KPC-cF: Korean Aspect-Based Sentiment Analysis via Pseudo-Classifier with Corpus Filtering for Low Resource Society

Kibeom Nam

Department of Data Science
Hongik University, 94, Wausan-ro, Mapo-gu, Seoul, 04066, Korea
skarlqja68@gmail.com

Abstract

Investigations into Aspect-Based Sentiment Analysis (ABSA) for Korean restaurant reviews are notably lacking in the existing literature. Our research proposes an intuitive and effective framework for ABSA in low-resource languages such as Korean. It optimizes prediction labels by integrating translated Benchmark and unlabeled Korean data. Using a model fine-tuned on translated data, we pseudo-labeled the actual Korean NLI set. Subsequently, we applied LaBSE and *MSP*-based filtering to this pseudo NLI set, enhancing its performance through additional training. Incorporating dual filtering, this model bridged dataset gaps, achieving positive results in Korean ABSA with minimal resources. Through additional data filtering and injecting pipelines, our approach aims to provide a cost-effective framework (e.g., human intervention and training resources) for data and model construction within communities, whether corporate or individual, in low-resource language countries. Compared to English ABSA, our framework showed an approximately 3% difference in F1 scores and accuracy. We will show the model[1] and data for Korean ABSA, publicly available at https://github.com/namkibeom/KPC-cF.

1 Introduction

In low-resource downstream tasks such as Korean ABSA, constraints exist in constructing ABSA systems that are socially and industrially beneficial (e.g., obtaining accurate labels and high-quality training data, building a efficient serving model). Addressing this challenge is fundamentally crucial for the practical implementation of multilingual ABSA leveraging the advantages of language models (Zhang et al. 2021; Lin et al. 2023). On the other hand, ABSA utilizing Large Language Models like ChatGPT can perform labeling through prompt tuning. however, it still shows limitations compared to small-scale models in terms of performance in diverse metrics and model serving costs (Wang et al. 2023; Wu et al. 2023; Dacon 2023). Therefore, in this study, we derive pseudo-labels for real Korean reviews using machine-translated English ABSA data, drawing inspiration from the research conducted by Balahur and Turchi (2012). Moreover, we employ LaBSE-based filtering on the actual Korean corpus transformed into an NLI task, thereby constructing an efficient Korean pseudo-classifier (Sun, Huang, and Qiu 2019; Feng et al. 2022). Through this process, we assess the impact of our constructed classifier on the practical classification performance of actual reviews. We validate that the pseudo-classifier, generated through the sentence-pair approach, outperforms the single approach when fine-tuning the translated dataset. Furthermore, using the model that predicts the translated dataset most effectively as a baseline, we generate pseudo-labels for actual data and conduct real-world testing of Korean ABSA. This involves subsequent fine-tuning the filtered corpus based on language-agnostic embedding similarity for review and aspect sentence pairs, along with setting a threshold for Maximum Softmax Probability (*MSP*) in pseudo-labels.

The main contributions of our work are:

- This is, to our knowledge, the first approach to generating a Pseudo classifier for automatic classification of aspect-based sentiment in the actual Korean domain.
- For actual review-based ABSA, we propose a filtered NLI corpus and additional training framework that allows stable fine-tuning in low-resource languages on models trained with high-resource translation data.
- A new challenging dataset of Korean ABSA, along with a KR3 and Translated benchmark correlated with cross-lingual understanding.

2 Background and Related Work

2.1 Task description

ABSA : In ABSA, Sun, Huang, and Qiu (2019) set the task as equivalent to learning subtasks 3 (Aspect Category Detection) and subtask 4 (Aspect Category Polarity) of SemEval-2014 Task 4 at the same time. Although there have been previous similar studies on Korean aspect-based sentiment classification in automotive domain datasets (Hyun, Cho, and Yu 2020), we perform a subtask method like Sun, Huang, and Qiu (2019) for Korean ABSA of restaurant reviews. A process of converting model and data to Korean is required. We aimed to identify a task-specific model through a comparison of two PLMs (mBERT, XLM-R$_{\text{Base}}$ in Sec. A 1), where there are no differences in the model structures other than those related to tokenization, vocabulary

