Measuring Your ASTE Models in The Wild: A Diversified Multi-domain Dataset For Aspect Sentiment Triplet Extraction

Ting Xu♠ , Huiyun Yang♣, Zhen Wu♠ , Jiaze Chen♣, Fei Zhao♠, Xinyu Dai♠
♠National Key Laboratory for Novel Software Technology, Nanjing University
♣ByteDance
{xut, zhaof}@smail.nju.edu.cn, {wuz, daixinyu}@nju.edu.cn
{yanghuiyun.11, chenjiaze}@bytedance.com

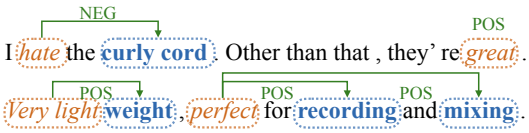
Abstract

Aspect Sentiment Triplet Extraction (ASTE) is widely used in various applications. However, existing ASTE datasets are limited in their ability to represent real-world scenarios, hindering the advancement of research in this area. In this paper, we introduce a new dataset, named DMASTE, which is manually annotated to better fit real-world scenarios by providing more diverse and realistic reviews for the task. The dataset includes various lengths, diverse expressions, more aspect types, and more domains than existing datasets. We conduct extensive experiments on DMASTE in multiple settings to evaluate previous ASTE approaches. Empirical results demonstrate that DMASTE is a more challenging ASTE dataset. Further analyses of in-domain and cross-domain settings provide promising directions for future research. Our code and dataset are available at https://github.com/NJUNLP/DMASTE.

1 Introduction

Aspect sentiment triplet extraction (ASTE; Peng et al., 2020), a fine-grained task in sentiment analysis (Hussein, 2018), has attracted considerable interest recently (Peng et al., 2020; Xu et al., 2020). The objective of this task is to extract the sentiment triplet, comprising of an aspect term, an opinion term, and a sentiment polarity, from a given review. As depicted in Figure 1, an example of the sentiment triplet is (*"curly cord"*, *"hate"*, *NEG*), representing a *negative* sentiment toward the aspect term *"curly cord"* using the opinion term *"hate"*. The ASTE task requires a deep understanding of linguistic forms and structures (e.g., aspect terms are usually nouns or verbs used as subjects or objects in a sentence), as well as the ability to identify the relationships between the various linguistic components (e.g., how to pair the aspect terms and opinion terms) in a given text.

Prior ASTE methods (Yan et al., 2021; Xu et al., 2021) have achieved promising results on existing academic datasets (Peng et al., 2020; Xu et al., 2020; Wu et al., 2020), greatly promoting the development and application of ASTE. However, the datasets employed in these studies remain comparatively uncomplicated, leading to disparities between these datasets and real-world settings in terms of various factors, such as length, expression diversity, domain distribution, etc. For instance, most reviews in existing datasets are of short length, with an average of 16 words per review, while reviews in real-world scenarios are longer (an average of more than 50 words). Additionally, expressions used in these datasets are typically simple and straightforward, with limited diversity in lexicality and syntactic. Furthermore, existing datasets typically contain two domains, i.e., restaurant and laptop, with very limited domain distributions. In a nutshell, these gaps hide the complexity of real-world scenarios, and therefore, impede the exploration to fully understand and address the challenges presented in real-world ASTE tasks.

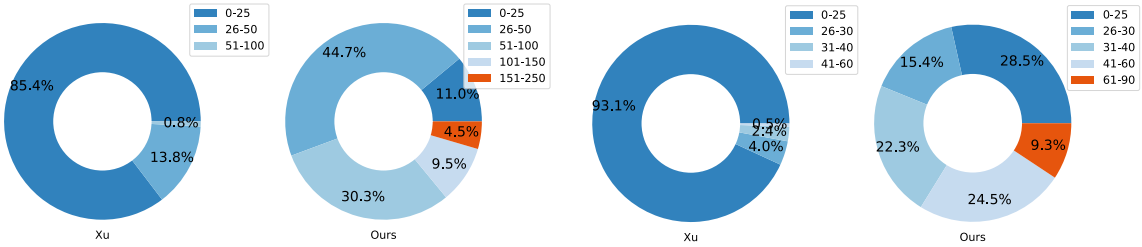


Triplets: (**curly cord**, *hate*, NEG); (**NULL**, *great*, POS);
(**weight**, *Very light*, POS); (**mixing**, *perfect*, POS);
(**recording**, *perfect*, POS).

Figure 1: An example of the ASTE task. The terms highlighted in blue are aspect terms. The terms in orange are opinion terms. The words in green are sentiment polarities. "NULL" denotes the implicit aspect.

In order to bridge the gap and better simulate real-world scenarios, we create a new dataset, named Diversified Multi-domain ASTE (DMASTE), which is manually annotated to provide a more diverse and realistic set of reviews for the task. As Table 1 and Figure 2 shows, the key

Findings of the Association for Computational Linguistics: ACL 2023, pages 2837–2853
July 9-14, 2023



(a) Distribution of review length.

(b) Distribution of the number of POS 2-gram.

Figure 2: Comparisons between Xu et al. (2020) and our dataset on review length and the number of POS 2-grams per review.

Dataset	#D	#R	#W/#R	#T	#IA	#T/#I	POS #n-gram	DP #n-gram	#vocab
$\mathcal{D}_P$	2	6037	16.49	9309	0	1.54	14.25/16.05/16.52	13.51/14.24/14.43	6512
$\mathcal{D}_W$	2	6009	16.46	10390	0	1.73	14.22/16.02/16.48	13.48/14.21/14.40	6484
$\mathcal{D}_X$	2	5989	16.43	10252	0	1.71	14.20/16.00/16.46	13.46/14.19/14.38	6467
DMASTE	8	7524	59.68	28233	11945	3.75	36.67/52.11/57.92	35.08/40.20/41.93	19226

Table 1: Comparisons between DMASTE and existing representative ASTE datasets, where $\mathcal{D}_P$, $\mathcal{D}_W$, $\mathcal{D}_X$ denote the datasets of Peng et al. (2020), Wu et al. (2020), and Xu et al. (2020). #D, #R, #W, #T, and #IA denote the numbers of domains, reviews, words, triplets, and triplets with implicit aspect terms, respectively. #n-gram denotes the numbers of 2/3/4-grams in part-of-speech sequences (POS) or dependency parsing trees (DP) per review. #vocab denotes the size of the vocabulary for each dataset.

characteristics of DMASTE can be summarized as follows: (1) **Various lengths**: DMASTE covers reviews of various lengths, ranging from 1 to 250 words, with an average of 59 words per review. (2) **Diverse expressions**: more part-of-speech and dependency n-grams show a wide variety of lexical bundles and syntactic in DMASTE, which can better represent the complexity and diversity of real-world scenarios. (3) **More aspect types**: DMASTE includes triplets annotated with both implicit and explicit aspect terms, providing a more comprehensive understanding of the target being discussed. (4) **More domains**: DMASTE covers eight domains, enabling more comprehensive research in ASTE like single-source and multi-source domain adaptation. To summarize, these characteristics make DMASTE a better benchmark to verify the ability of ASTE approaches in real-world scenarios.

