Grid Tagging Scheme for Aspect-oriented Fine-grained Opinion Extraction

Zhen Wu[1] **Chengcan Ying**[1] **Fei Zhao**[1] **Zhifang Fan**[1] **Xinyu Dai**[1] * **Rui Xia**[2]
[1]National Key Laboratory for Novel Software Technology, Nanjing University
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology
{wuz, yingcc, zhaof, fanzf}@smail.nju.edu.cn
daixinyu@nju.edu.cn, rxia@njust.edu.cn

Abstract

Aspect-oriented Fine-grained Opinion Extraction (AFOE) aims at extracting aspect terms and opinion terms from review in the form of opinion pairs or additionally extracting sentiment polarity of aspect term to form opinion triplet. Because of containing several opinion factors, the complete AFOE task is usually divided into multiple subtasks and achieved in the pipeline. However, pipeline approaches easily suffer from error propagation and inconvenience in real-world scenarios. To this end, we propose a novel tagging scheme, Grid Tagging Scheme (GTS), to address the AFOE task in an end-to-end fashion only with one unified grid tagging task. Additionally, we design an effective inference strategy on GTS to exploit mutual indication between different opinion factors for more accurate extractions. To validate the feasibility and compatibility of GTS, we implement three different GTS models respectively based on CNN, BiLSTM, and BERT, and conduct experiments on the aspect-oriented opinion pair extraction and opinion triplet extraction datasets. Extensive experimental results indicate that GTS models outperform strong baselines significantly and achieve state-of-the-art performance.

1 Introduction

Aspect-oriented Fine-grained Opinion Extraction (AFOE) aims to automatically extract opinion pairs (*aspect term*, *opinion term*) or opinion triplets (*aspect term*, *opinion term*, *sentiment*) from review text, which is an important task for fine-grained sentiment analysis (Pang and Lee, 2007; Liu, 2012). In this task, aspect term and opinion term are two key opinion factors. Aspect term, also known as opinion target, is the word or phrase in a sentence representing feature or entity of products or services. Opinion term refers to the term in a sentence used to express attitudes or opinions explicitly. For example, in the sentence of Figure 1, "*hot dogs*" and "*coffee*" are two aspect terms, "*top notch*" and "*average*" are two opinion terms.

Sentence: The hot dogs are top notch but average coffee	
Aspect Terms:	hot dogs, coffee
Opinion Terms:	top notch, average
Opinion Pairs:	(hot dogs, top notch), (coffee, average)
Opinion Triplets:	(hot dogs, top notch, positive), (coffee, average, neutral)

Figure 1: An example of aspect-oriented fine-grained opinion extraction. The spans highlighted in red are aspect terms. The terms in blue are opinion terms.

To obtain the above two opinion factors, many works devote to the co-extraction of aspect term and opinion term in a joint framework (Wang et al., 2016, 2017; Li and Lam, 2017; Yu et al., 2019; Dai and Song, 2019). However, the extracted results of these works are two separate sets of aspect term and opinion term, and they neglect the pair relation between them, which is crucial for downstream sentiment analysis tasks and has many potential applications, such as providing sentiment clues for aspect level sentiment classification (Pontiki et al., 2014), generating fine-grained opinion summarization (Zhuang et al., 2006) or analyzing in-depth opinions (Kobayashi et al., 2007), etc.

Opinion pair extraction (OPE) is to extract all opinion pairs from a sentence in the form of (*aspect term*, *opinion term*). An opinion pair consists of an aspect term and a corresponding opinion term. This task needs to extract three opinion factors, i.e., aspect terms, opinion terms, and the pair relation between them. Figure 1 shows an example. We can see that the sentence "*the hot dogs are top notch and great coffee!*" contains two opinion pairs, respectively (*hot dogs*, *top notch*) and (*coffee*, *average*) (the former is the aspect term, and latter

* Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2576–2585
November 16 - 20, 2020.

represents the corresponding opinion term). OPE sometimes could be complicated because an aspect term may correspond to several opinion terms and vice versa. Despite the great importance of OPE, it is still under-investigated, and only a few early works mentioned or explored this task (Hu and Liu, 2004; Zhuang et al., 2006; Klinger and Cimiano, 2013b; Yang and Cardie, 2013).

By reviewing the aspect-based sentiment analysis (ABSA) (Pontiki et al., 2014) research, we can summarize two types of state-of-the-art pipeline approaches to extract opinion pairs: (I). Co-extraction (Wang et al., 2017; Dai and Song, 2019)+Pair relation Detection (PD) (Xu et al., 2018); (II). Aspect term Extraction (AE) (Xu et al., 2018)+Aspect-oriented Opinion Term Extraction (AOTE) (Fan et al., 2019). Nevertheless, pipeline approaches easily suffer from error propagation and inconvenience in real-world scenarios.

