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

COLING2022

code: https://github.com/TOWESSL/TOWESSL

Reported by Xiaoke Li

Introduction

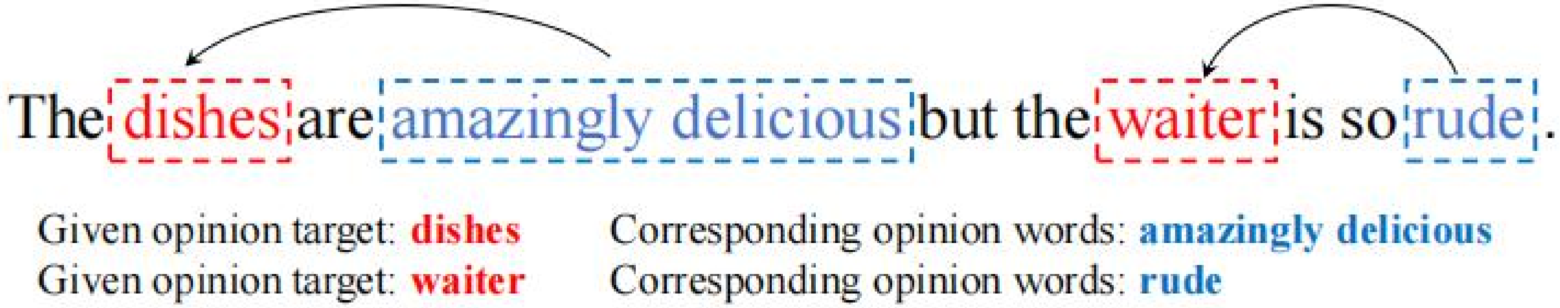


Figure 1: Example of TOWE. Words in red are opinion targets and words in blue are corresponding opinion words. TOWE extracts corresponding opinion words when given opinion targets.

Method

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.

