Refining and Synthesis: A Simple yet Effective Data Augmentation Framework for Cross-Domain Aspect-based Sentiment Analysis

Anonymous ACL submission

Abstract

Aspect-based Sentiment Analysis (ABSA) is extensively researched in the NLP community, yet related models face challenges due to data sparsity when shifting to a new domain. Hence, data augmentation for cross-domain ABSA has attracted increasing attention in recent years. However, two key points have been neglected in prior studies: First, target domain unlabeled data are labeled with pseudo labels by the model trained in the source domain with little quality control, leading to inaccuracy and error propagation. Second, the label and text patterns of generated labeled data are monotonous, thus limiting the robustness and generalization ability of trained ABSA models. In this paper, we aim to design a simple yet effective framework to address the above shortages in ABSA data augmentation, called Refining and Synthesis Data Augmentation (RSDA). Our framework roughly includes two steps: First, it refines generated labeled data using a natural language inference (NLI) filter to control data quality. Second, it synthesizes diverse labeled data via novel label composition and paraphrase approaches. We conduct experiments on 4 kinds of ABSA subtasks, and our framework outperforms 7 strong baselines, demonstrating its effectiveness.

AESC (source: laptop → target: restaurant)	
Source Domain Labeled Data (L)	•There is no number **pad** to the right of the keyboard.
Target Domain Unlabeled Text (R)	• The worst **pad tai**, I've ever had.
Target Domain Pseudo Label	• <neg> **pad**
Target Domain Generated Text	• Sometimes you have to tap the pad.

(a)

ATSE (source: laptop → target: restaurant)	
Target Domain Unlabeled Text (R)	• They pray to their Food Gods to make them into a **good pizza** like VT's . • Right off the L in Brooklyn this is a **nice cozy place** with **good pizza** .
Target Domain Pseudo Labels	• <pos> pizza <opinion> good • <pos> pizza <opinion> good
Target Domain Generated Texts	• The **pizza** is **good**. • The **pizza** is **good**.

(b)

Figure 1: The examples of error propagation and limited diversity in previous data augmentation work.

1 Introduction

Aspect-based Sentiment Analysis (ABSA) is a fundamental sentiment analysis task that aims to analyze sentiments at the aspect level (Liu, 2012; Xue and Li, 2018). It usually involves extracting several sentiment elements, including aspects, opinions, and sentiment polarities. For example, given a sentence: "*It is the best sushi I ever had*", the aspect term is "*sushi*", the corresponding opinion term is "*best*" and the sentiment polarity is "*positive*". ABSA has attracted more and more attention in the past decade (Nguyen and Shirai, 2015; Zhang et al., 2023b), with the development of deep learning, many models and methods can achieve good results on the aspect-level sentiment analysis dataset (Yadav et al., 2021; Zhang et al., 2023b). Most methods for model training only use same domain data and require fine-grained labeled data (Ding et al., 2017). This is problematic in nascent domains with little labeled data, impeding robust performance. Some studies focus on developing models with domain migration capabilities to address these challenges (Zhang et al., 2022) . Other works employ domain adaptation technology to transfer learned knowledge from labeled source domains to unlabeled target domains (Deng et al., 2023). However, the majority of these studies are based on discriminative models (Zhang et al., 2021a), necessitating customized design for specific tasks. In addition, other works resort to domain-specific dictionaries, using rule-based or neural network-based methods (Marcacini et al., 2018; Howard et al., 2022) to obtain external semantic dictionaries. While these approaches have demonstrated commendable performance on particular datasets, their excessive reliance on external knowledge impairs their generalization capacity.

Recently, the methods, which integrate various tasks into a unified framework by formalizing each task as a sequence-to-sequence problem, have achieved promising results (Wang and Wan, 2023). In the cross-domain low-resource scenario, the feasibility of the cross-domain data augmentation method based on such framework has also been verified (Yu et al., 2023; Ghosh et al., 2023a). Deng et al. (2023) proposes a data augmentation framework, which extracts pseudo-labels from target domain sentences and then generates new sentences based on these pseudo-labels to synthesize labeled data. Despite achieving promising results, the existing data augmentation frameworks primarily have the following shortcomings:

- **Low-quality Samples and Error Propagation** The target data generated by pseudo-labels is error-prone because the extraction model is trained using labeled data from the source domain. Figure 1(a) shows an incorrectly generated sentence due to error propagation caused by a pseudo label.
- **Limited Data Diversity** The diversity of the generated labeled data in the target domain is limited due to the constraints imposed by the scales of text generation models and the categories of pseudo-labels. Figure 1(b) shows that when two identical pseudo-labels are extracted, it often results in the generation of identical new sentences, even if their sources are different.

Towards this end, we propose a novel two-step data augmentation framework for cross-domain ABSA tasks named **Refining and Synthesis Data Augmentation** (**RSDA**). In the first step, our framework follows previous work (Deng et al., 2023) by extracting a pseudo label l from an unlabeled sentence t in the target domain. Subsequently, it generates a new sentence t' aligned with the pseudo label l, thereby producing a labeled sample (t', l) in the target domain. Then, our framework further employs an approach based on natural language inference (NLI), named the NLI filter (Sileo, 2023), to eliminate invalid samples by determining whether t and t' are in an entailment relationship. By employing this approach, we can obtain higher-quality labeled samples in the target domain.

