Audiences × Wikipedia 2030 Product Planning **2019**

Augmentation

We want the sum of all knowledge to be available to everyone in the world. We also want the process to assemble that knowledge to be inclusive, balanced, and safe for all participants.

01
Content Generation

02
Content Curation

03
Governance

04
Machine Translation

Published
January 2019

Authors
Marshall Miller
Jess Klein

Contributors
Dan Garry
David Goldberg
Aaron Halfaker
Danny Horn
Jon Katz
Josh Minor
Jonathan Morgan
Toby Negrin
Margeigh Novotny
Nirzar Pangarkar

The Wikimedia movement wants the sum of all knowledge to be available to everyone in the world. We also want the process to assemble that knowledge to be inclusive, balanced, and safe for all participants. But there is too much knowledge needed, in too many languages, for humans to do this alone. As an example, if we assume that a Wikipedia that covers a substantial amount of knowledge has 2 million articles (likely a low estimate), and we believe that 300 languages should have access to that knowledge, we should expect there to be 600 million articles. There are currently only 48 million articles[1], which is 8% of the way there. There are simply not enough potential human editors, especially in smaller languages, to get there. Whether or not we believe that long-form articles will be the medium of the future, this illustrates the problem we face.

Augmentation for contribution activities is our path to closing these gaps. Augmentation refers to any technology that helps humans do their work, and wikis have been using augmentation almost since their beginnings: Rambot created 34,000 articles from Census data in 2002[2], Twinkle has been automating repetitive tasks since 2007,[3] ClueBot has been reverting vandalism since 2011,[4] and the Content Translation tool has employed machine translation to generate content since 2016. Over the next three to five years, human editors will need to increasingly wield augmentation tools, especially those that incorporate artificial intelligence, to create content, curate content, and maintain a safe environment on the wikis. Artificial intelligence will not replace human editors -- it will allow human editors to focus on the most impactful and fulfilling work, and, if used correctly, will open up more avenues for more contributors.[5]

But although artificial intelligence is a powerful editing aide, it also has the potential to powerfully magnify the problems of bias and unfairness[6][7] that already exist in the wikis, and has the potential to discourage new editors.[8] Therefore, the role of human editors will change in the future to focus on wielding these tools safely to guard the wiki values that only humans understand.[5] In pursuing any augmentation technology, we should stick to the principles we apply to code and content: transparency and the ability for anyone to contribute. We should build closed-loop systems that essentially make augmentation "editable" by community-members, even non-technical ones. By making it possible for members of all communities to audit augmentation tools, contribute training data, flag errors, and tailor tools to their wikis, we will ensure that wikis are not unduly influenced by the smaller set of people who build the tools, while also opening up a new avenue of contribution.

In terms of capabilities we need to build, the Wikimedia Foundation should do two main things:

Build an infrastructure platform for many people to contribute augmentation tools.

Provide interfaces that make it possible for non-technical editors to apply, adjust and contribute to those tools.

The former would likely be pursued by the Technology department, while the latter would be pursued by the Audiences department. The Audiences work will create on-wiki tools that allow non-technical editors to record training data, identify errors in existing algorithms, and tune algorithms to fit their wiki's culture; surfacing those tasks as first-class wiki work that other editors can see. Through these interfaces, the shepherding of augmentation tools will become a new, major way of contributing that will ensure that machines are fair and healthy contributors to every wiki.

Assembling the platform and the interfaces that allow a feedback loop are the most important parts of this strategy -- more important than the particular applications of augmentation. That said, particular augmentation tools will generally fall into three aspects: content generation, content curation, and community conduct. We will need to develop design principles in each

of these aspects that ensure augmentation tools are transparent and editable; and that ensure augmentation respects the boundaries between human work and machine work. These principles should also govern the ways we incorporate augmentation resources from third parties not controlled by the Wikimedia movement, such as machine translation services.

And finally, in order to be successful with this strategy, we will need to continuously recognize and embrace augmentation as a major way to contribute to the wikis. We can do this through community capacity building, holding events, providing training, and encouraging discussion in the community.

