Physical Commonsense Reasoning Dataset for Hausa Language.

Adamu Labaran Mohammed
adlabaran@gmail.com

Abstract

We introduce Physical Commonsense Reasoning Dataset for the Hausa Language. The format of the dataset is similar to PIQA, a physical commonsense reasoning benchmark where each example consists of a prompt , two solutions(solution0 and solution1) and a label which tells the correct solution.The Datasets were authored in Hausa using a theme-driven generation process.The dataset is small and intended as a starting point; we encourage further annotation and scale up before broad benchmarking.

1 Introduction

Many languages lack culturally-specific evaluation datasets that are created by language community members themselves. The Hausa Language is widely spoken in West Africa and there's currently a scarce benchmark for it. This limits robust model evaluation and model development for the Hausa Language. Existing benchmarks are predominantly English. Additionally, the benchmark datasets typically omit low resource languages especially the ones spoken in the African continent. In this paper, we provide a compact, curated Hausa dataset in the PIQA format.

2 Dataset Construction

2.1 Datasets

We introduce a PIQA-style dataset for Hausa Language that will be used for benchmarking progress in physical commonsense understanding. The underlying dataset is a multiple question and answering dataset. Every question has two possible answers. Given a prompt q, and two possible solutions, s0 and s1. The task is to select the most suitable solution for the question.Questions were created using theme-by-theme approach. For each theme we wrote short situational prompts using **how-to , what you should do, best option, where to and what is the important of** as a scaffold to generate the prompt.

2.2 Theme

We used theme as a category to constrain and guide prompt generation so questions are focused and culturally relevant. We included Themes such as **Traveling, Eating, School, Exams, Driving and Health.** These Themes speed the prompt generation process, improve coverage of everyday physical situations and make it easier to produce coherent, comparable items across the dataset.

2.3 Prompt

Our main goal is to construct a prompt that requires concrete physical reasoning. To do so, we provide a theme and a preamble of question **(how-to , what you should do, best option, where to and what is the important of)** . We used the theme with the template question to write short Hausa situational question. This process speeds prompt generation, narrow the prompt space and reduces ambiguity.

Examples from file:

Prompt1: "zaaka fita kaje gari amma ana ruwa. Me za ka yi ?"
Solution0: "zaaka fita da umbrella"
Solution1: "zaaka fita bada umbrella ba"

Prompt3: "idan kana tafiya sai kudin ka ya fadi. Me za ka yi"
Solution0: "ka saya ka dauki kudin ka"
Solution1: "ka tafi ka bar kudin ka"

3 Conclusion

We provide a compact, well-structured Hausa physical-commonsense dataset in PIQA-like format. The dataset is intended as a seed resource for Hausa commonsense reasoning.