Reinforced Transformer with Cross-Lingual Distillation for Cross-Lingual Aspect Sentiment Classification

Hanqian Wu [1,2,]*[ID], Zhike Wang [1,2], Feng Qing [1,2] and Shoushan Li [3]

1   School of Computer Science and Engineering, Southeast University, Nanjing 210000, China;
    zhikerwang@163.com (Z.W.); qingfeng9101@foxmail.com (F.Q.)
2   Key Laboratory of Computer Network and Information Integration of Ministry of Education,
    Southeast University, Nanjing 210000, China
3   NLP Lab, School of Computer Science and Technology, Soochow University, Suzhou 215000, China;
    lishoushan@suda.edu.cn
*   Correspondence: hanqian@seu.edu.cn

Abstract: Though great progress has been made in the Aspect-Based Sentiment Analysis(ABSA) task through research, most of the previous work focuses on English-based ABSA problems, and there are few efforts on other languages mainly due to the lack of training data. In this paper, we propose an approach for performing a Cross-Lingual Aspect Sentiment Classification (CLASC) task which leverages the rich resources in one language (source language) for aspect sentiment classification in a under-resourced language (target language). Specifically, we first build a bilingual lexicon for domain-specific training data to translate the aspect category annotated in the source-language corpus and then translate sentences from the source language to the target language via Machine Translation (MT) tools. However, most MT systems are general-purpose, it non-avoidably introduces translation ambiguities which would degrade the performance of CLASC. In this context, we propose a novel approach called Reinforced Transformer with Cross-Lingual Distillation (RTCLD) combined with target-sensitive adversarial learning to minimize the undesirable effects of translation ambiguities in sentence translation. We conduct experiments on different language combinations, treating English as the source language and Chinese, Russian, and Spanish as target languages. The experimental results show that our proposed approach outperforms the state-of-the-art methods on different target languages.

Keywords: cross-lingual aspect sentiment classification; reinforced transformer; adversarial learning

1. Introduction

Aspect Sentiment Classification (ASC) aims to identify fine-grained polarity towards a specific aspect category (i.e., aspect). This task allows users to evaluate aggregated sentiments for each aspect of a given product or service and gain a more granular understanding. To date, a number of corpus-based approaches (Ma et al. [1]; Wang et al. [2]; Xu et al. [3]) have been developed for ASC. The approaches heavily rely on an adequate amount of manually annotated corpora for every domain. However, labeled data are not evenly distributed among languages and across domains. For a few rich-resource languages, including English, such labeled data are easily available. However, for many other languages, it is normal that only a limited number of labeled data exists (Lo et al. [4]). To leverage resources in the source language (e.g., English) to improve the aspect sentiment classification performance in the target language, we focus on the research task, namely Cross-Lingual Aspect Sentiment Classification (CLASC).

Existing methods in cross-lingual tasks either employ Machine Translation (MT) systems or combine a task-agnostic, pre-trained cross-lingual model with a task-specific neural architecture. There are limitations and challenges in both methods. MT is a common approach of bridging the gap between languages; however, there are some expected

drawbacks. Most MT systems (e.g., Google Translate) are trained on general text and thus introduce translation ambiguities which would lead to the domain shift problem (Fu et al. [5]) when translate domain-specific text. Moreover, redundant tokens are introduced, which would influence the distribution of word representations, and their sentiment polarity does not necessarily hold in translation contexts because of the imperfect natural language generation. These disadvantages degrade the performance of CLASC seriously because the sentiment polarity of aspect is related to specific words or phrases rather than the whole sentence.

Apart from MT-based approaches, multilingual representation is employed for cross-lingual tasks. Usually, paired sentences from parallel corpora are used to learn cross-lingual representations, eliminating the need of MT systems. However, pre-training cross-lingual language models from scratch are expensive and time-consuming. Most pre-trained cross-lingual language models rely on the large-scale task-unrelated parallel corpora and the general-purpose representations are far from satisfactory for the downstream task. Moreover, multilingual representation is not able to generalize equally well in all cases, it is constrained by typological similarity of languages (Pires et al. [6]).

In this paper, we propose the CLASC task that aims to help ASC in low-resource languages. We first build a bilingual lexicon for domain-specific training data to translate aspect. It is low cost because the number of aspect categories in a domain-specific corpus is always limited. Then, we employ Google Neural Machine Translation (GNMT) system to translate sentences into the target language. Next, we design approaches to overcome the problems caused by machine translation. Specifically, we adopt target-sensitive adversarial learning to solve domain shift problems, i.e., machine-translated training data and the target-language test data are not in exactly the same domain and genre (Duh et al. [7]). Adversarial learning has gained a lot of attention for domain adaptation by building domain-independent feature representations (Ganin et al. [8], Chen et al. [9]). Moreover, ASC is a fine-grained task which means that aspect sentiment polarity depends on specific tokens in the sentence sequence. For this, we propose a token selection model, namely Reinforced Aspect-guided Token Selector (RATS) to alleviate the effects of translation noise through discarding redundant token representations in a sentence-translation sequence. On the basis of RATS, we develop a Reinforced Transformer (In this paper, we use a transformer to denote the transformer encoder block.) with Cross-Lingual Distillation (RTCLD) approach to CLASC. Note that, we extend knowledge distillation to CLASC based on the intuition that the aspect sentiment distribution depends on semantic concepts other than specific languages and construct a well-trained source language classifier for guiding target language classifier to learn aspect-aware knowledge. On the whole, we propose a hybrid architecture, i.e., on the one hand, we combine the ASC model with MT tools for the CLASC task. On the other hand, we integrate deep learning and reinforcement learning in a single model and use a unified training framework. The main contributions in this paper are the following:

- We propose a novel approach for a CLASC task. Instead of large-scale parallel corpora, only annotated source-language corpora and source-to-target translations are required. Experiments demonstrate our approach outperforms state-of-the-art methods.
- We adopt target-sensitive adversarial learning to solve the distribution mismatch problem caused by domain shift. Furthermore, a innovative approach called RTCLD is proposed for the CLASC task, which distills aspect sentiment knowledge from the source to model aspect-aware representations in the target.

