Bidirectional Machine Reading Comprehension for Aspect Sentiment Triplet Extraction

Shaowei Chen[1], **Yu Wang**[1], **Jie Liu**[1,2*], **Yuelin Wang**[1]
[1]College of Artificial Intelligence, Nankai University, Tianjin, China
[2]Cloopen Research, Beijing, China
{shaoweichen, yuwang17, 1711496}@mail.nankai.edu.cn
jliu@nankai.edu.cn

Abstract

Aspect sentiment triplet extraction (ASTE), which aims to identify aspects from review sentences along with their corresponding opinion expressions and sentiments, is an emerging task in fine-grained opinion mining. Since ASTE consists of multiple subtasks, including opinion entity extraction, relation detection, and sentiment classification, it is critical and challenging to appropriately capture and utilize the associations among them. In this paper, we transform ASTE task into a multi-turn machine reading comprehension (MTMRC) task and propose a bidirectional MRC (BMRC) framework to address this challenge. Specifically, we devise three types of queries, including *non-restrictive extraction* queries, *restrictive extraction* queries and *sentiment classification* queries, to build the associations among different subtasks. Furthermore, considering that an aspect sentiment triplet can derive from either an aspect or an opinion expression, we design a bidirectional MRC structure. One direction sequentially recognizes aspects, opinion expressions, and sentiments to obtain triplets, while the other direction identifies opinion expressions first, then aspects, and at last sentiments. By making the two directions complement each other, our framework can identify triplets more comprehensively. To verify the effectiveness of our approach, we conduct extensive experiments on four benchmark datasets. The experimental results demonstrate that BMRC achieves state-of-the-art performances.

Introduction

Fine-grained opinion mining is an important field in natural language processing (NLP). It comprises various tasks, such as aspect term extraction (ATE) (Liu, Xu, and Zhao 2012; Xu et al. 2018; Li et al. 2018; Ma et al. 2019), opinion term extraction (OTE) (Fan et al. 2019; Wu et al. 2020b), and aspect-level sentiment classification (ASC) (Ma et al. 2017; Sun et al. 2019). Existing studies generally solve these tasks individually or couple two of them as aspect and opinion terms co-extraction task (Liu, Xu, and Zhao 2015; Wang et al. 2016, 2017; Dai and Song 2019), aspect term-polarity co-extraction task (Luo et al. 2019; Li et al. 2019a), and aspect-opinion pair extraction task (Chen et al. 2020; Zhao et al. 2020). However, none of these studies can identify aspects, opinion expressions, and sentiments in a complete solution. To deal with this problem, the latest literature (Peng et al. 2020) presents aspect sentiment triplet extraction (ASTE) task, which aims to identify triplets such as (*food, delicious, positive*) in Figure 1.

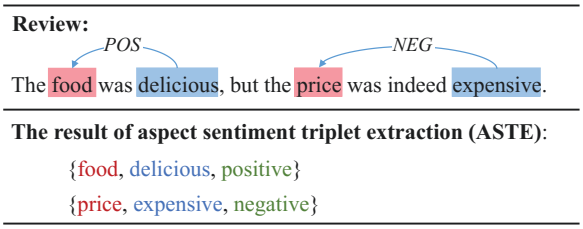


Figure 1: An example of ASTE task. The aspects, opinion expressions, and sentiments are marked with red, blue, and green, respectively.

Although these studies have achieved great progress, there are still several challenges existing in fine-grained opinion mining. **First**, aspects and opinion expressions generally appear together in a review sentence and have explicit corresponding relations. Hence, how to adequately learn the association between ATE and OTE and make them mutually beneficial is a challenge. **Second**, the corresponding relations between aspects and opinion expressions can be complicated, such as one-to-many, many-to-one, and even overlapped and embedded. Thus, it is challenging to flexibly and exactly detect these relations. **Third**, each review sentence may contain multiple sentiments. For example, given the review in Figure 1, the sentiments of *price* and *food* are negative and positive, respectively. These sentiments are generally guided by the corresponding relations between aspects and opinion expressions. Thus, how to properly introduce these relations to sentiment classification task is another challenge.

To address the aforementioned challenges, we deal with ASTE task and formalize it as a machine reading comprehension (MRC) task. Given a query and a context, MRC task aims to capture the interaction between them and extract specific information from the context as the answer. Different from the general MRC task, we further devise multi-turn queries to identify aspect sentiment triplets due to the complexity of ASTE. Specially, we define this formalization as

*Corresponding author.

multi-turn machine reading comprehension (MTMRC) task. By introducing the answers to the previous turns into the current turn as prior knowledge, the associations among different subtasks can be effectively learned. For example, given the review in Figure 1, we can identify the aspect *food* in the first turn and introduce it into the second turn query *What opinions given the aspect food?* to jointly identify the opinion expression *delicious* and the relation between *food* and *delicious*. Then, we can use the aspect *food* and the opinion expression *delicious* as the prior knowledge of the third turn query to predict that the sentiment of *food* is *positive*. According to these turns, we can flexibly capture the association between ATE and OTE, detect complex relations between opinion entities[1], and utilize these relations to guide sentiment classification.

