# Target-specified Sequence Labeling with Multi-head Self-attention for Target-oriented Opinion Words Extraction

**Yuhao Feng**[1], **Yanghui Rao**[1,*], **Yuyao Tang**[2], **Ninghua Wang**[2], **He Liu**[2]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Tencent Inc, Shenzhen, China
fengyh3@mail2.sysu.edu.cn; raoyangh@mail.sysu.edu.cn
{youngtang,ninghuawang,heziliu}@tencent.com

## Abstract

Opinion target extraction and opinion term extraction are two fundamental tasks in Aspect Based Sentiment Analysis (ABSA). Many recent works on ABSA focus on Target-oriented Opinion Words (or Terms) Extraction (TOWE), which aims at extracting the corresponding opinion words for a given opinion target. TOWE can be further applied to Aspect-Opinion Pair Extraction (AOPE) which aims at extracting aspects (i.e., opinion targets) and opinion terms in pairs. In this paper, we propose **T**arget-**S**pecified sequence labeling with **M**ulti-head **S**elf-**A**ttention (TSMSA) for TOWE, in which any pre-trained language model with multi-head self-attention can be integrated conveniently. As a case study, we also develop a **M**ulti-**T**ask structure named MT-TSMSA for AOPE by combining our TSMSA with an aspect and opinion term extraction module. Experimental results indicate that TSMSA outperforms the benchmark methods on TOWE significantly; meanwhile, the performance of MT-TSMSA is similar or even better than state-of-the-art AOPE baseline models.

## 1 Introduction

Aspect-Based Sentiment Analysis (ABSA) (Pontiki et al., 2014) has attracted much attention of researchers in recent years. In ABSA, aspect (or called opinion target) extraction and opinion term extraction are two fundamental tasks. Aspect is the word or phrase in the reviews referring to the object towards which users show attitudes, while opinion terms are those words or phrases representing users' attitudes (Wu et al., 2020). For example, in the sentence "*The dim sum is delicious.*", the phrase "*dim sum*" is an aspect and the word "*delicious*" is an opinion term. See the upper part of Table 1 for more examples. Plenty of works based on neural networks have been done in both aspect and opinion term extraction (Liu et al., 2015; Poria et al., 2016; Xu et al., 2018); moreover, some studies combine these two tasks into a multi-task structure to extract aspects and opinion terms simultaneously (Wang et al., 2016, 2017; Li and Lam, 2017; Dai and Song, 2019).

**Reviews:**
"Soooo great! The food is delicious and inexpensive, and the environment is in a nice. The only problem is that the soup and dessert are ordinary."

**Aspect-Opinion Pairs:**
food : [delicious, inexpensive] (*one-to-many*)
environment : [nice] (*one-to-one*)
soup, dessert : [ordinary] (*many-to-one*)

Table 1: The upper part is a restaurant review and the lower part shows the corresponding aspect-opinion pairs. Extracted aspects and opinion terms are marked in red and blue, respectively.

However, one critical deficiency in the researches mentioned above is that they ignore the relation of aspects and opinion terms, which leads to the birth of Target-oriented Opinion Words (or Terms) Extraction (TOWE) (Fan et al., 2019) for extracting the corresponding opinion terms of a given opinion target. Subsequently, Aspect-Opinion Pair Extraction (AOPE) (Chen et al., 2020) and Pair-wise Aspect and Opinion Terms Extraction (PAOTE) (Zhao et al., 2020) have emerged, which both aim at extracting aspects and opinion terms in pairs. AOPE and PAOTE are exactly the same task, only named differently. In the following, we use AOPE to denote this task for simplicity. It can be considered that AOPE contains aspect and opinion word extraction and TOWE. Since aspect extraction has been fully studied and satisfactory results have been obtained, TOWE, which aims at mining the relation between aspects and opinion terms, is the key to the AOPE task. As shown in the lower part of Table 1, the relational structure of

[*] The corresponding author.

*Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1805–1815
June 6–11, 2021. ©2021 Association for Computational Linguistics

the aspect-opinion pairs within a sentence can be complicated, including one-to-one, one-to-many, and many-to-one.

The challenge of TOWE is the learning of representations of the given opinion target accurately and a few works focus on this task. For instance, Fan et al. (2019) propose an Inward-Outward LSTM to pass target information to the left context and the right context of the target respectively, and then they combine the left, right, and global context to encode the sentence. Recently, SDRN (Chen et al., 2020) and SpanMlt (Zhao et al., 2020) both adopt a pre-trained language model to learn contextual representations for AOPE. In SDRN, a double-channel recurrent network and a synchronization unit are applied to extract aspects, opinion terms and their relevancy. In SpanMlt, the terms are extracted under annotated span boundaries with contextual representations, and then the relations between every two span combinations are identified. However, apart from hyper-parameters in the pre-trained language model, these two methods introduce many other hyper-parameters (e.g., the hidden size, thresholds and recurrent steps in SDRN, and the span length, top $k$ spans and the balanced factor of different tasks in SpanMlt). Some of these hyper-parameters have a significant impact on the model performance.

