KPC-cF: Korean Aspect-Based Sentiment Analysis via NLI-Based Pseudo-Classifier with Corpus Filtering

Anonymous ACL submission

Abstract

Previous research on Aspect-Based Sentiment Analysis (ABSA) for Korean reviews in the restaurant domain not has been conducted. Nowadays, most state-of-the-art results for a wide array of NLP tasks are achieved by utilizing pre-trained language representation. This paper seeks to develop PLM-based pseudo classifier that generates the best prediction labels by integrating translated data and unlabeled actual Korean data. We utilized the common ML concept of semi-supervised learning, along with LaBSE-based filtering, on the basis of transformation to the sentence-pair task and fine-tuned the crosslingual model. This achieved state-of-the-art results in Korean ABSA with low resources, showing approximately a 3% difference in F1 scores and accuracy compared to English ABSA results. We show the model and data for Korean ABSA, publicly available at https://huggingface.co/KorABSA.

1 Introduction

Over the past decade, Sentiment Analysis (SA) has been one of the most popular tasks in the Natural Language Processing (NLP) field due to the evolution of the Internet, particularly the increasing amount of user opinion content. Sentiment analysis has been widely used among companies to extract opinions about their products or services automatically. It aims to identify and extract user opinions (Erik et al., 2017), often in positive, neutral, and negative categories. Several companies may need more fine-grained analysis using aspect-based sentiment analysis (ABSA) because sentiment analysis may not be enough to respond to all real-world demands if the given text has more than one topic or aspect.

For example, we have a sentence, "This food is great, but the waitress has a bad attitude." In this example, we get two sentiment polarities toward two aspects: "food" is positive, and "service" is negative. In other words, the ABSA system returns output pairs of aspect and sentiment in the review text (Pontiki et al., 2014, 2015, 2016).

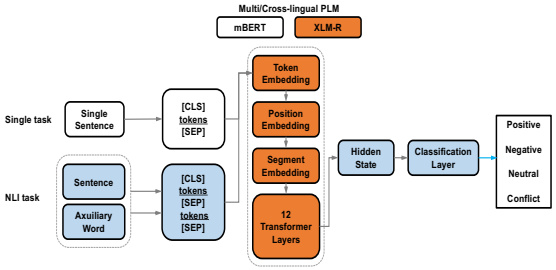

Figure 1: Overall structure of ABSA task using two classification task and PLM

Since devising deep learning models for ABSA recently received substantial attention (Zeng et al., 2019), building large-scale datasets in different languages has been an essential line of research (Rosenthal et al., 2019). However, such an approach requires domain-specific and manual training data. Due to the high human annotation cost, datasets' size and language are limited (Hyun et al., 2020). In addition, although transfer learning from a pre-trained language representation model (PLM) is a strong candidate because of its accurate performance with pre-trained data (Nurul Azhar and Leylia Khodra, 2021), there is still a problem of insufficient resources for accurate labels in order to apply it to downstream task like Korean ABSA. The process of improving the learning method in situations with insufficient labeled target language data has been found to be fundamentally challenging for the practical implementation of multilingual ABSA that leverages the advantages of language models (Lin et al., 2023; Zhang et al., 2021).

Therefore, in this paper, we generate pseudo-labels for actual Korean reviews using machine-translated English ABSA data, comparing with Balahur and Turchi (2012). We perform filtering based on LaBSE for the corpus transformed into an NLI

task, creating an effective Korean pseudo-classifier (Sun et al., 2019; Feng et al., 2022). Through this, we verify how our constructed classifier affects the actual review classification performance. We confirm that the pseudo-classifier generated by the sentence-pair approach is superior to the single approach in fine-tuning the translated dataset. Furthermore, using the top-performing model as a baseline, we generate pseudo-labels for actual review data. Subsequently, we conduct real-world testing of Korean ABSA by fine-tuning the filtered corpus based on language-agnostic embedding similarity for review and aspect sentence pairs, along with the threshold value of 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.
- We show insights into the selecting and fine-tuning PLM for effective Korean ABSA.
- For actual review-based ABSA, we propose a filtered NLI corpus framework that enables stable fine-tuning in low-resource languages.
- A new challenging dataset of Korean ABSA, along with a review of Korean nuances and Translated benchmark correlated with cross-lingual understanding.

