A Joint Training Dual-MRC Framework for Aspect Based Sentiment Analysis

Yue Mao, Yi Shen, Chao Yu, Longjun Cai
Alibaba Group, Beijing, China
{maoyue.my, sy133447, aiqi.yc, longjun.clj}@alibaba-inc.com

Abstract

Aspect based sentiment analysis (ABSA) involves three fundamental subtasks: aspect term extraction, opinion term extraction, and aspect-level sentiment classification. Early works only focused on solving one of these subtasks individually. Some recent work focused on solving a combination of two subtasks, e.g., extracting aspect terms along with sentiment polarities or extracting the aspect and opinion terms pair-wisely. More recently, the triple extraction task has been proposed, i.e., extracting the (aspect term, opinion term, sentiment polarity) triples from a sentence. However, previous approaches fail to solve all subtasks in a unified end-to-end framework. In this paper, we propose a complete solution for ABSA. We construct two machine reading comprehension (MRC) problems and solve all subtasks by joint training two BERT-MRC models with parameters sharing. We conduct experiments on these subtasks, and results on several benchmark datasets demonstrate the effectiveness of our proposed framework, which significantly outperforms existing state-of-the-art methods.

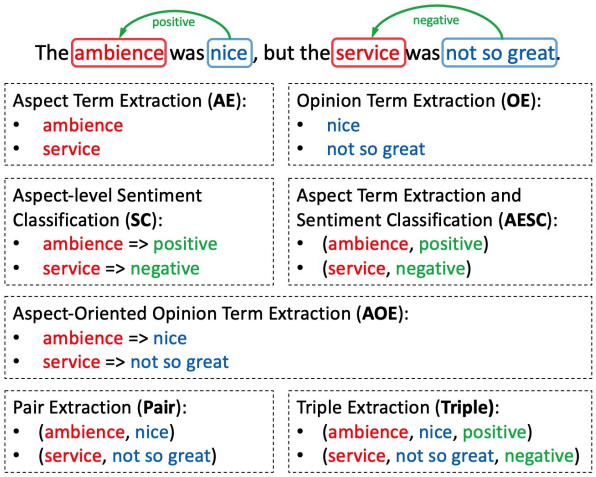

Figure 1: An illustrative example of ABSA substasks.

Introduction

Aspect based sentiment analysis (ABSA)[1] is an important research area in natural language processing. Consider the example in Figure 1, in the sentence "*The ambience was nice, but the service was not so great.*", the aspect terms (AT) are "*ambience/service*" and the opinion terms (OT) are "*nice/not so great*". Traditionally, there exist three fundamental subtasks: aspect term extraction, opinion term extraction, and aspect-level sentiment classification. Recent research works aim to do a combination of two subtasks and have achieved great progress. For example, they extract (AT, OT) pairs, or extract ATs with corresponding sentiment polarities (SP). More recently, some work that aims to do all related subtasks in ABSA with a unified framework has raised increasing interests.

For convenience, we assume the following abbreviations of ABSA subtasks as illustrated in Figure 1:

- ***AE***: AT extraction
- ***OE***: OT extraction
- ***SC***: aspect-level sentiment classification
- ***AESC***[2]: AT extraction and sentiment classification
- ***AOE***[3]: aspect-oriented OT extraction
- ***Pair***: (AT, OT) pair extraction
- ***Triple***: (AT, OT, SP) triple extraction.

We mainly focus on the task of extracting (a, o, s) triples since it is the hardest among all ABSA substasks. Peng et al. (2020) proposed a unified framework to extract (AT, OT, SP) triples. However, it is computationally inefficient as its framework has two stages and has to train three separate models.

In this paper, we propose a joint training framework to handle all ABSA subtasks (described in Figure 1) in one single model. We use BERT (Devlin et al. 2019) as our backbone network and use a span based model to detect the start/end positions of ATs/OTs from a sentence. Span based methods outperform traditional sequence tagging based methods for extraction tasks (Hu et al. 2019).

[1]It is also referred as target based sentiment analysis (TBSA).

[2]It is also referred as aspect based sentiment analysis (ABSA).

[3]It is also referred as target oriented opinion word extraction (TOWE).

Subtasks	Left-MRC	Right-MRC	
	Extraction	Classification	Extraction
AE	√		
AOE			√
SC		√	
AESC	√	√	
Pair	√		√
Triple	√	√	√

Table 1: Our proposed dual-MRC can handle all ABSA subtasks.

Following its idea, a heuristic multi-span decoding algorithm is used, which is based on the non-maximum suppression algorithm (NMS) (Rosenfeld and Thurston 1971).

We convert the original triple extraction task to two machine reading comprehension (MRC) problems. MRC methods are known to be effective if a pre-trained BERT model is used. The reason might be that BERT is usually pre-trained with the next sentence prediction to capture the pairwise sentence relations. Theoretically, the triple extraction task can be decomposited to subtasks *AE*, *AOE* and *SC* . Thus, we use the left MRC to handle *AE* and the right MRC to handle *AOE* and *SC*. Our main contributions in this paper are as follows:

- We show the triple extraction task can be jointly trained with three objectives.
- We propose a dual-MRC framework that can handle all subtasks in ABSA (as illustrated in Table 1).
- We conduct experiments to compare our proposed framework on these tasks. Experimental results show that our proposed method outperforms the state-of-the-art methods.

Related Work

Aspect-based sentiment analysis (ABSA) has been widely studied since it was first proposed in (Hu and Liu 2004). In this section, we present existing works on ABSA according to related subtasks.