Examples
Rambot (content generation)
Twinkle (content curation)
ClueBot (content curation)
SuggestBot (content generation)
HostBot (governance)
Bot approval process (governance)
ORES models in RecentChanges and Watchlist (content curation)
Content Translation tool (content generation)
Article Placeholder (content generation)

Areas of Impact
Wikidata[9]
ORES[10]
Experienced editors[11]
Volunteer developers[12]

Key External Factors
The rate of improvement to artificial intelligence, especially machine translation.[13]

Efforts by other tech companies to automatically translate English Wikipedia, or to otherwise make massive amounts of information available. [14]

The movement's ability to get top talent to work on these issues as staff or volunteers.[15]

01 Content Curation

Content curation is the aspect of wiki activity related to editing, refining, and cleaning up content that has been generated. The Wikimedia movement's ambitious aspiration to make the sum of all knowledge available to everyone in the world means that the movement has a tremendous amount of work to do with respect to making judgments about what information belongs, and how to organize, phrase, and cite it. Most of the hundreds of languages in the world have Wikipedias with less than 10% the number of articles that English Wikipedia has, and even the largest Wikipedias have serious gaps in terms of the depth of their articles, and the subject matter covered by their articles. As all that content gets added, the curation workload will increase beyond what humans are capable of doing.

Augmentation is a potential pathway to curating the massive amount of information needed in the Wikimedia projects. By applying algorithms and artificial intelligence in the right ways, human editors can be aided in making the most important judgments about the content in the wikis, allowing the content to be well-organized and reliable. This kind of human-machine partnership is not new in the wikis. Tools like Twinkle and Huggle have been helping to automate the tasks of reviewing recent changes and patrolling for vandalism since 2007 and 2008. ClueBot has been independently reverting vandalism since 2011. And in more recent history, ORES machine learning models have begun to surface the edits and pages most in need of attention.

As humans and machines work together to curate content, we can think about that interaction on a spectrum of how much work the human editor does and how much work the machine does. In some scenarios, the machine may just direct human attention to important curation needs. In other scenarios, the human may review tasks completed (e.g. edits reverted) by an algorithm. This paper explores some specific examples of content curation activities that can exist in the future, drawing from all along the spectrum of the human-machine partnership.

Augmentation strategy summary

To meet our movement's goal of making all the world's information available to everyone, we have more work to do than human editors can do alone. We need help in the form of augmentation, which is when humans and algorithms work together. Though augmentation in the wikis is not new, it will be a growing part of their future. To ensure that the contributions made by algorithms are productive, unbiased, and fair, we will need to stick to our movement's principles of openness, transparency, and the ability for anyone to contribute. We should build closed-loop infrastructure and interfaces that allow anyone to contribute new algorithms, and participate in training and tuning them. These principles would apply to augmentation as it is applied to content generation, content curation, and governing interactions between people.

Definition of content curation

Content curation is the activity of editing, refining, and cleaning up wiki content. This is in contrast with content generation, which is about adding new content, and with the interactions and communications between wiki editors (governance).

Aspiration

Making the sum of all knowledge available to everyone in the world is a tremendously ambitious goal. Today even the largest Wikipedias—such as the English wiki—have serious gaps in terms of the depth of their articles, and the subject matter covered by their articles. Assembling knowledge is about more than compilation—it means curation: making judgments about what information belongs, and how to organize, phrase, and cite it. Effective curation makes the sum of all knowledge more accessible, and also makes it more trustworthy. There is going to be too much curation work in the future for humans to do it unassisted.

Augmented content curation

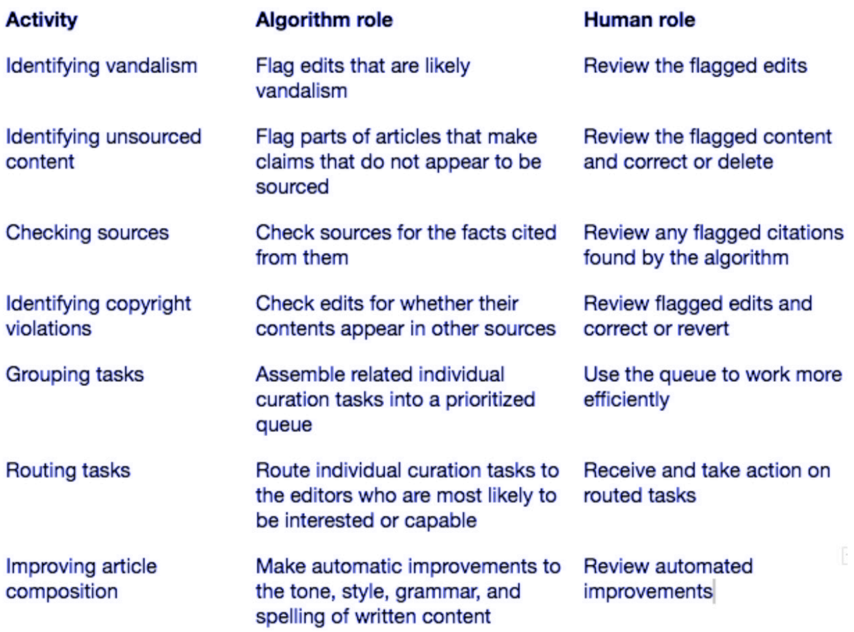
Augmentation is a potential pathway to curating the massive amount of information needed in the Wikimedia projects. The effective application of algorithms and artificial intelligence can help human editors make important judgments about the organization and reliability of wiki content, and human-machine partnerships are not new. Tools like Twinkle (2007) and Huggle (2008) have helped automate the tasks of reviewing recent changes and patrolling for vandalism. ClueBot has been independently reverting vandalism since 2011. More recently, ORES machine learning models have begun to surface the edits and pages most in need of attention.