To address the above issues and facilitate the research of AFOE, we propose a novel tagging scheme, **G**rid **T**agging **S**cheme (GTS), which transforms opinion pair extraction into one unified grid tagging task. In this grid tagging task, we tag all word-pair relations and then decode all opinion pairs simultaneously with our proposed decoding method. Accordingly, GTS can extract all opinion factors of OPE in one step, instead of pipelines. Furthermore, different opinion factors are mutually dependent and indicative in the OPE task. For example, if we know "*average*" is an opinion term in Figure 1, then "*coffee*" is probably deduced as an aspect term because "*average*" is its modifier. To exploit these potential bridges, we specially design an inference strategy in GTS to yield more accurate opinion pairs. In the experiments, we implement three GTS models, respectively, with CNN, LSTM, and BERT, to demonstrate the effectiveness and compatibility of GTS.

Besides OPE, we find that GTS is very easily extended to aspect-oriented Opinion Triplet Extraction (OTE), by replacing the pair relation detection of OPE with specific sentiment polarity detection. OTE is a new fine-grained sentiment analysis task and aims to extract all opinion triplets (*aspect term*, *opinion term*, *sentiment*) from a sentence (Peng et al., 2019). To tackle the task, Peng et al. (2019) propose a two-stage framework and still extract opinion pair (*aspect term*, *opinion term*) in pipeline, thus suffering from error propagation. In contrast, GTS can extract all opinion triplets simultaneously only with one unified grid tagging task.

The main contributions of this work can be summarized as follows:

- We propose a novel tagging scheme, Grid Tagging Scheme (GTS). To the best of our knowledge, GTS is the first work to address the complete aspect-oriented fine-grained opinion extraction, including OPE and OTE, with one unified tagging task instead of pipelines. Besides, this new scheme is easily extended to other pair/triplet extraction tasks from text.
- For the potential mutual indications between different opinion factors, we design an effective inference strategy on GTS to exploit them for more accurate extractions.
- We implement three GTS neural models respectively with CNN, LSTM, and BERT, and conduct extensive experiments on both tasks of OPE and OTE to verify the compatibility and effectiveness of GTS.

The following sections are organized as follows. Section 2 presents our proposed Grid Tagging Scheme. In Section 3, we introduce the models based on GTS and the inference strategy. Section 4 shows experiment results. Section 5 and Section 6 are respectively related work and conclusions. Our code and data will be available at https://github.com/NJUNLP/GTS.

2 Grid Tagging Scheme

In this section, we first give the task definition of Opinion Pair Extraction (OPE) and Opinion Triplet Extraction (OTE), then explain how the two tasks are represented in Grid Tagging Scheme. Finally, we present how to decode opinion pairs or opinion triplets according to the tagging results in GTS.

2.1 Task Definition

We first introduce the definition of the OPE task. Given a sentence $s = \{w_1, w_2, \cdots, w_n\}$ consisting n words, the goal of the OPE task is to extract a set of opinion pairs $\mathcal{P} = \{(a, o)_m\}_{m=1}^{|\mathcal{P}|}$ from the sentence s, where $(a, o)_m$ is an opinion pair in s. The notations a and o respectively denote an aspect term and an opinion term. They are two non-overlapped spans in s.

As for the OTE task, it additionaly extracts the corresponding sentiment polarity of each opinion pair (a, o), i.e., extracting a set of opinion tripelts

$\mathcal{T} = \{(a, o, c)_m\}_{m=1}^{|\mathcal{T}|}$ from the given sentence s, where c denotes the sentiment polarity and $c \in$ {positive, neutral, negative}.

2.2 Grid Tagging

To tackle the OPE task, Grid Tagging Scheme (GTS) uses four tags {A, O, P, N} to represent the relation of any word-pair (w_i, w_j) in a sentence. Here the word-pair (w_i, w_j) is unordered and thus word-pair (w_i, w_j) and (w_j, w_i) have the same relation. The meanings of four tags can be seen in Table 1. In GTS, the tagging result of a sentence is like a grid after displaying it in rows and columns. For simplicity, we adopt an upper triangular grid. Figure 2 shows the tagging results of the sentence of Figure 1 in GTS.

Tags	Meanings
A	two words of word-pair (w_i, w_j) belong to the same aspect term.
O	two words of word-pair (w_i, w_j) belong to the same opinion term.
P	two words of word-pair (w_i, w_j) respectively belong to an aspect term and an opinion term, and they form opinion pair relation.
N	no above three relations for word-pair (w_i, w_j).

Table 1: The meanings of tags for the OPE task.

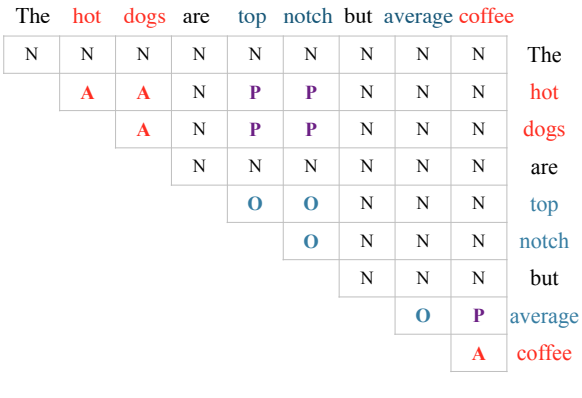

Figure 2: A tagging example with GTS for the OPE task. In the sentence, the spans highlighted in red are aspect terms and the spans in blue are opinion terms.

