How Wikimedia Understands the Text2Narrative Process

Leila Zia
Head of Research at the Wikimedia Foundation

Short Bio

Leila Zia is the Head of Research at the Wikimedia Foundation, the foundation that operates Wikipedia and its sister projects. Her research interests include quantifying and addressing the gaps of knowledge in Wikipedia and Wikidata, understanding Wikipedia's readers, and studying the contributor diversity on the Wikimedia projects. She has served as the general co-chair of The Web Conference 2019 (formerly known as WWW) and as the program committee co-chair of The Web Conference 2021. She received her PhD from Stanford University in Management Science and Engineering.

Text2Story21: A Conversation with Leila Zia

1 What can NLP/narrative extraction do for wikimedia?

1.1 What NLP does wikipedia use currently?

I'd like to start answering this question by sharing with you more about how Wikipedia, and for that matter other Wikimedia projects, work and how research and development is done on these projects. There are two aspects that are important to know:

- Who does research and development on Wikimedia projects?
- How do people surface the research and development they do?

Wikimedia Foundation is one of the major players in what many refer to as the Wikimedia Movement. However, we are not by any means the only players in this space. Research and development on the Wikimedia projects rely on a highly distributed network of volunteer developers and researchers, affiliates, user groups and organizations who contribute to the projects.

When we start talking about what NLP Wikipedia uses today, it's also important to remember how development on Wikipedia is done. There are two primary pathways. Development through the Wikimedia Foundation, and development through the volunteer developer communities who have access to public cloud resources and have the ability to write code that does things on Wikipedia. (for example, through bots.)

So, when I'm asked what NLP is used in Wikipedia today, I come to you from the place of humbleness to say that I won't be able to give you a comprehensive answer because of the highly distributed nature of research and development on Wikipedia and frankly because of how Wikipedia works.

With this in mind, let me answer your question: Generally speaking, Wikipedia today is offering more to the NLP research community than benefiting from it. You see a lot of use-cases of Wikipedia data for training models with applications outside of Wikipedia, than NLP being used in Wikipedia itself. This being said, we see some examples of NLP usage in Wikipedia:

In the Research team at the Wikimedia Foundation, and in collaboration with Leibniz University of Hannover we developed a neural network model to detect which statements in Wikipedia are in need of citations (in 3

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Finlayson (eds.): Proceedings of the Text2Story'21 Workshop, Online, 1-April-2021, published at http://ceur-ws.org

languages: English, Italian and French). The model is being used today by Citation Hunt, a tool developed by a dedicated Wikimedia volunteer developer, to flag statements that are in need of citations to the editors and encourage them to find and add appropriate citations[1].

It may also be interesting for you to know that Machine Translation is an NLP task that we're very interested in. The reasons that we have not been able to use some of the technologies developed outside of the Wikimedia Foundation are: proprietary code (what goes to production for Wikipedia needs to be open source), scaling challenges especially as we go to less common language pairs, and accuracy for encyclopedic use-cases.

OCR for Wikisource is another example.

1.2 Does wikimedia produce or envisage any kind of automatic summarization of multiple documents, showing timelines that link pages in one same story?

No, Wikipedia currently doesn't have this service/feature. The timelines are generally created by editors and manually.

We know that depending on which Wikipedia language the readers go to, somewhere between 15-45% of the time the readers want to read an overview or summary of the content that is available in a Wikipedia page [2].

In the Wikipedia world, the editors consider the lead paragraph of an article almost equal to its summary. Research that can shed light on whether the lead paragraph is a summary of an article and if not, in what ways it can be improved can be helpful.

So, yes. Providing summaries of articles to readers is an important avenue to explore. I would caution that the quantitative research in this space should be paired with qualitative research: what are the needs of the readers? What are they already satisfied with when they're trying to get an overview of one or more Wikipedia articles, and what's missing for them?

There are immediate places where a summary can potentially be used, depending on its quality. For example, through the "Page Preview" feature on Wikipedia where readers can preview the content of an article that is linked from the current article they're on without going to that article. There are potentially other applications for this kind work: for example in Abstract Wikipedia. An automatic summary of the article can be compared with the already existing article to identify missing components.

2 What can wikimedia do for narrative extraction?

2.1 Wikification is a powerful concept that can be useful for approximating grounding. Is wikimedia investing in becoming an Entity Linking standard?

Wikidata is the project where I would expect to see the development of entity linking standards. You can imagine Wikidata providing catalogues of entities to link to, for the central and widely used catalogues. What is important to keep in mind here is that such standards are best not to be developed in individual Wikipedia languages. Wikidata can allow for more entities to form, independent of language.

2.2 What other resources does wikimedia have that can be useful for this community of NarrativeExtraction from Texts?

Data. XML dumps is the place I recommend you start from. You can access them by going to dumps.wikimedia.org. They contain a complete copy of all Wikimedia projects. MediaWiki APIs are another asset for your community.