SC. Various neural models have been proposed for this task in recent years. The core idea of these works is to capture the intricate relationship between an aspect and its context by designing various neural architectures such as CNN (Huang and Carley 2018; Li et al. 2018a), RNN (Tang et al. 2016; Zhang, Zhang, and Vo 2016; Ruder, Ghaffari, and Breslin 2016), attention-based network (Ma et al. 2017; Du et al. 2019; Wang et al. 2016; Gu et al. 2018; Yang et al. 2017), memory network(Tang, Qin, and Liu 2016; Chen et al. 2017; Fan et al. 2018). Sun, Huang, and Qiu (2019) convert *SC* to a BERT sentence-pair classification task, which achieves state-of-the-art results of this task.

AE. As the pre-task of *SC*, *AE* aims to identify all aspect terms in a sentence (Hu and Liu 2004; Pontiki et al. 2014) and is usually regarded as a sequence labeling problem (Li et al. 2018b; Xu et al. 2018; He et al. 2017). Besides, (Ma et al. 2019) and (Li et al. 2020) formulated *AE* as a sequence-to-sequence learning task and also achieved impressive results.

AESC. In order to make *AESC* meet the needs of practical use, plenty of previous works make efforts to solve *AE* and *SC* simultaneously. Simply merging *AE* and *SC* in a pipeline manner will lead to an error-propagation problem (Ma, Li, and Wang 2018). Some works (Li et al. 2019a,b) attempt to extract aspects and predicting corresponding sentiment polarities jointly through sequence tagging based on a unified tagging scheme. However, these approaches are inefficient due to the compositionality of candidate labels (Lee et al. 2016) and may suffer the sentiment inconsistency problem. Zhou et al. (2019) and Hu et al. (2019) utilize span-based methods to conduct *AE* and *SC* at the span-level rather than token-level, which are able to overcome the sentiment inconsistency problem. It is worth noting that the information of opinion terms is under-exploited during these works.

OE. Opinion term extraction (*OE*) is widely employed as an auxiliary task to improve the performance of *AE* (Yu, Jiang, and Xia 2019; Wang et al. 2017; Wang and Pan 2018), *SC* (He et al. 2019) or both of them (Chen and Qian 2020). However, the extracted ATs and OTs in these works are not in pairs, as a result, they can not provide the cause for corresponding polarity of an aspect.

AOE. The task *AOE* (Fan et al. 2019) has been proposed for the pair-wise aspect and opinion terms extraction in which the aspect terms are given in advance. Fan et al. (2019) design an aspect-fused sequence tagging approach for this task. Wu et al. (2020) utilize a transfer learning method that leverages latent opinions knowledge from auxiliary datasets to boost the performance of *AOE*.

Pair. Zhao et al. (2020) proposed the *Pair* task to extract aspect-opinion pairs from scratch, they develop a span-based multi-task framework, which first enumerates all the candidate spans and then construct two classifiers to identify the types of spans (i.e. aspect or opinion terms) and the relationship between spans.

Triple. Peng et al. (2020) defined the triple extraction task for ABSA, which aims to extract all possible aspect terms as well as their corresponding opinion term and sentiment polarity. The method proposed in (Peng et al. 2020) is a two-stage framework, the first stage contains two separate modules, one is a unified sequence tagging model for *AE* and *SC*, the other is a graph convolutional neural network(GCN) for *OE*. In the second stage, all possible aspect-opinion pairs are enumerated and a binary classifier is constructed to judge whether the aspect term and opinion term match with each other. The main difference between our work and (Peng et al. 2020) is that we regard all subtasks as a question-answering problem, and propose a unified framework based on a single model.

Proposed Framework

Joint Training for Triple Extraction

In this section, we focus on the triple extraction task and the other subtasks can be regarded as special cases of it. Given a sentence x_j with max-length n as the input. Let $T_j = \{(a, o, s)\}$ be the output of annotated triples given the

input sentence x_j, where $s \in \{\text{Positive, Neutral, Negative}\}$ and (a, o, s) refers to (aspect term, opinion term and sentiment polarity). For the training set $\mathcal{D} = \{(x_j, T_j)\}$, we want to maximize the likelihood

$$L(\mathcal{D}) = \prod_{j=1}^{|\mathcal{D}|} \prod_{(a,o,s)\in T_j} P((a, o, s)|x_j). \quad (1)$$

Define

$$T_j|a := \{(o, s), (a, o, s) \in T_j\}, \quad k_{j,a} := |T_j|a|. \quad (2)$$

Consider the log-likelihood for x_j,

$$\begin{aligned}
\ell(x_j) &= \sum_{(a,o,s)\in T_j} \log P((a, o, s)|x_j) \\
&= \sum_{a\in T_j} \sum_{(o,s)\in T_j|a} \log P(a|x_j) + \log P((o, s)|a, x_j) \\
&= \sum_{a\in T_j} \left(\sum_{(o,s)\in T_j|a} \log P(a|x_j) \right) \\
&+ \sum_{a\in T_j} \left(\sum_{(o,s)\in T_j|a} \log P(s|a, x_j) + \log P(o|a, x_j) \right)
\end{aligned} \quad (3)$$

