Test-Time Code-Switching for Cross-lingual Aspect Sentiment Triplet Extraction

Dongming Sheng Kexin Han Hao Li Yan Zhang
Yucheng Huang Jun Lang Wenqiang Liu
Tencent
{hanssheng,masonqliu}@tencent.com

Abstract

Aspect Sentiment Triplet Extraction (ASTE) is a thriving research area with impressive outcomes being achieved on high-resource languages. However, the application of cross-lingual transfer to the ASTE task has been relatively unexplored, and current code-switching methods still suffer from term boundary detection issues and out-of-dictionary problems. In this study, we introduce a novel **T**est-**T**ime **C**ode-**SW**itching (**TT-CSW**) framework, which bridges the gap between the bilingual training phase and the monolingual test-time prediction. During training, a generative model is developed based on bilingual code-switched training data and can produce bilingual ASTE triplets for bilingual inputs. In the testing stage, we employ an alignment-based code-switching technique for test-time augmentation. Extensive experiments on cross-lingual ASTE datasets validate the effectiveness of our proposed method. We achieve an average improvement of 3.7% in terms of weighted-averaged F1 in four datasets with different languages. Additionally, we set a benchmark using ChatGPT and GPT-4, and demonstrate that even smaller generative models fine-tuned with our proposed TT-CSW framework surpass ChatGPT and GPT-4 by 14.2% and 5.0% respectively.

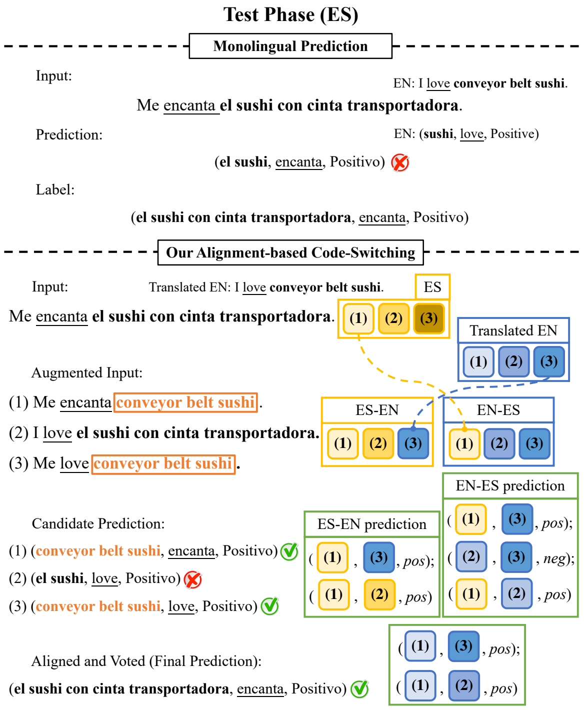

Test Phase (ES)
Monolingual Prediction
Input:
EN: I love conveyor belt sushi.
Me encanta el sushi con cinta transportadora.
Prediction:
EN: (sushi, love, Positive)
(el sushi, encanta, Positivo)
Label:
(el sushi con cinta transportadora, encanta, Positivo)
Our Alignment-based Code-Switching
Input:
Translated EN: I love conveyor belt sushi.
ES
Me encanta el sushi con cinta transportadora.
(1) (2) (3)
Translated EN
(1) (2) (3)
Augmented Input:
(1) Me encanta conveyor belt sushi.
(2) I love el sushi con cinta transportadora.
(3) Me love conveyor belt sushi.
ES-EN
(1) (2) (3)
EN-ES
(1) (2) (3)
EN-ES prediction
((1), (3), pos);
((2), (3), neg);
((1), (2), pos)
Candidate Prediction:
(1) (conveyor belt sushi, encanta, Positivo)
(2) (el sushi, love, Positivo)
(3) (conveyor belt sushi, love, Positivo)
ES-EN prediction
((1), (3), pos);
((1), (2), pos)
Aligned and Voted (Final Prediction):
(el sushi con cinta transportadora, encanta, Positivo)
((1), (3), pos);
((1), (2), pos)

Figure 1: An example of testing phase in cross-lingual ASTE task on Spanish dataset. Phrases with bold and underlined words represent **aspect** and opinion terms respectively. The substituted words are highlighted within the orange boxes. The diagram on the bottom right illustrates the pipeline of our proposed alignment-based code-switching method.

1 Introduction

Aspect sentiment Triplet Extraction (ASTE) task has drawn increasing attention in recent years (Peng et al., 2020; Xu et al., 2020; Zhang et al., 2023; Li et al., 2023). It aims at the co-extraction of aspect terms, opinion terms and sentiment polarities. Despite the success achieved on high-resource languages, it is still challenging to attain comparable performance for languages with limited annotation resources. This highlights the need for cross-lingual ASTE, an extended task commonly trained on languages with rich annotation resources (e.g., English) and tested on those with low resources (e.g., Basque and Catalan).

Recent studies have demonstrated that code-switching can effectively facilitate cross-lingual transfer for low-resource languages across various NLP tasks (Li et al., 2022; Zhang et al., 2021a; Qin et al., 2021; Zhu et al., 2023). However, in terms of cross-lingual ASTE task, current code-switching methods still suffer from two major issues.

Firstly, existing code-switching techniques are mainly used during the training phase as a method of data augmentation (Li et al., 2022; Zhang et al., 2021a), and the prediction is only based on the target language. However, the detection of term boundaries tends to be a challenge when conducting inference in a monolingual context for languages with limited annotation resources. As shown in Figure 1, *el sushi con cinta transportadora* means *conveyor belt sushi*(i.e., a type of sushi restaurant) in English. Due to limited annotation resources, the model may fail to recognize the aspect term *el sushi con cinta transportadora* as a whole, and instead predict *el sushi* only. This leads to the incorrect prediction of term boundaries.

