Target-oriented Opinion Words Extraction with Target-fused Neural Sequence Labeling

Zhifang Fan, Zhen Wu, Xin-Yu Dai*, Shujian Huang, Jiajun Chen
National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing, 210023, China
{fanzf,wuz}@nlp.nju.edu.cn;{daixinyu,huangsj,chenjj}@nju.edu.cn

Abstract

Opinion target extraction and opinion words extraction are two fundamental subtasks in Aspect Based Sentiment Analysis (ABSA). Recently, many methods have made progress on these two tasks. However, few works aim at extracting opinion targets and opinion words as pairs. In this paper, we propose a novel sequence labeling subtask for ABSA named TOWE (Target-oriented Opinion Words Extraction), which aims at extracting the corresponding opinion words for a given opinion target. A target-fused sequence labeling neural network model is designed to perform this task. The opinion target information is well encoded into context by an Inward-Outward LSTM. Then left and right contexts of the opinion target and the global context are combined to find the corresponding opinion words. We build four datasets for TOWE based on several popular ABSA benchmarks from laptop and restaurant reviews. The experimental results show that our proposed model outperforms the other compared methods significantly. We believe that our work may not only be helpful for downstream sentiment analysis task, but can also be used for pair-wise opinion summarization.

1 Introduction

Sentiment analysis, also known as opinion mining (Pang and Lee, 2007; Liu, 2012), has drawn increasing attention of researchers and industries in recent years. It can provide valuable information from user-generated reviews. However, sentiment analysis at sentence level or document level sometimes cannot provide more detailed information, thus a finer-grained task, Aspect-Based Sentiment Analysis (ABSA) (Pontiki et al., 2014), is proposed to identify the opinions of a specific target or aspect in reviews. ABSA consists of multiple subtasks including aspect category detection, opinion target extraction, aspect level sentiment classification etc. Opinion target extraction (OTE) and opinion words extraction (OWE) are two such fundamental subtasks. Opinion targets, sometimes called aspect terms, are the words or phrases in the sentence representing features or entities towards which users show attitude. Opinion words (or opinion terms) refer to those terms used to express attitude explicitly. For example, in the sentence "*The menu is limited but almost all of the dishes are excellent.*", the words "*menu*" and "*dishes*" are two opinion targets, and the words "*limited*" and "*excellent*" are opinion words. More examples can be found in the upper part of Figure 1.

Review:
"My friends and I were on vacation in NY and was referred to Chance by a friend. I found the food to be outstanding, particularly the salmon dish I had. I also ordered the Change Mojito, which was out of this world. My friends settled for rice dishes, but we came back the following day to try the dim sum, which was not outstanding, but good. We ate out in the back patio, which is worth it as it's cool and the music is hear well there. Overall, excellent restaurant!"

The list of extracted targets and opinion words as pairs:
- food : [outstanding]
- salmon dish : [outstanding]
- Change Mojito : [out of this world]
- dim sum : [not outstanding, good]
- back patio : [worth, cool]
- music : [hear well]
- restaurant : [excellent]

Figure 1: The upper part is a restaurant review and the lower part shows the pairs of extracted opinion targets (in red) and opinion words (in blue).

Recently, a great number of works based on neural networks have been done on these two subtasks (Liu et al., 2015; Poria et al., 2016; Xu et al., 2018). Furthermore, some works also integrate the two subtasks into a multi-task learning architecture to extract them jointly, which achieves great progress on both subtasks (Wang et al., 2016, 2017;

*Corresponding author.

Proceedings of NAACL-HLT 2019, pages 2509–2518
Minneapolis, Minnesota, June 2 - June 7, 2019.

Li and Lam, 2017). However, the extracted opinion targets and opinion words are not in pairs and the correspondence is not extracted. For instance, in the example sentence, ⟨*menu:limited*⟩ and ⟨*dishes:excellent*⟩ are two opinion pairs. Obviously, extracting them as pairs is significant for ABSA. Additionally, in Figure 1, the list of pairs extracted from the example review can be considered to be an extractive pair-wise opinion summarization.

Considering the significance of the pairs in reviews and promising results of targets extraction in previous works, in this paper, we propose a new subtask for ABSA named **TOWE** (Target-oriented Opinion Words Extraction). Given a review and a target in the review, the objective of TOWE is to extract the corresponding opinion words describing or evaluating the target from the review. Then, TOWE can form pairs of the given target and its corresponding opinion words.