The last equation holds because the opinion terms o and the sentiment polarity s are conditionally independent given the sentence x_j and the aspect term a. [4]

$$\begin{aligned}
\ell(x_j) &= \sum_{a\in T_j} k_{j,a} \cdot \log P(a|x_j) \\
&+ \sum_{a\in T_j} \left(k_{j,a} \cdot \log P(s|a, x_j) + \sum_{o\in T_j|a} \log P(o|a, x_j) \right)
\end{aligned} \quad (4)$$

We sum above equation over $x_j \in \mathcal{D}$ and normalize the both sides, then we get the log-likelihood of the following form

$$\begin{aligned}
\ell(\mathcal{D}) &= \alpha \cdot \sum_{j=1}^{|\mathcal{D}|} \sum_{a\in T_j} \left(\sum_{a\in T_j} \log P(a|x_j) \right) \\
&+ \beta \cdot \sum_{j=1}^{|\mathcal{D}|} \sum_{a\in T_j} \log P(s|a, x_j) \\
&+ \gamma \cdot \sum_{j=1}^{|\mathcal{D}|} \sum_{a\in T_j} \left(\sum_{o\in T_j|a} \log P(o|a, x_j) \right)
\end{aligned} \quad (5)$$

where $\alpha, \beta, \gamma \in [0, 1]$. The first term is repeated in order to match with the other two terms. From (5), we may conclude the triple extraction task *Triple* can be converted to the joint training of *AE*, *SC* and *AOE*.

Dual-MRC Framework

Now we are going to propose our joint training dual-MRC framework. As illustrated in Figure 2, our model consists of two parts. Both parts use BERT (Devlin et al. 2019) as their backbone models to encode the context information. Recall that BERT is a multi-layer bidirectional Transformer based language representation model. Let n denote the sentence length and d denote the hidden dimension. Suppose the last layer outputs for all tokens are $h^{l,s}, h^{r,s}, h^{l,e}, h^{r,e} \in \mathbb{R}^{(n+2)\times d}$ which are used for extraction, where l/r refer to the left/right part and s/e refer to the stard/end token. Suppose the output of BERT at the [CLS] token is $h^r_{cls} \in \mathbb{R}^{(n+2)\times d}$ which is used for classification.

The goal of the left part is to extract all ATs from the given text, i.e., the task *AE*. As we discussed previously, span based methods are proven to be effective for extraction tasks. We follow the idea in (Hu et al. 2019), for the left part, we obtain the logits and probabilities for start/end positions

$$g^{l,s} = W^{l,s}h^{l,s}, \quad p^{l,s} = softmax(g^{l,s}) \quad (6)$$

$$g^{l,e} = W^{l,e}h^{l,e}, \quad p^{l,e} = softmax(g^{l,e}) \quad (7)$$

where $W^{l,s} \in \mathbb{R}^{1\times d}$ and $W^{l,e} \in \mathbb{R}^{1\times d}$ are trainable weights and softmax is taken over all tokens. Define the extraction loss of the left part as

$$\mathcal{J}_{AE} = -\sum_i y_i^{l,s} \log(p_i^{l,s}) - \sum_i y_i^{l,e} \log(p_i^{l,e}) \quad (8)$$

where $y^{l,s}$ and $y^{l,e}$ are ground truth start and end positions for ATs.

The goal of the right part is to extract all OTs and find the sentiment polarity with respect to a given specific AT. Similarly, we obtain the logits and probabilities for start/end positions

$$g^{r,s} = W^{r,s}h^{r,s}, \quad p^{r,s} = softmax(g^{r,s}) \quad (9)$$

$$g^{r,e} = W^{r,e}h^{r,e}, \quad p^{r,e} = softmax(g^{r,e}) \quad (10)$$

where $W^{r,s} \in \mathbb{R}^{1\times d}$ and $W^{r,e} \in \mathbb{R}^{1\times d}$ are trainable weights and softmax is applied on all tokens. Define the extraction loss of the right part as

$$\mathcal{J}_{AOE} = -\sum_i y_i^{r,s} \log(p_i^{r,s}) - \sum_i y_i^{r,e} \log(p_i^{r,e}) \quad (11)$$

where $y^{r,s}, y^{r,e} \in \mathbb{R}^{(n+2)}$ are true start and end positions for OTs given a specific AT.

In addition, for the right part, we also obtain the sentiment polarity

$$p^r_{cls} = softmax(W^r_{cls}h^r_{cls} + b^r_{cls}) \quad (12)$$

The cross entropy loss for the classification is

$$\mathcal{J}_{SC} = CE(p^r_{cls}, y_{cls}) \quad (13)$$

where $y_{cls} \in \mathbb{R}^3$ represents the true labels for sentiment polarities. Then we want to minimize the final joint training loss

$$\mathcal{J} = \alpha \cdot \mathcal{J}_{AE} + \beta \cdot \mathcal{J}_{SC} + \gamma \cdot \mathcal{J}_{AOE} \quad (14)$$

where $\alpha, \beta, \gamma \in [0, 1]$ are hyper-parameters to control the contributions of objectives.