2 Related Works and Classifying Methods

2.1 Task description

ABSA In ABSA, Sun et al. (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 et al., 2020), we perform a subtask method like Sun et al. (2019) for Korean ABSA of restaurant reviews. A process of converting models and data to Korean is required.

2.2 Multi/Cross-lingual Model

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. Jafarian et al. (2021) and Azhar and Khodra (2020) 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. 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., 2019) 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-R_{Base}) coupled with an attention head classifier, the same optimizer. We aimed to identify a task-specific model through a comparison of two pre-trained models, where there are no differences in the model structures other than those related to tokenization (WordPiece, SPM), vocabulary size, and parameters.

2.3 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 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

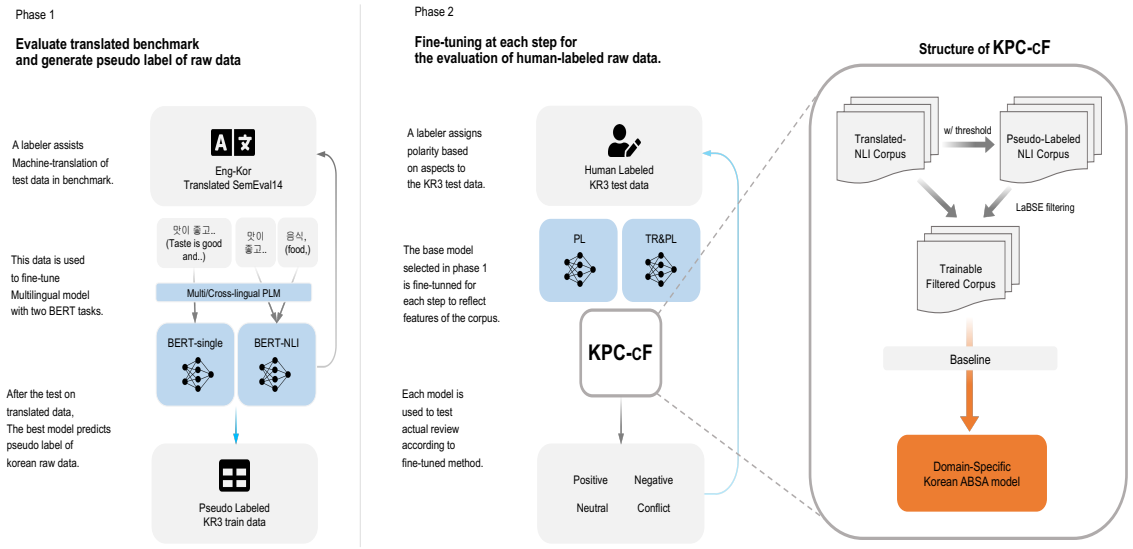

Figure 2: A diagram illustrating the two phase of our method: (1) Fine-tuning Kor-SemEval and generate pseudo labeled KR3, (2) Fine-tuning KR3 using either the untuned model or the model tuned on Kor-SemEval. We illustrated the filtering process (right) for fine-tuning KR3 data. Blue arrows indicate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

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_{Base}-single, mBERT-NLI, XLM-R_{Base}-NLI, and Figure 1 shows an overview of our models.

Ensemble Meanwhile, we additionally use a voting-based ensemble, a typical ensemble method. We first fine-tune two approaches of BERT (single, NLI) and two PLM (mBERT, XLM-R_{Base}). 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 classification method.

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. Figure 2 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 evaluate and compare the models fine-tuned for each corpus on the selected baseline using actual review data. During this process, thresholding of pseudo-labels and LaBSE filtering are performed to enhance the features of the corpus.

