

Dataset Description: Marathi Prompt–Solution Dataset

Anonymous ACL submission

001
002
003
004
005
006
007
008
009
010
011

012

013
014
015
016
017
018
019
020

021
022

023

024

025

026

027
028
029
030
031
032
033

Abstract

We present a manually constructed dataset of prompt–solution pairs in Marathi, with labels indicating the correct solution. The dataset is designed for tasks such as natural language understanding, commonsense reasoning, and question answering. It is formatted in TSV with four columns (prompt, solution0, solution1, label), and includes 150 entries authored in Standard Marathi with Puneri dialect influence.

1 Introduction

We present a manually constructed dataset of prompt–solution pairs in Marathi, with labels indicating the correct solution. The dataset is designed for tasks such as natural language understanding, commonsense reasoning, and question answering in Marathi.

The dataset is formatted in TSV with four columns:

- **prompt** – the question or instruction (in Marathi)
- **solution0** – candidate answer 1
- **solution1** – candidate answer 2
- **label** – the correct answer (0 or 1)

2 Language and Dialect

The dataset is written in **Standard Marathi**, specifically spoken in **Pune City, Maharashtra, India**. Sentences use commonly understood vocabulary and grammar from modern written Marathi. The **Puneri dialect** has been used to create these prompts, which has broader usability across Marathi speakers.

3 Dataset Construction

All entries were manually authored by fluent Marathi speakers. The prompts cover a wide variety of everyday domains, including:

- Education & Exams (study habits, science concepts)
- Cooking & Household Activities (recipes, safe practices)
- Sports & Games (cricket, football, volleyball, tennis, etc.)
- Shopping & Technology (gadgets, electronics, gold buying, vehicles)
- General Knowledge & Common Sense (weather, safety, language learning)

For each prompt, two possible solutions were written:

- One correct/appropriate answer (solution0 or solution1)
- One incorrect/inappropriate answer (contrastive choice)

The label column marks the correct choice with 0 or 1.

4 Data Statistics

- Format: TSV
- Columns: 4 (prompt, solution0, solution1, label)
- Language: Standard Marathi (Puneri dialect influence)
- Entries: 150

034
035
036
037

038
039

040
041

042
043

044
045

046
047

048
049

050
051

052
053

054
055

056

057

058
059

060
061

062

prompt	solution0	solution1	label
क्रिकेट खेळताना छक्का मारण्यासाठी चेंडूला योग्य दिशेने मारावे लागते कारण गुरुत्वाकर्षण शक्तीमुळे चेंडू परत खाली येतो	चेंडूला खालून वरच्या दिशेने मारावे जेणेकरून तो हवेत जास्त वेळ राहील.	चेंडूला वरून खालच्या दिशेने दाबून मारावे.	0

Table 1: Example from the dataset.

5 Example Entry

6 Conclusion

This Marathi Prompt–Solution dataset is a contribution to multilingual physical commonsense reasoning resources. It follows the rubrics provided for dataset creation and can serve as a foundation for advancing Marathi NLP research.