HEAT UP THE SENTIMENT LEARNING WITH ICE

Yao Yao[1,2], **Zuchao Li**[3,*] **and Hai Zhao**[1,2,*]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[3]National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University, Wuhan, 430072, P. R. China
yaoyao27@sjtu.edu.cn, zcli-charlie@whu.edu.cn,
zhaohai@cs.sjtu.edu.cn

ABSTRACT

Recently, dramatic gains have been made on the task of aspect sentiment triplet extraction (ASTE). In this paper, we introduce a straightforward pipeline model to perform two-stage sequence labeling, including aspect and opinion terms identification and aspect-opinion pair classification. To exploit the cross-sentence context information to the maximum extent possible, we propose the instance cooperative enhancement (ICE) by introducing unsupervised clustering methods. Through experimenting with various clustering methods, we found that GSDMM unleashes the potential of cross-sentence information to the most degree. Compared to current state-of-the-art models, the results show the effectiveness of our proposed framework on ASTE-Data-V2.

1 INTRODUCTION

ASTE aims at extracting aspect sentiment triplets in a sentence and each triplet is composed of an aspect term, an opinion term, and their sentiment dependency. An example of ASTE can be found in Appendix A. However, the recent work Zhong & Chen (2021) discovered that the contextual representations for the entity and relation models essentially capture distinct information. Inspired by them, we restore ASTE from joint model Xu et al. (2020) to pipeline model Li et al. (2019); Wang et al. (2017) and return to the original nature of ASTE by introducing a straightforward pipeline model for ASTE.

We also innovatively introduce instance cooperative enhancement (ICE) to help our model infiltrate the input with cross-sentence context. ICE allows instances to cooperate with each other and hence help our pre-trained encoder generate more specific representations. To be more specific, we first apply a collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (GSDMM) Yin & Wang (2014) to cluster similar sentences into groups. Then we concatenate the input sentence with cross-sentence context information within the same group. Many previous works proved that incorporating cross-sentence information can help model generate more accurate representations Wadden et al. (2019); Zhong & Chen (2021); Luan et al. (2019). To unleash the utmost potential of utilizing cross-sentence information, we conduct our experiments on various clustering methods, including simple clustering, k-means and GSDMM. We found that GSDMM shows great advantage in instance cooperative enhancement and effectively cope with the long sentence dependency problem.

2 METHOD

Inspired by Zhong & Chen (2021), Our proposed framework for ASTE consists of the following steps: 1) Instance cooperative enhancement: enhance the input sentence with cross-sentence context information. 2) Aspect and opinion identification: predict the aspect terms and opinion terms within the input sentence. 3) Aspect-opinion pair classification: predict the relation type between given aspect-opinion pair which is highlighted by special tokens `<S:label>`, `</S:label>`, `<O:label>` and `</O:label>`. The overview of our framework can be seen in Figure 1. The

[*] Corresponding author.

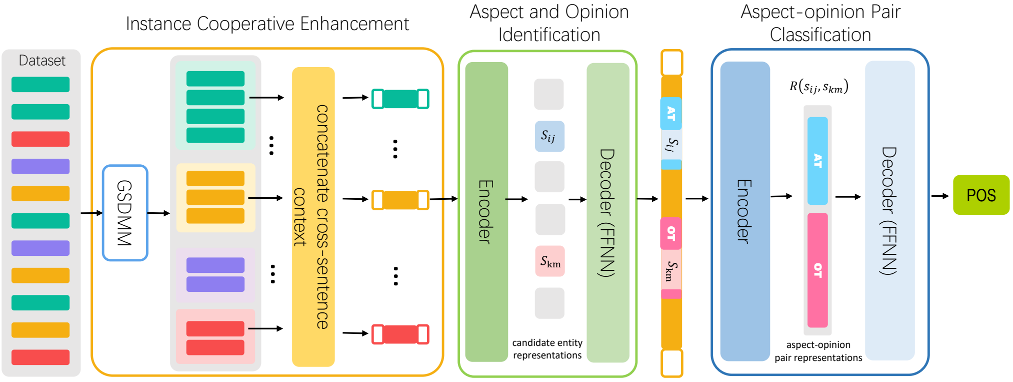
Instance Cooperative Enhancement
Dataset
GSDMM
concatenate cross-sentence context
Aspect and Opinion Identification
Encoder
s_{ij}
s_{km}
candidate entity representations
Decoder (FFNN)
AT
OT
Aspect-opinion Pair Classification
$R(s_{ij}, s_{km})$
aspect-opinion pair representations
POS

