Improving Generative Cross-lingual Aspect-Based Sentiment Analysis with Constrained Decoding

Jakub Šmíd[1,2][0000−0002−4492−5481], Pavel Přibáň[1][0000−0002−8744−8726], and Pavel Král[1,2][0000−0002−3096−675X]

[1] Department of Computer Science and Engineering, University of West Bohemia in Pilsen, Univerzitní, Pilsen, Czech Republic
[2] NTIS - New Technologies for the Information Society, University of West Bohemia in Pilsen, Univerzitní, Pilsen, Czech Republic
https://nlp.kiv.zcu.cz/
{jaksmid,pribanp,pkral}@kiv.zcu.cz

Abstract. While aspect-based sentiment analysis (ABSA) has made substantial progress, challenges remain for low-resource languages, which are often overlooked in favour of English. Current cross-lingual ABSA approaches focus on limited, less complex tasks and often rely on external translation tools. This paper introduces a novel approach using constrained decoding with sequence-to-sequence models, eliminating the need for unreliable translation tools and improving cross-lingual performance by 5% on average for the most complex task. The proposed method also supports multi-tasking, which enables solving multiple ABSA tasks with a single model, with constrained decoding boosting results by more than 10%.
We evaluate our approach across seven languages and six ABSA tasks, surpassing state-of-the-art methods and setting new benchmarks for previously unexplored tasks. Additionally, we assess large language models (LLMs) in zero-shot, few-shot, and fine-tuning scenarios. While LLMs perform poorly in zero-shot and few-shot settings, fine-tuning achieves competitive results compared to smaller multilingual models, albeit at the cost of longer training and inference times.
We provide practical recommendations for real-world applications, enhancing the understanding of cross-lingual ABSA methodologies. This study offers valuable insights into the strengths and limitations of cross-lingual ABSA approaches, advancing the state-of-the-art in this challenging research domain.

Keywords: Cross-lingual Aspect-Based Sentiment Analysis · Aspect-Based Sentiment Analysis · Large Language Models · Transformers · Constrained Decoding

1 Introduction

Aspect-based sentiment analysis (ABSA) is a natural language processing (NLP) task that focuses on identifying sentiment associated with specific aspects or features of a product or service, providing a more detailed examination than traditional sentiment analysis. ABSA finds practical applications in diverse fields

such as product marketing, customer feedback analysis, and reputation management. However, despite its significance, ABSA research has predominantly concentrated on English, leaving a gap in understanding the challenges of conducting ABSA in other languages, notably due to the lack of annotated data. Nevertheless, manual data annotation is resource-intensive and expensive, especially for languages with smaller speaker populations. To address this challenge, researchers have turned to cross-lingual sentiment analysis as a promising solution. This approach involves transferring knowledge from a *source language*, typically a resource-rich language with a large amount of annotated data, to a *target language*, which usually has limited resources, enabling the model to leverage the information of annotated data from the source language to perform sentiment analysis in the target language.

ABSA involves several sentiment elements [36]: 1) aspect term (a) is a word or phrase that represents the aspect within the text, 2) aspect category (c) defines unique aspects of an entity, and 3) sentiment polarity (p) indicates the orientation of the sentiment. For example, in a sentence *"Delicious tea"*, these elements are *"tea"*, *"drinks quality"*, *"Delicious"*, and *"positive"*. ABSA encompasses various tasks, including simple ones that focus on predicting a single sentiment element, such as aspect term extraction (ATE) [26] or aspect category detection (ACD) [26]. Recently, there has been a shift towards compound ABSA tasks that predict multiple sentiment elements simultaneously, making them more challenging as the number of elements to predict increases. These tasks include aspect category term extraction (ACTE) [25], aspect category sentiment analysis (ACSA) [28], end-to-end ABSA (E2E-ABSA) [38], and target-aspect-sentiment detection (TASD) [37]. Table 1 shows the output format of selected ABSA tasks.

Table 1. Output format for selected ABSA tasks for input: *"Delicious tea but pricey soup"*. Symbols a, c, and p denote aspect term, aspect category, and sentiment polarity, respectively.

Type	Task	Output	Example output
Simple	ATE	$\{a\}$	{"tea", "soup"}
	ACD	$\{c\}$	{drinks, food}
Compound	E2E-ABSA	$\{(a, p)\}$	{("tea", POS), ("soup", NEG)}
	ACSA	$\{(c, p)\}$	{(drinks, POS), (food, NEG)}
	ACTE	$\{(a, c)\}$	{("tea", drinks), ("soup", food)}
	TASD	$\{(a, c, p)\}$	{("tea", drinks, POS), ("soup", food, NEG)}

Multilingual pre-trained language models (mPLMs) based on the Transformer architecture [35], such as mBERT [6], XLM-R [4], and mT5 [41], have become the standard for cross-lingual transfer in NLP [12]. Typically, these multilingual models are fine-tuned on labelled data in the source language and applied directly to target language data for inference, leveraging the language knowledge acquired during pre-training. The zero-shot method uses only data from the source language for fine-tuning and relies solely on pre-training for

language-specific knowledge, which may not adequately cover low-resource languages. Using translated target language data with projected labels presents a potential solution, although its effectiveness depends on the quality of translation and label projection. However, this approach is also relatively expensive and complex.

Large language models (LLMs), such as ChatGPT [23] and LLaMA-based models [34], excel in zero-shot and few-shot scenarios across various NLP tasks [2]. LLMs typically comprise billions of parameters, with models exceeding 10 billion parameters considered large [43, 21][3]. Their size makes traditional fine-tuning challenging, making *prompting* a preferred alternative, where task descriptions guide model outputs without extensive fine-tuning. However, fine-tuning approaches on adequate data consistently outperform LLMs on compound ABSA tasks [43, 10]. Techniques like QLoRA [5] offer a way to fine-tune large models efficiently on a single GPU, yet fine-tuned open-source LLMs for cross-lingual ABSA are still underexplored [36].

Existing research on cross-lingual ABSA presents several key limitations. First, there is a considerably lower amount of cross-lingual ABSA research compared to monolingual ABSA. Second, the scope of cross-lingual ABSA tasks is narrow, particularly in compound tasks where multiple sentiment elements are involved. While recent English ABSA research has increasingly focused on these complex compound tasks, cross-lingual studies often address only simpler tasks and explores only one compound task, leaving most of the compound tasks underexplored. Third, most of the current cross-lingual ABSA research relies on older machine learning techniques, and there is a scarcity of studies employing modern Transformer-based models, with the exception of some works that use encoder-only Transformer models. In contrast, state-of-the-art performance in monolingual ABSA is frequently achieved with sequence-to-sequence approaches, which remain unexplored in cross-lingual settings. Fourth, most existing cross-lingual ABSA approaches depend heavily on external translation tools. Although these tools help bridge the language gap, they introduce additional complexity and potential errors, which can negatively affect the quality of the results. Finally, there has been little investigation into the potential of fine-tuned open-source LLMs for cross-lingual ABSA. While preliminary research [30] in monolingual settings suggests that LLMs can deliver state-of-the-art results for ABSA tasks in English, their cross-lingual applications remain largely unexplored.

The main motivation of this paper is to address the limited research on compound cross-lingual ABSA tasks and the absence of sequence-to-sequence approaches, which are critical in monolingual ABSA, while also reducing the reliance on external translation tools that introduce complexity and potential inaccuracies. Another motivation is the lack of research on cross-lingual capabilities of LLMs for ABSA. To address these issues, we introduce a novel sequence-to-sequence method enhanced with constrained decoding. This technique refines a model's token generation process to ensure that predictions adhere to the nec-

[3] This work considers models with 7 billion parameters and more as large models.

essary output structure by restricting them to acceptable tokens. Our approach eliminates the dependency on external translation tools, which can be unreliable and may introduce errors in cross-lingual ABSA tasks. Additionally, we evaluate several LLMs in cross-lingual settings and compare them to smaller multilingual models.

We achieve excellent results in zero-shot cross-lingual settings and evaluate our approach on seven languages and across six tasks, four of which are compound. Furthermore, given the limited research on LLMs for both monolingual and cross-lingual ABSA, we explore their capabilities for handling compound ABSA tasks, assessing their performance and potential in this context. Our main contributions are as follows:

- We propose a novel method that employs constrained decoding combined with sequence-to-sequence models. Constrained decoding significantly improves cross-lingual ABSA compared to the methods that do not utilize this enhancement. The proposed approach outperforms significantly state-of-the-art results for existing cross-lingual ABSA tasks and eliminates the need for external translation tools. To the best of our knowledge, this study represents the first application of sequence-to-sequence models and large language models for cross-lingual ABSA.
- We conduct extensive experiments on benchmark datasets across seven languages, evaluating performance on six distinct ABSA tasks. In addition to improving results on established cross-lingual and monolingual ABSA tasks, we pioneer the evaluation of several previously unexplored cross-lingual ABSA tasks.
- Our methodology supports multi-tasking capabilities, enabling the simultaneous solution of multiple ABSA tasks using a single model. Constrained decoding proves particularly effective in multi-tasking settings, consistently enhancing results by over 10%.
- We systematically evaluate the performance of various LLMs in zero-shot, few-shot, and fine-tuning scenarios across monolingual and cross-lingual settings. Our findings demonstrate that fine-tuning LLMs is necessary to achieve results comparable to those of smaller multilingual models. Additionally, we show that more advanced LLMs significantly outperform their counterparts, emphasizing the importance of model selection. This evaluation highlights the strengths and limitations of each approach, providing valuable insights for selecting appropriate models in different ABSA applications.
- We include a comparison of training and inference time requirements for different models. This analysis reveals that while some models achieve high performance, they come with substantial computational costs, highlighting the efficiency advantages of our method.
- Based on our extensive experimentation and analysis, we propose practical recommendations tailored for diverse scenarios in both cross-lingual and monolingual ABSA. These recommendations aim to guide researchers and practitioners towards effective model selection and deployment strategies in real-world applications.

This paper extends our previous publication [31] and together, the two works pioneer the use of sequence-to-sequence models and LLMs for cross-lingual ABSA. Our approach differs from much of the existing research, which often depends on external translation tools and yields mixed results. Moreover, we explore cross-lingual ABSA in greater depth, evaluating six ABSA tasks across seven languages. In contrast, prior studies typically focus on fewer languages and only a single compound ABSA task. We also offer practical guidance for selecting the most suitable approach to monolingual and cross-lingual ABSA, depending on task-specific constraints and real-world requirements. These recommendations enhances the flexibility of our study and provides actionable insights for practitioners seeking effective deployment strategies.[4]

Compared to our earlier work [31], this paper introduces three additional ABSA tasks (ATE, ACD, ACSA), one more language (Czech), broader combinations of source and target languages, an evaluation of different LLMs, and the introduction of multi-tasking models.

The rest of the paper is organized as follows. Section 2 reviews the related work in cross-lingual and monolingual ABSA. Section 3 describes our proposed methodology. Section 4 details our experimental setup and the datasets utilized. Section 5 presents the results and findings derived from our experiments. Section 6 thoroughly discusses the methodology's implications and offers practical recommendations based on our results. Finally, Section 7 summarizes our conclusions and highlights the contributions of this study to the field of ABSA.

2 Related Work

Modern monolingual ABSA is increasingly framed as a text generation task. Annotation-style and extraction-style paradigms show the viability of generative approaches [45]. Several works tackle sentiment quad prediction using natural language formats [42], multi-task frameworks with element prompts [9], or tree-based tuple generation [20]. Others explore the impact of sentiment element order [13] or use multiple permutations to improve prediction [10].

Early cross-lingual ABSA methods focus on single tasks and typically rely on translation followed by label projection, either directly or using alignment tools like FastAlign [8]. Data quality is improved through co-training [46], instance selection [15], or constrained SMT [16], while cross-lingual embeddings enable language-agnostic learning [1, 14, 39]. Recent work shifts towards E2E-ABSA using mPLMs like XLM-R [4] in combination with machine translation. Some approaches use only translated data [17], alignment-free label projection combined with distillation on unlabelled target language data [44], or contrastive learning [18], but they face challenges such as translation noise, reliance on external tools, and limited language coverage. Adapting mPLMs effectively for cross-lingual ABSA remains an open problem.

[4] We publish our codebase and data at https://github.com/biba10/Generative-Cross-lingual-ABSA.

LLMs like ChatGPT [23] have shown strong performance in zero-shot settings, but their effectiveness in ABSA drops compared to smaller fine-tuned models [10, 43]. This gap grows with task complexity, especially in compound ABSA. Nevertheless, fine-tuned LLaMA-based models achieve state-of-the-art results on several compound tasks [30]. A key limitation of many LLMs is their English-centric pre-training.

3 Methodology

This section describes our approach to handling the triplet task, i.e. the TASD task, which can be easily adapted for other tasks with slight modifications.

3.1 Problem Statement

Given a sentence as input, the objective is to predict all sentiment triplets $T = (a, c, p)$, where each triplet comprises an aspect term (a), aspect category (c), and sentiment polarity (p). We adopt approaches from prior research [10, 13, 42] to transform sentiment elements (a, c, p) into natural language representations (e_a, e_c, e_p).

For the aspect term a, the representation e_a is straightforward: it is the original aspect term, except in the case of a *"NULL"* aspect term, which we replace with *"it"*.

For the aspect category c, the original format is `ENTITY#ATTRIBUTE`. We transform it into the representation e_c by converting all letters to lowercase and replacing the "#" with a space. For example, the aspect category `FOOD#QUALITY` becomes *"food quality"*.

For the sentiment polarity p, we use the following mapping function $\mathcal{P}_p(p)$ to obtain the representation e_p:

$$\mathcal{P}_p(p) = \begin{cases} great & \text{if } p \text{ is } positive, \\ ok & \text{if } p \text{ is } neutral, \\ bad & \text{if } p \text{ is } negative. \end{cases} \tag{1}$$

3.2 Constructing Input and Output

In crafting the inputs and outputs for our model, we use element markers [10, 13] to denote each sentiment element: [A] for e_a, [C] for e_c, and [P] for e_p. These markers precede each element, collectively forming the target sequence. Additionally, we append these markers as a prompt to the input sequence s to guide the model towards producing the correct output as $x = s \mid$ [A] [C] [P]. Following the prioritization order $e_a > e_c > e_p$ recommended by a previous study [10], we create input-output pairs as shown below:

***Input (x)*:** Delicious tea but pricey soup | [A] [C] [P]

***Output (y)*:** [A] tea [C] drinks quality [P] great [;] [A] soup [C] food prices [P] bad

For sentences containing multiple sentiment tuples, we use the sequence [;] to concatenate their target schemes into the final target sequence.

3.3 Training

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

$$\mathcal{L} = -\sum_{i=1}^{n} \log p_{\mathbf{\Theta}}(y_i|\mathbf{e}, y_{<i}), \tag{2}$$

where n is the length of the target sequence y. Figure 1 shows the example of input creation, training and generation process.

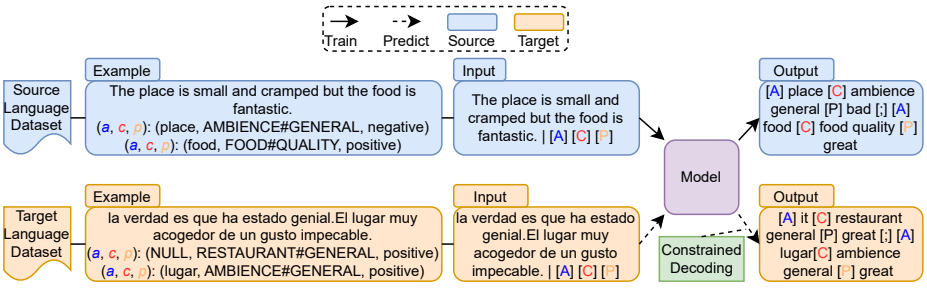


Fig. 1. Overview of our method including input creation, training and generation process with expected output [31].

