Towards Enhancing Knowledge Accessibility for Low-Resource Indian Languages: A Template Based Approach

Padakanti Srijith[*]
IIIT Hyderabad
padakanti.srijith
@research.iiit.ac.in

Aravapalli Akhilesh[*]
IIIT Hyderabad
aravapalli.akhilesh
@research.iiit.ac.in

Chelpuri Abhijith[*]
IIIT Hyderabad
abhijith.chelpuri
@research.iiit.ac.in

Radhika Mamidi
IIIT Hyderabad
radhika.mamidi@iiit.ac.in

Abstract

In today's digital age, access to knowledge and information is crucial for societal growth. Although widespread resources like Wikipedia exist, there is still a linguistic barrier to breakdown for low-resource languages. In India, millions of individuals still lack access to reliable information from Wikipedia because they are only proficient in their regional language. To address this gap, our work focuses on enhancing the content and digital footprint of multiple Indian languages.

The primary objective of our work is to improve knowledge accessibility by generating a substantial volume of high-quality Wikipedia articles in Telugu, a widely spoken language in India with around 95.7 million native speakers[1]. Our work aims to create Wikipedia articles and also ensures that each article meets necessary quality standards such as a minimum word count, inclusion of images for reference, and an infobox. Our work also adheres to the five core principles of Wikipedia[2]. We streamline our article generation process, leveraging NLP techniques such as translation, transliteration, and template generation and incorporating human intervention when necessary. Our contribution is a collection of 8,929 articles in the movie domain, now ready to be published on Telugu Wikipedia[3].

1 Introduction

Wikipedia has been a source of reliable structured information for much time now. Although English Wikipedia has abundant articles(6.8 million approx.) to cover many domains of knowledge, the scarcity of articles in many Indian languages is evident. This highlights the need for a robust knowledge base accessible to native speakers. In response, our research focuses on generating Wikipedia articles with consistency and structure, addressing this gap in information accessibility.Our chosen language for this work is Telugu, belonging to the Dravidian language family[4]. It is spoken by approximately 100 million people across Telangana and Andhra Pradesh in India.

After exploring various category pages on wikipedia and identifying reliable data sources on the internet, we chose the "movies" domain. This decision was based on the popularity of movie-related pages in India and the abundance of reliable sources available for this category. One of such sources is the IMDb, which holds a vast collection of movie data across languages. Leveraging data scraping techniques, we extract the necessary information from this webpage. We build a dataset comprising 8900 movies with 52 attributes each. Using Jinja [5] template creation, we transform this dataset into an XML dump of wiki articles. Each article is manually verified before being published on the global Wikipedia platform. Our template-driven method can be re-used to generate articles in any Indian language for many simple domains with little complexity with just changing the template.

2 Related Work

Significant progress has been made in the domain of Natural Language Generation(NLG) within the NLP community, ranging from rule-based methods to statistical methods(Mahapatra et al., 2016) to neural-network based techniques(Ji et al., 2020). However, for structured data, NLG techniques such as template-based generation, rule-based approaches, and data-to-text systems are mostly used (Oluwaseyi and Potter, 2023).

(Sitompul et al., 2021) have utilized a template-based approach to generate cohesive summaries

[*]These authors contributed equally to this work.
[1]https://en.wikipedia.org/wiki/Telugu_people
[2]https://en.wikipedia.org/wiki/Wikipedia:Five_pillars
[3]https://te.wikipedia.org
[4]https://en.wikipedia.org/wiki/Dravidian_languages
[5]https://jinja.palletsprojects.com

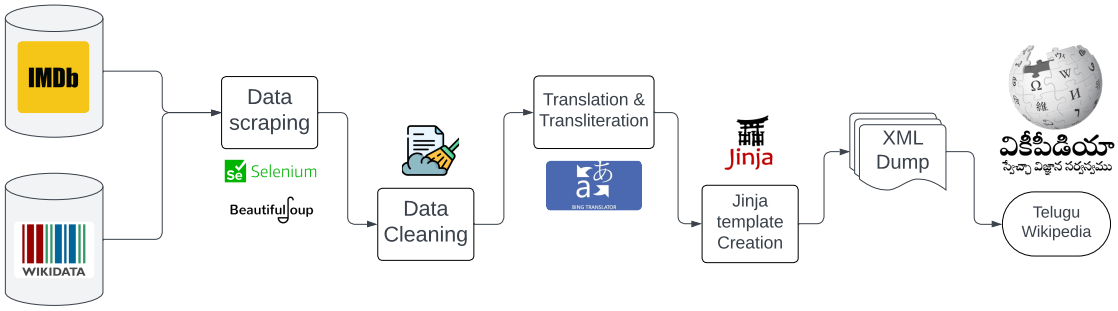

Figure 1: Process Flowchart for automatic article generation.

from laboratory findings to help young doctors.

