１　研究目的、研究方法など

本研究計画調書は「小区分」の審査区分で審査される。記述に当たっては、「科学研究費助成事業における審査及び評価に関する規程」（公募要領参照）を参考にすること。
本研究の目的と方法などについて、４頁以内で記述すること。
冒頭にその概要を簡潔にまとめて記述し、本文には、(1)本研究の学術的背景や本研究の着想に至った経緯、研究課題の核心をなす学術的「問い」、(2)本研究の目的および学術的独自性と創造性、(3)関連する国内外の研究動向と本研究の位置づけ(4)本研究で何をどのように、どこまで明らかにしようとするのか、(5)本研究の目的を達成するための準備状況について具体的かつ明確に記述すること。

（概要）

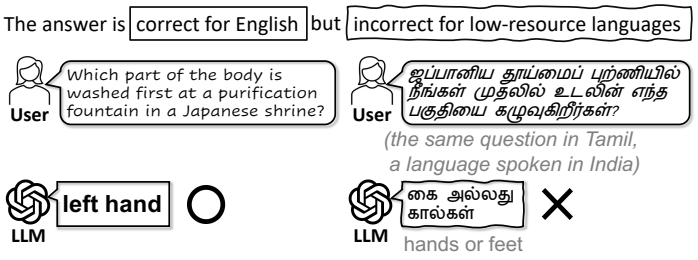
Large language models (LLMs) such as ChatGPT have achieved human-level performance for tasks written in English. However, the performance remains unsatisfactory for low-resource languages, which constitute the majority of 7,164 languages in the world. Past work improves the performance of specific languages by expensive data curation, making it hard to generalize to new languages.

This research aims a language-agnostic way to improve the performance of LLMs for low-resource languages. In our past experiments, we found the performance gap largely due to the difference of writing systems. Leveraging this observation, in this research we transfer knowledge from high-resource languages by surface-level script mapping and deep-level semantic representation mapping methods. Through this research, we enhance the performance of LLMs for long-tail, low-resource languages, and ultimately enable non-English speakers to benefit from the advancements in LLMs.

（本文）

① 本研究の学術的背景、着想に至った経緯、課題の核心をなす学術的「問い」

本研究の学術的背景



Large language models (LLMs) such as ChatGPT (OpenAI, 2024) have made remarkable progress across various tasks like question answering and machine translation. Nevertheless, current LLMs are English-centric. We found that they perform well on English queries but poorly on the same queries written in low-resource languages with limited publicly available data (Publication [9]).

What's more, there are only about 370 million people in the world (4.9% of the global population) who speak English as their native language. As a result, the majority of people in the world are unable to fully benefit from the convenience and productivity offered by the current English-centric LLMs.

While previous research (Llama Team, 2024) built multilingual LLMs for 8 languages, most languages in the world remain unsupported. Moreover, most of past research was language-specific and difficult to generalize to low-resource languages with diverse writing systems. Therefore, we aim to explore language-agnostic solutions to improve LLMs for low-resource languages.

本研究の着想に至った経緯

I have been working on low-resource language processing (Publication [1-8]) since master's course. During past research, we noticed that low-resource languages can benefit from related high-resource languages if we align their writing systems. We found many Japanese Kanji originated from Chinese Hanzi. Based on this, we mapped Chinese characters in Chinese sentences to the corresponding Japanese characters (such as 风→風) and created synthetic Japanese sentences. Results show that the

【１　研究目的、研究方法など（つづき）】

synthetic Japanese data improves Japanese↔English translation quality (Publication [3]).

We have been investigating the limitation of LLMs. In a collaboration project with MBZUAI, we tested LLMs on identical multi-choice questions written in English and 26 other languages, where I contributed to the creation of Japanese dataset. Results demonstrate a clear accuracy gap between questions in English and other languages (Publication [9]). This reveals that the knowledge stored in English data is inaccessible when queries are in other languages. Based on our script mapping work, we believe aligning scripts across different languages is a promising way toward multilingual LLMs.

However, language-agnostic alignment approaches are not explored yet, because mapping rules are limited to specific language groups such as between Japanese and Chinese, or among Indian languages (Hindi, Tamil, Bengali, etc.). Additionally, our past methods focus on surface-level alignment which are not effective enough. We are also interested in deeper-level representation alignment which focuses on aligning the semantic meanings of words across languages through their token embeddings or the activation patterns of neurons within each layer of LLMs.

