IJCNLP-AACL 2023, 1–4 November 2023

This work was supported by JSPS KAKENHI Grant Number JP23H00639.

Issues surrounding the use of ChatGPT in similar languages: The case of Malay and Indonesian

Hiroki Nomoto (nomoto@tufs.ac.jp)

東京外国語大学
Tokyo University of Foreign Studies

Synopsis

- **Language choice problem:** ChatGPT often responds to prompts in Malay (fewer speakers) in Indonesian (more speakers).
- Language identification (LangID) errors alone cannot explain the problem's severity.
- **Our claim:** The problem happens mainly because of ChatGPT's unequal treatment of the two languages. Malay is treated as if it were a non-standard variety of Indonesian.
- **Social issues caused by the problem:**
  1. Linguistic inequality and inequity
  2. Language shift
  3. Linguistic power imbalance
- These negative effects can be alleviated technologically and sociopolitically.

Malay and Indonesian

Malay (msa) ..... *macrolanguage*

Malay (zsm)	**Indonesian** (ind)	other Malay varieties
32 mil. speakers	270 mil. speakers	
Malaysia, Brunei Singapore	Indonesia	..... *official language varieties*

Estimated lexical difference = about 10% (Asmah 2001)

Language choice problem

- Prompts in Malay are often responded in Indonesian.
- The problem occurs predominantly with Malay. Therefore, Indonesians will not even notice the problem!

Experiment with 100 arbitrary Malay prompts

User: Apakah perbezaan bahasa Melayu dengan bahasa Indonesia?
ChatGPT: Bahasa Melayu dan bahasa Indonesia memiliki banyak persamaan karena keduanya berasal dari bahasa Melayu-Polinesia dan memiliki sejarah yang terkait. Namun, ada beberapa perbedaan dalam kosa kata, ejaan, tata bahasa, dan pengucapan antara keduanya. ...

ChatGPT's LangID ability

Initial hypothesis: Language choice problems are due to LangID errors. → Examination of ChatGPT's LangID ability

Methodology

- **Test data:** 600 sentences (arranged randomly & numbered) = 100 x 3 components (news, wiki, fiction) x 2 languages
- **Prompts:**

1–3: What languages are the following sentences written in, {1. Malay or Indonesian / 2. "id" or "ms" / 3. Malaysian or Indonesian }? For each sentence, choose one answer. No explanation is necessary.

4: Identify the languages of the following sentences. No explanation is necessary.

- **Experiment settings:**
  - Free versions of ChatGPT ("May 3", "May 12")
  - Data were split into small chunks (approx. 25 sentences)
  - Human native speakers: David Moeljadi (Indonesian), Farhan Athirah binti Abdul Razak (Malay)
  - Google Translate: `DETECTLANGUAGE` function

Results and discussions

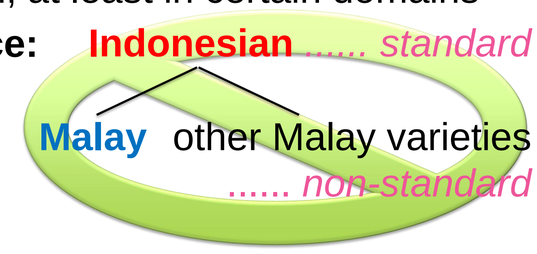
		Malay	Indonesian	Precision	Recall	F1
ChatGPT	Prompt 1	414	176	0.67	0.93	0.78
	Prompt 2	285	296	0.75	0.71	0.73
	Prompt 3	215	381	0.89	0.64	0.74
	Prompt 4	332	266	0.76	0.84	0.79
Human	Moeljadi	269	331	**0.96**	0.86	0.91
	Farhan	313	287	0.90	**0.94**	0.92
Google Translate		290	309	0.95	0.92	**0.94**

- ChatGPT's performance is lower than humans but too good.
- **Conclusion:** The main problem lies not in the LangID of prompts but in the language choice in response generation.
- ChatGPT can identify Malay prompts as Malay but chooses to respond in Indonesian, treating Malay not as a distinct language but a non-standard dialect of Indonesian.
- **Support 1:** ChatGPT was able to identify the language of all 100 Malay prompts correctly by Prompt 2.
- **Support 2:** ChatGPT cannot respond fully in Javanese due to insufficient training data, resulting in code-switching and code-mixing. This does not happen with Malay prompts.

Social issues & possible solutions

Social issues caused by ChatGPT

1. **Linguistic inequality and inequity:** Malay speakers often cannot receive responses in their language, but Indonesian speakers always can (inequality). Consequently, Malay speakers cannot receive the same amounts of benefit from ChatGPT as Indonesian speakers can (inequity).
2. **Language shift:**
   Malay speakers are disappointed & stop using Malay (L1) in favour of English (L2) → Decrease ChatGPT's Malay input → Deteriorate performance difference between L1 and L2
   ⇓
   Language shift to English, at least in certain domains
3. **Linguistic power imbalance:** ChatGPT creates a power imbalance that should not exist.



Possible solutions

- Introduce a language setting to prevent responses in an unwanted language.
- Make a list of languages that need to be treated separately.
- The Malaysian government can
  - encourage its citizens and companies to use more Malay (than English) to expand the amount of web data in Malay.
  - ask OpenAI and Common Crawl to make Malay represented equally as Indonesian.

  cf. ChatGPT's training dataset size by word count (Brown et al. 2020): Indonesian 0.05985% >> Malay 0.00685%