Text-2-Wiki: Summarization and Template-driven Article Generation

Jayant Panwar
LTRC, International Institute of
Information Technology, Hyderabad
jayant.panwar@research.iiit.ac.in

Radhika Mamidi
LTRC, International Institute of
Information Technology, Hyderabad
radhika.mamidi@iiit.ac.in

Abstract

Users on Wikipedia collaborate in a structured and organized manner to publish and update articles on numerous topics, which makes Wikipedia a very rich source of knowledge. English Wikipedia has the most amount of information available (more than 6.7 million articles); however, there are few good informative articles on Wikipedia in Indian languages. Hindi Wikipedia has approximately only 160k articles. The same article in Hindi can be vastly different from its English version and generally contains less information. This poses a problem for native Indian language speakers who are not proficient in English. Therefore, having the same amount of information in Indian Languages will help promote knowledge among those who are not well-versed in English.

Publishing the articles manually, like the usual process in Global English Wikipedia, is a time-consuming process. To get the amount of information in native Indian languages up-to-speed with the amount of information in English, automating the whole article generation process is the best option. In this study, we present a stage-wise approach ranging from Data Collection to Summarization and Translation, and finally ending with Template Creation. This approach ensures the efficient generation of a large amount of content in Hindi Wikipedia in less time. With the help of this study, we were able to successfully generate more than a thousand articles in Hindi Wikipedia with ease.

1 Introduction

In our study, we aim to populate our native language, i.e. Hindi's Wikipedia, on a certain domain by making use of a summarization and template-driven approach. We have selected Diseases and Medical conditions as our domain as they are a very important topic for readers of any language.

The gist of the study is to scrape and collect required data from online resources and then clean it accordingly. Then, we summarize the scraped data and organize it to give it some structure. We follow it with translation and transliteration of the data to Hindi. This stage involves both machine and manual translation and transliteration. Finally, we end the study by drafting up a template in which the data can be filled as fill-in-the-blanks. This ensures the automatic generation of a large amount of articles in accordance with the Wiki structure.

1.1 Related Work

The automatic generation of articles in Wikipedia has been a topic of interest to researchers for many years. We have seen some impressive approaches being researched upon and proven effective over the years.

Pochampally et al. (2021) showcased how we can make use of semi-supervised approaches for automatically generating articles on named entities, especially actors. For more neutral and factual domains like Science and Technology, Minguillón et al. (2017), showed how an entire corpus of Wikipedia articles on Science can be uncovered by studying the underlying graph structure of the strongly linked topics and categories. Treating the automatic generation of Wikipedia articles as a summarization task of long and detailed source documents (Liu et al., 2018) can also yield good quality results. More recently, Agarwal and Mamidi (2023) have shown how WikiData [1] can be utilized as a knowledge base to generate Hindi Wikipedia articles automatically.

We take inspiration of summarizing and simplifying the content information from Woodsend and Lapata (2011) wherein they showed how we can select the most relevant information from textual data and rewrite it in a manner understandable to even the non-native speakers of the language.

[1] https://www.wikidata.org/

2 System Overview

The system for our study can be broken up into 5 different stages. Firstly, we scrape relevant information from credible online sources, and then we proceed to clean the scraped unstructured data. After giving it some structure, we proceed to shorten the length of content by summarizing relevant information. Then, we proceed to translate the text into Hindi, and finally, we generate thousands of Wiki articles automatically by using a static template with variations.

2.1 Data Collection

As mentioned earlier, there were dedicated resources available for collecting information on Diseases and Conditions.
Firstly, SPARQL[2] queries were explored to learn about scraping data from Wikipedia specifically. The idea was to gather data from English Wikipedia about diseases. However, the diseases were not stored in specific categories and had varying sections. This made the data too unstructured, which would have caused a problem later on in the Template Creation stage. Another option to explore in Wikipedia itself was the translation of English Wikipedia articles on Diseases and Conditions into Hindi directly. However, this approach, too, had its fair share of issues. Firstly, even the English Wikipedia articles on Diseases are not as rich in information as some of the articles on online medical websites, and the depth of information varied wildly among the different articles on Wikipedia. Moreover, the number of Diseases written about in Wikipedia is limited. It was visible that many Diseases did not have a dedicated page in English Wikipedia. This would have limited our study to a very narrow use case, and the articles generated would fail to make the desired impact.

