A multi-task learning network using shared BERT models for aspect-based sentiment analysis

Quanzhen LIU † and Mizuho IWAIHARA ‡

† Graduate School of Information, Production and Systems, Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, Fukuoka, 808-0135 Japan

E-mail: † kenshin_ryuu@akane.waseda.jp, ‡ iwaihara@waseda.jp

Abstract Aspect-based sentiment analysis (ABSA) aims to predict the sentiment polarity of specific aspect words occurring in a text. ABSA includes aspect-category sentiment analysis (ACSA) and aspect-target sentiment analysis (ATSA). There have been many previous studies addressing both tasks through RNNs and other neural models. With BERT's remarkable performance on NLP tasks, several studies have enhanced aspect word extraction to solve ATSA by building new BERT-based models. But such an approach is not directly applicable to the ACSA task, because aspect words in ACSA are often not explicitly present in the text, so aspect word extraction becomes more difficult. In this paper, we propose a multi-task learning (MTL) approach to solve these problems. Our approach is based on a shared BERT model to construct a multi-task learning network, which is trained by strongly and weakly related tasks. We also use a multi-head self-attention layer to replace a linear layer in traditional multi-task learning networks, to enhance the ability to capture global semantics. We also propose a new fine-tuning strategy that can better improve the performance of the model. Experiments were conducted on four datasets from the ATSA task and four datasets from the ACSA task: laptop, restaurant, restaurant-2014, and restaurant-large from SemEval-2014. Our experimental results show that: In the ACSA task, our model outperformed all the baseline models, achieving the current state-of-the-art performance on the multiple datasets. For the ATSA task, our model performs close to the state-of-the-art, with much simpler architecture.

Keyword Aspect-based sentiment analysis, Multi-task learning, Fine tune, Shared BERT

1. Introduction

Aspect-based sentiment analysis (ABSA) is a fine-grained task to predict sentiment polarities specific to aspect words occurring in a text. It is more complicated than traditional sentiment analysis since ABSA requires a more in-depth analysis of contexts within a sentence or a document.

There are two subtasks in ABSA, namely aspect-category sentiment analysis (ACSA), and aspect-target sentiment analysis (ATSA). The ACSA task aims at predicting the sentiment polarity on given aspects. The aspects are in several pre-determined categories, and they may not appear in the sentence. The ATSA task aims at predicting the sentiment polarity of a target aspect word.

Table 1. Sentence contains different sentiment polarities towards two aspects.

Sentence	This dessert is delicious, but the price is a bit expensive.	
Category word	food	price
Target word	This dessert	price
Polarity	positive	negative

As shown in Table 1, the aspect-category sentiment analysis (ACSA) predicts sentiment polarity toward the aspect word *"food,"* which does not appear in the sentence. In contrast, aspect-target sentiment analysis (ATSA) aims to predict sentiment polarity toward the aspect word that is a part of the sentence. For example, aspect-target sentiment analysis would predict sentiment polarity for the target word *"This dessert,"* which is a part of the sentence.

Sentiment polarities in a sentence may be unequal when considering multiple aspect words. Therefore, a deep understanding of sentences for given aspect words is essential for ABSA. However, not all words in a sentence are useful for predicting sentiment polarities. For example, the words *"dessert"* and *"delicious"* are irrelevant for sentiment prediction for the aspect word *"price."* Without discerning the contexts of these words, the final sentiment prediction will fail.

Numerous existing models [7][23][26][30] typically use aspect-independent encoders to generate sentence representations, and then apply attention mechanisms [17] or gating mechanisms for feature selection and extraction, which are expected to produce noise-free representations. Besides, several models [19][29] concatenate aspect embeddings with each word embedding of a sentence and then use traditional long and short-term memories (LSTMs)

[9] to generate sentence representations. However, it is insufficient to exploit the given aspects and perform potentially complex feature selection and extraction [16]. Several approaches are proposed upon these studies, such as an aspect-guide GRU encoder based on a deep transition gated loop unit, which uses the given aspect to guide the earliest stage's sentence encoding process. The model is also forced to reconstruct the aspect-guided encoder's sentence representation for the given aspect word, yielding good results. However, traditional network designs have relied on RNNs. RNNs, such as LSTMs [9], are highly expressive but challenging to parallelize and require a vast memory space and computation time to propagate backward through time. Furthermore, every RNN's training algorithm is a truncated BPTT, which affects the model's ability to capture dependencies on longer time scales. While LSTM can alleviate the vanishing gradient problem, maintaining information over long distances typically requires large amounts of training data.