Motivated by the previous work and to address the challenges mentioned above, we propose a **T**arget-**S**pecified sequence labeling method based on **M**ulti-head **S**elf-**A**ttention (Vaswani et al., 2017) (TSMSA). The sentence is first processed in the format "[SEP] Aspect [SEP]" (e.g., "*The [SEP] food [SEP] is delicious.*"), which is inspired by Soares et al. (2019) who utilized a special symbol "[SEP]" to label all entities and output their corresponding representations. Then we develop a sequence labeling model based on multi-head self-attention to identify the corresponding opinion terms. By using the special symbol and self-attention mechanism, TSMSA is capable of capturing the information of the specific aspect. To improve the performance of our model, we apply pre-trained language models like BERT (Devlin et al., 2019) which contain a multi-head self-attention module as the encoder. As a case study, we integrate aspect and opinion term extraction, and TOWE into a **M**ulti-**T**ask architecture named MT-TSMSA to validate the effectiveness of our method on the AOPE task. In addition, apart from hyper-parameters in the pre-trained language model, we only need to adjust the balanced factor of different tasks in MT-TSMSA. In summary, our main contributions are as follows:

- We propose a target-specified sequence labeling method with multi-head self-attention mechanism to perform TOWE, which generates target-specific context representations for different targets in the same review with the special symbol and multi-head self-attention. Pre-trained language models can be conveniently applied to improve the performance.
- For our TSMSA and MT-TSMSA, only a small amount of hyper-parameters need to be adjusted when using pre-trained language models. Compared to the existing models for TOWE and AOPE, we alleviate the tradeoff issue between a model's complexity and performance.

Extensive experiments validate that our TSMSA can achieve the best performance on TOWE, and MT-TSMSA performs quite competitive on AOPE. The rest of this paper is organized as follows. Section 2 introduces the existing studies on TOWE and AOPE, respectively. Section 3 details the proposed TSMSA and MT-TSMSA. Section 4 presents our experimental results and discussions. Finally, we draw conclusions in Section 5.

## 2 Related Works

### 2.1 Target-oriented Opinion Words Extraction

Plenty of works have been carried out for aspect extraction and opinion term extraction. Early researches can be divided into unsupervised/semi-supervised methods (Hu and Liu, 2004; Zhuang et al., 2006; Qiu et al., 2011) and supervised methods (Jakob and Gurevych, 2010; Shu et al., 2017). With the development of neural networks, deep learning methods (Liu et al., 2015; Yin et al., 2016; Poria et al., 2016; Xu et al., 2018) have made impressive progress in recent years. Several works integrate aspect extraction and opinion term extraction into a co-extraction process. Qiu et al. (2011) expand the list of aspects and opinion terms in a bootstrapping method by double propagation. Some other works adopt the co-extraction structure in neural networks with multi-task learning (Wang et al., 2016, 2017; Li and Lam, 2017).

However, the above methods ignore the relation between aspects and opinion terms and only a few works focus on this field. Rule-based methods (Hu and Liu, 2004; Zhuang et al., 2006) are proposed to select corresponding opinion terms with distance rule and syntactic rule templates based on dependency parsing trees. However, the performance of these methods heavily relies on expert knowledge and these rules usually cover only a small amount of cases. Fan et al. (2019) carry out TOWE by extracting the corresponding opinion terms for a given aspect, and then utilize Inward-Outward LSTM to generate implicit representations of aspects. Nevertheless, this approach is not capable of applying powerful pre-trained language models like BERT as the encoder to perform better. Our model aims to extract corresponding opinion terms of the given aspect with explicit representations, in addition to boost performance by employing BERT as the encoder.

### 2.2 Aspect-Opinion Pair Extraction

Aspect-Opinion Pair Extraction (AOPE) (Chen et al., 2020) and Pair-wise Aspect and Opinion Terms Extraction (PAOTE) (Zhao et al., 2020) both aim at extracting aspects and opinion terms in pairs. AOPE and PAOTE are essentially the same task with different names, and they can be split into aspect extraction and TOWE. Chen et al. (2020) propose a Synchronous Double-channel Recurrent Network (SDRN) which consists of an opinion entity extraction unit, a relation detection unit, and a synchronization unit for pair extraction. Zhao et al. (2020) develop a span-based multi-task learning framework (SpanMlt) where the terms are extracted under annotated span boundaries, so as to identify the relations between every two span combinations.

However, SDRN contains a lot of hyper-parameters and SpanMlt generates a great many of candidate spans if the value of maximal length of a span is large or the sentence is too long. The advantage of our methods is that only a small amount of hyper-parameters adjustment is required and similar or even better performance can be achieved.

## 3 Methodology

### 3.1 Task Description

Given a sentence $s = \{w_1, w_2, ..., w_n\}$ consisting of $n$ words, an aspect (opinion target) $a = \{w_i, w_{i+1}, ..., w_{i+k}\}$, and an opinion term $o = \{w_j, w_{j+1}, ..., w_{j+m}\}$ ($a$ and $o$ are substrings of $s$), the probabilities of target-oriented opinion terms are defined as $p(o|s, a)$ in the TOWE task and the probabilities of aspect-opinion pairs are defined as $p(\langle a, o\rangle|s) = p(a|s) \times p(o|s, a)$ in the AOPE task. The BIO tagging scheme (Ramshaw and Marcus, 1995) and a special symbol "[SEP]" are applied to this task, where each word $w_i$ in the sentence $s$ is tagged as $y_i \in$ {B, I, O, [SEP]} (B: Beginning, I: Inside, O: Others, [SEP]: the tag of an aspect).

### 3.2 Framework

The structures of our **T**arget-**S**pecified sequence labeling method based on **M**ulti-head **S**elf-**A**ttention (TSMSA) and the **M**ulti-**T**ask version (MT-TSMSA) are shown in Figure 1 (c) and (d). As aforementioned, we first use a special symbol "[SEP]" to label each aspect. Next, the multi-head self-attention method is applied to capture the context representations of the specific aspect explicitly, then they are passed to a projection layer and a Conditional Random Field (CRF) (Lafferty et al., 2001) layer for sequence labeling. Furthermore, the aspect and opinion words extraction (task 0) as well as the target-oriented opinion words extraction (task 1) are combined for multi-task learning. These two tasks share the parameters of encoder but differ in projection and CRF layers.