研究課題の核心をなす学術的「問い」

Most factual knowledge in LLMs is embedded in the training data of high-resource languages, mainly in English. Our previous experiments show that the knowledge is inaccessible when the input queries are in low-resource languages because of script differences. To address this, we pose the question: How to transfer knowledge to low-resource languages with diverse writing systems in LLMs?

② 研究目的、独自性と創造性

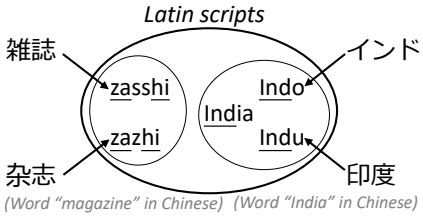
研究目的

The goal is to improve the performance of LLMs for low-resource languages with diverse writing systems in a language-agnostic way, where we transfer knowledge from high-resource languages by surface-level script alignment and deep-level semantic representation alignment approaches.

Ultimately, our goal is to enable non-English speakers to enjoy the convenience brought by the advancements in LLMs. Through this research, people speaking other languages can ask questions in their native languages and receive accurate and satisfactory answers from LLMs.

独自性

In our previous research, we applied script mapping from Chinese to Japanese and observed that enhancing the script similarity is a promising way of knowledge transfer (Publication [3]). Taking this one step further, we propose novel language-agnostic approaches to explore not only surface-level script similarity but also deep-level semantic representation similarity.



Romanization improves script similarity.

創造性

We creatively explore two script alignment and two semantic representation alignment approaches:

1. The script Romanization approach improves script similarity by mapping low-resource languages into the Latin script, which is semi-automated and linguistically motivated.
2. The pure byte-level inputs approach maps all languages into the same script space represented by bytes sequences, and captures fine-grained information compared to subword-based methods.

【１　研究目的、研究方法など（つづき）】

3. The word embedding mapping approach directly transfers semantic meaning between languages in the embedding layer, irrespective of differences in their writing systems.
4. The neuron activation pattern alignment approach aligns the hidden representations within intermediate layers of the model, which has not been studied before.

③ 関連研究の動向、本研究の位置付け

国内外の研究動向

Existing works to improve the LLMs performance for low-resource languages include extending the size of vocabulary to cover words in more languages, creating a more balanced training dataset across languages, and retrieving similar examples in high-resource language in the prompt during inference.

However, the vast number of languages with diverse writing systems makes vocabulary extension inefficient as it increases model parameter. While data curation is effective, it is expensive thus hard to generalize to new languages. And the improvement is limited without further training or fine-tuning.

本研究の位置付け

To address these, we explore (semi-)automated, language-agnostic and effective ways. Our approaches include both surface-level and deep-level knowledge transfer, where we first convert words to Latin characters, bytes, and embeddings automatically and then align them across various languages. We show the efficiency by combining the methods and train a small-scale LLM from scratch.

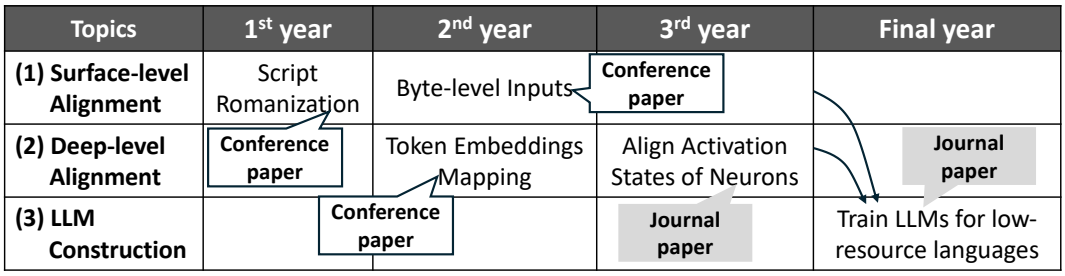
④ 本研究で何をどのように、どこまで明らかにしようとするのか

Our plan to transfer knowledge from high-resource to low-resource languages in LLMs includes:

(1) Two surface-level script alignment methods,

(2) Two deep-level alignment methods, and

(3) Integrating (1) and (2) to train an LLM and evaluate it on various low-resource languages.



(1) Surface-level alignment (2025/4–2026/9)

The research on surface-level alignment focuses on increasing the script similarity of low-resource and high-resource (primarily English) languages in the inputs. We first explore mapping sentences from all languages into Latin subword sequences in (1-a) and mapping into byte sequences in (1-b).

