A Multi-task Learning Framework for Opinion Triplet Extraction

Chen Zhang[1], Qiuchi Li[2], Dawei Song[1]*, Benyou Wang[2]
[1] Beijing Institute of Technology, Beijing, China.
[2] University of Padova, Padova, Italy.
{czhang,dwsong}@bit.edu.cn, {qiuchili,wang}@dei.unipd.it

Abstract

The state-of-the-art Aspect-based Sentiment Analysis (ABSA) approaches are mainly based on either detecting aspect terms and their corresponding sentiment polarities, or co-extracting aspect and opinion terms. However, the extraction of aspect-sentiment pairs lacks opinion terms as a reference, while co-extraction of aspect and opinion terms would not lead to meaningful pairs without determining their sentiment dependencies. To address the issue, we present a novel view of ABSA as an opinion triplet extraction task, and propose a multi-task learning framework to jointly extract aspect terms and opinion terms, and simultaneously parses sentiment dependencies between them with a biaffine scorer. At inference phase, the extraction of triplets is facilitated by a triplet decoding method based on the above outputs. We evaluate the proposed framework on four SemEval benchmarks for ASBA. The results demonstrate that our approach significantly outperforms a range of strong baselines and state-of-the-art approaches.[1]

1 Introduction

Aspect-based sentiment analysis (ABSA), also termed as Target-based Sentiment Analysis in some literature (Liu, 2012), is a fine-grained sentiment analysis task. It is usually formulated as detecting aspect terms and sentiments expressed in a sentence towards the aspects (Li et al., 2019; He et al., 2019; Luo et al., 2019; Hu et al., 2019). This type of formulation is referred to as *aspect-sentiment pair extraction*. Meanwhile, there exists another type of approach to ABSA, referred to as *aspect-opinion co-extraction*, which focuses on jointly deriving aspect terms (a.k.a. opinion targets) and opinion terms (a.k.a. opinion expressions) from sentences, yet without figuring out their sentiment dependencies (Wang et al., 2017; Li et al., 2018b). The compelling performances of both directions illustrate a strong dependency between aspect terms, opinion terms and the expressed sentiments.

Example sentence:	The atmosphere is attractive , but a little uncomfortable .
Aspect-sentiment pair extraction :	[(atmosphere, positive), (atmosphere, negative)]
Aspect-opinion co-extraction :	[atmosphere, attractive, uncomfortable]
Opinion triplet extraction :	[(atmosphere, attractive, positive), (atmosphere, uncomfortable, negative)]

Figure 1: Differences among aspect-sentiment pair extraction, aspect-opinion co-extraction, and opinion triplet extraction. Words in blue are aspect terms. Words in red are opinion terms. [ ] denotes a set of extracted patterns, and ( ) denotes an extracted pattern.

This motivates us to put forward a new perspective for ABSA as joint extraction of aspect terms, opinion terms and sentiment polarities,[2] in short *opinion triplet extraction*. An illustrative example of differences among aspect-sentiment pair extraction, aspect-opinion co-extraction, and opinion triplet extraction is given in Figure 1. Opinion triplet extraction can be viewed as an integration of aspect-sentiment pair extraction and aspect-opinion co-extraction, by taking into consideration their complementary nature. It brings in two-fold advantages: (1) the opinions can boost the expressive power of models and help better determine aspect-oriented sentiments; (2) the sentiment dependencies between aspects and opinions can bridge the gap of how sentiment decisions are made and further promote interpretability of models.

There is some prior research with a similar viewpoint. Peng et al. (2019) proposes to extract opin-

*Dawei Song is the corresponding author.

[1]Code and datasets for reproduction are available at https://github.com/GeneZC/OTE-MTL.

[2]For simplicity, these four concepts are hereafter referred to as aspect, opinion, sentiment, and triplet, respectively.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 819–828

ion tuples, i.e., (aspect-sentiment pair, opinion)s, by first jointly extracting aspect-sentiment pairs and opinions by two sequence taggers, in which sentiments are attached to aspects via unified tags,[3] and then pairing the extracted aspect-sentiments and opinions by an additional classifier. Despite of remarkable performance the approach has achieved, two issues need to be addressed.

The first issue arises from the prediction of aspects and sentiments with a set of unified tags thus degrading the sentiment dependency parsing process to a binary classification. As is discussed in prior studies on aspect-sentiment pair extraction (He et al., 2019; Luo et al., 2019; Hu et al., 2019), although the concerned framework with unified tagging scheme is theoretically elegant and mitigates the computational cost, it is insufficient to model the interaction between the aspects and sentiments (He et al., 2019; Luo et al., 2019).

