SJTU System Description for the WMT24 Low-Resource Languages of Spain Task

Tianxiang Hu[1*] **Pei Zhang**[2*] **Haoxiang Sun**[1] **Ruize Gao**[2] **Jialong Tang**[2]
Baosong Yang[2†] **Rui Wang**[1†]

[1]Shanghai Jiao Tong University, Shanghai, China
[2]Tongyi Lab, Hangzhou, China
{hutianxiang,wangrui12}@sjtu.edu.cn
{xiaoyi.zp,yangbaosong.ybs}@alibaba-inc.com

Abstract

This paper describes Shanghai Jiao Tong University low-resource languages of Spain translation systems for WMT24 shared task. We participate in the translation task on Spanish → Aragonese, Spanish → Aranese and Spanish → Asturian. Initially, we conduct preliminary experiments to assess the basic translation capabilities of various models and evaluate the impact of fine-tuning with different data types. We then choose to fine-tune the Qwen2-0.5B model using a forward synthesized pseudo-corpus from the Apertium translation system to replicate its fundamental performance. Building on this distillation model, we explore three optimization strategies across the three language directions: (1) Assembling the provided FLORES+ dev sets into a 5-shot format translation training dataset and performing few-shot fine-tuning to enhance model performance. (2) Utilizing the FLORES+ dev sets as training data and applying the Contrastive Preference Optimization (CPO) strategy for further refinement. (3) Retrieving the 20 most similar translation examples from the FLORES+ dev sets using the BM25 algorithm and performing 20-shot translations with the Claude 3.5-sonnet model. After evaluating these strategies, we select the best-performing approach for each language pair as our submission result.

1 Introduction

This paper introduces our submissions to the WMT24 Low-Resource Languages of Spain Task. We participate in the competitions for three translation directions: Spanish → Aragonese, Spanish → Aranese, and Spanish → Asturian. For the Spanish → Aragonese and Spanish → Aranese directions, we ultimately submit constrained results, while for Spanish → Asturian, we provide unconstrained (open system) results.

Neural machine translation (NMT) systems have achieved substantial advancements in recent years (Vaswani et al., 2017). However, training neural translation models typically necessitates large-scale parallel corpora (Ranathunga et al., 2021). In many low-resource scenarios, the availability of sufficient parallel data for training is limited, making low-resource translation a critical and valuable research area (Arivazhagan et al., 2019; Wang et al., 2021; Ranathunga et al., 2021). This competition task focuses on translating between Spanish and three other languages: Aragonese, Aranese, and Asturian. Of these, Aragonese and Aranese face particular challenges due to their relatively scarce parallel corpora. While the OPUS[1] website provides a considerable amount of parallel data, the quality of this data remains relatively low.

We initially conduct a preliminary evaluation of translation capabilities using models such as Apertium[2], GPT4 (Achiam et al., 2023), Llama-3 (AI@Meta, 2024), and Qwen2 (Yang et al., 2024) across the three language pairs. Our findings indicate that the Apertium translation system serves as a strong baseline, particularly in terms of BLEU (Papineni et al., 2002; Post, 2018) scores. Subsequently, we explore fine-tuning the Qwen2-0.5B model with various types of synthetic data and data from diverse domains. This exploration reveals that this task presents unique challenges compared to previous low-resource translation tasks. Specifically, forward-translated (Zhang and Zong, 2016) data and data from the OPUS NLLB corpus result in improved performance on dev test sets. We ultimately select the NLLB Spanish corpus from OPUS and perform forward translation using Apertium to generate the corresponding parallel pseudo-corpus. Fine-tuning Qwen2-0.5B with this synthetic data enables us to closely replicate the performance of the Apertium translation system.

[*] Tianxiang and Pei contributed equally. Work was done when Tianxiang was interning at Tongyi Lab.

[†] Rui Wang and Baosong Yang are co-corresponding authors.

[1] https://opus.nlpl.eu
[2] https://apertium.org

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 943–948
November 15–16, 2024.

Although simple forward distillation can effectively replicate the performance of the Apertium system, it does not exceed it, and the distilled model does not yield further performance improvements. To enhance the model's effectiveness, high-quality data is crucial. We randomly select a portion of the provided dev test set for additional fine-tuning, with the remaining portion designated as the new dev set. Building on this distilled model, we explore three optimization strategies across three language pairs: (1) We aggregate the provided FLORES+ dev sets into a 5-shot format translation training dataset and perform few-shot fine-tuning (Alves et al., 2023) to further refine the model. (2) We use the FLORES+ dev sets as training data and apply the Contrastive Preference Optimization (CPO) (Xu et al., 2024) strategy to improve model performance. (3) We retrieve the 20 most similar translation examples from the FLORES+ dev sets using the BM25 algorithm and employ the Claude 3.5-sonnet model[3] for 20-shot translations (Agrawal et al., 2022).

