INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

R&D SHOWCASE 2021
Technology, Social Impact

Enriching Wikipedia content in Indian languages

ABSTRACT

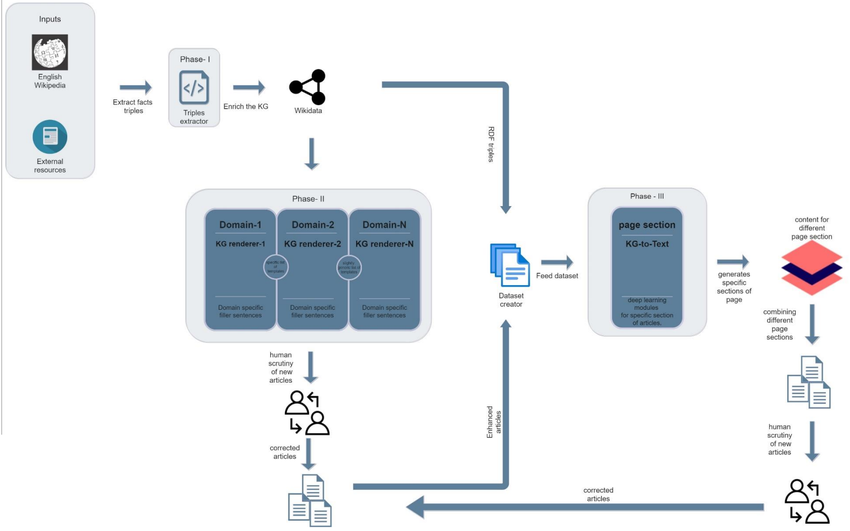
- Wikipedia is the most widely used resource of the Encyclopedic Knowledge, Education, and Literacy platform on the Web. It is a free, online encyclopedia with over six million articles just in English.
- Wikipedia content in Indian languages is minuscule compared with the number of Indian native speakers. 90% of India's population can only read in their native language couldn't benefit from this fantastic tool to access knowledge and learning.
- To facilitate article writing, we are building an automated system named **IndicWikiBot** to generate an initial version of the full Wikipedia article that humans can later edit or improve.
- In addition to this, we are also developing strategies to build a large community of human experts (**community building**) and improving present technology to enhance the productivity of the Wikipedia editors and volunteers (**content development**).

APPROACH

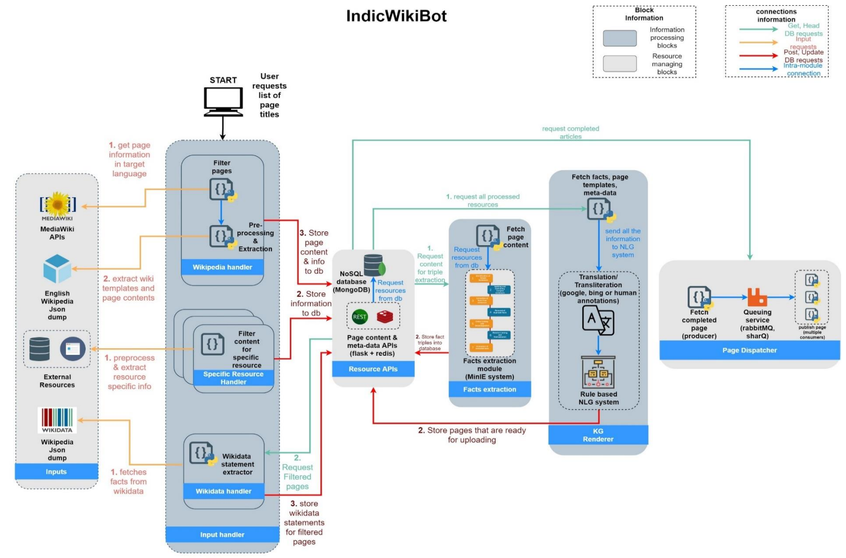
IndicWikiBot process both structured (Knowledge Bases, Databases, etc.) and unstructured information (running texts, webpages, etc.) to generate Wikipedia articles in Indian languages. It involves following steps:

1. Phase-I, Enrichment of Wikidata Knowledge Base (KB).
    1. Extraction of facts from unstructured text (most prominent source is English edition of Wikipedia).
2. Phase-II, Creation of Rule Based NLG system in Indian languages:
    1. Translation of facts stored in enriched Wikidata into native language.
    2. Creating of template (cloze style sentences) which are highly specific to one domain to create Wikipedia articles.
    3. Human evaluation and correction of the generated articles. Created articles are stored at IIIT's Wikipedia sandbox servers for further enhancement and then published to global Wikipedia.
3. Phase-III, Creation of Deep Learning based approaches
    1. Creation of sentence and fact triples pairs for specific Wikipedia page section using the enhanced articles from phase-II
    2. Designing and Training deep learning model on the above dataset to generate text in similar domain.

PROCESS FLOW



INDICWIKIBOT ARCHITECTURE



- Input specific resource handlers converts data (structured as well as unstructured data) to uniform machine-readable format that are stored to central database for further processing.
- Facts extraction module collects new facts from unstructured text (prominently from English Wikipedia). We have created a fact extraction pipeline that consists entity linking and disambiguation, relation linking and normalization, detection of complex and simple sentences, decomposition of complex sentence into simple sentences to increase the effectiveness of open information extraction system.
- Collected facts are translated to native language and rule based Natural Language Generation (NLG) system is used to create articles.

Authors: **Tushar A, Anubhav S, Bhavyajeet S, Divanshi G, Shivprasad S, Vasu S, Vasudeva V**

Research Center: Information Retrieval and Extraction Lab, LTRC, IIIT-H