Exploiting Unlabeled Data for Target-Oriented Opinion Words Extraction

Yidong Wang[1][*]   **Hao Wu**[1][*]   **Ao Liu**[1]   **Wenxin Hou**[2]

Zhen Wu[3][†]   **Jindong Wang**[4]   **Takahiro Shinozaki**[1]   **Manabu Okumura**[1]   **Yue Zhang**[5]

[1]Tokyo Institute of Technology   [2]Microsoft STCA
[3]Nanjing University   [4]Microsoft Research Asia   [5]Westlake University
{yidongwang37, wu.364371691}@gmail.com, wuz@nju.edu.cn

Abstract

Target-oriented Opinion Words Extraction (TOWE) is a fine-grained sentiment analysis task that aims to extract the corresponding opinion words of a given opinion target from the sentence. Recently, deep learning approaches have made remarkable progress on this task. Nevertheless, the TOWE task still suffers from the scarcity of training data due to the expensive data annotation process. Limited labeled data increase the risk of distribution shift between test data and training data. In this paper, we propose exploiting massive unlabeled data to reduce the risk by increasing the exposure of the model to varying distribution shifts. Specifically, we propose a novel Multi-Grained Consistency Regularization (MGCR) method to make use of unlabeled data and design two filters specifically for TOWE to filter noisy data at different granularity. Extensive experimental results on four TOWE benchmark datasets indicate the superiority of MGCR compared with current state-of-the-art methods. The in-depth analysis also demonstrates the effectiveness of the different-granularity filters. Our codes are available at https://github.com/TOWESSL/TOWESSL.

1 Introduction

Target-oriented Opinion Words Extraction (TOWE) (Fan et al., 2019) is an important subtask of aspect-based sentiment analysis (ABSA) (Pontiki et al., 2014), which aims to extract the corresponding opinion words for a given opinion target from the sentence. For the TOWE task, opinion targets, also called aspect terms, are the entities or objects in the sentence toward which users show attitudes; opinion words, sometimes known as opinion expressions, are those words explicitly mentioned in the sentence and used to express attitudes or opinions. Figure 1 shows an
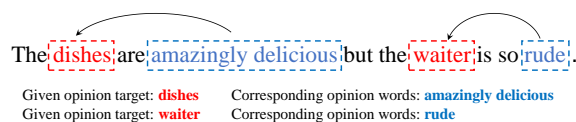

Figure 1: Example of TOWE. Words in red are opinion targets and words in blue are corresponding opinion words. TOWE extracts corresponding opinion words when given opinion targets.

example of the TOWE task. For the sentence "*The dishes are amazingly delicious but the waiter is so rude.*", the terms "*dishes*" and "*waiter*" are two opinion targets. The goal of TOWE is to extract "*amazingly delicious*" as the opinion words for the opinion target "*dishes*" and opinion word "*rude*" when given the opinion target "*waiter*".

While seminal work casts TOWE as a sequence labeling problem, using fully supervised learning methods to identify opinion words and phrases from sentences, recent work shows that external sources of information can be highly useful for improving the performance. In particular, both syntactic knowledge (Dai et al., 2022; Veyseh et al., 2020; Jiang et al., 2021; Zhang et al., 2021b) and sentiment information (Wu et al., 2020b) have been exploited, with the former helping to identify the correlation between opinion targets and opinion expressions, and the latter helping to identify words and phrases that are correlated with sentiment polarities. Existing work integrates these external features via representation structures (Dai et al., 2022; Veyseh et al., 2020; Jiang et al., 2021) and multi-task learning (Wu et al., 2020b; Zhang et al., 2021b).

Intuitively, the goal of opinion words extraction is to obtain structured knowledge from raw data, and therefore ideally the amount of test data, namely the data from which opinion words are mined, should be large. This can inevitably increase the risk of distribution shift between test instances and the training data, even if the test data

---

[*] Equal Contribution.
[†] Corresponding author.

is from the same domain ([Ganin et al., 2016](#)). Existing work, however, adopt a supervised training setting, with the model being tuned on a set of fixed training data. To address this issue, we consider making use of raw text inputs to increase the exposure of the model to varying distribution shifts. Our main idea is to (1) find raw data in a similar domain as the target-oriented opinion words extraction training data (i.e. gold data), (2) using a model trained on the gold data to assign silver labels, and then (3) design a set of filters to select the most useful set of silver data, so that (4) the selected silver data can be used together with the gold data for training a final model.

As a pre-processing step, we train a opinion target extraction model, which is used to label the set of raw data. Those raw sentences without any opinion target are filtered. The resulting raw data has the same format as the input structure of TOWE data. Then in our implementation, steps (2) to (4) above are done in a joint batch training process. First, we initialize a TOWE model using BERT. Then in each batch, we randomly sample a subset of gold data and a subset of raw data, using the current model to assign pseudo labels to the raw data. According to the current model probabilities and an external sentiment classifier, we filter the silver data by removing low-confidence sentences, as well as masking low-confidence words from the remaining sentences. Finally, the model parameters are updated by using the standard cross-entropy losses separately on both the gold and silver data, so that the next batch of training can start with the new model. The training process continues for a fixed number of iterations, and the model with the highest development scores are selected for testing.

