

Advancing Cross-Lingual Aspect-Based Sentiment Analysis with LLMs and Constrained Decoding for Sequence-to-Sequence Models

Jakub Šmíd^{1,2}^a, Pavel Příbání¹^b and Pavel Král²^c

¹Department of Computer Science and Engineering, University of West Bohemia in Pilsen, Univerzitní, Pilsen, Czech Republic

²NTIS - New Technologies for the Information Society, University of West Bohemia in Pilsen, Univerzitní, Pilsen, Czech Republic
{jaksmid, pribanp, pkrál}@kiv.zcu.cz

Keywords: Cross-Lingual Aspect-Based Sentiment Analysis, Aspect-Based Sentiment Analysis, Large Language Models, Transformers, Constrained Decoding.

Abstract: Aspect-based sentiment analysis (ABSA) has made significant strides, yet challenges remain for low-resource languages due to the predominant focus on English. Current cross-lingual ABSA studies often centre on simpler tasks and rely heavily on external translation tools. In this paper, we present a novel sequence-to-sequence method for compound ABSA tasks that eliminates the need for such tools. Our approach, which uses constrained decoding, improves cross-lingual ABSA performance by up to 10%. This method broadens the scope of cross-lingual ABSA, enabling it to handle more complex tasks and providing a practical, efficient alternative to translation-dependent techniques. Furthermore, we compare our approach with large language models (LLMs) and show that while fine-tuned multilingual LLMs can achieve comparable results, English-centric LLMs struggle with these tasks.

1 INTRODUCTION

Sentiment analysis aims to understand and quantify opinions expressed in text, playing a critical role in applications like customer feedback analysis, social media monitoring, and market research. Within this field, aspect-based sentiment analysis (ABSA) focuses on extracting fine-grained sentiment elements from text (Zhang et al., 2022). These elements include aspect term (a), aspect category (c), and sentiment polarity (p). For example, in the review “Tasty soup”, these elements are “soup”, “food quality”, and “positive”, respectively. Implicitly referenced aspect terms, as in “Delicious”, are frequently labelled as “NULL”.


Initially, ABSA research focused on individual sentiment elements, e.g. aspect term extraction and aspect category detection (Pontiki et al., 2014). Recent studies have shifted towards compound tasks involving multiple elements, such as end-to-end ABSA (E2E-ABSA), aspect category term extraction (ACTE), and target aspect category detection


(TASD) (Wan et al., 2020). Table 1 shows the output formats of these ABSA tasks.


Table 1: Output format for selected ABSA tasks for an input review: “Tasty soup but pricey tea”.

Task	Output	Example output
E2E-ABSA	$\{(a, p)\}$	$\{(\text{“soup”, POS}), (\text{“tea”, NEG})\}$
ACTE	$\{(a, c)\}$	$\{(\text{“soup”, food}), (\text{“tea”, drinks})\}$
TASD	$\{(a, c, p)\}$	$\{(\text{“soup”, food, POS}), (\text{“tea”, drinks, NEG})\}$

While ABSA research traditionally focuses on English, real-world applications demand multilingual capabilities. However, annotating multilingual data is costly and time-intensive. Although multilingual pre-trained models have become standard for cross-lingual transfer in natural language processing (NLP) tasks (Hu et al., 2020), applying them to cross-lingual ABSA presents challenges due to language-specific knowledge requirements. These models are usually fine-tuned on source language data and directly applied to target language data. However, they might lack language-specific knowledge for ABSA tasks involving user-generated texts with abbreviations, slang, and language-dependent aspects. A possible solution is using translated target language data with projected labels, but its effectiveness depends on

^a <https://orcid.org/0000-0002-4492-5481>

^b <https://orcid.org/0000-0002-8744-8726>

^c <https://orcid.org/0000-0002-3096-675X>

the quality of the translation and alignment.

Modern monolingual ABSA approaches use pre-trained sequence-to-sequence models, framing compound tasks as text generation problems. In contrast, cross-lingual ABSA research remains limited, focusing mainly on simple tasks and E2E-ABSA, with no studies employing the sequence-to-sequence methods that excel in monolingual ABSA.

Recent advancements in large language models (LLMs), such as GPT-4o (OpenAI, 2024) and LLaMA 3(AI@Meta, 2024), have achieved remarkable results across NLP tasks. However, fine-tuned models outperform LLMs on compound ABSA tasks (Zhang et al., 2024). Fine-tuned LLaMA-based models lead in English ABSA (Šmíd et al., 2024), but their cross-lingual performance remains unexplored.

The main motivation of this paper is the limited research on compound cross-lingual ABSA tasks, the absence of sequence-to-sequence approaches widely used in monolingual ABSA, and the reliance on external translation tools in related work, which adds complexity and potential error to the process. To address these shortcomings in existing works, we introduce a novel sequence-to-sequence method that achieves favourable results for compound ABSA tasks in cross-lingual settings without relying on external translation tools. Additionally, we explore the capabilities of several LLMs for cross-lingual ABSA, as their performance on this specific task has not been thoroughly investigated.