Secondly, the coupled aspect-sentiment formalization disregards the importance of their interaction with opinions. Such interaction has been shown important to handle the overlapping circumstances where different triplet patterns share certain elements, in other triplet extraction-based tasks such as relation extraction (Fu et al., 2019). To show why triplet interaction modelling is crucial, we divide triplets into three categories, i.e., aspect overlapped, opinion overlapped, and normal ones. Examples of these three kinds of triplets are shown in Figure 2. We can observe that two triplets tend to have the same sentiment if they share the same aspect or opinion. Hence, modelling triplet interaction shall benefit the ASBA task, yet it can not be explored with the unified aspect-sentiment tags in which sentiments have been attached to aspects without considering the overlapping cases.

To circumvent the above issues, we propose a multi-task learning framework for opinion triplet extraction, namely OTE-MTL, to jointly detect aspects, opinions, and sentiment dependencies. On one hand, the aspects and opinions can be extracted with two independent heads in the multi-head architecture we propose. On the other hand, we decouple sentiment prediction from aspect extraction. Instead, we employ a sentiment dependency parser as the third head, to predict word-level sentiment

[3]An aspect tag set {B, I, O} and a sentiment tag set {NEU, NEG, POS} are unified into the aspect-sentiment tag set {B-NEU, I-NEU, B-NEG, I-NEG, B-POS, I-POS, O}. Here, B, I, and O indicate begin, inside, and outside of a span. And NEU, NEG, and POS are neutral, negative, and positive.

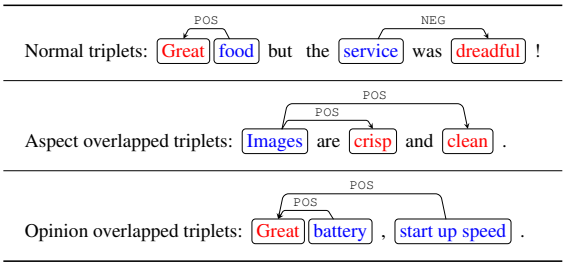

Figure 2: Categories of triplets. Spans in blue are aspects and spans in red are opinions. Arcs indicate sentiment dependencies and are always directed from an aspect to opinion.

dependencies, which will be utilized to further decode span-level[4] dependencies when incorporated with the detected aspects and opinions. In doing so, we expect to alleviate issues brought by the unified tagging scheme. Specifically, we exploit sequence tagging strategies (Lample et al., 2016) for extraction of aspects and opinions, whilst taking advantage of a biaffine scorer (Dozat and Manning, 2017) to obtain word-level sentiment dependencies. Additionally, since these task-heads are jointly trained, the learning objectives of aspect and opinion extraction could be considered as regularization applied on the sentiment dependency parser. In this way, the parser is learned with aspect- and opinion-aware constraints, therefore fulfilling the demand of triplet interaction modelling. Intuitively, if we are provided with a sentence containing two aspects but only one opinion (e.g., the third example in Figure 2), we can identify triplets with overlapped opinion thereby.

Extensive experiments are carried out on four SemEval benckmarking data collections for ABSA. Our framework are compared with a range of state-of-the-art approaches. The results demonstrate the effectiveness of our overall framework and individual components within it. A further case study shows that how our model better handles overlapping cases.

2 Proposed Framework

2.1 Problem Formulation

Given an input sentence $S = \{w_i\}_{i=1}^{|S|}$, our model aims to output a set of triplets $T = \{t_j\}_{j=1}^{|T|}$, where $|S|$, $|T|$ are the lengths of the sentence and the triplet set, respectively. A triplet t_j consists of three elements, i.e., $[m_j^{(ap)}, m_j^{(op)}, m_j^{(st)}]$,

[4]The aspects and opinions are usually spans over several words in the sentence

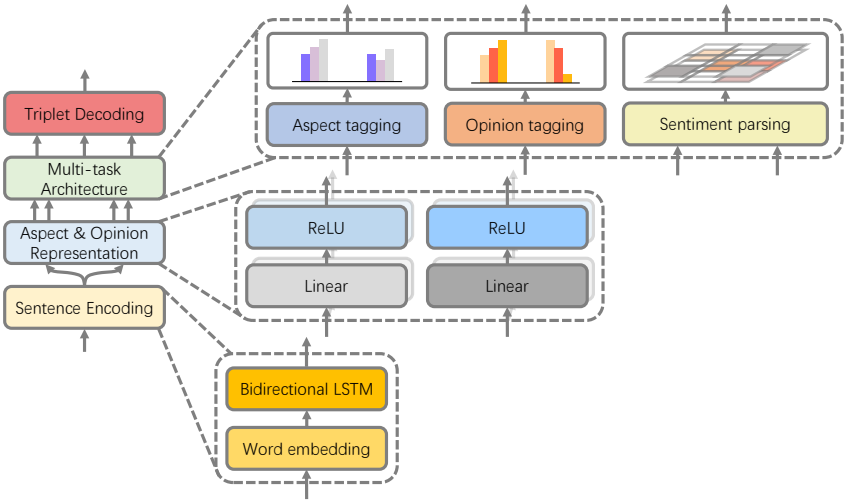

Figure 3: An overview of our proposed framework.