Results on four standard benchmarks show that our method gives significantly improved results when raw data are used. We achieve the best reported results on all the datasets. In addition, the critical ablation demonstrates our sentence-level and word-level noises filtering both bring improvements for the final results. In-depth analysis shows our method significantly reduces the extraction errors on different error types. To our knowledge, we are the first to consider the use of unlabeled data for TOWE, successfully giving state-of-the-art results on benchmarks. Our code and datasets are available at `https://github.com/TOWESSL/TOWESSL`.

Labeled data:
The/O entire/O dining/O experience/O was/O very/B wonderful/I !/O
Unlabeled data with the generated pseudo opinion target:
Their menu is too expensive for a bubble drink .

Figure 2: Examples of labeled data and unlabeled data with the pseudo opinion target. Words with underline indicate opinion targets. The span in the labeled data beginning with B and followed by I represent the corresponding opinion words.

2 Background

2.1 Task Formalization

The TOWE task requires the opinion target as the input and extracts the target-oriented opinion words from the sentence. It can be formulated as a target-oriented sequence labeling task. In this work, we use the notation s to represent a labeled sentence from the TOWE dataset, and use s^u to denote a unlabeled sentence. Formally, given an input sentence $s = \{w_1, w_2, \ldots, w_n\}$ consisting of n words and an opinion target w_t in s (here we notate an opinion target as one word for simplicity and t is the position of the opinion target in the sentence), the goal of TOWE is assign a corresponding label $y_i \in \{B, I, O\}$(B: **B**eginning, I: **I**nside, O: **O**thers) for each word w_i in s. The spans beginning with B and followed by I represent the corresponding opinion words of the opinion target w_t. Figure 2 shows an sequence labeling example of TOWE.

2.2 Pseudo Opinion Targets Labeling

Raw unlabeled data can not be directly used in the semi-supervised scenario for TOWE, as they lack the necessary annotations of opinion targets. We use the opinion targets in labeled TOWE data as the ground truth to train a BERT-based target extraction model.

Specifically, given an input sentence $s = \{w_1, w_2, \ldots, w_n\}$ from the TOWE dataset, we employ a BERT model ([Devlin et al., 2019](#)) to generate the context representation of the each word w_i as follows:

$$\mathbf{h}_1^{pt}, \ldots, \mathbf{h}_n^{pt} = \text{BERT}(w_1, \ldots, w_n) \quad (1)$$

Then the context representation $\mathbf{h}_i^{pt}$ of the word w_i is fed to a linear layer and a softmax layer to predict the corresponding label. Similar to TOWE, the BIO scheme is used. We train the opinion target extraction model by minimizing the cross-entropy loss between the predicted BIO label distribution
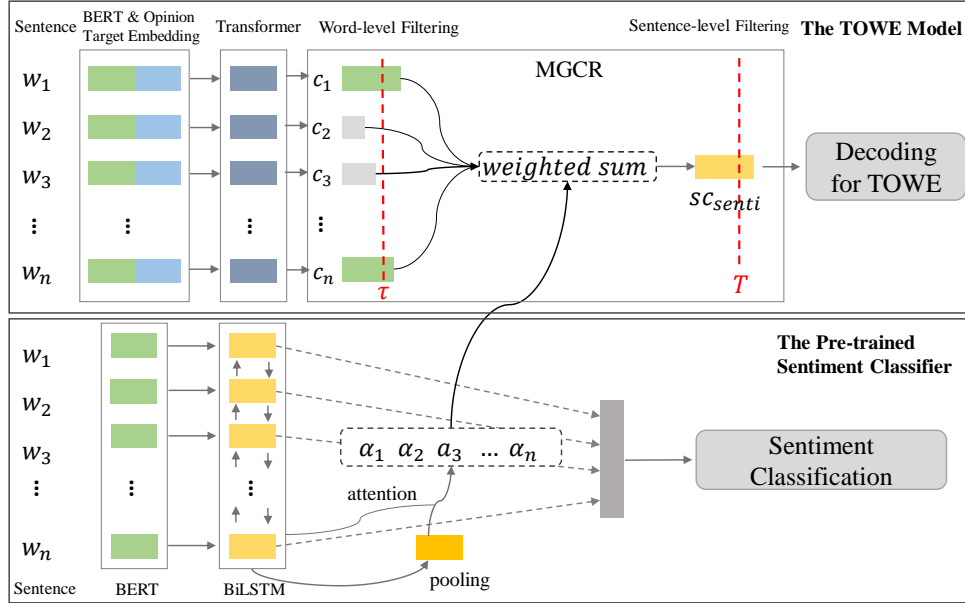
Figure 3: Overview of the architecture of multi-grained consistency regularization. For simplicity, we mark the confidence of i-th word as c_i. Note that the input sentence of the TOWE model is the same as the input sentence of the pre-trained sentiment classifier.

and the ground truth. After training, we can employ the opinion target extraction model to obtain the pseudo opinion targets of the unlabeled sentence s^u, which makes semi-supervised learning feasible for the TOWE task.