[4] Note (x_j, a) has all the information needed to determine s. The term o does not bring additional information as it can be implied by (x_j, a), therefore $P(s|x_j, a, o) = P(s|x_j, a)$

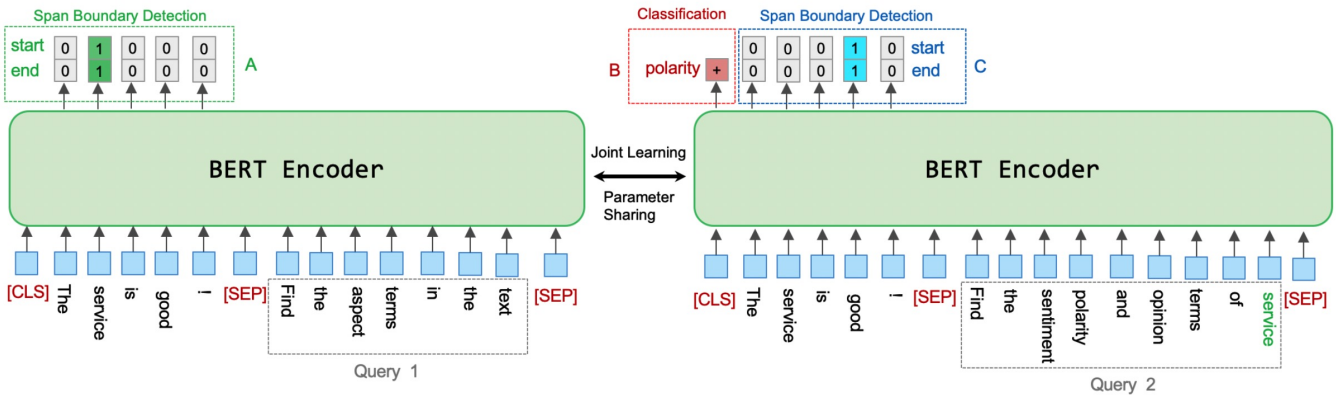

Figure 2: Proposed joint training dual-MRC framework.

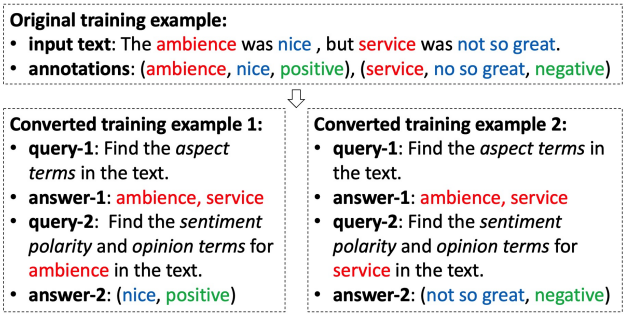

Figure 3: Dataset conversion.

MRC Dataset Conversion

As illustrated in Figure 3, the original triple annotations have to be converted before it is fed into the joint training dual-MRC model. Both MRCs use the input sentence as their contexts. The left MRC is constructed with the query

$$q_1 = \text{“Find the aspect terms in the text.”} \quad (15)$$

Then the answer to the left MRC is all ATs from the text. Given an AT, the right MRC is constructed with the query

$$q_2(AT) = \text{“Find the sentiment polarity and opinion terms for AT in the text.”} \quad (16)$$

The output to the right MRC is all OTs and the sentiment polarity with respect to the given AT. An important problem is that number of right MRCs equals the number of ATs, therefore, the left MRC is *repeated* for that number of times.

Inference Process

For *Triple*, we want to point out some differences between the training process and inference process. During the training process, the ground truth of all ATs are known, then the right MRC can be constructed based on these ATs. Thus, the training process is end-to-end. However, during the inference process, the ATs are the output of the left MRC.

Algorithm 1: The inference Process for Triple Extraction of the Dual-MRC Framework

Input: sentence x
Output: $T = \{(a, o, s)\}$ triples
Initialize $T = \{\}$;
Input x with the query q_1 described in (15) as the left MRC, and output the AT candidates A;
If $A = \{\}$, return T;
for $a_i \in A$ **do**
  Input x with the query q_2 described in (16) as the right MRC, and output the sentiment polarity s and OTs $\{o_j, j = 1, 2, ...\}$;
  $T \leftarrow T \bigcup \{(a_i, o_j, s), j = 1, 2, ...\}$
end
Return T.

Therefore, we inference the two MRCs in a pipeline, as in Algorithm 1.

The inference process of other tasks are similar. The task *AE* uses the span output from the left MRC. *AOE* and *SC* use the span and classification outputs from the right MRC. *AESC* and *Pair* use a combination of them. Please refer to Table 1 for details.

Experiments

Datasets

Original datasets are from the Semeval Challenges(Pontiki et al. 2014, 2015, 2016), where ATs and corresponding sentiment polarities are labeled. We evaluate our framework on three public datasets derived from them.

The first dataset is from (Wang et al. 2017), where labels for opinion terms are annotated. All datasets share a fixed training/test split. The second dataset is from (Fan et al. 2019), where (AT, OT) pairs are labeled. The third dataset is from (Peng et al. 2020) where (AT, OT, SP) triples are labeled. A small number of samples with overlapping ATs and OTs are corrected. Also, 20% of the data from the training set are randomly selected as the validation set. For detailed statistics of the datasets, please refer to the original papers.

	14res				14lap				15res			
	AE	OE	SC	AESC	AE	OE	SC	AESC	AE	OE	SC	AESC
SPAN-BERT	**86.71**	-	71.75	73.68	82.34	-	62.50	61.25	74.63	-	50.28	62.29
IMN-BERT	84.06	85.10	75.67	70.72	77.55	81.00	75.56	61.73	69.90	73.29	70.10	60.22
RACL-BERT	86.38	87.18	81.61	75.42	81.79	79.72	73.91	63.40	73.99	76.00	**74.91**	**66.05**
Dual-MRC	86.60	-	**82.04**	**75.95**	**82.51**	-	**75.97**	**65.94**	**75.08**	-	73.59	65.08

Table 2: Results for *AE*, *SC* and *AESC* on the datasets annotated by (Wang et al. 2017). *OE* is not applicable to our proposed framework. All tasks are evaluated with F1. Baseline results are directly taken from (Chen and Qian 2020). Our model is based on BERT-Large-Uncased. 20% of the data from the training set are randomly selected as the validation set. The results are the average scores of 5 runs with random initialization.