As humans and machines work together to curate content, we can think about that interaction on a spectrum of how much work the human editor does and how much work the machine does. In some scenarios, the machine may direct human attention to important curation needs, while in others the human reviews tasks completed by an algorithm (e.g. edits reverted). This table provides some specific examples of content curation activities that can exist in the future, drawing from all along the spectrum of the human-machine partnership.

Content curation strategy

The technical effort behind developing and deploying the human-machine partnership scenarios described above is only part of the challenge. The more important challenges involve defining technical frameworks and design principles that ensure that algorithms are implemented as forces for unbiased and fair curation.

Human-driven content curation necessarily reflects human biases. Because bias and unfairness already exists in Wikimedia projects, algorithms that learn from human work have the potential to magnify and exacerbate those problems. For instance, one human editor's preference for writing with a certain style might accidentally exclude edits done by members of other demographic backgrounds. The Wikimedia movement should confront this with the same principles that have led to our success in the past: transparency and the ability for anyone to contribute.



Activity	Algorithm role	Human role
Identifying vandalism	Flag edits that are likely vandalism	Review the flagged edits
Identifying unsourced content	Flag parts of articles that make claims that do not appear to be sourced	Review the flagged content and correct or delete
Checking sources	Check sources for the facts cited from them	Review any flagged citations found by the algorithm
Identifying copyright violations	Check edits for whether their contents appear in other sources	Review flagged edits and correct or revert
Grouping tasks	Assemble related individual curation tasks into a prioritized queue	Use the queue to work more efficiently
Routing tasks	Route individual curation tasks to the editors who are most likely to be interested or capable	Receive and take action on routed tasks
Improving article composition	Make automatic improvements to the tone, style, grammar, and spelling of written content	Review automated improvements

Concretely, for algorithms to successfully participate in content curation, the following standards should be adopted:

Algorithms should be able to be created and deployed by anyone. For example, if the creators of algorithms for identifying vandalism are all from the English-speaking world, the algorithms might do a poor job at identifying vandalism in other languages.

The provenance of Algorithmically-produced work should be clearly indicated. For example, the user who made an edit should be notified if an algorithm reverts it.

Algorithmically-produced works should always include a human in a "closed loop" of editing, improvement and auditing. For example, in the Recent Changes feed, ORES models suggest edits that need attention, but do not automatically take action. Further, there is no ay for humans to flag ORES suggestions that are incorrect so that ORES can be improved.

Shepherding, tuning, and training algorithms should be an important wiki role that non-technical editors can take on. Any editor should be able to wield, adjust, and provide data for improving augmentation. This work should "count" as wiki work, as actual edits, and editors should find their way to this augmentation niche. For example, flagging ORES judgments as incorrect should count as an edit.

As described in the overall augmentation theme strategy, the Wikimedia Foundation should do two concrete things to make the above possible:

One.

Build an infrastructure platform for many people to contribute augmentation tools, coupled with Wikidata (or something like it) to serve as a repository of facts.

