Sanskrit Question-Answering Framework

Automated Construction of Knowledge Graphs

Hrishikesh Terdalkar and Arnab Bhattacharya

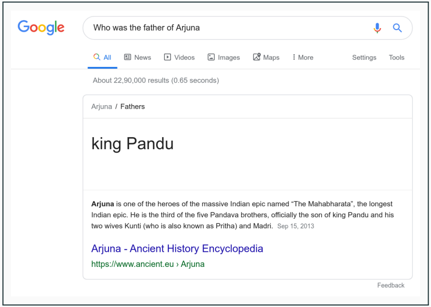
6th ISCLS, 2019

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

Introduction

Motivation

Who was the father of Arjuna



About 22,90,000 results (0.65 seconds)

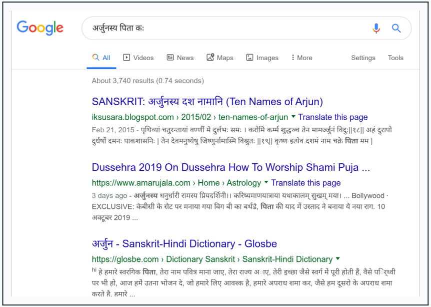
Arjuna / Fathers

king Pandu

Arjuna is one of the heroes of the massive Indian epic named “The Mahabharata”, the longest Indian epic. He is the third of the five Pandava brothers, officially the son of king Pandu and his two wives Kunti (who is also known as Pritha) and Madri. Sep 15, 2013

Arjuna - Ancient History Encyclopedia
https://www.ancient.eu › Arjuna

अर्जुनस्य पिता कः



About 3,740 results (0.74 seconds)

SANSKRIT: अर्जुनस्य दश नामानि (Ten Names of Arjun)
iksusara.blogspot.com › 2015/02 › ten-names-of-arjun ▾ Translate this page
Feb 21, 2015 - पृथिव्यां चतुरन्तायां वर्णी मे दुर्लभः समः । करोमि कर्म्म शुद्धञ्च तेन मामर्ज्जुनं विदुः||१८|| अहं दुरापो दुर्धर्षो दमनः पाकशासनिः | तेन देवमनुष्येषु जिष्णुर्नामास्मि विश्रुतः ||१९|| कृष्ण इत्येव दशमं नाम चक्रे **पिता** मम |

Dussehra 2019 On Dussehra How To Worship Shami Puja ...
https://www.amarujala.com › Home › Astrology ▾ Translate this page
3 days ago - **अर्जुनस्य** धनुर्धारी रामस्य प्रियदर्शिनी।। करिष्यमाणयात्राया यथाकालम् सुखम् मया। ... Bollywood · EXCLUSIVE: केबीसी के सेट पर मनाया गया बिग बी का बर्थडे, **पिता** की याद में उस्ताद ने बनाया ये नया राग. 10 अक्टूबर 2019 ...

अर्जुन - Sanskrit-Hindi Dictionary - Glosbe
https://glosbe.com › Dictionary Sanskrit › Sanskrit-Hindi Dictionary ▾
hi हे हमारे स्वरगिक **पिता**, तेरा नाम पवित्र माना जाए, तेरा राज्य अाए, तेरी इच्छा जैसे स्वर्ग में पूरी होती हैं, वैसे पृ्थ्वी पर भी हो, आज हमें उतना भोजन दे, जो हमारे लिए आवश्क है, हमारे अपराध शमा कर, जैसे हम दूसरो के अपराध शमा करते है. हमारे ...

Why not just use translations?

