ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446120

DECOUPLING AND REFILLING: A SIMPLE DATA AUGMENTATION METHOD FOR ASPECT TERM EXTRACTION

Jiaxiang Chen *Yu Hong*⋆ *Chaoqun Liu* *Qingting Xu* *Guodong Zhou*

School of Computer Science and Technology, Soochow University, SuZhou, China

ABSTRACT

Aspect term extraction (ATE) is an important Natural Language Processing task, which aims to extract aspect terms from reviews. Recently, data augmentation has emerged as a reliable approach for relieving data sparsity in the NLP area. For ATE, self-labeling and semi-generation methods have been proposed to implement effective data augmentation. However, they either rely on external data or a pretrained generation model. In this paper, we propose a simple and self-contained augmentation method, which produces new instances for augmentation by context decoupling and infrequent term refilling, without using external data and generation models. We conduct experiments on four benchmark SemEval datasets. The test results show that our method yields substantial improvements, and performs comparably to the state-of-the-art method which uses external data.

Index Terms— Aspect Term Extraction, Data Augmentation, Sequence Labeling, Natural Language Processing

1. INTRODUCTION

Aspect **T**erm **E**xtraction (ATE) is required to extract aspect terms from reviews [1], where an aspect term is specified as a word or phrase that depicts the specific property of an entity or product. ATE is useful in the studies of sentiment analysis [1] and knowledge graph completion [2].

We conduct ATE in an extractive mode, where ATE is tackled as a sequence labeling task using B/I/O tags [3, 4, 5]. Within the tags, "*B*" (Begining) denotes the start position of an aspect term, and "*I*" (Inside) marks the inner words in it, while "*O*" (Outside) signals other words outside the aspect. As demonstrated in (1), the words "*counter service*" are respectively labeled as "*B*" and "*I*", depicting an unabridged aspect term, while other words are uniformly labeled as "*O*".

(1) **Sample labeled with B/I/O**: *The* [O] ***counter*** **[**B**]** ***service*** **[**I**]** *is* [O] *bad* [O] . [O]

ATE suffers from data sparsity [6, 7, 8], which is known as the lack of observable ATE examples for training. Recently, there are some effective data augmentation methods have been proposed, which create new observable examples by self-labeling [4, 9] and semi-generation [10] (Section 2) without manual intervention. In addition, Large Language Models (LLMs) such as ChatGPT [11] and T5 [12] are used as a model-level solution within a generative ATE framework.

In this paper, we propose a simple and self-contained data augmentation method. It decouples aspect terms from their contexts, and couples infrequent aspect terms with the contexts by refilling without changing the positions. Briefly, we merely conduct aspect term replacement for producing new instances. As shown in Example (2), we replace the phrase "*counter service*" in Example (1) with "*white tuna sashimi*", which allows a new instance to be produced.

(2) **Augmented Sample**: *The* [O] ***white*** **[**B**]** ***tuna*** **[**I**]** ***sashimi*** **[**I**]** *is* [O] *bad* [O] . [O]

In our experiments, all the infrequent aspect terms are selected from the training set, in terms of statistical information. During training, we avoid the biased over-fitting problem by alternating optimization over new and original instances. In addition, Flooding strategy [13] is used to prevent overlearning on the new instances, as some of them are illogical (e.g., "bottle was attentive"). Our experiments on SemEval datasets show that our method allows BERT$_{base}$- and BERT$_{pt}$-based ATE models to be stronger, where substantial improvements are obtained. Besides, the enhanced BERT$_{pt}$ achieves state-of-the-art performance on the test sets R14 and R15.

Other contributions of this study are as follows: 1) we simplify the augmentation process, without using generation models (e.g., Seq2Seq generator in semi-generation [10]) and external data (e.g., Yelp in self-labeling [4, 9]), and 2) the infrequent terms in the long tail of aspect term list can be more sufficiently learned, along with the newly-coupled contexts.

