Response to Comments from Reviewers and Committees

Thanks for the constructive and valuable input from the reviewers and committees. We found the feedback very helpful in helping to clarify our research plan as well as further expanding and refining our research scope. We are extremely grateful to all the reviewers for their valuable input!

Below, we respond to individual comments and hope this will provide more information as well as greater clarity. Texts in Italic style are direct quotes from reviewers and texts in non-Italic style are our responses.

Reviewer 1

"The impact of tooling like translation can be quite nuanced, variable across communities, and difficult to capture through standard quantitative approaches."

Indeed, the impact of the tool for machine translation can be nuanced and heterogenous across different Wikipedia language editions. This presents us with both a challenge and a good opportunity to understand how the impact of the same technology can be different depends on the characteristics of the languages and characteristics of the communities. We will be mindful about this heterogeneity when conducting our
research and plan to investigate this from both quantitative and qualitative approaches.

For quantitative analysis, we will adopt modeling strategy from heterogenous treatment effect literature (Athey and Wagner, 2019; Athey et al., 2019; Mayer et al., 2020) and make sure we won't draw any one size-fit-all conclusions. In the statistical estimation and inference stage, we will explore heterogenous impact in both automatic and pre-defined subgroups of Wikipedia language editions. Specifically, we are particularly interested in how the impact of machine translation will differ for low-resourced vs. high resourced language. The reasoning being low-resourced language edition of Wikipedia face a bigger challenge in recruiting human editors and hence tools for making content contribution easier will benefit more for those communities. For the qualitative approach, please see ourresponse for the question below.

"My main ask would be to know what pathways the recipient might take for forming stronger bonds with Wikimedian communities such that they might help guide the research questions and analyses."

Thanks for the great suggestion. We agree that forming bonds with Wikipedian communities will be an essential component for understanding the impact of machine translation tool. We foresee three approaches we can take to achieve it.

- First, we could participate in the talk pages of translated articles and engage in dialogue with actual users of the Content Translation tool. This may allow us to gain insights about why they use machine translation as well as how and the extent to which they find machine translation helpful or unhelpful in their content production process.
- Second, some communities have their own policies regarding using machine translation in content creation. For instance, English Wikipedian community disables the Content Translation for regular users, presumably because they have relative more abundant human editor resources (hence less need for machine translation) and have higher quality standard for their articles. We

are in the process of figuring out a way to systematically collect the translation policies for all language editions and the community deliberation process behind them.

- Third, we plan to conduct interview with editors and community leaders to have a better understanding of their views regarding the role of machine translation in creating new articles in their language editions. We believe that this will help inform the research question and analysis. This is on our to-do list, and we would like to find out who to reach out and the best way to reach out to them. We believe this grant and the potential collaboration may help with this part. We welcome any other suggestions that can help us become more engaged with Wikipedia editors.

Reviewer 2

Comments in "Advancement of Knowledge" section regarding our research questions

RQ1: "'First, how does a better machine translation service enable knowledge transfer between different languages.' *On first pass it seems that this is already answered in the explanation in the first few paragraphs of the proposal: i.e. the number of translations have sharply increased, etc. What exactly is the question you want to answer?"*

Yes, we did find total number of new translated articles increased sharply after the integration of Google Translate (See Figure 1 in Project Description). We would like to go beyond total number translated articles across all language editions and have a closer and more nuanced look at information exchange between different language editions of Wikipedia. For instance, does machine translation 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? More content creation in less developed language editions of Wikipedia is desirable generally, but we need be cautious against the domination of English language and its point of view being extended via new and insidious way like machine translation. More broadly, we want to investigate what the distribution is of the target and source languages and how this distribution is affected by different characteristics of source and target languages. These questions have not been answered yet and will be valuable to understand the information exchange between languages due to machine translation.

