RCMT, RCMS, and RCDS 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: Root-Text and Commentary Matching Tibetan (RCMT) and Sanskrit (RCMS), and Root-Text and Commentary Detection Sanskrit (RCDS).

Goal: Identifying whether a given commentarial passage is commenting on (and therefore matches) a given verse or excerpt from a root-text (RCMS and RCMT). For Sanskrit, additionally detecting the precise boundaries (i.e., the span) of a commentarial passage on a given root-verse or excerpt within a potentially larger passage of text (RCDS).

Background: These tasks evaluate a language model's ability to identify commentarial relationships between root texts and their commentaries. These capabilities are particularly significant for Sanskrit and Tibetan literary traditions, where extensive commentarial literature forms the foundation of textual scholarship and interpretation.

Real-world relevancy: Success in this task demonstrates a model's potential for more sophisticated corpus-level operations, such as automatically identifying commentarial relationships and aligning related texts across large collections.

Example

(1) Root-Text and Commentary Matching Tibetan and Sanskrit (RCMT and RCMS)

Files are created with paired root-text and commentary passages:

File name: viṃśikā_1_root.txt

Content: 'di dag rnam par rig tsam nyid || yod pa ma yin don snang phyir || dper na rab rib can dag gis || skra zla la sogs med mthong bzhin ||

File name: viṃśikā_1_comm.txt

Content: theg pa chen po la khams gsum pa rnam par rig pa tsam du rnam par gzhag ste | mdo las | kye rgyal ba'i sras dag 'di lta ste | khams gsum pa 'di ni sems tsam mo zhes 'byung ba'i phyir ro || sems dang yid dang | rnam par shes pa dang | rnam par rig pa zhes bya ba ni rnam grangs su gtogs pa'o || sems de yang 'dir mtshungs par ldan pa dang bcas par dgongs pa'o || tsam zhes bya ba smos pa ni don dgag pa'i phyir ro || rnam par shes pa 'di nyid don du snang ba 'byung ste | dper na rab rib can rnams kyis skra zla la sogs pa med par mthong ba bzhin te | don gang yang med do ||

(2) Root-Text and Commentary Boundary Detection Sanskrit (RCDS)

File name: bca_5.1_root.txt

Content: śikṣāṃ rakṣitukāmena cittaṃ rakṣyaṃ prayatnataḥ | na śikṣā rakṣituṃ śakyā calaṃ cittam arakṣatā || 5.1 ||

File name: bca_5.1_comm.txt

Content: evam ātmabhāvādīnām utsargaṃ rakṣāṃ ca pratipādya punar vistareṇa rakṣāśodhanavardhanāni pratipādayitum upakramate. utpāditabodhicittena hi bodhisattvena utsṛṣṭasyāpi cātmabhāvasya rakṣāśodhanavardhanāni kāryāṇi. yasmāt—

> paribhogāya sattvānām ātmabhāvādi dīyate |
> arakṣite kuto bhogaḥ kiṃ dattaṃ yan na bhujyate ||
>
> tasmāt sattvopabhogārtham ātmabhāvādi pālayet |
> kalyāṇamitrānutsargāt sūtrāṇāṃ ca sadekṣaṇāt || (*Śikṣāsamuccaya* 5–6)

tac ca ātmabhāvādiparipālanādi śikṣārakṣaṇād eva syāt. anyathā narakādivinipātagamanāt tan na syāt. <commentary>ata idam abhidhīyate.

śikṣyate upādīyate gṛhītasaṃvaraṇeneti vihiteṣu karaṇīyatā, pratiṣiddheṣv akaraṇaṃ śikṣā, tāṃ rakṣituṃ paripālayituṃ kāmena icchatā bodhisattvena ātmacittaṃ rakṣitavyaṃ prayatnata iti kathayiṣyamāṇāt. atha śikṣārakṣaṇādhikāre kim iti cittaṃ rakṣyata ity āha—na śikṣeti. anyathā śikṣaiva rakṣitum aśakyā calam anāyattaṃ cittam arakṣatā, cittasya calatāyāṃ śikṣāyāḥ sthairyāyogāt.</commentary>

ito 'pi cittam eva rakṣaṇīyam ity āha. aparikarmitā mattavaravāraṇā na janayanti tāṃ pīḍām iha loke.

The relevant commentarial passage is marked with <commentary>...</commentary> tags.

Data collection guidelines

(1) For Matching Tasks (Tibetan and Sanskrit):

1. Select root-texts and one or more commentaries thereon from the Sanskrit or Tibetan corpora.

2. Prepare plain text files of the texts based on reliable editions or e-texts (such as from the ACIP repository).

(2) For Boundary Detection Task (Sanskrit):

1. Select a root text and a corresponding commentary.

2. Create one file with a verse or excerpt from the root-text.

3. Create a separate file with the corresponding commentary, which may begin a random number of sentences before and end a random number of sentences after the relevant commentarial passage.

4. The size of material before and after the start of the relevant commentarial passage should be randomly varied from 0 to 10 sentences or thereabouts.

Annotation guidelines

Annotation platform

Annotation for the Matching Tasks is achieved by properly segmenting texts and saving them in appropriately named files. Annotation for the Boundary Detection Task additionally involves the insertion of XML-style tags marking the beginning and end of the relevant commentarial passage.

(1) Matching Task (Tibetan and Sanskrit)

1. Read a unit (typically a single verse) from a root-text and read its corresponding commentary thoroughly.

2. Identify the beginning of the commentarial passage (including any introductory remarks) and its conclusion for the given root-text verse or excerpt.

3. Create a file (e.g., .txt) with the verse or excerpt from the root-text and another file with the corresponding commentarial passage.

4. Use clear file naming conventions (e.g., "TextName_VerseNumber_Root.txt" and "TextName_VerseNumber_Comm.txt").

5. Ensure proper file naming and pairing of root-text and commentary files.

6. Re-read and re-evaluate the pairing for accuracy.

(2) Boundary Detection Task (Sanskrit)

1. Read a unit from a root-text and read its corresponding commentary.

2. Identify the beginning of the commentarial passage (including any introductory remarks) and its conclusion for the given root-text verse or excerpt.

3. Create one file with the verse or excerpt from the root text.

4. Create a separate file with the corresponding commentary, which may begin a random number of sentences before and end a random number of sentences after the relevant commentarial passage.

5. The size of material before and after the start of the relevant commentarial passage should be randomly varied from 0 to 10 sentences or thereabouts.

6. Mark the beginning of the relevant commentarial passage with <commentary>.

7. Re-read and re-evaluate your annotations, focusing on boundaries.

Edge cases

1. Make sure the commentarial passage has enough information to be reasonably identifiable as explaining the given root-text passage.

2. Avoid commentarial passages with extremely long digressions.

3. In the case of ambiguity, please consult the team and the supervisors.

Something is off?

Reach out to us.

Thank you very much!