In the second step, we design two novel diversity enhancement modules to mitigate the duplication and oversimplification of model-generated target domain labeled samples. The first module is called *composition-based diversity enhancement*, which combines two selected labels into a longer one and then generates a new sentence by the generation model. The compositions of various labels will definitely increase the diversity of generated data. On the other hand, we propose a *paraphrase-based diversity enhancement* module that tackles data augmentation diversity from two perspectives, namely *label-variant paraphrase* and *label-invariant paraphrase*. The former directly paraphrases the unlabeled text in the target domain and extracts pseudo-labels from it, similar to the process in the first step. In contrast, the latter focuses on paraphrasing the target domain labeled text while retaining the original labels but altering their contextual representation. These two paraphrase methods complement each other effectively.

To validate the effectiveness of our framework, we conduct extensive cross-domain experiments on 4 ABSA subtasks. Our framework outperforms the strong baselines by at least 1.64%, 1.39%, 1.45% and 2.04% in averaged F1s of 4 subtasks.

Our main contributions can be summarized as follows:

- We introduce RSDA, a novel data augmentation framework designed for cross-domain ABSA. Unlike previous studies, RSDA prioritizes both data quality and diversity, which have been neglected in prior studies.
- To address the issue of limited diversity in generated data, our diversity enhancement method focuses on improving diversity from two angles: information density and expression variety.
- Our framework has been tested on 32 cross-domain experiments and the superior performances compared with 7 strong baselines demonstrate its effectiveness.[1]

2 Related work

2.1 Aspect-Based Sentiment Analysis

Aspect-based Sentiment Analysis (ABSA) (Liu, 2012; Xue and Li, 2018) is a well-established sentiment analysis task which encompasses various subtasks such as Aspect Sentiment Classification

[1]Our codes are available at RSDA for reviewing. They will be publicly released after the paper has been published.

Task	Output	Example
AE	(a)	I love **[pizza]**!
AESC	(s,a)	I love **[pizza]**$_{pos}$!
AOPE	(a,o)	I **[love]** **[pizza]**!
ASTE	(s,a,o)	I **[love]** **[pizza]**$_{pos}$!

Table 1: Four ABSA subtasks were investigated in this paper, where a, s and o denote aspect, sentiment polarity, and opinion respectively.

(AE), Aspect Extraction and Sentiment Classification (AESC), Aspect Opinion Pair Extraction (AOPE), and Aspect Sentiment Triple Extraction (ASTE) (Yan et al., 2021). Recent advancements, particularly with models like Bart or T5, have shifted towards end-to-end architectures (Lewis et al., 2020; Raffel et al., 2020). This transition has streamlined task forms, enhancing model adaptability to ABSA's multiple subtasks (Lu et al., 2022). However, in cross-domain settings, these methods encounter challenges due to limited training data in the target domain. The domain adaptation capability of these models requires further investigation.

2.2 Cross Domain Data Augmentation

Data augmentation is a widely used technique to address training data scarcity (Fadaee et al., 2017; Chao et al., 2023), proven effective in computer vision (Ye et al., 2019). However, for textual data with discrete and complex semantics, augmentation becomes challenging. Current methods include rule-based synonym replacement (Rennes and Jönsson, 2021) through conditional constraint generation or template-based approaches (Chen et al., 2021). Yet, these methods often rely on rigid rules or fixed templates, limiting model generalization and sample diversity. In the realm of cross-domain data augmentation, some approaches generate new samples by regenerating pseudo labels (Toledo-Ronen et al., 2022; Deng et al., 2023). While showing progress, these methods often neglect the issue of pseudo-label error propagation and struggle to accurately simulate real data distribution due to model and data limitations.

3 Methodology

3.1 Problem Definition

In this work, we focus on the unsupervised cross-domain setting (Sharma et al., 2018; Zhang et al., 2023a), which means that we only have labeled data in the source domain, namely $D^s = \{t, l\}$ where t and l denote a sentence and its corresponding label (e.g., aspect, opinion, sentiment polarity). Only unlabeled data is available in the target domain, denoted as $D^t = \{t\}$. The objective is to leverage the training data from both the source and target domains to predict the labels of the test set in the target domain. To validate the effectiveness of our framework, we employ 4 ABSA subtasks for experiments, as shown in Table 1.

3.2 Overview

Our RSDA framework mainly consists of two steps as shown in Figure 2: (1) The first step is *data generation and quality control*. Firstly, we obtain pseudo labels and corresponding generated samples of the target domain from the extraction and generation models trained on the source domain. Then, we use a Natural Language Inference (NLI) model as a filter to remove incorrect samples. (2) The second step is *data diversity augmentation*. In this step, we employ composition-based diversity enhancement to make generated samples contain multiple aspects to improve information density, and paraphrase-based diversity enhancement to generate new labels or change their context.

3.3 Data Generation and Quality Control

Data Generation: Following previous work (Deng et al., 2023), we train both the label extraction model $l = M_e(t)$ and text generation model $t = M_g(l)$ using the source domain data $D^s = \{t, l\}$. Both of them are based on T5-base (Raffel et al., 2020). Then, we utilize the extraction model M_e to extract the pseudo label l' from a target domain sentence t. Based on the pseudo label l', the sample generation model M_g can generate a new sentence t'. After data generation, we obtain a new target domain labeled dataset $D^t_p = \{t', l'\}$.

NLI-based Quality Control: As the generation model was trained on the source domain, it tends to generate text more aligned with the source domain, resulting in less fluent data. Moreover, the noise introduced by the extracted pseudo-labels can propagate into the generated text samples. To address these problems, we employ a Natural Language Inference (NLI) filter (Sileo, 2023) for quality control in the generation of data.