Motivated by the success of neural networks in natural language processing, we design a powerful sequence labeling neural network model to perform TOWE. The task TOWE aims to extract the target-oriented opinion terms. In the same review, for different targets, the model needs to output different results. Therefore, a core challenge is the learning of target-specific context representations. We design a neural encoder to incorporate target information and generate the target-fused context. To be specific, we propose an Inward-Outward LSTM to pass target information to the left context and the right context of the target respectively. Then we combine the left, right and global context to encode the sentence and make sequence labeling. It is essential and reasonable to formulate TOWE as a sequence labeling task because some opinion terms include several words and one opinion target may correspond to multiple opinion terms. We try two different decoding strategies in the experiment.

Our main contributions are summarized as follows:

- We propose a sequence labeling subtask for ABSA: TOWE (Target-oriented Opinion Words Extraction), which can offer assistance and interpretability for downstream tasks in ABSA.
- We design a novel sequence labeling neural network model to perform TOWE. It can generate target-specific context representations for different targets in the same review.
- We build four datasets from different domains serving as a benchmark for future works. We conduct extensive experiments on these datasets, and the results show that our model could significantly exceed a variety of baselines.

We release the datasets and our source code at https://github.com/NJUNLP/TOWE

2 Related works

A lot of works have been carried out for Opinion Targets Extraction. Traditional methods can be categorized into unsupervised/semi-supervised methods (Hu and Liu, 2004; Zhuang et al., 2006; Qiu et al., 2011) and supervised methods (Jakob and Gurevych, 2010; Shu et al., 2017). Recently, deep learning methods have also made progress in this task. Liu et al. (2015) apply a recurrent neural network with pre-trained word emebddings to solve this task. Yin et al. (2016) exploit a CRF with dependency-paths enhanced word embeddings for aspect term extraction. Poria et al. (2016) use a deep convolutional neural network (CNN) and Xu et al. (2018) propose a CNN model with double embeddings.

Some works extract the targets and opinion words jointly as a co-extraction strategy. Qiu et al. (2011) propose double propagation to expand opinion targets and opinion words lists in a bootstrapping way. Liu et al.(2013) extract the targets and opinion words jointly with modeling the relation from a statistical word alignment model. This co-extraction strategy can also be adopted in neural networks with multi-task learning (Wang et al., 2016, 2017; Li and Lam, 2017). However, in all these works, the extracted targets and opinion words are separated.

In the literature, only a few works discussed opinion pairs. Hu and Liu (2004) use the distance information and recognize the nearest adjective of target as the opinion words. Zhuang et al. (2006) utilize lexicons and human-built word lists to extract the targets and opinion words in the corpus, and then identify valid feature-opinion pairs with syntactic rule templates based on dependency parsing trees. However, these two methods heavily depend on the external resources such as parsers or lexicons and the performance of these approaches relies on the quality of parsing result. By contrast, our model is a purely data-driven supervised learning method and does not need any external

linguistic knowledge, lexicons or handcrafted templates. Moreover, in these two methods, the process of detecting opinion words and the process of discovering correspondence is separated into two tasks, which suffers from error propagation. Our model for TOWE aims at detecting the corresponding opinion words in one step with sequence labeling.

3 Our Methods

3.1 Task Formulation

Given a sentence $s = \{w_1, w_2, \ldots, w_i, \ldots, w_n\}$ consisting of n words, and a opinion target t in the sentence, the task is to make sequence labelling on the sentence to extract the target-oriented opinion words. We use the BIO tagging scheme (Ramshaw and Marcus, 1995) on this task. For each word w_i in the sentence s, it should be tagged as $y_i \in \{B, I, O\}$ (B: Beginning, I: Inside, O: Others).

For example, for different opinion targets, the sentence "*Waiters are very friendly and the pasta is out of this world .*" is tagged in w_i/y_i style as follows:

1. *Waiters/O are/O very/O* [*friendly/B*] *and/O the/O pasta/O is/O out/O of/O this/O world/O ./O* (Given opinion target: ***waiter***, extract ***"friendly"*** as corresponding opinion word).

2. *Waiters/O are/O very/O friendly/O and/O the/O pasta/O is/O* [*out/B of/I this/I world/I*] *./O* (Given Opinion target: ***pasta***, extract ***"out of this world"*** as corresponding opinion words).

3.2 Framework

Figure 2 shows the framework of our methods, which follows an encoder-decoder architecture. We propose a target-fused encoder to incorporate the target information into context and learn target-specific context representations, then pass them to the decoder for sequence labeling. In the target-fused encoder, we first use an Inward-Outward LSTM to model the left context and right context of the target, then combine them with the global context. In the decoder, we can adopt two different decoding strategies. We present the details of each component in the following sections.