Based on MTMRC, we propose a bidirectional machine reading comprehension (BMRC) framework [2] in this paper. Specifically, we design three-turn queries to identify aspect sentiment triplets. In the first turn, we design *non-restrictive extraction* queries to locate the first entity of each aspect-opinion pair. Then, *restrictive extraction* queries are designed for the second turn to recognize the other entity of each pair based on the previously extracted entity. In the third turn, *sentiment classification* queries are proposed to predict aspect-oriented sentiments based on the extracted aspects and their corresponding opinion expressions. Since there is no intrinsic order when extracting aspects and opinion expressions, we further propose a bidirectional structure to recognize the aspect-opinion pairs. In one direction, we first utilize a non-restrictive extraction query to identify aspects such as $\{food, price\}$ in Figure 2. Then, given the specific aspect like *food*, the second-turn query looks for its corresponding opinion expressions such as $\{delicious\}$ in Figure 2 via a restrictive extraction query. Similarly, the other direction extracts opinion expressions and their corresponding aspects in a reversed order. To verify the effectiveness of BMRC, we make comprehensive analyses on four benchmark datasets. The experimental results show that our approach substantially outperforms the existing methods. In summary, our contributions are three-fold:

- We formalize aspect sentiment triplet extraction (ASTE) task as a multi-turn machine reading comprehension (MTMRC) task. Based on this formalization, we can gracefully identify aspect sentiment triplets in a unified framework.
- We propose a bidirectional machine reading comprehension (BMRC) framework. By devising three-turn queries, our model can effectively build the associations among opinion entity extraction, relation detection, and sentiment classification.
- We conduct extensive experiments on four benchmark datasets. The experimental results demonstrate that our model achieves state-of-the-art performances.

[1]In this paper, we briefly note aspects and opinion expressions as opinion entities.

[2]Code is available at: https://github.com/NKU-IIPLab/BMRC.

Related Work

In this paper, we transform the aspect sentiment triplet extraction task into a multi-turn machine reading comprehension task. Thus, we introduce the related work from two parts, including fine-grained opinion mining and machine reading comprehension.

Fine-grained Opinion Mining

Fine-grained opinion mining consists of various tasks, including aspect term extraction (ATE) (Wang et al. 2016; He et al. 2017; Li et al. 2018; Xu et al. 2018; Li and Lam 2017), opinion term extraction (OTE) (Liu, Joty, and Meng 2015; Poria, Cambria, and Gelbukh 2016; Xu et al. 2018; Wu et al. 2020b), aspect-level sentiment classification (ASC) (Dong et al. 2014; Tang, Qin, and Liu 2016; Li et al. 2019c; He et al. 2018; Hazarika et al. 2018; Nguyen and Shirai 2015; Wang et al. 2018), etc. The studies solve these tasks individually and ignore the dependency between them.

To explore the interactions between different tasks, recent studies gradually focus on the joint tasks such as aspect term-polarity co-extraction (He et al. 2019; Mitchell et al. 2013; Li and Lu 2017; Li et al. 2019a), aspect and opinion terms co-extraction (Liu, Xu, and Zhao 2015; Wang et al. 2016, 2017; Dai and Song 2019), aspect category and sentiment classification (Hu et al. 2019), and aspect-opinion pair extraction (Chen et al. 2020; Zhao et al. 2020). Besides, there are also a lot of studies (Chen and Qian 2020; He et al. 2019) solving multiple tasks with a multi-task learning network. However, none of these studies could identify aspects, opinion expressions and sentiments in a unified framework. To deal with this issue, Peng et al. (2020) proposed a two-stage framework to solve aspect sentiment triplet extraction (ASTE) task, which aims to extract triplets of aspects, opinion expressions and sentiments. However, the model suffers from error propagation due to its two-stage framework. Besides, separating the extraction and pairing of opinion entities means that the associations between different tasks are still not adequately considered.

Machine Reading Comprehension

Machine reading comprehension (MRC) aims to answer specific queries based on a given context. Recent researches have proposed various effective architectures for MRC, which adequately learn the interaction between the query and context. For example, BiDAF (Seo et al. 2017) employs a RNN-based sequential framework to encode queries and passages, while QANet (Yu et al. 2018) employs both convolution and self-attention. Several MRC systems (Peters et al. 2018; Radford et al. 2019) adopt context-aware embedding as well and obtain comparable results, especially BERT-based MRC model (Devlin et al. 2019).

Recently, there is a tendency to apply MRC on many NLP tasks, including named entity recognition (Li et al. 2020), entity relation extraction (Li et al. 2019b; Levy et al. 2017), and summarization (McCann et al. 2018), etc. Due to the advantages of MRC framework, we naturally transform ASTE into a multi-turn MRC task to better construct the associations among aspects, opinions, aspect-opinion relations and

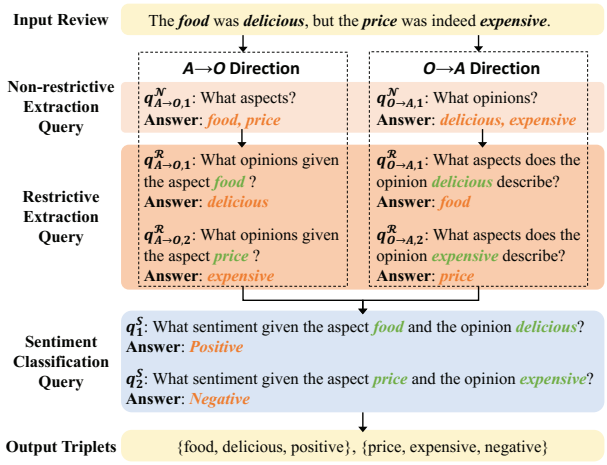


Figure 2: The bidirectional machine reading comprehension (BMRC) framework.

sentiments through well-designed queries. Different from the existing methods (Li et al. 2020, 2019b), we innovatively propose a bidirectional framework, which can identity triplets more comprehensively by making the two directions complement each other. This framework can be further extended to other tasks such as entity relation extraction.

Problem Formulation

Given a review sentence $X = \{x_1, x_2, ..., x_N\}$ with N tokens, ASTE task aims to identify the collection of triplets $T = \{(a_i, o_i, s_i)\}_{i=1}^{|T|}$, where a_i, o_i, s_i, and $|T|$ represent the aspect, the opinion expression, the sentiment, and the number of triplets[3], respectively.