2.3 Consistency Regularization

Consistency regularization is a semi-supervised learning method that shows effectiveness in a variety of fields (Sohn et al., 2020; Chen et al., 2020; Xie et al., 2020; Zhang et al., 2021a). It improves the generalization performance of a TOWE model by generating a perturbed version $\omega(s^u)$ of the original unlabeled sentence s^u and forcing the predicted category of each word to be the same, where ω is the perturbing function (Lee et al., 2013).

To perform semi-supervised learning for TOWE, we feed the unlabeled sentence s^u and the position t of the pseudo opinion target w_t^u to the TOWE model (see section 3.1), and obtain the predicted probability $p_i(y|\theta; s^u, t)$ of the word w_i^u, where θ represents the parameters of the TOWE model. A vanilla consistency loss for consistency regularization of TOWE is computed as:

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{H}(\hat{p}_i(y|\theta; s^u, t), p_i(y|\theta; \omega(s^u), t)), \quad (2)$$

where $\hat{p}_i(y|\theta; s^u, t) = \arg\max p_i(y|\theta; s^u, t)$ and $\hat{p}_i(y|\theta; s^u, t)$ denotes the predicted label of the w_i^u,

the $\mathcal{H}(\cdot, \cdot)$ refers to the cross-entropy loss. In this work, we use Random Mask and Random Synonym Replacement by using WordNet (Miller, 1995) as the perturbing function ω.

3 Method

Figure 3 shows the the framework of our *Multi-Grained Consistency Regularization (MGCR)* method. The TOWE model is a BERT-based neural sequence labeling network. As mentioned in the introduction, we initialize a BERT TOWE model (Section 3.1), and then iteratively conduct batch training. In each batch, we impose a standard cross-entropy loss on a set of sampled gold data (Eq 5), and a regularization loss on a set of randomly sampled raw data with targets (Eq 11). The latter is defined by using the current model to assign TOWE labels, and then adding a sentence-level (Section 9) and a word-level (Section 10) filter. Besides, as mentioned in the introduction, MGCR exploits latent opinion words from a pre-trained review sentiment classification (Wu et al., 2020b) to filter noises more accurately.

3.1 TOWE Model

For the sentence $s = \{w_1, w_2, \ldots, w_n\}$, the TOWE model first employs a pre-trained BERT to generate the context representations $\{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$ of s. To incorporate the opinion target information

into the sentence representations, we use the number 1 to represent the position of opinion target w_t in s and 0 to denote the positions of other words, and then map them to the position embedding sequence $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$. We integrate opinion target information by concatenating, i.e., $\widetilde{\mathbf{h}}_i = [\mathbf{h}_i; \mathbf{e}_i]$.

To better encode opinion target information, we additionally employ a multi-layer Transformer architecture (Vaswani et al., 2017) to generate the target-specific context representations as follows:

$$\mathbf{r}_1, \ldots, \mathbf{r}_n = \text{Transformer}(\widetilde{\mathbf{h}}_1, \ldots, \widetilde{\mathbf{h}}_n). \quad (3)$$

Finally, we use the representation $\mathbf{r}_i$ to predict the opinion word probability $p_i(y|\theta; s, t)$ of the word w_i when given the opinion target w_t:

$$p_i(y|\theta; s, t) = \text{softmax}(\mathbf{W}_r \mathbf{r}_i + \mathbf{b}_r), \quad (4)$$

where $\mathbf{W}_r$ and $\mathbf{b}_r$ are learnable weight and bias. The loss of supervised learning on the TOWE dataset is defined as:

$$\mathcal{L}_s = \frac{1}{n} \sum_{i=1}^{n} \mathcal{H}(y_i, p_i(y|\theta; s, t)). \quad (5)$$

3.2 Multi-Grained Consistency Regularization

3.2.1 Sentence-level Filtering

Given an unlabeled input sentence $s^u = \{w_1^u, w_2^u, \ldots, w_n^u\}$ with w_t^u as the pseudo opinion target position, the sentence-level confidence sc_{avg} is defined as follows:

$$sc_{avg} = \frac{1}{n} \sum_{i=1}^{n} \max(p_i(y|\theta; s^u, t)), \quad (6)$$

where $\max(p_i(y|\theta; s^u, t))$ (i.e., the maximum of the probabilities of B, I, and O) is the confidence of the i-th word. Sentences with confidences below the given threshold T are masked during training. In Eq. (6), different words in the sentence are treated equally, which ignores the importance of the opinion words. We highlight the confidences of opinion words identified by their larger sentiment scores. Specifically, we obtain the sentiment-attention scores from a pre-trained attention-based sentiment classifier parameterized by θ_{senti} from (Wu et al., 2020b)[1]. After acquiring the representations $\{z_1, z_2, \ldots, z_n\}$ of the unlabeled sentence,