3.3 Target-Fused Encoder

We first generate the input vectors for each word by using a embedding lookup table $\mathbb{L} \in \mathbb{R}^{d \times |V|}$, where d is the embedding dimension and $|V|$ is the vocabulary size. The embedding lookup table will map $s = \{w_1, w_2, \ldots, w_t, \ldots, w_n\}$ to a sequence of vectors $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_i, \ldots, \mathbf{e}_n\}$ as words representations where $\mathbf{e}_i \in \mathbb{R}^d$.

Typically, neural sequence labeling models use recurrent neural networks, such as LSTM (Hochreiter and Schmidhuber, 1997) or BiLSTM, to model the sentence. However, merely using BiLSTM to model the whole sentence is totally target-independent. For the different target terms in the same sentence, BiLSTM outputs equal representation and cannot generate target-specific results.

As mentioned before, the core challenge of TOWE is the learning of target-specific context representations. It is evident that different targets have different positions in the sentence and thus different contexts. So, we first split the sentence into three segments: left context $\{w_1, w_2, \cdots, w_l\}$, target term $\{w_{l+1}, \cdots, w_{r-1}\}$ and right context $\{w_r, \cdots, w_n\}$ and left and right contexts are target-specific.We use a left LSTM to model the left context plus target and a right LSTM to model the target plus right context respectively. In this way the target-specific contexts could generate target-specific context representations. However, the direction of the two LSTMs is a crucial problem.

3.3.1 Inward-LSTM

We can use a simple strategy called Inward-LSTM, which follows the design of TD-LSTM (Tang et al., 2016). As Figure 2 shows, Inward-LSTM runs the two LSTMs from the two ends of the sentence to the middle target respectively. It runs the left LSTM from the first word to opinion target as a forward-LSTM and a right LSTM from the last word to the opinion target as a backward-LSTM, so we call it as Inward. This is a process of passing the context to target. We obtain left context representations $\mathbf{H}^{\mathrm{L}}$ and right context representations $\mathbf{H}^{\mathrm{R}}$ as follows:

$$\mathbf{h}_i^{\mathrm{L}} = \overrightarrow{\mathrm{LSTM}}(\mathbf{h}_{i-1}^{\mathrm{L}}, \mathbf{e}_i), \forall i \in [1, \cdots, r-1]\,, \quad (1)$$

$$\mathbf{h}_i^{\mathrm{R}} = \overleftarrow{\mathrm{LSTM}}(\mathbf{h}_{i+1}^{\mathrm{R}}, \mathbf{e}_i), \forall i \in [l+1, \cdots, n]\,. \quad (2)$$

It is obvious that the words of opinion target $\{w_{l+1}, \cdots, w_{r-1}\}$ are represented twice in the left LSTM and right LSTM. We simply average the two representations for the same word to get the representation of target words:

$$\mathbf{h}_i^{\mathrm{LR}} = \frac{(\mathbf{h}_i^{\mathrm{L}} + \mathbf{h}_i^{\mathrm{R}})}{2}, \forall i \in [l+1, \cdots, r-1]\,. \quad (3)$$

Then the context representation is: $\mathbf{H}^{\mathrm{I}} = \{\mathbf{h}_1^{\mathrm{L}}, \cdots, \mathbf{h}_l^{\mathrm{L}}, \mathbf{h}_{l+1}^{\mathrm{LR}}, \cdots, \mathbf{h}_{r-1}^{\mathrm{LR}}, \mathbf{h}_r^{\mathrm{R}}, \cdots, \mathbf{h}_n^{\mathrm{R}}\}$.

