Cross-lingual Aspect-based Sentiment Analysis with Aspect Term Code-Switching *

Wenxuan Zhang[1], Ruidan He[2], Haiyun Peng[2], Lidong Bing[2] and Wai Lam[1]
[1]The Chinese University of Hong Kong
[2]DAMO Academy, Alibaba Group
{wxzhang,wlam}@se.cuhk.edu.hk
{ruidan.he,haiyun.p,l.bing}@alibaba-inc.com

Abstract

Many efforts have been made in solving the Aspect-based sentiment analysis (ABSA) task. While most existing studies focus on English texts, handling ABSA in resource-poor languages remains a challenging problem. In this paper, we consider the unsupervised cross-lingual transfer for the ABSA task, where only labeled data in the source language is available and we aim at transferring its knowledge to the target language having no labeled data. To this end, we propose an alignment-free label projection method to obtain high-quality pseudo-labeled data of the target language with the help of the translation system, which could preserve more accurate task-specific knowledge in the target language. For better utilizing the source and translated data, as well as enhancing the cross-lingual alignment, we design an aspect code-switching mechanism to augment the training data with code-switched bilingual sentences. To further investigate the importance of language-specific knowledge in solving the ABSA problem, we distill the above model on the unlabeled target language data which improves the performance to the same level of the supervised method.

1 Introduction

Aspect-based Sentiment Analysis (ABSA) is the task of extracting mentioned aspects from a given sentence and predicting their corresponding sentiment polarities[1] (Liu, 2012; Pontiki et al., 2014). Consider the following example, "*The food is great, but the service is kinda disappointing*", we can detect two mentioned aspect terms "*food*" and "*service*", and judge their corresponding sentiments as positive and negative, respectively. Given its wide application scenarios, it has attracted lots of attention in the NLP community in recent years (Li et al., 2019a; He et al., 2019; Hu et al., 2019; Chen and Qian, 2020; Liang et al., 2020; Mao et al., 2021; Zhang et al., 2021).

The majority of existing ABSA studies are conducted on English texts. However, in real-world scenarios such as the E-commerce website, users' opinions are usually expressed in different languages (Pontiki et al., 2016; Keung et al., 2020). Manually annotating a large quantity of ABSA data for every language can be extremely costly. In this work, we investigate the unsupervised cross-lingual transfer for the ABSA task, where we only have labeled data in the source language and aim to transfer the knowledge to target languages whose labeled ABSA data is unavailable.

Existing works on cross-lingual ABSA mainly focus on its subtasks, including cross-lingual aspect term extraction and aspect sentiment classification. Early studies usually adopt a *translate-then-align* strategy: a machine translation system is first used to translate the source sentence to the target language. Then, word alignment algorithms (Dyer et al., 2013) are applied to project the labels (*i.e.*, the position of the aspect term) to obtain labeled target language data (Zhou et al., 2015; Klinger and Cimiano, 2015). Later methods make use of the cross-lingual word embeddings (Ruder et al., 2019) trained on large parallel corpus to allow the model to be used in a language-independent manner, by simply switching the word embedding layer while keeping the model unchanged (Barnes et al., 2016; Akhtar et al., 2018; Jebbara and Cimiano, 2019) when adopted for different languages.

Recently, employing multilingual pre-trained models such as the multilingual BERT (Devlin et al., 2019) and XLM-Roberta (Conneau et al., 2020) has become the *de-facto* approach to tackle the cross-lingual transfer for many NLP tasks (Hu et al., 2020). Typically, the model is first fine-tuned

* Work done when Wenxuan Zhang was an intern at Alibaba. This work was supported by Alibaba Group through Alibaba Research Intern Program, and a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Codes: 14204418).

[1]Also called End-to-End ABSA or Unified ABSA

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9220–9230
November 7–11, 2021.

on labeled source language data and then can be directly used for inference on the target language data (*i.e.*, zero-shot approach), thanks to the language knowledge learned in the pre-training stage (Wu and Dredze, 2019; K et al., 2020).

There are some challenges for adopting such a paradigm to solve the cross-lingual ABSA task. The language-specific knowledge plays an essential role in tackling the ABSA problem, since the concerned texts are often written by ordinary users with all kinds of abbreviations or slang. The aspect terms and the opinion expressions may also be language-dependent. However, the language-specific knowledge of the zero-shot method purely comes from the pre-training process where the low-resource languages might be under-represented (Conneau et al., 2020; Pfeiffer et al., 2020). Utilizing the translated target language data with projected labels is a plausible idea to compensate the language-specific knowledge (Li et al., 2020). But the performance of such translation-based methods largely depends on the quality of the translation and label projection. The task-specific knowledge in the translated data would also be limited if the projected label quality is unsatisfactory.

