Exploring Multilingual Pre-trained Language Model for Aspect-based Sentiment Analysis

Ye Wang, Zhongqing Wang*, Ruijun Jiang and **Guodong Zhou**
Natural Language Processing Lab, Soochow University, Suzhou, China
{ywangxiaolu, rjjiang}@stu.suda.edu.cn
{gdzhou, wangzq}@suda.edu.cn

Abstract

Aspect-based sentiment analysis has garnered increasing attention in the research community; however, most studies have predominantly focused on English datasets, with other languages such as Chinese, Japanese, and German being neglected due to the limited availability of adequately labeled data. Even within English, labeled data is scarce. To address these challenges, this study investigates the utilization of a multilingual pre-trained setting to leverage resources from diverse languages for aspect-based sentiment analysis. Specifically, we propose a *Cross-lingual Knowledge Fusion* framework that explores various single-round and two-round bilingual pre-training configurations. This framework utilizes both the original and translated texts, along with their corresponding labels, to pre-train the multilingual model. Evaluation results reveal that our model significantly outperforms state-of-the-art performance across multiple languages, highlighting the effectiveness of the proposed multilingual pre-trained language model for aspect-based sentiment analysis.

1 Introduction

Aspect-based sentiment analysis (ABSA) represents a fine-grained approach to text sentiment analysis, which pinpoints sentiment information pertinent to specific aspects and offers businesses and organizations deeper market insights. In recent years, ABSA has attracted growing attention and interest within the research community.

Previous studies have demonstrated significant advancements in ABSA by utilizing pre-trained encoder-decoder language models (Zhang et al., 2021a). These studies have adopted various strategies, treating the class index (Yan et al., 2021), natural language descriptions (Zhang et al., 2021a), or the desired sentiment element sequence (Zhang et al., 2021c; Bao et al., 2022; Gou et al., 2023) as the target for the generation model. This innovative approach presents a promising avenue for ABSA research, as it seeks to alleviate the challenges posed by error propagation in traditional methods (Qiu et al., 2011; Cai et al., 2021; Fan et al., 2025; Su et al., 2025).

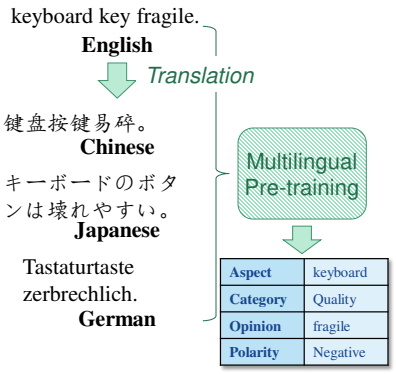

Figure 1: An example of multilingual pre-training.

However, most previous studies have primarily concentrated on English datasets, overlooking other languages like Chinese, Japanese, and German. This can be attributed to the scarcity of adequately labeled data in these languages. Moreover, even in English, the availability of labeled data is limited, as each review necessitates the annotation of a quadruple (i.e., aspect, opinion, category, polarity). This dearth of labeled data hinders the performance of ABSA across these languages. In English, too, the performance remains constrained, with the average F1 score even falling below 50%.

To address these limitations, we investigate the application of a *multilingual pre-trained setting* to harness the resources of different languages for ABSA. As shown in Figure 1, by extracting and integrating information from various languages and learning the intrinsic correlations among them, multilingual pre-trained models can effectively leverage the rich semantic representa-

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 6011–6023

tion inherent in multiple languages.

In particular, we propose a *Cross-lingual Knowledge Fusion* framework that leverages multilingual data to enhance the performance of aspect-based sentiment analysis. As illustrated in Figure 2, our approach begins by translating the original review text into multiple target languages, thereby enriching the input with diverse linguistic expressions. We then explore various single-round and two-round bilingual pre-training configurations, utilizing both the original and translated texts along with their corresponding labels to pre-train the multilingual model. Finally, the aspect-based sentiment analysis model is fine-tuned using the bilingual pre-trained configurations, allowing it to generalize better and achieve a deeper understanding of nuanced sentiments expressed in different languages.

The advantages of our proposed framework are twofold: these bilingual pre-training configurations not only enhance the model's capability to understand and generate translations across languages, but they also facilitate the alignment of sentences with labels across different languages. This, in turn, improves both intralingual and cross-lingual semantic representations for aspect-based sentiment analysis.

Detailed evaluations reveal that our model significantly outperforms the state-of-the-art performance across multiple languages. The results also highlight the effectiveness of the proposed multilingual pre-trained language model for aspect-based sentiment analysis.

2 Related Works

Aspect-based sentiment analysis (ABSA) encompasses multiple sub-tasks, including aspect term extraction (Qiu et al., 2011; Tang et al., 2016), aspect category identification (Zhou et al., 2015b), and sentiment polarity prediction (Wang et al., 2016). Recent research increasingly focuses on jointly modeling these elements to capture their interdependencies (Peng et al., 2020; Cai et al., 2021).

Advancements in ABSA leverage pre-trained encoder-decoder models, recasting the task as a generation problem. Outputs are generated as class indices (Yan et al., 2021), natural language descriptions (Zhang et al., 2021a), or structured sequences (Zhang et al., 2021c). Syntactic structures (Fan et al., 2025) and adaptive graph diffusion networks (Su et al., 2025) have further improved performance.