Concretely, we take the original target domain text t as the premise and the newly generated text t' as the hypothesis. The NLI filter can determine the relationship between a pair of premises and

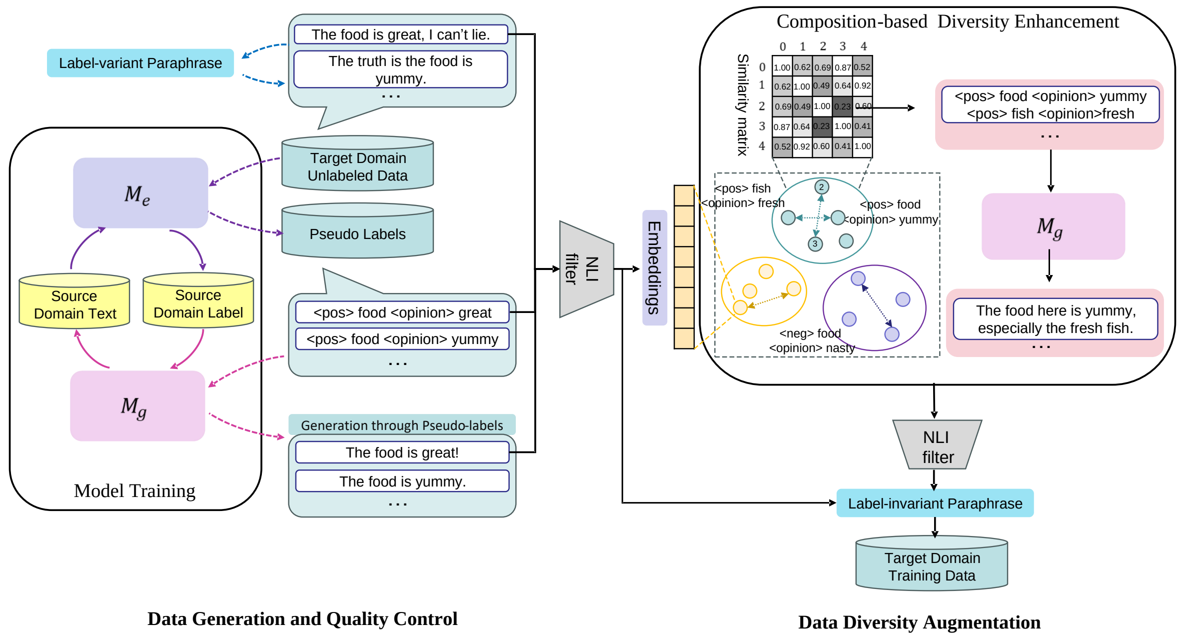

Figure 2: Overview of our proposed RSDA framework which includes two steps. We take examples from the ASTE task in the restaurant domain. In addition, M_e and M_g denote extraction and generation models where the solid line represents the training process and the dashed line represents data generation.

hypotheses, formulated as:

$$P(y \mid t, t') = f(t, t') \tag{1}$$

where y can be *Entailment*, *Contradiction* and *Neutral*, f denotes the NLI filter. When a contradiction relation arises between t and t', it suggests that the generated text t' is less likely to be inferred from the original text, which should be filtered out. After quality control, certain problematic samples are removed and the remaining high-quality labeled data are denoted as $D_f^t = \{t', l'\}$. The examples are in Appendix B.1.

3.4 Data Diversity Augmentation

In the previous step, we filtered model-generated target domain-labeled samples through the NLI filter. However, based on our manual observation, we identified two shortages that should be improved:

(1) As shown in Figure 1(b), the generation model tends to generate simple or duplicated sentences due to training resource limitations.

(2) Although the quality of generated data can be enhanced using the NLI filter, it sacrifices text expression and pattern diversity by filtering out part of the samples.

To address these issues, we focus on diversifying the data from two dimensions in the second step, namely information density and expression variety.

3.4.1 Composition-based Diversity Enhancement

First, we perform lexical clustering using the labels l' of the labeled target domain data $D_f^t = \{t', l'\}$. Specifically, we employ MiniLM-L6 [2] from Sentence Transformers to encode a label l'_i into its vector representation $h_{l'_i}$. Subsequently, the K-means clustering algorithm is applied to partition labels into K clusters, where the value of K is determined by the silhouette coefficient method (Rousseeuw, 1987). The calculation method is as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2}$$

$$K = \arg\max_k \left(\frac{1}{|D|} \sum_{i=1}^{|D|} s(i) \right) \tag{3}$$

where $a(i)$ and $b(i)$ represent the average distance between sample i and all other points within the same cluster and in different clusters, respectively. The silhouette score for sample i is denoted by $s(i)$, and $|D|$ represents the total number of samples.

Then, the semantic similarity between the text of each pair of data points in the same cluster is

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

measured by the cosine similarity, calculated as:

$$Sim(t'_i, t'_j) = \frac{h_{t'_i} \cdot h_{t'_j}}{\|h_{t'_i}\| \|h_{t'_j}\|} \quad (4)$$

where $h_{t'_i}$ and $h_{t'_j}$ denote the vector representation of the text pair t'_i and t'_j also encoded by MiniLM-L6. Concretely, we select the data points with the lowest semantic similarity to increase the diversity of data.

Next, we concatenate l'_i and l'_j and feed them into the generation model to obtain a synthesized text $t'' = M_g(l'') = M_g(l'_i \oplus l'_j)$. For example, after clustering and similarity calculation, we select two labels "*<pos> food <opinion> yummy*" and "*<pos> fish <opinion> fresh*". Then we concatenate them and input them into M_g to generate "*The food in here is yummy, especially the fresh fish*". By combining different labels, information density, and label diversity can be enhanced.

