Latent Opinions Transfer Network for Target-Oriented Opinion Words Extraction

Zhen Wu,* Fei Zhao,* Xin-Yu Dai,† Shujian Huang, Jiajun Chen
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, 210023, China
{wuz, zhaof}@smail.nju.edu.cn, {daixinyu, huangsj, chenjj}@nju.edu.cn

Abstract

Target-oriented opinion words extraction (TOWE) is a new subtask of ABSA, which aims to extract the corresponding opinion words for a given opinion target in a sentence. Recently, neural network methods have been applied to this task and achieve promising results. However, the difficulty of annotation causes the datasets of TOWE to be insufficient, which heavily limits the performance of neural models. By contrast, abundant review sentiment classification data are easily available at online review sites. These reviews contain substantial latent opinions information and semantic patterns. In this paper, we propose a novel model to transfer these opinions knowledge from resource-rich review sentiment classification datasets to low-resource task TOWE. To address the challenges in the transfer process, we design an effective transformation method to obtain latent opinions, then integrate them into TOWE. Extensive experimental results show that our model achieves better performance compared to other state-of-the-art methods and significantly outperforms the base model without transferring opinions knowledge. Further analysis validates the effectiveness of our model.

Introduction

Target-oriented opinion words extraction (TOWE) (Fan et al. 2019) is a new subtask of aspect-level sentiment analysis (ABSA) (Pang and Lee 2008; Liu 2012; Pontiki et al. 2014), which aims to extract the corresponding opinion words for a given opinion target from a review sentence. Opinion targets, also known as aspect terms, are the words or phrases in the sentence representing features or entities toward which users show attitudes. Opinion words refer to those terms of a sentence used to express attitudes or opinions explicitly. Figure 1 shows an example of TOWE. In the sentence "*waiters are very friendly and the pasta is out of this world.*", the terms "*waiters*" and "*pasta*" are two given opinion targets. TOWE needs to extract the word "*friendly*" as the opinion word of the opinion target "*waiters*" and the opinion words span "*out of this world*" for the target "*pasta*".

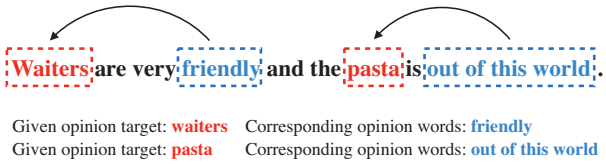

Figure 1: An example of TOWE. The words highlighted in red are two given opinion targets. TOWE task aims to extract the spans in blue as opinion words for the given targets. The arrows indicate the correspondence between opinion targets and opinion words. Note that opinion targets are given beforehand in the TOWE task.

Many downstream sentiment analysis tasks, e.g., target-oriented sentiment classification (Tang et al. 2016; Wang et al. 2016b; Xue and Li 2018) and pair-wise opinion summarization (Hu and Liu 2004; Zhuang, Jing, and Zhu 2006; Li et al. 2010), can benefit from TOWE as it provides explicit opinion pairs information. To study this task, Fan et al. (2019) released a benchmark corpus including four datasets and formalized TOWE as a problem of sequence labeling for given targets. Furthermore, they proposed a target-fused neural sequence labeling model and achieved state-of-the-art performance.

Despite the promising results of neural network methods, the lack of annotated data still heavily restricts the performance of TOWE. In practical scenarios, users usually refer to a considerable number of opinion targets in a review. It is extremely labor-intensive and time-consuming for annotators to identify all targets of a sentence and locate their corresponding opinion words. The difficulty of annotation causes the datasets of TOWE to be relatively scarce, which finally limits the effectiveness of neural models. In contrast, abundant labeled data of review sentiment classification are easily accessible at online review sites such as Amazon, Yelp and IMDB. Substantial opinions information and semantic patterns are naturally embodied in these reviews. Thus we propose to transfer them from large-scale sentiment classification datasets to the low-resource task TOWE. Although latent opinion knowledge is beneficial for TOWE, there are still two challenges remaining:

*Authors contributed equally.
†Corresponding author.

- The opinions information such as opinion words in sentiment classification datasets are latent and unannotated, we need to find them explicitly before transferring them.
- Since sentiment classification for reviews does not consider the target information, the latent-opinion information obtained is global and independent of the target. Thus, this information cannot be used directly by TOWE.

To address the above issues, we propose a novel model **L**atent **O**pinions **T**ransfer **N**etwork (LOTN) leveraging latent opinions knowledge from resource-rich review sentiment classification datasets to improve TOWE task. Specifically, we first pre-train an attention-based BiLSTM model on review sentiment classification datasets. The attention mechanism (Bahdanau, Cho, and Bengio 2015) is employed to extract possible opinion words through probabilistic weights. To solve the second issue, we design an effective transformation method to convert the global attention distribution over words in the sentiment classification model to latent target-dependent opinion words. Finally, we integrate these captured opinions words into TOWE model via an auxiliary learning signal. Additionally, we incorporate the encoder of the pretrained model to further guide TOWE model to learn latent opinions, which proves effective.