To thoroughly investigate the challenges introduced by DMASTE and explore promising directions for future ASTE research, we implement several representative methods and empirically evaluate DMASTE under multiple settings:

- In-domain results show that the performance of current models declines significantly on DMASTE. And analysis reveals that *long reviews, complex sentences, and implicit aspect terms make* DMASTE *a challenging dataset*.
- In the single-source domain adaptation setting, we observe a positive correlation between transfer performance and domain similarity. But simply learning domain-invariant features may lead to the loss of task-specific knowledge, which suggests that *reducing domain discrepancy while keeping the task-specific knowledge can be a future direction*.
- We observe the *negative transfer* (Rosenstein et al., 2005) in the multi-source domain adaptation setting, and find the negative transfer occurs mainly in dissimilar source-target pairs. This indicates that *the domain similarity may be a useful guideline for domain selection* in future multi-source cross-domain ASTE research.

Overall, the results of DMASTE on multiple settings provide a deeper understanding of the challenges and future directions for ASTE research. We believe this work will bring a valuable research resource and benchmark for the community.

2 Related Work

2.1 ASTE Datasets

Current ASTE datasets (Peng et al., 2020; Xu et al., 2020; Wu et al., 2020) share a common origin and are constructed through similar processes. Specifically, they originate from SemEval Challenge

datasets (Pontiki et al., 2014, 2015, 2016), which provide aspect terms and corresponding sentiments for reviews in the domains of restaurant and laptop. Based on the datasets, Fan et al. (2019) annotate the aspect-opinion pairs. To provide more detailed information about the review text, researchers resort to extracting sentiment triplets from the review, i.e., aspect term, opinion term, and sentiment polarity. Peng et al. (2020) and Wu et al. (2020) construct the ASTE datasets by aligning the aspect terms between the datasets of Fan et al. (2019) and original SemEval Challenge datasets. As noted by Xu et al. (2020), the dataset of Peng et al. (2020) does not contain cases where one opinion term is associated with multiple aspect terms. Xu et al. (2020) subsequently refine the dataset and release a new version.

However, all of these datasets contains reviews of limited diversity from only two domains. Additionally, they all require aspect terms to align the aspect-sentiment pair and aspect-opinion pair, thus they do not include implicit aspect terms (Poria et al., 2014). Our dataset, DMASTE, addresses these limitations by providing a more diverse set of reviews covering more domains and annotate triplets with both implicit and explicit aspect terms, making it better suited for real-world scenarios [1].

2.2 ASTE Methods

Corresponding solutions for ASTE can be divided into three categories: tagging-based (Li et al., 2019; Peng et al., 2020; Xu et al., 2020; Zhang et al., 2020; Wu et al., 2020; Xu et al., 2021), MRC-based (Chen et al., 2021; Mao et al., 2021) and generation-based (Zhang et al., 2021b; Yan et al., 2021; Fei et al., 2021; Mukherjee et al., 2021). The tagging-based method employs a sequence or grid tagging framework to extract the aspect and opinion terms, then combines them to predict the sentiment. The MRC-based method constructs a specific query for each factor in the triplet and extracts them through the answer to the query. The generation-based method transforms the ASTE task into a sequence generation problem and employs sequence-to-sequence (seq2seq) models. Then it decodes the triplets through a specifically designed algorithm. In this paper, we employ some representative methods in three categories and explore DMASTE in multiple settings.

3 Dataset

To construct a dataset that is more representative of real-world scenarios, we manually annotate a new dataset, named Diversified Multi-domain ASTE (DMASTE). In this section, we first present a detailed description of the data collection and annotation process. Then we demonstrate the superiority of DMASTE, through a comparison with previous datasets in terms of key statistics and characteristics.

3.1 Collection

The data collection process of DMASTE is carried out in three stages: (1) We select the Amazon dataset (Ni et al., 2019) as our source of data due to its large volume of reviews from various regions around the world, which aligns with the goal of creating a dataset that is more representative of real-world scenarios. (2) We select four of the most popular domains from the Amazon dataset (Appendix A), and randomly sample a portion of the data for annotation. (3) To further enable comprehensive exploration of ASTE like domain adaptation settings, we additionally sample four additional domains and annotate a smaller portion of the data for testing in domain adaptation settings.

3.2 Annotation

Simplified Annotation Guidelines [2]. Following Peng et al. (2020); Xu et al. (2020); Wu et al. (2020), we annotate *(aspect term, opinion term, sentiment polarity)* triplets for each review, which may include multiple sentences. An *aspect term* refers to a part or an attribute of products or services. Sometimes, the aspect term may not appear explicitly in the instance, i.e., implicit aspect terms (Poria et al., 2014). We keep the triplets with implicit aspect terms. An *opinion term* is a word or phrase that expresses opinions or attitudes toward the aspect term. A *Sentiment polarity* is the sentiment type of the opinion term, which is divided into three categories: positive, negative, and neutral. It is worth noting that one aspect term can be associated with multiple opinion terms and vice versa.

Annotation Process. In order to ensure the quality of the annotation, we employed 14 workers (an-

[1] Cai et al. (2021); Zhang et al. (2021a) can also be transformed into sentiment triplets. We omit the comparison with them since they share almost the same data origination with the existing triplet dataset.

[2] Due to the space limitation, we present detailed labeling guidelines in Appendix A.2.

Domain	Electronics	Fashion	Beauty	Home	Book	Pet	Toy	Grocery
Train	1395	851	535	1050	0	0	0	0
Dev	200	121	77	152	159	167	173	173
Test	399	245	154	301	325	340	354	353

Table 2: Statistics of DMASTE dataset.