Figure 1: Overview of ICE pipeline model for aspect sentiment term extraction

model at each stage is trained independently, and each model only relies on the output features provided by the previous stage. The detailed formulation of our Two Stage ASTE model can be found in Appendix C and experimental settings can be found in Appendix D

2.1 Instance Cooperative Enhancement

We introduce instance cooperative enhancement (ICE) to incorporate global context for the input sentence. Specifically, our model fixes the length of the input sentence to L. When given a sentence with a length of l, we concatenate the sentence with $\frac{L-l}{2}$ left and right context within the same group respectively. Whereas, ASTE dataset is created on sentence-level instead of document-level which makes it difficult for us to utilize cross-sentence context. Therefore, we employ different text clustering methods to cluster similar sentences into groups. We experiment the following clustering methods respectively:

• **Simple Cluster** Since the ASTE dataset is originally extracted from restaurant and laptop reviews and ordered by these two domains. We employed a simple cluster method which clusters five adjacent sentences into one group.

• **K-means** We also employ K-means MacQueen (1967) using cosine similarity for text clustering.

• **GSDMM** The collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (GSDMM) is a short text clustering method proposed by Yin & Wang (2014). A simple analogy model and algorithm to explain the GSDMM can be found in Appendix B

3 Results and Analysis

For ASTE task, our model achieves a new state-of-the-art in all four datasets of ASTE-Data-V2. By conducting experiments on all three clustering methods, we can see from Table 1 that in most of cases, GSDMM greatly outperforms simple clustering and K-means. Comparing to K-means, GSDMM provides a more stable improvement on all four datasets. According to the algorithm , we believe the main reason is that instead of previously defining the number of clusters as K-means did, GSDMM Yin & Wang (2014) can automatically calculate the number of clustered documents with a better balance between the completeness and homogeneity of the clustering results and hence wield a more stable and significant improvement. Further explorations on ICE and performance on different sentence lengths can be found in Appendix E

Model	14lap				14res				15res				16res			
	P	R	F1	Δ	P	R	F1	Δ	P	R	F1	Δ	P	R	F1	Δ
Xu et al. (2020)+BERT	55.39	47.33	51.04		70.56	55.94	62.40		64.45	51.96	57.53		70.42	58.37	63.83	
Yan et al. (2021)+BART	58.69	56.19	58.69		65.52	64.99	65.25		59.14	59.38	59.26		66.60	68.68	67.62	
Xu et al. (2021)+BERT	63.44	55.84	59.38		72.89	70.89	71.85		62.18	64.45	63.27		69.45	71.17	70.26	
Zhang et al. (2021)+BERT	-	-	60.78		-	-	72.16		-	-	62.10		-	-	70.10	
Ours(Simple)+BERT	65.58	56.01	60.42	-	72.15	68.81	70.44	-	67.59	60.21	63.69	-	71.51	69.84	70.67	-
Ours(Kmeans)+BERT	67.41	55.82	**61.07**	+0.65	72.60	69.32	70.92	+0.48	64.68	62.68	63.66	-0.03	73.94	67.90	70.79	+0.12
Ours(GSDMM)+BERT	63.49	57.86	60.54	+0.12	72.23	72.23	**72.23**	+1.79	66.45	62.47	**64.40**	+0.71	75.91	69.26	**72.43**	+1.76

Table 1: The overall results for ASTE task.