3.4.2 Paraphrase-based Diversity Enhancement

For paraphrase-based diversity enhancement, we design two methods to augment the target domain labeled data. Our core idea is to use paraphrase to rewrite labels or their context, thus generating new data. The details are explained below.

Label-variant Paraphrase Due to the non-linguistic nature of labels, we perform an indirect approach to implement label-variant paraphrase. As shown in Table 2, we apply a paraphrasing tool (Vladimir Vorobev, 2023) to the original target domain unlabeled text t and a piece of new paraphrased text can be generated, called t'. Then a pseudo label l' can be extracted using the extraction model M_e introduced in Section 3.3. Note that in this phase, because the paraphrase tool is directly applied to the raw text, all the words could be rewritten thus the extracted pseudo label could also be different from the original one. Afterwards, the generation model M_g will generate a new sentence t'' based on the pseudo label l'. This approach not only aligns with the label-invariant paraphrase procedure but also enhances the diversity and expression of D_p^t.

Label-invariant Paraphrase We also applied paraphrasing to the target domain labeled samples generated in previous steps to enrich their text patterns and avoid simple sentence structures. As shown in Table 2, we utilize prompts to encourage the paraphrase tool to include the label l' when it rewrites the text t' as t''. Then, we use post-processing methods to make sure that the paraphrased text t'' includes the label l'. In this phase, we try to keep the label in a target domain labeled sample unchanged and meanwhile transform its context, thus more diverse data could be synthesized.

4 Experiments

4.1 Experimental Settings

Datasets. To validate the effectiveness of our framework in cross-domain settings, we conduct extensive experiments on benchmark datasets from four domains: **Restaurant** (**R**), **Laptop** (**L**), **Device** (**D**) and **Service** (**S**). The specific dataset statistics are illustrated in Figure 7, which are widely-used for ABSA and sourced from the SemEval challenge 2014, 2015, and 2016 (Pontiki et al., 2014, 2015, 2016), reviews about digital devices (Toprak et al., 2010) and comments about web services (Hu and Liu, 2004).

Considering the domain similarities among datasets, we select several distinct source-to-target domain pairs for experimentation. In AE and AESC tasks, following prior work (Yu et al., 2021; Gong et al., 2020; Yu et al., 2023), we exclude experiments involving transfers between L and D domains due to their high similarity. Our experiments exclusively utilize the R and L datasets for AOPE and ASTE tasks, owing to limitations in data sources. Additionally, the experiments between R14, R15, and R16 were omitted because they share the same domain (Deng et al., 2023). Consequently, there are a total of 10 transfer experiments for AE, AESC and 6 transfer experiments for AOPE and ASTE.

Evaluation Metrics. We choose Micro-F1 as the primary evaluation metric for our experiments, considering a predicted label correct only when it fully matches the gold label (Lu et al., 2022). Additionally, to assess the diversity of enhanced samples, we employ diversity evaluation metrics, following Ghosh et al. (2023b); Yu et al. (2023). Specifically, $\mathcal{D}_a$ indicates the percentage of unique aspect terms among all aspect terms, while $\mathcal{D}_o$ represents the percentage of unique opinion terms among all opinion terms. For AE and AESC tasks, the diversity of generated samples is determined solely by $\mathcal{D}_a$, while for AOPE and ASTE tasks, the diversity is the average of $\mathcal{D}_a$ and $\mathcal{D}_o$.

	Before Paraphrase	After Paraphrase
Label-variant	**Text**: The **[cake]**$_{pos}$is **[yummy]**! **Label**: <pos> **cake** <opinion> **yummy**	**Text**: The **[cake]**$_{pos}$ of this restaurant is **[delicious]**! **Label**: <pos> **cake** <opinion> **delicious**
Label-invariant	**Prompt+Text**: screen, good must be included, now paraphrase:[The **screen** is **good**.] **Label**: <pos> **screen** <opinion> **good**	**Text**: After using it a couple of days, the **screen** is still **good**. **Label**: <pos> **screen** <opinion> **good**

Table 2: Examples for Paraphrase-based Diversity Enhancement.

AE	S→R	S→L	S→D	R→S	R→L	R→D	D→R	D→S	L→R	L→S	Avg.
BERT-UDA*	56.08	43.98	38.36	34.62	46.87	40.34	50.54	34.52	51.91	32.49	42.97
CDRG†	60.20	39.49	38.59	**49.97**	55.50	34.89	57.51	43.19	68.63	51.07	49.90
GAS*	54.61	35.12	35.81	30.99	43.50	39.29	53.40	33.34	49.06	29.64	40.48
DA^{2}LM†	**65.78**	44.96	**43.24**	43.41	54.55	**44.29**	63.86	38.20	68.72	41.06	50.80
BGCA*	63.20	46.15	38.24	45.86	57.13	37.15	65.33	54.07	**69.53**	44.85	52.15
RSDA	63.69	**47.47**	39.12	49.82	**58.15**	38.25	**66.74**	**54.45**	68.69	**51.48**	**53.79**
AESC	S→R	S→L	S→D	R→S	R→L	R→D	D→R	D→S	L→R	L→S	**Avg.**
BERT-UDA*	47.09	34.77	32.10	33.12	33.68	34.93	42.68	28.03	45.46	27.89	35.98
CDRG†	52.93	33.33	36.14	43.07	44.70	30.82	53.18	40.30	57.77	41.51	43.38
GAS*	54.61	35.12	35.81	30.99	43.50	39.29	53.40	33.34	49.06	29.64	40.48
DA^{2}LM†	**58.64**	**36.97**	**40.28**	40.44	42.91	**41.28**	58.98	35.75	60.39	36.84	45.24
BGCA*	56.39	36.40	36.57	43.20	45.52	34.16	59.12	47.94	61.69	39.76	46.07
RSDA	56.36	36.59	37.19	**44.84**	**46.85**	36.22	**59.79**	**48.66**	**62.78**	**45.27**	**47.46**

Table 3: Results on cross-domain AE and AESC tasks where † and * indicate that the results are sourced from Yu et al. (2023) and Deng et al. (2023). The results are the average F1s over 5 runs.