To solve this problem, we utilize the widely popular pretrained language model BERT [5], and design more sophisticated learning networks based on this model to achieve our goal. The BERT model can better compensate for the shortcomings of RNN networks.

In this paper, we propose a multi-task learning network based on the BERT model. We use four tasks for training alternately, so that the four tasks share parameters during the learning process. We also discuss fine-tuning the BERT model to make it more adopted to our task. We evaluate our model on multiple datasets for two subtasks of ABSA. Our experimental results demonstrate the effectiveness of our proposed approach.

2. Related Work

There are a variety of sentiment analysis tasks, such as document-level [27], sentence-level [33][34], aspect-level [20][28][31], and multimodal [1][3] sentiment analysis. For aspect-level sentiment analysis, previous work has typically applied attention mechanisms [17] combined with memory networks or gating units to address this task [2][6][10][12][23][25][30][32]. Several works have also used aspect weak association encoders to generate aspect-specific sentence representations [19][23][29]. All these approaches do not make sufficient use of the given aspect information. Also, they suffer from gradient disappearance and the absence of interaction between the target term and the context. Also, several approaches jointly extract aspect items (and opinion items) and predict their sentiment polarity [11][15][18][21].

3. Methodology

3.1. Problem definition

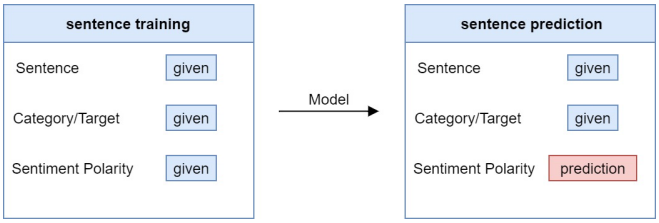

Fig. 1. Task definition

As shown in Figure 1, our task is for given aspect words, which express either category words or target words, to predict sentiment polarity in a sentence. In the training process, the model is trained with training samples (each consists of text, aspect words, and sentiment polarity). In the testing phase, the model predicts the sentiment polarity of the given aspect words in each test sample (which consists of text and aspect words). We use accuracy to evaluate the performance of the whole model on the task.

3.2. Model Overview

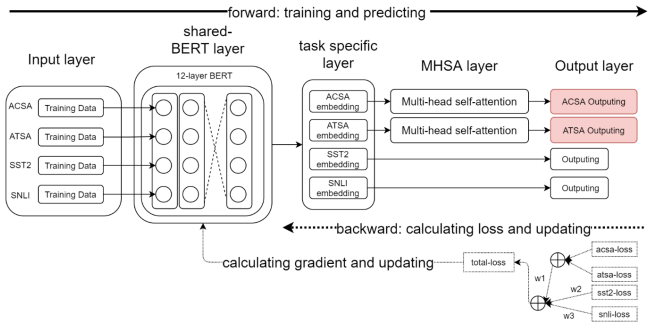

Fig. 2. Model overview

In this paper, we design a multi-task learning model based on BERT. Our model structure consists of four parts: The input layer, the shared-BERT layers, the MHSA layer, and the output layer. As shown in Figure 2, the input layer is supplied with datasets from multiple tasks. The multiple tasks consist of the main tasks and a set of related tasks. The related tasks are supposed to improve the main tasks' prediction performance through data augmentation and regularization. In this paper, our main tasks are ATSA and ACSA, and we choose two tasks, indicated as SST2 and SNLI in Figure 2, as related tasks. The datasets of the multiple tasks are entered into the shared BERT layers through unified processing. After the shared BERT layers, we obtain the output tensor of the multi-tasks. Then we extract the output tensor of the main tasks and supply it into the MHSA layers for final prediction.