Furthermore, in the context of the ASTE task, both opinion and aspect terms typically form phrases. This characteristic poses a challenge for existing code-switching methods that rely on bilingual dictionaries (Qin et al., 2021; Feng et al., 2022) or follow the translate-then-align procedure (Mayhew et al., 2017; Fei et al., 2020). For instance, *sound insulation* is not present in the bilingual dictionary, and translating *insulation* directly could result in semantic inaccuracies. Moreover, these terms can frequently be proper nouns, like the English brand name *Hard Rock Cafe*. However, these terms often fall outside the scope of bilingual dictionaries (i.e., out-of-dictionary issues), leading to issues of inconsistency and inaccuracies due to incorrect translations.

Based on the above observations, we propose a **T**est-**T**ime **C**ode-**SW**itching framework (**TT-CSW**) for the cross-lingual ASTE task. In our framework, the code-switching method offers a bilingual context in both training and testing phases. In this way, our framework can act as a bridge between the monolingual test-time prediction and the bilingual training phase.

For the training stage, our model learns to predict bilingual ASTE triplets based on code-switched inputs. To deal with the issue of out-of-dictionary and term boundary detection, we propose a boundary-aware code-switching method. This approach preserves the completeness of aspect and opinion terms during the translation process, circumventing problems associated with inconsistency and inaccuracies due to wrong translations from bilingual dictionaries. Consequently, it considerably enhances the alignment capability for models to understand bilingual context and predict term boundaries accurately.

During testing stage, our model can utilize knowledge from code-switching to generate triplets in the target language. To further address the incorrect prediction for term boundaries during test-time, we introduce a code-switching method based on alignment for test-time augmentation, as illustrated in Figure 1. A heuristic switching strategy is designed to generate a set of code-switched augmentation examples. The output triplets from these examples are then aligned into the target language for the final output. The integration of code-switching during the testing stage provides a bilingual multi-view of the input sentence, which incorporates information from source language with rich annotation resources, and improves performance for predicting term boundaries.

Extensive experiments on cross-lingual ASTE datasets validate the effectiveness of our proposed method. By integrating our method with various backbone models, we achieve an average improvement of 3.7% in terms of weighted-averaged F1 in four datasets with different languages. Furthermore, we benchmark ChatGPT [1] and GPT-4 [2], OpenAI's widely-used Large Language Model (LLM) and illustrate that small generative models finetuned with our proposed TT-CSW framework still outperform ChatGPT by 14.2% and 5.0% respectively in terms of weighted-averaged F1.

Our main contributions are as follows:

1) We propose a novel test-time code-switching framework for cross-lingual ASTE task, which can be easily integrated with various backbone models.

2) We develop a boundary-aware code-switching method based on translation system for solving the issue of out-of-dictionary and phrase code-switching.

3) We design an alignment method for test-time augmentation to improve term boundary prediction.

4) We benchmark ChatGPT and GPT-4 on cross-lingual ASTE task and show that small generative models finetuned with our TT-CSW framework can still outperform ChatGPT and GPT-4.

2 Methodology

Our proposed TT-CSW framework, as depicted in Figure 2, is made up of two key components: the training phase and the testing phase. The structure of the training phase, shown on the bottom side, is composed of three elements: boundary-aware code-switching, structural prediction, and alignment prediction. The layout of the testing phase, as illustrated on the upper side, involves two parts:

[1] https://chat.openai.com/

[2] https://openai.com/gpt-4

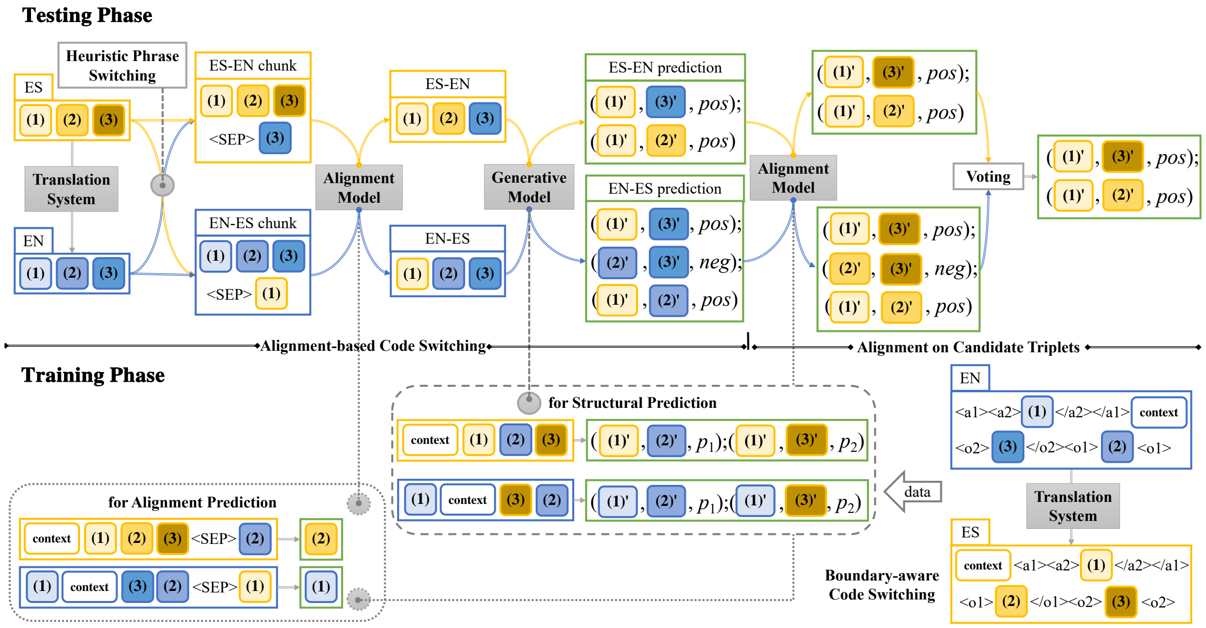


Figure 2: The overall architecture of our proposed TT-CSW framework.