2. RELATED WORK

We concentrate on the previous studies that strengthen ATE by overcoming the data sparsity. They can be divided into the modes of leveraging LLMs and data augmentation.

Leveraging LLMs: It is suggested that LLMs [14] are promisingly effective in solving data sparsity, as they have been pretrained with profound knowledge, and sufficiently instructed for inference. Scaria et al. [12]'s InstructABSA uses

⋆ Corresponding author.

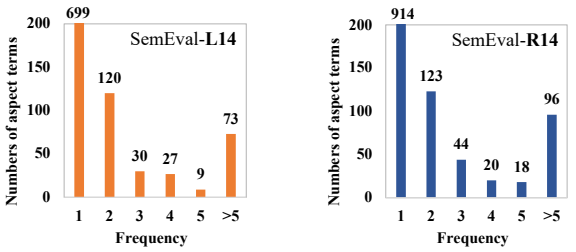


Fig. 1. The numbers of aspect terms occurring with different frequencies (Statistics on Restaurant-domain SemEval-R14 and Laptop-domain SemEval-L14 is shown respectively).

an elaborate prompt to instruct multi-task T5, which achieves the best performance in the generative ATE framework. Han et al. [11] leverage ChatGPT to perform generative ATE in zero-shot and few-shot scenarios. However, they demonstrate that ChatGPT fails to outperform some smaller language models (e.g., LSTM [15] and BERT$_{base}$ [16]) that are fine-tuned with in-domain or task-specific datasets.

Data Augmentation: Dai and Song [9]'s RINANTE is known as a self-labeling data augmentation method. It automatically concludes extraction rules from ground-truth data in terms of syntax, and applies the rules to annotate ATE examples in external datasets (e.g., Yelp) for data augmentation. Wang et al. [4] propose a progressive self-training method, which performs self-labeling over external datasets using the cyclically trained ATE models themselves. The self-labeled examples are utilized in the training-labeling cycle after being denoised. Li et al. [10] propose a self-generation method without using external data. It regenerates O-tagged fragments (i.e., non-aspect terms) in the ground-truth instances using an initialized-by-MASS Seq2Seq model [17], and uses the resultant new examples for data augmentation.

3. APPROACH

We follow the previous work [18, 19] to tackle the sentence-level ATE in an extractive model. Specifically, given a sentence X=$\{x_1, x_2, ..., x_n\}$ comprising n tokens, we predict a B/I/O tag for each token in X to output aspect terms.

We tend to strengthen ATE models (BERT$_{base}$ [16] and BERT$_{pt}$ [20]) by a novel data augmentation method, which is conditioned on an unbiased training. In the rest of this section, we respectively present the ATE models, data augmentation method and unbiased training process.

3.1. ATE Models

Structurally, both BERT$_{base}$ and BERT$_{pt}$ comprise 12-layer transformer encoders. The difference is that BERT$_{base}$ is pre-trained using open-domain data, while BERT$_{pt}$ is pretrained using in-domain data like Yelp and Amazon reviews.

Computationally, we use BERT to encode each token $x_i \in X$ into a hidden state $h_i \in \mathbb{R}^d$. Subsequently, we use a linear layer with Softmax activation to predict the B/I/O tag for x_i:

$$P(\hat{y}_i|x_i) = \text{Softmax}(h_i W + b) \quad (1)$$

where $h_i = \text{BERT}(x_i)$ and the function $\text{Softmax}(*)$ normalizes the logits of each token, converting them into probabilities $P(\hat{y}_i|x_i)$ of the B/I/O tags ($\hat{y}_i \in L$, $L = \{B, I, O\}$). Besides, $W \in \mathbb{R}^{d \times l}$ and $b \in \mathbb{R}^l$ (l=$\|L\|$) represent the trainable weights and biases, respectively.

