STAGE: Span Tagging and Greedy Inference Scheme for Aspect Sentiment Triplet Extraction

Shuo Liang\textsuperscript{1,2}, Wei Wei\textsuperscript{1,2,†}, Xian-Ling Mao\textsuperscript{3}, Yuanyuan Fu\textsuperscript{2,4}, Rui Fang\textsuperscript{2,4}, Dangyang Chen\textsuperscript{2,4}

\textsuperscript{1} Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology
\textsuperscript{2} Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL)
\textsuperscript{3} Department of Computer Science and Technology, Beijing Institute of Technology
\textsuperscript{4} Ping An Property & Casualty Insurance company of China, Ltd
shuoliang@hust.edu.cn, weiw@hust.edu.cn, maoxl@bit.edu.cn, fuyuanyuan83@gmail.com, fangrui051@pingan.com.cn, chendangyang273@pingan.com.cn

Abstract

Aspect Sentiment Triplet Extraction (ASTE) has become an emerging task in sentiment analysis research, aiming to extract triplets of the aspect term, its corresponding opinion term, and its associated sentiment polarity from a given sentence. Recently, many neural networks based models with different tagging schemes have been proposed, but almost all of them have their limitations: heavily relying on 1) prior assumption that each word is only associated with a single role (e.g., aspect term, or opinion term, etc.) and 2) word-level interactions and treating each opinion/aspect as a set of independent words. Hence, they perform poorly on the complex ASTE task, such as a word associated with multiple roles or an aspect/opinion term with multiple words. Hence, we propose a novel approach, \textit{Span Tagging and Greedy inference} (STAGE), to extract sentiment triplets in span-level, where each span may consist of multiple words and play different roles simultaneously. To this end, this paper formulates the ASTE task as a multi-class span classification problem. Specifically, STAGE generates more accurate aspect sentiment triplet extractions via exploring span-level information and constraints, which consists of two components, namely, \textit{span tagging} scheme and \textit{greedy inference} strategy. The former tag all possible candidate spans based on a newly-defined tagging set. The latter retrieves the aspect/opinion term with the maximum length from the candidate sentiment snippet to output sentiment triplets. Furthermore, we propose a simple but effective model based on the STAGE, which outperforms the state-of-the-arts by a large margin on four widely-used datasets. Moreover, our STAGE can be easily generalized to other pair/triplet extraction tasks, which also demonstrates the superiority of the proposed scheme STAGE.

Introduction

Aspect Sentiment Triplet Extraction (ASTE) aims to identify all sentiment triplets from an input sentence, namely, the aspect term ($a$) with its corresponding opinion term ($o$) and sentiment polarity ($s$) (as shown in Figure 1). It can be applied in many downstream tasks, such as dialogue generation (Liu et al. 2022; Wei et al. 2021, 2019) and recommendation system (Wang et al. 2022; Zou et al. 2022; Zhao et al. 2022). Indeed, ASTE can be decomposed into several subtasks, such as Aspect Term Extraction (ATE) (Li and Lam 2017; Xu et al. 2018; Dai and Song 2019), Opinion Term Extraction (OTE) (Fan et al. 2019; Dai and Song 2019), Aspect-Based Sentiment Classification (ABSC) (Zhang, Li, and Song 2019; Tang et al. 2020; Liang et al. 2022), and Aspect-Opinion Pair Extraction (AOPE) (Wang et al. 2017; Zhao et al. 2020). And a straightforward method is to independently extract elements of sentiment triplet through those subtasks in a pipeline fashion (Peng et al. 2020; Chen et al. 2021). However, the primary obstacle for those pipeline methods is the well-known error propagation problem.