Term	Aspect terms	Opinion terms	Triplets
IAA	0.666	0.664	0.593

Table 3: Inter-annotator agreement (IAA) of aspect terms, opinion terms, and sentiment triplets.

notators) to perform the annotation and 4 workers (verifiers) to perform the quality-assurance sampling. Both groups are compensated based on the number of annotations. The annotation process is carried out using a train-trial-annotate-check procedure. (1) *Train*: All workers are trained on the task of annotation. (2) *Trial*: All workers try to annotate a small portion of the data to familiarize themselves with the task and to receive feedback on their annotations. Workers are required to reach 95% accuracy in labeling before proceeding to the next phase. Following previous work in data annotation for ABSA (Barnes et al., 2018), we evaluate the inter-annotator agreement (IAA) using the AvgAgr metric (Esuli et al., 2008). The IAA scores of annotated aspect terms, opinion terms, and sentiment triplets at the trial stage are shown in Table 3. These scores are slightly lower than previous ABSA datasets (Barnes et al., 2018), which can be attributed to the inherent complexity of DMASTE. (3) *Annotate*: The annotators are responsible for annotating the data on a daily basis. (4) *Check*: The verifiers sampled 20% of the annotations each day to ensure accuracy. If the accuracy of an annotator is found to be below 95%, the data annotated by that worker for that day would be re-done until meeting the accuracy requirement. Otherwise, the annotations are accepted. To avoid false positives by the verifiers, we introduce an appeals mechanism. Please refer to the appendix A.2 for detailed information.

3.3 Statistics and Characteristics

To provide a deeper understanding of our proposed dataset, we present a series of statistics and characteristics in this section.

Various Lengths. Figure 2a illustrates the comparison of review length between Xu et al. (2020) and DMASTE. We find that DMASTE contains reviews of more varied lengths. And Table 1 demonstrates that the average length of DMASTE is 3.6 times that of Xu et al. (2020).

Diverse Expressions. We quantify the expression diversity by counting the vocabulary size, n-grams in part-of-speech (POS) sequences and dependency parsing (DP) trees [3]. Figure 2b shows the comparison of POS 2-grams between Xu et al. (2020) and DMASTE. The results indicate that DMASTE contains reviews with diverse expressions. And Table 1 shows that the number of n-grams of POS and DP trees in DMASTE is significantly higher than those of other datasets, with the number of 4-grams being about 3 times that of other datasets. Besides, the vocabulary size of DMASTE is about 3 times that of other datasets. These statistics indicate that the reviews in DMASTE are more diverse and complex.

More Aspect Types. Table 1 shows that DMASTE not only contains more explicit aspect terms than existing datasets but also includes annotations of implicit aspect terms, which constitute a large proportion. This provides a more comprehensive understanding of the target being discussed.

More Domains. Table 1 illustrates that our dataset comprises four times the number of domains compared to existing datasets. Furthermore, the first four domains presented in Table 2, which are characterized by a larger amount of annotated data, represent leading fields in e-commerce platforms (Appdenix A), providing a more realistic representation of popular topics. Additionally, we include an additional four domains with less annotated data to enable a more comprehensive analysis of the domain adaptation setting.

In summary, DMASTE is a more realistic and diversified dataset, providing a suitable testbed to verify the ability of ASTE methods in real-world scenarios.

[3] For the DP tree, we treat the parent node and child node as neighbors.

4 Experiment Settings

To thoroughly understand DMASTE and provide some promising directions for future ASTE research, we conduct experiments in multiple settings. This section will first provide an overview of the different experimental setups, then introduce the evaluation metric, and finally, describe the models employed in the experiments.

4.1 Setups

We conduct a series of experiments under comprehensive training and testing setups:

- *In-domain*: we train and test the models with data from the same domain.
- *Single-source Cross-domain*: we train the models with data from a single-source domain and test them on a different target domain.
- *Multi-source Cross-domain*: we train the models with data from multiple source domains and test them on a different target domain.

In the cross-domain setting, we regard Electronics, Fashion, Beauty, Home as the source domain, and Book, Pet, Toy, Grocery as the target domain. More training details are shown in Appendix B.1.

4.2 Evaluation Metric

Following Xu et al. (2021), we employ the F1 score to measure the performance of different approaches. All the experimental results are reported using the average of 5 random runs.

4.3 Models

This section presents the various approaches we evaluate on the DMASTE. We first introduce representative models for ASTE and employ them as our baseline models. Then for the single-source cross-domain setting, we utilize current methods for ASTE and integrate some of them with the adversarial training (Ganin et al., 2016), which is a widely-used technique in domain adaptation.

Baseline Models For ASTE. We implement several representative models from various frameworks, including tagging-based, MRC-based, and generation-based frameworks.

- Span-ASTE (Xu et al., 2021): a tagging-based method. It explicitly considers the span interaction between the aspect and opinion terms.
- BMRC (Chen et al., 2021): a MRC-based method. It extracts aspect-oriented triplets and opinion-oriented triplets. Then it obtains the final results by merging the two directions.
- BART-ABSA (Yan et al., 2021): a generation-based method. It employs a pointer network and generates indices of the aspect term, opinion term, and sentiment polarity sequentially.
- GAS (Zhang et al., 2021b): a generation-based method. It transforms the ASTE tasks into a text generation problem.

Baseline Models For Single-source Cross-domain Setting. We incorporate Span-ASTE and BMRC with adversarial training (AT), a common strategy in domain adaptation. Specifically, we apply a domain discriminator on different features for each method:

- BMRC+AT: We apply a domain discriminator on the token and [CLS] features. In this way, we can learn discriminative features by classifiers of the ASTE task and domain-invariant features by the domain discriminator induced by adversarial training.
- Span-ASTE+AT: As the extraction in this method is based on the prediction of the span and pair representation, which are derived from token representation, we apply adversarial learning to the token representations, similar to the BMRC+AT model.