[1] https://huggingface.co/KorABSA

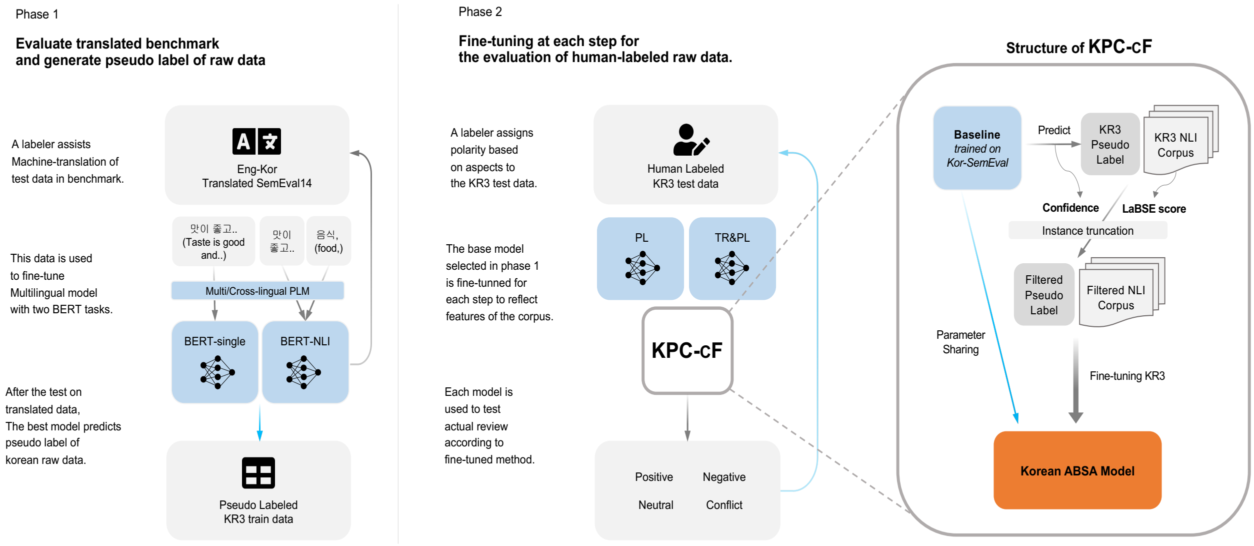
Phase 1
Evaluate translated benchmark and generate pseudo label of raw data
A labeler assists Machine-translation of test data in benchmark.
Eng-Kor Translated SemEval14
맛이 좋고.. (Taste is good and..)
맛이 좋고..
음식, (food,)
This data is used to fine-tune Multilingual model with two BERT tasks.
Multi/Cross-lingual PLM
BERT-single
BERT-NLI
After the test on translated data, The best model predicts pseudo label of korean raw data.
Pseudo Labeled KR3 train data
Phase 2
Fine-tuning at each step for the evaluation of human-labeled raw data.
A labeler assigns polarity based on aspects to the KR3 test data.
Human Labeled KR3 test data
The base model selected in phase 1 is fine-tuned for each step to reflect features of the corpus.
PL
TR&PL
KPC-cF
Each model is used to test actual review according to fine-tuned method.
Positive
Negative
Neutral
Conflict
Structure of KPC-cF
Baseline trained on Kor-SemEval
Predict
KR3 Pseudo Label
KR3 NLI Corpus
Confidence
LaBSE score
Instance truncation
Filtered Pseudo Label
Filtered NLI Corpus
Parameter Sharing
Fine-tuning KR3
Korean ABSA Model

Figure 1: A diagram illustrating the two phase of our method: (1) Fine-tuning Kor-SemEval and generate pseudo labeled KR3, (2) Fine-tuning KR3 using baseline model selected phase 1. We illustrated the filtering process (right) for fine-tuning KR3 data. Blue arrows (left & middle) indicate that this model is used to predict best label of review.

size, and model parameters. We refrained from defining a Unified-serving model using multi-label-multi-class classification from a task-oriented perspective due to the challenges associated with modifying pre-trained data and the ongoing injection of additional data. Consequently, we have redefined the problem into two BERT-based classification tasks as outlined below.

2.2 Classification approach

Single sentence Classification : BERT for single-sentence classification tasks. For ABSA, We fine-tune the pre-trained BERT model to train n_a (i.e., number of aspect categories) classifiers for all aspects and then summarize the results. The input representation of the BERT can explicitly represent a pair of text sentences in a sequence of tokens. A given token's input representation is constructed by summing the corresponding token, segment, and position embeddings. For classification tasks, the first word of each sequence is a unique classification embedding `[CLS]`. Segment embeddings in single sentence classification use one.

Sentence-pair Classification : Based on the auxiliary sentence constructed as aspect word text, we use the sentence-pair classification approach to solve ABSA. The input representation is typically the same with the single-sentence approach. The difference is that we have to add two separator tokens `[SEP]`, the first placed between the last token of the first sentence and the first token of the second sentence. The other is placed at the end of the second sentence after its last token. This process uses both segment embeddings. For the training phase in the sentence-pair classification approach, **we only need to train one classifier** to perform both aspect categorization and sentiment classification. Add one classification layer to the Transformer output and apply the softmax activation function. Corresponding to the combination of the multilingual pre-trained model and the presence of auxiliary sentences, we name the models: mBERT-single, XLM-R$_{\text{Base}}$-single, mBERT-NLI, XLM-R$_{\text{Base}}$-NLI.

