SAMBERT: Improve Aspect Sentiment Triplet Extraction by Segmenting the Attention Maps of BERT

Anonymous ACL submission

Abstract

Aspect Sentiment Triplet Extraction (ASTE) performs fine-grained sentiment analysis in a unified way through extracting sentiment triplets comprised of aspect terms, opinion spans, and their sentiment relations in sentences. The previous works show the adoption of BERT, which simply leverages its *last layer* output as the word representation, is beneficial for recognizing triplet elements. However, their methods limit the potential of pretrained knowledge in BERT, since the *different layers* can capture multi-level linguistic information existing in sentences, which are useful for ASTE as well. In this work, we explore to access the rich pretrained knowledge by fully leveraging its *attention maps* of different layers. To this end, we propose to Segment the Attention Maps of BERT (***SAMBERT***) by taking the merits of semantic segmentation, which can effectively discriminate the desired objects from others in an image. In this procedure, we can further reason over the knowledge of different levels in these attention maps to distinguish aspect terms, opinion spans and their sentiment relations from other parts, which results in a same-shape tagging matrix of word pairs for deriving sentiment triplets. Through the extensive experiments on four benchmarks, we demonstrate our method can achieve a new state of the art.

1 Introduction

Sentiment analysis (Liu, 2012; Feldman, 2013) is an important Natural Language Understanding task (NLU) to identify the sentiment from review sentences, which has been widely studied in many fields, e.g., E-commerce (Shivaprasad and Shetty, 2017) and social media (Agarwal et al., 2011). Recently, Aspect-based Sentiment Analysis (Pontiki et al., 2014; Ma et al., 2017) tries to perform sentiment analysis at the fine-grained level. It comprises several subtasks, such as Aspect Term Extraction (Li et al., 2018; Xue et al., 2017), Aspect Opinion Extraction (Fan et al., 2019; Pereg et al., 2020), and Aspect Sentiment Classification (Wang et al., 2016; Ruder et al., 2016). In order to provide a unified solution for these subtasks, Aspect Sentiment Triplet Extraction (ASTE) is proposed by (Peng et al., 2020) to extract sentiment triplets from review sentences, which contain all of the aspect terms, corresponding opinion spans, and sentiment polarities. For instance, given a review "*The barbecued salmon is elegantly spiced and not dry at all .*", the triplets of [*barbecued salmon*, *elegantly spiced*, *positive*] and [*barbecued salmon*, *not dry at all*, *positive*] should be extracted from this sentence.

To recognize the aspect term, opinion span and their sentiment relation, many efforts are devoted. (Peng et al., 2020; Xu et al., 2021) conduct ASTE in multiple stages, which firstly extract aspect terms and opinion spans, and then combine the valid pairs of them to decide their sentiment polarities. (Xu et al., 2020; Wu et al., 2020) jointly extract the triplet elements with their proposed unified tagging schemes in an end-to-end manner. Furthermore, they demonstrate that the adoption of BERT is beneficial for improving the performance of ASTE, by leveraging the contextual output of the last layer in BERT (Devlin et al., 2019) as their word embeddings, which follows the same strategy recommended by (Sun et al., 2019).

However, the way they use BERT may be not optimal since they ignore the rich pretrained knowledge existing in BERT. As the existing works (Jawahar et al., 2019; Clark et al., 2019) analyze, the different layers of BERT, which are comprised of multiple attention heads, can capture multi-level and multi-view knowledge existing in sentences. For example, (Jawahar et al., 2019) shows that the bottom layers of BERT focus more on phrase-level information (e.g., the opinion span "*not dry at all*" in Fig. 1), and the top layers mainly capture semantic features (e.g., the sentiment relation between "*barbecued salmon*" and "*elegantly spiced*"). *Therefore, we argue that the multi-level information can*

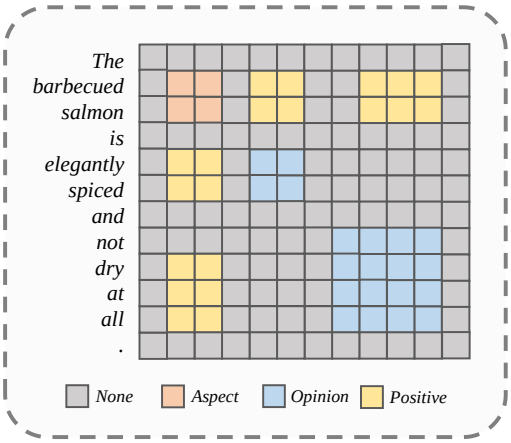

Figure 1: A review sentence of ASTE with its word pair form, which has the same shape of the attention maps. (1) The colored blocks in the diagonal correspond to the intra-associaitions of aspect terms and opinion spans. (2) The colored blocks in the non-diagonal are the sentiment relations (positive) between them. (3) The gray ones are undesired parts.

contain task-specific features, where making full and explicit use of them can be beneficial for recognizing triplet elements.