Cross-lingual ABSA research focuses on data alignment and embedding learning. Early methods use translation-based pseudo-labeling (Zhou et al., 2015a; Zhang et al., 2021b), while recent approaches emphasize shared semantic representations via multilingual embeddings (Ruder et al., 2019; Jebbara and Cimiano, 2019) and contrastive learning (Lin et al., 2023). Multi-scale optimization frameworks (Wu et al., 2025) enhance cross-lingual alignment. Recent studies explore using large language models (LLMs) to alleviate data scarcity in cross-lingual ABSA. Šmíd et al. (Šmíd et al., 2025) propose LACA, employing LLMs to generate high-quality pseudo-labeled data in the target language, enabling effective knowledge transfer without external translation systems. In a follow-up work, Šmíd et al. (Šmíd et al., 2025) investigate seq-to-seq formulations for compound ABSA tasks and introduce constrained decoding for valid cross-lingual outputs.

Different from these LLM-centric approaches that primarily focus on data augmentation or decoding constraints, our work emphasizes multilingual pre-training through cross-lingual knowledge fusion, leveraging both original and translated resources during pre-training to learn language-agnostic representations.

3 Cross-lingual Knowledge Fusion Framework

In this study, we propose a *Cross-lingual Knowledge Fusion* (CKF) framework to leverage multilingual data for enhancing aspect-based sentiment analysis. The core idea is to integrate knowledge across languages, thereby enriching the model's understanding of semantic patterns and improving its generalization ability.

As illustrated in Figure 2, our approach begins by translating the original review text into target languages. We then explore both single-round and two-round bilingual pre-training strategies, where the large language model is pre-trained on the original and translated texts along with their corresponding labels. After this bilingual pre-training phase, the model is fine-tuned for aspect-based sentiment analysis using the resulting pre-trained configurations. In the following subsections, we discuss these issues in detail. Our framework effectively leverages cross-lingual information to

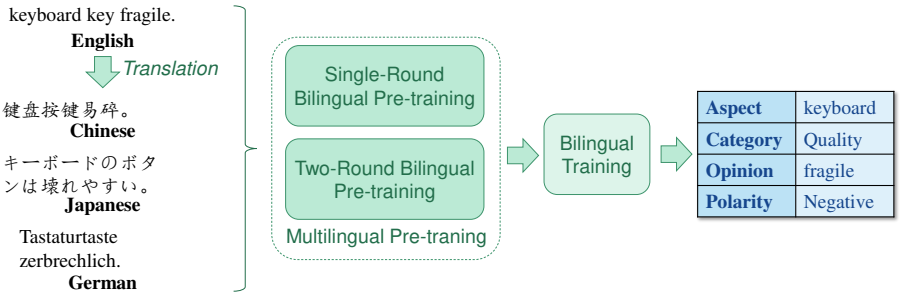

Figure 2: Overview of proposed framework.

capture richer semantic cues and enhance model performance across diverse linguistic contexts.

3.1 Translation of Original Review Text

For integrating multilingual information, it is essential to translate original text into target languages (Zhu et al., 2023). As an example, we take an English review text and translate it into Chinese, Japanese, and German. Specifically, we utilize a large language model for the translation task. To guide the model in generating high-quality translations while preserving label alignment, we adopt a ten-shot prompting approach. This approach leverages a small number of carefully chosen exemplars to instruct the model (Liu et al., 2021). The prompt template used in our method is structured as follows:

> Please translate the following sentences and their corresponding labels into target language. Ensure that the alignment of the labels in the translated text is maintained.

This template is designed to ensure that the model not only translates the sentences accurately but also preserves the semantic correspondence between the source and target languages. By providing ten exemplars that exhibit the desired translation style and alignment requirements, we enable the model to learn and apply these criteria effectively.

3.2 Single-Round Bilingual Pre-training

We initially present a variety of unique single-round pre-training configurations that incorporate not only the original sentences from the source language and their translated counterparts in the target language but also the respective labels during the pre-training phase. Specifically, given the source text T_S, the target text T_T, the source label Y_S, and the target label Y_T, these pre-training configurations are shown in Figure 3 and can be classified as follows:

- *Translation* ($T_S \rightarrow T_T$ and $T_T \rightarrow T_S$) configurations translate a sentence from the source language to the target language, and vice versa.
- *Sentence-To-Label* ($T_S \rightarrow Y_S$ and $T_T \rightarrow Y_T$) configurations maps a sentence to its target label.
- *Translated Sentence-To-Label* ($T_S \rightarrow Y_T$ and $T_T \rightarrow Y_S$) configurations map a source language sentence to its corresponding label in the target language, and vice versa, enabling cross-lingual label assignment.
- *Label-To-Sentence* ($Y_S \rightarrow T_S$ and $Y_T \rightarrow T_T$) configurations generate a language sentence from a corresponding label, reinforcing the model's ability to reconstruct sentences based on labels.
- *Translated Label-To-Sentence* ($Y_T \rightarrow T_S$ and $Y_S \rightarrow T_T$) configurations generate a sentence in the source language from a label in the target language, effectively mapping target labels onto the syntax and vocabulary of the source language, and vice versa.