To this end, we propose an *alignment-free* label projection method to obtain high-quality pseudo-labeled target language data. Different from the previous *translate-then-align* paradigm, our method does not rely on any word alignment tool for projecting the labels from the source to the translated target sentence, which avoids the mis-alignment issue brought in by this step. The high-quality labeled target data thus preserves more task-specific knowledge, helping establish a strong baseline by purely training on such pseudo-labeled data. Previous findings suggest that training on the bilingual corpus (*i.e.*, labeled source data and translated target data) often leads to better performance in cross-lingual transfer tasks (Hu et al., 2020). Inspired by this finding and to further enhance the interactions between the two languages, we propose an aspect code-switching (ACS) mechanism, which switches the aspect terms between the source and translated target sentences to construct two bilingual sentences. By training on the combinations of the monolingual sentences of the source/target languages and the code-switched bilingual sentences, the embedding space between the source and target languages can be better aligned with aspects as the anchor. It is natural to further extend our ACS method to the multilingual setting, assuming multiple translation engines are available. In this case, the target languages can benefit from the task-specific knowledge contained in different translations.

To further verify the importance of the language-specific knowledge for the cross-lingual ABSA task, we exploit the usage of the unlabeled target language data via knowledge distillation (Hinton et al., 2015). Concretely, we treat the proposed ACS method as the teacher model for predicting the probability distributions on the unlabeled target data. Then, a student model is trained with such soft-labeled data. The distilled student model can thus utilize both the task-specific knowledge from the teacher model and the language-specific knowledge from the unlabeled target language data.

In summary, our main contributions are: (1) We establish a strong translation-based baseline for the cross-lingual ABSA task based on an *alignment-free* label projection approach, which outperforms previous *translate-then-align* paradigm by a large margin. (2) We propose an aspect code-switching mechanism that makes better use of the source data and the translated target data via aspect terms as anchors. (3) We show that language-specific knowledge is essential for tackling the cross-lingual ABSA task. By distilling the proposed model on the unlabeled target data, the performance can be further improved. (4) We conduct extensive experiments on benchmark datasets of five languages and our method achieves new state-of-the-art results under both cross-lingual and multilingual settings.[2]

2 Methodology

2.1 Problem Formulation

We formulate the ABSA task as a sequence labeling problem (Li et al., 2019b; He et al., 2019). Given a sentence $\boldsymbol{x} = \{x_i\}_{i=1}^{L}$ with L tokens, the model predicts a label sequence $\boldsymbol{y} = \{y_i\}_{i=1}^{L}$ where $y_i \in \mathcal{Y} = \{\texttt{B}, \texttt{I}, \texttt{E}, \texttt{S}\}\text{-}\{\texttt{POS}, \texttt{NEU}, \texttt{NEG}\} \cup \{\texttt{O}\}$ denotes the aspect boundary and its sentiment polarity for the corresponding token x_i. For example, $y_i = \texttt{B-POS}$ means x_i is the beginning of a positive aspect term. In the cross-lingual transfer setting, we only have the sentence-label pair in the source language S, *i.e.*, $(\boldsymbol{x}^S, \boldsymbol{y}^S) \in D_S$ and aim to predict the label sequence $\boldsymbol{y}^T$ for the sentence $\boldsymbol{x}^T$ in the target language T.

[2]Our code is publicly available at https://github.com/IsakZhang/XABSA.

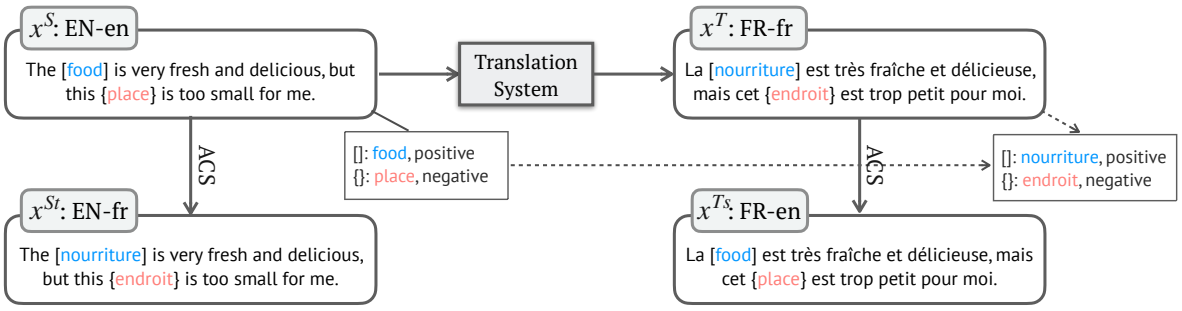

Figure 1: Example of the *alignment-free* label projection method (upper part) and the aspect code-switching strategy (lower part). Here we use English and French as the source and target language respectively.

2.2 Alignment-free Label Projection

To obtain the language-specific knowledge for the target languages, previous works usually first translate the source sentence with an off-the-shelf translation system, then word alignment tools such as fastAlign (Dyer et al., 2013) are used to map the token-level label from the source sentence to the translated sentence (Mayhew et al., 2017; Fei et al., 2020). Some heuristics are proposed for alleviating the alignment error, for example, by conducting a phrase-to-phrase mapping to refine the aspect boundary (Klinger and Cimiano, 2015; Li et al., 2020). However, the word or phrase alignment itself is a challenging task (Akbik and Vollgraf, 2018). The sentences of the ABSA task are usually user-generated (*e.g.*, product reviews and tweets) and informal, which further hinders the *translate-then-align* method to produce satisfactory pseudo-labeled target data (Lohar et al., 2019). The inaccurate pseudo labels inevitably limit the task-specific knowledge and lead to poor model performance.