---

[1]The sentiment classifier we use is the same as (Wu et al., 2020b), except that we use BERT as the word representation model while they use GloVe vectors (Pennington et al., 2014).

the attention score α_i of z_i is calculated as follows:

$$z_{avg} = \frac{1}{n} \sum_{i=1}^{n} z_i,$$
$$f(z_i, z_{avg}) = z_i \cdot \mathbf{W} \cdot z_{avg} + \mathbf{b}, \quad (7)$$
$$\alpha_i = \frac{e^{f(z_i, z_{avg})}}{\sum_{j=1}^{n} e^{f(z_j, z_{avg})}},$$

where $\mathbf{W}$ and $\mathbf{b}$ are learnable weight and bias. We then compute the sentiment-aware confidence based on the obtained attention scores as:

$$sc_{senti} = \sum_{i=1}^{n} \alpha_i \cdot \max(p_i(y|\theta; s^u, t)). \quad (8)$$

Similar to Eq. (6), any sentence whose sentiment-aware confidence is lower than the threshold T will be masked in consistency regularization. The final filtering mechanism of noisy sentences can be expressed as:

$$\mathbb{1}(sc_{senti} > T), \quad (9)$$

where $\mathbb{1}(\cdot > T)$ is the indicator function for confidence thresholding with T being the threshold.

3.3 Word-level Filtering

To further reduce the noise in the unlabeled data, we filter the noisy words with a more fine-grained confidence thresholding mechanism. The filtering of noisy words can be expressed as

$$\mathbb{1}(\max(p_i(y|\theta; s^u, t)) > \tau). \quad (10)$$

Any word with a confidence lower than τ is masked. Note that the word-level threshold τ can be different from the sentence-level threshold T.

3.4 Training Objective

For labeled sentences, the supervised loss is the same as Eq. (5). For unlabeled sentences, the consistency loss is:

$$\mathcal{L}_c = \mathbb{1}(sc_{senti} > T)$$
$$\cdot \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(\max(p_i(y|\theta; s^u, t)) > \tau) \right.$$
$$\left. \cdot \mathcal{H}(\hat{p}_i(y|\theta; s^u, t), p_i(y|\theta; \omega(s^u), t)) \right\}. \quad (11)$$

The final training objective is given by:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_c. \quad (12)$$

Datasets		#sentences	#opinion targets
14res	Train	1,627	2,643
	Test	500	864
15res	Train	754	1,076
	Test	325	436
16res	Train	1,079	1,512
	Test	329	457
14lap	Train	1,158	1,634
	Test	343	482
Yelp	Unlabeled	100,000	-
Amazon	Unlabeled	100,000	-

Table 1: Statistics of TOWE datasets and unlabeled datasets. For TOWE datasets, sentence may contain multiple opinion targets. For unlabeled datasets, we randomly sampled data from Yelp for 14res, 15res, 16res datasets and Amazon for 14lap dataset. The unlabeled data is available at https://github.com/TOWESSL/TOWESSL.

4 Experiments

4.1 Datasets and Metrics

Following previous studies (Fan et al., 2019; Wu et al., 2020b; Veyseh et al., 2020; Mao et al., 2021; Jiang et al., 2021; Feng et al., 2021; Zhang et al., 2021b), we conduct evaluations on four benchmark datasets for TOWE. The suffixes 'res' and 'lap' refer to restaurant reviews and laptop reviews, respectively. For the 14res, 15res, and 16res datasets, we use the unlabeled sentences from Yelp[2]. For the 14lap dataset, we use the unlabeled sentences from Amazon Electronics[3] (Ni et al., 2019). The statistics of these datasets are listed in Table 1.

We use the evaluation metrics of precision, recall, and F1 score to measure the performance of different methods following previous studies (Fan et al., 2019; Wu et al., 2020b; Jiang et al., 2021). An extraction is considered correct only if opinion words from the beginning to the end are all correctly predicted.

4.2 Experimental Settings

For our MGCR method, we set the hidden size of both the BERT and the Transformer to 512. The mini-batch sizes of labeled and unlabeled data are set to 16 and 96, respectively. All parameters are optimized using the AdamW optimizer (Loshchilov and Hutter, 2018) with an initial learning rate 2e-5 for BERT and 2e-4 for others. We randomly split

---

[2] https://www.yelp.com
[3] https://www.amazon.com

Hyperparameter	TOWE model	Sentiment Classifier
Batch size	16(96)	128
Epochs	50	-
Steps	-	3000
Learning rate (BERT)	2e-5	1e-5
Learning rate (Others)	2e-4	1e-4
Hidden dimension	512	512
Optimizer	AdamW	AdamW

Table 2: Experimental setting of the training of the TOWE model and the sentiment classifier. For the TOWE model, batch size for labeled data is 16 and 96 for unlabeled data.

20% of the training set as the validation set and used early stopping. We search different combinations of sentence-level and word-level confidence thresholds for each dataset and use the ones with best validation performances. The experimental results of different threshold combinations are provided in Table 5, Table 6, Table 7 and Table 8. The training details of the pseudo opinion targets generator and the sentiment classifier are provided in Table 2, respectively. We pre-train the sentiment classifier on the same data as the unlabeled data used for MGCR.