- Not always available
- Fail to convey the exact meaning

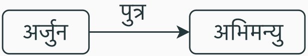
Contribution

- Automated construction of knowledge graphs
- Type of relationships
  - Human relationships from Rāmāyaṇa, Mahābhārata
  - Synonymous relationships from Bhāvaprakāś́a Nighaṇṭu
- Natural language question answering system (Sanskrit)
- Methods
  - Handcrafted rules
  - Heuristics based on linguistic information
  - Feature engineering
- 50% of the factoid questions answered
- Analysis of the shortcomings

Overview

Background

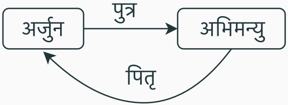
- Knowledge Graphs
  - Real-world entities as nodes
  - Relationships among the entities as directed edges
- Triplets (*subject, predicate, object*)
  - Common way of encoding the relationship information
  - Represents a directed edge
  - (Arjuna, has-son, Abhimanyu)



- Natural Language: Sanskrit (संस्कृतम्)
  - Morphologically rich
  - Abundance of compound words
  - Free word order, Strict grammatical rules

Human Relationships

- Relationship words corpus independent
  - **पितृ** (pitṛ, father), **मातृ** (mātṛ, mother), **पुत्र** (putra, son), etc.
- Synonyms to the relationship words
  - **दुहितृ, तनया, आत्मजा** are synonymous to **पुत्री**
- Inverse Relations
  - (Arjuna, has-son, Abhimanyu)



- Composite Relations
  - (Nakula, has-mother, Mādrī), (Mādrī, has-brother, Śalya)
  - **नकुलस्य मातुलः कः** (Who is the maternal uncle of Nakula?)
- Recursive Relations
  - *has-ancestor*, *has-descendant*

Question-Answering Framework

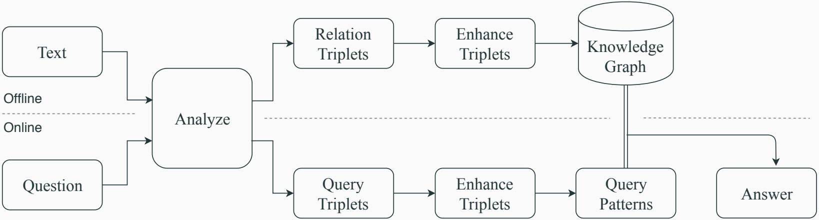


Figure 1: Overall framework of the QA system

Processing Sanskrit Text

- Sentence: **कर्णार्जुनयोः कः श्रेष्ठः**
- Splitting of samāsa and sandhi
  - *Sanskrit Sandhi and Compound Splitter* [1]
  - Output:
    - **कर्ण-अर्जुनयोः कः श्रेष्ठः**
- Semantic analysis of the word
  - *The Sanksrit Heritage Platform* [2]
    - case (vibhakti, **विभक्ति**)
    - number (vacana, **वचन**)
    - gender (liṅga, **लिङ्ग**)
  - Output:
    - **कर्ण** [‘voc.’, ‘sg.’, ‘m.’]
    - **अर्जुन** [‘loc.’, ‘du.’, ‘m.’]
    - **किम्** [‘nom.’, ‘sg.’, ‘m.’]
    - **श्रेष्ठ** [‘nom.’, ‘sg.’, ‘m.’]

[1] Oliver Hellwig, Sebastian Nehrdich: *Sanskrit Word Segmentation Using Character-level RNNs and CNNs.* EMNLP 2018.

[2] The Sanskrit Reader Companion, Heritage Platform, Gérard Huet, `https://sanskrit.inria.fr/DICO/reader.fr.html`

Knowledge Graph Construction

Building Knowledge Graph

- List of human relationship words and their synonyms (key-value)
- Map of Inferred Relations
  - Relation to Inverse Relation
  - Composite Relation to Constituent Relations

Finding Triplets

- Search for relationship words
- Proximity of subject and object (assumption)
  - Context window of 3 śloka
- Case based rules
  - *subject*: genitive case (ṣaṣṭhī vibhakti)
  - *predicate*: relationship word (various cases)
  - *object*: same case as the `predicate`