Parameter Settings. For extraction and generation models, we choose T5-base checkpoint from Hugging Face[3]. The architecture of the T5 model is based on the transformer model, comprising encoder and decoder components. We employ a T5-base paraphraser (Vladimir Vorobev, 2023) for paraphrase-based diversity enhancement. Moreover, we utilize the Adam optimizer with a learning rate of 3e-4 and a batch size of 16 for all tasks. All experiments are conducted on a single NVIDIA 3090 GPU. For further details, please refer to Appendix A.1.

4.2 Comparison with Other Approaches

4.2.1 Approach Introduction

For AE and AESC tasks, we follow previous work (Yu et al., 2021, 2023) and choose baselines including BERT-UDA (Gong et al., 2020), CDRG (Yu et al., 2021), GAS (Zhang et al., 2021b), DA2LM (Yu et al., 2023), BGCA (Deng et al., 2023). Both BERT-UDA and CDRG utilize the BERT base, while GAS, and BGCA are built upon the T5-base. The base models of baselines are of the same order of magnitude as the one we use.

As for ASTE, our selected baselines include RoBMRC (Liu et al., 2022), SpanASTE (Xu et al., 2021), GAS (Zhang et al., 2021b), and BGCA (Deng et al., 2023). For AOPE, we adapted RoBMRC and SpanASTE by excluding sentiment polarities to accommodate the task, following Deng et al. (2023).

4.2.2 Result Comparison

The overall results of AE and AESC tasks in the cross-domain setting are presented in Table 3. It can be observed that our proposed framework outperforms the state-of-the-art method BGCA in the majority of domain pairs across ten different cross-domain settings. Overall, our approach achieves a 1.64% absolute improvement in averaged Micro-F1 compared to BGCA in the AE task, and a 1.39% improvement in the AESC task.

For the AOPE and ASTE tasks, we conduct experiments on six different domain pairs as shown in Table 4. Our framework shows an averaged F1 improvement of 1.45% in the AOPE task and 2.04% in the ASTE task. Notably, it also achieves 1.02% $\sim$ 3.58% improvement in F1 score where R serves as the source domain and L as the target domain.

Additionally, through the experiments presented in Tables 3 and 4, we note the following observations:

(1) Our proposed framework consistently outperforms BGCA across all four tasks on average. We

[3]https://huggingface.co/google-t5/t5-base

AOPE	R14 → L14	R15 → L14	R16 → L14	L14 → R14	L14 → R15	L14 → R16	Avg.
SpanASTE*	51.90	48.15	47.30	61.97	55.58	63.26	54.69
RoBMRC*	52.36	46.44	43.61	54.70	48.68	55.97	50.29
GAS*	57.58	53.23	52.17	64.60	60.26	66.69	59.09
BGCA*	60.82	55.22	54.48	68.04	65.31	70.34	62.37
RSDA	**61.48**	**57.62**	**55.74**	**69.61**	**67.20**	**71.27**	**63.82**

ASTE	R14 → L14	R15 → L14	R16 → L14	L14 → R14	L14 → R15	L14 → R16	Avg.
SpanASTE*	45.83	42.50	40.57	57.24	49.02	55.77	48.49
RoBMRC*	43.90	40.19	37.81	57.13	45.62	52.05	46.12
GAS*	49.57	43.78	45.24	64.40	56.26	63.14	53.73
BGCA*	53.64	45.69	47.28	65.27	58.95	64.00	55.80
RSDA	**54.66**	**48.39**	**50.96**	**66.15**	**60.52**	**66.36**	**57.84**

Table 4: Results on cross-domain AOPE and ASTE tasks. The results are the average F1s over 5 runs and * indicates that the results are sourced from Deng et al. (2023).

	AE	AESC	AOPE	ASTE	Avg.
RSDA	**53.79**	**47.50**	**63.82**	**57.84**	**55.74**
w/o NLI filter	51.69	44.75	59.13	51.47	51.76
w/o Composition-based	52.26	45.28	61.49	54.94	53.49
w/o Paraphrase-based	53.21	46.86	63.45	57.22	55.19

Table 5: Ablation Study.

attribute this improvement to the incorporation of quality filtering and diverse enhancement strategies throughout the entire process, ensuring the quality of generated samples. Moreover, our framework is fully compatible with BGCA and serves as its further optimization.

(2) We observe that methods based on encoder-decoder structures such as T5 perform better than those based on BERT. We speculate that generative models, with their encoder-decoder architecture, excel in handling abstract tasks by better capturing contextual information. They comprehensively understand the entire text through self-attention mechanisms and recursive mechanisms.

(3) Our framework performs less effectively than DA^{2}LM in certain domain pairs, particularly in cross-domain experiments where S serves as the source domain. We identify a challenge in experiments where the data volume in the source domain is significantly lower than that in the target domain. In such cases, the model fails to receive sufficient training in both extraction and generation, limiting subsequent results in terms of extraction and generation capabilities.

