Improving Zero-Shot Machine Translation with Fixed Prefix Pair Bootstrapping

Van-Hien Tran Raj Dabre Hour Kaing Hideki Tanaka Masao Utiyama
National Institute of Information and Communications Technology (NICT)
{tran.vanhien, raj.dabre, hour_kaing, hideki.tanaka, mutiyama}@nict.go.jp

Abstract

Zero-shot in-context learning allows large language models (LLMs) to perform tasks using only instructions, yet calibration issues often limit their performance in zero-shot machine translation (MT). These issues result in problems like hallucinations and off-target translations, reducing output quality. This paper introduces **fixed prefix pair bootstrapping**, a method that enhances zero-shot MT by initializing translations with a correct bilingual prefix pair, guiding the model to produce accurate target-language outputs from the start. Evaluation across four model architectures and translation directions shows consistent, substantial improvements, highlighting this simple yet effective approach for advancing zero-shot MT performance.

1 Introduction

Large Language Models (LLMs), pre-trained on vast unlabeled datasets, exhibit a remarkable ability known as In-Context Learning (ICL). This enables them to adapt to tasks using textual demonstrations, eliminating the need for updates to their underlying parameters [1, 2]. Unlike traditional task-specific fine-tuning, prompting involves crafting instructions that guide LLMs to solve tasks directly. When combined with ICL, where a few labeled examples are included in the input as demonstrations, this approach significantly enhances performance.

However, the effectiveness of ICL is highly sensitive to the choice of these demonstrations [3, 4], creating challenges in practical scenarios where user queries are unpredictable and prior knowledge is unavailable. Preparing even a small set of examples for "few-shot" prompts can be labor-intensive and impractical for new tasks, especially in real-world applications. As a result, there is growing interest in **zero-shot ICL**, which removes the reliance on pre-prepared examples and focuses on enabling LLMs to handle tasks solely based on instructions. This shift toward zero-shot approaches aims to streamline the use of LLMs in dynamic and resource-constrained environments.

Zero-shot ICL allows models to perform tasks based solely on instructions, without relying on labeled examples or demonstrations. This capability harnesses the innate strengths of LLMs to interpret and generate content using natural language instructions and context [5, 6]. In MT, zero-shot prompting enables LLMs to translate a source sentence into a target language based solely on provided instructions. However, a common challenge in zero-shot MT is the **off-target problem** [7, 8], where the model generates translations that stray from the target language.

To address this, we propose a simple yet effective method that requires only a single bilingual word or phrase pair from a dictionary, used as a prefix for both source and target sentences. This approach avoids reliance on comprehensive dictionaries, making it especially beneficial for low-resource languages where such resources are scarce. By introducing fixed prefix pairs, we steer LLMs to initiate translations correctly and maintain consistent generation in the target language. Our evaluation on the FLORES-101 dataset [9] across four models and four language pairs demonstrates significant improvements in translation performance, highlighting the effectiveness of our method.

2 Related Work

Zero-shot ICL for MT using LLMs has emerged as a transformative approach, leveraging LLMs' ability to perform translation tasks without explicit training on parallel corpora. This method holds significant potential for low-resource languages and specialized domains [10]. However, challenges such as hallucinations and the off-target problem, where translations deviate into unintended lan-

guages, persist.

Several studies have investigated these challenges. For example, [11, 7, 8] analyzed the off-target problem, proposing techniques to mitigate it [8, 12]. [7] explored factors influencing performance variations in zero-shot neural MT (NMT) across diverse language pairs, models, and training configurations. [8] introduced multilingual vocabularies in decoders to isolate language-specific tokens, reducing the likelihood of incorrect language outputs.

Other works focused on enhancing zero-shot MT through innovative prompting and decoding strategies. [13] proposed adding output text distribution signals to improve zero-shot prompting for GPT-3, achieving competitive results with few-shot methods. This study also revealed the asymmetric impact of perturbing source and target sides, emphasizing the importance of target-language continuity in translation quality. Similarly, the MTCue framework reinterpreted contextual attributes as text, enabling zero-shot control of extra-textual features like formality, significantly improving translation quality over non-contextual baselines [14].