(1-a) Script Romanization (2025/4–2025/12)

Increasing script similarity with high-resource languages can improve the performance of low-resource languages (Publication [3]). Rather than relying on mapping rules of specific language pairs or language families, we plan to convert all languages into Latin script semi-automatically using input method editors. For example, the Japanese word "雑誌" will be mapped to "zasshi" and the Chinese word "杂志" to "zazhi", where there are four shared characters after conversion. What's more, the Japanese–Chinese character mapping table is not needed anymore. This kind of phonetic similarity is common across related languages, and Romanization converts phonetic similarity into script similarity,

【１　研究目的、研究方法など（つづき）】

which can be leveraged by LLMs. During the training of LLMs, the knowledge in high-resource language data will be encoded to tokens through back-propagation. These tokens are then shared with the Romanized low-resource languages, which will be better understood by LLMs during inference.

(1-b) Pure byte-level inputs (2026/1–2026/9)

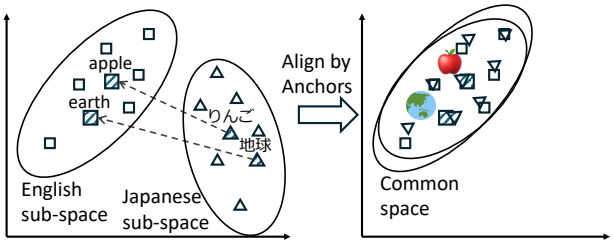
We aim for a fully automated script mapping approach by converting data in all languages into sequences of bytes. All languages are forced to share the same set of scripts that are 256 types of bytes. Furthermore, by discarding the subword tokenization, we do not need to expand the vocabulary as the number of languages increases, which saves model parameter. To analyze the effectiveness and efficiency of this method, we measure the degree of alignment by calculating the similarity of semantic representations of parallel sentences in various languages. Secondly, byte inputs results in longer inputs thus more calculation steps and require model to capture longer-term dependency. We alleviate these by replacing transformer architecture to Mamba architecture (Gu and Dao, 2024).

(2) Deep-level alignment (2026/10–2028/3)

In additional to aligning scripts, we plan to directly align the semantic meaning of words in the embedding layer (2-a) or in the activation pattern of neurons (2-b).

(2-a) Token embedding mapping (2026/10–2027/6)

This method aims to directly transfer meaning of words stored in embeddings of high-resource languages to low-resource languages. We first obtain seed word pairs using dictionaries or from web-crawled data. The seed words have similar semantic meanings across languages. Using seed words as anchors, we map the word embedding space of low-resource languages to that of high-resource languages such that all other non-anchor words will be roughly aligned automatically. As not all knowledge is from English data, we may need to combine this with surface-level alignment methods.



Token Embedding Space Before and After Mapping

(2-b) Align activation pattern of neurons (2027/7–2028/3)

Knowledge is not only stored in the first embedding layer, but also reflected in the pattern of how neurons are activated in the following layers. Motivated by this, we plan to analyze the activation states of neurons given the inputs in high-resource and low-resource languages and align the pattern.

If there is no clear activation pattern, we plan to align the sentence representations in the final layer.

(3) Integrating surface-level and deep-level methods to build a multilingual LLM (2028/4–2029/3)

We plan to train a multilingual LLM covering about 200 languages with 7B parameters from scratch, where we combine surface-level and deep-level alignments. With knowledge transfer, our model can achieve better performance on low-resource languages compared to LLMs of similar scales.

⑤ 目的を達成するための準備状況

Training LLMs is one of the foundational yet challenging part. We have successfully prompted and fine-tuned LLMs in various tasks (Publication [9, 14, 15]). There is an ongoing IndicLLM project with Indian Institutes of Technology (IIT), where I am building the pipeline from data curation to LLM training. I also have experience in low-resource language processing (Publications [1-9]), input tokenization (Publications [10-12]). Therefore, the initial implementation of this research will be smooth.

２　応募者の研究遂行能力及び研究環境

応募者の研究計画の実行可能性を示すため、(1)これまでの研究活動（主要な研究業績を含む）、(2)研究環境（研究遂行に必要な研究施設・設備・研究資料等を含む）について２頁以内で記述すること。
「(1)これまでの研究活動」の記述には、研究計画に関連した国際的な取組（国際共同研究の実施歴や海外機関での研究歴等）がある場合には必要に応じてその内容を含めること。また、研究活動を中断していた期間がある場合にはその説明などを含めてもよい。

① これまでの研究活動

The applicant has been working on topics related to this research: 1) low-resource languages 2) token representation, and 3) large language models. All publications listed here are peer-reviewed.