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

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

Dataset	Positive	Negative	Neutral	Conflict
Total sentiment for each aspect in ***Testset***				
Kor-SemEval	677	242	94	52
KR3	631	387	50	30

Table 2: Test data statistics of Kor-SemEval and KR3

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, can be applied not only to other monolingual tasks like STS but also to data filtering for creating high-quality training corpora. Therefore, to mitigate performance degradation caused by the linguistic gap between translated data and actual Korean data during fine-tuning, we introduce the following data filtering methods. 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[1], 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.

$$S_i = LaBSE\,(s_i) \quad (1)$$

$$A_i = LaBSE\,(a_i) \quad (2)$$

$$score_i = cosine\_similarity\,(S_i, A_i) \quad (3)$$

3.3 Dataset for Fine-tuning and Test

Kor-SemEval We translate the SemEval-2014 Task 4 (Pontiki et al., 2014) dataset[2]. Moreover, it is evaluated for Korean aspect-based sentiment analysis. The training data was machine-translated, and the test data was manually translated after machine translation.
Each sentence contains a list of aspects a with the sentiment polarity. Ultimately, given a sentence s in the sentence, we need to:

- detects the mention of an aspect ;
- 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). Afterward, train data was

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

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

learned using the multi, cross-lingual model and classification approach, and test data was evaluated.

KR3 Unlike the domains previously used for Korean sentiment classification (Ban, 2022; Lee et al., 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[3].

We configured the same number of training and test data as Kor-SemEval. Additionally, we configured an equal number of positive, negative, and neutral classes as mentioned in the existing KR3, and preprocessed them to ensure the representation of polarity in various sentence attributes. Afterward, the data were preprocessed in a form suitable for sentence pair and single sentence classification. To re-assign a polarity label for each aspect of KR3 data, pseudo-labeling was performed using the best model in testing Kor-SemEval. Training data was pseudo-labeled through the model with the best predictive performance for each single and NLI, and test data was manually re-labeled by researchers after pseudo-labeling.

Table 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. Table 2 shows the statistics of the test sets for each dataset. We have organized both Kor-SemEval and KR3 data as open-source to facilitate their use in various training and evaluation scenarios.

3.4 Metrics

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

3.5 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 et al., 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 et al., 2020), we fine-tuned with a batch size of 16, and the results from epoch 4 were reported.

4 Experiment

4.1 Exp-1: Kor-SemEval

We conducted evaluations for each of the mBERT-single, XLM-R_{Base}-single, mBERT-NLI, XLM-R_{Base}-NLI, and NLI-ensemble models. As there is no officially converged dataset and model research specifically for Korean ABSA, we included the results from the previous SemEval14 research and Kor-SemEval to compare and evaluate the performance in Korean.

4.2 Results

Results on Kor-SemEval are presented in Table 3 and Table 4. 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_{Base}

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

in NLI. The XLM-R_{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 evaluated on **Kor-SemEval***			
mBERT-single	92.16	77.95	84.46
XLM-R_{Base}-single	91.01	49.37	64.01
mBERT-NLI	91.10	79.90	85.14
XLM-R_{Base}-NLI	91.37	**83.71**	**87.37**
NLI-ensemble	**93.70**	81.27 (↓)	87.04 (↓)

Table 3: Test set results for Aspect Category Detection. We use the results reported in BERT-single and BERT-pair-NLI-M (Sun et al., 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 evaluated on **Kor-SemEval***			
mBERT-single	68.20	71.84	79.52
XLM-R_{Base}-single	62.93	66.29	75.20
mBERT-NLI	73.95	77.90	84.87
XLM-R_{Base}-NLI	**79.41**	**83.66**	**89.98**
NLI-ensemble	78.24 (↓)	82.43 (↓)	89.65 (↓)

Table 4: Test set accuracy (%) for Aspect Category Polarity. We use the results reported in BERT-single and BERT-pair-NLI-M (Sun et al., 2019) for English dataset together with our results.

