

Do Multilingual Language Models Think Better in English?

Julen Etxaniz¹ Gorka Azkune¹ Aitor Soroa¹ Oier Lopez de Lacalle¹ Mikel Artetxe²

¹HiTZ Center, University of the Basque Country UPV/EHU ²Reka AI

{julen.etxaniz,gorka.azkune,a.soroa,oier.lopezdelacalle}@ehu.eus mikel@reka.ai

Abstract

Translate-test is a popular technique to improve the performance of multilingual language models. This approach works by translating the input into English using an external machine translation system, and running inference over the translated input. However, these improvements can be attributed to the use of a separate translation system, which is typically trained on large amounts of parallel data not seen by the language model. In this work, we introduce a new approach called self-translate, which overcomes the need of an external translation system by leveraging the few-shot translation capabilities of multilingual language models. Experiments over 5 tasks show that self-translate consistently outperforms direct inference, demonstrating that language models are unable to leverage their full multilingual potential when prompted in non-English languages. Our code is available at <https://github.com/juletx/self-translate>.

1 Introduction

Multilingual autoregressive language models like XGLM (Lin et al., 2022), BLOOM (Scao et al., 2023) and PaLM (Chowdhery et al., 2022; Anil et al., 2023) have shown impressive capabilities on many tasks and languages. However, performance is usually lower for non-English languages, especially for low-resource ones (Ahuja et al., 2023). A common approach to mitigate this problem is to use translate-test, where the test data is translated into English using an external Machine Translation (MT) system, and then fed into the model. While primarily explored in the traditional pretrain/finetune paradigm (Ponti et al., 2021; Artetxe et al., 2023), early evidence has shown that translate-test can also bring sizeable improvements for few-shot learning with autoregressive language models (Shi et al., 2022).

However, translate-test relies on a separate MT system, which is usually trained on large amounts

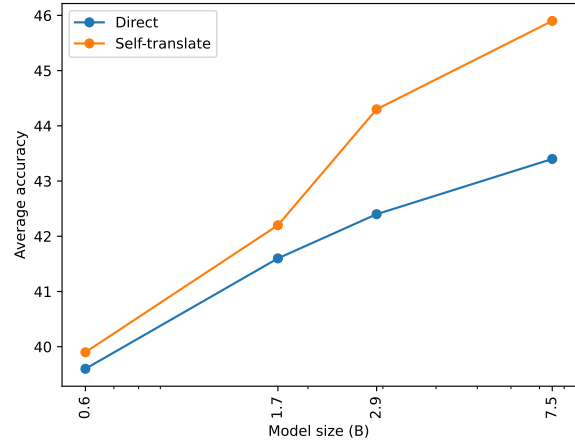


Figure 1: **XGLM results (average accuracy)**. We show that self-translate (using the model itself to translate the input into English) works better than using the original input in the non-English language.

of parallel data not seen by the primary model. In this paper, we investigate if the improvements from translate-test are solely due to the use of additional resources. To answer this question, we propose a new approach called self-translate, which leverages the few-shot translation capabilities of autoregressive language models (Vilar et al., 2023) instead of using an external system. More concretely, we prompt multilingual models to translate the input into English, and then feed the translated input to the same model to solve the task (Figure 2).

As shown in Figure 1, we find that self-translate works better than solving the task directly in the original language. This demonstrates that multilingual language models are unable to leverage their full potential when prompted in non-English languages. We find this phenomenon to be consistent across tasks, and more prominent for large models and high-resource languages. All in all, our work reveals an important limitation of multilingual language models, and prompts for future work to unleash their full potential without the need of intermediate inference steps.