2. Related Work

2.1. Cross-Lingual Sentiment Classification

Exiting studies for Cross-Lingual Sentiment Classification (CLSC) mainly focus on document-level or sentence-level. The approaches for CLSC can be divided into MT-based approached and cross-lingual representations.

MT-based approaches translate the source language into the target language (e.g., Chinese). More sophisticated algorithms including co-training (Wan [10]; Demirtas and Pechenizkiy [11]) and multi-view learning (Xiao and Guo [12]) have been shown to improve performance. Zhou et al. [13] proposed a combination CLSC model, which adopted denoising autoencoders to enhance the robustness to translation errors of the input. Zhou et al. [14] translated each document into the other language and enforce a bilingual constraint between the original document and the translated version. Xu et al. [15] conducted cross-lingual distillation by constructing pseudo parallel corpora from the source and translation. These methods construct pseudo-parallel corpus with the source sentence and translation version but all neglect the effect of noise words.

Cross-lingual representation is another approach to a cross-lingual task. Chen et al. [16] used bilingual word embedding to map documents in the source and target languages into the same semantic space, and adversarial training was applied to enforce the trained classifier to be language-invariant. Feng et al. [17] learned sentiment-specific word representations for CLSC without any cross-lingual supervision but relied on pre-trained bilingual representation. Keung et al. [18] proposed language-adversarial training during finetuning multilingual BERT, i.e., mBERT (Devlin et al. [19]) for cross-lingual classification and named-entity recognition. Dong et al. [20] proposed a robust self-learning framework for cross-lingual text classification which is based on pre-trained cross-lingual language model.

2.2. Cross-Lingual Aspect-Level Sentiment Classification

There are few studies that focus on cross-lingual aspect-level sentiment classification. Lambert [21] employed constrained Statistical Machine Translation (SMT) to translate the opinionated units (e.g., opinion holder, opinion target or opinion phrase) and train classifiers on translated opinionated units. However, this is a resource-limited SMT system which is not available in most language combinations. Barnes and Lambert et al. [22] explored distributional representations and machine translation for aspect-based CLSC, but all of the work was based on translated opinionated units without considering the sentence context.

Unlike all the above studies, this paper performs CLASC measuring the relationship of aspect category (i.e., aspect) and the sentence. Instead of relying on translated opinionated units, our method leverages a reinforced transformer to explore diverse interactions between aspect and sentence.

3. Proposed Method

Firstly, we introduce several notations used in our approach. We have training set in source language $L_{src} = \langle \mathcal{S}, \mathcal{A}, \mathcal{Y} \rangle$, where $\mathcal{S}$ denotes sentences, $\mathcal{A}$ denotes aspect categories, $\mathcal{Y}$ denotes the sentiment labels and $L_{trans} = \langle \mathcal{S}_{tr}, \mathcal{A}_{tr}, \mathcal{Y}_{tr} \rangle$ denotes the translation version of L_{src}. We then have our test set in the target language, given by $L_{tgt} = \langle \mathcal{S}_t, \mathcal{A}_t, \mathcal{Y}_t \rangle$ and adequate target-language sentences given by $U_{tgt} = \langle \mathcal{S}_u \rangle$. We assume L_{src}, L_{tgt} and U_{tgt} are in the same domain.

In this section, we first introduce the token selection model, i.e., RATS, which functions as a fundamental module of our approach to alleviate the effects of noisy representations. On the basis of RATS, we propose an RTCLD approach for CLASC which involves the target encoder and target classifier. Then, we introduce target-sensitive adversarial learning, i.e., target discriminator to solve distribution mismatch problems between the translated training set and the test set. Note that, we introduce Gradient Reversal Layer (Ganin et al. [23]) to update the parameters involved in the target encoder and target discriminator. Finally, we introduce our optimization strategy. The overall framework of our approach is shown in Figure 1.
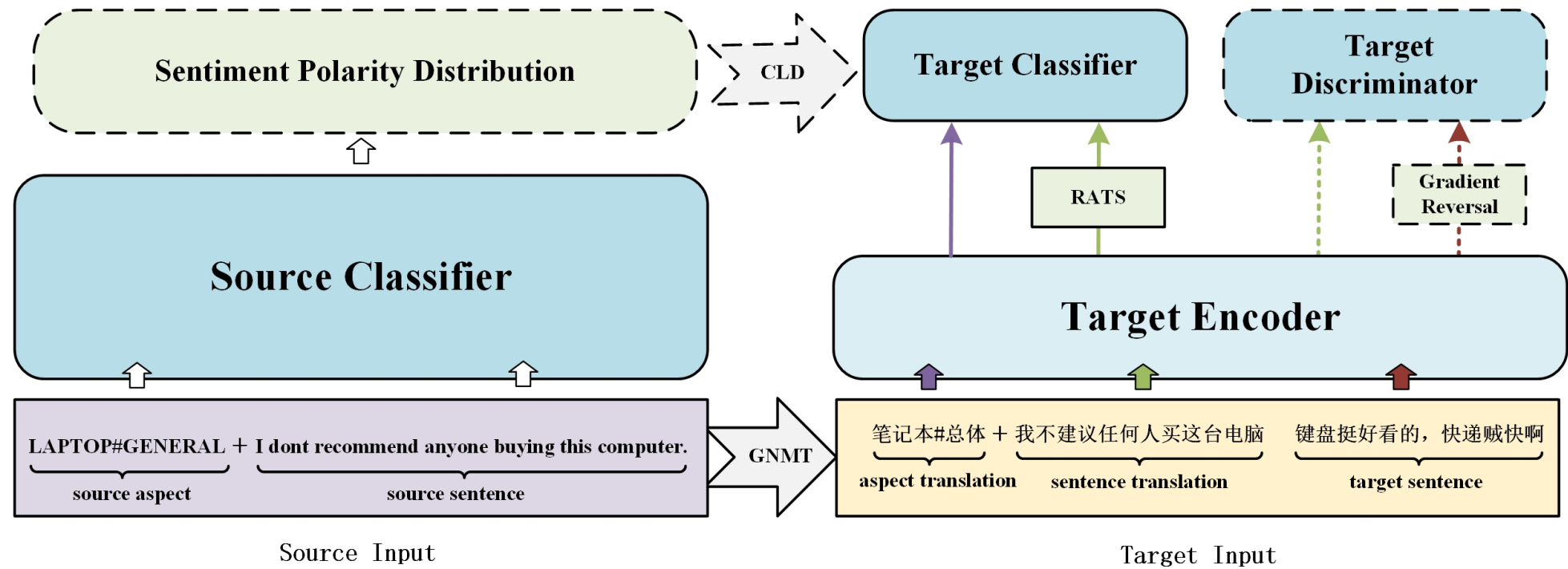
Figure 1. The overall framework of our approach. Note that we kept the parameter of source classifier constant in the training phase and the modules with dashed lines were not involved in the test phase.