RQ2: "'Second, how does Google Translate change the collaboration and coordination pattern between human editors and machine intelligence?' *This is an interesting question, and I wonder if this is intentionally meant to be specific to Google Translate, or whether this is meant to capture a different dimension?"*

Thanks for the question. This is a good place that we can clarify our research approach. In short, the specific research context we examine is the introduction of Google Translate to Wikipedia, but we intend to make wider inferences about the role of machine translation in content production by studying the impact of Google Translate.

- Before the integration of Google Translate in January 2019, Content Translation tool in Wikipedia has other machine translation engines available to its editors, e.g., Yandex, Apertium etc. Those are mostly rule-based or statistical machine translation algorithmsthat were popular before

2016. Neural machine translation gradually outperformed rule-based and statistical machine translation and deep learning become the dominant approach in the field of machine translation. - Google Translate represents state-of-the-art neural machine translation technology as it has both a larger coverage of language as well as better translation quality (as measured by BLEU score) for a wide range of language pairs. Ex ante, we were not sure whether the introduction of Google Translation make any difference in the Wikipedia ecosystem since it is not the first machine translation tool available in Content Translation. The sharp and discontinuous increase in total new translated articles we observed in our preliminary analysis(See Figure 1 in Project Description) demonstrates that it has a significant impact on the ecosystem and call for further investigation.

- By leveraging the exogenous variation in the natural experiment of Google Translate integration in January 2019, we will be able to make reliable causal inference about the impact of a better machine translation service. Because the timing of this partnership between Google and Wikimedia is exogenous to the outcomes variables we are interested in (e.g., number of translated articles, editing behavior, diffusion of local content, etc.), changes in quantities of interests before and after the event can be reliably attributed to the availability of Google Translate. The exogeneity of the Google Translate integration allow us to have causal interpretations for the inferred effect and it is the reason why we are very excited about this natural experiment and hope to make the most out of it for learning about the role of machine translation in narrowing knowledge gap between languages.

RQ3: *"Third, a large portion of each Wikipedia language edition is locally relevant and culture specific content. Does machine translation also help the exchange of local content?" I really like this question and its premise. I would like to go even further: How much does it help with the exchange of local content? What lessons can we learn from that and how can we improve this exchange? It would be also interesting to investigate some critical questions, e.g. what is the quality and completeness of the translations? What about propagation of updates on the source content, does this reach the translations? Or is translation often treated just as a number game, i.e. we count the number of articles but don't go into depth of the quality and coverage of the individual translation? What about changes in the translated content? Does it propagate to the source page? Etc. There are many questions around translations, and I would like to understand which ones this proposal is aiming at.*

We totally agree that there are many interesting questions around this aspect. We will incorporate your suggestions into our research plan. Particularly, quality and completeness are definitely aspects of translation we plan to examine in addition to “quantity” of translation. Completeness is relative easier to measure since translation is recorded at the level of article “section” in Wikimedia translation API. We can map sections in target articles back to sections in source articles and compute metrics for translation completeness. Quality of translation is relatively more difficult to measure, and it is especially a challenge in the multilingual context. Linguistic features of article quality are difficult to compare across languages. For now, we plan to measure translation quality by examining the difference in number of references, number of hyperlinks, number of images between source and target articles. Later in the project, we plan to develop more sophisticated measure of translation quality.

For the third research question, we are especially interested in the subset of Wikipedia articles that are about concepts or entities related to the languages or to the geographic regions where the language is spoken. This is what we refer to as locally relevant and culture-specific content or, in short, local content.

Expanding the availability of local content has been a focus of Wikimeida Foundation. Regarding the role of machine translation in growing local content, we are interested in will machine translation help spread of local content across different language editions once it has been created in one language of Wikipedia. To investigate this question, we will first need to classify each article as either local content or non-local content. Miquel-Ribé and Laniado (2019) has done great work in this regard, and we will follow approach. The classification of local content will enable us to investigate its diffusion and many interesting facets of this process.
Comments in “Revisions” section

I would suggest to detail how you are planning to leverage this unique natural experiment, and how you will apply techniques from econometrics modelling, causal inference, and natural language processing to answer the poised questions, and to also see a more sharper formulation of the questions being answered.