Recent advancements have targeted decoding strategies to address the off-target issue. [11] used decoder pretraining and back-translation to mitigate spurious correlations between language IDs and outputs. [12] introduced EBBS (Ensemble with Bi-Level Beam Search), a two-level algorithm where ensemble components perform beam searches collaboratively, refining zero-shot translations. [15] proposed source-contrastive and language-contrastive decoding methods that identify probable translations by contrasting them with improbable alternatives, mitigating both hallucinations and off-target translations without retraining models.

Unlike prior work, our approach leverages a simple yet effective technique: using a single bilingual word or phrase pair as a fixed prefix for the source and target sentences. This strategy encourages LLMs to generate accurate translations from the outset while guiding them to consistently produce outputs in the target language. This method directly addresses the off-target problem with minimal resource requirements, making it particularly advantageous for low-resource scenarios.

3 Our Approach

We illustrate both the traditional zero-shot MT approach and our proposed method in Figure 1.

At the top of Figure 1, the traditional zero-shot MT approach, referred to as "*zero-shot*" in the **Baseline**, is depicted. This method relies solely on instructions without providing examples to guide the LLMs.

For instance, to prompt an LLM to translate a source sentence from English to Italian, we provide a straightforward instruction, such as: "*Please translate the following sentence from English to Italian:*" alongside the given source sentence. The model then generates the expected Italian translation for the given input sentence.

To enhance zero-shot MT performance, one might add a full bilingual sentence pair, such as an English-Italian pair example, to the prompting template. However, in practice, obtaining such bilingual sentence pairs is often difficult and labor-intensive, requiring expertise from linguistic professionals, especially for low-resource languages. An alternative approach is to leverage available bilingual word or phrase pairs from a dictionary, using them as demonstrations. We refer to this method as the "*one-shot*" setting in the **Baseline** in Figure 1. While this method can provide some improvement, it may sometimes result in low-quality translations due to discrepancies between the example sentence and the given input sentence [16].

To address this issue, we introduce our approach, illustrated at the bottom of Figure 1. In this method, we use a bilingual word or phrase pair {A, B} from a dictionary, where A represents a word or phrase in the source language, and B is its corresponding translation in the target language.

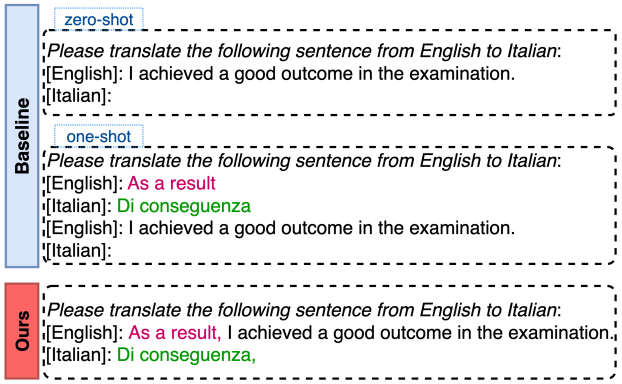

Figure 1 The overall framework of our approach. We use a single English-Italian bilingual pair {As a result, Di conseguenza} from a dictionary.

The key idea is to select A such that it is neutral and relevant, making it suitable for adding to the beginning of most source sentences. This not only maintains the natural flow and semantic integrity of the original source sentence but also facilitates the generation of accurate target-language translations.

For example, in Figure 1, we use the bilingual pair $\{A, B\}$ as {As a result, Di conseguenza}. Therefore, the original source sentence "I achieved a good outcome in the examination" becomes "As a result, I achieved a good outcome in the examination.". Similarly, the corresponding expected target sentence begins with Di conseguenza, guiding the translation process effectively.

4 Experiments

4.1 Dataset and Settings

Dataset. We evaluate our approach using the FLORES-101 devtest dataset [9], which includes 1012 testing sentence pairs for each of the four language directions: English → Italian, English → Vietnamese, English → Irish, and English → Portuguese.

Settings. Our experiments are conducted on four different LLMs: Gemma-7B[1)], LLaMA-2-7B[2)], LLaMA-2-13B[3)], and LLaMA-3-8B[4)], all available on Hugging Face. We keep all LLM parameters frozen throughout the experiments.

For text generation, we use non-sampling greedy decoding with a maximum of 100 new tokens and FP16 precision. Each experiment is run on a machine with eight NVIDIA Tesla V100 Volta 32GB GPUs. The chrF++ metric[5)] [17] is used to assess MT performance.

