SYNOPSIS

Name of Student: **Hrishikesh Rajesh Terdalkar** Roll No.: **14111265**

Degree for which submitted: **Doctor of Philosophy**

Department: **Computer Science and Engineering**

Thesis Title: **Sanskrit Knowledge-based Systems: Annotation and Computational Tools**

Name of Thesis Supervisor: **Prof. Arnab Bhattacharya**

Month and Year of Thesis Submission: **June, 2023**

A Knowledge Base (KB) is a representation of real-world knowledge in a particular domain in a computer system. Knowledge Graphs (KG) are KBs that use graph as the underlying data structure. Knowledge-based systems are software that utilize knowledge bases to solve problems. In the field of Natural Language Processing (NLP), Question Answering (QA) is a problem of finding answers to natural language questions posed by humans. KGs are an integral part of addressing the problem of question answering. Difficulty of constructing knowledge graphs depends greatly on the language and the resources available, such as datasets, tools and technologies.

Sanskrit is a classical language with a vast amount of written literature on a wide variety of topics. However, most of this literature is not available in a format that is readily usable by computer systems. As a result, from a computational perspective, Sanskrit is still considered a low-resource language. In this thesis, we make contributions towards the ultimate goal of question answering in Sanskrit through construction of various knowledge-based systems in Sanskrit.

1 Sanskrit Question Answering Framework

We first present a framework that attempts to answer factual questions through an automated construction of KGs. By leveraging the vast and varied literature in San-

skrit, including texts such as Mahābhārata and Rāmāyaṇa, and Bhāvaprakāśanighaṇṭu, we construct knowledge graphs specific to relationships found in these texts. Our natural language question answering system utilizes these knowledge graphs to answer factual questions, achieving a success rate of approximately 50%. We highlight the shortcomings and limitations of the state-of-the-art of Sanskrit NLP. The scarcity of appropriate datasets poses a significant challenge in the development and evaluation of automated systems for knowledge graph construction.

Human annotation plays an important role for the creation of such datasets. There is also a need for annotation tools with task-specific and intuitive interfaces to simplify the tedious task of manual annotation.

2 *Sangrahaka*: Annotation and Querying Tool for Knowledge Graphs

We present *Sangrahaka*, an annotator-friendly, web-based tool for ontology-driven annotation of entities and relationships towards the construction of knowledge graphs. It also supports querying. The tool is language and corpus-agnostic but customizable for specific needs. The web-based user interface has several components that are accessible to users based on their roles. Some of these components are shown in Figure 1.

We demonstrate the usefulness of the tool through a real-world annotation task on Bhāvaprakāśanighaṇṭu, an Āyurveda text. Through collaboration with Āyurveda experts, we have created a rich and extensive ontology suitable for the annotation of Bhāvaprakāśanighaṇṭu, an Āyurveda text detailing medicinal substances, their properties and medical applications. It consists of 300 node labels and 320 relationship labels. We use this ontology and a custom deployment of *Sangrahaka* to perform manual annotation on Bhāvaprakāśanighaṇṭu and subsequently construct a knowledge graph (KG), specifically focusing on the three chapters from Bhāvaprakāśanighaṇṭu, namely, Dhānyavarga, Śākavarga and Māṃsavarga. The constructed knowledge graph

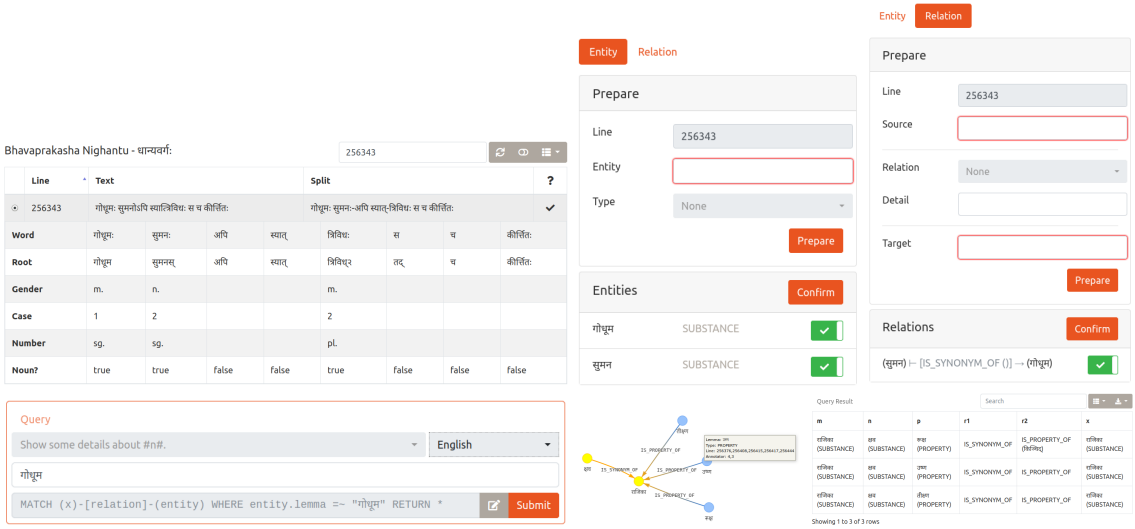

Figure 1: Corpus Viewer, Entity Annotator, Relation Annotator, Query Interface, Graphical Result Interface, Tabular Result Interface

contains 1606 entities and 1707 relationships, capturing the semantics of entity and relationship types present in the text. To facilitate querying the knowledge graph, we design 31 query templates that cover common question patterns.