4.3 Baselines

We compare our MGCR method with the following methods.

Distance-rule (Fan et al., 2019) uses POS tags and regards the nearest adjective to the opinion target as the corresponding opinion word.

Dependency-rule (Fan et al., 2019) builds templates from the training set with the POS tags of opinion targets and opinion words and the dependency path between them, then uses the high-frequency dependency templates for target-oriented opinion words extraction on the testing set.

TC-BiLSTM (Fan et al., 2019) follows the design of the work for target-oriented sentiment classification (Tang et al., 2016) and concatenate an opinion target embedding for each word position to perform sequence labeling.

IOG (Fan et al., 2019) employs six different positional and directional LSTMs to encode sentence and then extract the opinion words of the target.

LOTN (Wu et al., 2020b) transfers the latent opinion knowledge from sentiment classification task into the TOWE task via an auxiliary learning task.

ONG (Veyseh et al., 2020) leverages syntax-based opinion possibility scores and the syntactic

Methods	14res			15res			16res			14lap		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Distance-rule (Fan et al., 2019)	58.39	43.59	49.92	54.12	39.96	45.97	61.90	44.57	51.83	50.13	33.86	40.42
Dependency-rule (Fan et al., 2019)	64.57	52.72	58.04	65.49	48.88	55.98	76.03	56.19	64.62	45.09	31.57	37.14
TC-BiLSTM (Fan et al., 2019)	67.65	67.67	67.61	66.06	60.16	62.94	73.46	72.88	73.10	62.45	60.14	61.21
IOG (Fan et al., 2019)	82.85	77.38	80.02	73.24	69.63	71.35	76.06	70.71	73.25	**85.25**	78.51	81.69
LOTN (Wu et al., 2020b)	84.00	80.52	82.21	76.61	70.29	73.29	86.57	80.89	83.62	77.08	67.62	72.02
ONG (Veyseh et al., 2020)	83.23	81.46	82.33	76.63	81.14	78.81	87.72	84.38	86.01	73.87	77.78	75.77
Dual-MRC (Mao et al., 2021)	**89.79**	78.43	83.73	77.19	71.98	74.50	86.07	80.77	83.33	78.21	**81.66**	79.90
PER (Dai et al., 2022)	86.43	80.39	83.30	81.50	75.05	78.14	90.00	84.00	86.90	80.68	70.72	75.38
ARGCN (Jiang et al., 2021)	87.32	83.59	85.42	78.81	77.69	78.24	88.49	84.95	86.69	75.83	76.90	76.36
TSMSA (Feng et al., 2021)	-	-	86.37	-	-	81.64	-	-	89.20	-	-	82.18
MRC-MVT (Zhang et al., 2021b)	86.31	**89.42**	87.83	82.04	81.54	81.79	90.60	88.19	89.38	79.59	81.12	80.84
MGCR (ours)	88.65	89.36	**89.01**†	**84.29**	83.37	**83.80**†	**91.31**	**91.74**	**91.51**†	83.76	81.25	**82.47**†

Table 3: Main results (%) including recall, precision and F1-score. The best results are in bold and second-best results are underlined. Results of all comparison methods were copied from the original papers. The marker † represents that MGCR outperforms other methods significantly ($p < 0.01$).

Methods	14res			15res			16res			14lap		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MGCR	88.65	**89.36**	89.01	**84.29**	83.37	**83.80**	91.31	91.74	91.51	83.76	**81.25**	82.47
w/o Pre-trained Sentiment Classifier	87.69	89.03	88.35	82.79	82.89	82.77	90.67	90.60	90.63	**84.18**	79.19	81.05
w/o Filtering Noisy Unlabeled Sentences	**88.84**	88.00	88.41	80.13	**85.39**	82.62	89.59	91.68	90.62	82.84	78.83	80.77
w/o Filtering Noisy Unlabeled Words	87.29	88.12	87.70	80.10	85.33	82.66	91.02	91.30	91.16	81.99	80.19	81.07
w/o Consistency Regularization (Labeled Data Only)	87.34	87.05	87.19	82.42	81.81	82.11	87.19	88.38	87.76	81.70	77.89	79.70

Table 4: Ablation study results (%) when removing different components from MGCR method.

connections between the words for TOWE.

Dual-MRC (Mao et al., 2021) used BERT as the encoder and transforms the TOWE task into a question answering (QA) problem to solve.

PER (Dai et al., 2022) proposes a padding-enhanced reinforcement learning model on both sequential structure and syntactic structure to extract opinion words for opinion targets.

ARGCN (Jiang et al., 2021) proposes a directed syntactic dependency graph and a attention-based relational graph convolutional neural network to exploit syntactic information for TOWE.

TSMSA (Feng et al., 2021) design a target-specified sequence labeling with multi-head self-attention based on transformer architecture.