To formalize ASTE task as a multi-turn MRC task, we construct three types of queries[4], including non-restrictive extraction queries $Q^{\mathcal{N}} = \{q_i^{\mathcal{N}}\}_{i=1}^{|Q^{\mathcal{N}}|}$, restrictive extraction queries $Q^{\mathcal{R}} = \{q_i^{\mathcal{R}}\}_{i=1}^{|Q^{\mathcal{R}}|}$ and sentiment classification queries $Q^{\mathcal{S}} = \{q_i^{\mathcal{S}}\}_{i=1}^{|Q^{\mathcal{S}}|}$.

Concretely, in the first turn, each non-restrictive extraction query $q_i^{\mathcal{N}}$ aims to extract either aspects $A = \{a_i\}_{i=1}^{|A|}$ or opinion expressions $O = \{o_i\}_{i=1}^{|O|}$ from the review sentence to trigger aspect-opinion pairs. In the second turn, given the opinion entities recognized by $q_i^{\mathcal{N}}$, each restrictive extraction query $q_i^{\mathcal{R}}$ aims to identify either the corresponding aspects or the corresponding opinion expressions. To be more specific, given each aspect a_i extracted by $q_i^{\mathcal{N}}$, the restrictive extraction query extracts its corresponding opinion expressions $O_{a_i} = \{o_{a_i,j}\}_{j=1}^{|O_{a_i}|}$. In the final turn, each sentiment classification query $q_i^{\mathcal{S}}$ predicts the sentiment $s_{a_i} \in \{\text{Positive}, \text{Negative}, \text{Neutral}\}$ for each aspect a_i.

[3]The $|*|$ represents the number of elements in the collection $*$.

[4]We use superscripts $\mathcal{N}$, $\mathcal{R}$, and $\mathcal{S}$ to denote the query types.

Methodology

Framework

To deal with ASTE task, we propose a bidirectional machine reading comprehension (BMRC) framework. The overall framework is illustrated in Figure 2. Concretely, we first design non-restrictive extraction queries and restrictive extraction queries to extract aspect-opinion pairs. Considering that each pair can be triggered by an aspect or an opinion expression, we further construct a bidirectional structure. In one direction, the aspects are first extracted via non-restrictive extraction queries, and then the corresponding opinion expressions for each aspect are identified via restrictive extraction queries. We define the above process as A→O direction. Similarly, in O→A direction, the framework recognizes opinion expressions and their corresponding aspects in a reversed order. After that, we design sentiment classification queries to predict the sentiment polarity for each aspect. Furthermore, our model jointly learns to answer the above queries to make them mutually beneficial. Besides, we adopt BERT as the encoding layer for richer semantics representations. During inference, the model fuses the answers to different queries and forms the triplets.

Query Construction

In BMRC, we adopt a template-based approach to construct queries. Specifically, we first design the non-restrictive extraction query and the restrictive extraction query in the A→O direction as follows:

- **A→O non-restrictive extraction query** $q_{A\to O}^{\mathcal{N}}$: We design query '*What aspects?*' to extract the collection of aspects $A = \{a_i\}_{i=1}^{|A|}$ from the given review sentence X.
- **A→O restrictive extraction query** $q_{A\to O}^{\mathcal{R}}$: We design query '*What opinions given the aspect a_i?*' to extract the corresponding opinions $O_{a_i} = \{o_{a_i,j}\}_{j=1}^{|O_{a_i}|}$ for each aspect a_i and form aspect-opinion pairs.

Reversely, the O→A direction extraction queries are constructed as follows:

- **O→A non-restrictive extraction query** $q_{O\to A}^{\mathcal{N}}$: We use query '*What opinions?*' to extract the collection of opinion expressions $O = \{o_i\}_{i=1}^{|O|}$.
- **O→A restrictive extraction query** $q_{O\to A}^{\mathcal{R}}$: We design query '*What aspect does the opinion o_i describe?*' to recognize the corresponding aspects $A_{o_i} = \{a_{o_i,j}\}_{j=1}^{|A_{o_i}|}$ for each opinion expression o_i.

With the above queries, opinion entity extraction and relation detection are naturally fused, and the dependency between them is gracefully learned via the restrictive extraction queries. Then, we devise sentiment classification queries to classify the aspect-oriented sentiments as follows:

- **Sentiment Classification query** $q^{\mathcal{S}}$: We design query '*What sentiment given the aspect a_i and the opinion $o_{a_i,1}/.../o_{a_i,|O_{a_i}|}$?*' to predict sentiment polarity s_{a_i} for each aspect a_i.

With sentiment classification queries, the semantics of aspects and their corresponding opinion expressions can be adequately considered during sentiment prediction.

Encoding Layer

Given the review sentence $X = \{x_1, x_2, ..., x_N\}$ with N tokens and each query $q_i = \{q_{i,1}, q_{i,2}, ..., q_{i,|q_i|}\}$ with $|q_i|$ tokens, the encoding layer learns the context representation for each token. Inspired by the successful practice on many NLP tasks, we adopt BERT as the encoder. Formally, we first concatenate the query q_i and the review sentence X to obtain the combined input $I = \{[\text{CLS}], q_{i,1}, q_{i,2}, ..., q_{i,|q_i|}, [\text{SEP}], x_1, x_2, ..., x_N\}$, where [CLS] and [SEP] are the beginning token and the segment token. The initial representation $\mathbf{e}_i$ for each token is constructed by summing its word embedding $\mathbf{e}_i^w$, position embedding $\mathbf{e}_i^p$, and segment embedding $\mathbf{e}_i^g$. Then, BERT is used to encode the initial representation sequence $E = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{|q_i|+N+2}\}$ as the hidden representation sequence $H = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{|q_i|+N+2}\}$ with the stacked Transformer blocks.