### 3.3 Multi-Head Self-Attention

We describe the multi-head self-attention approach according to Vaswani et al. (2017) with the details shown in Figure 1 (a) and (b). For each attention head in the above approach, we first compute the scaled dot-product attention. Particularly, the input consists of a set of queries, keys, and values, where $d_k$ stands for the dimension of queries and keys, and $d_v$ represents the dimension of values. Then they are packed together into matrices $Q$, $K$, and $V$, respectively. The scaled dot-product attention is calculated as follows:

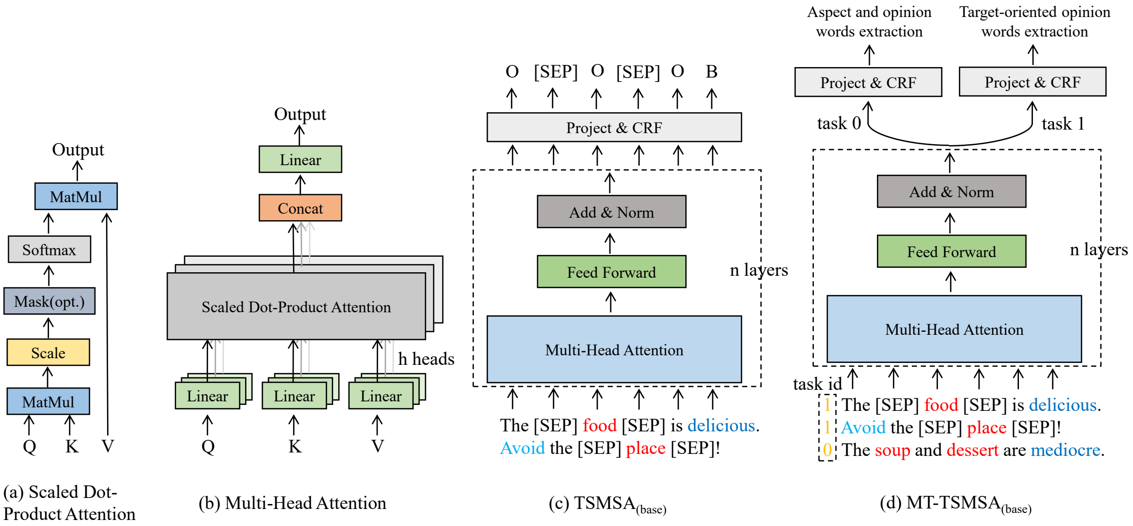
$$Attention(Q, K, V) = softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V. \tag{1}$$

Next, given the number of attention heads $h$, we can get the dimension of output $d_{model} = h \times d_v$. Finally, the multi-head attention is described as follows:

$$MH(I, h) = Concat(head_1, ..., head_h)W^O, \tag{2}$$

$$head_i = Attention(IW_i^Q, IW_i^K, IW_i^V), \tag{3}$$

where $I = \{\vec{i_1}, \vec{i_2}, ..., \vec{i_n}\}$ (the dimension of $\vec{i}$ is $d_{model}$) indicates the input and $n$ is the sequence



(a) Scaled Dot-Product Attention

(b) Multi-Head Attention

(c) TSMSA$_{(base)}$

(d) MT-TSMSA$_{(base)}$

Figure 1: The structures of our TSMSA and MT-TSMSA base models are presented in (c) and (d). For clarity, the details about Multi-head Attention are shown in (a) and (b).

length. The parameter matrices of projections are $W_i^Q \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^K \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^V \in \mathbb{R}^{d_{model}\times d_v}$, and $W_i^O \in \mathbb{R}^{d_{model}\times d_{model}}$.

### 3.4 Target-Specified Encoder

To start with, the input vector of each word is generated by utilizing a word embedding lookup table $\mathbb{L}_w \in \mathbb{R}^{r\times d_w}$ and a positional embedding lookup table $\mathbb{L}_p \in \mathbb{R}^{n\times d_p}$, where $d_w$ is the dimension of word embeddings, $r$ is the vocabulary size, and $d_p$ is the dimension of positional embeddings. These embedding lookup tables will map $s = \{w_1, ..., w_n\}$ to $\{\vec{e_1}^w, ..., \vec{e_n}^w\}$ and $\{\vec{e_1}^p, ..., \vec{e_n}^p\}$, respectively. For our base models (not using a pre-trained language model), $\vec{e_i}^w$ will be projected to a low dimensional vector $\vec{e_i}^{low}$ which is calculated as follows: $\vec{e_i}^{low} = \sigma(W^e\vec{e_i}^w)$, where $W^e \in \mathbb{R}^{d_{low}\times d_w}$ ($d_{low} < d_w$) denotes the matrix of projection and $\sigma(\cdot)$ is the activation function. In this case, $\vec{t_i}$ in the input $T = \{\vec{t_1}, ..., \vec{t_n}\}$ is represented by $[\vec{e_i}^{low}; \vec{e_i}^p]$ and $d_{model} = d_{low} + d_p$. For a pre-trained language model like BERT (Devlin et al., 2019), $\vec{t_i}$ equals the sum of $\vec{e_i}^w$, $\vec{e_i}^p$, and $\vec{e_i}^s$, where $e^s = \{\vec{e_1}^s, ..., \vec{e_n}^s\}$ (the dimension of $\vec{e_i}^s$ is $d_p$) represents segment embeddings, and $d_{model} = d_p = d_w$.