As a result, other resources dedicated solely to the medical field were explored. Web scraping using Selenium was carried out to collect the data. During the scraping phase, websites like PharmEasy and Netmeds were also considered, but due to Copyright restrictions, they were dropped from the list of potential resources. Thus, the two major resources that were used for this process were:

- Mayo Clinic[3]: The majority of the web scraping was done from this resource.
- National Health Service[4]: As some of the attributes had missing values, to gather information on some of them, this website was used.

All the required data was completely scraped from the above-mentioned resources. The scraped data was saved in a .CSV format file with the following 10 attributes:

Name, Link, Overview, Symptoms, Causes, Risk factors, Remedies, Diagnosis, Treatment, Medication

Infobox Scraping

Infoboxes are a very important part of any Wikipedia article as they help to give concise information in case the reader does not have the time to go through the entire article. Therefore, it was important to scrape information from the Diseases pages that existed on the Global English Wikipedia pages. Almost all the diseases that existed in our dataset had a page on the English Wikipedia as well. As a result, we were able to scrape all the information that existed in the diseases' infoboxes. Since the infoboxes on English Wikipedia had images as well, we scraped them too. As the images were affiliated with the WikiCommons Project, there will not be any copyright infringement issues. Ultimately, infobox scraping introduced 13 new variables in our dataset, namely:

Specialty, Symptoms, Usual onset, Duration, Causes, Risk factors, Diagnostic method, Prevention, Treatment, Prognosis, Frequency, Medication, ImageURL.

These variables were merged with the structured dataset and had the prefix *info-* attached to them in order to avoid confusion with variables of the same name (for ex: Causes, Symptoms) in the structured dataset.

2.2 Data Cleaning

The values for the non-infobox attributes listed above were basically a list of sentences. So, all the sentences were scraped, whether they were links to other sections, sponsored content, etc. This data needed cleaning, and for this, we ensured to remove links to other sections, footnotes, section headers,

[2] https://query.wikidata.org/
[3] https://www.mayoclinic.org/diseases-conditions
[4] https://www.nhs.uk/conditions/