Later, many efforts resort to extracting sentiment triplets in an end-to-end framework (Xu et al. 2020; Wu et al. 2020; Zhang et al. 2020; Yu Bai Jian et al. 2021; Xu, Chia, and Bing 2021; Mukherjee et al. et al. 2021; Zhang et al. 2021; Chen et al. 2022; Liu, Li, and Li 2022). Some works (Mukherjee et al. 2021; Zhang et al. 2021) formulate it as a generative problem and achieve good performance under the benefit of existing generative pre-trained models, such as BART (Lewis et al. 2020) and T5 (Raffel et al. 2020). In addition, most of end-to-end approaches focus on design-

<table>
<thead>
<tr>
<th>Sentences</th>
<th>Aspect Term</th>
<th>Opinion Term</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Love the stability of the Mac software and operating system.</td>
<td>stability</td>
<td>Love</td>
<td>Positive</td>
</tr>
<tr>
<td>After replacing the hard drive battery stopped working (3 months of use) which was frustrating.</td>
<td>hard drive</td>
<td>replacing battery</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

Figure 1: Examples of ASTE task from the real-world dataset. In each sentence, aspect terms are in blue and opinion terms are in rounded rectangles.
ing a new tagging scheme, converting ASTE into a classification problem towards each word (Xu et al. 2020) or each word pair (Wu et al. 2020; Chen et al. 2022). Despite their success, those word-level tagging schemes faces several problems: heavily relying on (1) Prior assumption that restricts the diversity of word roles. Those tagging schemes only work when each word is only associated with a single role (the aspect, or the opinion, etc.) or when each word pair corresponds to one specific relation. As a result, they cannot perform well when it comes to a word playing multiple roles or a word pair having multiple relations. (2) Word-level interactions, thus falling short of utilizing span-level information. Those methods only focus on the interactions between words and treat the multi-word term as a set of independent words, thus separating the semantic information of the whole span when facing multi-word aspect/opinion terms. It also causes difficulty in guaranteeing the sentiment consistency when extracting triplets. Without loss of generality, we illustrate those problems with examples in Figure 1: (1) Multiple roles of a word. “Stability” from sentence (a) has two roles simultaneously, that is, being (part of) the aspect term and the opinion term, which confuses the word-level tagging schemes when figuring out the proper tag; (2) Multiple relations of a word-pair. In sentence (b), not only “easy” and “use” are words of the same opinion term (i.e., easy to use), but they form a valid aspect-opinion pair as well, resulting in multiple relations of the word pair “easy-use”, which also challenges the existing tagging scheme; (3) Multi-word term. It is common that an aspect/opinion term contains multiple words, and a significant semantic difference may exist between the whole term and the words in it, such as “hard drive” from sentence (c). Treating those multi-word terms as a set of independent words, the word-level tagging schemes fail to utilize their complete semantics and consequently deteriorate the performance of the extraction task. Moreover, in word-level tagging methods, words that make up a term are classified independently and thus may contain different sentiment polarities. Therefore, extra effort is required to deal with this sentiment inconsistency problem when outputting triplets.

The aforementioned problems motivate us to explore a span-level tagging method and formulate it as a multi-class classification problem towards each span. Generally, it is observed that a span in the given sentences may have three roles: aspect term, opinion term, and a snippet that includes an aspect term and the corresponding opinion term, forming its boundaries (i.e., sentiment snippet), such as “love the stability” from Figure 1 (a). Note that a span can play different roles simultaneously.

Based on the above viewpoint, we pre-define three role dimensions (i.e., aspect term, opinion term, and sentiment snippet) and propose a novel span-level tagging-based approach, Span TAGging and Greedy inference (STAGE), to solve ASTE in an end-to-end fashion. By exploring span-level information and constraints, STAGE can efficiently generate more accurate sentiment triplets. Specifically, it consists of two components: (1) Span tagging scheme: which tags all possible spans based on the defined role dimensions, guaranteeing the diversity of span roles. Thus, it can depict a complete picture of each span, like how many and what roles a span plays. (2) Greedy inference strategy: which decodes all triplets simultaneously in a more accurate and efficient way by (i) considering the mutual span constraints between the sentiment snippets and aspect/opinion terms and (ii) only retrieving the aspect/opinion term with the maximum length from a sentiment snippet, compared to previous decoding algorithms that look into every possible pair combination of candidate aspects and opinions to see whether they are valid. Therefore, naturally performing at the span-level, STAGE can overcome the problems of existing word-level tagging schemes. Based on STAGE, we propose a simple model to demonstrate its effectiveness.

Moreover, with a minor tagging set change, STAGE is very easily generalized to other pair/triplet extraction tasks, such as entity and relation extraction (Zhong and Chen 2021; Yan et al. 2021), a fundamental task aiming to extract (head entity, tail entity, relation) triplets from the given sentence.

