Semantic Annotation and Querying Framework

based on Semi-structured Ayurvedic Text

Hrishikesh Terdalkar, Arnab Bhattacharya
Madhulika Dubey, Ramamurthy S and Bhavna Naneria Singh

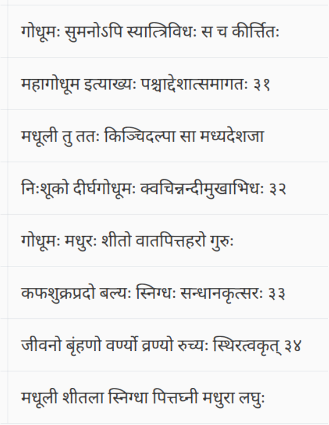
18th World Sanskrit Conference, 2022

Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur

Introduction

Knowledge Graph

Which representation is easier?



गोधूमः सुमनोऽपि स्यात्त्रिविधः स च कीर्त्तितः

महागोधूम इत्याख्यः पश्चाद्देशात्समागतः ३१

मधूली तु ततः किञ्चिदल्पा सा मध्यदेशजा

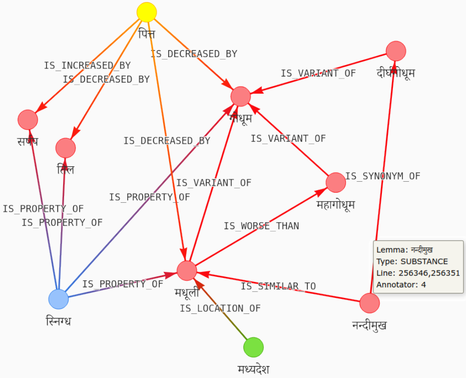
निःशूको दीर्घगोधूमः क्वचिन्नन्दीमुखाभिधः ३२

गोधूमः मधुरः शीतो वातपित्तहरो गुरुः

कफशुक्रप्रदो बल्यः स्निग्धः सन्धानकृत्सरः ३३

जीवनो बृंहणो वर्ण्यो व्रण्यो रुच्यः स्थिरत्वकृत् ३४

मधूली शीतला स्निग्धा पित्तघ्नी मधुरा लघुः



Lemma: नन्दीमुख
Type: SUBSTANCE
Line: 256346,256351
Annotator: 4

Knowledge Graph

- Real-world knowledge in structured format
- Graph data structure
- Nodes represent real-world entities
- Edges represent relationships between these entities
- Typically stored in (subject, predicate, object) triplet format, e.g., (Pāṇini, is-author-of, Aṣṭādhyāyī)

Knowledge Graph from Sanskrit

- Rich and varied literature
- Pragmatic choice – Āyurveda
- The Bṛhat-Trayī
    - Carakasaṃhitā, Suśrutasaṃhitā and Aṣṭāṅgahṛdayasaṃhitā
    - Voluminous and complex
- The Laghu-Trayī
    - Mādhavanidāna, Śārṅgadharasaṃhitā and Bhāvaprakāśa
- Bhāvaprakāśa
    - composed by Ācārya Bhāvamiśra (16th Century CE)
    - most recent of the classical treatises of Āyurveda
    - consists of 7 Bhāgas arranged in 3 Khaṇḍas
    - contains knowledge from almost all branches of Āyurveda
    - the main focus – medicine

Bhāvaprakāśanighaṇṭu

- **Bhāvaprakāśanighaṇṭu**– the glossary portion of **Bhāvaprakāśa**
  - included in the first **Bhāga** of **Pūrvakhaṇḍa**
  - consists of 2087 **ślokas** divided into 23 **varga**s
  - **varga** – a classification of substances with medicinal properties, as per their type, origin and medicinal activity.
- Contents of **Bhāvaprakāśanighaṇṭu**
  - various medicinal substances, both natural and prepared
  - synonyms, variants
  - identifying properties such as smell, color, texture, etc.
  - inherent properties such as effects on human body and effectiveness against specific symptoms or diseases
- Handy reference to practitioners and researchers of **Āyurveda**
- Ideal for construction of Knowledge Graph

Annotation

Definition

Annotation of a corpus is the process of highlighting and/or extracting objective information from it.