Our main contributions are as follows: 1) We introduce the first sequence-to-sequence approach for compound cross-lingual ABSA tasks, which does not rely on external translation tools. 2) We significantly improve zero-shot cross-lingual ABSA performance using constrained decoding. 3) We compare our method to LLMs, specifically GPT-4o mini and fine-tuned LLaMA 3 and LLaMA 3.1, showing that only fine-tuned multilingual LLaMA 3.1 achieves comparable results to our approach. 4) We conduct experiments on benchmark datasets in five languages and three compound ABSA tasks, achieving new state-of-the-art results in both cross-lingual and monolingual settings. To the best of our knowledge, we are the first to examine two compound cross-lingual ABSA tasks and the cross-lingual capabilities of LLMs for ABSA.

2 RELATED WORK

Cross-lingual ABSA research focuses on three main tasks: aspect term extraction (Klinger and Cimiano, 2015; Wang and Pan, 2018; Jebbara and Cimiano, 2019), aspect sentiment classification (Lambert,

2015; Barnes et al., 2016; Akhtar et al., 2018), and E2E-ABSA (Li et al., 2020; Zhang et al., 2021b; Lin et al., 2023; Lin et al., 2024). Of these tasks, only E2E-ABSA is compound task, i.e. it focuses on extracting more than one sentiment element simultaneously. Early methods relied on translation and word alignment tools like FastAlign (Dyer et al., 2013), with quality improvements through instance selection (Klinger and Cimiano, 2015) or constrained translation (Lambert, 2015). Others used cross-lingual embeddings trained on bilingual corpora for language-independent ABSA (Lambert, 2015; Barnes et al., 2016; Akhtar et al., 2018; Wang and Pan, 2018; Jebbara and Cimiano, 2019). Recent work focus on multilingual Transformer-based (Vaswani et al., 2017) encoder-only models, such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) combined with machine translation, with additional enhancements from parameter warm-up (Li et al., 2020), distillation on unlabelled target language data (Zhang et al., 2021b), contrastive learning (Lin et al., 2023), and dynamic loss weighting (Lin et al., 2024).

The latest monolingual ABSA research focuses primarily on text generation, exploring converting labels to natural language (Zhang et al., 2021c; Zhang et al., 2021a), multi-tasking (Gao et al., 2022), generating tuples as paths of a tree (Mao et al., 2022), element ordering (Hu et al., 2022b; Gou et al., 2023), and tagging-assisted generation (Xianlong et al., 2023).

Research shows that fine-tuned models outperform non-fine-tuned LLMs in compound ABSA tasks (Gou et al., 2023; Zhang et al., 2024), whereas fine-tuned LLaMA-based models achieve state-of-the-art results in English ABSA (Šmíd et al., 2024).

3 METHODOLOGY

This section presents our approach to addressing the triplet task (TASD), which can be easily modified for tuple tasks with minor adjustments. Figure 1 depicts the proposed approach.

3.1 Problem Definition

Given an input sentence, the aim is to predict all sentiment tuples $T = (a, c, p)$, each composed of an aspect term (a), aspect category (c), and sentiment polarity (p). Following prior works (Zhang et al., 2021a; Gou et al., 2023), we convert elements (a, c, p) into natural language (e_a, e_c, e_p) . For instance, we translate the “neutral” sentiment polarity into “ok” and the “NULL” aspect term into “it”, as shown in Figure 1.

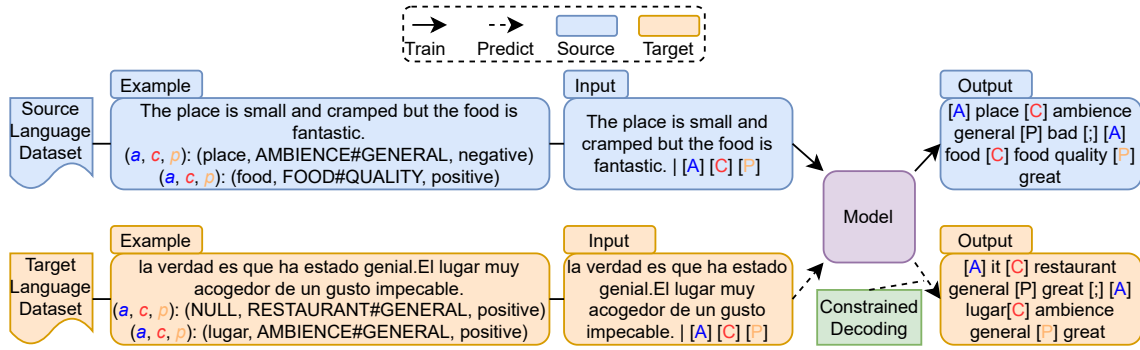


Figure 1: Overview of the proposed framework, which includes converting input labels to natural language phrases, fine-tuning on source language data, and making predictions on target language data using constrained decoding for enhancement.