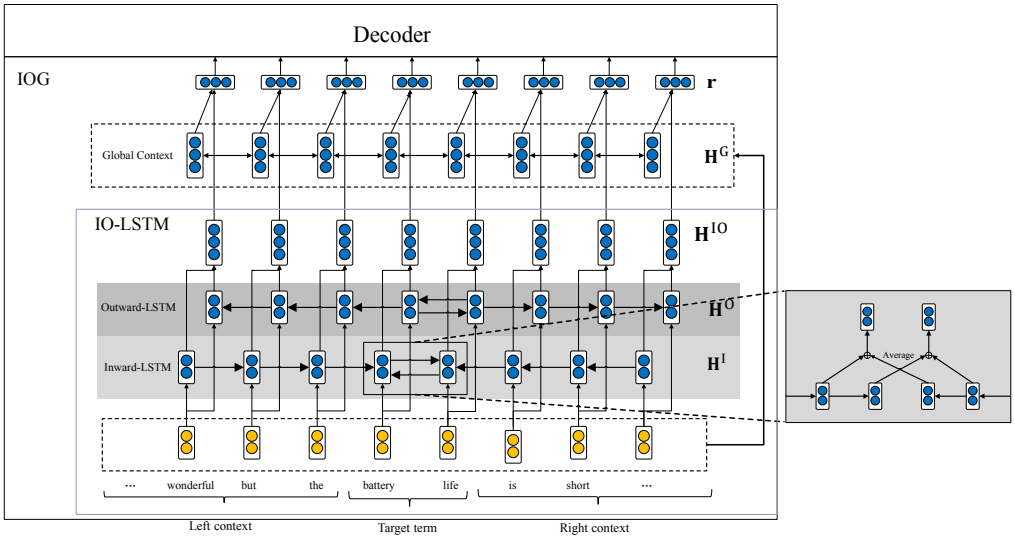

Figure 2: The framework of our method. The subfigure on the right side is an illustration for averaging the hidden states of the Opinion Target.

3.3.2 Outward-LSTM

Although passing contexts to the target in Inward-LSTM is a good strategy for encoding whole sentence representation, only using this strategy is not enough for TOWE because the target information is not passed to the left and right context. For example, in the sentence *"I found the food to be outstanding."*, the opinion target is *"food"*, the Inward-LSTM will first model *"outstanding"* and then model *"food"*. The representation of *"outstanding"* does not contain the information of *"food"*.

To solve this problem, we design a novel strategy specifically for TOWE, i.e., Outward-LSTM. The idea of the Outward-LSTM is to pass the target to the context, which we believe is a better choice. As Figure 2 shows, the Outward-LSTM starts two LSTMs from the target in the middle and run towards the both ends of the sentence, which means the left LSTM is a backward LSTM and the right LSTM is a forward LSTM. We average the duplicate target hidden states and get the target-fused context representations $\mathbf{H}^{\mathrm{O}} = \{\mathbf{h}_1^{\mathrm{L}}, \cdots, \mathbf{h}_l^{\mathrm{L}}, \mathbf{h}_{l+1}^{\mathrm{LR}}, \cdots, \mathbf{h}_{r-1}^{\mathrm{LR}}, \mathbf{h}_r^{\mathrm{R}}, \cdots, \mathbf{h}_n^{\mathrm{R}}\}$:

$$\mathbf{h}_i^{\mathrm{L}} = \overleftarrow{\mathrm{LSTM}}(\mathbf{h}_{i+1}^{\mathrm{L}}, \mathbf{e}_i), \forall i \in [1, \cdots, r-1], \quad (4)$$

$$\mathbf{h}_i^{\mathrm{R}} = \overrightarrow{\mathrm{LSTM}}(\mathbf{h}_{i-1}^{\mathrm{R}}, \mathbf{e}_i), \forall i \in [l+1, \cdots, n], \quad (5)$$

$$\mathbf{h}_i^{\mathrm{LR}} = \frac{(\mathbf{h}_i^{\mathrm{L}} + \mathbf{h}_i^{\mathrm{R}})}{2}, \forall i \in [l+1, \cdots, r-1]. \quad (6)$$

This concise and reasonable strategy can solve the problems remaining in the Inward-LSTM. As we start the LSTM from the target, the target's information is fused into each word in the sentence. Also, the Outward-LSTM ensures that for different targets each word has different representations. Take the sentence *"Its camera is wonderful but the battery life is short !"* as an example. For target *"camera"* or *"battery life"*, the target-fused representations for *"short"* are different and can generate target-specific results.

3.3.3 IO-LSTM

We can combine the both strategy and adopt an Inward-Outward LSTM (IO-LSTM). IO-LSTM concatenates the outputs of Outward-LSTM and Inward-LSTM. The output of Outward-LSTM is crucial for incorporating target information into context, while the Inwards-LSTM is included so they can complement each other and act as a Target-specific Bidirectional LSTM. The target-fused context representations are denoted as $\mathbf{H}^{\mathrm{IO}}$:

$$\mathbf{h}_i^{\mathrm{IO}} = \left[\mathbf{h}_i^{\mathrm{I}}; \mathbf{h}_i^{\mathrm{O}}\right]. \quad (7)$$