■ Our past research covers diverse languages including Japanese [1, 2, 3, 8], languages in Europe [4], low-resource African languages [5, 9], and Indian languages [6, 7], where we improve similarities in related languages [3], construct corpus [1, 8, 9] and pretrain or train models [2, 4, 5, 7].

[1] **Haiyue Song**, Raj Dabre, Atsushi Fujita, and Sadao Kurohashi. Coursera Corpus Mining and Multistage Fine-Tuning for Improving Lectures Translation. LREC 2020, Pages 3640–3649, (2020.5).

[2] Zhuoyuan Mao, Fabien Cromieres, Raj Dabre, **Haiyue Song**, and Sadao Kurohashi. JASS: Japanese-specific Sequence to Sequence Pre-training for Neural Machine Translation. LREC 2020, 3683-3691 (2020.5).

[3] **Haiyue Song**, Raj Dabre, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, and Eiichiro Sumita. Pre-training via Leveraging Assisting Languages for Neural Machine Translation. ACL SRW 2020, Pages 279–285, (2020.7).

[4] Zhuoyuan Mao, Chenhui Chu, Raj Dabre, **Haiyue Song**, Zhen Wan, and Sadao Kurohashi. When do Contrastive Word Alignments Improve Many-to-many Neural Machine Translation? Findings of NAACL, Pages 1766–1775 (2022.7).

[5] Francois Meyer, **Haiyue Song**, Abhisek Chakrabarty, Jan Buys, Raj Dabre, and Hideki Tanaka. NGLUEni: Benchmarking and Adapting Pretrained Language Models for Nguni Languages. LREC-COLING 2024, pages 12247–12258. **The best paper award at AfricaNLP Workshop 2024.** (2024.5).

[6] Abhisek Chakrabarty, **Haiyue Song**, Raj Dabre, Hideki Tanaka, and Masao Utiyama. Incorporating Hypernym Features for Improving Low-resource Neural Machine Translation. KEMT Workshop (2024.6)

[7] **Haiyue Song**, Raj Dabre. NICT's Cascaded and End-To-End Speech Translation Systems using Whisper and IndicTrans2 for the Indic Task. IWSLT, pages 17–22 (2024.8). **Ranking 1st out of 4 teams.**

[8] **Haiyue Song**, Raj Dabre, Chenhui Chu, Atsushi Fujita, and Sadao Kurohashi. Bilingual Corpus Mining and Multistage Fine-Tuning for Improving Machine Translation of Lecture Transcripts. Journal of Information Processing, Volume 32 Pages 628–640 (2024.8).

[9] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, ..., **Haiyue Song**, ..., and Alham Fikri Aji. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. (NeurIPS 2024 **under review**).

■ The applicant has a deep understanding of input/output token representation in language models and published works on linguistically motivated subword segmentation [10, 11], representation alignment of diverse segmentations [12], and decoding algorithms [13].

[10] **Haiyue Song**, Raj Dabre, Zhuoyuan Mao, Chenhui Chu, and Sadao Kurohashi. BERTSeg: BERT Based Unsupervised Subword Segmentation for Neural Machine Translation. AACL, Pages 85–94 (2022.11).

[11] **Haiyue Song**, ..., and Eiichiro Sumita. SelfSeg: A Self-supervised Sub-word Segmentation Method for Neural Machine Translation. ACM Trans. Asian Low-Resour. Lang. Inf. Process, Vol.22, Art. 215 (2023.8).

[12] **Haiyue Song**, Zhuoyuan Mao, Raj Dabre, Chenhui Chu, and Sadao Kurohashi. DiverSeg: Leveraging Diverse Segmentations with Cross-granularity Alignment for Neural Machine Translation. Journal of Natural

【２　応募者の研究遂行能力及び研究環境（つづき）】

Language Processing, Volume 31 Issue 1 Pages 155–188 (2024.4).

[13] **Haiyue Song**, Francois Meyer, Raj Dabre, Hideki Tanaka, Chenhui Chu, and Sadao Kurohashi. SubMerge: Merging Equivalent Subword Tokenizations for Subword Regularized Models in Neural Machine Translation. The 25th Annual Conference of the European Association for Machine Translation (2024.6).

■ Past research improves LLMs on relation extraction [14] and machine translation [15] tasks.