which separately stand for aspect span, opinion span, and sentiment. While the aspects and opinions are usually spans over several words in the sentence, we simplify the notation with the start position (denoted as sp) and end position (denoted as ep) of a span. Accordingly, $m_j^{(ap)}$ and $m_j^{(op)}$ can be represented as $(sp_j^{(ap)}, ep_j^{(ap)})$ and $(sp_j^{(op)}, ep_j^{(op)})$. Thus, the problem is formulated as finding a function $\mathcal{F}$ that accurately maps the sentence $S = \{w_i\}_{i=1}^{|S|}$ onto a triplet set $T = \{t_j \mid t_j = [(sp_j^{(ap)}, ep_j^{(ap)}), (sp_j^{(op)}, ep_j^{(op)}), m_j^{(st)}]\}_{j=1}^{|T|}$.

2.2 The OTE-MTL Framework

Our proposed OTE-MTL framework folds the triplet extraction process into two stages, i.e., prediction stage and decoding stage. An overview of our framework is presented in Figure 3. The prediction stage is parameterized by neural models and thus is trainable. It builds upon a sentence encoding module based on word embedding and a bidirectional LSTM structure, to learn an abstract representation of aspects and opinions. Underpinned by the abstract representation, there are three core components, accounting for three subgoals, i.e., aspect tagging, opinion tagging, and word-level sentiment dependency parsing. After the aspects, opinions and word-level dependencies have been detected, a decoding stage is then carried out to produce triplets based on heuristic rules.

2.3 Sentence Encoding

Context awareness is crucial for sentence encoding, i.e., encoding a sentence into a sequence of vectors. Hence, we adopt a bidirectional Long Short-term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997) as our sentence encoder, owing to the context modelling capability of LSTMs. In order to encode the input sentence, we first embed each word in a sentence to a low-dimensional vector space (Bengio et al., 2003) with pre-trained word embeddings[5]. With the embedded word representations $E = \{\mathbf{e}_i \mid \mathbf{e}_i \in \mathbb{R}^{d_e}\}_{i=1}^{|S|}$, the bidirectional LSTM is employed to attain contextualized representations of words $H = \{\mathbf{h}_i \mid \mathbf{h}_i \in \mathbb{R}^{2d_h}\}_{i=1}^{|S|}$ by the following operation:

$$\mathbf{h}_i = [\overrightarrow{\text{LSTM}}(\mathbf{e}_i) \oplus \overleftarrow{\text{LSTM}}(\mathbf{e}_i)] \quad (1)$$

where d_e and d_h denote the dimensionality of a word embedding and a hidden state from an unidirectional LSTM, while $\overrightarrow{\text{LSTM}}(\cdot)$ and $\overleftarrow{\text{LSTM}}(\cdot)$ stand for forward and backward LSTM, respectively. $\oplus$ means vector concatenation.

2.4 Aspect and Opinion Representation

We then extract the aspect- and opinion-specific features from the encoded hidden states, by applying dimension-reducing linear layers and nonlinear functions, rather than directly feeding the hidden states into the next components, for two reasons. First, the hidden states might contain superfluous information for follow-on computations, potentially causing a risk of overfitting. Second, such operations are expected to strip away irrelevant features for aspect tagging and opinion tagging. The computation process is formulated as

[5]In our experiments, GloVe vectors (Pennington et al., 2014) are used.

below:

$$\mathbf{r}_i^{(ap)} = g(\mathbf{W}_r^{(ap)}\mathbf{h}_i + \mathbf{b}_r^{(ap)}) \quad (2)$$

$$\mathbf{r}_i^{(op)} = g(\mathbf{W}_r^{(op)}\mathbf{h}_i + \mathbf{b}_r^{(op)}) \quad (3)$$

where $\mathbf{r}_i^{(ap)} \in \mathbb{R}^{d_r}$ and $\mathbf{r}_i^{(op)} \in \mathbb{R}^{d_r}$ are aspect and opinion representations, d_r is the dimensionality of the representation. $\mathbf{W}_r^{(ap)}$, $\mathbf{W}_r^{(op)} \in \mathbb{R}^{d_r \times 2d_h}$ and $\mathbf{b}_r^{(ap)}$, $\mathbf{b}_r^{(op)} \in \mathbb{R}^{d_r}$ are learnable weights and biases. Here, $g(\cdot)$ is a nonlinear function, which is ReLU, i.e., $\max(\cdot, 0)$, in our case.

Note that above representations are prepared for tagging. Likewise, we obtain another set of representations $\mathbf{r}_i^{(ap)\prime}, \mathbf{r}_i^{(op)\prime} \in \mathbb{R}^{d_r}$ for sentiment parsing, following the same procedure as Equation 2 and 3 but with different parameters.