- Annotation in the context of KG construction
- KG may use additional real-world information
- Domain knowledge plays a role, e.g.,
  - **vāta** has a general meaning as 'wind'
  - In Ayurvedic context, refers to the **tridoṣa** – **vāta**
  - Not directly mentioned in every Ayurvedic text
  - However, any domain expert is aware of this fact

Question

Why do we need manual annotation?

Introduction

Motivation

Relevant Sanskrit NLP Tasks

- Word Segmentation
- Morphological Parsing
- Dependency Parsing
- Poetry-to-prose Linearization
- Sentence Boundary Detection
- Named Entity Recognition
- Semantic Information Extraction

Semantic Information

Concept	Words or Phrases
increases **bala**	balya, balada, balāvaha, balaprada, balakara, balakṛt
increase **vāta**	vātala, vātakṛt, vātakara, vātajanaka, vātajananī, vātātikopana, vātaprakopaṇa, vātakopana, . . .
decreases **pitta**	pittaghna, pittapraṇāśana, pittapraśamana, pittahara, pittaghnī, pittāpaha, pittajit, pittahṛt, pittavināśinī, . . .
decreases **vāta** and **pitta**	vātapittaghna, pittavātaghna, pittavātavibandhakṛt, vātapittahara, vātapittahṛt

Table 1: Semantic variations in Sanskrit through examples from Dhānyavarga.

- Multiple ways of representing a single concept
- Samāsa for multiple increment or decrements at the same time
- Semantics based on context (e.g. -ghna)

Introduction

Overview

Contribution

- Process of KG construction through manual annotation
  - Capture semantic information that is otherwise hard to capture
  - Method for capturing unnamed entities
  - Curation process
  - Optimization for querying efficiency
- Ontology for **Bhāvaprakāśanighaṇṭu**
  - 25 entity types and 29 relationship types
  - Good starting point for other Ayurvedic texts
- Data and Framework
  - Manual annotation of **Dhānyavarga** from **Bhāvaprakāśanighaṇṭu**
  - Knowledge Graph consisting of 410 nodes and 764 relationships
  - Deployment of customized instance of *Sangrahaka*[1]
  - 31 query templates in Sanskrit and English

[1] `https://sanskrit.iitk.ac.in/ayurveda/`

Framework

Sangrahaka – (Terdalkar and Bhattacharya, FSE 2021)

a web-based tool for annotating entities and relationships from text corpora towards construction of a knowledge graph and subsequent querying using templatized natural language questions.

- Language and corpus agnostic tool
- Customized for our purpose
  - Enriched with output from Sanskrit specific tools
  - Auto-complete suggestions (transliteration schemes)
  - Live at `https://sanskrit.iitk.ac.in/ayurveda/`[2]
- Under active development
  - e.g., Graph Query Builder