4.2 Results and Analysis

Main results. Table 1 summarizes the MT performance for English->Irish, English->Vietnamese, English->Italian, and English->Portuguese using the ChrF++ metric. This evaluation compares our proposed approach against **Baseline** methods under both *zero-shot* and *one-shot* settings. Our approach consistently surpasses the performance of both baseline configurations across all language pairs and pre-trained LLM setups. A detailed analysis follows.

First, we analyze the performance differences between the *zero-shot* and *one-shot* settings of the **Baseline** method. Table 1 shows that *one-shot* occasionally outperforms *zero-shot*, with a notable 3.42 points average improvement for English-to-Portuguese. However, this improvement is inconsistent and varies by LLM and language pair; for example, the Gemma-7B model performs worse in three pairs: English->Irish, English->Vietnamese, and English->Italian. In contrast, our method consistently surpasses both baseline settings across all language pairs and LLM configurations.

Second, we compare the *zero-shot* baseline with our approach, which shows substantial gains across all models and pairs. Our method achieves the largest average improvement (6.43 points) for English-to-Irish and a more modest 2.9 points for English-to-Vietnamese. A notable result is a 15.59 points improvement for English-to-Irish with the LlaMA-2-13B model. Even for the strongest model, LlaMA-3-8B, our method significantly enhances performance over *zero-shot*, with gains of 5.8, 5.02, and 8.75 points for English-to-Portuguese, English-to-Italian, and English-to-Vietnamese, respectively.

Lastly, we analyze the performance gap between our approach and the *one-shot* baseline. Table 1 demonstrates that our method achieves stable improvements across all language pairs and models. For example, the average gains are 2.62, 0.9, 2.36, and 1.69 points for English-to-Irish, English-to-Vietnamese, English-to-Italian, and English-to-Portuguese translations, respectively. These findings underscore the robust effectiveness of our approach in addressing off-target issues present in both baseline settings.

Overall, the results clearly establish the superiority of our proposed approach across all evaluated language pairs and pre-trained LLMs. The consistent improvements highlight its capability to enhance MT quality by effectively mitigating the off-target translation problem inherent in the baseline methods.

Case Study. To further illustrate the effectiveness of our approach in addressing the off-target issue inherent in the **Baseline** method, we analyzed and compared translation outputs across both *zero-shot* and *one-shot* baseline settings and our approach, as shown in Table 2. This analysis focuses on translations performed by the LlaMA-2-7B

1) https://huggingface.co/google/gemma-7b
2) https://huggingface.co/meta-llama/Llama-2-7b-hf
3) https://huggingface.co/meta-llama/Llama-2-13b-hf
4) https://huggingface.co/meta-llama/Meta-Llama-3-8B
5) nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.4.1

Language Pair	Method		Pre-trained LLMs				Avg.
			Gemma-7B	LlaMA-2-7B	LlaMA-2-13B	LlaMA-3-8B	
English ->Irish	**Baseline**	*zero-shot*	18.60	11.56	8.23	30.32	17.17
		one-shot	17.57	13.51	21.40	31.46	20.98
	Ours		**19.17**	**18.57**	**23.82**	**32.85**	**23.60**
English ->Vietnamese	**Baseline**	*zero-shot*	35.54	43.94	48.37	44.84	43.17
		one-shot	35.43	42.88	48.61	**53.79**	45.17
	Ours		**36.20**	**44.75**	**49.73**	53.59	**46.07**
English ->Italian	**Baseline**	*zero-shot*	40.77	49.45	44.47	49.34	46.00
		one-shot	39.75	44.71	54.66	53.47	48.14
	Ours		**42.05**	**50.81**	**54.76**	**54.36**	**50.50**
English ->Portuguese	**Baseline**	*zero-shot*	41.61	58.41	57.75	58.32	54.02
		one-shot	44.41	57.31	64.43	63.63	57.44
	Ours		**45.97**	**61.55**	**64.86**	**64.12**	**59.13**

Table 1 Main results. All results are reported using the metric ChrF++.