4.3 Ablation Study

To analyze the effectiveness of our framework, we conduct ablation experiments using micro-F1 and diversity as metrics, and the specific results are shown in Table 5. Firstly, when we remove the NLI filter, we observe a significant drop of approximately 3.98% in F1 scores across all four tasks. This indicates the effectiveness of NLI-based quality control, as the NLI filter eliminates examples with semantic and format errors. The removal of composition-based diversity enhancement leads to an average decrease of approximately 2.25% in F1 scores, with particularly notable impacts observed in the ASTE and AOPE tasks. We speculate that the composition-based diversity enhancement has a more noticeable impact on tasks with richer label entailment information. Thirdly, removing paraphrase-based diversity enhancement leads to an average F1 score decrease of approximately 0.55% across all four tasks.

In addition, to assess the contributions of composition-based diversity enhancement, we conduct ablation experiments for it, the results are as shown in Figure 4. We use the proportion of generated samples with multi-aspect as a metric. Removing the composition-based diversity enhancement resulted in varying degrees of reduction in this metric across all four tasks, with ASTE and AOPE tasks experiencing nearly a 50% decrease, demonstrating that the composition-based diversity enhancement indeed enhances the information density of samples.

4.4 Further Analysis

4.4.1 Quality Assessment of Generated Data

To demonstrate the effectiveness of our framework in the cross-domain setting, we conduct quality assessments of generated samples for the AESC task across ten domain pairs.

We employ perplexity to measure the fluency of the generated samples and adopt GPT-2[4] for perplexity calculation following Yu et al. (2023).

[4]https://huggingface.co/evaluate-measurement

Diversity $\mathcal{D}$ ↑	S→R	S→L	S→D	R→S	R→L	R→D	D→R	D→S	L→R	L→S	**Avg.**
CDRG†	0.133	0.134	0.146	0.250	0.235	0.289	0.264	0.293	0.193	0.229	0.217
DA^{2}LM†	0.275	0.309	0.354	0.472	0.269	0.374	0.257	**0.503**	0.252	0.416	0.349
BGCA	0.247	**0.376**	0.378	0.366	0.288	0.487	0.375	0.386	0.289	0.504	0.370
RSDA	**0.282**	0.284	**0.452**	**0.397**	**0.337**	**0.599**	**0.386**	0.467	**0.315**	**0.595**	**0.411**

PPL ↓	S→R	S→L	S→D	R→S	R→L	R→D	D→R	D→S	L→R	L→S	**Avg.**
CDRG	613.2	675.4	323.6	567.2	839.5	927.1	400.7	746.3	313.6	461.7	587.3
DA^{2}LM	189.4	361.8	267.5	172.6	**244.2**	273.3	325.3	256.3	342.8	204.7	263.8
BGCA	79.8	**62.9**	**59.7**	186.4	419.3	160.9	284.5	153.2	167.2	217.4	179.1
RSDA	**73.1**	70.2	89.7	**118.1**	286.6	**110.3**	**112.6**	**156.8**	**88.0**	**134.5**	**123.9**

Table 6: Quality assessment of the generated data. PPL stands for perplexity and † indicates that the results are sourced from Yu et al. (2023).

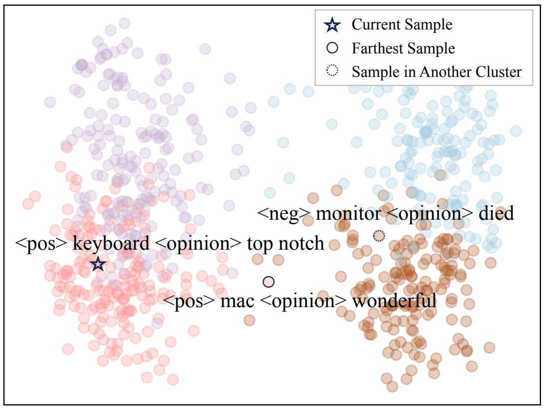

Figure 3: Clustering result visualization where the star symbol denotes the current sample point.

Given that the BGCA method did not generate additional data, resulting in a limited number of generated samples, for fairness, our experiments randomly select and test 500 samples generated by each method in perplexity testing. The results in Table 6 indicate that the perplexity of the samples generated by our framework is significantly lower than those of other methods. We speculate that the NLI filter effectively alleviates the issue of non-fluent generated samples caused by domain shift phenomena.

We also employ $\mathcal{D}_a$ to assess the distribution of generated samples. As shown in the last four rows of the table 6, our model exhibits higher diversity than other methods. It is noteworthy that, in the D→S task, although the DA^{2}LM method has a higher diversity value than our approach, our framework achieves an F1 score 12.91% higher. This suggests that our approach not only enhances the diversity of generated samples but also covers more aspect terms in the target domain.

4.4.2 Visual Case Study on Clustering

As described in Section 3.4.1, we conduct clustering on the filtered data and visualize the results as shown in Figure 3. The dataset for the target domain is denoted as L, focusing on the ASTE task. In the illustration, we have highlighted two samples from the same cluster with the farthest distance. It is evident from the figure that our algorithm not only ensures coherent labeling but also enhances diversity in the combinations, thereby achieving a balance between logical label pairing and increased variety.