Answer Prediction

Answer for Extraction Query For non-restrictive and restrictive extraction queries, the answers could be multiple opinion entities extracted from the review sentence X. For example, given the review in Figure 2, the aspects *food* and *price* should be extracted as the answer to the A→O non-restrictive extraction query $q^{\mathcal{N}}_{A\to O,1}$. Thus, we utilize two binary classifiers to predict the answer spans. Specifically, based on the hidden representation sequence H, one classifier predicts whether each token x_i is the start position of the answer or not, and another predicts the possibility that each token is the end position:

$$p\left(y_i^{start}|x_i, q\right) = \text{softmax}(\mathbf{h}_{|q|+2+i}W_s), \quad (1)$$

$$p\left(y_i^{end}|x_i, q\right) = \text{softmax}(\mathbf{h}_{|q|+2+i}W_e), \quad (2)$$

where $W_s \in \mathbf{R}^{d_h \times 2}$ and $W_e \in \mathbf{R}^{d_h \times 2}$ are model parameters, d_h denotes the dimension of hidden representations in H, and $|q|$ is the query length.

Answer for Sentiment Classification Query Following existing work (Devlin et al. 2019), the answer to sentiment classification query is predicted with the hidden representation of [CLS]. Formally, we append a three-class classifier to BERT for predicting the sentiment $y^{\mathcal{S}}$ as follows:

$$p\left(y^{\mathcal{S}}|X, q\right) = \text{softmax}(\mathbf{h}_1 W_c), \quad (3)$$

where $W_c \in \mathbf{R}^{d_h \times 3}$ is model parameter.

Joint Learning

To jointly learn the subtasks in ASTE and make them mutually beneficial, we fuse the loss functions of different queries. For non-restrictive extraction queries in both two directions, we minimize the cross-entropy loss as follows:

$$\mathcal{L}_{\mathcal{N}} = -\sum_{i=1}^{|Q^{\mathcal{N}}|}\sum_{j=1}^{N}[p\left(y_j^{start}|x_j, q_i^{\mathcal{N}}\right)\log\hat{p}\left(y_j^{start}|x_j, q_i^{\mathcal{N}}\right) + p\left(y_j^{end}|x_j, q_i^{\mathcal{N}}\right)\log\hat{p}\left(y_j^{end}|x_j, q_i^{\mathcal{N}}\right)], \quad (4)$$

where $p(*)$ represents the gold distribution, and $\hat{p}(*)$ denotes the predicted distribution.

Similarly, the loss of restrictive extraction queries in both two directions is calculated as follows:

$$\mathcal{L}_{\mathcal{R}} = -\sum_{i=1}^{|Q^{\mathcal{R}}|}\sum_{j=1}^{N}[p\left(y_j^{start}|x_j, q_i^{\mathcal{R}}\right)\log\hat{p}\left(y_j^{start}|x_j, q_i^{\mathcal{R}}\right) + p\left(y_j^{end}|x_j, q_i^{\mathcal{R}}\right)\log\hat{p}\left(y_j^{end}|x_j, q_i^{\mathcal{R}}\right)]. \quad (5)$$

For the sentiment classification queries, we minimize the cross-entropy loss function as follows:

$$\mathcal{L}_{\mathcal{S}} = -\sum_{i=1}^{|Q^{\mathcal{S}}|} p\left(y^{\mathcal{S}}|X, q_i^{\mathcal{S}}\right)\log\hat{p}\left(y^{\mathcal{S}}|X, q_i^{\mathcal{S}}\right). \quad (6)$$

Then, we combine the above loss functions to form the loss objective of the entire model:

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathcal{N}} + \mathcal{L}_{\mathcal{R}} + \mathcal{L}_{\mathcal{S}}. \quad (7)$$

The optimization problem in Eq.(7) can be solved by any gradient descent approach. In this paper, we adopt the AdamW (Loshchilov and Hutter 2017) approach.

Inference

During inference, we fuse the answers to different queries to obtain triplets. Specifically, in the A→O direction, the non-restrictive extraction query $q^{\mathcal{N}}_{A\to O}$ first identifies the aspect collection $A = \{a_1, a_2, ..., a_{|A|}\}$ with $|A|$ aspects. For each predicted aspect a_i, the A→O restrictive query $q^{\mathcal{R}}_{A\to O,i}$ recognizes the corresponding opinion expression collection and obtains the set of predicted aspect-opinion pairs $V_{A\to O} = [(a_k, o_k)]_{k=1}^{K}$ in the A→O direction. Similarly, in the O→A direction, the model identifies the set of aspect-opinion pairs $V_{O\to A} = [(a_l, o_l)]_{l=1}^{L}$ in a reversed order. Then, we combine $V_{A\to O}$ and $V_{O\to A}$ as follows:

$$V = V' \cup \{(a, o) \mid (a, o) \in V'', p(a, o) > \delta\}, \quad (8)$$

$$p(a, o) = \begin{cases} p(a)\,p(o|a) & if\ (a, o) \in V_{A\to O} \\ p(o)\,p(a|o) & if\ (a, o) \in V_{O\to A} \end{cases}, \quad (9)$$

where V' and V'' denote the intersection and difference set of $V_{A\to O}$ and $V_{O\to A}$, respectively. Each aspect-opinion pair in V'' is valid only if its probability $p(a, o)$ is higher than the given threshold δ. The probability of each opinion entity is calculated by multiplying the probabilities of its start and end positions.

Finally, we construct sentiment classification query $q_i^{\mathcal{S}}$ to predict the sentiment s_{a_i} of each aspect a_i. Based on these, the triplet collection $T = [(a_i, o_i, s_i)]_{i=1}^{|T|}$ can be obtained.

