QUDS 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: Quotation Detection Sanskrit (QUDS)

Goal: Identifying explicit citations within texts and extracting any mentioned authors or titles.

Background: This is a detection task that aims to identify explicit citations in texts and extract the authors and titles mentioned. Citations are defined here as direct speech attributed to another text, excluding silent borrowings, stock phrases or maxims, reported speech, dialogue, and root texts embedded within commentaries. The task evaluates a model's ability to recognize citation discourse structure and to associate bibliographic metadata whenever available.

Real-world relevancy: Citations play an essential role in Sanskrit and Tibetan textual studies, revealing an author's influences, points of reference, and often temporal and geographic localisation.

Example

The citations, authors, and titles are marked with XML-style tags. This is an example of a complex citation structure, where multiple bibliographic elements appear together:

<author id="a1">āryeṇāpy</author> uktaṃ <title id="t1">madhyamakaśāstre</title>—

> <quote id="q1" authorid="a1" titleid="t1">yaḥ pratītyasamutpādaḥ śūnyatāṃ tāṃ pracakṣmahe |
> sā prajñaptir upādāya pratipat saiva madhyamā ||</quote>

iti. <author id="a2">ācāryeṇāpy</author> uktam—

> <quote id="q2" authorid="a2">bhāvā yena nirūpyante tadrūpaṃ nāsti tattvataḥ |
> yasmād ekam anekaṃ ca rūpaṃ teṣāṃ na vidyate ||</quote>

iti.

Data collection guidelines

1. Collect texts or excerpts from a text rich in citations. These should be *śāstric* (technical) works, based on reliable editions, and encoded in plain text using IAST conventions.

2. Within a plain text file, divide the text into samples/prompts of no more than approximately. 2000 characters.
3. For negative examples, cut samples of texts into segments that do not contain citations.
4. For positive samples, identify passages that contain citations.

Annotation guidelines

Annotation platform

Texts are marked in a text editor (e.g., Neovim, VSCode) manually using XML-style tags. The validity of the tags is to be verified using a custom script.

Annotation guidelines

1. Mark citations (as defined above), authors, and titles using the following tags:
    - <quote>...</quote> identifies direct citation
    - <author>...</author> indicates author's name or epithet.
    - <title>...</title> identifies the title of a text.

2. Each tag is given a unique ID attribute: e.g, <quote id="q1">...</quote>. A citation from later or earlier within the root text of the commentary gets the id ROOT.

3. Authors and titles are connected to citations by their unique ID and the attribute "authorid" or "titleid" attached to the "quote" tag: e.g., <quote id="q1" authorid="a1" titleid="t1">...</quote>.

Edge cases

1. The attribute "authorid2" can be used if a single quotation is given with multiple authors: e.g., when an author gives the Bhagavān and Nāgārjuna as sources.
2. When vocalic *sandhi* obscures the boundary of a tagged item, the vowel in question should be included inside the tag.
3. If an author's name and title occurs in a single compound with vocalic sandhi in between, tag as <author id="a1">lūyīpādā</author><title id="t3" type="generic">bhisamaye</title>
4. Some verses in prose may be authorial: if there is a reasonable chance that a verse or passage is a quote, tag.

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

Something is off?

Reach out to us.

Thank you very much!