4 Conclusion

In this work, we propose a straightforward pipeline model with instance cooperative enhancement for aspect sentiment triplet extraction. To probe deep into the importance of cross-sentence context information, we experiment on various clustering methods and conclude that GSDMM can unleash the potential of context information to the most degree. We prove that ICE can also effectively solve the long-distance dependency problem. Without using any extrinsic information such as syntax, our model reaches competitive results on ASTE-Data-V2 dataset.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. A unified model for opinion target extraction and target sentiment prediction. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 6714–6721. AAAI Press, 2019.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. A general framework for information extraction using dynamic span graphs. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 3036–3046, 2019.

J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297, 1967.

Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Mach. Learn.*, 39(2/3):103–134, 2000.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. 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)*, November 2019.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In Satinder Singh and Shaul Markovitch (eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 3316–3322. AAAI Press, 2017.

Thomas Wolf, Lysandre Debut, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv*, 2019.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 2020.

Lu Xu, Yew Ken Chia, and Lidong Bing. 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)*, pp. 4755–4766, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.367. URL `https://aclanthology.org/2021.acl-long.367`.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 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)*, pp. 2416–2429, Online, August 2021.

Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani (eds.), *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pp. 233–242. ACM, 2014.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 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)*, pp. 504–510, August 2021.

Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2021.

A ASTE Example

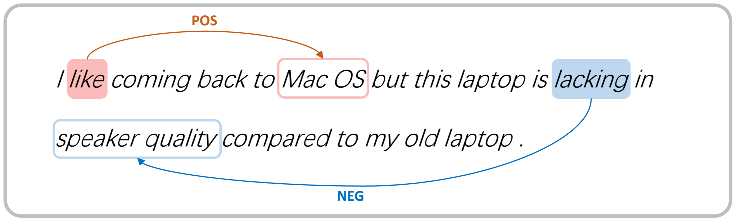
POS
I like coming back to Mac OS but this laptop is lacking in
speaker quality compared to my old laptop .
NEG

Figure 2: Examples for ASTE

B GSDMM

The authors Yin & Wang (2014) introduce a simple analogy model to explain the GSDMM called the Movie Group Process. Imagine that a professor starts a movie discussion. Before the class begins, the students make lists of their favorite films and are first randomly assigned to K tables. The professor decides to cluster the students into several groups. Students in the same group will share similar interests and vice versa. The professor then asks each student to re-join a table and the student are expected to choose a new table according to the following two rules: (1) the new table should have more student; (2) The students of the new table should watch more similar movie with him. By following these steps iteratively, the students would eventually fall into stable clusters.

The detailed algorithm is shown in Algorithm 1 where the $p(T_s = t|\mathbf{T}_{\neg}s, s)$ was derived from the Dirichlet Multinomial Mixture (DMM) model Nigam et al. (2000) and can be represented as follows:

$$p(T_s = t|\mathbf{T}_{\neg}s, s) \propto \frac{N_{d,\neg s} + \alpha}{S - 1 + K\alpha} \frac{\prod_{w \in s}\left(o^{w}_{d,\neg s} + \beta\right)}{\prod_{i=1}^{n_s}\left(n_{d,\neg s} + V\beta + i - 1\right)} \tag{1}$$

where S is the number of sentences in the dataset; n_s is the number of words in sentence s; V is the number of words in the vocabulary; α and β are hyperparameters which respectively control the probability of a sentence being clustered to an empty group and the probability of the sentence choosing a group with more similar sentences. We can see that Equation 1 follows two rules of the Movie Group Process model. Specifically, the first part follows the first rule where a sentence would be assigned to a group with more sentences and the second part of the Equation 1 satisfied the second rule where sentence will be clustered to a group which share similar words with it.