[14] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, **Haiyue Song**, Jiwei Li, and Sadao Kurohashi. GPT-RE: In-context Learning for Relation Extraction using Large Language Models. EMNLP, 3534–3547 (2023.12).

[15] Raj Dabre, **Haiyue Song**, Miriam Exel, Bianka Buschbeck, Johannes Eschbach-Dymanus, and Hideki Tanaka. How Effective is Synthetic Data and Instruction Fine-tuning for Translation with Markup using LLMs? Accepted to the Conference of the Association for Machine Translation in the Americas (2024.9).

■ Besides, the applicant is familiar with model architectures thus implements new ideas quickly.

[16] **Haiyue Song**, Xiang Song, Tianjian Li, Hao Dong, Naifeng Jing, Xiaoyao Liang, and Li Jiang. A FPGA Friendly Approximate Computing Framework with Hybrid Neural Networks (Abstract Only). In Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Page 286 (2018.2).

[17] **Haiyue Song**, Chengwen Xu, Qiang Xu, Zhuoran Song, Naifeng Jing, Xiaoyao Liang, and Li Jiang. Invocation-driven Neural Approximate Computing with a Multiclass-classifier and Multiple Approximators. In Proceedings of the International Conference on Computer-Aided Design, Page 50 (2018.11).

[18] Li Jiang, Zhuoran Song, **Haiyue Song**, Chengwen Xu, Qiang Xu, Naifeng Jing, Weifeng Zhang, and Xiaoyao Liang. Energy-Efficient and Quality-Assured Approximate Computing Framework Using a Co-Training Method. ACM Transactions on Design Automation of Electronic Systems, Pages 59:1–59:25 (2019.11).

[19] Zhuoyuan Mao, Raj Dabre, Qianying Liu, **Haiyue Song**, Chenhui Chu, and Sadao Kurohashi. Exploring the Impact of Layer Normalization for Zero-shot Neural Machine Translation. ACL, Pages 1300–1316 (2023.7)

[20] Weiqi Gu, **Haiyue Song**, Chenhui Chu, and Sadao Kurohashi. Spatial Hierarchical Attention Network Based Video-guided Machine Translation. Journal of Information Processing, Vol. 31 Pages 299–307 (2023.5).

■ International collaborations include works with SAP in Germany [15], the University of Cape Town in South Africa [5, 13], MBZUAI in Abu Dhabi [9], and ongoing one on IndicLLM with IIT in India.

■ Furthermore, the applicant has 8 non-peer-reviewed publications in domestic conferences, two tutorials [21, 22], and one patent in progress "BERTSeg: BERT Based Subword Segmentation". Part of these have been conducted under the Research Fellowships for Young Scientists DC1. Volunteer activities include co-organizing the Workshop on Asian Translation, mentoring students in the AACL 2020, and reviewing journal (TASLP, TALLIP) and conference (ACL, EMNLP, NAACL, ...) papers.

[21] **Haiyue Song**, Hour Kaing, and Raj Dabre. Linguistically Motivated Neural Machine Translation. (Tutorial) In the 25th Annual Conference of the European Association for Machine Translation (2024.6).

[22] Aditya Joshi, ..., and **Haiyue Song**. Connecting Ideas in 'Lower-Resource' Scenarios: NLP for National Varieties, Creoles, and Other Low-resource Scenarios. (Tutorial) Accepted to COLING 2025.

② 研究環境

Training LLMs requires computational resource and data. The institute provides the applicant 104 V100 GPUs, 40 A100 GPUs, and H100 GPUs will be available soon, and high-quality multilingual data. Therefore, based on the applicant's capabilities and environment, the research plan is fully feasible.

３　人権の保護及び法令等の遵守への対応（公募要領参照）

本研究を遂行するに当たって、相手方の同意・協力を必要とする研究、個人情報の取扱いの配慮を必要とする研究、生命倫理・安全対策に対する取組を必要とする研究など指針・法令等（国際共同研究を行う国・地域の指針・法令等を含む）に基づく手続が必要な研究が含まれている場合、講じる対策と措置を、１頁以内で記述すること。
個人情報を伴うアンケート調査・インタビュー調査・行動調査（個人履歴・映像を含む）、提供を受けた試料の使用、ヒト遺伝子解析研究、遺伝子組換え実験、動物実験など、研究機関内外の倫理委員会等における承認手続が必要となる調査・研究・実験などが対象となる。
該当しない場合には、その旨記述すること。

本項目に該当しない。