Method

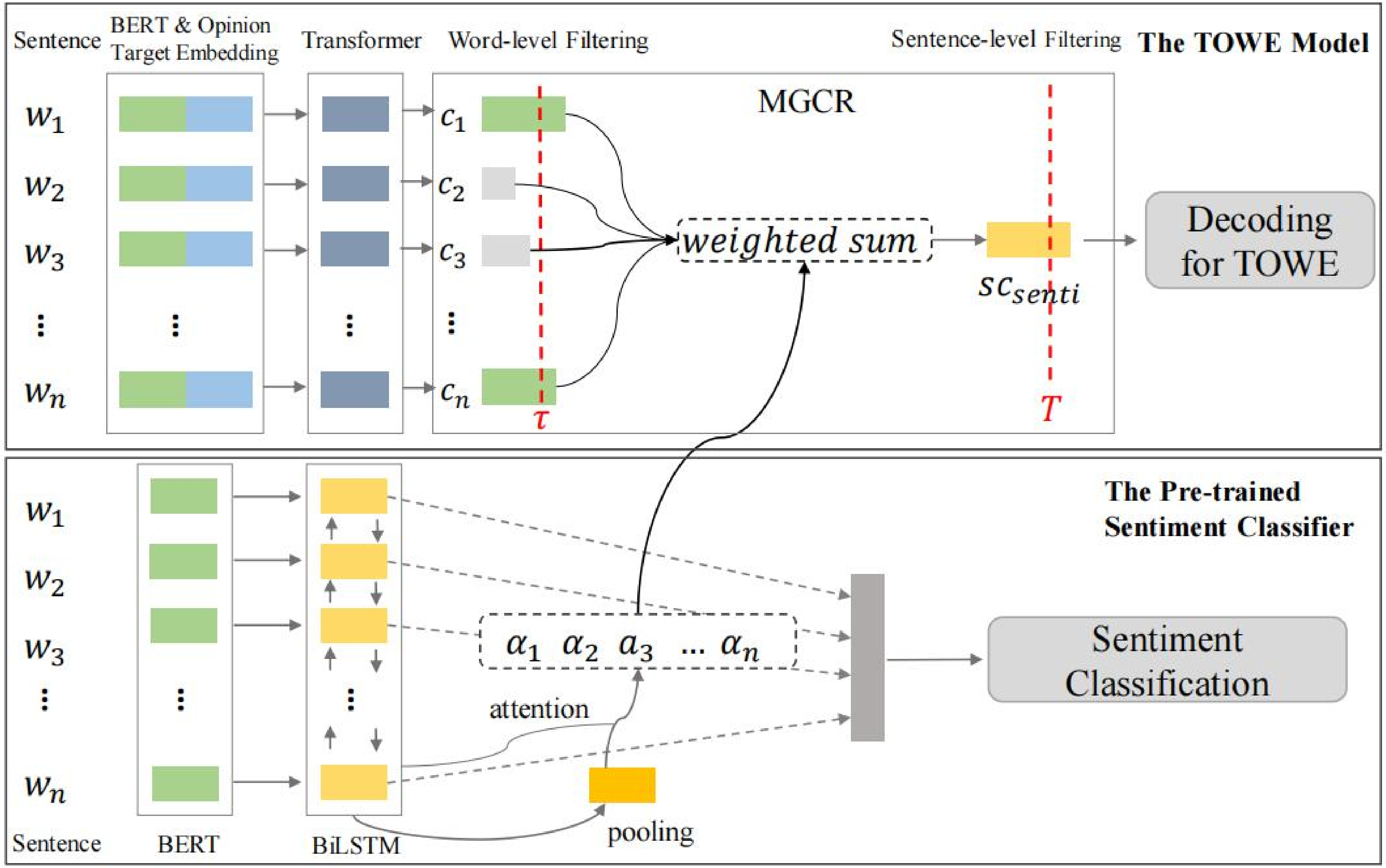
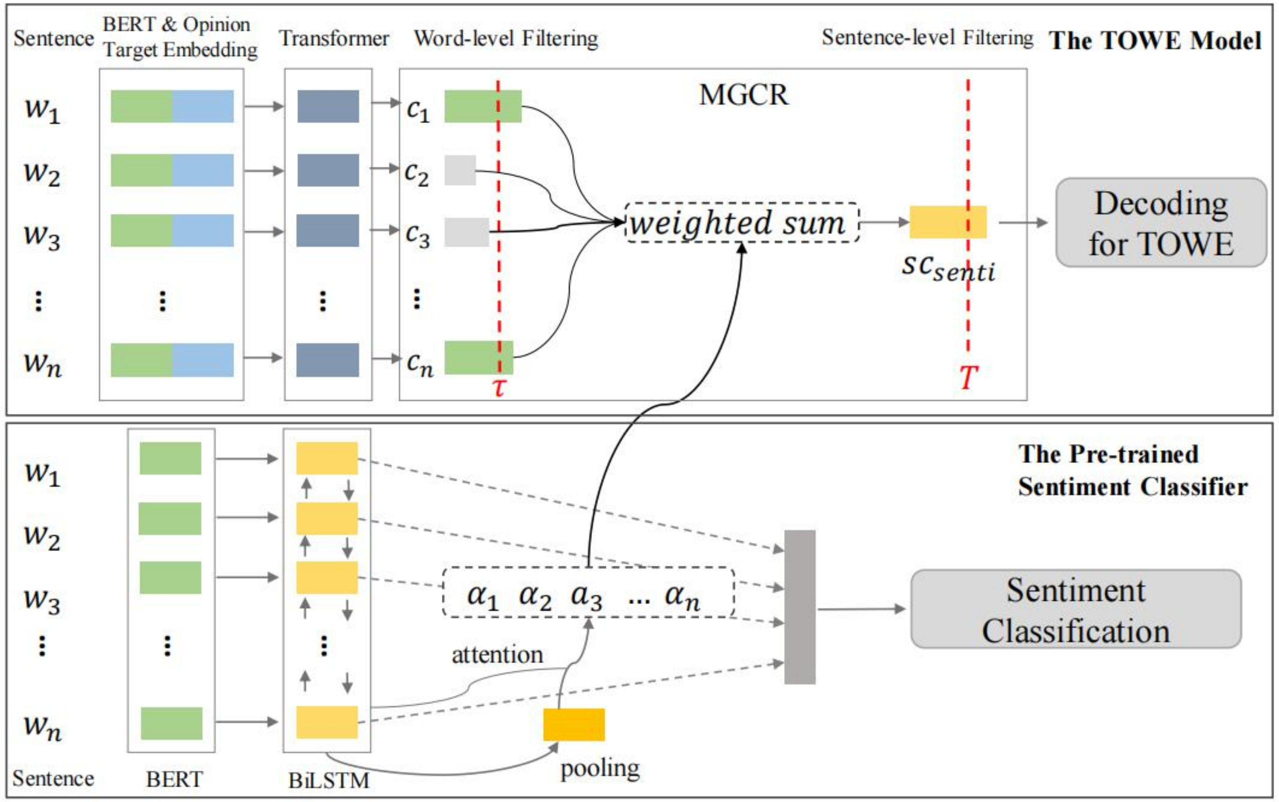