We propose an *alignment-free* label projection method for obtaining the pseudo-labeled data in the target language[3]. As depicted in the upper-left portion of Figure 1, we first mark each aspect term in the sentence with a special symbol (*e.g.*, different brackets like "{}" and "[]"), before feeding it into the translation system. If there are multiple aspect terms in one sentence, we mark them orderly with the predefined special symbol list. After getting the translation in the upper-right portion, we extract the spans with the special symbols and match them with the corresponding aspect terms in the source sentence so as to recover the aspect boundary and project the sentiment labels. Thus, a labeled sentence in the target language is obtained.

Formally, suppose in the source sentence, the i-th to j-th tokens $\boldsymbol{x}^S_{i:j}$ are marked with the special symbol t. After our label projection, we will label the span in the translated sentence $\boldsymbol{x}^T$ marked by the same special symbol t with the sentiment polarity of $\boldsymbol{x}^S_{i:j}$. Since we use different symbols for different aspect terms, the translated aspect terms would still be matched to the corresponding labels even if their orders are changed during the translation. In some cases, the special symbols might be ignored by the translation system. We thus count the number of special symbols after the translation and filter out those translated sentences with missing symbols[4]. We denote the pseudo-labeled training data of the target language as D_T.

2.3 Aspect Code-Switching

Explicitly mixing the data from different languages was proved effective for enhancing the cross-lingual capability (Singh et al., 2019). Therefore, after obtaining the labeled target language data, the combination of the source and target data is a natural choice to train models. We further design a fine-grained aspect term code-switching (ACS) mechanism to switch the aspect in the source and the translated data for better utilizing the available data to enhance the cross-lingual capability.

As shown in the lower portion of Figure 1, given the source sentence $\boldsymbol{x}^S$ and its translation $\boldsymbol{x}^T$, we switch the aspect terms in them to construct two bilingual sentences: The first one is derived from $\boldsymbol{x}^S$ with aspect terms now in the target language, denoted as $\boldsymbol{x}^{St}$; the other one is derived from $\boldsymbol{x}^T$ with aspect terms in the source language, denoted as $\boldsymbol{x}^{Ts}$. After the switching, we refine the corre-

[3]Note that any translation-based method is not truly alignment free, since the translation engines are trained with sentence pairs which contain implicit word alignments. Here the "alignment-free" refers to the absence of using word alignment tools during the label projection.

[4]This kind of cases only accounts for a very small amount in total, about 1% in our experiment.

sponding label sequences $\boldsymbol{y}^{St}$ and $\boldsymbol{y}^{Ts}$ according to the length of the switched aspect terms. We denote the constructed code-switched datasets as D_{St} and D_{Ts} respectively.

We can see that there are some interesting relations among these data. For example, $\boldsymbol{x}^{T}$ and $\boldsymbol{x}^{Ts}$ have the same sentence context but aspect terms in different languages; while $\boldsymbol{x}^{T}$ and $\boldsymbol{x}^{St}$ have different sentence context but the same aspect term in the target language. By training on the combination of these data, the embedding space of different languages can be better aligned via aspects as anchors.

Furthermore, assuming the translation system can translate from the source to multiple target languages, we can extend the proposed method to the multilingual setting. For the source sentence $\boldsymbol{x}^{S}$, we can obtain translations $\boldsymbol{x}^{T_1}, \boldsymbol{x}^{T_2}, \ldots, \boldsymbol{x}^{T_n}$ in n target languages. Similarly, their aspect terms can be switched with the source sentence, resulting in code-switched sentences $\boldsymbol{x}^{T_1 s}, \boldsymbol{x}^{T_2 s}, \ldots, \boldsymbol{x}^{T_n s}$ and $\boldsymbol{x}^{St_1}, \boldsymbol{x}^{St_2}, \ldots, \boldsymbol{x}^{St_n}$, respectively. For the i-th target language, we denote the translated data and code-switched data as D_{T_i}, $D_{T_i s}$, and D_{St_i} respectively. Since the translation and projection process is not flawless, the pseudo-labeled data in multiple target languages may compensate for each other. Therefore, the model trained on the combination of these data can benefit from the task-specific knowledge from multiple languages.

2.4 ABSA Model

We build our ABSA model with the pre-trained multilingual models as backbones. Given a text sequence $\boldsymbol{x} = \{x_i\}_{i=1}^{L}$ with L words, the backbone network encodes it into context-aware feature representations $\boldsymbol{h} = \{h_i\}_{i=1}^{L}$ with $h_i \in \mathbb{R}^{d_h}$ where h_i is the hidden feature representation for the corresponding token x_i and d_h is the dimension of the vector representation. We employ a simple linear classification layer built on top of the backbone to make the prediction. With the token representation h_i, the label distribution of x_i is computed as:

$$p_\theta(y_i|x_i) = \text{softmax}(W h_i + b) \qquad (1)$$

where $y_i \in \mathcal{Y}$, θ denotes all the parameters to be learned or fine-tuned, including task-specific ones (*i.e.,* W and b) and those of the backbone network.