To simplify the design of the model's overall loss

function, we adopt a weighted sum of the loss functions of the main and related tasks as the overall loss function.

3.3. Input layer

Regarding selection of tasks in the multi-task learning, we adopt the following policies.

First, ABSA is a sentence-level sentiment analysis based on given aspect words, where the ACSA task and the ATSA task are our main tasks. We choose to train both ACSA and ATSA together in our multi-task learning, to positively influence each other, instead of learning the main tasks separately.

For the ACSA task that predicts the sentiment polarity of the hypernym of the aspect word in a given sentence, we observe that the sentiment polarities of the hypernym (aspect category) and the hyponym (aspect target) are the same in most cases. When we train the ACSA task, we can positively influence the ATSA task on the same training sentence. Similarly, when we train the ATSA task, if the sentiment polarity of the hyponym and its hypernym are identical, at the same time the ACSA task can also be trained on the same training sentence.

There is a sentence *"I like eating cakes here but hating their service."* In the ACSA task, we can find *"positive"* for the category word *"sweet."* For the ATSA task, we can find *"positive"* on the target word *"cakes."* We can find ATSA task and ACSA task will positively influence each other on this sentence.

However, in some special situations, simultaneous learning for ATSA and ACSA is difficult. Here is a sentence *"I like eating apples but dislike bananas."* When we perform the ATSA task on this sentence, we can obtain corresponding sentiment polarities *"positive"* and *"negative"* for the target words *"apples"* and *"bananas,"* respectively. However, when we perform the ACSA task, we find that *"apple"* and *"banana"* belong to the same category word *"fruit."* Since an aspect word cannot have multiple polarities simultaneously, a new sentiment label *"conflict"* will be assigned to such an aspect word. We cannot expect mutual befits between ATSA and ACSA in such conflict situation.

In the evaluation, we conduct separate experiments whether including the conflict situation or not.

For the related tasks, we select SST-2 [36] and SNLI [37], based on the following points:

i. The SST-2 task is a sentiment analysis task that focuses on movie reviews to do sentiment classification, and it belongs to the text classification task of single sentences. It determines the sentiment polarity of the whole sentence by analyzing the sentiment words in the sentence, which has a strong correlation with our two main tasks.

ii. SNLI is a natural language inference task that analyzes the semantics of two sentences. Unlike our main task and the SST-2 task, it is not a sentiment analysis task. But it is also a classification task based on meanings of sentences. They are both sentence-level tasks in NLP.

iii. For the task selection, we try to choose subtasks that have the same evaluation metric and loss function, so that it is easy for us to construct the total loss function of the model.

By the above reasons, we choose SST-2 and SNLI as our related tasks; SST-2 is a binary sentiment classification task, and SNLI is a three-class natural language inference task.

3.4. Shared-BERT layers and multi-task learning

ACSA and ATSA are our main tasks, and we transform the context and aspect word of each sample of ACSA and ATSA into the format:

$$[CLS] + context + [SEP] + aspect\ word + [SEP].$$

Here, $[CLS]$ indicates that the feature is used for classification tasks, where the feature vector at the $[CLS]$ token is used for the final classification. The $[SEP]$ token denotes the split-sentence symbol, which is used to indicate the boundary of two segments in an input sequence.

SST-2 is a sentiment classification dataset and only performs binary sentiment classification at the sentence level. We transform samples of SST-2 into the format:

$$[CLS] + context + [SEP].$$

SNLI is a natural language inference task, and we transform samples into the format:

$$[CLS] + sentence1 + [SEP] + sentence2 + [SEP].$$

In our work, we only need to focus on the performance of the ACSA task and the ATSA task. Therefore, we do not need to focus on the SST-2 task and the SNLI task. Here, we use a weighted summation approach to strengthen the generalization ability of the model for the ACSA task and the ATSA task.