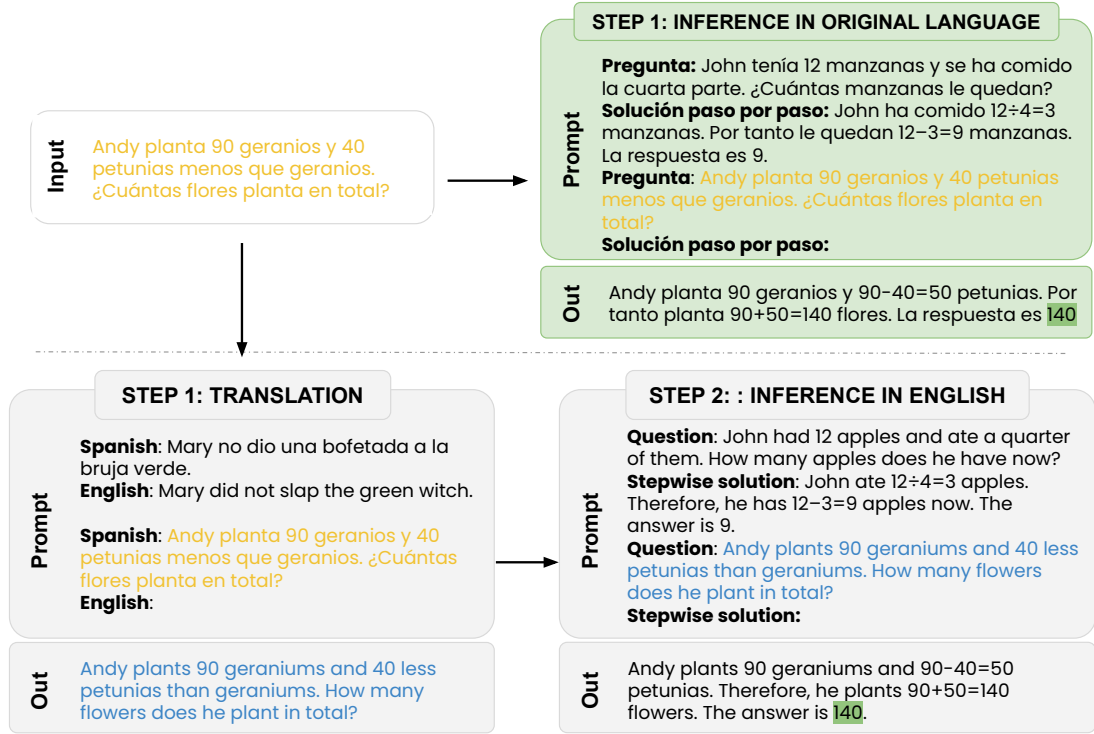


Figure 2: **Direct inference (top) vs. self-translate (bottom).** In direct inference (standard) the task is solved by prompting the model in the original language. In self-translate (proposed), we first translate the input into English by prompting the same model, and then solve the task in English.

2 Experimental settings

We next describe our experimental design, and report additional details in Appendix A.

Models. We experiment with 7 models from 2 families: the 564M, 1.7B, 2.9B and 7.5B models from **XGLM** (Lin et al., 2022), and the 7B, 13B and 30B models from **LLaMA** (Touvron et al., 2023a). XGLM has a multilingual focus and covers many languages, but is smaller in size and lags behind recent models in English. In contrast, LLaMA is primarily trained on English and is much stronger in this language, while also showing some multilingual capabilities. Appendix B reports additional results for BLOOM (Scao et al., 2023), LLaMA 2 (Touvron et al., 2023b), OpenLLaMA (Geng and Liu, 2023), OpenLLaMA V2 (Geng and Liu, 2023), Redpajama (Computer, 2023) and PolyLM (Wei et al., 2023).

Methods. As shown in Figure 2, we compare two methods for each model: **direct** inference, where we feed the original (non-English) input to the model, and **self-translate**, where we first translate the input into English using the model itself, and then feed this translated input to the same model to solve the task. For translation, we do 4-shot

prompting using examples from the FLORES-200 dataset (Costa-jussà et al., 2022), prepending each sentence with its corresponding language name. We select the first sentences from the development set, skipping those that are longer than 100 characters. We use greedy decoding, and translate each field in the input (e.g., the premise and hypothesis in XNLI) separately. For analysis, we additionally compare self-translate to using an external state-of-the-art MT system. To that end, we use the 3.3B NLLB-200 model (Costa-jussà et al., 2022).

Evaluation. We use the following tasks for evaluation: **XCOPA** (Ponti et al., 2020), a common sense reasoning task in 11 languages; **XStoryCloze** (Lin et al., 2022), a common sense reasoning task in 11 languages; **XNLI** (Conneau et al., 2018), a natural language inference task in 15 languages; **PAWS-X** (Yang et al., 2019), a paraphrase identification task in 7 languages; and **MGSM** (Shi et al., 2022), a mathematical reasoning task with grade school problems in 11 languages. For MGSM, we do 8-shot evaluation with a chain-of-thought prompt, and extract the answer using a regular expression. The rest of the tasks are not generative, so we feed each candidate in a zero-shot fashion and pick the one with the highest probability.