In summary, by engaging in tasks such as forward and reverse translation, as well as label-to-sentence and sentence-to-label mappings, these single-round pre-training configurations gains insights into structural patterns, semantic subtleties, and cross-lingual alignment, ultimately leading to a deeper understanding of language-specific aspect-based sentiment analysis systems.

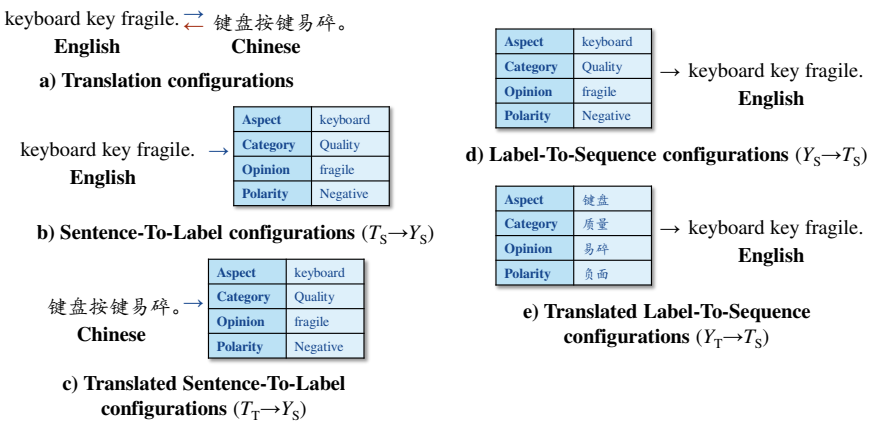

Figure 3: Examples of the single-round bilingual pre-training configurations.

3.3 Two-Round Bilingual Pre-training

We further propose two-round bilingual pre-training configurations to fully harness the advantages of cross-lingual knowledge transfer. Specifically, we design some distinct configurations, each meticulously designs to capitalize on the strengths exhibited by particular pre-training approaches that have proven highly effective in single-round pre-training scenarios.

Translation Fusion

The translation fusion approach seamlessly integrates the translation configurations: $T_T \rightarrow T_S$ and $T_S \rightarrow T_T$. This fusion strategy enhances the model's ability to reconstruct and translate sentences between the source and target languages, fostering bidirectional language understanding and generation.

Monolingual Reconstruction Fusion

The monolingual reconstruction fusion harmoniously integrates label-to-sentence configurations ($Y_T \rightarrow T_T$ and $Y_S \rightarrow T_S$). This method is designed to enhance the model's comprehension of the intricate structural and semantic relationships that exist within both the target and source languages.

By focusing on label-to-sentence reconstruction in both the target and source languages, this fusion approach enables the model to develop a deep understanding of the structural and semantic relationships within each language.

Cross-lingual Reconstruction Fusion

The cross-lingual reconstruction fusion approach is designed to seamlessly integrate the two translated label-to-sentence configurations ($Y_T \rightarrow T_S$ and $Y_S \rightarrow T_T$).

By combining these two tasks within a single framework, the model is compelled to learn a more comprehensive mapping between labels and sentences across languages. This integration fosters a deeper understanding of the semantic nuances and structural differences between the source and target languages.

3.4 Aspect-based Sentiment Analysis

After bilingual pre-training, we fine-tune the aspect-based sentiment analysis model using the corresponding monolingual settings. Specifically, we introduce a sequence-to-sequence model that is tailored to generate sentiment elements. This model employs a transformer-based encoder-decoder framework, as outlined in (Zhang et al., 2021a).

In particular, given the source language sequence $x = x_1, ..., x_{|x|}$ as input, the sequence-to-sequence model outputs the linearized representation $y = y_1, ..., y_{|y|}$. To this end, the sequence-to-sequence model first computes the hidden vector representation $H = h_1, ..., h_{|x|}$ of the input via a multi-layer transformer encoder:

$$H = \text{Encoder}(x_1, ..., x_{|x|}) \quad (1)$$

where each layer of Encoder is a transformer block with the multi-head attention mechanism.

After the input token sequence is encoded, the decoder predicts the output structure token-by-token with the sequential input tokens' hidden vectors. At the i-th step of generation, the self-attention decoder predicts the i-th token y_i in the

linearized form, and decoder state h_i^d as:

$$y_i, h_i^d = \text{Decoder}([H; h_1^d, ..., h_{i-1}^d], y_{i-1}) \quad (2)$$

where each layer of Decoder is a transformer block that contains self-attention with decoder state h_i^d and cross-attention with encoder state H.

The generated output structured sequence starts from the start token "$\langle bos \rangle$" and ends with the end token "$\langle eos \rangle$". The conditional probability of the whole output sequence $p(y|x)$ is progressively combined by the probability of each step $p(y_i|y_{<i}, x)$:

$$p(y|x) = \prod_{i=1}^{|y|} p(y_i|y_{<i}, x) \quad (3)$$

where $y_{<i} = y_1 ... y_{i-1}$, and $p(y_i|y_{<i}, x)$ are the probabilities over target vocabulary V normalized by softmax.