3 Two phase of Pseudo-Classifier

3.1 Motivation and Contribution

Our research aims to build a model that can perform the best ABSA in a simple way on actual data with Korean nuances. Past research by Balahur and Turchi (2012) has shown that Machine Translation (MT) systems can obtain training data for languages other than English in general sentiment classification. Also, although it was a different domain at Zhou et al. (2021), we found it necessary to investigate whether the concept of pseudo labels could help bridge the gap between translated data and actual target language data. Therefore, we attempted the following two phases to assess the impact of the generated pseudo-classifier, fine-tuned using translated datasets from the ABSA benchmark and pseudo-labeled actual review data, on Korean ABSA. Fig. 1 shows the two-phase pseudo-classifiers we will employ. In the first phase, the most effective baseline model is selected among the models trained and evaluated through the translation dataset. In Phase 2, We fine-tune the baseline model, which was effective in training on translated data, by additionally incorporating pseudo-labeled actual Korean reviews. Using this tuned model, we make predictions and evaluate on manually labeled real Korean reviews. During this process, thresholding of pseudo-labels and LaBSE filtering are performed to enhance the features of the corpus.

Kor-SemEval Train	Aspect	Polarity
서비스는 평범했고 에어컨이 없어서	가격 (price)	없음 (None)
편안한 식사를 할 수 없었습니다.	일화 (anecdotes)	없음 (None)
(*The service was mediocre*	음식 (food)	없음 (None)
and the lack of air conditioning made for	분위기 (ambience)	부정 (Negative)
a less than comfortable meal.)	서비스 (service)	중립 (Neutral)
KR3 Train	**Aspect**	**Polarity**
	Input form in NLI with ***Pseudo Label***	
가로수길에서 조금 멀어요 점심시간에 대기 엄청납니다	가격 (price)	없음 (None)
일행 모두 있어야 들어갈 수 있어요 맛은 보통이에요	일화 (anecdotes)	부정 (Negative)
(*It's a little far from Garosu-gil. There's a huge wait during lunch time.*	음식 (food)	없음(None)
You have to have everyone in your group to get in.	분위기 (ambience)	부정 (Negative)
The taste is average.)	서비스 (service)	없음 (None)

Table 1: Samples of Kor-SemEval and KR3 train dataset

3.2 LaBSE based Filtering

In this approach, we aim to extract good-quality sentences-pair from the pseudo-NLI corpus. Language Agnostic BERT Sentence Embedding model (Feng et al. 2022) is a multilingual embedding model that supports 109 languages, including some Korean languages. Feng et al. (2022) suggested that the dual-encoder architecture of the LaBSE model, originally designed for machine translation in source-target language data (Batheja and Bhattacharyya 2022, 2023), can be applied not only to other monolingual tasks like Semantic Textual Similarity (STS) but also to data (i.e., sentence pair-set) filtering for creating high-quality training corpora in terms of meaning equivalence. Therefore, to mitigate performance degradation caused by the linguistic gap between translated data and actual Korean data during fine-tuning, We introduce the following filtering method that enables the identification of meaning equivalence (e.g., connotation) in actual Korean sentence-pairs, even when viewed from the perspective of model trained on bilingual translation pairs.

We generate the sentence embeddings for the review text and aspect of the pseudo-NLI corpora using the LaBSE model. Then, we compute the cosine similarity between the review text and aspect sentence embeddings. After that, we extract good quality NLI sentences based on a threshold value of the similarity scores. We calculate the average similarity score on a dataset from the our KR3 NLI corpus. Our processed corpus consists of high-quality sentence pairs, so it helps us decide upon the threshold value.

LaBSE scoring : Let $D = \{(s_i, a_i)\}_{i=1}^{N}$ be a pseudo-NLI corpus with N examples, where s_i and a_i represents i^{th} review and aspect sentence respectively. We first feed all the review sentences present in the pseudo parallel corpus as input to the LaBSE model[2], which is a Dual encoder model with BERT-based encoding modules to obtain review sentence embeddings (S_i). The sentence embeddings are extracted as the l2 normalized `[CLS]` token representations from the last transformer block. Then, we feed all the aspect sentences as input to the LaBSE model to obtain aspect sentence embeddings (A_i). We then compute cosine similarity ($score_i$) between the review and the corresponding aspect sentence embeddings.

$$S_i = LaBSE\,(s_i) \tag{1}$$

$$A_i = LaBSE\,(a_i) \tag{2}$$

$$score_i = cosine\_similarity\,(S_i, A_i) \tag{3}$$

We aimed to apply the LaBSE scoring to the actual Korean dataset, KR3, intending to facilitate flexible learning compared to the translated dataset, Kor-SemEval.