During training, we use the cross-entropy loss to perform supervised fine-tuning for all parameters in the ATE models:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} P(y_i|x_i)\log(P(\hat{y}_i|x_i)) \quad (2)$$

where $P(y_i|x_i)$ is the l-dimensional vector mapped from the ground-truth L of x_i. In Section 3.3, we update the loss function to support unbiased training over the augmented data.

3.2. Data Augmentation

We augment the ATE training data with new instances, each of which is obtained using a simple term replacement method. Specifically, given a sentence X in the training set as well as the ground-truth aspect term $x_i \in X$, we decouple x_i from X to produce an exam question like that in the Cloze test. On this basis, we refill an infrequent aspect term into the blank of the exam question, which allows a new instance to be formed (see the example in (3)).

(3) **Sample labeled with B/I/O**: *The* [O] ***counter*** [B] ***service*** [I] *is* [O] *bad* [O] . [O]
Exam question: *The* _____ *is bad.*
New sample: *The white tuna sashimi is bad.*

The infrequent aspect term is selected from a list of terms ranked in descending order, where infrequent terms are of a lower order. The list comprises all aspect terms occurring in the ATE training set, and the frequency is obtained in terms of statistics on the set. Every time we conduct refilling, we select the most infrequent term from the bottom of the list. The frequency of such term will be subsequently increased by adding 1, and meanwhile the list will be updated by reranking all terms in both frequency order and alphabetical order. This ensures a dynamically changing list, and thus facilitates the selection of diverse infrequent terms for refilling.

The above augmentation method helps to relieve the weak supervision problem over the infrequent aspect terms, which is caused by unbalanced data. As shown in Fig. 1, there is a long tail of infrequent terms in the ATE training set, where 85.49% terms occur only once, while only 7.62% cases occur more than 5 times on L14. Coupling infrequent aspect terms with different contexts for producing additional training data is beneficial, i.e., contributing to supervised learning towards context-aware understanding of rare aspect terms.

Table 1. Quality verification results.

Annotator	#HLS	#IIP	#ILL
Human (on 500 new samples)	382	32	86
ChatGPT (on all new samples)	2,351	181	1,079

3.3. Unbiased Training

We verify the quality of the new ATE instances from three perspectives, including "Human-Like Sentence" (HLS), "Incorrect In Pragmatics" (IIP) and "ILLogical" (ILL). There are three annotators (2 masters and 1 expert) employed to annotate HLS, IIP and ILL for 500 new instances. They achieve a Cohen's kappa value [21] of 0.69 in the evaluation of tag consistency. The expert cooperates with them during calibration for controversial cases. We report the unanimous quality verification results in Table 1. In addition, ChatGPT is used to verify the quality for all the 3,611 new instances of R14. This global quality verification result is attached to Table 1. It can be observed that a considerable number of new instances are IIP or ILL. This easily results in misleading during training if overfitting and overlearning occur over the new instances.

To reduce data-specific overfitting, we train ATE models by alternating optimization. Specifically, we train them alternately using the original training data and newly-produced instances in the first $N/2$ epochs. In the remaining $N/2$ epochs, the models are exclusively trained using the original training data. N denotes the total number of training epochs.

To avoid overlearning on new instances, we use Flooding strategy [13] to prevent a quick convergence of the training process. By Flooding strategy, suboptimal parameters are obtained. It is implemented by firmly imposing a larger loss $\hat{\mathcal{L}}$ than the flooding level t upon the gradient approximation and back-propagation. $\hat{\mathcal{L}}$ is calculated as follows:

$$\hat{\mathcal{L}} = |\mathcal{L} - t| + t \ \ (t > 0) \tag{3}$$

where $\mathcal{L}$ is the actual loss, which is calculated by Equation 2.

4. EXPERIMENTATION

4.1. Datasets and Implementation Details

The benchmark datasets in our experiment are sourced from the SemEval, encompassing the domains of *Restaurant* and *Laptop* [1, 22, 23]. To provide a brief description, we utilize the mark "L" to denote "**L**aptop" and "R" to represent "**R**estaurant". Table 2 presents the statistics of the datasets. We evaluate all the models using F1-score.