Figure 3: Overview of the architecture of multi-grained consistency regularization. For simplicity, we mark the confidence of i-th word as c_i. Note that the input sentence of the TOWE model is the same as the input sentence of the pre-trained sentiment classifier.

Method



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

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

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

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

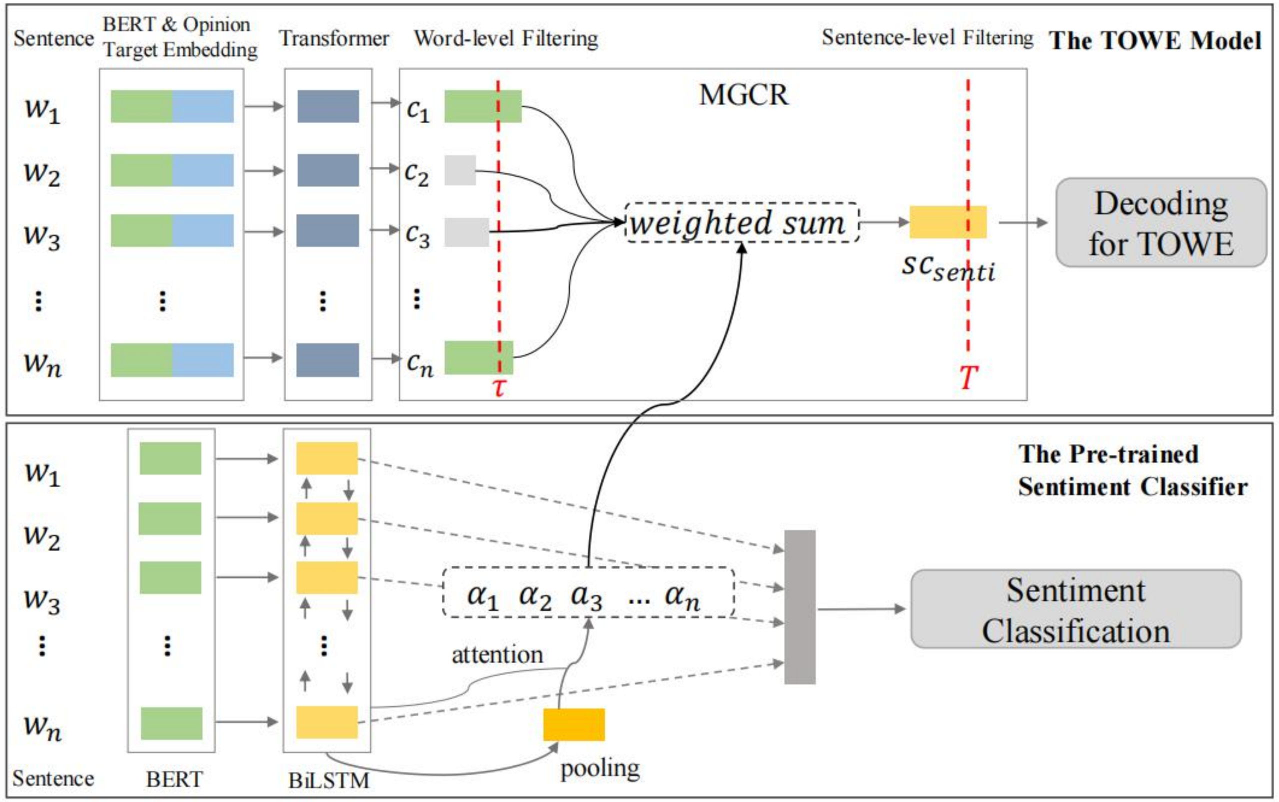
Then the context representation $\mathbf{h}_i^{pt}$ of the word w_i is fed to a linear layer and a softmax layer to predict the corresponding label.