3.1. Reinforced Aspect-Guided Token Selector (RATS)

Given an input sequence $x = \{x_1, x_2, ..., x_n\}$. RATS generates an equal-length sequence of binary random variables $a = \{a_1, a_2, ..., a_n\}$ where $a_i = 1$ implies that x_i is selected, whereas $a_i = 0$ indicates that x_i is discarded. The framework of RATS is shown in Figure 2.

In this way, RATS virtually functions as a hard attention mechanism to select a set of critical tokens according to a specific aspect. However, there are no ground labels to indicate whether or not a token should be selected, and the discrete random variables lead to a non-differentiable problem. Therefore, we employed the reinforcement learning algorithm, i.e., policy gradient to learn an optimal policy $\pi(a_{1:n})$ for RATS. The progress of learning the policy must rely on the local action-observation histories. To consider the full history, we adopted an LSTM network to model the policy network p_π for performing token selection over the token sequence x, denoted as LSTM_p.
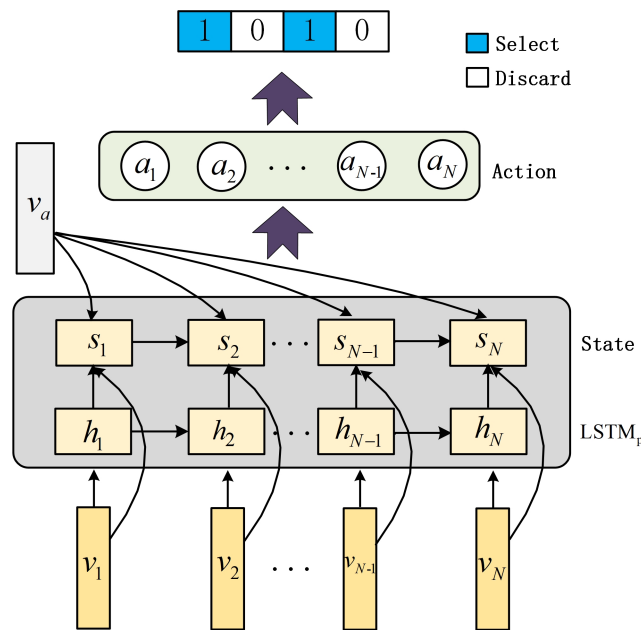
Figure 2. The framework of our proposed token selector, i.e., Reinforced Aspect-guided Token Selector (RATS).

The policy network p_π uses a **Reward** to guide the policy learning over token sequence x. It samples an **Action** a_i with the probability $p_\pi(a_i|s_i; \theta_r)$ at each **State** s_i to decide to select x_i or not. In this paper, state, action and reward are defined as follows.

● **State.** The state s_i at i-th time-step should provide adequate information for deciding to select the token x_i or not for aspect sequence x_A. Thus, the state $s_i \in \mathbb{R}^{3d}$ is composed of three parts, i.e., h_i, v_i, and v_A and the state s_i is formulated as follows:

$$s_i = h_i \oplus v_i \oplus v_A \tag{1}$$

where h_i is the hidden state of LSTM_p; v_i is the representation of token x_i; v_a is vector representation of x_A and $\oplus$ denotes vector concatenation.

● **Action.** Policy network p_π samples action $a_i \in \{0, 1\}$ with conditional probability $p_\pi(a_i|s_i; \theta_r)$. We use a logistic function to compute $p_\pi(a_i|s_i; \theta_r)$ as follows: where $\theta_r = \{W_r \in \mathbb{R}^{3d}, b_r \in \mathbb{R}\}$ is the parameter to be learned.

● **Reward.** To encourage the p_π to take better actions, we define an aspect-guided reward $\mathcal{R}$ which integrates cross-lingual distillation. Specifically, for each translated sample $(x_S, x_A, y) \in L_{trans}$ and the source version $(x'_S, x'_A, y) \in L_{src}$, the reward is formulated as follows:

$$\mathcal{R} = \log p(y|x_S, x_A; \theta_{tgt}) + p(y|x'_S, x'_A; \theta_{src}) \log p(y|x_S, x_A; \theta_{tgt}) - \gamma N'/N \tag{2}$$

where θ_{src} denotes the parameter of the source classifier and θ_{tgt} denotes the parameter of the target classifier. It's worthwhile to mention that we combine the cross-entropy of the target task with the cross-entropy of cross-lingual distillation (Section 3.2) as the delay reward. The intuition behind the definition is that integrating sentiment polarity knowledge captured by the source classifier can better guide the selector to select discriminative tokens. $\gamma N'/N$ is an additional term for limiting the number of selected tokens, where N' denotes the number of selected tokens and N denotes the number of total tokens, γ is a penalty weight (tuned to be 1×10^{-4} with the development set).