4.3 Exp-2: KR3 Test Set

Furthermore, based on the results from Kor-SemEval, we examined the task-specific dissimilarity between mBERT and XLM-R_{Base}. Accordingly, we opted for the XLM-R_{Base}-NLI approach, which demonstrated the best performance, as the base model for Phase 2.

We conducted evaluations on KR3 test data using the pseudo-labeled KR3 trainset (PL), the model trained with the original Kor-SemEval and additional fine-tuning with KR3 train (TR+PL), and corpus obtained through thresholding and LaBSE-based filtering on KR3 train (PL-CF).

4.4 Results

To investigate the effect of features for each corpus, we conduct tuning comparisons between the baseline's full data and the filtered pseudo data, as indicated in Table 5. The variants of our tuning framework includes:

- **Baseline+PL (Pseudo Labeled data)** : Fine-tuning the untuned baseline with the full pseudo KR3.
- **Baseline+PL-CF (Corpus Filtering)** : Fine-tuning the untuned baseline with the data obtained by **truncating the instance from entire pseudo KR3**, where the softmax threshold is less than 0.5 and the cosine similarity between LaBSE embeddings is less than 0.15.
- **Baseline+TR (TRanslated data)+PL** : Fine-tuning the tuned baseline from Experiment 1 on Kor-SemEval with the full pseudo KR3.
- **KPC-CF (Baseline+TR+PL-CF)** : Fine-tuning the tuned baseline from Experiment 1 on Kor-SemEval with PL-CF.

Results on the KR3 test set are presented in Table 5 and Figure 3. 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, while excelling at reflecting cross-lingual representations, shows an underfitting tendency in the context differences of aspect vocabulary in a single task. This can be understood as an issue of data scarcity relative to the model availability for each classifier, or as a limitation of the single text classification capability using the SPM in low-resource Korean ABSA. However, in the NLI task, it demonstrates potential by overpowering mBERT performance, guided by the instruction "aspect". mBERT, on the other hand, displays stable results in both single and NLI tasks, with an overall increase in accuracy, especially in the NLI task. Furthermore, in phase 2, it became evident

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 5: KR3 test set results for Aspect Category Detection (middle) and Aspect Category Polarity (right).

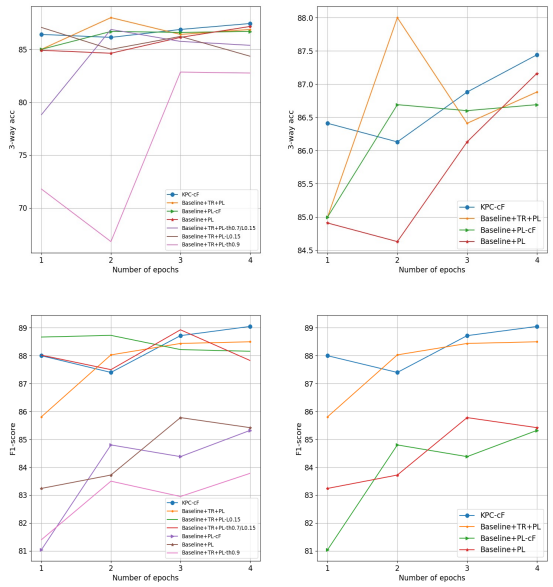

Figure 3: Performance of ACD and ACP during fine-tuning on KR3 test data. Left: results with the addition of other fine-tuned models; Right: four models compared in this paper. Blue line represents our proposed model, KPC-cF.

that the combination of NLI approach and translated data significantly influences the metrics of the model exploring aspects. Pseudo-labels in this phase contribute to improving the binary classification of sentiment, allowing classifiers to perform better. Moreover, the finely filtered pseudo-labels, unlike simply adding pseudo-labels to the translated data, contribute to maintaining and enhancing excellent accuracy and F1 score.