In summary, the key contributions are as follows:
(1) To the best of our knowledge, we make the first effort to explore the span-level tagging method and formulate the ASTE task as a multi-class span classification problem.
(2) We propose a novel end-to-end tagging-based approach. Span TAGging and Greedy inference (STAGE). Specifically, it consists of the span tagging scheme that considers the diversity of span roles, overcoming the limitations of existing tagging schemes, and the greedy inference strategy that considers the span-level constraints, generating more accurate triplets efficiently.
(3) We propose a simple but effective model based on STAGE and extensive experiments on four widely-used English datasets verify its effectiveness. Furthermore, it can be easily generalized to other pair/triplet extraction tasks, which also demonstrates the superiority of STAGE.

Related Work

We present some commonly used tagging schemes in the ASTE task, which can be categorized into two main kinds.

Sequence Tagging

The base tagging scheme is BIESO tagging2. Most pipeline models firstly adopt it to extract aspects and opinions separately and then determines whether any two of them can form a valid aspect-opinion pair with the corresponding sentiment. Generally, they view ASTE as two sequence labeling problems (i.e., classification problems towards each word) plus one classification problem towards each candidate aspect-opinion pair. Peng et al. (2020) propose a two-stage pipeline approach and utilize a unified tagging scheme to tag the boundary of aspect term with its sentiment and a common BIESO tagging scheme to extract opinion term.

However, the pipeline approaches ignore the interactions between three triplet elements and commonly suffer from error propagation. Besides, those tagging schemes encode no positional information and thus fail to model rich interactions among aspect and opinion terms. So Xu et al. (2020)

More details are shown in Section Analysis.

BIESO means “begin, inside, end, single, other”, respectively.
propose JET, an extended BIESO tagging scheme with position index information. However, it still fails to simultaneously tackle the situations that one aspect/opinion term relates to multiple opinion/aspect terms.

**Table Tagging**

The basic idea is to tag a \( |n| \times |n| \) table \( T \), where \( n \) is the length of the input text and \( T[i][j] \) corresponds to the relation between \( i \)-th word and \( j \)-th word. Wu et al. (2020) firstly propose a grid tagging scheme, utilizing a tag set \{A, O, POS, NEG, NEU, N\} to donate five possible relations of a word-pair. In order to utilize more relations between words, Chen et al. (2022) define ten relations and consider diverse linguistic features. The above approaches both perform in an end-to-end way, avoiding error prorogation and can simultaneously handle the case of one aspect term corresponding to multiple opinion terms and vice versa. However, as we mentioned before, they still face several limitations.

**Methodology**

In this section, we first introduce the ASTE task and then explain our proposed span tagging and greedy inference (STAGE) scheme in detail. Finally, we present our implemented model.

**Problem Statement**

Given the input sentence \( X = \{w_1, w_2, \ldots, w_n\} \), where \( w_i \) denotes a word and \( n \) is the sentence length. The goal of ASTE task is to extract a set of sentiment triplets \( T = \{(a, o, s)_{m=1}^{[T]}\} \) from the sentence \( X \), where \( a \), \( o \) and \( s \) represent aspect term, opinion term and sentiment polarity separately and \( s \in \{POS, NEG, NEU\} \). We use \( SP = \{SP_{i,1}, SP_{i,2}, \ldots, SP_{i,j}, \ldots, SP_{n,n}\} \) to illustrate all possible enumerated spans in \( X \), where \( i \) and \( j \) represent the start and end positions in \( X \), respectively. So the formal definition for the sentiment snippet is as follows, \( SP_{i,j} \) is an aspect snippet menas \( (SP_{i,k}, SP_{i,j}) \) is a valid (aspect, opinion) or (opinion, aspect) pair, where \( i \leq k, l \leq j \).

**STAGE: Span Tagging and Greedy inference**

**Span Tagging** Considering the diversity of span roles, our span tagging scheme depict each span (i.e., \( SP_{i,j} \)) in three role dimensions: whether \( SP_{i,j} \) is a valid aspect term, a valid opinion term and a valid sentiment snippet. We use \( role_{i,j}^a \), \( role_{i,j}^o \) and \( role_{i,j}^s \) to represent them separately, where \( role_{i,j}^a \in \{N,A\}, role_{i,j}^o \in \{N,O\}, role_{i,j}^s \in \{N, NEG, NEU, POS\} \).

