# WikiTransfer: Knowledge transfer from High-Resource to Low-Resource Language in Multilingual Wikipedia

**Paramita Das**
IIT Kharagpur, India

**Amartya Roy**
Bosch, India

**Animesh Mukherjee**
IIT Kharagpur, India

## Abstract

To address content disparities in multilingual Wikipedia versions, in this work, we propose a lightweight framework called **WikiTransfer** that can help enhance knowledge integrity across diverse linguistic communities. The framework employs powerful machine translation techniques to translate text from high-resource to low-resource language for transferring information. Our proposed framework aims to improve overall content quality, demonstrated through its effectiveness in enhancing the quality of Wikipedia articles at the section level. Although We have shown a case study of knowledge transfer from English to Hindi, WikiTransfer can be adapted to transfer potential knowledge across various language pairs.

**Keywords:** Multilingual Wikipedia, Low-resource language, Machine translation, Quality, Biographies.

## Overview

Wikipedia's decentralized structure and autonomous communities in various languages have positioned it as a "global repository of knowledge." Its widespread accessibility has resulted in its emergence as a popular encyclopedic source in many linguistic contexts especially in the field of Natural Language Processing (NLP) research and development, where Wikipedia serves as a fundamental training data source for large language models. However, previous studies (Miquel-Ribé and Laniado, 2020; Roy et al., 2020) have identified that the content of Wikipedia articles on a particular topic in different language versions can vary significantly. This content disparity can be attributed to several factors, such as the variability in the availability of information based on the cultural, historical, or regional relevance of that topic to the editors of a particular language. Secondly, the contributors to each language version may have different expertise levels, perspectives, or priorities, leading to variations in the coverage of an article in multiple languages. Identifying methodologies to address these disparities is a pivotal aspect of ensuring 'knowledge equity,' a concept introduced by the Wikimedia Foundation (Redi et al., 2020).

**Our contributions:**

Our goal is to address the content disparity between high and low-resourced languages, in which English is assumed the high-resourced language and Hindi is the low-resourced one in our work. Despite both languages being among the top ten most widely spoken in the globe, we have noticed a significant gap in content coverage in Hindi biographies compared to their English counterparts. In light of this observation, our research is guided by the following two key questions- **RQ1:** Given an article in two languages, i.e., Hindi and English, how can one determine which language version lacks sufficient information and knowledge on that specific article compared to the other? **RQ2:** How can one migrate content from the more enriched language version to the less enriched one, usually high-resource to low-resource?

To address RQ1, we utilized the quality scores of article revisions released by researchers (Das et al., 2024) and selected a set of articles that have less score of the Hindi version as compared to the English one. Our underlying hypothesis is that the lesser the quality score, there is a requirement for content enrichment. Next, to find the solution as mentioned in RQ2, we employed state-of-the-art machine translation models to facilitate the translation of Wikipedia article text for transferring missing content from English to Hindi. Our work-in-progress pipeline shows that our straightforward framework can assist the platform in automatically improving content in low-resourced languages by leveraging well-established content from high-resourced languages.

## Dataset Description

In our work, As we have concentrated on Hindi and English, a set of $4k$ of biographies available in both languages along with corresponding Wikidata IDs are extracted from the publicly available dataset by the authors in their work (Beytía et al., 2022). As previously mentioned, these $4k$ articles have the quality score of the English version more than the corresponding Hindi version. Next, we extracted the current version of these Wikipedia articles from the latest publicly available XML dumps as updated in December 2022 (as the quality scores are calculated at the same time). Later we pre-processed the

retrieved text and utilized the python package wikipedia[1] to extract the section headings for every article. Please note that we have not considered the sections– *See also*, *Notes*, *References*, *Further reading*, *External links* etc. in our pipeline.

## Framework

For a given biography article $P$, we have two versions in our dataset: the Hindi version $H_P$ and the English version $E_P$. Our objective is to enrich the content of sections in $H_P$ by incorporating content from the corresponding sections of $E_P$. The descriptive picture of the pipeline is shown in Figure 1.

**Sections mapping:** To achieve the mapping between the sections, we computed the embedding using the sentence transformer model [2] for each section heading and measured the cosine similarity of embedding between every pair of section headings in Hindi and English pages. We considered the section heading pairs with maximum similarity and thus a section heading in the English article, say $t_e$ is mapped to the Hindi section heading $t_h$.