To achieve this, instead of only utilizing the final generated word representation of BERT (Sun et al., 2019), we explore to fully leverage its pretrained knowledge *by explicitly accessing the attention maps of different layers.* By doing so, we can benefit from three sides: (1) For a single word pair, the attention scores between them can capture its hierarchical and diverse relations via different layers as well as different attention heads (Jawahar et al., 2019; Clark et al., 2019), which can be treated as our input feature to represent the connections at a word pair level. (2) For all the word pairs, their attention scores comprise a 2-D feature map (cf. the example in Fig. 1). In this feature map, the desired word pairs of the aspect terms and opinion spans form several blocks (i.e., the colored ones) scattered in other word pairs (i.e., the gray ones). (3) By using these word pair representations storing the hierarchical knowledge, we wish these desired word pairs can be distinguished from other parts for decoding the sentiment triplets, *which is inspired by the semantic segmentation task (Girshick et al., 2014) via a multi-scale context-aware pyramid structure (Zhao et al., 2017) to distinguish the object and background in an image.*

To this end, we propose ***SAMBERT***, which Segments the Attention Maps of BERT in different layers as a word pair level tagging matrix[1]. In this procedure, we can further reason over these attention maps and learn to distinguish the multi-level task-specific patterns of the aspect terms and opinion spans, as well as the sentiment relations between them from other parts (cf. Fig 1). Specifically, we first stack all the attention maps of the k layers in the bottom and top of BERT as our input feature maps, where k is a hyper-parameter used to control the number of layers we selected.[2] Then, a Convolutional Encoder-Decoder is leveraged to conduct the segmentation stage to model the task-specific patterns, and further strengthen these associations between the triplet elements in a sentence. Finally, a 2-D tagging matrix is predicted to decode sentiment triplets.

We conduct extensive experiments on four benchmarks (Peng et al., 2020; Xu et al., 2020), where our method can achieve a new state of the art. Also, further analysis verifies that with the segmentation paradigm, the task-specific features can be effectively distilled out from the attention maps of different layers for ASTE.

To summarize, our contributions are as follows:

- We are the first to explicitly leverage the attention maps of different layers in BERT to fully access the pretrained knowledge for ASTE.
- We formulate ASTE into a semantic segmentation paradigm, to further learn the task-specific patterns in these attention maps.
- The experimental results on four public benchmarks show that our method can achieve a new state of the art.

2 Related Works

Sentiment Analysis (Liu, 2012; Feldman, 2013) aims to automatically classify the sentiment polarity of a sentence (Maas et al., 2011; Yang and Cardie, 2014; Dai et al., 2020; Pontiki et al., 2014; Ma et al., 2017). While the sentence level sentiment analysis has been well studied, current literature of Sentiment Analysis tends to analyze the sentiment at a fine-grained level, i.e., analyzing the sentiment

[1]Thanks to the Grid Tagging Scheme (Wu et al., 2020; Chen et al., 2021b), a word pair level tagging matrix (cf. the right of Fig. 2), we can directly equip our method with it since they are naturally compatible due to the same shape of the attention maps of BERT and tagging matrix.

[2]We prefer not to use the intermediate layers of BERT since we find it is less helpful for ASTE according to our pre-experiment.

polarities of aspect terms with the specific opinion spans. In particular, Aspect-based Sentiment Analysis (ABSA) is divided into several subtasks like Aspect Term Extraction (Li et al., 2018; Xue et al., 2017), Aspect Opinion Extraction (Fan et al., 2019; Pereg et al., 2020), Aspect Sentiment Classification (Wang et al., 2016; Ruder et al., 2016) and Opinion Pair Extraction (Wang et al., 2017; Dai and Song, 2019; Wu et al., 2020). However, these subtasks only derive one or two elements of the aspect term, opinion span and sentiment polarity. To extract them all, Aspect Sentiment Triplet Extraction is proposed by (Peng et al., 2020) to generate triplets of all the elements.