Given a labeled training dataset $D = \{(\boldsymbol{x}, \boldsymbol{y})\}$, the training objective $\mathcal{L}_{CE}$ is computed as the cross-entropy loss between the predicted label distributions and the gold label in one-hot encoding:

$$\mathcal{L}_{CE} = \frac{1}{|D|} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in D} \left[-\frac{1}{L} \sum_{i=1}^{L} y_i \log p_\theta(y_i|x_i) \right] \qquad (2)$$

For our proposed ACS method in Sec 2.3, we can thus train a model using the combination of the bilingual data, as well as the code-switched data, *i.e.*, $D = D_S \cup D_T \cup D_{Ts} \cup D_{St} \equiv D_{ACS}$. In the multilingual setting, the code-switched data across multiple languages is used to train a multilingual model, *i.e.*, $D = D_S \cup \{D_{T_i}\}_{i=1}^{n} \cup \{D_{T_i s}\}_{i=1}^{n} \cup \{D_{St_i}\}_{i=1}^{n} \equiv D_{MACS}$, making it accessible to the knowledge from different language data.

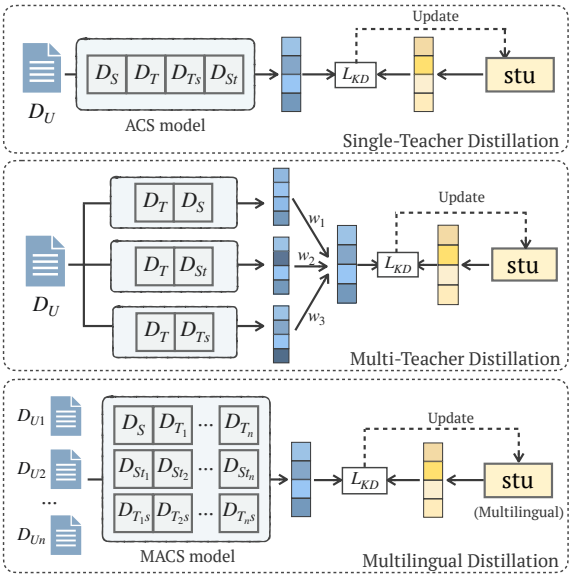

Figure 2: Distillation on the unlabeled target data.

2.5 Distillation on Unlabeled Target Data

The texts in the ABSA problem often involve many language-dependent expressions. To further investigate the importance of the language-specific knowledge in the cross-lingual ABSA task, we exploit to distill our proposed model on the unlabeled target data to utilize its rich language knowledge.

Single-teacher Distillation As shown in the upper portion of Figure 2, we treat the model trained with the data D_{ACS} as the teacher model which contains rich task-specific knowledge. Let $\bar{\boldsymbol{x}} \in D_U$ denotes an unlabeled sentence in the target language, we can obtain the soft labels produced by the teacher model and use it to distill a student model whose predicted probability distribution for the i-th token $p_{\theta_{stu}}(\bar{y}_i|\bar{x}_i)$ should approximate the

soft labels $p_{\theta_{tea}}(\bar{y}_i|\bar{x}_i)$ as follows:

$$\mathcal{L}_{KD} = \frac{1}{|D_U|} \sum_{\bar{\boldsymbol{x}} \in D_U} \left[\frac{1}{L} \sum_{i=1}^{L} \text{MSE}(p_{\theta_{tea}}, p_{\theta_{stu}}) \right] \tag{3}$$

Here we use the mean squared error loss ($\text{MSE}(\cdot)$) to measure the difference between the two probability distributions. Note that the soft labels (*i.e.*, predicted distributions) instead of the hard labels (*i.e.*, one-hot predictions after taking argmax) are used since the former contains much richer information than the latter (Hinton et al., 2015).

To produce $p_{\theta_{stu}}(\bar{y}_i|\bar{x}_i)$, the student model can actually take any form since we only aim to make its predictions mimic the soft labels. For simplicity, we use the same neural architecture described in Sec 2.4 as the student model. Finally, the trained student model can be used to make the prediction on the unseen data of the target language.

Multi-teacher Distillation As discussed previously, the relations between the code-switched data and bilingual data show some interesting characteristics. Specifically, we consider three types of combinations, including $D_1 = D_T \cup D_S$, $D_2 = D_T \cup D_{Ts}$, and $D_3 = D_T \cup D_{St}$. Each contains the translated dataset D_T to involve some language-specific knowledge and another dataset sharing the same sentence semantics (D_1), the same context sentence (D_2), and the same aspect term (D_3), respectively. To fully utilize the different characteristics in those datasets, we design a multi-teacher distillation with different teacher models separately trained on those three combinations, as shown in the middle portion of Figure 2. Let $p_{\theta_k}(\bar{y}_i|\bar{x}_i)$ denote the probability distributions of the unlabeled sentence $\bar{\boldsymbol{x}}$ given by the k-th model trained on D_k, we combine them into a single soft label:

$$p_{\theta_{tea}}(\bar{y}_i|\bar{x}_i) = \sum\nolimits_{k=1}^{3} w_k \; p_{\theta_k}(\bar{y}_i|\bar{x}_i) \tag{4}$$

where w_k is the weight for each teacher model. With the combined soft label $p_{\theta_{tea}}$, a student model can be trained in a similar way as Equation 3.

