Can Machine Translation Narrow Knowledge Gap across Languages? A Large-scale Multilingual Analysis of the Partnership between Google Translate and Wikipedia

Kai Zhu
McGill University
Montreal, Quebec, Canada
kai.zhu@mcgill.ca

Xin Yue Zhou
University of Oklahoma
Norman, USA
xy.zhou@ou.edu

ACM Reference Format:
Kai Zhu and Xin Yue Zhou. 2022. Can Machine Translation Narrow Knowledge Gap across Languages? A Large-scale Multilingual Analysis of the Partnership between Google Translate and Wikipedia. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Knowledge gap, often known as disparity in distribution of information and knowledge throughout a social system, widely exists in online space and in digital systems. Despite being one of the most successful open collaboration platforms, Wikipedia also suffers from this problem [1, 2]. With the objective of providing "free access to the sum of all human knowledge", Wikipedia is now part of the essential infrastructure of knowledge repositories in digital space. However, as pointed out by 2030 Wikimedia Strategic Direction [3], it is becoming increasingly critical to address the knowledge gap on Wikipedia so that it can better serve audiences, communities, and cultures that have been traditionally left out by structure of power and privilege. However, knowledge gap across languages is a notoriously challenging issue as it is difficult to recruit volunteers to contribute content in low-resource languages. In this study, we examine if and how state-of-art neural machine translation can narrow the knowledge gap across different language editions of Wikipedia.

Wikimedia Foundation leverages machine translation to support their editors by allowing them to create an initial translation of an article from other language editions of Wikipedia that the underlying "concept" has already existed. Wikipedia editors can select from several machine translation systems in its in-house Content Translation toolbox to support an initial article translation. After the draft is created via translation, editors can then review, edit, and improve. In January of 2019, Wikimedia Foundation integrated Google Translate to its Content Translation toolbox. The introduction of Google Translate as a state-of-art neural machine translation service enables editors to transfer knowledge to more target languages and with translations of higher quality. Despite the mixed sentiments among editors toward the role of machine intelligence in knowledge production on Wikipedia, we observe a large and sharp increase in the translation volume shortly after the roll-out of Google Translate on Wikipedia in our empirical analysis. In Figure 1, we show that the number of articles created with machine translation per month on Wikipedia increased immediately and steadily after the integration of Google Translate in January 2019.

This partnership between Wikipedia and Google Translate presents us with a great opportunity to gain a deeper understanding of if and how a new technology enabled by Artificial Intelligence can narrow the knowledge gap in social-technical systems. Leveraging this unique setting, we use tools and techniques from econometrics modeling, causal inference, and natural language processing to investigate three sets of closely related questions on the impact of machine translation on Wikipedia. First, how does a better machine translation service enable knowledge transfer between different languages? Does it mostly support knowledge outflow from a few major language editions like English and French? Or does it support a bidirectional and hence more mutual information exchange between different language editions of Wikipedia? Second, how does Google Translate change the collaboration and coordination pattern between human editors and machine intelligence? Specifically, how do the human editors change their roles in the process of content production when there is a good initial translation created by machines? Third, a large portion of articles from each Wikipedia language edition is locally-relevant and culture-specific content. Does machine translation also help the spreading of local content?

In this extended abstract, we provide some initial analysis of the above proposed questions. First of all, we present a counterfactual estimation of the impact of Google Translate on translation volume in the panel data setting in Figure 2. The stagger roll-out of Google Translate at different time to a subset of more than 300 language editions of Wikipedia enable us to estimate the average treatment effect for the treated language (ATT) by directly imputing the counterfactual outcomes where Wikipedia had not adopt Google Translate for a language at a given time. This is in analogy to the difference-in-difference analysis in spirit and the counterfactual estimation framework is more flexible and robust to heterogeneous treatment effect or when unobserved time-varying confounders exist. Figure 2 shows that the estimated ATT for a treated language edition of Wikipedia remains indistinguishable from zero before the treatment and becomes significantly above zero and steadily growing after the introduction of Google Translation. Averaged over the post-treatment period (i.e. between Jan 2019 and Dec 2021)

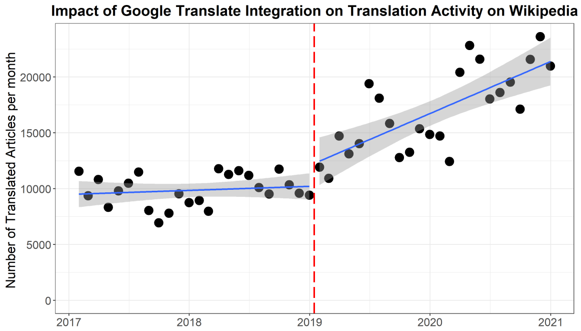

Figure 1: Google Translate was integrated to the Content Translation tool of Wikipedia on January 2019. It has an immediate impact on translation activity on Wikipedia