We evaluate the LOTN model on the four benchmark datasets. Results from extensive experiments indicate that our model achieves new state-of-the-art performance for TOWE and performs significantly better than our base model that does not transfer opinion knowledge. Further in-depth analysis also validates the effectiveness of our model. To the best of our knowledge, it is the first work to improve TOWE by transferring latent opinions knowledge of review sentiment classification datasets.

The main contributions of this work include:

- In tackling the problem of insufficient annotated data, we are the first to propose transferring latent opinion knowledge from resource-rich review sentiment classification datasets to the low-resource task of TOWE.
- To transfer opinion information effectively, we propose a novel model that obtains latent opinion words from a sentiment classification model and integrates them into TOWE via an auxiliary learning signal.
- The experiment results indicate that our model achieves better results compared to state-of-the-art methods. Extensive analysis validates the effectiveness of our model.

Preliminary

In this section, we will introduce the task formalization of TOWE and the pretrained sentiment classification model that is used for transferring latent opinions.

TOWE Formalization

TOWE aims to extract the corresponding opinion words for a given target from a sentence, which can be formalized as a task of sequence labeling for given targets. Specifically, given a review sentence $s = \{w_1, w_2, \cdots, w_n\}$ consisting n words and an opinion target w_t in the sentence s (**Note that**, we notate an opinion target as one word for simplicity and t is the position of the target in the sentence), the goal is to tag each word w_i in s with a label $y_i \in \{B, I, O\}$ (B: Beginning, I: Inside, O: Others). The spans composed by the tags B and I represent the corresponding opinion words of the target w_t. For example, the sentence in Figure 1 is tagged as w_i/y_i for different opinion targets as shown in Table 1.

Table 1: Different labeling results of a sentence when given different opinion targets. The opinion targets are highlighted in underline and the opinion words/phrases are in bold.

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
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

Pretraining Sentiment Classification Model

Review sentiment classification aims to detect overall sentiment polarity (e.g., positive or negative) of a review text. Before transferring latent opinions, we first pretrain an attention-based BiLSTM model on large-scale review sentiment classification datasets.

Specifically, we regard a review from sentiment classification datasets as a long sentence $\{w_1, w_2, \cdots, w_m\}$ consisiting m words, and map them into the corresponding vector representations $\{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m\}$ by looking up an embedding table. Then a BiLSTM network is applied to encode the word representations $\{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m\}$ and genenrate the context representations $\{\mathbf{h}_1^{sc}, \mathbf{h}_2^{sc}, \cdots, \mathbf{h}_m^{sc}\}$.

The attention mechanism is employed to capture the latent and global opinion words that are significant to sentiment classification. The attention weight α_i of $\mathbf{h}_i^{sc}$ is defined as:

$$u(\mathbf{h}_i^{sc}, \mathbf{h}_{avg}^{sc}) = \mathbf{h}_i^{sc} \cdot \mathbf{W}_u \cdot \mathbf{h}_{avg}^{sc} + b_u, \quad (1)$$

$$\alpha_i = \frac{\exp(u(\mathbf{h}_i^{sc}, \mathbf{h}_{avg}^{sc}))}{\sum_{j=1}^{m} \exp(u(\mathbf{h}_j^{sc}, \mathbf{h}_{avg}^{sc}))}, \quad (2)$$

where $\mathbf{h}_{avg}^{sc}$ is the average of all hidden states, i.e., $\mathbf{h}_{avg}^{sc} = \sum_{j=1}^{m} \mathbf{h}_j^{sc}/m$, $\mathbf{W}_u$ and b_u are the weight matrix and bias.

The review representation $\mathbf{r}_{sc}$ is a weighted sum of all hidden states:

$$\mathbf{r}_{sc} = \sum_{i=1}^{m} \alpha_i \mathbf{h}_i^{sc}. \quad (3)$$

Finally, the representation $\mathbf{r}_{sc}$ is fed to a linear layer and a softmax layer to predict the sentiment label of the review. We train the sentiment classification model by minimizing the cross-entropy loss between the predicted sentiment distribution and the ground truth. After pretraining is finished, all parameters in sentiment classification model are fixed.

Latent Opinions Transfer Network

In this section, we introduce our model Latent Opinions Transfer Network in details. We will present the overall architecture of the model, our base TOWE model and two proposed latent opinions transfer methods, as well as the final decoding and training.