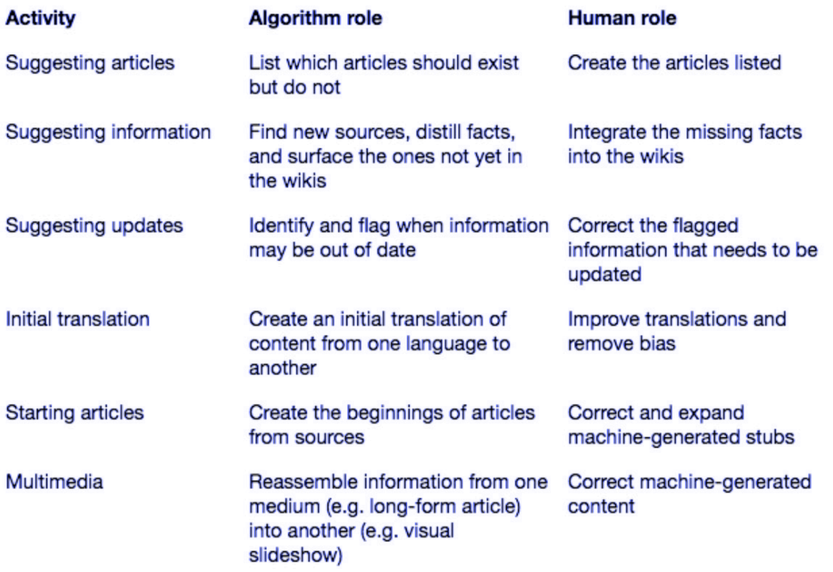
Two.

Provide interfaces that make it possible for non-technical editors to adjust and contribute to those tools.

02 Content Generation

Content generation is the aspect of wiki activity related to adding new facts, writing, translations, or images to the wikis. The Wikimedia movement's ambitious aspiration to make the sum of all knowledge available to everyone in the world means that the movement has a tremendous amount of work to do with regard to content parity across all wiki projects. Most of the hundreds of languages in the world have Wikipedias with less than 10% the number of articles that English Wikipedia has, and even the largest Wikipedias have serious gaps in terms of the depth of their articles, and the subject matter covered by their articles.

Augmentation is a potential pathway to closing the gaps described above. By applying algorithms and artificial intelligence in the right ways, human editors can be assisted in generating the most important content for the wikis, allowing us to close the most important gaps fastest. This kind of human-machine partnership is not new in the wikis.



Activity	Algorithm role	Human role
Suggesting articles	List which articles should exist but do not	Create the articles listed
Suggesting information	Find new sources, distill facts, and surface the ones not yet in the wikis	Integrate the missing facts into the wikis
Suggesting updates	Identify and flag when information may be out of date	Correct the flagged information that needs to be updated
Initial translation	Create an initial translation of content from one language to another	Improve translations and remove bias
Starting articles	Create the beginnings of articles from sources	Correct and expand machine-generated stubs
Multimedia	Reassemble information from one medium (e.g. long-form article) into another (e.g. visual slideshow)	Correct machine-generated content

Definition of content generation

Content generation is the activity of adding new facts, writing, translation, or media to a wiki. This is in contrast with content curation, which is about vetting, editing and organizing content, and with the interactions and communications between wiki editors (governance).

Augmented content generation

Augmentation is a potential pathway to generating the massive amount of information needed to close gaps in the Wikimedia projects. This kind of human-machine partnership is not new in the wikis. Tools like Rambot (2002) generated 32,000 stub articles in English Wikipedia using Census data, and in 2018, algorithms automatically translate thousands of articles. On the horizon are technologies like Quicksilver, which detects fact-related entities in news articles and collates them for human editors to turn into articles.

As humans and machines work together to generate content, we can think about that interaction on a spectrum of how much work the human editor does and how much work the machine does. In some scenarios, the machine may suggest a task that the human editor would execute. In other scenarios, the human may edit and improve on algorithmically-generated. The table provides specific examples of possible content generation activities that drawing from the full spectrum of human-machine partnership.

Content generation strategy

The technical effort behind developing and deploying the human-machine partnership scenarios described above is only part of the challenge. The more important challenges involve defining technical frameworks and design principles that ensure that algorithms generate high-quality and unbiased content. are implemented as forces for unbiased and fair curation.

Human-driven content curation necessarily reflects human biases. Because bias and unfairness already exists in Wikimedia projects, algorithms that learn from human work have the potential to magnify and exacerbate those problems.

Concretely, for algorithms to successfully participate in content generation, the following standards should be adopted:

Algorithms should be able to be created and deployed by anyone. For example, if the creators of algorithms for suggesting notable female scientists are all from the English-speaking world, it is possible that the algorithm neglects notable female scientists from outside the English-speaking world.

