



The Datasets of Human and AI Translation

DATA PAPER

YOKE LIAN LAU

SHIAW PHIN CHEE

RUTH HUI HUI CHUA

ZI HONG YONG

ING KET YONG

JEE CHIN TAN

HUI WEN YONG

ANNA LYNN ABU BAKAR

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The datasets outline a methodical approach for comparing translations from Mandarin to Malay completed by humans and AI. The datasets contain a framework with rubrics for a keyword detection template to discover common words used by both humans and AI in their translation tasks. The datasets include a Mandarin poem originally written in Mandarin, along with translations by a belt-and-road project translator, a Mandarin native speaker translator, and a Malay native speaker translator categorised as human translators. It also contains translations by ChatGPT 3.5 with 1 prompt, and ChatGPT 4.0 with four different prompts. All these prompts are also listed in the dataset.

CORRESPONDING AUTHOR:

Anna Lynn Abu Bakar

Centre for the Promotion of
Knowledge and Language
Learning, Universiti Malaysia
Sabah, 88400 Kota Kinabalu,
Sabah, Malaysia

annalynn@ums.edu.my

KEYWORDS:

Human translation; AI
translation; keyword detection;
framework; Mandarin

TO CITE THIS ARTICLE:

Lau, Y. L., Chee, S. P., Chua, R. H. H., Yong, Z. H., Yong, I. K., Tan, J. C., Yong, H. W., & Bakar, A. L. A. (2024). The Datasets of Human and AI Translation. *Journal of Open Humanities Data*, 10: 45, pp. 1–10. DOI: <https://doi.org/10.5334/johd.212>

(1) CONTEXT AND MOTIVATION

(1.1) AI TRANSLATION AND HUMAN TRANSLATION

The present article examines and contrasts the translation methodologies employed by human translators and AI translators. Human translators often consist of a Chinese native speaker and translators with expertise in the Belt and Road translation project. The research findings indicate that the current performance of ChatGPT translation is either comparable to or inferior to that of human translators. The statistical analysis indicated that there is no substantial disparity in the quality of the categories. Nevertheless, it is a formidable tool that would eventually surpass the level of translation quality attained by human translators (Lau et.al, 2024).

AI translation may miss textual and cultural details despite its speed and cost. Human translators are best for exact and culturally sensitive translations since they can accurately convey the original text's meaning and tone. (Mohammed Moneus & Sahari, 2023). AI appears to pose a threat to human translators, but it offers a more innovative approach and creates opportunities for collaboration with humans in the field of translation (Sundberg & Holmström, 2024). Both AI translators and human translators possess their own strengths. The strengths of human translators lie in their ability to create effective communication across diverse languages and cultures. However, these strengths are influenced by the distinct cognitive demands of different interpreting styles (Liu, Y & Liang J, 2024). The relationship between human and machine translation must be reevaluated as AI continues to develop. The interaction between human and machine translation is complex, and technology is thoroughly integrated into the translation process. The translation industry is in the process of changing, but machine translation continues to be constrained, and human translation continues to be subject to subjectivity. The future of machine and human translation, suggests a more equitable and liberated dynamic. Translators must maintain their own subjectivity and value while utilising technology to uncover truth to establish a successful partnership between humans and technology (Lin, 2023).

Lin (2023) highlights the constraints of machine translation (MT) algorithms, which become evident when trying to express comprehension in linguistic form. In the age of machine translation, it is essential for interpretation education to prioritise the importance of cultivating specialised abilities to efficiently enable contact between interpreters and individuals, thereby guaranteeing the preservation of the profession. Moreover, AI-powered technology has the capacity to aid those with cognitive impairments (Almufareh et al., 2023). The potential of AI-assisted human interpretation is evident; however, additional research is necessary to identify an approach that is both efficient and ergonomically optimal (Almufareh et al., 2023).