Experiments

Datasets

To verify the effectiveness of our proposed approach, we conduct experiments on four benchmark datasets[5] from

[5]https://github.com/xuuuluuu/SemEval-Triplet-data

Datasets	Train		Dev		Test	
	#S	#T	#S	#T	#S	#T
14-Lap (Pontiki et al. 2014)	920	1265	228	337	339	490
14-Res (Pontiki et al. 2014)	1300	2145	323	524	496	862
15-Res (Pontiki et al. 2015)	593	923	148	238	318	455
16-Res (Pontiki et al. 2016)	842	1289	210	316	320	465

Table 1: Statistics of datasets. #S and #T denote the number of sentences and triplets, respectively.

the SemEval ABSA Challenges (Pontiki et al. 2014, 2015, 2016) and list the statistics of these datasets in Table 1. Specifically, the golden annotations for opinion expressions and relations are derived from Fan et al. (2019). And we split the datasets as Peng et al. (2020) did.

Experimental Settings

For the encoding layer, we adopt the **BERT-base** (Devlin et al. 2019) model with 12 attention heads, 12 hidden layers and the hidden size of 768, resulting into 110M pretrained parameters. During training, we use AdamW (Loshchilov and Hutter 2017) for optimization with weight decay 0.01 and warmup rate 0.1. The learning rate for training classifiers and the fine-tuning rate for BERT are set to 1e-3 and 1e-5 respectively. Meanwhile, we set batch size to 4 and dropout rate to 0.1. According to the triplet extraction F_1-score on the development sets, the threshold δ is manually tuned to 0.8 in bound $[0, 1)$ with step size set to 0.1. We run our model on a Tesla V100 GPU and train our model for 40 epochs in about 1.5h.

Evaluation

To comprehensively measure the performances of our model and the baselines, we use *Precision*, *Recall*, and *F_1-score* to evaluate the results on four subtasks, including aspect term and sentiment co-extraction, opinion term extraction, aspect-opinion pair extraction, and triplet extraction. For reproducibility, we report the testing results averaged over 5 runs with different random seeds. At each run, we select the testing results when the model achieves the best performance on the development set.

Baselines

To demonstrate the effectiveness of BMRC, we compare our model with the following baselines:

- **TSF** (Peng et al. 2020) is a two-stage pipeline model for ASTE. In the first stage, TSF extracts both aspect-sentiment pairs and opinion expressions. In the second stage, TSF pairs up the extraction results into triplets via an relation classifier.
- **RINANTE+** adopts RINANTE (Dai and Song 2019) with additional sentiment tags as the first stage model to joint extract aspects, opinion expressions, and sentiments. Then, it adopts the second stage of TSF to detect the corresponding relations between opinion entities.
- **Li-unified-R+** jointly identifies aspects and their sentiments with Li-unified (Li et al. 2019a). Meanwhile, it predicts opinion expressions with an opinion-enhanced component at the first stage. Then, it also uses the second stage of TSF to predicts relations.
- **RACL+R** first adopts RACL (Chen and Qian 2020) to identify the aspects, opinion expressions, and sentiments. Then, we construct the query 'Matched the aspect a_i and the opinion expression o_j?' to detect the relations. Note that RACL is also based on BERT.

Results

The experimental results are shown in Table 2. According to the results, our model achieves state-of-the-art performances on all datasets. Although the improvements on aspect term and sentiment co-extraction and opinion term extraction are slight, our model significantly surpasses the baselines by an average of 5.14% F_1-score on aspect-opinion pair extraction and an average of 9.58% F_1-score on triplet extraction. The results indicate that extracting opinion entities and relations in pipeline will lead to severe error accumulation. By utilizing the BMRC framework, our model effectively fuses and simplifies the tasks of ATE, OTE, and relation detection, and avoids the above issue. It is worth noting that the increase in precision contributes most to the boost of F1-score, which shows that the predictions of our model own higher reliability than those baselines. Besides, RACL+R outperforms than other baselines because BERT can learn richer context semantics. TSF and Li-unified-R+ achieve better performances than RINANTE+ because TSF and Li-unified-R+ introduce complex mechanisms to solve the issue of sentiment contradiction brought by the unified tagging schema. Different from those approaches, our model gracefully solves this issue by transforming ASTE into a multi-turn MRC task.

Considering that the datasets released by Peng et al. (2020) remove the cases that one opinion expression corresponds to multiple aspects, we also conduct experiments on AFOE datasets [6] (Wu et al. 2020a) and report the results in Table 3. The AFOE datasets, which retains the above cases, are also constructed based on the datasets of Fan et al. (2019) and the original SemEval ABSA Challenges. And we further compare our model with two baselines, including IMN+IOG and GTS[7] (Wu et al. 2020a) . Specifically, IMN+IOG is a pipeline model which utilizes the interactive multi-task learning network (IMN) (He et al. 2019) as the first stage model to identity the aspects and their sentiments. Then, IMN+IOG use the Inward-Outward LSTM (Fan et al. 2019) as the second stage model to extract the aspect-oriented opinion expressions. And GTS is a latest model which proposes a grid tagging schema to identify the aspect sentiment triplets in an end-to-end way. Particularly, GTS also utilizes BERT as the encoder and designs an inference strategy to exploit mutual indication between different opinion factors. According to the results, our model and GTS significantly outperform IMN+IOG because the joint methods can solve the error propagation problem. Com-

[6] https://github.com/NJUNLP/GTS

[7] It worth noting that this paper has not been published when we submit our paper to AAAI 2021.