<table>
<thead>
<tr>
<th>Sub Tags</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>an aspect term</td>
</tr>
<tr>
<td>O</td>
<td>an opinion term</td>
</tr>
<tr>
<td>NEG</td>
<td>a sentiment snippet containing negative aspect-opinion pair.</td>
</tr>
<tr>
<td>NEU</td>
<td>a sentiment snippet containing neutral aspect-opinion pair.</td>
</tr>
<tr>
<td>POS</td>
<td>a sentiment snippet containing positive aspect-opinion pair.</td>
</tr>
<tr>
<td>N</td>
<td>nothing in the specific dimension</td>
</tr>
</tbody>
</table>

Table 1: Descriptions of sub tags.

**Greedy Inference** Given the tagging results in our span tagging scheme, in order to generate more accurate triplets efficiently, we propose a greedy inference strategy, where the “greedy” means retrieving the aspect/opinion term with the maximum length separately in a valid sentiment snippet.

The details are shown in Algorithm 1. Firstly, we obtain all possible aspects, opinions and sentiment snippets from the tagging results (line 1 to line 4). Then we look into every sentiment snippet (line 5 to line 24) and consider two cases:

1. **2D-version**: With tag set \{N, O, A\}-\{N, NEG, NEU, POS\}, it can be viewed as projecting the aspect and opinion role dimensions into one. The tag “N” in the first dimension denotes the span is neither an aspect nor an opinion. This variant cannot handle the situations that a span being aspect and opinion simultaneously.

2. **1D-version**: With tag set \{N, NEG, NEU, POS, O, A\}, it can be viewed as projecting the three role dimensions into one. The tag “N” means the span plays no roles. This variant fails to deal with the cases that a span has multiple roles.

- **love of mac software, easy to use.**

---

**Figure 2**: Example of 3D-version span tagging. Cells with gray background represent “N-N-N”.

13176
Algorithm 1: Greedy Inference.

Input:
The tagging results $P$ of the sentence $X$ with length $n$. $P_{i,j}$ represents the tag label of the span $SP_{i,j}$ and can be decomposed into three dimensions $role_{i,j}^a$, $role_{i,j}^o$ and $role_{i,j}^p$.

Output:
Triples $T$ of the given sentence:
1: Initialize $A = \{\}$, $O = \{\}$, $D = \{\}$, $T = \{\}$
2: $A = \{(i,j)\mid role_{i,j}^a = A, 0 \leq i \leq j \leq n\}$
3: $O = \{(i,j)\mid role_{i,j}^o = O, 0 \leq i \leq j \leq n\}$
4: $D = \{(i,j, role_{i,j}^p)\mid role_{i,j}^p \neq N, 0 \leq i \leq j \leq n\}$
5: for each $(i,j,s)$ in $D$ do
6:   // CASE-1: aspect term before opinion term;
7:   $CA = \{k \mid i \leq k \leq j, (i,k) \in A\}$
8:   $CO = \{l \mid i \leq l \leq j, (l,j) \in O\}$
9:   if $CA \neq \emptyset$ and $CO \neq \emptyset$ then
10:      remove $j$ (if exists) from $CA$ when $|CA| > 1$
11:      remove $i$ (if exists) from $CO$ when $|CO| > 1$
12:      $k = \text{max}(CA)$, $l = \text{min}(CO)$
13:      $T \leftarrow T \cup (SP_{i,k}, SP_{i,j})$
14:   end if
15:   // CASE-2: aspect term after opinion term;
16:   $CO = \{k \mid i \leq k \leq j, (i,k) \in O\}$
17:   $CA = \{l \mid i \leq l \leq j, (l,j) \in A\}$
18:   if $CO \neq \emptyset$ and $CA \neq \emptyset$ then
19:      remove $j$ (if exists) from $CO$ when $|CO| > 1$
20:      remove $i$ (if exists) from $CA$ when $|CA| > 1$
21:      $k = \text{max}(CO)$, $l = \text{min}(CA)$
22:      $T \leftarrow T \cup (SP_{i,j}, SP_{i,k})$
23:   end if
24: end for
25: return $T$;

Sentiment snippet (line 7 to line 8). If both candidate aspects and opinions exist, we then separately retrieve the one with the maximum length, forming a valid triplet (line 9 to line 14). Considering that the sentiment snippet itself can be the aspect/opinion, we only view it as an aspect/opinion when there is no other counterpart candidates (line 10/11).