Example - Building Knowledge Graph

- Line from śloka
  विराटस्य दुहितरमुत्तरां नामाभिमन्युरुपेयेमे
- After sandhi-samāsa splitting
  विराटस्य दुहितरम् उत्तराम् नाम अभिमन्युः उपेय इमे
- Semantic Analysis
  विराट {g. sg. m.}, **दुहितृ** {acc. sg. f.}, **उत्तरा** {acc. sg. f.}
- Relationship Triplet
  ('विराट', 'पुत्री', 'उत्तरा')
- Inverse Relationship Map:
  'पुत्री' → ['मातृ', 'पितृ']
- Enhanced Triplet:
  ('उत्तरा', 'पितृ', 'विराट')

Knowledge Graph Details

		Rāmāyaṇa	Mahābhārata
Time taken	Preprocessing	$\sim$ 3.5 days	$\sim$ 13 days
	Triplet Extraction	14.18 sec	57.19 sec
	Triplet Enhancement	0.40 sec	2.05 sec
Before enhancement	Entities (Nodes)	1,711	3,552
	Triplets (Edges)	6,155	18,936
	Type of Relations	24	25
After enhancement	Entities (Nodes)	1,711	3,552
	Triplets (Edges)	11,367	32,395
	Type of Relations	27	27

Table 1: Statistics of the knowledge graphs for the human relationships.

Question-Answering

Type of Questions

- Natural language questions (Saṃskṛta)
- Factoid questions
- Human relationships (Mahābhārata and Rāmāyaṇa)
- Query in `object`:
  अर्जुनस्य पिता कः? (Who was the father of Arjuna?)
- Query in `subject`:
  पुरुः कस्य भ्राता? (Whose brother was Puru?)
- Query in `predicate`:
  द्रौपदी अर्जुनस्य का (Who was Draupadī of Arjuna?)
- Complex Query
  कस्य पुत्रस्य विवाहः द्रौपद्या सह अभवत्? (Whose son married Draupadī?)

Identifying Query Triplets

- Pre-processed in the similar manner
  पुरोः भ्राता कः →
  पुरु ['g.', 'sg.', 'm.'], भ्रातृ ['nom.', 'sg.', 'm.'], किम् ['nom.', 'sg.', 'm.']
- Parsing the words and sequential processing to form triplets
  - Initialize blank triplet (_, _, _)
  - For each word, decide if subject, predicate or object
  - Decision based on case and linguistic rules
    - (पुरु, _, _)
    - (पुरु, भ्रातृ, _)
    - (पुरु, भ्रातृ, किम्)
  - Once a triplet is filled up, initialize a new blank one
- Collect all complete triplets

Example - Querying

- śloka from two different chapters
  पूरोर्भार्या कौसल्या बभूव तस्यामस्य जज्ञे जनमेजयः
  and
  शर्मिष्ठायाः सुतो द्रुह्युस्ततोऽनुः पूरुरेव च कथं ज्येष्ठानतिक्रम्य कनीयान्राज्यमर्हति

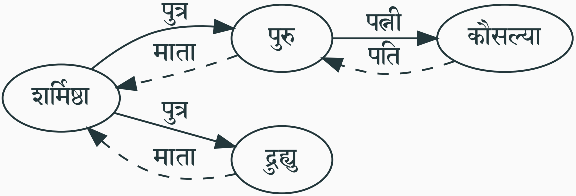
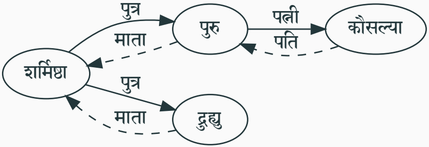


Figure 2: Knowledge Graph enchanced with Inverse Relations

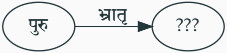
Example - Querying



Q1: पुरोः भ्राता कः

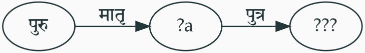
(Who was the brother of Puru?)

Triplet: [('पुरु', 'भ्रातृ', 'किम्')]



Composite Map:

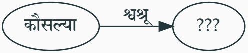
'भ्रातृ' → [('मातृ', 'पुत्र'), ('पितृ', 'पुत्र'), . . .]