3.2 Input and Output Building

To build our model’s inputs and outputs, we employ element markers to represent each sentiment element: [A] for e_a , [C] for e_c , and [P] for e_p . These markers prefix each element, forming the target sequence together. We also append these markers to the input sequence to guide the model for correct output. We follow the priority order $e_a > e_c > e_p$ recommended in prior research (Gou et al., 2023). For example, we create the following input-output pair:

Input (x): They offer a tasty soup | [A] [C] [P]

Output (y): [A] soup [C] food quality [P] great

For sentences with multiple sentiment tuples, we use the symbol [:] to concatenate their target schemes into the final target sequence. Different examples are depicted in Figure 1.

3.3 Constrained Decoding

To prevent the fine-tuned model from generating aspect terms in the source language instead of the target language, we have developed scheme-guided constrained decoding (CD) (Cao et al., 2021), which ensures that generated elements match their respective vocabulary sets by incorporating target schema information. This approach is beneficial in few-shot monolingual settings (Gou et al., 2023).

Constrained decoding dynamically adjusts candidate token lists based on the current state, enhancing control and accuracy in the generation process. For example, if the current token is ‘[’, the next token should be chosen from special terms: A, C, P, and ;. Additionally, it tracks previously generated output and current terms, guiding the decoding of subsequent tokens based on Table 2. Appendix 6 shows the algorithm in more detail.

Table 2: Candidate lists of tokens. <eos> indicates the end of a sequence, and “...” denotes arbitrary text.

Generated output	Candidate tokens
...	[
... [A / [C / [P / [;]
... [A]	Input sentence, “it”
... [C]	All categories
... [P]	great, ok, bad
... [A] ...	Input sentence, “it”, [
... [C] ...	All categories, [
... [P] ...	great, ok, bad, <eos> [
... [A] ... [C
... [C] ... [P
... [P] ... [;
... [:]	[
... [:] [A

3.4 Training

We fine-tune a pre-trained sequence-to-sequence model with provided input-output pairs. Sequence-to-sequence models consist of two components: the encoder, which transforms input sequence x into a contextualized sequence \mathbf{e} , and the decoder, which models the conditional probability distribution $P_{\Theta}(y|\mathbf{e})$ of the target sequence y based on the encoded input \mathbf{e} , where Θ represents the model’s parameters. At each decoding step i , the decoder generates the output y_i using previous outputs y_0, \dots, y_{i-1} and the encoded input \mathbf{e} . During fine-tuning, we update all model parameters and minimize the log-likelihood as

$$\mathcal{L} = - \sum_{i=1}^n \log p_{\Theta}(y_i | \mathbf{e}, y_{<i}), \quad (1)$$

where n is the length of the target sequence y .

Table 3: Dataset statistics for each language. POS, NEG and NEU denote the number of positive, negative and neutral examples, respectively.

	En	Es	Fr	Nl	Ru	Tr	
Train	Sentences	1,800	1,863	1,559	1,549	3,289	1,108
	Triplets	2,266	2,455	2,276	1,676	3,697	1,386
	Categories	12	12	12	12	12	12
	POS/NEG/NEU	1,503/672/91	1,736/607/112	1,045/1,092/139	969/584/124	2,805/641/250	746/521/119
	NULL aspects	569	700	694	513	821	135
Dev	Sentences	200	207	174	173	366	124
	Triplets	241	265	254	184	392	149
	Categories	11	11	12	11	12	10
	POS/NEG/NEU	154/77/10	189/67/8	115/120/15	94/62/28	298/68/26	74/65/10
	NULL aspects	58	83	66	64	109	15
Test	Sentences	676	881	694	575	1,209	144
	Triplets	859	1,072	954	613	1,300	159
	Categories	12	12	13	13	12	11
	POS/NEG/NEU	611/204/44	750/274/48	441/434/79	369/211/33	870/321/103	104/49/6
	NULL aspects	209	341	236	219	325	0

4 EXPERIMENTS

We report results with a 95% confidence interval from 5 runs with different seeds. The primary evaluation metric is the micro F1-score, the standard metric in ABSA research. We consider a predicted sentiment tuple correct only if all its elements exactly match the gold tuple.

4.1 Tasks and Dataset

We evaluate our method on the E2E-ABSA, ACTE, and TASD tasks (see Table 1 for task details).

We perform experiments on the standard SemEval-2016 dataset (Pontiki et al., 2016) with restaurant reviews in English (en), Spanish (es), French (fr), Dutch (nl), Russian (ru), and Turkish (tr), with provided training and test sets. We split the training data into a 9:1 ratio to create a validation set. We consider English as the source language and other languages as the target ones. Table 3 shows the data statistics for each language.

4.2 Prompts for LLMs

Figure 2 shows the prompt for LLMs for the TASD task, including one example for few-shot settings for Spanish. This prompt can be adapted for various tasks by excluding the unnecessary sentiment element for the specific task, such as the sentiment polarity for the ACTE task. For few-shot prompts, examples are drawn from the first 10 examples of the training dataset in the respective language.

