Antarlekhaka: A Comprehensive Tool for Multi-task Natural Language Annotation

Hrishikesh Terdalkar hrishirt@cse.iitk.ac.in
Arnab Bhattacharya arnabb@cse.iitk.ac.in
Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur, India

Aim

One-stop solution for Natural Language Annotation

Motivation

- World's ~7000 languages are **low-resource languages**
- Human annotation remains extremely relevant

Linguistic Phenomena

- Ambiguous or absent sentence boundaries
- Rearranging, splitting or merging tokens may be required
- Limited support in existing annotation tools

Punctuation and Word Order

" if no mistake you have made losing
you are a different game you should play "

" If no mistake you have made, losing you are.
A different game you should play. " – Yoda, Star Wars

" If you have made no mistake, you are losing.
You should play a different game. "

Sanskrit Example

- Majority of Sanskrit literature in poetry format
- Sentence boundary and token manipulation required

[na rocate mama-api-etad-ārye]$_1$ [yad-rāghavo vanam /
tyaktvā rājyaśriyaṃ gacchet]$_2$ [striyā vākyavaśaṃ gataḥ // 2
viparītaś ca vṛddhaś ca viṣayaiś ca pradharṣitaḥ /
nṛpaḥ kim iva na brūyāc codyamānaḥ samanmathaḥ // 3]$_3$
[...]

⇩

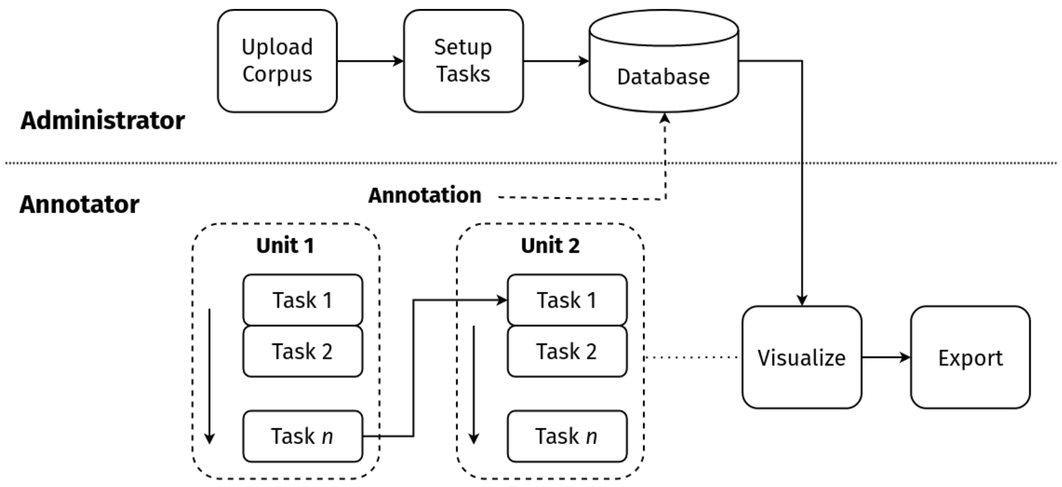
[ārye etad mama api na rocate]$_1$
[yad rāghavo rājyaśriyaṃ tyaktvā vanam gacchet]$_2$
[viparītaḥ vṛddhaḥ ca viṣayaiḥ pradharṣitaḥ ca codyamānaḥ
samanmathaḥ ca striyā vākyavaśaṃ gataḥ nṛpaḥ kim iva na brūyāt]$_3$
[...]

Example from Rāmāyaṇa

Requirements

- Intrinsic support for sentence boundaries and token order
- Comprehensive task coverage and task customization
- Multiple annotation tasks for same text

Architecture



Workflow of *Antarlekhaka*

Features

- Sequential annotation towards multiple NLP tasks
- **Eight categories** of tasks ⟹ Task-specific annotation interfaces
- Pluggable heuristics to aid annotators
- Task Management, Ontology Management, Progress Report, Clone Annotations
- Export in Human-readable and Machine-readable format
- Language agnostic, Unicode support

Task Categories

Sentence Boundary Detection
- Languages such as Sanskrit
- Corpora in poetry format

Token Manipulation: Addition, Exclusion, Merging, Splitting, Ordering
- Canonical Token Order
- Word Segmentation
- Word Grouping

Token Annotation
- Lemmatization
- Morphological Analysis
- Spelling Correction
- Phonetic Transcription

Token Classification
- Named Entity Recognition
- Part-of-speech Tagging
- Compound Classification