5 Results

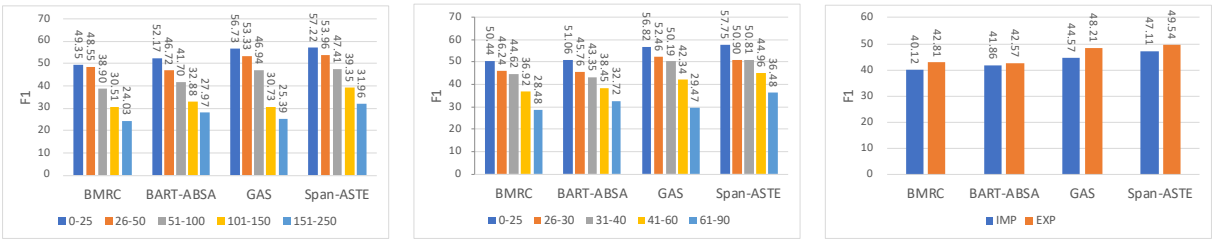
This section presents thorough analyses of the challenges of the DMASTE dataset and ASTE task, with the aim of better understanding these challenges and suggesting promising directions for future research. For each setting, we first show experimental results. Then we perform comprehensive analyses to investigate the challenges of the dataset and task, highlighting the limitations of current approaches. Finally, we provide promising directions for future research in this area.

5.1 In-domain

The overall in-domain experimental results are shown in Table 4. We first explore the limitations of baseline models to better understand challenges introduced by DMASTE. Then we compare two representative models to find promising directions for future research.

Method	Electronics	Beauty	Fashion	Home	Average
BMRC	41.95±0.34	38.57±0.97	44.87±0.69	41.18±0.66	41.64
BART-ABSA	43.38±1.37	41.13±1.14	43.89±0.82	40.56±1.04	42.24
GAS	47.10±0.64	44.32±0.52	47.80±1.07	47.22±1.13	46.61
Span-ASTE	**47.86**±0.74	**46.46**±0.66	**50.38**±0.68	**49.14**±0.41	**48.46**

Table 4: F1 scores of in-domain ASTE on DMASTE, and the best results are highlighted in bold font. Span-ASTE is significantly better than other methods with $p < 0.05$.



(a) F1 scores on reviews of different lengths.

(b) F1 scores on reviews of different numbers of POS 2-grams.

(c) F1 scores on reviews of different aspect types.

Figure 3: F1 scores on different review lengths, different numbers of POS 2-grams, and different aspect types. Results are average F1 scores on the four in-domain results. We can conclude that long reviews, complex sentences, and implicit aspect terms make DMASTE more challenging.

Challenges of DMASTE. Results of the in-domain experiments, reported in Table 4, show a performance drop compared with results on existing datasets (Span-ASTE gets a 59.38 score for the Laptop domain in Xu et al. (2020) and 47.86 for the Electronics domain in DMASTE). To further investigate the challenges of DMASTE, we further analyze the performance under different review lengths, sentence complexity, and aspect types.

- **Length.** Figure 3a illustrates the relationship between the review length and the model performance on DMASTE. Results show that the performance of the models decreases as the length of the review increases. This indicates long reviews present a significant challenge for current models.

- **Sentence Complexity.** We quantify the sentence complexity by calculating the number of 2-grams in the part-of-speech (POS) sequence of the review text. Then we analyze the relationship between the extraction performance and sentence complexity. As shown in Figure 3b, we observe a decline in performance with an increase in sentence complexity. This highlights the challenges posed by the diversified expression presented in DMASTE.

- **Aspect Types.** In Figure 3c, we analyze the performance of models on triplets with implicit and explicit aspect terms. Results demonstrate that the extraction of implicit aspect terms is more challenging than that of explicit aspect terms. The inclusion of both implicit and explicit aspect terms in DMASTE makes it more challenging for ASTE.

We can conclude that long reviews, complex sentences and implicit aspect terms make DMASTE a challenging dataset.

Model Comparison. Prior works have demonstrated that GAS (Zhang et al., 2021b) outperforms Span-ASTE (Xu et al., 2021) on the dataset of Xu et al. (2020). However, as shown in Table 4, Span-ASTE outperforms GAS on DMASTE. We conduct further analysis and discover that the reversal in performance can be attributed to the long instance lengths and complex sentence patterns presented in DMASTE. As illustrated in Figure 3, GAS achieves comparable results when the reviews are short and simple, but its performance declines more sharply when encountering long and complex reviews in comparison to Span-ASTE. This can be attributed to the generation-based nature of GAS, which is prone to forgetting the original text and misspellings or generating phrases that do not appear in the review when dealing with long reviews. In contrast, the tagging-based nature of Span-ASTE only identifies the start and end tokens for the aspect and opinion terms, which is less affected by the complexity and length of the

Domain	BMRC	BART-ABSA	GAS	Span-ASTE	BMRC+AT	Span-ASTE+AT
Electronics→ Book	33.74±0.64	35.43±0.61	35.57±0.76	**40.36**±1.59	32.56±1.57	39.58±0.70
Beauty→ Book	30.01±1.29	30.24±0.83	30.96±0.99	38.58±1.58	28.56±0.88	**38.79**±0.89
Fashion→ Book	31.71±1.37	32.45±1.77	36.26±0.64	**39.88**±0.74	30.80±1.59	39.60±0.69
Home→ Book	31.93±1.79	33.48±1.15	35.91±0.75	**39.45**±0.73	31.31±0.45	39.35±0.93
Electronics→ Grocery	39.68±0.70	40.39±1.28	39.16±1.25	**45.36**±1.12	40.15±0.55	43.90±0.47
Beauty→ Grocery	34.46±0.85	34.22±0.59	36.22±0.50	**40.32**±0.81	34.91±1.00	40.30±1.64
Fashion→ Grocery	37.18±0.50	37.27±1.08	40.13±0.38	**43.41**±0.83	37.56±0.29	42.29±0.79
Home→ Grocery	39.03±1.78	38.56±1.14	42.51±0.37	**43.74**±1.02	38.83±0.75	41.78±1.79
Electronics→ Toy	42.39±0.88	42.49±1.13	43.55±0.52	47.23±0.38	41.38±1.20	**47.43**±0.67
Beauty→ Toy	35.66±0.94	33.87±1.05	37.95±0.63	**41.19**±0.60	35.98±1.14	40.86±0.74
Fashion→ Toy	41.13±1.12	40.08±1.15	42.78±0.36	**46.83**±0.94	40.74±1.20	46.09±0.87
Home→ Toy	40.26±1.22	40.81±1.27	44.16±0.73	**47.60**±1.09	40.42±0.51	46.47±0.75
Electronics→ Pet	37.39±0.60	36.88±1.07	38.17±0.89	**41.04**±0.48	36.70±0.43	40.09±0.77
Beauty→ Pet	32.80±1.07	32.07±0.76	32.55±0.57	**36.41**±0.40	32.76±0.54	35.38±0.83
Fashion→ Pet	35.97±0.84	34.92±0.85	36.13±0.74	**40.57**±0.57	36.07±0.81	38.78±0.81
Home→ Pet	37.64±1.03	37.26±1.59	39.38±0.70	**41.42**±0.51	37.24±0.97	40.64±1.05
Average	36.31	36.28	38.21	**42.09**	36.00	41.33

Table 5: F1 scores of single-source cross-domain ASTE on DMASTE. We highlight the best results in bold. Span-ASTE performs significantly better than BMRC, BART-ABSA, and GAS with $p < 0.01$.