The provenance of Algorithmically-produced work should be clearly indicated. For example, if an article is primarily generated through machine translation (such as through the Content Translation tool), that characteristic should be clear to readers and editors. This will increase transparency, and potentially encourage human editors to improve the result.

Algorithmically-produced works should always include a human in a "closed loop" of editing, improvement and auditing. For example, in the Content Translation tool, users are required to review and correct the automated translation done by the algorithm.

Shepherding, tuning, and training algorithms should be an important wiki role that non-technical editors can take on. Any editor should be able to wield, adjust, and provide data for improving augmentation. This work should "count" as wiki work, as actual edits, and editors should find their way to this augmentation niche. For example, if a topic suggestion algorithm existed, editors should have a way to assess the notability of those topics in such a way that the algorithm could improve.

As described in the overall augmentation theme strategy, the Wikimedia Foundation should do two concrete things to make the above possible:

Build an infrastructure platform for many people to contribute augmentation tools, coupled with Wikidata (or something like it) to serve as a repository of facts.

Provide interfaces that make it possible for non-technical editors to adjust and contribute to those tools.

03 Governance

In order to meet our movement's goal of making all the world's information available to everyone, we have more work to do than human editors can do alone. We need help in the form of augmentation, which is when humans and algorithms work together. Though augmentation in the wikis is not new, it will be a growing part of their future. To ensure that the contributions made by algorithms are productive, unbiased, and fair, we will need to stick to our movement's principles of openness, transparency, and the ability for anyone to contribute. We should build closed-loop infrastructure and interfaces that allow anyone to contribute new algorithms, and for even non-technical editors to participate in training and tuning those algorithms. These principles would apply to all types of augmentation, whether it is in the aspect of content generation, content curation, or governing interactions between people.

Governance is a word meant to capture a broader scope than "Code of Conduct". It refers to all the ways that people interact with each other on wiki projects, in both constructive and unconstructive situations. Current newcomers rarely contribute past their initial edits because of bad reactions to quality control mechanisms, algorithmic tools (bots) or policy. We see augmented governance practices as the vehicle that will safeguard, and simultaneously empower the Wikimedia community to become that desired safe haven for knowledge discourse through a set of human-centered principles.

Aspiration

The Wikimedia movement wants the assembly of all the world's knowledge to be inclusive, balanced, and safe for all participants. Current newcomers rarely contribute more than initial edits because of bad reactions to quality control mechanisms, algorithmic tools (bots) or policy. We see augmented governance practices as the vehicle that will safeguard, and simultaneously empower the Wikimedia community to become that desired safe haven for knowledge discourse through a set of human - centered principles .

Augmented governance

Wikipedia wants to attract and retain a person who edits in good faith and has a relatively high quality of edits. Currently the interplay between augmentation and governance has been explored through bots such as HostBot, which welcome new contributors to Wikipedia, and through processes such as the Bot Approval Process. The reception of these machine-generated greetings and process guidance has been initially cold due to impersonal and sometimes ineffective guidance. However, the way forward is leveraging the technology to create a partnership spectrum in which bots and algorithmic tools support and enhance the role of humans within the Wikimedia communities.

This table provides some specific examples of content curation activities that can exist in the future, drawing from all along the spectrum of the human-machine partnership.

Governance Strategy

In the 1971 film RoboCop, a picture is painted of a dystopian city on the decline that is dealt with by employing an army of robots to brutally police the city. This is not the future of governance that we want at Wikimedia. The majority of AI systems and related tools that are being created in the world today are being put in place with minimal oversight, few accountability mechanisms and little research into their broader implications. As the table indicates, artificial intelligence is potentially a powerful editing aide, but each advantage is also has the potential to powerfully magnify the problems of bias, unfairness, and hostility to new editors that already exist in the wikis. Currently, there are no internally agreed-upon methods to measure and assess the social implications of governance augmentation. Therefore, we will need to develop ways to measure, audit, analyze and improve them. There are two ways that we can concretely do this: First, the Wikimedia community can generate a set of guiding principles to ensure that the wikis are authentic and sincere representations of the worlds' knowledge. These principles should focus on human-centered AI, learnability as a core metric and transparency. The Human-centered AI nudges and opens the door to human connection instead of closing it. In so doing, Wikimedia augmentation champions you—your security, privacy, and the quality of your online activity within our tools. Learnability as a core metric means optimizing Wikimedia algorithms for learning and assessing their capacity to do so in terms of how well it assisted in the learning experience. Transparency refers to a commitment to presenting the provenance of any augmentation tool's introduction and use.