3.3 Confidence score Filtering

Meanwhile, we need to develop a classifier capable of optimal predictions on the KR3 test set, which can be considered as out-of-distribution data separate from the translated data. Drawing on previous research (Arora, Huang, and He 2021), we expect that language shifts (i.e., translated data, actual korean data) embody both Background and Semantic shift characteristics. To ensure robust learning in both aspect detection and sentiment classification, we introduce additional thresholding on Maximum Softmax Probability (*MSP*; Hendrycks and Gimpel 2017) after LaBSE-based filtering on the KR3 train set. When considering an input $x = (s_i, a_i) \in \mathcal{X}$ and its corresponding pseudo label $y \in \mathcal{Y}$, the score $s(x)$ for *MSP* is expressed as:

$$s_{\textit{MSP}}(x) = \max_{k \in \mathcal{Y}} p_{model}(y = k \mid x). \tag{4}$$

Through this, we intended a dual scoring and filtering process to ensure that our classifier does not retrain on misplaced confidence or subpar prediction outcomes for out-of-distribution data.

3.4 Dataset for Fine-Tuning and Test

Kor-SemEval : We translate the SemEval-2014 Task 4 (Pontiki et al. 2014) dataset[3]. The training data was machine-translated (by Google Translate), and Test data was corrected manually only for fewer than 10 instances where

[2] https://huggingface.co/sentence-transformers/LaBSE

[3] http://alt.qcri.org/semeval2014/task4/

abnormal translations occurred after machine translation. Each sentence contains a list of aspect a with the sentiment polarity y. Ultimately, given a sentence s in the sentence, we need to:

- detects the mention of an aspect a;
- determines the positive or negative sentiment polarity y for the detected aspect.

This setting allows us to jointly evaluate Subtask 3 (Aspect Category Detection) and Subtask 4 (Aspect Category Polarity).

KR3 : Unlike the domains previously used for Korean sentiment classification (Ban 2022; Lee, Lim, and Choi 2020; Yang 2021), Korean Restaurant Review with Ratings (KR3) is a restaurant review sentiment analysis dataset constructed through actual certified map reviews. In the case of restaurant reviews, words and expressions that evaluate positive and negative are mainly included, and real users often infer what a restaurant is like by looking at its reviews. Accordingly, Jung et al.[3] constructed the KR3 dataset by crawling and preprocessing user reviews and star ratings of websites that collect restaurant information and ratings. KR3 has 388,111 positive and 70,910 negative, providing a total of 459,021 data plus 182,741 unclassified data, and distributed to Hugging Face[4].

We structured our training and test datasets to match the size of Kor-SemEval. Specifically, we addressed potential biases by randomly sampling indices from the original KR3, ensuring that evaluations for a specific restaurant were non-overlapping. Additionally, we maintained an even distribution of positive, negative, and neutral (ambiguous) classes, irrespective of the aspects indicated in the original KR3. This preprocessing step aimed to capture a comprehensive representation of sentiments across diverse attributes of sentences in the dataset. Subsequently, the data were configured to suit sentence pair classification (see Tab. 1). To allocate polarity labels for each aspect within the KR3 dataset, pseudo-labeling was conducted utilizing the optimal model identified during the Kor-SemEval performance evaluation. Pseudo labels were assigned to the KR3 training data, and post pseudo labeling, the test data underwent manual relabeling by researchers. Tab. 1 shows some Kor-SemEval and KR3 training data samples. In the case of KR3, the negative aspect is better reflected. Meanwhile, while Kor-SemEval gave neutrality to mediocre service, KR3 did not give neutrality to mediocre taste. While positive and negative data have been sufficiently accumulated and reflected, the tendency for a lack of neutral data can be confirmed in advance through some samples. We have organized both Kor-SemEval and KR3 data as open-source to facilitate their use in various training and evaluation scenarios.

3.5 Metrics

The benchmarks for SemEval-2014 Task 4 are the several best performing systems in Sun, Huang, and Qiu (2019), Wang et al. (2016) and Pontiki et al. (2014). When evaluating Kor-SemEval and KR3 test data with subtask 3 and 4, following Sun, Huang, and Qiu (2019), we also use Micro-F1 and accuracy respectively.

4 Experiments

4.1 Exp-I: Kor-SemEval

We conducted evaluations for each of the mBERT-single, XLM-R$_{\text{Base}}$-single, mBERT-NLI, XLM-R$_{\text{Base}}$-NLI, and NLI-ensemble models. We included the results from the previous SemEval14 research and Kor-SemEval to compare and evaluate the performance in Korean.