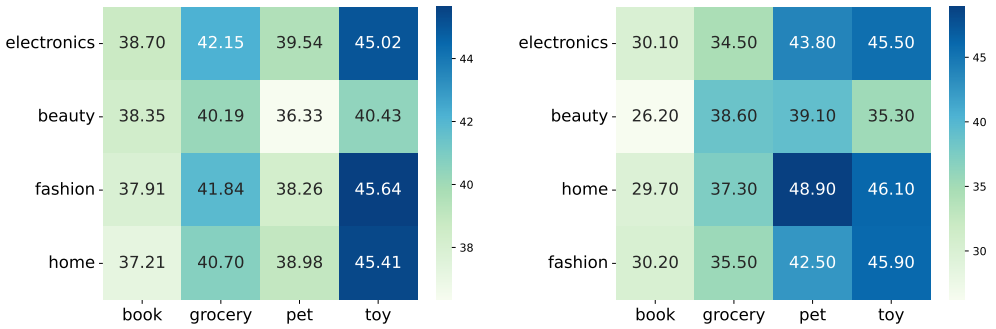
reviews. This analysis provides insights for future generation-based methods: *implementing a comparison algorithm between the generated output and the original input text and making modifications if they are mismatched.*

5.2 Single-source Cross-domain

We present the overall single-source cross-domain experimental results in Table 5 (additional training details of adversarial training are shown in Appendix B.2). We then conduct a correlation analysis to investigate the factors that impact the transfer performance. Additionally, we analyze current domain adaptation strategies and provide insights for future research in this area.

Model Performance v.s. Domain Similarity. Table 5 reveals that performance varies significantly when transferring from different source domains. For instance, the F1 score of Span-ASTE on Home→ Toy is 47.60, which is 6.41 higher than that on Beauty→Toy. To gain insights into the factors that impact the transfer performance, we conduct the Pearson Correlation Analysis (Benesty et al., 2009) on the relationship between the model performance and domain similarity based on Span-ASTE. Specifically, we fix the number of training data for each source domain to alleviate the impact of data volume. Following Liu et al. (2021), we measure the domain similarity by computing vocabulary overlaps, using the top 1k most frequent words in each domain excluding stopwords. Results in Figure 4 show that there is a positive correlation (with a 0.52 Pearson correlation coefficient) between model performance and domain similarity. This indicates that large domain discrepancy is a huge challenge for the single-source cross-domain ASTE task. Therefore, *reducing the domain discrepancy is a promising way to improve the transfer performance in cross-domain ASTE.*

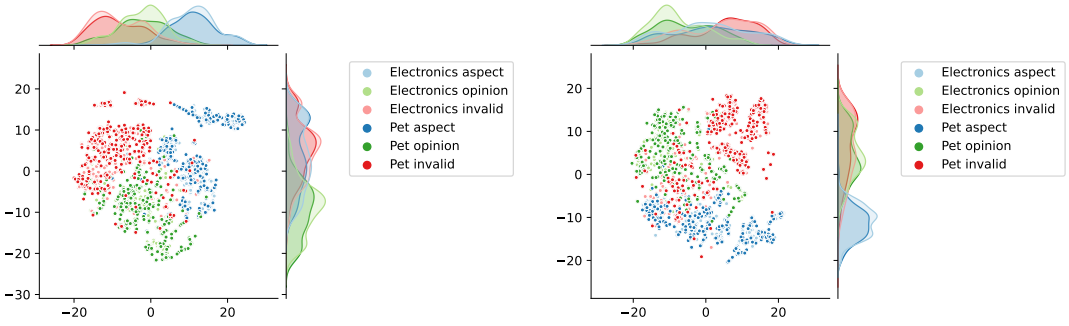
With AT v.s. Without AT. We compare the performance of BMRC and BMRC+AT, as well as Span-ASTE and Span-ASTE+AT in Table 5. Results indicate that adversarial training has a negative impact on the performance of cross-domain ASTE. To investigate the cause of performance degradation, we visualize the representations learned from the models when transferring from the Electronics to the Pet domain in Figure 5. Compared with Span-ASTE, the features of different categories in Span-ASTE+AT are less discriminable, especially on the x-axis. We attribute the performance drop to feature collapse (Tang and Jia, 2020) induced by adversarial training. This occurs when the model focuses on learning domain-invariant features while ignoring the discriminability of each category. This issue is particularly pronounced in the ASTE task, as it requires fine-grained discrimination for three factors. Future research in cross-domain ASTE could focus on *developing methods that can learn domain-invariant features while maintaining the discriminability of each category.*



(a) Model performances on single-source cross-domain ASTE.

(b) Vocabulary overlaps (%) between the source domains and target domains.

Figure 4: Correlation analysis between the model performance and domain similarity.



(a) Span-ASTE

(b) Span-ASTE+AT

Figure 5: T-SNE visualization of Span-ASTE and Span-ASTE+AT on Electonics→Pet. Compared with Span-ASTE, the features of different categories in Span-ASTE+AT are less discriminable, especially on the x-axis.

5.3 Multi-source Cross-domain

In this section, we conduct multi-source cross-domain experiments with the number of source domains varying from 2 to 4. The results are shown in Table 6. Our results reveal instances of negative transfer. For example, F + H → Book > E + F + H → Book.

Negative transfer indicates that transferring from some domains could harm the learning of the target domain (Guo et al., 2018). To further investigate the phenomenon of negative transfer, we conduct an analysis on the relationship between domain similarity and transfer performance. We further compare the results with the domain similarity in Figure 4b. The results indicate that half of the negative transfers are observed when transferring from the source domains with the least similarity to the target domain (e.g., F + H → Toy > B + F + H → Toy. Adding the Beauty domain leads to negative transfer and it is the least similar domain with Toy). This suggests that domain similarity can serve as a useful guideline for selecting source domains in future multi-source cross-domain research.