Evaluation	Models	14-Lap				14-Res				15-Res				16-Res			
		A-S	O	P	T	A-S	O	P	T	A-S	O	P	T	A-S	O	P	T
Precision	TSF	63.15	78.22	50.00	40.40	76.60	84.72	47.76	44.18	67.65	78.07	49.22	40.97	71.18	81.09	52.35	46.76
	RINANRTE+	41.20	78.20	34.40	23.10	48.97	81.06	42.32	31.07	46.20	77.40	37.10	29.40	49.40	75.00	35.70	27.10
	Li-unified-R+	66.28	76.62	52.29	42.25	73.15	81.20	44.37	41.44	64.95	79.18	52.75	43.34	66.33	79.84	46.11	38.19
	RACL+R	59.75	77.58	54.22	41.99	75.57	82.28	73.58	62.64	68.35	76.25	67.89	55.45	68.53	82.52	72.77	60.78
	Ours	**72.73**	**84.67**	**74.11**	**65.12**	**77.74**	**87.22**	**76.91**	**71.32**	**72.41**	**82.99**	**71.59**	**63.71**	**73.69**	**85.31**	**76.08**	**67.74**
Recall	TSF	61.55	71.84	58.47	47.24	67.84	80.39	68.10	62.99	64.02	78.07	65.70	54.68	72.30	86.67	70.50	62.97
	RINANRTE+	33.20	62.70	26.20	17.60	47.36	72.05	51.08	37.63	37.40	57.00	33.90	26.90	36.70	42.40	27.00	20.50
	Li-unified-R+	60.71	74.90	52.94	42.78	74.44	83.18	73.67	68.79	64.95	75.88	61.75	50.73	74.55	86.88	64.55	53.47
	RACL+R	**68.90**	**81.22**	**66.94**	51.84	**82.23**	**90.49**	67.87	57.77	**70.72**	**83.96**	63.74	52.53	**78.52**	**91.40**	71.83	60.00
	Ours	62.59	67.18	61.92	**54.41**	75.10	82.90	**75.59**	**70.09**	62.63	73.23	**65.89**	**58.63**	72.69	83.01	**76.99**	**68.56**
F_1-score	TSF	62.34	74.84	53.85	43.50	71.95	82.45	56.10	51.89	65.79	78.02	56.23	46.79	71.73	83.73	60.04	53.62
	RINANRTE+	36.70	69.60	29.70	20.00	48.15	76.29	46.29	34.03	41.30	65.70	35.40	28.00	42.10	54.10	30.70	23.30
	Li-unified-R+	63.38	75.70	52.56	42.47	73.79	82.13	55.34	51.68	64.95	77.44	56.85	46.69	70.20	83.16	53.75	44.51
	RACL+R	64.00	**79.36**	59.90	46.39	**78.76**	**86.19**	70.61	60.11	**69.51**	**79.91**	65.46	53.95	**73.19**	**86.73**	72.29	60.39
	Ours	**67.27**	74.90	**67.45**	**59.27**	76.39	84.99	**76.23**	**70.69**	67.16	77.79	**68.60**	**61.05**	73.18	84.13	**76.52**	**68.13**

Table 2: Experimental results (%). Specifically, 'A-S', 'O', 'P', and 'T' denote aspect term and sentiment co-extraction, opinion term extraction, aspect-opinion pair extraction, and aspect sentiment triplet extraction, respectively.

pared with GTS, our model still achieves competitive performances, which verify the effectiveness of our model.

Ablation Study

To further validate the origination of the significant improvement of BMRC, we conduct ablation experiments and answer the following questions:

- Does the restrictive extraction query build the association between opinion entity extraction and relation detection?
- Does the bidirectional structure promote the performance of aspect-opinion pair extraction?
- Do the relations between aspects and opinion expressions enhance the sentiment classification?
- How much improvement can the BERT bring?

Effect of the Restrictive Extraction Query

We first validate whether the restrictive extraction query could effectively capture and exploit the dependency between opinion entity extraction and relation detection for better performance. Accordingly, we construct a two-stage model similar to TSF, called 'Ours w/o REQ'. In the first stage, we remove the restrictive extraction query $Q^{\mathcal{R}}$ from BMRC for only the opinion entity extraction and sentiment classification. The stage-2 model, which is responsible for relation detection, is also based on MRC with the input query '*Matched the aspect* a_i *and the opinion expression* o_j*?*'. Experimental results are shown in Figure 3. Although the performances on aspect extraction and opinion extraction are comparable, the performances of 'Ours w/o REQ' on triplet extraction and aspect-opinion pair extraction are evidently inferior than BMRC. The reason is that with the removal of the restrictive extraction query, the opinion entity extraction and relation detection are separated and no dependency would be captured by 'Ours w/o REQ'. This indicates the effectiveness of the restrictive query at capturing the dependency.

Effect of the Bidirectional MRC Structure

To explore the effect of bidirectional MRC structure, we compare our model with two unidirectional models, including 'Ours w/o AO' and 'Ours w/o OA'. Concretely, 'Ours w/o AO' extracts triplets only through O→A direction, and 'Ours w/o OA' extracts triplets through A→O direction. As shown in Figure 3, 'Ours w/o OA' shows inferior performance on opinion term extraction without O→A direction MRC, while 'Ours w/o AO' shows worse performance on aspect term extraction. This further harms the performances on aspect-opinion pair extraction and triplet extraction. The reason is that both aspects and opinion expressions can initiate aspect-opinion pairs, and the model will be biased when relations are forced to be detected by either aspects or opinions only. By introducing the bidirectional design, the two direction MRCs can complement each other and further improve the performance of aspect-opinion pair extraction and triplet extraction.