Multilingual Distillation In the multilingual setting, we distill a multilingual model with the unlabeled data from multiple target languages. As shown in the lower portion of Figure 2, we treat the model trained on the multilingual code-switched data D_{MACS} as the teacher and compute soft labels for each sentence $\bar{\boldsymbol{x}}$ in the unlabeled data. Then a student model can be trained with such soft labels and used to make predictions for multiple target languages in a "*one model for all*" manner.

		EN	FR	ES	NL	RU
Train	# S	2000	1664	2070	1722	3655
	# A	1743	1641	1856	1231	3077
Test	# S	676	668	881	575	1209
	# A	612	650	713	373	949

Table 1: Statistics of the data in each language. # S and # A denotes the numbers of sentences and aspects in each set respectively.

3 Experiments

3.1 Dataset

We conduct experiments on the SemEval-2016 dataset (Pontiki et al., 2016), which includes real user reviews in English (EN), French (FR), Spanish (ES), Dutch (NL), and Russian (RU)[5]. The data in each language is already split into training and testing sets. We keep the original split and further sample 20% data from the training set as the validation set for model selection. We provide the summary data statistics in Table 1.

We treat English as the source language and other languages as targets. Following existing works to simulate true unsupervised setting (Jebbara and Cimiano, 2019; Hu et al., 2020), we use the English validation set in all experiments for the model selection. The original workshop also provides training data for each target language as well[6], we thus discard the label of the training set in each target language and use the raw sentences as the unlabeled data, similar with previous studies (Wang and Pan, 2018; Wu et al., 2020).

3.2 Experimental Settings

We conduct experiments based on two multilingual pre-trained models, including the cased multilingual BERT (**mBERT**) and the base XLM-Roberta model (**XLM-R**). Google translate API[7] is used for the translation process described in Sec 2.2. For the teacher model training, following Li et al. (2019b), we train the model based on mBERT and XLM-R

[5]There is one more language data namely Turkish is provided in the SemEval workshop. However, we leave it out in the experiments due to its extremely small testing set (less than 150 sentences).

[6]Note that the data for each target language is collected separately, which means they are not the "gold translations" of the source English reviews.

[7]https://translate.google.com/

Methods	mBERT					XLM-R				
	FR	ES	NL	RU	Avg	FR	ES	NL	RU	Avg
SUPERVISED	61.80	67.88	56.80	58.87	61.34	67.44	71.93	64.28	64.93	67.15
ZERO-SHOT†	45.60	57.32	42.68	36.01	45.40	56.43	67.10	59.03	56.80	59.84
TRANSLATION-TA†	40.76	50.74	47.13	41.67	45.08	47.00	58.10	56.19	50.34	52.91
BILINGUAL-TA†	41.00	51.23	49.72	43.67	46.41	49.34	61.87	58.64	52.89	55.69
TRANSLATION-AF	48.03	59.74	49.73	50.17	51.92	57.07	66.61	61.26	59.55	61.12
BILINGUAL-AF	48.05	60.23	49.83	51.24	52.34	57.91	68.04	60.80	60.81	61.89
ACS	49.65	59.99	51.19	52.09	53.23	59.39	67.32	62.83	60.81	62.59
ACS-DISTILL-S	52.23	62.04	52.72	53.00	55.00	**61.00**	68.93	62.89	60.97	63.45
ACS-DISTILL-M	**52.25**	**62.91**	**53.40**	**54.58**	**55.79**	59.90	**69.24**	**63.74**	**62.02**	**63.73**

Table 2: Main results of the cross-lingual ABSA task with English as the source language. We report results with the average F1 scores over five runs with different random seeds. The first row "SUPERVISED" gives results under the supervised setting, provided as an upper bound. † denotes results are from Li et al. (2020).

up to 2000 and 2500 steps respectively and conduct model selection on the last 500 steps. For the student model, we initialize it with the model parameters trained on the translated target data for a good starting point, and then continue the training on the soft labels by the teacher model. For the multi-teacher distillation, we treat each teacher equally, which means $w_i = 1/3$ in Equation 4.

We select the best training hyper-parameters by conducting a grid search on a combination of batch size and learning rate. The range of them are: learning rate {2e-5, 3e-5, 5e-5}; batch size {8, 16, 25}. The best choices are selected by the performance on the source language data. For mBERT, we use a learning rate being 5e-5 and the batch size being 16; for XLM-R, we use the learning rate of 2e-5 and the batch size being 8.

