CAN LLMS LEARN A NEW LANGUAGE ON THE FLY? A CASE STUDY ON ZHUANG

Chen Zhang, Mingxu Tao, Quzhe Huang, Zhibin Chen, Yansong Feng*
Peking University
{zhangch,thomastao,huangquzhe,czb-peking,fengyansong}@pku.edu.cn

ABSTRACT

Existing large language models still fail to support many low-resource languages. Especially for the extremely low-resource ones, there is hardly any training data to effectively update the model parameters. We thus investigate whether LLMs can learn a new language on the fly through in-context learning prompting. To study this question, we collect a tiny parallel corpus for Zhuang, a language supported by no LLMs currently. We study the performance of various LLMs on the Zhuang-Chinese translation task and find out the great potential of this learning paradigm.

1 INTRODUCTION

Existing large language models (LLMs) provide robust support for many high-resource languages, but their support for numerous low-resource languages is limited. To adapt LLMs to low-resource languages, continual pre-training or adaptors are commonly employed (Pfeiffer et al., 2020; Yong et al., 2023). However, a corpus of merely a few thousand sentences is insufficient for extremely low-resource languages to update the model parameters effectively (Joshi et al., 2020). Considering the inductive and mimicking capabilities of current LLMs, an interesting research question arises: Can LLMs learn a new low-resource language on the fly solely through prompting? This learning paradigm could enable more efficient utilization of limited resources and holds significant potential in areas such as language preservation and education.

To explore this question, we chose Zhuang, an extremely low-resource language, as our focus. There are no open-source NLP datasets in Zhuang, and existing LLMs do not support this language. Therefore, we curated ZHUANGBENCH-BETA, a collection for Zhuang, comprising a dictionary, a parallel corpus, and a machine translation test set[1]. This resource aids in enhancing Zhuang's accessibility in NLP research. It also provides a convenient research suite for investigating how models can learn an entirely new language via prompts. See more information about the Zhuang language in Appendix A.

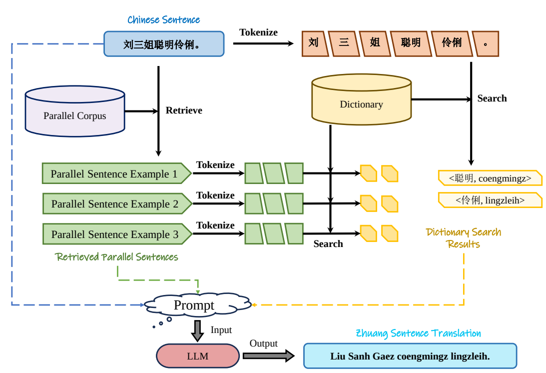

Figure 1: The framework of LLM-based translation by in-context learning.

We primarily focus on the translation tasks within on-the-fly language learning and model the task as in-context learning (ICL). We introduce word interpretations and similar sentence pairs in exemplars using the dictionary and parallel corpus. The LLMs are required to infer the morphological and syntactic rules from the examples in the prompt and organize the words into the final translation. This is a challenging task as it comprehensively evaluates the models' ability to follow instructions, extract rules and make inferences. Previous works have explored ICL for translation in high-resource languages (Ghazvininejad et al., 2023; Sia & Duh, 2023). and low-resource languages (Elsner & Needle, 2023; Tanzer et al., 2023),

*Corresponding author.
[1]Publicly available at https://github.com/luciusssss/ZhuangBench.

but the performance gains are modest, and relevant datasets are not open-sourced yet. By evaluating various models on ZHUANGBENCH-BETA, we provide several key findings on what scales or what types of models can learn a new language on the fly.

2 DATASET

We present ZHUANGBENCH-BETA, an NLP research suite for Zhuang. It consists of a Zhuang-Chinese dictionary, a Zhuang-Chinese parallel corpus, and a Zhuang-Chinese translation test set.

Dictionary. The Zhuang-Chinese dictionary is collected from an online dictionary site[2], with 16,031 Zhuang words. We also converted it to a Chinese-Zhuang dictionary with 13,618 Chinese words.

Parallel Corpus. The parallel corpus contains 3,587 Zhuang-Chinese sentence pairs collected from two sources. 2,135 pairs are obtained from the Chinese and Zhuang versions of the Government Work Reports in China. The remaining 1,452 pairs are collected from a Zhuang textbook.

