VPCS & MCS Annotation Guidelines

Hello, dear annotator, and thank you for your time!
Here is a brief background for the task and the specific annotation guidelines. This task is a part of the DharmaBench paper and evaluation benchmark for Sanskrit and Tibetan.

Introduction

Task: Verse vs. Prose Classification for Sanksrit (VPCS)

Goal: To distinguish verse from prose.

Background: It is a classification task aiming at assessing the model's ability to give a classification between prose and verse in any given Sanskrit text (either written exclusively in prose or verse composed intentionally following a certain metrical pattern). A verse is usually evenly divided into four parts (pādas), and each part, in most cases, follows the same metrical structure, and in some cases, follows different metrical structures, based on the number of syllables and the weight of each syllable, being guru (heavy/long) or laghu (light/short).

Real-world Relevancy: This can potentially serve as a way to pick up the root texts that are embedded in its commentary. Or further, in the case of multi-commentary texts, the model can collect and even collate the various commentaries on the same verse of the root-text. This task can also be linked to other related tasks, such as identifying quotations. Especially in the case that when it is NOT clearly indicated as a quote, by identifying that a certain portion of the text is a verse or even part of a verse, it leads to the awareness that this portion might be a quote.

Example

a. The following text from the Vigrahavyāvartanī is given as an example with the label "prose".

"na pratyayeṣu samagreṣu, na hetupratyayeṣu samagreṣu na hetupratyayavinirmuktaḥ pṛthag eva vā | yasmād atra sarvatra svabhāvo nāsti, tasmān niḥsvabhāvo 'ṅkuraḥ | yasmān niḥsvabhāvaḥ tasmāt śūnyaḥ | yathā cāyam aṅkuro niḥsvabhāvo niḥsvabhāvāc ca śūnyaḥ tathā sarvabhāvā niḥsvabhāvatvāc chūnyā iti |"

b. The following text from the Abhidharmakośabhāṣya is given as an example with the label "verse".

yaḥ sarvathā sarvahatāndhakāraḥ
saṃsārapaṃkāj jagad ujjahāra
tasmai namaskṛtya yathārthaśāstre
śāstraṃ pravakṣyāmy abhidharmakośam.

Task: Metre Classification Sanskrit (MCS)

Goal: To identify the metre of a certain given verse among ten different metres.

Background: There are more than 30 metres in Sanskrit commonly used in almost any type of Sanskrit literature. Metrical pattern is an important linguistic aspect of the Sanskrit language.

Real-world Relevancy: For philologists who edit a text, if the LLM can correctly identify a metrical pattern in a given text, it would be very helpful, because sometimes even rather common metrical patterns can be overlooked by even good human scholars. In the future, the LLM might be able to improve the quality of e-texts, when a metrical defect is detected by the machine.
The metrical pattern of a certain author can be regarded as a personal linguistic feature. Therefore, knowing a particular author's metrical habit can assist scholars in determining authorship.

Example

The following is given as an example with the label "sragdharā"
pādanyāsaiḥ pṛthivyāṃ vihitavighaṭanaṃ bhūbhṛtām aṭṭahasair
dṛktejaḥketughaṇṭādhvanibhir api nayannāśasṛṣṭīr jaganti |
bibhrāṇasyāvaliptapraśamanavidhaye bhīṣaṇān abhyupāyan
pāyād vo jainaguhyatrayahṛdayahṛdas pāṇḍavaṃ herukasya ||

Data Collection Guidelines

The two tasks share a dataset, consisting of 200 prose passages and 200 verses in the ten most commonly used metres, each with 20 samples, chosen from 55 works written in Classical Sanskrit composed or compiled in various time-periods before the 15th century CE, including belles-lettres, commentaries on belles-lettres, traditional Sanskrit grammar (vyākaraṇa), orthodox religious-philosophical treaties, non-orthodox religious-philosophical treaties (including Buddhist and Non-Buddhist), Buddhist scriptural literature, literary theory, and Sanskrit Epigraphy.

Annotation Guidelines

For **VPCS**, the aforementioned dataset, originally in various text formats, has been proofread and typed in TXT. format and is manually tagged by a human domain expert Shanshan Jia, with one of the two labels, either "prose" or "verse," and arranged in two folders. Harunaga Isaacson validates half of these samples.
For **MCS**, the 201 verses originally in various text formats are typed in txt. format and is manually tagged by a human domain expert Shanshan Jia

given one of the ten labels arranged in ten folders. Harunaga Isaacson validates half of these samples.

The ten tags are:

1. anuṣṭubh-pathyā
2. anuṣṭubh-vipulā
3. upajāti-family
4. śārdūlavikrīḍita
5. vasantatilakā
6. drutavilambita
7. sragdharā
8. śālinī
9. mandākrāntā
10. āryā

Apart from Vasantatilakā, which has 21 samples, twenty verses are given for each metre in other cases.

Something is off?

Reach out to us.

Thank you very much!