Attention-based Relational Graph Convolutional Network for Target-Oriented Opinion Words Extraction

Junfeng Jiang[*]
The University of Tokyo
Tokyo, Japan
jiangjf@is.s.u-tokyo.ac.jp

An Wang[*]
Tokyo Institute of Technology
Tokyo, Japan
wang.a.aa@m.titech.ac.jp

Akiko Aizawa[†]
National Institute of Informatics
Tokyo, Japan
aizawa@nii.ac.jp

Abstract

Target-oriented opinion words extraction (TOWE) is a subtask of aspect-based sentiment analysis (ABSA). It aims to extract the corresponding opinion words for a given opinion target in a review sentence. Intuitively, the relation between an opinion target and an opinion word mostly relies on syntactics. In this study, we design a directed syntactic dependency graph based on a dependency tree to establish a path from the target to candidate opinions. Subsequently, we propose a novel attention-based relational graph convolutional neural network (ARGCN) to exploit syntactic information over dependency graphs. Moreover, to explicitly extract the corresponding opinion words toward the given opinion target, we effectively encode target information in our model with the target-aware representation. Empirical results demonstrate that our model significantly outperforms all of the existing models on four benchmark datasets. Extensive analysis also demonstrates the effectiveness of each component of our models. Our code is available at https://github.com/wcwowwwww/towe-eacl.

1 Introduction

Target-oriented opinion words extraction (TOWE) (Fan et al., 2019) is a subtask of aspect-based sentiment analysis (ABSA) (Hu and Liu, 2004; Pontiki et al., 2016). Given a review and an opinion target in the sentence, the objective of TOWE is to extract the corresponding opinion words describing or evaluating the opinion targets from the review. Opinion targets are the words or phrases representing features or entities toward which users express their attitudes, whereas opinion words referring to those terms are used to express attitudes or opinions explicitly.

The food is tasty and portion sizes are appropriate.
Target: food Opinion: tasty

The food is tasty and portion sizes are appropriate.
Target: portion size Opinion: appropriate

Figure 1: Examples of TOWE task. The words highlighted in orange represent the given opinion targets, whereas the words in blue represent the corresponding opinion words.

Figure 1 shows two examples of TOWE. In the review "*The food is tasty and portion sizes are appropriate .*", the terms "*food*" and "*portion sizes*" are two given opinion targets. TOWE needs to extract the word "*tasty*" as the opinion word for the opinion target "*food*" and the opinion word "*appropriate*" for the opinion target "*portion sizes*".

Therefore, the first challenge is to effectively introduce the opinion target information into our model. Fan et al. (2019) designed the IO-BiLSTM to encode the context before and after the given opinion targets separately to represent the position of the existing opinion targets. Wu et al. (2020) introduced position embeddings based on the relative distance toward opinion targets. However, both studies only introduce parts of target information (the position information of targets). In this paper, we introduce the target-aware representation to fully exploit opinion target information in a concise way, which is especially important when our models are used for real-world reviews.

Becase TOWE can be viewed as a syntactic task, a natural solution is analysing the relationship between opinion targets and opinion words by dependency parsing. Recently, owing to the great success of graph convolutional networks (GCNs) in various fields (Kipf and Welling, 2016; Chen et al., 2018; Marcheggiani et al., 2018), a few researchers have attempted to encode the syntactic

[*] These authors contributed equally to this work; the order is random.

[†] Corresponding author.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1986–1997
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

dependency information with GCNs to build a robust dependency encoder. For example, GCNs over the dependency tree have been exploited to perform semantic role labelling (Marcheggiani and Titov, 2017) and named entity recognition (Cetoli et al., 2017).In addition, several studies explore GCNs over a dependency graph to complete the ABSA task (Sun et al. (2019), Zhang et al. (2019), Liang et al. (2020), Wang et al. (2020)).

However, it is worth mentioning that TOWE is defined as a sequence labelling task, and the manner in which GCNs are applied to TOWE effectively is yet to be explored. In this study, we first construct a directed graph based on a dependency tree to be more suitable for TOWE. Subsequently, we propose ARGCN, which can enhance our model by encoding syntactic information. ARGCN can be seen as extending the Relational Graph Convolutional Networks (R-GCNs) (Schlichtkrull et al., 2018) with the distance-aware attention mechanism. ARGCN can consider the semantic relevance and syntactic relevance between words simultaneously when it propagates information. In addition, sequential information is extremely important for sequence-labelling tasks. Therefore, after using multi-layer graph convolutions to encode syntactic information, we feed the syntactic representation to a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to capture the sequential information.

Experiments on four benchmark datasets demonstrate that our base model, Target-BiLSTM which is a BiLSTM with target-aware inputs has a similar or better performance than the state-of-the-art model, although we do not introduce extra external knowledge. In addiction, our full model ARGCN further improves the performance and significantly outperforms all of the existing models on four benchmark datasets. Furthermore, extensive experiments

demonstrate the effectiveness and necessity of all components in our full model. To the best of our knowledge, it is the first work on applying GCNs to the TOWE task.