The existing works of ASTE can be roughly divided into two categories, i.e., the multi-stage and one-stage methods. For the multi-stage method, (Peng et al., 2020) proposes to extract the elements at first, which will be combined into sentiment triplets later. (Chen et al., 2021a; Mao et al., 2021) transform ASTE task into a Machine Reading Comprehension (MRC) task to capture the connections among the subtasks of ASTE. (Huang et al., 2021) proposes a two-stage method to enhance the correlations between aspects and opinions. (Jian et al., 2021) proposes to regard the aspect and opinion terms as arguments of the expressed sentiment in a hierarchical reinforcement learning framework. (Xu et al., 2021) uses a span-level approach to explicitly consider the interactions between the whole spans of aspects and opinions when predicting their sentiment relations. However, these multi-stage methods can lead to error propagation.

To address this problem, the one-stage method is proposed: (Xu et al., 2020; Wu et al., 2020; Chen et al., 2021b) extract sentiment triplets in one stage by their proposed unified tagging schemes. (Xu et al., 2020) uses a word-level tagging scheme, but it is derived from the assumption that one aspect term corresponds to only one opinion span, which can not be always held in all possible scenarios. (Wu et al., 2020; Chen et al., 2021b) avoid this problem by a word pair level tagging scheme, which results in a 2-D tagging matrix. Besides, (Zhang et al., 2021b; Yan et al., 2021) both propose to extract the sentiment triplets via a generative way, where the sequence-to-sequence paradigm is used. Nevertheless, the exposure bias of the generative framework (Ranzato et al., 2016) can lead to a gap between training and inference.

Besides, (Liu et al., 2020; Zhang et al., 2021a) also formulate Incomplete Utterance Rewriting and Document-level Relation Extraction tasks as a semantic segmentation task. Nevertheless, the difference between our work and them (as well as all the aforementioned works) is that they only leverage the contextual representation of the last layer of Pretrained Models (Devlin et al., 2019; Lan et al., 2020; Liu et al., 2019) to further enhance the performance but ignore the rich hierarchical knowledge hidden in its different layers (Jawahar et al., 2019; Clark et al., 2019). In contrast, we can fully leverage the knowledge by explicitly unitizing the attention maps of different layers in Pretrained Models storing diverse associations between word pairs, which is analogous to the multi-scale context-aware pyramid structure (Zhao et al., 2017) used in Computer Vision.

3 Methodology

In this Section, we first describe the overall workflow of our method (Sec. 3.1). Then, we elaborate on each component, i.e., Review Encoder (Sec. 3.2), Segmentation Layer (Sec. 3.3), and Triplet Decoding procedure (Sec. 3.4).

3.1 Overall Workflow

As shown in Fig. 2, in the review encoding stage, we encode the review sentence with BERT (Devlin et al., 2019) to derive the attention maps of its bottom and top layers. Then, these attention features are stacked as an 2-D feature map, which is used to conduct the segmentation stage to reason over the task-specific patterns. Finally, a tagging matrix is predicted to decode the sentiment triplets. By this formulation, we can better reason over and refine the linguistic knowledge of different levels stored in the attention maps of BERT for ASTE.

The final segmentation classes (i.e., the tagging scheme) are inherited from (Wu et al., 2020; Chen et al., 2021b), i.e., $\{N, A, O, Pos, Neu, Neg\}$, where N means no association exists between a word pair; A means a word pair belongs to the same aspect term; O means a word pair belongs to the same opinion span; $\{Pos, Neu, Neg\}$ mean positive, neutral, and negative sentiment relations are expressed between a word pair. In addition, since the 2-D tagging matrix is symmetric, only the tags of the upper triangle part in the matrix are used for training and inference. Please refer to the tagging matrix in the right of Fig. 2 for better understanding.

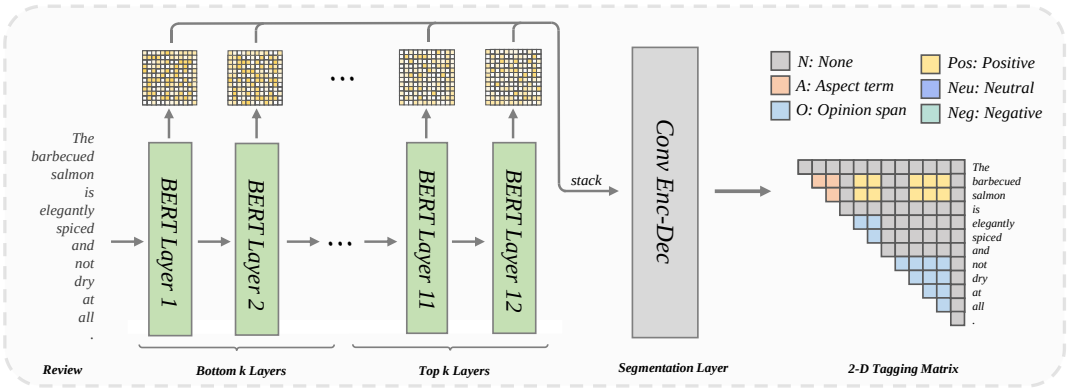

Figure 2: Overview of *SAMBERT*. The attention maps of the k bottom and top layers of BERT are stacked as a 2-D feature map. Then, a Convolutional Encoder-Decoder is leveraged to segment this feature and derive the Tagging Matrix (Wu et al., 2020) for decoding the final sentiment triplets, where the colored blocks represent the aspect terms, opinion spans and their sentiment relations. *Besides, only the upper triangle part in the tagging matrix are used due to the symmetry of word pairs.*