Second, developers and administrators will create and utilize open governance algorithms. If anyone can contribute to these tools, and if Wikimedia makes it possible for non-technical editors to watchguard them so that they aren't allocation or representation biased, we have the opportunity to define the role of human computer interaction within the context of governance within the communities that we create and run.

Guiding Questions

What do WE mean by fairness?

How might augmentation help us detect 'fake news' and misinformation campaigns from powerful malicious actors, like nation states?

How can we ensure that individual editors (new, experienced) feel that their contributions are still valued in a wiki where so much is created and curated by machines?

How can machines enhance the governance work done by editors?

03

Machine translation

Machine translation can contribute to the Movement's goal of making the sum of all knowledge available to everyone in the world by making the work done in one language available in others. Such an effort would dramatically accelerate the transmission of information, making more knowledge available and freeing contributors to work on original research. However, third party machine translation by the likes of Google and Facebook represents a threat to our model, potentially promotes English dominance, siphons traffic from our wikis, offers a poor experience, and discourages global contribution. We cannot ignore this change in the Internet's climate, but must adapt to it. No matter how fast humans create or translate content, we will not be able to create or, crucially, update content faster than a competitor who uses machine translation. We need to adopt a proactive, long-term approach to the widespread use of machine translation that serves our users, aligns with our values, and supports our ecosystem. If we do this correctly, machine translation could represent the ingredient that allows us to realize our vision of global knowledge sharing.

Machine translation is here

Imagine logging onto the internet, conducting a search, and seeing no results in your language. It is hard for English speakers to imagine, but this is the experience of many Internet users today.

Lack of relevant content was cited as one of the top reasons people don't read online by our New Reader's research. However, automated translation is changing how people read, write and relate to each other. In late 2016, Google announced that it was now using neural networks to power it's translation, as these had quickly surpassed traditional, algorithmic models. As a result, content translations for over 100 languages are now just 1-click away. As of 2017, Google Translate served 200M users a day and, as of 2018, translates 143 billion words a day. Meanwhile, the Facebook platform now performs 6 billion translations a day.

As impressive as this is, machine translation as it is deployed by third parties such as Google and Facebook is a threat to our model, our philosophy and our traffic. For some time, we have expected Google or someone else to provide translated versions of Wikipedia if we didn't. It is now clear that Google will be implementing a pilot of this in Bahasa Indonesian very soon, serving Wikipedia pages translated into Bahasa Indonesian from their English versions if a native language version doesn't exist.

We cannot ignore this trend, but must adapt to it. We are currently working with Google to ensure that translated pages provide an option to modify the translation and save it as a page on Bahasa Indonesian Wikipedia. However, this short-term solution may lead to static forks of pages that do not update over time because there are no human editors to attend to them or readers that know about them.

We need to adopt a proactive, long-term approach to machine translation that serves our users, aligns with our values and supports our ecosystem.

Short Term Necessity

If Google's pilot is successful, we can expect it to roll out in other languages. Google search users will be offered native language Wikipedia articles that have been machine-translated from the English language version. Internet wide deployment of this strategy would promote the epistemological dominance of English, divert traffic away from our wikis, offer an inconsistent experience and handicap our greatest strength: the invitation to contribute. Ideally Google would let us control this experience, but that is not an option and our licensing does not allow us to enforce it.

Instead, we are working with Google to address the issues of experience and community health. We are asking them to ensure that wikipedia users know that the content is not written by humans and offer them a way to modify the translation for addition to the wiki in their language. In the long-term, we hope they will choose to use articles from other languages when appropriate, but this is not currently on the table.

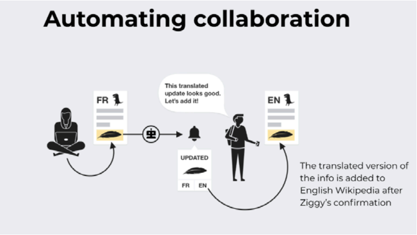
A Path to Leveraging Machine Translation

Google's efforts will change how people read Wikipedia globally. To take advantage of the opportunity, we will have to adapt our current approach from one in which content translation is a single fork of content, to one in which content flows readily from one wiki to another.