Among them, since SST-2 is a binary classification task, we use the following loss function:

$$L = \frac{1}{N}\sum_{i} L_i = \frac{1}{N}\sum_{i} -[y_i \cdot \log(p_i) + (1 - y_i) \cdot log(1 - p_i)]$$

Here, y_i denotes the label of sample i, where the positive label is 1, and the negative label is 0. p_i denotes

the probability that sample i is predicted to be positive. The other three tasks are three-class classification, so we use the following function:

$$L = \frac{1}{N}\sum_{i} L_i = \frac{1}{N}\sum_{i} -\sum_{c=1}^{M} y_{ic}\log(p_{ic})$$

Here, M is the number of categories. The variable y_{ic} equals 1 if the category c is the same as the category of sample i, and 0 otherwise, and p is the predicted probability of sample i belonging to category c.

We compute the four tasks' loss values and combine them by a linear function as:

$$total_{loss} = \mathrm{w}_1 * (acsa_{loss} + atsa_{loss}) + \mathrm{w}_2 * sst_{loss} + \mathrm{w}_3 * snli_{loss}$$

which gives a weight to each task loss, and then the overall loss is backpropagated to the BERT layers for the whole model update.

3.5. Multi-Head Self-Attention layers

By using shared-BERT layers, we can obtain the corresponding output tensor from each of the four tasks. The tensor of each of the main tasks, ACSA and ATSA, is extracted and entered into a multi-head self-attention layer.

As the task's training is completed in traditional multi-task models, we often directly extract the resulting tensor for the corresponding task and then make it pass through a linear layer before predicting the output. Although this simplifies the model's computational effort during training, we believe that such work can still be improved for tasks such as sentiment analysis, which needs to analyze long and complex contexts. Therefore, in our work, we add a multi-head self-attentive layer between the output layer and the shared-BERT layer, which is expected to further improve the understanding of global semantics.

Here, we have taken out the multi-head self-attention as a separate computational layer based on the transformer structure, which is involved in an additional computation before the output. It can further compute the tensor obtained after multi-task training to enhance the global semantics acquisition and then make the final prediction. This layer is similar in structure to the one in BERT. Multi-head self-attention in BERT is used for the computation of sequences during training, and it is involved in the training processing of the shared BERT layers. The multi-head self-attention layer here is used for the last tensor computation before output.

3.6. Output layer

In the output layer, the representation learned from the feature interaction layer is pooled by extracting the hidden state at the first token's corresponding position. Finally, the output layer is applied to predict the sentiment polarity.

3.7. Fine-tuning

For the shared BERT layers, we adopt *"bert-base-uncased,"* which has 12 layers. For fine-tuning the model through multi-task learning, we freeze the first eight layers of the model and fine-tune only its last four layers. Also, we use learning rate decay and weight decay to improve our fine-tuning effect further. Besides, to alleviate too few samples in our dataset, we also apply SemEval-2015 and SemEval-2016 datasets as data augmentation to fine-tune our model.

4. Datasets and baselines

4.1. Datasets and Metrics

Our experiments are evaluated on two datasets for ACSA and two datasets for ATSA. In these four datasets, the full datasets are given the symbol "DS." Also, to evaluate how a model can perform in difficult situations where a sentence contains opposite sentiment polarities on different aspects, a hard dataset is extracted from each full dataset and given the symbol "HDS," which consists of sentences having non-uniform sentiment labels on multiple aspects.

Aspect-Category Sentiment Analysis. We use the restaurant review dataset from SemEval-2014 Task 4 ("Restaurant-14") to evaluate the ACSA task. The dataset contains five predefined aspects and four sentiment labels. A larger dataset ("Restaurant-Large") covers restaurant reviews for the three years from 2014 to 2016. This dataset has eight predefined aspects and three labels. The statistics of the datasets are shown in Table 2.

Here, to facilitate comparison with baselines, we perform two sets of experiments: 3-class classification experiments where the conflict samples are removed, and 4-class classification experiments where the conflict samples are included. We refer to the treatment of AGDT and GCAE when we do the restaurant-large dataset experiments in the ACSA task and replace the conflict label with a neutral label. In the other tasks of ACSA, we remove the conflict samples for the corresponding experiments in the same way as other baselines.

