Toki Pali ("Word Maker"): A BERT Model for Toki Pona

Stan Lyakhov
Kendrea Beers
Gabriel Kulp

Project Goals and Methods

Implement a BERT model to do next-token prediction for the simple created language Toki Pona. Replace the tokenizer with one that only contains the Toki Pona vocabulary (so ~150 tokens).

Scrape Reddit or search Hugging Face for a dataset of material written in Toki Pona.

Many options for extensions!

- Create our own embedding--feasible because the language is so small.
- Take advantage of the small grammar to create decoding rules that will always yield correct grammar.
- Do sentiment analysis on the Toki Pona subreddit or Discord or something.
- Explore translation.
- Do image captioning--interesting because the words are so conceptually loaded, so this would test of the model's depth of understanding.
- Do speech detection, using next-token prediction to bias interpretation of next word. (Hard to get a dataset though.)

References

- original BERT paper
- Toki Pona website
- Extremely Small BERT Models from Mixed-Vocabulary Training