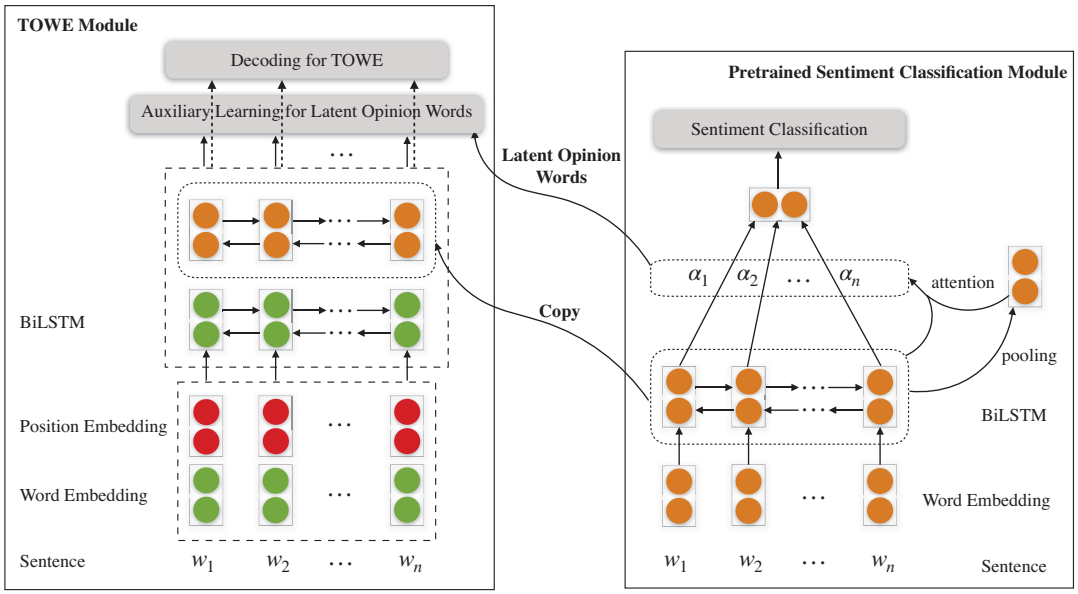

Figure 2: The architecture of Latent Opinions Transfer Network. Different opinion targets in a sentence have different position embeddings.

Overall Description

Figure 2 shows the overall architecture of Latent Opinions Transfer Network (LOTN). It consists of two components mainly: a TOWE module and a pretrained review sentiment classification module. We design a simple and effective network as our base TOWE module, namely position embedding based BiLSTM (PE-BiLSTM). The pretrained sentiment classification module is the aforementioned attention-based BiLSTM network. LOTN transfers latent opinions from the sentiment classification module to the TOWE module through two different perspectives.

Firstly, the BiLSTM layer of the pretrained sentiment classification module contains substantial implicit opinion information and semantic patterns. We integrate this information into the encoding layer of TOWE module to introduce external opinion knowledge.

Secondly, the latent opinion words captured by the attention process of the pretrained module are global and target-independent since review sentiment classification does not consider the target information. To address this issue, we propose a heuristic transformation method to convert global attention weights over words to latent target-dependent opinion words by considering the position information of the target and other words. Then we incorporate them into the TOWE module via an auxiliary learning signal.

Position Embedding based BiLSTM

In the base model, we use position embedding to model the target information instead the state-of-the-art model IOG (Fan et al. 2019) because IOG employs six different positional and directional LSTMs simultaneously and suffers from high model complexity. In contrast to IOG, position embedding is a simple and effective method of modeling position information and widely used in Natural Language Processing (Lin et al. 2016; Gehring et al. 2017; Vaswani et al. 2017; Gu et al. 2018).

Given the sentence $s = \{w_1, w_2, \cdots, w_n\}$ and the opinion target w_t in the sentence, we first generate the relative distance index $l_i = |i - t|$ of each word w_i in s by calculating the relative distance from w_i to the target w_t. Then the distance index l_i is mapped into the positional representation by using a position embedding tabel $\mathbf{E}_{pos} \in \mathbb{R}^{L \times d_1}$, where d_1 is the embedding dimension and L is the maximal position index. In addition, we also use a word embedding table $\mathbf{E}_{emb} \in \mathbb{R}^{|V| \times d_2}$ to obtain the semantic representation of word. The representation $\mathbf{e}_i$ of each word w_i is formed by concatenating the word vector and the corresponding position vector:

$$\mathbf{e}_i = [\mathbf{E}_{emb}(w_i); \mathbf{E}_{pos}(l_i)], \tag{4}$$

where the $[\cdot;\cdot]$ denotes the vector concatenation operation.

Finally, we employ a BiLSTM network to capture the contextual information of each word. The simplified update rule can be written as follows:

$$\mathbf{h}_i^t = \text{BiLSTM}(\mathbf{h}_{i-1}^t, \mathbf{e}_i, \theta_t), \tag{5}$$

where θ_t represents the parameters of BiLSTM.

In the base model, the context representation $\mathbf{h}^t$ can be used for predicting the opinion words of the given target.

