Issues Surrounding the Use of ChatGPT in Similar Languages: The Case of Malay and Indonesian

Hiroki Nomoto
Tokyo University of Foreign Studies
nomoto@tufs.ac.jp

Abstract

We report a problem that one faces when using ChatGPT in similar languages, taking Malay and Indonesian as examples: ChatGPT often responds to prompts in Malay (the language with fewer speakers) in Indonesian (the language with more speakers). We examined ChatGPT's identification (LangID) ability to find out whether this language choice problem arises from LangID errors. The results show that LangID errors alone cannot explain the problem's severity. By comparing the patterns of responses to Malay prompts and those to Javanese prompts, we conclude that the problem happens mainly because ChatGPT does not treat Malay and Indonesian equally as distinct languages. Rather, it behaves as if Malay were a non-standard variety of Indonesian. We also discuss social issues the language choice problem causes and possible solutions to them.

1 Introduction

As ChatGPT and other similar generative AIs become increasingly popular, their problems and limitations have come to be known and shared widely by the society. While some are general and relevant to all users, others are specific to a particular user group. The latter issues can remain unnoticed and overlooked because of their particularity. The present study discusses one such issue, namely wrong language choices involving similar languages, specifically Malay and Indonesian. The problem is that ChatGPT tends to provide responses to prompts in Malay in Indonesian, the language with more speakers and hence more data. We maintain that this technological problem in turn can lead to social issues. We make some concrete proposals to alleviate them.

The rest of the paper is organized as follows. Section 2 briefly explains the relation between Malay and Indonesian. We provide concrete examples of the language choice problem mentioned above in section 3. Then, we explore the possibility that language identification (LangID) failures cause the problem in section 4. Section 5 discusses social issues stemming from the language choice problem. Section 6 concludes the paper.

2 Malay and Indonesian

Linguistically, Malay (ISO693-3 zsm) and Indonesian (ISO693-3 ind) are two standard regional varieties of the same language, namely the macrolanguage Malay (ISO693-3 msa), which encompasses all Malay varieties in the Malay Archipelago. Note that the language name "Malay" is ambiguous. It may refer to the macrolanguage Malay (msa) or one of its varieties (zsm). In this paper, we use "Malay" to refer to the latter.

Malay is the official language of Malaysia and Brunei and one of the four official languages of Singapore. Indonesian is the official language of Indonesia. The numbers of speakers are approximately 32 million for Malay and 270 million for Indonesian. Malay and Indonesian share a significant portion of their lexicons and grammars. Asmah Haji Omar (2001) estimates the lexical difference is about 10%. Syntactic differences also exist although they are not as evident as lexical ones (see, e.g., Nomoto and Kartini Abd. Wahab, 2011).

3 Language Choice Problems

Language choice problems have been around with NLP applications in Malay and Indonesian for a long time. Machine translation applications still confuse Malay with Indonesian, despite their unprecedented improvement witnessed in recent years. For example, English-to-Malay translations sometimes contain Indonesian expressions. Crucially, the problem happens at the lexical level and rarely at the sentence level. That is, it is unlikely to receive a translation completely in Indonesian, unless the input is a very short sentence or phrase consisting only of a couple of words.

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 76–82
November 1–4, 2023.

The language choice problem with ChatGPT is far more serious. It is common that the response to a prompt in one language is given in the other language. The problem normally happens at the sentence level. It is worth noting that the problem occurs predominantly with prompts written in Malay. Therefore, Indonesian users will not even notice the problem.

To show how serious the problem is, we tested 100 arbitrary prompts in Malay.[1] The prompts were unambiguously written in Malay as much as possible. Their domains are varied, including science, health, history, politics, language, literature and religion. We used the free version of ChatGPT (May 3). As many as 66 responses were provided in Indonesian, as in (1), whereas 31 were provided in Malay. There were also three responses containing elements of both languages, as in (2). The italicized items with subscripts M and I are unique to Malay and Indonesian, respectively.

(1) **User:** Apakah *perbezaan*$_M$ bahasa Melayu dengan bahasa Indonesia? (What are differences between Malay and Indonesian?)
ChatGPT: Bahasa Melayu dan bahasa Indonesia memiliki banyak persamaan *karena*$_I$ *keduanya*$_I$ berasal dari bahasa Melayu-Polinesia dan memiliki sejarah yang *terkait*$_I$. Namun, ada beberapa *perbedaan*$_I$ dalam kosa kata, ejaan, *tata bahasa*$_I$, dan *pengucapan*$_I$ antara *keduanya*$_I$. ... (Malay and Indonesian have many similarities because both are from the Malayo-Polynesian language and their histories are connected. However, there are differences in vocabulary, orthography, grammar and pronunciation between them.)