Our model and corpus can be utilized in the following ways: When developing LMs for ABSA in personal research or industry, there is a significant challenge posed by the absolute lack of labels. Our pseudo-labeled NLI corpus has been meaningfully filtered from the perspective of language-agnostic embeddings. Tuned alongside translated data, our KPC-cF can not only be directly applied to the web but also serve as a foundational model effectively supporting the automatic labeling and classification tasks for constructing more meaningful Korean Aspect-Based Sentiment Analysis (ABSA) data from reviews.

Furthermore, we intend to utilize Kor-SemEval and KR3 for subsequent research in Korean ABSA. In the current situation, where there is a desire to enhance the cross-linguality of language models, it is crucial to accurately encompass the diversity of nuances in the target language and the quality polarity information that can exist within sentences in ABSA. In the future, after fully constructing KR3, we can compare it with translated data to propose a Korean benchmark for measuring aspects and polarity.

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 issue of label scarcity in Korean ABSA by constructing a translated dataset and a pseudo-labeled actual Korean dataset. We utilized the common ML concept of semi-supervised learning, along with LaBSE-based filtering, to fine-tune a crosslingual model for the sentence pair classification task in Korean ABSA, achieving state-of-the-art results. We compared the experimental results of single sentence classification and sentence pair classification, as well as the combination of mBERT and XLM-RoBERTa, analyzing the advantages of each classifying method to validate the effectiveness of the transformation approach in crosslingual models.

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, evalua-

tion results, and metrics.

Acknowledgements

We thank Chanseo Nam for manually translating 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.

References

Annisa Nurul Azhar and Masayu Leylia Khodra. 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)*, pages 1–6. IEEE.

Alexandra Balahur and Marco Turchi. 2012. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 52–60.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Cambria Erik, Poria Soujanya, Gelbukh Alexander, and Thelwall Mike. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Dongmin Hyun, Junsu Cho, and Hwanjo Yu. 2020. Building large-scale english and korean datasets for aspect-level sentiment analysis in automotive domain. In *Proceedings of the 28th international conference on computational linguistics*, pages 961–966.

Hamoon Jafarian, Amir Hossein Taghavi, Alireza Javaheri, and Reza Rawassizadeh. 2021. Exploiting bert to improve aspect-based sentiment analysis performance on persian language. In *2021 7th International Conference on Web Research (ICWR)*, pages 5–8. IEEE.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2020. Improving bert performance for aspect-based sentiment analysis. *arXiv preprint arXiv:2010.11731*.

Hyunjae Lee, Jaewoong Yoon, Bonggyu Hwang, Seongho Joe, Seungjai Min, and Youngjune Gwon. 2021. Korealbert: Pretraining a lite bert model for korean language understanding. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5551–5557. IEEE.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 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*, pages 1603–1610.

Nankai Lin, Yingwen Fu, Xiaotian Lin, Dong Zhou, Aimin Yang, and Shengyi Jiang. 2023. Cl-xabsa: Contrastive learning for cross-lingual aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Annisa Nurul Azhar and Masayu Leylia Khodra. 2021. Fine-tuning pretrained multilingual bert model for indonesian aspect-based sentiment analysis. *arXiv e-prints*, pages arXiv–2103.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Khoa Thi-Kim Phan, Duong Ngoc Hao, Dang Van Thin, and Ngan Luu-Thuy Nguyen. 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)*, pages 1–6. IEEE.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. *Proceedings of SemEval*, pages 19–30.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495.

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

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*.

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

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*.

Joanna Szołomicka and Jan Kocon. 2022. Multiaspectemo: Multilingual and language-agnostic aspect-based sentiment analysis. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 443–450. IEEE.

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

Yige Xu, Xipeng Qiu, Ligao Zhou, and Xuanjing Huang. 2020. Improving bert fine-tuning via self-ensemble and self-distillation. *arXiv preprint arXiv:2002.10345*.

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

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16):3389.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9220–9230.

Yan Zhou, Fuqing Zhu, Pu Song, Jizhong Han, Tao Guo, and Songlin Hu. 2021. An adaptive hybrid framework for cross-domain aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14630–14637.