Tools for Querying Indian Knowledge Systems

KID: 20250319 | Mr Sai Kasyap

Indian Knowledge Systems (IKS) encompass a vast body of philosophical, linguistic, and scientific traditions preserved in Sanskrit and other Indic languages. However, the complexities of Sanskrit morphology, free word order, and semantic density pose significant challenges for effective retrieval and knowledge extraction from these texts. Recent advances in Natural Language Processing (NLP) have provided a foundation for developing specialized tools to query IKS, with research spanning three domains: Named Entity Recognition (NER), semantic search engines, and speech technologies.

NER for Sanskrit

NER plays a crucial role in identifying and extracting entities such as deities, places, scholars, and canonical texts from IKS corpora. Unlike high-resource languages, Sanskrit presents unique challenges due to its sandhi formations, inflectional richness, and extensive use of compounds. Recent approaches have focused on pre-annotation and expert validation for constructing high-quality NER datasets, notably Sujoy et al. (2023), whose workflow emphasizes domain-grounded entity type design and accuracy improvement (Sujoy et al. 2023).

Comparative analysis of traditional and transformer-based NER methods on Indian epics highlights the superior adaptability of deep learning architectures for entity identification and classification tasks, especially in narrative, multilingual data (Sharma and Mohania 2022). Large-scale annotation efforts such as the "Naamapadam" dataset demonstrate the evolution of multi-language NER resources, extending entity coverage across Indic scripts and improving scalable model training for Sanskrit (Mhaske et al. 2022).

Transformer-based models, including BERT and XLM-R, have shown outstanding potential for entity recognition in morphologically rich languages by leveraging self-attention mechanisms to capture context over complex sentences (Devlin et al. 2019; Pande and Bhattacharya 2021). Fine-tuning these architectures on Sanskrit corpora aims to improve precision and recall, ensuring more reliable recognition of nuanced references embedded within verses and commentaries.

Semantic Search for IKS

Keyword matching methods are insufficient for Sanskrit literature due to lexical variation and context-specific meaning (Choudhury 2010). To overcome these challenges, specialized semantic search engines for IKS texts are under development, incorporating multilingual and cross-lingual features.
The semantic engine utilizes embedding models trained on Sanskrit–English parallel data, enabling cross-lingual semantic alignment and supporting queries in both languages. Various embedding models such as XLM-Roberta, LaBSE, and Qwen embeddings have been trained on the corpus (Reimers and Gurevych 2019; Mishra 2023).

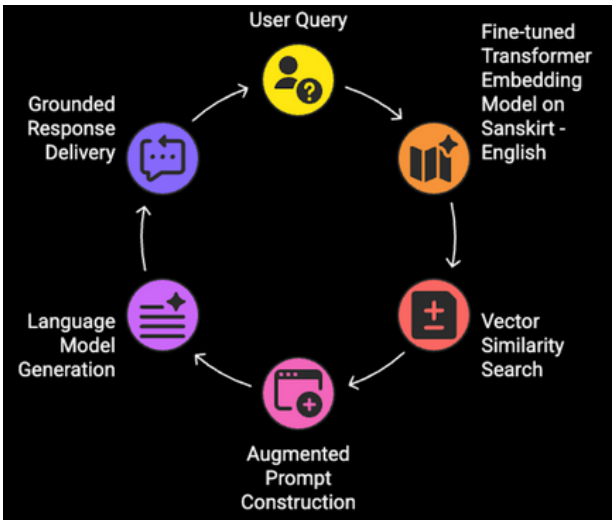

Fig 1. represents the RAG search pipeline that can be implemented after the fine tuning of the embedding model.

Speech Technologies for Sanskrit

Expanding accessibility further, speech technologies including Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) are integrated into IKS tools. ASR enables users to perform voice-based searches in Sanskrit or English, reducing the barrier posed by Indic script typing challenges (Rao and Murthy 2019). Recent advances leverage transformer-based models and optimized encoder networks for Sanskrit ASR and TTS (Sproat 2017; Joshi et al. 2021). This multimodal approach ensures interaction with IKS resources not only through text but also through spoken forms—for researchers, learners, and practitioners.

Towards an Integrated Framework

By combining NER optimization, semantic search, and speech technologies, a comprehensive framework for querying Indian Knowledge Systems can be realized. This unified approach that respects the linguistic, cultural, and oral dimensions of Sanskrit. This interdisciplinary effort enables the revitalization of IKS in the digital age, supporting both academic research and broad scholarly engagement.

Mr Sai Kasyap
Research Scholar, Dept of HST