Transferring Pretrained Encoder

In order to transfer latent opinions knowledge, we also feed the sentence s of TOWE task to the pretrained sentiment classification module to yield the corresponding hidden states $\{\mathbf{h}_1^{sc}, \mathbf{h}_2^{sc}, \cdots, \mathbf{h}_n^{sc}\}$ and attention weights $\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$.

From the semantic level view, the encoder of the pre-trained sentiment classification module holds substantial implicit opinion information, and thus we integrate it into the TOWE module by concatenating two hidden states:

$$\mathbf{r}_i = \left[\mathbf{h}_i^t; \mathbf{h}_i^{sc}\right], \tag{6}$$

where $\mathbf{r}_i$ contains both task-specific context representations and external opinions knowledge from review sentiment classification datasets.

Transferring Latent Opinion Words

To effectively transfer latent opinion words from sentiment classification module to TOWE module, we design a heuristic transformation method and an auxiliary learning signal respectively to capture and integrate them into TOWE.

Transformation Method As we have mentioned, the attention mechanism in the sentiment classification module captures latent opinion words in the style of probabilistic weights. However. these probabilistic opinions information are global and target-independent. Intuitively, the word that is closer to the opinion target is more likely to be the opinion word of the target. Thus we introduce the opinion target information into the attention distribution by a target-relevant distance weight c_i:

$$\alpha_i{}' = c_i \cdot \alpha_i, \tag{7}$$

$$c_i = 1 - \frac{|i-t|}{n}, \tag{8}$$

where n is the length of input sentence, t indicates the position of the opinion target w_t in the sentence, and $|i-t|$ denotes the relative distance between the word w_i and the target w_t. It can be observed that a closer word to the target has a bigger distance weight. To regain the probabilistic attention distribution, the target-dependent attention weight $\alpha_i{}'$ is re-normalized:

$$\beta_i = \frac{\alpha_i{}'}{\sum_{j=1}^{n} \alpha_j{}'}. \tag{9}$$

Finally, we use a heuristic strategy to convert the normalized attention weight β_i into the binary latent opinion words by the threshold $\frac{1}{n}$:

$$y_i^a = \begin{cases} 1 & \text{if } \beta_i \geq \frac{1}{n}, \\ 0 & \text{otherwise}, \end{cases} \tag{10}$$

where $y_i^a = 1$ denotes that the word w_i is a latent and target-relevant opinion word from the perspective of sentiment classification module and 0 indicates not.

Auxiliary Learning Signal In fact, the y_i^a is a pseudo label and represents the opinions knowledge from the sentiment classifcation module. We integrate these latent opinions into TOWE module by auxiliary learning signal:

$$\hat{\mathbf{y}}_i^a = \mathrm{softmax}(\mathbf{W}_\mathrm{a}\mathbf{r}_\mathrm{i} + \mathbf{b}_\mathrm{a}), \tag{11}$$

$$\mathcal{L}_a = -\sum_{i=1}^{n}\sum_{k=0}^{1} \mathbb{I}(y_i^a = k)\log(\hat{y}_{i,k}^a), \tag{12}$$

where $\mathbf{W}_a$ and $\mathbf{b}_a$ are the weight matrix and the bias, $\hat{\mathbf{y}}_i^a$ represents the prediction probability and $\mathbb{I}(\cdot)$ is the indicator function. LOTN embraces these latent opinions knowledge by optimizing the auxiliary loss $\mathcal{L}_a$, which helps TOWE moduel decode the opinion words of the target better.

Decoding and Training

Since the context representation $\mathbf{r}_i$ contains both task-specfic opinions and tranferred opinions knowledge, LTON finally use $\mathbf{r}_i$ to predict the tag $y_i \in \{B, I, O\}$ of the word w_i. It can regarded as a three-class classfication problem at each position of the sentence s. We use a linear layer and a softmax layer to compute prediction probaility $\hat{\mathbf{y}}_i$:

$$\hat{\mathbf{y}}_i = \mathrm{softmax}(\mathbf{W}_\mathrm{t}\mathbf{r}_\mathrm{i} + \mathbf{b}_\mathrm{t}), \tag{13}$$

where $\mathbf{W}_t$ is the weight matrix and $\mathbf{b}_t$ denotes the bias. The cross-entropy loss of TOWE task can be defined as follows:

$$\mathcal{L}_t = -\sum_{i=1}^{n}\sum_{k=0}^{2} \mathbb{I}(y_i = k)\log(\hat{y}_{i,k}), \tag{14}$$

here the tags $\{O, B, I\}$ are correspondingly converted into labels $\{0, 1, 2\}$ and y_i denotes the ground truth label.