(Pochampally et al., 2021) explores the creation of articles automatically looking at headings of various categories across Wikipedia. They propose a semi-supervised approach to generate articles automatically.

In (Agarwal and Mamidi, 2023) work, WikiData is be used as a reliable source of information for generating Wikipedia articles in Hindi.

In recent history, Sverker Johansson, a Swedish linguist, worked on creating Lsjbot[6], a program that automatically generated Wikipedia articles about living organisms and geographical entities (such as rivers, dams, and mountains).

We take inspiration from the many works that have utilized template-based text generation approaches and we focus our efforts on generating articles in Telugu Wikipedia, we employ a similar approach that would help in maintaining the consistency and reliability of Wikipedia articles. Our streamlined process is a comprehensive, end-to-end approach for generating articles across any language and domain, as long as the template used remains simple yet consistent. This flexibility and consistency make our methodology unique and sets it apart from other approaches.

3 Methodology

Our automated article generation process operates by populating a predefined text template with various attributes. The detailed methodology is illustrated in Figure 1.

3.1 Data Scraping

To create Wikipedia articles using the template-driven method, it is necessary to specifically scrape structured data in the form of an excel sheet from any particular domain. An article must meet several important requirements and policies set forth by the Wikipedia community to be accepted. The following are the main specifications: notability, trustworthy sources, neutral point of view, no personal opinions, citing sources, no plagiarism, avoiding conflict of interest, and content policies to name a few. Throughout our entire article creation process, we adhere to each of these requirements to ensure the quality and integrity of our contributions.

IMDb[7] is a reliable but huge repository of movies across languages. However, its size poses a problem in choosing which movies are noteworthy enough to be made into articles. To address this, we chose to employ the number of ratings as a proxy for a movie's popularity. We have chosen all the movies across languages which have more than 5000 votes. Using this strategy, we have consolidated a dataset of 8929 movies. To scrape all this data, we utilize a combination of Selenium[8] and BeautifulSoup[9], adapting to the static and dynamic characteristics of numerous pages across IMDb.

To develop a list of attributes required for a movie article, we draw inspiration from existing movie pages on English Wikipedia and align them with the data accessible on IMDb. These selected attributes are then extracted individually for each movie and subsequently merged. The attributes include: *IMDbID, Title, page_heading, Release_Year, Duration, eTitle, Rating, eRated, Rated, Genre, Director, Synopsis, Votes, Gross, Songs, countries, languages, release, film_locations, budget, opening_weekend, cumulative, production_company, sound_mix, colors, writers, cast, storyline, tagline, trivia, producer, composer, cinematography, film_editor, cast-*

[6]https://en.wikipedia.org/wiki/Lsjbot

[7]https://www.imdb.com/

[8]https://pypi.org/project/selenium/

[9]https://pypi.org/project/beautifulsoup4/

ing, production_design, art_design, set_decoration, eNominee, Nominee, eWinner, Winner, narrative_location, distributed_by, distribution_format, part_of_series, based_on, stars, main_subject, wikidata_url, wikipedia_url, poster.

While IMDb is a comprehensive dataset for movies, it lacks certain attributes like main subject, narrative location, source of distribution, and format of distribution which are present in English Wikipedia articles. To obtain these missing attributes, we turn to WikiData, a repository for structured data across Wikipedia. By utilizing IMDb IDs as primary keys, we link each movie entry in IMDb to its corresponding Wikipedia page using SPARQL[10], allowing us to extract the additional attributes needed from WikiData[11]. Including a reference image as part of each article is crucial for enhancing the reader's connection. However, while IMDb provides images or posters for each movie, their copyrighted status renders them unusable for extraction. Fortunately, existing Wikipedia pages offer an alternative source for images. These pages contain images that can be freely used in Telugu Wikipedia articles without any copyright concerns, as they belong to the same ecosystem. Usually it would be an easy task to map each of the movie to its corresponding Wikipedia page and extracted relevant information. However, the naming convention is a little inconsistent for pages across Wikipedia. For instance, while the movie "Slumdog Millionaire" may have a corresponding English Wikipedia page with the same name, "Dangal" might be listed under "Dangal (Film)." To address this issue, we employ a similar strategy as before, utilizing IMDb IDs and WikiIDs to bridge data across both ecosystems. This approach enables us to retrieve the corresponding Wikipedia page for each movie, from which we extract the required images effectively.