3.3.4 IOG: IO-LSTM + Global context

To extract the target-oriented opinion words, only considering the context of each side in isolation is not enough. The left context and right context in the IO-LSTM are separated, and the left LSTM and right LSTM only share the opinion target. It is important to understand the global meaning of the whole sentence while detecting the opinion words on the left and right context. So we introduce the

global context to further improve the IO-LSTM. We use a BiLSTM to model the whole sentence embeddings $\mathbf{e} = \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_i, \ldots, \mathbf{e}_n\}$ and obtain global contextualized representation $\mathbf{H}^{\mathrm{G}}$ as follows:

$$\mathbf{h}_i^{\mathrm{G}} = \left[\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i\right], \tag{8}$$

$$\overrightarrow{\mathbf{h}}_i = \mathrm{LSTM}(\overrightarrow{\mathbf{h}}_{i-1}, \mathbf{e}_i), \tag{9}$$

$$\overleftarrow{\mathbf{h}}_i = \mathrm{LSTM}(\overleftarrow{\mathbf{h}}_{i+1}, \mathbf{e}_i). \tag{10}$$

Then we combine left-right contexts from IO-LSTM and global context, as shown in Figure 2. This enables us to obtain the final target-specific contextualized representation $\mathbf{r}$ for each word:

$$\mathbf{r}_i = \left[\mathbf{h}_i^{\mathrm{IO}}; \mathbf{h}_i^{\mathrm{G}}\right]. \tag{11}$$

The final representation $\mathbf{r}$ is fused with both target information and global context information, which can be passed to the decoder for sequence labeling.

3.4 Decoder and Training

Given a sequential representation $\mathbf{r}$, we can use $\mathbf{r}$ to compute $p(\mathbf{y}|\mathbf{r})$ where $\mathbf{y} = \{y_1, \cdots, y_n\}$ are BIO-label sequence for the sentence and $y_i \in \{\mathrm{B}, \mathrm{I}, \mathrm{O}\}$. Two different decoding policies can be adopted.

3.4.1 Greedy decoding

The first is greedy decoding, formulated as a three-class classification problem at each position independently. We use softmax to compute the probability:

$$p(y_i|\mathbf{r}_i) = \mathrm{softmax}\left(\mathbf{W}_s\mathbf{r}_i + \mathbf{b}_s\right). \tag{12}$$

Greedy decoding just simply selects the tag with highest point-wise probability. It does not consider the dependencies between tags but runs faster. We use the negative log likelihood (NLL) as the loss for one sentence:

$$L(s) = -\sum_{i=1}^{n}\sum_{k=1}^{3}\mathbb{I}(y_i = k)\log p(y_i = k|w_i). \tag{13}$$

3.4.2 CRF

The second decoding method is to use Conditional Random Field (CRF) (Lafferty et al., 2001). CRF considers the correlations between tags in neighborhoods and score the whole sequence of tags. Specifically, we use a linear-chain CRF and score the tag sequence as conditional probability:

$$p(\mathbf{y}|\mathbf{r}) = \frac{\exp(s(\mathbf{r}, \mathbf{y}))}{\sum_{y' \in Y}\exp(s(\mathbf{r}, \mathbf{y}'))}, \tag{14}$$

where Y is the set of all possible tag sequences and $s(\mathbf{r}, \mathbf{y})) = \sum_i^n (\mathbf{A}_{y_{i-1}, y_i} + \mathbf{P}_{i, y_i})$ is the score function. $\mathbf{A}_{y_{i-1}, y_i}$ measures the transition score from y_{i-1} to y_i and $\mathbf{P}_i = \mathbf{W}_s\mathbf{r}_i + \mathbf{b}_s$. So we use negative log likehood as the loss of the sentence:

$$L(s) = -\log p(\mathbf{y}|\mathbf{r}). \tag{15}$$

When given a new sentence for decoding, we will output the tag sequence that maximizes the conditional probability with Viterbi algorithm.

Finally, we minimize the loss for training:

$$J(\theta) = \sum_{s}^{|D|} L(s). \tag{16}$$

4 Experiments

4.1 Datasets

We build the datasets based on the SemEval challenge 2014 Task4, SemEval Challenge 2015 task 12 and SemEval Challenge 2016 task 5 (Pontiki et al., 2014, 2015, 2016). The SemEval challenge provides several datasets from restaurant and laptop domain. These datasets are very popular benchmarks for many ABSA subtasks, including Aspect category detection, Opinion Target Extraction, Opinion Words Extraction and Target-Dependent Sentiment Analysis (TDSA).