6 Conclusion and Future Directions

We propose DMASTE, a manually-annotated ASTE dataset collected from Amazon. Compared with existing ASTE datasets, DMASTE contains reviews of various lengths, diverse expressions, more aspect types and covers more domains, which indicates DMASTE is a suitable testbed to verify the ability of ASTE methods in real-world scenarios. We explore the dataset in multiple scenarios, i.e., in-domain, single-source cross-domain, and multi-source cross-domain and provide some promising directions for future research. For the in-domain setting, we compare the results between existing datasets and DMASTE and find that the long reviews, complex sentences, and implicit aspect terms make DMASTE a challenging dataset. For the single-source cross-domain setting, we observe that domain similarity and cross-domain performance are positively correlated. Furthermore, analysis of adversarial training shows that simply learning domain-invariant features may lead to feature collapse and result in the loss of task-specific

Domain	Book	Grocery	Toy	Pet	Average
B + E	40.94±1.14	45.94±0.56	48.07±0.75	41.70±0.79	44.16
E + F	40.59±0.86	46.14±0.98	50.10±0.83	41.90±0.79	44.68
E + H	41.08±1.50	45.34±0.50	48.97±0.69	42.70±0.41	44.52
B + F	41.13±0.51	44.60±0.73	46.81±0.91	41.03±0.59	43.39
B + H	41.31±1.18	44.31±0.84	47.12±0.87	41.67±0.93	43.60
F + H	42.40±0.31	45.97±0.56	49.11±0.74	43.12±0.70	45.15
B + E + F	41.34±0.73	46.28±0.78	49.44±1.08	41.12±0.58	44.80
B + E + H	41.21±0.48	45.39±0.49	49.33±0.32	43.48±1.12	44.85
E + F + H	41.44±0.71	45.39±0.80	50.24±0.57	43.59±1.06	45.17
B + F + H	41.73±0.38	45.40±0.63	48.55±0.71	43.46±0.19	44.79
ALL	41.83±0.57	46.07±0.47	50.16±0.49	43.62±1.32	45.42

Table 6: Comparison results of multi-source cross-domain ASTE on Span-ASTE. We use abbreviations due to space limitation: B: Beauty, H: Home, F: Fashion, E: electronics, ALL: all the source domains.

knowledge. Therefore, it is important to design appropriate methods to reduce the domain discrepancy while preserving fine-grained task features for ASTE tasks. In multi-source domain adaptation, we find that most of the negative transfer comes from dissimilar source-target pairs, pointing out that domain similarity can be a domain selection guideline for future research. In conclusion, we hope that our dataset DMASTE and analyses will contribute to the promotion of ASTE research.

Limitations

We analyze the limitations of this study from the following perspectives:

- The ASTE task extracts the sentiment triplets from a review, while the Aspect Sentiment Quad Prediction (ASQP) task adds an aspect category based on the triplets and provides more comprehensive information. Defining the aspect category for each domain is also hard work. Future work can take the aspect category into consideration.

- All the models are evaluated by F1 score, in which only exact matching can be considered correct. This metric can not differentiate between partially matching and completely mismatching and is not the best choice for a challenging dataset like DMASTE. Future work can include some partially matching metrics for this task.

- There is no specifically designed method for cross-domain ASTE. But we analyze the challenges of this task in detail. We are planning to design a new method for cross-domain ASTE based on the analysis results.

Acknowledgments

This work is done during Ting Xu's internship at ByteDance. We would like to thank the anonymous reviewers for their insightful comments. Zhen Wu is the corresponding author. Ting Xu would like to thank Siyu Long for his constructive suggestions. This work is supported by National Natural Science Foundation of China (No. 62206126, 61936012, and 61976114).

References

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. 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 340–350, Online. Association for Computational Linguistics.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Proceedings Of The AAAI Conference On Artificial Intelligence*, volume 35, pages 12666–12674.

Andrea Esuli, Fabrizio Sebastiani, and Ilaria Urciuoli. 2008. Annotating expressions of opinion and emotion in the Italian content annotation bank. In *Proceedings of the Sixth International Conference on*

Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

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

Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. 2021. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium. Association for Computational Linguistics.

Doaa Mohey El-Din Mohamed Hussein. 2018. A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 30(4):330–338.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6714–6721.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13543–13551.

Rajdeep Mukherjee, Tapas Nayak, Yash Butala, Sourangshu Bhattacharya, and Pawan Goyal. 2021. PASTE: A tagging-free decoding framework using pointer networks for aspect sentiment triplet extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9291, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8600–8607.

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

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495.

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

Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, and Alexander Gelbukh. 2014. A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, pages 28–37.

Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. 2005. To transfer or not to transfer. *NIPS*.

Hui Tang and Kui Jia. 2020. Discriminative adversarial domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5940–5947.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. 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.

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, Online. 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)*, pages 2339–2349, Online. 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 (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

Chen Zhang, Qiuchi Li, Dawei Song, and Benyou Wang. 2020. A multi-task learning framework for opinion triplet extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 819–828, Online. 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*, pages 9209–9219, Online and Punta Cana, Dominican Republic. 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 (Volume 2: Short Papers)*, pages 504–510.

Yue Zhang, Quan Guo, and Parisa Kordjamshidi. 2021c. Towards navigation by reasoning over spatial configurations. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 42–52, Online. Association for Computational Linguistics.

A Dataset

A.1 Dataset Details

Domain Selection Process. First, we select the most popular four domains based on the reports of Top selling products on Amazon and the Internet. For these four domains, we annotate more data to enable research on ASTE with more realistic and representative reviews. Then, to enable more comprehensive research like the cross-domain setting, we randomly select another four domains and annotate fewer data for testing in the cross-domain setting.

Sampled Reviews. We sample two reviews for each domain and demonstrate them in Table 7. The first column displays the review text. And the second column shows the extracted triplets (aspect terms, opinion terms, sentiment polarity).

Annotator Compensation. As described in Section 3.2, we hire 18 workers in the annotation and follow strict quality control to ensure the quality of the annotation. We ensure the privacy right of workers is respected in the annotation process. All workers have been paid above the local minimum wage and agree to use the dataset for research purposes.

License. DMASTE will be publicly avaliable under the terms of CC BY-NC-SA 4.0 License. The dataset is for academic use, which is consistent with its origination Amazon (Ni et al., 2019) dataset.