4.3 Compared Methods

We compare our method with constrained decoding against models without it. For E2E-ABSA—the only task with available related work—we evaluate against approaches using decoder-only models. One such method (Li et al., 2020) operates in a true zero-shot setting, training only on source language data without machine translation. Other methods integrate machine translation with additional enhancements, including alignment-free projection and aspect code-switching for interchanging aspect terms between languages with distillation on unlabelled target language data (Zhang et al., 2021b), contrastive learning (Lin et al., 2023), and dynamically weighted loss combined with anti-decoupling to improve semantic information utilization and address class imbalances (Lin et al., 2024).

It is important to note that these methods use slightly different task definitions and datasets. Specifically, previous work employs decoder-only models, excludes implicit aspect terms (“NULL”), and restricts each aspect term to a single sentiment polarity. For example, prior research (Zhang et al., 2021b) reports 612 tuples in the English test set after filtering “NULL” aspect terms and merging sentiment polarities for each aspect term, whereas we report 859 tuples. In contrast, our approach predicts “NULL” aspect terms and allows multiple sentiment polarities per aspect term, making the task inherently more challenging.

4.4 Experimental Details

We use the large mT5 (Xue et al., 2021), selected based on related work for English (Zhang et al., 2021c; Zhang et al., 2021a; Gao et al., 2022; Gou

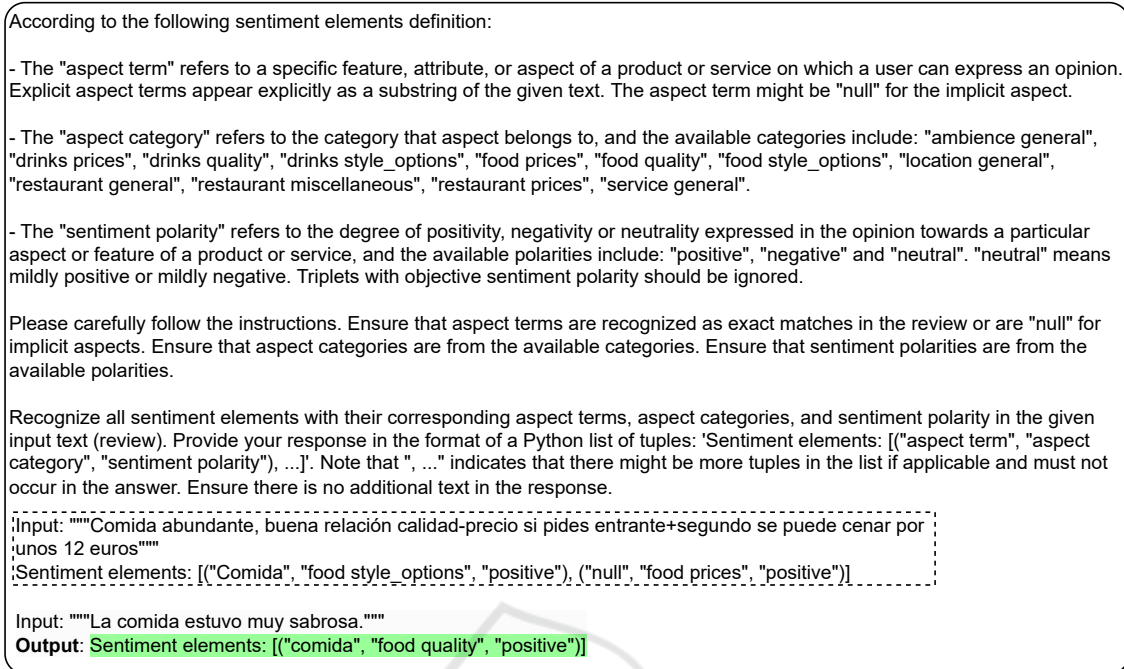


Figure 2: Prompt for the TASD task with example input, expected output in a green box, and one demonstration in Spanish enclosed in a dashed box. The demonstrations are used solely in few-shot scenarios.

et al., 2023) employing monolingual T5 (Raffel et al., 2020), and the large mBART (Tang et al., 2020) to evaluate our method across different architectures. Both models are from the HuggingFace Transformers library¹ (Wolf et al., 2020). We fine-tune the models for all experiments over 20 epochs with a batch size of 16 and employ greedy search for decoding. For mT5, we use a learning rate of 1e-4 and the Adafactor optimizer (Shazeer and Stern, 2018). For mBART, we use a learning rate of 1e-5 and the AdamW optimizer (Loshchilov and Hutter, 2017). These settings were chosen based on consistent performance on validation data across all languages and tasks.