**Machine translation of English content to Hindi:** After establishing the section mapping between English and Hindi pages, the corresponding English content is translated into Hindi using the language model named IndicTrans2 (Gala et al., 2023). Mathematically, the English page $E_P$, which has $e$ sentences in section $t_e$ are translated into $e$ Hindi sentences, denoted as $Hindi(e)$ and are then appended to the existing $h$ sentences in mapped section $t_h$ of Hindi Page $H_P$. However, we did not append the whole translated content (the English section content) to the existing Hindi section content. We measured the similarity of the translated content and the existing content of a section (in the Hindi version) and identified candidate Hindi-translated sentences that are semantically dissimilar to the existing Hindi sentences of the section. These filtered sentences are considered as a piece of new knowledge to be added to the Hindi article.

## Result

To evaluate the relatedness and quality of generated content, we employed three Information Quality (IQ) metrics: Informativeness, Readability, Understandability, and Quality, as outlined in Sugandh et al.'s work (Sugandhika and Ahangama, 2022). Due to the scarcity of resources, particularly unsupervised lexical methods for assessing Hindi text quality, we measured quality in the English domain instead. This involved translating the relevant portion of Hindi text into English and computing the metrics for both old and new English content. Sections with new content scores surpassing their old content scores were deemed suitable for addition. Average values of Informativeness, Readability, Understandability, and Quality for these sections are presented in Table 1. To evaluate the quality of the generated content, we randomly select the content of 50 sections. Each sample corresponds to a set of two paragraphs in the Hindi language– *before* and *after* running our WikiTransfer framework and the annotators had to evaluate the *after* paragraph in terms of Informativeness, Readability, and Understandability. The inter-annotators agreement measured by Cohen's Kappa is as follows– Informativeness (0.61), Readability (0.58) and Understandability (0.28).

## Future Direction

In our study, we introduced a lightweight approach aimed at improving information across diverse linguistic communities. The results presented in Table 1 demonstrate a significant increase in the quality of old content in Hindi by leveraging knowledge transfer from high-resource languages like English. Operating at the section level of Wikipedia articles, our framework is adaptable and can be applied to enhance content quality across any language pair. In the future, we plan to include incorporating knowledge from external book corpora, such as biographies and autobiographies, into low-resourced language Wikipedia content using our pipeline. In this context, we aim to tackle the challenge of *neutral point of view* (NPOV) issue. Furthermore, we plan to integrate additional evaluation techniques and large-scale human assessments to ensure the coherence of newly generated content with the existing one.

## References

[Beytía et al.2022] Pablo Beytía, Pushkal Agarwal, Miriam Redi, and Vivek K. Singh. 2022. Visual gender biases in wikipedia: A systematic evaluation across the ten most spoken languages. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):43–54, May.

[Das et al.2024] Paramita Das, Isaac Johnson, Diego Saez-Trumper, and Pablo Aragón. 2024. Language-agnostic modeling of wikipedia articles for content quality assessment across languages. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1924–1934.

[Gala et al.2023] Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

[Miquel-Ribé and Laniado2020] Marc Miquel-Ribé and David Laniado. 2020. The wikipedia diversity ob-

[1] https://pypi.org/project/wikipedia/
[2] https://sbert.net/

servatory: A project to identify and bridge content gaps in wikipedia. New York, NY, USA. Association for Computing Machinery.

[Redi et al.2020] Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. 2020. A taxonomy of knowledge gaps for wikimedia projects (second draft). *arXiv preprint arXiv:2008.12314*.

[Roy et al.2020] Dwaipayan Roy, Sumit Bhatia, and Prateek Jain. 2020. A topic-aligned multilingual corpus of Wikipedia articles for studying information asymmetry in low resource languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2373–2380, Marseille, France, May. European Language Resources Association.

[Sugandhika and Ahangama2022] Chinthani Sugandhika and Supunmali Ahangama. 2022. Assessing information quality of wikipedia articles through google's e-a-t model. *IEEE Access*, 10:52196–52209.

| Metric | Old content | New content |
|---|---|---|
| Informativeness | 34.88 | 55.02 |
| Readability | 4.78 | 4.84 |
| Understandability | 17.69 | 17.45 |
| Quality | 21.88 | 26.92 |

Table 1: Table showing the averaged value of evaluation metrics for old and new content.

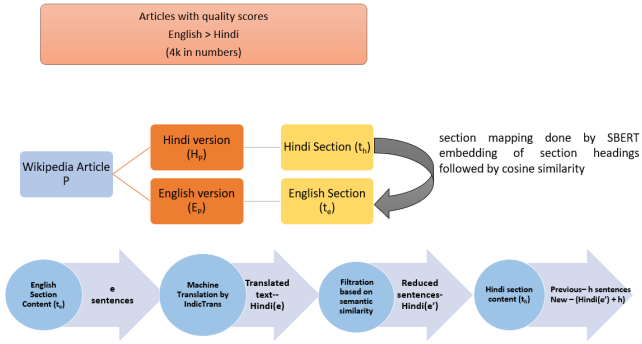


Figure 1: Block diagram of the pipeline and its different modules