3.2 Review Encoder

As (Clark et al., 2019; Jawahar et al., 2019) discuss, the Self Attention mechanism (Vaswani et al., 2017) in different layers of Pretrained Models (Devlin et al., 2019; Lan et al., 2020; Liu et al., 2019) focus on different levels of the linguistic knowledge. Therefore, our Review Encoder aims to fully access and explicitly leverage these rich linguistic features when the review sentence is encoded.

To this end, this work explores to *directly utilize the attention maps in different layers* of Pretrained Models. Here we choose BERT as the representative of them, in order to align with the previous works. Specifically, given one review sentence $\mathcal{S} = [w_1, w_2, ..., w_n]$, we first obtain its input embedding sequence. That is, $H_0 = [e_1, e_2, ..., e_n]$ ($e_i = w_i + p_i$), where w_i and p_i are the word embedding and position embedding of the i-th word. Then, the input embedding sequence is feed into BERT to obtain its attention maps:

$$H_i, A_i = BERT\_Layer_i(H_{i-1}), i \in [1, N],$$
$$A = [A_1; ...; A_k; A_{N-k}...; A_N],$$

where $A_i \in \mathbb{R}^{h \times n \times n}$ is the derived h head attention maps of i-th layers. Here we stack both bottom and top k layers as our feature map $A \in \mathbb{R}^{(2*k*h) \times n \times n}$. Please note that here we prefer not to use the intermediate layers of BERT, since we find it less helpful according to our pre-experiment. A reasonable explanation is that the information in the bottom and top layers is enough for ASTE. Hence, we use a hyper-parameter k to select the BERT layers we used, which is a simple way to avoid irrelevant information and we leave other advanced selection methods (e.g., the attention mechanism) for future work.

Note that A has a shape of $n \times n$. Each A_{ij} represents the attention scores of different layers that contains diverse associations between the i-th and j-th words. Therefore, the word pair representations within/between aspect terms and opinion spans (v.s. the objects in an image) can provide task-specific features and be distinguished from other parts of the feature map (v.s. the background in an image) in the downstream segmentation stage.

We argue that this is a more effective way to leverage the pretrained knowledge in BERT, since the attention maps of different layers originally store the multi-level and multi-view knowledge via the pretraining paradigm (Jawahar et al., 2019; Clark et al., 2019). In contrast, all the existing works only use the word representations of last layer, which can result in losing task-specific features for ASTE.

3.3 Segmentation Layer

After obtaining the 2-D feature map A, it needs to be mapped into a same-shape tagging matrix $M \in \mathbb{R}^{n \times n}$ (cf. the right of Fig. 2), where each A_{ij} is mapped to a predefined tag $\in \{N, A, O, Pos, Neu, Neg\}$.

In this procedure, it should not only consider the information in the word pair representation itself

Dataset	Res14				Lap14				Res15				Res16			
	#Sent.	#Pos.	#Neu.	#Neg.	#Sent.	#Pos.	#Neu.	#Neg.	#Sent.	#Pos.	#Neu.	#Neg.	#Sent.	#Pos.	#Neu.	#Neg.
Train	1266	1692	166	480	906	817	126	517	605	783	25	205	857	1015	50	329
Dev	310	404	54	119	219	169	36	141	148	185	11	53	210	252	11	76
Test	492	773	66	155	328	364	63	116	322	317	25	143	326	407	29	78

Table 1: The detailed statistics of ASTE-Data-V2, where #Sent. denotes the number of sentences, and #Pos., #Neu., and #Neg. denote the numbers of the positive, neutral, and negative triplets in each dataset.

(i.e., the attention scores between this word pair), where the information contained in its adjacent word pairs is also useful for prediction, since the word pairs belonging to the same triplet element are consistent to share the similar information of the same tags (cf. the right of Fig. 2). An inductive example of that is, only using the information of one pixel in an image is hard to tell what the object is, while looking at a larger region is much easier.

