Tagalog PIQA Dataset

This PIQA dataset is a human-made questionnaire of 100 items written in the Tagalog Language. Items were made to reflect common spoken Tagalog (Northern and Manila dialects) and Filipino dialect as accurately as possible. Items may include grammatical errors which should be understandable by native speakers. Writing style varies between normal *street-spoken* Tagalog and formal Tagalog.

Making the Dataset

The first 66 items were inspired by daily life in the agricultural town of Talavera, Nueva Ecija showcasing normal, every day situations (things to do with a phone) and interacting with the environment (how to fish or cook).

Meanwhile, items 67 to 100 were inspired by Instructables posts that don't seem too far-fetched to be seen in Talavera. Locally-known nouns are used in place of some terms, such as *mighty bond* in place of *malakas na pandikit* (super glue), or *rugby* instead of *pandikit ng mga balat* (surface adhesive for leather).

Sources for the Instructables inspiration posts are visible in the Google Sheets and CSV versions of this dataset, which can be found in the Jupyter Notebook documentation hosted on Google Colab.

Controlling for Unconscious Biases in Item-writing

While making the dataset, I made sure to always write the correct answer first before writing the wrong answer. In the Google Sheets version of this dataset, `solution0` is always the correct answer and is properly marked by the `label` column.

This serves to control for me unconsciously writing the solutions in a certain style when the correct answer is in `solution1` and vice versa.

In the Jupyter Notebook for the TSV and JSON versions, the correct answer is randomly flipped with `solution1`. A seed is placed to make this pattern repeatable, although anyone using the Notebook can change the seed as needed.

Conclusion

This document details the creation of a Tagalog PIQA-style dataset inspired by daily life in the agricultural town of Talavera, Nueva Ecija. The dataset is available in spreadsheet, CSV, TSV, and JSON formats.