4.1.1 Results

Results on Kor-SemEval are presented in Tab. 2 and Tab. 3. Similar to the SemEval results, it was confirmed that tasks converted to NLI tasks tend to be better than single tasks, with mBERT achieving better results in single and XLM-R$_{\text{Base}}$ in NLI. The XLM-R$_{\text{Base}}$-NLI model performs best, excluding precision for aspect category detection. It also works best for aspect category polarity. The NLI-ensemble model was the best in precision but performed poorly in other metrics.

Model	**SemEval-14**		
	Precision	Recall	Micro-F1
BERT-single	92.78	89.07	90.89
BERT-pair-NLI-M	93.15	90.24	91.67
*Models trained & evaluated on **Kor-SemEval***			
mBERT-single	92.16	77.95	84.46
XLM-R$_{\text{Base}}$-single	91.01	49.37	64.01
mBERT-NLI	91.10	79.90	85.14
XLM-R$_{\text{Base}}$-NLI	91.37	**83.71**	**87.37**
NLI-ensemble	**93.70**	81.27	87.04

Table 2: Test set results for Aspect Category Detection. We use the results reported in BERT-single and BERT-pair-NLI-M (Sun, Huang, and Qiu 2019) for English dataset together with our results.

Model	**SemEval-14**		
	4-way acc	3-way acc	Binary
BERT-single	83.7	86.9	93.3
BERT-pair-NLI-M	85.1	88.7	94.4
*Models trained & evaluated on **Kor-SemEval***			
mBERT-single	68.20	71.84	79.52
XLM-R$_{\text{Base}}$-single	62.93	66.29	75.20
mBERT-NLI	73.95	77.90	84.87
XLM-R$_{\text{Base}}$-NLI	**79.41**	**83.66**	**89.98**
NLI-ensemble	78.24	82.43	89.65

Table 3: Test set accuracy (%) for Aspect Category Polarity. We use the results reported in BERT-single and BERT-pair-NLI-M (Sun, Huang, and Qiu 2019) for English dataset together with our results.

[4] https://huggingface.co/datasets/leey4n/KR3

Model	#Sample Capacity/Count	Pre-tuning	Aspect Category			Polarity		
			Precision	Recall	Micro-F1	4-way acc	3-way acc	Binary
Baseline+PL	4.60MB 15.23K	un-tuned	91.82	79.85	85.42	**84.78**	**87.16**	**91.55**
Baseline+PL-cF	2.15MB 6.08K	un-tuned	91.72	79.76	85.32	84.32	86.69	90.86
Baseline+TR+PL	6.14MB 30.45K	Kor-SemEval	**92.03**	**85.23**	**88.50**	84.50	86.88	90.37
KPC-cF	3.69MB 21.30K	Kor-SemEval	**92.79** (↑)	**85.60** (↑)	**89.05** (↑)	**85.05** (↑)	**87.44** (↑)	**91.65** (↑)

Table 4: KR3 test set results for Aspect Category Detection (middle) and Aspect Category Polarity (right). We reported the total number and capacity of the trained samples for each model in the **#Sample**. Specifically, for **Baseline+PL**, it refers to the total number (i.e., equivalent to Kor-SemEval) and capacity of samples from the KR3 train that we initially established. for Additionally, **KPC-cF**, this refers to the total number and capacity of samples from the filtered KR3 train and Kor-SemEval train set.

4.2 Exp-II: KR3 Test Set

Furthermore, based on the results from Kor-SemEval, we examined the dissimilarity specific to the translation task between mBERT and XLM-$\mathrm{R}_{\mathrm{Base}}$. Accordingly, we opted for the XLM-$\mathrm{R}_{\mathrm{Base}}$-NLI approach, which demonstrated the best performance, as the base model for Phase 2 (see Sec. 3.1, Fig. 1). We conducted evaluations on KR3 test data using the KR3 train set (PL), the model trained on Kor-SemEval and additional fine-tuning with KR3 train (TR+PL), and corpus obtained through Confidence thresholding and LaBSE-based filtering on KR3 train (PL-cF). More details are described in Sec. 4.2.1, Tab. 5.

4.2.1 Results

To investigate the effect of features for each corpus, we conduct baseline tuning comparisons between the PL and the PL-cF (i.e., All data follows the NLI format), as indicated in Tab. 4. The variants of our tuning framework includes:

- **Baseline+PL (Pseudo Labeled data):** Fine-tuning XLM-$\mathrm{R}_{\mathrm{Base}}$ with pseudo KR3.
- **Baseline+PL-cF (Corpus Filtering):** Fine-tuning XLM-$\mathrm{R}_{\mathrm{Base}}$ with the data obtained by truncating instance from pseudo KR3, where the threshold of *MSP* (Hendrycks and Gimpel 2017) is less than 0.5 and the cosine similarity between LaBSE embeddings is less than 0.15.
- **Baseline+TR (TRanslated data)+PL:** Fine-tuning XLM-$\mathrm{R}_{\mathrm{Base}}$-NLI (pre-tuned on Kor-SemEval) with pseudo KR3.
- **KPC-cF (Baseline+TR+PL-cF):** Fine-tuning XLM-$\mathrm{R}_{\mathrm{Base}}$-NLI (pre-tuned on Kor-SemEval) with PL-cF.