5 Conclusion

In this paper, we propose a two-step data augmentation framework for cross-domain ABSA tasks. The first step controls sample quality and filters low-quality pseudo labels using NLI filter. The second step enhances the diversity of augmented data using label composition and paraphrase methods. We conduct 32 experiments in cross-domain settings to demonstrate the effectiveness of our framework, which outperforms 7 strong baselines. Our approach not only mitigates error propagation caused by incorrect pseudo-labels but also enhances the diversity and fluency of the generated labeled data in the target domain. It is simple yet effective to implement and extend to other domains and tasks without much effort. In the future, we will explore the generalization ability of our framework to other structural information extraction tasks.

Limitations

While our method has achieved promising results, there are still several limitations to be addressed. Although our approach effectively enhances information density through composition-based diversity enhancement, this advantage is more pronounced when label information is abundant. Further investigation is needed on how to improve performance in scenarios with limited label information. Additionally, our framework has only been tested on sentiment analysis datasets, and its applicability to other tasks remains to be explored.

Ethics Statement

We conduct extensive experiments on benchmark datasets from four domains: **Restaurant** (**R**), **Laptop** (**L**), **Device** (**D**) and **Service** (**S**), which are widely used for ABSA tasks. These datasets do not include personal information or contain sensitive content. In the process of generating data, we employ constrained decoding and quality control methods, which to some extent mitigate the presence of harmful content. However, human review is necessary when using these data in real-world applications.

References

Guoqing Chao, Jingyao Liu, Mingyu Wang, and Dianhui Chu. 2023. Data augmentation for sentiment classification with semantic preservation and diversity. *KNOWLEDGE-BASED SYSTEMS*, 280(2):111038.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Data augmentation for cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5346–5356.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Bidirectional generative framework for cross-domain aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 12272–12285.

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*, pages 3436–3442.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 567–573.

Sreyan Ghosh, Utkarsh Tyagi, Manan Suri, Sonal Kumar, Ramaneswaran S, and Dinesh Manocha. 2023a. ACLM: A selective-denoising based generative data augmentation approach for low-resource complex NER. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 104–125.

Sreyan Ghosh, Utkarsh Tyagi, Manan Suri, Sonal Kumar, Ramaneswaran S, and Dinesh Manocha. 2023b. ACLM: A selective-denoising based generative data augmentation approach for low-resource complex NER. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 104–125.

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*, pages 7035–7045.

Phillip Howard, Arden Ma, Vasudev Lal, Ana Paula Simões, Daniel Korat, Oren Pereg, Moshe Wasserblat, and Gadi Singer. 2022. Cross-domain aspect extraction using transformers augmented with knowledge graphs. In *Proceedings of the 31st International Conference on Information Knowledge Management*, pages 780–790.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence*, pages 755–760.

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.

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

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*, pages 272–278.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5755–5772.

Ricardo Marcondes Marcacini, Rafael Geraldeli Rossi, Ivone Penque Matsuno, and Solange Oliveira Rezende. 2018. Cross-domain aspect extraction for sentiment analysis: A transductive learning approach. *Decis. Support Syst.*, 114(1):70–80.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. PhraseRNN: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514.

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*, 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*, 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*, 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. *Journal of Machine Learning Research*, 21(2):140:1–140:67.

Evelina Rennes and Arne Jönsson. 2021. Synonym replacement based on a study of basic-level nouns in swedish texts of different complexity. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics*, pages 259–267.

P. J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20(2):53–65.

Raksha Sharma, Pushpak Bhattacharyya, Sandipan Dandapat, and Himanshu Sharad Bhatt. 2018. Identifying transferable information across domains for cross-domain sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 968–978.

Damien Sileo. 2023. tasksource: Structured dataset preprocessing annotations for frictionless extreme multi-task learning and evaluation. *arXiv preprint arXiv:2301.05948*.

Orith Toledo-Ronen, Matan Orbach, Yoav Katz, and Noam Slonim. 2022. Multi-domain targeted sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2751–2762.

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.

Maxim Kuznetsov Vladimir Vorobev. 2023. A paraphrasing model based on chatgpt paraphrases.

Ke Wang and Xiaojun Wan. 2023. Counterfactual representation augmentation for cross-domain sentiment analysis. *IEEE Transactions on Affective Computing*, 14(1):1979–1990.

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 (Volume 1: Long Papers)*, pages 4755–4766.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2514–2523.

Rohan Kumar Yadav, Lei Jiao, Ole-Christoffer Granmo, and Morten Goodwin. 2021. Human-level interpretable learning for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14203–14212.

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*, pages 2416–2429.

Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 6210–6219.

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*, pages 4767–4777.

Jianfei Yu, Qiankun Zhao, and Rui Xia. 2023. Cross-domain data augmentation with domain-adaptive language modeling for aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1456–1470.

Jingyi Zhang, Ying Zhang, Yufeng Chen, and Jinan Xu. 2023a. Structure and label constrained data augmentation for cross-domain few-shot NER. In *Findings of the Conference on Empirical Methods in Natural Language Processing*, pages 518–530.

Kai Zhang, Qi Liu, Zhenya Huang, Mingyue Cheng, Kun Zhang, Mengdi Zhang, Wei Wu, and Enhong Chen. 2022. Graph adaptive semantic transfer for cross-domain sentiment classification. In *Proceedings of the 45th International Conference on Research and Development in Information Retrieval*, pages 1566–1576.

Mao Zhang, Yongxin Zhu, Zhen Liu, Zhimin Bao, Yunfei Wu, Xing Sun, and Linli Xu. 2023b. Span-level aspect-based sentiment analysis via table filling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9273–9284.

Shaokang Zhang, Lei Jiang, Huailiang Peng, Qiong Dai, and Jianlong Tan. 2021a. Discriminative representation learning for cross-domain sentiment classification. In *Proceedings of the 25th conference on Pacific Asia Knowledge Discovery and Data Mining Part II*, pages 54–66.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. 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*, pages 504–510.