3.5 Bilingual Training

For bilingual training, we initially pre-train monolingual models, each with its own set of parameters, θ_T for the target language and θ_S for the source language, utilizing their respective monolingual corpora. Following the independent training of these models, we proceed to integrate their parameters in order to establish a unified bilingual model.

To achieve this, we introduce low-rank updates $\Delta\theta_T$ and $\Delta\theta_S$ using the LoRA (Hu et al., 2021) method:

$$\theta'_T = \theta_T + \Delta\theta_T, \quad \theta'_S = \theta_S + \Delta\theta_S \quad (4)$$

The final merged model is obtained via a weighted combination of the adapted parameters:

$$\theta_{\text{Fusion}} = \alpha_T \theta'_T + \alpha_S \theta'_S \quad (5)$$

where α_T and α_S are scaling factors that balance the contributions from the two pre-trained models. To ensure the merged model maintains the knowledge from both pre-training phases while being functionally coherent, we minimize the weighted sum of task-specific losses:

$$\mathcal{L}_{\text{Fusion}}(\theta) = \sum_t \lambda_t \sum_{(x,y)\in\mathcal{D}_t} \text{loss}(f_\theta(x), y) \quad (6)$$

Here, λ_t controls the relative importance of each task during loss computation. After merging, we apply a lightweight adaptation step, where only the LoRA-updated parameters are optimized:

$$\theta_{\text{Fusion}} \leftarrow \theta_{\text{Fusion}} - \eta \nabla_\theta \mathcal{L}_{\text{Fusion}}(\theta) \quad (7)$$

Language	Train	Test	Quadruples
English	2,934	816	3658
Chinese	2,798	799	11463
Japanese	2,566	642	6230
German	3,119	1,028	3572

Table 1: Number of sentences and extracted quadruples in the proposed dataset for four languages.

4 Experiments

In this section, we introduce the datasets used for evaluation and the baseline methods employed for comparison. We then report the experimental results conducted from different perspectives, and analyze the effectiveness of the proposed model with different factors.

4.1 Data

In this study, we utilize datasets in four languages, each covering a distinct domain. Specifically, the *English* dataset is sourced from the laptop domain within the ACOS dataset (Cai et al., 2021), the *Chinese* dataset is derived from the mixed Chinese CCD dataset (Wang et al., 2023), the *Japanese* dataset pertains to the economic domain in the chABSA dataset[1], and the *German* dataset originates from the transportation domain in the MobASA dataset (Gabryszak and Thomas, 2022). The detailed statistics for each language are presented in Table 1. For training, the original training data of each dataset was independently split into a 90% training set and a 10% validation set.

4.2 Setting

Since the sentiment elements in the datasets for different languages vary, we conduct the Aspect-Sentiment-Quad Prediction task for English and Chinese, and employ the Target Aspect Sentiment Detection task for the remaining two languages. For each task, we utilize LLaMA-8B-Instruct[2] (Touvron et al., 2023) as the pre-trained generation model. The model is fine-tuned using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 3e-4 and a batch size of 16 for 10 epochs. All experiments are conducted on a single NVIDIA RTX 4090 GPU. The reported results are averaged over ten runs with random initialization.

[1] https://github.com/chakki-works/chABSA-dataset

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Methods	English			Chinese			Japanese			German		
	P.	**R.**	**F1.**	**P.**	**R.**	**F1.**	**P.**	**R.**	**F1.**	**P.**	**R.**	**F1.**
TAS-BERT	47.15	19.22	27.31	33.50	32.99	33.24	30.96	30.29	30.62	46.45	25.73	33.12
Extract-Classify	45.46	29.48	35.80	47.44	41.93	44.51	-	-	-	-	-	-
GAS	41.60	42.75	42.17	73.13	70.96	72.03	68.99	61.61	65.09	78.81	73.53	76.08
Paraphrase	41.77	45.04	43.34	74.32	76.05	75.17	62.75	65.38	64.04	77.60	77.73	77.67
Seq2Path	42.51	43.17	42.84	72.81	76.42	74.51	64.12	65.25	64.48	75.31	80.88	77.96
MvP	44.20	43.70	43.94	76.60	76.17	76.38	68.05	67.14	67.59	79.20	78.89	79.04
Seq2Seq w/ CD	42.89	42.05	42.46	74.50	76.29	75.38	63.67	64.91	64.27	77.61	77.95	77.78
LACA*	45.32	44.46	44.88	77.19	74.86	76.02	69.17	70.90	70.01	81.26	82.34	81.78
DeepSeek-V3	19.64	20.16	19.89	37.39	28.70	32.47	25.12	30.02	27.35	49.76	46.86	48.27
ChatGPT-5	15.34	17.40	16.30	32.04	41.11	36.01	21.62	26.05	23.63	54.10	60.78	57.25
LLaMA-3.1-8B	43.44	43.25	43.34	75.14	75.19	75.17	68.74	70.51	69.48	80.93	81.56	81.24
Qwen-2.0-7B	43.75	42.54	43.14	75.38	77.18	76.27	68.15	68.15	68.15	80.05	80.60	80.32
Ours	**46.63**	**46.39**	**46.51**	**75.99**	**78.55**	**77.25**	**71.43**	**72.94**	**72.18**	**82.94**	**84.05**	**83.49**

Table 2: Comparison with baselines. Some baseline results are taken from prior published work, and others are reproduced by us. The symbol "–" indicates that the method is not applicable to the Target Aspect Sentiment Detection task. LACA* denotes our adaptation that specifically utilizes its core LLM-based data augmentation mechanism.