Multi-Task Learning We enable multi-task learning by modifying the prompt appended to the input sentence, allowing the model to predict multiple tasks simultaneously. Depending on the additional markers included in the input prompt, the model selects the task to perform. Figure 2 illustrates an example of multi-task learning.

3.4 Scheme-Guided Constrained Decoding

In refining our model, we encountered issues where the fine-tuned model might not follow the desired output format, occasionally generate aspect terms in the

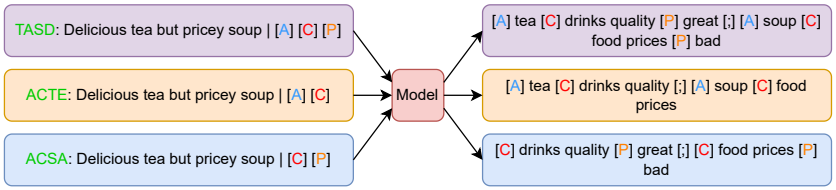


Fig. 2. Example of multi-task learning. The model simultaneously addresses multiple ABSA tasks based on different input prompts.

source language rather than the target language or generate text not present in the original review. To address this, we introduce scheme-guided constrained decoding (CD) [3], a method designed to ensure the generated elements align with their intended vocabulary sets by leveraging target schema information. This approach proves particularly beneficial in scenarios with limited training data in monolingual settings [10].

Constrained decoding solves common decoding problems, where the model searches the entire vocabulary for the subsequent token. By employing constrained decoding, we prevent the generation of undesired sequences that fail to meet our specifications. This technique operates dynamically, adjusting the list of potential candidate tokens based on the current state and checking the tokens individually to ensure a more controlled and precise generation process. For instance, when the current token is '[', the subsequent token selection should be restricted to special terms: 'A', 'C', 'P', and ';'. Furthermore, constrained decoding monitors the generated output and current term, guiding the selection of subsequent tokens according to the criteria outlined in Table 2.

Table 2. Candidate lists of tokens for the TASD task. `<eos>` indicates the end of a sequence, and "..." denotes arbitrary text.

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

For example, when the model generates an aspect term, constrained decoding limits the available tokens to only those present in the input (target) language

sentence. Similarly, the available tokens are restricted to predefined aspect categories when generating an aspect category. The same principle applies to sentiment polarity, where the tokens can only come from permissible sentiment polarities. This method ensures that the generated output is consistent with the expected format and content.

Algorithm 1 shows the pseudo-code of proposed CD 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 "[";
Count "[" and "]" tokens and find last "[";
Get last generated token;
if *fewer "]" than "[" and last generated token is special* **then**
  return "]";
if *last generated token is "["* **then**
  if *last special token is ";" or none* **then**
    return "A";
  if *last special token is "A"* **then**
    return "C";
  if *last special token is "C"* **then**
    return "P";
  if *last special token is "P"* **then**
    return ";";
if *last special token is ";"* **then**
  return "[";
Initialize result as an empty list;
if *last special token is "A"* **then**
  Add input sentence tokens and "it" to result;
if *last special token is "C"* **then**
  Add category tokens to result;
if *last special token is "P"* **then**
  Add sentiment tokens to result;
if *last generated token is not "]"* **then**
  Add "]" to result;
  if *last special token is "P"* **then**
    Add "⟨eos⟩" to result;
return result;

Algorithm 1: Proposed constrained decoding for the TASD task [31].

3.5 Large Language Models Prompts

Figure 3 illustrates the prompt for the TASD task with expected input, output, and few-shot demonstrations in Czech. The prompt is adaptable for various tasks

by omitting the unnecessary sentiment element for the specific task, such as the sentiment polarity for the ACTE task. Few-shot examples are drawn from the first ten examples of the training dataset in the respective language for a fair evaluation. The distribution of labels in these examples is similar to the entire dataset, ensuring a random and representative sample.

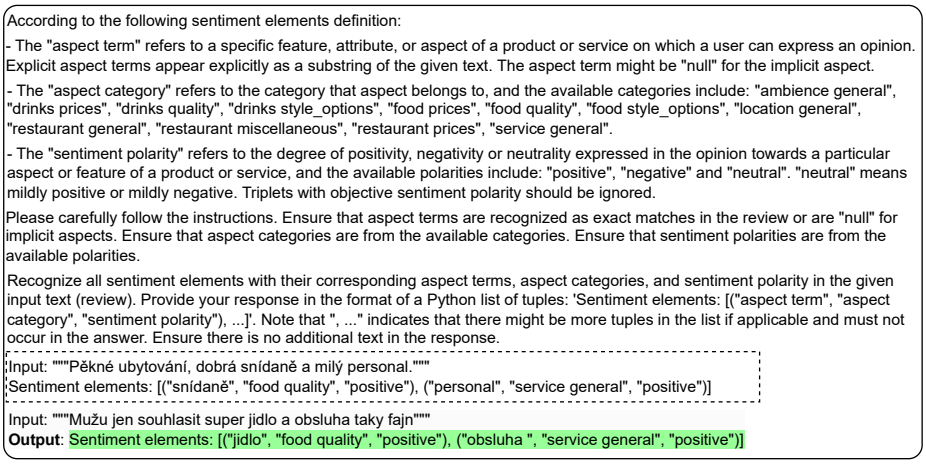


Fig. 3. Prompt for the TASD task with example input, expected output in a green box, and three demonstrations in Czech enclosed in a dashed box. The demonstrations are used solely in few-shot scenarios [31].

4 Experiment Setup

This section describes the datasets used in our experiments, the ABSA tasks we tackle, and the detailed experimental setup, including hyperparameters and evaluation metrics.

4.1 Data & Tasks

For the experiments, we use the SemEval-2016 Task 5 dataset [24], which contains restaurant reviews in English (en), Spanish (es), French (fr), Dutch (nl), Russian (ru), and Turkish (tr). Each dataset is pre-split into training and test sets. We further divide the training data into a 9:1 ratio to obtain a validation set. Additionally, we use the CsRest-M dataset [32] that contains restaurant reviews in Czech (cs). This dataset is provided with training, validation and test splits. Table 3 shows the statistics for each dataset.

We address six ABSA tasks: ATE, ACD, ACSA, E2E-ABSA, ACTE, and TASD.

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

		Cs	En	Es	Fr	Nl	Ru	Tr
Train	Sentences	2,151	1,800	1,863	1,559	1,549	3,289	1,108
	Triplets	4,386	2,266	2,455	2,276	1,676	3,697	1,386
	Categories	12	12	12	12	12	12	12
	POS/NEG/NEU	2,663/1,338/385	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	961	569	700	694	513	821	135
Dev	Sentences	240	200	207	174	173	366	124
	Triplets	483	241	265	254	184	392	149
	Categories	12	11	11	12	11	12	10
	POS/NEG/NEU	278/161/44	154/77/10	189/67/8	115/120/15	94/62/28	298/68/26	74/65/10
	NULL aspects	104	58	83	66	64	109	15
Test	Sentences	798	676	881	694	575	1,209	144
	Triplets	1,609	859	1,072	954	613	1,300	159
	Categories	12	12	12	13	13	12	11
	POS/NEG/NEU	972/497/140	611/204/44	750/274/48	441/434/79	369/211/33	870/321/103	104/49/6
	NULL aspects	342	209	341	236	219	325	0

4.2 Experiment Details

We employ two models, large mT5 [41] and large mBART [33], from the Huggingface Transformers library[5] [40]. The choice of the mT5 model follows prior English research [9, 10, 13, 42, 45] utilizing the monolingual T5 model [27]. The inclusion of the mBART model aims to assess the robustness of our approach with a different backbone model. All experiments are conducted using an NVIDIA A40 with 48 GB GPU memory.

We maintain the same settings across all experiments, chosen based on consistent validation performance across all languages and tasks. We use a batch size 16 and conduct training for 20 epochs using a greedy search for decoding in all experiments. For the mT5 model, we set the learning rate to 1e-4 and utilize the Adafactor optimizer [29]. For the mBART model, we set the learning rate to 1e-5 and employ the AdamW optimizer [19]. Given that all examples fit within the maximum length of 512, there is no need to trim the input to meet the maximum length requirements of the models. During the fine-tuning process, we update all model parameters. Multi-tasking models are fine-tuned on all six tasks simultaneously using the same hyperparameters as single-task models. Since mT5 outperformed mBART in preliminary experiments and to reduce the number of experiments, we use only mT5 for multi-tasking.

4.3 Large Language Models

We compare our method with several LLMs, – LLaMA 2 [34], LLaMA 3 [7], Orca 2 [22], and ChatGPT (`gpt-3.5-turbo`) [23]. Notably, we assess both the 13B and 7B versions of LLaMA 2 and Orca 2, as well as the 8B version of LLaMA 3. Our evaluation encompasses zero-shot and few-shot prompts across compound tasks. Moreover, we conduct instruction tuning for Orca 2, LLaMA 2, and LLaMA 3 in monolingual and cross-lingual settings.

[5] https://github.com/huggingface/transformers

For ChatGPT, we utilize the official paid API[6]. For other models, which are open-source, unlike ChatGPT, we apply 4-bit quantization to fit the model into GPU memory. Preliminary results indicate that 4-bit quantization performs comparably to 8-bit quantization.

Instruction Tuning We employ QLoRA [5] with 4-bit NormalFloat quantization to fine-tune the LLMs. This technique utilizes a quantized 4-bit frozen backbone LLM with a small set of learnable LoRA [11] weights, enabling fine-tuning of LLMs on a single consumer GPU. Following the recommendations in the QLoRA paper, we adopt a batch size of 16, a constant learning rate of 2e-4, AdamW optimizer, and apply LoRA adapters on all linear transformer block layers with LoRA settings of $r = 64$ and $\alpha = 16$. Utilizing the zero-shot prompt (without demonstrations) shown in Figure 3, we fine-tune the model for up to 5 epochs, selecting the best-performing model based on validation loss.

4.4 Evaluation Metrics

The primary evaluation metric used is the micro F1 score, a standard metric in ABSA research [10, 42, 44]. We define a predicted sentiment tuple as correct only if all its elements precisely match the gold tuple. Results are presented with a 95% confidence interval derived from 5 runs with different random seeds.

4.5 Compared Methods

Where possible, we compare our method to existing cross-lingual works. This includes the ATE task and E2E-ABSA tasks. However, the compared methods for the ATE task involves older machine learning approaches, as mentioned in Section 2. For the E2E-ABSA tasks, the methods use encoder-based Transformer models. This lack of research on compound ABSA tasks is one of the motivations for this paper.

5 Results

This section presents results for six ABSA tasks using mT5 and mBART models, and four compound ABSA tasks using LLMs. In the tables, notation such as en→cs indicates English as the source language and Czech as the target. We begin with simple ABSA tasks, followed by pair extraction tasks, and then provide a detailed analysis of the most complex task, TASD. Next, we present results across all LLMs, expanding on the earlier focus on ChatGPT and Orca 2 13B, which generally perform best. Finally, we discuss training and inference speeds for different models and provide an error analysis highlighting key challenges in the tasks.

[6] https://platform.openai.com/

5.1 Simple ABSA Tasks

Table 4 shows the results for simple ABSA tasks, highlighting the effectiveness of constrained decoding for the cross-lingual ATE task. This advantage arises from instances where the model correctly identifies the aspect term but in the source language instead of the target one. For example, the model might predict *"place"* instead of the Czech *"místo"* when Czech is the target language. Constrained decoding mitigates this problem effectively by restricting the available tokens for generation to only those from the input target language sentence. The results indicate that mT5 generally outperforms mBART and benefits more from constrained decoding. For the ACD task, such issues with generation do not occur since aspect categories are always predicted in English, resulting in no significant improvement from constrained decoding. Furthermore, constrained decoding does not significantly enhance monolingual results, as the model has no problems generating the aspect terms in a given source language.

Table 4. Monolingual (for target languages) and cross-lingual F1 scores for for simple ABSA tasks. **Bold** indicates significant improvements with constrained decoding (CD) over without. The best cross-lingual result per task and language pair is underlined. Asterisks (*) denote multi-tasking models. The compared works have the same data splits and task definition.