	14res			14lap			15res			16res		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
IOG	82.38	78.25	80.23	73.43	68.74	70.99	72.19	71.76	71.91	84.36	79.08	81.60
LOTN	84.00	**80.52**	82.21	77.08	67.62	72.02	76.61	70.29	73.29	**86.57**	**80.89**	**83.62**
Dual-MRC	**89.79**	78.43	**83.73**	**78.21**	**81.66**	**79.90**	**77.19**	**71.98**	**74.50**	86.07	80.77	83.33

Table 3: Results for *AOE* on the datasets annotated by (Fan et al. 2019). Baseline results are directly taken from (Wu et al. 2020). Our model is based on BERT-Base-Uncased.

Subtasks and Baselines

There exist three research lines in ABSA where each research line with different data annotations, ABSA substasks, baselines and experimental settings. To fairly compare our proposed framework with previous baselines, we should specify them clearly for each research line.

Using the dataset from (Wang et al. 2017), the following baselines were evaluated for *AE*, *OE*, *SC* and *AESC*:

- **SPAN-BERT** (Hu et al. 2019) is a pipeline method for *AESC* which takes BERT as the backbone network. A span boundary detection module is used for *AE*, then followed by a polarity classifier based on span representations for *SC*.
- **IMN-BERT** (He et al. 2019) is an extension of IMN (He et al. 2019) with BERT as the backbone. IMN is a multi-task learning method involving joint training for *AE* and *SC*. A message-passing architecture is introduced in IMN to boost the performance of *AESC*.
- **RACL-BERT** (Chen and Qian 2020) is a stacked multi-layer network based on BERT encoder and is the state of the art method for *AESC*. A Relation propagation mechanism is utilized in RACL to capture the interactions between subtasks (i.e. *AE*, *OE*, *SC*).

Using the dataset from (Fan et al. 2019), the following baselines were evaluated for *AOE*:

- **IOG** (Fan et al. 2019) is the first model proposed to address *AOE*, which adopts six different BLSTMs to extract corresponding opinion terms for aspects given in advance.
- **LOTN** (Wu et al. 2020) is the state of the art method for AOE, which transfer latent opinion information from external sentiment classification datasets to improve the performance.

Using the dataset from (Peng et al. 2020), the following baselines were evaluated for *AESC*, *Pair* and *Triple*:

- **RINANTE** (Dai and Song 2019) is a weakly supervised co-extraction method for *AE* and *OE* which make use of the dependency relations of words in a sentence.
- **CMLA** (Wang et al. 2017) is a multilayer attention network for *AE* and *OE*, where each layer consists of a couple of attentions with tensor operators.
- **Li-unified-R** (Peng et al. 2020) is a modified variant of Li-unified(Li et al. 2019a), which is originally for *AESC* via a unified tagging scheme. Li-unified-R only adapts the original *OE* module for opinion term extraction.
- **Peng-two-stage** (Peng et al. 2020) is a two-stage framework with separate models for different subtasks in ABSA and is the state-of-the-art method for *Triple*.

Model Settings

We use the BERT-Base-Uncased[5] or BERT-Large-Uncased as backbone models for our proposed model depending on the baselines. Please refer to (Devlin et al. 2019) for model details of BERT. We use Adam optimizer with a learning rate of $2e^{-5}$ and warm up over the first 10% steps to train for 3 epochs. The batch size is 12 and a dropout probability of 0.1 is used. The hyperparameters α, β, γ for the final joint training loss in Equation 14 are not sensitive to results, so we fix them as $1/3$ in our experiments. The logit thresholds of heuristic multi-span decoding algorithms (Hu et al. 2019) are very sensitive to results and they are manually tuned on each dataset, and other hyperparameters are kept default. All experiments are conducted on a single Tesla-V100 GPU.

Evaluation Metrics

For all tasks in our experiments, we use the precision (P), recall (R), and F1 scores [6] as evaluation metrics since a pre-

[5]https://github.com/google-research/bert

[6]We use F1 as the metric for aspect-level sentiment classification following (Chen and Qian 2020)