Inspired by the semantic segmentation task (Girshick et al., 2014) in Computer Vision, which aims to distinguish desired objects from others in images, we also formulate this stage as a *segmentation* task, which refines the attention maps as a 2-D tagging matrix M to discriminate the desired triplet elements.

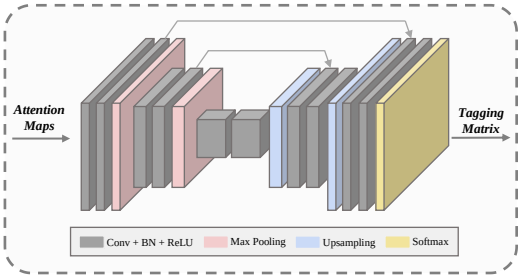

Figure 3: The backbone we used for the segmentation of stacked attention maps in BERT.

Specifically, we use a Convolutional Encoder-Decoder framework (Ronneberger et al., 2015) to perform this stage. As shown in Fig. 3, in the encoding stage, the attention maps are encoded by two convolutional layers with the kernel size of 3×3, while the channels of input are doubled and its size is halved by the max-pooling operation. Then, in the decoding stage, twice up-sampling operations are leveraged to cooperate with two convolutional layers (3×3 kernels) to make the size of output consistent with the feature map, where the channels of input are halved. Besides, the residual connection between the encoding layer and decoding layer is also used for better training. Furthermore, the final probability matrix of the tagging classes is obtained by a fully-connected layer.

3.4 Triplet Decoding

After the segmentation stage, the derived tagging matrix is used to decode sentiment triplets. We use the same decoding strategy as (Wu et al., 2020; Chen et al., 2021b) did: (1) we first search the elements of word pairs on the main diagonal of the tagging matrix, where the continuous tags of $\{A, O\}$ are recognized as an aspect term or opinion span. (2) Then, we count the tags $\{Pos, Neu, Neg\}$ of the corresponding word pairs between the recognized aspect terms and opinion spans, the most predicted tag is assigned as the sentiment polarity to this triplet. If these tags do not belong to $\{Pos, Neu, Neg\}$, then this triplet is dropped.

4 Experimental Setup

4.1 Datasets

There are two versions of datasets for ASTE: one (named ASTE-Data-V1) is released by (Peng et al., 2020) and another (named ASTE-Data-V2) is released by (Xu et al., 2020). ASTE-Data-V1 does not contain cases where one opinion span is associated with multiple targets, but these cases are very common in the real world. ASTE-Data-V2 refines the V1 version with these additional missing triplets. Therefore, we mainly use ASTE-Data-V2 for our experiments, which is more general. Note that some works (Mao et al., 2021; Chen et al., 2021a) use ASTE-Data-V1 for the experiments. We also report the results of our method on ASTE-Data-V1 to fairly compare with them. The detailed statistics of ASTE-Data-V2 are listed in Tab. 1.

4.2 Implementation Details

The hyper-parameters in our experiment are tuned over the development set by grid search. We use *bert-base-uncased* as our Review Encoder to be consistent with the previous works of ASTE. The learning rate of BERT is set to $5e-5$ with gradient clip selected from 1 to 5. The learning rate of the

Models	Res14			Lap14			Res15			Res16		
	P.	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*
ASTE-Data-V2 + Static Word Embeddings (GloVe)												
Peng-two-stage	43.24	63.66	51.46	37.38	50.38	42.87	48.07	57.51	52.32	46.96	64.24	54.21
OTE-MTL	62.70	57.10	59.71	49.62	41.07	44.78	55.63	42.51	47.94	60.95	53.35	56.82
JET^o	61.50	55.13	58.14	53.03	33.89	41.35	64.37	44.33	52.50	70.94	57.00	63.21
GTS	66.13	57.91	61.73	53.35	40.99	46.31	60.10	46.89	52.66	63.28	58.56	60.79
Span-ASTE	72.52	62.43	67.08	59.85	45.67	51.80	64.29	52.12	57.56	67.25	61.75	64.37
ASTE-Data-V2 + Pretrained Model (BERT)												
JET^o	70.56	55.94	62.40	55.39	47.33	51.04	64.45	51.96	57.53	70.42	58.37	63.83
GTS	67.76	67.29	67.50	57.82	51.32	54.36	62.59	57.94	60.15	66.08	69.91	67.93
Span-ASTE	72.89	70.89	71.85	63.44	55.84	59.38	62.18	64.45	63.27	69.45	71.17	70.26
SAMBERT	70.29	74.92	**72.53**	62.26	59.15	**60.66**	65.12	63.51	**64.30**	68.01	75.44	**71.53**

Table 2: The overall evaluation results on ASTE-Data-V2. *P.* and *R.* are Precision and Recall respectively. The best results are in **bold font** and the second-best ones are underlined. The results of OTE-MTL and GTS are adopted from (Xu et al., 2021).