(2) **User:** Bagaimana *Jepun*$_M$ menjajah Malaysia? (How did Japan colonize Malaysia?)
ChatGPT: *Jepun*$_M$ tidak pernah menjajah Malaysia dalam *arti*$_I$ tradisional. Namun, pada masa Perang Dunia II, *Jepun*$_M$ memasuki Malaysia dan menguasai wilayah tersebut selama *tiga setengah tahun*$_I$, dari 1941 hingga 1945. Ini adalah *periode*$_I$ yang penuh dengan penderitaan dan kesengsaraan bagi penduduk Malaysia, terutama yang hidup di bawah pemerintahan *Jepun*$_M$. ... (Japan has never colonized Malaysia in the traditional sense. However, during Second World War, Japan entered Malaysia and controlled the region for three and a half years, from 1941 to 1945. This was a period full of suffering and misery for Malaysians, especially those living under Japanese rule.)

We should note that the responses were surprisingly natural unless code-mixing, as shown in (2), occurs, despite the tiny proportions of Malay and Indonesian data in ChatGPT's training dataset (Table 1). Furthermore, code-mixing, which is thought to stem from insufficient target language knowledge, occurred only in three of our 100 responses.

Language	# of words	% of total words
Dutch	669,055,061	0.34244%
Japanese	217,047,918	0.11109%
Indonesian	116,930,321	0.05985%
Malay	13,389,340	0.00685%
Javanese	1,448	0.00000%

Table 1: ChatGPT's training dataset sizes by word count (Brown et al., 2020)[2]

4 ChatGPT's LangID Ability

Why does the language choice problem discussed above happen? We hypothesized that the answer lies in LangID. Specifically, our hypothesis is that when ChatGPT responds to Malay prompts in Indonesian, it misidentifies the language of the prompt as Indonesian and decides to continue the exchange in Indonesian. We therefore examined ChatGPT's LangID ability. We also examined those of native speakers and Google Translate for comparison.

4.1 Methodology

We asked ChatGPT to identify the languages of 600 sentences (300 sentences for each language).

Test data The sentences were taken from the test data Nomoto et al. (2018) used. Their data consists of three components: news, wiki and fiction. Each component contains 100 sentences per language. The news component consists of articles from the online version of two local newspapers,

[1]The same test was not conducted for Indonesian because Indonesian does not seem to suffer from the relevant language choice problem.

[2]https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv, accessed on 16 September 2023

namely *Sinar Harian* for Malay and *Kompas* for Indonesian. The wiki component is created from the Asian Language Treebank Parallel Corpus (Riza et al., 2016),[3] which was built by translating English *Wikinews* articles. The fiction component consists of short stories called "cerpen" collected from the online *cerpen* collection *Penulisan2u* for Malay and *Cerpenmu* for Indonesian. For each component, we took the first two sentences from the first 50 files unless the sentence was shorter than four words or an English sentence resulting from code-switching, in which case we took the next sentence. The collection of 600 sentences thus collected were arranged randomly and numbered.

Prompts We used four prompts in English, assuming that ChatGPT performs best in English. Prompts 1–3 differ in the expressions referring to the languages. They start with the instructions in (3), followed by the test sentences. Prompt 4 does not specify the language options at all, as in (4).

(3) **Prompts 1–3:** What languages are the following sentences written in, $\left\{\begin{array}{l}1.\ \text{Malay or Indonesian}\\ 2.\ \text{"id" or "ms"}\\ 3.\ \text{Malaysian or Indonesian}\end{array}\right\}$?
For each sentence, choose one answer. No explanation is necessary.

(4) **Prompt 4:** Identify the languages of the following sentences. No explanation is necessary.

Prompt 2 uses the ISO693-1 language codes because we supposed that the data on which ChatGPT was trained contain these language codes. Malay is referred to as "Malaysian" in Prompt 3 to avoid potential confusion between the macrolanguage "Malay" (msa) and "Malay" as the standard variety used in Malaysia (zsm) (see section 2 for this ambiguity).

Experiment settings The free version of ChatGPT was used. The version of ChatGPT was "May 3" when the experiments with Prompts 1–2 were conducted. The version had been upgraded to "May 12" when the experiments with Prompts 3–4 were conducted. Due to the prompt length limit, the test sentences had to be split into small chunks consisting of approximately 25 sentences.