alignment-based code-switching, and alignment of candidate triplets. In this section, we begin by providing a formal definition of the cross-lingual ASTE task, then we delve into the specifics of our proposed TT-CSW framework.

2.1 Task Definition

We denote an monolingual ASTE dataset with N samples as $D = \{D_1, D_2, ..., D_N\}$. For each sample $D_i = \{s_i, RT_i\}$, s_i represents the input sentence, and $RT_i = \{T_1, ..., T_n\}$ represents the ground truth triplet list for the input sentence. Each triplet $T_i = \{a_i, o_i, p_i\}$ consists of aspect term (a), opinion term (o) and sentiment polarity (p). The ASTE task aims to predict a list of m predicted triplets $PT_i = \{T_1, T_2, ..., T_m\}$ for each of the input sentence s_i. For cross-lingual ASTE, we denote the dataset for source language as $D^{(src)}$, and the dataset for target language as $D^{(tgt)}$. We need to train our model M on $D^{(src)}$ and perform inference on $D^{(tgt)}$.

2.2 Training Phase

For the training phase, we first create a bilingual code-switching dataset using our proposed boundary-aware code-switching method. Following this, we train two separate models utilizing this dataset: a generative bilingual model and a bilingual alignment model. The generative bilingual model is designed to produce bilingual ASTE triplets grounded on the bilingual context. Meanwhile, the role of bilingual alignment model is to convert the bilingual candidate triplets into the same language during the testing phase.

2.2.1 Boundary-aware Code-switching

Traditional methods for creating code-switching context rely on bilingual dictionary (Qin et al., 2021; Feng et al., 2022), or follow the procedure of translate-then-align (Mayhew et al., 2017; Fei et al., 2020), which utilizes word alignment tools after translation. Inspired by Zhang et al. (2021a), we propose a boundary-aware code-switching method via the translation system without the use of bilingual dictionaries or word alignment tools, as shown on the bottom right of Figure 2.

Given that there are often multiple triplets within a single sentence, we introduce the HTML tags to locate these triplets. We employ <a_i> and </a_i> to indicate the start and end of the i-th aspect term, and <o_i> and </o_i> to denote the boundaries of the i-th opinion term. Therefore, we can distinctively differentiate multiple aspect and opinion term pairs during the translation, while preserving their original boundaries. With the help of the HTML tags, we can easily construct bilingual parallel phrases for training the bilingual alignment model in section 2.2.3. Since the HTML tags are not part of the original sentence, we can easily remove them after the translation. The sentiment polarity remains unchanged during the translation, providing extra training examples for both the bilingual generative model and the bilingual alignment model.

2.2.2 Bilingual Structural Prediction

We develop a bilingual generative model trained on the boundary-aware code-switching dataset to perform the structural prediction task, as shown on the bottom center of Figure 2. To serialize the output triplets, we opt not to use commas or semicolons as separators or connectors within and between triplets, which is a standard practice in the GAS-Extraction format (Zhang et al., 2021b). The reason for this is that these symbols could also be present in aspect or opinion terms, leading to format confusion. Instead, we choose to use the special tokens <split> and <join> for this purpose. As an example, for a input sequence with two triplets (a_1, o_1, p_1) and (a_2, o_2, p_2), we use the following format for structural prediction: (a_1<split>o_1<split>p_1) <join> (a_2<split>o_2<split>p_2).

2.2.3 Bilingual Alignment Prediction

We use bilingual parallel phrases from boundary-aware code-switching dataset to train a bilingual alignment model, which will be used for testing phase in section 2.3. We use the mT5-base model as the backbone. For a pair of parallel aspect term a and $a^{(T)}$, we append the translated term $a^{(T)}$ to the end of original sentence and use a special token <SEP> to separate them, as depicted on the bottom left of Figure 2. We enhance computational efficiency and improve diversity by segmenting the original sentence into multiple chunks, each with a maximum length of 128, using a sliding window method. If the sentence chunk does not contain the original term a after segmentation, we treat it as a negative sample by setting the ground-truth label to *None*. Additionally, to further generate negative samples and improve robustness of our alignment model, we substitute 10% of the translated term $a^{(T)}$ with a random token from the vocabulary. The ground-truth label for these randomly substituted terms is also set to *None*.

2.3 Testing Phase

During the testing phase, we initially leverage the bilingual alignment model developed during the training phase to generate code-switch augmented examples. Subsequently, we utilize the bilingual generative model to produce a group of bilingual candidate triplets. Finally, using the bilingual alignment model, we align these candidate triplets into the target language. Through candidate voting, we obtain the final output triplets. For ease of demonstration, we denote the input sentence in target language during testing phase as $s^{(tgt)}$, and the input sentence in source language as $s^{(src)}$.