		14res			14lap			15res			16res		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
AESC	RINANTE	48.97	47.36	48.15	41.20	33.20	36.70	46.20	37.40	41.30	49.40	36.70	42.10
	CMLA	67.80	73.69	70.62	54.70	59.20	56.90	49.90	58.00	53.60	58.90	63.60	61.20
	Li-unified-R	73.15	74.44	73.79	66.28	60.71	63.38	64.95	**64.95**	64.95	66.33	**74.55**	70.20
	Peng-two-stage	74.41	73.97	74.19	63.15	61.55	62.34	**67.65**	64.02	**65.79**	**71.18**	72.30	**71.73**
	Dual-MRC	**76.84**	**76.31**	**76.57**	**67.45**	**61.96**	**64.59**	66.84	63.52	65.14	69.18	72.59	70.84
Pair	RINANTE	42.32	51.08	46.29	34.40	26.20	29.70	37.10	33.90	35.40	35.70	27.00	30.70
	CMLA	45.17	53.42	48.95	42.10	46.30	44.10	42.70	46.70	44.60	52.50	47.90	50.00
	Li-unified-R	44.37	73.67	55.34	52.29	52.94	52.56	52.75	61.75	56.85	46.11	64.55	53.75
	Peng-two-stage	47.76	68.10	56.10	50.00	58.47	53.85	49.22	**65.70**	56.23	52.35	70.50	60.04
	Dual-MRC	**76.23**	**73.67**	**74.93**	**65.43**	**61.43**	**63.37**	**72.43**	58.90	**64.97**	**77.06**	**74.41**	**75.71**
Triple	RINANTE	31.07	37.63	34.03	23.10	17.60	20.00	29.40	26.90	28.00	27.10	20.50	23.30
	CMLA	40.11	46.63	43.12	31.40	34.60	32.90	34.40	37.60	35.90	43.60	39.80	41.60
	Li-unified-R	41.44	68.79	51.68	42.25	42.78	42.47	43.34	50.73	46.69	38.19	53.47	44.51
	Peng-two-stage	44.18	62.99	51.89	40.40	47.24	43.50	40.97	**54.68**	46.79	46.76	62.97	53.62
	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**

Table 4: Results for *AESC*, *Pair* and *Triple* on the datasets annotated by (Peng et al. 2020). Baseline results are directly taken from (Peng et al. 2020). Our model is based on BERT-Base-Uncased.

dicted term is correct if it exactly matches a gold term.

Main Results

As mentioned previously, there are three research lines with different datasets, ABSA substasks, baselines and experimental settings. For each research line, we keep the same dataset and experimental setting, and compare our proposed dual-MRC framework with the baselines and present our results in Table 2, Table 3 and Table 4.

First, we compare our proposed method for *AE*, *SC* and *AESC* on the dataset from (Wang et al. 2017). *OE* is not applicable to our proposed framework [7]. Since the pair-wise relations of (AT, OT) are not annotated in this dataset, we use the right part of our model for classification only. 20% of the data from the training set are randomly selected as the validation set. The results are the average scores of 5 runs with random initialization and they are shown in Table 2. We adopt BERT-Large-Uncased as our backbone model since the baselines use it too. All the baselines are BERT based and our results achieve the first or second place comparing to them. Recall that our approach is inspired by SPAN-BERT, which is a strong baseline for extraction tasks. Our results are close to SPAN-BERT in *AE*. However, with the help of MRC, we achieve much better results in *SC* and *AESC*.

Second, we compare our proposed method for *AOE* on the dataset from (Fan et al. 2019), where the pair-wise (AT, OT) relations are annotated. This task can be viewed as a trivial case of our proposed full model. The results are shown in Table 3. BERT-Base-Uncased is used as our backbone model. Although the result for 16res is a little bit lower than LOTN, most of our results significantly outperform the previous baselines. It indicates our model has advantage in matching AT and OT. In particular, our model performs much better than baselines on lap14. It is probably due to the domain difference between the laptop (14lap) comments and the restaurant comments (14res/15res/16res).

[7] If needed, we can train a separate model with the query "*Find the opinion terms in the text.*" for *OE*.

Third, we compare our proposed method for *AESC*, *Pair* and *Triple* on the dataset from (Peng et al. 2020). The full model of our proposed framework is implemented. The results are shown in Table 4. BERT-Base-Uncased is used as our backbone model. Our results significantly outperform the baselines, especially in the precision scores of extraction the pair-wise (AT, OT) relations. Note that Li-unified-R and Peng-two-stage both use the unified tagging schema. For extraction tasks, span based methods outperform the unified tagging schema for extracting terms, probably because determining the start/end positions is easier than determining the label for every token. More precisely, for the unified tagging schema, there are at 7 possible choices for each token, say {B-POS, B-NEU, B-NEG, I-POS, I-NEU, I-NEG, O}, so there are 7^n total choices. For span based methods, there are at 4 possible choices for each token, say {IS-START, NOT-START, IS-END, NOT-END}, then there are $4^n (\ll 7^n)$ total choices. Our proposed method combines MRC and span based extraction, and it has huge improvements for *Pair* and *Triple*.

Analysis on Joint Learning

We give some analysis on the effectiveness of joint learning. The experimental results on the dataset from (Peng et al. 2020) are shown in Table 6. Overall, from the experimental results, adding one or two learning objectives does not affect much in F-1 scores. However, joint learning is more efficient and it can handle more tasks with one single model.

For the task *AESC*, we compare the results with or without the span based extraction output from the right part of our model. By jointly learning to extract the opinion terms for a given aspect, the result of aspect-level sentiment classification is improved a little bit. It makes sense because extracted OTs are useful for identifying the sentiment polarity of the given AT.

For the task *Pair*, we compare the results with or without the classification output from the right part of our model. The F-1 scores for OT extraction decrease a little bit when

Example	Ground Truth	Our model	Peng-two-stage
Rice is too dry, tuna was n't so fresh either.	(Rice, too dry, NEG) (tuna, was n't so fresh, NEG)	(Rice, too dry, NEG) (tuna, was n't so fresh, NEG)	(Rice, too dry, NEG) (tuna, was n't so fresh, NEG), (Rice, was n't so fresh, NEG)✗ (tuna, too dry, NEG)✗
I am pleased with the fast log on, speedy WiFi connection and the long battery life.	(log on, pleased, POS) (log on, fast, POS) (WiFi connection, speedy, POS) (battery life, long, POS)	(log on, pleased, POS) (log on, fast, POS) (WiFi connection, speedy, POS) (WiFi connection, pleased, POS)✗, (battery life, long, POS)	(log, pleased, POS)✗, (log, fast, POS)✗ (WiFi connection, speedy, POS) (battery life, long, POS)
The service was exceptional - sometime there was a feeling that we were served by the army of friendly waiters.	(service, exceptional, POS) (waiters, friendly, POS)	(service, exceptional, POS) (waiters, friendly, POS)	(service, exceptional, POS) (waiters, friendly, POS)

Table 5: Case study of task *Triple*. Wrong predictions are marked with ✗. The three examples are extractly the same as the ones selected by (Peng et al. 2020).