In evaluation, a prediction is considered correct if and only if all elements within the predicted structure exactly match those in the corresponding gold structure, including their combination. On this basis, we calculate the Precision and Recall, and use F1 score as the final evaluation metric for aspect sentiment extraction (Cai et al., 2021; Zhang et al., 2021a).

4.3 Main Results

As shown in Table 2, we compare our proposed model against several robust baseline approaches, which can be classified into three categories: classification-based models, pre-trained generative models, and large language models. The classification-based models include **TAS-BERT** (Wan et al., 2020), and **Extract-Classify** (Cai et al., 2021). The pre-trained generative models comprise **GAS** (Zhang et al., 2021c), **Paraphrase** (Zhang et al., 2021a), **Seq2Path** (Mao et al., 2022), **MvP** Gou et al. (2023), **Seq2Seq w/ CD** (Šmíd et al., 2025), **LACA** (Šmíd et al., 2025). The large language models used for comparison include **ChatGPT-5** (Brown et al., 2020), and **DeepSeek-V3** (DeepSeek-AI et al., 2025). These models are evaluated under a ten-shot learning setting, with prompts designed following. In addition, we also compare with **LLaMA-3.1-8B** (Touvron et al., 2023) and **Qwen-2.0-7B** (Yang et al., 2024), both of which are fine-tuned using the LoRA method (Hu et al., 2021).

Based on the results, it is clear that generative methods outperform discriminative methods by employing a unified architecture that effectively incorporates rich label semantics into the output. On the other hand, large language models may face challenges such as contextual ambiguity, difficulties in accurately identifying specific aspects, and constraints in the diversity of training data. Moreover, our proposed model significantly outperforms all baseline methods ($p < 0.05$) across all the languages. This not only highlights the efficacy of our model but also emphasizes the importance of integrating multilingual information in aspect-based sentiment analysis.

4.4 Impact of Pre-Training Methods

We subsequently explore different pre-training methods mentioned in Section 3, including all single-round and two-round bilingual pre-training approaches, with Chinese as the target language.

As shown in Table 3, cross-lingual pre-training substantially improves performance over the monolingual ABSA baseline. Among single-round settings, $Y_S \rightarrow T_S$ yields the best English F1, showing its strength in reconstructing sentiment-rich text from labels. Fusion-based strategies further boost results across all languages, with our method achieving the highest overall performance. These findings demonstrate that integrating complementary semantic signals from both source and target domains enhances cross-lingual sentiment alignment, particularly for distant language pairs.

Methods		English	Chinese	Japanese	German
Monolingual ABSA	-	43.34	75.17	69.48	81.24
Translation	$T_S \rightarrow T_T$	44.63	76.66	70.94	82.09
	$T_T \rightarrow T_S$	45.44	76.31	70.50	82.49
Sentence-To-Label	$T_S \rightarrow Y_S$	43.38	75.41	69.86	81.51
	$T_T \rightarrow Y_T$	45.51	76.27	68.70	81.47
Translated Sentence-To-Label	$T_S \rightarrow Y_T$	45.67	77.14	71.13	82.58
	$T_T \rightarrow Y_S$	42.57	74.89	65.69	80.26
Label-To-Sentence	$Y_S \rightarrow T_S$	46.18	76.52	70.98	82.38
	$Y_T \rightarrow T_T$	44.80	76.10	70.77	82.11
Translated Label-To-Sentence	$Y_T \rightarrow T_S$	45.55	76.37	71.46	82.47
	$Y_S \rightarrow T_T$	45.33	76.66	70.89	81.98
Translation Fusion	$T_T \rightarrow T_S + T_S \rightarrow T_T$	45.96	76.95	71.24	82.61
Monolingual Reconstruction Fusion	$Y_T \rightarrow T_S + Y_S \rightarrow T_S$	46.07	76.73	71.61	82.66
Ours	$T_S \rightarrow Y_T + T_T \rightarrow Y_S$	**46.51**	**77.25**	**72.18**	**83.49**

Table 3: Impact of different pre-training configurations on the English, Chinese, Japanese and German datasets. T_S and T_T denote the source and target texts, respectively, while Y_S and Y_T denote the source and target labels, respectively. The superior results are highlighted in red, and the optimal performance is indicated in bold.

5 Analysis and Discussion

In this section, we further give several analyses and discussions to show the importance of proposed framework.

5.1 Impact of Target Languages

As shown in Table 4, we further investigate the impact of different target languages using English, Chinese, Japanese, and German as source languages. The results reveal that linguistic proximity plays a key role in cross-lingual transfer.

