A Dataset in Malayalam for Physical Commonsense Reasoning

by Dhyuthy Krishna Kumar

Dataset description

Common sense reasoning is a crucial part of natural language understanding that continues to pose a challenge to current systems of intelligence. While several benchmark datasets exist for English, one in Malayalam targeting common sense reasoning remains to be developed. A manually constructed physical common sense reasoning dataset in Malayalam is presented here with the intention of bridging this gap. This contribution represents a step towards building more equitable and multilingual AI systems.

The dataset is entirely manually constructed with topics spanning across culture, local cuisine, etiquette, superstitions, religion, life hacks, logical reasoning etc. Since the focus is physical common sense, efforts were taken to include as many objects of daily use as possible, as found around a Malayali household. To achieve this, I situated myself in different household locations and reflected on everyday interactions with objects, as well as lesser-known facts about them.

When writing prompts about objects for which English terms are more common among native speakers, their Malayalam transliterations were used. Translation in Malayalam was provided only where it did not sound contrived to a native speaker. Direct translation from English to Malayalam while writing prompts was avoided entirely, as this was thought to diminish the nuance that underpins a Malayali's experiences. For the benefit of future users unfamiliar with the language, rough English translations are included for all the prompts in the dataset.

The dataset comprises 100 manually constructed prompts, all following the PIQA format, as specified in the guidelines. Additional columns for translations are also present. A good percentage of the dataset consists of prompt-solution pairs that are longer than 25 words. Most prompts also consist of more than one sentence, in order to provide adequate context. Solutions given in the dataset differ by no more than three words, chosen to highlight a

physical property or action surrounding an object or concept. Some prompts contain blanks, followed by a suggestion to fill them using the solutions.

Several prompts also illustrate linguistic features unique to Malayalam. For instance, the distinction between inclusive and exclusive plural first-person pronouns in Malayalam was highlighted through the solutions for prompt number 16. The key differing word in the solution was sometimes given a near-minimal pair counterpart (see prompt number 25). Popular tourist spots in Kerala make an appearance in some prompts (prompt number 27). For many prompts, both solutions represent technically achievable results. However, accurate distinguishing of the right solution in such cases requires a basic understanding of physical properties of the object and what the common practices are surrounding it.

Sufficient context has been provided within prompts, wherever necessary, so that the question does not seem irrelevant or displaced from the accompanying solutions. During training, this ensures that the model becomes familiar with how descriptions are provided in Malayalam, and the different modifiers typically used with different objects.

Quality control for the dataset was ensured through repeated proofreading by the author and two additional native speakers of Malayalam. All feedback regarding improving readability and accuracy was incorporated. Any remaining errors are solely the responsibility of the author.

[End of document]