[2] **Login**: demo, **Password**: wsc22demo

Outline

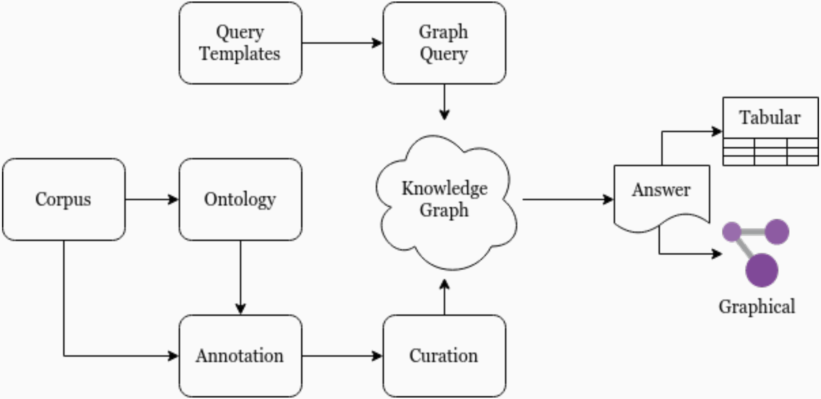


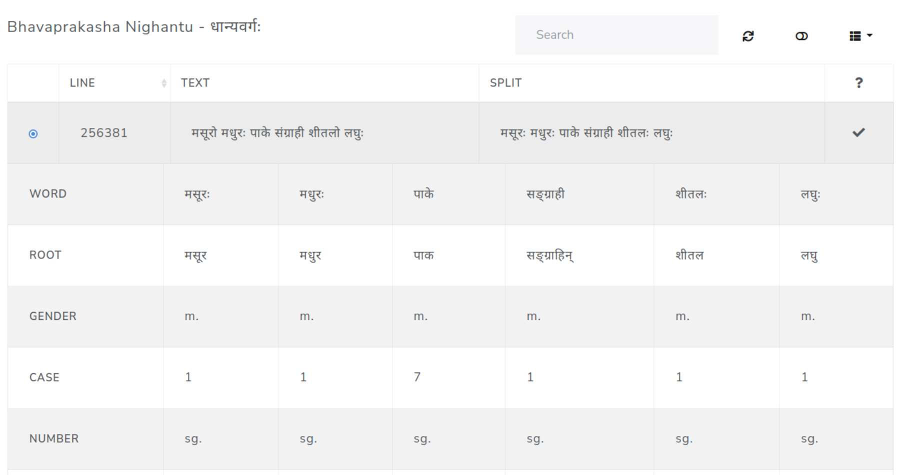
Figure 1: Workflow of semantic annotation for KG construction and querying

Annotation

Annotation Process

- Annotation for the purpose of building a KG
- Ontology-driven
- Five annotators with basic knowledge of Sanskrit and Ayurveda
- Careful reading of each line from Dhānyavarga
  - Entities – Substances, Properties, . . .
  - Relationships
- References: Translations in Hindi and English
  - Hindi version often uses the Sanskrit words as they are
  - English version has several errors
  - Both versions are consulted only as a reference

Corpus Interface



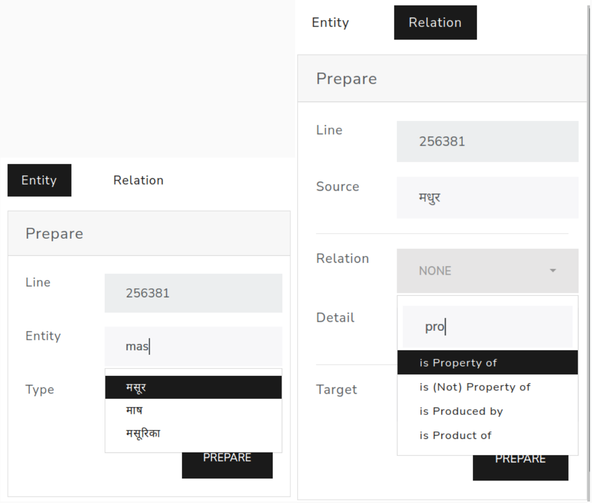
Bhavaprakasha Nighantu - धान्यवर्गः

	LINE	TEXT	SPLIT	?
◉	256381	मसूरो मधुरः पाके संग्राही शीतलो लघुः	मसूरः मधुरः पाके संग्राही शीतलः लघुः	✓

WORD	मसूरः	मधुरः	पाके	सङ्ग्राही	शीतलः	लघुः
ROOT	मसूर	मधुर	पाक	सङ्ग्राहिन्	शीतल	लघु
GENDER	m.	m.	m.	m.	m.	m.
CASE	1	1	7	1	1	1
NUMBER	sg.	sg.	sg.	sg.	sg.	sg.

Figure 2: Sample text from **Dhānyavarga** with linguistic information

Annotation Interface



Line 256381

Entity mas

Type

Line 256381

Source मधुर

Relation NONE

Detail pro

Target

Figure 3: Modified annotation interface with multi-transliteration-based suggestions

Auto-complete Suggestions

- For every *Devanagari* entity that gets annotated,
- Maintain index of transliterations to several standard schemes[3]
- e.g., Consider a word in Devanagari 'माष'
- Transliterations: '**mASa**' (HK), '**mASha**' (ITRANS), '**māṣa**' (IAST), '**maa.sa**' (Velthuis), '**mARa**' (WX) and '**mAza**' (SLP1).
- User may enter at least 3 starting characters from any of the scheme, e.g., 'mas', 'maa', 'maz', 'mar', etc. and
- Devanagari word 'माष' will be in the suggestions