The contributions of this paper can be summarized as follows.

- We propose target-aware representation to effectively introduce opinion target information. An empirical study shows it is significant and extensible for the TOWE task.
- We exploit syntactic dependency graphs of sentences and establish the relations between opinion targets and the corresponding opinion words.
- We propose a novel attention-based relational graph convolutional network, ARGCN, an extension of R-GCNs suited to encode syntactic dependency information.
- We propose an ARGCN-based TOWE model. Experimental results show that it significantly outperforms the state-of-the-art model on all datasets of the TOWE task.

2 Related Work

As subtasks of ABSA, a series of early studies focused on opinion targets extraction, including unsupervised/semi-supervised methods (Qiu et al., 2011; Liu et al., 2012, 2013) and supervised methods (Jakob and Gurevych, 2010; Li et al., 2010). Some recent studies extracted opinion targets and opinion words jointly in a uniform framework and achieved promising results (Wang et al., 2016; Li and Lam, 2017). However, they did not extract the corresponding relation between opinion targets and opinion words. Moreover, studies on extracting paired opinion relations are rare (Hu and Liu, 2004; Zhuang et al., 2006). Because it is important for downstream sentiment analysis and real-world applications, Fan et al. (2019) proposed a new subtask of ABSA, target-oriented word extraction, aiming to extract the corresponding opinion words for the given opinion targets in a review. They released four benchmark datasets for evaluation, designed a target-fused model, and achieved excellent performance. Wu et al. (2020) adopted transfer learning to transfer latent opinion information from the sentiment analysis model to the TOWE model. In this study, we also focus on the TOWE task.

Since Kipf and Welling (2016) proposed their GCN with some simplifications on ChebNet (Defferrard et al., 2016), a variety of graph convolutional networks appeared (Veličković et al., 2018; Schlichtkrull et al., 2018; Busbridge et al., 2019) and achieved great success in many fields, including computer vision (Chen et al., 2018; Garcia and Estrach, 2018; Wang et al., 2019), natural language processing (Marcheggiani and Titov, 2017; Marcheggiani et al., 2018; Yao et al., 2019; Wang et al., 2020) and even in chemistry (De Cao and Kipf, 2018). One of the reasons why GCNs work well in several fields is that they can naturally process the graph-structured data to greatly exploit

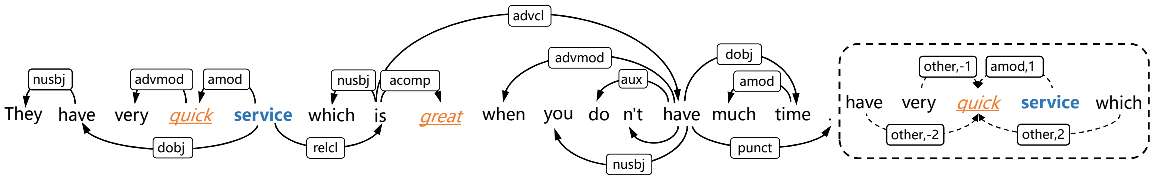

Figure 2: An example of the syntactic dependency graph based on dependency tree generated by spaCy[*]dependency parser. “service” is the given opinion target. “quick” and “great” are two corresponding opinion words. In the left figure, we show the reshaped dependency graph before adding extra edges for those who do not have dependency relation but have close distance (closer than a threshold D). In the right figure, we can see that each edge has two features: dependency relation and distance. And we show all edges coming towards the word ”quick” in the final graph as an example.

the latent information behind the graph structure. Therefore, they are proven to be efficient especially with only a small amount of data (Kipf and Welling, 2016; Garcia and Estrach, 2018).

Recently, a few studies have tried applying GCNs over the dependency graph to complete some ABSA tasks. Sun et al. (2019) proposed the CDT to perform GCN over a dependency tree together with contextual representations extracted by BiLSTM. Liang et al. (2020) introduced dependency relational embedding to GCN (Kipf and Welling, 2016) to complete ABSA with their DREGCN. Specially, the R-GAT-ABSA (Wang et al., 2020) is a newly proposed architecture for the ABSA task. It focuses on the GAT (Veličković et al., 2018) and extends it by introducing relational embedding for calculating relational attention.

3 Our Methods

3.1 Task Formalization

TOWE aims to extract corresponding opinion words based on the given opinion targets. Formally, we have a review sentence $s = \{w_1, w_2, ..., w_n\}$ containing n words. Then, we adopt the BIO tagging scheme (Ramshaw and Marcus, 1999) as Fan et al. (2019) do in their paper. For each words in the sentence, we tag them as $y_i \in \{B, I, O\}$ (B: Beginning, I: Inside, O: Others). For example, the sentence in figure 1 is tagged as “*The/O food/O is/O [tasty/B] and/O portion/O sizes/O are/O appropriate/O ./O*”, indicating the opinion word “*tasty*” for target “*food*”.