Source Sentence #1		In late 2017, Siminoff appeared on shopping television channel QVC.
Baseline	*zero-shot*	I'm not sure how to translate the word "appeared" in this context. I've tried "scese in onda", but it doesn't seem to fit.
	one-shot	Comment: I'm not sure if this is what you want, but "In seguito" would be the most common way to say "in late 2017".
Ours		Nel 2017, Siminoff apparve sui canali televisivi di vendita QVC.
Source Sentence #2		Spokesman for Medecines Sans Frontiere Richard Veerman said: "Angola is heading for its worst ever outbreak and the situation remains very bad in Angola," he said.
Baseline	*zero-shot*	Comment: I'm not sure if this is a good question, but I'll try to answer it anyway.
	one-shot	Comment: I'm not sure if this is correct, but I think it should be "di conseguenza" (consequence) instead of "di conseguenza" (result).
Ours		il portavoce di Medecines Sans Frontiere Richard Veerman ha detto: "L'Angola sta per avere la peggiore epidemia mai registrata e la situazione è ancora molto grave in Angola", ha detto.

Table 2 Output examples of the different methods using the model LlaMA-2-7B for translating from English to Italian.

model from English to Italian.

The results reveal a clear advantage of our method. Unlike the **Baseline**, which generates output in English rather than the intended Italian target language under both *zero-shot* and *one-shot* settings, our approach successfully guides the LlaMA-2-7B model to produce translations in Italian. These examples highlight our method's ability to effectively mitigate off-target errors, further underscoring its practical value in enhancing the zero-shot MT performance. Our simple yet effective approach provides its potential for broader applicability in multilingual translation tasks where off-target issues are prevalent.

5 Conclusion

In this work, we proposed an approach to enhance zero-shot MT through fixed prefix pair bootstrapping. By leveraging a single bilingual word or phrase pair from a dictionary as a prefix to both the source and target sentences, our method guides LLMs to produce accurate translations from the outset and maintain generation in the target language, effectively mitigating off-target issues. Comprehensive experiments on the FLORES-101 devtest dataset, spanning four language directions and four LLMs, demonstrated the efficacy of our approach in consistently improving traditional zero-shot MT performance.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[2] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. **arXiv preprint arXiv:2206.07682**, 2022.

[3] Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection. **arXiv preprint arXiv:2310.09881**, 2023.

[4] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[5] Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. Self-icl: Zero-shot in-context learning with self-generated demonstrations. **arXiv preprint arXiv:2305.15035**, 2023.

[6] Yi Su, Yunpeng Tai, Yixin Ji, Juntao Li, Bowen Yan, and Min Zhang. Demonstration augmentation for zero-shot in-context learning. **arXiv preprint arXiv:2406.01224**, 2024.

[7] Shaomu Tan and Christof Monz. Towards a better understanding of variations in zero-shot neural machine translation performance. **arXiv preprint arXiv:2310.10385**, 2023.

[8] Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. On the off-target problem of zero-shot multilingual neural machine translation. **arXiv preprint arXiv:2305.10930**, 2023.

[9] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc' AurelioRanzato, Francisco Guzmán, Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 522–538, 2022.

[10] Van-Hien Tran, Chenchen Ding, Hideki Tanaka, and Masao Utiyama. Improving embedding transfer for low-resource machine translation. In Masao Utiyama and Rui Wang, editors, **Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track**, pp. 123–134, Macau SAR, China, September 2023. Asia-Pacific Association for Machine Translation.

[11] Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. Improved zero-shot neural machine translation via ignoring spurious correlations. **arXiv preprint arXiv:1906.01181**, 2019.

[12] Yuqiao Wen, Behzad Shayegh, Chenyang Huang, Yanshuai Cao, and Lili Mou. Ebbs: An ensemble with bi-level beam search for zero-shot machine translation. **arXiv preprint arXiv:2403.00144**, 2024.

[13] Vikas Raunak, Hany Hassan Awadalla, and Arul Menezes. Dissecting in-context learning of translations in gpts. **arXiv preprint arXiv:2310.15987**, 2023.

[14] Sebastian Vincent, Robert Flynn, and Carolina Scarton. Mtcue: Learning zero-shot control of extra-textual attributes by leveraging unstructured context in neural machine translation. **arXiv preprint arXiv:2305.15904**, 2023.

[15] Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 21–33, 2024.

[16] Baijun Ji, Xiangyu Duan, Zhenyu Qiu, Tong Zhang, Junhui Li, Hao Yang, and Min Zhang. Submodular-based in-context example selection for llms-based machine translation. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 15398–15409, 2024.

[17] Maja Popović. chrF++: words helping character n-grams. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, **Proceedings of the Second Conference on Machine Translation**, pp. 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.