[3] All available schemes in `indic-transliteration` package

Annotation

Ontology

Ontology Creation

- Careful examination of several chapters of Bhāvaprakāśanighaṇṭu
- Factors
    - Importance of the concept
    - Frequency of its occurrence
    - Relationship with other concepts
    - Nature of frequently asked questions
- 25 Entity Types
- 29 Relationship Types

Ontology

Entities (25)	Substance, Part of a Substance, Compound Substance, Prepared Substance, Collection of Substances, Tridoṣa, Property, Effect, Disease, Symptom, Product/Waste of Human Body, Part of Human Body, Person, Animal, Plant, Source, Animal Source, Plant Source, Quantity, Method or Preparation, Usage, Location, Time, Season, Others
Relationships (29)	is Synonym of, is Type of, is Variant of, is Property of, is (Not) Property of, is Similar to, is Better/Larger/Greater than, is Worse/Smaller/Lesser than, is Newer than, is Older than, is Best/Largest/Greatest among, is Medium among, is Worst/Smallest/Least among, is Ingredient of, is Part of, is (Not) Part of, is Disease of, is Caused by, is (Not) Caused by, is Benefited by, is Harmed by, is Produced by, is Removed/Cured by, is Increased by, is Decreased/Reduced by, is Preparation of, is (Absence/Lack of) Preparation of, is Location of, is Time of

Entity Annotation

Example – śloka-31

godhūmaḥ sumano'pi syāttrividhaḥ sa ca kīrttitaḥ.

Meaning: **Godhūma** (wheat) is also called **sumana**, and it is said to be of three kinds.

- Two words – **godhūmaḥ** and **sumanaḥ**
- **Prātipadika** – **godhūma** and **sumana**
- Both of type “Substance”
- Needs to be added explicitly only the first time it is encountered

Entity Annotation

Example – śloka-33 – Compound Word

godhūmaḥ madhuraḥ śīto vātapittaharo guruḥ.

Meaning: Godhūma is sweet, cold, hard to digest and removes (decreases) vāta and pitta.

- Often **samāsa** is used to indicate an effect on an entity
- Identify relevant word(s) from the segmentation
- **vātapittaharaḥ** – a single word
- indicates that **vāta** and **pitta** are reduced by **godhūma**
- **vātapittahara** not added as an entity
- entities **vāta** and **pitta** are recognized

Relation Annotation

Example

śloka-31 line-1
godhūmaḥ sumano'pi syāttrividhaḥ sa ca kīrttitaḥ.
śloka-33
godhūmaḥ madhuraḥ śīto vātapittaharo guruḥ.
kaphaśukraprado balyaḥ snigdhaḥ sandhānakṛtsaraḥ.

- Relevant relations marked for lines as and when encountered

(31.1) sumana $\vdash$ is Synonym of $\rightarrow$ godhūma

- Details can be added on relations

(33.1) madhura $\vdash$ is (rasa) Property of $\rightarrow$ godhūma

- For compound words, relations with each relevant word

(33.1) vāta $\vdash$ is Decreased by $\rightarrow$ godhūma
(33.1) pitta $\vdash$ is Decreased by $\rightarrow$ godhūma

- Subject-word or object-word might be absent from the line

(33.2) kapha $\vdash$ is Increased by $\rightarrow$ godhūma

Unnamed Entities

Example **śloka-39**

mudgo bahuvidhaḥ śyāmo haritaḥ pītakastathā.
śveto raktaśca teṣāntu pūrvaḥ pūrvo laghuḥ smṛtaḥ. ||39||

- At times, an entity may be referenced by its properties only, and not named at all in the text
- Five colored variants of **mudga**, but they are not named
- We create *unnamed entities* (denoted by `X`-prefixed nodes)
- Each given a unique identifier, e.g., `X1-256358`, `X2-256358`, . . .
- Relations to describe the properties, e.g.,
  śyāma ⊢ is (varṇa) Property of → `X1-256358`
  harita ⊢ is (varṇa) Property of → `X2-256358`
- Word **teṣām** in second line refers to the five varieties
- Relations between unnamed entities can now be captured
  `X1-256358` ⊢ is Better (in property laghu) than → `X2-256358`
- Anonymous nodes are treated like any other node

