Research into Translation Imbalances

Abhishek Bhardwaj
Cluster Innovation Centre, University of Delhi

Abstract

This study aims to investigate translation imbalances within the Content Translation tool, specifically examining the significant disparities, often at a ratio of 100:1 (Link), between translations made in each direction of a language pair. The research endeavors to address the underlying cause of these imbalances, seeking to determine whether they arise organically or result from potential biases inherent in the current UI/UX and underlying algorithms of the tool.
By analyzing the patterns and factors influencing translation ratios, this study aims to uncover whether these disparities are a product of natural user behavior or if they indicate inadvertent biases towards specific languages. The findings will shed light on potential areas for improvement in the tool's design and functionality to ensure a more equitable and balanced translation experience across all language pairs.

Introduction

- **Problem Statement:**
  The research aims to investigate the notable imbalances observed in translations facilitated by the Content Translation tool. There exists a considerable discrepancy, often at a ratio of 100:1, between translations made in different directions within language pairs. This study seeks to address why these imbalances occur—whether they stem from inherent biases within language selection algorithms or result from natural user behavior.
- **Significance for Wikimedia Projects**:
  The impact of these imbalances on Wikimedia projects is substantial, particularly for communities associated with smaller languages. Evidence suggests that smaller Wikipedias receive fewer translations into other languages, limiting access to unique articles they possess. Addressing this issue is pivotal for creating a more inclusive translation environment. Successful conclusions from this research will benefit Wikimedia communities by enabling increased translation of exclusive articles, fostering a more comprehensive and diverse knowledge repository.
- **Research Questions:**
  - **Translation Dynamics**: What were the language options available to translators at a given time, their proficiency in these languages, and the availability of articles in those languages during the translation timestamp? This analysis aims to discern whether translators explore all available options or predominantly select from a limited pool due to specific reasons.
  - **Temporal Changes:** How have user preferences for article choices evolved over time

within each language since the inception of the Content Translation tool?
    - **Article Suggestor Influence**: To what extent does the article suggestor impact users' language choice during translation processes?

- **Research Duration:**
  This proposal spans from June 1, 2024, to August 31, 2024, aiming to gather comprehensive data and derive insights to address the outlined research questions.

Related work

1. Self-Reported Proficiency and Language Choice:
During a past internship, I investigated how self-reported translator proficiency influences language selection, revealing insights into how personal language skills shape translation choices.
2. "Digital Divisions of Labor and Informational Magnetism":
This study mapped editor participation across Wikipedia languages, providing a broader understanding of participation dynamics within Wikimedia projects.
Contribution and Advancement:
Existing studies laid groundwork by exploring language choice, article preferences, and participation dynamics. Our research aims to delve deeper, investigating the specific causes behind translation imbalances in the Content Translation tool. By examining language options, temporal preferences, and article suggestor influence, our study aims to offer nuanced insights into these imbalances, contributing to a more comprehensive understanding of translation dynamics within Wikimedia.

Methods

- **Data Collection:**
    - **Sources**: Scraping Wikipedia pages, querying Wikimedia and Wikipedia databases, and using Wikipedia dumps.
    - **Collection Approach:** Utilizing web scraping, database queries, and dumps to gather comprehensive data on language options, article availability, and user choices.
- **Tools and Analysis:**
    - **Data Processing:** Python, R, and Orange for cleaning, processing, and transforming the dataset.
    - **Visualization:** Python, R, and Orange for clear visual representations of translation patterns and trends.
    - **Hosting:** Project materials hosted on GitHub for transparency and collaboration in further research and development.
- **Analysis Approach:**
    - **Exploratory Data Analysis:** Descriptive statistics to explore language dynamics and user behaviors.
    - **Inferential Analysis:** Statistical methods to infer relationships between translator behavior and language/article availability.

Expected output

- **Insights to Inform Decision Making:**
    - **Primary Audience:** Language Engineering Team, Wikimedia Community

    - **Benefit**: Improved language selection algorithms within the Content Translation tool, enhancing user experience and facilitating more balanced translations across language pairs.
- **Scientific Publication:**
    - **Primary Audience:** Researchers, Academics, and Interested Public
    - **Benefit**: Sharing research findings through publication in relevant journals or conferences, contributing to the academic understanding of translation dynamics and biases in digital environments.
- **Public Datasets and Python Modules:**
    - **Primary Audience**: Researchers, Developers, Wikimedia Community
    - **Benefit**: Providing accessible public datasets for further research and analysis, facilitating transparency and collaboration within the research community. Additionally, reusable Python modules will aid in replicating or extending the study for future investigations.
- **Potential Publication Venues:**
    - **Scientific Publication:** Conferences such as the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), the International Conference on Web and Social Media (ICWSM).

Community impact plan

- **Open Access to Resources:**
    - **Public Datasets and Tools:** Making research datasets and developed Python modules openly accessible on platforms like GitHub for researchers, developers, and Wikimedia contributors to explore, utilize, and contribute to further improvements.
    - **Documentation and Tutorials:** Creating user-friendly documentation and tutorials to assist Wikimedia volunteers or developers in understanding and utilizing the datasets and tools effectively.
- **Contributions to Wikimedia Improvement:**
    - **Contribution to Tool Enhancements**: Offering suggestions and potential improvements derived from the research outcomes, aiming to directly impact the design and functionality of translation tools within Wikimedia projects.

Evaluation

- **Algorithm Improvement:**
    - Enhanced language selection, resulting in more balanced translations.
- **Dissemination and Adoption:**
    - High visibility and utilization of research outputs.
- **Engagement and Collaboration:**
    - Active participation and contributions from Wikimedia communities and the Language Engineering Team.
- **Contribution to Knowledge:**
    - Recognition in academic circles for insights into translation dynamics and biases.

Budget

Stipend - $10,000
System requirements - $2,000

Prior contributions

In a prior Outreachy research role, I scrutinized knowledge distribution within Wikipedia, focusing on imbalances in translation flows using the Content Translation tool. I identified technical and social factors contributing to disparities in translation ratios. This experience underpins my commitment to rectifying these imbalances and promoting more equitable knowledge dissemination on Wikipedia, aligning with the current proposal's goal.

References

- Research:Increasing article coverage
- Graham, Mark; Straumann, Ralph K.; Hogan, Bernie (2015-11-02). "Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia". *Annals of the Association of American Geographers* **105** (6): 1158–1178. ISSN 0004-5608. doi:10.1080/00045608.2015.1072791.
- McDonough Dolmaya, Julie (2017-04-03). "Expanding the sum of all human knowledge: Wikipedia, translation and linguistic justice". *The Translator* **23** (2): 143–157. ISSN 1355-6509. doi:10.1080/13556509.2017.1321519.