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

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.

$$\widetilde{\mathbf{h}}_i = [\mathbf{h}_i; \mathbf{e}_i].$$

Method



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

$$z_{avg} = \frac{1}{n}\sum_{i=1}^{n} z_i,$$

$$f(z_i, z_{avg}) = z_i \cdot \mathbf{W} \cdot z_{avg} + \mathbf{b}, \qquad (7)$$

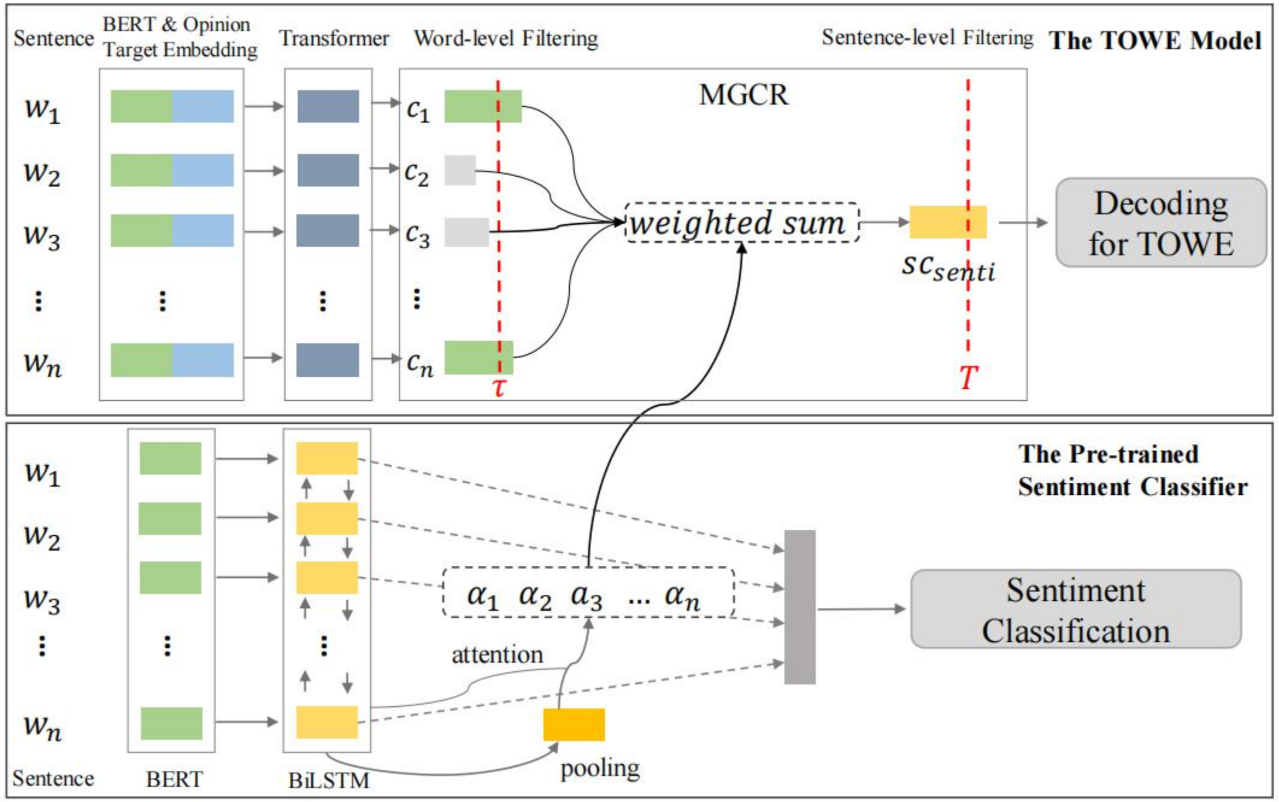
$$\alpha_i = \frac{e^{f(z_i, z_{avg})}}{\sum_{j=1}^{n} e^{f(z_j, z_{avg})}},$$