3.2. Reinforced Transformer with Cross-Lingual Distillation (RTCLD)

The fundamental idea behind this paper is that policy network functions as a hard attention mechanism to discard redundant tokens which may degrade the soft attention mechanism (e.g., self-attention) effectiveness. Based on this idea, we combined the RATS with a stacked transformer block to model interactions between aspect representations and sentence representations, i.e., RTCLD as shown in Figure 3.

In this paper, we adopted BERT as the target encoder, which calculates the token-level representations by using multi-head self-attention layers. Given a training sample $(x_S, x_A) \in L_{trans}$, where x_S denotes the token sequence of the sentence and x_A denotes the token sequence of the aspect. Let $E_S = [e_S^1, ..., e_S^N]$ denote the input representation of the sentence, $E_A = [e_A^1, ..., e_A^M]$ denotes the input representation of aspect, where N, M are the respective maximum lengths. Let $\mathbf{BERT}(\cdot)$ be the pre-trained BERT model, we can obtain the hidden representations of x_S as $H_S = [h_S^1, ..., h_S^N]$ and the hidden representations of x_A as $H_A = [h_A^1, ..., h_A^M]$. Note that, we calculate token representations of the aspect and sentence separately rather than construct a sentence pair as the whole input. In order to alleviate the effects of noisy tokens, we employ the RATS module (as introduced in Section 3.1) to perform token selection over a translated sentence, i.e.,

$$\mathbf{a} = [a_1, a_2, ..., a_n] \sim \mathbf{RATS}(H_S, v_A) \tag{3}$$

where $\mathbf{a}$ denotes the sampling result from the output of RATS, $\sim$ denotes the discrete *action* sampling operation, v_A denotes the vector representation of x_A, which is obtained by mean-pooling operation over H_A. We then append the selected representations to H_D, which denotes the denoised hidden representations of x_S.

Once we obtain the hidden representations H_A of the aspect and the denoised hidden representation H_D of the sentence, we employed an additional transformer block to model interactions between H_A and H_D, i.e.,

$$H = \mathbf{Transformer}(H_A \oplus H_D) \tag{4}$$

where $\mathbf{Transformer}(\cdot)$ was constructed by stacking L transformer blocks to generate aspect-aware hidden representations. We then employed a pooling layer and a linear layer to calculate logits $\mathbf{q} = [q_1, q_2, ..., q_K]$ of each class, where K denotes the class number. The logits are converted into probabilities of classes through the softmax layer, by normalizing each q_i with all other logits, i.e.,

$$p_i = \frac{exp(q_i/T)}{\sum_{k=1}^{K} exp(q_k/T)} \tag{5}$$

where T is a temperature and is normally set to 1. Using a higher value of T generates a softer probability distribution over classes.
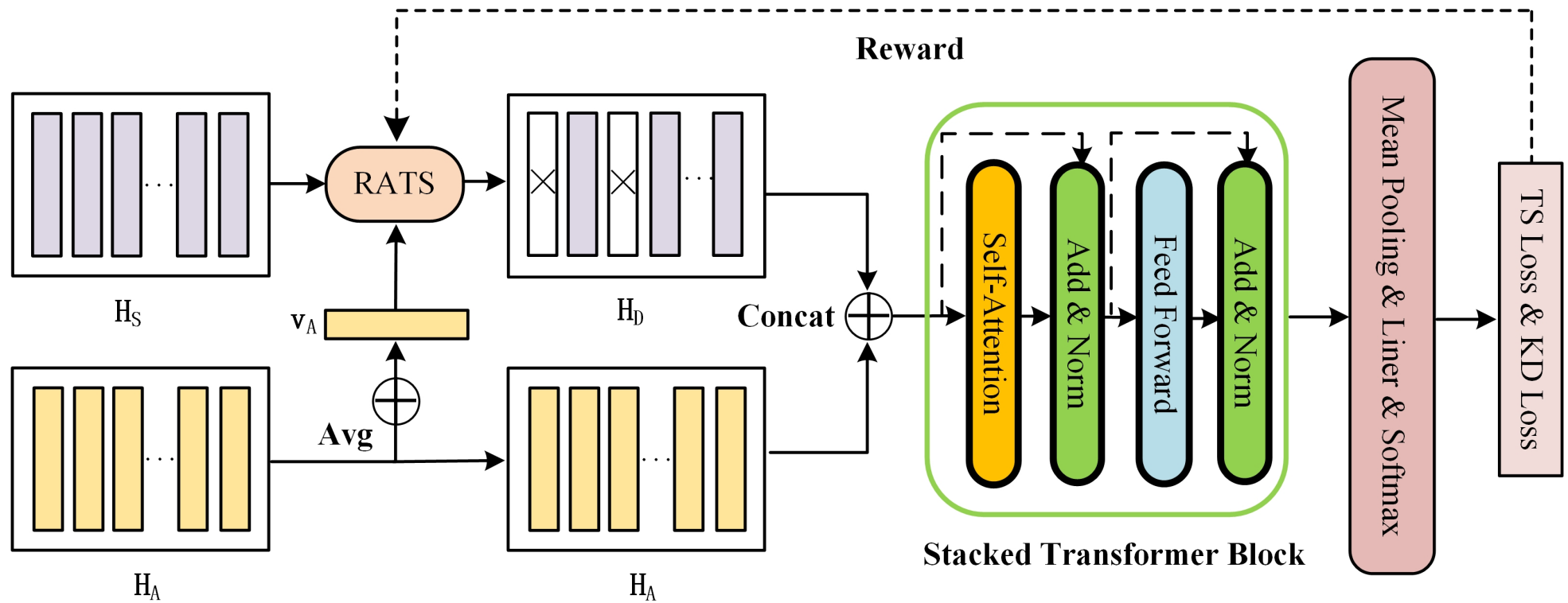
Figure 3. The framework of our proposed Reinforced Transformer with Cross-Lingual Distillation approach. TS Loss denotes task-specific loss and KD Loss denotes knowledge distillation loss.