Translation Test Set. To ensure the correctness of the evaluation samples, the machine translation test set is collected from the official Zhuang Language Proficiency Test (Vahcuengh Sawcuengh Suijbingz Gaujsi, V.S.S.G.) in China. In total, we collected 60 sentence pairs.

3 METHOD AND EXPERIMENT

Method. We cast the on-the-fly machine translation as an ICL task. For a Zhuang sentence to be translated into Chinese, we provide several exemplars in the prompt. Each exemplar is a Zhuang-Chinese sentence pair retrieved from our parallel corpus by BM25 (Robertson et al., 2009), together with the meaning of each word in the dictionary. We use a similar prompt for Chinese-to-Zhuang translation. See the framework in Figure 1 and the prompt example in Appendix B.2.

Model	zh2za		za2zh	
	BLEU	**chrF**	**BLEU**	**chrF**
LLaMA-2-7B-Chat	3.2	29.5	6.6	8.1
LLaMA-2-13B-Chat	4.1	31.9	9.3	10.2
LLaMA-2-70B-Chat	4.9	33.7	12.9	12.8
Baichuan-2-7B-Chat	3.1	30.7	12.8	12.9
Baichuan-2-13B-Chat	5.9	32.3	20.0	18.4
GPT-3.5-Turbo-1106	4.4	32.7	17.4	16.4
GPT-4-Turbo-1106	7.5	36.4	29.0	25.0

Table 1: 5-shot performance on the test set of ZHUANGBENCH-BETA. zh denotes Chinese and za denotes Zhuang.

Experiment Setup. We mainly use three types of models for experiments: (1) LLaMA-2-Chat (Touvron et al., 2023), an open-source English-centric model, (2) Baichuan-2-Chat (Yang et al., 2023), a bilingual model for English and Chinese, and (3) GPT-3.5-Turbo and GPT-4 (OpenAI, 2023), two commercial multilingual models. We use BLEU and chrF for metrics, implemented by Post (2018).

Results and Analyses. We report the main experiment results in Table 1. In terms of model scales, We observe that the performance is steadily improved with the increase of model parameters for LLaMA-2 and Baichuan-2. Since Baichuan-2 has a better Chinese capability than the English-centric LLaMA-2, a 13B Baichuan-2 model can outperform a 70B LLaMA-2. GPT-4 outperforms all the other models, demonstrating its excellent reasoning ability. It is worth noting that GPT-4 achieves 29.0 BLEU on Zhuang-to-Chinese translation, which is qualified for practical use.

4 CONCLUSION

In this paper, we investigate whether LLMs can learn a completely new language on the fly. Our experiment on the Zhuang language shows that current models can pick up the language quickly through proper ICL. Although still challenging for current LLMs, language learning through prompting shows great potential in adapting LLMs to low-resource languages. We hope that our ZHUANGBENCH-BETA can encourage more research on this topic.

[2] https://zha_zho.en-academic.com/

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Micha Elsner and Jordan Needle. Translating a low-resource language using GPT-3 and a human-readable dictionary. In Garrett Nicolai, Eleanor Chodroff, Frederic Mailhot, and Çağrı Çöltekin (eds.), Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 1–13, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sigmorphon-1.2. URL https://aclanthology.org/2023.sigmorphon-1.2.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. Dictionary-based phrase-level prompting of large language models for machine translation. arXiv preprint arXiv:2302.07856, 2023.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560.

OpenAI. Gpt-4 technical report, 2023.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL https://aclanthology.org/2020.emnlp-main.617.

Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval, 3(4):333–389, 2009.

Suzanna Sia and Kevin Duh. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. arXiv preprint arXiv:2305.03573, 2023.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning to translate a new language from one grammar book. arXiv preprint arXiv:2309.16575, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305, 2023.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 11682–11703, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.653. URL `https://aclanthology.org/2023.acl-long.653`.

A THE ZHUANG LANGUAGE

Zhuang is a group of Kra–Dai languages spoken by the Zhuang people of Southern China in the province of Guangxi and adjacent parts of Yunnan and Guangdong. It is used by more than 16 million people. The current official writing system for Zhuang is the Latin script. Zhuang is considered an isolating language with few inflectional morphology. In our work, we focus on Standard Zhuang, the official standardized form of the Zhuang language.