Task	Settings	Model		En→cs	En→es	En→fr	En→nl	En→ru	En→tr	Cs→en	Es→en	Fr→en	Nl→en	Ru→en	Tr→en
ACD	Monolingual	mT5	Without CD	$84.5^{\pm0.4}$	$82.3^{\pm0.7}$	$77.4^{\pm0.4}$	$81.8^{\pm0.6}$	$85.7^{\pm0.4}$	$80.7^{\pm2.4}$	$84.7^{\pm0.6}$	$84.7^{\pm0.6}$	$84.7^{\pm0.6}$	$84.7^{\pm0.6}$	$84.7^{\pm0.6}$	$84.7^{\pm0.6}$
			With CD	$84.2^{\pm0.6}$	$82.3^{\pm0.6}$	$76.9^{\pm0.5}$	$80.9^{\pm0.6}$	$85.8^{\pm0.8}$	$81.3^{\pm0.7}$	$84.8^{\pm0.7}$	$84.8^{\pm0.7}$	$84.8^{\pm0.7}$	$84.8^{\pm0.7}$	$84.8^{\pm0.7}$	$84.8^{\pm0.7}$
			Without CD*	$84.5^{\pm0.4}$	$82.3^{\pm0.7}$	$77.4^{\pm0.4}$	$81.8^{\pm0.6}$	$85.7^{\pm0.4}$	$80.7^{\pm2.4}$	$84.7^{\pm0.6}$	$84.7^{\pm0.6}$	$84.7^{\pm0.6}$	$84.7^{\pm0.6}$	$84.7^{\pm0.6}$	$84.7^{\pm0.6}$
			With CD*	$84.2^{\pm0.6}$	$82.3^{\pm0.6}$	$76.9^{\pm0.5}$	$80.9^{\pm0.6}$	$85.8^{\pm0.8}$	$81.3^{\pm0.7}$	$84.8^{\pm0.7}$	$84.8^{\pm0.7}$	$84.8^{\pm0.7}$	$84.8^{\pm0.7}$	$84.8^{\pm0.7}$	$84.8^{\pm0.7}$
		mBART	Without CD	$81.5^{\pm0.4}$	$79.6^{\pm1.1}$	$75.2^{\pm0.6}$	$77.9^{\pm1.2}$	$85.0^{\pm1.3}$	$74.9^{\pm3.4}$	$82.3^{\pm1.2}$	$82.3^{\pm1.2}$	$82.3^{\pm1.2}$	$82.3^{\pm1.2}$	$82.3^{\pm1.2}$	$82.3^{\pm1.2}$
			With CD	$81.6^{\pm0.9}$	$78.8^{\pm1.1}$	$74.0^{\pm0.9}$	$77.3^{\pm0.8}$	$83.9^{\pm1.6}$	$78.4^{\pm2.0}$	$83.2^{\pm1.3}$	$83.2^{\pm1.3}$	$83.2^{\pm1.3}$	$83.2^{\pm1.3}$	$83.2^{\pm1.3}$	$83.2^{\pm1.3}$
	Cross-lingual	mT5	Without CD	$77.2^{\pm1.2}$	$76.3^{\pm0.7}$	$73.9^{\pm0.9}$	$74.7^{\pm0.3}$	$80.2^{\pm1.6}$	$76.5^{\pm1.6}$	$\underline{80.9}^{\pm0.7}$	$\underline{80.5}^{\pm0.2}$	$81.5^{\pm0.6}$	$80.3^{\pm0.6}$	$\underline{81.1}^{\pm0.6}$	$\underline{79.3}^{\pm1.0}$
			With CD	$77.3^{\pm0.8}$	$77.3^{\pm0.3}$	$74.8^{\pm0.8}$	$75.4^{\pm1.4}$	$80.2^{\pm0.9}$	$76.4^{\pm1.6}$	$80.8^{\pm0.9}$	$80.4^{\pm0.8}$	$81.3^{\pm0.6}$	$\underline{80.7}^{\pm0.5}$	$80.9^{\pm0.4}$	$78.8^{\pm0.6}$
			Without CD*	$76.9^{\pm1.1}$	$76.8^{\pm0.6}$	$\underline{75.9}^{\pm1.3}$	$75.3^{\pm1.2}$	$80.3^{\pm0.6}$	$76.0^{\pm2.4}$	$79.9^{\pm0.8}$	$80.1^{\pm0.5}$	$80.3^{\pm1.1}$	$80.5^{\pm0.6}$	$80.8^{\pm0.5}$	$78.2^{\pm0.9}$
			With CD*	$\underline{77.8}^{\pm0.9}$	$\underline{77.7}^{\pm1.0}$	$75.7^{\pm0.5}$	$\underline{75.6}^{\pm1.2}$	$\underline{80.8}^{\pm0.7}$	$\underline{77.1}^{\pm2.0}$	$80.6^{\pm0.6}$	$80.1^{\pm0.3}$	$\underline{81.6}^{\pm0.8}$	$80.4^{\pm1.3}$	$81.0^{\pm0.7}$	$78.5^{\pm0.8}$
		mBART	Without CD	$69.2^{\pm2.2}$	$69.7^{\pm0.8}$	$68.4^{\pm0.9}$	$68.9^{\pm2.8}$	$77.7^{\pm1.6}$	$60.6^{\pm6.5}$	$74.3^{\pm2.4}$	$75.7^{\pm1.5}$	$75.6^{\pm2.5}$	$77.0^{\pm1.9}$	$75.0^{\pm1.4}$	$72.9^{\pm1.7}$
			With CD	$67.5^{\pm1.7}$	$70.7^{\pm1.4}$	$69.1^{\pm1.1}$	$69.5^{\pm2.8}$	$77.8^{\pm1.5}$	$63.5^{\pm3.0}$	$75.7^{\pm2.0}$	$76.9^{\pm0.9}$	$76.9^{\pm1.2}$	$77.4^{\pm1.3}$	$78.0^{\pm1.7}$	$74.7^{\pm1.5}$
ATE	Monolingual	mT5	Without CD	$81.4^{\pm0.4}$	$80.4^{\pm0.5}$	$76.2^{\pm0.8}$	$80.0^{\pm0.4}$	$79.9^{\pm0.6}$	$68.1^{\pm2.4}$	$83.7^{\pm0.3}$	$83.7^{\pm0.3}$	$83.7^{\pm0.3}$	$83.7^{\pm0.3}$	$83.7^{\pm0.3}$	$83.7^{\pm0.3}$
			With CD	$81.5^{\pm0.7}$	$80.7^{\pm0.9}$	$77.3^{\pm0.4}$	$72.8^{\pm0.9}$	$79.6^{\pm0.6}$	$68.0^{\pm2.5}$	$83.6^{\pm0.8}$	$83.6^{\pm0.8}$	$83.6^{\pm0.8}$	$83.6^{\pm0.8}$	$83.6^{\pm0.8}$	$83.6^{\pm0.8}$
			Without CD*	$81.9^{\pm0.7}$	$81.2^{\pm0.8}$	$78.0^{\pm1.1}$	$81.9^{\pm1.7}$	$80.6^{\pm0.2}$	$69.6^{\pm1.6}$	$84.2^{\pm0.7}$	$84.2^{\pm0.7}$	$84.2^{\pm0.7}$	$84.2^{\pm0.7}$	$84.2^{\pm0.7}$	$84.2^{\pm0.7}$
			With CD*	$82.4^{\pm0.7}$	$80.9^{\pm0.6}$	$78.7^{\pm1.2}$	$75.1^{\pm1.3}$	$81.1^{\pm0.5}$	$68.6^{\pm1.4}$	$84.3^{\pm0.2}$	$84.3^{\pm0.2}$	$84.3^{\pm0.2}$	$84.3^{\pm0.2}$	$84.3^{\pm0.2}$	$84.3^{\pm0.2}$
		mBART	Without CD	$79.2^{\pm0.9}$	$79.8^{\pm1.2}$	$74.6^{\pm1.0}$	$79.0^{\pm1.3}$	$78.6^{\pm0.8}$	$65.2^{\pm5.0}$	$80.9^{\pm1.5}$	$80.9^{\pm1.5}$	$80.9^{\pm1.5}$	$80.9^{\pm1.5}$	$80.9^{\pm1.5}$	$80.9^{\pm1.5}$
			With CD	$77.7^{\pm0.7}$	$77.8^{\pm2.1}$	$72.0^{\pm0.6}$	$71.1^{\pm1.7}$	$75.4^{\pm1.0}$	$64.0^{\pm1.4}$	$80.8^{\pm0.8}$	$80.8^{\pm0.8}$	$80.8^{\pm0.8}$	$80.8^{\pm0.8}$	$80.8^{\pm0.8}$	$80.8^{\pm0.8}$
		[14]		–	68.0	–	60.0	56.0	48.0	66.0	66.0	66.0	66.0	66.0	66.0
	Cross-lingual	mT5	Without CD	$65.7^{\pm0.2}$	$63.4^{\pm0.6}$	$67.2^{\pm0.5}$	$62.3^{\pm2.2}$	$65.7^{\pm1.3}$	$52.2^{\pm1.6}$	$75.3^{\pm1.0}$	$73.1^{\pm1.8}$	$75.2^{\pm1.0}$	$70.8^{\pm0.5}$	$75.9^{\pm0.1}$	$65.9^{\pm1.1}$
			With CD	$\mathbf{68.6}^{\pm1.1}$	$\mathbf{74.9}^{\pm1.6}$	$66.5^{\pm0.6}$	$\mathbf{65.2}^{\pm0.6}$	$\mathbf{70.1}^{\pm1.3}$	$\underline{\mathbf{54.5}}^{\pm0.4}$	$\mathbf{77.3}^{\pm0.8}$	$\mathbf{76.3}^{\pm1.0}$	$\underline{\mathbf{76.8}}^{\pm0.3}$	$\underline{\mathbf{77.5}}^{\pm0.6}$	$\mathbf{77.8}^{\pm0.6}$	$66.7^{\pm0.6}$
			Without CD*	$61.8^{\pm1.5}$	$63.6^{\pm1.0}$	$63.2^{\pm1.2}$	$59.6^{\pm1.0}$	$46.0^{\pm2.3}$	$41.2^{\pm1.9}$	$77.6^{\pm0.4}$	$76.6^{\pm1.2}$	$75.6^{\pm1.2}$	$71.0^{\pm0.9}$	$77.0^{\pm1.0}$	$69.2^{\pm1.5}$
			With CD*	$\underline{\mathbf{71.9}}^{\pm1.4}$	$\underline{\mathbf{78.4}}^{\pm1.0}$	$\underline{\mathbf{71.5}}^{\pm0.8}$	$\underline{\mathbf{67.9}}^{\pm1.1}$	$\underline{\mathbf{71.2}}^{\pm1.3}$	$\mathbf{51.9}^{\pm2.2}$	$\underline{78.3}^{\pm1.5}$	$\underline{77.2}^{\pm1.2}$	$76.1^{\pm1.0}$	$\mathbf{77.1}^{\pm1.3}$	$\underline{78.8}^{\pm0.8}$	$\underline{69.4}^{\pm0.8}$
		mBART	Without CD	$62.1^{\pm1.9}$	$70.3^{\pm2.0}$	$60.9^{\pm4.3}$	$60.4^{\pm2.8}$	$66.5^{\pm2.4}$	$35.6^{\pm4.2}$	$76.2^{\pm2.3}$	$72.1^{\pm1.2}$	$73.4^{\pm1.5}$	$68.2^{\pm2.3}$	$77.1^{\pm0.6}$	$65.2^{\pm2.3}$
			With CD	$60.6^{\pm1.9}$	$71.5^{\pm1.8}$	$58.6^{\pm0.6}$	$59.7^{\pm3.1}$	$66.1^{\pm2.5}$	$38.6^{\pm3.8}$	$74.0^{\pm1.4}$	$73.4^{\pm2.7}$	$72.7^{\pm1.1}$	$71.7^{\pm1.5}$	$76.0^{\pm1.4}$	$63.4^{\pm1.6}$
		[39]		–	50.5	50.0	–	–	–	–	44.1	50.3	–	–	–
		[14]		–	50.0	–	46.0	37.0	17.0	–	43.0	–	45.0	42.0	33.0

The performance of multi-tasking models is noteworthy, as they often achieve the best overall cross-lingual results across various language pairs. Constrained decoding is especially beneficial for multi-tasking models in the ATE task, often yielding 10% or more improvements over models without constrained decoding when English is the source language. In most cases, multi-tasking models with

constrained decoding outperform or perform on par with models specialized in a single task, allowing for using a single model for multiple tasks without sacrificing performance.

When comparing monolingual and cross-lingual performance, cross-lingual results are generally about 4–6% worse than monolingual results for the ACD task. The results for all other source languages are very similar when English is the target language. However, there are significant performance drops in some cross-lingual scenarios compared to monolingual results for the ATE task. For instance, using Turkish as both the source and target language results in a performance decrease of more than 10% compared to monolingual results. Nevertheless, for most other language pairs, the cross-lingual results are only around 5% worse than monolingual results, demonstrating the effectiveness of constrained decoding.

Additionally, the table includes comparisons with existing results for the ATE task [14, 39]. Our approach, especially with constrained decoding, significantly outperforms previous benchmarks, with improvements exceeding 20% and even 30% in most cases.

In summary, the results emphasize the effectiveness of constrained decoding in improving model performance in cross-lingual settings for aspect term detection. Multi-tasking models with constrained decoding are robust and capable, often surpassing specialized models.

5.2 Pair Extraction ABSA Tasks

Table 5 presents the results for the three pair-extraction ABSA tasks: E2E-ABSA, ACSA, and ACTE. These tasks differ in complexity, but several trends are consistent across them.

In monolingual settings, ChatGPT consistently achieves the best results in both zero-shot (ZS) and few-shot (FS) configurations. The superior performance of ChatGPT can be attributed to its extensive parameter count and comprehensive training data. Few-shot learning improves performance across all models and tasks. Fine-tuning offers the significant gains, as shown for Orca 2 13B. Among sequence-to-sequence models, mT5 generally performs well despite its smaller size compared to LLMs, outperforming Orca 2 in certain lower-resource languages like Czech and Turkish. The mBART model performs consistently worse than mT5 across all tasks and languages. Constrained decoding does not affect monolingual results.

In the cross-lingual setting, mT5 generally outperforms large language models when English is the source language. All evaluated LLMs perform particularly poorly when Turkish is the target language. However, when English is the target language, Orca 2 often performs best. The difference in performance in English may be related to the predominant pre-training of models like Orca 2 on English data, which gives them a significant advantage in that language. Nonetheless, using English as the target language is less practical due to its resource-rich status and is, therefore, more commonly used as the source language in cross-language tasks. Similar to monolingual results, mT5 outperforms mBART in

Table 5. Monolingual (for target languages) and cross-lingual F1 scores for pair-wise tasks. **Bold** indicates significant improvements with constrained decoding (CD) over without. The best cross-lingual result per task and language pair is underlined. Asterisks (*) denote multi-tasking models. ZS and FS stand for zero-shot and few-shot (10 examples). Compared E2E-ABSA baselines differ in data size and task definition.