In Sun's (2021) believes that machine translation still falls short of the proficiency exhibited by expert translators in their manual work. In comparison to the substantial demand in the translation market, the available number of translators can be considered insignificant. Machine translation engines have revolutionised translation in multilingual organisations and global enterprises, as they provide a distinct edge over human translation. According to Yu (2024), the broad adoption of AI-based translation technology can be attributed to advancements in computer and AI technologies. This method employs computer technology and AI to analyse unprocessed data, allowing for independent translation.

(1.2) WORD DETENTION IN TRANSLATION

One of the most significant and straightforward methods for determining the accuracy of translation is the word detention method (Zhili & Qian, 2024). Wong & Kit (2009), Anahita (2015), Livia Kelebercová & Fero Forgac (2022), Kumar & Sahula (2022) have used the word detention method in their research. Given the abundance of words, it is possible to identify the exact same word in both the source and target languages. Lin & Mitamura (2004), among the earliest studies to investigate the general issue of keyword translation in a multilingual open-domain question-answering system. Their research demonstrates that the accuracy of keywords can be enhanced by utilizing multiple machine translation systems and the query sentence to perform sense disambiguation. Kuang et al. (2022) creates experiments to illustrate a model which is more effective at capturing local keyword information than the current state-of-the-art models. The method of improving relation classification by capturing keyword features of entity relation dependencies is effective. Zhili & Qian (2024) conclude that the semantic similarity determination approach is believed to enhance the translation

system’s comprehension and processing of keywords in the source language, hence optimizing the translation process. When encountering polysemous words or words with various meanings, one can enhance the accuracy and naturalness of translation by selecting the most contextually appropriate translation based on the semantic similarity of keywords.

McKellar & Puttkammer (2020) provide a dataset that can be used to evaluate the accuracy of machine translation between any pair of the 11 South African languages. These data are available for use by any researcher specialising in machine translation of South African languages. Having a single aligned evaluation set will enable accurate comparison of any future machine translation systems for South African languages. Data labelling enables the comparison of important technologies and applications in natural language processing across different languages. The main distinction between the McKellar & Puttkammer (2020) dataset and our datasets is that we offer a detailed, systematic guide for generating AI translations and human translations for poems. In contrast, the McKellar & Puttkammer dataset consists of translations for 500 source sentences, which were produced by a different professional human translator.

(1.3) EXPLANATION OF THE DATASETS

The datasets utilized in this inquiry were acquired from the research conducted by Lau et al. (2024). The datasets were created with the purpose of facilitating the reuse of templates for undertaking comparative translation tasks between human translators and AI translation systems within the same study topic. They were easily accessible and reusable and could also be utilized for other language translation tasks.

The design was also specifically intended to aid academics in analyzing and conducting studies on translation activities performed by both human translators and AI (Lau et al., 2024). A prior analysis utilizing this dataset showcased the vital function of human translators in the translation sector and underscored the deficiencies of AI translators that necessitate future rectification. These files include unprocessed data for anybody interested in doing a comparative comparison between human and AI translation.

However, due to the limited availability of data, it is used as pilot research for this project. This article includes two datasets. Figure 1 shows the two datasets and their respective files. The initial dataset comprises a PDF document of a poem written by Fang Mei in Mandarin, translations by three human translators, and translations by four different ChatGPT prompts. The second dataset is a PDF file with rubrics and templates utilized for keyword recognition to evaluate human and AI translations. The file name for Dataset 1 is “POEM HUMAN AND