2.3.1 Alignment-based Code-switching

As depicted on the upper part of Figure 2, our code-switching method for test-time augmentation contains four steps. Initially, we use an off-the-shelf translation system to convert the target language into English. Then, we construct alignment inputs for the bilingual alignment model. A heuristic method is designed to select phrases with a maximum of 3-grams from the translated sentence. The criterion for selection is that bilingual alignment model should not predict *None* as output. The top-10 longest phrases are chosen to construct the alignment inputs. There are two types of alignment inputs: $s^{(tgt)}$ <SEP> $t^{(src)}$ and $s^{(src)}$ <SEP> $t^{(tgt)}$, where $t^{(src)}$ and $t^{(tgt)}$ represent aspect term or opinion term in source language and target language respectively. Subsequently, we use the bilingual alignment model to get the augmented code-switching sentences. Finally, these sentences are used to create a set of candidate triplets with the help of the bilingual generative model.

2.3.2 Alignment on Candidate Triplets

As discussed in section 2.3.1, we generate two types of alignment inputs: $s^{(tgt)}$<SEP>$t^{(src)}$ and $s^{(src)}$<SEP>$t^{(tgt)}$. Consequently, we manage to create candidate triplets in two distinct languages. The former situation is straightforward because we aim for the triplets to be in the target language, and wev́e already computed the bilingual parallel terms. However, for the latter situation, we still need to align the remaining terms from the source language into the target language. For this alignment, we continue to employ the bilingual alignment model that we obtain during the training time to ensure that the candidate triplets conform to a single language. Finally, to get the final output triplets, we use a voting mechanism to decide which of the terms in the candidate triplets are the most likely to be correct.

3 Experiments

3.1 Datasets

We conduct experiments on the publicly available datasets from Semeval-2022 task 10: structured sentiment analysis (Barnes et al., 2022). We use the English OpeNER dataset (Agerri et al., 2013) for training, and perform cross-lingual validation

Dataset	Language	Train			Validation			Test		
		# s	# a	# o	# s	# a	# o	# s	# a	# o
OpeNER$_{\text{EN}}$	English	2494	3850	4150	2494	3850	4150	2494	3850	4150
NoReC$_{\text{Fine}}$	Norwegian	-	-	-	-	-	-	11437	8923	11115
MultiB$_{\text{EU}}$	Basque	-	-	-	-	-	-	1521	1775	2328
MultiB$_{\text{CA}}$	Catalan	-	-	-	-	-	-	1678	2336	2756
OpeNER$_{\text{ES}}$	Spanish	-	-	-	-	-	-	2057	3980	4388

Table 1: Dataset statistics. # s, # a and # o refer to the number of sentences, the number of aspect terms, and the number of opinion terms respectively. For cross-lingual evaluation, training and validation sets are not available for the non-English datasets, which we denote as '-'.

on four datasets in other languages respectively, i.e., Spanish (Agerri et al., 2013), Catalan (Barnes et al., 2018), Basque (Barnes et al., 2018) and Norwegian (Øvrelid et al., 2020) datasets. For reproducibility, we use the same train/validation/test split as the official datasets. The statistics of the datasets are listed in Table 1.

The original datasets contain four types of annotations: holders, targets, expressions and polarities. As for the task of aspect sentiment triplet extraction, we treat the target annotations as aspect terms (a), and the expression annotations as opinion terms (o). The holder annotations are not used in our experiments. We did not use the multi-lingual dataset released in Semeval-2016 task 5 (Pontiki et al., 2016) because it does not contain opinion term annotations.

3.2 Implementation Details

All our experiments are conducted on a single NVIDIA Tesla P40 GPU with 24GB of GPU memory. We set the maximum sequence length to 128 and the training batch size to 8. We use AdamW optimizer with a learning rate of 1e-4. The model is trained for 10 epochs and checkpoints with the best performance on validation set are selected for the final predictions on test set. We use Google translate API [3] as the translation model in our experiments.

3.3 Compared Methods

We use the following models in our experiments:

mT5-base (Xue et al., 2021) mT5 is a multilingual variant of T5 (Raffel et al., 2020). T5 is a large-scale pre-trained language model with encoder-decoder architecture, and is trained with the span corruption task. mT5-base is pre-trained on a new Common Crawl-based dataset (mC4) covering 101 languages.

m2m100_418M (Fan et al., 2021) m2m100 is a variant of mBART (Liu et al., 2020b). mBART is a multi-lingual sequence-to-sequence model aimed for machine translation task. Compared to mBART, m2m100 is designed to be a many-to-many multilingual translation model that can translate directly between any pair of 100 languages. It is trained with the sequence-to-sequence denoising auto-encoding task.

For performing cross-lingual ASTE task on m2m100, we need to append an additional language token to the input sentence, and set target language id to be the first generated token. As the original settings in m2m100 does not consider the bilingual code-switched context, we manually set the source language as English, and we use the target language id as the first generated token when generating triplets. We use the spanish language id as the first generated token for Basque dataset, for the reason that the m2m100 model does not contain Basque language id.

ChatGPT & GPT-4 ChatGPT and GPT-4 are large language models developed by OpenAI. ChatGPT is trained with both supervised fine-tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF). GPT-4 surpasses ChatGPT in its advanced reasoning capabilities. It can solve difficult problems with greater accuracy, resulting from its broader general knowledge and problem solving abilities.

3.4 Evaluation Metrics

We use weighted-averaged precision, recall and F1-score to evaluate the performance of our model. Our evaluation metrics are calculated in the same way as Sentiment Graph F1 (Barnes et al., 2021), with the exception that we do not utilize the graph-based structure for triplet representation. This graph-based structure requires additional alignment

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

for generative models, which we sidestep by directly forming pairs based on the number of overlapping words between the predicted triplets and the ground-truth triplets. When calculating the precision score, we identify the most similar ground-truth triplet for each predicted triplet. Conversely, when calculating the recall score, we focus on pairing each ground-truth triplet with the most similar predicted triplet. The details of calculation can be found in A.