Convolutional Encoder-Decoder is selected from the range of [5e-4, 1e-4, 5e-5]. The Adam optimizer (Kingma and Ba, 2015) is used for model optimization. Besides, since the information of some layers in BERT can be irrelevant to ASTE, we only leverage the first k bottom layers and last k top layers as input, where we set k to 4 in all the experiments. Our implementation is based on PyTorch (Paszke et al., 2019) and HuggingFace's transformers library (Wolf et al., 2020).

4.3 Evaluation Metrics

Following the existing works (Peng et al., 2020; Xu et al., 2020; Wu et al., 2020; Chen et al., 2021a), we use precision, recall, and F1 score as the metrics to evaluate the performance of ASTE. A correct triplet requires an exact match between the prediction of the aspect term, opinion span, and the sentiment polarity with the ground truth. Note that the F1 score takes into account both precision and recall, which can be regarded as a harmonic average of them. Therefore, we focus on the F1 score in following experiments.

4.4 Baselines

Our method is compare to the following methods.

- Peng-two-stage: (Peng et al., 2020) extracts the sentiment triplets in two stages, which first extract the elements and then combine them into sentiment triplets.
- OTE-MTL: (Zhang et al., 2020) proposes a multi-task learning framework to jointly extract aspect terms and opinion spans with parsing the sentiment polarities between them simultaneously.
- JET: (Xu et al., 2020) proposes to extract the sentiment triplets by a word-level position-aware tagging scheme.
- GTS: (Wu et al., 2020) uses a grid tagging scheme and an inference strategy for extracting the sentiment triplets.
- Span-ASTE: (Xu et al., 2021) proposes a span-level approach to explicitly consider the interaction between the whole span of the aspect and opinion when predicting their sentiment.
- Dual-MRC: (Mao et al., 2021) proposes a dual-MRC framework to handle ASTE task, by jointly training two BERT-MRC models with parameter sharing.
- BMRC: (Chen et al., 2021a) proposes a bidirectional MRC framework to capture and utilize the associations among ASTE subtasks.

5 Results

5.1 Overall Evaluation

As reported in Tab. 2 and Tab. 3, our method outperforms all of the existing state-of-the-art methods. Specifically, for ASTE-Data-V2 (cf. Tab. 2), our method surpasses all of the non-Bert-based and Bert-based methods. Compared to the multi-stage method *Span-ASTE* (Xu et al., 2021), we can averagely outperform it by 1.07% on the four datasets. Besides, compared to the one-stage method *GTS* (Wu et al., 2020), our method can also boost the performance by 4.77 points on average.

In addition, to keep consistency and fairly compare with (Mao et al., 2021; Chen et al., 2021a),

Models	Res14			Lap14			Res15			Res16		
	P.	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*
ASTE-Data-V1 + Pretrained Model (BERT)												
Dual-MRC	71.55	69.14	70.32	57.39	53.88	55.58	63.78	51.87	57.21	68.60	66.24	67.40
BMRC	71.32	70.09	70.69	65.12	54.41	59.27	63.71	58.63	61.05	67.74	68.56	68.13
SAMBERT	75.15	72.97	**74.04**	63.03	57.14	**59.96**	61.97	60.88	**61.42**	68.12	73.98	**70.93**

Table 3: The overall evaluation results on ASTE-Data-V1.

Model	Res14			Lap14			Res15			Res16		
	P.	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*
SAMBERT	70.29	74.92	**72.53**	62.26	59.15	**60.66**	65.12	63.51	**64.30**	68.01	75.44	**71.53**
only top layers	68.48	73.00	70.66	59.57	51.76	55.39	57.93	61.03	59.44	64.78	73.49	68.86
only bottom layers	69.53	59.09	63.89	46.17	42.33	44.17	58.82	47.42	52.51	60.81	58.67	59.72
MHSA over word rep	65.07	71.47	68.12	56.23	56.75	56.49	56.66	58.76	57.69	61.73	72.32	66.61
word rep concat	69.80	71.57	70.68	58.49	57.30	57.89	63.31	60.83	62.04	56.08	69.40	67.17
w/o segmentation	56.75	46.50	51.11	49.51	28.10	35.85	42.27	40.00	41.10	56.08	46.78	51.01

Table 4: The ablation study of our method. These experiments are based on ASTE-Data-V2.

we also report the results of our method on ASTE-Data-V1, which are shown in Tab. 3. It's observed that, although these two methods use BERT (Devlin et al., 2019) in a MRC way to leverage its capability of deep language understanding, our method can also improve the performance by 1.85% on average.