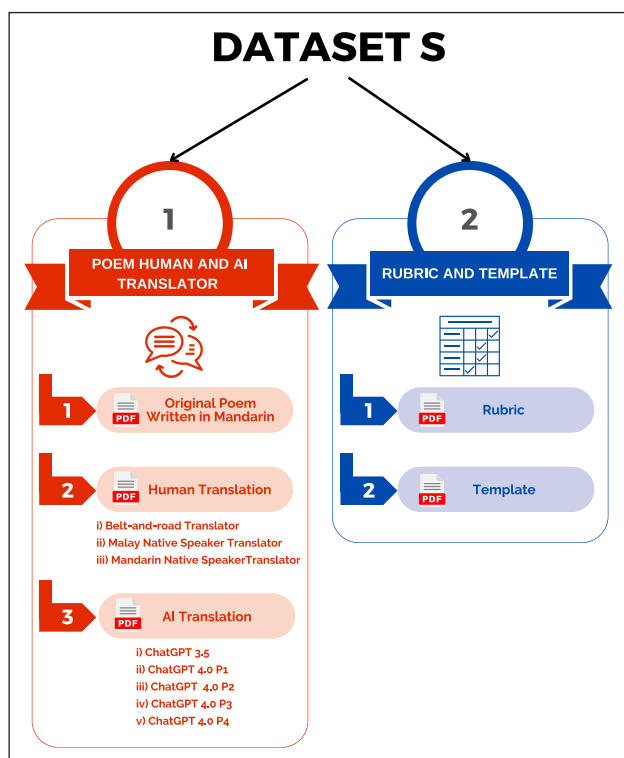


Figure 1 Datasets and their accompanying information.

AI TRANSLATOR” (Lau, 2024a) and for Dataset 2 (Lau, 2024b) is “RUBRIC AND TEMPLATE.” Dataset 1 consists of 9 tables. Table 1 presents the original poem in Mandarin by Fang Mei. The poem consists of 9 stanzas, with each stanza containing one sentence. Table 2 shows the translation done by the Belt and Road translator, Table 3 shows the translation by a Malay native speaker, Table 4 shows the translation by a Chinese native speaker, Table 5 shows the translation by ChatGPT 3.5, Table 6 shows the translation by ChatGPT 4.0 with prompt 1, Table 7 shows the translation by ChatGPT with prompt 2, Table 8 shows the translation by ChatGPT with prompt 3, and the final table shows the translation by ChatGPT with prompt 4. Table 1 presents the comprehensive data from dataset 1.

DETAIL OF DATA	TABLE 1	TABLE 2	TABLE 3	TABLE 4	TABLE 5	TABLE 6	TABLE 7	TABLE 8	TABLE 9
Total sentence	9	9	9	9	9	9	9	9	9
Written language	Mandarin	Malay	Malay	Malay	Malay	Malay	Malay	Malay	Malay
Translator	Original Write (Author)	Belt and Road	Malay native speaker	Chinese native speaker	ChatGPT 3.5	ChatGPT 4.0 P1	ChatGPT 4.0 P2	ChatGPT 4.0 P3	ChatGPT 4.0 P4

(2) DATASET DESCRIPTION

Table 1 Displays the detailed information and data from dataset 1.

REPOSITORY LOCATION

Dataset 1: Data Mendeley: <https://data.mendeley.com/datasets/vc5wc8rymx/1>;

DOI: [10.17632/vc5wc8rymx.1](https://doi.org/10.17632/vc5wc8rymx.1)

Dataset 2: Data Mendeley: <https://data.mendeley.com/datasets/s6bx8wyvwg/1>;

DOI: [10.17632/s6bx8wyvwg.1](https://doi.org/10.17632/s6bx8wyvwg.1)

REPOSITORY NAME

[DataMendeley.com](https://data.mendeley.com)

OBJECT NAME

1. POEM HUMAN AND AI TRANSLATOR,
2. RUBRIC AND TEMPLATE

FORMAT NAMES AND VERSIONS

PDF

CREATION DATES

2023-06-21 TO 2024-03-25

DATASET CREATORS

Yoke Lian Lau (Conceptualization; methodology; investigation; data curation; writing original draft preparation and visualization)

LANGUAGE

Malay, Mandarin, English

LICENSE

CC BY Attribution 4.0 International

PUBLICATION DATE

2024.03.25

(3) METHOD

The study gathered translation works from three human translators at Universiti Malaysia Sabah, with the translation work of a Malay native speaker serving as the benchmark for comparison. The focus of the study revolved around a 9-sentence poem that was originally written in Mandarin. The researchers collected 27 translated sentences from a total of three human translators. Five ChatGPT prompts were assembled for the purpose classifying as an AI translator. A total of 45 translated sentences were collected by the study team.

