the original labels can be directly transferred to the target sentence due to the word-to-word generation.

As shown in Table 3, we evaluate the generated data with three metrics including F1-score, BLEU[1], and Perplexity[2], and report the average results on six cross-domain pairs. For F1-score, we observe that our proposed model GCDDA achieves the best result. It is due to the fact that we add the label embedding for GCDDA to capture the consistency between tokens and labels, which leads to more precise and controllable generation of target aspects or opinions in the appropriate place conditioned on the source-domain context and its labels. Additionally, compared with only using source labeled data, the significant improvement shows that our method GCDDA is indeed helpful for the domain adaptation problem.

Moreover, GCDDA obtains lower BLEU than CDDA-MLM, which indicates that our method generates more diverse sentences. This is because that GCDDA decodes the whole sentence while CDDA-MLM generates new words only on the $[mask]$

[1] We choose the source data as the reference and use BLEU-1 as the final result. The lower value means more differences from the source data.

[2] We calculate the perplexity with a pre-train language model GPT. The more fluent sentence has a lower value.

	Source	CDDA-MLM	GCDDA
R->L	0.3502	0.1796	**0.1128**
R->D	0.3897	0.2284	**0.1598**
L->R	0.3306	0.2315	**0.1488**
L->D	0.0826	0.0667	**0.0575**
D->R	0.3888	0.2685	**0.1429**
D->L	0.1078	0.0812	**0.0701**
AVE	0.2749	0.1760	**0.1153**

Table 4: Maximum Mean Discrepancy (MMD) between the target-domain test set and the dataset generated from CDDA-MLM, GCDDA. The lower is better.

Source	the food was extremely tasty , creatively presented and the wine excellent .
CDDA-MLM	the screen is extremely well ##sty , creative ##ly presented and the are excellent .
GCDDA	the screen is extremely lightweight , creatively presented and the battery excellent .
Source	the food is prepared quickly and efficiently .
CDDA-MLM	the received is up quickly and efficiently .
GCDDA	the computer runs quickly and efficiently .
Source	i recommend the jelly fish , drunken chicken and the soupy dumplings , certainly the stir fry blue crab .
CDDA-MLM	i recommend the color ##screen , other 8 , the really key ##ba also , certainly the anti the hybrid ##nes .
GCDDA	i recommend the backlit , wireless keyboard , the pre-loaded mbp applications , certainly the toshiba satellite .
Source	we had the lobster sand and it was fantastic .
CDDA-MLM	i had it head basement and it was fantastic .
GCDDA	i bought this netbook and it was fantastic .
Source	instead of wasting your time here : support restaurants that care about food .
CDDA-MLM	instead of wasting your time here : call people that care about you .
GCDDA	instead of wasting your time here : buy computers that care about gaming .

Table 5: Comparison examples from source-domain labeled data and generated data by CDDA-MLM, GCDDA. Words in blue denote the aspect terms and words in red denote the opinion terms. All generated samples are extracted from domain pair (R->L).

R->L	dvd games, wireless devices, wireless keyboard, toshiba speakers, digital graphics, hp g73, hard drive applications, pre-loaded applications,light weight, precise, over weight, ingenious, well built
L->R	corn sauce, house vibe, wine quality, chili sauce, grilled meat, outdoor dining atmosphere, service provider, internal decor, heating system, steak platter, cleaner, much larger, softer
R->D	ipod plug, battery capacity, radio signal, sound quality of music, built-in speakerphone, usb connection, sony camera, black card with fm transmitter, 2mb flash card, amazing great , user friendly

Table 6: New aspect and opinion terms generated by GCDDA which have not appeared in the training data of both domains. These examples are extracted from cross-domain pairs (R->L, L->R, R->D) with three different target domains.

position. Furthermore, each masked token is predicted independently in CDDA-MLM, which may generate incoherent collocation as the length of the mask segment grows. Instead, the auto-regressive model BART in GCDDA takes into account the coherence and the consistency of sentences. Thus, GCDDA obtains a better result than CDDA-MLM in terms of Perplexity.

Moreover, we measure the distribution distance between our generated target dataset and the test set from the target domain. Specifically, we use the BERT model to obtain the representation of each sentence in two datasets, and then calculate the distribution distance by Maximum Mean Discrepancy (MMD[3]). The comparison results are shown in Table 4. The results, on the other hand, proves that our method can be used to alleviate the domain discrepancy issue in domain adaptation.

Case Study: In Table 5, we show several examples generated by CDDA-MLM and GCDDA on the cross-domain pair R->L. It can be obviously found that our method produces more precise aspects or opinions related to the domain laptop. We summarize this into two reasons: 1) integrating the BART model with label embedding has captured the consistency between token and its label; and 2) the other is owed to the domain prompt in the front of the decoder, which controls the domain where the generated words come from. Moreover, the label prediction from the decoder makes it possible to generate more flexible sentence while obtaining the adaptive token tags. We also present aspects and opinions extracted from our generated dataset in Table 6. These words or phrases have not appeared in the source and target training datasets, which shows that the generated dataset expands the vicinity distribution for aspect/opinion identification. To sum up, our method GCDDA not only exploits the rich contexts from the source domain, but also produces new aspects and opinions to generate more diverse target-domain samples. The two factors are integrated to train a more robust model for the target domain.