Table 2. Statistics of datasets for the aspect-category sentiment analysis task.

		Positive		Negative		Neutral		Conflict		Total	
		DS	HDS	DS	HDS	DS	HDS	DS	HDS	DS	HDS
Restaurant-14	train	2179	139	839	136	500	50	195	40	3713	365
	test	657	32	222	26	94	12	52	19	1025	89
Restaurant-large	train	2710	182	1198	178	757	107	-	-	4665	467
	test	1505	92	680	81	241	61	-	-	2426	234

Table 3. Statistics of datasets for the aspect-target sentiment analysis task.

		positive		negative		Neutral		Conflict		Total	
		DS	HDS	DS	HDS	DS	HDS	DS	HDS	DS	HDS
Restaurant	train	2164	379	805	323	633	293	91	43	3693	1038
	test	728	92	196	62	196	83	14	8	1134	245
Laptop	train	987	159	866	147	460	173	45	17	2358	496
	test	341	31	128	25	169	49	16	3	654	108

Aspect-Target Sentiment Analysis. We use the restaurant and laptop review datasets of SemEval-2014 Task 4 to evaluate the ATSA task. Both datasets contain four sentiment labels. Here, to facilitate comparison with baselines, we perform two sets of experiments: The 3-class classification experiments with the conflict label removed, and the 4-class classification experiments with the conflict label included. The statistics of the datasets are shown in Table 3.

Metrics. The evaluation metric is accuracy.

4.2. Baselines

ATAE-LSTM [29] This is an attention-based model of LSTM, which sums the given aspect embeddings with each word embedding, and then uses the jointly named embeddings as input to the LSTM. The output of LSTM is appended with aspect embedding again.

CNN [13]. The model focuses on extracting n-gram features to generate sentence representations for sentiment classification.

GCAE [30]. This model extracts features using CNNs and then uses two Gated Tanh-ReLu units to selectively output sentiment information flow aspects for sentiment label prediction.

IAN [18] The model uses two LSTMs and an interactive attention mechanism to learn sentence and aspect representations and concatenate them for sentiment prediction.

RAM [3] The model applies multiple attentional and memory networks to generate sentence representations.

TD-LSTM [24] The model uses two LSTMs to capture the left and right contexts of terms to generate target-relevant representations for sentiment prediction.

AGDT [16] The model uses a given aspect to reconstruct a given aspect with the generated sentence representation by bootstrapping the sentence encoding from scratch via a depth transformation architecture.

IRAM [19] IRAM is a model that leverages recurrent memory networks with a multihop attention mechanism.

Tnet [14] The model proposes a new component based on LSTM to better integrate information on the target words. CNN is also used as a feature extractor for the classifier.

VAE [4] The model is experimentally analyzed for the ATSA task. It is based on transformers to build encoders and decoders.

PBAN [8] The model is based on Bi-GRU, which uses the positional embedding of aspect words to calculate the corresponding weights. The method also uses a bi-attention mechanism to model the relationship between sentences and different aspect words while using positional information to determine the sentiment polarity of the aspect words.

AOA [12] An attention over attention model is proposed to learn one sentence's aspect words and critical parts.

MGAN [7] A new attention mechanism is proposed for this model to fuse aspect words with words in context to capture the interaction between words.

DAuM [35] A new neural network with an auxiliary memory function is proposed for this model to handle the sentiment classification task.

4.3. Other BERT-based models

BERT-pair-QA-B [22] The model constructs an auxiliary sentence from the aspect and convert the ACSA task into a sentence-pair classification task.

LCF-ATEPC [32] This model uses the local context focus mechanism and firstly proposes a multi-task learning model for Chinese-oriented aspect-based sentiment analysis.

5. Results and Analysis

We perform comparative experiments to show the superiority of our approach. BERT-noMTL-Fit refers to

the model that does not use the multitasking learning method, and a BERT model is fine-tuned to a single task. BERT-MTL-4-Fit refers to the model that uses multi-task learning of a shared BERT model with four tasks. BERT-MTL-2-Fit refers to the model that uses multi-task learning of a shared BERT model with ACSA task and ATSA task.