Specifically, the tag A represents that the two words of word-pair (w_i, w_j) belong to the same aspect term. For example, the position of word-pair (*hot*, *dogs*) in Figure 2 is the tag A. Similarly, the tag O indicates that the two words of word-pair (w_i, w_j) exist in the same aspect term. Notably, GTS also considers the word-pair (w_i, w_i), i.e., the relation of each word to itself, which can help represent a single-word aspect term or opinion term. The tag P represents that two words of word-pair (w_i, w_j) respectively belong to an aspect term and an opinion term, and the two terms are an opinion pair, such as the word-pair (*hot*, *top*) and (*dogs*, *top*) in Figure 2. The last tag N denotes no relation between word-pair (w_i, w_j).

To deal with the OTE task, GTS replaces the previous P tag with the specific sentiment label. To be specific, GTS adopts the tag set {A, O, Pos, Neu, Neg, N} to denote the relation of word-pair in the OTE task. The three tags Pos, Neu, Neg respectively indicate positive, neutral, or negative sentiment expressed in the opinion triplet consisting of the word-pair (w_i, w_j). A tagging example of the OTE task is shown in Figure 3.

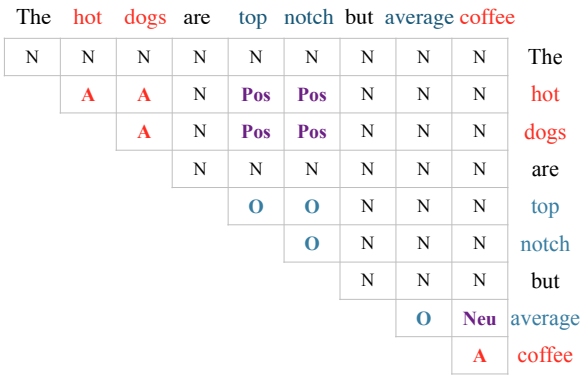

Figure 3: A tagging example for the OTE task.

It can be concluded that Grid Tagging Scheme successfully transforms end-to-end aspect-oriented fine-grained opinion extraction into a unified tagging task by labeling the relations of all word-pairs.

2.3 Decoding Algorithm

In this subsection, we focus on how to decode the final opinion pairs or opinion triplets according to the tagging results of all word-pairs. In fact, various methods can be applied to obtaining these tagging results, and we adopt neural network models in this work (see Section 3).

After obtaining the predicted tagging results of a sentence in GTS, we can extract opinion pairs or opinion triplets by strictly matching the relations of word-pairs as in Figure 2 and Figure 3. However, it might get low recall due to abundant N tags in GTS. To address this issue, we relax matching constraints and design a simple but effective method to decode opinion pair or opinion triplet.

The decoding details for the OPE task are shown in Algorithm 1. Firstly, we use the predicted tags of all (w_i, w_i) word-pairs on the main diagonal to

Algorithm 1 Decoding Algorithm for OPE

Input: The tagging results T of a sentence in GTS. $T(w_i, w_j)$ denotes the predicted tag of the word-pair (w_i, w_j).
Output: Opinion pair set $\mathcal{P}$ of the given sentence.

1: Initialize the aspect term set $\mathcal{A}$, opinion term set $\mathcal{O}$, and opinion pair set $\mathcal{P}$ with $\varnothing$.
2: **while** a span left index $l \leq n$ and right index $r \leq n$ **do**
3: **if** all $T(w_i, w_i) = \texttt{A}$ when $l \leq i \leq r$, meanwhile $T(w_{l-1}, w_{l-1}) \neq \texttt{A}$ and $T(w_{r+1}, w_{r+1}) \neq \texttt{A}$ **then**
4: Regard the words $\{w_l, \cdots, w_r\}$ as an aspect term a, $\mathcal{A} \leftarrow \mathcal{A} \cup \{a\}$
5: **end if**
6: **if** all $T(w_i, w_i) = \texttt{O}$ when $l \leq i \leq r$, meanwhile $T(w_{l-1}, w_{l-1}) \neq \texttt{O}$ and $T(w_{r+1}, w_{r+1}) \neq \texttt{O}$ **then**
7: Regard the words $\{w_l, \cdots, w_r\}$ as an opinion term o, $\mathcal{O} \leftarrow \mathcal{O} \cup \{o\}$
8: **end if**
9: **end while**
10: **while** $a \in \mathcal{A}$ and $o \in \mathcal{O}$ **do**
11: **while** $w_i \in a$ and $w_j \in o$ **do**
12: **if** any $T(w_i, w_j) = \texttt{P}$ **then**
13: $\mathcal{P} \leftarrow \mathcal{P} \cup \{(a, o)\}$
14: **end if**
15: **end while**
16: **end while**
17: **return** the set $\mathcal{P}$

recognize aspect terms and opinion terms, without considering other word-pair constraints. As line 2 to line 9 of Algorithm 1 shows, the spans comprised of continuous A tags are regarded as aspect terms, and spans consisting of continuous O are detected as opinion terms. For an extracted aspect term a and an opinion term o, we think they form an opinion pair on condition that at least one word-pair (w_i, w_j) is labeled with the tag P when $w_i \in a$ and $w_j \in o$, as shown in line 11 to line 15.