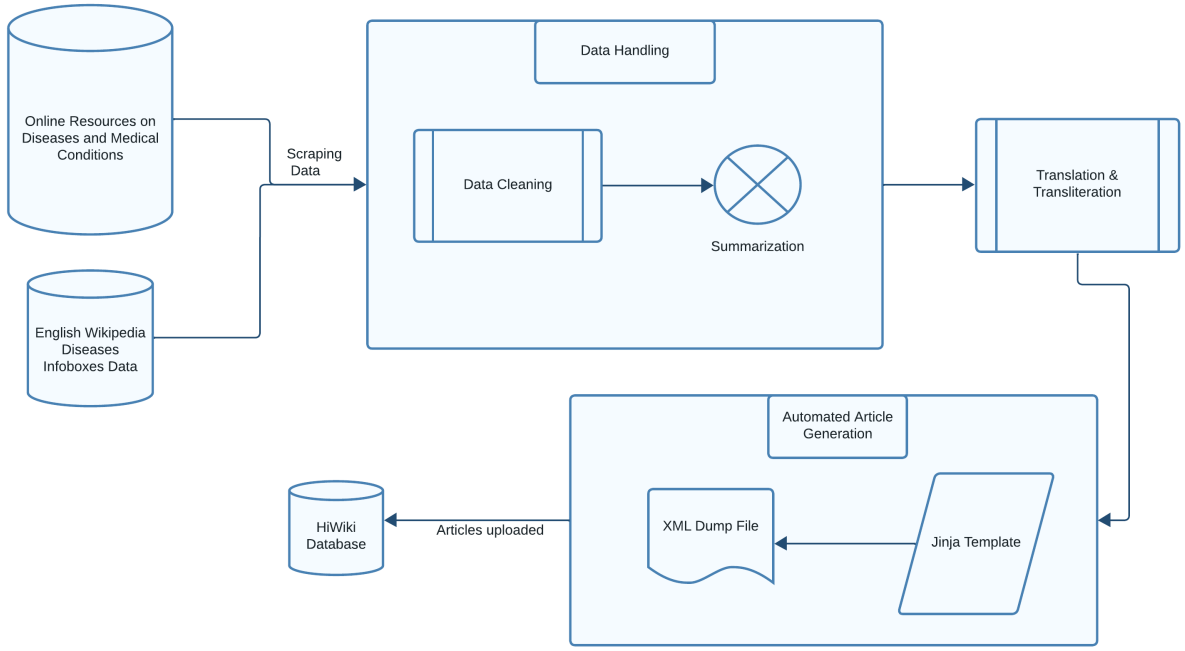

Figure 1: System Architecture: Presenting an alternate way to create Wiki Articles

duplicate sentences, and language pertinent to the website (for example, "Try our voice app", etc.)

2.3 Summarization

The third stage of our approach involved removing languages that were not in the third-person narrative and were too informal. The final article needs to have structured and formal sentences like the English Wikipedia contains for Diseases and Conditions. Therefore, a small summarizing script was deployed on the values of our dataset (sentences), and finally, only those sentences were left in the final version that were structured and contributed to the meaning of the paragraph. The concept behind the summarizing script was to pick those sentences that had words with the highest frequency, and their sentence value (sum of frequencies of words) was greater than the average sentence value.

At last, the dataset had the same attributes, but their values were fine-tuned and polished according to the study's requirements. A considerable amount of content was lost, but the informativeness and coherence did not decrease.

2.4 Translation & Transliteration

After merging the structured and infobox dataset, we had these final 24 attributes in our final merged dataset:

Name, link, link2, Symptoms, Overview, Causes, Risk factors, Diagnosis, Treatment, Remedies, Medication, info-Specialty, info-Symptoms, info-Usual onset, info-Duration, info-Causes, info-Risk factors, info-Diagnostic method, info-Prevention, info-Treatment, info-Prognosis, info-Frequency, info-Medication, ImageURL

Out of these variables, only 21 need to be translated/transliterated to Hindi as the other three, namely: *link, link2, ImageURL*, were hyperlinks. For translation, the Google Translate library was used, and for transliteration, the Indic-transliteration library was used. We experimented with both Bing[5] and Google[6] translating tools for this stage, and we found that Bing translator gives better performance when context is involved in sentences, but as our dataset sentences were context-free, Google translator gave more than a satisfactory performance.

Generally, the translation and transliteration were of a good standard, but the names of about 30 diseases were not transliterated properly. For those, the transliteration was done manually. The transliteration was shown to native Hindi speakers, who could comprehend it without any issues. All in all, the transliteration was impressively clear, making the content of the article comprehensible

[5] https://www.bing.com/translator
[6] https://translate.google.com/

for those with limited to moderate understanding of the disease or medical terminology.

2.5 Template-driven Generation

The next stage involves coding the Jinja2 template. This template, along with Python scripting, would help to generate thousands of articles automatically. In the template, all the structured dataset variables like *Overview, Causes, Medication, etc.* function like the sections for our wiki articles. For the infobox template, the official Wikipedia medical condition infobox format[7] was used. Using the official format gives the advantage of automatic translation when the Wiki Sandbox has other languages' extensions embedded.

The very final stage of the study involves the creation of an XML file that will contain all the thousands of generated articles in a single XML format file. For the same, an XML header file is created that has all the XML meta-info of the TeWiki (our copy of the Wiki environment to test and review Articles) website. Whenever a single article is generated, it is appended to this same header file. This process keeps on repeating until the very last article has been generated.

3 Results

Finally, the XML dump for all 1154 articles was generated and shared with the coordinators. The articles were imported onto the TeWiki website without any issues and can be found here.

The criteria for evaluation is entirely manual. The articles are reviewed by the editors manually and quite thoroughly. It is a lengthy process which takes time. The small errors relating to the manual aspect of the methodology, for example, transliteration, were corrected manually. Apart from those, there were no major concerns or issues raised as such, the articles were incorporated into the TeWiki website and, in due time, will be uploaded to Global Hindi Wikipedia as well. If the reviewers may have missed some errors, it would not be a major concern as the Wikipedia platform is meant for users to not only read the articles but also help improve them by pointing out mistakes and correcting them manually.