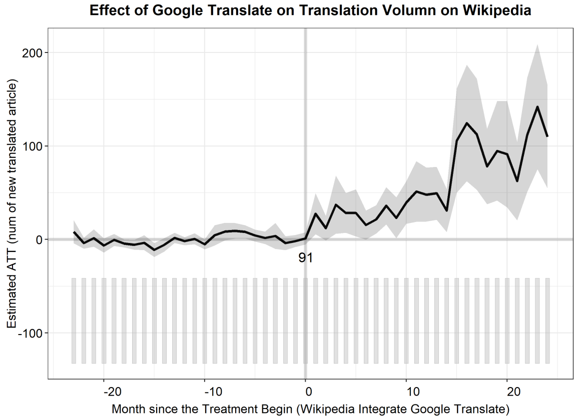

Figure 2: Based on the counterfactual estimation framework, Google Translate increase 62 new articles for treated Wikipedia language edition. The standard error for this estimated ATT is 15 (articles).

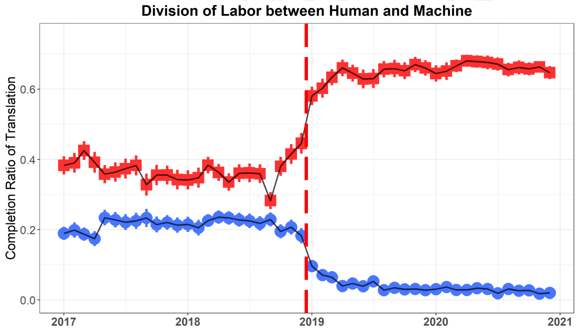

Figure 3: Division of labor between human intelligence and machine intelligence changes after the integration of Google Translate

and over treated languages, Google Translate increased 62 new articles for treated Wikipedia language edition. The standard error for this estimated ATT is 15 (articles).

In addition to the volume of knowledge transfer through machine translation, Google Translate as a new technology also changes the collaboration and coordination pattern between human editors and machine intelligence. As an initial peak into this question, we analyzed the division of labor between human and machine when translating an article. Wikipedia articles break down to "sections". When editors initialize a translation of an article, they can decide to translate as many or as few sections in the source article as they want as well as how to translate those sections (i.e. using machine translation systems or translate by themselves). We computed the completion ratio of translation of machine/human as number sections translated by machine/human divided by total number of sections in source article. As shown in Figure 3, there is a clear shift in the division of labor between machine intelligence and human editors (Red squares represent machine translation and blue dots represent translation by human editors. Error bars are two standard errors of the mean in a given month). We can see in Figure 3 that as the technology advances, as in the case of adoption of Google Translate, a larger portion of the knowledge production process (via translation) is allocated to machine intelligence. Right now, we allocate the credits to machines as long as human editors request a machine translation draft of a section. However, human editors often are also involved as they will review, revise, and improve the results from machine translation. For the next step of analysis, we will have a deeper look into how human editors interact with the draft from machine translation as another subtle and interesting way of coordination and collaboration between human and machine intelligence.

In the time leading up to the conference date of Wiki Workshop 2022, we will analyze the up-to-date revision history of translated articles and examine how other editors continue building upon the initial content created with machine translation. A positive growth dynamic would shed light on another channel through which machine translation can benefit content growth. Moreover, we will dive deeper into the textual and topical content of translated articles. In particular, we will investigate that if machine translation help the dissemination of locally-relevant and culture=specific content across languages. This will gain us insights regarding if machine intelligence can contribute to diversity and inclusion of knowledge on digital platforms.

REFERENCES

[1] Mark Graham, Bernie Hogan, Ralph K Straumann, and Ahmed Medhat. 2014. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers* 104, 4 (2014), 746–764.

[2] Kai Zhu, Dylan Walker, and Lev Muchnik. 2020. Content Growth and Attention Contagion in Information Networks: Addressing Information Poverty on Wikipedia. *Information Systems Research* 31, 2 (2020), 491–509.

[3] Leila Zia, Isaac Johnson, Bahodir Mansurov, Jonathan Morgan, Miriam Redi, Diego Saez-Trumper, and Dario Taraborelli. 2019. Knowledge Gaps – Wikimedia Research 2030. (2019). https://upload.wikimedia.org/wikipedia/commons/3/31/Knowledge_Gaps_%E2%80%93_Wikimedia_Research_2030.pdf