LOTN also integrates latent opinions through auxiliary learning signal $\mathcal{L}_a$. Thus the final loss is defined as follows:

$$J = \mathcal{L}_t + \lambda\mathcal{L}_a, \tag{15}$$

where λ measures the importance of auxiliary learning and can be adjusted. We minimize the loss J to optimize the LOTN model.

Experiments

Datasets and Metrics

We evaluate our model on four benchmark datasets (Fan et al. 2019). The statistics of the datasets are summarized in Table 2. The datasets 14res and 14lap are derived from SemEval Challenge 2014 task 4 (Pontiki et al. 2014), 15res and 16res are respectively from SemEval Challenge 2015 task 12 (Pontiki et al. 2015) and SemEval Challenge 2016 task 5 (Pontiki et al. 2016). The suffixes "res" and "lap" respectively represent reviews from restaurant domain or laptop domain. The original SemEval Challenge datasets are very popular benchmarks for ABSA subtasks. They provide the annotation of opinion targets for each sentence, but not the corresponding opinion words. To perform TOWE task, Fan et al. (2019) annotate the corresponding opinion words for the given targets from sentences and remove the cases without explicit opinion words.

To pretrain the sentiment classification model, we use the two datasets respectively from Amazon Review and Yelp Review. The Yelp Review is applied to transfer latent opinions for TOWE datasets 14res, 15res, and 16res. The Amazon Review is used for the dataset 14lap. Table 3 shows the statistics of Amazon Review and Yelp Review.

Following the state-of-the-art work (Fan et al. 2019), we use the metrics precision, recall, and F1-score to measure the performance of different methods. An opinion word/phrase is deemed to be correct on the condition that the starting and ending positions of the prediction are both the same as those of the golden word/phrase.

Table 2: Statistics of TOWE datasets. A sentence may contain multiple opinion targets.

Datasets		#sentences	#targets
14res	Train	1,627	2,643
	Test	500	864
14lap	Train	1,158	1,634
	Test	343	482
15res	Train	754	1,076
	Test	325	436
16res	Train	1,079	1,512
	Test	329	457

Table 3: Statistics of the two datasets Amazon Review and Yelp Review.

Datasets	#positive	#negative	#total
Yelp Review	266,041	177,218	443,259
Amazon Review	277,228	277,769	554,997

Experiment Settings

We initialize word vectors with 300-dimension word embeddings from GloVe (Pennington, Socher, and Manning 2014). The word vectors are fixed and not fine-tuned during the training stage. We set the dimension of position embeddings to be 300. The position embeddings, all weight matrices and biases are randomly initialized by a uniform distribution $U(-0.01, 0.01)$. The dimension of LSTM cell is set to 200. We adopt Adam optimizer (Kingma and Ba 2015) to update parameters of models. The initial learning rate is 0.001 and mini-batch size is set to 25. The dropout (Hinton et al. 2012) is applied on embedding layer with probability 0.5. We randomly select 20% of training set as validation set for tuning hyper-parameters and early stopping. Finally, we run each model five times and report the average result of them.

Compared Methods

We compare our model with the following methods:

- **Distance-rule** first performs POS tagging on the sentence, then regards the nearest adjective to the opinion target as the corresponding opinion word (Hu and Liu 2004).
- **Dependency-rule** collects the POS tags of opinion targets and opinion words and the dependency path between them as rule templates from the training set. Then the hign-frequency dependency templates are applied to detect the corresponding opinion words of opinion targets for the testing set (Zhuang, Jing, and Zhu 2006).
- **LSTM/BiLSTM** employs word embeddings to represent words, then uses LSTM/BiLSTM network to capture the context information of input. Finally, each hidden state is fed to a softmax classifier for three-class classification to extract the opinion words of the given target (Liu, Joty, and Meng 2015).
- **Pipeline** is a combination method of BiLSTM and Distance-rule method (Fan et al. 2019). It first trains a BiLSTM model on the training set. During the testing stage, the Pipeline method uses BiLSTM model to extract all the opinion words spans, then select the nearest span to the target as the extraction result.
- **TC-BiLSTM** follows the design of the work for target-oriented sentiment classification (Tang et al. 2016). This method adopts the average pooling to obtain dimension-fixed target vector, then concatenate it with word vector at each position of the sentence. Finally, the concatenation of target vectors and word vectors are fed to BiLSTM for sequence labeling.
- **IOG** adopts six different positional and directional LSTMs to extract the opinion words of the target. This model achieves very powerful performance and is the state-of-the-art method in TOWE (Fan et al. 2019).
- **PE-BiLSTM** is our base model. It incorporates the target information into TOWE by position embedding, then uses a BiLSTM to extract the opinion words.

Main Results

The main experiment results are shown in Table 4. We can observe that pure rule-based methods perform very poorly compared to the supervised learning models. The method Distance-rule achieves the worst recall and F1-score in most of the datasets since it only detects the single word as opinion word. Dependency-rule method obtains some improvements, but it is still worse than the sequence labeling models.