Results on the KR3 test set are presented in Tab. 4 and Fig. 2. We find that the KPC-cF approach achieved good and stable trained results in both subtasks for the actual korean data. The model pre-tuned with Kor-SemEval achieves the best performance in Aspect Category Detection (ACD). For Aspect Category Polarity (ACP), it performs exceptionally well in the tuning of Pseudo Labels, especially in the Binary setting. Filtered Pseudo Labels preserve this characteristic well and amplify the performance of all metrics within ACP.

5 Discussion

In Phase 1, XLM-R, known for its proficiency in capturing cross-lingual representations, exhibits an underfitting tendency concerning the contextual disparities in aspect vocabulary within a single task. This can be attributed to data scarcity relative to model availability for each classifier or viewed as a limitation in single text classification using SPM in low-resource Korean ABSA. Nevertheless, in the NLI task, it showcases potential by outperforming mBERT, guided by the instruction ”aspect.” Conversely, mBERT demonstrates stable results in both single and NLI tasks, exhibiting an overall accuracy increase, particularly in the NLI task. Furthermore, Phase 2 reveals that the combination of the NLI approach and translated data significantly impacts the metrics of model exploration in aspects. Pseudo-labels in this phase contribute to enhancing the binary classification of sentiment, resulting in improved classifier performance. Notably, finely filtered pseudo-labels, unlike a mere addition to translated data, play a crucial role in maintaining and enhancing accuracy and F1 score, even with fewer training resources (3.69MB/21.30K). In essence, the language-agnostic perspective of the filtered NLI set (i.e., Sentence-pair & Label), combined with threshold filtering for pseudo-labels, facilitates performance improvement without compromising pre-tuned parameters during training. This superiority is evident compared to other models, particularly within 3 to 4 epochs.

6 Conclusion

Aspect Based Sentiment Analysis (ABSA) has been recognized as one of the most attractive subareas in text analytics and NLP. However, obtaining high-quality or ample-size label data has been one of the most essential issues hindering the development of ABSA. In this paper, We addressed the language gap issue in ABSA by building a pseudo-classifier. This involved fine-tuning an NLI model with translated data, performing LaBSE scoring on Korean NLI pairs, and further fine-tuning with optimal pseudo-labels. Additionally, we presented Kor-SemEval (translated) and KR3 train (pseudo labeled & filtered), testset (Gold Label) composed of actual Korean nuances, developing a fine-tuned model and data that can provide powerful assistance in Korean ABSA. We invite the community to extend Korean ABSA by providing new datasets, trained models, evaluation results, and metrics.

Acknowledgements

We thank Chanseo Nam for manually correcting and annotating the Kor-SemEval and KR3 testset used to test our model, during his undergraduate research internship. This research was supported by Brian Impact, a non-profit organization dedicated to the advancement of science and technology.

A. Appendix

A 1. Additional information

mBERT : Multilingual BERT is a BERT trained for multilingual tasks. It was trained on monolingual Wikipedia articles in 104 different languages. It is intended to enable mBERT finetuned in one language to make predictions for another. Azhar and Khodra (2020) and Jafarian et al. (2021) show that mbert performs effectively in a variety of multilingual Aspect-based sentiment analysis. It is also actively used as a base model in other tasks of Korean NLP (Lee et al. 2021; Park et al. 2021), but is rarely confirmed in Korean ABSA tasks. Thus, our study used the pre-trained mBERT base model with 12 layers and 12 heads (i.e., 12 transfomer encoders). This model generates a 768-dimensional vector for each word. We used the 768-dimensional vector of the Extract layer to represent the comment. Like the English language subtasks, a single Dense layer was used as the classification model.

XLM-R : XLM-RoBERTa (Conneau et al. 2020) is a cross-lingual model that aims to tackle the curse-of-multilingualism problem of cross-lingual models. It is inspired by RoBERTa (Liu et al. 2019), trained in up to 100 languages, and outperforms mBERT in multiple cross-lingual ABSA benchmarks (Zhang et al. 2021; Phan et al. 2021; Szołomicka and Kocon 2022). However, like mBERT, Korean ABSA has yet to be actively evaluated, so we used it as a base model. We use the base version (XLM-$\mathrm{R}_{\mathrm{Base}}$) coupled with an attention head classifier, the same optimizer as mBERT.

