AI FORUM

Year: 2025 | Edition: 14 |
Cambodia | 9 May 2025 |

INSIGHTS

Machine Translation: Bridging Khmer Language and Large Language Models (LLMs)

BUOY Rina[1], PhD (Informatics)

CHENDA Sovisal[2], MA (Advanced Computing Systems)

TAING Nguonly[3], PhD (Computer Science)

KONG Marry[4], PhD (Telecommunications Technology and Information Technology)

[1] ***BUOY Rina*** *is a Research Fellow, AI & Linguistics, AI Forum*
[2] ***CHENDA Sovisal*** *is an Applied Artificial Intelligence Associate Researcher, Techo Startup Center*
[3] ***TAING Nguonly*** *is an Executive Director, Techo Startup Center*
[4] ***KONG Marry*** *is a Secretary of State, Ministry of Economy and Finance*

Abstract

While the development of large language models (LLMs), specifically for the Khmer language, is constrained by both computational resources and the availability of quality training data, this paper presents a strategic direction highly relevant to artificial intelligence and machine learning engineers and developers working with the Khmer language. This paper highlights the significant role of high-quality machine translation in advancing Khmer natural language processing (KNLP). It argues that improving machine translation between Khmer and high-resource languages is currently the most viable approach to fully leverage LLMs and other advanced models for the Khmer language, at least in the short term.

Introduction

While large language models (LLMs) have significantly transformed the landscape of natural language processing (NLP) in a narrow sense, and the entire field of artificial intelligence in a broader sense, LLMs such as Phi-3 (Abdin et al., 2024) and Llama-3 (Grattafiori et al., 2024) are limited to only a few high-resource languages such as English. In addition, open-source LLMs such as Sea-LLMs (Nguyen et al., 2024) and Sea-lion (Ong & Limkonchotiwat, 2024) are still brittle in their support for the Khmer language, while the proprietary models such as ChatGPT and Claude are too expensive to utilize due to an unfavorable tokenization scheme, not to mention performance issues. Training LLMs specifically on high-quality Khmer training data can enhance performance and reliability; nonetheless, this effort is constrained by a lack of data, computational resources, and skilled engineers. This article presents a strategic direction highly relevant to artificial intelligence and machine learning engineers and developers working with the Khmer language. This article argues that advancing machine translation between Khmer and other high-resource languages is the only viable option to fully leverage LLMs and other advanced models for the Khmer language, at least in the short term.

Machine Translation

Machine translation (MT) is defined as the use of digital computers to translate from a source language to a target language. MT is commonly used to provide information access, reducing the digital divide (Jurafsky & Martin, 2024). High-quality MT allows speakers of low- resource languages to access vast amounts of information that would otherwise be available only in

high-resource languages. While languages are structurally similar as a means of communication, they can differ idiosyncratically and systematically (Jurafsky & Martin, 2024). These divergences present challenges for high-quality MT. Moreover, translation is an asymmetric challenge: translating from English into Khmer is relatively less challenging than the reverse.

The attention-based encoder-decoder neural network architectures (Bahdanau, & Bengio, 2014; Vaswani et al., 2017), illustrated in Figure 1, have been a standard modelling architecture for MT tasks. The encoder encodes the source sentence while the decoder decodes the target sentence one word or token at a time. During decoding, the decoder utilizes cross-attention mechanisms to selectively attend to relevant inputs before generating the next word. To train an MT model, a large parallel corpus is required. While large parallel corpora are available for high-resource languages, such corpora are scarce for low-resource languages such as Khmer. To overcome this data challenge, multilingual MT models can be efficiently fine-tuned via transfer learning on limited Khmer parallel corpora. Alternatively, multilingual LLMs can be further instruction-fine-tuned to perform MT tasks for the Khmer language, thanks to their vast amount of general pretraining knowledge, which may include Khmer content. To turn an LLM into a Khmer-English machine translator, we can utilize the following training prompt template: <user>ខ្ញុំទៅសាលារៀន។<assistant>I go to school.<eos>.