2 Preliminary Experiment

In this section, we first investigate the basic translation capabilities of various models and identify the Apertium translation system as a particularly strong baseline. We then examine the fine-tuning of the Qwen2-0.5B model using different types of data, which reveals that this task presents unique challenges compared to previous low-resource scenarios. Ultimately, we select the NLLB[4] Spanish corpus from OPUS, forward-translate it using Apertium to create a parallel pseudo-corpus, and fine-tune Qwen2-0.5B with this synthetic data.

Data The results presented in this section are derived from experiments conducted on the official FLORES+ dev test sets[5], which come from Pan-Iberian Language Archival Resource (PILAR). The three language pairs under consideration are Spanish → Aragonese (spa-arg), Spanish → Aranese (spa-arn), and Spanish → Asturian (spa-ast), each comprising 997 sentences.

2.1 Translation capabilities of different models

We begin by evaluating the BLEU (Papineni et al., 2002; Post, 2018) performance of five models (Apertium, GPT-4, Llama3-8B, Llama3-70B, and Qwen2-0.5B) on the three language pairs in this task using the FLORES+ dev sets. For the 1-shot scenario, the format used is as follows: "Translate the following sentence from <src lang> into <tgt lang>.\n <src lang>: <src example1>.\n <tgt lang>: <tgt example1>.\n \n Translate the following sentence from <src lang> into <tgt lang>.\n <src lang>: <src sentence>.\n <tgt lang>:". In the 5-shot scenario, this format is extended by providing five examples instead of one. The few-shot examples are randomly sampled from the corresponding language FLORES+ dev sets without repetition.

As shown in Table 1, our results indicate that the Apertium translation system serves as a very strong baseline, significantly outperforming other large models in BLEU scores for the Spanish → Aragonese (spa-arg) and Spanish → Aranese (spa-arn) language pairs. Notably, even the widely used GPT-4 scores considerably lower in BLEU compared to the Apertium system. This superior performance of Apertium may be attributed to the fact that the dev test sets for these two language pairs were derived from Apertium's translations with post-editing. Additionally, we observed that increasing the number of example shots in translation leads to a substantial improvement in performance. This suggests that, for these low-resource languages, providing translation examples enhances the ability of large models to learn and perform the translation task more effectively.

	spa-arg	spa-arn	spa-ast
Apertium	66.0	38.0	17.1
GPT4 1shot	35.9	16.1	18.6
GPT4 5shot	37.4	17.7	19.1
Llama3-8B 1shot	36.3	7.8	16.6
Llama3-8B 5shot	41.0	10.6	18.3
Llama3-70B 1shot	46.4	15.6	19.4
Llama3-70B 5shot	52.4	19.9	22.4
Qwen2-0.5B 1shot	22.7	4.1	8.6
Qwen2-0.5B 5shot	22.7	4.2	8.9

Table 1: BLEU evaluation of different models on dev test sets for three language pairs. Apertium translation system demonstrates a strong baseline.

2.2 Effects of different types of data

To explore the types of data that can be used for fine-tuning the base model, we conduct preliminary experiments focusing exclusively on Aragonese.

[3] https://claude.ai
[4] https://opus.nlpl.eu/NLLB/corpus/version/NLLB
[5] https://github.com/transducens/PILAR

As shown in Table 2, we evaluate the impact of different data types on fine-tuning performance. Our findings indicate that forward translation (FT) (Zhang and Zong, 2016) outperforms back translation (Sennrich et al., 2016). This result may be attributed to the fact that the dev test set is derived from Apertium with post-editing, which means that the Aragonese side of the dev test set reflects Apertium's translation style rather than the natural language style of Aragonese. In contrast, back translation targets the authentic Aragonese language style, which does not align with the style of the dev test set, potentially leading to BLEU scores that do not accurately represent the actual translation quality. However, due to the extremely low-resource nature of this language, we have to rely on the official dev test set and BLEU scores for optimization.

Additionally, the table highlights another critical factor affecting performance: the source of the fine-tuning data. Using Spanish monolingual data from the OPUS NLLB corpus[6] provides a noticeable performance advantage over using WMT news[7], Pilar[8] or random samples from OPUS[9]. This suggests that the domain of the dev test set is more closely aligned with the OPUS NLLB corpus, facilitating better adaptation to the dev set for this task. Furthermore, we observe that mixing data from different domains or simultaneously using both BT and FT does not enhance performance, despite increasing the volume of data. In fact, this approach slightly degrades the original performance.