We mainly compare our method to self-labeling (namely PST in [4]) and self-generation (CDA [10]), which have been used to enhance BERT$_{base}$ and BERT$_{pt}$ based models in the extractive ATE framework. Accordingly, we use the same optimizer and learning rate as that in [4]. Though, we set the number N of epochs to 12, while batch size is set to 10. The flooding level t is set to 0.01.

Table 2. Statistics of ATE datasets. #Sent and #Asp respectively denote the number of sentences and aspects.

		Train	Test			Train	Test
L14	#Asp	2,373	654	R14	#Asp	3,695	1,134
	#Sent	3,048	800		#Sent	3,044	800
R15	#Asp	1,199	542	R16	#Asp	1,743	612
	#Sent	1,315	685		#Sent	2,000	676

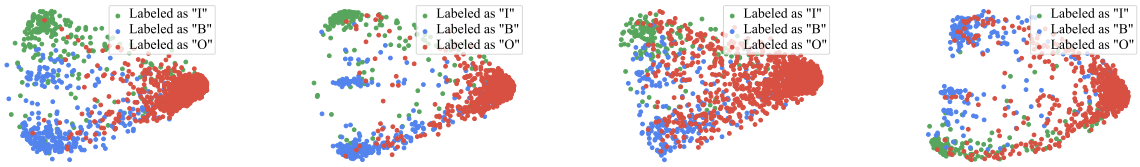
Table 3. $F1$-scores (%) of ATE models, where the mark "+" denotes the enhancement by a specific method. "$*$" denotes that our method yields significant improvements ($p < 0.05$) [24] compared to InstructABSA, BERT$_{base}$ and BERT$_{pt}$.

	L14	R14	R15	R16
ChatGPT Zero-shot [11]	43.03	55.65	40.33	-
ChatGPT 5-shot ICL [11]	48.19	70.99	53.49	-
ChatGPT 5-shot COT [11]	54.50	72.41	59.27	-
DECNN [25]	81.59	-	-	74.37
+CDA [10]	81.58	-	-	75.19
+Repositioning [19]	84.17	84.55	72.03	75.40
+PrototypeE [18]	83.19	87.39	73.27	76.98
BiLSTM-CRF [9]	73.47	84.06	66.17	-
+RINANTE [9]	80.16	86.45	69.90	-
InstructABSA [12]	72.33	81.10	64.35	67.66
+**Ours**	84.49*	88.46*	73.34*	77.80*
BERT$_{base}$ [16]	79.86	86.58	68.08	73.50
+CDA [10]	81.14	-	-	75.89
+PST [4]	84.17	87.63	72.81	77.09
+**Ours**	84.31*	88.27*	72.10*	78.27*
BERT$_{pt}$ [20]	84.37	88.41	73.66	78.29
+CDA [10]	85.33	-	-	80.29
+PST [4]	**86.91**	88.75	75.82	**82.56**
+**Ours**	86.05*	**89.13***	**76.09***	81.89*

In addition, we compare with InstructABSA, a generative ATE model that uses instruction-tuned Vanilla T5 [12]. For comparison purpose, we follow the common practice [4, 10, 11] to reevaluate InstructABSA using an exact-matching metric instead of partial matching. Besides, we retrain InstructABSA using our augmented data without changing its hyperparameters, and report the performance in the same way. We also compare to ATE performance reported in [11], where ChatGPT is prompted by 5-shot In-Context Learning (ICL) and Chain-Of-Thought (COT). All experiments are conducted on a GeForce RTX 2080 Ti GPU.