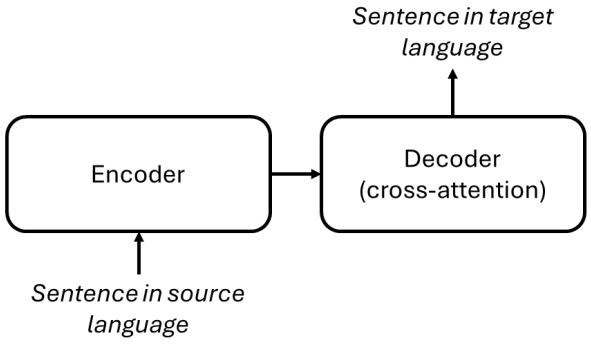

Figure 1: Attention-based encoder-decoder MT architecture. The model takes an input sentence in a source language and maps it to an output sentence in a target language.

MT-Embedded LLMs and Use Cases for Khmer Language

With high-quality MT capability for the Khmer language, it is possible to fully leverage English-dominated LLMs or any powerful English-based multimodal artificial intelligence (AI) models when dealing with Khmer text, as illustrated in Figure 2. For example, to summarize a long Khmer article into a few sentences while preserving semantic meaning, we can leverage a state-of-the-art (SOTA) text summarization model, such as BART (Lewis et al., 2019), which is not available for the Khmer language. The article is first translated from Khmer into English using a Khmer-English MT model, and the translated text is then summarized using the text summarization model. The summarized text is finally translated back from English to Khmer using an English-Khmer MT model. The quality of the resulting Khmer summarization depends on two factors: the text summarization model and the MT model. The same architecture proposed in Figure 2 can also be utilized for various other downstream applications, such as a rural healthcare chatbot system where a user in a remote area can get real-time health-related consultation in Khmer.

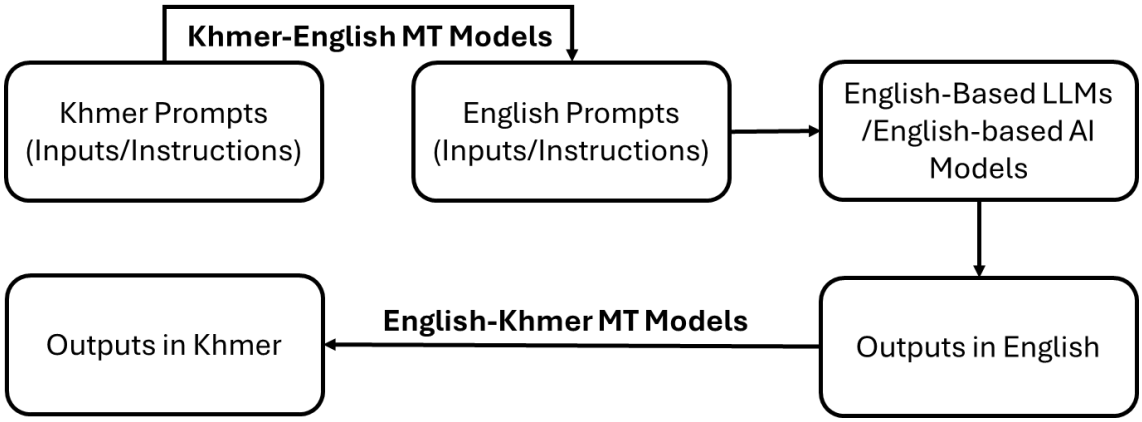

Figure 2: MT-Embedded LLMs for Khmer Language. In the proposed architecture, the user inputs in Khmer are first translated into English for the English-based LLMs/models. The English outputs are then translated back into Khmer to users.

In addition, by utilizing the proposed workflow illustrated Figure 2, it is possible to unlock a range of powerful features of LLMs for the Khmer language, including personal assistants, content creation, and NLP tasks such as semantic search, retrieval-augmented generation (RAG), text summarization, and reading comprehension. It also opens doors for Khmer speakers to utilize and interact with other advanced visual-language models (VLMs), such as Florence-2 (Xiao et al., 2024), which understand both images and text simultaneously. A visual

question-answering (VQA) task is an example of visual-language tasks performed by VLMs. In a VQA task, users input image(s) and a text prompt, and the VLM model completes the prompt using its visual understanding of the input image(s). An example of MT-embedded VLMs for Khmer Language is illustrated in Figure 3. The same architecture proposed in Figure 3 can also be utilized for various other downstream applications, such as a visual cultural question-answering system where for example, a user provides inputs in terms of cultural images and questions, the system can generate answers based on the provided images and questions.