For example, today, the article about Genetically Modified Organisms (GMOs) might be brought from English to Hebrew. However, from this point on, the development of the two articles is forked. In English, several paragraphs might be added about the history of GMOs. In Hebrew, someone might add a paragraph about the economics of GMOs. At this point, neither wiki is benefiting from the scholarship of the other.

As articles change, their counterparts in other languages should have the option to import the new material. For example, editors would be informed about changes to articles in other languages (sections added, facts, etc).

Here is a symbolic flow of what that might look like:

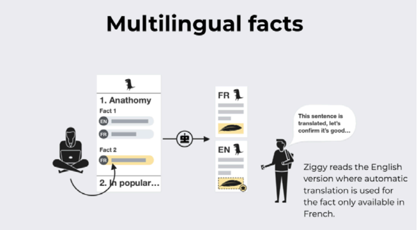


Automating collaboration

The translated version of the info is added to English Wikipedia after Ziggy's confirmation

(From Pau Giner's "A Multilingual Wikipedia")

Another approach would be to focus on generating facts which can be migrated from language to language via machine translation. Whether or not Wikidata, already a global repository, is used as a semantic storage and mediation platform between wikis is somewhat contested.



Multilingual facts

Ziggy reads the English version where automatic translation is used for the fact only available in French.

(From Pau Giner's "A Multilingual Wikipedia")

In either approach, we are taking the best of augmentation: using machines to replicate existing efforts and bringing in humans to confirm. This will require a shift in the kinds of work that needs to be done and the kinds of people who work on the encyclopedia. For every 1 writer who has a book in their hand and cites it, there need to be 100 other editors whose job is to import the new content into the appropriate place in the destination wikis.

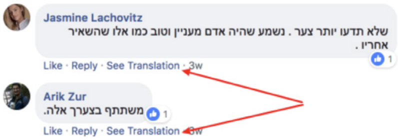
Machine translation will impact different wikis in different ways at different times. An automated translation approach would focus mostly on the wikis for which machine translation is good enough– as measured by our users. Other segments won't be as immediately impacted by machine translation.

The Glorious Future

The future is already here — it's just not very evenly distributed.

— William Gibson

Machine translation will make possible real-time collaboration across languages in any given digital scenario. Facebook and Google have already built early versions of such technology into their services. Here is an example where a woman was able to share the news of her father's death one-time in Hebrew and have her English-speaking and Hebrew-speaking friends discuss on the same thread:

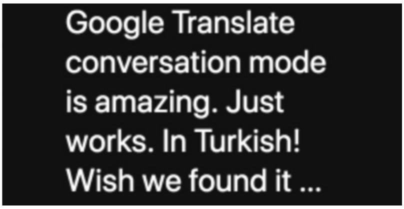


Jasmine Lachovitz
שלא תדעו יותר צער . נשמע שהיה אדם מעניין וטוב כמו אלו שהשאיר
אחריו .
1
Like · Reply · See Translation · 3w

Arik Zur
משתתף בצערך אלה.
1
Like · Reply · See Translation · 3w

Screenshot from Facebook, pulled November 1st, 2018

Facebook's next step would be to automatically translate other languages into the language of the reader. This began earlier this year in select places. Google implemented real-time translation of speech in 2015 and made this functionality a key feature of the earbuds they introduced in 2017.[13] A friend travelling around the world just posted this from Turkey:



Google Translate conversation mode is amazing. Just works. In Turkish! Wish we found it ...

Screenshot from Facebook, pulled November 9th, 2018

It is possible that we will eventually arrive at a single knowledge corpus, offering a variety of perspectives and written and consumed in many languages. In fact, such a system was recently proposed in a paper by Denny Vrandečić and shared on wikimedia-l.

The idea of a more global Wikipedia was also promoted by a community member in response to our announcement of a new machine translation service being available:

"I believe that if in the future it is possible to write the wikipedia articles in a universal language (I am not referring to English, Latin or something like that, but something for computer), and any change made in any language was visible to all…" This would be a more efficient way to amass the sum of human knowledge, but whether or not this idea is actually feasible will depend in large part on the interests of our communities. The time for that conversation is far away.

Happening Now

Today, Amisha, a biology student in Indonesia, looks up Cystoseira baccata, a species of brown seaweed, on Google. There is no Wikipedia article in Bahasa Indonesia for this topic, so Google provides the Wikipedia article in English in the results page.