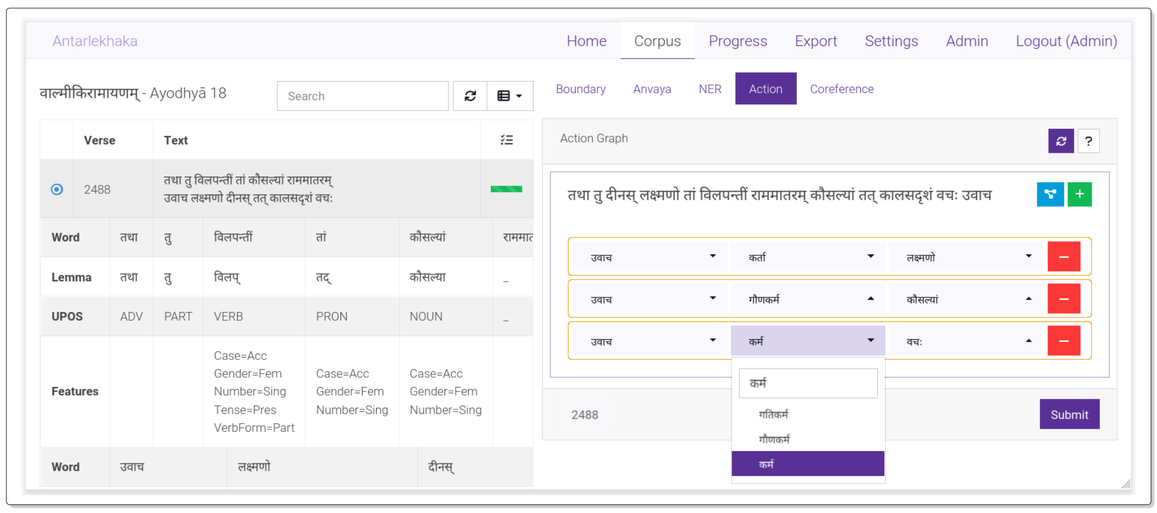
Token Graph
- Dependency Parsing
- Constituency Parsing
- Semantic Graph
- Action Graph

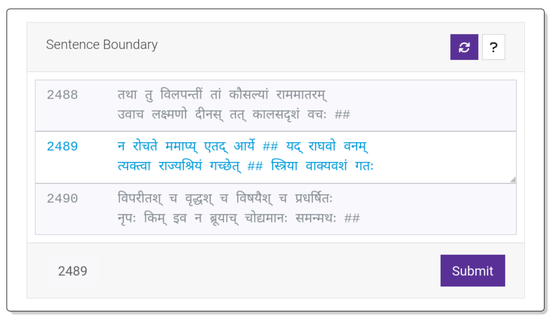
Token Connection
- Co-reference Resolution
- Interaction Networks
- Text Clustering

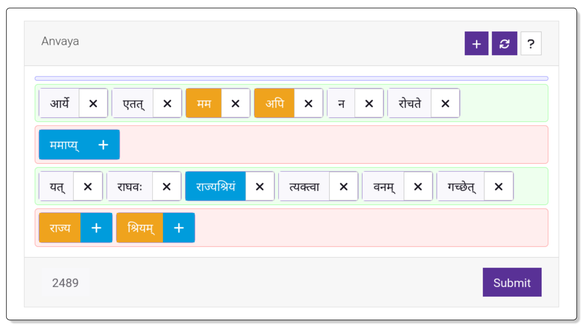
Sentence Classification
- Sentiment Detection
- Sarcasm Detection
- Spam Detection

Sentence Graph
- Discourse Graph
- Timeline Annotation

Annotation Interface

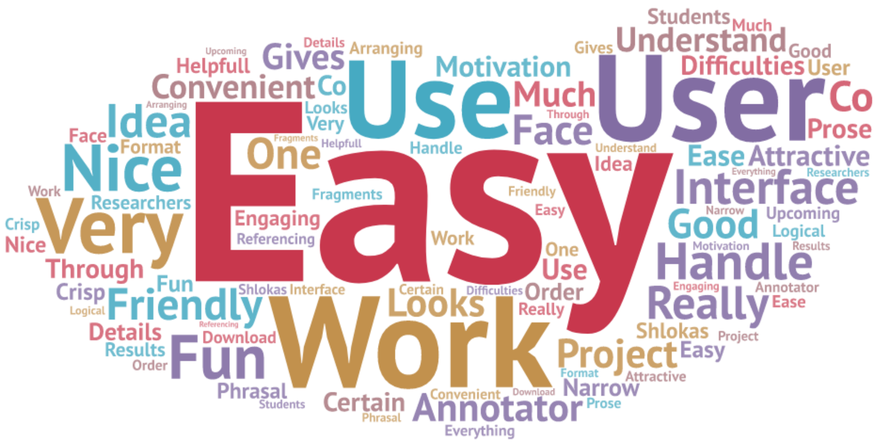






Evaluation

- **Total 29 Criteria**: Technical, Functional, Data Related, Task Related
- **Scores**: Antarlekhaka (**0.79**), INCEpTION (0.74), Sangrahaka (0.74), FLAT (0.71)
  - Only Antarlekhaka supports token ordering



Wordcloud of Survey Responses

Open Source Software

Antarlekhaka/code

Acknowledgements

We thank Chaitali Dangarikar, Shubhangi Agarwal, V S D S Mahesh Akavarapu, and Pralay Manna for their valuable feedback.

powered by Flask SQLite Jinja HTML B JS