We evaluate GPT-4o mini (OpenAI, 2024) using zero- and few-shot prompts. Additionally, we fine-tune the 8B versions of LLaMA 3 (AI@Meta, 2024) and LLaMA 3.1 (Dubey et al., 2024), employing QLoRA (Dettmers et al., 2023) with 4-bit NormalFloat quantization, a batch size of 16, a constant learning rate of 2e-4, AdamW optimizer, LoRA adapters (Hu et al., 2022a) on all linear Transformer block layers, and LoRA $r = 64$ and $\alpha = 16$. Utilizing the zero-shot prompt shown in Figure 2, i.e. without the demonstrations, we fine-tune the model for up to 5 epochs, selecting the best-performing model based on validation loss. All experiments are conducted using an NVIDIA A40 with 48 GB GPU.

¹<https://github.com/huggingface/transformers>

5 RESULTS

Table 4 presents the results. Some key observations include the following:

1) Constrained decoding significantly improves cross-lingual ABSA by up to 10% over baseline models, effectively mitigating the issue where the model predicts aspect terms in English instead of the target language (e.g. “*place*” instead of “*lugar*”). The improvement is most noticeable in Spanish and Russian. Constrained decoding is unnecessary in monolingual experiments as the problem does not occur.

2) MT5 generally outperforms mBART and benefits more from constrained decoding.

3) GPT-4o mini performs relatively well for the E2E-ABSA task, which excludes aspect categories. However, its performance is notably lower on the ACTE and TASD tasks, indicating that identifying aspect categories is challenging for the model. Incorporating few-shot prompts boosts performance by 10–20% in most cases.

4) Our approach consistently outperforms GPT-4o mini across all tasks and languages. Notably, we achieve approximately a 20% improvement on the TASD task across all languages in zero-shot cross-lingual settings. Even when GPT-4o mini is enhanced with 10-shot prompts, our results remain significantly superior. The biggest difference is generally for Rus-

Table 4: F1 scores for zero-shot cross-lingual ABSA with English as the source language and other languages as target languages compared to monolingual results and GPT-4o mini. The compared works have different models and E2E-ABSA definitions. **Bold** results indicate significant improvements using constrained decoding (CD). Underlined results are the best absolute results for each language and task in both monolingual and cross-lingual settings.

		E2E-ABSA				ACTE				TASD							
		Es	Fr	Nl	Ru	Tr	Es	Fr	Nl	Ru	Tr	Es	Fr	Nl	Ru	Tr	
Monolingual	mT5	w/o CD	74.4±0.6	69.9±0.5	71.6±1.0	72.4±0.2	60.1±1.7	70.4±0.7	63.7±0.8	68.8±0.5	73.2±0.5	59.1±0.5	65.8±0.4	59.0±0.6	62.9±1.4	67.0±0.9	54.1±3.0
		w/ CD	75.3±0.6	69.8±1.4	67.0±0.4	72.2±0.4	<u>60.7±1.1</u>	69.9±0.4	64.9±0.5	62.9±0.5	72.8±1.0	60.4±2.1	66.2±0.5	58.9±1.1	57.6±0.5	66.4±0.4	53.9±1.5
	mBART	w/o CD	73.0±0.5	66.4±1.1	68.9±1.2	68.7±1.6	56.0±2.7	66.4±1.6	61.1±1.6	64.1±1.2	70.9±0.6	56.8±2.2	62.9±1.2	54.8±0.9	57.6±0.9	62.6±0.7	49.3±3.1
		w/ CD	71.9±1.3	64.0±1.7	61.6±1.0	66.2±1.1	54.4±2.3	66.8±1.5	58.2±1.2	58.0±1.2	67.4±0.3	55.3±1.5	61.5±1.4	52.4±0.6	52.1±1.0	60.1±1.9	47.6±2.7
	GPT-4o mini	0-shot	55.7	52.5	45.8	48.3	47.9	34.8	31.7	27.9	31.7	27.4	32.4	31.8	29.1	31.0	32.2
		10-shot	63.8	55.2	58.9	58.9	51.0	56.2	46.1	49.5	45.8	47.5	50.9	40.9	46.0	42.1	44.3
	LLaMA 3	70.0±2.0	63.1±1.8	66.0±1.2	60.7±2.3	48.6±2.0	59.8±2.9	54.9±1.2	58.1±4.3	62.1±1.7	46.3±2.9	57.2±1.7	48.2±2.4	55.4±2.8	53.7±2.9	39.1±3.2	
	LLaMA 3.1	<u>77.4±0.5</u>	<u>71.2±0.5</u>	<u>74.0±0.5</u>	<u>73.7±0.3</u>	59.2±1.6	<u>70.9±0.6</u>	<u>66.8±2.3</u>	<u>70.6±0.8</u>	<u>75.3±0.9</u>	<u>60.7±0.5</u>	65.8±1.2	<u>62.0±0.6</u>	<u>65.7±1.1</u>	<u>68.4±1.0</u>	<u>58.6±1.0</u>	
Cross-lingual	mT5	w/o CD	59.2±0.5	57.8±1.2	57.1±0.9	56.4±2.1	44.4±1.4	52.5±1.0	55.8±0.7	52.3±1.3	55.0±2.7	41.4±1.4	48.3±0.5	50.4±1.4	47.7±1.1	48.6±2.0	39.1±3.6
		w/ CD	69.3±1.0	61.1±1.2	60.8±0.3	63.7±1.3	48.9±1.4	62.8±1.4	57.5±0.3	54.1±0.2	60.4±0.9	49.0±0.9	57.6±0.6	50.4±0.8	50.4±1.3	54.9±2.0	43.8±0.8
	mBART	w/o CD	61.1±2.6	49.4±3.8	51.6±2.7	57.1±1.4	31.6±3.9	52.5±1.4	49.3±1.5	44.5±1.4	53.8±1.5	31.1±2.1	47.6±1.9	39.6±0.8	39.1±0.9	48.5±1.1	23.5±2.6
		w/ CD	61.7±2.7	49.2±4.1	50.1±3.5	57.8±1.8	30.3±3.0	54.8±0.4	49.2±0.6	46.9±0.9	55.9±0.2	34.7±1.1	51.1±1.2	39.9±0.6	38.9±0.9	50.5±0.7	27.3±1.1
	LLaMA 3	47.2±5.0	41.7±3.0	43.3±3.5	53.1±2.2	26.6±9.7	44.8±8.5	38.1±3.5	40.0±6.0	50.6±4.1	29.9±7.1	39.5±7.5	31.5±3.2	29.8±5.7	46.5±2.6	22.8±7.2	
	LLaMA 3.1	<u>73.4±0.7</u>	<u>68.1±0.5</u>	<u>64.2±0.8</u>	58.8±1.0	48.1±4.1	<u>64.9±1.1</u>	<u>60.9±3.0</u>	<u>55.7±1.5</u>	59.2±1.6	47.7±2.1	<u>61.3±0.6</u>	<u>57.3±1.2</u>	<u>54.8±0.8</u>	53.1±1.2	41.4±0.7	
	(Li et al., 2020)	67.1	56.4	59.0	56.8	46.2	-	-	-	-	-	-	-	-	-	-	
	(Zhang et al., 2021b)	69.2	61.0	63.7	62.0	-	-	-	-	-	-	-	-	-	-	-	
	(Lin et al., 2023)	61.6	49.5	51.0	50.8	-	-	-	-	-	-	-	-	-	-	-	
	(Lin et al., 2024)	69.6	60.7	61.3	62.3	-	-	-	-	-	-	-	-	-	-	-	