When English is the source, German performs best, likely due to typological similarity. While Japanese and Chinese underperform individually, their combination achieves the highest F1 score in multilingual settings, suggesting that typologically distant languages can offer complementary signals. However, combining German with either Japanese or Chinese causes slight performance drops, indicating interference between similar and distant language pairs. Similar trends emerge when using Chinese, Japanese, or German as the source, with overall gains linked to linguistic similarity and alignment quality. Overall, the results indicate that linguistically similar languages provide more stable transfer gains, whereas distant languages offer complementary benefits only when combined judiciously.

5.2 Analysis of Translation Quality

We conduct an in-depth analysis of translation quality across different models. In addition to our translation model, we evaluate Qwen-Plus (Yang et al., 2024), LLaMA-3.1-8B (Touvron et al., 2023), and ChatGPT-4o-mini (Brown et al., 2020). We randomly sample 500 instances and manually annotate reference translations for each target language. BLEU scores (Papineni et al., 2002) are computed by comparing model outputs with human-annotated references following the standard BLEU protocol.

As shown in Table 5, all models achieve reasonably strong translation performance across languages, indicating that the overall translation quality is sufficient for downstream processing. Nevertheless, our model consistently achieves the best or highly competitive BLEU scores across all four languages. This observation is further reflected in its superior performance on downstream tasks, suggesting a positive correlation between translation quality and task effectiveness.

5.3 Results of Different Multilingual Integration Methods

To deeply analyze the effect of our proposed multilingual pre-training framework, we compared our model with several other multilingual integration methods, which include: 1) *MultiInput*: directly concatenate the target language with the source language in the input; 2) *MultiOutput*: directly output quadruples with both the source language and target languages; 3) *MultiInOut* combine the methods from 1) and 2). In particular, English is used as the source language with Chinese as the target. Similarly, for Chinese, Japanese, and German, English serves as the target language.

Methods	EN	Methods	CN	Methods	JP	Methods	DE
Monolingual	43.34	Monolingual	75.17	Monolingual	69.48	Monolingual	81.24
+ JP	44.48	+ EN	76.19	+ EN	70.77	+ EN	82.11
+ CN	44.80	+ JP	76.27	+ CN	70.89	+ JP	81.19
+ DE	45.99	+ DE	75.76	+ DE	71.07	+ CN	82.20
+ JP & CN	46.25	+ JP & EN	75.98	+ CN & EN	71.13	+ CN & EN	82.24
+ JP & DE	44.85	+ JP & DE	76.04	+ CN & DE	71.25	+ CN & JP	82.87
+ CN & DE	45.76	+ EN & DE	76.83	+ EN & DE	71.15	+ EN & JP	81.35

Table 4: F1 scores of different multilingual integration methods across English (EN), Chinese (CN), Japanese (JP) and German (DE) datasets.

Model	English				Chinese			
	P	R	F1	BLEU	P	R	F1	BLEU
Qwen-Plus	45.89	44.84	44.36	36.61	75.66	77.49	76.56	20.22
LLaMA-3.1-8B	44.91	46.15	45.52	37.31	75.78	78.91	77.32	**22.14**
ChatGPT-4o-mini	45.48	44.62	45.04	32.87	75.11	78.25	76.65	18.76
Ours	46.63	46.39	**46.51**	**37.71**	75.99	78.55	**77.25**	21.81
Model	**Japanese**				**German**			
	P	R	F1	BLEU	P	R	F1	BLEU
Qwen-Plus	67.85	69.50	68.67	28.78	82.02	81.74	81.88	43.77
LLaMA-3.1-8B	70.92	72.21	71.46	30.64	81.89	82.96	82.42	48.69
ChatGPT-4o-mini	68.74	70.92	69.81	27.35	81.52	81.30	81.41	43.50
Ours	71.43	72.94	**72.18**	**31.41**	82.94	84.05	**83.49**	**51.99**

Table 5: Quantitative evaluation of translation quality. We report Precision (P), Recall (R), F1-score, and BLEU for four languages. BLEU scores are computed on translations generated via a ten-shot prompt for each model.

Methods	EN	CN	JP	DE
Monolingual	43.34	75.17	69.48	81.24
+ *MultiInput*	42.49	76.61	71.52	81.75
+ *MultiOutput*	43.54	76.02	69.13	81.46
+ *MultiInOut*	44.21	76.45	69.91	82.49
Ours	**46.51**	**77.25**	**72.18**	**83.49**

Table 6: F1 scores of different multilingual integration methods across English (EN), Chinese (CN), Japanese (JP) and German (DE) datasets.

Table 6 compares various multilingual integration strategies. All approaches outperform the monolingual baseline, with *MultiInOut* achieving the best results by combining input and output augmentation. Our proposed multilingual pre-training framework, however, consistently surpasses all baselines across English, Chinese, Japanese, and German. This advantage primarily stems from our specially designed cross-lingual alignment pre-training tasks. Notably, even though other methods use the same bilingual data, their gains remain relatively modest, highlighting the distinctive effectiveness of our pre-training design.

The more detail analysis of case study, computational cost analysis, and the application of our proposed model on NER can be found in Appendix.