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

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

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

Method



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

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

Experiments

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.

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.

Experiments

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 ($p < 0.01$) .

Experiments

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.

Experiments

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.

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.

Experiments

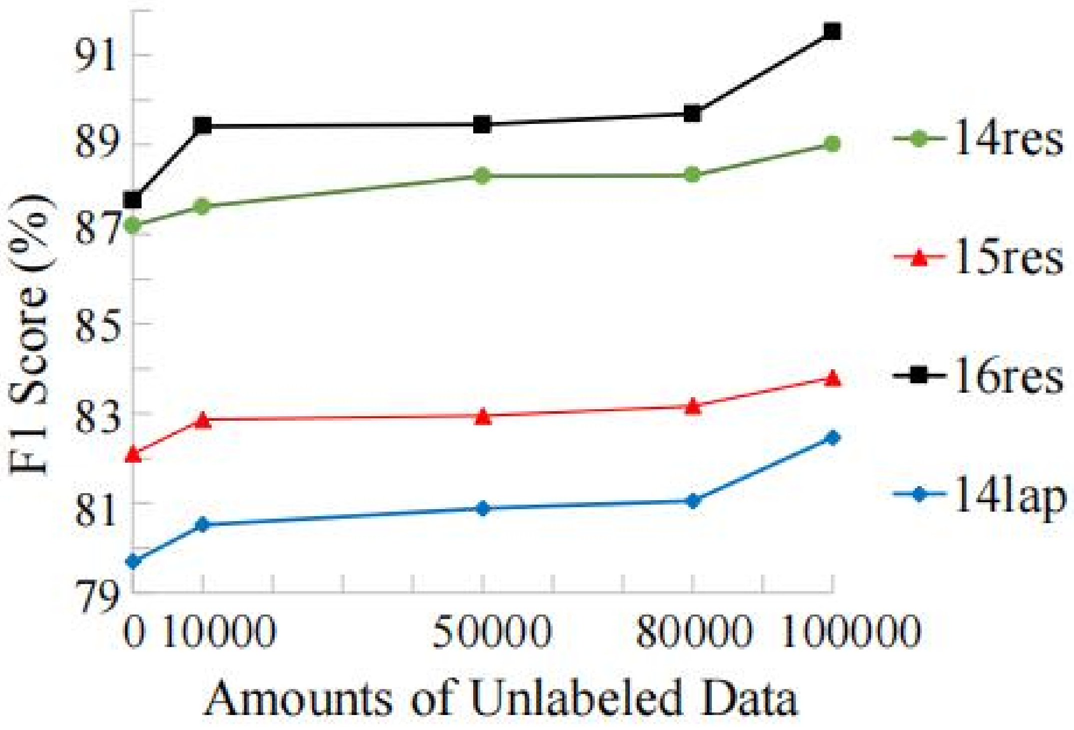


Figure 4: F1-score (%) on four TOWE datasets with varying amounts of unlabeled data.

Experiments

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.

Thanks