Model	Size	Method	XStoryC	XCOPA	XNLI	PAWS-X	MGSM	Avg
XGLM	0.6B	Direct	53.5	54.9	39.4	48.4	1.7	39.6
		Self-translate	52.8 (-0.8)	53.4 (-1.5)	41.5 (+2.1)	50.6 (+2.2)	1.4 (-0.3)	39.9 (+0.3)
	1.7B	Direct	56.5	57.1	41.9	50.7	1.7	41.6
		Self-translate	55.9 (-0.6)	58.4 (+1.3)	44.9 (+3.0)	50.2 (-0.5)	1.7 (+0.0)	42.2 (+0.6)
	2.9B	Direct	58.2	58.5	43.0	50.8	1.4	42.4
		Self-translate	58.2 (+0.0)	62.5 (+4.0)	46.2 (+3.2)	53.2 (+2.4)	1.6 (+0.2)	44.3 (+1.9)
	7.5B	Direct	59.9	60.6	44.0	51.6	0.8	43.4
		Self-translate	60.9 (+1.0)	64.4 (+3.8)	48.9 (+4.9)	55.4 (+3.8)	0.1 (-0.7)	45.7 (+2.3)
LLaMA	7B	Direct	53.6	53.9	37.1	53.2	5.0	40.6
		Self-translate	55.8 (+2.2)	54.9 (+1.0)	43.0 (+5.9)	57.0 (+3.8)	6.1 (+1.1)	43.4 (+2.8)
	13B	Direct	54.8	54.7	34.2	49.5	7.4	40.1
		Self-translate	57.7 (+2.9)	56.5 (+1.8)	35.1 (+0.9)	52.1 (+2.6)	10.0 (+2.6)	42.3 (+2.2)
	30B	Direct	56.7	55.2	37.0	50.9	15.5	43.1
		Self-translate	59.0 (+2.3)	58.4 (+3.2)	43.5 (+6.5)	55.6 (+4.7)	16.3 (+0.8)	46.6 (+3.5)

Table 1: **Main results (accuracy)**. Task performance in terms of accuracy for different sizes of XGLM and LLaMA, using **direct** inference and **self-translate**. The last column shows the average accuracy over all tasks. We highlight the best results for each model and task in bold.

3 Results

Table 1 reports our main results, and Figure 1 visualizes the average accuracy of XGLM as a function of scale. Figure 3 compares the downstream performance and translation quality of self-translate and NLLB, grouped by low-resource and high-resource languages. Additional results are reported in Appendix B. We next summarize our main findings:

Self-translate outperforms direct inference. We find that self-translate works better than direct inference in average for all models. The results are also consistent across tasks, with only a few exceptions for the smaller XGLM models. This proves that multilingual language models are more capable than immediately obvious in non-English languages, but unveiling their full potential requires performing intermediate steps.

Multilingual language models do transfer capabilities across languages. One possible explanation for the previous finding is that language models acquire capabilities separately for each language, without any effective cross-lingual transfer. However, a closer comparison of LLaMA and XGLM refutes this hypothesis. In particular, we observe that LLaMA is much better than XGLM in MGSM despite being worse in other tasks. This is because MGSM is an emergent task (Wei et al., 2022), and XGLM, being smaller and less capable, obtains near 0 accuracy. In contrast, LLaMA is more capable at solving math word problems, and it is able to leverage this capability even if prompted

in other languages. The superior performance of self-translate shows that this cross-lingual transfer is not fully effective, but our results suggest that it does happen to a large extent.

Self-translate is more effective for high-resource languages and large models. Figure 1 shows that the gap between self-translate and direct inference gets larger at scale. Similarly, as shown by Table 1, it is the largest LLaMA model that obtains the biggest absolute gains over direct inference. At the same time, Figure 3 (top) shows that the effect of scale is bigger for high-resource languages and, for the largest model sizes, high-resource languages benefit more from self-translate than low-resource languages. This suggests that the effectiveness of self-translate is not explained by the limited capacity of smaller models, and can be expected to increase at scale.

MT outperforms self-translate, but the gap narrows at scale. As shown by Figure 3 (top), NLLB performs better than self-translate, meaning that it can still be beneficial to use an external MT system. However, the gap narrows at scale, as the translation capabilities of the largest models approach NLLB (Figure 3, bottom). Given the recent claims that state-of-the-art multilingual language models are competitive with traditional MT systems (Vilar et al., 2023; Hendy et al., 2023), this suggests that stronger language models would not require an external MT system for best results.