MATO: A Model-Agnostic Training Optimization for Aspect Sentiment Triplet Extraction

Shaopeng Tang[1], Lin Li[1,*], Xiaohui Tao[2], Leqi Zhong[1], Qing Xie[1]
[1]Wuhan University of Technology, China, [2]University of Southern Queensland, Australia
{karitown, cathylilin, zlq_lucky, felixxq}@whut.edu.cn, xiaohui.tao@unisq.edu.au

Abstract

As an important fine-grained sentiment analysis task, aspect sentiment triplet extraction (ASTE) aims to identify three elements, i.e., aspect, opinion and sentiment polarity as a triplet. Advanced ASTE researches have mostly explored triplet-wise ability to achieve superior improvement. However, existing models with strong in-house performances may struggle to generalize to the challenging cases with the diverse expression of inter-triplet and intra-triplet elements. To this end, we propose a **M**odel-**A**gnostic **T**raining **O**ptimization (**MATO**) to improve ASTE model inference consistent with expected results facing triplet element diversity. Specifically, to indicate the capacity to accommodate the diverse elements, we design inter-triplet and intra-triplet metamorphic relations (MRs), and calculate the violation rate (VR) on each element of one triplet through metamorphic testing (MT). Moreover, we propose an element-wise diversity-aware loss based on the VRs of aspect, opinion and sentiment, which can be jointly trained with existed ASTE models via uncertainty weighing. Conducted on four benchmark datasets and seven ASTE models, experimental results show that our MATO can enhance their diversity capacity, decreasing the average element-wise VRs by 3.28% to 15.36%. Meanwhile, our MATO is comparable to or better than those in terms of F1-score.

1 Introduction

Aspect sentiment triplet extraction (ASTE) aims to identify three elements, i.e., aspect term, opinion term and sentiment polarity as a triplet. As in the example "*The sound is nice and loud; I do n't have any problems with hearing anything.*" in Fig. 1 (c), its goal is to extract two triplets "(*sound*, *nice*, POS)" and "(*sound*, *loud*, POS)".

Many approaches to ASTE have been proposed successively. Peng et al. (2020) introduced the ASTE task at the first time and provided a two-stage framework in a pipeline approach to accomplish the extraction of aspect term and opinion term successively as well as the classification of sentiment. To overcome the error propagation problem, subsequent works adopted *table filling* representation to jointly model the ASTE task (Wu et al., 2020b; Chen et al., 2022b; Zhang et al., 2022; Sun et al., 2024). Some works used *sequence tagging* to enrich label representation to enhance representation learning (Xu et al., 2020, 2021; Liang et al., 2023; Li et al., 2023). Besides, there were some studies that tried to convert the ASTE task into a machine reading comprehension (*MRC*) task (Mao et al., 2021; Chen et al., 2021a; Zhai et al., 2022). Recently, generative model has gained significant results on many tasks, and equally some works have addressed the ASTE task with *generative manner* (Zhang et al., 2021a; Zhou and Qian, 2023).

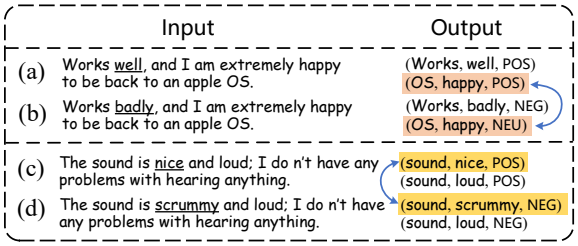

Figure 1: Two sets of examples for some ASTE models. (a) and (c) are from LAP14 while inter-triplet and intra-triplet diverse expressions appear in (b) and (d), respectively. The underlined indicates the distinctions.

The above methods explored the ability to improve the model's feature representation, feature learning and the ability to inference etc., and have obtained superior performance. However, they may struggle to generalize to the challenging cases with the diverse expression of inter-triplet and intra-triplet elements. As shown in Fig. 1 (a) and (b), the extraction of the triplet "(*OS*, *happy*, POS)" is affected just by changing the other triplet's opinion, which indicates that the resistance of ASTE models

* Corresponding author.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1648–1662
April 29 - May 4, 2025

to **inter-triplet** diverse expression is not powerful enough. Similarly, as shown in Figs. 1 (c) and (d), simply making a synonym transformation [1] for the opinion of the triplet "(*sound*, *nice*, POS)" leads to an inversion of the output sentiment, which suggests that the perceptual ability of ASTE models to **intra-triplet** diverse expression is also insufficient.

To address the aforementioned problem, we propose a **M**odel-**A**gnostic **T**raining **O**ptimization (**MATO**) to improve an ASTE model inference consistent with expected results facing diversity. Specifically, we firstly design **inter-triplet and intra-triplet metamorphic relations (MRs)** from the perspective of potential causes affecting the extraction result of one triplet. Based on these MRs, we introduce a metric for assessing the diversity and use it for training optimization. **The violation rate (VR) conducted with metamorphic testing (MT)**, can be calculated by comparing the target triplets between their originals and metamorphosis to check whether it follows the MR. For example, we can establish an inter-triplet MR based on (a) and (b) in Fig. 1. Return to MRs, we can analyze the element (i.e., aspect, opinion, and sentiment) of the diverse expression. And thus statistically obtain the VR on each element, indicating the capacity to diversity of triplet elements. And the paired Wilcoxon signed rank tests (Corder and Foreman, 2014) are performed on the MT results to ensure that the MRs are highly qualified.

Secondly, in order to make an ASTE model more focused on the triplet itself and shield from other triplets when generating triplets, we introduce three discerners to identify aspect, opinion, and sentiment, and propose **an element-wise diversity-aware loss based on VRs**. In particular, we sum three losses from the three discerners with the weights from the element-wise VRs. Finally, to better simultaneously learn ASTE triplet extraction and element-wise diversity awareness, an uncertainty-based weighting is applied to jointly train **diversity-aware loss and ASTE loss**, that is, MATO can work with most ASTE models.

Extensive experiments are conducted with seven ASTE SOTA models on four benchmark datasets. Our finding is that the capacity of the seven models facing inter-triplet and intra-triplet diverse expression remains significantly weak. The addition of our MATO to those ASTE SOTA models can enhance their capacity, decreasing the average element-wise VRs by 3.28% to 15.36%. Meanwhile, our MATO is comparable to or better than those in terms of F1-score.

[1] According to the Oxford English Dictionary (https://www.oed.com), scrummy means excellent, marvellous, enjoyable and delicious etc., representing nonstandard London speech.

2 Related Work

2.1 ASTE Models

Aspect Sentiment Triplet Extraction (ASTE) is a typical task in current research of aspect-based sentiment analysis (ABSA), proposed by Peng et al. (2020). ABSA is a traditional fine-grained sentiment analysis (Pontiki et al., 2014; Schouten and Frasincar, 2016; Xue and Li, 2018; Liu et al., 2020; Chen et al., 2022a; Liu et al., 2023; Li et al., 2024). The early work of ABSA involved three basic tasks, including aspect term extraction (Yin et al., 2016; Xu et al., 2018; Dai and Song, 2019; Chen and Qian, 2020; Li et al., 2020), opinion term extraction (Wan et al., 2020; Wu et al., 2020a) and aspect-level sentiment classification (Wang et al., 2016; Tang et al., 2016; Li et al., 2021; Brauwers and Frasincar, 2023).