Task	Left	Right		14res	14lap	15res	16res
	e	c	e				
AESC	√	√		76.31	63.95	65.43	69.48
	√	√	√	76.57	64.59	65.14	70.84
Pair	√		√	76.33	65.26	65.21	76.61
	√	√	√	74.93	63.37	64.97	75.71
AE	√			82.80	78.35	78.22	82.16
	√	√	√	82.93	77.31	76.08	81.20

Table 6: Results on the analysis of joint learning for *AESC* and *Pair* on the dataset from (Peng et al. 2020). In the table, the letter e stands for extraction and the letter c stands for classification.

the sentiment classification objective is added. The reason might be that the sentiment polarity can point to multiple OTs in a sentence where some OTs are not paired with the given AT.

Case Study

To validate the effectiveness of our model, we compare our method based on exactly the same three examples in the baseline (Peng et al. 2020) as its source code is not public. The results are shown in Table 5.

The first example shows our MRC based approach performs better in matching AT and OT. Peng's approach matches "*tuna*" and "*too dry*" by mistake while our approach converts the matching problem to a MRC problem. The second example shows the span based extraction method is good at detecting boundaries of entities. Our approach successfully detects "*log on*" while Peng's approach detects "*log*" by mistake. Moreover, the sentiment classification result indicates that our MRC based approach is also good at *SC*.

We plot in Figure 4 the attention matrices from our fine-tuned model between the input text and the query. As we can see, the "*opinion term*" has high attention scores with "*fresh*", and "*sentiment*" has high attention scores with "*food/fresh/hot*". As a result, the queries can capture important information for the task via self-attentions.

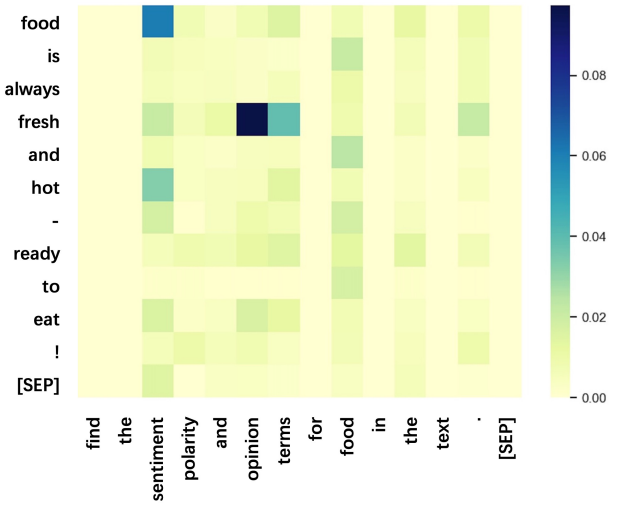

Figure 4: An example of attention matrices for the input text and query.

Conclusions

In this paper, we propose a joint training dual-MRC framework to handle all ABSA subtasks of aspect based sentiment analysis (ABSA) in one shot, where the left MRC is for aspect term extraction and the right MRC is for aspect-oriented opinion term extraction and sentiment classification. The original dataset is converted and fed into dual-MRC to train jointly. For three research lines, experiments are conducted and are compared with different ABSA subtasks and baselines. Experimental results indicate that our proposed framework outperforms all compared baselines.

References

Chen, P.; Sun, Z.; Bing, L.; and Yang, W. 2017. Recurrent Attention Network on Memory for Aspect Sentiment Anal-

ysis. In *EMNLP*, 452–461.

Chen, Z.; and Qian, T. 2020. Relation-Aware Collaborative Learning for Unified Aspect-Based Sentiment Analysis. In *ACL*, 3685–3694.

Dai, H.; and Song, Y. 2019. Neural aspect and opinion term extraction with mined rules as weak supervision. In *ACL*, 5268–5277.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Du, C.; Sun, H.; Wang, J.; Qi, Q.; Liao, J.; Xu, T.; and Liu, M. 2019. Capsule Network with Interactive Attention for Aspect-Level Sentiment Classification. In *EMNLP*, 5488–5497.

Fan, C.; Gao, Q.; Du, J.; Gui, L.; Xu, R.; and Wong, K. 2018. Convolution-based Memory Network for Aspect-based Sentiment Analysis. In *SIGIR*, 1161–1164.

Fan, Z.; Wu, Z.; Dai, X.; Huang, S.; and Chen, J. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *NAACL-HLT*, 2509–2518.

Gu, S.; Zhang, L.; Hou, Y.; and Song, Y. 2018. A Position-aware Bidirectional Attention Network for Aspect-level Sentiment Analysis. In *COLING*, 774–784.

He, R.; Lee, W. S.; Ng, H. T.; and Dahlmeier, D. 2017. An Unsupervised Neural Attention Model for Aspect Extraction. In *ACL*, 388–397.