For the OTE task, the decoding part is different from the OPE task from line 11 to line 15 of Algorithm 1. Specifically, we count the predicted tags of all word-pairs (w_i, w_j) when $w_i \in a$ and $w_j \in o$. The most predicted sentiment tag $c \in \{\texttt{Pos}, \texttt{Neu}, \texttt{Neg}\}$ is regarded as the sentiment polarity of the opinion triplet (a, o, c). If their predicted tags do not belong to $\{\texttt{Pos}, \texttt{Neu}, \texttt{Neg}\}$, we think a and o cannot form an opinion triplet.

3 Validation Models

To verify the effectiveness and good compatibility of GTS, we respectively tried three typical neural networks, i.e., CNN, LSTM, and BERT, as encoder implementations of GTS (Section 3.1). Besides, different opinion factors in AFOE mutually rely on and can benefit each other. Therefore, we design an inference strategy to exploit these potential indications in Section 3.2. Figure 4 shows the overall architecture of GTS models.

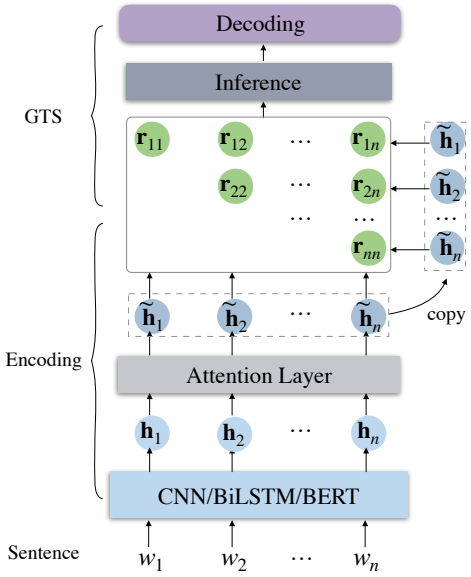

Figure 4: The overall architecture of neural models based on GTS.

3.1 Encoding

Given a sentence $s = \{w1, w_2, \cdots, w_n\}$, CNN, BiLSTM or BERT can be used as the encoder of GTS to generate the representation $\mathbf{r}_{ij}$ of the word-pair (w_i, w_j).

CNN. We follow the design of state-of-the-art aspect term extraction model DE-CNN (Xu et al., 2018). It employs 2 embedding layers and a stack of 4 CNN layers to encode the sentence s, then generates the feature representation $\mathbf{h}_i$ for each word w_i. Dropout (Srivastava et al., 2014) is applied after the embedding and each ReLU activation. The details can be found in Xu et al. (2018).

BiLSTM. BiLSTM employs a standard forward Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and a backward LSTM to encode the sentence, then concatenate the hidden states in two LSTMs as the representation $\mathbf{h}_i$ of each word w_i.

BERT. BERT adopts subwords embedding, position embedding and segment embedding as the representation of subword, then employs a multi-layer bidirectional Transformer (Vaswani et al., 2017) to generate the contextual represenations $\{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n\}$ of the given sentence s. For a more comprehensive description, readers can refer to Devlin et al. (2019).

To obtain a robust representation for word-pair (w_i, w_j), we additionally employ an attention layer to enhance the connection between w_i and w_j. The

details are as follows:

$$u_{ij} = \mathbf{v}^\top(\mathbf{W}_{a1}\mathbf{h}_i + \mathbf{W}_{a2}\mathbf{h}_j + \mathbf{b}_a), \quad (1)$$

$$\alpha_{ij} = \frac{\exp(u_{ij})}{\sum_{k=1}^{n}\exp(u_{ik})}, \quad (2)$$

$$\widetilde{\mathbf{h}}_i = \mathbf{h}_i + \sum_{j=1}^{n}\alpha_{ij}\mathbf{h}_j, \quad (3)$$

where $\mathbf{W}_{a1}$ and $\mathbf{W}_{a2}$ are weight matrices, and $\mathbf{b}_a$ is the bias. Note that, the above attention is not applied on the representations of BERT, because BERT itself contains multiple self-attention layers.

Finally, we concatenate the enhanced representations of w_i and w_j to represent the word-pair (w_i, w_j), i.e., $\mathbf{r}_{ij} = [\widetilde{\mathbf{h}}_i; \widetilde{\mathbf{h}}_j]$, where $[\cdot;\cdot]$ denotes the vector concatenation operation.

3.2 Inference on GTS

As aforementioned, different opinion factors of AFOE are mutually indicative. Therefore, we design the inference strategy in GTS to exploit these potential indications for facilitating AFOE.