Recent ASTE studies consider the integrity among the three elements and can be classified into five streams, that is, pipeline (Peng et al., 2020), table filling (Wu et al., 2020b; Chen et al., 2022b; Zhang et al., 2022; Sun et al., 2024), sequence tagging (Xu et al., 2020; Liang et al., 2023; Li et al., 2023), MRC-based (Mao et al., 2021; Chen et al., 2021a; Zhai et al., 2022) and generative manner (Yan et al., 2021; Zhang et al., 2021a,b; Zhou and Qian, 2023). These methods explored the ability to improve the model's feature representation, feature learning and the ability to inference etc., and have obtained superior performance. But few focuses enough on the element-wise diversity capacity.

2.2 Metamorphic Testing in NLP

In software engineering, metamorphic testing (MT) is the process of testing a program by examining the metamorphic relation (MR) between the results of multiple executions of the program to find and correct defects and errors in the software (Chen et al., 2018). MR is task-specific, and many works designed specific MRs for different tasks. Jiang et al. (2021) identified six types of MRs for the machine translation task, covering a wide range of properties that most NLI tasks are expected to have. The experimental results could explain the capabilities

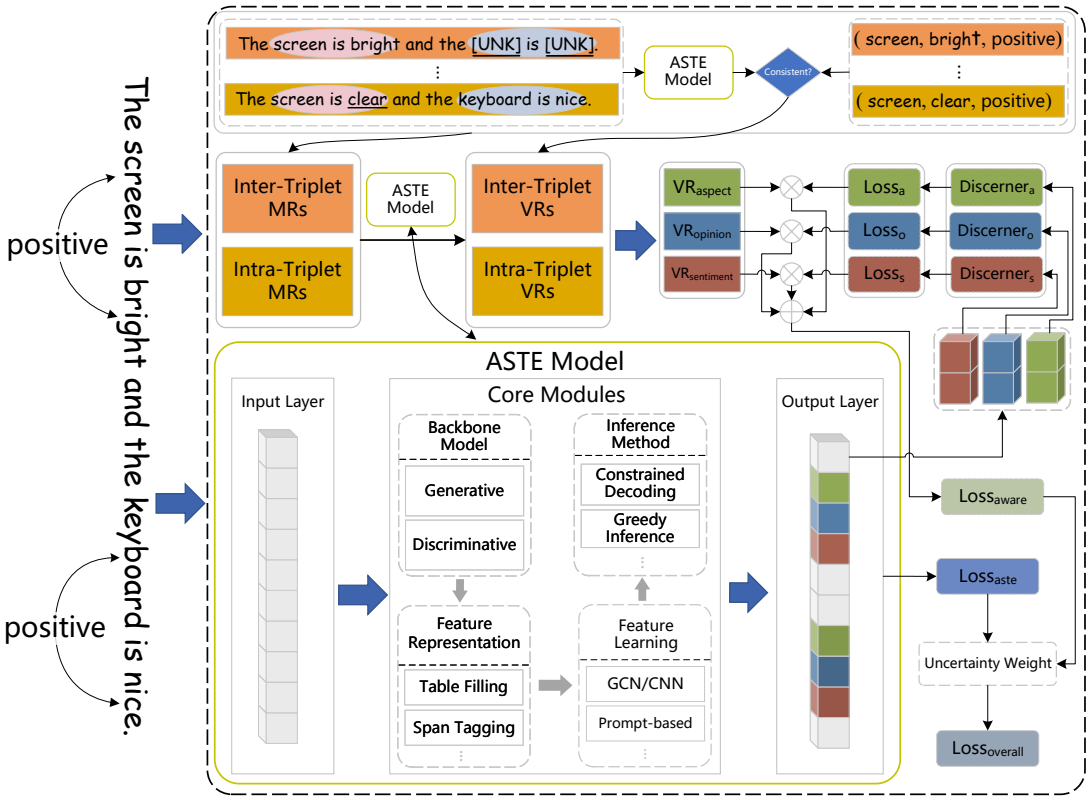

Figure 2: The overview of our MATO, a model-agnostic working with ASTE models

of the NLI model in different dimensions. Manino et al. (2022) proposed three MRs, which addressed the properties of systematicity, compositionality and transitivity. Manino et al. (2022) tested the internal consistency of state-of-the-art NLP models, and they did not always behave according to their expected linguistic properties. Hyun et al. (2024) proposed a framework with MT for analyzing large language model to address the limited coverage of quality attributes. Pu et al. (2023) adopted MT to evaluate the robustness of hand pose estimation models and provided suggestions on the choice of HPE models for different applications. Recent studies have found that the **property-based** validation method (such as violation rate based on MT) is more flexible than the traditional **reference-based** validation method (precision, recall and F1-score etc.) in revealing the actual language understanding capability of the NLP models (Chen et al., 2021b; Aleti, 2023; Wang et al., 2024).

To our literature review, our work is the first to consider ASTE models enhanced with MT. The major challenge is to design suitable MRs to reflect inter-triplet and intra-triplet diverse expressions and the violation rate based on MT can be used for training optimization.

3 Proposed Method

In this section, we present our MATO in details as shown in Fig. 2. ASTE model can generate the hidden representations for aspect, opinion and sentiment through feature representation, feature learning and inference stages. ASTE model computes ASTE loss by comparing the labels. We use MT to indicate the capacity to diversity of ASTE model and introduce diversity-aware loss to enhance the perception ability. Finally, an uncertainty weighing is applied to jointly train diversity-aware loss and ASTE loss. This process does not depend on the specific ASTE model (i.e., model-agnostic).

3.1 Task Description

Given a sentence $X = \{w_1, w_2, \ldots, w_n\}$ with n words, the goal of the ASTE model is to output all triplets $\mathcal{T} = \{(a, o, s)_i\}_{i=1}^{m}$ in the sentence, where a and o denote aspect term and opinion term respectively, and they both come from the sentence X. The sentiment polarity s belongs to the label set $S = \{POS, NEU, NEG\}$, and m is the number of triplets in the sentence.

3.2 Metamorphic Relations Design

The current mainstream ASTE models suffer from the problem of weak capacity when facing element-wise diverse expression, while MT based on task-specific MRs are able to cover such linguistic property. Therefore, we regard to design appropriate MRs to alleviate this diversity problem. As shown in the example in Table 1, we design inter-triplet and intra-triplet MRs from the viewpoint of possibly causing an unexpected change in the output for one triplet (denoted as target triplet).

3.2.1 Inter-triplet MRs

The expression of triplets aside from the target triplet is individual to individual, i.e. there is inter-triplet diverse expression. The output result of target triplet may be affected by external diversity, and an ASTE model with diversity capacity should avoid such cases. We principally consider the influence of other triplets on target triplet and design inter-triplet MRs, and then determine if there is a violation against one MR by comparing the consistency of the real output with the expected output.

MR1-1: According to the relative independence among triplets, the synonym transformation [2] to the opinion of other triplets in the sentence should not affect the output of the target triplet, i.e., its expected output is consistent to its original output.

Example for MR1-1:
Source Input: The screen is bright and the keyboard is nice.
Follow-up Input: The screen is bright and the keyboard is good.

MR1-2: When the sentiment of other triplets in the sentence is inverted (i.e., the opinion undergoes an antonym transformation), the extraction of the target triplet should not be affected, i.e., its expected output is consistent to its original output.

Example for MR1-2:
Source Input: The screen is bright and the keyboard is nice.
Follow-up Input: The screen is bright and the keyboard is bad.

MR1-3: In order to bring in more factors that may affect the target triplet extraction results, we add to the sentences some phrases [3] consisting of the triplet with the opposite sentiment of the target triplet. This should not affect the extraction of the target triplet, i.e., its expected output is consistent to its original output.

Example for MR1-3:
Source Input: The screen is bright and the keyboard is nice.
Follow-up Input: The screen is bright and the keyboard is nice. windows 7 is slow.