Ensemble : Meanwhile, we additionally use a voting-based ensemble, a typical ensemble method. The ensemble can confirm generalized performance based on similarity of model results in NLI task (Xu et al. 2020). So, We add separate power-mean ensemble result to identify a metric that amplifies probabilities based on the Pre-trained Language Models (PLMs). we reported the ensemble results of the top-performing models trained on NLI tasks for each PLM.

A 2. Hyperparameter

All experiments are conducted on two pre-trained cross-lingual models. The XLM-RoBERTa-base and BERT-base Multilingual-Cased model are fine-tuned. The number of Transformer blocks is 12, the hidden layer size is 768, the number of self-attention heads is 12, and the total number of parameters for the XLM-RoBERTa-base model is 270M, and BERT base Multilingual-Cased is 110M. When fine-tuning, we keep the dropout probability at 0.1 and set the number of epochs to 2 and 4. The initial learning rate is 2e-5, and the batch size is 3 and 16.

In the translated dataset, Kor-SemEval, we aimed to introduce a solid regularization effect for the incoherence of the trained data by using a small batch size (Sekhari, Sridharan, and Kale 2021). Additionally, for fair comparison, we set the batch size to 3, allowing variability in the training pattern of the input form in NLI. This setting was applied to both single and NLI tasks. The max length was set to 512, and for epochs beyond 3, no significant performance improvement was observed, so the results from epoch 2 were noted. Subsequently, in KR3, following the pattern of the previous experiments (Karimi, Rossi, and Prati 2021), we fine-tuned with a batch size of 16, and the results from epoch 4 were reported. Each reported metric is the average of three runs with three different random seeds to mitigate the effects of random variation of the results.

A 3. Datasets

Dataset	Count	L-avg	*MSP*-avg
KR3 Train (**PL**)	15.23K	0.15	0.87
PL-LaBSE 0.15-th 0.5 (**PL-cF**)	6.08K	0.21 (↑)	0.84 (↓)

Table 5: Number of instances and average scores (LaBSE, *MSP*) for KR3 & Filtered KR3 fine-tuning set

A 4. Additional Results

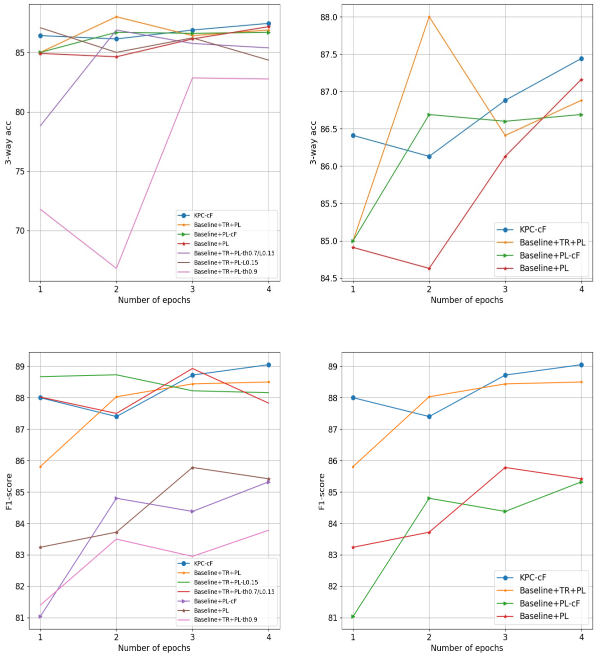

Figure 2: Performance of ACD and ACP during fine-tuning on KR3 test data. Left: results with the addition of other fine-tuned models. **th** denotes the threshold for confidence of pseudo labeling, and **L** denotes the threshold for filtering of LaBSE scoring; Right: four models compared in this paper. Blue line represents our proposed model, **KPC-cF**.

References

Arora, U.; Huang, W.; and He, H. 2021. Types of Out-of-Distribution Texts and How to Detect Them. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10687–10701. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Azhar, A. N.; and Khodra, M. L. 2020. Fine-tuning pretrained multilingual bert model for indonesian aspect-based sentiment analysis. In *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–6. IEEE.

Balahur, A.; and Turchi, M. 2012. Multilingual Sentiment Analysis using Machine Translation? In Balahur, A.; Montoyo, A.; Barco, P. M.; and Boldrini, E., eds., *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, 52–60. Jeju, Korea: Association for Computational Linguistics.

Ban, B. 2022. A survey on awesome korean nlp datasets. In *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 1615–1620. IEEE.