In Grid Tagging Scheme, let us consider what is helpful to detect the relation of word-pair (w_i, w_j). First, relations between w_i and other words (except w_j) can help detection. For example, if predicted tags of word-pairs consisting of w_i contain A, the tag of word-pair (w_i, w_j) is less possible to be O and vice versa. So does the word w_j. Second, the previous prediction for (w_i, w_j) heps infer the tag of (w_i, w_j) of the current turn. To this end, we propose an inference strategy on GTS to exploit these indications by iterative prediction and inference. In the t-th turn, the feature representation $\mathbf{z}_{ij}^t$ and predicted probability distribution $\mathbf{p}_{ij}^t$ of word-pair (w_i, w_j) can be calculated as follows:

$$\mathbf{p}_i^{t-1} = \text{maxpooling}(\mathbf{p}_{i,:}^{t-1}), \quad (4)$$

$$\mathbf{p}_j^{t-1} = \text{maxpooling}(\mathbf{p}_{j,:}^{t-1}), \quad (5)$$

$$\mathbf{q}_{ij}^{t-1} = [\mathbf{z}_{ij}^{t-1}; \mathbf{p}_i^{t-1}; \mathbf{p}_j^{t-1}; \mathbf{p}_{ij}^{t-1}], \quad (6)$$

$$\mathbf{z}_{ij}^t = \mathbf{W}_q\mathbf{q}_{ij}^{t-1} + \mathbf{b}_q, \quad (7)$$

$$\mathbf{p}_{ij}^t = \text{softmax}(\mathbf{W}_s\mathbf{z}_{ij}^t + \mathbf{b}_s). \quad (8)$$

In the above process, $\mathbf{p}_{i,:}^{t-1}$ represents all predicted probability between the word w_i and other words. In fact, $\mathbf{p}_{i,:}^{t-1} = (\mathbf{p}_{1:i,i}^{t-1}, \mathbf{p}_{i,i:n}^{t-1})$ in GTS as we use the upper triangular grid. Equation 4 and 5 aim to help infer the possible tags for (w_i, w_j) by observing predictions between w_i/w_j and other words. The initial predicted probability $\mathbf{p}_{ij}^0$ and representation $\mathbf{z}_{ij}^0$ of (w_i, w_j) is set as:

$$\mathbf{p}_{ij}^0 = \text{softmax}(\mathbf{W}_s\mathbf{r}_{ij} + \mathbf{b}_s), \quad (9)$$

$$\mathbf{z}_{ij}^0 = \mathbf{r}_{ij}. \quad (10)$$

Finally, the prediction $\mathbf{p}_{ij}^L$ in the final turn is used to extract fine-grained opinions according to Algorithm 1. The L is a hyperparameter denoting the inference times.

3.3 Training Loss

We use y_{ij} to represent the ground truth tag of the word-pair (w_i, w_j). The unified training loss for AFOP is defined as the cross entropy loss between grouhd truth distribution and predicted tagging distribution $\mathbf{p}_{ij}^L$ of all word-pairs:

$$\mathcal{L} = -\sum_{i=1}^{n}\sum_{j=i}^{n}\sum_{k\in C}\mathbb{I}(y_{ij}=k)\log(p_{i,j|k}^L), \quad (11)$$

where $\mathbb{I}(\cdot)$ is the indicator function, and C denotes the label set. In the OPE task, C is $\{\texttt{A}, \texttt{O}, \texttt{P}, \texttt{N}\}$. For the OTE task, the set C is $\{\texttt{A}, \texttt{O}, \texttt{Pos}, \texttt{Neu}, \texttt{Neg}, \texttt{N}\}$.

4 Experiments

4.1 Datasets and Metrics

Datasets		#S	#A	#O	#P	#T
14res	Train	1,259	2,064	2,098	2,356	2,356
	Dev	315	487	506	580	580
	Test	493	851	866	1,008	1,008
14lap	Train	899	1,257	1,270	1,452	1,452
	Dev	225	332	313	383	383
	Test	332	467	478	547	547
15res	Train	603	871	966	1,038	1,038
	Dev	151	205	226	239	239
	Test	325	436	469	493	493
16res	Train	863	1,213	1,329	1,421	1,421
	Dev	216	298	331	348	348
	Test	328	456	485	525	525

Table 2: Statistics of aspect-oriented fine-grained opinion extraction datasets. Here "#S", "#A", "#O", "#P", and "#T" respectively denote the numbers of sentence, aspect term, opinon term, opinion pair, and opinion triplet. The "res" and "lap" represent datasets from restaurant domain or laptop domain.

To study aspect-oriented opinion term extraction, Fan et al. (2019) annotate and release four opinion pair datasets[1] based on SemEval Challenges (Pontiki et al., 2014, 2015, 2016). However, they do not annotate the sentiment polarity of

