Attention Transfer Network for Aspect-level Sentiment Classification

Fei Zhao[*] **Zhen Wu**[*] **Xinyu Dai**[†]
National Key Laboratory for Novel Software Technology, Nanjing University, China
Collaborative Innovation Center of Novel Software Technology and Industrialization, China
{zhaof, wuz}@smail.nju.edu.cn, daixinyu@nju.edu.cn

Abstract

Aspect-level sentiment classification (ASC) aims to detect the sentiment polarity of a given opinion target in a sentence. In neural network-based methods for ASC, most works employ the attention mechanism to capture the corresponding sentiment words of the opinion target, then aggregate them as evidence to infer the sentiment of the target. However, aspect-level datasets are all relatively small-scale due to the complexity of annotation. Data scarcity causes the attention mechanism sometimes to fail to focus on the corresponding sentiment words of the target, which finally weakens the performance of neural models. To address the issue, we propose a novel Attention Transfer Network (ATN) in this paper, which can successfully exploit attention knowledge from resource-rich document-level sentiment classification datasets to improve the attention capability of the aspect-level sentiment classification task. In the ATN model, we design two different methods to transfer attention knowledge and conduct experiments on two ASC benchmark datasets. Extensive experimental results show that our methods consistently outperform state-of-the-art works. Further analysis also validates the effectiveness of ATN. Our code and dataset are available at https://github.com/1429904852/ATN.

1 Introduction

Aspect-level sentiment classification (ASC) is a fundamental task in sentiment analysis (Pang et al., 2008; Liu, 2012; Pontiki et al., 2014), which aims to infer the sentiment polarity (e.g. positive, neutral, negative) of a given opinion target in a review sentence. An opinion target, also known as aspect term, refers to a word or a phrase in review describing an aspect of an entity. For example, the sentence "*The* ***tastes*** *are great, but the* ***service*** *is dreadful*" consists of two opinion targets, namely "*tastes*" and "*service*". User's sentiment towards the opinion target "*tastes*" is positive while negative in terms of target "*service*". Traditional methods usually focus on designing a set of features such as bag-of-words or sentiment lexicon to train a classifier (e.g., SVM) for ASC (Jiang et al., 2011; Kiritchenko et al., 2014). Motivated by the great success of deep learning in computer vision (Krizhevsky et al., 2012), speech recognition (Dahl et al., 2012) and natural language processing (Bengio et al., 2003), recent works use neural networks to learn low-dimensional and continuous text representations without any feature engineering, and achieve competitive results on the ASC task (Tang et al., 2016a).

From the above example, we can see that a sentence sometimes refers to several opinion targets and they may express different sentiment polarities, thus one main challenge of ASC is to separate different opinion contexts for different targets. To this end, abundant state-of-the-art works employ attention mechanism (Bahdanau et al., 2014) to capture sentiment words related to the given target, and then aggregate them to make sentiment prediction (Wang et al., 2016; Tang et al., 2016b; Ma et al., 2017; Chen et al., 2017; Majumder et al., 2018; Fan et al., 2018). Despite the effectiveness of attention mechanism, we argue that it fails to reach the full potential due to the limited ASC labeled data. It is well-known that the promising results of deep learning heavily rely on sufficient training data. However, the annotation

[*] Authors contributed equally.

[†] Corresponding author.

Proceedings of the 28th International Conference on Computational Linguistics, pages 811–821
Barcelona, Spain (Online), December 8-13, 2020

of ASC data is very labour-intensive and expensive in real-world scenarios, because annotators need to not only identify all opinion targets in a sentence but also determine their corresponding sentiment polarity. The difficulty of annotation leads to that existing public aspect-level datasets are all relatively small-scale, which finally limits the potential of attention mechanism.

Despite the lack of ASC data, enormous labeled data of document-level sentiment classification (DSC) are available at online review sites such as Amazon and Yelp. These reviews contain substantial sentiment knowledge and semantic patterns. Therefore, one meaningful but challenging research question is how to leverage resource-rich DSC data to improve the low-resource task ASC. For this purpose, He et al. (2018) design the PRET+MULT framework to transfer sentiment knowledge from DSC data to ASC task through sharing shallow embedding and LSTM layer. Inspired by the capsule network (Sabour et al., 2017), Chen and Qian (2019) propose TransCap to share bottom three capsule layers, then separate two tasks only in the last ClassCap layer. Fundamentally, PRET+MULT and Transcap improve ASC by sharing parameters and multi-task learning, but they cannot accurately control and interpret what knowledge to be transferred. In this work, we directly focus on the aforementioned attention issue in the ASC task and propose a novel framework, **A**ttention **T**ransfer **N**etwork (ATN), to explicitly transfer attention knowledge from the DSC task for improving the attention capability of the ASC task. Compared with PRET+MULT and Transcap, our model achieves better results and retains good interpretability.