Q2: कौसल्यायाः श्वश्रूः का

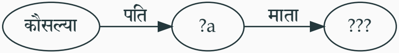
(Who was the mother-in-law of Kausalyā?)

Triplet: [('कौसल्या', 'श्वश्रू', 'किम्')]



Composite Map:

'श्वश्रू' → [('पति', 'मातृ'), ('पत्नी', 'मातृ')]



Question-Answering Tasks for Human Relationships

Questions

- Collected from 12 different users (5-10 per user)
- 35 questions from Rāmāyaṇa
- 45 questions from Mahābhārata

Tasks

- QParse: (query parsing task)
  success, if the query pattern is correctly formed from the natural language question;
  failure, otherwise.
- QCond: (conditional question answering task)
  success only if the QParse is successful and answer is found;
- QAll: (overall question answering task)

Performance on Human Relationships

Text	Task	Total	Found	Correct	Precision	Recall	F1
Rāmāyaṇa	QParse	35	33	27	0.82	0.77	0.79
	QCond	27	19	09	0.47	0.33	0.39
	QAll	35	20	10	0.50	0.29	0.37
Mahābhārata	QParse	45	45	41	0.91	0.91	0.91
	QCond	41	36	22	0.61	0.54	0.57
	QAll	45	40	23	0.58	0.51	0.54
Combined	QParse	80	78	68	0.87	0.85	**0.86**
	QCond	60	55	31	0.56	0.46	**0.50**
	QAll	80	60	33	0.55	0.41	**0.47**

Table 2: Performance of the question-answering tasks.

Errors in Knowledge Graph and Question-Answering

- Errors in parsing the question
  - **कर्णार्जुनयोः कः सम्बन्धः** → **[किम्, किम्, सम्बन्ध]**
  - Due to unhandled pattern
  - Easy to resolve, if found
- Errors in answering
  - **हनुमतः पिता कः** → **[हनुमत्, पितृ, किम्]**
  - Answer triplet **[मारुति, पितृ, पवन]** exists
  - **मारुति** is another name of **हनुमत्**
  - Use of dictionaries, thesauri ‘might’ help
  - Corpus-dependent

Other Errors

- Errors in text
    - [चन्द्रि का] चर्महन्त्री च पशुमेहनकारिका
    - चन्द्रि का → चन्द्रिका
- Errors in semantic analysis
    - नन्दिनी → नन्दिन् ['acc.', 'du.', 'n.']
    - Correct: नन्दिनी ['nom.', 'sg.', 'f.']
- Oversplitting sandhi and samāsa
    - कारवी → का रवी
- Errors in analysis of split samāsa
    - कारवी → का रवी → किम् ['nom.', 'sg.', 'f.'], रवि ['acc.', 'du.', 'm.']
    - Correct: कारवी ['nom.', 'sg.', 'f.']

Technical Texts

Technical Texts

- Corpus
    - Bhāvaprakāśa Nighaṇṭu from Āyurveda
    - Glossary chapter
- Structure
    - Similar substances (dravya, द्रव्य) in one chapter
    - Various *blocks* (sets of consecutive śloka about one substance)
    - Internal components of a block
        - Synonyms of the concerned substance
        - Where that substance can be found
        - Properties of the substance. e.g., colour, smell, texture, composition and other medicinal properties
        - Differences between the different varieties of the substance
- Deviation from structure exists.

Types of Nouns

- Substances
  Names of medicinal herbs and substances, or their synonyms
- Property Words
  - Words describing names of various properties of susbstances
    e.g. colour, smell, texture, etc.
  - Values of these properties
    e.g. red, sweet, rough, etc.
- Frequency Analysis
  - $\sim 19k$ nouns ($\sim 3.5k$ unique)
    ('पित्त', 461), ('कफ', 438), ('गुरु', 254), ('उष्ण', 240), ('तिक्त', 237)
- Heuristic
  - Top-N (50) frequent nouns as *property words*