MRC-MVT (Zhang et al., 2021b) leverages a machine reading comprehension model trained with a multiview paradigm to extract target-oriented opinion words.

4.4 Main Results and Discussion

Table 3 shows main results of different methods on four benchmarks. These results demonstrate that MGCR achieves the best F1-score on all datasets. Concretely, MGCR has the following advantages:

- MGCR significantly outperforms other methods. In these methods, rule-based methods

(Distance-rule and Dependency-rule) achieve the worst performance, since they lack robustness and only cover a small number of cases. By contrast, neural methods obtain obvious improvements by introducing the external knowledge (e.g., LOTN and ARGCN) or other solution paradigm (e.g., Dual-MRC and MRC-MVT). Among the neural methods, MRC-MVT achieves very competitive results on all datasets with using machine reading framework. Nevertheless, our MGCR still outperforms it by 1.18%, 2.01%, 2.13%, and 1.63% in F1-score respectively on 14res, 15res, 16res and 14lap datasets. These comparisons demonstrate the great superiority of MGCR in exploiting unlabeled data to reduce distribution shift for TOWE, thereby successfully boosting the extraction performance.

- With few labeled TOWE data, MGCR can also bring great performance gain. As illustrated in Table 1, the labeled training sentences of the 14res dataset is about twice the amount of the 15res dataset, while MGCR outperforms MRC-MVT by 1.18% and 2.01% in F1-score respectively on 14res and 15res datasets. The more improvement on 15res proves that MGCR is promising because it

τ \ T	0.5			0.7			0.9		
	P	R	F1	P	R	F1	P	R	F1
0.5	89.55	86.53	88.02	88.60	88.28	88.44	87.88	89.29	88.55
0.7	89.58	87.63	88.60	87.80	88.77	88.28	**88.65**	**89.36**	**89.01**
0.9	88.64	87.83	88.23	87.65	89.28	88.45	87.73	88.35	88.03

Table 5: Results (%) of combinations of sentence-level threshold and word-level threshold on 14res. T and τ represent sentence-level threshold and word-level threshold respectively.

τ \ T	0.5			0.7			0.9		
	P	R	F1	P	R	F1	P	R	F1
0.5	81.31	82.62	81.94	81.17	83.57	82.33	82.16	84.17	83.12
0.7	80.13	85.39	82.62	81.84	84.04	82.92	**84.29**	**83.37**	**83.80**
0.9	81.38	83.43	82.38	81.41	84.38	82.81	81.35	84.24	82.73

Table 6: Results (%) of combinations of sentence-level threshold and word-level threshold on 15res. T and τ represent sentence-level threshold and word-level threshold respectively.

perform well even with few labeled sentences.

4.5 Ablation Study

To evaluate the effectiveness of each component of MGCR, we conduct an ablation study. As shown in Table 4, we observe different levels of performance degradation on the four TOWE datasets when removing the components from MGCR. Specifically, when removing the pre-trained sentiment classifier, the F1 score of MGCR drops from 0.5% to 1.5% on all datasets, indicating that using sentiment-attention scores to emphasize latent opinion words helps better filter the noisy unlabeled sentences. After removing either of the coarse-grained and fine-grained filtering processes, the performance declines on all datasets, demonstrating that these two confidence-based thresholding mechanisms alleviate the issue of confirmation bias caused by noisy training signals in consistency regularization. In addition, we find that removing the consistency regularization (i.e., supervised learning only) worsens the performance most.

4.6 Results of Different Thresholds Combinations

We present the results of different combinations of sentence-level threshold and word-level thresholds on four TOWE datasets. The range of each threshold is from 0.5 to 0.9. The detailed results for the 14res, 15res, 16res, and 14lap dataset are listed in Tables 5, 6, 7 and 8, respectively. An interesting finding is that using a high T with a low τ is the best strategy on most datasets, which means
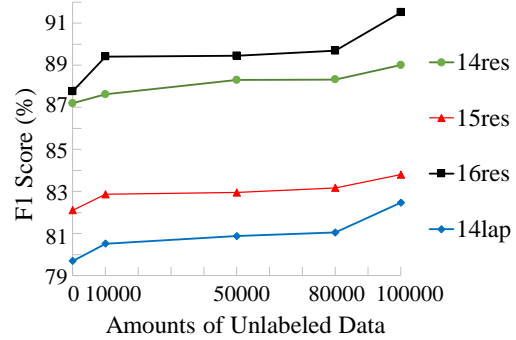

Figure 4: F1-score (%) on four TOWE datasets with varying amounts of unlabeled data.

τ \ T	0.5			0.7			0.9		
	P	R	F1	P	R	F1	P	R	F1
0.5	90.44	90.66	90.55	**91.31**	**91.74**	**91.51**	91.09	90.79	90.94
0.7	90.21	90.72	90.46	89.85	91.68	90.76	90.73	90.72	90.72
0.9	89.45	90.35	89.88	90.16	91.36	90.75	90.90	91.36	91.13

Table 7: Results (%) of combinations of sentence-level threshold and word-level threshold on 16res. T and τ represent the sentence-level threshold and word-level threshold respectively.