Based on the intuition that the sentiment polarity distribution is language-independent and semantic-related, the knowledge captured in the source model can be transferred to the target model. We adopted cross-lingual distillation, which leverages a well-trained source classifier (i.e., teacher) to guide the target classifier (i.e., student). In the cross-lingual distillation framework, we first predicted a soft class distribution by source classifier with high temperature. Then, we minimized the cross-entropy of soft distributions produced by the source and target classifiers and the task-specific cross-entropy simultaneously. More formally, we optimized θ_{tgt} according to the following loss function,

$$\mathcal{J}_T = - \sum_{(x_S, x_A, y) \in L_{tgt}} \log p(y|x_S, x_A; \theta_{tgt})(1 + p(y|x'_S, x'_A; \theta_{src})) \tag{6}$$

where θ_{src} denotes the parameter of the source classifier, which has been frozen, $\theta_{tgt} = \{\theta_e, \theta_y\}$ contains the parameter θ_e of the target encoder and the parameter θ_y of the target classifier. During distillation, the same high temperature was used for training the target model and the temperature was set to $T = 6$. After it was trained, we set the temperature to 1 for testing.

3.3. Target-Sensitive Adversarial Learning

Although we leveraged a policy network to discard noisy token representations of the translated sentence and transfer knowledge from the source to target classifiers, all of the work was based on sentences in L_{src} and L_{trans}. However, we use L_{tgt} in the test phase, which have non-negligible distribution mismatch with L_{trans} because of domain shift. The distribution divergence may cause the policy network to fail to make the correct decision in the test phase, thereby reducing the overall performance.

To address this problem, we added a target discriminator module, which uses vector representations generated by the target encoder and a mean-pooling layer to classify whether the input sequence is the translation version or the native target version, i.e.,

$$\mathcal{J}_D = - \sum_{x_S \in L_{trans}} \log p(0|x_S; \theta_{adv}) - \sum_{x_S \in U_{tgt}} \log p(1|x_S; \theta_{adv}) \tag{7}$$

where $\theta_{adv} = \{\theta_d, \theta_e\}$. We then seek the parameter θ_d of the target discriminator to minimize $\mathcal{J}_D$, while simultaneously seeking the parameter θ_e of the target encoder to maximize $\mathcal{J}_D$ to adapt the target encoder to alleviate distribution divergence. In fact, we can leverage the gradient reversal layer to update those parameters by gradient descent in an objective function, i.e., the gradient reversal layer between target encoder and target discriminator could reverse the gradient w.r.t. θ_e by multiplying it by $-\lambda$ and passing it to the preceding layer. Therefore, when we minimized the loss function $\mathcal{J}_D$ of the target discriminator, the parameter θ_d is updated to minimize $\mathcal{J}_D$, while the parameter θ_e maximized $\mathcal{J}_D$ as a result of the existing gradient reversal layer. In this way, we can optimize a single objective function to achieve the adversarial training.

3.4. Optimization Strategy

The parameters in our approach were divided into two parts according to the optimization strategy, θ_r for the RATS module and θ for the remaining parts, which included the parameter θ_e for the target encoder, the parameter θ_y for the target classifier and the parameter θ_d for the target discriminator.

Optimizing θ_r was formulated as a reinforcement learning problem solved by the policy gradient method (Sutton et al. [24]). In detail, we first obtained an aspect-guided reward $\mathcal{R}$ according to Equation (2). Then, the objective of learning θ_r maximized the expected reward $\mathcal{J}_R(\theta_r)$ and the policy gradient w.r.t. θ_r was computed as follows:

$$\nabla_{\theta_r} \mathcal{J}_R(\theta_r) = -\frac{1}{D} \sum_{i=1}^{D} \sum_{t=1}^{N} \mathcal{R} \nabla_{\theta_r} \log p_\pi(a_t^{(i)}|s_t^{(i)}) \tag{8}$$

where D denotes the number of training samples and N is the length of sentences.

For θ, we optimized it with back-propagation. In detail, we sought the parameter θ_e to minimize $\mathcal{J}_T$ and maximize $\mathcal{J}_D$, while simultaneously seeking the parameter θ_y and θ_d to minimize their corresponding loss function. We can minimize $\mathcal{J}_T$ and update the parameter θ_y and θ_e by gradient descent. Likewise, we minimized $\mathcal{J}_D$ and updated the parameter θ_d and θ_e. Note that both involve the parameter θ_e, and we used λ (mentioned in Section 3.3) as a hyper-parameter to balance the relative importance. Thus, we take the sum of two loss functions as the final optimization objective, i.e.,

$$\mathcal{J}(\theta) = \mathcal{J}_T + \mathcal{J}_D \tag{9}$$

During model training, θ_r is not updated in early stage, which means that the RATS selects all tokens in the sentence sequence. When θ is optimized for several beginning epochs until the loss over development set does not decrease significantly, we begin to optimize θ_e, θ_y and θ_r simultaneously.

4. Experiment

4.1. Experimental Settings

- **Data Settings.** We conducted experiments on the SemEval 2016 Task 5 review dataset. In order to evaluate our approach on the CLASC task, we treated English as the source language, others as the target language and experiment on datasets from two domains, i.e., restaurants and laptops. To test our model, we selected parts of samples from the annotated samples in the target language as the test set and the rest as unlabeled data. In order to alleviate the effect of unbalanced data distribution, we discarded the samples with a neutral label and aspect categories, which contained less than 20 samples. It is noted that we annotated the Chinese test data in the laptops domain according to annotation guidelines of SemEval 2016, because there is no public annotated Chinese corpus in the laptops domain. The statistics of datasets are shown in Table 1.