In the ATN framework, we adopt two attention-based BiLSTM networks, respectively, as the DSC module and base ASC module, and propose two different methods to transfer attention from DSC to ASC. The first transfer approach is called *Attention Guidance*. Specifically, we first pre-train an attention-based BiLSTM on large-scale DSC data, then exploit the attention weights from the DSC module as a learning signal to guide the ASC module to capture sentiment clues more accurately, thereby acheiving improvements. The second approach adopts the way of *Attention Fusion*, and directly incorporates the attention weights of the DSC module into the ASC module. The two approaches work in different ways and have their different advantages. *Attention Guidance* aims to learn the attention ability of the DSC module and has faster inference speed, since it does not use external attention from DSC during the testing stage. In contrast, *Attention Fusion* can leverage the attention knowledge of the DSC module during the testing stage and make more comprehensive predictions.

We conduct experiments on two benchmark datasets to evaluate different methods. The results indicate that the ATN model can be substantially improved by incorporating the two attention transfer approaches, and outperforms all compared methods on the ASC task.

2 Model

Figure 1 shows the overall architecture of the Attention Transfer Network (ATN). It mainly consists of four parts: the pre-trained DSC module, the base ASC module, and two attention transfer approaches. In this section, we will first give the task formalization of ASC and DSC, then introduce the attention-based pre-trained DSC module and base ASC module. Finally, we present the details of our proposed two attention transfer approaches, namely *Attention Guidance* and *Attention Fusion*.

2.1 Task Formalization

ASC Formalization Formally, given a sample $< s, t >$ from the ASC dataset $\mathcal{A}$, $s = \{w_1, w_2, ..., w_n\}$ is a review sentence consisting of n words and $t = \{w_l, w_{l+1}, ..., w_r\}$ is a given opinion target containing $|r - l|$ words. The opinion target t is a continuous subsequence of s. The goal of ASC is to predict the sentiment polarity (i.e., positive, neutral and negative) of the opinion target t in the sentence s.

DSC Formalization For a review document d from the DSC dataset $\mathcal{D}$, we regard it as a special long sentence $\{w^d_1, w^d_2, ..., w^d_n\}$ consisting of n words. DSC aims to determine the overall sentiment polarity of the review document d.

2.2 Pre-trainig DSC Module

Before transferring attention knowledge, we first pre-train a DSC module on the large-scale DSC dataset $\mathcal{D}$. In this work, we employ a conventional attention-based BiLSTM as our DSC module.

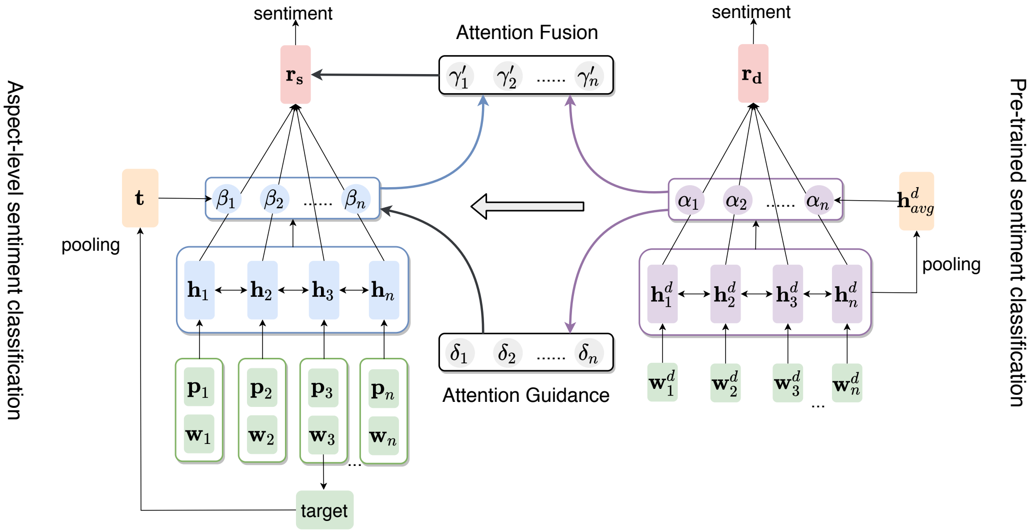
sentiment
Attention Fusion
sentiment
$\mathbf{r_s}$
γ'_1 γ'_2 ...... γ'_n
$\mathbf{r_d}$
Aspect-level sentiment classification
Pre-trained sentiment classification
t
β_1 β_2 ...... β_n
α_1 α_2 ...... α_n
$\mathbf{h}^d_{avg}$
pooling
pooling
$\mathbf{h}_1$ $\mathbf{h}_2$ $\mathbf{h}_3$ $\mathbf{h}_n$
$\mathbf{h}^d_1$ $\mathbf{h}^d_2$ $\mathbf{h}^d_3$ $\mathbf{h}^d_n$
δ_1 δ_2 ...... δ_n
Attention Guidance
$\mathbf{p}_1$ $\mathbf{p}_2$ $\mathbf{p}_3$ $\mathbf{p}_n$
$\mathbf{w}_1$ $\mathbf{w}_2$ $\mathbf{w}_3$ $\mathbf{w}_n$
$\mathbf{w}^d_1$ $\mathbf{w}^d_2$ $\mathbf{w}^d_3$ $\mathbf{w}^d_n$
target