Steps – The Mandarin version of the poem is the original composition by Fang Mei in Mandarin. A Belt and Road translator is a translator who is involved in translating project related to the One Belt and One Road China Malaysia Children Literature initiative. A Malay native speaker was a Malaysian Malay translator proficient in both Mandarin and Malay. A Chinese native speaker was a translator who is a native speaker of the Chinese language. Three of them were from Universiti Malaysia Sabah. All three translators were categorized as human translators. ChatGPT 3.5 is the free version of ChatGPT that utilizes prompts to generate translations. ChatGPT 4.0 P1, P2, P3, and P4 are distinct versions of ChatGPT that require payment and offer various translated works based on different prompts. The prompts, created by a professional Malaysia publisher chief editor, can be utilized by other researchers. It has been tested and proven to be beneficial and efficient.

The original poetry written in Mandarin is provided solely for reference purposes. The objective of the research is to compare the Malay translation produced by human translators with AI translators. Therefore, to simplify the process, future study might refrain from consulting the original Mandarin poems contained in the datasets. The original poetry written in Mandarin is provided solely for reference purposes. The objective of the research is to compare the Malay translation produced by human translators with AI translators. Therefore, to simplify the process, future study might refrain from consulting the original Mandarin poems contained in the datasets.

Table 2 displays the five prompts utilized by the group of AI translations.

	PROMPT	HOW TO USE THE PROMPT
ChatGPT 3.5	Translate this poem into Malay	After copying and pasting the original poetry into ChatGPT 3.5, proceed to copy and paste this prompt into ChatGPT 3.5 Chabot.
ChatGPT 4.0 P1	Translate this poem into Malaysia Malay	After copying and pasting the original poetry into ChatGPT 4.0, proceed to copy and paste this prompt into ChatGPT 4.0 Chabot.
ChatGPT 4.0 P2	Translate this poem into Malaysia Malay with a focus on capturing the essence and style appropriate for Malay language expression	After copying and pasting the original poetry into ChatGPT 4.0, proceed to copy and paste this prompt into ChatGPT 4.0 Chabot.
ChatGPT 4.0 P3	Translate this children's poem into Malaysia Malay with an interpretive approach, preserving its poetic essence and ensuring it aligns with the Malay cultural context	After copying and pasting the original poetry into ChatGPT 4.0, proceed to copy and paste this prompt into ChatGPT 4.0 Chabot.
ChatGPT 4.0 P4	Translate this children's poem into Malaysia Malay using poetic language and the Malay way of thinking, interpretatively translate it, and then embellish the poem	After copying and pasting the original poetry into ChatGPT 4.0, proceed to copy and paste this prompt into ChatGPT 4.0 Chabot.

Table 2 ChatGPT and its Prompts.

SAMPLING STRATEGY

The study utilized ChatGPT versions 3.5 and 4.0 to construct five distinct ChatGPT translations. This was achieved by applying five unique prompts. The suggestions were provided by an experienced and proficient chief editor from a reputable Malaysian publisher.

The research team developed a framework for evaluating and contrasting human and AI translation endeavors using the keyword detection method. The template was equipped with flexible and customizable criteria to accommodate the diverse requirements and research goals of different researchers.

(4) IMPLICATIONS/APPLICATIONS

Dataset 2 has two tables. One is a rubric, while the other is a template used to compare human and AI translation endeavors. The rubric details the process for using the template. The template shows the analysis translation work. Researchers can use the template by filling in the columns and using their own translation efforts. The rubrics can be easily modified to suit the researcher's needs, making them customizable and flexible. Table 3 displays the rubrics utilized. Other researchers are welcome to modify or elaborate on it with more detail and use the symbol that is more often utilized. For example: Replace the word “v” with “x” in the Malay native speaker's translation check.