In terms of the amount of information, the generated articles had more information than their Global Hindi Wikipedia counterparts and, in some cases, even more than Global English Wikipedia pages. This is primarily due to the fact that our sources for information were NHS and Mayo Clinic medical articles, and they many a time contain more information on a particular disease than any Global English Wikipedia article as they are authored by certified medical professionals and other subject matter experts.

Finally, we consider the contribution of our study to the Wikipedia platform. As discussed earlier, Hindi Wikipedia has only a fraction of the amount of articles of that of English Wikipedia. The same holds true for the "Human Diseases and Disorders" category. English Wikipedia has approximately 1500 articles on the subject, while Hindi Wikipedia's count comes to around 500 articles. With our contribution, this number will rise to a little over a thousand articles. Even if we consider 500 articles to be repeated in both versions, our contribution still adds more than 600 articles while also adding extra information to about 500 persisting articles in Hindi Wikipedia. This clearly highlights that the automated generation of Wikipedia articles can help close or at least shorten the vast gap that exists between information available in English and native Indian languages.

4 Conclusion

Through this study, we were able to explore the effectiveness of the summarization and template-driven techniques when it comes to the automated generation of a large amount of Wikipedia articles. We demonstrated a step-by-step approach to convert unstructured textual data to some form of a mixture of structured and unstructured data and ultimately utilizing translation libraries to enrich the Indian language Wikipedia.

This goes on to show that the techniques indeed complement each other in a good way. With this study, we are able to demonstrate a very important application of NLP techniques in the modern day, i.e., producing a large quantity of good quality content in the native language of the readers on a platform which follows strict guidelines when it comes to the neutral tone of the articles. This method may be used to generate Wiki articles in any number of languages, given the availability of required Machine Translation and Transliteration tools and expert assistance for that language.

[7] https://en.wikipedia.org/wiki/Template:Infobox_medical_condition

Limitations

Like any other research study, ours, too, is filled with limitations. Overcoming some of these would directly result in better quality of articles, while some others may increase the amount of information present in the articles.

First of all, we can try to better our summarization approach by making use of famous neural-network-based approaches. This would help us to lengthen or shorten our sections appropriately.

We can also try to use abstractive summarization rather than just extractive. Implementing abstractive summarization while keeping Wikipedia's strict neutral tone in articles in mind is a challenge worth taking on. The results of such a study could prove beneficial for anyone interested in the automated generation of Wikipedia articles.

Lastly, we can try to include more linguistic variation in our template to make sure the articles produced have neutral but varied tones, as reading monotonous articles will not prove beneficial for native language readers.

References

Aditya Agarwal and Radhika Mamidi. 2023. Automatically generating hindi wikipedia pages using wikidata as a knowledge graph: A domain-specific template sentences approach. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 11–21, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences.

Julií Minguillón, Maura Lerga, Eduard Aibar, Josep Lladós-Masllorens, and Antoni Meseguer-Artola. 2017. Semi-automatic generation of a corpus of wikipedia articles on science and technology. *Profesional de la información*, 26(5):995–1005.

Yashaswi Pochampally, Kamalakar 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*, 10(2):72–79.

Kristian Woodsend and Mirella Lapata. 2011. Wikisimple: Automatic simplification of wikipedia articles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1):927–932.

Text-2-Wiki: Summarization and Template-driven Article Generation

Jayant Panwar
LTRC, International Institute of Information Technology, Hyderabad
jayant.panwar@research.iiit.ac.in

Radhika Mamidi
LTRC, International Institute of Information Technology, Hyderabad
radhika.mamidi@iiit.ac.in

A Appendix

Sample Generated Hindi Wiki Article on Migraine.
Click here for the full version of the article.