In the original datasets of SemEval challenge, the opinion targets (aspect terms) are annotated, but the opinion words and the correspondence with targets are not provided. So we annotate the corresponding opinion words for the annotated targets. Every sentence is annotated by two people, and the conflicts will be checked. Each instance of the datasets consists of a sentence, the position of the target and the positions of the corresponding opinion words. Note that we only keep the sentences that contain pairs of target and opinion words. The sentences without targets or with implicit opinion expressions are not included.

Finally, we generate four datasets: **14res** and **14lap** from SemEval 2014, **15res** from SemEval 2015 and **16res** from SemEval 2016. **14res**, **15res**, and **16res** contain reviews from restaurant domain. The sentences in **14lap** come from laptop domain. The statistics of the four datasets is shown in Table 1.

4.2 Settings

In our experiments, we initialize word embedding vectors with 300-dimension GloVe vectors which

Dataset		#sentences	#targets
14res	Training	1627	2643
	Testing	500	865
14lap	Training	1158	1634
	Testing	343	482
15res	Training	754	1076
	Testing	325	436
16res	Training	1079	1512
	Testing	329	457

Table 1: Statistics of datasets. The number of targets is identical to the number of pairs and instances

are pre-trained on unlabeled data of 840 billion tokens (Pennington et al., 2014). The word embeddings are fixed and not fine-tuned during the training stage. The dimension of hidden states in all the LSTM cell is set as 200. Adam (Kingma and Ba, 2015) is chosen as the optimization method with the default setting in the original paper. We randomly split 20% of the train set as dev set for tunning the hyperparameters and early stopping. Then we test the models on testing sets and the average result of five runs is reported.

4.3 Evaluation Metrics

Precision, recall and F1 score are used as the metrics to measure the performance of models. An extracted opinion words span is regarded as a correct prediction when the starting and ending offset of the predicted span are both identical to those of a golden opinion words span. We compute Precision, Recall and F1 with the span as the unit.

4.4 Compared Methods

Since we are the first to study this sequence labeling task, there is no available sequence labeling model in the literature to be compared. Although there are a number of complicated models in TDSA, the task is different. Those TDSA models focus on sentence-level representations for sentiment classification, while for TOWE the representation on token-level representations is more crucial. Simply transferring the TDSA models for TOWE is not suitable.

Except for two rule-based methods, we can only design and implement the baselines for TOWE ourselves. Our final model is the IOG encoder with a greedy decoding strategy. We compare it with the following baselines:

- **Distance-rule**: Hu and Liu(2004) use the distance and POS tags to determine the opinion words. Following this idea, we first use the nltk toolkit to make part-of-speech tagging on each word and select the nearest adjective from the target as the corresponding opinion word.

- **Dependecy-rule**: We adopt the strategies proposed in (Zhuang et al., 2006) which uses dependency-tree based templates to identify opinion pairs. The POS tag of opinion targets and opinion words and the dependency path between them in the training set are recorded as rule templates.[1] The high-frequency dependency templates are used for detecting the related opinion words in the testing set.

- **LSTM/BiLSTM**: This method is an LSTM/BiLSTM network built on top of word embeddings proposed by (Liu et al., 2015). We pass the whole sentence into the LSTM/BiLSTM and each hidden state is fed to a softmax layer for three-class classification, which works as sentence-level opinion words extraction.

- **Pipeline**: This method combines BiLSTM and Distance-rule method in a pipelined way. We first train a sentence-level opinion words extraction model with BiLSTM and extract all the opinion words in the test sentences; then we select the closest extracted opinion words of the target as the result.

- **Target-Concatenated BiLSTM (TC-BiLSTM)**: This method incorporates the target information into sentence by concatenation. A target vector is obtained by the average pooling of target word embeddings. The word representation at each position is the concatenation of word embedding and target vector, which is then fed into a BiLSTM for sequence labeling.