Micro-F1 is employed as the evaluation metric where a prediction will be judged as correct only if both its boundary and sentiment polarity are correct. For all experiments, we report the average F1 scores over 5 runs with different random seeds.

3.3 Compared Methods

We adopt the following approaches for comparisons and also revealing the characteristics of the cross-lingual ABSA task: ZERO-SHOT, a method utilizing the labeled source data to fine-tune the model and directly conduct inference on the target data, which has shown to be a strong baseline for the cross-lingual adaptation (Wu and Dredze, 2019; Conneau et al., 2020). To compare with the previous translation-based method, we adopt the baseline that utilizes the pseudo-labeled data with the *Translate-then-Align* paradigm (Klinger and Cimiano, 2015; Li et al., 2020) (TRANSLATION-TA) and the combination of the source data with such translated data (BILINGUAL-TA).

For our method, we report the performance of the model trained on the pseudo-labeled target data with the proposed *Alignment-Free* label projection method (TRANSLATION-AF) and the combination of the translated data and source language data (BILINGUAL-AF); the results of the aspect code-switching method (ACS) trained on D_{ACS} as described in Sec 2.4; the results with the single-teacher distillation (ACS-DITILL-S, *i.e.* the upper model in Figure 2) and multi-teacher distillation (ACS-DISTILL-M, *i.e.* the middle model in Figure 2) introduced in Sec 2.5.

In addition to the above cross-lingual setting (*i.e.* from one source to one target), we also evaluate the multilingual setting. We mainly compare with MTL-WS (Li et al., 2020), a previous state-of-the-art method using a parameter warm-up mechanism. For our proposed method, we report the results using the combination of the source data and the multilingual translated data with the *alignment-free* label projection paradigm (MTL-AF); the results with our multilingual aspect code-switching method (MTL-ACS) trained on D_{MACS} as described in Sec 2.4, and the results with multilingual distillation (MTL-ACS-D, *i.e.* the lower model in Figure 2) introduced in Sec 2.5.

We also present the results under the supervised setting (SUPERVISED) where the model is trained with the training data in the corresponding language. It provides an upper bound for us to measure the cross-lingual transfer performance.

3.4 Cross-lingual ABSA Results

We present the cross-lingual ABSA results in Table 2. There are some notable observations: 1) The ZERO-SHOT method based on mBERT is relatively weak, while with the XLM-R backbone, it becomes a competitive baseline. The main reason might be that XLM-R was pre-trained on a much larger multilingual corpus, leading to richer language knowledge and cross-lingual ability. 2) BILINGUAL-TA which is trained with both the source data and labeled target data from the *translate-then-align* paradigm performs even worse than the ZERO-SHOT method, implying that the low-quality target data actually limits the task-specific knowledge and hurts the performance. 3) Our proposed *alignment-free* label projection method ("TRANSLATION-AF") establishes a strong baseline for the cross-lingual ABSA problem. It not only outperforms the ZERO-SHOT method based on either mBERT or XLM-R; but also achieves much better performance than the previous translation-based approach. Compared with TRANSLATION-TA, it obtains 6.84% and 8.21% absolute performance gains based on mBERT and XLM-R, respectively. This shows that our method constructs high-quality labeled target data by alleviating the mis-alignment issues. 4) Our proposed ACS method further outperforms the TRANSLATION-AF baseline, showing that explicitly switching the aspect terms between two languages is an effective approach to utilize the source and translated data for further enhancing the alignment between them. 5) By distilling the model on the unlabeled data of the target language, the proposed single-teacher distillation (ACS-DISTILL-S) and multi-teacher distillation (ACS-DISTILL-M) both achieve greater performance. This verifies our assumption that the language-specific knowledge is essential for tackling the cross-lingual ABSA task, even distilling the model on the unlabeled target data, the performance could be further improved. Specifically, the model trained with multiple teachers achieves slightly better performance than the single-teacher model. This is likely due to the reason that those multiple teachers capture different characteristics, thus the soft labels combining their predictions can better "teach" the student model.

3.5 Multilingual ABSA Results

In addition to the cross-lingual results (*i.e.*, from one source language to one target language), we also report the results under multilingual setting in Table 3. It can be noticed that training on the multilingual pseudo-labeled translated target data from our label projection method (*i.e.*, MTL-AF) sets up a quite strong baseline which can already outperform MTL-WS, a previous state-of-the-art model and surpasses its counterpart MTL-TA method by a large margin. Similar to the observation in the cross-lingual transfer, our proposed aspect code-switching method and distillation on unlabeled data further improve the adaptation performance, thanks to the better alignment of different languages and the utilization of the language-specific knowledge.