MR1-4: For a further analysis of the impact of other triplets on target triplet extraction in a sentence, we substitute aspect term and opinion term involved in other triplets with "[UNK]" to mask the semantic information brought by the other triplets. The expected output of the target triplet should be consistent to its original output without this part of the semantic information.

Example for MR1-4:
Source Input: The screen is bright and the keyboard is nice.
Follow-up Input: The screen is bright and the [UNK] is [UNK].

3.2.2 Intra-triplet MRs

ASTE models with diversity capacity should not only be able to extract target triplet easing the influence of other triplets, but more importantly focus on target triplet own diverse information. For such considerations, we check whether the ASTE model is able to respond correctly to its own changes by introducing diversity to the target triplet.

MR2-1: The sentiment is dependent, so we first consider making some changes to the holder of the sentiment, i.e., making a synonym/hypernym transformation [4] to aspect. The aspect in the target triplet output should change accordingly, the rest should be consistent.

Example for MR2-1:
Source Input: The screen is bright and the keyboard is nice.
Follow-up Input: The monitor is bright and the keyboard is nice.

MR2-2: In natural language, there are a variety of expressions that convey the approximate meaning. The model needs to be able to maintain a diversity-aware performance facing different opinon expressions, and understand the semantic information in the text. We apply a synonym transformation to the opinion of the target triplet. The opinion in the target triplet output should change accordingly, the rest should be consistent.

[2]The synonyms and antonyms in this work are derived from NLTK (https://www.nltk.org) and an online dictionary (https://github.com/meetDeveloper/freeDictionaryAPI).

[3]The generation of phrases from triplets in the datasets are automatically implemented with Parrot (https://github.com/PrithivirajDamodaran/Parrot_Paraphraser).

[4]The hypernyms in this work are also derived from NLTK (https://www.nltk.org).

Inter/Intra	MR type	Follow-up Input	Expected Output
Inter-triplet	MR1-1	The screen is bright and the keyboard is good.	(screen, bright, POS)
	MR1-2	The screen is bright and the keyboard is bad.	
	MR1-3	The screen is bright and the keyboard is nice. windows 7 is slow.	
	MR1-4	The screen is bright and the [UNK] is [UNK].	
Intra-triplet	MR2-1	The monitor is bright and the keyboard is nice.	(monitor, bright, POS)
	MR2-2	The screen is clear and the keyboard is nice.	(screen, clear, POS)
	MR2-3	The monitor is clear and the keyboard is nice.	(monitor, clear, POS)
	MR2-4	The screen is unclear and the keyboard is nice.	(screen, unclear, NEG)

Table 1: For the target triplet "(screen, bright,POS)" in the original outputs of the source input "The screen is bright and the keyboard is nice.", the follow-up inputs and the expected outputs corresponding to target triplet following the inter-triplet and intra-triplet MRs. The underlined parts indicate the distinctions.

Example for MR2-2:
Source Input: The screen is bright and the keyboard is nice.
Follow-up Input: The screen is clear and the keyboard is nice.

MR2-3: It is not sufficient to transform either aspect or opinion alone, so we consider synonym/hypernym transformations for both at the same time. The aspect and opinion in the target triplet output should change accordingly, the sentiment should be consistent.

Example for MR2-3:
Source Input: The screen is bright and the keyboard is nice.
Follow-up Input: The monitor is clear and the keyboard is nice.

MR2-4: Antonym transformation leads to changes in the meaning of the source input, and the model needs to understand the new semantics and accurately capture the sentiment polarity corresponding to the opinion. The opinion in the target triplet output should change accordingly, the aspect should be consistent. Moreover, The sentiment should be reversed.

Example for MR2-4:
Source Input: The screen is bright and the keyboard is nice.
Follow-up Input: The screen is unclear and the keyboard is nice.

We can generate a large amount of data to test the model based on the above MRs. According to the comparison of the output of the target triplet and the expected output, we can calculate the corresponding VRs: $\{VR_1, VR_2, \ldots, VR_8\}$, indicating the capacity of the ASTE model facing inter-triplet and intra-triplet diverse expression.

3.3 Diversity Awareness

Given an input sentence $X = \{w_1, w_2, \ldots, w_n\}$ with n tokens and the triplets $\mathcal{T} = \{(a, o, s)_i\}_{i=1}^{m}$. The last hidden layer representation sequence for aspect, opinion and sentiment in the ASTE model are denoted separately as:

$$H_{l_i} = \{h_1, h_2, \ldots, h_{l_i}\}, i \in \{1, 2, 3\} \quad (1)$$

where l_1, l_2 and l_3 denote the length of the last hidden layer representation sequence for aspect, opinion and sentiment, respectively.

We acquire the hidden representations of which represent aspect, opinion and sentiment:

$$H_e = \{h_{e_1}, h_{e_2}, \ldots, h_{e_m}\}, e \in \{a, o, s\} \quad (2)$$

To make the ASTE model more focus on the above representations, we introduce three discerners (i.e., three linear layers) and sigmoid activation function to identify aspect, opinion and sentiment:

$$\hat{y}_e = Sigmoid(discerner_e(H_e)), \quad e \in \{a, o, s\} \quad (3)$$

From these, three identification binary cross-entropy (BCE) losses can be obtained:

$$Loss_e = BCE(y_e, \hat{y}_e), e \in \{a, o, s\} \quad (4)$$

where the shape of y_e is the same with $\hat{y}_e$ and its values are fully 1.

We can summarize to derive VRs on aspect, opinion and sentiment:

$$VR_e = \frac{1}{|map_e|} \sum_{i \in map_e} VR_i, e \in \{a, o, s\} \quad (5)$$

where $|map_e|$ denotes the length of map_e. map_a is $\{3, 4, 5, 7\}$, map_o is $\{1, 2, 3, 4, 6, 7, 8\}$ and map_s is $\{2, 3, 4, 8\}$ based on the type of the diverse element in the above MRs.

Dataset	Train		Dev		Test	
	#S	#T	#S	#T	#S	#T
LAP14	906	1460	219	346	328	543
RES14	1266	2338	310	577	492	994
RES15	605	1013	148	249	322	485
RES16	857	1394	210	339	326	514

Table 2: Statistics of four ASTE datasets (Xu et al., 2020). #S and #T represent the number of sentences and triplets, respectively.

So far, we can obtain the diversity-aware loss:

$$Loss_{aware} = \sum_{e \in \{a,o,s\}} VR_e Loss_e \quad (6)$$

Diversity-aware loss makes the ASTE model percieve element-wise diverse expression.

3.4 Model-Agnostic Autoweighted Training

Let $f^W(X)$ be the output of the ASTE model with weights W on input X. The loss of the ASTE model can be uniformly formulated as $Loss_{aste}$.

Diversity-aware loss is model-agnostic and can be jointly trained with $Loss_{aste}$. Kendall et al. (2018) proposed to use uncertainty to weigh multiple losses to achieve multi-task learning for scene geometry and semantics. According to the proof of Kendall et al. (2018). we can approximate the overall loss for ASTE task:

$$Loss_{overall} \approx \frac{1}{\sigma_1^2} Loss_{aste} + \frac{1}{\sigma_2^2} Loss_{aware} + log\sigma_1 + log\sigma_2 \quad (7)$$

where the positive scalars σ_1 and σ_2 can be learnt from triplet extraction output and diversity-aware output distributions, respectively, where the parameter determines how "uniform" (flat) the discrete distribution is. This relates to its uncertainty, as measured in entropy (Kendall et al., 2018).

Later, Liebel and Körner (2018) have improved to avoid the loss value becoming negative during training. In practical, the unified training of ASTE triplet extraction and triplet focusing for multiple objectives is based on AutomaticWeightedLoss [5].

4 Experiments

This paper aims to address three research questions.