5.1. Aspect-Category Sentiment Analysis

5.1.1. Without "conflict" label

The results of ACSA in Table 4 show that our BERT-MTL-Fit outperforms all baseline models on both datasets, "Restaurant-14" and "Restaurant-Large." AGDT uses a given aspect to reconstruct a given aspect with the generated sentence representation by bootstrapping the sentence encoding from scratch via a depth transformation architecture. BERT-pair-QA-B constructs an auxiliary sentence from the aspect and converts the ACSA task to a sentence-pair classification task.

Compared with ATAE-LSTM, our BERT-MTL-4-Fit achieves a performance improvement of 6.24% in the "DS" part of the restaurant datasets. Compared with BERT-pair-QA-B, our BERT-MTL-4-Fit achieves a performance improvement of 0.34% in the "DS" part of the restaurant datasets.

Our BERT-MTL-4-Fit achieves a performance improvement of 1.69% in the "DS" part of the restaurant-large datasets on the restaurant-large dataset.

"HDS" aims to measure whether a model can well discriminate different sentiment polarities in one sentence, consisting of sentences with disagreeing polarities of multiple aspects. Our BERT-MTL-4-Fit outperforms AGDT by a margin on the restaurant-large dataset by 4.53%, which illustrates the remarkable advantage of our multitasking strategy on specific aspects of words.

Also, comparing BERT-noMTL-Fit, BERT-MTL-2-Fit, BERT-MTL-4-Fit, we find that the BERT-MTL-4-Fit has the best prediction performance, which indicates that our multitasking strategy is effective.

5.1.2. With the label "conflict"

We show our model's overall performance and the baseline models in Table 5. Compared with AGDT, our BERT-MTL-4-Fit achieves a performance improvement of 3.67% in the "DS" part of restaurant datasets. Compared with BERT-pair-QA-B, our BERT-MTL-4-Fit still has a 0.45% difference in the "DS" part of restaurant datasets. In the HDS part, our model achieved an improvement of 1.62%.

Table 4. The accuracy (%) result of the ACSA task without label "conflict." "*" refers to citing from AGDT. '-' means not reported.

ACSA task without label 'conflict'			
	restaurant	restaurant-large	
	DS	DS	HDS
ATAE-LSTM[29]*	84	83.91	66.32
CNN[13]*	-	84.28	50.43
GCAE[30]*	-	85.92	70.75
AGDT[16]*	-	87.55	75.73
other BERT models			
BERT-pair-QA-B[22]	89.9	-	-
our model			
BERT-MTL-4-Fit	**90.24**	**89.24**	**80.26**
BERT-noMTL-Fit	86.54	88.78	77.68
BERT-MTL-2-Fit	89.72	89.07	79.40

Table 5. The accuracy (%) result of the ACSA task with the label "conflict." "*" refers to citing from AGDT. '-' means not reported.

ACSA task with label 'conflict.'		
	restaurant	
	DS	HDS
ATAE-LSTM[29]*	78.29	45.62
CNN[13]*	79.47	44.94
GCAE[30]*	79.35	50.55
AGDT[16]*	81.78	62.02
other BERT models		
BERT-pair-QA-B[22]	**85.9**	-
our model		
BERT-MTL-4-Fit	85.45	**63.64**
BERT-noMTL-Fit	82.7	62.5
BERT-MTL-2-Fit	85.35	62.5

5.2. Aspect-Target Sentiment Analysis

5.2.1. Without "conflict" label

As shown in Table 6 on ATSA results, our BERT-MTL-4-Fit model consistently outperforms all the comparison methods in both datasets, except LCF-ATEPC.

In the "DS" datasets, our BERT-MTL-4-Fit model outperforms all baseline models except LCF-ATEPC, suggesting that the BERT-MTL-4-Fit model is better able to predict the sentiment polarity of multifaceted words and that fine-tuning helps the model to learn further.