Task	Setup		Model	En→cs	En→es	En→fr	En→nl	En→ru	En→tr	Cs→en	Es→en	Fr→en	Nl→en	Ru→en	Tr→en
E2E-ABSA	ZS		ChatGPT	44.6	42.4	37.1	31.7	35.1	37.9	55.5	55.5	55.5	55.5	55.5	55.5
			Orca 2 13B	22.7	27.4	22.5	18.3	20.0	19.2	42.2	42.2	42.2	42.2	42.2	42.2
	FS		ChatGPT	54.8	59.1	51.7	51.6	51.9	42.7	62.2	62.2	62.2	62.2	62.2	62.2
			Orca 2 13B	41.0	50.6	43.9	46.1	38.7	32.0	62.4	62.4	62.4	62.4	62.4	62.4
	Monolingual Fine-tuning		Orca 2 13B	$71.8^{\pm1.1}$	$74.8^{\pm1.0}$	$69.8^{\pm2.3}$	$76.2^{\pm0.3}$	$72.2^{\pm1.2}$	$52.2^{\pm1.6}$	$82.3^{\pm0.6}$	$82.3^{\pm0.6}$	$82.3^{\pm0.6}$	$82.3^{\pm0.6}$	$82.3^{\pm0.6}$	$82.3^{\pm0.6}$
		mT5	Without CD	$73.4^{\pm0.8}$	$74.4^{\pm0.6}$	$69.9^{\pm0.5}$	$71.6^{\pm1.0}$	$72.4^{\pm0.2}$	$60.1^{\pm1.7}$	$77.7^{\pm0.4}$	$77.7^{\pm0.4}$	$77.7^{\pm0.4}$	$77.7^{\pm0.4}$	$77.7^{\pm0.4}$	$77.7^{\pm0.4}$
			With CD	$73.5^{\pm0.4}$	$75.3^{\pm0.6}$	$69.8^{\pm1.4}$	$67.0^{\pm0.4}$	$72.2^{\pm0.4}$	$60.7^{\pm1.1}$	$77.4^{\pm0.5}$	$77.4^{\pm0.5}$	$77.4^{\pm0.5}$	$77.4^{\pm0.5}$	$77.4^{\pm0.5}$	$77.4^{\pm0.5}$
			Without CD*	$73.8^{\pm0.5}$	$75.3^{\pm0.6}$	$69.4^{\pm0.8}$	$73.2^{\pm1.4}$	$72.2^{\pm0.4}$	$63.4^{\pm1.0}$	$78.1^{\pm1.1}$	$78.1^{\pm1.1}$	$78.1^{\pm1.1}$	$78.1^{\pm1.1}$	$78.1^{\pm1.1}$	$78.1^{\pm1.1}$
			With CD*	$74.6^{\pm0.5}$	$74.9^{\pm1.3}$	$69.9^{\pm0.9}$	$67.4^{\pm1.3}$	$73.2^{\pm0.7}$	$63.4^{\pm1.7}$	$78.0^{\pm0.8}$	$78.0^{\pm0.8}$	$78.0^{\pm0.8}$	$78.0^{\pm0.8}$	$78.0^{\pm0.8}$	$78.0^{\pm0.8}$
		mBART	Without CD	$69.1^{\pm0.3}$	$73.0^{\pm0.5}$	$66.4^{\pm1.1}$	$68.9^{\pm1.2}$	$68.7^{\pm1.6}$	$56.0^{\pm2.7}$	$74.3^{\pm1.4}$	$74.3^{\pm1.4}$	$74.3^{\pm1.4}$	$74.3^{\pm1.4}$	$74.3^{\pm1.4}$	$74.3^{\pm1.4}$
			Without CD	$68.8^{\pm0.8}$	$71.9^{\pm1.3}$	$64.0^{\pm1.7}$	$61.6^{\pm1.0}$	$66.2^{\pm1.1}$	$54.4^{\pm2.3}$	$72.0^{\pm1.4}$	$72.0^{\pm1.4}$	$72.0^{\pm1.4}$	$72.0^{\pm1.4}$	$72.0^{\pm1.4}$	$72.0^{\pm1.4}$
	Cross-lingual Fine-tuning		Orca 2 13B	$57.6^{\pm0.4}$	$69.5^{\pm0.6}$	$65.6^{\pm0.7}$	$61.2^{\pm0.3}$	$58.9^{\pm1.7}$	$33.8^{\pm0.8}$	$76.5^{\pm0.6}$	$72.0^{\pm0.2}$	$75.6^{\pm0.3}$	$66.3^{\pm0.2}$	$79.0^{\pm0.6}$	$67.5^{\pm1.1}$
		mT5	Without CD	$57.3^{\pm1.4}$	$59.2^{\pm0.5}$	$57.8^{\pm1.2}$	$57.1^{\pm0.9}$	$56.4^{\pm2.1}$	$44.4^{\pm1.4}$	$68.8^{\pm1.2}$	$65.4^{\pm1.0}$	$69.1^{\pm0.9}$	$63.3^{\pm0.5}$	$68.0^{\pm1.7}$	$59.4^{\pm1.1}$
			With CD	$\mathbf{62.4}^{\pm1.6}$	$\mathbf{69.3}^{\pm1.0}$	$\mathbf{61.1}^{\pm1.2}$	$\mathbf{60.8}^{\pm0.3}$	$\mathbf{63.7}^{\pm1.3}$	$\mathbf{48.9}^{\pm1.4}$	$68.7^{\pm0.9}$	$\mathbf{68.9}^{\pm1.1}$	$68.3^{\pm0.7}$	$\mathbf{69.1}^{\pm0.5}$	$\mathbf{70.6}^{\pm0.7}$	$\mathbf{61.5}^{\pm0.7}$
			Without CD*	$54.7^{\pm1.3}$	$58.2^{\pm0.9}$	$55.1^{\pm1.3}$	$53.8^{\pm1.9}$	$40.2^{\pm1.9}$	$38.2^{\pm2.7}$	$44.3^{\pm2.9}$	$44.7^{\pm1.4}$	$59.3^{\pm1.7}$	$54.2^{\pm1.9}$	$35.7^{\pm1.0}$	$42.9^{\pm1.0}$
			With CD*	$\mathbf{64.0}^{\pm1.0}$	$\mathbf{72.0}^{\pm1.0}$	$\mathbf{62.7}^{\pm1.4}$	$\mathbf{62.2}^{\pm1.0}$	$\mathbf{63.4}^{\pm0.9}$	$\mathbf{47.0}^{\pm0.6}$	$\mathbf{65.0}^{\pm1.7}$	$\mathbf{64.9}^{\pm1.3}$	$\mathbf{67.3}^{\pm0.8}$	$\mathbf{67.9}^{\pm1.6}$	$\mathbf{58.9}^{\pm2.7}$	$\mathbf{58.5}^{\pm1.7}$
		mBART	Without CD	$51.5^{\pm3.3}$	$61.1^{\pm2.6}$	$49.4^{\pm3.8}$	$51.6^{\pm2.7}$	$57.1^{\pm1.4}$	$31.6^{\pm3.9}$	$64.2^{\pm2.2}$	$64.5^{\pm1.6}$	$64.9^{\pm3.2}$	$59.5^{\pm1.8}$	$68.4^{\pm1.2}$	$57.6^{\pm0.8}$
			With CD	$48.8^{\pm2.7}$	$61.7^{\pm2.7}$	$49.2^{\pm4.1}$	$50.1^{\pm3.5}$	$57.8^{\pm1.8}$	$30.3^{\pm3.0}$	$65.0^{\pm1.3}$	$64.9^{\pm2.4}$	$63.4^{\pm1.1}$	$\mathbf{64.1}^{\pm1.3}$	$68.5^{\pm0.7}$	$56.7^{\pm0.8}$
			[44]	–	69.2	61.0	63.7	62.0	–	–	–	–	–	–	–
			[17]	–	67.1	56.4	59.0	56.8	46.2	–	–	–	–	–	–
			[18]	–	61.6	49.5	51.0	50.8	–	–	–	–	–	–	–
ACSA	ZS		ChatGPT	57.3	60.9	58.1	56.0	55.8	54.6	60.3	60.3	60.3	60.3	60.3	60.3
			Orca 2 13B	49.5	50.4	46.3	47.6	51.0	33.0	59.6	59.6	59.6	59.6	59.6	59.6
	FS		ChatGPT	61.6	67.6	62.2	63.9	64.9	58.8	64.7	64.7	64.7	64.7	64.7	64.7
			Orca 2 13B	63.4	67.0	55.7	61.1	59.8	45.6	65.2	65.2	65.2	65.2	65.2	65.2
	Monolingual Fine-tuning		Orca 2 13B	$75.4^{\pm0.6}$	$80.4^{\pm0.4}$	$78.1^{\pm0.4}$	$76.1^{\pm0.7}$	$80.2^{\pm0.6}$	$69.3^{\pm0.8}$	$84.3^{\pm0.9}$	$84.3^{\pm0.9}$	$84.3^{\pm0.9}$	$84.3^{\pm0.9}$	$84.3^{\pm0.9}$	$84.3^{\pm0.9}$
		mT5	Without CD	$76.6^{\pm0.2}$	$77.1^{\pm0.2}$	$69.2^{\pm0.8}$	$74.1^{\pm0.3}$	$78.0^{\pm0.8}$	$74.2^{\pm1.5}$	$78.6^{\pm1.3}$	$78.6^{\pm1.3}$	$78.6^{\pm1.3}$	$78.6^{\pm1.3}$	$78.6^{\pm1.3}$	$78.6^{\pm1.3}$
			With CD	$76.5^{\pm1.1}$	$77.4^{\pm0.6}$	$69.0^{\pm0.7}$	$74.3^{\pm0.9}$	$77.7^{\pm0.5}$	$74.4^{\pm4.6}$	$78.0^{\pm1.4}$	$78.0^{\pm1.4}$	$78.0^{\pm1.4}$	$78.0^{\pm1.4}$	$78.0^{\pm1.4}$	$78.0^{\pm1.4}$
			Without CD*	$75.9^{\pm0.5}$	$76.8^{\pm0.5}$	$69.5^{\pm0.7}$	$74.1^{\pm0.7}$	$77.4^{\pm0.5}$	$74.1^{\pm1.8}$	$79.0^{\pm0.9}$	$79.0^{\pm0.9}$	$79.0^{\pm0.9}$	$79.0^{\pm0.9}$	$79.0^{\pm0.9}$	$79.0^{\pm0.9}$
			With CD*	$76.5^{\pm0.7}$	$76.7^{\pm0.8}$	$68.9^{\pm0.6}$	$75.2^{\pm1.5}$	$77.6^{\pm0.5}$	$75.0^{\pm1.8}$	$78.3^{\pm1.0}$	$78.3^{\pm1.0}$	$78.3^{\pm1.0}$	$78.3^{\pm1.0}$	$78.3^{\pm1.0}$	$78.3^{\pm1.0}$
		mBART	Without CD	$72.6^{\pm1.3}$	$73.2^{\pm1.4}$	$65.0^{\pm0.9}$	$70.2^{\pm1.9}$	$73.3^{\pm1.2}$	$66.1^{\pm4.0}$	$74.7^{\pm1.7}$	$74.7^{\pm1.7}$	$74.7^{\pm1.7}$	$74.7^{\pm1.7}$	$74.7^{\pm1.7}$	$74.7^{\pm1.7}$
			With CD	$71.5^{\pm0.8}$	$73.3^{\pm0.8}$	$63.7^{\pm1.0}$	$68.3^{\pm2.9}$	$72.8^{\pm1.0}$	$64.4^{\pm5.1}$	$74.9^{\pm1.7}$	$74.9^{\pm1.7}$	$74.9^{\pm1.7}$	$74.9^{\pm1.7}$	$74.9^{\pm1.7}$	$74.9^{\pm1.7}$
	Cross-lingual Fine-tuning		Orca 2 13B	$70.6^{\pm0.1}$	$74.8^{\pm0.7}$	$73.8^{\pm0.5}$	$70.8^{\pm0.4}$	$75.6^{\pm1.6}$	$62.9^{\pm2.1}$	$79.6^{\pm0.2}$	$81.4^{\pm0.3}$	$83.8^{\pm0.7}$	$81.0^{\pm0.8}$	$82.3^{\pm0.4}$	$80.5^{\pm0.2}$
		mT5	Without CD	$68.0^{\pm1.1}$	$71.1^{\pm0.8}$	$63.7^{\pm0.7}$	$70.4^{\pm0.7}$	$71.5^{\pm1.1}$	$70.5^{\pm2.7}$	$73.8^{\pm1.1}$	$73.5^{\pm0.2}$	$74.2^{\pm0.6}$	$72.1^{\pm0.7}$	$74.0^{\pm0.7}$	$71.3^{\pm1.2}$
			With CD	$67.7^{\pm1.1}$	$71.0^{\pm1.1}$	$64.4^{\pm1.3}$	$70.4^{\pm1.1}$	$71.4^{\pm0.3}$	$70.7^{\pm3.0}$	$72.9^{\pm0.3}$	$73.5^{\pm0.5}$	$74.7^{\pm0.8}$	$72.9^{\pm1.1}$	$74.2^{\pm0.3}$	$71.8^{\pm0.8}$
			Without CD*	$67.4^{\pm1.1}$	$70.1^{\pm0.9}$	$64.8^{\pm1.9}$	$68.9^{\pm0.6}$	$71.1^{\pm0.9}$	$69.3^{\pm1.9}$	$76.0^{\pm0.5}$	$76.2^{\pm1.6}$	$76.6^{\pm1.4}$	$75.2^{\pm1.3}$	$75.9^{\pm1.2}$	$74.0^{\pm1.4}$
			With CD*	$68.2^{\pm0.4}$	$70.9^{\pm1.6}$	$65.2^{\pm1.1}$	$69.2^{\pm0.7}$	$71.3^{\pm1.1}$	$69.8^{\pm2.7}$	$75.2^{\pm0.8}$	$75.9^{\pm1.0}$	$76.7^{\pm1.1}$	$76.2^{\pm0.9}$	$74.5^{\pm1.1}$	$75.1^{\pm0.4}$
		mBART	Without CD	$55.2^{\pm1.4}$	$61.7^{\pm2.5}$	$53.7^{\pm2.4}$	$58.4^{\pm2.1}$	$65.6^{\pm1.9}$	$50.7^{\pm2.6}$	$66.3^{\pm2.3}$	$67.1^{\pm2.0}$	$66.7^{\pm1.6}$	$66.7^{\pm1.8}$	$65.0^{\pm1.3}$	$63.8^{\pm2.5}$
			With CD	$53.8^{\pm4.0}$	$63.0^{\pm2.3}$	$53.0^{\pm1.8}$	$59.4^{\pm2.0}$	$66.4^{\pm1.7}$	$49.3^{\pm2.7}$	$67.3^{\pm2.1}$	$68.3^{\pm1.8}$	$\mathbf{69.8}^{\pm1.3}$	$\mathbf{69.6}^{\pm0.5}$	$\mathbf{68.6}^{\pm0.8}$	$64.6^{\pm1.4}$
ACTE	ZS		ChatGPT	26.3	30.1	27.8	20.9	26.4	26.3	31.7	31.7	31.7	31.7	31.7	31.7
			Orca 2 13B	13.7	15.3	14.1	10.1	12.4	10.4	19.3	19.3	19.3	19.3	19.3	19.3
	FS		ChatGPT	45.8	50.7	41.3	42.4	43.5	38.8	45.5	45.5	45.5	45.5	45.5	45.5
			Orca 2 13B	35.1	41.3	31.3	31.5	30.8	27.3	44.8	44.8	44.8	44.8	44.8	44.8
	Monolingual Fine-tuning		Orca 2 13B	$72.6^{\pm0.6}$	$68.5^{\pm0.9}$	$62.9^{\pm3.5}$	$70.2^{\pm0.5}$	$70.4^{\pm0.2}$	$54.0^{\pm0.4}$	$80.7^{\pm1.0}$	$80.7^{\pm1.0}$	$80.7^{\pm1.0}$	$80.7^{\pm1.0}$	$80.7^{\pm1.0}$	$80.7^{\pm1.0}$
		mT5	Without CD	$73.5^{\pm0.8}$	$70.4^{\pm0.7}$	$63.7^{\pm0.8}$	$68.8^{\pm0.5}$	$73.2^{\pm0.5}$	$59.1^{\pm0.5}$	$75.7^{\pm0.4}$	$75.7^{\pm0.4}$	$75.7^{\pm0.4}$	$75.7^{\pm0.4}$	$75.7^{\pm0.4}$	$75.7^{\pm0.4}$
			With CD	$73.6^{\pm0.5}$	$69.9^{\pm0.4}$	$64.9^{\pm0.5}$	$62.9^{\pm0.5}$	$72.8^{\pm1.0}$	$60.4^{\pm2.1}$	$75.7^{\pm1.3}$	$75.7^{\pm1.3}$	$75.7^{\pm1.3}$	$75.7^{\pm1.3}$	$75.7^{\pm1.3}$	$75.7^{\pm1.3}$
			Without* CD	$73.1^{\pm0.5}$	$70.1^{\pm0.8}$	$64.7^{\pm0.4}$	$69.3^{\pm1.1}$	$73.0^{\pm0.4}$	$62.9^{\pm1.7}$	$75.9^{\pm0.7}$	$75.9^{\pm0.7}$	$75.9^{\pm0.7}$	$75.9^{\pm0.7}$	$75.9^{\pm0.7}$	$75.9^{\pm0.7}$
			With CD	$73.7^{\pm0.7}$	$69.7^{\pm0.7}$	$64.5^{\pm0.8}$	$64.0^{\pm1.9}$	$73.2^{\pm0.6}$	$60.7^{\pm0.8}$	$76.1^{\pm0.9}$	$76.1^{\pm0.9}$	$76.1^{\pm0.9}$	$76.1^{\pm0.9}$	$76.1^{\pm0.9}$	$76.1^{\pm0.9}$
		mBART	Without CD	$70.1^{\pm0.8}$	$66.4^{\pm1.6}$	$61.1^{\pm1.6}$	$64.1^{\pm1.2}$	$70.9^{\pm0.6}$	$56.8^{\pm2.2}$	$71.6^{\pm1.3}$	$71.6^{\pm1.3}$	$71.6^{\pm1.3}$	$71.6^{\pm1.3}$	$71.6^{\pm1.3}$	$71.6^{\pm1.3}$
			Without CD	$68.4^{\pm0.9}$	$66.8^{\pm1.5}$	$58.2^{\pm1.2}$	$58.0^{\pm1.2}$	$67.4^{\pm0.3}$	$55.3^{\pm1.5}$	$70.3^{\pm0.7}$	$70.3^{\pm0.7}$	$70.3^{\pm0.7}$	$70.3^{\pm0.7}$	$70.3^{\pm0.7}$	$70.3^{\pm0.7}$
	Cross-lingual Fine-tuning		Orca 2 13B	$54.6^{\pm0.4}$	$62.6^{\pm1.2}$	$57.1^{\pm2.4}$	$53.4^{\pm3.8}$	$59.3^{\pm1.1}$	$38.0^{\pm2.2}$	$73.8^{\pm0.2}$	$68.9^{\pm0.1}$	$71.8^{\pm0.1}$	$63.9^{\pm0.7}$	$73.1^{\pm0.4}$	$63.3^{\pm0.2}$
		mT5	Without CD	$54.3^{\pm1.6}$	$52.5^{\pm1.0}$	$55.8^{\pm0.7}$	$52.3^{\pm1.3}$	$55.0^{\pm2.7}$	$41.4^{\pm1.4}$	$55.8^{\pm2.0}$	$53.6^{\pm1.7}$	$61.2^{\pm1.3}$	$57.6^{\pm2.4}$	$52.8^{\pm1.0}$	$44.2^{\pm1.2}$
			With CD	$\mathbf{58.7}^{\pm1.0}$	$\mathbf{62.8}^{\pm1.4}$	$\mathbf{57.5}^{\pm0.3}$	$\mathbf{54.1}^{\pm0.2}$	$\mathbf{60.4}^{\pm0.9}$	$\mathbf{49.0}^{\pm0.9}$	$\mathbf{66.1}^{\pm1.3}$	$\mathbf{65.3}^{\pm1.2}$	$\mathbf{65.1}^{\pm0.8}$	$\mathbf{65.5}^{\pm0.7}$	$\mathbf{65.9}^{\pm0.7}$	$\mathbf{55.8}^{\pm0.6}$
			Without CD*	$51.1^{\pm1.1}$	$50.2^{\pm0.3}$	$52.5^{\pm1.5}$	$47.4^{\pm1.4}$	$37.9^{\pm1.1}$	$33.7^{\pm1.2}$	$41.5^{\pm2.1}$	$42.2^{\pm1.4}$	$54.3^{\pm3.1}$	$50.0^{\pm1.0}$	$31.9^{\pm0.9}$	$39.6^{\pm1.3}$
			With CD*	$\mathbf{59.5}^{\pm1.4}$	$\mathbf{63.3}^{\pm1.0}$	$\mathbf{58.2}^{\pm0.7}$	$\mathbf{53.8}^{\pm0.9}$	$\mathbf{60.6}^{\pm1.1}$	$\mathbf{44.8}^{\pm1.6}$	$\mathbf{60.5}^{\pm1.0}$	$\mathbf{61.7}^{\pm1.2}$	$\mathbf{63.3}^{\pm1.1}$	$\mathbf{65.0}^{\pm0.5}$	$\mathbf{53.9}^{\pm2.5}$	$\mathbf{56.4}^{\pm1.5}$
		mBART	Without CD	$48.6^{\pm2.7}$	$52.5^{\pm1.4}$	$49.3^{\pm1.5}$	$44.5^{\pm1.4}$	$53.8^{\pm1.5}$	$31.1^{\pm2.1}$	$58.4^{\pm3.5}$	$59.4^{\pm4.1}$	$62.0^{\pm3.5}$	$56.6^{\pm1.9}$	$56.7^{\pm3.4}$	$50.9^{\pm3.8}$
			With CD	$45.8^{\pm4.5}$	$\mathbf{54.8}^{\pm0.4}$	$49.2^{\pm0.6}$	$\mathbf{46.9}^{\pm0.9}$	$\mathbf{55.9}^{\pm0.2}$	$\mathbf{34.7}^{\pm1.1}$	$61.7^{\pm2.1}$	$59.4^{\pm2.3}$	$59.5^{\pm0.7}$	$\mathbf{59.6}^{\pm1.0}$	$\mathbf{61.3}^{\pm1.1}$	$\mathbf{55.7}^{\pm0.7}$