RQ1: Is the traditional reference-based validation method (i.e., F1-score) effective enough in revealing the diversity capacity of the ASTE model?

RQ2: Is the property-based validation method (i.e., VR) effective?

RQ3: Does our MATO significantly improve the capacity of the ASTE model when facing inter-triplet and intra-triplet diverse expression?

We conduct a thorough assessment of the quality of our designed MRs following Chen et al. (2021b), the results of which are shown in Appendix A.3.

4.1 Datasets and Baselines

We conducted extensive experiments on four ASTE benchmark datasets refined by (Xu et al., 2020). All datasets are sourced from SemEval Challenges (Pontiki et al., 2014, 2015, 2016), including reviews in the laptop and restaurant domains. The statistics of the four datasets are shown in Table 2.

For analyzing the capacity of various types of models in solving ASTE task as well as validating the effectiveness of our MATO, we select seven SOTA models as our baselines. (1) **Table filling**: EMCGCN (Chen et al., 2022b), BDTF (Zhang et al., 2022), MiniConGTS (Sun et al., 2024). (2) **Sequence tagging**: STAGE (Liang et al., 2023), SimSTAR (Li et al., 2023). (3) **MRC-based**: COM-MRC (Zhai et al., 2022). (4) **Generative manner**: SLGM (Zhou and Qian, 2023). Detailed descriptions for these baselines can be found in the Appendix A.1.

4.2 Experimental Settings

We run the above seven models corresponding to the parameter settings in their papers. For each synonym/antonym/hypernym transformation, we obtain at most 10 synonyms/antonyms/hypernyms from NLTK or the online dictionary. To minimize the influence of randomness, we ran run each model 5 times with different random initializations and then report the average results. All experiments are conducted on an NVIDIA TITAN XP GPU.

4.3 RQ1: Overall Performance with F1-score

Table 3 lists the experimental comparisons of the seven models before and after the introduction of MATO. MATO can improve the F1-score (green background) of the six models on the four datasets, and slightly reduce the F1-score (red background) of the COM-MRC, which suggests that our MATO can improve the traditional performance of most of ASTE models. Another interesting finding is that the addition of MATO is able to reduce the gap between precision and recall while keeping the performance of the F1-score stable.

[5] https://github.com/Mikoto10032/AutomaticWeightedLoss

Model	LAP14				RES14				RES15				RES16			
	P	R	F1	$\Delta F1$	P	R	F1	$\Delta F1$	P	R	F1	$\Delta F1$	P	R	F1	$\Delta F1$
EMCGCN (ACL, 2022b)	**61.70**	56.26	58.81	+0.89	71.21	72.39	71.78	+0.78	61.54	**62.47**	61.93	+0.15	65.62	**71.30**	68.33	+0.09
+ MATO	60.26	**59.15**	**59.70**		**71.70**	**73.45**	**72.56**		**61.89**	62.27	**62.08**		**65.88**	71.15	**68.42**	
COM-MRC (EMNLP, 2022)	**62.35**	58.16	**60.17**	-0.27	**75.46**	68.91	**72.01**	-0.34	**68.35**	61.24	**64.53**	-0.41	**71.55**	71.59	**71.57**	-0.59
+ MATO	59.58	**60.23**	59.90		73.65	**69.80**	71.67		65.99	**62.38**	64.12		68.90	**73.21**	70.98	
BDTF (EMNLP, 2022)	**68.94**	55.97	61.74	+0.96	**75.53**	73.24	74.35	+0.19	68.76	63.71	66.12	+1.33	71.44	73.13	72.27	+0.65
+ MATO	66.46	**59.33**	**62.70**		75.15	**73.94**	**74.54**		**69.67**	**65.36**	**67.45**		**71.75**	**74.12**	**72.92**	
SimSTAR (SIGIR, 2023)	**66.46**	58.23	62.07	+0.68	**76.23**	71.63	73.86	+0.30	**71.71**	59.59	65.09	+0.18	**72.02**	74.12	73.06	+0.64
+ MATO	65.64	**60.08**	**62.75**		75.20	**73.14**	**74.16**		70.14	**61.03**	**65.27**		71.55	**75.98**	**73.70**	
STAGE (AAAI, 2023)	**71.98**	53.86	61.58	+1.31	78.58	69.58	73.76	+0.93	**73.63**	57.90	64.79	+0.22	**76.67**	70.12	73.24	+0.65
+ MATO	68.99	**57.79**	**62.89**		**78.95**	**70.91**	**74.69**		71.35	**59.73**	**65.01**		75.70	**72.16**	**73.89**	
SLGM (ACL, 2023)	**70.54**	60.74	65.27	+0.15	78.84	72.70	75.64	+0.70	69.75	**66.85**	68.27	+0.70	75.86	75.76	75.80	+0.94
+ MATO	70.11	**61.33**	**65.42**		**79.01**	**73.84**	**76.34**		**71.21**	66.80	**68.97**		**76.45**	**77.04**	**76.74**	
MiniConGTS (EMNLP, 2024)	**66.82**	60.68	63.61	+1.10	**76.10**	75.08	75.59	+0.64	**66.50**	63.86	65.15	+0.70	**75.52**	74.14	74.83	+0.38
+ MATO	66.00	**63.46**	**64.71**		74.38	**78.20**	**76.23**		65.37	**66.34**	**65.85**		74.83	**75.61**	**75.21**	

Table 3: Experimental results(%). The score marked as **bold** means the better performance in pairwise comparison.

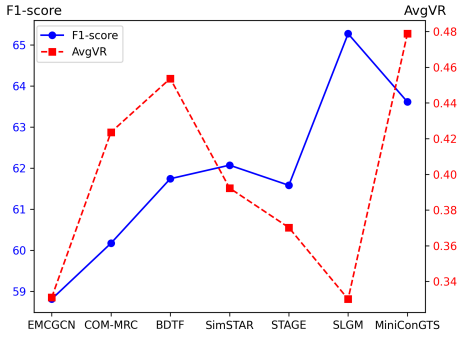

Figure 3: The relation of average VR and F1-score of each model without MATO on LAP14

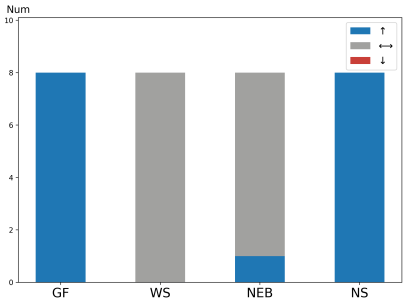

Figure 4: Statistics of the paired Wilcoxon signed-rank test results on SLGM (Here red color is few and more results of other six models are at Appendix A.3).

In order to analyze whether it is appropriate to use the traditional reference-based metric (e.g., F1-score) directly to evaluate the diversity capacity of the model, we discuss the relation between VR and F1-score. Fig. 3 shows the average VR and F1-score of each model on LAP14, and it is not difficult to find out that the VR does not always decrease with the growth of F1-score, which suggests that it is inappropriate to use F1-score directly to assess the capacity of the ASTE model.

4.4 RQ2: Diversity Capacity Comparison between F1-score and VR

To ensure the effectiveness of our designed MRs, evaluation for MRs is essential. Chen et al. (2021b) introduced multiple mutants to MRC models to evaluate MRs. They compared whether there was a significant difference (i.e., **better**, **similar**, or **worse**) between two VRs with the paired Wilcoxon signed-rank test. One VR was calculated from the output of the samples which the ASTE model without any mutants can pass under F1-score before and after the introduction of mutants. The other VR was calculated from the output of the metamorphosis data generated following the MRs before and after the introduction of mutants. Inspired by this, and considering that our baselines are transformer-based models, we adopt the four mutants: Gaussian Fuzzing (GF), Weight Shuffling (WS), Neuron Effect Blocking (NEB) and Neuron Switch (NS), and details are in Appendix A.2.