Then, the input vector $T$ is passed to multi-head self-attention modules, where a feed-forward network and an add-norm network are combined in sequence to generate the context representation of each layer $H = \{H^1, ..., H^l\}$, where $l$ is the number of multi-head attention layers and $H^i = \{\vec{H_1}^i, ..., \vec{H_n}^i\}$. $H^i$ can be calculated as follows:

$$O^i = MH(H^{i-1}, h), \tag{4}$$

$$FFN^i = max(0, O^iW_1^i + b_1^i)W_2^i + b_2^i, \tag{5}$$

$$H^i = LN(H^{i-1} + FFN^i), \tag{6}$$

where $h$ is the number of attention heads, $H^0 = T$, the matrices $W_1^i \in \mathbb{R}^{d_{model}\times d_{ff}}$ and $W_2^i \in \mathbb{R}^{d_{ff}\times d_{model}}$ represent mappings from $d_{model}$ to $d_{ff}$ and back to $d_{model}$. $LN(\cdot)$ is a layer normalization method applying to sequential data (Ba et al., 2016). Finally, the output of the encoder is $H^l$, i.e., the last layer of $H$.

### 3.5 Decoder and Training

Given a sequential representation $H^l$ and a sequential label $Y = \{y_1, ..., y_n\}$ ($y_i \in$ {B, I, O, [SEP]} or $y_i \in$ {B-ASP, I-ASP, B-OP, I-OP, O}[1]), we can use $H^l$ to compute $p(Y|H^l)$. Greedy decoding or CRF can be adopted in the decoding process. CRF is chosen as our decoding strategy because CRF has the ability to capture the correlations between tokens and labels and the correlations between adjacent labels simultaneously. Given a new sentence, we use Viterbi algorithm (Viterbi, 1967) to predict the label sequence by maximizing the conditional probability $p(Y|H^l)$ in the decoding process.

#### 3.5.1 Single-Task Version

The single-task version of our approaches is TSMSA. Given a predicted label sequence $Y$ and

[1] B-ASP: beginning of an aspect, I-ASP: inside of an aspect, B-OP: beginning of an opinion term, I-OP: inside of an opinion term, and O: others.

a sequential representation $H^l$, the score function $S(H^l, Y)$ can be defined as follows:

$$S(H^l, Y) = \sum_{i=1}^{n} Q_{y_{i-1},y_i} + \sum_{i=1}^{n} P_{i,y_i}, \quad (7)$$

$$P = H^l W_p + b_p, \quad (8)$$

where the matrix $Q \in \mathbb{R}^{k\times k}$ captures the relation of adjacent labels, the matrix $P \in \mathbb{R}^{n\times k}$ learns the relation of tokens and labels, and the matrices $W_p \in \mathbb{R}^{d_{model}\times k}$ and $b_p \in \mathbb{R}^{n\times k}$ indicate a projection operation from dimension $d_{model}$ to dimension $k$. In the above, $k$ means the dimension of the label space. Then, the linear-chain CRF is exploited to calculate the conditional probability of the predicted sequence $Y$ as follows:

$$p(Y|H^l) = \frac{\exp(S(H^l, Y))}{\sum_{\tilde{Y}\in Y_{all}} \exp(S(H^l, \tilde{Y}))}, \quad (9)$$

where $Y_{all}$ denotes the set of all possible sequential labels. So the loss of a sentence can be calculated by the negative log likelihood as follows:

$$L(s) = -\log p(Y|H^l). \quad (10)$$

#### 3.5.2 Multi-Task Version

By integrating aspect and opinion term extraction (task 0) and TOWE (task 1) into a multi-task architecture, we propose a MT-TSMSA method for AOPE. MT-TSMSA can be defined as using a sentence $H^l$ and a task id $\in \{0, 1\}$ to calculate the conditional probability $p(Y|H^l, id)$. When the task id equals 0, it means aspect and opinion term extraction. For TOWE, the task id is 1. Some examples are shown in Figure 1 (d). Aiming at handling different tasks, different score functions $S_0(H^l, Y^0)$ and $S_1(H^l, Y^1)$ are defined, where $S_0(\cdot)$ and $S_1(\cdot)$ have different parameter matrices, $Y^0$ ($Y_i^0 \in$ {B-ASP, I-ASP, B-OP, I-OP O}) and $Y^1$ ($Y_i^1 \in$ {B, I, O, [SEP]}) represent the sequential labels of aspect and opinion term extraction, and TOWE, respectively. So the conditional probabilities of the predicted sequences $Y^0$ and $Y^1$ can be calculated as follows:

$$p(Y^0|H^l, id = 0) = \frac{\exp(S_0(H^l, Y^0))}{\sum_{\tilde{Y}\in Y_{all}^0} \exp(S_0(H^l, \tilde{Y}))}, \quad (11)$$

$$p(Y^1|H^l, id = 1) = \frac{\exp(S_1(H^l, Y^1))}{\sum_{\tilde{Y}\in Y_{all}^1} \exp(S_1(H^l, \tilde{Y}))}, \quad (12)$$

where $Y_{all}^0$ denotes the set of all possible sequential labels of task 0 and $Y_{all}^1$ represents the set of all possible sequential labels of task 1. The loss of a sentence is also calculated by the negative log likelihood as follows:

$$L(s, id) = -\log p(Y|H^l, id). \quad (13)$$

Given $M$ sentences $S = \{s_1, s_2, ..., s_M\}$ with $id = \{id_1, ..., id_M\}$, we can minimize the loss for training:

$$J(\theta) = \sum_{k=1}^{M} ((1 - id_k)\lambda + id_k)L(s_k, id_k), \quad (14)$$

where $\lambda$ is the hyper-parameter used to balance these two tasks.