	FR	ES	NL	RU	Avg
*Based on **mBERT**:*					
MTL-TA†	40.72	54.14	49.06	43.89	46.95
MTL-WS†	46.93	58.18	49.87	44.88	49.96
MTL-AF	50.00	59.31	53.16	50.04	53.13
MTL-ACS	50.74	59.59	53.33	51.61	53.82
MTL-ACS-D	**53.56**	**62.05**	**53.56**	**53.87**	**55.76**
*Based on **XLM-R**:*					
MTL-TA†	52.80	63.56	60.37	55.67	58.10
MTL-WS†	57.96	68.60	61.24	59.74	61.89
MTL-AF	59.68	68.20	63.33	61.02	63.06
MTL-ACS	59.19	68.88	63.06	61.92	63.26
MTL-ACS-D	**62.17**	**70.38**	**65.98**	**62.79**	**65.33**

Table 3: Multilingual results with mBERT and XLM-R as backbone respectively. The best baseline results are underlined and best results by our model are in bold. † denotes results are from Li et al. (2020).

Compared with the bilingual transfer, the multilingual model can be used in a "*one model for all*" manner, *i.e.*, the same model can be applied for multiple target languages. Moreover, the model can benefit from the task-specific knowledge of the pseudo-labeled data in multiple target languages. With the XLM-R as backbone, our model distilled on the multilingual data (MTL-ACS-D) achieves 65.33 average F1 scores, which is even close to the F1 scores under the supervised setting (*i.e.* 67.15), showing the superiority of the proposed approach.

3.6 Discussions and Analysis

Ablation of the ACS method Table 4 gives the results from the three teacher models in the multi-teacher distillation method described in Sec 2.5, which can be regarded as the ablated variants of our proposed ACS model. We can see that all these three variants are not as competitive as ACS which is trained with all the code-switched bilingual data and the original two monolingual datasets. Among

	FR	ES	NL	RU	Avg
*Based on **mBERT**:*					
ACS	**49.65**	59.99	**51.19**	**52.09**	**53.23**
$D_S + D_T$	48.05	**60.23**	49.83	51.24	52.34
$D_{St} + D_T$	48.27	60.08	50.81	50.33	52.38
$D_{Ts} + D_T$	47.01	59.87	49.69	51.10	51.92
*Based on **XLM-R**:*					
ACS	**59.39**	67.32	**62.83**	**60.81**	**62.59**
$D_S + D_T$	57.91	**68.04**	60.80	60.81	61.89
$D_{St} + D_T$	56.26	65.47	61.03	60.00	60.69
$D_{Ts} + D_T$	56.13	66.80	60.36	59.69	60.75

Table 4: Ablation Study on the ACS method.

	ENTIRE	PARTIAL	BOUNDARY	TOTAL
TA	7.9%	13.4%	6.4%	27.8%
AF	0.7%	2.1%	0.0%	2.8%

Table 5: Error analysis of label projection methods, where the ratios are relative to the total 140 aspects.

them, combining the translated data and the source data is the most powerful method. Specifically, we can see that it achieves the best performance when the target language is Spanish. We conjecture the reason is that among the four target languages, Spanish is the most similar language with the source language (*i.e.*, English), their embedding spaces are already relatively well-aligned in the pre-trained language models. Therefore, using our proposed aspect switching strategy to explicitly enhance the cross-lingual alignment does not help in this case.

Impact of label projection methods Comparing *X*-AF (models using our *alignment-free* paradigm) and *X*-TA (models using the *translate-then-align* paradigm) in Tables 2 and 3, our *alignment-free* label projection achieves significantly higher performance than its counterpart. Here we conduct a detailed error analysis by randomly sampling 50 sentences in English and examining the pseudo-labeled data from ***AF*** and ***TA***. We manually categorize the errors in the 200 pseudo-labeled sentences of the four target languages and report the ratio for each error type in Table 5.

We can see that the alignment errors of the previous ***TA*** method lead to aspect term partially missed (PARTIAL), or incorrect aspect boundary which includes non-aspect words as a part of the aspect (BOUNDARY). Even worse, 7.9% of the aspects are entirely mismatched (ENTIRE) which would greatly hurt the model performance. Compared with it, our ***AF*** method largely alleviates those issues. We present an example case in Table 6 where ***TA*** only labels "*fruits*" in the translated sentence as a positive aspect, due to the incomplete alignment. While our method correctly matches "*seafood*" in English to "*fruits de mer*" in French, which provides the correct labeled target data. Such high-quality labeled data preserves the task-specific knowledge in the target language to establish a strong baseline for the cross-lingual ABSA task. We notice that our AF also produces 2.1% partially missed aspects, especially when facing long aspect terms. For example, "*shank*" is missed in the translated sentence for the aspect "*braised lamb shank in red wine*". It is due to the difficulty to translate such cases with an off-the-shelf translation system.

4 Related Work

Existing works on cross-lingual ABSA mainly focus on its sub-tasks including the cross-lingual aspect term extraction (Lin et al., 2014; Klinger and Cimiano, 2015) and aspect sentiment classification (Lambert, 2015; Barnes et al., 2016). To obtain language knowledge of the target languages, translation systems are used to obtain pseudo-parallel data (Zhou et al., 2015). A word or phrase alignment algorithm such as fastAlign (Dyer et al., 2013) is then utilized to project the label from the source to the target sentence. Since the performance of such methods heavily depends on the quality of the translation and alignment, different strategies are proposed to further improve the data quality. Klinger and Cimiano (2015) conduct an instance selection process to filter out low-quality target data. Li et al. (2020) propose a span-to-span mapping heuristics to refine the aspect boundary.