2.5 Multi-task Architecture

The multi-task architecture includes two parts: aspect and opinion tagging, and word-level sentiment dependency parsing.

Aspect and Opinion Tagging. Following the {B, I, O} tagging scheme, we tag each word in the sentence with two taggers, i.e., one tagger for aspect, and the other for opinion. In particular, we receive two series of distributions over {B, I, O} tags $\mathbf{p}_i^{(ap)}$ and $\mathbf{p}_i^{(op)} \in \mathbb{R}^3$ through:

$$\mathbf{p}_i^{(ap)} = \text{softmax}(\mathbf{W}_t^{(ap)}\mathbf{r}_i^{(ap)} + \mathbf{b}_t^{(ap)}) \quad (4)$$

$$\mathbf{p}_i^{(op)} = \text{softmax}(\mathbf{W}_t^{(op)}\mathbf{r}_i^{(op)} + \mathbf{b}_t^{(op)}) \quad (5)$$

where $\mathbf{W}_t^{(ap)}$, $\mathbf{W}_t^{(op)} \in \mathbb{R}^{3\times d_r}$ and $\mathbf{b}_t^{(ap)}$, $\mathbf{b}_t^{(op)} \in \mathbb{R}^3$ are trainable parameters.

Accordingly, we can deduce the loss function, typically cross entropy with categorical distribution, for tagging as:

$$\begin{aligned}\mathcal{L}_{tag} = &-\frac{1}{|S|}\sum_i\sum_k \hat{\mathbf{p}}_{i,k}^{(ap)}\log(\mathbf{p}_{i,k}^{(ap)}) \\ &-\frac{1}{|S|}\sum_i\sum_k \hat{\mathbf{p}}_{i,k}^{(op)}\log(\mathbf{p}_{i,k}^{(op)})\end{aligned} \quad (6)$$

where $\hat{\mathbf{p}}_i^{(ap)}$ and $\hat{\mathbf{p}}_i^{(op)}$ respectively denote the ground truth aspect and opinion tag distributions of each word, and k is an enumerator over each item in a categorical distribution.

Word-level Sentiment Dependency Parsing. There are $|S|^2$ possible word pairs (including self-pairing cases) in each sentence and we intend to determine dependency type of every word pair. The set of dependency types is defined as {NEU, NEG, POS, NO-DEP}, so as to address all kinds of dependencies. Here, NO-DEP denotes no sentiment dependency. In addition, inspired by the table filling methods (Miwa and Sasaki, 2014; Bekoulis et al., 2018), sentiment dependencies are considered only for a pair of words that are exactly the last word of an aspect and the last word of an opinion in a triplet. Recall the example sentence "*Great battery, start up speed.*". For the triplet (*start up speed*, *great*, POS), the sentiment dependency is simplified to (*speed*, *great*, POS). As such, the learning redundancy for the parser is much reduced, while the span-level sentiment dependency is still available when it is combined with extracted aspect and opinion spans.

We utilize a biaffine scorer to capture the interaction of two words in each word pair, due to its proven expressive power in syntactic dependency parsing (Dozat and Manning, 2017). The score assignment to each word pair is as below:

$$\begin{aligned}\tilde{\mathbf{s}}_{i,j,k} &= [\mathbf{W}^{(k)}\mathbf{r}_i^{(ap)\prime} + \mathbf{b}^{(k)}]^\top \mathbf{r}_j^{(op)\prime} \\ &= [\mathbf{W}^{(k)}\mathbf{r}_i^{(ap)\prime}]^\top \mathbf{r}_j^{(op)\prime} + \mathbf{b}^{(k)\top}\mathbf{r}_j^{(op)\prime}\end{aligned} \quad (7)$$

where $\tilde{\mathbf{s}}_{i,j,k}$ stands for score of the k-th dependency type for a word pair (w_i, w_j). $\mathbf{W}^{(k)}$ and $\mathbf{b}^{(k)}$ are trainable weight and bias for producing the k-th score, respectively. Moreover, we use $\mathbf{s}_{i,j}$ to indicate a softmax-normalized vector of scores, which contains probabilities of all dependency types for the word pair (w_i, w_j):

$$\mathbf{s}_{i,j,k} = \text{softmax}(\tilde{\mathbf{s}}_{i,j,k}) \quad (8)$$

As observed from the factorization in Equation 7, conceptually the biaffine scorer can not only model the likelihood of w_i receiving w_j as a dependent of a specific type (the first term), but also include the prior probability of w_j being a dependent of such type (the second term). When it is implemented, the scorer is essentially an affine transform followed by matrix multiplication.