Curation

Equivalent Entities

- Adjectives in different genders – e.g., **grāhin** ↔ **grāhiṇī**
- Multiple names for same concept – e.g., **anila** ↔ **vāta**
- Add '**is Synonym of**' for these relations as well

Inconsistent Node Categories

- Differences of opinions between annotators, e.g.,
- An entity **jvara** (fever) marked both as a "Symptom" and "Disease"
- Resolved through discussion among the curators

Missing Node Categories

- Entities can be mentioned in the relationships directly
- Set of inference rules
- e.g., source of the relation '**is Property of**' should be a "Property".

Annotation

Symmetric Relationships

Symmetric Relationships – Problem

- Relation **is Synonym of** is symmetric
- A is a synonym of B ⇔ B is a synonym of A
- Several synonyms of each substance
  e.g. **rājikā** ↔ **kṣava** ↔ **kṣutābhijanaka** ↔ **kṛṣṇīkā** ↔ **kṛṣṇasarṣapa** ↔ **rājī** ↔ **kṣujjanikā** ↔ **āsurī** ↔ **tīkṣṇagandhā** ↔ **cīnāka**
- **Annotation**
  uṣṇa ⊢ **is Property of** → **rājikā**
- **Query**
  Find all properties of **cīnāka**.
- **Problem**
  - Relations might be connected to each other only in a chain.
  - Potentially 10 edge traversal required!

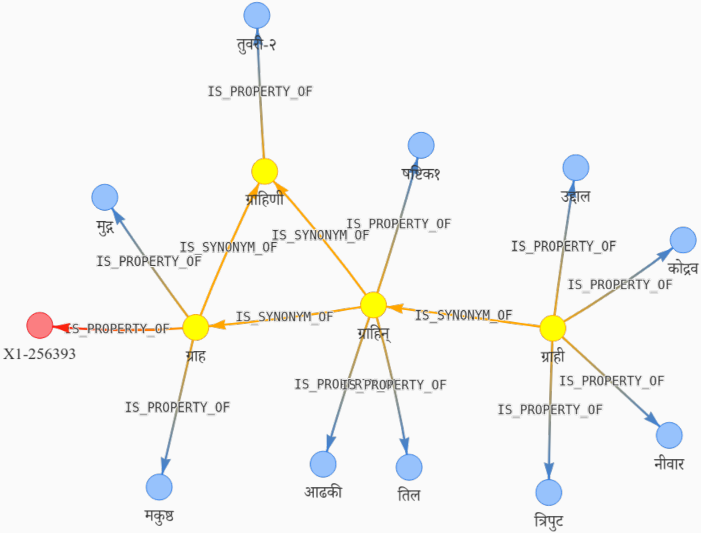
Symmetric Relationships – Solution

- For each node, identify group of nodes connected to it by paths of specific symmetric relations (e.g. is Synonym of)
- Choose a canonical node (e.g. one with the highest out-degree)
- Transfer all edges from each node in the group to the canonical node

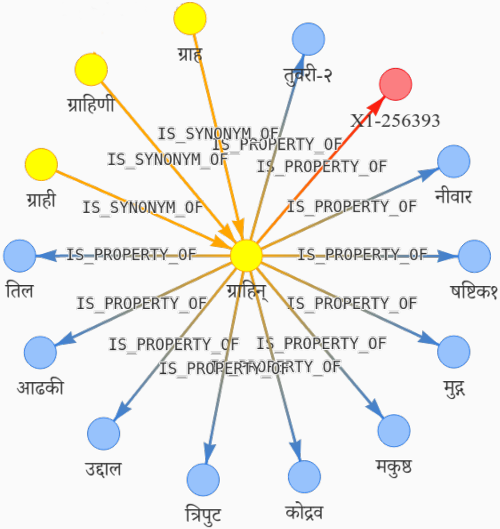
Effect

- Every node connected to canonical node.
- Thus, at most 1 extra edge traversal required.
- Initial computation cost for efficient querying.