(2) The aspect term appears after the opinion term (line 16 to line 23), which is very similar to the previous case, except the candidate aspects share the end index and the candidate opinions share the start index with current sentiment snippet.

Finally, we can obtain all valid sentiment snippets after examining all sentiment snippets. Compared to previous decoding algorithms, our strategy has advantages in:

(1) Accuracy: The mutual span-level constraints guarantee the accuracy of the extracted sentiment triplets. For the sentiment snippet, which also indicates the pairing of the aspect and opinion, not only itself should have correct sentiment polarity, but also it should contain valid aspects and opinions that simultaneously sharing one of its boundaries. For the aspect/opinion term, not only themselves should be classified correctly, but also the boundaries of them should form a valid sentiment snippet.

(2) Efficiency: Contrary to the common practice that looks into every possible pair combination of candidate aspects and opinions to see whether they are valid, which contributes to heavy computational overhead, our method extracts the triplets based on the predicted sentiment snippets, the number of which is much less. More specifically, the time complexity is $O(n^2)$ with greedy inference and $O(n^4)$ with common practice, where $n$ is length of input text.

STAGE Model

Based on STAGE, we implement a simple BERT-based model. Shown in Figure 3, the overall model architecture contains an encoder that outputs contextual word representations for generating span representations and a classifier that predicts the tag of each span. The inference module is used to generate triplets from the tagging results.

Encoding  The key point of this process is to obtain span representations. Firstly, we adopt BERT (Devlin et al. 2019) plus an additional fully connected layer as our encoder to generate contextual word representations $h_i$ by,

$$\hat{h} = BERT([CLS] + \{w_i\} + [SEP]),$$

$$h_i = \sum_{k \in BERT(w_i)} \frac{1}{BertT(w_i)} \hat{h}_k + b_{fe},$$

where $BertT(w_i)$ returns the index set of $w_i$’s sub-words in BERT sequence, and $\cdot \parallel \cdot$ returns its length. $W_{fe}$ and $b_{fe}$ are the parameters of the fully connected layer, which is used to project the word representations from BERT to a lower dimension space. Then we form the span representations by,

$$sp_{i,j} = h_i \oplus h_j \oplus \sum_{k=i}^{j} h_k,$$

where $\oplus$ means vector concatenation. In this way, the model can utilize both the boundary information and the whole semantics of the span.
### Role distributions of spans (aspect-opinion-sentiment snippet)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#S</th>
<th>#T</th>
<th>#Pos.</th>
<th>#Neu.</th>
<th>#Neg.</th>
<th>A-N-N</th>
<th>N-O-N</th>
<th>N-N-S</th>
<th>A-O-N</th>
<th>A-N-S</th>
<th>N-O-S</th>
<th>A-O-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>906</td>
<td>1460</td>
<td>817</td>
<td>126</td>
<td>517</td>
<td>1279</td>
<td>1263</td>
<td>1454</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Dev</td>
<td>219</td>
<td>345</td>
<td>169</td>
<td>36</td>
<td>140</td>
<td>296</td>
<td>304</td>
<td>344</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Test</td>
<td>328</td>
<td>541</td>
<td>364</td>
<td>63</td>
<td>114</td>
<td>463</td>
<td>473</td>
<td>540</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Train</td>
<td>1266</td>
<td>2337</td>
<td>1691</td>
<td>166</td>
<td>480</td>
<td>2043</td>
<td>2072</td>
<td>2313</td>
<td>0</td>
<td>7</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Dev</td>
<td>310</td>
<td>577</td>
<td>404</td>
<td>54</td>
<td>119</td>
<td>497</td>
<td>498</td>
<td>569</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Test</td>
<td>492</td>
<td>994</td>
<td>773</td>
<td>66</td>
<td>115</td>
<td>840</td>
<td>850</td>
<td>984</td>
<td>1</td>
<td>7</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Statistics of datasets. #S, #T mean the total number of sentences and triplets. #Pos., #Neu. and #Neg. denote the number of positive, neutral, and negative sentiment triplets respectively. Note that we remove the duplicate ones. Role distributions of spans show the number of the spans playing different roles from three dimensions (aspect, opinion and sentiment snippet), where N means nothing in that dimension, A, O and S denote aspect, opinion and sentiment snippet, respectively.