3.5 Main Results

The main results are illustrated in Table 2. Based on the results, we have the following observations:

3.5.1 Our TT-CSW framework boosts the performance of backbone models on cross-lingual ASTE.

As shown in Table 2, the cross-lingual ASTE results of mT5-base and m2m100 are significantly improved after applying our proposed TT-CSW framework. Specifically, compare to the original cross-lingual results, the weighted-averaged F1 on Spanish, Basque, Catalan datasets are improved by 15.2%, 13.9% and 14.4% respectively when using mT5-base as the backbone model. As for m2m100 backbone, the weighted-averaged F1 on Spanish, Basque, Catalan datasets are improved by 43.8%, 27.8% and 41.8% respectively. This proves that the combination of training phase bilingual code-switching and testing phase alignment-based code-switching can significantly improve cross-lingual understanding of backbone models.

As for the results on Norwegian dataset, we can observe some abnormal phenomena: all the models perform worse than the all-null baseline (i.e., outputs an empty list for all the test samples). We notice that 47% of the test samples in Norwegian dataset do not contain any aspect or opinion terms. It is an unusual high rate of empty labels compared to the other three datasets, which are 11.7%, 21.3% and 16.1% for Spanish, Basque and Catalan datasets respectively. We suspect that the labeling standard for Norwegian dataset is different from the other three datasets, which drops irrelevant aspects and opinions during the annotation process.

3.5.2 Test time augmentation further improves performance.

As depicted in Table 2, we can observe that the performance of complete translation (CT) and code-switching (CSW) are both improved after applying our proposed test-time augmentation method. For mT5-base backbone, the average wF1 on four datasets improved by 1.8% and 1.6%, as compared to the original CT and CSW results. As for m2m100 backbone, the improvements are 2.8% and 2.6% respectively. By combining training and testing phases of code-switching, we can achieve an improvement of 3.7% and 4.0% on average wF1 for mT5-base and m2m100 respectively. The bilingual multi-view of the input sentence introduced by our proposed test-time augmentation method can reduce the ambiguity of the input sentence, therefore further enhancing model performance.

3.5.3 Our TT-CSW framework surpasses evaluation results of both ChatGPT and GPT-4.

We use the same zero-shot prompt as in Gou et al. (2023) for cross-lingual ASTE task, which briefly describes the task and the definition of the output triplet first, and then provides the format for the output. As for the few-shot prompt, we randomly select 10 samples from the english training set and use them across all four datasets. The details of the prompts are listed in Appendix B. The results are listed in Table 2. We can observe that our proposed TT-CSW framework outperforms ChatGPT on all four datasets. Specifically, the average wF1 score of our proposed TT-CSW framework with mT5-base as backbone model is 46.6%, which is 15.7% higher than ChatGPT-0 and 14.3% higher than ChatGPT-10. This evidence indicates that even though ChatGPT can perform zero-shot cross-lingual transfer on ASTE tasks, its efficiency is still significantly less than that of smaller, fine-tuned models using our proposed TT-CSW framework. When it comes to GPT-4, except for the Spanish dataset, our TT-CSW framework outperforms its 10-shot performance. We surmise this could be because GPT-4 has extensive knowledge of the Spanish language, which is not the case for the other three languages with scarce annotation resources.

3.6 Boundary Prediction Analysis

To examine the effectiveness of our proposed boundary-aware code-switching method, we conduct a further analysis on the boundary prediction results on the Spanish, Basque and Spanish datasets. In specific, we use an evaluation metric called non-polar weighted-averaged F1 (NP-wF1). This metric is similar to wF1 as defined in section A, except that we ignore the sentiment polarity part

		Spanish			Basque			Catalan			Norwegian			**AVG**
		wP	wR	wF1	wP	wR	wF1	wP	wR	wF1	wP	wR	wF1	
all-null		11.7	4.8	6.8	21.3	12.9	16.1	16.1	9.4	11.8	47.0	32.6	**38.5**	18.3
ChatGPT	-0	42.7	35.8	38.9	25.2	19.1	21.8	39.6	34.7	37.0	25.3	26.5	25.9	30.9
	-10	48.9	48.2	48.5	24.0	26.8	25.3	41.3	41.9	41.6	12.4	15.7	13.9	32.3
GPT-4	-10	61.0	50.3	**55.1**	33.4	29.2	31.2	48.8	43.3	45.9	38.3	30.7	34.1	41.6
mT5-base (Xue et al., 2021)	CL	45.5	29.4	35.7	35.1	28.1	31.2	42.6	34.5	38.1	39.3	33.6	36.2	35.3
	CT	52.9	42.9	47.4	44.1	42.6	43.3	49.8	47.7	48.7	32.7	31.2	31.9	42.8
	+tta	54.6	44.5	49.1	46.2	42.6	44.4	51.9	50.5	51.2	35.8	32.1	33.8	44.6
	CSW	54.8	44.4	49.1	42.6	44.1	43.3	50.8	49.4	50.1	44.0	32.3	37.3	44.9
	+tta	58.0	45.4	50.9	45.0	45.1	**45.1**	54.4	50.8	**52.6**	44.7	32.6	37.7	**46.6**
m2m100 (Fan et al., 2021)	CL	11.1	4.6	6.5	20.7	12.4	15.5	14.5	8.3	10.5	44.0	30.6	36.1	17.1
	CT	52.7	47.0	49.7	31.9	33.8	32.8	47.3	51.1	49.1	31.2	31.8	31.5	40.8
	+tta	55.6	47.5	51.2	39.1	36.9	38.0	49.7	54.4	51.9	34.0	32.7	33.3	43.6
	CSW	53.6	46.9	50.0	35.6	37.7	36.6	51.5	48.6	50.0	33.9	30.7	32.2	42.2
	+tta	53.8	47.2	50.3	43.5	43.2	43.3	53.4	51.3	52.3	35.7	31.0	33.2	44.8