- **Model Details.** We used pre-trained English BERT as the source classifier, i.e., teacher classifier and fine-tune pre-trained BERT on the English dataset for ASC via constructing the auxiliary sentence (Sun et al. [25]). For the target classifier, we achieved the best results when $L = 4$ and tuned the hyper-parameters on the development set. Specifically, we adopted the Adam optimizer with an initial learning rate of 1×10^{-5} for the target classifier, 1×10^{-6} for the target discriminator, 1×10^{-6} for the policy network. λ was tuned to be 5×10^{-7} for target-sensitive adversarial learning. The temperature T for cross-lingual distillation was tuned to be 6. The number of training epochs, batch size and the dropout rate was respectively set as 10, 32 and 0.3. The maximum length of the aspect and sentence inputs were set as M = 10 and N = 50, respectively.

Table 1. Corpus statistics. Note that *Asp* denotes the number of the aspect category.

Languages	Domain	Positive	Negative	All	Asp
Chinese	Laptops	1205	819	2024	20
English	Laptops	1430	923	2330	20
English	Restaurants	1657	749	2406	12
Russian	Restaurants	2533	626	3159	12
Spanish	Restaurants	1926	674	2600	12

4.2. Baselines

- **Machine Translation Baselines.** We evaluated our approach against MT-based approaches in detail. **MT-DAN** (Chen et al. [16], 2018) translated the target-language

test set into the source language and employed a Deep Averaging Network (DAN) for CLSC. **BERT** (Devlin et al. [19], 2018) translated the source-language training set into the target language and fine-tuning BERT based on a translated dataset for CLASC. **CLD-KCNN** (Xu et al. [15], 2018) conducted cross-lingual distillation by constructing pseudo-parallel corpora from the source and translation. **Dual BERT** (Cui et al. [26], 2019) leveraged cross-lingual attention mechanism to model interactions between the source and translation. Since the input of all these approaches should be a single sentence sequence, we concatenate the aspect and sentence sequence to generate a single sequence.

- **Cross-Lingual Transfer Baselines.** We also compared our approach against cross-lingual transfer approaches. **ADAN** (Chen et al. [16], 2018) leveraged bilingual representations and language-adversarial training for CLSC. **mBERT** (Devlin et al. [19], 2018), i.e., multilingual version of BERT demonstrated the ability to perform cross-lingual classification. **AmBERT** (Keung et al. [18], 2019) further improved the cross-lingual performance of mBERT via language adversarial training. **SL-mBERT** (Dong et al. [20], 2019) presented a robust self-learning framework for cross-lingual classification. Cross-lingual transfer approaches achieve comparable performance to MT-based approaches, which demonstrates the potential of multilingual representations for cross-lingual tasks.

4.3. Experimental Results

For comparison, we implemented several state-of-the-art approaches to CLASC as baselines. Table 2 shows the performances of different approaches to CLASC. As seen from the table, our proposed approach RTCLD outperforms all the baseline methods for the CLASC task.

Table 2. The results of all the methods. The best scores are shown in bold.

Approaches	EN-SP		EN-RU		EN-CH	
	Acc	F1	Acc	F1	Acc	F1
MT-DAN (Chen et al.)	0.796	0.736	0.772	0.707	0.763	0.702
CLD-KCNN (Xu et al.)	0.808	0.749	0.780	0.735	0.749	0.730
BERT (Devlin et al.)	0.830	0.801	0.790	0.775	0.758	0.757
Dual BERT (Cui et al.)	0.858	0.828	0.810	0.781	0.794	0.778
ADAN (Chen et al.)	0.826	0.747	0.806	0.776	0.765	0.740
mBERT (Devlin et al.)	0.816	0.753	0.770	0.728	0.747	0.734
AmBERT (Keung et al.)	0.828	0.784	0.790	0.752	0.781	0.767
SL-mBERT (Dong et al.)	0.842	0.790	0.818	0.769	0.794	0.769
RTCLD (ours)	**0.878**	**0.838**	**0.822**	**0.798**	**0.803**	**0.799**

In MT-based approaches, **MT-DAN** is a source-language classifier and back-translates the test set from the target language into the source language. Back-translate approaches for a cross-lingual task depend on machine translation tools during the test phase. **BERT** is trained as a target-language classifier via translating the source training set into the target language. Both of the above methods only construct monolingual classifiers without any multilingual interaction. **Dual BERT** proposes a cross-lingual attention mechanism to simultaneously model the training data in both source and target language to better exploit the relations among different languages. We can see that **Dual BERT** achieved an improvement of 2.8% (Accuracy) and 2.7% (F1) in EN-SP, 0.2% (Accuracy) and 0.6% (F1) in EN-RU, 3.6% (Accuracy) and 2.1% (F1) in EN-CH compared with **BERT**, which proves that it is beneficial to adopt a source language to improve the performance of ASC in other languages. Although these approaches are all based on machine translation, **RTCLD** significantly outperforms all the state-of-the-art approaches, which proves the significance of considering domain shift and noise words in CLASC tasks.

mBERT is another baseline that improves the performance of the cross-lingual task by leveraging multilingual representations. The results show that **mBERT** is able to perform cross-lingual generalization well and achieves an almost comparable performance to MT-based methods, which proves that multilingual representations are also effective for cross-lingual tasks. **AmBERT** leverage the addition of a language-adversarial task during fine-tuning mBERT and achieved an improvement of 1.2% (Accuracy) and 3.1% (F1) in EN-SP, 2.0% (Accuracy) and 2.4% (F1) in EN-RU, 3.4% (Accuracy) and 3.3% (F1) in EN-CH. **ADAN** also leverages the adversarial language discriminator to achieve a significant improvement compared with **MT-DAN**. Different from language adversarial learning, we propose target-sensitive adversarial learning to alleviate a distribution mismatch between the target and translation. **SL-mBERT** offers a robust self-learning framework to include target-language samples in the fine-tuning process of mBERT. Based on the cross-lingual prediction ability of mBERT, this elegantly simple framework had the best performance of all cross-lingual transfer baselines.