3.2 Data Cleaning

Once we have compiled all the information extracted from various sources, our next step involves cleaning this dataset for further utilization. We employ regular expressions (Regex) to remove various instances of special characters. Additionally, we tackle noise extracted due to inconsistencies in webpages using specially designed regex functions. Null values within the dataset are effectively handled using Python and Pandas. Furthermore, the dataset is checked for irregularities and inconsistencies manually and corrected.

3.3 Translation and Transliteration

Our work primarily focuses on generating articles in Telugu Wikipedia, so this necessitates translation/transliteration of all of the attributes into Telugu. While attributes such as plot, tagline, and synopsis can be directly translated, attributes such as movie name, cast, and awards, need to be transliterated because of their linguistic nature.

For translation, both the Google Translate and Bing Translate APIs were evaluated for their accuracy in cases with large contexts. We have chosen Bing Translate as it was retaining maximum context in most cases. We utilized Bing Translate API to translate all required attributes.

For transliteration, we have employed DeepTranslit[12], a python library developed for transliteration of Indian languages. However, transliterating from English, which is alphabetic in nature and where one alphabet can represent multiple sounds, often leads to inaccuracies in Indian languages. For instance, words like "The" and "To" were incorrectly transliterated as "Tē" and "Tō", respectively, due to the differing pronunciations of the letter "t". Additionally, the model also overlooked abbreviations and treated them as regular words, such as "UK" being transliterated as "uk". To address these errors, mis-transliterated words were manually identified and corrected.

3.4 Template Generation

The most crucial aspect of this paper is the generation of text and articles for the Telugu Wikipedia, for which the Jinja2 template was used. This tool was selected to convert structured data into coherent articles, with a particular focus on the infobox, which provides key information at a glance.

The final template consists of two paragraphs for introduction/summary, an infobox, plot, crew, songs table, technical details, production box office, awards, critics' responses, other info, and related categories/references section as part of each article.

Existing infobox templates from other articles were utilized, and their attributes were systematically listed. To maintain consistency across infoboxes of all movies, all the images/posters are resized to a fixed size during rendering. The infobox section includes attributes such as movie

[10]https://query.wikidata.org/

[11]https://www.wikidata.org/wiki/Wikidata:Main_Page

[12]https://pypi.org/project/deeptranslit/

name, director, writer, producer, actors, music composer, cinematographer, film editor, year of release, runtime, budget, country, language, gross revenue, distributors, and movie poster.

To enhance the readability of the articles, elements such as awards, details, and results (winner/nominee) were represented in a table format. Scrollability is added to maintain the page size across articles ensuring the information remains accessible and well-organized as shown in Table 1.

Ultimately, a Jinja template adhering to all our requirements is developed and then generated movie articles using Python. Our template is also designed to produce articles that mimic human-like writing, ensuring diversity and dynamism to avoid detection as bot-generated content by Wikipedia.

In accordance with Wikipedia's policies, we adhere to the requirement of providing references for every assertion made in each article. We have compiled IMDb references for assertions regarding songs, title pages, introduction paragraphs, and awards. Additionally, images are sourced and referenced from Wikipedia. For properties such as production design, set decoration, art design, distribution details, and series affiliation, which are derived from WikiData using wptools, references to Wikipedia and other related categories are also provided.

We have adhered to the five core pillars of Wikipedia throughout our article generation process. To maintain neutrality, we used a neutral point of view and consistently utilized passive voice in the generated content. All data sources used for extraction are reputable and recognized as authoritative within their respective domains. Additionally, we ensured that no personal opinions were incorporated into the content, as the only text added during the process is structured through a predefined template. Furthermore, our streamlined workflow includes automated citation generation for each article, proper attribution is provided in accordance with Wikipedia's guidelines. Through these measures, we have ensured that our contributions align with Wikipedia's standards and its five foundational principles.

The sample template can be found here[13]. Utilizing the dataset that we have created and the template designed, an XML dump is generated which contains all the articles that are ready to be published to global Wikipedia.

[13]https://github.com/aforakhilesh/movies-indicWiki

పురస్కారము	క్యాటగిరి	గ్రహీత (లు)	ఫలితము
గోల్డెన్ మూన్ అవార్డ్ (Golden Moon Award)	బెస్ట్ యాక్టర్ (Best Actor)	డానిలో 'బాటా' స్టోజోకేవిక్ (Danilo 'Bata' Stojokovic)	విన్నర్ (winner)
జ్యూరీ ప్రైజ్ (Jury Prize)	బాజీడర్ 'బోటా' నికాలిక్ (Bajider 'Bota' Nikalik)	డ్యూశాన్ కొవాసెవిక్ (Dusan Kovacevic)	విన్నర్ (winner)
గోల్డెన్ అరేన (Golden Arena)	బెస్ట్ యాక్టర్	daనిలో 'బాటా' స్టోజోకేవిక్	విన్నర్ (winner)
గోల్డెన్ అరేన (Golden Arena)	బెస్ట్ ఫిల్మ్ (Best Film)	డ్యూశాన్ కొవాసెవిక్ (డైరెక్టర్) (Dusan Kovacevic(Director))	విన్నర్ (winner)
గోల్డెన్ మూన్ అవార్డ్ (Golden Moon Award)	బెస్ట్ ఫిల్మ్ (Best Film)	డ్యూశాన్ కొవాసెవిక్ (Dusan Kovacevic)	నామినేట్ (nominate)

Table 1: Table showing the awards, categories, awardees, and results.