However, our results are lower than LCF-ATEPC. The model applies self-attention and local context focus techniques to aspect word extraction tasks and fully explores their potential in aspect word extraction. This approach fully combines the two tasks of target word extraction and text analysis while using the self-attention mechanism to make them better together, achieving the best results so far. In the current ATSA task, target word extraction is the most effective solution. But it has a disadvantage that it cannot be used in ACSA tasks because the target words in the ACSA task are not necessarily words or phrases in a sentence, which makes the target

word extraction much difficult.

5.2.2. With the label "conflict"

As shown in Table 7 on ATSA results, our BERT-MTL-4-Fit model consistently outperforms all the comparison methods in both datasets.

In the "DS" datasets, our BERT-MTL-Fit model outperforms all baseline models, suggesting that the BERT-MTL-4-Fit model is better able to predict the sentiment polarity of aspect words and that fine-tuning helps the model to learn further.

In the "HDS" datasets, our BERT-MTL-4-Fit model's accuracy is 14.26% higher than AGDT in the restaurant dataset and 15.06% higher than AGDT in the laptop dataset, indicating that our model is more capable than AGDT in aspect sentiment problems. These results further demonstrate that our model works well on a variety of tasks and datasets.

Table 6. The accuracy (%) result of the ATSA task without label "conflict." "*" refers to citing from AGDT.

ATSA task without label 'conflict.'		
	restaurant	laptop
	DS	DS
IARM[19]*	80	73.8
Tnet[14]*	80.79	76.54
VAE[4]*	81.1	75.34
PBAN[8]*	81.16	74.12
AOA[12]*	81.2	74.5
MGAN[7]*	81.25	75.39
DAuM[35]*	82.32	74.45
other BERT models		
LCF-ATEPC[32]	**90.18**	**82.29**
our model		
BERT-MTL-4-Fit	84.99	81.16
BERT-noMTL-Fit	83.47	78.96
BERT-MTL-2-Fit	84.09	79.18

Table 7. The accuracy (%) result of the ATSA task with label "conflict." "*" refers to citing from AGDT.

ATSA task with label 'conflict.'				
	restaurant		laptop	
	DS	HDS	DS	HDS
TD-LSTM[24]*	73.44	56.48	62.23	46.11
ATAE-LSTM[29]*	73.74	50.98	64.38	40.39
IAN[18]*	76.34	55.16	68.49	44.51
RAM[3]*	76.97	55.85	68.48	45.37
GCAE[30]*	77.28	56.73	69.14	47.06
AGDT[16]*	78.85	60.33	71.5	51.3
our model				
BERT-MTL-4-Fit	**81.64**	**74.59**	**77.18**	**66.36**
BERT-noMTL-Fit	81.02	69.67	75.19	63.55
BERT-MTL-2-Fit	81.53	71.72	76.26	64.17

6. Conclusion and Future work

In this paper, we proposed a BERT-based multi-task learning model called BERT-MTL-Fit. Empirical studies from four datasets show that BERT-MTL-Fit significantly improves existing aspect-based sentiment analysis models, especially on the aspect-category sentiment analysis task. However, experiments with multiple datasets show that our model still cannot effectively solve the problem of aspect words that possess multiple sentiment polarities. It is a current challenge in the field of sentence-level sentiment analysis.

Our multi-task model also uses the most straightforward way of weight summation when constructing the loss function for generalizing the overall model performance. This approach makes it easy to decide the weights for each task. But it still has large space for improvement. For example, we can adjust each task's weights in real-time by the speed of fitting each task during the training process, when the whole training process is dynamic. We can also generalize the model's performance better by changing the relevant tasks or adding more subtasks. These directions would achieve further improvements.

References

[1] Akhtar, Md Shad, et al. "Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis." 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). 2019.

[2] Bao, Lingxian, Patrik Lambert, and Toni Badia. "Attention and lexicon regularized LSTM for aspect-based sentiment analysis." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. 2019.

[3] Chen, Peng, et al. "Recurrent attention network on memory for aspect sentiment analysis." Proceedings of the 2017 conference on empirical methods in natural language processing. 2017.

[4] Cheng, Xingyi, et al. "Variational Semi-Supervised Aspect-Term Sentiment Analysis via Transformer." Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). 2019.