Algorithm 1 GSDMM

Annotation: S: sentences; N_d: number of sentences in group d; n_s: number of words in sentence s; o_d^w: number of occurrences of word w in group d; $\mathbf{T}$: group labels for each sentences; $\mathbf{T}_\neg s$: $\mathbf{T}$ without sentence s group label

Input: $\mathbf{S}$: input sentences

Output: $\mathbf{T} = \{T_1, T_2 ... T_n\}$; T_s is the group label of sentences s

```
for group $d \in D$ do
    initialize $N_d \leftarrow 0$,$n_d \leftarrow 0$, $o_d^w \leftarrow 0$ ;
end for
for $s \in \mathbf{S}$ do
    sample sentence $s$ a group
    $T_s \leftarrow t \sim Multinomial(1/K)$;
    update group properties
    $N_d \leftarrow N_d + 1$; $n_d \leftarrow n_d + o_s$
    for every word $w \in s$ do
        $o_d^w \leftarrow o_d^w + N_s^w$
    end for
end for
for iteration$i \in [1, I]$ do
    for every sentence $s \in \mathbf{S}$ do
        record the current group of sentence $s$: $t \leftarrow t_s$
        remove $s$ from group $d$ and update the group properties:
        $N_d \leftarrow N_d - 1$; $n_d \leftarrow n_d - o_s$
        for every word $w \in s$ do
            $o_d^w \leftarrow o_d^w - N_s^w$
        end for
        sample a new group for $s$
        $T_s \leftarrow t \sim p(T_s = t|\mathbf{T}_\neg s, s)$
        $N_d \leftarrow N_d + 1$; $n_d \leftarrow n_d + o_s$
        for every word $w \in s$ do
            $o_d^w \leftarrow o_d^w + N_s^w$
        end for
    end for
end for
```

C Two Stage ASTE model

Following Zhong & Chen (2021), our model consists of two stage (1) Aspect and opinion identification; (2) Aspect-opinion pair classification.

C.0.1 Aspect and Opinion Identification

In the first stage, our model aims at predicting the aspect terms and opinion terms of the input sentence $X = \{x_1, x_2, ..., x_n\}$. We first build contextualized representations $H = \{h_1, h_2, ..., h_n\}$ for the input sentence using pretrained language model. For every candidate span with length less than L, we construct its span representation:

$$s_{ij} = [h_i; h_j; \phi(i,j)] \tag{2}$$

where $\phi(i,j)$ is the distance feature vector which is used to encode the span width. We then use a feedforward neural network as a decoder to classify each span. The probability distribution of span types is calculated by:

$$p(y|s_{ij}) = softmax(\mathbf{FFNN}(s_{ij})) \tag{3}$$

where $y \in Y = \{AT, OT, \epsilon\}$

C.0.2 Aspect-opinion Pair Classification

In the pair classification model, we first insert special tokens to highlight aspect and opinion terms predicted by the first stage. For instance, for spans s_{ij} and s_{km} with the labels of y_i and $y_k \in Y$, our highlighted input sentence X can be denoted as:

$$\begin{aligned} X =&\{x_1, ..., < s : y_i >, x_i, ..., x_j, < /s : y_i >, ..., \\ &< o : y_k >, x_k, ..., x_m, < /s : y_k >, ..., x_n\} \end{aligned} \tag{4}$$

We then use another encoder to build contextualized representations H for highlighted input X.

$$H =\{..., h^s_i, h_i, ..., h_j, h^e_i, ..., h^s_k, h_k, ..., h_m, h^e_k, ...\} \tag{5}$$

The relation pair representation of span s_{ij} and s_{km} is calculated by :

$$R(s_{ij}, s_{km}) = [h^s_i, h^s_k] \tag{6}$$

Similar to stage 1, we also use a feedforward neural network to calculate the probability distribution of relation types.