We execute the evaluation experiments following the methodology of Chen et al. (2021b) in both 2-tailed and 1-tailed manners to calculate p values, at the σ level of 0.05. Thus, we have:

- In the 2-tailed test, VR calculated by a particular MR is **similar** (denoted as $\longleftrightarrow$) to F1-score in revealing the mutants when $p \geq \sigma$.
- In the 1-tailed test, VR calculated by a particular MR is **better** (denoted as $\uparrow$) than F1-score in revealing the mutants when $p < \sigma$.
- Otherwise, VR calculated by a particular MR is **worse** (denoted as $\downarrow$) than F1-score in revealing the mutants.

Dataset	Model	MR1-1	MR1-2	MR1-3	MR1-4	MR2-1	MR2-2	MR2-3	MR2-4	Aspect	Opinion	Sentiment
LAP14	SLGM	0.2503	0.1549	0.2866	0.4110	0.3273	0.3371	0.2548	**0.6186**	0.3199	0.3305	0.3678
	+MATO	**0.2029**	**0.1240**	**0.2242**	**0.3821**	**0.2832**	**0.2657**	**0.1843**	0.6244	**0.2685**	**0.2868**	**0.3387**
RES14	SLGM	0.1807	0.1318	0.2838	0.2798	0.2423	0.2216	0.1855	**0.5392**	0.2479	0.2603	0.3087
	+MATO	**0.1384**	**0.0973**	**0.1922**	**0.2413**	**0.1804**	**0.1618**	**0.1241**	0.5581	**0.1845**	**0.2162**	**0.2722**
RES15	SLGM	0.2274	0.1125	0.3148	0.3943	0.3435	0.2356	0.2247	0.4214	0.3193	0.2758	0.3108
	+MATO	**0.1906**	**0.0650**	**0.2740**	**0.3549**	**0.3056**	**0.1933**	**0.1661**	**0.4201**	**0.2752**	**0.2377**	**0.2785**
RES16	SLGM	0.1632	0.1004	0.2526	0.3075	0.2604	0.1903	0.1739	**0.5988**	0.2486	0.2552	0.3148
	+MATO	**0.1236**	**0.0838**	**0.1787**	**0.2332**	**0.1896**	**0.1503**	**0.1326**	0.6029	**0.1835**	**0.2150**	**0.2747**

Table 4: The MT results (i.e., VRs) on SLGM (other six models see Appendix A.4). The value of VR ranges from 0 to 1, with smaller values indicating a greater capacity facing the inter-triplet and intra-triplet diverse expression. The score marked as **bold** means the better performance in pairwise comparison.

No.	Input	SLGM Output	Violate ?	SLGM+MATO Output	Violate ?
(1)	Works well, and I am extremely happy to be back to an apple OS.	(OS, happy, POS)	Yes	(OS, happy, POS)	No
(2)	Works badly, and I am extremely happy to be back to an apple OS.	(OS, happy, NEU)		(OS, happy, POS)	
(3)	The sound is nice and loud; I do n't have any problems with hearing anything.	(sound, nice, POS)	Yes	(sound, nice, POS)	No
(4)	The sound is scrummy and loud; I do n't have any problems with hearing anything.	(sound, scrummy, NEG)		(sound, scrummy, POS)	

Table 5: Case study for the target triplets "(OS, happy, POS)" and "(sound, nice, POS)" when SLGM faces inter-triplet and intra-triplet diverse expression, respectively. The underlined parts indicate the distinctions.

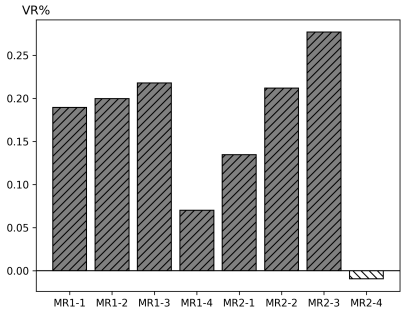

Figure 5: MATO reduces the VR magnitudes of all MRs on LAP for SLGM

The paired Wilcoxon signed-rank test results are in the Appendix A.3. Taking the test on SLGM as an example, as shown in Fig. 4, the VRs calculated by MRs are better or similar in revealing the mutants, indicating that the property-based validation method VR based the inter-triplet and intra-triplet MRs are more effective than the traditional reference-based validation method F1-score.

4.5 RQ3: Diversity Capacity Gain from MATO

The MT results on SLGM are in Table 4 and the MT results on the others are in the appendix A.4.

To start with, we can find that SLGM's VRs of for four datasets on MRs except MR1-2 is higher than 0.2, which indicates that there is still much room for improvement in the capacity of SLGM. We can find that the performance of these models varies in facing inter-triplet and intra-triplet diverse expression. For example, the VRs of SLGM in MR1-2 are all below 0.2 while the VRs of SLGM in MR2-4 are all above 0.4 in Table 4. This suggests that there is capacity difference on SLGM when facing inter-triplet and intra-triplet diversity.

Next, we can enhance the capacity of different dimensions of the ASTE model significantly by introducing diversity-aware loss. Take SLGM as an example on LAP14, as shown in Fig. 5, we can find that MATO significantly reduces the VR of MR1-1, MR1-2, MR1-3, MR2-1, MR2-2 and MR2-3 (all reduced by more than 10%). This suggests that MATO is able to keep the extraction process more capacity by having the ASTE model focus more on triplet representations. The introduction of MATO slightly increases the VR of MR2-4 on LAP14 (from 0.6186 to 0.6244, VR% below 1%), which does not affect the capacity of SLGM.

4.6 Case Study

We perform a case study on the response of SLGM when facing inter-triplet and intra-triplet diverse expression for the target triplet. From the results of (1) and (2) in Table 5, SLGM suffers from impacts outside the target triplet when solving the ASTE task, whereas our MATO enables the model to be shielded from the impacts when extracting the target triplet. From the results of (3) and (4) in

No.	Input	GPT-4o	Violate ?	Qwen2.5	Violate ?	SLGM+MATO	Violate ?
(1)	The sound is nice and loud; I do n't have any problems with hearing anything.	(sound, nice, POS)	No	(sound, nice, POS)	No	(sound, nice, POS)	No
(2)	The sound is scrummy and loud; I do n't have any problems with hearing anything.	(sound, scrummy, POS)		(sound, scrummy, POS)		(sound, scrummy, POS)	
(3)	It 's fast, light, and simple to use.	(use, fast, POS)	Yes	(use, fast, POS)	Yes	(use, fast, POS)	No
(4)	It 's fast, light, and [UNK] to use.	(It, fast, POS)		(It, fast, POS)		(use, fast, POS)	

Table 6: Case study for the target triplets "(sound, nice, POS)" and "(use, light, POS)" when GPT-4o and Qwen2.5 face inter-triplet and intra-triplet diverse expression, respectively. The underlined parts indicate the distinctions.

Table 5, SLGM is not able to perceive the change of the target triplet and reacts incorrectly, while our MATO is able to enhance the model's ability to perceive the target triplet.

As shown in Table 6, we perform a case study on two representative large language models (LLMs). From the results of (1) and (2) in Table 6, the two LLMs can easily determine the sentiment polarity of "scrummy" based on its rich knowledge storage. However, from the results of (3) and (4) in Table 6, after masking some semantic information with "[UNK]", the two LLMs both fail to identify the aspect "use", and mistakes "It " as aspect, while SLGM+MATO can still identify aspect correctly.