Compared to other neural network methods, LSTM and BiLSTM achieve poor performance because they are target-independent. They extract the same span as opinion words for different targets in a sentence. Pipeline uses BiLSTM to extract the spans of opinion words, then selects the closest span to the target as the final result. In fact, it is a target-dependent method and the distance strategy makes Pipeline method obtain nearly 20% improvement of precision over BiLSTM in the dataset 14res. By contrast, TC-BiLSTM performs worse than Pipeline and even is inferior to BiLSTM in 14lap. One possible reason for this is that concatenating the same target vector at each position is not a good approach to incorporate the target information. IOG employs six different positional and directional LSTMs to generate rich target-dependent context representations, achieving very powerful results in all datasets. However, it also suffers from high model complexity.

PE-BiLSTM is our base method that adopts position embedding to incorporate the target information. We can observe that this simple method obtain competitive performance and is only inferior to IOG and LOTN. The experiment results show that our model LOTN achieves the best F1-score in all datasets. Compared to its base version PE-BiLSTM, LOTN obtains about 4%$\sim$5% improvements in F1-score. In addition, LOTN outperforms the previous state-of-the-art method IOG by 1.98% and 2.02% F1-score respectively in the datasets 14res and 16res. These observations demonstrate that our model can effectively transfer the latent opinions knowledge from sentiment classification datasets to the TOWE task.

Table 4: Main experiment results(%). Best results are in bold (P, R, and F1-score, the larger is the better). The marker † represents that LOTN outperforms other methods significantly ($p < 0.01$) .

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
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
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
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	82.38	78.25	80.23	73.43	**68.74**	70.99	72.19	**71.76**	71.91	84.36	79.08	81.60
PE-BiLSTM	80.10	76.51	78.26	72.01	64.20	67.83	70.36	65.73	67.96	82.27	74.95	78.43
LOTN	**84.00**†	**80.52**†	**82.21**†	**77.08**†	67.62	**72.02**†	**76.61**†	70.29	**73.29**†	**86.57**†	**80.89**†	**83.62**†

Table 5: Experiment results of adding the transferred encoder or auxiliary learning on PE-BiLSTM(%).

Models	14res			14lap			15res			16res		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PE-BiLSTM	80.10	76.51	78.26	72.01	64.20	67.83	70.36	65.73	67.96	82.27	74.95	78.43
+transferred encoder	84.57	79.54	81.97	77.50	67.47	72.13	75.90	69.00	72.26	86.05	79.81	82.79
+auxiliary learning	84.10	77.20	80.49	75.63	66.42	70.71	76.31	68.67	72.29	86.77	79.46	82.93
LOTN	84.00	80.52	82.21	77.08	67.62	72.02	76.61	70.29	73.29	86.57	80.89	83.62

The Effects of Tranferring Encoder and Latent Opinion Words

To investigate the effects of transferring the encoder and latent opinion words, we conduct the following experiments:

- **PE-BiLSTM+transferred encoder**: We only transfer the encoder of the sentiment classification model to TOWE based on the model PE-BiLSTM.
- **PE-BiLSTM+auxiliary learning**: This method only incorporates the latent opinion words from sentiment classification model into PE-BiLSTM via auxiliary learning.

Table 5 shows the experiment results. Compared to the base model PE-BiLSTM, we can find that PE-BiLSTM+transferred encoder and PE-BiLSTM+auxiliary learning both achieve significant and consistent improvements on all datasets. The comparisons validate the effectiveness of transferring the encoder and latent opinion words from sentiment classification model for the TOWE task. After integrating both the transferred encoder and auxiliary learning, the model LOTN obtains further improvements. The results indicate that the proposed two methods are useful for the final model LOTN and they transfer opinions knowledge from different perspectives.

The Effect of the Hyper-parameter λ

To analyze the effect of different λ on our model, we adjust λ of Equation 15 in $(0, 1)$ to conduct experiments and the step is 0.05. Figure 3 shows the results of LOTN with different λ on four datasets.

We can observe that LOTN achieves the relatively stable performance with varying λ on the datasets 14res, 15res and 16res, which indicates the robustness of our method. In general, the performance of LOTN has a downward trend with an increase of λ since the bigger λ has a negative effect on the decoding of the model. Finally, we set λ to be 0.1 on 14lap and 0.2 on other datasets.

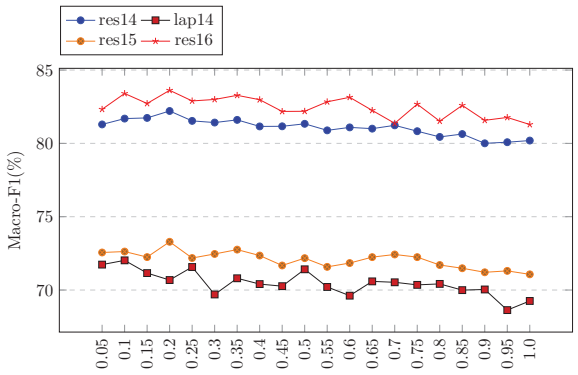

Figure 3: The effect of different hyper-parameter λ.