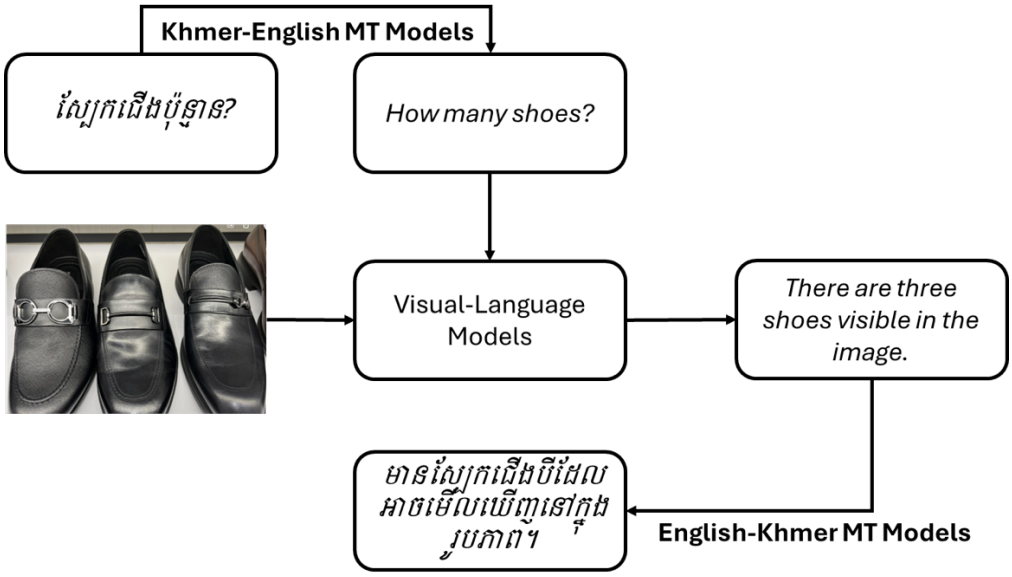

Figure 3: MT-Embedded VLMs for Khmer Language. In the proposed architecture, the user inputs in Khmer are first translated into English for the English-based LLMs/models. The English outputs are then translated back into Khmer to users.

Conclusion

While the lack of data, computational resources, and skilled engineers poses a challenge for developing Khmer-dedicated LLMs or advanced AI models, high-quality MT for the Khmer language can be viewed as a bridge between Khmer and advanced LLMs. By removing the language support barrier, it enables Khmer speakers to utilize and interact with LLMs and other advanced AI models. Such utilization and interaction can lead to a range of creative and entrepreneurship-oriented AI applications for the Khmer language. Nonetheless, to achieve high-quality MT for the Khmer language, future research should focus on improving

performance of the MT models, and establish open-source, diverse, standard, high-quality MT training and evaluation datasets for the Khmer language.

Reference

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D.,… Zhou, X. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv*. https://arxiv.org/abs/2404.14219

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. https://doi.org/10.48550/arXiv.1409.0473

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., AI-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Kinsvark, A.,… Ma, Z. (2024). The llama 3 herd of models. arXiv:2407.21783

Jurafsky, D., & Martin, J. H. (2024). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed., draft). Stanford University. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. https://arxiv.org/abs/1910.13461

Nguyen, X.-P., Zhang, W., Li, X., Aljunied, M., Hu, Z., Shen, C., Chia, Y. K., Li, X., Wang, J., Tan, Q., Cheng, L., Chen, G., Deng, Y., Yang, S., Liu, C., Zhang, H., & Bing, L. (2023). *SeaLLMs--Large Language Models for Southeast Asia*. arXiv. https://doi.org/10.48550/arXiv.2312.00738

Ong, D. T.-W., & Limkonchotiwat, P. (2023). Southeast Asia LLMs: SEA-LION and Wangchan-LION. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)* (p. 246). Association for Computational Linguistics. https://aclanthology.org/2023.nlposs-1.26.pdf

Vaswani, A. Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. arXiv:1706.03762

Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng M., Liu, C., & Yuan, L. (2024). Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4818-4829)*.

** The opinions expressed are the authors' own and do not reflect the views of AI Forum.*