### 3.6 Inference Process

For TOWE, a sentence with a given aspect (i.e., target) is first processed into target-specified mode ("[SEP] Aspect [SEP]") with the special symbol "[SEP]" and then passed into TSMSA, the outputs of which are the target-oriented opinion terms. For AOPE, MT-TSMSA generates aspect-opinion pairs by a two-stage inference process. Firstly, a sentence is passed into MT-TSMSA, where aspects are extracted in task 0. Secondly, given extracted aspects, repeating the inference process of TOWE, MT-TSMSA outputs the target-oriented opinion terms from task 1. Accordingly, the combinations of aspects from task 0 and target-orient opinion terms from task 1 are aspect-opinion pairs.

## 4 Experiments

### 4.1 Datasets

To evaluate the performance of our model[2], we conduct experiments on two public datasets from laptop and restaurant domains. These two datasets were respectively built by Fan et al. (2019) for TOWE and Chen et al. (2020) for AOPE based on SemEval Challenge 2014 Task 4, SemEval Challenge 2015 Task 12, and SemEval Challenge 2016 Task 5 (Pontiki et al., 2014, 2015, 2016). For the first dataset, every sentence was annotated by two people, and the conflicts were checked and eliminated manually. The second dataset was developed by extending the first one. The statistics of these benchmark datasets are shown in Table 2, from

[2] The code of our model is available in public at: https://github.com/fengyh3/TSMSA.

which we can observe that the second dataset includes many negative samples for AOPE (i.e., the sentences only contain aspects and opinion terms, without any aspect-opinion pairs). Note that these negative samples will also be considered when testing our model on AOPE.

| Provider | Datasets | | #Sent | #Target | #A | #None |
|---|---|---|---|---|---|---|
| (Fan et al., 2019) | **14lap** | train | 1158 | 1634 | 1626 | 0 |
| | | test | 343 | 482 | 481 | 0 |
| | **14res** | train | 1627 | 2643 | 2638 | 0 |
| | | test | 500 | 865 | 864 | 0 |
| | **15res** | train | 754 | 1076 | 1076 | 0 |
| | | test | 325 | 436 | 436 | 0 |
| | **16res** | train | 1079 | 1512 | 1512 | 0 |
| | | test | 329 | 457 | 456 | 0 |
| (Chen et al., 2020) | **14lap** | train | 3045 | 1535 | 2359 | 1297 |
| | | test | 800 | 380 | 653 | 334 |
| | **14res** | train | 3041 | 2809 | 3693 | 902 |
| | | test | 800 | 936 | 1134 | 219 |
| | **15res** | train | 1315 | 1231 | 1205 | 556 |
| | | test | 685 | 516 | 542 | 352 |

Table 2: Statistics of datasets. #Sent, #Target, #A, and #None represent the numbers of sentences, relations, and aspects, and the sentences without any aspects and opinion terms, respectively.

## 4.2 Baselines

Fan et al. (2019) have employed various baselines in TOWE, including **Distance-rule** (Hu and Liu, 2004), **Dependency-rule** (Zhuang et al., 2006), **BiLSTM + Distance-rule**, and **TC-BiLSTM**, except for BERT-based methods. To achieve comprehensive comparative analysis, we develop baselines of **BERT + Distance-rule** and **Target-fused BERT** (TF-BERT) for this task. The former trains a sentence-level opinion term extraction model by BERT, and the target-oriented opinion term is the one nearest to each aspect. The latter utilizes the average pooling of target word embeddings to represent the target information. The word representation at each position is the addition of word embedding and target information, which is fed into BERT to extract target-oriented opinion terms. Zhao et al. (2020) have applied some baselines in AOPE, including **HAST** (Li et al., 2018) **+ IOG** and **JERE-MHS** (Bekoulis et al., 2018). Besides the above methods, we also employ the following baselines:

- **IOG** (Fan et al., 2019) utilizes an Inward-Outward LSTM and a Global LSTM to capture the information of aspects and global information respectively, then it combines these information for sequence labeling.
- **SpanMlt** (Zhao et al., 2020) is a span-based multi-task learning framework where the terms are extracted with annotated span boundaries and then the relations between combinations of every two spans are identified.
- **SDRN** (Chen et al., 2020) utilizes BERT as the encoder which consists of an opinion entity extraction unit, a relation detection unit, and a synchronization unit for the AOPE task. In the case of TOWE, this model extracts the target-oriented opinion terms with given correct aspects.

## 4.3 Hyper-parameter Settings

For the TOWE task, Fan et al. (2019) utilize 300-dimension GloVe (Pennington et al., 2014) vectors which are pre-trained on unlabeled data of 840 billion tokens to initialize word embedding vectors in IOG. The word embeddings are fixed at the stage of training. For fair comparison, we use the same fixed word embeddings in TSMSA(Base). We randomly select 20% of the training set as the development set for adjusting all hyper-parameters. The value of $d_{model}$ is 128, and the numbers of attention heads and layers are 4 and 6, respectively. In addition, the dropout rate, learning rate, and maximal sequence length are set to 0.5, 0.001, and 100, respectively. Adam optimizer (Kingma and Ba, 2015) is adopted to optimize our model. Pre-trained language models like BERT (Devlin et al., 2019) can be applied to our methods, and we adopt BERT-base[3] model, where $d_{model}$ is 768 and the number of attention heads and layers are both 12. Other hyper-parameters include the learning rate of BERT and CRF, the maximal sequence length, and the number of epochs. Based on the development set, these hyper-parameters are set to 5e-5, 2e-4, 100, and 8, respectively. Unless otherwise mentioned, $\lambda$ is set to 1.