A.2 Annotation Guidelines

The purpose of this annotation guideline is to provide a consistent and accurate method for annotating reviews. The task of the annotator is to identify triplets comprising of the following types:

- **Aspect Term.** An Aspect term refers to a part or an attribute of products or services. Sometimes, the aspect term may not appear explicitly in the instance, i.e., implicit aspect (Poria et al., 2014). We keep the triplets with implicit aspects.
- **Opinion Term.** An opinion term is a word or phrase that expresses opinions or attitudes towards aspects.
- **Sentiment Polarity.** Sentiment polarity is the sentiment type of the opinion term, which is divided into three categories: positive, negative, and neutral.

Annotation Guidelines for Aspect Terms. Aspect terms can be divided into three types: (1) a part of the product, (2) an attribute of the product or store, (3) an attribute of a part of the product. Note, if the review expresses opinions toward some targets which is not explicitly mentioned, we annotate them as implicit aspect (NULL). If we have two aspect terms split with "and", ",", etc, label them separately. *Examples of aspect terms: battery, price, outlook, taste, customer support.*

Annotation Guidelines for Opinion Terms. Opinion terms usually expressed opinions or attitudes. (1) Label as much information as possible about "opinion of the product, user experience, an emotion about buying the product", etc. (2) The information selected should not change the original semantics. (3) Opinion terms are usually adjectives/adverbs, and they can be a phrase or a single word. (4) "definitely" or "very" alone are not enough to be labeled as an opinion term. (5) If we have two opinion terms split with "and", ",", etc, label them separately. *Examples of opinion terms: so good, expensive, love, unfriendly.*

Annotation Guidelines for Sentiment Polarities. Sentiment polarities can be divided into three types: (1) positive: the review expresses a positive attitude toward a specific aspect, (2) negative: the review expresses a negative attitude toward a specific aspect, (3) neutral: the review expresses no obvious positive or negative attitudes but it expresses an attitude. *Examples of sentiment polarities: (price, expensive, negative), (battery life, long, positive), (outlook, just okay, neutral).*

Abandon Cases. If the review is just some meaningless word, e.g. hhhh, ahhhh, oooooo, then abandon it.

Annotations Steps. When annotating the reviews, please follow the steps below:

1. Start annotating the reviews from left to right one by one using appropriate order from aspect terms, opinion terms to sentiment polarities.
2. Please make sure to select the right text boundary for aspect terms and opinion terms before clicking the "Mark this" button.

3. Please make sure to select a suitable category for sentiment polarities.

4. After double-checking the current triplet, click the "complete" button to proceed to the next triplet.

5. If the review contains more than one triplet, go through the same steps to annotate them and click "Submit" to proceed to the next review.

6. If the review satisfies the condition of abandon case, click "abandon" and proceed to the next review.

Appeal Process. During the check phase, the verifiers are not always right. Specifically, we introduce an Appeal process to ensure the high quality of dataset in the check phase. The detailed process is as follows:

1. If the verifier agrees with the annotation of the annotator, the annotation is added to the dataset. Otherwise, the verifier will annotate the data and send it to the annotator as feedback. If the annotator accepts the feedback, the annotation from the verifier is added to the dataset.

2. If the annotator disagrees with the feedback of the verifier, the annotator will appeal. The corresponding data will be discussed by all verifiers and annotators until they reach an agreement. After that, the new annotation is added to the dataset.

From the above process, verifiers sometimes play a role in annotators and they work together to ensure the high quality of dataset. We will add the above details in the revision.

B Experiments

B.1 Training Details

We utilize the pretrained model provided by HuaggingFace and run all the experiments on NVIDIA A100 GPU with pytorch. For the hyper-parameters of the baseline models, we follow the original settings in their paper (Chen et al., 2021; Yan et al., 2021; Zhang et al., 2021c; Xu et al., 2021). For adversarial training, we follow the implementation in Ganin et al. (2016). One hyper-parameter in this method is α, the ratio of training the generator to the discriminator. We search α in $\{1, 3, 5, 7, 10, 15, 20, 30, 50, 100\}$ and $\{1, 10, 30, 50, 100, 500, 700, 800, 1000, 1500\}$ for Span-ASTE + AT and BMRC + AT, respectively. For each value, we conduct experiments with 5 random seeds and set α by the F1 score on the development set. We set $\alpha = 10$ for Span-ASTE+AT and $\alpha = 800$ for BMRC+AT. The parameter search costs about 1000 GPU hours.

B.2 Detailed Results For Adversarial Training

We search the hyper-parameter α for adversarial learning on the development set. Detailed experiment results are shown in Table 8 and Table 9. We can observe that adversarial training is parameter-sensitive.

Electronics	
Good case . It has minimum padding that makes the phone feel secure but too much to be burdensome . The card slots are convenient .	(card slots, convenient, POS); (case, Good, POS); (padding, minimum, POS).
They work very well . The connections are nice and tight and I loose no quality over the length of the cord .	(connections, nice, POS); (connections, tight, POS); (length of the cord, loose no quality over, POS); (work, very well, POS).
Beauty	
This smells so good exactly like the body lotion and spray mist . It is a soft femine fragrance not too overpowering . I would highly recommend !	(NULL, would highly recommend, POS); (fragrance, not too overpowering, POS); (fragrance, soft femine, POS); (smells, so good, POS).
Revision products are awesome . I switch away from this in the winter months since it s not as moisturizing as an emollient based cleanser .	(Revision, awesome, POS); (NULL, not as moisturizing, NEG).
Home	
I think i just started a new hobby , this Victorinox is pretty awesome . I might have to get a few more designs .	(Victorinox, pretty awesome, POS).
Too big and cumbersome to be useful . If design were downsized a couple of inches , they would be perfect . Much too big for my salad bowls .	(NULL, Much too big, NEG); (NULL, Too big, NEG); (NULL, cumbersome, NEG).
Fashion	
the fabric is very , very thin . It should be underwear , not a regular shirt . fits too snug and is see through	(NULL, see through, NEG); (fabric, very thin, NEG); (fits, too snug, NEG).
What can I say that you don t already know ? Classic Chucks . One of the best , most versatile , and most durable sneakers out there .	(Chucks, Classic, POS); (sneakers, One of the best, POS); (sneakers, most durable, POS); (sneakers, most versatile, POS).
Book	
It was a good book , haven t read it before bow . Very interesting , with good suspense and humor mixed in ! 10 / 10	(NULL, Very interesting, POS); (book, good, POS); (humor, good, POS); (suspense, good, POS).
Challenging to read but worth it . Persevere to the end of the book . Ask God to open your mind and heart .	(NULL, Challenging to read, NEU); (NULL, worth, POS).
Grocery	
These are the best Slim Jims every , I have tried them , but the taste , the texture of the Honey BBQ is the Greatest .	(Slim Jims, best, POS); (taste, Greatest, POS); (texture, Greatest, POS).
A good almond extract , and I like that it s organic . I took off one star for high price . Should not cost that much .	(almond extract, good, POS); (organic, like, POS); (price, high, NEG).
Pet	
This is a good prefilter . The clear plastic adapters are a little brittle so be careful when attaching to your filter input .	(clear plastic adapters, a little brittle, NEG); (prefilter, good, POS).
Bought these as cat treats . Have to break them up , but my cat goes crazy for them , and this tub is lasting forever .	(NULL, goes crazy for, POS); (tub, lasting forever, POS).
Toy	
Super simple dynamics , and a great game for friends , family , kids , etc . ! Easy to learn , and variety of play is good .	(NULL, Easy to learn, POS); (dynamics, Super, POS); (dynamics, simple, POS); (game, great, POS); (variety of play, good, POS).
My 1 year old loved this for his birthday . Such a fun , easy toy . I would buy this again and again .	(NULL, loved, POS); (toy, easy, POS); (toy, fun, POS).