Thereafter, the loss function for word-level sentiment dependency parsing is a cross entropy function given below:

$$\mathcal{L}_{dep} = -\frac{1}{|S|^2}\sum_{(i,j)}\sum_k \hat{\mathbf{s}}_{i,j,k}\log(\mathbf{s}_{i,j,k}) \quad (9)$$

where $\hat{\mathbf{s}}_{i,j}$ is the ground-truth dependency distribution for each word pair (w_i, w_j).

Overall Learning Objective. Ultimately, we can conduct joint training of the multi-task learning framework with the following objective:

$$\min_{\theta} \mathcal{L} = \min_{\theta} \mathcal{L}_{tag} + \alpha \mathcal{L}_{dep} + \gamma ||\theta||_2 \quad (10)$$

where α is a trade-off term to balance the learning between tagging and sentiment dependency parsing. θ stands for trainable parameters. $||\theta||_2$ and γ are L_2 regularization of θ and a controlling term, respectively.

2.6 Triplet Decoding

Upon obtaining the extracted aspects, opinions, and word-level sentiment dependencies, we conduct a triplet decoding process using heuristic rules. Basically, we view the sentiment dependencies resulted from the biaffine scorer as pivots, and carry out a reverse-order traverse on tags generated by the aspect and opinion taggers.

For example, from word sequence "*Great battery , start up speed .*", we get aspect tags {O, B, O, B, I, I, O}, opinion tags {B, O, O, O, O, O, O}, and a word-level sentiment dependency, which is represented in index form, (6, 1, POS). The yielded sentiment dependency typically means that the last word of aspect is the 6-th word (*speed*), the last word of opinion is the 1-th word (*Great*), and they together form a positive sentiment. The traverse is conducted based on the aspect and opinion index (pivots) and the word sequence following stop-on-non-I criterion. And the final output should be [(4, 6), (1, 1), POS]. Details of the algorithm is shown in 1.

Algorithm 1 Decoding w/ stop-on-non-I criterion.

Input: aspect tags $\{g_i^{(ap)}\}_{i=1}^n$, opinion tags $\{g_i^{(op)}\}_{i=1}^n$, sentiment dependency (j, k, p).
Output: triplet t

$j' \leftarrow j$
while $g_{j'}^{(ap)}$ is I **do** ◁ stop on B and O.
  $j' \leftarrow j' - 1$
  if $j' \leq 0$ **then** ◁ or exceeding boundary.
    break
$k' \leftarrow k$
while $g_{k'}^{(op)}$ is I **do**
  $k' \leftarrow k' - 1$
  if $k' \leq 0$ **then**
    break
$t \leftarrow [(j', j), (k', k), p]$

3 Experimental Setup

3.1 Datasets and Evaluation Metrics

We conduct experiments on three datasets in the "restaurant" domain from SemEval 2014, 2015 and 2016 (Pontiki et al., 2014, 2015, 2016), and one dataset in the "laptop" domain from SemEval 2014. Hereafter, we will refer to them as REST14, REST15, REST16, and LAPTOP14 respectively. Since they are originally annotated with aspects and sentiments only, we additionally adopt annotations of opinion terms from Wang et al. (2017) and Peng et al. (2019). Each dataset is split to three subsets, namely, training set, validation set, and test set. The statistics of these datasets are shown in Table 1. It is worth noting that, in (Peng et al., 2019), the opinion overlapped triplets (in short OOTs) are removed from all four datasets in the preprocessing step. However, these cases are preserved in our setting. A key observation from the statistics is that there are large amounts of overlapping cases in the datasets, on average accounting for 24.2% of the total number of triplets across all four datasets. This phenomenon suggests the need and significance of triplet interaction modelling.

Moreover, we adopt precision, recall, and micro F1-measure as our evaluation metrics for triplet extraction. Only exactly matched triplets, i.e., with all of the aspect, opinion and sentiment matched against gold standards, are viewed as true positives during evaluation. All results are reported by averaging 10 runs with random initialization. Paired t-test is used to examine statistical significance of the results.

3.2 Implementation Details

In our experiments, the word embeddings are initialized with pretrained GloVe word vectors (Pennington et al., 2014). The dimensionalities of embeddings d_e, hidden states d_h, aspect and opinion representations d_r are set to 300, 300, 100, respectively. The trade-off term in learning objective, i.e., α, is set to be 1. The coefficient for L_2 regularization, i.e., γ, is 10^{-5}. Dropout is applied on embeddings to avoid overfitting and the drop rate is 0.5. The learning rate during training is 10^{-3} while the batch size is 32. All the parameters are initialized with uniform distribution and optimized with the Adam optimizer. Besides, we set a patience number 5, so that we could stop the learning process early if there is no further performance improvement on validation set.