6 Conclusion

In this study, we investigate the utilization of a multilingual pre-trained setting to leverage resources from diverse languages for aspect-based sentiment analysis. Specifically, we propose a *Cross-lingual Knowledge Fusion* framework that explores various single-round and two-round bilingual pre-training configurations. This framework utilizes both the original and translated texts, along with their corresponding labels, to pre-train the multilingual model. Evaluation results reveal that our model significantly outperforms state-of-the-art performance across multiple languages, highlighting the effectiveness of the proposed multilingual pre-trained language model for aspect-based sentiment analysis.

Limitations

In this study, we focus on leveraging multilingual pretraining for aspect-based sentiment analysis. However, multilingual pretraining inevitably increases computational cost, and although LoRA alleviates parameter overhead, the overall training and inference time still grows as more languages are incorporated.As a result, scaling the proposed approach to a larger set of languages remains challenging in resource-limited environments. In addition, ensuring high-quality translations requires human verification, which introduces additional time consumption and manual effort. Finally, it is necessary to further evaluate the proposed model across a wider range of domains and languages to better assess its generalization capability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62376178), Jiangsu Key Laboratory of Language Computing(JSLCKeyLab 202500003) and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.This work was supported by the National Natural Science Foundation of China (No. 62376178), Jiangsu Key Laboratory of Language Computing(JSLCKeyLab 202500003) and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Xiaoyi Bao, Wang Zhongqing, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4044–4050. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Tom B. Brown, Ben Mann, Nick Ryder, Amanda Subbiah, Jack Kaplan, Prafulla Dhariwal, A. Neelakantan, K. Zirnheld, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Rui Fan, Shu Li, Tingting He, and Yu Liu. 2025. Aspect-based sentiment analysis with syntax-opinion-sentiment reasoning chain. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3123–3137, Abu Dhabi, UAE. Association for Computational Linguistics.

Aleksandra Gabryszak and Philippe Thomas. 2022. MobASA: Corpus for aspect-based sentiment analysis and social inclusion in the mobility domain.

In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 35–39, Marseille, France. European Language Resources Association.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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, Volume 1 (Long and Short Papers)*, pages 2486–2495, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

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*, 31:2935–2946.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2Path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2215–2225, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Jakub Šmíd, Pavel Priban, and Pavel Kral. 2025. LACA: Improving cross-lingual aspect-based sentiment analysis with LLM data augmentation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–853, Vienna, Austria. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto de Alencar Lotufo. 2019. Portuguese named entity recognition using bert-crf. *ArXiv*, abs/1909.10649.

Guixin Su, Yongcheng Zhang, Tongguan Wang, Mingmin Wu, and Ying Sha. 2025. Unified grid tagging scheme for aspect sentiment quad prediction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3997–4010, Abu Dhabi, UAE. Association for Computational Linguistics.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan. The COLING 2016 Organizing Committee.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hugo Touvron, David Vial, François Yvon, David Grangier, Sergey Edunov, and et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z. Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9122–9129.

Ye Wang, Yuan Zhong, Xu Zhang, Conghui Niu, Dong Yu, and Pengyuan Liu. 2023. *CCD-ASQP: A Chinese Cross-Domain Aspect Sentiment Quadruple Prediction Dataset*, pages 233–245.

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, Austin, Texas. Association for Computational Linguistics.

Chengyan Wu, Bolei Ma, Ningyuan Deng, Yanqing He, and Yun Xue. 2025. Multi-scale and multi-objective optimization for cross-lingual aspect-based sentiment analysis.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for 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 (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021b. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021c. 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 (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

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

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2015b. Representation learning for aspect category detection in online reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Wenhao Zhu, Hao Zhou, Changjiang Gao, Sizhe Liu, and Shu-Hua Huang. 2023. Research development of machine translation and large language models. In *China National Conference on Chinese Computational Linguistics*.

Jakub Šmíd, Pavel Přibáň, and Pavel Král. 2025. Advancing cross-lingual aspect-based sentiment analysis with llms and constrained decoding for sequence-to-sequence models. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*, page 757–766. SCITEPRESS - Science and Technology Publications.

A Case Study

To further investigate the effectiveness of cross-lingual knowledge fusion, we present two examples where our method successfully predicts all quadruples from the test dataset, as illustrated in Figure 4.

In the first example, the sentence "The keyboard is stiff and unresponsive" poses a challenge for models to differentiate between design-related and performance-related aspects. While the baseline model incorrectly categorizes both "stiff" and "unresponsive" under the general attribute quality of keyboard, the Chinese translation "键盘有时僵硬且对我的打字无响应" provides additional clarity. Specifically, the term "僵硬" (stiff) pertains to design features of the keyboard, whereas "无响应" (unresponsive) indicates an issue related to operation performance. This demonstrates how cross-lingual information can refine attribute categorization.