Case Study

In order to compare different methods and validate the effectiveness of our model, we present some extracted results of the dataset 14res in Table 6.

The first example shows that the Distance-rule method cannot extract opinion phrase for the given target and thus makes the wrong prediction. Comparing the second and third examples, we can find that BiLSTM gives the same predictions for the different opinion targets since it ignores the target information. In the last two examples, the state-of-the-art model IOG and our base model PE-BiLSTM both make incorrect predictions in complicated cases, while our proposed model LOTN extracts target-dependent opinion words successfully. The results demonstrate that our model can leverage latent opinions information from sentiment

Table 6: Examples of the extracted results in different methods. The opinion targets are in underline and the corresponding golden opinion words are in bold. The "NULL" represents that the prediction is empty.

Sentence	Distance-rule	BiLSTM	IOG	PE-BiLSTM	LOTN	
					Latent Opinion Words	Target Decoding
*The bread is **top notch** as well.*	*top*✗	*top notch*✓	*top notch*✓	*top notch*✓	*top notch*	*top notch*✓
***Great** food but the service was dreadful!*	*Great*✓	*dreadful*✗	*Great*✓	*Great*✓	*Great*	*Great*✓
*Great food but the service was **dreadful**!*	*dreadful*✓	*dreadful*✓	*dreadful*✓	*dreadful*✓	*dreadful*	*dreadful*✓
***Good** for a **quick** sushi lunch.*	*quick*✗	*Good, quick*✓	*quick*✗	*quick*✗	*Good*	*Good, quick*✓
*Their twist on pizza is healthy, but **full** of flavor.*	*full*✓	*healthy*✗	*healthy, full*✗	NULL✗	*full*	*full*✓

classification datasets to improve the TOWE task.

Error Analysis

We count the distribution of different error types in the dataset 14res to analyze the weakness of LOTN. The results are given in Table 7. The "NULL" represents the prediction is empty. The "Under-extracted" and "Over-extracted" respectively mean that LOTN extracts the part of the ground truth and extra words besides the golden opinion words.

Table 7: Statistics of different error types for PE-BiLSTM and LOTN in the dataset 14res.

Models	NULL	Under-extracted	Over-extracted	Others	Total
PE-BiLSTM	76	107	49	34	266
LOTN	65	85	62	31	243

It can be observed that PE-BiLSTM and LOTN do not extract any opinion words in more than a quarter of error cases. In two models, the under-extracted error is the main error type. Compared to PE-BiLSTM, LOTN makes fewer mistakes in the NULL and under-extracted type. In contrast, PE-BiLSTM makes fewer over-extracted predictions. The three comparisons consistently indicate that LOTN tends to decode more opinion words under the influence of latent opinions from the sentiment classification model.

In addition, we find that the sentence is often quite long and the golden opinion words are far away from the target in NULL error. As for under-extracted cases, the models usually ignore modifiers such as the word "*most*" in "*most delicious*" or negators, e.g., only extracting "*best*" from "*not the best*". The over-extracted predictions are common in the samples containing multiple targets.

Although our model improves the overall performance of TOWE, LOTN makes some wrong predictions due to transferring some noise opinions. For example, about 5.7% samples of the dataset 14res are predicted successfully by the model PE-BiLSTM but incorrectly by LOTN. The heuristic transformation method is unable to consistently yield correct opinion words for a given target, thus we inevitably introduce some noise during converting global attention knowledge into latent target-dependent opinion words.

Related Work

The early works devote to the research of opinion targets extraction, including unsupervised methods (Popescu and Etzioni 2005; Wang and Wang 2008; Qiu et al. 2011; Liu, Xu, and Zhao 2012) and supervised methods (Jin, Ho, and Srihari 2009; Li et al. 2010; Liu, Joty, and Meng 2015; Poria, Cambria, and Gelbukh 2016; Xu et al. 2018). Recently, some works extract opinion targets and opinion words jointly in a unified framework and achieve promising results (Qiu et al. 2011; Liu, Xu, and Zhao 2012; Liu et al. 2013; Wang et al. 2016a; 2017; Li and Lam 2017). However, these works does not extract the corresponding relation between targets and opinion words.

