Wiki Workshop (10th edition) – May 11, 2023

Detecting Cross-Lingual Information Gaps in Wikipedia

Vahid Ashrafimoghari
School of Business
Stevens Institute of Technology
Hoboken, NJ, USA

Jordan W. Suchow
School of Business
Stevens Institute of Technology
Hoboken, NJ, USA

Abstract

An information gap exists across Wikipedia's language editions, with a considerable proportion of articles available in only a few languages. As an illustration, it has been observed that 10 languages possess half of the available Wikipedia articles, despite the existence of 330 Wikipedia language editions. To address this issue, this study presents an approach to identify the information gap between the different language editions of Wikipedia. The proposed approach employs Latent Dirichlet Allocation (LDA) to analyze linked entities in a cross-lingual knowledge graph to determine topic distributions for Wikipedia articles in 28 languages. The distance between paired articles across language editions is then calculated. The potential applications of the proposed algorithm to detecting sources of information disparity in Wikipedia are discussed, and directions for future research are put forward.

Keywords: Document-level Similarity, Computational Linguistics, Topic Modeling, Wikipedia, Cross-lingual Gap

Introduction

Wikipedia articles in different language editions that are purportedly about the same topic are created by different editors and thus reflect the biases and perspectives of their respective editors (Callahan and Herring, 2011). Therefore, there is a significant gap in the information presented across different language editions of Wikipedia (Wulczyn et al., 2016). Some portions of the content may be direct translations while others may have been created independently and have little thematic or lexical overlap. The similarity between texts can impact the quality of resources generated for tasks such as Cross-Language Information Retrieval (CLIR) or Statistical Machine Translation (SMT) as non-similar documents can introduce additional noise into Machine Translation (MT) task and affect its performance(Bonab et al., 2020).

This paper introduces a new method to measure the information gap present across language editions of Wikipedia. By allowing Wikipedia users of all language editions to access information that would otherwise be unavailable, we expect this work to significantly promote global knowledge sharing and bridge the information divide between different communities and regions.

Methods

Data Retrieval

To implement our proposed approach, we first selected 28 languages: English, German, Swedish, Chinese, Japanese, French, Spanish, Farsi, Arabic, Polish, Hebrew, Italian, Vietnamese, Waray, Russian, Dutch, Ukrainian, Portuguese, Korean, Turkish, Finish, Serbian, Catalan, Indonesian, Albanian, Romanian, Czech and Greek. Next, we retrieved the data dumps of these language editions (2022-06-01 release) from the Wikimedia data repository. Then, we extracted page IDs from the XML data dumps and matched page IDs with QIDs before applying the similarity measure to calculate pairwise distances between articles. Moreover, to obtain an overview of the current status of article imbalance among various Wikipedia language editions, the English edition was chosen as the baseline, and the information covered in all the other editions was compared against it with the help of Wikimedia Statistics.

Document-level Similarity

To compare the content across paired articles, we used Polyglot Dirichlet Allocation (WikiPDA), which is a cross-lingual topic model that works at the QID level and deploys Latent Dirichlet Allocation (LDA) to learn a representation of Wikipedia articles as distributions over a common set of language-independent topics derived from the link structure of Wikipedia (Piccardi and West, 2021).

To measure the similarity between the topic distributions, we employed cosine similarity which measures the similarity between two non-zero vectors by determining their orientation, not magnitude. The smaller the cosine similarity, the greater the dissimilarity between the vectors.

Human-based Evaluation

When presented with a pair of Wikipedia articles in a specific language, native speakers of the non-English language and fluent in English are asked to read the articles and provide their opinion on the overall similarity and specific aspects such as structure, length, interlinked entities, and content overlap regarding the cross-lingual information gap taxonomy proposed in (Johnson and Lescak, 2022).

MT-based Evaluation

We use Google Translate and set English as the reference language to create a comparable corpus of English and non-English article pairs. Subsequently, the Latent Dirichlet Allocation (LDA) method obtains monolingual topic distributions for the selected article pairs. Finally, the cosine similarity metric is applied in this study to compare a pair of articles based on their topic distributions.

Results

Table 1 summarizes descriptive statistics pairwise similarity calculation for all article pairs in 28 language editions of Wikipedia. We calculated the degree of similarity between all language editions by computing the average cosine similarity between the topic vectors of each paired article, as depicted in Figure 1. We then extracted a list of candidate article pairs ranked according to the degree of the information gap. With the exception of the English and Arabic editions, the top five articles in the other languages mainly pertained to renowned personalities, organizations, locations, or scientific topics. Notably, among the top-five highly distant articles for the English–Farsi language pair, there was a topic concerning headscarves, where courageous women in Iran are presently demonstrating against their government to secure the right to choose their attire. This finding highlights that authoritarian regimes may utilize Wikipedia to manipulate their societies' perceptions and propagate their ideology online. Thus, reducing the information gap can be viewed as a countermeasure against censorship, enabling people in communities where access to information is restricted to obtain information that is not biased by propaganda narratives.