Question-Answering Task

- Implicit questions
- Relationship: `is-synonym-of`
- Triplets: (`substance-1`, `is-synonym-of`, `substance-2`)
- Finding pairs of synonyms
    - Finding śloka containing synonyms
    - Given such a śloka, finding pairs of synonyms

Synonym Identification

- Synonym śloka Identification
  - Realized as binary classification problem
  - Structural information to identify synonyms
  - Extract linguistic features
  - 42 dimensional feature vector for each śloka
    - #words, #nouns, #properties, various ratios, etc
  - Created ground truth for 2 chapters
  - Out-of-the-box classifiers
- Synonym Pair Identification
  - List of nouns $\{n_1, n_2, \ldots, n_k\}$
  - Exclude property words
  - *Synonym Pair*: (n_i, n_j) such that both n_i and n_j have
    same case (**विभक्ति**)
    same number (**वचन**)
  - Synonyms can be in different genders

Feasibility of Classifiers

- Does the structure change with chapters?
- Various training-testing set choices
- Precision: ~ 0.74, Recall: ~ 0.65, F1: ~ 0.69

Scenario	Training Set	Testing Set
S1	20% of adhyāya 1	80% of adhyāya 1
S2	20% of adhyāya 2	80% of adhyāya 2
S3	adhyāya 1	adhyāya 2
S4	adhyāya 2	adhyāya 1

Table 3: Training and testing scenarios on Bhāvaprakāśa Nighaṇṭu.

Group Coverage

- **Synonym Group**
  Set of synonyms of a particular substance
- **Coverage**
  A *synonym group* is said to be **covered** if *at least two* from the group are detected as synonyms.

	Synonym śloka	Groups present	Groups found	Group coverage
adhyāya 1	90	87	60	0.69
adhyāya 2	54	53	39	0.74

Table 4: Group coverage in synonym pair identification.

Summary

Summary

- Framework to build knowledge graph from Saṃskṛta texts
- Multiple rule-based and heuristic-based components
- A step towards building full-fledged knowledge graphs

Future Work

- Improving individual components
- Utilisation of dictionaries, thesauri
- Reachability queries to improve searching for relations
- Identifying properties of substances to complete herbal database

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`

Thank you!

Questions?

Dataset Statistics

Dataset	Rāmāyaṇa	Mahābhārata	Bhāvaprakāśa Nighaṇṭu
Type	Classical	Classical	Technical
Chapters	7 (kāṇḍa)	18 (parvan)	23 (adhyāya)
Documents	606	2,327	23
śloka	23,934	81,603	4,244
Words (total)	2,69,603	17,49,709	31,532
Words (unique)	16,083	55,366	5,976
Nouns (total)	1,52,878	6,36,781	19,689
Nouns (unique)	9,553	20,545	3,684

Table 5: Statistics of the various datasets used.

Features of śloka

Counts	Words, Nouns, Properties, Non-Properties, Special Words, Pronouns, Verbs, Case-*i* Nouns, Number-*j* Nouns
Ratio to Words	Nouns, Properties, Non-Properties, Special Words
Ratio to Nouns	Properties, Non-Properties, Special Words, Case-*i* Nouns, Number-*j* Nouns
Other Ratios	Properties to Non-Properties, Non-Properties to Properties, Special Words to Properties, Special Words to Non-Properties

Table 6: Features of a śloka.

Performance of Classifiers

Scenario	Train	Test	P	P'	TP	Accuracy	Precision	Recall	F1
S1	52	209	84	56	42	0.73	0.75	0.50	0.60
S2	26	105	44	43	31	0.76	0.72	0.71	0.71
S3	261	131	54	45	36	0.79	0.80	0.67	0.73
S4	131	261	90	99	66	0.78	0.67	0.73	0.70

Table 7: Performance of classifiers in identifying synonym śloka.

	Synonym śloka	Groups present	Groups found	Group coverage
adhyāya 1	90	87	60	0.69
adhyāya 2	54	53	39	0.74

Table 8: Group coverage in synonym pair identification.