Figure 1: An illustration of our attention transfer network. The left one is the aspect-level sentiment classification, the right one is the pre-trained DSC module, and the middle part presents two proposed attention transfer approaches.

For a review document $d = \{w^d_1, w^d_2, ..., w^d_n\}$, we map it into the corresponding word representations $\{\mathbf{w}^d_1, \mathbf{w}^d_2, ..., \mathbf{w}^d_n\}$ by looking up an embedding table $\mathbf{E}_{emb} \in \mathbb{R}^{|V| \times d_e}$, where $|V|$ is the vocabulary size and d_e denotes the word embedding dimension. Then a BiLSTM network is applied to capture the contextual information for each word and generate a sequence of hidden states $\{\mathbf{h}^d_1, \mathbf{h}^d_2, ..., \mathbf{h}^d_n\}$. To obtain the document representation $\mathbf{r}_d$, we employ the attention mechanism to aggregate the sentiment words that are significant for sentiment classification as follows:

$$\mathbf{r}_d = \sum_{i=1}^{n} \alpha_i \mathbf{h}^d_i, \tag{1}$$

where α_i is the attention weight of $\mathbf{h}^d_i$ and defined as:

$$\alpha_i = \frac{\exp(f(\mathbf{h}^d_i, \mathbf{h}^d_{avg}))}{\sum_{j=1}^{n} \exp(f(\mathbf{h}^d_j, \mathbf{h}^d_{avg}))}, \tag{2}$$

$$f(\mathbf{h}^d_i, \mathbf{h}^d_{avg}) = \mathbf{h}^d_i \cdot \mathbf{W}_d \cdot \mathbf{h}^d_{avg} + \mathbf{b}_d, \tag{3}$$

where $\mathbf{h}^d_{avg}$ is the average of all the hidden states, i.e., $\mathbf{h}^d_{avg} = \sum_{i=1}^{n} \mathbf{h}^d_i / n$, $\mathbf{W}_d$ and $\mathbf{b}_d$ are respectively the weight matrix and bias.

Finally, the representation $\mathbf{r}_d$ is fed to a linear layer and a softmax layer to predict the sentiment label of the review document d. We pre-train the DSC module by minimizing the cross-entropy loss between the predicted sentiment distribution and the ground truth. After pre-training is finished, all parameters in the DSC module are fixed.

2.3 Base ASC Module

As shown in the left part of Figure 1, the base ASC module has a similar architecture to the DSC module. The difference is that the ASC task needs to model opinion target information. To obtain target-aware context representations, we additionally employ position embedding besides word embedding, which is an effective method of modeling position information (Lin et al., 2016; Gehring et al., 2016). Therefore, the base ASC module is an attention-based BiLSTM network enhanced with position embedding.

Specifically, given a sentence $s = \{w_1, w_2, ..., w_n\}$ and an opinion target $t = \{w_l, w_{l+1}, ..., w_r\}$ in s, we first map each word w_i into its word embedding representation $\mathbf{w}_i$ by using the word embedding table.