### Training

The span representations are fed to a fully connected layer (i.e., classifier) with a softmax activation function, generating the probabilities over role tags:

\[
P(s_{pi,j}) = \text{softmax}(W_p s_{pi,j} + b_p),
\]

where \(W_p, b_p\) are parameters of the classifier.

The training loss is defined as the cross-entropy loss between golden labels and predicted label distributions of all spans:

\[
L(\theta) = -\sum_{i=1}^{n} \sum_{j=1}^{n} \text{loss}(P(s_{pi,j}), y_{i,j}),
\]

where \(y_{i,j}\) represent the gold label for \(S_{pi,j}\), \(\text{loss}\) is the standard cross-entropy loss and \(\theta\) represents model parameters.

### Implementation Details

The proposed STAGE\(^4\) model contains a BERT-base-uncased model\(^3\) and a fully connected layer as encoder, with hidden state dimension of 768 and 200 respectively. The dropout rate is 0.5. The model is trained in 100 epochs with a batch size of 16. AdamW optimizer (Loshchilov and Hutter 2017) is adopted with a learning rate of \(2 \times 10^{-5}\) for BERT fine-tuning and \(10^{-3}\) for other trainable parameters. For each dataset, we select the model with the best \(F_1\) scores on the development set and report the average results of five runs with different random seeds.

### Baselines

We compare our model with the following baselines:

- **Pipeline**: CMLA+ (Wang et al. 2017) considers the aspect-opinion interactions by an attention mechanism.
- **RI-NANTE+** (Dai and Song 2019) uses a BiLSTM-CRF and models the dependency relations, and Li-unified-R (Li et al. 2019) adopts a unified tagging method to jointly extract the aspect term and its sentiment, all modified by Peng et al. (2020) to perform ASTE task.
- **Peng-two-stage** (Peng et al. 2020) models the interaction between aspects and opinions.

- **End-to-End**: OTE-MTL (Xu et al. 2020) adopts a multi-task learning framework with a biaffine scorer.
- **JET-BERT** (Xu et al. 2020) introduces position information into the sequence tagging scheme.
- **GTS-BERT** (Wu et al. 2020) performs ASTE task by predicting the sentiment relations of word pairs.
- **Span-ASTE** (Xu, Chia, and Bing 2021) learns span interactions between aspects and opinions.
- **BMRC** (Chen et al. 2021) views ASTE as a multi-turn machine reading comprehension problem.
- **ASTE-RL** (Yu Bai Jian et al. 2021) adopts a hierarchical reinforcement learning framework by considering the aspect and opinion terms as arguments of the expressed sentiment.
- **EMC-GCN** (Chen et al. 2022) proposes to fully utilize the relations between words and diverse linguistic features.