As for the comparison experiments with native speakers, David Moeljadi and Farhan Athirah binti Abdul Razak,[4] who are native speakers of Indonesian and Malay, respectively, manually classified the test sentences as Malay or Indonesian. For the experiment with Google Translate, the `DETECTLANGUAGE` function was used.

4.2 Results and discussion

Table 2 shows the overall results. We provide more detailed results in Appendix A. We calculated the evaluation metrics with Malay as the positive class. Even though we explicitly instructed ChatGPT to choose from Malay and Indonesian, it returned other languages too, including English (Prompt 1: 7; Prompt 2: 16; Prompt 3: 4; Prompt 4: 2), Malay/English (Prompt 1: 3), Indonesian/English (Prompt 2: 1) and Indonesian/Malay (Prompt 2: 2). Google Translate identified one sentence as Hawaiian, which is why the numbers in Table 2 do not add up to 600 for ChatGPT and Google Translate.

The native speakers' performance demonstrates the high similarity between the two languages. Some sentences can pass as either language. It is impossible to classify them into one of the two languages. Identification errors for such sentences are therefore not real errors but arise due to the task design, in which the option "Malay/Indonesian" is not available, in addition to "Malay" and "Indonesian." That said, ChatGPT's performance is significantly lower than that of human speakers.

Our hypothesis that ChatGPT responds to Malay prompts in Indonesian because of LangID errors is only partially correct. The very high recall of Prompt 1 is contrary to the fact that language choice problems are quite common in Malay but absent in Indonesian. The recalls of Prompts 2–4 are still high, given the results of our test using 100 Malay prompts presented in section 3, in which ChatGPT responded to as many as 66 of 100 prompts in Indonesian.

We therefore conclude that the main problem lies not in the LangID of prompts but in the language choice in response generation. Confirming this conclusion, surprisingly, ChatGPT was able to identify the language of all 100 Malay prompts correctly in our test by using Prompt 2.

Another support for our conclusion comes from

[3] `http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html`, downloaded on 19 July 2017

[4] Unfortunately I was not able to include them as authors due to my failure to register them in the submission system.

		Malay	Indonesian	Precision	Recall	F1
ChatGPT	Prompt 1	414	176	0.67	0.93	0.78
ChatGPT	Prompt 2	285	296	0.75	0.71	0.73
ChatGPT	Prompt 3	215	381	0.89	0.64	0.74
ChatGPT	Prompt 4	332	266	0.76	0.84	0.79
Human	Moeljadi	269	331	0.96	0.86	0.91
Human	Farhan	313	287	0.90	0.94	0.92
Google Translate		290	309	0.95	0.92	0.94

Table 2: LangID ability of ChatGPT, human native speakers and Google Translate

Javanese, a language spoken in Indonesia related to Malay and Indonesian. ChatGPT often does not respond to Javanese prompts completely in Javanese, and only the first few sentences are in Javanese, with the rest in Indonesian. For example, when we asked, "How do you make a kebaya shirt?" in Javanese (*kepriye carane nggawe baju kebaya?*), ChatGPT only provided the first three sentences of the response in Javanese, but unwanted code-switching occurred after that, and the remaining nine sentences were all in Indonesian. We therefore continued by instructing ChatGPT to answer in Javanese (*jawaben nganggo basa jawa*). The response contained no Indonesian sentence, but we found Indonesian words and phrases here and there. ChatGPT seems to regard Javanese as a distinct language from Indonesian. However, it failed to respond fully in Javanese, presumably due to insufficient training data (cf. Table 1), resulting in code-switching and code-mixing.

In the case of Malay, code-switching in the middle of a response does not seem to occur. Problematic responses are either fully in Indonesian (1) or in Malay, mixed with Indonesian words and phrases (2). This difference from Javanese suggests that ChatGPT does not treat Malay as a distinct language from Indonesian in the same way as it treats Javanese. Rather, it treats Malay as if it were a non-standard dialect or a non-formal register of Indonesian. It is known that ChatGPT responds in the standard formal variety of a language, regardless of the dialect and register of the prompt. It is true that in terms of linguistic characteristics, Malay and Indonesian are two varieties of the same language (see section 2). However, it is inadequate to treat one as standard and the other as non-standard. They are both standard varieties and therefore must be clearly distinguished in practical applications, such as ChatGPT.

5 Social Issues and Possible Solutions

The language choice problem discussed above is not just technological but leads to social issues. In this section, we discuss some of them and suggest possible solutions to overcome them.