To incorporate opinion target information with position embedding, we calculate the relative distance l_i of each word w_i to the opinion target t:

$$l_i = \begin{cases} l - i & \text{if } i < l, \\ 0 & \text{if } l \leq i \leq r, \\ i - r & \text{otherwise .} \end{cases} \tag{4}$$

The distance index l_i is mapped into the positional representation $\mathbf{p}_i$ by looking up a position embedding table $\mathbf{E}_{pos} \in \mathbb{R}^{L \times d_p}$, where L denotes the maximal position index and d_p is the embedding dimension. Then we concatenate the word embedding representation $\mathbf{w}_i$ and position embedding representation $\mathbf{p}_i$ as the repsentation $\mathbf{e}_i$ of the word w_i, i.e., $\mathbf{e}_i = [\mathbf{w}_i; \mathbf{p}_i]$, where $[\cdot;\cdot]$ denotes the vector concatenation operation. Similarly, we employ a BiLSTM to receive the word represenations $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n\}$ as input and generate target-aware context representations $\{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n\}$. Different from the attention part of the DSC module, we use the opinion target represenation $\mathbf{t} = \sum_{i=l}^{r} \mathbf{h}_i/(r-l)$ as query in the ASC task to extract target-dependent sentiment clues:

$$f(\mathbf{h}_i, \mathbf{t}) = \mathbf{h}_i \cdot \mathbf{W}_s \cdot \mathbf{t} + \mathbf{b}_s, \tag{5}$$

$$\beta_i = \frac{\exp(f(\mathbf{h}_i, \mathbf{t}))}{\sum_{j=1}^{n} \exp(f(\mathbf{h}_j, \mathbf{t}))}, \tag{6}$$

$$\mathbf{r}_s = \sum_{i=1}^{n} \beta_i \mathbf{h}_i, \tag{7}$$

where $\mathbf{W}_s$ and $\mathbf{b}_s$ are respectively the weight matrix and bias.

Finally, the target-dependent sentence representation $\mathbf{r}_s$ is used for detecting the sentiment polarity of the target t, and the base ASC module can optimized by minimizing the following cross-entropy loss:

$$\hat{y}_i = \text{softmax}(\mathbf{W}_o \mathbf{r}_s + \mathbf{b}_o), \tag{8}$$

$$\mathcal{L}_o = -\sum_{i \in \mathcal{A}} y_i log(\hat{y}_i), \tag{9}$$

where $\hat{y}_i$ and y_i respectively are the predictive class distribution and golden class distribution.

2.4 Attention Guidance

To leverage the attention knowledge of the DSC module, we simultaneously input the sentence s into the base ASC module and the pre-trained DSC module when performing the ASC task, generating the attention weights β_i in Equation 6 and α_i in Equation 2.

As mentioned before, the attention mechanism of the ASC module cannot reach full potential due to limited training data, which means that the attention weights β_i may fail to capture target-relevant sentiment words. In contrast, sufficient DSC data enables the DSC module to extract sentiment words more accurately. Thus we propose the *Attention Guidance* approach to guide the learning of the attention weights β_i with the help of α_i. Nevertheless, there is a tiny gap between the attention weights α_i and β_i. Since the DSC task only detects the overall sentiment of a review, the sentiment words captured by α_i are global and target-irrelevant. To make up the gap, we use a heuristic method to transform target-irrelevant attention weight α_i into target-relevant weight δ_i:

$$\alpha'_i = \frac{1}{2^{(l_i - 1)}} \alpha_i, \tag{10}$$

$$\delta_i = \frac{e^{\alpha'_i}}{\sum_{i=1}^{n} e^{\alpha'_i}}, \tag{11}$$

where l_i denotes the relative distance between the word and the target as in Equation 4. We can see that a word nearer to the target receives a higher attention weight according to δ_i, because the closer word has a bigger probability of modifier relation to the target.

Finally, we apply KL (Kullback–Leibler divergence) to describe the differences between attention distributions β and δ:

$$KL(\delta||\beta) = \sum_{i=1}^{n} \delta_i log \frac{\delta_i}{\beta_i}, \tag{12}$$

$$= \sum_{i=1}^{n} (\delta_i log \delta_i - \delta_i log \beta_i). \tag{13}$$

In the pre-trained DSC module, the above term $\sum_{i=1}^{n} \delta_i log \delta_i$ in Equation 13 is invariant for the given sentence s and the opinion target t. Therefore, we can minimize the loss $\mathcal{L}_a = \sum_{i=1}^{n} -\delta_i log \beta_i$ to guide the ASC module to focus on target-relevant sentiment words. In the *Attention Guidance* approach, the final loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_o + \lambda \mathcal{L}_a. \tag{14}$$

where λ is the hyperparameter that controls the importance of $\mathcal{L}_a$.