NO	EXPLANATION
RUBRIC TO IDENTIFY THE KEYWORDS.	
1	<ul style="list-style-type: none"> a. The word that is the same as the Malay native speaker's translation check v. b. The phrase that is the same as the Malay native speaker's version check v. c. Roof term (without regard to tenses, etc.) that is the same as the Malay native speaker's translation will check v. d. The tenses or gramma differences will be listed after the “v” in brackets (). e. The sequence of word/phrase difference but appear in the same sentence identical as the Malay native speaker's translation check “v” but after “v” will insert the word in {}.
2	The score is calculated using the formula $\text{total v} \div \text{total words/segment} \times 100\%$.
3	The sentence with the highest score will be highlighted in grey.

Table 3 Rubric.

The template is designed to be very user-friendly and easily accessible. Table 4 displays the template for contrasting the aims of human and AI translation efforts.

SENTENCE 1	KAMU	ADALAH	PARI-PARI	DI	DALAM	AIR
L&A	v		bidadari	v	v	v
CNS	v	ialah	dongeng	v	v	v
CGPT 3.5	Anda	v	peri	v	v	v
CGPT 4.0 P1	Engkau		bidadari	v	v	v
CGPT 4.0 P2	Engkau		peri air	yang	lembut	
CGPT 4.0 P3	Engkau		peri cantik	v	v	v
CGPT 4.0 P4	Engkau	ibarat	Peri yang menari	v	v	v

Table 4 Template.

Sentence 1 serves as the primary sentence utilized as a model or target for translation efforts. This template utilized the version provided by a proficient Malay linguist as the benchmark for the translation task. Arrange each phase or word in an individual column. A six-word message will be shown in six columns. The column below sentence 1 presents the classifications of translator. If the research includes five translators, then five columns will be utilized. The translation efforts of seven translators in this template will be compared with the translation efforts of a native Malay speaker. A checkmark will be added in the column if the keyword chosen matches that of the Malay native speaker. If the translators L&A and CNS utilize the same keyword as the Malay native speaker, a checkmark will be added below the term “Kamu” for both translators L&A and CNS. Researchers can conduct an experiment by comparing two or three translated versions of a basic poem to assess the effectiveness of the template.

(5) RESULTS AND DISCUSSION

The potential of AI research lies in its ability to facilitate the generation of novel applications for the amplification mechanism. To accomplish this, it is imperative that we possess a comprehensive comprehension of AI's capacity to comprehend both substance and context. Effectively managing and utilising AI in both digital and real-world contexts require meticulous preparation and consideration. This is crucial for leveraging LLMs to enhance AI development and ensure its reliability.

(5.1) “GOOD” AND “POOR” TRANSLATION

Li (2024) research shows that several Neural Machine Translation (NMT) systems will be capable of carrying out a range of uncomplicated translation tasks. LLMs have recently demonstrated significant potential due to advancements in AI technology. This implies that they have the potential to surpass NMT in terms of translation performance. LLMs do not surpass Neural Machine Translation (NMT) in terms of translation accuracy. All the selected systems exhibit superior performance in non-literary translation compared to literary translation. The use of these strategies yields superior results in Chinese-English translations compared to English Chinese translations.

LLMs remain the most sophisticated form of artificial intelligence technology due to their ability to not only translate diverse texts, but also perform a wide array of other tasks, surpassing the capabilities of NMT systems.

The results from Lin (2023) indicated that human interpreters possess certain abilities that enable them to achieve effective communication across diverse languages and cultures. However, these abilities are influenced by certain cognitive requirements that are unique to different interpreting modalities.

In this study, the translations that have been accepted by the publisher and published serve as the exemplar or goal for translating into the target language. The translation, whether done by humans or AI, that achieves a better score (Lau et al., 2024) in the comparison can be deemed as a “good” translation. To enhance the user-friendliness of the framework provided in the datasets, future researchers should prepare a published translation, a human translation, and an AI translation for the purpose of comparison.