the majority of cases. While mBART also benefits from constrained decoding, it does so to a lesser extent than mT5.

Constrained decoding proves critical for tasks that involve aspect term prediction (E2E-ABSA and ACTE) in cross-lingual settings. It effectively addresses the issue of models generating aspect terms in the source language rather than the target language – a problem that also occurs in the ATE task. This issue frequently arises in zero-shot cross-lingual transfer and is particularly problematic for multi-tasking models, where constrained decoding often leads to over 10% absolute improvements in F1 score, bringing their performance on par with task-specific models.

The ACSA task, in contrast, does not benefit from constrained decoding. Since ACSA does not involve aspect term prediction, the primary source of cross-lingual transfer errors addressed by constrained decoding is absent. This aligns with findings from simpler tasks like ACD (see Section 5.1). Surprisingly, even multi-tasking models see no improvement from constrained decoding in ACSA. Nevertheless, the general performance pattern holds: ChatGPT leads in zero-shot and few-shot settings, and fine-tuned Orca 2 13B performs best in most cases. The mT5 model shows strong results in some languages, particularly Turkish and Czech. ACSA is comparatively simpler due to the absence of aspect terms, as the label space for aspect categories and sentiment polarities is more limited than the open-ended label space for aspect terms.

We improve on previous cross-lingual E2E-ABSA results [17, 18, 44], except for the en→nl combination. However, this comparison is not straightforward. Prior works use encoder-only models, are restricted to one sentiment polarity per aspect term, and do not predict *"NULL"* (implicit) aspect terms. In contrast, our method supports multiple sentiment polarities per aspect and includes implicit aspect prediction, making the task more complex. This also changes the number of tuples to be predicted: for instance, our English test set contains 859 tuples, compared to 612 in [44], where implicit aspects are excluded and sentiment polarities merged. Despite these differences, our constrained decoding approach achieves comparable or superior results in zero-shot cross-lingual settings.

Importantly, our approach does not rely on external translation systems. Previous methods depend on translation tools and are subject to translation quality. In contrast, constrained decoding offers a language-agnostic and implementation-friendly solution. Given its simplicity and effectiveness, our method provides a practical alternative for cross-lingual ABSA tasks.

5.3 TASD Results

This section focuses on the TASD task in more detail, given that this task offers the most comprehensive analysis of reviews involving the simultaneous prediction of three sentiment elements. Table 6 presents the results of the TASD task.

The monolingual results for the TASD task show that ChatGPT consistently performs the best across different settings. However, zero-shot performance is generally weak, with F1 scores often falling below 10% for non-English languages for models other than ChatGPT, and under 20% for English. Even ChatGPT does not exceed 30% for any language in zero-shot scenarios. Few-shot settings yield better results but remain under 50% for all languages while often not

Table 6. Monolingual (for target languages) and cross-lingual F1 scores for the TASD task. **Bold** indicates significant improvements with constrained decoding (CD) over without. The best cross-lingual result per task and language pair is underlined. Asterisks (*) denote multi-tasking models. ZS and FS stand for zero-shot and few-shot (10 examples).

Setup	Model		En→cs	En→es	En→fr	En→nl	En→ru	En→tr	Cs→en	Es→en	Fr→en	Nl→en	Ru→en	Tr→en
Monolingual ZS	ChatGPT		25.4	27.5	25.1	18.6	22.2	19.2	27.7	27.7	27.7	27.7	27.7	27.7
	Orca 2 13B		11.2	12.8	10.5	9.1	9.3	4.6	18.7	18.7	18.7	18.7	18.7	18.7
Monolingual FS	ChatGPT		42.6	47.7	37.1	40.0	37.0	35.8	42.4	42.4	42.4	42.4	42.4	42.4
	Orca 2 13B		32.6	44.2	30.0	26.5	28.3	24.2	46.3	46.3	46.3	46.3	46.3	46.3
Monolingual Fine-tuning	Orca 2 13B		$65.6^{\pm 0.4}$	$66.2^{\pm 0.4}$	$63.0^{\pm 1.0}$	$64.5^{\pm 1.8}$	$64.5^{\pm 1.0}$	$48.8^{\pm 0.9}$	$77.3^{\pm 1.8}$	$77.3^{\pm 1.8}$	$77.3^{\pm 1.8}$	$77.3^{\pm 1.8}$	$77.3^{\pm 1.8}$	$77.3^{\pm 1.8}$
	mT5	Without CD	$66.9^{\pm 0.3}$	$65.8^{\pm 0.4}$	$59.0^{\pm 0.6}$	$62.9^{\pm 1.4}$	$67.0^{\pm 0.9}$	$54.1^{\pm 3.0}$	$71.4^{\pm 0.9}$	$71.4^{\pm 0.9}$	$71.4^{\pm 0.9}$	$71.4^{\pm 0.9}$	$71.4^{\pm 0.9}$	$71.4^{\pm 0.9}$
		With CD	$67.1^{\pm 1.3}$	$66.2^{\pm 0.5}$	$58.9^{\pm 1.1}$	$57.6^{\pm 0.5}$	$66.4^{\pm 0.4}$	$53.9^{\pm 1.5}$	$70.4^{\pm 0.8}$	$70.4^{\pm 0.8}$	$70.4^{\pm 0.8}$	$70.4^{\pm 0.8}$	$70.4^{\pm 0.8}$	$70.4^{\pm 0.8}$
		Without CD*	$66.3^{\pm 0.5}$	$65.6^{\pm 0.5}$	$57.9^{\pm 0.6}$	$62.8^{\pm 0.9}$	$65.7^{\pm 0.8}$	$58.0^{\pm 0.8}$	$70.5^{\pm 0.3}$	$70.5^{\pm 0.3}$	$70.5^{\pm 0.3}$	$70.5^{\pm 0.3}$	$70.5^{\pm 0.3}$	$70.5^{\pm 0.3}$
		With CD*	$67.3^{\pm 0.5}$	$64.6^{\pm 0.4}$	$58.0^{\pm 0.8}$	$58.0^{\pm 1.7}$	$66.5^{\pm 0.4}$	$57.0^{\pm 0.5}$	$70.8^{\pm 0.8}$	$70.8^{\pm 0.8}$	$70.8^{\pm 0.8}$	$70.8^{\pm 0.8}$	$70.8^{\pm 0.8}$	$70.8^{\pm 0.8}$
	mBART	Without CD	$62.6^{\pm 0.7}$	$62.9^{\pm 1.2}$	$54.8^{\pm 0.9}$	$57.6^{\pm 0.9}$	$62.6^{\pm 0.7}$	$49.3^{\pm 3.1}$	$66.1^{\pm 1.4}$	$66.1^{\pm 1.4}$	$66.1^{\pm 1.4}$	$66.1^{\pm 1.4}$	$66.1^{\pm 1.4}$	$66.1^{\pm 1.4}$
		With CD	$61.9^{\pm 1.6}$	$61.5^{\pm 1.4}$	$52.4^{\pm 0.6}$	$52.1^{\pm 1.0}$	$60.1^{\pm 1.9}$	$47.6^{\pm 2.7}$	$64.8^{\pm 1.4}$	$64.8^{\pm 1.4}$	$64.8^{\pm 1.4}$	$64.8^{\pm 1.4}$	$64.8^{\pm 1.4}$	$64.8^{\pm 1.4}$
Cross-lingual Fine-tuning	Orca 2 13B		$49.7^{\pm 0.5}$	$\underline{58.0}^{\pm 1.1}$	$\underline{56.1}^{\pm 1.0}$	$50.2^{\pm 1.1}$	$\underline{55.6}^{\pm 1.4}$	$31.4^{\pm 1.6}$	$\underline{71.7}^{\pm 0.2}$	$\underline{67.2}^{\pm 0.3}$	$\underline{68.3}^{\pm 1.6}$	$59.1^{\pm 1.4}$	$\underline{69.2}^{\pm 0.6}$	$\underline{60.4}^{\pm 0.1}$
	mT5	Without CD	$50.2^{\pm 0.9}$	$48.3^{\pm 0.5}$	$50.4^{\pm 1.4}$	$47.7^{\pm 1.1}$	$48.6^{\pm 2.0}$	$39.1^{\pm 3.6}$	$54.9^{\pm 1.9}$	$50.7^{\pm 1.7}$	$57.4^{\pm 1.0}$	$53.3^{\pm 0.8}$	$50.1^{\pm 2.0}$	$43.2^{\pm 0.6}$
		With CD	$\mathbf{53.3}^{\pm 1.5}$	$\mathbf{57.6}^{\pm 0.6}$	$50.4^{\pm 0.8}$	$\underline{\mathbf{50.4}}^{\pm 1.3}$	$\mathbf{54.9}^{\pm 2.0}$	$\underline{\mathbf{43.8}}^{\pm 0.8}$	$\mathbf{60.1}^{\pm 0.8}$	$\mathbf{59.7}^{\pm 0.4}$	$\mathbf{59.9}^{\pm 0.6}$	$\underline{\mathbf{59.9}}^{\pm 0.3}$	$\mathbf{61.8}^{\pm 0.8}$	$\mathbf{52.3}^{\pm 0.5}$
		Without CD*	$45.4^{\pm 1.5}$	$46.2^{\pm 0.7}$	$44.8^{\pm 1.6}$	$43.4^{\pm 0.8}$	$33.6^{\pm 1.3}$	$24.7^{\pm 4.0}$	$36.5^{\pm 2.1}$	$37.7^{\pm 0.9}$	$49.8^{\pm 2.3}$	$44.7^{\pm 0.9}$	$28.7^{\pm 0.7}$	$35.3^{\pm 0.8}$
		With CD*	$\underline{\mathbf{53.4}}^{\pm 1.5}$	$\mathbf{57.8}^{\pm 0.5}$	$\mathbf{50.8}^{\pm 1.1}$	$\mathbf{49.3}^{\pm 0.6}$	$\mathbf{53.6}^{\pm 0.9}$	$\mathbf{40.2}^{\pm 1.6}$	$\mathbf{55.7}^{\pm 1.3}$	$\mathbf{56.1}^{\pm 0.6}$	$\mathbf{58.0}^{\pm 0.8}$	$\mathbf{58.0}^{\pm 1.1}$	$\mathbf{50.2}^{\pm 2.7}$	$\mathbf{50.5}^{\pm 1.3}$
	mBART	Without CD	$40.4^{\pm 3.0}$	$47.6^{\pm 1.9}$	$39.6^{\pm 0.8}$	$39.1^{\pm 0.9}$	$48.5^{\pm 1.1}$	$23.5^{\pm 2.6}$	$54.3^{\pm 0.8}$	$55.4^{\pm 2.9}$	$56.0^{\pm 2.3}$	$50.4^{\pm 1.4}$	$54.5^{\pm 3.2}$	$49.1^{\pm 0.8}$
		With CD	$39.3^{\pm 1.0}$	$\mathbf{51.1}^{\pm 1.2}$	$39.9^{\pm 0.6}$	$38.9^{\pm 0.9}$	$\mathbf{50.5}^{\pm 0.7}$	$\mathbf{27.3}^{\pm 1.1}$	$54.2^{\pm 1.7}$	$55.1^{\pm 1.9}$	$55.0^{\pm 1.3}$	$\mathbf{54.4}^{\pm 1.1}$	$\mathbf{59.0}^{\pm 0.9}$	$48.8^{\pm 0.8}$

reaching even 30%, highlighting the inherent difficulty of the TASD task. Fine-tuning improves performance significantly, with fine-tuned versions of Orca 2 and mT5 achieving the best monolingual results.

In cross-lingual settings, similar patterns emerge as observed in other tasks. Constrained decoding significantly enhances performance. This improvement is particularly pronounced because the task involves predicting aspect terms, which may be generated in the source language rather than the target language, an issue mitigated by constrained decoding. For multi-tasking models, constrained decoding provides even more substantial benefits. Generally, multi-tasking models perform on par with specialized models. Orca 2 usually achieves the best results when English is the target language. When English is the source language, both mT5 and Orca 2 perform well, except for Turkish as the target language, where mT5 outperforms Orca 2 by 12%. Compared to monolingual results, cross-lingual results are often around 10% worse.