The results on the two versions of datasets demonstrate that our method, which makes full and explicit use of the pretrained knowledge in BERT and further equip it with a segmentation paradigm, is more effective to tackle ASTE and can achieve a new state of the art.

5.2 Effects of Different Components

In this Subsection, we discuss the effects of the two crucial components of our method. For the attentions maps of BERT, we replace them with two variants:

- *only top layers*: the attention maps of bottom layers are removed to prove the information it contains can help with the improvement.
- *only bottom layers*: only the attention maps of the bottom BERT layers are leveraged to show the information in top layers is necessary.

Also, other two operations are used to calculate the feature maps by the word representations of last layer in BERT, which aim to verify the superiority of leveraging the attention features originally derived by BERT:

- *MHSA over word rep*: we replace the attention maps derived within BERT with the post-calculated attention maps over its word representations.
- *word rep concat*: we concatenate the pairs between the word representations (Wu et al., 2020) to substitute our attention maps of different layers.

Besides, the Segmentation Layer is replaced with a vanilla fully-connected classifier to verify the effectiveness of the semantic segmentation, dubbed as *w/o segmentation*. The experiments in this Subsection are all based on ASTE-Data-V2.

As reported in Tab. 4, when only leveraging the attention maps of top layers or bottom layers, the F1 score drops by 3.67% or 12.18% on average. That means the top layers contain most of the required information for ASTE, but the bottom layers can also provide some task-specific features. In other words, only using the information stored in top layers can maintain a high-level performance of ASTE, but the knowledge in bottom layers can also supplement the effective information to further boost the performance.

When replacing the derived attention maps of BERT with post-calculated attention maps[3] over the output of word representations of BERT, the performance averagely declines by 5.03 points. That indicates the knowledge originally stored in different layers via the pretraining paradigm can not be easily obtained by its last layer output. In contrast, directly and explicitly leveraging these attention

[3]Here we use a 12-head Self Attention to calculate the attention maps.

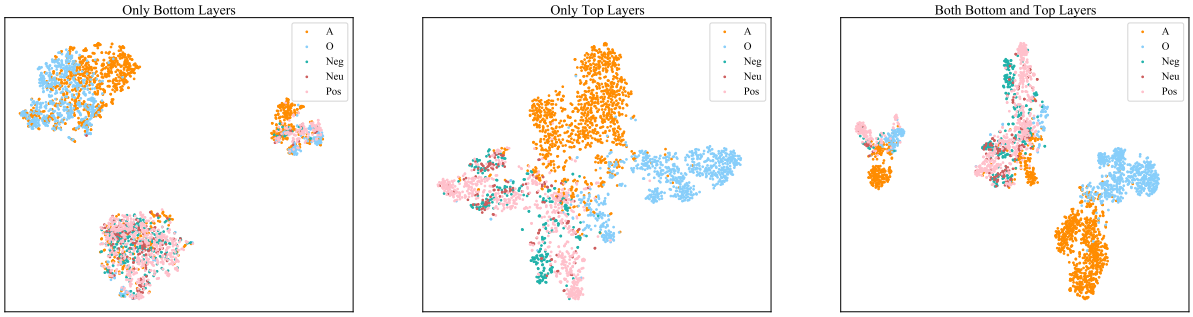

Figure 4: The t-SNE visualization of the word pair representations consisting of only bottom layers (left), only top layers (middle), and both bottom and top layers (right). We only visualize these representations whose classes belong to $\{A, O, Pos, Neu, Neg\}$ and the label $None$ is omitted since its number is enormous and can interfere with observation.

maps derived by BERT is a better way to access the hierarchical features. In addition, we follow (Wu et al., 2020) to concatenate the pairs of contextualized representations of BERT as the input to Segmentation Layer, where the performance also drops by 2.81 points. The conclusion is the same as the one we draw above.

Finally, we verify the effectiveness of the Segmentation Layer. When the variant uses a vanilla classifier instead of the Segmentation Layer, the performance of the F1 score dramatically degrades by 22.49 points on average. That indicates the semantic segmentation paradigm is important to perceive and utilize the information existing in other adjacent word pairs and further boost the performance of ASTE.

The ablation studies from different perspectives imply that both components are useful to improve the performance of ASTE, and the combination of them can further reach their full potential.