3.2 Target-Aware Representation

As described above, we should extract the corresponding opinion words based on the given opinion targets. Therefore, our model should be aware of which words are the opinion targets and identify the corresponding opinion words. All previous studies only encode the position information for targets. In contrast, we directly introduce category embeddings with respect to the target tag of words to fully introduce target information in the TOWE model. Figure 3 shows an overview of our model.

We denote the category embedding table as $\mathbf{T}^t \in \mathbb{R}^{3 \times d^t}$, where d^t is the dimension of the category embedding. Next, we can obtain the target embedding of each word and form a target embedding matrix of a sentence as $\mathbf{E}^t = [\mathbf{e}_1^t; \mathbf{e}_2^t; \cdots ; \mathbf{e}_n^t]$.

To retain the target information clearly when feeding to the next module, we concatenate it together with the word representation.

$$\mathbf{e}_i = [\mathbf{e}_i^w, \mathbf{e}_i^t] \tag{1}$$

where e_i^w is word representation of word i, [,] represents the concatenation operation.

Thus, our model can understand which words are opinion targets. The target embedding table is jointly optimized during training so that our model can learn the proper target embeddings specifically for the TOWE task.

For simplicity, we denote our target-aware representation as $\mathbf{E} = [\mathbf{e}_1; \mathbf{e}_2; \cdots ; \mathbf{e}_n]$ and then feed it to the following modules.

3.3 Syntactic Dependency Graph

In this section, we provide a detailed description of our method of building a suitable syntactic dependency graph for the TOWE task. For a given sentence $s = \{w_1, w_2, \cdots , w_n\}$, after dependency parsing, we obtain a dependency tree. Figure 2 is the original dependency tree of the sentence “The food is tasty and portion sizes are appropriate .”. Next, we add some edges whose relative distance in the sentence is smaller than a given threshold D.

We formally define the directed graph as $G = \{V, E, \mathcal{R}, \mathcal{P}\}$, where $V = \{v_i\}_{i=1}^n$ is the set of nodes, which are words in a sentence, $E =$

[*] https://spacy.io/

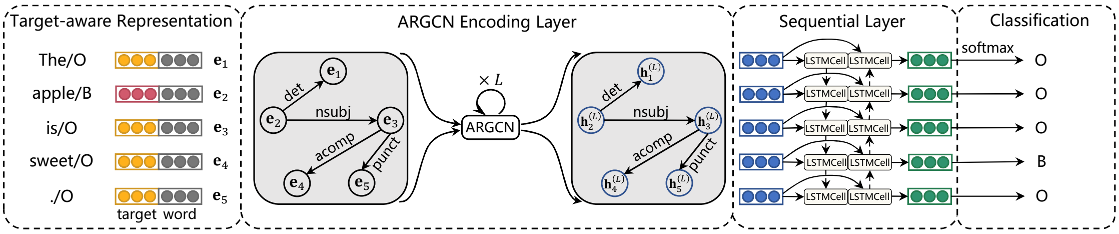

Figure 3: Overview of ARGCN. We generate the target-aware representation as the input node representation. Then, L-layers of ARGCN are applied over our syntactic dependency graph. After encoding, we capture sequential information with BiLSTM. Finally, we perform prediction with softmax classifier. Because of space limits, we omit other edges except for those who have dependency relations.

$\{e_{ij}\}_{i,j=1}^n$ is the set of edges, and $\mathcal{R} = \{r_{ij}\}_{i,j=1}^n$ is the set of edge relational types, where r_{ij} is the corresponding dependency relation from v_i to v_j. If there is not any dependency relation between v_i and v_j whose relative distance is smaller than D, we add an edge between them and set a special edge type for it, such as *other*. $\mathcal{P} = \{p_{ij}\}_{i,j=1}^n$ represents the set of relative positions, and p_{ij} is the relative position from v_i to v_j in the sentence. Note that e_{ij} indicates v_i is the neighbour of v_j.

To ensure the target information can correctly propagate to the latent opinion words, we redirect some specific dependency relations linking to the target words. Regarding dependency trees, when the edge type is *nsubj* or *dobj*, the direction of the edge is from predicate to subject or object. Hence, the information of the subject or object cannot flow through the predicate. Thus, we reverse the dependency edge when it links target words and its type is *nsubj* (nominal subject) or *dobj* (direct object). In addition, we remove the *root* relation because it is a self-loop, which is not helpful for our model.

3.4 Attention-Based Relational Graph Convolutional Network (ARGCN)

To encode the well-designed syntactic dependency graph, we begin from R-GCNs (Schlichtkrull et al., 2018) and extend it with a distance-aware attention mechanism. In this paper, we propose an attention-based relational graph convolutional network (ARGCN). The main purpose of our model is to consider semantic and syntactic relevance between words simultaneously.