A Experimental Details

A.1 Training Settings

Task	Datasets	Train	Dev	Test
AE&AESC	L	3,045	304	800
	R	3,877	387	2,158
	D	2,557	255	1,279
	S	1,492	149	747
AOPE	L14	1035	116	343
	R14	1,462	163	500
	R15	678	76	325
	R16	971	108	328
ASTE	L14	906	219	328
	R14	1,266	310	492
	R15	605	148	322
	R16	857	210	326

Table 7: Basic statistics of the datasets.

For the extraction model, we adopted constrained decoding, confined to the vocabulary of the target domain. In selecting the training epochs for the four tasks, we drew inspiration from the approaches of Zhang et al. (2021b) and Deng et al. (2023), and extended our experimentation beyond the {15, 20, 25, 30} epochs. As for the paraphrasing model, to generate more diverse text, we set the temperature to 0.7. To ensure fairness, the amount of data generated by our method is kept at the same order of magnitude as the BGCA method. Furthermore, we performed post-processing(Deng et al., 2023) on the generated data to further ensure sample accuracy, including deduplication, format checking, and regeneration as needed.

A.2 Supplementary Experiments

Figure 4 shows the ablation experiments of composition-based sition on four tasks, with the metric being the proportion of samples with multiple aspects in the generated data.

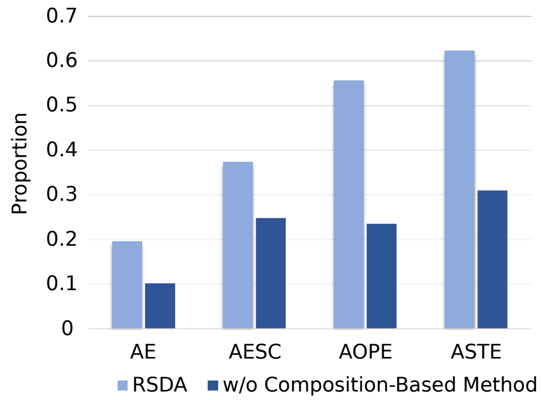

Figure 4: Ablation study about composition-based diversity enhancement.

B Case Studies

B.1 Applications of NLI Filter

Figure 5 illustrates two applications of the NLI filter. As for (a) in Figure 5, the diagram depicts how the generation of new samples in the target domain can be influenced by data from the source domain, resulting in domain shift phenomena that may significantly affect the fluency of generated samples. However, the NLI filter effectively identifies and filters out such examples promptly. In (b), the diagram shows an entailment relationship between the unlabeled text in the target domain and the newly generated text, indicating that such examples should be retained. Through NLI filter processing, we can filter out samples with semantic inconsistencies, such as those influenced by the source domain or model hallucinations.

B.2 Examples for Composition-based Diversity Enhancement

The merging process is illustrated in the figure 6. We concatenate the labels and texts of the farthest two examples in the same cluster, resulting in l_c and t_c. Then, we utilize the generation model M_g to obtain a t_n that is smoother than t_c. This constitutes a newly labeled sample (t_n, l_c). Finally, all samples undergo quality control again with the NLI filter. For tasks like AE with scarce label information, we adopt a random token selection approach to augment the label information in both training and inference processes.

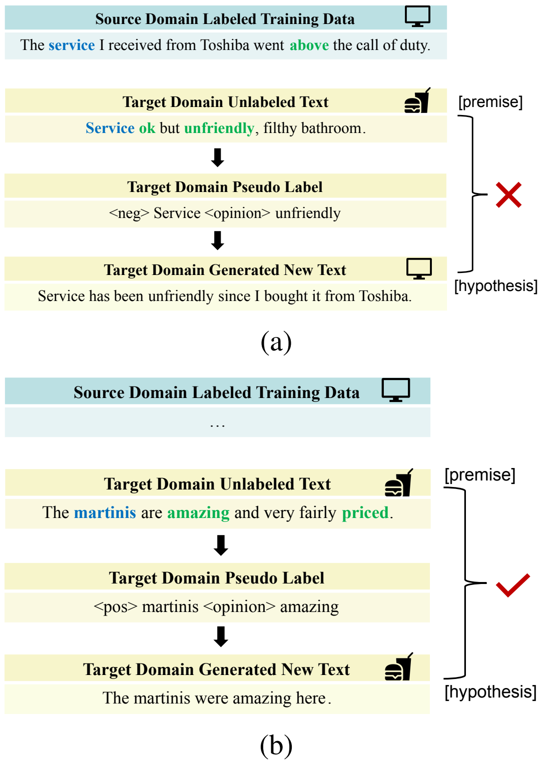

Figure 5: The examples of NLI filter where the cross symbol indicates a contradiction between the two, while the checkmark indicates entailment.

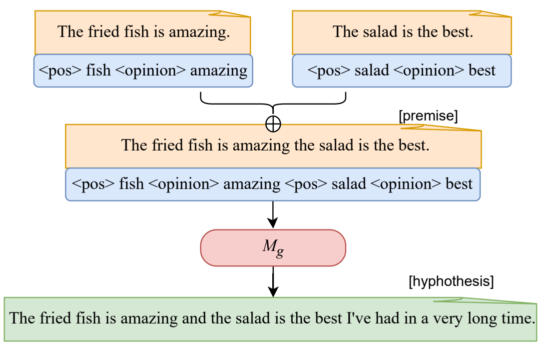

Figure 6: An example for composition-based diversity enhancement.