5.1 Social issues caused by ChatGPT

Linguistic inequality and inequity Malay speakers often cannot receive responses in their language whereas Indonesian speakers always can (inequality). Consequently, Malay speakers cannot receive the same amounts of benefit from ChatGPT as Indonesian speakers can (inequity). Malay speakers could avoid Indonesian responses by prompt engineering (e.g. "Answer in Malay, but not Indonesian"). The problem with this solution is that the extra effort is not necessary for Indonesian speakers. It is the service provider's social responsibility to ensure equality and equity among the speakers.

Language shift If ChatGPT keeps responding in a language different from the language used in the prompt, the speakers of the latter language will be disappointed. Because most societies in the world are bi- or multilingual, many are likely to stop using their first language (L1) in favour of their second language (L2). This will decrease the input of L1 to ChatGPT and deteriorate the performance difference between L1 and L2, which in turn could motivate some speakers to shift from L1 to L2, at least in certain domains, including IT services. In the case of Malay, the speakers will most likely shift to English. We believe that IT services are one of the most important domains that affect a language's vitality.

Linguistic power imbalance At the end of section 4, we pointed out the possibility that ChatGPT does not treat Malay and Indonesian equally. Although Malay and Indonesian are both the official

language of a country or countries, ChatGPT behaves as if the former were a non-standard variety of the latter. Thus, ChatGPT creates a power imbalance between the two languages that should not exist. It is easy to imagine why ChatGPT exhibits such a behaviour. Indonesian has far more speakers than Malay (see section 2), hence far more training data (cf. Table 1). Therefore, without deliberate human intervention, ChatGPT will continue to widen the disparity.

5.2 Possible solutions

First, introducing a language setting whereby the individual user can specify the language can prevent responses in an unwanted language. However, this is not an ideal solution if only the speakers of a "dominant" language (Indonesian in our case) can enjoy the automatic language detection function. Therefore, LangID ability must be improved at the same time. A specialized LangID module can be incorporated into ChatGPT. Existing language identifiers are better than ChatGPT and, as far as Malay and Indonesian are concerned, can achieve human-level performance (cf. Table 2) although much room for improvement remains (Caswell et al., 2020).

Second, relating to the first point, there should be a list of languages that need to be treated separately. If both Malay and Indonesian are listed there, they will be treated equally as distinct languages rather than one being standard and the other non-standard. Such lists are already available in machine translation services such as Bing Translate and Google Translate. However, it is not always transparent why some languages are included (and provided with particular additional features) but others are not, which causes various speculations regarding the service provider's attitude towards different languages and their speakers.

Lastly, governments can also take action, so their citizens can benefit from ChatGPT. The necessary actions vary from country to country. For instance, the government of Iceland partnered with OpenAI to improve GPT's ability to handle the Icelandic language.[5] Because ChatGPT is already able to handle Malay well, the Icelandic strategy is irrelevant to the Malaysian government. Instead, it can encourage its citizens and companies to use more Malay on the internet to expand the amount of web data in Malay. Currently, many corporate websites are only available in English, even though English is neither an official language nor the national language of the country. In addition, the Malaysian government, perhaps in tandem with the governments of Brunei and Singapore, can ask OpenAI and Common Crawl, the primary source of ChatGPT's training dataset, to make Malay represented equally as Indonesian and the official languages of other countries. The current situation, as Table 1 shows, is evidently Eurocentric. It is unclear based on what criteria (aside from being an European language) the data sizes of various languages are determined. Table 3 shows three socioeconomic indicators that could be relevant to various data sizes, namely population, GDP and GDP per capita. However, none of them explains the actual data size differences.

Country	Population	GDP (bil.)	GDP per capita
Netherlands	17,703,090	884	49,979
Japan	125,124,989	4,508	36,032
Indonesia	275,501,339	1,122	4,073
Malaysia	33,938,221	385	11,372

Table 3: Socioeconomic statistics of the countries speaking the languages in Table 1 in 2022. The units for GDP and GDP per capita are (constant 2015) USD. Source: The World Bank Open Data[6]

6 Conclusion

This paper reported a problem with the use of ChatGPT in Malay and Indonesian, namely that ChatGPT often responds to Malay prompts in Indonesian. The problem occurs partially due to LangID errors, but its main source is the unequal treatment of the two languages. Specifically, Malay is treated as if it were a non-standard variety of Indonesian. The problem is not only technological but also has negative social effects, which can be alleviated technologically and sociopolitically. The present study thus contributes to ongoing debates on responsible AI development. Although it is concerned with Malay and Indonesian, the issues and solutions discussed there could apply to other sets of similar languages, such as Bosnian, Croatian and Serbian as well as Brazilian and European Portuguese.

