SDT 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: Similes Detection Tibetan (SDT)

Goal: Identifying similes in text.

Background: Tibetan literature extensively employs similes to illustrate a wide range of philosophical themes. This task aims to evaluate models' ability to identify explicit similes in Tibetan literature, both allochthonous and autochthonous. These similes are typically marked by indicators such as "is like," "resembles," or "is similar to" ('dra ba, dang 'dra, bzhin, ji ltar, de ltar, lta bu).

Real-world relevancy: Identifying similes across a large corpus of texts represents the foundational step toward understanding how similes develop and function in Tibetan literature. This analysis could facilitate future research endeavors, such as tracing the evolution of specific ideas expressed through various literary devices, including similes. When combined with other analytical and computational capabilities, this approach can also help trace the historical development of particular concepts, including the range of similes used to express them and their transformation over time. Since this task is applied to both Sanskrit and Tibetan texts, the ultimate objective is to identify similes both monolingually within each corpus and cross-lingually across the two corpora.

Example

The similes are highlighted. This is an example of more complex simile, where a single object "*sems*" i.e., "the mind" is compared to four other objects: "*glog,*" "*sprin,*" "*rlung,*" and "*rgya mtsho chen po yi ni rlabs*" i.e., "lightning," "a cloud," "the wind," and "the waves on the ocean" using two different indicators, "*mtshungs*" and "*'dra*" which both can be translated as "similar to" or "like." While the example itself is not ambiguous or inherently complex to the reader, it is a sample that was challenging to annotate because of its extended simile.

1. sems ni glog dang sprin dang rlung dang mtshungs||
   rgya mtsho chen po yi ni rlabs dang 'dra||
   sgyu can 'dod dgur yul la mngon dga' ba||
   g.yo zhing 'phyan pa nges par 'dul bar bya||

Data collection guidelines

1. Go over texts from the defined sources list.
2. For negative examples, cut samples of texts into segments no longer than two verses consisting of four lines each, which do not contain a simile.
3. For positive samples, detect similes that appear in the texts.
4. Collect all in a CSV file with a column for source text name and the cut text chunk.

Annotation guidelines

Annotation platform

We will be using the Label Studio app. We will upload the collected CSV files and set up the annotation project.

Preparation

1. Log in to our Label Studio organization here.
2. Let us know, and we will approve your user and assign you to the respective project.

Annotation guidelines

1. Read the sentence/paragraph thoroughly.
2. Identify the simile(s) in the sentence.
3. Mark with “SIM” label the minimal span which contains the simile, if found.
4. Re-read and re-evaluate your annotation/s, focusing on boundaries.
5. Annotation Speed vs. Accuracy Tradeoff: Emphasize **accuracy over speed** if necessary.

Edge cases

1. In the case of ambiguity, please consult the team and the supervisors.
    a. If you can’t reach an agreement, discard the sample.

Something is off?

Reach out to us.

Thank you very much!