3 *Antarlekhaka*: Comprehensive Natural Language Annotation Tool

Additionally, we present *Antarlekhaka*, a versatile general-purpose multi-task annotation system that supports the manual annotation of a comprehensive set of NLP tasks. This system allows users to annotate small units of text with multiple categories of NLP tasks in a sequential manner. The system not only addresses the standard NLP tasks but also provides support for specific tasks such as identifying sentence boundaries and establishing canonical word order, making it especially useful for Sanskrit and other poetic corpora. It supports a total of eight generic categories of annotation tasks: sentence boundary detection, canonical word ordering, token annotation, token classification, token graph, token connection, sentence classification and sentence graph, amounting to the support for a much larger set of NLP tasks. Each category of tasks has a unique user-friendly and intuitive interface for

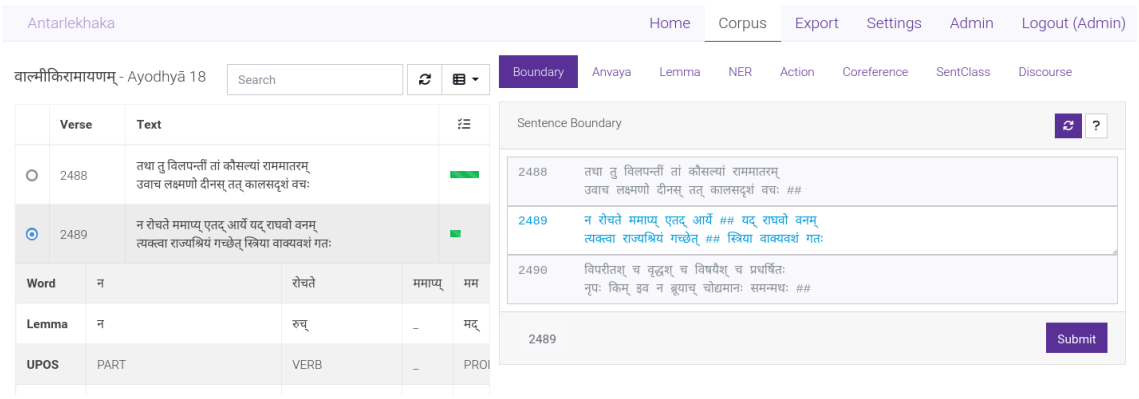

Figure 2: Annotation Interface: Corpus area shows text split into small units, and annotation area highlights various annotation task tabs

annotation, making the tedious task of annotation much more accessible. Figure 2 displays the overall annotation interface of *Antarlekhaka*.

We highlight the utility of the tool through the application of the tool for the annotation of Vālmīki Rāmāyaṇa, resulting in datasets for NLP tasks of sentence boundary detection, canonical word ordering, named entity recognition and co-reference resolution.

Both *Sangrahaka* and *Antarlekhaka* are presented as full-stack web-based software to support distributed annotation. They are designed to be easily configurable, web-deployable, customizable and with a multi-tier permission system. They are actively being used in real-world annotation tasks. The annotator-friendly and intuitive annotation interfaces of these tools have received positive feedback from the users, and they outperform other annotation tools in objective evaluation.

4 Chandojñānam: Sanskrit Meter Identification and Utilization

Sanskrit text corpora have undergone large-scale digitization efforts using OCR technology, inadvertently leading to the introduction of various errors. In this thesis, we present Chandojñānam, a system for identifying and utilizing Sanskrit meters. Apart from its core functionality of meter identification, the system also enables finding