5 Conclusions

In this paper, we design metamorphic relations to evaluate the capacity of seven SOTA ASTE models by metamorphic testing and find that most of the models have weak capacity in dealing with diverse expression in sentiment analysis. Meanwhile, we propose a model-agnostic diversity-aware method to improve ASTE model inference consistent with expected results facing triplet diversity. Experiments show that our MATO can significantly enhance the diversity capacity of ASTE models.

Limitations

Although the MRs we design are able to expose problems regarding the diversity capacity of the ASTE model and MATO can improve the capacity, there are some limitations. On the one hand, the MRs we design only introduce diversity into triplets, while there are much diversity in the real world, such as changes in sentence structure and typos etc., and we will follow up with deeper work in this problem. On the other hand, from the results of MT, we can find that MATO does not bring considerable enhancement on MR2-4, and our subsequent work will explore how to make the model able to capture such intra-triplet sentiment inversion changes.

Acknowledgments

This work has been partially supported by the National Natural Science Foundation of China (NSFC) with Grant No. 62276196.

References

Aldeida Aleti. 2023. Software testing of generative AI systems: Challenges and opportunities. In *IEEE/ACM International Conference on Software Engineering: Future of Software Engineering, ICSE-FoSE 2023, Melbourne, Australia, May 14-20, 2023*, pages 4–14. IEEE.

Gianni Brauwers and Flavius Frasincar. 2023. A survey on aspect-based sentiment classification. *ACM Comput. Surv.*, 55(4):65:1–65:37.

Chenhua Chen, Zhiyang Teng, Zhongqing Wang, and Yue Zhang. 2022a. Discrete opinion tree induction for aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2051–2064. Association for Computational Linguistics.

Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022b. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2974–2985. Association for Computational Linguistics.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021a. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12666–12674. AAAI Press.

Songqiang Chen, Shuo Jin, and Xiaoyuan Xie. 2021b. Validation on machine reading comprehension software without annotated labels: a property-based method. In *ESEC/FSE '21: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, August 23-28, 2021*, pages 590–602. ACM.

Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. 2018. Metamorphic testing: A review of challenges and opportunities. *ACM Comput. Surv.*, 51(1):4:1–4:27.

Zhuang Chen and Tieyun Qian. 2020. Enhancing aspect term extraction with soft prototypes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2107–2117. Association for Computational Linguistics.

Gregory W Corder and Dale I Foreman. 2014. *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons.

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

Sangwon Hyun, Mingyu Guo, and M Ali Babar. 2024. Metal: Metamorphic testing framework for analyzing large-language model qualities. In *2024 IEEE Conference on Software Testing, Verification and Validation (ICST)*, pages 117–128. IEEE.

Mingyue Jiang, Houzhen Bao, Kaiyi Tu, Xiao-Yi Zhang, and Zuohua Ding. 2021. Evaluating natural language inference models: A metamorphic testing approach. In *32nd IEEE International Symposium on Software Reliability Engineering, ISSRE 2021, Wuhan, China, October 25-28, 2021*, pages 220–230. IEEE.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7482–7491. Computer Vision Foundation / IEEE Computer Society.

Dongxu Li, Zhihao Yang, Yuquan Lan, Yunqi Zhang, Hui Zhao, and Gang Zhao. 2023. Simple approach for aspect sentiment triplet extraction using span-based segment tagging and dual extractors. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2374–2378. ACM.

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, ACL 2020, Online, July 5-10, 2020*, pages 7056–7066. Association for Computational Linguistics.

Lin Li, Shaopeng Tang, and Renwei Wu. 2024. Majority rules guided aspect-category based sentiment analysis via label prior knowledge. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10952–10957. ELRA and ICCL.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard H. Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6319–6329. Association for Computational Linguistics.

Shuo Liang, Wei Wei, Xian-Ling Mao, Yuanyuan Fu, Rui Fang, and Dangyang Chen. 2023. STAGE: span tagging and greedy inference scheme for aspect sentiment triplet extraction. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13174–13182. AAAI Press.

Lukas Liebel and Marco Körner. 2018. Auxiliary tasks in multi-task learning. *CoRR*, abs/1805.06334.

Dan Liu, Lin Li, Xiaohui Tao, Jian Cui, and Qing Xie. 2023. Descriptive prompt paraphrasing for target-oriented multimodal sentiment classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4174–4186. Association for Computational Linguistics.

Haoyue Liu, Ishani Chatterjee, MengChu Zhou, Xiaoyu Sean Lu, and Abdullah Abusorrah. 2020. Aspect-based sentiment analysis: A survey of deep learning methods. *IEEE Trans. Comput. Soc. Syst.*, 7(6):1358–1375.

Edoardo Manino, Julia Rozanova, Danilo S. Carvalho, André Freitas, and Lucas C. Cordeiro. 2022. Systematicity, compositionality and transitivity of deep NLP models: a metamorphic testing perspective. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2355–2366. Association for Computational Linguistics.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based

sentiment analysis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13543–13551. AAAI Press.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8600–8607. AAAI Press.

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

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

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

Muxin Pu, Chun Yong Chong, and Mei Kuan Lim. 2023. Robustness evaluation in hand pose estimation models using metamorphic testing. In *8th IEEE/ACM International Workshop on Metamorphic Testing, MET@ICSE 2023, Melbourne, Australia, May 14, 2023*, pages 31–38. IEEE.

Kim Schouten and Flavius Frasincar. 2016. Survey on aspect-level sentiment analysis. *IEEE Trans. Knowl. Data Eng.*, 28(3):813–830.

Qiao Sun, Liujia Yang, Minghao Ma, Nanyang Ye, and Qinying Gu. 2024. MiniConGTS: A near ultimate minimalist contrastive grid tagging scheme for aspect sentiment triplet extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2817–2834, Miami, Florida, USA. Association for Computational Linguistics.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 214–224. The Association for Computational Linguistics.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z. Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9122–9129. AAAI Press.

Jun Wang, Yanhui Li, Zhifei Chen, Lin Chen, Xiaofang Zhang, and Yuming Zhou. 2024. Knowledge graph driven inference testing for question answering software. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering, ICSE 2024, Lisbon, Portugal, April 14-20, 2024*, pages 119:1–119:13. ACM.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 606–615. The Association for Computational Linguistics.

Meixi Wu, Wenya Wang, and Sinno Jialin Pan. 2020a. Deep weighted maxsat for aspect-based opinion extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5618–5628. Association for Computational Linguistics.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020b. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. Association for Computational Linguistics.

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

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet

extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4755–4766. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2339–2349. Association for Computational Linguistics.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2514–2523. Association for Computational Linguistics.

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

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

Zepeng Zhai, Hao Chen, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. COM-MRC: A context-masked machine reading comprehension framework for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3230–3241. Association for Computational Linguistics.

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

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, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 504–510. Association for Computational Linguistics.

Yice Zhang, Yifan Yang, Yihui Li, Bin Liang, Shiwei Chen, Yixue Dang, Min Yang, and Ruifeng Xu. 2022. Boundary-driven table-filling for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6485–6498. Association for Computational Linguistics.

Shen Zhou and Tieyun Qian. 2023. On the strength of sequence labeling and generative models for aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12038–12050. Association for Computational Linguistics.

A Appendix

A.1 The baselines

The seven baselines are described as follows.

MiniConGTS: It improves and utilizes pre-trained representations by integrating a minimalist tagging scheme and a novel token-level contrastive learning strategy (Sun et al., 2024).

SLGM: It proposes a sequence labeling enhanced generative model to enhance the mutual informative clues between aspect and opinion terms (Zhou and Qian, 2023).

STAGE: It proposes span tagging and greedy inference to extract sentiment triplets in span-level, where each span might consist of multiple words and play different roles simultaneously (Liang et al., 2023).