Compute resources. If you are working in low resource environments and you want to work on research projects that directly benefit the Wikimedia communities, check out Wikimedia Cloud Services to take advantage of a humble cloud resource environment for all.

Research and Software Grants. Wikimedia Foundation gives small grants for researchers to do research and development that can help improve Wikimedia projects.

The volunteer community of editors. If you intend to have research that is useful for Wikipedia, I highly encourage you to get in touch with your language community and get involved. Wikimedia editors, especially in

[1] Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli,*Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability*, 2019.

[2] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia, *Why the world reads wikipedia: Beyond english speakers*, 2018.

languages that receive less attention, are incredible assets for researchers as they have depth of knowledge about the Wikimedia projects and can be your partners.

2.3 Is wikimedia producing datasets that can help with your challenges? Do you encourage researchers to annotate wikipedia resources? Can these datasets be freely published?

Yes, releasing public data-sets is one of the ways we encourage the acceleration of research and development in specific research areas. WikiReliability data-set is a recent example of a data-set our team has released.

There may be applications outside of Wikipedia where annotating the Wikipedia articles can be useful. We generally encourage research to look into crowd-annotated data that already exists in Wikipedia and try to find ways to use that. This data is highly curated as Wikipedia editors care about the projects and attention to details is their strength. For example, many editors use specific templates to signal content issues. You can use those as high quality labels in many cases. (Check out the WikiReliability data-set linked above for an example.)

3 How does wikimedia handle dis-information and bias? Can narrative extraction help here?

You must have attacks that inject fake news in wikipedia. Is this solved organically? Do you use any AI tools for this? Is story generation of texts faced as a problem (due to fake news possibility) or understood as a viable solution to "replace" or complement current Wikipedia editors?

Wikipedia works based on the assumption that when given the opportunity, the majority of the people will choose to do the right thing. That is why the "Edit" button is available to the readers, and in some languages they don't even have to log in to edit. But it's important to know that Wikipedia is not naive. There are systems built in place, by the volunteer editors of Wikipedia, to assure quality checks and course correction.

Transparency: Every article has a History page where you can see which username and at what time has done what exact edit to the page. You can revert their edit, or help improve them. Wikipedia:verifiability calls for backing up the statements by reliable secondary sources. Wikipedia:consensus sets the decision making model on Wikipedia consensus based. Wikipedia:ContentForking prohibits the creation of forks when disagreements arise. You will need to stay around, discuss, and arrive at consensus with other editors.

With this in mind: Yes. Content moderation in Wikipedia is done primarily through Wikipedia volunteer editors. Every edit to Wikipedia is added to a backlog of edits that need to be checked, even when it goes live immediately. The use of AI for curbing the spread of disinformation on the projects is more limited today, although this is an active area of research for our team. Sophisticated models and tools do get used in other areas though, such as for identifying copyright violations. Examples: sockpuppet detection, controversy detection, mismatch detection, etc.

Story generation as a problem or a solution? Generally speaking, Wikipedia editors welcome the use of technology to improve their workflows. What they don't appreciate is the type of technology that is black box and they cannot scrutinize its quality and working. Remember that transparency is key in the Wikipedia world and that applies to models, too. There is also different tolerance and acceptance of more advanced technologies in the different Wikipedia languages. In smaller Wikipedia languages where the editor resources are more constrained, technology can be a big help and the editors appreciated it.

When it comes to text generation, however, across the board editors are concerned about the maintenance costs. It is one thing to develop a sophisticated model with high performance that can generate summary articles in a language, it's another thing to assure that in the highly distributed world of Wikipedia the model can be maintained for the coming decade(s).

4 Future Challenges

NLP-wise and in terms of Information Retrieval and Extraction, which challenges do you see the community could take?

Almost all the classical NLP tasks could be useful for the Wikimedia projects. Think about information retrieval to improve search, named entity recognition to enrich article annotation, automatic fact checking to help editors audit quality of references on Wikipedia, etc.

Open Source: If I can tell you one thing about the challenge that a project such as Wikipedia faces today is that a lot of research and development is done in ways that are not compatible with Wikipedia and its ecosystem. So first and foremost, I encourage you all to seriously consider licensing your code under a free license.

Open Data: When you do research, consider publishing the valuable data-sets that you produce, under a compatible free license.

Abstract Wikipedia is the youngest member of the Wikimedia projects and the research and development in this space can open up new pathways for more NLP applications in Wikipedia. Particularly, the project will require advancements in NLG, especially in many of the languages that we don't have any Wikipedia in today or we have a very small community of Wikipedia editors.

Misinformation, Disinformation and content quality on Wikipedia: Text alignment across languages has many applications, one of which is to be able to detect, at scale, where there may be signs of misinformation or disinformation, or simply where we can improve the quality of content.