4.2. Main Results

The test results are shown in Table 3. It can be observed that our method significantly improves the baseline ATE models (InstructABSA, BERT$_{base}$ and BERT$_{pt}$) on all the SemEval test sets (L14-15 and R16). Besides, our method outperforms the data augmentation methods CDA and PST on L14 and R15-16 when BERT$_{base}$ is used for ATE. On the other hand, our method achieves better performance than CDA on L14



(a) Baseline on L14 (b) Our method on L14 (c) Baseline on R15 (d) Our method on R15

Fig. 2. Visualization of the distribution of each word.

Table 4. Results of ablation experiments.

	L14	R14	R15	R16
Ours	**84.31**	**88.27**	**72.10**	**78.27**
w/o Re	83.36	87.01	70.66	76.54
w/o Fl	83.86	88.13	71.91	77.70
w/o Al	82.93	86.92	70.53	76.81
w/o all	79.86	86.58	68.08	73.50

Table 5. Comparison of various training data.

	L14	R14	R15	R16
Gold	**79.86**	**86.58**	**68.08**	73.50
New	79.18	85.42	67.14	**74.17**

and R16, as well as comparable performance to PST on all test sets when BERT$_{pt}$ is used as the backbone.

Briefly, the comparison results can be concluded as follows. The current in-domain task-specific ATE models are more effective than instruction-tuned open-domain ChatGPT. Although simple, our method is effective for enhancing ATE models, and performs comparably to or even better than the augmentation methods employing external data.

4.3. Ablation Study

Table 4 shows the test results in the ablation experiments. We ablate different components of our method, including 1) replacing infrequent term **Re**filling by random refilling, 2) removing the constraint of **Fl**ooding strategy on overlearning, and 3) abolishing the **Al**ternating training, but instead training ATE models using the hybrid data (original instances plus our newly produced cases) in every epoch. It can be found that ablation causes performance degradation. The negative effect of ablating alternating training is the most significant.

4.4. Reliability of New Data

Although containing IIP and ILL cases (as founded in Section 3.3), the newly produced ATE instances (by refilling) are worth being independently used. As shown in Table 5, training the BERT$_{base}$ models only on the new instances actually causes a comparable performance to the ones trained on the gold SemEval training data. It can be also suspected that the current BERT-based ATE models are "tolerant" to some extent, which are less severely interfered by IIP and ILL cases.

Table 6. Statistics of salvage terms. The mark "m" denotes the missed cases by the baseline, while "s" denotes the salvaged cases by our approach.

	L14	R14	R15	R16
Unseen Term (s/m)	25/95	29/89	16/86	36/90
Long-tail Term (s/m)	8/19	5/11	5/13	5/13
Common Term (s/m)	10/16	10/26	9/43	18/50

4.5. Impact of Data Augmentation

We use Principal Component Analysis (PCA) [26] to investigate distributions of the test data in the feature space. PCA provides a 2-dimensional distribution map by dimensionality reduction of hidden states. We consider the hidden states produced by the BERT$_{base}$ model that is optimized on the gold training data, as well as that by training using our method. Fig. 2 shows the results obtained on L14 and R15. It can be observed that our method reduces the number of uncertain tokens occurring in the central area of "B", "I" and "O"-classes. This implies the ability of recognizing rare aspect terms.

Besides, Table 6 shows that there is a larger number of unseen and long-tail terms that can be salvaged.

5. CONCLUSION

We propose a simple external-data-independent data augmentation method. Experiments show that it is generally effective for enhancing both extractive and generative aspect term extraction models, and achieves comparable performance to the self-labeling (PST) method that uses external data, while outperforms the self-generation (CDA) method. In the future, we will study a twin agent-based extraction approach which prompts two LLM-based agents to produce highly reliable evidence for B/I/O tagging within a generative rebuttal scenario.

6. ACKNOWLEDGMENT

The research is supported by National Science Foundation of China (62376182, 62076174).

7. REFERENCES

[1] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect Based Sentiment Analysis," in *SemEval 2014: 27–35*.

[2] M. Zamini, H. Reza, and M. Rabiei, "A Review of Knowledge Graph Completion," *Information*, vol. 13, no. 8, pp. 396, 2022.