[1]https://github.com/NJUNLP/TOWE

Methods	14res			14lap			15res			16res		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Pipeline: Co-extraction+The Pair Relation Detection												
CMLA+Dis-BiLSTM	**77.21**	52.14	62.24	59.47	45.23	51.17	64.86	44.33	52.47	66.29	50.82	57.33
CMLA+C-GCN	72.22	56.35	63.17	60.69	47.25	53.03	64.31	49.41	55.76	66.61	59.23	62.70
RINANTE+C-GCN	71.07	59.45	64.69	67.38	52.10	58.76	65.52	42.74	51.73	-	-	-
Pipeline: Aspect Term Extraction+Aspect-oriented Opinion Term Extraction												
BiLSTM-ATT+Distance	47.09	39.40	42.90	38.85	29.20	33.34	39.63	33.95	36.57	43.60	39.65	41.53
BiLSTM-ATT+Dependency	56.31	48.93	52.36	31.58	28.84	30.15	58.26	42.19	48.94	64.48	48.85	55.59
BiLSTM-ATT+IOG	69.99	61.58	65.46	64.93	44.56	52.84	59.14	56.38	57.73	66.07	62.55	64.13
DE-CNN+IOG	67.70	69.41	68.55	59.59	51.68	55.35	56.18	60.08	58.04	62.97	66.22	64.55
RINANTE+IOG	70.16	65.47	67.74	61.76	53.11	57.10	63.24	55.57	59.16	-	-	-
Our End-to-End GTS Models												
GTS-CNN	74.13	69.49	71.74	**68.33**	55.04	60.97	66.81	61.34	63.96	70.48	72.39	71.42
GTS-BiLSTM	71.32	67.07	69.13	61.53	54.31	57.69	**67.76**	63.19	65.39	70.32	70.46	70.39
GTS-BERT	76.23	**74.84**	**75.53**	66.41	**64.95**	**65.67**	66.40	**68.71**	**67.53**	**71.70**	**77.79**	**74.62**

Table 3: The experiment results on the OPE task (%). Best and second-best results are respectively in bold and underline. The marker "-" represents that the original code of RINANTE method does not contain necessary resources for running on the dataset 16res.

each opinion pair. The original SemEval Challenge datasets provide the annotation of aspect terms and the corresponding sentiment, while not the corresponding opinion terms. Thus we align the datasets of Fan et al. (2019) and original SemEval Challenge datasets to build AFOE datasets. Table 2 shows their statistics, and we can observe that one sentence may contain multiple aspect terms or opinion terms. Besides, one aspect term may correspond to multiple opinion terms and vice versa.

To evaluate the performance of different methods, we use precision, recall, and F1-score as the evaluation metrics. The extracted aspect terms and opinion terms are regarded as correct only if predicted and ground truth spans are exactly matched.

4.2 Experimental Settings

Following the design of DE-CNN (Xu et al., 2018), we use double embeddings to initialize the word vectors of GTS-CNN and GTS-BiLSTM, which contains a domain-general embedding from 300-dimension GloVe (Pennington et al., 2014) pre-trained with 840 billion tokens and a 100-dimension domain-specific embedding trained with fastText (Bojanowski et al., 2017). The CNN kernel size on domain-specific embedding is 3 and others are 5. In GTS-BiLSTM, the dimension of LSTM cell is set to 50. We adopt Adam optimizer (Kingma and Ba, 2015) to optimize networks and the initial learning rate is 0.001. The dropout (Srivastava et al., 2014) is applied after embedding layer with probability 0.5. As for GTS-BERT, we use uncased $\text{BERT}_{\text{BASE}}$ version[2] and set the learning rate to 5e-5. The mini-batch size is set to 32. The development set is used for early stopping. We run each model five times and report the average result of them.

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

4.3 Results of Opinion Pair Extraction

Compared Methods We summarize the ABSA studies and combine the state-of-the-art methods as our strong OPE baselines. They include: (I). CMLA (Wang et al., 2017) and RINANTE (Dai and Song, 2019) for the co-extraction of aspect term and opinion term (Co-extraction), Dis-BiLSTM and C-GCN (Zhang et al., 2018) for the Pair relation Detection (PD); (II). BiLSTM-ATT and DE-CNN (Xu et al., 2018) for Aspect term Extraction (AE), Distance (Hu and Liu, 2004), Dependency (Zhuang et al., 2006), and IOG (Fan et al., 2019) for Aspect-oriented Opinion Term Extraction (AOTE). Note that, our GTS models do not use sentiment labels information when performing the OPE task. Table 3 shows the experiment results of different methods.

Observing two types of pipeline methods, we can find that the pipeline of AE+AOTE seems to perform better than Co-extraction+PD. Specifically, the method RINANTE+IOG outperforms RINANTE+C-GCN significantly on the datasets 14res and 15res, though C-GCN is a strong relation classification model. This indicates that the detection of opinion pair relation might be more difficult than aspect-oriented opinion term extraction. Besides, RINANTE+IOG also achieves better performances than another strong method DE-CNN+IOG respectively by the F1-score of 1.75% and 1.12%

Methods	14res			14lap			15res			16res		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Li-unified-R+PD§	41.44	68.79	51.68	42.25	42.78	42.47	43.34	50.73	46.69	38.19	53.47	44.51
Peng-unified-R+PD§	44.18	62.99	51.89	40.40	47.24	43.50	40.97	54.68	46.79	46.76	62.97	53.62
Peng-unified-R+IOG	58.89	60.41	59.64	48.62	45.52	47.02	51.70	46.04	48.71	59.25	58.09	58.67
IMN+IOG	59.57	63.88	61.65	49.21	46.23	47.68	55.24	52.33	53.75	-	-	-
GTS-CNN	70.79	61.71	65.94	55.93	47.52	51.38	60.09	53.57	56.64	62.63	**66.98**	64.73
GTS-BiLSTM	67.28	61.91	64.49	**59.42**	45.13	51.30	**63.26**	50.71	56.29	66.07	65.05	65.56
GTS-BERT	**70.92**	**69.49**	**70.20**	57.52	**51.92**	**54.58**	59.29	**58.07**	**58.67**	**68.58**	66.60	**67.58**

Table 4: The experiment results on the OTE task (%). Best and second-best results are respectively in bold and underline. The results with § are retrieved from Peng et al. (2019). The marker "-" represents that the original code of IMN method does not contain necessary resources for running on the dataset 16res.

on the datasets 14lap and 15res, which validates the facilitation of co-extraction strategy for the aspect term extraction.