We can also see that the performance of above the baselines varies on different target languages, especially the cross-lingual transfer baselines. A possible explanation for this is typological similarity. English and Chinese have a different order of subject, verb and object, while English and Spanish have similar orders, and mBERT may have trouble generalizing across different orderings.

4.4. Ablation Studies

In this section, we ablate important components in our model to explicitly demonstrate its effectiveness. The ablation results are depicted in Table 3. As we can see that, **RTCLD w/o Adv** reduces the performance without target-sensitive adversarial learning, suggesting that it is beneficial to alleviate the distribution mismatch between the translation and target. **RTCLD w/o CLD** which removes cross-lingual distillation also harms overall performance, demonstrating that aspect-aware knowledge extracted from the source language can improve the performance of the target classifier. Besides, **RTCLD w/o RATS** degrades the performance significantly without a reinforced token selector, which means discarding redundant token representations is beneficial to model aspect-aware representations.

To verify the effect of noise tokens on CLASC tasks, we conducted further experiments and the results are shown in Figure 4. Most previous works (Sun et al. [25], Xu et al. [3] and so on) treat aspect sentiment classification as a sentence-pair classification task. However, when the sentence contains noise tokens, the performance of this approach will degrade significantly. To illustrate this, we compared the results of two approaches, one of which directly calculates sentence-pair representations and the other one calculates the representations of aspect and sentence separately, and then they adopt the transformer block and pooling layer to generate the final vector representation. The results show that the latter has better performance, which means that premature interactions between the aspect and sentence will reduce the quality of token representations because of the noise tokens in the sentence.

Table 3. The results of ablation studies. The best scores are shown in bold.

Approaches	EN-SP		EN-RU		EN-CH	
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**
RTCLD	**87.8**	**83.8**	**82.2**	**79.8**	**80.3**	**79.9**
RTCLD w/o Adv	85.6 (−2.2)	82.1 (−1.7)	80.6 (−1.6)	78.7 (−1.1)	78.1 (−2.2)	77.7 (−2.2)
RTCLD w/o CLD	86.2 (−1.6)	83.2 (−0.6)	81.2 (−0.1)	79.0 (−0.8)	78.7 (−1.6)	78.4 (−1.5)
RTCLD w/o RATS	85.8 (−2.0)	82.0 (−1.8)	81.8 (−0.4)	79.0 (−0.8)	77.8 (−2.5)	77.6 (−2.3)
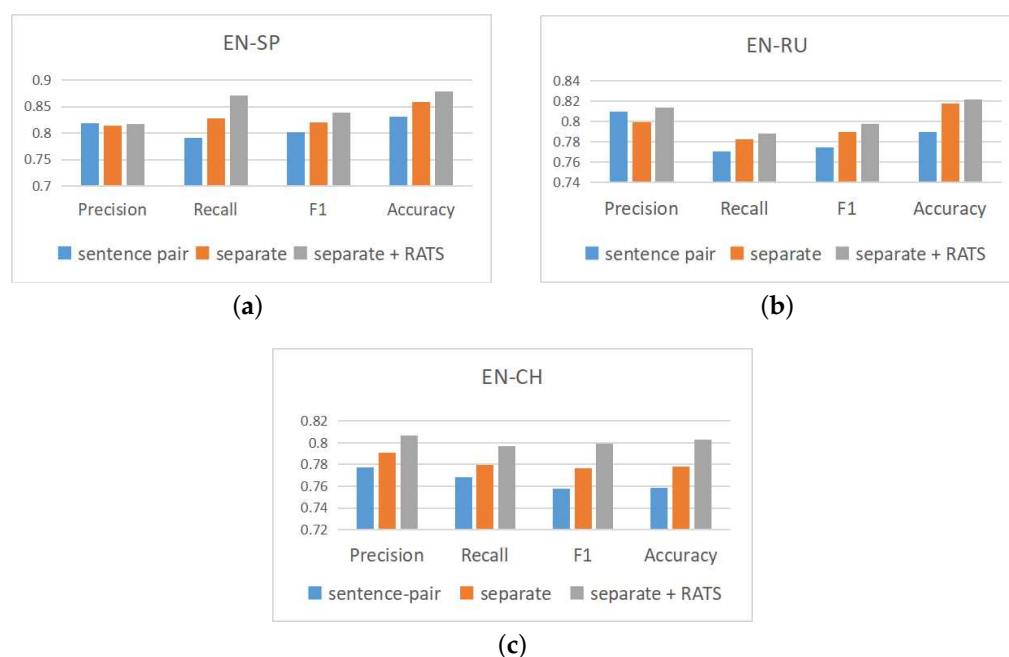
Figure 4. Comparison of models trained on sentence-pair input, separate input, and separate input with RATS.

4.5. Error Analysis

We randomly analyze 100 error cases in the experiments, which can be roughly categorized into 5 types. (1) 27% errors are because of the differences in expression. An example is "太坑了(so bad)". Our approach fails to understand "坑" means bad. (2) 24% errors were due to negation words. An example is "the price is not cheap". Our approach failed to select the word "not" and incorrectly predicted positive polarity. This inspires us to optimize our approach so as to capture the negation scope better in the future. (3) 21% errors were due to implicit opinion words. An example is "I spent 2200 dollars on a "top of the line laptop"". Our approach incorrectly predicted positive for aspect "price". (4) 19% errors were due to the wrong prediction for recognizing neutral instances. The shortage of neutral training examples made the prediction of neutral instances very difficult. (5) Finally, 9% errors were due to the sentence being too long and the opinion word was truncated when constructing the input.