Table 7 shows the results for the TASD task, comparing the performance of the mT5 model with and without constrained decoding alongside the Orca 2 13B model across all language combinations. The results indicate that constrained decoding generally enhances performance across most language pairs. On average, constrained decoding improves the results by almost 5%. Notably, the mT5 model consistently outperforms the Orca 2 13B model for some target languages, especially for Turkish and Czech. In contrast, the Orca 2 model demonstrates superiority when French and English are the target languages. On average, the Orca 2 model outperforms the mT5 model with constrained decoding by 0.6% (0.4% in cross-lingual settings), which is primarily due to English as the target language, where it outperforms the mT5 model by more than 7% on average. No-

tably, the mT5 model has over ten times fewer parameters, yet its performance is nearly equivalent to that of the Orca 2 model. Overall, English emerges as the most favourable target language overall, yielding the highest scores across multiple source languages. On the other hand, Turkish consistently yields weaker results, suggesting potential challenges due to limited data availability or greater linguistic divergence. Turkish is the only language in the study that does not come from an Indo-European family.

Table 7. Comparison of mT5 models with and without constrained decoding (CD), and Orca 2 13B across various language combinations for the TASD task. Results are reported in terms of F1 scores, with target languages in columns and source languages in rows. AVG* excludes monolingual results, i.e. results where the source and target languages are the same.

			Cs	**En**	**Es**	**Fr**	**Nl**	**Ru**	**Tr**	**AVG**	**AVG***
Cs	mT5	Without CD	66.9	54.9	50.1	41.9	45.1	44.4	39.2	48.9	45.9
		With CD	67.1	60.1	57.4	46.9	46.0	55.1	41.8	53.5	51.2
	Orca 2 13B		65.6	71.7	60.7	50.8	46.3	55.5	38.4	55.6	53.9
En	mT5	Without CD	50.2	71.4	48.3	50.4	47.7	48.6	39.1	50.8	47.4
		With CD	53.3	70.4	57.6	50.4	50.4	54.9	43.8	54.4	51.7
	Orca 2 13B		49.7	77.3	58.0	56.1	50.2	55.6	31.4	54.0	50.1
Es	mT5	Without CD	50.8	50.7	65.8	43.2	43.1	49.8	38.5	48.9	46.0
		With CD	55.5	59.7	66.2	49.9	45.1	52.1	43.3	53.1	50.9
	Orca 2 13B		52.1	67.2	66.2	54.2	49.0	51.4	30.9	53.0	50.8
Fr	mT5	Without CD	47.8	57.4	46.6	59.0	47.4	46.1	32.5	48.1	46.3
		With CD	53.3	59.9	57.6	58.9	48.3	50.5	35.1	52.0	50.8
	Orca 2 13B		51.5	68.3	60.1	63.0	50.2	52.3	30.7	53.7	52.2
Nl	mT5	Without CD	43.7	53.3	48.8	43.2	60.3	43.6	29.0	46.0	43.6
		With CD	49.1	59.9	57.1	48.0	58.5	46.1	34.5	50.5	49.1
	Orca 2 13B		42.4	59.1	53.2	49.1	63.3	41.1	25.1	47.6	45.0
Ru	mT5	Without CD	41.6	51.3	52.5	44.5	41.6	67.0	37.0	47.9	44.8
		With CD	53.8	59.4	55.4	46.8	47.4	66.4	40.4	52.8	50.5
	Orca 2 13B		54.7	67.7	57.0	50.9	49.2	64.5	29.8	53.4	51.5
Tr	mT5	Without CD	45.2	43.2	37.4	30.7	36.1	42.4	54.1	41.3	39.2
		With CD	47.4	52.3	40.9	32.2	39.1	43.7	53.9	44.2	42.6
	Orca 2 13B		49.2	60.4	50.8	36.4	36.0	47.4	48.8	47.0	46.7
AVG	mT5	Without CD	49.5	54.6	49.9	44.7	45.9	48.9	38.5	47.4	–
		With CD	54.2	60.2	56.0	47.6	47.8	52.7	41.8	51.5	–
	Orca 2 13B		52.2	67.4	58.0	51.5	49.2	52.5	33.6	52.1	–
AVG*	mT5	Without CD	46.6	51.8	47.3	42.3	43.5	45.8	35.9	–	44.7
		With CD	52.1	58.6	54.3	45.7	46.1	50.4	39.8	–	49.6
	Orca 2 13B		49.9	65.7	56.6	49.6	46.8	50.6	31.1	–	50.0

5.4 Detailed LLM Results

Table 8 shows the monolingual results for LLMs. The best results are generally achieved with the Orca 2 13B and LLaMA 3 8B models in zero-shot and few-shot settings. The LLaMA 3 is a more modern model, which may contribute to its good performance despite having fewer parameters than the LLaMA 2 13B and Orca 2 13B. Orca 2 is based on the LLaMA 2 with some improvements, which could explain why Orca 2 models outperform the same-sized LLaMA 2 models. The Orca 2 13B model generally achieves the best results with fine-tuning. Interestingly, the Orca 2 7B model often outperforms the larger LLaMA 2 13B model with fine-tuning, suggesting that smaller but more advanced models can outperform larger, less sophisticated ones.

Table 8. Zero-shot, few-shot and monolingual results for compound ABSA tasks with five different LLMs. The best zero-shot, few-shot and fine-tuning results for each combination of task and language are in **bold**.

Model	Settings	Task	En	Cs	Es	Fr	Nl	Ru	Tr
LLaMA 2 13B	Zero- / few-shot	ACSA	43.3 / 60.6	32.7 / 49.0	37.8 / 60.2	34.0 / 43.3	36.5 / 52.3	35.5 / 54.8	24.5 / 34.9
		E2E	34.7 / 46.5	19.0 / **44.0**	26.2 / 52.3	24.2 / 38.2	18.7 / 39.7	13.9 / 34.0	13.8 / 23.3
		ACTE	17.9 / 34.7	9.6 / 37.7	8.6 / 42.7	8.8 / 32.1	6.8 / 25.0	5.4 / 29.4	4.4 / 24.0
		TASD	17.5 / 33.1	10.2 / **35.7**	11.1 / 34.6	8.4 / 24.1	8.0 / 23.9	6.1 / 24.6	**6.9** / 18.7
	Fine-tuning	ACSA	$82.5^{\pm 0.9}$	$73.2^{\pm 1.7}$	$78.4^{\pm 1.3}$	$69.3^{\pm 3.2}$	$74.7^{\pm 1.7}$	$77.2^{\pm 1.7}$	$67.0^{\pm 2.2}$
		E2E	$77.8^{\pm 4.1}$	$69.4^{\pm 0.7}$	$69.4^{\pm 1.2}$	$65.8^{\pm 0.7}$	$71.3^{\pm 1.8}$	$67.3^{\pm 1.4}$	$48.0^{\pm 2.8}$
		ACTE	$75.6^{\pm 1.2}$	$67.0^{\pm 0.9}$	$65.4^{\pm 3.1}$	$58.2^{\pm 5.1}$	$66.0^{\pm 1.3}$	$68.8^{\pm 1.3}$	$53.7^{\pm 1.8}$
		TASD	$72.0^{\pm 0.8}$	$60.7^{\pm 0.9}$	$61.6^{\pm 1.8}$	$59.3^{\pm 0.8}$	$60.3^{\pm 0.9}$	$62.1^{\pm 1.8}$	$41.6^{\pm 1.6}$
LLaMA 2 7B	Zero- / few-shot	ACSA	13.3 / 57.3	14.1 / 41.0	17.8 / 50.7	14.1 / 42.8	15.3 / 14.4	14.4 / 37.0	19.5 / 38.7
		E2E	24.4 / 40.9	12.2 / 35.0	15.2 / 42.9	13.8 / 32.4	10.7 / 4.1	6.5 / 26.2	11.4 / 18.7
		ACTE	11.7 / 35.1	6.0 / 29.1	5.6 / 30.0	5.8 / 28.0	2.7 / 10.1	1.7 / 14.3	4.3 / 17.5
		TASD	7.9 / 30.6	3.6 / 25.9	4.8 / 23.5	3.1 / 13.9	2.0 / 15.2	1.0 / 11.4	2.7 / 11.7
	Fine-tuning	ACSA	$82.1^{\pm 1.2}$	$73.3^{\pm 0.7}$	$74.3^{\pm 3.8}$	$66.6^{\pm 0.7}$	$71.4^{\pm 1.0}$	$74.8^{\pm 3.1}$	$61.0^{\pm 3.3}$
		E2E	$75.5^{\pm 2.4}$	$66.5^{\pm 0.4}$	$67.8^{\pm 3.9}$	$60.6^{\pm 0.6}$	$64.8^{\pm 1.2}$	$65.3^{\pm 1.5}$	$44.4^{\pm 1.1}$
		ACTE	$73.4^{\pm 1.1}$	$66.1^{\pm 0.7}$	$65.3^{\pm 0.9}$	$56.4^{\pm 4.4}$	$62.7^{\pm 1.7}$	$66.7^{\pm 1.2}$	$46.9^{\pm 1.8}$
		TASD	$70.3^{\pm 1.7}$	$60.2^{\pm 0.8}$	$58.1^{\pm 4.5}$	$54.0^{\pm 0.7}$	$55.5^{\pm 0.3}$	$58.8^{\pm 1.0}$	$36.9^{\pm 1.9}$
LLaMA 3 8B	Zero- / few-shot	ACSA	58.9 / 62.5	**53.3** / 59.0	**54.5** / **67.2**	**51.3** / 37.7	**48.7** / 57.7	50.5 / **60.3**	**44.7** / **57.3**
		E2E	40.9 / 55.7	**34.5** / 42.3	**40.6** / **55.5**	**27.5** / 41.4	**30.9** / 40.9	**28.8** / **42.0**	**20.0** / **45.5**
		ACTE	18.4 / **49.8**	**16.7** / **39.4**	**19.3** / **50.3**	12.5 / **35.2**	**13.5** / 31.6	7.9 / **37.1**	7.3 / **42.0**
		TASD	10.9 / 43.7	7.9 / 34.6	**13.5** / **49.0**	4.8 / 27.1	8.3 / **32.8**	3.6 / **35.5**	5.4 / **38.3**
	Fine-tuning	ACSA	$81.1^{\pm 2.1}$	$70.8^{\pm 2.1}$	$73.2^{\pm 0.5}$	$69.5^{\pm 3.7}$	$71.5^{\pm 3.5}$	$71.6^{\pm 2.3}$	$65.5^{\pm 1.6}$
		E2E	$71.4^{\pm 2.8}$	$63.0^{\pm 1.1}$	$70.0^{\pm 2.0}$	$63.1^{\pm 1.8}$	$66.0^{\pm 1.2}$	$60.7^{\pm 2.3}$	$48.6^{\pm 2.0}$
		ACTE	$69.2^{\pm 1.7}$	$63.1^{\pm 0.6}$	$59.8^{\pm 2.9}$	$54.9^{\pm 1.2}$	$58.1^{\pm 4.3}$	$62.1^{\pm 1.7}$	$46.3^{\pm 2.9}$
		TASD	$62.5^{\pm 2.4}$	$56.8^{\pm 1.4}$	$57.2^{\pm 1.7}$	$48.2^{\pm 2.4}$	$55.4^{\pm 2.8}$	$53.7^{\pm 2.9}$	$39.1^{\pm 3.2}$
Orca 2 13B	Zero- / few-shot	ACSA	**59.6** / **65.2**	49.5 / **63.4**	50.4 / 67.0	46.3 / **55.7**	47.6 / **61.1**	**51.0** / 59.8	33.0 / 45.6
		E2E	**42.2** / **62.4**	22.7 / 41.0	27.4 / 50.6	22.5 / **43.9**	18.3 / **46.1**	20.0 / 38.7	19.2 / 32.0
		ACTE	**19.3** / 44.8	13.7 / 35.1	15.3 / 41.3	**14.1** / 31.3	10.1 / 31.5	**12.4** / 30.8	**10.4** / 27.3
		TASD	**18.7** / **46.3**	**11.2** / 32.6	12.8 / 44.2	**10.5** / **30.0**	**9.1** / 26.5	**9.3** / 28.3	4.6 / 24.2
	Fine-tuning	ACSA	$\mathbf{84.3}^{\pm 0.9}$	$\mathbf{75.4}^{\pm 0.6}$	$\mathbf{80.4}^{\pm 0.4}$	$\mathbf{78.1}^{\pm 0.4}$	$\mathbf{76.1}^{\pm 0.7}$	$\mathbf{80.2}^{\pm 0.6}$	$\mathbf{69.3}^{\pm 0.8}$
		E2E	$\mathbf{82.3}^{\pm 0.6}$	$\mathbf{71.8}^{\pm 1.1}$	$\mathbf{74.8}^{\pm 1.0}$	$\mathbf{69.8}^{\pm 2.3}$	$\mathbf{76.2}^{\pm 0.3}$	$\mathbf{72.2}^{\pm 1.2}$	$\mathbf{52.2}^{\pm 1.6}$
		ACTE	$\mathbf{80.7}^{\pm 1.0}$	$\mathbf{72.6}^{\pm 0.6}$	$68.5^{\pm 0.9}$	$62.9^{\pm 3.5}$	$\mathbf{70.2}^{\pm 0.5}$	$70.4^{\pm 0.2}$	$\mathbf{54.0}^{\pm 0.4}$
		TASD	$\mathbf{77.3}^{\pm 1.8}$	$\mathbf{65.6}^{\pm 0.4}$	$\mathbf{66.2}^{\pm 0.4}$	$\mathbf{63.0}^{\pm 1.0}$	$\mathbf{64.5}^{\pm 1.8}$	$\mathbf{64.5}^{\pm 1.0}$	$\mathbf{48.8}^{\pm 0.9}$
Orca 2 7B	Zero- / few-shot	ACSA	38.8 / 59.7	35.7 / 54.3	30.6 / 59.0	22.9 / 47.9	23.5 / 45.9	22.3 / 51.7	18.9 / 32.3
		E2E	36.5 / 49.6	16.1 / 33.0	24.4 / 46.3	22.1 / 30.6	13.7 / 32.9	10.2 / 26.9	7.4 / 24.5
		ACTE	16.6 / 35.2	7.6 / 28.4	9.4 / 36.3	10.6 / 24.8	6.5 / **35.8**	3.5 / 22.8	9.8 / 21.8
		TASD	13.7 / 39.5	4.6 / 26.7	7.7 / 38.1	7.4 / 25.1	3.1 / 32.0	3.0 / 18.4	2.2 / 16.6
	Fine-tuning	ACSA	$83.5^{\pm 0.6}$	$74.3^{\pm 0.4}$	$77.4^{\pm 0.5}$	$72.9^{\pm 0.6}$	$75.9^{\pm 0.8}$	$78.1^{\pm 0.6}$	$66.2^{\pm 0.8}$
		E2E	$81.4^{\pm 0.5}$	$70.8^{\pm 0.2}$	$74.5^{\pm 0.1}$	$67.7^{\pm 0.1}$	$64.1^{\pm 0.3}$	$69.3^{\pm 0.2}$	$47.8^{\pm 1.0}$
		ACTE	$79.9^{\pm 1.3}$	$70.0^{\pm 0.3}$	$\mathbf{68.6}^{\pm 0.2}$	$\mathbf{65.1}^{\pm 0.2}$	$65.5^{\pm 0.2}$	$\mathbf{70.5}^{\pm 0.1}$	$51.3^{\pm 0.6}$
		TASD	$74.8^{\pm 0.3}$	$63.4^{\pm 0.3}$	$63.4^{\pm 0.3}$	$58.8^{\pm 3.9}$	$60.3^{\pm 1.6}$	$62.1^{\pm 0.3}$	$42.0^{\pm 0.8}$

Despite strong zero-shot and few-shot results with LLaMA 3, its fine-tuned version often underperforms compared to other models. This may be due to suboptimal fine-tuning hyperparameters, which could be less compatible with the newer architecture. Overall, adding few-shot examples substantially improves performance, while the best results are achieved with fine-tuning.