माइग्रेन

From tewiki

Contents [hide]
1 अवलोकन
2 लक्षण
3 कारण
4 जोखिम कारक
5 निदान
6 इलाज
7 उपचार
8 दवाई
9 सन्दर्भ

माइग्रेन

माइग्रेन

Specialty	Lua error in package.lua at line 80: module 'strict' not found.
Symptoms	सिरदर्द, मतली, प्रकाश के प्रति संवेदनशीलता, ध्वनि के प्रति संवेदनशीलता, गंध के प्रति संवेदनशीलता
Usual onset	यौवन के आसपास
Duration	आवर्तक, दीर्घकालिक
Causes	पर्यावरण और आनुवंशिक
Risk factors	पारिवारिक इतिहास, महिला
Prevention	मेटोप्रोलोल, वैल्प्रोएट, टोपिरामेट
Medication	इबुप्रोफेन, पेरासिटामोल (एसिटामिनोफेन), ट्रिप्टान, एर्गोटामाइन
Frequency	12.6%

अवलोकन [edit | edit source]

एक माइग्रेन गंभीर धड़कते दर्द या एक स्पंदन सनसनी पैदा कर सकता है, आमतौर पर सिर के एक तरफ। इसमें दृश्य गड़बड़ी भी शामिल हो सकती है, जैसे प्रकाश की चमक या अंधे धब्बे, या अन्य गड़बड़ी, जैसे चेहरे के एक तरफ या हाथ या पैर में झुनझुनी और बोलने में कठिनाई।

लक्षण [edit | edit source]

माइग्रेन, जो अक्सर बचपन, किशोरावस्था या शुरुआती वयस्कता में शुरू होता है, चार चरणों में आगे बढ़ सकता है: प्रोड्रोम, ऑरा, अटैक और पोस्ट-ड्रोम। माइग्रेन के हमले के बाद, एक व्यक्ति एक दिन तक थका हुआ, भ्रमित और धुला हुआ महसूस कर सकता है।

कारण [edit | edit source]

हालांकि माइग्रेन के कारणों को पूरी तरह से समझा नहीं गया है, आनुवंशिकी और पर्यावरणीय कारक एक भूमिका निभाते हैं। ब्रेनस्टेम में परिवर्तन और ट्राइजेमिनल तंत्रिका के साथ इसकी बातचीत, एक प्रमुख दर्द मार्ग, शामिल हो सकता है। कुछ न्यूरोट्रांसमीटर माइग्रेन के दर्द में भूमिका निभाते हैं, जिसमें कैल्सीटोनिन जीन-संबंधित पेप्टाइड (सीजीआरपी) शामिल है। महिलाओं में हार्मोनल परिवर्तन सहित कई माइग्रेन ट्रिगर हैं। हार्मोनल दवाएं, जैसे कि मौखिक गर्भ निरोधकों और हार्मोन रिप्लेसमेंट थेरेपी, भी माइग्रेन को खराब कर सकती हैं। हालांकि, कुछ महिलाएं इन दवाओं को लेते समय अपने माइग्रेन को कम बार पाती हैं। तेज रोशनी और सूरज की चकाचौंध से तेज आवाज के साथ-साथ माइग्रेन भी हो सकता है। परफ्यूम, पेंट थिनर, सेकेंड हैंड स्मोक और अन्य सहित तेज गंध कुछ लोगों में माइग्रेन को ट्रिगर करती है। नींद न आना, बहुत अधिक नींद लेना या जेट लैग कुछ लोगों में माइग्रेन को ट्रिगर कर सकता है। यौन गतिविधि सहित तीव्र शारीरिक परिश्रम, माइग्रेन को भड़का सकता है। मौखिक गर्भनिरोधक और वैसोडिलेटर, जैसे नाइट्रोग्लिसरीन, माइग्रेन को बढ़ा सकते हैं। कारणों में कई खाद्य पदार्थों में पाए जाने वाले स्वीटनर एस्पार्टेम और परिरक्षक मोनोसोडियम ग्लूटामेट (एमएसजी) भी शामिल हैं।

जोखिम कारक [edit | edit source]

पारिवारिक इतिहास सहित कई कारक एक व्यक्ति को माइग्रेन होने का अधिक खतरा बनाते हैं। माइग्रेन किसी भी उम्र में शुरू हो सकता है, हालांकि पहली बार किशोरावस्था के दौरान होता है। माइग्रेन 30 के दशक के दौरान चरम पर होता है और बाद के दशकों में धीरे-धीरे कम गंभीर और कम

Figure 1: Snapshot of a Sample Generated Wiki Article.