To be consistent with various baselines (Fan et al., 2019; Chen et al., 2020; Zhao et al., 2020), the term-level F1 score is used as the evaluation metric for both TOWE and AOPE tasks. Term-level means that the boundaries of the span are the same as the ground-truth. For the AOPE task, the consistency of a predicted aspect-opinion pair with the labeled pair indicates the correctness of prediction.

## 4.4 Results and Analysis

### 4.4.1 Target-oriented Opinion Words Extraction

Table 3 presents the performance of different models on TOWE. Firstly, the F1 scores of rule-

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

| Models | (Fan et al., 2019) | | | | (Chen et al., 2020) | | |
|---|---|---|---|---|---|---|---|
| | 14lap | 14res | 15res | 16res | 14lap | 14res | 15res |
| Distance-rule | 40.42 | 49.92 | 45.97 | 51.83 | 47.68* | 56.24* | 51.15* |
| Dependency-rule | 37.14 | 58.04 | 55.98 | 64.62 | 43.23* | 62.17* | 59.76* |
| BiLSTM + Distance-rule | 63.38 | 69.18 | 66.97 | 74.01 | 66.77* | 72.54* | 69.42* |
| TC-BiLSTM | 61.21 | 67.61 | 62.94 | 73.10 | 65.83* | 71.23* | 65.55* |
| IOG | **71.35** | 80.02 | 73.25 | **81.69** | 76.43* | **83.24*** | 76.63* |
| TSMSA(Base) | 71.10 | **80.31** | **75.38** | 80.68 | **77.66** | 82.35 | **77.52** |
| BERT + Distance-rule | 70.54* | 76.23* | 71.26* | 79.53* | 73.84* | 78.92* | 76.57* |
| TF-BERT | 72.26* | 78.23* | 71.58* | 79.23* | 74.32* | 79.28* | 76.94* |
| SDRN | 80.24* | 83.53* | 80.18* | 86.72* | 87.54* | 86.72* | 85.17* |
| TSMSA(BERT) | **82.18** | **86.37** | **81.64** | **89.20** | **88.63** | **90.03** | **87.30** |

Table 3: Experimental results (F1 score, %) of different models on TOWE. The methods in the upper part do not utilize the pre-trained language model (i.e., BERT) and BERT is applied in the lower part. The results with '*' are reproduced by us, and others are released from Fan et al. (2019). Best results are marked in bold.

based methods are poor because the rules only cover a small number of cases. By utilizing BiLSTM or BERT as the encoder to extract opinion terms, the BiLSTM/BERT + Distance-rule perform much better than other rule-based methods. However, these methods cannot deal with the one-to-many case. Secondly, TC-BiLSTM and TF-BERT extract static word embeddings for aspects and then incorporate them into sentence representation by concatenation or addition. Nevertheless, the results of TC-BiLSTM and TF-BERT are still over 10% lower than IOG/TSMSA(Base) and SDRN/TSMSA(BERT), respectively. It reveals that the static word embedding is not a good representation of the aspect and the concatenation/addition operation is not good enough to represent the specific aspect. Finally, IOG is a state-of-the-art baseline method for TOWE and the performance of TSMSA(Base) trained by the same word embedding is similar to IOG, which indicates the effectiveness in capturing the representation of a specific aspect with the symbol "[SEP]".

Furthermore, the pre-trained language model BERT can be applied to our basic method. The F1 score of TSMSA(BERT) is in average 8% higher than TSMSA(Base) and IOG. SDRN, which also exploits BERT as the encoder, passes the information of the aspect through a synchronization unit and utilizes supervised self-attention to capture this information. Nevertheless, it represents the specific aspect implicitly, which might have an negative impact on capturing the information of targets. In average, the performance of SDRN is 2% lower than TSMSA(BERT). The overall results reveal that our proposed method achieves state-of-the-art performance on TOWE.

### 4.4.2 Aspect-Opinion Pair Extraction

As mentioned above, our method can be applied to AOPE by combining TOWE with aspect and opinion term extraction. We here compare the performance of our multi-task model (i.e., MT-TSMSA) with the following competitive models: HAST + IOG, JERE-MHS, SpanMlt, and SDRN. The results are shown in Table 4. Note that the overlapping ratios of pairs in 14lap, 14res, and 15res are 78.8%, 92%, and 99.8% for (Fan et al., 2019), and 87.1%, 86.2%, and 86.4% for (Chen et al., 2020), respectively. Thus, there is a difference (within 2% mostly) between the results on these two datasets.

| Models | (Fan et al., 2019) | | | | (Chen et al., 2020) | | |
|---|---|---|---|---|---|---|---|
| | 14lap | 14res | 15res | 16res | 14lap | 14res | 15res |
| HAST + IOG | 53.41 | 62.39 | 58.12 | 63.84 | 58.97* | 63.14* | 58.84* |
| JERE-MHS | 52.34 | 66.02 | 59.64 | 67.65 | 58.69* | 67.81* | 60.17* |
| SpanMlt | 68.66 | 75.60 | 64.68 | 71.78 | - | - | - |
| SDRN | 68.50* | 74.91* | **70.08*** | 76.92* | 67.13 | 76.48 | 70.94 |
| MT-TSMSA(BERT) | **69.33** | **78.37** | 69.13 | **78.39** | **68.18** | **76.69** | **71.64** |

Table 4: F1 scores (%) of aspect-opinion pairs extraction. The results with '*' are reproduced by us, and others are released from Zhao et al. (2020) and Chen et al. (2020). Best results are marked in bold.