In fact, there are only a few works focusing on the paired opinion relations. Hu and Liu (2004) propose to use association rule mining for extracting opinion targets and regard the nearest adjective of targets as the corresponding opinion words. Zhuang, Jing, and Zhu (2006) adopt WordNet (Miller 1995) and human-built word lists to find targets and opinion words, then apply dependency-tree templates to extract the valid target-opinion pairs. These two unsupervised methods heavily depend on pre-defined rules and external resources such as syntax parser. Because of the significance of the paired relations of targets and opinion words for downstream sentiment task, Fan et al. (2019) propose a new subtask of ABSA, target-oriented opinion words extraction (TOWE), to extract the corresponding opinion words for a given opinion target from a review. They also design a target-fused neural sequence labeling model and achieve competitive results.

Conclusion

Insufficiency of labeled data heavily restricts the effectiveness of the neural models for TOWE. In this paper, we propose a novel model to transfer latent opinions knowledge from resource-rich review sentiment classification datasets to improve the low-resource task TOWE. Specifically, we propose an effective method to convert the attention knowledge in the sentiment classification model into target-dependent opinion words, then integrate them into TOWE via auxiliary learning signal. In addition, we also integrate the encoder of the sentiment classification model to further improve TOWE. Results from numerous experiments indicate that our approach achieves better performance than other state-of-the-art methods. Extensive analysis also demonstrates the effectiveness of our model.

Acknowledgments

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

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Fan, Z.; Wu, Z.; Dai, X.; Huang, S.; and Chen, J. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *NAACL-HLT*, 2509–2518.

Gehring, J.; Auli, M.; Grangier, D.; and Dauphin, Y. N. 2017. A convolutional encoder model for neural machine translation. In *ACL*, 123–135.

Gu, S.; Zhang, L.; Hou, Y.; and Song, Y. 2018. A position-aware bidirectional attention network for aspect-level sentiment analysis. In *COLING*, 774–784.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR* abs/1207.0580.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *ACM SIGKDD*, 168–177.

Jin, W.; Ho, H. H.; and Srihari, R. K. 2009. A novel lexicalized hmm-based learning framework for web opinion mining. In *ICML*, 465–472. Citeseer.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Li, X., and Lam, W. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *EMNLP*, 2886–2892.

Li, F.; Han, C.; Huang, M.; Zhu, X.; Xia, Y.; Zhang, S.; and Yu, H. 2010. Structure-aware review mining and summarization. In *COLING*, 653–661.

Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; and Sun, M. 2016. Neural relation extraction with selective attention over instances. In *ACL*, 2124–2133.

Liu, K.; Xu, H. L.; Liu, Y.; and Zhao, J. 2013. Opinion target extraction using partially-supervised word alignment model. In *IJCAI*, 2134–2140.

Liu, P.; Joty, S. R.; and Meng, H. M. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *EMNLP*, 1433–1443.

Liu, K.; Xu, L.; and Zhao, J. 2012. Opinion target extraction using word-based translation model. In *EMNLP-CoNLL*, 1346–1356.

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

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends®in Information Retrieval* 2(1–2):1–135.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *SemEval@COLING 2014*, 27–35.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Manandhar, S.; and Androutsopoulos, I. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *SemEval@NAACL-HLT*, 486–495.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Al-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; Clercq, O. D.; Hoste, V.; Apidianaki, M.; Tannier, X.; Loukachevitch, N. V.; Kotelnikov, E. V.; Bel, N.; Zafra, S. M. J.; and Eryigit, G. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *SemEval@NAACL-HLT*, 19–30.

Popescu, A., and Etzioni, O. 2005. Extracting product features and opinions from reviews. In *HLT/EMNLP*, 339–346.

Poria, S.; Cambria, E.; and Gelbukh, A. F. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowl.-Based Syst.* 108:42–49.

Qiu, G.; Liu, B.; Bu, J.; and Chen, C. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics* 37(1):9–27.

Tang, D.; Qin, B.; Feng, X.; and Liu, T. 2016. Effective lstms for target-dependent sentiment classification. In *COLING*, 3298–3307.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Wang, B., and Wang, H. 2008. Bootstrapping both product features and opinion words from chinese customer reviews with cross-inducing. In *IJCNLP*, 289–295.

Wang, W.; Pan, S. J.; Dahlmeier, D.; and Xiao, X. 2016a. Recursive neural conditional random fields for aspect-based sentiment analysis. In *EMNLP*, 616–626.

Wang, Y.; Huang, M.; Zhu, X.; and Zhao, L. 2016b. Attention-based LSTM for aspect-level sentiment classification. In *EMNLP*, 606–615.

Wang, W.; Pan, S. J.; Dahlmeier, D.; and Xiao, X. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *AAAI*, 3316–3322.

Xu, H.; Liu, B.; Shu, L.; and Yu, P. S. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. In *ACL*, 592–598.

Xue, W., and Li, T. 2018. Aspect based sentiment analysis with gated convolutional networks. In *ACL*, 2514–2523.

Zhuang, L.; Jing, F.; and Zhu, X. 2006. Movie review mining and summarization. In *ACM CIKM*, 43–50.