In the second example, the sentence "The battery lasts a while but drains faster than expected" often leads baseline models to misclassify the sentiment as neutral due to the initial positive clause "lasts a while". However, the Japanese translation 「バッテリーはしばらく持つが、予想より早く消耗する」 makes the negative sentiment

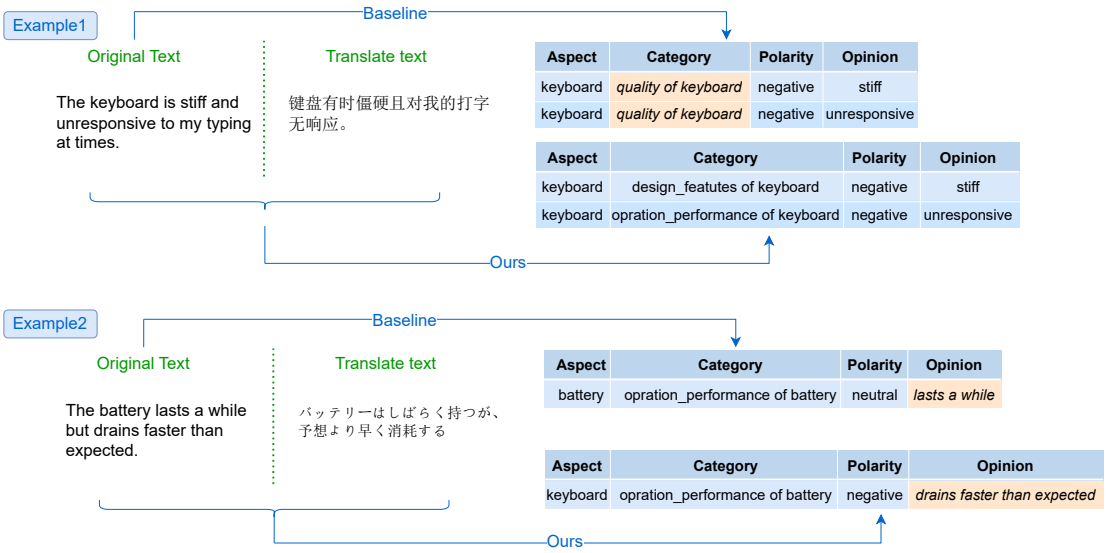

Figure 4: Examples of case study.

Model	English		Chinese	
	Train Time	Max GPU Usage	Train Time	Max GPU Usage
LLaMA-3.1-8B	41.50 min	19709 MiB	64.57 min	20905 MiB
Translation Fusion	83.93 min	22229 MiB	128.07 min	22583MiB
Ours	129.84 min	22833 MiB	190.48 min	23655MiB
Model	**Japanese**		**German**	
	Train Time	**Max GPU Usage**	**Train Time**	**Max GPU Usage**
LLaMA-3.1-8B	56.08 min	20839 MiB	46.54 min	18241 MiB
Translation Fusion	110.65 min	21787 MiB	100.90 min	20523 MiB
Ours	160.42 min	22941 MiB	155.42 min	22561 MiB

Table 7: Training Time and Peak GPU Memory Usage across English, Chinese, Japanese, and German Datasets.

more explicit through: (1) the strong contrastive particle 「が」 (but), which more sharply divides positive and negative clauses than English "but", and (2) the explicitly negative technical term 「消耗する」 (drains), which carries stronger negative connotations in product reviews. Our model correctly identifies the overall negative sentiment and associates it with the operation performance of the battery, demonstrating how Japanese grammatical markers can clarify ambiguous polarity cues in English.

B Computational Cost Analysis

Table 7 reports the training time and peak GPU memory usage of different models across four languages. Compared with the LLaMA-3.1-8B baseline, our cross-lingual knowledge fusion framework requires additional training time due to the incorporation of bilingual data and multi-round pre-training. However, the peak GPU memory usage remains largely comparable to the baselines. This indicates that our approach does not introduce extra model parameters or memory-intensive architectural components. Instead, the computational overhead mainly arises from increased training iterations and data processing, while maintaining similar memory efficiency.

C Application on NER

Named Entity Recognition (NER) involves identifying and classifying specific entities mentioned in text. We further evaluate our proposed multilingual pre-training model on the NER task (Peters et al., 2017; Souza et al., 2019).

Specifically, we use the CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003) to train and evaluate our model. For comparison, we select Qwen-Plus (Yang et al., 2024),

Method	F1.
Qwen-plus	31.65
Deepseek-v3	41.94
ChatGPT-4o-mini	30.00
T5-base	60.26
LLaMA-3.1-8B	79.13
Ours	**82.00**

Table 8: Results of different named entity recognition models.

DeepSeek-V3.1 (DeepSeek-AI et al., 2025), T5-base (Raffel et al., 2019), LLaMA-3.1-8B (Touvron et al., 2023), and ChatGPT-4o-mini (Brown et al., 2020) as baseline models. Among these baselines, Qwen, DeepSeek, and ChatGPT are evaluated in a few-shot setting, where the models rely on a small number of labeled examples for NER. In contrast, T5-base, LLaMA-3.1-8B, and our model are fully fine-tuned on the dataset.

As shown in Table 8, we can observe that our model significantly outperforms the other baselines. The substantial improvement of our multilingual pre-training model demonstrates its effectiveness in capturing entity-specific representations, enabling it to significantly enhance NER performance. These results highlight that our approach provides notable gains over both few-shot and standard fine-tuned models, underscoring the practical benefits of cross-lingual pre-training for downstream NLP tasks.