The performance of JERE-MHS is better than HAST + IOG, which indicates that the degree of error propagation in the separate training model might be smaller than it in the model of joint training. Moreover, SpanMlt, SDRN, and MT-TSMSA(BERT) use powerful pre-trained language models, which have a significant improvement in the performance on AOPE. We observe that SDRN and MT-TSMSA(BERT) perform better than SpanMlt, showing that selecting top $k$ spans from candidate spans as pairs might miss some correct pairs. Compared to SDRN, MT-TSMSA(BERT) performs better on three datasets and nearly the same on four datasets. Overall, MT-TSMSA achieves quite competitive performance on AOPE by simply incorporating our TSMSA into a multi-task structure.

### 4.5 Ablation Experiments

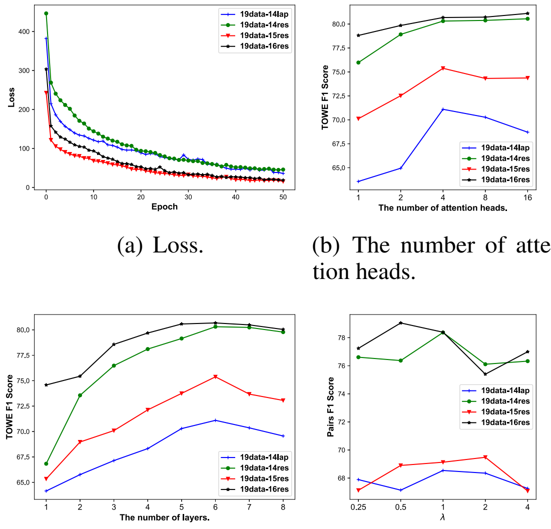
To evaluate the impacts of different word embeddings and training strategies on our models, we conduct ablation experiments by varying the above factors. The results shown in Table 5 indicate that a suitable word embedding is capable of improving the performance of our models. Firstly, BERT embedding shows poor performance when compared to Glove. We conjecture that BERT embedding needs to cooperate with the pre-trained encoder of BERT to perform better on TOWE. Secondly, applying the word embedding and the encoder of BERT without fine-tuning also fails to work on TOWE. The reason may be that the encoder of BERT without fine-tuning cannot capture the information of the specific aspect with the symbol "[SEP]". Furthermore, opinion terms extracted from task 0 help to identify the corresponding opinion terms in task 1, which means that the multi-task structure is able to achieve better results than the single-task structure on TOWE. Although the improvement is not significant in average, we observe that the former structure can achieve more stable performance than the latter one.

| Models | (Fan et al., 2019) | | | |
|---|---|---|---|---|
| | 14lap | 14res | 15res | 16res |
| TSMSA(random initialized) | 56.29 | 69.05 | 59.44 | 71.59 |
| TSMSA(Glove embedding) | 71.10 | 80.31 | 75.38 | 80.68 |
| TSMSA(BERT embedding) | 61.23 | 70.12 | 62.12 | 72.47 |
| TSMSA(BERT fixed) | 65.13 | 72.37 | 66.79 | 73.80 |
| TSMSA(BERT fine-tuned) | 82.18 | 86.37 | 81.64 | 89.20 |
| MT-TSMSA(BERT fine-tuned) | 82.41 | 86.52 | 81.92 | 89.56 |

Table 5: Results of ablation experiments (F1 score, %) on TOWE. The different word embeddings and training strategies of the models are described in parentheses.

### 4.6 Convergence and Sensitivity Studies

The results of convergence and sensitivity studies are shown in Figure 2. Figure 2 (a) reveals that our model gradually converges as the number of epochs increases. Although the dropout rate is set to 0.5, it also converges smoothly. Figure 2 (b) shows the effect of the number of attention heads. When the number of attention heads is 4, TSMSA(Base) achieves stable and good performance, and as the value increased, the performance might be better. Figure 2 (c) shows that the best performance is achieved when the number of multi-head self-attention layers is 6, and as the number increased, the model might be confronted with overfitting. Figure 2 (d) indicates the impact of $\lambda$ on our model which influences the learning of different tasks. Stable and good results can be obtained when $\lambda = 1$, and better performance can be achieved when the value is set to 0.5 or 2. Compared with other hyper-parameters, the results also indicate that $\lambda$ has a relatively small impact on the model performance.



(a) Loss.

(b) The number of attention heads.

(c) The number of layers.

(d) $\lambda$.

Figure 2: (a) is the decline trend of loss. (b), (c), and (d) are the comparisons among different values of layers, attention heads, and $\lambda$, respectively.

### 4.7 Visualization of Attention

In this part, we apply an open source tool[4] to visualize the attention scores of TSMSA(BERT) and describe two attention heads on the tenth layer in Figure 3 (a) and (b), where attention scores less than 0.1 and unimportant words are not displayed. As we can see, the words "nice" and "great" are both close to the aspect "food", but "nice" will not pay attention to this aspect. In addition, "great" and "reasonable" focus on the special symbol "[SEP]" and the specific aspect "food", as shown in Figure 3 (a). At the same time, "food" gives attention to "great" and "reasonable" on different attention heads, as described in Figure 3 (b). All these instances reveal that multi-head self-attention mechanism is capable of capturing the representation of a