Parameter Study: As mentioned in Section 4.3, we mask sentences with different seeds to produce more target samples. Figure 2 shows the average F1 score of generating different sizes of augmented data from CDDA-MLM and GCDDA for aspect and opinion co-extraction. It can be seen that the two models consistently achieve the best result when the size is 3, i.e., generating three different target reviews from one labeled source review.

[3] Maximum Mean Discrepancy (MMD) reflects the similarity of distribution between two datasets.

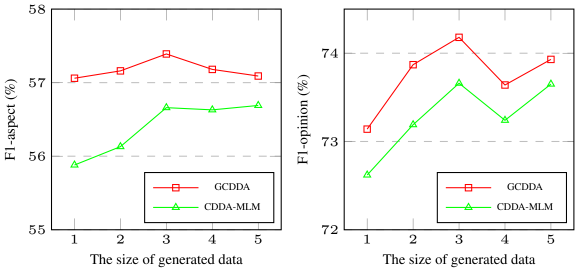

Figure 2: Average F1-score on different size of samples generated by CDDA-MLM and GCDDA for aspect and opinion co-extraction.

Index	Method	F1	
		ASP	OP
0	GCDDA (ours)	57.39	74.18
1	**W/O** data augmentation	54.27	73.56
2	BART ⇒ Bi-LSTM	56.78	73.58
3	**W/O** data filter	55.50	72.71
4	1∼4-gram ⇒ 1-gram	56.92	73.55
5	**W/O** source data	56.71	73.80
6	**W/O** encoder label embedding	55.88	73.04

Table 7: Ablation study of our Generative Cross-Domain Data Augmentation method.

5.6 Ablation Study

Finally, to better analyze the effectiveness of each key component in our method, we conduct the ablation study and show the results in Table 7.

We first remove the whole cross-domain data augmentation framework in index 1, and only use the pseudo labeled data D_{TP} to train the main task classifier. It shows a considerable performance drop due to the label noise in D_{TP}. This indicates that cross-domain data augmentation needs to be applied, which can produce high-quality target samples by combining the contexts from D_S with the weak supervision signals in D_{TP}. In index 2, we replace the pre-trained model BART with Bi-LSTM. Compared with Bi-LSTM, the pre-trained BART model obtains a better result, because it possesses rich language knowledge and a more complex model structure for generation tasks.

Index 3 proves the importance of data filter in the post-processing step. As shown in Table 5, our generated samples usually have the same syntactic structure as the labeled source data. Therefore, the base classifier trained on D_S can be employed to filter out some noisy samples in our generated target-domain data. In index 4, compared with our 1∼4-gram mask, the single word mask strategy limits the diversity of generated sentences, and thus decreases the model performance. Moreover, the performance drops when we remove the labeled source data for the reconstruction task, which can capture the domain-invariant features for the BART model. Finally, removing the label embedding layer in the BART encoder also leads to the performance decline.

6 Conclusion

In this paper, we proposed a Generative Cross-Domain Data Augmentation Framework for unsupervised domain adaptation, which leverages the labeled source-domain data to directly generate labeled target-domain data based on a fine-tuned sequence-to-sequence model BART. Experiments on three benchmark datasets show that our generative approach generates high-quality target-domain data with fine-grained annotation and outperforms previous domain adaptation methods for aspect and opinion co-extraction.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments. This work was supported by the Natural Science Foundation of China (62076133 and 62006117), and the Natural Science Foundation of Jiangsu Province for Young Scholars (BK20200463) and Distinguished Young Scholars (BK20200018).

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020. Local additivity based data augmentation for semi-supervised ner. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1241–1251.

Zhuang Chen and Tieyun Qian. 2021. Bridge-based active domain adaptation for aspect term extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 317–327.

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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Ying Ding, Jianfei Yu, and Jing Jiang. 2017. Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544, Florence, Italy. Association for Computational Linguistics.

Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. Kingdom: Knowledge-guided domain adaptation for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3198–3210.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*.

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

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Semantics-preserved data augmentation for aspect-based sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4417–4422.

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*, pages 1035–1045.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

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*, pages 7871–7880.

Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang, and Xiaoyan Zhu. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 410–419.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018a. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7056–7066.

Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang. 2019. 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)*.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018b. Hierarchical attention transfer network for cross-domain sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Pengfei Liu, Shafiq Joty, and Helen 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*, pages 1433–1443.

Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760.

Oren Pereg, Daniel Korat, and Moshe Wasserblat. 2020. Syntactically aware cross-domain aspect and opinion terms extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1772–1777.

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, 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, Dublin, Ireland. Association for Computational Linguistics.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.

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 (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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7192–7199.

Wenya Wang and Sinno Jialin Pan. 2020. Syntactically meaningful and transferable recursive neural networks for aspect and opinion extraction. *Computational Linguistics*, 45(4):705–736.

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 AAAI Conference on Artificial Intelligence*, volume 31.

Meixi Wu, Wenya Wang, and Sinno Jialin Pan. 2020. Deep weighted maxsat for aspect-based opinion extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5618–5628.

Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 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, Volume 1 (Long and Short Papers)*, pages 2324–2335.

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)*, pages 2979–2985.

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.

Jianfei Yu and Jing Jiang. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 236–246.

Jianfei Yu, Jing Jiang, and Rui Xia. 2018. Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):168–177.

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

Yftah Ziser and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1241–1251.