---

\(^3\)ASTE-Data-V2 from https://github.com/xuuuluuu/SemEval-Triplet-data

\(^4\)Code is available at https://github.com/CCIIPLab/STAGE

**References**

- Dai and Song (2019)
- Peng et al. (2020)
- Wang et al. (2017)
- Loshchilov and Hutter (2017)
- Li et al. (2019)
- Xu et al. (2020)
- Chen et al. (2021)
- Yu Bai Jian et al. (2021)
- Xu, Chia, and Bing (2021)
- Chen et al. (2022)
Overall Performance
The overall performances are shown in Table 3. The observations are that: (1) Under the $F_1$ scores, our STAGE models (including two variants) all outperform state-of-the-art baselines by large margins in four datasets. And compared to previous best results, STAGE-3D significantly exceeds their $F_1$ scores by an average of 2.22%. And two variants (STAGE-2D and STAGE-1D) also exceed by an average of 2.17% and 2.10%, respectively. Considering our model architecture is quite simple, those significant improvements can well verify the effectiveness of our STAGE scheme. (2) Compared with STAGE-3D, the variant STAGE-2D achieves competitive results, even though it cannot deal with the cases that a span is the aspect and opinion simultaneously. It may result from the very limited training samples in those cases (as shown in Table 2), which causes STAGE-3D cannot learn such knowledge well. (3) STAGE-1D exceeds GTS-BERT (Wu et al. 2020) by an average 5.18% under the $F_1$ score. Considering that the two models have the same tag set (i.e., $\{A, O, POS, NEU, NEG, N\}$) but with different tagging schemes, it verifies that span-level tagging is superior to word-level tagging. Moreover, all our models outperform two state-of-the-art tagging-based baselines GTS-BERT and EMC-GCN significantly and consistently. (4) Although exceeding most baselines, our recall results are slightly lower than the best. But our precision results are better. One reason is that considering the mutual span-level constraints makes the process of retrieving and pairing much stricter during the inference.

Experiments on Subtasks
To further investigate the effectiveness of STAGE, we conduct experiments on three subtasks of ASTE, that is, aspect term extraction (ATE), opinion term extraction (OTE) and aspect-opinion pair extraction (AOPE). Note that Our method can directly address those subtasks without additional modifications. Specifically, we compare our method with two state-of-the-art tagging-based methods GTS-BERT (Wu et al. 2020) and EMC-GCN (Chen et al. 2022) by using their official released codes and configurations, for a fair comparison. The results are shown in Table 4 and the observations are that: (1) On ATE and OTE task, our models achieve the best results consistently on four datasets, which indicates that considering the complete semantics of each span via span tagging scheme, our method benefits the extraction of aspect and opinion terms. (2) On AOPE task, our models also exceed the compared models by a large margin. It proves that not only our span tagging scheme can better capture aspect/opinion terms, but the greedy inference that considers the span-level constraints can improve the pairing performance as well.

Table 3: Experimental results on the test set. The best $F_1$ results are in bold. The second best $F_1$ results are underlined. The “$t$” and “$s$” mean that the results are retrieved from Xu, Chia, and Bing (2021) and Chen et al. (2022), respectively. The “⊥” means we reproduce the results using the official released implementation code and configuration.

<table>
<thead>
<tr>
<th>Task</th>
<th>Model</th>
<th>14lap $P$</th>
<th>14res $P$</th>
<th>15res $F_1$</th>
<th>16res $F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATE</td>
<td>GTS-BERT</td>
<td>80.44</td>
<td>82.76</td>
<td>78.27</td>
<td>81.13</td>
</tr>
<tr>
<td></td>
<td>EMC-GCN</td>
<td>81.67</td>
<td>84.68</td>
<td>77.62</td>
<td>80.57</td>
</tr>
<tr>
<td></td>
<td>STAGE-1D</td>
<td>82.96</td>
<td>85.24</td>
<td>80.91</td>
<td>83.60</td>
</tr>
<tr>
<td></td>
<td>STAGE-2D</td>
<td>82.64</td>
<td>85.23</td>
<td>81.39</td>
<td>83.95</td>
</tr>
<tr>
<td></td>
<td>STAGE-3D</td>
<td>82.62</td>
<td>84.99</td>
<td>81.21</td>
<td>83.86</td>
</tr>
<tr>
<td>OTE</td>
<td>GTS-BERT</td>
<td>77.82</td>
<td>84.85</td>
<td>77.53</td>
<td>84.36</td>
</tr>
<tr>
<td></td>
<td>EMC-GCN</td>
<td>78.85</td>
<td>85.62</td>
<td>78.97</td>
<td>85.33</td>
</tr>
<tr>
<td></td>
<td>STAGE-1D</td>
<td>80.68</td>
<td>86.36</td>
<td>79.47</td>
<td>85.97</td>
</tr>
<tr>
<td></td>
<td>STAGE-2D</td>
<td>80.82</td>
<td>86.02</td>
<td>79.57</td>
<td>86.51</td>
</tr>
<tr>
<td></td>
<td>STAGE-3D</td>
<td>80.39</td>
<td>85.70</td>
<td>80.03</td>
<td>85.72</td>
</tr>
<tr>
<td>AOPE</td>
<td>GTS-BERT</td>
<td>66.66</td>
<td>74.63</td>
<td>67.52</td>
<td>74.20</td>
</tr>
<tr>
<td></td>
<td>EMC-GCN</td>
<td>67.94</td>
<td>76.33</td>
<td>67.26</td>
<td>74.15</td>
</tr>
<tr>
<td></td>
<td>STAGE-1D</td>
<td>69.98</td>
<td>77.42</td>
<td>71.35</td>
<td>79.41</td>
</tr>
<tr>
<td></td>
<td>STAGE-2D</td>
<td>69.70</td>
<td>77.87</td>
<td>70.60</td>
<td>79.98</td>
</tr>
</tbody>
</table>