Table 7: Sampled reviews for each domain in DMASTE.

α	1	3	5	7	10	15	20	30	50	100
E→K	33.79	37.46	38.36	38.74	38.74	39.46	35.68	38.99	35.17	35.70
B→K	31.55	35.20	37.95	36.63	37.83	36.77	35.37	34.99	36.47	28.41
F→K	32.15	39.83	39.74	39.93	39.95	39.38	38.88	39.86	40.13	38.93
H→K	35.15	35.64	40.12	38.12	39.96	38.10	38.45	37.40	32.99	38.49
E→G	39.86	44.71	44.86	45.11	44.39	45.66	44.23	44.59	44.46	45.48
B→G	18.92	40.90	40.69	42.16	41.01	40.39	40.49	41.41	40.51	41.49
F→G	38.23	43.98	43.36	43.88	44.89	42.65	44.40	44.23	43.97	43.78
H→G	27.58	44.18	45.00	43.74	44.38	44.46	44.25	44.15	43.52	44.12
E→T	33.21	47.82	48.65	48.48	47.23	47.95	47.57	48.26	48.64	47.33
B→T	30.05	41.54	41.12	43.13	42.65	42.03	41.76	41.35	41.52	42.45
F→T	33.63	46.47	47.12	47.65	47.16	47.73	48.00	47.31	46.53	47.08
H→T	42.09	48.27	48.24	48.30	48.44	48.93	46.19	47.01	48.41	48.91
E→P	33.51	41.18	41.95	41.19	41.43	41.64	41.81	42.01	42.36	42.06
B→P	18.03	35.96	37.05	35.71	36.92	36.08	36.26	36.72	37.15	36.10
F→P	35.44	39.30	39.87	39.98	40.00	40.07	40.24	39.61	40.41	40.40
H→P	35.41	41.28	41.55	42.24	41.35	42.51	41.41	37.80	41.25	41.48
AVE	32.41	41.48	42.23	42.19	**42.27**	42.11	41.56	41.61	41.47	41.39

Table 8: F1 scores for Span-ASTE + AT. All the results are obtained on the development set by an average of 5 random runs. We highlight the best average results in bold.

α	1	10	30	50	100	500	700	800	1000	1500
E→K	0.75	2.22	6.18	16.90	6.61	28.08	31.86	31.67	30.64	30.23
B→K	0.30	1.04	9.57	17.53	26.56	27.93	28.62	27.62	29.46	29.88
F→K	0.22	7.66	8.05	14.82	27.14	30.34	32.09	30.72	31.21	31.09
H→K	0.37	9.81	11.22	18.96	22.74	26.66	27.92	31.23	28.52	28.93
E→G	2.43	12.42	24.97	19.98	35.43	38.71	39.14	39.48	39.05	38.13
B→G	2.13	20.84	33.99	32.90	32.63	33.92	33.84	32.88	32.89	32.78
F→G	2.42	37.73	27.40	34.71	38.03	38.98	37.99	37.56	38.62	38.20
H→G	4.55	30.18	32.95	37.50	37.72	37.62	37.86	38.48	37.65	37.85
E→T	4.57	31.09	35.18	38.12	42.73	43.35	42.74	42.86	43.01	42.59
B→T	1.65	7.45	23.93	33.70	35.53	36.41	35.39	35.46	35.90	35.73
F→T	4.10	37.50	38.43	40.38	41.23	42.66	41.65	42.26	42.21	42.58
H→T	5.11	35.05	41.77	43.45	43.22	44.54	44.43	44.27	43.79	44.72
E→P	10.47	29.64	37.53	36.99	38.14	37.62	37.78	38.59	38.03	37.65
B→P	3.15	26.35	30.14	32.04	31.64	31.55	32.29	32.44	33.25	32.20
F→P	8.24	33.79	36.40	37.27	36.35	36.72	37.20	36.79	36.82	37.19
H→P	17.51	32.01	39.06	37.24	37.80	38.30	38.33	38.87	39.24	39.29
AVE	4.25	22.17	27.30	30.78	33.34	35.84	36.20	**36.32**	36.27	36.19

Table 9: F1 scores for BMRC + AT. All the results are obtained on the development set by an average of 5 random runs. We highlight the best average results in bold.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

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

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

B ☑ Did you use or create scientific artifacts?

Section 3 for Dataset.

☑ B1. Did you cite the creators of artifacts you used?
Section 3 for Dataset and Appendix A for Dataset.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A for Dataset.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A for Dataset.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3 for Dataset and Appendix A for Dataset.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3 for Dataset and Appendix A for Dataset.

C ☑ Did you run computational experiments?

Section 4 for Experiment Settings, Section 5 for Results, and Appendix B for Experiments.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B for Experiments.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 for Experiment Settings and Appendix B for Experiments.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 for Experiment Settings, Section 5 for Results, and Appendix B for Experiments.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix B for Experiments.

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

Section 3 for Dataset and Appendix A for Dataset.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix A for Dataset.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix A for Dataset.

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

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