[3] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Coupled Multi-layer Attentions for Co-Extraction of Aspect and Opinion Terms," in *AAAI 2017: 3316–3322*.

[4] Q. Wang, Z. Wen, Q. Zhao, M. Yang, and R. Xu, "Progressive Self-training with Discriminator for Aspect Term Extraction," in *EMNLP 2021: 257–268*.

[5] A. Klein, O. Pereg, D. Korat, V. Lal, M. Wasserblat, and I. Dagan, "Opinion-based Relational Pivoting for Cross-domain Aspect Term Extraction," in *WASSA 2022: 104–112*.

[6] X. Li, L. Bing, P. Li, W. Lam, and Z. Yang, "Aspect Term Extraction with History Attention and Selective Transformation," in *IJCAI 2018: 4194–4200*.

[7] D. Ma, S. Li, F. Wu, X. Xie, and H. Wang, "Exploring Sequence-to-Sequence Learning in Aspect Term Extraction," in *ACL 2019: 3538–3547*.

[8] Y. Yang, K. Li, X. Quan, W. Shen, and Q. Su, "Constituency lattice encoding for aspect term extraction," in *COLING 2020: 844–855*.

[9] H. Dai and Y. Song, "Neural Aspect and Opinion Term Extraction with Mined Rules as Weak Supervision," in *ACL 2019: 5268–5277*.

[10] K. Li, C. Chen, X. Quan, Q. Ling, and Y. Song, "Conditional Augmentation for Aspect Term Extraction via Masked Sequence-to-Sequence Generation," in *ACL 2020: 7056–7066*.

[11] R. Han, T. Peng, C. Yang, B. Wang, L. Liu, and X. Wan, "Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors," *arXiv:2305.14450*, 2023.

[12] K. Scaria, H. Gupta, S. Goyal, S. A. Sawant, S. Mishra, and C. Baral, "InstructABSA: Instruction Learning for Aspect Based Sentiment Analysis," *arXiv:2302.08624*, 2023.

[13] T. Ishida, I. Yamane, T. Sakai, G. Niu, and M. Sugiyama, "Do We Need Zero Training Loss After Achieving Zero Training Error?," in *ICML 2020: 4604–4614*.

[14] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *NeurIPS 2022: 27730–27744*.

[15] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," *Neural computation*, vol. 9, pp. 1735–1780, 1997.

[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL 2019:4171–4186*.

[17] K. Song, X. Tan, T. Qin, J. Lu, and T. Liu, "MASS: Masked Sequence to Sequence Pre-training for Language Generation," in *ICML 2019: 5926–5936*.

[18] Z. Chen and T. Qian, "Enhancing Aspect Term Extraction with Soft Prototypes," in *EMNLP 2020:2107–2117*.

[19] Z. Wei, Y. Hong, B. Zou, M. Cheng, and J. Yao, "Don't Eclipse Your Arts Due to Small Discrepancies: Boundary Repositioning with a Pointer Network for Aspect Extraction," in *ACL 2020: 3678–3684*.

[20] H. Xu, B. Liu, L. Shu, and P. Yu, "BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis," in *NAACL 2019: 2324–2335*.

[21] J. Cohen, "Kappa: Coefficient of concordance," *Educ Psych Measurement*, vol. 20, no. 37, pp. 37–46, 1960.

[22] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval-2015 task 12: Aspect based sentiment analysis," in *SemEval 2015: 486–495*.

[23] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. D. Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryiğit, "SemEval-2016 task 5: Aspect based sentiment analysis," in *SemEval 2016:19–30*.

[24] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, "The hitchhiker's guide to testing statistical significance in natural language processing," in *ACL 2018: 1383–1392*.

[25] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction," in *ACL 2018: 592–598*.

[26] I. T. Jolliffe, "Principal component analysis," *Technometrics*, vol. 45, no. 3, pp. 276, 2003.