(5.2) POTENTIAL ETHICAL CONSIDERATIONS

We concur with the findings of Essel et al. (2024). To ensure the responsible implementation of artificial intelligence (AI) in education, it is crucial to comprehend the ethical ramifications and drawbacks associated with it. Subsequent investigations could explore the impact of socio-cultural and personal variables on creative cognition by utilising ChatGPT. This may entail assessing the cognitive aptitude for innovative thinking among pupils with diverse socio-cultural and personal backgrounds in comparison to individuals who have utilised ChatGPT. An attractive research issue involves comparing Language Models such as Google Bard in educational environments to identify the most effective AI solutions for specific educational contexts. All the datasets are exclusively owned by the researcher and their team. However, it is free to reuse, but it is necessary to ensure that the shared data and template are not abused or misused.

(5.3) AI FOR TRANSLATION IN CULTURALLY SENSITIVE CONTEXT

AI translation has the capability to translate culturally sensitive words, but it requires a human translator to cross-check and ensure that the translated version aligns with the desired quality and adheres to the rules and political stance of the country. If we overly rely on AI translation, human translators will lose the ability to perceive the nuances in this context.

(5.4) CHALLENGE OF WORKING WITH MALAY TRANSLATIONS

The study by Bakarola & Nasriwala (2021) demonstrates the application of attention processes in encoder-decoder models for sequence-to-sequence learning on a bilingual corpus of Sanskrit-to-Hindi translations. Utilising distinct networks for encoding and decoding the source and target

vectors, respectively, enhances the combined conditional probability. The attention mechanism demonstrates encouraging outcomes by directing its emphasis towards specific components of the provided Sanskrit sentences. In 2020, Wilhelmina Nekoto and her colleagues introduced a participatory approach to broaden NLP research to languages with limited resources. Upon identifying the primary contributors to machine translation (MT) advancement, they establish an inclusive African MT community. Their analysis uncovered highly effective techniques for cultivating and synchronising growth, communicating information, and developing models. Furthermore, they demonstrate how the collaborative design of the community enables researchers to evaluate model outputs through human assessment, while also providing benchmarks and datasets for languages that have not been well studied. This addresses an issue pertaining to NLP algorithms that have limited resources.

Anuraj and Goutam (2024) assert that machine translation, which is a component of natural language processing (NLP), helps to surmount linguistic barriers. In their work, they research optimal methodologies, fostering impactful translation initiatives for underprivileged language populations.

Malay is also classified as a low-resource language when compared to Chinese. Engaging in translation duties will lead to a decrease in the overall performance of ChatGPT. This issue necessitates a comprehensive analysis due to the disparity in resource availability between the source language and the target language, characterised by their varying levels of resources. Certain intricate words may necessitate the aid of a human translation for rectification. Nevertheless, even though the datasets are sourced from a book of children's literature and the publisher requires the use of uncomplicated and readily comprehensible language, ChatGPT is nevertheless capable of doing translation work of exceptional quality. NLP has aided in addressing the challenge of low-resource languages in the field of AI.

(5.5) CONCLUSION

While the current progress in AI offers great possibilities for innovation, effectively utilising this potential requires cautious supervision. It is necessary to have human supervision and emphasise cooperative endeavours between humans and artificial intelligence. By considering these issues, companies and communities can employ AI to foster substantial and morally sound progress (Sundberg & Holmström, 2024). Ensuring a harmonious collaboration between artificial intelligence and humans is crucial, whether in the field of translation or any other domain. One can utilise the formidable proficiency of AI, yet it is unwise to place complete reliance on AI for everything. It is important to always keep in mind that AI is simply a tool designed to assist us in simplifying tasks, rather than allowing it to exert control over us. Without AI, however, we would not lose our ability to think altogether. Humans must be vigilant. Ensure that in the event of AI's disappearance, human translators are still able to rely on their own cognitive abilities to do translations. According to Yu's (2024) research findings, MT has gained dominance in recent years, supplanting the conventional rule-based MT approach.

Our study has determined that, although machine translation technology is being used more often, there is a statistically significant disparity in the quality of translations compared to human translation. To improve the quality and efficiency of translation systems, it is crucial to leverage the advantages provided by artificial intelligence technology in combination with human translation.