Thanks for the suggestion. Due to the space limit, we were very brief in in Stage I proposal. Hope our response to the above questions will provide more information and greater clarity. Further information please see the Project Description in this current proposal.

I would also suggest to have outputs beyond a paper analyzing the situation, but also develop something the community can use in order to continue improving the situation, e.g. dashboards, a gap detector, a way to explore the data being gathered, etc.

This is a great idea. The direct outcomes of this study are empirical observations and insights that hopefully will be able to guide policy and design of machine translation tool to better serve the Wikipedia communities. We do hope to actively engage with the communities and actual users of Content Translation to both help guide our research question and analysis as well as eventually disseminate our findings that could be helpful for improving knowledge equity.

Reviewer 3

As mentioned before, the proposal is very compelling but falls short in explaining how the proposed methods can be used to answer the research questions. The overall recommendation is to a) re-scale ambitions to a more feasible study, e.g., a pilot, and b) better explain which methods to use and why in view of it.

Thanks for the suggestion. Indeed, we were very brief in Stage I proposal due to the space limit. In the Project Description of this current proposal, we provide more information about how we apply the proposed method to answer our research questions.

Regional Committee and Wikimedia Communities

We are extremely grateful for the insightful comments from regional committee and Wikimedia communities. Those inputs bring us both new insights and vindications of some of the considerations that inspire us to start the project. Specifically, some key messages we learn from the feedback include:

1) machine translation has been used extensively in a lot of Wikipedia communities; 2) it is especially important for smaller and under-resourced language projects; 3) knowledge equity and disparity are an important consideration when it comes to machine translation tool.

One of the comments from regional committee mentioned that “However, the aims and research questions are already somewhat biased towards an unfounded optimism in Google Translate”. Thanks for the suggestion and reminder. We will be mindful about any potential unfounded optimism or pessimism and consider both positive and negative impact when investigating the role of machine translation in content production on Wikipedia. We do have some hypothesis pointing to the potential undesirable outcomes. For instance, if Google Translate accelerate or attenuate the domination of English language and its point of view in online knowledge repository is still an open question. To empirically examine this question, we need to investigate if machine translation mostly supports knowledge outflow from a few major language editions like English and French or it promote a bidirectional and hence more mutual information exchange between different language editions of Wikipedia.

Project Description

Knowledge gap, often known as disparity in distribution for 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 and part of the essential infrastructure of knowledge repositories in digital space, Wikipedia also suffer with the knowledge gap problem (Graham et al. 2014, Zhu et al. 2020). As pointed out by 2030 Wikimedia Strategic Direction (Zia et al. 2019), 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-resourced 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 another language edition 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 allow editors to translate content 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 preliminary 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.