Compared with the strong pipelines DE-CNN+IOG and RINANTE+IOG, our three end-to-end GTS models all achieve obvious improvements, especially on the datasets 15res and 16res. Despite RINANTE using weak supervision to extend millions of training data, GTS-CNN and GTS-BiLSTM still obtain obvious improvements only through one unified tagging task without additional resources. This comparison shows that error propagations in pipeline methods limit the performance of OPE. There is no doubt that GTS-BERT achieves the best performance because of the powerful ability to model context. The results in Table 3 and above analysis consistently demonstrate the effectiveness of GTS for the OPE task.

4.4 Results of Opinion Triplet Extraction

Compared Method We use the latest OTE work proposed by Peng et al. (2019) as the compared method. In addition, we also employ the state-of-the-art work IMN (He et al., 2019) and the first step of Peng et al. (2019) for extracting the (*aspect term*, *sentiment*) pair, then combine them with IOG as strong baselines. The experiment results are shown in Table 4.

We can observe that IMN+IOG outperforms Peng-unified-R+IOG obviously on the datasets 14res and 15res, because IMN uses multi-domain document-level sentiment classification data as auxiliary tasks. In contrast, GTS-CNN and GTS-BiLSTM still obtain about 3% improvements in F1-score than IMN+IOG without requiring additional document-level sentiment data. The overall experiment results on the OTE task again validate the effectiveness of GTS. Furthermore, GTS-BERT outperforms GTS-CNN and GTS-BiLSTM only about 2%-3% on the datasets 15res and 16res, which to some extent shows the ability of the proposed tagging scheme itself besides BERT encoder.

Methods	14res		15res	
	A	O	A	O
BiLSTM-ATT	79.03	80.55	73.59	73.01
DE-CNN	81.90	80.57	75.24	73.07
CMLA	81.22	80.48	76.03	74.67
RINANTE	81.34	83.33	73.38	75.40
GTS-CNN	81.82	83.07	77.33	75.23
GTS-BiLSTM	81.10	82.62	78.44	75.63
GTS-BERT	**83.82**	**85.04**	**78.22**	**79.31**

Table 5: The results of different methods on the extractions of aspect term and opinion term (%) . The abbreviations "A" and "O" respectively denote the aspect term extraction and opinion term extraction.

4.5 Results of Aspects Term Extraction and Opinion Term Extraction

To further analyze the performance of different methods, we also compare them on extractions of aspect term and opinion term. We only report F1-score of datasets 14res and 15res for limited space. The experiment results are shown in Tabel 5.

Compared to GTS-CNN and GTS-BiLSTM, we can see that RINANTE achieves comparable or better results on the datasets 14res, while it performs worse on the OPE task. This comparison indicates that pipeline methods suffer from error propagation. According to the results on the dataset 15res, our GTS models not only can address the OPE task and OTE task in an end-to-end way, but also improve the performance of aspect term extraction and opinion term extraction. This is because our novel tagging scheme and inference strategy can exploit potential connections between different opinion factors to facilitate extraction.

4.6 Ablation Study

To investigate the effects of the attention mechanism and inference strategy on GTS models, we conduct ablation study on the OPE task. The experiment results are shown in Table 6.

Methods	14res	14lap	15res	16res
	F1	F1	F1	F1
GTS-CNN	71.74	60.97	63.96	71.42
w/o attention	70.33	60.49	63.09	70.88
w/o inference	68.92	57.03	61.81	66.66
GTS-BiLSTM	69.13	57.69	65.39	70.39
w/o attention	68.74	56.73	64.97	69.39
w/o inference	67.55	55.94	62.99	67.06

Table 6: Ablation study on the OPE task (%) .

After removing the attention mechanism, the performance of the model GTS-CNN and GTS-BiLSTM drop slightly, which indicates that the attention mechanism enhances the connection between words. Comparing the full models with the versions w/o inference, we find that the former outperforms the latter significantly on all datasets. It is reasonable because the proposed inference strategy can leverage the potential bridges between different opinion factors and makes more comprehensive predictions. As for the model GTS-BERT w/o inference, it represents that the inference times is 0, and we show its results in the next section.

4.7 Effects of Inference Times

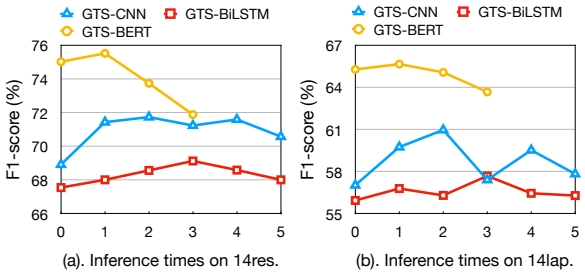

Figure 5: Effects of inference times on GTS models for the OPE task.

