ABSTRACT

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. We first present a framework that attempts to answer factual questions through an automated construction of KGs. We highlight the shortcomings and lim-

itations 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.

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. We demonstrate the usefulness of the tool through a real-world annotation task on Bhāvaprakāśanighaṇṭu, an Āyurveda text. We showcase a carefully constructed extensive ontology suitable for this task, resulting in annotations that contribute to the development of a knowledge graph and querying framework. These contributions are based on three chapters from Bhāvaprakāśanighaṇṭu.

Then, we present *Antarlekhaka*, a general purpose multi-task annotation system for manual annotation of a comprehensive set of NLP tasks. The system supports annotation towards multiple categories of NLP tasks: sentence boundary detection, canonical word ordering, token annotation, token classification, token graph, sentence classification and sentence graph. The annotation is performed in a sequential manner for small logical units of text (e.g., a verse). 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, action graph construction 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.

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 fuzzy matches based on sequence matching, thereby facilitating the correction of inaccuracies in digital corpora. The user-friendly interface of Chandojñānam displays the scansion, a graphical representation of the metrical pattern. 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.

Finally, 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 comprehending 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.

In conclusion, this thesis addresses 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.