An experimental evaluation was conducted on a sample of 100 article pairs from the English-Farsi language pair. This language pair was selected due to the author's fluency in Farsi and English as a native Farsi speaker. The evaluation process involved thoroughly examining each article pair by the author, who then assigned a similarity score on a scale of 0 to 100. The author meticulously recorded any possible sources or causes of dissimilarity encountered during the evaluation. The evaluation results, as revealed by the Spearman non-parametric test, indicated that the similarity scores generated by the proposed algorithm were significantly correlated with human judgment (ρ=0.353, $p < 0.001$).

Finally, a detailed examination of a sample of paired articles from the English and Farsi language editions revealed that mismatches between the interlinked entities of articles on a given topic are likely the most prevalent causes of information disparities between the two editions. These mismatches stem from differences in the size, age, and composition of the Farsi edition's editor community and the editors' areas of interest. Additional factors contributing to information disparities include the presence of outdated, culturally dependent, bot-generated, or geographically dependent content, censorship and propaganda, the unavailability of sources in one of the language editions, the presence of controversial topics, mislabeling, and instances of vandalism.

Conclusions & Future Work

This work presented a novel algorithm for detecting information disparities between paired articles from different language editions of Wikipedia. In a preliminary evaluation, the proposed algorithm was found to be useful for classifying the potential sources of information gaps in Wikipedia. Future research avenues include improving the accuracy of the proposed algorithm using statistical measures and evaluating it using additional language pairs and machine-based methods.

References

[Bonab et al.2020] Hamed Bonab, Sheikh Muhammad Sarwar, and James Allan. 2020. Training effective neural clir by bridging the translation gap. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 9–18.

[Callahan and Herring2011] Ewa S Callahan and Susan C Herring. 2011. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915.

[Johnson and Lescak2022] Isaac Johnson and Emily Lescak. 2022. Considerations for multilingual wikipedia research. *arXiv preprint arXiv:2204.02483*.

[Piccardi and West2021] Tiziano Piccardi and Robert West. 2021. Crosslingual topic modeling with wikipda. In *Proceedings of the Web Conference 2021*, pages 3032–3041.

[Wulczyn et al.2016] Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. 2016. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 975–985.

	Similarity Scores
Count	103,144,900
Mean	0.71
Std	0.3
Min	0
25%	0.55
50%	0.83
75%	0.96
Max	1

Table 1: Descriptive statistics of pairwise similarity calculations for all article pairs in 28 languages

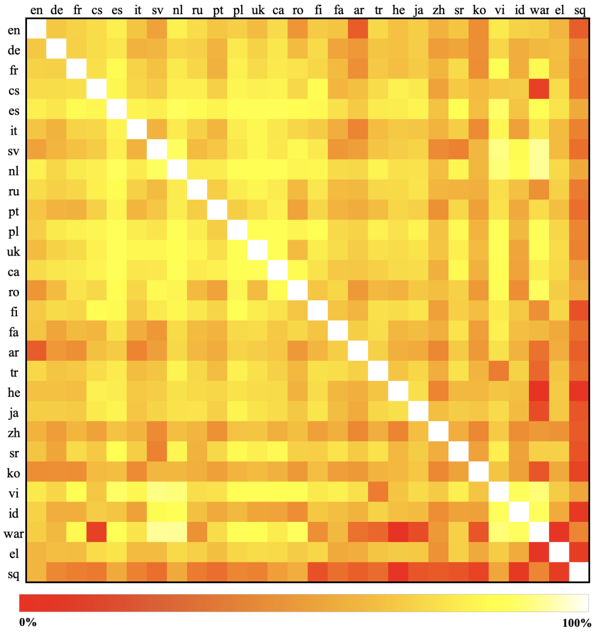

Figure 1: The degree of similarity between paired articles across 28 Wikipedia language editions

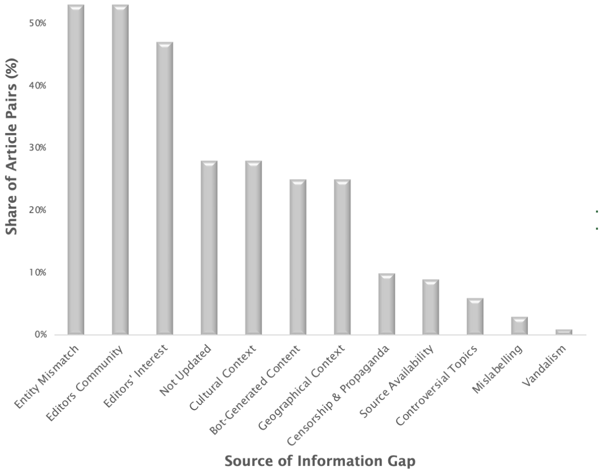

Figure 2: Share of potential sources of information gap between English and Farsi language editions