2.5 Attention Fusion

Attention Guidance learns the attention ability of the DSC module through an auxiliary supervision signal. However, it cannot leverage the attention weights from the DSC module during the testing stage and wastes the pre-trained knowledge. To make full use of the additional attention capacity, we further propose the *Attention Fusion* approach to incorporate them directly.

Specifically, we design a fusion gate g to integrate the global attention weight α_i from the DSC module and the target-dependent attention weight β_i from the ASC module, thereby generating more comprehensive and accurate attention weight γ_i':

$$g = \sigma(\mathbf{W}_g[\alpha_i; \beta_i]), \tag{15}$$

$$\gamma_i = g\alpha_i + (1-g)\beta_i, \tag{16}$$

$$\gamma_i' = \frac{e^{\gamma_i}}{\sum_{i=1}^{n} e^{\gamma_i}}, \tag{17}$$

where σ denotes sigmoid function and $\mathbf{W}_g$ is the weight matrix.

Finally, we replace β_i in Equation 7 with the new attention weight γ_i' to obtain the target-dependent sentence representation $\mathbf{r}_s$ for sentiment prediction.

3 Experiments

3.1 Datasets and Metrics

We evaluate our model on two ASC benchmark datasets from SemEval 2014 Task 4 (Pontiki et al., 2014). They respectively contain reviews from *Restaurant* and *Laptop* domains. Following previous studies (Tang et al., 2016b; Chen et al., 2017; He et al., 2018), we remove samples with conflicting polarities in all datasets. The statistics of the ASC datasets are shown in Table 1.

To pre-train the DSC module, we employ two larget-scale DSC datasets, respectively *Yelp Review* and *Amazon Review* (Li et al., 2018a). The DSC dataset *Yelp Review* is applied to transfer attention knowledge for the ASC dataset *Restaurant*. The *Amazon Review* is used for the dataset *Laptop*. Table 2 shows their statistics. In this work, we adopt Accuracy and Macro-F1 score as the metrics to evaluate the performance of different methods on the ASC task.

3.2 Experimental Settings

In our experiments, word embeddings are initialized by 300-dimension GloVe (Pennington et al., 2014). After initialization, the word vectors are fixed and not fine-tuned during the training stage. All the weight matrices and biases are given the initial value by sampling from the uniform distribution $U(-0.1, 0.1)$. The dimension of LSTM cell hidden states is set to 300. We employ stochastic gradient descent (SGD)

Dataset	#Pos	#Neg	#Neu	#Total
Restaurant-Train	2164	807	637	3608
Restaurant-Test	728	196	196	1120
Laptop-Train	994	870	464	2328
Laptop-Test	341	128	169	638

Table 1: Statistics of the ASC datasets.

Datasets	#Pos	#Neg	#Total
Yelp Review	266k	177k	443k
Amazon Review	277k	277k	554k

Table 2: Statistics of the DSC datasets.

with momentum (Qian, 1999) to train models. The initial learning rate and momentum parameter are respectively set to 0.1 and 0.9. In addition, we apply dropout (Hinton et al., 2012) with probability 0.5 on embedding layer as a regularizer. The parameter λ in *Attention Guidance* approach is set to 0.4. All hyper-parameters were tuned on 20% randomly held-out training data. Finally, we run each model five times and report the average result of them.

3.3 Compared Methods

We divide compared methods into two groups according to whether using transferred knowledge.
(I). The first group contains some classic methods for the ASC task:

Majority assigns each instance in the test set with the most frequent sentiment label in the training set.

Feature-based SVM (Kiritchenko et al., 2014) is the top system of SemEval 2014 Task 4. It uses n-gram features, parse features and lexicon features to train an SVM classifier.

TD-LSTM (Tang et al., 2016a) applies two LSTM networks to model the left context and right context of opinion target respectively, then concatenates their last hidden states for sentiment prediction.

ATAE-LSTM (Wang et al., 2016) concatenates the word embedding and target embedding as the input of LSTM, then employs the attention mechanism to capture target-dependent sentiment information.

IAN (Ma et al., 2017) proposes the interactive attention to interactively learn representations of the context and target. The two representations are then concatenated for prediction.

MemNet (Tang et al., 2016b) uses multi-hops attention on the word embeddings to generate the target-dependent sentence representation.

RAM (Chen et al., 2017) works similar to the method MemNet. It employs BiLSTM to build memory and applies GRU-based multi-hops attention.

IARM (Majumder et al., 2018) incoporates the neighboring targets-related information for ASC by using memory networks.

MGAN (Fan et al., 2018) proposes a fine-grained attention mechanism to capture the word-level interaction between target and context, then combines it with coarse-grained attention for ASC.