Table 9 presents the cross-lingual results with LLMs. Similar to the monolingual results with fine-tuning, the Orca 2 13B model performs best in most cases across all tasks and language combinations.

Table 9. F1 scores for compound ABSA tasks with five different LLMs in cross-lingual fine-tuning settings. The best results for each task and language combination is in **bold**.

Model	Task	En→cs	En→es	En→fr	En→nl	En→ru	En→tr	Cs→en	Es→en	Fr→en	Nl→en	Ru→en	Tr→en
LLaMA 2 13B	ACSA	$66.6^{\pm 2.0}$	$72.9^{\pm 1.5}$	$69.8^{\pm 1.2}$	$\mathbf{71.8}^{\pm 1.4}$	$74.0^{\pm 1.2}$	$52.0^{\pm 3.9}$	$76.9^{\pm 1.7}$	$79.7^{\pm 2.4}$	$77.5^{\pm 3.9}$	$77.9^{\pm 0.5}$	$81.0^{\pm 0.6}$	$78.8^{\pm 1.1}$
	E2E	$48.0^{\pm 2.8}$	$58.0^{\pm 0.7}$	$52.3^{\pm 5.3}$	$55.4^{\pm 3.7}$	$53.3^{\pm 0.9}$	$29.5^{\pm 4.9}$	$69.0^{\pm 0.8}$	$62.7^{\pm 1.7}$	$62.6^{\pm 5.7}$	$56.9^{\pm 1.0}$	$68.8^{\pm 1.6}$	$57.1^{\pm 1.0}$
	ACTE	$44.5^{\pm 0.8}$	$52.1^{\pm 2.6}$	$49.4^{\pm 2.1}$	$47.7^{\pm 2.0}$	$50.6^{\pm 3.1}$	$31.4^{\pm 2.5}$	$61.5^{\pm 3.8}$	$59.6^{\pm 1.3}$	$58.8^{\pm 4.6}$	$57.5^{\pm 1.4}$	$54.7^{\pm 1.7}$	$57.7^{\pm 0.9}$
	TASD	$40.4^{\pm 1.9}$	$49.3^{\pm 1.4}$	$45.3^{\pm 2.2}$	$41.8^{\pm 2.0}$	$43.7^{\pm 3.4}$	$25.0^{\pm 2.6}$	$62.4^{\pm 0.9}$	$55.9^{\pm 0.8}$	$61.6^{\pm 2.8}$	$51.8^{\pm 0.6}$	$53.5^{\pm 3.6}$	$49.8^{\pm 1.7}$
LLaMA 2 7B	ACSA	$64.7^{\pm 2.2}$	$71.6^{\pm 1.3}$	$68.0^{\pm 1.2}$	$65.4^{\pm 1.8}$	$71.6^{\pm 0.6}$	$46.1^{\pm 2.0}$	$79.5^{\pm 0.4}$	$74.3^{\pm 1.4}$	$73.6^{\pm 1.4}$	$77.5^{\pm 0.8}$	$78.0^{\pm 1.4}$	$73.7^{\pm 1.8}$
	E2E	$42.7^{\pm 2.5}$	$54.0^{\pm 1.2}$	$49.0^{\pm 0.6}$	$49.3^{\pm 1.5}$	$43.5^{\pm 1.3}$	$22.7^{\pm 5.2}$	$58.0^{\pm 1.5}$	$53.7^{\pm 1.1}$	$55.0^{\pm 3.8}$	$54.5^{\pm 0.7}$	$63.0^{\pm 1.0}$	$53.7^{\pm 1.2}$
	ACTE	$37.4^{\pm 1.0}$	$42.0^{\pm 2.8}$	$43.7^{\pm 1.4}$	$42.5^{\pm 0.9}$	$41.5^{\pm 2.3}$	$25.4^{\pm 2.1}$	$60.9^{\pm 1.6}$	$54.6^{\pm 5.0}$	$56.9^{\pm 3.2}$	$54.3^{\pm 1.6}$	$49.8^{\pm 3.4}$	$47.2^{\pm 0.7}$
	TASD	$34.6^{\pm 1.0}$	$36.6^{\pm 2.7}$	$36.9^{\pm 3.0}$	$36.7^{\pm 1.3}$	$36.3^{\pm 2.0}$	$17.3^{\pm 1.8}$	$56.7^{\pm 0.3}$	$50.9^{\pm 1.5}$	$55.4^{\pm 2.0}$	$49.5^{\pm 2.6}$	$45.8^{\pm 1.7}$	$47.0^{\pm 1.3}$
LLaMA 3 8B	ACSA	$65.2^{\pm 1.3}$	$71.4^{\pm 1.3}$	$66.8^{\pm 3.3}$	$67.5^{\pm 3.5}$	$71.1^{\pm 2.6}$	$61.1^{\pm 3.0}$	$79.7^{\pm 1.6}$	$75.6^{\pm 2.0}$	$79.5^{\pm 2.3}$	$76.8^{\pm 2.8}$	$77.8^{\pm 2.5}$	$77.3^{\pm 1.1}$
	E2E	$37.9^{\pm 3.0}$	$47.2^{\pm 5.0}$	$41.7^{\pm 3.0}$	$43.3^{\pm 3.5}$	$53.1^{\pm 2.2}$	$26.6^{\pm 9.7}$	$63.3^{\pm 5.8}$	$60.6^{\pm 3.6}$	$61.5^{\pm 3.7}$	$55.5^{\pm 1.6}$	$64.2^{\pm 2.8}$	$62.0^{\pm 1.7}$
	ACTE	$45.3^{\pm 4.7}$	$44.8^{\pm 8.5}$	$38.1^{\pm 3.5}$	$40.0^{\pm 6.0}$	$50.6^{\pm 4.1}$	$29.9^{\pm 7.1}$	$60.1^{\pm 3.2}$	$54.5^{\pm 2.7}$	$51.0^{\pm 4.2}$	$50.7^{\pm 1.0}$	$48.5^{\pm 4.6}$	$51.3^{\pm 2.8}$
	TASD	$33.3^{\pm 2.9}$	$39.5^{\pm 7.5}$	$31.5^{\pm 3.2}$	$29.8^{\pm 5.7}$	$46.5^{\pm 2.6}$	$22.8^{\pm 7.2}$	$55.8^{\pm 3.6}$	$54.0^{\pm 3.1}$	$49.5^{\pm 3.9}$	$47.1^{\pm 0.7}$	$42.1^{\pm 6.5}$	$47.6^{\pm 3.5}$
Orca 2 13B	ACSA	$\mathbf{70.6}^{\pm 0.1}$	$\mathbf{74.8}^{\pm 0.7}$	$\mathbf{73.8}^{\pm 0.5}$	$70.8^{\pm 0.4}$	$\mathbf{75.6}^{\pm 1.6}$	$\mathbf{62.9}^{\pm 2.1}$	$79.6^{\pm 0.2}$	$\mathbf{81.4}^{\pm 0.3}$	$\mathbf{83.8}^{\pm 0.7}$	$\mathbf{81.0}^{\pm 0.8}$	$\mathbf{82.3}^{\pm 0.4}$	$\mathbf{80.5}^{\pm 0.2}$
	E2E	$\mathbf{57.6}^{\pm 0.4}$	$\mathbf{69.5}^{\pm 0.6}$	$\mathbf{65.6}^{\pm 0.7}$	$\mathbf{61.2}^{\pm 0.3}$	$\mathbf{58.9}^{\pm 1.7}$	$\mathbf{33.8}^{\pm 0.8}$	$\mathbf{76.5}^{\pm 0.6}$	$72.0^{\pm 0.2}$	$\mathbf{75.6}^{\pm 0.3}$	$\mathbf{66.3}^{\pm 0.2}$	$\mathbf{79.0}^{\pm 0.6}$	$\mathbf{67.5}^{\pm 1.1}$
	ACTE	$\mathbf{54.6}^{\pm 0.4}$	$\mathbf{62.6}^{\pm 1.2}$	$57.1^{\pm 2.4}$	$\mathbf{53.4}^{\pm 3.8}$	$\mathbf{59.3}^{\pm 1.1}$	$\mathbf{38.0}^{\pm 2.2}$	$\mathbf{73.8}^{\pm 0.2}$	$\mathbf{68.9}^{\pm 0.1}$	$\mathbf{71.8}^{\pm 0.1}$	$\mathbf{63.9}^{\pm 0.7}$	$\mathbf{73.1}^{\pm 0.4}$	$\mathbf{63.3}^{\pm 0.2}$
	TASD	$\mathbf{49.7}^{\pm 0.5}$	$\mathbf{58.0}^{\pm 1.1}$	$\mathbf{56.1}^{\pm 1.0}$	$\mathbf{50.2}^{\pm 1.1}$	$\mathbf{55.6}^{\pm 1.4}$	$\mathbf{31.4}^{\pm 1.6}$	$\mathbf{71.7}^{\pm 0.2}$	$\mathbf{67.2}^{\pm 0.3}$	$\mathbf{68.3}^{\pm 1.6}$	$\mathbf{59.1}^{\pm 1.4}$	$\mathbf{69.2}^{\pm 0.6}$	$\mathbf{60.4}^{\pm 0.1}$
Orca 2 7B	ACSA	$70.4^{\pm 0.7}$	$74.6^{\pm 0.5}$	$71.8^{\pm 1.5}$	$67.7^{\pm 0.9}$	$75.0^{\pm 0.8}$	$59.9^{\pm 1.3}$	$\mathbf{81.6}^{\pm 0.9}$	$79.9^{\pm 0.5}$	$81.8^{\pm 0.5}$	$80.0^{\pm 1.1}$	$81.3^{\pm 1.2}$	$79.0^{\pm 0.2}$
	E2E	$55.2^{\pm 1.4}$	$69.2^{\pm 1.3}$	$63.4^{\pm 1.2}$	$60.6^{\pm 1.0}$	$54.0^{\pm 1.7}$	$29.3^{\pm 0.9}$	$76.2^{\pm 0.3}$	$\mathbf{75.0}^{\pm 1.0}$	$73.5^{\pm 0.2}$	$59.7^{\pm 6.4}$	$77.6^{\pm 0.9}$	$64.4^{\pm 1.1}$
	ACTE	$54.2^{\pm 0.2}$	$61.1^{\pm 1.2}$	$\mathbf{57.4}^{\pm 1.7}$	$52.6^{\pm 0.3}$	$55.3^{\pm 0.7}$	$29.8^{\pm 5.5}$	$72.8^{\pm 0.4}$	$68.7^{\pm 1.0}$	$71.0^{\pm 1.1}$	$59.2^{\pm 0.4}$	$67.8^{\pm 0.3}$	$62.6^{\pm 1.1}$
	TASD	$45.1^{\pm 1.0}$	$55.9^{\pm 0.6}$	$49.5^{\pm 1.1}$	$45.2^{\pm 1.2}$	$47.6^{\pm 5.1}$	$24.8^{\pm 1.8}$	$70.0^{\pm 0.3}$	$66.6^{\pm 1.0}$	$66.1^{\pm 3.4}$	$56.0^{\pm 1.1}$	$68.0^{\pm 0.1}$	$60.2^{\pm 0.3}$

5.5 Inference and Training Speed

Table 10 presents the average absolute and relative times for training one epoch and inference time per example for various models on the TASD task, with English as the source language and Czech as the target language. The absolute times are measured in seconds, while the relative times indicate a comparison against the baseline model, mT5. All experiments were performed on a same hardware for comparability.

The mT5 model serves as the reference, with a relative training time of 1.00. The multi-tasking variant of mT5, while more computationally demanding during training (7.45 times slower than mT5), exhibits a similar inference time, as inference follows the same procedure as the baseline mT5 model. The Orca 2 13B model is significantly slower, requiring 9.40 times the training time of mT5 and a much higher inference time, 15.32 times that of mT5. This indicates that while larger models like Orca 2 13B may offer performance gains, they come with substantial computational costs during both training and inference.

Table 10. 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 Czech as the target language.

Model	Training Time Per Epoch		Inference Time Per Example	
	Absolute [s]	Relative	Absolute [s]	Relative
mT5	178	1.00	0.28	1.00
Multi-tasking mT5	1,326	7.45	0.27	0.96
Orca 2 13B	1,674	9.40	4.29	15.32

5.6 Error Analysis

To gain insights into the challenges of sentiment prediction, we conduct an error analysis focusing on identifying the most difficult sentiment elements to predict. We manually investigate 100 random test samples with predictions from the best-performing run of the mT5, both with and without constrained decoding, for a few language combinations (cs→en, en→cs, en→es, en→nl, en→fr) for the TASD task. Figure 4 shows the analysis for the en→es combination, alongside results with Orca 2 13B.

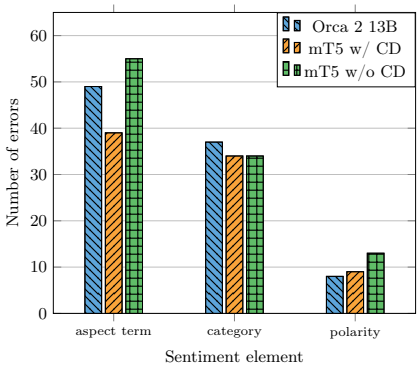


Fig. 4. Number of error types for Orca 2 13B and mT5 with and without constrained decoding (CD) on the Spanish target language and the TASD task.

Output Format One key challenge is producing the correct output format, which is crucial for target extraction. The models occasionally struggle with this, sometimes duplicating outputs and reducing the diversity of generated sentiment triplets. Although we ensure that identical triplets are not counted multiple times (thus not impacting the results), this repetition limits the models from generating unique outputs. It potentially causes them to miss specific prediction targets.

Aspect Term Prediction The primary source of error lies in aspect term prediction. As mentioned in the results for each task, the model sometimes generates the aspect term in the source language instead of the target language, a problem mitigated by constrained decoding. Constrained decoding helps mitigate this problem by restricting the generated tokens to only those in the input target language sentence, significantly reducing the available tokens pool.

Constrained decoding also helps reduce other errors, such as correcting typos and inventing words. For example, if the review contains the typo *"se**vr**ice"*, the model might generate the corrected word *"service"*. Without constrained decoding, the model also sometimes invents words. For instance, some reviews contain implicit opinions about the ambience, leading the model to generate *"ambience"* instead of *"it"* (an implicit aspect term). Constrained decoding reduces the generation of text that is not present in the original review or in a modified format.

Additionally, the models frequently generate partial aspect terms instead of complete ones, such as *"steak"* instead of *"Rump steak"*. Furthermore, the models may blend parts of the review, leading to outputs that do not match the original text's specific form. For instance, instead of *"Green Curry Ramen"*, the model might generate *"Green Ramen"*, a phrase not in the original review. Furthermore, the models occasionally produce lowercase output even when the original text contains uppercase letters.

Aspect Categories In terms of aspect categories, errors are less frequent than in the case of aspect terms. The models frequently omit the less common categories, such as *"location general"* or *"drinks style_ options"*. The models often confuse the *"restaurant miscellaneous"* and *"restaurant general"* classes, which are often inconsistent in the annotations. Some categories occur only in one or a few languages; for instance, *"food general"* appears solely in the Dutch test set, making it impossible for the classifier to learn from other source languages.

Sentiment Polarity The most common error concerning sentiment polarity is in predicting the *"neutral"* class, possibly due to imbalanced label distribution, since the *"neutral"* class is the least frequent in all datasets.