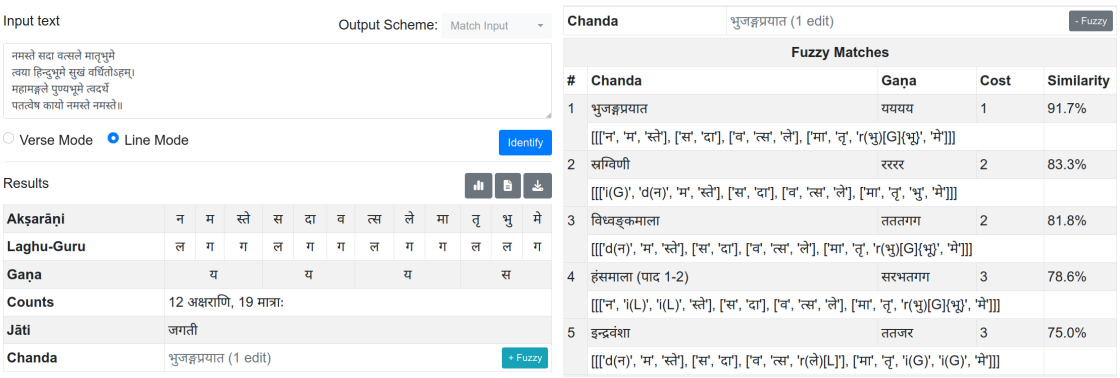

Figure 3: Meter identification with fuzzy matching and suggestions

fuzzy matches based on sequence matching, thereby facilitating the correction of inaccuracies in digital corpora. Chandojñānam displays the entire result is displayed in a neat tabular format with scansion, a graphical representation of the metrical pattern, and fuzzy matches. This can be seen in Figure 3. Additionally, the system supports meter identification from uploaded images through the utilization of optical character recognition (OCR) engines. The text can be processed in either line-by-line mode or verse-by-verse mode.

5 Miscellaneous Computational Tools for Sanskrit

As part of our research contribution, we offer an extensive range of web-interfaces, tools, and software libraries specifically designed to highlight and utilize the computational aspects of Sanskrit. This diverse compilation includes Jñānasaṅgrahaḥ, a comprehensive web-based collection of various computational applications dedicated to the Sanskrit language. The overarching aim of Jñānasaṅgrahaḥ is to present the features of the Sanskrit language in an accessible manner, even for enthusiastic users with limited Sanskrit backgrounds. Within this collection, you will find Saṅkhyāpaddhatiḥ, a web-interface that encompasses three ancient numeral systems, enabling the representation of numbers as text. Additionally, we offer Chandojñānam, a system for Sanskrit meter identification and utilization, as well as Varṇajñānam, a utility pertaining to varṇa, a phonetic unit of the Sanskrit language. Furthermore, our contributions extend to a Telegram bot designed to assist learners in com-

prehending Sanskrit grammar. Lastly, we have developed a set of Python libraries to aid programmers in working with Sanskrit corpora. These include *PyCDSL*, a Python library and a Command Line Interface (CLI) to simplify the processes of downloading, managing, and accessing Sanskrit dictionaries, *Heritage.py*, a Python interface to The Sanskrit Heritage site and *sanskrit-text*, a library for the manipulation of Sanskrit alphabet. Collectively, these resources serve as catalysts for encouraging and enabling a wider audience to delve into the richness of Sanskrit and its profound cultural heritage.

6 Conclusions

We, in this thesis, address the challenges and opportunities in the development of knowledge systems for Sanskrit, with a focus on question answering. By proposing a framework for the automated construction of knowledge graphs, introducing annotation tools for ontology-driven and general-purpose tasks, and offering a diverse collection of web-interfaces, tools, and software libraries, we have made significant contributions to the field of computational Sanskrit. These contributions not only enhance the accessibility and accuracy of Sanskrit text analysis but also pave the way for further advancements in knowledge representation and language processing. Ultimately, this research contributes to the preservation, understanding, and utilization of the rich linguistic information embodied in Sanskrit texts.

Publications

[1] **Hrishikesh Terdalkar** and Arnab Bhattacharya. Framework for question-answering in Sanskrit through automated construction of knowledge graphs. In *Proceedings of the 6th International Sanskrit Computational Linguistics Symposium*, pages 97–116, IIT Kharagpur, India, October 2019. Association for Computational Linguistics.

[2] **Hrishikesh Terdalkar** and Arnab Bhattacharya. Sangrahaka: A tool for annotating and querying knowledge graphs. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2021, page 1520–1524, New York, NY, USA, 2021. Association for Computing Machinery.

[3] **Hrishikesh Terdalkar**, Arnab Bhattacharya, Madhulika Dubey, S Ramamurthy, and Bhavna Naneria Singh. Semantic annotation and querying framework based on semi-structured Ayurvedic text. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 155–173, Canberra, Australia, January 2023. Association for Computational Linguistics.

[4] Jivnesh Sandhan, Ashish Gupta, **Hrishikesh Terdalkar**, Tushar Sandhan, Suvendu Samanta, Laxmidhar Behera, and Pawan Goyal. A novel multi-task learning approach for context-sensitive compound type identification in Sanskrit. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4071–4083, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[5] **Hrishikesh Terdalkar** and Arnab Bhattacharya. Chandojnanam: A Sanskrit meter identification and utilization system. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 113–127, Canberra, Australia, January 2023. Association for Computational Linguistics.

[6] **Hrishikesh Terdalkar** and Arnab Bhattacharya. Antarlekhaka: A comprehensive tool for multi-task natural language annotation. In *Proceedings of the 3rd Workshop on NLP Open Source Software at the 2023 Conference on Empirical Methods in Natural Language Processing*, NLP-OSS @ EMNLP, Singapore, December 2023. Association for Computational Linguistics.