[5] https://openai.com/customer-stories/government-of-iceland, accessed on 23 May 2023

[6] https://data.worldbank.org, accessed on 16 September 2023

Limitations

We examined ChatGPT's LangID ability by asking it to identify the language of a sentence. However, this method does not target the language ChatGPT actually identifies directly but guesses it indirectly based on the assumption that it will be reflected in the response. This assumption could be wrong. Moreover, we used the free version of ChatGPT (GPT-3.5). Some of the issues discussed in this study may not be replicable in the paid version (GPT-4).

Ethics Statement

The study reported in this paper was conducted solely by the authors, and no research assistant was involved. We used Google Translate to prepare the English translations of the Malay sentences in (1)–(2). We added necessary edits to the translations Google Translate provided. Section 5 of this paper discusses social issues arising from the use of ChatGPT in similar languages. We hope that our paper will raise awareness of those issues amongst NLP researchers and practitioners as well as the policy makers of the countries to which the issues are relevant.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23H00639.

References

Asmah Haji Omar. 2001. The Malay language in Malaysia and Indonesia: From lingua franca to national language. *The Aseanists ASIA*, II. Accessed 13/05/2017.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hiroki Nomoto, Shiro Akasegawa, and Asako Shiohara. 2018. Reclassification of the Leipzig Corpora Collection for Malay and Indonesian. *NUSA*, 65:47–66.

Hiroki Nomoto and Kartini Abd. Wahab. 2011. Konstruksi *kena* dalam bahasa Indonesia: Perbandingan dengan bahasa Melayu. *Linguistik Indonesia*, 29(2):111–131.

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian Language Treebank. In *Oriental COCOSDA*.

A More Detailed Version of Table 2

Table 2 reports the overall results of the LangID experiments, in which the test data's three components are combined. Tables 4–6 report the results for the three components separately. Table 4 shows that ChatGPT's LangID ability is in fact not very bad for news, which suggests that it is trained primarily on news data. ChatGPT's performance is much lower than that of the human native speakers and Google Translate for wiki (Table 5) and fiction (Table 6). In Table 6, the scores are generally lower than those in the other tables. Therefore, it can be said that the LangID task is more difficult in the fiction genre than in the journalism genre. The fact that the same trend is also observed with human native speakers means that fewer differences exist between Malay and Indonesian in the language of fiction stories.

		Malay	Indonesian	Precision	Recall	F1
ChatGPT	Prompt 1	118	80	0.84	0.99	0.91
ChatGPT	Prompt 2	112	85	0.84	0.94	0.89
ChatGPT	Prompt 3	94	106	0.96	0.90	0.93
ChatGPT	Prompt 4	109	91	0.86	0.94	0.90
Human	Moeljadi	98	102	0.98	0.96	0.97
Human	Farhan	106	94	0.92	0.98	0.95
Google Translate		101	99	0.97	0.98	0.98

Table 4: LangID ability of ChatGPT, human native speakers and Google Translate: News

		Malay	Indonesian	Precision	Recall	F1
ChatGPT	Prompt 1	145	47	0.62	0.90	0.73
ChatGPT	Prompt 2	99	86	0.71	0.70	0.70
ChatGPT	Prompt 3	57	139	0.81	0.46	0.67
ChatGPT	Prompt 4	116	82	0.70	0.81	0.75
Human	Moeljadi	100	100	0.95	0.95	0.95
Human	Farhan	93	107	0.97	0.90	0.93
Google Translate		100	100	0.97	0.97	0.97

Table 5: LangID ability of ChatGPT, human native speakers and Google Translate: Wiki

		Malay	Indonesian	Precision	Recall	F1
ChatGPT	Prompt 1	151	49	0.66	0.90	0.72
ChatGPT	Prompt 2	74	125	0.68	0.50	0.57
ChatGPT	Prompt 3	64	136	0.86	0.55	0.67
ChatGPT	Prompt 4	107	93	0.72	0.77	0.74
Human	Moeljadi	71	129	0.96	0.68	0.80
Human	Farhan	114	86	0.82	0.93	0.87
Google Translate		89	110	0.91	0.81	0.86

Table 6: LangID ability of ChatGPT, human native speakers and Google Translate: Fiction