4 Extending our work

Expanding our research scope, we have extended our efforts in generating articles across various domains of knowledge. Our contributions are listed as follows:

- In **plants** domain, we have successfully created over 2140 articles with 59 attributes each that are ready to be published in the Global Wikipedia. Sources such as USDA[14] plants database and JSTOR[15] global plants database were utilized.
- In **animals** domain, 4928 Wikipedia articles are ready to be published to Telugu Wikipedia. relevant data was extracted from reliable and trusted sources like IUCN[16].
- Similarly, in **volcanoes** domain, we have 8676 articles that are ready to be added into Telugu Wikipedia. Smithsonian[17], a well-grounded source was used to extract all the relevant information related to the domain.

5 Results & Conclusion

To assess the accuracy of the translated text, we employed BERTScore(Zhang et al., 2019), a metric that evaluates both the semantic meaning and syntactic structure of sentences. Given the significant risk of meaning loss when translating to or from Indian languages, we prioritized measuring semantic similarity to evaluate the quality of machine-translated text.We achieved an average BERTScore of 0.79 for articles generated using a reference of

[14]https://plants.usda.gov/home
[15]https://plants.jstor.org/
[16]https://iucn.org/
[17]https://volcano.si.edu/

their corresponding articles from Global Wikipedia. The quality & human-like characteristics of the text generated were evaluated by 4 annotators who rated each article within a range of 1-5. To demonstrate consistency & inter-annotator agreement, we also report Fleiss' Kappa score to be at **0.82**.

As of early 2023, Telugu Wikipedia hosts around 95,599 articles[18]. Our work has produced approximately close to 25000 articles ready for publication in Global Telugu Wikipedia across multiple domains. This makes our contribution to the Telugu community further significant. Furthermore, we believe this work serves as motivation for the research community to take initiative in developing and enhancing natural language resources for various low-resource languages. Such efforts are crucial in bridging existing gaps and making a meaningful impact on society. In conclusion, our paper has demonstrated how employing data scraping techniques coupled with template generation can produce articles in Telugu Wikipedia, thereby significantly enhancing the accessibility of knowledge for millions of native Telugu speakers. We show how NLP can be leveraged to bridge linguistic and knowledge gaps across low-resource languages, ultimately creating a more inclusive society. We remain committed to our goal of enhancing the digital footprint of many more Indian languages.

References

Aditya Agarwal and Radhika Mamidi. 2023. Automatically generating hindi wikipedia pages using wikidata as a knowledge graph: A domain-specific template sentences approach. *Proceedings of Recent Advances in Natural Language Processing*.

Yangfeng Ji, Antoine Bosselut, Thomas Wolf, and Asli Celikyilmaz. 2020. The amazing world of neural language generation. In *EMNLP (Tutorial Abstracts)*, pages 37–42. Association for Computational Linguistics.

Joy Mahapatra, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2016. Statistical natural language generation from tabular non-textual data. In *International Conference on Natural Language Generation*.

Joseph Oluwaseyi and Kaledio Potter. 2023. Exploring natural language generation (nlg) methods for generating human-like text from structured or unstructured data. *Journal of Machine to Machine Communications*.

Yashaswi Pochampally, K. Karlapalem, and Navya Yarrabelly. 2021. Semi-supervised automatic generation of wikipedia articles for named entities. *Proceedings of the International AAAI Conference on Web and Social Media*.

Opim Salim Sitompul, Erna Budhiarti Nababan, Dedy Arisandi, Indra Aulia, and Hengky Wijaya. 2021. Template-based natural language generation in interpreting laboratory blood test. *IAENG International Journal of Computer Science*, 48(1):57–65.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Acknowledgments

We would like to thank the MT-NLP Lab, LTRC, IIIT-Hyderabad for their valuable efforts in reviewing and manually correcting the articles. We also extend our gratitude to the IndicWiki team for hosting our articles on tewiki.iiit.ac.in

[18] https://meta.wikimedia.org/wiki/List_of_Wikipedias_by_language_group