Table 4: Test $F_1$ scores on ATE, OTE and AOPE tasks, with the best in bold and second best underlined.
### Analysis

#### Scheme Generalization

In this part, we extend our STAGE in entity and relation extraction task. Similar to the ASTE task, the span in the given sentence may have two roles: an entity term or a relation snippet that includes a pair of entities with a specific relation, forming its boundaries. So, we can define two role dimensions (entity and relation snippet) for entity and relation extraction task. But unlike the ASTE task, the “relation” here has “direction” (i.e., from head entity to tail entity). Thus, we adopt the span tagging scheme with the following tag set: \( \{ N, E_1, ..., E_{\left|E\right|} \} \rightarrow \{ N, R_1^{(H-T)}, R_2^{(T-H)}, ..., R_{\left|R\right|}^{(T-H)} \} \), where \( E = \{ E_1, E_2, ..., E_{\left|E\right|} \} \) denotes entity categories, \( R = \{ R_1, R_2, ..., R_{\left|R\right|} \} \) denotes relation types, \( N \) means nothing in that role dimension, \( (R-H) \) and \( (T-H) \) discriminate the relation directions as the former indicates head entity appears before the tail entity and the latter shows the opposite, and \( \cdot \) represents the set size.

We conduct experiments on SciERC dataset (Luan et al. 2018), a widely-used dataset including 500 scientific articles with annotations on the scientific terms and their relations. We follow the experiment settings of previous work (Zhong and Chen 2021) and utilize micro F1 score \( (F_1) \) as the evaluation metric. The experimental results in Table 6 show our model also achieves good performance, which verifies the good generalization ability of STAGE.

#### Conclusion

In this paper, we propose a novel approach, Span Tagging and Greedy inference (STAGE), to solve the ASTE task in an end-to-end manner. To the best of our knowledge, it is the first effort to explore a span-level tagging method and formulate the ASTE task as a multi-class span classification problem. Naturally performing on span-level, STAGE can overcome the limitations of previous tagging methods and effectively generate more accurate sentiment triplets by exploring span-level information and constraints. Specifically, STAGE consists of two components, namely span tagging scheme and greedy inference strategy. The former considers the diversity of span roles and tags the span based on three pre-defined role dimensions. The latter considers the mutual span-level constraints and retrieves the aspect/opinion term with the maximum length from the candidate sentiment snippet. Based on STAGE, we propose a simple but effective model, which outperforms the state-of-the-arts by a large margin on four widely-used datasets. Moreover, it can be easily generalized to other pair/triplet extraction tasks, indicating the superiority of the scheme STAGE.

---

**Table 5:** Test \( F_1 \) scores of ASTE under four settings, with the best in bold. Single. denotes triplets with single-word aspect and opinion. Multi. denotes triplets that tripllets with multi-word aspects or multi-word opinions. Multi. A./Multi. O. denotes triplets with multiple-word aspects/opinions.

**Table 6:** Test \( F_1 \) scores on SciERC dataset with the best in bold. \( \dagger \) means using cross-sentence context. Ent denotes the entity boundaries evaluation. Rel denotes the entity boundaries and relation being correct and Rel+ denotes the entity boundaries, entity types and the relation being correct.
Acknowledgments
This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110, Grant No.61772076, in part by CCF-AFSG Research Fund under Grant No.RF20210005, and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References