[4] https://github.com/jessevig/bertviz

| |
|---|
| Case 1: The receiver was full of superlatives for the quality and performance.<br>**SDRN**: (quality, superlatives), (performance, superlatives) ✓<br>**MT-TSMSA(BERT)**: (quality, superlatives), (performance, superlatives) ✓ |
| Case 2: The selection of food is excellent, and the atmosphere is great.<br>**SDRN**: (selection of food, excellent), (atmosphere, great) ✓<br>**MT-TSMSA(BERT)**: (selection of food, excellent), (atmosphere, great) ✓ |
| Case 3: The bartenders and the managers are really nice and the decor is very comfy and laid-back, all the while being trendy.<br>**SDRN**: (bartenders, nice), (managers, nice), (decor, comfy), (decor, laid-back), (decor, trendy) ✓<br>**MT-TSMSA(BERT)**: (bartenders, nice), (managers, nice), (decor, comfy), (decor, laid-back), (decor, trendy) ✓ |
| Case 4: Additionally, there is barely a ventilation system in the computer, and even the simple activity of watching videos let alone playing steam games causes the laptop to get very very hot, and in fact impossible to keep on lap.<br>**SDRN**: (ventilation system, barely), (ventilation system, hot), (watching videos, simple),<br>(playing steam games, hot) missed (watching videos, hot)<br>**MT-TSMSA(BERT)**: (ventilation system, barely), (ventilation system, hot), (watching videos, simple),<br>(watching videos, hot), (playing steam games, hot) ✓ |
| Case 5: Every time I log into the system after a few hours, there is this endlessly frustrating process that I have to go through.<br>**SDRN**: (log into the system, frustrating) ✓<br>**MT-TSMSA(BERT)**: missed (log into the system, frustrating) |

Table 6: Case study results. The aspect and opinion terms are highlighted in green and blue, respectively. The extracted pairs from SDRN and MT-TSMSA are shown in parenthesis, where the missed pairs are marked in red.

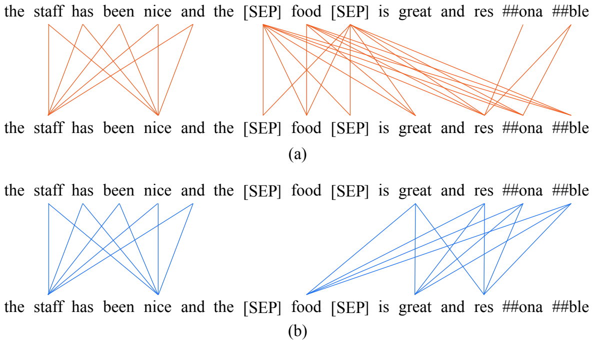

the staff has been nice and the [SEP] food [SEP] is great and res ##ona ##ble
the staff has been nice and the [SEP] food [SEP] is great and res ##ona ##ble
(a)
the staff has been nice and the [SEP] food [SEP] is great and res ##ona ##ble
the staff has been nice and the [SEP] food [SEP] is great and res ##ona ##ble
(b)

Figure 3: Visualization of multi-head self-attention mechanism. A line represents that a word from the bottom sentence pays close attention to the word from the top sentence.

specific aspect.

### 4.8 Case Study

To further compare our MT-TSMSA(BERT) with the best-performing baseline of SDRN, we here conduct a case study by following (Chen et al., 2020). As shown in Table 6, both SDRN and MT-TSMSA(BERT) perform well in extracting aspect-opinion pairs from complicated relations. But in some cases like Case 4, SDRN misses the pair of (watching videos, hot). The reason may be that the massive hyper-parameters in SDRN have a great impact on the effect. For example, the threshold $\beta$ in the relation synchronization mechanism of SDRN will largely affect the results of the model. On the other hand, our method can extract all the pairs because it introduces fewer hyper-parameters, which leads to stable results. However, in Case 5, our method cannot extract the pair. The reason is that task 0 of MT-TSMSA(BERT) fails to extract the aspect term "log into the system". Moreover, the in-depth reason is that for the aspect term extraction task, the performance of SDRN (i.e., 83.67%, 89.49%, and 74.05%) is better than that of MT-TSMSA(BERT), i.e., 83.11%, 84.85%, and 72.69% on the datasets from (Chen et al., 2020).

## 5 Conclusions

In this paper, we propose a target-specified sequence labeling method based on multi-head self-attention (TSMSA) and a multi-task version (MT-TSMSA) to deal with TOWE and AOPE, respectively. In our methods, the encoder is capable of capturing the information of the specific aspect which is labeled by a special symbol "[SEP]". Experimental results demonstrate that TSMSA and MT-TSMSA achieve quite competitive performance in most cases. When combining aspect and opinion words extraction with TOWE, our MT-TSMSA can slightly improve the performance as compared with TSMSA. In the future, we plan to extend our approaches to sentiment classification of pairs and explore an efficient model with a one-stage inference process to reduce the time complexity on AOPE.

## Acknowledgment

We are grateful to the reviewers for their constructive comments and suggestions on this study. This work has been supported by the National Natural Science Foundation of China (61972426) and Guangdong Basic and Applied Basic Research Foundation (2020A1515010536).

## References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Syst. Appl.*, 114:34–45.

Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi. 2020. Synchronous double-channel recurrent network for aspect-opinion pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6515–6524.

Hongliang Dai and Yangqiu Song. 2019. Neural aspect and opinion term extraction with mined rules as weak supervision. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 5268–5277.

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

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

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4194–4200.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2892.

Pengfei Liu, Shafiq R. Joty, and Helen M. Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT*, pages 486–495.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING*, pages 27–35.

Soujanya Poria, Erik Cambria, and Alexander F. Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowl. Based Syst.*, 108:42–49.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Comput. Linguistics*, 37(1):9–27.

Lance A. Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the 3rd Workshop on Very Large Corpora, VLC@ACL*.

Lei Shu, Hu Xu, and Bing Liu. 2017. Lifelong learning CRF for supervised aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 148–154.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2895–2905.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 5998–6008.

Andrew J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, 13(2):260–269.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 616–626.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3316–3322.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. *CoRR*, abs/2010.04640.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 592–598.

Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised word and dependency path embeddings for aspect term extraction. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 2979–2985.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3239–3248.

Li Zhuang, Feng Jing, and Xiaoyan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 2006 ACM International Conference on Information and Knowledge Management*, pages 43–50.