Models	14-Lap*	14-Res*	15-Res*	16-Res*
IMN + IOG	47.68	61.65	53.75	-
GTS	54.58	**70.20**	58.67	**67.58**
Ours	**57.83**	70.01	**58.74**	67.49

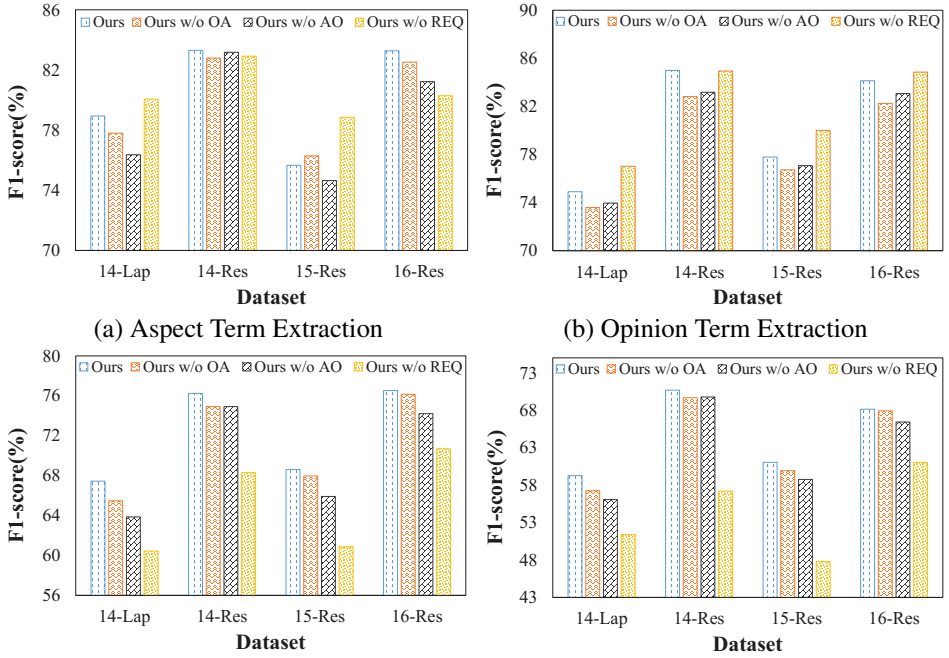
Table 3: Experimental results of aspect sentiment triplet extraction on the AFOE datasets. (F_1*-score, %*).

Datasets	A		A-S	
	Ours	Our w/o REQ	Ours	Ours w/o REQ
14-Lap	78.94	**80.06**	**67.27**	61.61
14-Res	**83.31**	82.73	**76.39**	66.26
15-Res	75.67	**79.00**	**67.16**	56.82
16-Res	**83.28**	80.60	**73.18**	68.82

Table 4: Experimental results of the ablation study on relation-aware sentiment classification (F_1*-score, %*). Specifically, 'A' and 'A-S' stand for aspect term extraction and aspect term and sentiment co-extraction, respectively.

Effect of Relation-Aware Sentiment Classification

In order to examine the benefit that the relations between aspects and opinion expressions provide for the sentiment classification, we compare the performances of our model and 'Ours w/o REQ'. Experimental results on aspect term



(a) Aspect Term Extraction (b) Opinion Term Extraction

(c) Aspect-Opinion Pair Extraction (d) Triplet Extraction

Figure 3: Experimental results of ablation study on the restrictive extraction query and the bidirectional structure.

Models	14-Lap	14-Res	15-Res	16-Res
TSF	43.50	51.89	46.79	53.62
Ours w/o BERT	48.15	63.32	53.77	63.16
Ours w/o REQ	51.40	57.20	47.79	61.03
Ours	**59.27**	**70.69**	**61.05**	**68.13**

Table 5: Experimental results of the ablation study on aspect sentiment triplet extraction (F_1-*score, %*), which aims to analyze the effect of BERT.

extraction and aspect term and sentiment co-extraction are shown in Table 4. Since 'Ours w/o REQ' separate relation detection and sentiment classification in two stages, the detected relations cannot directly provide assistance to sentiment classification. According to the results, although removing relation detection from joint learning does not harm the performance of aspect term extraction seriously, the performances of aspect term and sentiment co-extraction are all significantly weakened. This clearly indicates that the relations between aspects and opinion expressions can effectively boost the performance of sentiment classification.

Effect of BERT

We analyze the effect of BERT and our contributions from two perspectives. First, we construct our model based on BiDAF, which is a typical reading comprehension model without BERT, and refer it to 'Ours w/o BERT'. According to the results shown in Table 5, it significantly surpasses TSF by an average of 8.15% F1-score on triplet extraction, which shows that our model can achieve SOTA performance without BERT. Besides, compared with 'Ours w/o BERT', our model further improves 7.69% F1-score, which is brought by BERT. Second, we compare our model with 'Ours w/o REQ' and TSF. The ablation model 'Ours w/o REQ' can be regarded as an implementation version of TSF based on BERT and MRC framework. By comparing it with TSF, the results show that BERT based MRC model can bring an average of 5.41% F1-score improvement on triplet extraction against the counterpart model without BERT. By further introducing the bidirectional MRC structure and three types of queries, our model further outperforms 'Ours w/o REQ' by 10.4% F1-score. These two-fold analyses indicate that our contributions play a greater role in improving performances than BERT.

Conclusion

In this paper, we formalized the aspect sentiment triplet extraction (ASTE) task as a multi-turn machine reading comprehension (MTMRC) task and proposed the bidirectional MRC (BMRC) framework with well-designed queries. Specifically, the non-restrictive and restrictive extraction queries are designed to naturally fuse opinion entity extraction and relation detection, enhancing the dependency between them. By devising the bidirectional MRC structure, it can be ensured that either an aspect or an opinion expression can trigger an aspect-opinion pair just like human's reading behavior. In addition, the sentiment classification query and joint learning manner are used to further promote sentiment classification with the incorporation of relations between aspects and opinion expressions. The empirical study demonstrated that our model achieves state-of-the-art performance.

Acknowledgments

This research is supported by the National Natural Science Foundation of China under grant No. 61976119 and the Major Program of Science and Technology of Tianjin under grant No. 18ZXZNGX00310.

References

Chen, S.; Liu, J.; Wang, Y.; Zhang, W.; and Chi, Z. 2020. Synchronous Double-channel Recurrent Network for Aspect-Opinion Pair Extraction. In *ACL 2020*, 6515–6524.

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

Dai, H.; and Song, Y. 2019. Neural Aspect and Opinion Term Extraction with Mined Rules as Weak Supervision. In *ACL 2019*, 5268–5277.