Dataset Labelling Additionally, we identified mistakes in the dataset labels. For example, in the test part of the English dataset, the aspect *"Service"* in the sentence *"Worst Service I Ever Had"* is labelled as *"positive"*, despite being clearly *"negative"*. Similarly, we noticed inconsistencies in the datasets, such as in the sentence *"One of the best hot dogs I have ever eaten"*, where the expression *"hot dogs"* is not labelled as an aspect term for the *"food quality"* category; instead, it is labelled as an implicit aspect term (*"NULL"*), contrary to other examples. These labelling errors could negatively impact the final scores of evaluated models.

6 Discussion & Recommendations

The results presented in Section 5 underscore the effectiveness of constrained decoding in cross-lingual aspect-based sentiment analysis, providing a practical alternative to translation tools, which can be tricky [44] or ineffective [17]. Constrained decoding significantly improves sequence-to-sequence models by addressing errors in aspect term prediction. These errors include generating the aspect terms in the source language instead of the target language and predicting aspect terms not present in the original review text. On average, constrained decoding improves the results by 5% in cross-lingual settings.

Sequence-to-sequence models offer advantages over encoder-based models used in previous cross-lingual ABSA studies due to their capability to detect implicit aspect terms and assign multiple sentiment polarities to a single aspect term, thereby providing a more comprehensive evaluation. Additionally, these models can be adapted for various ABSA tasks through straightforward changes to the output format. In contrast, encoder-based models require specialized architectures for complex ABSA tasks involving multiple sentiment elements. Moreover, sequence-to-sequence models facilitate multi-task fine-tuning, allowing simultaneous predictions for different tasks. These attributes make sequence-to-sequence models preferable for ABSA applications compared to encoder-based models. Among the evaluated models, mT5 consistently outperforms mBART.

Constrained decoding plays a crucial role in achieving competitive results with multi-tasking models compared to specialized models in aspect term prediction tasks, where it improves the results by more than 10%. While our findings do not favour multi-tasking over single-task specialization, a consistent trend suggests that multi-tasking matches or surpass single-task results in most cases. Given the substantial additional resources required for fine-tuning multi-tasking models – particularly the sixfold increase in training examples compared to single-task models, which prolongs fine-tuning duration – alongside the absence of clear performance advantages, we recommend fine-tuning specialized models when focusing on specific ABSA tasks. However, multi-tasking models offer a viable option for applications requiring simultaneous handling of multiple ABSA tasks, performing comparably to specialized models without significant performance trade-offs.

Using large language models in zero-shot and few-shot settings yields poor results for compound ABSA tasks but provides quick results without fine-tuning. ChatGPT consistently outperforms other evaluated LLMs in zero-shot and few-shot scenarios across all tasks and languages. However, fine-tuning LLMs can perform well in monolingual settings, often outperforming the mT5 model except for some languages.

Fine-tuning LLMs in cross-lingual settings can achieve performance similar to the mT5 model with constrained decoding in some cases. Fine-tuned LLMs generally perform better than mT5 when English is the target language. Nevertheless, using English as the target language is impractical and uncommon in real-world scenarios where English is typically the source language. For the more common scenario where English is the source language, mT5 often outper-

forms LLMs, sometimes by more than 10% for certain language combinations. The choice of LLM is crucial. Only Orca 2 13B achieves comparable results to mT5 among the evaluated LLMs. Since LLMs are predominantly pre-trained on English data[7] and multilingual open-source LLMs are still evolving, we recommend using mT5 with constrained decoding. Additionally, fine-tuning LLMs on consumer GPUs requires special techniques such as quantization and parameter-efficient fine-tuning, whereas fine-tuning the mT5 model does not require these techniques. Moreover, the training and inference times of LLMs are significantly higher compared to the mT5 model, which could be a considerable drawback in scenarios where rapid processing is crucial.

The error analysis presented in Section 5.6 reveals recurring dataset issues, including inconsistent labelling of aspect terms and categories and mislabelled sentiment polarity. These findings highlight the pivotal role of high-quality, consistent datasets in the training and evaluation of models. Addressing these challenges through enhanced data curation and rigorous cleaning processes can significantly improve model performance.

Based on our findings, we provide recommendations for various scenarios in Table 11 to enhance cross-lingual and monolingual ABSA.

Table 11. Recommendations for cross-lingual and monolingual ABSA.

Scenario	Recommendation
Quick results regardless of performance	Use LLMs with zero-shot or preferably few-shot prompts; larger models generally perform better.
High-performance monolingual results	Fine-tune LLMs, mT5 or a monolingual sequence-to-sequence model if available.
High-performance cross-lingual results	Use mT5 with constrained decoding or fine-tune the Orca 2 13B model.
High-performance cross-lingual results for multiple tasks simultaneously	Use multi-tasking mT5 with constrained decoding.
Quick training and inference	Fine-tune mT5 or a monolingual sequence-to-sequence model if available.

7 Conclusion

This paper presents a comprehensive study of cross-lingual and monolingual aspect-based sentiment analysis, with a primary focus on cross-lingual analysis, leveraging sequence-to-sequence models and large language models. We conduct extensive experiments across seven languages and six different ABSA tasks using restaurant domain datasets. Notably, we are the first to access four out of these six ABSA tasks evaluated. Our pioneering use of sequence-to-sequence models for cross-lingual ABSA, alongside the first application of fine-tuned LLMs in this domain, expands the frontier of cross-lingual ABSA research.

[7] For example, 90% of pre-training data for LLaMA 2 models is in English [34].

Our approach, centred on constrained decoding combined with sequence-to-sequence models, yields significant improvements in cross-lingual ABSA performance while eliminating reliance on external translation tools. Specifically, using constrained decoding improves the target-aspect-sentiment detection task results by 5% compared to not using it, demonstrating its effectiveness in enhancing model accuracy across diverse languages.

We surpass previous state-of-the-art results in cross-lingual ABSA tasks, showcasing the effectiveness of our proposed methodology. Moreover, our method supports multi-tasking capabilities, enabling simultaneous resolution of multiple ABSA tasks. Constrained decoding enhances the performance of multi-tasking models by more than 10%, underscoring its efficacy in addressing complex linguistic tasks across diverse languages.

In addition to evaluating LLMs in zero-shot, few-shot, and fine-tuning scenarios in both monolingual and cross-lingual settings, we demonstrate that while LLMs struggle in zero-shot and few-shot contexts compared to smaller fine-tuned models, fine-tuning boosts their performance significantly in monolingual tasks. However, LLMs generally lag behind smaller models equipped with constrained decoding in cross-lingual settings.

Furthermore, we provide detailed error analysis to highlight the primary challenges in cross-lingual ABSA, and we compare different models in terms of training and inference speed. Based on our extensive experimentation and evaluation, we propose a comprehensive set of recommendations tailored for various real-world scenarios in monolingual and cross-lingual ABSA applications. These recommendations aim to guide practitioners and researchers towards effective model selection and deployment strategies in diverse linguistic contexts.

For future research, potential directions could include verifying the effectiveness of our cross-lingual ABSA methods across different domains and languages, expanding beyond the restaurant domain. These efforts would involve creating and utilizing datasets designed explicitly for cross-lingual aspect-based sentiment analysis in various contexts. Currently, the availability of datasets in different languages for various domains is limited. Additionally, exploring cross-domain, cross-lingual ABSA could simultaneously assess the potential for transferring sentiment-related knowledge across different domains and languages. This challenging task could lead to significant advancements in cross-lingual sentiment analysis with practical applications across diverse fields.

Acknowledgments. This work has been supported by the Grant No. SGS-2025-022 – New Data Processing Methods in Current Areas of Computer Science. 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.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akhtar, M.S., Sawant, P., Sen, S., Ekbal, A., Bhattacharyya, P.: Solving data sparsity for aspect based sentiment analysis using cross-linguality and multilinguality. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 572–582. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1053, https://aclanthology.org/N18-1053
2. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q.V., Xu, Y., Fung, P.: A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In: Park, J.C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., Krisnadhi, A.A. (eds.) Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 675–718. Association for Computational Linguistics, Nusa Dua, Bali (Nov 2023). https://doi.org/10.18653/v1/2023.ijcnlp-main.45, https://aclanthology.org/2023.ijcnlp-main.45/
3. Cao, N.D., Izacard, G., Riedel, S., Petroni, F.: Autoregressive entity retrieval. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=5k8F6UU39V
4. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747
5. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: efficient finetuning of quantized llms. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
7. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
8. Dyer, C., Chahuneau, V., Smith, N.A.: A simple, fast, and effective reparameterization of IBM model 2. In: Vanderwende, L., Daumé III, H., Kirchhoff, K. (eds.) Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 644–648. Association for Computational Linguistics, Atlanta, Georgia (Jun 2013), https://aclanthology.org/N13-1073

9. Gao, T., Fang, J., Liu, H., Liu, Z., Liu, C., Liu, P., Bao, Y., Yan, W.: 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. pp. 7002–7012. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (Oct 2022), https://aclanthology.org/2022.coling-1.610
10. Gou, Z., Guo, Q., Yang, Y.: MvP: Multi-view prompting improves aspect sentiment tuple prediction. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4380–4397. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.240, https://aclanthology.org/2023.acl-long.240
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
12. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 4411–4421. PMLR (13–18 Jul 2020), https://proceedings.mlr.press/v119/hu20b.html
13. Hu, M., Wu, Y., Gao, H., Bai, Y., Zhao, S.: Improving aspect sentiment quad prediction via template-order data augmentation. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 7889–7900. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.538, https://aclanthology.org/2022.emnlp-main.538
14. Jebbara, S., Cimiano, P.: Zero-shot cross-lingual opinion target extraction. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 2486–2495. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1257, https://aclanthology.org/N19-1257
15. Klinger, R., Cimiano, P.: Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In: Proceedings of the Nineteenth Conference on Computational Natural Language Learning. pp. 153–163. Association for Computational Linguistics, Beijing, China (Jul 2015). https://doi.org/10.18653/v1/K15-1016, https://aclanthology.org/K15-1016
16. Lambert, P.: Aspect-level cross-lingual sentiment classification with constrained SMT. In: Zong, C., Strube, M. (eds.) 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). pp. 781–787. Association for Computational Linguistics, Beijing, China (Jul 2015). https://doi.org/10.3115/v1/P15-2128, https://aclanthology.org/P15-2128
17. Li, X., Bing, L., Zhang, W., Li, Z., Lam, W.: Unsupervised cross-lingual adaptation for sequence tagging and beyond (2021), https://arxiv.org/abs/2010.12405
18. Lin, N., Fu, Y., Lin, X., Yang, A., Jiang, S.: Cl-xabsa: Contrastive learning for cross-lingual aspect-based sentiment analysis (2023)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019)

20. Mao, Y., Shen, Y., Yang, J., Zhu, X., Cai, L.: Seq2Path: Generating sentiment tuples as paths of a tree. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022. pp. 2215–2225. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.174, https://aclanthology.org/2022.findings-acl.174
21. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J.: Large language models: A survey (2024)
22. Mitra, A., Corro, L.D., Mahajan, S., Codas, A., Simoes, C., Agarwal, S., Chen, X., Razdaibiedina, A., Jones, E., Aggarwal, K., Palangi, H., Zheng, G., Rosset, C., Khanpour, H., Awadallah, A.: Orca 2: Teaching small language models how to reason (2023)
23. OpenAI: Openai: Introducing chatgpt (Nov 2022), https://openai.com/blog/chatgpt
24. 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., Eryiğit, G.: SemEval-2016 task 5: Aspect based sentiment analysis. In: Bethard, S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P., Zesch, T. (eds.) Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 19–30. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/S16-1002, https://aclanthology.org/S16-1002
25. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2015 task 12: Aspect based sentiment analysis. In: Nakov, P., Zesch, T., Cer, D., Jurgens, D. (eds.) Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 486–495. Association for Computational Linguistics, Denver, Colorado (Jun 2015). https://doi.org/10.18653/v1/S15-2082, https://aclanthology.org/S15-2082
26. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: SemEval-2014 task 4: Aspect based sentiment analysis. In: Nakov, P., Zesch, T. (eds.) Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 27–35. Association for Computational Linguistics, Dublin, Ireland (Aug 2014). https://doi.org/10.3115/v1/S14-2004, https://aclanthology.org/S14-2004
27. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140:1–140:67 (2020), http://jmlr.org/papers/v21/20-074.html
28. Schmitt, M., Steinheber, S., Schreiber, K., Roth, B.: Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1109–1114. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1139, https://aclanthology.org/D18-1139/
29. Shazeer, N., Stern, M.: Adafactor: Adaptive learning rates with sublinear memory cost (2018)
30. Šmíd, J., Priban, P., Kral, P.: LLaMA-based models for aspect-based sentiment analysis. In: De Clercq, O., Barriere, V., Barnes, J., Klinger, R., Sedoc, J., Tafreshi, S. (eds.) Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis. pp. 63–

70. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024), https://aclanthology.org/2024.wassa-1.6
31. Šmíd, J., Priban, P., Kral, P.: Advancing cross-lingual aspect-based sentiment analysis with llms and constrained decoding for sequence-to-sequence models. In: Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART. pp. 757–766. INSTICC, SciTePress (2025). https://doi.org/10.5220/0013349400003890
32. Šmíd, J., Přibáň, P., Prazak, O., Kral, P.: Czech dataset for complex aspect-based sentiment analysis tasks. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 4299–4310. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.384
33. Tang, Y., Tran, C., Li, X., Chen, P.J., Goyal, N., Chaudhary, V., Gu, J., Fan, A.: Multilingual translation from denoising pre-training. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 3450–3466. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.304, https://aclanthology.org/2021.findings-acl.304/
34. Touvron, H., Martin, L., Stone, K., Albert, P., et al.: Llama 2: Open foundation and fine-tuned chat models (2023)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
36. Šmíd, J., Král, P.: Cross-lingual aspect-based sentiment analysis: A survey on tasks, approaches, and challenges (2025). https://doi.org/10.1016/j.inffus.2025.103073, https://www.sciencedirect.com/science/article/pii/S1566253525001460
37. Wan, H., Yang, Y., Du, J., Liu, Y., Qi, K., Pan, J.Z.: Target-aspect-sentiment joint detection for aspect-based sentiment analysis. Proceedings of the AAAI Conference on Artificial Intelligence **34**(05), 9122–9129 (Apr 2020). https://doi.org/10.1609/aaai.v34i05.6447, https://ojs.aaai.org/index.php/AAAI/article/view/6447
38. Wang, F., Lan, M., Wang, W.: Towards a one-stop solution to both aspect extraction and sentiment analysis tasks with neural multi-task learning. In: 2018 International joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2018)
39. Wang, W., Pan, S.J.: Transition-based adversarial network for cross-lingual aspect extraction. In: IJCAI. pp. 4475–4481 (2018)
40. 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., Rush, A.: Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6/
41. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: 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., Zhou, Y. (eds.) Proceedings of the 2021

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 483–498. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.41, https://aclanthology.org/2021.naacl-main.41

42. Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., Lam, W.: Aspect sentiment quad prediction as paraphrase generation. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 9209–9219. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.726, https://aclanthology.org/2021.emnlp-main.726
43. Zhang, W., Deng, Y., Liu, B., Pan, S., Bing, L.: Sentiment analysis in the era of large language models: A reality check. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024. pp. 3881–3906. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.findings-naacl.246, https://aclanthology.org/2024.findings-naacl.246
44. Zhang, W., He, R., Peng, H., Bing, L., Lam, W.: Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 9220–9230. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.727, https://aclanthology.org/2021.emnlp-main.727
45. Zhang, W., Li, X., Deng, Y., Bing, L., Lam, W.: 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). pp. 504–510. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-short.64, https://aclanthology.org/2021.acl-short.64
46. Zhou, X., Wan, X., Xiao, J.: Clopinionminer: Opinion target extraction in a cross-language scenario. IEEE/ACM Transactions on Audio, Speech, and Language Processing **23**(4), 619–630 (2015)