R-GCNs (Schlichtkrull et al., 2018) updated the hidden states of nodes by aggregating node representations of their neighbours according to the edge type of their connections,

$$\mathbf{h}_i' = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\mathbf{x}_{ij,r} + \mathbf{W}_1\mathbf{h}_i\right) \tag{2}$$

$$\mathbf{x}_{ij,r} = \frac{1}{c_{i,r}}\mathbf{W}_r\mathbf{h}_j \tag{3}$$

where $\mathcal{R}$ denotes the set of relations, $\mathbf{h}_i$ is the input representation of node v_i, $\mathbf{h}_i'$ is the output representation of node v_i, $\mathcal{N}_i^r$ is the set of neighbours of v_i under relation $r \in \mathcal{R}$, $\mathbf{W}_r$ and $\mathbf{W}_1$ are trainable parameters, and $c_{i,r}$ is a problem-specific normalization constant, which is usually assigned as the number of neighbours of v_i under relation r. Moreover, σ is an element-wise activation function.

In Equation (5), each relation r corresponds to a relation-specific matrix $\mathbf{W}_r$. To reduce the parameter number, we perform a basis decomposition (Schlichtkrull et al., 2018). In particular, we set the number of bases as one:

$$\mathbf{W}_r = b_r\mathbf{W}_0 \tag{4}$$

where b_r is the coefficient depending on r. In this way, every $\mathbf{W}_r$ shares $\mathbf{W}_0$ as the basis, thereby the number of parameters is greatly reduced. On the other hand, b_r denotes the influence with respect to relation types.

In ARGCN, we introduce a distance-aware attention mechanism to enhance the power of RGCN:

$$\mathbf{x}_{ij,r} = \alpha_{ij,r}\mathbf{W}_0\mathbf{h}_j \tag{5}$$

$$\alpha_{ij,r} = \sigma(\mathbf{c}^T[b_r, \beta_{ij}]) \tag{6}$$

where β_{ij} is the attention coefficient between v_i and v_j, and $\mathbf{c}$ is a trainable vector, which can adjust the influence of the relation and the attention coefficient. σ is an activation function, and we choose to use ReLU in ARGCN layers.

We assume that the attention coefficients between two nodes are based on the features of nodes and the relative position in the sentence. First, we obtain query and key by project node features $\mathbf{h}_i$ and $\mathbf{h}_j$ by multiplying the same projection matrix $\mathbf{W}_1$. Next, we get relative positional encoding $\mathbf{p}$ by a sinusoid encoding matrix as in Dai et al. (2019).

Then, we use a shared attention mechanism to perform attention on the query, key, and relative positional encoding:

$$o_{ij} = \sigma(\mathbf{a}^T[\mathbf{W}_1\mathbf{h}_j, \mathbf{W}_1\mathbf{h}_i, \mathbf{p}]) \tag{7}$$

where **a** is a trainable vector mapping the concatenated representation to a scalar.

Finally, we normalize o_{ij} across all neighbours of v_i using the softmax function:

$$\beta_{ij} = \frac{exp(o_{ij})}{\sum\limits_{k \in \mathcal{N}_i} exp(o_{ik})} \tag{8}$$

where β_{ij} indicates the importance of v_i toward v_j with respect to the node representations and the relative position.

In addition, extending our mechanism to employ multi-head attention helps to stabilize the learning process and enhances the performance. Specifically, K independent attention mechanisms execute the transformation of Equation (5).

$$\mathbf{x}_{ij,r}^k = \alpha_{ij,r}^k \mathbf{W}_0^k \mathbf{h}_j \tag{9}$$

where $\mathbf{W}_0^k \in \mathbb{R}^{d_j \times d_k}$, d_j is the input dimension, and d_k is the dimension of each head. Then, the output of the multi-head attention mechanism is

$$\mathbf{x}_{ij,r} = \mathbf{W}_d[\mathbf{x}_{ij,r}^1, \mathbf{x}_{ij,r}^2, ..., \mathbf{x}_{ij,r}^K] \tag{10}$$

where $\mathbf{W}_d \in \mathbb{R}^{Kd_k \times d_{j+1}}$.

Unlike Vaswani et al. (2017), who chose to use d_{j+1}/K as a dimension of each head, we set $d_k = d_{j+1}$, which leads to slight performance gains based on preliminary experiments.

We find that aspect and opinion terms often have direct or indirect relations in the graph based on the syntactic dependency tree. For example, Figure 2 shows that the relation between "service" and "quick" is direct whereas that between "service" and "great" is indirect. To capture these direct or indirect relations, we use L-layers of ARGCN, because L successive ARGCNs result in the propagation of information across the L-th order neighbour.