Another line of work uses the cross-lingual word embeddings trained on large parallel bilingual corpus (Ruder et al., 2019). By switching the word embeddings between different languages, the model can be used in a language-agnostic manner (Barnes et al., 2016; Akhtar et al., 2018; Wang and Pan, 2018; Jebbara and Cimiano, 2019). Wang and Pan (2018) propose a transition-based aspect extraction model which aligns the representations in different languages through an adversarial network. Jebbara and Cimiano (2019) consider the zero-shot transfer for aspect term extraction task with two types of cross-lingual embeddings.

Recently, transformer-based models pre-trained

***Example**: Labeled target data given by different label projection methods*	
Source English Sentence	[Mermaid Inn]$_{\texttt{POS}}$ is an overall good restaurant with really good [seafood]$_{\texttt{POS}}$.
Labeled French Sentence with *translate-then-align* Method	[Mermaid Inn]$_{\texttt{POS}}$ est un bon restaurant dans l' ensemble avec de très bons [fruits]$_{\texttt{POS}}$ de mer .
Labeled French Sentence with *alignment-free* Method	[Mermaid Inn]$_{\texttt{POS}}$ est un bon restaurant dans l' ensemble avec de très bons [fruits de mer]$_{\texttt{POS}}$.

Table 6: Example of different label projection methods with French as the target language. We use the bracket to highlight the aspect term, the corresponding sentiment polarities are shown as the subscript for each aspect.

on large multilingual corpus, such as the multilingual BERT (Devlin et al., 2019) and XLM-Roberta (Conneau et al., 2020), have shown significant improvements for various cross-lingual NLP tasks. Thanks to the language knowledge learned in the pre-training process, fine-tuning the model on the labeled source language data and directly conducting the inference on the target data can achieve impressive cross-lingual adaptation performance (Wu and Dredze, 2019; Pires et al., 2019; K et al., 2020). Some studies further utilize the translation system together with the pre-trained models, for example, by direct data transfer (Fei et al., 2020; Hu et al., 2020), data augmentation (Singh et al., 2019), and parameter warm-up (Li et al., 2020).

5 Conclusions

We tackle the cross-lingual ABSA task in this paper. To obtain high-quality labeled target data, we design an alignment-free label projection method which establishes a strong translation-based baseline. We further propose an aspect code-switching strategy to enhance the cross-lingual alignment and distill our method on the unlabeled target data to verify the importance of the language-specific knowledge for the ABSA problem. Experiments with five languages show the effectiveness of our methods.

References

Alan Akbik and Roland Vollgraf. 2018. ZAP: an open-source multilingual annotation projection framework. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*.

Md. Shad Akhtar, Palaash Sawant, Sukanta Sen, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Solving data sparsity for aspect based sentiment analysis using cross-linguality and multi-linguality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2018*, pages 572–582.

Jeremy Barnes, Patrik Lambert, and Toni Badia. 2016. Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification. In *26th International Conference on Computational Linguistics, COLING 2016*, pages 1613–1623.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 3685–3694.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, NAACL*, pages 644–648.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 7014–7026.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 504–515.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119, pages 4411–4421.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 537–546.

Soufian Jebbara and Philipp Cimiano. 2019. Zero-shot cross-lingual opinion target extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2019*, pages 2486–2495.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 4563–4568.

Roman Klinger and Philipp Cimiano. 2015. Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015*, pages 153–163.

Patrik Lambert. 2015. Aspect-level cross-lingual sentiment classification with constrained SMT. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015*, pages 781–787.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In *AAAI*, pages 6714–6721.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019b. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text, W-NUT@EMNLP 2019*, pages 34–41.

Xin Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2020. Unsupervised cross-lingual adaptation for sequence tagging and beyond. *CoRR*, abs/2010.12405.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2020. An iterative knowledge transfer network with routing for aspect-based sentiment analysis. *CoRR*, abs/2004.01935.

Zheng Lin, Xiaolong Jin, Xueke Xu, Weiping Wang, Xueqi Cheng, and Yuanzhuo Wang. 2014. A cross-lingual joint aspect/sentiment model for sentiment analysis. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014*, pages 1089–1098.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies.

Pintu Lohar, Maja Popović, Haithem Alfi, and Andy Way. 2019. A systematic comparison between smt and nmt on translating user-generated content.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. In *AAAI*.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, pages 2536–2545.

Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: an adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 7654–7673.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 4996–5001.

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 V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016*, pages 19–30.

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

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.*, 65:569–631.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: cross-lingual data augmentation for natural language inference and question answering. *CoRR*, abs/1905.11471.

Wenya Wang and Sinno Jialin Pan. 2018. Transition-based adversarial network for cross-lingual aspect extraction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 4475–4481.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jianguang Lou. 2020. Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3926–3932.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 2019*, pages 833–844.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 504–510.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2015. Clopinionminer: Opinion target extraction in a cross-language scenario. *IEEE ACM Trans. Audio Speech Lang. Process.*, 23(4):619–630.