2.3 Final Distillation Experiment

Based on the experimental results discussed above, we first perform basic filtering on the NLLB Spanish corpus from OPUS and then randomly sample 1 million sentences. We use the Apertium translation system to translate these 1 million Spanish sentences into the three target languages, creating a parallel pseudo-corpus. We then fine-tune the open-source Qwen2-0.5B model separately for each language using this pseudo-corpus. During training, we fine-tune the model for 1.5 epochs with a batch size of 64, a learning rate of 1e-05, and a weight decay of 0.1. For decoding, we employ beam search with a beam size of 4. As shown in Table 3, this approach effectively replicates the baseline performance of the Apertium translation system.

[6] https://opus.nlpl.eu/NLLB/corpus/version/NLLB

[7] https://www.statmt.org/wmt11/translation-task.html

[8] https://github.com/transducens/PILAR

[9] https://opus.nlpl.eu

Data size	Data source	Data type	BLEU
16k	OPUS	bilingual	37.7
16k	OPUS	FT	61.7
16k	News	FT	53.0
16k	OPUS NLLB	FT	63.8
16k	OPUS	BT	41.1
16k	Pilar	BT	34.3
32k	OPUS	FT+BT	59.6
32k	OPUS+News	FT	59.8

Table 2: BLEU evaluation on fine-tuning Qwen2-0.5B using different types of data. Data size refers to the training data size. FT refers to forward translation of Spanish to comprise synthesized parallel data; BT refers to backward translation of Aragonese to comprise synthesized parallel data; News referes to the WMT news.

	spa-arg	spa-arn	spa-ast
Apertium	66.0	38.0	17.1
distillation model	66.0	38.0	17.0

Table 3: BLEU evaluation of the distillation model on dev test sets for three language pairs. We have replicated the baseline capability of Apertium translation system.

3 Method

In Section 3, we initially replicate the performance of the strong baseline system Apertium using the Qwen2-0.5B model but are unable to surpass it. We also observe that fine-tuning the model with filtered bilingual data resulted in decreased BLEU scores, likely due to the low quality of available bilingual data. The synthetic pseudo-corpus generated through forward translation reach its performance limits, as further improvements could not be achieved with the distilled model. To address this, we randomly select 700 sentences from the provided dev test set for additional fine-tuning, reserving the remaining 297 sentences as the new dev set. We next explore three optimization strategies to further enhance translation performance for these three language pairs.

3.1 Dev 5shot SFT

In Table 1, we observe that providing few-shot examples to large language models improves trans-

lation performance. However, supervised fine-tuning can reduce some of these few-shot capabilities (Alves et al., 2023). To maintain consistency in the inference format, we structure the fine-tuning data into a 5-shot format during training. For inference, we also use the 5-shot format, with few-shot examples randomly selected from the dev test set. The fine-tuning data consists of 700 sentences from the previously mentioned dev test set.

3.2 Dev CPO

Given that the official dev test set is derived from post-edited results, our goal is to assist the model in learning the subtle distinctions between pre-edited and post-edited translations, thereby enhancing its translation capabilities. DPO (Rafailov et al., 2023) is a training strategy focused on optimizing preferences, while CPO (Xu et al., 2024) builds upon DPO by providing further refinements. The following is the formulation of CPO loss:

$$\mathcal{L}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\pi_\theta(y_w|x) - \beta\log\pi_\theta(y_l|x)\Big)\Big], \quad (1)$$

$$\min_\theta \underbrace{\mathcal{L}(\pi_\theta)}_{\mathcal{L}_{\text{prefer}}} \underbrace{-\mathbb{E}_{(x,y_w)\sim\mathcal{D}}[\log\pi_\theta(y_w|x)]}_{\mathcal{L}_{\text{NLL}}}, \quad (2)$$

where x is source sentence, y_w is preferred translation, y_l is less preferred translation, $\mathcal{D}$ is a dataset of comparisons.

In our approach, we use translations produced by Apertium as negative examples and the corresponding results from the dev test set as positive examples. This CPO training allows the model to learn the nuanced differences between positive and negative instances.

3.3 Dev fewshot BM25 with LLM

Previous research suggests that providing similar parallel translation pairs as guidance can improve translation quality with large language models (Agrawal et al., 2022). To leverage this, we use the BM25 algorithm to retrieve several of the most similar translation examples from the dev test set based on the source sentences. These examples are concatenated into the previously described few-shot translation format and positioned before the sentence to be translated. We then employ state-of-the-art LLMs, such as GPT-4 and Claude-3.5, for the translation process.