B IMPLEMENTATION DETAILS

B.1 TOKENIZATION

The Zhuang language adopts a Latin script. Although not tailored for the Zhuang language, the default tokenizers of LLaMA-2 and Baichuan can tokenize Zhuang texts into subwords, or characters at least. They will not produce UNK tokens. For example, for the sentence *Liuz Sanhcej coengmingz lingzleih.*, the tokenizer of LLaMA-2 outputs [‘_Li’, ‘uz’, ‘_San’, ‘h’, ‘cej’, ‘_co’, ‘eng’, ‘ming’, ‘z’, ‘_ling’, ‘z’, ‘le’, ‘ih’, ‘.’].

B.2 PROMPT

In Table 2 we provide an example of the prompt for Chinese-to-Zhuang translation.

C ADDITIONAL EXPERIMENTS

C.1 NON-LLM BASELINES

Here we provide results for a commonly-used non-LLM baseline based on mT5 (Xue et al., 2021), a series of multilingual sequence-to-sequence models. We finetune mT5 with the 3.6K parallel sentences and report the results on the test set in Table 3. With only a few thousand parallel sentences, the non-LLM baseline can hardly model an unseen language or learn the mapping between a high-resource language and the unseen one. This result further demonstrates the advantage of adopting LLMs for understanding extremely low-resource languages.

C.2 ORDER OF WORD SENSES

Furthermore, we want to highlight the importance of using dictionaries properly. We find that LLMs fail to disambiguate the different senses of a word with limited context and that they often use the first sense provided for translation. A simple strategy of sorting the senses according to their frequencies in the corpus helps. For example, this strategy improves the performance of Baichuan-2-13B-Chat by 1.6 BLEU on Zhuang-to-Chinese translation.

请仿照样例，参考给出的词汇，将汉语句子翻译成壮语。 *(Please follow the example and refer to the given vocabulary to translate the Chinese sentences into Zhuang.)* ## 请将下面的汉语句子翻译成壮语：好。明天你就要回去了，今天晚上我让我妻子弄几个菜，咱们喝两杯。*(Please translate the following Chinese sentence into Zhuang: OK. You're going back tomorrow. I'll ask my wife to prepare some dishes tonight and we'll have a drink.)* ##在上面的句子中，汉语词语"好"在壮语对应的词是"ndei"或"baenz"；汉语词语"明天"在壮语对应的词是"ngoenzcog"或"ngoenzbyug"； ... *(In the above sentence, the Chinese word "good" corresponds to the Zhuang word "ndei" or"baenz"; the Chinese word "tomorrow" corresponds to the Zhuang word "ngoenzcog" or "ngoenzbyug";...)* ## 所以，完整的壮语翻译是：Ndei. Ngoenzcog mwngz couh yaek baema lo, haemhneix gou heuh yah gou loengh geij yiengh byaek, raeuz ndoet song cenj. *(So, the complete Zhuang translation is: Ndei. Ngoenzcog mwngz couh yaek baema lo, haemhneix gou heuh yah gou loengh geij yiengh byaek, raeuz ndoet song cenj.)* *(More exemplars here)* ## 请将下面的汉语句子翻译成壮语：现在，农村许多老年人年轻人会使用手机来买卖商品，很方便。*(Please translate the following Chinese sentence into Zhuang: Nowadays, many elderly and young people in rural areas use mobile phones to buy and sell goods, which is very convenient.)* ##在上面的句子中，汉语词语"现在"在壮语对应的词是"seizneix"或"neix"； ... *(In the above sentence, the Chinese word "nowadays" corresponds to the Zhuang word "seizneix" or"neix"; ...)* ## 所以，完整的壮语翻译是： *(So, the complete Zhuang translation is:)*

Table 2: An example of the prompt for Chinese-to-Zhuang translation.

Model	zh2za		za2zh	
	BLEU	**chrF**	**BLEU**	**chrF**
mT5-base	0.1	8.7	0.6	1.5
mT5-large	0.8	16.6	0.7	2.8

Table 3: Performance of mT5 on the test set of ZHUANGBENCH-BETA. zh denotes Chinese and za denotes Zhuang.