Dataset		# sentence	# triplet	# sentence w/ overlap	# triplet w/ overlap
REST14	train	1300	2409	437	578
	val.	323	590	92	147
	test	496	1014	193	389
REST15	train	593	977	151	189
	val.	148	160	42	62
	test	318	479	68	71
REST16	train	842	1370	208	256
	val.	210	334	52	61
	test	320	507	77	120
LAPTOP14	train	920	1451	263	365
	val.	228	380	80	101
	test	339	552	103	140

Table 1: Statistics of datasets. Sentence w/ overlap means sentence containing overlapped triplets and triplet w/ overlap denotes triplet that overlaps with other triplets.

3.3 Baselines and Variants

To perform a systematic comparison, we introduce a variety of baselines, which can be classified into two groups, i.e., pipeline methods proposed in Peng et al. (2019) and joint methods we adapted from previous aspect-opinion co-extraction systems based on our framework **OTE-MTL**.

First, we list the baselines with a pipeline structure. (1) **Pipeline** (Peng et al., 2019) decomposes triplet extraction to two stages: stage one for predicting unified aspect-sentiment and opinion tags, while stage two for pairing the two results from stage one. We further include three models adjusted in accordance with Pipeline: (2) **Unified+** (Li et al., 2019) is a typical aspect-sentiment pair extraction system, in which the unified tagging scheme is used. (3) **RENANTE+** (Dai and Song, 2019) is originally an aspect-opinion co-extraction system in a weakly-supervised manner. (4) **CMLA+** (Wang et al., 2017) is an aspect-opinion co-extraction system modelling the interaction between the aspects and opinions. Additionally, we adapt two extra baseline models to the multi-task leaning, resulting in: (5) **CMLA-MTL** and (6) **HAST-MTL** (Li et al., 2018b), which are extended from existing state-of-the-art aspect-opinion co-extraction systems.

We also propose a list of variants of our proposed OTE-MTL framework to examine the efficacy of different components in it. (a) **OTE-MTL-Inter** feeds the prediction of aspects and opinions to the biaffine scorer by imposing tag embedding and concatenating tag embeddings to the input of the scorer. (b) **OTE-MTL-Concat** replaces the biaffine scorer with an activated linear layer applied on the concatenated vectors of aspect and opinion representations. (c) **OTE-MTL-Unified** uses unified aspect-sentiment tagging scheme and degrades the biaffine scorer to a binary pair classifier, which is similar to Pipeline but is jointly trained. (d) **OTE-MTL-Collapsed** combines the aspect and opinion tagging components into one single module via a collapsed tag set {`B-AP`, `I-AP`, `B-OP`, `B-OP`, `O`}, thus is forced to account for the constraint that aspects and opinions would never overlap.

4 Results and Analysis

4.1 Quantitative Evaluation

Comparison with Baselines. The results in comparison with baselines are shown in Table 2, both on datasets with and without OOTs for a fair comparison. Our propose model OTE-MTL consistently outperforms all state-of-the-art baselines on all datasets with and without OOTs. Thus, we conclude OTE-MTL is effective in dealing with opinion triplet extraction task.

We observe that the results of OTE-MTL on datasets without OOTs are generally better than those with OOTs except for LAPTOP14, implying that datasets without OOTs is comparably simpler and easier to achieve a good performance. Hence, we believe that overlapping cases bring challenges and can be partly addressed via triplet interaction modelling. Nevertheless, CMLA+ presents a worse performance in contrast to superior performance produced by CMLA-MTL. This fact suggests that, through decoupling aspect and sentiment predictions and puting them under the multi-task learning framework, the model can be enhanced and gain better results.

Comparison with Variants. The comparison with variants of OTE-MTL shown in Table 2 aims to verify the effectiveness of different components of OTE-MTL. As a whole, OTE-MTL surpasses all its variants. Specifically, OTE-MTL is slightly better than OTE-MTL-Inter, however, OTE-MTL exceeds other variants by large margins.

Rather than implicitly modelling the interaction between tagging and sentiment dependency parsing, OTE-MTL-Inter explicitly feeds emebddings of predicted tags to the biaffine scorer. It gets an inferior performance. We conjecture the reason lies in the latent error propagation when tags are par-