GCAE (Xue and Li, 2018) uses a convolutional neural network (CNN) with gating mechanisms to perform the ASC task.

TNet (Li et al., 2018b) proposes target specific transformation component to integrate target information into the word representation.

(II). Besides, we also compare two existing methods using transferred knowledge from large-scale DSC data to facilitate the ASC task:

PRET+MULT (He et al., 2018) shares shadow embedding and LSTM layers between the ASC model and the DSC model through multi-task learning.

TransCap (Chen and Qian, 2019) employs capsule network to share the bottom features between the ASC task and the DSC task.

3.4 Main Results and Analysis

The main results are shown in Table 3. We classify the results into three groups: the first lists the classic methods for the ASC task, the second presents two existing transfer-based methods, and the last is our base ASC model and enhanced versions with transferring attention knowledge. We use ATN-AG and ATN-AF respectively to represent ATN using *Attention Guidance* and *Attention Fusion*.

Method	Restaurant		Laptop	
	Acc.	Macro-F1	Acc.	Macro-F1
Majority	65.00	33.33	53.50	33.33
Feature-SVM (Kiritchenko et al., 2014)	80.16	N/A	70.49	N/A
ATAE-LSTM (Wang et al., 2016)	77.20	N/A	68.70	N/A
TD-LSTM (Tang et al., 2016a)	78.00	66.73	71.83	68.43
IAN (Ma et al., 2017)	78.60	N/A	72.10	N/A
MemNet (Tang et al., 2016b)	80.32	N/A	72.37	N/A
RAM (Chen et al., 2017)	80.23	70.80	74.49	71.35
IARM (Majumder et al., 2018)	80.00	N/A	73.80	N/A
MGAN (Fan et al., 2018)	81.25	71.94	75.39	72.47
GCAE (Xue and Li, 2018)	77.43	66.24	71.03	64.43
TNet (Li et al., 2018b)	80.79	70.84	76.01	71.47
PRET+MULT (He et al., 2018)	79.98	69.39	74.14	69.14
TransCap (Chen and Qian, 2019)	80.72	71.98	74.92	70.21
Base ASC model	80.38	70.69	73.52	70.78
ATN-AG	$81.39^{\dagger}$	$72.44^{\dagger}$	$76.41^{\dagger}$	$72.59^{\dagger}$
ATN-AF	$\mathbf{82.36}^{\dagger}$	$\mathbf{74.00}^{\dagger}$	$\mathbf{76.48}^{\dagger}$	$\mathbf{72.60}^{\dagger}$

Table 3: Main experiment results (%). The base ASC model is attention-based BiLSTM enhanced with position embedding. AT-AG and ATN-AF respectively refer to ATN model using *Attention Guidance* and *Attention Fusion*. The best performances are marked in bold. The marker † represents that ATN-AG and ATN-AF outperform the compared methods significantly ($p < 0.05$).

The method Feature-SVM obtains competitive results on the restaurant dataset but performs poorly on the laptop dataset. This may be attributed to that the performance of simple feature-based methods heavily relies on the quality of hand-crafted features. IAN achieves better performance than TD-LSTM and ATAE-LSTM by using the interactive attention mechanism to learn the representations of context and opinion target. With combining of fine-grained and coarse-grained attention mechanisms, MGAN achieves the best performance among all pure attention-based models. Among the memory-based methods, it can be observed that RAM outperforms MemNet and IARM on the laptop dataset, which validates the effectiveness of multi-hops attention based on recurrent network. GCAE performs poorly compared with other neural methods, as CNN is not good at capturing the long-term dependencies between context words. TNet achieves state-of-the-art performance by designing target-specific transformation mechanism between LSTM and CNN.

PRET+MULT and Transcap transfer knowledge implicitly from large-scale DSC data to the ASC task through sharing parameters and multi-task learning. They show superiority compared to some methods without transferring knowledge. For example, the base model of PRET+MULT is an attention-based LSTM similar to ATAE-LSTM. We can observe that PRET+MULT outperforms ATAE-LSTM significantly, and achieves 2.78% and 5.44% accuracy improvements respectively on the restaurant and laptop datasets. Transcap obtains better results compared to PRET+MULT, which verifies the effectiveness of capsule network for capturing shared features.