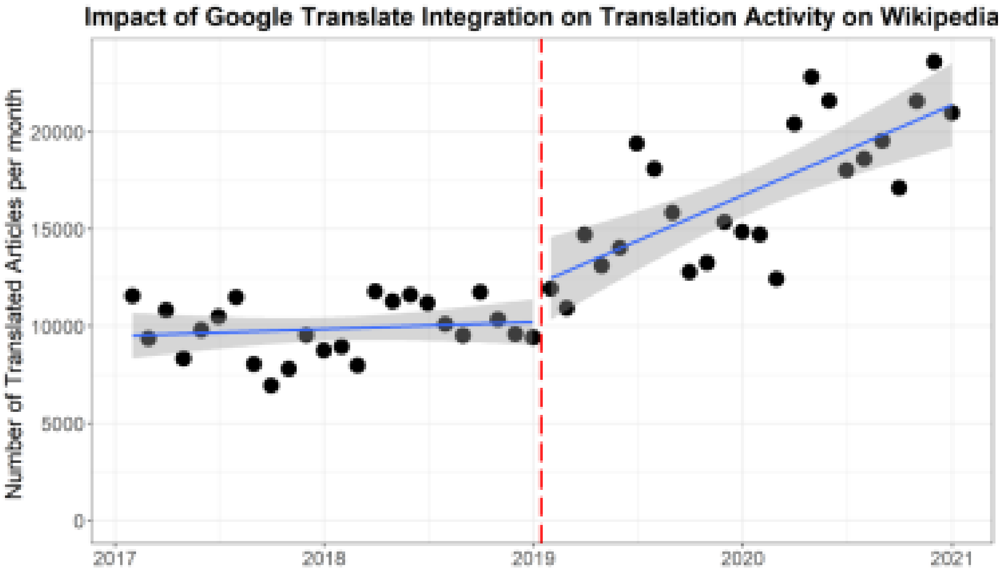

Figure 1: Google Translate was integrated to the Content Translation tool of Wikipedia in January 2019. It has an immediate impact on translation activity on Wikipedia

This partnership between Wikipedia and Google Translate presents us with a great opportunity to gain a better understanding of how a new technology enabled by state-of-the-art deep learning techniques may change the content production model on Wikipedia and if it can narrow the knowledge gap in social technical systems. Before the integration of Google Translate in January 2019, Content Translation tool in Wikipedia has other machine translation engines available to its editors, e.g., Yandex, Apertium etc. Those are mostly rule-based orstatistical machine translation algorithms that were popular approach before the wide application of neural network in machine translation. Since around 2016, neural machine translation begins outperforming rule-based and statistical machine translation in providing translation of higher quality. Google Translate represents state-of-the-art neural machine translation technology as it has both a larger coverage of language as well as better translation quality (as measured by BLEU score) for a wide range of language pairs. However, ex ante, we were not sure whether the introduction of Google Translation will make any difference in the Wikipedia ecosystem since it is not the first machine translation service available in Content Translation. The sharp and discontinuous increase in total new translated articles we observed in Figure 1 demonstrates that it clearly has a significant impact on the ecosystem and call for further investigation.

Through the natural experiment of Google Translation integration, we would like to learn about the role of machine translation for content production in open collaboration system more broadly. By leveraging the exogenous variation in the introduction of Google Translate, we will be able to make reliable causal inference about the impact of a better machine translation service. Because the timing of this partnership between Google and Wikimedia is exogenous to the outcomes variables we are interested in (e.g., number of translated articles, editing behavior, diffusion of local content, etc.), changes in quantities of interests

before and after the event can be reliably attributed to the availability of Google Translate free of unobserved confounders. The exogeneity of the Google Translate integration allow us to have causal interpretations for the inferred effect and hence provide credible policy and practical implication.

We aim to answer three sets of closely related research questions regarding the impact of Google Translate on Wikipedia. First, how does a better machine translation service affect the pattern of information exchange between different languages editions of Wikipedia? Going beyond the increase in total number of translated articles that we have observed in Figure 1, a lot of nuances in the pattern of source-to-target language translation could have different practical implication. For instance, does machine translation mostly support knowledge outflow from a few major language editions like English and French to smaller wikis? Or does it support a bidirectional and hence more mutual information exchange between language pairs on Wikipedia? It is possible that part of the impact of Google Translate is further enhancing the domination of English language and its point of view in online knowledge repository. We plan to use counterfactual estimation in the panel data setting to infer the causal impact of Google Translate. 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 adopted 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.

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? Google Translate as a new technology that automat part of human tasks bring a new working mode for Wikipedia editors. They are freed from manually translate the source article from scratch and instead need to review, revise, and improve the results from machine translation. We are interested in the “division of labor” between human and machine when they together translate 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 each section (i.e., using machine translation systems or translate by themselves). We can compute the completion ratio of translation of machine/human as number sections translated by machine/human divided by total number of sections in source article. Moreover, Wikimedia Foundation records detailed textual content for both raw machine translation output and the final text for the section after human editing. By comparing these two parallel text corpora using techniques from natural language processing (e.g., text embedding and text distance), we aim to gain insights on what kind of human editing was involved to complement machine translation results and if the current system incentivizes the desirable behavior.