Moreover, with the deepening network layers, ARGCN tends to be over-smooth. In order to alleviate this problem, we add a residual connection on each ARGCN layer:

$$\mathbf{h}_i^{l+1} = \mathbf{h}_i^l + \mathbf{h}_i'^l \tag{11}$$

where $\mathbf{h}_i^l$ is the input of v_i in l-th layer of ARGCN, and $\mathbf{h}_i'^l$ is the output of v_i in l-th layer of ARGCN. Thus, $\mathbf{h}_i^{l+1}$ is the input of $(l+1)$-th layer of ARGCN.

3.5 Sequential Layer

The insufficiency of ARGCN is that it cannot encode the sequential information, which is extremely important for the TOWE task because it is defined as a sequence-labelling task. Intuitively, prediction relies on the prediction label of the words before and after the current word. Therefore, the performance of the model will not be satisfactory without capturing sequential information.

Consequently, we feed the syntactic representation extracted from L-layers of ARGCN to a BiLSTM to capture the sequential information:

$$\hat{\mathbf{h}}_i = \text{BiLSTM}(\mathbf{h}_i^{(L)}, \hat{\mathbf{h}}_{i-1}) \tag{12}$$

where $\hat{\mathbf{h}}_i$ is the concatenation of the forward and backward output vectors at time-step i.

Many other studies that used GCNs over the dependency graph (Marcheggiani and Titov, 2017; Sun et al., 2019; Wang et al., 2020) often applied LSTM to encode the sequential information, fed the obtained contextual representation to GCNs, and used them for predictions. We also attempted to first encode the sequential relationship by LSTM and then feed them to ARGCN to finally predict the labels of words. However, the performance was impaired. We believe the reason is that the sequential relationship is essential for sequence-labelling tasks. If we collect it before encoding the dependency information, it will be confused through aggregation, leading to poor performance.

3.6 Model Training

After collecting the sequential information, we simply mapped the representations to the output space with a fully connected layer and calculated the probability of the labels of words with the softmax function:

$$\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{W}_{fc}\hat{\mathbf{h}}_i + \mathbf{b}_{fc}) \tag{13}$$

where $\mathbf{W}_{fc}$ and $\mathbf{b}_{fc}$ are the trainable parameters of the fully connected layer.

Next, the cross-entropy loss is defined as

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

and minimized during training. Here, the opinion word tags $\{O, B, I\}$ are correspondingly numeralized as labels $\{0, 1, 2\}$, respectively, and y_i denotes the gold label.

4 Experiments

4.1 Datasets and Metrics

Following the previous works (Fan et al., 2019; Wu et al., 2020), we evaluate the models on four benchmark datasets, including **14res**, **14lap**, **15res** and **16res**. Explicitly, the datasets **14res** and **14lap** are annotated from SemEval Challenge 2014 task 4 (Pontiki et al., 2014). The **15res** and **16res** are annotated from SemEval Challenge 2015 task 12 (Pontiki et al., 2015) and SemEval Challenge 2016 task 5 (Pontiki et al., 2016) respectively. The suffixes "**res**" and "**lap**" indicate they are collected from restaurant reviews and laptop reviews, respectively.

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

Table 1: Statistics of the four benchmark datasets.

The original SemEval challenge datasets are very popular for ABSA subtasks. However, they only contain annotations of aspect terms. Therefore, Fan et al. (2019) extended the annotation to further annotate the corresponding opinion words based on the given opinion targets and ignored the cases without explicit opinion words. Detailed statistics are shown in Table 1.

For the classification task, we adopted commonly used evaluation metrics: precision, recall, and F1-score. An extraction is considered as correct only when the opinion words from the beginning to the end are all predicted exactly as the ground truth.

4.2 Experimental Settings

For ARGCN and Target-BiLSTM, we adopted 300-dimension GloVe word embeddings (Pennington et al., 2014) as our word representations. For ARGCN-bert and Target-BiLSTM-bert, we adopted the last hidden states of the pre-trained BERT (Devlin et al., 2018) as word representations and fine-tuned it jointly. Inspired by Xu et al. (2018), we fine-tuned the GloVe vectors during training to obtain a domain-specific representation. The dimension of target embedding was 3 and 100 for our base model and GCNs-based models, respectively. We implemented our models with PyTorch (Paszke et al., 2019). We introduced 10 layers of ARGCN with 128 channels, 8 attention heads and set the hidden size of BiLSTM to 128.

We used spaCy (Honnibal and Johnson, 2015) as our dependency parser. To improve the generalization of ARGCN, dropout (Hinton et al., 2012) layers were applied after the activation with the probability of 0.5. The threshold of relative distance was set to be 3. All of the parameters were optimized by Adam optimizer (Kingma and Ba, 2014). The initial learning rate was 1×10^{-3}. We randomly split 20% of the training set as the validation set to fine-tune the hyperparameters and apply early stopping. Subsequently, we tested our models and averaged the results of 5 runs.

4.3 Compared Methods

We compare our model with several methods which can be categorized into three groups.