ACKNOWLEDGEMENTS

Special thanks to Dr Chen Eng Chia, a chief editor of Malaya Press who provided the translations ChatGPT prompts.

FUNDING INFORMATION

The research was funded by China Ningbo Publishing House, grant number: TLA2104. The researchers belong to the Language and Multilinguistic Research Group at Centre for The Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Author contributions Conceptualization, ALAB and YLL; methodology, ZHY and YLL; formal analysis, ZHY and ALAB; investigation, YLL and ZHY; resources, ALAB and ZHY; data curation, YLL; writing original draft preparation, YLL, JCT and ZHY; writing review and editing, IKY, RHH, SPC, HWY, JCT and ALAB; visualization, ZHY and YLL; supervision, ALAB. All authors have read and agreed to the published version of the manuscript.

AUTHOR AFFILIATIONS

Yoke Lian Lau  orcid.org/0000-0002-7949-7683

Centre for the Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

Shiaw Phin Chee  orcid.org/0000-0003-4382-4870

Centre for the Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

Ruth Hui Hui Chua  orcid.org/0000-0003-4364-8758

Centre for the Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

Zi Hong Yong  orcid.org/0000-0002-0531-2420

Centre for the Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

Ing Ket Yong  orcid.org/0000-0003-0865-2751

Centre for the Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

Jee Chin Tan  orcid.org/0009-0001-3932-0955

Centre for the Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

Hui Wen Yong  orcid.org/0009-0007-6948-8605

Centre for the Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

Anna Lynn Abu Bakar  orcid.org/0000-0002-2540-3637

Centre for the Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

REFERENCES

- Almufareh, M. F., Tehsin, S., Humayun, M., & Kausar, S. (2023). Intellectual disability and technology: An artificial intelligence perspective and framework. *Journal of Disability Research*, 2(4). DOI: <https://doi.org/10.57197/jdr-2023-0055>
- Anahita, M. B. (2015). *Word Alignment for Statistical Machine Translation Using Hidden Markov Models*. [Report, Simon Fraser University].
- Anuraj, B., & Goutam, M. (2024). A case study on tools and techniques of machine translation of Indian low resource languages. In P. Pakray, P. Dadure & S. Bandyopadhyay (Eds.), *Empowering low-resource languages with NLP solutions* (pp. 51–85). IGI Global. DOI: <https://doi.org/10.4018/979-8-3693-0728-1.ch004>
- Bakarola, V., & Nasriwala, J. (2021). Attention-based neural machine translation approach for low-resourced Indic languages—A case of Sanskrit to Hindi translation. *Smart Systems: Innovations in Computing*, 565–572. DOI: https://doi.org/10.1007/978-981-16-2877-1_52
- Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers and Education: Artificial Intelligence*, 6, 100198. DOI: <https://doi.org/10.1016/j.caeai.2023.100198>
- Kuang, H., Chen, H., Ma, X., & Liu, X. (2022). A keyword detection and context filtering method for document level relation extraction. *Applied Sciences*, 12(3), 1599. DOI: <https://doi.org/10.3390/app12031599>
- Kumar, R., & Sahula, V. (2022). Word translation using cross-lingual word embedding: Case of Sanskrit to Hindi translation. In *2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*. DOI: <https://doi.org/10.1109/aisp53593.2022.9760564>