He, R.; Lee, W. S.; Tou Ng, H.; and Dahlmeier, D. 2019. An Interactive Multi-Task Learning Network for End-to-End Aspect-Based Sentiment Analysis. In *ACL*, 504–515.

Hu, M.; and Liu, B. 2004. Mining and summarizing customer reviews. In *KDD*, 168–177.

Hu, M.; Peng, Y.; Huang, Z.; Li, D.; and Lv, Y. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *ACL*, 537–546.

Huang, B.; and Carley, K. M. 2018. Parameterized Convolutional Neural Networks for Aspect Level Sentiment Classification. In *EMNLP*, 1091–1096.

Lee, K.; Kwiatkowski, T.; Parikh, A. P.; and Das, D. 2016. Learning Recurrent Span Representations for Extractive Question Answering. *CoRR* abs/1611.01436.

Li, K.; Chen, C.; Quan, X.; Ling, Q.; and Song, Y. 2020. Conditional Augmentation for Aspect Term Extraction via Masked Sequence-to-Sequence Generation. In *ACL*, 7056–7066.

Li, X.; Bing, L.; Lam, W.; and Shi, B. 2018a. Transformation Networks for Target-Oriented Sentiment Classification. In *ACL*, 946–956.

Li, X.; Bing, L.; Li, P.; and Lam, W. 2019a. A unified model for opinion target extraction and target sentiment prediction. In *AAAI*, 6714–6721.

Li, X.; Bing, L.; Li, P.; Lam, W.; and Yang, Z. 2018b. Aspect Term Extraction with History Attention and Selective Transformation. In *IJCAI*, 4194–4200.

Li, X.; Bing, L.; Zhang, W.; and Lam, W. 2019b. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. In *W-NUT@EMNLP*, 34–41.

Ma, D.; Li, S.; and Wang, H. 2018. Joint Learning for Targeted Sentiment Analysis. In *EMNLP*, 4737–4742.

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

Ma, D.; Li, S.; Zhang, X.; and Wang, H. 2017. Interactive Attention Networks for Aspect-Level Sentiment Classification. In *IJCAI*, 4068–4074.

Peng, H.; Xu, L.; Bing, L.; Huang, F.; Lu, W.; and Si, L. 2020. Knowing What, How and Why: A Near Complete Solution for Aspect-Based Sentiment Analysis. In *AAAI*, 8600–8607.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; AL-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; Hoste, V.; Apidianaki, M.; Tannier, X.; Loukachevitch, N.; Kotelnikov, E.; Bel, N.; Jiménez-Zafra, S. M.; and Eryiğit, G. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *SemEval@NAACL- HLT*, 19–30.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Manandhar, S.; and Androutsopoulos, I. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *SemEval@NAACL-HLT)*, 486–495.

Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *SemEval@COLING*, 27–35.

Rosenfeld, A.; and Thurston, M. 1971. Edge and curve detection for visual scene analysis. *IEEE Transactions on computers* 100(5): 562–569.

Ruder, S.; Ghaffari, P.; and Breslin, J. G. 2016. A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis. In *EMNLP*, 999–1005.

Sun, C.; Huang, L.; and Qiu, X. 2019. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In *NAACL-HLT*, 380–385.

Tang, D.; Qin, B.; Feng, X.; and Liu, T. 2016. Effective LSTMs for target-dependent sentiment classification. In *COLING*, 3298–3307.

Tang, D.; Qin, B.; and Liu, T. 2016. Aspect Level Sentiment Classification with Deep Memory Network. In *EMNLP*, 214–224.

Wang, W.; and Pan, S. J. 2018. Recursive Neural Structural Correspondence Network for Cross-domain Aspect and Opinion Co-Extraction. In *ACL*, 2171–2181.

Wang, W.; Pan, S. J.; Dahlmeier, D.; and Xiao, X. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *AAAI*, 3316–3322.

Wang, Y.; Huang, M.; Zhu, X.; and Zhao, L. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *EMNLP*, 606–615.

Wu, Z.; Zhao, F.; Dai, X.-Y.; Huang, S.; and Chen, J. 2020. Latent Opinions Transfer Network for Target-Oriented Opinion Words Extraction. In *AAAI*, 9298–9305.

Xu, H.; Liu, B.; Shu, L.; and Yu, P. S. 2018. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. In *ACL*, 592–598.

Yang, M.; Tu, W.; Wang, J.; Xu, F.; and Chen, X. 2017. Attention Based LSTM for Target Dependent Sentiment Classification. In *AAAI*, 5013–5014.

Yu, J.; Jiang, J.; and Xia, R. 2019. Global Inference for Aspect and Opinion Terms Co-Extraction Based on Multi-Task Neural Networks. *IEEE ACM Trans. Audio Speech Lang. Process.* 27(1): 168–177.

Zhang, M.; Zhang, Y.; and Vo, D. 2016. Gated Neural Networks for Targeted Sentiment Analysis. In *AAAI*, 3087–3093.

Zhao, H.; Huang, L.; Zhang, R.; Lu, Q.; and Xue, H. 2020. SpanMlt: A Span-based Multi-Task Learning Framework for Pair-wise Aspect and Opinion Terms Extraction. In *ACL*, 3239–3248.

Zhou, Y.; Huang, L.; Guo, T.; Han, J.; and Hu, S. 2019. A Span-based Joint Model for Opinion Target Extraction and Target Sentiment Classification. In *IJCAI*, 5485–5491.