Batheja, A.; and Bhattacharyya, P. 2022. Improving Machine Translation with Phrase Pair Injection and Corpus Filtering. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5395–5400. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Batheja, A.; and Bhattacharyya, P. 2023. "A Little is Enough": Few-Shot Quality Estimation based Corpus Filtering improves Machine Translation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 14175–14185. Toronto, Canada: Association for Computational Linguistics.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Online: Association for Computational Linguistics.

Dacon, J. 2023. Are You Worthy of My Trust?: A Socioethical Perspective on the Impacts of Trustworthy AI Systems on the Environment and Human Society. *arXiv preprint arXiv:2309.09450.*

Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; and Wang, W. 2022. Language-agnostic BERT Sentence Embedding. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 878–891. Dublin, Ireland: Association for Computational Linguistics.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Hyun, D.; Cho, J.; and Yu, H. 2020. Building Large-Scale English and Korean Datasets for Aspect-Level Sentiment Analysis in Automotive Domain. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 961–966. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Jafarian, H.; Taghavi, A. H.; Javaheri, A.; and Rawassizadeh, R. 2021. Exploiting BERT to improve aspect-based sentiment analysis performance on Persian language. In *2021 7th International Conference on Web Research (ICWR)*, 5–8. IEEE.

Karimi, A.; Rossi, L.; and Prati, A. 2021. Improving BERT Performance for Aspect-Based Sentiment Analysis. In Abbas, M.; and Freihat, A. A., eds., *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, 39–46. Trento, Italy: Association for Computational Linguistics.

Lee, H.; Yoon, J.; Hwang, B.; Joe, S.; Min, S.; and Gwon, Y. 2021. Korealbert: Pretraining a lite bert model for korean language understanding. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 5551–5557. IEEE.

Lee, Y.-J.; Lim, C.-G.; and Choi, H.-J. 2020. Korean-specific emotion annotation procedure using N-gram-based distant supervision and Korean-specific-feature-based distant supervision. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1603–1610.

Lin, N.; Fu, Y.; Lin, X.; Zhou, D.; Yang, A.; and Jiang, S. 2023. Cl-xabsa: Contrastive learning for cross-lingual aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Park, S.; Moon, J.; Kim, S.; Cho, W. I.; Han, J. Y.; Park, J.; Song, C.; Kim, J.; Song, Y.; Oh, T.; Lee, J.; Oh, J.; Lyu, S.; Jeong, Y.; Lee, I.; Seo, S.; Lee, D.; Kim, H.; Lee, M.; Jang, S.; Do, S.; Kim, S.; Lim, K.; Lee, J.; Park, K.; Shin, J.; Kim, S.; Park, L.; Oh, A.; Ha, J.-W.; and Cho, K. 2021. KLUE: Korean Language Understanding Evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Phan, K. T.-K.; Hao, D. N.; Van Thin, D.; and Nguyen, N. L.-T. 2021. Exploring zero-shot cross-lingual aspect-based sentiment analysis using pre-trained multilingual language models. In *2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 1–6. IEEE.

Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; and Manandhar, S. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Nakov, P.; and Zesch, T., eds., *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35. Dublin, Ireland: Association for Computational Linguistics.

Sekhari, A.; Sridharan, K.; and Kale, S. 2021. Sgd: The role of implicit regularization, batch-size and multiple-epochs. *Advances In Neural Information Processing Systems*, 34: 27422–27433.

Sun, C.; Huang, L.; and Qiu, X. 2019. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 380–385. Minneapolis, Minnesota: Association for Computational Linguistics.

Szołomicka, J.; and Kocon, J. 2022. MultiAspectEmo: Multilingual and Language-Agnostic Aspect-Based Sentiment Analysis. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 443–450. IEEE.

Wang, Y.; Huang, M.; Zhu, X.; and Zhao, L. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In Su, J.; Duh, K.; and Carreras, X., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 606–615. Austin, Texas: Association for Computational Linguistics.

Wang, Z.; Xie, Q.; Ding, Z.; Feng, Y.; and Xia, R. 2023. Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339*.

Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.-L.; and Tang, Y. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5): 1122–1136.

Xu, Y.; Qiu, X.; Zhou, L.; and Huang, X. 2020. Improving bert fine-tuning via self-ensemble and self-distillation. *arXiv preprint arXiv:2002.10345*.

Yang, K. 2021. Transformer-based Korean pretrained language models: A survey on three years of progress. *arXiv preprint arXiv:2112.03014*.

Zhang, W.; He, R.; Peng, H.; Bing, L.; and Lam, W. 2021. Cross-lingual Aspect-based Sentiment Analysis with Aspect Term Code-Switching. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9220–9230. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Zhou, Y.; Zhu, F.; Song, P.; Han, J.; Guo, T.; and Hu, S. 2021. An adaptive hybrid framework for cross-domain aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14630–14637.