Our base ASC model attention-based BiLSTM enhanced with position embedding performs better than some attention-based models, such as ATAE-LSTM and IAN. This result indicates that position embedding is beneficial for modeling target information in the ASC task. On this basis, our attention transfer models ATN-AG and ATN-AF respectively achieve about 1% and 2% improvements in accuracy on the restaurant dataset, and over 2.8% improvements on the laptop dataset. In addition, they surpass two existing methods that use transferred knowledge obviously, i.e., PRET+MULT and Transcap. These comparisons demonstrate the effectiveness of our proposal of explicitly transferring attention knowledge from resource-rich DSC data to the ASC task. Compared with ATN-AG, ATN-AF achieves better performance on the restaurant dataset. It is reasonable because ATN-AG cannot leverage the attention weights of the DSC module during the testing stage. Nevertheless, ATN-AG still obtains comparable results on the laptop dataset and has a faster inference speed than ATN-AF.

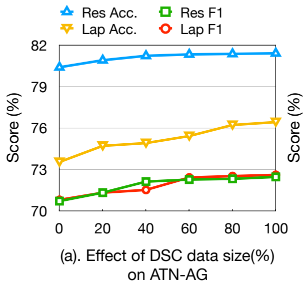

Figure 2: Performance of ATN-AG and ATN-AF with different percentages of DSC data.

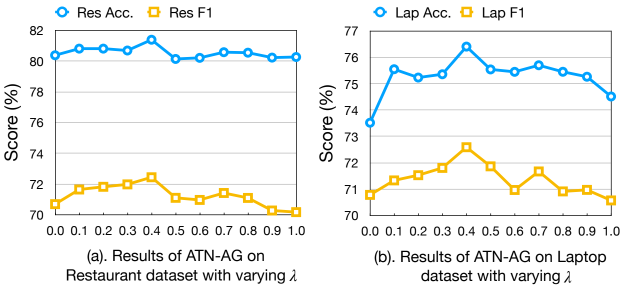

Figure 3: Effect of hyper-parameter λ on ATN-AG.

Base model	I use it mostly for **[content creation]** (Audio , video , photo editing) and its reliable .	Neagtive✘
ATN-AG	I use it mostly for **[content creation]** (Audio , video , photo editing) and its reliable .	Positive✔
ATN-AF	I use it mostly for **[content creation]** (Audio , video , photo editing) and its reliable .	Positive✔
Base model	Did not enjoy the new Windows 8 and **[touchscreen functions]**	Positive✘
ATN-AG	Did not enjoy the new Windows 8 and **[touchscreen functions]**	Negative✔
ATN-AF	Did not enjoy the new Windows 8 and **[touchscreen functions]**	Negative✔

Table 4: Attention visualization of ATN-AG and ATN-AF. The spans in bold are opinion targets. A darker color indicates a higher attention weight.

3.5 Effect of DSC Data Size

To investigate the effect of DSC data size on our approaches, we vary the percentage of DSC data from 0% to 100% to report the results of ATN-AG and ATN-AF. The critical values 0% and 100% respectively mean no DSC data and using the complete DSC dataset. The results are shown in Figure 2.

We can observe that our approaches ATN-AG and ATN-AF both achieve very stable improvements on the two datasets with the increase of DSC data size. This indicates that the ASC task indeed benefits from the transferred attention knowledge from the pre-trained DSC module. Consistent and stable improvements show the robustness of our approaches.

3.6 Effect of Hyper-parameter λ

To analyze the effect of hyper-parameter λ in Equation 14 on ATN-AG, we adjust it in [0, 1] to conduct experiments and the step is 0.1. Figure 3 shows the performance of ATN-AG with different λ on the restaurant and laptop datasets.

We can see that the curves on two datasets have an overall upward trend when $\lambda < 0.4$, but become flat or downward once $\lambda > 0.4$. In the upward part, the attention knowledge from the DSC module is a useful guidance signal to help the ASC module to focus on sentiment words more accurately, thus improve the performance of ASC. Once the weight λ exceeds 0.4, the transferred attention knowledge begins to dominate the attention process while the ASC module loses the mastership and perform worse. Therefore, we finally set λ to be 0.4 on two datasets.

3.7 Case Study

In the ATN model, we propose the approaches *Attention Guidance* and *Attention Fusion* to help the ASC module to capture sentiment clues more accurately. To verify this, we analyze some dozens of instances from the test set. Compared with the base ASC model, we find that our attention transfer methods can deal with low-frequency sentiment words and complex sentiment patterns such as negation. Table 4 shows the attention visualizations of two examples and the corresponding sentiment predictions under the base model, ATN-AG and ATN-AF. Note that the darker color means higher attention weight.