Table 2: Main results on four datasets with different languages on cross-lingual ASTE task. wF1 scores are reported; the best results are in bold, and the second best are underlined. AVG represents the average wF1 score on all four datasets. ChatGPT-0 and ChatGPT-10 refer to zero-shot and 10-shot results of ChatGPT respectively. CL: cross lingual result; CT: complete translation, i.e., translate-train; CSW: code-switching. +tta refers to the results after combining our proposed test-time augmentation method.

		Spanish			Basque			Catalan			**AVG**
		NP-wP	NP-wR	NP-wF1	NP-wP	NP-wR	NP-wF1	NP-wP	NP-wR	NP-wF1	
mT5-base	CL	48.31	31.10	37.84	36.72	28.95	32.38	46.73	37.86	41.83	37.35
	dict_csw (static)	58.39	43.00	49.53	40.07	41.63	40.83	54.31	45.81	49.70	46.69
	dict_csw (dynamic)	59.18	44.10	50.54	40.54	37.39	38.90	55.56	43.91	49.05	46.16
	CT	56.31	45.41	50.28	46.36	44.39	**45.36**	55.43	52.57	53.96	49.87
	our CSW	56.73	46.00	50.81	43.79	45.80	44.77	55.47	53.90	**54.67**	**50.08**
m2m100	CL	12.68	5.17	7.35	20.98	12.75	15.86	15.52	9.01	11.40	11.54
	dict_csw (static)	36.86	19.12	25.18	23.05	14.46	17.77	33.13	18.67	23.88	22.28
	dict_csw (dynamic)	29.18	19.26	23.20	22.21	13.57	16.85	27.84	19.20	22.72	20.93
	CT	56.89	50.53	53.52	35.61	36.98	36.28	51.95	56.14	53.96	47.92
	our CSW	57.86	50.52	**53.94**	40.23	42.42	41.30	55.97	52.86	54.37	49.87

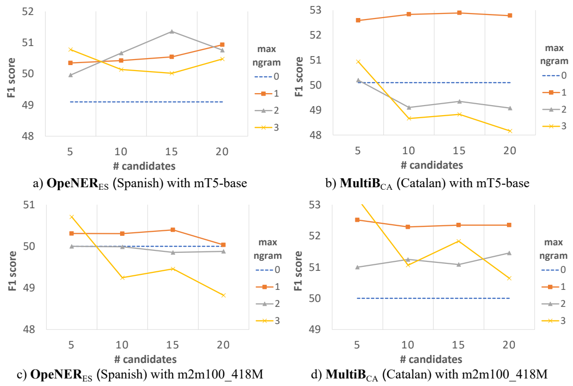
Table 3: Boundary prediction results with non-polar wF1 on three datasets with different languages on cross-lingual ASTE task. dict_csw refers to the results of dictionary-based code-switching. Static and dynamic in the parentheses refer to different strategies for loading the bilingual dictionary.

during the matching between predicted triplets and ground-truth triplets. In this way, we can focus on evaluating term boundaries. For dictionary-based code-switching, we use the bilingual dictionary released by Qin et al. (2021), which is based on MUSE (Lample et al., 2018). We use two different strategies for loading the bilingual dictionary: static and dynamic. Static refers to the strategy that we construct the code-switched samples before the training phase, and the switched words are fixed during the training phase. Dynamic refers to the strategy that we reconstruct the code-switched samples at the start of each epoch during the training phase. We keep a ratio of 0.3 for the probability of switching each word based on the bilingual dictionary.

The results are listed in Table 3. We can observe that for both mT5-base and m2m100, our proposed boundary-aware code-switching method outperforms the dictionary-based code-switching method and the complete translation method. For dictionary-based code-switching, static strategy performs better than dynamic strategy. Also, m2m100 struggles to predict terms accurately given the dictionary-based code-switched context. We suspect that this is due to the fact that bilingual dictionary contains some noisy translations, which may lead to incorrect term boundaries. Overall, the results prove the efficacy of improving term boundaries prediction with our proposed boundary-aware code-switching method.

3.6.1 Effect Analysis

We conduct an analysis on the effect of maximum n-gram and number of candidates for code-switching in test phase. The results are depicted in Figure 3. When the number of candidates is relatively small (i.e., 5), increasing maximum n-gram helps to im-



a) **OpeNER**$_{\mathrm{ES}}$ (Spanish) with mT5-base

b) **MultiB**$_{\mathrm{CA}}$ (Catalan) with mT5-base

c) **OpeNER**$_{\mathrm{ES}}$ (Spanish) with m2m100_418M

d) **MultiB**$_{\mathrm{CA}}$ (Catalan) with m2m100_418M

Figure 3: Effect of maximum n-gram and number of candidates for code-switching in test phase on Spanish and Catalan datasets. "# candidates" refers to the number of augmented input sentence.

prove performance. However, when the number of candidates is larger, the improvement is not stable and the performance even decreases. We suspect that this is because the number of candidates is already large enough to cover the possible code-switched context, and increasing maximum n-gram may introduce more noise.