[5] Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL-HLT (1). 2019.

[6] Duan, Junwen, Xiao Ding, and Ting Liu. "Learning sentence representations over tree structures for target-dependent classification." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018.

[7] Fan, Feifan, Yansong Feng, and Dongyan Zhao. "Multi-grained attention network for aspect-level sentiment classification." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[8] Gu, Shuqin, et al. "A position-aware bidirectional attention network for aspect-level sentiment analysis." Proceedings of the 27th International Conference on Computational Linguistics. 2018.

[9] H. Sepp and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[10] He, Ruidan, et al. "Effective attention modeling for aspect-level sentiment classification." Proceedings of the 27th international conference on computational linguistics. 2018.

[11] Huang, Binxuan, and Kathleen M. Carley. "Parameterized Convolutional Neural Networks for Aspect Level Sentiment Classification." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[12] Huang, Binxuan, Yanglan Ou, and Kathleen M. Carley. "Aspect level sentiment classification with attention-over-attention neural networks." International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer, Cham, 2018.

[13] Kim, Yoon. "Convolutional Neural Networks for Sentence Classification." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

[14] Li, Xin, et al. "Transformation Networks for Target-Oriented Sentiment Classification." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.

[15] Li, Xin, et al. "A unified model for opinion target extraction and target sentiment prediction." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.

[16] Liang, Yunlong, et al. "A Novel Aspect-Guided Deep Transition Model for Aspect Based Sentiment Analysis." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[17] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.

[18] Ma, Dehong, Sujian Li, and Houfeng Wang. "Joint learning for targeted sentiment analysis." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.

[19] Majumder, Navonil, et al. "IARM: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis." Proceedings of the 2018 conference on empirical methods in natural language processing, 2018.

[20] Pontiki, Maria, et al. "SemEval-2014 Task 4: Aspect Based Sentiment Analysis."

[21] Schmitt, Martin, et al. "Joint Aspect and Polarity Classification for Aspect-based Sentiment Analysis with End-to-End Neural Networks." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[22] Sun, Chi, Luyao Huang, and Xipeng Qiu. "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence." 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). 2019.

[23] Tang, Duyu, Bing Qin, and Ting Liu. "Aspect Level Sentiment Classification with Deep Memory Network." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016.

[24] Tang, Duyu, et al. "Effective LSTMs for Target-Dependent Sentiment Classification." Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016.

[25] Tang, Jialong, et al. "Progressive Self-Supervised Attention Learning for Aspect-Level Sentiment Analysis." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[26] Tay, Yi, Luu Anh Tuan, and Siu Cheung Hui. "Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

[27] Thongtan, Tan, and Tanasanee Phienthrakul. "Sentiment classification using document embeddings trained with cosine similarity." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. 2019.

[28] Wang, Jingjing, et al. "Aspect sentiment classification towards question-answering with reinforced bidirectional attention network." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[29] Wang, Yequan, et al. "Attention-based LSTM for aspect-level sentiment classification." Proceedings of the 2016 conference on empirical methods in natural language processing. 2016.

[30] Xue, Wei, and Tao Li. "Aspect Based Sentiment Analysis with Gated Convolutional Networks." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.

[31] Yang, Chao, et al. "Aspect-based sentiment analysis with alternating coattention networks." Information Processing & Management 56.3 (2019): 463-478.

[32] Yang, Heng, et al. "A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction." arXiv preprint arXiv:1912.07976 (2019).

[33] Zhang, Liwen, Kewei Tu, and Yue Zhang. "Latent variable sentiment grammar." arXiv preprint arXiv:1907.00218 (2019).

[34] Zhang, Yuan, and Yue Zhang. "Tree communication models for sentiment analysis." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[35] Zhu, Peisong, and Tieyun Qian. "Enhanced aspect level sentiment classification with auxiliary memory." Proceedings of the 27th International Conference on Computational Linguistics. 2018.

[36] "Sentiment Analysis", https://nlp.stanford.edu/sentiment/index.html

[37] "The Stanford Natural Language Inference (SNLI) Corpus", https://nlp.stanford.edu/projects/snli/