$$p(y_{rel}|(s_{ij}, s_{km}) = softmax(\mathbf{FFNN}(R(s_{ij}, s_{km}))) \tag{7}$$

where $y_{rel} \in Y_{rel} = \{POS, NEG, NEU, \epsilon\}$

D Experiments

We experiment our model on ASTE-Data-V2 dataset Xu et al. (2020). Our models are implemented based on HuggingFace's Transformers library Wolf et al. (2019). For all stages in our pipelined model, we use the pre-trained language model bert-based-uncased Devlin et al. (2018) as the base encoders, and all two models are trained for 100 epochs, and with a learning rate of 1e-5, a batch size of 16, and a dropout probability of 0.1 for the classifier. We fix the length of the input sentence to $L = 80$ using instance cooperative enhancement and for GSDMM, we set $K = 121, \alpha = 0.1, \beta = 0.1$ and train the model for 100 iterations. We evaluate our model with all three cluster methods.

E Further Exploration

E.1 Instance Cooperative Enhancement or Not?

	14lap			
	P	R	F1	Delta
Ours -wo ICE(L=0)	58.02	57.49	57.75	-
-w ICE(L=70)	63.06	57.12	59.94	+2.19
-w ICE(L=80)	65.58	56.01	**60.42**	+2.67
-w ICE(L=90)	65.49	55.08	59.84	+2.09
-w ICE(L=100)	66.59	53.79	59.51	+1.76
-w ICE(L=200)	63.75	55.27	59.21	+1.46

Table 2: Comparison between with (-w) or without (-wo) ICE

To prove that the instance cooperative enhancement (ICE) efficiently improves the overall performance, we use original sentences as input without concatenating the left and right context within the same document. To further probe into the influence ICE exerts on the results, we experiment the ICE with different input length L. The results are shown in Table 2. We can see that ICE is of great importance for improvement of the model performance. Models with ICE greatly outperforms those without ICE by a large margin of 2.67 when L is set to 80. We find that the F1 score reaches the peak value when $L = 80$ and do not further increases when the input length grows. We believe that adding cross-sentence information on one hand can help our model capture deeper relations between entities, but on the other hand, too much cross-sentence information enhancement also introduces noises and hence hurts the overall performance. Therefore, we use $L = 80$ for our ASTE experiments.

E.2 Performance On Different Sentence Lengths

Since we employ ICE to incorporate richer context information, we speculate that the most performance improvement from our model may attribute to its solved long-distance dependency problem.

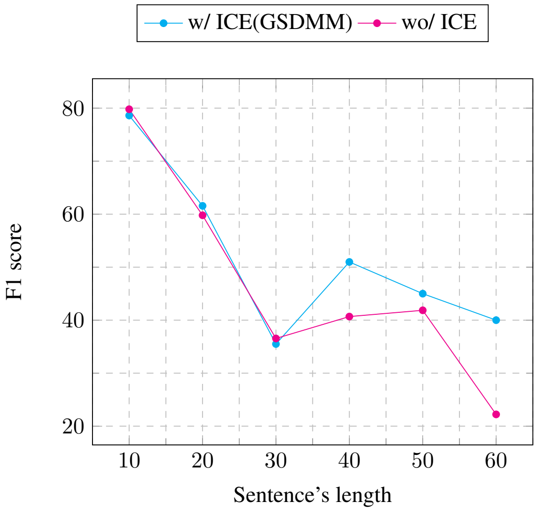

Figure 3: Performance of different length ranges in 14lap dataset.

To verify our hypothesis, we deprive our model of ICE and compare it with our full model. We perform their statistics on the 14lap dataset for different length ranges, and the results are shown in Figure 3.

According to the curves in the figure, it is obvious that the deprived model and the full model shares similar results in short sentences (< 30), while when not using ICE, the model suffers a significant decrease as the sentence length grows. In contrast, our model has the smallest decrease and obtains the best results, thus indicating that our method can effectively cope with the long sentence dependency problem.