SimSTAR: It proposes a span-based solution with segment tagging and dual extractors to address inadequate representation of the information through previous tagging schemes and insufficient usage of all available sentiment data (Li et al., 2023).

BDTF: It proposes boundary-driven table-filling, which represents each triplet as a relation region in the 2D table and transforms the ASTE task into detection and classification of relation regions (Zhang et al., 2022).

COM-MRC: It proposes a novel context-masked MRC framework to address the failure of the MRC-based methods when facing multiple aspect terms, due to the interference from other aspect terms (Zhai et al., 2022).

EMCGCN: It proposes an enhanced multi-channel graph convolutional network model to fully

utilize the relations between words (Chen et al., 2022b).

A.2 The mutants

The four mutants are described as follows.

Gaussian Fuzzing: Weights of the neurons act as the key to control the decision logic of the neural networks. This mutant fuzzes the values of the weights for all the target neurons to change the connection importance they represent through adding Gaussian noise.

Weight Shuffling: The output of a neuron is usually determined by the neurons in the previous layer through the connections with weights. This mutant shuffles the weights of randomly picked target neurons to disturb their connections with their previous layers.

Neuron Effect Blocking: Every neuron in a neural network contributes to its final decision to some extent. This operator removes the influence of randomly picked target neurons to the final decision by resetting their connection weights of the next layers to zeros to block the propagation of their effects.

Neuron Switch: Different neurons in one neural network layer usually play different roles on the connected neurons in the next layer. This operator switches the weights of two neurons in the same layer to exchange their effects for next layer.

A.3 The statistics of the paired Wilcoxon signed-rank test results for the other six models

On the whole, the data generated by the inter-triplet and intra-triplet MRs we design are better than the original data in revealing the mutants (i.e., Gaussian Fuzzing, Weight Shuffling, Neuron Effect Blocking and Neuron Switch) as shown in Fig. 6. GF, WS, NEB and NS denote Gaussian Fuzzing, Weight Shuffling, Neuron Effect Blocking and Neuron Switch respectively. Better conclusion shows that the data generated by the corresponding MR is better than the original data in detecting the corresponding mutant, which indicates that the designed MR is effective in analyzing the model diversity capacity.

There are two significant findings. To begin with, most of the 2-tailed p values in the "better" test are 0.0020, and most of the 1-tailed p values are 0.0010 (both extreme borderline values), indicating that the generated data are significantly better than the original data in revealing the mutants (Corder and Foreman, 2014).

Furthermore, we find that almost all of the "worse" test results occur in Weight Shuffling and Neuron Effect Blocking, which may be due to the fact that these seven transformer-based models are the deep neural network.

A.4 The MT results for the other six models

The MT results for the other six models are in Table 7. It can be observed that the introduction of our MATO is able to significantly reduce the VRs of the six models on all except MR2-4, as well as significantly reduce the VRs on aspect term, opinion term and sentiment polarity, which suggests that our MATO is able to enhance their capacity.

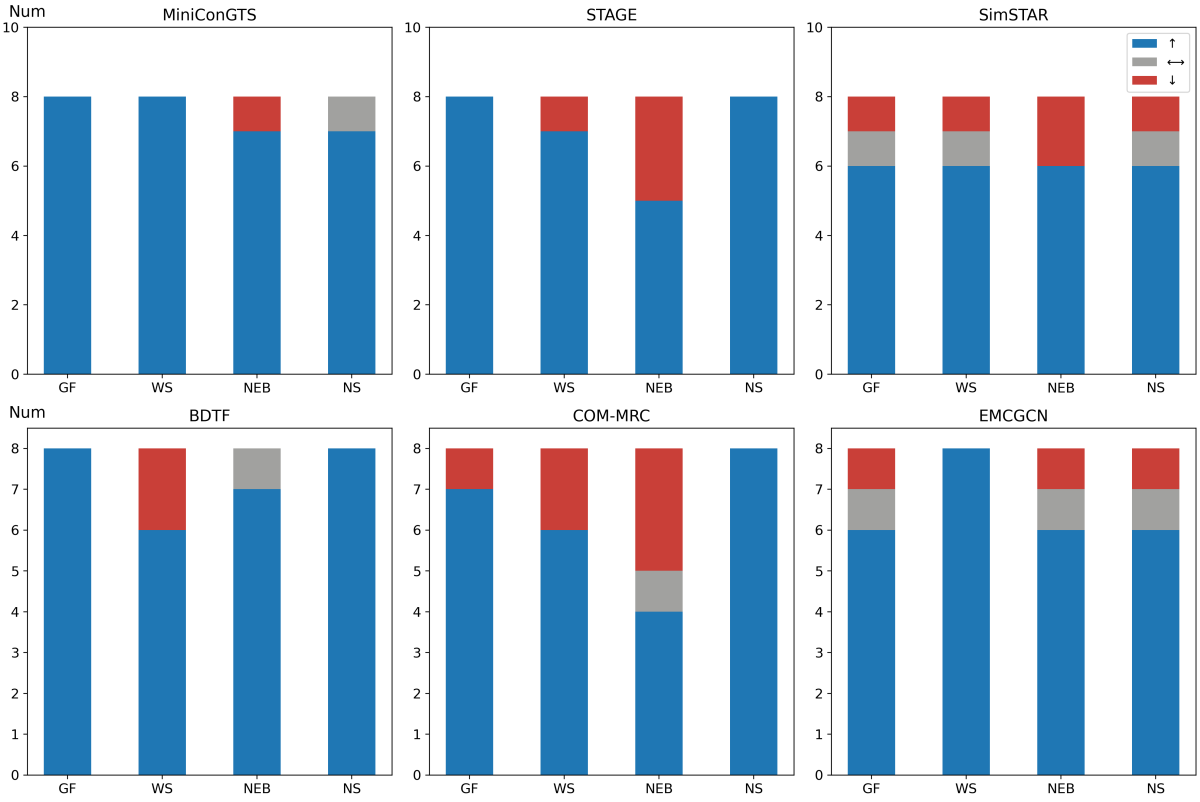

Figure 6: Statistics of the paired Wilcoxon signed-rank test results on other six models