4.5 Results and Discussion

The main results can be found in Table 2. Note that all the neural models in Table 2 adopt greedy decoding. The performance of Distance-rule method is not satisfactory and the worst among all the methods; its recall score is especially low. IOG obtains an F1 score with a greater-than 30% improvement over the Distance-rule method. Dependency-rule method obtains a general improvement than Distance-rule, but it was still lower than the below sequence-labeling based methods. This reveals the

[1] We use the parsers in spaCy: https://spacy.io

Models	14res			14lap			15res			16res		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Distance-rule	58.39	43.59	49.92	50.13	33.86	40.42	54.12	39.96	45.97	61.90	44.57	51.83
Dependency-rule	64.57	52.72	58.04	45.09	31.57	37.14	65.49	48.88	55.98	76.03	56.19	64.62
Pipeline	77.72	62.33	69.18	72.58	56.97	63.83	74.75	60.65	66.97	81.46	67.81	74.01
LSTM	52.64	65.47	58.34	55.71	57.53	56.52	57.27	60.69	58.93	62.46	68.72	65.33
BiLSTM	58.34	61.73	59.95	64.52	61.45	62.71	60.46	63.65	62.00	68.68	70.51	69.57
TC-BiLSTM	67.65	67.67	67.61	62.45	60.14	61.21	66.06	60.16	62.94	73.46	72.88	73.10
IOG	**82.85**	**77.38**	**80.02**	**73.24**	**69.63**	**71.35**	**76.06**	**70.71**	**73.25**	**85.25**	**78.51**	**81.69**

Table 2: Main Results in terms of Precsion, Recall and F1-score. Best results are in bold. IOG outperms all the baselines significantly ($p < 0.01$).

Models	14res			14lap			15res			16res		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Inward-LSTM	53.64	70.06	60.67	55.49	54.96	55.18	53.84	68.60	60.32	66.46	73.83	69.92
Outward-LSTM	81.08	73.65	77.17	70.34	64.90	67.48	74.18	65.60	69.62	81.12	77.68	79.35
IO-LSTM	82.09	75.44	78.62	72.90	66.49	69.51	74.71	67.22	70.74	83.31	78.67	80.91
IOG	82.85	77.38	80.02	73.24	69.63	71.35	76.06	70.71	73.25	85.25	78.51	81.69
IOG + CRF	82.97	77.73	80.24	74.19	68.96	71.39	75.50	71.68	73.51	84.41	79.43	81.84

Table 3: Comparisions for different model design in terms of Precsion, Recall and F1-score.

lack of robustness in rule-based approaches. The error propagation from syntactic parsers is also a reason for poor performance.

The Pipeline model performs much better than rule-based methods, obtaining an especially high precision, showing that machine-learning methods can obtain better opinion words extraction. However, pipeline model is still not ideal, and the F1-score is approximately 10% lower than our proposed model in several datasets. This reflects that the distance information is not sufficient for detecting target-oriented opinion words while IOG could better handle long distance dependency problem. Also, this strategy cannot solve the cases where one target corresponds to more than one opinion term. It also suffers from error propagation.

LSTM and BiLSTM are both target-independent leading to low precision, and their performance is even worse than pipeline method. IOG outperforms BiLSTM by about 15% averagely, which indicates the target information should be included.

TC-BiLSTM includes the target information by concatenation and obtains better general performance than LSTM and BiLSTM. However, TC-BiLSTM is still over 10% lower than IOG and is slightly inferior to Pipeline, showing that the concatenation is not a good way to incorporate the target information for TOWE. We believe that the problem is that the concatenated target may interfere with the other targets in the same sentence.

IOG achieves the best performance on all datasets from different domains compared to both the rule-based methods and neural models. We can conclude that IOG can learn target-specific representations more effectively and can better capture the correspondence between targets and opinion words.

4.6 Model Analysis

To compare the different design of our model and provide more compared models, we also report the results of the variants of our models in Table 3.

- **Inward-LSTM**: $\mathbf{H}^{\mathrm{I}}$ computed from (1), (2), (3) are fed to the greedy decoder for sequence labeling.
- **Outward-LSTM**: $\mathbf{H}^{\mathrm{O}}$ computed from (4), (5), (6) are fed for greedy decoding
- **IO-LSTM**: Combining Inwards-LSTM and Outwards-LSTM, $\mathbf{H}^{\mathrm{IO}}$ is obtained by concatenation of $\mathbf{H}^{\mathrm{I}}$ and $\mathbf{H}^{\mathrm{O}}$ in (7), which is then used for greedy decoding.
- **IOG+CRF**: Passing the representations $\mathbf{r}$ in IOG to a CRF decoder.

The performance of Inward-LSTM is inferior, similar to the target-independent BiLSTM. This demonstrates that only passing the context to target is similar to not considering the target information owing to the problems we discussed before.