In the first example, the base ASC model mainly focuses on the adverb "*mostly*", while fails to capture the critical sentiment clue "*reliable*". According to the statistics, the word "*reliable*" only appears five times in the training set. This indicates that the base model is not good at catching low-frequency sentiment words, thus makes wrong sentiment predictions. In contrast, the enhanced models ATN-AG and ATN-AF with transferred attention knowledge both successfully capture the informative word "*reliable*", and give the right predictions.

From the second example, we can see that the base ASC model mainly focuses on the word "*enjoy*" rather than the sentiment negator "*not*". It is hard for the base model to learn the negation with the insufficient labeled dataset. With the help of the external attention knowledge, our approaches ATN-AG and ATN-AF pay more attention to the negator "*not*", and make correct sentiment predictions.

The above observations show that our approaches indeed improve the low-resource task ASC with the transferred attention knowledge and retain good interpretability.

4 Related Work

4.1 Aspect-level Sentiment Classification

Early works adopt supervised learning and devote to designing effective features for the ASC task, such as n-gram features (Kiritchenko et al., 2014) and sentiment lexicons (Vo and Zhang, 2015). The performance of these methods heavily depends on labor-intensive feature engineering. With the development of deep learning, Tang et al. (2016a) use two Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks to respectively model the left context and right context of the given opinion target. However, it cannot capture the association between the context and opinion target. To address the issue, recent works employ the attention mechanism to catch target-dependent sentiment context and achieve very promising resutls (Wang et al., 2016; Ma et al., 2017; Fan et al., 2018). Instead of single attention, some works propose multi-hops attention based on memory networks (Sukhbaatar et al., 2015) to detect more powerful sentiment clues (Tang et al., 2016b; Chen et al., 2017; Majumder et al., 2018).

Despite attention-based models showing the potential for ASC, they highly rely on data-driven attention mechanism. Unfortunately, public ASC datasets are all small-scale because of the complexity of annotation. Insufficient labeled data finally limits the effectiveness of attention mechanism for the ASC task. Different from the above methods, we improve the attention capacity of the ASC model in this work, by transferring substantial attention knowledge from the DSC model pre-trained with resource-rich document-level sentiment classification data.

4.2 Transfer Learning

Transfer learning aims to extract knowledge from one or more source tasks and then apply them to a target task. Neural transfer learning has proven effective for image recognition (Donahue et al., 2014) and natural language processing tasks (Mou et al., 2016; Dong and De Melo, 2018; Wu et al., 2020). He et al. (2018) are the first to transfer knowledge from document-level review data to improve the ASC task through sharing embedding and LSTM layers. Chen and Qian (2019) employ capsule network to share bottom features between the ASC task and DSC task. In this work, we aim to transfer attention knowledge from the DSC model explicitly to improve the effectiveness of attention mechanism for the ASC task. In contrast to the two existing works, our proposed approaches show better performance and good interpretability.

5 Conclusion

Insufficient labeled data limits the effectiveness of attention-based models for the ASC task. In this paper, we propose a novel attention transfer framework, in which two different attention transfer methods are designed to exploit attention knowledge from resource-rich document-level sentiment classification corpus to enhance the attention process of resource-poor aspect-level sentiment classification, finally achieving the goal of improving the performance of ASC. Experimental results indicate that our approaches outperform the state-of-the-art works. Further analysis validates the effectiveness and benefits of transferring the attention knowledge from DSC data for the ASC task.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. This work was supported by the NSFC (No. 2018YFB1005102) and National Key R&D Program of China (No. 61976114, 61936012).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Zhuang Chen and Tieyun Qian. 2019. Transfer capsule network for aspect level sentiment classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 547–556.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.

George E. Dahl, Dong Yu, Li Deng, and Alex Acero. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Audio, Speech & Language Processing*, 20(1):30–42.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655.

Xin Dong and Gerard De Melo. 2018. A helping hand: Transfer learning for deep sentiment analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2524–2534.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442.

Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Exploiting document knowledge for aspect-level sentiment classification. *arXiv preprint arXiv:1806.04346*.

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

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

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018a. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018b. Transformation networks for target-oriented sentiment classification. *arXiv preprint arXiv:1805.01086*.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, Md Shad Akhtar, Erik Cambria, and Asif Ekbal. 2018. Iarm: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3402–3411.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in nlp applications? *arXiv preprint arXiv:1603.06111*.

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

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

Ning Qian. 1999. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *COLING*.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *EMNLP*.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Zhen Wu, Fei Zhao, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2020. Latent opinions transfer network for target-oriented opinion words extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9298–9305. AAAI Press.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. *arXiv preprint arXiv:1805.07043*.