3.4 Post-processing

We observe that translations produced by large language models may encounter issues such as omissions, over-translation, and non-following with instructions (Jiao et al., 2023; Xu et al., 2023). To address these issues, we apply the following rule-based post-processing:

1. For translations generated by the Dev 20-shot BM25 method with LLMs, if the output fails to adhere to instructions, for instance, if it includes phrases such as "I apologize" or "sorry", we perform a retranslation. If the correct translation is not achieved after three attempts, we revert to the translation produced by the Apertium software.

2. Replace any translations where language detection is incorrect with those generated by Apertium software.

3. Replace any translations where the ratio of the length of the source text to the translated text is less than 0.75 or greater than 1.3 with translations generated by Apertium software.

4 Experiment

Data In this Section, we randomly selected 700 sentences from the provided dev test set for additional fine-tuning, leaving the remaining 297 sentences as the new dev set.

Experiment Details For SFT, we fine-tune the distillation model for 5 epochs with a batch size of 8, a learning rate of 1e-05, and a weight decay of 0.1. For decoding, we use beam search with a beam number of 4. For few-shot BM25, we use the BM25 algorithm to select a number of the most similar examples (excluding the sentence itself) from the 997 sentences in the dev set for few-shot translation.

Results As illustrated in Table 4, the BLEU scores for the three language pairs across various methods demonstrate noticeable performance improvements over the Distillation model. Specifically, the best performance for Spanish → Aragonese (spa-arg) is achieved with the Distillation model + dev 5-shot SFT, for Spanish → Aranese (spa-arn) with the Distillation model + dev CPO, and for Spanish → Asturian (spa-ast) with Claude 3.5-sonnet + 20-shot BM25.

Furthermore, the dev 5-shot SFT method yields a more consistent performance improvement compared to direct dev SFT. Among the models evaluated, Claude 3.5-sonnet generally outperforms GPT-4-turbo across these three low-resource language pairs, and BM25 retrieval of similar examples significantly boosts translation performance.

	spa-arg	spa-arn	spa-ast
Distillation model	67.4	39.5	17.0
+ dev SFT	69.3	40.5	17.3
+ dev 5shot SFT	**69.9**	40.8	17.4
+ dev CPO	69.7	**41.4**	17.3
GPT4-turbo			
+ 5shot	40.5	32.3	20.1
+ 5shot BM25	44.3	33.2	20.5
+ 20shot BM25	47.5	33.6	21.4
Claude3.5-sonnet			
+ 5shot	47.8	35.2	22.9
+ 5shot BM25	53.6	37.4	24.2
+ 20shot BM25	59.9	38.1	**25.2**

Table 4: BLEU evaluation of different methods on partitioned dev test sets for three language pairs. Our methods all achieve certain performance improvements. For the Aragonese language pair, the best strategy is dev 5-shot SFT. For the Aranese language pair, the optimal strategy is dev CPO. For Asturian language pair, the best approach is using Claude 3.5-sonnet for 20-shot BM25 translation.

5 Conclusion

This paper presents the Shanghai Jiao Tong University translation systems for low-resource Spanish languages in the WMT24 shared task. We first create synthetic data through forward distillation using the Apertium translation system, then fine-tune the Qwen2-0.5B model to establish a basic baseline capability. Subsequently, we apply three optimization strategies using the dev test sets: 5-shot format fine-tuning, Contrastive Preference Optimization, and 20-shot translation with BM25 retrieval. Our experiments demonstrate that all three methods lead to performance improvements.

Acknowledgements

The work was supported by Alibaba Innovative Research Program, the General Program of National Natural Science Foundation of China (62176153), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), the Science and Technology Development Fund, Macau SAR (Grant Nos. FDCT/060/2022/AFJ, FDCT/0070/2022/AMJ), and the Multi-year Research Grant from the University of Macau (Grant No. MYRG-GRG2023-00006-FST-UMDF).

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2023. Gpt-4 technical report.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

AI@Meta. 2024. Llama 3 model card.

Duarte M. Alves, Nuno M. Guerreiro, Joao Alves, José P. Pombal, Ricardo Rei, Jos'e G. C. de Souza, Pierre Colombo, and André Martins. 2023. Steering large language models for machine translation with fine-tuning and in-context learning. In *Conference on Empirical Methods in Natural Language Processing*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. ParroT: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for low-resource languages: A survey. *ArXiv preprint*, abs/2106.15115.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. In *International Joint Conference on Artificial Intelligence*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *ArXiv preprint*, abs/2309.11674.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.