Model	REST14			REST15			REST16			LAPTOP14		
	pre.	rec.	f1.	pre.	rec.	f1.	pre.	rec.	f1.	pre.	rec.	f1.
RENANTE+$^{\dagger *}$	30.90	38.30	34.20	29.40	26.90	28.00	27.10	20.50	23.30	23.10	17.60	20.00
CMLA+$^{\dagger *}$	38.80	47.10	42.50	34.40	37.60	35.90	43.60	39.80	41.60	31.40	34.60	32.90
Unified+$^{\dagger *}$	43.83	62.38	51.43	43.34	50.73	46.69	38.19	53.47	44.51	42.25	42.78	42.47
Pipeline$^{\dagger *}$	42.29	64.07	50.90	40.97	54.68	46.79	46.76	62.97	53.62	40.40	47.24	43.50
OTE-MTL (ours)*	66.04	56.25	**60.62**‡	57.51	43.96	**49.76**‡	64.68	54.97	**59.36**‡	50.52	39.71	**44.31**‡
CMLA-MTL	43.24	44.95	43.97	35.87	39.85	37.55	44.22	46.43	45.01	33.61	36.11	34.68
HAST-MTL	58.97	46.75	52.04	41.48	37.58	39.32	52.32	48.56	49.92	47.70	25.74	33.24
OTE-MTL (ours)	64.54	55.57	**59.67**‡	54.18	45.20	**48.97**‡	58.16	54.02	**55.83**‡	48.17	42.43	**45.05**‡
OTE-MTL-Inter	66.24	54.38	59.61	49.32	46.12	47.33	57.71	53.06	55.17	47.66	41.85	44.43
OTE-MTL-Concat	48.79	48.28	48.46	46.88	42.61	44.53	52.55	48.03	50.09	46.81	38.46	42.14
OTE-MTL-Unified	51.19	44.65	47.64	40.32	34.38	37.01	48.52	40.30	43.85	37.42	34.17	35.54
OTE-MTL-Collapsed	45.38	36.26	40.19	32.55	29.52	30.68	37.86	33.06	35.19	32.56	27.23	29.60

Table 2: Quantitative evaluation results (%). Results of models with marker * are reported on datasets without OOTs. Results of models with marker † are directly cited from Peng et al. (2019). F1 measures in **bold** are the best performing numbers on each dataset. F1 measures with marker ‡ are significantly better than other numbers on each dataset with paired t-test ($p < 0.01$).

tially wrong, therefore hinting implicit modelling is a promising choice. The failure of OTE-MTL-Concat, which cannot model priors, supports the idea of leveraging biaffine scorer as word-level sentiment dependency parser. The result of OTE-MTL-Unified indicates that coupling aspect and sentiment extraction is suboptimal. Furthermore, we use OTE-MTL-Collapsed to account for non-overlap constraint of aspects and opinions, however, it obtains unexpectedly poor results. A possible explanation is that simultaneously collapsing aspect and opinion representations into one space may cause limited capacity for expressiveness.

4.2 Qualitative Evaluation

Case Study. To understand in what way our framework overwhelms the other unified tagging-based approaches, we perform a case study on three representative examples from test sets, as displayed in Table 3.

We notice that both OTE-MTL-Unified and OTE-MTL are working well for the first case which involves no overlapping. Nonetheless, OTE-MTL-Unified performs less well when faced with the second sample which contains aspect overlapped triplets and requires triplet interaction modelling. This case also shows conflicting opinions to an aspect (Tan et al., 2019), which is not covered by the training set but exists in real-world applications. It cannot be coped with by coupled aspect-sentiment tags since a tag should not have diverse sentiments. Thus decoupling sentiments from aspect tags is necessary. In the third example with long-range dependency, both aspect overlap and opinion overlap exist. For this case, OTE-MTL is not strong enough to make all correct predictions, but still seems to work better than OTE-MTL-Unified.

Error Analysis. To further find out the strengths and limitations of OTE-MTL, we conduct a detailed analysis of false positives (extracted by the system but not existing in ground truth) and false negatives (not extracted by the system but existing in ground truth) on REST14. For false positives, we categorize them into four classes: false aspect, false opinion, false sentiment, and other (mixed) case. For false negatives, we divide them according to categories of overlap (i.e., aspect overlapped, opinion overlapped, normal).

Figure 4 shows the analysis result. False positives are largely triggered by only one false element, especially, aspect or opinion, of an extracted triplet, motivating us to develop more robust span detection algorithms. In addition, the circumstance might also reflect that exact match is not an ideal metric when systems are evaluated, since minor discrepancy in a span may be harmless for opinion interpretation in practice, as we could observe in Table 3. Likewise, from Figure 4, we posit that overlapping cases are still non-trivial to solve given they have almost taken half of the false negatives.

5 Related Work

5.1 Aspect-based Sentiment Analysis

Our work falls in the broad scope of ABSA. As we have previously discussed, there are two types

Case	Ground truth	OTE-MTL-Unified	OTE-MTL
Great food but the service was dreadful !	[(food, Great, POS), (service, dreadful, NEG)]	[(food, Great, POS), (service, dreadful, NEG)]	[(food, Great, POS), (service, dreadful, NEG)]
The atmosphere is attractive , but a little uncomfortable .	[(atmosphere, attractive, POS), (atmosphere, uncomfortable, NEG)]	[(atmosphere, attractive, POS), (atmosphere, uncomfortable, POS$^{\times}$)]	[(atmosphere, attractive, POS), (atmosphere, uncomfortable, NEG)]
I am pleased with the fast log on , speedy WiFi connection and the long battery life .	[(log on, fast, POS), (WiFi connection, speedy, POS), (battery life, long, POS), (log on, pleased, POS), (WiFi connection, pleased, POS), (battery life, pleased, POS)]	[(log$^{\times}$, fast, POS), (WiFi connection, speedy, POS), (battery life, long, POS), (log$^{\times}$, pleased, POS), ()$^{\times}$, ()$^{\times}$]	[(log$^{\times}$, fast, POS), (WiFi connection, speedy, POS), (battery life, long, POS), (log$^{\times}$, pleased, POS), (WiFi$^{\times}$, pleased, POS), ()$^{\times}$]

Table 3: Case study. Marker $^{\times}$ indicates incorrect predictions.