No.	Dataset	Model	MR1-1	MR1-2	MR1-3	MR1-4	MR2-1	MR2-2	MR2-3	MR2-4	Aspect	Opinion	Sentiment
1	LAP14	MiniConGTS	0.3782	0.2519	0.5505	0.6525	0.5123	0.4854	0.3734	0.6263	0.5222	0.4740	0.5203
		+MATO	**0.3161**	**0.1865**	**0.5284**	**0.6173**	**0.5039**	**0.4528**	**0.3510**	**0.5667**	**0.5002**	**0.4313**	**0.4747**
	RES14	MiniConGTS	0.3764	**0.1963**	0.5944	0.6539	0.5889	0.4261	0.4012	0.5593	0.5596	0.4582	0.5010
		+MATO	**0.3494**	0.2066	**0.5561**	**0.6304**	**0.5650**	**0.4037**	**0.3792**	**0.5267**	**0.5327**	**0.4360**	**0.4800**
	RES15	MiniConGTS	0.3656	**0.1958**	0.6828	0.7345	0.6061	0.4223	0.3558	0.4299	0.5948	0.4552	0.5108
		+MATO	**0.3449**	0.1966	**0.6509**	**0.6982**	**0.5496**	**0.3781**	**0.3193**	**0.4148**	**0.5545**	**0.4290**	**0.4901**
	RES16	MiniConGTS	0.3634	0.2104	0.6046	0.6536	0.5542	0.4282	0.3884	0.5988	0.5502	0.4639	0.5169
		+MATO	**0.3372**	**0.1842**	**0.5792**	**0.6207**	**0.5182**	**0.4040**	**0.3626**	**0.5829**	**0.5202**	**0.4387**	**0.4918**
2	LAP14	STAGE	0.2772	0.1606	0.3476	0.5341	0.3943	0.3819	0.2652	**0.6006**	0.3853	0.3667	0.4107
		+MATO	**0.2446**	**0.1396**	**0.3397**	**0.4668**	**0.3264**	**0.3433**	**0.2356**	0.6224	**0.3421**	**0.3417**	**0.3921**
	RES14	STAGE	0.2087	0.1456	0.2899	0.3573	0.2849	0.2365	0.1944	**0.5465**	0.2816	0.2827	0.3348
		+MATO	**0.1754**	**0.1256**	**0.2569**	**0.3146**	**0.2542**	**0.2081**	**0.1726**	0.5501	**0.2496**	**0.2576**	**0.3118**
	RES15	STAGE	0.2445	0.1073	0.4204	0.5330	0.4548	0.3260	0.2822	0.4332	0.4226	0.3352	0.3735
		+MATO	**0.2286**	**0.0936**	**0.3766**	**0.4784**	**0.3966**	**0.3026**	**0.2541**	**0.4283**	**0.3764**	**0.3089**	**0.3442**
	RES16	STAGE	0.1744	0.1066	0.2946	0.3586	0.2701	0.2172	**0.1848**	**0.5865**	0.2770	0.2747	0.3366
		+MATO	**0.1644**	**0.0991**	**0.2766**	**0.3284**	**0.2542**	**0.2113**	0.1891	0.5906	**0.2621**	**0.2656**	**0.3237**
3	LAP14	SimSTAR	0.3090	0.1878	0.3650	0.5565	0.4043	0.4116	0.2756	0.6282	0.4004	0.3905	0.4344
		+MATO	**0.2575**	**0.1485**	**0.3156**	**0.4786**	**0.3380**	**0.3693**	**0.2572**	**0.6128**	**0.3474**	**0.3485**	**0.3889**
	RES14	SimSTAR	0.1847	0.1325	0.2618	0.3227	0.2527	0.2064	0.1678	0.5538	0.2513	0.2614	0.3177
		+MATO	**0.1745**	**0.1281**	**0.2467**	**0.3016**	**0.2367**	**0.1943**	**0.1603**	**0.5511**	**0.2363**	**0.2509**	**0.3069**
	RES15	SimSTAR	0.2460	0.1318	0.4068	0.5037	**0.3833**	0.3008	**0.2213**	**0.4136**	0.3788	0.3177	0.3640
		+MATO	**0.2415**	**0.1170**	**0.3579**	**0.4638**	0.3859	**0.2738**	0.2232	0.4283	**0.3577**	**0.3008**	**0.3418**
	RES16	SimSTAR	0.1626	0.1066	**0.2616**	0.3167	0.2521	0.1896	0.1688	**0.5988**	0.2498	0.2578	0.3209
		+MATO	**0.1570**	**0.0960**	0.2699	**0.2872**	**0.2250**	**0.1723**	**0.1420**	0.6070	**0.2310**	**0.2473**	**0.3150**
4	LAP14	BDTF	0.3690	0.1807	0.4913	0.6497	**0.4776**	0.4848	0.3502	0.6263	0.4922	0.4503	0.4870
		+MATO	**0.3549**	**0.1759**	**0.4800**	**0.6368**	0.4831	**0.4694**	**0.3474**	**0.6224**	**0.4868**	**0.4410**	**0.4788**
	RES14	BDTF	0.3934	0.2621	0.5737	0.6749	0.6066	0.4406	0.4134	**0.5593**	0.5672	0.4739	0.5175
		+MATO	**0.3762**	**0.2483**	**0.5478**	**0.6541**	**0.5831**	**0.4223**	**0.3925**	0.5599	**0.5444**	**0.4573**	**0.5025**
	RES15	BDTF	0.3704	0.1805	0.5597	0.7244	0.5818	0.4039	0.3290	0.4185	0.5487	0.4266	0.4708
		+MATO	**0.3574**	**0.1716**	**0.5517**	**0.6936**	**0.5531**	**0.3705**	**0.2909**	**0.4085**	**0.5223**	**0.4063**	**0.4564**
	RES16	BDTF	0.3896	0.2142	0.5685	0.6891	0.5854	0.4757	0.4283	0.5988	0.5678	0.4806	0.5177
		+MATO	**0.3646**	**0.1992**	**0.5554**	**0.6567**	**0.5574**	**0.4560**	**0.3994**	0.5988	**0.5422**	**0.4614**	**0.5025**
5	LAP14	COM-MRC	0.2986	0.1647	0.4284	0.5874	0.5139	0.4245	0.3702	**0.6013**	0.4750	0.4107	0.4455
		+MATO	**0.2795**	**0.1520**	**0.3834**	**0.5289**	**0.4059**	**0.3771**	**0.2716**	0.6032	**0.3975**	**0.3708**	**0.4169**
	RES14	COM-MRC	0.1919	0.1425	0.3004	0.3453	0.2939	0.2231	0.2016	**0.5508**	0.2853	0.2794	0.3348
		+MATO	**0.1728**	**0.1168**	**0.2749**	**0.3159**	**0.2538**	**0.2020**	**0.1697**	0.5520	**0.2536**	**0.2577**	**0.3149**
	RES15	COM-MRC	0.2388	0.1080	0.4700	0.5067	0.4148	0.2883	0.2283	0.4348	0.4050	0.3250	0.3799
		+MATO	**0.2152**	**0.1030**	**0.3698**	**0.4543**	**0.3920**	**0.2561**	**0.2182**	**0.4103**	**0.3586**	**0.2896**	**0.3344**
	RES16	COM-MRC	0.1483	0.0892	0.2719	0.3369	0.2778	0.2045	0.1862	0.5865	0.2682	0.2605	0.3211
		+MATO	**0.1433**	**0.0780**	**0.2366**	**0.2919**	**0.2361**	**0.1791**	**0.1551**	0.5865	**0.2299**	**0.2386**	**0.2983**
6	LAP14	EMCGCN	0.2396	0.1300	0.3054	0.4627	0.3356	0.3421	0.2380	**0.5949**	0.3354	0.3304	0.3733
		+MATO	**0.2249**	**0.1263**	**0.3025**	**0.4315**	**0.3256**	**0.3073**	**0.2163**	0.6083	**0.3190**	**0.3167**	**0.3672**
	RES14	EMCGCN	**0.1493**	**0.1117**	0.2286	0.2683	0.2158	0.1610	0.1381	0.5511	0.2127	0.2297	0.2899
		+MATO	0.1577	0.1160	**0.2195**	**0.2666**	**0.2079**	**0.1567**	**0.1298**	**0.5474**	**0.2060**	**0.2277**	**0.2874**
	RES15	EMCGCN	0.2222	0.1051	**0.3546**	0.4382	0.3672	0.2447	0.1951	0.4332	0.3388	0.2847	0.3328
		+MATO	**0.1914**	**0.0981**	0.3598	**0.4042**	**0.3490**	**0.2195**	**0.1895**	**0.4234**	**0.3256**	**0.2694**	**0.3214**
	RES16	EMCGCN	0.2167	0.1284	0.3103	0.3868	0.3292	0.2303	0.2217	**0.5865**	0.3120	0.2972	0.3530
		+MATO	**0.1572**	**0.0963**	**0.2395**	**0.2966**	**0.2396**	**0.1713**	**0.1514**	0.5906	**0.2318**	**0.2433**	**0.3058**

Table 7: The MT results of other six models