τ \ T	0.5			0.7			0.9		
	P	R	F1	P	R	F1	P	R	F1
0.5	83.29	79.65	81.42	84.85	78.07	81.32	84.43	77.51	80.82
0.7	83.78	79.48	81.56	84.54	79.36	81.86	83.29	78.48	80.78
0.9	82.43	78.83	80.77	84.21	78.82	81.36	**83.76**	**81.25**	**82.47**

Table 8: Results (%) of combinations of sentence-level threshold and word-level threshold on 14lap. T and τ represent the sentence-level threshold and word-level threshold respectively.

we should filter noisy unlabeled sentences more strictly than filtering noisy unlabeled words.

4.7 Effect of Amounts of Unlabeled Data

We conduct experiments by varying the amounts of unlabeled data to investigate the effect of different data amounts. The results are shown in Figure 4. Compared with supervised training only, even a few unlabeled data can improve the performance on all datasets, which validates the effectiveness of utilizing semi-supervised consistency regularization. Besides, the generalization performance is better with more unlabeled data used, because more data effectively reduce the risk of domain shift between test data and training data. This trend demonstrates the potential of MGCR, which are likely to be further improved with even more unlabeled sentences and computing resources.

Methods	NULL	Under-extracted	Over-extracted	Others	Total
MGCR	**2**	**9**	**24**	8	**43**
MGCR **w/o** Pre-trained Sentiment Classifier	3	12	29	11	54
MGCR **w/o** Filtering Noisy Unlabeled Sentences	5	**9**	29	**7**	55
MGCR **w/o** Filtering Noisy Unlabeled Words	4	14	38	11	67
MGCR **w/o** Consistency Regularization (Labeled Data Only)	8	11	44	15	70

Table 9: Statistics of different error types of our MGCR method and different ablation versions on the 16res dataset.

4.8 Error Analysis

We count all the error instances on the 16res dataset to analyze the distribution of different error types. We categorize errors into four types: 'NULL': no opinion word is extracted, 'Under-extracted': only parts of opinion words are extracted from the input sentences, 'Over-extracted': the model extracts redundant opinion words, and 'Others': wrong extraction and other error types. As shown in Table 9, training with only labeled data suffers from the severe overfitting problem, which leads to many 'Over-extracted' errors. For instance, in the sentence "*open & cool place with the best pizza and coffee.*", the overfitted TOWE model extracts both "*cool*" and "*best*" as opinion words for the opinion target "*coffee*", actually the word "*cool*" is the opinion word for "*place*". Our MGCR reduces the number of 'Over-extracted' errors from 44 to 24, indicating that introducing high-quality unlabeled data improves the generalization performance of the TOWE model. It is worth noting that MGCR also reduces 40% of total errors compared with training with only labeled data.

5 Related Work

5.1 TOWE

Aspect-based Sentiment Analysis (ABSA) contains a set of various subtasks (Pontiki et al., 2014; Liu et al., 2015; Tang et al., 2016; Wang et al., 2016; Ning et al., 2018; Zhao et al., 2020; Wu et al., 2020a). Fan et al. (2019) first propose TOWE as a new subtask to expand the ABSA research, which aims to extract opinion words for a given opinion target from a sentence. They employ several LSTM networks and propose a target-fused neural sequence labeling model that achieves promising results on TOWE. Following the idea, many works design advanced multi-head self-attention or multi-view deep attention mechanism to generate target-specific context representation (Feng et al., 2021; Zhang et al., 2021b). On this basis, some works incorporate syntactic knowledge (Feng et al., 2021;

Zhang et al., 2021b) or sentiment knowledge (Wu et al., 2020b) to further improve the performance of TOWE. There are also some other works that transform TOWE into a question answering problem and adopt the framework of machine reading comprehension (MRC) to solve TOWE (Mao et al., 2021; Zhang et al., 2021b), achieving very competitive results. Different from these works, we argue that insufficient labeled data greatly increase the risk of distribution shift for the TOWE task, and thus propose making use of massive unlabeled raw text to reduce the shift risk.

5.2 Semi-supervised Learning

Low-resource learning including semi-supervised learning (Sohn et al., 2020; Zhang et al., 2021a; Xie et al., 2020; Wang et al., 2022; Wang et al.; Oliver et al., 2018) has proven effective in the natural language processing (Peters et al., 2017; He et al.; Cheng et al., 2019; Gururangan et al., 2019; Izmailov et al., 2020; Sintayehu and Lehal, 2021; Clark et al., 2018; Ruder and Plank, 2018; Liu et al., 2022; Lu et al., 2022; Yang et al., 2021), especially ABSA tasks (Marcacini et al., 2018; Xu and Tan, 2019; Augenstein et al., 2018; Cheng et al., 2019; Li et al., 2020). Consistency regularization (Sajjadi et al., 2016) is a very popular semi-supervised learning technique, which has been widely applied in various tasks (Berthelot et al., 2019b,a; Sohn et al., 2020; Xie et al., 2020; Chen et al., 2020). Its core idea is to improve the robustness of the model by minimizing the discrepancy between data and its perturbation. Miao et al. (2020) first applied consistency regularization to aspect sentiment classification task by interpolating embeddings of input sentences. By contrast, we are the first to successfully apply semi-supervised learning consistency regularization to TOWE.