To investigate the effects of inference times on performance, we report the results of GTS models for the OPE task on the datasets 14res and 14lap with different inference times in Figure 5.

It can be observed that the inference strategy brings significant improvements for the model GTS-CNN. On the whole, GTS-CNN and GTS-BiLSTM achieve the best results respectively with 2 and 3 inference times on two datasets, and GTS-CNN performs better than GTS-BiLSTM in different inference times. In contrast, GTS-BERT reaches a crest only with 1 time of inference because BERT has contained rich context semantics.

5 Related Work

In literature, only a few works mentioned or explored the opinion pair extraction. Hu and Liu (2004) employ frequent pattern mining to extract aspect terms, then regard the closest adjective to aspect term as the corresponding opinion term. Zhuang et al. (2006) adopt dependency-tree based templates to identify opinion pairs after extracting the aspect term set and opinion term set. Recently, some works adopt neural networks to perform the subtasks of OPE, such as co-extraction of aspect term and opinion term (Wang et al., 2017; Dai and Song, 2019) (Xu et al., 2018). aspect term extraction (Xu et al., 2018), and aspect-oriented opinion term extraction (Fan et al., 2019; Wu et al., 2020), and finally combine them to accomplish OPE in pipeline. To avoid the error propagation of pipeline methods, some studies use joint learning based on traditional machine learning algorithms and hand-crafted features, including Imperatively Defined Factor graph (IDF) (Klinger and Cimiano, 2013a), joint inference based on IDF (Klinger and Cimiano, 2013b), and Integer Linear Programming (ILP) (Yang and Cardie, 2013). However, these methods heavily depend on the quality of hand-crafted features and sometimes perform worse than pipeline methods (Klinger and Cimiano, 2013b).

The opinion triplet extraction is a new aspect-oriented fine-grained opinion extraction task (Peng et al., 2019). Inspired by extracting (*aspect term*, *sentiment*) pair in a joint model (Li et al., 2019; Luo et al., 2019; He et al., 2019), Peng et al. (2019) propose a two-stage framework to extract opinion triplets. In the first stage, they first use a neural model to extract the pair (*aspect term*, *sentiment*) and unpaired opinion terms, then detect the pair relation between aspect term and opinion terms in the second stage. We can see that the key opinon pair extraction of aspect term and opinion term is still accomplished in pipeline and their approach also suffers from error propagation.

6 Conclusions

Aspect-oriented fine-grained opinion extraction (AFOE), including opinion pair extraction (OPE) and opinion triplet extraction (OTE), is usually

achieved in the pipeline because of referring to multiple opinion factors, thereby suffering from error propagation. In this paper, we propose a novel scheme, Grid Tagging Scheme (GTS), to address this task in an end-to-end way. Through tagging the relations between all word-pairs, GTS successfully includes all opinion factors extraction of AFOE into a unified grid tagging task, and then uses the designed decoding algorithm to generate opinion pairs or opinion triplets. To exploit the potential mutual indications between different opinion factors, we design an effective inference strategy on GTS. Three different GTS models respectively based on CNN, BiLSTM, and BERT consistently indicate that our methods outperform strong baselines and achieve state-of-the-art performance on the opinion pair extraction and opinion triplet extraction. Further analysis also validates the effectiveness of GTS and the inference strategy.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback. This work was supported by the NSFC (No. 61936012, 61976114) and National Key R&D Program of China (No. 2018YFB1005102).

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

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, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2509–2518.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 504–515.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Roman Klinger and Philipp Cimiano. 2013a. Bidirectional inter-dependencies of subjective expressions and targets and their value for a joint model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 848–854, Sofia, Bulgaria. Association for Computational Linguistics.

Roman Klinger and Philipp Cimiano. 2013b. Joint and pipeline probabilistic models for fine-grained sentiment analysis: Extracting aspects, subjective phrases and their relations. In *13th IEEE International Conference on Data Mining Workshops, ICDM Workshops, TX, USA, December 7-10, 2013*, pages 937–944.

Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 1065–1074.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6714–6721.

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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2886–2892.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Huaishao Luo, Tianrui Li, Bing Liu, and Junbo Zhang. 2019. DOER: dual cross-shared RNN for aspect term-polarity co-extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 591–601.

Bo Pang and Lillian Lee. 2007. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2019. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *CoRR*, abs/1911.01616.

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, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 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 2016, San Diego, CA, USA, June 16-17, 2016*, 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 2015, Denver, Colorado, USA, June 4-5, 2015*, 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 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

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, EMNLP 2016, November 1-4, 2016*, 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 Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3316–3322.

Zhen Wu, Fei Zhao, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2020. Latent opinions transfer network for target-oriented opinion words extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9298–9305.

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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 592–598.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1640–1649.

Jianfei Yu, Jing Jiang, and Rui Xia. 2019. Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. *IEEE ACM Trans. Audio Speech Lang. Process.*, 27(1):168–177.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2205–2215.

Li Zhuang, Feng Jing, and Xiaoyan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, November 6-11, 2006*, pages 43–50.