Third, a large portion of articles from each Wikipedia language edition is locally relevant and culture specific content (or, in short, local content). Does machine translation help the diffusion such local content? Specifically, for this research question, we are interested in the subset of Wikipedia articles that represent its associated cultural context. Those are articles about concepts or entities related to the languages or to the geographic regions where the language is spoken. Expanding the availability of local content has been a focus of Wikimedia Foundation. To investigate this question, we will first need to classify each article as either local content or non-local content. Wikipedia Diversity Observatory (Miquel-Ribé and Laniado 2019, 2020) has done great work in this regard. We will make use of the data set they have made public and classify articles base on a rich set of features. The identification of local content will enable us to further investigate its diffusion and many interesting facets of this process. For instance, do the articles created with machine translation tend to be about universal concepts or concepts not represented or not shared across languages? We hypothesize that machine translation may also play an important role in helping the spread of local content across different language editions once it has been created in one language of Wikipedia.

References

Graham M, Hogan B, Straumann RK, Medhat A (2014) Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty. *Annals of the Association of American Geographers* 104(4):746–764.

Miquel-Ribé M, Laniado D (2019) Wikipedia Cultural Diversity Dataset: A Complete Cartography for 300 Language Editions. *ICWSM* 13:620–629.

Miquel-Ribé M, Laniado D (2020) The Wikipedia Diversity Observatory: A Project to Identify and Bridge Content Gaps in Wikipedia. *Proceedings of the 16th International Symposium on Open Collaboration*. OpenSym 2020. (Association for Computing Machinery, New York, NY, USA), 1–4.

Zhu K, Walker D, Muchnik L (2020) Content Growth and Attention Contagion in Information Networks: Addressing Information Poverty on Wikipedia. *Information Systems Research* 31(2):491–509. Zia L, Johnson I, Bm JM, Redi M, Saez-Trumper D, Taraborelli D (2019) Knowledge Gaps--Wikimedia Research 2030.

Budget Description

Ph.D. student: The project will involve one PhD student assisting data collection and data analysis. I already have the student enrolled in our PhD program and need to seek one year of funding for her. At the university I am working, the supervisor (the grant applicant) needs to share the cost for about $20,500. The supplemental private health insurance is included in this cost.	$20,500
Summer Support for applicant: To support research activity of the applicant (a faculty member) during summer period.	$12,500

Research equipment: one-time expenditure to purchase a powerful research computer for data collection, data processing, and modeling. The research computer needs to have 1) a large RAM (e.g., 128GB); 2) a large storage space (at least several TB), and 3) a Graphic Processing Unit (GPU). A large RAM is necessary as I will need to use it to process large-scale structured data and unstructured text (textual content for more than 1 million translated article). A GPU is needed as we will need to train neural embedding models. Large hard drive for data storage is necessary as size of the raw data of the Wikipedia dump is large. It is important to keep all the raw data so the results can be later replicated. Note: The applicant of this proposal is in his early career and does not have such a machine right now. This will be critical to executing the proposed project.	$4,500
Travel for academic conference: To present the findings of the study and get feedback, the applicant will attend one academic conference. The target conferences are the International Conference on Information Systems (ICIS) and Conference of Information Systems and Technology (CIST) among others. Based on the past experience, the expenditures are estimated to as following: - Conference registration: $500 - Airplane ticket, economy class: $1,000 - Hotel: $200/night*3 nights: $600 - Per diem: 4 days x $60/day: $240	$2,340
Open access publishing cost: The goal of the projects is to have two publications in academic journal. For the journals that I am targeting to publish (e.g., Management Science or Information System Research), the open access fee is $3,000 per article. Note: Those journals are not open-access journal and they do not charge for publishing. If we do not pay, the paper will only be available to institutional users who have subscription to the journals. The $3,000 fee will make the publication available to everyone even though the journal itself is not open access	$6,000
Total:	$45,840