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

Dong, L.; Wei, F.; Tan, C.; Tang, D.; Zhou, M.; and Xu, K. 2014. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In *ACL 2014*, 49–54.

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 2019*, 2509–2518.

Hazarika, D.; Poria, S.; Vij, P.; Krishnamurthy, G.; Cambria, E.; and Zimmermann, R. 2018. Modeling Inter-Aspect Dependencies for Aspect-Based Sentiment Analysis. In *NAACL 2018*, 266–270.

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

He, R.; Lee, W. S.; Ng, H. T.; and Dahlmeier, D. 2018. Exploiting Document Knowledge for Aspect-level Sentiment Classification. In *ACL 2018*, 579–585.

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

Hu, M.; Zhao, S.; Zhang, L.; Cai, K.; Su, Z.; Cheng, R.; and Shen, X. 2019. CAN: Constrained Attention Networks for Multi-Aspect Sentiment Analysis. In *EMNLP-IJCNLP 2019*, 4600–4609.

Levy, O.; Seo, M.; Choi, E.; and Zettlemoyer, L. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *CoNLL 2017*, 333–342.

Li, H.; and Lu, W. 2017. Learning Latent Sentiment Scopes for Entity-Level Sentiment Analysis. In *AAAI 2017*, 3482–3489.

Li, X.; Bing, L.; Li, P.; and Lam, W. 2019a. A Unified Model for Opinion Target Extraction and Target Sentiment Prediction. In *AAAI 2019*, 6714–6721.

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

Li, X.; Feng, J.; Meng, Y.; Han, Q.; Wu, F.; and Li, J. 2020. A Unified MRC Framework for Named Entity Recognition. In *ACL 2020*, 5849–5859.

Li, X.; and Lam, W. 2017. Deep Multi-Task Learning for Aspect Term Extraction with Memory Interaction. In *EMNLP 2017*, 2886–2892.

Li, X.; Yin, F.; Sun, Z.; Li, X.; Yuan, A.; Chai, D.; Zhou, M.; and Li, J. 2019b. Entity-Relation Extraction as Multi-Turn Question Answering. In *ACL 2019*, 1340–1350.

Li, Z.; Wei, Y.; Zhang, Y.; Zhang, X.; and Li, X. 2019c. Exploiting Coarse-to-Fine Task Transfer for Aspect-Level Sentiment Classification. In *AAAI 2019*, 4253–4260.

Liu, K.; Xu, L.; and Zhao, J. 2012. Opinion Target Extraction Using Word-Based Translation Model. In Tsujii, J.; Henderson, J.; and Pasca, M., eds., *EMNLP-CoNLL 2012*, 1346–1356.

Liu, K.; Xu, L.; and Zhao, J. 2015. Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model. *IEEE Trans. Knowl. Data Eng.* 27(3): 636–650.

Liu, P.; Joty, S.; and Meng, H. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *EMNLP 2015*, 1433–1443.

Loshchilov, I.; and Hutter, F. 2017. Fixing Weight Decay Regularization in Adam. *CoRR* abs/1711.05101.

Luo, H.; Li, T.; Liu, B.; and Zhang, J. 2019. DOER: Dual Cross-Shared RNN for Aspect Term-Polarity Co-Extraction. In *ACL 2019*, 591–601.

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

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

McCann, B.; Keskar, N. S.; Xiong, C.; and Socher, R. 2018. The Natural Language Decathlon: Multitask Learning as Question Answering. *CoRR* abs/1806.08730.

Mitchell, M.; Aguilar, J.; Wilson, T.; and Durme, B. V. 2013. Open Domain Targeted Sentiment. In *EMNLP 2013*, 1643–1654.

Nguyen, T. H.; and Shirai, K. 2015. PhraseRNN: Phrase Recursive Neural Network for Aspect-based Sentiment Analysis. In *EMNLP 2015*, 2509–2514.

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 2020*, 8600–8607.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *NAACL-HLT 2018*, 2227–2237.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Al-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; Clercq, O. D.; Hoste, V.; Apidianaki, M.; Tannier, X.; Loukachevitch, N. V.; Kotelnikov, E. V.; Bel, N.; Zafra, S. M. J.; and Eryigit, G. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016*, 19–30.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Manandhar, S.; and Androutsopoulos, I. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015*, 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 *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014*, 27–35.

Poria, S.; Cambria, E.; and Gelbukh, A. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems* 108: 42–49.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog 2019* 1(8): 9.

Seo, M. J.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional Attention Flow for Machine Comprehension. In *ICLR 2017*.

Sun, K.; Zhang, R.; Mensah, S.; Mao, Y.; and Liu, X. 2019. Aspect-Level Sentiment Analysis Via Convolution over Dependency Tree. In *EMNLP-IJCNLP 2019*, 5678–5687.

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

Wang, S.; Mazumder, S.; Liu, B.; Zhou, M.; and Chang, Y. 2018. Target-Sensitive Memory Networks for Aspect Sentiment Classification. In *ACL 2018*, 957–967.

Wang, W.; Pan, S. J.; Dahlmeier, D.; and Xiao, X. 2016. Recursive Neural Conditional Random Fields for Aspect-based Sentiment Analysis. In *EMNLP 2016*, 616–626.

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 2017*, 3316–3322.

Wu, Z.; Ying, C.; Zhao, F.; Fan, Z.; Dai, X.; and Xia, R. 2020a. Grid Tagging Scheme for Aspect-oriented Fine-grained Opinion Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2576–2585.

Wu, Z.; Zhao, F.; Dai, X.; Huang, S.; and Chen, J. 2020b. Latent Opinions Transfer Network for Target-Oriented Opinion Words Extraction. In *AAAI 2020*, 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 2018*, 592–598.

Yu, A. W.; Dohan, D.; Luong, M.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. V. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. In *ICLR 2018*.

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 2020*, 3239–3248.