- **Early Solutions**: Some early solutions including rule-based methods and trivial deep learning methods are assigned to the first group. Inspired by Hu and Liu (2004) and Zhuang et al. (2006), Fan et al. (2019) proposed the Distance-rule and Dependency-rule as two representative rule-based methods. Following Liu et al. (2015) and Tang et al. (2016), Fan et al. (2019) proposed LSTM/BiLSTM and the TC-BiLSTM as some trivial deep learning methods. Besides, Fan et al. (2019) combined BiLSTM and Distance-rule method to complete TOWE in a pipelined way, which is named as Pipeline in the experiments.

- **TOWE models**: IOG is the first TOWE model proposed by Fan et al. (2019). It adopts six different positional and directional LSTMs to extract the opinion words. PE-BiLSTM is the base model of the LOTN (Wu et al., 2020). They introduced target information of TOWE by position embedding and extracted opinion words with a BiLSTM. Wu et al. (2020) proposed an effective transfer learning method LOTN to identify latent opinions from the sentiment analysis model. Next, they integrated it with the PE-BiLSTM to achieve the state-of-the-art performance in TOWE.

- **AOPE model**: Aspect-opinion pair extraction (AOPE) task which aims at extracting aspects

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
SDRN+bert[‡]	91.14	76.37	83.10	84.37	65.42	73.69	83.57	70.33	76.38	91.13	80.34	85.40
Target-BiLSTM	84.00	79.34	81.58	75.35	**69.93**	72.50	76.95	71.62	74.14	87.85	80.99	84.24
ARGCN	86.67	**82.72**	**84.65**	79.45	**71.60**	**75.32**	76.57	**76.88**	**76.72**	86.16	**84.19**	85.16
Target-BiLSTM+bert	86.72	78.64	82.48	75.50	**72.84**	**74.15**	81.25	71.20	75.89	86.58	**84.76**	85.66
ARGCN+bert	87.32	**83.59**	**85.42**	75.83	**76.90**	**76.36**	78.81	**77.69**	**78.24**	88.49	**84.95**	**86.69**

Table 2: Main Experimental Results(%). Comparison between our proposed models and baselines on four benchmark datasets. P, R and F1 are *precision*, *recall* and *F1-score*, respectively. The result in bold indicates that the model outperforms all of the baselines above significantly ($p < 0.01$). The results are averaged scores of 10 runs. The results of baselines are copied from the previous work (Wu et al., 2020). Noted that the experiment results of SDRN are obtained by using their released codes to train and evaluate on TOWE datasets.

and opinion expressions in pairs, is a similar task as TOWE. SDRN, which is the state-of-the-art AOPE model proposed by Chen et al. (2020), mainly consists of an opinion entity extraction unit, a relation detection unit, and a synchronization unit. The synchronization unit could enhance the mutual benefit on the opinion entity extraction unit and a relation detection unit. As a baseline, it extracts the target and opinion. Subsequently, it collect the corresponding target-opinion pairs based on the predicted relations to complete the TOWE task.

- **Base model**: To show the effectiveness of the target-aware representation, we propose our base model, Target-BiLSTM. A BiLSTM receives the target-aware representation as the input and then predicts after a fully-conneceted layer and a softmax layer.

4.4 Results and Discussion

Table 2 shows the main experimental results of the baselines and our models on four benchmark datasets. We can observe that under the same condition of using GloVe for word representation, our base model Target-BiLSTM outperforms PE-BiLSTM with large improvements ranging from 3.31% to 6.18% on F1-score. Note that PE-BiLSTM uses position embedding. Instead, Target-BiLSTM introduces target embedding, which is the evidence of the effectiveness of our target-aware representation. Moreover, it not only performs similarly with LOTN on 14res and 14lap but also significantly outperforms LOTN on 15res and 16res, which introduces a large-scale sentiment analysis dataset for transfer learning. In contrast, our base model does not require additional resources except for the pre-trained word embeddings. Besides, our full model ARGCN outperforms Target-BiLSTM by a large margin on the four datasets. Therefore, we conclude that the syntactic information ARGCN encoded over the dependency graph is helpful to TOWE. Furthermore, ARGCN significantly outperforms LOTN, with large improvements of F1-score ranging from 1.54% to 3.43%, which proves its effectiveness on the TOWE task. With a pre-trained representation model, BERT, Target-BiLSTM achieves the state-of-the-art performance by significant margins, demonstrating the power of the pre-trained language model in this task. In addition, when we apply BERT as the representation layer for ARGCN, it achieves a further state-of-the-art performance, which demonstrates the effectiveness of capturing important syntactic information for sentiment analysis.

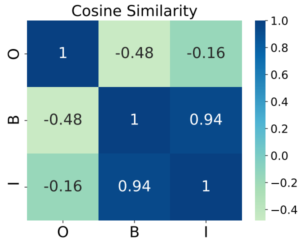

Figure 4: We train a Target-BiLSTM on 14res. After convergence, we collect the target embeddings and calculate their cosine similarities.