4 Related Works

4.1 Aspect Sentiment Triplet Extraction

Aspect Sentiment Triplet Extraction (ASTE) task was first proposed by Peng et al. (2020) as a subtask of aspect-based sentiment analysis (ABSA), and has drawn increasing attention in recent years with various kinds of methods. Peng et al. (2020), Xu et al. (2020) and Liang et al. (2023) proposed to use tagging-based approaches to solve this task. Graph-based encoding methods are also proposed for modeling the relationships between words (Barnes et al., 2021; Chen et al., 2022). There also exists some works that tried to formalize this task as a machine reading comprehension (MRC) task (Chen et al., 2021; Liu et al., 2022; Zhai et al., 2022). With the fashion of multi-task learning, generative methods are proposed to solve not only ASTE task, but also other ABSA subtasks together in a unified framework (Yan et al., 2021; Zhang et al., 2021b; Gao et al., 2022; Gou et al., 2023).

4.2 Cross-lingual Transfer

Even though supervised methods have attained remarkable results on high-resource languages, it is still a challenge to attain comparable performance for languages with limited annotated resources. Cross-lingual transfer is one of the solution to this data scarcity issue on low-resource languages. It aims to solve the problem by leveraging the knowledge from high-resource languages (Schuster et al., 2019; Lin et al., 2019). Existing methods on cross-lingual transfer can be roughly divided into two categories: data transfer and representation transfer. Data transfer methods usually rely on pesudo-labels on target language generated from machine translation tools (Fei et al., 2020; Zhang et al., 2021a) or knowledge-distillation methods (Liu et al., 2020a; Ge et al., 2023). Representation transfer methods try to align the representations of source and target languages in a shared space and exploit the language-independent features (Nooralahzadeh et al., 2020; Huang et al., 2021, 2023). Existing works on cross-lingual ABSA mainly focus on sentiment polarity part, which utilize translation-based methods (Barnes et al., 2016; Zhang et al., 2021a) or teacher-student distillation (Lin et al., 2023). However, few attempts have been made to apply cross-lingual transfer to ASTE task.

5 Conclusion

In this study, we present a new code-switching framework for the cross-lingual aspect-based sentiment extraction (ASTE) task that can be easily incorporated with a variety of generative backbone models. It bridges the gap between the bilingual training phase and the monolingual test-time prediction. Our approach includes a boundary-aware code-switching method via the translation system, significantly improving the accurate determination of term boundaries. Additionally, we have designed a test-time augmentation alignment method that minimizes the ambiguity of the input sentence and further boosts model performance. Our proposed Test-time Code-Switching Framework (TT-CSW) has been thoroughly evaluated under four cross-lingual ASTE datasets with different languages, demonstrating its effectiveness. By integrating our method with several benchmark models, we attain an average improvement of 3.7% on weighted F1-score. We also evaluate ChatGPT and GPT-4, two commonly used Large Language Model (LLM) developed by OpenAI. Furthermore, we prove that small generative models, when combined with our proposed TT-CSW framework, can exceed the performance of ChatGPT and GPT-4 by 14.2% and 5.0% respectively.

Limitations

Despite the promising results, our proposed TT-CSW framework still has some limitations for future work. Firstly, our proposed boundary-aware code-switching method relies on the translation system, which may introduce translation errors. Secondly, our proposed test-time augmentation method may introduce additional computational cost, which requires a trade-off between performance and efficiency for real-time applications. Lastly, we only evaluate our proposed framework on cross-lingual ASTE task, further experiments are needed to expand the scope of our proposed framework to other cross-lingual tasks.

Ethics Statement

Our experiments are conducted using publicly accessible datasets, ensuring no personal information is gathered. There's no utilization of sensitive or private data in our research processes. We maintain a strict policy against the use of any data that could potentially harm an individual, group, or the environment.

References

Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural*, (51):215–218.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. 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 3387–3402.

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

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. Semeval 2022 task 10: structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295.

Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2974–2985.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12666–12674.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

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

Yukun Feng, Feng Li, and Philipp Koehn. 2022. Toward the limitation of code-switching in cross-lingual transfer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5966–5971.

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. Lego-absa: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *Proceedings of the 29th international conference on computational linguistics*, pages 7002–7012.

Ling Ge, Chunming Hu, Guanghui Ma, Hong Zhang, and Jihong Liu. 2023. Prokd: an unsupervised prototypical knowledge distillation network for zero-resource cross-lingual named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12818–12826.

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.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1684–1697.

Yucheng Huang, Wenqiang Liu, Xianli Zhang, Jun Lang, Tieliang Gong, and Chen Li. 2023. PRAM: An end-to-end prototype-based representation alignment model for zero-resource cross-lingual named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3220–3233, Toronto, Canada. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Pan Li, Ping Li, and Kai Zhang. 2023. Dual-channel span for aspect sentiment triplet extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 248–261.

Zhi Li, Xing Gao, Ji Zhang, and Yin Zhang. 2022. Multi-label masked language modeling on zero-shot code-switched sentiment analysis. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2663–2668.

Shuo Liang, Wei Wei, Xian-Ling Mao, Yuanyuan Fu, Rui Fang, and Dangyang Chen. 2023. Stage: span tagging and greedy inference scheme for aspect sentiment triplet extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13174–13182.

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*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, volume 57.

Junhao Liu, Linjun Shou, Jian Pei, Ming Gong, Min Yang, and Daxin Jiang. 2020a. Cross-lingual machine reading comprehension with language branch knowledge distillation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2710–2721.