5.3 Visualization

Besides, to demonstrate the knowledge stored in the attention maps of different layers is beneficial to obtain informative and discriminative representations for ASTE, we also apply t-SNE (van der Maaten and Hinton, 2008) to these word pair level representations comprised of attention scores, and plot their 2-dimensional vectors.[4]

Specifically, we visualize three types of attention features, i.e., only bottom layers, only top layers, and both bottom and top layers. As shown in Fig. 4, it is obvious that (1) Only using the bottom layers can easily tell the difference between intra-associations (i.e., $\{A, O\}$) and inter-relations (i.e., $\{Pos, Neu, Neg\}$) of the aspect terms and opinion spans with a large margin. That indicates the bottom layers do capture some task-specific information existing in sentiment triplets. (2) Although only using the top layers can better recognize both intra- and inter-associations of the triplet elements, the clusters are less compact than only using the bottom layers. (3) When both the bottom and top layers are used, the representations of word pairs can result in more compact clusters and clearer boundaries between different classes than only using the bottom or top layers. That suggests the features in bottom and top layers are complementary to each other, which are helpful to decide the classes the word pairs belong to. Without any part of them can result in the situation of losing task-specific information.

[4]Due to the space limitation, we only illustrate Lap14 dataset, where other datasets have the same performance.

6 Conclusion

In this work, we propose a novel framework, i.e., *SAMBERT*, to Segment the Attention Maps of BERT, which aims to fully and explicitly leverage the rich pretrained knowledge stored in its different layers. By formulating ASTE as a semantic segmentation task, we can further reason over the knowledge of different levels and views in these attention maps, so as to distinguish aspect terms, opinion spans and their sentiment relations from other parts. That results in a same-shape tagging matrix of word pairs, which is used to derive the sentiment triplets of review sentences. Through the experiments on four public benchmarks, we demonstrate that our method can achieve a new state of the art. The further analyses in both quantitative and qualitative perspectives verify the effectiveness of the proposed components of our method.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, page 30–38, USA. Association for Computational Linguistics.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021a. Bidirectional machine reading comprehension for aspect sentiment triplet extraction.

Zhexue Chen, Hong Huang, Bang Liu, Xuanhua Shi, and Hai Jin. 2021b. Semantic and syntactic enhanced aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 1474–1483, Online. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Hongliang Dai and Yangqiu Song. 2019. Neural aspect and opinion term extraction with mined rules as weak supervision. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5268–5277, Florence, Italy. Association for Computational Linguistics.

Yong Dai, Jian Liu, Xiancong Ren, and Zenglin Xu. 2020. Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7618–7625.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2509–2518, Minneapolis, Minnesota. Association for Computational Linguistics.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 580–587, USA. IEEE Computer Society.

Lianzhe Huang, Peiyi Wang, Sujian Li, Tianyu Liu, Xiaodong Zhang, Zhicong Cheng, Dawei Yin, and Houfeng Wang. 2021. First target and opinion then polarity: Enhancing target-opinion correlation for aspect sentiment triplet extraction.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Samson Yu Bai Jian, Tapas Nayak, Navonil Majumder, and Soujanya Poria. 2021. Aspect sentiment triplet extraction using reinforcement learning. *arXiv preprint arXiv:2108.06107*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of the International Conference on Learning Representations*.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4194–4200. AAAI Press.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020. Incomplete utterance rewriting as semantic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4068–4074.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.

Oren Pereg, Daniel Korat, and Moshe Wasserblat. 2020. Syntactically aware cross-domain aspect and opinion terms extraction. In *Proceedings of the International Conference on Computational Linguistics*, pages 1772–1777, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks.

O. Ronneberger, P.Fischer, and T. Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351 of *LNCS*, pages 234–241. Springer.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 999–1005, Austin, Texas. Association for Computational Linguistics.

T. K. Shivaprasad and Jyothi Shetty. 2017. Sentiment analysis of product reviews: A review. In *Proceedings of the International Conference on Inventive Communication and Computational Technologies*, pages 298–301.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *Proceedings of the China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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*, 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 Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 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 Conference on Empirical Methods in Natural Language Processing*, pages 2339–2349, Online. Association for Computational Linguistics.

Wei Xue, Wubai Zhou, Tao Li, and Qing Wang. 2017. MTNA: A neural multi-task model for aspect category classification and aspect term extraction on

restaurant reviews. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 151–156, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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 Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 2416–2429, Online. Association for Computational Linguistics.

Bishan Yang and Claire Cardie. 2014. Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 325–335.

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*, pages 819–828, Online. Association for Computational Linguistics.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021a. Document-level relation extraction as semantic segmentation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 504–510, Online. Association for Computational Linguistics.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6230–6239.