Example – Before Optimization



Example – After Optimization



Querying

Querying System

- Neo4j Graph Database
- Total 31 query templates

Example – Query Template

Sanskrit: के पदार्थाः {0} इति दोषस्य वर्धनं कुर्वन्ति।

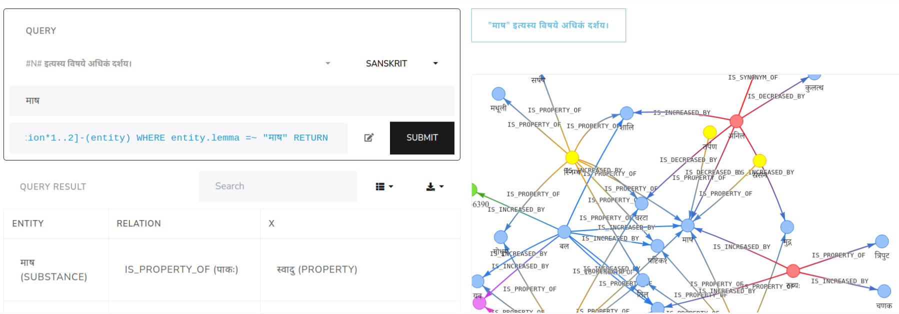
English: `Which entities increase the dosha {0}?`

Cypher:

```
MATCH (dosha:TRIDOSHA)-[r:IS_INCREASED_BY]->(entity)
WHERE dosha.lemma =  "{0}"
RETURN entity
```

- Generalized Query Templates
  - Which entity is related to entity **{0}** by relation **{1}**?
  - How is entity **{0}** related to entity **{1}**?
  - Show all matches where an entity of type **{0}** has relation **{1}** with an entity of type **{2}**.

Query Interface



QUERY

#N# इत्यस्य विषये अधिकं दर्शय।

SANSKRIT

माष

:ion*1..2]-(entity) WHERE entity.lemma =~ "माष" RETURN

SUBMIT

"माष" इत्यस्य विषये अधिकं दर्शय।

QUERY RESULT

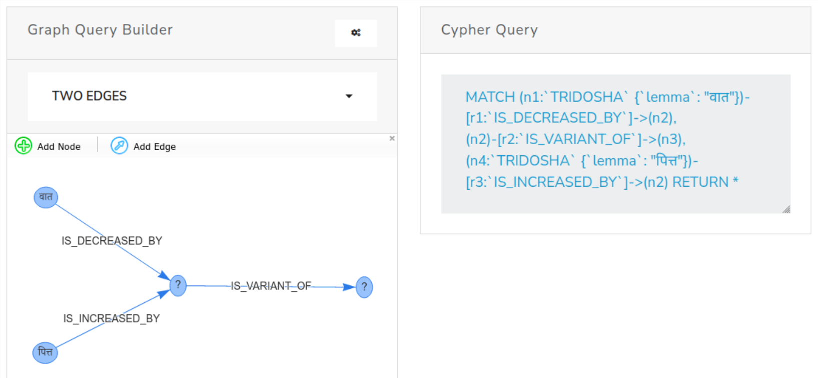
Search

ENTITY	RELATION	X
माष (SUBSTANCE)	IS_PROPERTY_OF (पाकः)	स्वादु (PROPERTY)

Figure 4: Output using query interface featuring Sanskrit query templates

- Results available in tabular and graphical format
- Tabular results can be exported as JSON, CSV, TXT, EXCEL
- Graphical results can be exported as PNG

Graph Builder



TWO EDGES

```
MATCH (n1:`TRIDOSHA` {`lemma`: "वात"})-
[r1:`IS_DECREASED_BY`]->(n2),
(n2)-[r2:`IS_VARIANT_OF`]->(n3),
(n4:`TRIDOSHA` {`lemma`: "पित्त"})-
[r3:`IS_INCREASED_BY`]->(n2) RETURN *
```

Figure 5: Graph Builder Interface

Conclusions

Conclusion and Future Work

- Construction of a knowledge graph (KG) through manual annotation process with a special focus on capturing semantic information.
- Introduce a mechanism to handle unnamed entities in a KG.
- Created an ontology for Bhāvaprakāśanighaṇṭu
- Performed semantic annotations on a chapter – Dhānyavarga
- Deployment: `https://sanskrit.iitk.ac.in/ayurveda/`