- Lau, Y. L., Yong, Z. X., Chia, C. E., Yong, Z. H., Abu Bakar, A. L., Ku, C. J., Nasir, E., & Arumugam, B. (2024). Comparing translation accuracy in Belt and Road Malaysia children's literature: Malay and Chinese native speakers vs ChatGPT. *Forum for Linguistic Studies*, 6(1), 2069. Retrieved from <https://ojs.acad-pub.com/index.php/FLS/article/view/v6i1.2069>. DOI: <https://doi.org/10.59400/fls.v6i1.2069>
- Lau, Y. L. (2024a). *Poem human and AI translator* [Data set]. Mendeley Data. DOI: <https://doi.org/10.17632/vc5wc8rymx.1>
- Lau, Y. L. (2024b). *Rubric and template* [Data set]. Mendeley Data. DOI: <https://doi.org/10.17632/s6bx8wyvwg.1>
- Li, X. (2024). Comparison of translation quality between large language models and neural machine translation systems: A case study of Chinese-English language pair. *International Journal of Education and Humanities*, 4(2), 121–128. DOI: [https://doi.org/10.58557/\(ijeh\).v4i2.213](https://doi.org/10.58557/(ijeh).v4i2.213)
- Lin, F., & Mitamura, T. (2004). Keyword translation from English to Chinese for multilingual QA. *Machine Translation: From Real Users to Research*, 164–176. DOI: https://doi.org/10.1007/978-3-540-30194-3_19
- Lin, Y. (2023). The relationship between machine translation and human translation in the era of artificial intelligence machine translation. *Applied and Computational Engineering*, 5(1), 133–138. DOI: <https://doi.org/10.54254/2755-2721/5/20230547>
- Liu, Y., & Liang, J. (2024). Multidimensional comparison of Chinese-English interpreting outputs from human and machine: Implications for interpreting education in the machine-translation age. *Linguistics and Education*, 80, 101273. DOI: <https://doi.org/10.1016/j.linged.2024.101273>
- Kelebercová, L., & Forgac, F. (2022). Keyword extraction for automatic evaluation of machine translation.
- McKellar, C. A., & Puttkammer, M. J. (2020). Dataset for comparable evaluation of machine translation between 11 South African languages. *Data in Brief*, 29, 105146. DOI: <https://doi.org/10.1016/j.dib.2020.105146>
- Mohammed Moneus, A., & Sahari, Y. (2023). Artificial intelligence and human translation: A contrastive study based on legal texts. DOI: <https://doi.org/10.2139/ssrn.4441379>
- Sun, W. (2021). Integration of machine translation and manual translation in translation practice based on artificial intelligence and big data technology. In *2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture* (pp. 1895–1897). DOI: <https://doi.org/10.1145/3495018.3495510>
- Sundberg, L., & Holmström, J. (2024). Innovating by prompting: How to facilitate innovation in the age of generative AI. *Business Horizons*. DOI: <https://doi.org/10.1016/j.bushor.2024.04.014>
- Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohunge, T., Akinola, S. O., Muhammad, S. H., Kabongo, S., Osei, S., Freshia, S., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Meressa, M., Adeyemi, M., Mokgesi-Seling, M., Okegbemi, L., ... Martinus, L. J. (2020). Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. *Findings of the Association for Computational Linguistics: EMNLP*, 2144–2160. DOI: <https://doi.org/10.18653/v1/2020.findings-emnlp.195>
- Wong, B., & Kit, C. (2009). ATEC: Automatic evaluation of machine translation via word choice and word order. *Machine Translation*, 23(2–3), 141–155. DOI: <https://doi.org/10.1007/s10590-009-9061-x>
- Yu, Y. X. (2024). Application of translation technology based on AI in translation teaching. *Systems and Soft Computing*, 6, 200072. DOI: <https://doi.org/10.1016/j.sasc.2024.200072>
- Zhili, W., & Qian, Z. (2024). A deep learning-based method for determining semantic similarity of English translation keywords. *International Journal of Advanced Computer Science and Applications*, 15(5). DOI: <https://doi.org/10.14569/ijacsa.2024.0150531>

TO CITE THIS ARTICLE:

Lau, Y. L., Chee, S. P., Chua, R. H. H., Yong, Z. H., Yong, I. K., Tan, J. C., Yong, H. W., & Bakar, A. L. A. (2024). The Datasets of Human and AI Translation. *Journal of Open Humanities Data*, 10: 45, pp. 1–10. DOI: <https://doi.org/10.5334/johd.212>

Submitted: 29 March 2024

Accepted: 17 July 2024

Published: 31 July 2024

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.