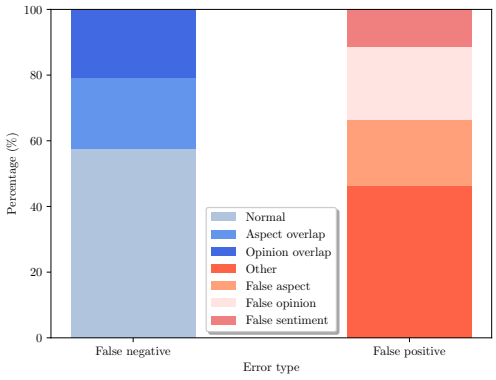

Figure 4: Components of false positives and false negatives.

of approaches in ABSA: aspect-sentiment pair extraction that concentrates on collaboratively detecting aspects and attached sentiment orientations (Li et al., 2019; He et al., 2019; Luo et al., 2019; Hu et al., 2019), and aspect-opinion co-extraction that tends to co-extract aspects and opinions (Wang et al., 2017; Li et al., 2018b). Alternatively, ABSA is also formulated as determining sentiment polarity of a given aspect in a sentence (Jiang et al., 2011; Dong et al., 2014; Tang et al., 2016a,b; Li et al., 2018a; Zhang et al., 2019), which is inflexible for practical use since aspects are not naturally accessible.

In this paper, we unify the aspect-sentiment pair extraction and aspect-opinion co-extraction, and formulate them as a triplet extraction problem. Our work is also aimed at addressing several issues in Peng et al. (2019), as discussed in the Introduction Section.

5.2 Triplet Extraction-based Task

Other than ABSA, a majority of triplet extraction-based tasks lies in the area of natural language processing. For example, Joint Entity and Relation Extraction (JERE) aims at detecting a pair of entity mentions in a sentence and predicting relation between the two. Approaches to JERE can be sorted into four streams: pipeline-based, table filling-based (Miwa and Sasaki, 2014; Bekoulis et al., 2018; Fu et al., 2019), tagging-based (Zheng et al., 2017), and encoder decoder-based (Zeng et al., 2018). Our work is motivated by table filling methods in Miwa and Sasaki (2014) and Bekoulis et al. (2018). We decompose triplet extraction to three subtasks, in which word-level sentiment dependency parsing can actually be viewed as a table filling problem, and solve them jointly in a multi-task learning framework.

6 Conclusions and Future Work

Our work put forwards an opinion triplet extraction perspective for aspect-based sentiment analysis. Existing works that are applicable to opinion triplet extraction have been shown insufficient, owing to the use of unified aspect-sentiment tagging scheme and ignorance of the interaction between elements in the triplet. Thus, we propose a multi-task learning framework to address the limitations by highlighting the uses of joint training, decoupled aspect and sentiment prediction, and regularization among correlated tasks during learning. Experimental results verify the effectiveness of our framework in comparison with a wide range of strong baselines. Comparison results with different variants of the proposed framework signify the necessity of the core components in the framework.

Based on the observations from a case study and error analysis, we plan to carry out further research in the following aspects: (1) more robust taggers for aspect and opinion extraction, (2) more flexible evaluation metric for triplet extraction, and (3) more mighty triplet interaction mechanism (e.g., encoder decoder structure).

Acknowledgments

This work is supported by The National Key Research and Development Program of China (grant No. 2018YFC0831704) and Natural Science Foundation of China (grant No. U1636203, U1736103).

References

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

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

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

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 49–54.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418.

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 Annual Meeting of the Association for Computational Linguistics*, pages 504–515.

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

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018a. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6714–6721.

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

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

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 Annual Meeting of the Association for Computational Linguistics*, pages 591–601.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1858–1869.

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)*, pages 1532–1543.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

Maria Pontiki, Dimitrios Galanis, Harris 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 2015)*, pages 486–495.

Maria Pontiki, Dimitrios 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 2014)*, pages 27–35.

Xingwei Tan, Yi Cai, and Changxi Zhu. 2019. Recognizing conflict opinions in aspect-level sentiment classification with dual attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3426–3431.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4560–4570.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236.