5. Conclusions

In this study, we proposed a Reinforced Transformer with Cross-Lingual Distillation, i.e., RTCLD approach for the CLASC task. Specifically, we adopted target-sensitive adversarial learning to adapt target encoder to alleviate distribution mismatch and Reinforced Aspect-guided Token Selector (i.e., RATS) to discard redundant token representations. On the basis of RATS, we adopt the transformer block with cross-lingual distillation to generate an aspect-aware representation. The experimental results show that our proposed method outperforms several state-of-the-art baselines. To date, there are few studies that focus on CLASC tasks, the main limitations for cross-lingual tasks are the scarce resources. In fact, a large number of parallel corpora without annotations or a small number of target language corpora with annotations are all beneficial to cross-lingual tasks. Unlike document-level or sentence-level annotation, aspect-level annotation work is more difficult and time-consuming, and detailed annotation specifications need to be developed. In our future work, we intend to solve other challenges in aspect-level cross-lingual sentiment analysis such as term-based cross-lingual sentiment classification and cross-lingual aspect term extraction.

Author Contributions: Conceptualization, H.W., F.Q. and S.L.; Formal analysis, H.W.; Methodology, Z.W. and F.Q.; Resources, S.L.; Software, Z.W.; Supervision, H.W. and F.Q.; Writing—original draft, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This work is supported in part by the Industrial Prospective Project of the Jiangsu Technology Department under grant BE2017081 and the National Natural Science Foundation of China under grant 61572129.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ma, D.; Li, S.; Zhang, X.; Wang, H. Interactive Attention Networks for Aspect-Level Sentiment Classification. In Proceedings of the International Joint Conferences on Artifical Intelligence(IJCAI), Melbourne, VIC, Australia, 19–25 August 2017; pp. 4068–4074.
2. Wang, Y.; Huang, M.; Zhu, X.; Zhao, L. Attention-based LSTM for Aspect-level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 606–615.
3. Xu, H.; Liu, B.; Shu, L.; Yu, P.S. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 6–7 June 2019; pp. 2324–2335.
4. Lo, S.L.; Cambria, E.; Chiong, R.; Cornforth, D. Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artif. Intell. Rev.* **2017**, *48*, 499–527. [CrossRef]
5. Fu, Y.; Hospedales, T.M.; Xiang, T.; Gong, S. Transductive Multi-View Zero-Shot Learning. *IEEE Trans.* **2015**, *37*, 2332–2345. [CrossRef] [PubMed]
6. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual is Multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July 2019; pp. 4996–5001.
7. Duh, K.; Fujino, A.; Nagata, M. Is Machine Translation Ripe for Cross-Lingual Sentiment Classification? In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 429–433.
8. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V.S. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
9. Chen, X.; Cardie, C. Multinomial Adversarial Networks for Multi-Domain Text Classification. In Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LO, USA, 1–6 June 2018; pp. 1226–1240.
10. Wan, X. Co-Training for Cross-Lingual Sentiment Classification. In Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, CA, USA, 11–17 July 2009; pp. 235–243.
11. Demirtas, E.; Pechenizkiy, M. Cross-lingual polarity detection with machine translation. In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, Chicago, IL, USA, 11 August 2013; pp. 91–98.
12. Xiao, M.; Guo, Y. Multi-View AdaBoost for Multilingual Subjectivity Analysis. In Proceedings of the 24th International Conference on Computational Linguistics, Mumbai, India, 11–17 December 2012; pp. 2851–2866.
13. Zhou, H.; Chen, L.; Huang, D. Cross-Lingual Sentiment Classification Based on Denoising Autoencoder. In Proceedings of the conference on Natural Language Processing and Chinese Computing, Shenzhen, China, 5–9 December 2014; pp. 181–192.
14. Zhou, X.; Wan, X.; Xiao, J. Cross-Lingual Sentiment Classification with Bilingual Document Representation Learning. In Proceedings of the annual meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 481–492.
15. Xu, R.; Yang, Y. Cross-lingual Distillation for Text Classification. In Proceedings of the annual meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1415–1425.
16. Chen, X.; Sun, Y.; Athiwaratkun, B.; Cardie, C.; Weinberger, K.Q. Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 557–570. [CrossRef]
17. Feng, Y.; Wan, X. Learning Bilingual Sentiment-Specific Word Embeddings without Cross-lingual Supervision. In Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 6–7 June 2019; pp. 420–429.
18. Keung, P.; Lu, Y.; Bhardwaj, V. Adversarial Learning with Contextual Embeddings for Zero-resource Cross-lingual Classification and NER. In Proceedings of the EMNLP-IJCNLP-2019, Hong Kong, 3–7 November 2019; pp. 1355–1360.
19. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT-2019, Minneapolis, MN, USA, 6–7 June 2019; pp. 4171–4186.
20. Dong, X.; de Melo, G. A Robust Self-Learning Framework for Cross-Lingual Text Classification. In Proceedings of the EMNLP-IJCNLP-2019, Hong Kong, 3–7 November 2019; pp. 6305–6309.

21. Lambert, P. Aspect-Level Cross-lingual Sentiment Classification with Constrained SMT. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguisticsand the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 781–787.

22. Barnes, J.; Lambert, P.; Badia, T. Exploring Distributional Representations and Machine Translation for Aspect-based Cross-lingual Sentiment Classification. In Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016; pp. 1613–1623.

23. Ganin, Y.; Lempitsky, V.S. Unsupervised Domain Adaptation by Backpropagation. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1180–1189.

24. Sutton, R.S.; McAllester, D.A.; Singh, S.P.; Mansour, Y. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In Proceedings of the 12th International Conference on Neural Information Processing Systems, Denver, CO, USA, 29 November–4 December 1999; pp. 1057–1063.

25. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 6–7 June 2019; pp. 380–385.

26. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; Hu, G. Cross-Lingual Machine Reading Comprehension. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, 3–7 November 2019; pp. 1586–1595.