[‡]SDRN is an APOE model but evaluated on TOWE task.

Models	14res			14lap			15res			16res		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
R-GCN (#basis=1)	85.23	81.67	83.40	76.24	70.26	73.12	74.08	72.74	73.35	86.25	80.76	83.39
GAT	81.30	74.27	77.63	66.73	65.43	66.07	68.91	70.59	69.74	81.80	72.76	77.02
RGAT	82.99	74.37	78.44	74.38	62.96	68.19	77.67	66.33	71.55	87.18	77.71	82.18
ARGCN (original)	84.79	82.59	83.65	77.79	70.79	74.06	74.83	74.06	74.43	85.64	83.47	84.53
ARGCN	86.67	82.72	84.65	79.45	71.60	75.32	76.57	76.88	76.72	86.16	84.19	85.16

Table 3: Ablation study results (%). LSTM-ARGCN denotes the model that places the BiLSTM before ARGCN. R-GCN (#basis=1) denotes using R-GCNs with basis decomposition and the number of basis is one. ARGCN (original) denotes using original dependency tree.

4.5 Visualization on Target Embedding

We also designed an experiment to evaluate if our model can learn suitable target embeddings during training, thereby it can benefit from the target-aware representation.

Intuitively, a good target embedding should have such a property: the representation of tag “O” is significantly different from that of tags “B” and “I”. However, representations of “ B” and “I” are similar. However, after training, as we can observe from Figure 4, the cosine similarity between “B” and “I” is close to 1 (0.94), whereas the similarity between “B” and “O” is even smaller than 0, and that between “I” and “O” has the same property. Therefore, we conclude that our model can learn to generate suitable target embeddings during training, which confirms the effectiveness and interpretability of our target-aware representation.

4.6 Ablation Study

To evaluate the influence of each component of ARGCN, we conducted an ablation study on ARGCN. As shown in Table 3, we observe performance drops on the four datasets when replacing ARGCN layers with R-GCN layers following Equations (2) and (3), which verifies the effectiveness of employing the distance attention mechanism in ARGCN. In addition, we also find that ARGCN outperforms GAT(Veličković et al., 2018), which proves that specifying the dependency relational type is crucial for applying the dependency graph to the TOWE task. Moreover, we compared ARGCN with an ABSA model, RGAT (Busbridge et al., 2019), which is similar to our model. We observe that our ARGCN performs much better than RGAT. These results prove that in TOWE task, the approach to encode syntactic information in ARGCN is more suitable than the approach used in RGAT.

In addition, to confirm that the syntactic graph that we constructed is effective and reasonable, we compare the ARGCN over the original dependency tree and our reshaped graph. The results show that the latter model outperforms the former one, which proves the effectiveness of our reshaped syntactic dependency graph.

4.7 Model Analysis

We further analyzed the effect of the layers number of ARGCN, the number of attention heads and the threshold of relative distance in our model by using different hyper-parameters but keeping the other hyper-parameters unchanged as the experimental settings mentioned above.

Because ARGCN involves an L-layer GCN, we investigated the effect of the layers number L with the final performance of ARGCN. Basically, we varied the value of L in the set $\{2, 4, 6, 8, 10, 12\}$ and showed the corresponding F1-score of ARGCN on the 14res dataset.

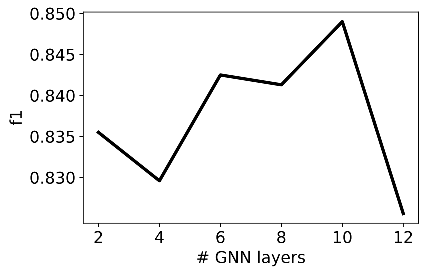

Figure 5: Effect of the number of GCN layers.

The results are illustrated in Figure 5, which shows that ARGCN achieves the best performance when $L = 10$. In this sense, our model can benefit from the increasing number of layers. However, when the number of layers is larger than 10, our model will tend to be over-smoothing which makes the performance drop dramatically.

As for the effect of the number of attention heads, we also varied the value of attention head number K in the set $\{1, 2, 4, 6, 8, 10, 12, 14\}$ and showed the corresponding F1-score of ARGCN on the 14res dataset.

Models	14res			14lap			15res			16res		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
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
RGCN (syntax only)	78.03	67.57	72.42	69.61	52.91	60.12	71.69	64.71	68.02	77.48	72.76	75.05
RGAT (syntax only)	78.59	61.65	69.10	61.87	45.50	52.44	68.09	58.42	62.88	74.62	65.52	69.78
ARGCN (syntax only)	78.51	71.65	74.92	68.23	56.44	61.78	74.36	65.31	69.55	82.11	72.57	77.05

Table 4: Evaluation of syntactic information (%). *Syntax only* means these models use only target embedding without word representation.