6 Conclusion

The TOWE task suffers from the risk of distribution shift which arises from scarce labeled data. In

this paper, we propose the novel MGCR method to increase the exposure of the model to varying distribution shifts by exploiting unlabeled data, and naturally the risk can be reduced. In the MGCR method, two different-grained filters, i.e., sentence-level (coarse-grained) and word-level (fine-grained) confidence-based thresholding, are designed to filter noisy sentences and words for the high-quality exposure. To further underline the possible opinion words during learning, we employ a pre-trained sentiment classifier and incorporate sentiment-attention scores into the sentence-level filter. Experimental results indicate that our MGCR method significantly outperforms all other TOWE methods and achieves state-of-the-art performance on four TOWE datasets. The in-depth analysis demonstrate the effectiveness of each component in MGCR.

7 Acknowledgements

References

Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. 2018. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1896–1906.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. 2019a. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019b. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157.

Xingyi Cheng, Weidi Xu, Taifeng Wang, Wei Chu, Weipeng Huang, Kunlong Chen, and Junfeng Hu. 2019. Variational semi-supervised aspect-term sentiment analysis via transformer. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 961–969.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925.

Yaqing Dai, Pengfei Wang, and Xiaofei Zhu. 2022. Reasoning over multiplex heterogeneous graph for target-oriented opinion words extraction. *Knowledge-Based Systems*, 236:107723.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Zhifang Fan, Zhen Wu, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2509–2518.

Yuhao Feng, Yanghui Rao, Yuyao Tang, Ninghua Wang, and He Liu. 2021. Target-specified sequence labeling with multi-head self-attention for target-oriented opinion words extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1805–1815.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A Smith. 2019. Variational pretraining for semi-supervised text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5880–5894.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. Adaptive semi-supervised learning for cross-domain sentiment classification.

Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. 2020. Semi-supervised learning with normalizing flows. In *International Conference on Machine Learning*, pages 4615–4630. PMLR.

Junfeng Jiang, An Wang, and Akiko Aizawa. 2021. Attention-based relational graph convolutional network for target-oriented opinion words extraction.

In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1986–1997.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *International Conference on Machine Learning*.

Ning Li, Chi-Yin Chow, and Jia-Dong Zhang. 2020. Seml: A semi-supervised multi-task learning framework for aspect-based sentiment analysis. *IEEE Access*, 8:189287–189297.

Ao Liu, An Wang, and Naoaki Okazaki. 2022. Semi-supervised formality style transfer with consistency training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4689–4701.

Pengfei Liu, Shafiq Joty, and Helen 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.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Jinghui Lu, Linyi Yang, Brian Namee, and Yue Zhang. 2022. A rationale-centric framework for human-in-the-loop machine learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6986–6996.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13543–13551.

Ricardo Marcondes Marcacini, Rafael Geraldeli Rossi, Ivone Penque Matsuno, and Solange Oliveira Rezende. 2018. Cross-domain aspect extraction for sentiment analysis: A transductive learning approach. *Decision Support Systems*, 114:70–80.

Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. Snippext: Semi-supervised opinion mining with augmented data. In *Proceedings of The Web Conference 2020*, pages 617–628.

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

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197.

Tianhao Ning, Zhen Wu, Xin-Yu Dai, Jiajun Huang, Shujian Huang, and Jiajun Chen. 2018. Improving aspect identification with reviews segmentation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 416–428. Springer.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31.

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.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *SemEval@COLING 2014*, pages 27–35.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in neural information processing systems*, pages 1163–1171.

H Sintayehu and GS Lehal. 2021. Named entity recognition: a semi-supervised learning approach. *International Journal of Information Technology*, 13(4):1659–1665.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Amir Pouran Ben Veyseh, Nasim Nouri, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020. Introducing syntactic structures into target opinion word extraction with deep learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8947–8956.

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.

Yidong Wang, Hao Chen, Yue Fan, SUN Wang, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, et al. Usb: A unified semi-supervised learning benchmark.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. 2022. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020a. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585.

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

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.

Weidi Xu and Ying Tan. 2019. Semi-supervised target-oriented sentiment classification. *Neurocomputing*, 337:120–128.

Linyi Yang, Jiazheng Li, Pádraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 306–316.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2021a. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34.

Jingyuan Zhang, Zequn Zhang, Zhi Guo, Li Jin, Kang Liu, and Qing Liu. 2021b. Enhancement of target-oriented opinion words extraction with multiview-trained machine reading comprehension model. *Computational Intelligence and Neuroscience*, 2021.

Fei Zhao, Zhen Wu, and Xinyu Dai. 2020. Attention transfer network for aspect-level sentiment classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 811–821.