There is an offer to translate it. If she instead clicks on the link, she is taken to an article in English and her browser may or may not offer to translate it.

There is a Google effort to improve this experience. In the near future, if Amisha conducts the same search, Google will by default offer up a translated version of the English page, hosted by Google with a Google header, yet with of our site's appearance and branding.

Here is one possible view:

This offering has been tested, and Google has suggested that readers reacted very positively to it. They plan on doing this for any page in which there isn't a suitable Bahasa Indonesian article. Bahasa Indonesian is the only language they are currently applying this to, because it is a fairly well-structured and easy-to-translate language. However, it is obvious that Google's ambitions do not stop at Indonesian or with Wikipedia. Their goal is to make the world's knowledge available in every language on the web.

Guiding principles

To serve a more adaptive, fluid strategy for our engagement with widespread machine translation, here is a set of principles we stand by:

More knowledge availability is good

Contributions from people all over the world are necessary for capturing valuable perspectives and because a single region cannot capture all the world's knowledge

Machines are neither good nor bad, but they require oversight.

All wikis should have the chance to export content other languages. Default translations should be derived from the most compelling content from whichever (translatable) language provides it.

Wikimedia Foundation decisions regarding MT need to be informed by what readers and community members want with respect to machine translation usage. Deciding based only on our values or perception has the potential to exacerbate unhealthy global power dynamics.

Risks and Open Questions

We still are many questions about machine translation that we will have to address over the coming months and years. The following section presents a set of risks and mitigations that are significant but over the horizon.

Risk: Disintermediation. Even if Google agrees to provide the users options to circle back to edit articles on their destination wiki, it is possible the mechanism will not be sufficient and local wiki communities will stagnate. Similarly, if users never reach Wikimedia-hosted site, we will not be able to fundraise.

Defense: Work with partners to establish the necessary entry points into our system As rapidly as possible fill out wiki content using machine translation to augment human effort.

Risk: Machine translation enforces bias. If we continue to rely on third party machine translation services, we are subject to the unknown biases built into those tools. In the case of language, some specific examples include using the male form of "prime minister" by default in gendered languages.

Defense: There aren't great options here. Work with tool providers to ensure that feedback is registered and to promote transparency. Using Wikidata's label system might help here. Another response might be to create our own, open machine translation tool.

Risk: Dependence on a third party tool makes us vulnerable. If a fact is created on Wikipedia and then is replicated 100x using machine translation, most of the work is now being done by a third party tool. This dependency means we risk losing the tool at any time and would be vulnerable to the demands of a few key players, such as Google or Yandex.

Defense: There aren't great options here. Work with tool providers to establish terms up front. Push towards storing content as structured data. Another response might be to create our own, open machine translation tool.

Risk: Low quality machine translation. If machine translation is pushed on users by a third party and creates incomprehensible text or even promulgates falsehoods, we risk harming our users and Wikipedia's reputation.

Defense: Work with Google to push back where necessary. Research tools and listen to user feedback. We are currently planning an investigation into how machine translation is perceived.

Summary

We need to multiply our efforts on the translation front in three areas:

Real-time translation: Google is using machine translation to provide meaningful default articles where the users' primary language version is non-existent (or maybe even a short stub), using the English article. Eventually move to some notion of article quality and translation quality to choose the source language. We use default articles as funnel to contribution. "Improve this translation". A key component of this is working with Google.

Syncing articles: Harness machine translation to improve cross-wiki collaboration. Move from forking articles to syncing articles, Expand the use cases so that contributors do not need to know more than one language to propagate changes from one wiki to another.

Potential Global Corpus: Eventually, we approach a state where there is a global corpus and when someone does the research to expand an existing Wikipedia entry, that research and writing doesn't necessarily need to be manually distributed in 200 separate places. We are able to examine whether we want different perspectives to be reflected along ethnic or philosophical lines, rather than along language lines (which have, by necessity,served as a convenient, but imperfect proxy for "perspective").

1 https://en.wikipedia.org/wiki/Wikipedia:Twinkle

2 https://en.wikipedia.org/wiki/Wikipedia:Huggle

3 https://www.mediawiki.org/wiki/ORES

4 https://en.wikipedia.org/wiki/Wikipedia:Twinkle

5 https://en.wikipedia.org/wiki/Wikipedia:Huggle

6 https://en.wikipedia.org/wiki/User:ClueBot_NG

7 https://www.mediawiki.org/wiki/ORES