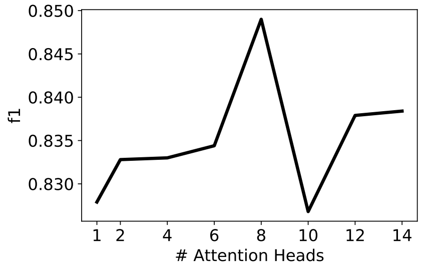

Figure 6: Effect of the number of Attention Heads.

The results are illustrated in Figure 6, which shows that ARGCN achieves the best performance when $K = 8$, which justifies the selection on the number of attention heads in the experimental settings. Comparing with the cases between $K = 1$ and $K = 8$, we found that the model with 8 attention heads performed better than that with only one attention head. This experiment demonstrated the necessity of the multi-head attention mechanism in our ARGCN.

As for the effect of threshold of relative distance, we performed experiments with different threshold D ranging from 1 to 6 and showed the corresponding F1-score of ARGCN on the 14res dataset.

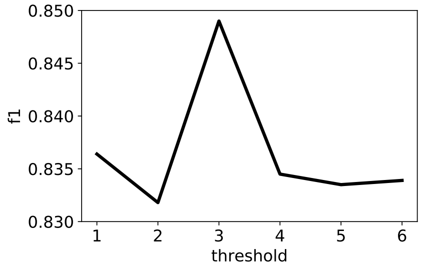

Figure 7: Effect of threshold.

The results are illustrated in Figure 7, where we observe that $D = 3$ is best value for the threshold in ARGCN.

4.8 Evaluation of syntactic information

To understand the role of syntactic information in the TOWE task and measure the ability of ARGCN to encode syntactic information, we removed word representation from our models, leaving the target embedding only. We performed some experiments on evaluating our model on the TOWE task only with syntactic information and position information. The results are shown in Table 4.

We notice that the GNN models perform better than the dependency-rule model, which indicates that the GNN models can exploit the syntactic information well from dependency graph. Furthermore, our well-designed ARGCN outperforms other GNN models including the latest one, RGAT. The reason is that ARGCN considers the relative position of words in the sentence and dependency relation type at the same time when it propagate the information.

5 Conclusions

In this paper, we proposed a target-aware representation to efficiently introduce opinion target information to our TOWE model. Moreover, we proposed ARGCN by extending the R-GCNs with a distance-aware attention mechanism. Because the sequential information is essential for such a sequence-labelling task, we captured the sequential information with BiLSTM after ARGCN layers and then completed the TOWE task. Empirical results show that our model significantly outperforms all baselines, including state-of-the-art, with large margins, which strongly proves the effectiveness of our model. The extensive analysis also demonstrated the effectiveness and necessity of all components in our model. In addition, we found that GNN model, especially a well-designed GNN model, such as ARGCN, is suitable for encoding syntactic information. We hope that these findings can be insightful for other researchers in the community.

References

Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y Hammerla. 2019. Relational graph attention networks.

Alberto Cetoli, Stefano Bragaglia, Andrew D O'Harney, and Marc Sloan. 2017. Graph convolutional networks for named entity recognition. *arXiv preprint arXiv:1709.10053*.

Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi. 2020. Synchronous double-channel recurrent network for aspect-opinion pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6515–6524, Online. Association for Computational Linguistics.

Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. 2018. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7239–7248.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Nicola De Cao and Thomas Kipf. 2018. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Victor Garcia and Joan Bruna Estrach. 2018. Few-shot learning with graph neural networks. In *6th International Conference on Learning Representations, ICLR 2018*.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

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

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1373–1378.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1035–1045.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Yingju Xia, Shu Zhang, and Hao Yu. 2010. Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 653–661.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2886–2892.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2020. A dependency syntactic knowledge augmented interactive architecture for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:2004.01951*.

Kang Liu, Heng Li Xu, Yang Liu, and Jun Zhao. 2013. Opinion target extraction using partially-supervised word alignment model. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Kang Liu, Liheng Xu, and Jun Zhao. 2012. Opinion target extraction using word-based translation model. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1346–1356. Association for Computational Linguistics.

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, Lisbon, Portugal. Association for Computational Linguistics.

Diego Marcheggiani, Joost Bastings, and Ivan Titov. 2018. Exploiting semantics in neural machine translation with graph convolutional networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 486–492.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

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

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019. Aspect-level sentiment analysis via convolution over dependency tree. 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 5683–5692.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*. Accepted as poster.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.12362*.

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.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):1–12.

Zhen Wu, Fei Zhao, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2020. Latent opinions transfer network for target-oriented opinion words extraction. *arXiv preprint arXiv:2001.01989*.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, Melbourne, Australia. Association for Computational Linguistics.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. *arXiv preprint arXiv:1909.03477*.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, page 43–50, New York, NY, USA. Association for Computing Machinery.