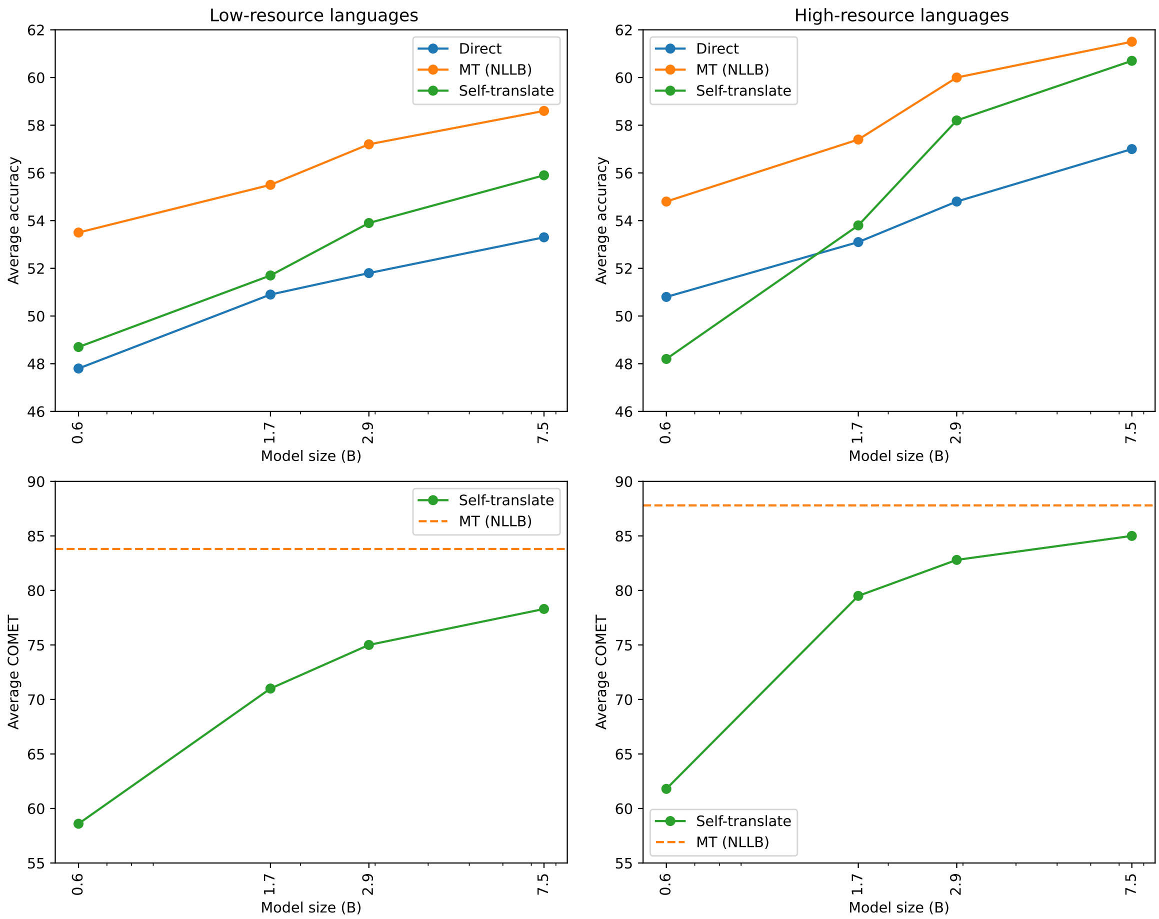

Figure 3: **Downstream (top) and MT (bottom) performance, grouped by low-resource (left) and high-resources (right) languages.** For downstream, we report average accuracy over XStoryCloze, XCOPA and XNLI, which have the most language variety. Low- and high-resource languages follow Lin et al. (2022), merging the low and ex-low categories. For MT, we report COMET (Rei et al., 2022), using the target language text for each field in those datasets as the source, and the English text as the reference.

4 Related work

Translate-test is a strong baseline in the traditional pretrain/finetune paradigm (Ponti et al., 2021; Artetxe et al., 2023). Early evidence shows that it is also effective for prompting autoregressive language models (Lin et al., 2022; Shi et al., 2022), as these models have irregular performance depending on the input language (Bang et al., 2023). Recent work has shown that multilingual language models are good translators (Zhang et al., 2023; Hendy et al., 2023; Vilar et al., 2023), which our approach exploits to replace the external MT system in translate-test. Concurrent to our work, Huang et al. (2023) propose a more complex prompting method that involves translating the input, but they only experiment with proprietary models and do not study the role of translation in isolation. Finally, Reid and Artetxe (2023) show that using synthetic parallel data from unsupervised MT can improve the performance of multilingual models, but they focus on pretraining seq2seq models.

5 Conclusion

We have proposed a new method called self-translate, where we use a multilingual language model to translate the test data into English, and then feed the translated data to the same model to solve the task. Self-translate consistently outperfoms the standard direct inference approach, which directly feeds the test data in the original language. Our approach does not involve any additional data or training, showing that language models are not able to leverage their full multilingual potential when prompted in non-English languages. In the future, we would like to explore training methods to mitigate this issue without the need of intermediate inference steps.

Limitations

Despite consistently outperforming direct inference, self-translate is substantially slower due to the cost of the translation step.

Our goal was to study a fundamental limitation of multilingual language models, and we decided to use base models to that end. In practice, instruction-tuned models would remove the need for few-shot prompts and make self-translate more efficient, as well as enabling to translate and solve the task in a single step.

Finally, all the datasets that we use were created through (human) translation, which can result in evaluation artifacts for methods involving machine translation (Artetxe et al., 2020). A more realistic scenario would be to use datasets that are natively written in different languages, but such datasets are scarce and not standard for evaluating autoregressive language models.

Acknowledgements

Julen is funded by a PhD grant from the Basque Government (PRE_2022_1_0047). This work is partially supported by projects founded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR (DeepR3 TED2021-130295B-C31, AWARE TED2021-131617B-I00, and DeepKnowledge PID2021-127777OB-C21), and the Basque Government (IXA excellence research group IT1570-22, IKER-GAITU 11:4711:23:410:23/0808 and NEL-GAITU).

References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv*.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. *arXiv*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv*.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Xinyang Geng and Hao Liu. 2023. Openllama: An open reproduction of llama.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. 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.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv*.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer. *arXiv*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Machel Reid and Mikel Artetxe. 2023. On the role of parallel data in cross-lingual transfer learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5999–6006, Toronto, Canada. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *arXiv*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study.

A Experimental details

In this section, we report additional experimental details that cover the evaluation library, task descriptions and prompts.

A.1 Evaluation library

We use LM Evaluation Harness (Gao et al., 2021) for evaluation. There were no multilingual tasks in the library, so we decided to add them so that our results can be replicated and extended to more models. For self-translate and MT, we define another evaluation task that uses a different dataset format. We created a fork of the evaluation library that includes these additional tasks at `https://github.com/juletx/lm-evaluation-harness/tree/translation`. All the translations generated with self-translate and MT are available at `https://huggingface.co/juletxara`.

A.2 Prompts

For self-translate and MT, we used the same English prompts used in XGLM to evaluate most tasks (Table 2). For direct inference, we use multilingual prompts, which are already available in some datasets (e.g. MGSM). When multilingual prompts are not available, we create them by translating English prompts to each language, using Google Translate. Note that this is suboptimal because translations are generally not as good as native prompts. Another option would be to always use English prompts, but this is also unnatural because it adds English tokens in the middle of other languages. All the multilingual prompts are available in the evaluation library above.