Future Work

- Complete the annotation of the rest of the Bhāvaprakāśanighaṇṭu
- Explore more classical texts such as Rāmāyaṇa and Mahābhārata
- Annotating more general kinds of relationships

References

References

- Oliver Hellwig, Sebastian Nehrdich: *Sanskrit Word Segmentation Using Character-level Recurrent and Convolutional Neural Networks*. EMNLP 2018.
- The Sanskrit Reader Companion, Heritage Platform, Gérard Huet, `https://sanskrit.inria.fr/DICO/reader.fr.html`
- Hrishikesh Terdalkar, Arnab Bhattacharya: *Framework for Question-Answering in Sanskrit through Automated Construction of Knowledge Graphs*. ISCLS 2019.
- Hrishikesh Terdalkar, Arnab Bhattacharya: *Sangrahaka: A Tool for Annotating and Querying Knowledge Graphs*. ESEC/FSE 2021.
- Amba Kulkarni, *Samsaadhanii: A Sanskrit computational toolkit*. 2016.
- Amrith Krishna, Bishal Santra, Pavankumar Satuluri, Sasi Prasanth Bandaru, Bhumi Faldu, Yajuvendra Singh, and Pawan Goyal, *Word segmentation in sanskrit using path constrained random walks*. COLING 2016.
- Amrith Krishna, Bishal Santra, Ashim Gupta, Pavankumar Satuluri, and Pawan Goyal, *A graph-based framework for structured prediction tasks in sanskrit*. Computational Linguistics, 2021.

Thank you!

Questions?

State-of-the-art

Word Segmentation (WS) and Morphological Parsing (MP)

- Highly inflectional language
- Heavy use of compound words in sandhi and samāsa
- Notable Works
  - The Sanskrit Heritage Platform (SHP) (Huet, 2009; Goyal, 2012)
  - Samsaadhanii (Kulkarni, 2016)
  - Sanskrit Sandhi and Compound Splitter (SSCS) (Hellwig and Nehrdich, 2018)
  - Word segmentation using path constrained random walks (Krishna et al., 2016)
  - Graph based framework for structured prediction (Krishna et al., 2021)

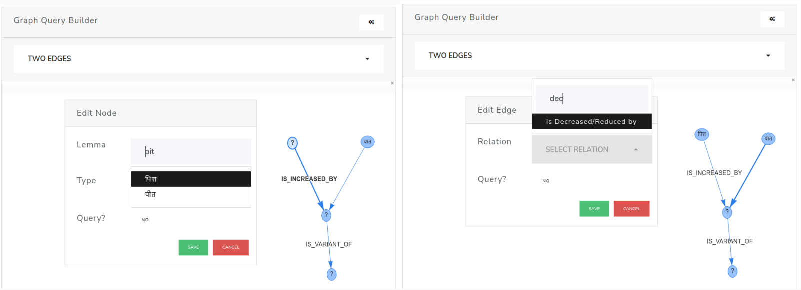
Performance and Issues

- WS task as splitting both sandhi and samāsa can be problematic
  - If passed to a morphological analyzer afterwards
- Multiple morphological analyses possible
- Sanskrit WSMP Hackathon[4]
  - T1: WS *(F1: 97.478)*
  - T2: MP (on segmented output) *(F1: 69.327)*
  - T3: Combined WS and MP *(F1: 80.018)*
- Not sufficient for downstream tasks
- Dependency parsing
  - Samsaadhanii (Kulkarni 2016) – requires prose order
- Poetry-to-prose linearization
  - (Krishna et al. 2021) – could not obtain code to evaluate
- NER, Sentence boundary detection, . . .

[4] `https://competitions.codalab.org/competitions/35744#results`

Additional Screenshots

Graph Builder



Graph Query Builder

TWO EDGES

Edit Node

Lemma: pit

पित्त
पीत

Type

Query? NO

SAVE CANCEL

Graph Query Builder

TWO EDGES

Edit Edge

dec

is Decreased/Reduced by

Relation: SELECT RELATION

Query? NO

SAVE CANCEL

Figure 6: Graph Builder Interface