The F1-score of Outward-LSTM exceeds that of the Inward-LSTM by more than 10%. This shows

Sentence	Distance-rule	Dependency rule	Pipeline	BiLSTM	TC-BiLSTM	IOG
The bread is top notch as well .	*top* ✗	NULL ✗	*top notch* ✓	*top notch* ✓	NULL ✗	*top notch* ✓
BEST spicy tuna roll, great asian salad .	*asian* ✗	*"great", "asian"* ✗	*great* ✓	*asian* ✗	*"BEST", "great"* ✗	*great* ✓
I love the drinks , esp lychee martini , and the food is also VERY good .	*lyche* ✗	*love* ✓	*love* ✓	*"love","good"* ✗	*"love","good"* ✗	*love* ✓
Food was decent , but not great .	*decent* ✗	*"decent", "great"* ✗	*decent* ✗	*"decent", "not great"* ✓	*decent* ✗	*"decent", "not great"* ✓
The food was excellent - authentic Italian cuisine made absolutely fresh .	*excellent* ✓	*NULL* ✗	*excellent* ✓	*excellent* ✓	*"excellent", "fresh"* ✗	*excellent* ✓
The food was excellent - authentic Italian cuisine made absolutely fresh .	*Italian* ✗	*"authentic", "Italian "* ✗	*authentic* ✗	*excellent* ✗	*"authentic", "excellent ", "fresh"* ✗	*"authentic", "fresh"* ✓

Table 4: Examples for the extracted result, the target terms are in red and the golden corresponding opinion words are in blue.

that passing target into context is a better choice and learning the target-specific word representations is crucial. In fact, Outward-LSTM has already outperformed all the previous baselines, which indicates that this is a really good design for TOWE. IO-LSTM which combine the Inward and Outward is slightly better than Outward-LSTM, showing that Inward-LSTM can still provide supplementary information for Outward-LSTM. Through combining global context with IO-LSTM as IOG model, we roughly obtain a further 1% improvement.

We also test our model with a linear Conditional-Random-Field as the decoder. CRF considers the label dependencies. It can be observed that IOG with CRF obtains a slight improvement.

4.7 Case Study

To demonstrate the effectiveness of our model, we pick some examples in the test dataset in **14res** and show the extracted results of different models.

In the first sentence, since the Distance-rule cannot extract phrases, the extraction it makes is incorrect. In addition, merely selecting the nearest adjective using the Distance-rule approach does not enable coverage in all cases, as shown in the second and third sentence (e.g., the *"asian"* and *"lyche"*). Dependency-rule in some cases fails to extract any word owing to the error of parser and no template to match. Pipeline method has the problem that it cannot handle the cases that one target corresponds to multiple opinion terms (e.g., *"not great"* is not extracted in the fourth sentence). The drawback of BiLSTM is that it does not include target information, so it extracts both *"love"* and *"good"* in the third sentence while only *"love"* is the corresponding opinion word for *"drinks"*. Although TC-BiLSTM is a target-specific model, it tends to extract irrelevant opinion words because of the interference from concatenation. In the last two rows of Table 4, we show the same sentence with two different targets and only IOG does not make mistakes for both targets. IOG outputs the correct results for all the sentences in the table.

5 Conclusion and Future Works

In this paper, we propose a novel subtask for aspect-based sentiment analysis: Target-oriented Opinion Words Extraction (TOWE) which aims at extracting the corresponding opinion words for a given opinion target. We design a novel neural model IOG to solve this task. IOG can effectively encode target information into left and right context respectively. Then we combine the left and right context of the opinion target and global context for extracting the corresponding opinion word in the decoder. We contribute four datasets based on several benchmarks. The experimental results demonstrate that our model achieves the best performance across all the datasets from different domains.

In future works, TOWE could be utilized to further improve the performance on downstream sentiment analysis tasks with building a more interpretable model, such as enhanced-feature or multi-task learning. In addition, an end-to-end opinion extractive summary method without given golden targets is also a future work.

Acknowledgements

The authors would like to thank Fang Qian and Fei Zhao for their contribution to building the datasets, and Robert Ridley for his comments on this paper. We also express the gratitude to the anonymous reviewers for their valuable feedback. This work is supported by the National Natural Science Foundation of China (No. 61672277, U1836221), the Jiangsu Provincial Research Foundation for Basic Research (No. BK20170074).

References

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

1780.

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

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

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

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

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

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

Kang Liu, Heng Li Xu, Yang Liu, and Jun Zhao. 2013. Opinion target extraction using partially-supervised word alignment model. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2134–2140.

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

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

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

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy 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*, 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*, 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*, pages 27–35.

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

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

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

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

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

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

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

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

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

Li Zhuang, Feng Jing, and Xiaoyan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 43–50.