sian and French.

5) Our method consistently exceeds the performance of fine-tuned LLaMA 3 across all tasks and language combinations, likely due to LLaMA 3 being primarily pre-trained for English.

6) LLaMA 3.1 extends language support to include Spanish and French but lacks compatibility with Russian, Dutch and Turkish. It achieves the best results in most tasks and languages in monolingual settings, although mT5 performs similarly when accounting for confidence intervals. In cross-lingual scenarios, mT5 with constrained decoding usually outperforms LLaMA 3.1 for Russian and Turkish and demonstrates similar or slightly inferior performance for other languages. While LLaMA 3.1 does not officially support Dutch, it likely benefits from linguistic similarities with its supported languages.

7) Results for Turkish are consistently worse than for other languages, which might be because it is the only language not from the Indo-European family.

As mentioned in Section 4, comparing our E2E-ABSA results with prior research is challenging due to methodological differences. Previous works exclude implicit aspect terms, limit one sentiment polarity per aspect term, and use decoder-only models. Moreover, prior approaches rely on external translation tools and can be affected by the translation quality, whereas our method avoids external tools entirely. In contrast, our approach predicts implicit aspect terms, allows multiple sentiment polarities per aspect term, and avoids external tools entirely. Despite these challenges, our method with constrained decoding achieves comparable results and proves. We find constrained decoding to be more practical than relying on external translation tools.

No prior research exists for ABSA tasks beyond E2E-ABSA in cross-lingual settings, which serves as one of the primary motivations for this paper.

5.1 Inference and Training Speed

Table 5 summarizes the average training time per epoch and inference time per example (both absolute in seconds and relative) for various models on the TASD task, with English as the source language and Spanish as the target language.

Table 5: Average absolute and relative training time per epoch and inference time per example for different models on the TASD task, with English as the source language and Spanish as the target language.

Model	Training Time Per Epoch		Inference Time Per Example	
	Absolute [s]	Relative	Absolute [s]	Relative
mT5	210	1.00	0.24	1.00
LLaMA 3.1	924	4.40	1.04	4.33

The mT5 model is the reference, with a relative training time of 1.00. The LLaMA 3.1 model is significantly slower, requiring 4.40 times the training time of mT5 and a much higher inference time, 15.32 times that of mT5. This comparison indicates that while larger models like LLaMA 3.1 may offer performance gains, they come with substantial computational costs during both training and inference.