Shu Liu, Kaiwen Li, and Zuhe Li. 2022. A robustly optimized bmrc for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 272–278.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2536–2545.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8600–8607.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2021. Cosda-ml: multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3853–3860.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

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.

Zepeng Zhai, Hao Chen, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Com-mrc: A context-masked machine reading comprehension framework for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3230–3241.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021a. 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.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. 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.

Yice Zhang, Yifan Yang, Meng Li, Bin Liang, Shiwei Chen, and Ruifeng Xu. 2023. Target-to-source augmentation for aspect sentiment triplet extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12165–12177.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023. Enhancing code-switching for cross-lingual slu: A unified view of semantic and grammatical coherence. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7849–7856.

A Calculation of Evaluation Metrics

We use the same type of symbols as defined in section 2.1. The similarity score between a pair of triplets T_1 and T_2 can be calculated as shown in Equation 1, where $overlap(a, b)$ represents the number of overlapping words between a pair of string a and b. If the sentiment polarity part is not correctly predicted, we consider it as an incorrect prediction and the similarity score is set to 0. We omit the respective terms in Equation 1 when OT_1 or AT_1 is left empty. For the special case when terms in both triplets are empty, we regard it as an exact match. The weighted-averaged precision and recall score are calculated as shown from Equation 2 to 4 respectively. The RT_i and PT_i denote the ground truth triplet list and the predicted triplet list for sample D_i.

$$\begin{aligned} sim(T_1, T_2) = & \frac{overlap(OT_1, OT_2)}{2len(OT_1)} \\ & + \frac{overlap(AT_1, AT_2)}{2len(AT_1)} \end{aligned} \tag{1}$$

$$wP = \frac{\sum_{i=1}^{N} \sum_{T_j \in PT_i} \max_{T_k \in RT_i}(sim(T_j, T_k))}{TP + FP} \tag{2}$$

$$wR = \frac{\sum_{i=1}^{N} \sum_{T_j \in RT_i} \max_{T_k \in PT_i}(sim(T_j, T_k))}{TP + FN} \tag{3}$$

$$wF1 = \frac{2wP \cdot wR}{wP + wR} \tag{4}$$

B Prompts for ChatGPT and GPT-4

As shown in Listing 1 and 2, we list the zero-shot and few-shot prompts for ChatGPT and GPT-4 in our experiments on the cross-lingual ASTE task. We use the same few-shot prompt across all the four datasets with different languages.

Listing 1: Zero-shot Prompt for cross-lingual ASTE task.

```
According to the following sentiment elements definition:

– The ’aspect term’ refers to a specific feature, attribute, or aspect of a product or service that a user may express an opinion about.
– The ’opinion term’ refers to the sentiment or attitude expressed by a user towards a particular aspect or feature of a product or service.
– The ’sentiment polarity’ refers to the degree of positivity, negativity or neutrality expressed in the opinion towards a particular aspect or feature of a product or service, and the available polarities inlcudes: ’positive’, ’negative’ and ’ neutral’.

Recognize all sentiment elements with their corresponding aspect terms, opinion terms and sentiment polarity in the following text with the format of [(’aspect term’, ’opinion term’, ’sentiment polarity’), ...]:
```

Listing 2: Few-shot Prompt (10 shots) for cross-lingual ASTE task.

```
According to the following sentiment elements definition:

– The ’aspect term’ refers to a specific feature, attribute, or aspect of a product or service that a user may express an opinion about.
– The ’opinion term’ refers to the sentiment or attitude expressed by a user towards a particular aspect or feature of a product or service.
– The ’sentiment polarity’ refers to the degree of positivity, negativity or neutrality expressed in the opinion towards a particular aspect or feature of a product or service, and the available polarities inlcudes: ’positive’, ’negative’ and ’ neutral’.

Recognize all sentiment elements with their corresponding aspect terms, opinion terms and sentiment polarity in the following text with the format of [(’aspect term’, ’opinion term’, ’sentiment polarity’), ...]:
Text: Although I wouldn ’t say this was a cheap holiday , it didn ’t break the bank either , so if you want a guaranteed tan go to Egypt , if not and its your main summer / yearly holiday , I wouldn ’t recommended it ... I , d go somewhere else to avoid disappointment .
[(’it’, "wouldn ’t recommended", ’negative’), (’somewhere else’, ’d go’, ’negative’)]

Text: The Frankfurter Hof is surrounded by some of Europe ’s most impressive skyscrapers .
[]

Text: Near hotel there are many bars , pubs , clubs .
[(’clubs’, ’Near hotel there are’, ’positive’), (’many bars’, ’Near hotel there are’, ’positive’), (’pubs’, ’Near hotel there are’, ’ positive’)]

Text: Great central location
[(’location’, ’central’, ’positive’), (’location’, ’Great’, ’positive’)]

Text: Ofitsyanty while working well .
[(’’, ’working well’, ’positive’)]

Text: Never worth the money we had to pay for the room and the stay !!!
[(’the money we had to pay’, ’Never worth’, ’negative’)]

Text: You can pay exrra for these by paying for a ’ gold all inclusive ’ package
[]

Text: Very good parking possibilties .
[(’parking possibilties’, ’Very good’, ’positive’)]

Text: Nevertheless , because of its good location aside the liverpool One Shopping centre , with a lot of bars and restaurant , I continue going there when travelling to Liverpool .
[(’location’, ’good’, ’positive’), (’location’, ’continue going there’, ’positive’)]

Text: ( we had earplugs and used them !
[]
```