Proceedings of the Society for Computation in Linguistics

Volume 6 Article 44

6-1-2023

Do All Minority Languages Look the Same to GPT-3? Linguistic (Mis)information in a Large Language Model

Sydney Nguyen
Wellesley College, sn102@wellesley.edu

Carolyn Jane Anderson
Wellesley College, carolyn0jane0anderson@gmail.com

Recommended Citation

Nguyen, Sydney and Anderson, Carolyn Jane (2023) "Do All Minority Languages Look the Same to GPT-3? Linguistic (Mis)information in a Large Language Model," *Proceedings of the Society for Computation in Linguistics*: Vol. 6, Article 44.
DOI: https://doi.org/10.7275/xdf4-mh72
Available at: https://scholarworks.umass.edu/scil/vol6/iss1/44

Do All Minority Languages Look the Same to GPT-3? Linguistic (Mis)information in a Large Language Model

Sydney Nguyen and Carolyn Anderson
Wellesley College

Chatbots built atop large language models, like ChatGPT, have been proposed as a replacement for search engines. Such chatbots generate answers rather than referring users to existing resources; troublingly, these generated answers are often coherent and confident even when incorrect. The machine translation capabilities of GPT-3 (Brown et al., 2020), the base model for ChatGPT, have previously been explored (Hendy et al., 2023). In this poster, we focus specifically on how it propagates linguistic misinformation. We assess GPT-3's responses to translation queries drawn from Swadesh lists for 1,419 language varieties and document concerning patterns of behavior for low-resource languages, including mistranslating, translating into unrelated languages, and denying the existence of words.

Dataset

To test how much multilingual knowledge GPT-3 retains from training, we use a dataset of **Swadesh lists** (Swadesh, 1952), meanings proposed to be universal, from the IDS (Key and Comrie, 2021) and the Rosetta Project[1], for a total of 1,419 language varieties.[2] Though GPT-3's training data is not public, these lists were online during GPT-3's training period, and 176 are in Wikipedia, a common training data source. We use the 5 most common meanings across lists: *eye*, *fire*, *sun*, *moon*, *water*.

Evaluating linguistic (mis)information

We evaluate the correctness and consistency of GPT-3's multilingual knowledge in three tasks:

Translation We evaluate GPT-3's ability to generate and understand words in the target language in two tasks. In the *form-to-meaning* task, the model is given the language name and word form, and asked to generate its meaning. In the *meaning-to-form task*, the model's goal is to produce the correct form, given the language and target meaning. In Figure 1, the model succeeds at the form-to-meaning task and the first meaning-to-form task, but generates an incorrect form in the second (*lua* instead of *mahina*).

Backtranslation GPT-3 might translate a meaning consistently even when it does not translate correctly. We evaluate the model's consistency via a *backtranslation* task. We take each generated translation from the meaning-to-form task and ask GPT-3 to translate back into English. In Figure 1, the model fails to translate *lua* back to moon, showing that it is both incorrect and inconsistent.

Language identification We are also interested in whether GPT-3 systematically conflates languages: when it mistranslates, does it reliably swap words from particular languages? We assess this in two language identification tasks. To measure how often GPT-3 confuses forms in different languages, we ask it to identify the language for each form in a target language. We also use the model to identify the language for each generated translation from the meaning-to-form task.

Evaluation

We take the most likely completions generated by the Text-Davinci-002 GPT-3 model at temperature 0.5. We postprocess to allow for orthographic differences and to extract language names.

[1] `https://archive.org/details/rosettaproject`

[2] Some languages have multiple dialects represented; we query each dialect individually.

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 400-402.
Amherst, Massachusetts, June 15-17, 2023

	N=0	N=1	N=2	N=3	N=4	N=5
Languages with N meanings correctly translated	1272	62	24	17	15	29
Languages with N forms correctly translated	1209	89	40	32	29	20
Languages with N meanings consistently translated	811	207	116	118	98	69
Languages with N forms correctly language-IDed	1350	23	14	15	8	9

Table 1: Number of correct and consistent translations and language identifications by frequency

Meaning-to-form translation:
Translate the following words into Hawaiian:
1. sun
2. moon
1. la
2. lua

Form-to-meaning translation:
What does the word mahina mean in Hawaiian?
The word mahina in Hawaiian means moon.

Backtranslation:
What does the word lua mean in Hawaiian?
The word lua in Hawaiian means two.

Language ID:
What language is the word mahina in?
Language:
Hawaiian
What language is the word lua in?
Language:
Lua

Figure 1: Example prompts (italicized) and model responses (blue)

Translation accuracy and consistency Table 1 summarizes the model accuracy and consistency in translation and language identification across languages in our dataset. We observe that the model is more consistent than it is accurate: in many cases, it provides an incorrect form for the target language, but backtranslates it to the intended meaning. Manual inspection reveals that in many of the cases, the model provides a form for the intended meaning that is valid in a different language, as in the example in Figure 1, where the model gives the Portuguese word for moon instead of translating into Hawaiian. When we analyze by language family, we see that the model tends to mistranslate into higher-frequency members of the same family: for instance, Austronesian languages are most frequently translated into Indonesian and Niger-Congo languages into Swahili.

Language (mis)identification Language identification is challenging because forms may belong to multiple languages; it is also a task that users are likely to ask search chatbots to do. We observe harmful patterns in GPT-3's language identification. The model identifies most Pama-Nyungan language forms as Dinka. In 30 cases, the model classifies forms as *Aboriginal* or *Aboriginal Australian*. In 40 cases, the model claims that the word does not exist. The model also generates names of programming languages (Lua), language games (Pig Latin, Gibberish), and fictional languages (*Sindarin*) and places (*Wakanda*). By denying the existence or legitimacy of low-resource languages, these responses constitute representational harms to users of those languages.

Conclusion

We identify concerning patterns in how GPT-3 translates low-resource languages, including producing inconsistent translations or faulty translations, translations into unrelated languages or language games, and claims that forms are not real or belong to fictional languages. Stemming from poor online representation, these findings suggest that deploying large language models as alternatives to search engines will amplify representational harms to low-resource languages.

References

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33*, 1877–1901.

Hendy, A., M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla (2023). How good are gpt models at machine translation? a comprehensive evaluation.

Key, M. R. and B. Comrie (Eds.) (2021). *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Swadesh, M. (1952). Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society 96*, 452–463.