5.2 Recommendations

In summary, constrained decoding improves cross-lingual results significantly. Overall, we recommend using the mT5 model with constrained decoding in most scenarios, while LLaMA 3.1 is prefer-

able for languages it specifically supports. The mT5 model consistently outperforms the English-specific LLaMA 3 across all tasks and languages. Moreover, mT5 delivers excellent results compared to the multilingual LLaMA 3.1 despite having approximately seven times fewer parameters while also requiring less training time per epoch and offering faster inference. Additionally, fine-tuning LLaMA models on consumer GPUs demands specialized techniques, which can be a limiting factor. However, LLaMA 3.1 may be the preferred choice if hardware resources are sufficient and training time and inference are not major concerns.

5.3 Error Analysis

We perform an error analysis to understand the challenges of sentiment prediction better. Specifically, we manually examine the predictions for the first 100 test samples of the Spanish dataset in the TASD task, using the best-performing runs of LLaMA 3.1 and mT5 with and without constrained decoding. Figure 3 depicts the results of the error analysis.

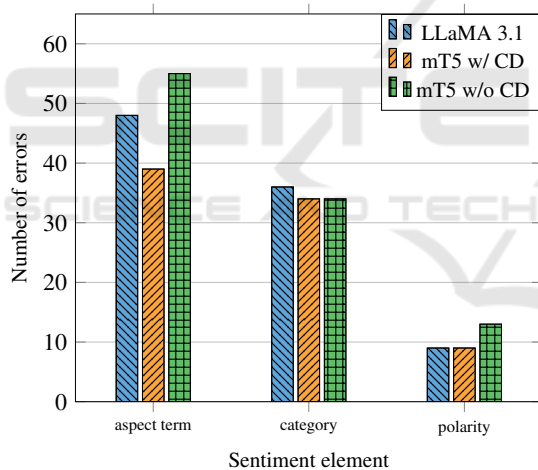


Figure 3: Number of error types for LLaMA 3.1 and mT5 with and without constrained decoding (CD) on the Spanish target language and the TASD task.

The primary source of errors lies in aspect term prediction, where the model often misses some aspect terms, generates additional ones, or produces incomplete terms instead of full ones. Without constrained decoding, the mT5 model may also generate text that does not appear in the original review, alter its format, or use the source language instead of the target language (e.g. “service” instead of “servicio”). Constrained decoding addresses these issues effectively, reducing the number of errors.

Further challenges arise from less frequent aspect categories and inconsistent annotations, partic-

ularly for categories like “restaurant general” and “restaurant miscellaneous”, which impact performance. Another common error is the confusion between “restaurant prices” and “food prices” categories. Additionally, some categories appear in only one or a few languages, such as “food general”, found exclusively in the Dutch test set, limiting the classifier’s ability to learn from other source languages.

Sentiment polarity prediction is generally less challenging than other sentiment elements, with errors primarily occurring in misclassifying the “neutral” polarity.

6 CONCLUSION

This paper addresses three compound cross-lingual ABSA tasks using sequence-to-sequence models and constrained decoding without needing external translation tools. Through extensive experiments in six languages, we emphasize the effectiveness of constrained decoding in enhancing zero-shot cross-lingual ABSA. The proposed approach offers a practical alternative to external translation tools, demonstrating robustness and effectiveness across various language pairs and models, and opens up new possibilities for advanced cross-lingual ABSA. Additionally, we compare our method to modern LLMs, finding that older multilingual models outperform fine-tuned English-centric and closed-source LLMs. We show that fine-tuning multilingual LLMs boosts performance significantly, surpassing smaller models in supported languages.

Future research could explore multi-task learning, where a single model is trained simultaneously on multiple tasks, enabling a unified approach to handling diverse ABSA challenges. Additional experiments could examine various source-target language pair combinations to assess cross-lingual adaptability further. Finally, investigating tasks involving a fourth sentiment element (opinion terms) would be valuable, though current efforts are constrained by the limited availability of annotated data, particularly in languages other than English.

ACKNOWLEDGEMENTS

This work was created with the partial support of the project R&D of Technologies for Advanced Digitalization in the Pilsen Metropolitan Area (DigiTech) No. CZ.02.01.01/00/23_021/0008436 and by the Grant No. SGS-2022-016 Advanced methods

of data processing and analysis. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

REFERENCES

- AI@Meta (2024). Llama 3 model card.
- Akhtar, M. S., Sawant, P., Sen, S., Ekbal, A., and Bhat-tacharyya, P. (2018). Solving data sparsity for aspect based sentiment analysis using cross-linguality and multi-linguality. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 572–582, New Orleans, Louisiana. Association for Computational Linguistics.
- Barnes, J., Lambert, P., and Badia, T. (2016). Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification. In Matsumoto, Y. and Prasad, R., editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1613–1623, Osaka, Japan. The COLING 2016 Organizing Committee.
- Cao, N. D., Izacard, G., Riedel, S., and Petroni, F. (2021). Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dubey, A., Jauhri, A., Pandey, A., et al. (2024). The llama 3 herd of models.
- Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In Vanderwende, L., Daumé III, H., and Kirchoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.
- Gao, T., Fang, J., Liu, H., Liu, Z., Liu, C., Liu, P., Bao, Y., and Yan, W. (2022). LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7002–7012, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Gou, Z., Guo, Q., and Yang, Y. (2023). MvP: Multi-view prompting improves aspect sentiment tuple prediction. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *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.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022a). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- Hu, M., Wu, Y., Gao, H., Bai, Y., and Zhao, S. (2022b). Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jebbara, S. and Cimiano, P. (2019). Zero-shot cross-lingual opinion target extraction. In Burstein, J., Doran, C., and Solorio, T., editors, *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.
- Klinger, R. and Cimiano, P. (2015). Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 153–163, Beijing, China. Association for Computational Linguistics.
- Lambert, P. (2015). Aspect-level cross-lingual sentiment classification with constrained SMT. In Zong, C. and

- Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 781–787, Beijing, China. Association for Computational Linguistics.
- Li, X., Bing, L., Zhang, W., Li, Z., and Lam, W. (2020). Un-supervised cross-lingual adaptation for sequence tagging and beyond. *arXiv preprint arXiv:2010.12405*.
- Lin, N., Fu, Y., Lin, X., Zhou, D., Yang, A., and Jiang, S. (2023). Cl-xabsa: Contrastive learning for cross-lingual aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2935–2946.
- Lin, N., Zeng, M., Liao, X., Liu, W., Yang, A., and Zhou, D. (2024). Addressing class-imbalance challenges in cross-lingual aspect-based sentiment analysis: Dynamic weighted loss and anti-decoupling. *Expert Systems with Applications*, 257:125059.
- Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Mao, Y., Shen, Y., Yang, J., Zhu, X., and Cai, L. (2022). Seq2Path: Generating sentiment tuples as paths of a tree. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2215–2225, Dublin, Ireland. Association for Computational Linguistics.
- OpenAI (2024). *GPT-4o*. Accessed November 2024.
- Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., and Eryigit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In Bethard, S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P., and Zesch, T., editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.
- Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In Nakov, P. and Zesch, T., editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Shazeer, N. and Stern, M. (2018). Adafactor: Adaptive learning rates with sublinear memory cost. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.
- Šmíd, J., Priban, P., and Kral, P. (2024). LLaMA-based models for aspect-based sentiment analysis. In De Clercq, O., Barriere, V., Barnes, J., Klinger, R., Sedoc, J., and Tafreshi, S., editors, *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 63–70, Bangkok, Thailand. Association for Computational Linguistics.
- Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, page 6000–6010. Curran Associates, Inc.
- Wan, H., Yang, Y., Du, J., Liu, Y., Qi, K., and Pan, J. Z. (2020). Target-aspect-sentiment joint detection for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9122–9129.
- Wang, W. and Pan, S. J. (2018). Transition-based adversarial network for cross-lingual aspect extraction. In *IJCAI*, pages 4475–4481.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Xianlong, L., Yang, M., and Wang, Y. (2023). Tagging-assisted generation model with encoder and decoder supervision for aspect sentiment triplet extraction. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2093, Singapore. Association for Computational Linguistics.
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., and Lam, W. (2021a). Aspect sentiment quad prediction as paraphrase generation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Nat-*

ural Language Processing, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhang, W., Deng, Y., Liu, B., Pan, S., and Bing, L. (2024). Sentiment analysis in the era of large language models: A reality check. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Zhang, W., He, R., Peng, H., Bing, L., and Lam, W. (2021b). Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *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.

Zhang, W., Li, X., Deng, Y., Bing, L., and Lam, W. (2021c). Towards generative aspect-based sentiment analysis. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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.

Zhang, W., Li, X., Deng, Y., Bing, L., and Lam, W. (2022). A survey on aspect-based sentiment analysis: tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

APPENDIX

Constrained Decoding

Algorithm 1 shows the pseudo-code of proposed constrained decoding algorithm.

Data: Generated sequence, Input sentence tokens, Special token map
Result: Candidate tokens for the next step
 Get positions of “[” and “]” in the generated sequence;
if no “[” tokens generated then
 | **return “[”;**
end
 Count “[” and “]” tokens and find last “[”;
 Get last generated token;
if fewer “]” than “[” and last generated token is special then
 | **return “]”;**
end
if last generated token is “[” then
 | **if last special token is “;” or none then**
 | | **return “A”;**
 | **end**
 | **if last special token is “A” then**
 | | **return “C”;**
 | **end**
 | **if last special token is “C” then**
 | | **return “P”;**
 | **end**
 | **if last special token is “P” then**
 | | **return “;”;**
 | **end**
end
if last special token is “;” then
 | **return “[”;**
end
 Initialize result as an empty list;
if last special token is “A” then
 | Add input sentence tokens and “it” to result;
end
if last special token is “C” then
 | Add category tokens to result;
end
if last special token is “P” then
 | Add sentiment tokens to result;
end
if last generated token is not “]” then
 | Add “]” to result;
 | **if last special token is “P” then**
 | | Add “(eos)” to result;
 | **end**
end
return result;

Algorithm 1: Proposed constrained decoding for the TASD task.